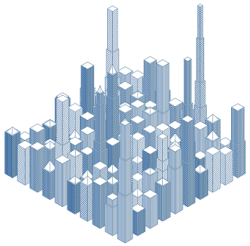
Chapter 6 Notes

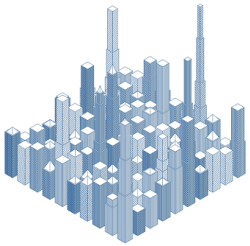
Samples and Observational Studies
STP-231

Arizona State University



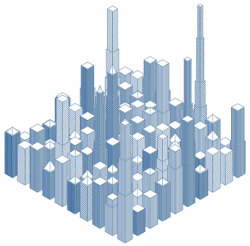
Objectives

- What is a population versus a sample?
- What is an observational study? An experiment?
- Randomness, bias, simple random samples (SRS)
- Other probability samples
- Sample surveys
- Comparitive observational studies



Noise versus Signal

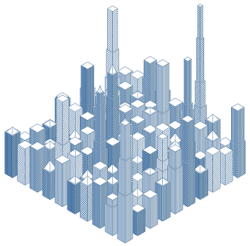
- Variability in data is inevitable, but it is important to the difference between noise and signal
- **Noise**: Is the variability we would expect by chance
- **Signal**: This is the variability due to a certain characteristic. In other words, this is a "real effect", not a statistical anamoly



Example

Below is a random sample of US lunch restaurant goers. Is there a difference between what men and women eat for lunch?

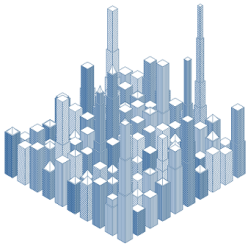
Diet	Men	Women	Total
Vegetarian	184	225	409
Non-Vegetarian	316	275	591
Total	500	500	1000



Observational Studies vs Experimental Studies

- **Observational study**: Record data on individuals without attempting to influence the responses

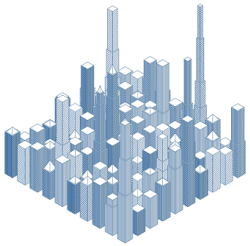
- For example, analyzing data from people who attended a certain university and their outcomes after the fact



Observational Studies vs Experimental Studies

- **Experimental study**: Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.

- For example, a study over whether or not a new toothpaste helps peoples tooth health places half of the participants (randomly) into a control group who receive a placebo toothpaste, and the treatment group who receives the real experimental toothpaste



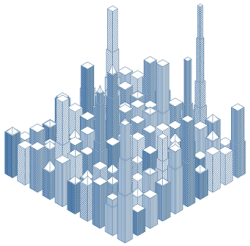
Observation vs Experimental

Observational Study

- We ONLY observe the subject
- Conclusions can be drawn about an association between two variable

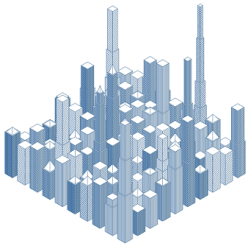
Experimental Design/Study

- We impose conditions on the subjects, i.e. we ask them to do something
- Conclusions can be drawn about cause & effect relationship between two variables



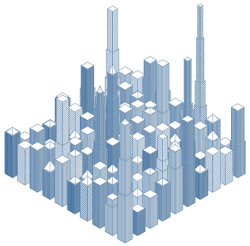
When to use each

- Use an experimental study if we wanna establish cause and effect
- Use observational study when:
  - Sometimes experimental studies are not ethical
  - Experimental studies take too long to complete, are too expensive, hard to properly design experiment
  - Sometimes causality is not that important, an association is enough
  - If we want to look at a historical study



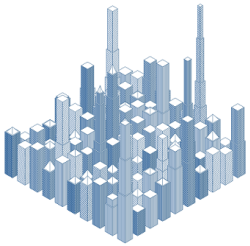
Observational Studies or Experiment?

- A 2013 Gallup study investigated how phrasing affects opinions of Americans regarding physician-assisted suicide. Telephone interviews were conducted with a random sample of 1,535 national adults. Using random assignment, 719 heard the question in form A and 816 the question in form B.
- The different forms worded the question different and 70% in form A said "should be allowed" versus 51% in form B.



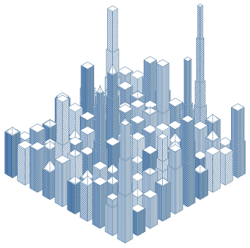
Observational Studies or Experiment?

- A 2013 Gallup study investigated how phrasing affects opinions of Americans regarding physician-assisted suicide. Telephone interviews were conducted with a random sample of 1,535 national adults. Using random assignment, 719 heard the question in form A and 816 the question in form B.
- The different forms worded the question different and 70% in form A said "should be allowed" versus 51% in form B.
- This is a randomized experiment because the groups were randomly assigned



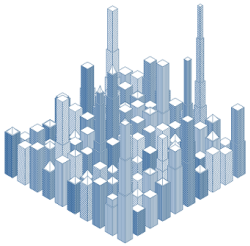
Components in Experimental Study

- **Response variable**: Outcome of interest
- **Treatment**: Conditions imposed on the subject. Treatment groups are the group of units in experiment who receive treatment, such as medication
- **Placebo**: The control group, which does not receive treatment. Sometimes placebo effect can lead to psuedo-treatment effect
- **Blind study**: Subjects do not know the treatment they are receiving
- **Double blind study**: Neither subject nor experimenter know which treatment subject receives
- **Panel bias**: A subject may behave differently if they participate in an experimental study



Confounding

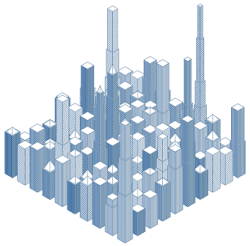
- Two variables are **confounded** when their effects on response variable cannot be distinguished. If the confounder affects both the treatment and outcome.
- Therefore, we cannot determine how one variable affects another if there is a third variable affecting both
- The "lurking" variable can affect the outcome, as long it is not associated with the "treatment". This can be accomplished if controlled for (we'll talk about this later)



Confounding Example

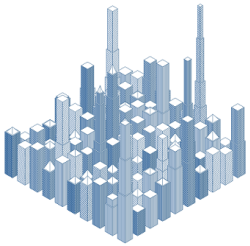
- Does cold weather cause people to be sick?
- Maybe it is because of the confounder that people go inside more when it is cold

Populations & Samples



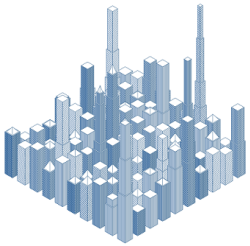
Population versus Sample

- **Population**: The entire group of individuals in which we are interested in but can't usually assess directly
- A **parameter** is a number summarizing a characteristic of the population.
- **Sample**: The part of the population we actually examine and for which we have data
- A **statistic** is a number summarizing a characteristic of a sample
- Parameters are denoted by Greek letters, i.e. μ for mean and lowercase English letters for statistics, such as $\overline{x}$ for sample mean



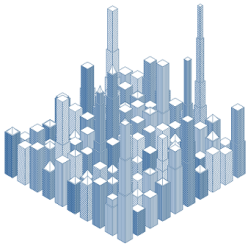
Role of Randomness in Sampling

- How do you select the individuals/units in a sample?
- **Probability sampling**: individuals or units are randomly selected; the sampling process is **unbiased**.
- **bias** is the systematic tendency for a study to favor certain outcomes.
- Think of bias as the accuracy of your study. How close is your estimate of a parameter (your statistic) to the true value of the parameter?



Bad Sampling

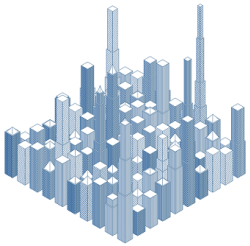
- Ann Landers summarizing 70% of responses of parents wrote in to say having kids was not worth it. But a random sample showed only 9% of parents believe this!
- This is because newsletters readers are not necessarily representative. This particular bunch was potentially disgruntled
- Another example. You are tasked with asking 200 people what they think about the legalization of marijuana and trying to predict how the state will vote. Is sampling the 200 from a college campus a representative sample? From a nursing home? From a mall?



Are these good samples?

Say you have some free time on a Tuesday night, and you want to estimate how much Netflix ASU students watch a week. You want a representative sample of the whole population, i.e. the student body. How good are these samples?

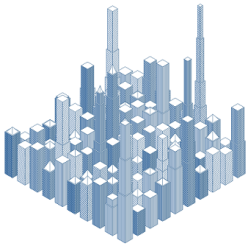
- You sample 100 students at Pete's trivia night on Mill avenue
- You sample 100 students at the library
- You ask 100 students at the gym
- You ask 100 people at the MU



The Simple Random Sample

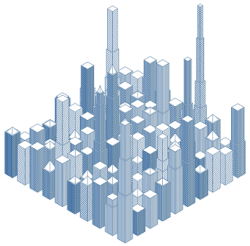
- A **Simple random sample** (**SRS**) is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample. $\binom{N}{n}$ possible samples with n/N probability of being the sample drawn. All the possible samples of size n have the same chance of being the sample drawn
- How do we draw an SRS? Usually we assign random numbers to every unit and draw till we get the desired size. Without a computer, we could draw from a hat, or use a table.
- Example in *R*:

```
set.seed(12296)
names<-randomNames::randomNames(35)
our_sample<-names[sample(35,10)];our_sample
```



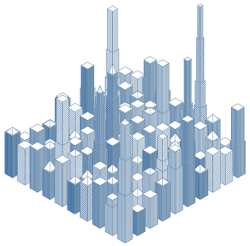
The Simple Random Sample

- How do we draw an SRS? Usually we assign random numbers to every unit and draw till we get the desired size. Without a computer, we could draw from a hat, or use a table.
- Example. You have the following 8 cities: 1. Boston, 2. New York, 3. Houston, 4. Phoenix, 5. San Diego, 6. Chicago, 7. Philadelphia, 8. Las Vegas
- You are given the following numbers: 22838 26302 47615
- Read the numbers from left to right until you select four cities



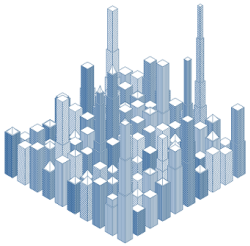
The Simple Random Sample

- How do we draw an SRS? Usually we assign random numbers to every unit and draw till we get the desired size. Without a computer, we could draw from a hat, or use a table.
- Example. You have the following 8 cities: 1. Boston, 2. New York, 3. Houston, 4. Phoenix, 5. San Diego, 6. Chicago, 7. Philadelphia, 8. Las Vegas
- You are given the following numbers: 22838 26302 47615
- Read the numbers from left to right until you select four cities
- ANSWER: New York, Las Vegas, Houston, Chicago



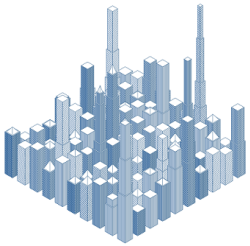
The Simple Random Sample-Exercise

- Using Number generator, record your number. If you have 5 or less, go to left side of room. Otherwise, go to right.
- How should the two sides differ? Should they at all?
- How will this affect what we infer from a sample? Keep in mind for later on in the course...



Other Probability Samples

- A **stratified random sample**: make sure your sample has known percentages of individuals of certain types (strata)
- America's State of Mind report was based on a probability sample of Medco's de-identified database of members with 24 months of continuous insurance enrollment. Sampling was stratified by age group and sex to match the demographics of the whole customer base.
- A **multistage sample**: select your final sample in stages, by sampling successively within a sample within a sample

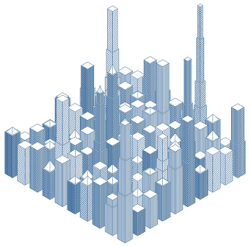


Example of 2-stage Sample

The National Youth Tobacco Survey administered in schools use a sampling procedure to generate a nationally representative sample of students in grades 6-12. Sampling is probabilistic and consists of selecting:

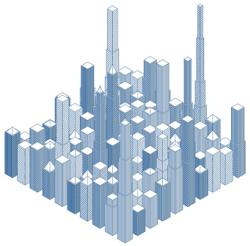
- Counties as primary sampling units (PSU)
- Schools within each selected PSU
- Classes within each selected school

These studies are cheaper to conduct (for example have to sample less counties, meaning less distance between schools), but estimates of population totals and means vary more (bad)!



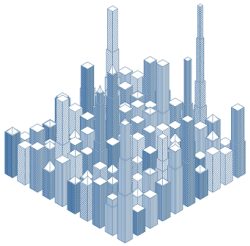
Sample Surveys

- A **sample survey** is an observational study that relies on a random sample drawn from the entire population
- Opinion polls are sample surveys that typically use voter registries or telephone numbers to select their samples
- In epidemiology, sample survey are used to establish the **incidence** (rate of new cases per year) and the **prevalence** (rate of all cases at one point in time) of various medical conditions, diseases, and lifestyles. These are typically stratified or multistage samples.



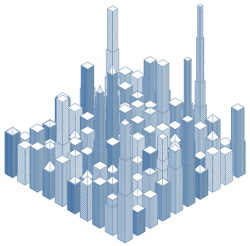
Some Challenges

- **Undercoverage**: Parts of the population are systematically left out
- **Nonresponse**: Some people choose not to answer/participate
- **Wording effects**: Biased or leading questions, and complicated/confusing statements can influence survey results
- **Response bias**: If people lie, forget, or misanswer
- **Endogeneity**: Outcome of survey affects estimand survey is trying to estimate!



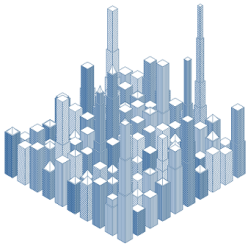
How are we biasing

- Try to think about how your sampling issues may bias your estimate of the population
- For example, if we over sample from a certain population how might these affect our estimate?
- However, trying to predict the direction of bias could be an issue. At some point we are just guessing and our predispositions could affect analysis. Almost impossible to predict confounding!



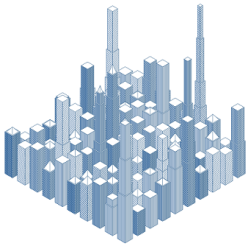
Undercoverage

- Polls and surveys may have a harder time reaching younger people who do not use traditional devices that are used as mediums for surveys, like landlines
- Certains groups of people that may be disenfranchised may be harder to reach or not a group that is reached out to
- Implication of undercoverage is other groups are **over – sampled**. Weighting is one way to correct this



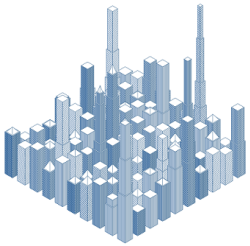
Nonresponse

- The Census Bureau's American Community Survey is about 97.5% via mail with reminders, with mandatory response
- University of Chicago's General Social Survey (GSS): Has about 70% response in person
- Pew Research Center methodology survey has about 10% response
- Private polling firms such as SurveyUSA has about 10% response in 2002. Even lower in 2020 with shift away from landlines. Online polls have low participation rates, but high amount of users (could affect how representative your sample is)



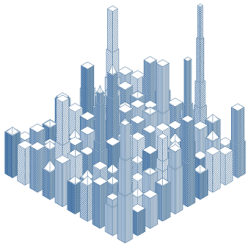
Wording Effects

- A Gallup 2013 study investigated how phrasing effects the opinions of Americans regarding physician-assisted suicide. Recall, we had form A and form B
- Form A: When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient's life by some painless means if the patient and his family request it?
- Form B: When a person has a disease that cannot be cured and is living in severe pain, do you think the doctors should or should not be allowed by law to assist the patients to commit suicide if the patient requests it?
- Even though this is an experiment that was randomly assigned, the wording makes the conclusion less robust



Comparitive Observational Studies

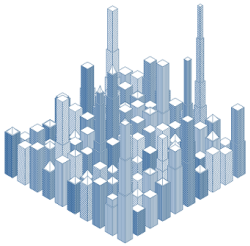
- **Case – control studies** start with 2 random samples of individuals with different outcomes, and look for exposure factors in the subjects' past ("retrospective")
- Individuals with the condition are cases, and those without are controls
- Good for studying rare conditions. Selecting controls is challenging
- **Cohort studies** enlist individuals of common demographic and keep of them over a long period of time ("prospective"). Individuals who later develop a condition are compared to those who don't develop the condition.
- Cohort studies examine the compounded effect of factors over time. Good for studying common conditions.



A Case-Control Study Example

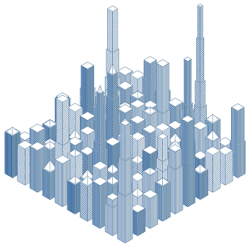
- Alfatoxins are secreted by a fungus found in damaged crops and can cause severe poisoning and death.

- The Kenya Ministry of Health investigated a 2004 outbreak of aflatoxicosis resulting in over 300 cases of liver failure. A sample of 40 case patients and 80 healthy controls were asked how they had stored and prepared their maize.



A Case-Control Study Example

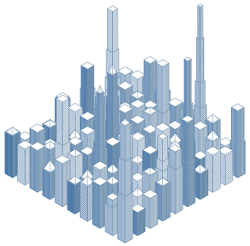
- The case patients were randomly selected from a list of individuals admitted to a hospital during the 2004 outbreak for for unexplained acute jaundice.
  Control individuals were selected to be as similar to the case patients as possible, yet randomly selected.
- Preliminary data suggested that soil, microclimate, and farming practices mights have played a role, but not age or gender
- For each case patient, two individuals from the patients' village with no history of jaundice symptoms were randomly selected.



Example of Cohort Study (1)

- The Nurses' Health Study is one of the largest prospective observational studies designed to examine factors that may affect major chronic diseases in women.

- Since 1976, the study has followed a cohort of over 100,000 registered nurses. Every two years, they receive a follow-up questionnaire about diseases and health-related topics, with 90% response rate each time.



Example of Cohort Study (2)

- 2007 Report on Age-Related Memory Loss: About 20,000 women ages 70+ had completed telephone interviews every two years to assess their memory with a set of cognitive tests. One of the findings: the more women walked during their late 50s and 60s, the better their memory score was at age 70 and older.

- However, because this is an observational study, we cannot conclude a causal effect of walking on protecting against memory loss.

Chapter 1 Notes

Picturing Distributions with Graphs
STP-231

Arizona State University

Introduction
Key Terms

- We have individuals and variables. Variables change value and take on individual values.
- Categorical variables: Variables take on set values, i.e. categories
- Quantitative variables: Variables that record an amount

Individuals
Potential Outcomes

- The objects (or units) described by a set of data.
- They are the individual units within a sample or a population.
- Examples:
  - People
  - Animals
  - Plants, etc.

Variables

- A characteristic of an individual that can be assigned a number (i.e. height) or category (i.e. color, yes/no).
- Not a specific value until **observed** from an individual.
- Capital letters denote variables and lower case letters denote observations. Ex:

$$\text{Y} = \text{number of hours you sleep per night}$$
$$y_1 = 8, y_2 = 7, y_3 = 8, \ldots, y_n = 7$$

Quantitative Variables

- **Discrete**: A variable with finite number of possible values that we could list. Ex: The number of texts you send on a given day

- **Continuous**: A variable with an infinite number of possible values Measured on continuous scale. Ex: weights of babies

Quantitative Variables
Examples

Are these discrete or continuous?

- Number of books in Hayden library?
- The time between bus arrivals
- Flip a coin 20 times and count the number of heads obtained
- The number of days an ant lives (rounded to the nearest day)

Categorical Variables
Examples

We have **Ordinal** and **Nominal** variables

- Ordinal: Ranked categorical variables with meaningful order
- Ex: Grading scale (A-F)
- class year (Fresh, soph, etc.)
- Nominal Unordered categorical variables
- Ex: the brand of your phone, favorite animal, the state you live in

Categorical Variables
Examples

Are these numeric or categorical?

- A biologist measured the number of leaves on each of 25 plants
- The temperature was recorded everyday for a month
- A conservationist recorded the weather (clear, cloudy, partly cloudy, rainy) and the number of cars parked at a trailhead on each of 18 days
- The months of the year
- Nationality

Exploratory Data Analysis

Frequency Distributions

- **Frequency** How often a value occurs for a categorical or quantitative variable within our data,
- A frequency distribution is a listing of distinct values from the data set and their number of occurrences

Example 2.2.1 **Color of Poinsettias** Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.[1] The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1. ■

Table 2.2.1 Color of 182 poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Samuels, Myra L., et al. *Statistics for the Life Scienc*

Relative Frequency Distributions

Relative frequency: The ratio of the total number of observations in the data set

- Let n be the sample size, then

$$\text{relative frequency} = \frac{\text{frequency}}{n}$$

- Multiply by 100% to represent as a percentage

Frequency Distributions
Example

- Representing the same data as relative frequencies/percentages

Table 2.2.1 Color of 182 poinsettias

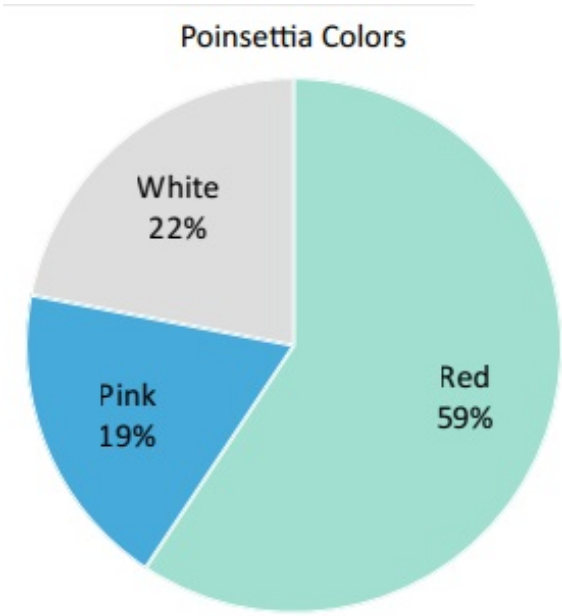
Color	Frequency (number of plants)	Relative Frequency	Percentage
Red	108	0.593	59%
Pink	34	0.186	18%
White	40	0.219	21%
Total	182	0.998?	99.8?

- Roundoff error Makes the data easier to read but why percentages may not add up to 100%.

Displaying Categorical Data
Objective: To Organize Categorical Data

- Organize qualitative data by constructing either the frequency distribution or relative frequency distribution
- Oranize categorical data by graph
- Bar chart, pie chart, waffle chart

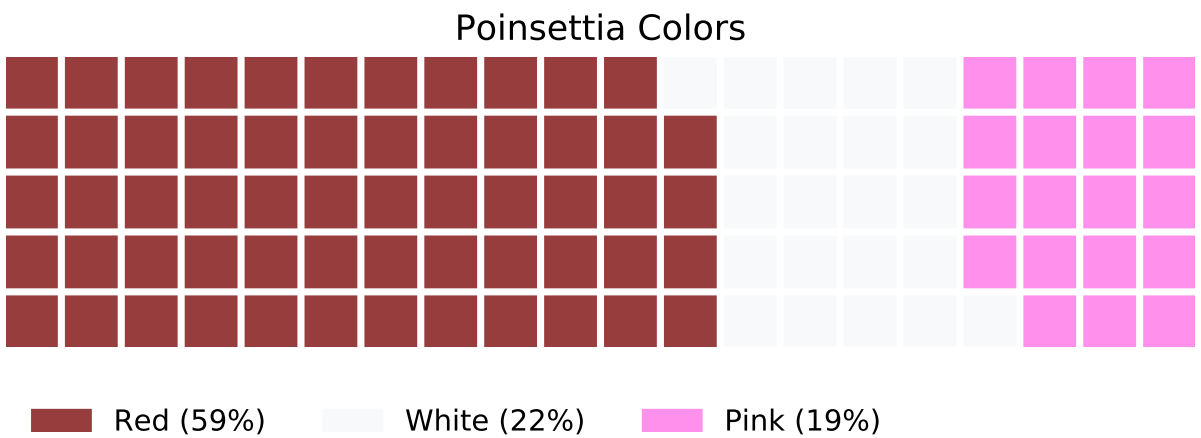
Pie chart



- A circle divided into sectors, each sector shows relative frequency of each category.
- Select category relationship to the whole
- Best used for a small amount of categories or when one is much bigger.

Waffle Chart

Counting rectangles easier than seeing percentage on circle



Bar Graph

- Represent frequency or relative frequency per category through bar height
- Decreasing order of magnitude (height) points out relative importance
- Separated bars for categorical - no order or connection between groups

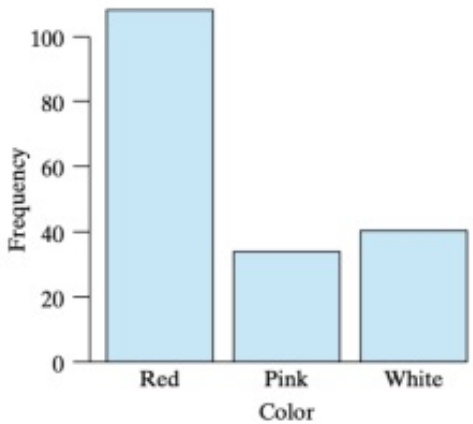
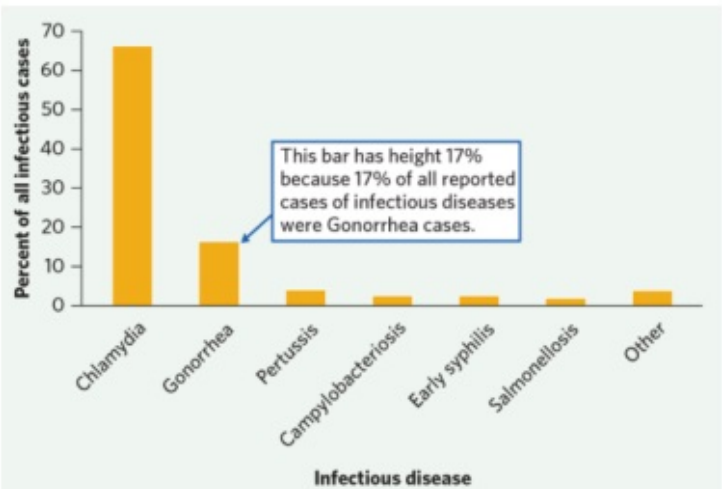


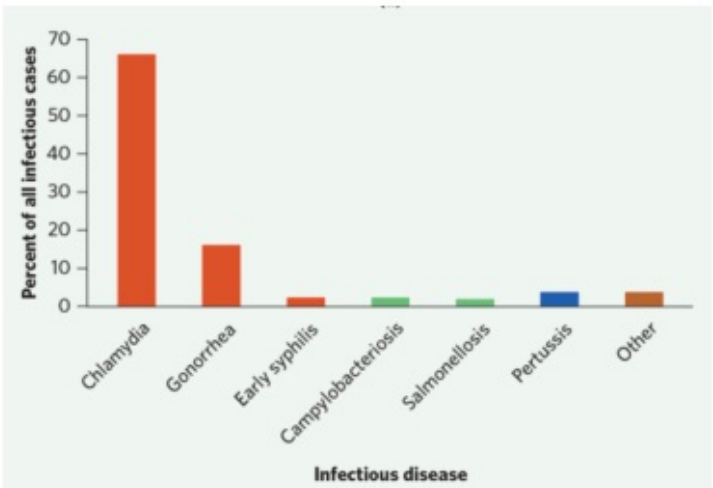
Table 2.2.1 Color of 182 poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Example

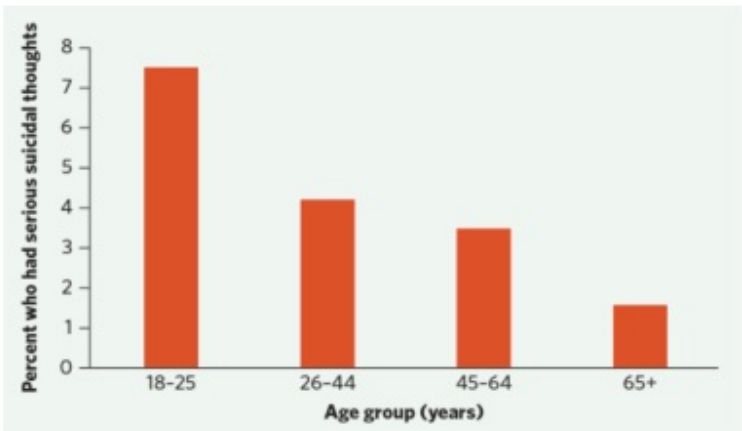
- In 2014, California had a total of 262,780 reported infectious cases.

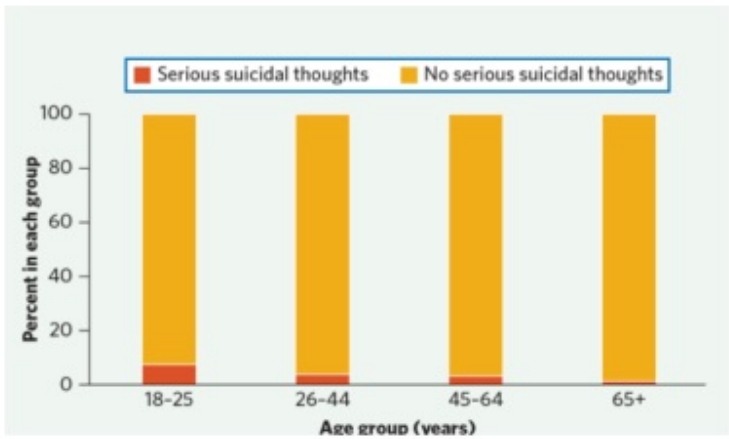




Bar Charts vs Pie Charts

- Pie charts handle all categories for one variable, easier interpretation
- Bar charts are more flexible, maybe less interpretable





Bar Charts vs Pie Charts
Continues

- Bar chart flexibility can enhance understanding, for example if we unstack the bars:

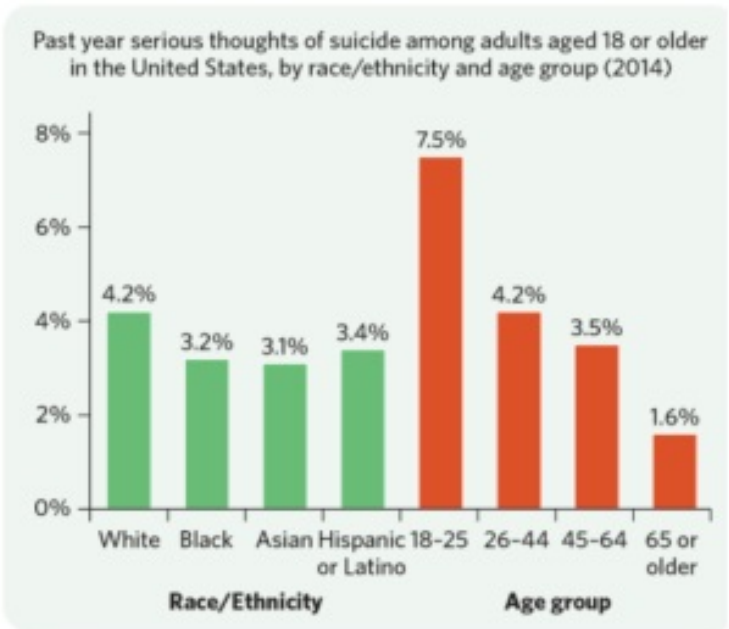
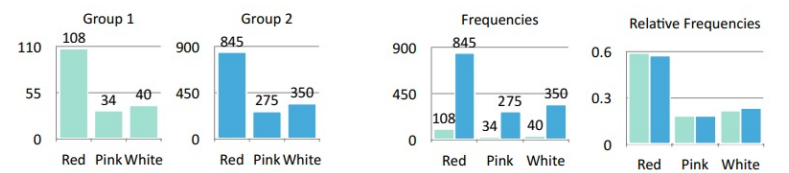


Figure 1.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Relative Frequency to Compare Groups

- The relative frequency scale is useful if several data sets of different sizes (i.e. changing n) are to be displayed together for comparison.
- Poinsettia data (with sample sizes 182 and 1470 respectively)
- Placing group 1 frequencies on group 2 frequencies is not informative, but relative is



Displaying Quantitative Data
Organize Quantitative Data

- Stem and Leaf and dot plots are useful
- Can display all the data, but not informative with big data. There, grouping is useful, i.e. a histogram
- Time plots are useful too.

Stem and Leaf Plot

- Visual display of the "raw" data
- Each possible value is split into a "stem" (the first digit or digits) and a "leaf" (the last digit)
- We can one line per stem or two lines per stem

Stem and Leaf Plot
One line per stem

- Rearrange number in ascending order, split first digits (stem) from the last digit (leaf) for each observation.
- Stems may have multiple leaves, and place stems in ascending order vertically, place leaves in ascending order horizontally

52	68	74	79	88
63	69	77	82	93
52	65	70	77	82

Stem	Leaf				
5	2	2			
6	3	5	8	9	
7	0	4	7	7	9
8	2	2	8		
9	3				

Stem and Leaf Plot
Two lines per stem

- Rearrange numbers in ascending order, split first digits (stem) from the last digit (leaf) for each observation.
- Write each stem twice in ascending order vertically. First stem is leaves <5, 2nd ≥ 5.

2.8	3.1	3.5	3.3	3.3
3.3	3.4	3.8	3.9	4.0
2.0	2.5	2.2	2.7	2.7

2	0 2
2	5 7 7 8
3	1 3 3 3 4
3	5 8 9
4	0
4	

Stem: Ones

Leaf: First decimal place

Example

Example 7 from Elementary Statistics by Neil A. Weiss

2.65 Ages of Baseball Players. From *MLB Roster Analysis* on the ESPN Web site, we found the average age of the players on each of the 30 major league baseball teams, as of May 2, 2005, to be as follows.

26.6	27.9	27.9	29.9	29.3	28.1
28.4	28.9	27.7	28.7	30.5	29.8
28.5	27.9	30.9	29.3	28.8	28.6
29.1	31.0	30.7	30.3	29.7	31.0
29.4	29.8	29.4	32.7	34.0	31.8

Construct a stem-and-leaf diagram for these data using

a. one line per stem. b. two lines per stem.

c. Which stem-and-leaf diagram do you find more useful? Why?

Dot Plot

- Visual display of the "Raw" data
- To make, sort the data set and plot each observation according to numerical value along a labeled scale axis.
- Each dot represents one observation in the data set, identical observations usually stacked

Table 2.2.3 Infant mortality in seven South Asian countries

Country	Infant mortality rate (deaths per 1,000 live births)
Bangladesh	47.3
Bhutan	40.0
India	44.6
Maldives	25.5
Nepal	41.8
Pakistan	59.4
Sri Lanka	9.2

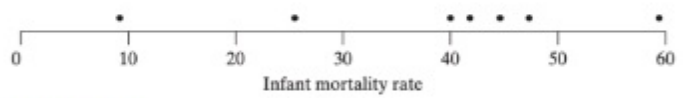
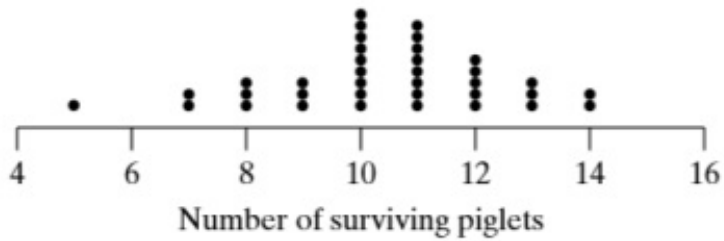


Figure 2.2.3 Dotplot of infant mortality in seven South Asian countries

Dot Plot
Example

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36



Histogram

- Numerical data using vertical bars. The height of the bar represents the frequency or relative frequency of each possible value/range of values
- Values of numerical variable is on horizontal axis, frequency on vertical axis
- Like a bar chart but representing numeric variable w/ natural order and scale
- Scale of variable determines where bars are placed & no space between bars

Dot plot to histogram

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

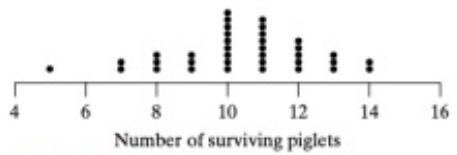


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

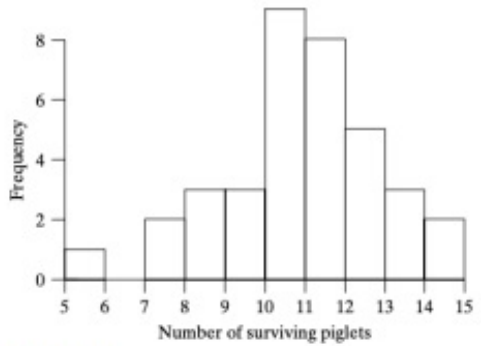


Figure 2.2.5 Histogram of number of surviving piglets of 36 sows

Grouped Frequency Distributions

- Organized quantitative data by dividing the observations into groups called classes
- Group data shows more information about the shape of the distribution rather than an individual observation

Grouped Frequency Distributions
Example

Serum CK Creatine phosphokinase (CK) is an enzyme related to muscle and brain function. As part of a study to determine the natural variation in CK concentration, blood was drawn from 36 male volunteers. Their serum concentrations of CK (measured in U/l) are given in Table 2.2.6.[5]

Table 2.2.6 Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Hist. without grouping:

Grouped Frequency Distributions
Example (continued)

Table 2.2.7 shows these data grouped into **classes**. For instance, the frequency of the class [20,40) (all values in the interval $20 \leq y < 40$) is 1, which means that one CK value fell in this range. The grouped frequency distribution is displayed as a histogram in Figure 2.2.7.

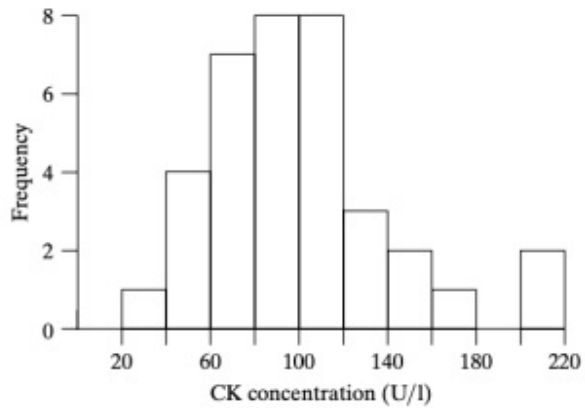


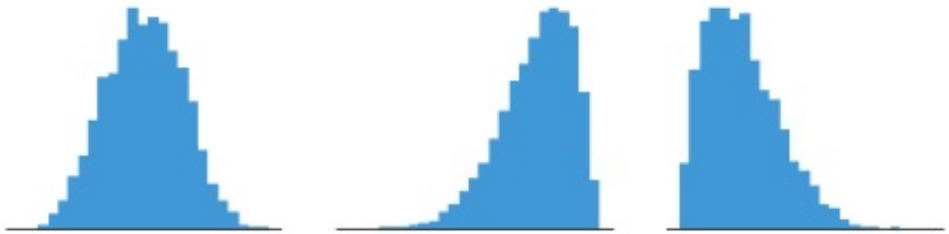
Table 2.2.7 Frequency distribution of serum CK values for 36 men

Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

Interpreting Histograms

- Any graph is used to identify patterns or deviations from patterns
- Patterns in histograms are described by:
  - Shape (modality, skewness)
  - Center (mean, median, mode)
  - Spread (standard deviation, percentiles)
  - Outliers: an individual value that falls outside the overall pattern

Modality



Unimodal - One Peak

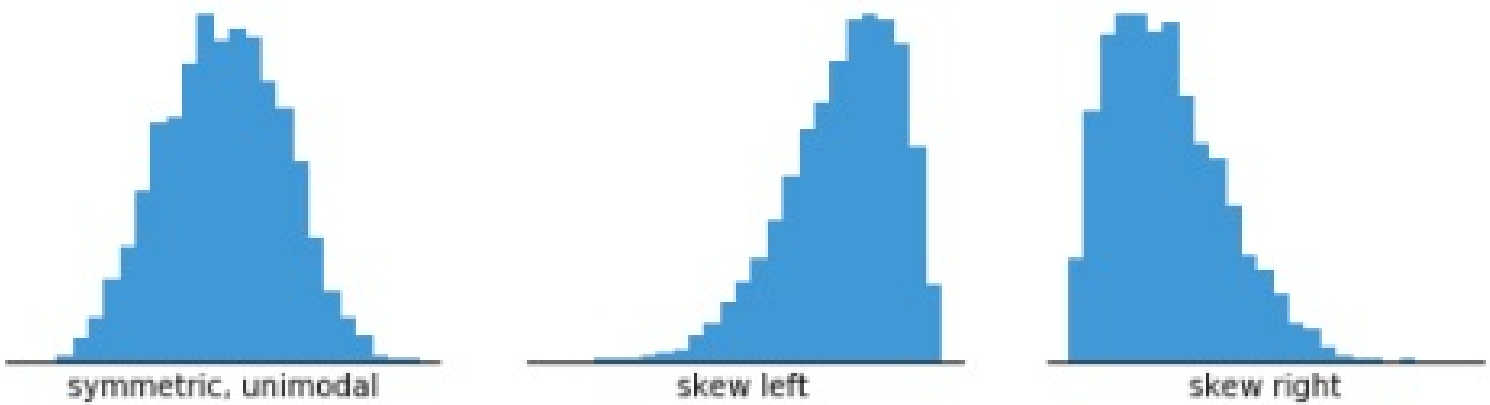
Uniform

Bimodal - Two Peaks

Multimodal - Several Peaks

Skewness

- Skewed to the right (positive skew): The right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side
- Skewed to the left (negative skew): The left size of the histogram extends much farther out than the right side
- Symmetric: Left and right hand side of the distribution are basically same



Cut Point Grouping

Definitions: A class is the same as a bin.

- Lower class cut point: The smallest value that can go into a class
- Upper class cut point: The smallest value that can go into the next higher class
- Class width: The difference between the cut points in a class
- Class midpoint: The average of two cut points in a class

Cut Point Grouping
Continued

- Number of classes should be between 5 and 20, and you approximate the number of classes
- All classes should share the same width
- All values must be included and each value belongs ONLY to one class
- An approximate class width is:

$$\frac{\text{Maximum value-Minimum value}}{\#\text{ of classes}}$$

Cut Point Grouping
Procedure

- Calculate the approximate class width, if not a whole number, round up
- Choose a number for the lower cut point of the first class (must be less than or equal to minimum observation)
- Obtain the other lower cut points by successively adding the chosen class width
- Specify all classes
- Determine which observations belong to each class and count frequencies

Cut Point Grouping
Example

87	81	86	90	88
90	86	86	87	88
90	81	89	89	83
89	85	86	86	90

Table: Ages of 20 people living in a retirement home

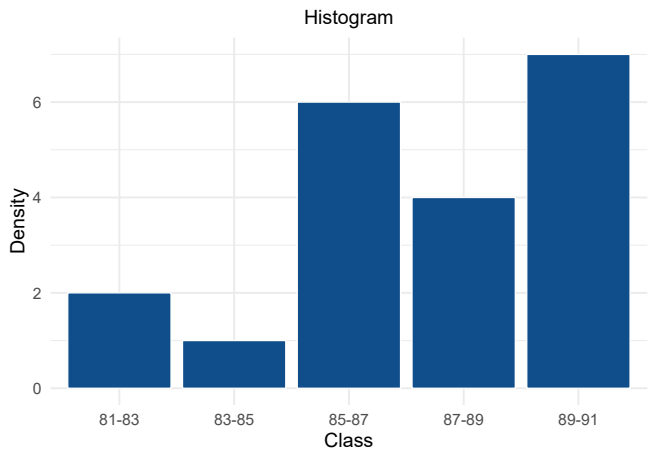
- Assume 5 classes (bins). We have

$$\frac{\text{Maximum value-Minimum value}}{\#\text{ of classes}} = \frac{90-81}{5} = 1.8 \uparrow 2$$

- In the last step we rounded up to 2. Now, we choose 81 as the lowest cut point
- Then the 5 cut points are each 2 units apart, i.e. we have cut points at 81, 83, 85, 87, 89, and 91

Cut Point Grouping
Example

Class	Frequency	Relative Frequency
81-under 83	2	0.1
83-under 85	1	0.05
85-under 87	6	0.3
87-under 89	4	0.2
89-under 91	7	0.35



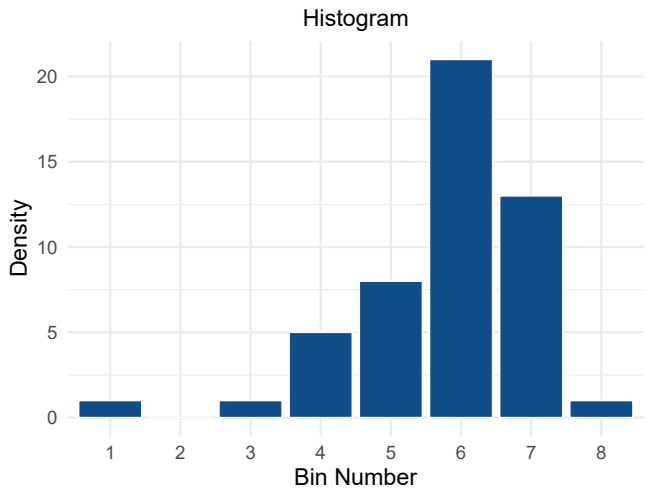
Cut Point Grouping
Example

Researcher want to know how often students buy used textbooks. They ask 50 graduates how many used textbooks they purchased during their undergraduate study. Results are summarized in the following table. Given that one of the intervals is 9 to 12, complete the frequency and relative frequency distribution table

13	16	21	17	12	13	17	19	19	20
17	16	14	11	20	16	22	20	15	17
18	21	19	19	19	18	22	3	17	18
19	17	20	20	13	14	17	22	21	16
19	25	21	17	16	17	8	18	20	17

Example continued

Bin	Class	Frequency	Relative Frequency
1	3-under 6	1	0.02
2	6-under 9	0	0
3	9-under 12	1	0.02
4	12-under 15	5	0.1
5	15-under 18	8	0.16
6	18-under 21	21	0.42
7	21-under 24	13	0.26
8	24-under 27	1	0.02



New Example

To improve egg production, producers often test alternative feeds and enhanced nutrients. Suppose a new enzyme is being tested and 20 eggs are randomly selected and weighed. The resulting weight (in grams) are given in the following table. Given that one of the intervals is 60 to under 62, complete the frequency and relative frequency distribution

58.8	59.9	60.7	59.0	55.9	56.5	54.4	60.6	62.1	59.9
57.1	58.9	58.1	56.6	55.2	57.7	55.8	56.2	57.9	56.5

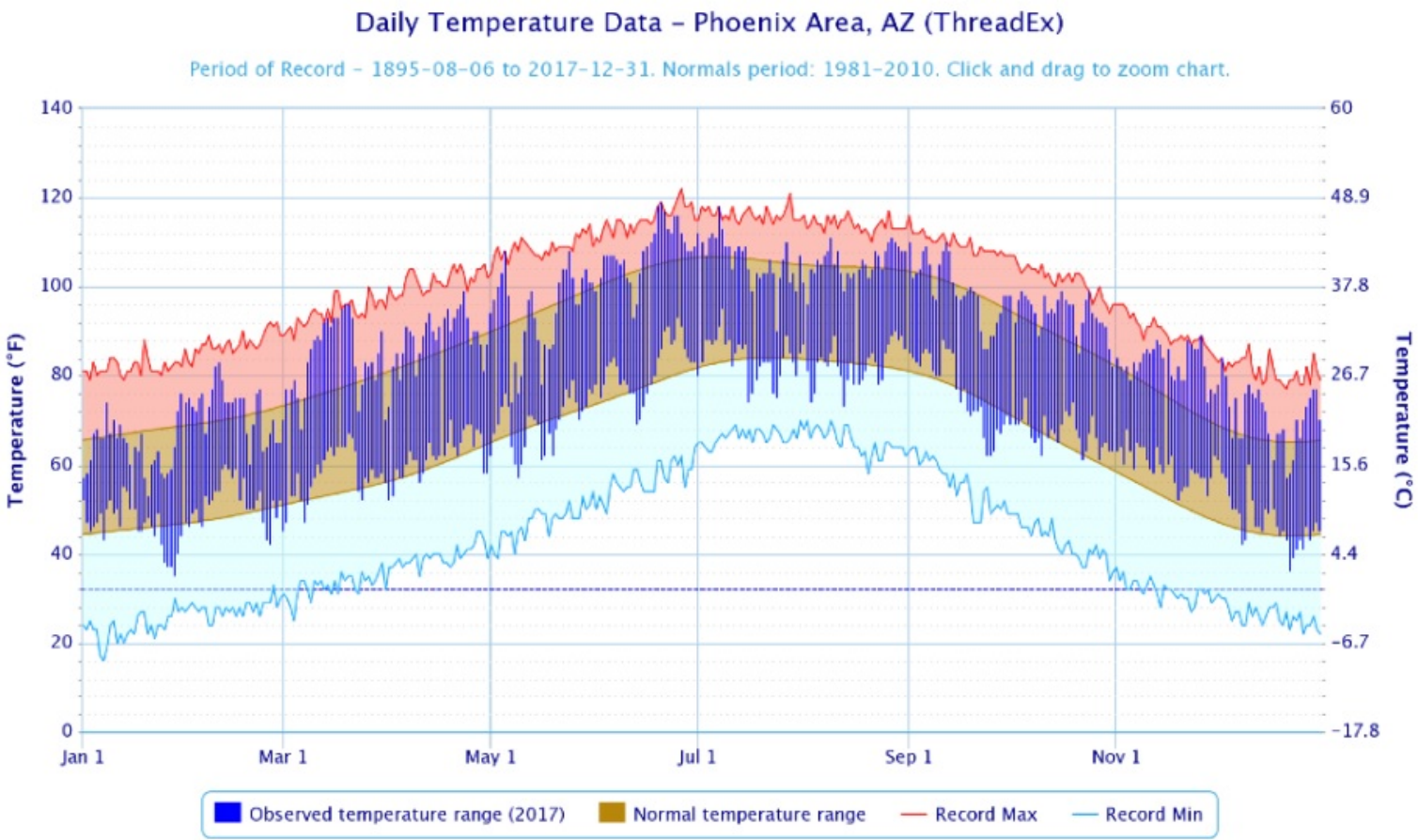
New Example

Since we know one interval is 60-62, so we start from the minimim value and increment by 2 till we capture all the data

Class	Frequency	Relative Frequency
54-under 56	4	0.20
56-under 58	7	0.35
58-under 60	6	0.30
60-under 62	2	0.10
62-under 64	1	0.05

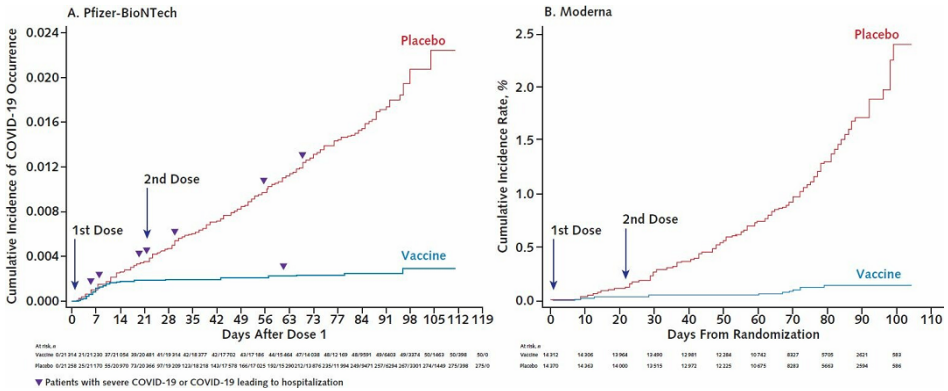
Time Plots

- A time plot of variables plots each observation against the time at which it was measured
- Put time on horizontal, variable of interest on vertical



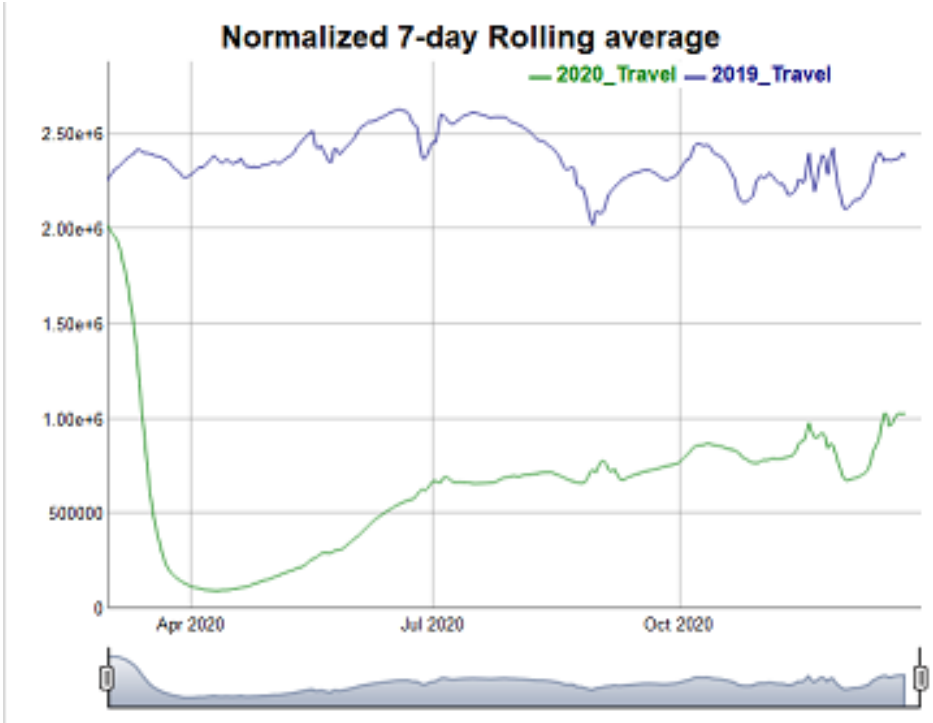
Comparing two Time plots

- The y-axes are on different scales. No good!



US Air Travel 2020 vs 2019

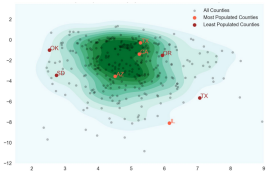
- This is a weekly average of 2020 vs 2019 from March on. Lag adjusted so we match day of week vs calendar date



Chapter 2 Notes

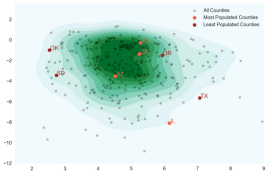
Describing Quantitative Distributions with Numbers
STP-231

Arizona State University



Introduction
Key Terms

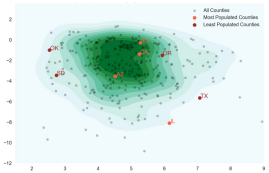
- A **Statistic** is a numerical measure calculated from sample data
- **Descriptive Statistics** define characteristics of data that describe it. Observational unit inference is made (if possible)
- We will look at the mean, median, and the mode



Sample Median

Split the ordered data into two equal halves

- Half of observations in the sample are above, half below
- To find the median, called $\tilde{y}$, rearrange the values of the data set in ascending order.
- For an odd number of observations, the median is the middle value, located at the $n/2$ element in the list
- If even, the median is the average of the 2 middle values.



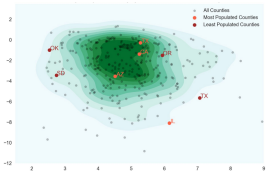
Sample Median
Examples

2.3.3 from Statistics for the Life Sciences by Myra Samuels 2016

- A researcher applied the carcinogenic compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:

$$6.3 \quad 5.9 \quad 7.0 \quad 6.9 \quad 5.9$$

Find the median



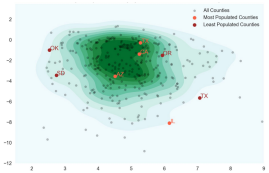
Sample Mean
Aka the average

- The sum of all the values divided by the total number of observations (where y_i is an observation)

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

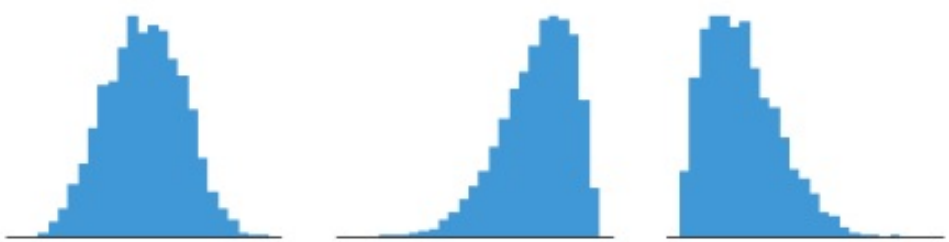
- For example, from the previous example

$$\overline{y} = \frac{1}{5}(6.3 + 5.9 + 7.0 + 6.9 + 5.9) = 6.4$$



Sample Mode

The most frequent value(s)

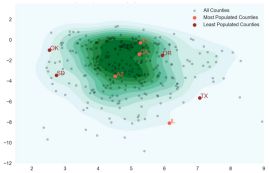


Unimodal - One Peak

Uniform

Bimodal - Two Peaks

Multimodal - Several Peaks



Example

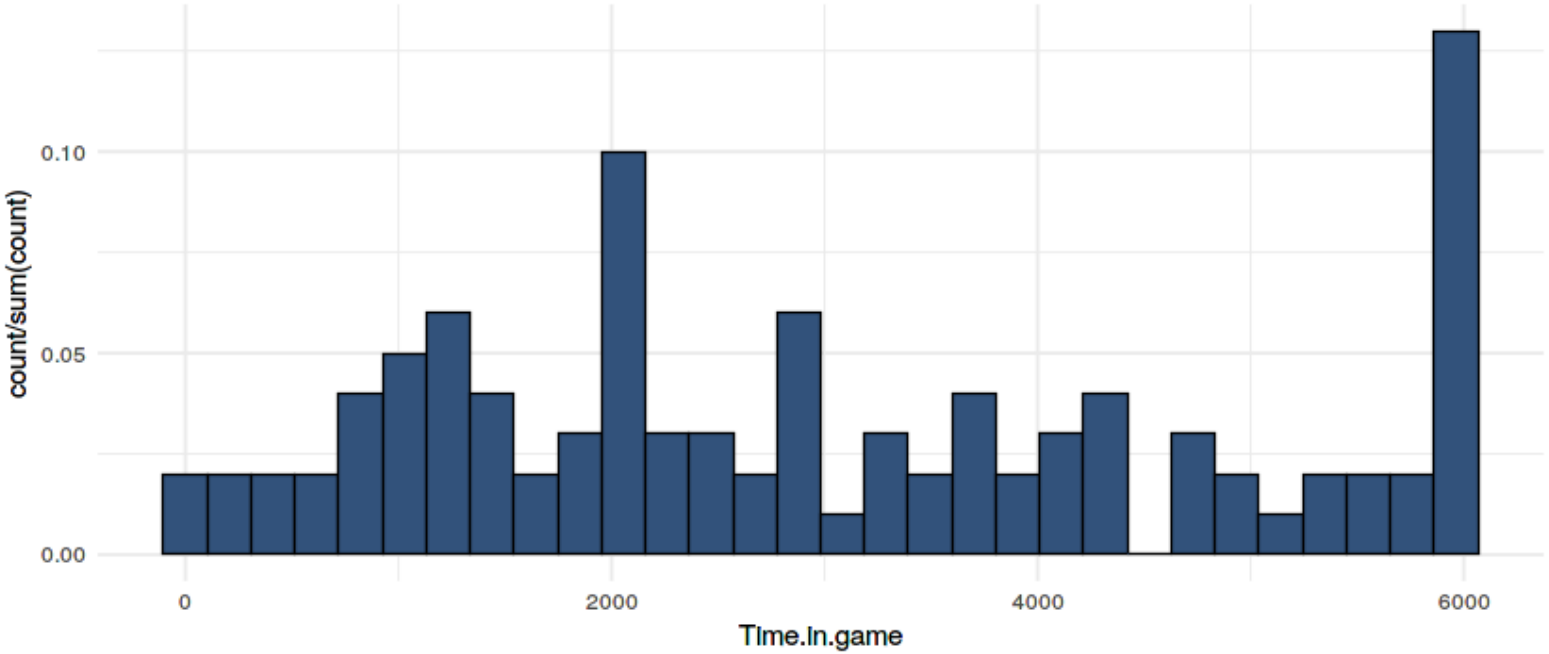
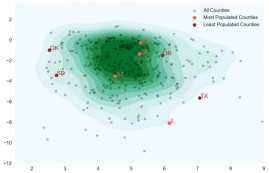
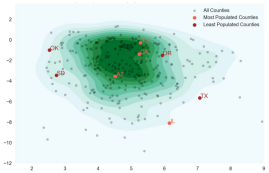


Figure: Lacrosse penalty data over time. Where would the mean, median, mode be? Which is most useful?



Resistant Statistics

- Resistant if extreme values have little to no influence on its outcome
- Median is better than mode in this regard
- However, if skewed, both will be pulled towards the longer tail
- Mean is usually pulled more than the median



Resistant Statistics
Example

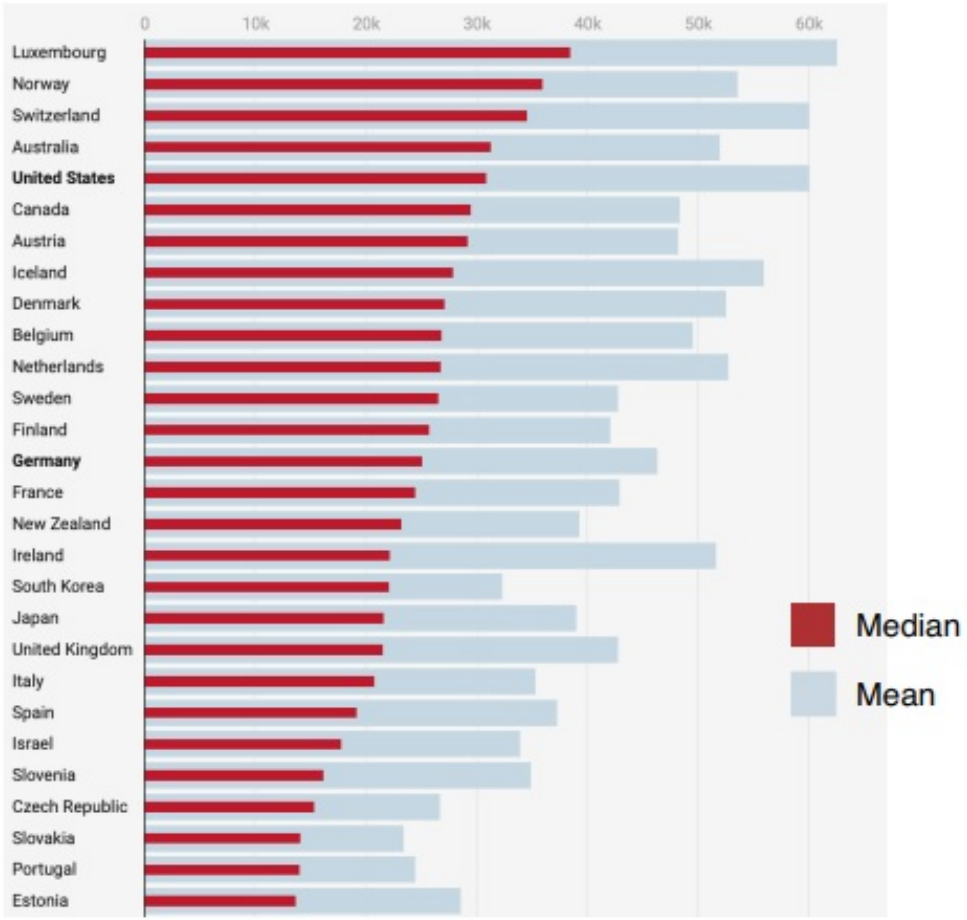
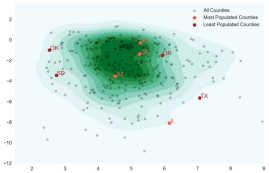


Figure: Median and mean income by countries, 2012/2014 (PPP).
https://blog.datawrapper.de/weekly-chart-income/



Deviation

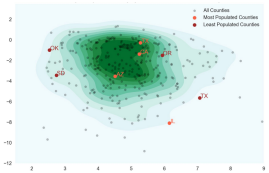
- Difference between a sample data point and the mean of the sample

$$y_i - \overline{y}$$

- The mean is uniquely defined as the value that "balances" the deviations.

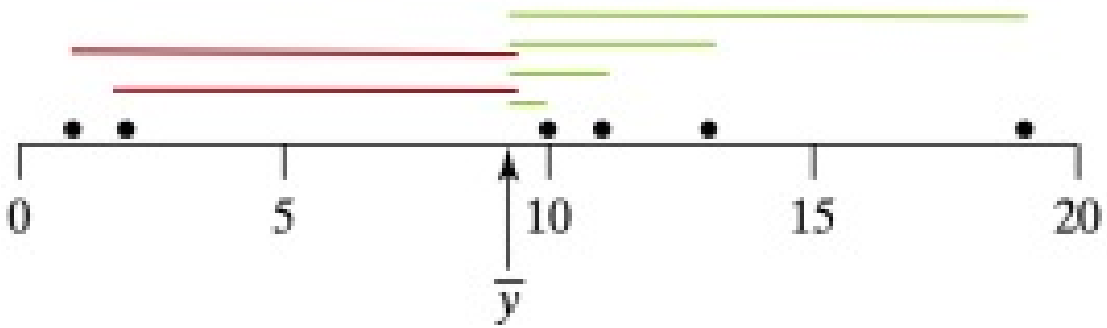
$$\sum_{i=1}^{n}(y_i - \overline{y}) = 0$$

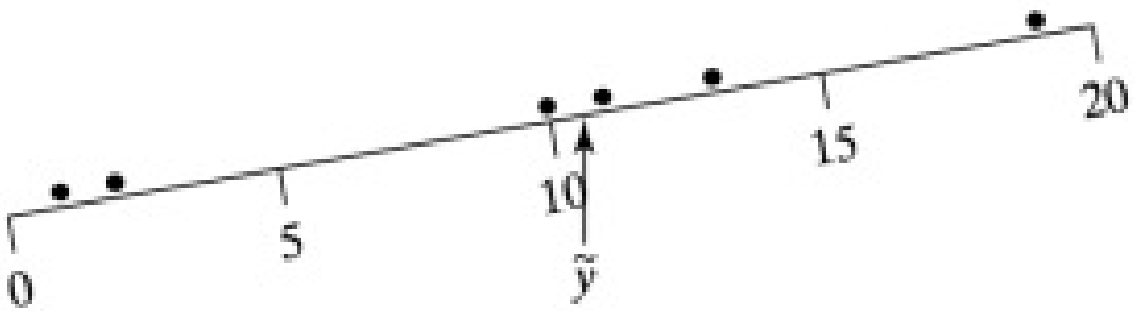
$$\begin{aligned}
&y_1 - \frac{1}{n}(y_1 + y_2 + \ldots + y_n) + \ldots + y_n - \frac{1}{n}(y_1 + y_2 + \ldots + y_n) \\
&= \left(\frac{ny_1}{n} - \frac{1}{n}(y_1 + y_2 + \ldots + y_n)\right) + \ldots + \left(\frac{ny_n}{n} - \frac{1}{n}(y_1 + y_2 + \ldots + y_n)\right) \\
&= \frac{y_1(n-1)}{n} - \left(\frac{y_2}{n} + \ldots + \frac{y_n}{n}\right) + \ldots + \frac{y_n(n-1)}{n} - \left(\frac{y_1}{n} + \ldots + \frac{y_{n-1}}{n}\right) \\
&= \frac{(n-1)y_1 - (n-1)y_1}{n} + \frac{(n-1)y_2 - (n-1)y_2}{n} + \ldots + \frac{(n-1)y_n - (n-1)y_n}{n} = 0
\end{aligned}$$

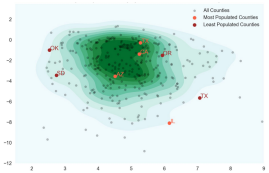


Deviation
Continued

- Consider the lamb data



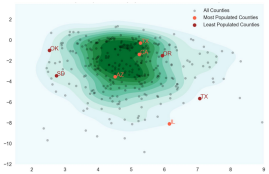




Example

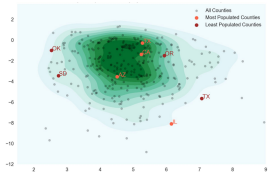
- An experiment of a weight loss drug finds that the group receiving the drug lost on average 10 pounds, but a median value of about 7 pounds lost.
- Can you explain whats going on?
- How would you find the mode in this example?
- Individual cases vary, but on **average** a participant on the drug would expect to lose 10 pounds

Measures of Spread



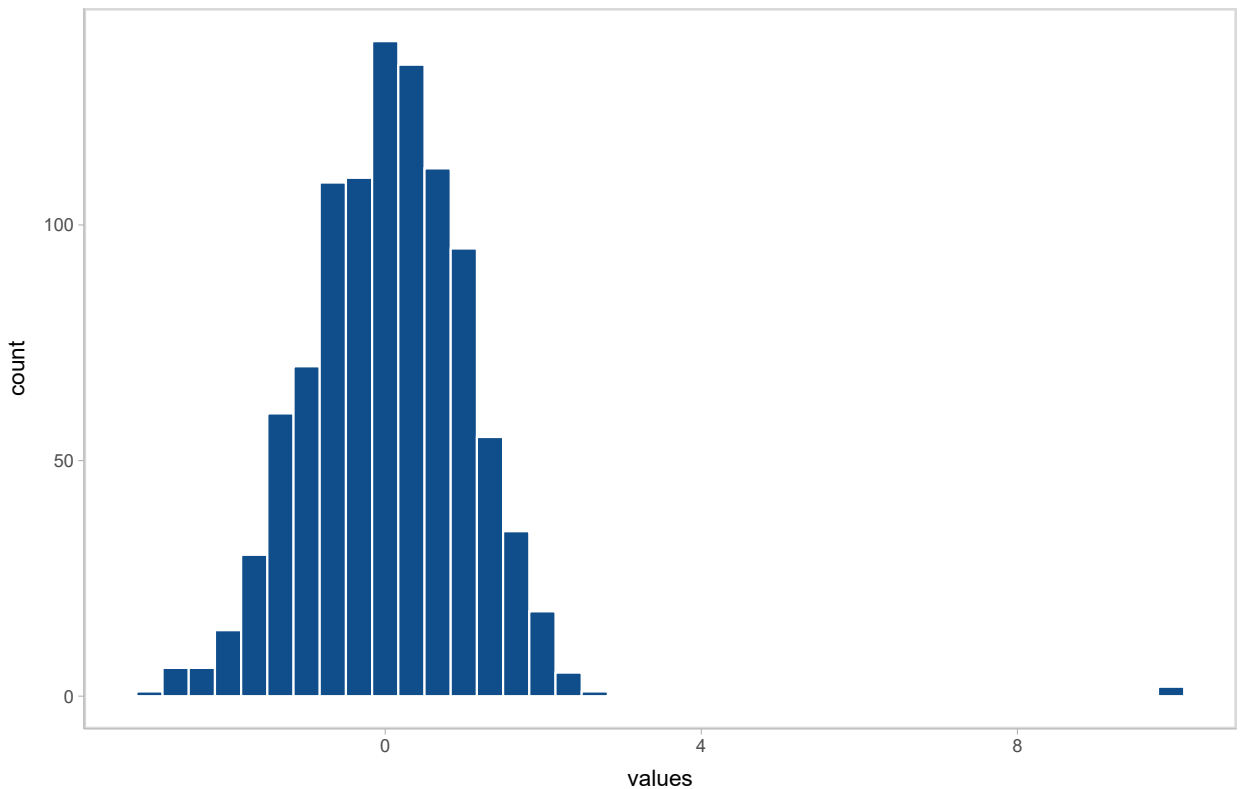
Measures of Spread

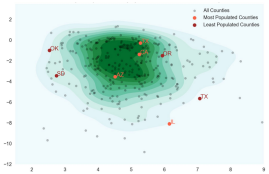
- We are still discussing statistics rather than parameters.
- A histogram is defined by not only its center, but its spread as well
- We can look at quartiles and standard deviation (as well as variation)



The Range

- The simplest way to define the spread, the distance from minimum to maximum. . . great unless we have outliers





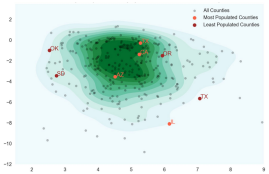
Quartiles
Examples

Note, if odd number of observations, do not include median in quartile calculations. If even, include the left side of median in Q_1 calculation, right median for Q_3 calculation.

- Q_1: (25th percentile) median of all data to the left of overall median Q_2
- Q_3: (75th percentile) median of the data to the right of the overall median Q_2
- Interquartile range (IQR):

$$\text{IQR} = Q_3 - Q_1$$

  Contains 50% of data.
- The five number summary
- Minimum, first quartile, median, third quartile, maximum

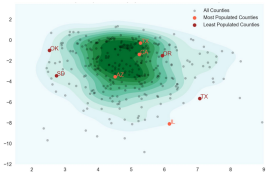


Quartiles
Examples

Note, if odd number of observations, do not include median in quartile calculations. If even, include the left side of median in Q_1 calculation, right median for Q_3 calculation.

- Q_1 is not strictly the 25th percentile (if we don't have enough data!) but as close as possible. This is why we do not include the median when an odd total n.

- We calculate the first quartile as the median of the first half and the third quartile as the median of the second half of our data.

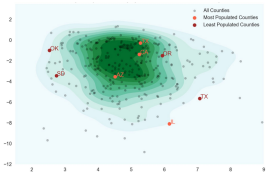


Quartiles
Example

Find the quartiles and IQR

- In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (in litres) were as follows[1]:

[1] Statistics for Life Sciences, Myra Samuels 2016



Quartiles
Example

Find the quartiles and IQR

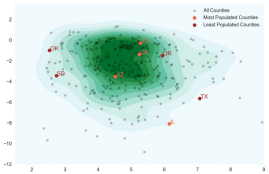
- In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (in litres) were as follows[1]:

  56.5 89.8 110.1 65.6 63.7 82.6 75.1 91.5 102.9 44.4 108.1

  Order them as:

  $$44.4, 56.5, 63.7, 65.6, 75.1, 82.6, 89.8, 91.5, 102.9, 108.1, 110.1$$

[1] Statistics for Life Sciences, Myra Samuels 2016



Quartiles
Example

Find the quartiles and IQR

- In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (in litres) were as follows[1]:

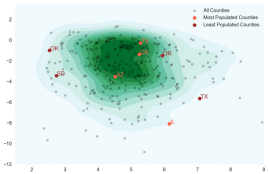
  56.5 89.8 110.1 65.6 63.7 82.6 75.1 91.5 102.9 44.4 108.1

  Order them as:

  $$44.4, 56.5, 63.7, 65.6, 75.1, 82.6, 89.8, 91.5, 102.9, 108.1, 110.1$$

  The median is 82.6, the 1st quartile is 63.7, and the last is 102.9. We find the median to the left of the median and the median to the right of the median, excluding the median!

[1] Statistics for Life Sciences, Myra Samuels 2016



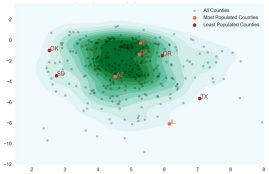
Quartiles
Example

Find the quartiles and IQR with an even number of data points

- The following is a list of 12 wait times at the grocery store (in minutes)

$$1, 2, 4, 4, 5, 8, 9, 10, 12, 13, 14, 15$$

- The first quartile is the median of



Quartiles
Example

Find the quartiles and IQR with an even number of data points

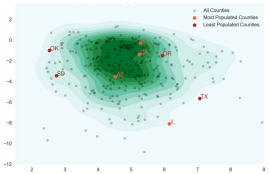
- The following is a list of 12 wait times at the grocery store (in minutes)

$$1, 2, 4, 4, 5, 8, 9, 10, 12, 13, 14, 15$$

- The first quartile is the median of

$$1, 2, 4, 4, 5, 8 = 4$$

- The third quartile is the median of



Quartiles
Example

Find the quartiles and IQR with an even number of data points

- The following is a list of 12 wait times at the grocery store (in minutes)

$$1, 2, 4, 4, 5, 8, 9, 10, 12, 13, 14, 15$$

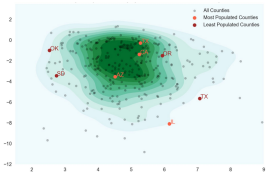
- The first quartile is the median of

$$1, 2, 4, 4, 5, 8 = 4$$

- The third quartile is the median of

$$9, 10, 12, 13, 14, 15 = (12 + 13)/2 = 12.5$$

- The median is



Quartiles
Example

Find the quartiles and IQR with an even number of data points

- The following is a list of 12 wait times at the grocery store (in minutes)

$$1, 2, 4, 4, 5, 8, 9, 10, 12, 13, 14, 15$$

- The first quartile is the median of

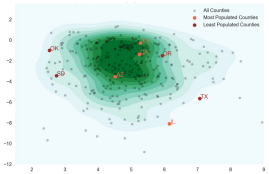
$$1, 2, 4, 4, 5, 8 = 4$$

- The third quartile is the median of

$$9, 10, 12, 13, 14, 15 = (12 + 13)/2 = 12.5$$

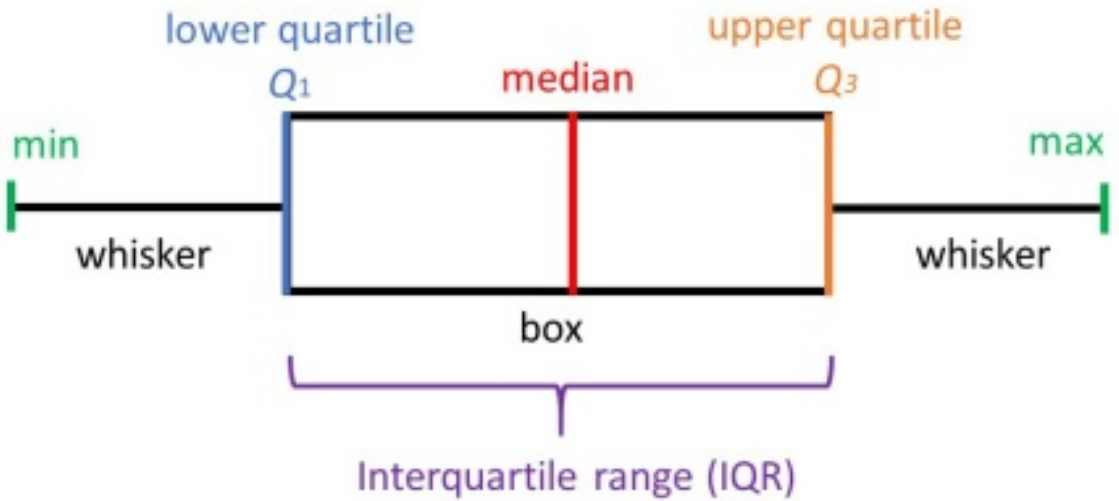
- The median is

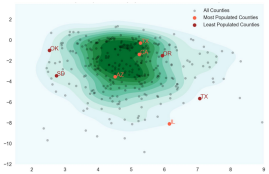
$$(8 + 9)/2 = 8.5$$



Box Plot
Aka box and whisker plot

- Used as a visual representation of the five-number summary





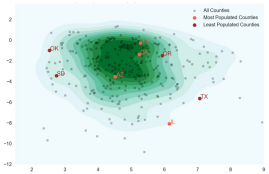
Outliers

- Observations that are far outside the overall pattern
- Worthy of investigation. Data entry error, strange phenomena? Even if we can explain them, could be an issue when analyzing data
- How do we decide what is classified as an outlier?
- Define lower and upper fences as
  $Q_1 - 1.5 * \text{IQR}$=lower fence and $Q_3 + 1.5 * \text{IQR}$=lower fence

Five Number Summary

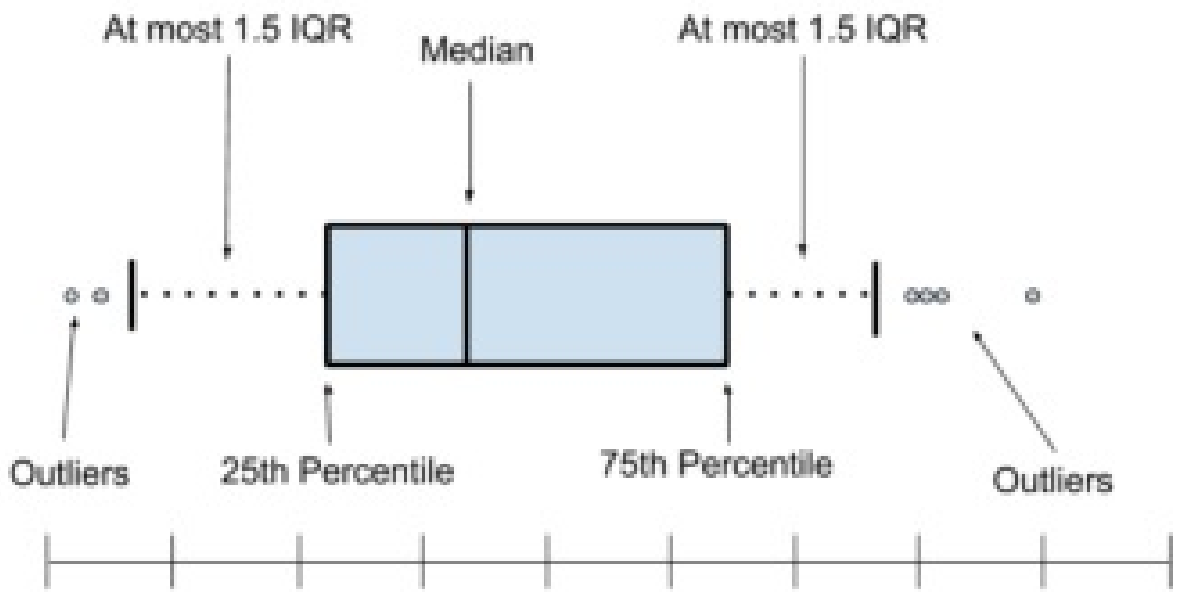
0.4 1.6 4.1 6.0 17.1

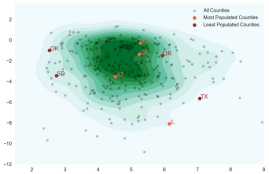
Figure 2.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



Modified Box Plot

- Box remains in its normal position
- Whiskers are updated, by drawing lower whiskers to the lowest data that remains above the lower fence
- Draw upper whiskers to highest data point that remains below upper fence

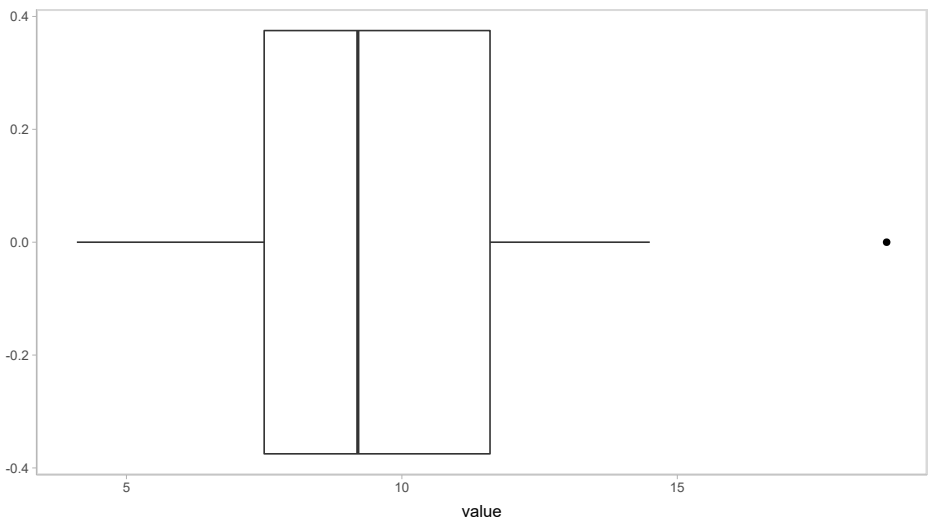


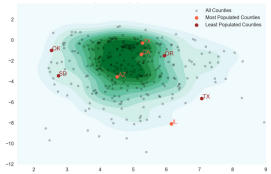


Example

MAO (Monoamine Oxidase) enzyme levels of 18 people were measured. The results (expressed as number of moles of benzaldehyde product per 108 platelets):

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	4.1	5.2	6.8	7.3	7.4	7.8	7.8	8.4	8.7	9.7	9.9	10.6	10.7	11.9	12.7	14.2	14.5	18.8

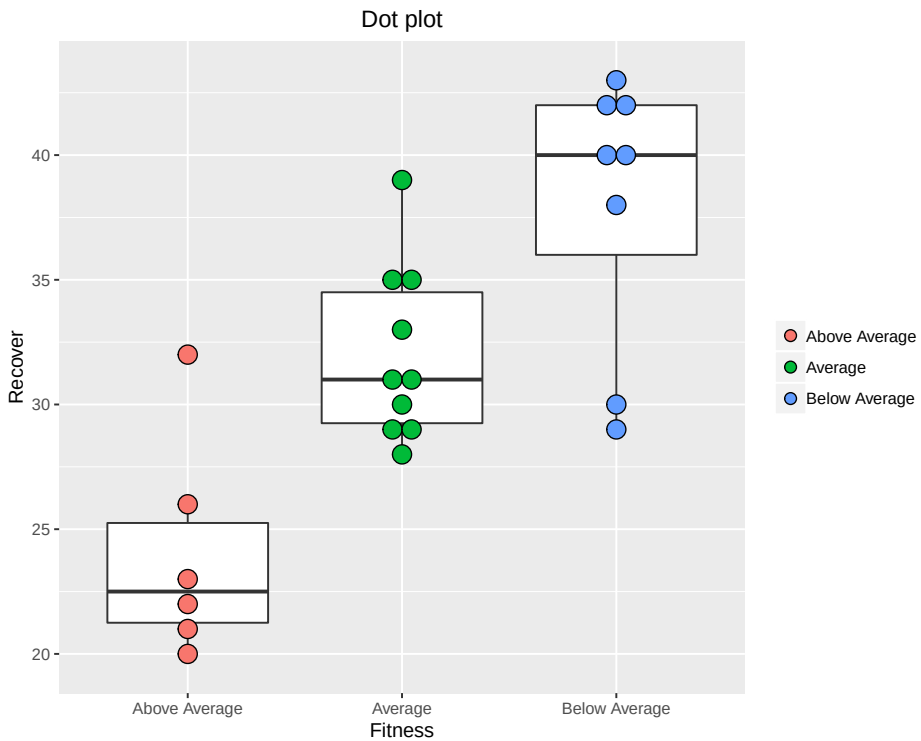


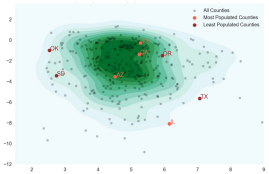


Example
Using Box plot with dot plot

Below is a table of recovery time of people after exercise based on their self-described fitness level

Recovery time	fitness level
29.0	below average
42.0	below average
⋮	⋮
33.0	average
⋮	⋮
22.0	above average





Histogram vs box plot

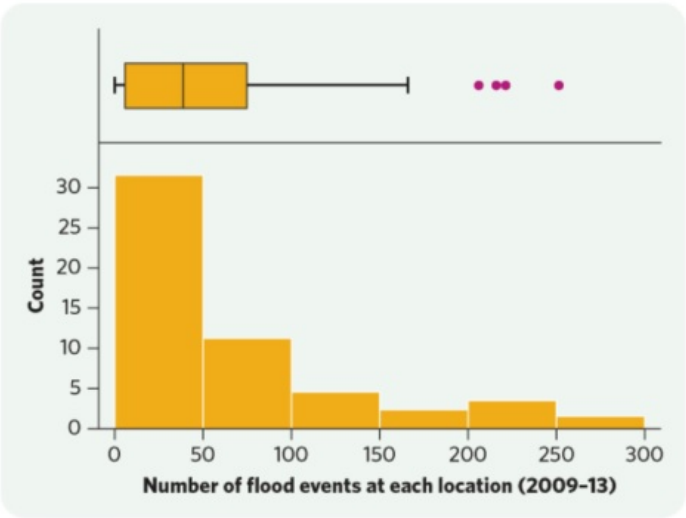
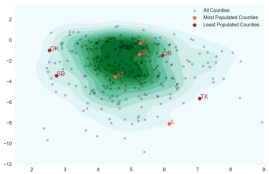


Figure 2.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

- Histograms show modality, skewness, range, and spread
- Boxplots show exact median, exaxt 50% spread, and easier to see (suspected) outliers



Sample Standard Deviation (s)

- Roughly defined as the average distance between a point and the sample mean The standard deviation is the square root of the variance.

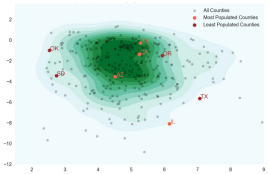
$$s = \sqrt{s^2} = \sqrt{\text{variance}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

- From a frequency distribution (where m is the number of unique values):

$$s = \sqrt{\frac{\sum_{i=1}^{m} f_i \times (y_i - \overline{y})^2}{n-1}}$$

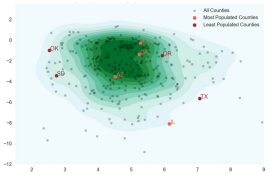
- And the **Coefficient of Variation**

$$\text{Coefficient of Variation} = \frac{s}{\overline{y}}$$



Properties of standard deviation

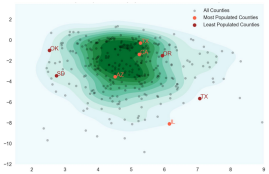
- Does not exist for a sample with a single point
- Represented in the units of the variable
- Less resistant than IQR more resistant than the range, but still pretty affected by outliers and skewness
- The square makes the variance more sensitive because of the square operation



Calculate some Values

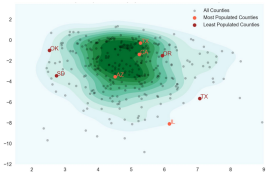
Plaza Gallery in Chicago sells home furniture online and is concerned some of their shipments to customers arriving late. Six days were randomly selected and the number of late shipment complaints were recorded. The observations were 3,5,5,6,6,8. Find the range, sample variance, and sample standard deviation for these data. R-code to calculate quickly:

```
data<-c(3,5,5,6,6,8)
var_ex<-var(data); print(var_ex)
sd_ex<-sd(data); print(sd_ex)
range_ex<-range(data); print(range_ex)
```



Calculate Variance

- Our data is $y = (3, 5, 5, 6, 6, 8)$. The mean is

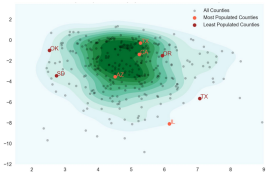


Calculate Variance

- Our data is $y = (3, 5, 5, 6, 6, 8)$. The mean is 5.5
- The sum of the squared deviations is:

$$(3-5.5)^2+(5-5.5)^2+(5-5.5)^2+(6-5.5)^2+(6-5.5)^2+(8-5.5)^2 = 13.5$$

- Which means the variance is



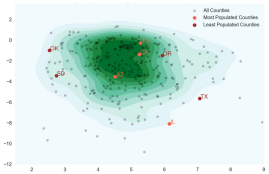
Calculate Variance

- Our data is $y = (3, 5, 5, 6, 6, 8)$. The mean is 5.5
- The sum of the squared deviations is:

$$(3-5.5)^2+(5-5.5)^2+(5-5.5)^2+(6-5.5)^2+(6-5.5)^2+(8-5.5)^2 = 13.5$$

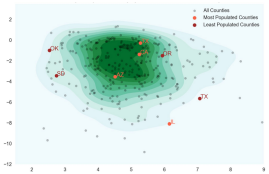
- Which means the variance is

$$\frac{13.5}{6-1} = 2.7 \xrightarrow{\text{standard deviation}} \sqrt{2.7} = 1.64$$



Degrees of Freedom

- Maybe you are wondering why we divide by $n-1$ not n for the standard deviation estimate
- The sum of deviations is always 0
- Then $n-1$ observations can vary freely but one is constrained since they must sum to zero, so only $n-1$ terms contriute to information about deviation
- Therefore, dividing by $n-1$ estimates the true population variance from the sample



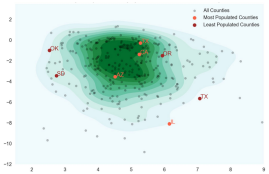
More Formally (skip if confusing!)

In general, for any a,

$$\begin{aligned}\sum_{i=1}^{n}(y_i - a)^2 &= \sum_{i=1}^{n}(y_i - \overline{y} + \overline{y} - a)^2 \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \overline{y}) + \sum_{i=1}^{n}(\overline{y} - a)^2 \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 + \sum_{i=1}^{n}(\overline{y} - a)^2 \text{ because the deviations sum to 0}\end{aligned}$$

Which is minimized when $\overline{y} = a$. If $a = 0$, then

$$(n-1)s^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$$

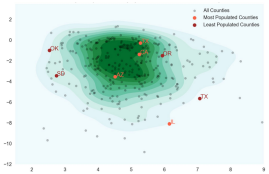


More Formally (skip if confusing!)

Note $\mu = E(\overline{y})$, the true parameter mean and $\sigma^2 = \text{Variance}(y_1)$ (assuming variance of all y_i's the same a common assumption) the true parameter standard deviation (where E is an expected value which we learn about later) Using the results from slide 31, we have

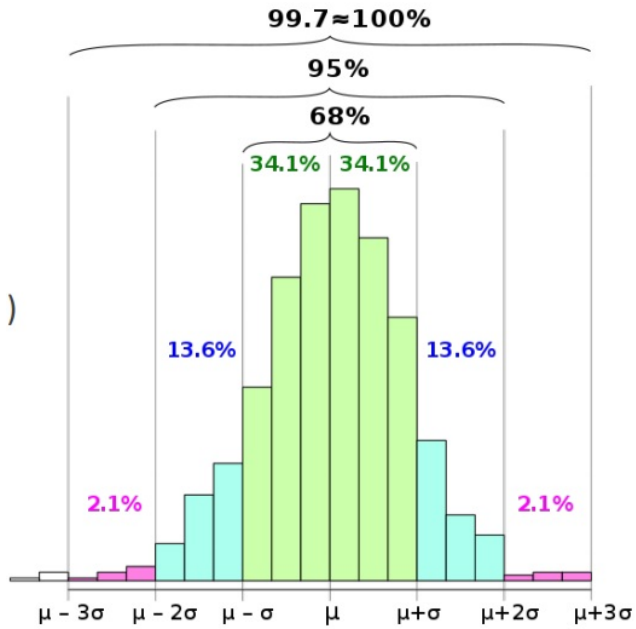
$$\begin{aligned} E(s^2) &= \frac{1}{n-1}\sum(y_i^2 - n\overline{y}^2) \\ &= \frac{1}{n-1}\left(nEy_1^2 - nE(\overline{y}^2)\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2 \end{aligned}$$

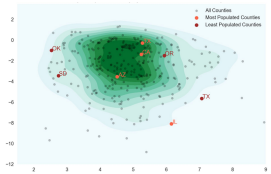
Therefore we divide by $n-1$ so there is no bias (i.e. we have an accurate estimate) when using the sample standard deviation to estimate the truth



Empirical Rule for a Sample

- For any sample of observations with a symmetric and unimodal distribution (and big enough n), we expect to find
  - 68% of all values fall within $(\overline{y} - s, \overline{y} + s)$
  - 95% of all values fall within $(\overline{y} - 2s, \overline{y} + 2s)$
  - 99.7% of all values fall within $(\overline{y} - 3s, \overline{y} + 3s)$





Example

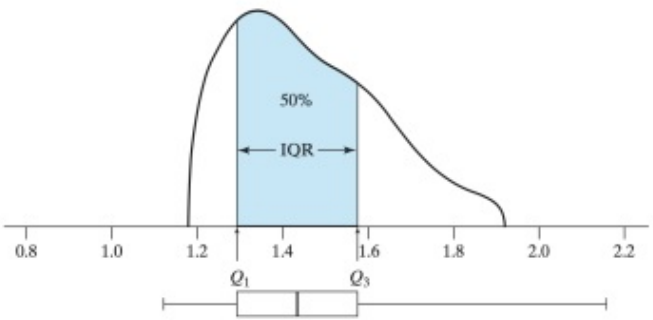
Example

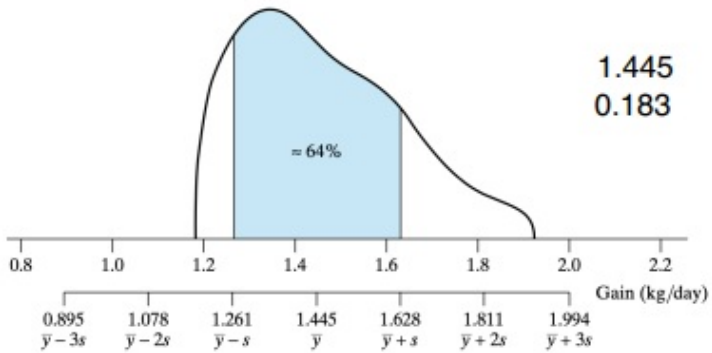
Example 2.6.6

Daily Gain of Cattle The performance of beef cattle was evaluated by measuring their weight gain during a 140-day testing period on a standard diet. Table 2.6.2 gives the average daily gains (kg/day) for 39 bulls of the same breed (Charolais); the observations are listed in increasing order.[36] The values range from 1.18 kg/day to 1.92 kg/day. The quartiles are 1.29, 1.41, and 1.58 kg/day. Figure 2.6.3 shows a histogram of the data, the range, the quartiles, and the interquartile range (IQR). The shaded area represents the middle 50% (approximately) of the observations.

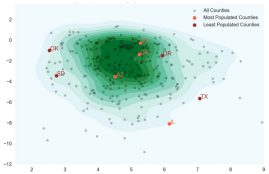
Table 2.6.2 Average daily gain (kg/day) of 39 Charolais bulls

1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	





Samuels, Myra L., et al. *Statistics for the Life Sciences*. Pearson, 2016.



Example II

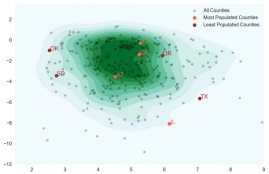
2.6.11 from the book

- Listed in increasing order are the serum creatine phophokinase (CK) levels (U/I) of 36 healthy men (these are the data from example 2.2.6):

(***)25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
*57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/I and the SD is 40.4 U/I. What %-age of the observations are within

  - 1 SD of the mean?



Example II

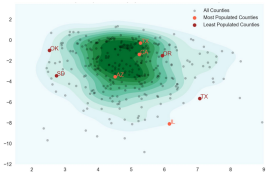
2.6.11 from the book

- Listed in increasing order are the serum creatine phophokinase (CK) levels (U/I) of 36 healthy men (these are the data from example 2.2.6):

(***)25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
*57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/I and the SD is 40.4 U/I. What %-age of the observations are within

  - 1 SD of the mean? 26/36
  - 2 SDs of the mean?



Example II

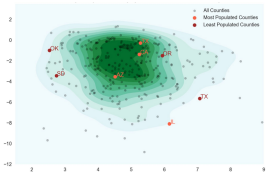
2.6.11 from the book

- Listed in increasing order are the serum creatine phophokinase (CK) levels (U/I) of 36 healthy men (these are the data from example 2.2.6):

(***)25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
*57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/I and the SD is 40.4 U/I. What %-age of the observations are within

  - 1 SD of the mean? 26/36
  - 2 SDs of the mean? 34/36
  - 3 SDs of the mean?



Example II

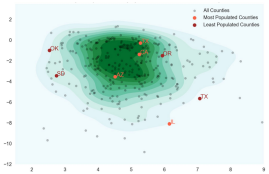
2.6.11 from the book

- Listed in increasing order are the serum creatine phophokinase (CK) levels (U/I) of 36 healthy men (these are the data from example 2.2.6):

(***)25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
*57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/I and the SD is 40.4 U/I. What %-age of the observations are within

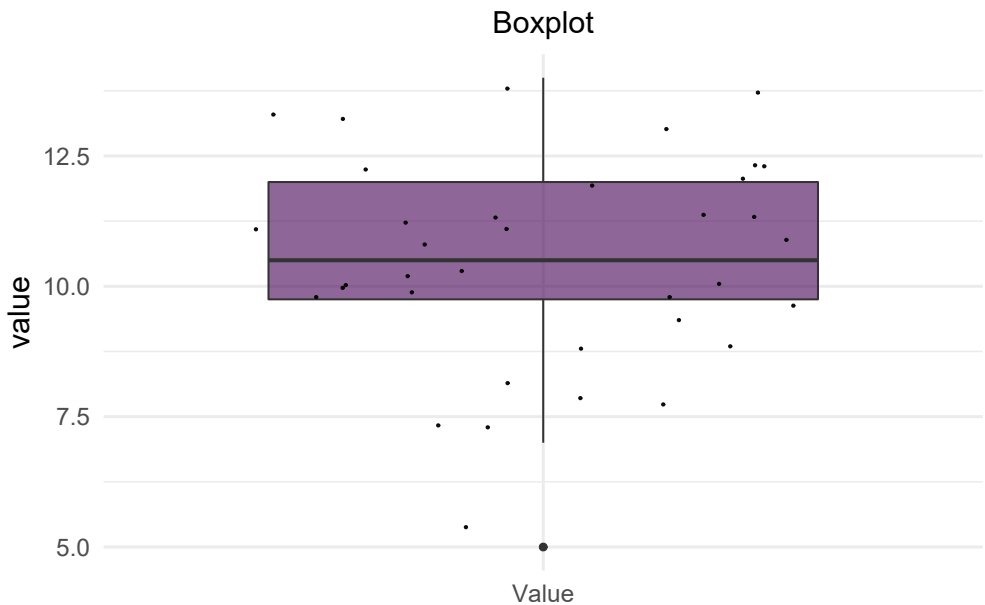
  - 1 SD of the mean? 26/36
  - 2 SDs of the mean? 34/36
  - 3 SDs of the mean? 36/36

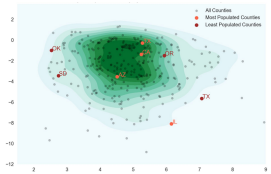


Example

Find the mean, standard deviation, and construct a boxplot:

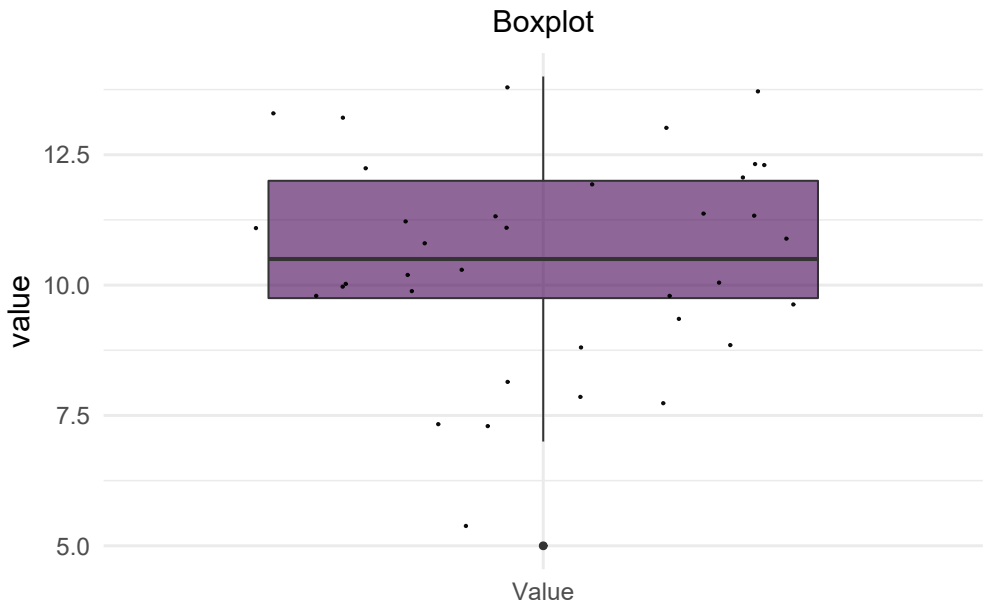
Number of piglets	Frequency (number of sows)	$f_i * y_i$
5	1	5
6	0	0
7	2	14
8	3	24
9	3	27
10	9	90
11	8	88
12	5	60
13	3	39
14	2	28
	$n = 36$	$\sum f_i \times y_i = 375$





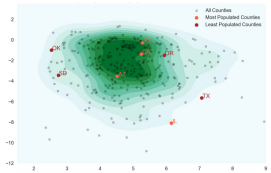
Example

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2



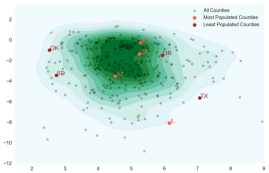
(a) $s = 1.99$ and $\overline{y} = 10.42$

Transformations



Data Transformations

- Big part of any statistical analysis
- Why? Maybe you need to change the scale, maybe you wanna transform the shape (i.e. to go from exponential to linear), or maybe you wanna change your units

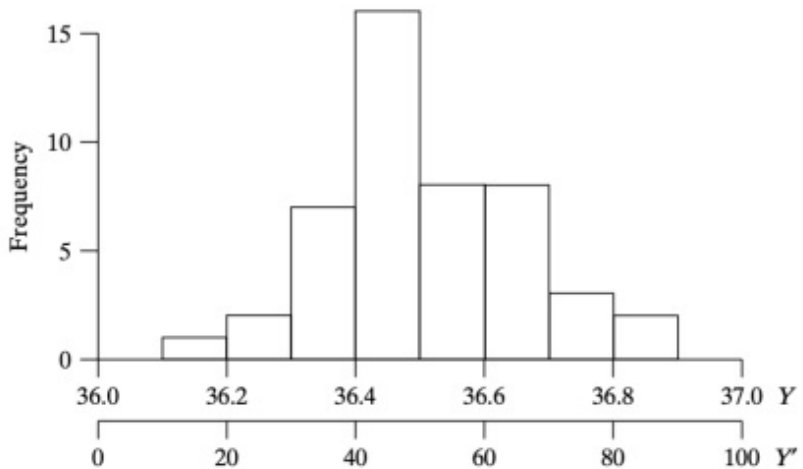


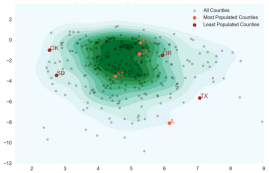
Linear Transformations

- Generally, of the classical linear form y=mx+b
- Does not change the shape of the distribution
- Scale data by multiplication/division or shift data by adding/subtracting (or both)

Figure 2.7.1 Distribution of 47 temperature measurements showing original and linearly transformed scales

$$Y' = (Y - 36) \times 100$$





Linear Transformations
Continued

- We can add or subtract a constant to the original variable

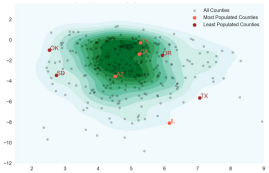
$$\text{If } Y' = Y \pm C \implies \overline{y}' = \overline{y} \pm C \text{ and } s' = s$$

- Multiply by a constant to the original variable

$$\text{If } Y' = mY \implies \overline{y}' = m \cdot \overline{y} \text{ and } s' = m \cdot s$$

- Multiply by a constant to the original variable and then add/subtract a constant

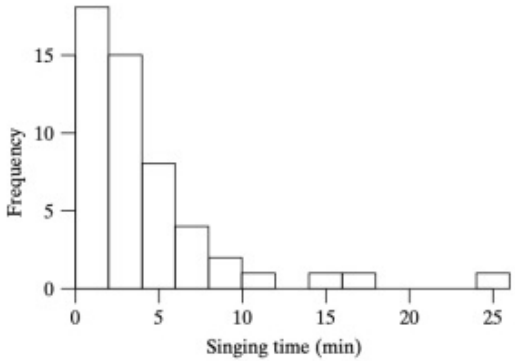
$$\text{If } Y' = mY \pm C \implies \overline{y}' = m \cdot \overline{y} \pm C \text{ and } s' = m \cdot s$$

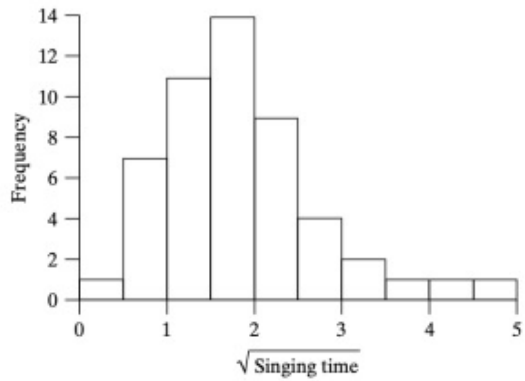


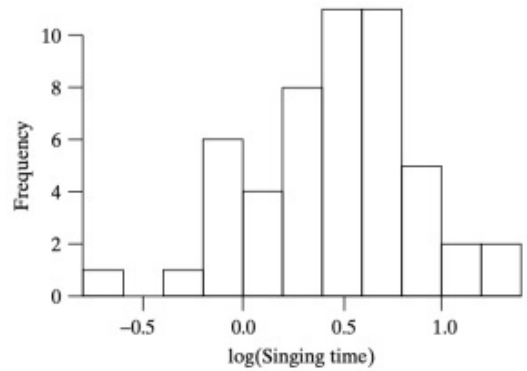
Other Transformations

$$Y' = \sqrt{Y}$$
$$Y' = \ln(Y)$$
$$Y' = \frac{1}{Y}$$
$$Y' = Y^2$$

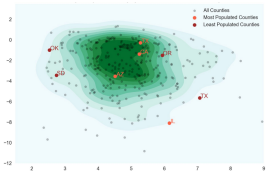
Square root and log transformations pull right tail inward and push out left tail





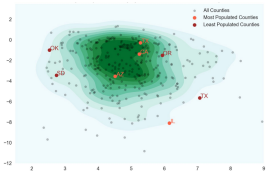


Statistical Inference



Inference

- **Statistical Inference**: Methods for making predictions on population based on data from a sample
- Goal: results from any sample would be identical or nearly identical to the results obtained from the population
- The population has its own distribution
- For a simple random sample, the sample distribution approximates the population distribution
- The larger the sample size, the better the approximation
- Statistics describes a sample. Parameters describe a population



Inference (continued)

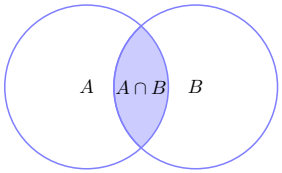
- Statistics desribe sample characteristics, $\overline{y}$ and s are sample mean and sample standard deviation
- Parameters describe population characteristics. Usually trying to estimate these
- Proportions: are a relative frequency, can be for population or sample

Measure	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\overline{y}$	μ
Standard deviation	s	σ

Chapter 9 (Part 1) Notes

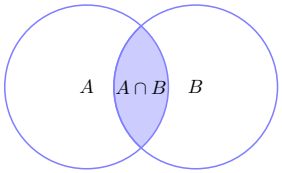
Essential Probability Rules
STP-231

Arizona State University



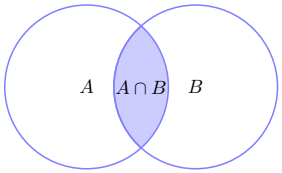
Probability Theory

- **Probability theory**: the science of uncertainty or chance
- **Randomness**: individual outcomes are uncertain, but there is nonetheless a regular distribution of outcomes with repetition
- **Probability of an outcome**: for random phenomenon, the proportion of times the outcome occurs over many repetitions



Probability and the Life Sciences

- Probability can determine the likelihood of a certain event
- In life sciences: example is is the likelihood of twins being born
- Do outside factors influence this occurence? Do twins run in mother;s line of the family? Has mother taken fertility drugs? Is the mother a hollywood actress?
- Conclusions in statistical analysis are reported in probabilities:
- What is the likelihood that a particular sample is different from a population or from another sample?
- How likely or unlikely is an experimental result



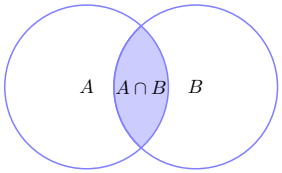
Frequentist vs Personal Probabilities

Frequentist approach

- A very large random sample can be used to approximate probabilities of random phenomenon
- Long run probability

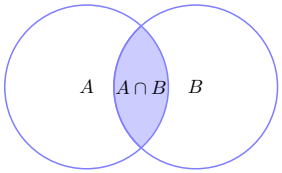
Personal

- Subjective probability based on one's own experience and judgement
- A probability nonetheless



Probability Models

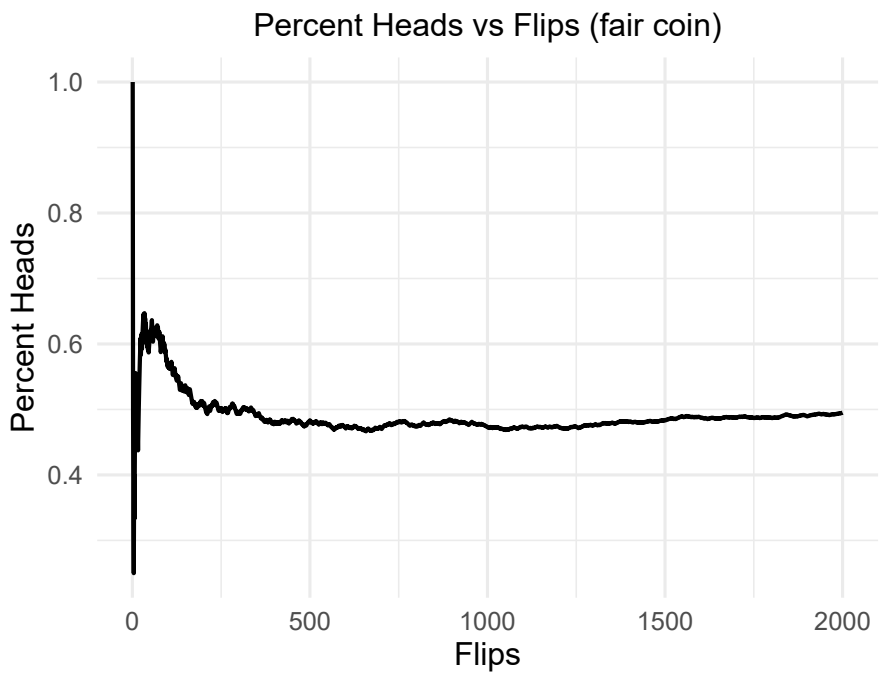
- **Sample space** ($\mathcal{S}$): Includes all possible outcomes of a random phenomenon
- **Event**: An outcome or set of outcomes of a random phenomenon, and is a subset of the sample space
- A probability model of a random phenomenon mathematically defines a sample space $\mathcal{S}$.
- How do we assign probabilities of events within $\mathcal{S}$?

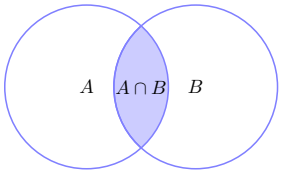


Example: Flipping a coin

Example: flipping a coin. $\mathcal{S}$ is the possible outcomes, i.e. some enumeration of H and T (heads/tails). Denote as $\mathcal{S} = \{H, T\}$. The event E in this example is how many heads we get,

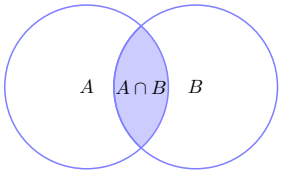
$$\Pr(E) \leftrightarrow \frac{\text{Number of times we flip head}}{\text{Number of times we flip the coin}}$$





Probability Rules

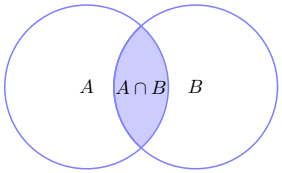
- Probability of an event, $\Pr(E)$, is a value that is always between 0 and 1, inclusive
- Probability of an impossible event is 0
- Probability of a certain event is 1 (event will always happen)
- The sum of probabilities of all events equals 1
- When the two events have no outcomes in common, they can never happen together, i.e. their joint probability is zero
- The probability that one or the other occurs is the sum of their individual probabilities minus probability both happen, $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$
- The probability an event does not occur is 1 minus the probability event does occur, i.e. $\Pr(E^c) = 1 - \Pr(E)$



Discrete Probability Models

- A probability model with a sample space made up of alist of every individual outcome

- The probability of any event is the sum of the probabilities of the outcomes making up the event



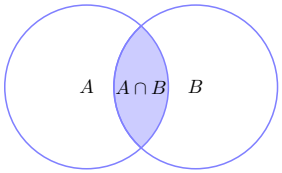
Equal Likelihood Model

- Every possible sample from the total number of possible outcomes is equally likely to be chosen
- The probability of an event E is defined as

$$\Pr(E) = \frac{\#\text{ of outcomes in event}}{\text{Total outcomes in sample space}}$$

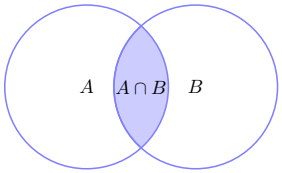
The procedure:

- Identify all possible outcomes in the sample space
- Identify the event
- Identify the outcomes that belong in the event
- Calculate ratio



Example

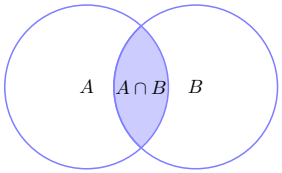
A test consists of two true or false problems. Suppose the answers to both questions are true. A student guesses both questions by flipping a coin. If it lands on heads, the student chooses true as the final answer and if lands on tails, the student chooses false as the final answer. What is the probability that a student gets exactly one out of two questions correct?



Example 2

Five students, Riley (R), Toro (T), Eddie (E), Jas (J), and Kelly (K), are all allergic to flaxseed. Two out of the five will be tested on a new genetically altered flaxseed to see if the two students are still allergic.

- What is the probability that Toro and Kelly (event A) are chosen for the study?



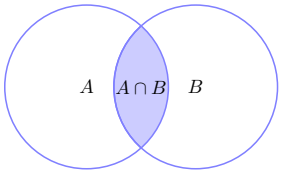
Example 2

Five students, Riley (R), Toro (T), Eddie (E), Jas (J), and Kelly (K), are all allergic to flaxseed. Two out of the five will be tested on a new genetically altered flaxseed to see if the two students are still allergic.

- What is the probability that Toro and Kelly (event A) are chosen for the study? 1/10
  The ten events in $\mathcal{S}$ are

$$\mathcal{S} = \Big\{(R,T),(R,K),(R,J),(R,E),(T,K),(T,J),(T,E),(K,J),(K,E),(J,E)\Big\}$$

  The probability either or is in the study?



Example 2

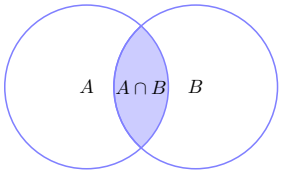
Five students, Riley (R), Toro (T), Eddie (E), Jas (J), and Kelly (K), are all allergic to flaxseed. Two out of the five will be tested on a new genetically altered flaxseed to see if the two students are still allergic.

- What is the probability that Toro and Kelly (event A) are chosen for the study? 1/10
  The ten events in $\mathcal{S}$ are

$$\mathcal{S} = \Big\{(R,T),(R,K),(R,J),(R,E),(T,K),(T,J),(T,E),(K,J),(K,E),(J,E)\Big\}$$

  The probability either or is in the study? 7/10
  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$

- What is the probability that Riley (Event B) is chosen for the study?



Example 2

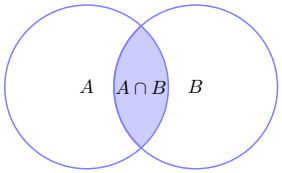
Five students, Riley (R), Toro (T), Eddie (E), Jas (J), and Kelly (K), are all allergic to flaxseed. Two out of the five will be tested on a new genetically altered flaxseed to see if the two students are still allergic.

- What is the probability that Toro and Kelly (event A) are chosen for the study? 1/10
  The ten events in $\mathcal{S}$ are

$$\mathcal{S} = \Big\{(R,T),(R,K),(R,J),(R,E),(T,K),(T,J),(T,E),(K,J),(K,E),(J,E)\Big\}$$

  The probability either or is in the study? 7/10
  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$

- What is the probability that Riley (Event B) is chosen for the study? 4/10. Write out all the combinations see Riley shows up 4 times.



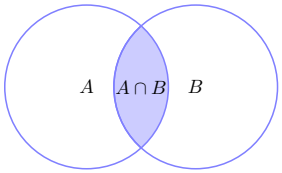
Random Sampling and Probabilities

When sampling from a population

- The probability for randomly choosing an individual with a certain characteristic is equivalent to the population proportion (relative frequency) of individuals with that charactertistic in the population

- EXAMPLE: A large population of the fruitflu Drosophila melanogaster is maintained in a lab. 30% of individuals are black, 70% are grey in the population. Suppose one fly is chosen at random from the population. Then the probability that a black fly is chosen is 0.3. More formally, define

$$E\text{: sampled fly is black, then } \Pr(E) = 0.3$$



Sampling from Populations
Continued

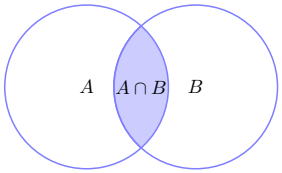
In a certain population of the freshwater sculpin, Cottus rotheus, the distribution of the number of tail vertebrae is shown in the table[1]

# of vertebrae	% of fish
20	3
21	51
22	40
23	6

Find the probability that the number of tail vertebrae in a fish randomly chosen from the population

- Equals 21
- Is less than or equal to 22
- is greater than 21
- is no more than 21

[1] 3.2.1 from Statistics for the Life Sciences by Samuels

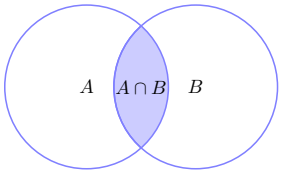


Sampling from Populations
Another Example

The effectives of a seed eating weevil on population control of a nonnative invasive species of a tree in South America called the Paraserianthes Iopantha was studied.

% Seed damage	# of trees
0-9	19
10-19	2
20-29	5
30-39	3
40-49	6
50-59	2
60-69	2

- Find probability of event that the tree has 20-29% seed damage
- At least 50% seed damage
- Less than 40% seed damage
- At least 30% but less than 59%
- Probability of all the previous 4 events occurring together



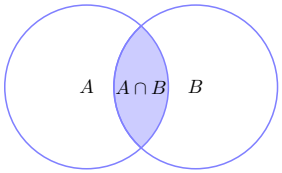
Continuous Probability Models

Density Curve

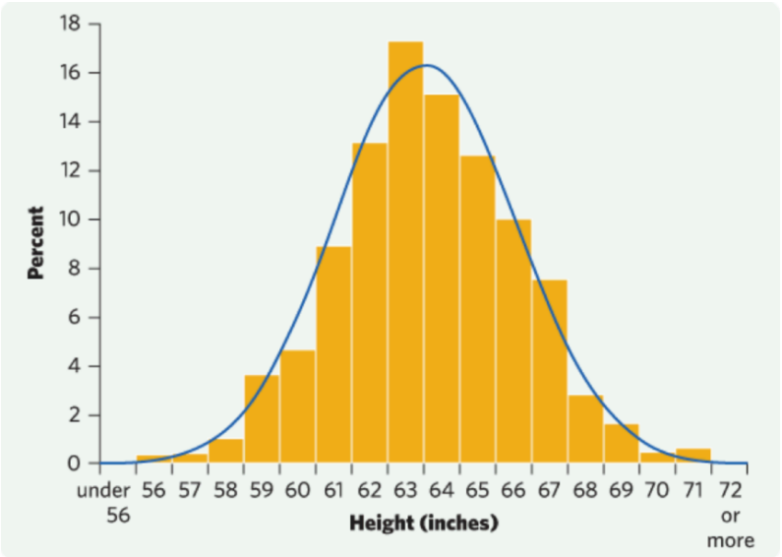
- Is always on or above the horizontal axis
- has area exactly 1 underneath
- Area under the curve and above any range of values on the horizontal axis is the proportion of all observations that fall within that range

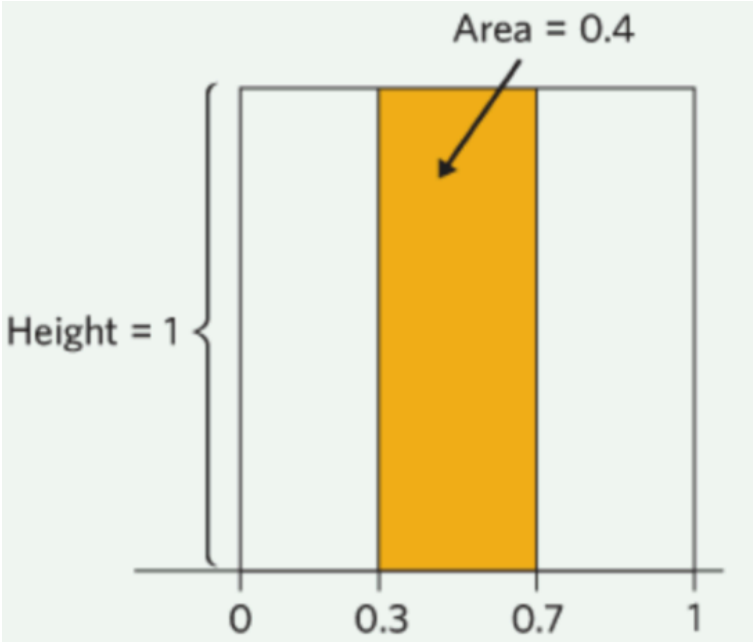
Continuous probability model:

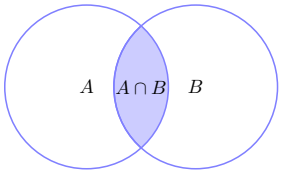
- Assigns probabilities as areas under a density curve



Pictoral Representation





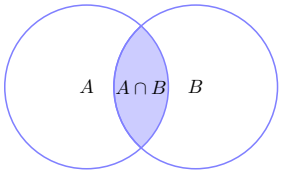


Example

Example 9.11 in the text, from previous slide.

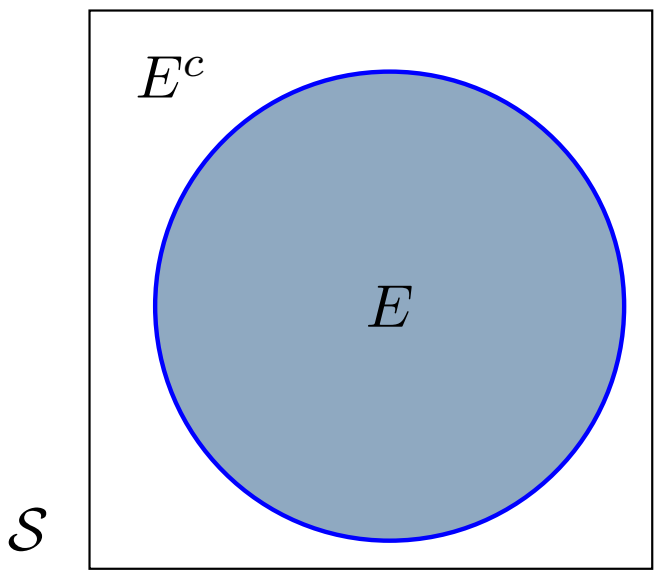
- $\mathcal{S} = \{\text{all numbers between 0 and 1}\}$
- Consider a random number generator that chooses numbers in $\mathcal{S}$. The random number generator will spread its output uniformly. The density curve is on the right.
- How do we assign probabilities to intervals?
- How do we assign probabilities individual values?

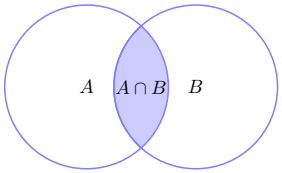
Venn Diagrams



Venn Diagrams

- Visuals to display events and relationships among events
- The sample space, $\mathcal{S}$, is the space that includes all possible outcomes in an experiment, and an event, E, is a subset of the sample space



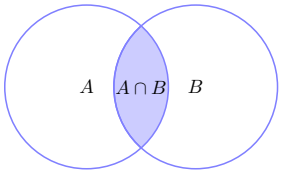


The Complement of E

- E^c : the event E does not occur
- For any event E:

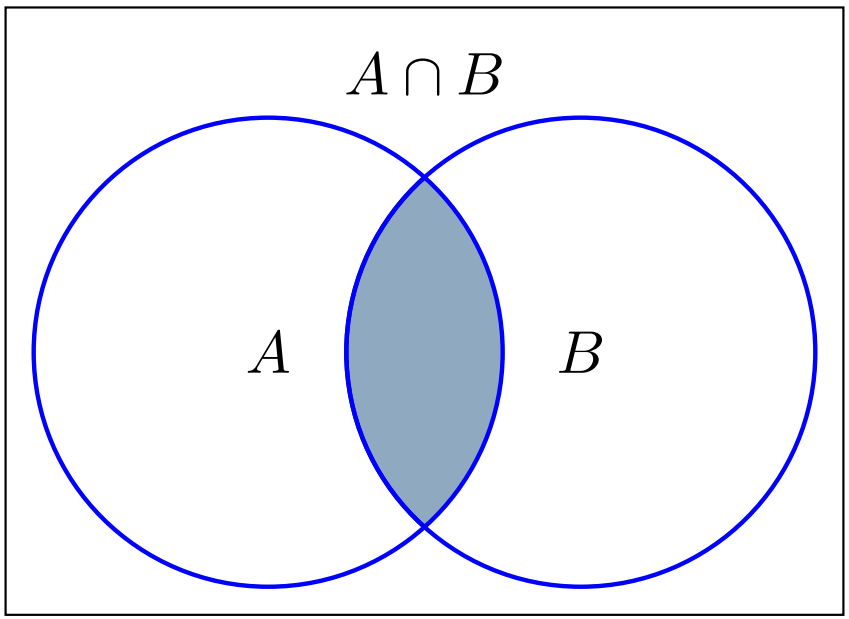
$$\Pr(E) = 1 - \Pr\left(E^c\right)$$
$$\Pr(E^c) = 1 - \Pr(E)$$

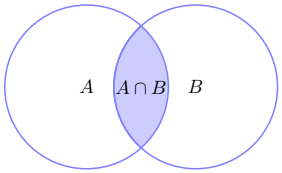
- For example, if event A is the suit hearts in a deck of cards, the complement is clubs, diamonds, and spades
- If event B is a worker making at least 40,000 dollars annually, the complement is workers making less than that



Intersection of Events

- A and B: The event both A and B occur, denoted $A \cap B$

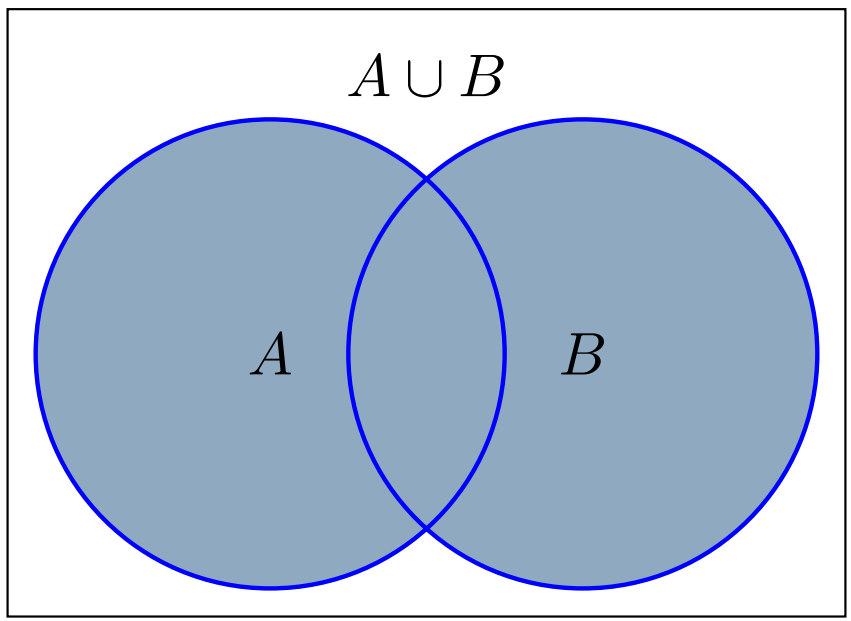


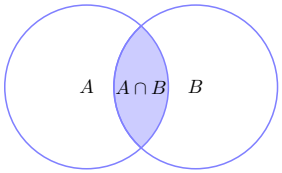


Union of Events

- A or B: The event either A or B occur, denoted $A \cup B$.

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$$



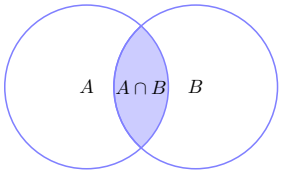


Example

When we roll a die, there are six possible outcomes. Define

- Event A: The die comes up even
- Event B: The die comes up odd
- Event C: The die comes up 6

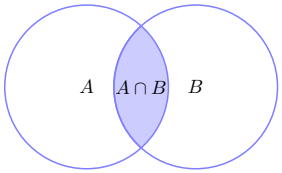
List the outcomes in not A, $A\&C$, A or C, B & C



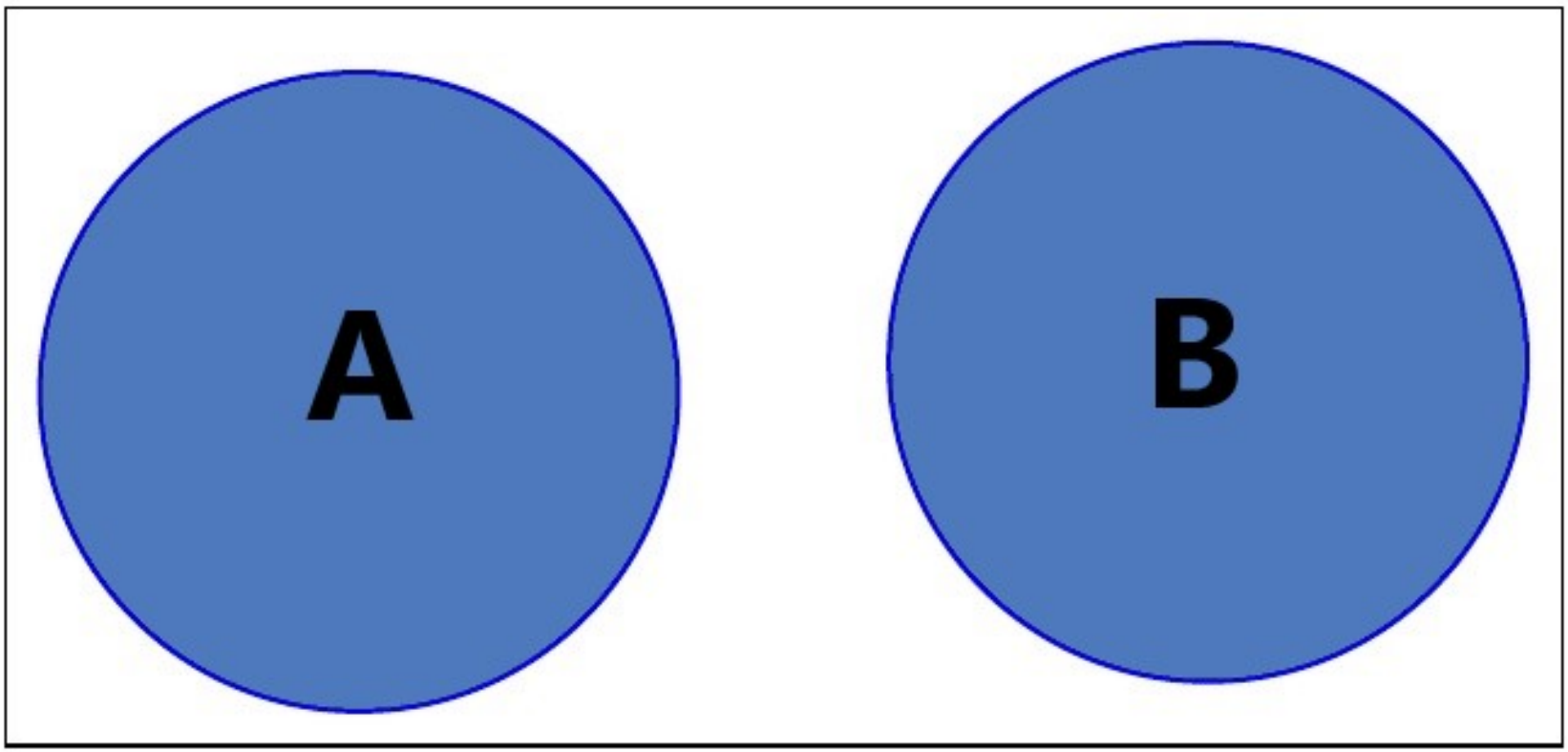
Mutually Exclusive

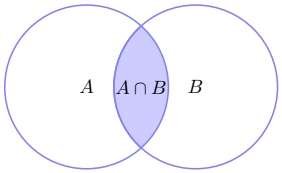
Mutually exclusive events: Two or more events that can not occur together. If A and B are mutually exclusive

- $\Pr(A\&B) = 0$
- $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$
- In general, if $A, B, C, \ldots$ are all mutually exclusive:
- $\Pr(A\&B\&C\&\ldots) = 0$
- $\Pr(A \text{ or} B \text{ or } C \text{ or } \ldots) = \Pr(A) + \Pr(B) + \Pr(C) + \ldots$



Mutually Exclusive
Picture

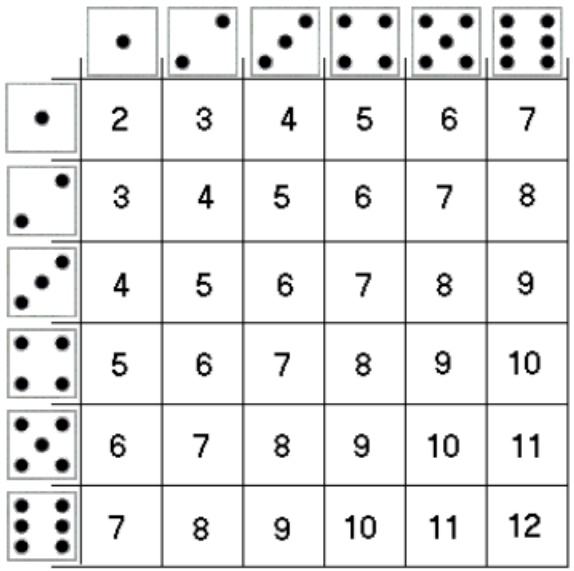




Example

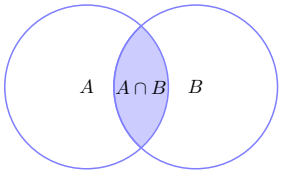
Roll two dice, 36 equally likely outcomes

- A: Event that sum of the dice is 7
- B: Event that sum of the dice is 11
- C: Event that roll of the dice is odd
- D: Event that sum of the dice is 8
- E Event that roll of the dice is doubles



	2	3	4	5	6	7
	3	4	5	6	7	8
	4	5	6	7	8	9
	5	6	7	8	9	10
	6	7	8	9	10	11
	7	8	9	10	11	12

Find the probabilities $\Pr(A)$, $\Pr(A \text{ or } B)$, $\Pr(C \text{ and } D)$, $\Pr(D \text{ and } E)$, $\Pr(D \text{ or } E)$

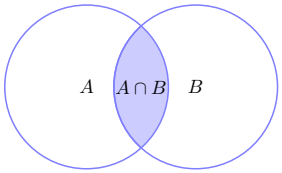


Example 2

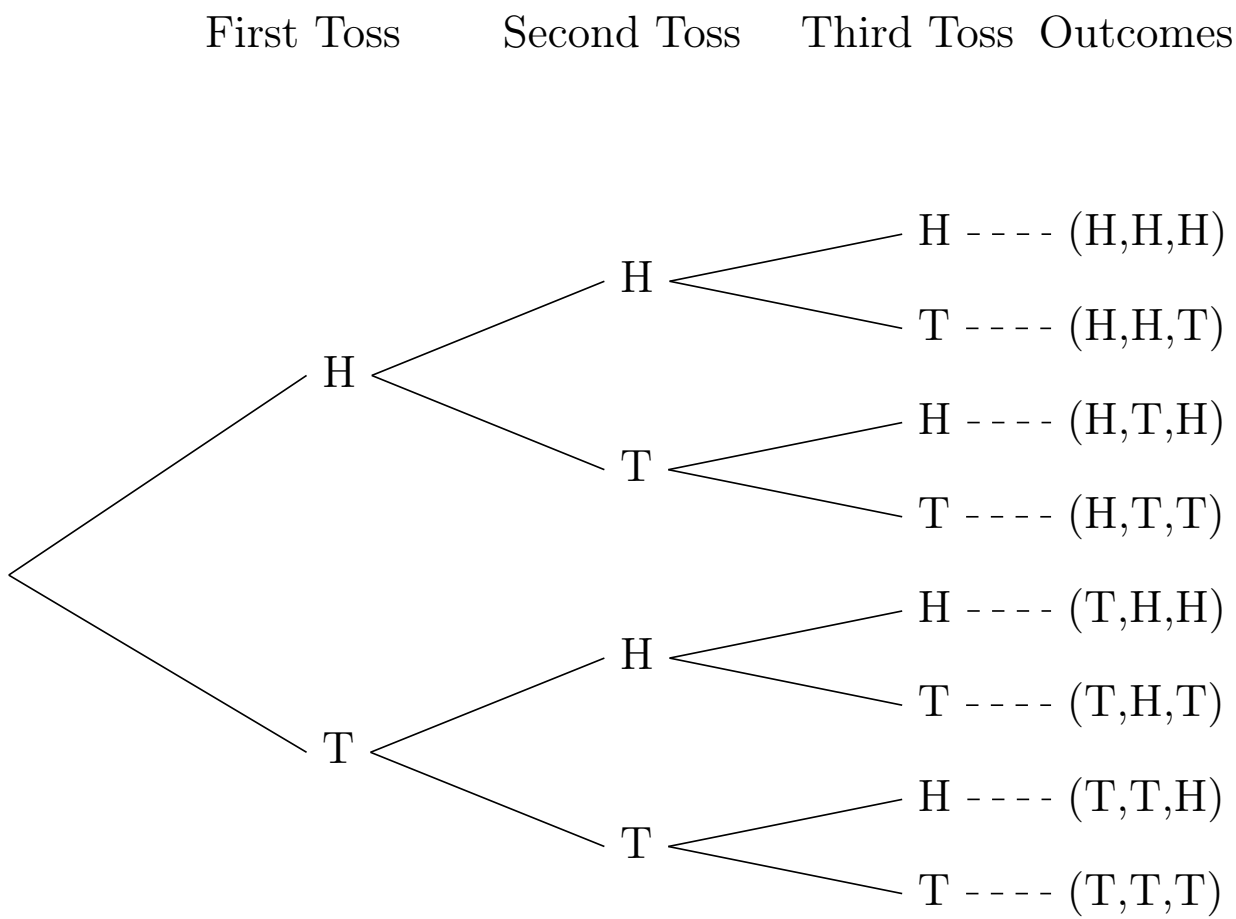
There are five associates on duty in a Staples office supply store: three women (Maggie, Linda, Vanessa) and two men (Gary, Pablo). An experiment consists of classifying the next customer's action. They will make a purchase from exactly one of the sales associates or buy nothing. What is the probability the next customer buys from Maggie or Pablo?

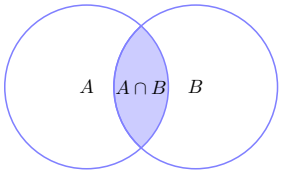
Action	Probability
Buy from Maggie	0.08
Buy from Linda	0.12
Buy from Vanessa	0.10
Buy from Gary	0.25
Buy from Pablo	0.15
Buy nothing	0.30

Probability Trees



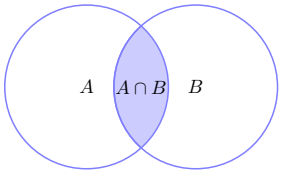
Probability Tree





Using Probability Trees

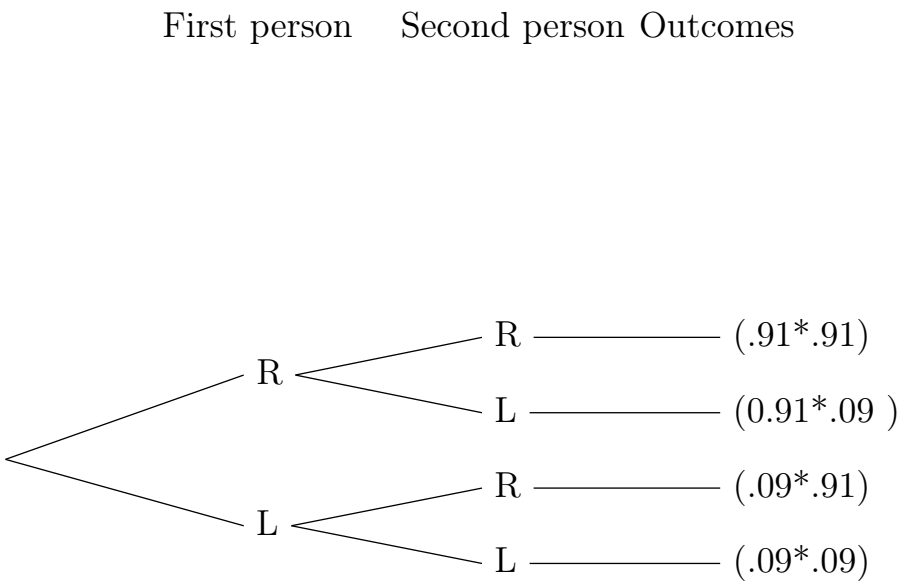
- Outcome probabilities:
- Travel ALONG branches
- When you are traveling along branches you multiply the probabilities
- Event probabilities
- Choose outcomes that satisfy event
- Sum those probabilities

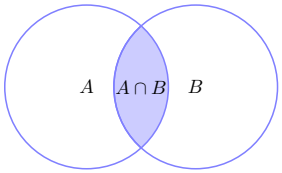


Example

About 9% of people are left-handed. You randomly select two people

- What is the probability they are both right-handed?
- What is the probability that at least one right-handed person is selected?





Another Example

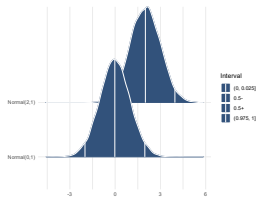
From 3.2.4 in Samuels Statistics for the Life Sciences

- Suppose that a disease is inherited via a sex-linked mode of inheritance so that a male offspring has a 50% chance of inheriting the disease. Further suppose that 51.3% of births are male. What is the probability that a randomly chosen child will be affected by the disease?

Chapter 9 (Part 2) Notes

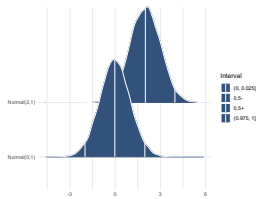
Random Variables and Density Curves
STP-231

Arizona State University



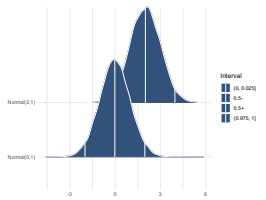
Relative Frequency Hist./Density Curves

- Consider a relative frequency histogram as an approximation of the underlying true
- It is often desirable to describe a population frequency distribution with a smooth curve (especially for continuous variables)
- We can idealize a density curve as a relative frequency histogram with very narrow classes



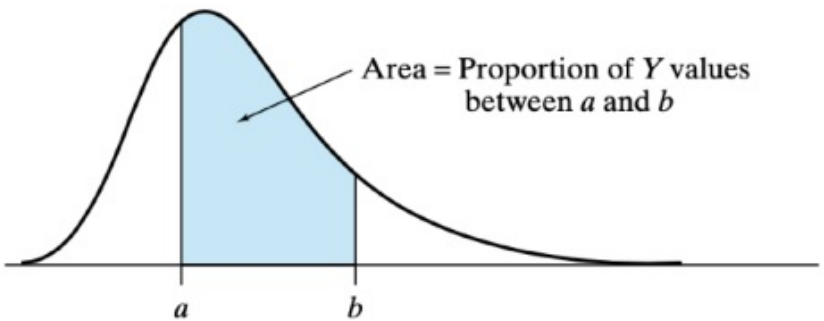
Density Curves

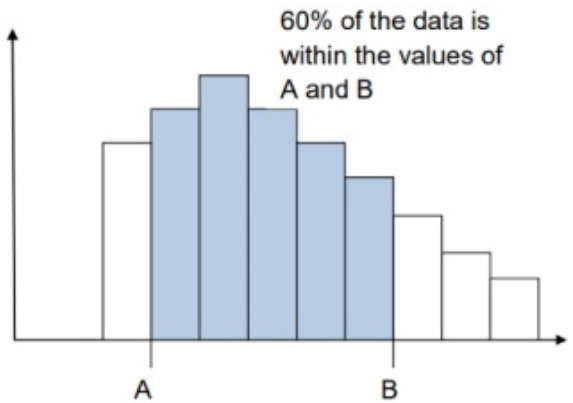
- **Density Curve**: a smooth curve representing a frequency distribution
- **Density Scale**:
- The plot of the vertical coordinates on the density curve are plotted on this scale
- Relative frequencies are represented as areas under the curve

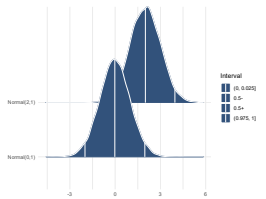


Relative Frequency Expansion

- Relative frequencies are probabilities
- Relative frequencies add up to one
- The area under a density curve or a relative frequency histogram adds up to one
- We can use this to find proportions between particular values

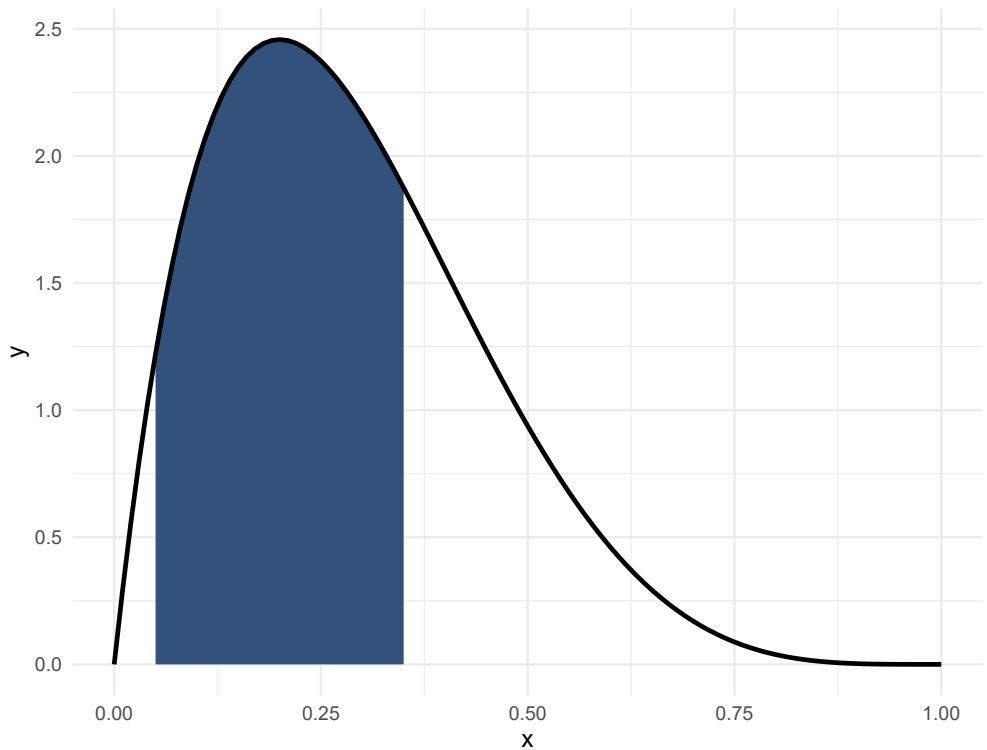


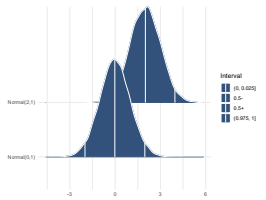




Distribution

- For any two numbers A and B, the area under the density curve between A and B is the same as the proportion of $y-$values between A and B
- The total area under the curve is 1

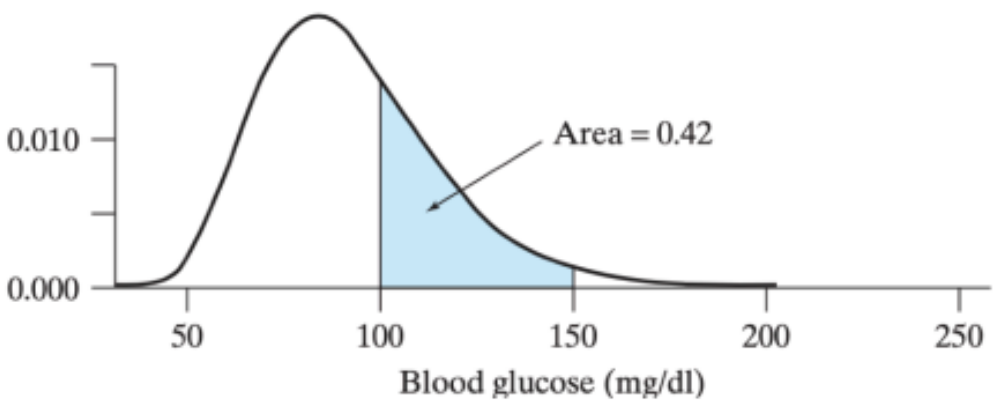


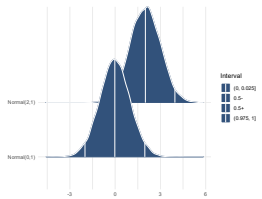


Probabilities and Density Curves

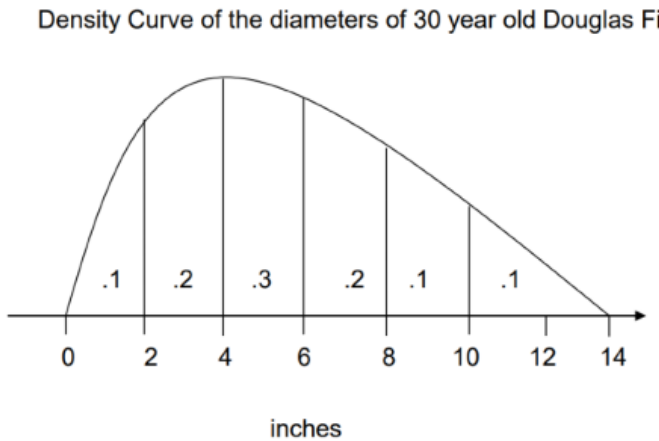
- The relative frequency of a specific Y value is zero, i.e. $\Pr(Y = \#) = 0$. Therefore, we do not discuss the relative frequency of a single Y value
- However, we can assign probabilities to intervals, defined as the area under the curve

$$\Pr(A \leq Y \leq B) = \Pr(A < Y < B)$$



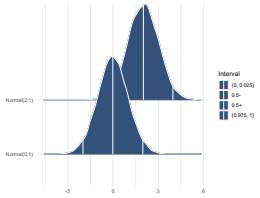


Example



- Find the probabilities $\Pr(d = 4)$, $\Pr(0 < d < 4)$, $\Pr(d < 6)$, $\Pr(d > 8)$, and $\Pr(d < 10)$
- Now assume we take a sample of two trees, which we consider as independent events. Find the probability:
- both trees have diameter less than 4".
- Diameter of first tree less than 8", 2nd tree greater than 8".
- Exactly one tree has diameter less than 4" and exactly one tree has diameter greater than 8".

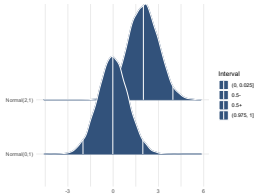
Random Variables



Random Variables

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon

- Can either represent discrete or continuous values
- Example: In a population of flies, the random variable X represents the amount of flies that are still alive after 24 hours. X takes on values $\{0, 1, 2, \ldots\}$
- Random number generatror generates numbers in the interval [0,1]. The random variable Y represents a number chosen and takes on values in [0,1].



Probability Distribution

The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values

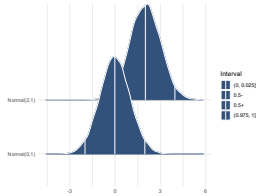
- Example: Roll a die. Y represents the number of spots on the side facing you. Therefore, Y=1,2,3,4,5,6. We do not know what Y will be till we toss the die. For all we know it could be a weighted dice.
- Note, this is a **discrete** probability distribution, so the random variable Y can take specific values with a probability attached.

Fair Dice

y_i	1	2	3	4	5	6
$\Pr(Y = y_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Weighted Die

y_i	1	2	3	4	5	6
$\Pr(Y = y_i)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$



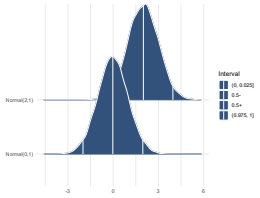
Probability Distribution (Continued)

- 3.5.2 from Statistics for the Life sciences: Y denotes nymber of children in a family chosen at random. $Y = 0, 1, 2, 3, \ldots$. The probability Y has a particular value is equal to the %-age of families with that many children. For example, if 23% of families have 2 kids, then

$$\Pr(Y = 2) = 0.23$$

- Example 3.5.3 from text. Y is random variable denoting number of medications a patient is given following cardiac surgery. If 52% of all patients are given 2,3,4, or 5 medications, then

$$\Pr(2 \leq Y \leq 5) = 0.52$$

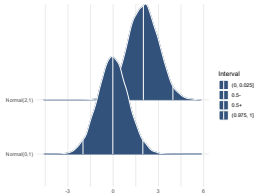


Mean of a Discrete Random Variable

- Let X be a discrete random variable with k elements $x_1, x_2, \ldots, x_k$ in its sample space $\mathcal{S}$. The mean of X is

$$\sum_{i=1}^{k} x_i \times \Pr(X = x_i)$$

- This is also called the expected value: $E(X) = \mu_X$
- Named expected value because over several repetitions of the random event we expect the value of X to be μ_X



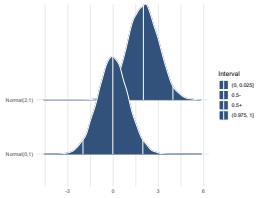
Example 1

Die Roll	Probability $\Pr(Y = y_i)$	Winnings (y_i)
1	0.5	\$1
2	0.3	\$5
3	0.2	\$-10

- Imagine we have a game with a 3-sided die. What is the expected value?

$$E(Y) = \sum_{i=1}^{3} y_i \cdot \Pr(Y_i = i) = (1 \cdot 0.5) + (5 \cdot 0.3) + (-10 \cdot 0.2) = 0$$

- So if you played enough times you'd expect to win 0 dollars. Would you still play one time though? Is the 80% chance of winning something enough to outweigh the fear of losing big 20% of the time?



Variance & std. deviation (Discrete R.V.)

- Let X be a discrete random variable with k elements $x_1, x_2, \ldots, x_k$ in its sample space $\mathcal{S}$. The variance of X is:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \times \Pr(X = x_i)$$

- The standard deviation is the square root of the variance

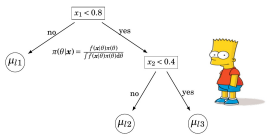
$$\sigma = \sqrt{\sigma^2}$$

- The standard deviation is a measure of dispersion away from the mean that takes into account both how far each value is from the mean and how likely each value is

Chapter 10 Notes

Conditional Probability, Independent Events, & Bayes' Formula
STP-231

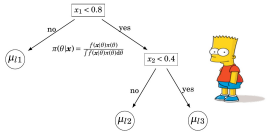
Arizona State University



Independence

- **Dependent events**: The occurrence of one event **changes** the probability of a different event occurring
- Example: The probability Riley gets an A on their test depends on whether or not they study more than 3 hours

- **Independent events**: The occurrence of event does not change the probability of a second event occurring
- Example: We assume the probability of a flipping a coin tails is independent of the previous flip



Independence
Continued

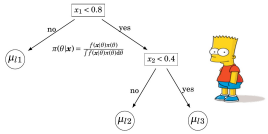
- If events A and B are independent, then:

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$$

- If they are not independent, then:

$$\Pr(A \text{ and } B) \neq \Pr(A) \times \Pr(B)$$

- Independence is often assumed in a probability model where the events seem to have no connection. Also, its convenient and can make the problem easier, though we need just justification to claim independence!



Mutually Exclusive vs Independent Events
Not the same!

- Example: $\Rightarrow$ Roll a die, A is event of rolling a 2,4, or 6, and event B is rolling a 6. Since the rolls are independent, this is an independent event, i.e.

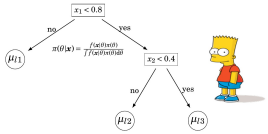
$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12} \neq 0$$

Independent, but not mutually exclusive

- Example: $\Leftarrow$ Again roll a die. Let A be event your roll is an odd, and event B you roll a 6. These are mutually exclusive, but

$$\Pr(B) = 1/6 \qquad \Pr(A) = 1/2 \text{ but } \Pr(A \cap B) = 0$$

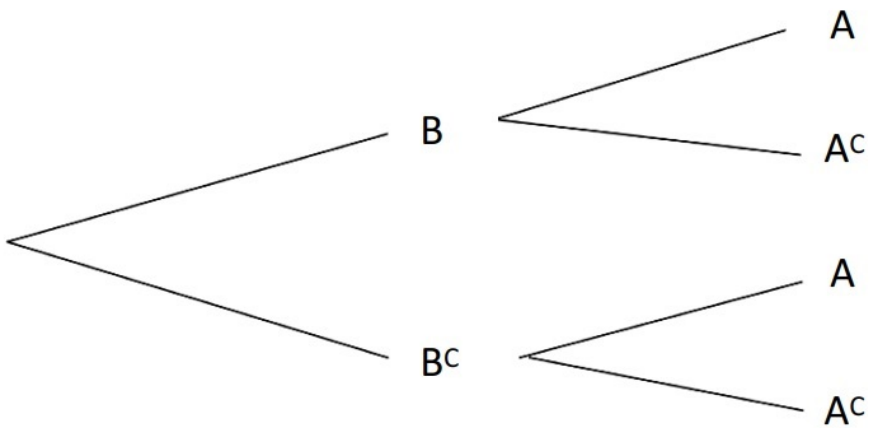
But these events are dependent, as rolling a 6 means you roll an even, and event A means you rolled an odd, so these two can't happen simultaneously. In general, mutually exclusive events are dependent



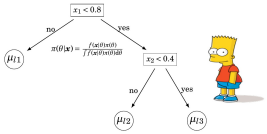
Conditional Probability

- Probability of an event occurring given that another event has already occurred:

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}, \quad \text{provided } \Pr(B) > 0$$



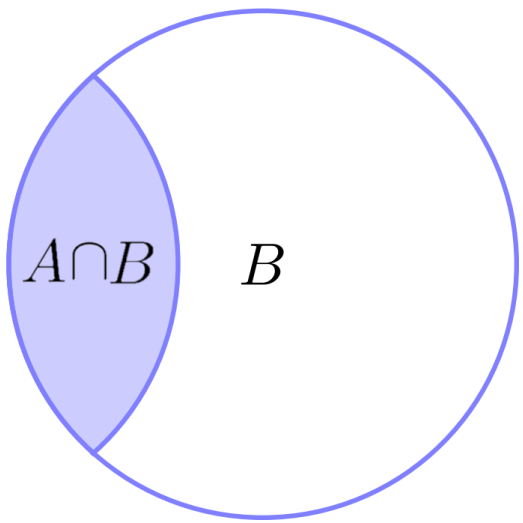
- Example: Roll fair dice once. Given an odd number was rolled, what is the probability of rolling a one. Given an even number was rolled, what is the probability of rolling a two or six?

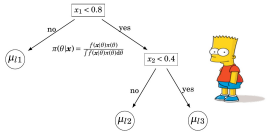


Venn Diagram Approach

We know with certainly the B event did occur, so the probability of A in this situation is the shaded region in the picture (the only place where A occurs given that B for sure occurred) divided by the total probability of B, i.e.

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$



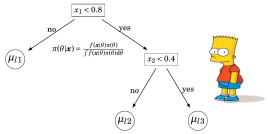


Example

The culinary herb cilantro is very polarizing: Some love it and others hate it (some say it tastes like soap). Here are the results from a survey asking 12087 people whether or not they like/dislike cilantro

	# of Men	# of Women
Dislikes cilantro	1632	1549
Likes cilantro	5165	3741

- What is the approximate conditional probability that an adult likes cilantro given they are a man

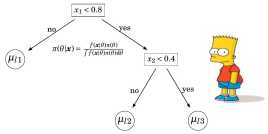


Example

The culinary herb cilantro is very polarizing: Some love it and others hate it (some say it tastes like soap). Here are the results from a survey asking 12087 people whether or not they like/dislike cilantro

	# of Men	# of Women
Dislikes cilantro	1632	1549
Likes cilantro	5165	3741

- What is the approximate conditional probability that an adult likes cilantro given they are a man 5165/(1632+5165)
- What is the approximate probability this person is a man and likes cilantro?

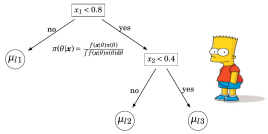


Example

The culinary herb cilantro is very polarizing: Some love it and others hate it (some say it tastes like soap). Here are the results from a survey asking 12087 people whether or not they like/dislike cilantro

	# of Men	# of Women
Dislikes cilantro	1632	1549
Likes cilantro	5165	3741

- What is the approximate conditional probability that an adult likes cilantro given they are a man 5165/(1632+5165)
- What is the approximate probability this person is a man and likes cilantro? 5165/12087

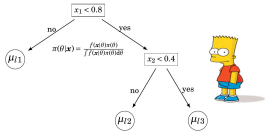


Another Example

- Your friend tells you they will visit you one week-day in the coming week, with a probability of 0.20 of visiting on any given week day.

- Given that it is Thursday and your friend has yet to visit, what is the probability they will visit today?

$$\begin{aligned}\Pr(A \mid B) &= \frac{\Pr(A \cap B)}{\Pr(B)} \\ &= \frac{\Pr(\text{visit today and not M, T, or W})}{\Pr(\text{did not visit M, T, or W})} \\ &= \frac{0.20}{1 - 0.60} = 0.50\end{aligned}$$



Conditional Probability Rules

- General multiplication rule: the probability of events A and B happening together is (both are equivalent):

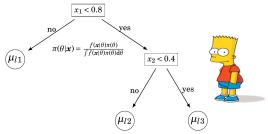
$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B \mid A) \quad \text{or} \quad \Pr(B) \times \Pr(A \mid B)$$

- We also know from earlier rules that:

$$\Pr(A^c \mid B) = 1 - \Pr(A \mid B)$$
$$\Pr(A \mid B) = 1 - \Pr(A^c \mid B)$$

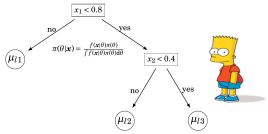
- If two events A and B are independent, then

$$\Pr(A \mid B) = \Pr(A)$$



Diagnostic Tests

- True Positive: The test states that the person/item has the disease when they really have the disease
- False Negative: The test states that the person/item does not have the disease when they do have the disease
- False Positive: The test states that the person/item has the disease when they really do not have the disease
- True Negative: The test states that the person/item does not have the disease when they do not have the disease



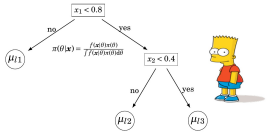
Diagnostic Tests
Example

- Suppose a medical test has a 92% chance of detecting a disease given the person has the disease and a 94% chance of correctly indicating that the disease is absent if the person really does not have the disease. Suppose 10% of the population has the disease
- What is the probability that a randomly chosen person will test positive? (True positive + false positive)

$$0.10 * 0.92 + (1 - 0.94) * (1 - 0.10) = 0.146 = 14.6\%$$

- Suppose a randomly chosen person does test positive. What is the probability that this person really has the disease?

$$0.10/0.146 = 0.685 = 68.5\%$$

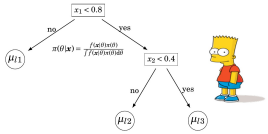


Contingency Tables

- Useful to display bivariate data. Also helpful to determine prediction accuracy, to calculate joint and conditional probabilities, and see balance in data outcomes[1]

	Have young kid		
Sleep $<$ 7 hours	Yes	No	Total sleep
Yes	21	122	143
No	70	493	563
Total Had kid	91	615	706

[1]Data from Wooldridge Economics

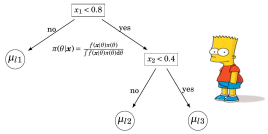


Contingency Tables
Example

In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions. The following data table is taken from the study, relating to the comparison between income and the amount of stress reported by the people in the study

	Income			
	Low	Medium	High	Total
Stressed	526	274	216	1,016
Not Stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

- Pr(Low Income)

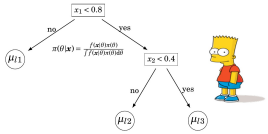


Contingency Tables
Example

In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions. The following data table is taken from the study, relating to the comparison between income and the amount of stress reported by the people in the study

	Income			
	Low	Medium	High	Total
Stressed	526	274	216	1,016
Not Stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

- Pr(Low Income) = 2480/6549
- Pr(Stressed and Low Income)

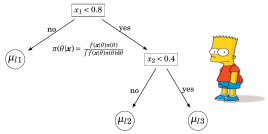


Contingency Tables
Example

In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions. The following data table is taken from the study, relating to the comparison between income and the amount of stress reported by the people in the study

	Income			
	Low	Medium	High	Total
Stressed	526	274	216	1,016
Not Stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

- Pr(Low Income) = 2480/6549
- Pr(Stressed and Low Income)= 2480/6549
- Pr(Stressed | Low Income)

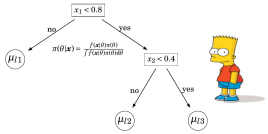


Contingency Tables
Example

In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions. The following data table is taken from the study, relating to the comparison between income and the amount of stress reported by the people in the study

	Income			
	Low	Medium	High	Total
Stressed	526	274	216	1,016
Not Stressed	1,954	1,680	1,899	5,533
Total	2,480	1,954	2,115	6,549

- Pr(Low Income) = 2480/6549
- Pr(Stressed and Low Income)= 2480/6549
- Pr(Stressed | Low Income) = 526/2480

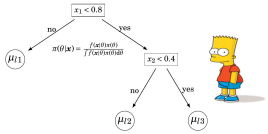


Contingency Tables
Another Example

Random sample of residents was selected and each response was categorized according to revenue preference and age. The questions were related to the legalization of gambling.

Age	Gambling	Liquor Stores	Other	Total
18-20	33	68	12	113
21-30	55	121	50	226
31-44	117	109	132	358
At least 45	158	110	90	358
Total	363	408	284	1055

- What is probability resident is in favor of legalized gambling?
- What is probability person is 31-44 given they are in favor of state-owned liquor stores?
- What is the probability person is 21-30 and in favor of other?
- What is the probability of being 18-20 or 45+?
- Are the events favoring liquor stores and being 31-44

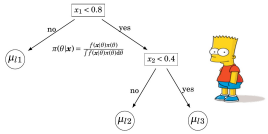


Example

Smoking Status	Upper	Middle	Lower	Total
Current	51	22	43	116
Former	55	121	50	141
Never	117	109	132	99
Total	211	52	93	356

- $\Pr(\text{Current smoker} \mid \text{upper class}) = \frac{51}{211}$
- $\Pr(\text{never smoked \& lower class}) = \frac{22}{356}$
- $\Pr(\text{middle class} \mid \text{former smoker}) = \frac{21}{141}$
- $\Pr(\text{former smoker}) = \frac{141}{356}$
- $\Pr(\text{Upper class \& current smoker}) = \frac{51}{356}$
- $\Pr(\text{Upper class}) \times \Pr(\text{Current smoker}) = \frac{211}{356} \times \frac{116}{356} \approx 0.19$

Bayes Theorem

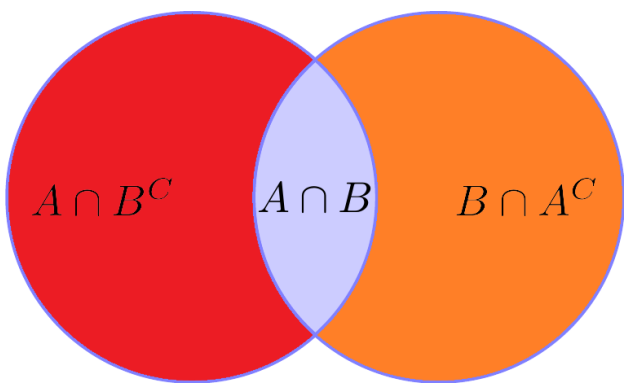


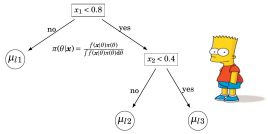
Law of Total Probability
For two Events

For two events A and B, we can write $\Pr(A) = \Pr(A \text{ and } B) + \Pr(A \text{ and } B^c)$ (and vice versa for $\Pr(B)$), which we can rewrite with conditional probabilities:

$$\Pr(A) = \Pr(B) \times \Pr(A \mid B) + \Pr(B^c) \times \Pr(A \mid B^c)$$
$$\Pr(B) = \Pr(A) \times \Pr(B \mid A) + \Pr(A^c) \times \Pr(B \mid A^c)$$

This figure shows why $\Pr(A) = \Pr(A \text{ and } B) + \Pr(A \text{ and } B^c)$ and vice versa $\Pr(B) = \Pr(B \text{ and } A) + \Pr(B \text{ and } A^c)$



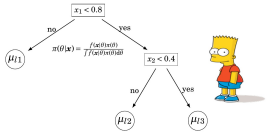


Bayes' Theorem

Suppose that $A_1, A_2, \ldots, A_k$ are disjoint events whose probabilities are not 0 and sum to 1
That is, any outcome has to be exactly in one of these events.
Then if B is any other event whose probability is not 0 or 1, then

$$\Pr(A_i \mid B) = \frac{\Pr\big(B|A_i\big)\Pr(A_i)}{\Pr\big(B|A_1\big)\Pr(A_1)+\Pr\big(B|A_2\big)\Pr(A_2)+\ldots+\Pr\big(B|A_i\big)+\ldots+\Pr\big(B|A_k\big)\Pr(A_k)}$$

Notice, that $1 \leq i \leq k$, meaning that i is just some index between 1 and k. In this class, our two categories are the event and the complement, so $k = 2$.



Bayes' Theorem

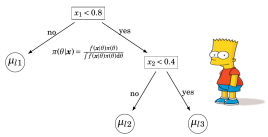
This comes from conditional probability:

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} \tag{1}$$

$$\Pr(B \mid A) = \frac{\Pr(A \text{ and } B)}{\Pr(A)} \implies \Pr(A \text{ and } B) = \Pr(B \mid A)\Pr(A) \tag{2}$$

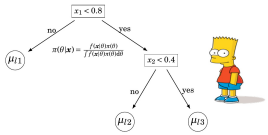
$$\implies \Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \text{ (2) in num, (1) in denom} \tag{3}$$

But, $\Pr(B) = \sum_{j=1}^{k} \Pr(A_j \cap B) = \sum_{j=1}^{k} \Pr(B \mid A_j)\Pr(A_j)$ (from law of total probability) which is how we get the theorem (notice we index over j not i, because we are summing over all the possible events A we can condition on)



Example for the Denominator

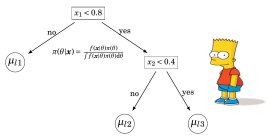
- Say we took a poll of the bio and non-bio students at ASU and asked if they were left-handed. What we get then is $\Pr(L \mid B)$ and $\Pr(L \mid B^c)$ where B^c means not a bio student and L is a variable for left handed. Say, however, that we want the total probability of left-handed people. Then we can use the equation from the previous slide: Assume $\Pr(L \mid B) = 0.20$, $\Pr(L \mid B^c) = 0.10$, $\Pr(B) = 0.05$, and $\Pr(B^c) = 1 - 0.05 = 0.95$.
- We could reasonably expect to know the conditional probabilities by having polls in specific classes, and we can use school data to calculate for $\Pr(B)$, so this is an example where we can combine those two in a sort of "weighted average" to get the total probability we care about!



Example Continued

We use the equation for the marginal probability of L from slide 19

$$\begin{aligned}\Pr(L) &= \Pr(L \mid B)\Pr(B) + \Pr(L \mid B^c)\Pr(B^c) \\ &= 0.20 * 0.05 + 0.10 * 0.95 \\ &= 0.105\end{aligned}$$



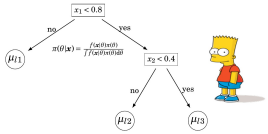
Bayes' Theorem

If we are given $\Pr(B)$, then:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

However, even if we don't know $\Pr(B)$, we're okay. If we let A_i we just have two options, an event and its compliment, i.e. A and A^c,

$$\boxed{\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}} \tag{4}$$



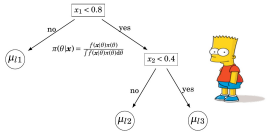
An Example

Suppose:

$$\begin{aligned}\Pr(\text{dangerous fire}) &= 0.01\\ \Pr(\text{smoke}) &= 0.10\\ \Pr(\text{A dangerous fire makes smoke}) &= 0.90\end{aligned}$$

What is the probability of dangerous fire when there is smoke?

$$\begin{aligned}\Pr(\text{Fire} \mid \text{Smoke}) &= \frac{\Pr(\text{Fire})\Pr(\text{Smoke} \mid \text{Fire})}{\Pr(\text{Smoke})}\\ &= \frac{0.01 \cdot 0.90}{0.10}\\ &= 0.09\end{aligned}$$



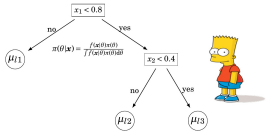
Examples

Suppose that a medical test has a 99% chance of detecting a disease given the person actually has the disease. The test has a 90% chance of correctly telling someone they do not have the disease when they in fact do not have the disease, i.e. 10% of people are falsely told they have the disease. Now, suppose 5% of the population actually has the disease.

- What is the probability that a randomly chosen person will test positive?

$$\begin{aligned}\Pr(P) &= \Pr(P \mid D)\Pr(D) + \Pr(P \mid D^c)\Pr(D^c) \\ &= 0.99 \cdot 0.05 + 0.10 \cdot 0.95 \\ &= 0.1445\end{aligned}$$

i.e. even though only 5% of people have disease, 14.45% of tests come back positive!



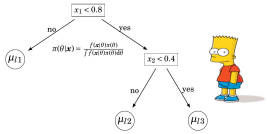
Example Continued

- Suppose that a person does test positive. What is the probability that this person really has the disease?

- We still need to account for false negatives in the numerator, but basically we are dividing the number of people who really have the disease over the total number of positive

$$\Pr(D \mid P) = \frac{\Pr(P \mid D)\Pr(D)}{\Pr(P \mid D)\Pr(D) + \Pr(P \mid D^c)\Pr(D^c)} = \frac{0.99 * 0.05}{14.45} = 0.3$$

- So most of error comes from rarity of disease and false positives not the test missing the disease
- We can apply Bayes rule again with $\Pr(D) = 0.343$ and see the probability we really have the disease if we get ANOTHER positive test (it'll be higher). Try it out!



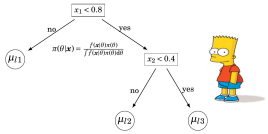
Extra Example

Overall, suppose 1/4 of students get a B on an exam. Now suppose 2/3 of students do not carefully read exam questions, and in that case 1/5 of them get a B.

- Prob a student who read instructions correctly gets a B: Let B be the event of getting a B, and R be the event of reading instructions.

$$\Pr(B \mid R) = \frac{\frac{1}{4} - \frac{1}{5} \cdot \frac{2}{3}}{\frac{1}{3}} = \frac{7}{20} = 0.35$$

- The numerator is the proportion of total students who got a B minus the proportion who got a B while not reading the instructions. This gives us the proportion of students who got a B while reading correctly. However, we divide by the prob a student read correctly to get the probability given that conditional



Pr(B|R) mathematically

- We want $\Pr(B \mid R)$. Bayes theorem applied directly gives us

$$\Pr(B \mid R) = \frac{\Pr(B)\Pr(R \mid B)}{\Pr(R)}$$

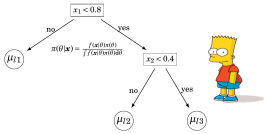
But we don't know $\Pr(R \mid B)$.

- Recall, (see slides 16-18)

$$\Pr(B) = \Pr(B \mid R)\Pr(R) + \Pr(B \mid R^c)\Pr(R^c)$$

so that means, solving for $\Pr(B \mid R)$:

$$\Pr(B \mid R) = \frac{\Pr(B) - \Pr(B \mid R^c)\Pr(R^c)}{\Pr(R)}$$



Continued

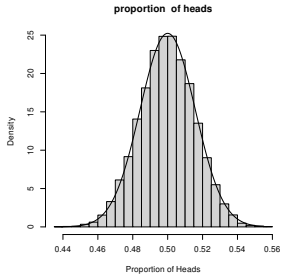
- Prob a student who gets a B has read the instructions correctly:
  $\Pr(B) = 1/4$, $\Pr(B \mid R^c) = (1/5) \cdot (2/3) = 2/15$. Using Bayes rule:

$$\Pr(R \mid B) = \frac{\Pr(B|R)\Pr(R)}{\Pr(B|R)\Pr(R)+\Pr(B|R^c)\Pr(R^c)} = \frac{(7/20)\cdot(1/3)}{(7/20)*(1/3)+(1/5)*(2/3)} = \frac{7}{15}$$

Chapter 11 Notes

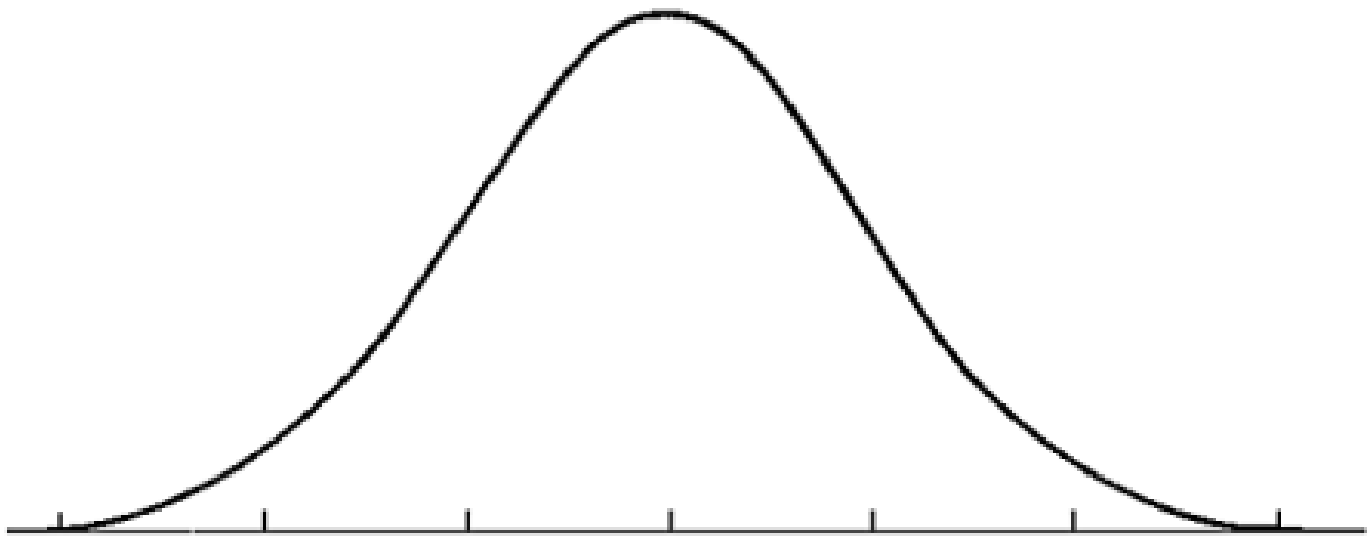
The Normal Distribution
STP-231

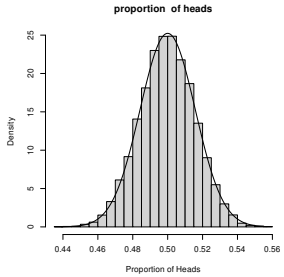
Arizona State University



The normal Curve

- The normal density curve has a symmetric "bell-shaped" curve
- It is symmetric and unimodal





The Normal Distribution

- If a random variable Y follows a normal distribution, then it is distributed as

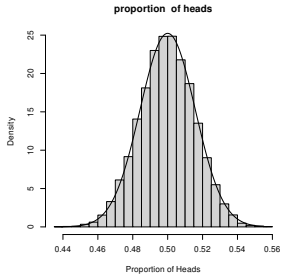
$$Y \sim N(\mu, \sigma)$$

Where N means the distribution is normal. More formally,

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

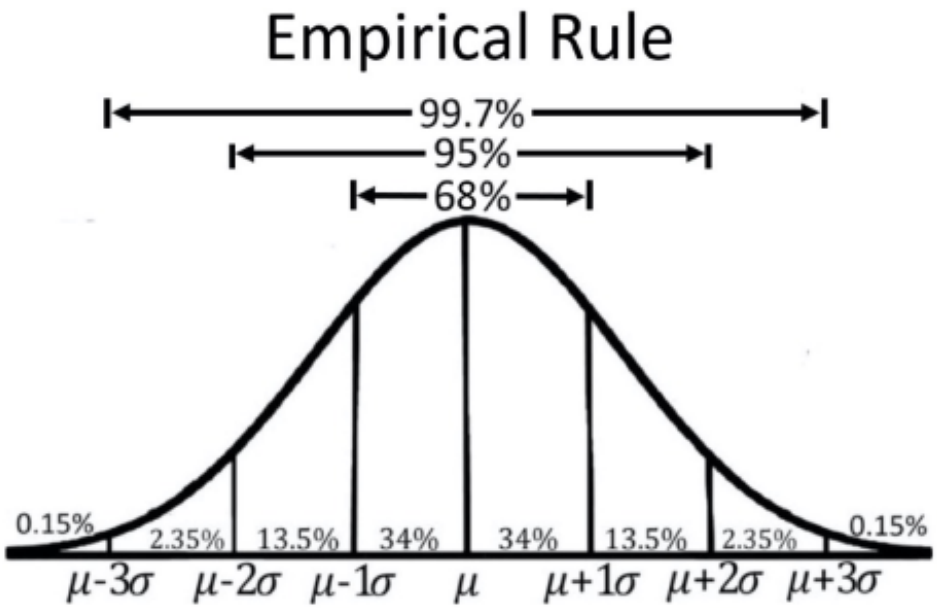
- μ represents the population mean, which can be positive, negative, or zero. This shifts where the peak is left or right

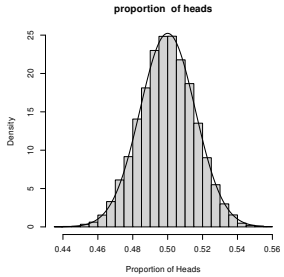
- σ represents the standard deviation, which is always greater than zero. The widens/thins our curve



Basis for the Empirical Rule

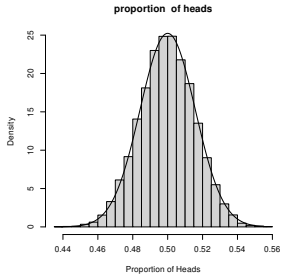
- Recall the empirical rule, that 68% of data is within 1 standard deviation, 95% within 2, 99.7 within 3





Mean and Median

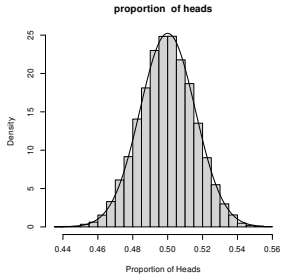
- They are the same! The normal curve is symmetric and unimodal
- The highest peak is at the mean. That is, the mode=median=mean.
- The distribution is symmetric to the mean divides the distribution in half
- Since it is symmetric around the 50%-ile, that is why the mean=median



Standardization and Z-score

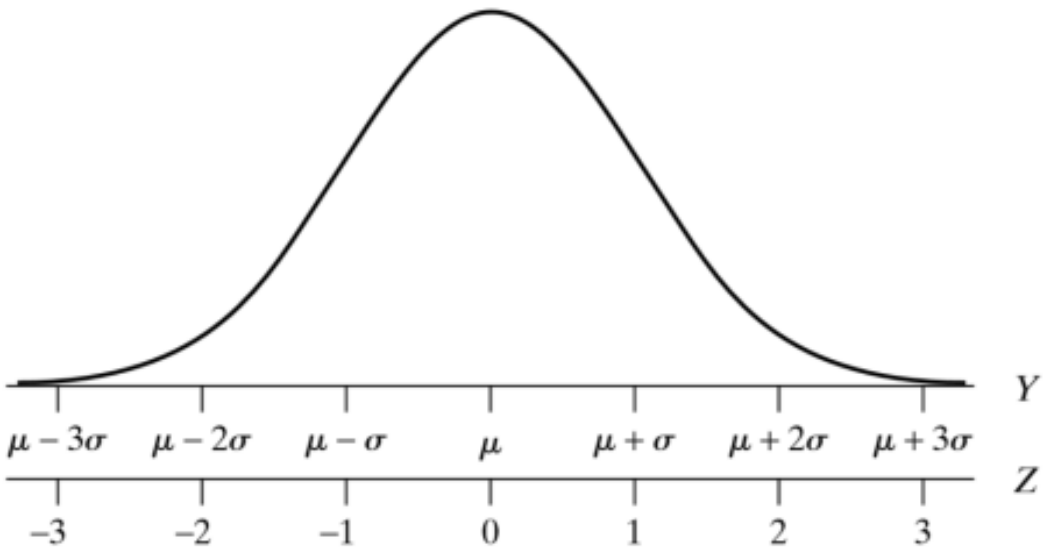
- The Z-score represents values in relation to μ and σ
- How many standard deviations above or below the mean is a particular value
- Unit of measurement does not affect the z-score value
- The transformation is of the form:

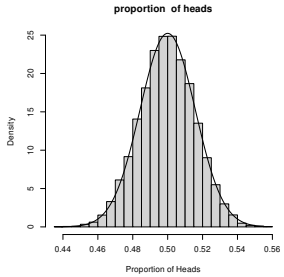
$$z = \frac{y - \mu}{\sigma}$$



Y vs Z

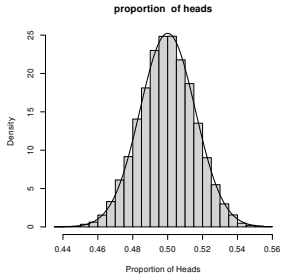
- If we recall from earlier chapter, a linear transformation does not affect the shape of distribution





Calculating Probability from Z-scores

- Draw a standard normal curve
- Label the z-score(s) on the curve
- Shade in the region of interest
- Determining the corresponding area under the standard normal curve using a Z-table

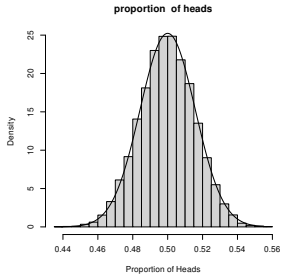


Using the Calculator

Finding the Area to the left of an observation from a Normal

1. Press 2nd key
2. Press VARS
3. Scroll to normalcdf
4. The form for normalcdf is normalcdf(lower bound, upper bound, mean, standard deviation)

By default, it uses $\mu = 0$ and $\sigma = 1$, so we either use our transformed z score or plug in the appropriate μ and σ from given problem. For lower and upper bound, use -1E99 (lower) or 1E99 if we want all the area to left or right. E99 is 2nd comma symbol in the calculator.

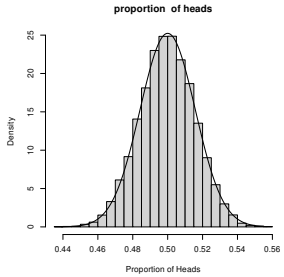


Using the Calculator

Finding a value from a normal distribution given area to its left

1. Press 2nd key
2. Press VARS
3. Scroll to invNorm
4. The form for normalcdf is invNorm(area to the left, mean, standard deviation)

By default, it uses $\mu = 0$ and $\sigma = 1$, so we either use our transformed z score or plug in the appropriate μ and σ from given problem.



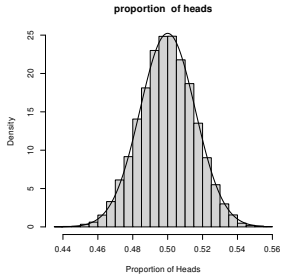
Example

Find the area to the left of specified z-scores

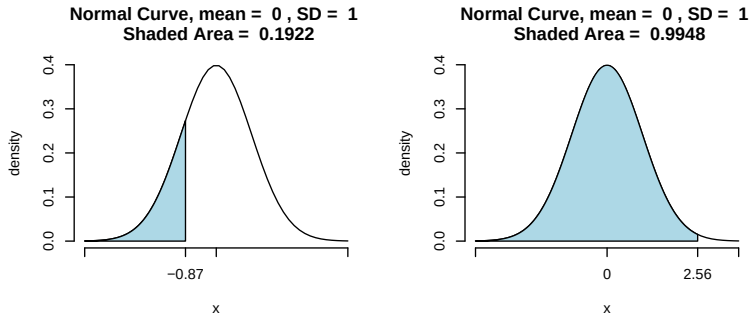
- -0.87
- 2.56
- 5.12

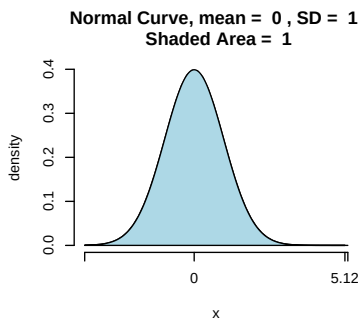
Find the area to the right of specified z-scores

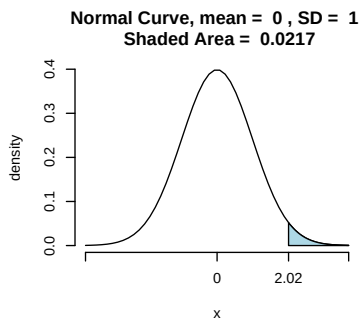
- 2.02
- -0.56
- -4

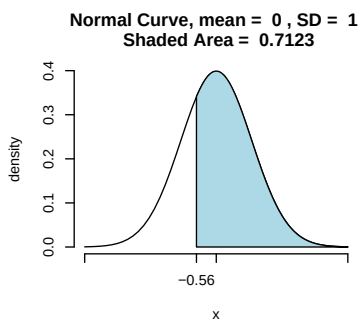


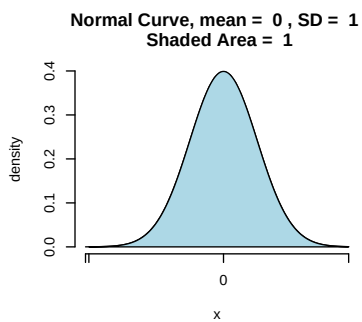
Example Answer

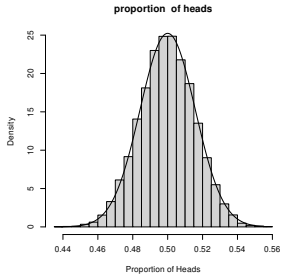






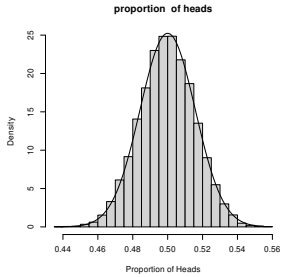






New Example: Area between curves

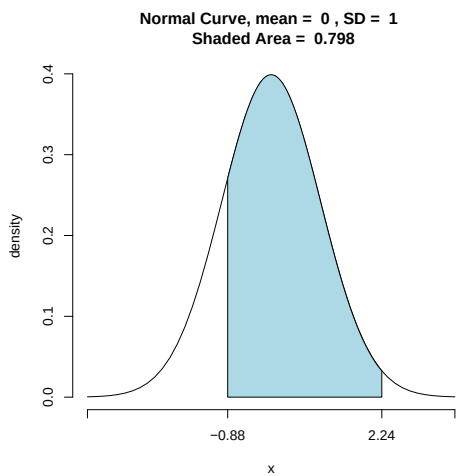
Determine the area under the standard normal curve that lies between -0.88 and 2.24

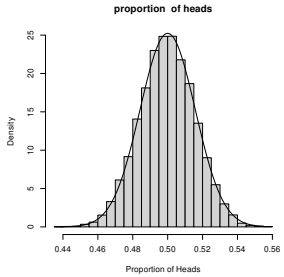


New Example: Area between curves

Determine the area under the standard normal curve that lies between -0.88 and 2.24

$$\begin{aligned}\Pr(-0.88 \leq Z \leq 2.24) &= \Pr(Z \leq 2.24) - \Pr(Z \leq -0.88) \\ &= \Pr(Z < 2.24) - \Pr(Z < -0.88) \\ &\approx 0.9875 - 0.1894 = 0.7984\end{aligned}$$



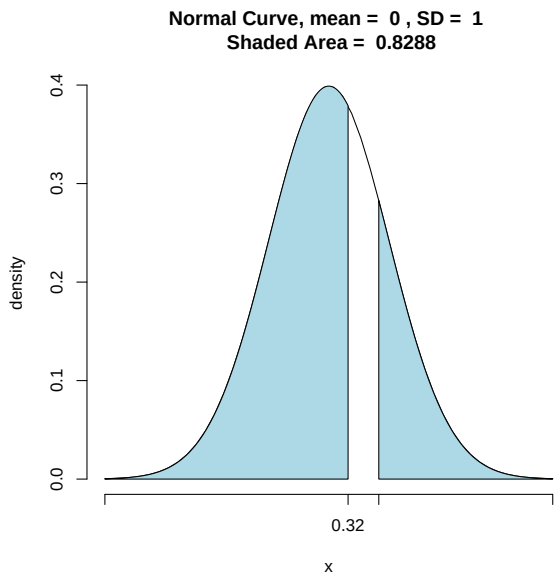


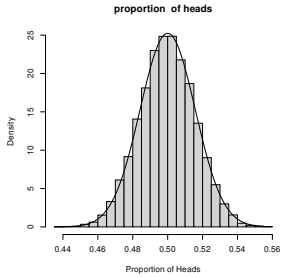
Another Example

Find the area below 0.32 and above 0.83 under the curve

$$\Pr(Z < 0.32) + \Pr(Z > 0.83) \approx 0.6255 + (1 - \Pr(Z < 0.83) \approx 0.8288$$

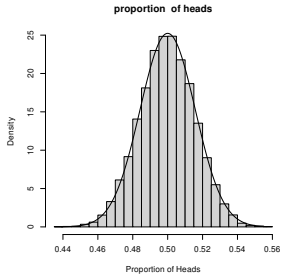
$$\text{OR } 1 - \Pr(0.32 < Z < 0.83)$$





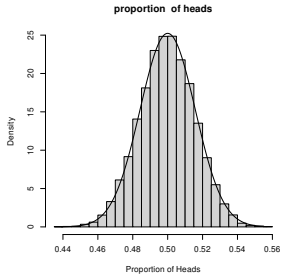
Probability for a Normally Distributed Variable

1. Sketch the normal curve associated with the variable
2. Shade the region of interest and mark its delimiting y-value(s)
3. Find the z-score(s) for the delimiting y-value(s) found in step 2
4. Use the table to find the area under the standard normal curve delimited by the z-score(s) found in step 3.



Another Example

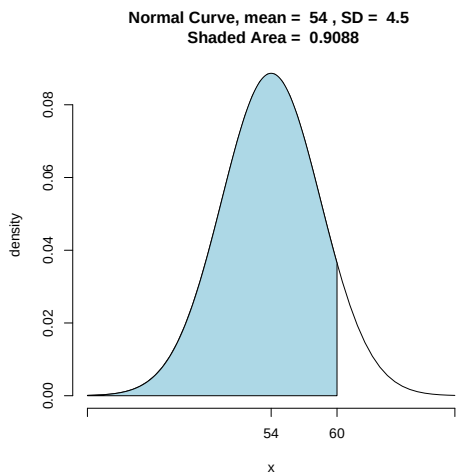
In a certain population of the herring Pomolobus aestivalis, the lengths of the individual fish follow a normal distribution. The mean length of the fish if 54.0 mm, and the standard deviation is 4.5mm. What percentage of fish are less than 60 mm long?

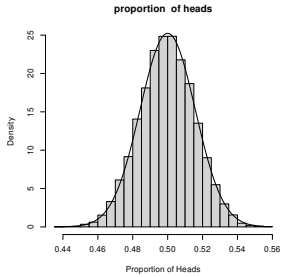


Another Example

In a certain population of the herring Pomolobus aestivalis, the lengths of the individual fish follow a normal distribution. The mean length of the fish if 54.0 mm, and the standard deviation is 4.5mm. What percentage of fish are less than 60 mm long?

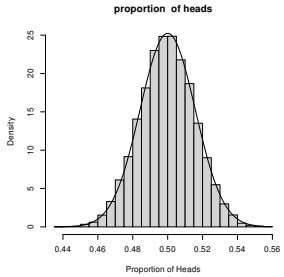
$$z = \frac{y - \mu}{\sigma} = \frac{60 - 54}{4.5} = 1.33$$





More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

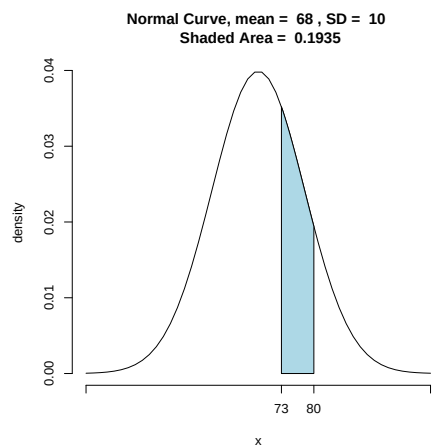


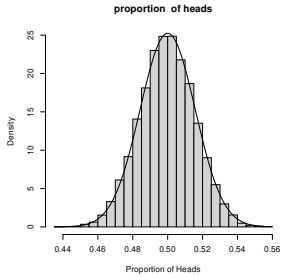
More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

1. Lie between 73 and 80

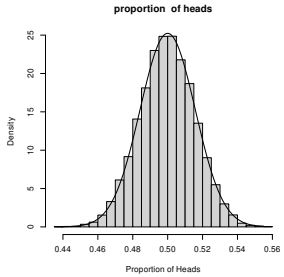
$$\Pr(73 < Y < 80) - \Pr\left(\frac{73-68}{10} < Z < \frac{80-68}{10}\right)$$
$$= \Pr(Z < 1.2) - \Pr(Z < 0.5) \approx 0.1935$$





More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

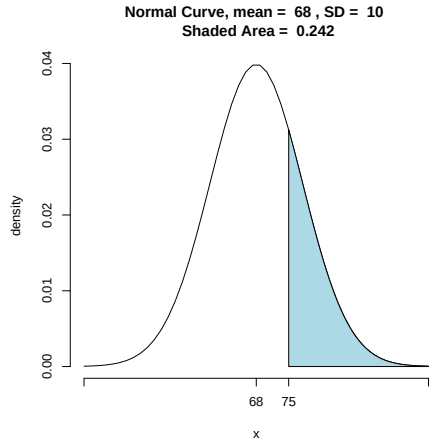


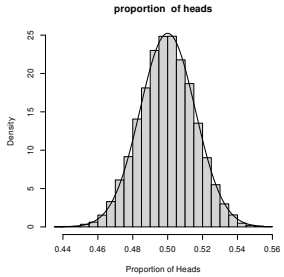
More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

2. Are at least 75

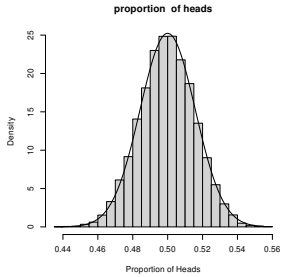
$$\Pr(Y > 75) = 1 - \Pr(Y < 75) = 1 - \Pr(Z < 0.7)$$
$$= \approx 0.242$$





More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

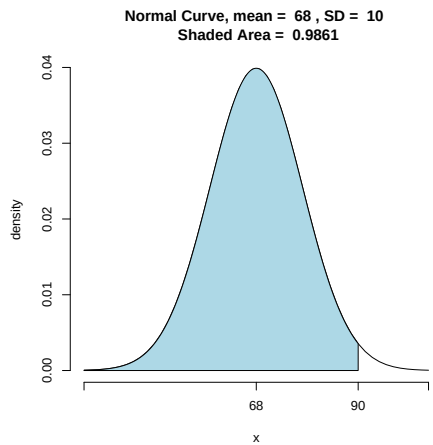


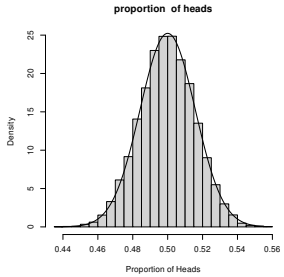
More Examples

A variable is normally distributed with mean 68 and standard deviation 10. Find the percentage of all possible values of the variable that

3. Are at most 90

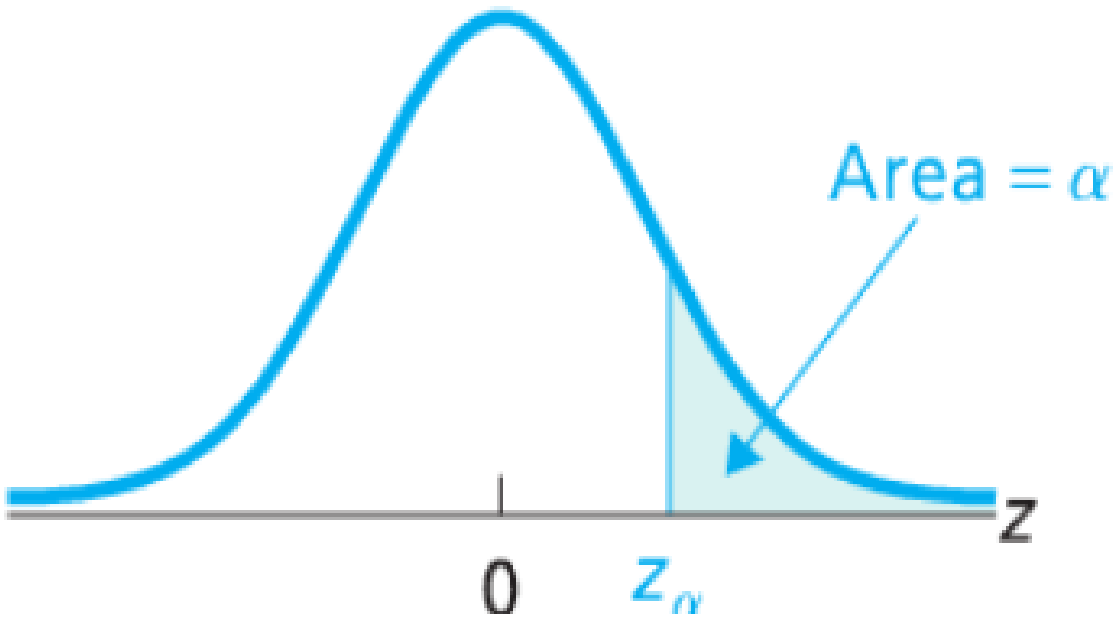
$$\Pr(Y < 90) = \Pr\left(Z < \frac{90 - 68}{10}\right)$$
$$= \Pr(Z < 2.2) \approx 0.9861$$

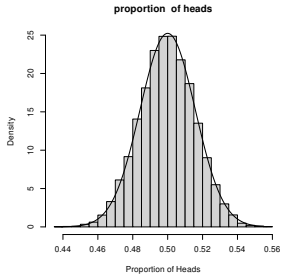




Z_α Notation

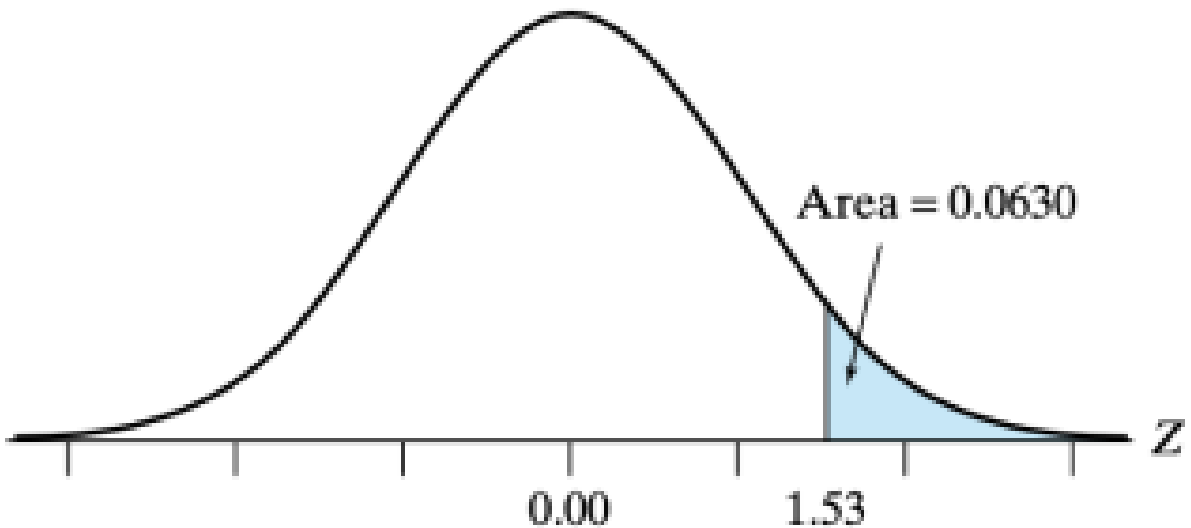
- Z_α is used to denote the z-score that has an area of α to the right under the standard normal curve
- α is a probability. Z_α is a z-score, but not a probability...why is that?

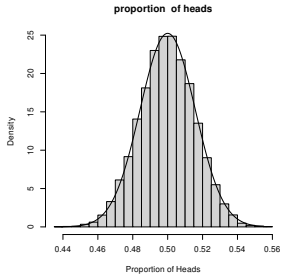




Example

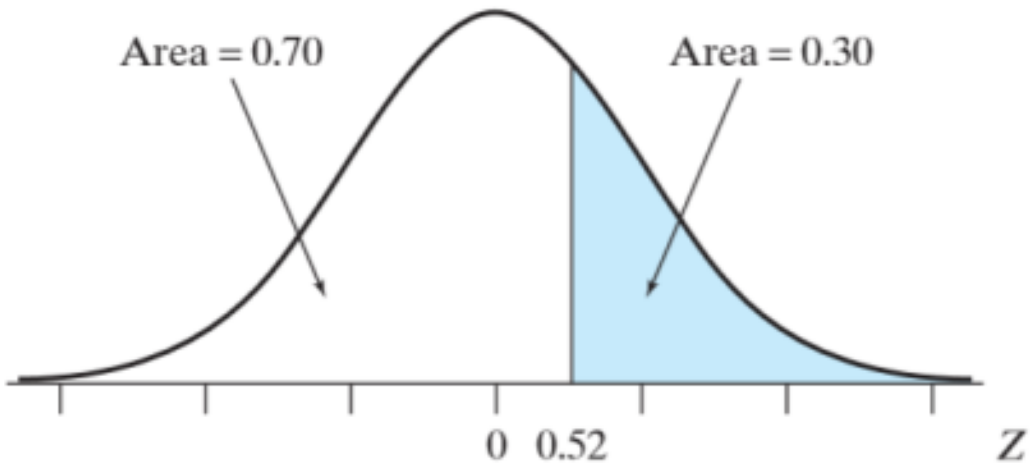
- $Z_{0.0630} = 1.53$

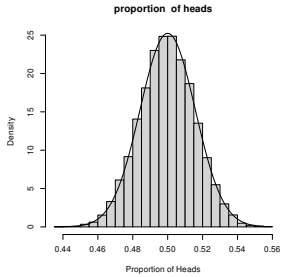




Example

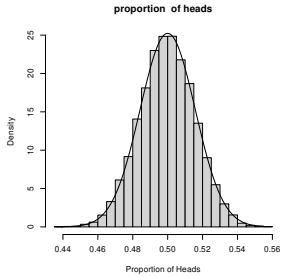
- $Z_{0.30} = 0.52$





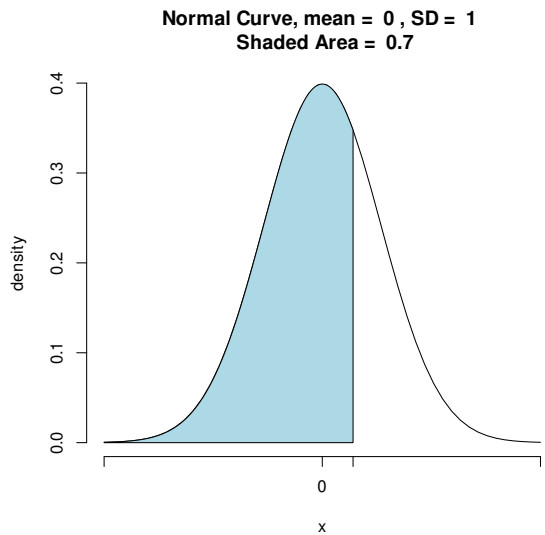
Percentiles

- Percentiles divide the distribution into 100 equal parts.
- Indicates the value below which a given percentage of observations fall
- We can compare to $Z_{\alpha'}$, but here we consider area of α to the left

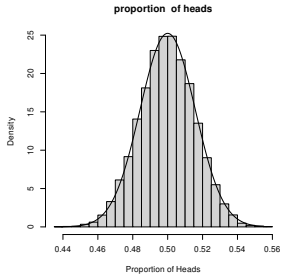


Percentiles Example

Suppose we want to find the 70th percentile of a standard normal distribution. We want to find the z-value that divides the bottom 70% from the top 30%. What is the value?

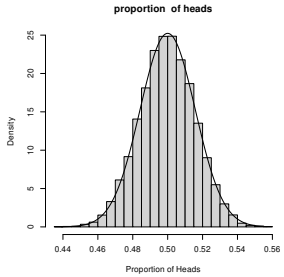


This is at $Z_{0.30} = 0.524$.



%-iles for any Normally Distributed Variable

1. Sketch the normal curve associated with the variable
2. Shade the region of interest
3. Use the table to find the z-score(s) for the delimiting region found in step 2
4. Find the y-value(s) having the z-score(s) found in step 3



How to quickly generate in R

Y is normally distributed with mean 68 and standard deviation 10. Find the value of the 99th percentile:

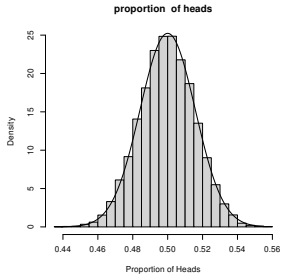
- We note the 99th percentile is also

$$Z_{0.01} = 2.325$$

on the Z-scale. So we transform Y to Z:

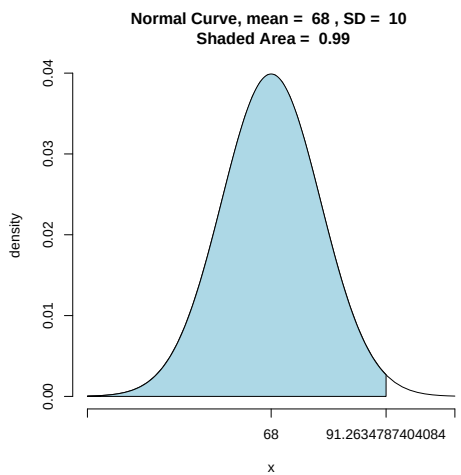
$$z = 2.325 = \frac{y - 68}{10} \implies Y_{.01} = 91.25$$

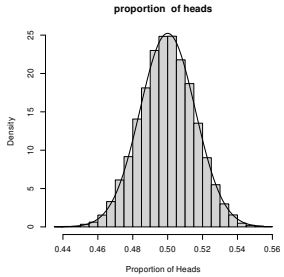
Because $y = \sigma z + \mu$ when we transform back!



Graphically

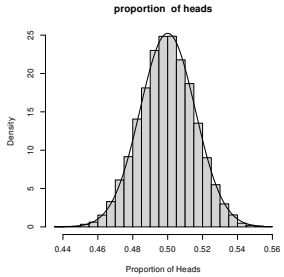
```
library(tigerstats)
#use qnorm() for percentile, pnorm() for area
#code to get 99th percentile
pnormGC(qnorm(0.99), region="below", mean=0,
sd=1,graph=TRUE)

```





Where do quartiles fit in?

Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:



Where do quartiles fit in?

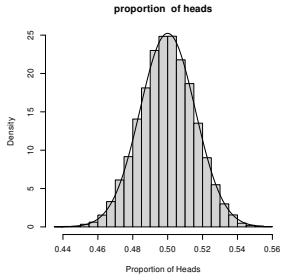
Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:

$$Q_1 = Z_{0.75} = -0.675 \xrightarrow{\text{Transform Y}} Q_1[Y] = -0.675 \cdot (10) + 68 = 61.26$$

$$Q_2 = Z_{0.50} = 0 \xrightarrow{\text{Transform Y}} Q_2[Y] = 68$$

$$Q_3 = Z_{0.25} = 0.674 \xrightarrow{\text{Transform Y}} Q_3[Y] = 0.675 \cdot (10) + 68 = 74.74$$

Find the value that 85% of all possible values of the variable exceed



Where do quartiles fit in?

Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the quartiles:

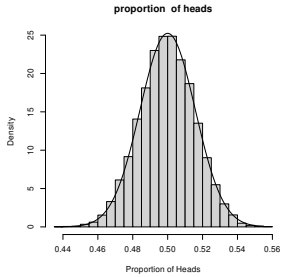
$$Q_1 = Z_{0.75} = -0.675 \xrightarrow{\text{Transform Y}} Q_1[Y] = -0.675 \cdot (10) + 68 = 61.26$$

$$Q_2 = Z_{0.50} = 0 \xrightarrow{\text{Transform Y}} Q_2[Y] = 68$$

$$Q_3 = Z_{0.25} = 0.674 \xrightarrow{\text{Transform Y}} Q_3[Y] = 0.675 \cdot (10) + 68 = 74.74$$

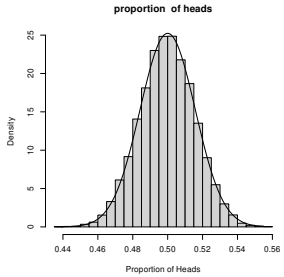
Find the value that 85% of all possible values of the variable exceed If 85% exceed then 15% don't. So we want

$Z_{0.85} = -1.35$. Which means $Y_{0.15} = -1.036 \cdot 10 + 68 = 57.6$



Example continued

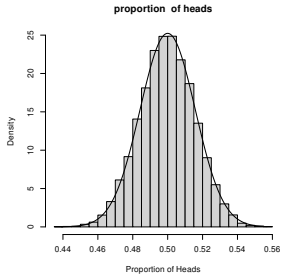
Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the two values that divide the area under the corresponding normal curve into a middle area of 0.90 and two outside areas of 0.05.



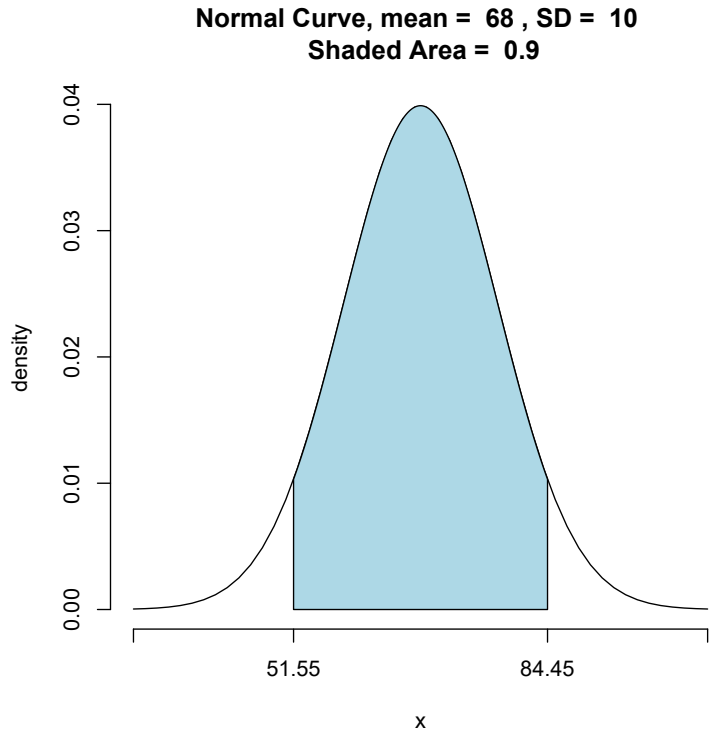
Example continued

Again assume a variable is normally distributed with mean 68 and standard deviation 10. Find the two values that divide the area under the corresponding normal curve into a middle area of 0.90 and two outside areas of 0.05.

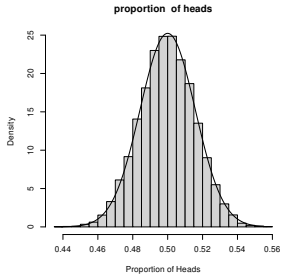
Note because of symmetry $\Pr(X \geq a) = \Pr(X \leq -a)$ for some a, so its also true $Z_\alpha = -Z_{1-\alpha}$. We see that in this example. We want the 5th and 95th percentile, which are respectively -1.645 and 1.645.



Pictoral Representation



Then, we find $y_{\text{lower}} = \sigma z_{\text{lower}} + \mu = 10 * -1.645 + 68 = 51.55$ and $y_{\text{upper}} = \sigma z_{\text{upper}} + \mu = 10 * 1.645 + 68 = 84.45$

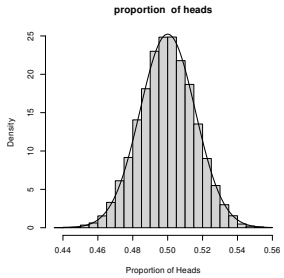


Word Problem

The salinity, or salt content, in the ocean is expressed in parts per thousand (ppt). The number varies due to depth, rainfall, evaporations, river runoff, and ice formation. During January and February, the mean salinity in a region of the northeast continental shelf was 34.08 ppt. The distribution of salinity is normal w/ standard deviation 0.52 ppt. Suppose a random sample of ocean water from this region is obtained.

- What is the probability the salinity is more than 35 ppt?
- A certain species of fish can only survive if the salinity is between 33pt and 35 ppt. What is the probability this species can survive in a randomly selected area?
- Find the salinity that corresponds to the 65th percentile

Assessing Normality



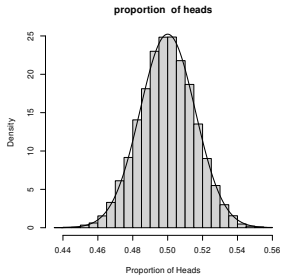
Verify Empirical Rule

Any normally distributed random variable has the following properties:

- 68% of all possible observations lie within one standard deviation to either side of the mean, that is, between $\mu - \sigma$ and $\mu + \sigma$

- 95% of all possible observations lie within two standard deviation to either side of the mean, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$

  99.7% of all possible observations lie within 3 standard deviations to either side of the mean, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$

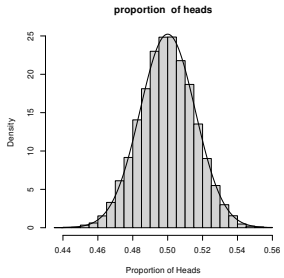


Example: Number of Siblings

Number of Siblings	Frequency $\Pr(Y = y_i)$	
0	4	$\frac{4}{29}$
1	9	$\frac{9}{29}$
2	10	$\frac{10}{29}$
3	3	$\frac{3}{29}$
4	1	$\frac{1}{29}$
5	0	$\frac{0}{29}$
6	0	$\frac{0}{29}$
7	2	$\frac{2}{29}$

How well does the empirical rule estimate this data? $\mu_Y = 1.931$ and $\sigma_Y = 1.680$

- 75% of the data lies within 1 standard deviation from the mean, i.e. $(0.251, 3.611)$
- 93% of the data lies within 2 standard deviations from the mean, i.e. $(-1.43, 5.29)$
- 93% of the data lies within 3 standard deviations from the mean, i.e. $(-3.11, 6.97)$



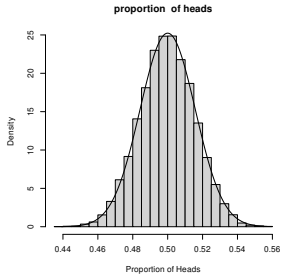
Assessing Samples (Large vs Small)

If we have a large data set:

- Plot a histogram
- Analyze the shape
- Is it normal or approximately normal looking?

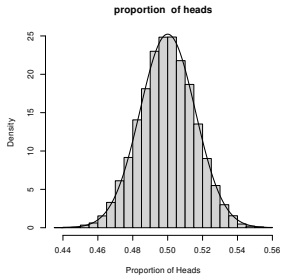
For a small data set

- Make a normal probability plot (aka the quantile plot), because the histogram may not be entirely useful if data-set too small



Normal Quantile Plots (QQ Plots)

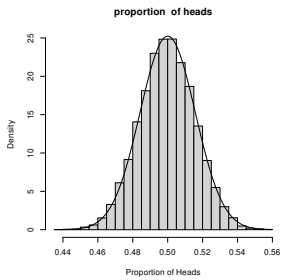
- Statistical graphs that assess normality
- Compare observations with the values we would expect for a standard normal distribution
- Focus on proportions for percentiles rather than the empirical rule
- Normal score: the data points we expect to obtain if our data was normal



QQ Plot Example

- Height of 11 women has mean 65.5 inches and standard deviation 2.9 inches. The smallest observation is 61 inches tall, which means our sample predicts 1/11th of women are 61 inches or shorter in population, i.e. 9.09th percentile

[1] We will use adjusted percentiles in reality



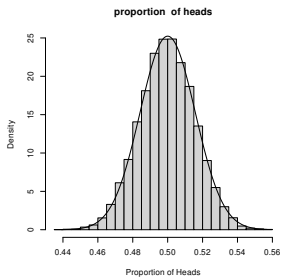
QQ Plot Example

- Height of 11 women has mean 65.5 inches and standard deviation 2.9 inches. The smallest observation is 61 inches tall, which means our sample predicts 1/11th of women are 61 inches or shorter in population, i.e. 9.09th percentile
- If data were truly from a normal distribution (our big assumption), then[1]

$$\mu + Z_{1-0.0909}\sigma = 65.5 - 1.34 \cdot 2.9 = 61.6$$

- Repeat this calculation for rest of the data. Then plot the actual quantiles vs the expected
- Normal score: the data points we expect to obtain if our data was normal

[1] We will use adjusted percentiles in reality

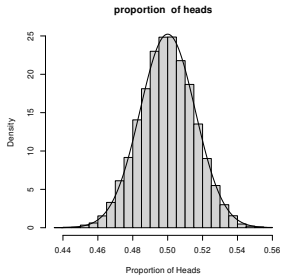


How to Construct a Quantile Plot

1. Rank the observed values from smallest to largest
2. Percentiles=$100 \times \frac{i}{n}$, where i is the ranked observation index and n is the sample size

   We solve for the adjusted percentiles $\left(\frac{i-0.5}{n}\right) \times 100$

   Where we adjust because the max of the sample should not equate to the 100th percentile
3. Find the corresponding expected normal scores (or z-scores) for the adjusted percentiles
4. Plot the observed values on the y-axis and the normal scores (or z-score) on the x-axis
5. If the plot is roughly linear, you can assume that the plot is approximately normally distributed

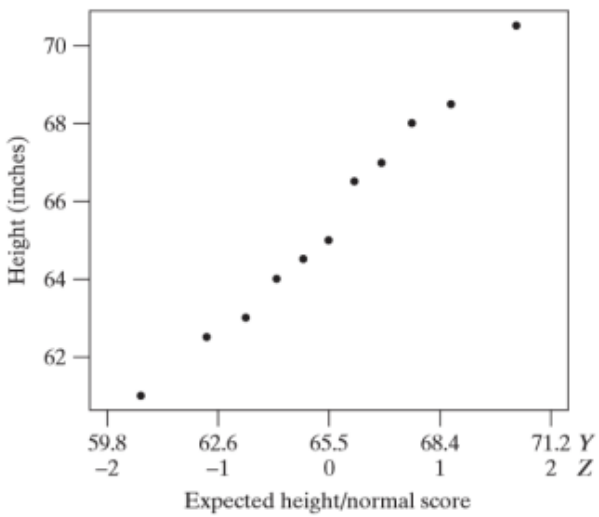


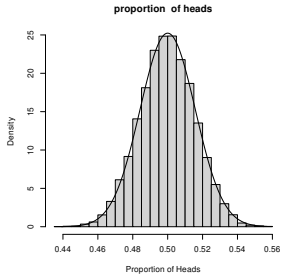
Example

We observe the heights of the 11 women:

$$61, 62.5, 63, 64, 64.5, 65, 66.5, 67, 68, 68.5, 70.5$$

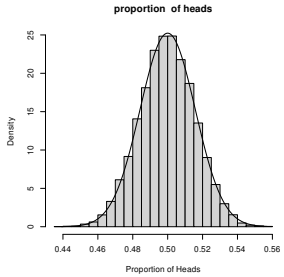
i	1	2	3	4	5	6	7	8	9	10	11
Observed Height	61.0	62.5	63.0	64.0	64.5	65.0	66.5	67.0	68.0	68.5	70.5
Adjusted percentile	4.55	13.64	22.73	31.82	40.91	50.00	59.09	68.18	77.27	86.38	95.45
z	-1.69	-1.10	-0.75	-0.47	-0.23	0.00	0.23	0.47	0.75	1.10	1.69
Theoretical height	60.6	62.3	63.4	64.1	64.8	65.5	66.2	66.9	67.6	68.7	70.4



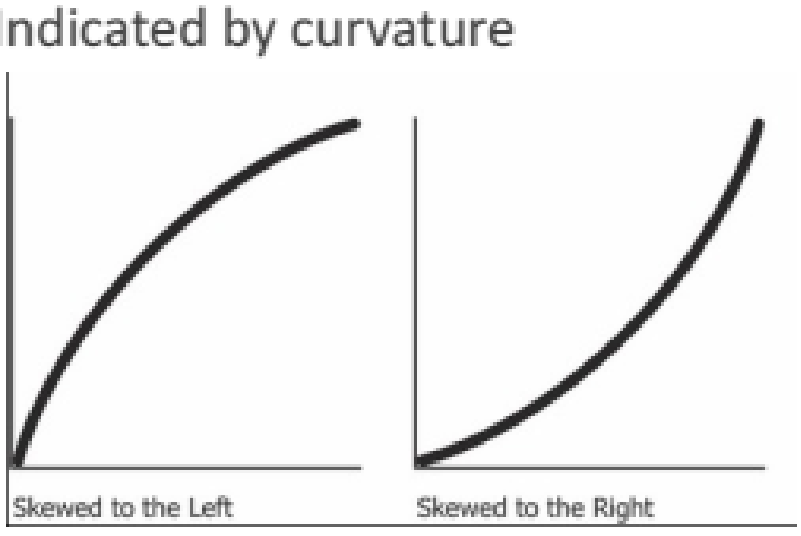


Interpret QQ Plots

- The plotted points fall along an imaginary straight line through (0,mean) when comparing z-scores to normal scores, (0,0) when comparing z-scores to z-scores, or (mean, mean) when comparing normal scores to normal scores
- If the plot is roughly linear, we conclude our data is roughly normally distributed
- Interpret loosely for small samples; adhere strictly for large samples
- Check for curvature at ends (i.e. different tails) and for outliers

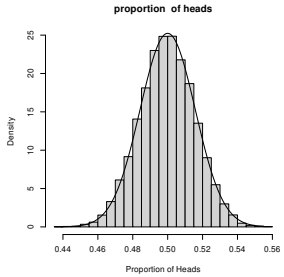


Skewness



The transformations $\sqrt{y}$ or $\ln(y)$ have mild skewness, whereas $\frac{1}{\sqrt{y}}$ and $\frac{1}{y}$ have extreme skew

Normal Approximation of Binomial

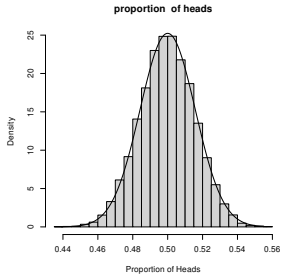


Normal Approximation

- As the number of observations n gets larger, the binomial distribution gets close to normal distribution
- Consider a binomial random variable X. Recall this has n observations, p as a success probability, an expected value and standard deviation of

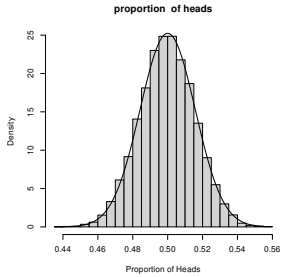
$$\mu_X = np \text{ and } \sigma_X = \sqrt{np(1-p)}$$

- If n is large, then X is approximately $N(np, \sqrt{np(1-p)})$. That is, with enough trials, our outcome of interest ends up having a normal curve describing its potential values
- We consider a sample large enough if $np \geq 10$ and $n(1-p) \geq 10$.



Example

Let's say 1/10 people have been to Antarctica (probably way too high, but let's go with it). What is the probability you know 1 (or more) person who has been to Antarctica, assuming you know 200 people.



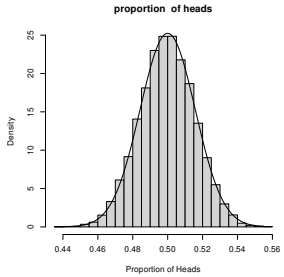
Example

Let's say 1/10 people have been to Antarctica (probably way too high, but let's go with it). What is the probability you know 1 (or more) person who has been to Antarctica, assuming you know 200 people.

- Binomial calculation: Let X be the random variable denoting how many people you know in Antarctica

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \binom{200}{0} 0.1^0 (1-p)^{200} \approx 1$$

So not super helpful...



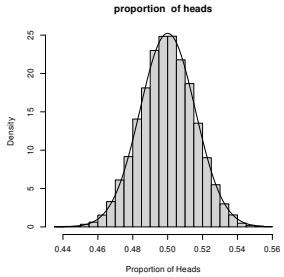
Example Continued

- But the approximation (check the conditions are satisfied) is helpful, because we can now use our tables to calculate probabilities. We can approximate the number of people we know who have been to Antarctica as

$$X \sim N(.1 * 200, \sqrt{20(1 - .1)})$$

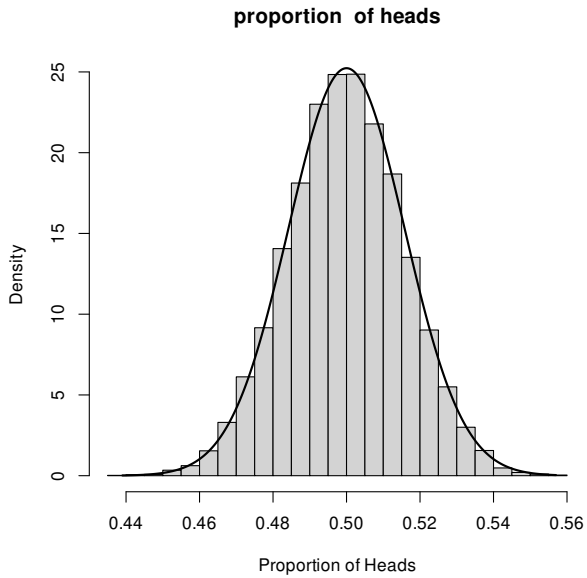
$$\begin{aligned} \Pr(Y < 0) &= \Pr\left(Z < \frac{0 - 20}{\sqrt{18}}\right) \\ &= \Pr(Z < -4.71) \approx 1.21 \times 10^{-6} \end{aligned}$$

- So the probability we know at least one person is approximately $1 - 1.21 \times 10^{-6} \approx 1$.

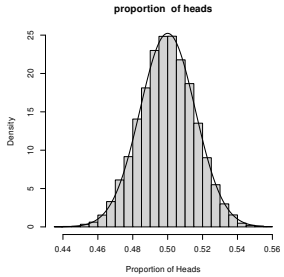


Coin Flip Example

Say we flip a fair coin 1000 times (some large number). We tabulate the number of heads we get, and repeat this experiment 10000 times. (These numbers need not be the same)



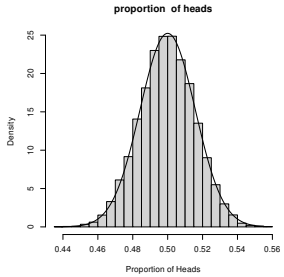
Note, when plotting, we converted the scale from total to relative frequency for the sake of visual clarity



Code to Simulate this Example

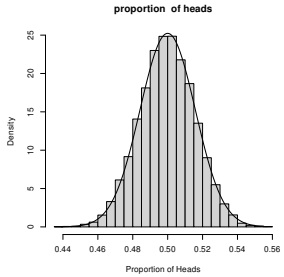
This code runs in base R with no necessary packages. Takes a few seconds to run however.

```
n=1000
p=0.5
sig=sqrt(n*p*(1-p))
aa=sapply(1:10000, function(i) mean(rbinom(1000, 1,
  .5)))
#b/c mean is same as proportion of 1's
h <- hist(aa, 40,
col = "lightgray", xlab = "Proportion of Heads", main
  = "proportion  of heads", freq=F)
xfit <- seq(min(aa), max(aa), length = n)
yfit <- dnorm(xfit, mean = n*p/n, sd = sig/n)
lines(xfit, yfit, col = "black", lwd = 2)

```



New Example

500 guests will attend a dinner event, and it is estimated that 10% of those attendees require a vegan option. What is the probability that the caterers need to have at least 55 plates designated for the vegan option. Use the estimation method: Note, the actual binomial probability is 0.2477



New Example

500 guests will attend a dinner event, and it is estimated that 10% of those attendees require a vegan option. What is the probability that the caterers need to have at least 55 plates designated for the vegan option. Use the estimation method: Note, the actual binomial probability is 0.2477

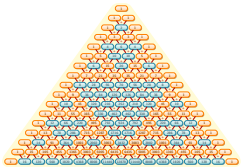
- We can approximate as $N(500 \cdot 0.1, 500 \cdot 0.1(1-0.1))$. Then,

$$\begin{aligned}\Pr(Y \geq 55) &= 1 - \Pr\left(Z < \frac{55-50}{\sqrt{45}}\right) \\ &= 1 - \Pr(Z < 0.745) = 0.228\end{aligned}$$

Chapter 12 Notes

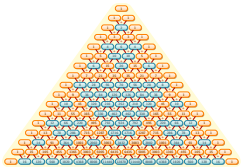
The Binomial Distribution
STP-231

Arizona State University



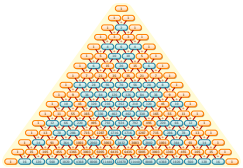
Discrete Probability Distributions

- The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values
- In the discrete case, all possible values (and subsequently their probabilities) can be listed
- Let the random variable X be the number of heads when tossing a coin twice (2 Bernoulli trials, hence a Binomial)
- Let the random variable Y be the number of heads when tossing a coin n times (n Bernoulli, also Binomial)



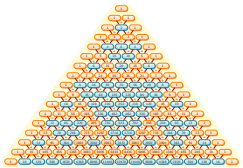
Discrete Probability Distributions
Continued

- For probabilities on intervals of possible values we take note of which outcomes are defined in the interval
- Note, that unlike the continuous case $\Pr(a < X < b)$ may not be the same $\Pr(a \leq X \leq b)$ The same is true for other combinations of inclusion
- For example, let Y be the count of heads that come up after flipping a coin 10 times
- Then $\Pr(Y > 8) \neq \Pr(Y \geq 8)$



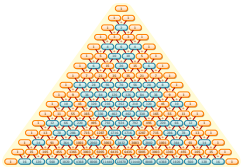
Bernoulli Random Variables

- The simplest type of random variable
- Can take on two values: "success" or "failure"
- We define p to be the probability of success; meaning the probability of failure is $1 - p$. A single coin toss is an example
- success and failure defined arbitrarily. A tail could be success/failure depending on how you define it



Binomial Distribution Definition

- Assume a series of n independent Bernoulli experiments is conducted
- Each independent repetition of the experiment is a trial
- Each experiment results in either "success" or "failure"
- n is the number of trials, k the number of successes, and $n - k$ is the number of failures
- The probability of success is the same for each experiment, regardless of the outcomes of the other trials (i.e. independence)



The Binomial Coefficient

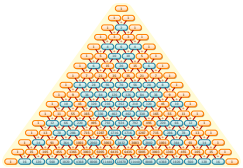
- Say we have n total items and we want to choose k. Let's say we have 5 total letters to choose from (A, B, C, D, E) and choose 3. For our first choice, we have 5 options, then for the second we have 4, then 3. The number of ways to choose these k items in conjunction would the product of the previous arrangement, i.e.

$$5 \cdot 4 \cdot 3 = \frac{5!}{2!} \text{ This is True if order matters}$$

- If order does not matter, this is not true because we double count many arrangements. For example

$$\text{A, B, C} = \text{C, B, A} = \text{B, A, C}$$

because we choose the same letters each time. Therefore, we have to divide by all possible "repeat arrangements", which is $3 \cdot 2 \cdot 1 = 3! = 6$ in this example.



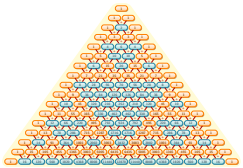
Binomial Coefficient

- Therefore, we have

$$\binom{5}{3} = \frac{5!}{(5-3)!(3)!} = 10$$

- In general:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

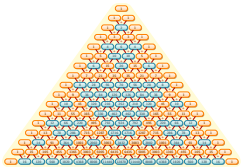


The Binomial Distribution

- The probability distribution for the number of successes in a sequence of Bernoulli trials
- For a binomial random variable Y, the probability that the n trials results in k successes and $n-k$ failures is given by:

$$\Pr(Y = k) = \Pr(k \text{ successes}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \quad (1)$$

- The distribution reflects the probabilities of the k successes and the $n-k$ failures (the p^k and the $(1-p)^{n-k}$ are from the independence assumption)
- Multiplied by the unique number of ways that those n trials can be arranged



Example

Describe the distribution of 2 coin tosses

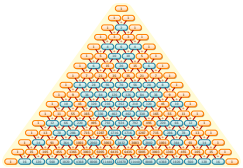
- Let's say tails is a success, and heads is a failure. Instead of counting out the outcomes, lets let Y be the number of tails

$$\Pr(Y = 0) = \binom{2}{0} \cdot p^0 \cdot (1-p)^{2-0} = (1-p)^2$$

$$\Pr(Y = 1) = \binom{2}{1} \cdot p^1 \cdot (1-p)^{2-1} = 2 \cdot p \cdot (1-p)$$

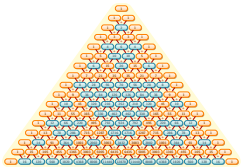
$$\Pr(Y = 2) = \binom{2}{2} \cdot p^2 \cdot (1-p)^{2-2} = p^2$$

- If $p = 0.5$, i.e. a fair coin, what are the probabilities?



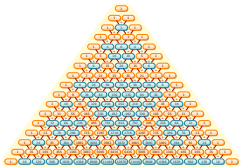
Conditions to satisfy for a Binomial RV

- Each trial only has two possible outcomes: success or failure
- Trials are independent of each other
- The number of trials is fixed at n
- The probability of success, denoted by p, is the same for all trials



Sampling Distribution of a Count

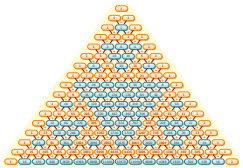
- Choose and SRS (simple random sample) of size n from a population with proportion p successes
- Recall: the probability for randomly choosing an individual with a certain characteristic is equivalent to the population proportion (relative frequency) of individuals with that characteristic in the population
- When the population is much larger than the sample, the count X of successes in the sample has approximately the binomial distribution with parameters n and p



Example

Assume we have one mole of air in a room (6.02×10^{23} air particles). Suppose we split the room in half. Then they are 6.02×10^{23} ways to have all but 1 of the 6.02×10^{23} air particles on one side of the room. This would almost surely suffocate any living thing on the 1-particle side of the room. And there are so many ways for this to happen!
So why don't we spontaneously choke more often? Because there are WAY more ways to have equilibrium. This is the idea of entropy in physics[1]

[1] Independence assumption is iffy



Example

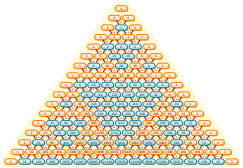
Ladies Home Journal magazine in 1993 that 66% of all dog owners greet their dog before greeting their spouse when they return home at the end of the workday. Suppose that 12 dog owners are selected at random.

(a) What is the probability exactly 4 dog owners greet their dog before greeting their spouse?

(b) What is the probability between 3 and 5 (including 3 and 5) dog owners greet their dog before their spouse?

(c) What is the probability at least 1 dog owner greets their dog before greeting their spouse?

(d) What is the probability that 50% of these dog owners greet their dog before greeting their spouse?



Solutions

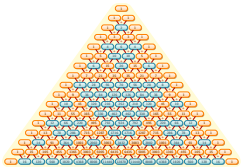
Let X be the R.V. describing how many dog owners greet their dogs, with a success being greeting their dog first

(a)

$$\begin{aligned}\Pr(X=4) &= \frac{n!}{k!(n-k!)}p^k(1-p)^{n-k} \\ &= \frac{12!}{4!(12-4)!}(0.66)^4(1-0.66)^{12-4} \\ &= 0.0168\end{aligned}$$

(b)

$$\begin{aligned}\Pr(3 \leq X \leq 5) &= \Pr(X=3)+\Pr(X=4)+\Pr(X=5) \\ &= 0.00384+0.0168+0.0521 \\ &= 0.0727\end{aligned}$$



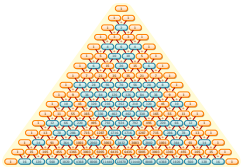
Solutions
Continued

(c)

$$\begin{aligned}\Pr(X \geq 1) &= 1 - \Pr(X = 0) \\ &= 1 - \frac{12!}{0!(12-0)!}(0.66)^0(1-0.66)^{12-0} \\ &= 1 - 0.000002386 \\ &= 0.999997614\end{aligned}$$

(d) 50% of dog owners means number of successes is 6

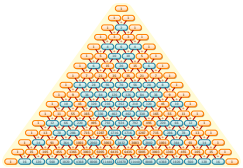
$$\begin{aligned}\Pr(X = 6) &= \frac{12!}{6!(12-6)!}(0.66)^6(1-0.66)^{12-6} \\ &= 0.11798\end{aligned}$$



Example

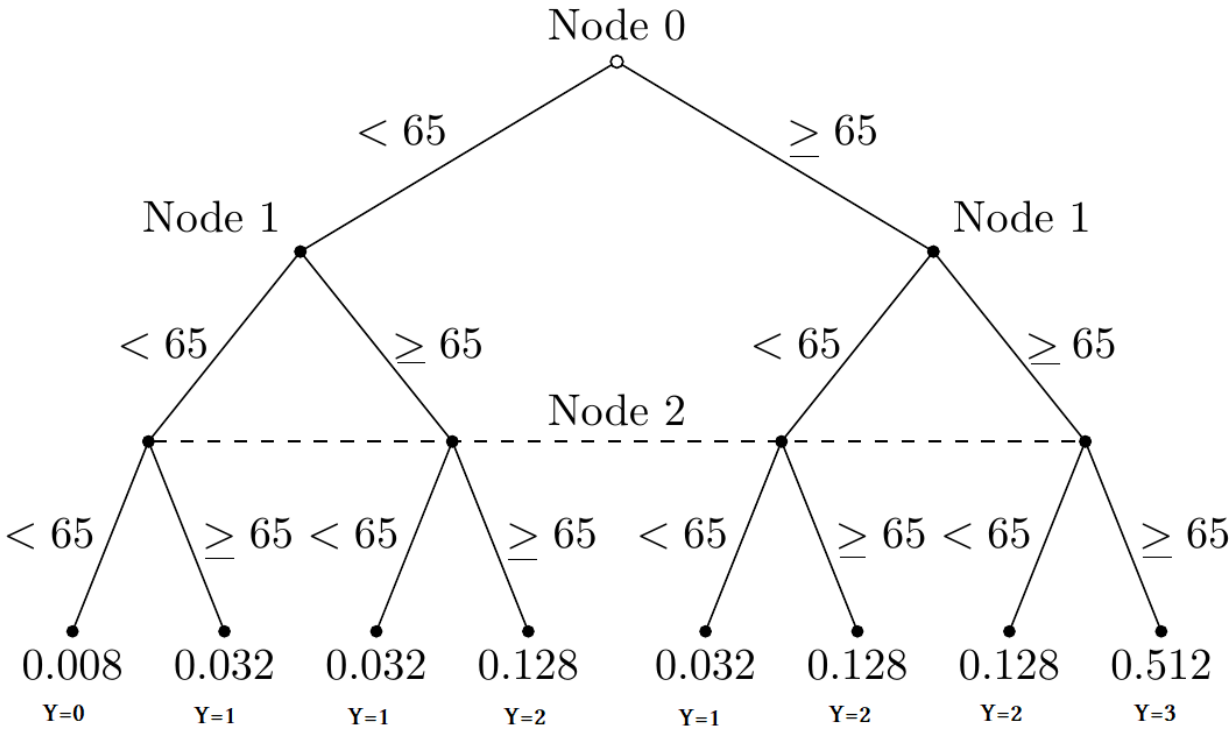
U.S. National Center for Health Statistics states that there is an 80% that a person aged 20 will be alive at age 65. If we randomly select 3 people, and Y is the event that we select a person who is alive at 65, we can create a tree diagram to determine the possible outcomes.

- Determine the probability for each outcome using the results for the tree diagram.
- Use the Binomial distribution formula to show how we can get the same results
- Find the probability that at least one person out of the three was alive at age 65
- Find $\Pr(Y > 2)$ and $\Pr(Y < 1)$

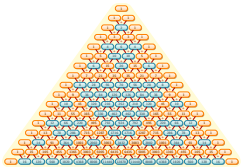


The tree

< 65 means $p = 0.2$ and ≥ 65 means $p = 0.8$. Multiply down the branches



There are multiple ways to get $Y = 1$ or $Y = 2$ (3 of each). We must sum these probabilities.



The Binomial Distribution

The $\binom{n}{k}$ term gives us the number of ways to get $Y = k$ from the tree and the $p^k(1-p)^{n-k}$ gives us the probabilities from the tree.

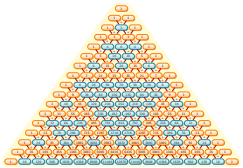
$$\Pr(Y=0) = \binom{3}{0}0.8^0(1-0.8)^{3-0} = 0.2^3 = 0.008$$

$$\Pr(Y=1) = \binom{3}{1}0.8^1(1-0.8)^{3-1} = 3 \cdot 0.8 \cdot 0.2^2 = 0.096$$

$$\Pr(Y=2) = \binom{3}{2}0.8^2(1-0.8)^{3-2} = 3 \cdot 0.8^2 \cdot 0.2 = 0.384$$

$$\Pr(Y=3) = \binom{3}{3}0.8^3(1-0.8)^{3-3} = 0.8^3 = 0.512$$

- Make sure they sum to 1! A little bit easier to use the formula, but not quite as clear



Which Event is a Success?

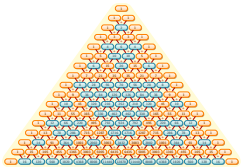
Suppose a quality control inspector selects a random sample of 10 motherboards from a very large shipment for inspection. Unknown to the inspector, only 60% of the motherboards meet specifications. What is the probability that no more than 1 of the 10 motherboards in the sample fails the inspection?

- Say X is the R.V. which describes the number of motherboards that fail to meet specifications, i.e. failing to meet specifications is the success, i.e.

$$\Pr(X = k) = \binom{10}{k} 0.4^k (0.6)^{10-k}$$

- Y could also be the # that do meet specs, i.e.

$$\Pr(Y = k) = \binom{10}{k} 0.6^k (0.4)^{10-k}$$



Mean and Variance of Binomial

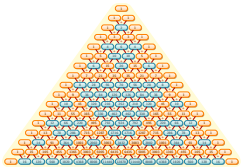
- The mean is

$$\mu = np \tag{2}$$

- The variance is

$$\sigma^2 = np(1-p)$$

- What are the mean and variance from the Ladies Home Journal magazine dog example?



Using the Calculator

Finding $\Pr(X = k)$ from binomial

1. Press 2nd key
2. Press VARS
3. Scroll to "bindomPDF"
4. The form is binomdPDF(n,p, k)

Finding $\Pr(X < k)$ from binomial

1. Press 2nd key
2. Press VARS
3. Scroll to "bindomCDF"
4. The form is binomdCDF(n,p, k)

This will not include the k'th term. To get $\Pr(X \geq k)$, do 1-the answer from the above procedure

Chapter 13 Notes

Sampling Distributions
STP-231

Arizona State University

Parameters vs Statistics

- Statistics desribe sample characteristics, $\overline{Y}$ and S are sample mean and sample standard deviation
- Parameters describe population characteristics. Usually trying to estimate these
- Proportions: are a relative frequency, can be for population or sample

Measure	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\overline{Y}$	μ
Standard deviation	S	σ
Variance	S^2	σ^2

Sampling

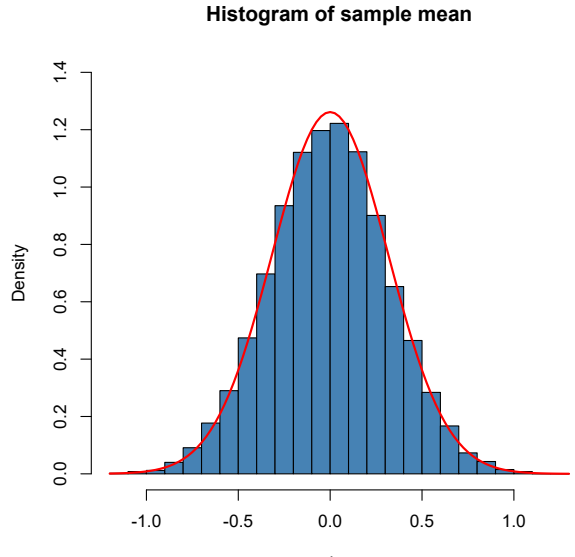
- We conduct an experiment on a random sample, hoping to be representative of a population
- Cannot expect sample to perfectly approximate population
- Statistics is distinguishing whether differences between sample and population are from random chance or if there is a real effect
- **Sampling error**: discrepancy between the sample and population

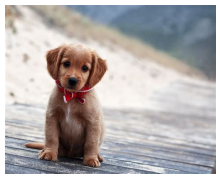
Sampling (continued)

If a sample is chosen randomly:

- Statistics themselves have their own distribution, i.e. $\hat{p}$ and $\overline{Y}$ are random variables themselves

- For example, the picture below is the distribution of the sample mean of a random sample from a normal distribution ($\overline{Y} \sim N(\mu_Y, \sigma_Y^2/n)$)





Sampling (continued)

If a sample is chosen randomly:

- We can interpret the sampling distribution of a statistic as "what would happen if we sample many times", though this is often not feasible

- Simulation is one solution, we can use software to imitate the random behavior

Sampling Distribution of a statistic

- Distribution of all of the possible observations of a statistic for samples of a given size from a population
- Ideal pattern that would emerge if we could view all samples of a given size. Shape, center, and spread are still useful descriptive properties
- The goal is to see how closely does our sample resemble the population

Notation

- Population parameters will either be subscripted with the random variable in question, "pop", or not at all, i.e.

$$\mu_Y \text{ or } \mu_{\text{pop}} \text{ or } \mu$$

same deal with σ

- The parameters of the sampling distribution of any statistic will be subscripted with the statistic in question, i.e. $\mu_{\overline{Y}}$ or $\sigma_{\overline{Y}}$ if we are looking at the sample mean distribution, or $\mu_{\hat{p}}$ if looking at sample proportion for example

- The sampling distribution of the mean is $N(\mu, \sigma/\sqrt{n})$ when the underlying population is normal

Meta Study

- Experiments where every possible sample that can be taken from a population is indeed taken
- Collecting data from every sample is unlikely
- Theoretical experiments, not often done in practice

Example 1: Meta Study

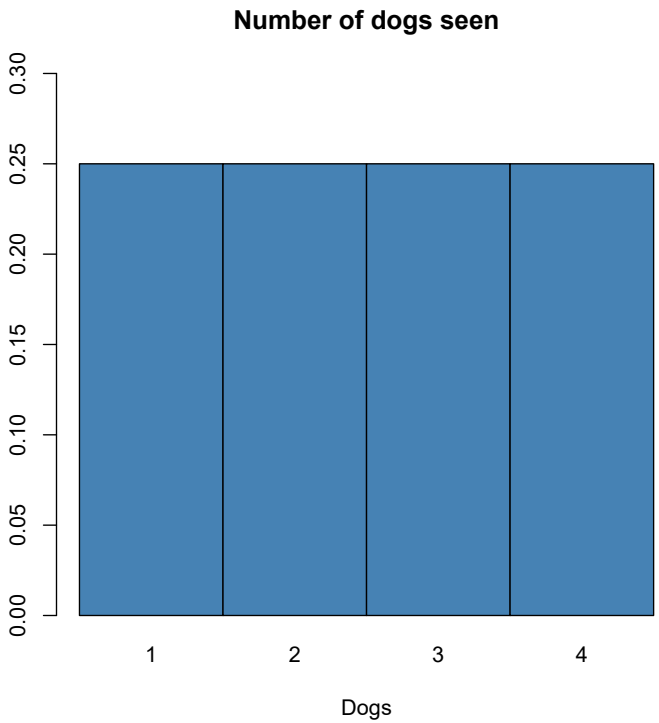
Imagine we are part of a class project in a veterinary class. We want to find out how many pups ASU students have.
Unfortunately, we cannot ask every student, so we merely ask 2.
Assume

- The actual number of dogs follows a uniform(1,4) distribution (same probability for each whole number 1,2,3, or 4)
- Let X be the # of dogs people have, where $\Pr(X = 1) = \Pr(X = 2) = \Pr(X = 3) = \Pr(X = 4)$
- This means nobody has 0 dogs, and nobody has more than 4 (assume this is population truth)

Population Characteristics

$$\mu_X = 1 * .25 + 2 * .25 + 3 * .25 + 4 * .25 = 2.5 \qquad \sigma_X = \sqrt{\frac{15}{12}}$$

Where σ_X is calculated similarly.



Possible Samples of Size $n = 2$

- If we choose 2 people and ask how many dogs they saw, we have 4^2 possible combinations answers they give us. This is because here we sample with replacement (i.e. an observation can be chosen again once its in the sample). This is so we can find a probability of that specific combination

- The 16 possible samples:

$$\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(2,4),(3,1),(3,2),(3,3),(3,4),(4,1),(4,2),(4,3),(4,4)\}$$

Sampling Distribution

The order of the data for each sample is negligible, i.e. (1,2) is the same as (2,1)

Sample	Pr(Sample)
1 and 1	0.0625
2 and 1	0.125
3 and 1	0.125
4 and 1	0.125
2 and 2	0.0625
3 and 2	0.125
4 and 2	0.125
3 and 3	0.0625
4 and 3	0.125
4 and 4	0.0625

Sampling Distribution of $\overline{X}$

Sample Mean	Sample
1	1 and 1
1.5	2 and 1
2	1 and 3; 2 and 2
2.5	1 and 4; 2 and 3
3	2 and 4; 3 and 3
3.5	4 and 3
4	4 and 4

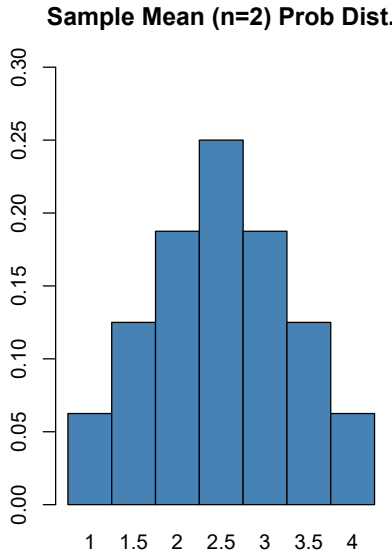
Sampling Distirbution of $\overline{X}$ Continued

Sample Mean	Pr(Sample Mean)
1	0.0625
3/2	0.125
2	0.1875
5/2	0.25
3	0.1875
7/2	0.125
4	0.0625

Example:

$$\Pr\left(\overline{X} = 2\right) = \Pr(\text{sample with observations “2” and “2”}) + \Pr(\text{sample with observations “1” and “3”}) = 0.1875$$

Summary Measures of Sampling Distribution of $\overline{X}$



$$\mu_{\overline{X}} = 2.50 \qquad \sigma = \sqrt{0.67} = 0.81 \approx \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

where $n = 2$. However, using a normal approximation implies we'd have a sample mean of 4.1 dogs or higher in 2.5% of our samples. Not the best!

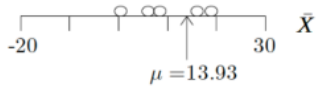
Example 1: Conclusions

- We constructed the sampling distribution of the mean by exhausting all possible samples within our study.
- The sampling distribution mean was equal to the mean, and the standard deviation was a function of the population standard deviation
- The sample distribution appears approximately normal. If you were to choose multiple random samples from the population, the sample mean would be exactly the population mean 25% of the time.

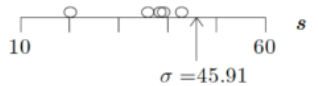
Example: Variability of Random Samples

Dot plots of sample statistics from 5 random samples of size $n = 25$.

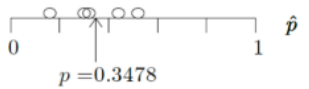
- **Five sample means:**



- **Five sample standard deviations:**



- **Five sample proportions:**



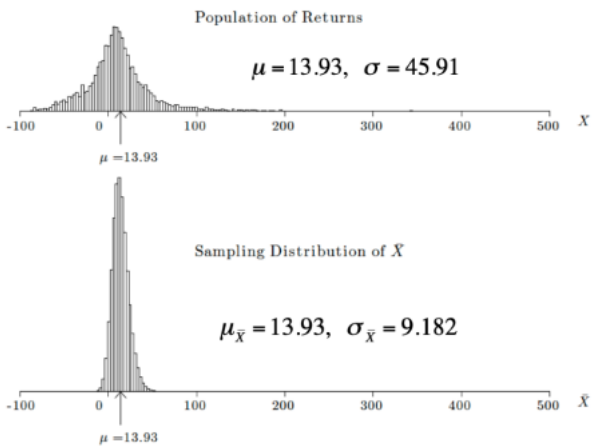
Notice the placement of the pop mean, pop proportion, and pop standard deviation relative to stats

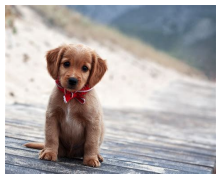
Takeaways

- Statistics vary from sample to sample
- Statistics of samples are themselves random variables and have distributions. This allows us to generalize our results without taking every possible sample, of which there are a lot
- We can answer questions like how close to the true μ is $\overline{X}$ likely to be?
- What is the shape, center, and spread of the distribution?

Example 2

Take 1,000,000 samples of size $n = 25$ (top) and $n = 625$ (bottom) and plot the distribution of the resulting sample averages. Notice, the mean location is the same, but the variance is tighter. Notice, the number of samples is there to show the behavior is true, but the variance only depends on sample size!





Properties of sampling distribution of $\overline{X}$

- Assume our population is normally distributed. Then:
- The mean of the sampling distribution of $\overline{X}$ is the population mean, i.e. it's unbiased
- The spread of the sampling distribution of $\overline{X}$ is given by the standard deviation, which is

$$\sigma_{\overline{X}} = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

Central Limit Theorem

- In fact, for any distribution as the sample gets larger:
- The standard deviation of the sampling distribution decreases
- The sampling distribution becomes normal, specifically of the sample mean
- n=10 pretty well usually CLT viz

Central Limit Theorem Continued

- Regardless of shape, $n = 10$ provides an approximately normal sampling distribution
- A distribution tends to look very normal for $n = 30$
- Regardless of normaility of the population, the sampling distribution of the sample mean will be normal provided n is large enough

Example 3

Phone call lengths are distributed normally with $\mu = 8$ and $\sigma = 2$ minutes. If you select a random sample of 25 calls, what percentage of the sample means would be between 7.8 and 8.2 minutes?

Example 3

Phone call lengths are distributed normally with $\mu = 8$ and $\sigma = 2$ minutes. If you select a random sample of 25 calls, what percentage of the sample means would be between 7.8 and 8.2 minutes?
With a sample of 25, we'd expect a sampling distribution of the mean to be

$$N(8, 2/\sqrt{25}) = N(8, 0.40)$$

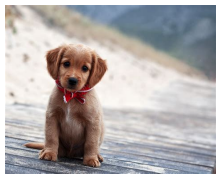
So we use normcdf(7.8, 8.2, 8, 0.40), which from the calculator is ≈ 0.38 . So about 38% of the data would be in this range.

Sampling Distribution of $\hat{p}$

- $\hat{p}$ is the proportion of observations that satisfy a condition within a sample

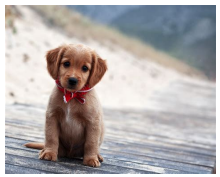
- Let X be the count of the occurrences of some categorical outcome in a fixed number of observations. Let n be the number of observations. Then the sample proportion is

$$\hat{p} = \frac{X}{n}$$



Sampling Distribution of $\hat{p}$

- Choose an SRS of size n from a large population that contains population proportion p of successes. Then:
- $\hat{p} = \frac{\#\text{ of successes in a sample}}{\text{sample size}}$
- The mean of the sampling distribution if p, so its unbiased. nice
- The standard deviation of the sampling distribution if $\sqrt{p(1-p)/n}$. As the sample size increases, the sampling distribution $\hat{p}$ becomes approximately normal



Sampling Distribution of Variance

- The sample variance S^2 has a mean of σ^2.
- We'll see this later, but $(n-1)S^2/\sigma^2$ is distributed χ^2_{n-1}, which will be useful later on in the semester

Law of Large Numbers

- The standard deviations of the sampling distributions are functions of sample size

$$\sigma_{\hat{p}} = \sqrt{p(1-p) \cdot n}$$

$$\sigma_{\overline{X}} = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

- If we continue increasing n then the variability in the distributions will be reduced

- Reduce in spread means statistics will get closer and closer to their respective parameters,

$$\overline{x} \to \mu_X$$

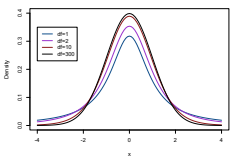
$$\hat{p} \to p$$

Our pupper friend

Chapter 14 Notes

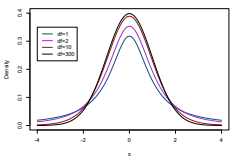
Confidence Intervals for One Mean
STP-231

Arizona State University



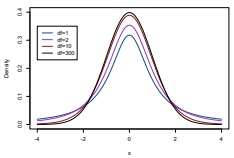
Statistical Inference

- We have essentially worked so far to use descriptive measures to discuss characteristics of observed data. We have discussed theoretical parameters of a population
- We have also discussed the limitations of using samples to estimate populations. Now, we more thoroughly quantify this
- **Statistical Inference** describes methods for making predictions about a population based on information collected from a sample
- For example, we will look at statistical estimation, confidence intervals, and hypothesis testing



Statistical Estimation

- In statistical estimation, we make inferences about data to:
- Determine an estimate of some value of the population (i.e. a statistic to estimate a parameter)
- Determine how precise the estimate is (i.e. figure out the standard deviation of our estimator/statistic)
- A **point estimate** is a value of a statistic that is used as an estimate of a parameter, i.e. our "best guess"

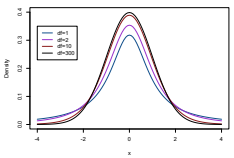


Example 1

As part of a larger study of body composition, researchers captuered 14 Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in cm^2)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wing size	33.9	33.0	30.6	36.6	36.5	34.0	36.1	32.0	28.0	32.0	32.2	32.2	32.3	30.0

Calculate the mean and standard deviation



Example 1

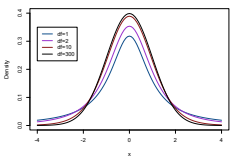
As part of a larger study of body composition, researchers captuered 14 Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in cm^2)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wing size	33.9	33.0	30.6	36.6	36.5	34.0	36.1	32.0	28.0	32.0	32.2	32.2	32.3	30.0

Calculate the mean and standard deviation Let Y be the random variable representing the wing area. What are the units on each?

$$\overline{Y} \approx 32.81 \qquad s \approx 2.48$$

both are in units cm^2



Standard Error of the Mean

- The standard deviation of $\overline{Y}$ is $\frac{\sigma}{\sqrt{n}}$ Why? Because $\text{Var}(a \cdot X) = a^2\text{Var}(X)$, that is:

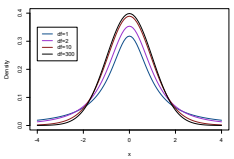
$$\text{Var}(\overline{Y}) = \text{Var}(\sum_{i=1}^{n} \frac{y_i}{n}) \tag{1}$$

$$= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} y_i) \text{ Since constants square out} \tag{2}$$

$$= \frac{1}{n^2} n\text{Var}(Y) \text{ (Since } y_i \text{ identically distributed)} \tag{3}$$

$$= \frac{n}{n^2}\text{Var}(Y) \implies \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} \tag{4}$$

where the last line is b/c $\text{Var}(Y) = \sigma^2$.



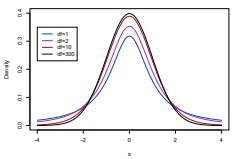
Standard Error of the Mean

- Say we are using $\overline{Y}$ as a point estimate of μ. We use S as an estimate of σ

- Similarly, $\frac{\sigma}{\sqrt{n}}$, the standard deviation of the sampling distribution can be estimated with $\frac{S}{\sqrt{n}}$ We call this the

  standard error of the mean,

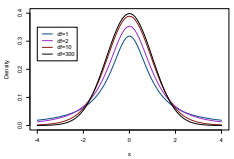
$$\text{SE}_{\overline{Y}} = \text{SE} = \frac{S}{\sqrt{n}}$$



Standard Deviation vs Standard Error

- Standard deviation: Inherent variation (or within sample variation), refers to spread of the sample values. What we "expect" within a sample
- Standard error: Uncertainty due to sampling error in the mean of the data
- Measuring reliability that our calculated sample means are close to the actual population mean. Factors in within sample variation and the sample size (sample to sample variation)

Confidence Intervals

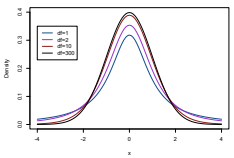


Confidence Intervals

- **Confidence Interval**: Also known as CI, is an interval of values obtained from a point estimate of a parameter
- Usually takes the form

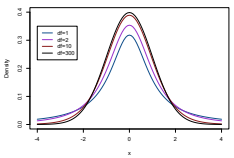
$$\text{point estimate} \pm \text{margin of error}$$

- **Confidence Level**: How sure we are the parameter lies in the confidence interval
- **Margin of error**: Some measure of the sampling error
- **Confidence – interval estimate**: The confidence level and interval



Confidence Level

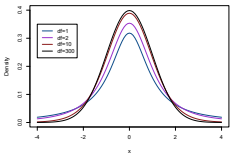
- **Confidence level**: the confidence we have that the parameter lies in the confidence interval
- Written in the form $1 - \alpha$, where α is a number between 0 and 1
- α is called the significance level
- $1 - \alpha$ is the success rate of the method that produces the interval



Example 2

The number of offspring of Eastern Cotton Mouth snakes is believed to be smaller b/c of human encroachment. 44 female snakes were randomly samples and the number of offspring from each snake was counted. The data is below:

5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6



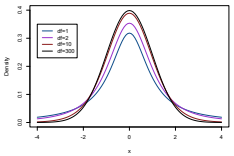
Example 2

The number of offspring of Eastern Cotton Mouth snakes is believed to be smaller b/c of human encroachment. 44 female snakes were randomly samples and the number of offspring from each snake was counted. The data is below:

5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6

This yields $\overline{Y} = 7.59$ young/litter and $\sigma = 2.4$ (assuming we know standard deviation)

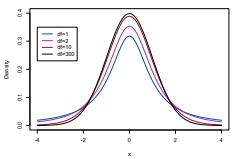
- We know $\overline{Y}$ is a point estimate for μ. How likely is it that they are exactly equal? We can create a range of values where we have some confidence that μ will fall into



Example 2 continued

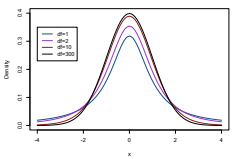
5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6

- Consider two standard deviations away from $\overline{Y}$. We are 95.44% confident that the mean falls within these limits
- The CI is $\overline{Y} \pm 2 \times \frac{\sigma}{\sqrt{n}} = 7.59 \pm 2 \times \frac{2.4}{\sqrt{44}}$, which is (6.87, 8.31)
- **Interpretation** "If we were to do this study 100 times, approximately 4-5 of them would not contain the true population mean, i.e. 4-5 of them would have sample means less than 6.87 or greater than 8.31".
- For example, if sample mean was 6.86 from a certain sample, it's 2 σ CI would be (6.14, 7.58) which does not contain the true mean of 7.59.



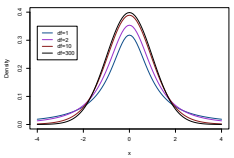
Interpreting a Confidence Interval

- The confidence level is the success rate of the method that produced the interval. We don't know if the $1-\alpha$ CI for particular sample will be one of these successes. We do know:
- $(1-\alpha) \times 100\%$ if intervals will be success
- $\alpha \times 100\%$ will not be
- Example: To say we are 95% confident that the unknown value of μ falls within a 1-0.05 confidence interval is saying: "We got these intervals using a method that gives correct results 95% of the time."
- **The confidence level should never be interpreted as the probability that a parameter is within a specific confidence interval"
- cool link



Example 3: Incorrect Interpretation

- We obtain a 90% confidence interval for a population parameter μ : $(8.13, 10.4)$
- Based on one sample alone. We are not interested in quantifying $\Pr(8.13 < \mu < 10.4)$ because μ is fixed and so is the sample for this confidence interval in particular
- The probability mentioned above is either 0 or 1, the mean is in the interval or it is not
- We do consider how often this method does contain the population parameter
- In this way the confidence level is the proportion of successes that we would count over many repetitions of this method
- "90% of the time the mean would be contained in an interval formed this way

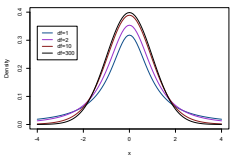


Confusing

- Confidence intervals confuse a lot of people. Case in point:

> The uncertainty around a point estimate can be small or large. Scientists represent this uncertainty by calculating a range of possibilities, which they call a confidence interval. One way of thinking of a confidence interval is that we can be 95 percent confident that the efficacy falls somewhere inside it. If scientists came up with confidence intervals for 100 different samples using this method, the efficacy would fall inside the confidence intervals in 95 of them.

- These are conflicting interpretations (from the New York Times)



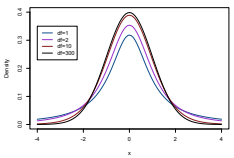
Why is Bayesian Statistics Different?

Assume we take a test that is 97% true positive and 2% false positive. Assume 60% of people taking the test have the disease of interest. What is the probability that if someone tests positive twice in a row they actually have the disease?
Let P be the random variable having the positive test. Then from bayes rule:

$$\begin{aligned}\Pr(D \mid P) &= \frac{\Pr(P \mid D)\Pr(D)}{\Pr(P \mid D) + \Pr(N \mid D)\Pr(N)} \\ &= \frac{0.97 \times 0.60}{0.97 \times 0.60 + 0.02 \times 0.4} = 0.9864\end{aligned}$$

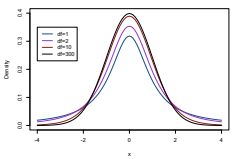
Now, for the second text, $\Pr(D)$ becomes 0.9864, i.e. we are updating our *prior* belief, yielding

$$\Pr(D \mid P, P) = \frac{0.97}{0.97 \times 0.9864 + 0.02 + 0.01364} = 0.9997$$



Bayesian interpretation part 2

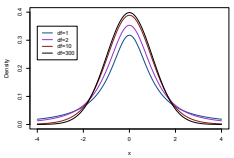
In terms of Bayesian statistics, the parameter is not **fixed** but itself a random variable which is updated based on what our prior belief is (probability of D we assumed before seeing any data)



Confidence Interval for the Pop. Mean (σ known)

- Assume a simple random sample
- Also assume the population is normally distributed and we have a large sample
- If we know the population standard deviation, then the $(1-\alpha)\times 100\%$ confidence interval for the population is:

$$\overline{Y} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$



Intuition

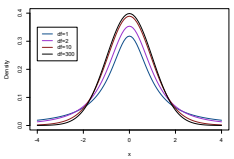
- Let C be any confidence level. We call $z^* = z_{\alpha/2}$ and $-z^* = -z_{\alpha/2}$ the critical values. They are the values that provide probability C within their range under the curve.
- If we start at the sample mean and move outward by z^* standard deviations on either side, we get an interval that contains the population mean μ in a proportion C of all samples
- Recall:

$$z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \implies \mu = \overline{Y} - z \times \frac{\sigma}{\sqrt{n}}$$

solve:

$$z_{\alpha/2} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \text{ and } -z_{\alpha/2} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

for μ



Visual Intuition

We use α instead of C.

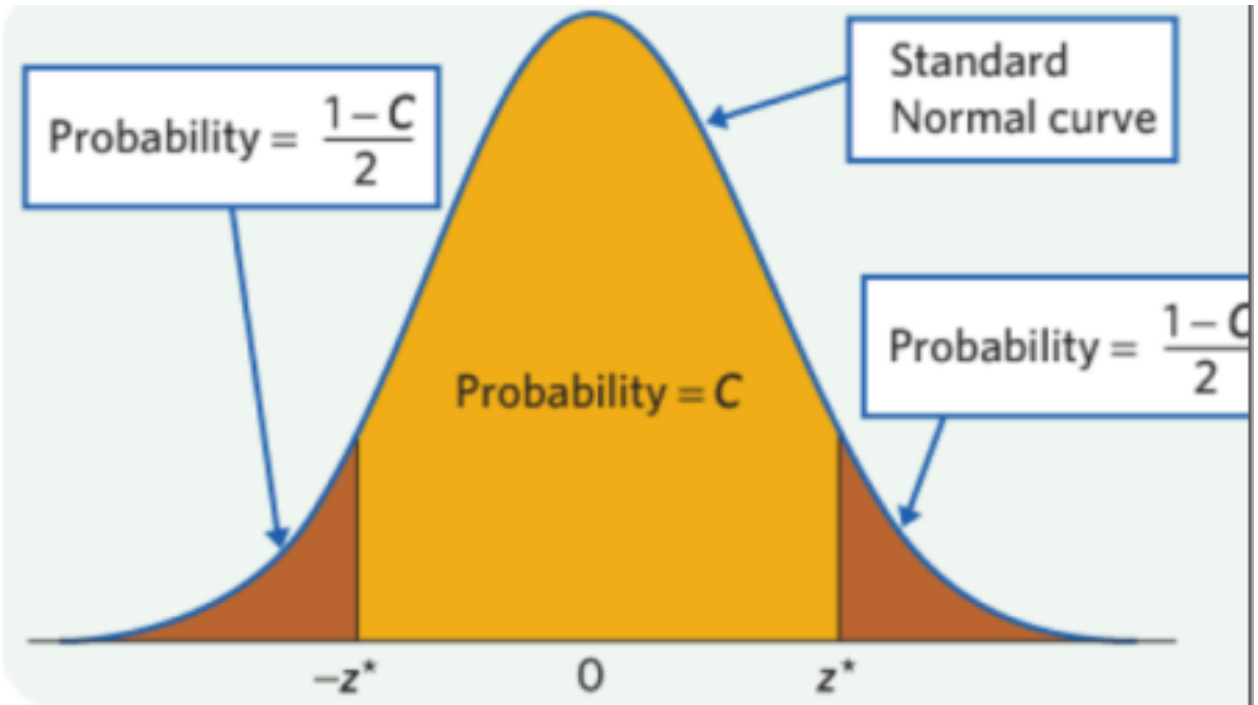
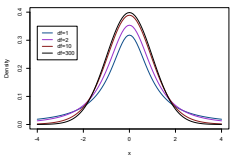


Figure 14.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 20
W. H. Freeman and Company



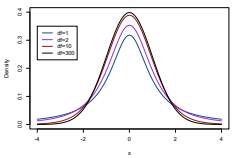
Example 4

Researchers are trying to predict the mean body temperature of humans. They obtained the body temperature of 93 humans. The mean body temperature if 98 degrees fahrenheit. Assume population standard deviation is 0.63 degrees fahrenheit (maybe we know from past studies). Find a 95% confidence interval for the mean body temperature. What is the correct interpretation?

$$\overline{Y} \pm 1.96 \times \frac{0.63}{\sqrt{93}} \longrightarrow (97.87, 98.13)$$

But isn't the agreed upon mean body temp 98.6? A sample with that sample mean would fall outside the CI. So we'd conclude the value from old studies is wrong...but we use old studies σ to prove new studies are wrong, so using known σ is dangerous!

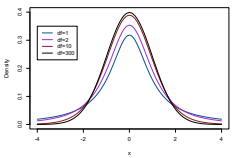
When σ is unknown



When σ is unknown

- This is more realistic
- In this case, we use the sample standard deviation as an estimate of the truth, i.e. S instead of σ
- Instead of the standard z, we now use the studentized version of $\overline{Y}$, which means we use a t-distribution to approximate a normal

$$z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \longrightarrow t = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

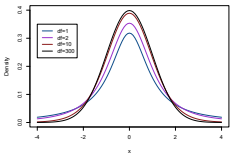


Student's t-distribution

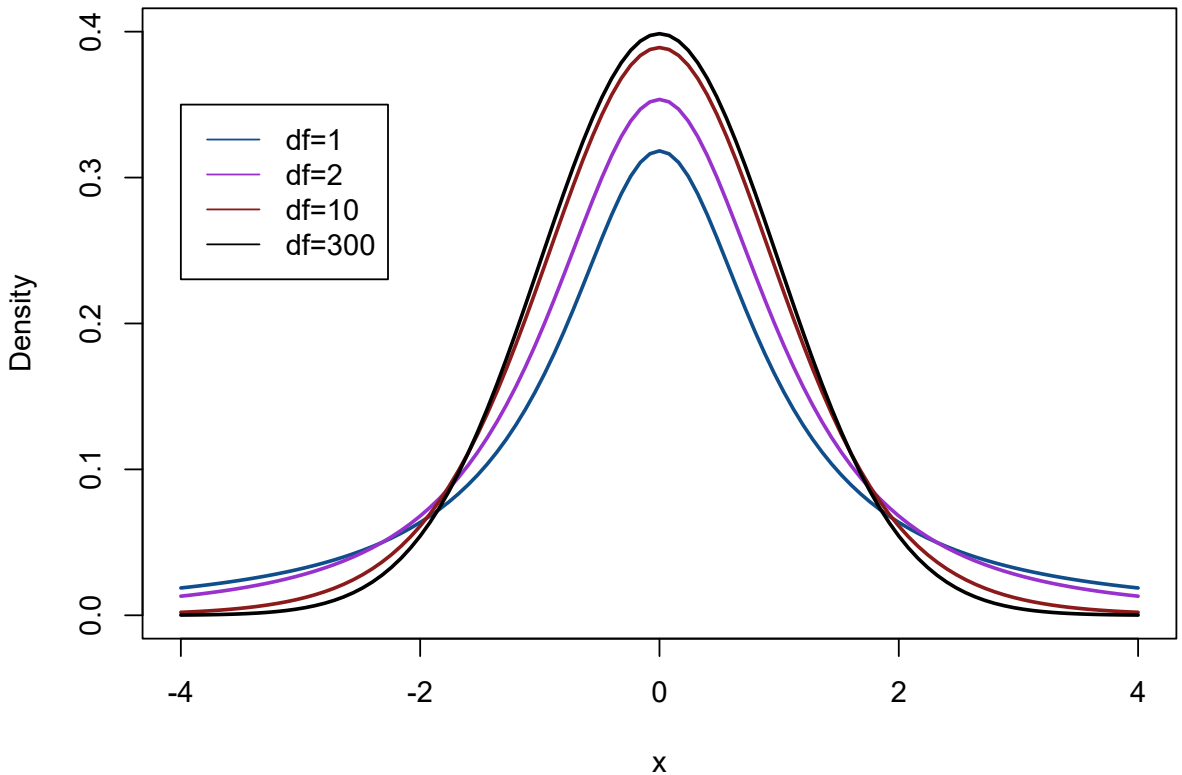
- Bell-shaped curve centered at 0, with thicker tails
- Uses degrees of freedom, where df=$n-1$ for n data points in the sample
- As n increases, looks more and more like normal distribution
- In all it's glory, the distribution is given by (where $\nu = n-1$ degrees of freedom) Wiki article

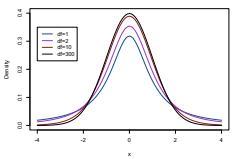
$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}$$

$\Gamma(n) = (n-1)!$ if integer. O.w. $\Gamma(x) = \frac{1}{x}\Gamma(x+1)$



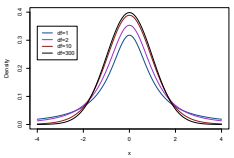
Visual t-distribution





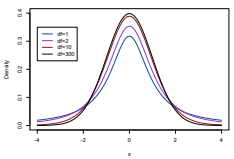
How to make the plot

```
#make the curves
curve(dt(x, df=1), lwd=2,lty=1,from=-4, to=4, col='dodgerblue4', ylim=c(0, 0.4), ylab='Density')
#need add=T to get multiple.
curve(dt(x, df=2), lwd=2,lty=1,from=-4, to=4, col='darkorchid', add=T, ylim=c(0, 0.4), ylab='Density')
curve(dt(x, df=10), lwd=2, lty=1,from=-4, to=4, col='firebrick4', add=TRUE, ylim=c(0,0.4), ylab='Density')
curve(dt(x, df=300), lwd=2,lty=1,from=-4, to=4, col='black', add=TRUE, ylim=c(0, 0.4), ylab='Density')
#add legend
legend(-4, .35, legend=c("df=1", "df=2",  "df=10", "df=300"),
col=c("dodgerblue4","darkorchid",  "firebrick4", "black"), lty=c(1,1,1,1), cex=1.)
```



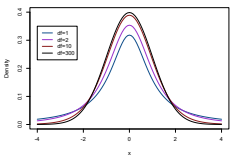
More on the t-distribution

- Use the t-table for the area under the curve
- Area under the curve is the right tail for df=$n-1$ where $1-\alpha$ is the confidence level and α is the significance level. Find:
- $t_{0.10}$ for $n = 10$
- $t_{0.05}$ for $n = 26$
- $t_? = 1.948$ for $n = 11$
- $t_? = 1.973$ for $n = 10$



More on the t-distribution

- Use the t-table for the area under the curve
- Area under the curve is the right tail for df=$n-1$ where $1-\alpha$ is the confidence level and α is the significance level. Find:
- $t_{0.10}$ for $n = 10$
- $t_{0.05}$ for $n = 26$
- $t_? = 1.948$ for $n = 11$
- $t_? = 1.973$ for $n = 10$
- Use the t-table for the area under the curve
- Area under the curve is the right tail for df=$n-1$ where $1-\alpha$ is the confidence level and α is the significance level. Find:
- $t_{0.10}$ for $n = 10$=-1.38
- $t_{0.05}$ for $n = 26$=-1.71
- $t_? = 1.948$ for $n = 11$=0.96

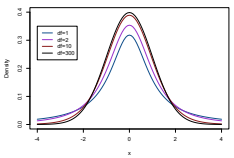


CI for pop mean (σ unknown)

We make the following assumptions:

- Take a simple random sample (SRS)
- Normal population or a large sample
- σ is unknown
- For a confidence level of $1-\alpha$, we can generate a confidence interval for μ by

$$\overline{Y} \pm t_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

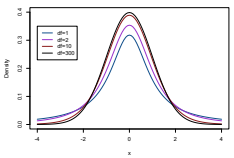


Example 4

Snakes deposit chemical trails as they travel through their habitats. These trails are often detected and recognized by lizards which are potential prey. The ability to recognize their predators via tongue flicks can often mean life or death for lizards. Scientists were interested in quantifying the responses of the common lizard to natural predator cues to determine whether the behavior is learned or congenital. Seventeen juvenile common lizards were exposed to the chemical cues of the viper snake. Their responses in number of tongue flicks per twenty minutes are:

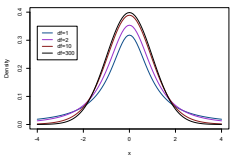
425	510	629	236	654	200
276	501	811	332	424	674
676	694	710	662	663	

Construct a 90% confidence interval and interpret



Example 5

The subterranean coruro (Spalacopus cyanus) is a social rodent that lives in large colonies in underground burrows that can reach lengths of up to 600 meters. A sample of 51 burrows had an average depth of 15.05 centimeters with a sample standard deviation of 2.50 centimeters. Construct a 95 % confidence interval for the mean burrow depth for all subterranean coruros.

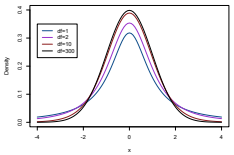


Example 5

The subterranean coruro (Spalacopus cyanus) is a social rodent that lives in large colonies in underground burrows that can reach lengths of up to 600 meters. A sample of 51 burrows had an average depth of 15.05 centimeters with a sample standard deviation of 2.50 centimeters. Construct a 95 % confidence interval for the mean burrow depth for all subterranean coruros. Note, $S = 2.50$. Then:

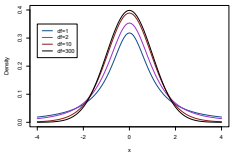
$$\begin{aligned} 15.05 \pm \frac{2.50}{\sqrt{51}} \times t_{0.025,51-1} \\ = 15.05 \pm -2.01 \times 0.35 \implies 14.35 < \mu < 15.75 \end{aligned}$$

Planning Study Sample Size



Planning a Study for Estimating μ

- Find sufficient sample size to estimate the parameter with an acceptable confidence level
- After spending time and resources it is upsetting to realize there is an insufficient amount of data to draw a reasonable conclusion
- Reduce variability: Example, for a Breast Cancer study on five-year survival rates the data on the patients could be organized into groups such as the following:
- Stage I-IV diagnosis
- Pre-menopausal or post-menopausal
- Estrogen, progesterone, or HER2 negative/positive
- Chemotherapy-yes or no
- Radiation-yes or no



Determining Sample Size

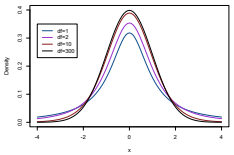
- What sample size will be sufficient to achieve a desired degree of precision in estimation of the population mean?

- Use the standard error as our measure of precision

$$\text{SE}_{\overline{Y}} = \frac{S}{\sqrt{n}}$$

- The required sample size is then determined from the following equation

$$\text{Desired Standard Error} = \frac{\text{Guess standard deviation}}{\sqrt{n}}$$

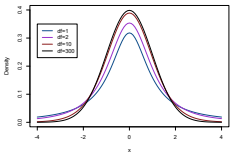


Example 6

Assume we have the following data on butterfly wings.

$$\overline{Y} = 32.81 \text{ cm}^2$$
$$S = 2.48 \text{ cm}^2$$
$$\text{SE} = 0.66 \text{ cm}^2$$

Suppose the researcher is now planning a new study of butterflies and has decided that it would be desirable that the SE be no more than 0.4 cm^2. What n would we need?



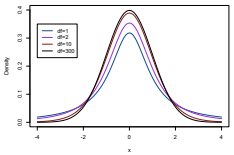
Margin of Error, Precision, Sample Size

- Length of a confidence interval is a measure of the precision with which $\overline{Y}$ estimates μ
- For a fixed confidence level, increasing the sample size improves the precision
- If a confidence level and margin of error is given, then the appropriate sample size needed to meet those specifications must be determined from the formula:

$$E = t_{\alpha/2} \times \frac{S}{\sqrt{n}} \rightarrow n = \left(\frac{t_{\alpha/2,n-1} \cdot S}{E} \right)^2$$

If σ is known, then the equation is

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$



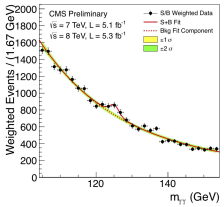
Example 7

Physical therapy students during their graduate-school years were studied by the College of Health at the University of Nevada, Las Vegas. The researchers were interested in the fact that, although graduate physical therapy students are taught the principles of fitness, some have difficulty finding the time to implement those principles. Assuming that percent body fat of graduate physical-therapy students is normally distributed with standard deviation 4.10 percent body fat. Determine the sample size required to have a margin of error of 1.25 percent body fat with 95% confidence level.

Chapter 14 Part 2 Notes

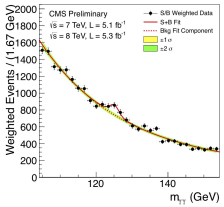
Hypothesis Testing for One Mean
STP-231

Arizona State University



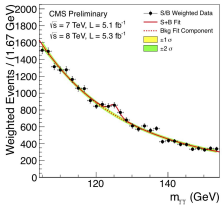
Hypothesis Tests (significance tests)

- Shift focus from estimating parameters (CI) to judging claims about a population
- The goal is the assess the strength of the evidence provided by data against some claim about a population
- We want to determine if vitamin C helps against colds. Take the mean of people's fevers. Then we give some people in the sample the treatment, vitamic C
- We have by now enough studies that know what the mean fever should be in a population so we compare out our sampled vitamin C peoples mean fever to that value. Is there significance?



Terminology

- **Hypothesis**: is a statement that something is true. We typically consider:
- **Null hypothesis**: A hypothesis to be tested denoted H_0
- **Alternative hypothesis**: A hypothesis to be considered as an alternative to the null hypothesis denoted H_a
- **Hypothesis test**: We decide if there is evidence to reject the null hypothesis in favor of the alternative hypothesis



Symbolically

- For the null hypothesis, there is no effect or difference given a treatment or not, i.e.

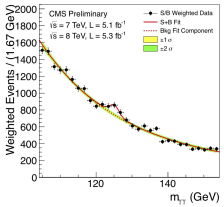
$$\mu = \mu_0$$

- For the alternative, if two-tailed (both sides add up), the parameter μ is different from the null value, i.e.

$$\mu \neq \mu_0 \implies \mu > \mu_0 \text{ or } \mu < \mu_0$$

If one tailed (only one side of the area curve matters) the parameter μ is greater than (or smaller than) the null value, i.e.

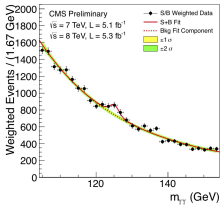
$$\mu > \mu_0 \text{ greater (right tail)}$$
$$\mu < \mu_0 \text{ smaller (left tail)}$$



Example 1

Dementia is the loss of the intellectual and social abilities severe enough to interfere with judgment, behavior, and daily functioning. Alzheimer's disease is the most common type of dementia. In the article "Living with Early Onset Dementia: Exploring the Experience and Development Evidence-Based Guidelines for Practice" the researchers explored the experience and struggles of people diagnosed with dementia and their families. A hypothesis test is to be performed to decide whether the mean age at diagnosis of all people with early-onset dementia is less than 55 years old.

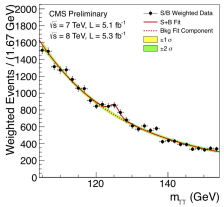
- Determine the null hypothesis
- Determine the alternative hypothesis
- Classify the test as two tailed, left-tailed, or right-tailed



Example 2

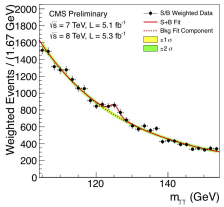
A study on "Heat Stress Evaluation and Worker Fatigue in a Steel Plant" assessed fatigue in steel-plant workers due to heat stress. Among other things, the researchers monitored the heart rates of a random sample of 29 casting workers. A hypothesis test is to be conducted to decide whether the mean post-work heart rate of casting workers exceeds the normal resting heart rate of 72 beats per minute (bpm).

- Determine the null hypothesis
- Determine the alternative hypothesis
- Classify the test as two tailed, left-tailed, or right-tailed



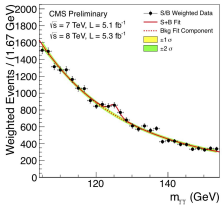
Intuition

- Say we sample from a population. Is the data consistent with the null hypothesis? If so, do not reject the null hypothesis. If not and supports the alternative, reject the null hypothesis in favor of the alternative
- We make the claim H_0 in hope that we can reject the claim.
- We use the evidence provided by the data to refute it and use probabilities as a measure of the strength of the evidence to reject



Intuition continued

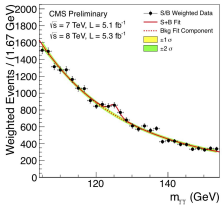
- Hypothesize that the post work heart rate mean is the same as the normal population mean (alternative: the mean isn't the same): $H_0 : \mu = 72, H_a : \mu > 72$
- Assume the null is true, i.e. $\mu = 72$. Then the sampling distribution of $\overline{X}$ at $n = 29 \sim N(72, \sigma/\sqrt{n})$
- If $\overline{X}$ is close to the hypothesized mean, within the central bulk of the hypothesized sampling distribution, this indicates that such a sample mean is likely to happen by chance when the population mean is 72.



Intuition continued

- Our logic is: observing an outcome that would rarely happen if a hypothetical claim was
  true is good evidence that this claim is not true

- In the example: If the sample statistic was found to be far above the true population mean
  even though this value has very small probability of occurring by chance we should be
  suspicious that our original assumption was true



Test Statistics and p-values

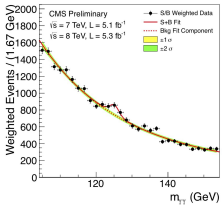
Test statistic

- By definition a function of our data. Measures how far the data diverge from the null hypothesis. Large values indicate the data are far from what we would expect if H_0 was true. The location on the x-axis of the density curve of interest

p – value

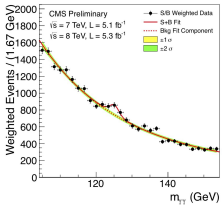
- Probability that the test statistic would take a value at least as extreme as the one actually observed (in the direction of H_a) i.e. the area under the curve (in specified direction) at the test statistic.

- Smaller p-value means more evidence against the null With small p-values we reject the null hypothesis and results are called **statistically significant**



p-value Meaning and Significance Level

- With small p-values we reject the null hypothesis and say are results are statistically significant
- Failing to reject the null hypothesis because a p-value is not small enough does not imply that H_0 is true
- **Significance level**: Denoted α, an arbitrary threshold to determine if the p-value is statistically significant
- If p-value $< \alpha$, the results are statistically significant

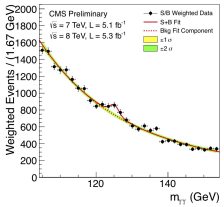


Conclusions

- If the null hypothesis is rejected, state "we conclude that the data provide sufficient evidence to support the alternative hypothesis and we reject the null"

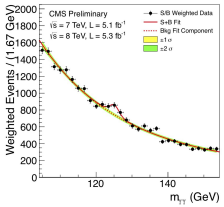
- If the null hypothesis is not rejected, state "we conclude that the data does not provide sufficient evidence to support the alternative hypothesis and we do not reject the null"

p-values



Hypothesis testing for p-value approach

- State the hypothesis
- State significance level α
- Compute value of test statistic
- Find the p-value and compare it to α. If it's smaller, reject the null
- State whether you reject H_0 or fail to reject H_0
- Interpret your results in the context of the problem

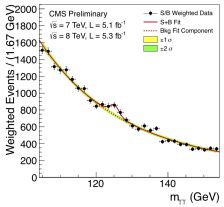


Z test for population mean (σ is known)

Begin with the usual assumptions that we have a simple random sample, a normal population or large sample, and σ is known. Then the test statistic is

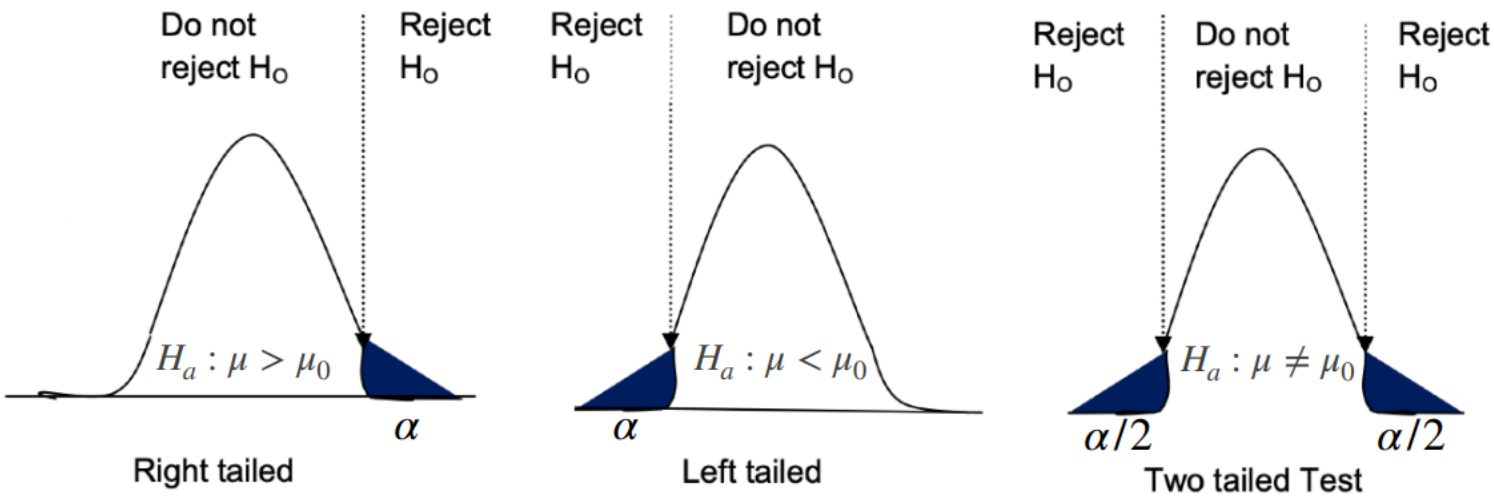
$$z^* = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

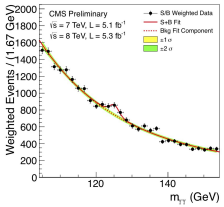
Compare this to the z value for which you'd reach the appropriate significance level



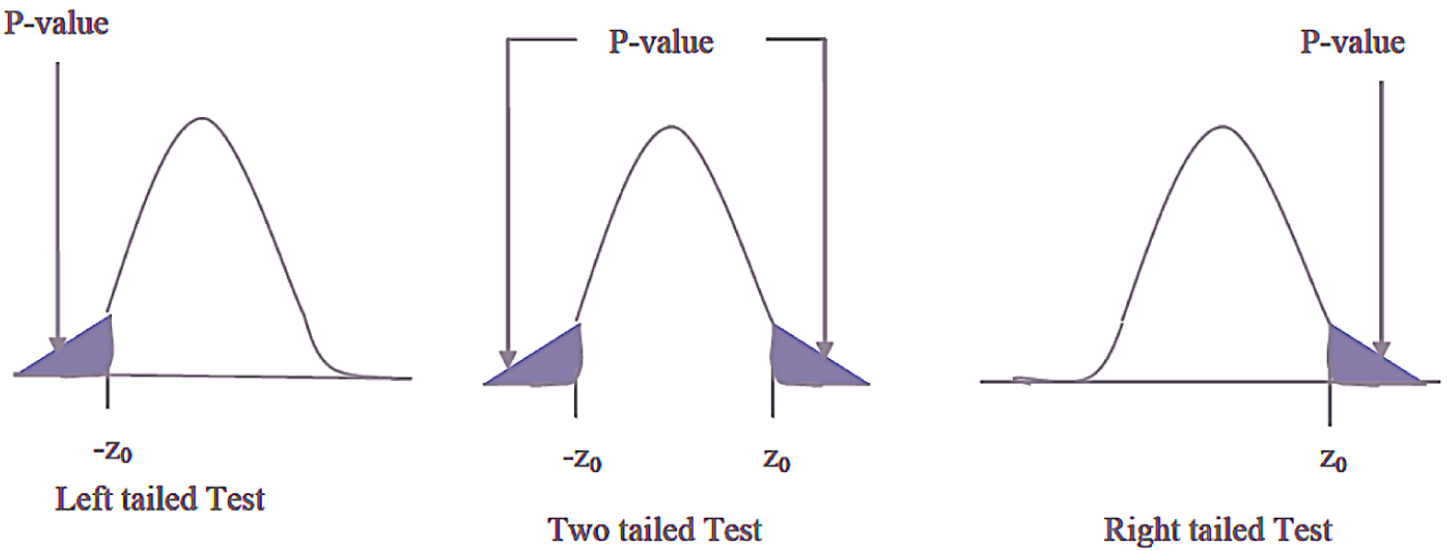
Z test for population mean (σ is known)

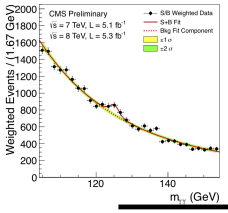
- Z follows a standard normal and we can determine if p-value $\leq \alpha$, i.e. the **rejection region**. Specifically, this is the set of values for the test-statistic that leads to rejection of the null hypothesis
- The **non − rejection region** is the set of values for the test statistic that leads to non-rejection of the null





P-values for z-test



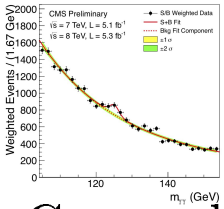


How to get P-values in Calculator

Test	Null	Alternative	Level	Reject if
Right sided	$\mu = \mu_0$	$\mu > \mu_0$	$1 - \alpha$	$t^* > t_{1-\alpha,\text{df}}$
Left sided	$\mu = \mu_0$	$\mu < \mu_0$	α	$t^* < t_{\alpha,\text{df}}$
Two sided	$\mu = \mu_0$	$\mu \neq \mu_0$	$\alpha/2$	$\lvert t^* \rvert > \lvert t_{\alpha/2,\text{df}} \rvert$

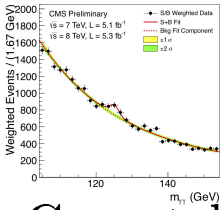
This is specifically for 1-sample, but the general idea holds for the one sample, just instead we compare μ_1 to μ_2 (chapter 18) and are not looking at a function of two means. In general, to calculate our p-values (replace tcdf(lower bound, upper bound, df) with normalcdf(lower bound, upper bound, 0,1) and our t^* with the z^* statistic

$$\text{Left-tail} \longrightarrow \text{tcdf}(-1E99, t^*, \text{df})$$
$$\text{Right-tail} \longrightarrow \text{tcdf}(t^*, 1E99, \text{df})$$
$$\text{Two-Tail} \longrightarrow 2 * \text{tcdf}(|t^*|, 1E99, \text{df})$$



Example 4

Gary tells his friends that he averages 5 miles a run. We gets data from a random sample of 20 runs of his runs, with sample mean of 4.8 miles, and known true standard deviation of all runs is 1.2 miles. At the 10% significance level, does the data provide sufficient evidence to conclude that Gary's running distance is less than 5 miles? Hint: we wanna reject the claim Gary is telling the truth



Example 4

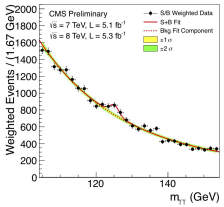
Gary tells his friends that he averages 5 miles a run. We gets data from a random sample of 20 runs of his runs, with sample mean of 4.8 miles, and known true standard deviation of all runs is 1.2 miles. At the 10% significance level, does the data provide sufficient evidence to conclude that Gary's running distance is less than 5 miles? Hint: we wanna reject the claim Gary is telling the truth The null is $H_0 : \mu = 5$, alternative $H_a : \mu < 5$. The test statistic is

$$z^* = \frac{4.8 - 5}{\frac{1.2}{\sqrt{20}}} = -0.7453 \implies \text{p-value} = \text{normcdf(-1E99, -0.745, 5, } \frac{1.2}{\sqrt{20}})$$

which equals 0.228 which is not less than 0.10, so we don't reject. This is a left-sided test where we want $p < 0.10$. This is true if

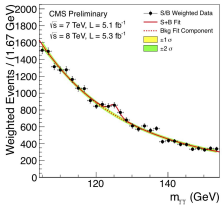
$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{0.10} \implies \frac{\overline{X} - 5}{1.2/\sqrt{20}} < -1.28$$

which means if his running mean was less than 4.656 (invnorm(0.1, 5, $1.2/\sqrt{20}$)) we'd reject the claim his true running distance is 5 miles.



Example 5

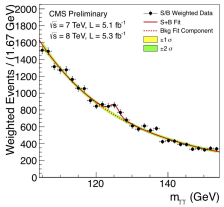
Job seekers are told by company A that on average, their entry level financial analysts have 2 years of relevant work experience. A random sample of 31 employees from company A are interviewed. All 31 employees started working at company A as entry level financial analyst and they had, on average, 3 years of relevant work experience at the time of hire. The known standard deviation is 1.5. Use α=0.1. Is there any evidence to suggest the true mean years of work experience is different from 2? Hint, since this is two sided, compare $|z^*|$ to $\left|z_{\alpha/2}\right|$



When to use z-test (σ known)

- $n < 15$ only when the variable under consideration is normally distributed or very close to being so
- $15 < n < 30$ can be used unless the data contains outliers or the variable under consideration is farm from being normally distributed
- If $n > 30$ we can essentially use the z-test without restriction, because S approximated σ and the t-distribution looks like the z-distribution in this case
- However, if outliers are present and their removal is not justified, we could be in trouble
- To remedy this, compare the z-test obtained with and without the outliers and see if the difference is influential. May need to use a different sample (if possible) depending on the scale of the issue

t-test for population mean

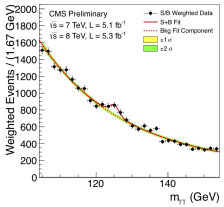


t-test for population mean

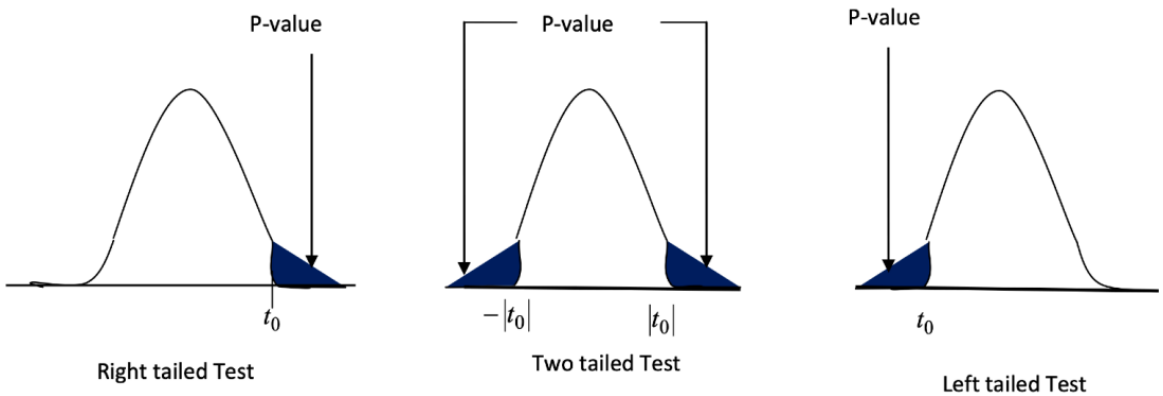
Begin with the assumptions:

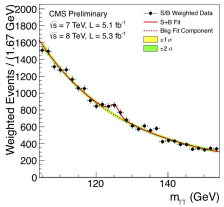
- We have a simple random sample, a normal population or large sample, and unknown σ
- Use the studentized version of $\overline{X}$
- The test statistic is:

$$t^* = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$



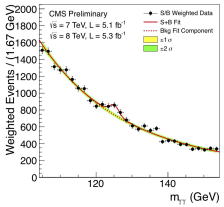
t-test for population mean





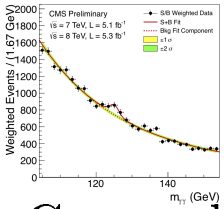
Example 6

Active management of labor (AML) is a group of interventions designed to help reduce the length of labor and the rate of cesarean deliveries. Physicians were interested in determining whether AML would translate into a reduced cost for delivery. According to the article published on this study, 200 AML deliveries had a mean cost of \$2480 with a standard deviation of \$776. Atthe time of the study, the average cost of having a baby in a U.S. hospital was \$2528. At the 5% significance level, do the data provide sufficient evidence to conclude that, on average, AML reduces the cost of having a baby in a U.S. hospital?



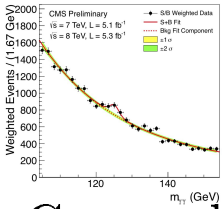
Example 7

According to the document Consumer Expenditures, a publication of the Bureau of Labor Statistics, the average consumer unit spent \$1874 on apparel and services in 2006. The same year, 25 consumer units in the Northeast had the following annual expenditures, in dollars, on apparel and services. At the 5% significance level, do the data provide sufficient evidence to conclude that the 2006 mean annual expenditure on apparel and services for consumer units in the Northeast differed from the national mean of \$1874. The sample mean and sample standard deviation of the data are \$2060.76 and \$350.90.



Example 8

Gary tells his friends that he averages 5 miles a run. We gets data from a random sample of 20 runs of his runs, with sample mean of 4.8 miles, and sample standard deviation of all runs is 1.2 miles. At the 10% significance level, does the data provide sufficient evidence to conclude that Gary's running distance is less than 5 miles? Hint: we wanna reject the claim Gary is telling the truth



Example 8

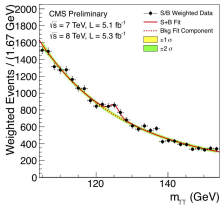
Gary tells his friends that he averages 5 miles a run. We gets data from a random sample of 20 runs of his runs, with sample mean of 4.8 miles, and sample standard deviation of all runs is 1.2 miles. At the 10% significance level, does the data provide sufficient evidence to conclude that Gary's running distance is less than 5 miles? Hint: we wanna reject the claim Gary is telling the truth The null is $H_0 : \mu = 5$, alternative $H_a : \mu < 5$. The test statistic is now t

$$t^* = \frac{4.8 - 5}{\frac{1.2}{\sqrt{20}}} = -0.7453 \implies \text{p-value} = \text{tcdf(-1E99,-.7453, 20-1)}$$

which equals 0.232 which is not less than 0.10, so we don't reject. This is a left-sided test where we want $p < 0.10$. This is true if

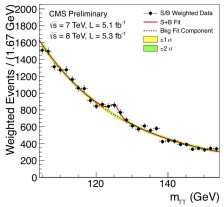
$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < t_{0.10} \implies \frac{\overline{X} - 5}{1.2/\sqrt{20}} < -1.33$$

which means if his running mean was less than 4.644 we'd reject the claim his true running distance is 5 miles.



Tests from confidence intervals

A level α two sided hypothesis test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when μ_0 falls outside a $1 - \alpha$ confidence interval for μ. They really are kinda the same thing!

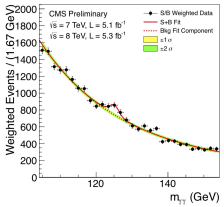


Type I and Type II error

- **Type I Error**: rejecting the null hypothesis when it is in fact true

- **Type II Error**: not rejecting the null hypothesis when it is in fact false

Decision	True	False
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision



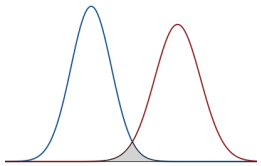
Type I and Type II error

- **Significance level** α: the probability of making a Type I error (rejecting a true null hypothesis)
- β: the probability of making a Type II error
- The **power** of a test against any specific alternative is 1 minus the probability of a Type II error for that alternative
- What is the relationship? The smaller we specify the significance level α the larger the probability of β of not rejecting a false null hypothesis will be

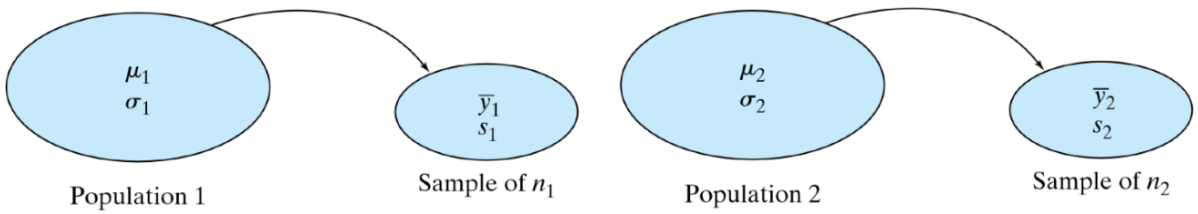
Chapter 18 Notes

Inference for Two Means (Independent Samples)
STP-231

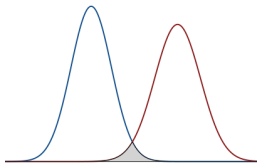
Arizona State University



Conditions for Comparing Two Means



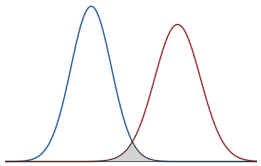
- Say we have two simple random samples
- These are distinct populations that are independent, but we want to study the same quantitative variable of interest
- Assume both parameters are normally distributed, with unknown parameters



Comparing two Means

We compare two populations by:

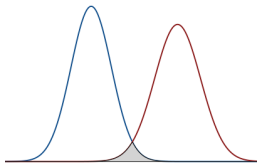
- Confidence intervals on the difference of their means $\mu_1 - \mu_2$
- Hypothesis tests with $H_0 : \mu_1 = \mu_2$ which can also be stated as $H_0 : \mu_1 - \mu_2 = 0$
- To perform inference on $\mu_1 - \mu_2$, we start with the difference $\overline{Y}_1 - \overline{Y}_2$



Sampling Distribution of $\overline{Y}_1 - \overline{Y}_2$

Previously we have seen the sampling distributions of $\overline{Y}_1$ and $\overline{Y}_2$ have:

- means μ_1 and μ_2 respectively
- standard deviations $\frac{\sigma_1}{\sqrt{n_1}}$ and $\frac{\sigma_2}{\sqrt{n_2}}$
- The center of the sampling distribution of $\overline{Y}_1 - \overline{Y}_2$ is $\mu_{\overline{Y}_1 - \overline{Y}_2} = \mu_1 - \mu_2$
- Similarly, the standard deviation of the sampling distribution is $\sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

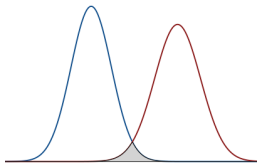


Standard Error of $\overline{Y}_1 - \overline{Y}_2$

- Recall, for the sample mean, the standard error is a measure of precision $\text{SE}_{\overline{Y}} = \frac{S}{\sqrt{n}}$

- Typically the parameters σ_1 and σ_2 will not be known so we estimate them using S_1 and S_2. The standard error of $\overline{Y}_1 - \overline{Y}_2$ is

$$\text{SE}_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\text{SE}_{\overline{Y}_1}^2 + \text{SE}_{\overline{Y}_2}^2}$$



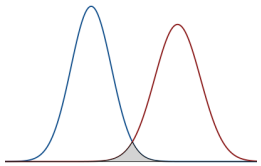
Two Sample t-statistic

- Standardizing the statistic $\overline{Y}_1 - \overline{Y}_2$ we get

$$t^* = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- This statistic approximately follows a t-distribution with degrees of freedom (round down to nearest whole number, hence approximately)

$$\boxed{\text{df} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}} \tag{1}$$



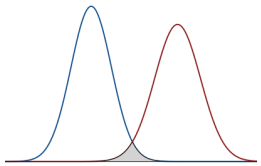
Confidence Interval for $\mu_1 - \mu_2$

- We make the now familiar assumptions of taking data with a simple random sample, the data coming from a true population that is normally distributed or from a large sample, and the samples are independent

- For a confidence level of $1 - \alpha$, a confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

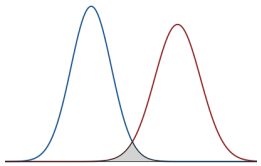
- $t_{\alpha/2}$ is determined from Student's t-distribution with degrees of freedom from equation(1)



Example In-Class

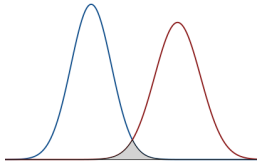
Suppose we survey coffee and non drinkers and measure their typing skills online game. We get a wpm (words per minute) metrics for all the users summarized below:

Had Breakfast	No Breakfast
62	55
81	65
61	107
95	60
69	50
51	61
	52
	68



Example In Class Continued

Calculate the necessary information, n_1, n_2, $\overline{Y}_1$, $\overline{Y}_2$, S_1^2, S_2^2, SE, and df (the degrees of freedom equation from equation(1).



Example In Class Continued

Calculate the necessary information, n_1, n_2, $\overline{Y}_1$, $\overline{Y}_2$, S_1^2, S_2^2, SE, and df (the degrees of freedom equation from equation(1).

$$n_1 = 6$$

$$n_2 = 8$$

$$\overline{Y}_1 = 69.8$$

$$\overline{Y}_2 = 64.7$$

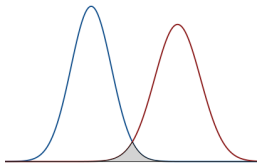
$$S_1^2 = 15.8^2 = 250.6$$

$$S_2^2 = 18.15^2 = 329.6$$

$$\text{SE} = \sqrt{\frac{250.6}{6} + \frac{329.6}{8}} \approx 9.11$$

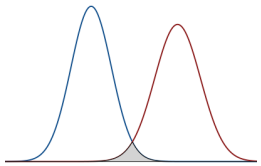
$$\text{df} = \frac{\left(\frac{250.6}{6} + \frac{329.6}{8}\right)^2}{\frac{1}{5}\left(\frac{250.6}{6}\right)^2 + \frac{1}{7}\left(\frac{329.6}{8}\right)^2} \approx 11.6 \downarrow 11$$

Do we think we'll find statistical evidence that the means between the two groups are different? Why or why not?



Example In Class Continued

- Construct a 95% confidence interval for the difference in means.

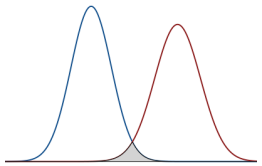


Example In Class Continued

- Construct a 95% confidence interval for the difference in means. Because it is a 2-sided test, we want $t_{\alpha/2,\text{df}=11}$

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha/2,11} \cdot \text{SE} \rightarrow 5.1 \pm (2.2 \cdot 9.11)$$

$$-14 < \mu_1 - \mu_2 < 25.1$$



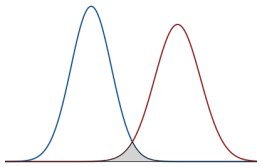
Example In Class Continued

- Construct a 95% confidence interval for the difference in means. Because it is a 2-sided test, we want $t_{\alpha/2,\text{df}=11}$

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha/2,11} \cdot \text{SE} \rightarrow 5.1 \pm (2.2 \cdot 9.11)$$

$$-14 < \mu_1 - \mu_2 < 25.1$$

- Conduct a hypothesis test $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 \neq \mu_2$ at $\alpha = 0.05$ level. Why did the result of the confidence interval tell you what to expect?



Example In Class Continued

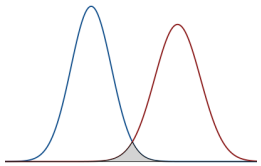
- Construct a 95% confidence interval for the difference in means. Because it is a 2-sided test, we want $t_{\alpha/2,\text{df}=11}$

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha/2,11} \cdot \text{SE} \rightarrow 5.1 \pm (2.2 \cdot 9.11)$$

$$-14 < \mu_1 - \mu_2 < 25.1$$

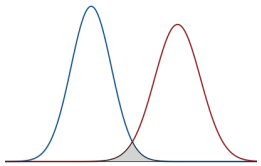
- Conduct a hypothesis test $H_0 : \mu_1 - \mu_2 = 0$ and $H_a : \mu_1 \neq \mu_2$ at $\alpha = 0.05$ level. Why did the result of the confidence interval tell you what to expect? Calculate the $t_* = \dfrac{\left(\overline{Y}_1 - \overline{Y}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$.

  Because our null hypothesis assumes $\mu_1 = \mu_2$, then we have this equal to $t^* = \frac{5.1}{9.11} = 0.56$. We compare this to $\left|t_{0.025,11}\right| \approx 2.2$. Since $t^* < t_{\text{critical}}$, we fail to reject the null, which makes sense because the confidence interval contained 0, the value of our expected hypothesis. If our confidence interval contains the "null-mean", then we are within statistical noise!



Example In Class p-value

- Find the p-value of your result. Does it fall in line with the results from the first two parts?

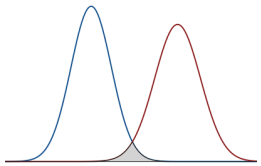


Example In Class p-value

- Find the p-value of your result. Does it fall in line with the results from the first two parts?
  Recall, to find the p-value we want to use our tables to find the probability of obtaining test results at least as extreme as the results actually observed (i.e. t^*), under the assumption that the null hypothesis is correct. Therefore, we wanna see the probability of $t^* = 0.56$, which we can find from our tables or calculators.
  In the calculator (since this is two-sided), we do

  $2*(1-\text{tcdf}(-1E99, |0.56|, 11)) \approx 0.59$ or $2*\text{tcdf}(|0.56|, 1E99, 11) \approx 0.59$

  which is higher than the 0.05 threshold (so we fail to reject). We take the absolute value to ensure we have are on the right side of the t-curve, we subtract from 1 to get the area to the right (intuitively because a really big score would be unlikely), and multiply by 2 because we have a 2-sided test.

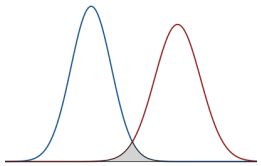


Example 1

Suppose we survey coffee and non drinkers and measure their typing skills online game. We get a wpm (words per minute) metrics for all the users summarized below:

Coffee Drinkers	Non-Coffee Drinkers
$\overline{Y}_1$=63.4 wpm	$\overline{Y}_2$=56.2 wpm
$S_1 = 3.3$ wpm	$S_2 = 5.2$ wpm
$n_1 = 27$	$n_2 = 16$

Compute a 95% CI for $\mu_1 - \mu_2$



Example 1 Solution

Coffee Drinkers	Non-Coffee Drinkers
$\overline{Y}_1$=63.4 wpm	$\overline{Y}_2$=56.2 wpm
$S_1 = 3.3$ wpm	$S_2 = 5.2$ wpm
$n_1 = 27$	$n_2 = 16$

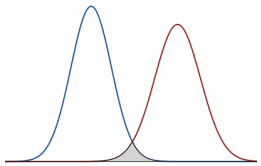
$$\text{df} = \lfloor 22.28 \rfloor = 22 \implies \left| t_{\alpha/2,\text{df}} \right| = 2.074$$

$$\text{SE} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 1.44$$

$$\overline{Y}_1 - \overline{Y}_2 = 7.2$$

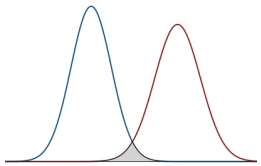
So our confidence interval is:

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \longrightarrow 3.87 < \mu_1 - \mu_2 < 10.53$$



Example 1

- What does this confidence interval tell us?
- Because the confidence interval does not contain 0, what does this tell us about a hypothesis test $\mu_1 = \mu_2$?
- Imagine we change n_1 or n_2 or S_1 etc. How would changing these numbers change our confidence intervals? How would we get wider confidence intervals (i.e. the initial big difference between the groups may be coincidental, not meaningful)



Hypothesis Tests for Two Means

- The null hypothesis is that there is no difference in population means

$$H_0 : \mu_1 = \mu_2 \text{ or } H_0 : \mu_1 - \mu_2 = 0$$

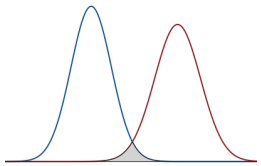
- The alternative can be two tailed meaning there is a difference in means μ_1 and μ_2

$$H_a : \mu_1 \neq \mu_2 \text{ or } H_a : \mu_1 - \mu_2 \neq 0$$

- If one sided:

Right tailed: $\mu_1 > \mu_2$ ($H_a : \mu_1 > \mu_2$ or $H_a : \mu_1 - \mu_2 > 0$)

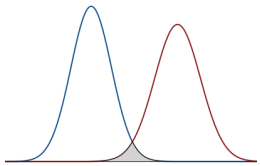
Left tailed: $\mu_1 < \mu_2$ ($H_a : \mu_1 < \mu_2$ or $H_a : \mu_1 - \mu_2 < 0$)



Example 2

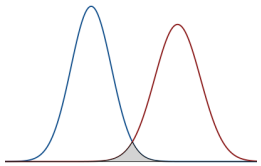
An issue of USA Today discussed the amounts spent by teens and adults at shopping malls. Suppose that we want to perform a hypothesis test to decide whether the mean amount spent by teens is less than the mean amount spent by adults.

- Identify the variable
- Identify the two populations
- Determine the null and alternative hypothesis
- Classify the hypothesis test as two tailed, left tailed, and right tailed



Hypothesis Testing for p-value Approach

- State the hypothesis
- State significance level α
- Compute value of test statistic
- Find the p-value and compare it to α. If it's smaller, reject the null
- State whether you reject H_0 or fail to reject H_0
- Interpret your results within the context of the problem

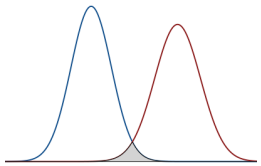


t-test for two Population Means

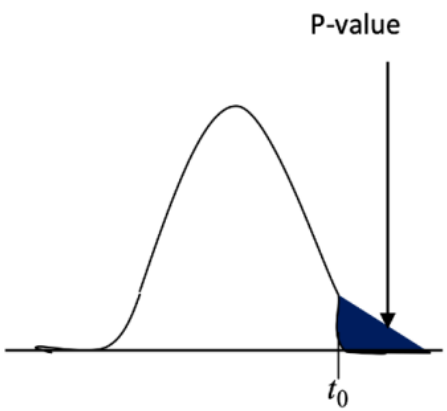
- We make the now familiar assumptions of taking data with a simple random sample, the data coming from a true population that is normally distributed or from a large sample, and the samples are independent

- Our test statistic is:

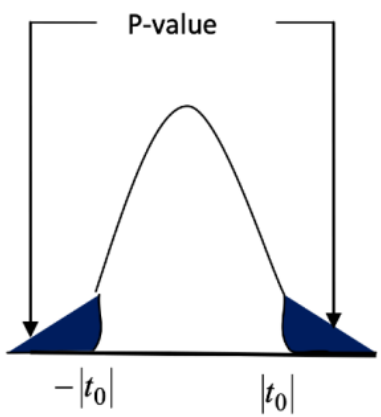
$$t^* = \frac{(\overline{Y}_1 - \overline{Y}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{2}$$



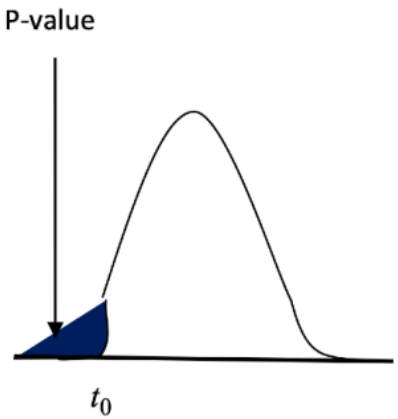
Pictorally



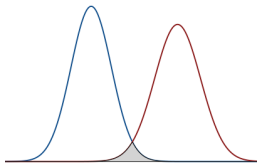
Right tailed Test



Two tailed Test



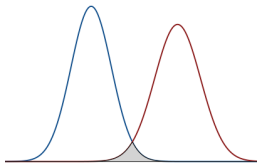
Left tailed Test



Example 3

The seagrass Thalassia testudinum is an integral part of the Texas coastal ecosystem. Essential to the growth of T. Testudinum is ammonium. Researchers K. Lee and K. Dunton of the Marine Science Institute of the University of Texas at Austin noticed that the seagrass beds in Corpus Christi Bay (CCB) were taller and thicker than those in Lower Laguna Madre (LLM). They compared the sediment ammonium concentrations in the two locations and published their findings in Marine Ecology Progress Series. At the 1% significance level, is there sufficient evidence to conclude that the mean sediment ammonium concentration in CCB exceeds that in LLM?

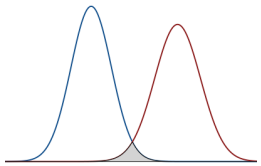
CCB	LLM
$\overline{Y}_1$=115.1	$\overline{Y}_2$=24.3
$S_1 = 79.4$	$S_2 = 10.5$
$n_1 = 51$	$n_2 = 19$



Example 3 Solution

CCB	LLM
$\overline{Y}_1$=115.1	$\overline{Y}_2$=24.3
$S_1 = 79.4$	$S_2 = 10.5$
$n_1 = 51$	$n_2 = 19$

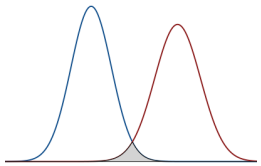
Yes there is evidence, because our t-test statistic from equation(2) is 90.8/11.38=7.982$>t_{\alpha/2} = 2.67$. In fact, such a t-statistic corresponds to a p-value of essentially 0, meaning the chance we see this event by random chance is essentially 0. See the R-code for an easy way to convert tables to tests!



Example 4

A random sample of human vegetarians and omnivores was taken to determine the difference in the daily protein level intake measured in grams between the two. With a significance level of 0.05, determine if the daily mean protein level intake for omnivores and vegetarians is different.

CCB	LLM
$\overline{Y}_1$=49.92 grams	$\overline{Y}_2$=39.04 grams
$S_1 = 18.97$ grams	$S_2 = 18.82$ grams
$n_1 = 53$	$n_2 = 51$



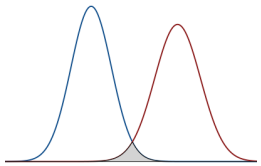
Example 5

A sample of 15 children from New York State showed that the mean time they spend watching television is 28.50 hours per week with a standard deviation of 4 hours. Another sample of 16 children from California showed that the mean time spent by them watching television is 23.25 hours per week with a standard deviation of 5 hours. With a significance level of 0.025, is there any evidence to suggest that the mean time spent watching television by children in New York State is greater than that for children in California? (Assume the times spent watching television by children for both states follow a normal distribution)

Chapter 19

Confidence Intervals for One Population Proportion
STP-231

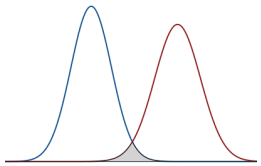
Arizona State University



Proportions

- Dichotomous observations: when only two types of observations exist
- The binomial distribution
- Categories: "success" and "failure"
- We can discuss proportions for these categories
- p, the population proportion
- Sample proportion (a point estimate of p)

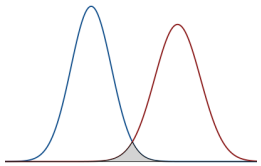
$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{total number of individuals in the sample}}$$



Example 1

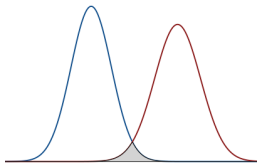
At any given time, soft-drink dispensers may harbor bacteria such as Chrseobacterium meningosepticum that can cause illness. To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with Chryseobacterium meningosepticum. Thus the sample proportion of contaminated dispensers is

$$\hat{p} = \frac{5}{30} = 0.167$$



Wilson Adjusted Sample proportion (plus 4 estimate)

- Rather than using $\hat{p}$ for inference we use $\tilde{p}$ for added accuracy
- $\tilde{p} = \frac{\text{number of successes in the sample} + 2}{n+4}$
- Increases the number of observations with the particular attribute by 2
- Increases the total number of observations by 4
- Explanation



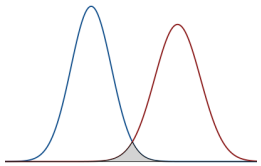
95% Confidence Interval for p

- Standard error

$$\text{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

- Confidence interval

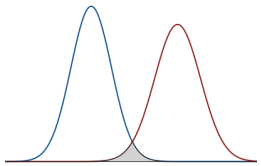
$$\tilde{p} \pm 1.96 \times \text{SE}_{\overline{p}}$$



Example 2

In an experiment with a certain mutation in the fruit fly Drosophila, n individuals were examined; of these 20% were found to be mutants. Determine the standard error of $\tilde{p}$ if

- $n = 100$
- $n = 400$
- For $n = 100$ and $n = 400$ construct a 95% confidence interval for the population proportion of mutants



Example 2

In an experiment with a certain mutation in the fruit fly Drosophila, n individuals were examined; of these 20% were found to be mutants. Determine the standard error of $\tilde{p}$ if

- $n = 100$
- $n = 400$
- For $n = 100$ and $n = 400$ construct a 95% confidence interval for the population proportion of mutants

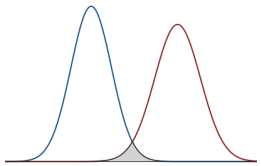
First, note $\tilde{p} = \frac{100*0.2+2}{100+4} = 0.212$

$$0.212 \pm 1.96 \times \sqrt{\frac{0.212 * (1 - 0.212)}{100 + 4}}$$

$$0.212 \pm 1.96 * 0.04$$

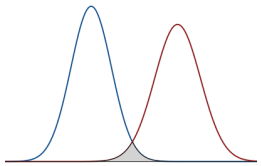
$$0.1336 < p < 0.2906$$

Repeat for $n = 400$



Example 3

In a study of in vitro fertilization, 264 women ages 40-44 underwent a procedure known as elective single embryo transfer (eSET) to attempt to get pregnant. Sixty of these women successfully became pregnant and gave birth. Use these data to construct a 95% confidence interval for the probability of success using eSET for a woman ages 40-44.



Example 3

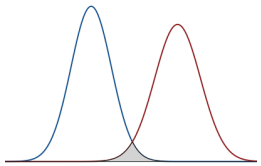
In a study of in vitro fertilization, 264 women ages 40-44 underwent a procedure known as elective single embryo transfer (eSET) to attempt to get pregnant. Sixty of these women successfully became pregnant and gave birth. Use these data to construct a 95% confidence interval for the probability of success using eSET for a woman ages 40-44.

First, note $p = \frac{60}{264} = 0.227$ and $\tilde{p} = \frac{60+2}{264+4} = 0.231$

$$0.231 \pm 1.96 \times \sqrt{\frac{0.231 * (1 - 0.231)}{264 + 4}}$$

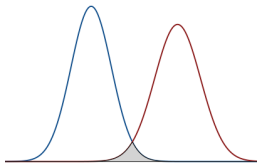
$$0.231 \pm 1.96 * 0.02994$$

$$0.181 < p < 0.281$$



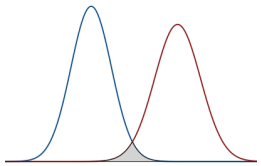
Example 4

Researchers tested patients with cardiac pacemakers to see if use of a cellular telephone interferes with the operation of the pacemaker. There were 959 tests conducted with one type of cellular telephone; interference with the pacemaker (detected with electrocardiographic monitoring) was found in 15.7% of these tests. Use these data to construct an appropriate 95% confidence interval.



Planning a study to Estimate p

- What sample size will be sufficient to achieve a desired degree of precision in estimation of p?
- Use the standard error as our measure of precision
- $$\text{Desired SE} = \sqrt{\frac{\text{Guessed } \tilde{p}(1 - \text{guessed } \tilde{p})}{n+4}}$$



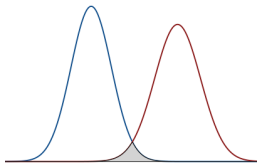
Example 5

In a survey of 136 students at a U.S. college, 19 of them said that they were vegetarians. The sample estimate of the proportion is

$$\tilde{p} = \frac{19+2}{136+4} = 0.15$$

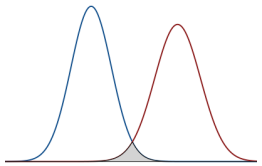
Suppose we regard these data as a pilot and we now wish to plan a study large enough to estimate p with a standard error of two percentage points, that is, 0.02. We choose n to satisfy the following relation:

$$\sqrt{\frac{0.15 * 0.85}{n+4}} \leq 0.02$$



Example 6

In the population of the snail Cepaea, the shells of some individuals have dark bands, while other individuals have unbanded shells. Suppose that a biologist is planning a study to estimate the percentage of banded individuals in a certain natural population and that they want to estimate the percentage-which she anticipates will be in the neighborhood of 60%-with a standard error not to exceed 4 percentage points. How many snails should she plan to collect.



Example 6

In the population of the snail Cepaea, the shells of some individuals have dark bands, while other individuals have unbanded shells. Suppose that a biologist is planning a study to estimate the percentage of banded individuals in a certain natural population and that they want to estimate the percentage-which she anticipates will be in the neighborhood of 60%-with a standard error not to exceed 4 percentage points. How many snails should she plan to collect.

$$\text{Desired SE} = \sqrt{\frac{\text{Guessed } \tilde{p}(1 - \text{guessed } \tilde{p})}{n+4}}$$

The lhs is 0.04, and we guess $\tilde{p}$ as 0.60.

$$0.04^2 = \frac{0.6 * (1 - 0.6)}{n+4}$$

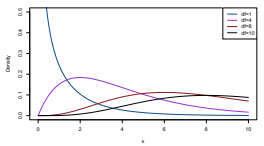
$$n + 4 = \frac{0.24}{0.0016}$$

$$n + 4 = 150$$

Chapter 21

Chi-Square Goodness of Fit
STP-231

Arizona State University

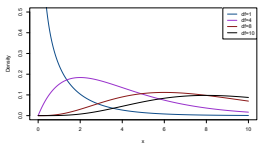


Proportions

- Dichotomous observations: when only two types of observations exist
- The binomial distribution
- Categories: "success" and "failure"
- We can discuss proportions for these categories
- p, the population proportion
- Sample proportion (a point estimate of p)

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{total number of individuals in the sample}}$$

- We are interested on inference on **population** proportion as well
- What if we wanna study all proportions/frequencies of variables with two or more categories



Example 1

A cross between white and yellow summer squash gave progeny of the following colors:

Color	White	Yellow	Green
Number of progeny	155	40	10

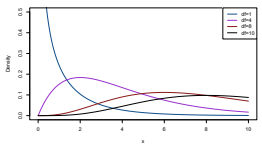
We are interested in if this data is consistent with proportions predicted by a certain genetic model. Those proportions being:

$$p_{\text{white}} = \frac{12}{16} \qquad p_{\text{yellow}} = \frac{3}{16} \qquad p_{\text{green}} = \frac{1}{16}$$

We perform a test where the null hypothesis is that the data is consistent with the proposed proportions

$$H_0 : p_{\text{white}} = \frac{12}{16} \qquad p_{\text{yellow}} = \frac{3}{16} \qquad p_{\text{green}} = \frac{1}{16}$$

H_a : The null hypothesis is false

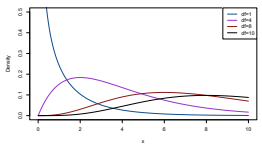


χ^2 distribution

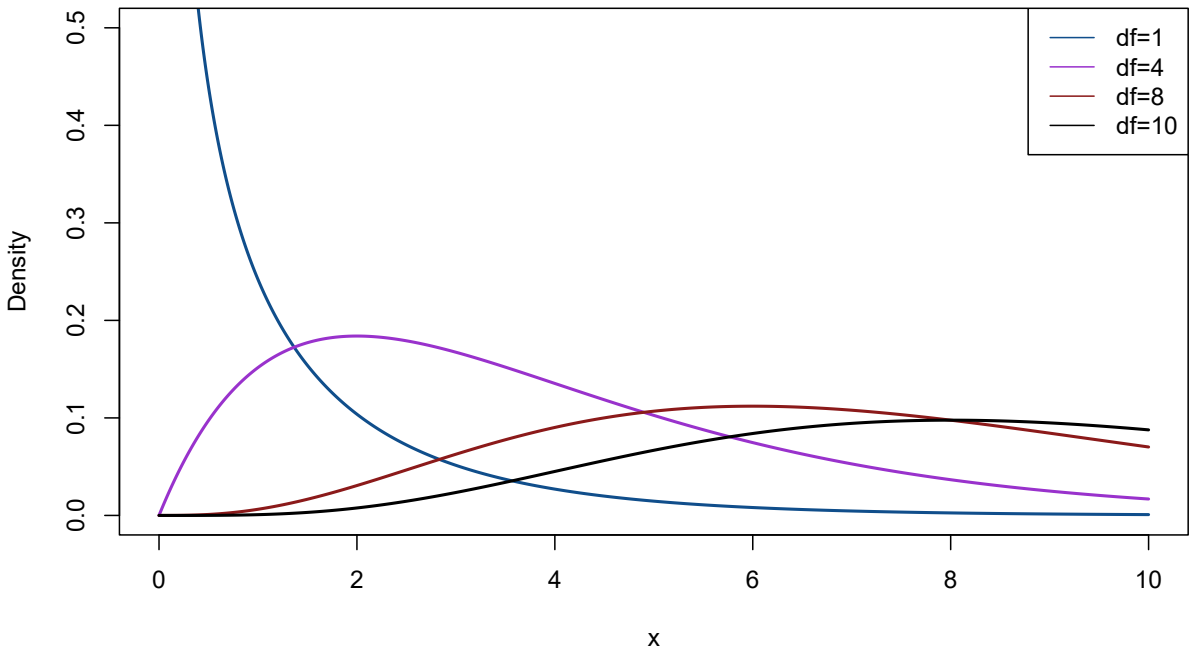
- To execute tests we discuss some properties of a chi-square distribution. A χ^2 distribution with k degrees of freedom is given by:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

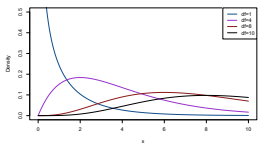
  $\Gamma(k) = (k-1)!$ if integer. O.w. $\Gamma(x) = \frac{1}{x}\Gamma(x+1)$
- Starts out 0. Is NOT symmetric (at least until high degrees of freedom)
- Looks like a normal curve with enough degrees of freedom
- In fact, the form is derived from the normal approximation to the binomial. The square of a normal distribution is a chi-squared with 1 degree of freedom!



Pictorally



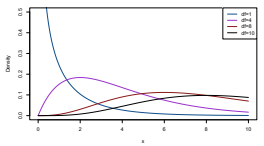
This curve is always positive because we are going to look at a squared statistic (spoilers!)



χ^2 Probabilities and Quantiles

Find the following: (use table or calculator). Form is $\chi^2_{\alpha,\text{df}}$

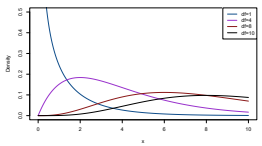
- $\chi^2_{0.05,10}$
- $\chi^2_{0.005,100}$
- $\Pr\left(\chi^2 > 65\right)$, df=50
- $\Pr\left(\chi^2 < 6.1\right)$, df=17



χ^2 Probabilities and Quantiles Solutions

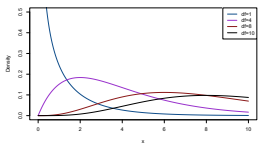
Find the following: (use table or calculator). Form is $\chi^2_{\alpha,\text{df}}$ and we want the area TO THE RIGHT!

- $\chi^2_{0.05,10} = 18.31$ Use the table. If using invχ^2, make sure you plug in 0.05 as calculator defaults area to the left
- $\chi^2_{0.005,100} = 140.2$
- $\Pr\left(\chi^2 > 65\right)$, df=50, and the value is 1-0.925=0.075. Use χ^2cdf(65, 1E99, 50) in the calculator (the blue E above the calculator)
- $\Pr\left(\chi^2 < 6.1\right)$, df=17=0.008 Use χ^2cdf(0, 6.1, 17) in the calculator (the lower bound is now 0)



Observed Freq vs Expected Freq

- If a variable has k possible outcomes, there is a probability (proportion) associated with each $p_1, p_2, \ldots, p_k$
- We test $H_0 : p_1 = p_{1_0}, p_2 = p_{2_0}, \ldots, p_k = p_{k_0}$ vs H_a: the null is false by calculating expected frequencies
- **Observed frequencies**: The frequencies from a sample (O_i). To calculate (if not given) we do total number observed * observed proportion
- **Expected Frequencies**: The frequencies from the proposed model
- Expected frequency for outcome $i = np_{i_0}$. Call this E_i. O_i is $n \cdot p_i$. This is total number observed times the expected proportion. We call this E_i.



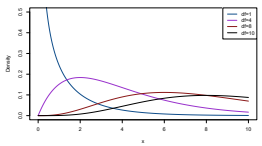
Chi-square Statistic

- The chi-square statistic is a measure of how far the observed counts in a random sample are from the expected counts defined by the null hypothesis. The formula for the statistic is (for k categories)

$$\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i}$$

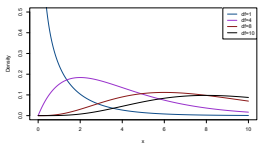
$$\chi^2_* = \sum_{i=1}^{k} \frac{[\text{Observed}_i - \text{Expected}_i]^2}{\text{Expected}_i}$$

$$\chi^2_* = \frac{[\text{observed}_1 - \text{expected}_1]^2}{\text{Expected}_1} + \frac{[\text{observed}_2 - \text{expected}_2]^2}{\text{Expected}_2} + \frac{[\text{observed}_3 - \text{expected}_3]^2}{\text{Expected}_3} + \ldots + \frac{[\text{observed}_k - \text{expected}_k]^2}{\text{Expected}_k}$$



Why we divide by E_i

- We divide by E_i to make sure we take scale into account. For example, if we expect a count of 1000 and see 995, that is a difference of 5 but not super far off from 1000. If we expect 7 and see 12, that is more unexpected, despite being the same difference.

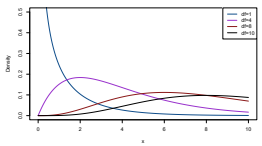


Why we square

Here is a (fake) example of why we square. Suppose we are tracking the number of rainy days in Tempe in 2021 as compared to the usual (gathered from historical weather).

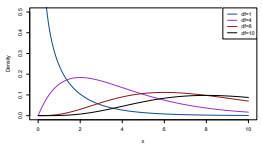
Month	Observed rainy days	Expected Rainy Days	$O_i - E_i$
January	17	7	10
February	1	5	-4
March	2	6	-4
April	3	5	-2

In this case, it's clear that January was a very wet month followed by 3 sorta dry months. This seems abnormal, but the sum of the differences is $10 - 4 - 4 - 2 = 0$, which would indicate business as usual!
Note, we cannot use the chi-square goodness of fit test here because we violate assumptions in slide 14



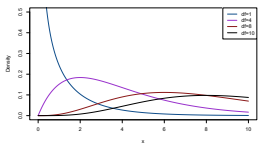
Intuition behind the Statistic

- $\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i}$ is the sum of squared deviances from H_0
- Small χ^2_* would imply there is little difference from the null and observed data
- Large implies the opposite
- Large χ^2_* gives evidence that observed counts are far from expected counts if the model proposed H_0 were true which in turn tell us H_0 is not appropriate



Hypothesis testing for p-value approach

- State the hypotheses
- State the significance level α
- Compute the value of the test statistic
- Find the p-value and compare it to α
- State whether you reject H_0 or fail to reject H_0
- Interpret your results in the context of the problem



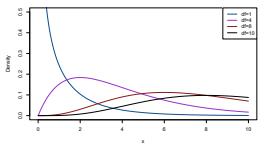
χ^2 Goodness of Fit Test

Assumptions

- All expected frequencies are 1 or greater
- At most 20% of the expected frequencies are less than 5
- Simple random sample
- The test statistic is

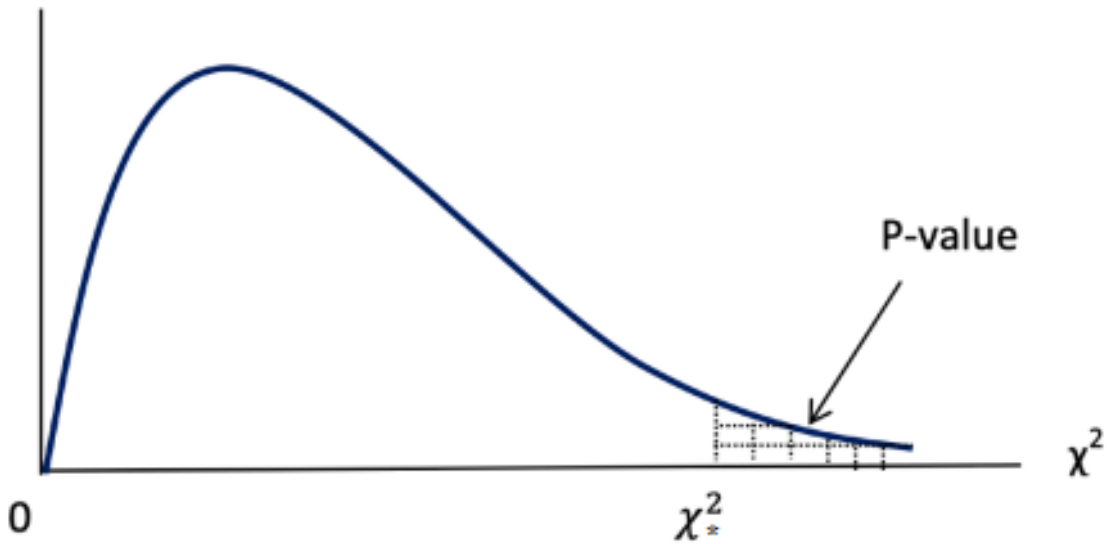
$$\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i} \tag{1}$$

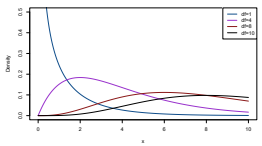
where the degrees of freedom is equal to $k - 1$.



Reject the Null based on p-value

If $p-$value is less than or equal to α, reject the null hypothesis





Example 2

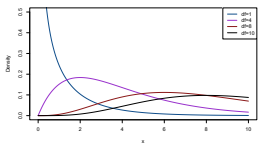
Find the expected frequencies:

Color	White	Yellow	Green
Number of progeny	155	40	10

The proportions from some model are:

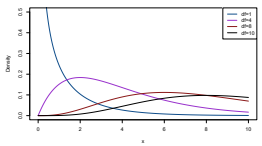
$$p_{\text{white}} = \frac{12}{16} \qquad p_{\text{yellow}} = \frac{3}{16} \qquad p_{\text{green}} = \frac{1}{16}$$

Now, we wanna do a χ^2 goodness of fit test (if not specified assume all the p's are equal to $1/k$)



Example 2

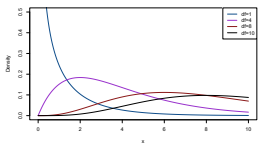
i	O	p	$E = np$	$E - O$	$(E - O)^2$	$(E - O)^2/E$
white 1 Yellow 2 Green 3 k=3, n=205						



Example 2

i	O	p	$E = np$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
white 1	155	12/16	153.75	-1.25	1.56	0.0101
Yellow 2	40	3/16	38.44	-1.56	2.43	0.0632
Green 3	10	1/16	12.81	2.81	7.90	0.6167
k=3, n=205						sum=0.69=χ^2_*

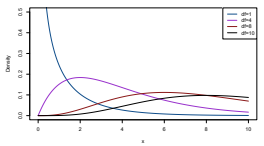
Degrees of freedom=k-1=2. What is the p-value? in R, pchisq(0.6900, 2)=0.292. However, we want area to right, which is 1-0.292=$\boxed{0.708}$. In calculator, we want 1-2nd χ^2cdf(0, 0.69, 2) (or we can do χ^2cdf(0.69, 1E99, 2)), because we want area to right.



Degrees of Freedom

- How could we have more of an effect? Changing the number of white flowered expected minus observed $(E - O)$ would the have the most effect because that is the highest expected value. Yellow would change less and green is entirely dependent on the values of the other two.

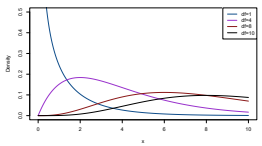
- This is why our degrees of freedom is $k - 1$, because on of the categories is just the "leftovers" of the rest in a sense.



Example 3

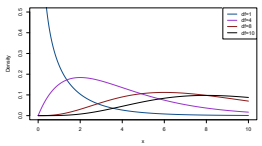
Does fire affect dear behavior? Six months after a fire burned 730 acres of deer habitat, researchers surveyed a 3,000 acre parcel surrounding the area, which they divided into four regions: 1) inner burn, 2) inner edge, 3) outer edge, 4) outer unburned. The null hypothesis is that show no preference to any particular type of burned/unburned habitat (i.e. the deer are randomly distributed across the regions)

Region	Acres	Proportion	Observed
Inner burn	520	0.173	2
Inner edge	210	0.070	12
outer edge	240	0.080	18
outer unburned	2,030	0.677	43
TOTAL	3,000	1.000	75



Example 3

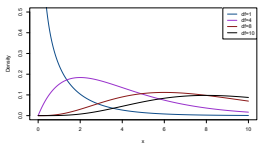
i	O	p	$E = np$	$E - O$	$(E - O)^2$	$(E - O)^2/E$
inner burn						
inner edge						
outer edge						
outer unburned						
n=75, k=4						sum



Example 3

i	O	p	$E = np$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
inner burn	2	0.173	12.975	10.975	120.45	9.283
inner edge	12	0.070	5.25	-6.75	45.56	8.678
outer edge	18	0.080	6	-12	144	24
outer unburned	43	0.677	50.775	7.775	60.45	1.19
n=75, k=4						sum=43.15=χ^2_*

So the χ^2_* statistic is 43.151. Degrees of freedom is 3. Find the p-value:



Example 3

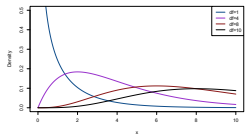
i	O	p	$E = np$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
inner burn	2	0.173	12.975	10.975	120.45	9.283
inner edge	12	0.070	5.25	-6.75	45.56	8.678
outer edge	18	0.080	6	-12	144	24
outer unburned	43	0.677	50.775	7.775	60.45	1.19
n=75, k=4						sum=43.15=χ^2_*

So the χ^2_* statistic is 43.151. Degrees of freedom is 3. Find the p-value: The p-value is $\chi^2\text{cdf}(43.151, 1E99, 4-1) \approx 0$.

Chapter 22

Chi-Square Test of Association
STP-231

Arizona State University

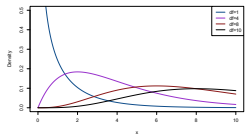


How to use the Table

Note: Column headings are non-directional (omni-directional) *P*-values. If H_A is directional (which is only possible when df = 1), the directional *P*-values are found by dividing the column headings in half.

	TAIL PROBABILITY						
df	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	1.64	2.71	3.84	5.41	6.63	10.83	15.14
2	3.22	4.61	5.99	7.82	9.21	13.82	18.42
3	4.64	6.25	7.81	9.84	11.34	16.27	21.11
4	5.99	7.78	9.49	11.67	13.28	18.47	23.51
5	7.29	9.24	11.07	13.39	15.09	20.51	25.74
6	8.56	10.64	12.59	15.03	16.81	22.46	27.86
7	9.80	12.02	14.07	16.62	18.48	24.32	29.88
8	11.03	13.36	15.51	18.17	20.09	26.12	31.83
9	12.24	14.68	16.92	19.68	21.67	27.88	33.72
10	13.44	15.99	18.31	21.16	23.21	29.59	35.56
11	14.63	17.28	19.68	22.62	24.72	31.26	37.37
12	15.81	18.55	21.03	24.05	26.22	32.91	39.13
13	16.98	19.81	22.36	25.47	27.69	34.53	40.87
14	18.15	21.06	23.68	26.87	29.14	36.12	42.58
15	19.31	22.31	25.00	28.26	30.58	37.70	44.26
16	20.47	23.54	26.30	29.63	32.00	39.25	45.92
17	21.61	24.77	27.59	31.00	33.41	40.79	47.57
18	22.76	25.99	28.87	32.35	34.81	42.31	49.19
19	23.90	27.20	30.14	33.69	36.19	43.82	50.80
20	25.04	28.41	31.41	35.02	37.57	45.31	52.39

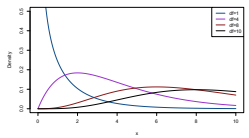
In calculator, to get p-values, use 2nd χ^2cdf(χ^2_*, 1E99, df)



Multiple Categorical Variables

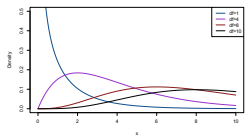
- Contingency tables highlight the dependence or association between variables in the column and the row variables. Since we do not have all of the data from the population, we have to use a sample and apply inferential methods to determine if an association exists between the two variables!

	Drink Coffee		
Better grades	Yes	No	Row totals
Yes	41	15	56
No	8	11	19
Column totals	49	26	$\boxed{n = 75}$



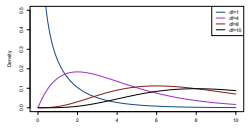
Chi-Square Independence Test Hypotheses

- H_0 : There is no relation between the two variables in question (independent)
- H_a : The two variables in question are associated with each other
- Notice, we don't say what type of relationship they have, just state that there is some association between the two in the alternative
- We create a test based on expected frequencies for each cell



Expected Cell Frequency

- We have observed cell frequencies from sample data (O). We calculate expected cell frequencies assuming that the row variable and column variable are independent.
- Let R_j be the row totals and C_m be the column totals for the specified cells
- Expected cell frequency$=\frac{R_j \times C_m}{n}$, where n is the sample size. We call this E_i. See slide 14 for an example



Chi-Square Statistic

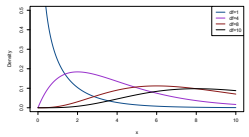
This statistic is a measure of how far the observed counts in each cell are from the expected counts defined by the null hypothesis (if the column and row variable are independent). The formula for the statistic is

$$\chi^2_* = \sum_{i=1}^{r \cdot k} \frac{[O_i - E_i]^2}{E_i}$$

$$\chi^2_* = \sum \frac{[\text{Observed}_i - \text{Expected}_i]^2}{\text{Expected}_i}$$

$$\chi^2_* = \frac{[\text{observed}_1 - \text{expected}_1]^2}{\text{Expected}_1} + \frac{[\text{observed}_2 - \text{expected}_2]^2}{\text{Expected}_2} + \frac{[\text{observed}_3 - \text{expected}_3]^2}{\text{Expected}_3} + \ldots + \frac{[\text{observed}_{rk} - \text{expected}_{rk}]^2}{\text{Expected}_{rk}}$$

which follows a $\chi^2_{(r-1)(k-1)}$ distribution, i.e. df$=(r-1)(k-1)$



Intuition for Cell Expected Frequencies

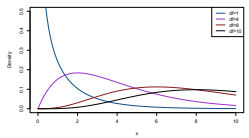
Recall that if events A and B are independent, then $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$. For any cell from a contingency table if H_0 was true:

$$\Pr(\text{row variable and column variable}) = \Pr(\text{row variable}) \times \Pr(\text{column var})$$

$$= \frac{R_j}{n} \times \frac{C_m}{n} = \frac{R_j \cdot C_m}{n^2}$$

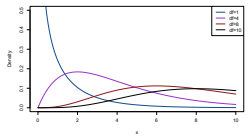
To turn this into a frequency, we would multiply by n. Therefore, the expected frequency if H_0 was true:

$$n \times \frac{R_j \cdot C_m}{n^2} = \frac{R_j \cdot C_m}{n}$$



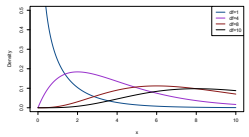
Intuition behind the Statistic

- $\chi^2_* = \sum \frac{[O_i - E_i]^2}{E_i}$ is the sum of squared deviances from H_0
- Small χ^2_* would imply there is little difference from the null and observed data
- Large implies the opposite
- Large χ^2_* gives evidence that observed counts are far from expected counts if the model proposed H_0 were true which in turn tell us H_0 is not appropriate



Hypothesis testing for p-value approach

- State the hypotheses
- State the significance level α
- Compute the value of the test statistic
- Find the p-value and compare it to α
- State whether you reject H_0 or fail to reject H_0
- Interpret your results in the context of the problem



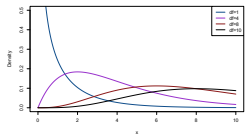
χ^2 Independence Test

Assumptions

- All expected frequencies are 1 or greater
- At most 20% of the expected frequencies are less than 5
- Simple random sample
- The test statistic is

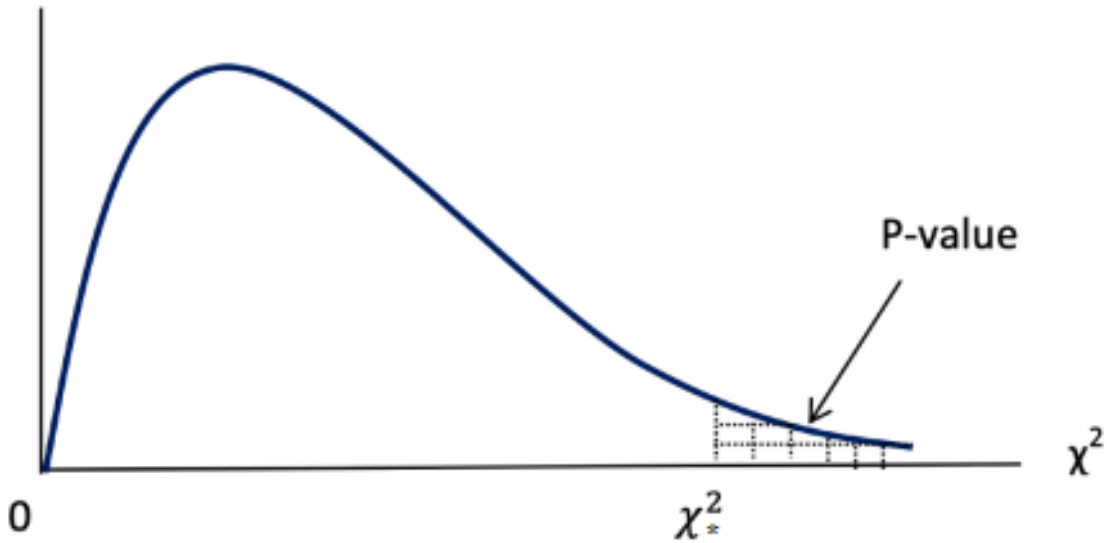
$$\chi^2_* = \sum_{i=1}^{r \cdot k} \frac{[O_i - E_i]^2}{E_i} \tag{1}$$

where the degrees of freedom is equal to $(k-1)(r-1)$. where k is the number of columns and r the number of rows.

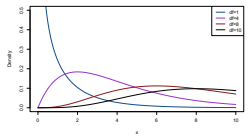


Reject the Null based on p-value

If $p-$value is less than or equal to α, reject the null hypothesis



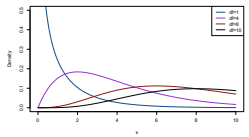
χ^2_* is the test statistics. Use 1 - 2nd distr χ^2cdf(0, χ^2_*, df) or 2nd distr χ^2cdf(χ^2_*, $1E99$, df) in calculator. Use table by finding where you're df are, tracing across the row till you find your test statistic, then going up to see what α you're at (usually given).



Example 1

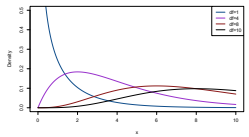
Traditionally red wine goes with red meat, and white wine goes with fish and poultry. A random sample of diners at four-star restaurants was obtained, and each diner was classified according to the food and wine ordered. Is there any evidence that food and wine choice are dependent? Use $\alpha = 0.005$

	Red Wine	White Wine	Row Total
Red Meat	86	46	132
Fish & Poultry	50	64	114
Column Total	136	110	246



Example 1 Continued

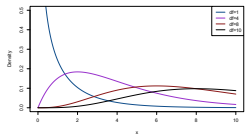
i	O	R_j	C_m	$E = R_j \cdot C_m / n$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
Red meat, red wine Red meat, white win fish, red wine fish, white wine							
n=246, k=2, r=2							sum=



Example 1 Solution

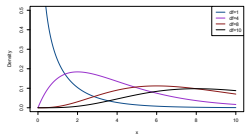
cells	O	R_j	C_m	$E = R_j \cdot C_m/n$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
Red meat, red wine	86	132	136	72.98	13.02	169.5	2.32
Red meat, white wine	46	132	110	59.02	-13.02	169.63	2.87
fish, red wine	50	114	136	63.02	-13.02	169.63	2.69
fish, white wine	64	114	110	50.98	13.02	169.63	3.32
n=246, k=2, r=2							sum=11.21=χ^2_*

Use the p-value approach. Compare pvalue of χ^2 for 11.21 with degrees of freedom (2-1)*(2-1)=1 to 0.005. Use table or calculator χ^2cdf(11.21,1E99, 1)$\approx$ 0.0008, so we reject our null.



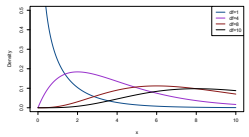
Example 2

It has been suspected that prolonged use of a cellular telephone increases the chance of developing brain cancer due to the microwave-frequency signal that is transmitted by the cell phone. According to this theory, if a cell phone is repeatedly held near one side of the head, the brain tumors are more likely to develop on that side of the head. To investigate this, a group of patients were studied who had used cell phone for a least six months prior to developing brain tumors. The patients were asked whether they routinely held the cell phone to a certain ear and, if so, which ear. The 88 responses (from those who preferred one side over the other) are shown in the following table. Do the data provide sufficient evidence to conclude that the use of cellular telephones leads to an increase in brain tumors on that side of the head? Use a chi-square test to test if there is an association between what side of the head a phone is held and where the brain tumor is located. Let $\alpha = 0.05$



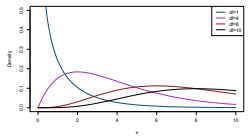
Example 2 Continued

	Phone side		
Tumor side	Left	Right	Row total
Left	14	28	42
Right	19	27	46
Column Total	33	55	$\boxed{n = 88}$



Example 2 Continued

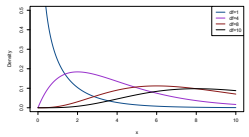
i	O	R_j	C_m	$E = R_j \cdot C_m/n$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
Tumor Left, phone left	14	42	33				
tumor left, phone right	28	42	55				
tumor right, phone left	19	46	33				
tumor right, phone right	27	46	55				
n=88, k=2, r=2							sum=



Example 2 Solution

i	O	R_j	C_m	$E = R_j \cdot C_m/n$	$E - O$	$(E-O)^2$	$(E-O)^2/E$
Tumor Left, phone left	14	42	33	15.75	1.75	3.0625	0.194
tumor left, phone right	28	42	55	26.25	-1.75	3.0625	0.117
tumor right, phone left	19	46	33	17.25	-1.75	3.0625	0.178
tumor right, phone right	27	46	55	28.75	1.75	3.0625	0.107
n=88, k=2, r=2							sum=0.596=χ^2_*

Now, as usual, we compare to the χ^2_{critical} (or calculate a p-value) in our calculator. χ^2cdf(0.596, $1E99$, df=(2-1)*(2-1))=0.440. The critical value, from the tables, is at the 0.05 level, 3.84.



How to use the Table

Note: Column headings are non-directional (omni-directional) *P*-values. If H_A is directional (which is only possible when df = 1), the directional *P*-values are found by dividing the column headings in half.

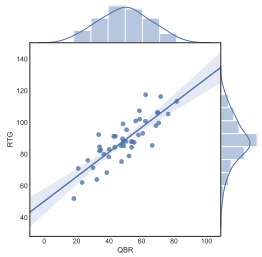
	TAIL PROBABILITY						
df	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	1.64	2.71	3.84	5.41	6.63	10.83	15.14
2	3.22	4.61	5.99	7.82	9.21	13.82	18.42
3	4.64	6.25	7.81	9.84	11.34	16.27	21.11
4	5.99	7.78	9.49	11.67	13.28	18.47	23.51
5	7.29	9.24	11.07	13.39	15.09	20.51	25.74
6	8.56	10.64	12.59	15.03	16.81	22.46	27.86
7	9.80	12.02	14.07	16.62	18.48	24.32	29.88
8	11.03	13.36	15.51	18.17	20.09	26.12	31.83
9	12.24	14.68	16.92	19.68	21.67	27.88	33.72
10	13.44	15.99	18.31	21.16	23.21	29.59	35.56
11	14.63	17.28	19.68	22.62	24.72	31.26	37.37
12	15.81	18.55	21.03	24.05	26.22	32.91	39.13
13	16.98	19.81	22.36	25.47	27.69	34.53	40.87
14	18.15	21.06	23.68	26.87	29.14	36.12	42.58
15	19.31	22.31	25.00	28.26	30.58	37.70	44.26
16	20.47	23.54	26.30	29.63	32.00	39.25	45.92
17	21.61	24.77	27.59	31.00	33.41	40.79	47.57
18	22.76	25.99	28.87	32.35	34.81	42.31	49.19
19	23.90	27.20	30.14	33.69	36.19	43.82	50.80
20	25.04	28.41	31.41	35.02	37.57	45.31	52.39

In calculator, to get p-values, use 2nd χ^2cdf(χ^2_*, 1E99, df)

Chapter 4 Notes

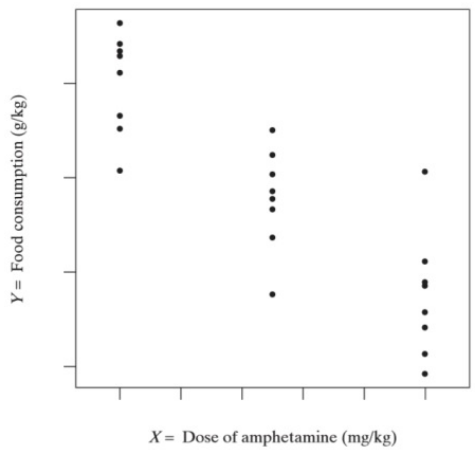
Bivariate Data, Linear Correlation, and Regression
STP-231

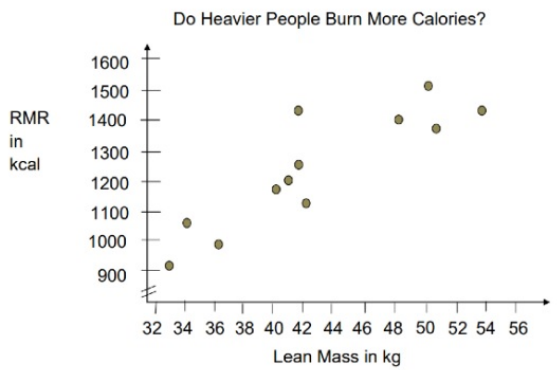
Arizona State University

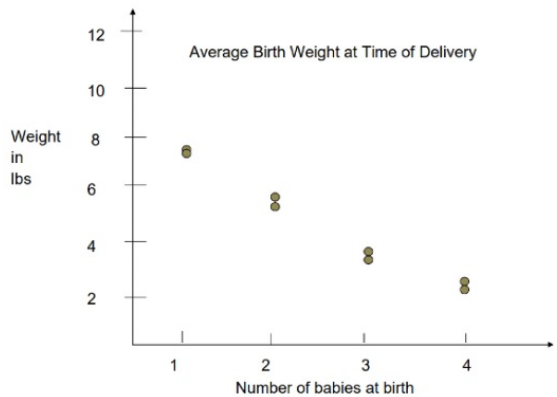


Scatter Plot
Key Terms

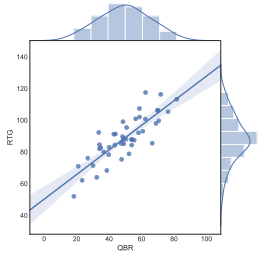
- A **Scatter plot** is a plot all of the data points and it used to help determine if there is any linear relationship between the two variables and if there are any anamolies.







The Linear Equation



Linear Equation

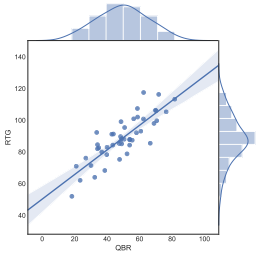
- The general form of a linear equation with independent variable can be written as:

$$y = b_0 + b_1 x$$

- Where x is the independent (explanatory) variable, y is the dependent (response) variable, b_0 is the $y-$intercept, and b_1 is the slope
- **slope**: Change in the response variable for every unit increase in the explanatory variable

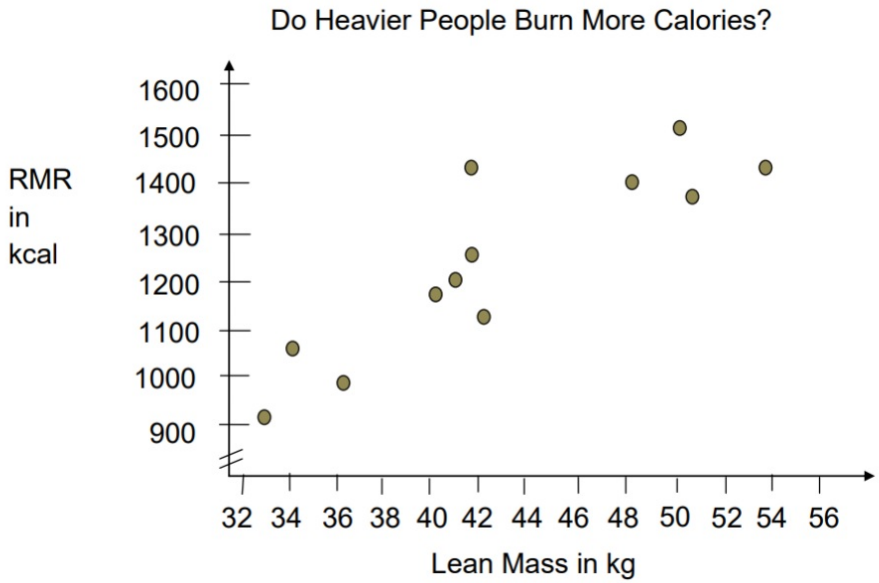
$$b_1 \text{ units} = \frac{\text{units of the response y}}{\text{units of the independent variable x}}$$

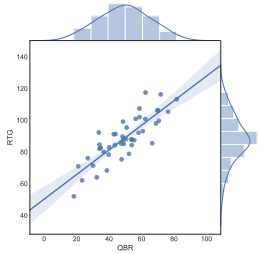
- **Y – intercept**: Value of the response variable when the value of the explanatory variable is 0. The sign of b_1 holds basic information about the linear relationships between x and y.



Positive Slope

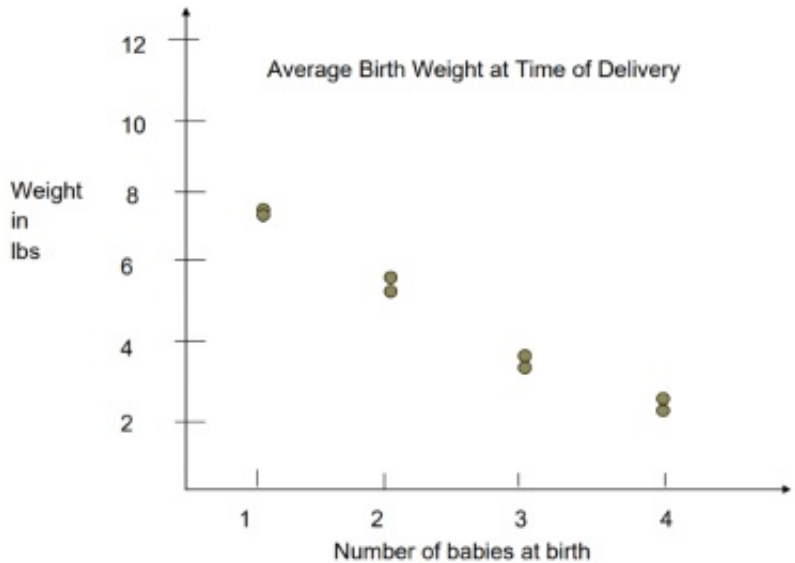
- If $b_1 > 0$
  - As the independent variable (x) increases the dependent variable (y) increases
  - If the slope of the line appears positive, then the correlation is positive

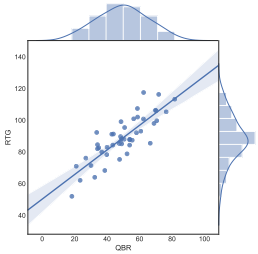




Negative Slope

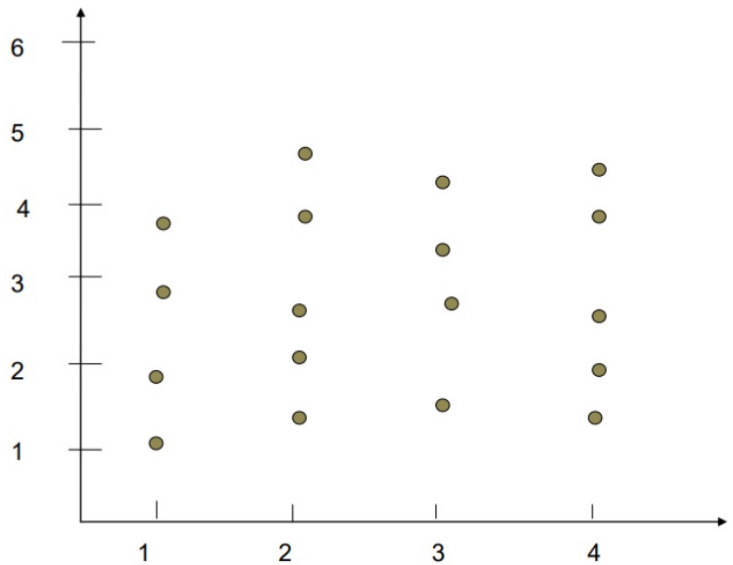
- If $b_1 < 0$
  - As the independent variable (x) decreases the dependent variable (y) decreases
  - If the slope of the line appears negative, then the correlation is negative

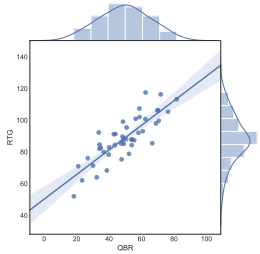




No Slope

- If $b_1 = 0$
  - There is not a linear relationship between the 1 independent variable (x) and the dependent variable (y)
  - There is also not a correlation between the x and y variables



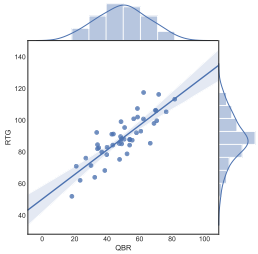


Correlation

- Contains information of the direction (positive or negative) and strength (weak, moderate, or strong) of linear association
- unitless
- How you assign explanatory and response variables does not affect the correlation
- **Linear Correlation Coefficient**

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x s_y}$$

- s_x and s_y denote the sample standard deviations of x and y respectively



Example 1 Calculation

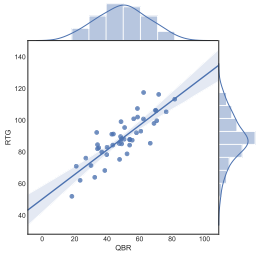
- We measure length and weight of 9 snakes[1]

Length X (cm)	Weight Y (g)
60	136
69	198
66	194
64	140
54	93
67	172
59	116
65	174
63	145

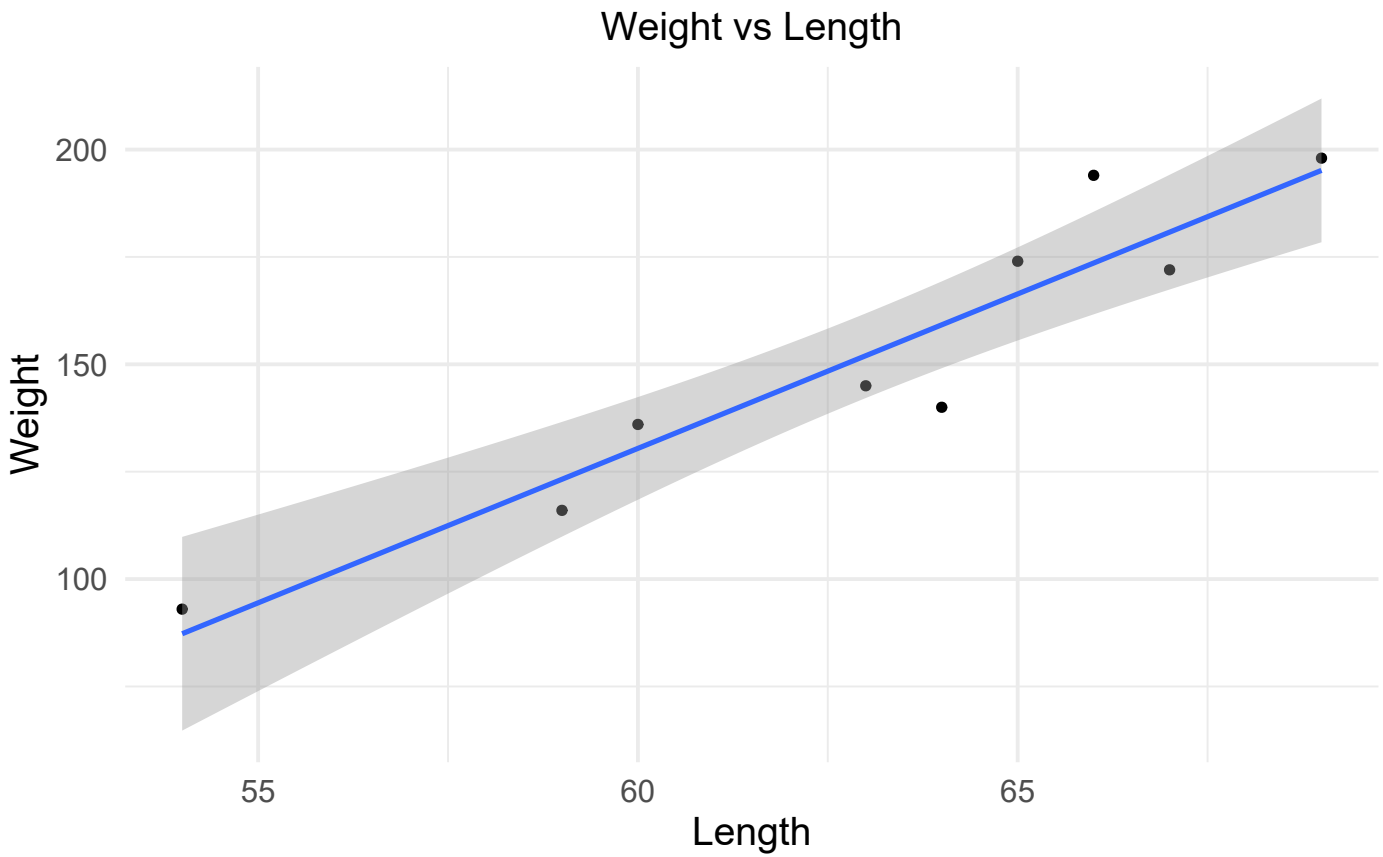
- $s_x = 4.6368$, $s_y = 35.3376$, and $r = 0.9437$. Verify hand calculations in R:

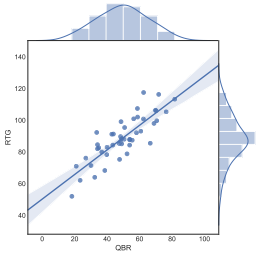
```
X=c(60,69,66,64,54, 67, 59, 65, 63)
Y=c(136,198, 194, 140, 93, 172, 116, 174,
    145)
cor(X, Y)
```

[1]Samuels, Myra L. et al. Statistics for Life Sciences. Pearson, 2016.



Example 1 Continued



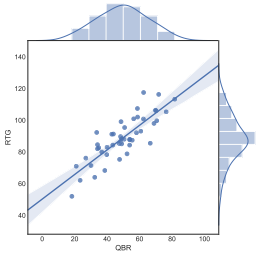


Interpreting Correlation

From definition, $-1 \leq r \leq 1$ The general rule of thumb:

- $0.8 < r < 1$: Strong, positive relationship
- $0.4 < r < 0.8$: Moderate, positive relationship
- $0 < r < 0.4$: Weak, positive relationship
- $-0.4 < r < 0$: Weak, negative relationship
- $-0.8 < r < -0.4$: Moderate, negative relationship
- $-1 < r - 0.8$: Strong, negative relationship

However, this can be misleading



Different Correlations

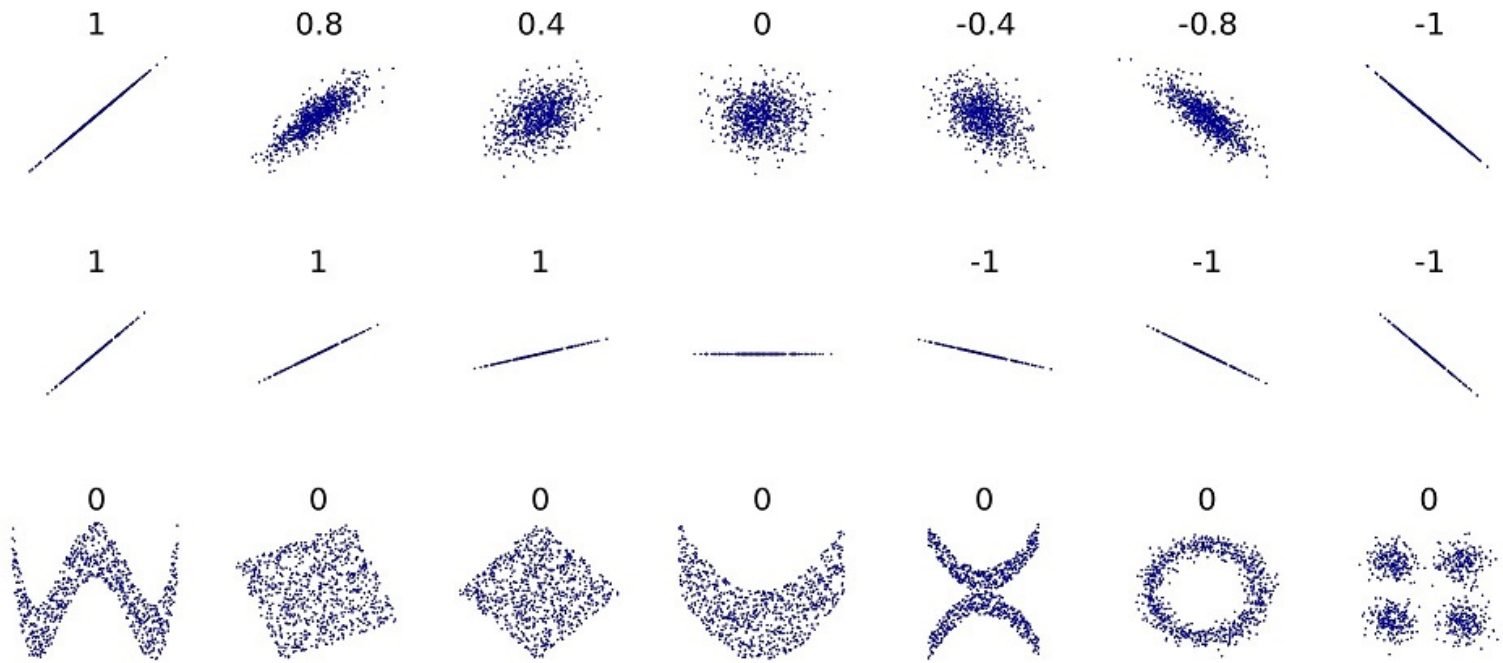
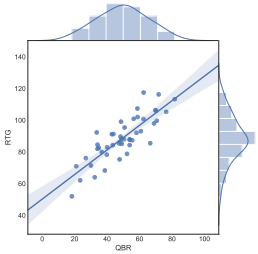
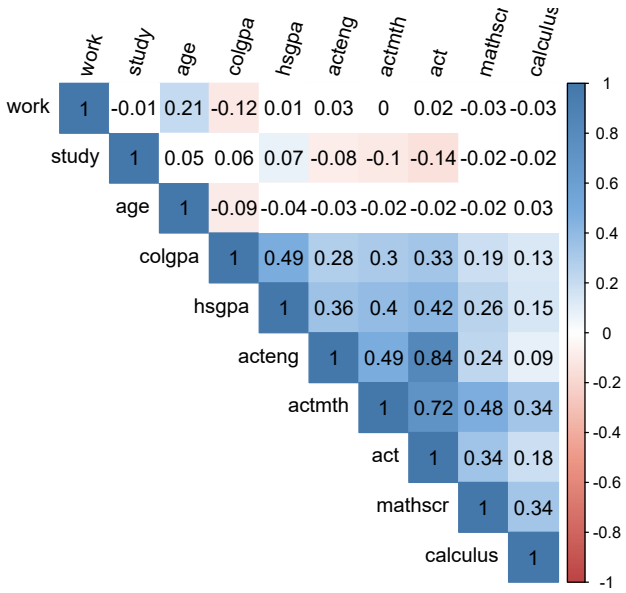
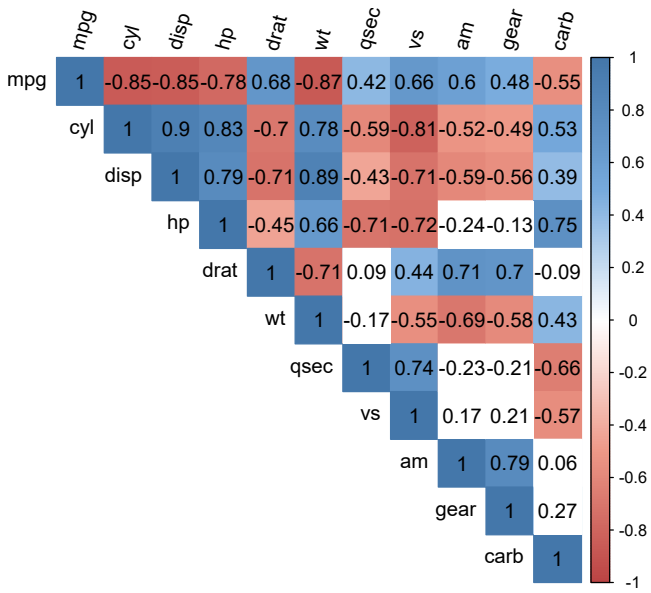


Figure from Wikipedia

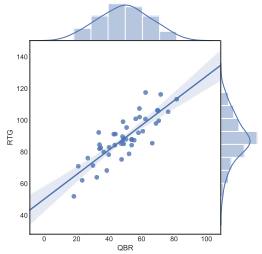


Correlations between variables



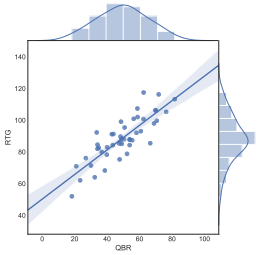


- On the left, correlation between different measurements of econ students from wooldridge::econmath in R. On the right, is comparison of correlations between various attributes of cars from mtcars in R.



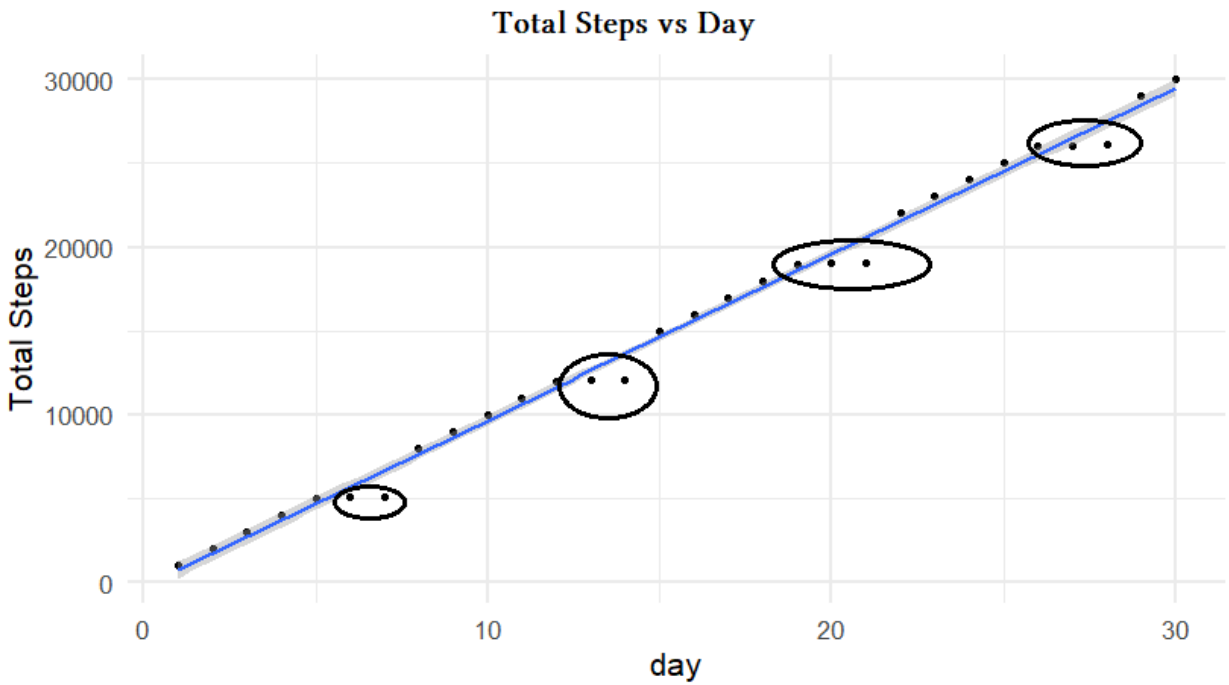
Coefficient of Determination

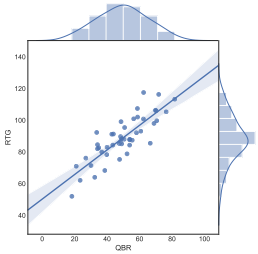
- Denoted by r^2, where $0 \leq r^2 \leq 1$.
- Determines the percentage of variation in the observed values of the repsonse variable that is explained by the regression line
- Interpretation: "r^2 of the variability in the response variable can be explained by the explanatory variable"
- No tried and true way of telling if you have a "good" r^2.
- The value of the correlation squared



Interpreting Correlation and r^2 values
Word of caution

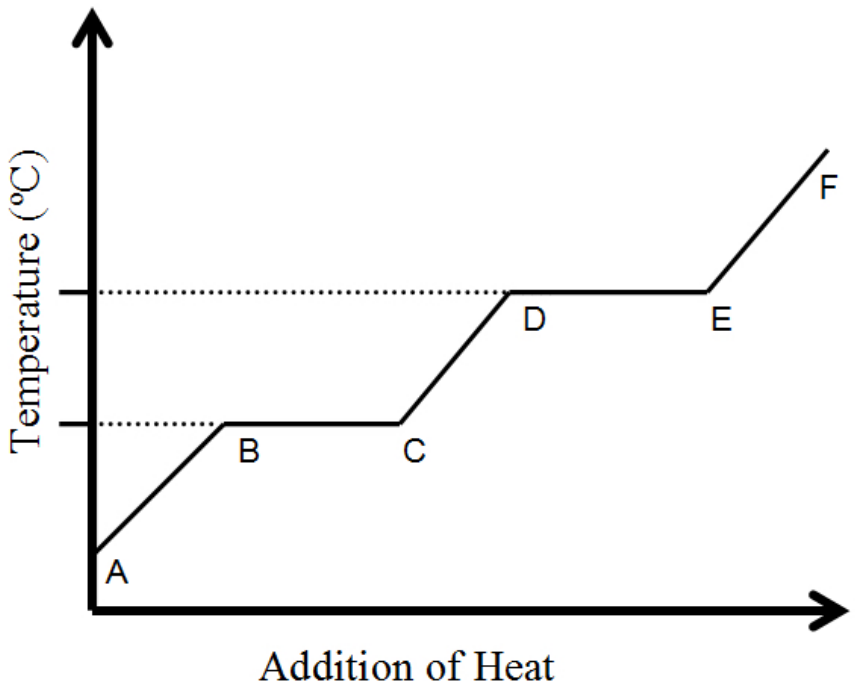
- Kept track of total steps for a day. Monday-Friday, took approximately 1000 steps, then only about 100 a day on Sat. and Sun.
- Correlation of 0.997, r^2 of 0.994, because the weekdays dominate, even the trend is clearly NOT true on weekends.

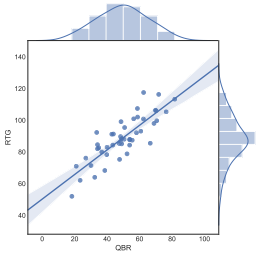




Example 2

The previous example is similar to the following picture of phase diagrams (thanks Avery!)



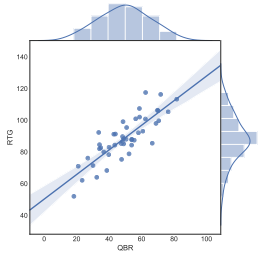


Interpreting r^2
Misleading

- Say we are measuring the speed of a marble falling over time through air and have this data:

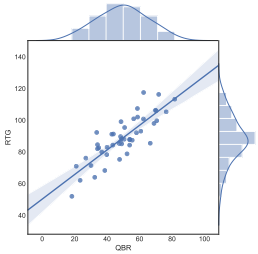
Time (s)	Speed (m/s)
0	0
1	9
2	21
3	28
4	36
5	49

- This has an r^2 of 0.988. Pretty good...BUT this should be a perfect 1 since we've known this relationship for hundreds and years. In fact, this data would estimate Earth's gravity, g, at $9.5\,\mathrm{m/s^2}$, well off the 9.81 that is well established!



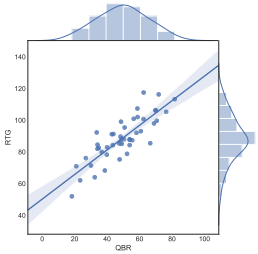
Where these Summary Statistics go wrong
Some places we can err

- We have shown two examples where the correlation and r^2 are misleading
- If the linear equation, $y = b_0 + b_1 x$ is **deceptively** correct, like the walking count example
- If the linear equation is fully right, being off even by a little could be consequential and you can be fooled into thinking you performed a good experiment, when in fact you may receive a low grade on your physics lab!



Correlation != Causation

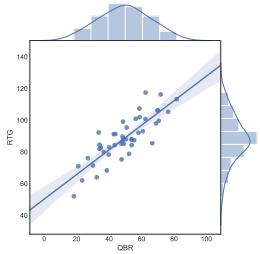
- Cool website: Spurious correlations Some correlations are completely random.
- Note, on that site, the two lines would normally be plotted against one another on a Y-X plane. Try it yourself!
- However, sometimes there is real causation. A seperate field of statistics we will not be covering



HOWEVER!

- Just because correlation doesn't imply causation...does not mean causation doesn't exist!
- Some things do cause other things...it's just a hard problem to solve!
- A meme here would be nice

Regression



A Hypothesis Test on ρ

- ρ is the population correlation coefficient. We assume simple random samples and normal distributions
- Hypotheses:

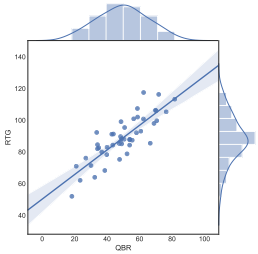
  $H_0 : \rho = 0$ x and y are uncorrelated in the population

  $H_a : \rho \neq 0$ x and y are correlated in the population

- First specify the level you want, α (usually $\alpha = 0.05$).
- The test statistic is (t-distribution with df=n-2)

$$t^* = r\sqrt{\frac{n-2}{1-r^2}}$$

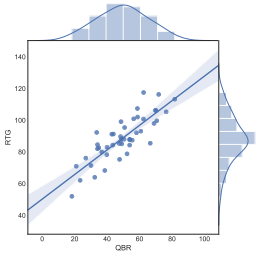
- Reject if p-value is $\leq \alpha$, p-value is equal to 2*tcdf($|t^*|$, 1E99, n-2), or if using table, find 2*(1-area of $|t^*|$)
- Caveats: As n gets large, the test statistic can get very big.



Example 3

	X	Y
	6	6
	1	7
	3	3
	2	2
	5	14
Mean	3.4	6.4
SD	2.1	4.7

- Plot the data. Does there appear to be a relationship between X and Y? Is it linear or nonlinear? Weak or strong?
- Compute the sample correlation coefficient between X and Y.
- Is there significant evidence that X and Y are correlated? Conduce a test using $\alpha = 0.05$. Use your tables.

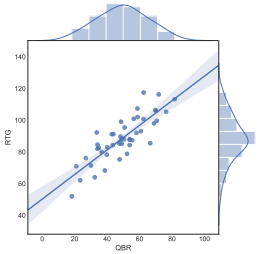


Last part continued

- We actually fail to reject the null partly because the n value is so small

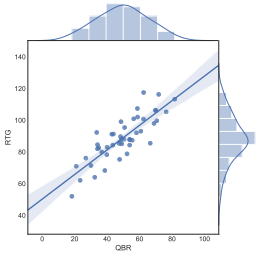
```
X=c(6,1,3,2,5)
Y=c(6,7,3,2,14)
n=length(X)
rr=cor(XX,YY)
alpha=0.05
tstat=rr*sqrt((n-2)/(1-rr^2));2*(1-pt(tstat,  ))<
    alpha
```

Use the equation $t^* = r\sqrt{\frac{n-2}{1-r^2}}$, where $r = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y}$



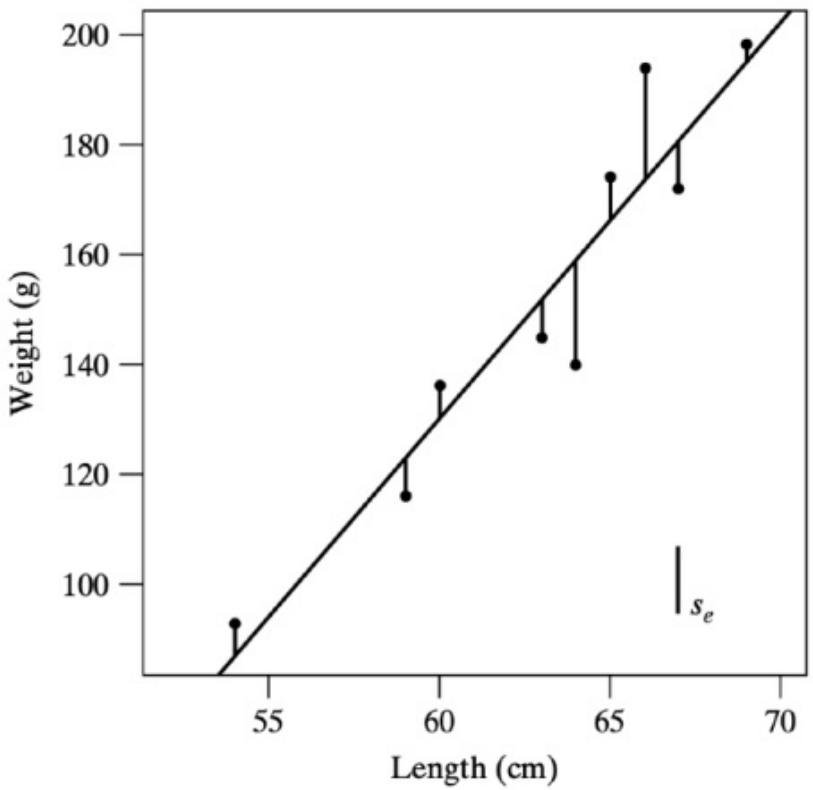
Regression Line

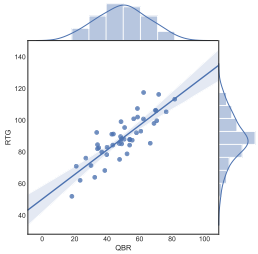
- A straight line that summarize the linear relationship between two variables when one of the variables is thought to help to explain or predict the other
- Shows how the response variable y changes on average as the explanatory variable x changes
- We can predict expected value of y for given value of x



Least Squares Regression Line

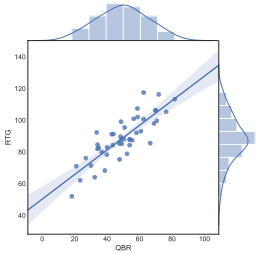
- The line that makes the sum of the squared vertical distances of the data points from the line as small as possible. Note, this line is with respect to y-axis, not perpendicular to the line itself!





Equation of Least Squares Regression Line

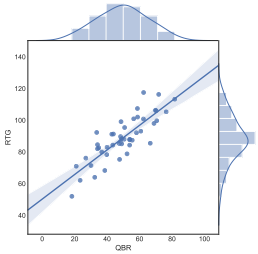
- If we have data on explanatory variable x and response y then the least squares regression line can be found with $\overline{x}$, $\overline{y}$, s_x, s_y, and r. The equation is
- $\hat{y} = b_0 + b_1 x$
- Slope:$b_1 = r\frac{s_y}{s_x} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
- Intercept: $b_0 = \overline{y} - b_1\overline{x} = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$
- Notice: The line that is determine will pass through $(\overline{x}, \overline{y})$. $\hat{y}$ is a predicted value for each x, the observed values will not exactly be $\hat{y}$.



Example 4

The lean body mass in kilograms and resting metabolic rate in kilocalories for 12 subjects who are in a study on dieting was recorded. Lean body mass is the weight without the fat. The resting metabolic rate (RMR) was measured in kilocalories (Kcal) over a 24 hour period.

$$\overline{x} = 43.0333 \text{ kg}$$
$$s_x = 6.8684 \text{ kg}$$
$$\overline{y} = 1235.08333 \text{ Kcal}$$
$$s_y = 188.28289 \text{ Kcal}$$
$$r = 0.08765$$



Residuals

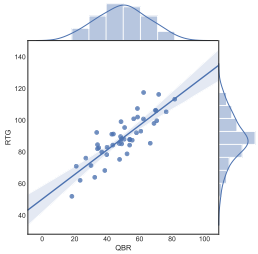
- The distance from a predicted response and observed response

$$e_i = y_i - \hat{y}_i$$

- Residuals sum of squares:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- If the value of the residual sum of square is small the data will fall close to the regression line
- The "best straight line" is the one that minimizes the residual sum of squares

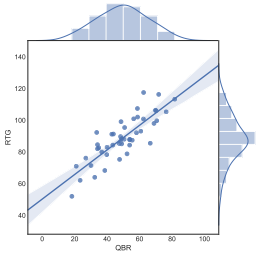


Residual Standard Deviation

- Another thing that we include in our regression analysis summarization is a measure of how close the actual data points are in relation to the fitted line:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

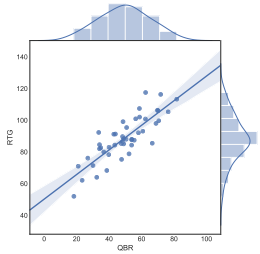
- The measurement tells us how far above or below the regression line the data points tend to be



Empirical Rule

We can also tie the residual standard deviation back to the Empirical Rule. If somewhat normal (Bell shaped) data, then we expect:

- 68% of the observations (y-values) to be within ± 1 residual standard deviation of the regression line
- 95% of the observations to be within ± 2 residual standard deviation of the regression line
- Over 99% of the observations to be within ± 3 residual standard deviation of the regression line



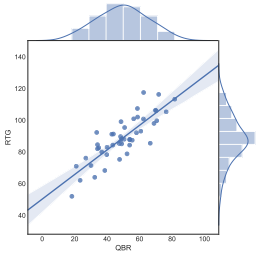
Correlation and Residual Standard Deviation

- r is the correlation coefficient and describes how closely the linear relationship between the X and Y variables is. Related to the slope of the regression line

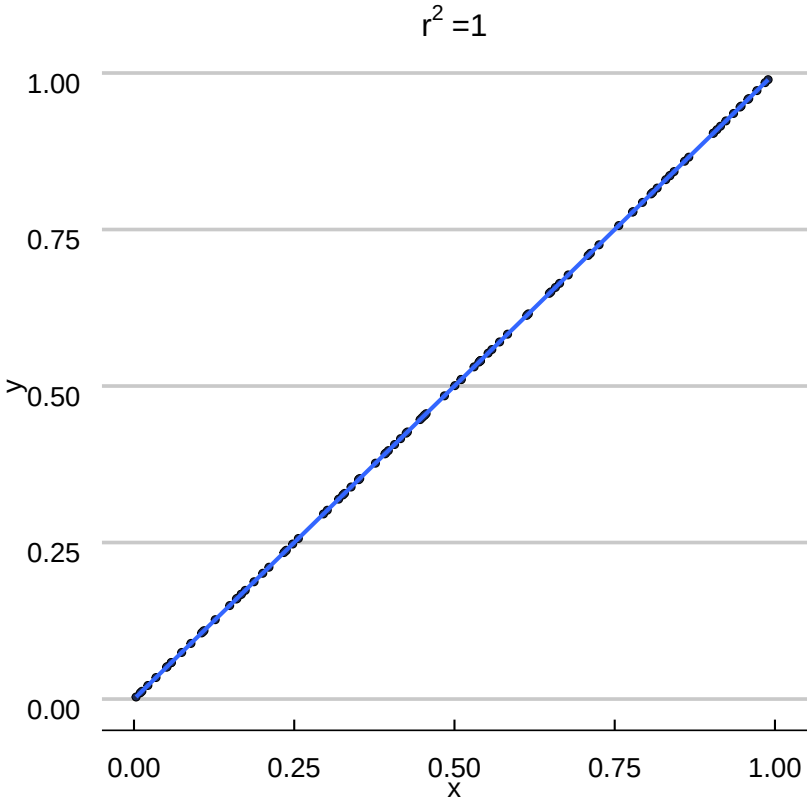
- r^2 is the coefficient of determination and describes the proportion of the variance in Y that is explained by the linear relationship between Y and X

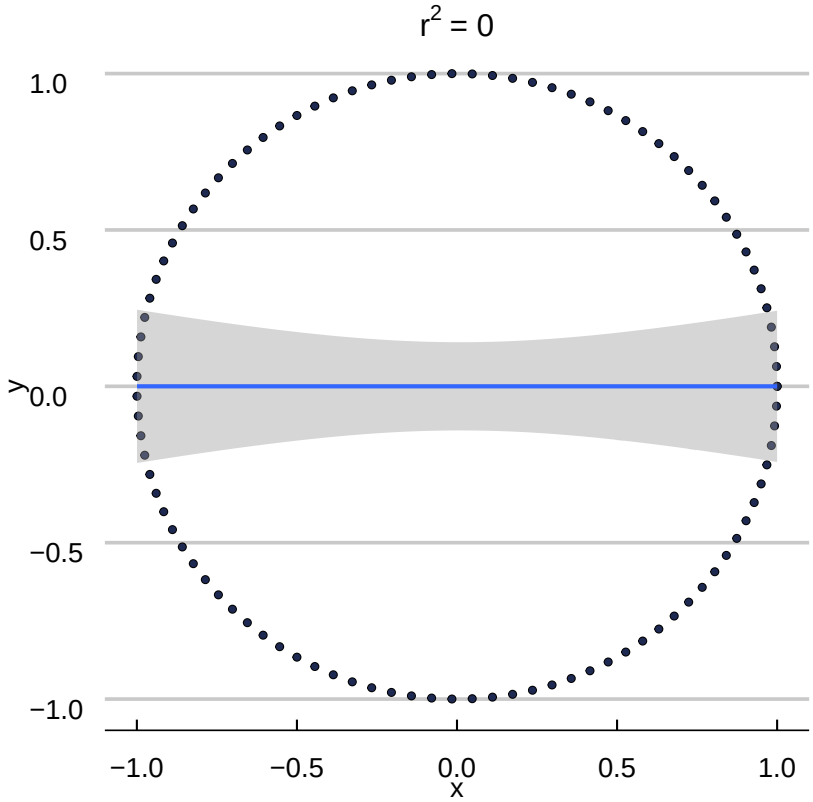
$$r^2 = \frac{\sum(y_i - \overline{y})^2 - \sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2} \approx \frac{s_y^2 - s_e^2}{s_y^2}$$

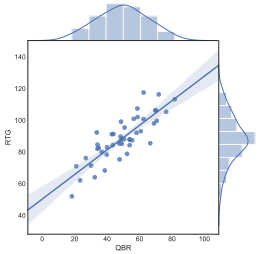
When $s_y^2 = s_e^2$, it means the straight line fit is just the mean of y, i.e. just a flat line horizontally.



Perfect fit is a straight line, 0 is the average of y!

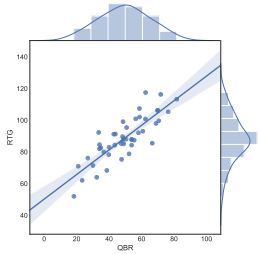






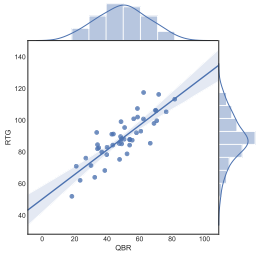
Correlation and Residual Standard Deviation Continued

- If s_e is very small, then $\frac{s_e^2}{s_y^2}$ will be very close to 0, and r^2 will be close to 1
- If s_e is very close to s_y, then $\frac{s_e^2}{s_y^2}$ will be very close to 1, and r^2 will be close to 0
- Typically, we want r^2 close to 1, though as we have seen that's not sufficient to mean you've done a "good" analysis



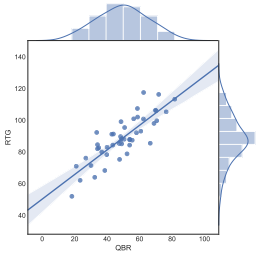
Additional Cautions about Regression

- What if the relationships are not linear? (this is common)
- The correlation statistic and least squares are both not resistant to outliers. Why we always plot first
- If we removed some data points and it greatly changes our results, this shows the methodology may be limited
- We want to avoid extrapolation. The the use of regression for prediction well beyond whats observed is very uncertain
- Lurking variables (aka confounding variables). A lurking variable is a variable that is not one of the explanatory or response variables in a study, yet may influence the interpretation of relationships of those variables



Lurking Variables
Example 5

Exploring relation between drinking orange juice and health, but how much you exercise could be a lurking variable



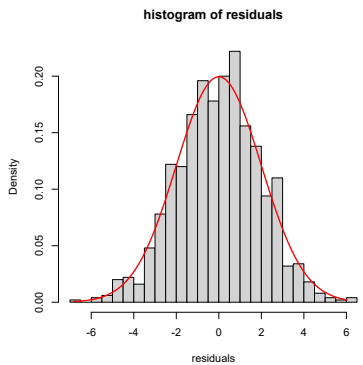
Simple Linear Regression

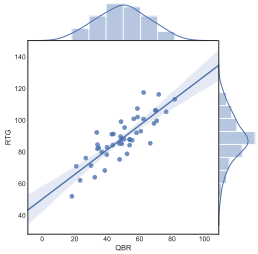
- The model is

$$y = \underbrace{b_0 + b_1 x}_{\hat{y}} + \underbrace{e}_{\text{residuals}}$$

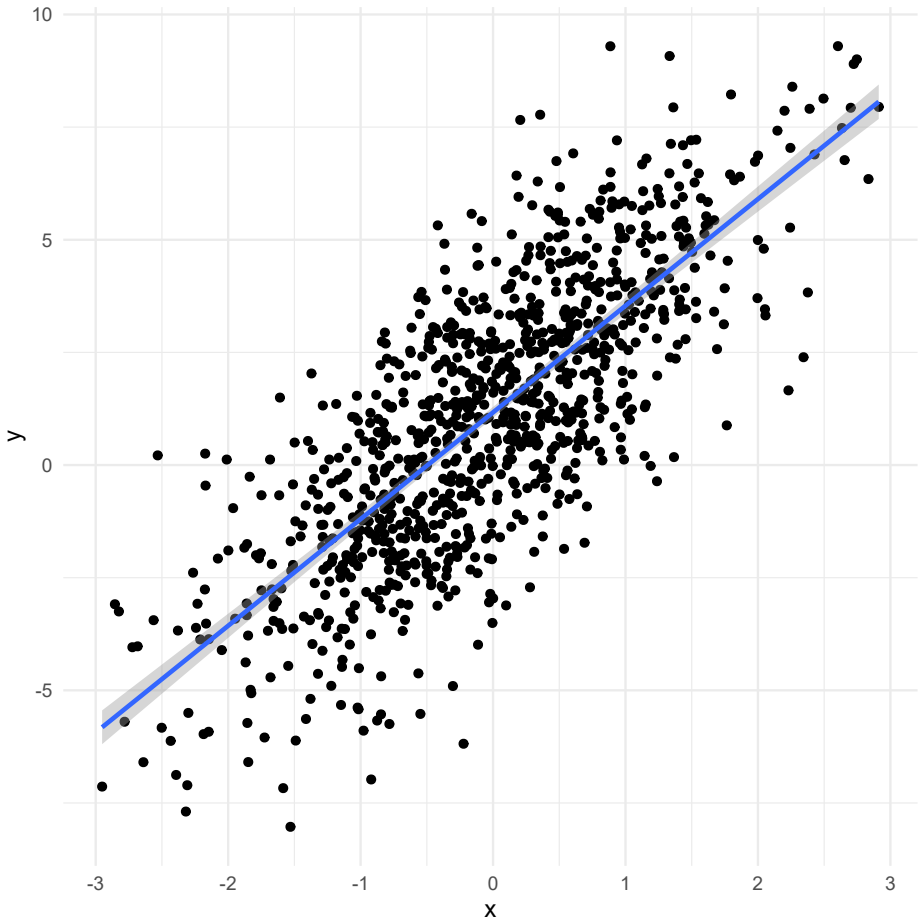
where $e \sim N(0, \sigma)$. This is not a new regression line, it's saying y is the model line we fit plus the error of that line (the difference between the line and the y_i points=e_i). Those differences are normally distributed if the model assumptions are met

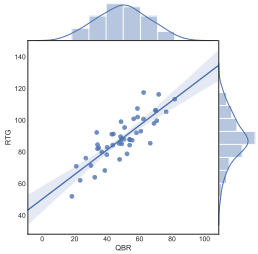
- That means we fit a line to the data and if we plot how much we miss, we get a normal distribution i.e. we are normally wrong (clever pun?) app





Simple Linear Regression





Simpson Paradox

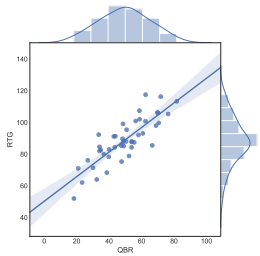
See Wiki link

- Example: Derek Jeter had a .250 batting average in 1995 and a .314 in 1996. David Justice had a .253 in 1995 and .321 in 1996.
- However, Jeter's combined average between 1995 and 1996 was .310 versus .270 for Justice
- How is this possible?

Chapter 23 Notes

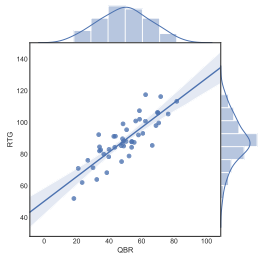
Inference for Regression
STP-231

Arizona State University



Regression Parameters

- Previously, our regression line is calculated from statistics
- $\overline{x}$, $\overline{y}$, s_x, s_y, and r change if different samples are chosen
- So does the line $\hat{y} = b_0 + b_1 x$
- To perform inference we think of b_0 and b_1 as estimates of the regression parameters that describe the entire population

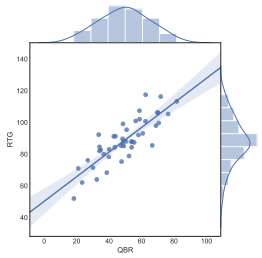


Conditions for Regression Inference

- For inference on the relationship between explanatory variable x and response variable y assume:
- y is normally distributed at any fixed value of x
- The mean response μ_y has a linear relationship with x given by the population regression line

$$\mu_y = \alpha + \beta x$$

- The parameters α and β represent the unknown intercept and the unknown slope respectively.
- The standard deviation of y, σ_y is unknown, but the same for all values of x



Pictoral

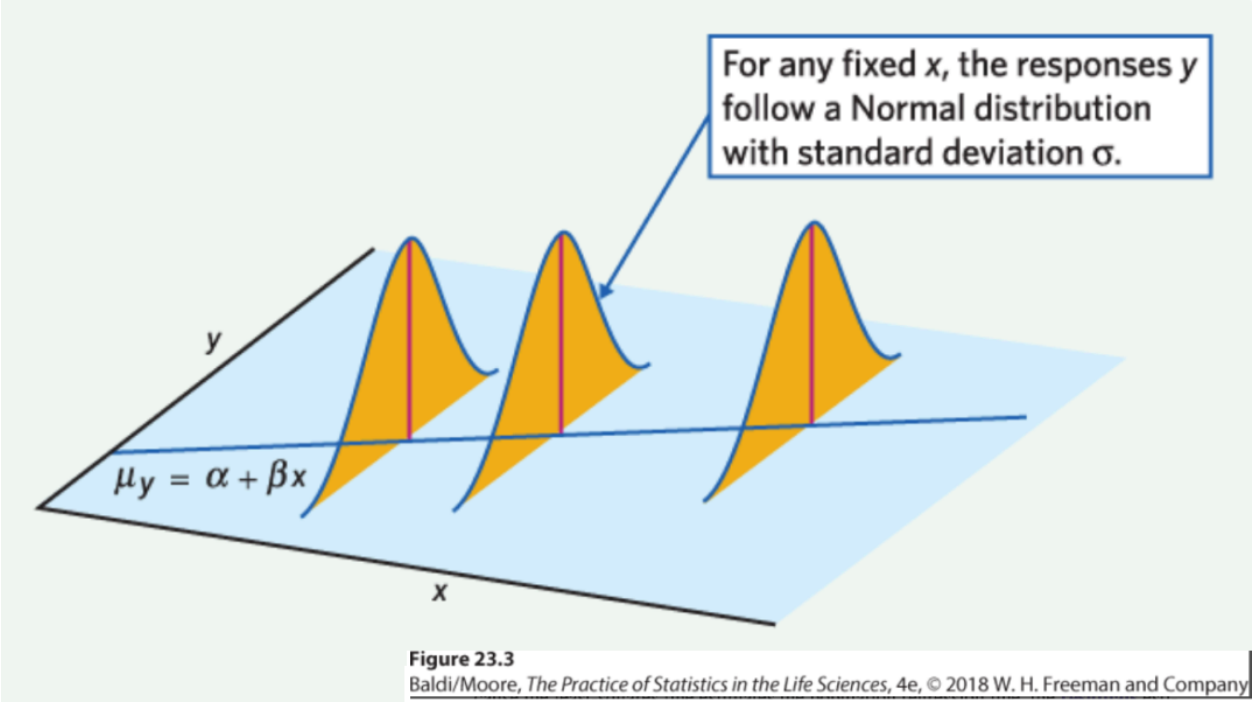
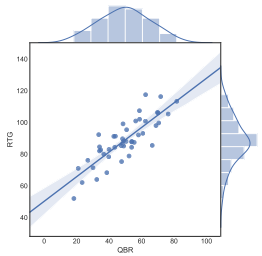
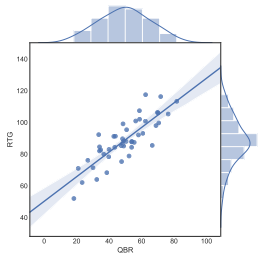


Figure 23.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



Point Estimates

- For n observations of explanatory variable x and response variable y:
- b_0 estimates the unknown intercept parameter α
- b_1 estimates unknown slope parameter β
- s_e estimates unknown standard deviation σ_y



Testing hypothesis of no linear relationship

We test the hypothesis:

$H_0 : \beta = 0$ The slope is zero, no linear relationship

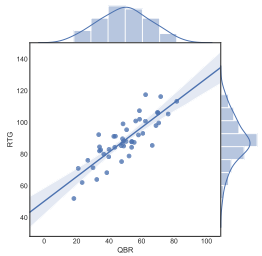
$H_A : \beta \neq 0$ The slope is not zero, there is linear relationship

The test statistic is

$$t^* = \frac{b_1}{\text{SE}_{b_1}} \sim t(\text{df}=n\text{-}2)$$

And the standard error of the least squares slope SE_b is

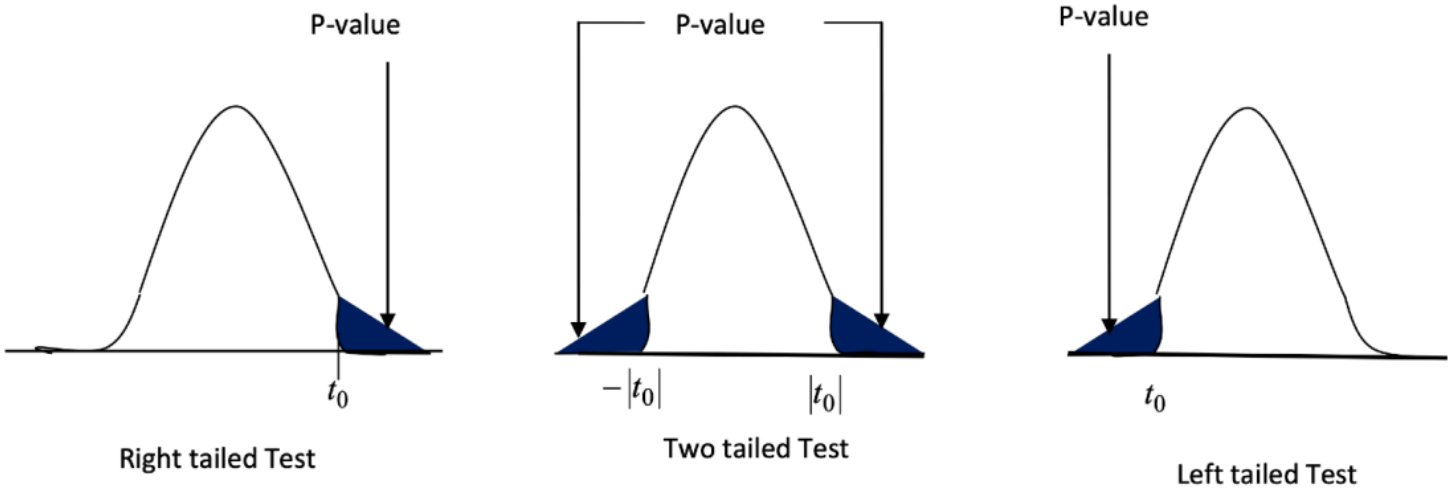
$$\text{SE}_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{s_e}{s_x \times \sqrt{n-1}} = \frac{\text{residual std dev}}{\text{std dev of } x \times \sqrt{n-1}}$$

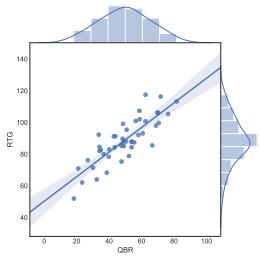


A familiar face appears

Compare the p-value to the significance level α. Is the p-value $\leq \alpha$. Recall, $H_a : \beta \neq 0$, and we get the p-value from the 2-sided approach, where (T being the random variable from the t-dist)

$$\text{p-value} = 2 \cdot \Pr\Big(T > \big|t^*\big|\Big)$$



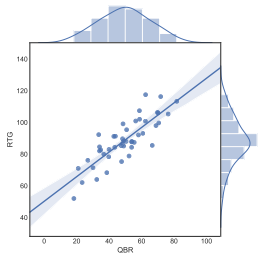


Confidence interval for regression slope

- A $1-\alpha$ confidence interval for β is:

$$b_1 \pm t_{\alpha/2,n-2}\text{SE}_{b_1}$$

- where $t_{\alpha/2,n-2}$ is the $\alpha/2$ quantile from a t-distribution with $n-2$ degrees of freedom

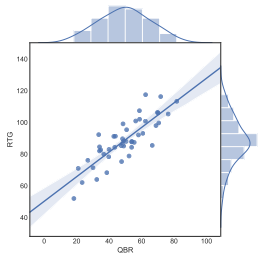


Test of lack of correlation

Linear relationship described by slope b is closely related to r

- When the correlation is 0, the slope will be exactly 0
- When the correlation is not 0, slope will not be exactly 0
- Repeat testing the hypothesis of no linear relationship
- Let ρ be the population correlation coefficient
- The hypotheses:

$$H_0 : \rho = 0 \text{ there is no correlation}$$
$$H_a : \rho \neq 0 \text{ there is correlation}$$



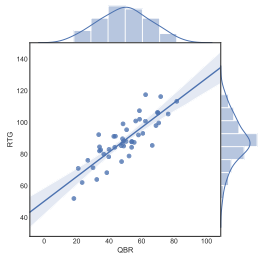
Test of lack of correlation (alternative)

ρ the population correlation coefficient. Assumptions:

- Simple random samples and normal distributions for x and y

$$H_0 : \rho = 0 \; x \text{ and } y \text{ uncorrelated in the population}$$
$$H_a : \rho \neq 0 \; x \text{ and } y \text{ correlated in the population}$$

- The test statistic is $t^* = r\sqrt{\frac{n-2}{1-r^2}}$ following a t-distribution with df=$n-2$. Reject if p-value $\leq \alpha$.



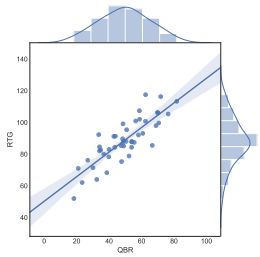
Inference on Prediction: Confidence Intervals

- A $1-\alpha$ confidence interval for μ_y at $x = x^*$ is:

$$\hat{y} \pm t_{\alpha/2,n-2}\text{SE}_{\hat{\mu}}$$

where $t_{\alpha/2,n-2}$ is the $\alpha/2$ quantile from a t distribution with $n-2$ degrees of freedom and

$$\text{SE}_{\hat{\mu}} = s_e\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$



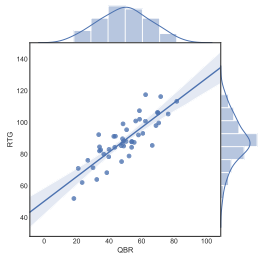
Inference on Prediction: Prediction Intervals

- A $1-\alpha$ prediction interval for single observation y at $x = x^*$ is:

$$\hat{y} \pm t_{\alpha/2,n-2}\text{SE}_{\hat{y}}$$

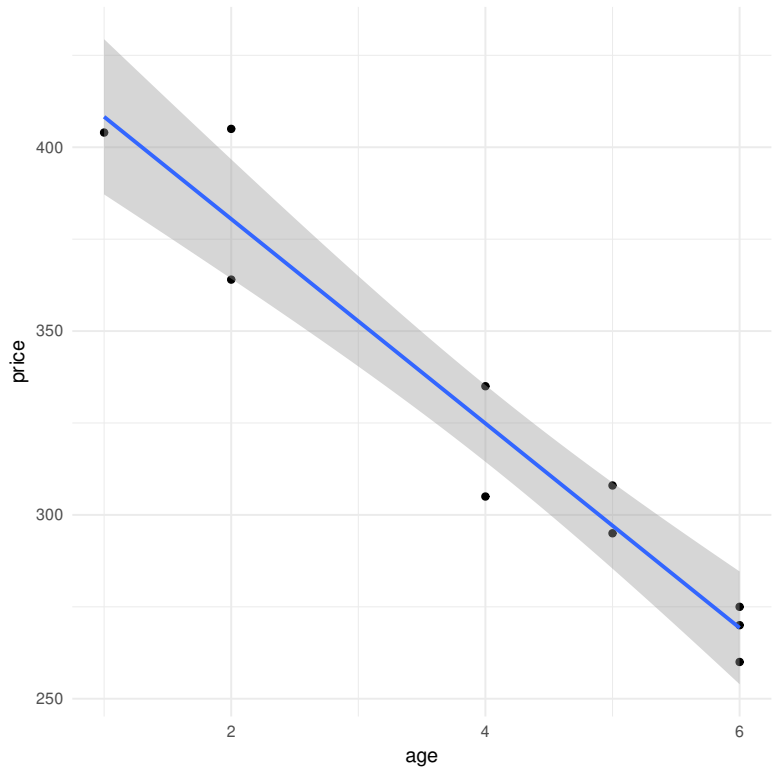
where $t_{\alpha/2,n-2}$ is the $\alpha/2$ quantile from a t distribution with $n-2$ degrees of freedom and

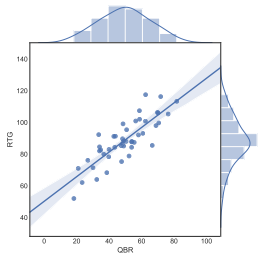
$$\text{SE}_{\hat{y}} = s_e\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$



Example 1

age	6	6	6	2	2	5	4	5	1	4
price	270	260	275	405	364	295	335	308	404	305

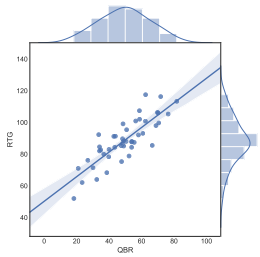




Example 1 Continued

age	6	6	6	2	2	5	4	5	1	4
price	270	260	275	405	364	295	335	308	404	305

Find the slope and intercept



Example 1 Continued

age	6	6	6	2	2	5	4	5	1	4
price	270	260	275	405	364	295	335	308	404	305

Find the slope and intercept

$$\overline{x} = 4.1 \qquad \overline{y} = 322.1$$

$$s_x = 1.85 \qquad s_y = 53.25$$

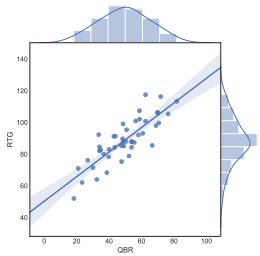
$$r = -0.967499 \quad r^2 = 0.9361$$

$$\sum_{i=1}^{n}(y - \hat{y})^2 = 1631.7$$

$$b_1 = r\frac{s_y}{s_x} = -0.9675 \cdot \frac{53.25}{1.85} = -27.8$$

$$b_0 = \overline{y} - b_1\overline{x} = 322.1 - (-27.8) \cdot 4.1 = 436.1$$

$$\hat{y} = 436.1 - 27.8x$$



Example 1 Continued

age	6	6	6	2	2	5	4	5	1	4
price	270	260	275	405	364	295	335	308	404	305

Test the hypothesis of no linear correlation at significant level 0.05. $t^* = r\sqrt{\frac{n-2}{1-r^2}}$

$t^* = -10.82$ compare to $t_{.025,10-2} = |t^*| = 10.82 > |t_{.025,10-2}| = 2.31$

The p-value is 2*tnorm(|−10.82|, 10-2)=4.69×10^{-6} (using R) in calculator, 2*tcdf(|−10.82|, 1E99, 10-2) or
$2 * (1 - \text{tcdf}(-1E99, |-10.82|, 10 - 2))$