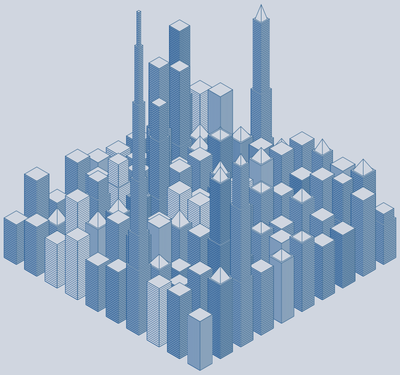
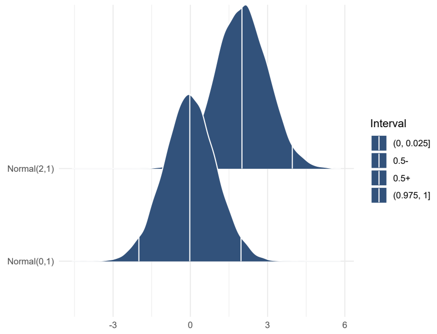


Statistics Notes

ASU

Contents



1 Introduction

The statistics we cover in this course are broadly broken up into three parts: summary, evaluation, and prediction.

For the summary section, we cover what most people traditionally define as statistics. We will go over terms like mean, median, mode. We will look at many different charts and graphs. We will make tables and look at data. Terms like sampling, polling, and surveys fall into this category of statistics. We will describe the difference between a sample vs a population, and how this shapes much of modern statistics, even in an age of "Big Data". These concepts are the baseline of modern statistics, and thus will be the foundation of course.

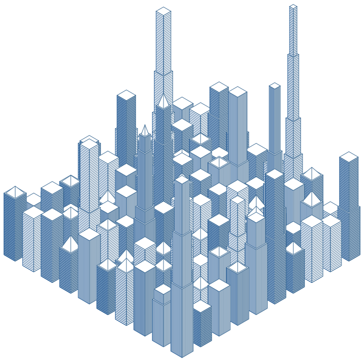
Next, we turn to evaluation, whose more technical term is **Inference**. Inference is the process of looking at the statistics that summarize our data and seeing if any effects we have are real. Inference is based on **probability** theory, a deep and interesting field of mathematics in its own right. Essentially, the study of probability is the study of randomness and chance. At the crux of one of the main philosophies of statistics, **frequentism**, is that we observe random data that over enough time stipulates that some events should randomly happen. The question we then ask is how much more/less often did our observed event happen compared to what was expected. This is where terms like "p-value" and "hypothesis tests" dominate. The other main philosophy is **Bayesian** statistics, where we stipulate that the truth is random, and our final conclusions are guided by what we believed prior to witnessing our data *and* the data itself.

Finally, we will look at prediction. This is probably many people's main interest in statistics.

1.1 Review

Sampling, statisticians have told us, is a much more effective way of getting a good census.

Rob Lowe



2

An Introduction to Sampling (Chapter 6)

Equations and Format

- The test will be 12 questions, 10 multiple choice and 2 free response. The multiple choice will be a mix of graphing, conceptual, and calculations. These are the formulas you will receive:

$$
\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 \\
s &= \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2} \\
\text{coefficient of variation} &= \frac{s}{\overline{y}} \\
\text{Range} &= \text{Max-Min} \\
\text{Bin width} &= \frac{\text{Max-Min}}{\text{\# Bins}} \\
\text{IQR} &= Q_3 - Q_1 \\
\text{Lower Limit} &= Q_1 - 1.5\cdot(\text{IQR}) \\
\text{Upper Limit} &= Q_3 + 1.5\cdot(\text{IQR})
\end{aligned}
$$

In an **observational study**, data is only observed and no conditions are manipulated. In an **experiment**, the conditions are imposed by the researcher. Typically, assignment in an experiment is random, but this need *not* be the case. We just need the researcher to control who is in which group in the experiment, whereas in the observational study we observe the data without controlling who receives which conditions.

Sampling methods, such as a simple random sample, are used either to select who is to participate in an experiment OR who to participate in an observational study. An example of this is a longitidunal observational study. Researchers could choose which schools to observe behavior/results in an observational study. An experiment would require assigning (in this example) certain schools to some treatment (such as giving every student a laptop in one school and none in the other). Whereas in an observational study, the researcher is like an outside observer, simply watching what occurs without choosing who receives what. In an observational study, we can only draw **association**, whereas in an experiment (with random assignment) we can not only see association, but also **causation**. Do not assume just because we have an observational study we cannot have an association. We can, and often do, we just do not want to automatically assume the association is a causation. Similarly, an experiment does not guarantee causation, as there are issues such as administration of the experiment and whether or not groups were randomly assigned.[1]

We learned about 3 ways to sample: simple random samples (SRS), stratified sampling, and cluster sampling. Simple random samples are the easiest and most common for us. Stratified sampling is useful when we have certain subgroups in our population we want to make sure we sample appropriately, such as stratifying by age groups. This lowers our variance, but is harder to sample. Cluster sampling is used when SRS or stratified sampling is too expensive, but this type of sampling does a worse job estimating our population.

Recall, the population is the "ground truth". That is, the population is the entire group we want to estimate with samples, and parameters are the true values we want to estimate with statistics. We draw a sample of size n to model a population of a size that is either unknown or too big to realistically sample. The sample size is the number of people/units (total) in the sample.

Some practice questions!

1. **Sort of a trick question. Just saw it on the weather channel describing the winter storm Orlina. Gerry starts a study looking at car accidents and snowfall. Gerry finds that 70% of accidents occur in snow storms with less than 2 inches of snow. You conclude that there less snow means more slipperyness and write an article**

[1] For an interesting read into studying causation in observational studies, read this. Very cool!

stating that less snow *causes* more accidents with this as the causal mechanism. However, upon further reading their study, Riley says "what nonsense!". She sees that 90% of snow storms result in less than 2 inches of snow. Riley writes that concluding causation was wrong since this was an observational study, and further goes on to say there was *no* association between amount of snowfall and car accidents, and that less snowfall does not make the road more slippery than heavier snowfall totals. Which of the following statements is correct comparing Riley and Gerry's interpretations.

(a) Gerry's intepretation was fully correct.

(b) Riley is correct that the causation cannot be drawn, but there is an association between light snowfall amounts and accidents, although it is negative! There is not enough information to say anything about how slippery the snow is.✓

(c) Riley is correct that the causation cannot be drawn and that there is no association. There is not enough information to say anything about how slippery the snow is.

(d) Riley is correct that the causation cannot be drawn, but there is an association between light snowfall amounts and accidents, although it is negative! There is enough information to say anything about how slippery the snow is, so Riley is wrong about that.

<u>Answer:</u> This is an example of where you might see a headline and wanna read deeper. People also are less likely to drive to begin with when there is a heavy amount of snow. This is a confounder in an observational study that we can reasonably assume biases our estimate of how many accidents occur upwards. Therefore, if anything, the association between low snowfall and more accidents is even *more* negative, which confirms intuition for anyone that's ever driven in the snow before.

2. **A study of 300 college students finds that 240 of them drank coffee before they were 18. Of those 240, 180 wore glasses. Of the 60 who did not drink coffee before 18, 15 wore glasses. Answer the following questions:**

   (a) Is this an observational study or experiment?

   <u>Answer:</u> Observational. No researcher chose which groups were gonna drink coffee and which were not.

   (b) What is the sample size? The proportion (aka the relative frequency) of wearing glasses given that you drank coffee before 18? Given that you did not drink coffee before 18?

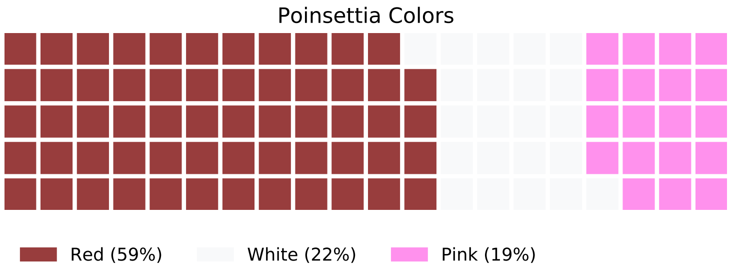
   <u>Answer:</u> The sample size is 300, the proportions are 75% and 25%.

   (c) Is there an association between drinking coffee as a child and wearing glasses in college? Can we draw a causal conclusion? If no, what are some potential confounders?

Answer: Yes, there is an association because the proportion of coffee drinkers wearing glasses is 75% vs 25% for non-coffee drinkers (just a coincidence these add up to 100%). Because this is an observational study, we *cannot* draw a causal conclusion. Some potential confounders include perhaps people drank coffee because they were studying a lot in high school, and they studied a lot because they wanted to buy new glasses (this part is jusr for fun). It doesn't really seem like there is a connection between coffee and eyesight, so this result, causally at least, should seem dubious from the onset.

Exploratory data analysis is an underrated part of a statistical analysis

A lot of data scientists



3

Picturing Distributions with Graphs (Chapter 1)

3.0.1 Charts and Graphs

Recall we looked at many different types of charts and graphs in our class so far to look at frequency distributions. For categorical variables, these include bar charts, pie charts, and the waffle chart[1]. For numeric data, we looked at dot plots and stem and leaf charts, which as our data grew evolved into histograms and boxplots. Histograms are useful for looking at big data and make the shape of our distribution more obvious, whereas boxplots are more useful for locating outliers and are easier to draw.

In order to make a histogram, we have to administer cut point grouping. To do so, we do the following, given the number of bins (or classes, a bin/class is equivalent) we need. First, we calculate the approximate bin width from

$$\frac{\text{Max value- minimum value}}{\text{\# of bins}}$$

If it is already a whole number, this is your bin width. Otherwise, round up to the next integer. Obtain the lower cut points by successively adding the chosen bin width. Specify all the bins, and the determine which observations belong to each class and count frequencies. We will have an example at the end.

[1] waffles are the coolest?

3.0.2 Types of Variables

We have looked at quantitative/numeric variables and categorical. The main difference is we used numbers to describe quantitative and categories for categorical. For example, a quantitative variable could be the amount of snowfall a city averages in a year, or the the number of cities in a county. The first is a continous quantitative variable, whereas the second is discrete because it takes on an integer (whole number) value. A categorical variable can be, for example, your class standing, i.e. freshmen, sophomore, junior, senior. This is an ordinal variable because there is an ordering based on college experience. A nominal categorical variable does not have an order, such as the state you live in. Note, if we used class year in terms of numbers (i.e. first grade, second grade, etc.) you could argue that is a discrete numeric variable. The distinction is kind of moot, so for this class just know if you see it described as a number to think numeric, otherwise categorical.

3.0.3 Skewness and Modality

We also talked about skewness and modality. If we are **left skewed**, then the left side of the histogram, containing the half of the observations with smaller values, extends much farther out than the left side. Additionally, this typically means the mean is *smaller* than the median. Why? Because the outliers are to the left, then they are less positive, and since the mean is more affected by outliers than the median, the mean will be smaller.

If we are **right skewed**, then the righer side of the histogram, containing the half of the observations with larger values, extends much farther out than the left side. Additionally, this typically means the mean is *bigger* than the median. Why? Because the outliers are to the right, then they are more positive, and since the mean is more affected by outliers than the median, the mean will be bigger. If the left and right side are basically mirrored, we have a **symmetric** distribution.

The terms left and right skew can be confusing. A right skew means you have more data points far out to the right, meaning your largest **modes** will likely be on the left, and vice versa. So be careful about that.

1. **With the following data, create a grouped frequency distribution with 5 classes (bins) and plot with a histogram:**

1.79	2.48	4.73	6.12	8.18	8.80	9.51	10.18	10.74	10.93	11.12	12.29	12.32	12.52	12.75	12.91	13.24	13.94	14.10

Answer: The class width is

$$\text{class width} = \frac{14.10 - 1.79}{5} = 2.46 \uparrow 3$$

Therefore, our bins are (because we get to choose the lowest point, we choose the first integer beneath our lowest value)

Bin	Interval	Frequency
1	1-less than 4	2
2	4-less than 7	2
3	7-less than 10	3
4	10-less than 13	9
5	Greater than or equal to 13	3

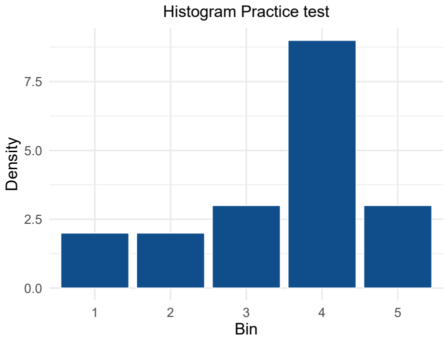
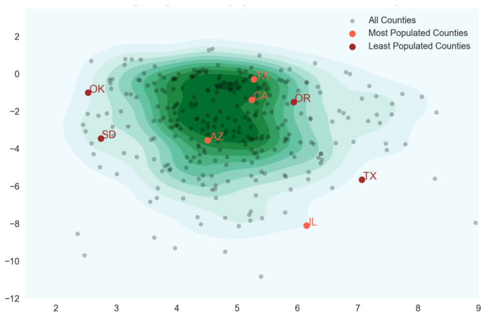


Figure 3.1: Not the best title, but don't worry too much about those. Typically you want to draw the bars touching each other to distinguish from a bar chart. But, in practice, you wouldn't use a histogram with this few bins.

Even a broken watch is right twice a day



4

Describing Quantitative Distributions with Numbers(Chapter 2)

4.0.1 Measures of Spread

We talked about the measures of center and spread in class. These include the mean, median, mode for measures of center and Quartiles, range, variance, standard deviation, and coefficient of variation for measures of spread. Typically, the mean is the least resistant, as we saw in the slides, whereas the mode and median are situationally affected by outliers (in fact the mode can be an outlier!). For the measures of spread, the standard deviation is probably the most informative, is more resistant than the range, but less resistance than the IQR. We will do some examples to calculate these values at the practice questions section.

1. **You receive the following data. It is the high temperature in Phoenix for the last 14 days, in Farhenheit :)** [1] **The data is as follows: Calculate the variance, the standard deviation, the median, the interquartile range, the mean, and the coefficient of variation.**

49	54	65	67	69	70	72	74	75	76	76	77	78	81

[1] Typically favor the metric system, but Fahrenheit is more spread out which makes it slightly easier to detect changes in temperature. For example a 5°C is the same as a 9°F, which seems more pronounced. However, Fahrenheit not being centered at 0 and 100 degrees for freezing and boiling water is dumb.

Answer: Conveniently, the data is already ordered. First, lets do the hard ones. For the Interquartile range, we need to calculate Q_1 and Q_3. Those are as follows:

$$Q_1 = \cancel{49}, \cancel{54}, \cancel{65}, 67, \cancel{69}, \cancel{70}, \cancel{72}$$
$$Q_3 = \cancel{74}, \cancel{75}, \cancel{76}, 76, \cancel{77}, \cancel{78}, \cancel{81}$$

So

$$\boxed{\text{IQR} = 76 - 67 = 9}$$

The median is calculated by the 2 middle values (since we are using an even number of observations), which is the average of 72 and 74, so

$$\boxed{\text{median} = Q_2 = 73}$$

For the other measures we have:

$$\begin{aligned} s &= 9.12^\circ F \\ s^2 &= 9.12^2 = 83.3 \\ \bar{y} &= 70.2^\circ F \\ \text{coefficient of variation} &= \frac{s}{\bar{y}} = 0.13 \end{aligned}$$

2. Calculate the median, mean, and IQR from the following stem and leaf plot: Are there any outliers?

Stem	Leaf
9	1 2 4 5
8	0
7	2 3 4
6	1 5
5	1 4 5 5
4	2 4
2	2
1	0 6
0	4 8

Answer: We first write these in order

$$4, 8, 10, 16, 22, 42, 44, 51, 54, 55, 55, 61, 65, 72, 73, 74, 80, 91, 92, 94, 95$$

This gives: (remember for the IQR when we have an odd number of total observations we *do NOT* include the median point, i.e. the 11th observation here)

$$Q_1 = \cancel{4}, \cancel{8}, \cancel{10}, \cancel{16}, 22, 42, \cancel{44}, \cancel{51}, \cancel{54}, \cancel{55} = 32$$
$$Q_3 = \cancel{61}, \cancel{65}, \cancel{72}, \cancel{73}, 74, 80, \cancel{91}, \cancel{92}, \cancel{94}, \cancel{95} = 77$$

meaning

$$\boxed{\text{IQR} = 77 - 32 = 45}$$

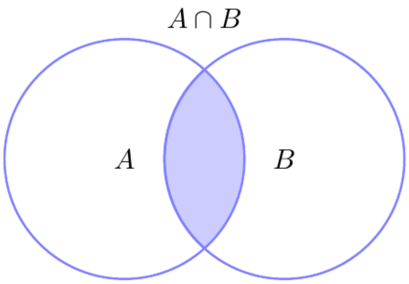
The median is the 11th value which is 55, and mean is 55.14. Therefore, this is roughly symmetric. A boxplot or histogram would show that as well. There are no outliers because nothing is outside $Q_1 - 1.5 \cdot \text{IQR}$ or $Q_3 + 1.5 \cdot \text{IQR}$. To draw a boxplot, we note there are two kinds, the regular or the modified. For a regular boxplot, draw the whiskers to the min on the left, and the max on the right. A modified boxplot has whiskers to Q_1-1.5*IQR on the left (with outliers beyond this as dots) and Q_3+1.5*IQR on the right(with outliers beyond this as dots).

3. **We measured the average commute time of all the students in our class that live on campus. Let's assume most peoples commute is from their bed to zoom console. So 33 people are between 0 and 1 minutes. However, 4 people (students plus teacher + tech admin). Is this distribution symmetric or skewed? Which way is it skewed? Is the mean bigger than the median? What is the mode?**

   <u>Answer:</u> We'll use the poll results from zoom, but lets just assume the mode is around 1 minute, same as the median, but the mean will be higher, so this is right skewed. This also shows the mean is a less resilient measure and more affected by outliers.

God does not play dice with the Universe

Einstein



5 Essential Probability Rules(Chapter 9)

Equations and Format

- The test will be 14 questions, 11 multiple choice and 3 short response.

$$\begin{aligned}\Pr(E^c) &= 1 - \Pr(E)\\ \Pr(E_1 \text{ or } E_2) &= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \text{ and } E_2)\\ \Pr(E_1 \mid E_2) &= \frac{\Pr(E_1 \text{ and } E_2)}{\Pr(E_2)}\\ \Pr(E_1 \text{ and } E_2) &= \Pr(E_1) \times \Pr(E_2) \text{ If independepent probabilities}\end{aligned}$$

- **Binomial Distribution**

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \mu = np \quad \sigma = \sqrt{np(1-p)}$$

- **General Random Variables**

$$\mu_Y = \sum \Big(y_i \times \Pr(Y = y_i)\Big) \qquad \sigma_Y = \sqrt{\sum \Big[(y_i - \mu_Y)^2 \cdot \Pr(Y = y_i)\Big]}$$

- **Descriptive measures**

$$\overline{y} = \frac{\sum_{\text{sample}} y_i}{n}$$
$$\mu = \frac{\sum_{\text{pop.}} y_i}{N}$$
$$\sigma = \sqrt{\frac{\sum (y_i - \mu)^2}{N}}$$
$$z = \frac{y - \mu}{\sigma}$$

5.1 Chapter 9: Probability

In this chapter, we studied how probabilities are defined and introduced some concepts. In general, probability is the study of random events. Individual outcomes are uncertain, but we still have a rough idea of how to quantify what we expect to happen upon repetition. To make this more clear we define:

1. A **Sample Space**, aka $\mathcal{S}$, which includes all of the possible outcomes of a random phenomenon.
2. An **event** is an outcome or set of outcomes of a random phenomenon. Thus, an event belongs *in* our sample space.

Some rules about probabilities we should know: The probability of an event, $\Pr(E)$ is between 0 and 1, i.e. $0 \leq \Pr(E) \leq 1$. The probability of an impossible event is 0, and a certain event is 1. In other words, a probability of 0 implies an event cannot happen, whereas a probability of 1 implies it will happen. The sum of the probabilities of all events is 1, which means that the probability that something in your sample space happens is 1. For example, if we want to roll a die, there are 6 possibilities. We can roll a 1, a 2, a 3, a 4, a 5, or a 6. The probability we roll a 1,2,3, 4, 5, or 6 is 1, whereas the probability we roll any other number is 0. This means our sample space is:

$$\mathcal{S} = \{\text{roll a } 1, 2, 3, 4, 5, 6\}$$

When two events have no outcomes in common, they can never happen together. We call this **mutually excusive** or **disjoint**. If we only have two events, call them A and B, then the probability one or the other happens, or both, is

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

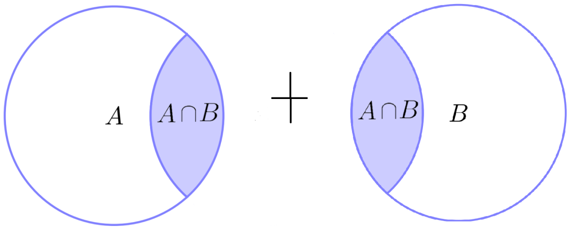


Figure 5.1: Where the equation for $\Pr(A \text{ or } B)$ comes from

Why is this? If we draw $\Pr(A)$ and $\Pr(B)$ as circles in a venn diagram, then we see that we count the overlapping area twice! So we must subtract one of them. See figure 5.1 Notice, if the events are **disjoint/mutually exclusive**, then

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

because there is no overlap.

Finally, the probability an event does not occur is 1 minus the probability the event does occur, i.e.

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

Y	0	1	2	3	4	5
Probability	0.02	0.2	0.5	0.22	?	0.03

Table 5.1: What is the probability $Y = 4$? Assume these events encompass all of the sample space $\mathcal{S}$.

Example In table 5.1, what is the probability $Y = 4$? We find this by doing

$$\Pr(Y = 4) = 1 - \Pr(Y = 0) + \Pr(Y = 1) + \Pr(Y = 2) + \Pr(Y = 3) + \Pr(Y = 5) = 0.03$$

5.1.1 Random Variables

Random variables are the building block of probability theory. They can either represent discrete or continuous values. A discrete random variable can take on specific values. For example, in a population of flies, if the random variable X represents the number of flies alive after 24 hours, this is a discrete random variable because the values X can take are whole numbers, i.e. $\{0, 1, 2, \ldots, \}$. A continous random variable, on the other hand, represents data where specific values are impossible. In that case, we are interested in ranges of values. For example, if the random variable Y represents a number chosen between $[0, 1]$, then this is a continuous random variable because there are infinite values between

0 and 1 and it is impossible to fall exactly on a single value (assuming no rounding). For example, if the numbers represents the % of battery life left on a computer (say its on a 0-1 scale, not a 0-100), then the number is constantly changing. At one second, it may be at 0.672742882818382... (multiply by 100 to get into a %-age point) battery level, and there is no clear cutoff as to what exactly the value is. In this case, we do not assign a probability to a value, but rather a probability to an interval, i.e.

$$\Pr(a \leq Y \leq b) = \Pr(a < Y < b) \tag{5.1}$$

See figure 5.2 The equations for the expected value (i.e. the theoretical mean of a distribution) and the theoretical standard deviation are:[1]

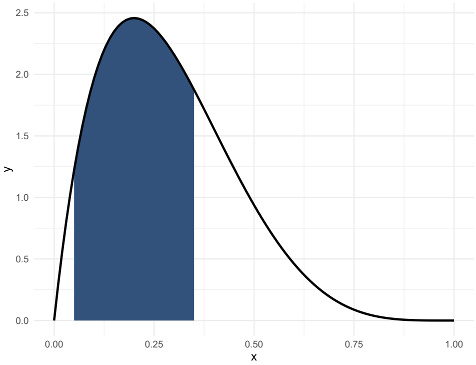


Figure 5.2: This is a distribution of possible values on the horizontal axis and their densities on the vertical axis. The shaded area represents the probability to be between points A and B. This is in fact a Beta distribution, which is what the Polya urn converges to. That is the process of having an urn with an equal number of red and blue balls. Every time you select a ball, you replace it with 2 balls of the same color. Weirdly, the final distribution of colors of balls is a random variable, and the entire process is a martingale, i.e. $E(X_{n+1} \mid X_1, \ldots X_n) = X_n$.

$$\mu_Y = \sum \Big(y_i \times \Pr\big(Y = y_i\big) \Big) \qquad \sigma_Y = \sqrt{\sum \Big[(y_i - \mu)^2 \cdot \Pr\big(Y = y_i\big) \Big]}$$

1. **Suppose we poll people on the proper social protocol for grabbing cookies at a free food event. The proportion of people who answered 0 cookies is 0.02, 0.42 for 1 cookie, and so on, with the appropriate answers given in table ??. Let X be the number of cookies people think is a cool amount with $\Pr(X)$ denoting the probability that someone would agree with your choice of cookies to grab as being the right amount.**

   Find the mean number of cookies you should take to appease the crowd.

   Answer: First, we need to find $\Pr(X = 3) = 1 - 0.02 + 0.42 + 0.45 + 0.04 = 0.07$. Then the expected value of X is

   $$\mu_X = 0 \cdot 0.02 + 1 \cdot 0.42 + 2 \cdot 0.45 + 3 \cdot 0.07 + 4 \cdot 0.04 = 1.69$$

[1] The difference between theoretical and observed is the theoretical refers to the population and the observed is the sample.

X	$\Pr(X)$
0	0.02
1	0.42
2	0.45
3	???
4	.04

2. **Event A depends on event B. Are A and B mutually exclusive (aka disjoint) events?**

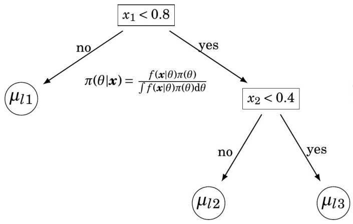
   Answer:

   (a) No, because independence implies they cannot be mutually exclusive ✓

   (b) Yes, because independence implies they are be mutually exclusive

Persi, it soundedas if you were apologizing for being a Bayesian.You don't have to apologize, it's the only sensiblestatistical theory.

David Blackwell



$x_1 < 0.8$
no
yes
$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta}$
μ_{l1}
$x_2 < 0.4$
no
yes
μ_{l2}
μ_{l3}

6 Conditional Probability and Bayes Rule (Chapter 10)

6.1 Conditional Probabilities & Bayes

If an event is **independent**, then the probability A and B happen is the product, i.e.

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B) \tag{6.1}$$

First, we define conditional probability as

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} \tag{6.2}$$

See figure 10.2 for an explanation of where this equation comes from. Note, the $|$ symbol means given. So the probability A given that B has happened is the ratio that both happen divided by the probability what we have seen would happen in total.

If events A and B are independent, then the probability of A is the same where or not we have seen B or not. Therefore, in that case,

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

So the $\Pr(A \mid B) = \Pr(A)$ because A does not depend on B. Notice how this also shows that independence does not mean disjoint. Because if two events are disjoint, $\Pr(A \text{ and } B) = 0$, whereas here $\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$.

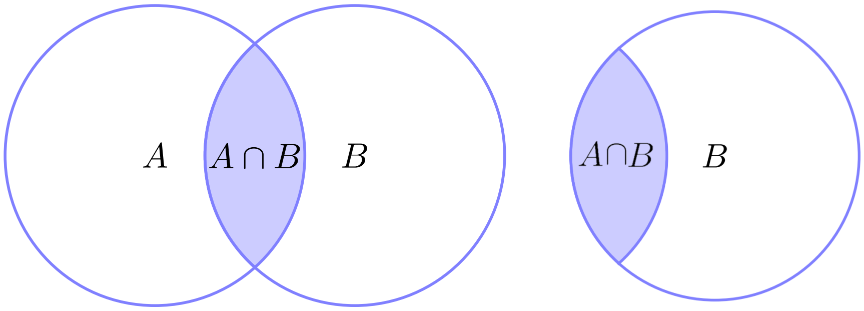


Figure 6.1: On the left is when we see both events, on the right we know the B event did occur, so the probability of A in this situation is the shaded region (the only place where A occurs given that B for sure occurred) divided by the total probability of B.

		Income		
	Low	Medium	High	Total stress
Stressed	526	274	216	1,016
Not Stressed	1,954	1,680	1,899	5,533
Total income group	2,480	1,954	2,115	$\boxed{6,549}$

Table 6.1: Sum across rows to get total stressed. Down columns for income group.

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. See table 6.1

- $\Pr(\text{Low Income})$ To do this, we note there are 2480 total low income people out of 6549. So
$$\Pr(\text{Low Income}) = 2480/6549 = 0.379$$
- $\Pr(\text{Stressed and Low Income})$ Here, we want the section where both of these are true, which is the first column and first row. We want to divide this by the total, i.e.
$$\Pr(\text{Stressed and Low Income}) = \frac{526}{6549} = 0.08$$
- $\Pr(\text{Stressed} \mid \text{Low Income})$ We use the answer above to calculate the probability someone is stressed and low income, which we found to be 0.08. , We know probability of being low income as 2480/6549=0.379. Given that we eliminate the possibilities to be medium or high stress, and that the total 6549 is consistent between the numerator and denominator of the equation for conditional probability (6.2), then this number divides out. Therefore,
$$\Pr(\text{Stressed} \mid \text{Low Income}) = \frac{526/6549}{2480/6549} = \frac{526}{2480}0.212$$

- Pr(stressed)?

  What is the probability of being stressed? Well, we sum across the row and divide by the total:

  $$\Pr(\text{stressed}) = \frac{1016}{6549} = 0.155$$

  Therefore, we conclude stress level depends on income, since the $\Pr(\text{stressed}) \neq \Pr(\text{Stressed} \mid \text{Low Income})$. You could also show

  $$\Pr(\text{stressed and low income}) \neq \Pr(\text{stress}) \cdot \Pr(\text{low income})$$

6.1.1 Bayes Theorem

For two events A and B

$$\begin{aligned}\Pr(A) &= \Pr(B) \times \Pr\big(A \mid B\big) + \Pr\big(B^c\big) \times \Pr\big(A \mid B^c\big)\\ \Pr(B) &= \Pr(A) \times \Pr\big(B \mid A\big) + \Pr\big(A^c\big) \times \Pr\big(B \mid A^c\big)\end{aligned}$$

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

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

Notice, the i index in the numerator is not the same as the index in the denominator!

This comes from conditional probability:

$$\Pr\big(A \mid B\big) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} \tag{6.3}$$

$$\Pr\big(B \mid A\big) = \frac{\Pr(A \text{ and } B)}{\Pr(A)} \implies \Pr(A \text{ and } B) = \Pr\big(B \mid A\big)\Pr(A) \tag{6.4}$$

$$\implies \Pr\big(A \mid B\big) = \frac{\Pr\big(B \mid A\big)\Pr(A)}{\Pr(B)} \text{ (2) in num, (1) in denom} \tag{6.5}$$

But, $\Pr(B) = \sum_{j=1}^{k} \Pr\Big(A_j \cap B\Big) = \sum_{j=1}^{k} \Pr\Big(B \mid A_j\Big)\Pr\Big(A_j\Big)$ (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).

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

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

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,

$$\Pr\big(A \mid B\big) = \frac{\Pr\big(B \mid A\big)\Pr(A)}{\Pr\big(B \mid A\big)\Pr(A) + \Pr\big(B \mid A^c\big)\Pr(A^c)}$$

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.

Example of Bayes Theorem

- 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

$$\Pr\big(D \mid P\big) = \frac{\Pr\big(P \mid D\big)\Pr(D)}{\Pr\big(P \mid D\big)\Pr(D) + \Pr\big(P \mid D^c\big)\Pr(D^c)}$$

- So most of error comes from rarity of disease and false positives not the test missing the disease

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

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

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

Bonus Bayes Question 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.

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

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

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

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

- Recall, (see slides 16-18)

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

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

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

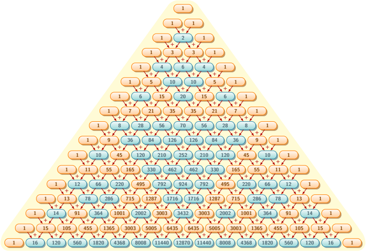
- Prob a student who gets a B has read the instructions correctly:

  $\Pr(B) = 1/4$, $\Pr\big(B \mid R^c\big) = (1/5) \cdot (2/3) = 2/15$. Using Bayes rule:

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

Flip a coin. When it's in the air, you'll know which side you're hoping for.

Arnold Rothstein



7 The Binomial Distribution(Chapter 12)

7.1 The Binomial Distribution

The **binomial distribution** is one of the first distributions we learn about. The binomial distribution is a sequence of n independent Bernoulli experiments. Each individual repetition is a trial, and each experiment results in either a "success" or a "failure". The terms success and failure need not be taken literally, we just need one of the outcomes to be a success and the other a failure (there are only two outcomes). The binomial has n trials, k number of successes, and $n - k$ number of failures. What becomes tricky is to count how many of those outcomes there are. For example, if our binomial distribution is referring to flipping a coin n times, and Y is the number of tails we flip, we need a way to count all the equivalent ways of getting y heads.[1]

In the slides, there is a more thorough explanation of where this term comes from, but the binomial coefficient is defined as:

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

where $n! = n \cdot n - 1 \cdot n - 2 \cdot \ldots \cdot 2 \cdot 1$ With this in mind, the binomial distribution can be calculated as:

$$\Pr(Y = k) = \Pr(k \text{ successes}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \tag{7.1}$$

[1] Notice that y here is lower case. That is because it is a specific value the random variable Y can take.

The main assumptions we make when using the binomial distribution are:

1. Each trial has only two possible outcomes, which again we define as successes or failures.

2. The trials are independent of each other. That is, the probability of a success during your current trial was not changed by any previous trials nor will it change the probability of any future trials.

3. The number of trials does not change, and remains n throughout the experiment.

4. Similarly, p is a constant. The probability of success is the same throughout all the trials.

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.

  See figure 7.1. Not only do we need to calculate each probability uniquely, we must them up. This tree example shows how the binomial coefficients are indeed correct however.

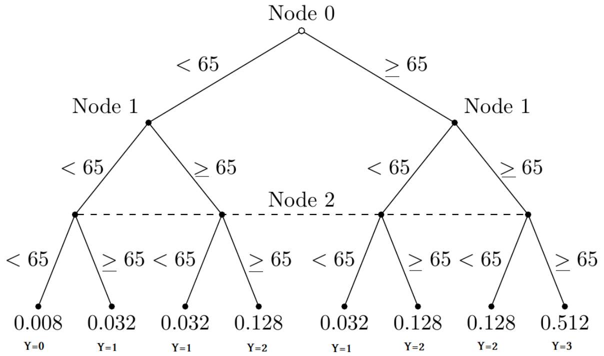


Figure 7.1: Using a tree diagram to express the binomial theorem. Recall, for tree diagrams, multiple down the branch to get joint probabilities, because all the events are independent. This is longer for sure!

- Use the Binomial distribution formula to show how we can get the same results

Here, we can use the binomial distribution formula and individually plug in all the probabilities.

$$
\begin{aligned}
\Pr(Y=0) &= \binom{3}{0}0.8^0(1-0.8)^{3-0} = 0.2^3 = 0.008 \\
\Pr(Y=1) &= \binom{3}{1}0.8^1(1-0.8)^{3-1} = 3\cdot 0.8\cdot 0.2^2 = 0.096 \\
\Pr(Y=2) &= \binom{3}{2}0.8^2(1-0.8)^{3-2} = 3\cdot 0.8^2\cdot 0.2 = 0.384 \\
\Pr(Y=3) &= \binom{3}{3}0.8^3(1-0.8)^{3-3} = 0.8^3 = 0.512
\end{aligned}
$$

- Find the probability that at least one person out of the three was alive at age 65

  We can either add $\Pr(Y=1)+\Pr(Y=2)+\Pr(Y=3)$, or we can do $1-\Pr(Y=0)$.

The expected value of a binomial distribution is np and the standard deviation $\sqrt{np(1-p)}$. That is

$$\mu_{\text{bin}} = np \tag{7.2}$$

$$\sigma_{\text{bin}} = \sqrt{np(1-p)} \tag{7.3}$$

What does that mean? Does it make sense? The mean might at least. It says if you repeat an experiment n times, the expected number of successes is $n\cdot p$. For example, if we flip a coin 1000 times, we'd expect to see 500 heads.

7.1.1 Binomial on calculators

If you have a graphing calculator, the way to find

$$\Pr(Y=k) = \text{2nd binompdf}(n,p,k)$$

For example, in the previous example, what if we wanted to know the probability we get exactly 500 heads? We don't really wanna calculate out the binomial distribution equation in (7.1), so we could plug in

$$\text{2nd binompdf(1000, 0.5, 500)} = 0.025$$

At big n, the mode of the binomial distribution equals its mean because the binomial is approximated as the normal distribution, so this means that the most likely scenario has a probability of 0.025, or occurs about 2.5% of the time. Interesting.

Sometimes we want the sum of events, not just a single event. What if we wanted to know the probability of having more than 500 heads? On the calculator,

$$\text{2nd binomcdf(1000, 0.5, 500)} \approx 0.51$$

Notice, that binomcdf($n, p, \leq y$) *includes* the value on the right when it sums.

1. **Our class has 35 people. Let's say there is a 64% attendance rate on average. What is the probability that 29 students show up on a given day? The probability at least 8 show up?**

   Answer: This is binomial with $n = 35$ and $p = 0.64$ being the probability of success (i.e. a student shows up) The first question can be solved with 2nd binompdf(35, 0.64, 29)

   $$\text{binom}(35, 29, 0.64) = \binom{35}{29} \cdot 0.64^{29} \cdot 0.36^{6}$$

   For the second question, we can plug in 1-2nd binomcdf(35, 0.64, 8) on our calculators

   $$1 - \Pr(X \leq 8) \approx 1$$

2. **You are an avid basketball fan. You're favorite player makes 75% of their free throws. Suppose that they missed 4 in a row. What is the probability of this? (Hint: the shots are independent events)**

   Answer:

   (a) 0.0039=0.39% ✓

   (b) 1=100%

   (c) 0

   (d) 0.316=31.6%

   For this question, notice that this is a binomial(4, 0, 0.75), where 0.75 is the probability of a success. (if you define a miss as a success, then this is binomial(4,4,0.25)). Therefore, to miss 4 consecutive shots is equivalently to miss 1 with prob 0.25 times missing shot 2 with probability 0.25 multiplied by missing shot 3 with probability 0.25 multiplied by missing shot 4 with probability 0.25, aka

   $$\Pr\big(\text{miss 4 straight shots}\big) = 0.25^{4} = 0.0039$$

   There is only 1 possible way to miss 4 consecutive shots, which is why we just multiply the probabilities together and do not need to worry about the binomial coefficient.

3. **Suppose planes overbook their planes because they assume people will not show up. So far a plane that holds 191 people, they actually sell 194 seats to try to secure some extra cheddar. In the case that more than 191 people show up, people may get "bumped" and moved to a different flight. What is the probability at least one person is bumped from their seat?**

Answer:

For every ticket, the outcomes are showing up on time (a failure) or not (a success). These are backwards here, but it'll be easier to proceed this way. Since all of these are independent, this is binomial, with $n = 194$ and $p = 0.01$. $n \neq 191$, which is the number of seats on the plane. That means if all 194 people show up, 3 people will be bumped. However, we are interested in the probability that there is 1 bump, which means that 2 people will show up late, i.e. we want a binomial experiment with $n = 194, p = 0.01, k = 2$.

$$\begin{aligned}\Pr\big(\text{one bump}\big) &= \text{Binomial}(194, 0.01, 2)\\ &= \binom{194}{2} 0.01^2 (1-0.01)^{194-2}\end{aligned}$$

You can also use binompdf(n, p, k) in your calculator.

Now, the probability two people are bumped is the same as the probability one person misses their flight, i.e. one "success"

$$\Pr\big(\text{two bumps}\big) = \text{Binomial}(194, 0.01, 1)$$

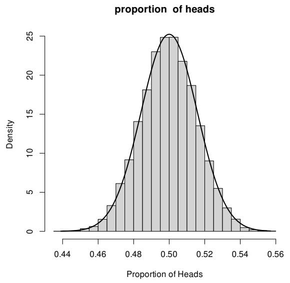
The probability 3 people are bumped is gonna be true when all 194 people show up for the flight but there are only 191 seats, so 3 people will have to lose their seats. This is the case when every experiment in the binomial distribution is a failure, where we defined "failure" as the event where the person shows up on time (which is confusing, but otherwise you'd have to do 1 - in the probability statements).

$$\Pr\big(\text{three bumps}\big) = \text{Binomial}(194, 0.01, 0)$$

Finally, the probability at least one person is bumped is the same as the sum of the previous 3 answers.

Normally wrong

Cool stats podcast name



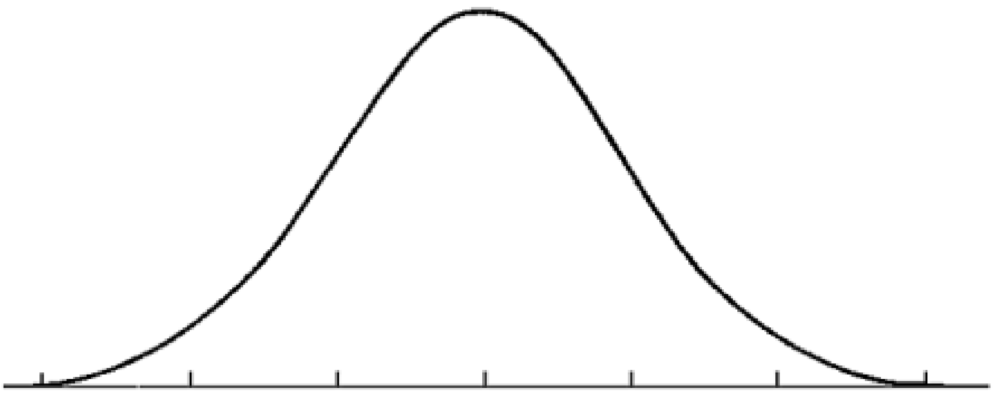
8

The Normal Distribution(Chapter 11)

8.1 The Normal Distribution

Some important properties of the normal distribution:

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



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

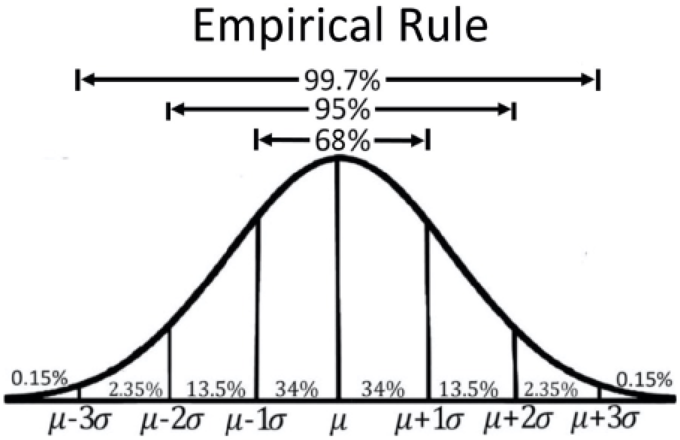
$$Y \sim N(\mu, \sigma)$$

Where N means the distribution is normal. More formally,

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

You don't need to know this.

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



- The mean and the median 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

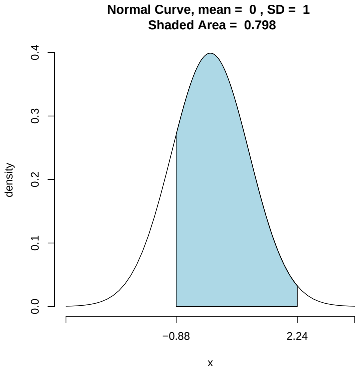
8.1.1 Standardization and Z-score

The $Z-$score is a really important tool we use a lot. Essentially is moves the mean of data to the center and then tells us how many standard deviations away from that center we are. It scales so that it's standard deviation is exactly 1, so being 1.5 σ away means that we are exactly 1.5 standard deviations away. This is not necessarily the case when you are given arbitrary Y data. In equation form

$$z = \frac{y - \mu}{\sigma} \tag{8.1}$$

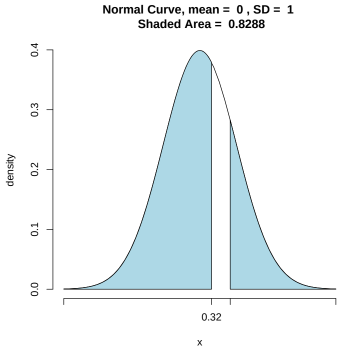
Here are some examples: Determine the area under the standard normal curve that lies between -0.88 and 2.24

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



Find the area below 0.32 and above 0.83 under the curve

$$\begin{gathered} \Pr(Z < 0.32) + \Pr(Z > 0.83) \approx 0.6255 + (1 - \Pr(Z < 0.83) \approx 0.8288 \\ \text{OR } 1 - \Pr(0.32 < Z < 0.83) \end{gathered}$$



In general, these are the steps we would take:

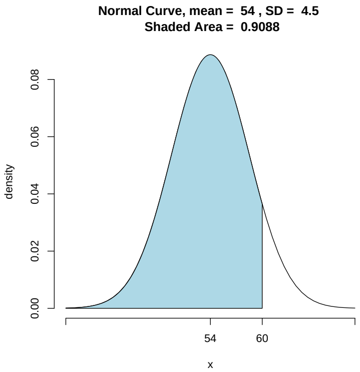
1. Sketch the normal curve associated with the variable
2. Shade the region of interest and mark its delimiting y-value(s)
3. Find the z-score(s) for the delimiting y-value(s) found in step 2
4. Use the table to find the area under the standard normal curve delimited by the z-score(s) found in step 3.

Example with Words 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?

First, we need to convert from y to z. This is because although we know the mean and standard deviation, we can not easily calculate the probability of meaning less than or equal to a certain value when the mean is not 0 and the standard deviation is not 1. For more details on that, go to the Normal wiki and look at the solution of the cumulative distribution function, $\Pr\big(Y < y\big)$. It is a complicated function that must be approximated, and this is easier to do when more terms are 0 or 1.

Once we convert to the z score, we note that:

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



How did we find the probability associated with $z = 1.33$? Using the z-tables.pdf on canvas, we can locate where the row=1.3 and the column=0.03 which is 0.9099. This the probability $\Pr(Z < 1.33)$, which is shaded to the left of $z = 1.33$. Additionally, on the calculator, you can use

2nd normalcdf(-1e10, z, 0, 1)

where on the left is just a really small negative number, the second term is your z-score, and the third and fourth terms are μ and σ. Something that is kind of cool is you can check your work and skip the z-score calculation if you plug in

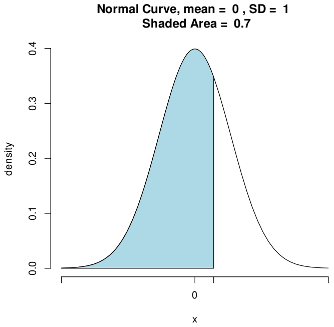
$$\text{2nd normalcdf(-1e-10, 60, 54, 4.5)} = 0.909$$

8.1.2 Z_α Scores

A Z_α score refers to the z-score that will give you an area of α to the *right* of the Z_α score. That's a little confusing at first. For example, $Z_{0.30} = 0.52$, meaning if we want an area of 0.30 to the right of a certain z value, that z value would be 0.52. For this reason, these scores are useful in finding 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

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 $Z_{0.30} = 0.524$. Use the z-table to locate where $1 - \alpha = 1 - 0.3 = 0.70$ and write down the corresponding z-value.

The general procedure for finding percentiles for any normally distributed random variable is given by:

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

Quartiles: Percentiles we have seen before Assume a variable is normally distributed with mean 68 and standard deviation 10, i.e.

$$Y \sim N(68, 10)$$

Find the quartiles:

$$\begin{aligned} Q_1 &= Z_{0.75} = -0.675 \xrightarrow{\text{Transform Y}} Q_1[Y] = -0.675 \cdot (10) + 68 = 61.26 \\ Q_2 &= Z_{0.50} = 0 \xrightarrow{\text{Transform Y}} Q_2[Y] = 68 \\ Q_3 &= Z_{0.25} = 0.674 \xrightarrow{\text{Transform Y}} Q_3[Y] = 0.675 \cdot (10) + 68 = 74.74 \end{aligned}$$

We found the Z_α values from our table, or we can back solve using the calculator with the following command:

$$\text{2nd invNorm}(\alpha, \mu, \sigma)$$

In the first example, we plugged in

$$\text{2nd invNorm(0.75, 0, 1)} \approx 0.674$$

Equivalently, to skip the $z-$transformation if we plug in

$$\text{2nd invNorm(0.75, 68, 10)} \approx 74.74$$

which is actually the final answer conveniently.

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 $Y_{0.15} = -1.35 \cdot 10 + 58 = 54.5$

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. See figure8.1

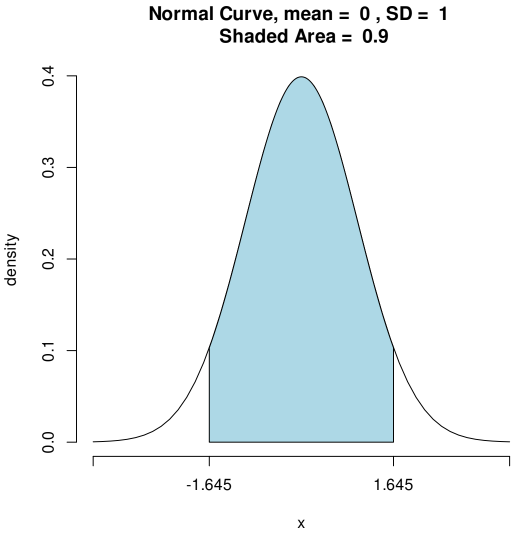


Figure 8.1: Choosing the α's s.t. 90% of the area is in between.

Some practice questions!

1. **The weight of adult jaguars are normally distributed,** $Y \sim N(109, 13)$**. What is the weight that would serve as a cut-point for the top 13% of jaguars?**

   Answer: First, find $Z_\alpha = Z_{1-0.13}$ because Z_α scores are areas to the right. Then

   $$z_{0.87} \approx 1.13 \qquad y = z_{0.87}\sigma + \mu = 123.6$$

   Alternatively, plug in 2nd invnorm(0.87, 109, 13).

2. **The weight of adult jaguars are normally distributed,** $Y \sim N(109, 13)$**. What percentage of jaguars weigh less than 103 pounds?**

   Answer: First, convert y to z

   $$z = \frac{103 - 109}{13} = -0.462$$

   This is our z-score. Find this on the table or just plug in 2nd normalcdf(-1E10, -0.46, 0, 1)=0.322. Then we know 32.2% of jaguars weigh less than 103 pounds. Additionally, you could just plug in 2nd normalcdf(-1E10, 103, 109, 13)=0.322

This is a cute dog.

9

Sampling Distributions(Chapter 13)

Chapter 13: Sampling Distributions, CLT

In this chapter we discuss the concept of statistics being random variables themselves. What does that mean? Well, were we to take every possible sample of size n from a population, and record statistics such as the sample mean, what we expect the different sample means to look like?

We focus mainly on the sample mean and the sample proportion, which (given large enough n) have the following distribution:

$$\overline{Y} \sim N(\mu_Y, \frac{\sigma_Y}{\sqrt{n}}) \tag{9.1}$$

$$\hat{p} \sim N\left(p, \sqrt{\frac{p \cdot (1-p)}{n}}\right) \tag{9.2}$$

This motivates the thinking in chapter 14 and onwards of confidence intervals for our true parameters which we derive from the approximate distributions of our sample statistics.

Example Suppose you are a laboratory scientist measuring the UV-blocking capabilities of different sunglasses. Given that you cannot test all sunglasses, you take a sample of 27 sunglasses and measure the mean blocking capability. We know the true mean blocking capabilities are 44 (no units as this is a fake problem) with standard deviation 23.6, because the FDA tests all sunglasses every year. What is the probability that the sample mean would be greater than 59 (fake units)?

First, because the sample mean is approximately normal

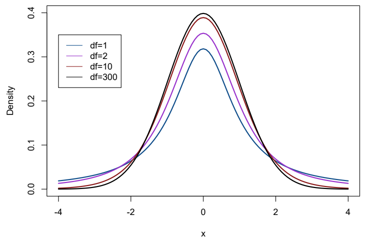
$$\overline{Y} \sim N(44, 23.6/\sqrt{27}) = N(44, 4.54)$$

Then, we use the tools from chapter 11 (tables/calculator) to calculate

$$z = \frac{y - \mu}{\sigma'} \implies \frac{59 - 44}{4.54} = 3.304$$

what α is associated with $z_\alpha = 3.304$? Recall, because we want the probability to be *greater* than 59, we want the right tail probability, so if we use the z-table, remember to do 1-probability from the table, because the z-table reports probabilities to the left of the associated z-value. 2ndnormalcdf(3.304, 1E99, 0, 1)=4.77E-4, or 1-2ndnormalcdf(-1E99, 3.304, 0,1). Alternatively, you could just plug in 2ndnormalcdf(59,1E99, 44, 4.54), where the σ' you plug in is $\sigma/\sqrt{n}$ because we are looking at the distribution of the sampling mean.

We can say with 95% confidence that our confidence interval will be misinterpreted.



10

Confidence Intervals (Chapter 14)

We are interested in one mean in this chapter. This means we are interested in confidence intervals of the form

$$\overline{Y} - \text{critical value} \cdot \text{standard error} < \mu < \overline{Y} + \text{critical value} \cdot \text{standard error}$$

where a critical value is $\left|t_{\alpha/2,\text{df}}\right|$ if σ is unknown and $\left|z_{\alpha/2}\right|$ if σ is known. Confidence intervals get wider at higher confidence levels, which are dictated by $1-\alpha$, where α is specified. See figure 10.1. For testing, see 10.2 for a picture of when to use α or $\alpha/2$. For confidence intervals, we are equivalent in a sense to 2-sided t-tests. For example, if your null hypothesis is $\mu = \mu_0$ at a level α, and if μ_0 is not in your confidence interval, you are equivalently rejecting your null hypothesis. The interpretation of confidence intervals is that at a level α, $(1-\alpha) * 100\%$ of samples taking would have sample means within the confidence intervals. To calculate a critical value, use 2nd invT(α, df) for the t-test or 2nd invNorm(α) if it's a z-test. We use invNorm if you're passing a probability in and want a number out, use 2ndcdf if you want to pass in a number and get a probability. This is how you can tell if you are using the correct one, as probabilities have to be between 0 and 1. Otherwise, using the table, the row gives the df, and the column gives the specified α[1], then find the closest value in the table. Some other stuff from chapter 14:

- **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

[1] divide your given α by 2 if a 2-sided t-test, if a right sided test, not $1-\alpha = -\alpha$, because of the symmetry of the t and z curves, so always just use the α specified for those

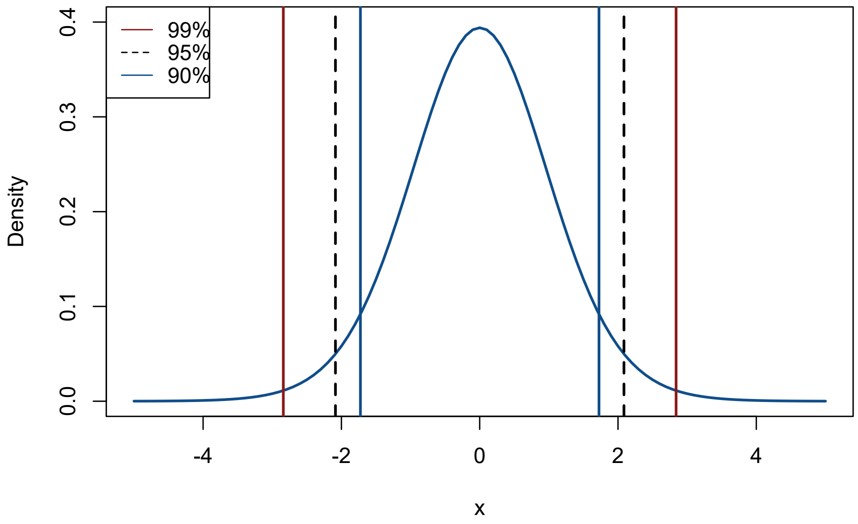


Figure 10.1: Confidence intervals get *wider* at higher % confidence intervals. df=20 and a t-curve is shown here. This plot indicates various critical values. This is the $t_{\alpha/2,20}$ values multiplied by the standard error and subtracted/added from the sample mean to get the confidence intervals.

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

Table 10.1: Top line across the columns indicates whether or not H_0 is true or false.

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

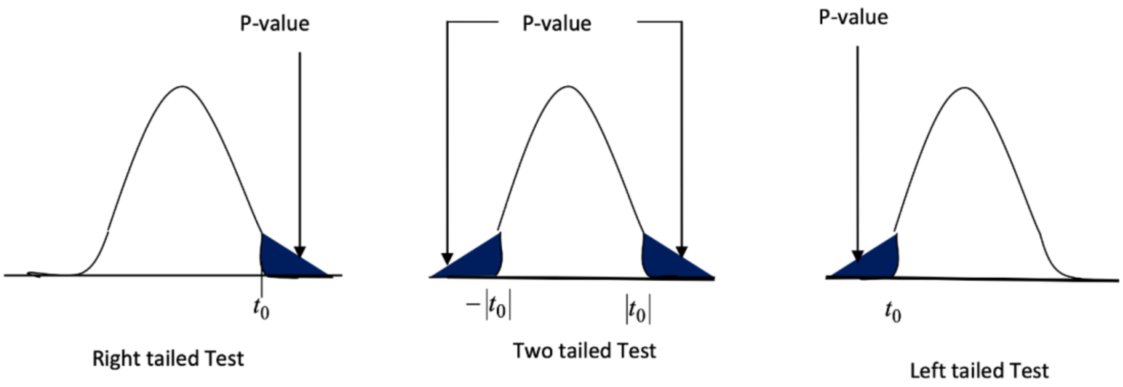


Figure 10.2: Pictoral representation of testing.

Chapter 18: Testing for difference in population means

Independent Samples

If two samples are measured independently of one another (an assumption that is reasonably valid often) we are left with the following equations: (**Please note in the degree of freedom calculation you must round down to the nearest integer)

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

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

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

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

Table 5.1 gives an idea of how to do hypothesis tests and which direction we want to be in:

Test	Null	Alternative	Level	Reject if
Right sided	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$1 - \alpha$	$t^* > t_{1-\alpha,\text{df}}$
Left sided	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	α	$t^* < t_{\alpha,\text{df}}$ or $\lvert t^*\rvert > \lvert t_{\alpha,\text{df}}\rvert$
Two sided	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\alpha/2$	$t^* > \lvert t_{\alpha/2,\text{df}}\rvert$

Table 10.2: How we use tests. This is specifically for 2-samples, but the general idea holds for the one sample, just instead we compare μ to μ_0 and are not looking at a function of two means.

Dependent or Pooled Samples

Dependent or pooled (or paired) samples are in a sense like looking at one sample, because they are dependent on one another (not quite true). This similarity manifests itself in the degrees of freedom, which is $n_d - 1$, where n_d is just the sample size of paired observations.

$$\begin{aligned}
\text{SE of } \overline{d} = \text{SE}_{\overline{D}} &= \frac{S_d}{\sqrt{n_d}} = \frac{\sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}}}{\sqrt{n_d}} \\
\text{df} &= n_d - 1 \\
\overline{d} &= \frac{\sum_{i=1}^{N} d_i}{n} \\
t^* &= \frac{\overline{d}}{\text{SE}_{\overline{d}}} \\
\text{CI for } \mu_d & \; \overline{d} \pm t_{\alpha/2, n_d - 1} \cdot \text{SE}_{\overline{d}}
\end{aligned}$$

1. **We randomly sample 54 ASU students and take their blood temperature. We find a sample mean of 97.4 with a sample standard deviation of 0.5.**

   (a) **State a hypothesis test relating the sample mean of the blood temperature to the "true" 98.6**

   Answer:

   $$H_0 : \mu = 98.6 \qquad H_a : \mu \neq 98.6$$

   (b) **Find a 95% CI for the mean. Interpret this in the context of the hypothesis test...would you reject the NULL?**

   Answer: The confidence interval is

   $$\overline{Y} - \left|t_{.025,53}\right| \cdot \frac{0.5}{\sqrt{54}} < \mu < \overline{Y} - \left|t_{.025,53}\right| \cdot \frac{0.5}{\sqrt{54}} \implies 97.26 < \mu < 97.56$$

   GASP! we must then reject the null hypothesis from the first part. The interpretation is that if we took 100 samples, 95 of them would not contain the true mean of 98.6.

2. **We are studying how much television people born between 1995 and 2000 (zillenials) watch on the daily. We assume the amount of minutes/day is approximately normal and we know from past studies the standard deviation is 3 minutes. If we want a margin of error of 1 minute for a 95% confidence interval, what sample size would we need? How would this change if instead we had a sample standard deviation (guessed) versus the assumed standard deviation?**

   Answer: From the equation in the sheet, since σ is known

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies n = \left(\frac{\sigma}{E}z_{\alpha/2}\right)^2 = 34.57$$

   which we round up to $n = 35$. If we only had a sample standard deviation, we replace $z_{\alpha/2}$ with 2 and σ with 2.

3. **You decide to measure the length of two different types of fish, fish A and fish B. Suppose you want to see if fish A is on average longer than fish B. You measure fish A with a sample mean of 21.2cm with sample standard deviation 2.2. Meanwhile, fish B comes in at 19.6cm with sample standard deviation 2.6cm. You measure 30 of fish A and 24 of fish B. From this data, it is known that degrees of freedom is equal to 45. Use level=0.05**

   (a) **State the hypothesis test:**

   Answer:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 < \mu_2$$

   (b) **Conduct the test. Do you reject the null?**

   Answer: The standard error is

$$\text{SE} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n}} = 0.666$$

   This means our t-test statistic is

$$t^* = \frac{21.2 - 19.6}{0.666} = 2.404$$

   We compare this to $t_{1-0.05,45} = 1.68$. Since this is greater, we reject the null and conclude in favor of the alternative. **BONUS** the p-value is 0.02.

4. **You have been told your whole life that 8 hours of sleep is a healthy and necessary amount for adults. Now, assume you are quite fond of sleep and want to prove that in order to be healthy you need more than 8 hours of sleep. Ignoring what you've learned in this course, you pull together a group of sleep heavy, healthy friends. Of these 23 friends, you find a sample sleep time of 9.2 hours with a sample standard deviation of 1.4 hours.**

(a) Write a hypothesis test about your findings (use significance level 0.05)

Answer: The hypothesis implies a right tailed test:

$$H_0 : \mu = \mu_0 \qquad H_a : \mu > \mu_0$$

where $\mu_0 = 8$.

(b) Calculate a $p-$value for your data.

Answer: Our test statistic is

$$t^* = \frac{9.2 - 8}{1.4/\sqrt{23}} = 4.11$$

Where does this $t-$ value lie? Use 2nd tcdf(-1E99, 4.11, df=22)=0.99976. Because this is a right tailed test,

$$\text{p-value} = 1 - 0.99976 = 2.31E - 4$$

which means we reject the NULL hypothesis at 0.05 significance level because p-value is less than 0.05. Alternatively, $t^* = 4.11 > t_{0.95,22} = 1.71$. **NOTE** for right sided tests, use α, left sided tests use $1 - \alpha$, and for two-sided use $\alpha/2$.

5. **A company wanted to know if attending a course on "how to be a successful salesperson" affects the average sales of its employees. The company sent six of its salespersons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended this course. Standard deviation of the paired differences is 3.366502. Assume that the population of paired differences is normally distributed. Is there any evidence to suggest that the mean weekly sales for all salespersons changed as a result of attending this course? Use $\alpha = 0.01$.**

Before	12	18	25	9
After	18	24	24	14

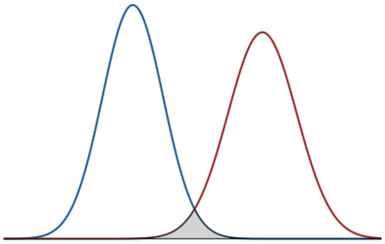
Answer: Our d_i are $6, 6, -1, 5$, meaning $\sum d_i = 16$, $\overline{d} = 4$, and $\sum d_i^2 = 98$. Then

$$\text{SE} = \frac{\sqrt{\frac{98 - \frac{16^2}{4}}{3}}}{\sqrt{4}} \approx 1.683$$

as said in the problem. The null is $H_0 : \mu_1 = \mu_2$ with $t^* = \frac{\overline{d}}{\text{SE}_{\overline{d}}} = 2.376$. Compare this to the critical value at $\alpha = 0.01$, which is $\left|t_{.01/2}, 4 - 1\right| = 5.84$. We do not reject because $t^* < 5.84$.

My happiness grows in direct proportion to my acceptance, and in inverse proportion to my expectations.

Michael J Fox



11

Chapter 19: Confidence Intervals for One Population Proportion

We continue the previous chapters and discuss asymptotic (large n) properties of the sample proportion. When studying these, we are allowed to make inference on the population proportion. We are interested in:

- 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)

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{total number of individuals in the sample}}$$

However, 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

This has standard error

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

and confidence interval

$$\tilde{p} \pm 1.96 \times \mathrm{SE}_{\overline{p}}$$

Equations and Format

● The test will be 12 questions, 10 multiple choice (50 pts) and 2 short response (52 pts) total=102. ●**Goodness of Fit Test**

$$E_i = n \cdot p_i \qquad \text{df}=k-1$$

$$\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i}$$

● **Chi-Square Test**

$$E = \frac{(\text{Row total}) \cdot (\text{column total})}{\text{Grand total}} \qquad \text{df=(r-1)(k-1)}$$

$$\chi^2_* = \sum_{i=1}^{r \cdot k} \frac{[O_i - E_i]^2}{E_i}$$

● **Correlation Coefficient**

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x-\overline{x}}{s_x}\right) \cdot \left(\frac{y-\overline{y}}{s_y}\right) = \frac{1}{n-1} \sum (Z_X \cdot Z_Y)$$

$$t^* = r\sqrt{\frac{n-2}{1-r^2}} \qquad \text{with df}=n-2$$

The coefficient of determination is exactly $r^2 = (r)^2$

The coefficient of determination is also (when fititng regression) $r^2 = \dfrac{s_y^2 - s_e^2}{s_y^2} = 1 - \dfrac{s_e^2}{s_y^2}$

The Fitted Regression Line

$$\hat{y} = b_0 + b_1 x \qquad b_1 = r\left(\frac{s_y}{s_x}\right) \qquad b_0 = \overline{y} - b_1\overline{x} \qquad e_i = y_i - \hat{y}_i$$

$$\text{SS(resid)} = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y})^2 \qquad s_e = \sqrt{\frac{\text{SS(resid)}}{n-2}}$$

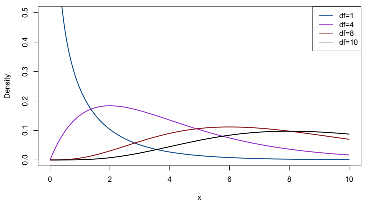
Inference

$$\text{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}$$

$$b_1 \pm t_{0.025,n-2} \cdot \text{SE}_{b_1}$$

$$t^* = \frac{b_1}{\text{SE}_{b_1}} = r\sqrt{\frac{n-2}{1-r^2}} \qquad \text{df=}n-2$$

Can't think of a quote



12

Chapter 21: Chi-Square Goodness of Fit

We now wanna continue studying categorical variables, *but* we now wanna study all proportions/frequencies of variables with two or more categories. To execute studies of this situation, we first introduce the chi-square distribution (aka the χ^2_k distribution, which has k degrees of freedom) 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. $\Gamma(x) = \frac{1}{x}\Gamma(x+1)$ This distribution starts out 0. Is NOT symmetric (at least until high degrees of freedom) It 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! Figure 12.1 shows what different values for the degrees of freedom do to the distribution:

Example Find the following: (use table or calculator). Form is $\chi^2_{\alpha,\text{df}}$ and we want the area TO THE RIGHT! In general, if finding the critical values, find column, go down, then left to row. If finding p-value, like the second two examples below, find row, go to right, and then go up the column (may need to be between columns).

- $\chi^2_{0.05,10} = 18.31$

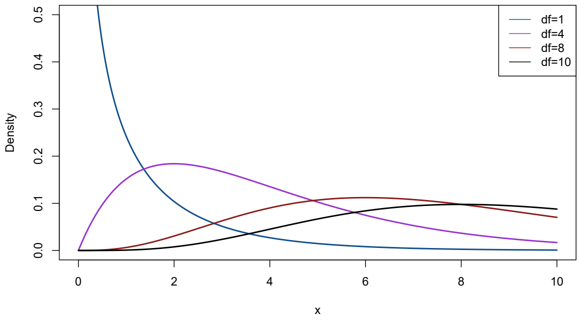


Figure 12.1: Different degrees of freedom for chi-square distribution.

- $\chi^2_{0.005,100} = 140.2$

- $\Pr\left(\chi^2 > 65\right)$, df=50=1-0.925=0.075

- $\Pr\left(\chi^2 > 6.1\right)$, df=17=0.992

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. The frequencies from a sample (O_i) **Observed frequencies** The frequencies from the proposed model are given by: **Expected Frequencies**: Finally, the expected frequency for outcome $i = np_{i_0}$. Call this E_i. O_i is $n \cdot p_i$.

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)

$$\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i}$$

$$\chi^2_* = \sum_{i=1}^{k} \frac{\left[\text{Observed}_i - \text{Expected}_i\right]^2}{\text{Expected}_i}$$

$$\chi^2_* = \frac{\left[\text{observed}_1 - \text{expected}_1\right]^2}{\text{Expected}_1} + \frac{\left[\text{observed}_2 - \text{expected}_2\right]^2}{\text{Expected}_2} + \frac{\left[\text{observed}_3 - \text{expected}_3\right]^2}{\text{Expected}_3} + \ldots + \frac{\left[\text{observed}_k - \text{expected}_k\right]^2}{\text{Expected}_k}$$

Intuitively, $\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i}$ is the sum of squared deviances from H_0, so small χ^2_*

would imply there is little difference from the null and observed data, while large implies the opposite. In fact, 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. After comparing observed and expected, we perform a chi-square goodness of fit test instead just say the expected does not equal the observed. The statistics behind it is that we wanna say if we are within reasonable uncertainty . To conduct a test we must first check that our assumptions are correct:

****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
$$\chi^2_* = \sum_{i=1}^{k} \frac{[O_i - E_i]^2}{E_i} \tag{12.1}$$
where the degrees of freedom is equal to $k - 1$.

The steps when we test are:

- 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

If $p-$value is less than or equal to α, reject the null hypothesis. See figure 12.2. Notice, the area we care about is to the right!

Example 2 Find the expected frequencies:

The proportions for the model are:

$$p_{\text{white}} = \frac{12}{16} \qquad p_{\text{yellow}} = \frac{3}{16} \qquad p_{\text{green}} = \frac{1}{16}$$

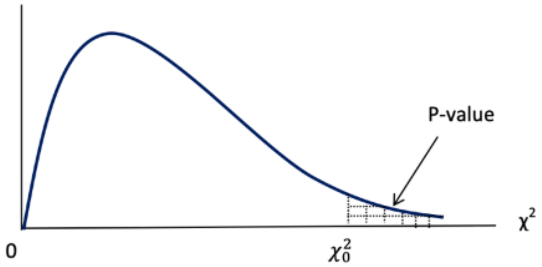


Figure 12.2: Notice the area we care about is *to the right*.

Color	White	Yellow	Green
Number of progeny	155	40	10

Now, we wanna do a χ^2 goodness of fit test 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=0.708. In calculator, we want 1-2nd χ^2cdf(0, 0.69, 2), because we want area to right.

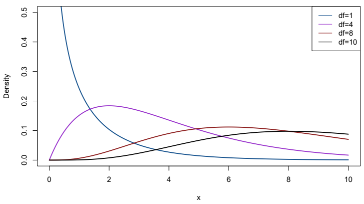
Example Does fire affect dear behavior? Six months after a fire burned 730 acres of deer habitat, researchers surveyed a 3,000 acre parcel surrounding the area, which they divided into four regions: 1) inner burn, 2) inner edge, 3) outer edge, 4) outer unburned. The null hypothesis is that show no preference to any particular type of burned/unburned habitat (i.e. the deer are randomly distributed across the regions)

So the χ^2_* statistic is 43.151. Degrees of freedom is 3. Find the p-value

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

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=43.151

can't think of a quote



13 Chapter 22: Chi-Square Test of Association

We shift our attention to multiple categorical variables. Our goal here is to see whether or not these variables are associated with one another, similar to the last chapter, but with some differences.[1]

We follow the following procedure when testing for associations:

- 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

Like the previous chapter, We have observed cell frequencies from sample data (O). We calculate expected cell frequencies assuming that the row variable and column variable are independent. The main difference is now we go across rows and down columns because we care about the interactions as well. Let R_j be the row totals and C_m be the column

[1] As we've discussed, **association** just means two variables have some relation. **Correlation** refers to a linear association between variables, and has a strict mathematical formulation. Finally, **causation**, the Holy Grail of sorts in statistics, refers to the concept of one variable *causing* another. This we do not discuss in this course too much, as it is fairly difficult to quantify causal effects. If interested, a course in causal inference may be useful.

totals for the specified cells. Expected cell frequency=$\frac{R_m \times C_j}{n}$, where n is the sample size. We call this E_i. 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

$$\chi^2_* = \sum_{i=1}^{r\cdot k} \frac{[O_i - E_i]^2}{E_i}$$

$$\chi^2_* = \sum \frac{\left[\text{Observed}_i - \text{Expected}_i\right]^2}{\text{Expected}_i}$$

$$\chi^2_* = \frac{\left[\text{observed}_1 - \text{expected}_1\right]^2}{\text{Expected}_1} + \frac{\left[\text{observed}_2 - \text{expected}_2\right]^2}{\text{Expected}_2} + \frac{\left[\text{observed}_3 - \text{expected}_3\right]^2}{\text{Expected}_3} + \ldots + \frac{\left[\text{observed}_{rk} - \text{expected}_{rk}\right]^2}{\text{Expected}_{rk}}$$

which follows a $\chi^2_{(r-1)(k-1)}$ distribution, i.e. df=$(r-1)(k-1)$. 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:

$$\Pr(\text{row variable and column variable}) = \Pr(\text{row variable}) \times \Pr(\text{column var})$$
$$= \frac{R_j}{n} \times \frac{C_m}{n} = \frac{R_j \cdot C_m}{n^2}$$

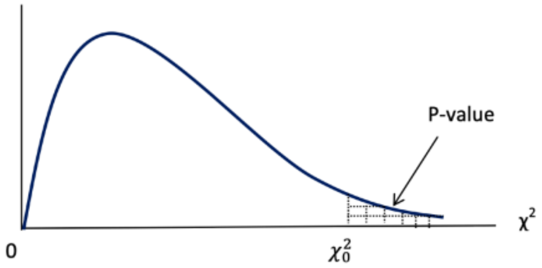
To turn this into a frequency, we would multiply by n. Therefore, the expected frequency if H_0 was true:

$$n \times \frac{R_j \cdot C_m}{n^2} = \frac{R_j \cdot C_m}{n}$$

This gives a sorta intuitive idea behind why we are doing what we are doing. Furthermore, $\chi^2_* = \sum \frac{[O_i - E_i]^2}{E_i}$ is the sum of squared deviances from H_0. Small χ^2_* would imply there is little difference from the null and observed data, while 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. For the p-value approach, we follow the usual procedure:

- 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

Recall, however, when using the chi-square table we are looking for the **area to the right**. If $p-$value is less than or equal to α, reject the null hypothesis



χ_0^2 is the critical value. Use 1 - 2ndcdf χ^2cdf(0, χ_*^2, df) or 2ndcdf 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.

What are the assumptions we make?

- 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

$$\chi_*^2 = \sum_{i=1}^{r \cdot k} \frac{[O_i - E_i]^2}{E_i} \tag{13.1}$$

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.

Example 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	$\boxed{246}$

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

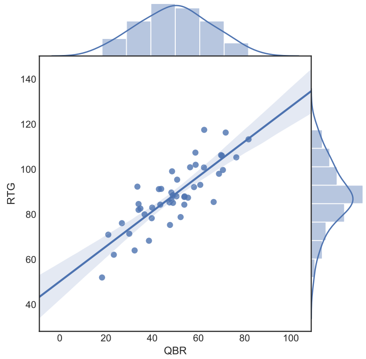
We make our usual table, except now we must account for the row-column pairs (all 4 of them)

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. Using the table (or with R 1-pchisq(11.21, 1)) we get a p-value of 0.00082. Therefore, we reject the null hypothesis in favor of the alternative.

What's the difference between chapter 21 and chapter 22? In chapter 21, we simply look at whether or not there is some effect across multiple categories at the same time. However, in chapter 22, for each of the categories, we have different variables as well and are looking for associations amongst the variables. Then we care about the interactions between the categories/the variables in a sense. z

With regression we were led to believe we can predict the future. We were misled.

Person



14

Bivariate Data, Linear Correlation, and Regression(Chapter 4)

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. $b_1 > 0$ means positive linear relationship, $b_1 < 0$ means negative linear relationship, and 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. What is the correlation (sample)? Call it r. The **correlation** contains information of the direction (positive or negative) and strength (weak, moderate, or strong) of linear association. 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, i.e.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

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

We can do hypothesis testing on ρ: ρ is the population correlation coefficient. We assume simple random samples and normal distributions. Our hypothesis is:

$$H_0 : \rho = 0 \text{ } x \text{ and } y \text{ are uncorrelated in the population}$$
$$H_a : \rho \neq 0 \text{ } x \text{ and } y \text{ are correlated in the population}$$

First specify the level you want, α (usually $\alpha = 0.05$). The test statistic is (t-distribution with df=n-2)

$$t^* = r\sqrt{\frac{n-2}{1-r^2}}$$

We 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. Other variables could also influence response

The **coefficient of determination** is denoted by r^2, where $0 \leq r^2 \leq 1$. It determines the percentage of variation in the observed values of the repsonse variable that is explained by the regression line The interpretation: "r^2 of the variability in the response variable can be explained by the explanatory variable" However, there is no tried and true way of telling if you have a "good" r^2. Importantly, the coefficient of determination is the value of the correlation squared.

To calculate the equation of the least squares line, we first denote what the regression line is for. Think of it as 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. [1] It 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.

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

$$\hat{y} = b_0 + b_1 x$$

with slope:

$$b_1 = r\frac{s_y}{s_x} = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

The intercept is

$$b_0 = \overline{y} - b_1\overline{x} = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

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 $\hat{y}$.

Since we fit a line to the data, we are likely not fitting perfectly, which leads us to the concept of **residuals**, which we can think of as the distance from a predicted response and observed response

$$e_i = y_i - \hat{y}_i$$

Residuals sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

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. 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:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n-2}}$$

The measurement tells us how far above or below the regression line the data points tend to be. These values (the residuals) should be normally distributed around zero with some standard deviation. In fact, we can also tie the residual standard deviation back to the Empirical Rule. We expect:

- 68% of the observations (y-values) to be within ± 1 residual standard deviation of the regression line

[1] In the sense, the parameters are linear, not the explanatory variable necessarily. I.e. the slope b_1 is a linear term.

- 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

There is also a relationship between the residuals and the correlation: 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

$$r^2 = \frac{s_y^2 - s_e^2}{s_y^2}$$

- 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

Reference figure 14.1 Let (sometimes see m and b instead of b_1 and b_0)

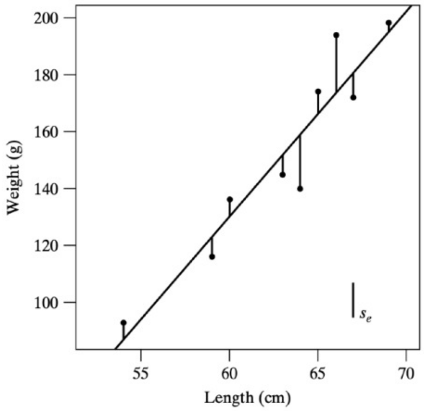


Figure 14.1: Vertical distance between line & points, not perpendicular!

$$\Delta y_i = y_i - \hat{y}(x_i) = y_i - (b_1 x_i + b_0)$$

We want the sum over all the i from 1 to n to be as small as possible. Therefore, it makes sense to minimize $y_i - (b_1 x_i + b_0)$ right? Well, sorta. Because that term can be negative or positive, we will never find a true local minimum, so we need to make sure the term is always positive. One option is the absolute value, but that function is mathematically trickier for a variety of reasons than taking the *square* of that value. Instead, then we aim to minimize the function

$$\begin{aligned} S(b_1, b_0) &= \sum_{i=1}^{N} (\Delta y_i)^2 \\ &= \sum_{i=1}^{N} \left[y_i - \hat{y}(x_i) \right]^2 \\ &= \sum_{i=1}^{N} \left[y_i - (b_1 x_i + b_0) \right]^2 \end{aligned} \tag{14.1}$$

Where N is the total number of observations, and b_1 and b_0 are the slope and intercept terms we will solve for. To do this, we set each derivative equal to zero (if you haven't taken calculus you can skip this part)

$$\begin{aligned} 0 &= \frac{\partial S(b_1, b_0)}{\partial b_1} \\ &= \frac{\partial}{\partial b_1} \sum_{i=1}^{N} \left[y_i - (b_1 x_i + b_0) \right]^2 \\ &= \sum_{i=1}^{N} 2 \left[y_i - (b_1 x_i + b_0) \right] (-x_i) \\ 0 &= \sum_{i=1}^{N} (x_i y_i) - b_1 \sum_{i=1}^{N} (x_i)^2 - b_0 \sum_{i=1}^{N} x_i \end{aligned}$$

and for the intercept term:

$$\begin{aligned} 0 &= \frac{\partial S(b_1, b_0)}{\partial b_0} \\ &= \frac{\partial}{\partial b_0} \sum_{i=1}^{N} \left[y_i - (b_1 x_i + b_0) \right]^2 \\ &= \sum_{i=1}^{N} 2 \left[y_i (-b_1 x_i + b_0) \right] (-1) \\ 0 &= \sum_{i=1}^{N} y_i - b_1 \sum_{i=1}^{N} x_i - N b_0 \end{aligned}$$

Solving these simultaneously yields:

$$\begin{aligned} b_1 &= \frac{N \sum x_i y_i - \sum x_i \sum y_i}{N \sum x_i^2 - \sum {x_i}^2} \\ b_0 &= \frac{\sum {x_i}^2 \sum y_i - \sum x_i \sum x_i y_i}{N \sum x_i^2 - \sum {x_i}^2} \end{aligned}$$

In matrix form,

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \varepsilon \\ E(\boldsymbol{y}) &= \boldsymbol{X\beta} \\ \boldsymbol{y} &\sim N(\boldsymbol{X\beta}, \boldsymbol{\sigma^2 I}_n) \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ \hat{\boldsymbol{y}} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{y} = \boldsymbol{Hy} \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the estimator. Note, the $\boldsymbol{\beta}$ and $\boldsymbol{y}$ are obvious, but the $\boldsymbol{X}$ matrix isn't:

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \dots & X_{1,p-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \dots & X_{n,p-1} \end{pmatrix}$$

Where $X_{i,j}$ X_i refers to the the ith data point, and X_j refers to the the "x estimate". Ie, if we want to predict peoples weight based on their weight and height, $X_{1,1}$ would be the weight of person 1, $X_{1,2}$ would be the height of person 1 and so on. We can also use this matrix form to add non linear terms. In the case of a single variate quadratic regression, this would look like

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1}^2 & X_{1,2} \\ 1 & X_{2,1}^2 & X_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{n,1}^2 & X_{n,2} \end{pmatrix}$$

Notice that multiplying this by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

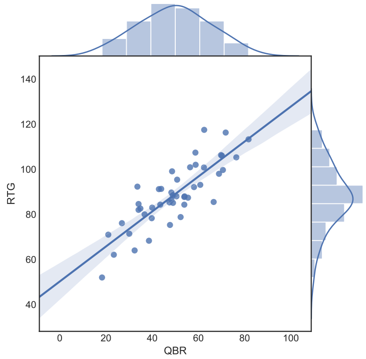
will give us $\boldsymbol{y}$, which is $n \times 1$, X in this case is $n \times 3$ and $\boldsymbol{\beta}$ is 3×1. If we are minimizing least square error, ie running a specific regression, $\boldsymbol{\beta}$ becomes an estimator, $\hat{\boldsymbol{\beta}}$.

and for least squares the normal equations in matrix mode are

$$\begin{aligned} n\hat{\beta_0} + \hat{\beta_1}\sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ \hat{\beta_0}\sum_{i=1}^{n} X_i + \hat{\beta_1}\sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}$$

With regression we were led to believe we can predict the future. We were misled.

Person



15

Chapter 23: Inference for Regression

We can also think of inference from regression. Yes, that means hypothesis testing! Because we mentioned how residuals should be approximately normally distributed (given sufficient n), we have a basis for how to evaluate inference on our slope and correlation. Previously, our regression line is calculated from statistics $\overline{x}$, $\overline{y}$, s_x, s_y, and r change if different samples are chosen and so does the line $\overline{y} = b_0 + b_1 x$. Because these vary from sample to sample, we want to implement statistical inference/ To perform inference we think of b_0 and b_1 as estimates of the regression parameters that describe the entire population. 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

Figure 15.1 gives a good visual of this idea:

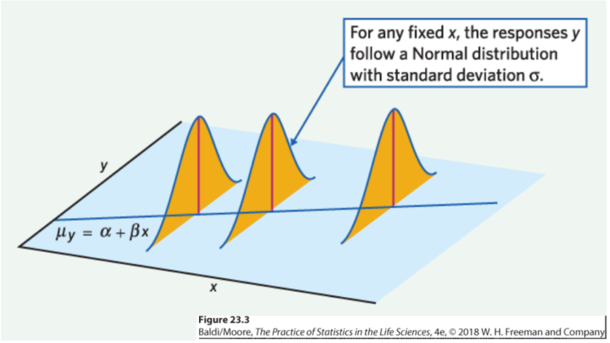

For any fixed x, the responses y follow a Normal distribution with standard deviation σ.
y
$\mu_y = \alpha + \beta x$
x
Figure 23.3
Baldi/Moore, The Practice of Statistics in the Life Sciences, 4e, © 2018 W. H. Freeman and Company

Figure 15.1: Pictoral representation of regression line at various x.

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, and s_e estimates unknown standard deviation σ_y

We can also test hypothesis of no linear relationship. We test the hypothesis:

$$H_0 : \beta = 0 \text{ The slope is zero, no linear relationship}$$
$$H_A : \beta \neq 0 \text{ The slope is not zero, there is linear relationship}$$

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

$$\text{SE}_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (x - \overline{x})^2}} = \frac{s_e}{s_x \times \sqrt{n-1}} = \frac{\text{residual std dev}}{\text{std dev of } x \times \sqrt{n-1}}$$

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)

$$\text{p-value} = 2 \cdot \Pr\Big(T > \big|t^*\big|\Big)$$

From this, we can also calculate confidence intervals using similar thinking as previous chapters: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.

Linear relationship described by slope b is closely related to r

- When the correlation is 0, the slope will be exactly 0

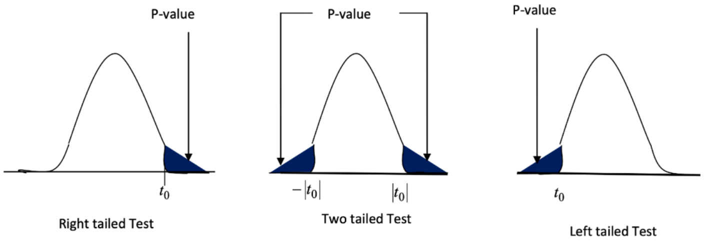


Figure 15.2: The now familiar photo

- 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:

$$H_0 : \rho = 0 \text{ there is no correlation}$$
$$H_a : \rho \neq 0 \text{ there is correlation}$$

We can test of lack of correlation (alternative) by looking at 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 $\leq \alpha$.

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

$$\text{SE}_{\hat{\mu}} = s_e\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_{i=1}^{n}(x - \overline{x})^2}}$$

Similarly, but not quite the same, 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

$$\mathrm{SE}_{\hat{y}} = s_e\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{\sum_{i=1}^{n}(x-\overline{x})^2}}$$

The prediction interval thus gives interval for y, whereas the confidence interval gives an interval for $E(y \mid x)$. In that sense, even with a perfect regression where we estimate $E(y \mid x)$, the "average of y" perfectly, we would still have some uncertainty about y itself, so our prediction interval would have some width, despite the confidence interval being just a point. See here if confused by that. Note, because of the +1, the prediction interval will *always be wider* than the confidence interval.

Example Imagine we have the age of phone (x variable) and want to see the response of price (y variable) to age (dollars vs years)

age	6	6	6	2	2	5	4	5	1	4
price	270	260	275	405	364	295	335	308	404	305

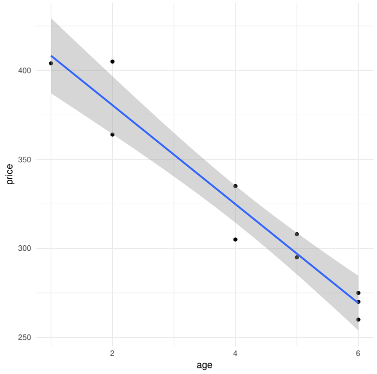


Figure 15.3: Plotting data from example

Find the slope and intercept

$$\begin{aligned}
\overline{x} &= 4.1 \qquad \overline{y} = 322.1\\
s_x &= 1.85 \qquad s_y = 53.25\\
r &= -0.967499 \quad r^2 = 0.9361\\
\sum_{i=1}^{n}&(y-\hat{y})^2 = 1631.7\\
b_1 &= r\frac{s_y}{s_x} = -0.9675\cdot\frac{53.25}{1.85} = -27.8\\
b_0 &= \overline{y} - b_1\overline{x} = 322.1 - (-27.8)\cdot 4.1 = 436.1\\
\hat{y} &= 436.1 - 27.8x
\end{aligned}$$

Test the hypothesis of no linear correlation at significant level 0.05. $t^* = r\sqrt{\frac{n-2}{1-r^2}}$

$$t^* = -10.82 \text{ compare to } t_{.025,10-2} = |t^*| = 10.82 > |t_{.025,10-2}| = 2.31$$

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)

Some practice questions!

1. **We investigate whether screen time is associated with migraine levels. We ask 72 people with varying screen time (the explanatory variable, in hours per day) and the migraine level of those people (on a scale of 0-100).We fit a regression with the line $\hat{y} = 2.21x - 3.31$ with $r = 0.76$, $s_e = 2.67$, $s_x = 0.51$, and $n = 72$.**

   (a) **We want to test whether or not $\beta = 0$. Calculate the test statistic and the p-value at $\alpha = 0.05$.**

   Answer: Note, $t^* = \frac{b_1}{\text{SE}_{b_1}}$ where

   $$\text{SE}_{b_1} = \frac{s_e}{s_x\sqrt{n-1}} = 0.621$$

   So

   $$t^* = 3.55$$

   Now, we can compare this to $t_{0.975,68} = 1.995$, and since $|t^*| > t_{0.975,68}$, we reject the null hypothesis. The p-value is 0.0007, which is less than 0.05, so we reject (recall you must multiply by 2).

   (b) **Perform a test of whether or not there is a linear correlation.**

   Answer: The null is $H_0 : \rho = 0$ meaning x and y are uncorrelated in the population. The test statistic is

   $$t^* = r\sqrt{\frac{n-2}{1-r^2}} = 0.76 \cdot 12.69 = 9.64$$

   This will also be rejected.

2. **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 $\alpha = 0.005$ What are the degrees of freedom for this example?**

	Red Wine	White Wine	Row Total
Red Meat	86	46	132
Fish & Poultry	50	64	114
Column Total	136	110	$\boxed{246}$

   (a) **2**

   (b) **1** ✓

   (c) **4**

	Yes	No	Row Total
Millenial	71	29	100
Zillenial	34	47	81
Gen Z	11	61	72
Column Total	116	137	$\boxed{253}$

(d) 3

3. **We are studying whether or not people born between 1995 and 2000 (zillenials), 1985-1995 (millenials), and 2000+ (gen-zers) read daily.**

   (a) What is the expected count for millenial yes readers?

   Answer:

   $$E = (100 * 137)/253 = 45.85$$

   (b)

4. **You are given the equation line $\hat{y} = 2x + 3$. This has $r^2 = 0.88$ and was fit between $x = 1$ and $x = 10$ (the domain). Now say I ask you to give me the predicted value at $x = 20$. What is that?**

   (a) 3

   (b) 40

   (c) 43

   (d) Since 20 is outside the domain of x, wouldn't recommend extrapolating $\checkmark$