Chapter 1 Notes

Picturing Distributions with Graphs
STP-231

Arizona State University

Introduction
Key Terms

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

Individuals
Potential Outcomes

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

Variables

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

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

Quantitative Variables

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

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

Quantitative Variables
Examples

Are these discrete or continuous?

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

Categorical Variables
Examples

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

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

Categorical Variables
Examples

Are these numeric or categorical?

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

Exploratory Data Analysis

Frequency Distributions

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

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

Table 2.2.1 Color of 182 poinsettias

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

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

Relative Frequency Distributions

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

- Let n be the sample size, then

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

- Multiply by 100% to represent as a percentage

Frequency Distributions
Example

- Representing the same data as relative frequencies/percentages

Table 2.2.1 Color of 182 poinsettias

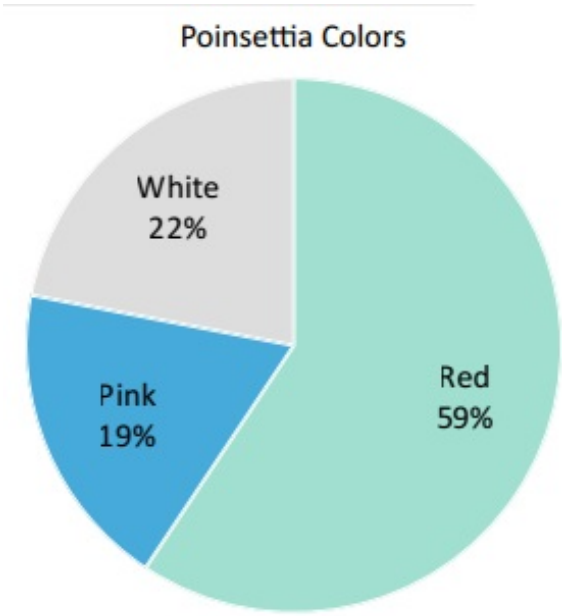
Color	Frequency (number of plants)	Relative Frequency	Percentage
Red	108	0.593	59%
Pink	34	0.186	18%
White	40	0.219	21%
Total	182	0.998?	99.8?

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

Displaying Categorical Data
Objective: To Organize Categorical Data

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

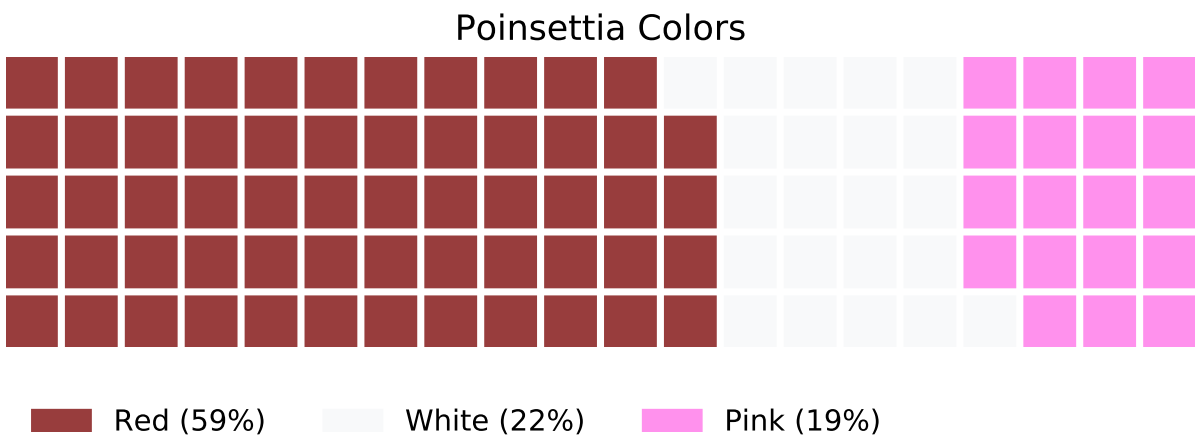
Pie chart



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

Waffle Chart

Counting rectangles easier than seeing percentage on circle



Bar Graph

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

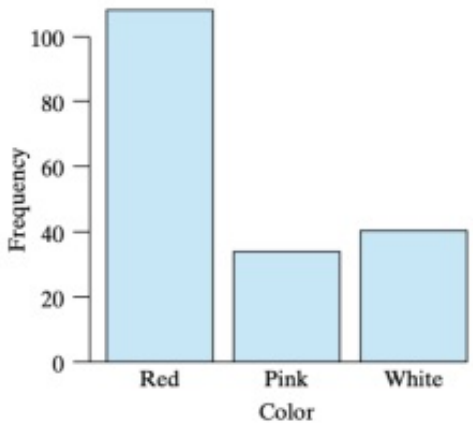
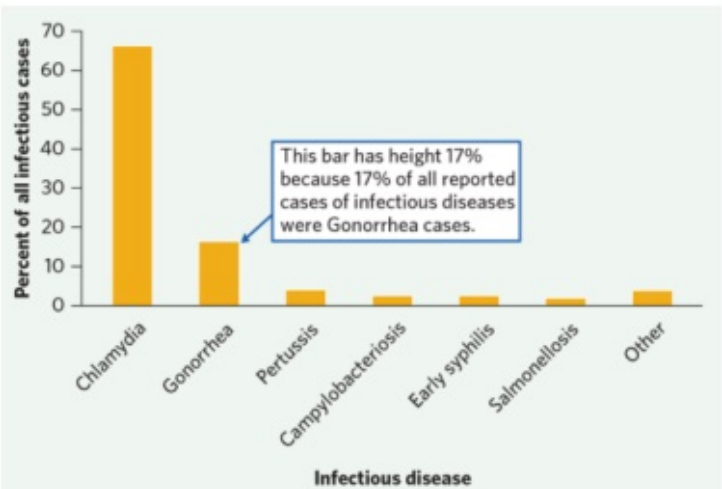


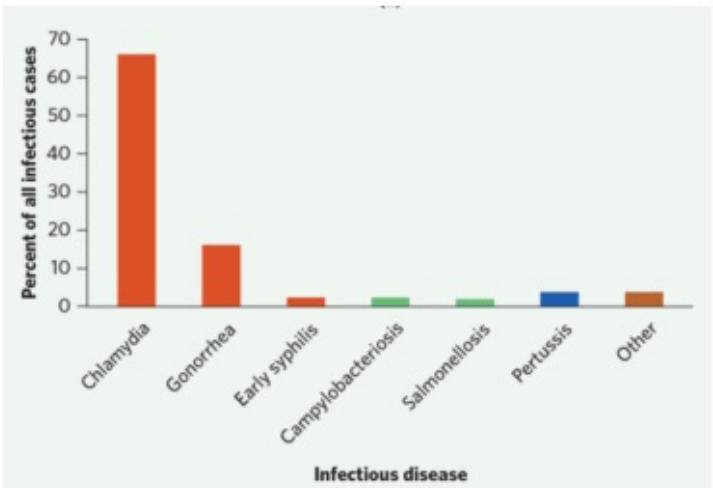
Table 2.2.1 Color of 182 poinsettias

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

Example

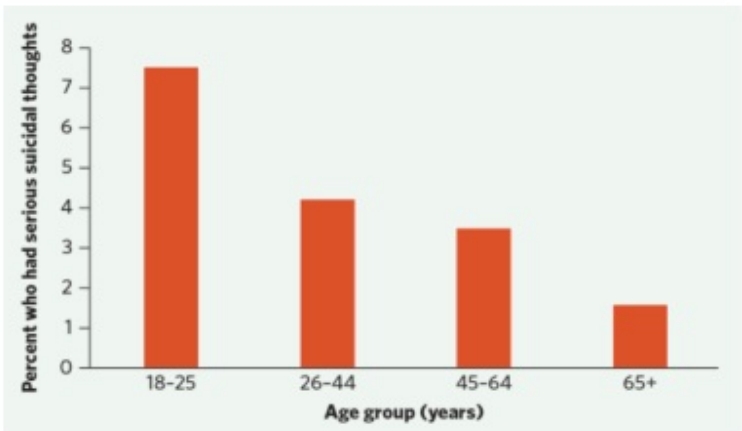
- In 2014, California had a total of 262,780 reported infectious cases.

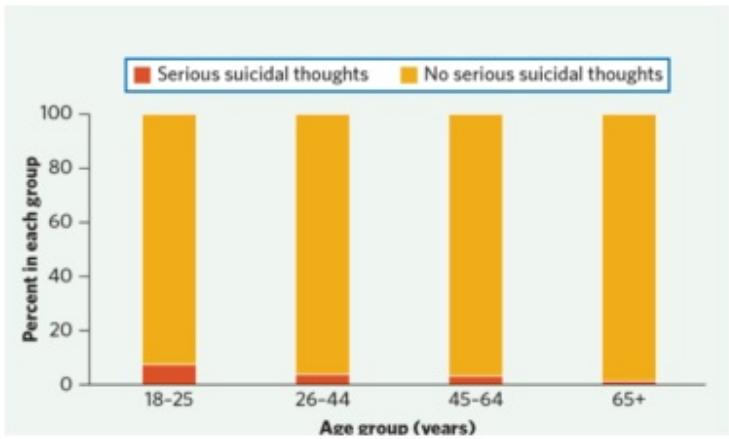




Bar Charts vs Pie Charts

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





Bar Charts vs Pie Charts
Continues

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

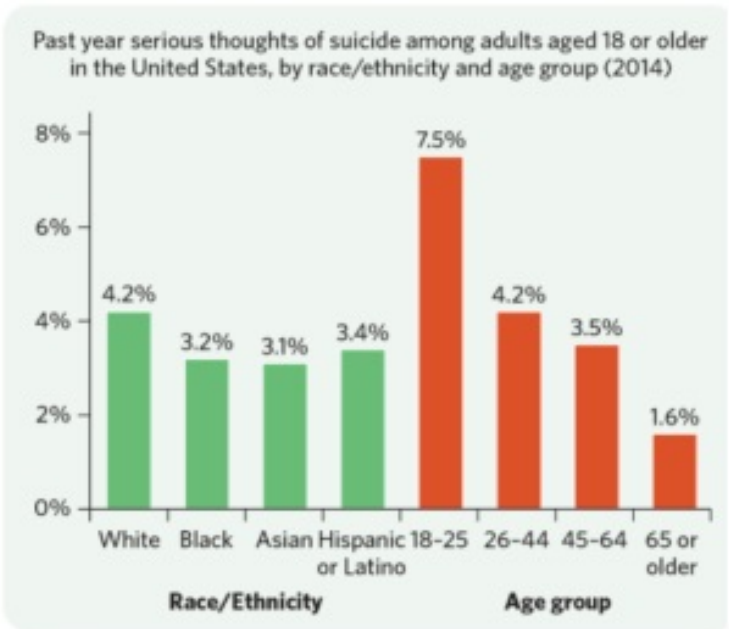
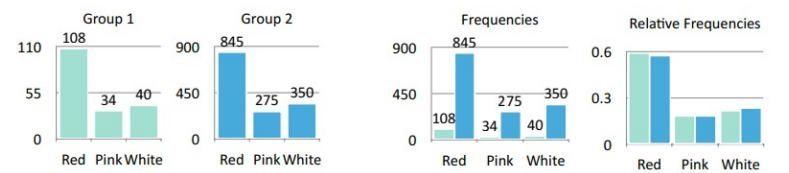


Figure 1.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Relative Frequency to Compare Groups

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



Displaying Quantitative Data
Organize Quantitative Data

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

Stem and Leaf Plot

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

Stem and Leaf Plot
One line per stem

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

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

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

Stem and Leaf Plot
Two lines per stem

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

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

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

Stem: Ones

Leaf: First decimal place

Example

Example 7 from Elementary Statistics by Neil A. Weiss

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

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

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

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

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

Dot Plot

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

Table 2.2.3 Infant mortality in seven South Asian countries

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

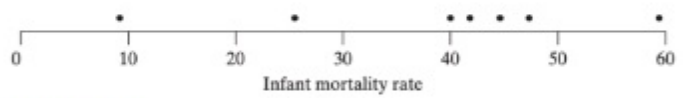
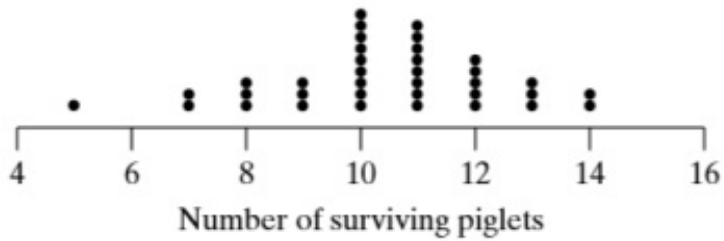


Figure 2.2.3 Dotplot of infant mortality in seven South Asian countries

Dot Plot
Example

Table 2.2.4 Number of surviving piglets of 36 sows

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



Histogram

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

Dot plot to histogram

Table 2.2.4 Number of surviving piglets of 36 sows

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

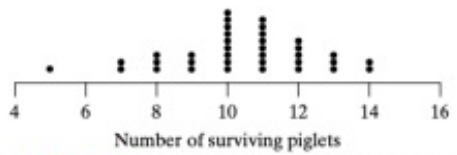


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

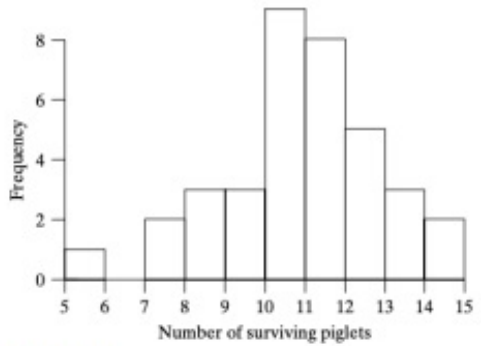


Figure 2.2.5 Histogram of number of surviving piglets of 36 sows

Grouped Frequency Distributions

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

Grouped Frequency Distributions
Example

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

Table 2.2.6 Serum CK values for 36 men

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

Hist. without grouping:

Grouped Frequency Distributions
Example (continued)

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

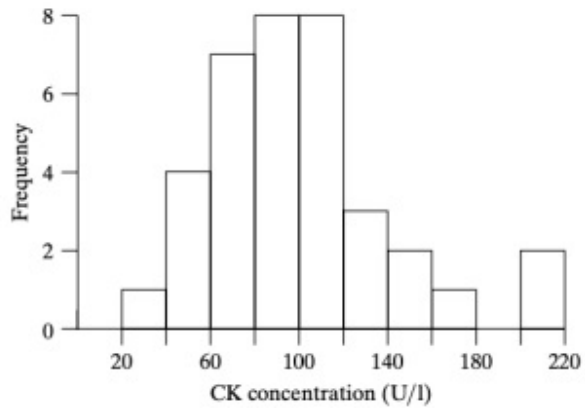


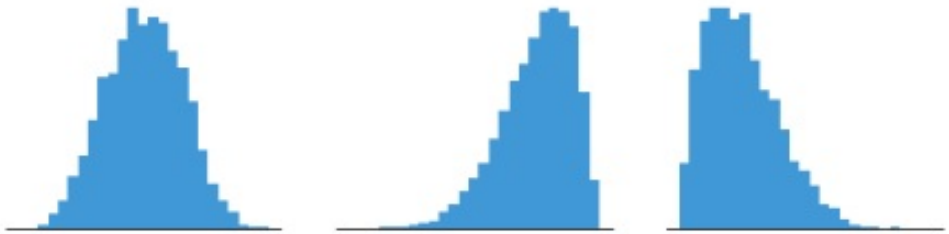
Table 2.2.7 Frequency distribution of serum CK values for 36 men

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

Interpreting Histograms

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

Modality



Unimodal - One Peak

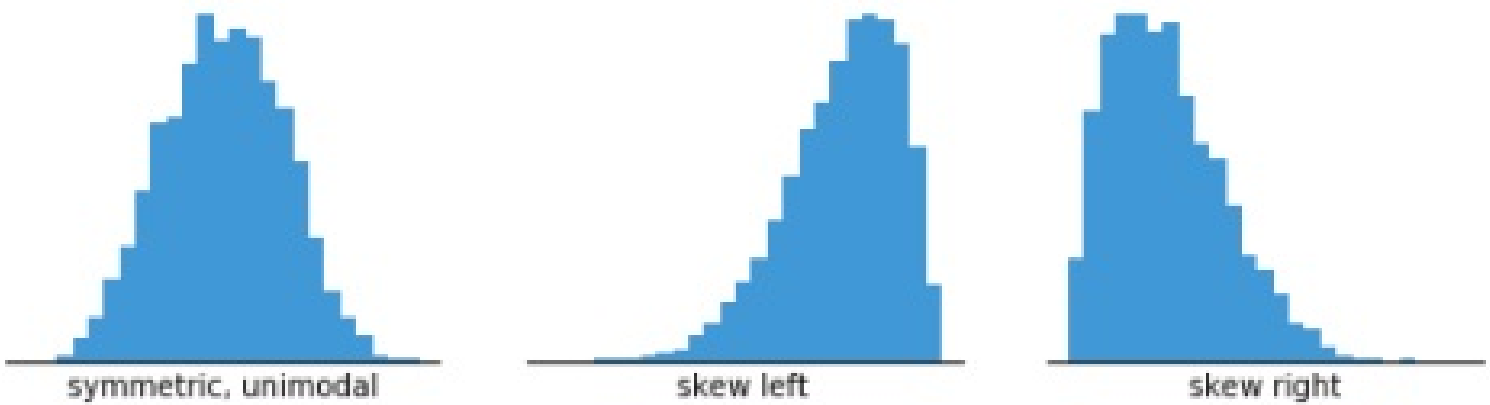
Uniform

Bimodal - Two Peaks

Multimodal - Several Peaks

Skewness

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



Cut Point Grouping

Definitions: A class is the same as a bin.

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

Cut Point Grouping
Continued

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

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

Cut Point Grouping
Procedure

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

Cut Point Grouping
Example

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

Table: Ages of 20 people living in a retirement home

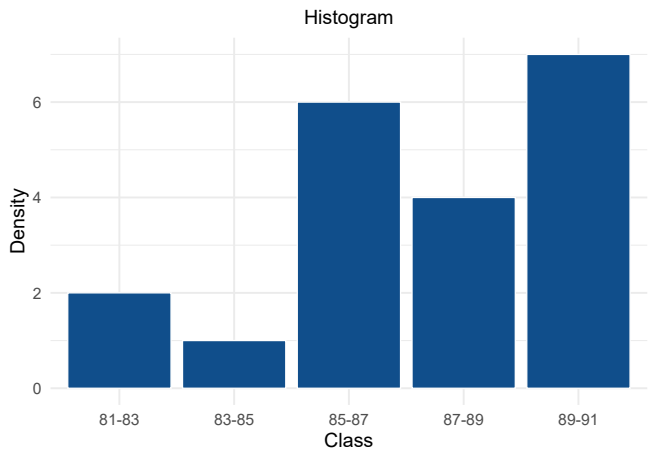
- Assume 5 classes (bins). We have

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

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

Cut Point Grouping
Example

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



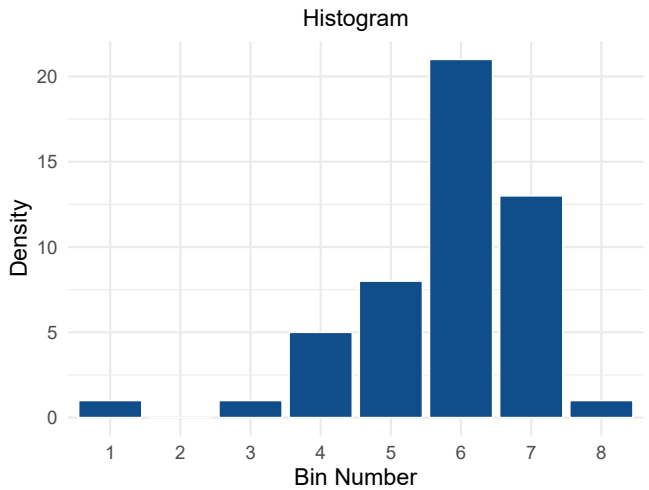
Cut Point Grouping
Example

Researcher want to know how often students buy used textbooks. They ask 50 graduates how many used textbooks they purchased during their undergraduate study. Results are summarized in the following table. Given that one of the intervals is 8 to 11, complete the frequency and relative frequency distribution table

13	16	21	17	12	13	17	19	19	20
17	16	14	11	20	16	22	20	15	17
18	21	19	19	19	18	22	3	17	18
19	17	20	20	13	14	17	22	21	16
19	23	21	17	16	17	8	18	20	17

Example continued

Bin	Class	Frequency	Relative Frequency
1	2-under 5	1	0.02
2	5-under 8	0	0
3	8-under 11	1	0.02
4	11-under 14	5	0.1
5	14-under 17	8	0.16
6	17-under 20	21	0.42
7	20-under 23	13	0.26
8	23-under 26	1	0.02



New Example

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

48.8	59.9	60.7	59.0	55.9	56.5	54.4	60.6	62.1	59.9
57.1	58.9	58.1	56.6	55.2	57.7	55.8	56.2	57.9	56.5

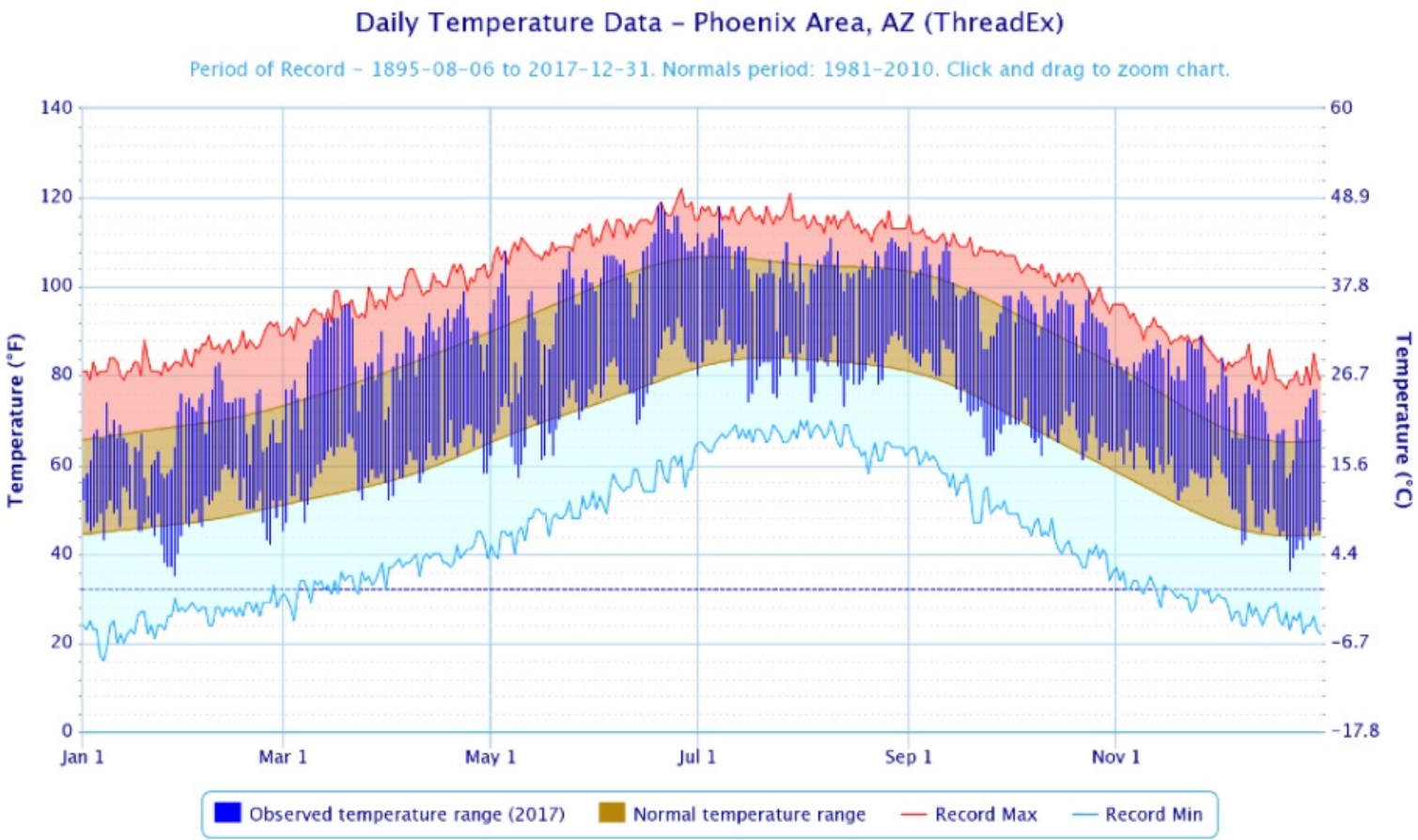
New Example

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

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

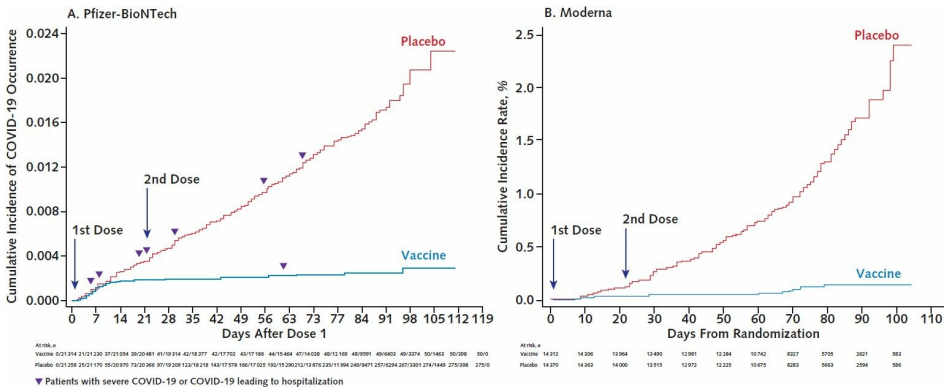
Time Plots

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



Comparing two Time plots

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



US Air Travel 2020 vs 2019

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