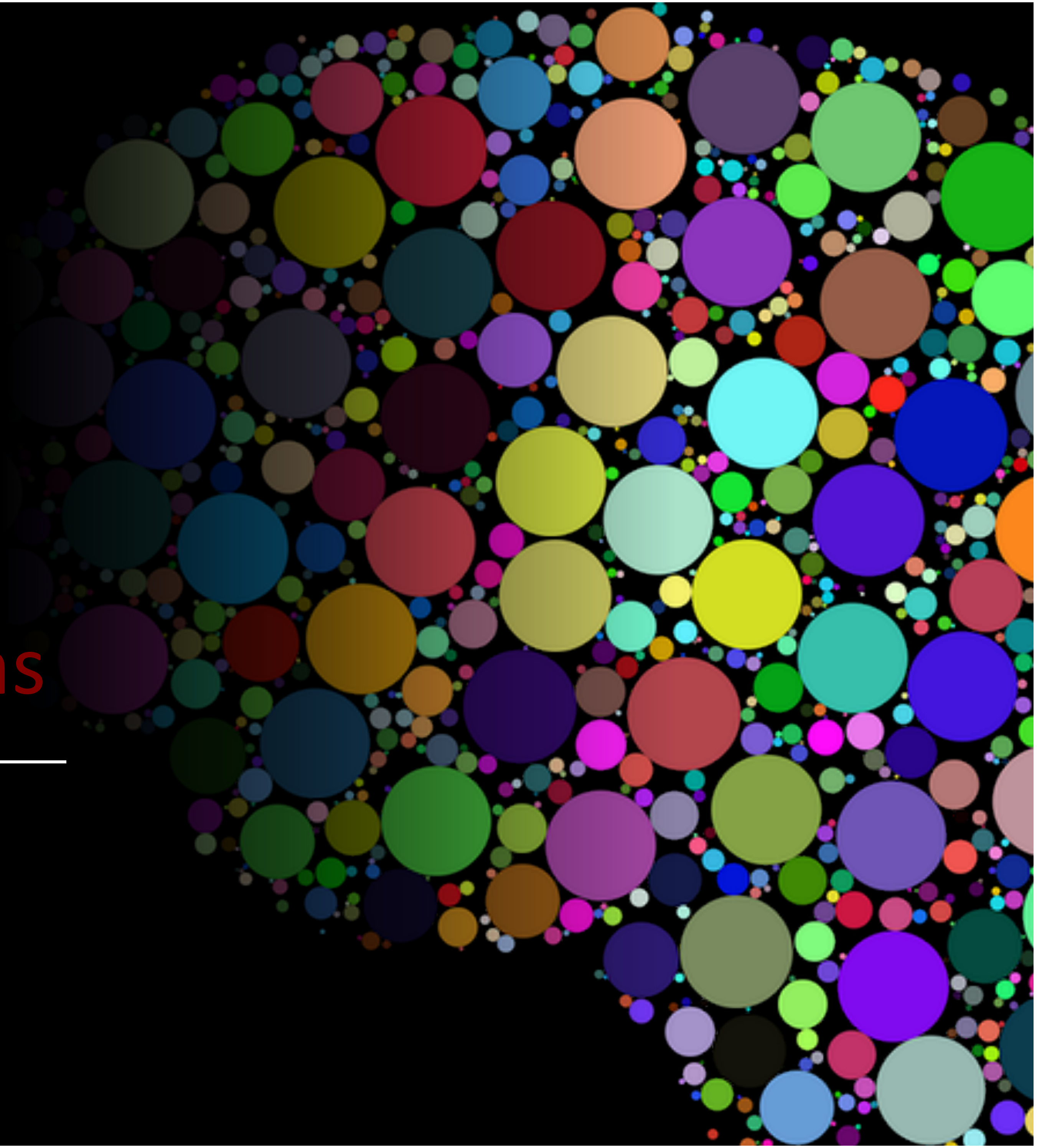
Frequency Distributions

Chapter 2



Descriptive Statistics

- Distributions are part of descriptive statistics…we are learning how to describe some data by first graphing it in a meaningful way.

Descriptive Statistics

- Frequency distribution – describes the pattern of a set of numbers by displaying a count/percent for the values of the variables.

Frequency Distributions

- Four different ways to visually describe just one variable:
  - Frequency table
  - Grouped frequency table
  - Frequency histograms
  - Frequency polygon

Open up R!

- We are going to use some built in data in R to build frequency tables and histograms.
  - Use data(airquality) to load the airquality dataset
  - At first it will say *promise* next to it, but once we use it, it will pop up and be viewable in the environment window.

Frequency Tables

- Frequency tables often include:
  - Values (all the possible numbers)
  - Frequency (how many times each number appears)
  - Percent/proportion
- Why percent when we have frequency?

Frequency Tables

- Grouped frequency tables are frequency tables where information has been clumped together.
  - For example, ABCD breakdowns instead of each grade individually. Or income ranges rather than each income separately.
  - Very useful for data with decimals and wide ranges of values.

How-To R

- Frequency tables:
  - table(*column name*) function
  - The table function requires you to give it *vector/column* of data.
  - Remember that if that column of data is in a dataframe, then you have to tell R where the data is stored.

How-To R

- Don’t do this:
  - table(airquality)
    - Airquality is a bunch of columns.
  - table(Temp)
    - R doesn’t know that Temp is hiding in airquality
- Do this:
  - table(airquality$Temp)

How-To R

- What if you wanted percentages instead of raw frequencies?
  - You can divide the table by N!
  - table(*column name*) / length(*column name*) * 100
  - The length() function tells you how many items are in that function.

How-To R

- table will give you a frequency table of each individual value … what about grouping them?
- stem(*column name*, scale = *#*)
  - Use this function to try making a stem and leaf plot, which will allow you to group like values together.
  - For the scale = #, try playing around with the numbers until you get the grouping you think is best.

Histograms

- Histograms are frequency tables in graph form (basically they are turned on their side and made into a chart).
- Gives you an idea of the shape of the distribution.

Frequency Polygon

- Frequency polygons are histograms that show a smoothed line instead of the bars for a histogram.
  - So why use these?
  - They often give you a better picture of the data, since histograms can be changed based on binwidth (more on this idea in a minute).

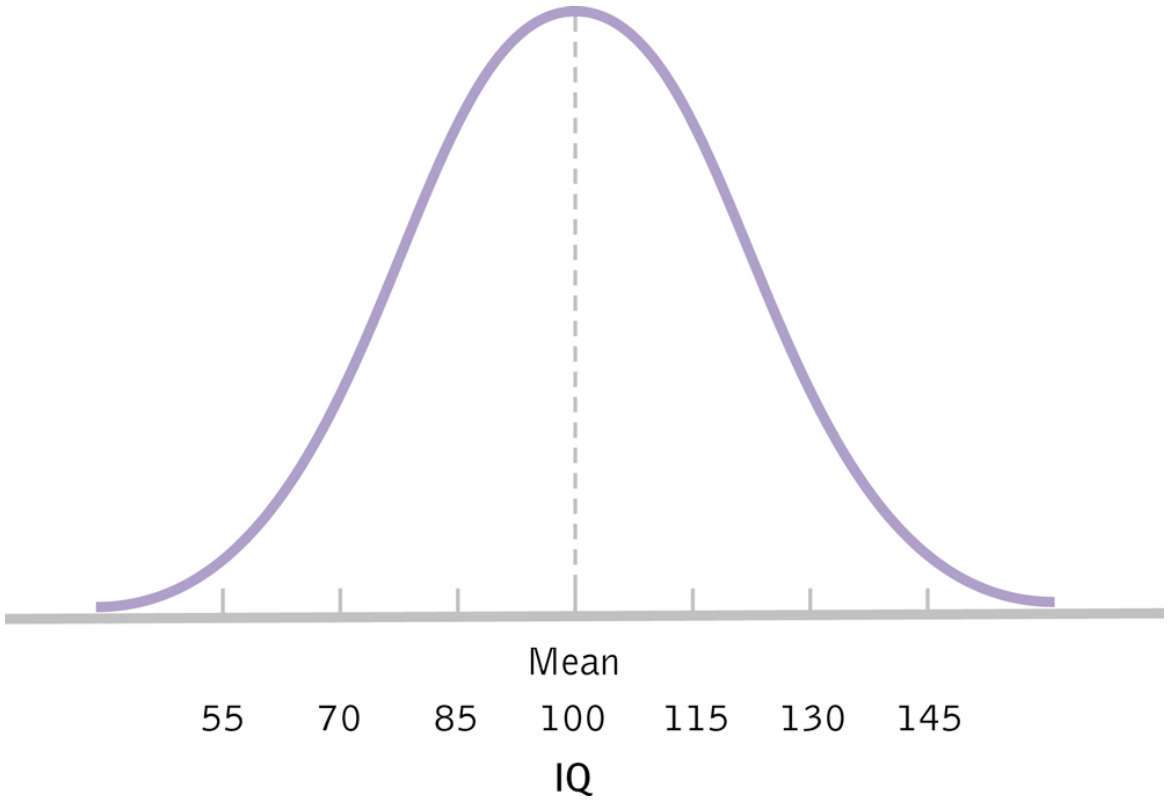
Shapes of Distributions

- Unimodal
  - (normal distribution)
- Bimodal
- Multimodal
- Rectangular

Shapes of Distributions

- Normal distributions: Specific frequency distribution
  - Bell shaped
  - Symmetrical
  - Unimodal

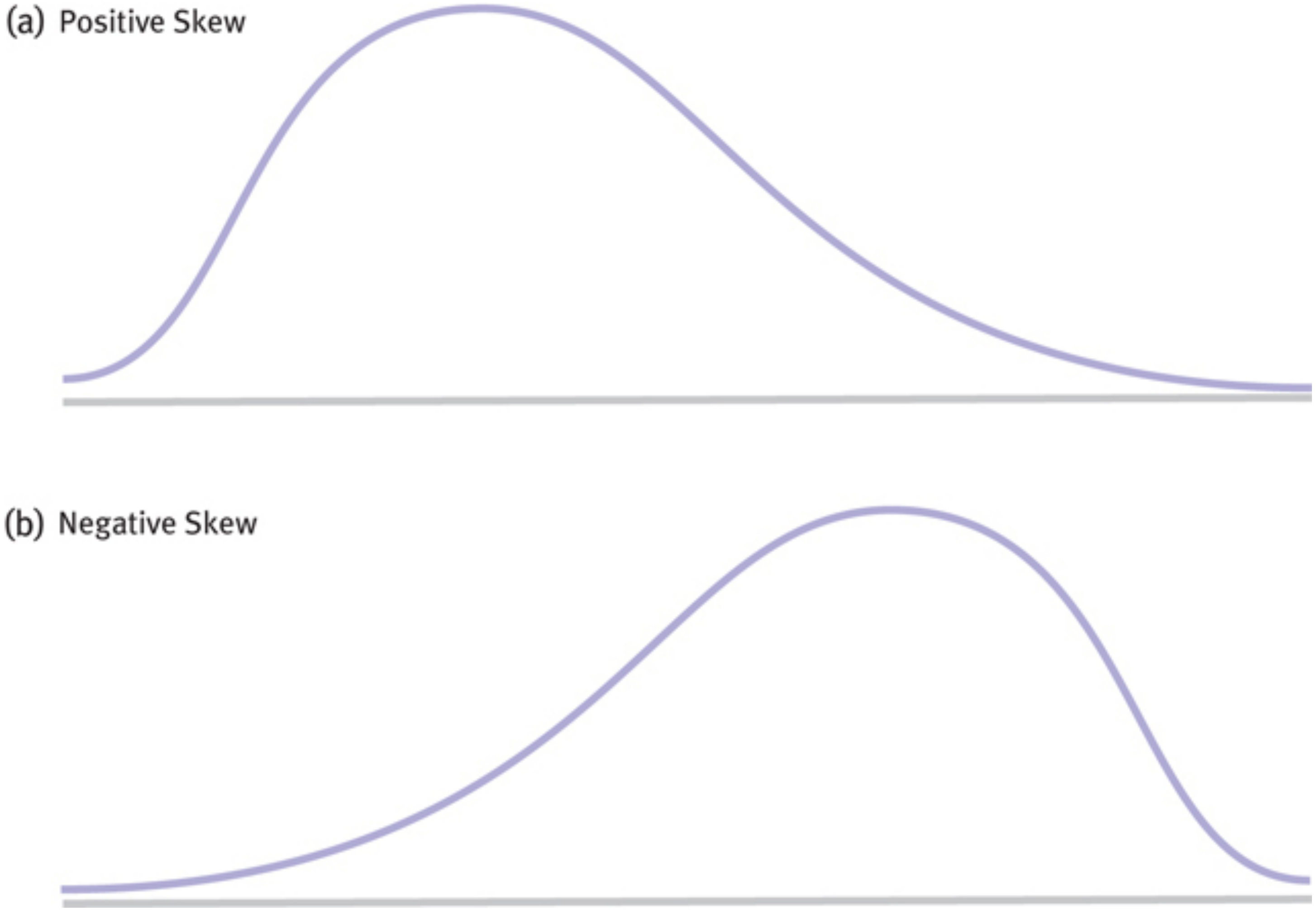
The Normal Distribution



Skewed Distributions

- When our data are not symmetrical
  - Positive: tail to the right
    - May represent floor effects
  - Negative: tail to the left
    - May represent ceiling effects
  - Memory hint: skew is where the tail is (the cat!)

Two Kinds of Skew

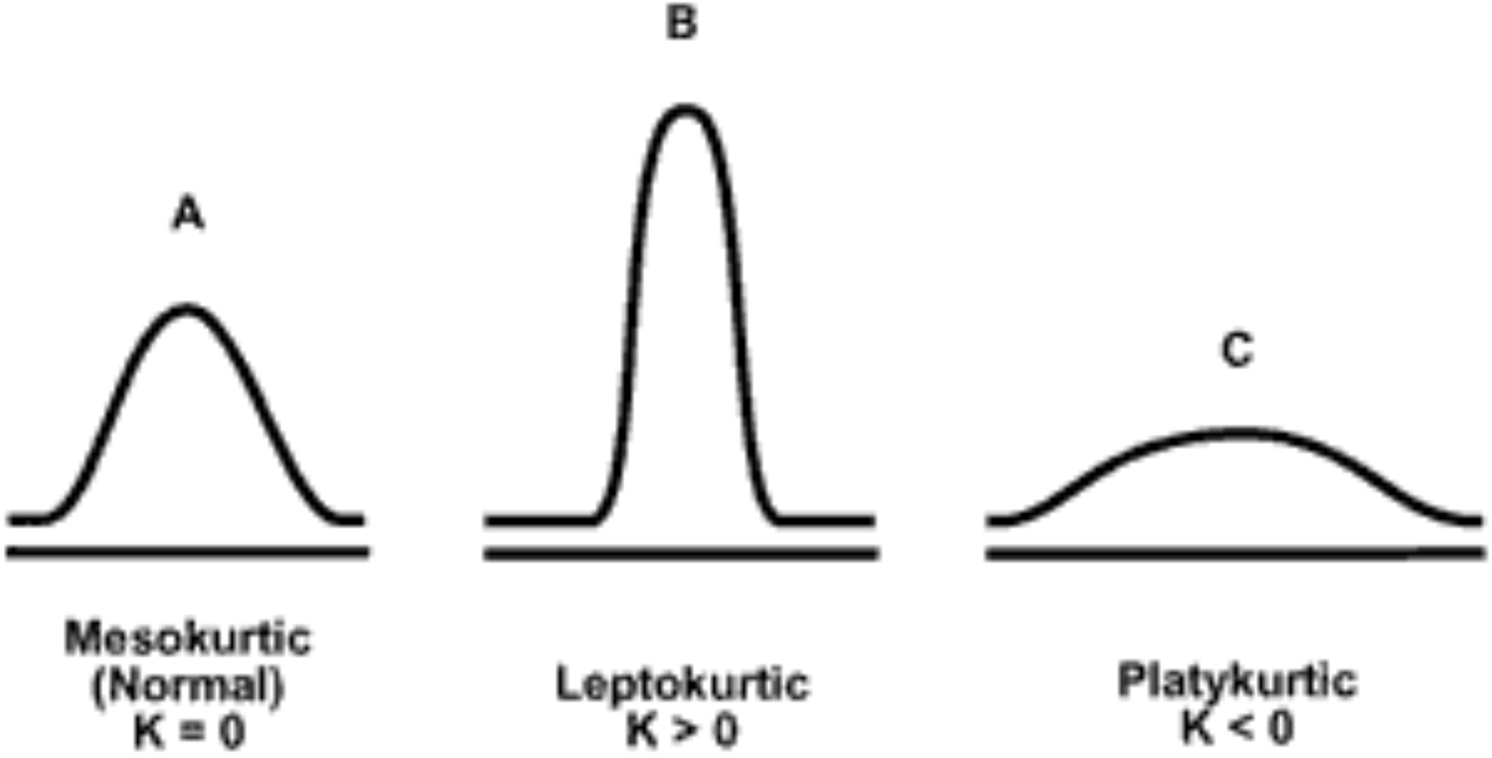


(a) Positive Skew

(b) Negative Skew

Kurtotic Distributions

- Kurtosis – how much a distribution deviates from normal by looking at spread



How-To R

- First, install the ggplot2 package.
  - Click packages in the bottom right window > install > type in ggplot2
  - Wait for it to do its thing…
- Load the library
  - library(ggplot2)
  - Or you won’t get very far…

How-To R

- Ggplot2 is a package that does lots of cool graphing, which we will use a lot in chapter 3.
- It requires several steps to make a plot.
  - Think about it like a transparency.
  - You first tell it what you want to use in your plot.
  - Then you draw the pieces of the plot one line at a time.

How-To R

- Basic histograms:
  - First, include the variables you want to use.
  - Save the plot.
- Something like this:
  - myplot = ggplot(*dataset*, aes(*column name*))
    - Note: here because you have told it the dataset, you do not have to do the $ thing. In table(), we did because we hadn't mentioned the dataset.

How-To R

- Now, let’s add things to myplot
- myplot + geom_histogram()
  – Geom_histogram creates a histogram.
  – Run!
  – Play with binwidth = X

How-To R

- Let’s try making a frequency polygon
- myplot + geom_freqpoly()

How-To R

- So what’s a good binwidth?
- 5-15? (seems to work for me)