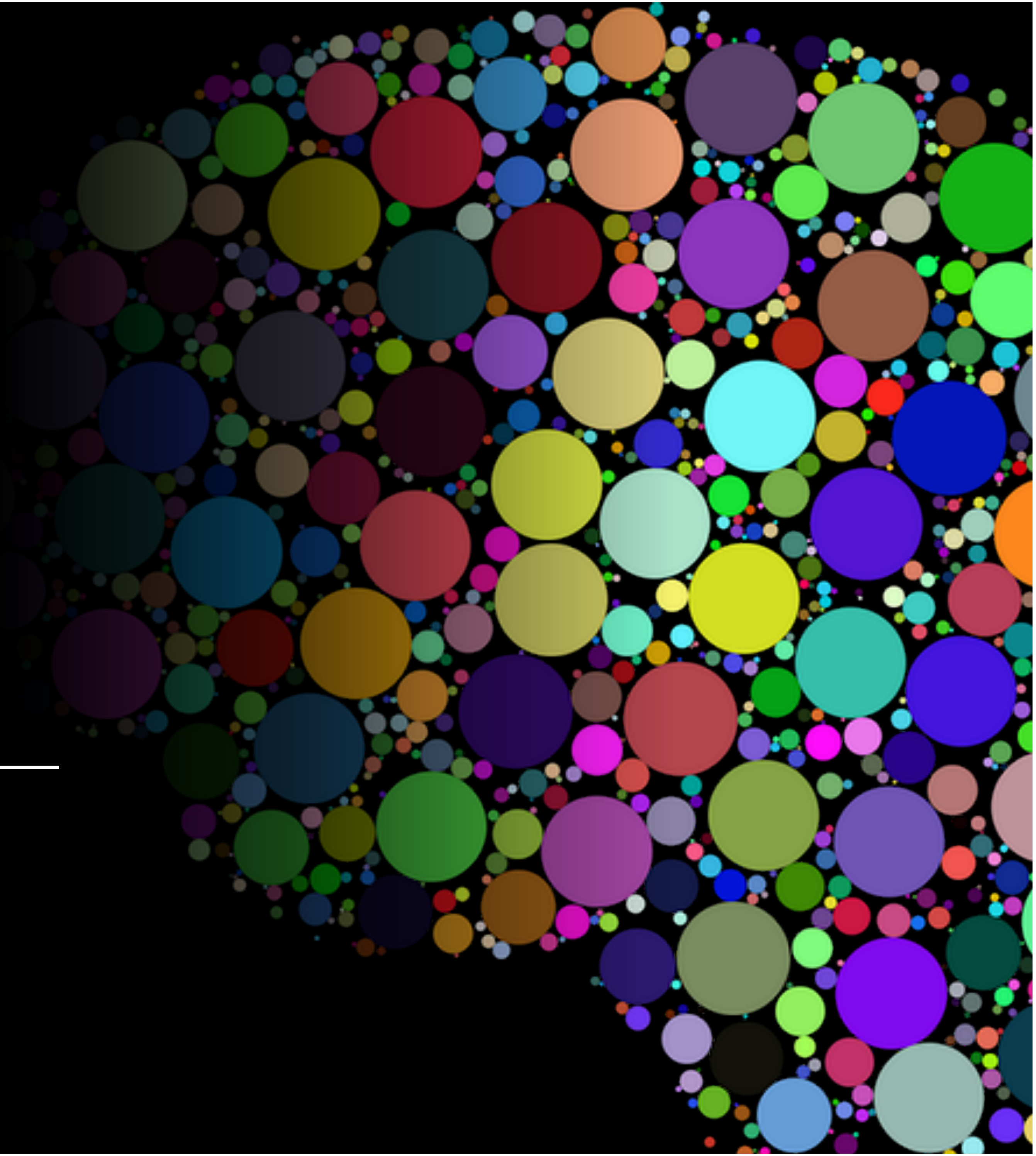
Central Tendency and Variability

Chapter 4



Variability

- In reality – all of statistics can be summed into one statement:
  - Variability matters.
  - (and less is more!, depending).
  - (and error happens).

Central Tendency

- Definition: descriptive stat that best represents the center of a distribution of data.
- Mean: arithmetic average
  - “Typical score”
  - Often described as the “middle” of the scores, so don’t confuse this with medians.

Calculating the Mean

- Add up all scores
- Divide by number of scores

$$\overline{X} = \frac{\sum X}{N}$$

Traditionally, we use *M*
as the symbol for sample means.

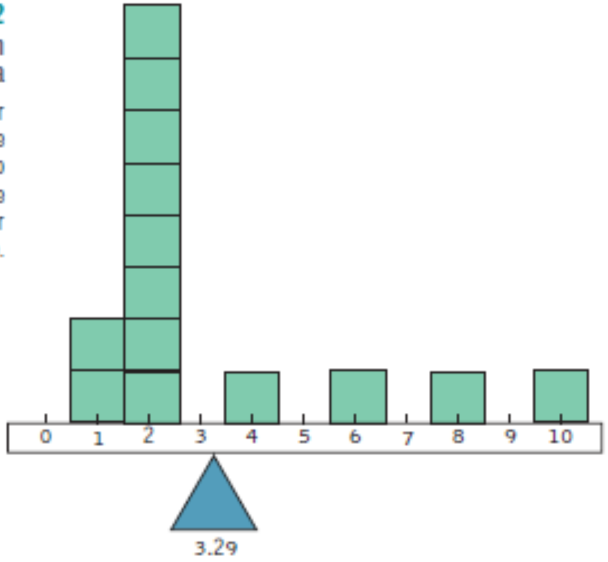


FIGURE 4-2
The Mean as the Fulcrum of Our Data

The mean, 3.29, is the balancing point for all the scores for top finishes for the countries that competed in World Cup soccer tournaments. Mathematically, the scores always balance around the mean for any sample.

Note on Symbols

- Usually Latin letters (normal alphabet) are used for samples
  - *M, SD*
  - Sample statistics
- Greek letters are used for populations
  - *μ, σ*
  - Population parameters.
- All statistical letters are italicized.

Get some data!

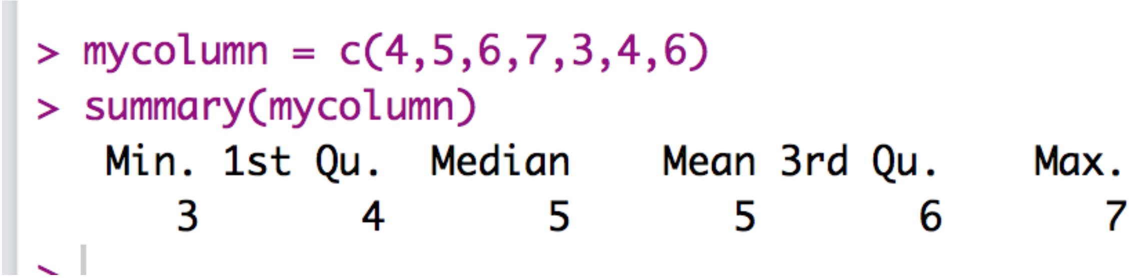
- Go to http://www.sporcle.com/games/RobPro/animal-logos
- Take the quiz!
- Give your score!

How-To R

- Entering raw data.
  - You can enter the data from the board by creating your own individual columns of data.
  - mycolumn = c(*#, #, #*)
  - How to reference that column? You do NOT need the $ operator.
    - Why not?

How-To R

- How to calculate the mean:
  - Two ways:
  - summary(*column name*)
  - mean(*column name,* na.rm = T)



```
> mycolumn = c(4,5,6,7,3,4,6)
> summary(mycolumn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      3       4       5       5       6       7 
>
```

Central Tendency

- Median: middle score when ordered from lowest to highest
  - No real symbol, but you can abbreviate *mdn*

Calculating the Median

- Line up the scores in ascending order
- Find the middle number
  - For an odd number of scores, just find the middle value.
  - For an even number of scores, divide number of scores by two.
  - Take the average of the scores around this position.

How-To R

- You will get the median with the summary() function.
- Or you can use:
  - median(*column*, na.rm = T)

Central Tendency

- Mode: most common score
  - It’s the value:
    - With the largest frequency (or percent on a table).
    - The highest bar on a histogram depending on binwidth.
    - The highest point on a frequency polygon.
- Note…sometimes there are multiple modes.

Calculating the Mode

- Line up the scores in ascending order.
- Find the most frequent score.
- That’s the Mode!

Aka, book notes can be silly sometimes.

Mode + Distributions

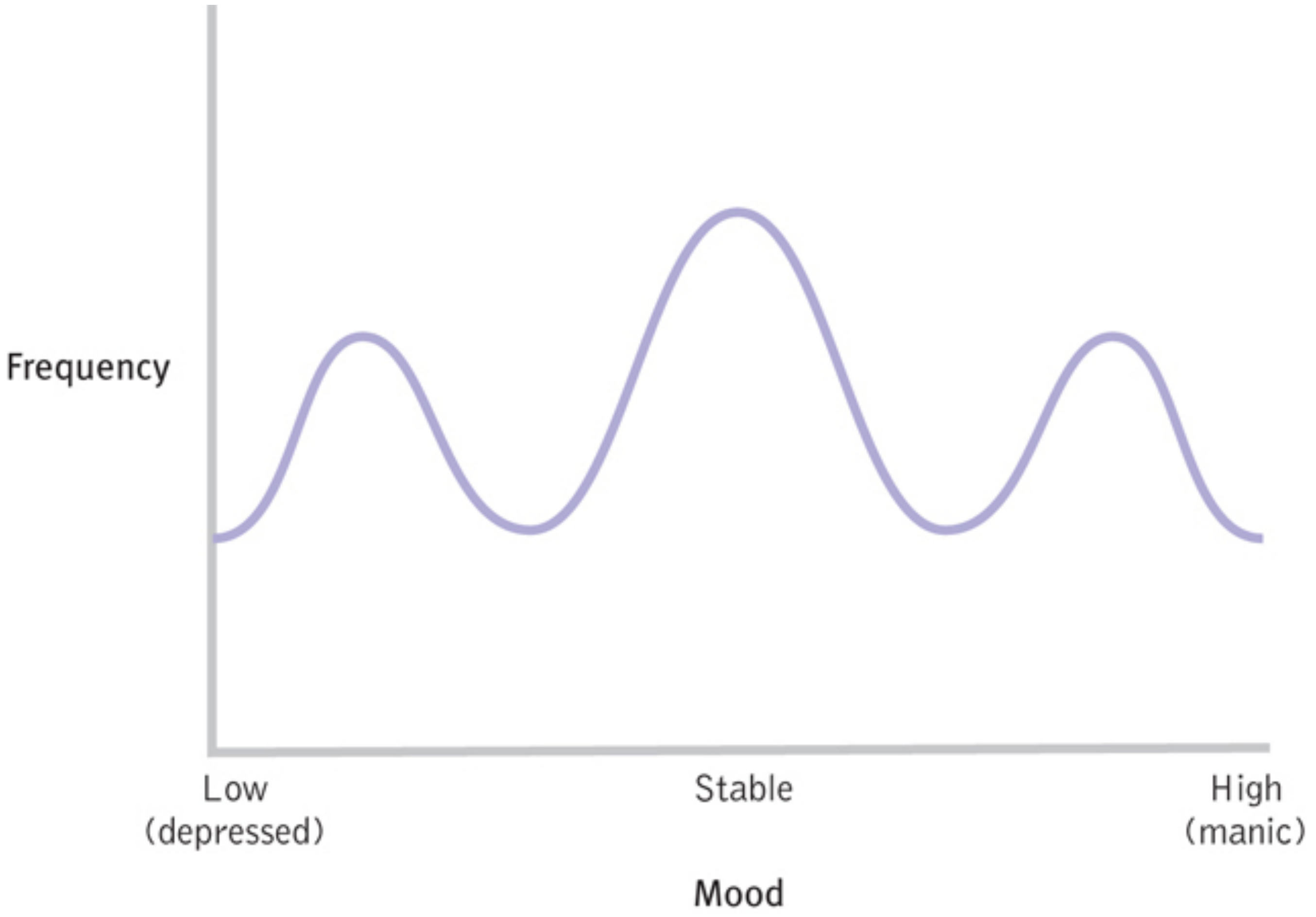
- We talked about this before but:
  - Unimodal = one hump distributions with one mode.
  - Bimodal = distributions with two modes.
  - Multimodal = distributions with three+ mode.
- Remember we talked about traditionally how if there are 10 5s and 10 6s (that is technically two modes) that people consider that unimodal because they are so close together.

How-To R

- Not as easy ☹
- temp <- table(as.vector(*column*))
- names(temp)[temp == max(temp)]

Why central tendency is not always the best answer:

Figure 4-4: Bipolar Disorder and the Modal Mood



Outliers and the Mean

- An early lesson in lying with statistics
  - Which central tendency is “best”: mean, median, or mode?
  - Depends!

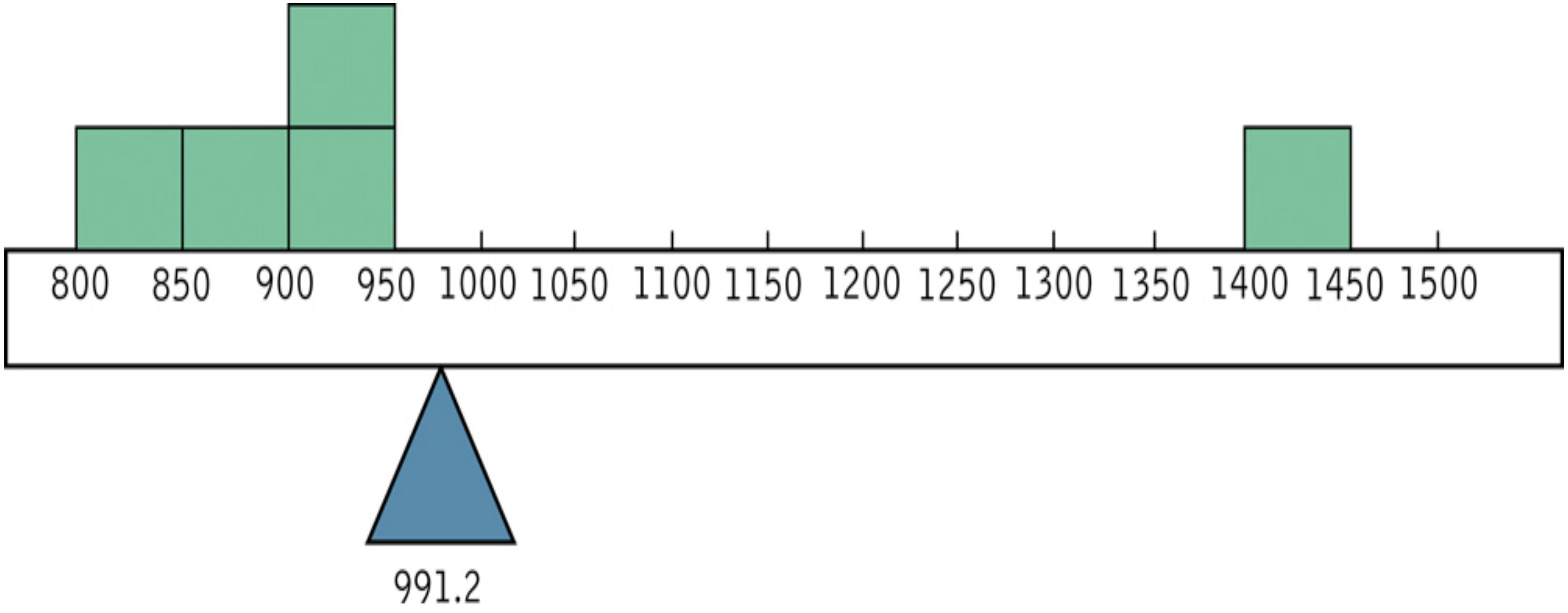
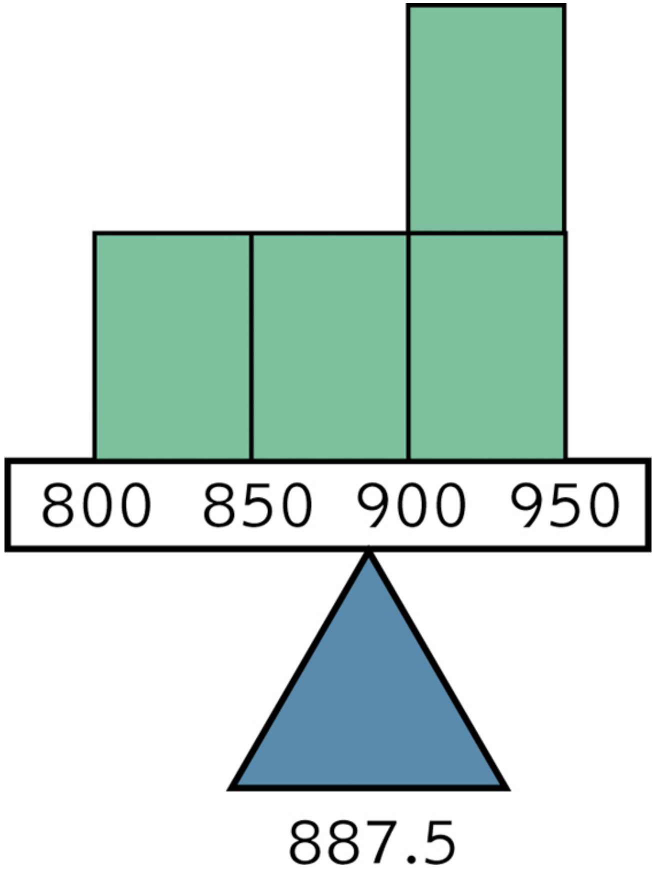


Figure 4-6:

The Mean without the Outlier



Let’s try it!

- Add an outlier to our data.
- outlier = c(mycolumn, #)
- Rerun the mean, median, mode.
- What happened?

Test with Outliers

- So what happens if we delete our outliers?
- Summary:
  - Mean is most affected by outliers (moved up or down, can be by a lot).
    - Best for symmetric distributions.
  - Median may change slightly one number up or down.
    - Best for skewed distributions or with outliers.
  - Generally the mode will not change. Uses:
    - One particular score dominates a distribution.
    - Distribution is bi or multi modal
    - Data are nominal.

Measures of Variability

- Variability: a measure of how much spread there is in a distribution
- Range
  - From the lowest to the highest score

Calculating the Range

- Determine the highest score
- Determine the lowest score
- Subtract the lowest score from the highest score

$$Range = x_{Highest} - x_{Lowest} = 10 - 1 = 9$$

How to Range

- Use the summary() function to get the min and max and subtract.
- Or do this:
  - max(*column*, na.rm = T) – min(*column*, na.rm = T)

Measures of Variability

- Variance
  - Average squared deviation from the mean
  - How much, on average, do people vary from the middle?

Calculating the Variance

- Subtract the mean from each score
- Square every deviation from the mean
- Sum the squared deviations
- Divide the sum of squares by N

$$SD^2 = \frac{\sum (X - M)^2}{N}$$

Super special notes right here about N versus N-1.

Measures of Variability

- Standard deviation
  - (square root of variance)
  - Typical amount that each score deviates from the mean.
  - Most commonly used statistic with the mean.
  - Why use this when variance says the same thing?
    - Standardized – brings the numbers back to the original scaling (since they were squared before).
    - Still biased by scale.

Calculating the Standard Deviation

- Typical amount the scores vary or deviate from the sample mean
  - This is the square root of variance

$$SD = \sqrt{\frac{\sum (X - M)^2}{N}}$$

Quick Notes about Formulas

- Samples usually use N – 1 as the denominator
  - UNBIASED
  - var(*column name*, na.rm = T)
  - sd(*column name,* na.rm = T)

Quick Notes about Formulas

- Populations usually use N as the denominator
  - BIASED
- Run this code as is:
  - pop.var <- function(x) var(x) * (length(x)-1) / length(x)
  - pop.sd <- function(x) sqrt(pop.var(x))

Quick Notes about Formulas

- Populations usually use N as the denominator
  - BIASED
- Run this code as is:
  - pop.var(*column name*)
  - pop.sd(*column name*)
- We created our own functions!
  - So, you will have to run it every time you open R and want to use it.

Maybe start a “I need this” file with libraries, themes, and special functions?

Interquartile Range

- Measure of the distance between the 1st and 3rd quartiles.
- 1st quartile: 25th percentile of a data set
- The median marks the 50th percentile of a data set.
- 3rd quartile: marks the 75th percentile of a data set

Calculating the Interquartile Range

- Subtract: 75th percentile – 25th percentile.
  - You can look at these numbers in the summary() function.
- IQR(*column name*, na.rm = T)