Chapter 5: Descriptive Statistics

Virginia Price

7/1/2019

5.1 Measures of Central Tendency

"Central tendency" -> measure of the "middle" or "average" of a data set

Here's our data set for this chapter; it includes the teams that played in the playoffs, as well as the margin by which they won.

```
load("../Navarro-data/aflsmall.Rdata")
```

Mean

The average! Or the "center of gravity" of the data set notation: * total number of observations: N * each observation = X; individual observations, $X_1, X_2 ... X_N$ * Mean is $\bar{X}$, given by

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

So for 5 observations,

```
(56+31+56+8+32)/5        #ineffieceint and non-generalizable
```

```
## [1] 36.6
```

```
sum(afl.margins[1:5])/5 # way better
```

```
## [1] 36.6
```

```
mean(afl.margins[1:5])  # the best
```

```
## [1] 36.6
```

Buuuut R has an inbuilt function for this so let's just find the mean of errything

```
mean(afl.margins)
```

```
## [1] 35.30114
```

Median

The observation in the middle ("the middle name")

For our 5-observation subset,

8,31,**32**,56,56 -> 32 is the median

Adding a 6th game, 8,14,**31**,**32**,56,56 -> 31.5 is the median (average of the two middle numbers)

R has an inbuilt function for this too:

```
median( afl.margins )
```

```
## [1] 30.5
```

When do we use these different central measures?

- *interval data* – choose between median and mean! Note that means can be pretty sensitive to outliers. Median is good if you want "typical", mean is good if you want overall
- *ordinal data* – ranked data, like a Likert scale – use the median!
- *nominal data* – that is, data that is not weighted by its number (e.g., words!) - Don't use central measures! Focus on grouping analyses

Example: an Australian housing market bank used the mean instead of the median and ended up with a much lower housing:income ratio because it averaged rich peoples' incomes compared it to the median house price. This skewed the dataset from 9:1 -> 5:1!

Trimmed mean

Dataset: -100,2,3,4,5,6,7,8,9,1

-100 is probably an outlier! But what if the 'outlier' is -15 instead? Should we still include it? At what point does that data become an outlier?

Use the median, or the "trimmed mean" by discarding the most extreme examples *on both ends*. Generally uses more info than the median to get a better idea of the dataset.

```r
dataset <- c( -15,2,3,4,5,6,7,8,9,12)
mean(dataset)
```

```
## [1] 4.1
```

```r
median(dataset)
```

```
## [1] 5.5
```

```r
mean(dataset, trim=0.1) #trims 10% of the dataset: compare to median!
```

```
## [1] 5.5
```

```r
#For the margins data:
mean(afl.margins,trim=0.05)
```

```
## [1] 33.75
```

Mode

The value that occurs most frequently! R can make frequency tables!

```r
table( afl.finalists )
```

```
## afl.finalists
##         Adelaide         Brisbane          Carlton      Collingwood
##               26               25               26               28
##         Essendon          Fitzroy        Fremantle          Geelong
##               32                0                6               39
##         Hawthorn        Melbourne  North Melbourne    Port Adelaide
##               27               28               28               17
##         Richmond         St Kilda           Sydney       West Coast
##                6               24               26               38
## Western Bulldogs
##               24
```

R doesn't have a built-in function to calculate the mode, but there's functions in the lsr library for that:

```r
modeOf(afl.finalists) # [1] "Geelong"
```

```
## [1] "Geelong"
```

```r
# Finds the thing that appears most frequently
# Geelong has played the most in the finals between 1987-2010

maxFreq( afl.finalists ) # [1] 39
```

```
## [1] 39
```

```r
# Tells you the number of times the most popular team appeared
```

5.2 Measures of Variability

variability: How "spread out" the data are.

Range

Range = Biggest value - smallest value. It's a *terrible* measure of variability (aka, not robust.)

For example, our set [-100,2,3,4,5,6,7,8,9,10] has a range of 110, but would only have a range of 8 if the outlier was removed.

```r
max(afl.margins)
```

```
## [1] 116
```

```r
min(afl.margins)
```

```
## [1] 0
```

Interquartile Range (IQR)

Better than regular range: finds the range between the 25th and 75th percentile (aka as a "quantile"). Think about it as the "middle half" of the data.

The 10th quantile/percentile is the smallest number x such that 10% of the data is less than x. The 50th percentile is also known as the median!!

```r
quantile( x = afl.margins, probs=.5)
```

```
##  50%
## 30.5
```

We can put in lots of quantiles at once by specifiying a vector for the `probs` argument, then subtract them…

OR we can just use `IQR`.

```r
quantile( x=afl.margins, probs = c(.25,.75) )
```

```
##   25%   75%
## 12.75 50.50
```

```r
50.50-12.75
```

```
## [1] 37.75
```

```r
IQR( x = afl.margins )
```

```
## [1] 37.75
```

Mean Absolute Deviation

Rather than looking over the entire spread of the data, we can look at "typical" deviations from a specific reference point, like the mean or the median. This is called the **mean absolute deviation**. We're going to call this $AAD(X)$.

$$AAD(X) = \frac{1}{N}\sum_{i=1}^{N}|X_i - \bar{X}|$$

The steps to calculate $AAD(X)$ might look something like this:

```r
X <- afl.margins[1:5]
X.bar <- mean(X)
AD <- abs( X- X.bar)
AAD <- mean(AD)
```

```r
print(AAD)
```

```
## [1] 15.52
```

This is quite wordy, but yay, the `lsr` package has an `aad()` function:

```r
aad( X )
```

```
## [1] 15.52
```

AAD is nice, but could be better. It's better to use squared deviations, aka, the… ### Variance Very nice because Navarro Says So. It's basically the AAD, but we used "squared deviations" instead just regular mean deviations. So sometimes caled the "mean square deviation."

$$Var(X) = s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

R has an inbuilt function, yay:

```r
mean( (X - mean(X) )^2)
```

```
## [1] 324.64
```

```r
var( X )
```

```
## [1] 405.8
```

Wait, what? Why are these so different?

Let's do the full set of 175 games to make sure we aren't crazy:

```r
mean( (afl.margins - mean(afl.margins) )^2)
```

```
## [1] 675.9718
```

```r
var( afl.margins )
```

```
## [1] 679.8345
```

Okay, so they're different, but not by much. R uses a slightly different formula:

$$Var(X) = s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Why? Well, statistics reasons which will be explored in Chapter 10. Something about calculating a "sample statistic" vs a "population parameter". For now, just trust R.

Variances are also very hard to talk about in human language, since it has such gibberish units. So the standard deviation is the **square root** of that, which makes much more sense to humans.

Standard Deviation

Effectively solves our "variance is confusing AF problem" by taking the square root of the variance.

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Standard deviation ($\hat{\sigma}$) in R is calculated with `sd`:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

```
sd( afl.margins )
```

```
## [1] 26.07364
```

To interpret it, use arule of thumb: abou 68% of the data should fall within 1 σ of the mean, 95% within 2σ, and 99.7% within 3σ. This works well most of the time, but it's not exact: it's based on an assumption that the histogram is symmetric and and bell-shaped (in math-y terms, "normally distributed").

Median Absolute Deviation

Basically the idea behind AAD, but the reference point is the median instead of the mean.

Comparing the two:

```
# Mean absolute deviation from the mean:
mean( abs(afl.margins - mean(afl.margins)) )
```

```
## [1] 21.10124
```

```
# Median absolute deviation from the median:
median( abs(afl.margins - median(afl.margins)) )
```

```
## [1] 19.5
```

MAD describes a *typical deviation from a typical value* in the dataset. e.g. A typical game involved a winning margin of 30 points, but any individual value typically varied from this median by about 19-20 points.

R has a built-in function for caluclating mad, and it's called `mad()`, of course. But it doesn't exactly use the above formula. It uses a constant that is set to `constant = 1.4826` to basically look like a standard deviation. This does also rely on the assumption that the data are shaped like a bell curve, though, so be cautious.

```r
mad(x = afl.margins, constant=1)
```

```
## [1] 19.5
```

```r
mad( afl.margins)
```

```
## [1] 28.9107
```

Which spread measure should I use?

- *range*: Gives full spread of the data, often not used unless you have good reason to care about extremes.
- *Interquartile range*: Tells you where the "middle half" of the data sits. It's nice pretty robust and used often.
- *Mean Absolute Deviation (AAD)*: How far "on average" observations are from the mean. Interpretable, but has a few minor issues that make it less attrative than standard deviation.
- *Variance*: average squared deviaton from the mean. Mathematically "right", but hard to interpret so very rarely reported directly.
- *standard deviation*: Easy to interpret, and straight up most popular measure.
- *Median Absolute Deviation (MAD)*: Typical deviation from the median value. In raw form, is simple and interpretable. Corrected form is a robust way to estimate the standard deviation for some data sets. Not used often.

So, in conclusion, use IQR and standard deviation unless you have a reason to use the others.

Vocabulary

- **descriptive statistics**: Finding ways of summarizing the data in a compact and easily understood fashion.
- **central tendency**: A measure that describes the "average" or "middle" of the data. Usually mean, median, and mode.
- **mean**: The average value of a set of observations
- **median**: The middle value of a set of observations
- **outlier**: A value that doesn't really belong with the others
- **robust**: A measure that is actually representative of the dataset, such as the mean of a set with lots of outliers.
- **frequency table**: A table showing how often certain values appear in a data set
- **variability**: how 'spread out' the data are.
- **range**: biggest value minus smallest value in a dataset. The numbers the data "ranges over", geddit
- **interquartile range (IQR)***: calculates the difference between the 25th and 75th percentile of the data.
- **quantile**: percentile
- **mean absolute deviation**: The average of the distances of each datapoint away from a reference point, such as the mean or median.
- **variance**: mean absolute deviation, but squared in the middle instead of with absolute values. Confusing in human terms.
- **standard deviation**: or "root mean squared deviation" (RSMD). Square root of variance.