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.

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

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

```
median(dataset)
```

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

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

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

```
#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!

```
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 *cannot* calculate the mode, but there's functions in the lsr library for that:

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

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

```
# 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
```

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

5.2 Measures of Variability