Stat 202 - 2015 XD W1 - Tues

Review

cases - objects described by data
variable - a characteristic of a case
label - variable(s) used to distinguish cases

categorical variable - places case into one of several categories

quantitative variable - a variable that takes numerical values for which arithmetic operations make sense,

ordinal categorical - natural order
nominal categorical - in name only

Distribution of a variable tells us what values it takes and how often it takes them

Exploratory Data Analysis
I) Study each variable by itself
II) Study relationships among variables
For each: (a) start with a graph or graph
(b) add numerical summaries
graphs for categorical: bar plot, pie chart } both show distribution

Graphs for
Quantitative variable

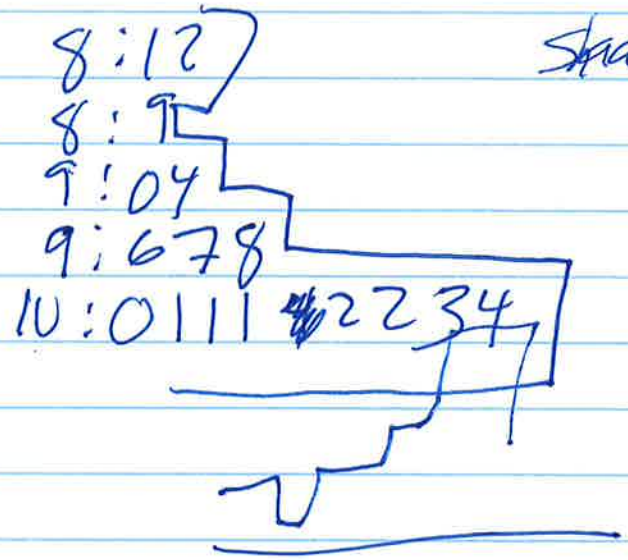
stemplot histogram

See homework 2; pass out extra copies
Project → if necessary

Do 1.7

The other way of plotting the distribution of a single ~~categorical~~ quantitative variable is with a histogram

You can make a histogram from a stemplot as follows



Slide in and rotate

pass out homework #3

bin with 5

Histograms are conventionally drawn vertically and not horizontally

Stemplots can be made with either by hand of software

Histograms should be made with software

Pg 4 ~~Pg 10~~

Stemplots preferred when you have a little bit of data

Histograms preferred (necessary when you have lots)

10K data point stemplot?

Stemplots have bin widths that are either 2, 5, 10, 20, 50, 100, etc

2 and 5 divide 10

Histograms can have any bin width

The counts of observations in bins are called frequences

80-85	2	this is called a frequency table
85-90	1	
90-95	2	
95-100	2	
100-105	8	

A histogram is essentially a bar ~~graph~~ plot of the freq table EXCEPT that customarily we do not put spaces between bar in a histogram (as we do for barplot)

Pg 7 ~~Pg 11~~

Homework 3 CO2

Load and plot ~~a~~ histogram

No time yet to work on homework #3

Pg 6 ~~Pg 12~~

The height of a bar can be (in StatCrunch)

Frequency (# of datapoints in bin)

Relative frequency $\left(\frac{\text{Frequency}}{\text{total \# of data points in data set}}\right)$

density $\left(\frac{\text{relative frequency}}{\text{bin width}}\right)$

For these choices the shape of the histogram doesn't change only the vertical scale

Choice of bin width

NO one right choice

too large (all data fall into one or a few bins) "Sky scraper"

too small (each datum falls into its own bin, each bin has only one data point) "pancake"

neither are desirable - use your best judgement

Note: changing bin width changes shape of histogram

Pg 7 ~~Pg 13~~

Another difference between bar graphs and histograms

A bar ppt need not have any measurement scale on the horizontal axis (if the variable is categorical) It simply identifies the items being compared.

A histogram's horizontal scale indicates possible values of a quantitative variable

No space between bars is suggestive of the fact that all values are covered.

The horizonal scale of a histogram usually has cuts.

~~[redacted]~~

~~Tomorrow we will look at call center, feel free to play around with it~~

Homework #3

We have data on the length of all 31,492 calls made to the customer service center of a small bank in a month.

Load and look at call center data set.

How do you analyze a dataset like this?

- Understand the background of the dataset (cases, variables, units of measure)

  (calls, length of call, seconds)

- Plot data
- Look for an overall pattern

  bin width 10
  where length ≤ 100 ≤ 1000 ≤ 5000 ≤ 2000

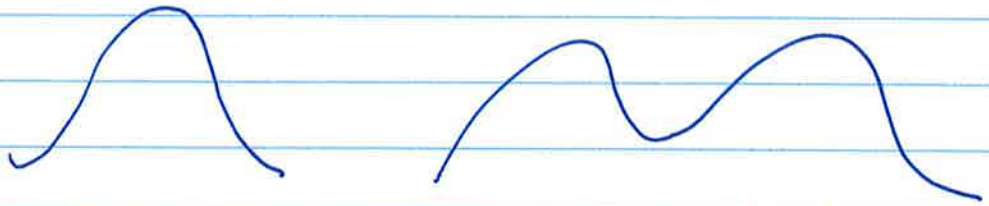
- Look for strikings deviations from the pattern

what is called a long tail (lots of individual values much greater than most others)

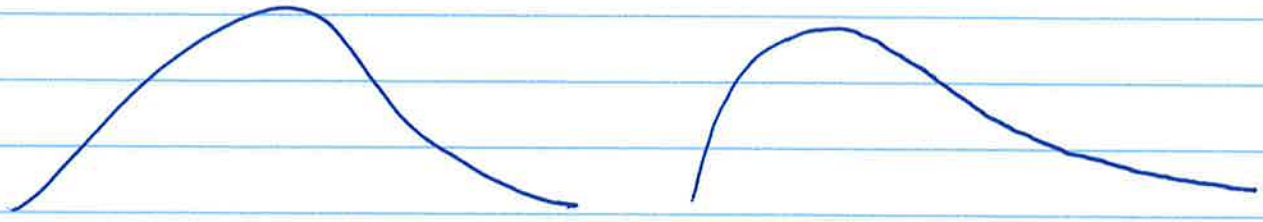
explain 1st bin

- Describe the overall pattern in terms of
  - shape
  - center
  - spread

Shape • the number of peaks or **modes**
• symmetric or skewed



1 peak two peaks



Symmetric (about mid point) Skewed to right

Call center two peaks (bimodal) skewed
Lack IQ (one mode symmetric

Skewed to left (analogous)

Tail – extreme values are said to be within the tails of the distribution

Center – "midpoint" of distribution

Spread – How "wide" or "narrow" the distribution is

In a minute we'll make these notions precise

Recall after we make graphs and desribe in terms of shape, center, spread we add numerical summaries,
Shape is more qualitative
for center and spread we ~~get~~ can derive precise numerical summaries

Start with center: there is two commonly used measures of center

Mean and Median

Mean – to find the mean of a set of observations, add their values and divide by number of observations

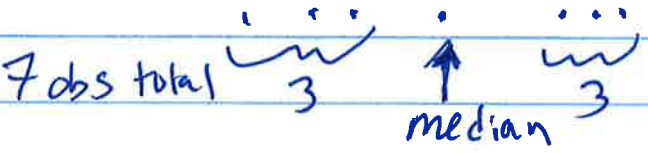
notation for mean → $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

Σ-notation → $\bar{x} = \frac{1}{n} \sum x_i$ limits of sum are 1 to n unless specified

Median - The median is the midpoint of the distribution, Half the obs are less than the median, the other half are larger

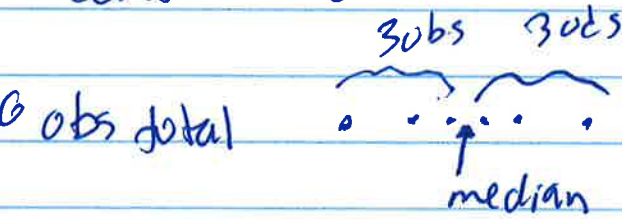
To find median

1. sort the obs from smallest to largest
2. IF n (number of obs) is odd median is center



not always equally spaced

IF n is even median is halfway between center two



Show mean and median on StatCrunch for IQ and call center.

Mean is sensitive to outliers
Median is resistant to outliers

Lets say there are 999 datapoints and all 999 are equal to 10

Whats the mean? 10
Whats the median? 10

Now suppose we add an outlier - a 1000th data point equal to 1 trillion 10^{12}

Whats the mean

$$\frac{10+10+10+\dots+10^{12}}{1000} = \frac{10^{12}}{1000} + \frac{9990}{1000}$$

$$= 10^9 + 9.99$$
$$\approx 1 \text{ billion}$$

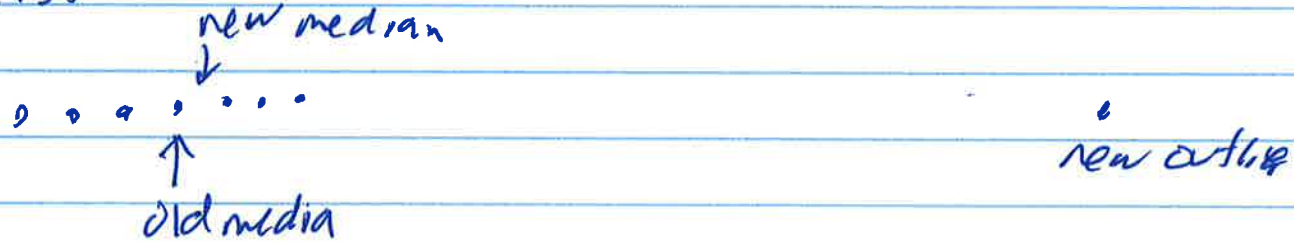
What's the median

Middle points both 10 so 10.

Doesn't change at all

One enormous outlier changes the mean tremendously but one enormous outlier usually does not change the median much In this case not at all.

A single outlier can change the median by at most one data point in the sorted list



Median is resistant to outliers (or just resistant)
Mean is sensitive to outliers.

Important distinction

If outliers are present it is often best to use resistant measures

Median and mean are measures for center of distribution. This doesn't tell whole story

- median family income doesn't tell your about extremes of wealth and poverty
- A drug may have correct mean concentration of an active ingredient but if some batches are way too high and others are way too low, that's a problem.

Need a measure of spread of a distribution

Spread (Resistant to outliers)
- Upper Quartile
- Lower Quartile
- Min
- Max

Spread (Sensitive to outliers) ← more commonly used
- Standard deviation

Upper Quartile (aka Third Quartile) (denoted Q_3)

is the median of the upper half of the observations (i.e. the median of the observations above the overall median)

Also called the 75th percentile.

Lower Quartile (First Quartile)

Median of lower half of obs
Also called the 25th percentile

The pth percentile – the value such that p percent of observations fall at or below it

Median is the 50th ~~Quartile~~ percentile or 2nd quartile Q_2 (but this notation is not usually used M is used.

Alert for percentiles: There might not be an obs such that exactly p percent of obs fall at or below it.

Example 11 obs find 27th percentile

Book says: take nearest obs

Software: different programs might give slightly different values.

For 25th 50th, 75th percentiles there is an exact defn based on the defn of median (middle obs or midpoint of middle two)

But even in this case software can give different answers

OK to report whatever software gives, If publishing, say what software used.

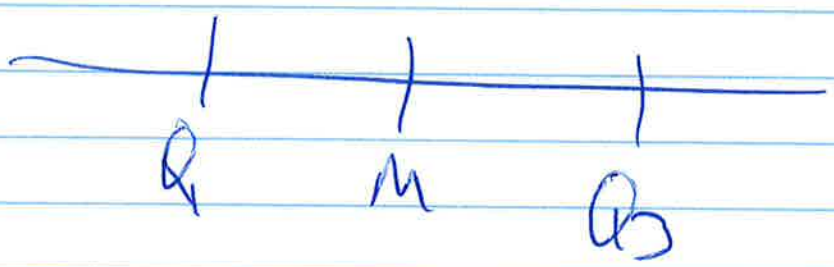
Quartile and percentiles are <u>resistant</u> to outliers

For describing center and spread of a distribution there is something called the

5-number summary

of a set of observations

Three of these numbers are



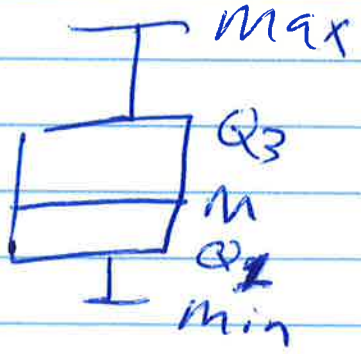
The other two are

Min
(Smallest obs)

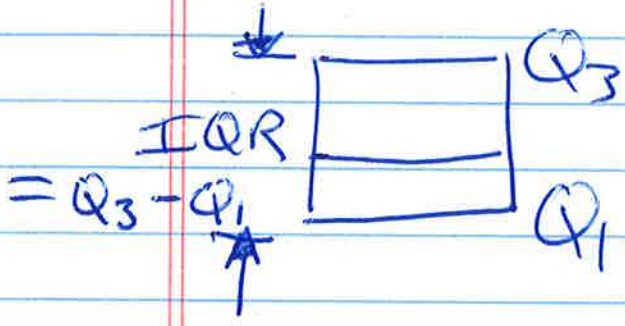
Max
(largest obs)

pg 18

You can display the 5-number summary with a box plot



The distance between Q_1 and Q_3 is called the inner quartile range (IQR)

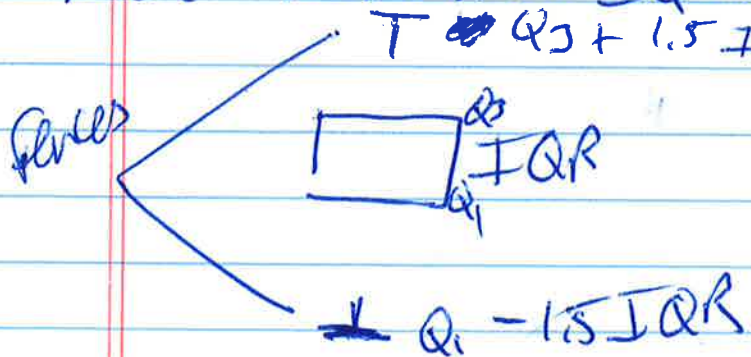


$$IQR = Q_3 - Q_1$$

The IQR is useful for flagging potential outliers

1.5 IQR Rule for Outliers

Call an obs a suspected outlier if it falls more than $1.5 \times IQR$ above Q_3 or below Q_1

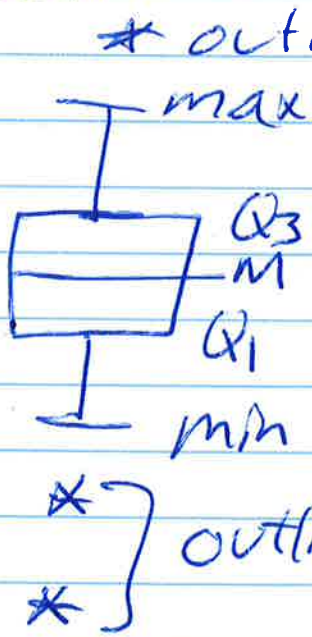


points that fall beyond fences are flagged as potential outliers

A modified boxplot also draws outliers

* Don't include suspected outliers in min/max

* Draw asterisks at suspected outliers



IQR
Show call center

Pass out and do homework #4

Talk about log transform