Stat 202 - 20155 - W2 - Wed

Review

Outlier - an individual value that falls outside of the overall pattern

tails of a distribution - region of extreme values (including outliers)

Symmetric distribution - values smaller and larger than midpoint are mirror images

Modes - peaks of distribution
bimodal - two mods
unimodal - one mode
multimodal - more that one mode (two or more)

Skewed to right right tail much longer
Skewed to left left tail much longer

Today

- Mean
- Median
- and more

Mean - To find the mean of a set of observations, add their values and divide by number of observation

written $\bar{x}$ →

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

or in Σ-notation

$$\bar{X} = \frac{1}{n}\sum X_i$$

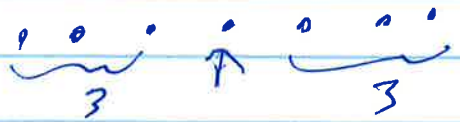
limits of sum are 1 to n if not specified

Median The median is the midpoint of the distribution, have the obs are less than the median, rest are larger

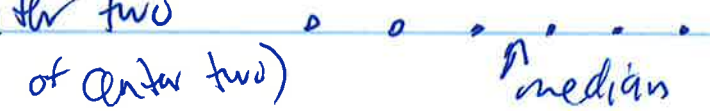
To find median,

1. Sort the obs from smallest to largest
2. If n (number of obs always) is odd median is center of list

these aren't necessarily equally spaced →



3. If n is even median is halfway between center two (mean of center two)



Mean versus Median

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

What's the mean 10
What's the median 10

Now suppose we add an outlier a 1000th datapoint equal to 1 milion

What's the mean?

$$\frac{10+10+\cdots 10+10^{12}}{1000} \simeq 10^9$$

about 1 billion

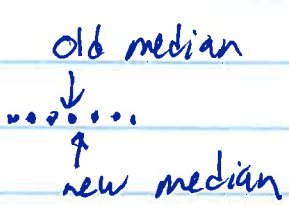
What's the median? 10

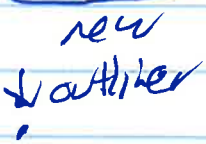
One ~~enormous~~ enormous outlier changes the mean tremendously but one enormous outlier usually doesn't change the median much. In this case not at all.

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

~~[illegible]~~

Stat 202-20155-W2-Wed





mean is changed dramatically

Median is resistant to outliers

Mean is sensitive to outliers

important distinction
If outliers are presents 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 you about extremes of wealth and poverty

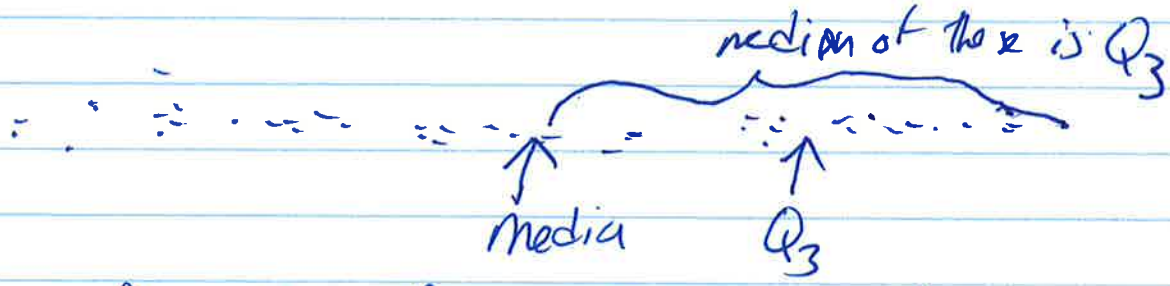
- A drug may have correct mean concentration of an active ingredient but if some batches are way too high and some are way to low, that's a problem.

Need a measure of spread of a distribution; not just center.

Upper Quartile (Third quartile) 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) Q_1

Median of lower half of obs.

Also called the 25th percentile

The p^{th} percentile - The the value such that p percent of observations fall at or below it

Median is the 50th percentile

Stat 202 - 2015S - W2 - Wed

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

Eg 11 obs 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 defn of median given (midpoint when even number of obs). But even then software can give slightly different values.

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

Question: Are quartiles and percentiles resistant to outliers or sensitive?

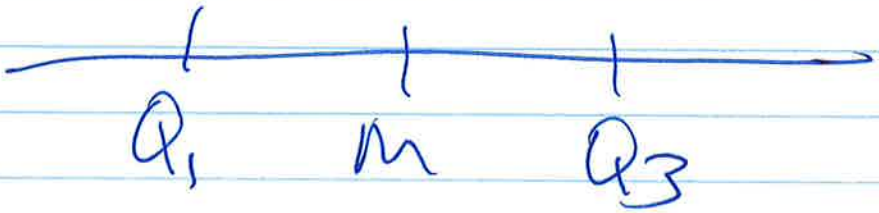
Stat 202-20155-W2-Wed (Pg 7)

There is something called the

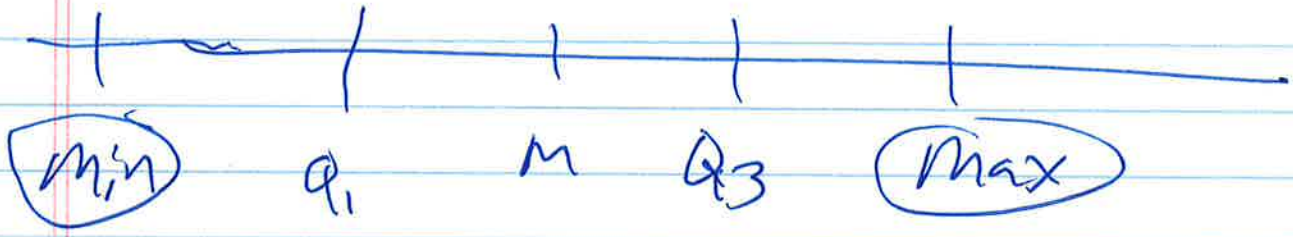
5-number summary

of a set of observation

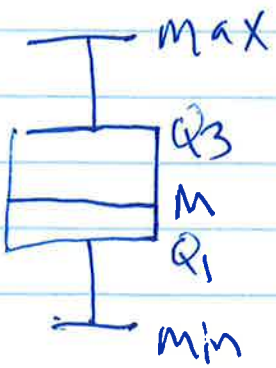
Three of these numbers are



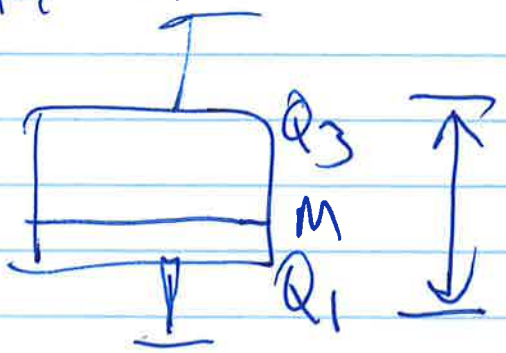
Any guesses for the other two numbers



You can display the 5 number summary with a box plot



or in Statcrunch



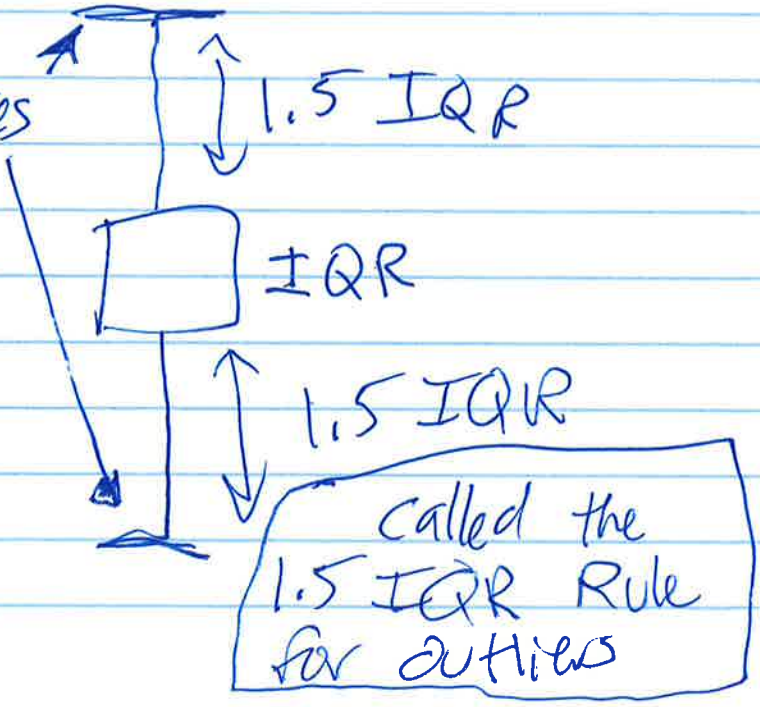
$Q_3 - Q_1 = IQR$

The difference between Q_3 and Q_1 (i.e. $Q_3 - Q_1$) is called the interquartile range (IQR)

The IQR is useful for flagging potential outliers.

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

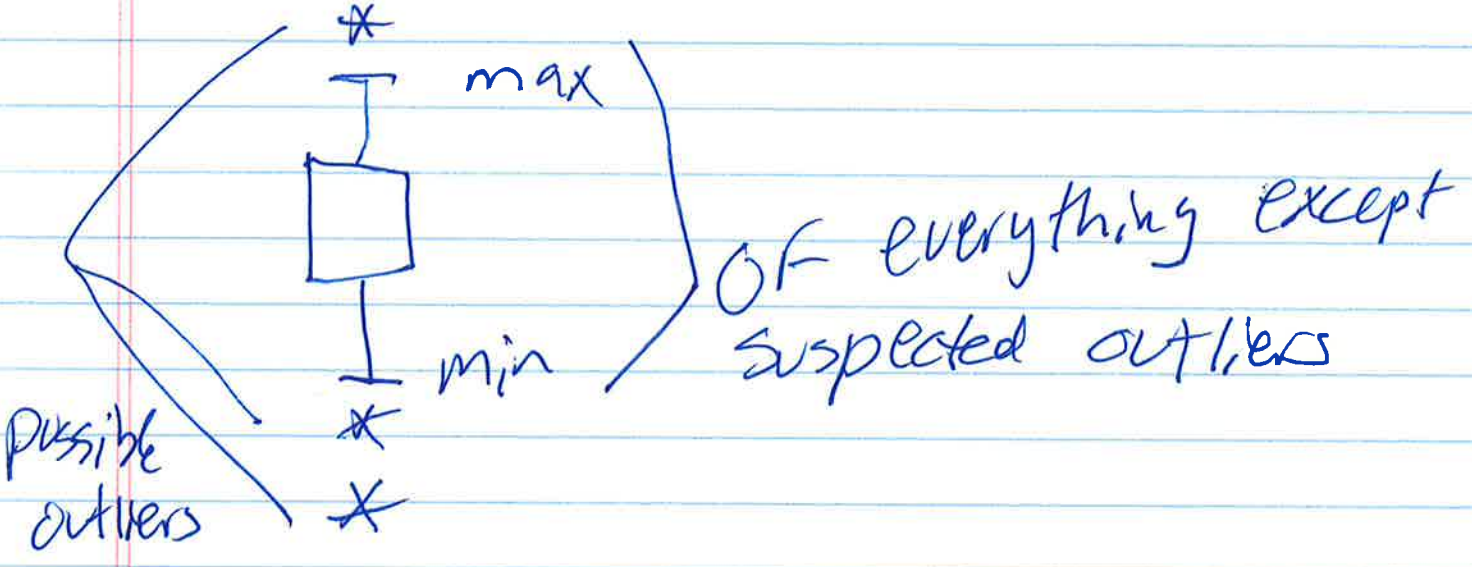
points that fall beyond fences are flagged as potential outlies



Modified Boxplot

* Don't include suspected outliers in min/max

* Draw asterisks at suspected outliers



Standard Deviation

The Five number summary is <u>not</u> the most common numerical description of a distribution

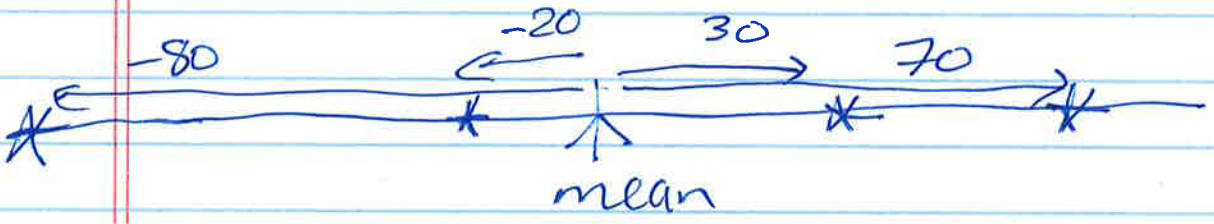
5-number summary $[\min, Q_1, M, Q_3, \max]$

The most common is {mean $\bar{x}$
and (standard deviation s.

How do you find the standard deviation?

Find the variance s^2 and take the square root.

How do you find the variance?



1) Find the deviations from the mean (these will always sum to zero)
2) Square the deviations
3) Average* the squared deviations

Deviations: $-80, -20, 30, 70$

Squared deviations:

$(-80)^2, (-20)^2, (30)^2, (70)^2$

$= 6400 \quad 400 \quad 900 \quad 4900$

Averaged* squared deviation

$$\frac{6400 + 400 + 900 + 4900}{3}$$

This average is different because you divide by 3 here instead of 4

Divide by 3 because 3 is the number of degrees of freedom (deviations add to zero)

Steps 1-3 compute s^2 (variance)

to compute s, one more step

step 4) $s = \sqrt{s^2}$

Properties of the standard deviation

* s measure spread about mean and should be used only when mean is chosen as the measure of center.

* $s = 0$ only when there is no spread (all obs have same value)

* otherwise $s > 0$, As spread becomes larger s becomes larger

* Like $\bar{x}$, s is not resistant (to outliers)

* s has same units as data; s^2 has square of units of data (data = height in inches s^2 = inches^2) s is usually preferred to s^2 for this reason

* The 5-number summary is usually better than (mean, std) for describing a skewed distribution or a distribution with strong outliers