# Stat 202-2015XD W1-Wed

## Review

For plotting distribution of quantitative variable we had
- stemplot
- histogram

Any questions?

In looking at these plots we describe distribution in terms of
- shape Unimodal or bimodal / Symmetric or Skewed
- center
- Spread

We actually have statistics to precisely quantify center and spread

Center: Mean and Median

Mean $\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

$= \frac{1}{n}\sum x_i$

Median - midpoint of data

Mean is sensitive to outlier
Median is resistant to outliers
(sometimes said to just be "resistant")

## Spread

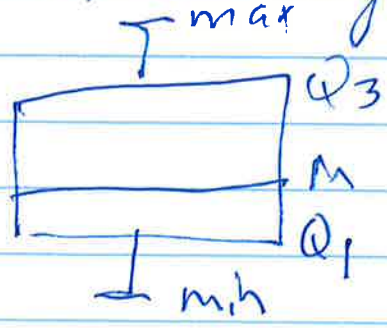
Upper Quartile (Third Quartile)
$Q_3$ - median of upper half of data)
Lower " ( First " )
$Q_1$ - median of lower half of data

## 5-number summary

Min, $Q_1$, M, $Q_3$, Max

## How to Find in Statcrunch

Box plot conveys the 5-number summary




$Q_3 - Q_1$ is IQR

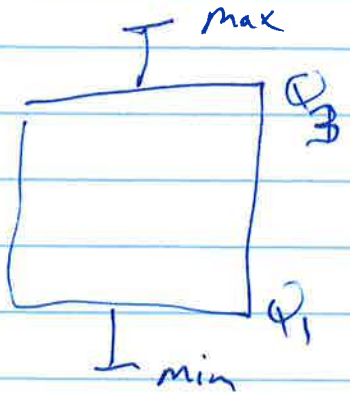
The min and max are obviously sensitive to outliers. Indeed they are equal to the most extreme outlier on either end.

So it would be nice to make the min and max more resistant to outliers.

That's done with the modified box plot, and the 1.5 IQR rule.

The 1.5 IQR Rule puts fences

— Fence $Q_3 + 1.5\,IQR$




— Fence $Q_1 - 1.5\,IQR$

* ] suspected outliers
*

Anything beyond the fences is flagged as a suspected outlier. Min and Max include only datapoints inside fences, more resistant to outliers. Suspected outliers are drawn as asterisks.

Do homework #4

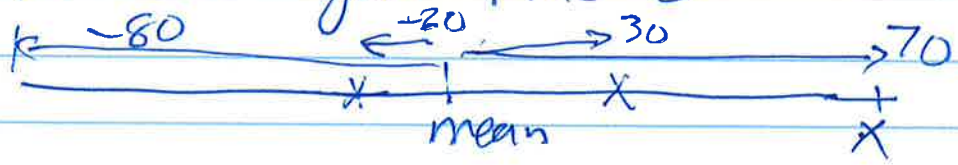
And the first two problems of homework #5

The 5-number summary is not the most common numerical description of a distribution
[min, $Q_1$, M, $Q_3$, max]
The most common numerical description is $[\bar{x}, s]$
mean and standard deviation

A related measure is $s^2$, the square of s, called the variance.

How do you find s?
First find $s^2$ then take the square root

How do you find $s^2$




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

Deviations from mean:
$-80, -20, 30, 70$
(Sum to zero)
(because of property of mean

Squared deviations
$(-80)^2, (-20)^2, (30)^2, (70)^2$
$= 6400 \quad 400 \quad 900 \quad 4900$

Averaged squared deviations

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

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

When averaging square deviations from mean, you always divide by one less than the number of data points instead of number of data points

One less than the number of data points is the number of degrees of freedom which has to do with the fact that

The degrees of freedom are less than n because the numbers sum to zero,

Why divide by $n-1$ instead of $n$.
There is a better match to theoretical distribution and theoretical Variance when you divide by $n-1$

apparently
Khan academy has a video on this which may provide a better explanation,

To compute $s$ not $s^2$
Do steps 1-3 then

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

Properties of the standard deviation

* $s$ measures spread about mean and should be used only when mean is chosen as the measure of center (use 5-number summary if median chosen)

* $s = 0$ only when there is no spread in data (i.e. all observations have same value)

* Otherwise $s > 0$, as spread becomes larger $s$ becomes larger

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

* $s$ has same units as data
  $s^2$ has square of units as data
  (e.g. if data are heights in cm
  $s$ is in cm
  $s^2$ is in $cm^2$
  It is nice to have measures with same units as data. For this reason $s$ is usually preferred to $s^2$

* The 5-number summary is usually better than $(\bar{x}, s)$ for describing a skewed distribution or a distribution with strong outliers

* Formula for $s$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

limits of $i = 1$ to $n$

$(x_i - \bar{x})$: deviations

Square deviation

average* of squared deviations $= s^2$

Square root of $s^2$

This is a situation in which it is helpful to use $\Sigma$-notation, otherwise

$$s = \sqrt{\frac{1}{n-1}\left((x_1-\bar{x})^2+(x_2-\bar{x})^2+\dots+(x_n-\bar{x})^2\right)}$$

This is harder to read.

A <u>transformation</u> is a function that transforms an old variable into a new variable

$$x_{new} = F(x_{old})$$

A transformation is just another name for a function

When we use the word transformation we think of the function as transforming the old variable into the new

(A) Example 1: If $x_{old}$ is distance in km and $x_{new}$ is distance in miles

$$x_{new} = .62\, x_{old}$$

In familiar function notation

$$y = .62x \quad \text{or} \quad F(x) = .62x$$

(B) Example 2:

$x_{old}$ is temperature measured in deg F
$x_{new}$ is temperature measured in deg C

$$x_{new} = \frac{5}{9}(x_{old} - 32)$$

or

$$y = 5/9\,(x - 32) \quad \text{or} \quad F(x) = 5/9\,(x - 32)$$

Both of these examples are examples of linear transformations

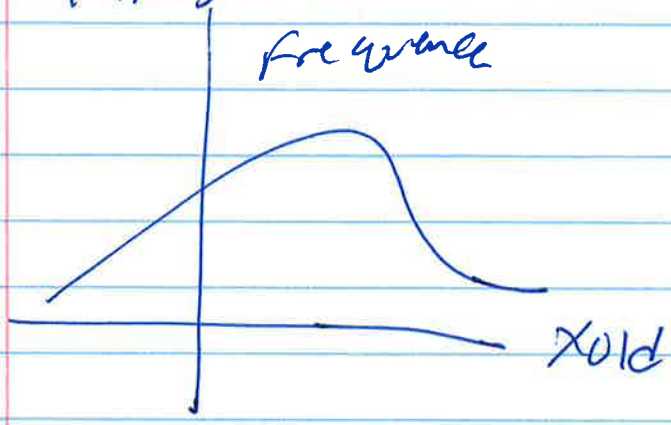
$y = mx + b$ in familiar notation

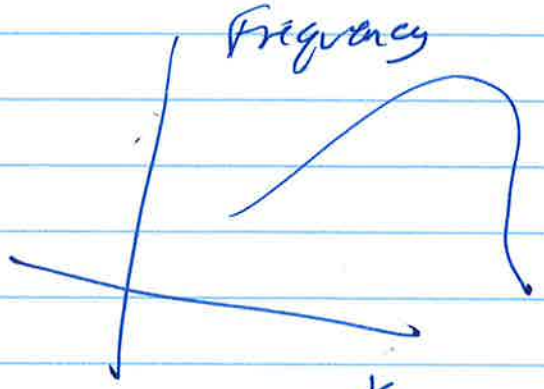
$x_{new} = a + b\, x_{old}$ in books notation

Linear transformations do not change the shape of a distribution

$X_{old}$ is right skewed iff $X_{new}$ is right skewed
$X_{old}$ is symmetric iff $X_{new}$ is symmetric
$X_{old}$ is bimodal iff $X_{new}$ is bimodal

Histograms




$X_{new} = a + b X_{old}$

a shifts histogram

b shrinks histogram if $|b| < 1$
expand " if $|b| > 1$
flips histogram about horizontal axis if $b < 0$

peaks, gaps and skewness remain

# Density Curves

A density curve is a smooth approximation to the irregular bars of a histogram

For a histogram there were three plotting alternatives

A: Frequency – count of obs in bin
B: Relative Frequency – Freq/total # of obs
C: Density – Relative Freq/bin width

Why pick density? Suppose you collect more and more data. You are likely to see less and less irregularity in the histogram. With A you will get more and more obs in each bin. Therefore the vertical scale will keep increasing. Not so with B and C.

Another thing to do is to decrease the width of the bin as you increase the number of obs. This will give a smoother and smoother histogram.

B will change vertical scale as you decrease bin width. C won't.

That's why we like C

As the amount of data grows and as the bin width decreases the histogram gets closer and closer to the density curves

Show with StatCrunch

For A the sum of bin heights is n
" B " " " 1
" C the area of bins is 1

For density curves (like for C)
1) The curve is always on or above the horizontal axis
2) the area under the curve and above the horizontal axis is 1

A density curve is any curve that satisfies (1) and (2).

A distribution is completely described by its density curve.

So instead of giving the 5-number summary of a variable (data) it would be better to give density curve (histogram plotted as density with an infinite amount of data)

Problem though: You can always plot the histogram but you can never have enough data to be sure you know the density curve.

There are ways to smooth the histogram and estimate density curve from data Stat Crunch can't do this though

What StatCrunch will do instead is overlay your histogram of data with a density curve of a standard distribution,

There are infinitely many distributions (any density curve satisfying conditions 1 and 2 will deform the one)

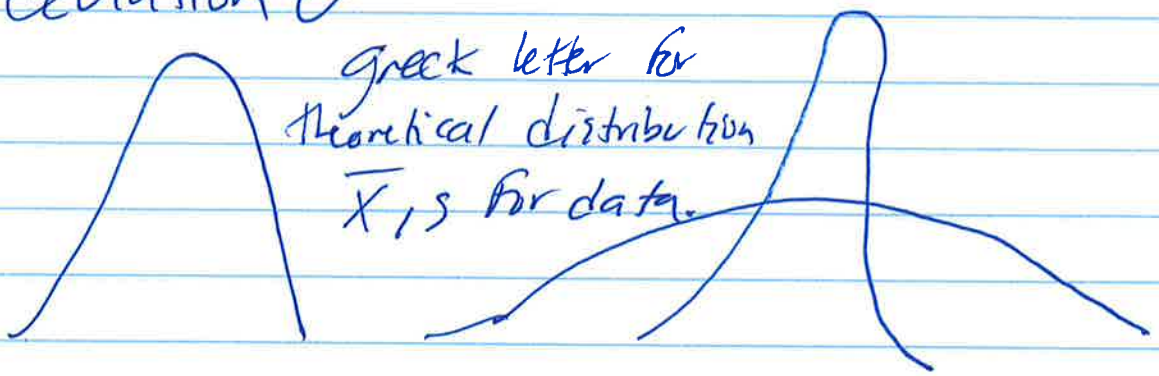
Only a few have names, These are called Standard distributions

The most important standard distribution is the Normal Distribution

Its density curve is the bell curve,

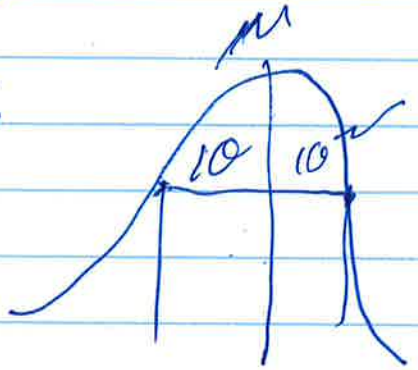
Bell curves are determined by there mean and standard deviation

Bell curves are determined by there mean $\mu$ and standard deviation $\sigma$




The area under a bell curve (or any density curve) is 1 therefore more narrow bell curves are taller.

The mean is the peak.




The inflection points measure 1 standard deviation from mean (Normal / Bell curves only)

HWHS