Stat 202-20155 - W2 - Fri (Pg 1)

Review

We had two ways of describing the center and spread of a distribution

(1) $\bar{x}, s$

mean, standard deviation

if either are present not as useful {
- sensitive to outliers (both measures)
- doesn't show skewness in distribution
  - eg. skewed to left
  - or skewed to right

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \qquad \frac{\text{sum of obs}}{\text{num of obs}}$$

$$= \frac{1}{n}\sum x_i \qquad \text{Sigma-notation}$$

$$s = \underbrace{\sqrt{\underbrace{\frac{1}{n-1}\underbrace{\sum \underbrace{\underbrace{(x_i - \bar{x})}_{\text{deviations}}{}^2}_{\text{squared deviations}}}_{\text{sum of squared deviations}}}_{\text{average of squared deviations dividing by } n-1 \text{ instead of } n}}}_{\text{square root}}$$

s^2 is variance

Stat 202-2015S W2-Fri (Pg 2)

② min, Q_1, M, Q_3, max

a.k.a 5 number summary

minimum, 1st Quartile, Median, 3rd Quartile, maximum
0th percentile, 25th percentile, 50th percentile, 75th per, 100th

resistant to outliers (all measures, except min and max)

→ but in a modified box-plot you exclude outliers (or suspected outliers) ~~from~~ min and max in that case the min and max ~~would~~ as displayed would be resistant (to outliers)

Median - point where half the obs are above half are below

Q_1 - Median of ~~upper~~ lower half of obs

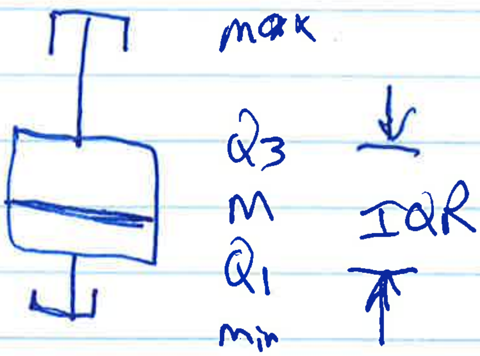
Q_3 - Median of upper half of obs

min/max the smallest and largest obs (in a modified box plot you exclude suspected outliers from the min and max)

Stat 202 - 2015S W_2-Friday

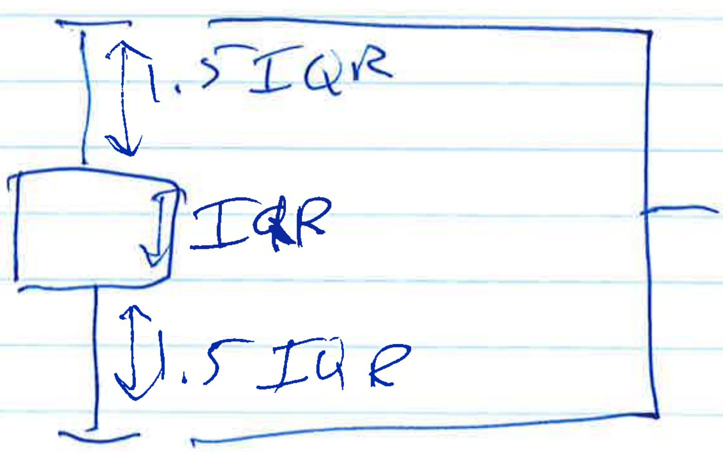
- The five number summary shows skewness of a distribution.

- The five number summary is displayed with a box plot



$Q_3 - Q_1 = IQR$ = Inner-quartile range.

The IQR is useful for flagging suspected outliers



fences beyond which we deem points suspected outliers.

In a modified box plot the plotted min and max are inside fences and obs outside are plotted as asterisks.

Stat 202 2015S W2 - Friday (Pg4)

Show homework 4

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

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

Transformation is just another name for a function.

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

<u>Examples</u>

(a) IF X_{old} is distance measured in kilometers and X_{new} is distance measured in miles

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

In familiar function notation

$$y = .62X = F(x)$$

Stat 202-2015S - W2 - Friday Pg5

Example 2

X_{old} = temperature measured in deg F°
X_{new} = temperature measured in deg C°

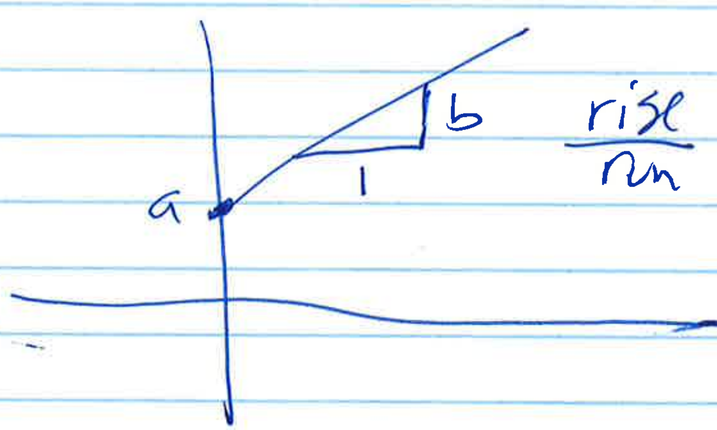
$$X_{new} = \frac{5}{9}(X_{old} - 32)$$
$$y = \frac{5}{9}x - 32$$

Unit conversions are usually linear functions (graphs are lines)

These are called $\boxed{\text{linear transformations}}$

$y = mx + b$ in familar notation

$X_{new} = a + bX_{old}$ book's notation



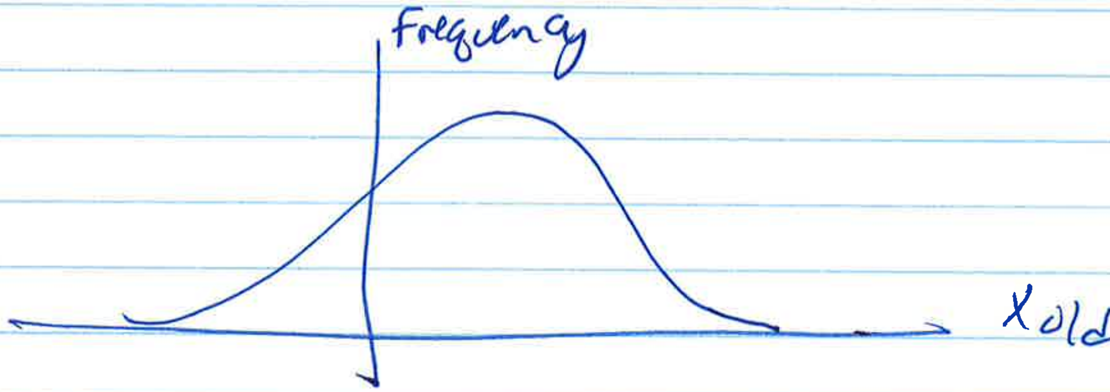
$$\frac{\text{rise}}{\text{run}} = \frac{b}{1} = b$$

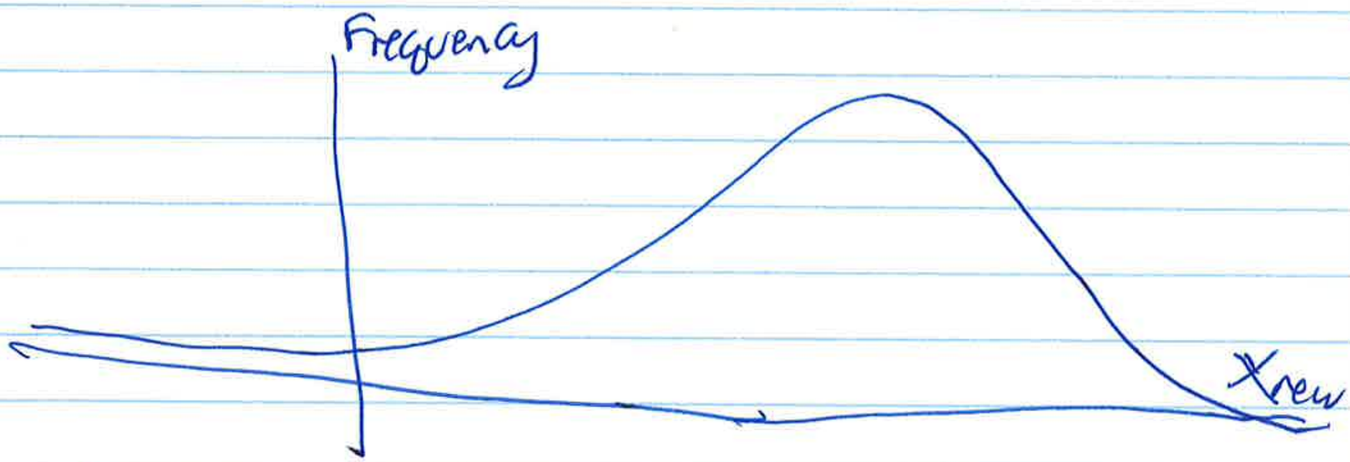
Stat 202 2015S – W2 – Fri

Linear transformations do not change the shape of a distribution

X_{old} right skewed $\Leftrightarrow$ X_{new} right skewed

X_{old} symmetric unimodal $\Leftrightarrow$ X_{new} symetric unimodal





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

a shifts histogram

b shrinks histogram if $|b| < 1$

expands " $|b| > 1$

flips if $b < 0$

peaks, gaps, and skewness remain.

Stat 202-2015S W2 - Friday

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 numer 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 noise in 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 bin as you increase the number of data points. This will give a smoother and smoother histogram.

B will change (vertical scale) as you decrease bin width
C wont.

Stat 202 2015S W2 Friday pg 8

Thats why we like C.

As we approach a histogram
with more and more data and
smaller and smaller bin width
we approach a histogram

Show with stat crunch

For A (Frequency) the sum of bin
heights is n

For B the sum of bin hieights is 1.

For C, the area of the bins is 1

For density curves, like for C.

(1) the area under curve is 1
(2) the curve is always on or above
the horizonal axis.

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 it would be better to give the density curve.

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

There are ways to estimate density curves from data. But StatCrunch does not do this

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 determines one)

Only a few have names (the standard distributions)

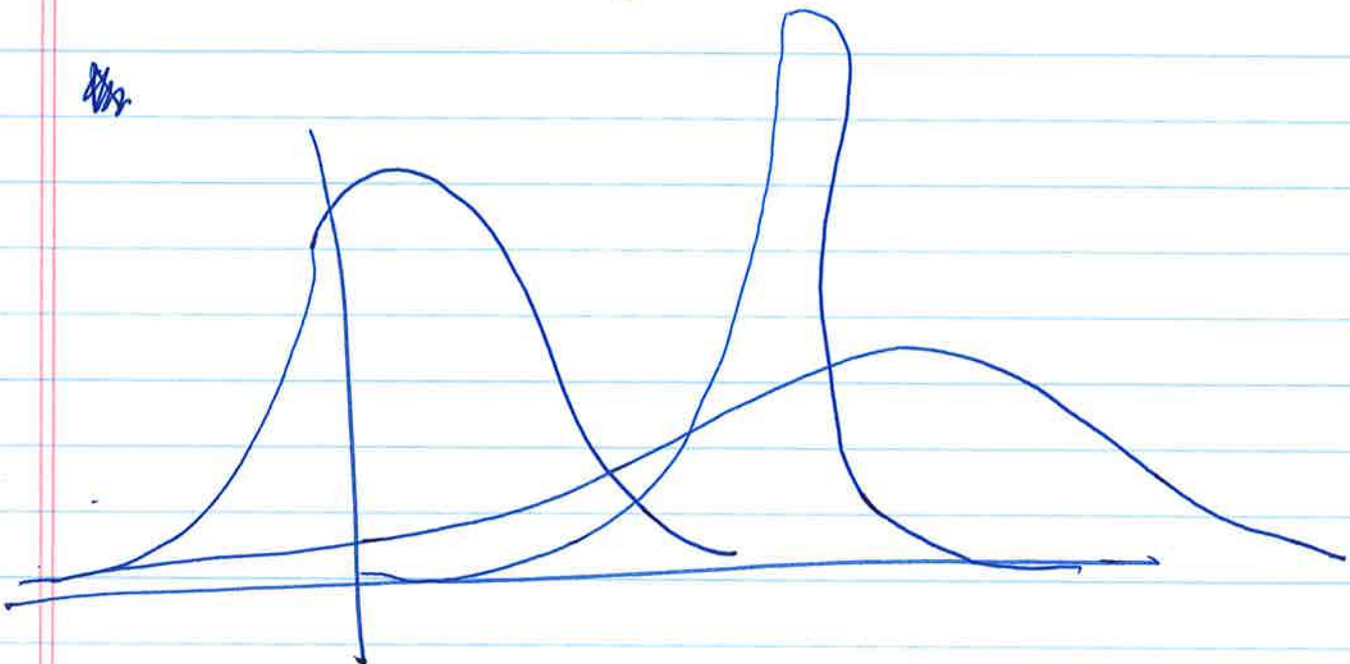
The most important standard distribution is the <u>Normal Distribution</u>

Its density curve is the <u>bell curve</u>.

Stat 202-2015S-W2-Fri

Pg10

Bell Curves are determined by their mean and standard deviation



Mean of data written $\bar{x}$
Standard deviation of data written s

Mean of density curve written μ
Std dev σ

We'll give precise defns next week