Stat 202 - 2015XD - W1 - Thurs (Pg 1)

## Review

Two common ways to report center and spread of a distribution (data)
(min, $Q_1$, M, $Q_3$, max) ← 5 number
($\bar{X}$, S)
↑ mean; ↑ standard deviation

$$S = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

A related measure is $S^2$, variance.

A transformation is a function for transforming variables

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

old transformed to new

A linear transformation is a transformation in the form

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

{ b is slope; a is y intercept }

Examples - unit conversions
Km → mi   °F → °C

A linear transformation shifts, shrinks expands or flips a histogram but doesn't change its shape

Peaks, gaps and skewness remain

## Density Curves

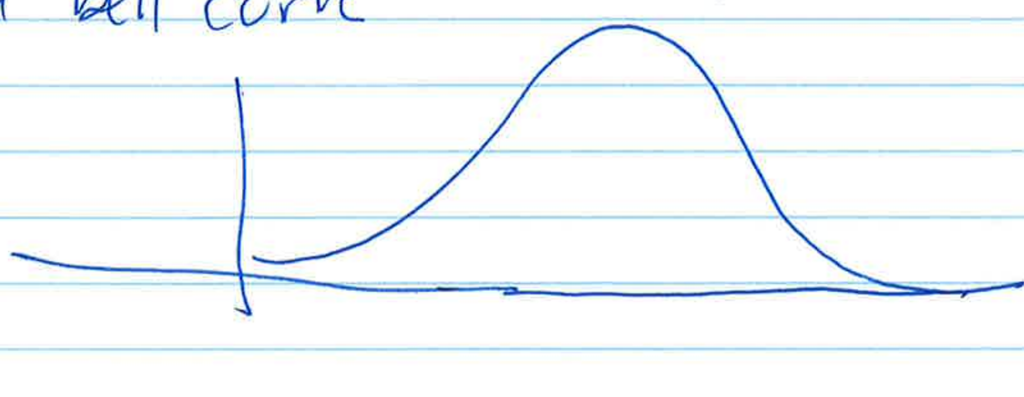
Distributions for quantitative variables are described by density curve.

density curve — theoretical construct
histogram — derived from data

both tell the same story

the density curve is a smoothed version of the ~~hi~~ histogram for more and more data and smaller and smaller bin widths.

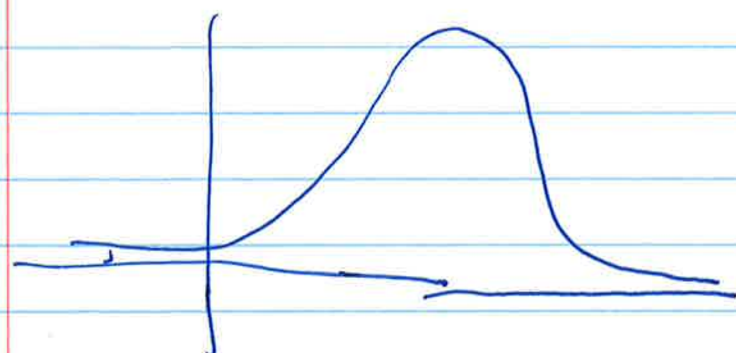
The most common density curve is a bell curve



~~##~~ Like for a histogram the horizontal axis of a density curve is the possible values of the quantitative variable.

The vertical axis on a density curve is density

$$\text{density} = \frac{\text{Proportion of obs in an interval}}{\text{length of interval}}$$



We have a density curve that could look something like this.

Two things have to be true about the curve

(1) The density curve is above the horizontal axis
(2) The area between the density curve and the horizontal axis is one.

Any curve that satisfies (1) and (2) is a density curve for some distribution

There are infinitely many density curves but a few of them actually have names — these are the standard distributions.

The most important standard distribution is the Normal Distribution
Its density curve is the bell curve

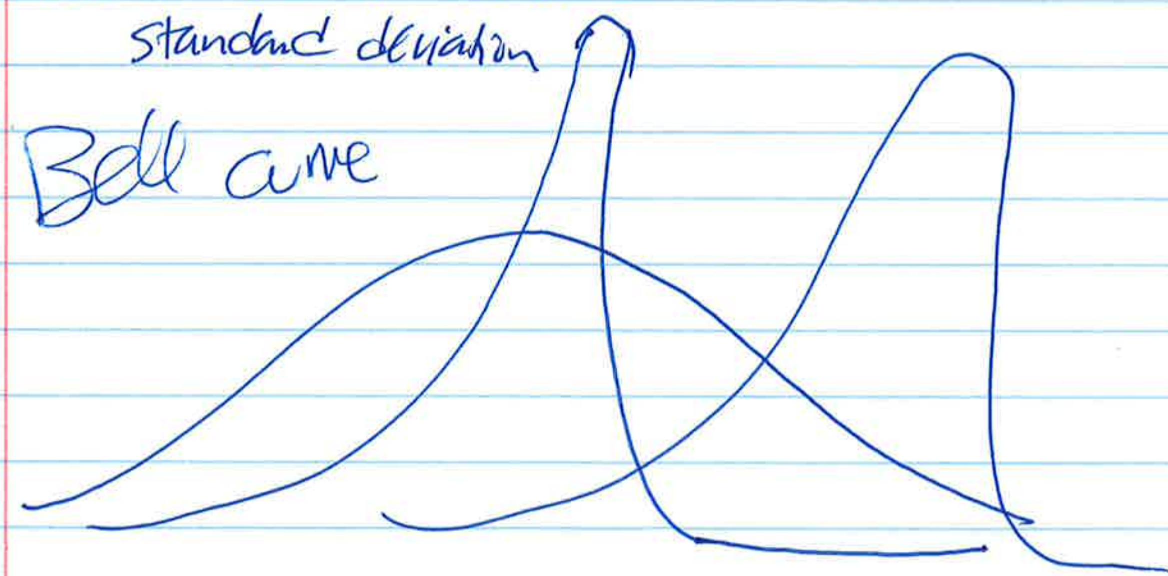
Bell curves are determined by their mean $\mu$ and standard deviation $\sigma$

greek letters [ $\mu$ and $\sigma$ are used for theoretical distributions

$\bar{x}$ and $s$ are used for data

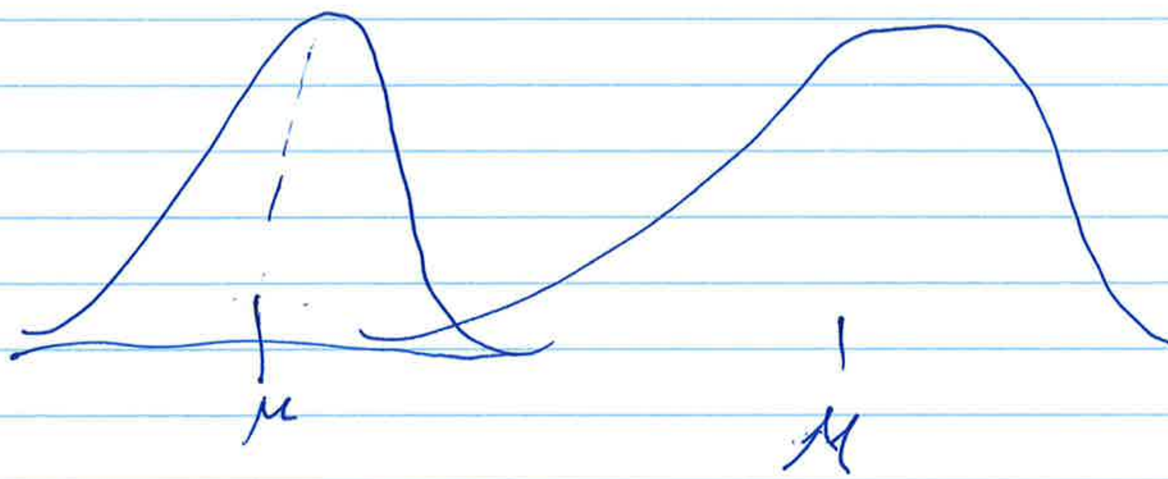
Both are notations for mean and standard deviation

Bell curve

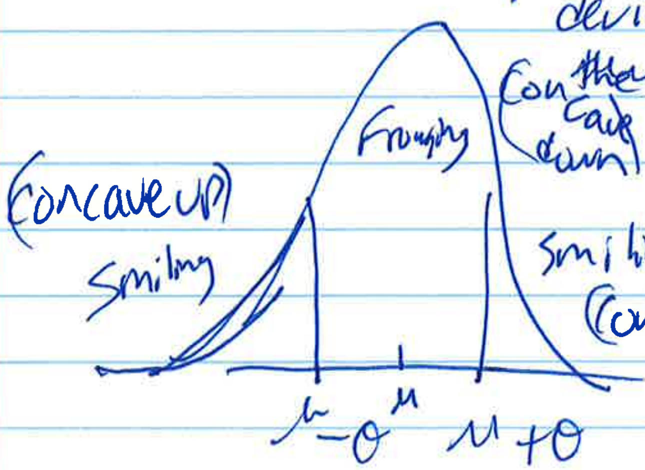


The more narrow the bell curve, the higher it is. That is because the area under the bell curve and above the horizontal axis is 1.

The peak occurs at the mean




For bell curves / normal distributions only the inflection points are 1 standard deviation above and below the mean



# Homework 5

While still at computers...
Now lets do an exercise in StatCrunch

Theoretical Distribution is Normal
Density is bell Curve

Using the computer we are going to generate data according to this distribution and look at the histogram

1000 datapoints → | Change bin width
100 →
100,000 data points

overlay bell curve

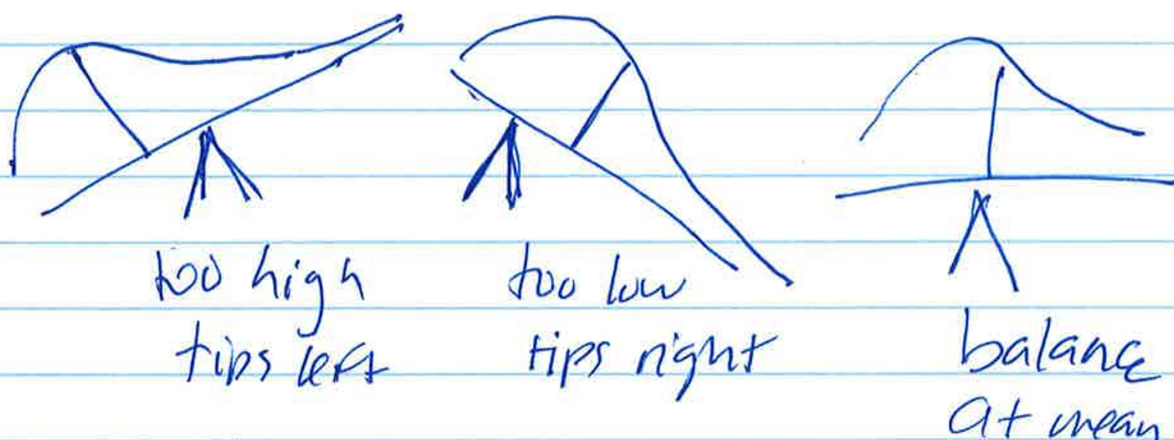
Measuring the center and spread for density curves

A mode of a distribution is a peak point of its density curve

The median is the point with half the total are on each side (total area is 1 so area 1/2 on each side)

The quartiles are found by dividing the area into quarters

The mean of a distribution is the balance point of the density curves



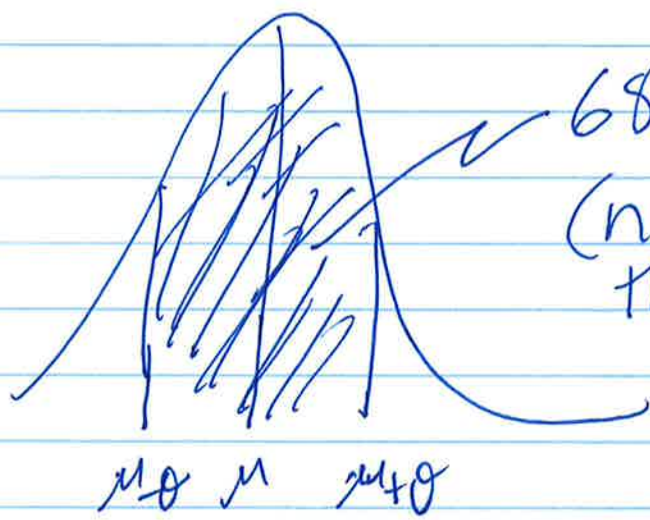

As already said for normal distributions/bell curves the inflection points are one standard deviation above and below the mean. This is for Normal only

Other distributions don't have an easy way of finding standard deviation

# The 68-95-99.7 Rule

For all normal distributions
(must be normal)

Approximately 68% of the area under its density curve lies within 1 standard deviation of mean.



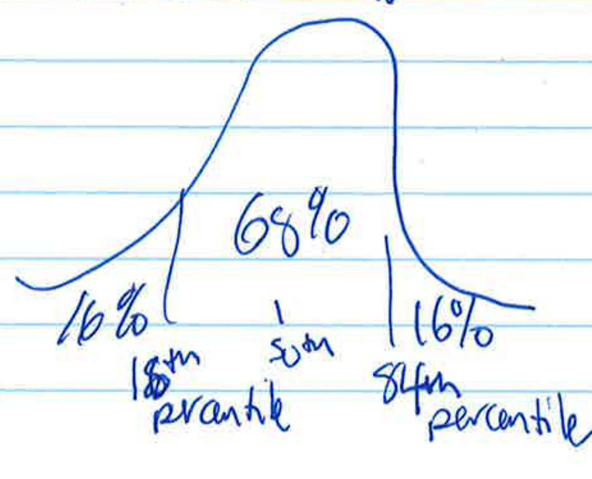

68% of area

(not exactly 68% though. This is rounded number when real number is irrational.

This means that approximately 68% of obs fall within 1 standard deviation from the mean and approximately 32% fall farther from the mean.

16% fall below 1 std
16% fall ~~below~~ above

16 + 68 + 16 = 100%

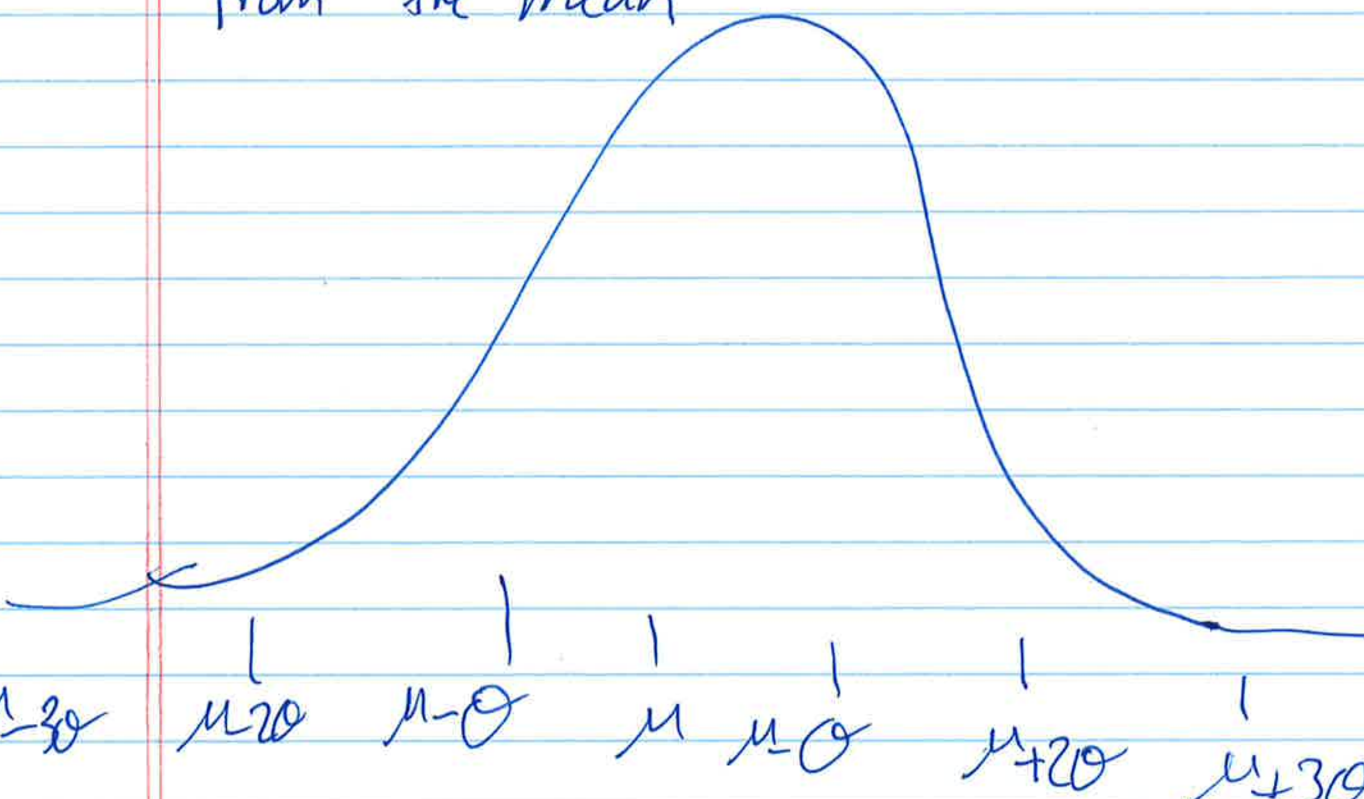


Normal only

Approximately 95% of obs fall within 2 standard deviations from the mean

That means 5% fall beyond two standard deviations 2.5% below and 2.5% above

Approximately 99.7% of obs fall within 3 standard deviation from the mean




0.3% fall beyond 3σ from mean
0.15% below
0.15% above

thats the 68-95-99.7 Rule

As suggested by the 68-95-99.7 Rule, all that matters for a Normal distribution is how close you are to the mean, relative to the standard deviation.

We can actually transform (with a linear transformation) a normal variable with any mean and std. dev. into one with the "standard" mean (zero) and the "standard" standard deviation (one)

Performing this transformation is called **Standardizing the observations**

The "new" variable is called the Z-score and is written Z

All Z-scores have mean 0 and std dev 1

The linear transformation that standardizes the data is

$$Z = \frac{X - \mu}{\sigma}$$

where $\mu$ is old mean
$\sigma$ is old std dev
X is old data

new mean is 0 new std dev is 1

The transformation

$X_{new} = X - \mu$

Shifts the density curve so that the mean is 0


The tranformation $Z = \frac{X-\mu}{\sigma}$

Scales the density curve

Shrinks if $\sigma > 1$ to $\sigma_{new} = 1$
or
expand if $\sigma < 1$ to $\sigma_{new} = 1$

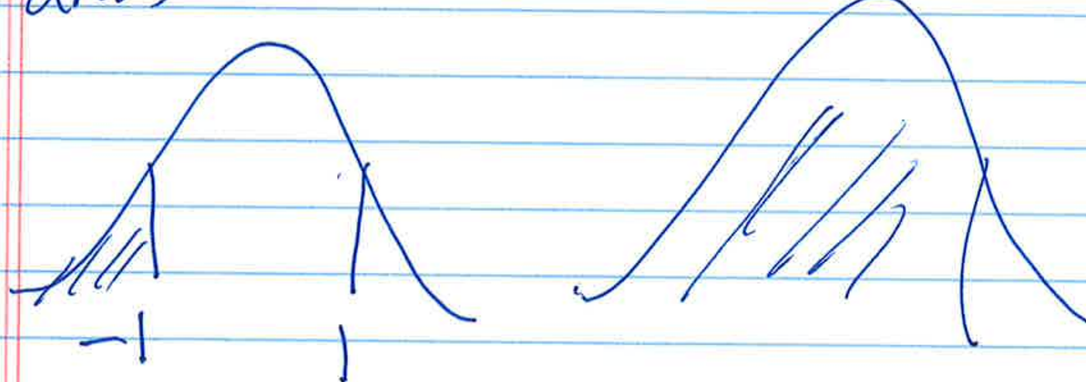
The Normal distribution can be represented symbollically. To know everything about a Normal distribution we only need to know its mean and standard deviation $\mu$ and $\sigma$

The normal distribution with mean $\mu$ and std. dev. $\sigma$ is denoted

$$N(\mu, \sigma)$$

Thus $N(1201, 320)$ is a normal distribution with mean 1201 and std. dev. 320

A table for the normal distribution will give you z scores and corresponding areas




z score of -1 has area $.16 = \frac{100\% - 68\%}{2}$

Show Stat Crunch HW 6 / HW 7