***COURSERA: STATS W/ R SPECIALIZATION***

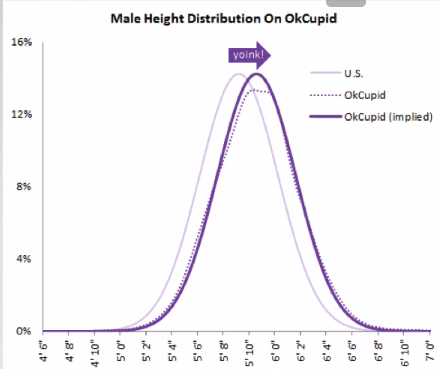
***COURSE 1 - Introduction to Probability and Data***

**WEEK 4- Probability Distributions**

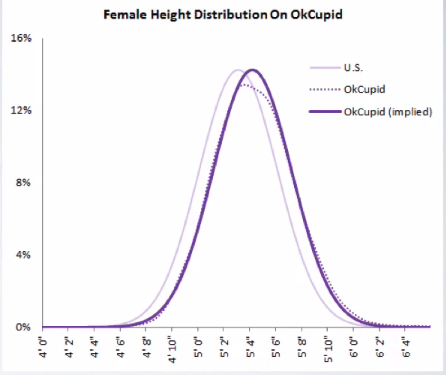
***4.1 The Normal Distribution***

**Normal Distribution**

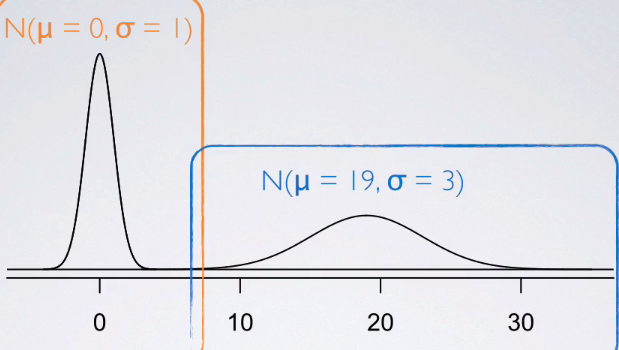
* Many variables in nature = nearly normally distributed, such as heights
* Ex: distribution of recorded heights of members of an online dating website OkCupid + since members of this website are US residents + likely represent a random sample from the US population, we expect their heights to follow the same height distribution of all Americans.
* However, a closer look shows that's not exactly the case.



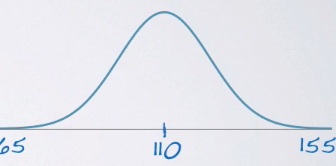
* Heights reported by men on OkCupid very nearly follow the expected normal distribution, except shifted to the right of where it should be.
* Appears males on OkCupid add on average a couple inches to their heights.
* Additionally, starting at about 5'8", the top of the dotted curve tilts even *further* rightward
* Indicates that closer to the 6 feet, males start round up a bit more than usual, which the OkCupid blog interprets as stretching for that coveted psychological benchmark of being 6 ft. tall
* Similar height exaggeration w/ females but w/out the lurch towards a benchmark height.



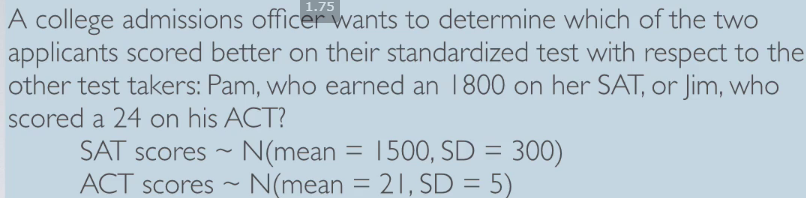
* **normal distribution/bell curve =** unimodal + symmetric
* Follows very strict guidelines about how variably data are distributed around the mean.
* While many variables are *nearly* normal, none are *exactly* normal due to these strict guidelines.
* normal distribution has 2 parameters, mean, μ, + SD, δ 🡪 N(μ, δ)



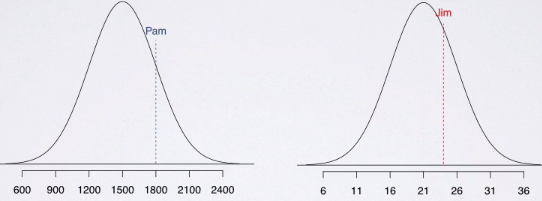
* See 2 normal distributions, 1 centered at 0 w/w SD = 1 + the other centered at 19 w/ SD = 3
* Good representation of how changing the center + spread of a distribution changes overall So what are these strict rules that govern
* For nearly normally distributed data:
* 68% falls within 1 SD of the mean. 95% falls within 2 SDs, + 99.7% falls within 3 SDs
* Possible for observations to fall 4, 5, or even more SDs away from the mean, but these occurrences are very rare if the data are nearly normal.
* 
* 
* We can also use this rule to estimate the SD of a normal model given just a few parameters about the distribution of the data.
* Ex: Doctor collects a large set of HR measurements that approximately follow a normal distribution. He only reports 3 statistics 🡺 mean = 110 BPM, the minimum = 65 BPM, + the maximum = 155 BPM. Which of the following is most likely to be the SD of the distribution?
* Told the distribution is normal 🡪 So the very 1ST thing we do = draw the normal curve + mark our mean = 110 in the center w/ min = 65 + max = 155.



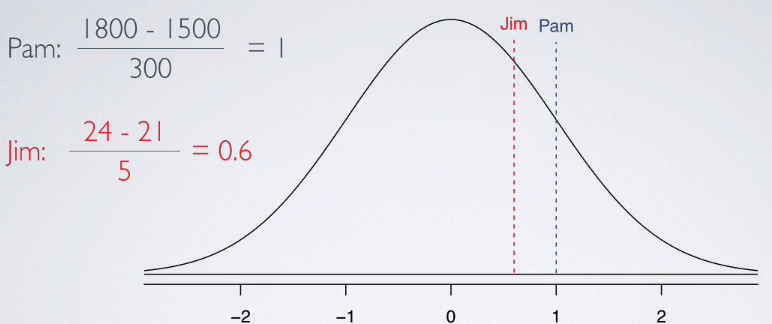
* Almost all of the data lie w/in 3 SDs of the mean.
* If SD = 5, the expected min + max = 110 +/- (3x5) 🡪 95 + 125
* If SD = 15, the expected min + max = 110 +/- (3x15) 🡪 65 + 155

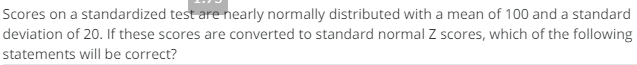


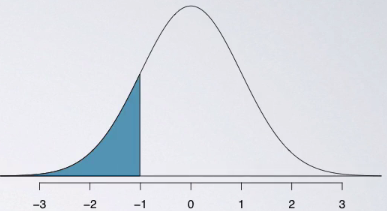
* Can draw the distribution of SAT scores + see Pam scored 300 points above the mean + Jim scored only 3 points above the mean



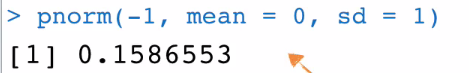
* However, can't just compare raw scores of 1800 vs 24 since we have different scales
* Instead, figure out how many SDs above the respective means of their distributions Pam + Jim scored
* SD of SAT = 300, so Pam scored 1 SD above the mean 🡺 (1800 – 1500)/300 = 1
* SD of ACT = 5, so (24 – 21)/5 🡺 3/5 🡺 Jim only scored 0.6 SD above the mean.
* Plotting these values on the same distribution: see Pam indeed do better than Jim.



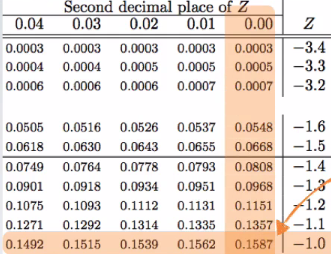
* These values = **standardized score** or **Z-score** = # of SDs an observation falls below/above the mean
* **Z-score** of an observation = that observation - the mean divided by the SD.
* By definition, the Z-score of the mean = 0
* Standardized scores are useful for IDing unusual observations (usually absolute Z-scores > 2, so either 2 SDs below or above the mean, or something beyond that)
* Z-scores are actually defined for distributions of any type, not just normal.
* Every distribution will have a mean + a SD, therefore for any observation, whatever distribution that random variable follows, we could calculate a Z-score.
* 
* 
* But when the distribution is normal, Z-scores can also be used to calculate **percentiles** = the % of observations that fall below a given data point.
* Graphically = the area below the probability distribution curve, **AUC**, to the left of that observation.

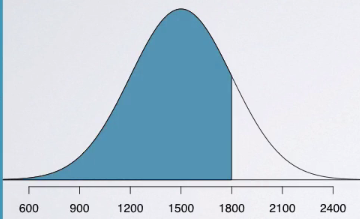


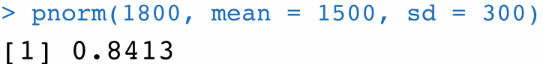
* Why can only use Z-scores under normal curves, but not in a distribution of a different shape?
* Can always calculate percentiles for *any* sort of distribution, *except* if the distribution does NOT follow a nice unimodal symmetric normal shape (need calculus for that)
* In this day + age, percentiles are easily calculated using computation (in R 🡪 **pnorm()** gives the percentile of an observation, given the mean + the SD of the distribution



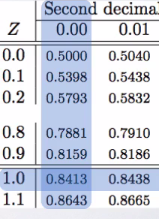
* Can avoid computation altogether + use a normal probability table
* Locate the Z-score on the edges of the table + grab the associated percentile value given in the center of the table.



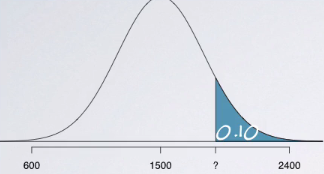
* While a computational approach is a little less archaic, the tables are actually very useful for getting a conceptual understanding of what we mean by AUC.
* Ex: We know that SAT scores are distributed normally w/ mean 1500 + SD 300 + also know Pam earned an 1800. Want to find out her percentile score.
* As we find out the distribution is normal: draw the curve, mark the mean, + shade the area of interest (scores below 1800)
* 



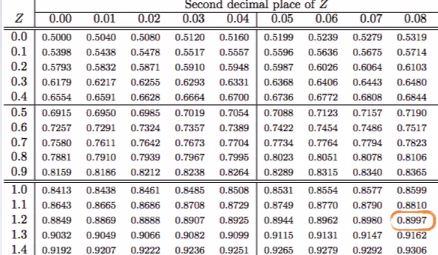
* Get a percentile of 0.8413, meaning Pam scored better than 84.13% of SAT takers.
* Could also use the table to arrive at the same conclusion.
* First, calculate the Z-score as the observation 🡪 1800 – 1500 / 300 = 1
* In the table, look for the Z-score of 1 + get the same probability, 0.8413 = probability of obtaining a Z-score < 1, which basically means the same thing that the shaded AUC,



* Note that both the table + pnorm always yield the AUC *below* the given observation.
* To find out the area *above* the observation, simply take the **complement** of this value since total AUC is always 1.
* So Pam scored worse than 1 - 0.8413 = 15.87% of the test takers.
* 
*  
* 
* We can also use the same properties of the **standard normal distribution** (the distribution of Z-scores) to find **cutoff values** corresponding to a desired percentile.
* Ex: A friend scored in the top 10% on the SAT, what is the lowest possible score she could’ve gotten?
* Remember, SAT scores are normally distributed w/ mean 1500 + SD 300.
* Looking for the cutoff value for the top 10% of the distribution
* This time we don't know the value of the observation of interest.
* But we DO know (or at least we can) get its **percentile score**.
* Since total AUC = 1, percentile score associated w/ the cutoff value for top 10% = 1 - .10 = .90



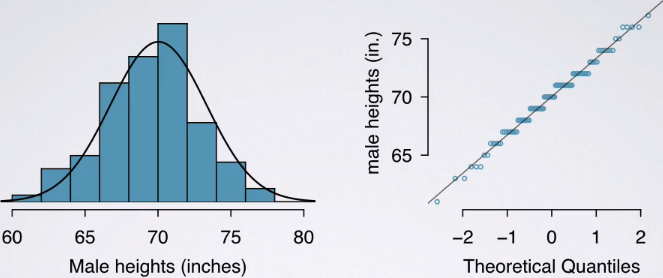
* We know the mean + the SD + using the table we can find the Z-score associated w/ the 90th percentile.



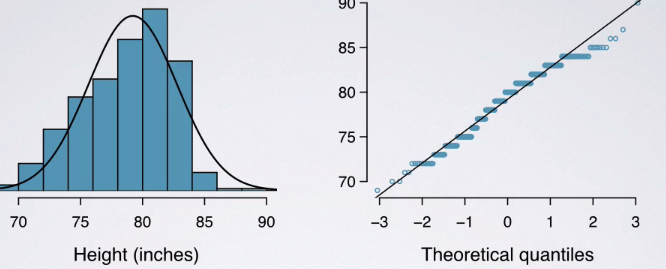
* Closest we can get is 0.8997 = the Z-score is 1.28.
* We know 1.28 = the unknown observation, X, minus the mean divided by the SD.
* 1.28 = (X – mean) / SD = 1.28(300) + 1500 = **1884 =** the cutoff value for top 10%, or bottom 90%, of the distribution of SAT scores
* If you have scored above 1884, you know you're in the top 10% of the distribution.

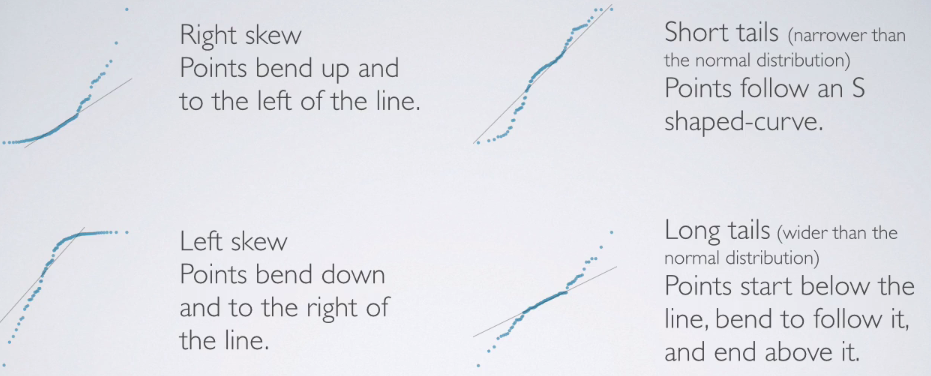
**Evaluating the Normal Distribution**

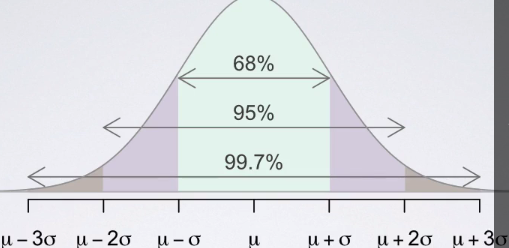
* How to evaluate whether a distribution is nearly normal or not.



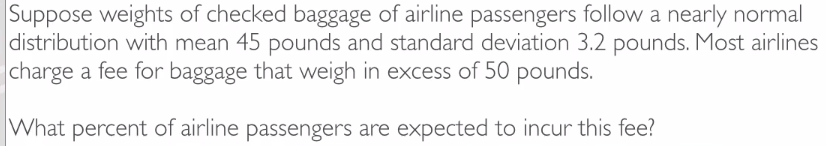
* Histogram + a **normal probability plot** of a sample of 100 male heights.
* On a **normal probability plot**, data are plotted on the y-axis + theoretical quantiles (quantiles that follow a normal distribution/what would be expected under the normal distribution) are plotted on the x-axis.
* If there is a one-to-one relationship between the data + the theoretical quantiles, the data follow a nearly normal distribution.
* Since a one-to-one relationship appears as a straight line on a scatterplot, the closer points are to a perfect straight line, the more confident we can be the data follow a normal model.
* So when looking at a normal probability plot 🡺 looking for straight lines.
* Constructing a normal probability plot = calculating percentiles + corresponding z-scores for *each* observation in a data set, which can be quite tedious, especially w/ a large sample (what we like)
* Therefore, generally rely on software when making these plots.
* Example of data that do NOT really follow a normal distribution = height of NBA players from the
* 2008 and 2009 season.



* Since NBA players tend to be disproportionately taller compared to the general population, the distribution of their heights is left-skewed.
* On a normal probability plot, left skew appears as points bending *down* + to the right of the normal line.
* Can also see that these points have jumps 🡪 due to rounding when reporting heights.
* Just like with histograms, normal probability plots also reveal shapes of distributions.
* In a right-skewed distribution, points bend up + to the left of the line.
* If the distribution's left skewed, points bend down + to the right
* Distributions w/ short tails (narrower/skinnier than a normal distribution) follow an S shaped curve.
* Those w/ long tails (wider than a normal distribution) start below the line, bends to follow it, + ends above it.
* 
* Can also use the 68-95-99.7% rule to evaluate normality by assessing whether the distribution follows what's required by this rule.

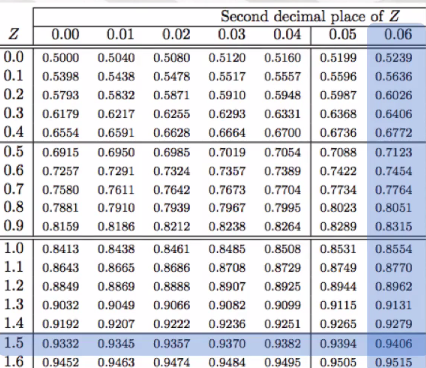


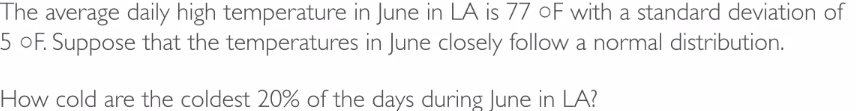
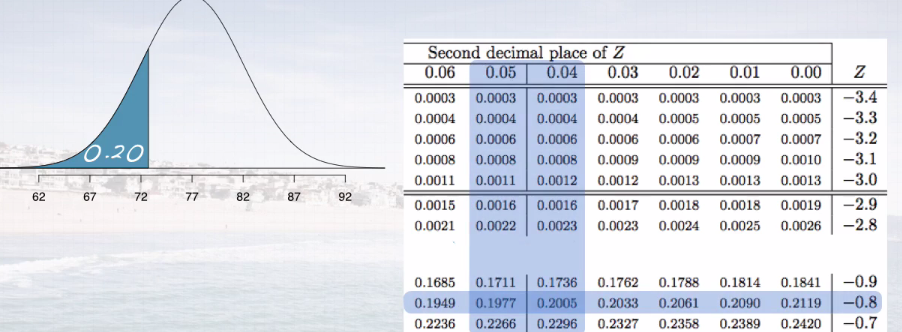
**Working W/ the Normal Distribution**

* 
* Told baggage weights are nearly normally distributed w/ mean 45 + SD 3.2
* Could write out our normal model 🡺 baggage ~ N(45, 3.2)
* 
* Roughly 5.91% of the passengers are expected to have baggage that weigh > 50 pounds.
* Can also do this calculation by hand using z-scores + the normal probability table.



* Z score = 1.56 🡪 refer to the table

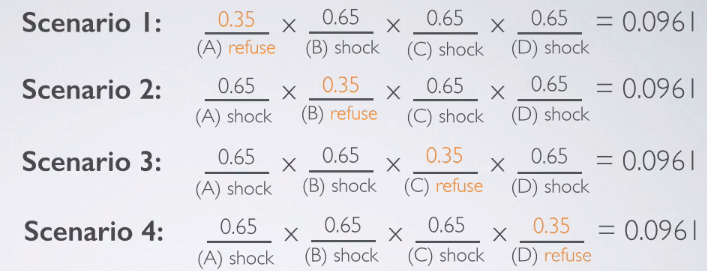


* Corresponds to the value = 0.9406 🡺 the AUC *below 50*.
* Find the complement 1 - .9406 = ~5.9
* 
* 
* By Z-scores:
* 
* Z = comes from somewhere between .1977 + .2005.
* Go w/ the closest to what we want (.2) 🡪 .2005 = a Z-score of -0.84.
* -.84 = (x – 77)/5 = -.84(5) + 77 = 72.8 degrees Fahrenheit.

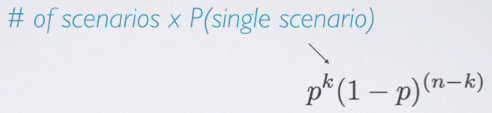
***4.2 The Binomial Distribution***

**Binomial Distribution**

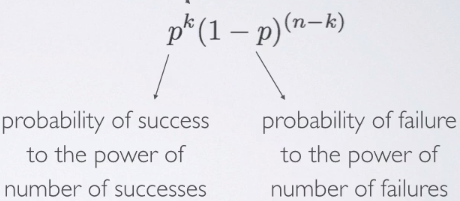
* **Milgram Experiment**, started by a Yale University psychologist in ‘60s = measured willingness of participants to obey authority figure instructing them to perform acts conflicting w/ conscience.
* Experimenter orders teacher to give severe shocks to a learner each incorrectly question answer
* The teacher = subject of the study + learner = just an actor + electric shocks are not real but pre-recorded sound played each time the teacher administers a shock.
* Found that ~65% of people obey authority + give such shocks.
* Over the years, additional research suggested this number is approximately consistent across communities + time.
* Each person in Milgram's experiment = a **trial** + is labeled a success = refuses to administer a shock + failure if administering a shock.
* Since only 35% of people refused to administer such a shock, P(S) = .35
* When an individual trial has only 2 possible outcomes, it is called a **Bernoulli random variable**.
* Suppose we randomly select 4 individuals (A, B, C, D) to participate in this experiment. What is the probability exactly 1 of them will refuse to administer the shock?
* In scenario 1 🡪 4 people in experiment + say the 1st person refuses + remainder all shock.
* Probability associated w/ refusing = 0.35 + probability associated w/ the rest = 0.65.
* Since we say the 1st person will refuse + 2nd-4th will shock, multiply these probabilities b/c these are independent trials since these are a random sample of people
* P(1) = .35\*.65\*.65\*.65 = 0.09611875 = 9% chance 1st person refuses + everybody else
* Continue on w/ other scenarios
* Even though order changes, overall probability has not changed b/c order in which you multiply numbers does not change the product.



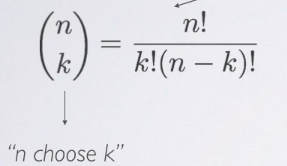
* Possible scenarios could be any of the 4 (OR)
* Since these are **disjoint scenarios**/**outcomes** 🡪 can't all happen at the same time.
* Therefore, add the probabilities + the overall probability exactly 1 person out of 4 refuses to administer the shock = **0.3844.**
* Could’ve actually arrived at this probability w/ P(S1) \* number of scenarios.
* This is a perfect setting for the **binomial distribution** = *describes the probability of having exactly* ***k*** *successes in* ***n*** *independent Bernouilli trials with probability of success,* ***p***.
* This probability can be calculated = # of scenarios \* probability of a single scenario
* The probability of a single scenario is simply **p^(k)\*(1 – p)^(n – k)**



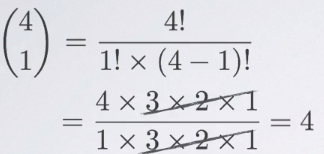
* This means the probability of success to the power of # of successes (our k) multiplied by the probability of failure to the power of # of failures.



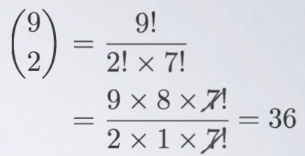
* To find the # of scenarios:
* To enumerate each possible scenario is only feasible w/ a small n
* Alternative approach = the **choose function** (n choose k) = useful for calculating # of ways to choose k successes in n trials



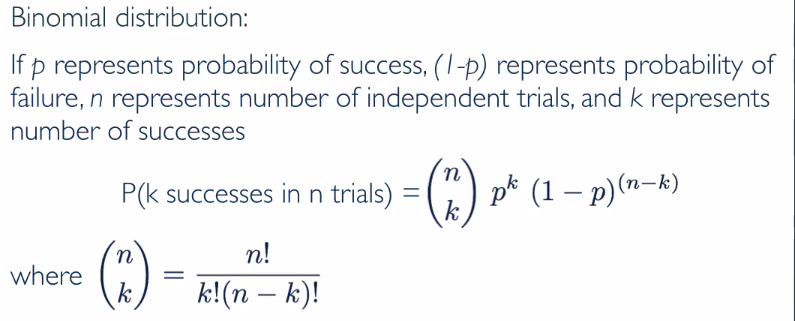
* Ex: Want to find how many scenarios yield 1 success in 4 trials.
* Here n = 4, k = 1; therefore, n choose k = 4! / (1! \* (3)!) = 4

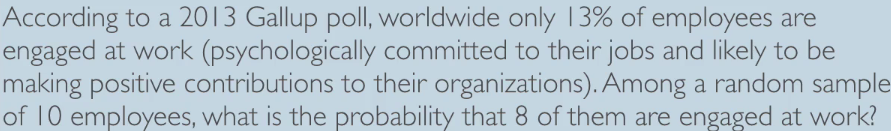


* Ex: n = 9 trials, k = 2 successes (*how many trials lead to 2 successes in 9 trials?*)

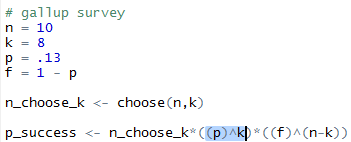


* 
* 
* Putting all of this together:



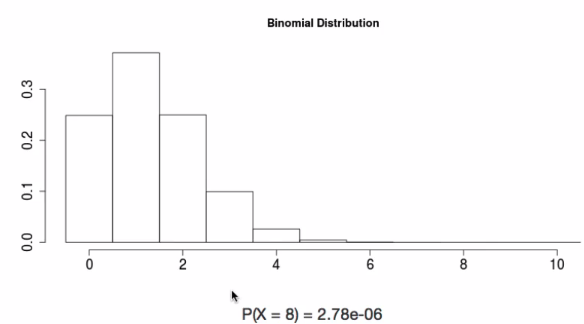
* What does it take for a random variable to follow a binomial distribution?
* 1) Trials must be independent.
* 2) Number of trials, n, must be *fixed*
* 3) Each trial outcome must be classified as either a success or a failure
* 4) Probability of success, p, must be same for each trial.
* goes hand in hand w/ the 1ST one, B/C if you have independent trials, you can be reasonably certain the probability of success is going to be the same for each
* 
* 1ST, parse through the given info
* n = 10, k = 8, p = .13, 1- p = .87





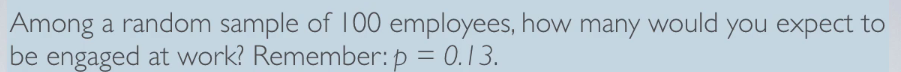
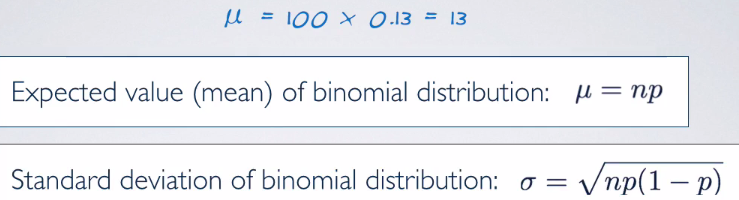
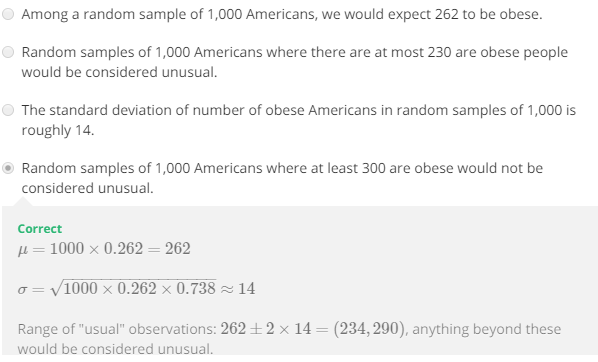






* Out of 10 employees, we’d expect so much fewer employees to be engaged than 8 if probability of success is only 13% = why we're looking for a highly unlikely outcome = a very low probability.
* 

* 
* Expected number of engaged employees, u, = 100\*0.13 = 13
* More formally, the **expected value (mean of the binomial distribution) is simply = n\*p**.
* This DOESN'T mean in every random sample of 100 employees exactly 13 will be engaged at work.
* In some samples the number of engaged employees will be fewer, and in others, more.
* How much would we expect this value to vary? 🡺 quantify variability around the mean using SD
* For a binomial distribution:
* 
* Plug in the values from the original survey = expect 13/100 employees are expected to be at engaged at work, give or take approximately 3.36.
* ***Note:*** Mean + SD of a binomial might not always be whole numbers + that's alright.
* These values represent what we’d expect to see *on average.*
* 
* 

**Normal Approximation to Binomial**

**Working W/ the Binomial Distribution**