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

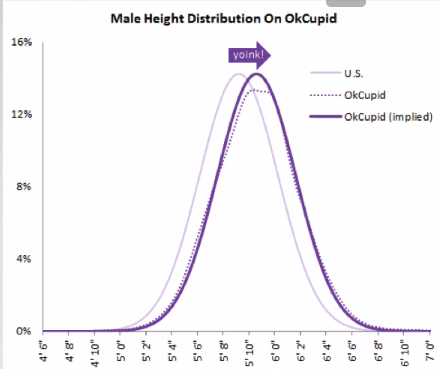
***COURSE 1 - Introduction to Probability and Data***

**WEEK 4- Probability Distributions**

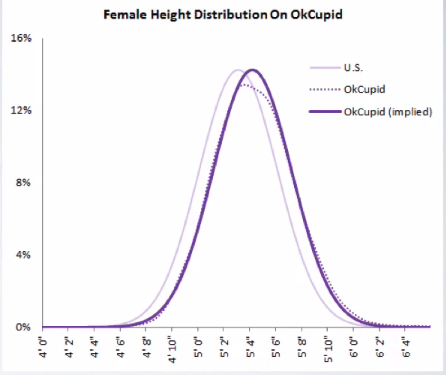
***4.1 The Normal Distribution***

**Normal Distribution**

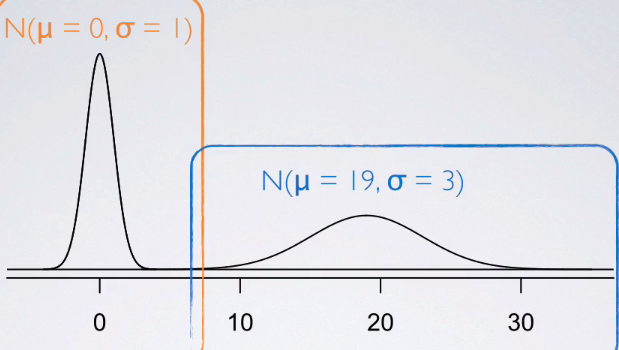
* Many variables in nature = nearly normally distributed, such as heights
* Ex: distribution of recorded heights of members of OkCupid + since members of this website are US residents + likely represent a random sample from the US population, we expect user heights to follow the same height distribution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



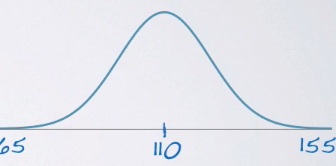
* Heights reported by men on OkCupid very nearly follow expected normal distribution, except are shifted to the right of where it should be.
* Appears males on OkCupid add on average a couple inches to their heights.
* Additionally, starting at about 5'8", the top of the dotted curve tilts even *further* rightward
* Indicates that closer to 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.

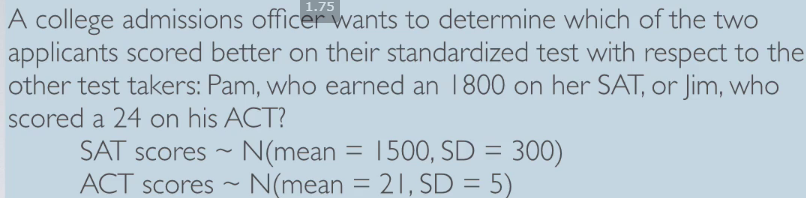


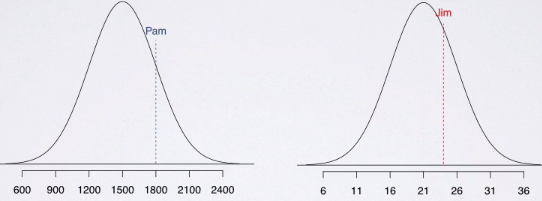
* **normal distribution/bell curve =** unimodal + symmetric
* Follows very strict guidelines about how variably data are distributed around the mean.
* While many variables are *nearly* normal, none are *exactly* normal due to these strict guidelines.
* Normal distribution has 2 parameters, mean, **μ**,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



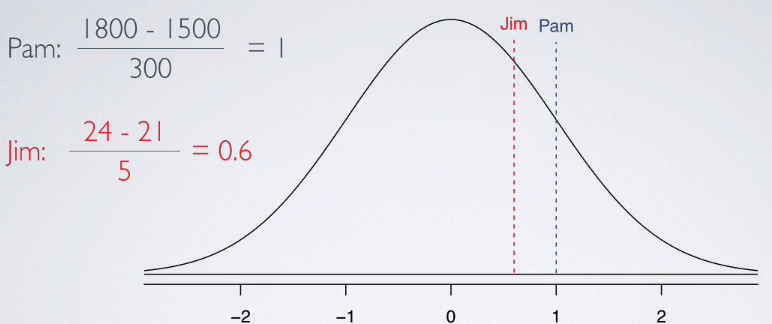
* See 2 normal distributions, 1 centered at 0 w/w SD = 1 + the other centered at 19 w/ SD = 3
* Good representation of how changing center + spread of a distribution changes overall shape
* For nearly normally distributed data:
* 68% falls within 1 SD of the mean. 95% falls within 2 SDs, + 99.7% falls 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 + only reports 3 statistics 🡺 mean = 110 BPM, min = 65 BPM, + max = 155 BPM.
* Which of the following is most likely to be the SD of the distribution?
* Given distribution is normal 🡪 So, the very 1ST thing to do = draw the normal curve + mark mean = 110 in the center w/ min = 65 + max = 155.



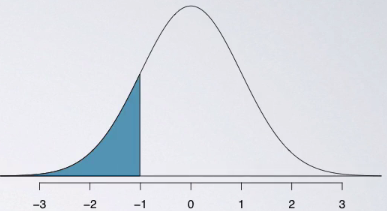
* *Almost all of the data lie w/in 3 SDs of the mean.*
* If SD = 5, the expected min + max = 110 +/- (3x5) 🡪 95 + 125
* **If SD = 15, the expected min + max = 110 +/- (3x15) 🡪 65 + 155**
* 
* Can draw the distribution of SAT scores + see Pam = 300 points above the mean + Jim = only 3 points above the 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 P + J 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.



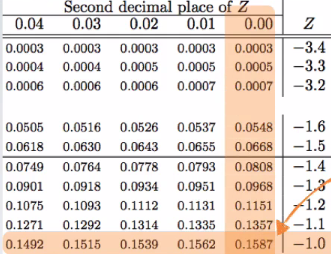
* These values = **standardized score** or **Z-score** = # of SDs an observation falls below/above the mean
* **Z-score** of an observation = that observation - the mean divided by the SD.
* By definition, the Z-score of the mean = 0
* Standardized scores are useful for IDing unusual observations = usually |Z-scores| > 2 = 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 DP.
* Graphically = area below the probability distribution curve, **AUC**, to the left of that observation.

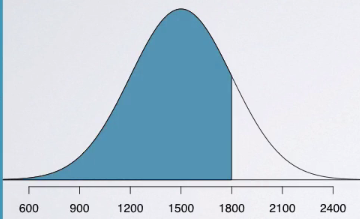


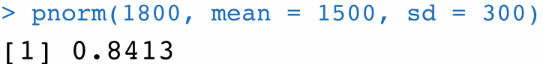
* *Why can we only use Z-scores under normal curves, but not in a distribution of a different shape?*
* Can *always* calculate *percentiles* for *any* sort of distribution, *except* if the distribution does NOT follow a nice unimodal symmetric normal shape (need calculus for that = integral)
* In this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



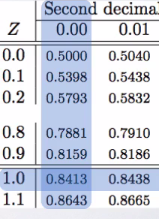
* Can avoid computation altogether + use a normal probability table
* Locate the Z-score on the edges of the table + grab the associated percentile value given in the center of the table.



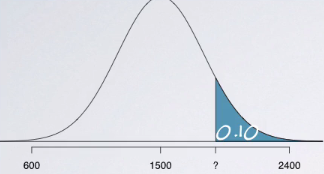
* Computational approach is a little less archaic, but tables = actually very useful for getting a conceptual understanding of AUC.
* Ex: We know SAT scores are distributed normally w/ mean 1500 + SD 300 + also know Pam earned an 1800 🡪 find out her percentile score.
* Given/Assume the distribution is normal 🡪 draw the curve, mark the mean, + shade the area of interest (scores below 1800)
* 



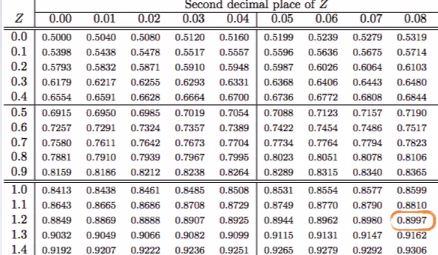
* Get a percentile of 0.8413, meaning Pam scored better than 84.13% of SAT takers.
* Could also use the table to arrive at the same conclusion.
* First, calculate the Z-score as the observation 🡪 1800 – 1500 / 300 = 1
* In the table, look for the Z-score of 1 + get the same probability, 0.8413 = probability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



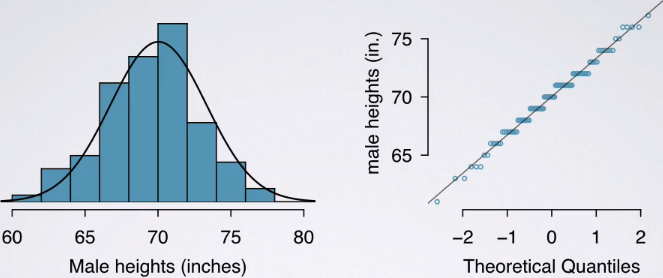
* We know mean + SD 🡪 using the table, find the Z-score associated w/ the 90th percentile.



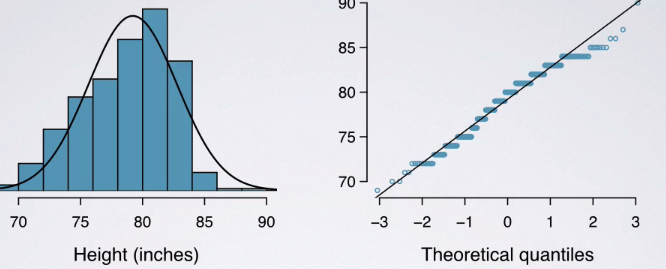
* Closest we can get is 0.8997 🡪 Z-score = 1.28.
* We know 1.28 == the unknown observation, X, minus the mean, divided by the SD.
* 1.28 = (X – mean) / SD == 1.28\*(300) + 1500 == **1884 =** the cutoff value for top 10%, or bottom 90%, of the distribution of SAT scores
* If you have scored above 1884,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

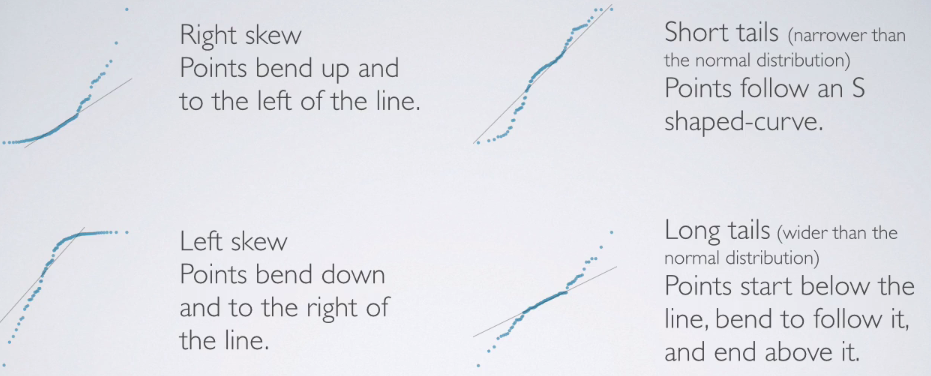
**Evaluating the Normal Distribution**

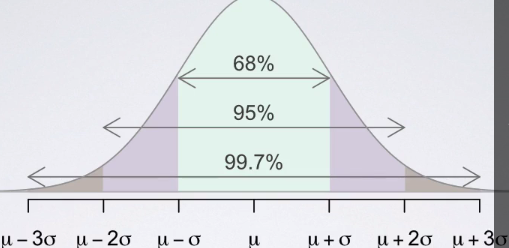
* How to evaluate whether a distribution is nearly normal or not.



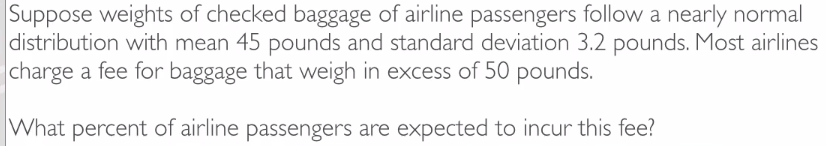
* Histogram + a **normal probability/quantile-quantile (QQ) plot** of a sample of 100 male heights.
* **Normal probability/QQ plot** = actual data is plotted on the y-axis + *theoretical* quantiles (quantiles that follow a normal distribution/what would be expected under the normal distribution) are 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 = data is normal if line is straight
* *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 QQ plot involves 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 2008-2009 season.



* Since NBA players tend to be disproportionately taller compared to general population (nearly normal), the distribution of their heights is left-skewed.
* On a normal probability/QQ 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.
* So just like with histograms, normal probability/QQ plots also reveal shapes of distributions.
* In a right-skewed distribution, points bend up + to the left of the line.
* In a left-skewed distribution, points bend down + to the right of the line.
* Distribution w/ short tail (narrower/skinnier than normal distribution) follows an S shaped curve
* Distribution w/ long tail (wider than a normal distribution) start below the line, bend to follow it, + end 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**

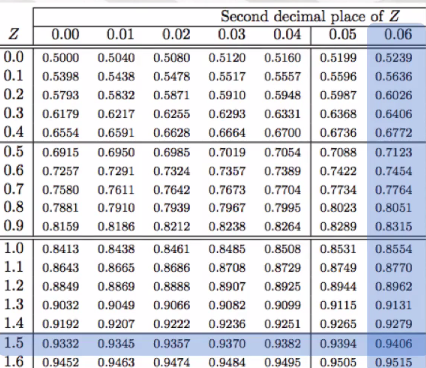
* 
* Told baggage weights are nearly normally distributed w/ mean 45 + SD 3.2
* Could write out our normal model 🡺 **baggage ~ N(45, 3.2)**

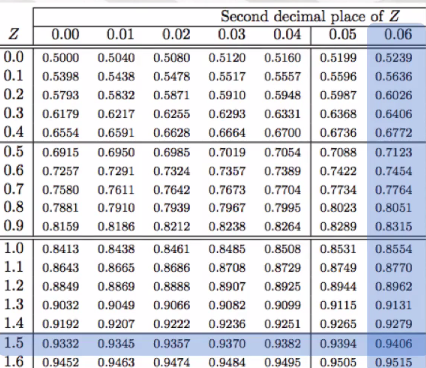


* Roughly 5.91% of the passengers are expected to have baggage that weigh > 50 pounds.
* Can also do this calculation by hand using z-scores + the normal probability table.



* Z score = 1.56 🡪 refer to the table

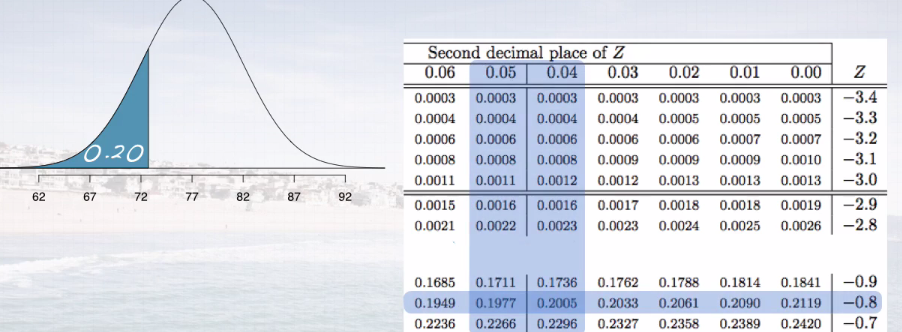




* Corresponds to the value = 0.9406 🡺 the AUC *below 50*.
* Find the complement 1 - .9406 = ~5.9
* 



* By Z-scores:

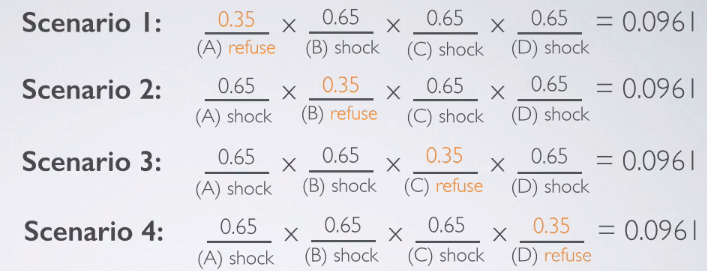


* Z = comes from somewhere between .1977 + .2005.
* Go w/ the closest to what we want (.2) 🡪 .2005 = a Z-score of -0.84.
* -.84 = (x – 77)/5 = -.84(5) + 77 = 72.8 degrees Fahrenheit.

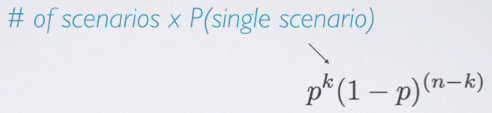
***4.2 The Binomial Distribution***

**Binomial Distribution**

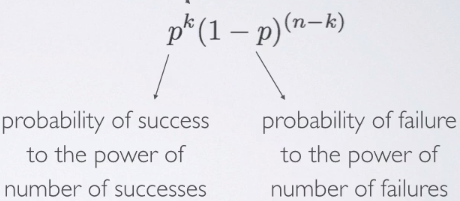
* **Milgram Experiment**, started by a Yale University psychologist in ‘60s = measured willingness of participants to obey authority figure instructing them to perform acts conflicting w/ conscience.
* Experimenter orders teacher to give severe shocks to learner for each incorrect question answer
* Teacher = subject of the study + learner = an actor + electric shocks are not real but pre-recorded sounds 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** when a teacher refuses to administer a shock + **failure** if a teacher administers 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**.
* *Fixed number of trials (a fixed SS, n)*
* *In each trial, event of interest either occurs or does not.*
* *Probability of occurrence (or not) is same on each trial.*
* 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(1st) = .35\*.65\*.65\*.65 = 0.096119 = **9% chance 1st person refuses + everybody else shocks**
* 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.



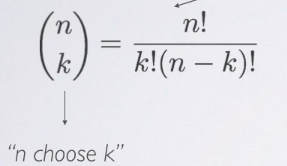
* Possible scenarios could be any of the 4 (it’s an OR problem)
* These are **disjoint scenarios**/**outcomes** 🡪 can't all happen at the same time.
* Therefore, *add* the probabilities 🡪 overall probability exactly 1 person out of 4 refuses to administer the shock == .961\*4 = **0.3844.**
* Could’ve actually arrived at this probability w/ P(1st) \* number of scenarios.
* This is a perfect setting for the **binomial distribution** = *describes the probability of having exactly* ***k*** *successes in* ***n*** *independent Bernoulli trials given 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)**



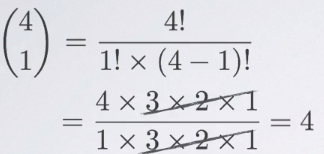
* This means the probability of success to the power of # of successes (k) multiplied by the probability of failure to the power of # of failures.



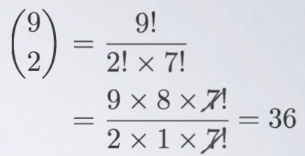
* How to find the # of scenarios:
* To be able o enumerate each possible scenario is only feasible w/ a small n
* Alternative approach = **choose function** (n choose k) = useful for calculating # of ways to choose k successes in n trials



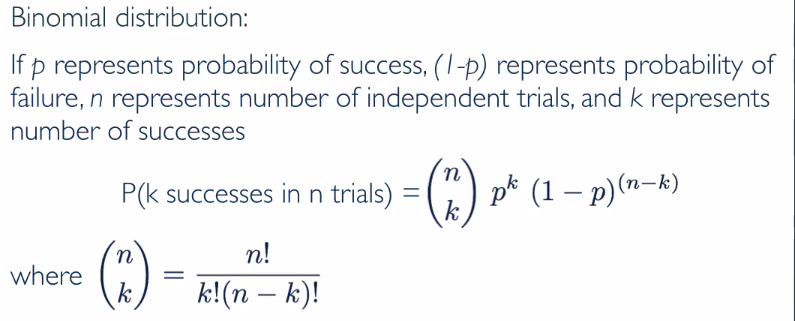
* Ex: How many scenarios yield 1 success in 4 trials?
* Here n = 4, k = 1; therefore, n choose k = **4! / (1! \* (3)!) = 4**

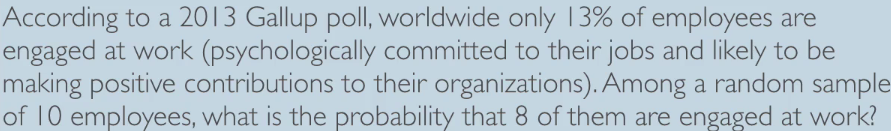


* Ex: n = 9 trials, k = 2 successes (*how many trials lead to 2 successes in 9 trials?*)

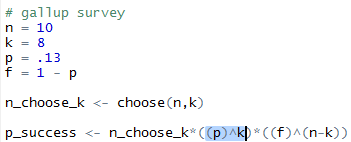


* 
* 
* Putting all of this together:



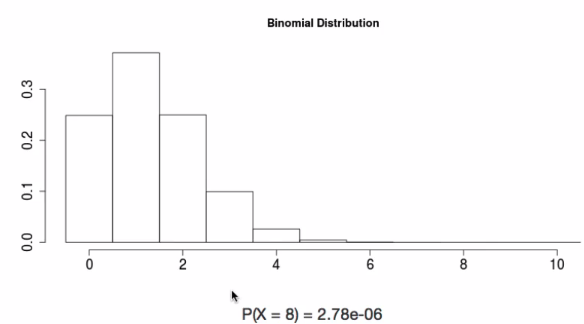
* What does it take for a random variable to follow a binomial distribution?
* 1) Trials must be independent.
* 2) Number of trials, n, must be *fixed*
* 3) Each trial outcome must be classified as either a success or a failure
* 4) Probability of success, p, must be same for each trial.
* goes hand 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





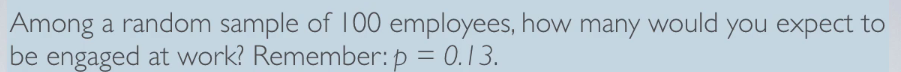
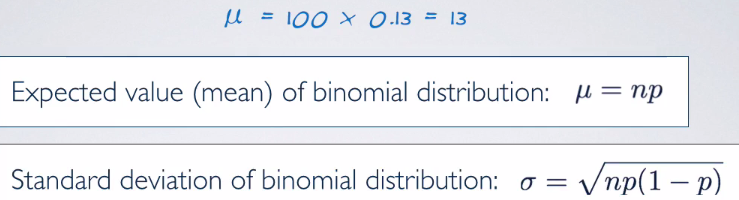
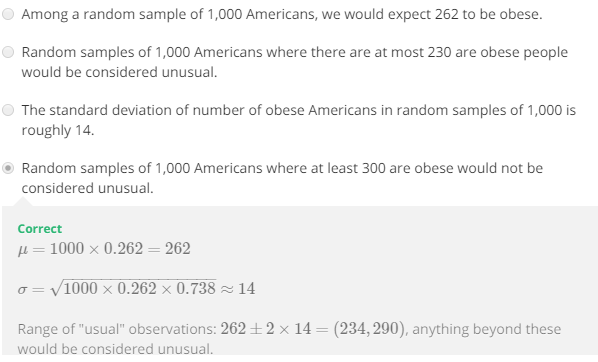






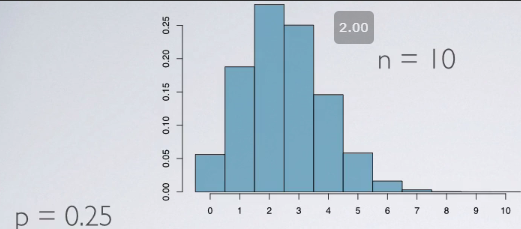
* Out of 10 employees, we’d expect so much fewer employees to be engaged than 8 if probability of success is only 13% = why we're looking for a highly unlikely outcome = a very low probability.
* 

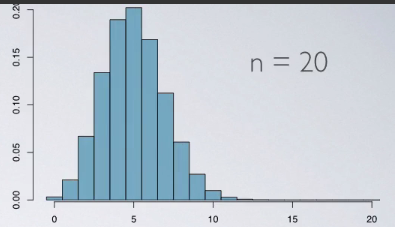
* 
* Expected number of engaged employees, u, = 100\*0.13 = 13
* More formally, the **expected value (*mean of the binomial distribution*) is simply = n\*p**.
* DOESN'T mean in every random sample of 100 employees *exactly 13* will be engaged at work.
* In some samples, number of engaged employees will be fewer, and in others, more.
* How much would we expect this value to vary? 🡺 must 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**

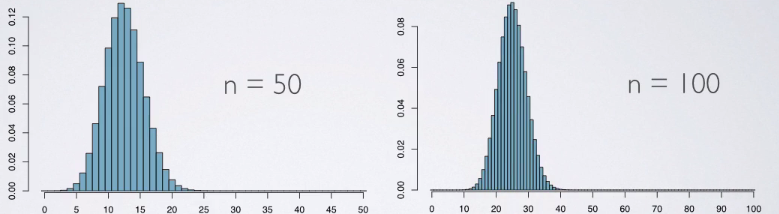
* Shapes of binomial distributions change as we tweak parameters (# of trials or probability of success) + when # of trials increases, the shape of the binomial *actually starts looking closer to a full normal distribution*
* Ex: Binomial random variable w/ P(success) = .25
* Here is the distribution when n = 10.



* Each bar represents a potential outcome 🡪 # of successes could range anywhere from 0-10, so therefore we have 11 bars
* Heights of bars = *likelihood* of outcomes
* Ex: P(k = 0) 🡺 P(failures)^(n – k) = P(.75)^10 = .056
* W/ n = 10 + p = 0.25, expected number of successes is 2.5 (highest bar + distribution is centered around this value)
* So, the binomial distribution w/ p = 0.25 + n = *10* is *right-skewed*
* Increase SS a bit keeping p constant at 0.25 to n = *20*:

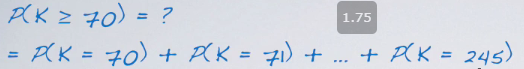


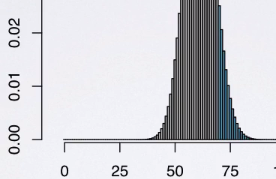
* see a change in the center of the distribution = expected since n\*p is now different
* Also see a change in shape = much less skewed



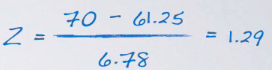
* Increase n more = Distribution looks no different than the normal distribution.
* Why this might be of interest?
* Ex: Recent study found FB users “get more than they give.”
* 40% of FB users in a sample made a friend request, but 63% *received* at least 1 request.
* Users pressed “like” to friends' content an average of 14 times, but had *their* content liked an average of 20 times.
* Users sent 9 PM’s on average but received 12.
* 12% of users tagged friends in a photo, but 35% were themselves tagged in a photo.
* What explains this phenomenon? 🡪 **power users** = contribute much more content than typical user
* Other findings = 25% of FB users are considered power users + an average FB user has 245 friends
* Looking for probability an average FB user w/ 245 friends has 70+ friends who are power users
* P(power) = .25% + average FB user has 245 friends 🡪 n = 245 (b/c looking at average users)
* Probability interested in = P(70+ PU friends) 🡪 translates to # of successes >= 70.
* n = 245 trials, a *fixed #* + each trial outcome is classified as *success* or *failure* (PU or not)
* Probability of success = *same* for each trial, 25%
* Assume trials are *independent* (might not be in reality, since if you're the type of person to have some friends who are PUs, the others might be more likely to be as well, but assume independence for the sake of example)



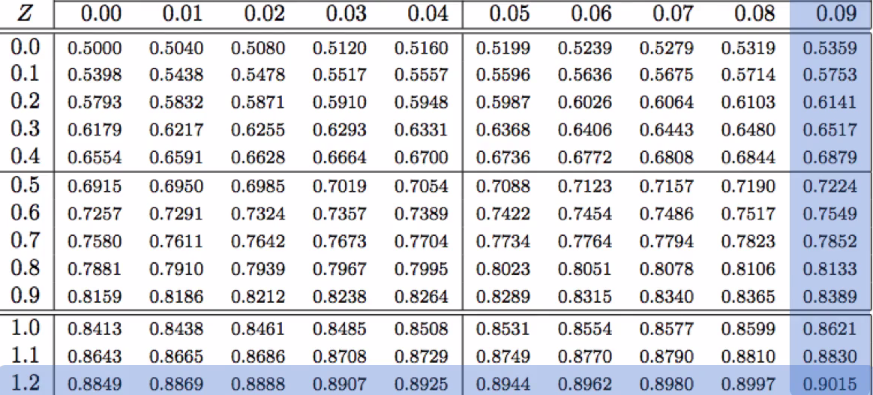
* We're interested in P(70+) = 70+ PUFs among 245 friends
* Interested in the *sum* of probabilities of *each* outcome 🡪 70-245.
* Can calculate each probability using the binomial formula + add them up, but this is tedious
* 
* *This is where the resemblance between the binomial + normal distributions comes in very handy*.



* Blue-shaded area of interest can just as well be calculated as the AUC of the smooth normal curve that *closely resembles* the more-jagged binomial distribution.
* Calculating AUC for the normal curve = much simpler task than calculating individual binomial probabilities for all outcomes + adding them up
* To calculate a normal probability 🡪 need a little more info on parameters of the normal distribution
* Can be *estimated* by the mean + the SD of the *original* binomial distribution 🡪 N(mean, SD)
* Estimate mean = n\*p 🡪 245\*0.25 = **61.25** Estimate SD = square root(245\*.25\*.75) = **6.78**
* *So among 245 friends, we expect 61.25 PUs, give or take 6.78.*
* Given an observation, the mean, + the SD, can calculate AUC via a z score



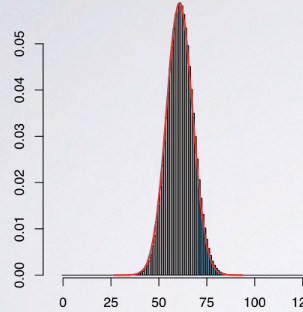
* Then find probability of z-score > 1.29, since we shaded the AUC *beyond* the observation of interest



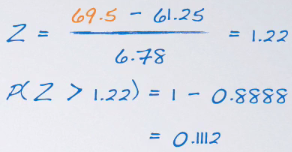
* See 0.9015 🡪 P(z > 1.29) = 1 - .9015 = **0.0985** = ***9.85% chance an average FB user (245 friends) has at least 70 friends who are considered power users***
* Can also directly calculate this probability using R



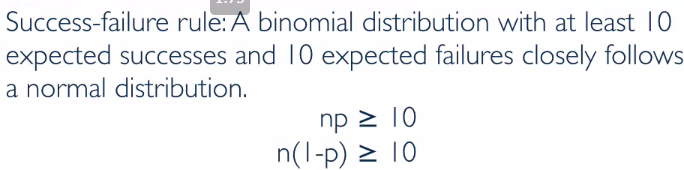
* 11.3% is different than 9.85% found before b/c the approach is the **normal APPROXIMATION** to the binomial after = just an approximation + not an exact result.
* If we need the EXACT probability, this difference may be frustrating.
* Let's take a closer look at the binomial distribution + the normal approximation to it.



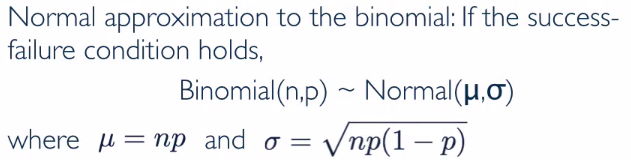
* The red normal curve is slightly different than the bars representing the EXACT binomial probabilities = falls a little bit short.
* Also, under a CONTINUOUS NORMAL distribution, probability of *exactly* 70 successes is undefined
* So the shaded area above 70 doesn't *exactly* include the probability of 70 successes.
* **A common fix to this problem is a 0.5 adjustment to the observation of interest.**
* So calculate a z-score using x = 69.5 as opposed to 70 which yields an **adjusted z score** of 1.22
* Everything else about the method stays the same + the result we get is much closer to the exact result from the binomial distribution, **0.1112 vs 0.113.**



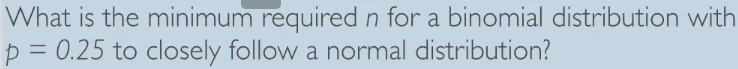
* What if we couldn't plot the binomial distribution? What are some guidelines we can use to determine whether a SS (# of trials) is large enough, such that we can be *confident in estimating the binomial distribution using the normal*?
* In other words, *how can we tell if the shape of the binomial distribution is going to be unimodal + symmetric + closely follow the normal distribution?*
* Rule of thumb = the **Success-Failure Condition** = a binomial distribution *w/ at least 10 expected successes* + *10 expected failures* closely follows a normal distribution.
* So BOTH n\*p needs to be >= 10 AND n\*(1 – p) needs to be >= 10



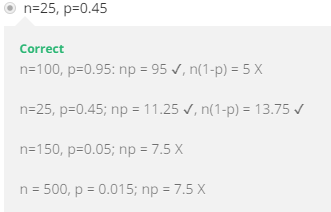
* In cases where it DOES, we CAN approximate the binomial distribution w/ the normal, where the parameters of the normal distribution = the estimated mean + estimated SD of the binomial



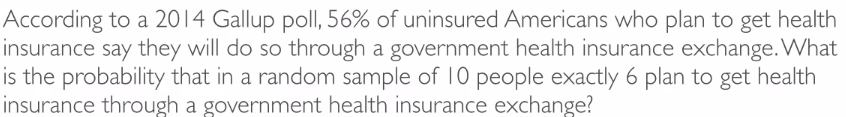
* The **0.5 adjustment** 🡪 makes probabilities calculated using the normal *approximation* much closer to the *exact* probabilities from the binomial distribution.
* Don’t focus on those details a whole lot, but instead try to focus on the bigger picture.
* Remember **the binomial distribution w/ sufficient SS starts to look nearly normal.**
* This is important + we're emphasizing this b/c when doing **inference** for categorical variables w/ 2 outcomes (like Bernoulli outcomes that follow a binomial distribution
* Going to make use of the fact that the distributions start to looks, nearly normal
* Going to apply methods based on the normal distribution to do inference for these variables
* Quick practice problem:



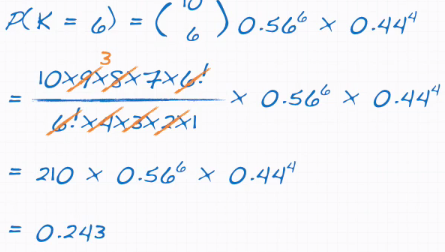
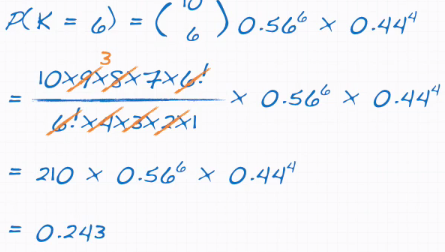
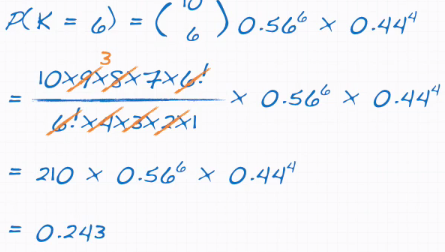
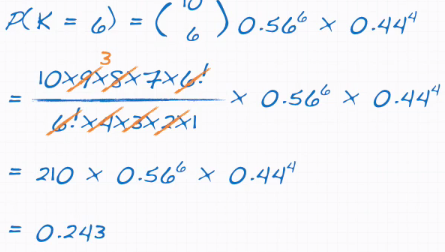
* We know n\*p and n\*(1 – p) both need to be >= 10
* So for both of these equations we want to solve for n + then take the MAXIMUM of those since, that's going to be the MINIMUM required SS
* 
* Solving, n needs to be >= 40 or >= 13.33.
* So **we need at least 40 observations for a binomial distribution w/ p equals 0.25, to closely follow a normal distribution.**
* 

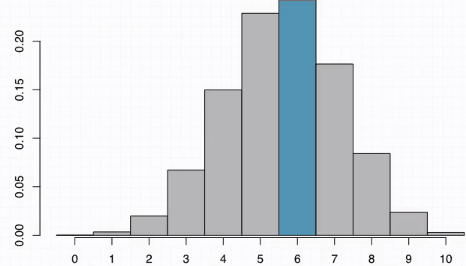


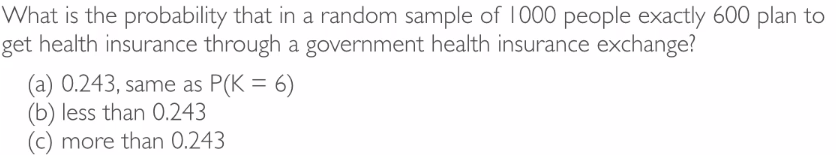
**Working W/ the Binomial Distribution**

* 
* Given: n = 10, k = 6, P(s) = .56

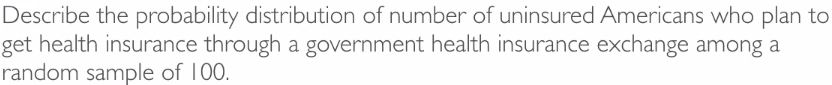


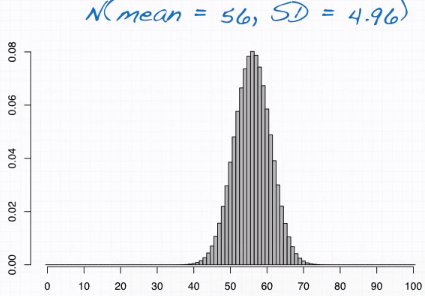
*  

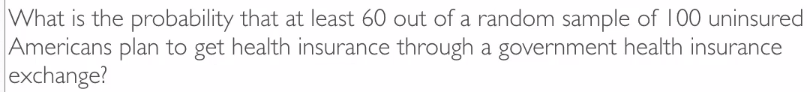


* 
* Once again, P(S) =.56 + before, n = 10 so the expected # of successes = 10\*0.56 = 5.6.
* The difference between expected + desired number of successes was 6 - 5.6 = 0.4.
* In *this* exercise, n = 1000, so expected # of successes = 1000\*0.56 = 560.
* The difference between expected + desired # of successes is 600 – 560 = 40.
* Here*, the desired outcome is much farther than the expected outcome*
* Based on law of large numbers, obtaining 600 successes when expected is 560 should be a much less likely outcome than obtaining 6 successes when the expected is 5.6.
* Therefore the answer is something less than 0.243.

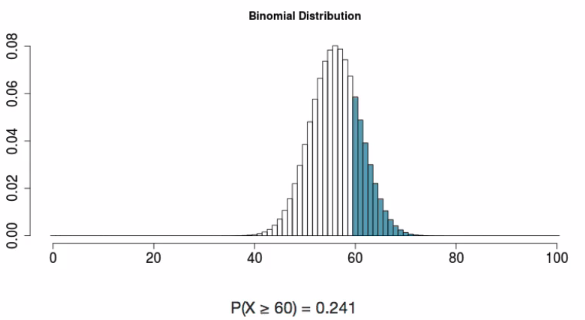


* 
* Still, P(S) = .56, n = 100 which seems larger than 1st, lower than 2nd example
* Expected # of success = 56 Expected # of failures = 44
* Both are > 10 🡪 **Success-Failure Condition** = this binomial distribution closely follows a normal
* Calculate estimated parameters to fully describe the distribution
* Expected mean = n\*p = 56 Expected SD = = = 4.96

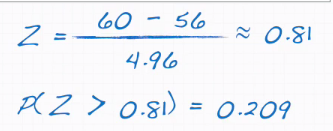


* 

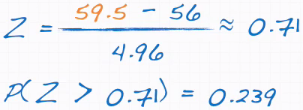


* When finding normal probabilities, calculate the z-score = 0.81 + find it on a table 🡪 find the probability to be roughly 0.209.
* Once again, this is a little lower than calculated w/ R.



* Remember this discrepancy is mostly due to the fact that under the normal distribution, probability of *exactly* 60 successes is undefined.
* To account for that, apply a 0.5 correction to the observation of interest + work through the problem again.



* The updated probability = 0.239, much closer to the exact probability calculated using the exact binomial distribution