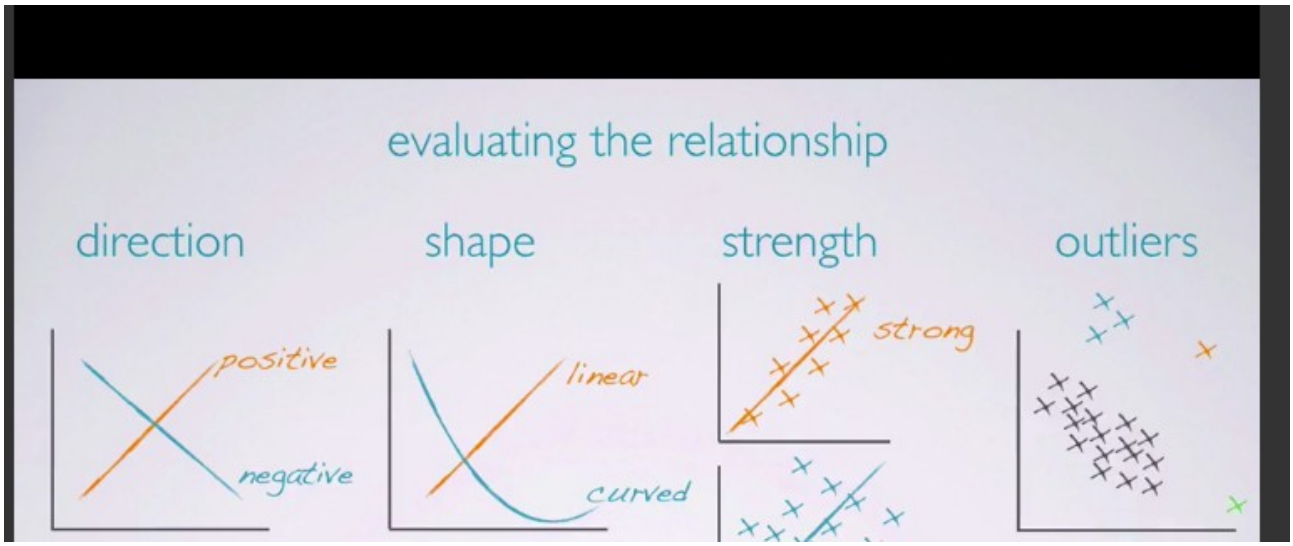
COURSE 1 → WEEK 2 NOTES.

Topic 1.6 of the book.
Visualizing the Numeric data:

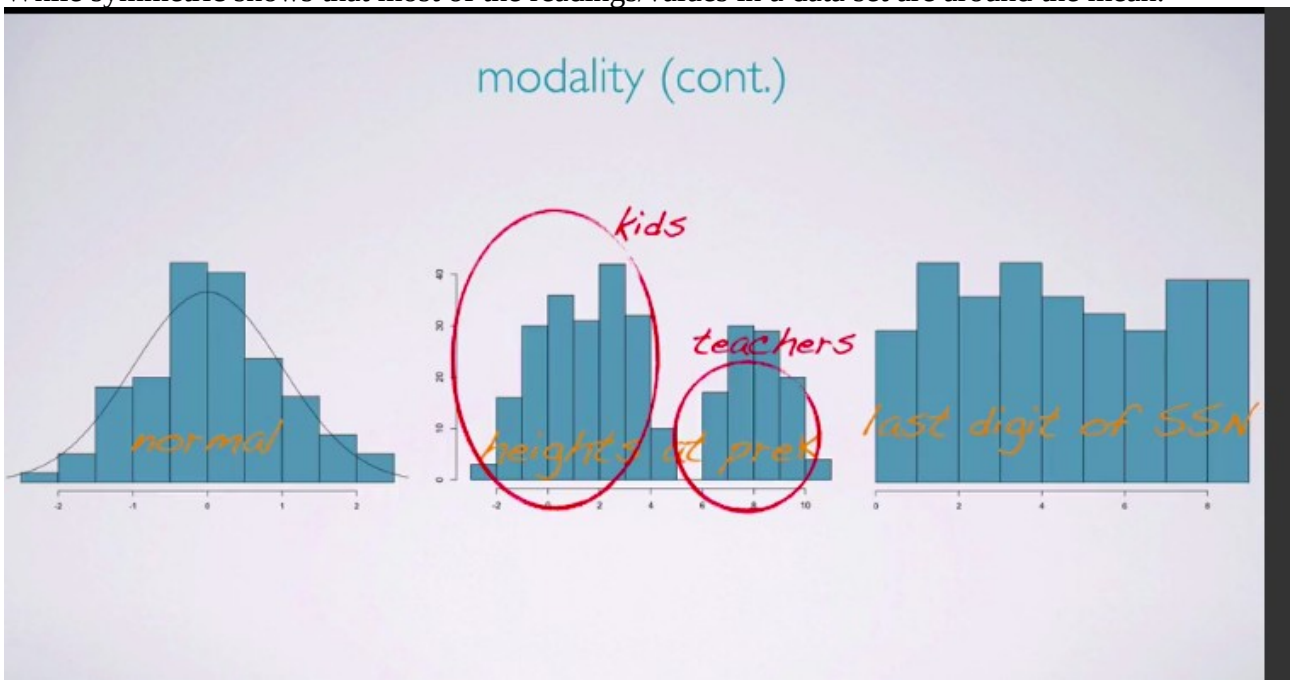


histogram helps us to determine the dist of the data, it provides a view of data density.

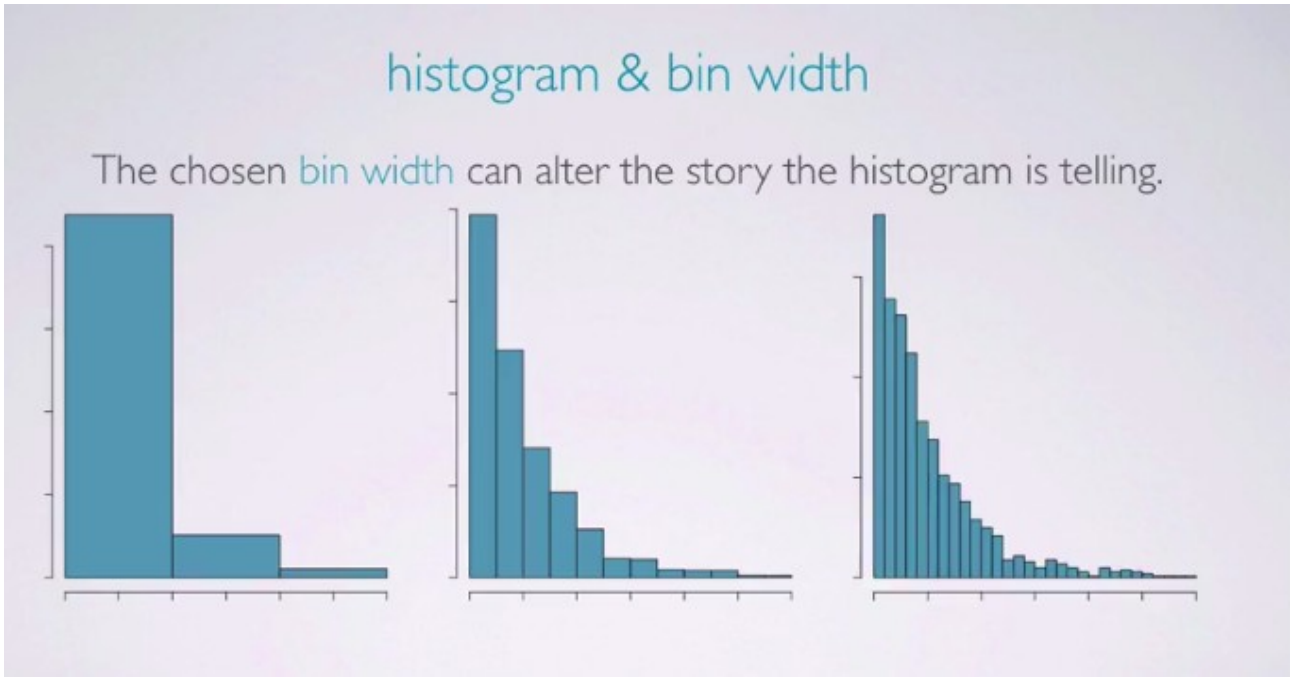
Skewness : dists are skewed to the side of the long tail.

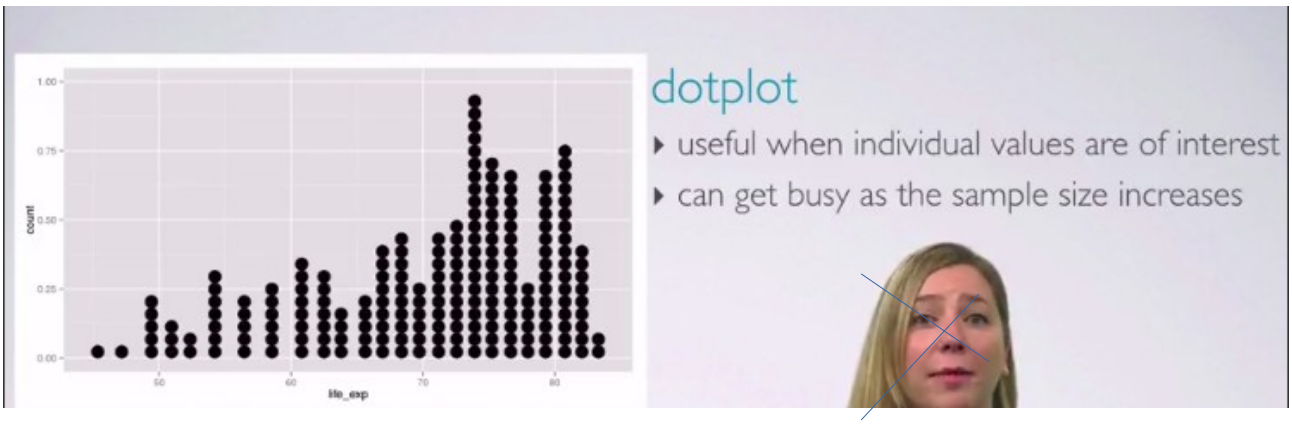
The bimodal shows that there are 2 groups in a data set while uniform shows that there's no trend every value in a data is equally likely to occur.

While symmetric shows that most of the readings/values in a data set are around the mean.

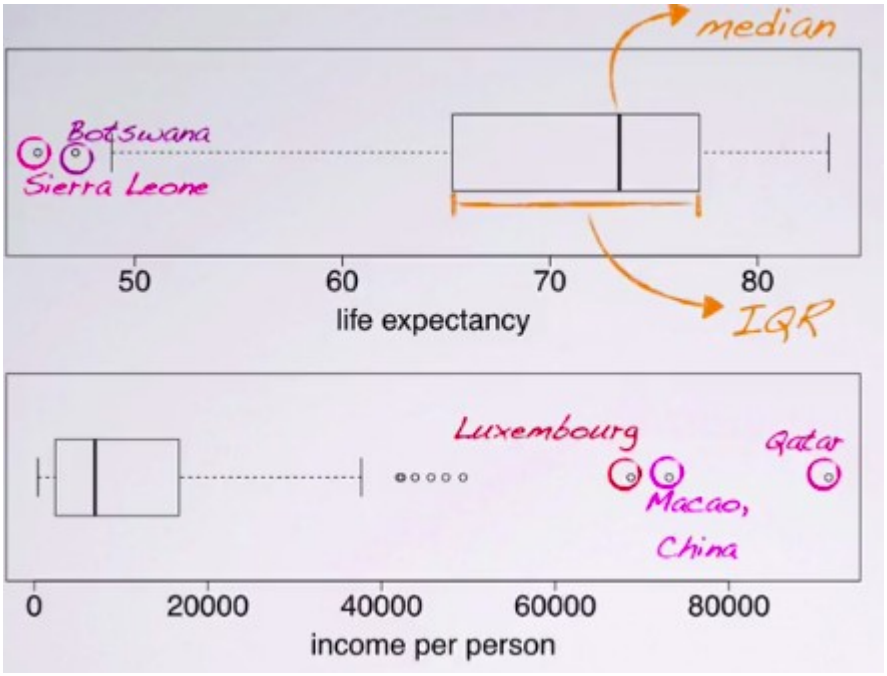


Here the last graph of the above pic is the plot of the last digit of Social security number of US citizen ,which is a Uniform dist as expected.

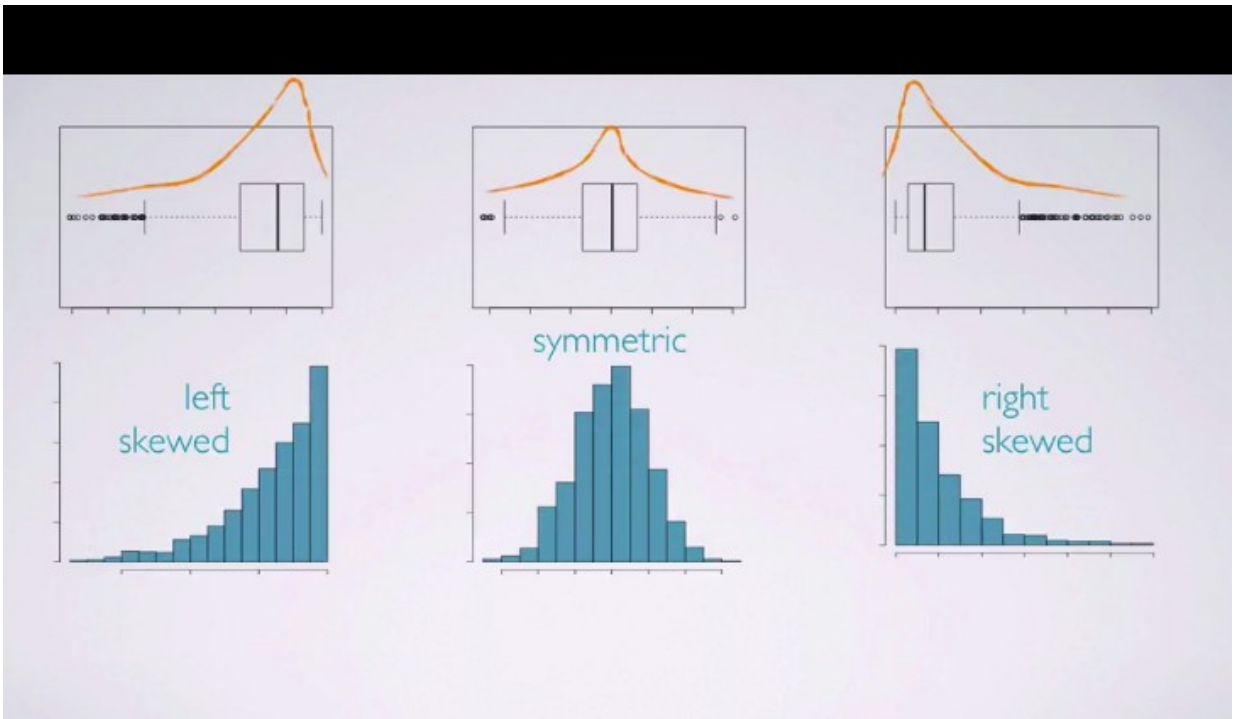




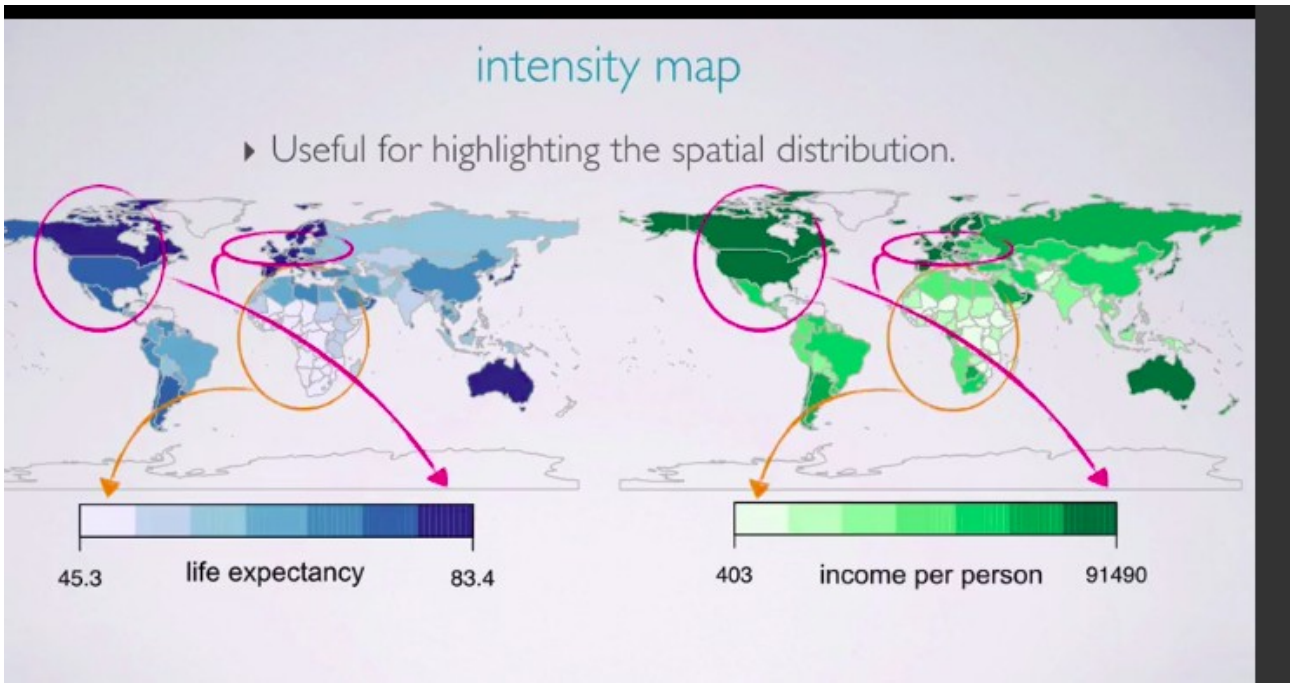
BOX PLOTS: useful for highlighting outliers and IQR.(inter quartile range):



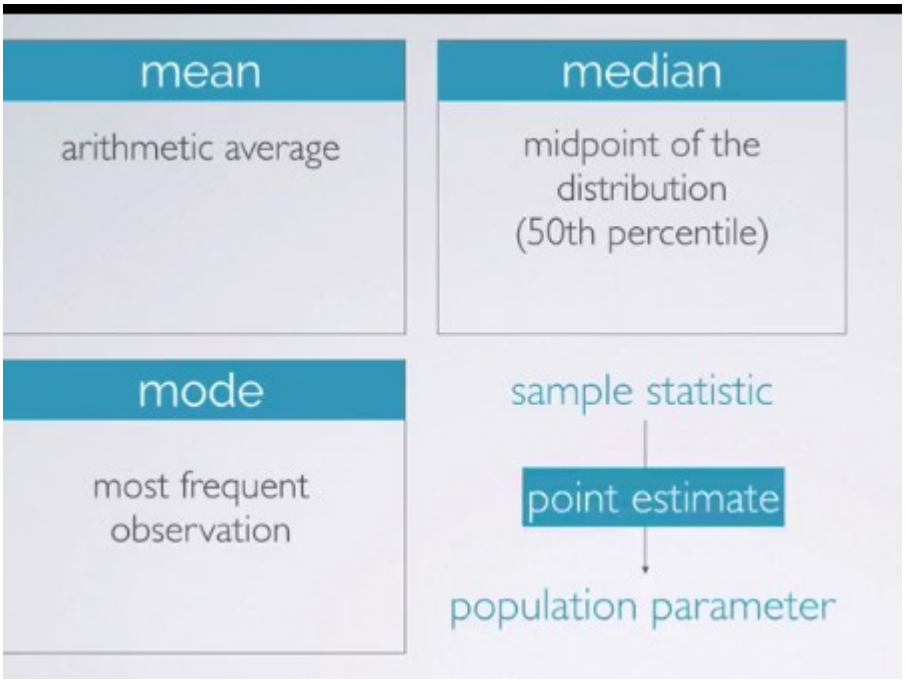
they donot show the modality, but they do show skewness.



INTENSITY MAP:



MEASURE OF CENTER:



Another key characteristic that is of interest is the center of the distribution, commonly used measures of center are the mean, which is simply the arithmetic average. The median, which is the mid point of the distribution or in other words the 50th percentile and the mode which is the most frequent observation. If these measurements are calculated from a sample, they're called sample statistics. Sample statistics are point estimates for the unknown population parameters.

We usually use letters from the Latin alphabet when denoting sample statistics, and letters from the Greek alphabet when denoting population parameters. For example, the sample mean is x bar and the population mean is mu.

Which statement is true:

(Hint: Sketching the distributions might be useful.)

- In a symmetric distribution, more than 50% of the data are below and less than 50% are above the mean.
- In a left skewed distribution, roughly 50% of the data are below and 50% are above the mean.
- In a right skewed distribution, less than 50% of the data are below the mean.
- In a left skewed distribution, less than 50% of the data are below the mean.

less than 50 is noticable.

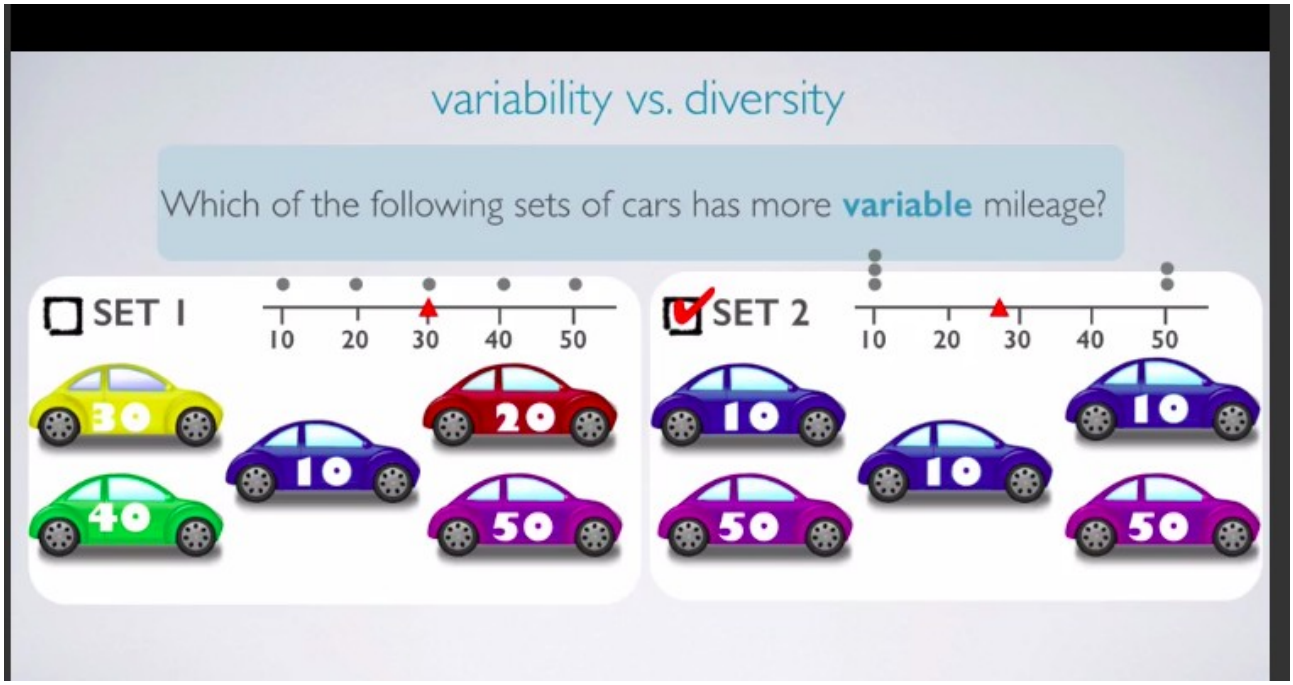
Correct

Since the median marks the 50th percentile, and in a left-skewed distribution the mean is smaller than the median, less than 50% of the data will be smaller than the mean.

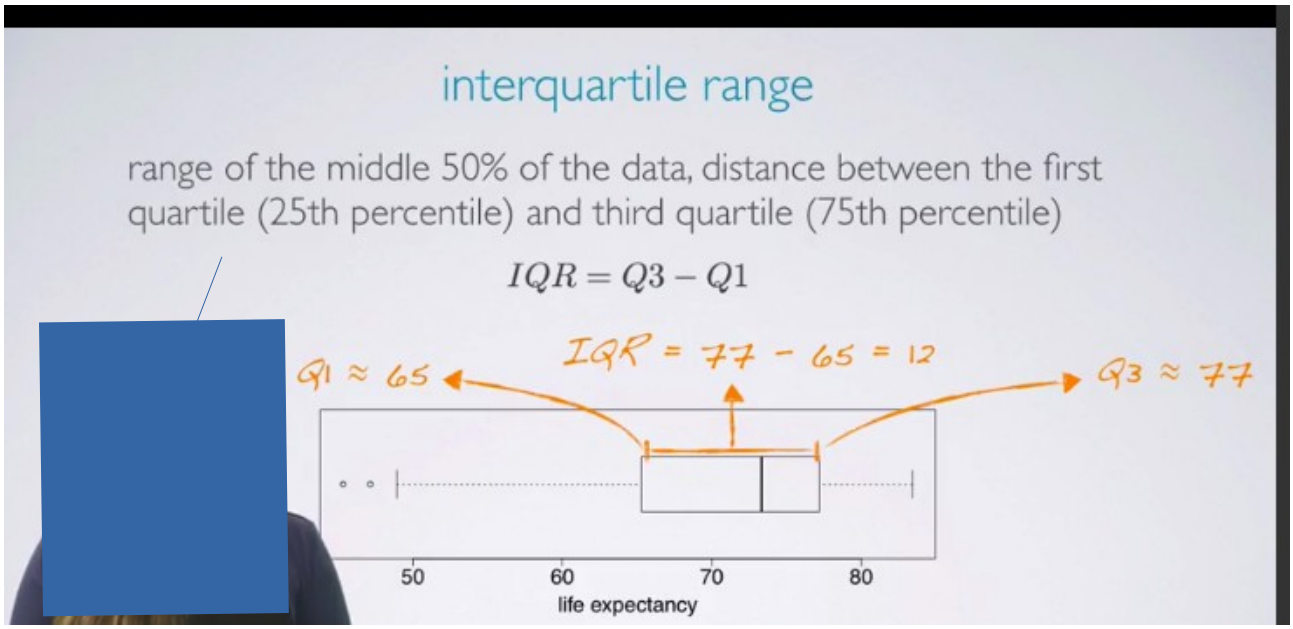
Measure of spreads:

range,variance,std and IQR.

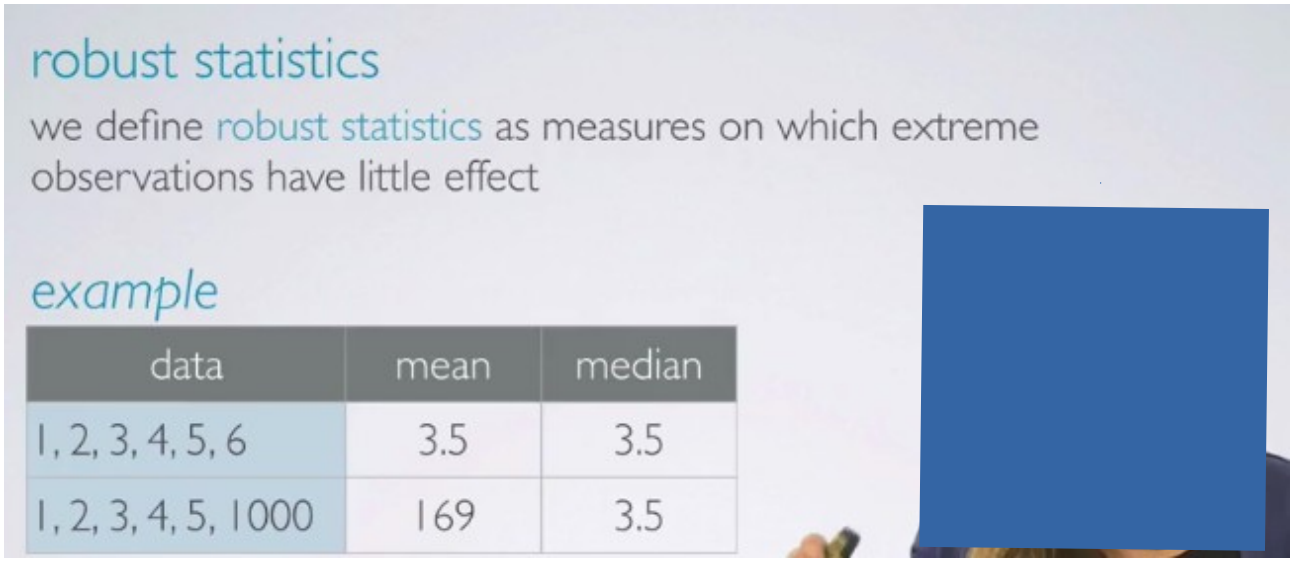
The set 1 is more diverse but the set 2 is more variable bcz the ***means*** are 30 and 26 which is



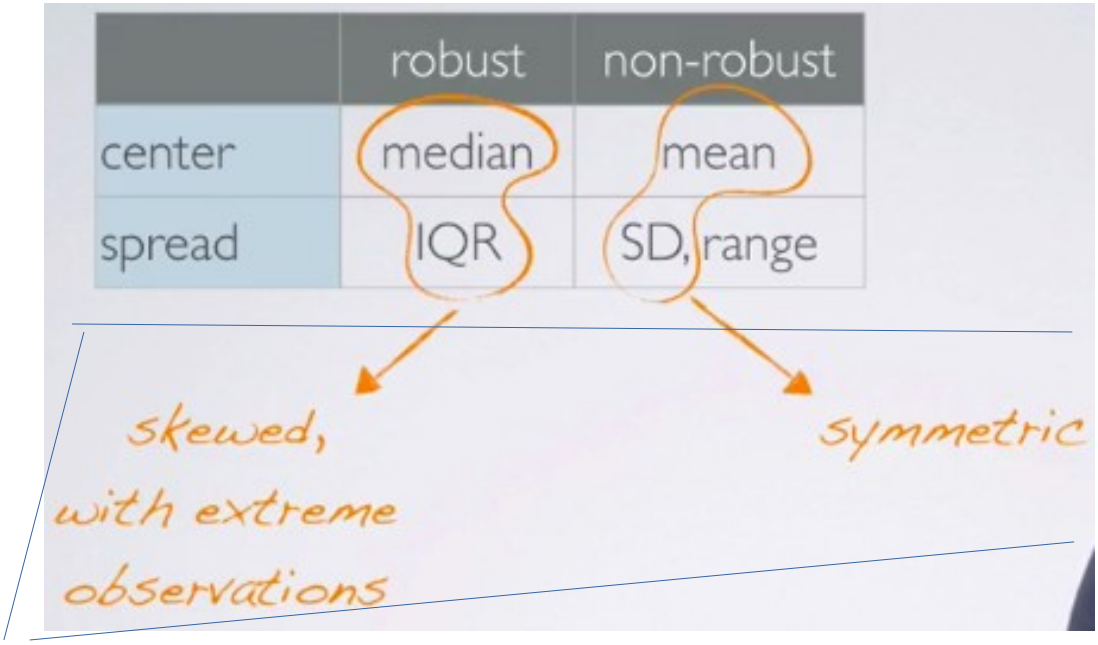
roughly equal but the values aren't centered in the set 2 they are at end points.



Robust statistics:



median is more robust then the mean because it doesn't relies on the extreme values.



Robusts are used for skewed distributions while we use mean and std for symmetric.

You have collected annual salary data from a large company with many employees who make below $100,000 per year, a fewer number of managers with salaries between $100,000 - $150,000, and a few high-level executives whose salaries can go beyond $1 million per year. Determine what shape the distribution of these salaries would be expected to follow, and accordingly decide whether the median or the mean would best represent a typical salary for an employee working at this company.

- right skewed, mean is a better measure of typical salary
- right skewed, median is a better measure of typical salary

Correct
Majority of the distribution is below $100,000 and as the salary increases the number of employees who make as high as salary decreases, hence giving the distribution a long tail on the right (right skewed). For skewed distributions the median is a better measure of the typical observation.

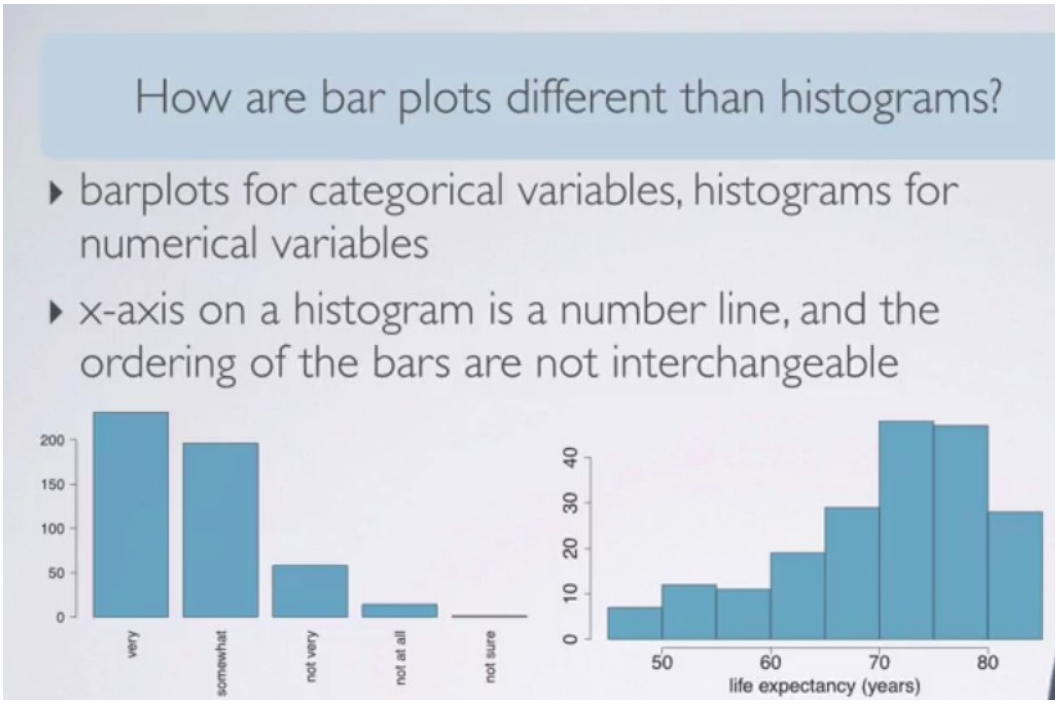
- symmetric, mean is a better measure of typical salary
- symmetric, median is a better measure of typical salary
- left skewed, mean is a better measure of typical salary
- left skewed, median is a better measure of typical salary

TRANSFORMING DATA:

<u>Goals of transforming the data:</u>

We might want to see the data structure a little differently.
We might want to reduce skew to assist in modeling
We might want to straighten a nonlinear relationship in a scatterplot, so that we can model the relationship with simpler methods.

CATEGORICAL VARIABLES:



This video is must for exploring categorical variables.

For single categorical variable we use frequency distribution table while for more than 2 categorical variable we use contingency table.

We use segmented barplot for contingency table.

STATISTICAL INFERENCE:

We saw a difference of roughly 30% between the proportion of male and female files that are promoted. Based on this information, which of the below might be true? Select all that apply.

- [ ] If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

Un-selected is correct

- [x] Males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.

Correct
Both of these options are possible – the difference might be indicative of discrimination against women in promotion decisions, or it might just be due to chance. We need further analysis to make a decision between these two competing claims.

- [x] The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination.

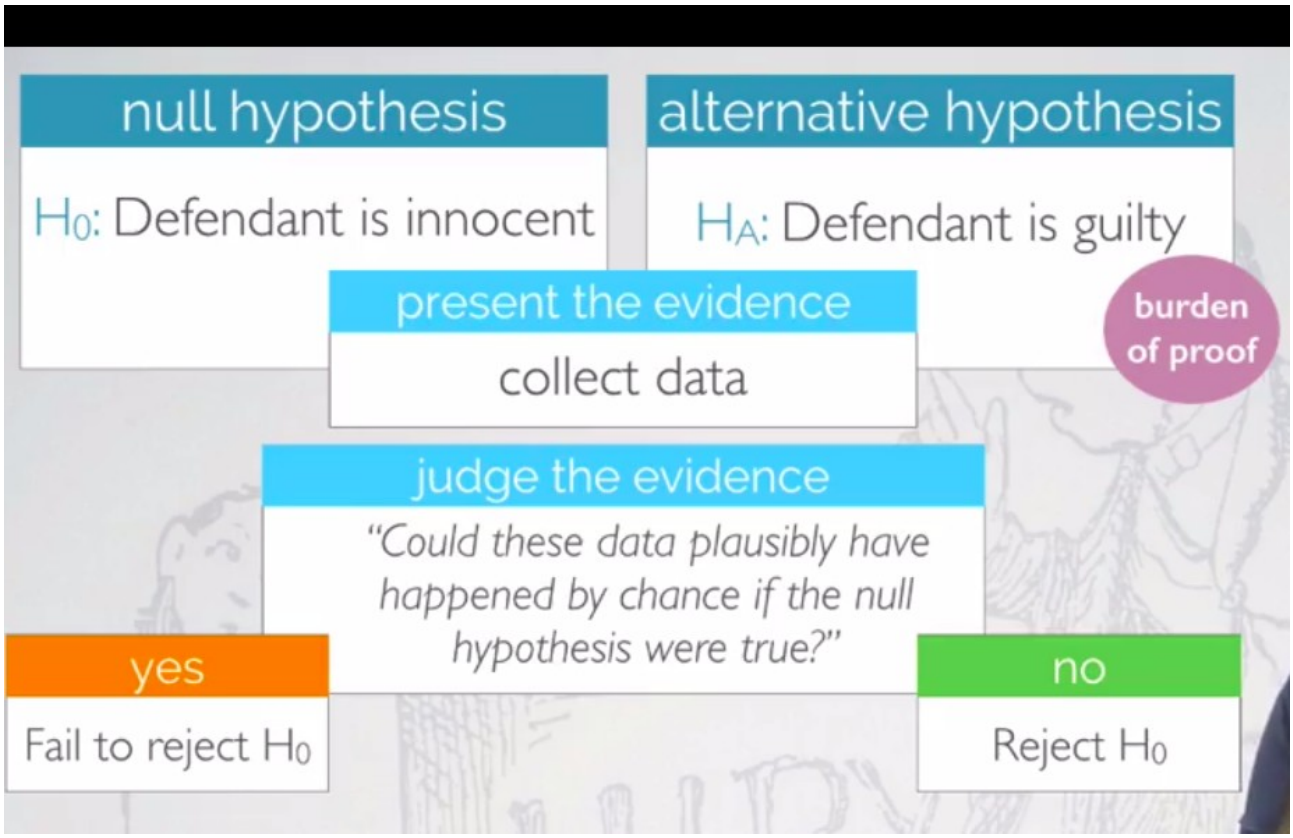
Correct
Both of these options are possible – the difference might be indicative of discrimination against women in promotion decisions, or it might just be due to chance. We need further analysis to make a decision between these two competing claims.

- [ ] Women are less qualified than men, and this is why fewer females get promoted.

Un-selected is correct

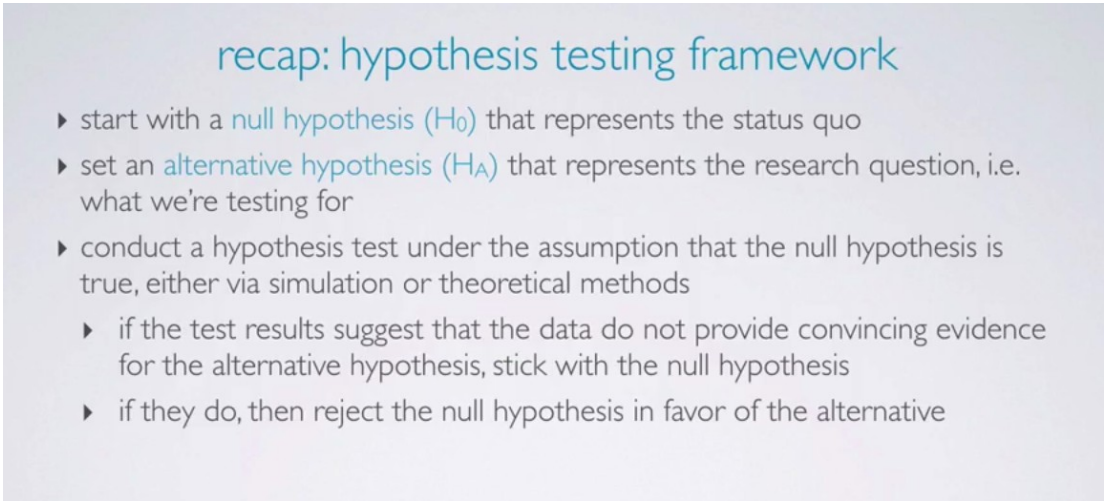
if the null hypothesis is true than the Z statistic has a std normal distribution

Z-scores can be calculated for distributions that are not normal.



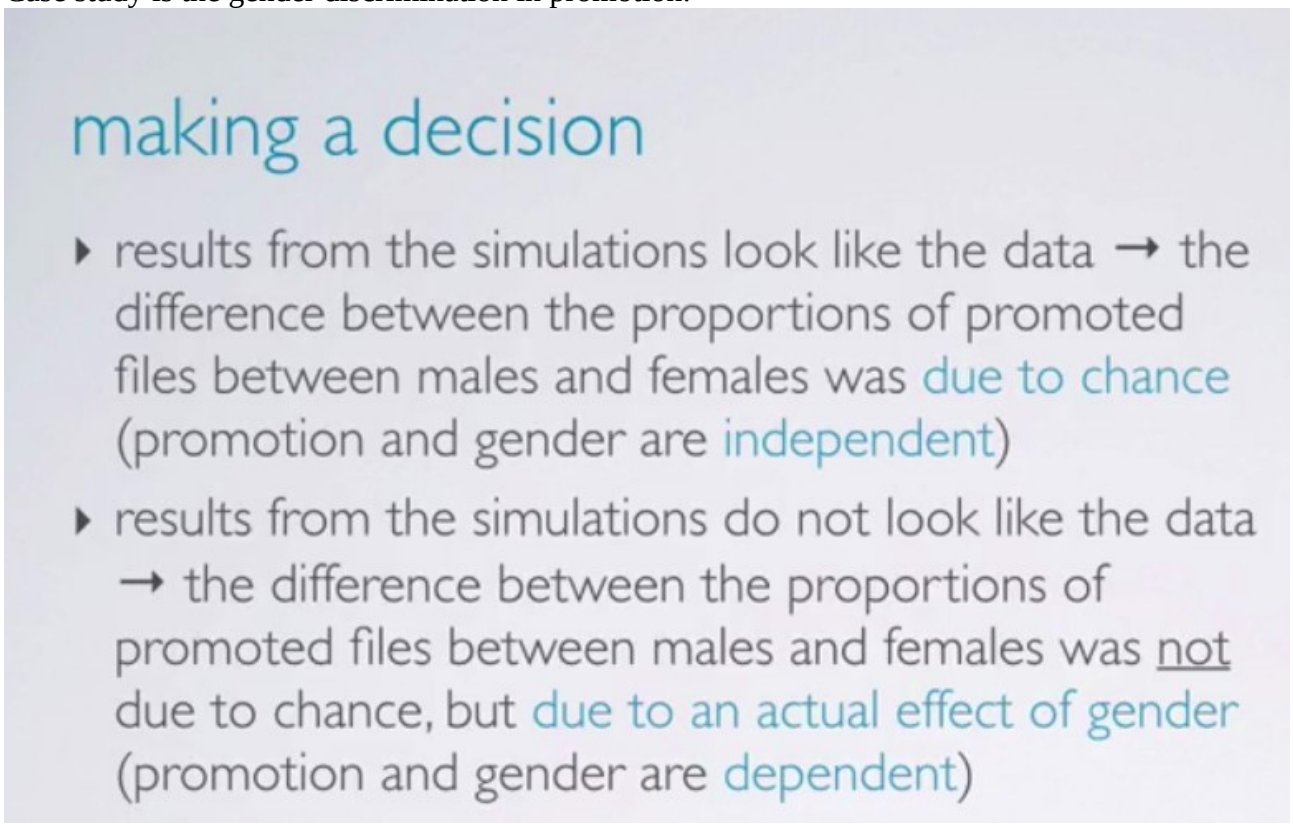
Hypothesis testing is very much like a court trial in the US. The null hypothesis says that the defendant is innocent and the alternative hypothesis says that the defendant is guilty. We then present evidence or, or in other words, collect data. Then, we judge this evidence and ask ourselves the question, could these data plausibly have happened by chance if the null hypothesis were true?

If the data were likely to have occurred under the assumption that the null hypothesis were true, then we would fail to reject the null hypothesis, and state that the evidence is not sufficient to suggest that the defendant is guilty. Note that when this happens, the jury returns with a verdict of not guilty. The jury does not say the defendant is innocent, just that there is not enough evidence to convict. The defendant may in fact be innocent but the jury has no way of being sure. Said statistically, we fail to reject the null hypothesis. We never declare the null hypothesis to be true. Because we do not know and cannot prove whether it's true or not. Therefore, we also never say that we would accept the null hypothesis. If the data were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis, and hence we reject the null hypothesis



how to make decision?

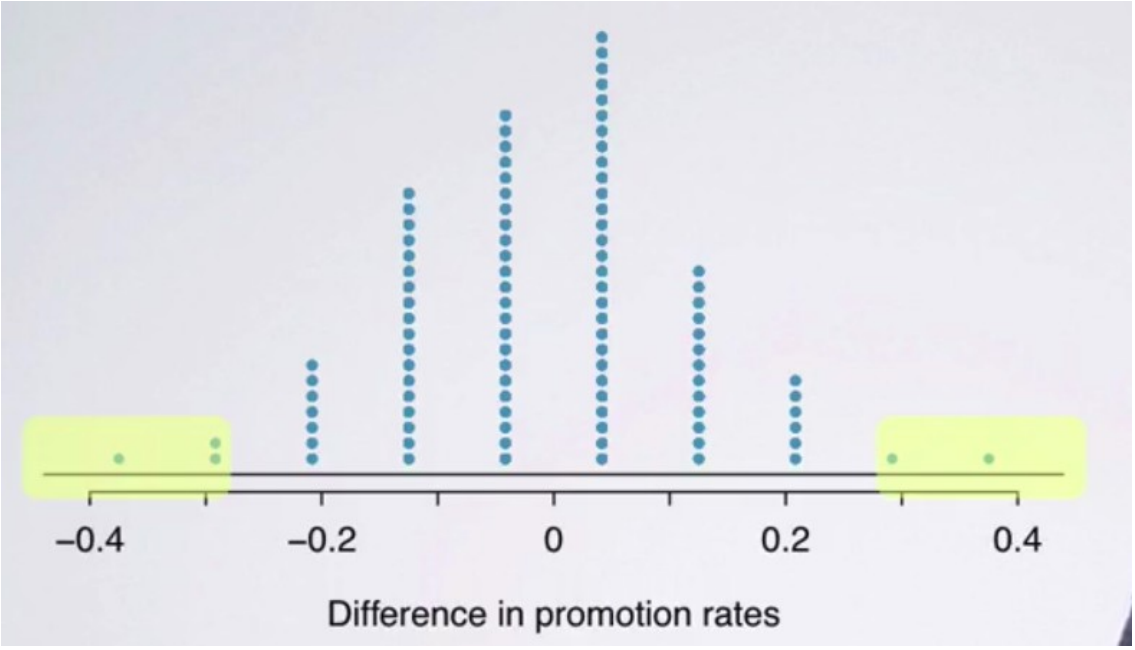
Case study is the gender discrimination in promotion.



We simulated the experiment and obtained 100 samples I.e randomly assigning the files to the officers for promotion of males and females and we observed the diff between the proportions of them being promoted or not we then took the difference between their proportions and plot it using dot plot.

The plot shows that the promotion is independent of gender because the difference is highly centered around the zero. And there's only 5% of the time when the difference between them is 30%.

therefore we reject our null hypothesis that the promotion is dependent on gender discrimination.



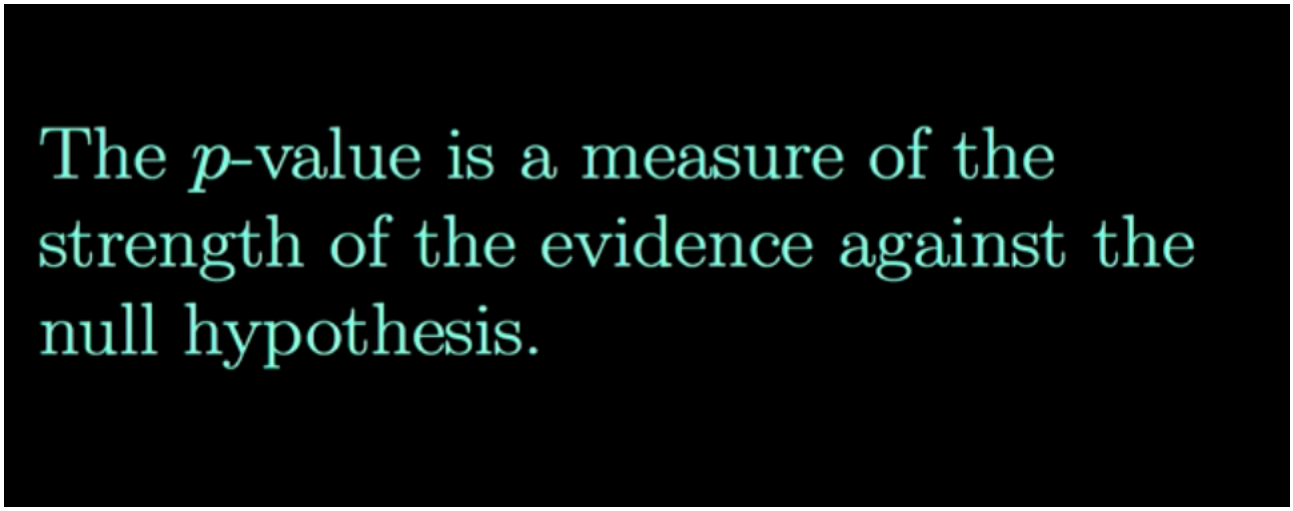
summary

- set a null and an alternative hypothesis
- simulate the experiment assuming that the null hypothesis is true
- evaluated the probability of observing an outcome at least as extreme as the one observed in the original data

p-value

- and if this probability is low, reject the null hypothesis in favor of the alternative

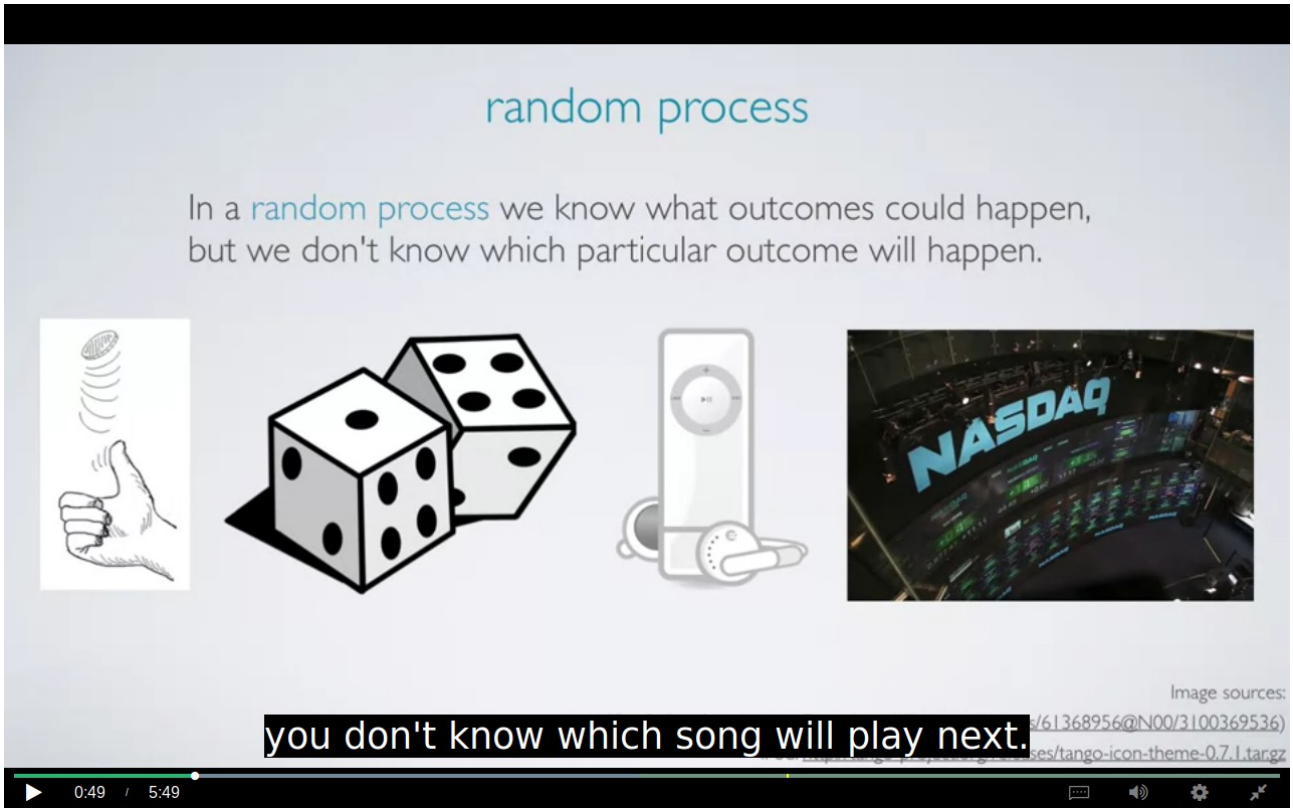
P-VALUE



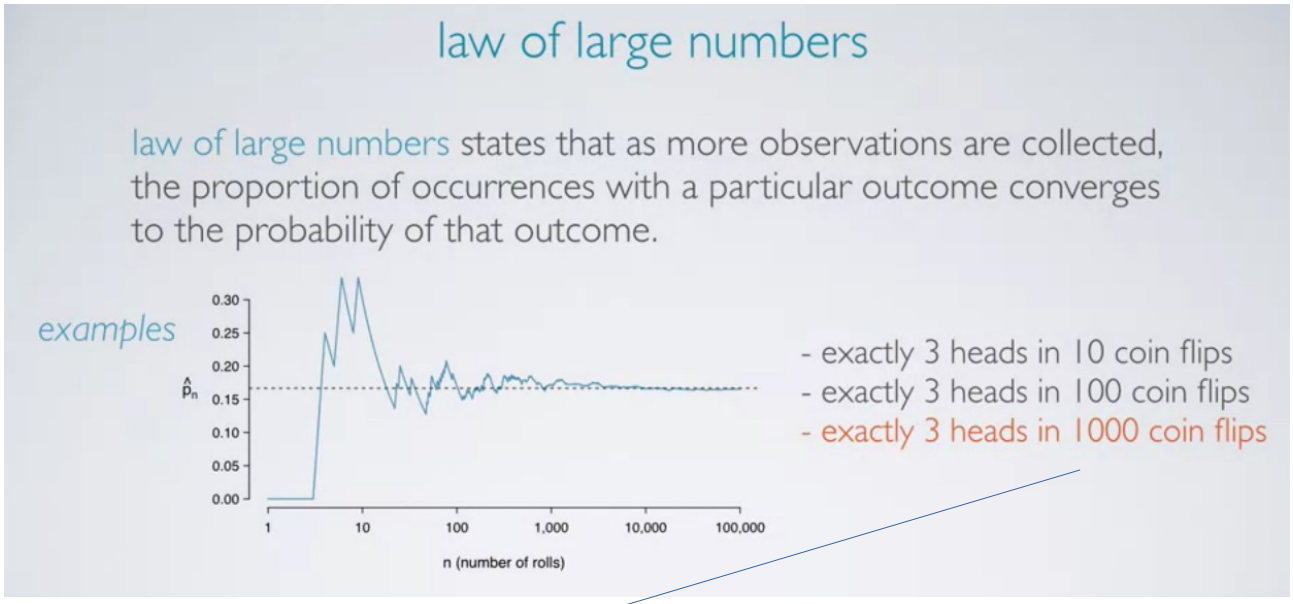
if the null hypothesis is true than the Z statistic has a std normal distribution

if p value is < alpha then we will reject the null hypothesis.

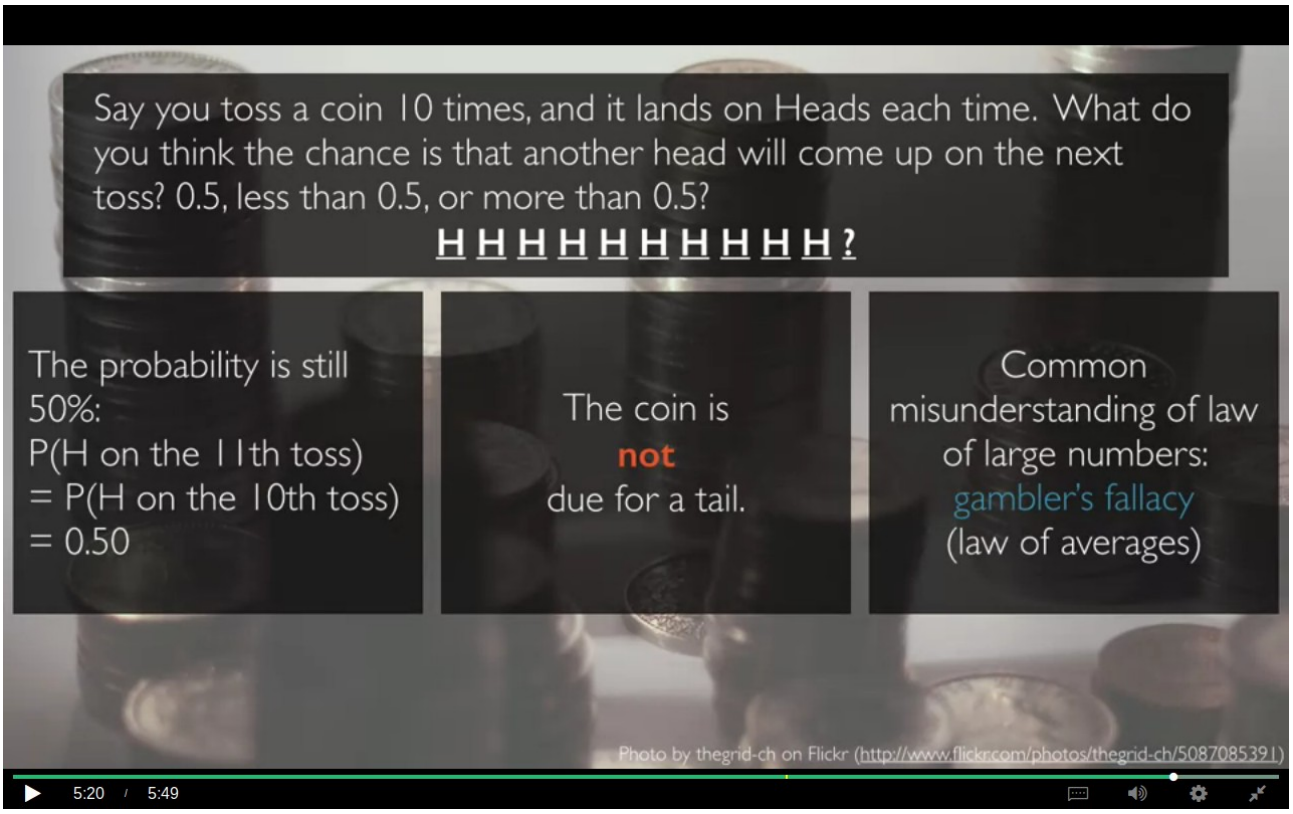
Week 3

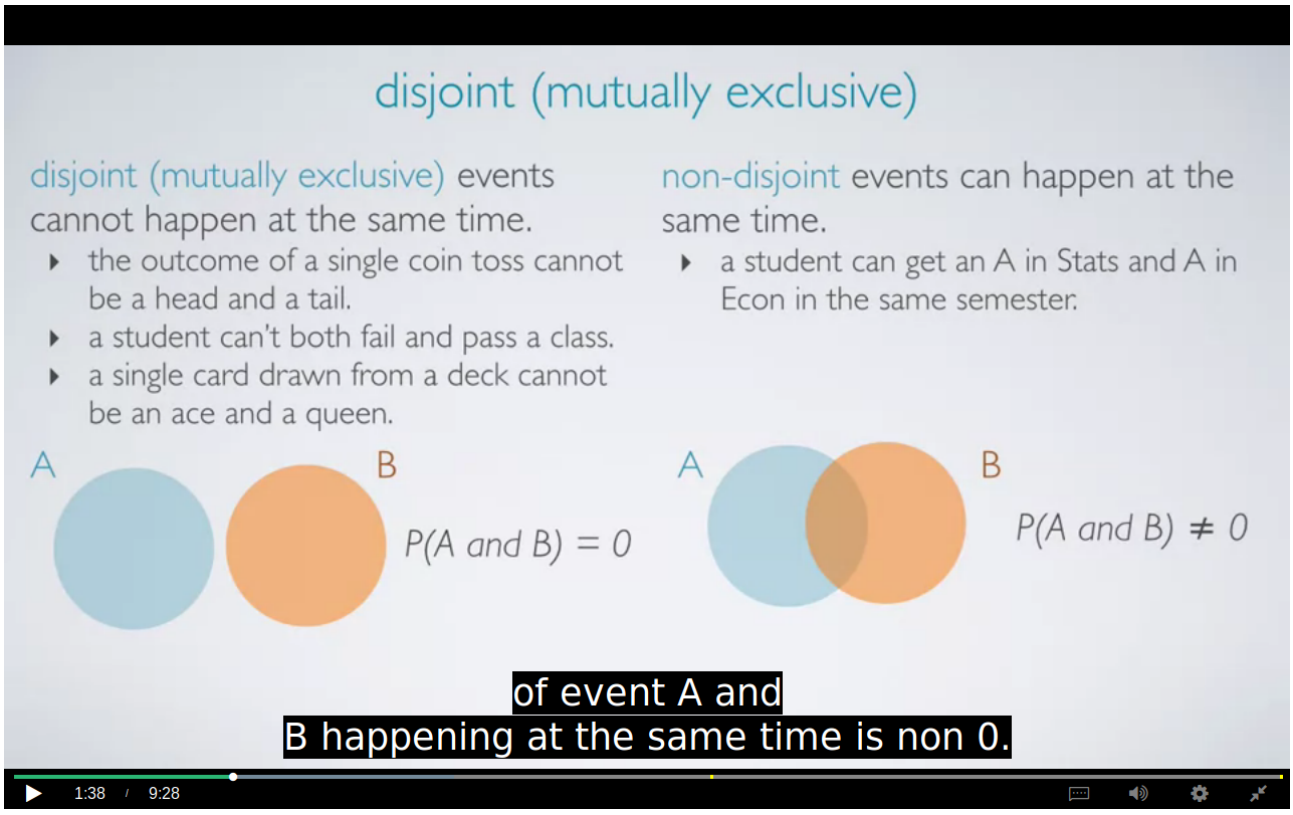


in bayesian interpretation 2 people can assign diff probability to the same event.



This is more unusual


Say you toss a coin 10 times, and it lands on Heads each time. What do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?
H H H H H H H H H H ?
The probability is still 50%:
P(H on the 11th toss)
= P(H on the 10th toss)
= 0.50
The coin is
not
due for a tail.
Common misunderstanding of law of large numbers:
gambler's fallacy
(law of averages)
Photo by thegrid-ch on Flickr (http://www.flickr.com/photos/thegrid-ch/5087085391)
5:20 / 5:49


disjoint (mutually exclusive)
disjoint (mutually exclusive) events cannot happen at the same time.
the outcome of a single coin toss cannot be a head and a tail.
a student can't both fail and pass a class.
a single card drawn from a deck cannot be an ace and a queen.
non-disjoint events can happen at the same time.
a student can get an A in Stats and A in Econ in the same semester.
A
B
P(A and B) = 0
A
B
P(A and B) ≠ 0
of event A and
B happening at the same time is non 0.
1:38 / 9:28

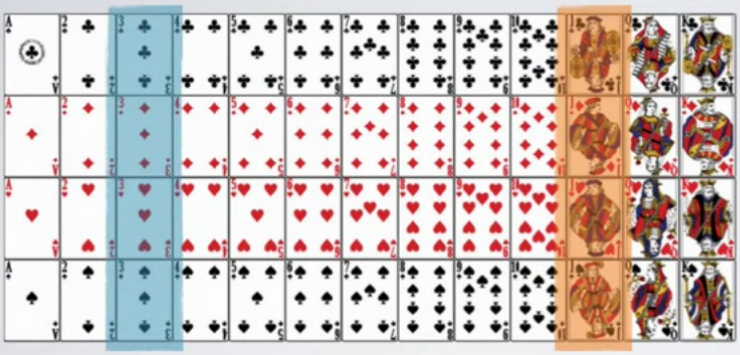
Can an event be disjoint and independent?

3 Answers. **Two disjoint events can** never be **independent**, except in the case that one of the **events** is null. ... **Events** are considered **disjoint** if they never occur at the same time. For example, being a freshman and being a sophomore would be considered **disjoint events**. Jun 20, 2016

Probability: Are disjoint events independent? - Mathematics Stack ...
https://math.stackexchange.com/questions/.../probability-are-disjoint-events-independent

union of disjoint events

What is the probability of drawing a Jack or a three from a well shuffled full deck of cards?



$P(J \text{ or } 3)$
$= P(J) + P(3)$
$= (4/52) + (4/52)$
≈ 0.154

For disjoint events A and B,
P(A or B) = P(A) + P(B)

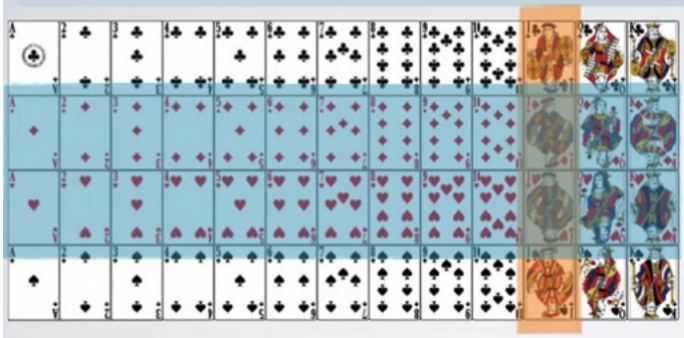
Image source: http://svg-cards.sourceforge.net/

2:48 / 9:28

union of non-disjoint events

What is the probability of drawing a Jack or a red card from a well shuffled full deck of cards?

$$P(J \text{ or red}) = P(J) + P(\text{red}) - P(J \text{ and red})$$
$$= (4/52) + (26/52) - (2/52)$$
$$\approx 0.538$$



For non-disjoint events A and B,
P(A or B) = P(A) + P(B) - P(A and B)

Image source: http://svg-cards.sourceforge.net/

4:26 / 9:28

What is the probability that a randomly sampled student thinks marijuana should be legalized or they agree with their parents' political views?

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes			
Total			

- (40+36-78) / 165
- (114+118-78) / 1

Correct
Use the general

- 78 / 165
- 78 / 188
- 11 / 47

probability distributions

a probability distribution lists all possible outcomes in the sample space, and the probabilities with which they occur.

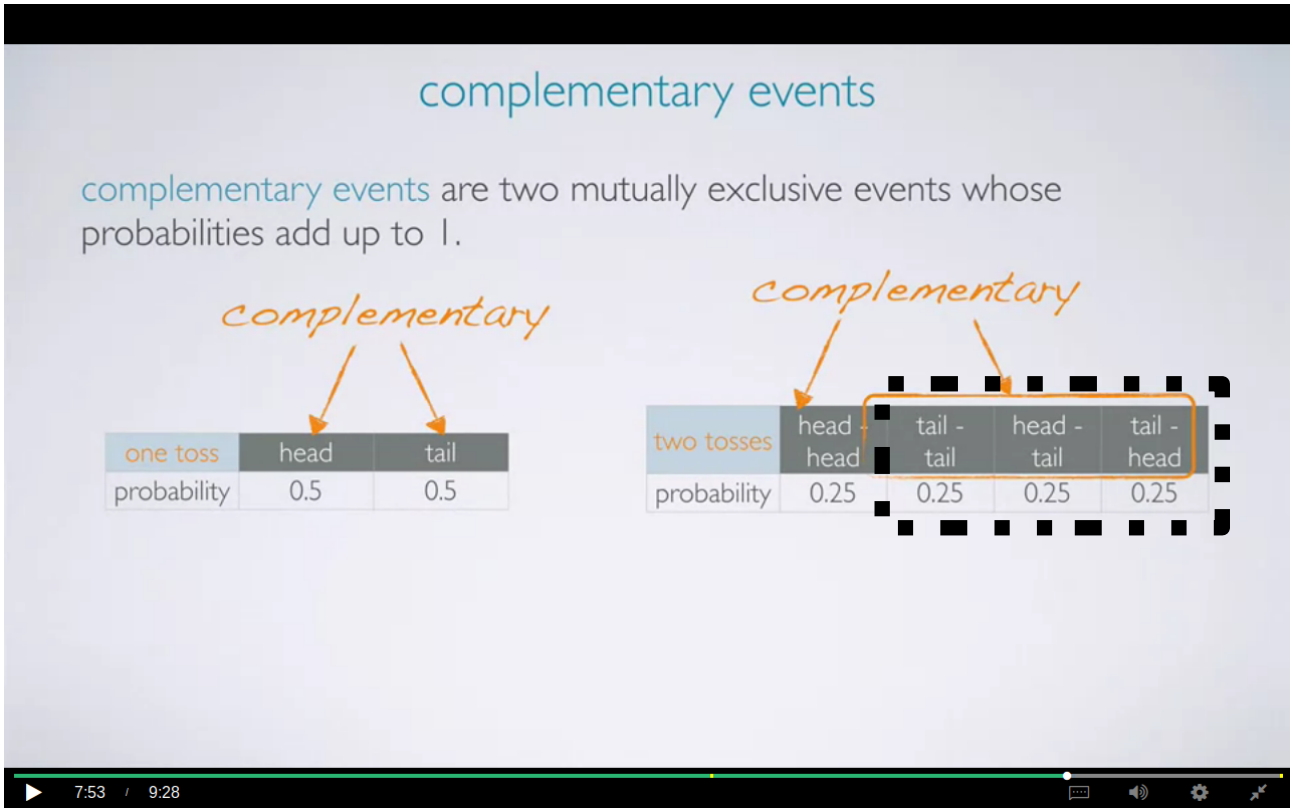
one toss	head	tail
probability	0.5	0.5

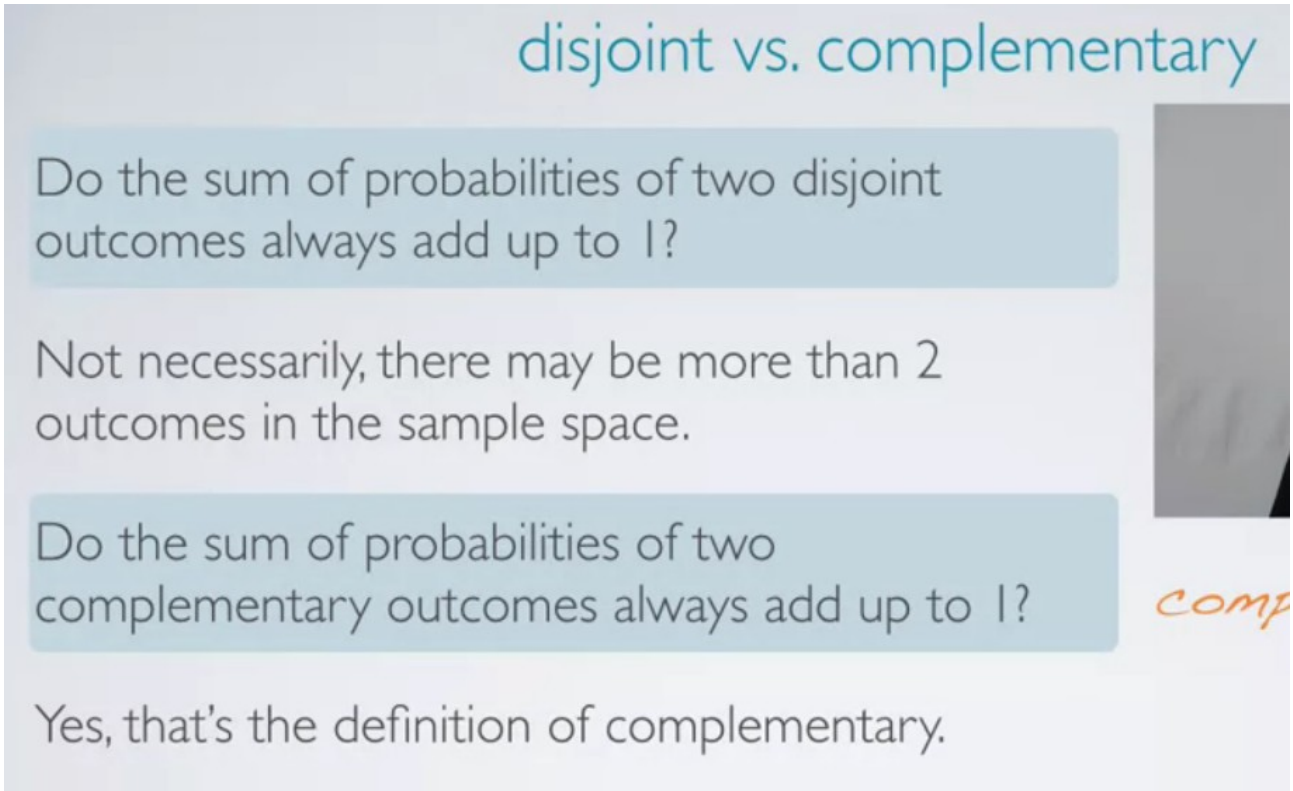
two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

rules

1. the events listed must be disjoint
2. each probability must be between 0 and 1
3. the probabilities must total 1

7:17 / 9:28


complementary events
complementary events are two mutually exclusive events whose probabilities add up to 1.
complementary
one toss | head | tail
probability | 0.5 | 0.5
complementary
two tosses | head - head | tail - tail | head - tail | tail - head
probability | 0.25 | 0.25 | 0.25 | 0.25
7:53 / 9:28


disjoint vs. complementary
Do the sum of probabilities of two disjoint outcomes always add up to 1?
Not necessarily, there may be more than 2 outcomes in the sample space.
Do the sum of probabilities of two complementary outcomes always add up to 1?
Yes, that's the definition of complementary.

Complementary are always disjoints but disjoints aren’t always complementary.

2.1.5 Complement of an event

S
Sample space

Rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

A^c
Complement of outcome A

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 2.9 shows the relationship between D, D^c, and the sample space S.

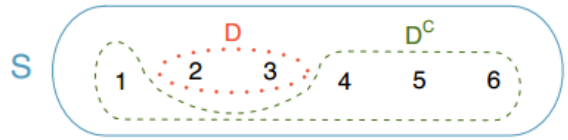
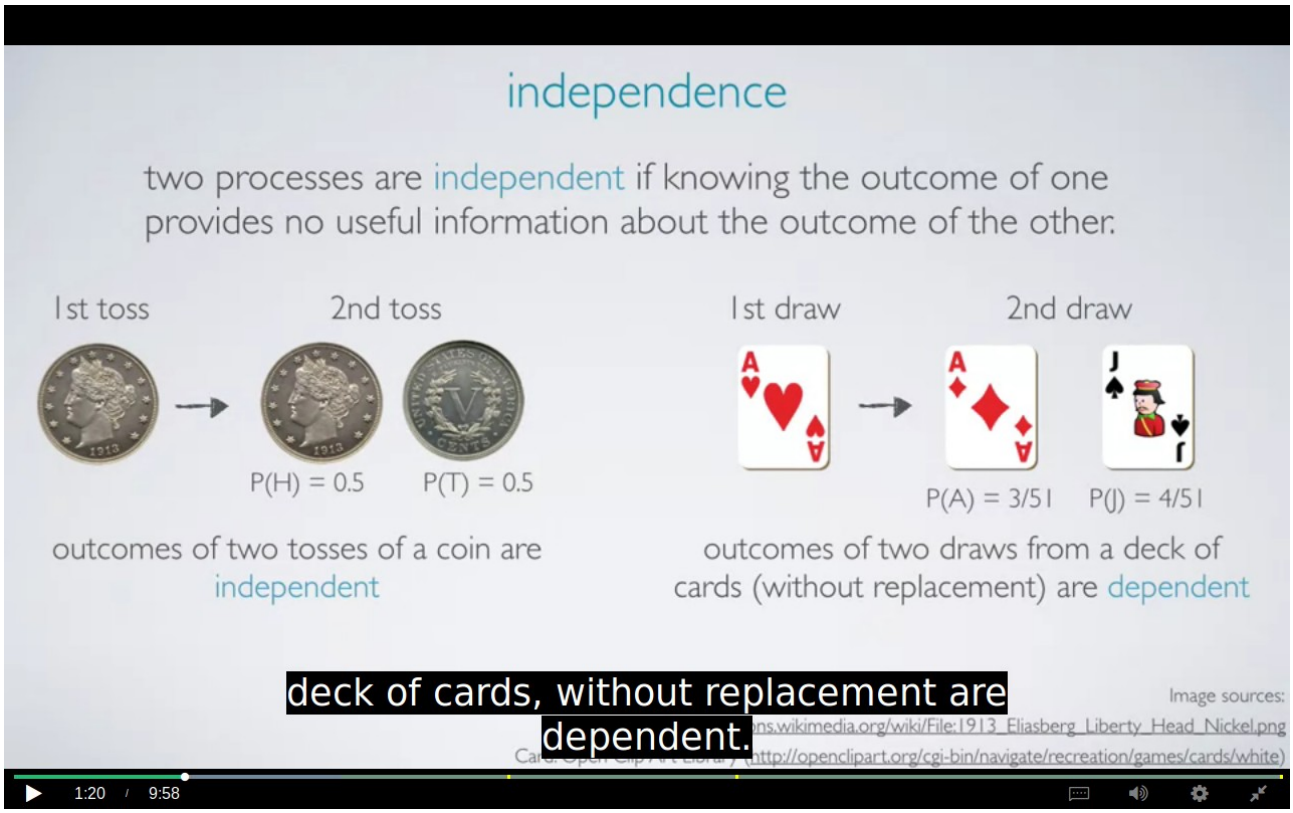
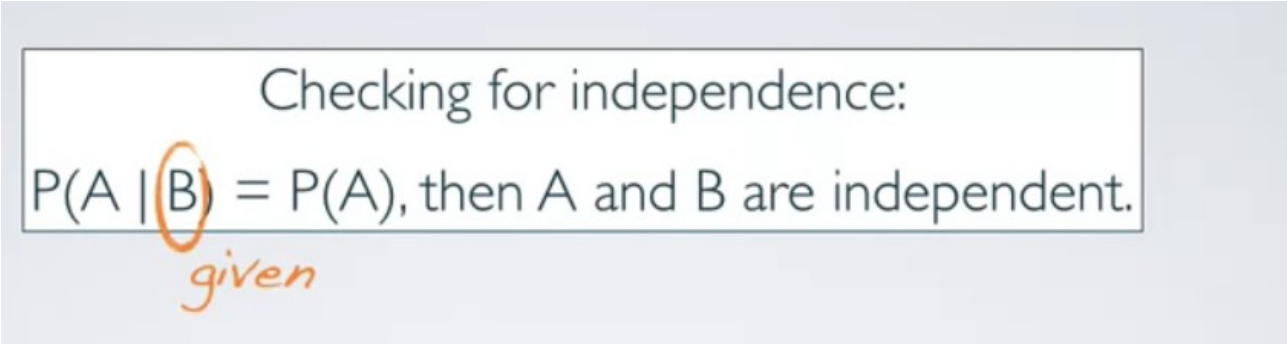


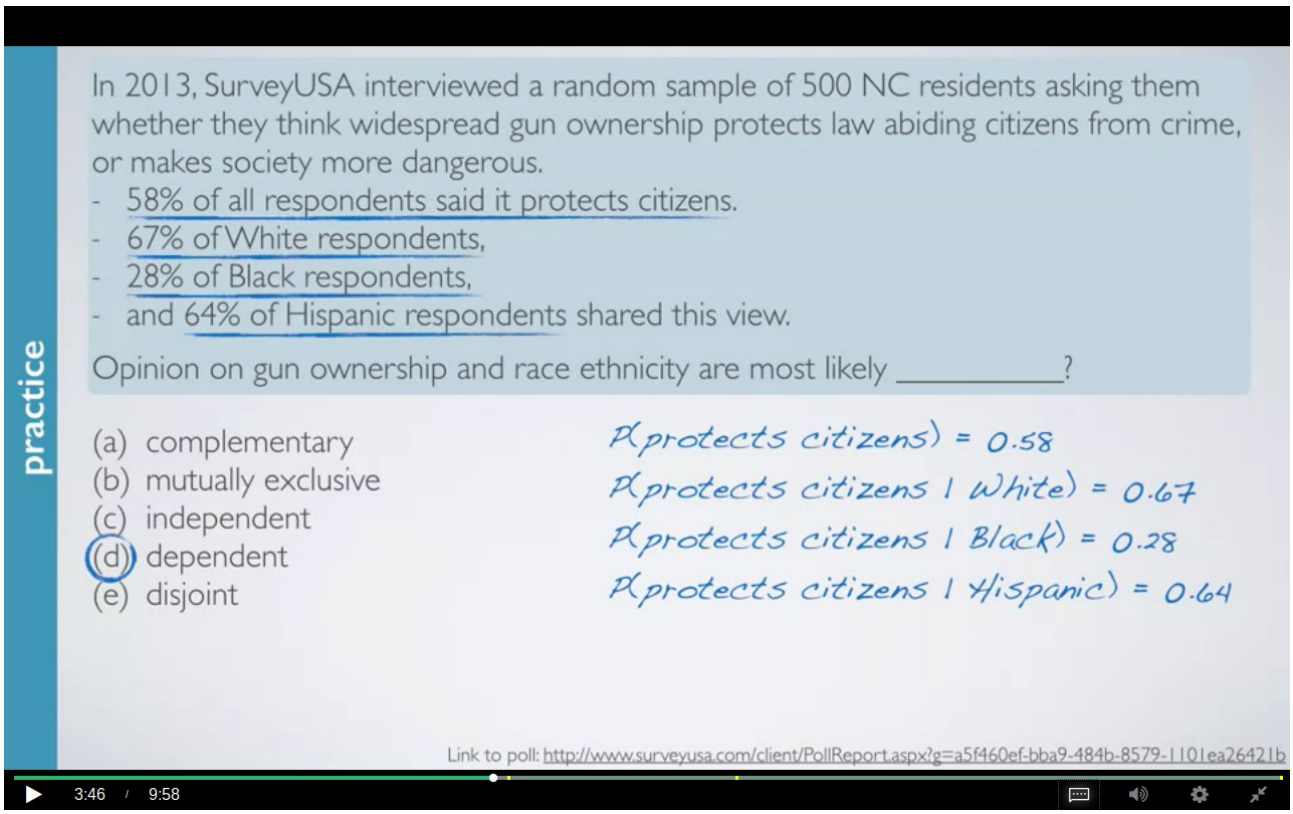
Figure 2.9: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible outcomes.



If the probability of an event A occurring, given that event B occurred is the same as the probability of event A occurring in the first place, then events A and B are said to be independent. *(as in the case of coin tossing)*

This rule basically says that knowing B tells us nothing about A





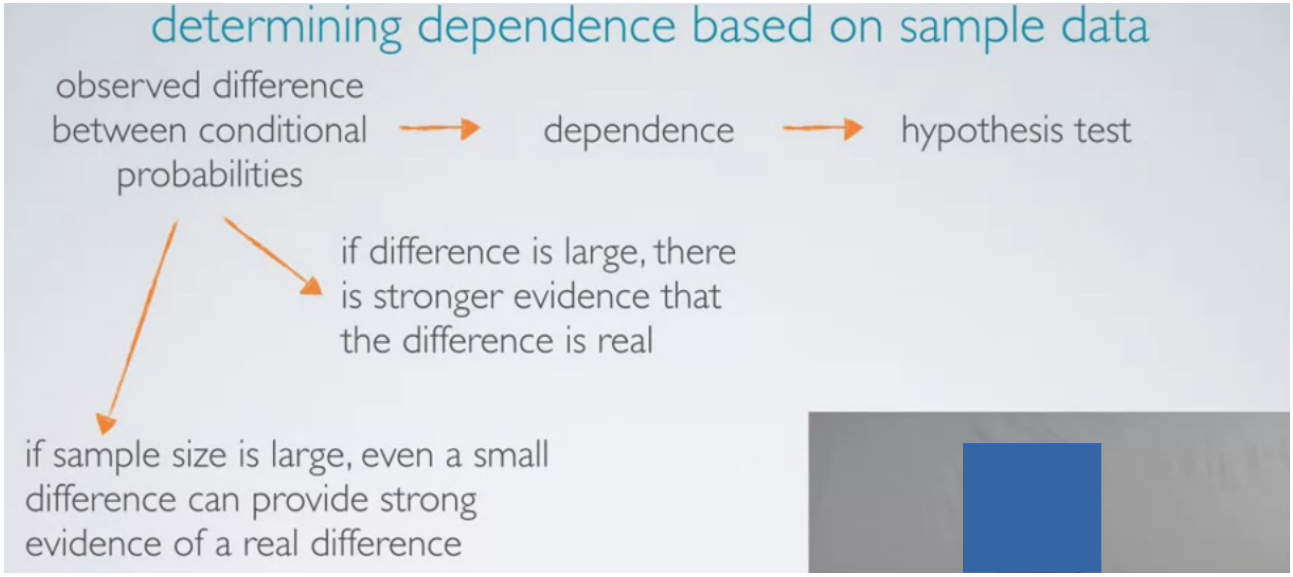
P(citizen who thinks that gun protects given that he is white) is 0.67.

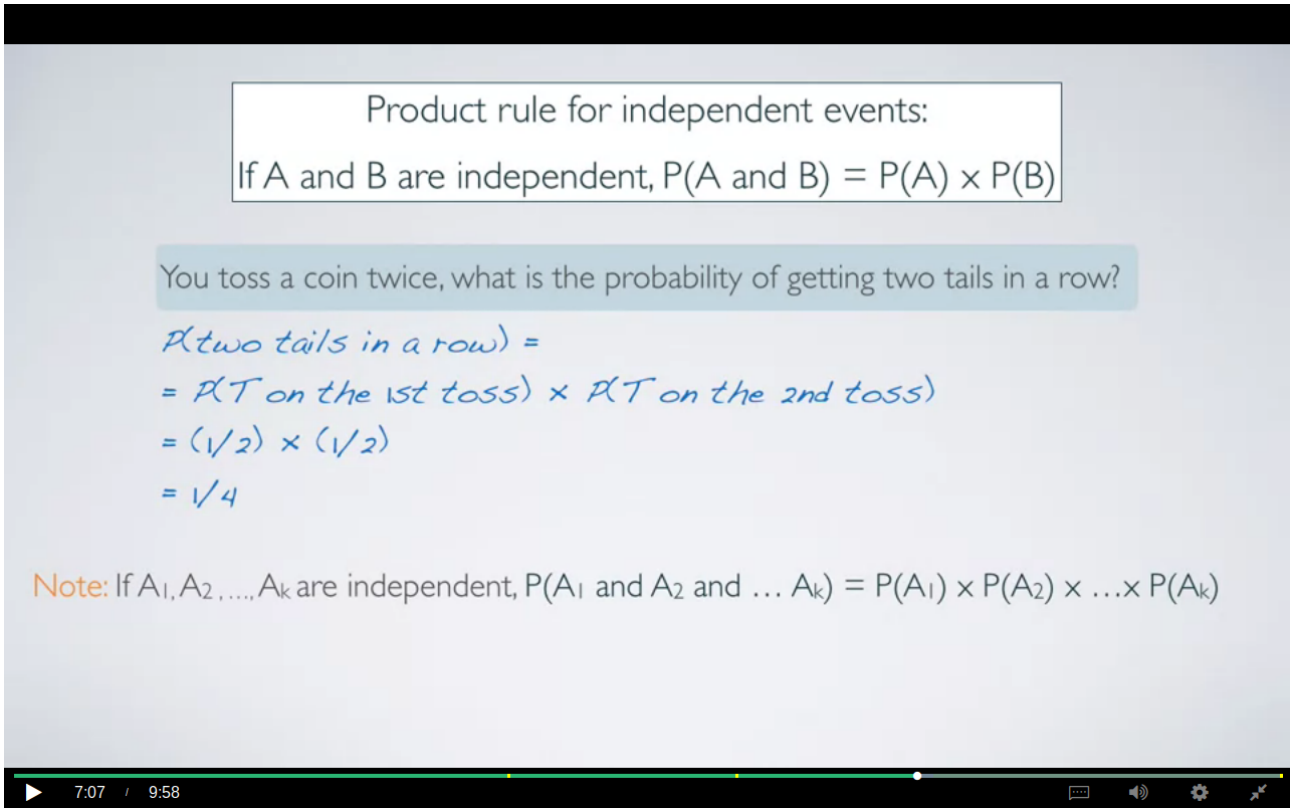
Earlier we saw that P(protects citizens | White) = 0.67 and P(protects citizens | Hispanic) = 0.64. Under which sample size would you be more convinced of a real difference between the proportions of Whites and Hispanics who think the widespread gun ownership protects citizens?

- n = 500
- n = 50,000

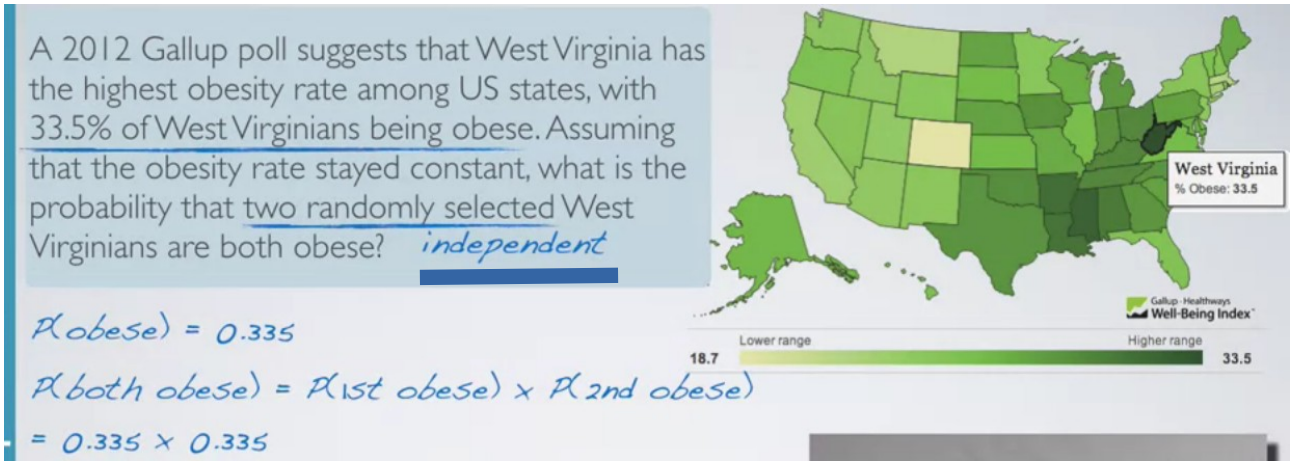
Correct

Larger the sample size, the more convincing the evidence from that sample.





if A1, A2 all the way through Ak are independent, then probability of all of these events happening at once is simply going to be the product of the individual probabilities of the events



Given the probabilities below and the accompanying Venn diagram, what percent of the world population have a university degree or higher and disagree with the statement about men having more right to a job than women?

P(agree) = 0.362

P(uni. degree) = 0.138

P(agree & uni. degree) = 0.036

(This is a long question, please scroll on the right to see the options.)

- 1 - 0.138 = 0.862 → 86.2%
- 0.102 → 10.2%

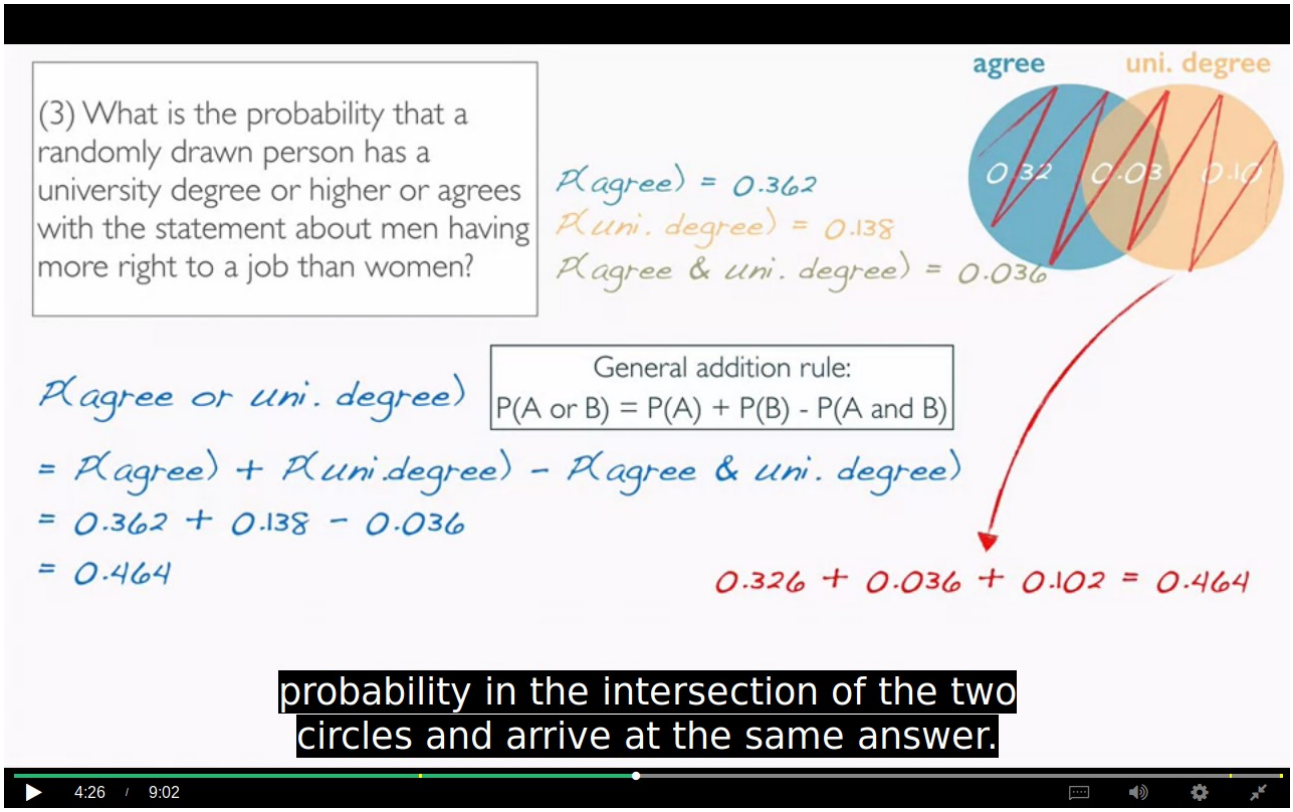
Correct

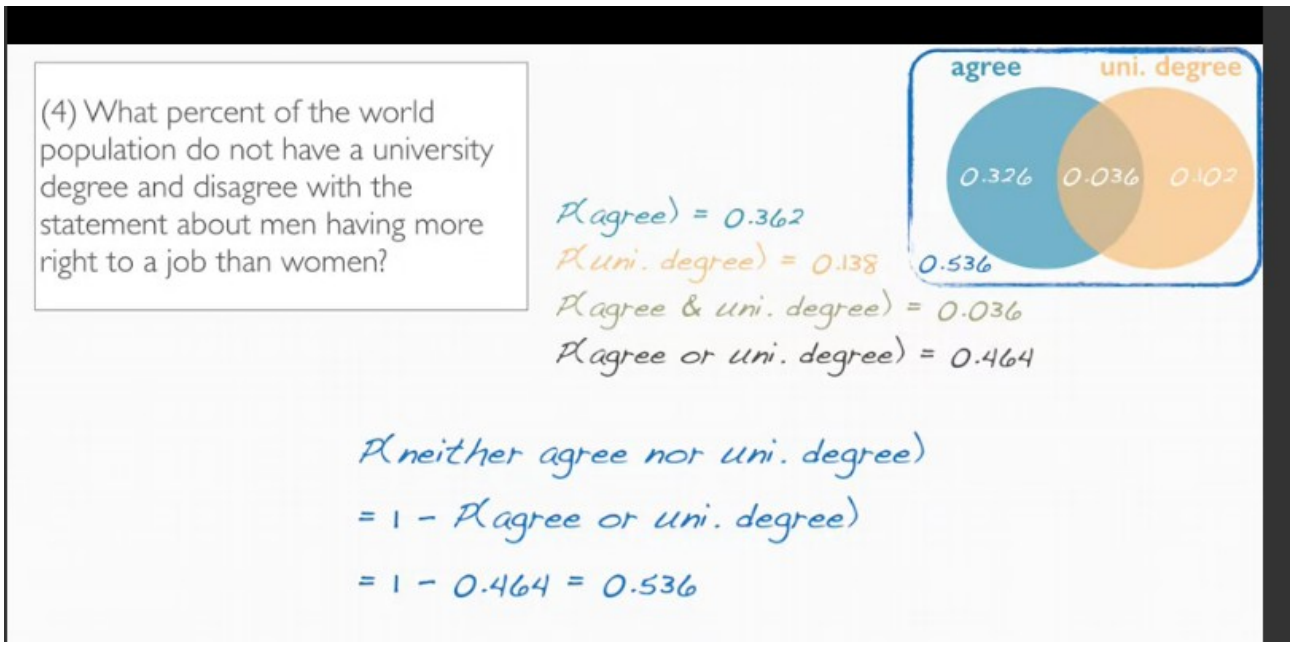
P(uni degree & disagree)

= P(uni degree) – P(uni degree & agree)

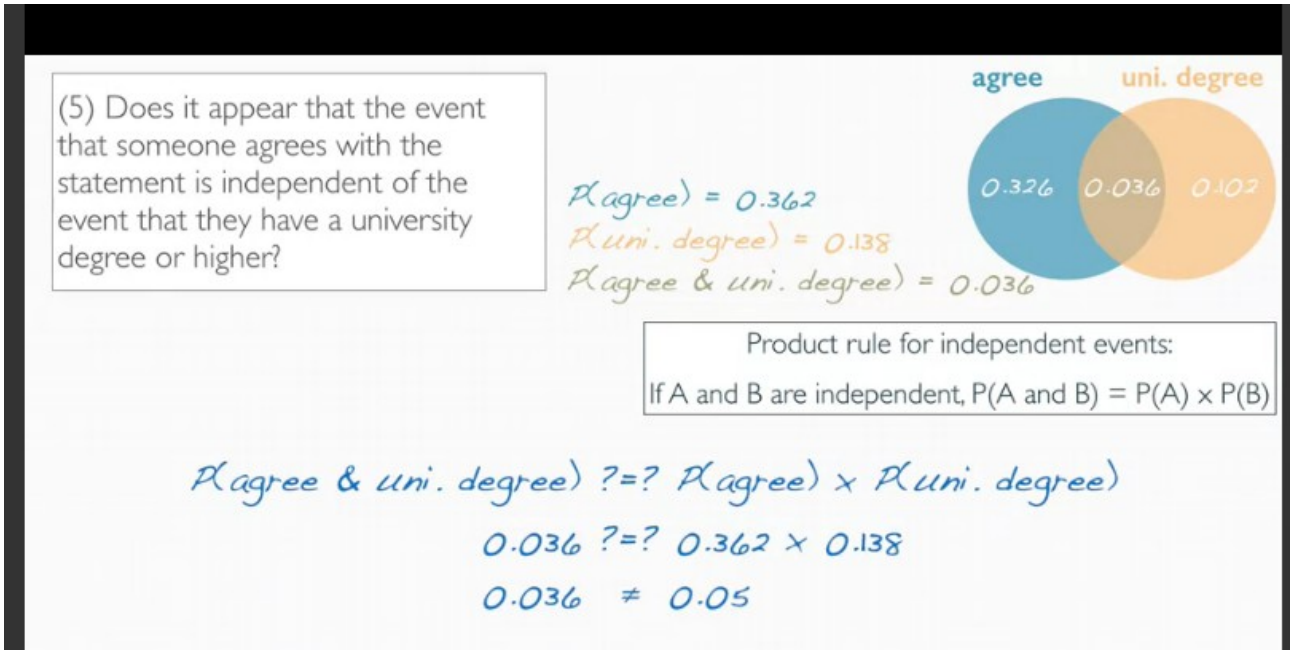
= 0.138 – 0.036 = 0.102

below are the 2 ways for calculating the probability by using addition rule and venn diagram.

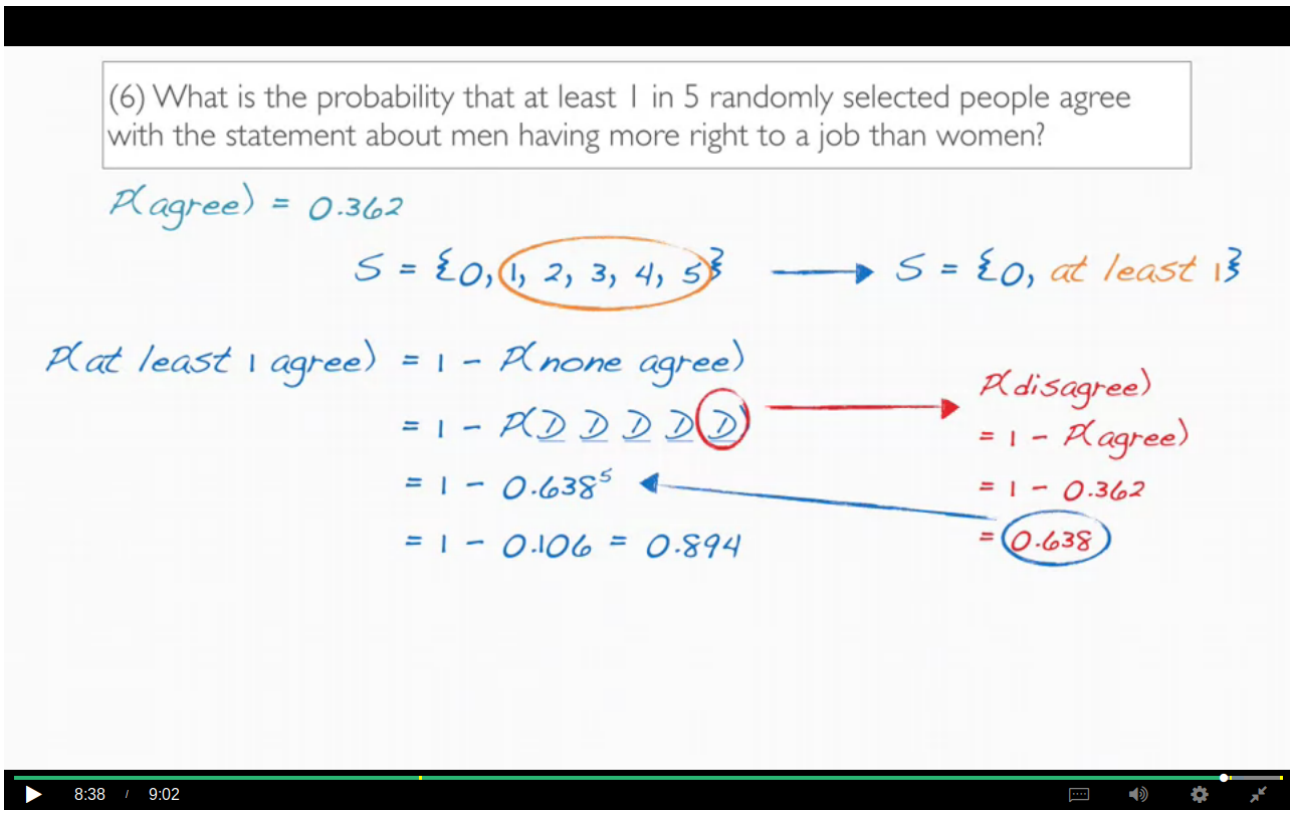




How to check independence ?



They aren't independent as the occurrence of event doesn't give us any hint about the occurrence of second event.



Which of the following is true? Assume a fair coin.

- Outcomes of one coin toss are disjoint and independent; outcomes of two coin tosses are dependent.
- Outcomes of one coin toss are disjoint and dependent; outcomes of two coin tosses are independent.

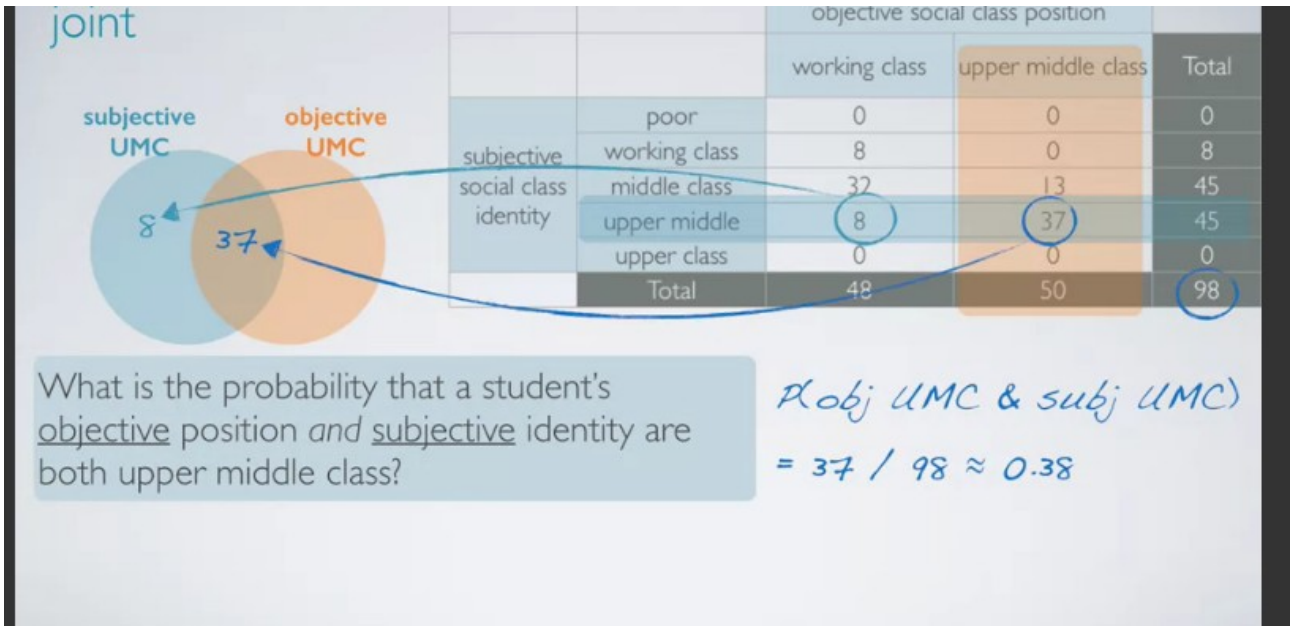
Correct
For a single coin toss, if we know that it's heads, then we know that it's not tails, hence dependent. For two consecutive coin tosses, knowing the result of the first toss does not tell us anything about the result of the second toss, hence independent.

- Outcomes of one coin toss are disjoint and dependent; outcomes of two coin tosses are also dependent.

We can generalize this to say that disjoint events with non-zero probability are always dependent on each other. Because if we know that one happened, we know that the other one cannot happen.

Conditional probability

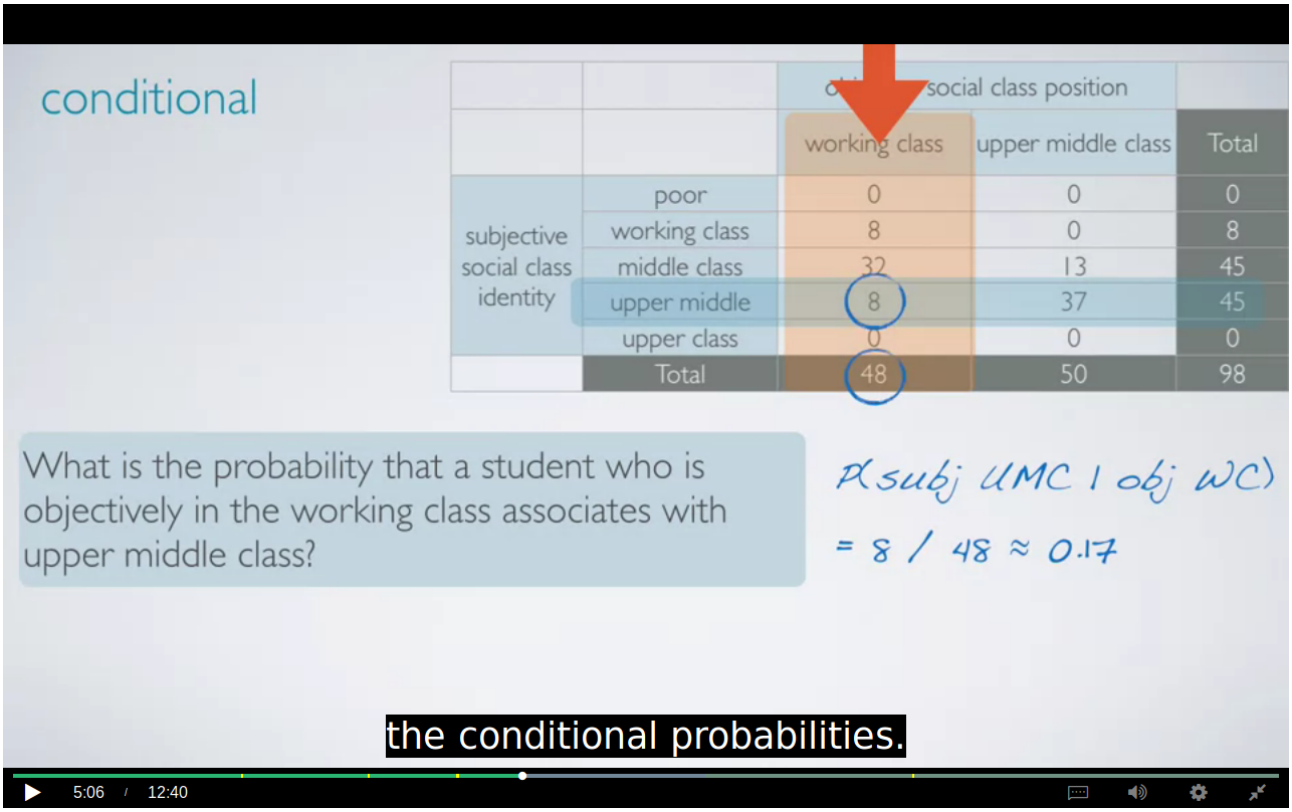
joint probability:



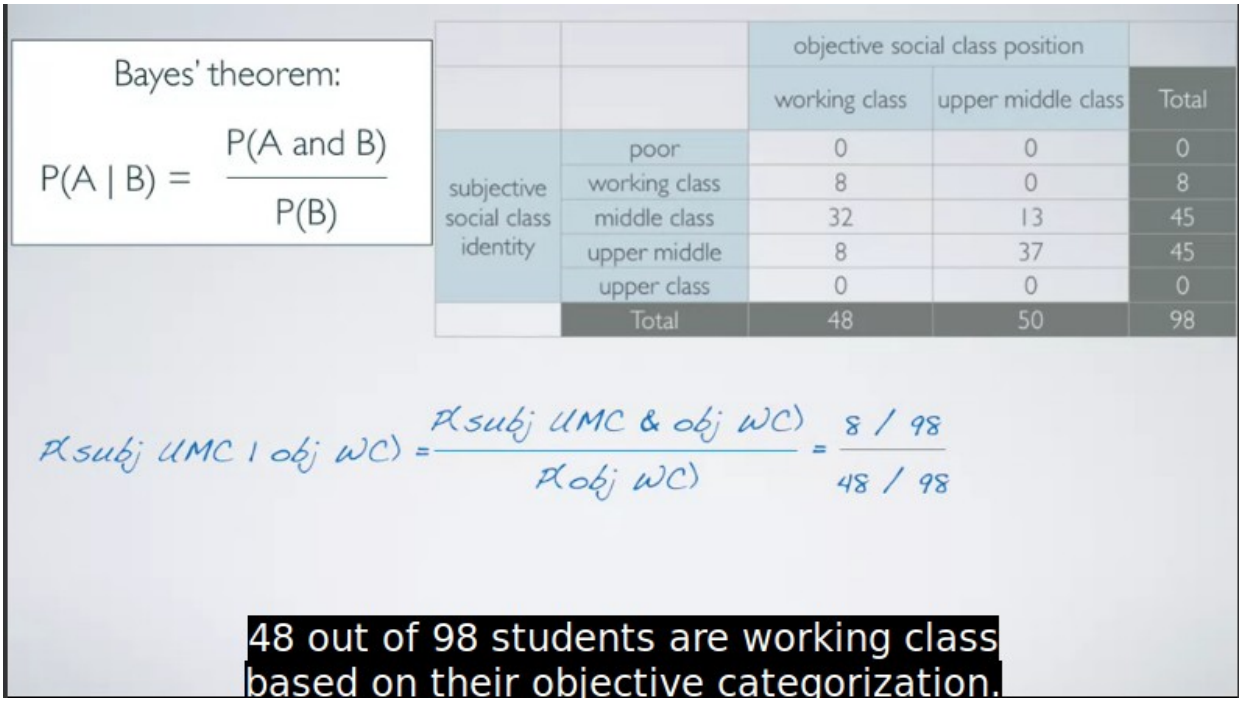
The term joint probability comes from the fact that we're considering the students who are at the intersection of the two events of interest.

CONDITIONAL :

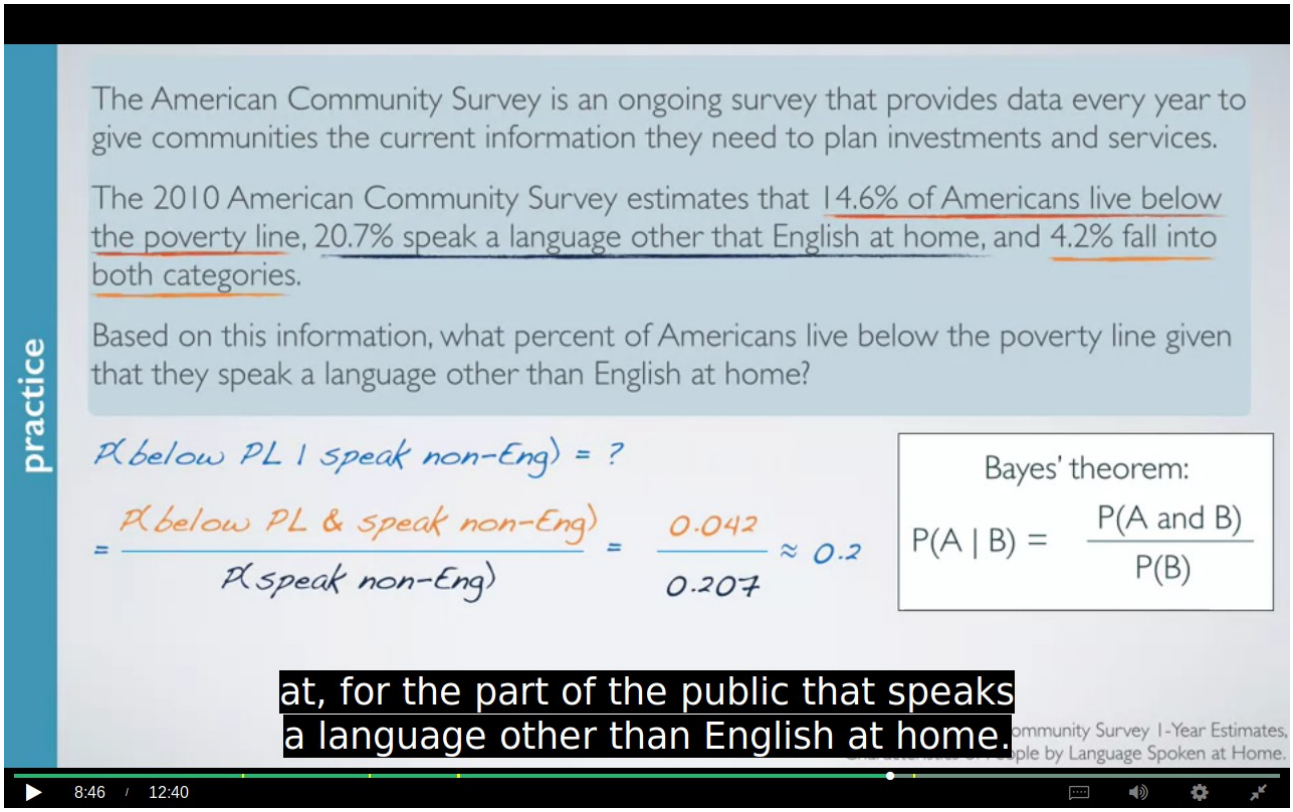
read it from the book too.



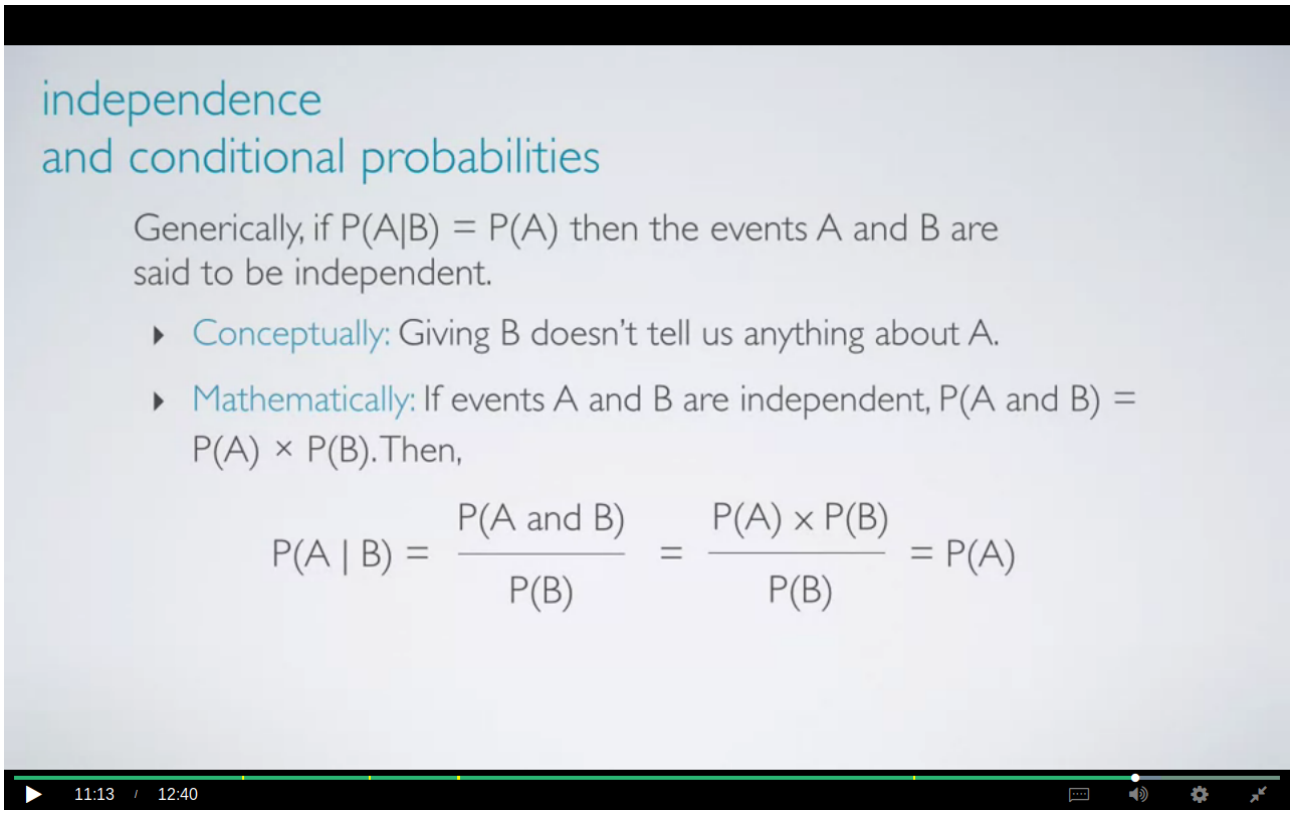
When the counts are not given we use bayes theorm to calculate the conditional probabilities.



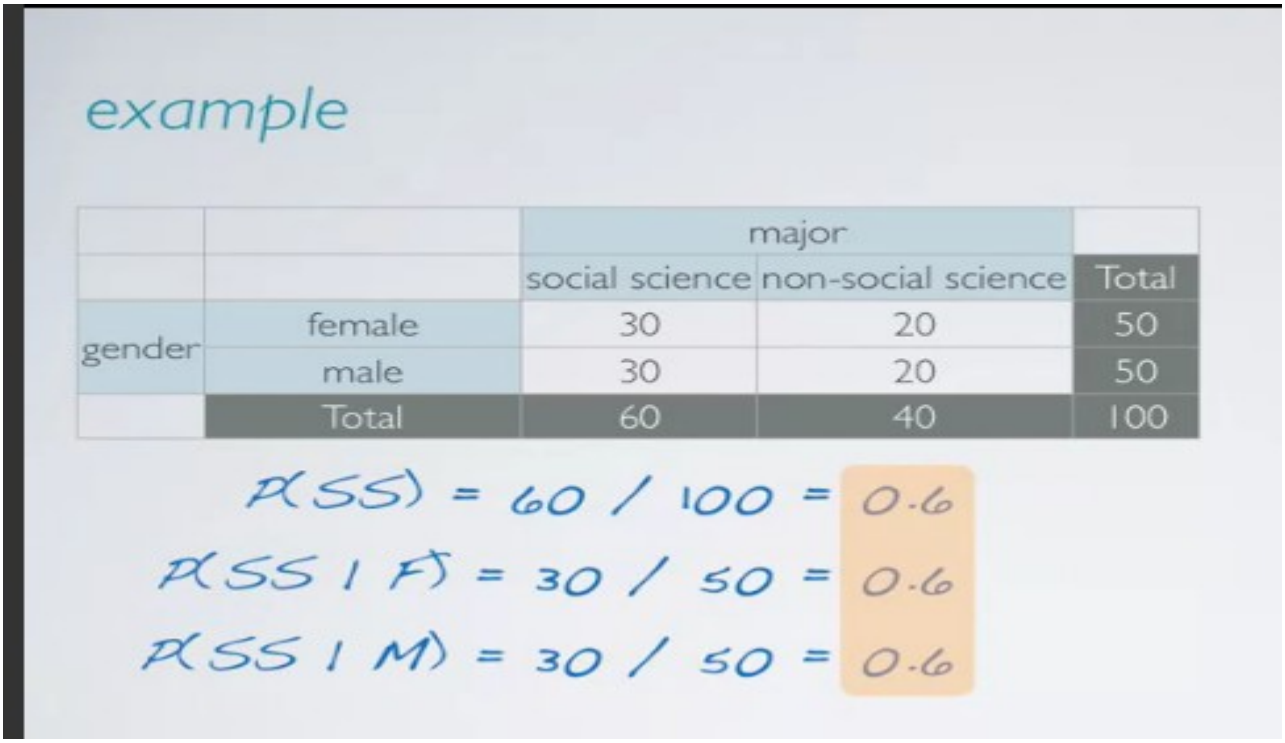
The numerator is the joint probability while the denominator is the marginal .



roughly 20% of Americans who speak a language other than English at home also live below the poverty line.



IMPORTANT EXAMPLE:



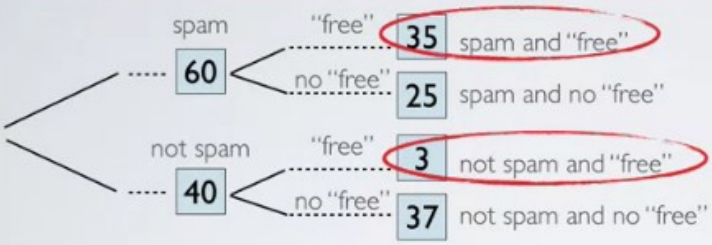
So what we're seeing here is that all of these probabilities are exactly the same. So this goes back to probability of a given b. If that equals probability of a, then we know that the events are independent. In this case, probability of social science equals probability of social science given female or social science given male.So we would determine that the two variables, gender and major are independent of each other, given this hypothetical distribution

PROBABILITY TREES

they are useful when we have to calculate the reverse of what we are given.

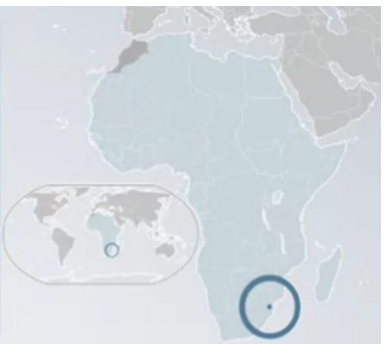
$$P(A \mid B) \rightarrow P(B \mid A)$$

You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word "free". Of the rest, 3 contain the word "free". If an email contains the word "free", what is the probability that it is spam?



$$P(\text{spam} \mid \text{"free"}) = \frac{35}{35 + 3} = 0.92$$

As of 2009, Swaziland had the highest HIV prevalence in the world. 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

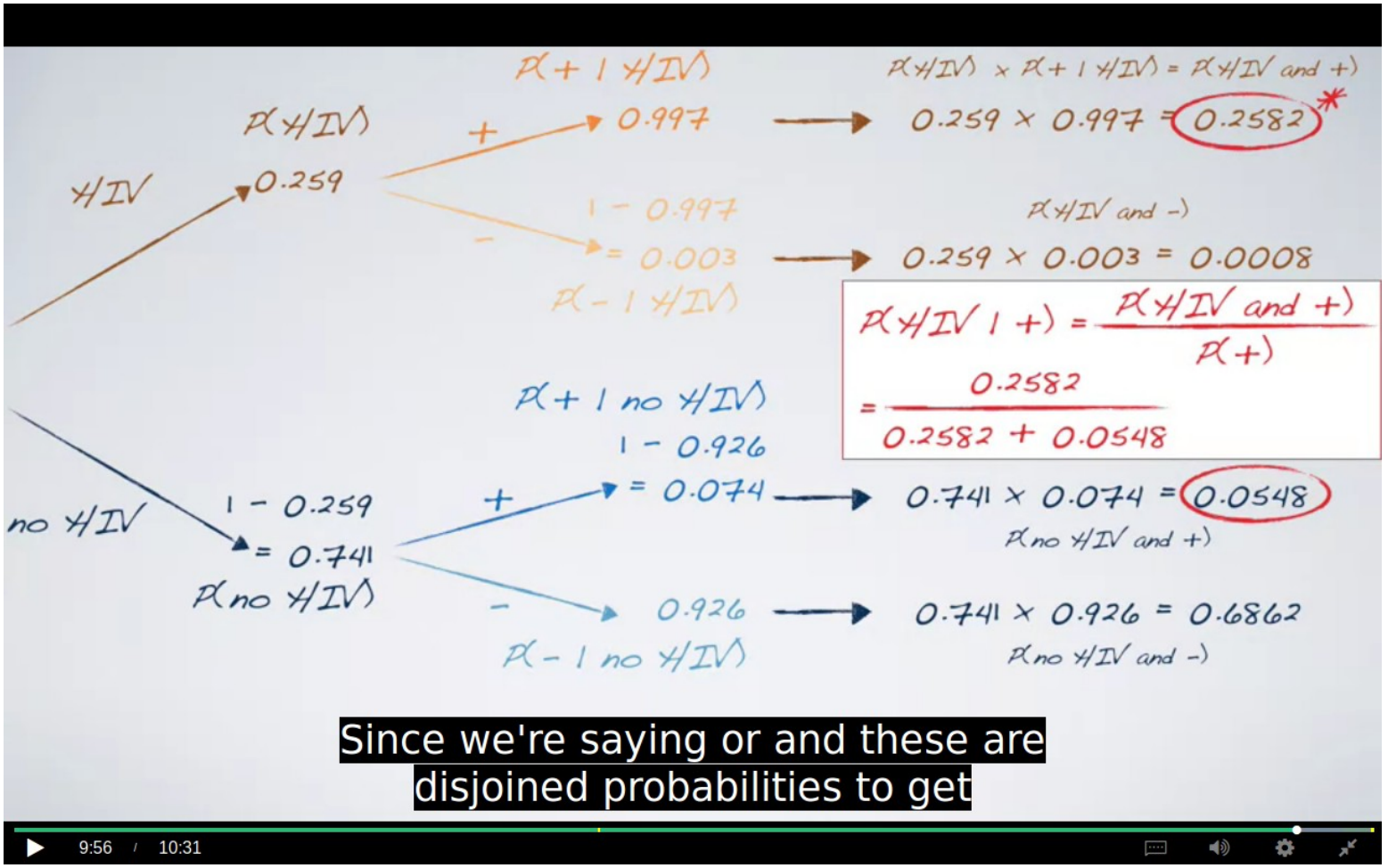


$P(HIV) = 0.259$

$P(+ \mid HIV) = 0.997$ $P(- \mid \text{no } HIV) = 0.926$

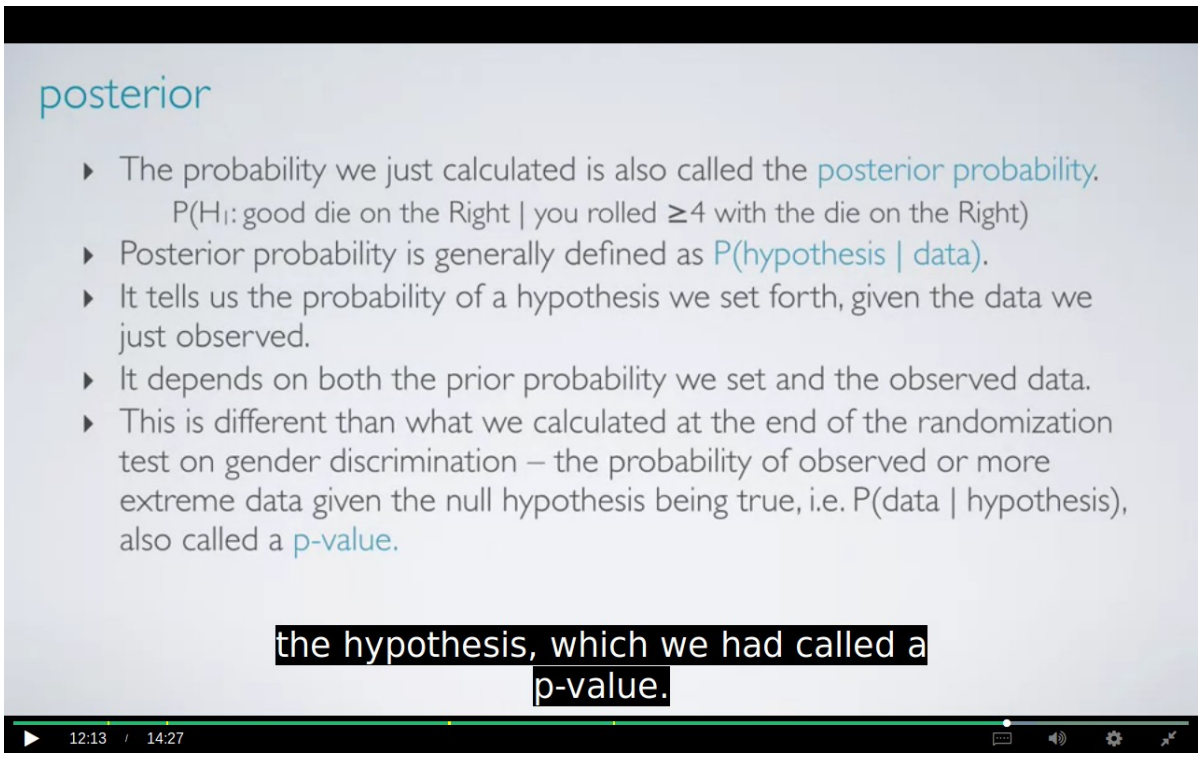
tree diagram!

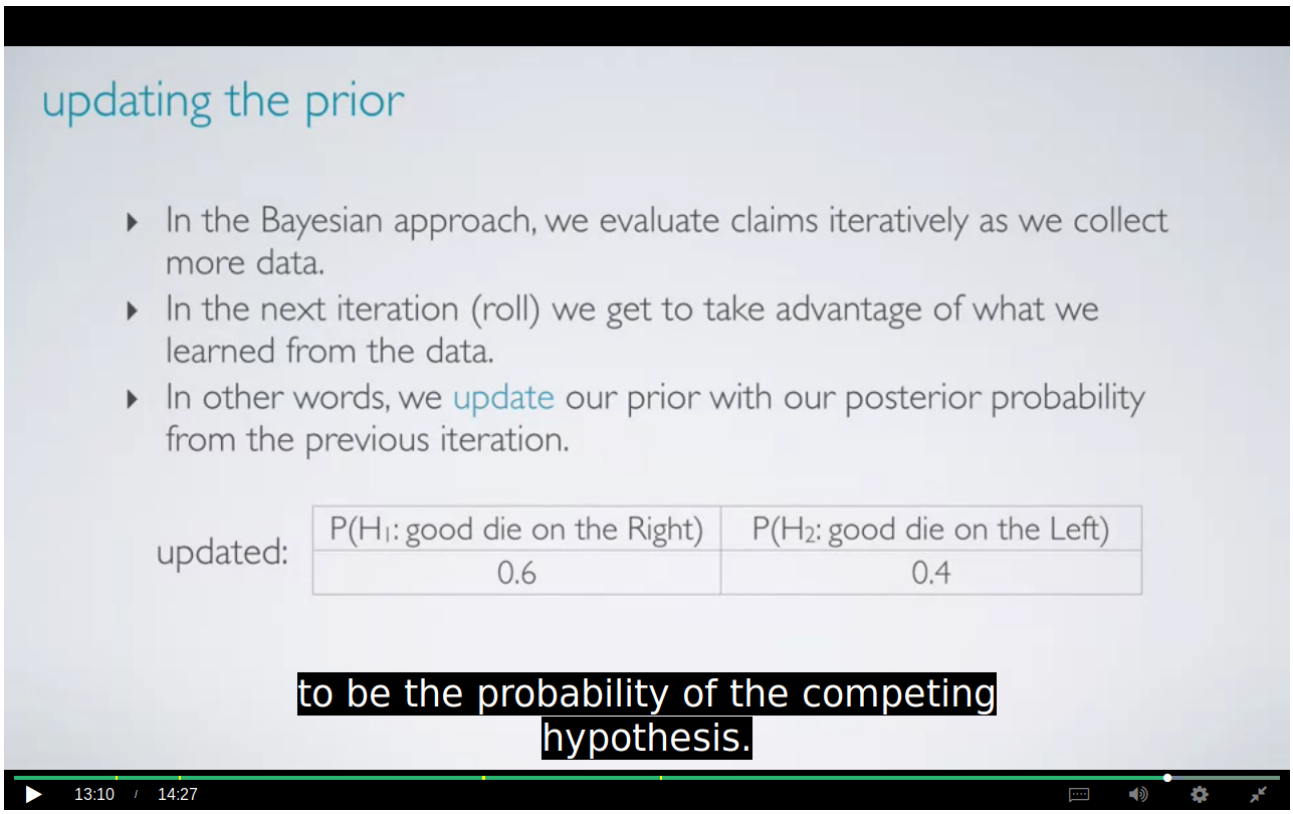
$P(HIV \mid +) = ?$

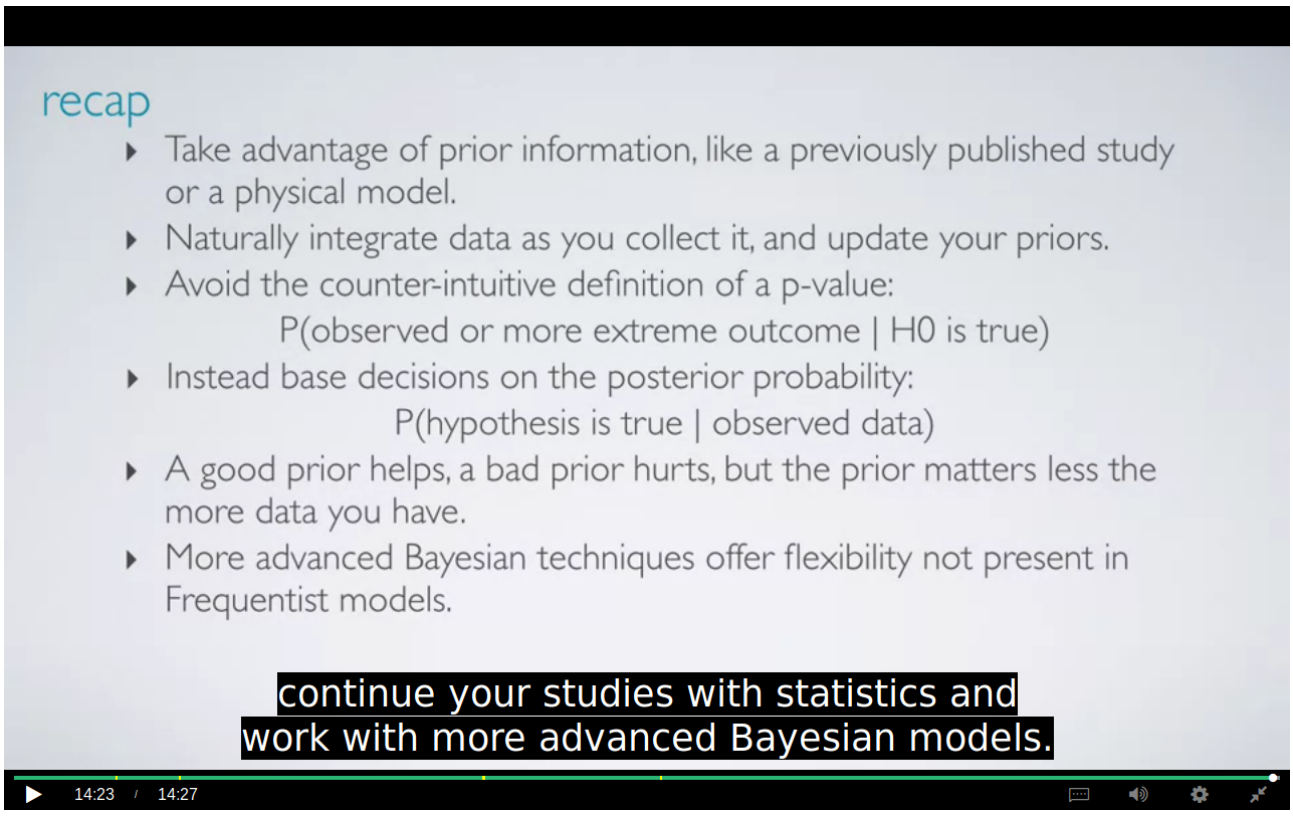


Note that the marginal probability [P(+)] is the sum of [P(+ and hiv) + P(+ and no hiv)].

BAYESIAN INFERENCE



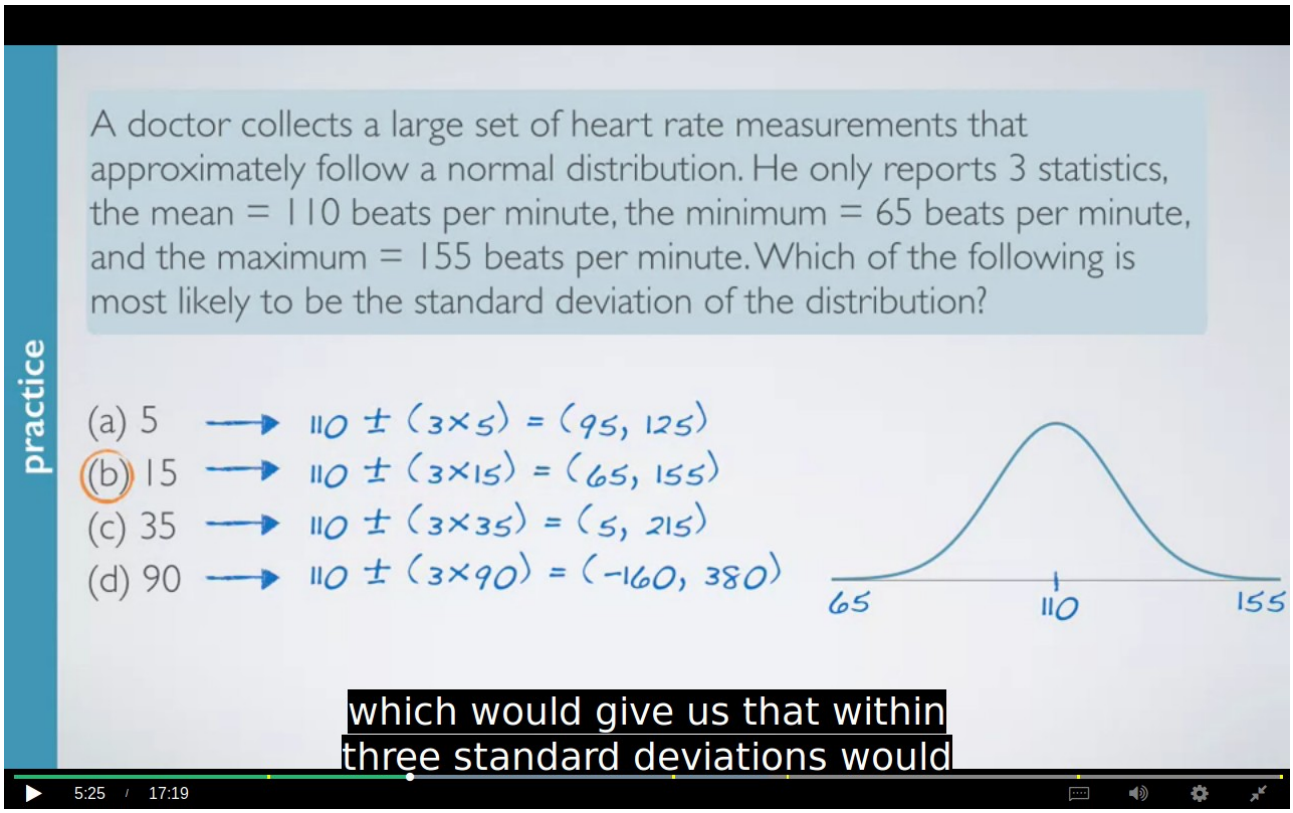




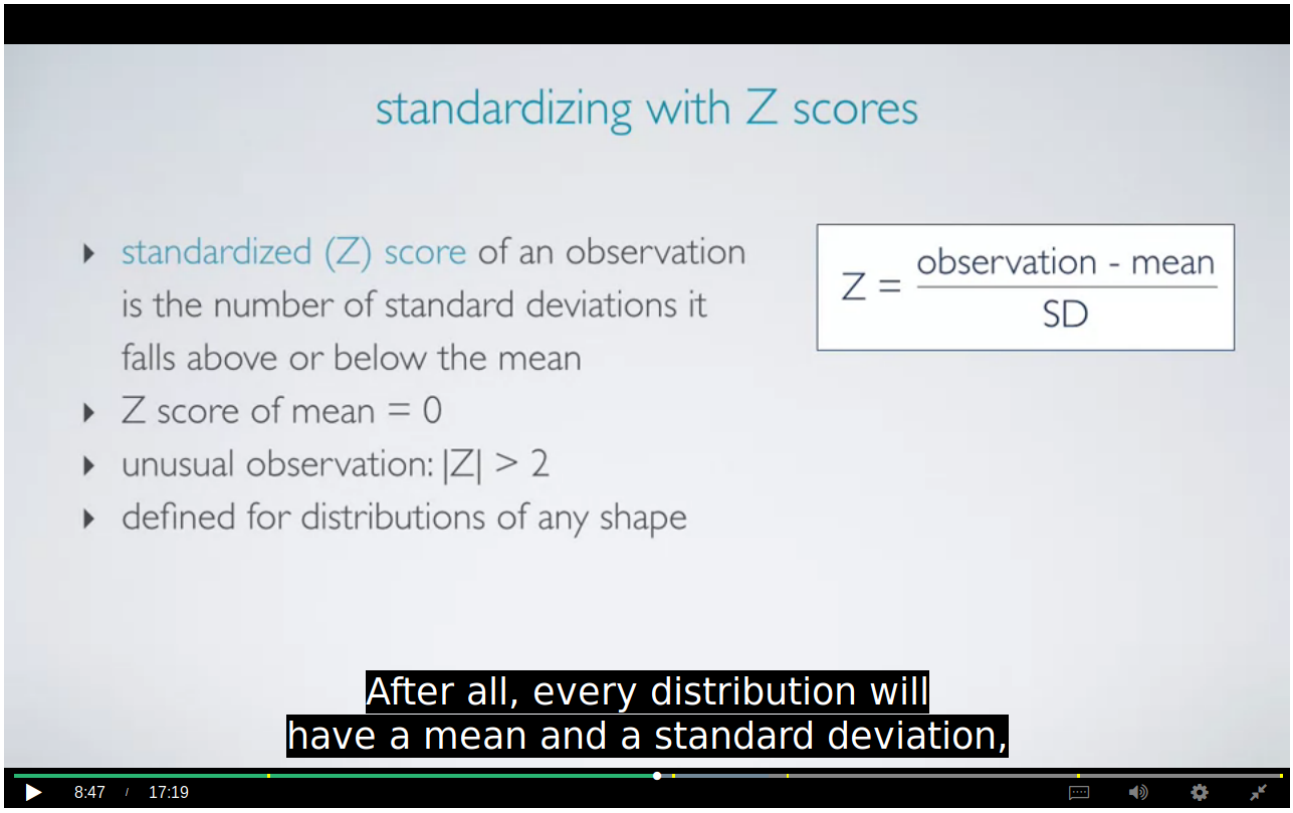
if you didn't have a great prior to begin with, as you collect my, more data, you're going to be able to converge to the right probabilities.

NORMAL DISTRIBUTION

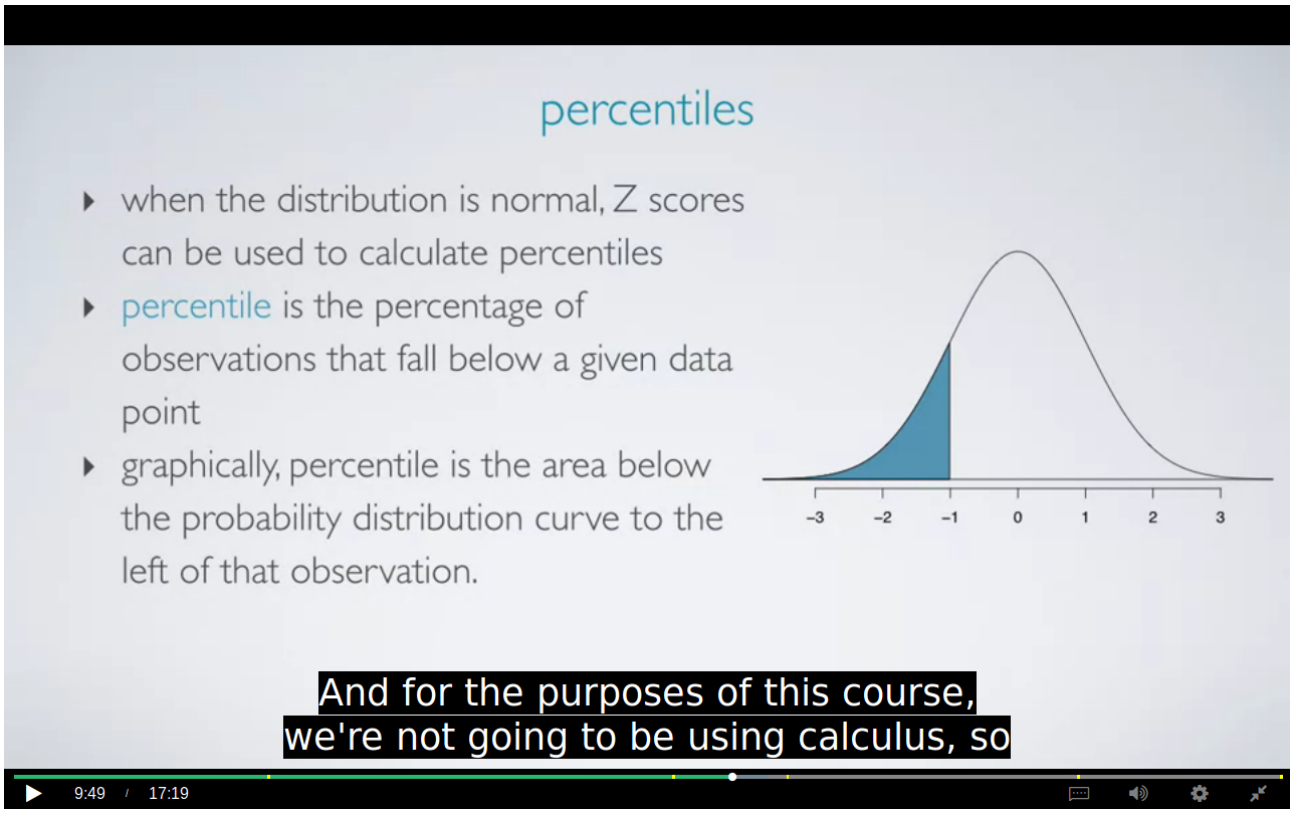
we calculate the STD with min,max and mean.

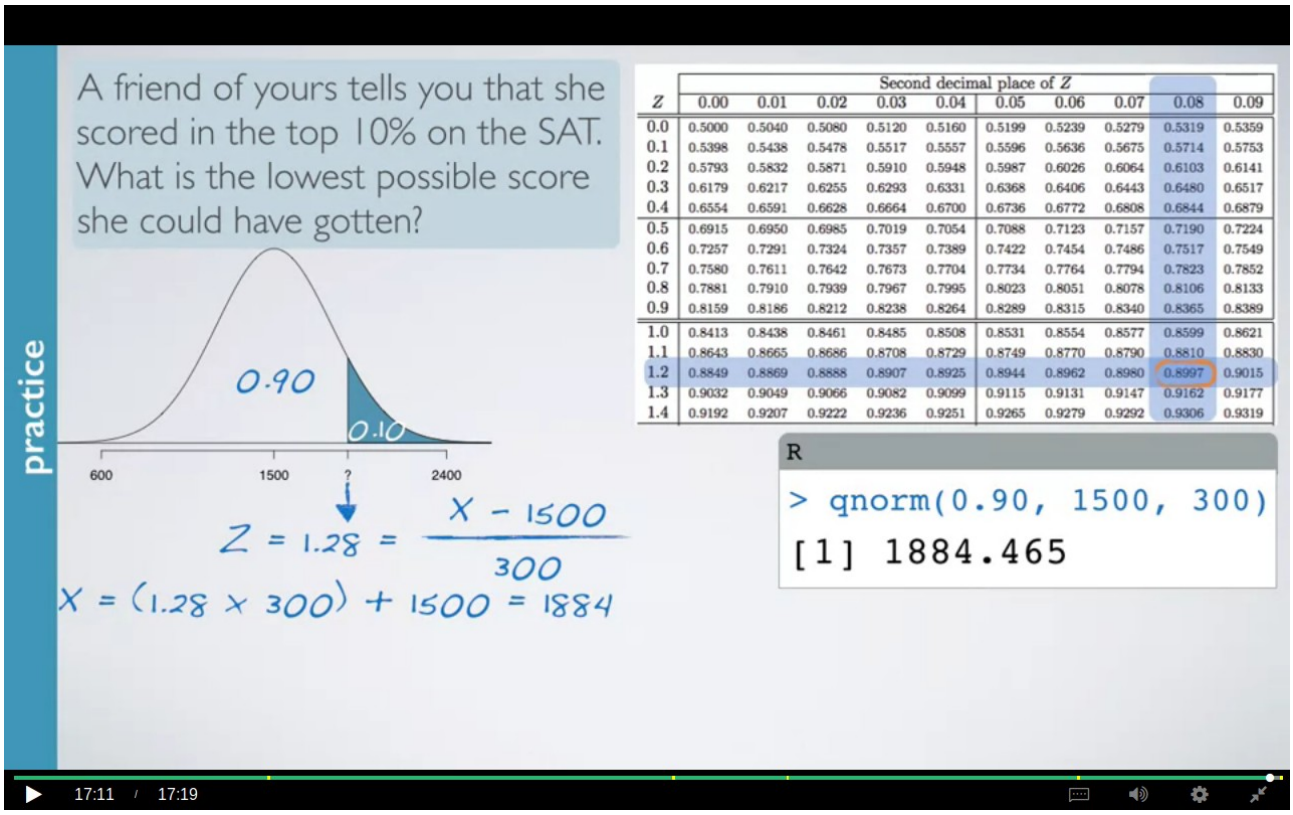


We can calculate Z scores for any distribution.

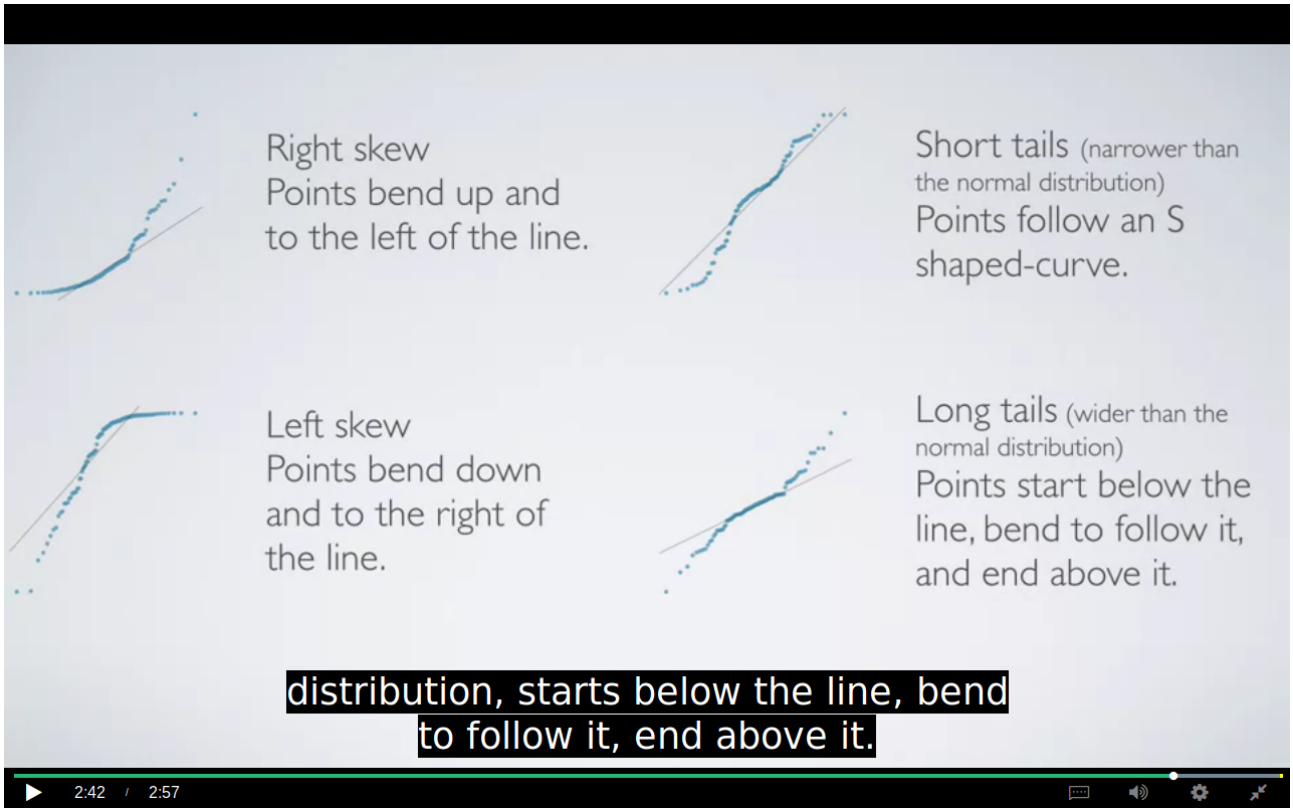


Percentiles can be calculated for dist of any shape but if it is not symmetric then we have to use calculus.





Shape of dist using Normal probability plot.



Distribution of number of hours of sleep college students get has a mean of 7 hours. 62% of students sleep between 6 and 8 hours, 92% of students sleep between 5 and 9 hours, and 95% sleep between 4 and 10 hours. Which of the following is true?

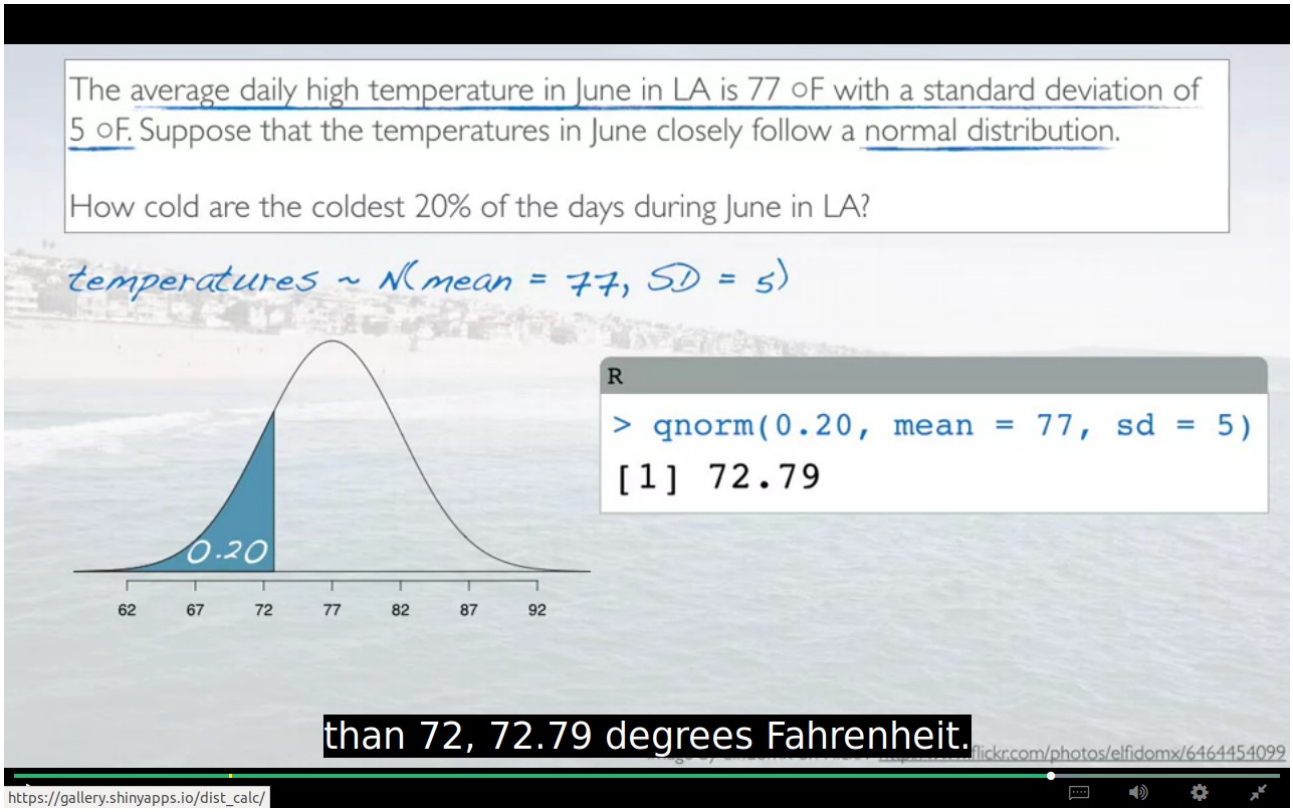
- The distribution is more variable than a normal distribution with mean 7 and standard deviation 1

Correct
In a normal distribution with mean 7 and standard deviation 1, we would expect 68% of the data to be between 6-8, 95% of the data between 5-9, and 99.7% of the data between 4-10 hours. In this data set a lower percentage than expected fall within these ranges, and hence the data are more spread out, i.e. more variable, than N(mean = 7, SD = 1).

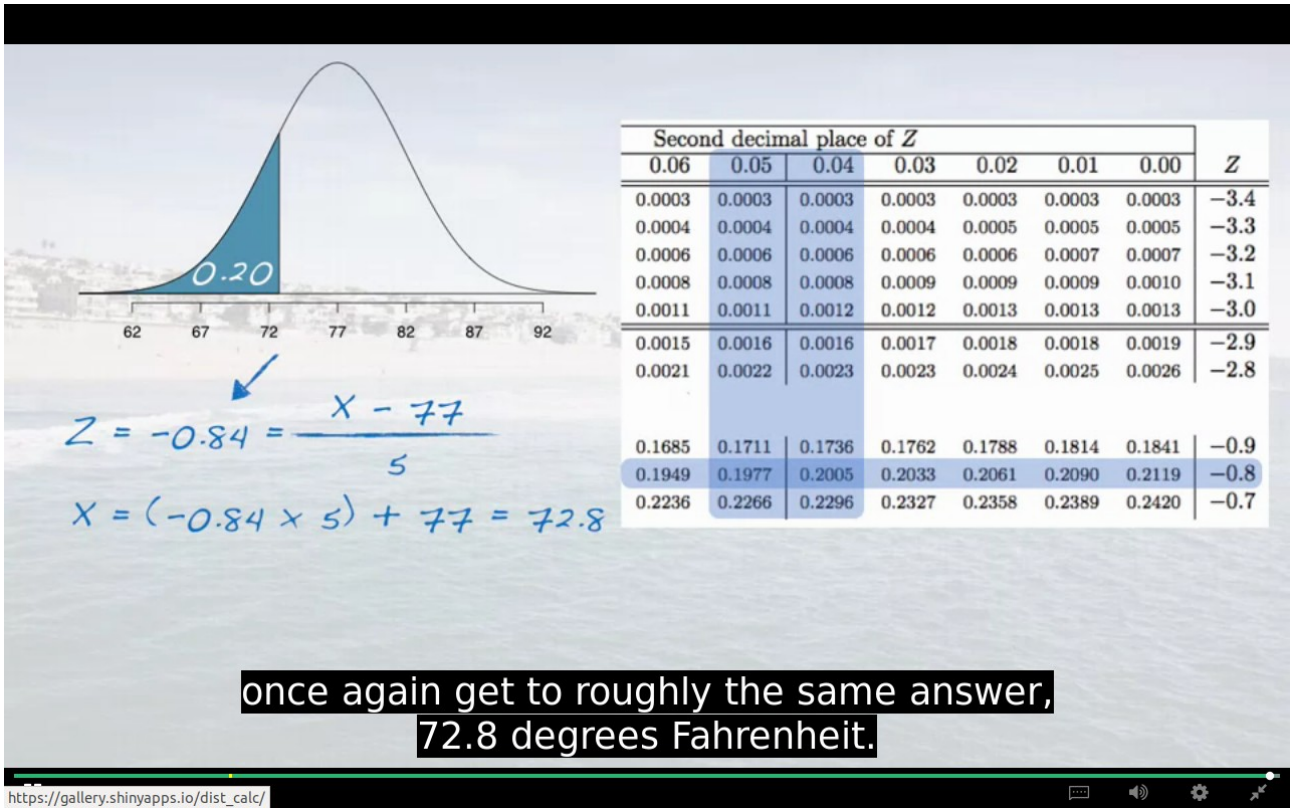
- The distribution is less variable than a normal distribution with mean 7 and standard deviation 1
- The distribution is nearly normal

SOLUTION:

```
> 1-pnorm(50,45,3.2)
[1] 0.05908512
```



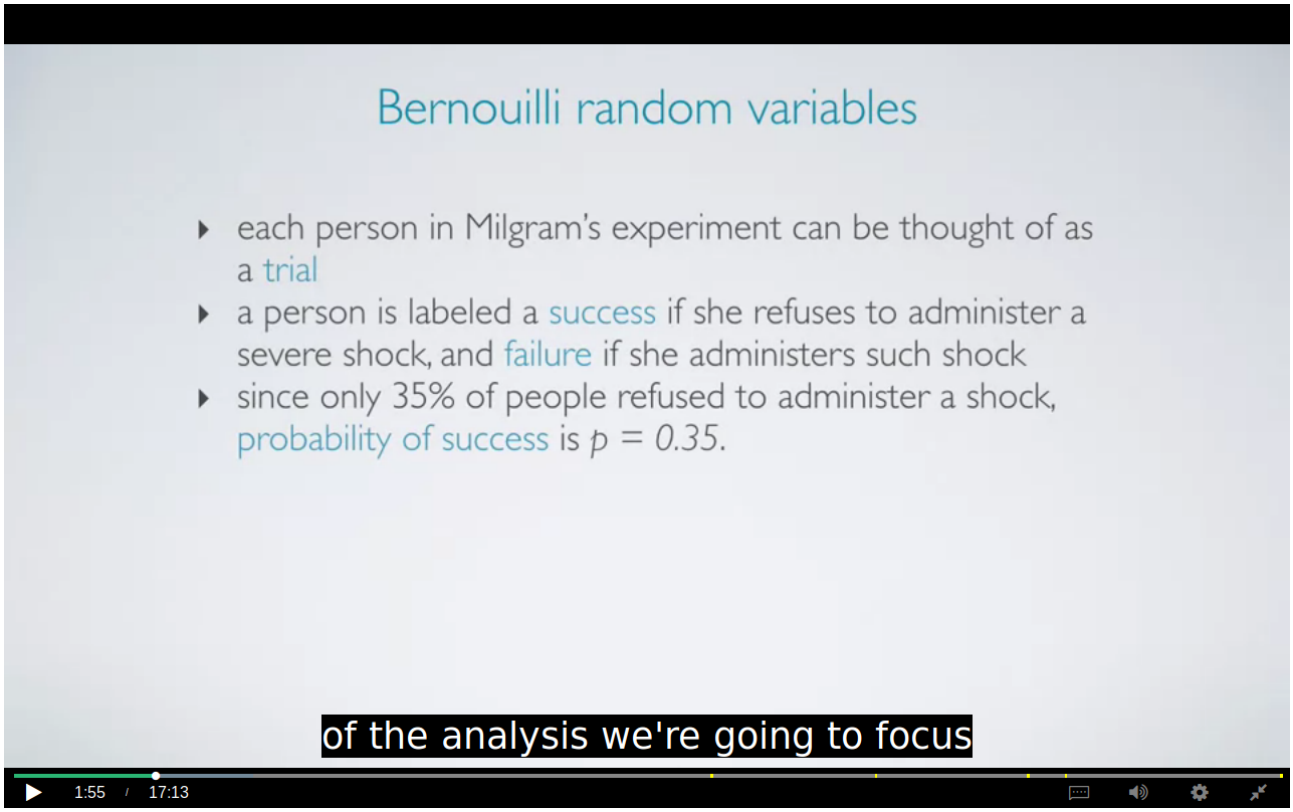
another approach for solving using Z table

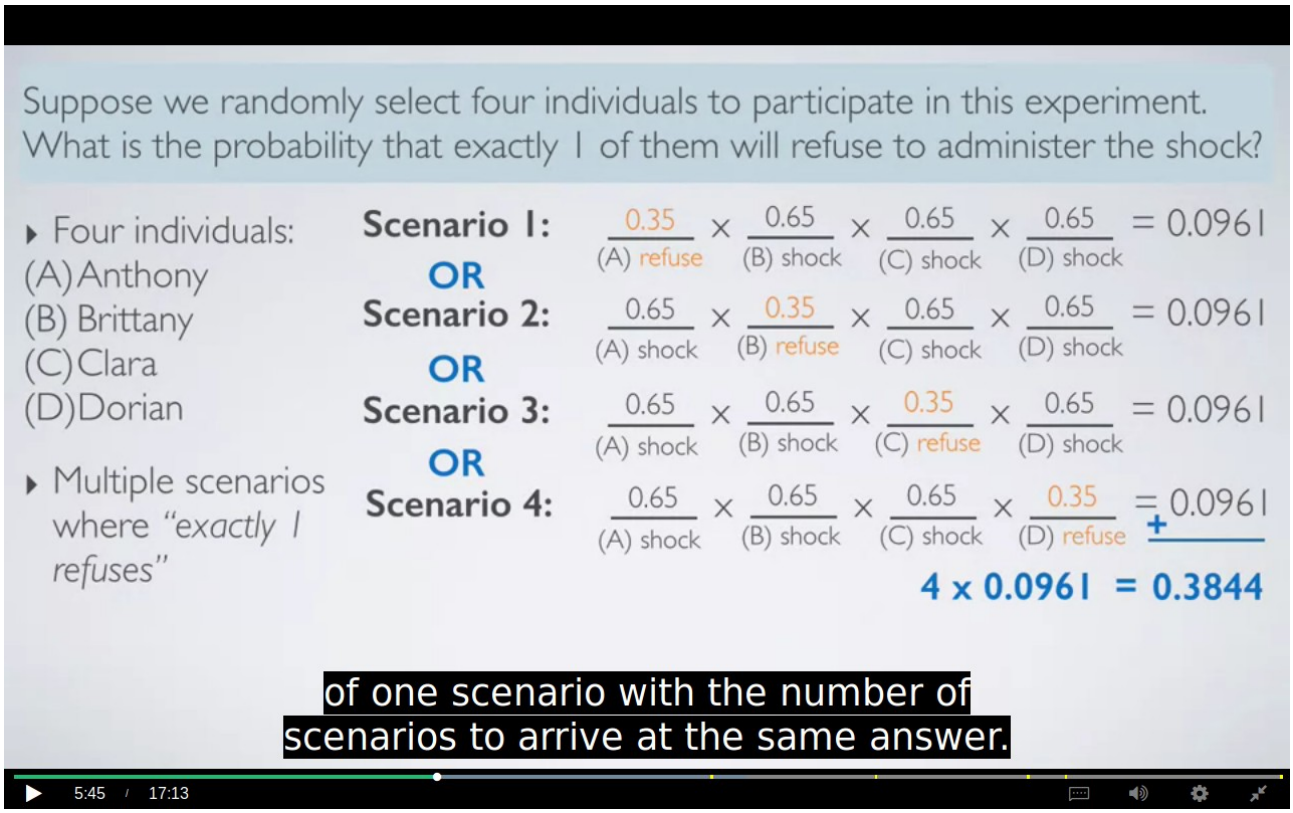


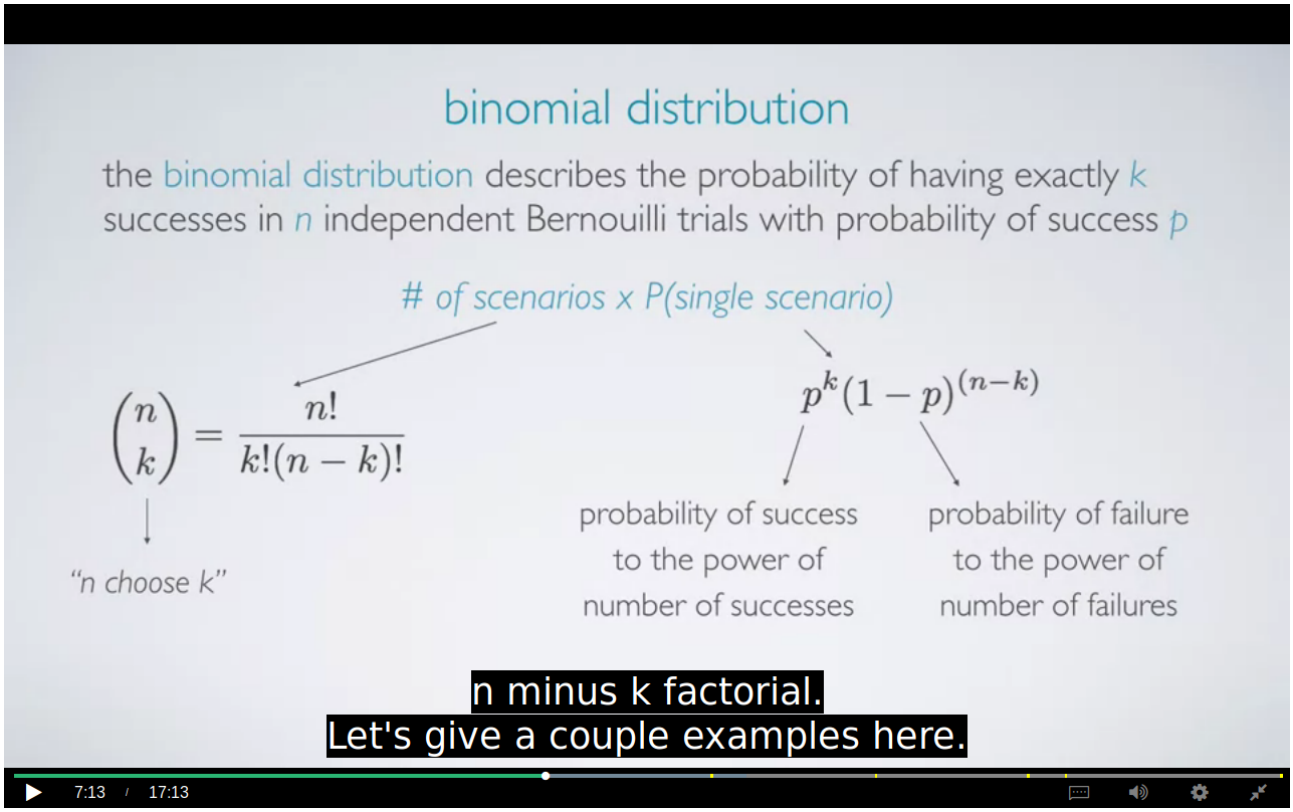
Test yourself: *True/False: In a right skewed distribution the Z score of the median is positive.*

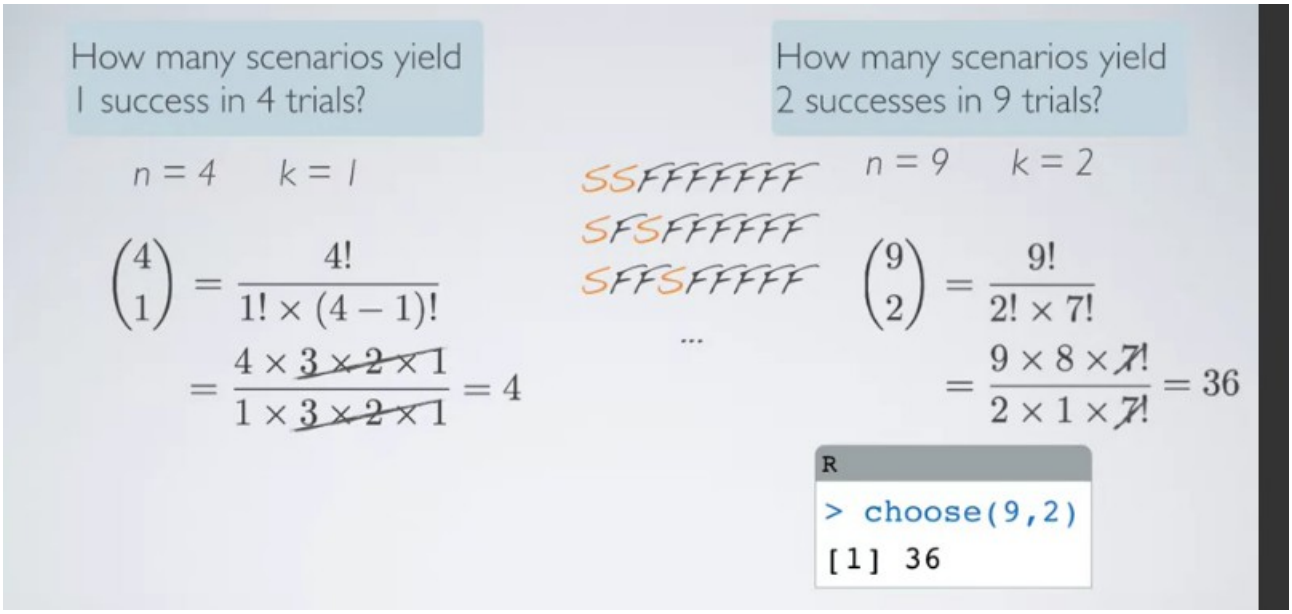
the answer is false bcz in Right skewed distribution the mean is > median.

BINOMIAL DISTRIBUTION









Binomial distribution:

If p represents probability of success, $(1-p)$ represents probability of failure, n represents number of independent trials, and k represents number of successes

$$\text{P(k successes in n trials)} = \binom{n}{k} p^k (1-p)^{(n-k)}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

binomial conditions:

1. the trials must be independent
2. the number of trials, n, must be fixed
3. each trial outcome must be classified as a success or a failure
4. the probability of success, p, must be the same for each trial

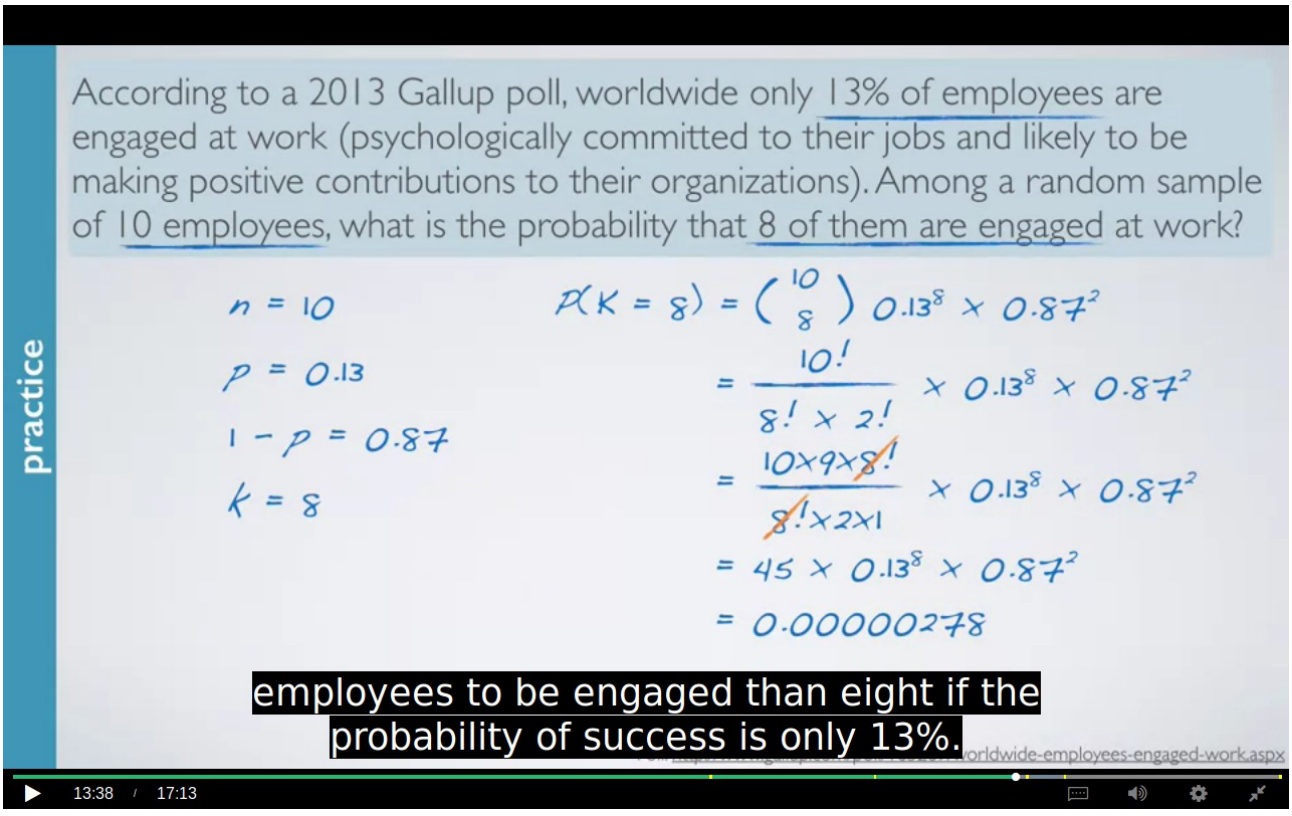
Same data:According to a 2013 Gallup poll, worldwide only 13% of employees are engaged at work (psychologically committed to their jobs and likely to be making positive contributions to their organizations). We are interested in finding the probability that among a random sample of 10 employees, what is the probability that 8 of them are engaged at work.

Without doing any calculations, would you expect this probability to be pretty low or pretty high?

- pretty low

Correct
Among a sample of 10, we would only expect $10 \times 0.13 = 1.3$ people to be engaged at work, so getting a random sample where 8 out of 10 are engaged would be pretty unlikely.

- pretty high



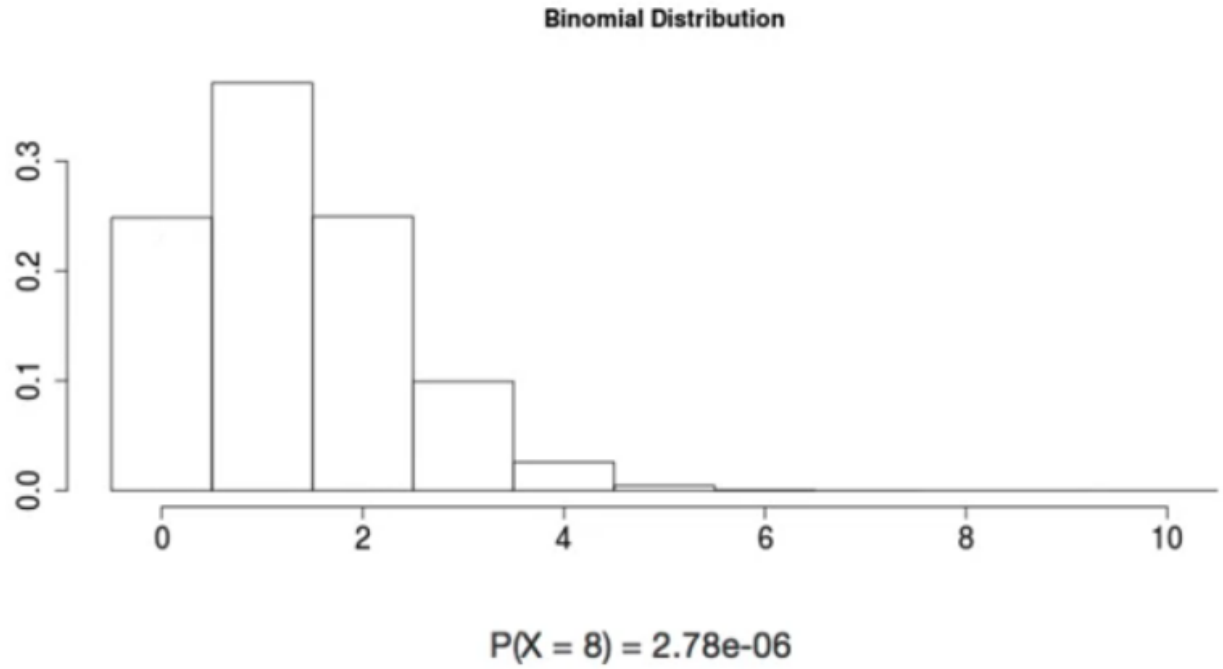
R command:

> choose(10,8) * (0.13^8) * (0.87^2)

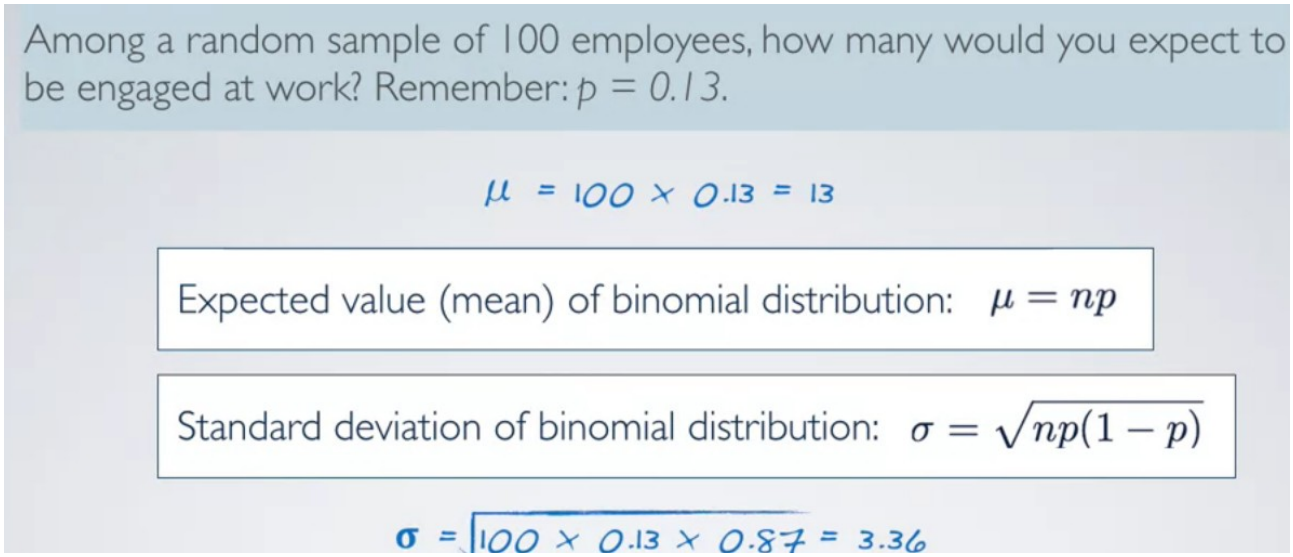
[1] 2.77842e-06

> dbinom(8,size=10,p=0.13)

plot for above question using applet.



The probability for 8 is very low therefore there is no bar above it. The height of the bar shows probability.



This means that 13 out of 100 employees are expected to be at engaged at work, give or take approximately 3.36. Note that the mean on the standard deviation of a binomial, might not always be whole numbers, and that's alright. These values represent what we would expect to see on average.

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Which of the following is false?

- Among a random sample of 1,000 Americans, we would expect 262 to be obese.
- Random samples of 1,000 Americans where there are at most 230 are obese people would be considered unusual.
- The standard deviation of number of obese Americans in random samples of 1,000 is roughly 14.
- Random samples of 1,000 Americans where at least 300 are obese would not be considered unusual.

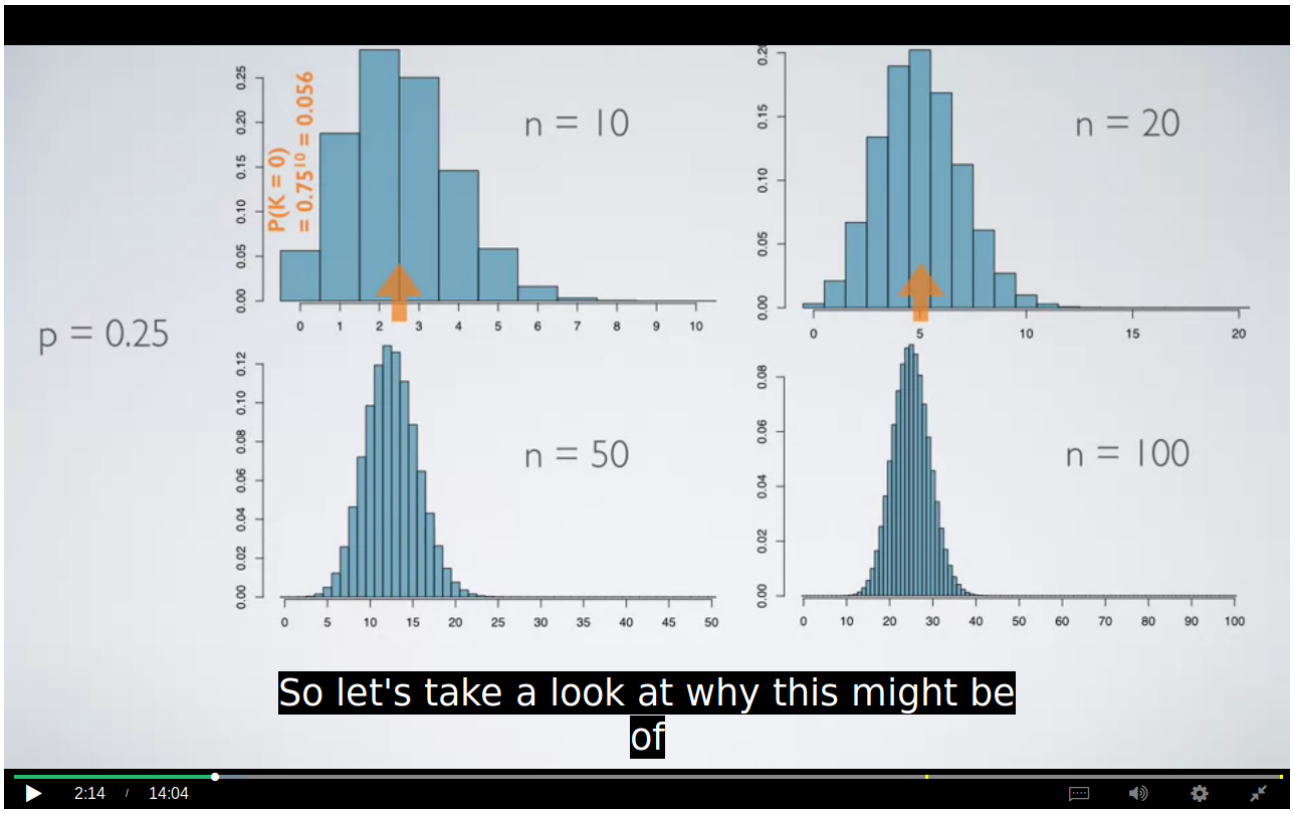
Correct

$\mu = 1000 \times 0.262 = 262$

$\sigma = \sqrt{1000 \times 0.262 \times 0.738} \approx 14$

Range of "usual" observations: $262 \pm 2 \times 14 = (234, 290)$, anything beyond these would be considered unusual.

NORMAL APPROXIMATION TO BINOMIAL:



THE CENTER OF THE DISTRIBUTION CHANGES AS WE ARE INCREASING THE NO OF TRIALS.

Calculating the area under the curve is much more simpler then calculating the individual probabilities therefore we use normal approximation

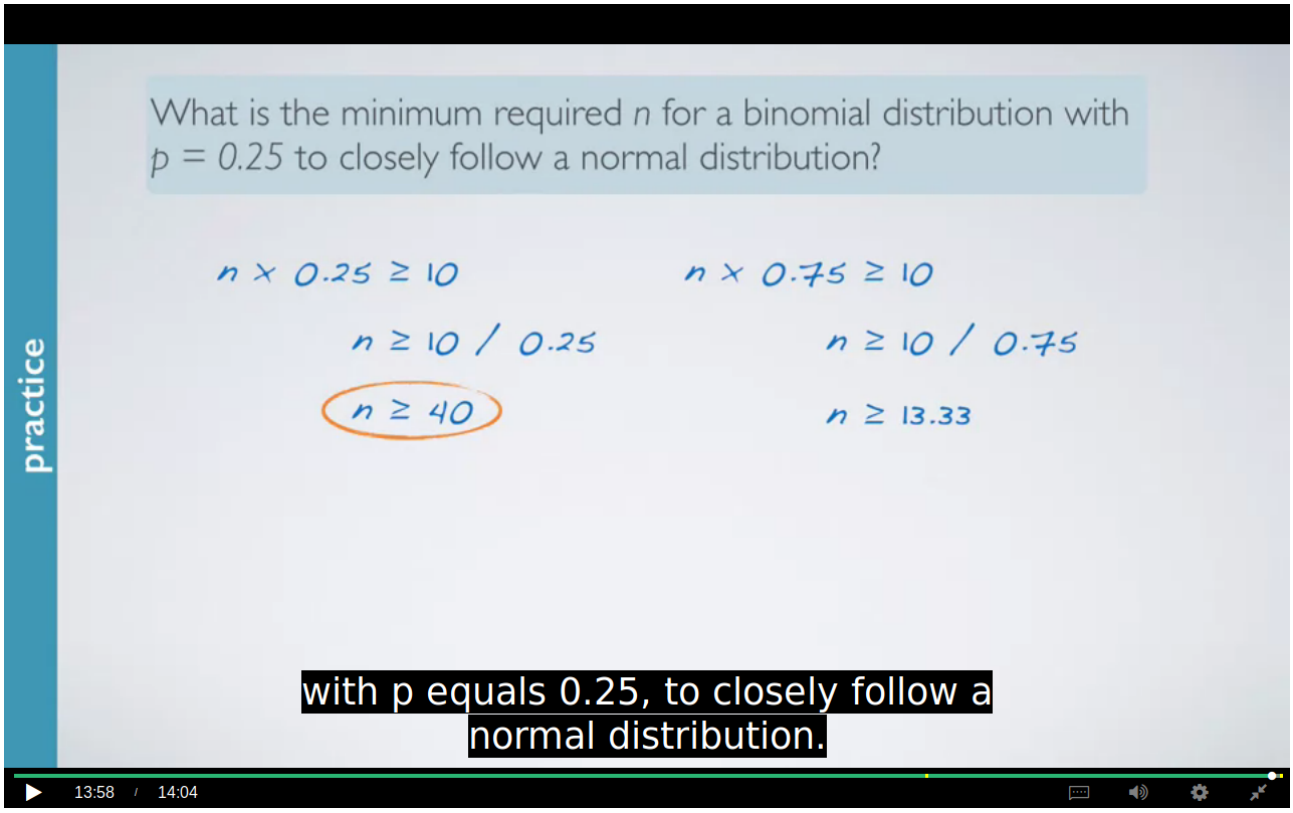
Success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution.

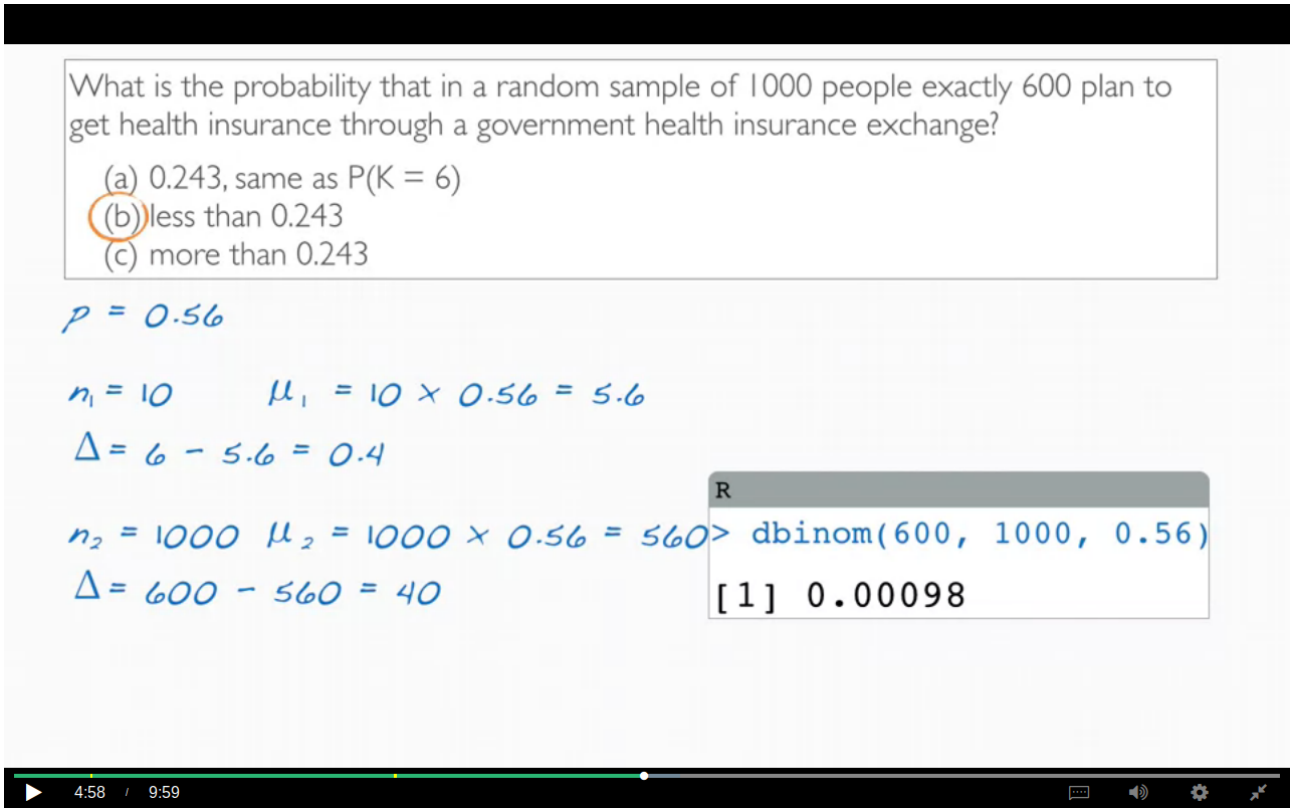
$$np \geq 10$$
$$n(1-p) \geq 10$$

Normal approximation to the binomial: If the success-failure condition holds,

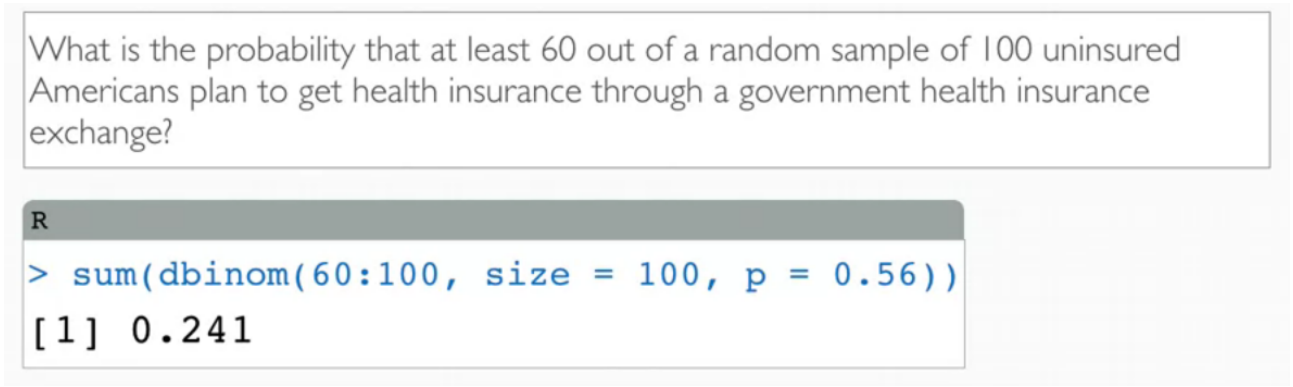
$$\text{Binomial}(n,p) \sim \text{Normal}(\mu,\sigma)$$

where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$

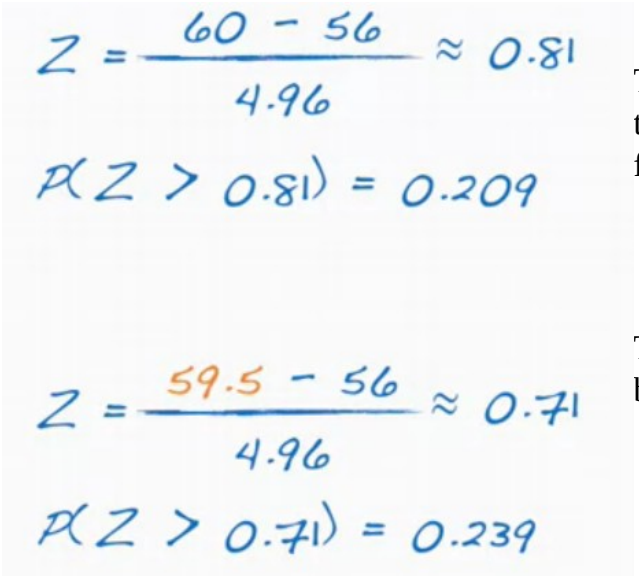




the diff between the expected and desired is much lower in the case of n=10 while it is much larger in n=1000 therefore acc to law of large number the desired outcome is pretty unusual.



Alternate Solution for same question using 0.5 correction by normal approximation.



The probability was .20 which differs from the one calculated with the binomial formula.

The probability is almost equal to the binomial after using 0.5 correction.

Important question

3. You are about to take a multi-day tour through a national park which is famous for its wildlife. The tour guide tells you that on any given day there's a 61% chance that a visitor will see at least one "big game" animal, and a 39% chance they'll see no big game animals; when the tour guide says "big game", he refers to either a moose or a bear. The guide assures you that big game sightings on a single day are independent of any other day's sightings. Given the information from the tour guide, which of the following calculations cannot be performed using a binomial distribution?

- Calculate the probability that you see at least 4 big game animals on the first day of a 5-day trip.

Correct

This question refers to the following learning objective: Determine if a random variable is binomial using the four conditions.

- The trials are independent.
- The number of trials, n, is fixed.
- Each trial outcome can be classified as a success or failure.
- The probability of a success, p, is the same for each trial.

You cannot calculate the probability of seeing at least 4 big game animals on a particular day because not enough information was given in the problem. In other words, the problem gave information on the probability of a success and failure where "success" means seeing any (i.e. at least 1) big game animals on a particular day - we have no information on the exact number of big game animals seen on a particular day.

6. About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex, half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

- ○ 33%
- ○ 50%
- ● 46%

Correct

This question refers to the following learning objective: Distinguish between marginal and conditional probabilities. Construct tree diagrams to calculate conditional probabilities and probabilities of intersection of non-independent events using Bayes' theorem: P(A | B) = P(A and B) / P(B)

A good strategy is to first decide what quantity we're being asked to calculate. Then we'll use Bayes Theorem to write the quantity in terms of other quantities which we hope to be given directly in the problem. So, we're asked to calculate P(I | G) where I is the event of known twins being identical and G the event that both twins are girls. Then use Bayes' Theorem to write $\frac{P(G|I)P(I)}{P(G|I)P(I)+P(G|F)P(F)}$. Let F be the event that known twins are fraternal. It turns out that we're given all of the quantities on the right-hand side of that equation, so the result is P(I | G) = $\frac{.5\times.3}{.5\times.3+.25\times.7}$. Your answer may vary slightly due to rounding.

CLT:

INCLUDE THE BELOW WITH CLT TOPIC:

4.5 Hen eggs. The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a

distribution of sample means.

(a) What is this distribution called?
(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.
(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.
(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

ANS 4.5:

(a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution.

(b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric.

(c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. SE = 18.2/ 45 =2.713.

(d) The sample means will be more variable with the smaller sample size.

4.13 Waiting at an ER, Part I. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.
(a) This confidence interval is not valid since we do not know if the population distribution of the ER wait times is nearly Normal.

(b) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

(c) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

(d) 95% of random samples have a sample mean between 128 and 147 minutes.

(e) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

(f) The margin of error is 9.5 and the sample mean is 137.5.

(g) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size.

4.13 (a) False. Provided the data distribution is not very strongly skewed (n = 64 in this sample, so we can be slightly lenient with the skew), the sample mean will be nearly normal, allowing for the method normal approximation described.
(b) False. Inference is made on the population parameter, not the point estimate. The point estimate is always in the confidence interval.
(c) True.
(d) False. The confidence interval is not about a sample mean.
(e) False.To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake.
(f) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval.
(g) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample 2^2= 4 times the number of people in the initial sample.