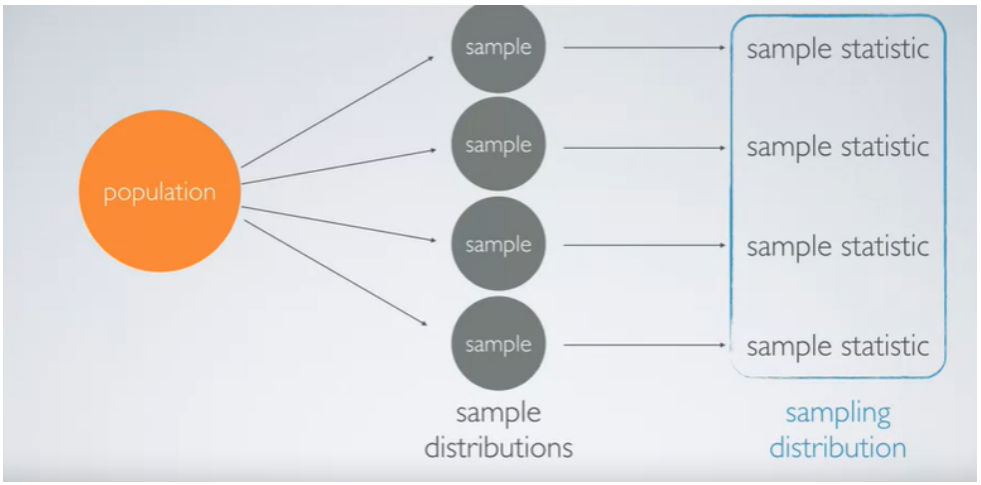
[Study note] Statistics with R specialization - 2. Inferential Statistics

2021 October - written by Mihyeon Jeon

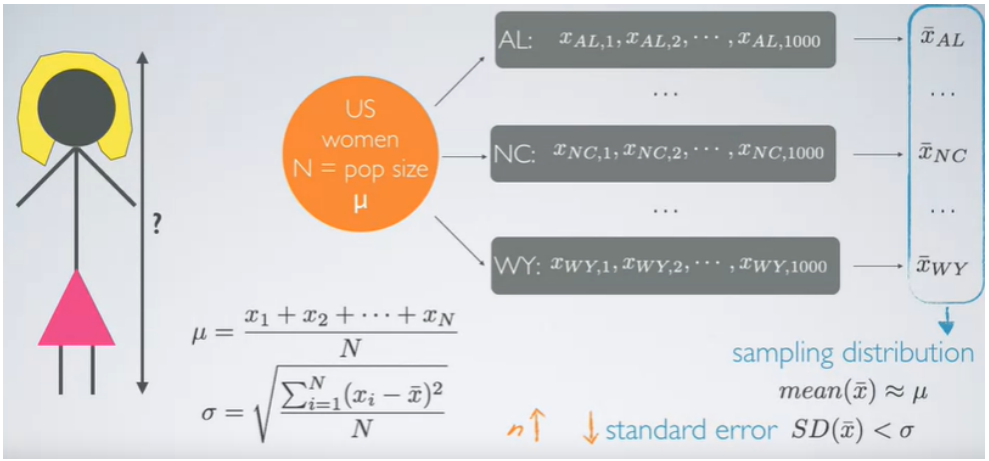
Week1

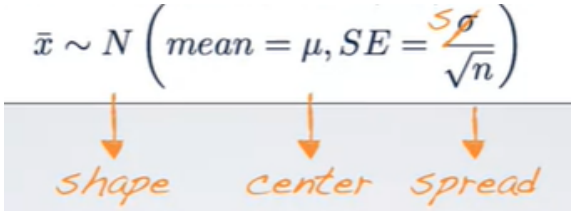
[CLT and Sampling]

1. Sampling Variability & CLT

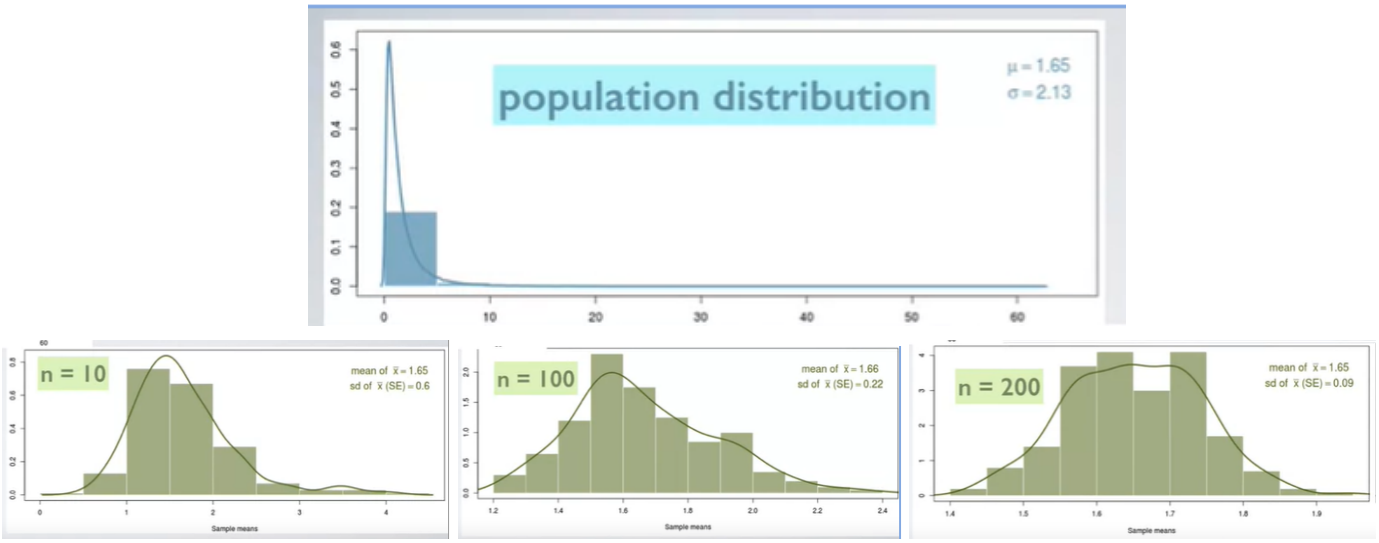


- Sample distributions and sampling distribution sound similar but different concepts.
- Ex : Height of women in US



- https://gallery.shinyapps.io/CLT_mean/
- If you take a bigger sample size in each sample population, the standard error decreases.
- **As the sample size increases, we would expect samples to yield more consistent sample means, hence the variability among the sample means would be lower, which results in a lower standard error.**
- Central Limit Theorem (CLT) : The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.
- 

- **Conditions for the CLT:**
    1. Independence : Sampled observations must be independent.
        - Random sample/assignment
        - If sampling without replacement, n < 10% population
    2. Sample size/skew : Either the population distribution is normal, of if the population distribution is skewed, the sample size is large (rule of thumb : n > 30)
- This distribution of the population is also something very difficult to verify because we often do not know what the population looks like. That's why we're doing this investigation in the first place, but we can check it using the sample data. And assume that the sample mirrors the population so if you make a plot of your sample distribution and it looks nearly normal then you might be fairly certain that the parent population distribution is coming from is nearly normal as well.
- Why 10% condition - “if sampling without replacement, n<10% or the population” ?
  : if you grab too large a sample size, each individual is not likely to be independent. It’s good to have a large size of sample, but we also do not want our samples to be much larger than or any more than 10% of our population.
- Sample size / skew condition



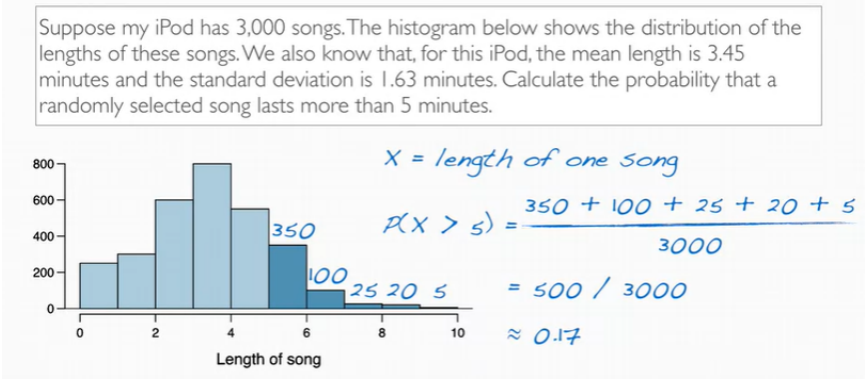
: As the sample size is bigger, the sampling distribution starts to resemble a closely normal distribution
Once we have a normal distribution, calculating probabilities which will later serve as P values in hypothesis tests are relatively simple. Having a nearly normal distribution that relies on CLT is going to open up a bunch of doors for us for doing statistical inference using confidence intervals and hypothesis tests using normal distribution theory.

- The more the skew, the higher the sample size you need for the central limit theorem to kick in.

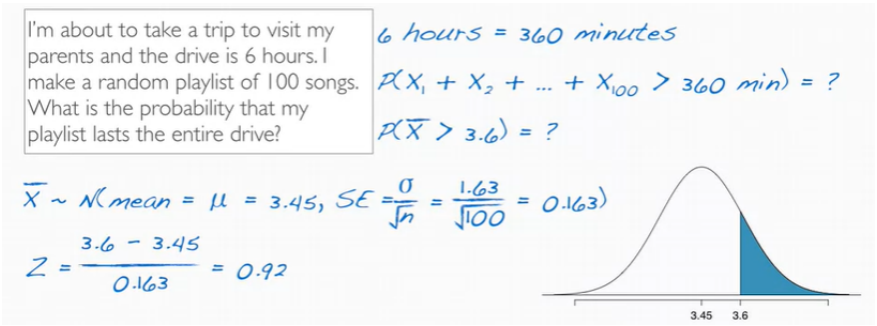
+ Standard error : 표준 오차(평균의 SE)는 같은 모집단에서 여러 표본을 추출하는 경우 얻게 될 표본 평균 간의 변동성을 추정한다.

2. CLT(for the mean) examples

1)

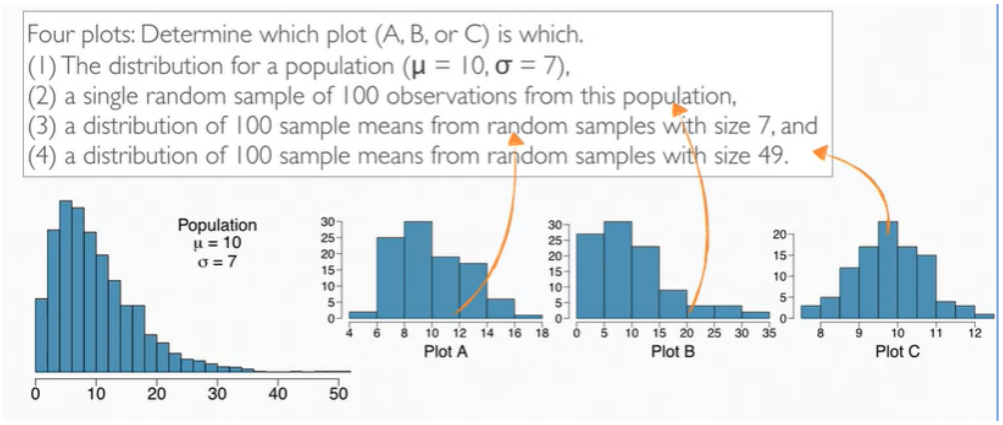


2)



- You should use the stand error, not the standard deviation of the entire population as we are interested in X bar!
- P(Z>0.92) = 0.179

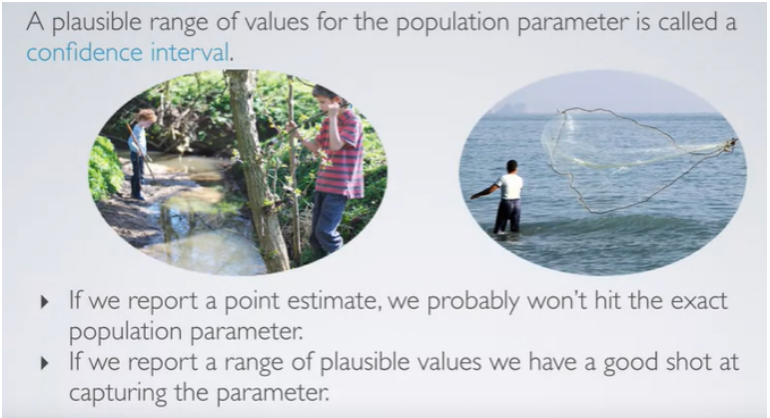
3)

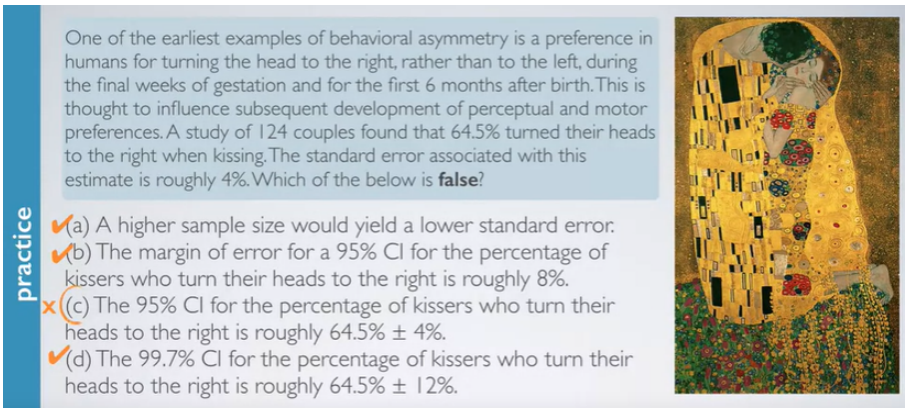


[Confidence Intervals(신뢰구간)]

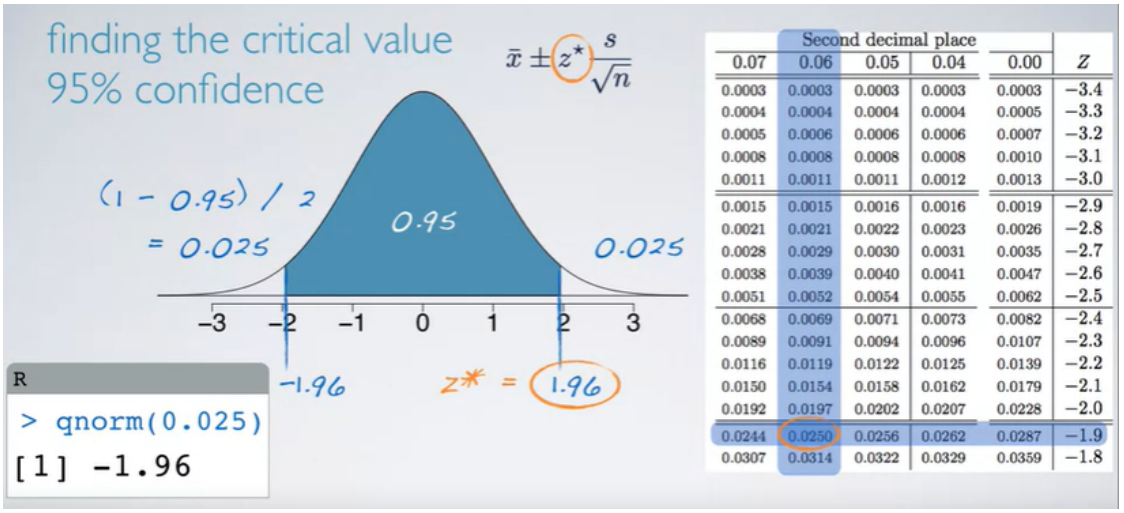
1. Confidence Interval (for a mean)

- Confidence interval : A plausible range of values for the population parameter





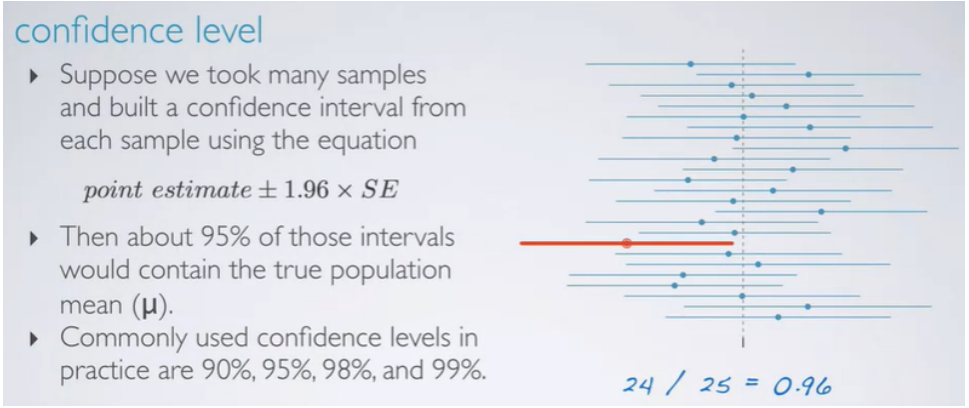
- Confidence interval for a population mean : Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).
- **Conditions for this confidence interval** :
    1. Independence : Sampled observations must be independent.
        - random sample/assignment
        - if sampling without replacement, n<10% of population
    2. Sample size/skew : n >= 30, larger if the population distribution is very skewed.

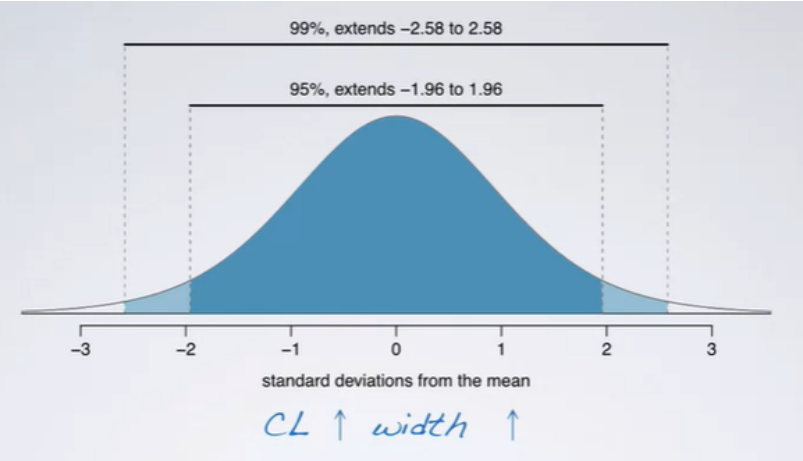


	Second decimal place				
0.07	0.06	0.05	0.04	0.00	z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

- If we are looking for a critical value, we are always going to need the positive version of the number in R.

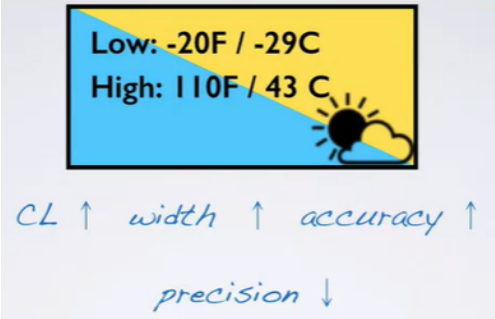
2. Accuracy vs. Precision





More accurate means a higher confidence level.
If we want higher accuracy, it means we also need to increase the confidence level.

- What drawbacks are associated with using a wider interval?



The width of the confidence interval increases as well which then increases the accuracy. However, the precision goes down.

- How can we get the best of both worlds - higher precision and higher accuracy?
  : Increase sample size!

ex :

practice

The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from 1,154 US residents. Based on the survey results, a 95% confidence interval for the average number of hours Americans have to relax or pursue activities that they enjoy after an average work day was found to be 3.53 to 3.83 hours. Determine if each of the following statements are true or false.

F (a) 95% of Americans spend 3.53 to 3.83 hours relaxing after a work day.
T (b) 95% of random samples of 1,154 Americans will yield confidence intervals that contain the true average number of hours Americans spend relaxing after a work day.
F (c) 95% of the time the true average number of hours Americans spend relaxing after a work day is between 3.53 and 3.83 hours.
F (d) We are 95% confident that Americans in this sample spend on average 3.53 to 3.83 hours relaxing after a work day.

- The confidence interval is not about the sample mean, but is instead about the population mean.
- The population parameter is not this moving target that is sometimes within an interval and sometimes outside of it.

3. Required Sample Size for ME

backtracking to n for a given ME

given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

$$ME = z^\star \frac{s}{\sqrt{n}} \rightarrow n = \left(\frac{z^\star s}{ME}\right)^2$$

ex :

practice

A group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this medication during pregnancy.

Previous studies suggest that the SD of IQ scores of three-year-old children is 18 points.

How many such children should the researchers sample in order to obtain a 90% confidence interval with a margin of error less than or equal to 4 points?

$ME \leq 4$ pts

$CL = 90\%$

$z^* = 1.65$

$\sigma = 18$

$$4 = 1.65 \frac{18}{\sqrt{n}} \rightarrow n = \left(\frac{1.65 \times 18}{4}\right)^2 = 55.13$$

We need at least 56 such children in the sample to obtain a maximum margin of error of 4 points.

- Even though mathematically this number will be rounded to 55, actually in calculations of minimum required sample size, regardless of the value of the decimal, we always want to round up.
- What if we want lower ME?

practice

We found that we needed at least 56 children in the sample to achieve a maximum margin of error of 4 points. How would the required sample size change if we want to further decrease the margin of error to 2 points?

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{n}} \frac{1}{2}$$

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{4n}}$$

4. CI (for the mean) examples

ex 1)

The General Social Survey asks: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. Interpret this interval in context of the data.

= We are 95% confident that Americans on average have 3.40 to 4.24 bad mental health days per month.

- In this context, what does a 95% confidence level mean?
  : 95% of random samples of 1151 Americans will yield CIs that capture the true population mean of number of bad mental health days per month.

ex 2)

A sample of 50 college students were asked how many exclusive relationships they've been in so far. The students in the sample had an average of 3.2 exclusive relationships, with a standard deviation of 1.74. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average number of exclusive relationships based on this sample using a 95% confidence interval.

1. random sample & 50 < 10% of all college students
We can assume that the number of exclusive relationships one student in the sample has been in is independent of another.
2. $n > 30$ & not so skewed sample
We can assume that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal.

$n = 50$
$\bar{x} = 3.2$
$s = 1.74$

$n = 50$
$\bar{x} = 3.2$
$s = 1.74$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$$

$$\bar{x} \pm z^* SE = 3.2 \pm 1.96\,(0.246)$$
$$= 3.2 \pm 0.48$$
$$= (2.72, 3.68)$$

We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.

Quizz

This question refers to the following learning objective(s):

Recognize
that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

- In the case of the mean the CLT tells us that if

(1a) the sample size is sufficiently large $(n \geq 30)$ and the data are not extremely skewed or

(1b) the population is known to have a normal distribution, and

(2) the observations in the sample are independent,

then
the distribution of the sample mean will be nearly normal, centered at the true population mean and with a standard error of $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

- When
the population distribution is unknown, condition (1a) can be checked using a histogram or some other visualization of the distribution of the observed data in the sample.
- The larger the sample size
(n), the less important the shape of the distribution becomes, i.e. when n is very large the sampling distribution will be nearly normal regardless of the shape of the population distribution.

Review the associated learning objective.

5. The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. In 2010, the survey collected responses from over a thousand US residents. The survey is conducted face-to-face with an in-person interview of a randomly-selected sample of adults. One of the questions on the survey is "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?"

   Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. Given this information, which of the following statements would be most appropriate to make regarding the true average number of days of "not good" mental health in 2010 for US residents?

   - ○ For all US residents in 2010, based on this 95% confidence interval, we would reject a null hypothesis stating that the true average number of days of "not good" mental health is 5 days.
   - ○ There is not sufficient information to calculate the margin of error of this confidence interval.
   - ◉ For all US residents in 2010, there is a 95% probability that the true average number of days of "not good" mental health is between 3.40 and 4.24 days.
   - ○ For these 1,151 residents in 2010, we are 95% confident that the average number of days of "not good" mental health is between 3.40 and 4.24 days.

   ✕ **Incorrect**

   This question refers to the following learning objective(s):

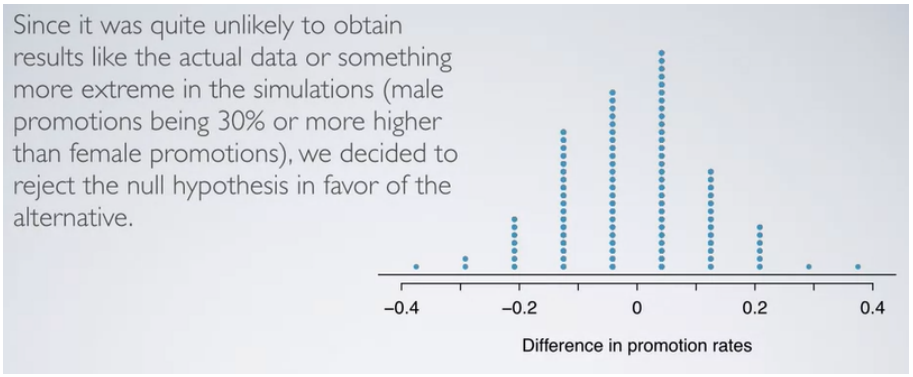
   - Interpret a confidence interval as "We are XX% confident that the true population parameter is in this interval", where XX% is the desired confidence level.
   - Define margin of error as the distance required to travel in either direction away from the point estimate when constructing a confidence interval.

   The confidence level is not about the probability of the population parameter being in one single confidence interval, instead it's about the percentage of confidence intervals at that confidence level that are expected to capture the true parameter.

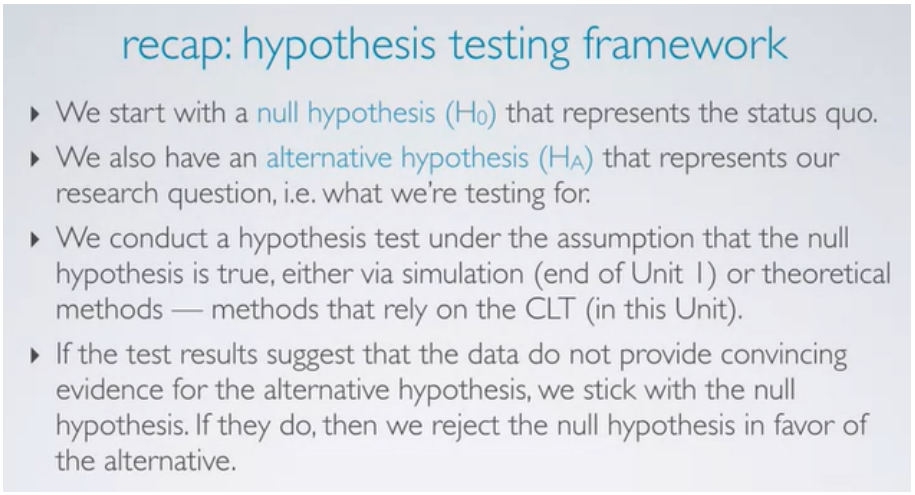
Week2

[Hypothesis Testing]

1. Another Introduction to Inference



- Each one of these dots here in the dot plot represents a one simulated difference between the proportions of males and females getting promoted.



2. Hypothesis Testing (for a mean)

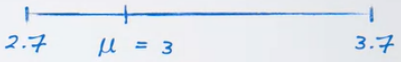
- Hypotheses
- null - H_0 : Often either a skeptical perspective or a claim to be tested (=)
- alternative - H_A: Represents an alternative claim under consideration and is often represented by a range of possible parameter values. (<, >, !=)
- The skeptic will not abandon the H_0 unless the evidence in favor of the H_A is so strong that she rejects H_0 in favor of H_A

ex)

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

H_0: $\mu = 3$ College students have been in 3 exclusive relationships, on average.

H_A: $\mu > 3$ College students have been in more than 3 exclusive relationships, on average.



- The interval says that any value within it could conceivably be the true population mean.
- Therefore cannot reject the null hypothesis in favor of the alternative.
- This does not tell us the p-value so it's quick but not trustworthy.
- The hypotheses are always about the population parameters, not about the sample statistics.
-

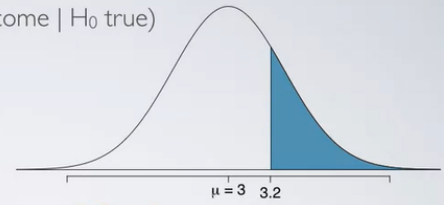
- p-value : P(observed or more extreme outcome | H_0 true)

p-value

P(observed or more extreme outcome | H_0 true)

$P(\bar{X} > 3.2 \mid H_0: \mu = 3)$

$\bar{X} \sim N(\mu = 3, SE = 0.246)$



$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

$SE = 0.246$

test statistic $Z = \frac{3.2 - 3}{0.246} = 0.81$

$p\text{-value} = P(Z > 0.81) = 0.209$

decision based on the p-value

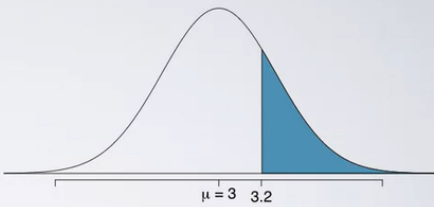
- We used the test statistic to calculate the p-value, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

- Since p-value is high (more than 0.05), we do not reject H_0.

interpreting the p-value

- If in fact college students have been in 3 exclusive relationships on average, there is a 21% chance that a random sample of 50 college students would yield a sample mean of 3.2 or higher.
- This is a pretty high probability, so we think that a sample mean of 3.2 or more exclusive relationships is likely to happen simply by chance.

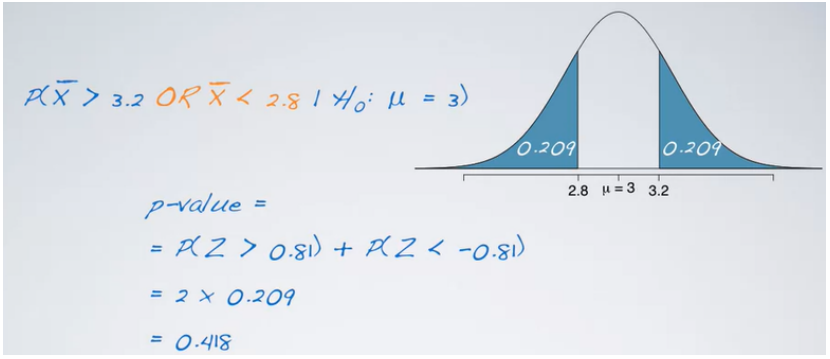
p-value = 0.209 ≈ 0.21



- making a decision
- Since the p-value is high (higher than 5%), we fail to reject H0.

- These data do not provide convincing evidence that college students have been in more than 3 relationships on average.
- The difference between the null value of 3 relationships and the observed sample mean of 3.2 relationships is due to chance or sampling variability.

- two-sided tests
- Often instead of looking for a divergence from the null in a specific direction, we might be interested in divergence in any direction.
- We call such hypothesis tests two-sided (or two-tailed)
- The definition of a p-value is the same regardless of doing a one or two-sided test, however the calculation is slightly different since we need to consider “at least as extreme as the observed outcome” in both directions.



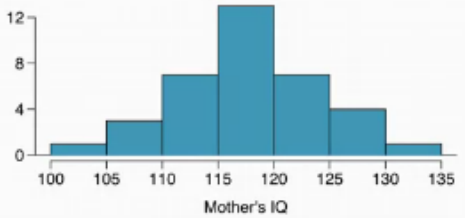
Hypothesis testing for a single mean:

1. Set the hypotheses: $H_0 : \mu = null\ value$
   $H_A : \mu < or > or \neq null\ value$
2. Calculate the point estimate: $\bar{x}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\bar{x} - \mu}{SE}$, $SE = \frac{s}{\sqrt{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value < α, reject H_0; the data provide convincing evidence for H_A.
   - If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.

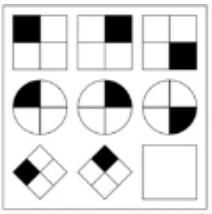
3. Hypothesis Testing (for a mean) examples

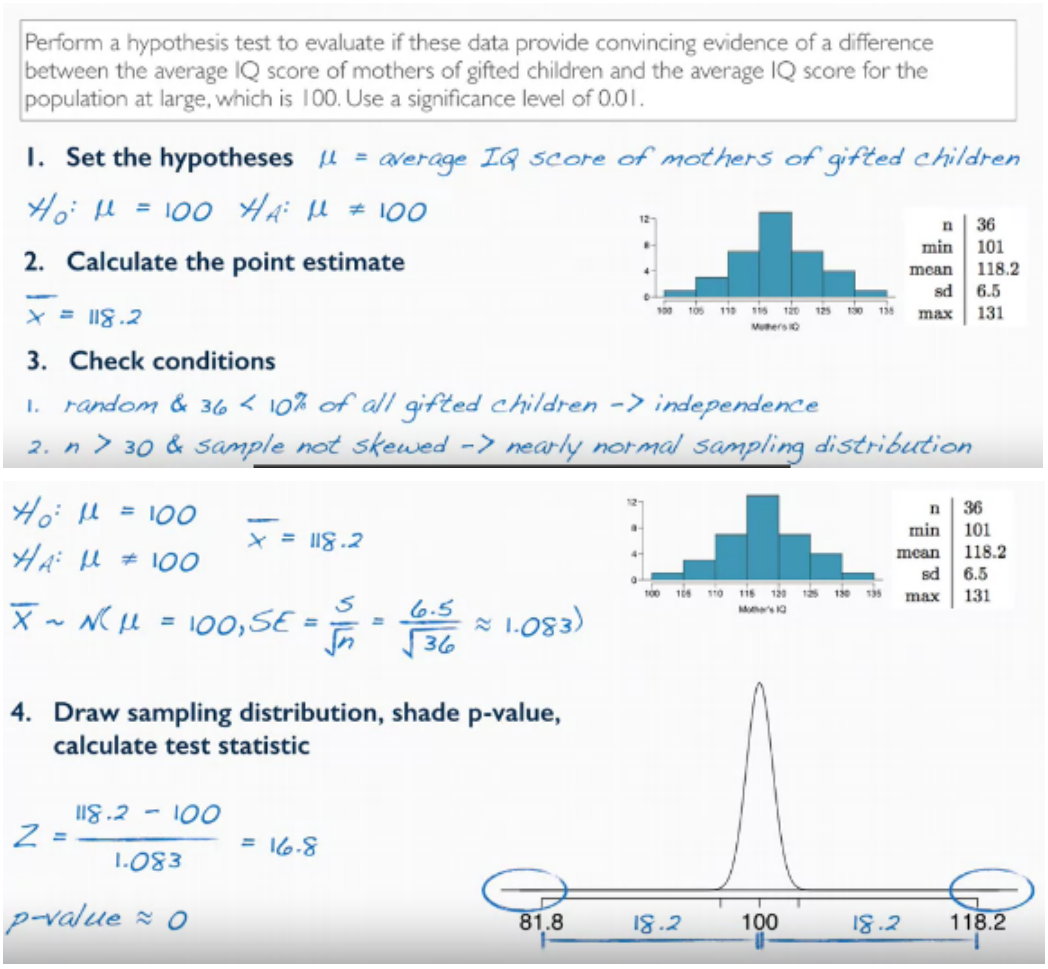
ex 1)

Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. In this study, along with variables on the children, the researchers also collected data on their mothers' IQ scores. The histogram shows the distribution of these data, and also provided are some sample statistics.



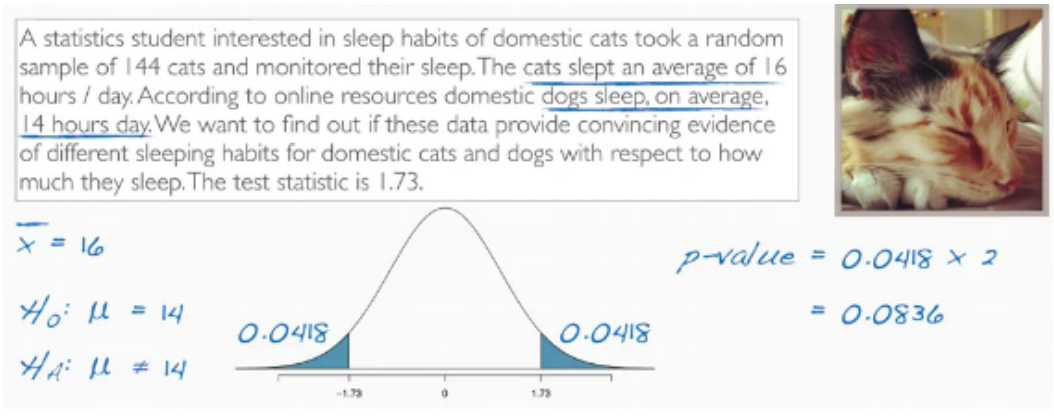
n	36
min	101
mean	118.2
sd	6.5
max	131





-> p-value is very low. = strong evidence against the null hypothesis
-> We reject the null hypothesis and conclude that the data provide convincing evidence of a difference between the average IQ score of mothers of gifted children and the average IQ score for the population at large.

ex 2)



-> p-value means P(observed or more extreme outcome | H0 true)
-> P(obtaining a random sample of 144 cats that sleep 16 hours or more or 12 hours or less, on average, if in fact cats truly slept 14 hours per day on average) = 0.0836
(고양이가 평균 14시간을 잘 때 144마리의 고양이 집단이 16시간 이상 혹은 12시간 이하로 잘 확률, 즉 null 가정이 맞다고 가정할 때, 평균에서 벗어날 확률 -> 이 가능성이 작을수록 null hypothesis를 뒷받침할 수 없음.)

[Significance]

1. Inference for Other Estimators

nearly normal sampling distributions

sample mean	$\bar{x}$
difference between sample means	$\bar{x}_1 - \bar{x}_2$
sample proportion	$\hat{p}$
difference between sample proportions	$\hat{p}_1 - \hat{p}_2$

- Unbiased estimator
- An important assumption about point estimates is that they are unbiased, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimates.
- That is, an unbiased estimate does not naturally over or underestimate the parameter, it provides a “good” estimate.
- The sample mean is an example of an unbiased point estimate, as well as others we just listed.

ex 1)

practice

A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show (an American late-night TV show). The standard error of this estimate is 0.014. Estimate the 95% confidence interval for the proportion of college graduates who watch The Daily Show.

$\hat{p} = 0.33$

$SE = 0.014$

$\hat{p} \pm z^* SE$

$0.33 \pm 1.96 \times 0.014$

0.33 ± 0.027

$(0.303, 0.357)$

ex 2)

practice

The 3rd NHANES collected body fat percentage (BF%) and gender data from 13,601 subjects ages 20 to 80. The average BF% for the 6,580 men in the sample was 23.9, and this value was 35.0 for the 7,021 women. The standard error for the difference between the average male and female BF%s was 0.114. Do these data provide convincing evidence that men and women have different average BF%s. You may assume that the distribution of the point estimate is nearly normal.

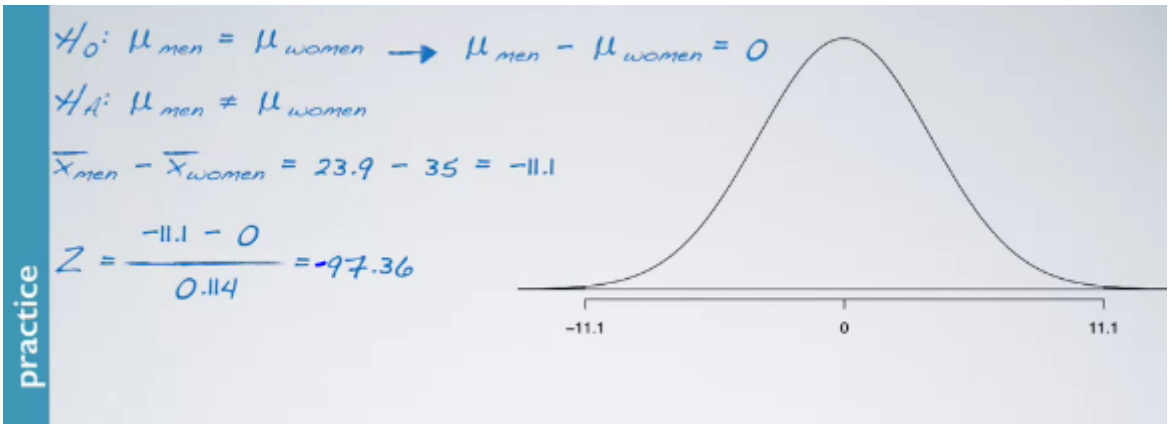
1. Set the hypotheses

$H_0: \mu_{men} = \mu_{women}$ $H_A: \mu_{men} \neq \mu_{women}$

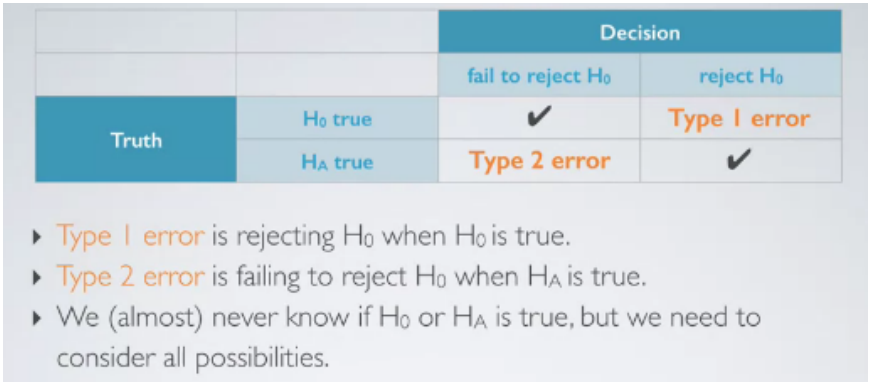
2. Calculate the point estimate

$\bar{x}_{men} - \bar{x}_{women} = 23.9 - 35 = -11.1$

3. Check conditions

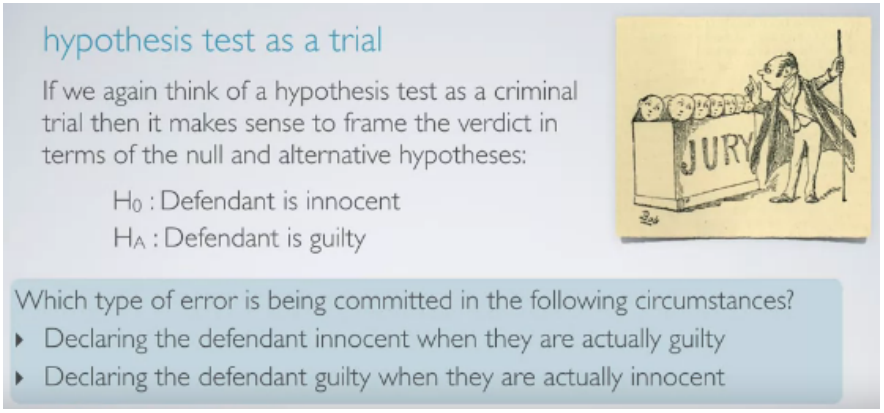


2. Decision errors



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✔	Type I error
	H_A true	Type 2 error	✔

- Type I error is rejecting H_0 when H_0 is true.
- Type 2 error is failing to reject H_0 when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.



- Type 2 error and Type 1 error.

- **Type 1 error rate**
- We reject H0 when the p-value is less than 0.05 ($\alpha = 0.05$)
- This means that, for those cases where H0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about a 5% change of making a Type 1 error if the null hypothesis is true.
- P(Type 1 error | H0 true) = α
- This is why we prefer small values of α - increasing α increases the Type 1 error rate.
- **Choosing α**
- If Type 1 error is dangerous or especially costly, choose a small significance level (e.g. 0.01)
  -> Goal : We want to be very cautious about rejecting H_0, so we demand very strong evidence favoring H_A before we would do so.

- If Type 2 error is dangerous or especially costly, choose a small significance level (e.g. 0.1)

-> Goal : We want to be cautious about failing to reject H_0 when the null is actually false.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level).
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$

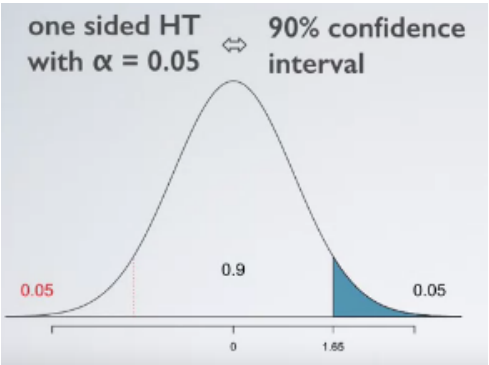
- Type 2 error rate
- If the alternative hypothesis is actually true, what is the chance that we make a Type 2 error, i.e. we fail to reject the null hypothesis even when we should reject it?
- The answer is not obvious.
- If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null value, it will be easier to detect a difference.
- Clearly, β depends on the effect size(δ), difference between point estimate and null value.

3. Significance vs. Confidence level

- Two-sided hypothesis test
- alpha of 0.05 : At each tail, you can afford to have about 0.025. Confidence interval is 95%.

- One-sided hypothesis test
- alpha of 0.05 : Confidence level is actually going to 90% because we're allowing for the 5% at the one end and other 5% at the other end.

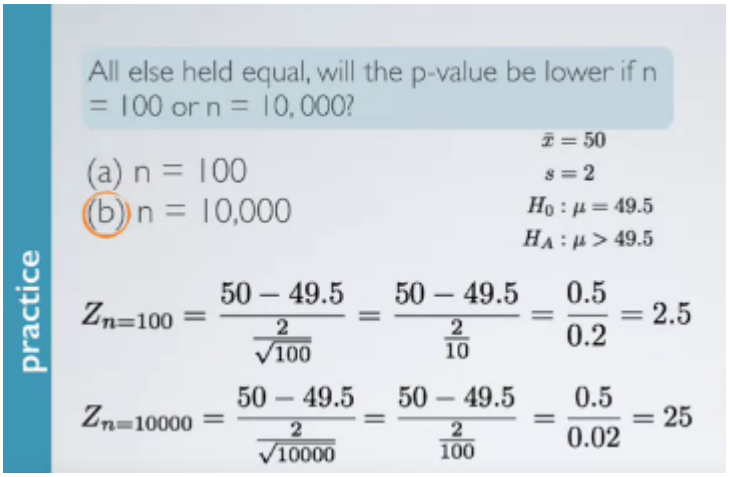
-



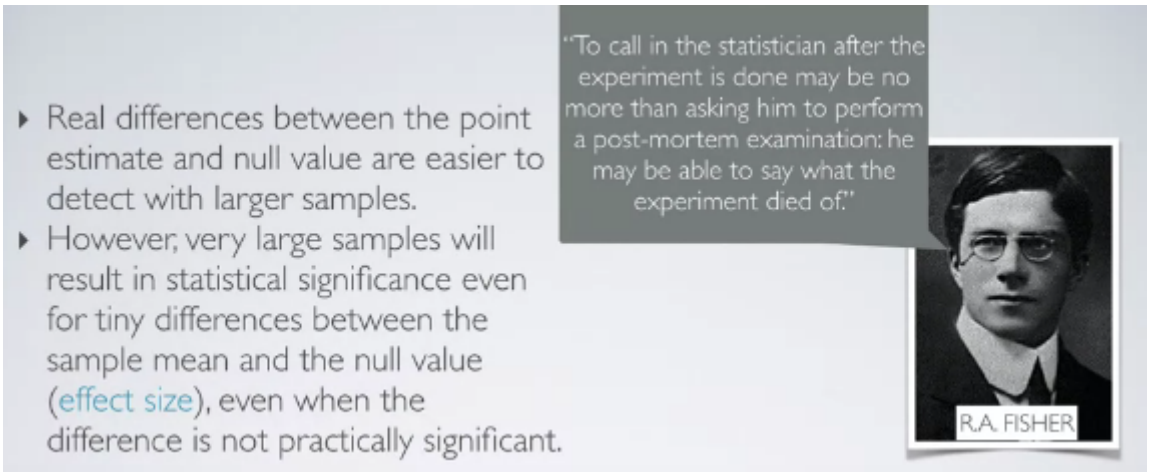
agreement of CI and HT

- A two sided hypothesis with threshold of α is equivalent to a confidence interval with CL = $1 - \alpha$.
- A one sided hypothesis with threshold of α is equivalent to a confidence interval with CL = $1 - (2 \times \alpha)$.
- If H_0 is rejected, a confidence interval that agrees with the result of the hypothesis test should not include the null value.
- If H_0 is failed to be rejected, a confidence interval that agrees with the result of the hypothesis test should include the null value.

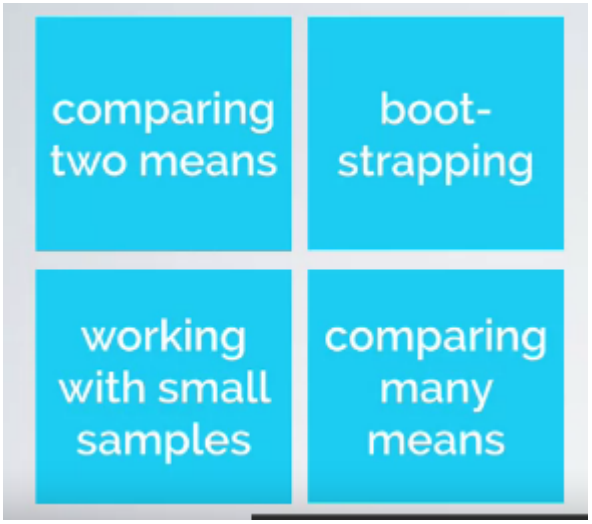
4. Statistical vs. Practical significance



- When sample size goes up, p-value is lower.



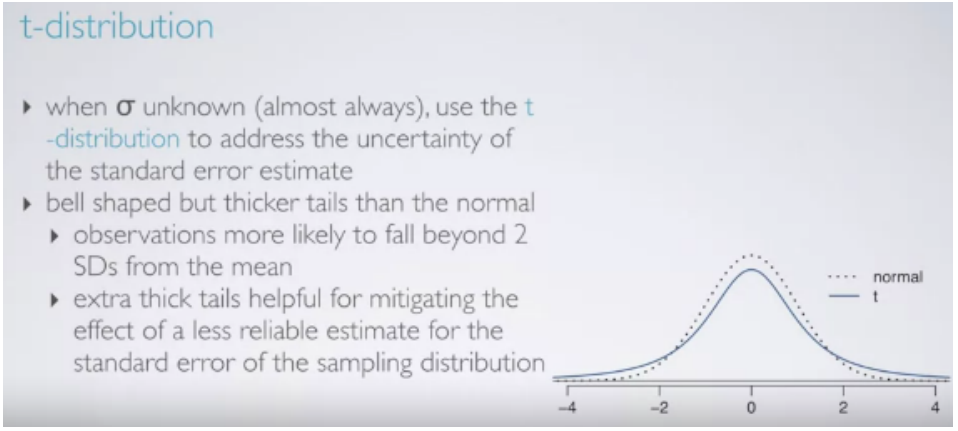
Week3



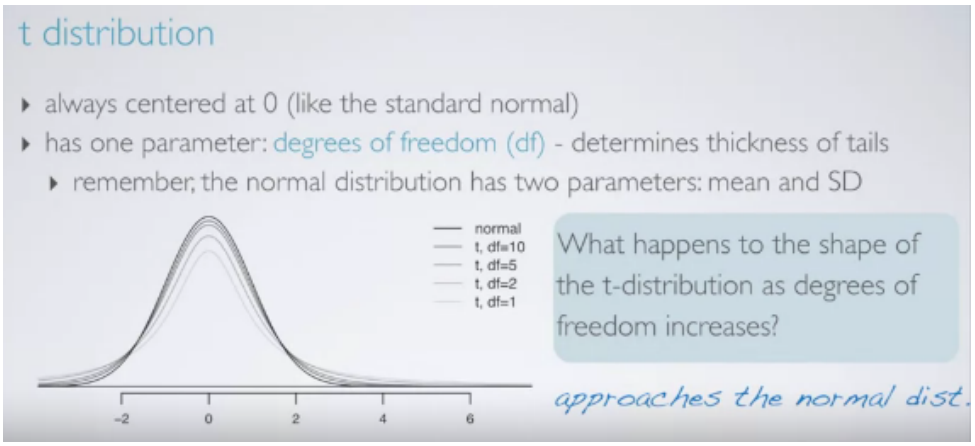
[t-distribution and Comparing Two means]

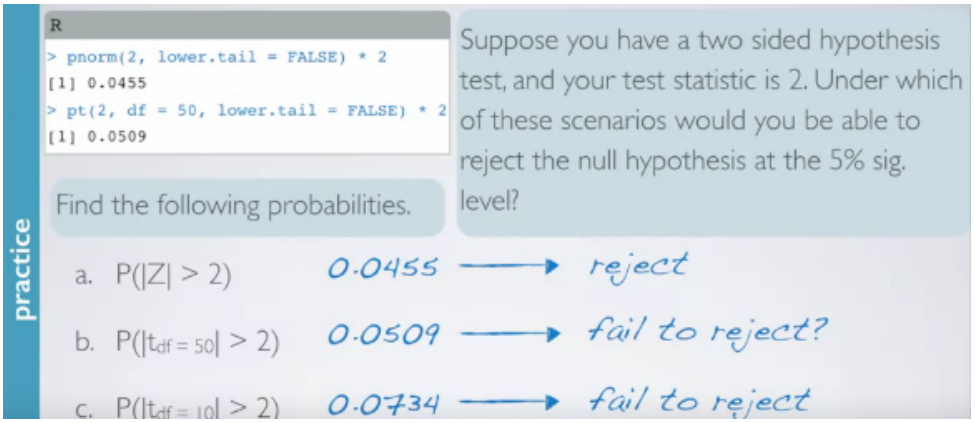
1. t-distribution

- t-distribution is useful for describing the distribution of the sample mean **when the population standard deviation, sigma is unknown.**
- Remember! What purpose does a large sample serve?
    - As long as observations are independent, and the population distribution is not extremely skewed,
    - The sampling distribution of the mean is nearly normal
    - The estimate of the standard error is reliable : s / sqrt(n)



- It is always centered at 0.
- It has one parameter : degrees of freedom(df) -> determines thickness of tails
    - remember, the normal distribution has two parameters: mean and SD
- As degrees of freedom increase, it approaches the normal distribution.

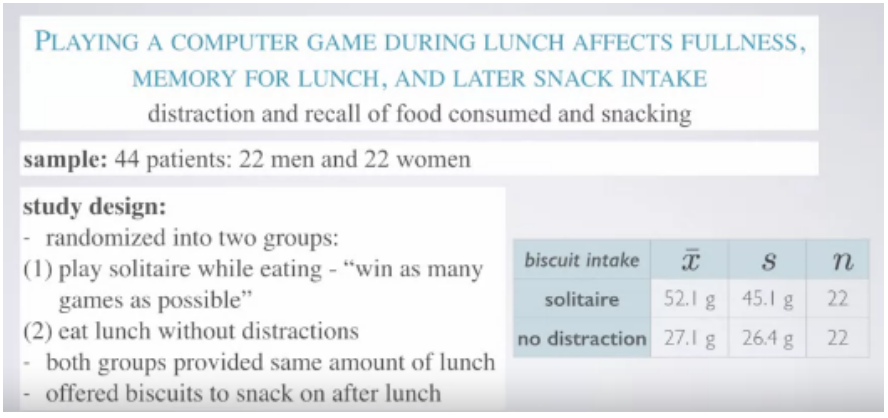


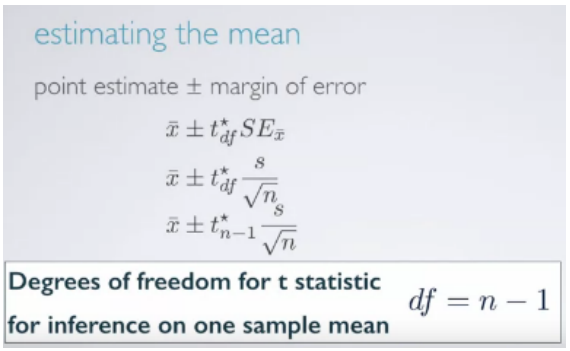


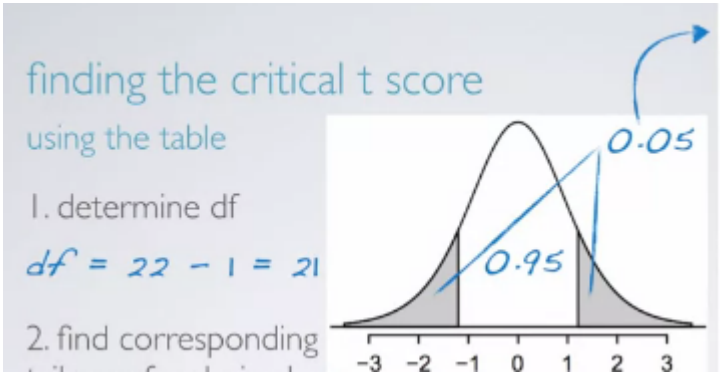
- We get more conservative with a t distribution with lower degrees.
- We also become less likely to be able to reject the null hypothesis.
- Generally degrees of freedom are tied to sample size. If your sample size is low, it's not easy to reject the null hypothesis and stronger evidence is needed to be able to do so.

2. Inference for a mean

study case)

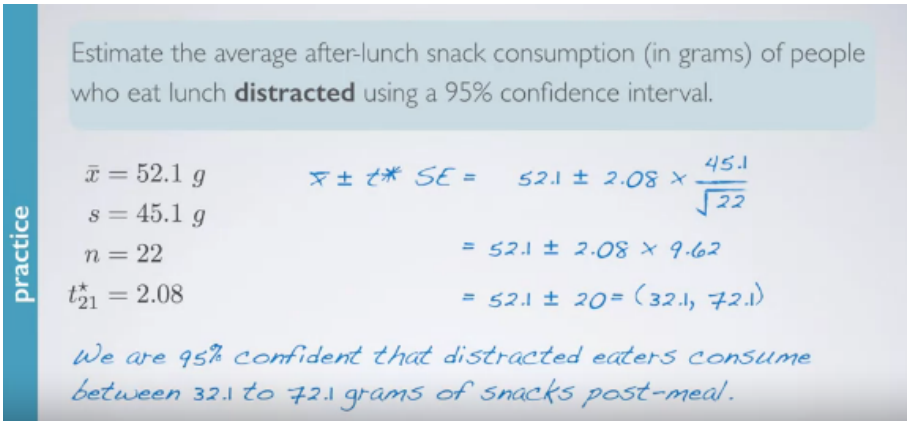


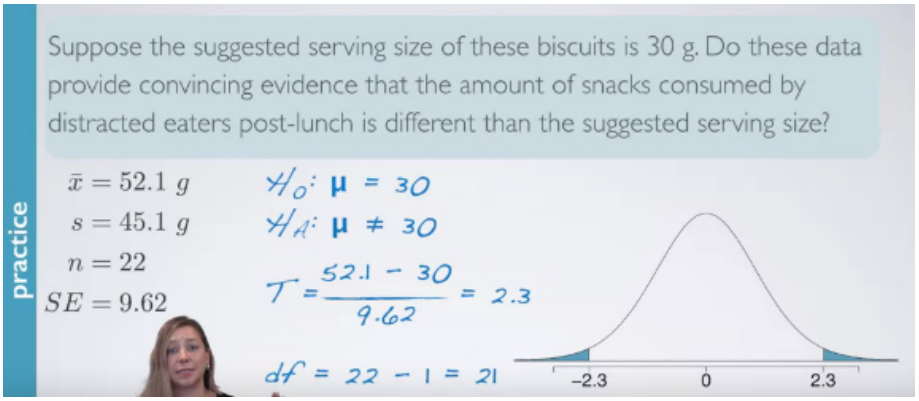




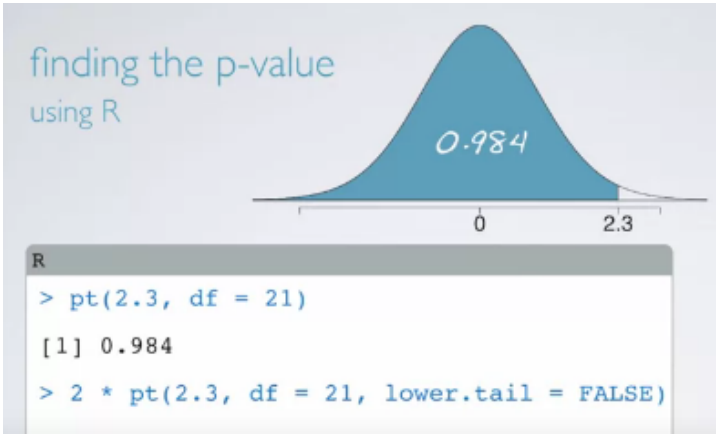
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79

- Estimate the average with 95% confidence interval

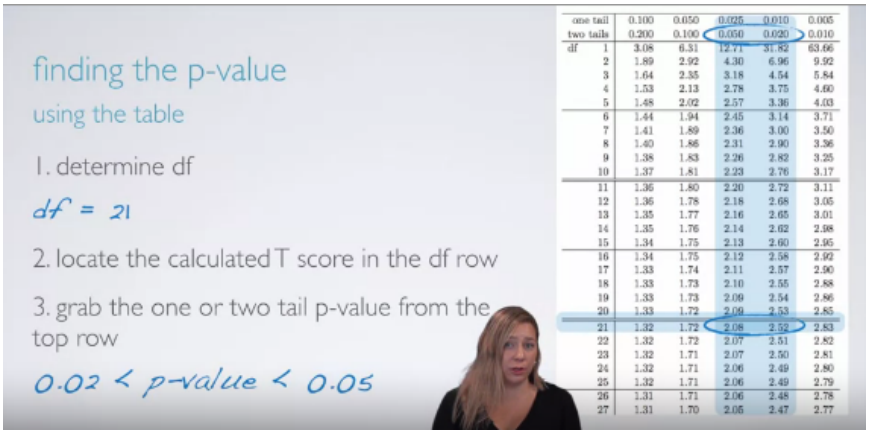




- Finding the p-value (1) using R

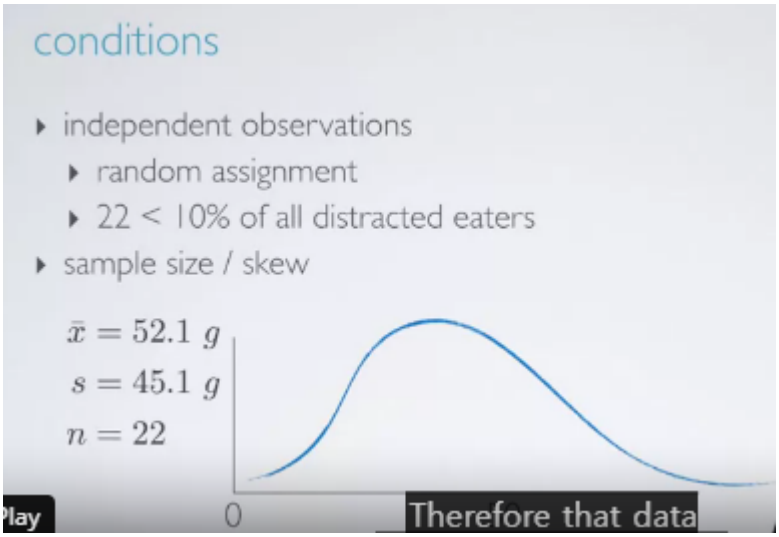


- Finding the p-value (2) using the table



: 95% confidence interval (32.1g, 72.1g)

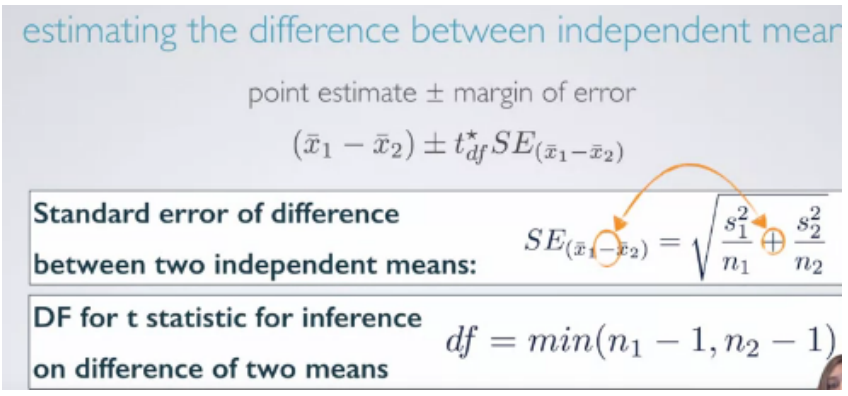
- Reject H0
- The confidence interval does not contain the value of 30.



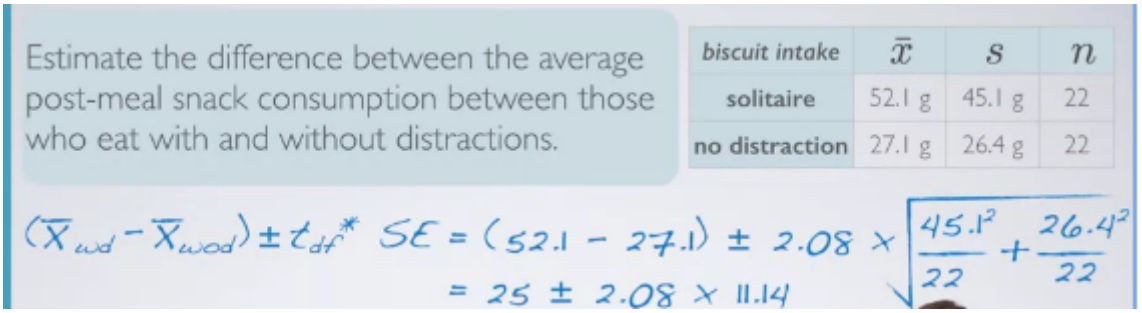
- The sample mean is 52 and there's a natural boundary at 0 since one cannot eat less than 0 g of biscuit. The 68, 95, 99.7 rule doesn't apply here.
- Therefore that data is likely right skipped.

3. Inference for comparing two independent means

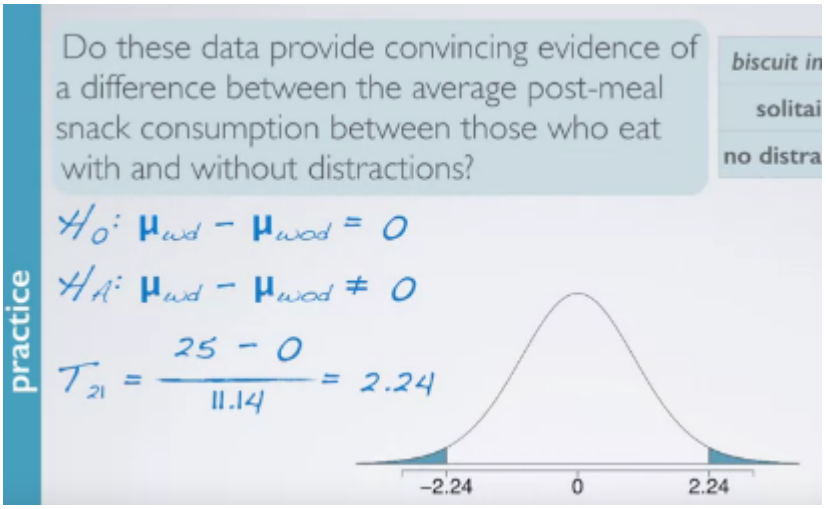
- Suppose we want to estimate how much more, or less, distracted eaters snack compared to non-distracted eaters.
- **The point estimate : the difference between the two sample average**

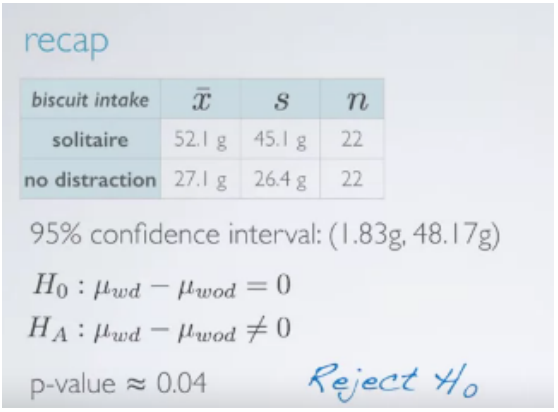


- Conditions for inference for comparing two independent means:
    1. Independence:
    - Within groups : sampled observations must be independent
        - random sample/assignment
        - if sampling without replacement, n < 10% population
    - Between groups : the two groups must be independent of each other(non-paired)
    2. Sample size/skew : The more skew in the population distributions, the higher the sample size needed



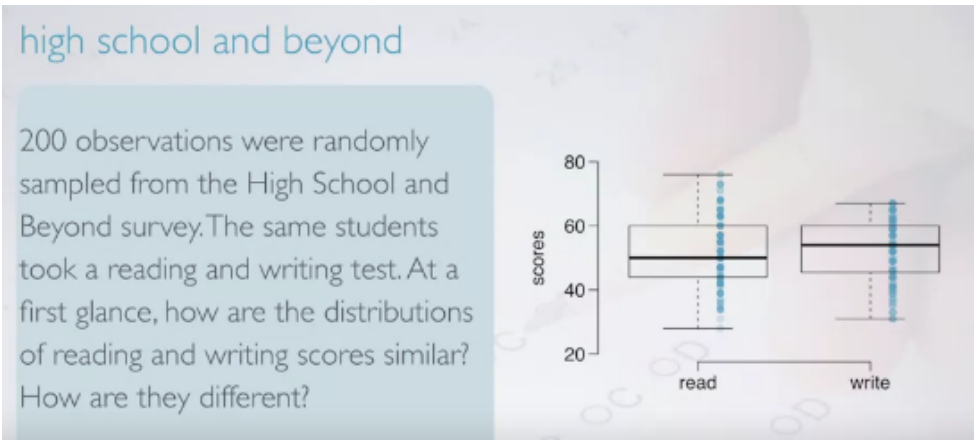
- (1.83, 48.17)
- We are 95% confident that those who eat with distractions consume 1.83 g and 48.17 g more snacks than those who eat without distractions, on average.





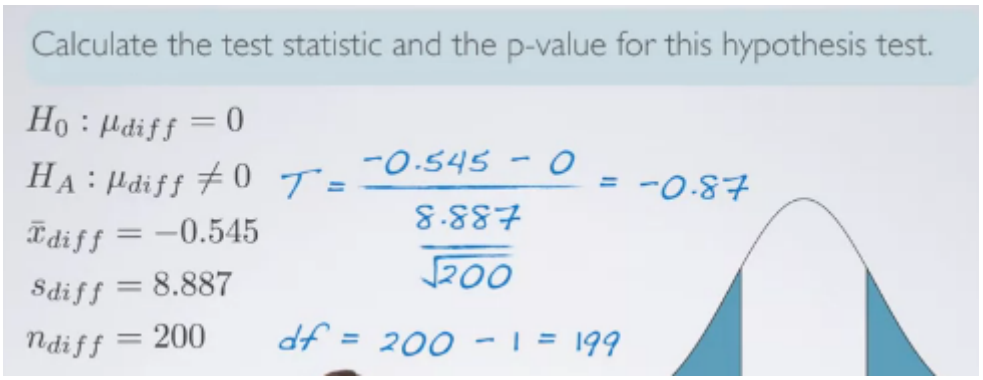
- The confidence interval for the average difference was 1.83 to 48.17 and the hypothesis test evaluating a difference between the two means yielded a p-value of roughly 4%. Which means that we would reject the null hypothesis and conclude that these data do indeed provide convincing evidence that there is a difference between the average snack intake of distracted and non-distracted eaters.

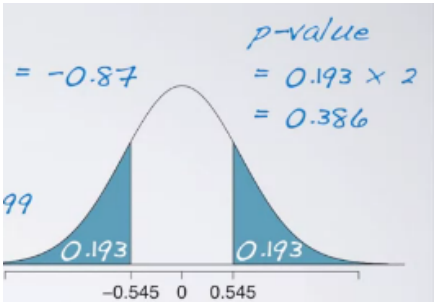
4. Inference for comparing two paired means (dependent)



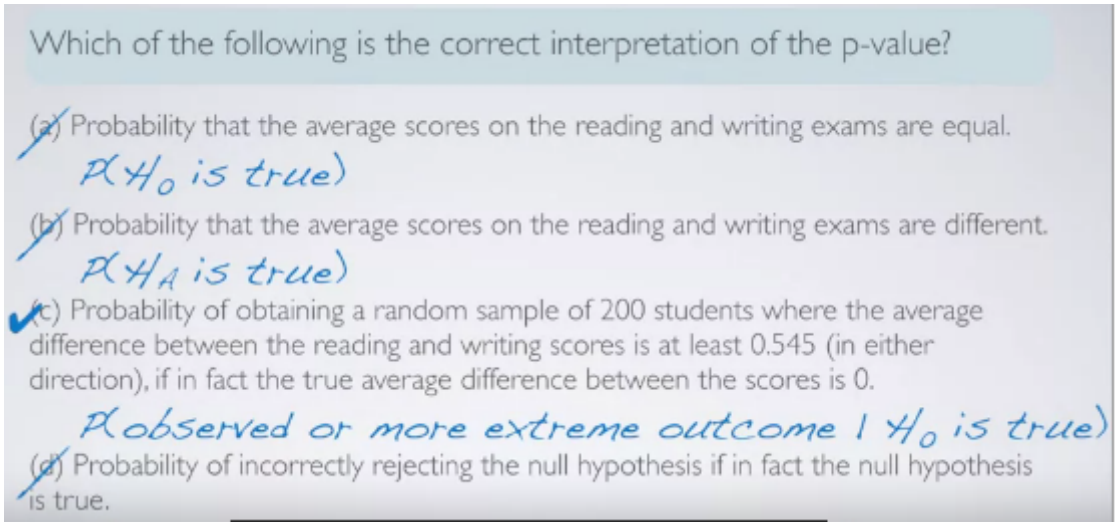
- Analyzing paired data
  - When two sets of observations have this special correspondence(not independent), they are said to be paired.
  - To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations : diff = read - write
  - **parameter of interest** : Average difference between the reading and writing scores of all high school students
  - **point estimate** : Average difference between the reading and writing scores of sampled high school students

- Hypothesis for paired means
  - H_0 : $\mu diff = 0$; There is no difference between the average reading and writing scores.
  - H_0 : $\mu diff \neq 0$; There is a difference between the average reading and writing scores.





- The interpretation of the p-value

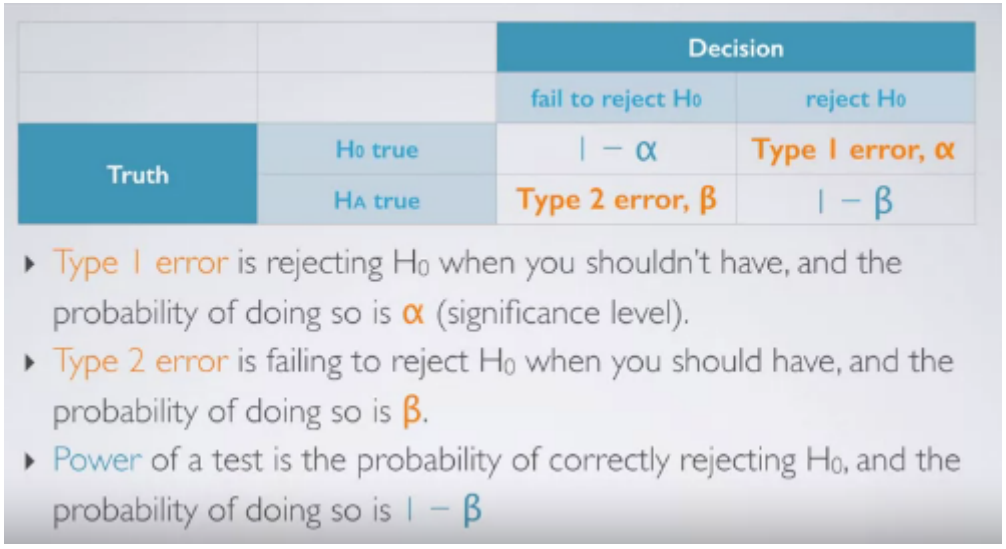


- Summary
  - paired data (2 vars.) -> differences (1 var.)

- most often : H0 : = 0
- same individuals: pre-post studies , repeated measures, etc.
- different (but dependent) individuals:twins, partners, etc.

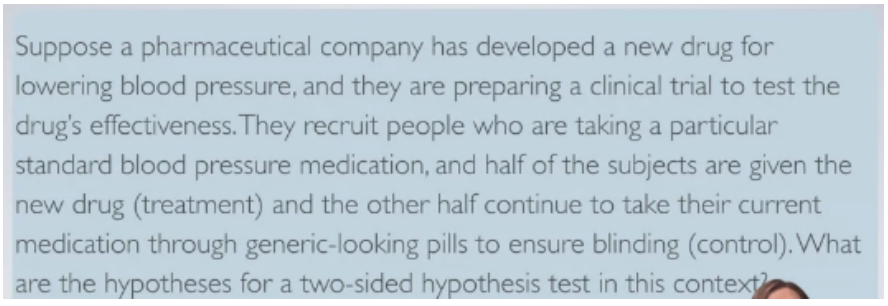
5. Power

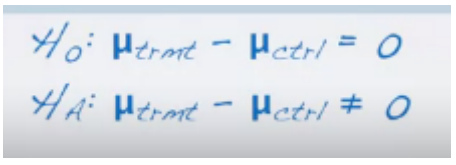
- Oftentimes, in experiment planning, there are two competing considerations.
    1. We want to collect enough data to detect important effects but it can be expensive.
    2. In experiments involving people, there may be some risk to patients.
- We will find the required sample size that will result in a test with 80% power. It seems arbitrary, but it is commonly required power for most experiments.

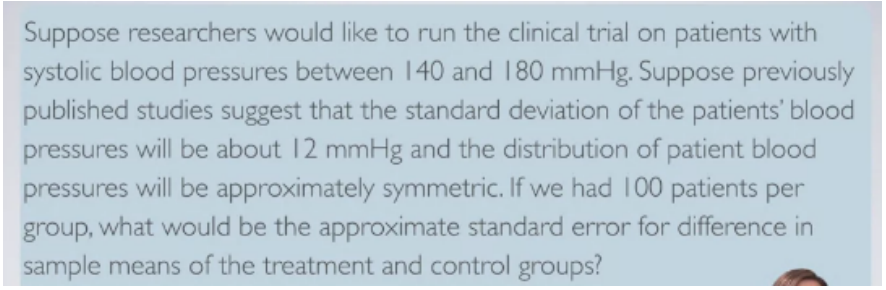


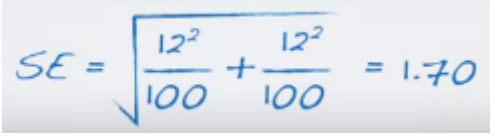
- Power of a test is the probability of correctly rejecting H0, and the probability of doing so is 1-β.
- Therefore, keeping the Type2 error rate low increases the power, which is a desirable outcome.
- goal : keep α and β low
- but decreasing one increases the other.
    - One solution for this is getting a larger sample size.
    -

ex)



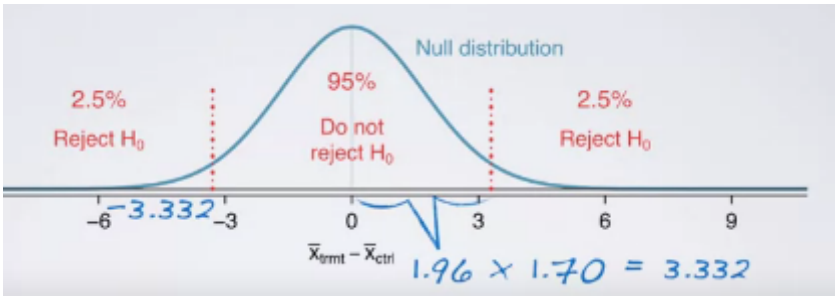






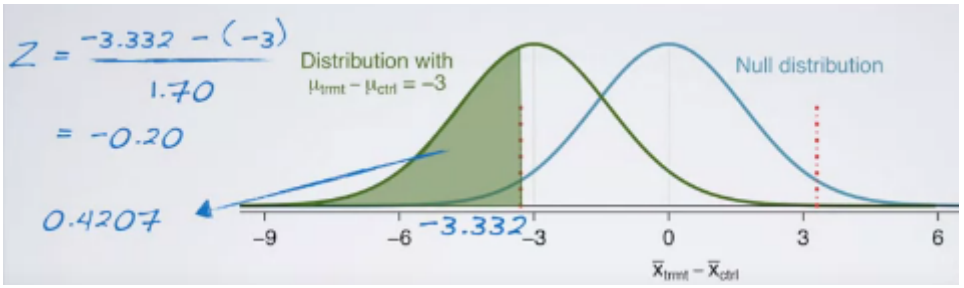
For what values of the difference between the observed averages of blood pressure in treatment and control groups (effect size) would we reject the null hypothesis at the 5% significance level?

↓



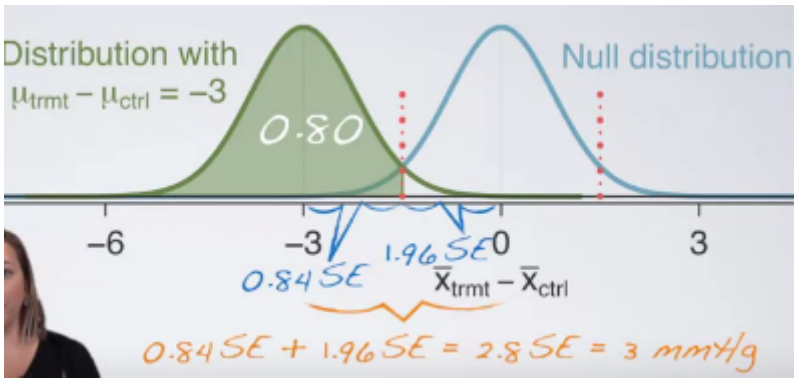
Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. What is the power of the test that can detect this effect?

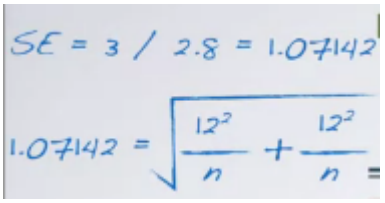
↓



- In other words, 3 mm of mercury is the minimum effect size of interest and we want to know how likely we are to detect this size o f an effect in this study.
- It means the observed distribution of differences in average blood pressures between the two groups will be shifted from the null by 3 mm of mercury, as shown in the plot.
- The shaded green area is approximately 0.4207. **Therefore, the power of the test is about 42% when the effect size is -3 and each group has a sample size of 100.**

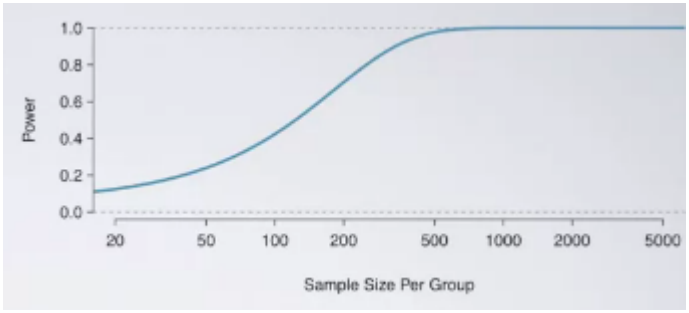
● What sample size will lead to a power of 80% for this test?





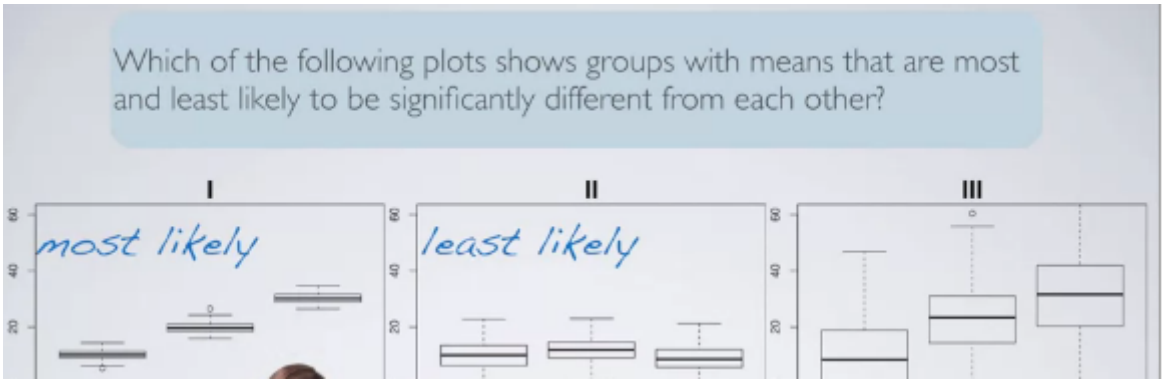
- n = 250.88

● Summary
  - calculate required sample size for a desired level of power
  - calculate power for a range of sample sizes, and choose target power

- 

[ANOVA and Bootstrapping]

1. Comparing more than two means



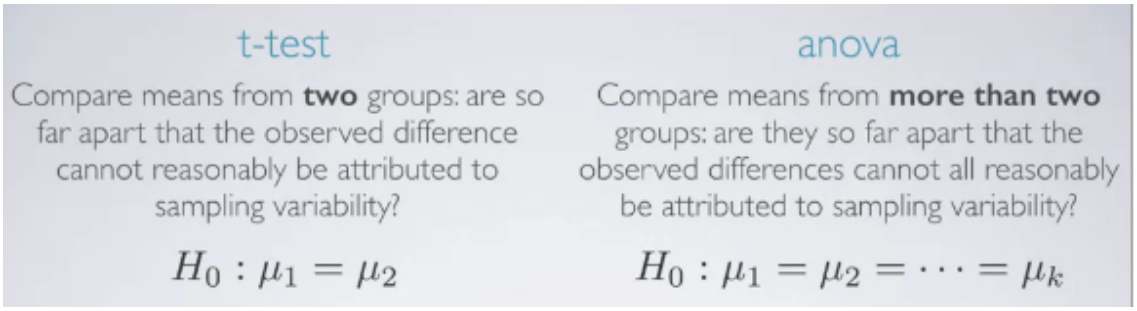
- Our goal : **to find out if there's a difference between the average vocabulary scores of Americans from the different classes.**
  - Compare means of 2 groups using a T statistic.
  - Compare means of 3_ groups using a new test called analysis of variance (ANOVA) and a new statistic called F.
- ANOVA (3개 이상 그룹 비교할 때 사용하는 가설검정방법)
  - H_0 : The mean outcome is the same across all categories

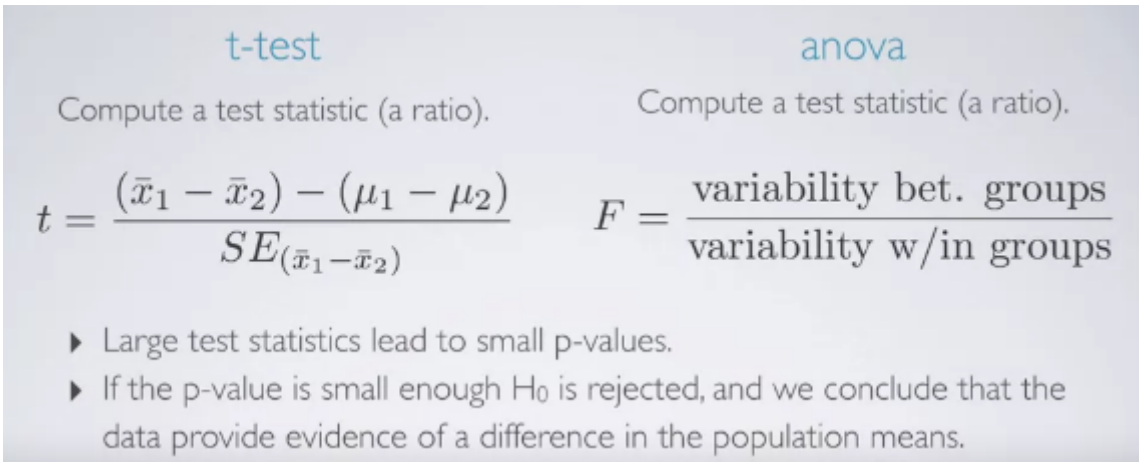
$$\mu_1 = \mu_2 = \cdots = \mu_k$$

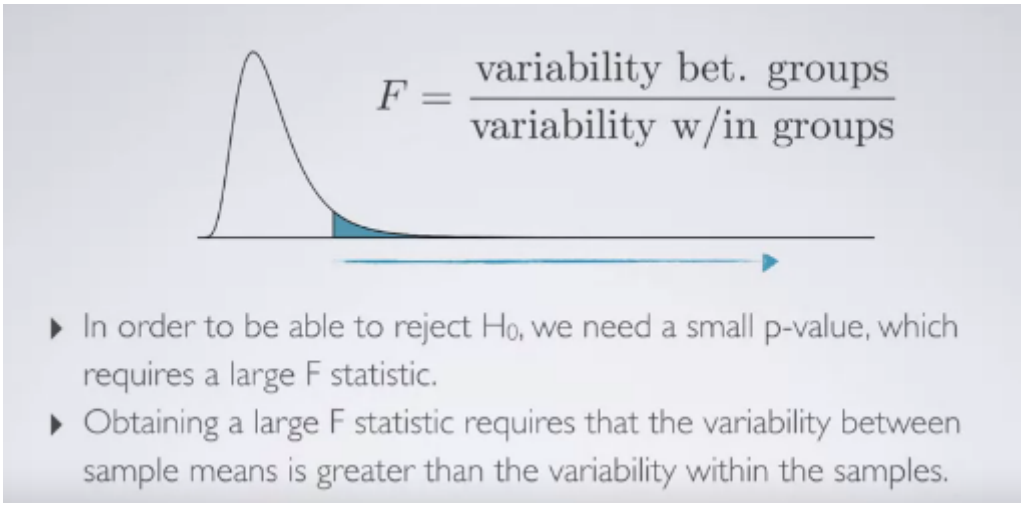
μ_i : mean of the outcome for observations in category i

k : number of groups

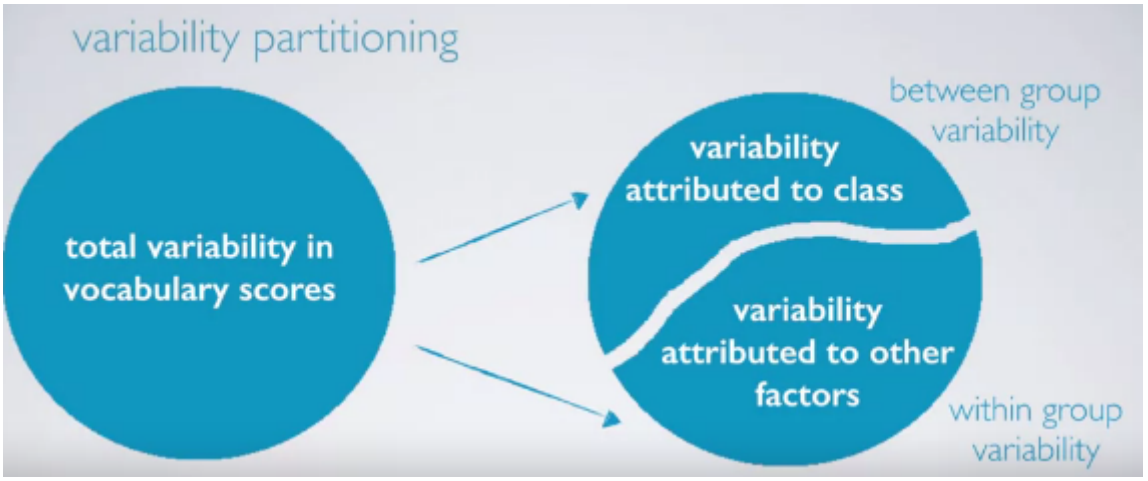
  - H_A : At least one pair of means are different from each other

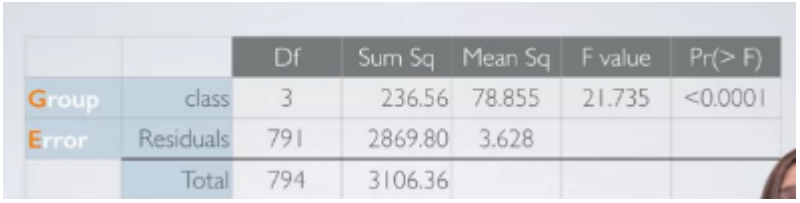






2. ANOVA





		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- The first row (group row) : about between group variability
- the second row (error row) : within group variability

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.80			
	Total		3106.36			

- SST (sum of squares total)
  - measures the total variability in the response variable
  - calculated very similarly to variance (except not scaled by the sample size)

Sum of squares total (SST):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i : value of the response variable for each observation

$\bar{y}$: grand mean of the response variable

- SSG (sum of squares groups)
  - measures the variability between groups
  - explained variability : squared deviation of group means from overall mean, weighted sample size

Sum of squares group (SSG):

$$SSG = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

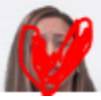
n_j : number of observations in group j

$\bar{y}_j$: mean of the response variable for group j

$\bar{y}$: grand mean of the response variable

	n	mean	sd
lower class	41	5.07	2.24
working class	407	5.75	1.87
middle class	331	6.76	1.89
upper class	16	6.19	2.34
overall	795	6.14	1.98

$$SSG = (41 \times (5.07 - 6.14)^2) + (407 \times (5.75 - 6.14)^2) + (331 \times (6.76 - 6.14)^2) + (16 \times (6.19 - 6.14)^2) \approx 236.56$$



- SSE (sum of squares error)
- measures the variability within groups
- unexplained variability : unexplained by the group variable, due to other reasons

Sum of squares error (SSE):

$$SSE = SST - SSG$$

$3106.36 - 236.56 = 2869.8$

- MSE (mean square error)

mean square error

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855		
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

Mean squares: Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

- group: $MSG = SSG/df_G$ → 236.56 / 3 ≈ 78.855
- error: $MSE = SSE/df_E$ → 2869.8 / 791 ≈ 3.628

- F statistic

F statistic

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

F statistic: Ratio of the average between group and within group variabilities:

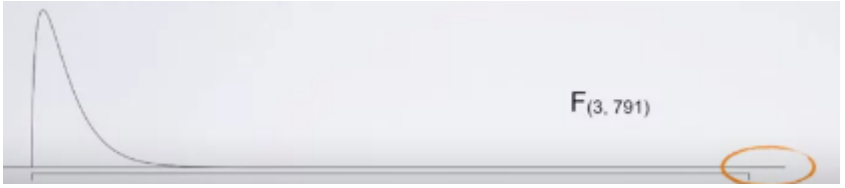
$$F = \frac{MSG}{MSE} \longrightarrow \frac{78.855}{3.628} \approx 21.735$$

- p-value

p-value

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class	3	236.56	78.855	21.735	<0.0001
Error	Residuals	791	2869.80	3.628		
	Total	794	3106.36			

- p-value is the probability of at least as large a ratio between and within group variabilities if in fact the means of all groups are equal.
- area under the F curve, with degrees of freedom dfG and dfE, above the observed F statistic.



- The F statistic is always positive, a more extreme statistic will always be more extreme in the positive direction.

- Conclusion

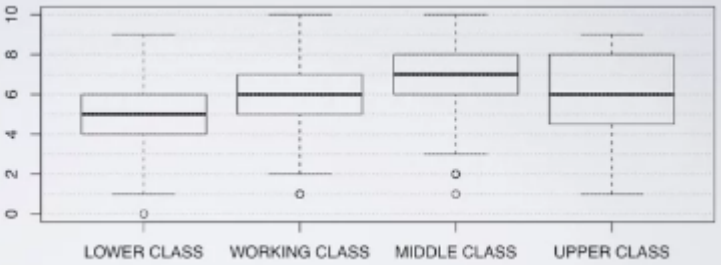
conclusion

- If p-value is small (less than α), reject H_0.
  - The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one).
- If p-value is large, fail to reject H_0.
  - The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

3. Conditions for ANOVA

1. Independence
- within groups : sampled observations must be independent
    - random sample/assignment
    - each nj less than 10% of respective population
    - always important, but sometimes difficult to check
- between groups : the groups must be independent of each other (non-paired)
    - carefully consider whether the groups may be dependent
    - repeated measures anova
2. Approximate normality
    - distributions should be nearly normal within each group
    - especially important when sample sizes are small
3. Equal variance : groups should have roughly equal variability
    - variability should be consistent across groups: homoscedastic groups
    - especially important when sample sizes differ between groups
    -

	n	sd
lower class	41	2.24
working class	407	1.87
middle class	331	1.89
upper class	16	2.34
overall	795	1.98



4. Multiple comparisons

- Which means are different?
    - two sample t tests for differences in each possible pair of groups
    - multiple tests -> inflated Type 1 error rate (H_0 true인데 기각하는 것)
    - **solution : use modified significance level**

- Multiple comparisons
    - testing many pairs of group
    - The Bonferroni correction suggests that a more stringent significance level is more appropriate for these tests.
        - Adjust α by the number of comparisons being considered
    -

Bonferroni correction:

$$\alpha^\star = \alpha/K \quad K\text{: number of comparisons, } K = \frac{k(k-1)}{2}$$

The social class variables has 4 levels. If $\alpha = 0.05$ for the original ANOVA, what should the modified significance level be for two sample t tests for determining which pairs of groups have significantly different means?

$k = 4$

$K = \frac{4 \times 3}{2} = 6$

$\alpha^* = 0.05 / 6 \approx 0.0083$

- pairwise comparisons
    - constant variance(등분산성) -> rethink standard error and degrees of freedom:
        - use consistent standard error and degrees of freedom for all tests

- compare the p-values from each test to the modified significance level

Standard error for multiple pairwise comparisons:

$$SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

indep. groups test:

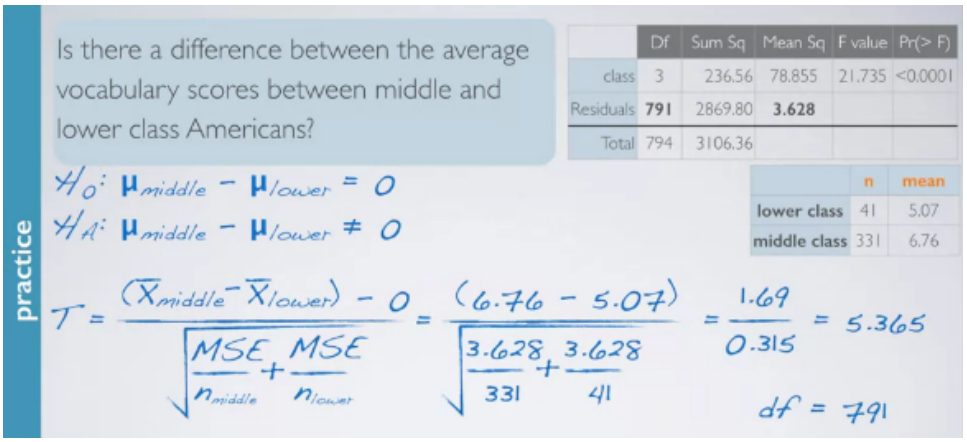
$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

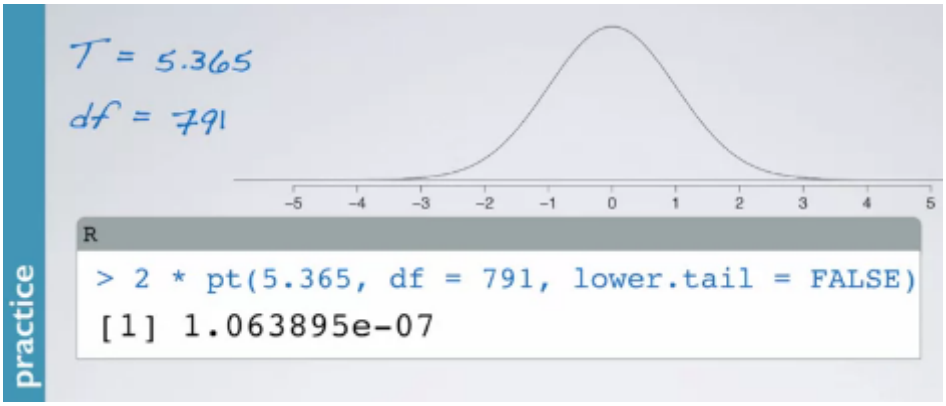
- MSE is actually the average within group variance. (평균제곱오차)

Degrees of freedom for multiple pairwise comparisons:

$$df = df_E$$

$$df = min(n_1 - 1, n_2 - 1)$$





- Controlling the Type 1 error rate

5. Bootstrapping

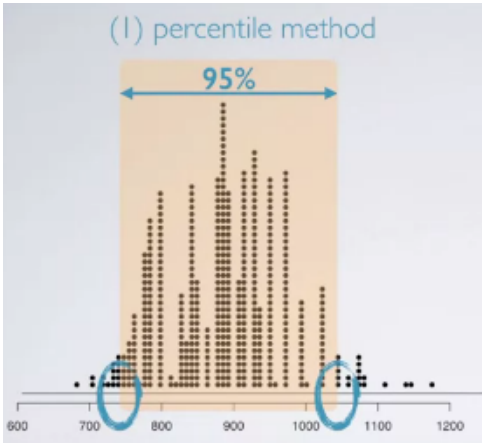
- When the distribution is skewed, using median is a better measure of the typical observation.
- We can't apply CLT so we need another method.

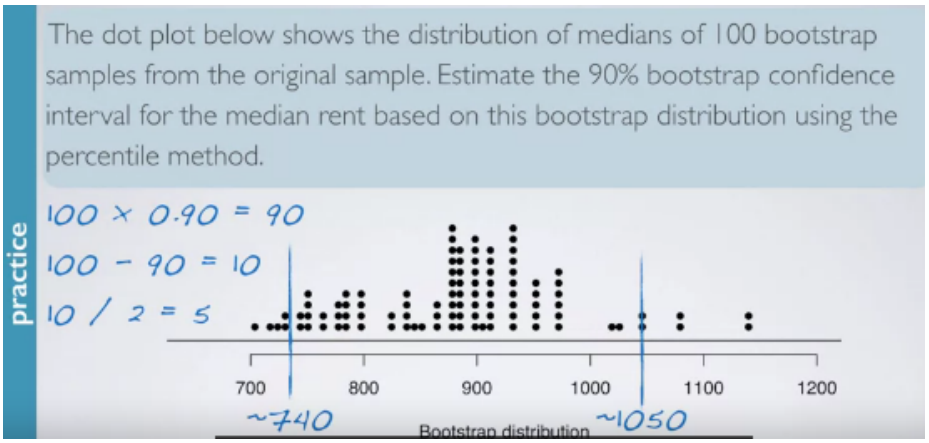
- Bootstrapping : a metaphor for accomplishing an impossible task without any outside help
- Bootstrapping scheme

1) take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
3) repeat steps 1 and 2 many times to create a bootstrap distribution - a distribution of bootstrap statistics

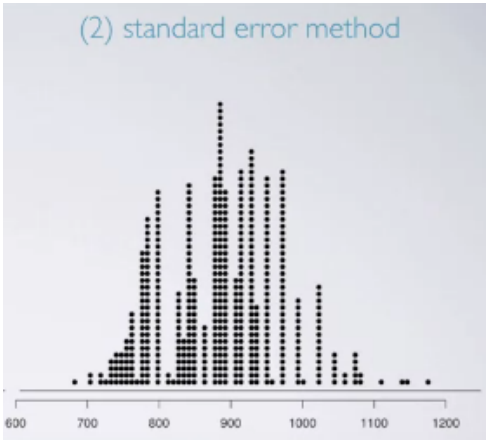
- Calculating a confidence level in two ways

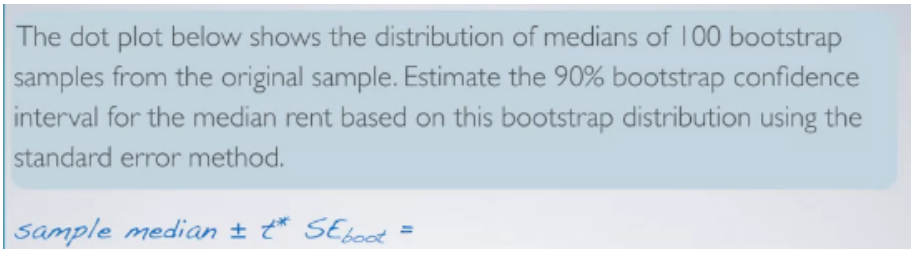
1) percentile method





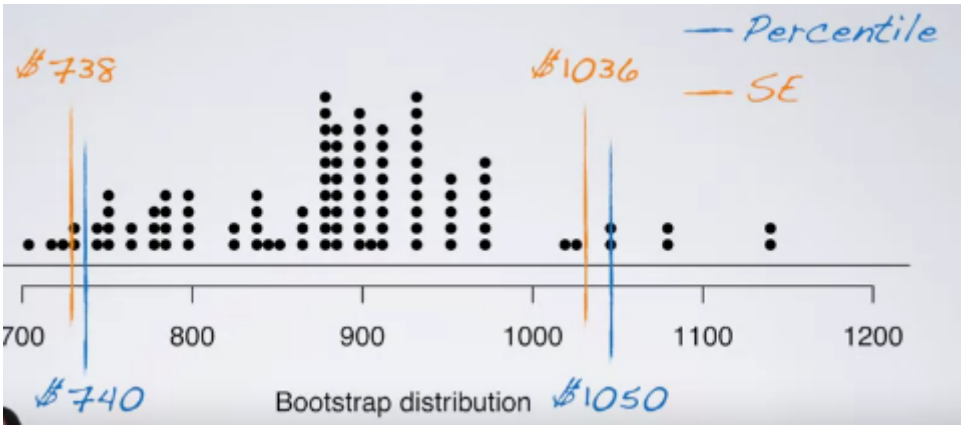
2) standard error method





= 887 ± 1.73*89.5758 = (732, 1042)

- Comparisons : percentile vs. SE methods



- It's pretty close to each other even though they're not exactly the same.

- bootstrapping limitation

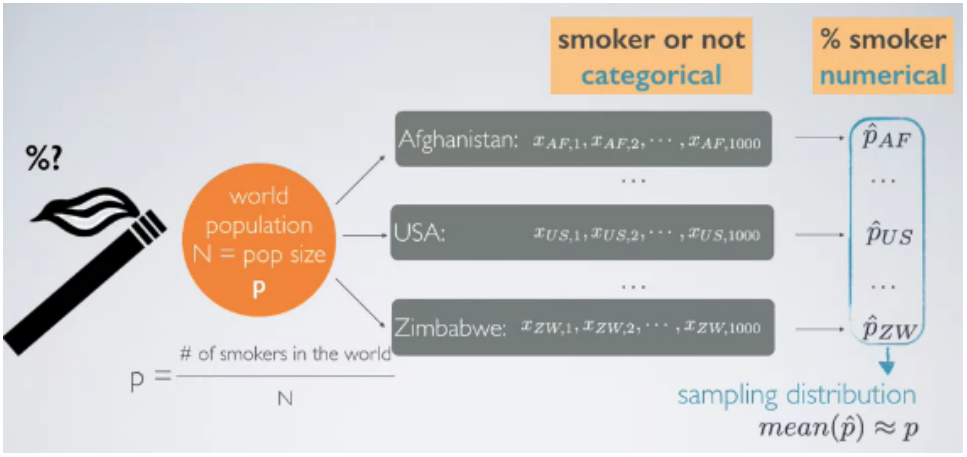
    - Not as rigid conditions as CLT based methods
    - If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable.
    - A representative sample is still required - if the sample is biased, the estimates resulting from this sample will also be biased.

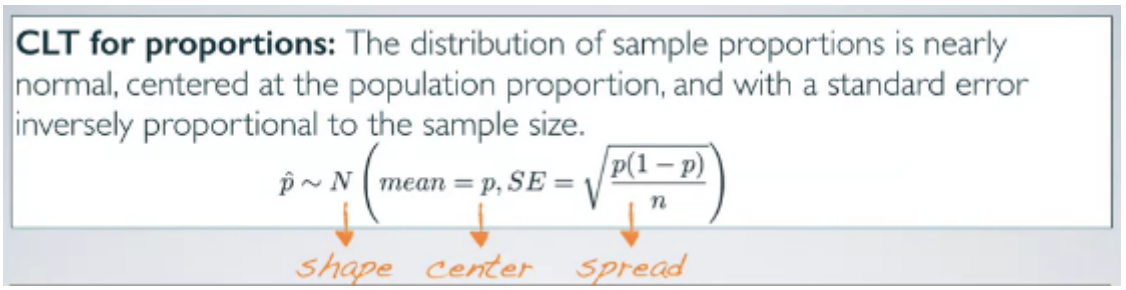
- Bootstrap vs. sampling distribution
    - Sampling distribution created using sampling (with replacement) from the population
    - Bootstrap distribution created using sampling (with replacement) from the sample
    - Both are distributions of sample statistics.

Week4

[Inference for proportions]

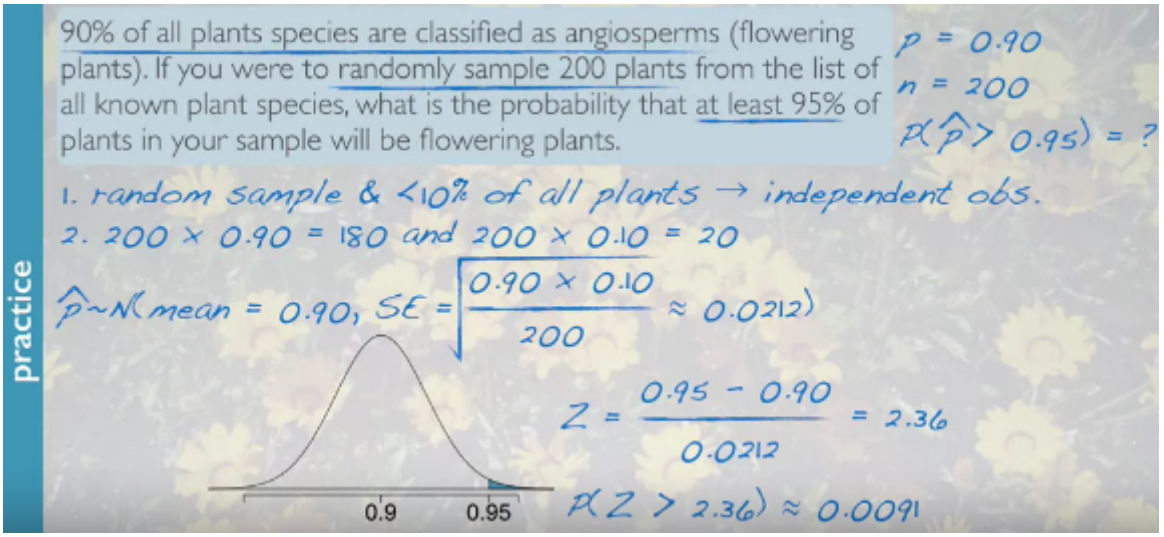
1. Sampling Variability and CLT for Proportions

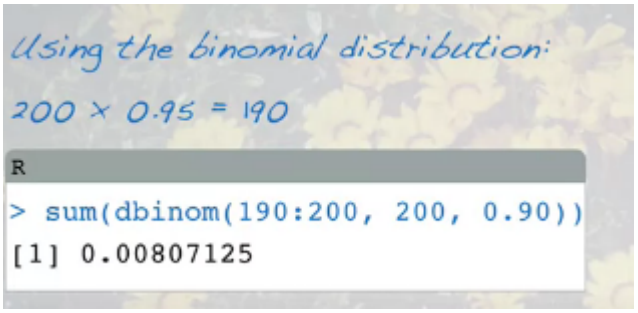




- Conditions for the CLT
  - Independence : Sampled observations must be independent
    - random sample/assignment
    - if sampling without replacement, n < 10% population
  - Sample size / skew : There should be at least 10 successes and 10 failure in the sample:
    - np >= 10 and n(1-p)>=10
    - if p is unknown, use $\hat{\rho}$

ex)

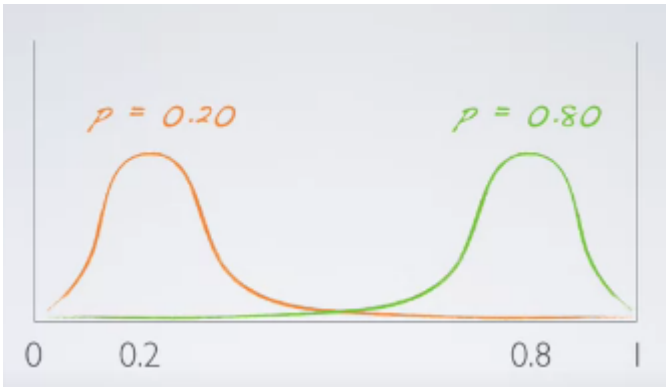




- If you were to randomly sample 200 plants from the list of all known plant species, would it be considered unusual if 87.5% of the plants in a random sample of 200 were angiosperms? (remember, 90% of all plants species are classified as a angiosperms)

-> No, it's usual within two standard errors. Z = (0.875-0.9) / 0.0212 = -1.18

- What if the success-failure condition is not met:
  - the center of the sampling distribution will still be around the true population proportion
  - the spread of the sampling distribution can still be approximated using the same formula for the standard error
  - the shape of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1
- shape of the sampling distribution



-> when the success-failure condition is not met.

- If the condition is met, it's going to yield a smaller SE and be looking more and more symmetric as the sample size increases.

2. Confidence Interval for a Proportion

case)

from the 2010 GSS

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

(a) All 1000 get the drug
(b) 500 get the drug, 500 don't

experimental design	
bad intuition	99
good intuition	571
total	670

What percent of Americans have good intuition about experimental design?

parameter of interest

Percentage of **all** Americans who have good intuition about experimental design.

p

point estimate

Percentage of **sampled** Americans who have good intuition about experimental design.

$\hat{p}$

estimating a proportion

point estimate ± margin of error

$$\hat{p} \pm z^\star SE_{\hat{p}}$$

Standard error for a proportion, for calculating a confidence interval: $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

The GSS found that 571 out of 670 (~85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

1. independence: 670 < 10% of Americans, and GSS samples randomly Whether one American in the sample has good intuition about experimental design is independent of another.

2. sample size / skew: 571 successes, 670 − 571 = 99 failures Since the success-failure condition is met, we can assume that the sampling distribution of the proportion is nearly normal.

$$\hat{p} \pm z^* SE = 0.85 \pm 1.96\sqrt{\frac{0.85 \times 0.15}{670}}$$

$$= 0.85 \pm 1.96 \times 0.0138$$

$$= 0.85 \pm 0.027$$

$$= (0.823, 0.877)$$

We are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design.

The margin of error for the previous confidence interval was 2.7%. If, for a new confidence interval based on a new sample, we wanted to reduce the margin of error to 1% while keeping the confidence level the same, at least how many respondents should we sample?

$$ME = 0.01 = 1.96\sqrt{\frac{0.85 \times 0.15}{n}}$$

$$0.01^2 = \frac{1.96^2 \times 0.85 \times 0.15}{n}$$

$$n = \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} = 4898.04 \rightarrow \text{at least } 4899$$

calculating the required sample size for desired ME

remember $ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- if there is a previous study that we can rely on for the value of $\hat{p}$ use that in the calculation of the required sample size
- if not, use $\hat{p} = 0.5$
  - if you don't know any better, 50-50 is a good guess
  - gives the most conservative estimate – highest possible sample size

3. Hypothesis Test for a Proportion

Hypothesis testing for a single proportion:

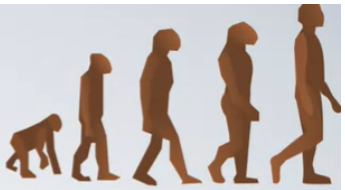
1. Set the hypotheses: $H_0 : p = null\ value$
   $H_A : p <\ or\ >\ or\ \neq\ null\ value$
2. Calculate the point estimate: $\hat{p}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** $np \geq 10$ and $n(1-p) \geq 10$
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value < α, reject H_0; the data provide convincing evidence for H_A.
   - If p-value > α, fail to reject H_0 the data *do not* provide convincing evidence for H_A.

$\hat{p}$ vs. p

	confidence interval	hypothesis test
success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

practice)

A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?



$H_0: p = 0.5$

$H_A: p > 0.5$

$\hat{p} = 0.6$

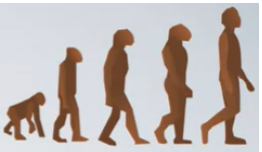
$n = 1983$

1. independence: 1983 < 10% of Americans & random sample
   Whether one American in the sample believes in evolution is independent of another.
2. sample size / skew: $1983 \times 0.5 = 991.5 > 10$

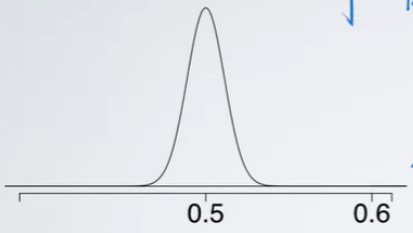
S-F condition met → nearly normal sampling distribution

$H_0: p = 0.5$ $\quad \hat{p} = 0.6$

$H_A: p > 0.5$ $\quad n = 1983$



$\hat{p} \sim N(mean = 0.5, SE = \sqrt{\frac{0.5 \times 0.5}{1983}} \approx 0.0112)$



$Z = \frac{0.6 - 0.5}{0.0112} \approx 8.92$

$p\text{-value} = P(Z > 8.92)$

$= almost\ 0$

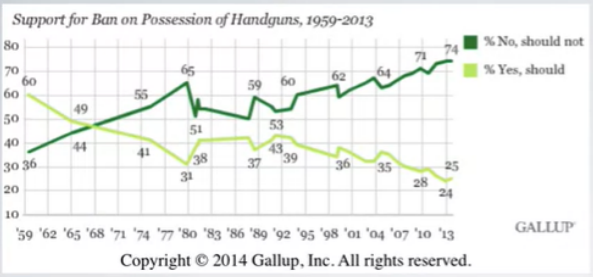
4. Estimating the Difference Between Two Proportions

practice)

In early October 2013, a Gallup poll asked "Do you think there should or should not be a law that would ban the possession of handguns, except by the police and other authorized persons?"

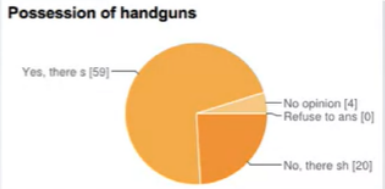
(a) No, there should not be such a law
(b) Yes, there should be such a law
(c) No opinion

US



Coursera

No, there should not be such a law	20	24%
Yes, there should be such a law	59	71%
No opinion	4	5%



	$suc.$	n	$\hat{p}$
US	257	1028	0.25
Coursera	59	83	0.71

How do Coursera students and the American public at large compare with respect to their views on laws banning possession of handguns?

parameter of interest

Difference between the proportions of **all** Coursera students and **all** Americans who believe there should be a ban on possession of handguns.

$$p_{Coursera} - p_{US}$$

point estimate

Difference between the proportions of **sampled** Coursera students and **sampled** Americans who believe there should be a ban on possession of handguns.

$$\hat{p}_{Coursera} - \hat{p}_{US}$$

estimating the difference between two proportions

point estimate ± margin of error

$$(\hat{p}_1 - \hat{p}_2) \pm z^\star SE_{(\hat{p}_1 - \hat{p}_2)}$$

Standard error for difference between two proportions, for calculating a confidence interval:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions for inference for comparing two independent proportions:

1. ***Independence:***
   - ✓ **within groups:** sampled observations must be independent within each group
     - random sample/assignment
     - if sampling without replacement, n < 10% of population
   - ✓ **between groups:** the two groups must be independent of each other (non-paired)
2. ***Sample size/skew:*** Each sample should meet the success-failure condition:
   - ✓ $n_1p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
   - ✓ $n_2p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Using a 95% confidence interval, estimate how Coursera students and the American public at large compare with respect to their views on laws banning possession of handguns.

	suc.	n	$\hat{p}$
US	257	1028	0.25
Coursera	59	83	0.71

1. *independence:* ✓ random sample: yes for US, no for Coursera
   ✓ 10% condition: met for both

   Sampled Americans independent of each other, sampled Courserians may not be.
2. *sample size / skew:* ✓ US: 257 successes, 1028 − 257 = 771 failures
   ✓ Coursera: 59 successes, 83 − 59 = 24 failures

We can assume that the sampling distribution of the difference between two proportions is nearly normal.

	suc.	n	$\hat{p}$
US	257	1028	0.25
Coursera	59	83	0.71

$$(\hat{p}_{Coursera} - \hat{p}_{US}) \pm z^* SE =$$

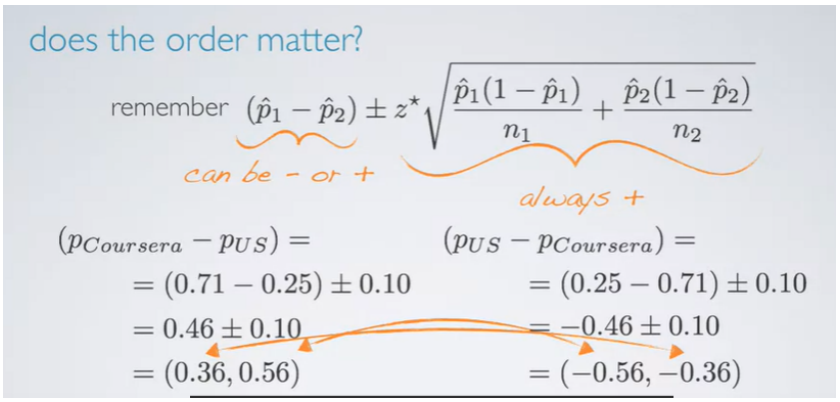
$$= (0.71 - 0.25) \pm 1.96 \sqrt{\frac{0.71 \times 0.29}{83} + \frac{0.25 \times 0.75}{1028}}$$

$$= 0.46 \pm 1.96 \times 0.0516$$

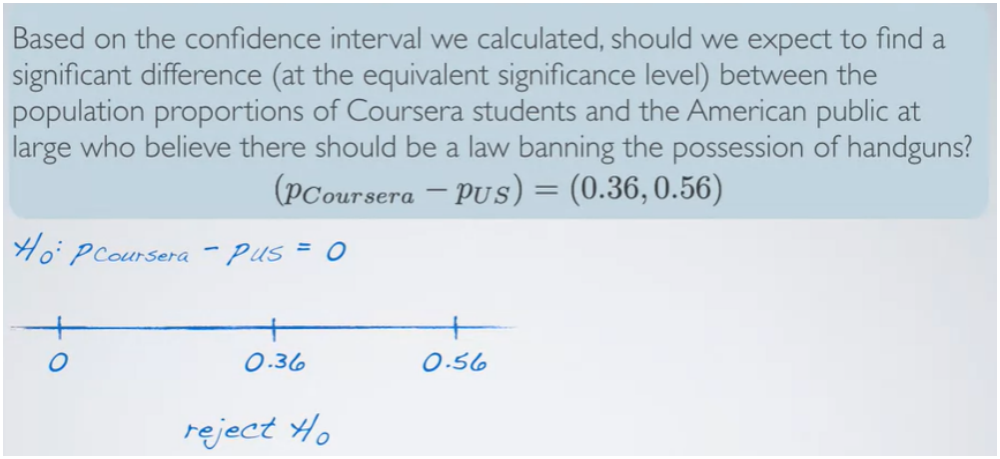
$$= 0.46 \pm 0.10$$

$$= (0.36, 0.56)$$

- It means that we are 95% confident that the proportion of Coursera students who believe there should be a ban on possession of handguns, is 36 to 56% higher than the proportion of Americans who do.
- That's a huge difference even when we factor in the variability, around the point estimate.
- This is probably expected based on how different the composition of the two populations are.



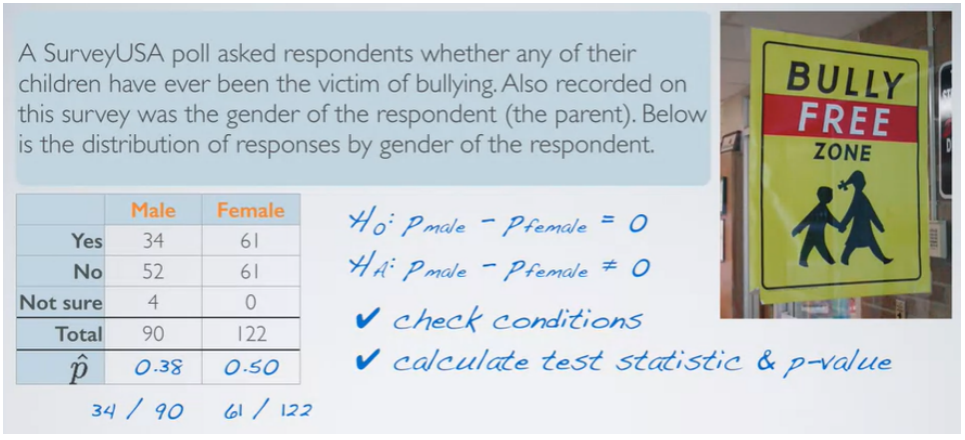
- The order doesn't matter as long as you interpret correctly.



: In this hypothesis test the null value for the difference between the two population proportions would be 0, and 0 isn't in the interval, hence we should expect to find a difference.

5. Hypothesis Test for Comparing Two Proportions

practice)



working with two proportions: $\hat{p}$ vs. p

	observed confidence interval	expected hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $n_2(1-\hat{p}_2) \geq 10$	$H_0 : p_1 = p_2$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	

- At no point do we define what these should be equal to. We don't have a readily available null value.
- What do we do? -> We make one up, the pooled proportion.
  - Could we actually come up with a best guess for what these could be equal to under the assumption of the null hypothesis.

pooled proportion

$H_0 : p_1 = p_2$ = ?

Pooled proportion:

$$\hat{p}_{pool} = \frac{total\ successes}{total\ n} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

Calculate the estimated pooled proportion of males and females who said that at least one of their children has been a victim of bullying.

$$\hat{p}_{pool} = \frac{34 + 61}{90 + 122} \approx 0.45$$

	Male	Female
Yes	34	61
No	52	61
Not sure	4	0
Total	90	122
$\hat{p}$	0.38	0.50

revisit: working with two proportions: $\hat{p}$ vs. p

	observed confidence interval	expected hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $n_2(1-\hat{p}_2) \geq 10$	$n_1\hat{p}_{pool} \geq 10$ $n_2\hat{p}_{pool} \geq 10$ $n_1(1-\hat{p}_{pool}) \geq 10$ $n_2(1-\hat{p}_{pool}) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$

what about means?

parameter of interest: μ

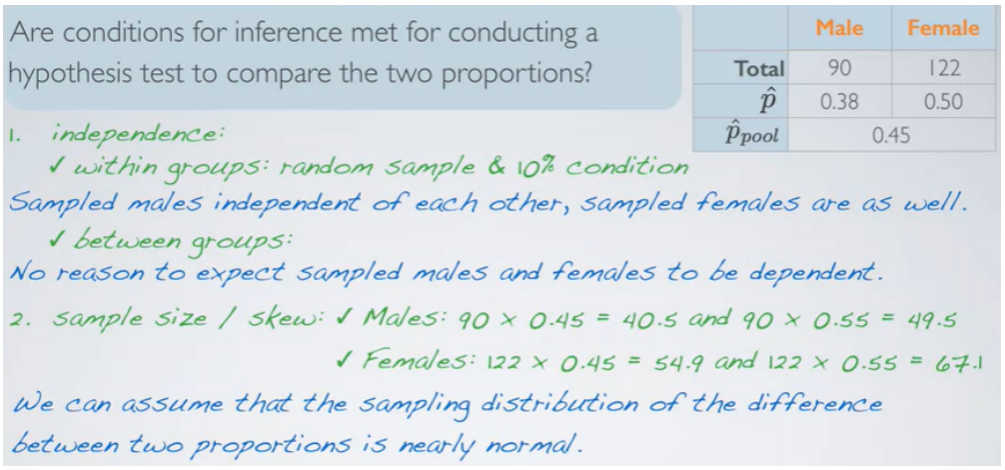
$H_0 : \mu = null\ value$

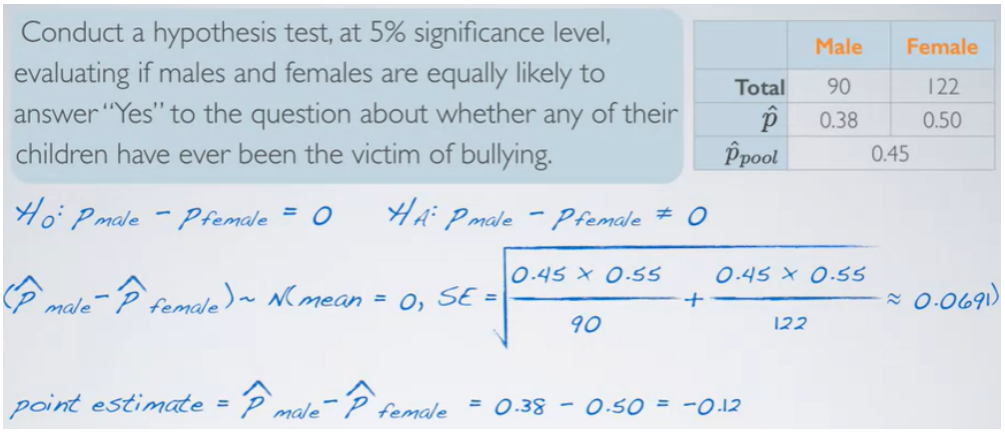
$SE = \frac{s}{\sqrt{n}}$ — μ doesn't appear in SE

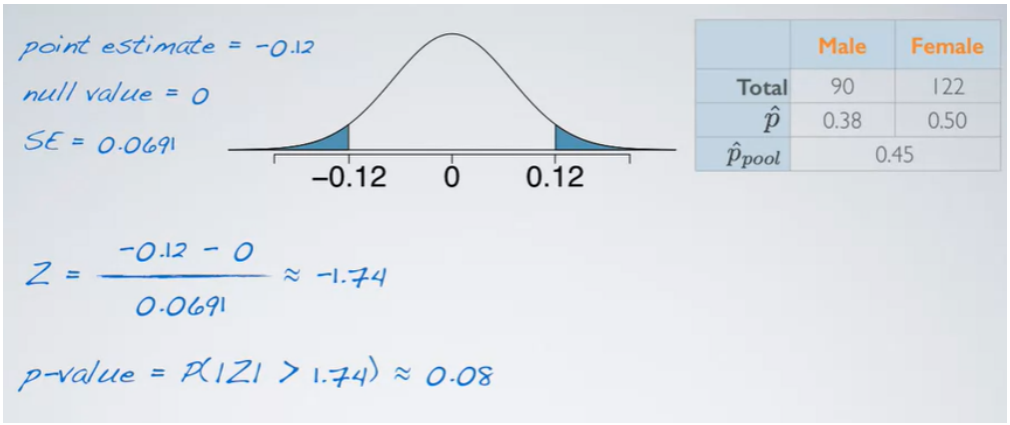
parameter of interest: p

$H_0 : p = null\ value$

$SE = \sqrt{\frac{p(1-p)}{n}}$ — p appears in SE







- There is no difference in males and females with respect to likelihood of reporting their kids being bullied.

[Simulation based inference for proportions and chi-square testing]

1. Small Sample Proportions

Paul the Octopus predicted 8 World Cup games, and predicted them all correctly. Does this provide convincing evidence that Paul actually has psychic powers, i.e. that he does better than just randomly guessing?

$H_0: p = 0.5$

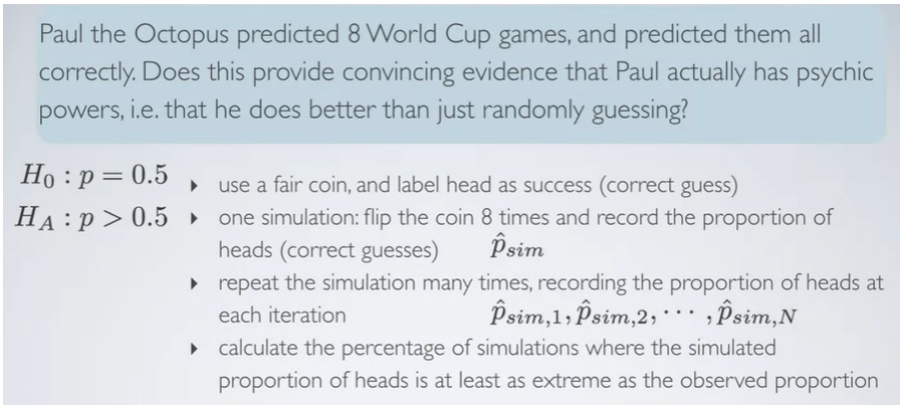
$H_A: p > 0.5$

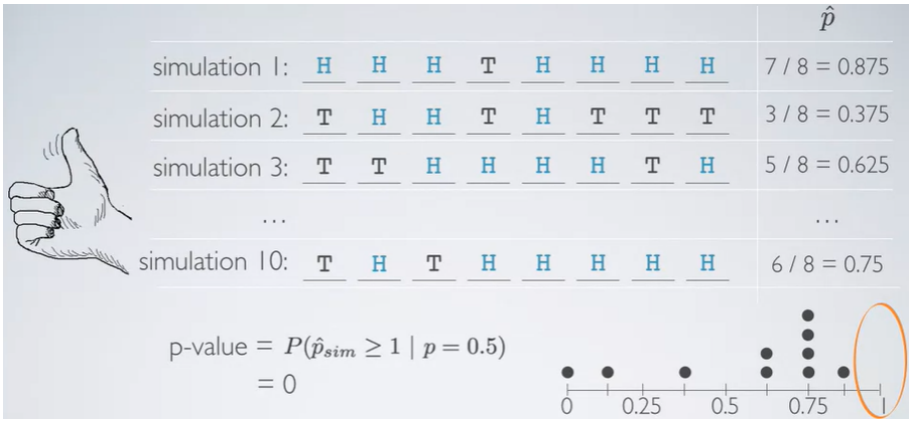
$n = 8$

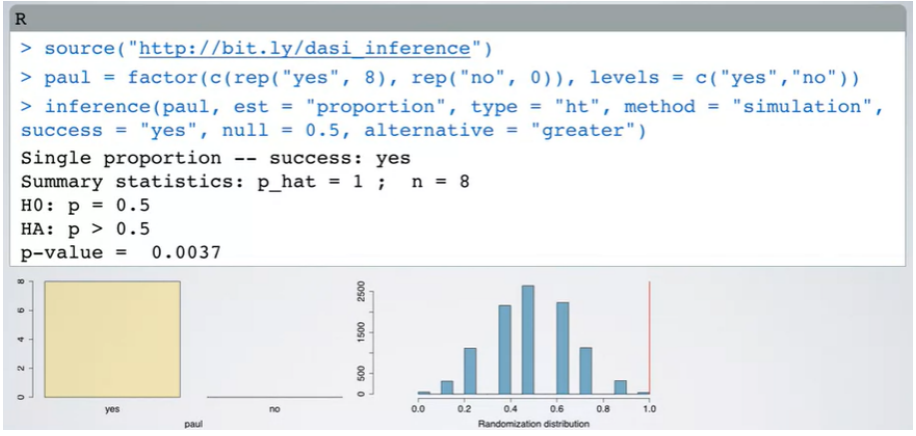
$\hat{p} = 1$

1. independence:
we can assume that his guesses are independent
2. sample size / skew: $8 \times 0.5 = 4$ → not met
distribution of sample proportions cannot be assumed to be nearly normal

- Inference via simulation
  - The ultimate goal of a hypothesis test is a p-value.
    - p-value : P(observed or more extreme outcome | H0 true)
  - devise a simulation scheme that assumes the null hypothesis is true
  - repeat the simulation many times and record relevant sample statistic
  - calculate p-value as the proportion of simulations that yield a result favorable to the alternative hypothesis







- What does rejection of the null hypothesis here mean?
    - Chances are we've made some sort of an error where the null hypothesis should not have been rejected. We had a pretty small sample size.
    - We might be making a Type 1 error.
    - The possibility would be to try to collect a little more data from Paul.

2. Examples

There is a saying in English *"to know something like the back of your hand"*, which means to know something very well. MythBusters (a popular TV show) put to test the validity of this saying.

They recruited 12 volunteers, each of whom were shown 10 pictures of backs of hands (while wearing gloves so they couldn't see their own hands), and asked them to identify their own hand among the 10 pictures. 11 out of 12 people completed the task successfully.

What are the hypotheses for evaluating whether these data provide convincing evidence of the validity of the saying, i.e. that people do better than random guessing when it comes to recognizing the back of their own hand?

$H_0: p = 0.1$

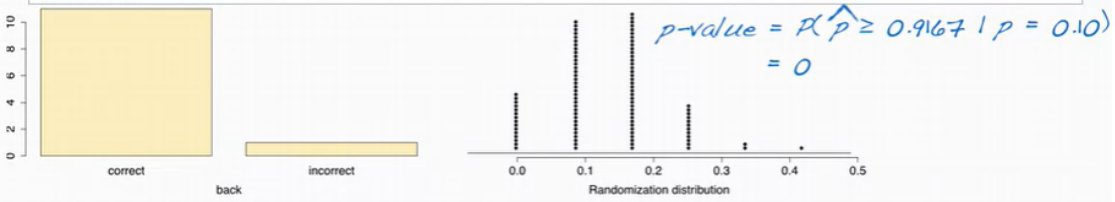
$H_A: p > 0.1$

Fill in the blanks below:

1. Use a 10-sided fair die to represent the sampling space, and call 1 a success (guessing correctly), and all other outcomes failures (guessing incorrectly).
2. Roll the die 12 times (each representing one of 12 people in the experiment), count the number rolls that resulted in ones, and calculate the proportion of correct guesses in one simulation of 12 rolls.
3. Repeat step (2) 100 times, each time recording the proportion of simulated successes in a series of 12 rolls of the die.
4. Create a dot plot of the simulated proportions from step (3) and count the number of simulations where the proportion is 11/12 or greater (the observed proportion).

R

```
> source("http://bit.ly/dasi_inference")
> back = factor(c(rep("correct", 11), rep("incorrect", 1)))
> inference(back, est = "proportion", type = "ht", method = "simulation",
success = "correct", null = 0.1, alternative = "greater", nsim = 100)
Single proportion -- success: correct
Summary statistics: p_hat = 0.9167 ;  n = 12
H0: p = 0.1
HA: p > 0.1
p-value =  0
```



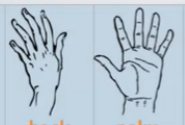
$p\text{-value} = P(\hat{p} \geq 0.9167 \mid p = 0.10) = 0$

- There is a 0% chance of 11 or more out of 12 people recognizing the backs of their hands in this experiment if in fact they were randomly guessing.

3. Comparing Two Small Sample Proportions

from MythBusters

"to know something like the back of your hand"



	back	palm	total
correct	11	7	18
incorrect	1	5	6
total	12	12	24
$\hat{p}$	0.9167	0.5833	0.75

Do these data provide convincing evidence that there is a difference in how good people are at recognizing the backs and the palms of their hands?

$H_0: p_{back} - p_{palm} = 0$ $H_A: p_{back} - p_{palm} \neq 0$

1. independence:
   - ✓ within groups: within each group we can assume that the guess of one subject is independent of another.
   - ✓ between groups: No, same people guessing – assume to be met for illustrative purposes
2. sample size / skew: $12 \times 0.75 = 9$ and $12 \times 0.25 = 3$ – not met, use simulation methods

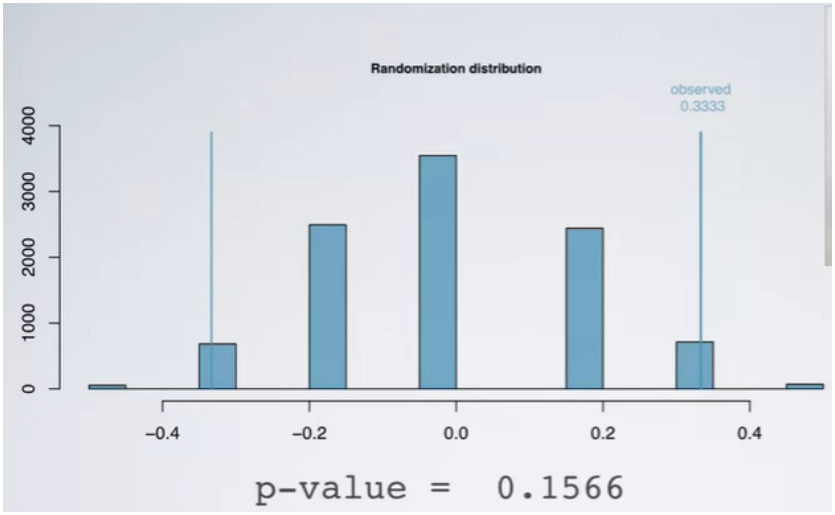
simulation scheme

1. Use 24 index cards, where each card represents a subject.
2. Mark 18 of the cards as "correct" and the remaining 6 as "wrong".
3. Shuffle the cards and split into two groups of size 12, for back and palm.
4. Calculate the difference between the proportions of "correct" in the back and palm decks, and record this number.
5. Repeat steps (3) and (4) many times to build a randomization distribution of differences in simulated proportions.

- Interpreting the simulation results

: simulate the experiment under the assumption of independence, i.e. leaving things up to chance

- results from the simulations look like the data -> the difference between the proportions of correct guesses in the two groups was due to chance.
- results from the simulations do not look like the data -> the difference between the proportions of correct guesses in the two groups was not due to chance, but because people actually know the backs of their hands better.



- The height of the bars represent what percent of the time, or how many times within these 10,000 simulations a particular simulated p hat was achieved.
- The success rate in the back of the hand group, and those in the palm of the hand group
- The difference between the two proportions come out to be roughly 33% away from the center of the distribution.
- With a p-value of 0.16, we would fail to reject the null hypothesis and say that there isn't actually convincing evidence that people are between at or worse or at least there's some difference in how they recognize the backs versus the palms of their hands.

4. Chi-Square GOF Test

jury selection

- In a county where jury selection is supposed to be random, a civil rights group sues the county, claiming racial disparities in jury selection.
- Distribution of ethnicities of the people in the county who are eligible for jury duty (based on census results):

ethnicity	white	black	nat. amer.	asian & PI	other
%in population	80.29%	12.06%	0.79%	2.92%	3.94%

- Distribution of 2500 people who were selected for jury duty the previous year:

ethnicity	white	black	nat. amer.	asian & PI	other
observed #	1920	347	19	84	130

jury selection

The court retains you as an independent expert to assess the statistical evidence that there was discrimination. You propose to formulate the issue as an hypothesis test.

H_0 (nothing going on): People selected for jury duty are a simple random sample from the population of potential jurors. The observed counts of jurors from various race/ethnicities follow the same ethnicity distribution in the population.

H_A (something going on): People selected for jury duty are not a simple random sample from the population of potential jurors. The observed counts of jurors from various ethnicities do not follow the same race/ethnicity distribution in the population.

evaluating the hypotheses

- quantify how different the observed counts are from the expected counts
- large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- called a goodness of fit test since we're evaluating how well the observed data **fit** the expected distribution

Calculate expected number of jurors from each ethnicity if in fact the jury selection is random.

ethnicity	white	black	nat. amer.	asian & PI	other	total
%in population	80.29%	12.06%	0.79%	2.92%	3.94%	100%
expected #	2007 + 302 + 20 + 73 + 98					= 2500

2500 × 0.8029

2500 × 0.1206

Conditions for the chi-square test:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
   - each case only contributes to one cell in the table
2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

anatomy of a test statistic

general form of a test statistic

$$\frac{point\ estimate - null\ value}{SE\ of\ point\ estimate}$$

1. Identifying the difference between a point estimate and an expected value if the null hypothesis were true
2. standardizing that difference using the standard error of the point estimate

chi-square statistic

when dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the chi-square (χ^2) statistic.

χ^2 statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$

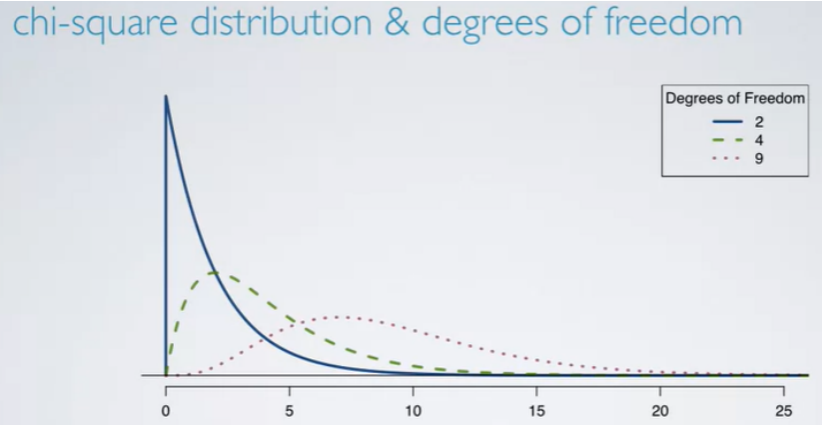
O : observed
E : expected
k : number of cells

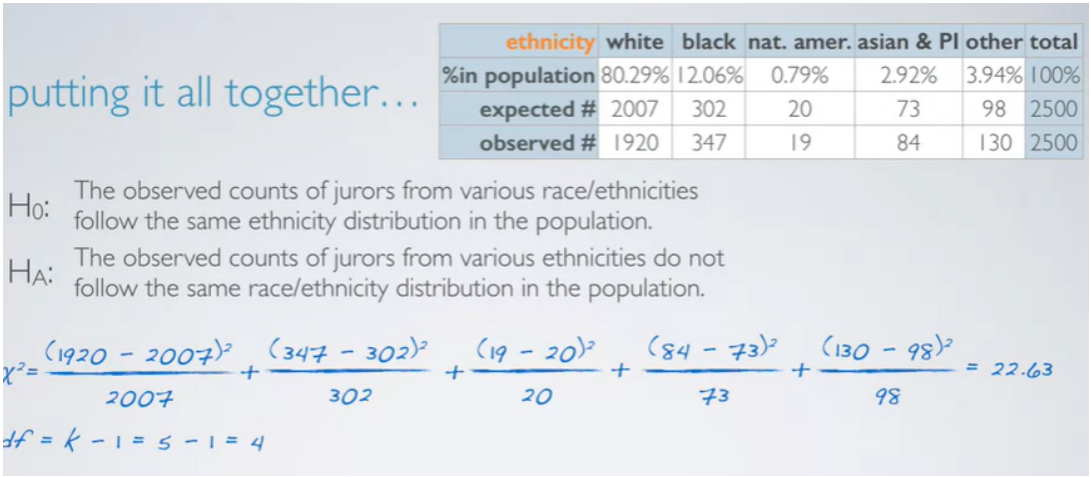
- Why square?
  - positive standardized difference
  - highly unusual differences between observed and expected will appear even more unusual

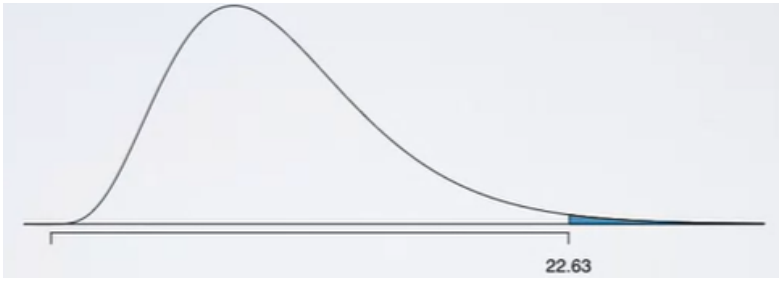
- degrees of freedom
  - to determine if the calculated χ^2 statistic is considered unusually high or not we need to first describe its distribution
  - chi-square distribution has just 1 parameter:
    - degrees of freedom (df) : influences the shape, center and spread

χ^2 **degrees of freedom for a goodness of fit test:** $df = k - 1$

k : number of cells





- p-value
  - p-value for a chi-square test is defined as the tail area above the calculated test statistic
  - because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis.
  - 

p-value

$\chi^2 = 22.63$ $df = 4$

using R

```
R
> pchisq(22.63, 4, lower.tail = FALSE)
[1] 0.0002
```

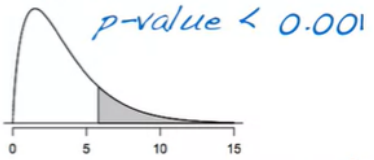
using the applet

http://bitly.com/dist_calc

using the table

Chi-square probability table

p-value < 0.001



Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

- With such a small p-value, we would reject the null hypothesis, which in this context means that the data provide convincing evidence that the observed distribution of the counts of race ethnicities of jurors does not follow the distribution in the population.

5. The Chi-Square Independence Test

study results

	dating	cohabiting	married	total
obese	81	103	147	331
not obese	359	326	277	962
total	440	429	424	1293

Does there appear to be a relationship between weight and relationship status?

hypotheses

H_0 (nothing going on): Weight and relationship status are independent. Obesity rates do not vary by relationship status.

H_A (something going on): Weight and relationship status are dependent. Obesity rates do vary by relationship status.

- evaluating the hypotheses
  - quantify how different the observed counts are from the expected counts
  - large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
  - called an independence test since we're evaluating the relationship between two categorical variables

chi-square test of independence

χ^2 test of independence:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

$$df = (R-1) \times (C-1)$$

O : observed E : expected
k : number of cells
R : number of rows
C : number of columns

Conditions for the chi-square test:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
   - each case only contributes to one cell in the table
2. ***Sample size:*** Each particular scenario (i.e. cell) must have at least 5 expected cases.

What is the overall obesity rate in the sample?

331 / 1293 = 0.256

	dating	cohabiting	married	total
obese	81	103	147	331
not obese	359	326	277	962
total	440	429	424	1293

If in fact weight and relationship status are independent (i.e. if in fact H_0 is true) how many of the dating people would we expect to be obese? How many of the cohabiting and married?

dating: 440 × 0.256 ≈ 113

cohabiting: 429 × 0.256 ≈ 110

married: 424 × 0.256 ≈ 108

expected counts in two-way tables

$$expected\ count = \frac{(row\ total) \times (column\ total)}{table\ total}$$

Test the hypothesis that relationship status and obesity are associated at the 5% significance level.

	dating	cohabiting	married	total
obese	81 **(113)**	103 **(110)**	147 **(108)**	331
not obese	359 **(327)**	326 **(319)**	277 **(316)**	962
total	440	429	424	1293

$$\chi^2 = \frac{(81-113)^2}{113} + \frac{(103-110)^2}{110} + \frac{(147-108)^2}{108} + \frac{(359-327)^2}{327} + \frac{(326-319)^2}{319} + \frac{(277-316)^2}{316}$$

$$= 31.68$$

$$df = (2-1) \times (3-1) = 1 \times 2 = 2$$

R

```
> pchisq(31.68, 2, lower.tail = FALSE)
[1] 1.320613e-07
```

- So we reject H0, which means these data provide convincing evidence that relationship status and obesity are associated.
- Can we conclude from these data that living with someone is making some people obese, and that marrying someone is making people even more obese? -> No! It is an observational study; we can't do causal study here.

chi-square tests

- goodness of fit: comparing the distribution of one categorical variable (with more than 2 levels) to a hypothesized distribution
- independence: evaluating the relationship between two categorical variables (at least one with more than 2 levels)

8. Suppose in a population 20% of people wear contact lenses. What is the expected shape of the sampling distribution of proportion of contact lens wearers in random samples of 1000 people from this population?

- ○ uniform
- ○ left-skewed
- ◉ right-skewed
- ○ nearly normal

⊗ **Incorrect**

The question refers to the following learning objective(s):

Note
that if the CLT doesn't apply and the sample proportion is low (close to 0) the sampling distribution will likely be right skewed, if the sample proportion is high (close to 1) the sampling distribution will likely be left skewed.

S-F condition met.