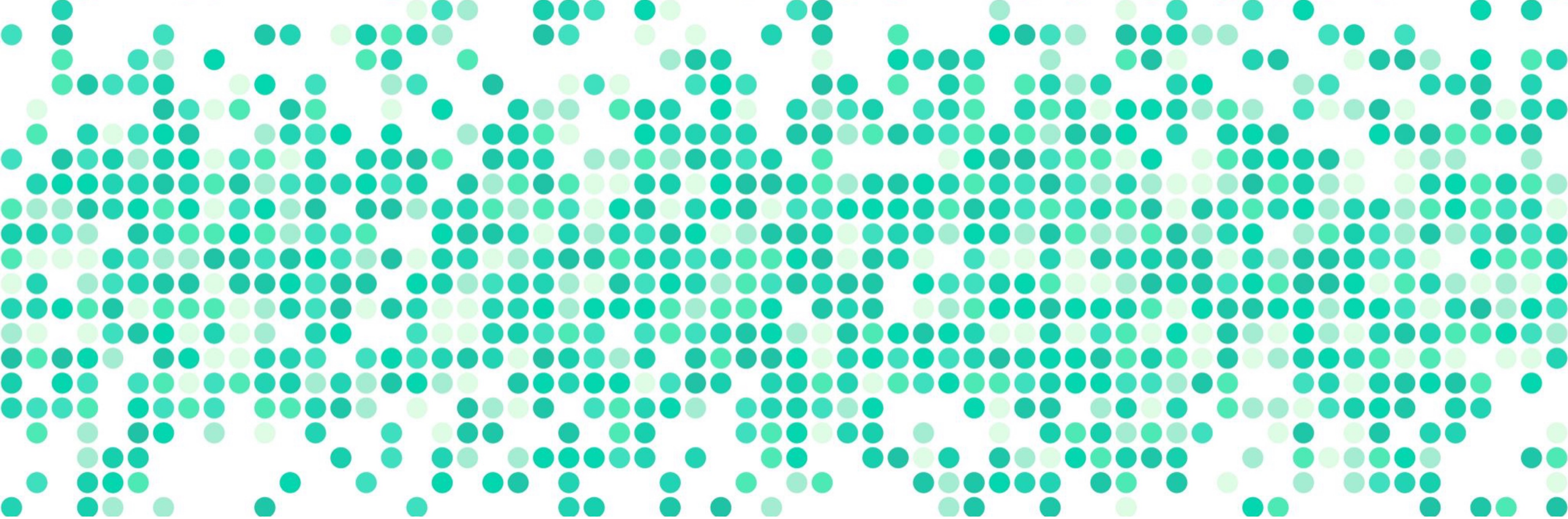


DATA ANALYSIS

Week 7: Hypothesis testing

logistics

midterm 1

- all scores up on Canvas

office hours

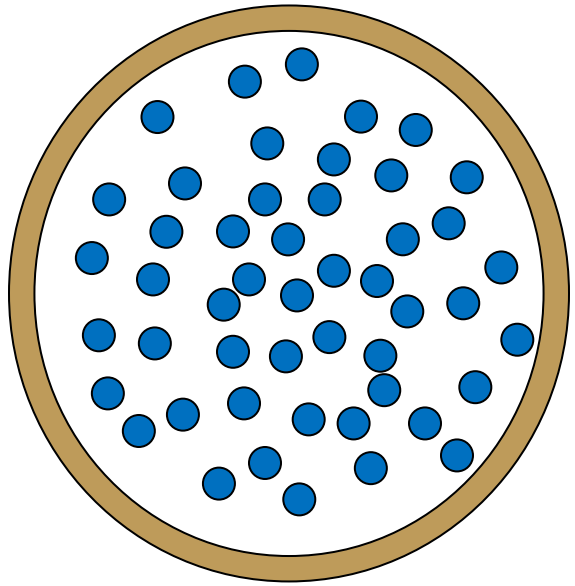
- Wed (today, Kanbar 217): 1.15-2.45 pm
- Thurs (virtual): 2-3 pm, 10-12 pm

problem sets

- opt-out deadline now March 12
- PS4 is also due March 12

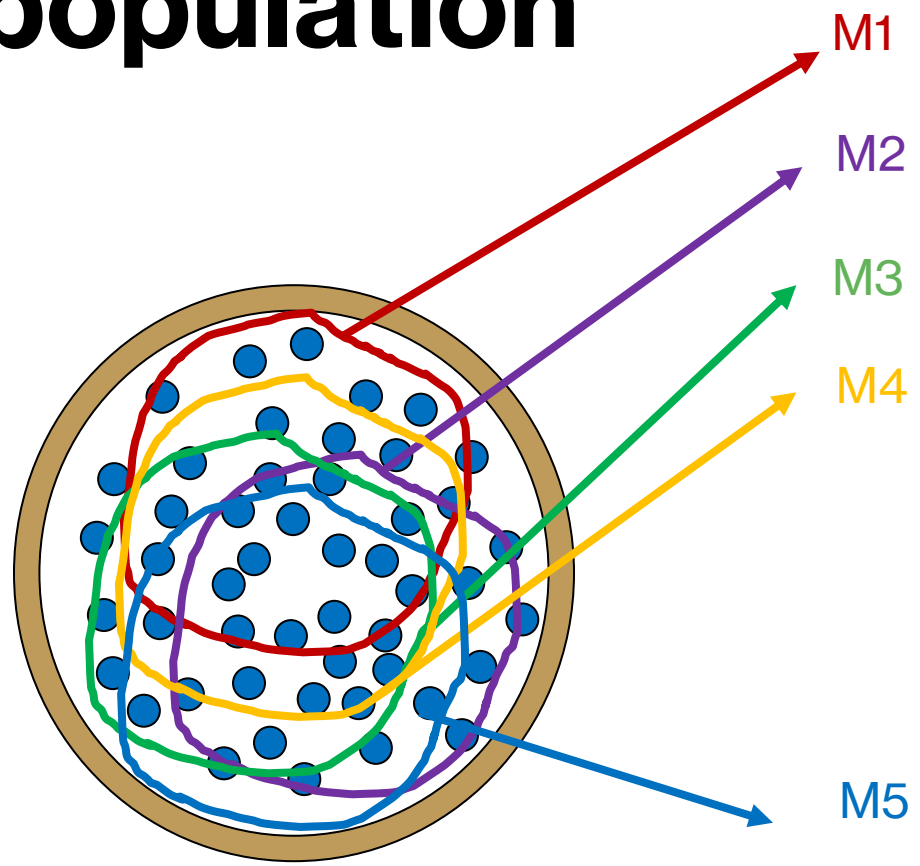
recap: sampling from a population

- each time we sample, we compute some type of statistic (e.g., mean, correlation, etc.)
- sampling error: the discrepancy between the sample statistic and the true population parameter it is estimating
- sampling distribution: distribution of all possible values of the **statistic** obtained from multiple samples of a given size
- distribution of sample means contains all sample means of a size ***n*** that can be obtained from a population



recap: sampling from a population

- each time we sample, we compute some type of statistic (e.g., mean, correlation, etc.)
- sampling error: the discrepancy between the sample statistic and the true population parameter it is estimating
- sampling distribution: distribution of all possible values of the **statistic** obtained from multiple samples of a given size
- distribution of sample means contains all sample means of a size ***n*** that can be obtained from a population



what does the sampling distribution of means look like??

recap: central limit theorem

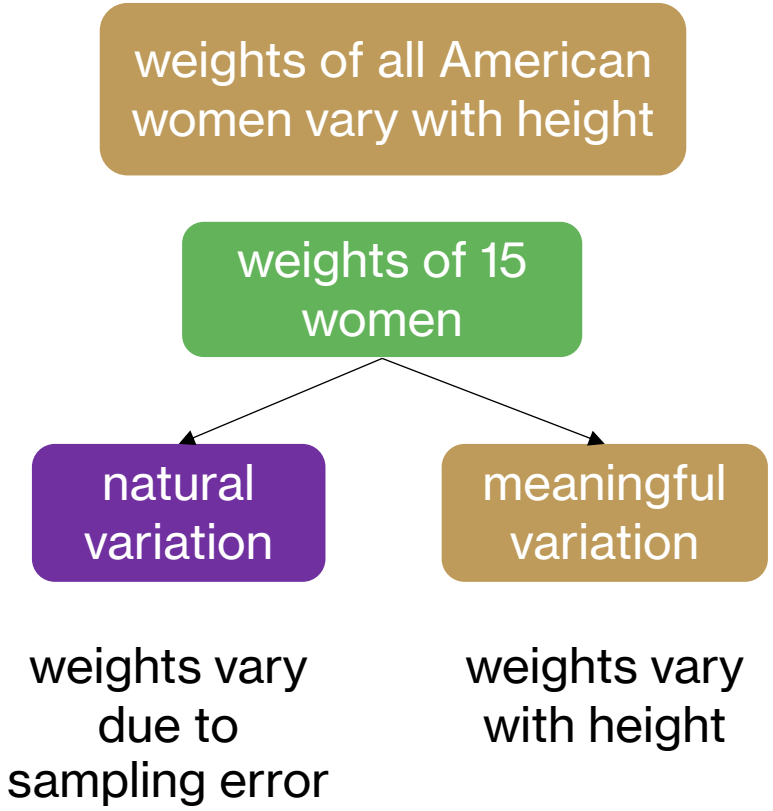
- the central limit theorem states that for **any** population with mean μ and standard deviation σ , the distribution of sample means for sample size *n* will have:
  - a mean of $\mu_M = \mu$ = expected value of M
  - a standard deviation of $\sigma_M = \frac{\sigma}{\sqrt{n}}$ = standard error of the mean or M
  - will approach a normal distribution as n approaches infinity
  - distribution of sample means will be normally distributed **even if the population was not normally distributed (if n is large enough!)**
  - typically n (number of scores in a sample) around 30 yields a reasonably normal distribution
- CLT only applies to the distribution of sample means, i.e., if our sample statistic is NOT a mean, and our hypotheses are NOT about means, then we must use a different sampling distribution for the null hypothesis

hypothesis testing: fundamentals

- our goal is to evaluate the likelihood of the hypothesis, given the sample statistic we have obtained, i.e., how likely is my hypothesis?
- P (your hypothesis, given the data sample)

  = P (your hypothesis | sample statistic)

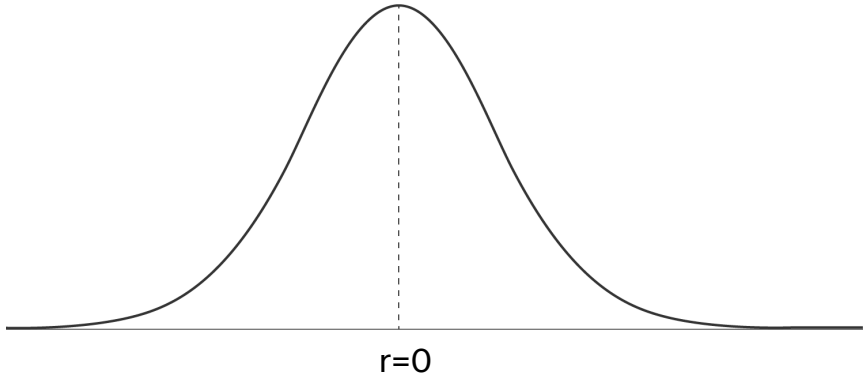
  = P (weights vary with height | r = 0.995)
- we make use of the sampling distribution of the sample statistic to reason about our hypothesis



from samples to populations

- we can start by assuming that our hypothesis is wrong
  - H_0: null hypothesis: there is no meaningful relationship between Y (weight) and X (height)
  - population parameter, $\rho = 0$
- we generate a sampling distribution of the statistic given that the null hypothesis is true, and examine where our sample statistic lies within this distribution

ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



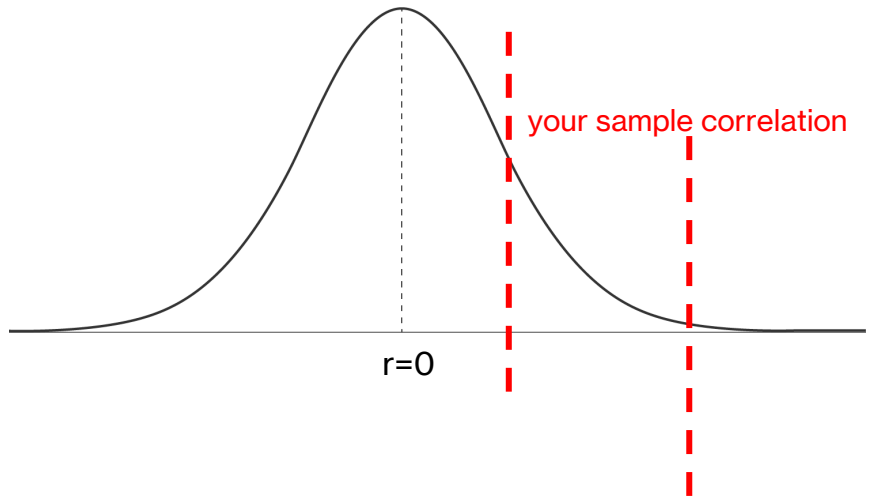
from samples to populations

- once we know the form of the sampling distribution under the null hypothesis, we want to know how likely is the sample statistic you obtained
- P (your sample correlation | true correlation = 0)

  = P (your sample correlation | null hypothesis)

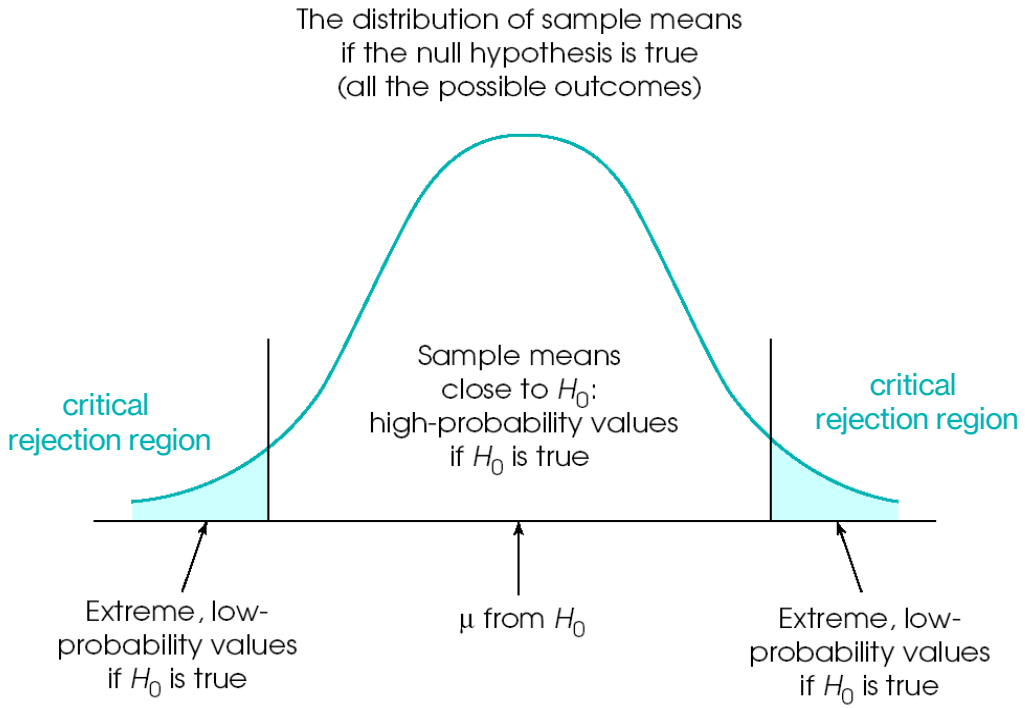
  = P (r= 0.995 | null hypothesis is true)
- if this probability is really low, we can infer that the null hypothesis may not be true, and subsequently infer that your actual hypothesis may be true!
  - H_1: alternative hypothesis ($\rho \neq 0$)

ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



why the null?

- why do we rely on the null hypothesis? because then we can estimate what would be the natural variation expected in samples via the sampling distribution
- remember that for large samples, the sampling distribution of a statistic is normal even if the population is not (CLT)
- which criteria should we use to assess statistical significance?
- α-level is called the significance-level criteria, typically set to 0.05, if P (data | null) is in the extreme 5% of the distribution, we will reject the null hypothesis in favor of the alternative hypothesis



summary of NHST

example

- the average mother sea turtle lays $\mu = 80$, $\sigma = 6$ eggs per mating season. We work for an endangered species foundation, and are testing the effectiveness of a new hormone (X15) on turtle fertility. We predict that turtles treated with the hormone will produce *different* nest sizes from the average turtle (no direction). We collect a sample of n=30 turtles from the above population and treat them with the hormone. We then count the number of eggs in their nest and get a mean of 84.
- is the fertility hormone effective?
- conduct hypothesis testing: use the α =.05 significance level

framing the problem

- population characteristics (usual turtles): $\mu = 80, \sigma = 6$
- sample characteristics (our 30 turtles): *M* = 84
- two possible explanations for the difference in sample mean (M) and population mean (μ)
  - sampling error (H_0: null hypothesis)
  - true effect of hormone (H_1:alternative hypothesis)
- we assume the null hypothesis is true and set out to reject this assumption with our evidence and chance model

step 1: stating the hypotheses

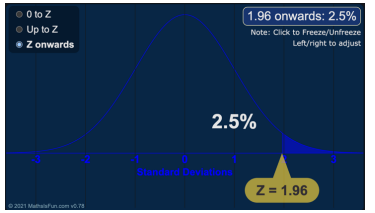
- null hypothesis: X15 does not have an effect on nest size, i.e., it stays the same and any variation observed in samples is simply due to sampling error

$$H_o: \mu = 80$$

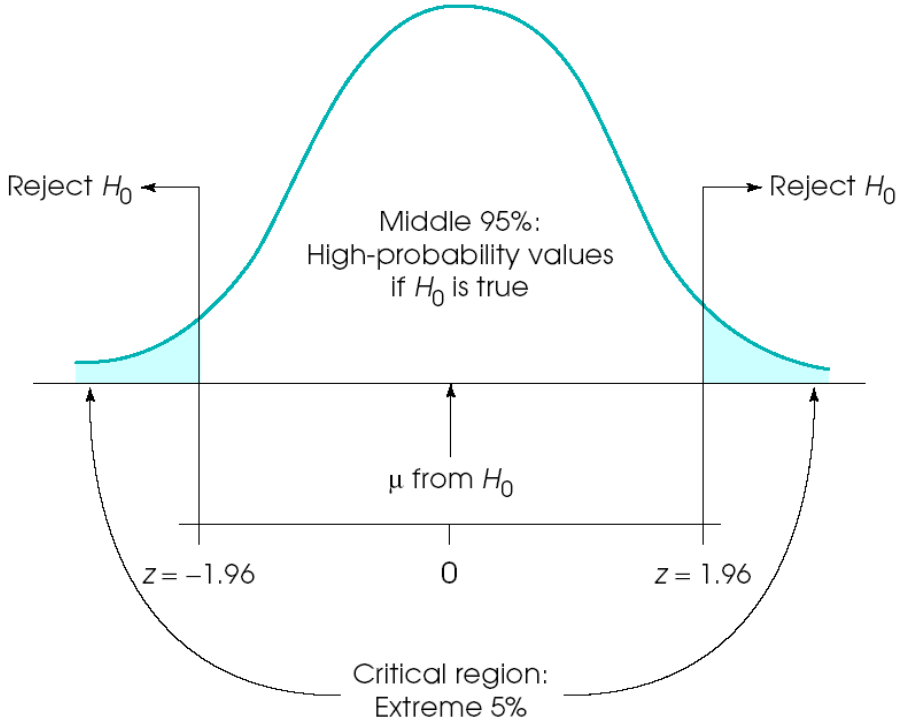
- alternative hypothesis: X15 has an effect on nest size

$$H_1: \mu \neq 80$$

step 2: set criteria for decision

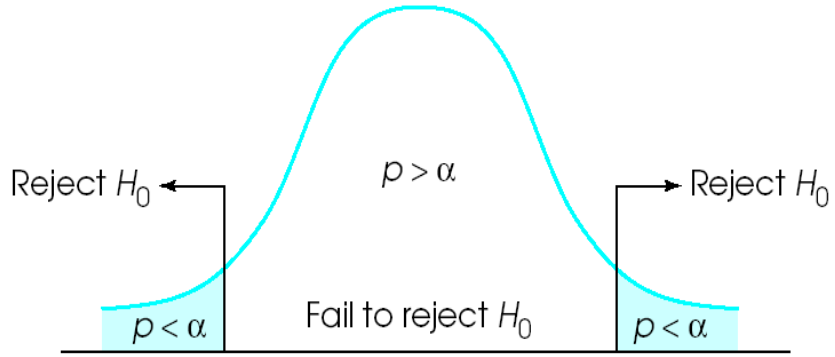


- examine the distribution of sample means
- the central limit theorem states that for **any** population with mean μ and standard deviation σ , the distribution of sample means for sample size *n* will have:
  - a mean of $\mu_M = \mu$ = expected value of M = 80
  - a standard deviation of $\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{30}}$ = 1.095
- represent the problem graphically
- criteria: α-level = .05 (extreme 5%)
- find $z_{critical} = \pm 1.96$, $p_{critical} = .05$
- for a p-value smaller than .05, the obtained sample mean is unlikely to have come from this expected sampling distribution for the null hypothesis



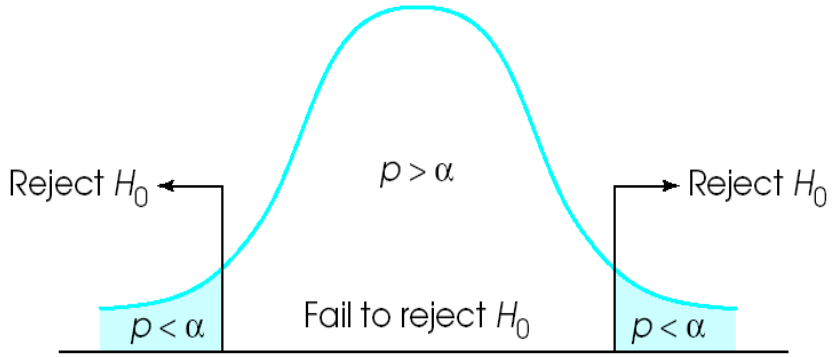
step 3 and 4: collect data and decide!

- we collect the data and evaluate the probability of obtaining the data as extreme as this, under the null hypothesis
  - P (data | null)
  - compute z-score of sample mean under the sampling distribution
  - $z_{observed} = \frac{M - \mu}{\sigma_M} = \frac{84 - 80}{1.095} = 3.65$
  - remember that $z_{critical} = \pm 1.96$
  - look up the probability, $p_{observed} < .001$ and $p_{critical} = .05$
- this sample is very rare if the null hypothesis was true
- conclusion: we reject the null hypothesis that the hormone does not produce a difference
- reporting: X15 has a significant effect on sea turtle fertility ($z = 3.65$, $p < .001$)



statistical significance

- typically, when a hypothesis test is conducted, you report the "p – value" associated with that test, which denotes the likelihood of obtaining the data under the null hypothesis: P(data | null)
- findings are said to be statistically significant if the null hypothesis ("no effect") has been rejected
- reporting results:
  - treatment (X15) was effective, $z = 3.65$, $p < .001$
  - treatment (A) produced no effect, $z = 0.67$, $p = .251$



statistical significance ≠practical significance!

- we found that the effect of the X15 hormone was “statistically significant” ($p < .001$)
- should we provide this hormone to all turtles? why or why not?
- statistical significance only tells us the probability of the data under the null hypothesis; it does not tell us the *how important* the finding is or how practically significant this effect may be
- what are some reasons that this finding may not be practically significant?

common misconceptions

- we find a significant p-value of 0.03. does it mean:
- the probability of the null hypothesis being true is .03?
- the probability that you are making the wrong decision is .03?
- if you ran the study again, you would obtain the same result 97% of the time?

activity (Q9a from Ch 8)

- The psychology department is gradually changing its curriculum by increasing the number of online course offerings. To evaluate the effectiveness of this change, a random sample of n = 36 students who registered for Introductory Psychology is placed in the online version of the course. At the end of the semester, all students take the same final exam. The average score for the sample is M = 76. For the general population of students taking the traditional lecture class, the final exam scores form a normal distribution with a mean of $\mu = 71$.
- If the final exam scores for the population have a standard deviation of $\sigma = 12$, does the sample provide enough evidence to conclude that the new online course is significantly different from the traditional class? Use a two-tailed test with $\alpha = .05$.

step 1: stating the hypotheses

- null hypothesis: the new online course does not produce any change in scores compared to the traditional class and any variation observed in samples is simply due to sampling error

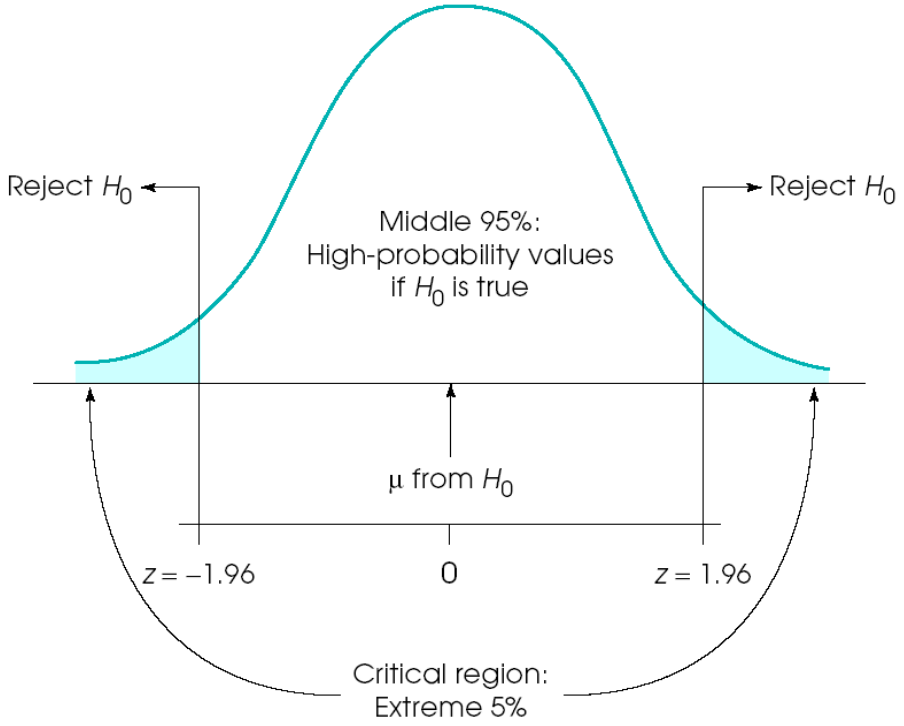
$$H_o: \mu = 71$$

- alternative hypothesis: the new online course is significantly different from the traditional class

$$H_1: \mu \neq 71$$

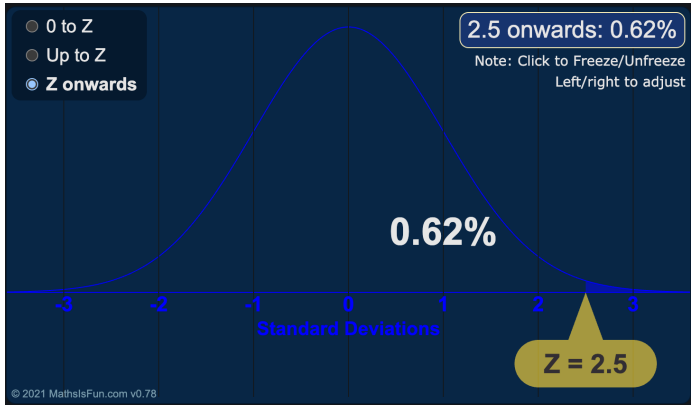
step 2: set criteria for decision

- examine the distribution of sample means
- n = 36
- $\mu_M = \mu$ = expected value of M = 71
- $\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = 2$
- criteria:
  - α-level = .05
  - $z_{critical} = \pm 1.96$ (extreme 5%)
  - $p_{critical} = .05$



step 3 and 4: collect data and decide!

- compute z-score of this mean under the sampling distribution
  - $z_{observed} = \frac{M - \mu}{\sigma_M} = \frac{76 - 71}{2} = 2.5$
  - $p_{observed} = .0062 + .0062 = .012$ (both sides)
  - remember that $z_{critical} = \pm 1.96$ and $p_{critical} = .05$
- this sample is very <u>rare</u> if the null hypothesis was true
- conclusion: we reject the null hypothesis
- reporting: online course has a significant effect on scores ($z = 2.5$, $p = .012$)

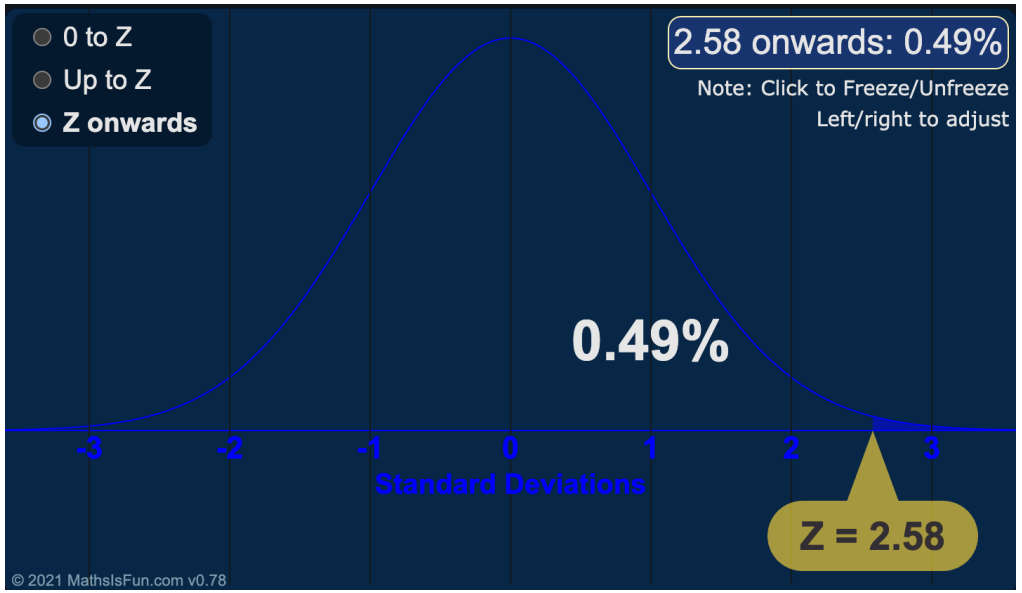


activity (Q9 from Ch 8)

- The psychology department is gradually changing its curriculum by increasing the number of online course offerings. To evaluate the effectiveness of this change, a random sample of n = 36 students who registered for Introductory Psychology is placed in the online version of the course. At the end of the semester, all students take the same final exam. The average score for the sample is M = 76. For the general population of students taking the traditional lecture class, the final exam scores form a normal distribution with a mean of $\mu = 71$.
- If the final exam scores for the population have a standard deviation of $\sigma = 12$, does the sample provide enough evidence to conclude that the new online course is significantly different from the traditional class? Use a two-tailed test with $\alpha = .01$.

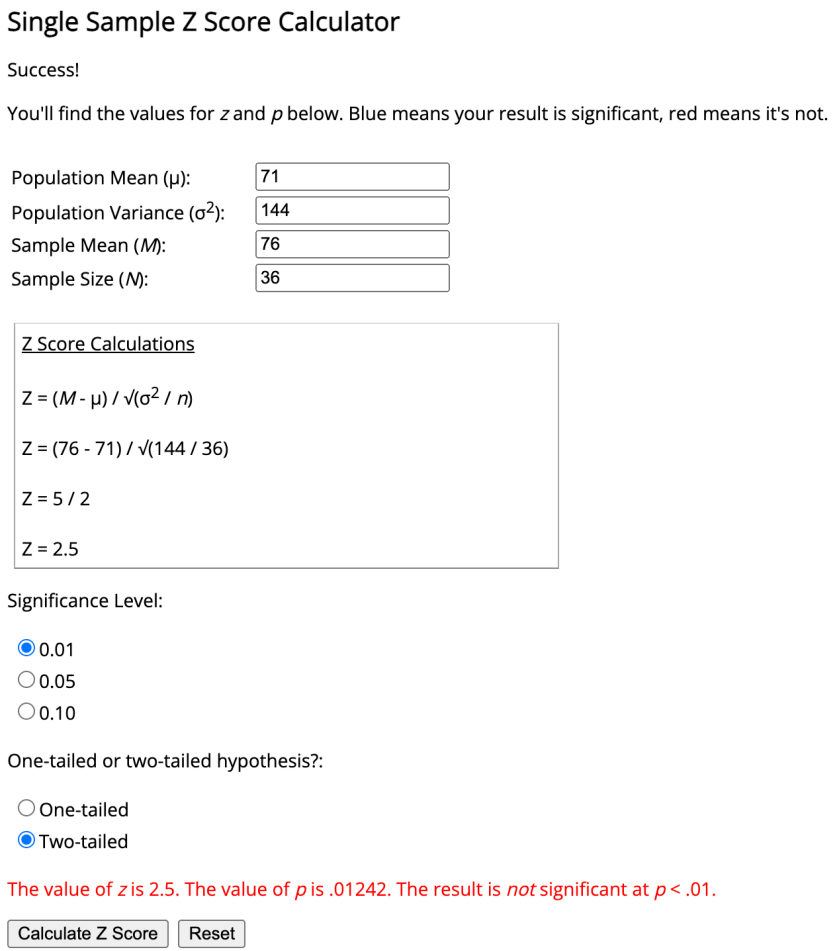
what changes?

- step 1: stating the hypotheses (same!)
- step 2: set criteria for decision
  - n = 36
  - $\mu_M = \mu$ = expected value of M = 71
  - $\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{36}} = 2$
  - $z_{critical}$ based on .5% on each side = $\pm$ 2.58
  - $p_{critical}$ = .01
- step 3: collect data (same!)
  - $z_{observed} = \frac{M - \mu}{\sigma_M} = \frac{76 - 71}{2} = 2.5, p_{observed} = .012$
- step 4: decide
  - $z_{observed} < z_{critical}$ and $p_{observed} > p_{critical}$
  - cannot reject the null hypothesis!



using online calculator (z-test)

- [z-test calculator]
- can provide population parameters and sample statistics and specify the criterion to obtain z and p-values
- good check when you do manual calculations (showing your work is required in problem sets and exams!)
- google sheets formula: ZTEST(range, mean, sd)
  - assumes α-level = .05

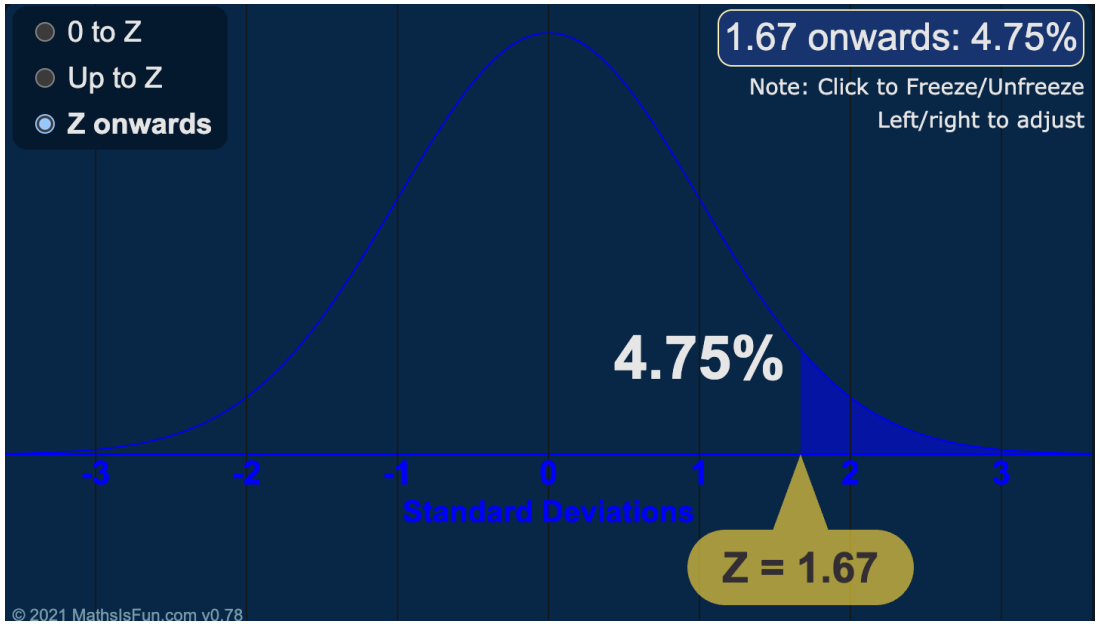


activity (Q9b from Ch 8)

- The psychology department is gradually changing its curriculum by increasing the number of online course offerings. To evaluate the effectiveness of this change, a random sample of n = 36 students who registered for Introductory Psychology is placed in the online version of the course. At the end of the semester, all students take the same final exam. The average score for the sample is M = 76. For the general population of students taking the traditional lecture class, the final exam scores form a normal distribution with a mean of $\mu = 71$.
- If the population standard deviation is $\sigma = 18$, is the sample sufficient to demonstrate a significant difference? Again, use a two-tailed test with $\alpha = .05$.

what changes?

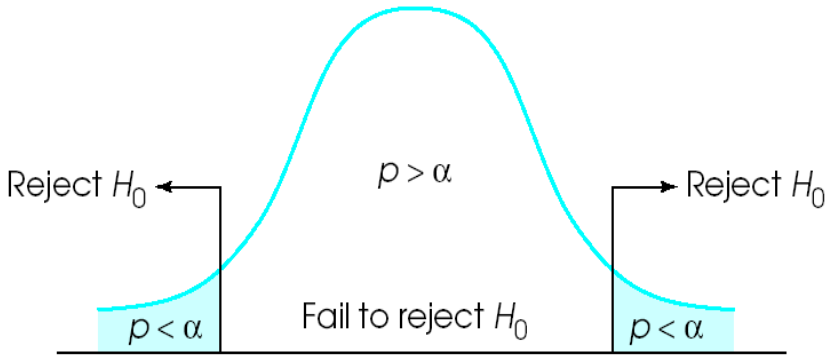
- step 1: stating the hypotheses (same!)
- step 2: set criteria for decision
  - n = 36
  - $\mu_M = \mu$ = expected value of M = 71
  - $\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{36}} = 3$, $z_{critical} = \pm 1.96$, $p_{critical} = .05$
- step 3: collect data
  - $z_{observed} = \frac{M - \mu}{\sigma_M} = \frac{76 - 71}{3} = 1.67$
  - $p_{observed} = .0475 + .0475 = .095$
- step 4: decide
  - $z_{observed} < z_{critical}$
  - cannot reject the null hypothesis!



summary

- making the critical region region smaller (by decreasing the α-level) makes the z-test more conservative, i.e., $z_{critical}$ will be higher, making it harder to reject the null hypothesis
- higher standard errors (due to high population variance σ or low sample size *n*) will impact the sample z-score ($z_{observed}$), i.e., where the sample statistic lies relative to the distribution

$$z = \frac{M - \mu}{\sigma_M} \; and \; \sigma_M = \frac{\sigma}{\sqrt{n}}$$



one vs. two-tailed tests

- two-tailed tests make no assumptions about directionality when discussing the hypotheses
  - H_o: $\mu = 80$, H_1: $\mu \neq 80$ (sea turtles example)
  - $\alpha = 0.05$ splits the null distribution into two regions (corresponding to p < .025 and p > .025)
- one-tailed (directional) tests specify a direction in the hypotheses, i.e., an increase or decrease in the population parameter
  - H_o: $\mu \leq 80$, H_1: $\mu > 80$
  - $\alpha = 0.05$ is restricted to only ONE part of the null distribution, leading to a larger area
  - more sensitive but also less conservative

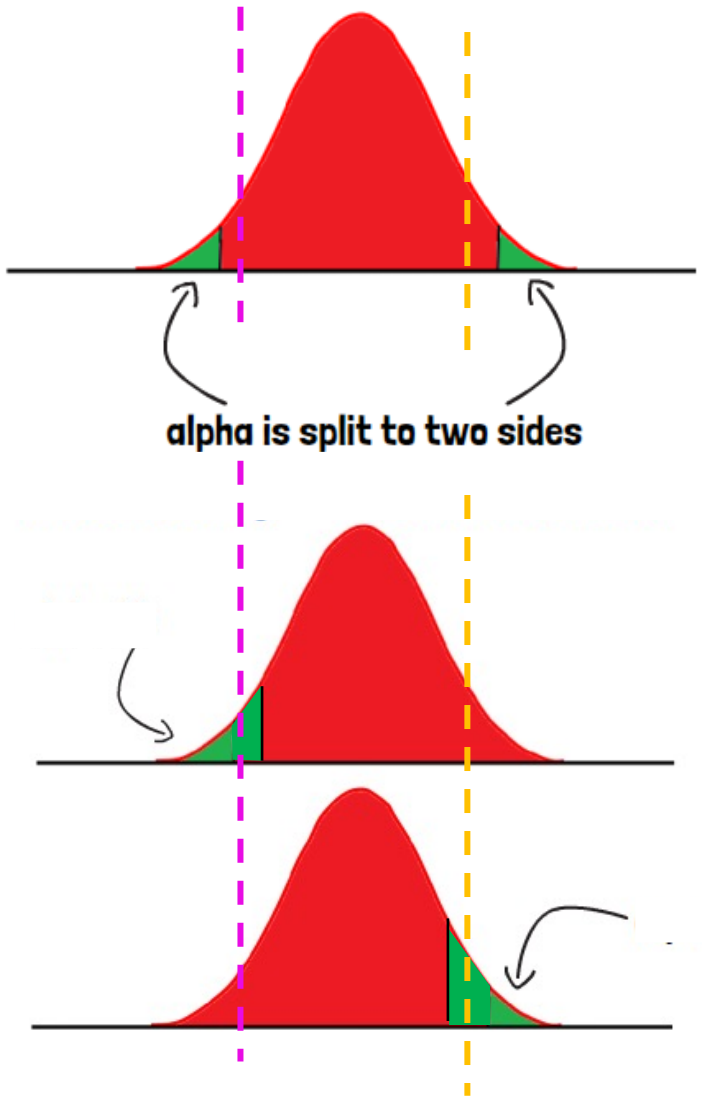


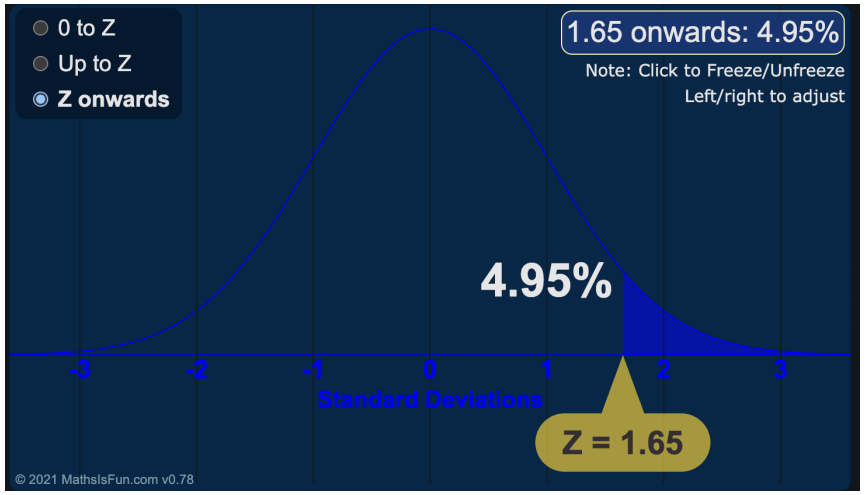
figure source

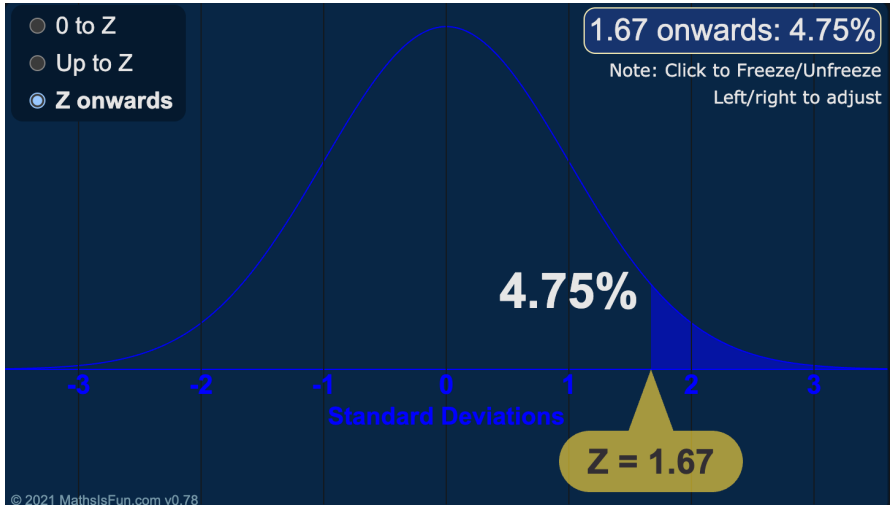
activity

- The psychology department is gradually changing its curriculum by increasing the number of online course offerings. To evaluate the effectiveness of this change, a random sample of n = 36 students who registered for Introductory Psychology is placed in the online version of the course. At the end of the semester, all students take the same final exam. The average score for the sample is M = 76. For the general population of students taking the traditional lecture class, the final exam scores form a normal distribution with a mean of μ = 71.
- If the population standard deviation is σ = 18, is the sample sufficient to demonstrate that online courses **increases** the score compared to traditional class? Use a one-tailed test with α = .05.

what changes?

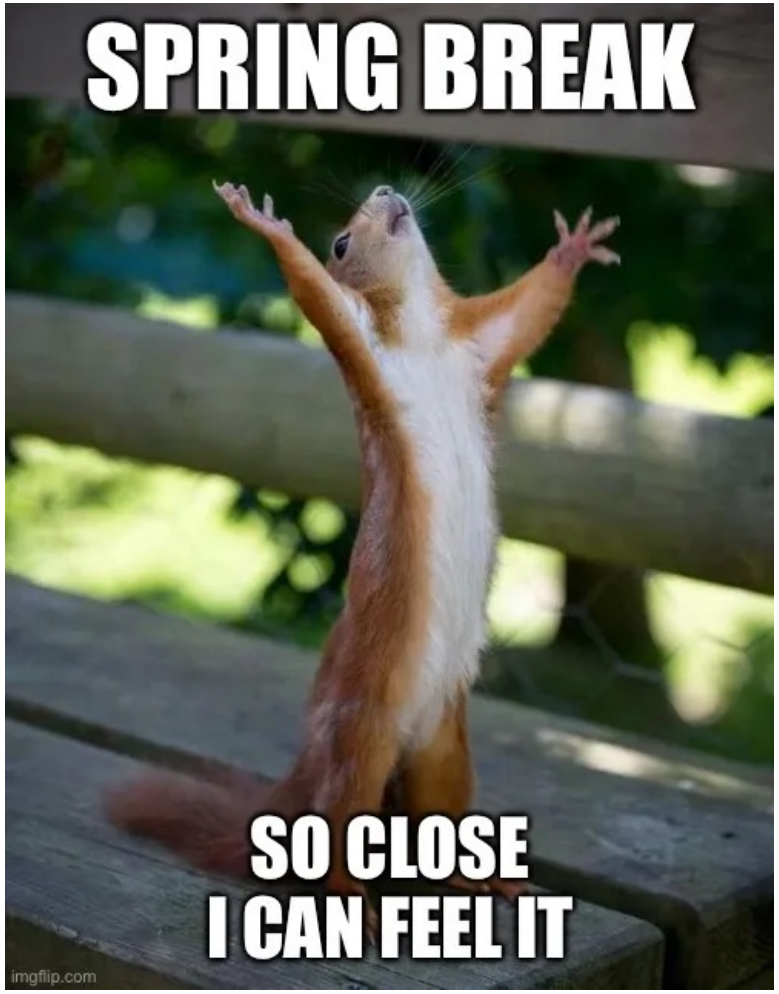
- step 1: stating the hypotheses
  - H_1: $\mu > 71$ and H_0: $\mu \leq 71$
- step 2: set criteria for decision
  - n = 36, $\mu_M = \mu$ = expected value of M = 71
  - $\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{36}} = 3$
  - $z_{critical}$ = +1.65 (only one side), $p_{critical}$ = 0.05
- step 3: collect data
  - $z_{observed} = \frac{M - \mu}{\sigma_M} = \frac{76 - 71}{3} = 1.67$
  - $p_{observed} = .0475$
- step 4: decide
  - $z_{observed} > z_{critical}$
  - we can reject the null hypothesis if we use a one-tailed test!





next time

- **before** class
  - *prep*: textbook chapters
  - *try*: PS4 problems (Ch 8)
  - *apply*: optional meme
- **during** class
  - more on hypothesis testing



https://lolalambchops.com/best-spring-break-memes/