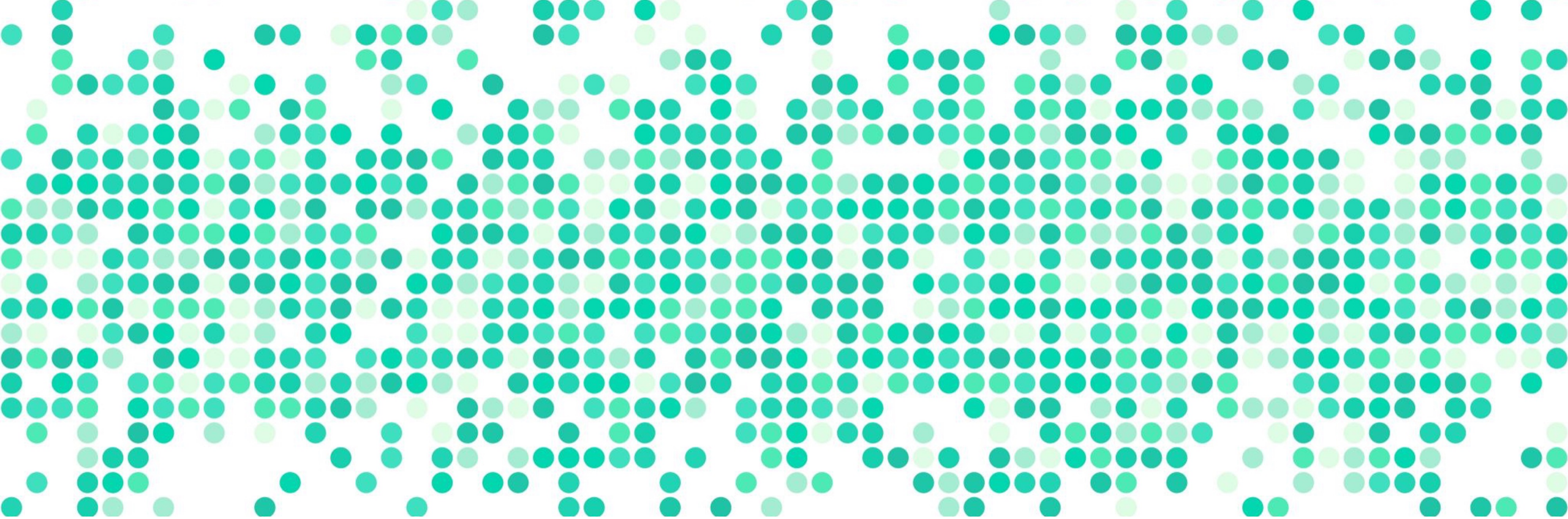


DATA ANALYSIS

Week 7: Hypothesis testing continued...

today's agenda

errors in hypothesis testing

assumptions and extensions

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

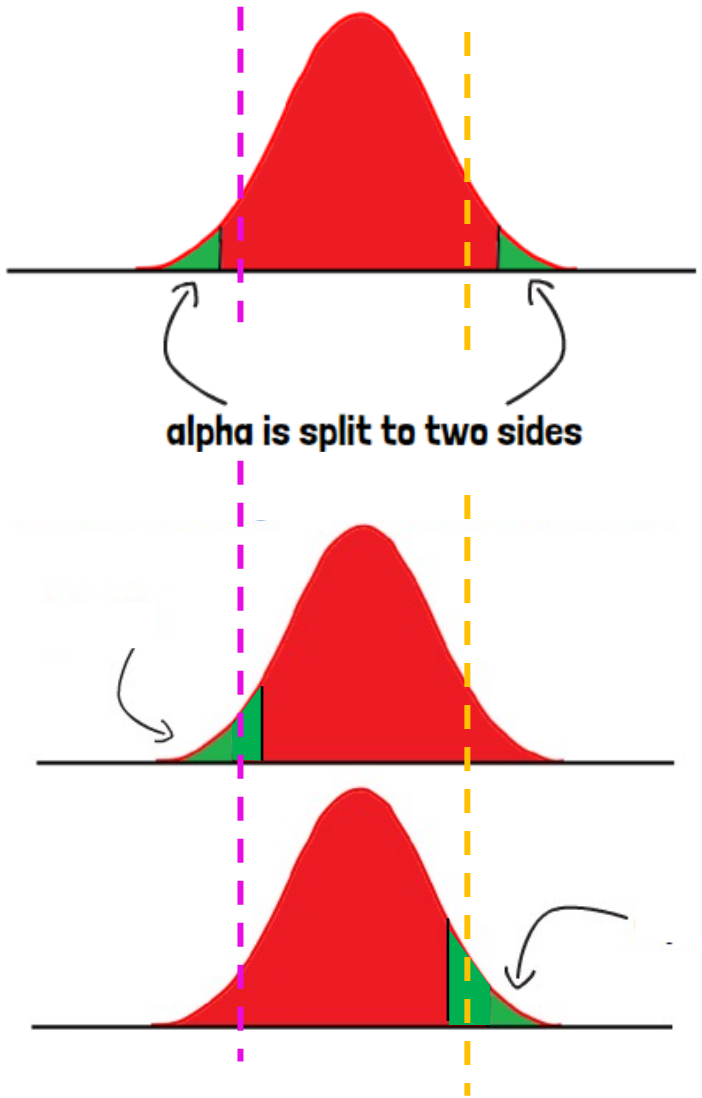
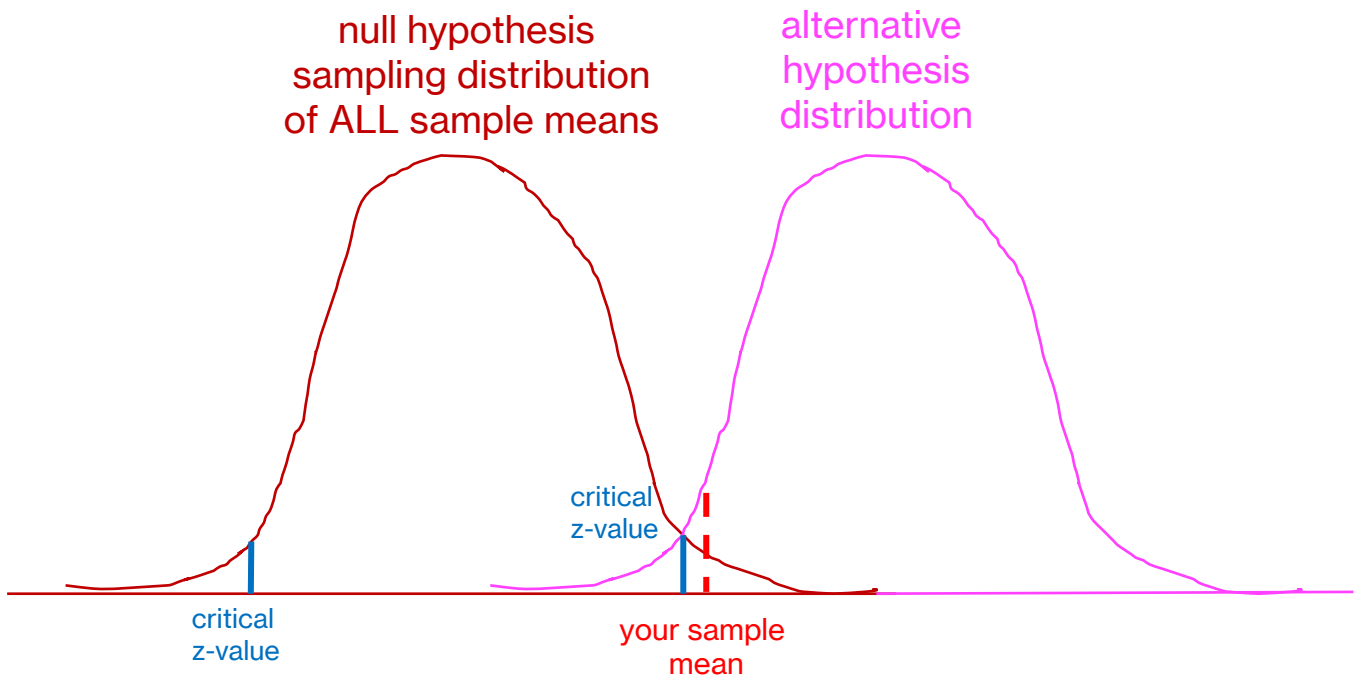


figure source

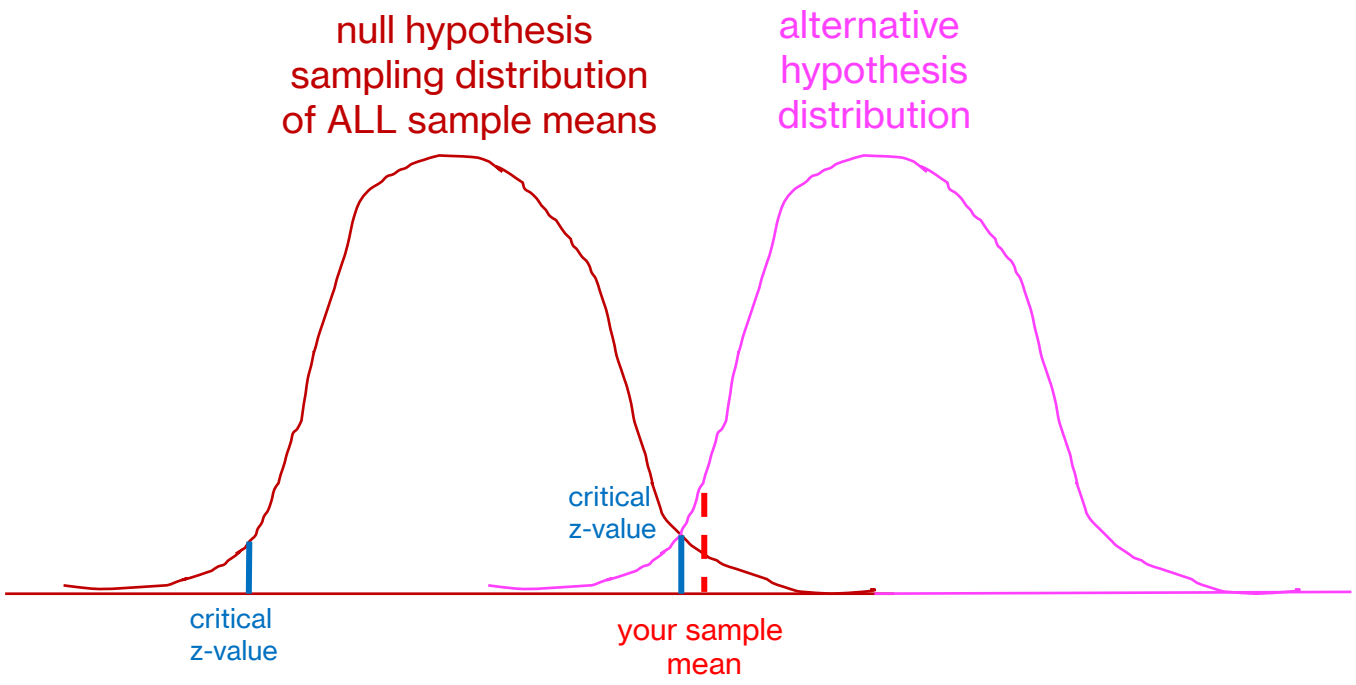
rejecting the null hypothesis

- when a null hypothesis is rejected based on a given α-level, you are making an inference that the sample statistic you have obtained is highly unlikely based on the null hypothesis, i.e., it is not likely to be part of the null distribution
- this implies that the sample statistic you have obtained is consistent with a different sampling distribution, i.e., the one suggested by the alternative hypothesis



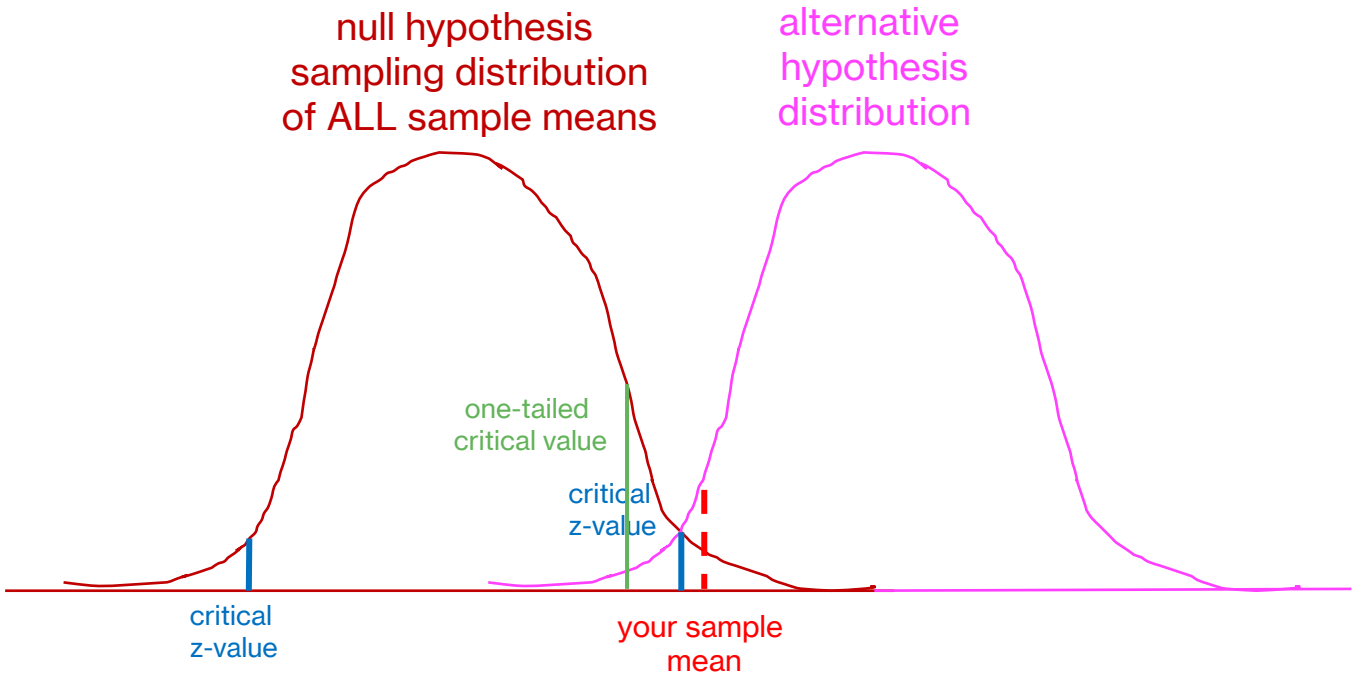
rejecting the null hypothesis

- BUT highly unlikely ≠ impossible!
- rare events are still possible events, i.e., your sample statistic could STILL be part of the null distribution
- For α=.05, there is a 5% chance of obtaining a rare sample that lies in the tails of the null distribution
- there is a 5% chance that you may be making an error when rejecting the null hypothesis in favor of the alternative hypothesis
- this is called a Type I error



one vs. two tailed test

- in a two tailed test, your α-level is split across two sides and your chance of making a Type I error is lower, compared to a one-tailed test
- lower values of α make it harder to reject the null hypothesis!
- you are essentially less likely to reject a null hypothesis in a two-tailed test, i.e., it is more stringent than a one-tailed test
- in practice, you are expected to report the two-tailed test even if you have a directional hypothesis



signal detection and hypothesis testing

- the process of evaluating the effect of a manipulation in experimental research is quite similar to differentiating a signal from noise, an idea with roots in signal detection theory with broad applications

	world truth	
your data	true effect	noise
effect found	hit	false alarm
effect not found	miss	correct rejection

activity: signal vs. noise

- groups of 2
- you will be presented with a research situation and you need to come up with a signal vs. noise formulation of the situation

activity: signal vs. noise

- a researcher is trying to understand if there are gender differences in the ability to manage money. They conduct an experiment where they record the amount of money saved by different genders in a given month. Their hypothesis is that men and women do not save equal amounts of money.

activity: signal vs. noise

- construct the signal vs. noise table for this experiment design

	world truth	
your data	true effect ???	noise ????
effect found ???	hit	false alarm
effect not found ???	miss	correct rejection

activity: signal vs. noise

- construct the signal vs. noise table for this experiment design

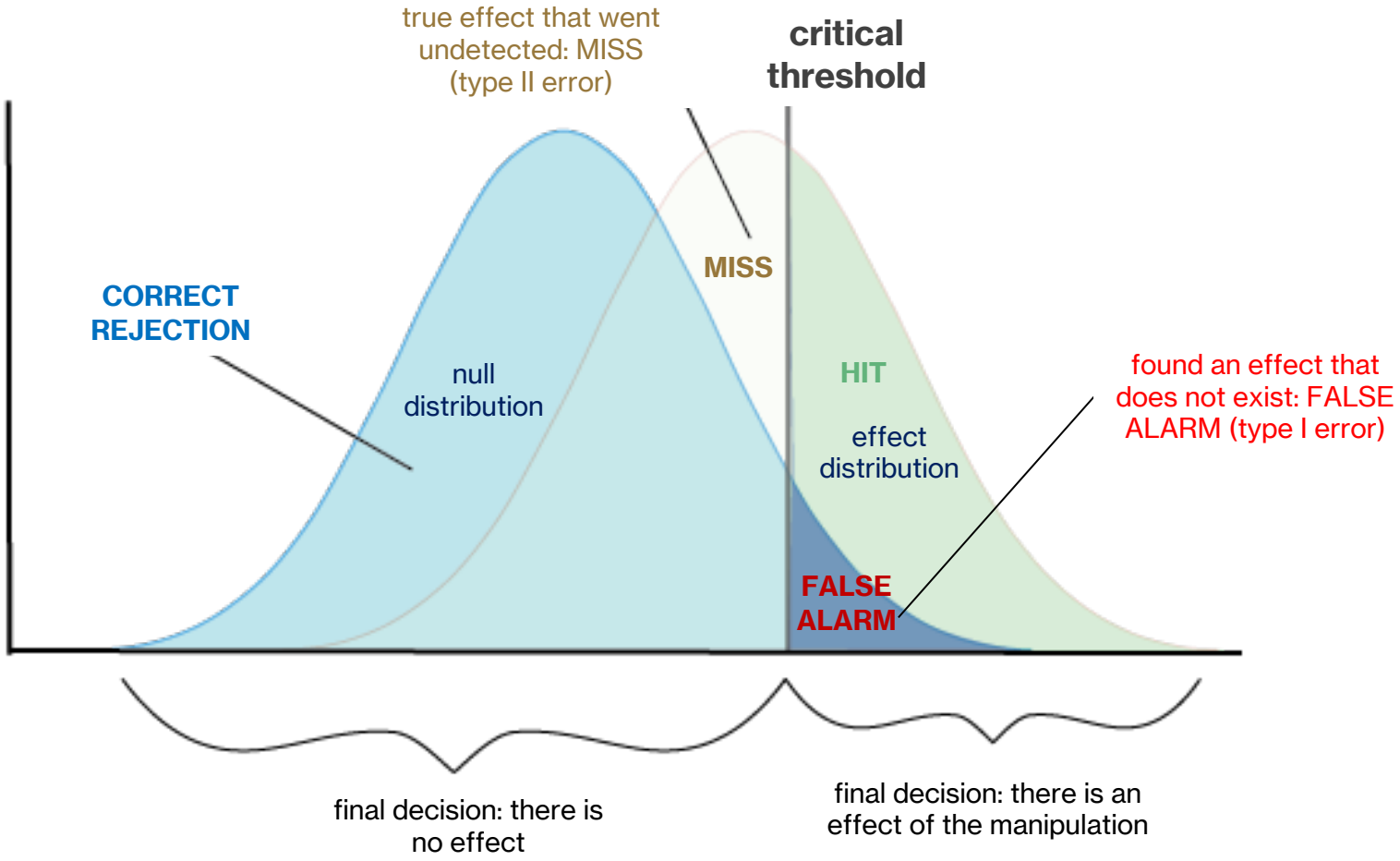
	world truth	
your data	true effect (there are gender differences in the real world)	noise (there are no gender differences)
effect found (the data sample shows gender differences)	hit	false alarm
effect not found (the data sample shows no gender differences)	miss	correct rejection

signal vs. noise

- construct the signal vs. noise table for this experiment design

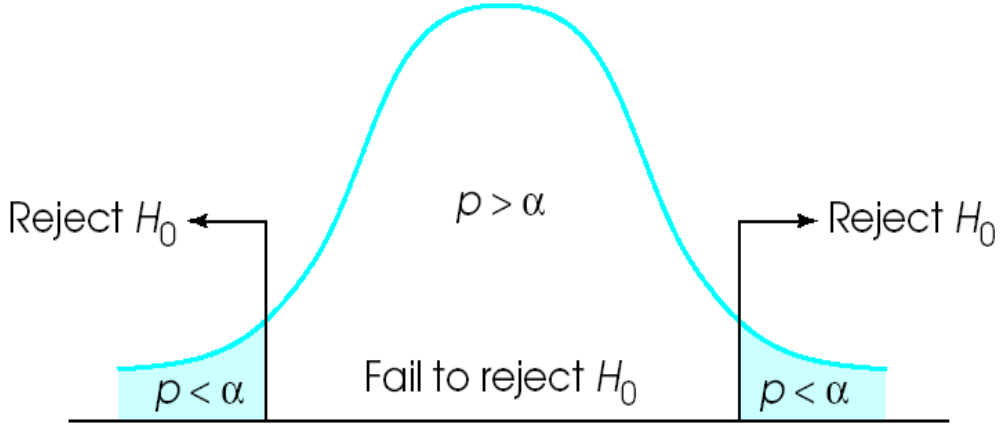
	world truth	
your data	true effect (there are gender differences in the real world)	noise (there are no gender differences)
effect found (the data sample shows gender differences)	hit (power)	false alarm (type I error)
effect not found (the data sample shows no gender differences)	miss (type II error)	correct rejection

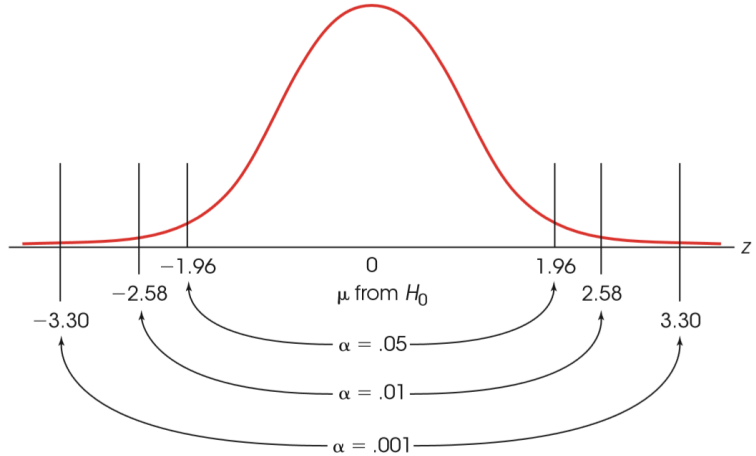
summary of errors



type I error: false alarms

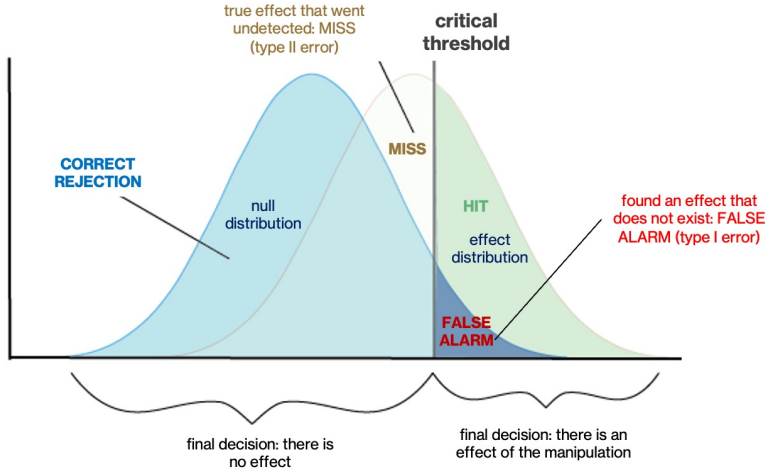
- type I errors occur when H_0 (null hypothesis) is rejected when it should not have been, i.e., your sample is simply a rare one within the null distribution
- what determines the rejection of H_0?
  - α (critical threshold): probability of making a type I error
  - lower the threshold (typically $\alpha < .05$) lower your chance of making a type I error
  - but not too low, else it will be impossible to find evidence to reject H_0

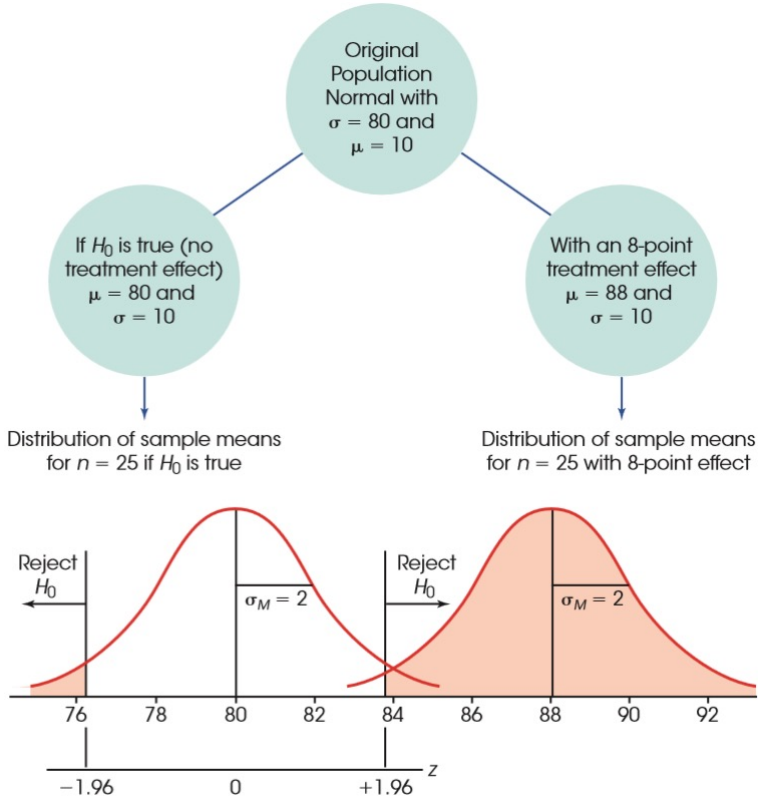




type II error: miss

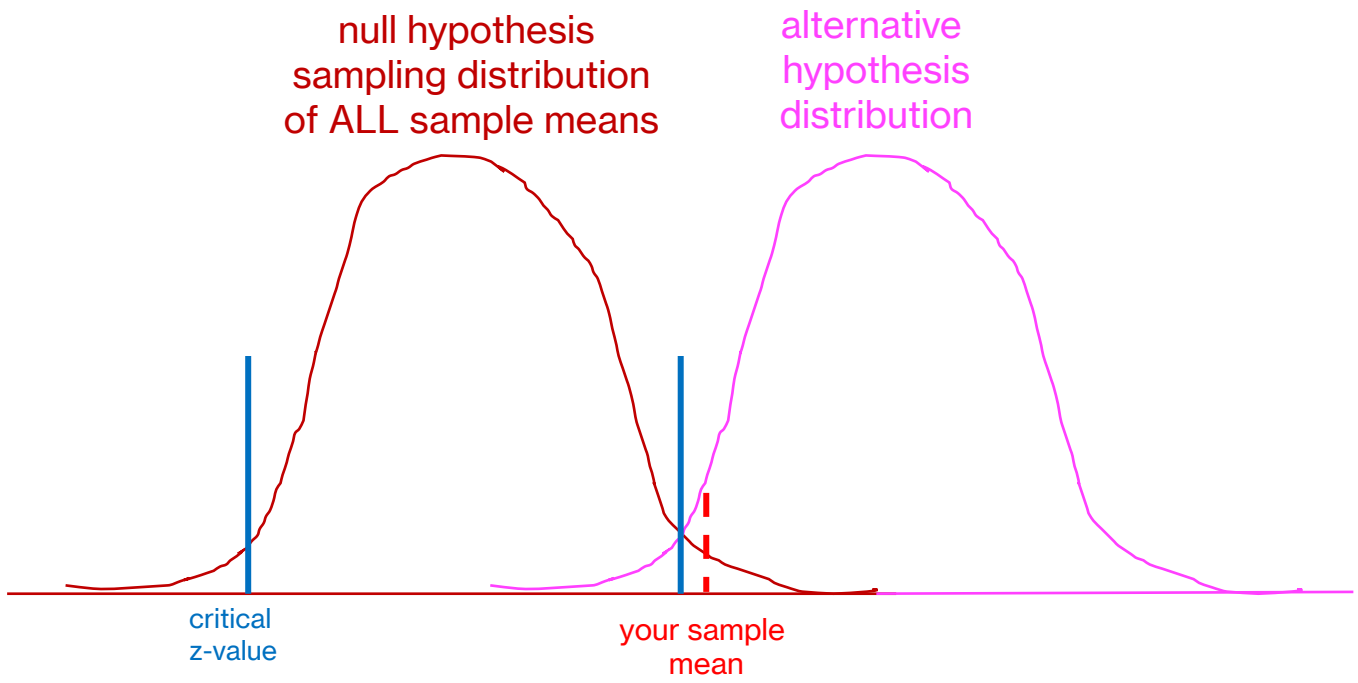
- type II errors occur when you fail to reject H_0 (null hypothesis) while there was an effect, i.e., the sample actually belonged to the alternative hypothesis distribution but was in the overlapping area with the null distribution
- β: probability of making a type II error
- power: $1 - \beta$, probability that a test will correctly reject the null hypothesis
  - typically calculated *before* data collection
  - depends on the (1) α level, (2) one vs. two-tailed test, (3) sample size, (4) effect size,





statistical power: α-level / tails

- increasing the α-level OR using a one-tailed test increases the power because we are more likely to reject the null hypothesis in both cases

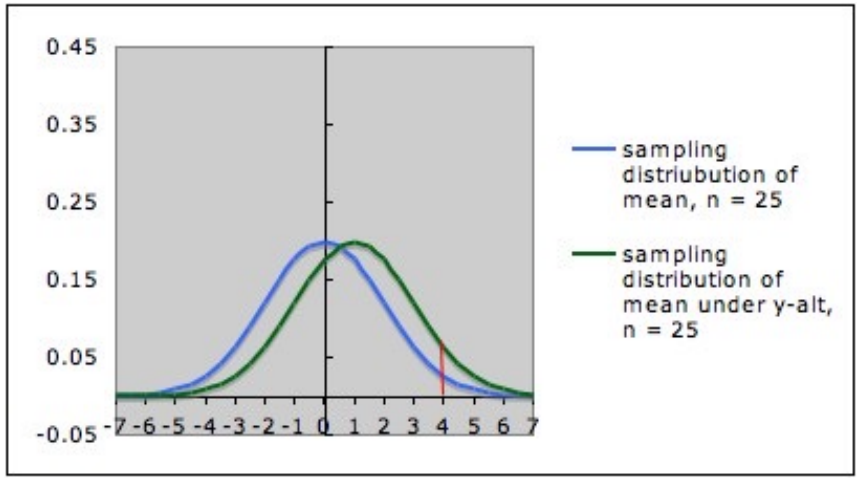


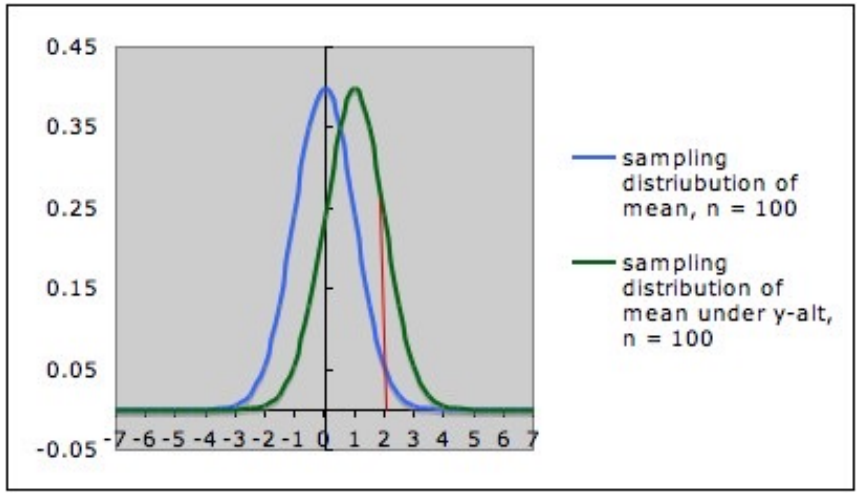
statistical power: sample size

- higher sample size decreases the standard error of the mean (σ_M)

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

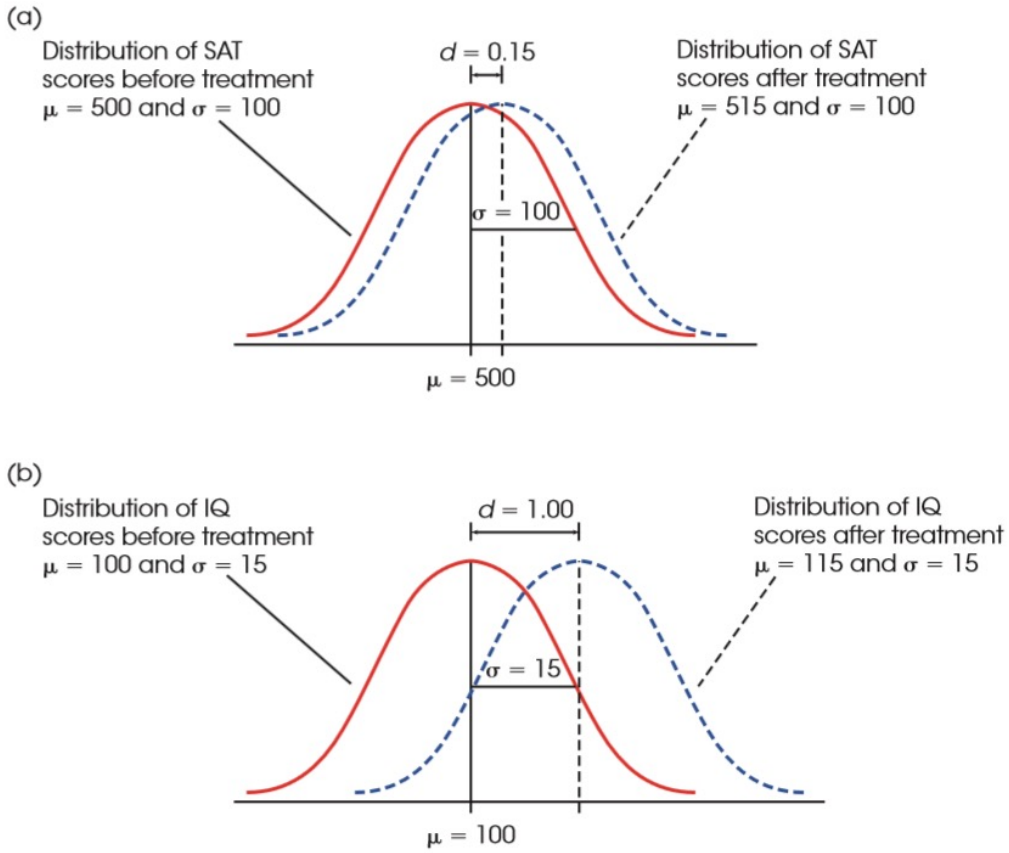
- this leads to narrower distributions, and decreasing the overlap between the two distributions, again leading to higher power





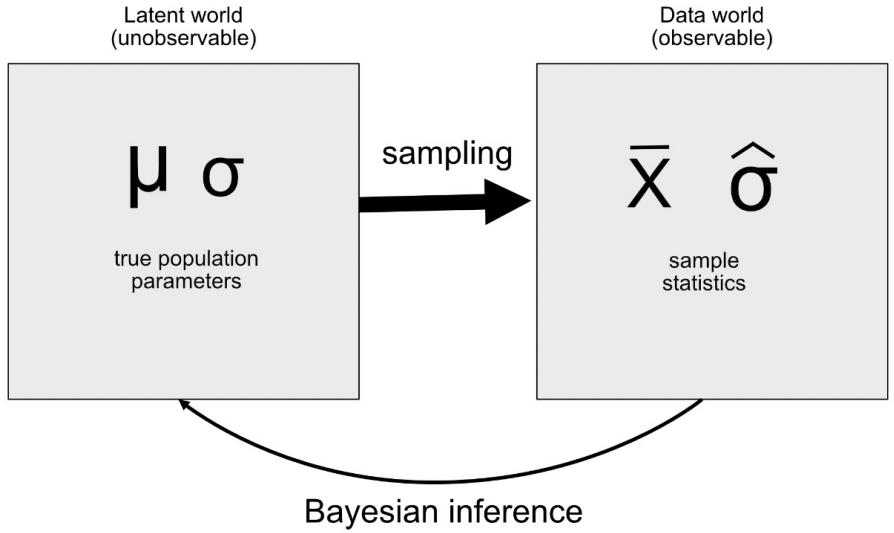
statistical power: effect size

- if the separation between the two distributions is greater to begin with, power is higher
  - this is called effect size!
- effect size is an estimate of the difference between what would be expected by chance (null) vs. what is observed due to our manipulation (effect), irrespective of the sample size
- Cohen's *d* is a measure of how different the observed mean is relative to the population mean
- populations: $d = \frac{mean\ difference}{standard\ deviation} = \frac{\mu_{alternative} - \mu_{null}}{\sigma}$
- samples: $estimated\ d = \frac{M-\mu}{s}$



other issues

- under NHST, we never actually test the likelihood of our hypothesis!
  - we obtain P(data | null hypothesis)
  - we want P(alternative hypothesis | data)
- NHST is limited because you cannot make actual inferences about the hypothesis you want to test
- alternative framework: Bayesian statistics!
  - Bayesian statistics allows you to evaluate P(alternative hypothesis | data)

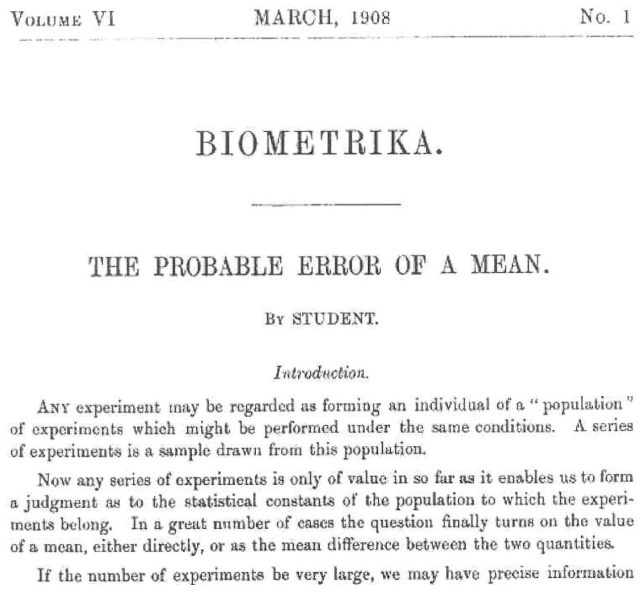


z to t-distribution

- inferential statistics = from samples to populations
- but...we need information about the population to make inferences, i.e., we need to know the mean (μ) and standard deviation (σ) of the population! this information is usually not known!
- when σ is unknown and we have to rely on sample standard deviation (s) as an estimator, we cannot use the normal distribution as our sampling distribution

$$\sigma_M = \frac{\sigma}{\sqrt{n}} \text{ vs. } s_M = \frac{s}{\sqrt{n}}$$

- we instead use the student's *t* distribution
- originally employed by Guinness Brewery, Dublin, Ireland for dealing with small samples in brewing quality control



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

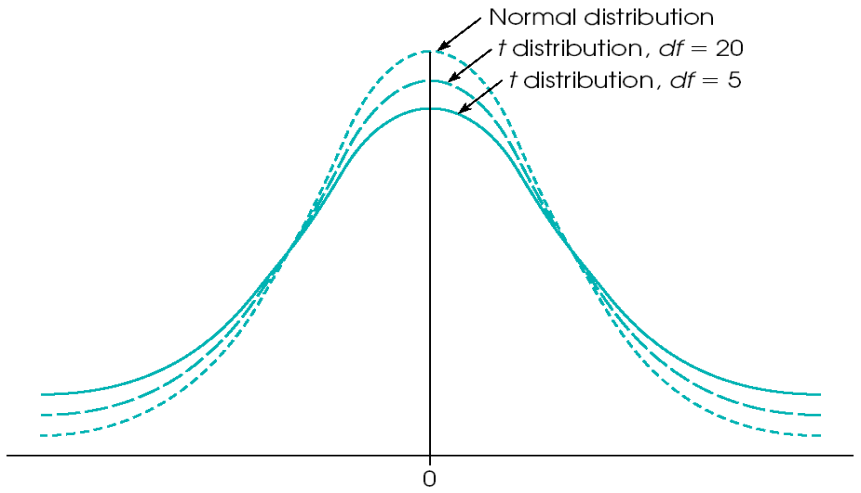
If the number of experiments be very large, we may have precise information

t-statistic

- the t-statistic approximates the z-score

$$t = \frac{M - \mu}{s_M} = \frac{sample\ statistic - population\ parameter}{standard\ error}$$

- how good is this approximation? depends on the sample size (n)
  - each t-curve is defined by degrees of freedom, *df* = n-1
  - for large dfs, the t distribution approximates the normal distribution
- hypothesis testing with t-distributions
  - instead of a z-score, you now calculate a t-score!



z-test vs. one-sample t-test

z-tests

- **when**: population mean and standard deviation are known
- **want to compare**: sample mean to population mean

one sample t-test

- **when**: population standard deviation is unknown
- **want to compare**: sample mean to population mean

example (Ch9-Q23a)

- research examining the effects of preschool childcare has found that children who spent time in day care, especially high-quality day care, perform better on math and language tests than children who stay home with their mothers (Broberg, Wessels, Lamb, & Hwang, 1997). In a typical study, a researcher obtains a sample of n = 10 children who attended day care before starting school. The children are given a standardized math test for which the population mean is $\mu = 50$. The scores for the sample are as follows: 53, 57, 61, 49, 52, 56, 58, 62, 51, 56.
- Is this sample sufficient to conclude that the children with a history of preschool day care are significantly different from the general population? Use a two-tailed test with $\alpha = .01$.
- use data here

framing the problem

- population
  - mean μ = 50
  - standard deviation (σ) unknown, we cannot use a z-test
- sample
  - n = 10, sample scores: X = 53, 57, 61, 49, 52, 56, 58, 62, 51, 56
  - sample mean: $M = \frac{\sum X}{n}$ = 55.5
  - sample standard deviation: $s = \sqrt{\frac{(X-M)^2}{n-1}} = 4.249183$
- we need to use a t-test!
  - degrees of freedom: $df = n - 1 = 9$
  - $s_M = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{(4.249183)^2}{10}} = 1.34371$

sheets solution

hypothesis testing

- step 1: stating the hypotheses
  - H_0: $\mu = 50$; H_1: $\mu \neq 50$
- step 2: setting decision criteria
  - two-tailed test (t-value calculator)
  - $t_{critical}(9) = \pm 2.2626$ for $\alpha = .05$
- step 3: collect data
  - $t_{observed} = \frac{M - \mu}{s_M} = \frac{55.5 - 50}{1.34371} = 4.09$
  - $p_{observed} = 0.0027$ (obtained from p-value calculator for t-score)
- step 4: decide!
  - $p_{observed} < .05 \; and \; t_{observed} > t_{critical}(9)$
  - reject H_0 and conclude that children with a history of preschool day care are significantly different from the general population, t (9) = 4.09, p = .003.

sheets solution

example (Ch9-Q23b)

- research examining the effects of preschool childcare has found that children who spent time in day care, especially high-quality day care, perform better on math and language tests than children who stay home with their mothers (Broberg, Wessels, Lamb, & Hwang, 1997). In a typical study, a researcher obtains a sample of n = 10 children who attended day care before starting school. The children are given a standardized math test for which the population mean is $\mu = 50$. The scores for the sample are as follows: 53, 57, 61, 49, 52, 56, 58, 62, 51, 56.
- Compute Cohen's d to measure the size of the preschool effect.

example (Ch9-Q23b)

- $estimated\ d = \frac{M-\mu}{s}$
- population mean μ = 50
- sample mean: $M = \frac{\sum X}{n}$ = 55.5
- sample standard deviation: $s = \sqrt{\frac{(X-M)^2}{n-1}} = 4.249183$
- $d = \frac{M-\mu}{s} = \frac{55.5-50}{4.249} = 1.29$
- Math achievement scores for children with a history of preschool day care are significantly different from the general population, *t* (9) = 4.09, *p* = .003, *d* = 1.29.

where are we going next?

- what kinds of sample statistics have we examined so far?
  - means / medians / modes
  - correlations / slopes
- what might the null/alternative hypotheses look like for these?
  - mean: $H_o: \mu = 0$ and $H_1: \mu \neq 0$
  - correlation: $H_o: \rho = 0$ and $H_1: \rho \neq 0$
  - regression: $H_o: b = 0$ and $H_1: b \neq 0$
- in coming weeks, we will find the underlying sampling distribution and P (data | null)!

happy spring break!

- **before you go**
  - *finish*: Week 7 quiz
  - *submit*: PS4 or opt-out (due March 12)
  - *apply*: optional meme