Week 4: Frequentism, hypothesis tests, & P-values

ANTH 674: Research Design & Analysis in Anthropology

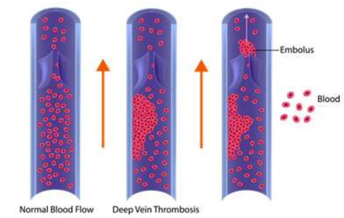
Professor Andrew Du

Andrew.Du2@colostate.edu

1

Statistical vignette

On October 18, 1995, UK Committee on Safety of Medicines (CSM) issued a letter, warning: “New evidence has become available, indicating that the chance of a **thrombosis** occurring in a vein is increased around **two-fold** for some types of [contraceptive] pills compared with others.”



2

Statistical vignette

On October 18, 1995, UK Committee on Safety

Government Warns Some Birth Control Pills May Cause Blood Clots

EDITH M. LEDERER October 19, 1995

Click to copy

LONDON (AP) _ The government warned Thursday that a new type of birth control pill used by 1.5 million British women may cause blood clots, according to new, unpublished studies.

https://apnews.com/article/4e07b291c887bb4b5659beab81ffb015

3

What happened?

- 12% of users stopped taking the pill or switched
- In England & Wales, 26,000 more babies conceived in 1996 compared to 1995
- 13,600 more abortions in 1996 than 1995
- All in all, total number of prevented embolism deaths was estimated to be…”possibly one”
- Risk of thrombosis w/ other pills: 1 in 7,000
- Doubled with newer, riskier pills: 2 in 7,000

4

What happened?

- 12% of users stopped taking the pill or switched
- In Eng… concе…s
- 13,60…95
- All in all, total number of prevented embolism deaths was estimated to be…"possibly one"
- Risk of thrombosis w/ other pills: 1 in 7,000
- Doubled with newer, riskier pills: 2 in 7,000

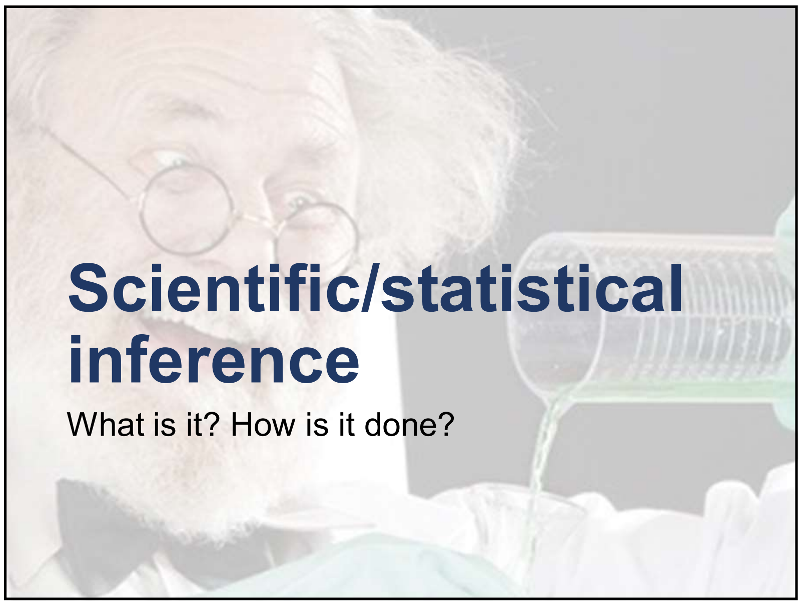
Statistical significance
≠
Scientific significance!

5

Lecture outline

1. Quick intro: scientific/statistical inference
2. The frequentist perspective
3. Hypothesis tests
   1. Confidence intervals
   2. P-values

6



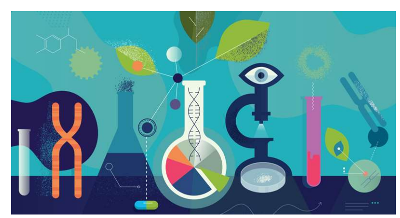
Scientific/statistical inference

What is it? How is it done?

7

What is scientific inference?

- Inferring something about some **LARGER** process or pattern using a **SMALLER** sample of data
- A key step in this endeavor is statistical inference



8

What is statistical inference?

From Wikipedia:

- Process of using data analysis to [infer] properties of an *underlying distribution of probability*
- Inferential statistical analysis infers properties of a *population*, for example by *testing hypotheses* and *deriving estimates*
- It is assumed that the observed data set is sampled from a *larger population*

9

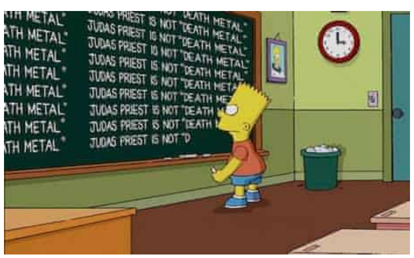
The frequentist perspective

What is it? How is it used in probability and inferential statistics?

10

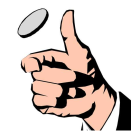
What is frequentism?

- Emphasizes frequency of some event or measure, repeated ***MANY*** times over



11

An example: coin flips

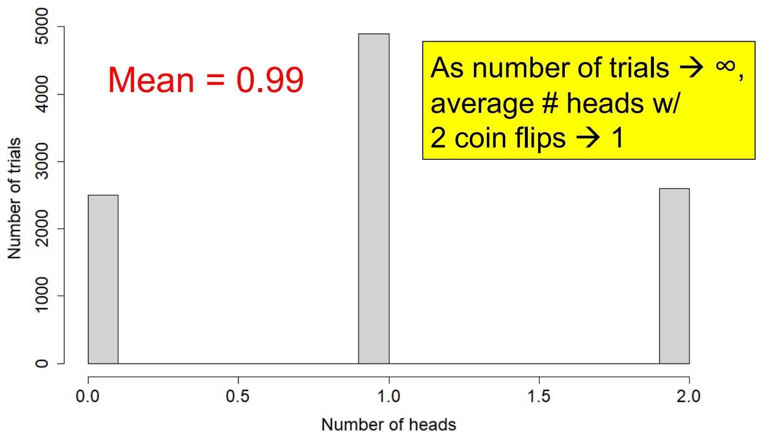


- What is the probability of getting heads?
- With two flips, would you be surprised if you got {H,H} or {T,T}?
- What if I did two coin flips, and repeated this 10,000 times (i.e., 10,000 trials or *replicates*)?

{H,T}, {T,H}, {H,H}, {T,H}, {T,H} ……

12

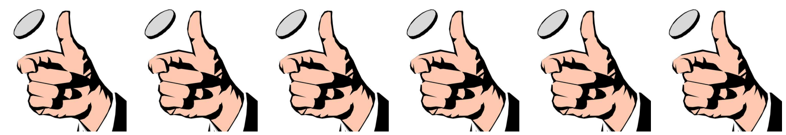
Two coin flips, repeated 10,000x



13

Frequentist definition of probability

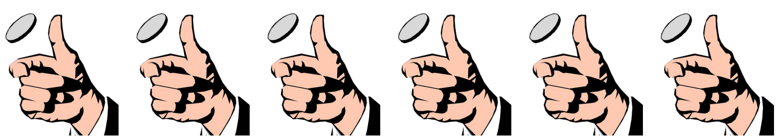
- Relative frequency of some outcome based on an infinitely large number of trials
- How quickly observed frequency converges on true frequency depends on how variable underlying process/pattern is



14

Can also flip a coin many, many times (one trial)

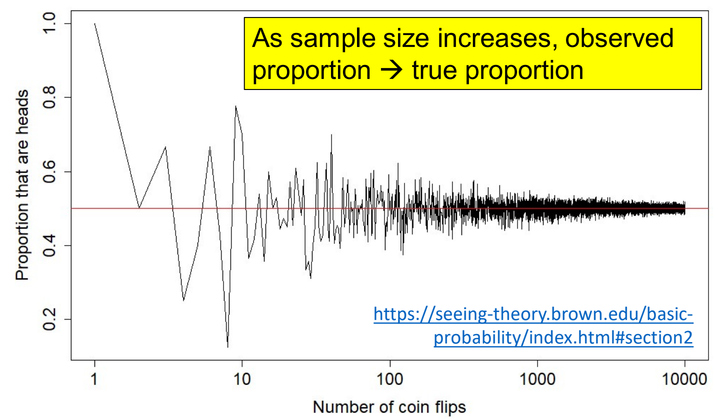
- I flipped a coin 10,000 times and got 5,067 heads (0.5067)



https://seeing-theory.brown.edu/basic-probability/index.html#section1

15

One trial each



16

Estimating true probability

- Get an accurate estimate if you take the average of many, many trials (e.g., two flips & 5,000 trials)
- Or have large enough sample size (e.g., 10,000 flips & one trial)

Law of large numbers: with large numbers, get a good estimate of the true value (i.e., *parameter*)

17

Questions?

18

Frequentism: measurements

- What if we wanted to know the average height of all human adults on Earth?
- How would we go about figuring this out (not feasible to measure every single person)?

19

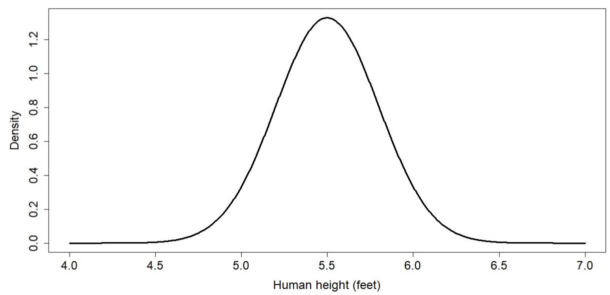
Frequentism: measurements

- Can measure a large, representative **sample** of people from the **larger population** of interest (i.e., **statistical population**)
- Can measure a representative smaller **sample** from the **statistical population** w/ multiple trials/replicates

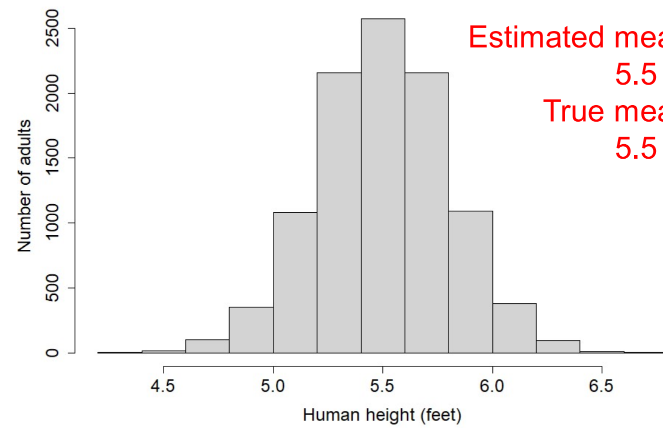
20

Average height of all adults

- Let’s say we’re all-knowing beings who know the true mean of the statistical population of heights is **<u>5.5 feet</u>** w/ a SD of 0.3 feet



If I sample 10,000 adults (one trial/replicate)



Estimated mean = 5.5 feet
True mean = 5.5 feet

If I sample 100 adults (100 trials/replicates)

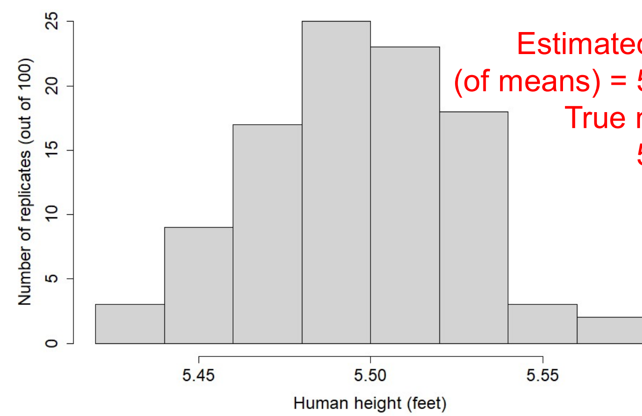
Length of vector = 100

1. {5.3, 5.5, 5.6, 5.6, 5.8, 5.1, 5.1, …}; mean = 5.4
2. {5.8, 5.5, 5.0, 5.5, 5.6, 5.2, 5.4, …}; mean = 5.5
3. {6.0, 5.6, 5.4, 5.6, 4.9, 5.4, 4.9, …}; mean = 5.5

⋮

100. {5.8, 5.0, 5.5, 5.1, 5.5, 6.1, 5.2..}; mean = 5.5

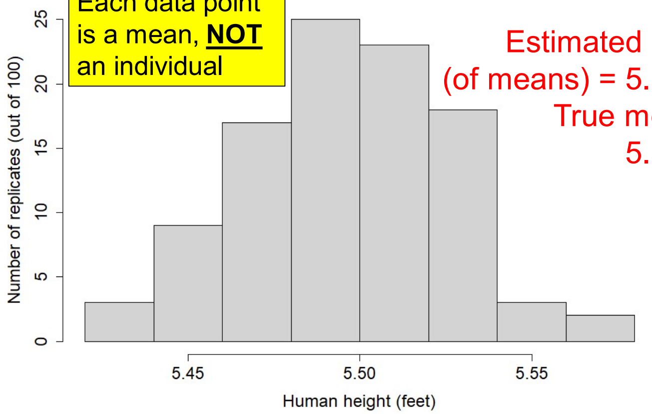
If I sample 100 adults (100 trials/replicates)



Estimated mean (of means) = 5.5 feet
True mean = 5.5 feet

Sampling distribution of mean

Each data point is a mean, **NOT** an individual

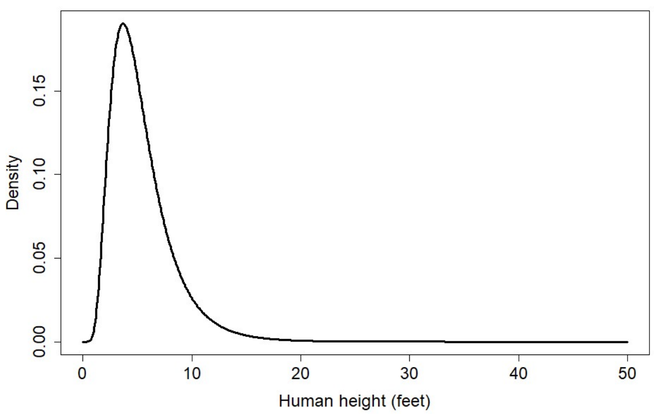


Estimated mean (of means) = 5.5 feet
True mean = 5.5 feet

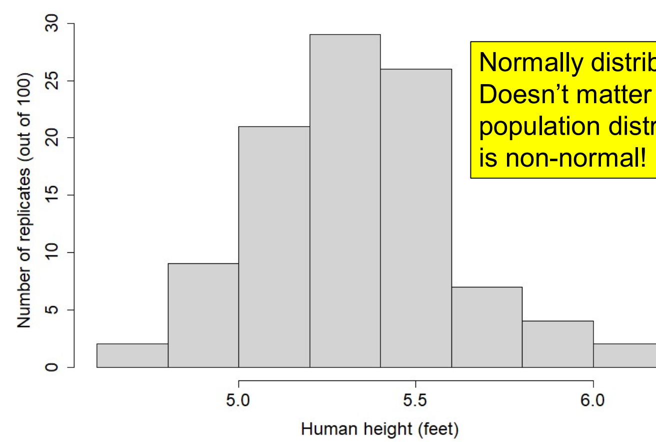
What is a sampling distribution?

- Distribution of a **statistic** (e.g., mean, median) obtained from a large number of samples drawn from the population
- Sampling distributions lie at the heart of statistical inference!

What if human height was lognormally distributed?



Sampling distribution of mean



Normally distributed! Doesn't matter if population distribution is non-normal!

Central limit theorem

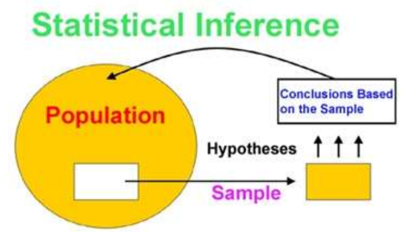
- Sampling distribution of mean (or sum) approximates a normal distribution as sample size gets larger, **no matter what the population distribution is**
- Many statistics rely on normally distributed sampling distributions, so this is a good thing!

https://seeing-theory.brown.edu/probability-distributions/index.html#section3

29

Goals of statistical inference

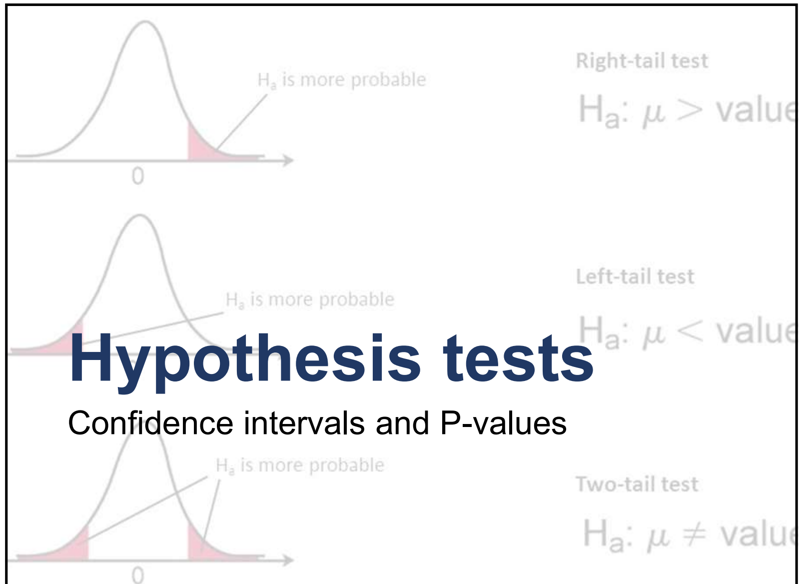
- To understand properties of some larger statistical population by analyzing a smaller sample from said population
- Key in this process is using sampling distributions to test hypotheses



30

Questions?

31



Hypothesis tests

Confidence intervals and P-values

32

What is a hypothesis?

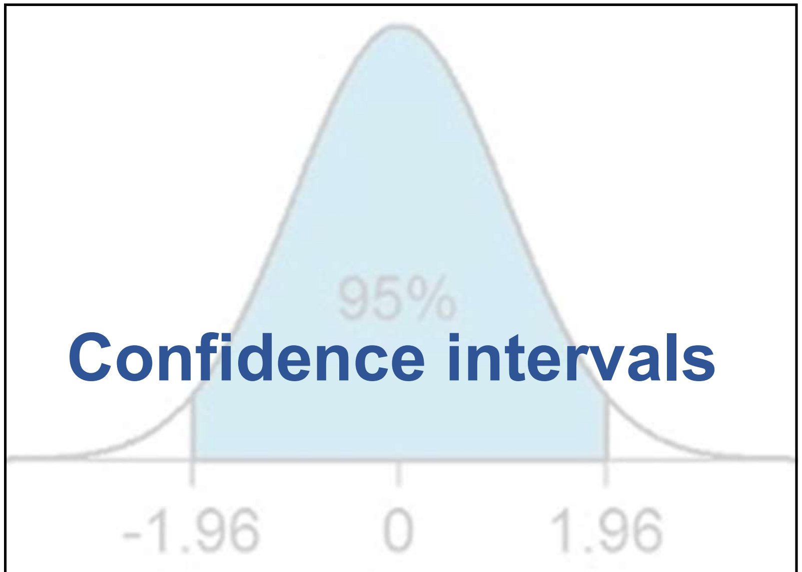
- A proposed, falsifiable explanation for some observation
- E.g., "Because of better nutrition, this group of adults should be 5.55 feet on average "
- Falsified if mean height is not 5.55 feet

33

Testing hypotheses w/ statistics

- Formalized way of comparing data to expectations derived from hypothesis
- Are the data consistent with or significantly different from expectations?
- Evaluated using confidence intervals and P-values (which derive from sampling distributions)

34



Confidence intervals

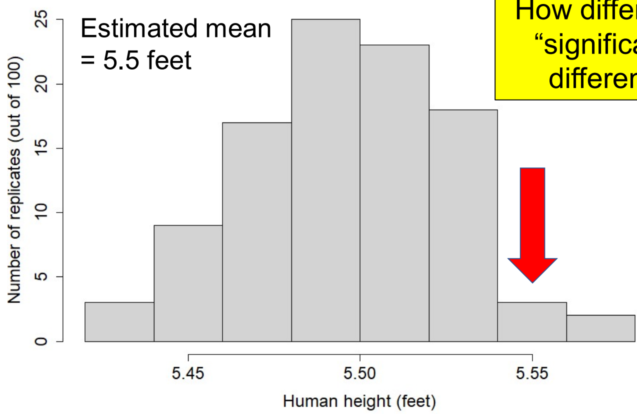
35

What are confidence intervals (CI)?

- Wikipedia: range of plausible values for an unknown parameter (e.g., population mean)
- CI use sampling distribution to quantify the variability around estimate of true parameter
- CI used to evaluate, e.g., if our estimated mean height from sample is consistent with or significantly different from our expectation of 5.55 feet

36

Sampling distribution of mean (from previous example)



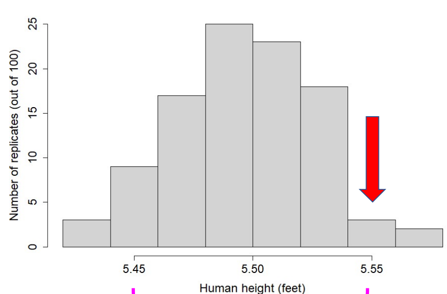
How different is "significantly different"?

Significance level (α)

- α = 0.05 in most sciences, though the cut-off is ultimately arbitrary
- Means we will falsely reject hypothesis 5% of the time if hypothesis is true
- α = 0.05 translates to 95% confidence intervals (CI) and $P < 0.05$
- 95% CI circumscribe the middle 95% of sampling distribution

Sampling distribution of mean (from previous example)

Estimated mean = 5.5



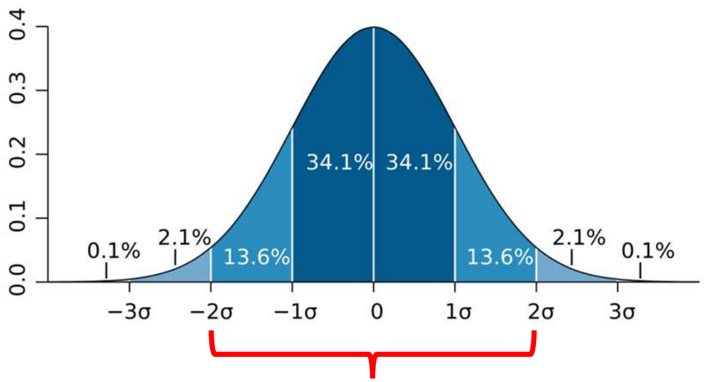
95% CI = 5.442 – 5.546

- Our hypothesis of 5.55 feet is rejected, since it falls outside the 95% CI of the mean (i.e., it is an unlikely value)
- Statistics formalizes the falsification of hypotheses!

How to calculate 95% CI of mean?

- Saving the sampling distribution as object x
- `quantile(x, c(0.025, 0.975))`
  `# middle 95% of sampling dist.`
- Standard deviation of sampling distribution is known as the *standard error* (SE)
- `se = sd(x);`
  95% CI = `mean(x)-1.96*se`, `mean(x)+1.96*se`
- Remember the 68–95–99.7 rule?

68–95–99.7 rule



± 2 SD = middle 95.4% of distribution

What if we don't have replicates?

- Sampling distribution of mean height created by taking 100 different samples from the population and calculating the mean each time
- In practice, we usually have only one sample/replicate
- E.g., fossil record gives us only one sample of *Australopithecus afarensis* crania; how do we calculate 95% CI of ECV with that one replicate?

Calculating standard error with one replicate

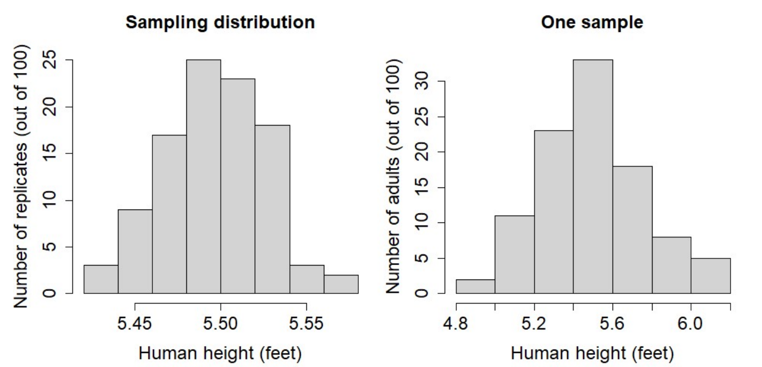
- Before computers, very smart statisticians developed ways to estimate properties of sampling distribution using one sample only
- Standard error of the mean can be calculated as:

Standard deviation of sample

$$\frac{sd}{\sqrt{n}}$$

Sample size

Calculating standard error (SE)



SE = SD = 0.03

SE = SD / √n = 0.3 / √100 = 0.03

Standard error ≠ standard deviation!

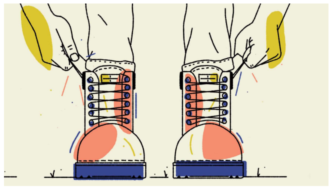
- Standard deviation is SD of **one** sample of observations
- Standard error is SD of **sampling distribution** of statistic

$$se = \frac{sd}{\sqrt{n}}$$

45

These days, can bootstrap

- Resampling your data with replacement
- This is a *Monte Carlo method*: class of methods that use repeated **random sampling** to obtain numerical results
- Only possible with powerful computers



46

Bootstrapping CIs

https://seeing-theory.brown.edu/frequentist-inference/index.html#section3

- Sample is treated as proxy for population & resampling mimics multiple samples from pop.

Data = {5.6, 5.0, 5.6, 5.2, 5.5, 5.9, 5.6, 5.3}

#1: {5.9, 5.6, 5.6, 5.2, 5.6, 5.6, 5.3, 5.3} mean = 5.5
#2: {5.9, 5.6, 5.3, 5.2, 5.0, 5.6, 5.6, 5.5} mean = 5.5
#3: {5.2, 5.3, 5.9, 5.6, 5.6, 5.6, 5.9, 5.9} mean = 5.6

Repeat many, many times (at least 1,000x)

- Get a sampling distribution of means

47

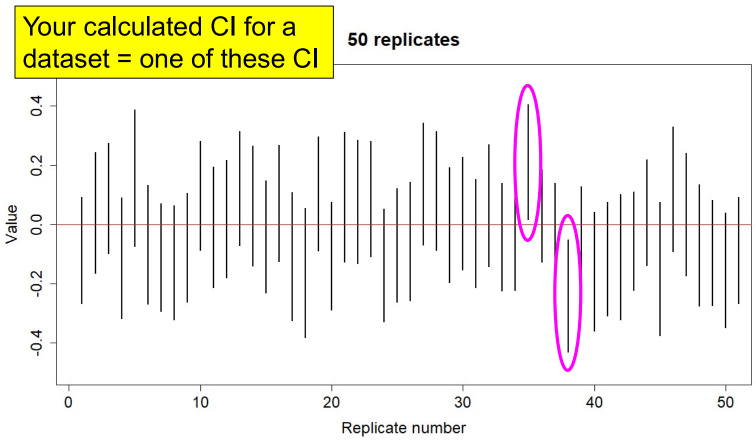
How to interpret confidence intervals?

- Quantifies variability around some estimated parameter (e.g., true mean of population)
- "95% probability that true parameter value is in 95% CI"
- **WRONG**
- For every CI, the fixed, true parameter value is either inside or not
- 5% of replicated confidence intervals will miss the true parameter value

48

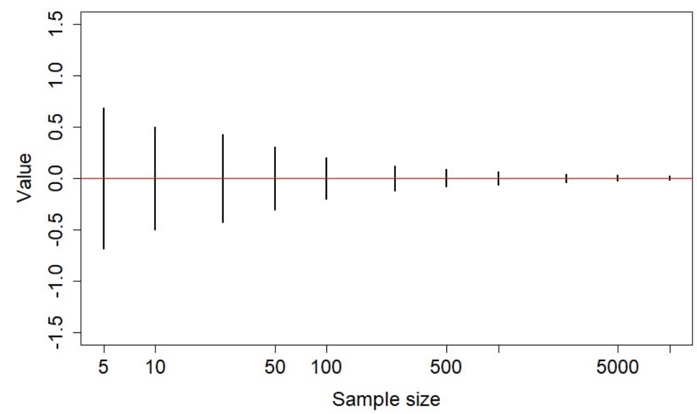
https://seeing-theory.brown.edu/frequentist-inference/index.html#section2

True mean = 0



49

Confidence intervals & sample size



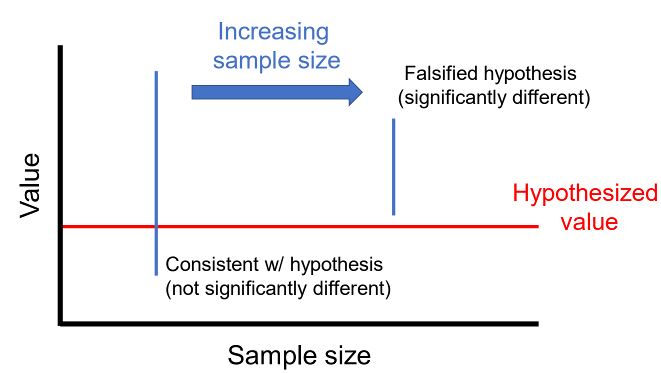
50

Confidence intervals & sample size

- CI shrinks as sample size increases due to law of large numbers
- With larger sample size, parameter estimate will be significantly different from virtually every number except the true value
- Emphasizes that a value falling within CI is not evidence that parameter is that value (might exclude it w/ larger sample size)
- Smaller CI better for demonstrating consistency with some hypothesis

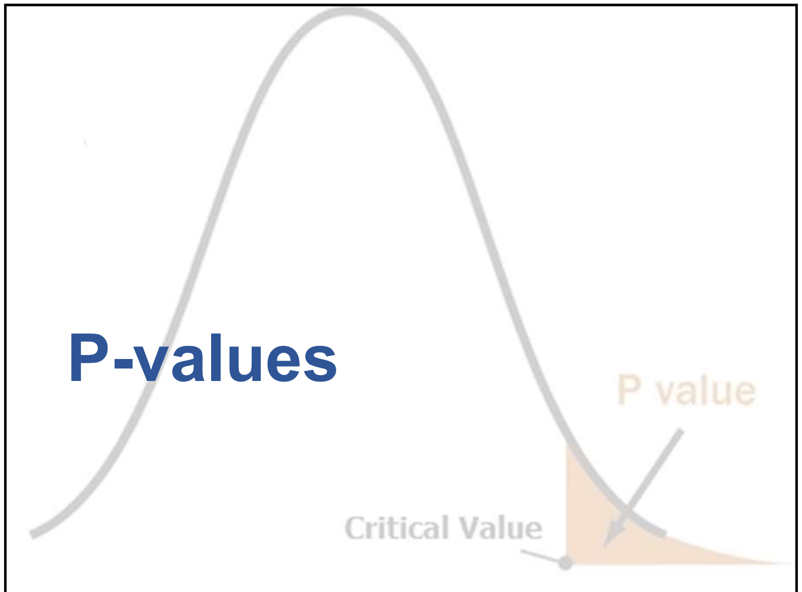
51

Confidence intervals & sample size



52

Questions?



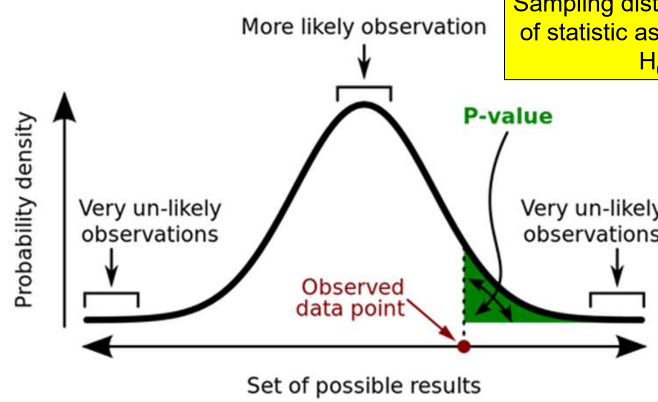
P-values

What are P-values?

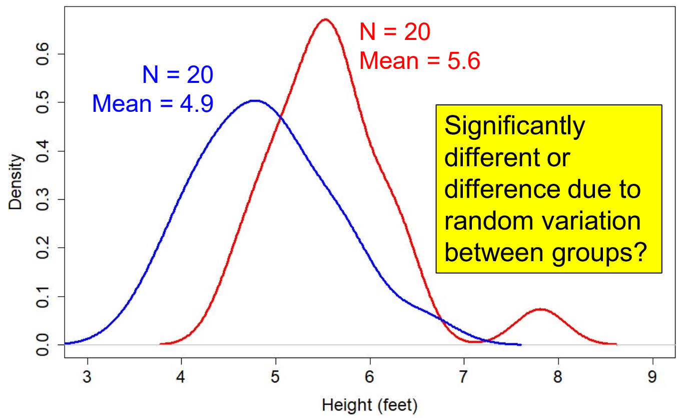
- Quantifies the probability of getting the data (or more extreme), given some hypothesis
- P(D | H)
- Direct test of hypothesis! Hypothesis is falsified if obtaining the data are unlikely ($P < 0.05$)
- Most common hypothesis is a *null hypothesis* (H_0), e.g., true mean = 0, mean difference between groups = 0
- SUPER common in research, but commonly misunderstood!

What are P-values?

Sampling distribution of statistic assuming H_0 is true



Example: mean difference between groups



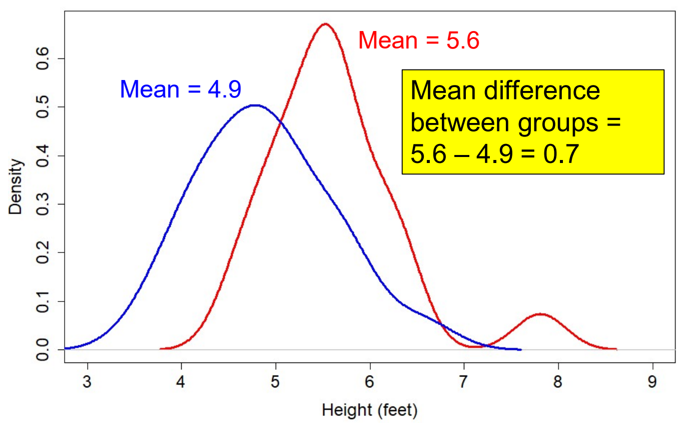
57

Recipe for null hypothesis tests

1. Pick a test statistic (e.g., mean difference)
2. *Assume* null hypothesis is true (e.g., groups come from same population)
3. Create the null sampling distribution of test statistic, given Step 2
4. Calculate the probability of getting the observed test statistic or more extreme (i.e., tail probability), given the null distribution (i.e, P-value)

58

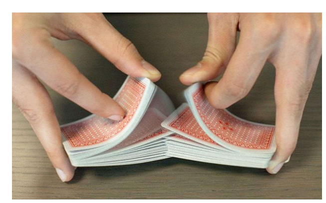
1. Pick a test statistic



59

Steps 2 & 3

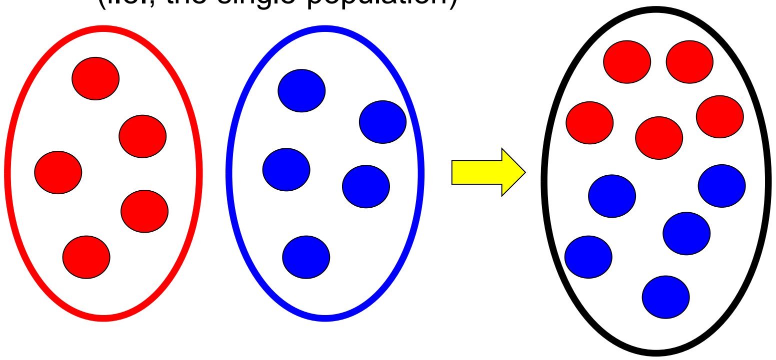
- Create sampling distribution of mean difference, assuming null hypothesis is true (i.e., groups sample same population)
- How do we do that?
- Can use Monte Carlo methods (bootstrap)



60

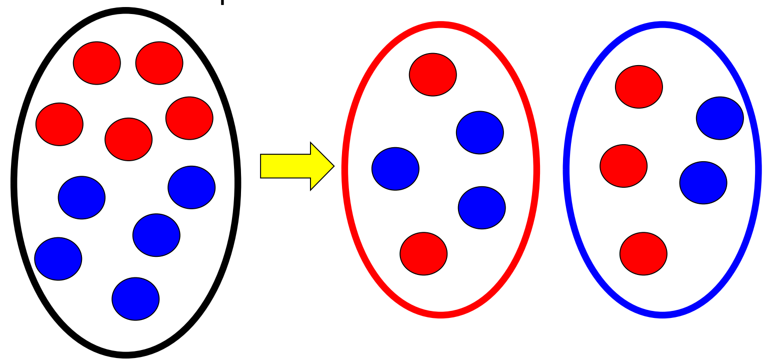
Creating the null distribution

1. Combine groups' vectors into one vector (i.e., the single population)



Creating the null distribution

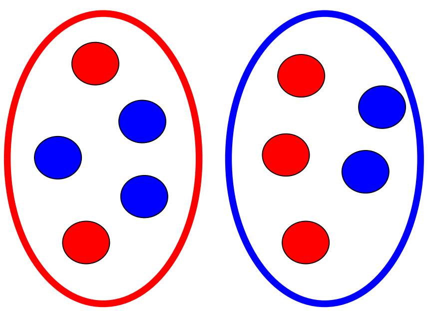
2. Randomly sample (with replacement) from combined vector for each group according to their sample size



Creating the null distribution

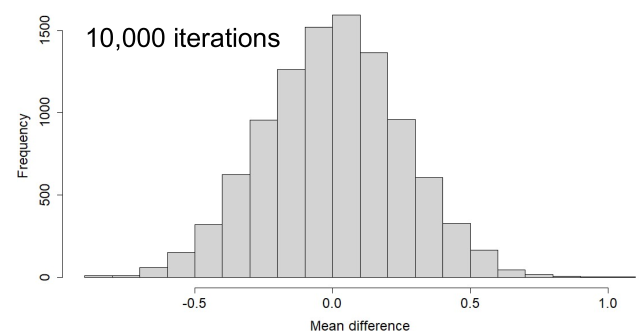
3. Calculate mean difference and save result

Mean minus Mean



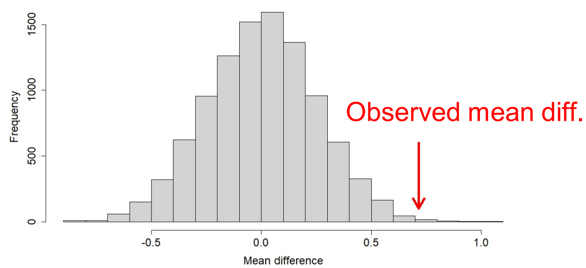
Creating the null distribution

4. Repeat steps 2-3 at least 1,000 times to get a null distribution of mean differences



4. Calculate P-value

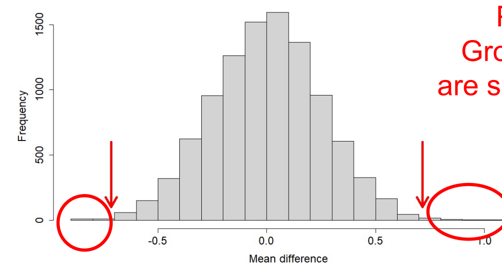
- Probability of getting observed data or more extreme, given null distribution
- Calculated by counting # iterations where simulated data more extreme than observed and dividing by # iterations



65

Two-tailed or one-tailed?

- Two-tailed: don't care about direction of difference
- That is, just interested in whether group means are different, not whether one group is bigger than another

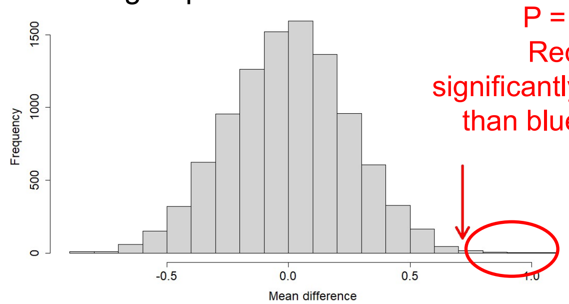


P = 0.0039
Group means are significantly different

66

Two-tailed or one-tailed?

- One-tailed: DO care about direction of difference
- Interested in whether red group mean is bigger than blue group mean



P = 0.0024
Red group significantly larger than blue group

67

Two-tailed or one-tailed?

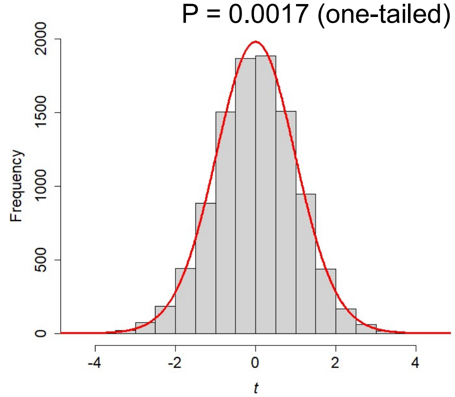
- Which one to pick depends on your research question!

68

Before computers: t-test

- Test statistic is t
- $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}}$
- Histogram: Monte Carlo methods
- Red curve: t-distribution

t-test
P = 0.0033 (two-tailed)
P = 0.0017 (one-tailed)



69

How to interpret P-values?

- "P-value is the probability the null hypothesis is correct"
- **WRONG**
- Remember, P-value is $P(D|H_0)$
- First bullet point is $P(H_0|D)$
- This is a logical fallacy known as **affirming the consequent**

70

Affirming the consequent: an example

- Let's say your friend is not returning your calls
- What's the probability that your friend doesn't call you back, given that they're mad at you?
- P(no call | mad): low, medium, or high?
- What you really want to know is P(mad | no call)
- Is this low, medium, or high?

71

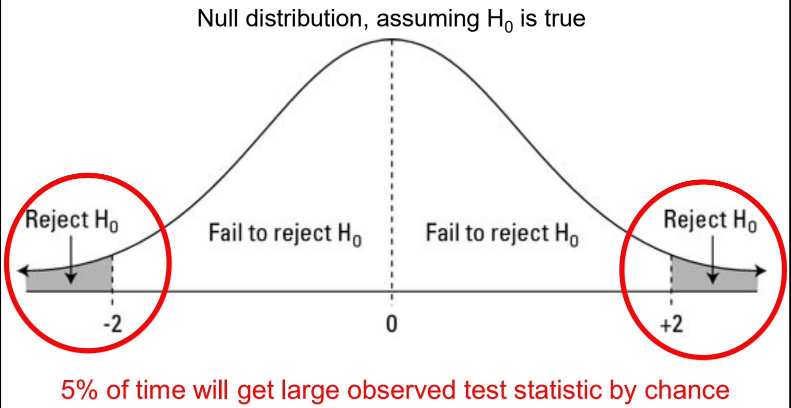
Type I error

- If null hypothesis is true, you will get $P < 0.05$, 5% of the time on average

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

72

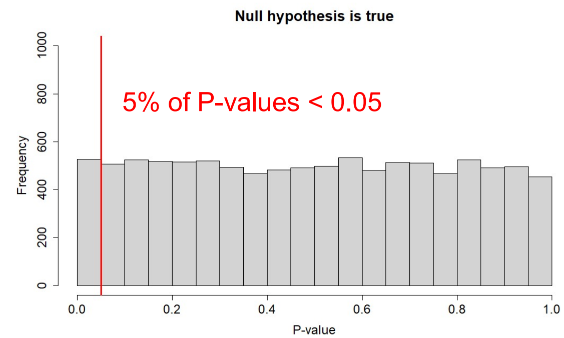
Type I error



73

Type I error

- Believe it or not, P-values have sampling distributions too



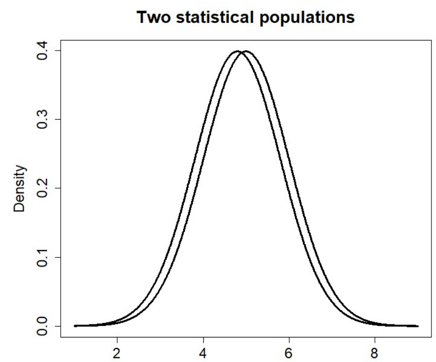
74

What affects P-values?

- P-values go down if:
- Mean difference is larger (difference between groups is clearer)
- Variation among observations decreases (difference between groups is clearer)
- Sample size increases (estimated means more accurate & precise due to law of large numbers)

75

P-values and sample size

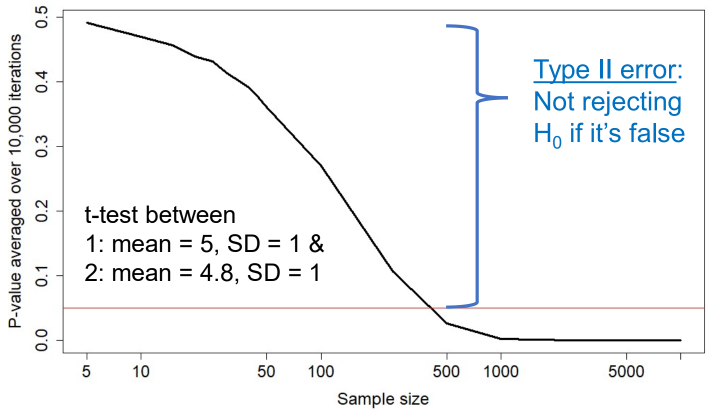


- With sample size n = 10, P = 0.91
- If n = 20, P = 0.56
- If n = 50, P = 0.35
- If n = 100, P = 0.03*

Just like how CI shrinks & significantly differs from values w/ increasing sample size!

76

P-values and sample size



77

P-values and sample size

- If you have a large enough sample size, any small difference will be significant (law of large numbers → means are estimated precisely so sampling distribution of mean diff. will be narrow & exclude obs. mean diff.)

Government Warns Some Birth Control Pills May Cause Blood Clots

EDITH M. LEDERER October 19, 1995

Click to copy

LONDON (AP) _ The government warned Thursday that a new type of birth control pill used by 1.5 million British women may cause blood clots, according to new, unpublished studies.

78

P-values and sample size

- If you have a large enough sample size, any small difference will be significant (law of large numbers → means are estimated precisely so sampling distribution of mean diff. will be narrow & exclude obs. mean diff)
- I will make the (untestable) assertion that <u>NOTHING</u> in biology/anthropology is *exactly* zero
- Therefore, $P < 0.05$ w/ a large enough sample size, and $P > 0.05$ if sample size is too small

79

P-values and sample size

- So, P-values are just measuring statistical power (i.e., the ability to reject H_0 if it's false), which is correlated with sample size
- *Always* interpret the mean difference itself in addition to P-values!
- E.g., Group 1 (mean height = 5.5 feet) is statistically different from Group 2 (mean height = 5.6 feet)
  - Interesting or not?

80

Questions?

Summary

- Statistical inference is key in scientific inference
- Frequentism is most common framework (taking many samples from a theoretical population)
- The sampling distribution of a statistic is the statistic calculated on each of these samples
- Important for standard errors, confidence intervals, and P-values (i.e., hypothesis testing)
- Can use R and simulations to understand how these statistics behave!

- To Fisher, the significance test makes sense only in the context of a sequence of experiments, all aimed at elucidating the effects of specific treatments. Reading through Fisher's applied papers, one is led to believe that he used significance tests to come to one of three possible conclusions. If the p-value is very small (usually less than .01), he declares that an effect has been shown. If the p-value is large (usually greater than .20), he declares that, if there is an effect, it is so small that no experiment of this size will be able to detect it. If the p-value lies in between, he discusses how the next experiment should be designed to get a better idea of the effect. Except for the above statement, Fisher was never explicit about how the scientist should interpret a p-value. What seemed to be intuitively clear to Fisher may not be clear to the reader.
- Salsbury (2001:100)

- For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning .... It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true. In fact that “errors of the second kind” are committed only by those who misunderstand the nature and application of tests of significance.
- Fisher (1935)