Announcements

- No class on Monday 20th January
- ECON 270 Pre-req issue.

Lecture 2

Statistics Review I

How to get credible casual effects?

- Randomization: The Gold Standard?

  -Example: Randomly give people (who are similar in characteristics) health insurance, and then compare health status of those insured with the uninsured.

  -How can it go wrong? What are some criticisms of randomization?

  -Think about it and we will discuss it in later classes

- Non-experimental methods:
  - Regression with control variables/matching
  - Differences-in-differences
  - Instrumental variables
  - Regression discontinuity designs

Non-experimental methods

- Selection on the observables

  -We assume treatment status is determined by observable variables.

  e.g. The decision to buy health insurance is determined by family income, education, race (observed)

  -Then we can “control” for these variables in some way and estimate a causal effect.

  -E.g. Regressions with control variables/matching

- Selection on the unobservables

  -Here we acknowledge that treatment status is determined by factors that we can’t measure.

  e.g. The decision to go to Harvard is determined by one’s innate ability (unobserved)

  -Then we can still find good counterfactuals to “control” for the unobservables

  -E.g. Instrumental variable, regression discontinuity designs, differences-in-differences

Outline

Statistics review

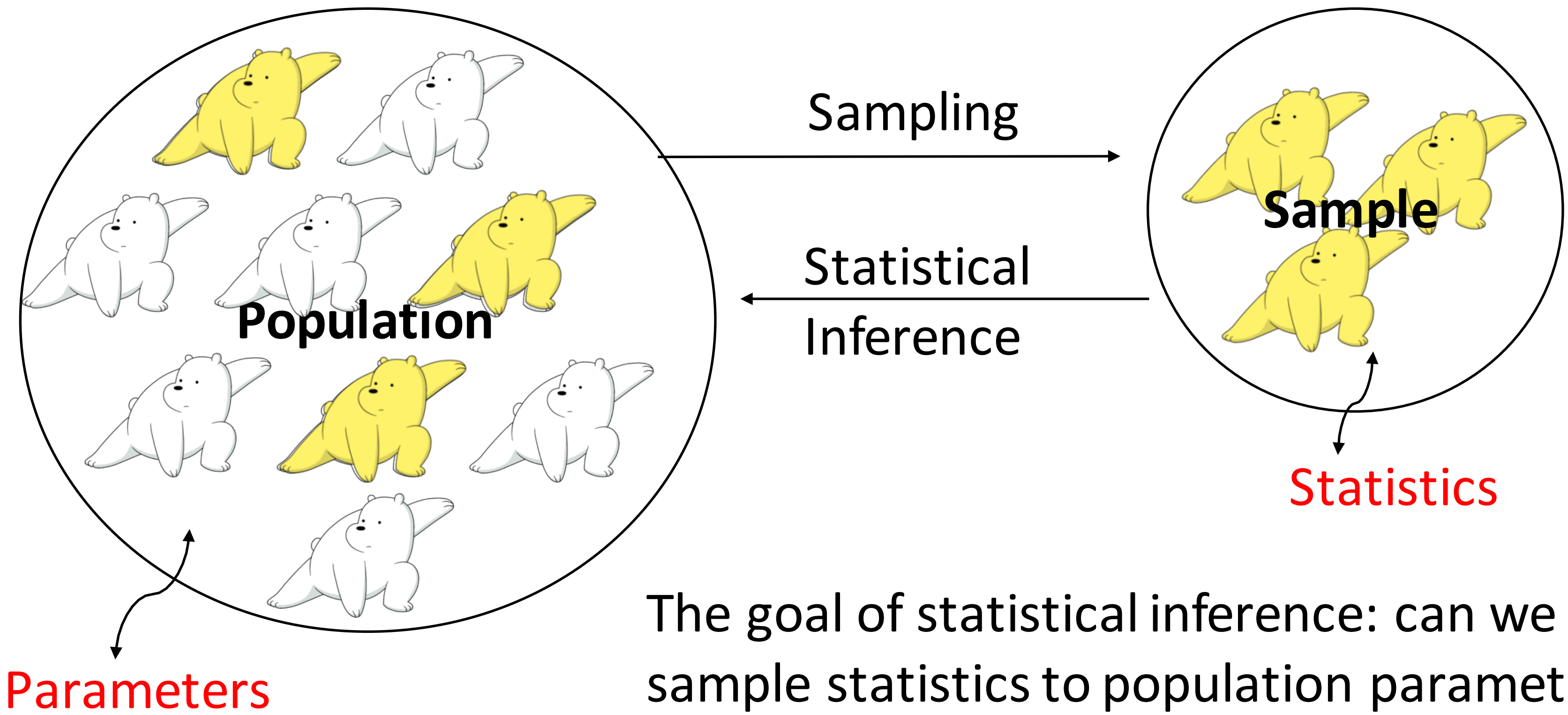
- Sampling distribution of the sample mean
- Central limit theorem
- Statistical inference (P-values)

Statistics Review I

Basic concepts: population vs. sample

- **Populations** contain all of the items or individuals we are interested in
- **Samples** are subsets of population
- **Parameters** are measures describing populations
- **Statistics** are measures describing samples
- **Sampling** is the selection of samples from a population
- **Statistical inference** is the process of drawing conclusions about population from samples. This is the core part of statistics

Population vs. Sample



The goal of statistical inference: can we generalize sample statistics to population parameters?

Mean, variance and standard deviation

- For all three measures, they can either describe a population or sample. For population, they are called **parameters**; for sample, they are called **statistics**.
- Population mean: E(X) $= \frac{\sum_{i=1}^{N} X_i}{N}$ (Read: The expectation of X)
- Sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (Σ means sum)
- Population variance: $\mathrm{v}ar(X) = \frac{\sum_{i=1}^{N}[X_i - E(X)]^2}{N}$; $\mathrm{S.D.} = \sqrt{Var(X)}$
- Sample variance: S^2=$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$; S=$\sqrt{S^2}$

Covariance

- Measures how much two numerical variables change together
- Measures the **direction** and strength of **linear** relationship of two numerical variables
- Population covariance :$cov(X,Y) = \frac{\sum_{i=1}^{N}[X_i - E(X)][Y_i - E(Y)]}{N}$
- Sample covariance:$S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{n-1}$

Example: training performance

- Mary and Emily like to keep fit. Mary lift weights and Emily runs. The following table shows their hours of training and results for 5 days. What is the covariance between training hours and exercise performance for Mary?

	Training hours	**Mary-max weight lifted (lb)**	**Emily-fastest 400m race time (sec)**
Day1	2	70	50
Day2	1	60	54
Day3	1.5	65	52
Day4	2.4	80	48
Day5	1.8	60	50

Answer : $Sxy = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Exercise: Find the covariance for Emily

Mary:

- Find covariance between training hour and max weight lifted
- Step1: find the mean of training hour ($\bar{X}$)and max weight ($\bar{Y}$)
- Step2:

	X- $\bar{X}$	**Y- $\bar{Y}$**	(X- $\bar{X}$)*(Y- $\bar{Y}$)
Day1	2-1.74	70-67	0.26*3
Day2	1-1.74	60-67	-0.74*(-7)
Day3	1.5-1.74	65-67	-0.24*(-2)
Day4	2.4-1.74	80-67	0.66*13
Day5	1.8-1.74	60-67	0.06*(-7)
	$\bar{X}$=1.74	$\bar{Y}$=67	Σ(X- $\bar{X}$)*(Y- $\bar{Y}$)=14.6

- Step3:

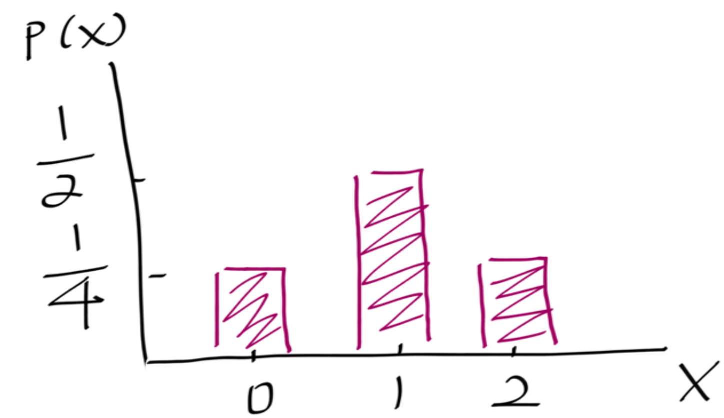
$$Sxy = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{14.6}{5-1} = 3.65$$

Random variable and probability distribution

- Random variable is a variable whose value is a **numerical** outcome of a random phenomenon
- What is the difference between a random variable and a regular variable?
  - Random variable always has a probability distribution associated with it
- e.g. Toss the coin twice
  - Define the random variable X= # of heads
  - Now X can be 0, 1 or 2
  - P(X=0)=1/4
  - P(X=1)=1/2
  - P(X=2)=1/4

prob. distribution :

Toss a coin twice – X = # of heads

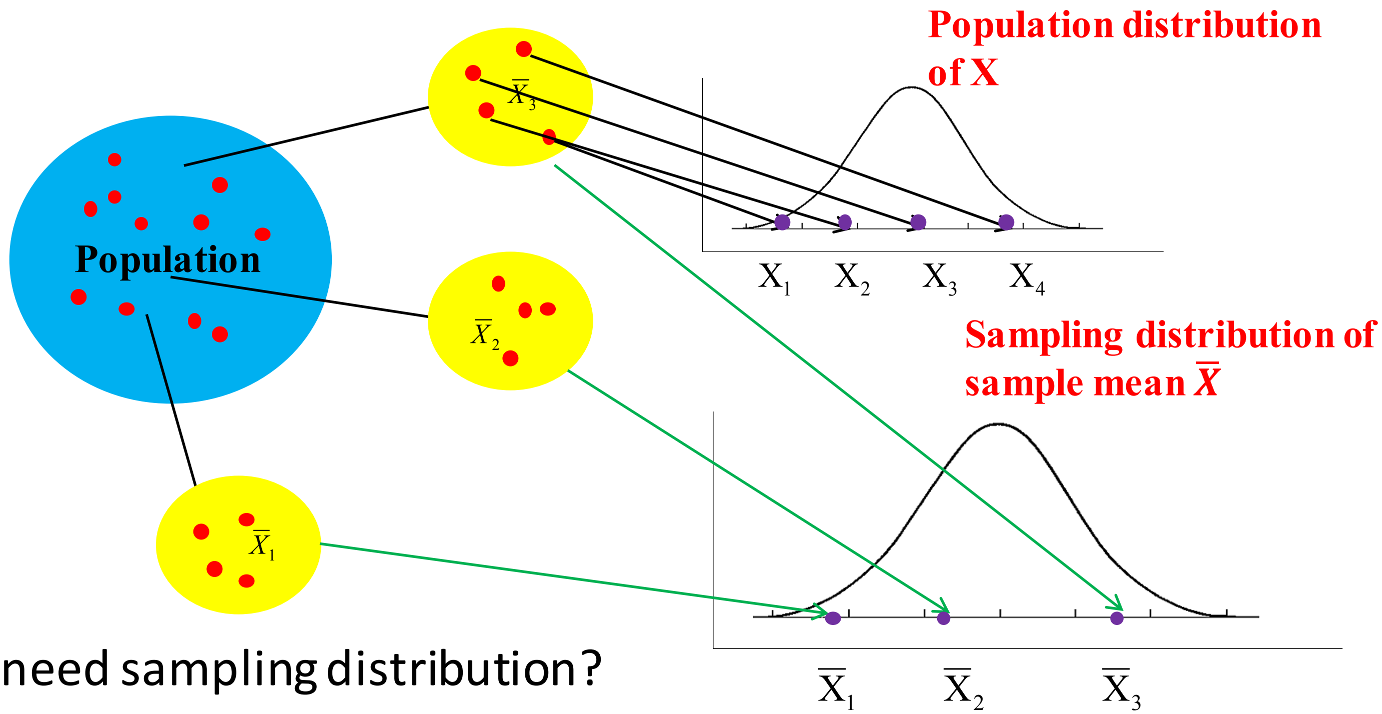


Population distribution vs. sampling distribution

- **Population distribution** of a random variable (X) is the distribution of its values for all members of the population.
  - Example: Height of individuals in the entire country.
- **Sampling distribution** is the probability distribution of a **statistic (e.g. mean ($\overline{X}$))**.
  - The average height of a class follows normal distribution-sampling distribution

Sampling distribution of the sample mean

- A graphical comparison between population distribution and sampling distribution:



- Why do we need sampling distribution?
  - To determine whether sample mean is a good measure of population mean, we need to know its distribution-sampling distribution

Central Limit Theorem

- A video of CLT https://vimeo.com/75089338

- If the population distribution is normal, i.e. $X \sim N(\mu, \sigma^2)$. Sampling distribution is normal too, i.e. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.
  - What does it mean?

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

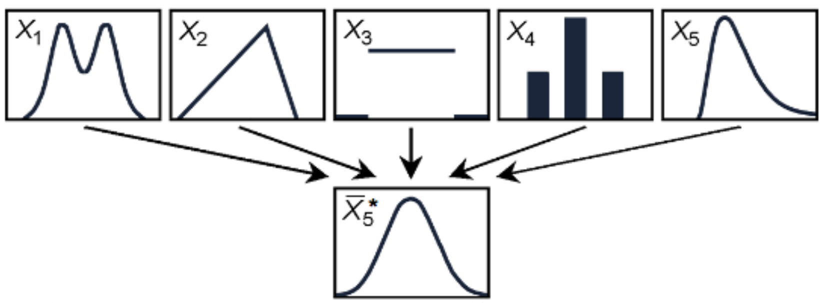
- As sample size n increases, the standard deviation of sample mean (known as **standard error)** decreases.
  - It is used to determine how far away the mean of each sample is from the true population mean.

Exercise

- You draw random samples of size n=36 from a population with mean 240 and standard deviation 18. Find the mean and standard error of the sampling distribution
- Repeat the calculation for a sample size of 144. Explain the effect of sample size on standard error

Central limit theorem

- Why is it important?
- Allows us to use the normal distribution for statistical inference in situations where the underlying distribution is **<u>not</u>** normal.

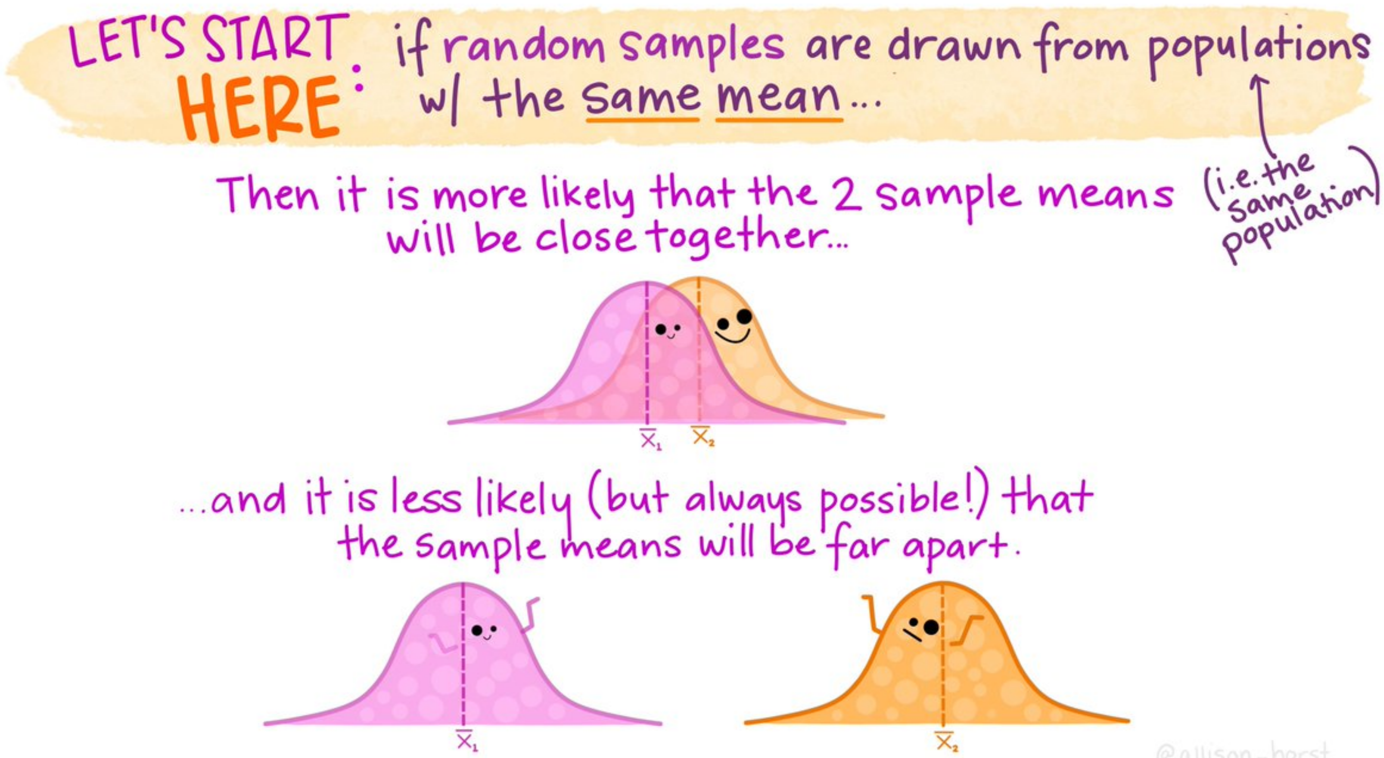


- How big is “sufficiently large?” Typically we think n about 30 is sufficient.

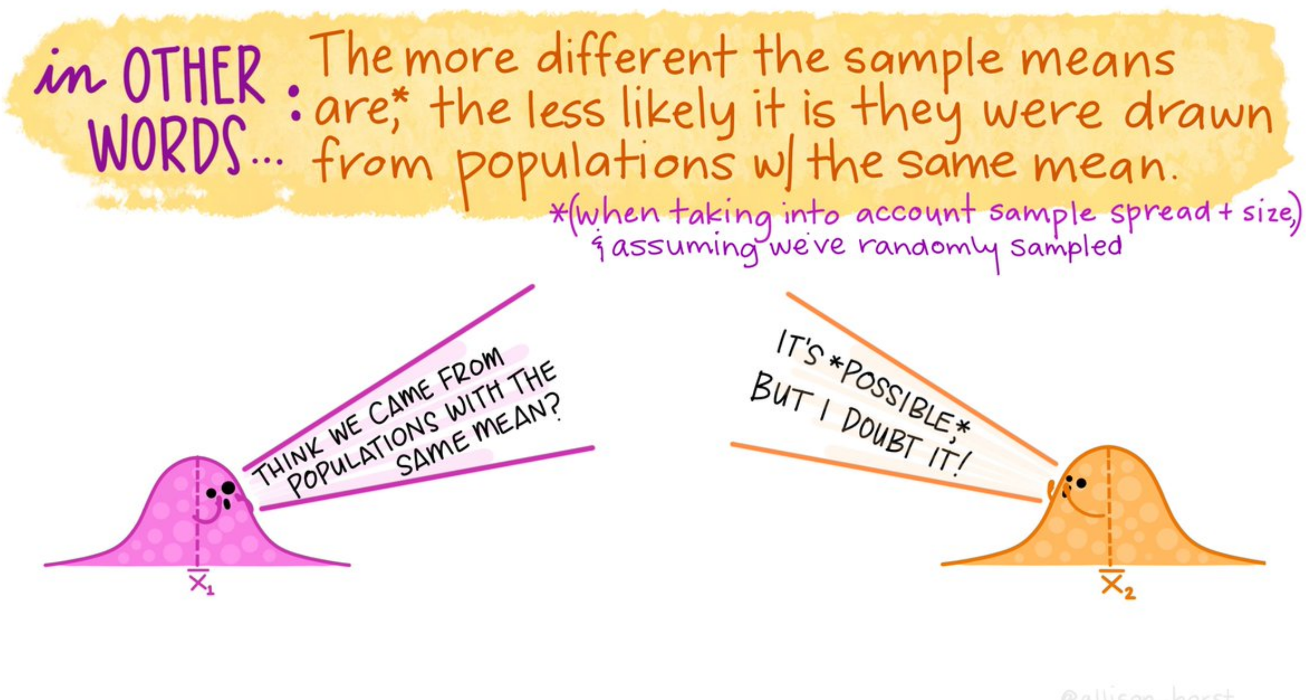
Population and sampling distribution

	POPULATION	SAMPLING DISTRIBUTION
Mean	μ	$\mu_{\bar{x}} = \mu$
Standard Deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Shape	Normal	Normal
	Undetermined (skewed, etc.)	If n is "small" shape is similar to shape of original graph *OR* If n is "large" (rule of thumb: $n \geq 30$) shape is approximately normal (central limit theorem)

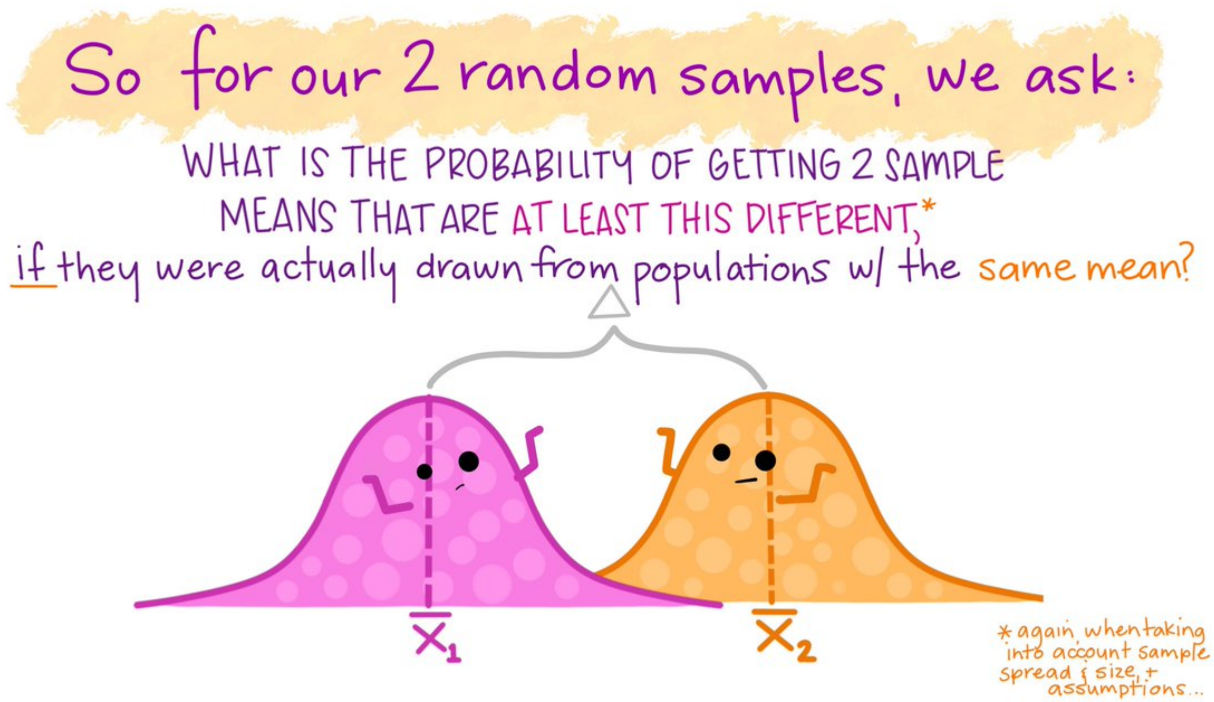
Statistical inference (borrowed from @allison_horst)



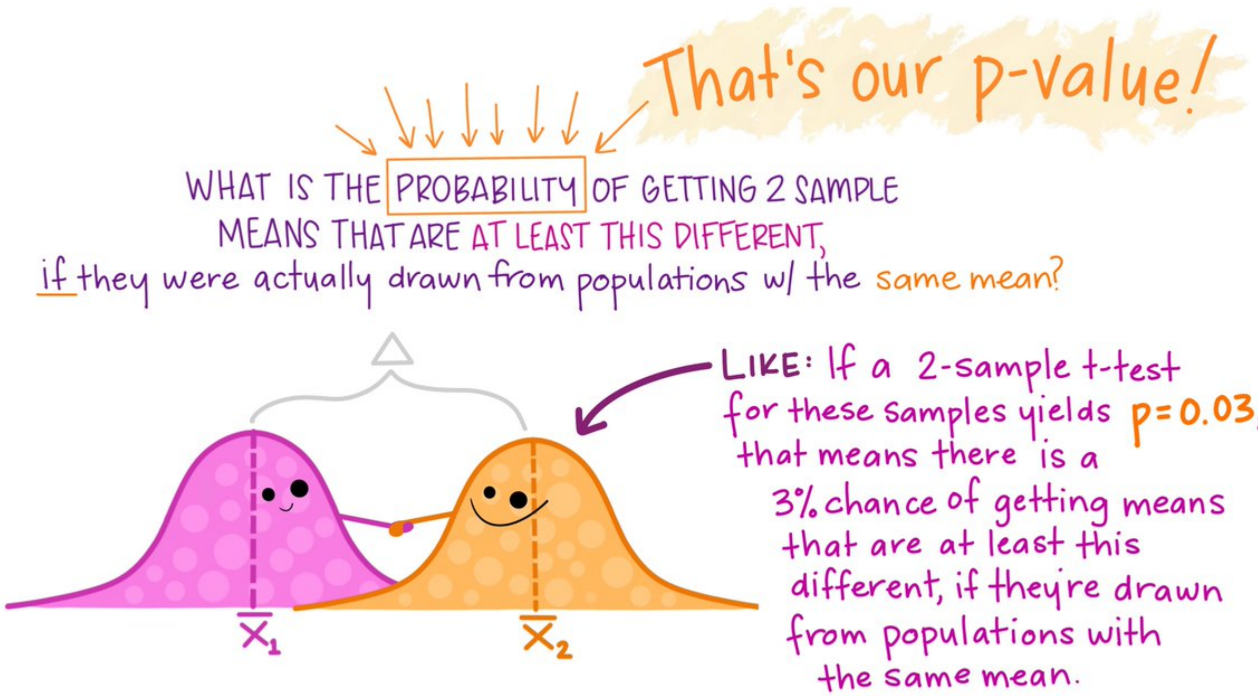
Statistical inference



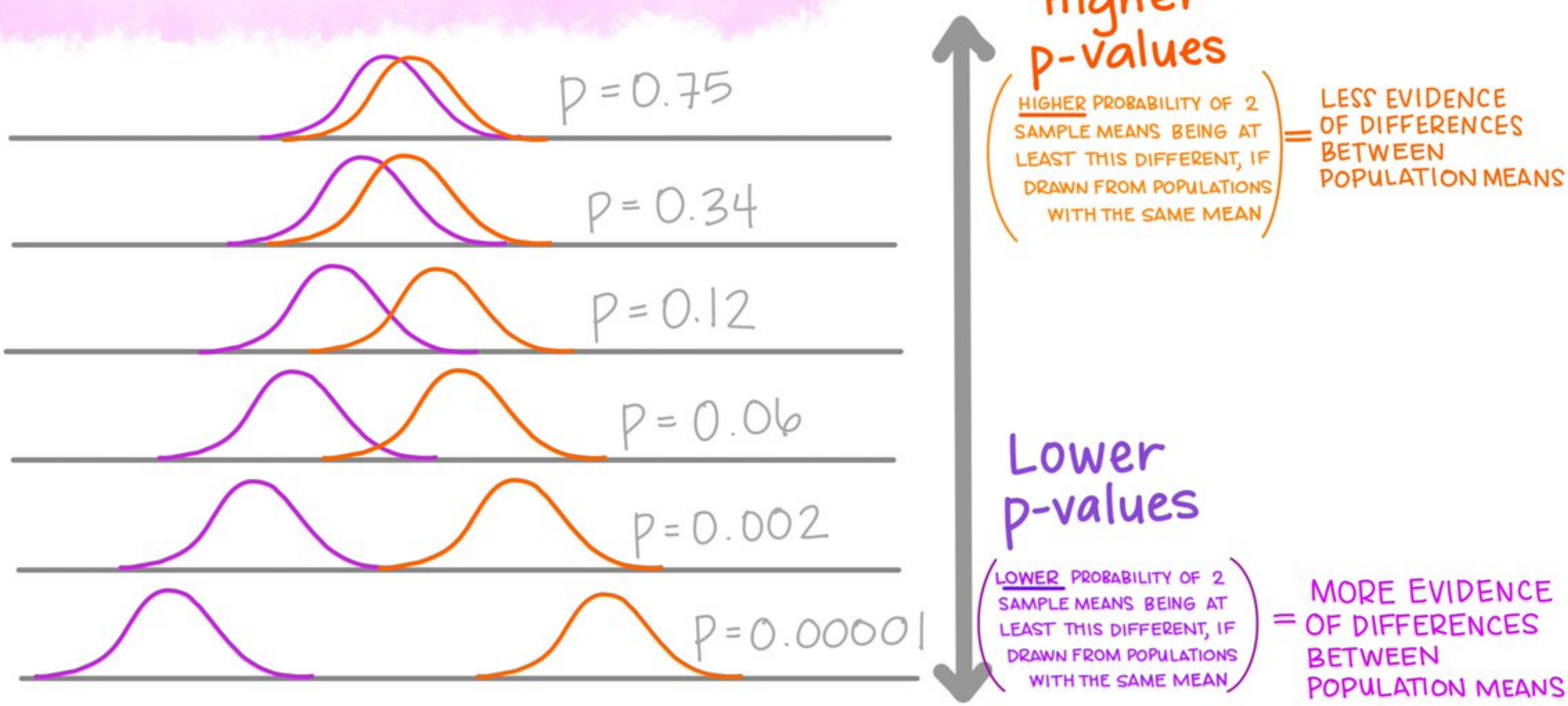
Meet p-values



P-values continued



P-VALUES, SCHEMATICALLY:



Question:

WHEN DO WE HAVE ENOUGH EVIDENCE TO SAY THERE IS A SIGNIFICANT DIFFERENCE?

Answer:

WHEN OUR P-VALUE IS BELOW OUR SELECTED SIGNIFICANCE LEVEL (α), USUALLY (BUT NOT ALWAYS) = 0.05.

Which means:

IF THE PROBABILITY (p-value) OF FINDING AT LEAST OUR DIFFERENCE IN SAMPLE MEANS (IF THEY WERE DRAWN FROM POPULATIONS WITH THE SAME MEANS) IS LESS THAN 5%, THAT'S ENOUGH EVIDENCE FOR US TO DECIDE THEY ARE LIKELY FROM POPULATIONS WITH UNEQUAL MEANS.

Statistical inference

- Statistical inference: draw conclusions about population from sample.
- Example
  - Parameter: average height of adults in the US
  - Statistic : average height of 50 randomly selected people from the population

- Suppose scores on an IQ test are normally distributed. 10 people are randomly selected and tested. The mean and standard deviation in the sample group is 95 and 15. Construct a 95% confidence interval for the true population mean.
- Extract information:
- n=10,degree of freedom=n-1=9; $\bar{X} = 95$;S=15; confidence level (C)=0.95

Steps to construct Confidence Interval(CI)

- Each confidence level C corresponds to a tail probability $\alpha = 1 - C$
- Given α and the degree of freedom (n-1), find $t_{\alpha/2}$ using the t-table
- Construct margin of error m=$t_{\alpha/2}$*s/$\sqrt{n}$
- CI=$\bar{X}$ $\pm$m=$\bar{X} \pm t_{\alpha/2}$*s/$\sqrt{n}$

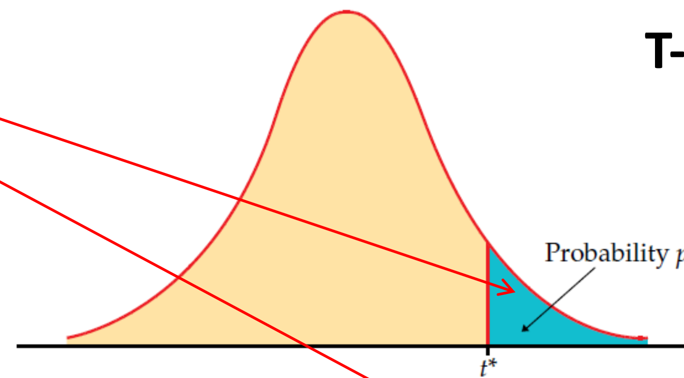
- Where s is the standard deviation of the sample, and $\bar{X}$ is the mean. n is sample size.

Answer

- Step 1: Convert C=0.95 to t-score using t table.
- C=0.95, $\alpha = 1 - C$=0.05, $\frac{\alpha}{2} = 0.025$
- look it up in t-table, $t_{\alpha/2}$=2.262
- Step 2: Construct margin of error.
- m=$t_{\alpha/2}$*s/$\sqrt{n}$=2.262*15/$\sqrt{10}$=10.73
- Step 3: Construct CI
- $\bar{X} \pm$m=95 $\pm$10.73
- **<u>Translate the CI into real world meaning</u>**

\- You are 95 percent confident that the true mean is within 84.27-105.73

T-Table



df	Upper-tail probability p						
	.25	.20	.15	.10	.05	.025	.02
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359

Exercise

- The Nielsen company conducted a survey of 64 mobile phone subscribers and find on average they spend 4 hours watching videos on their phone. The sample standard deviation is 3. Let's determine a 99% confidence interval for the average time phone subscribers spend watching videos over phone. Explain what the CI means

Review for quizzes

- Be able to calculate mean, variance, S.D, and covariance
- Understand basic concepts about sample and population
- Sampling distribution and standard error: Why sample mean is not representative of population mean
- What CLT is and why it is important
- What are p-values?