INTRODUCTION TO DATA SCIENCE

This lecture is based on course by
M. Cetinkaya-Rundel, Duke University
Data Analysis and Statistical Inference

21/01/2020

WFAiS UJ, Informatyka Stosowana
I stopień studiów

Statistical inference

- **Lets start with small case study:**
  - **gender discrimination**

- 48 male bank supervisors given the same personnel file, asked to judge whether the person should be promoted
- files were identical, except for gender of applicant
- random assignment
- 35 / 48 promoted
- are females are unfairly discriminated against?

Statistical inference: case study

data

		promotion		
		promoted	not promoted	total
gender	male	21	3	24
	female	14	10	24
	total	35	13	48

% of males promoted = 21/24 ≈ 88%

% of females promoted = 14/24 ≈ 58%

Statistical inference: case study

two competing claims

null hypothesis

"There is nothing going on"

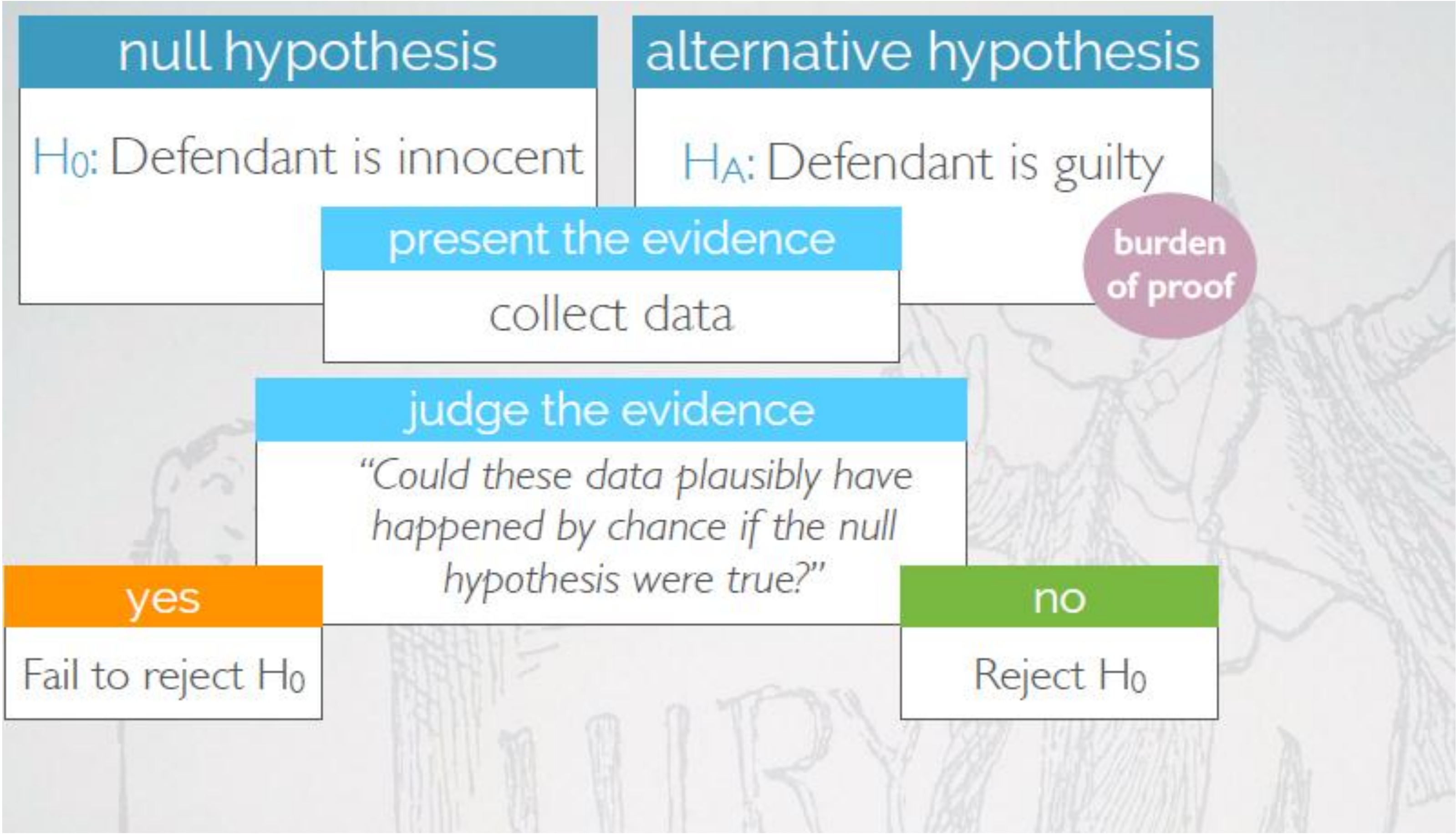
promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance

alternative hypothesis

"There is something going on"

promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance.

Statistical inference: case study



Statistical inference: case study

recap: hypothesis testing framework

- start with a null hypothesis (H_0) that represents the status quo
- set an alternative hypothesis (H_A) that represents the research question, i.e. what we're testing for
- conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods
  - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - if they do, then reject the null hypothesis in favor of the alternative

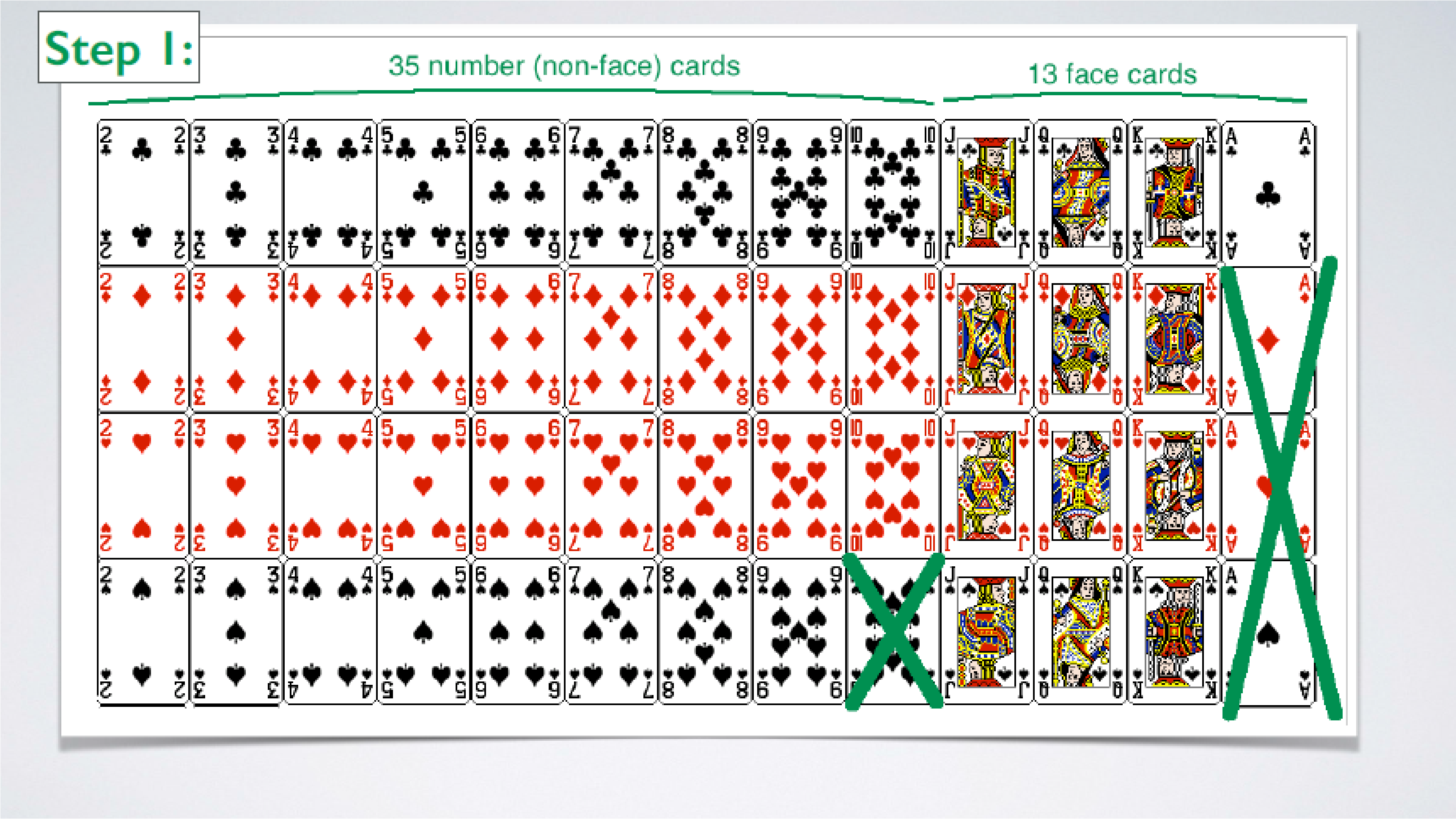
Statistical inference: case study

simulation scheme

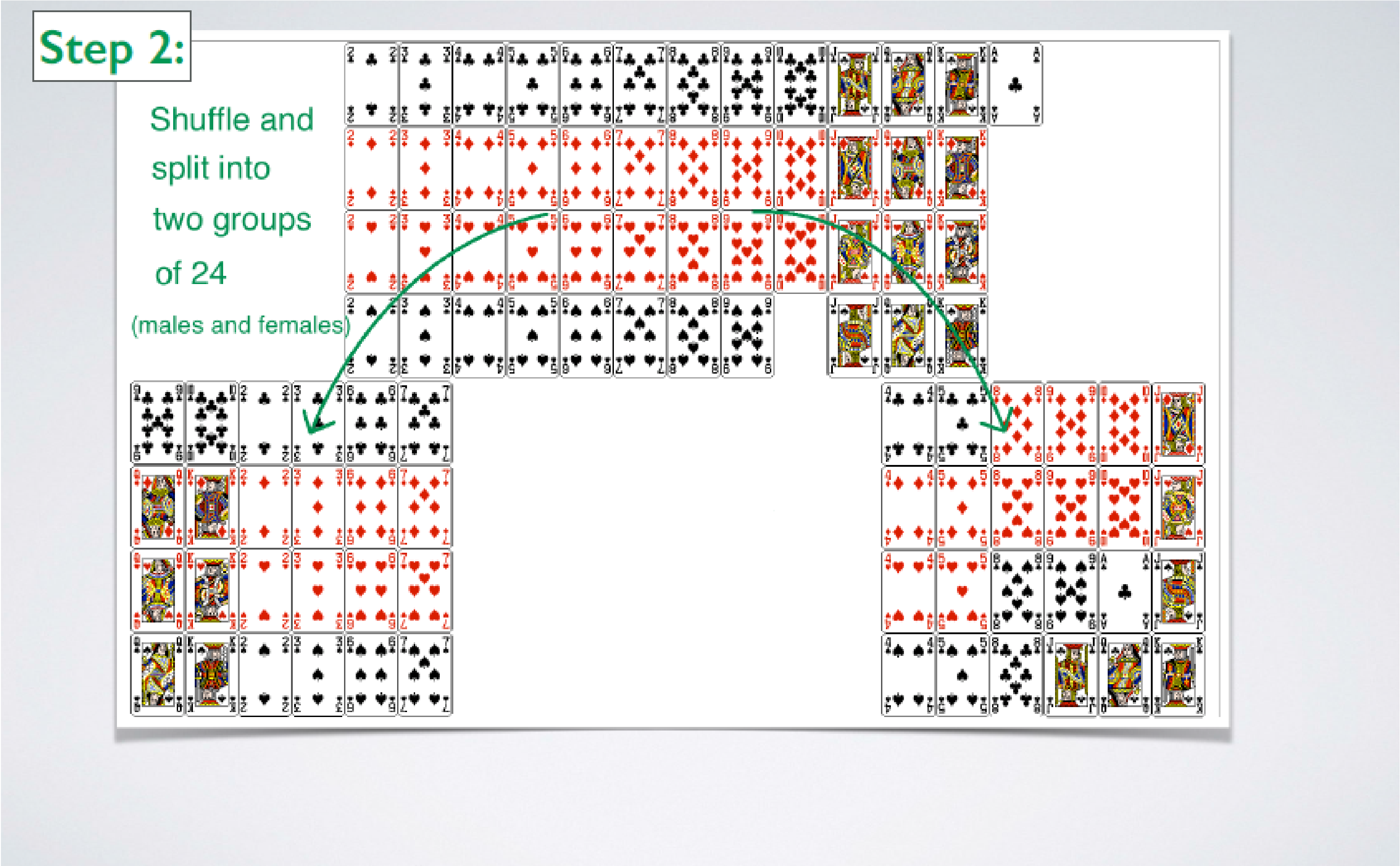
[use a deck of playing cards to simulate this experiment]

1. face card: not promoted, non-face card: promoted
   - set aside the jokers, consider aces as face cards
   - take out 3 aces → exactly 13 face cards left in the deck (face cards: A, K, Q, J)
   - take out a number card → 35 number (non-face) cards left in the deck (number cards: 2-10)
2. shuffle the cards, deal into two groups of size 24, representing males and females
3. count how many number cards are in each group (representing promoted files)
4. calculate the proportion of promoted files in each group, take the difference (male - female), and record this value
5. repeat steps 2 - 4 many times

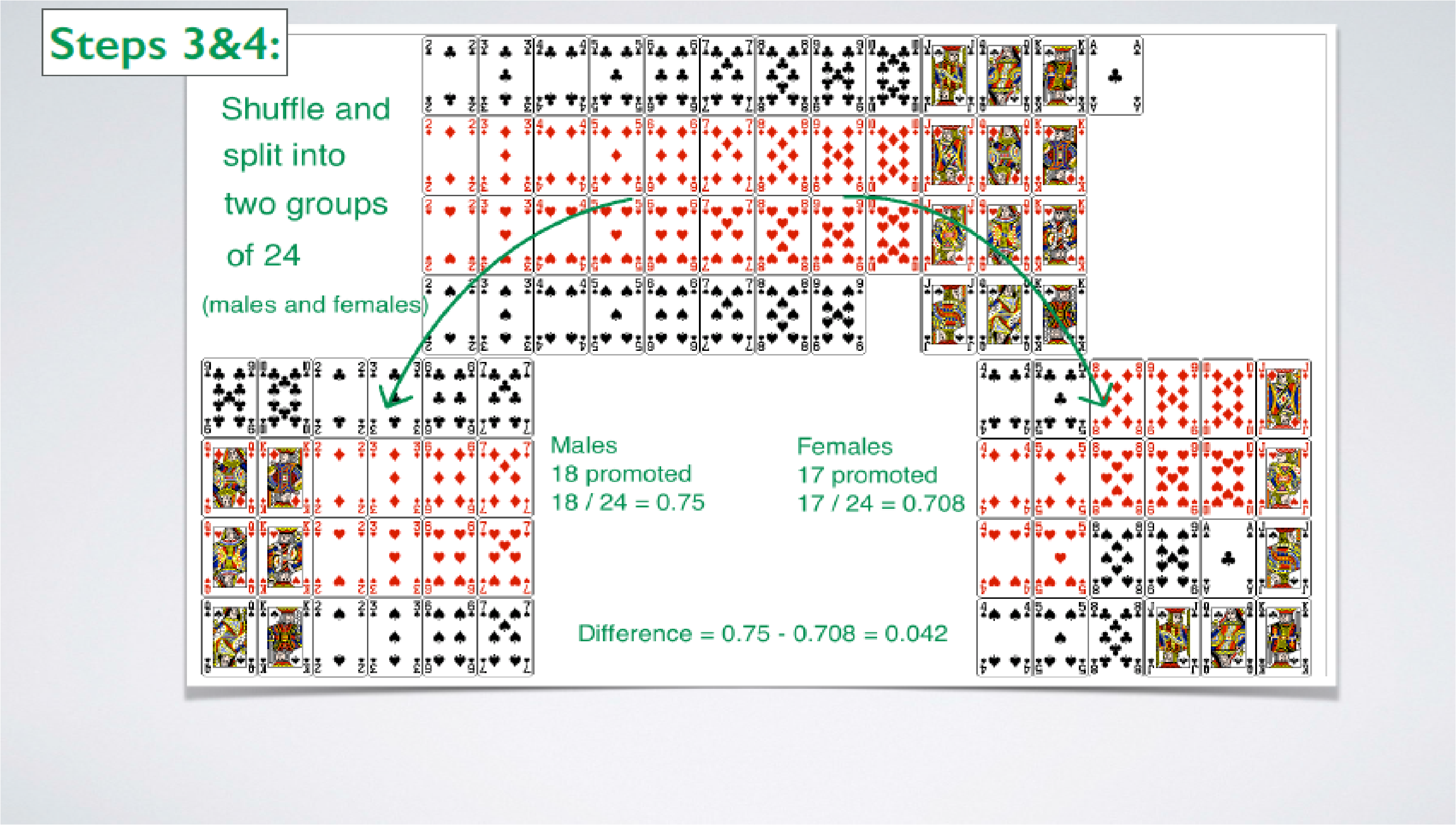
Statistical inference: case study



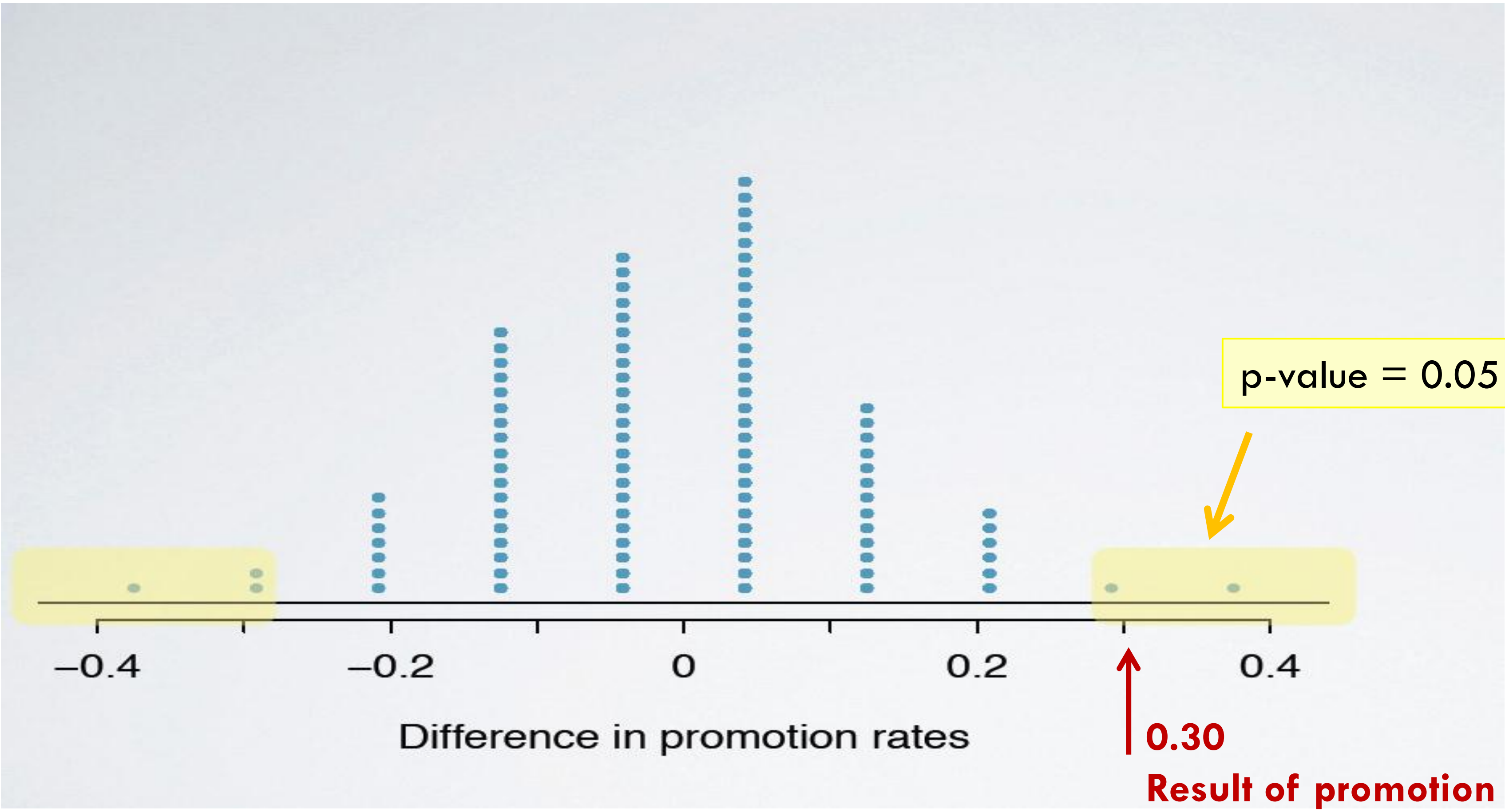
Statistical inference: case study



Statistical inference: case study



Statistical inference: case study



Statistical inference: case study

making a decision

- results from the simulations look like the data → the difference between the proportions of promoted files between males and females was due to chance (promotion and gender are independent)
- results from the simulations do not look like the data → the difference between the proportions of promoted files between males and females was <u>not</u> due to chance, but due to an actual effect of gender (promotion and gender are dependent)

Statistical inference: case study

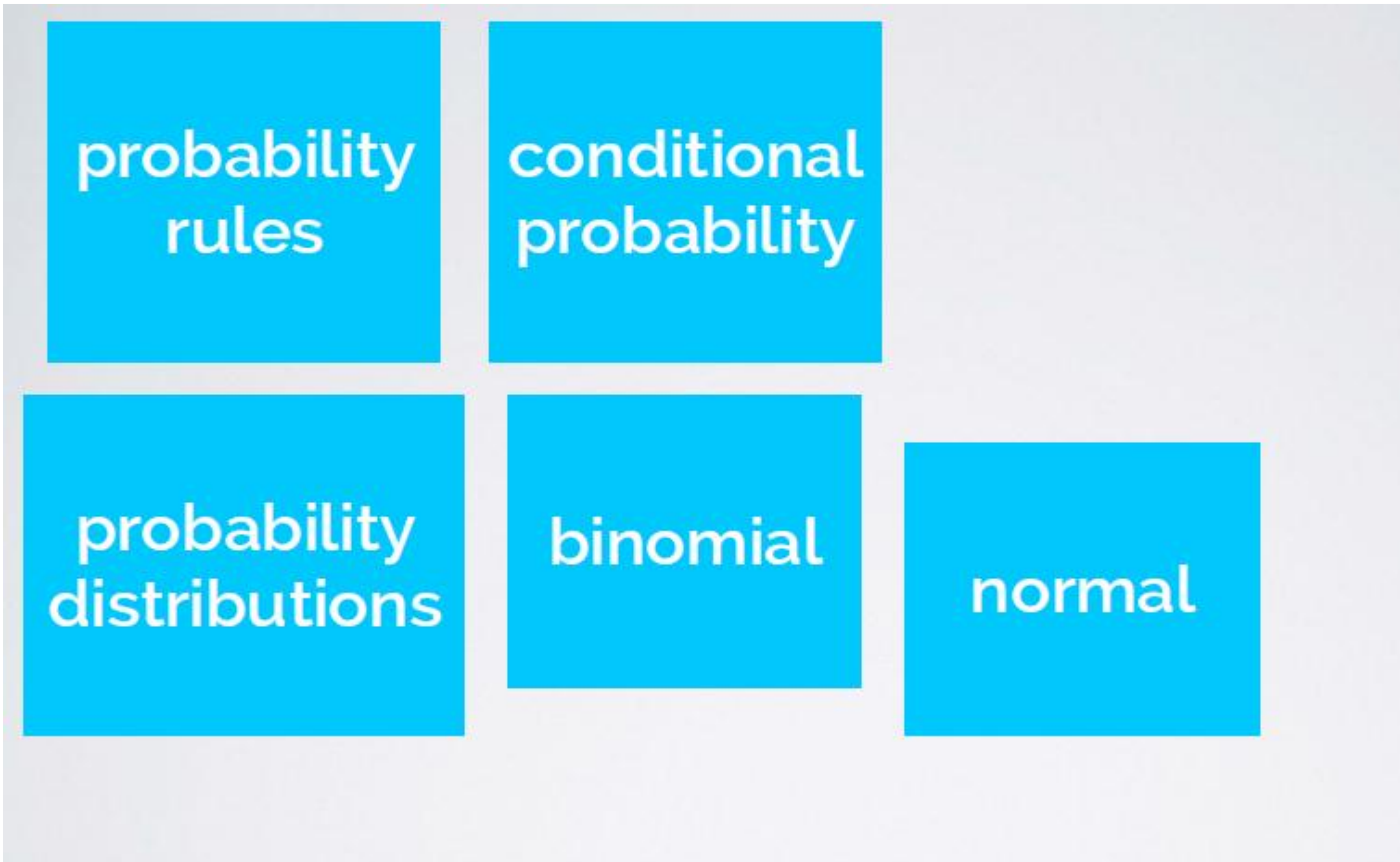
summary

- set a null and an alternative hypothesis
- simulate the experiment assuming that the null hypothesis is true
- evaluated the probability of observing an outcome at least as extreme as the one observed in the original data

p-value

- and if this probability is low, reject the null hypothesis in favor of the alternative

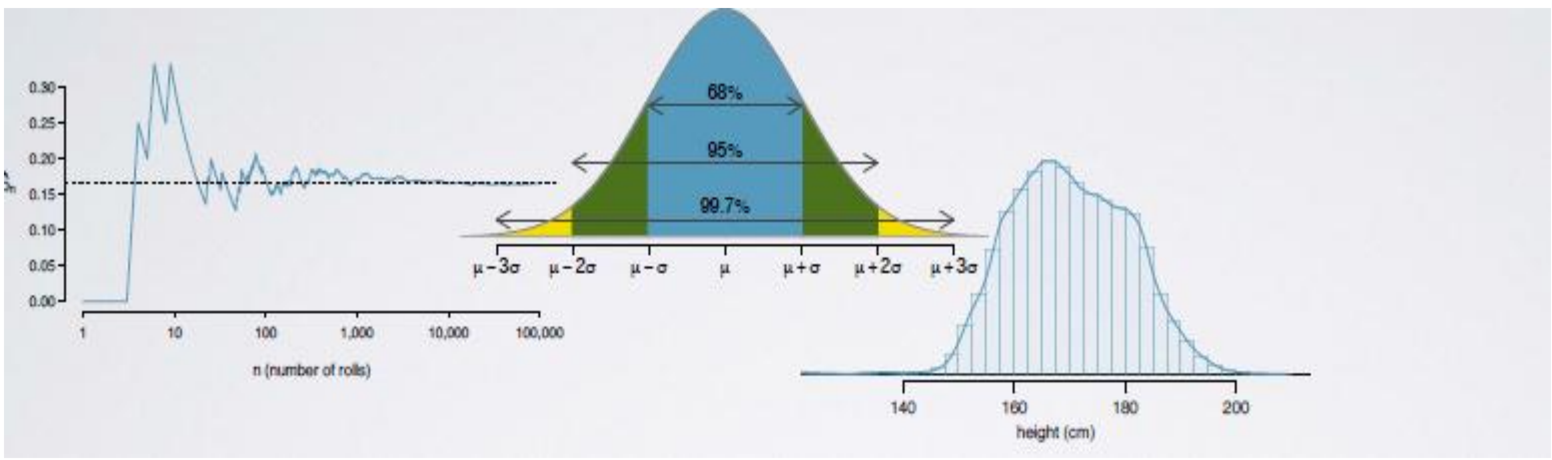
Probability and distributions



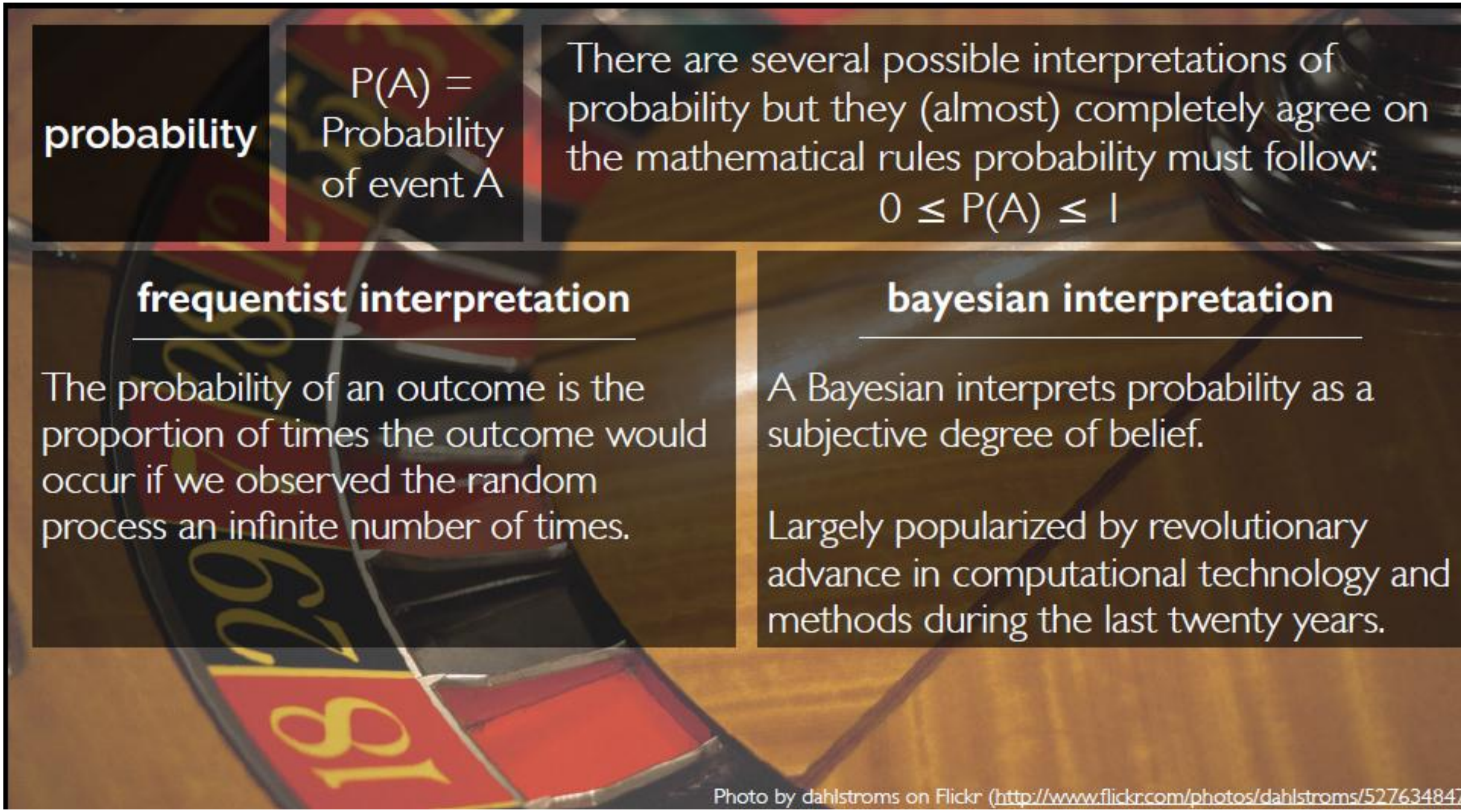
Probability and distributions

random process

In a random process we know what outcomes could happen, but we don't know which particular outcome will happen.



Probability and distributions

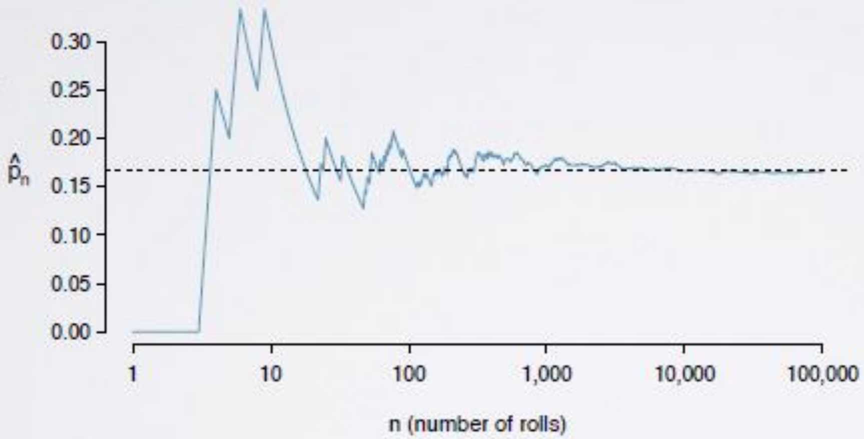


Probability and distributions

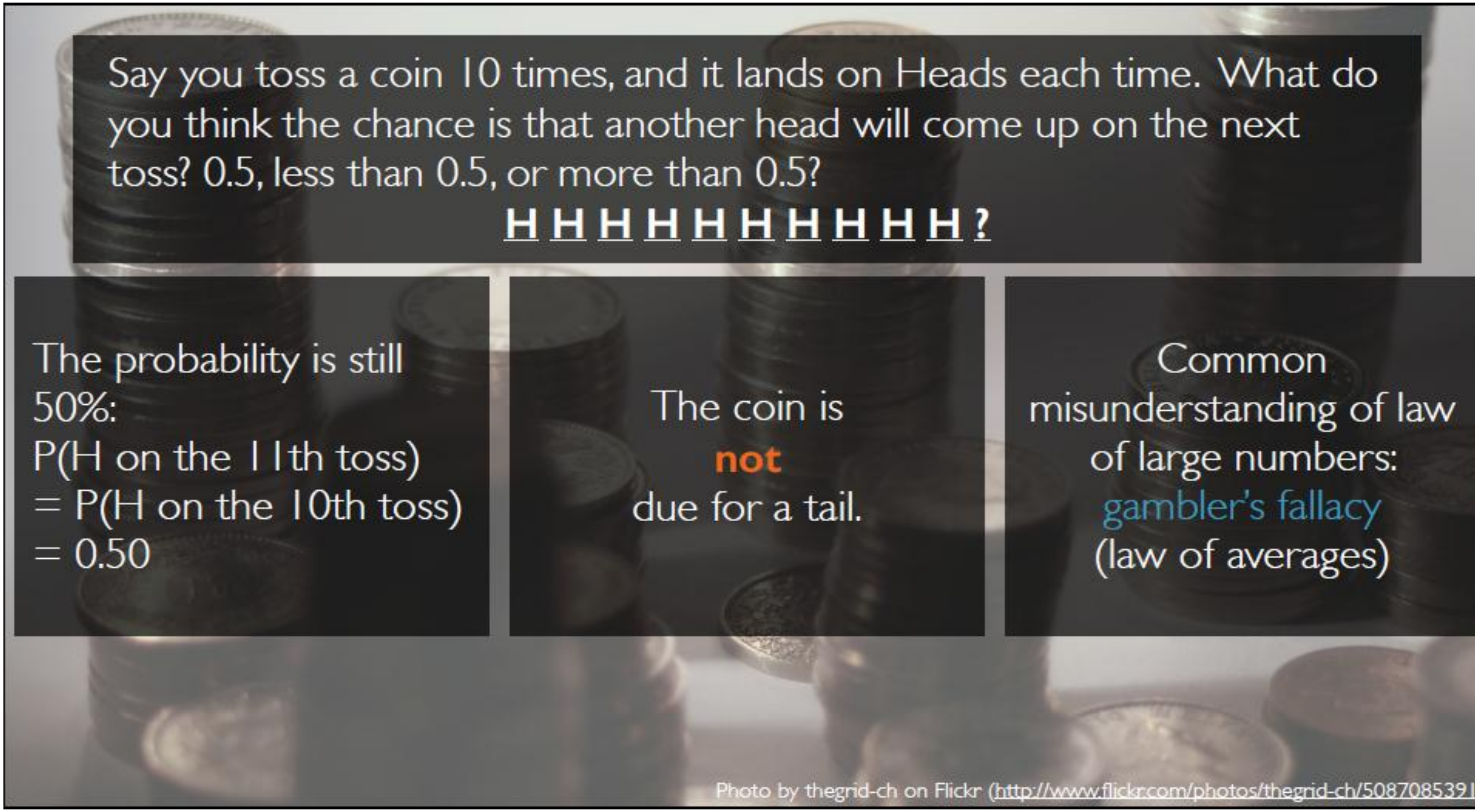
law of large numbers

law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome.

examples



Probability and distributions



Disjoint (mutually exclusive)

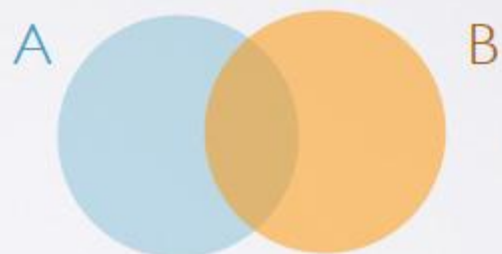
disjoint (mutually exclusive) events cannot happen at the same time.

- the outcome of a single coin toss cannot be a head and a tail.
- a student can't both fail and pass a class.
- a single card drawn from a deck cannot be an ace and a queen.

$P(A \text{ and } B) = 0$

non-disjoint events can happen at the same time.

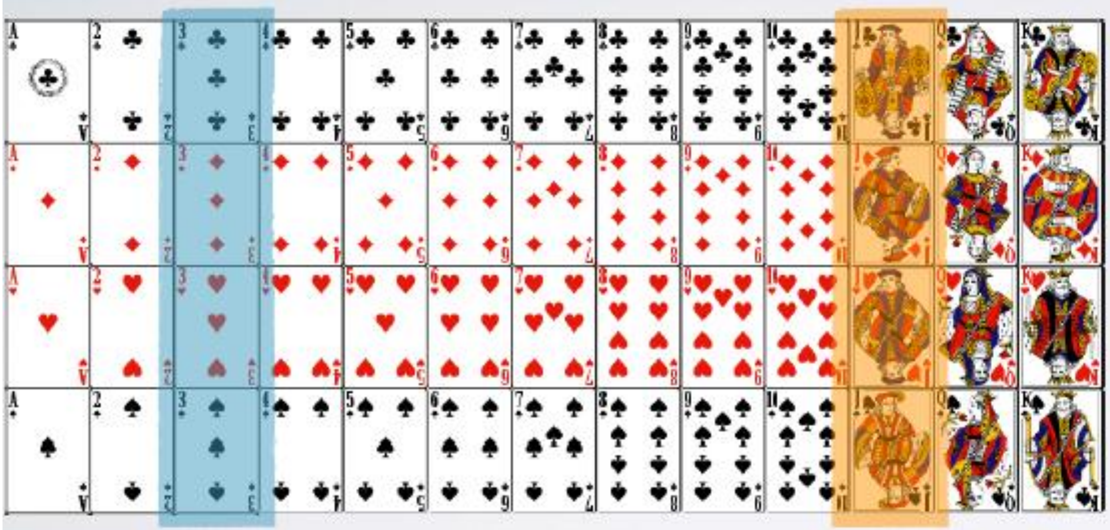
- a student can get an A in Stats and A in Econ in the same semester.



$P(A \text{ and } B) \neq 0$

Union of disjoint events

What is the probability of drawing a Jack or a three from a well shuffled full deck of cards?



$P(J \text{ or } 3)$

$= P(J) + P(3)$

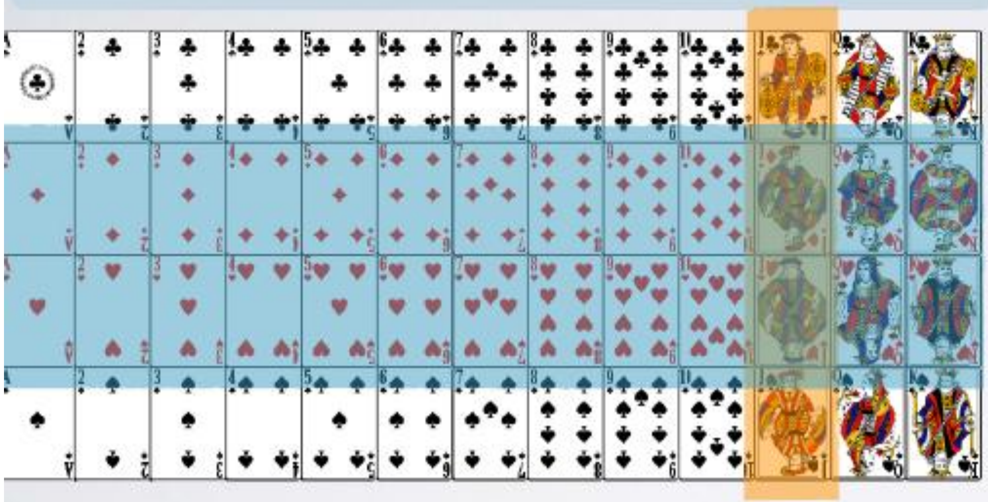
$= (4/52) + (4/52)$

≈ 0.154

For disjoint events A and B,
P(A or B) = P(A) + P(B)

Union of ono-disjoint events

What is the probability of drawing a Jack or a red card from a well shuffled full deck of cards?



$P(J \text{ or red})$

$= P(J) + P(\text{red}) - P(J \text{ and red})$

$= (4/52) + (26/52) - (2/52)$

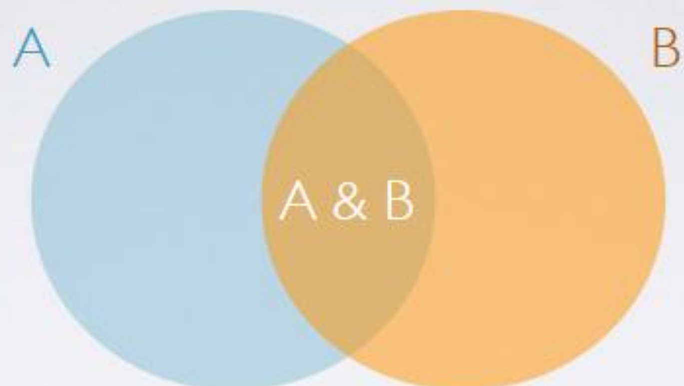
≈ 0.538

For non-disjoint events A and B,
P(A or B) = P(A) + P(B) - P(A and B)

General addition rule

General addition rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Note: When A and B are disjoint, $P(A \text{ and } B) = 0$, so the formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$.

Sample space

a sample space is a collection of all possible outcomes of a trial.

A couple has two kids, what is the sample space for the sex of these kids? For simplicity assume that sex can only be male or female.

$$S = \{MM, FF, FM, MF\}$$

Probability distributions

a probability distribution lists all possible outcomes in the sample space, and the probabilities with which they occur.

one toss	head	tail
probability	0.5	0.5

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

rules

1. the events listed must be disjoint
2. each probability must be between 0 and 1
3. the probabilities must total 1

Complementary events

complementary events are two mutually exclusive events whose probabilities that add up to 1.

complementary

one toss	head	tail
probability	0.5	0.5

complementary

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

Disjoint vs complementary

Do the sum of probabilities of two disjoint outcomes always add up to 1?

Not necessarily, there may be more than 2 outcomes in the sample space.

Do the sum of probabilities of two complementary outcomes always add up to 1?

Yes, that's the definition of complementary.

complementary

disjoint

Independence

two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

1st toss → 2nd toss

P(H) = 0.5 P(T) = 0.5

outcomes of two tosses of a coin are independent

1st draw → 2nd draw

P(A) = 3/51 P(J) = 4/51

outcomes of two draws from a deck of cards (without replacement) are dependent

Image sources:
Coin: http://commons.wikimedia.org/wiki/File:1913_Eliasberg_Liberty_Head_Nickel.png
Card: Open Clip Art Library (http://openclipart.org/cgi-bin/navigate/recreation/games/cards/white)

Independence

Checking for independence:

$P(A \mid B) = P(A)$, then A and B are independent.

given

Independence

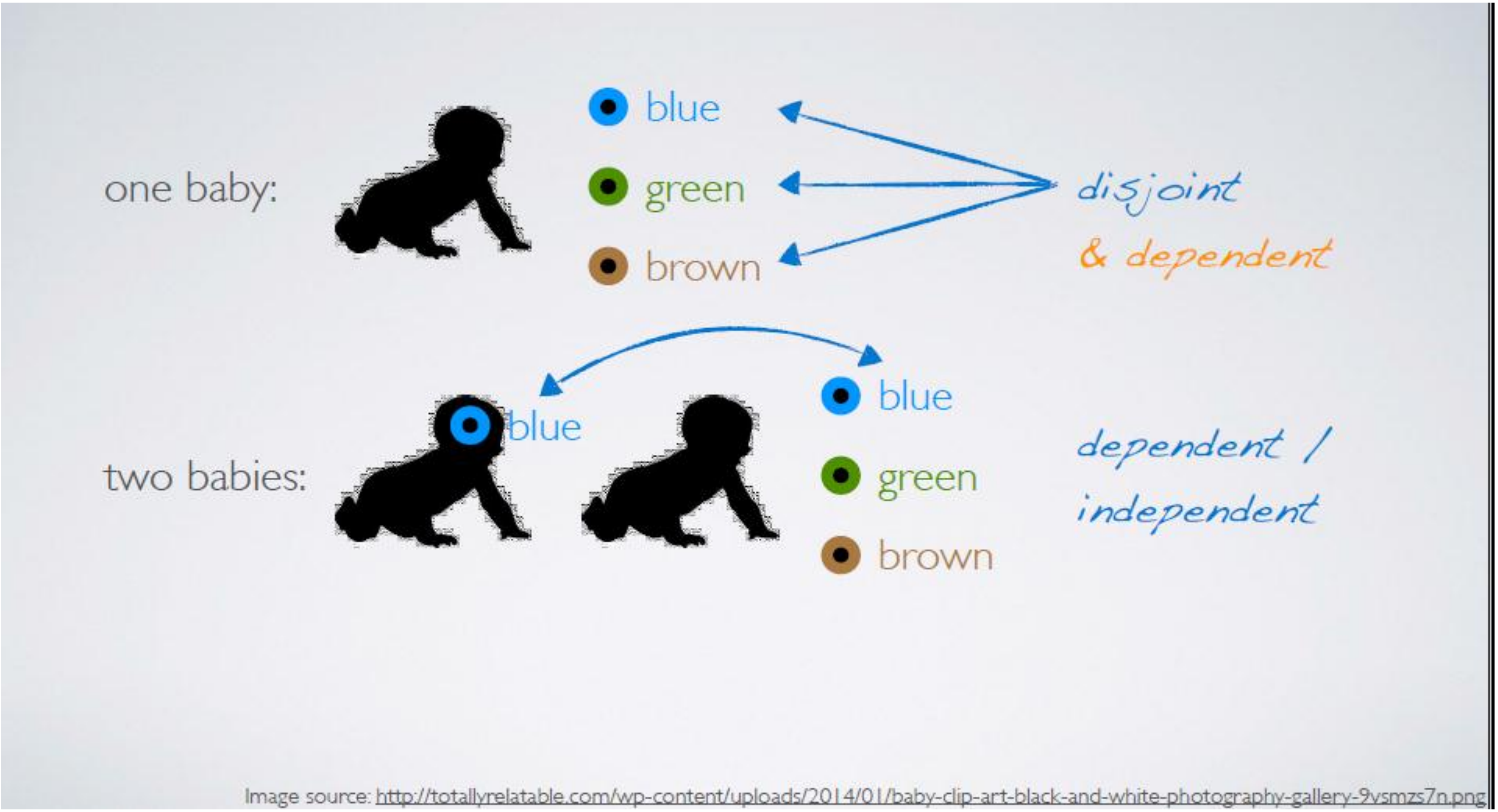
two events that are disjoint (mutually exclusive) cannot happen at the same time

$P(A \text{ and } B) = 0$

two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other

$P(A \mid B) = P(A)$

Independence



Practice

In 2013, SurveyUSA interviewed a random sample of 500 NC residents asking them whether they think widespread gun ownership protects law abiding citizens from crime, or makes society more dangerous.

- 58% of all respondents said it protects citizens.
- 67% of White respondents,
- 28% of Black respondents,
- and 64% of Hispanic respondents shared this view.

Opinion on gun ownership and race ethnicity are most likely __________?

(a) complementary
(b) mutually exclusive
(c) independent
(d) dependent
(e) disjoint

$P(\text{protects citizens}) = 0.58$

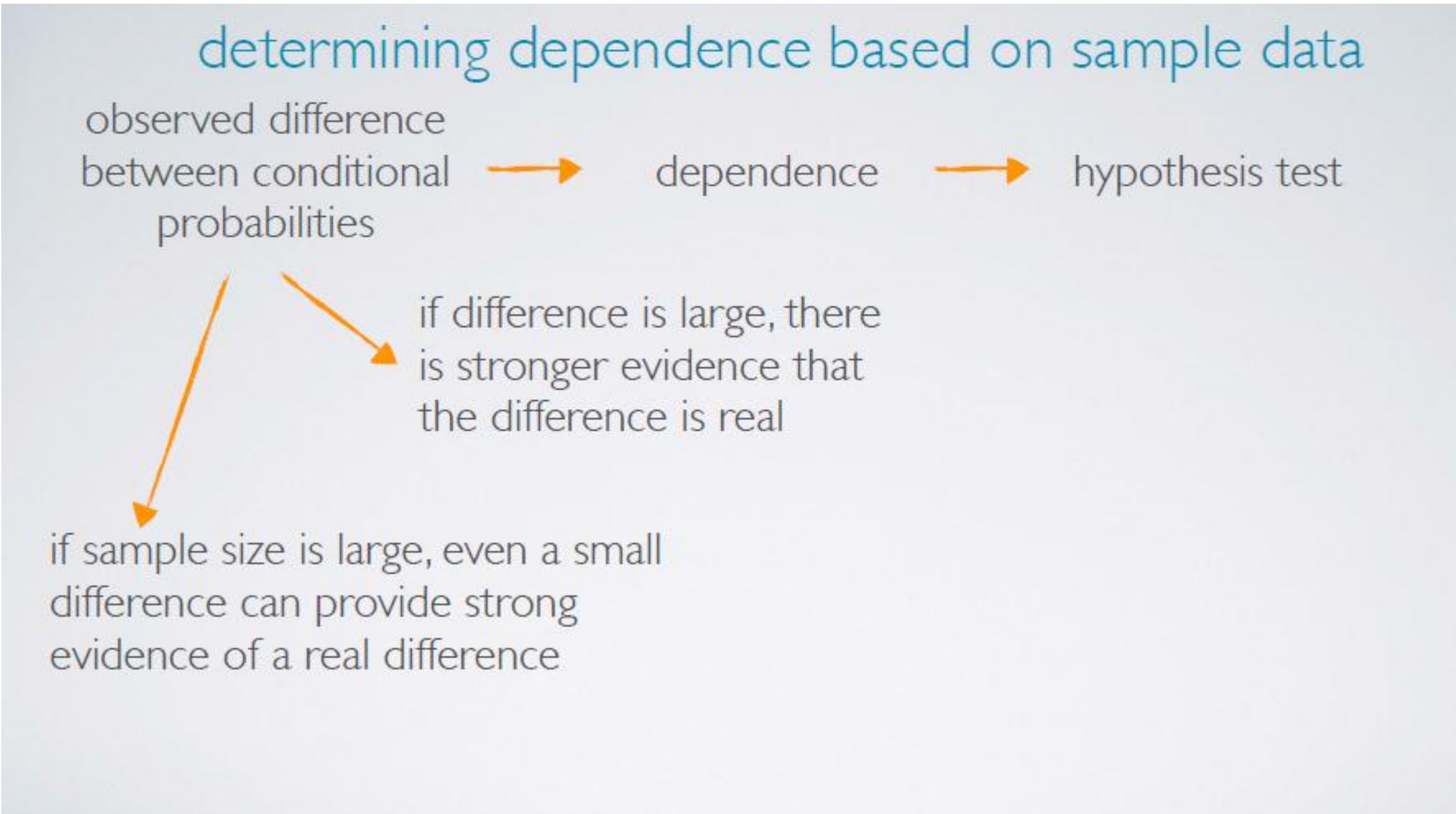
$P(\text{protects citizens} \mid \text{White}) = 0.67$

$P(\text{protects citizens} \mid \text{Black}) = 0.28$

$P(\text{protects citizens} \mid \text{Hispanic}) = 0.64$

Link to poll: http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea26421b

Determining dependence



Determining dependence

Product rule for independent events:

If A and B are independent, P(A and B) = P(A) x P(B)

You toss a coin twice, what is the probability of getting two tails in a row?

P(two tails in a row) =

= P(T on the 1st toss) × P(T on the 2nd toss)

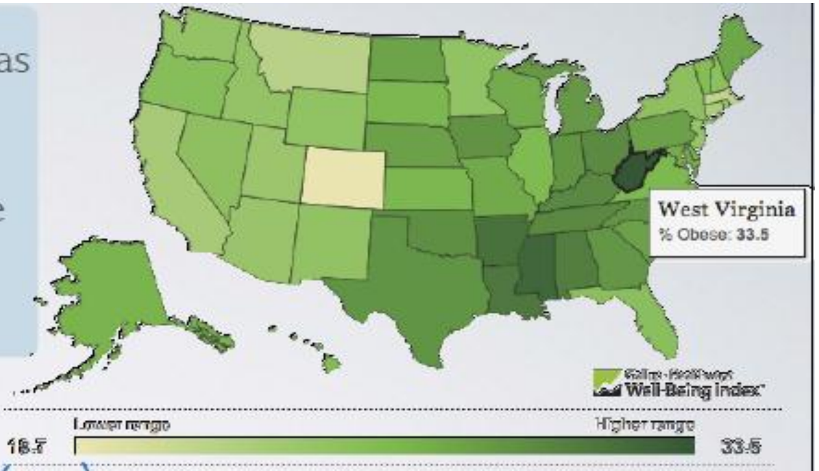
= (1/2) × (1/2)

= 1/4

Note: If $A_1, A_2, \ldots, A_k$ are independent, $P(A_1 \text{ and } A_2 \text{ and } \ldots A_k) = P(A_1) \times P(A_2) \times \ldots \times P(A_k)$

Practice

A 2012 Gallup poll suggests that West Virginia has the highest obesity rate among US states, with 33.5% of West Virginians being obese. Assuming that the obesity rate stayed constant, what is the probability that two randomly selected West Virginians are both obese? *independent*



$P(\text{obese}) = 0.335$

$P(\text{both obese}) = P(\text{1st obese}) \times P(\text{2nd obese})$

$= 0.335 \times 0.335$

≈ 0.11

Image source + Link to poll: http://www.surveyusa.com/client/PollReport.aspx?g=a5f460ef-bba9-484b-8579-1101ea2642

Example: probability

- sample spaces
- disjoint, complementary, and independent events
- addition rule for unions of events
- multiplication rule for joint probabilities for independent events

Example

The World Values Survey is an ongoing worldwide survey that polls the world population about perceptions of life, work, family, politics, etc.

The most recent phase of the survey that polled 77,882 people from 57 countries estimates that a 36.2% of the world's population agree with the statement "Men should have more right to a job than women."

The survey also estimates that 13.8% of people have a university degree or higher, and that 3.6% of people fit both criteria.

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036$

Survey: http://www.worldvaluessurvey.org/

Example

(1) Are agreeing with the statement "Men should have more right to a job than women" and having a university degree or higher disjoint events?

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036 \neq 0 \longrightarrow$ not disjoint

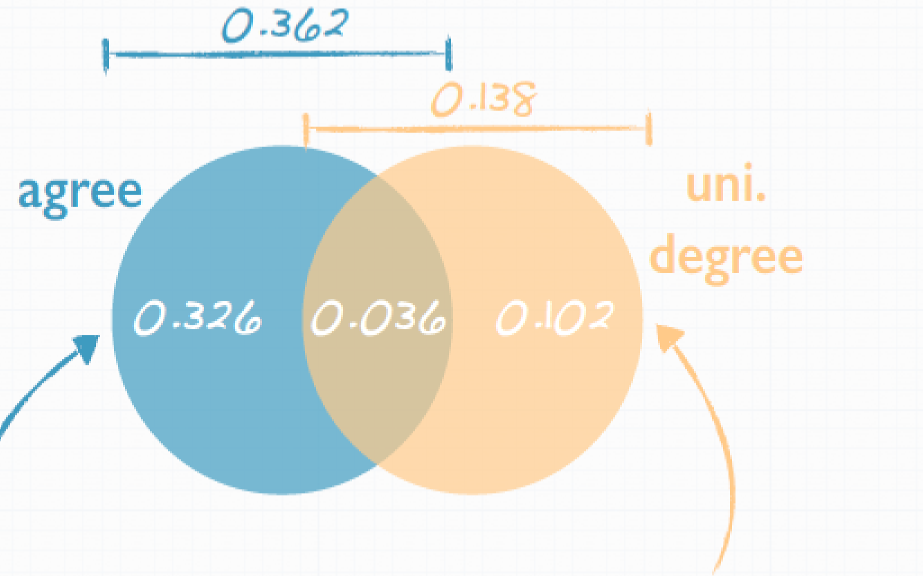
Example

(2) Draw a Venn diagram summarizing the variables and their associated probabilities.

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036$



$0.362 - 0.036 = 0.326$

$0.138 - 0.036 = 0.102$

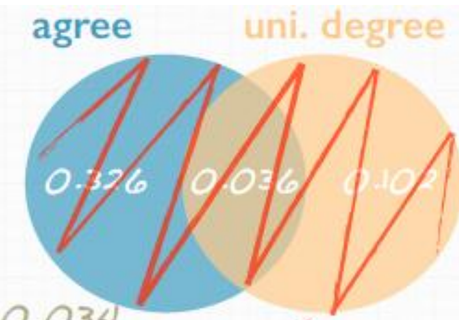
Example

(3) What is the probability that a randomly drawn person has a university degree or higher or agrees with the statement about men having more right to a job than women?

$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.034$



General addition rule:
P(A or B) = P(A) + P(B) - P(A and B)

$P(\text{agree or uni. degree})$

$= P(\text{agree}) + P(\text{uni. degree}) - P(\text{agree \& uni. degree})$

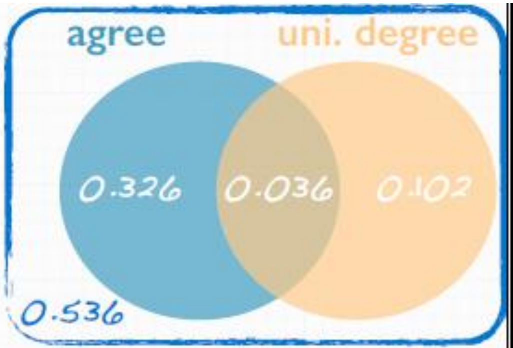
$= 0.362 + 0.138 - 0.036$

$= 0.464$

$0.326 + 0.036 + 0.102 = 0.464$

Example

(4) What percent of the world population do not have a university degree and disagree with the statement about men having more right to a job than women?



P(agree) = 0.362

P(uni. degree) = 0.138

P(agree & uni. degree) = 0.036

P(agree or uni. degree) = 0.464

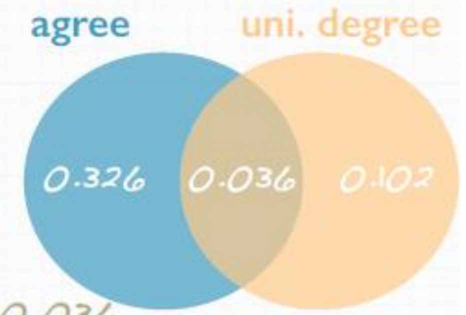
P(neither agree nor uni. degree)

= 1 − P(agree or uni. degree)

= 1 − 0.464 = 0.536

Example

(5) Does it appear that the event that someone agrees with the statement is independent of the event that they have a university degree or higher?



$P(\text{agree}) = 0.362$

$P(\text{uni. degree}) = 0.138$

$P(\text{agree \& uni. degree}) = 0.036$

Product rule for independent events:

If A and B are independent, P(A and B) = P(A) x P(B)

$$P(\text{agree \& uni. degree}) \ ?=? \ P(\text{agree}) \times P(\text{uni. degree})$$

$$0.036 \ ?=? \ 0.362 \times 0.138$$

$$0.036 \neq 0.05 \longrightarrow \text{not independent}$$

Example

(6) What is the probability that at least 1 in 5 randomly selected people agree with the statement about men having more right to a job than women?

$P(\text{agree}) = 0.362$

$S = \{0, 1, 2, 3, 4, 5\} \longrightarrow S = \{0, \text{at least } 1\}$

$P(\text{at least 1 agree}) = 1 - P(\text{none agree})$

$= 1 - P(\underline{D}\ \underline{D}\ \underline{D}\ \underline{D}\ \underline{D}) \longrightarrow$

$P(\text{disagree})$

$= 1 - P(\text{agree})$

$= 1 - 0.362$

$= 0.638$

$= 1 - 0.638^5$

$= 1 - 0.106 = 0.894$

Conditional probability

study

ADOLESCENTS' UNDERSTANDING OF SOCIAL CLASS

study examining teens' beliefs about social class

sample: 48 working class and 50 upper middle class 16-year-olds

study design:
- "objective" assignment to social class based on self-reported measures of both parents' occupation and education, and household income
- "subjective" association based on survey questions

Study reference: Goodman, Elizabeth, et al. "Adolescents' understanding of social class: a comparison of white upper middle class and working class youth." Journal of adolescent health 27.2 (2000): 80-83.

Conditional probability

results:		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

Marginal probability

marginal

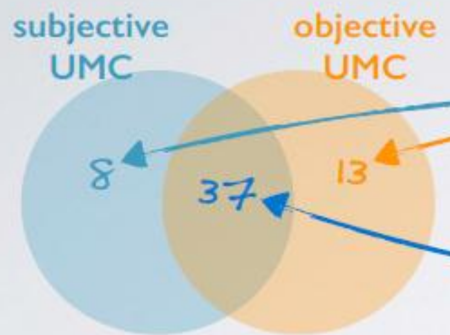
		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle class	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student's <u>objective</u> social class position is upper middle class?

$P(\text{obj UMC})$
$= 50 / 98 \approx 0.51$

Joint probability

joint



		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student's objective position *and* subjective identity are both upper middle class?

$P(\text{obj UMC \& subj UMC})$

$= 37 / 98 \approx 0.38$

Conditional probability

conditional

		[illegible] social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

What is the probability that a student who is objectively in the working class associates with upper middle class?

P(subj UMC | obj WC)

$= 8 / 48 \approx 0.17$

Conditional probability

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

		objective social class position		
		working class	upper middle class	Total
subjective social class identity	poor	0	0	0
	working class	8	0	8
	middle class	32	13	45
	upper middle	8	37	45
	upper class	0	0	0
	Total	48	50	98

$$P(\text{subj UMC} \mid \text{obj WC}) = \frac{P(\text{subj UMC \& obj WC})}{P(\text{obj WC})} = \frac{8 / 98}{48 / 98} = 8 / 48 \approx 0.17$$

Practice

The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services.

The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English at home, and 4.2% fall into both categories.

Based on this information, what percent of Americans live below the poverty line given that they speak a language other than English at home?

$P(\text{below PL} \mid \text{speak non-Eng}) = ?$

$$= \frac{P(\text{below PL \& speak non-Eng})}{P(\text{speak non-Eng})} = \frac{0.042}{0.207} \approx 0.2$$

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Data source: U.S. Census Bureau, 2010 American Community Survey 1-Year Estimates, Characteristics of People by Language Spoken at Home.

Practice

Product rule for independent events:

If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

General product rule:

$$P(A \text{ and } B) = P(A \mid B) \times P(B)$$

Practice

independence and conditional probabilities

Generically, if $P(A|B) = P(A)$ then the events A and B are said to be independent.

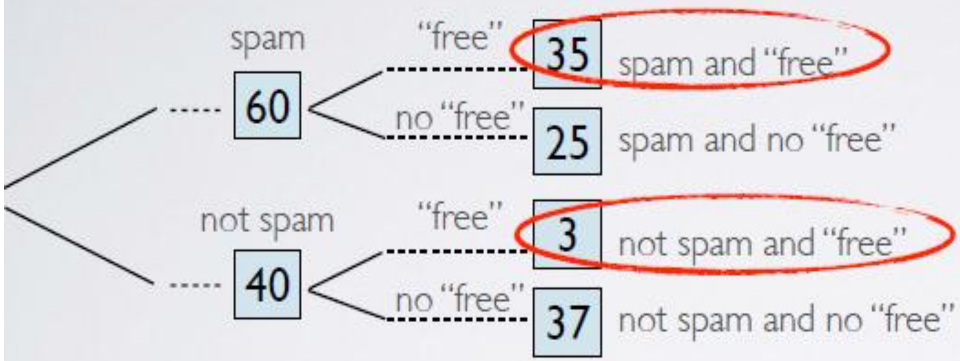
- Conceptually: Giving B doesn't tell us anything about A.
- Mathematically: If events A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$. Then,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Probability trees

P(A | B) → P(B | A)

You have 100 emails in your inbox: 60 are spam, 40 are not. Of the 60 spam emails, 35 contain the word "free". Of the rest, 3 contain the word "free". If an email contains the word "free", what is the probability that it is spam?



$$P(\text{spam} \mid \text{"free"}) = \frac{35}{35 + 3} = 0.92$$

Probability trees

As of 2009, Swaziland had the highest HIV prevalence in the world. 25.9% of this country's population is infected with HIV. The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

$P(HIV) = 0.259$

$P(+ \mid HIV) = 0.997$ $P(- \mid \text{no } HIV) = 0.926$

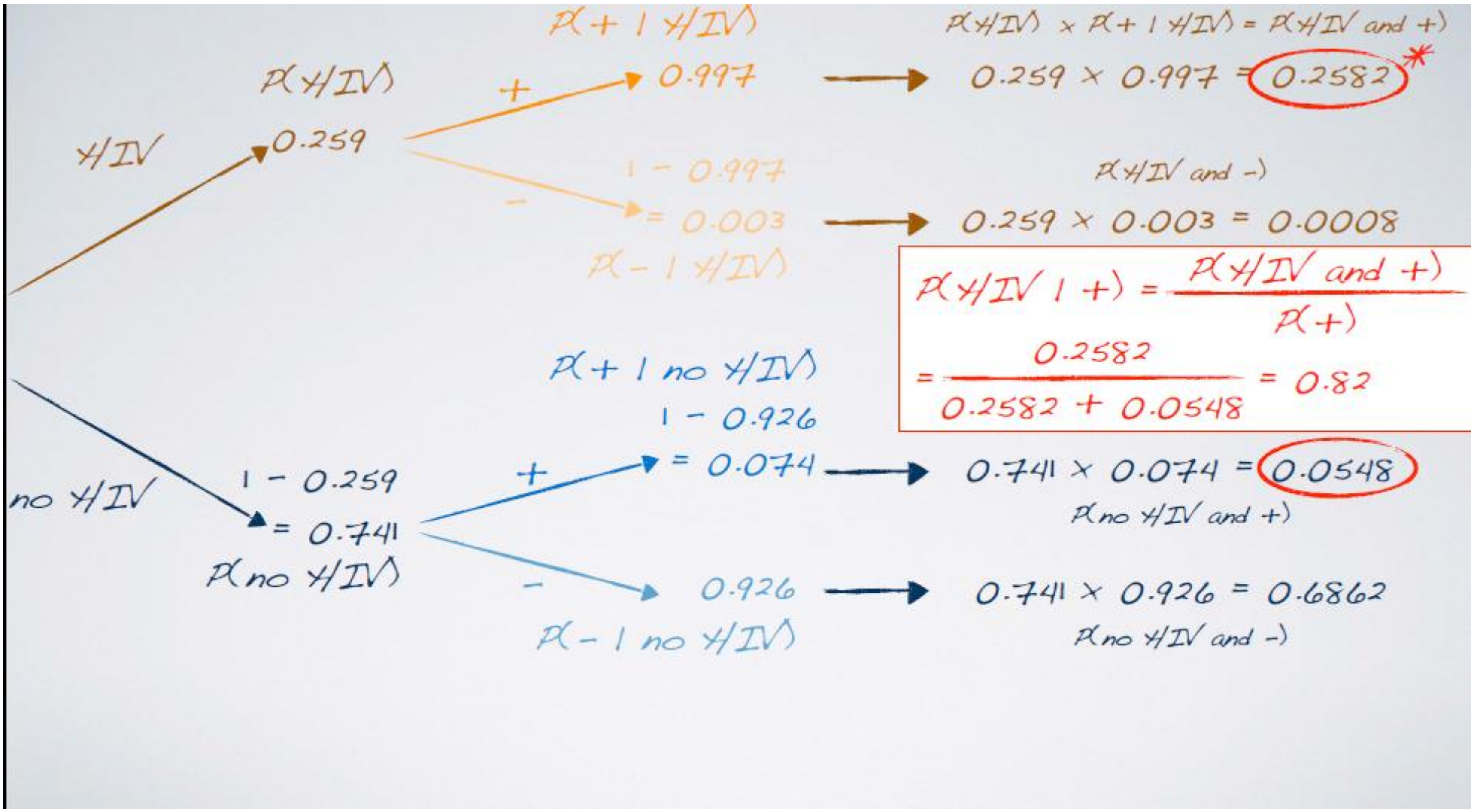
tree diagram!

$P(HIV \mid +) = ?$

Image source: http://en.wikipedia.org/wiki/File:Location_Swaziland_AU_Africa.svg

Data source: CIA Factbook, Country Comparison: HIV/AIDS - Adult Prevalence Rate https://www.cia.gov/library/publications/the-world-factbook/rankorder/2155rank.html

Probability trees



Probability trees

If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

$$P(HIV \mid +) = 0.82$$

There is an 82% chance
that an individual from Swaziland
who has tested positive
actually carries HIV.

Bayesian inference

What is the probability of rolling ≥4 with a 6-sided die?

$S = \{1,2,3,\boxed{4,5,6}\}$

$P(\geq 4) = 3/6 = 1/2 = 0.5$

What is the probability of rolling ≥4 with a 12-sided die?

$S = \{1,2,3,\boxed{4,5,6,7,8,9,10,11,12}\}$

$P(\geq 4) = 9/12 = 3/4 = 0.75$

Bayesian inference

"good die"

Say you're playing a game where the goal is to roll ≥ 4. If you could get your pick, which die would you prefer to play this game with?

good die

(a)

(b)

$P(\geq 4) = 0.5$

$P(\geq 4) = 0.75$

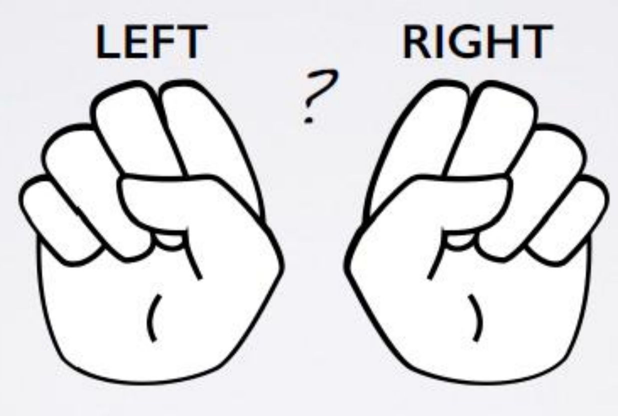
Bayesian inference

rules

hypotheses and decisions

		Truth	
		Right good, Left bad	Right bad, Left good
Decision	pick Right	You win the game!	You lose :(
	pick Left	You lose :(	You win the game!

Bayesian inference



before you collect data

Before we collect any data, you have no idea if I am holding the good die (12-sided) on the right hand or the left hand. Then, what are the probabilities associated with the following hypotheses?

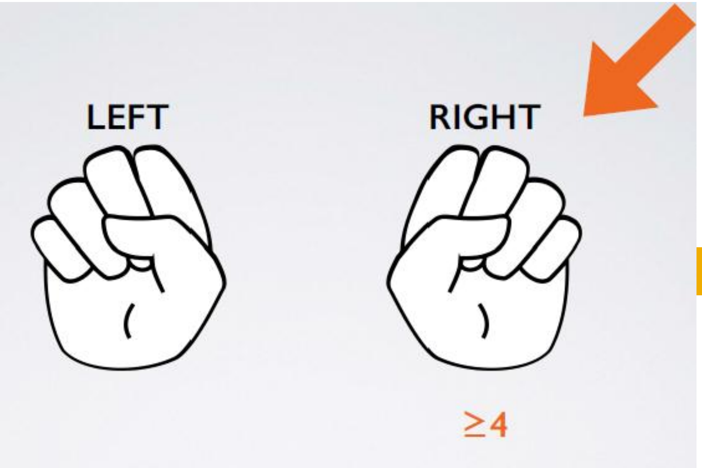
H_1: good die on the Right (bad die on the Left)

H_2: good die on the Left (bad die on the Right)

	P(H_1: good die on the Right)	P(H_2: good die on the Left)
(a)	0.33	0.67
(b)	0.5	0.5
(c)	0	1
(d)	0.25	0.75

prior

Bayesian inference



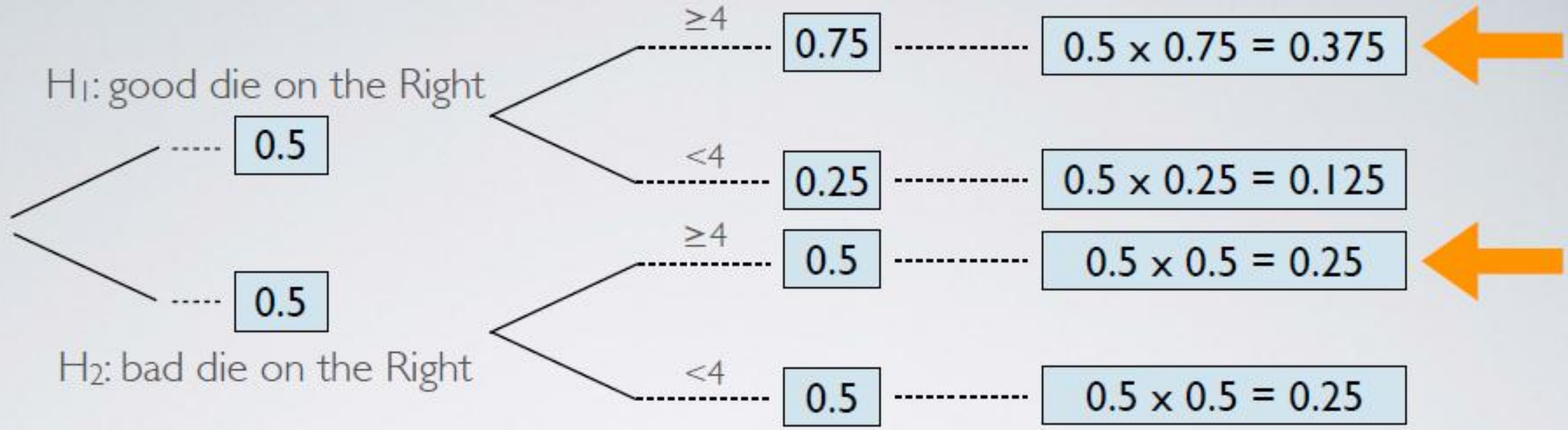
after you see the data

You chose the right hand, and you won (rolled a number ≥4). Having observed this data point how, if at all, do the probabilities you assign to the same set of hypotheses change?

H_1: good die on the Right (bad die on the Left)
H_2: good die on the Left (bad die on the Right)

	P(H_1: good die on the Right)	P(H_2: good die on the Left)
(a)	0.5	0.5
(b)	more than 0.5	less than 0.5
(c)	less than 0.5	more than 0.5

Bayesian inference



$$P(H_1\text{: good die on the Right} \mid \text{you rolled} \geq 4 \text{ with the die on the Right}) =$$

$$= \frac{P(\text{good Right} \ \& \geq 4 \text{ Right})}{P(\geq 4 \text{ Right})} = \frac{0.375}{0.375 + 0.25} = 0.6$$

Bayesian inference

posterior

- The probability we just calculated is also called the posterior probability.

  P(H_1: good die on the Right | you rolled ≥4 with the die on the Right)
- Posterior probability is generally defined as P(hypothesis | data).
- It tells us the probability of a hypothesis we set forth, given the data we just observed.
- It depends on both the prior probability we set and the observed data.
- This is different than what we calculated at the end of the randomization test on gender discrimination – the probability of observed or more extreme data given the null hypothesis being true, i.e. P(data | hypothesis), also called a p-value.

Bayesian inference

updating the prior

- In the Bayesian approach, we evaluate claims iteratively as we collect more data.
- In the next iteration (roll) we get to take advantage of what we learned from the data.
- In other words, we update our prior with our posterior probability from the previous iteration.

updated:

$P(H_1$: good die on the Right)	$P(H_2$: good die on the Left)
0.6	0.4

Bayesian inference

recap

- Take advantage of prior information, like a previously published study or a physical model.
- Naturally integrate data as you collect it, and update your priors.
- Avoid the counter-intuitive definition of a p-value:

  P(observed or more extreme outcome | H0 is true)
- Instead base decisions on the posterior probability:

  P(hypothesis is true | observed data)
- A good prior helps, a bad prior hurts, but the prior matters less the more data you have.
- More advanced Bayesian techniques offer flexibility not present in Frequentist models.

Example: Bayesian inference

- setting a prior
- collecting data
- obtaining a posterior
- updating the prior with the previous posterior

Example

American Cancer Society estimates that about 1.7% of women have breast cancer.
http:// www.cancer.org/ cancer/ cancerbasics/ cancer-prevalence

Susan G. Komen For The Cure Foundation states that mammography correctly identifies about 78% of women who truly have breast cancer.
http:// ww5.komen.org/ BreastCancer/ AccuracyofMammograms.html

An article published in 2003 suggests that up to 10% of all mammograms are false positive.
http:// www.ncbi.nlm.nih.gov/ pmc/ articles/ PMC1360940

$P(bc) = 0.017$

$P(+ \mid bc) = 0.78$

$P(+ \mid no\ bc) = 0.10$

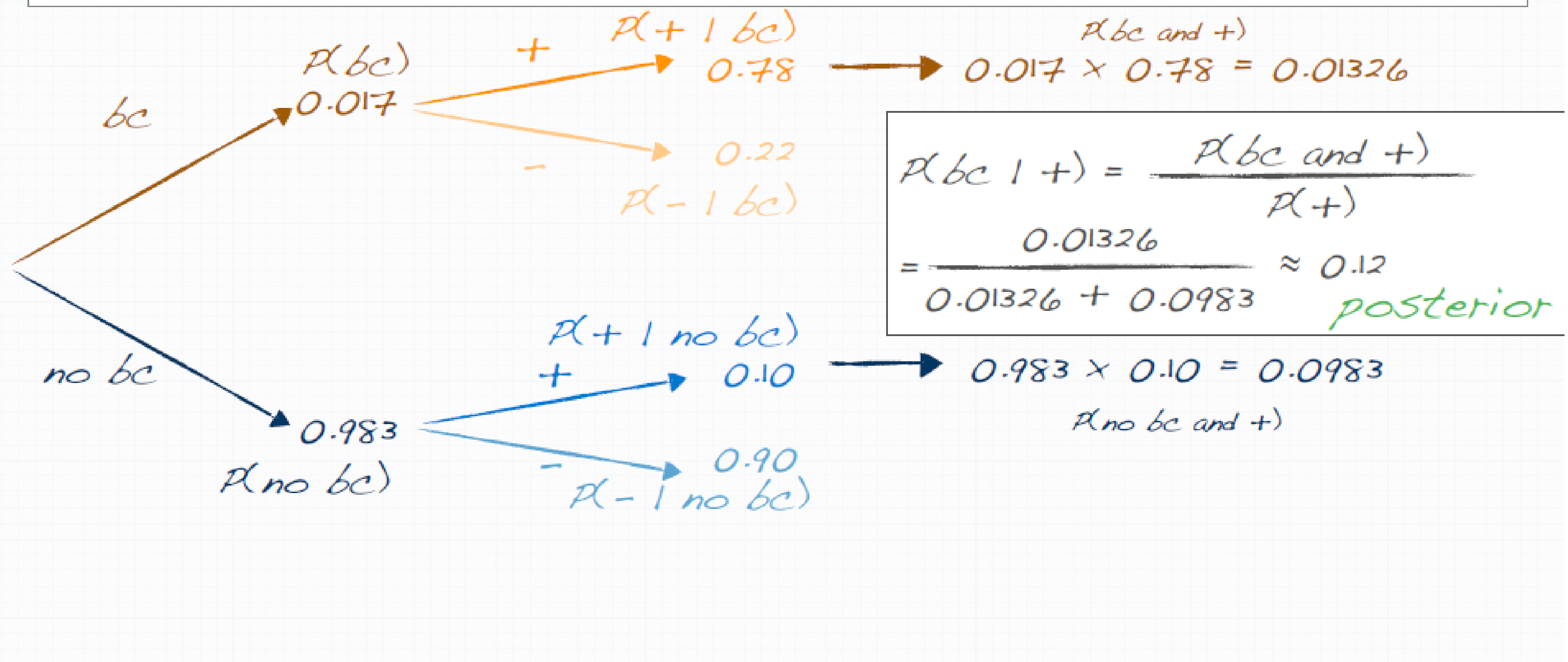
Example

Prior to any testing and any information exchange between the patient and the doctor, what probability should a doctor assign to a female patient having breast cancer?

$P(bc) = 0.017 \longrightarrow$ prior

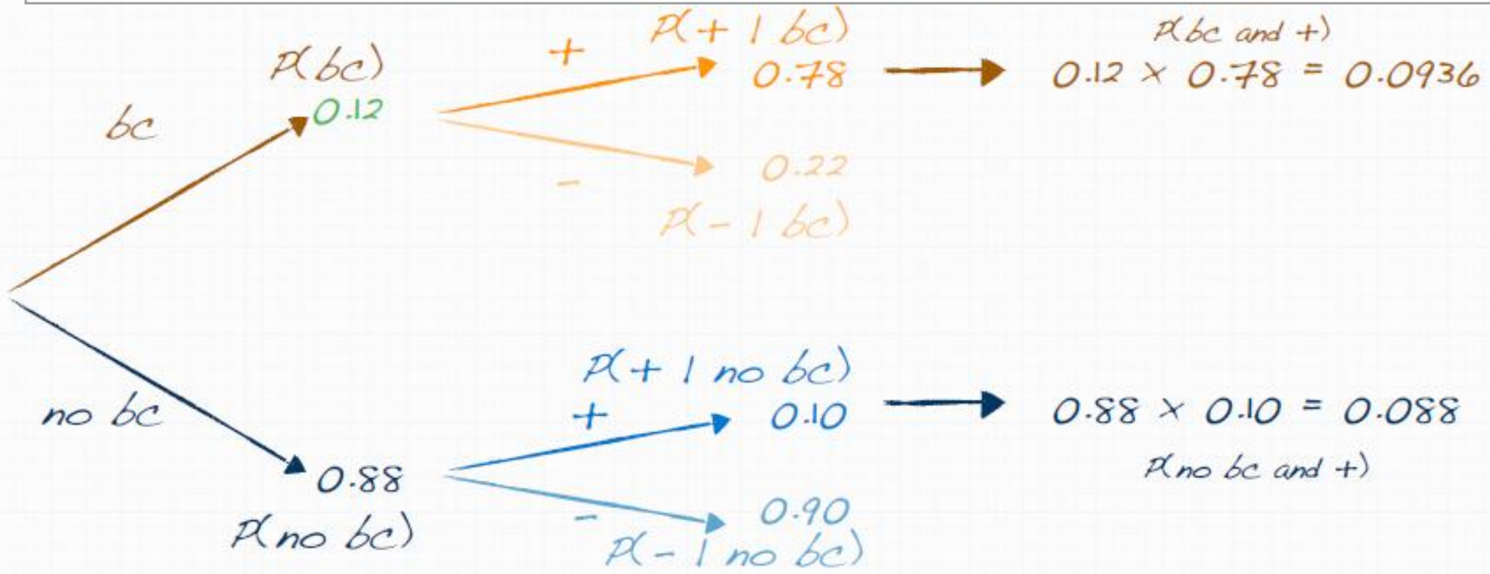
Example

When a patient goes through breast cancer screening there are two competing claims: patient has cancer and patient doesn't have cancer. If a mammogram yields a positive result, what is the probability that patient has cancer? $P(bc \mid +) = ?$



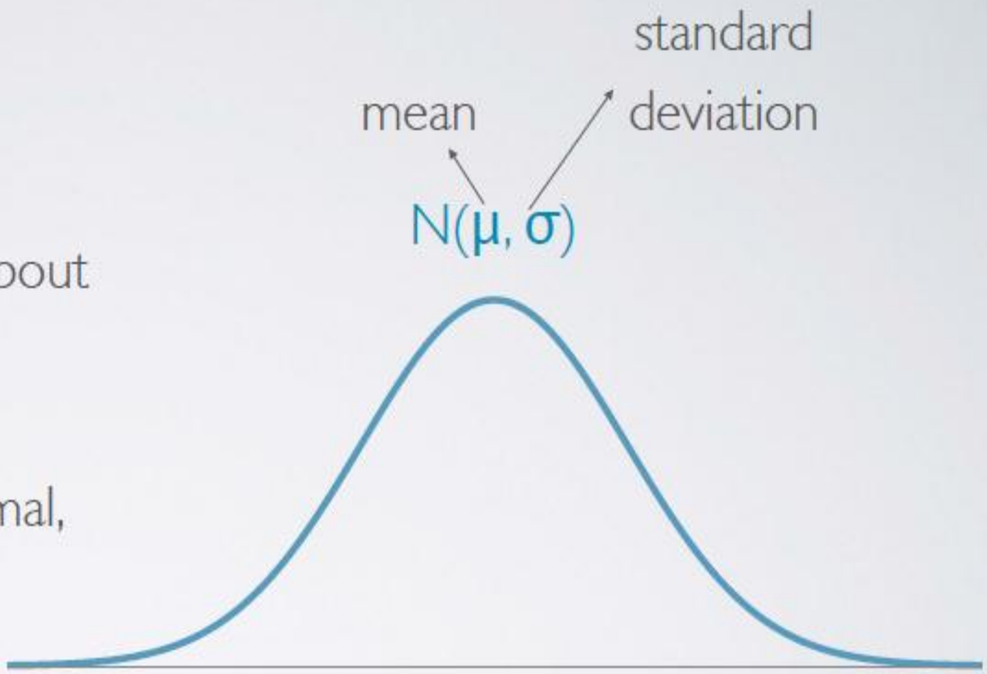
Example

Since a positive mammogram doesn't necessarily mean that the patient actually has breast cancer, the doctor might decide to re-test the patient. What is the probability of having breast cancer if this second mammogram also yields a positive result?

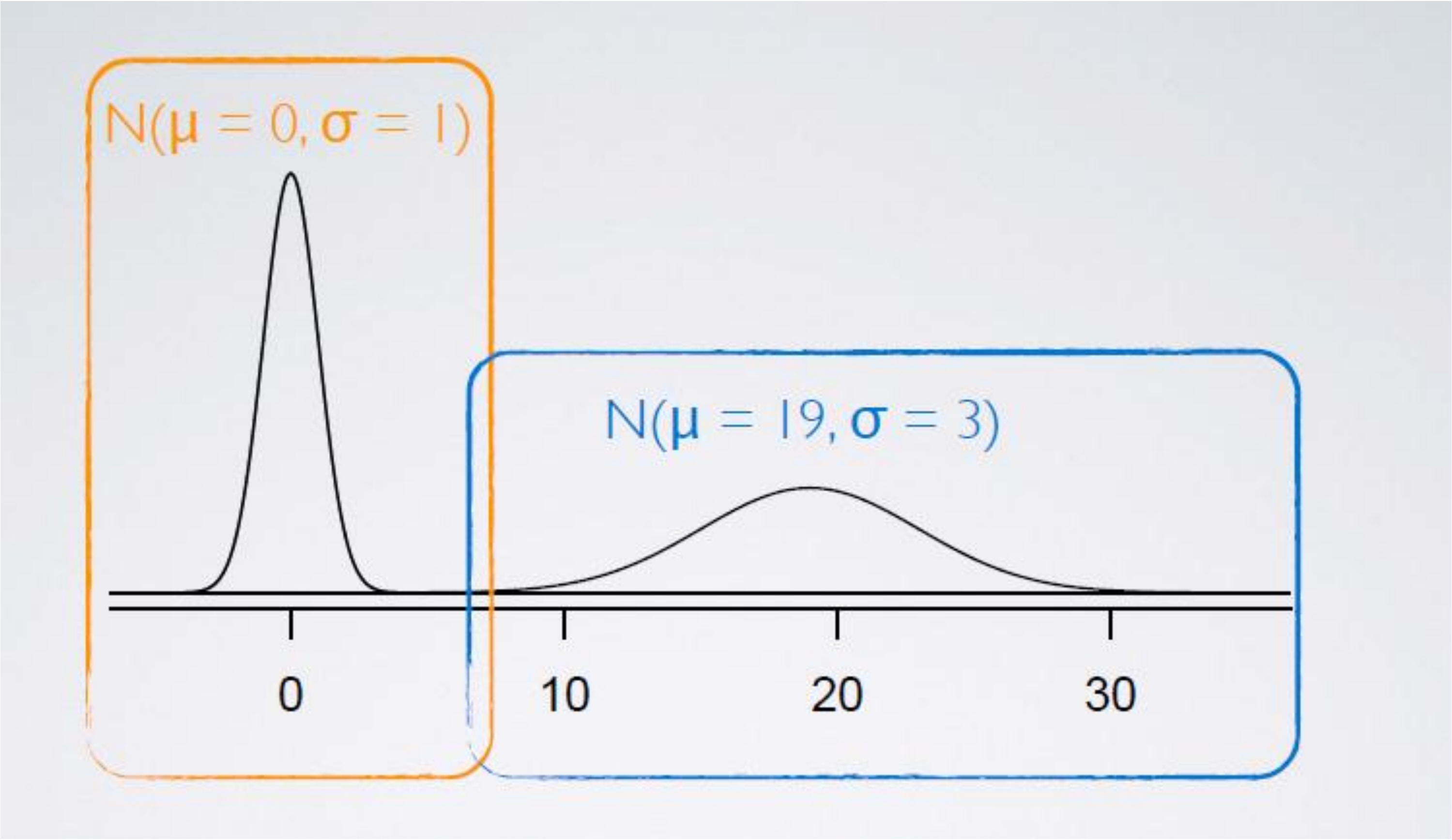


Normal distribution

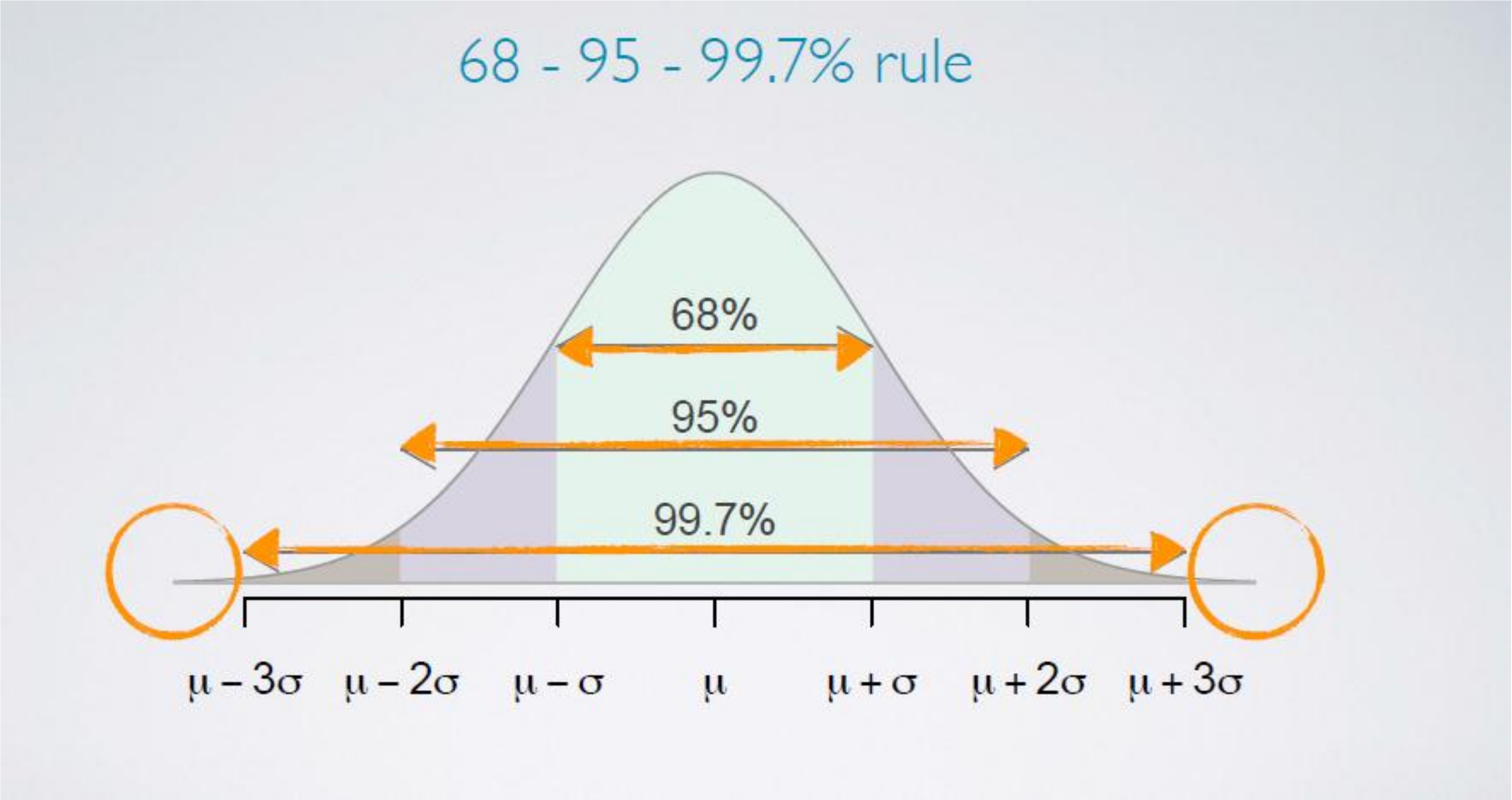
- unimodal and symmetric
  - bell curve
- follows very strict guidelines about how variably the data are distributed around the mean
- many variables are nearly normal, but none are exactly normal



Normal distribution



Normal distribution



Practice

A doctor collects a large set of heart rate measurements that approximately follow a normal distribution. He only reports 3 statistics, the mean = 110 beats per minute, the minimum = 65 beats per minute, and the maximum = 155 beats per minute. Which of the following is most likely to be the standard deviation of the distribution?

(a) 5 → $110 \pm (3 \times 5) = (95, 125)$

(b) 15 → $110 \pm (3 \times 15) = (65, 155)$

(c) 35 → $110 \pm (3 \times 35) = (5, 215)$

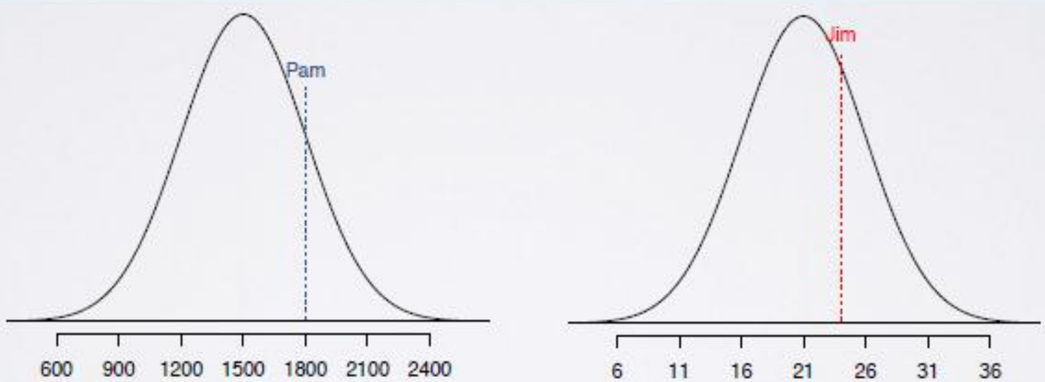
(d) 90 → $110 \pm (3 \times 90) = (-160, 380)$

Practice

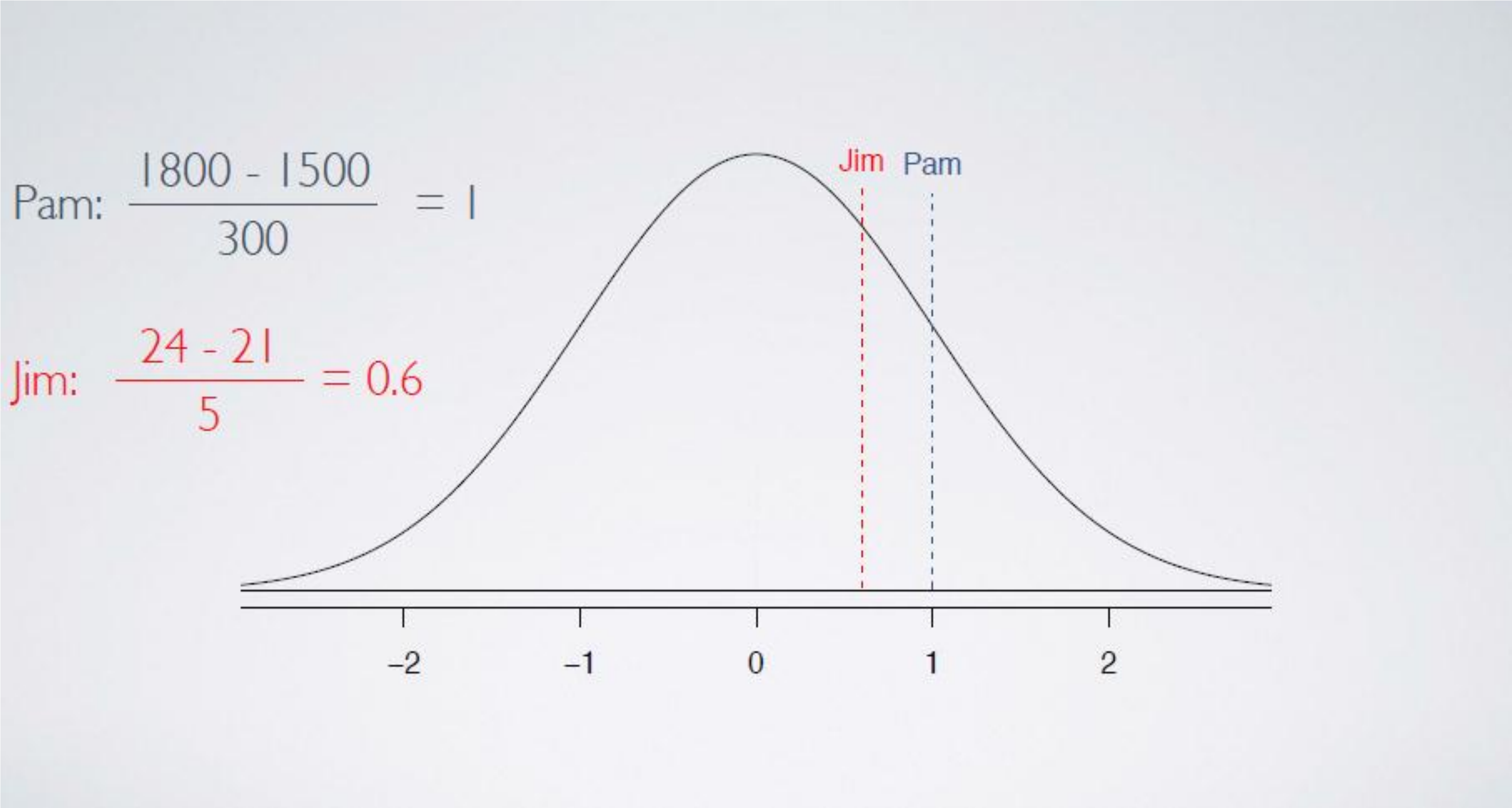
A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

SAT scores ~ N(mean = 1500, SD = 300)

ACT scores ~ N(mean = 21, SD = 5)



Practice



Practice

standardizing with Z scores

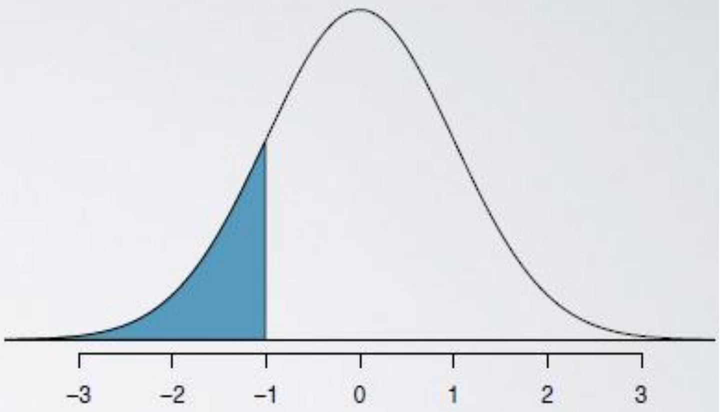
- standardized (Z) score of an observation is the number of standard deviations it falls above or below the mean
- Z score of mean = 0
- unusual observation: $|Z| > 2$
- defined for distributions of any shape

$$Z = \frac{\text{observation - mean}}{\text{SD}}$$

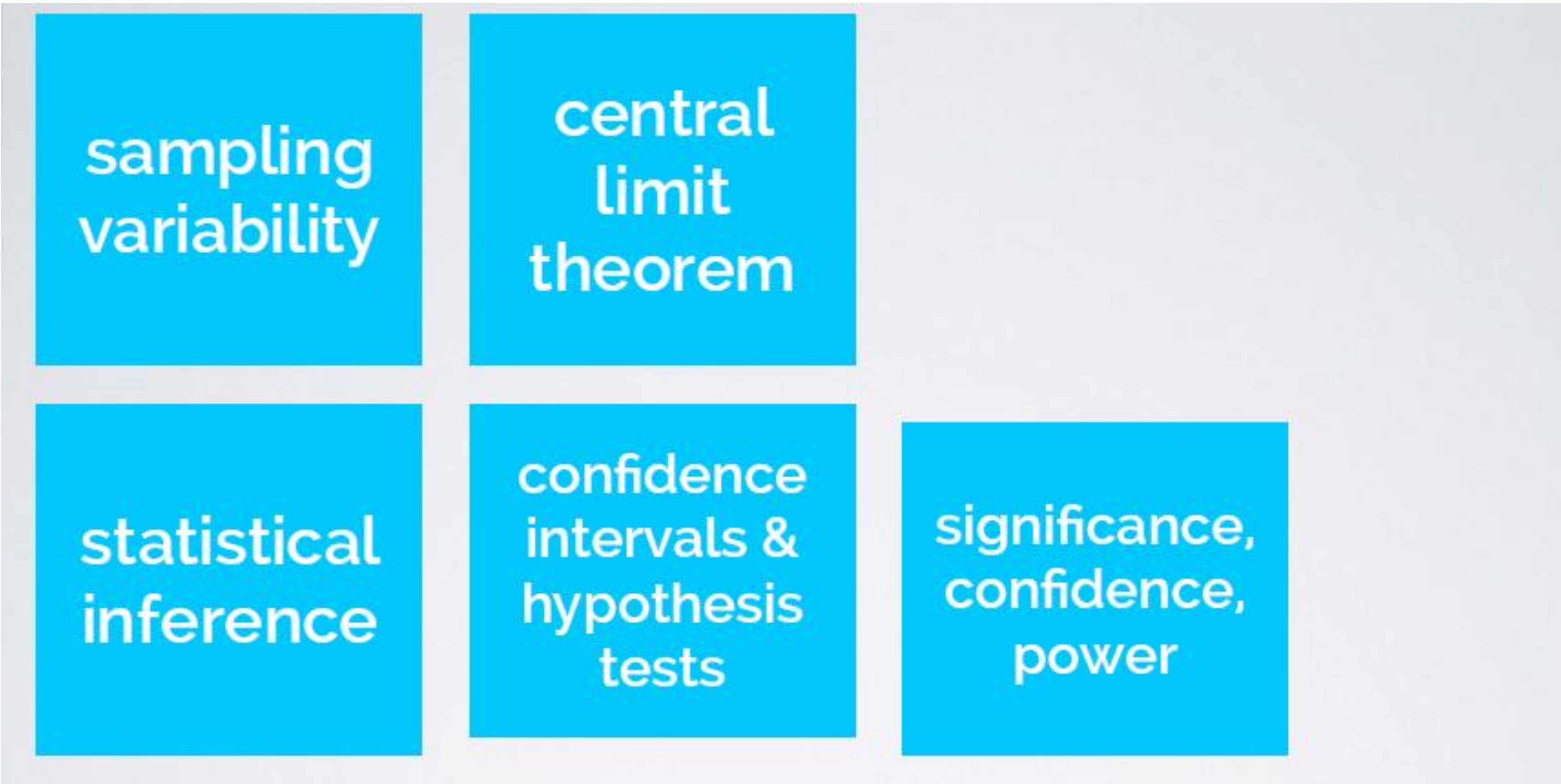
Practice

percentiles

- when the distribution is normal, Z scores can be used to calculate percentiles
- percentile is the percentage of observations that fall below a given data point
- graphically, percentile is the area below the probability distribution curve to the left of that observation.



Foundation for inference



Young, Unemployed, and Optimistic

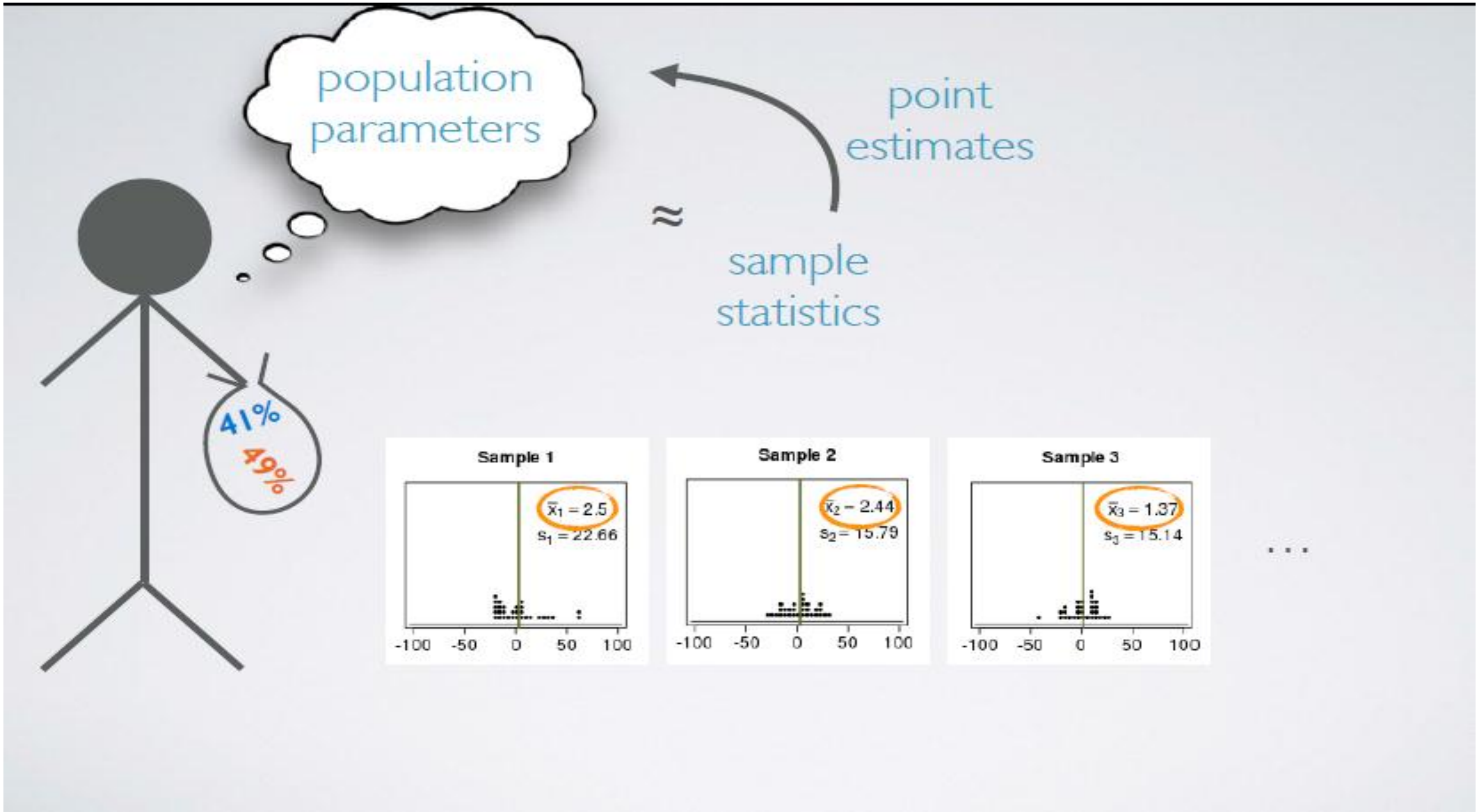
Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

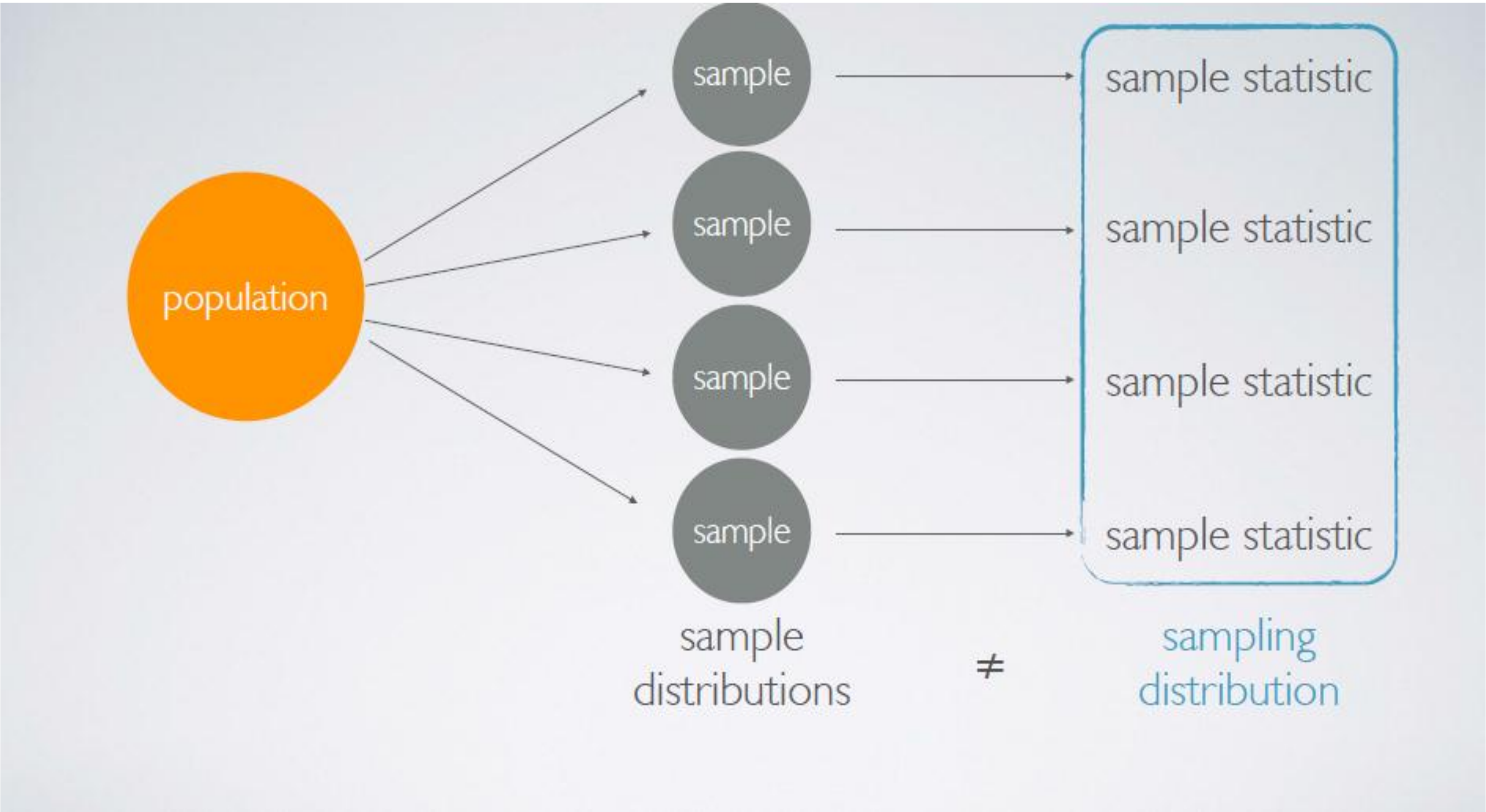
Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience.

The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States [...] Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

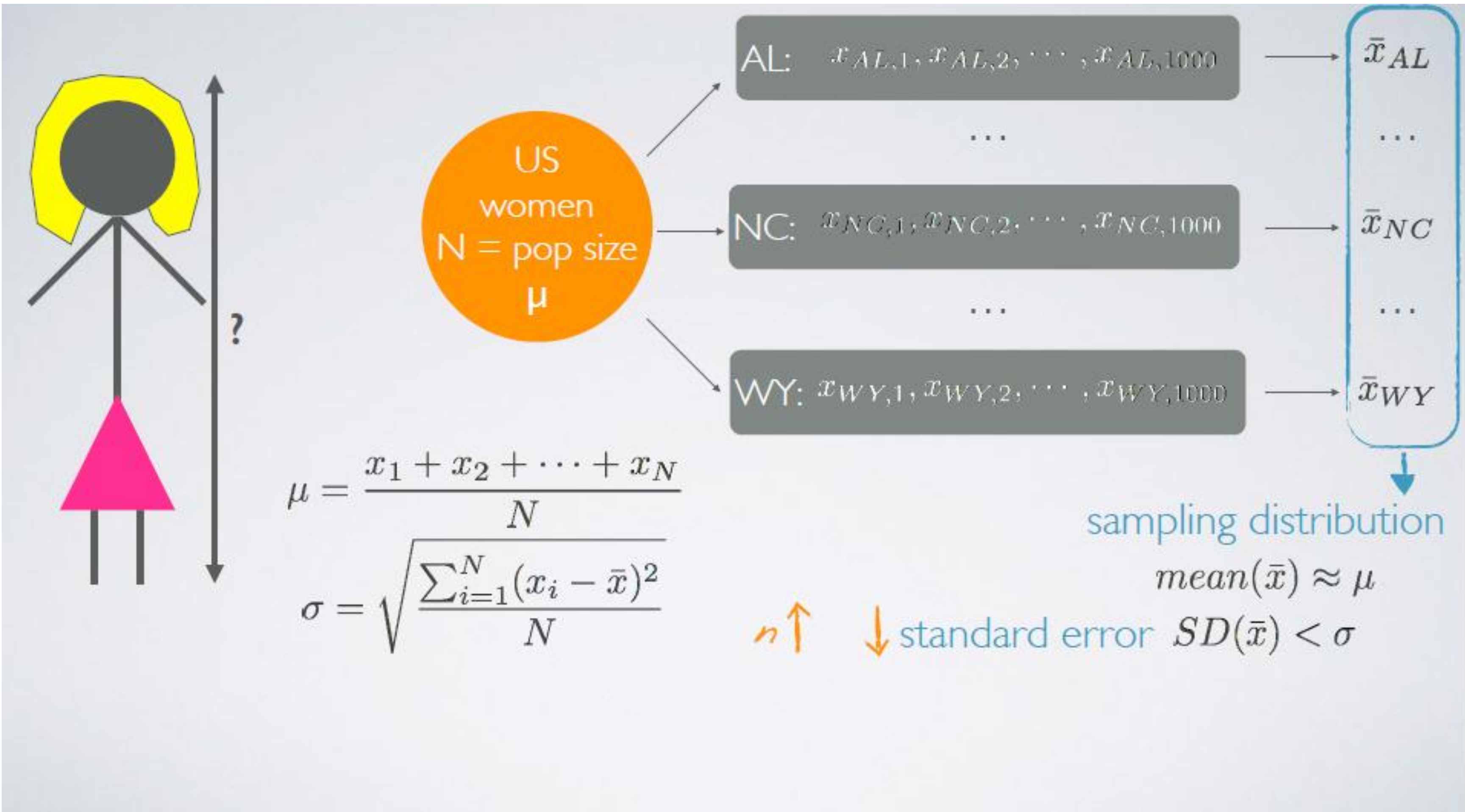
- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.
- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.


population
parameters
point
estimates
≈
sample
statistics
41%
49%
Sample 1
$\bar{x}_1 = 2.5$
$s_1 = 22.66$
-100 -50 0 50 100
Sample 2
$\bar{x}_2 = 2.44$
$s_2 = 15.79$
-100 -50 0 50 100
Sample 3
$\bar{x}_3 = 1.37$
$s_3 = 15.14$
-100 -50 0 50 100
. . .

Sampling distribution



Sampling distribution



Central Limit Theorem

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.

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

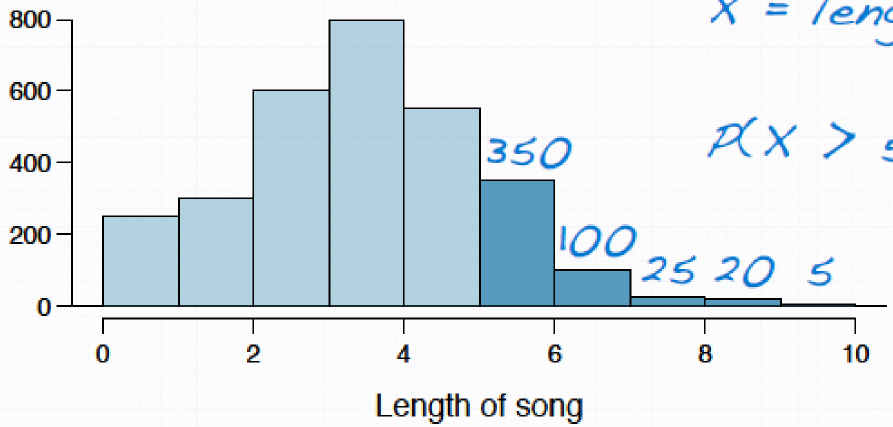
shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

Example

Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



X = length of one song

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000}$$

$= 500 / 3000$

≈ 0.17

Example

I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

6 hours = 360 minutes

$P(X_1 + X_2 + \dots + X_{100} > 360 \text{ min}) = ?$

$P(\bar{X} > 3.6) = ?$

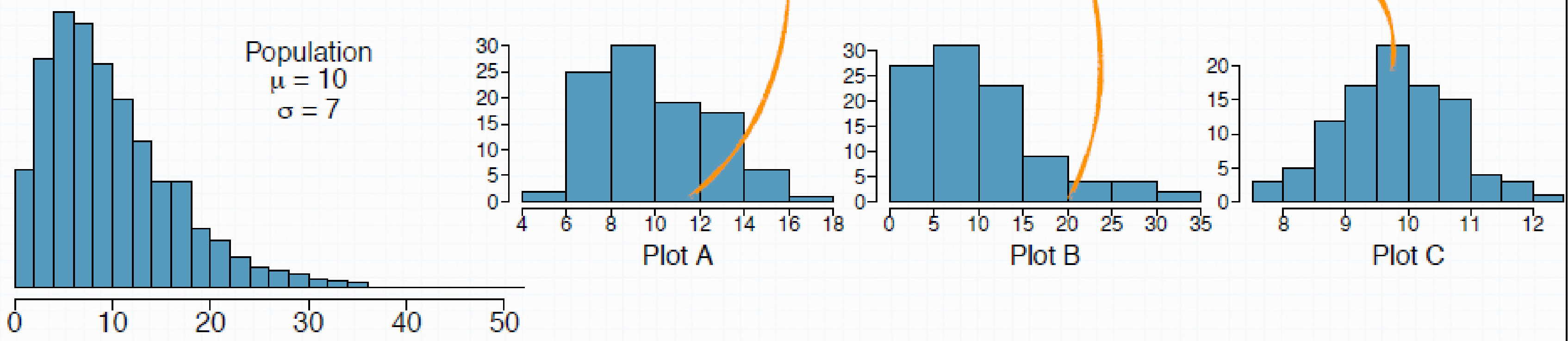
$\bar{X} \sim N(\text{mean} = \mu = 3.45, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

$Z = \frac{3.6 - 3.45}{0.163} = 0.92$

$P(Z > 0.92) = 0.179$

Example

Four plots: Determine which plot (A, B, or C) is which.
(1) The distribution for a population ($\mu = 10$, $\sigma = 7$),
(2) a single random sample of 100 observations from this population,
(3) a distribution of 100 sample means from random samples with size 7, and
(4) a distribution of 100 sample means from random samples with size 49.



Confidence interval (for a mean)

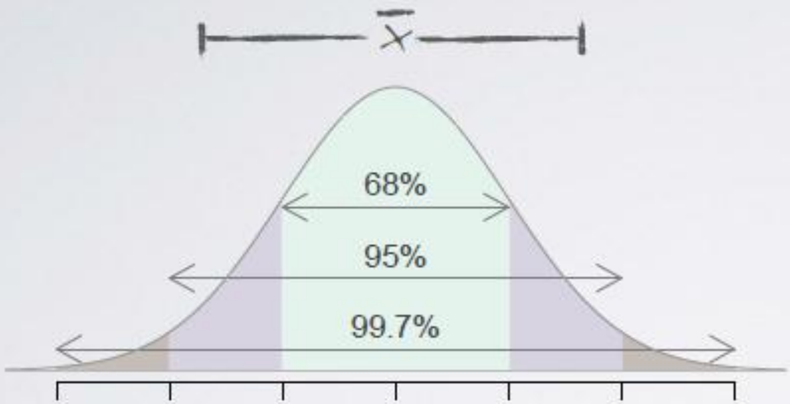
A plausible range of values for the population parameter is called a confidence interval.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Confidence interval

Central Limit Theorem (CLT):

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



approximate 95% CI: $\bar{x} \pm \boxed{2SE}$

margin of error (ME)

Confidence interval

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).

$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$

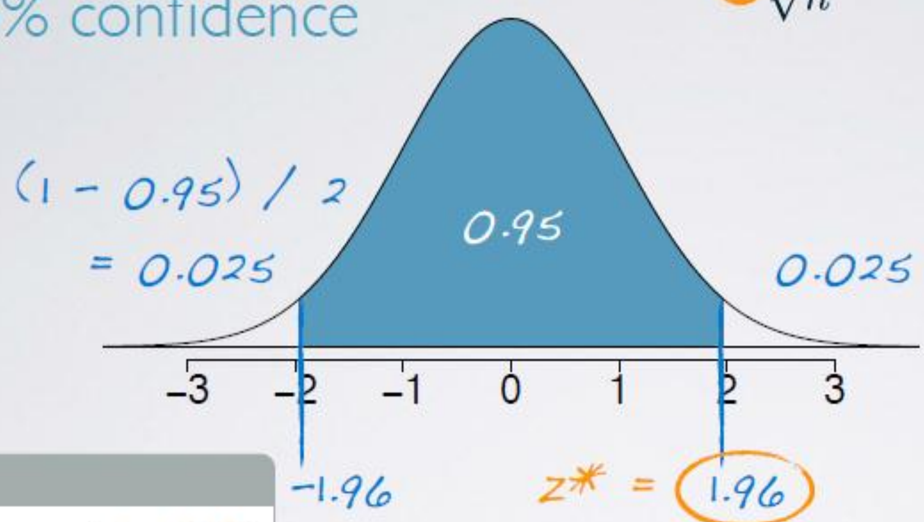
Conditions for this confidence interval:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
2. ***Sample size/skew:*** $n \geq 30$, larger if the population distribution is very skewed.

Confidence interval

finding the critical value
95% confidence

$$\bar{x} \pm z^{\star}\frac{s}{\sqrt{n}}$$



R
```
> qnorm(0.025)
[1] -1.96
```

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

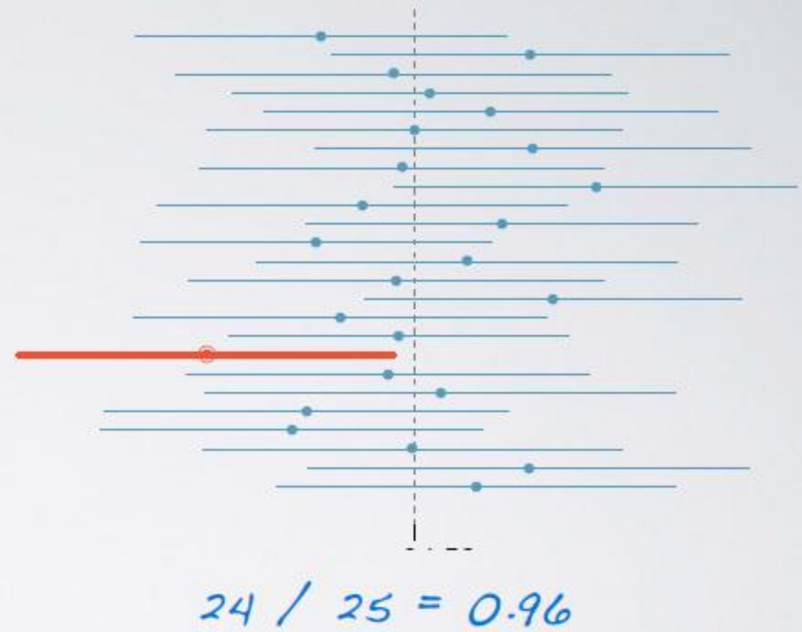
Confidence level

confidence level

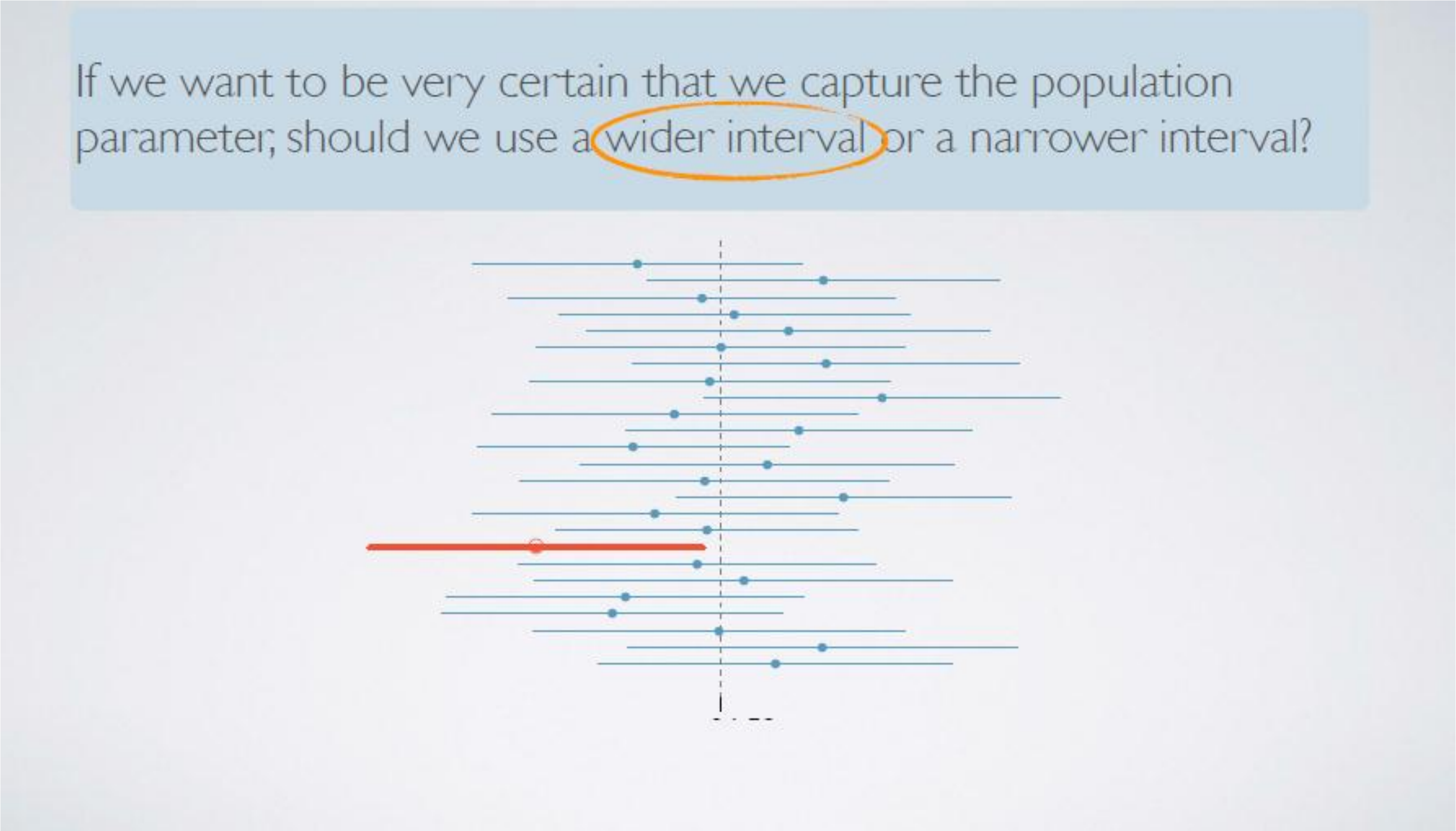
- Suppose we took many samples and built a confidence interval from each sample using the equation

$$point\ estimate \pm 1.96 \times SE$$

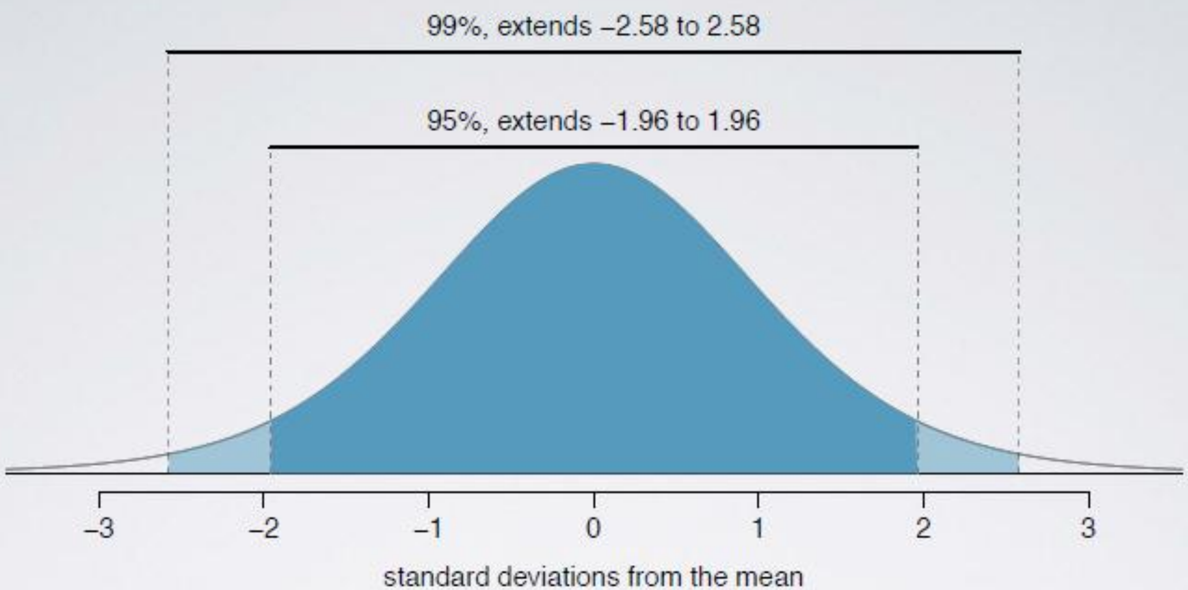
- Then about 95% of those intervals would contain the true population mean (μ).
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.



Confidence level

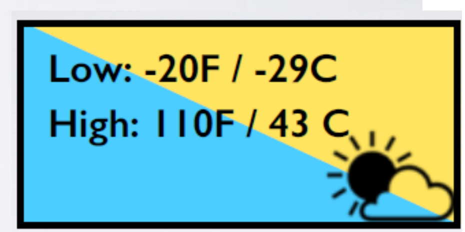


Confidence level



CL ↑ width ↑ accuracy ↑

precision ↓



Confidence level

How can we get the best of both worlds — higher precision and higher accuracy?

increase sample size

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.

Required sample size

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.

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

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

$CL = 90\%$

$z^* = 1.65$

$\sigma = 18$

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

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

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?

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

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

$$4n = 56 \times 4 = 224$$

Examples: Confidence interval

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.

Examples: Confidence interval

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.

In this context, what does a 95% confidence level mean?

95% of random samples of 1,151 Americans will yield CIs that capture the true population mean of number of bad mental health days per month.

Examples: Confidence interval

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.

Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be narrower or wider than the 95% confidence interval?

As CL increases so does the width of the confidence interval, so wider.

Examples: Confidence interval

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.

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

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.

Heart: http://commons.wikimedia.org/wiki/File:Heart-padlock.svg

Examples: Confidence interval

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

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

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

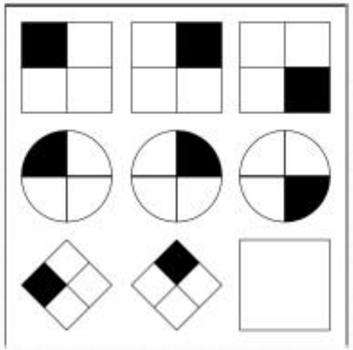
We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.

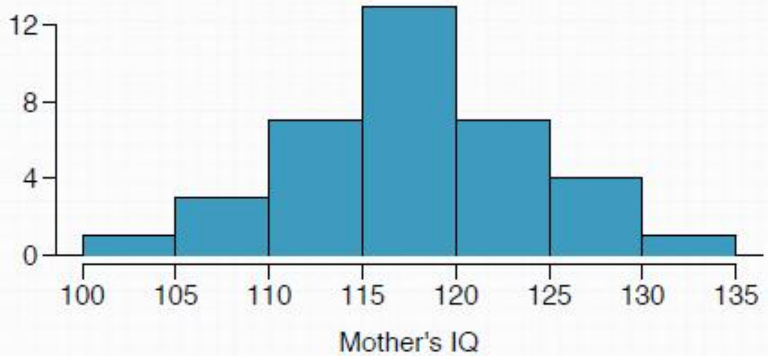
Hypothesis testing framework

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods — methods that rely on the CLT
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Example

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.





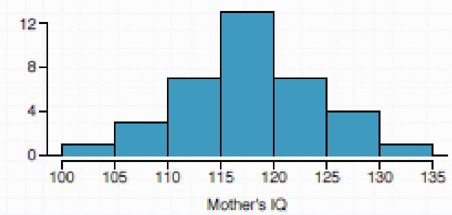
n	36
min	101
mean	118.2
sd	6.5
max	131

Example

Perform a hypothesis test to evaluate if these 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, which is 100. Use a significance level of 0.01.

1. **Set the hypotheses** μ = average IQ score of mothers of gifted children

$H_0: \mu = 100 \quad H_A: \mu \neq 100$



n	36
min	101
mean	118.2
sd	6.5
max	131

2. **Calculate the point estimate**

$\bar{x} = 118.2$

3. **Check conditions**

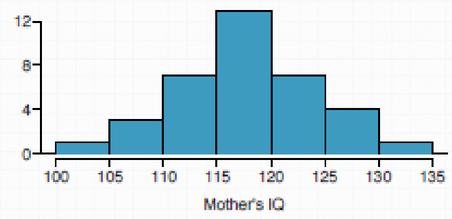
1. random & 36 < 10% of all gifted children -> independence
2. n > 30 & sample not skewed -> nearly normal sampling distribution

Example

$H_0: \mu = 100$

$H_A: \mu \neq 100$

$\bar{x} = 118.2$



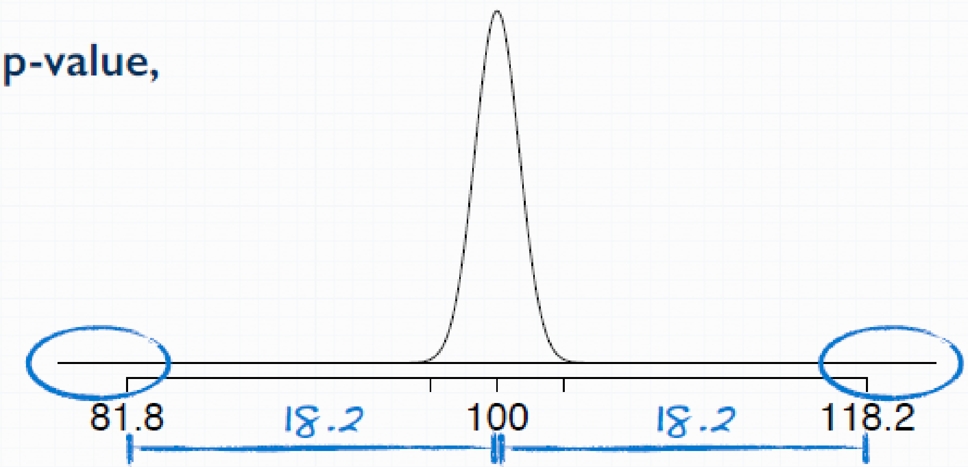
n	36
min	101
mean	118.2
sd	6.5
max	131

$\bar{X} \sim N\left(\mu = 100, SE = \frac{s}{\sqrt{n}} = \frac{6.5}{\sqrt{36}} \approx 1.083\right)$

4. **Draw sampling distribution, shade p-value, calculate test statistic**

$Z = \frac{118.2 - 100}{1.083} = 16.8$

$p\text{-value} \approx 0$



Example

5. **Make a decision, and interpret it in context of the research question**

p-value is very low -> strong evidence against the null

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.

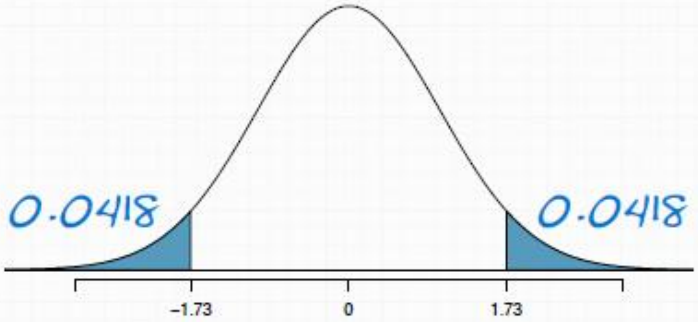
Example

A statistics student interested in sleep habits of domestic cats took a random sample of 144 cats and monitored their sleep. The cats slept an average of 16 hours / day. According to online resources domestic dogs sleep, on average, 14 hours day. We want to find out if these data provide convincing evidence of different sleeping habits for domestic cats and dogs with respect to how much they sleep. The test statistic is 1.73.

$\bar{x} = 16$

$H_0: \mu = 14$

$H_A: \mu \neq 14$



$p\text{-value} = 0.0418 \times 2$

$= 0.0836$

Example

What is the interpretation of this p-value in context of these data?

$= P($observed or more extreme outcome $| H_0$ 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$) = 0.0836$

$n = 144$

$\bar{x} = 16$

$H_0: \mu = 14$

$H_A: \mu \neq 14$

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	$\hat{p}_1 - \hat{p}_2$

Inference for other estimators

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.

Inference for other estimators

confidence intervals
for nearly normal point estimates

$$point\ estimate \pm z^{\star} \times SE$$

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.

$\hat{p} = 0.33$

$SE = 0.014$

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

$$0.33 \pm 1.96 \times 0.014$$

$$0.33 \pm 0.027$$

$$(0.303, 0.357)$$

Practice

hypothesis testing
for nearly normal point estimates

$$Z = \frac{point\ estimate - null\ value}{SE}$$

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**

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

2. **Calculate the point estimate**

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

3. **Check conditions**

Practice

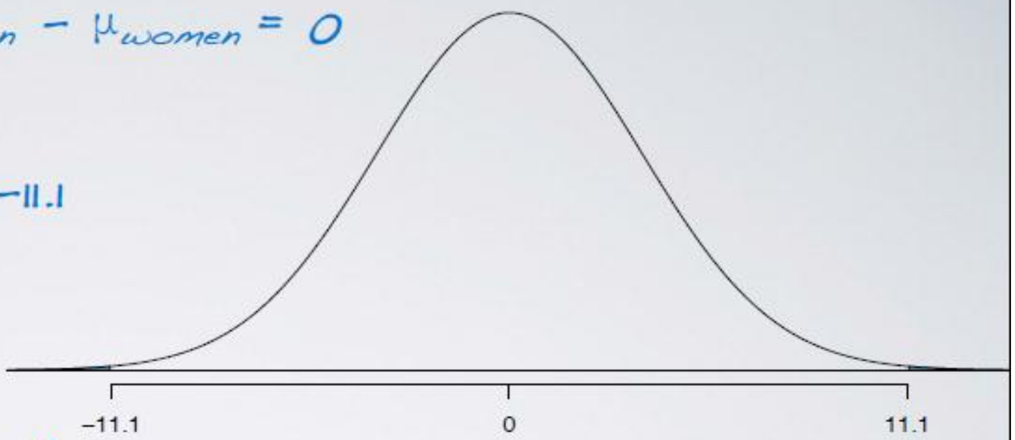
$H_0: \mu_{men} = \mu_{women} \longrightarrow \mu_{men} - \mu_{women} = 0$

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

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

$$Z = \frac{-11.1 - 0}{0.114} \approx -97.36$$

p-value $\approx 0 \longrightarrow$ Reject H_0



These data provide convincing evidence that the average BF% of men and women are different.

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.

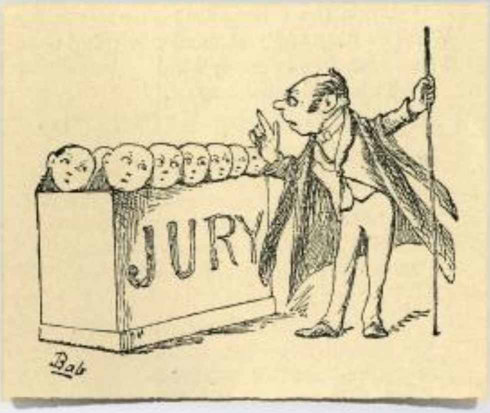
Decision errors

hypothesis test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0 : Defendant is innocent

H_A : Defendant is guilty



Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty Type 2 error
- Declaring the defendant guilty when they are actually innocent Type 1 error

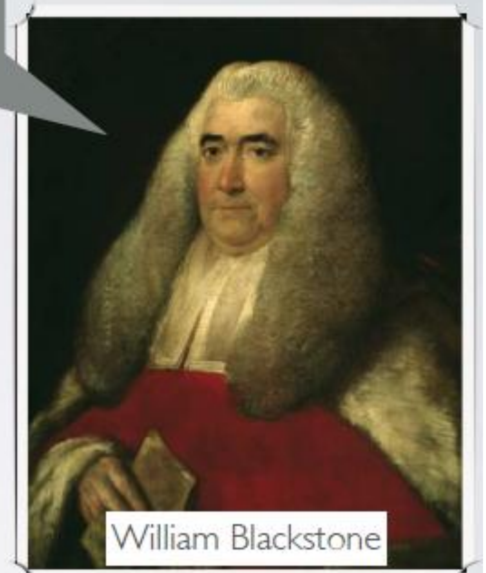
Jury: http://upload.wikimedia.org/wikipedia/commons/5/5d/Trial_by_Jury_Usher.jpg

Decision errors

"better that ten guilty persons escape than that one innocent suffer"

Which error is the worst error to make?

- Type 2 : Declaring the defendant innocent when they are actually guilty
- Type 1 : Declaring the defendant guilty when they are actually innocent


William Blackstone

William Blackstone: http://en.wikipedia.org/wiki/File:SirWilliamBlackstone.jpg

Decision errors

type 1 error rate

- We reject H_0 when the p-value is less than 0.05 ($\alpha = 0.05$).
- This means that, for those cases where H_0 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 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α – increasing α increases the Type 1 error rate.

Decision errors

If Type I 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.

choosing α

If a Type 2 Error is relatively more dangerous or much more costly, choose a higher significance level (e.g. 0.10).

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

Scale: http://commons.wikimedia.org/wiki/File:US_Department_of_Justice_Scales_Of_Justice.svg

Decision errors

goal:
keep α and β low

		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 $1 - \beta$

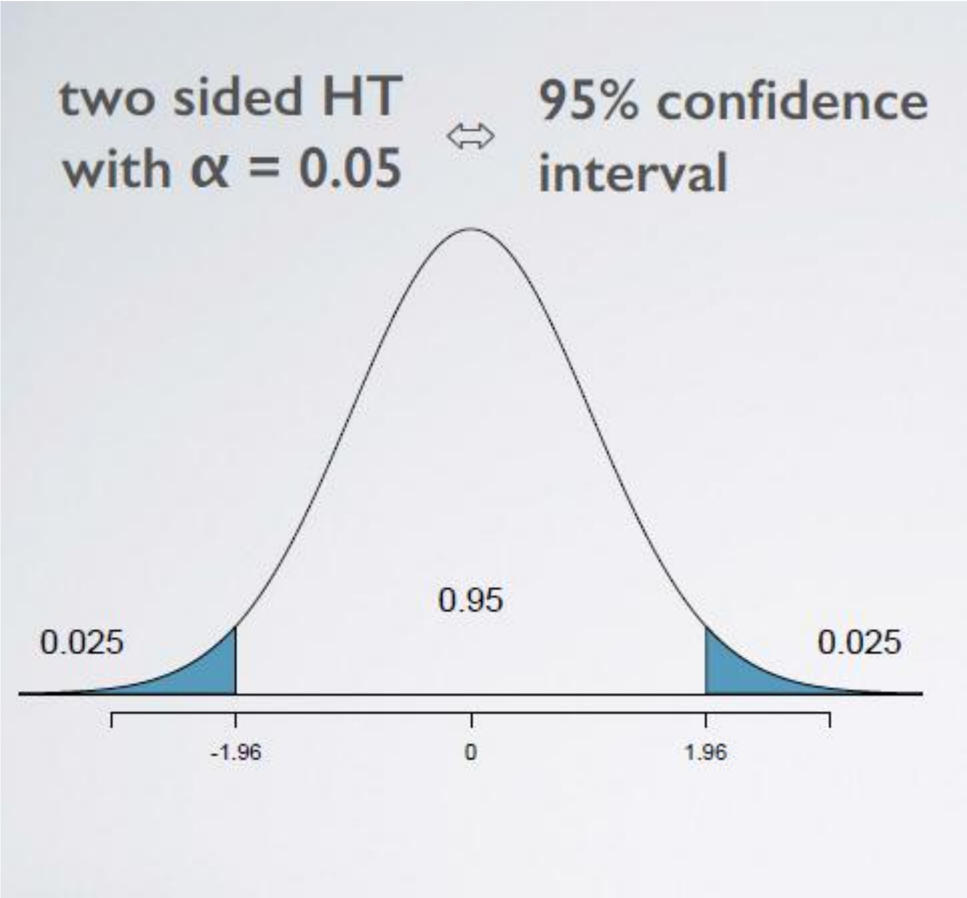
Decision errors

type 2 error rate

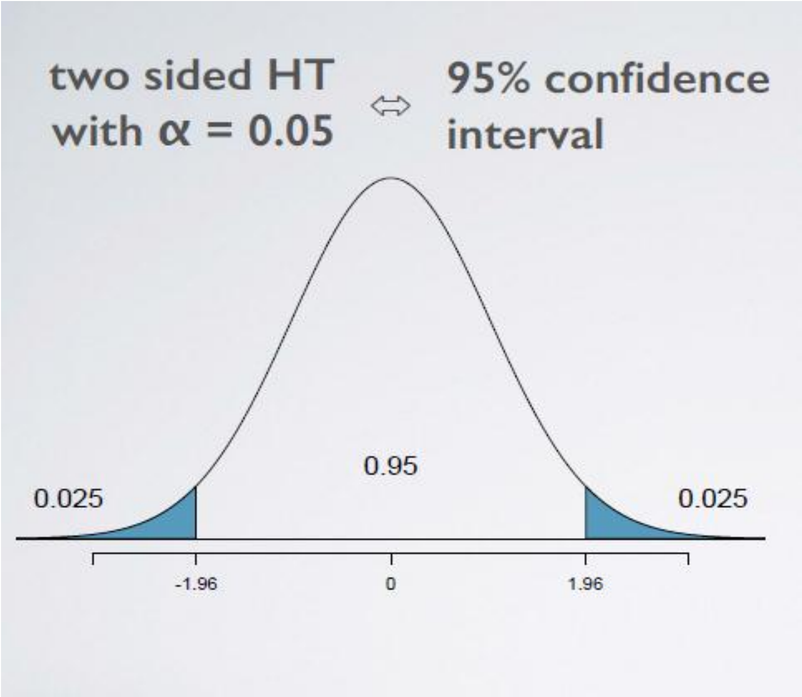
If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

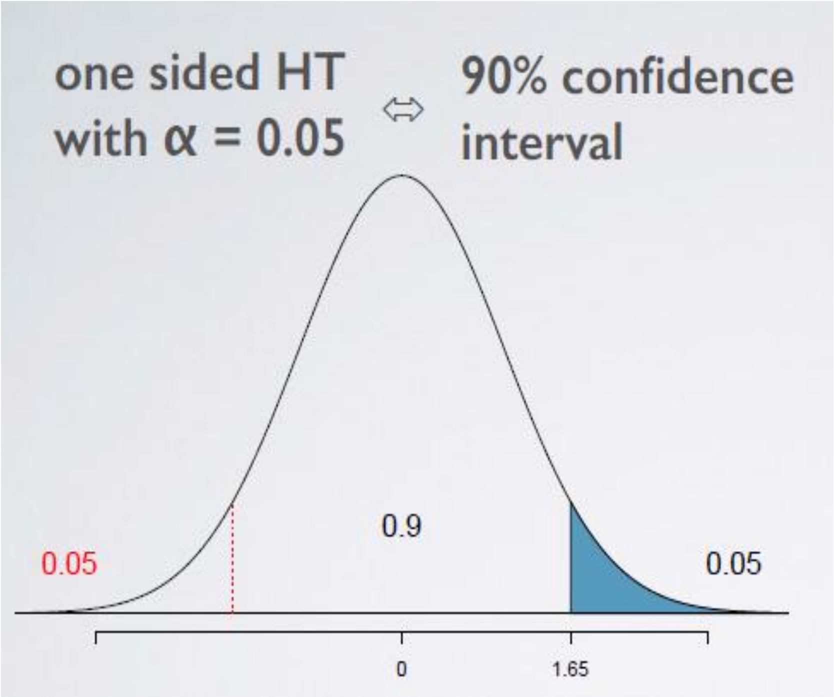
- The answer is not obvious.
- If the true population average is very close to the null value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null value, it will be easier to detect a difference.
- Clearly, β depends on the effect size (δ), difference between point estimate and null value.

Significance vs confidence level



Significance vs confidence level





Significance vs confidence level

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.

Statistical vs. practical significance

All else held equal, will the p-value be lower if n = 100 or n = 10, 000?

(a) n = 100
(b) n = 10,000

$$\bar{x} = 50$$
$$s = 2$$
$$H_0 : \mu = 49.5$$
$$H_A : \mu > 49.5$$

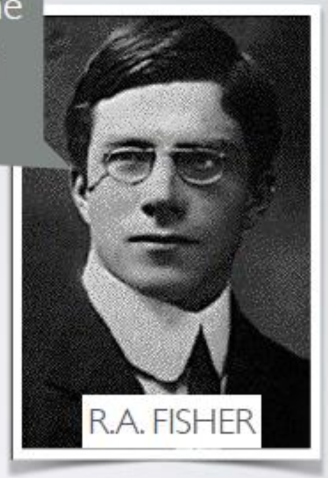
$$Z_{n=100} = \frac{50 - 49.5}{\frac{2}{\sqrt{100}}} = \frac{50 - 49.5}{\frac{2}{10}} = \frac{0.5}{0.2} = 2.5$$

$$Z_{n=10000} = \frac{50 - 49.5}{\frac{2}{\sqrt{10000}}} = \frac{50 - 49.5}{\frac{2}{100}} = \frac{0.5}{0.02} = 25$$

Statistical vs. practical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."



Fisher: http://upload.wikimedia.org/wikipedia/en/2/21/RonaldFisher1912.jpg

Inference for numerical variables

- comparing two means
- boot-strapping
- working with small samples
- comparing many means

Hypothesis testing for paired data

high school and beyond

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test. At a first glance, how are the distributions of reading and writing scores similar? How are they different?

Hypothesis testing for paired data

Given that the same students took the reading and the writing tests, are the reading and writing scores of each student independent of each other?

	ID	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
…	…	…	…
200	137	63	65

Hypothesis testing for paired data

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:

$$\text{diff} = \text{read} - \text{write}$$

- It is important that we always subtract using a consistent order.

	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
...	...	...	...	...
200	137	63	65	-2

Hypothesis testing for paired data

parameter of interest

Average difference between the reading and writing scores of **all** high school students.

μ_{diff}

point estimate

Average difference between the reading and writing scores of **sampled** high school students.

$\bar{x}_{diff}$

Hypothesis testing for paired data

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

	ID	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
...	...	...	...	...
200	137	63	65	-2

$$\bar{x}_{diff} = -0.545$$

$$s_{diff} = 8.887$$

$$n_{diff} = 200$$

Hypothesis testing for paired data

hypotheses for paired means

$H_0 : \mu_{diff} = 0$ There is no difference between the average reading and writing scores.

$H_A : \mu_{diff} \neq 0$ There is a difference between the average reading and writing scores.

Hypothesis testing for paired data

nothing new!

one numerical variable

diff
5
11
19
-5
...
-2

hypothesis about the mean

$$H_0 : \mu_{diff} = 0$$
$$H_A : \mu_{diff} \neq 0$$

Hypothesis testing for paired data

Hypothesis testing for a ~~single mean~~: difference between paired means

1. Set the hypotheses: $H_0 : \cancel{\mu}\ \mu_{diff} = null\ value$

   $H_A : \cancel{\mu}\ \mu_{diff} < or > or \neq null\ value$
2. Calculate the point estimate: $\cancel{\bar{x}}\ \bar{x}_{diff}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, $\cancel{n}$ < 10% of population)
   2. ***Sample size/skew:*** $\cancel{n}\ n_{diff} \geq 30$, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic

$$Z = \frac{x_{diff} - \mu_{diff}}{SE_{\bar{x}_{diff}}}$$

5. Make a decision, and interpret it in context of the research question:

Hypothesis testing for paired data

summary

- paired data (2 vars.) → differences (1 var.)
- most often $H_0 : \mu_{diff} = 0$
- same individuals: pre-post studies, repeated measures, etc.
- different (but dependent) individuals: twins, partners, etc.

Practice

Describe the sampling distribution of the differences between the paired means of reading and writing scores.

$H_0 : \mu_{diff} = 0$

$H_A : \mu_{diff} \neq 0$

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$

$\bar{X}_{diff} \sim N(mean = 0, SE = \frac{8.887}{\sqrt{200}} \approx 0.628)$

Practice

Calculate the test statistic and the p-value for this hypothesis test.

$H_0 : \mu_{diff} = 0$

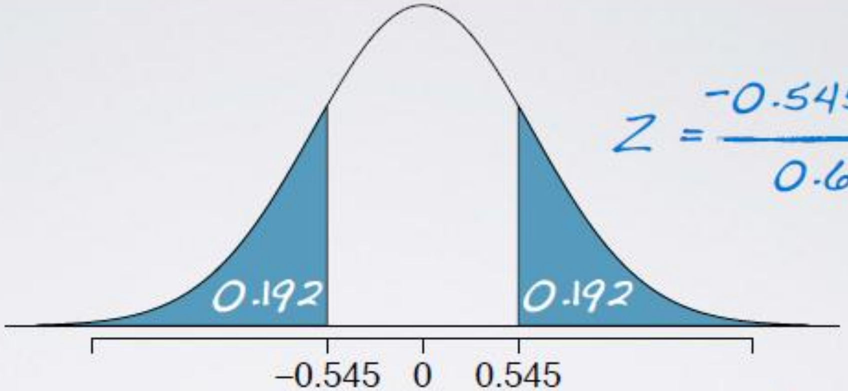
$H_A : \mu_{diff} \neq 0$

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

$n_{diff} = 200$

$\bar{x}_{diff} \sim N(mean = 0, SE = 0.628)$



$$Z = \frac{-0.545 - 0}{0.628} = -0.87$$

$$p\text{-value} = 0.192 \times 2 = 0.384$$

https://bitly.com/dist_calc

Practice

Which of the following is the correct interpretation of the p-value?

~~(a)~~ Probability that the average scores on the reading and writing exams are equal.

$P(H_0 \text{ is true})$

~~(b)~~ Probability that the average scores on the reading and writing exams are different.

$P(H_A \text{ is true})$

✔(c) Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0.

$P(\text{observed or more extreme outcome} \mid H_0 \text{ is true})$

~~(d)~~ Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true. $P(\text{reject} \mid H_0 \text{ is true}) = P(\text{Type 1 error})$

Bootstrapping

Twenty 1+ bedroom apartments were randomly selected on raleigh.craigslist.org. (keyword: `Durham`). Is the mean or the median a better measure of typical rent in Durham?

Can we apply CLT based methods we have learned so far to construct confidence intervals for both?

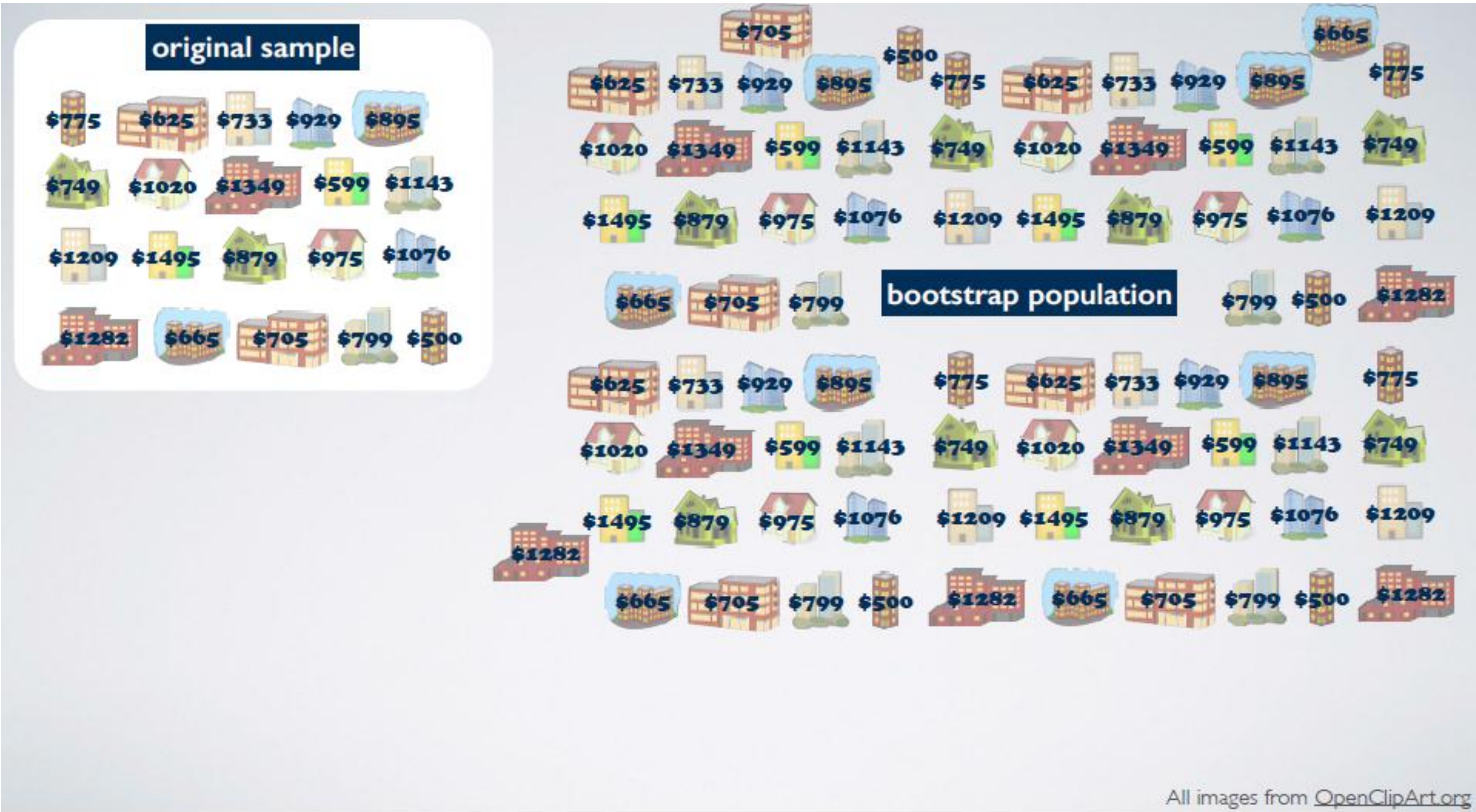
Bootstrapping

- An alternative approach to constructing confidence intervals is bootstrapping.
- This term comes from the phrase *"pulling oneself up by one's bootstraps"*, which is a metaphor for accomplishing an impossible task without any outside help.
- In this case the ~~im~~possible task is estimating a population parameter, and we'll accomplish it using data from only the given sample.

Boots: http://openclipart.org/detail/26401/-by--26401

Bootstrapping

Bootstrapping



Bootstrapping

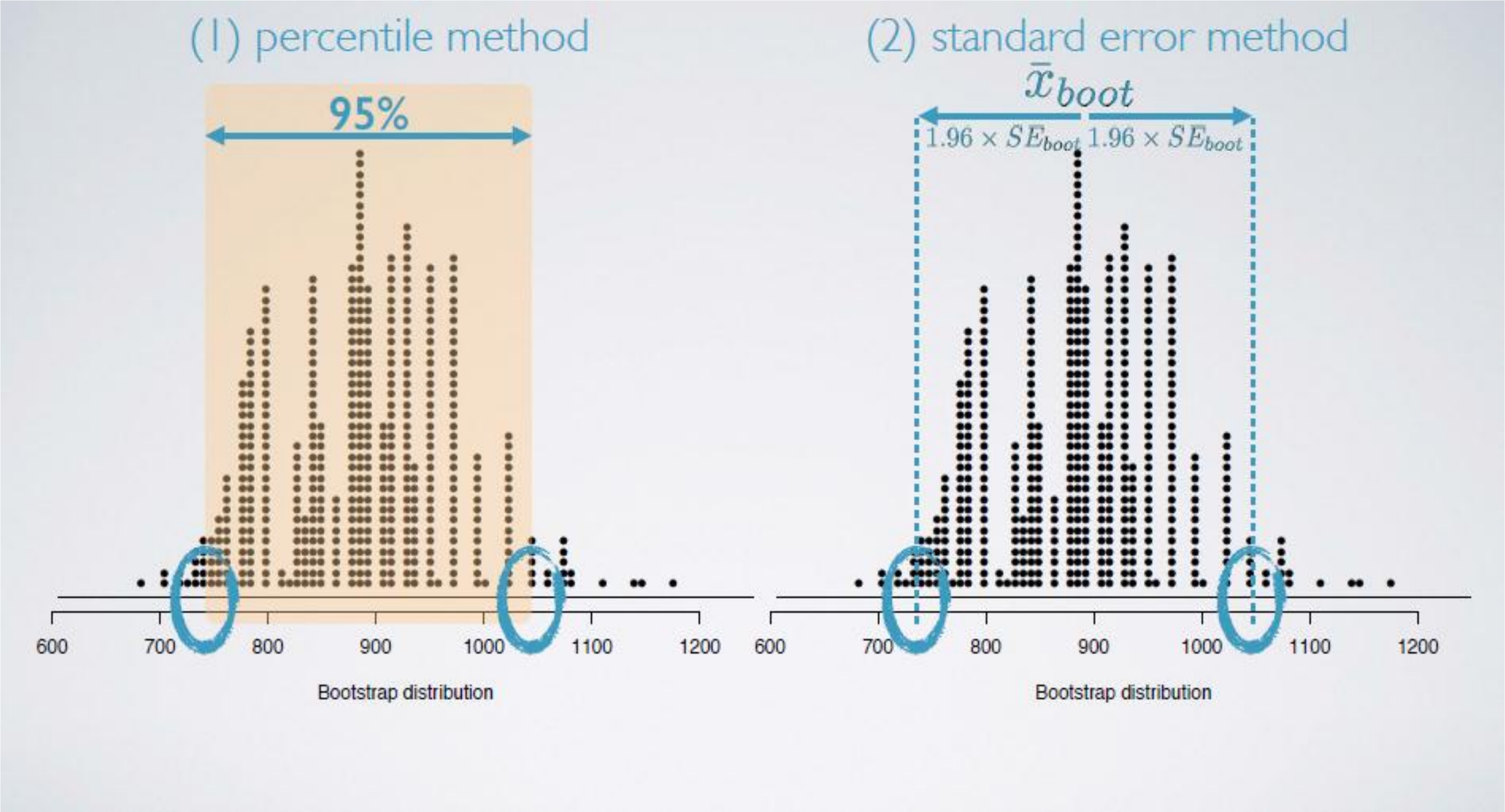
bootstrapping scheme

(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

Bootstrapping



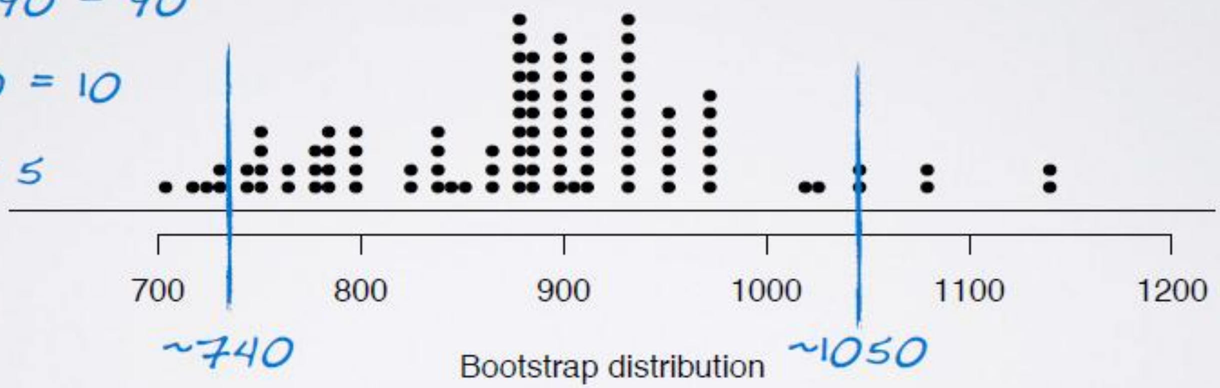
Practice

The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the percentile method.

$100 \times 0.90 = 90$

$100 - 90 = 10$

$10 / 2 = 5$



Practice

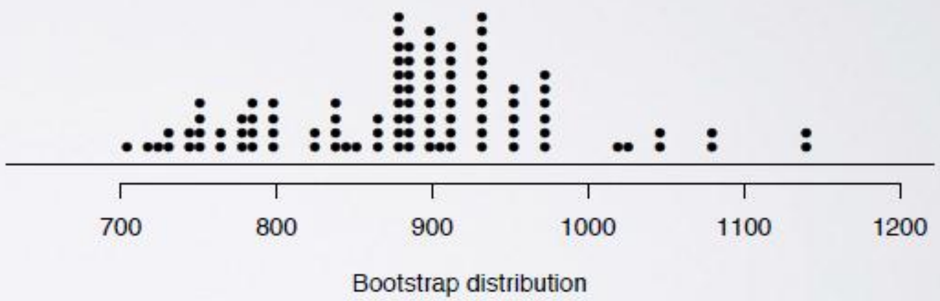
The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the standard error method.

$\bar{x}_{boot} \pm z^* \, SE_{boot} =$

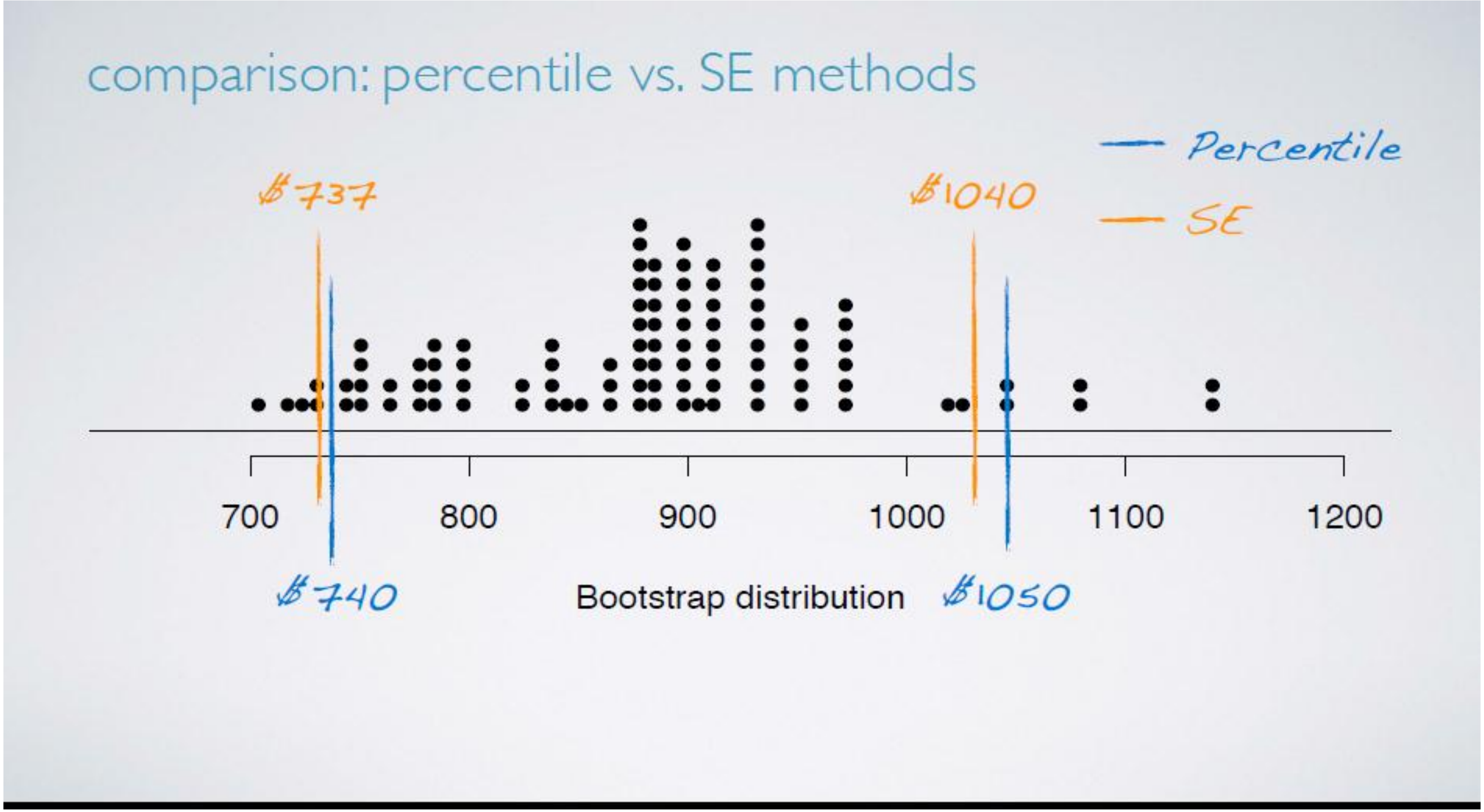
$= 882.515 \pm 1.65 \times 89.5758$

$\approx (734.7, \; 1030.3)$

```
Boot. mean = 882.515
Boot. SE = 89.5758
```



Practice


comparison: percentile vs. SE methods
Percentile
SE
$737
$1040
700
800
900
1000
1100
1200
$740
Bootstrap distribution
$1050

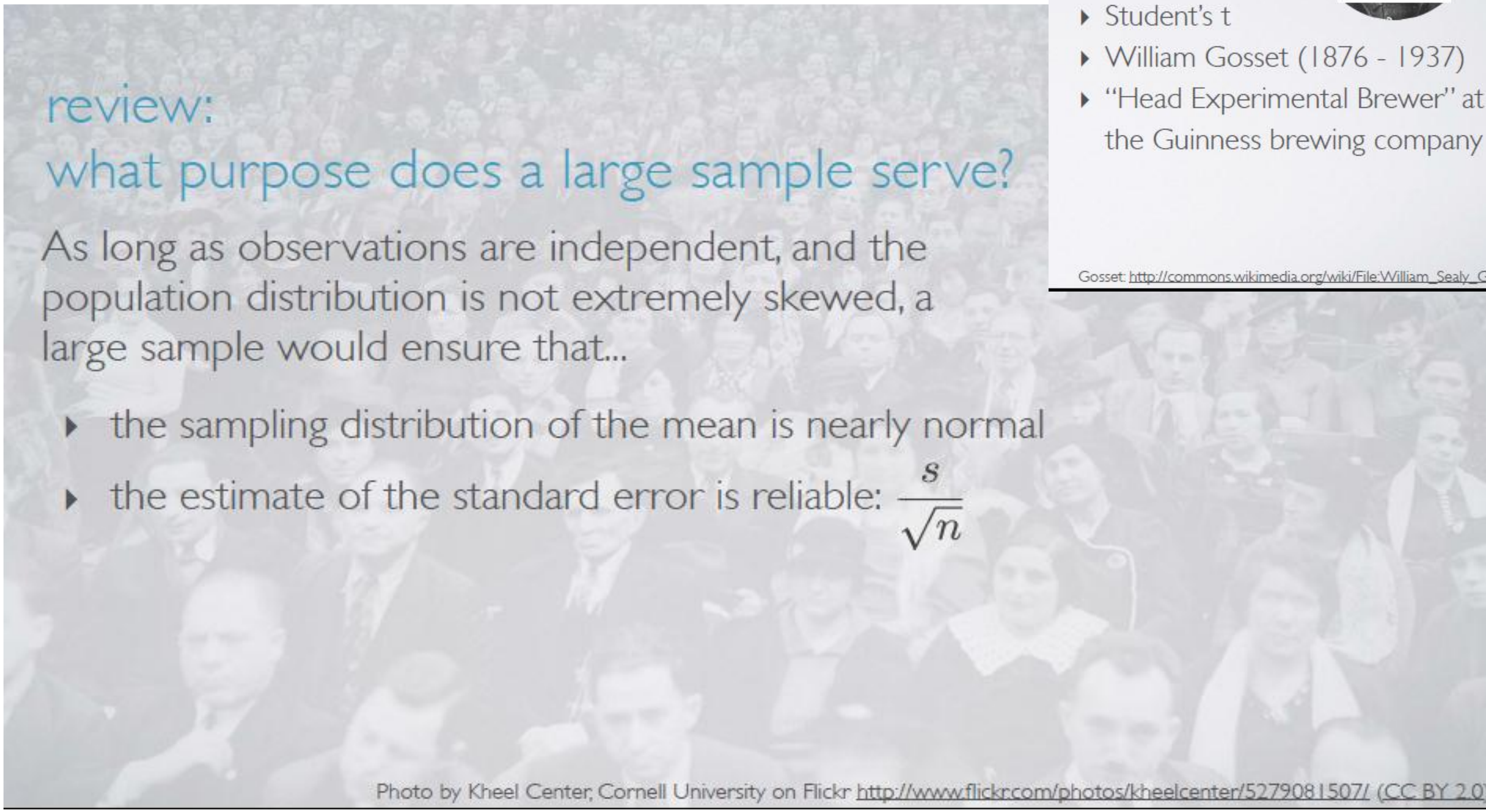
Bootstrapping limitations

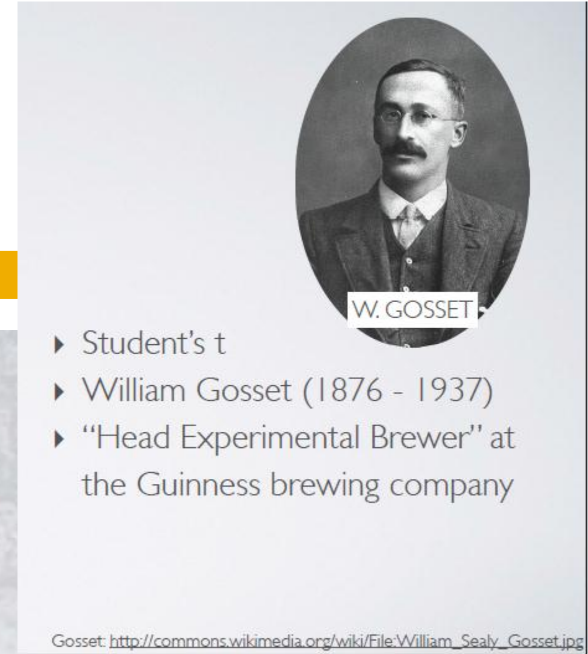
- Not as rigid conditions as CLT based methods.
- However if the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable.
- A representative sample is required for generalizability. If the sample is biased, the estimates resulting from this sample will also be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

Bootstrapping 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.

t distribution

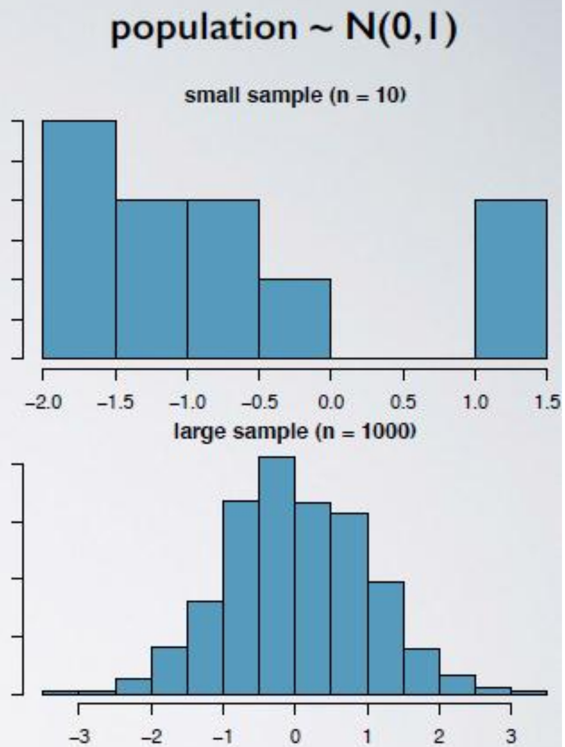




t distribution

review: normality of sampling distributions

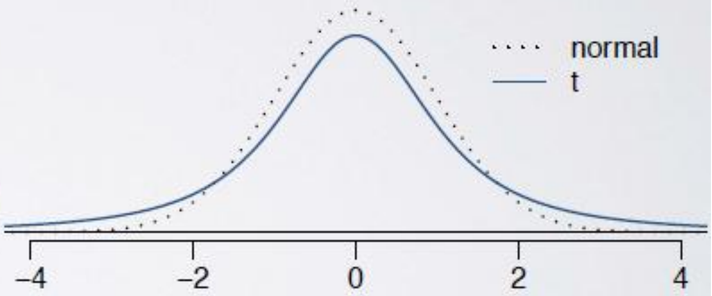
- CLT: sampling distributions are nearly normal as long as the population distribution is nearly normal, for **any** sample size.
- Helpful special case, but difficult to verify normality in small data sets.
- Careful with the normality condition for small samples: don't just examine the sample, also think about where the data come from.
  - *"Would I expect this distribution to be symmetric, and am I confident that outliers are rare?"*



t distribution

t distribution

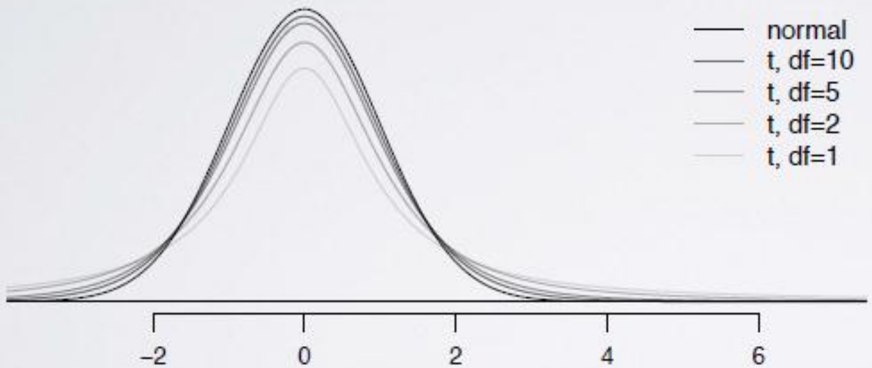
- n is small & σ unknown (almost always), use the t distribution to address the uncertainty of the standard error estimate
- bell shaped but thicker tails than the normal
  - observations more likely to fall beyond 2 SDs from the mean
  - extra thick tails helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution



t distribution

t distribution

- always centered at 0 (like the standard normal)
- has one parameter: degrees of freedom (df) - determines thickness of tails
  - remember, the normal distribution has two parameters: mean and SD



What happens to the shape of the t-distribution as degrees of freedom increases?

approaches the normal dist.

t distribution

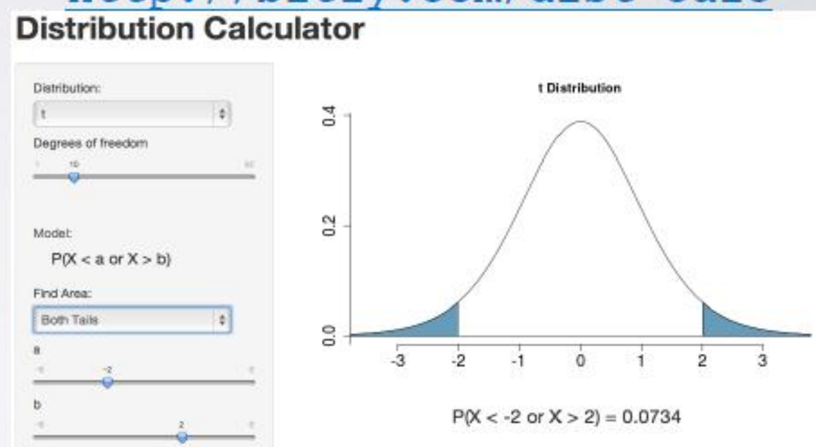
t statistic

- for inference on a mean where
  - σ unknown
  - $n < 30$
- calculated the same way

$$T = \frac{obs - null}{SE}$$

- p-value (same definition)
  - one or two tail area, based on H_A
  - using R, applet, or table

http://bitly.com/dist_calc



Practice

Find the following probabilities.

Say you have a two sided hypothesis test, and your test statistic is 2. Under which of these scenarios would you be able to reject the null hypothesis at the 5% sig. level?

a. $P(|Z| > 2)$ 0.0455 → reject

b. $P(|t_{df=50}| > 2)$ 0.0509 → fail to reject?

c. $P(|t_{df=10}| > 2)$ 0.0734 → fail to reject

Inference for a small sample mean

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:
- randomized into two groups:
(1) play solitaire while eating - "win as many games as possible"
(2) eat lunch without distractions
- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Study reference: Oldham-Cooper, Rose E., et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake." The American journal of clinical nutrition 93.2 (2011): 308-313.

Inference for a small sample mean

estimating the mean (based on a small sample)

point estimate $\pm$ margin of error

$$\bar{x} \pm t^{\star}_{df} SE_{\bar{x}}$$

$$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^{\star}_{n-1} \frac{s}{\sqrt{n}}$$

Degrees of freedom for t statistic for inference on one sample mean $\quad df = n - 1$

Inference for a small sample mean

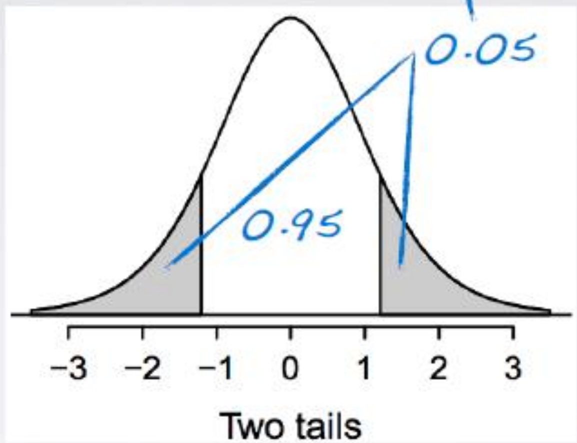
finding the critical t score

using the table

1. determine df

$df = 22 - 1 = 21$

2. find corresponding tail area for desired confidence level



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
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77

Practice

Estimate the average after-lunch snack consumption (in grams) of people who eat lunch **distracted** using a 95% confidence interval.

$\bar{x} = 52.1\ g$

$s = 45.1\ g$

$n = 22$

$t^{\star}_{21} = 2.08$

$\bar{x} \pm t^* SE = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$

$= 52.1 \pm 2.08 \times 9.62$

$= 52.1 \pm 20 \quad = (32.1, 72.1)$

We are 95% confident that distracted eaters consume between 32.1 to 72.1 grams of snacks post-meal.

Practice

Suppose the suggested serving size of these biscuits is 30 g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$\bar{x} = 52.1\ g$

$s = 45.1\ g$

$n = 22$

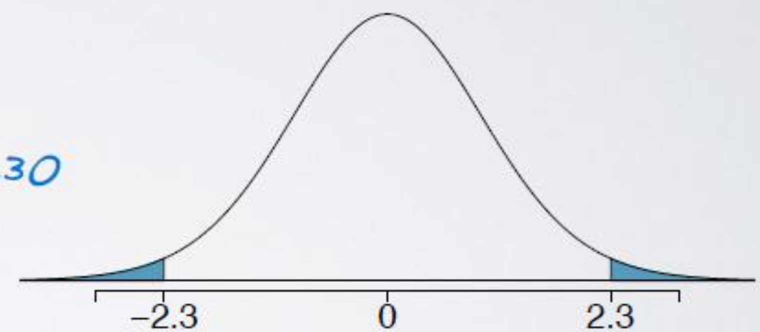
$SE = 9.62$

$H_0: \mu = 30$

$H_A: \mu \neq 30$

$$T = \frac{52.1 - 30}{9.62} = 2.30$$

$df = 22 - 1 = 21$



Practice

finding the p-value

using the table

1. determine df

df = 21

2. locate the calculated T score in the df row

3. grab the one or two tail p-value from the top row

$0.02 < \text{p-value} < 0.05$

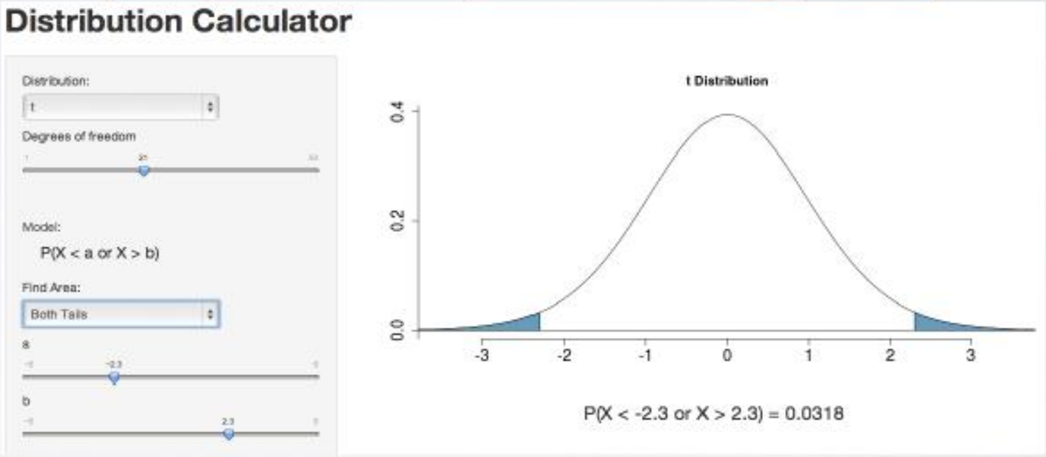
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
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
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77

Practice

finding the p-value

using the applet

http://bitly.com/dist_calc



Practice

recap

$\bar{x} = 52.1\ g$

$s = 45.1\ g$

$n = 22$

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

$H_0 : \mu = 30$

$H_A : \mu \neq 30$

p-value ≈ 0.0318

Reject H_0

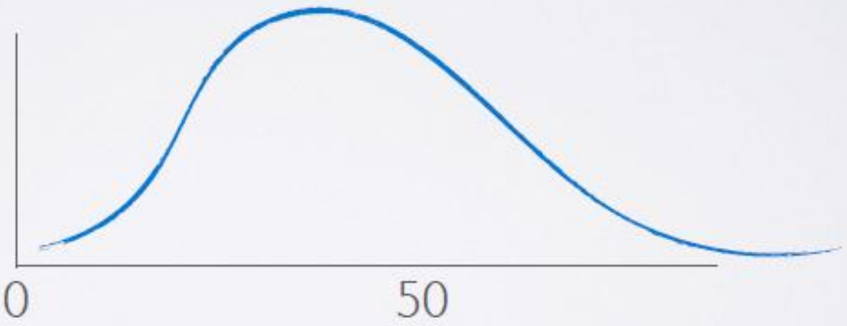
agree

Practice

conditions

- independent observations
  - random assignment
  - 22 < 10% of all distracted eaters
- sample size / skew

$$\bar{x} = 52.1\ g$$
$$s = 45.1\ g$$
$$n = 22$$



Inference for comparing two small sample means

PLAYING A COMPUTER GAME DURING LUNCH AFFECTS FULLNESS, MEMORY FOR LUNCH, AND LATER SNACK INTAKE

distraction and recall of food consumed and snacking

sample: 44 patients: 22 men and 22 women

study design:

- randomized into two groups:

(1) play solitaire while eating - “win as many games as possible”

(2) eat lunch without distractions

- both groups provided same amount of lunch
- offered biscuits to snack on after lunch

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

Study reference: Oldham-Cooper, Rose E., et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake." The American journal of clinical nutrition 93.2 (2011): 308-313.

Inference for comparing two small sample means

comparing means based on small samples

confidence interval

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^{\star}_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$$

hypothesis test

$$T_{df} = \frac{obs - null}{SE}$$

$$T_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}}$$

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

DF for t statistic for inference on difference of two means $df = min(n_1 - 1, n_2 - 1)$

Practice

Do these data provide convincing evidence of a difference between the average post-meal snack consumption between those who eat with and without distractions?

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

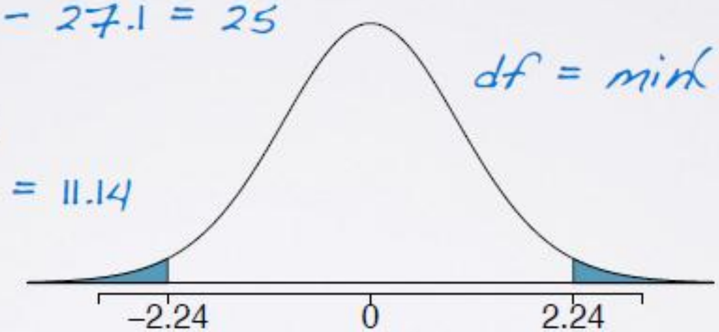
$H_0: \mu_{wd} - \mu_{wod} = 0$ $H_A: \mu_{wd} - \mu_{wod} \neq 0$

$$T = \frac{25 - 0}{11.14} = 2.24$$

$(\bar{X}_{wd} - \bar{X}_{wod}) = 52.1 - 27.1 = 25$

$df = min(22 - 1, 22 - 1) = 21$

$$SE = \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}} = 11.14$$



Practice

Estimate the difference between the average post-meal snack consumption between those who eat with and without distractions?

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

$$\bar{x}_{wd} - \bar{x}_{wod} = 25$$

$$SE = 11.14$$

$$(\bar{X}_{wd} - \bar{X}_{wod}) \pm t^* SE = 25 \pm 2.08 \times 11.14$$

$$= 25 \pm 23.17$$

$$= (1.83, 48.17)$$

Practice

recap

biscuit intake	$\bar{x}$	s	n
solitaire	52.1 g	45.1 g	22
no distraction	27.1 g	26.4 g	22

95% confidence interval: (1.83g, 48.17g)

$$H_0 : \mu_{wd} - \mu_{wod} = 0$$
$$H_A : \mu_{wd} - \mu_{wod} \neq 0$$

p-value ≈ 0.04

Reject H_0

agree

Comparing more than two means

vocabulary score and class

from the 2010 GSS

	wordsum	8
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
…	…	…
795	9	middle class

10 question vocabulary test (scores range from 0 to 10)

self identified social class (lower, working, middle, upper)

Comparing more than two means

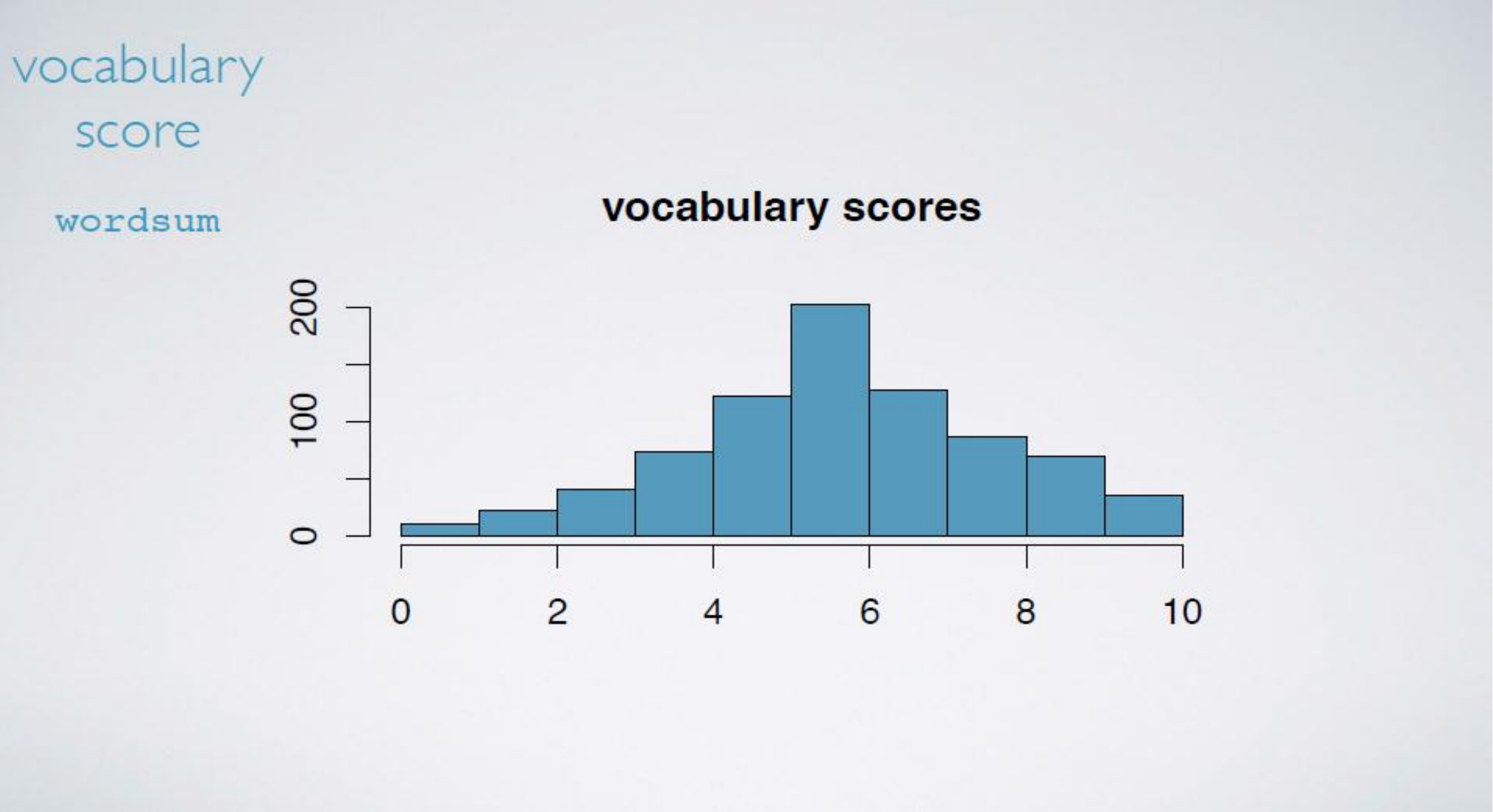
vocabulary score

Choose a word from a list of provided options that comes closest to the meaning of the first word provided in capital letters.

wordsum

1. SPACE (school, noon, captain, room, board, don't know)
2. BROADEN (efface, make level, elapse, embroider, widen, don't know)
3. EMANATE (populate, free, prominent, rival, come, don't know)
4. EDIBLE (auspicious, eligible, fit to eat, sagacious, able to speak, don't know)
5. ANIMOSITY (hatred, animation, disobedience, diversity, friendship, don't know)
6. PACT (puissance, remonstrance, agreement, skillet, pressure, don't know)
7. CLOISTERED (miniature, bunched, arched, malady, secluded, don't know)
8. CAPRICE (value, a star, grimace, whim, inducement, don't know)
9. ACCUSTOM (disappoint, customary, encounter, get used to, business, don't know)
10. ALLUSION (reference, dream, eulogy, illusion, aria, don't know)

Comparing more than two means



Comparing more than two means

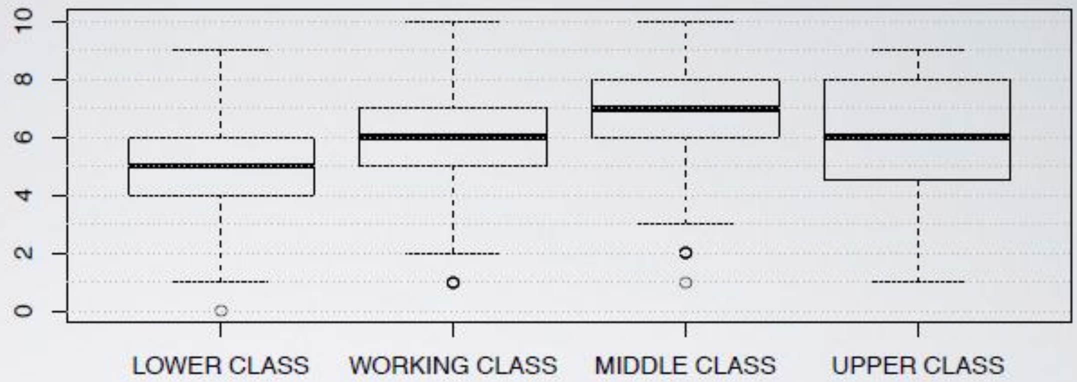
self identified social class

`class`

If you were asked to use one of four names for your social class, which would you say you belong in: the lower class, the working class, the middle class, or the upper class?

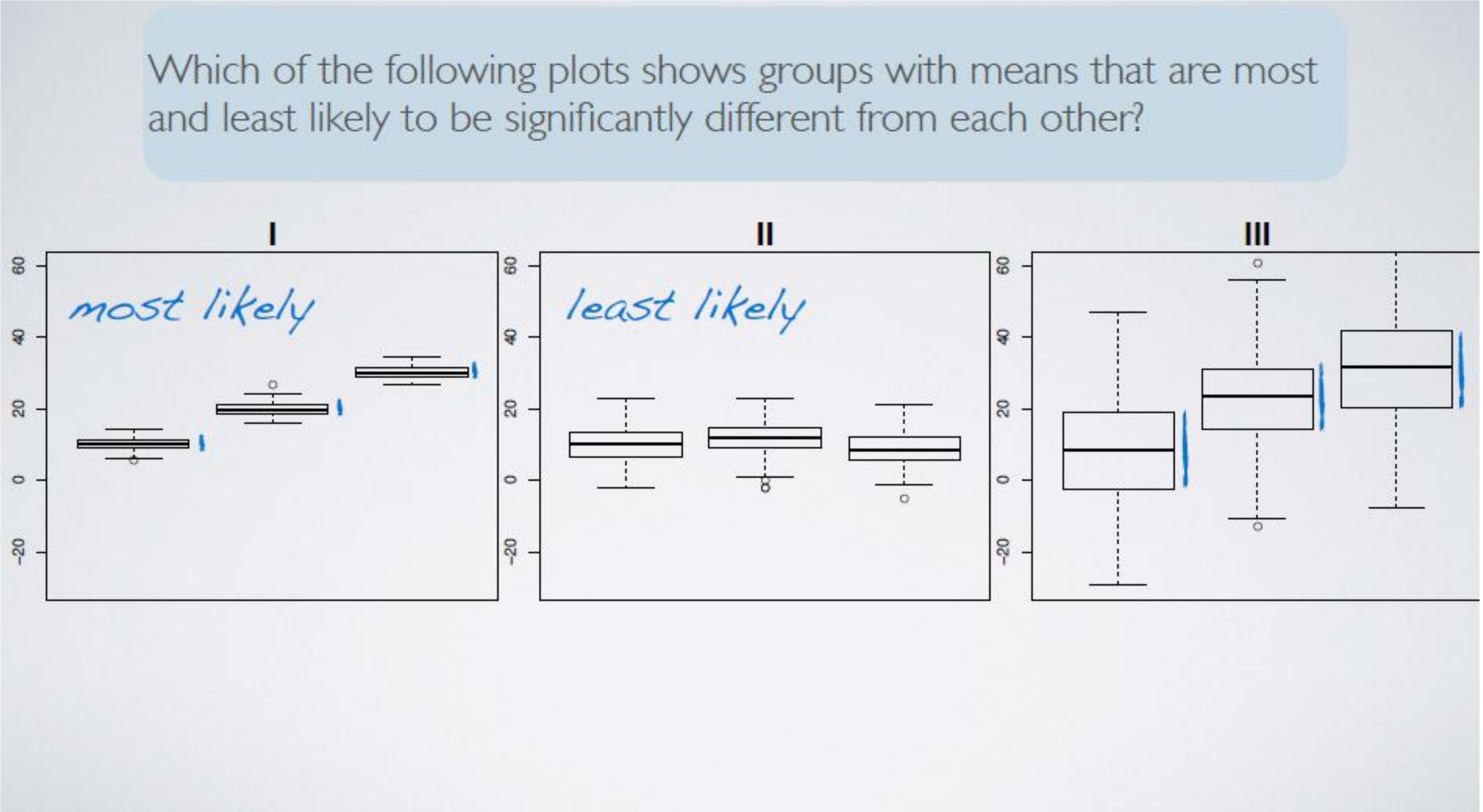
Comparing more than two means

exploratory analysis



	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

Comparing more than two means



Comparing more than two means

Is there a difference between the average vocabulary scores of Americans from different (self reported) classes?

- To compare means of 2 groups we use a Z or a T statistic.
- To compare means of 3+ groups we use a new test called analysis of variance (ANOVA) and a new statistic called F.

Comparing more than two means

anova

H_0: The mean outcome is the same across all categories

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

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

μ_i : mean of the outcome for observations in category *i*

k : number of groups

Comparing more than two means

z / t test

Compare means from **two** groups: are so far apart that the observed difference cannot reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2$$

anova

Compare means from **more than two** groups: are they so far apart that the observed differences cannot all reasonably be attributed to sampling variability?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

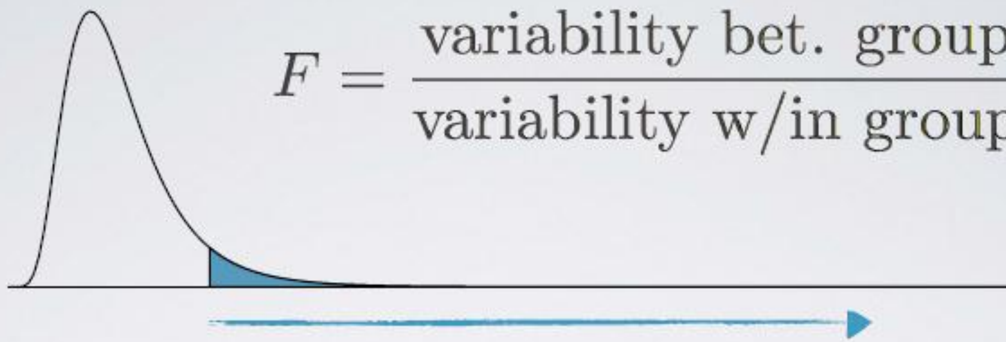
Comparing more than two means

z / t test	anova
Compute a test statistic (a ratio).	Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{(\bar{x}_1 - \bar{x}_2)}} \qquad F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, and we conclude that the data provide evidence of a difference in the population means.

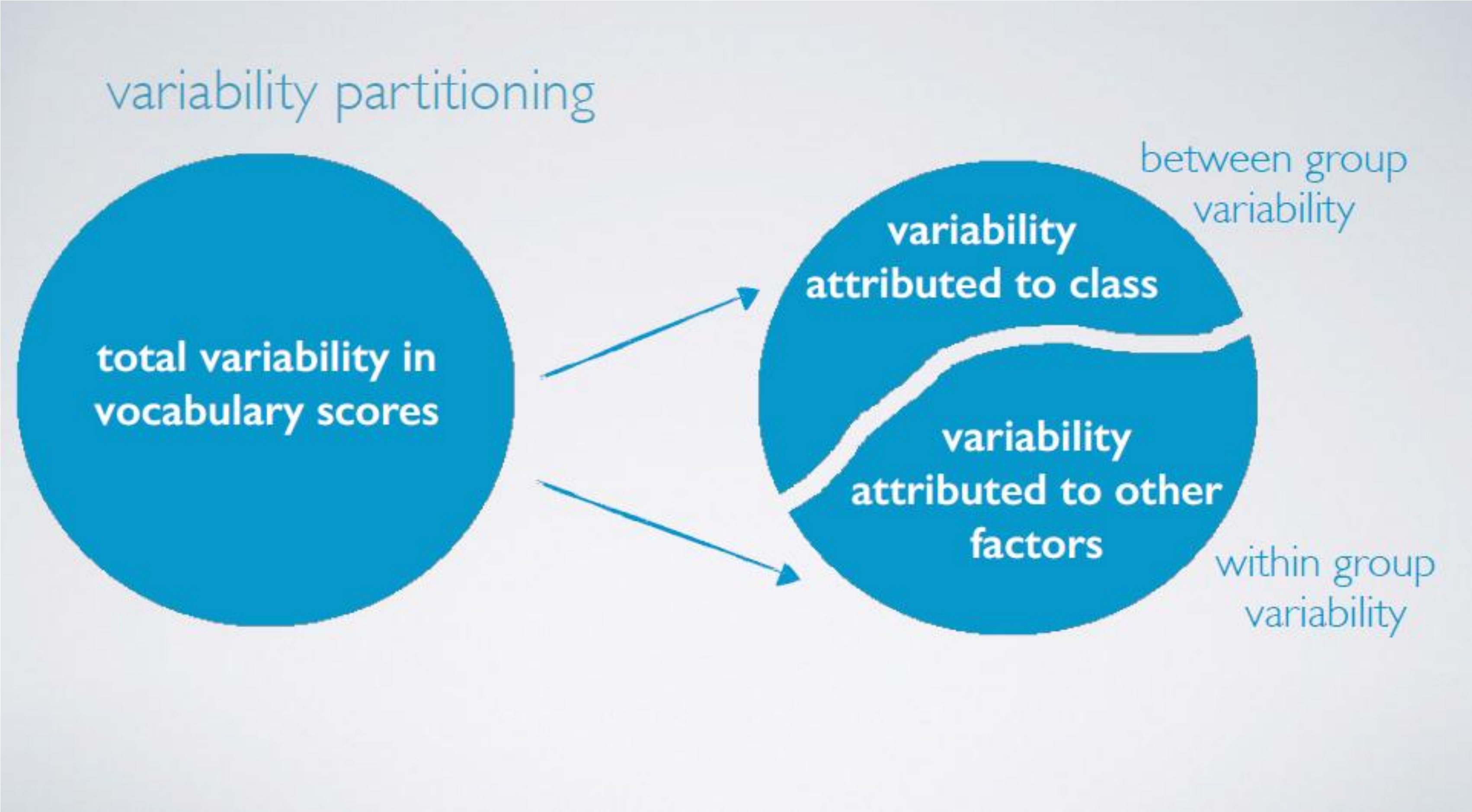
Comparing more than two means



$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}} \quad \begin{matrix} ? \\ ? \end{matrix}$$

- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

ANOVA


variability partitioning
total variability in vocabulary scores
variability attributed to class
variability attributed to other factors
between group variability
within group variability

ANOVA

vocabulary score and class

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
4	5	working class
5	6	working class
6	6	working class
...	...	...
795	9	middle class

	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

H_0: The mean outcome is the same across all categories

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

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

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			

ANOVA

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

sum of squares total (SST)

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

ANOVA

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

	wordsum	class
1	6	middle class
2	9	working class
3	6	working class
...	...	...
795	9	middle class

	n	mean	sd
overall	795	6.14	1.98

$$SST = (6-6.14)^2 + (9-6.14)^2 + (6-6.14)^2 + \cdots + (9-6.14)^2 = 3106.36$$

ANOVA

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

sum of squares groups (SSG)

- measures the variability **between groups**
- **explained variability:** deviation of group mean from overall mean, weighted by sample size

ANOVA

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

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

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56			
Error	Residuals		2869.8			
	Total		3106.36			

Sum of squares error (SSE):

$$SSE = SST - SSG$$

3106.36 − 236.56 = 2869.8

sum of squares error (SSE)

- measures the variability **within groups**
- **unexplained variability:** unexplained by the group variable, due to other reasons

ANOVA

		Df	Sum Sq	Mean Sq	F value	Pr(> F)
Group	class		236.56	?		
Error	Residuals		2869.8	?		
	Total		3106.36	?		

- now we need a way to get from these measures of total variability to average variability
- scaling by a measure that incorporates sample sizes and number of groups → degrees of freedom

degrees of freedom

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

Degrees of freedom associated with ANOVA:

- total: $df_T = n - 1$ → $795 - 1 = 794$
- group: $df_G = k - 1$ → $4 - 1 = 3$
- error: $df_E = df_T - df_G$ → $794 - 3 = 791$

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: $MSG = SSG/df_G$ → $236.56 \; / \; 3 \approx 78.855$
- error: $MSE = SSE/df_E$ → $2869.8 \; / \; 791 \approx 3.628$

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			

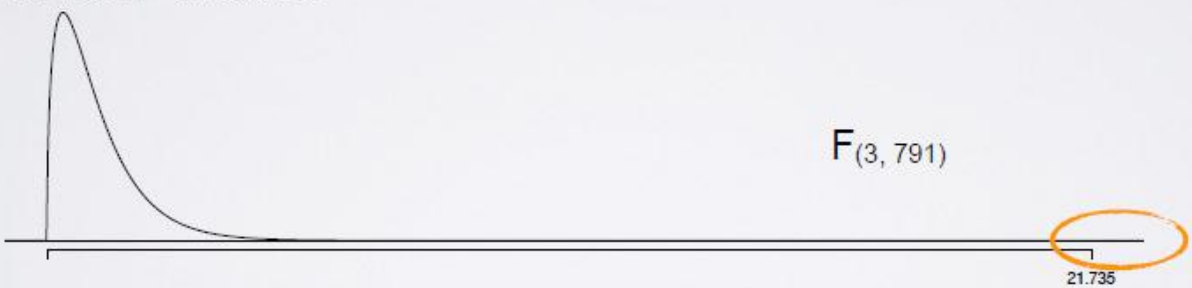
F statistic: Ratio of the between group and within group variability:

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

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 the "between" and "within" group variabilities if in fact the means of all groups are equal
- area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



conclusion

		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			

- 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 one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

Conditions for ANOVA

Conditions for ANOVA

1. ***Independence:***
   - ✓ **within groups:** sampled observations must be independent
   - ✓ **between groups:** the groups must be independent of each other (non-paired)
2. ***Approximate normality:*** distributions should be nearly normal within each group
3. ***Equal variance:*** groups should have roughly equal variability

Conditions for ANOVA

(1) independence

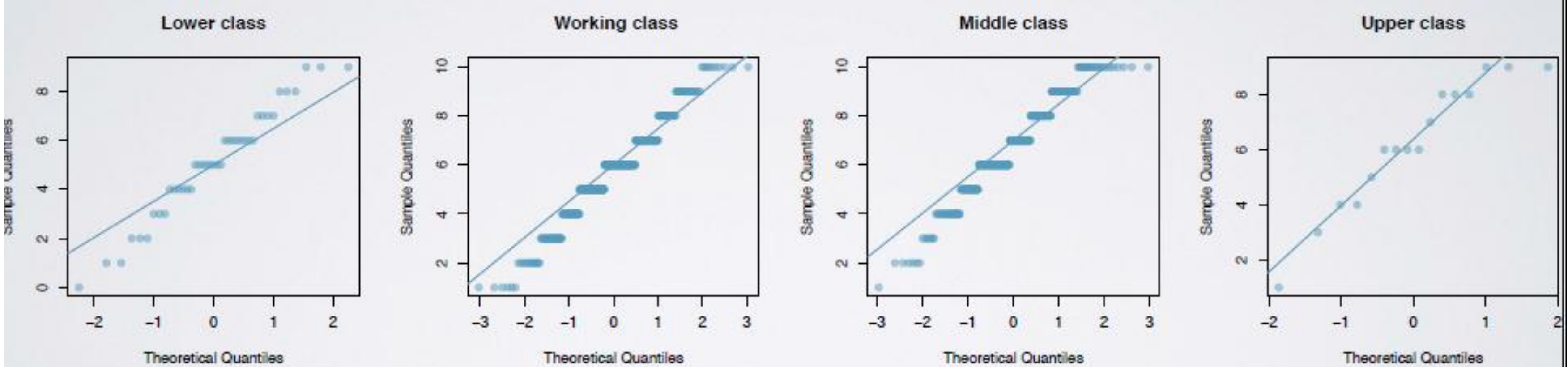
sampled observations must be independent of each other

- random sample / assignment
- each n_j less than 10% of respective population
- carefully consider whether the groups may be independent (e.g. no pairing) → repeated measures anova
- always important, but sometimes difficult to check

Conditions for ANOVA

(2) approximately normal

- distribution of response variable within each group should be approximately normal
- especially important when sample sizes are small

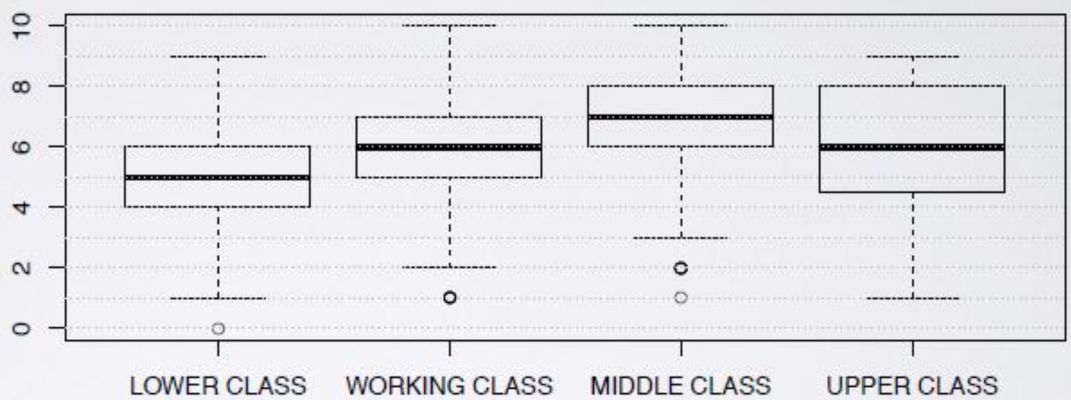


Conditions for ANOVA

(3) constant variance

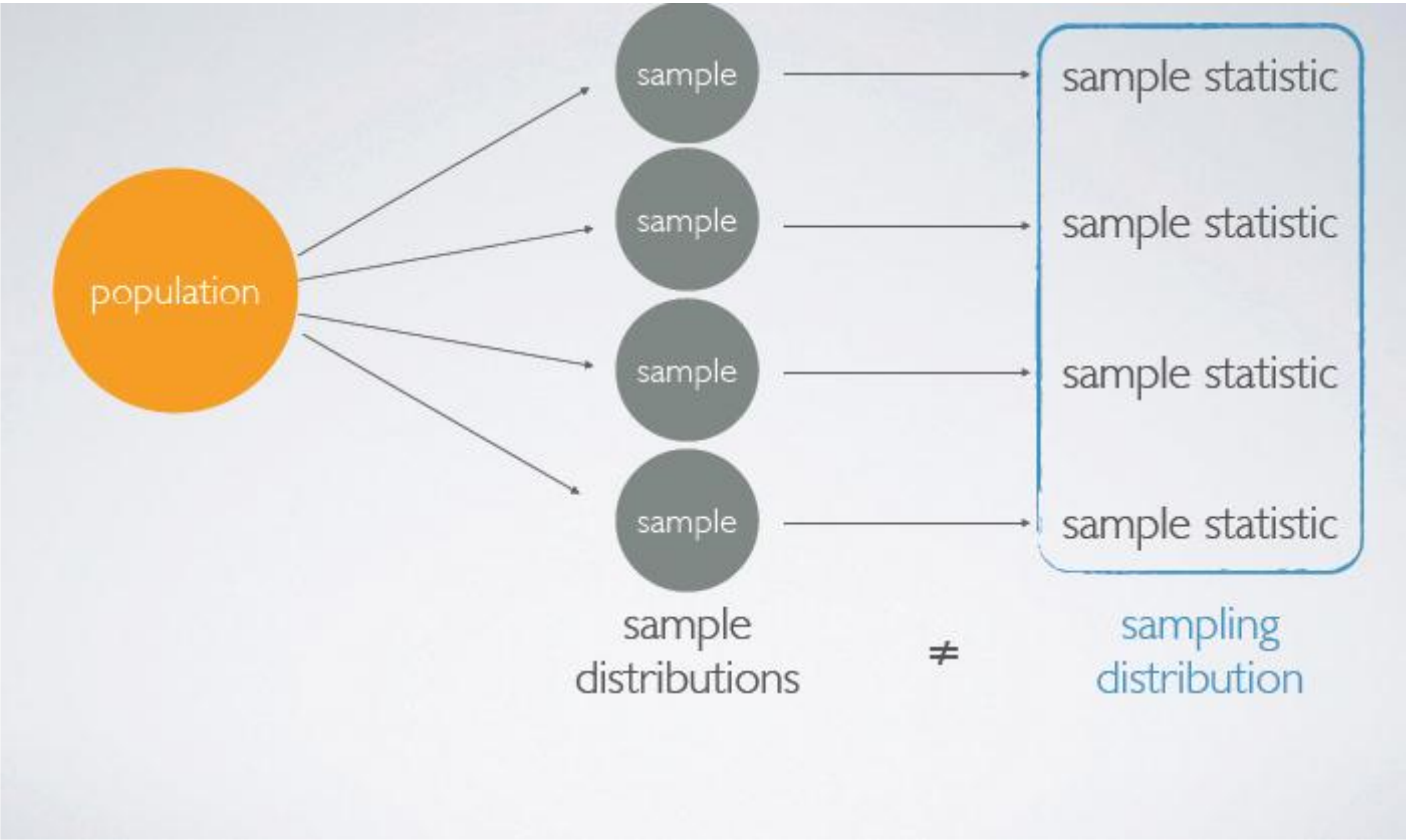
- variability should be consistent across groups: homoscedastic groups
- especially important when sample sizes differ between groups

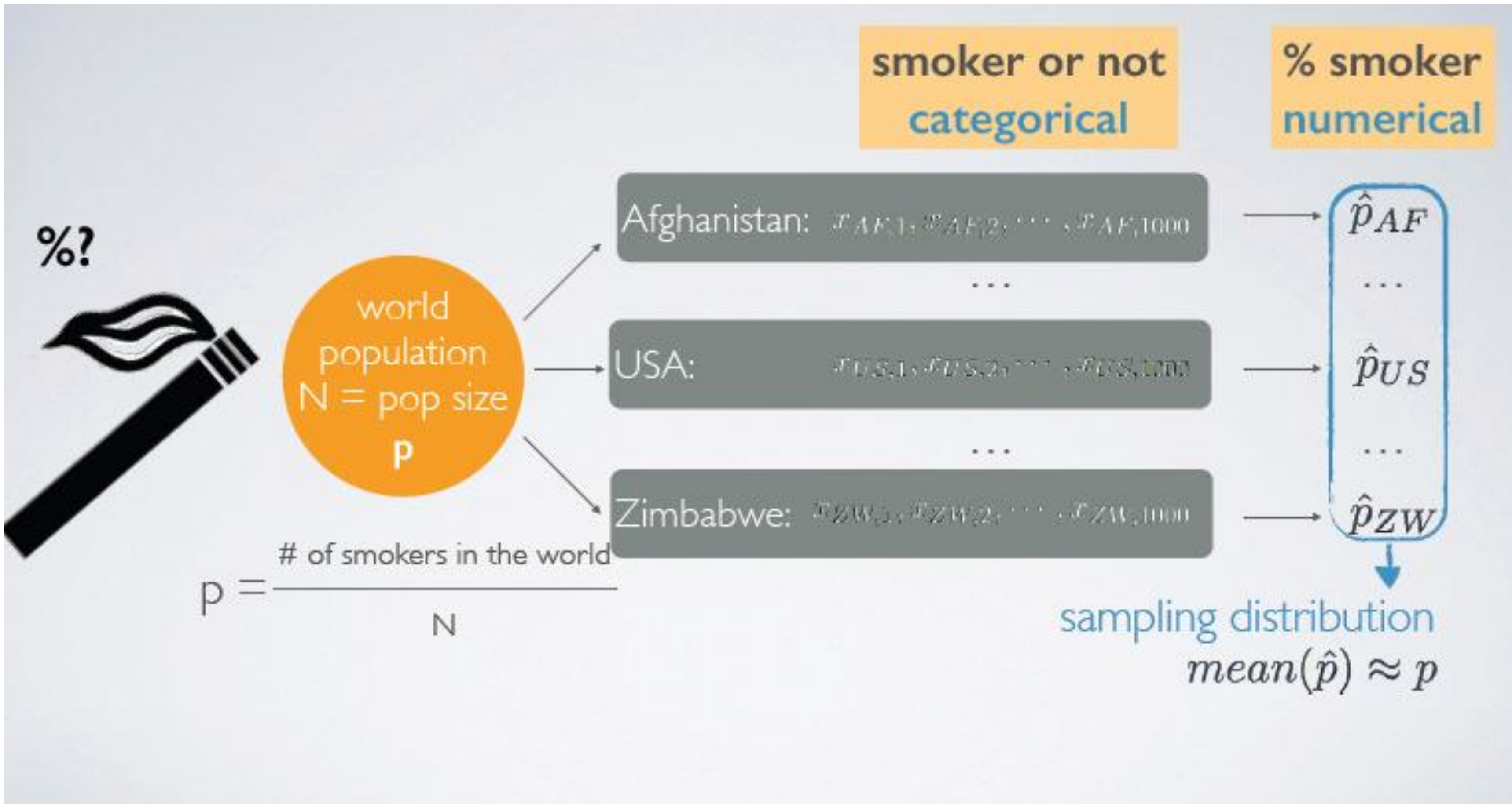
	n	sd
lower class	41	2.24
working class	407	1.87
middle class	331	1.89
upper class	16	2.34
overall	795	1.98



Inference for categorical variables

Sampling variability & CLT for proportions




smoker or not
categorical
% smoker
numerical
%?
world
population
N = pop size
p
Afghanistan: $x_{AF,1}, x_{AF,2}, \cdots, x_{AF,1000}$
USA:
Zimbabwe:
$\hat{p}_{AF}$
$\hat{p}_{US}$
$\hat{p}_{ZW}$
$p = \frac{\text{\# of smokers in the world}}{N}$
sampling distribution
$mean(\hat{p}) \approx p$

CLT for proportions: The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** There should be at least 10 successes and 10 failures in the sample: $np \geq 10$ and $n(1-p) \geq 10$.

if p unknown, use $\hat{p}$

Practice

90% of all plants species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants.

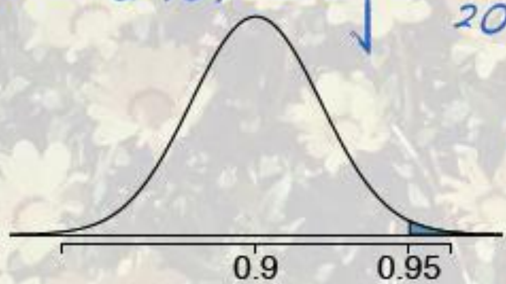
$p = 0.90$

$n = 200$

$P(\hat{p} > 0.95) = ?$

1. random sample & <10% of all plants → independent obs.
2. $200 \times 0.90 = 180$ and $200 \times 0.10 = 20$

$$\hat{p} \sim N\left(\text{mean} = 0.90,\ SE = \sqrt{\frac{0.90 \times 0.10}{200}} \approx 0.0212\right)$$



$$Z = \frac{0.95 - 0.90}{0.0212} = 2.36$$

$P(Z > 2.36) \approx 0.0091$

Practice

90% of all plants species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants.

$p = 0.90$

$n = 200$

$P(\hat{p} > 0.95) = ?$

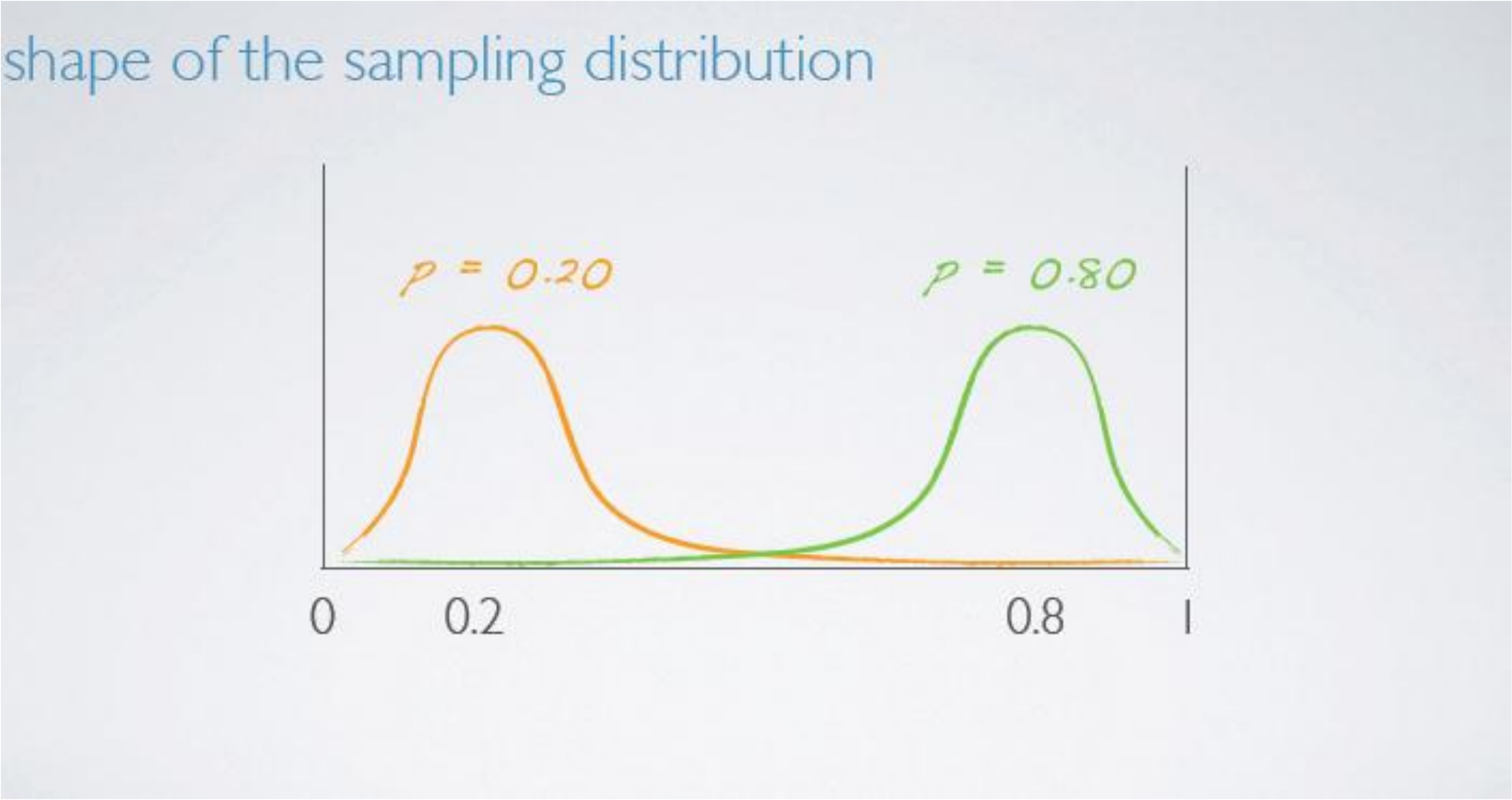
Using the binomial distribution:

$200 \times 0.95 = 190$

What if

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
p = 0.20
p = 0.80
0
0.2
0.8
1

Hypothesis testing for a proportion

Hypothesis testing for a single proportion:

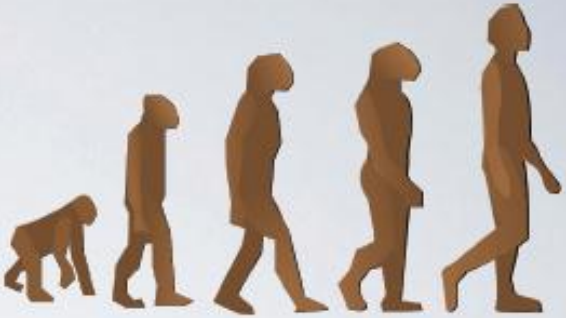
1. Set the hypotheses: $H_0 : p = null\ value$
   $H_A : p < \ or\ > \ or\ \neq \ null\ value$
2. Calculate the point estimate: $\hat{p}$
3. Check conditions:
   1. ***Independence:*** Sampled observations must be independent (random sample/assignment & if sampling without replacement, n < 10% of population)
   2. ***Sample size/skew:*** *np ≥ 10* and *n(1-p) ≥ 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

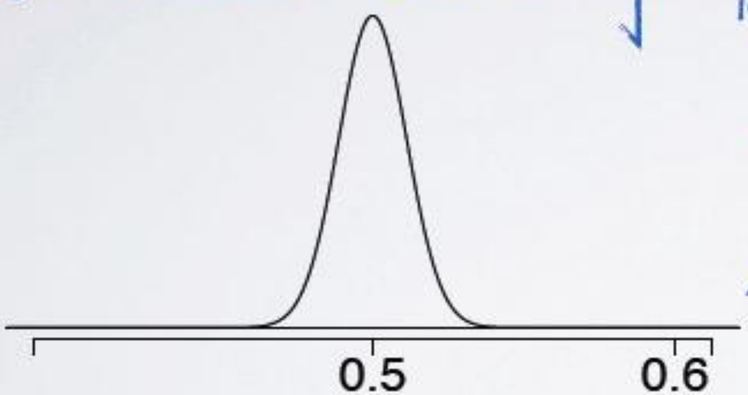
Image source: http://openclipart.org/detail/12755/evolution-steps-by-anonymous-12755

Practice

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

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

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



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

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

$$= \text{almost } 0 \rightarrow \text{reject } H_0$$

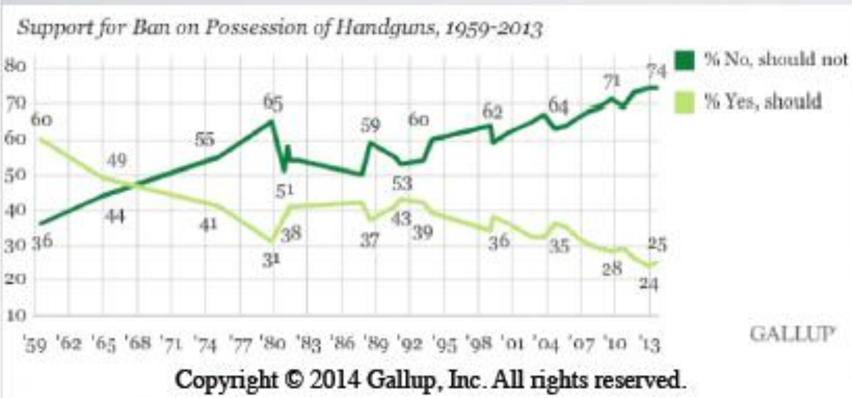
Estimating the difference between two proportions

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

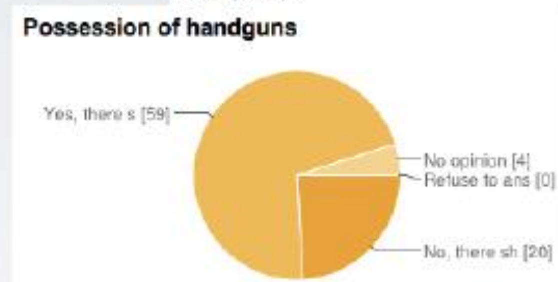
Link to poll: http://www.gallup.com/poll/165563/remains-divided-passing-stricter-gun-laws.aspx
Image source: http://openclipart.org/detail/14213/not-for-sale-by-cibo00-14213

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 diference between two proportions

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:

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

Estimating diference between two proportions

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:
   - ✓ $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
   - ✓ $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Practice

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.

Practice

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

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

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

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

$$= 0.46 \pm 0.10$$

$$= (0.36, 0.56)$$

Practice

does the order matter?

remember $$\underbrace{(\hat{p}_1 - \hat{p}_2)}_{\text{can be - or +}} \pm z^\star \underbrace{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}_{\text{always +}}$$

$$\begin{aligned}(p_{Coursera} - p_{US}) &= \\ &= (0.71 - 0.25) \pm 0.10 \\ &= 0.46 \pm 0.10 \\ &= (0.36, 0.56)\end{aligned}$$

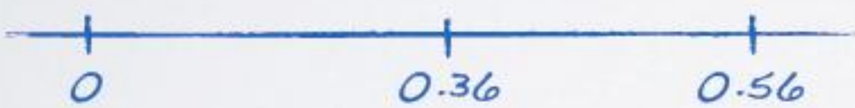
$$\begin{aligned}(p_{US} - p_{Coursera}) &= \\ &= (0.25 - 0.71) \pm 0.10 \\ &= -0.46 \pm 0.10 \\ &= (-0.56, -0.36)\end{aligned}$$

Practice

Based on the confidence interval we calculated, should we expect to find a significant difference (at the equivalent significance level) between the population proportions of Coursera students and the American public at large who believe there should be a law banning the possession of handguns?

$$(p_{Coursera} - p_{US}) = (0.36, 0.56)$$

$H_0: p_{Coursera} - p_{US} = 0$



reject H_0

Hypothesis tests for comparing two proportions

A SurveyUSA poll asked respondents whether any of their children have ever been the victim of bullying. Also recorded on this survey was the gender of the respondent (the parent). Below is the distribution of responses by gender of the respondent.

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

34 / 90   61 / 122

$H_0: p_{male} - p_{female} = 0$

$H_A: p_{male} - p_{female} \neq 0$

✔ check conditions

✔ calculate test statistic & p-value

Link to poll: http://www.surveyusa.com/client/PollReport.aspx?g=1823ef50-44c7-4d2a-9efc-ead711b4ad9c flashback to working with one proportion: $\hat{p}$ vs. p

	observed confidence interval	expected 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	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

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

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

$$H_0 : p_1 = p_2$$

Pooled proportion

pooled proportion

$$H_0 : p_1 = p_2 = ?$$

Pooled proportion:

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

Practice

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

Practice

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

	observed confidence interval	*expected* hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_2(1-\hat{p}_2) \geq 10$	$n_1\hat{p}_{pool} \geq 10$ $n_1(1-\hat{p}_{pool}) \geq 10$ $n_2\hat{p}_{pool} \geq 10$ $n_2(1-\hat{p}_{pool}) \geq 10$
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

Practice

Are conditions for inference met for conducting a hypothesis test to compare the two proportions?

	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

1. independence:
   ✓ within groups: random sample & 10% condition

Sampled males independent of each other, sampled females are as well.

✓ between groups:

No reason to expect sampled males and females to be dependent.

2. sample size / skew: ✓ Males: $90 \times 0.45 = 40.5$ and $90 \times 0.55 = 49.5$
   ✓ Females: $122 \times 0.45 = 54.9$ and $122 \times 0.55 = 67.1$

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

Conduct a hypothesis test, at 5% significance level, evaluating if males and females are equally likely to answer "Yes" to the question about whether any of their children have ever been the victim of bullying.

	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

$H_0: p_{male} - p_{female} = 0$ $\quad H_A: p_{male} - p_{female} \neq 0$

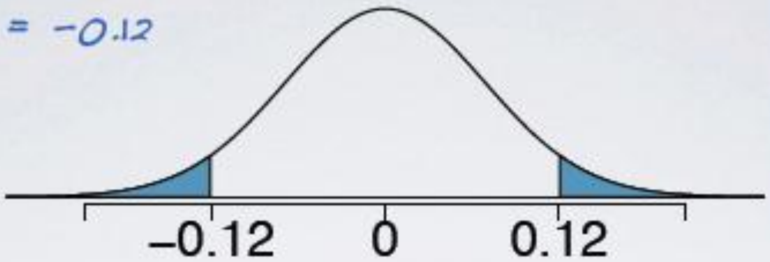
$$(\hat{p}_{male} - \hat{p}_{female}) \sim N\left(mean = 0,\ SE = \sqrt{\frac{0.45 \times 0.55}{90} + \frac{0.45 \times 0.55}{122}} \approx 0.0691\right)$$

$$\text{point estimate} = \hat{p}_{male} - \hat{p}_{female} = 0.38 - 0.50 = -0.12$$

point estimate = -0.12

null value = 0

$SE = 0.0691$



	Male	Female
Total	90	122
$\hat{p}$	0.38	0.50
$\hat{p}_{pool}$	0.45	

$$Z = \frac{-0.12 - 0}{0.0691} \approx -1.74$$

$$\text{p-value} = P(|Z| > 1.74) \approx 0.08$$