Bootstrap: getting a lot of data is hard, mimics drawing samples from the population which can be costly / take a long time

Taking samples from the original with replacement

Each bootstrap is centered on the sample mean, not the true mean

Bootstrap distribution: distribution of the statistic of interest from the bootstrap data, sampling distribution & bootstrap means are different but spreads are same

Confidence Intervals: how confident you are that the statistic of interest falls in a certain range, original sample must be random, larger samples are better

- common confidence intervals: 90, 95, 99

Need a large number of resamples to get a good distribution

- 15,000 times

- if original sample uses good practices (representative of population etc.) bootstrap can be generalized to the population

The bootstrap distribution allways mirrors the sampling distribution

Probability - always between [0, 1]

If mutually exclusive, $P(A \text{ and } B) = 0$

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

Independence = if knowing the outcome of one provides no useful info abt the outcome of another

Complement: $P(A) = 1 - P(A^c)$ A^c = anything other than A

Conditional: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$ Conditional compliment: $1 = P(A^c|B) + P(A|B)$

Multiplication: $P(A \text{ and } B) = P(A|B)P(A)$

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

Total Probability: $P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \ldots + P(B|A_k) \cdot P(A_k)$

$\hat{P} = \frac{X}{n}$ — X — number of times we observed the event; n — the number of trials

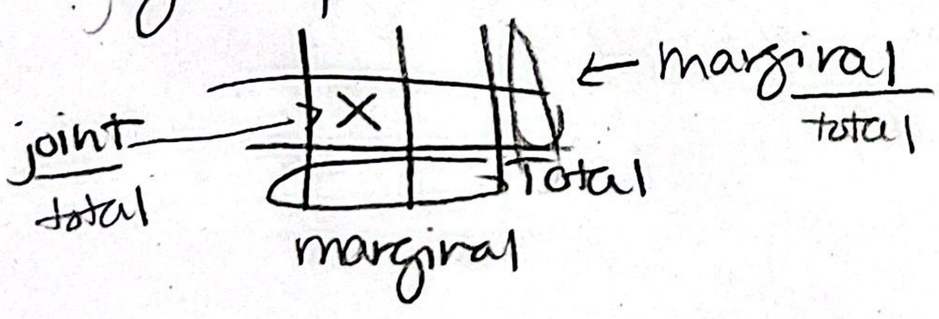
$\hat{P}$ = empirical P = Theoretical

Theoretical Probabilities — A system of ideas intended to explain something

Empirical Probabilities — ~~Based on/concerned with~~ Estimated based on observed data

joint probability: $P(A \text{ and } B)$

Contingency table ex:



α

Confidence Intervals

We can have confidence intervals of different sizes, common ones include 90%, 96%, 99%

z scores are → $z = \frac{x - \mu}{\sigma}$

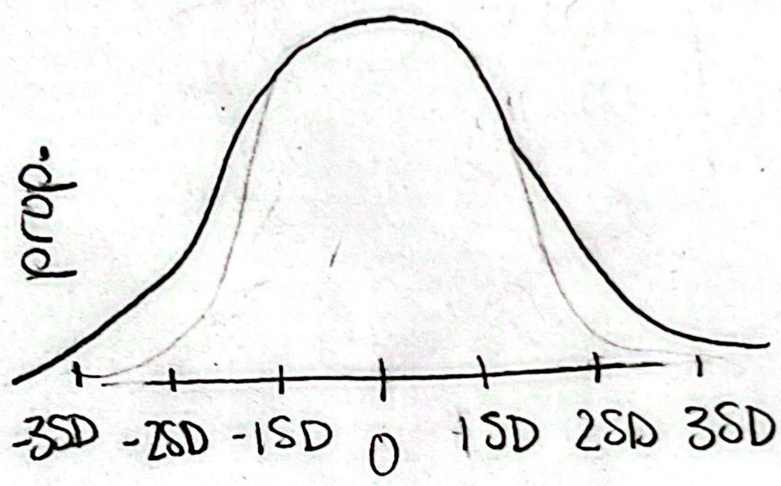
$z_{\alpha/2}$ { ($\alpha = .1$) 90% → 1.645
($\alpha = .05$) 95% → 1.96
($\alpha = .005$) 99% → 2.576

(how many standard deviations you fall above/below the mean)

lower bound = $\mu - z_{\alpha/2}(SE)$
upper bound = $\mu + z_{\alpha/2}(SE)$

$SE = \frac{\sigma_x}{\sqrt{n}}$

emperical rule:



mean

68% mean ± 1SD

95% mean ± 2 SD

99.7% mean ± 3 SD

α = 100 - % of CI

for example,
$\alpha = 0.1$
95% CI, $\alpha = 0.05$,

so $\alpha/2 = 0.025$
$\alpha/2 = 0.05$

$P(|Z| \leq z_{\alpha/2}) = (1 - \alpha)$

a higher confidence interval means there is a wider range of values

"We are ___% confident that the confidence interval [lower, upper] captures the true (mean or proportion) of (context)

The statistic is $\bar{x}$ or $\hat{p}$ w/ a ___% CI of [upper bound, lower bound]

ex: Our 95% CI [lower, upper] doesn't contain 40. Thus, we have evidence against H_0 + conclude the avg hrs worked per week for adults in the US is likely not 40 hrs.

← Writing Conclusions

- what's included: what you checked, favor/against H_0, in context (avoiding technical + definitive language)

Hypothesis Testing

→ Testing between 2 competing outcomes

* Null Hypothesis (H_0) → a statement that you are trying to disprove.
* Alternative Hypothesis (H_A) → what the problem is asking / anything but the null hypothesis

Process of Hypothesis Testing:

1) Frame research question ⟶ figure out null & alternative
2) Collect Data → - random sample, - large sample } representative sample
3) Model Randomness → will do more with this step in future units
4) Analyze Data ⟶ confidence intervals (or p-values or test statistics, coming soon)
5) Form conclusion

* Type 1 Error: when we're rejecting the null hypothesis but it's actually true.
* Type 2 Error: when we're rejecting the alternative hypothesis but it is actually true.

→ Hypothesis tests can be framed

Choice \ truth	H_0	H_A
H_0	correct	type II ERROR
H_A	type I Error	correct

→ Two Sided hypotheses
ex: $H_0 = \mu$, $H_A \neq \mu$

→ one sided hypotheses
ex: $H_0 \geq \mu$, $H_A < \mu$
$H_0 \leq \mu$, $H_A > \mu$

Whether type 1 or type 2 error is worse depends on context

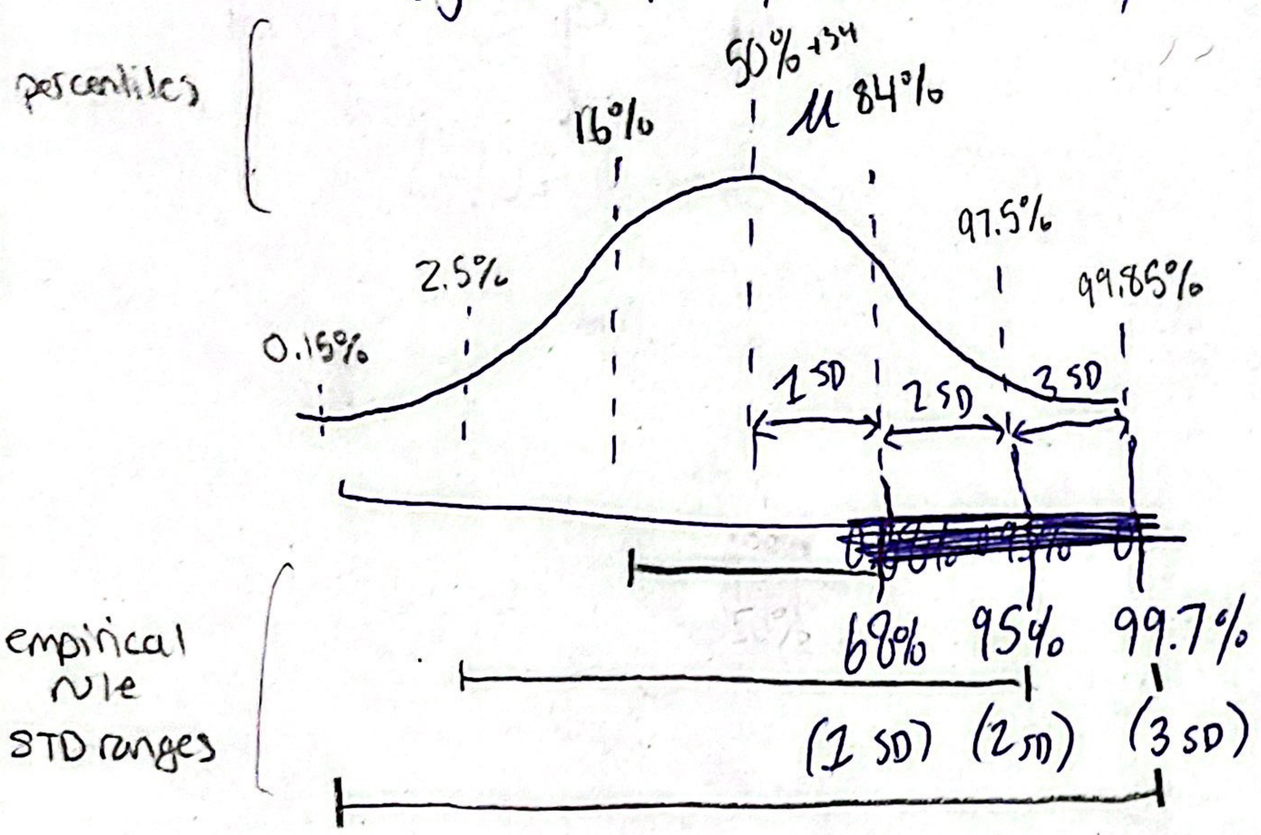
significance test → method of using data to summarize the evidence about a hypothesis

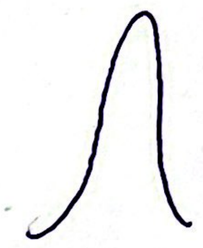
FTS!

estimated mean $\{ \bar{x} \to \mu, \hat{p} \to p \}$ True mean

Normal Distribution → characterized by its mean (μ, the center) and the standard deviation (σ, variability)

symmetric, unimodal, continuous probability distribution



more data: 

less data: 

#FTS!! Fundamental Theorem of statistics

$z_{0.05} = 1.645$ 90% CI

$z_{0.025} = 1.96$ 95% CI

$z_{0.005} = 2.576$ 99% CI

K-th percentile represents a value at which K percent of values fall below

z-score formula:

$$z = \frac{x - \mu}{\sigma} \sim N(1, 0)$$

Remember: you need at least Thirty data points to provide a normal distribution

Standard Error:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

confidence intervals:

lower = point estimate $- z_{\alpha/2} \times SE$

upper = point estimate $+ z_{\alpha/2} \times SE$

z-scores not effected by unit

the area under a normal distribution always equals 1

shorthand notation → $N(\mu, \sigma)$

μ = mean

σ = SD

$z_{\alpha/2}$ is the $(1 - \frac{\alpha}{2})\ 100^{th}$ percentile

In R: pnorm() gives percentiles

in R: normTail(z, σ, SD) gives shaded

qnorm() gives z score given percentile

PDFs → represented by histograms + ridge/density plots

PMFs → represented by plots, tables, + the complicate function below (see star)

Random Variables

P(X = x) always equals 1

A random process with a numeric outcome

↳ Discreet & continuous types

Typically denoted with a capital letter → possible outcomes notated w/a lowercase letter

Discrete random variables have probability mass functions (PMF)

PMF Continuous random variables have Probability density functions (PDF)

↳ comes with a histograms/ridge plot!

PMF Function: $P(X=x) = \frac{n!}{x!(n-x)!} p^x(1-p)^x$ ⊛

Expected value of a discrete random variable formula:

$$E(X) = \sum_{x} x \times P(X=x)$$

→ Pdfs tell us Relative probability

(weighted average of all possible values x might take on)

→ PMFs have a direct relationship w/ Probability

support = all possible values of a random variable

→ pdf's involve calc, pms do not

- E(X) is a constant ↑ a constant

$SD(X) = \sqrt{var(X)} = \sqrt{\sigma^2} = \sigma$

- E(x) is based on mathematical relationships (not empirical)

- Variance characterizes the spread of a random variable

[OI 3.13] Joint and conditional probabilities. $P(A) = 0.3$, $P(B) = 0.7$

a) Can you compute $P(A \text{ and } B)$ if you only know $P(A)$ and $P(B)$? *NO, we don't know if independent

b) Assuming that events A and B arise from independent random processes,

i) what is $P(A \text{ and } B)$? $P(A) \times P(B) = .21$

ii) what is $P(A \text{ or } B)$? $P(A) + P(B) - P(A \text{ and } B) = .79$

iii) what is $P(A|B)$? $\frac{P(A \text{ and } B)}{P(B)} = \frac{.21}{.7} = .3$

c) If we are given that $P(A \text{ and } B) = 0.1$, are the random variables giving rise to events A and B independent? not, given $\neq .21$ $P(A \text{ and } B) \neq P(A) \times P(B)$

d) If we are given that $P(A \text{ and } B) = 0.1$, what is $P(A|B)$?

$$\frac{P(A \text{ and } B)}{P(B)} = \frac{.1}{.7} = .14$$

a. No, because we don't know if it is independent or not

b.

i. $P(A) \times P(B) = 0.21$

ii. $P(A) + P(B) - P(A \& B)$

$0.3 + 0.7 - 0.21 = 0.79$

iii. $P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{0.21}{0.7} = 0.3$

C. It would be dependent, because $P(A \& B) \neq P(A) \times P(B)$

d. $\frac{0.1}{0.7} = 0.14$

[OI 3.18] Assortative mating. Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

WE LIKE IT :)

a) $P(\text{Blue or Blue}) = P(\text{Blue M}) + P(\text{Blue F}) - P(\text{Blue and Blue}) = \frac{114}{204} + \frac{108}{204} - \frac{78}{204} = \frac{144}{204} \approx 0.71$

b) $P(\text{Blue F} \mid \text{Blue M}) = \frac{P(\text{Blue and Blue})}{P(\text{Blue M})} = \frac{78}{114} \approx 0.68$

c) $P(\text{Blue F} \mid \text{Brown M}) = \frac{P(\text{Blue F and Brown M})}{P(\text{Brown M})} = \frac{19}{54} \approx 0.35$

$P(\text{Blue F} \mid \text{Green M}) = \frac{P(\text{Blue F and Green M})}{P(\text{Green M})} = \frac{11}{36} \approx 0.31$

d) $P(\text{Blue F}) = \frac{108}{204} \approx 0.53$

Answers for (b) and (c) differ and P(Blue F) differs, indicating the two variables are dependent because the value of one variable varies based on the value of the other.

if independent: $P(B) = P(B|A)$

$.53 \neq 0.71$

[OI 3.32] Is it worth it? Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an extra \$20 if he draws the ace of clubs.

a) Create a probability model and find Andy's expected profit per game.

b) Would you recommend this game to Andy as a good way to make money? Explain.

a)

Profit	2-10	face	ace	ace of clubs
	\$-2	\$1	\$3	\$23
Probability	$36/52$	$12/52$	$3/52$	$1/52$

$$E[X] = -2\left(\frac{36}{52}\right) + 1\left(\frac{12}{52}\right) + 3\left(\frac{3}{52}\right) + 23\left(\frac{1}{52}\right)$$
$$= -.54$$

b) NO, BECAUSE THE EXPECTED VALUE IS A LOSS OF MONEY

a.

Profit	-2	1	\$3	23
Prob.	$36/52$ .69%	23%	6%	2%

$$E[X] = \sum_x x \times P(X = x)$$

$$E[X] = -2(0.69) + 1(.23) + 3(0.06) + 23(0.02)$$
$$= -.54$$

b. No, the expectation is to lose around \$.54 a game

[OI 4.4] Triathlon times, Part I. In triathlons, it is common for racers to be placed into age and gender groups. Suppose two friends competed in the above below. Leo completed the race in 4948 seconds, while Mary completed the race in 5513 seconds. Here is some information on the performance of two groups for a particular race:

- Men, Ages 30 - 34: mean of 4313 seconds with a standard deviation of 583 seconds. (4313, 583)
- Women, Ages 25 - 29: mean of 5261 seconds with a standard deviation of 807 seconds. (5261, 807)
- The distributions of finishing times for both groups are approximately Normal.

a) Write down the short-hand for these two normal distributions.

b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

d) What percent of the triathletes did Leo finish faster than in his group?

e) What percent of the triathletes did Mary finish faster than in her group?

f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change?

(a) Men: $N(\mu = 4313, \sigma = 583)$, Women: $N(\mu = 5261, \sigma = 807)$

(b) $z = \frac{x - \mu}{\sigma}$ $\quad z_{Leo} = \frac{4948 - 4313}{583} \approx 1.089$

$z_{mary} = \frac{5513 - 5261}{807} \approx 0.312$

z-scores tell you how many standard deviations away from the mean they are.

(c) Mary did better because her z-score is lower. Higher z-score = slower

(d) In R: pnorm(4948, mean = 4313, sd = 583)

=> 0.8620 → 86.2% were faster than Leo

so Leo was faster than 13.8%

sub in zscore ← pnorm(1.089) → .862

(e) In R:

pnorm(5513, mean = 5261, sd = 807)

=> 0.6226 → 62.3% were faster than Mary,

so Mary was faster than 37.7%

sub zscore ← Pnorm(.312) → .6226

(f) a - c stays same, d - e change

[OI 4.6] Triathlon times, Part II. In Exercise 4.4 we saw two distributions for triathlon times: $N(\mu = 4313, \sigma = 583)$ for Men, Ages 30 - 34 and $N(\mu = 5261, \sigma = 807)$ for the Women, Ages 25 - 29 group. Times are listed in seconds. Use this information to compute each of the following:

a) The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

b) The cutoff time for the slowest 10% of athletes in the women's group.

(a) In R: qnorm(0.05, mean = 4313, sd = 583)
=> 3354.05 seconds

(b) In R: qnorm(0.9, mean = 5261, sd = 807)
=> 6295.2 seconds

[IMS 13.5] Repeated water samples. A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

a) What is this distribution called?

b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

c) What is the name of the variability of this distribution.

d) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

a) sample distribution

b) symmetric because its a distribution of sample proportions, assuming that there are? at least 10 that satisfies the condition of our sample (5%-30%

c) Standard Error

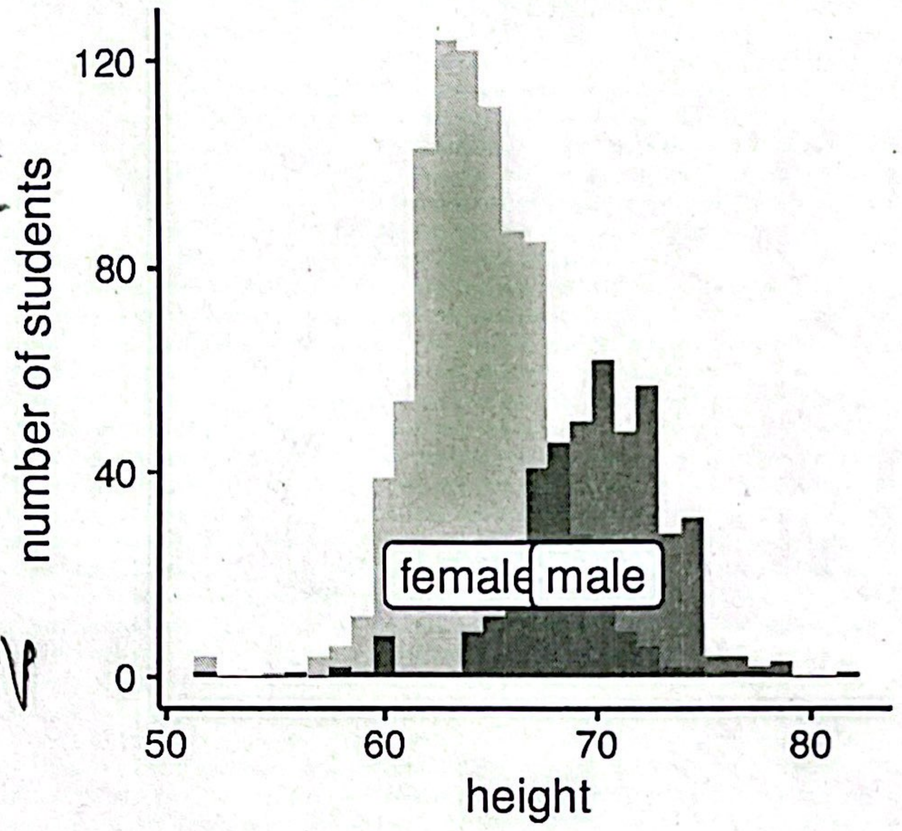
d) larger variability with fewer samples.

We will consider the `speed_gender_height` data from the `openintro` package. This data set contains the recordings of 1,292 UCLA students. These students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender. Here are the first ten rows of data.

Question 1

a. Consider the histogram of heights of male and female participants from a UCLA survey below created with the raw data. Do the two distributions represent data distributions, sampling distributions, or bootstrap distributions?

data distribution b/c tracks survey data and does not predict general pop. values + is not from a bootstrap study



b. Does it seem like a normal distribution would be a good fit for these distributions? yes, both graphs follow somewhat of a normal curve

c. Consider the summary statistics below. Assuming the two groups are normally distributed, write the shorthand notation for distribution for each of the male and female heights of students from this survey.

gender	Mean	Std.Dev	n
female	64.35	2.99	863
male	69.64	3.54	439

female ~ N(64.35, 2.99)
male ~ N(69.64, 3.54)

d. Calculate a 95% confidence interval for each group.

e. You will not be able to bootstrap on the exam, but you can still consider the bootstrap distribution. Suppose all female-identifying UCLA students are the population, and recall our sample has 863 female-identifying students in it. If you were to guess, what would your bootstrap distribution look like for mean hieght of female-identifying students? Write the notation. ? similar to data distribution but narrower distribution, approx same mean

female ~ N(64.35, 1.5)

f. Consider two subsets of the `speed_gender_height` data. The first subset contains all 863 female-identifying students from the original data set. The second subset only contains the 30 female-identifying students from the original data set. Two bootstrap distributions ($B = 10,000$) were created from these data sets and are plotted below. How do the distributions differ?

female

$SE = \frac{2.99}{\sqrt{863}} = 0.102$

lower → $64.35 - 1.96(0.102) = 64.15$

upper → $64.35 + 1.96(0.102) = 64.55$

[64.15, 64.55]

male

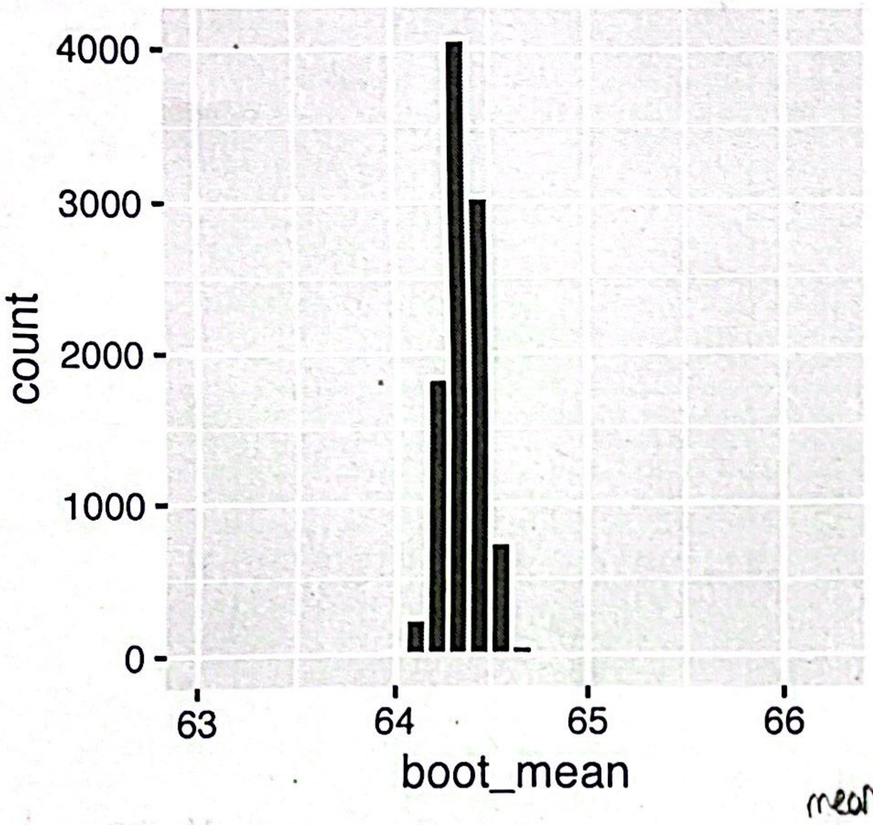
$SE = \frac{3.54}{\sqrt{439}} = 0.169$

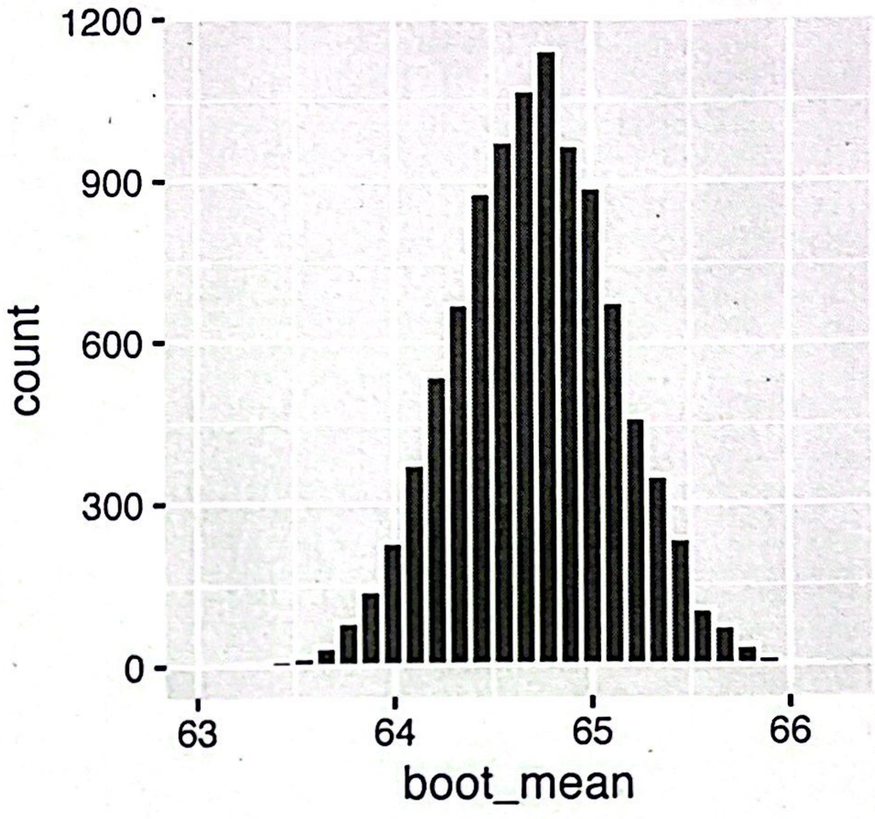
lower → $69.64 - 1.96(0.169) = 69.31$

upper → $69.64 + 1.96(0.169) = 69.97$

[69.31, 69.97]

The bootstrap w/ less observations creates a much wider + more symmetrical graph than the bootstrap w/ more observations.





larger sample size means that bootstrap means will be closer to population mean

g. If we calculated a 95% bootstrap confidence interval of heights using the $n = 863$ female-identifying students how would it compare to the confidence interval we generated in question 1c? It would be centered around the sample mean as opposed to the true mean.

h. If we make a 99% bootstrapped confidence interval how will that compare to the bootstrap confidence interval in question 1g? It would be wider b/c it includes more data but still centered around the sample mean.

i. Interpret your confidence interval for female-identifying students created in 1c.

j. A friend claims that female-identifying students at UCLA have an average height of 70 inches. Write out the null and alternative hypotheses that correspond to this claim. Using a 95% confidence interval, would you believe your friend's claim?

→ We are 95% confident that the confidence interval [64.15, 64.55] captures the avg height of female-identifying UCLA students.

→ μ = avg height of female-identifying UCLA students

$H_0: \mu = 70$

$H_A: \mu \neq 70$

Our 95% confidence interval [64.15, 64.55] doesn't contain 70. Thus, we have evidence against H_0 and conclude the avg height for female-identifying UCLA students is not 70 inches.

Question 2

In a recent study 109 moderately obese subjects were given a low-carbohydrate diet. The prediction was that the subjects would lose weight on the average. After two years, the mean change was -5.5 kg with a standard deviation of 7.0 kg.

a) What population is under consideration in the data set?

b) What parameter is being estimated?

c) What is the point estimate for the parameter?

d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

e) Compute the value from part 2d for this context.

f) A recent magazine claimed that following a low-carbohydrate diet can help people lose 2 kg per week. How would you set up your hypothesis to test if the claimed rate agrees with the data observed from the study?

a) The population is moderatley obese adults

b) The mean weight change ($\bar{x}$)

(c) $-5.5\ kg(\bar{x})$

(d) standard error

(e) $\frac{7.0}{\sqrt{109}} = 0.67$, $SE = \frac{sd}{\sqrt{sample\ \#}}$

(f) 1) null hypothesis = $H_0 = -2\ kg/week$

$H_A \neq -2\ kg/week$

2) collect data of different subjects who are still within the moderatley obese adults.

3) Confidence Interval

→ see if null hypothesis is within the confidence interval.