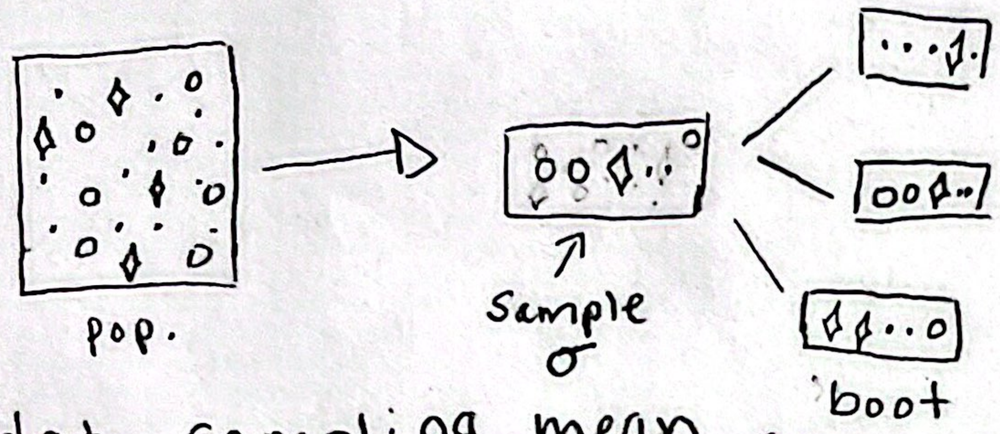
Bootstrap

- Reselecting original sample size w/ replacement to simulate a statistic distribution
- can be skewed by outliers in original sample, centered around sample mean or proportion



- essential that sample is a random sample from the population
- data sampling mean should be the same as the sampling distribution mean
  - the bootstrap mean should hover around the means
- The standard dev. of a sampling distribution and a bootstrap distribution are the same, though centers may differ
- the more times bootstrapping is done, the more accurate center and spread you can get
  - ↳ >5000 bootstraps
- bootstrap distribution gives confidence rather than probability.
- big data leads to less variability within bootstrap ($n > 30$)

Scanned with CamScanner

Probability

$P(A \text{ and } B) = P(A) + P(B) - P(A \text{ or } B)$ → General Addition Rule

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

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

complement: $P(A) = 1 - P(A^c)$

conditional: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$ ← if A and B are dependent

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

total probability: $P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_k)P(A_k)$

*if A and B are independent, $P(A|B) = P(A)$

$\text{Var}(X) = \sum_x (x - E[x])^2 \cdot P(X = x)$ ← variance

Scanned with CamScanner

Confidence Intervals: An interval containing the most plausible values for a parameter

- Common confidence levels:
  - 90%, $\alpha = 0.1$, $z_{\alpha/2} = 1.645$
  - 95%, $\alpha = 0.05$, $z_{\alpha/2} = 1.96$
  - 99%, $\alpha = 0.01$ $z_{\alpha/2} = 2.576$

$$z\text{score} = \frac{x - \mu}{\sigma}$$

Z-scores measures the #SDs it falls above/below the mean.

- With bootstraps:
  - Construct confidence intervals using percentile

- Interpreting a Confidence Interval:
  - The statistic is $\bar{x}$ or $\hat{p}$ with a 95% confidence interval of [lower bound, upper bound]

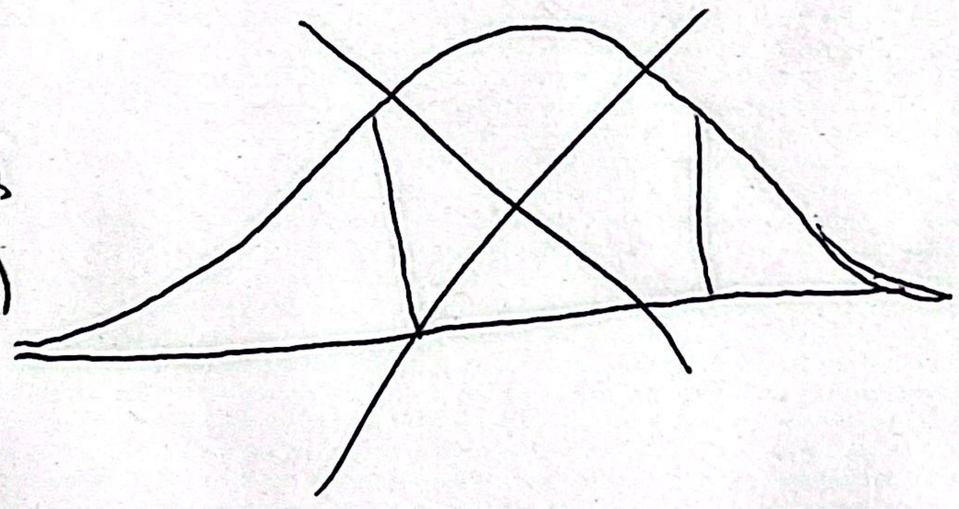
- Confidence Interval mathematically:
  - lower bound = point estimate $- z_{\alpha/2} \times SE$
  - upper bound = point estimate $+ z_{\alpha/2} \times SE$

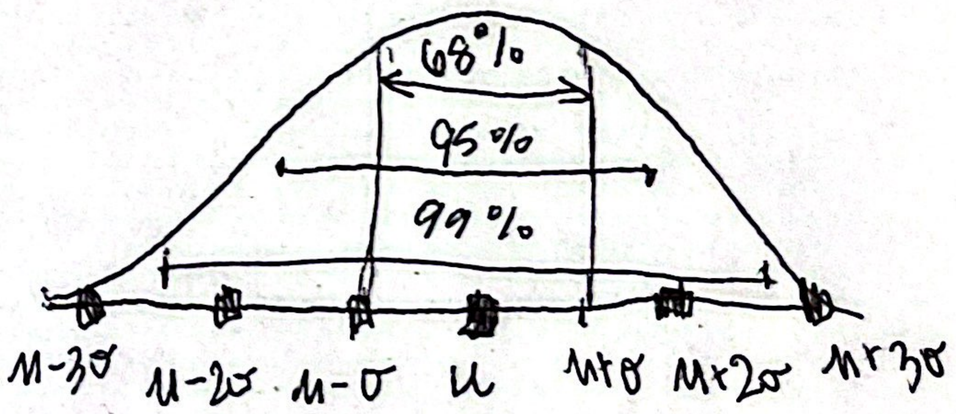
$$SE(\sigma_{\bar{x}}) = \frac{\sigma_x}{\sqrt{n}}$$

or

We are ___% confident that CI [ ] captures the (true mean or proportion)

... add context!





Upper/Lower Bounds
- 68%: $\mu \pm \sigma$
- 95%: $\mu \pm 2\sigma$
- 99%: $\mu \pm 3\sigma$

is this percentiles or CI's?

Scanned with CamScanner

Hypothesis Testing

1. frame research Question in terms of hypotheses (H_0 and H_A)
   - H_0 is usually the skeptical perspective
   - H_A is any other option
2. collect data. H_0 is usually $=, \leq,$ or $\geq$
   - independent, enough, and representative data
3. model randomness if H_0 was true
   $$\bar{X} \sim N\left(\mu, \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$$ • assume/fit
4. analyze data
   - use confidence intervals
     - calculate a $(1-\alpha)100\%$ CI
5. form a conclusion
   - if μ is in the interval, we have evidence to support the claim

   include:
   - what you checked (CI, p-value, test statistic)
   - favor / against H_0
   - context

	Truth: H_0:	H_A:
choice: H_0	correct	Type 2
choice: H_A	Type 1	correct

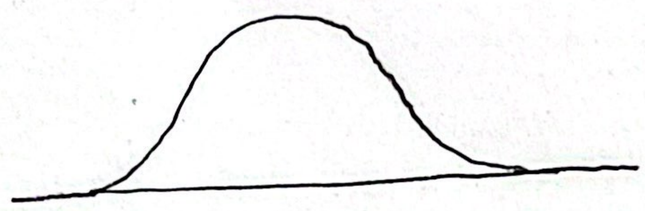
Scanned with CamScanner

Normal Distribution

- Symmetric, unimodal, bell-shaped, continuous, probability distribution

- $N(\mu, \sigma)$ → where μ = mean and σ = standard deviation
- Sampling distribution of $\hat{p}$ converges to a normal distribution as n increases

- z score: observation of x that follows a distribution with μ & σ
  - ↳ $z = \frac{x - \mu}{\sigma}$



* z-scores transform a random variable $X \sim N(\mu, \sigma)$ into a standard normal distribution where $z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

- bootstrapping a distribution leads to normal curve if $n \geq 30$
  - ↳ $\bar{x}$ and $\hat{p}$ will be normally distributed
- individual variables can follow normal distribution

  - Standard error: when sampling mean = $\bar{x}$

$$\bar{x} = N(\mu, \sigma_{\bar{x}}) \quad SE = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

(obtain from data distribution)

- 68% of data falls within 1 SD, 95% within 2, 99.7% within 3
- For a 90% CI = $z_{.05} = 1.645$, 95% CI = $z_{.025} = 1.96$, 99% CI = $z_{.005} = 2.576$

* These principles only work under the presumption that the distribution is "approximately normal"

Scanned with CamScanner

Random Variables

A random process or variable with a numerical outcome

- can be discrete or continuous

$P(X=x) = \frac{n!}{x!(n-x)!} p^x (1-p)^x$ - probability mass function

$Var(aX + bY) = a^2 \cdot var(x) + b^2 \times var(Y)$ - linear combos of random variables

$E[X] = \sum_x x \cdot P(X=x)$ ← expected val. of discrete random var.

table form

description		
X		
P(X=x)		

- variance properties
  - $Var(c) = 0$ → c is a constant
  - $var(x + c) = Var(x)$
  - $var(xc) = c^2 Var(x)$

Expected values properties
$E[c] = c$
$E[cX] = c\,E[x]$
$E[x+Y] = E[X] + E[Y]$

X = random variable

x = observable value

Scanned with CamScanner

[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 $P(B)$?

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

   i) what is $P(A \text{ and } B)$?

   ii) what is $P(A \text{ or } B)$?

   iii) what is $P(A|B)$?

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?

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

a) No, you cannot compute P(A and B) if only given P(A) and P(B) because you don't know their relationship. Only if they are independent can you compute P(A and B).

b) i) $P(A \text{ and } B) = 0.3(0.7) = 0.21$

ii) $P(A \text{ or } B) = 0.3 + 0.7 - 0.21 = 0.79$

iii) $P(A|B) = P(A) = 0.3$

c) No, not independent because calculated $P(A \text{ and } B) = 0.21$, which $\neq 0.1$.

d) $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)} = \dfrac{0.1}{0.7} = 0.143$

(a) You can only compute then if they are independent of each other.

(b) i. $P(A \text{ and } B) = 0.3(0.7) = 0.21$

ii. $P(A \text{ or } B) = (0.3 + 0.7) - 0.3(0.7)$

$= .79$

iii. $P(A/B) = (0.3(0.7)) / 0.7$

$= 0.3$

(c) No, $P(A \text{ and } B) = 0.2 \neq 0.1$

(d) $P(A/B) = \dfrac{0.1}{0.7} = 0.143$

Scanned with CamScanner

[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

* this contingency table is counted as couples

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.

a) A → male is the one w/ blue eyes in the couple
B → female is the one w/ blue eyes in the couple

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

$= \frac{108}{204} + \frac{114}{204} - \frac{78}{204} = \frac{144}{204}$ ✓

$\left(\frac{114}{204}\right) + \frac{108}{204} - \frac{78}{204} =$

b) $P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{78/204}{114/204} = \frac{78}{114}$ ✓

$\frac{78}{204} \div \frac{114}{204} = \frac{78}{204} \cdot \frac{204}{114}$

c) C → male is the one w/ brown eyes in couple
B → female is the one w/ blue eyes in couple

1. $P(B|C) = \frac{P(C \text{ and } B)}{P(C)} = \frac{19/204}{54/204} = \frac{19}{54}$ 2. $\frac{11}{54}$ $\frac{11}{36}$

d) You can do a math check: $P(A)P(B) = P(A \text{ and } B)$

ex: $\frac{108}{204} \times \frac{114}{204} = 0.296$ $\frac{78}{204} = 0.382$

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

$0.382 \neq 0.296$, so they are not independent

Scanned with CamScanner

[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.

DESCRIPTION

	cards 2-10	Face card Jacks, Queens	Ace	Ace of clubs
x	-2	1	3	23
P(x)=x	36/52	12/52	3/52	1/52

b) $E(x) = -2(36/52) + 1(12/52) + 3(3/52) + 23(\frac{1}{52})$

$= -\$0.54$

WOULD NOT RECOMMEND! LOSE MONEY

The expected earnings, on average, will be negative.

description	# C2-10	face	ace	ace club
X	0	3	5	25
P(X=x)	32/52	12/52	3/52	1/52

* other way to do problem

← remember to subtract cost to play (\$2)

a) $E(x) = 0(32/52) + 3(12/52) + 5(3/52) + 25(1/52)$

$= 1.46 - 2 = \$0.54$

b) expected winnings are below how much it costs to play = LOSE MONEY

Scanned with CamScanner

[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.
- Women, Ages 25 - 29: mean of 5261 seconds with a standard deviation of 807 seconds.
- 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?

Men:
a) $X \sim N(4313, 583)$

Women
$x \sim N(5261, 807)$

b) Z-scores:

Leo $z = \frac{4948 - 4313}{583} = 1.089$ Mary: $\frac{5513 - 5261}{807} = 0.31$

These z scores tell you how many standard deviations above (from) the mean Leo and Mary finished. *or how many standard deviations away from the mean that Leo and Mary's times are.*

c) Mary ranked better in her respective group because her time is less standard deviations above the mean,

d) 1 − pnorm(1.089, mean = 0, sd = 1) = 0.138 (4313, 583)

e) 1 − pnorm(0.312, mean = 0, sd = 1) = 0.378 (5261, 807)

f) Part b and c can still be calculated. The relative z scores, however, cannot be compared across the two different races However the z-scores still be compared and have little to no meaning. Thus, c, d and e cannot be calculated. Because the distributions are not normal, you cannot find percentiles.

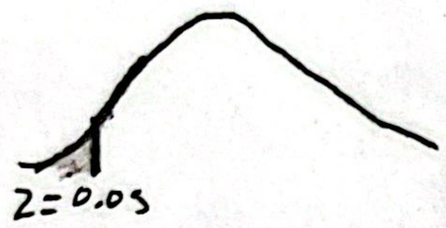
Scanned with CamScanner

[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:

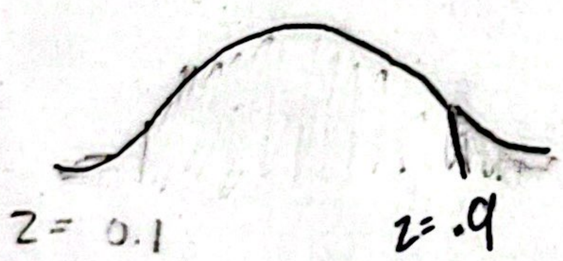
a) The cutoff time for the fastest 5% of athletes in the men's group, i.e. those who took the shortest 5% of time to finish.

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

a) qnorm (0.05, μ = 4313, σ = 583) = 3354 seconds



b) qnorm (0.9, μ = 5261, σ = 807) = 6295 seconds



lower (0.95

upper (0.05)

Scanned with CamScanner

[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) ~~Like a Bootstrap dist~~ sampling distribution. ✓

b) $n \geq 30$ thus ~~the bootstrap~~ will be a normal distribution (symmetric). (when just means, but diffent for proportion)

c) Standard Error measures how many standard deviations the statistic (sample proportion(s)) is from the true parameter. This measures the variability of this distribution. ✓

d). sample size (n) ↓ ⇒ ↑ variability

---

b) would need point estimate to better estimate the shape of the distribution. so we would assume it is symmetric as there is a range for the estimate.

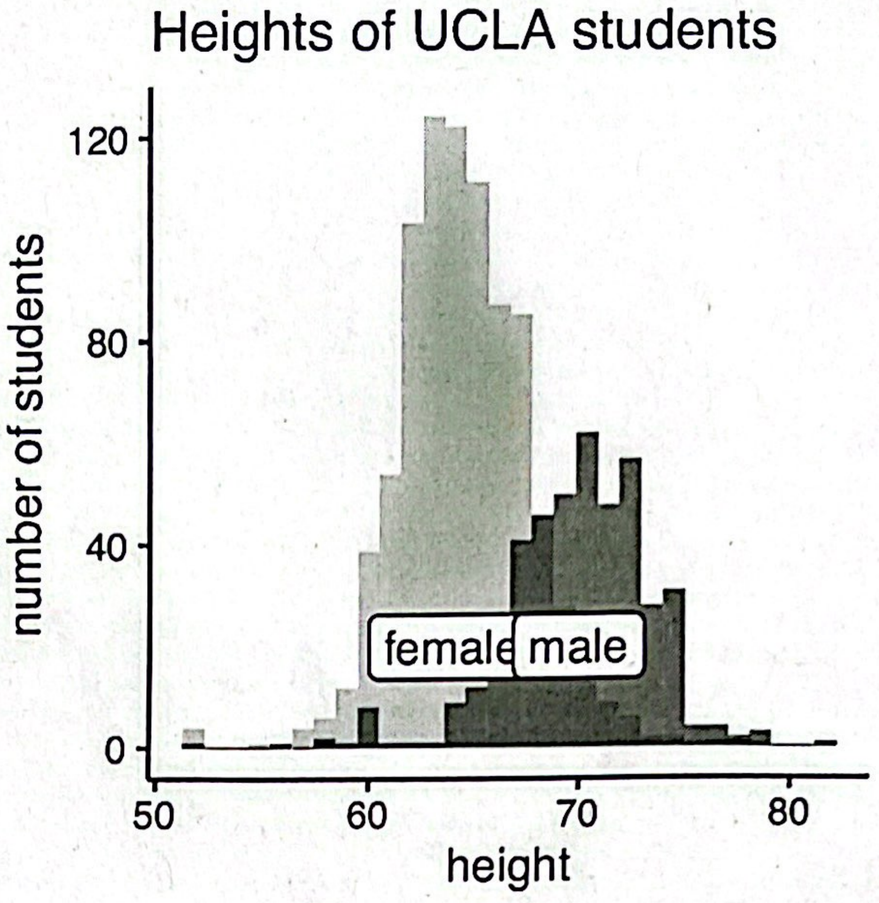
Scanned with CamScanner

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. Does it seem like a normal distribution would be a good fit for these distributions? yes

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

X = women

Y = men

$X \sim N(\mu = 64.35, \sigma = 2.99)$

$Y \sim N(\mu = 69.64, \sigma = 3.54)$

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

women = [58.49, 70.2], men = [62.7, 76.58]

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.

$X \sim N(\mu = 64.35, \sigma_{\bar{x}} = 0.108)$ $Y \sim N(\mu = 69.64, \sigma_{\bar{x}} = 0.171)$

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?

d)

64.35 − 1.96(2.99) = 58.49] women

64.35 − 1.96(2.99) 70.2]

69.64 − 1.96(3.54) = 62.7

69.64 + 1.96(3.54) = 76.58

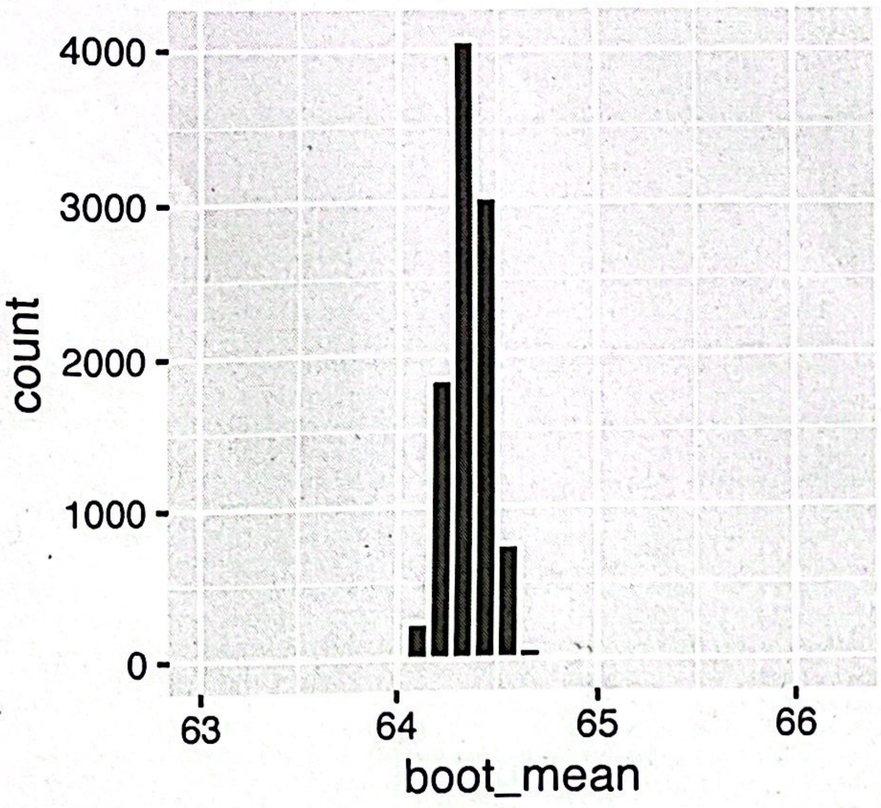
e) $\frac{2.99}{\sqrt{863}} = 0.108$

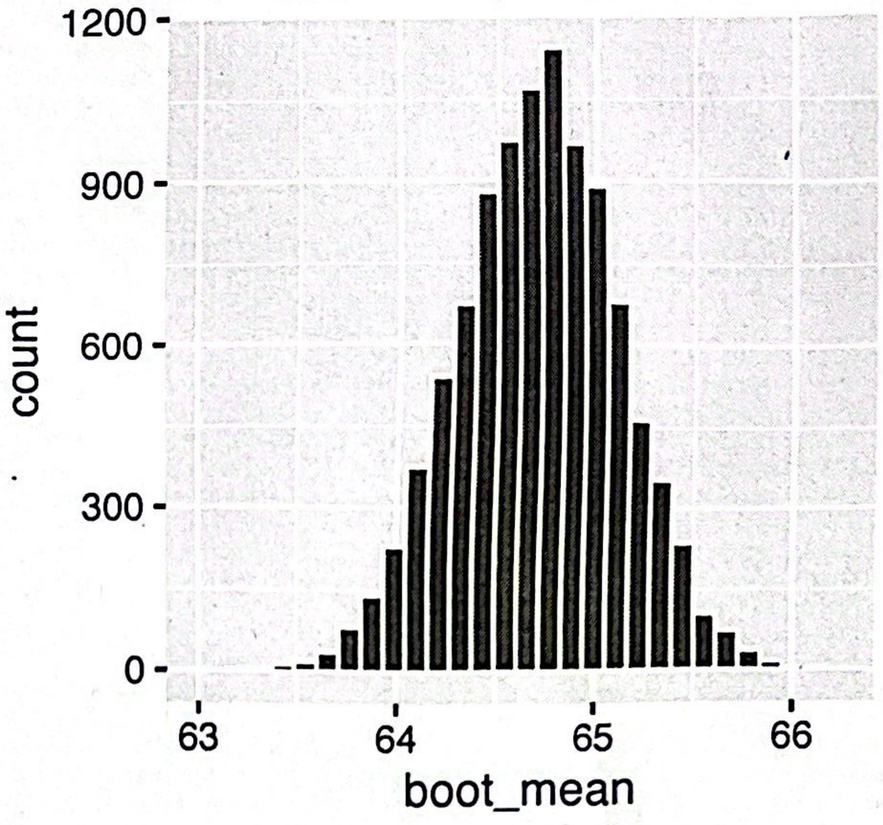
$\frac{3.54}{\sqrt{439}} = 0.171$

f) The distribution with the n=30 appears to be normal, but has less observations

n=30 → has a larger standard error due to smaller # of observations

Scanned with CamScanner





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? smaller standard error because the standard error is smaller for all n>1 (compared to the Standard distribution)

h. If we make a 99% bootstrapped confidence interval how will that compare to the bootstrap confidence interval in question 1g? It would be a wider range of values

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?

i.) We are 95% confident that the CI [58.49, 70.2] captures the true mean height of female-identifying students at UCLA.

j.) $H_0 = 70$ in.

$H_A \neq 70$ in

It falls within the confidence interval so it could be a reasonable claim

Scanned with CamScanner

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) ~~109~~ moderatley obese subjects

b) Mean change in Weight

c) -5.5 kg

d) Standard Error?

e) $\frac{7}{\sqrt{109}} = 0.67$

f) $H_0 = -2$

$H_A \neq -2$

a) moderately obese subjects

b)

a) moderatley obese people

b) Mean change in weight (kg)

c) $-5.5\text{kg} = \bar{x}$

d) standard error

e) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7.0}{\sqrt{109}} = 0.67$

f) $H_0 = \mu = 2$

$H_A = \mu \neq -2$

f) $\frac{-2\text{kg}}{\text{week}} \cdot \frac{52\text{ week}}{\text{year}} \cdot 2\text{ years} =$

$-208\text{ kg in } 2\text{ yrs}$

Agree

$H_0: \mu = -208\text{kg}$

$H_a: \mu \neq -208\text{kg}$

Because the problem asks for the claim "-2 kg per week" and the study/data is over 2 yrs (104 wks), the

Scanned with CamScanner