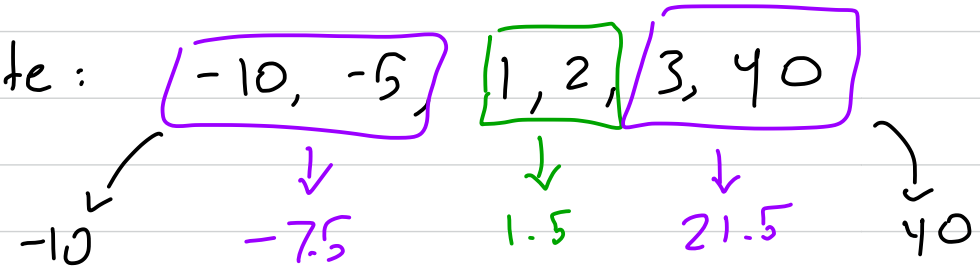
Ex: Find the 5-number summary of

1, 2, 3, 40, -10, -5

Rewrite: -10, -5, 1, 2, 3, 40



5-num: -10, -7.5, 1.5, 21.5, 40

outliers? $IQR = Q_3 - Q_1 = 21.5 - (-7.5)$

$= 29$

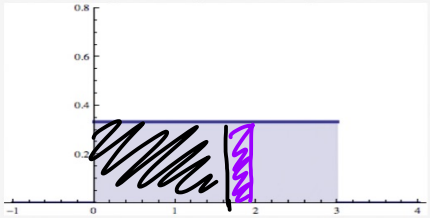
$Q_3 + 1.5\,IQR = 21.5 + 43.5 = 65$

$Q_1 - 1.5\,IQR = -7.5 - 43.5 = -51$

No outliers

(1 point) **Oregon/MA243/Moore5-3.2.pg**

Examining the location of accidents on a level, 3-mile bike path shows that they occur uniformly along the length of the path. This figure



displays the density curve that describes the distribution of accidents.

(a) The proportion of accidents that occur up to mile 1.7 of the path is the area under the density curve between 0 miles and 1.7 miles. What is this area?

(b) Sue's property adjoins the bike path between the 1.7 mile mark and the 1.9 mile mark. What proportion of accidents happen in front of Sue's property?

(a)

(b)

total area = 1

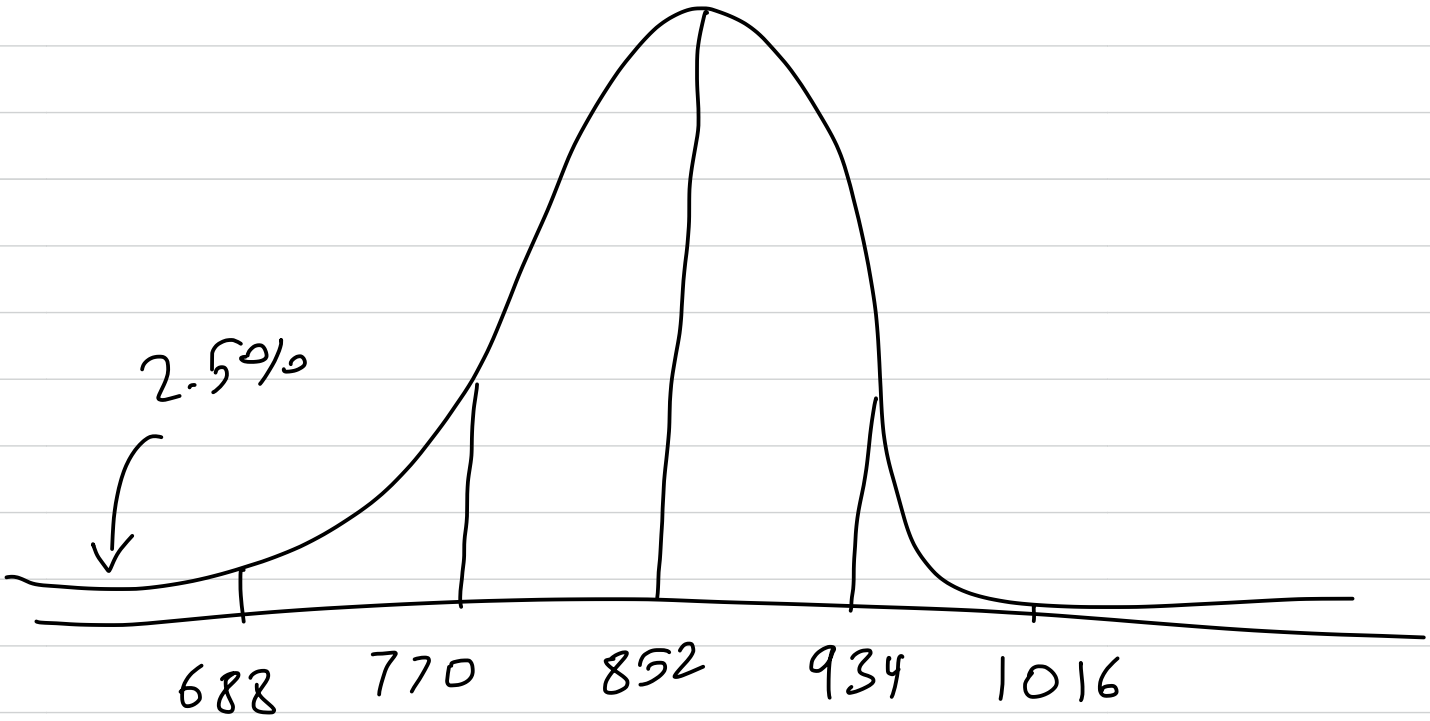
proportion before 1.7 is $\frac{1.7}{3}$

$\frac{.2}{3}$ ← 1.9 − 1.7

80% rain in summer

monsoon: $N(852, 82)$

Between what for 95%?



Chapter 9: Experiments

Def: An observational study observes individuals and measures certain variables, but doesn't attempt to change any of them.

Def: An experiment deliberately assigns some individuals to treatments to study whether the treatments cause changes in certain variables.

Types of data collection (so far):

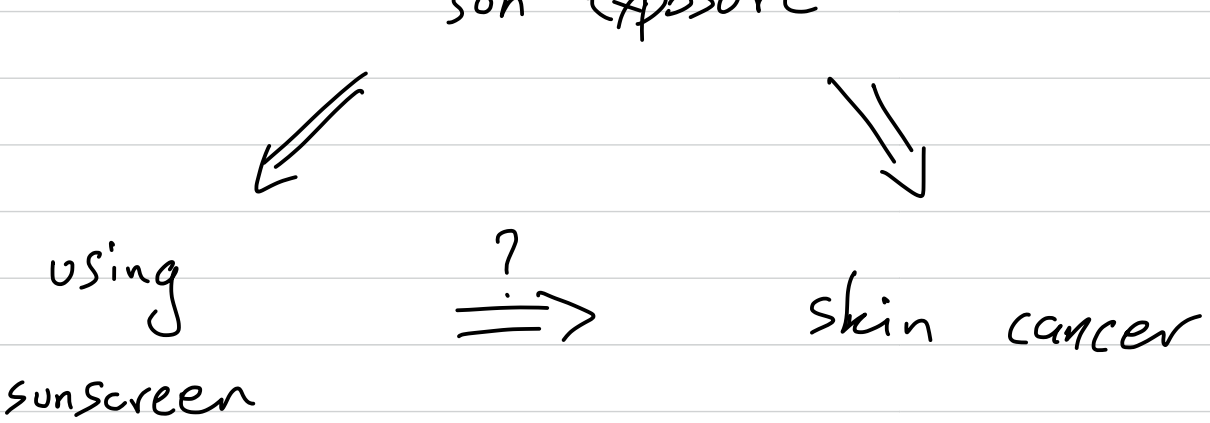
Surveys

Observational studies

Experiments ← best for determining cause and effect

Question: in an experiment, why do we need to only give the treatment to some individuals?

Ex: it's easy to find a link between increased use of sunscreen and increased rate of skin cancer if you don't know what you're doing.



Def: We call sun exposure in the previous example a lurking variable.

Def: Similarly, we call sunscreen an explanatory variable and skin cancer a response variable. These are just hypotheses.

Ex: a group of 10892 middle-aged adults was studied over 9 years. People in the group who began as smokers and quit had a higher risk of diabetes within three years of quitting than nonsmokers or continuing smokers.

observational study

What type of data collection is this? What are the explanatory and response variables?

explanatory → quitting smoking

response → diabetes risk

Do you think there is a cause and effect relationship? What might some lurking variables be?

Does this data show that there

is a cause-and-effect relationship?

no – because it's not an experiment

Two things to think about: quitting smoking often causes weight gain, which increases diabetes risk.

Also: a common cause of quitting smoking is health problems

Corellation does not imply causation

Def: In an experiment, the individuals we study are called subjects, the explanatory variables are called factors,

and the different values each factor can take are called levels. A treatment is a specific experimental condition applied to a subject.

Ex: to study the effects of different harvesting conditions on mangoes, they were harvested at 80, 95, or 110 days after setting (i.e. turning from flower to fruit). Then they were stored at 20, 30, or 40 °C. For each harvest time and each storage temperature, a random sample of mangoes was selected, and the time to ripen was measured.

What are the factors, treatments, levels, and response variables?

Factors: harvest time, temp

Levels: 80, 95, 110 days and 20, 30, 40°C

Treatment: specific selection of harvest time and temp

Response variable: ripening time.

Commit: Basic principles of experiments

Control: comparing effects of treatments to non-treatments or different ones helps avoid the effects of lurking variables.

Randomization: using random chance

to assign treatments helps avoid bias.

Replication: using enough subjects in each group, and repeating the entire experiment in multiple locations helps avoid coincidences associated with small numbers.

Control: one way to implement control is via a control group, which is a group that receives either a placebo (a non-function treatment) or the current standard of treatment

Also: blocking (later)

Interpreting results: for the mangoes, either changes in ripening times were caused by random chance, or they were caused by the treatments. When we see that change, we use probability to determine how likely it was to just be random. We say a result is statistically significant if it is so strange that would rarely occur by chance.

Placebo: fake treatment that mimics the real one. Can help remove the

factor of subjects' expectations.

Double blind: neither the subjects nor the scientists know which treatment which subject is getting.

those who interact with the subjects

Blocking: subjects are grouped into similar categories, and then individuals per category are randomly assigned to treatments. Helps avoid variability and is a form of control: e.g. "we controlled for age"

Matched pairs: pairs similar subjects and assigns treatment to exactly one subject per pair.

Chapter 12: Intro to Probability

Ex: 304 people are interviewed before going into the movies. 48 have tickets to see Wonder Woman. Based on this information, the approximate probability that a randomly selected person is going to see

Wonder Woman is $48/304 \approx .158$.

Ex: The same survey also measured all theaters in the county. 36517 people were surveyed, and 6573 were seeing Wonder Woman. The probability is therefore approximately $6573/36517 \approx .18$. This is a much better approximation.

Def: A phenomenon is random if its invidual outcomes are uncertain, but the pattern of outcomes follows a predictable distribution. The probability of an outcome is the proportion of times that it would occur in a

very long sequence of repitions.

Ex: Flipping a coin: the outcome of one flip is unknown, but over time, the proportions of heads and tails tend to .5 each.

Def: The sample space of a phenomenon is the set S of possible outcomes.

Ex: $S = \{\text{heads}, \text{tails}\}$

Def: Any outcome or group of outcomes is called an event.

Ex: some events: flipping heads, flipping tails, or both (?)

Def: Two events that cannot occur at the same time are called disjoint.

Ex: heads and tails are disjoint.

Def: We can denote events with capital letters like A and their probabilities by $P(A)$.

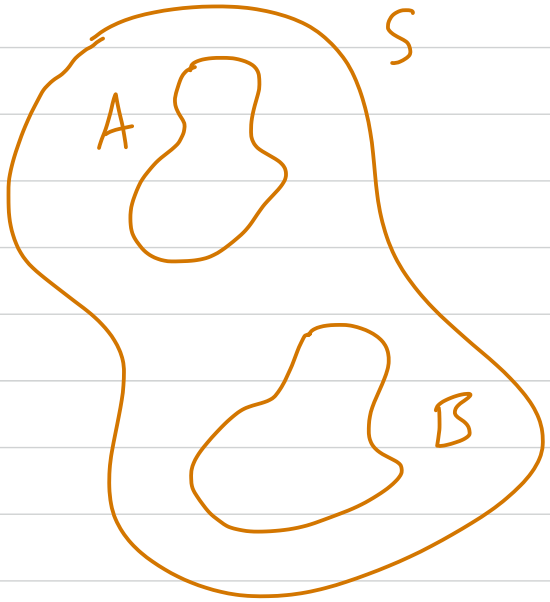
Ex: Denote heads by H and tails by T. Then $P(H) = .5$ and $P(T) = .5$.

Prop (Rules of Probability): Let E be an event.

1. $0 \leq P(E) \leq 1$.

② $P(S) = 1$, where S is the sample space.

③ If A and B are disjoint, then $P(A \text{ or } B) = P(A) + P(B)$.



④ The probability that an event A does <u>not</u> occur is $P(\text{not } A) = 1 - P(A)$

Ex: $P(T) = P(\text{not } H) = 1 - P(H) = 1 - .5 = .5$

Ex: We roll two four-sided dice and record the results of each die separately.

1. What is the sample space?
2. If the dice are fair, what is the probability of each outcome?
3. What is the probability that the sum of the two numbers rolled is exactly 5?
4. What is the probability that the sum is at least 3?

1. $\{(1,1), (1,2), (1,3), (1,4),$
$(2,1), (2,2), (2,3), (2,4),$
$(3,1), (3,2), (3,3), (3,4),$
$(4,1), (4,2), (4,3), (4,4)\}$

2. $1/16$

3. Purple outcomes have a sum of 5, and they are disjoint, so we can add their probabilities:

$$P(\text{sum of } 5) = P(1,4) + P(2,3) + P(3,2) + P(4,1)$$
$$= \frac{1}{4}.$$

4. $$P(\text{sum} \geq 3) = 1 - P(\text{sum} < 3)$$
$$= 1 - P(1,1)$$
$$= 1 - \frac{1}{16} = \frac{15}{16}.$$

Def: A random variable (typically denoted by a capital letter like X) is a placeholder for the outcome of a random phenomenon. It could take on any of the possible outcomes.

Ex: If X is the sum of the two four-sided dice, then X could take any value between 2 and 8.

$$P(X = 2) = 1/16$$

$$P(X = 5) = 1/4$$

$$P(X \geq 3) = 15/16$$

Def: A probability distribution is a list of what values a random variable

could take and their probabilities of occurring.

Ex:

X	P(X)

← nonsense!

Write tables like this:

x	P(X = x)
2	1/16
3	1/8
4	3/16
5	1/4
6	3/16
7	1/8
8	1/16

Def: A probability distribution with finitely many outcomes* is called discrete.

One in which the probabilities are

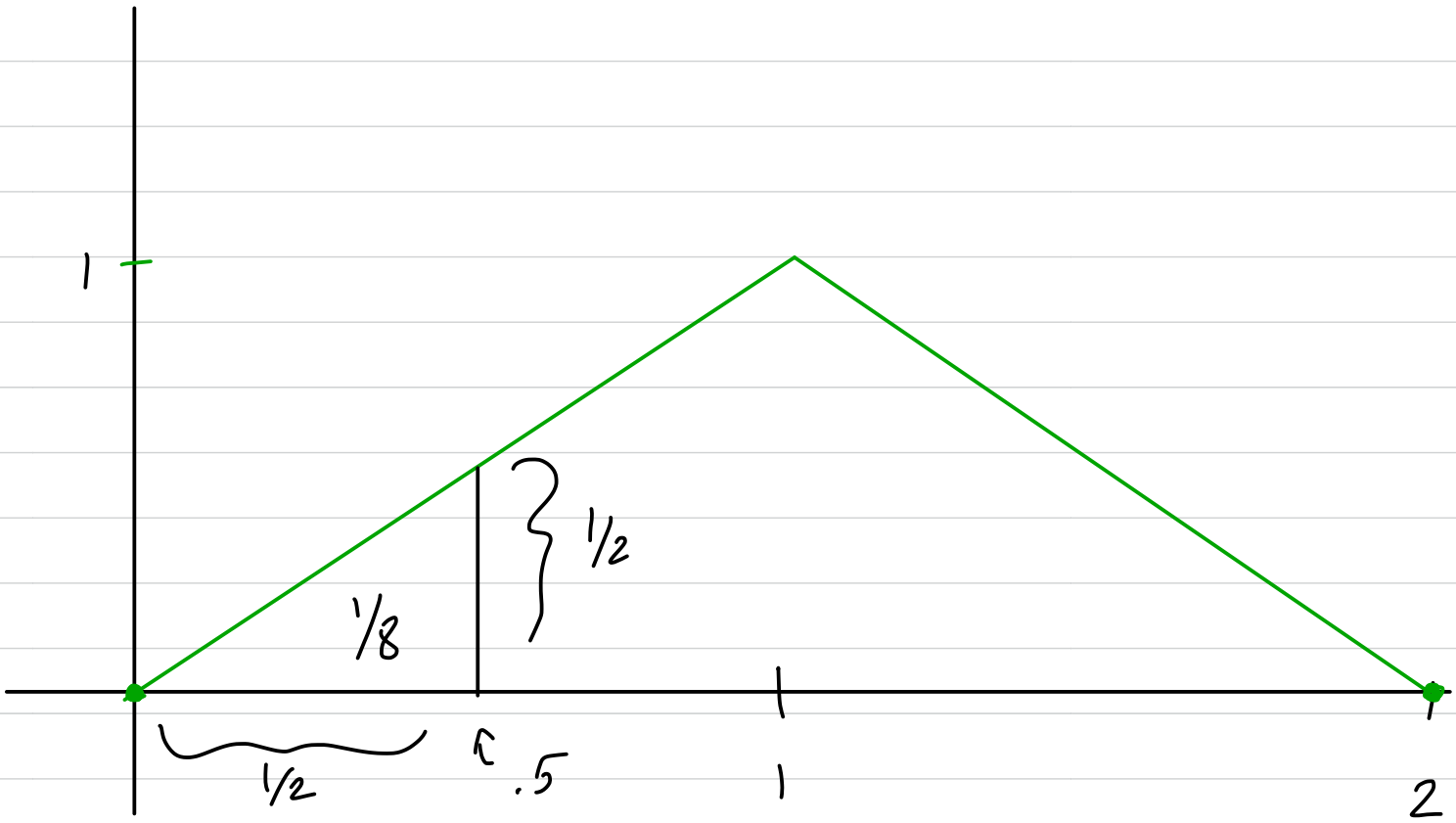
given by a density curve is called continuous.

* there are other discrete distributions, but they're out of the scope of this class.

Ex: Let X be the sum of two uniformly random numbers in [0, 1]. Then X is a continuous random variable, since it can take any value in [0, 2]. Draw the probability distribution curve for X.

The way we want this to work is for the total area of the

curve to be 1, and the area to the left of any outcome to be the probability of getting that outcome or less.



$P(X \leq .5) = \frac{1}{8}$

This is not a bell curve! There is no nice geometric representation of the

mean or standard deviation, and in general, these curves aren't symmetric. Most importantly, z-scores don't mean anything.

Chapter 13: Rules of Probability

Ex: The molecule PTC has an unusual property: about 70% of the population find it to have a bitter taste, and the other 30% find to be tasteless what is the probability that two randomly selected people

both can't taste PTC?

Def: Two events A and B are independent if knowing that one occurs doesn't change the probability that the other one does. If this is the case, then:

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

Ex: Flip a fair coin twice. Let A be the event of flipping heads on the first flip and B be the event of flipping heads on the second. Are A and B independent? They are, because knowing that you flipped heads on the first flip

has no impact on the result of the second.

Ex: Pick two sisters and ask them if they can taste PTC. Let A be event of the first sister not being able to PTC and B the event of the second not being able to taste it. Are A and B independent?

Probably not - since siblings share a gene pool, knowing one can't taste PTC probably raises the chance that the other one can't either.

Ex: You draw three cards off the top off a shuffled deck. Let A be the event of getting a heart on the first draw and B the event of getting a spade on the third. Are A and B independent?

No, because getting a heart on draw one removes it from the deck, making the proportion of spades in the remaining deck higher.

Ex: $P(\text{two heads in a row}) = (.5)(.5)$

$= .25$

P(drawing a heart and then drawing a

spade two cards later) $\neq \frac{1}{4} \cdot \frac{1}{4}$

Ex: The probability of winning a certain lottery is 5%. What is the probability of playing 30 times and never winning? All 30 plays are independent, so the probability is

$.95 \cdot .95 \cdot \cdots \cdot .95 = (.95)^{30} \approx .215$

Ex: What is the probability of winning at least once if you play 4 times?

P(winning on try 1 or on try 2 or try 3 or try 4)

Bad approach: because the chance of winning each attempt is .05, the chance of winning at least one is .05 + .05 + .05 + .05 = .2. But, these events are not disjoint!

Disjoint vs. Independent:

Disjoint: two events cannot both occur at once.

Independent: one event occuring doesn't change the probability the other one does

Ex: To actually find the chance of winning at least once, we can leverage the fact that winning

(and therefore losing) on each attempt is independent.

P(winning at least once) =

1 - P(never winning) =

1 - P(losing 4 times in a row) =

$1 - (.95)^4$ = (because of independence)

1 - .815 =

.185

↑ not that different from .2 because the chance of winning more than once is so small.

Ex: A study of distance learning courses found that most of the students enrolled in them were older than the average age among all enrolled students.

Let A = a randomly chosen student is 25 or older

B = a randomly chosen student is local

$P(A) = .7$ ← Found by survey

$P(B) = .25$

$P(A \text{ and } B) = .05$

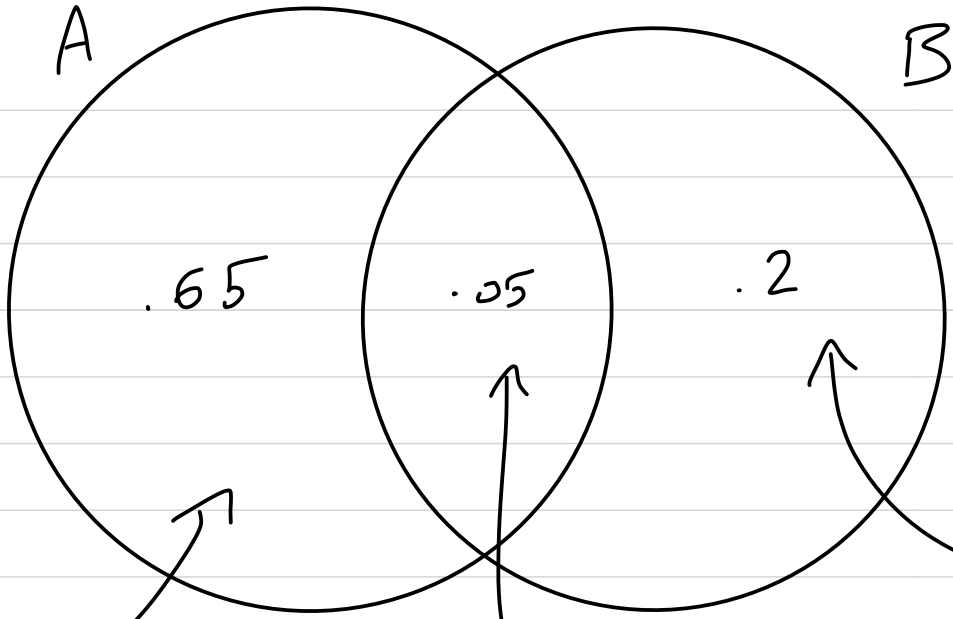
What is $P(A \text{ or } B)$?

Not disjoint, since some students are both 25 or older and local

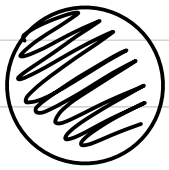
Not independent, since knowing that a student is 25 or older affects the chance that they're enrolled in distance learning.

Def: A Venn diagram is a figure that plots the probabilities of events occuring. To draw one, draw one circle for each event and write the corresponding probabilities in each region.

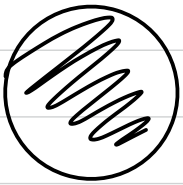
Ex:



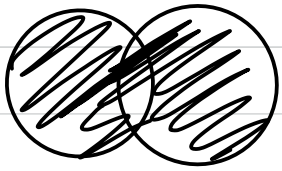
P(A):



P(B):



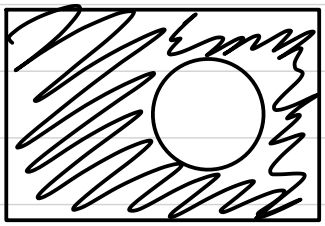
P(A or B):



: .65 + .05 + .2 = .9

P(A and B):

P(not B):



Def: Let A and B be events. The probability that A occurs, given that B has already occured, is $P(A \mid B)$ (read P of A given B).

Comment: By definition, if A and B are independent, then $P(A \mid B) = P(A)$.

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

Ex: Find $P(A \mid B)$ from the distance learning example.

$$P(A \mid B) = \frac{.05}{.25} = .2.$$

Def: When talking about conditional probability, we sometimes say that $P(A)$ is the prior and $P(A|B)$ is the posterior. The idea here is that you update your information by knowing that B has occured, so $P(A)$ is your best guess prior to that update, and $P(A|B)$ is your better guess afterward.

Ex: Table of proportions of cars in the US:

	Domestic	Imported	Total
Light trucks	.43	.07	.5
Cars	.33	.17	.5
Total	.76	.24	1

Find $P(\text{truck} \mid \text{imported})$

$$= \frac{P(\text{truck and imported})}{P(\text{imported})} = \frac{.07}{.24} = .292$$

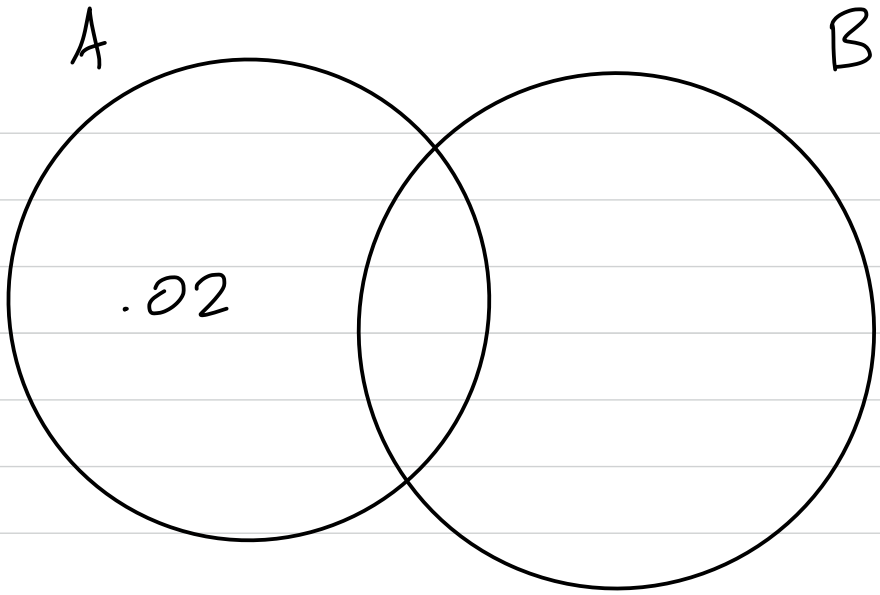
Find $P(\text{imported} \mid \text{truck})$

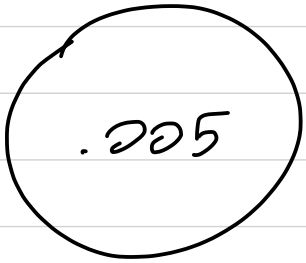
$$= \frac{P(\text{imported and truck})}{P(\text{truck})} = \frac{.07}{.5} = .14$$

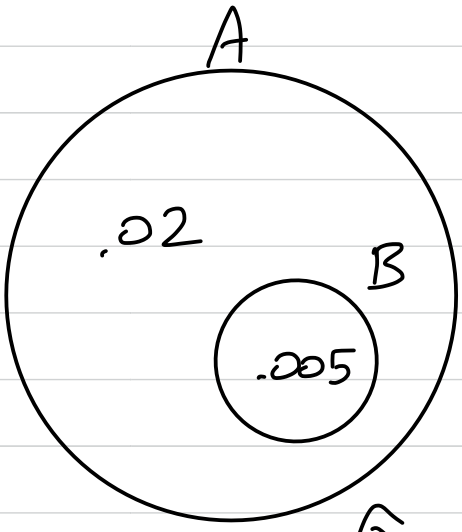
Comment: Bayes' Theorem lets you calculate $P(B|A)$ if you know $P(A)$, $P(B)$, and $P(A|B)$.

Ex: Let's say that a drug test has a 2% false positive rate (2% of the time, someone not using this type of drug is mistakenly labeled as using it). Also, suppose that .5% of the population uses this drug. Given that someone has a positive drug test, what are the chances they are actually using the drug?

If A is the event of getting a positive test and B is the event of using the drug, then







$P(A \mid B) = 1$

Want: $P(B \mid A)$

$P(A) = .02 + P(A \text{ and } B)$

$P(A \text{ and } B) = .005$

$P(A) = .025$

correct once we know all the probabilities

Thm (Bayes' Rule): $P(B|A) = \dfrac{P(A|B) \cdot P(B)}{P(A)}$

$$P(B|A) = \frac{1 \cdot .005}{.025} = \frac{1}{5}$$

The probability of a positive test result actually implying that this drug is being used is only 20%.

Chapter 15: Sampling Distributions

Def: A parameter is a variable

corresponding to a population.

Def: A statistic is a variable corresponding to a sample.

Typically, parameters are unknown and statistics are used to estimate them.

Ex: The mean wage for individuals in the US was $43041.39, and a survey of 100 households found that the mean income was $46 091. The median for the entire US was $28031.02. If we're studying the US, which values are parameters

and which are statistics?

Parameters Statistics

Ex: A group of middle school students measures the height of dandelions in a field. Let's assume that those heights are Normally distributed with some mean μ.

Group 1: 13.7, 15.3, 9.1, 16.0, 16.1

Group 2: 13.9, 9.5, 12.7, 16.9, 13.5

Group 3: 17.3, 10.8, 17.7, 11.8, 21.7

Group 4: 13.2, 13.4, 20.1, 15.8, 16.2

What is $\bar{x}$ for group 1? 14.04

What is $\bar{x}$ for groups 1 and 2 combined?

13.67

Which one better approximates μ?

The second $\bar{x}$ is a better approximation.

Why?

Thm (The Law of Large Numbers):
As a sample size n gets larger, the sample mean $\bar{x}$ approaches the population mean μ.

Ex: $\bar{x}$ for all 4 groups combined is 14.735.

This is likely a better approximation than either 13.67 or 14.04. Note: this probably implies that 13.67 was a worse approximation of μ than 14.04.

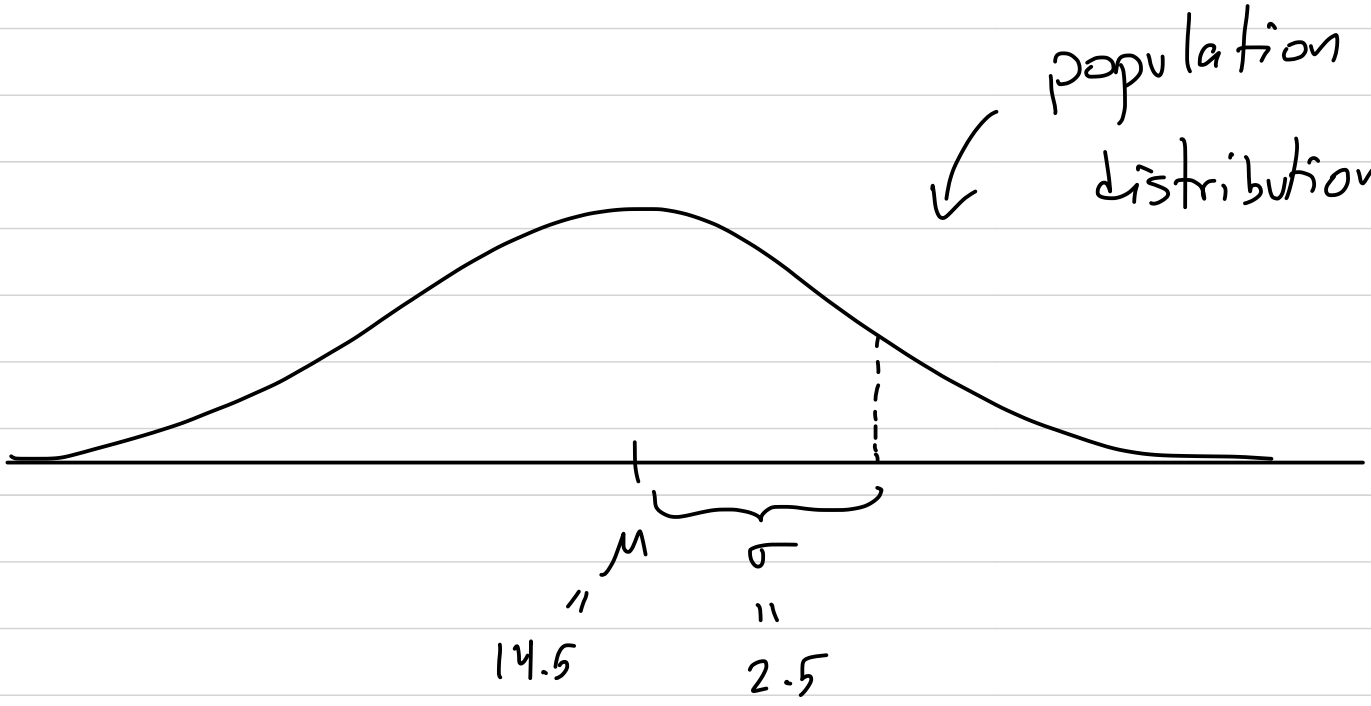
Def: A sampling distribution is the distribution of a statistic: for example, $\bar{x}$ for all possible samples of a fixed size.

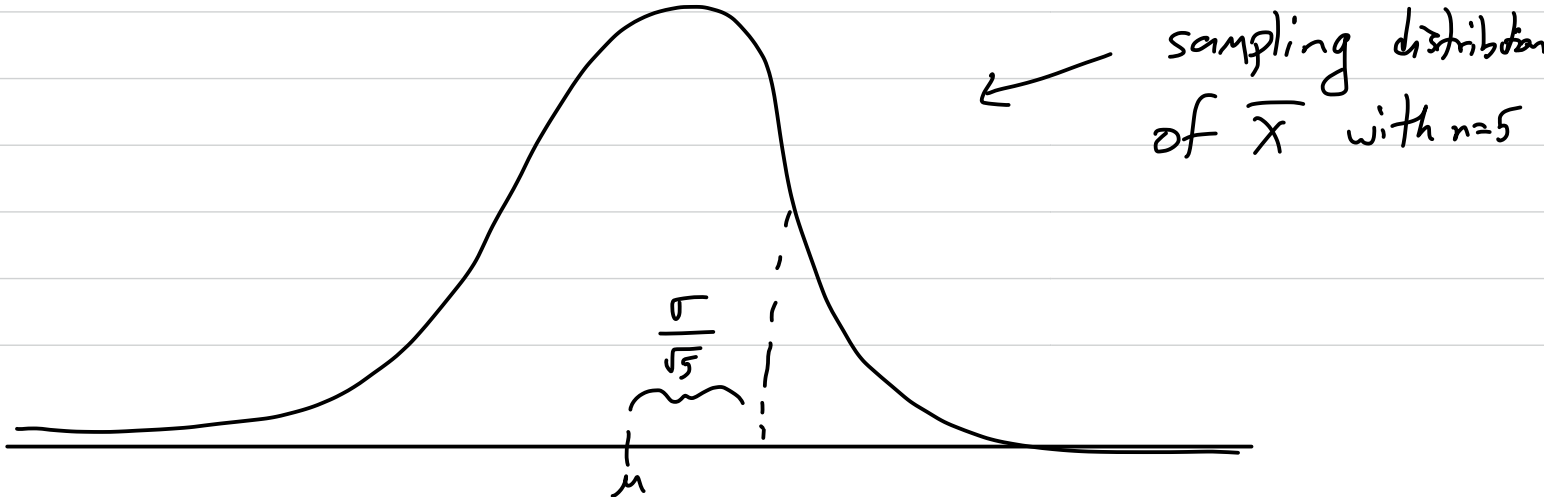
Comment: Note the difference: we have the distribution of the parameter, with mean μ and standard deviation σ (e.g. the heights of all

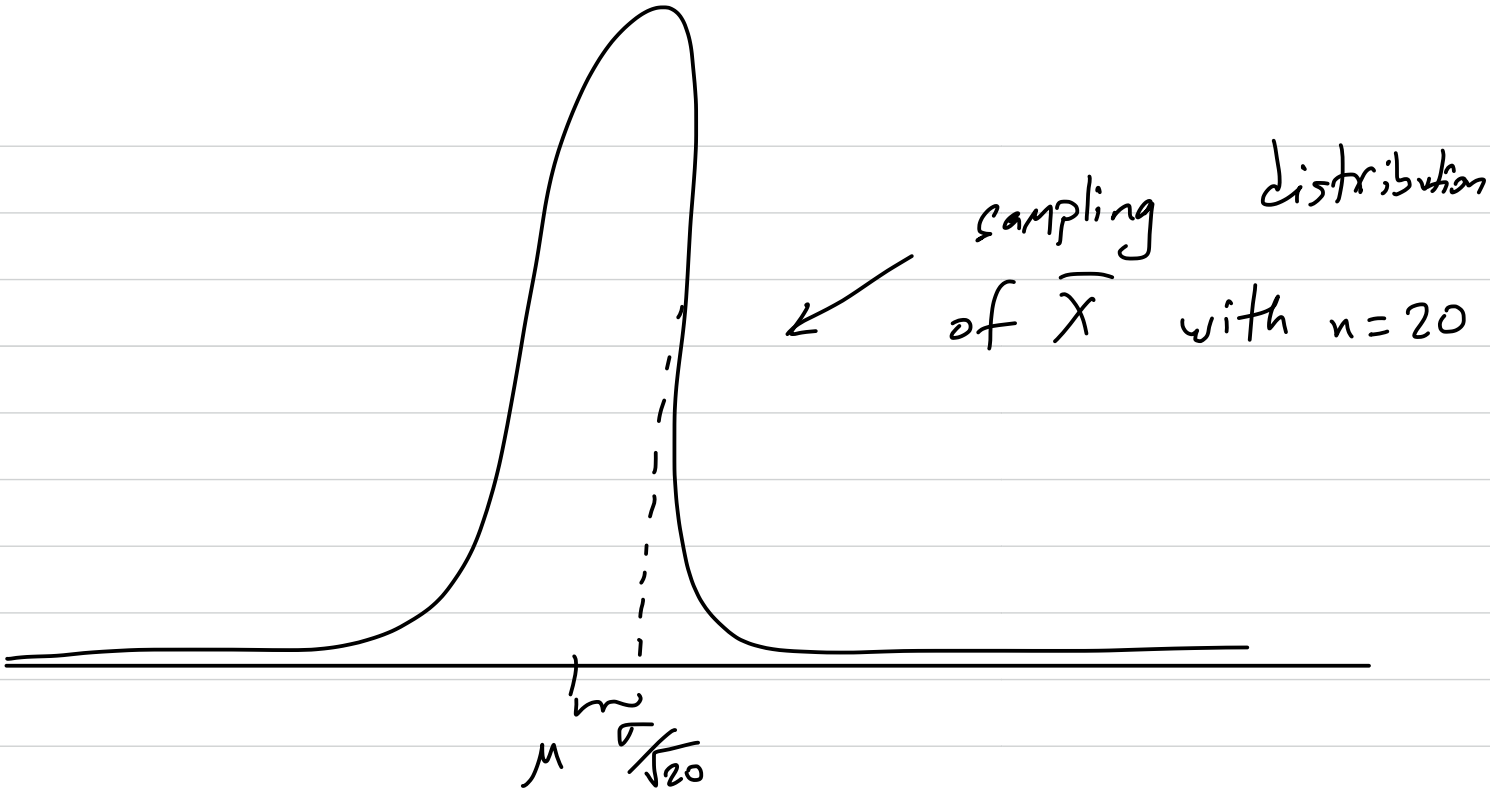
dandelions). We also have the sampling distribution of the statistic, which (if we fix 20 individuals per group) is the distribution of possible values for (in our case) $\bar{x}$.

Thm: Let $\bar{X}$ be the mean of a SRS of size n drawn from a large population has mean μ and standard deviation σ. Then the sampling distribution of $\bar{X}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Ex: Say the actual mean of the dandelion heights is 14.5 and the standard deviation is 2.5. How could we calculate that 14.5 number with samples?







Ex: We take a sample of 4 observations from a Normal distribution with mean 4.5 and standard deviation 2. What distribution does $\bar{X}$ follow?

What is the probability that a single observation in the population has a value greater than 6?

What is the probability that a sample mean is greater than 6?

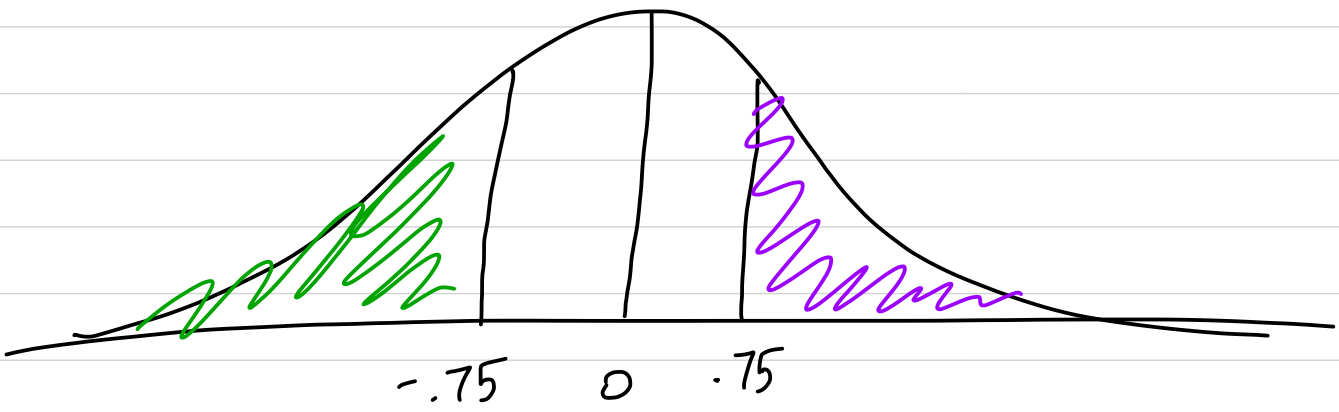
exactly Normal b/c original dist was Normal

$\overline{X}$ has a distribution of $N\left(4.5, \frac{2}{\sqrt{4}}\right)$,

so $N(4.5, 1)$.

In $N(4.5, 2)$, 6 has a z-score of

$z = \frac{6 - 4.5}{2} = .75.$



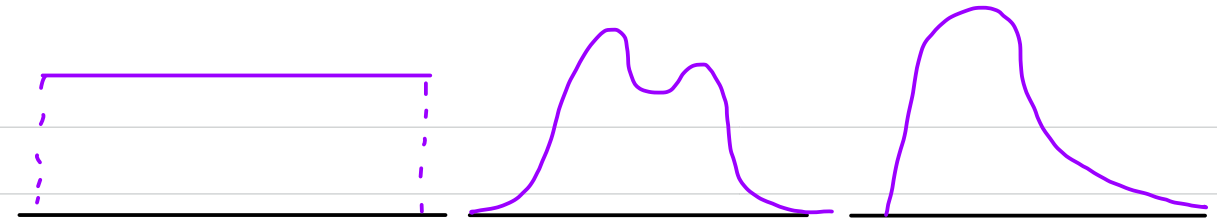
By the table, $-.75$ corresponds to .2266,

so there is a 22.66% chance that a given individual has value larger than 6.

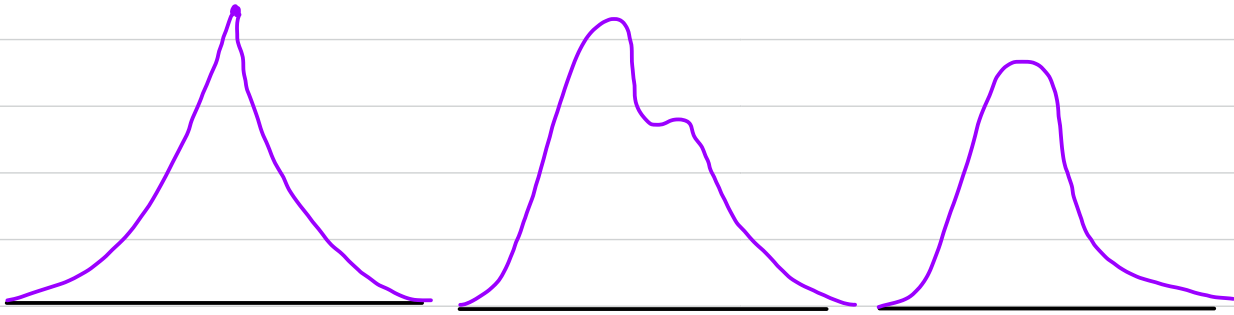
For the sampling distribution, the only change is that the standard deviation is now 1, so $z = \frac{6-4.5}{1} = 1.5$. By the table, there is only a .0668 chance that a given sample has mean larger than 6.

Thm (The Central Limit Theorem): As the sample size increases, the sampling distribution approaches a normal distribution.

Population



$\bar{X}$ w/ $n = 5$



$\bar{X}$ w/ $n = 30$

approximately Normal

Midterm: Chapters 1, 2, 3, 8, 9, 12

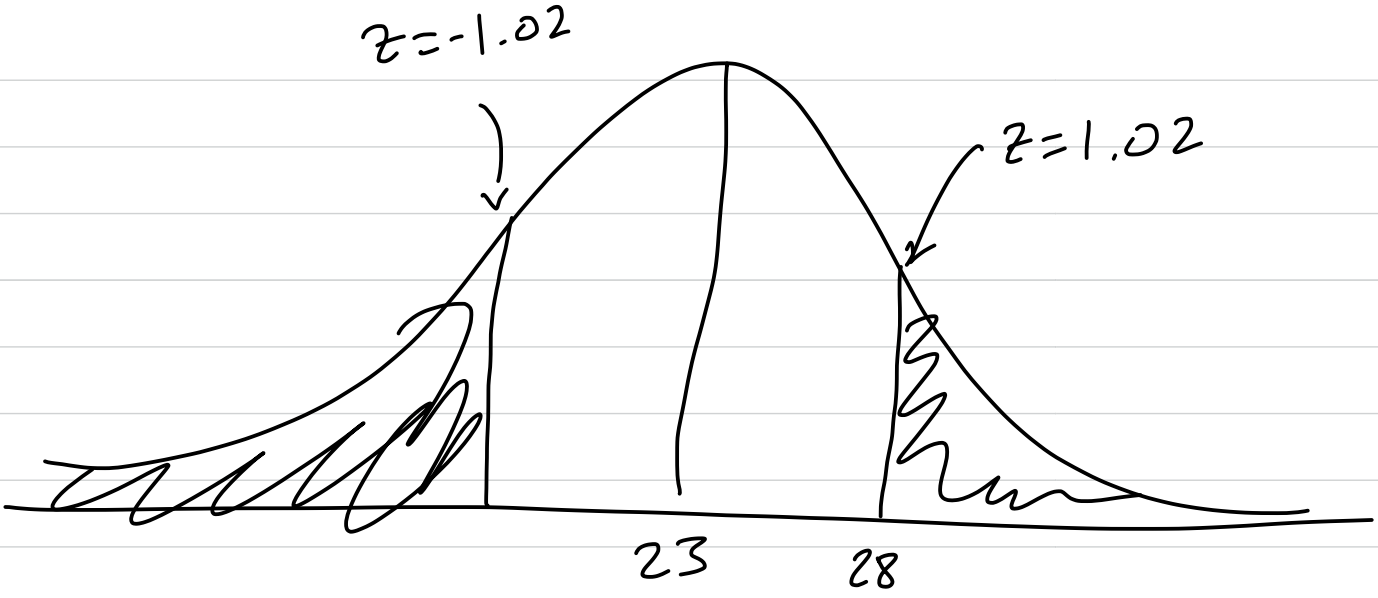
10:00 - 11:50

Open book, open notes, closed internet and other people

28 mpg

Among all cars: $N(23, 4.9)$

What % have higher than 28 mpg?



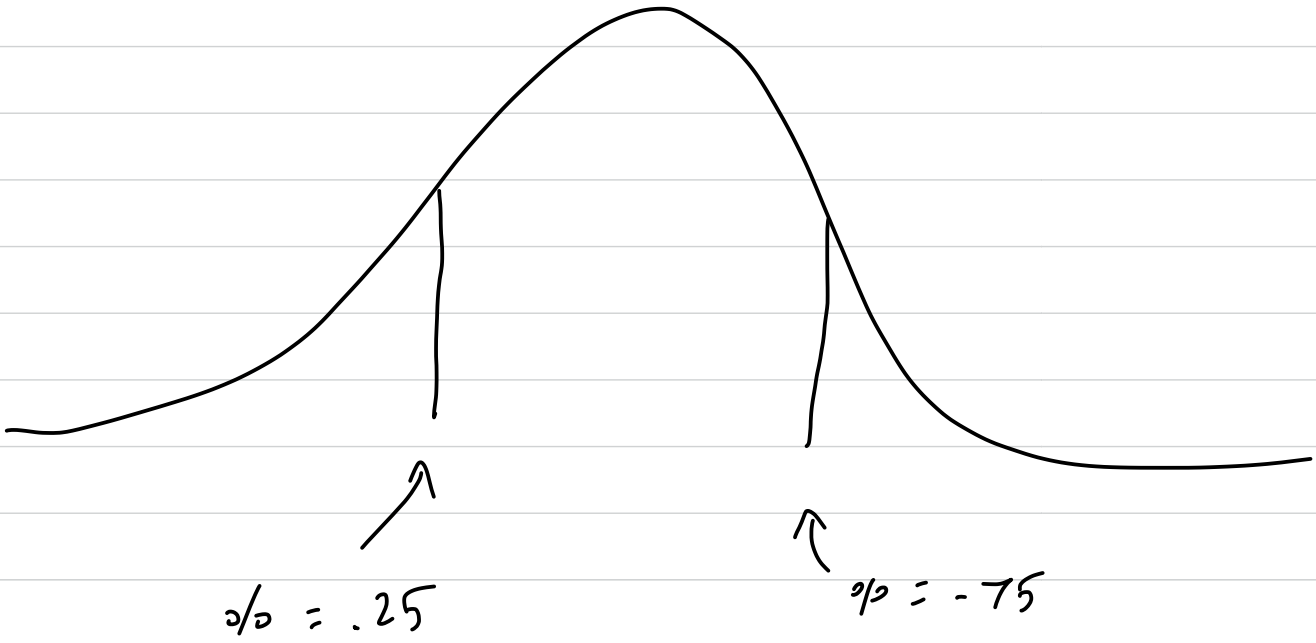
$$z = \frac{28 - 23}{4.9} = 1.02$$

$z = -1.02$ corresponds to .1539, so roughly 15.39% of all cars have mpg higher than 28.

.1492 corresponds to a z-score of -1.04, so we want 1.04

$$1.04 = \frac{x - 23}{4.9}$$

$$x = 28.096$$



.2514 occurs at a z-score of -.67,

so a z-score of .67 corresponds to

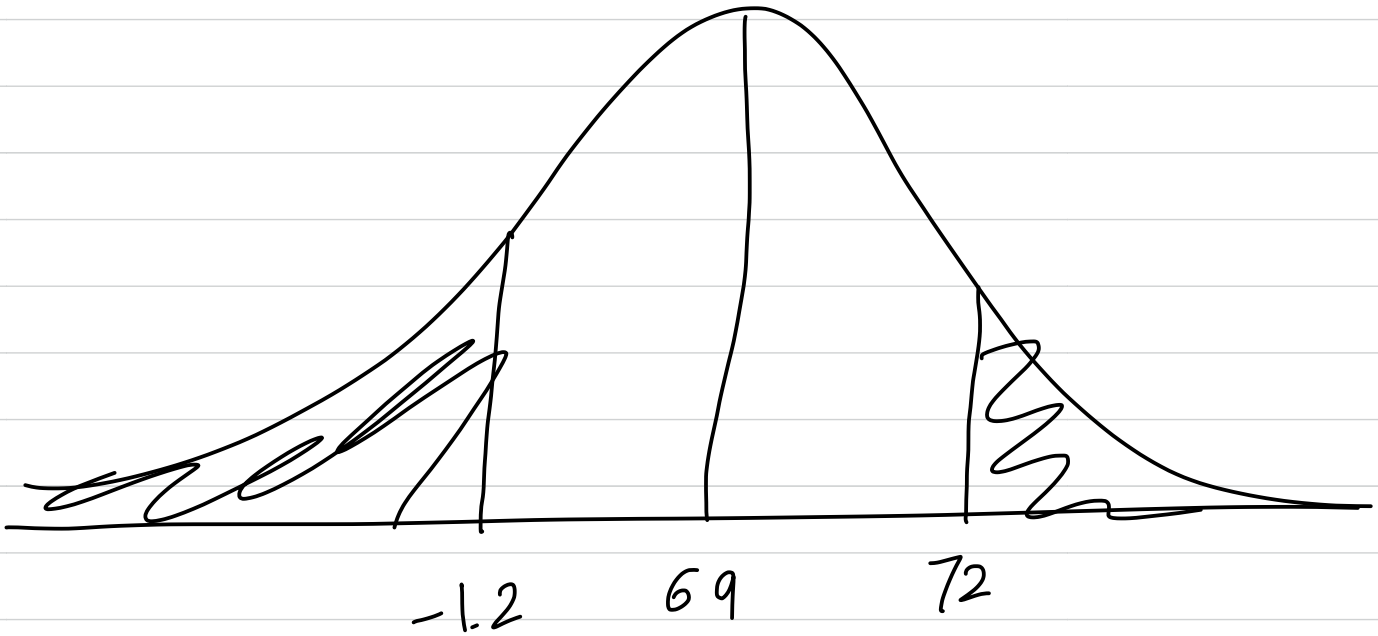
$1 - .2514 = .7486$

$$-.67 = \frac{Q_1 - 23}{4.9}$$

$19.72 = Q_1$

$.67 = \dfrac{Q_3 - 23}{4.9}$

$Q_3 = 26.28$

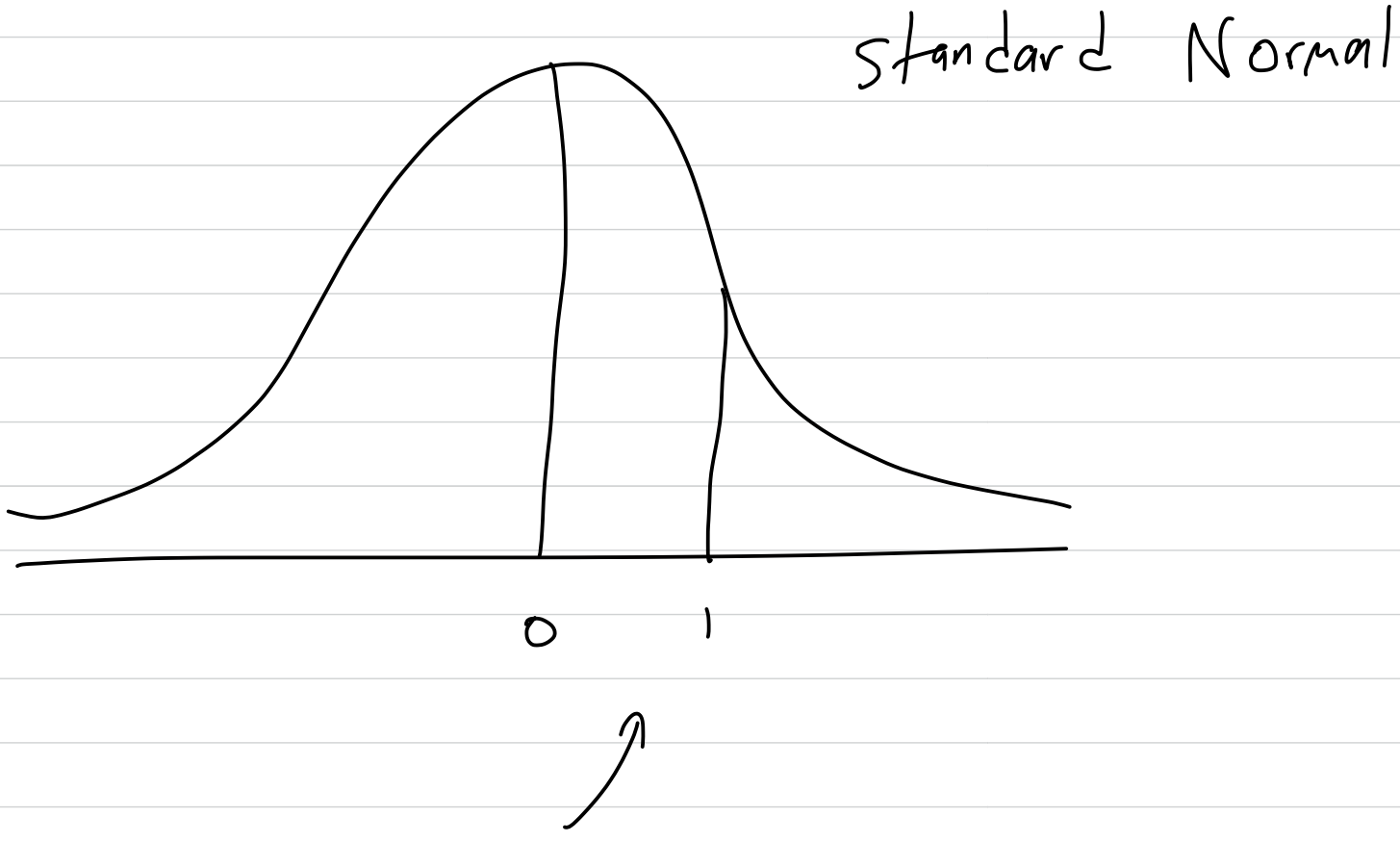


$z = \dfrac{72 - 69}{2.5} = \dfrac{3}{2.5} = 1.2$

Table: $z = -1.2 \iff .1151$

So 11.51% have height at least 6 feet.

$.2005 \iff \boxed{z = -.84}$



NCSU:			
	26%	A	4
	42%	B	3
	20%	C	2
	10%	D	1
	2%	F	0

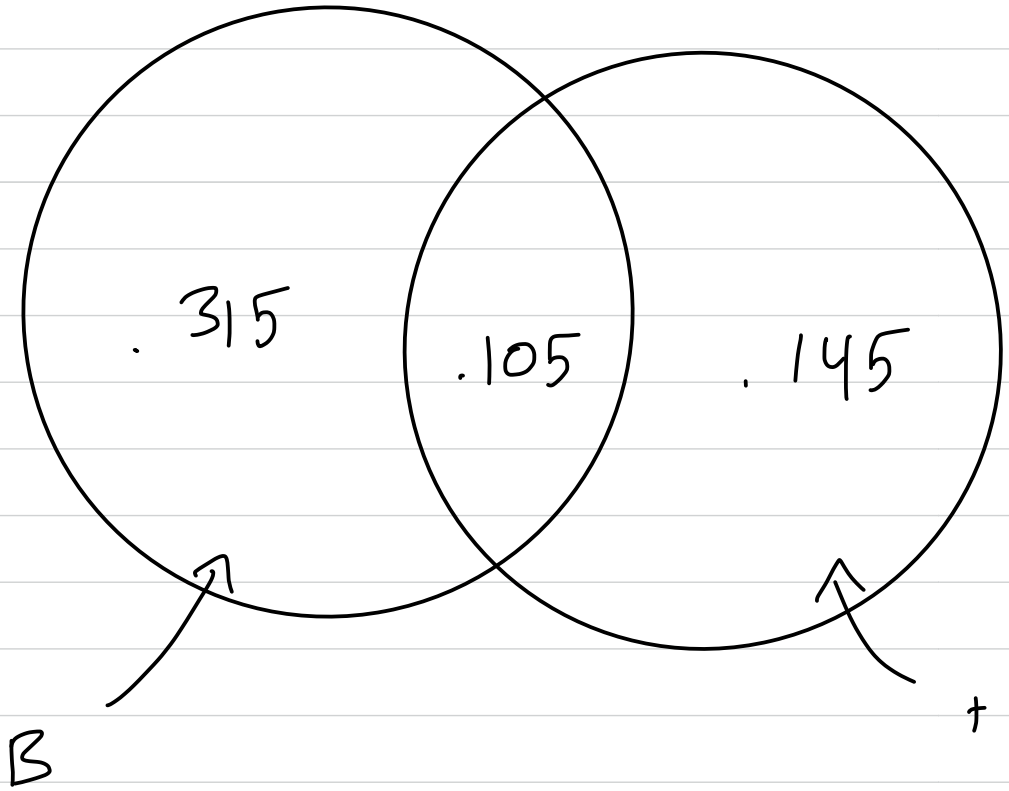
x	P(X = x)
0	.02
1	.1
2	.2
3	.42
4	.26

Probability of getting a + is 25%,

a − is 25%, and neither is 50%

Find probability of getting either a (B−, B, or B+) or (getting a +).

Assume that your letter grade has no effect on the likelihood that you get a +/−.



$$P(B+) = \underbrace{P(\text{any } B) \cdot P(\text{any } +)}_{\text{b/c independent}}$$

$$= (.42)(.25) = .105$$

$$P(\text{any } B \text{ or any } +) = .315 + .105 + .145$$

$$= .565$$