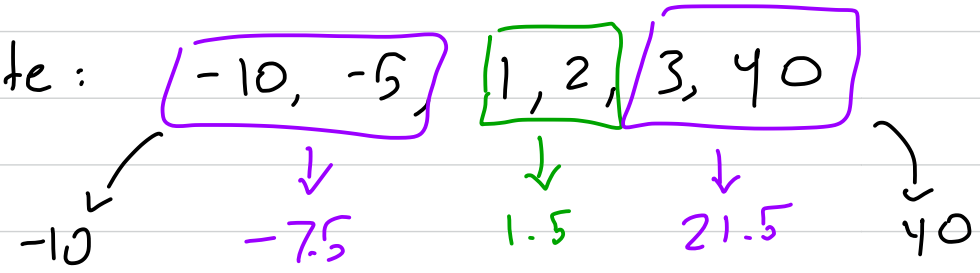
Ex: Find the 5-number summary of

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

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



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

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

$= 29$

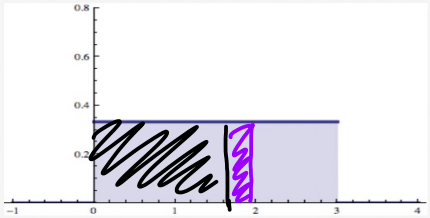
$Q_3 + 1.5\,IQR = 21.5 + 43.5 = 65$

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

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

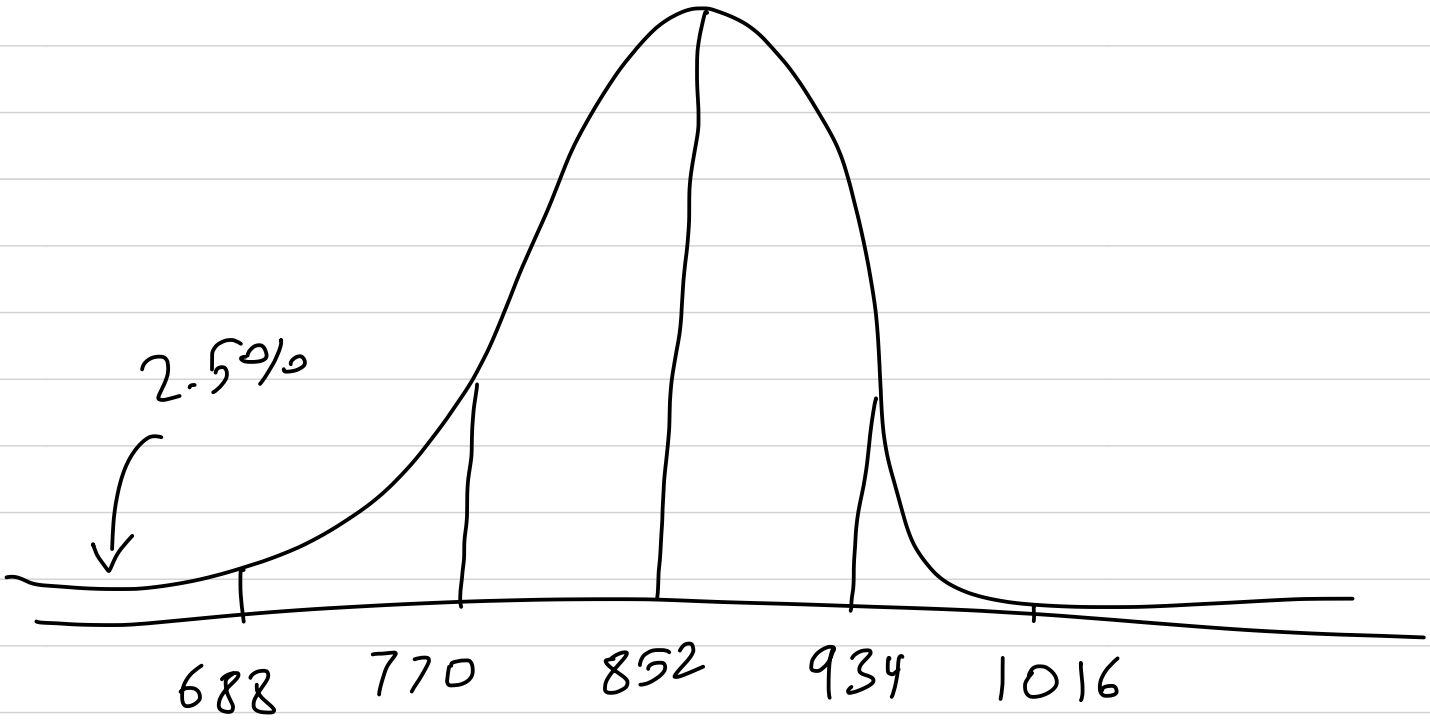
proportion before 1.7 is $\frac{1.7}{3}$

$\frac{.2}{3}$ ← 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):

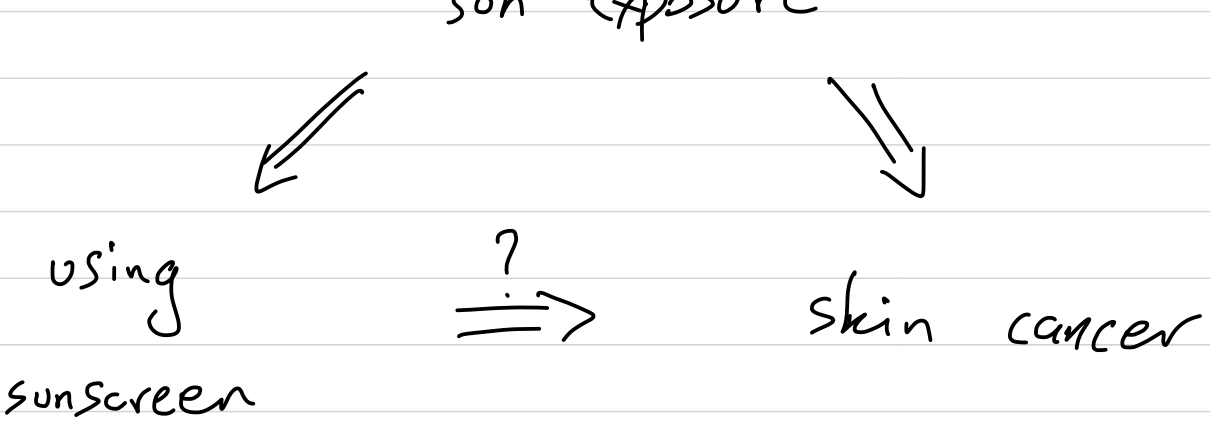
Surveys

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 $P(A)$.

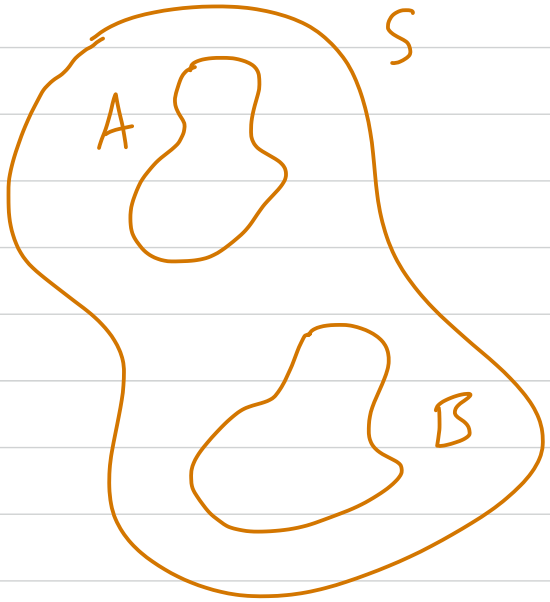
Ex: Denote heads by H and tails by T. Then $P(H) = .5$ and $P(T) = .5$.

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

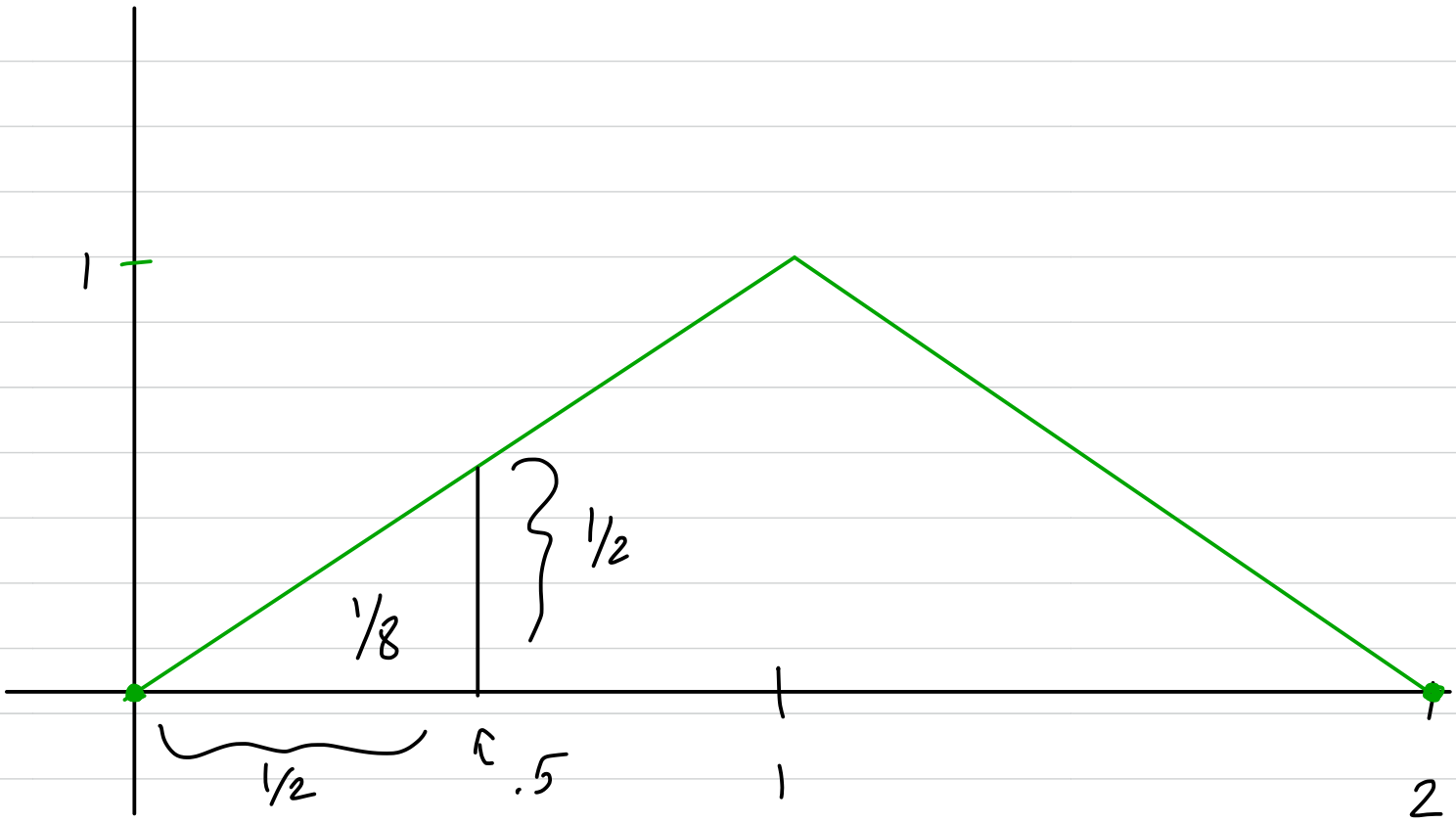
given by a density curve is called continuous.

* 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.