Stat 201: Statistics I
Week 3

Week 3
More Probability, Sampling methods and Types of Studies

Section 3.1
Conditional Probability and Bayes Theorem

Formal definition of conditional probability

Recall the multiplication rule,

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

Formal definition of conditional probability

Recall the multiplication rule,

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

From this, the formal definition of conditional probability is

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

Intuitive approach to conditional probability

An intuitive approach to $P(B \mid A)$ is to assume A has occurred, then count the instances of B. A is, in a sense, the new sample space.

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

Practice: Cancer screening

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

Practice: Cancer screening

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

- $A =$ Subject has cancer
  $B =$ Positive test result

Practice: Cancer screening

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

- $A =$ Subject has cancer
  $B =$ Positive test result
- Find $P(B \mid A)$, formally,

Practice: Cancer screening

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

- A = Subject has cancer
  B = Positive test result
- Find $P(B \mid A)$, formally,

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.074}{0.087} = 0.851$$

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

- $A =$ Subject has cancer
  $B =$ Positive test result
- Find $P(B \mid A)$, intuitive approach,

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a positive test result if the subject has cancer?

- $A =$ Subject has cancer
  $B =$ Positive test result
- Find $P(B \mid A)$, intuitive approach,

$$P(B \mid A) = \frac{\text{number of } B \text{ and } A}{\text{number of } A} = \frac{74}{87} = 0.851$$

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a negative test result if the subject does not have cancer?

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a negative test result if the subject does not have cancer?

- $A =$ Subject does not have cancer
  $B =$ Negative test result

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a negative test result if the subject does not have cancer?

- A = Subject does not have cancer
  B = Negative test result
- Find $P(B \mid A)$, formally

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.887}{0.913} = 0.972$$

Practice: Cancer screening, cont.

	Positive	Negative	Total
Cancer	74 (0.074)	13 (0.013)	87 (0.087)
No cancer	26 (0.026)	887 (0.887)	913 (0.913)

What is the probability of a negative test result if the subject does not have cancer?

- $A =$ Subject does not have cancer
  $B =$ Negative test result
- Find $P(B \mid A)$, intuitive approach,

$$P(B \mid A) = \frac{\text{number of } B \text{ and } A}{\text{number of } A} = \frac{887}{913} = 0.972$$

Sensitivity and specificity

The proceeding examples have specific terms when used with diagnostic tests.

Sensitivity and specificity

The proceeding examples have specific terms when used with diagnostic tests.

- **Sensitivity** is the probability of a positive test result for a subject which has the conditions, $P(\text{positive test} \mid \text{has disease})$.

Sensitivity and specificity

The proceeding examples have specific terms when used with diagnostic tests.

- **Sensitivity** is the probability of a positive test result for a subject which has the conditions, $P(\text{positive test} \mid \text{has disease})$.
- **Specificty** is the probability of a negative test result for a subject which does not have the conditions, $P(\text{negative test} \mid \text{does not have disease})$.

Sensitivity and specificity

The proceeding examples have specific terms when used with diagnostic tests.

- **Sensitivity** is the probability of a positive test result for a subject which has the conditions, $P(\text{positive test} \mid \text{has disease})$.
- **Specificty** is the probability of a negative test result for a subject which does not have the conditions, $P(\text{negative test} \mid \text{does not have disease})$.

Many diagnostic tests work by measuring the level of a certain chemical and returning a positive result if it is above a designated threshold. Adjusting this threshold to increase sensitivity will decrease specificity, and vice versa. There is always a trade-off.

Sensitivity and specificity, examples

Example

Screening tests for prostate cancer measure levels of Prostate Specific Antigen (PSA). The sensitivity and specificity of the test depends on the cutoff point used.

	< 4.0 ng/mL	< 3.0 ng/mL
Sensitivity (%)	21	32
Specificity (%)	91	85

Sensitivity and specificity, examples

Example

Accuracy of tests often depend also on the population being screened. The sensitivity of mammograms is different for different age groups.

	40-49 years	50-59 years
Sensitivity (%)	77	88

Screening tests for rare events

Example

Suppose there is a screening test for a rare disease which has a prevalence of 0.3%. The screening test has 99% sensitivity and 99% specificity. 100,000 people are screened.

	Positive	Negative	Total
Disease	297	3	300
No disease	997	98703	99700
Total	1294	98706	100,000

What is the probability that someone who tested positive does not have the disease?

Screening tests for rare events

Example

Suppose there is a screening test for a rare disease which has a prevalence of 0.3%. The screening test has 99% sensitivity and 99% specificity. 100,000 people are screened.

	Positive	Negative	Total
Disease	297	3	300
No disease	997	98703	99700
Total	1294	98706	100,000

What is the probability that someone who tested positive does not have the disease?

$$P(\text{no disease} \mid \text{positive}) = \frac{997}{1294} = 0.770$$

Screening tests for rare events

Example

Suppose there is a screening test for a rare disease which has a prevalence of 0.3%. The screening test has 99% sensitivity and 99% specificity. 100,000 people are screened.

	Positive	Negative	Total
Disease	297	3	300
No disease	997	98703	99700
Total	1294	98706	100,000

What is the probability that someone who tested positive does not have the disease?

$$P(\text{no disease} \mid \text{positive}) = \frac{997}{1294} = 0.770$$

- The complement, $P(\text{disease} \mid \text{positive}) = 0.23$, is known as the **precision** or the **positive predictive value (PPV)** of the test.

Screening tests for rare events, cont.

- This does not mean screening tests are not useful. Often they are a first step before tests that are more accurate, but also more expensive and/or more invasive.
  - Cancer screening, followed by biopsy for confirmation

Screening tests for rare events, cont.

- This does not mean screening tests are not useful. Often they are a first step before tests that are more accurate, but also more expensive and/or more invasive.
  - Cancer screening, followed by biopsy for confirmation
- Sometimes tests like these can have profound consequences for peoples lives.
  - Drug screening for jobs
  - Vetting for refugees or immigrants
  - etc.

Screening tests for rare events, cont.

- This does not mean screening tests are not useful. Often they are a first step before tests that are more accurate, but also more expensive and/or more invasive.
  - Cancer screening, followed by biopsy for confirmation
- Sometimes tests like these can have profound consequences for peoples lives.
  - Drug screening for jobs
  - Vetting for refugees or immigrants
  - etc.
- It is important to remember that no test is perfect and there are often trade-offs (sensitivity / specificity).

Bayes Theorem

Consider again the multiplication rule,

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

Bayes Theorem

Consider again the multiplication rule,

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

It could also be expressed with equal validity as,

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

Bayes Theorem

Consider again the multiplication rule,

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

It could also be expressed with equal validity as,

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

With some algebra,

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

Bayes Theorem

Consider again the multiplication rule,

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

It could also be expressed with equal validity as,

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

With some algebra,

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

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

Bayes Theorem, cont.

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

This equation is known as **Bayes Theorem**.

Thomas Bayes was a Presbyterian minister and amateur mathematician who lived 1701 - 1761. The early form of the theorem that bears his name was published posthumously, though it has been refined by many people since..

Bayes Theorem, example

Example

According to the Minnesota Department of Public Safety 2017 statistics, there were 78,465 motor vehicle crashes, 341 of them involving fatalities. Seat belts were used in 54.1% of the fatal crashes (in 13.6% of fatal crashes, seat belt use was unknown). Overall, the rate of seat belt use in MN was 92.0%.

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

Bayes Theorem, example

Example

According to the Minnesota Department of Public Safety 2017 statistics, there were 78,465 motor vehicle crashes, 341 of them involving fatalities. Seat belts were used in 54.1% of the fatal crashes (in 13.6% of fatal crashes, seat belt use was unknown). Overall, the rate of seat belt use in MN was 92.0%.

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

- $A =$ A crash results in fatalities. $P(A) = \dfrac{341}{78465} = 0.0043$

Bayes Theorem, example

Example

According to the Minnesota Department of Public Safety 2017 statistics, there were 78,465 motor vehicle crashes, 341 of them involving fatalities. Seat belts were used in 54.1% of the fatal crashes (in 13.6% of fatal crashes, seat belt use was unknown). Overall, the rate of seat belt use in MN was 92.0%.

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

- $A =$ A crash results in fatalities. $P(A) = \dfrac{341}{78465} = 0.0043$
- $B =$ Car occupants use seat belts. $P(B) = 0.92$

Bayes Theorem, example

Example

According to the Minnesota Department of Public Safety 2017 statistics, there were 78,465 motor vehicle crashes, 341 of them involving fatalities. Seat belts were used in 54.1% of the fatal crashes (in 13.6% of fatal crashes, seat belt use was unknown). Overall, the rate of seat belt use in MN was 92.0%.

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

- $A =$ A crash results in fatalities. $P(A) = \dfrac{341}{78465} = 0.0043$
- $B =$ Car occupants use seat belts. $P(B) = 0.92$
- $B \mid A =$ Occupants used seat belts given the crash involved fatalities. $P(B \mid A) = 0.541$

Bayes Theorem, example

Example

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

$$P(A) = 0.0043 \qquad P(B) = 0.92 \qquad P(B \mid A) = 0.541$$

Bayes Theorem, example

Example

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

$$P(A) = 0.0043 \qquad P(B) = 0.92 \qquad P(B \mid A) = 0.541$$

- Find $P(A \mid B)$

Bayes Theorem, example

Example

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

$$P(A) = 0.0043 \qquad P(B) = 0.92 \qquad P(B \mid A) = 0.541$$

- Find $P(A \mid B)$

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

Bayes Theorem, example

Example

What is the probability a motor vehicle crash with occupants wearing seat belts results in deaths?

$$P(A) = 0.0043 \qquad P(B) = 0.92 \qquad P(B \mid A) = 0.541$$

- Find $P(A \mid B)$

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

$$P(A \mid B) = \frac{0.0043 \times 0.541}{0.92} = 0.0025$$

Bayes Theorem, interpretation

There are two main ways to think about Bayes Theorem:

Bayes Theorem, interpretation

There are two main ways to think about Bayes Theorem:

- Update a probability with new information.

  If you know a car is involved in a crash, the probability it resulted in a death is 0.0043. However, if you further learn that the occupants were wearing seat belts, that probability drops to 0.0025. If you learn more information, such as the age of the driver, you could further refine the probability of fatalities.

Bayes Theorem, interpretation

There are two main ways to think about Bayes Theorem:

- Update a probability with new information.

  If you know a car is involved in a crash, the probability it resulted in a death is 0.0043. However, if you further learn that the occupants were wearing seat belts, that probability drops to 0.0025. If you learn more information, such as the age of the driver, you could further refine the probability of fatalities.

- Reverse a known conditional probability.

  If we know the probability of seat belt use given the crash involved a fatality (and the marginal probabilities of fatal crashes and seat belt use overall), we can figure out the probability of fatalities given seat belt use.

Why learn about Bayes Theorem?

- In simple cases, probabilities might be easier to calculate using tree diagrams. However, in more complicated scenarios, Bayes Theorem can become an important tool.

Why learn about Bayes Theorem?

- In simple cases, probabilities might be easier to calculate using tree diagrams. However, in more complicated scenarios, Bayes Theorem can become an important tool.
- There are two main schools of statistics. This class, and undergraduate statistics in general, utilize **frequentist** statistics. A more recent and more complicated approach is known as **bayesian** statistics, which is based, as you might expect, on Bayes Theorem.

Group work

- Complete all parts of question 1.
- Probabilities can be expressed as fractions.

Section 3.2
Sampling Methods and Types of Studies

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.
- Therefore, how we pick our sample is very important in how valid the interpretation of our results are.

Samples

- Recall, when we want to know something about a population and we can't collect data from the entire population, we can collect data from a subset, or a **sample**, of the population instead.
- We can then use statistics to learn something about the whole population.
- Therefore, how we pick our sample is very important in how valid the interpretation of our results are.

Example

- Suppose an organization is interesting in the taco consumption by Metro State students. It would be difficult, if not impossible, to ask every student about their taco eating habits. A sample is needed.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

A **simple random sample** is a sample selected such that every possible sample of a specific size has an equal chance of being selected.

- These are the “best” kind of samples for producing valid, unbiased results, but they are not always easy to get.

Types of samples: Random sample

A **random sample** is a sample selected such that every individual member of a population has an equal chance of being included.

A **simple random sample** is a sample selected such that every possible sample of a specific size has an equal chance of being selected.

- These are the "best" kind of samples for producing valid, unbiased results, but they are not always easy to get.

Example

- Given an alphabetical list of students, use a random number generator to select a sample.

Types of samples: Systematic sampling

Systematic sampling is a method where every kth member of of a population is selected.

- These samples are ofter easier to produce, but can lead to biased samples.

Types of samples: Systematic sampling

Systematic sampling is a method where every kth member of of a population is selected.

- These samples are ofter easier to produce, but can lead to biased samples.

Example

- Given an alphabetical list of students, select every fifth student until you have a sample of the desired size.

Types of samples: Convenience sampling

Convenience sampling is a method of choosing members of a population that are nearby or easy to access.

- The easiest of all methods, but by far the lowest quality data for producing results.
- On the other hand, convenience samples are sometimes the only possible sample.

Types of samples: Convenience sampling

Convenience sampling is a method of choosing members of a population that are nearby or easy to access.

- The easiest of all methods, but by far the lowest quality data for producing results.
- On the other hand, convenience samples are sometimes the only possible sample.

Example

- Wander the halls before class, asking students who happen to walk by.
- Put a poll on the Metro State website.
- Everyone who is diagnosed with a rare disease at a particular clinic.

Types of samples: Stratified sampling

Stratified sampling is a method where the population is divided into groups and samples are selected from each group.

- Useful when you want to ensure that a factor of interest has enough representation, but it is not a random sample as we have defined it.

Types of samples: Stratified sampling

Stratified sampling is a method where the population is divided into groups and samples are selected from each group.

- Useful when you want to ensure that a factor of interest has enough representation, but it is not a random sample as we have defined it.

Example

- If we have particular interest in the taco consuming difference between graduate students and undergrads, select a sample from each group.

Types of samples: Cluster sampling

Cluster sampling is a method where the population is divided into sections or clusters. Then, a number of clusters are randomly selected and all members of the clusters are included in the sample.

- More convenient than some methods, but better randomization the pure convenience sampling.

Types of samples: Cluster sampling

Cluster sampling is a method where the population is divided into sections or clusters. Then, a number of clusters are randomly selected and all members of the clusters are included in the sample.

- More convenient than some methods, but better randomization the pure convenience sampling.

Example

- Choose 5 random classes, and survey all the students in those classes.

Types of samples: Multistage sampling

Multistage sampling is a when a combination of methods are used to produce a sample.

Types of samples: Multistage sampling

Multistage sampling is a when a combination of methods are used to produce a sample.

Example

- Choose random classes by cluster sampling, and then take a simple random sample of students from each chosen class.

Types of studies

In an **observational study** data is collected from a sample without trying to modify behavior or results.

Types of studies

In an **observational study** data is collected from a sample without trying to modify behavior or results.

In an **experiment** a change (treatment) is made to some or all of sample and then data is collected in order to detect changes.

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

A **retrospective** study collects data from the past, whether from recollections or by examining records.

- Also know as: case-control

Types of observational studies

A **cross-sectional** study measures and collects data from one point in time (the present).

A **retrospective** study collects data from the past, whether from recollections or by examining records.

- Also know as: case-control

A **prospective** study follows subjects into the future to measure and collect data.

- Also know as: longitudinal study, cohort study

Experimental design: Controlling

An experiment is **controlled** when at least one group of subjects are not given any experimental treatments. The control group might receive no treatments, a placebo treatment (see blinding) or a standard-of-care treatment. Controlling an experiment allows a direct measurement of any possible treatment effects.

Example

The World Health Organization says the average case fatality rate for Ebola virus disease (EVD) is 50%, with fatality rates of individual breakouts ranging from 25% to 90%.

PREVAIL II, a controlled trial of a new drug cocktail for EVD, found a fatality rate in the control group was 37% and 22% in the treatment group.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.
- A **double blinded** study is one where both the subjects and the researchers administrating treatment and gathering results don't know which treatment the subjects are receiving.

Experimental design: Blinding

Blinding is the process of hiding which treatment or lack of treatment a subject is receiving from one or more groups of study participants. This is done in order to reduce bias in the results.

- A **single blinded** study is one where the subjects don't know what treatments they are receiving.
- A **double blinded** study is one where both the subjects and the researchers administrating treatment and gathering results don't know which treatment the subjects are receiving.

The **placebo effect** is a phenomenon where people who believe they are being treated demonstrate improvement.

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

- Experimental studies should have adequate sample sizes to ensure that observed effects are "true" effects and not due to individual characteristics or chance.

Experiment design: Replication

Replication is the repetition of the experiment on more than one individual or in more than one study.

- Experimental studies should have adequate sample sizes to ensure that observed effects are "true" effects and not due to individual characteristics or chance.
- Experimental studies should be, but rarely are, repeated by different researchers to verify results.

Experimental design: Randomization

Randomization is the process selecting samples and assigning treatment groups randomly. This is done to ensure that samples are representative of populations and that characteristics are evenly distributed among treatment groups.

Experimental design: Randomization

Randomization is the process selecting samples and assigning treatment groups randomly. This is done to ensure that samples are representative of populations and that characteristics are evenly distributed among treatment groups.

Confounding variables (or just confounders) are unmeasured and possible unknown factors that affect the experimental outcome.

Group work

- Answer all parts of question 2.