

# STOR 155 - Lecture 1

## Case Study

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

August 21, 2023

# Treating Chronic Fatigue Syndrome

- ▶ **Objective:** Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
- ▶ **Participant pool:** 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
- ▶ **Actual participants:** Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

# Study design

- ▶ Patients randomly assigned to treatment and control groups, 30 patients in each group:
  - ▶ **Treatment group:** Cognitive behavior therapy – collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
  - ▶ **Control group:** Relaxation – No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.
- ▶ 7 patients dropped out of the study later: 3 from the treatment and 4 from the control group.

## Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		<i>Good outcome</i>		Total
		Yes	No	
<i>Group</i>	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

## Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		<i>Good outcome</i>		Total
		Yes	No	
<i>Group</i>	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

## Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		<i>Good outcome</i>		Total
		Yes	No	
<i>Group</i>	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

- Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

- Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

## Understanding the results

Do the data show a “real” difference between the groups?

# Understanding the results

Do the data show a “real” difference between the groups?

- ▶ Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
- ▶ The observed difference between the two groups ( $70 - 19 = 51\%$ ) may be real, or may be due to natural variation.
- ▶ Since the difference is quite large, it is more believable that the difference is real.
- ▶ We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.



## Generalizing the results

Are the results of this study **generalizable** to all patients with chronic fatigue syndrome?

# Generalizing the results

Are the results of this study **generalizable** to all patients with chronic fatigue syndrome?

- ▶ These patients had specific characteristics and volunteered to be a part of this study.
- ▶ Therefore, maybe not representative of all patients with chronic fatigue syndrome.
- ▶ While we cannot immediately generalize the results to all patients, this first study is encouraging.
- ▶ The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.

# References

- ▶ Section 1.1 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 2

## Data Basics

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

August 23, 2023

# Classroom survey

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

- ▶ `gender`: What is your gender?
- ▶ `intro_extra`: Do you consider yourself introverted or extraverted?
- ▶ `sleep`: How many hours do you sleep at night, on average?
- ▶ `bedtime`: What time do you usually go to bed?
- ▶ `countries`: How many countries have you visited?
- ▶ `dread`: On a scale of 1-5, how much do you dread being here?

# Data matrix

Data collected on students in a statistics class on a variety of variables:

*variable*

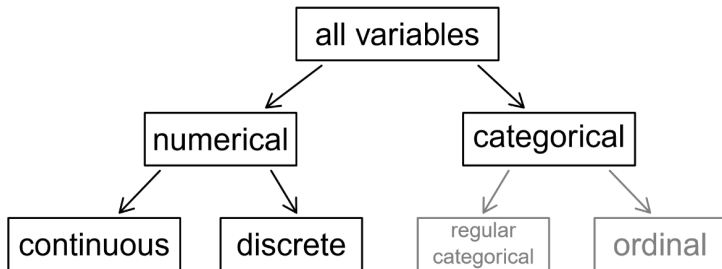
↓

Stu.	gender	intro_extra	...	dread
1	male	extravert	...	3
2	female	extravert	...	2
3	female	introvert	...	4
4	female	extravert	...	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	...	3

← *observation*

# Types of variables

- We classify variables into the following types:



## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

► gender:



## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

► gender: *categorical*

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep:

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime:

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime: *categorical, ordinal*

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime: *categorical, ordinal*
- ▶ countries:

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime: *categorical, ordinal*
- ▶ countries: *numerical, discrete*

## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime: *categorical, ordinal*
- ▶ countries: *numerical, discrete*
- ▶ dread:



## Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

- ▶ gender: *categorical*
- ▶ sleep: *numerical, continuous*
- ▶ bedtime: *categorical, ordinal*
- ▶ countries: *numerical, discrete*
- ▶ dread: *categorical, ordinal - could also be used as numerical*

# Practice

What type of variable is a telephone area code?

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical
- (d) categorical, ordinal

What type of variable is a telephone area code?

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical
- (d) categorical, ordinal

## Practice

What type of variable represents a student's grade level (e.g., freshman, sophomore, junior, senior)?

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical, nominal
- (d) categorical, ordinal

What type of variable represents a student's grade level (e.g., freshman, sophomore, junior, senior)?

(a) numerical, continuous

(b) numerical, discrete

(c) categorical, nominal

(d) categorical, ordinal

# Practice

What type of variable is the count of apples in a basket?

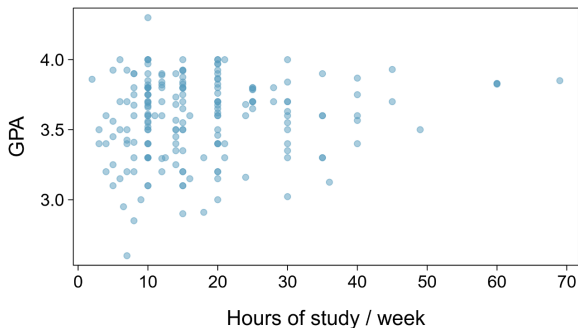
- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical, nominal
- (d) categorical, ordinal

What type of variable is the count of apples in a basket?

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical, nominal
- (d) categorical, ordinal

# Scatterplot

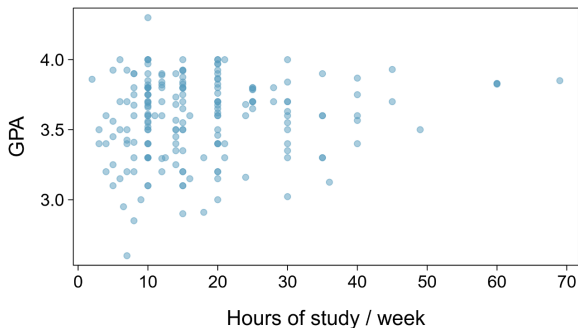
- Does there appear to be a relationship between GPA and number of hours students study per week?





# Scatterplot

- ▶ Does there appear to be a relationship between GPA and number of hours students study per week?

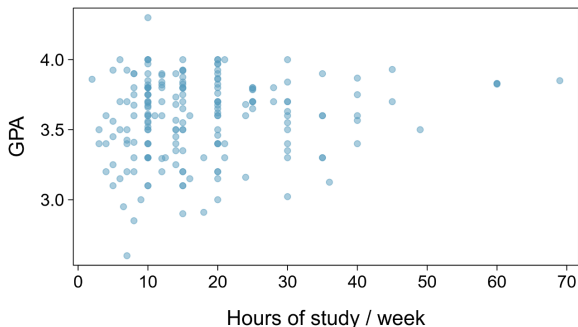


- ▶ Can you spot anything unusual about any of the data points?



# Scatterplot

- Does there appear to be a relationship between GPA and number of hours students study per week?



- Can you spot anything unusual about any of the data points?
- *There is one student with  $GPA > 4.0$ , which is not possible. This looks like a data error.*

# Explanatory and Response Variables

- ▶ **Explanatory Variables (Predictors):** These are the variables that are used to predict or explain variations in the response variable. Often referred to as independent variables.
- ▶ **Response Variable (Outcome):** This is the variable that we want to predict or explain. It is also known as the dependent variable.
- ▶ **Relationship:** The relationship between explanatory and response variables may be causal or correlative and is often expressed in a mathematical or statistical model.

# Explanatory and response variables

- ▶ To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

explanatory variable  $\xrightarrow{\text{might affect}}$  response variable

- ▶ Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables.
- ▶ We use these labels only to keep track of which variable we suspect affects the other.
- ▶ **Correlation DOES NOT imply causation!**

## Two primary types of data collection

- ▶ **Observational studies:** Collect data in a way that does not directly interfere with how the data arise (e.g., surveys).
  - ▶ Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

## Two primary types of data collection

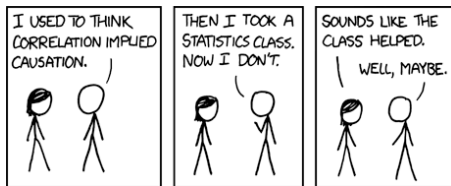
- ▶ **Observational studies:** Collect data in a way that does not directly interfere with how the data arise (e.g., surveys).
  - ▶ Can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.
- ▶ **Experiment:** Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.
  - ▶ In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

# Correlation vs. Causation

- ▶ When two variables show some connection with one another, they are called **correlated/associated** variables.
  - ▶ Associated variables can also be called **dependent** variables and vice-versa.
- ▶ If two variables are not associated, i.e., there is no evident connection between the two, then they are said to be **independent**.
- ▶ In general, correlation does not imply causation, and causation can only be inferred from a randomized experiment.

# Correlation vs. Causation

- ▶ When two variables show some connection with one another, they are called **correlated**/**associated** variables.
  - ▶ Associated variables can also be called **dependent** variables and vice-versa.
- ▶ If two variables are not associated, i.e., there is no evident connection between the two, then they are said to be **independent**.
- ▶ In general, correlation does not imply causation, and causation can only be inferred from a randomized experiment.





# Correlation

- ▶ A statistical measure that describes the extent to which two variables change together.
- ▶ If one variable tends to go up when the other goes up, there is a positive correlation.
- ▶ If one variable tends to go down when the other goes up, there is a negative correlation.

# Correlation

- ▶ A statistical measure that describes the extent to which two variables change together.
- ▶ If one variable tends to go up when the other goes up, there is a positive correlation.
- ▶ If one variable tends to go down when the other goes up, there is a negative correlation.
- ▶ *Example:* Imagine you're examining the relationship between the amount of time students spend studying for a test and their respective test scores. You might find that, generally, as the number of hours spent studying increases, test scores also increase. This would be a positive correlation.

# Causation

- ▶ A change in one variable is responsible for a change in another. Three conditions must be met:

# Causation

- ▶ A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.

# Causation

- ▶ A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.

# Causation

- ▶ A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.
  3. There aren't any other factors that could explain the relationship.

# Causation

- ▶ A change in one variable is responsible for a change in another. Three conditions must be met:
  1. There's a correlation between the two variables.
  2. The cause happens before the effect.
  3. There aren't any other factors that could explain the relationship.
- ▶ *Example:* Taking a painkiller (like ibuprofen) to reduce a headache. If headaches consistently diminish or disappear after taking the pill, and no other factors are at play (e.g., you didn't also rest in a dark room), then we can infer a causal relationship between the painkiller and the reduction of headache pain.

# Correlation $\nRightarrow$ Causation

## Example

- ▶ Ice Cream Sales and Drowning Incidents.
- ▶ As ice cream sales increase, the number of drowning incidents also tends to increase.
- ▶ Does this mean buying more ice cream causes more drownings?



# Correlation $\nRightarrow$ Causation

## Example

- ▶ Ice Cream Sales and Drowning Incidents.
- ▶ As ice cream sales increase, the number of drowning incidents also tends to increase.
- ▶ Does this mean buying more ice cream causes more drownings?
- ▶ Of course not! Potential reason?

# Correlation $\nRightarrow$ Causation

## Example

- ▶ Ice Cream Sales and Drowning Incidents.
- ▶ As ice cream sales increase, the number of drowning incidents also tends to increase.
- ▶ Does this mean buying more ice cream causes more drownings?
- ▶ Of course not! Potential reason?
- ▶ The **confounding variable** of temperature.
- ▶ During the summer, when it's hot, people are more likely to buy ice cream and also more likely to go swimming, which increases the chances of drowning.

# References

- ▶ Section 1.2 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 3

## Sampling Principles and Strategies

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

August 25, 2023

## Anecdotal evidence and early smoking research

- ▶ Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.

## Anecdotal evidence and early smoking research

- ▶ Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- ▶ Anti-smoking research was faced with resistance based on anecdotal evidence such as “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that might not be representative of the population.

## Anecdotal evidence and early smoking research

- ▶ Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- ▶ Anti-smoking research was faced with resistance based on anecdotal evidence such as “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that might not be representative of the population.
- ▶ It was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”

## Anecdotal evidence and early smoking research

- ▶ Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
- ▶ Anti-smoking research was faced with resistance based on **anecdotal evidence** such as “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that might not be representative of the population.
- ▶ It was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”
- ▶ In time, researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.



# Census

- ▶ Wouldn't it be better to just include everyone and “sample” the entire population?
  - ▶ This is called a **census**.

# Census

- ▶ Wouldn't it be better to just include everyone and “sample” the entire population?
  - ▶ This is called a **census**.
- ▶ There are problems with taking a census:

# Census

- ▶ Wouldn't it be better to just include everyone and “sample” the entire population?
  - ▶ This is called a **census**.
- ▶ There are problems with taking a census:
  1. It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*

# Census

- ▶ Wouldn't it be better to just include everyone and “sample” the entire population?
  - ▶ This is called a **census**.
- ▶ There are problems with taking a census:
  1. It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  2. Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.

# Census

- ▶ Wouldn't it be better to just include everyone and “sample” the entire population?
  - ▶ This is called a **census**.
- ▶ There are problems with taking a census:
  1. It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  2. Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
  3. Taking a census may be more complex than sampling.

# Exploratory analysis to inference

- ▶ Sampling is natural.

## Exploratory analysis to inference

- ▶ Sampling is natural.
- ▶ Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.

# Exploratory analysis to inference

- ▶ Sampling is natural.
- ▶ Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- ▶ When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's **exploratory analysis**.



# Exploratory analysis to inference

- ▶ Sampling is natural.
- ▶ Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- ▶ When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's **exploratory analysis**.
- ▶ If you generalize and conclude that your entire soup needs salt, that's an **inference**.

# Exploratory analysis to inference

- ▶ Sampling is natural.
- ▶ Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- ▶ When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's **exploratory analysis**.
- ▶ If you generalize and conclude that your entire soup needs salt, that's an **inference**.
- ▶ For your inference to be valid, the spoonful you tasted (the sample) needs to be **representative** of the entire pot (the population).
  - ▶ If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - ▶ If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

# Sampling bias

- ▶ **Non-response:** If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
  - ▶ *Example:* A sample of students in the class are given a survey form but only one or two students fill it up.

# Sampling bias

- ▶ **Non-response:** If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
  - ▶ *Example:* A sample of students in the class are given a survey form but only one or two students fill it up.
- ▶ **Voluntary response:** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.
  - ▶ *Example:* A survey asks whether the students are enjoying STOR 155. Those who don't like it are more eager to complain, and fill up the survey, than those who like the course.

# Sampling bias

- ▶ **Non-response:** If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
  - ▶ *Example:* A sample of students in the class are given a survey form but only one or two students fill it up.
- ▶ **Voluntary response:** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.
  - ▶ *Example:* A survey asks whether the students are enjoying STOR 155. Those who don't like it are more eager to complain, and fill up the survey, than those who like the course.
- ▶ **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample.
  - ▶ *Example:* A survey has a question: "Do you like answering surveys?"

# Observational studies

- ▶ Researchers collect data in a way that does not directly interfere with how the data arise (e.g., questionnaire).

# Observational studies

- ▶ Researchers collect data in a way that does not directly interfere with how the data arise (e.g., questionnaire).
- ▶ Results of an observational study can generally be used to establish an **association** between the explanatory and response variables.

# Observational studies

- ▶ Researchers collect data in a way that does not directly interfere with how the data arise (e.g., questionnaire).
- ▶ Results of an observational study can generally be used to establish an **association** between the explanatory and response variables.
- ▶ They **cannot** be used to establish **causation**.



# Obtaining good samples

- ▶ Almost all statistical methods are based on the notion of implied **randomness**.

# Obtaining good samples

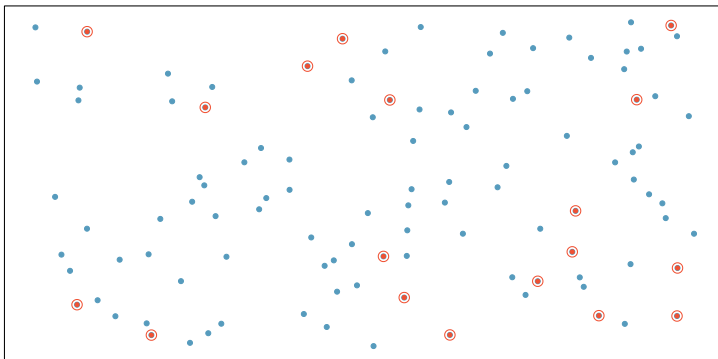
- ▶ Almost all statistical methods are based on the notion of implied **randomness**.
- ▶ If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.

# Obtaining good samples

- ▶ Almost all statistical methods are based on the notion of implied **randomness**.
- ▶ If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
- ▶ Most commonly used random sampling techniques are **simple**, **stratified**, **cluster**, and **multistage** sampling.

# Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



## Example

- ▶ Objective: study the salaries of MLB players.

## Example

- ▶ Objective: study the salaries of MLB players.
- ▶ Each player is a member of one of the league's 30 teams.

## Example

- ▶ Objective: study the salaries of MLB players.
- ▶ Each player is a member of one of the league's 30 teams.
- ▶ Suppose we want to take a random sample of 120 players.

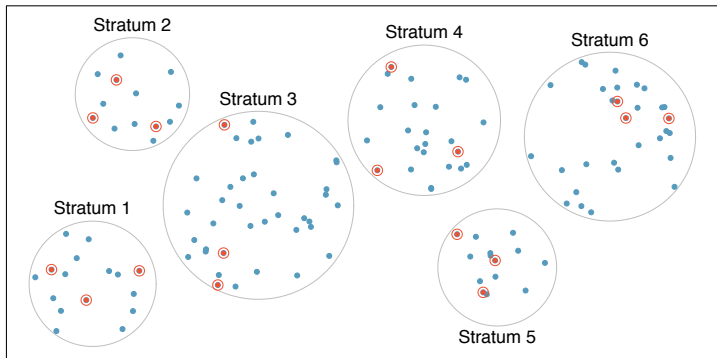
## Example

- ▶ Objective: study the salaries of MLB players.
- ▶ Each player is a member of one of the league's 30 teams.
- ▶ Suppose we want to take a random sample of 120 players.
- ▶ Execution: write all the players' names on slips of paper, mix them inside a bucket, and select 120 of them.



# Stratified sample

**Strata** are made up of similar observations. We take a simple random sample from each stratum.



## Example

- ▶ Back to the MLB players example.
- ▶ Goal: sample 120 players.

## Example

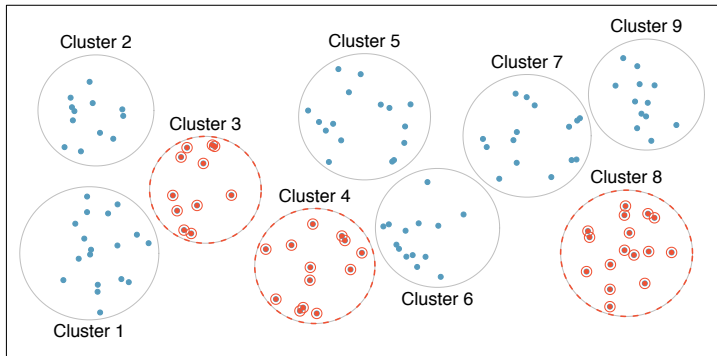
- ▶ Back to the MLB players example.
- ▶ Goal: sample 120 players.
- ▶ Insight: teams might differ a lot in terms of compensation, so it makes sense to use different teams as different strata.

## Example

- ▶ Back to the MLB players example.
- ▶ Goal: sample 120 players.
- ▶ Insight: teams might differ a lot in terms of compensation, so it makes sense to use different teams as different strata.
- ▶ Stratified sampling: split the players population naturally in 30 teams, and then sample 4 players from each team.

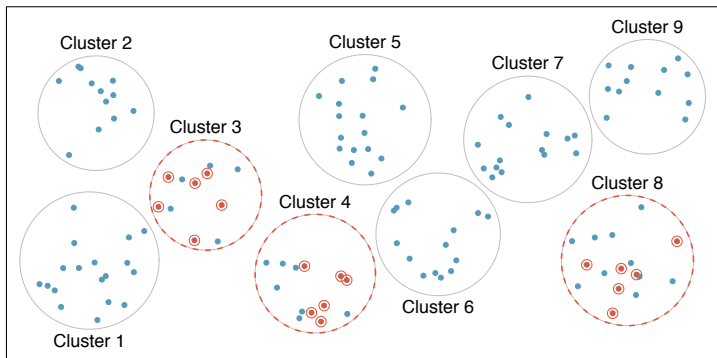
# Cluster sample

**Clusters** are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then sample all observations in that cluster. Usually preferred for economical reasons.



# Multistage sample

**Clusters** are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.



# References

- ▶ Section 1.3 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 4

## Experiments

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 1, 2023



# Experiments

- ▶ Studies where the researchers assign treatments to cases are called **experiments**.

# Experiments

- ▶ Studies where the researchers assign treatments to cases are called **experiments**.
- ▶ When this assignment includes randomization, e.g., using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**.

# Experiments

- ▶ Studies where the researchers assign treatments to cases are called **experiments**.
- ▶ When this assignment includes randomization, e.g., using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**.
- ▶ Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

# Principles of experimental design

Randomized experiments are generally built on four principles:

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.

# Principles of experimental design

Randomized experiments are generally built on four principles:

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.
2. **Randomize:** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.

# Principles of experimental design

Randomized experiments are generally built on four principles:

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.
2. **Randomize:** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. **Replicate:** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.

# Principles of experimental design

Randomized experiments are generally built on four principles:

1. **Control:** Control for the (potential) effect of variables other than the ones directly being studied.
2. **Randomize:** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. **Replicate:** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. **Block:** If there are variables that are known or suspected to affect the response variable, first group subjects into **blocks** based on these variables, and then randomize cases within each block to treatment groups.

# Controlling

- ▶ Researchers assign treatments to cases, and they do their best to control any other differences in the groups.



# Controlling

- ▶ Researchers assign treatments to cases, and they do their best to control any other differences in the groups.
- ▶ *Example:* When patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

# Randomization

- ▶ Researchers randomize patients into treatment groups to account for variables that cannot be controlled.

# Randomization

- ▶ Researchers randomize patients into treatment groups to account for variables that cannot be controlled.
- ▶ *Example:* Some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

# Replication

- ▶ The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response.

# Replication

- ▶ The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response.
- ▶ In a single study, we replicate by collecting a sufficiently large sample.

# Replication

- ▶ The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response.
- ▶ In a single study, we replicate by collecting a sufficiently large sample.
- ▶ Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

# Blocking

- ▶ Sometimes, variables other than the explanatory ones might affect the response. These are called **blocking variables**.

# Blocking

- ▶ Sometimes, variables other than the explanatory ones might affect the response. These are called **blocking variables**.
- ▶ To account for them, we might first group the subjects into **blocks** and then randomize cases within each block to the treatment groups. This process is called **blocking**.



# Blocking

- ▶ Sometimes, variables other than the explanatory ones might affect the response. These are called **blocking variables**.
- ▶ To account for them, we might first group the subjects into **blocks** and then randomize cases within each block to the treatment groups. This process is called **blocking**.
- ▶ *Example:* If we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group.

## Blocking - Example

- ▶ We would like to design an experiment to investigate if energy gels makes you run faster:

## Blocking - Example

- ▶ We would like to design an experiment to investigate if energy gels makes you run faster:
  - ▶ Treatment: energy gel
  - ▶ Control: no energy gel

## Blocking - Example

- ▶ We would like to design an experiment to investigate if energy gels makes you run faster:
  - ▶ Treatment: energy gel
  - ▶ Control: no energy gel
- ▶ It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:

## Blocking - Example

- ▶ We would like to design an experiment to investigate if energy gels makes you run faster:
  - ▶ Treatment: energy gel
  - ▶ Control: no energy gel
- ▶ It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - ▶ Divide the sample to pro and amateur
  - ▶ Randomly assign pro athletes to treatment and control groups
  - ▶ Randomly assign amateur athletes to treatment and control groups
  - ▶ Pro/amateur status is equally represented in the resulting treatment and control groups

# Difference between blocking and explanatory variables

- ▶ **Factors** are conditions we can impose on the experimental units.

# Difference between blocking and explanatory variables

- ▶ **Factors** are conditions we can impose on the experimental units.
- ▶ **Blocking variables** are characteristics that the experimental units come with, that we would like to control for.

# Difference between blocking and explanatory variables

- ▶ **Factors** are conditions we can impose on the experimental units.
- ▶ **Blocking variables** are characteristics that the experimental units come with, that we would like to control for.
- ▶ Blocking is like stratifying, except used in experimental settings when **randomly assigning**, as opposed to when **randomly sampling**.



# Experimental Design terminology

- ▶ **Placebo:** fake treatment, often used as the control group for medical studies.

# Experimental Design terminology

- ▶ **Placebo:** fake treatment, often used as the control group for medical studies.
- ▶ **Placebo effect:** experimental units showing improvement simply because they believe they are receiving a special treatment.

# Experimental Design terminology

- ▶ **Placebo:** fake treatment, often used as the control group for medical studies.
- ▶ **Placebo effect:** experimental units showing improvement simply because they believe they are receiving a special treatment.
- ▶ **Blinding:** when experimental units do not know whether they are in the control or treatment group.

# Experimental Design terminology

- ▶ **Placebo:** fake treatment, often used as the control group for medical studies.
- ▶ **Placebo effect:** experimental units showing improvement simply because they believe they are receiving a special treatment.
- ▶ **Blinding:** when experimental units do not know whether they are in the control or treatment group.
- ▶ **Double-blind:** when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group.

# Random Assignment vs. Random Sampling

<i>ideal experiment</i>	Random assignment	No random assignment	<i>most observational studies</i>
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
<i>most experiments</i>	Causation	Correlation	<i>bad observational studies</i>

# References

- ▶ Section 1.4 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 5

## Numerical Data: centrality and dispersion measures

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 6, 2023

# Centrality and Dispersion Measures

- ▶ **Centrality measures** are important statistical tools to analyze numerical data. They provide insight into the “central” or typical values of datasets.
- ▶ Common centrality measures: **mean, median, mode**
- ▶ **Measures of dispersion** are non-negative real numbers that help to gauge the spread of data about a central value. These measures help to determine how stretched or squeezed the given data is.
- ▶ Common measures of dispersion: **range, variance, standard deviation**



# Mean

- ▶ The **mean**, often referred to as the **average**, is calculated by summing all values in the dataset and dividing by the total number of observations.
- ▶ It is highly sensitive to outliers.
- ▶ The mean is an appropriate measure of central tendency when the data is roughly symmetrically distributed.

- ▶ The **sample mean**, denoted as  $\bar{x}$ , can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where  $x_1, x_2, \cdots, x_n$  represent the  $n$  observed values.

- ▶ The **population mean** is also computed the same way but is denoted as  $\mu$ . It is often not possible to calculate  $\mu$  since population data are rarely available.
- ▶ The sample mean is a **sample statistic**, and serves as a **point estimate** of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

# Median

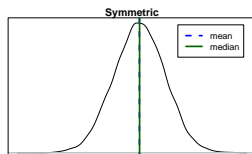
- ▶ The **median** is the value that separates the higher half from the lower half of a dataset.
- ▶ If the dataset has an odd number of observations, the median is the middle number. If it's even, the median is the average of the two middle numbers.
- ▶ Unlike the mean, the median is not affected by extreme values (outliers), i.e., it is a **robust** statistic.

To *compute* the median:

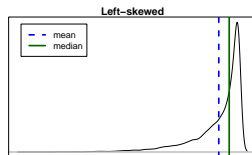
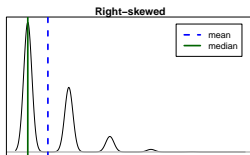
1. Order the numbers in increasing order.
2. If the multitude of numbers is odd, then the median is the number in the middle.
3. If the multitude of numbers is even, then the median is the average of the two middle numbers.
4. *Example:* The median of the 8 numbers 2, 5, 3, 8, 1, 6, 3, 2 can be found by ordering them first. The ordered numbers are 1, 2, 2, 3, 3, 5, 6, 8. There is an even (8) number of numbers, hence the median is the average of the two middle numbers 3 and 3, i.e.,  $(3 + 3)/2 = 3$ .

# Mean vs. median

- ▶ If the distribution is **symmetric**, the center is often defined as the mean:  $\text{mean} \approx \text{median}$



- ▶ If the distribution is **skewed** or has extreme outliers, center is often defined as the median
  - ▶ Right-skewed:  $\text{mean} > \text{median}$
  - ▶ Left-skewed:  $\text{mean} < \text{median}$



# Mean vs. Median: A Comparison

- ▶ **Sensitivity to Outliers:**

- ▶ Mean: Highly sensitive
- ▶ Median: Not sensitive (Robust)

- ▶ **Applicability:**

- ▶ Mean: Best for symmetric distributions
- ▶ Median: Useful for skewed distributions

- ▶ **Calculation Complexity:**

- ▶ Mean: Requires all data points
- ▶ Median: If the data is ordered, not all points are required

- ▶ **Representation:**

- ▶ Mean: Might not be an actual data point
- ▶ Median: Always an actual data point for odd sample sizes

## Q1 and Q3

- ▶ The 25<sup>th</sup> percentile is also called the first quartile, **Q1**. It is computed as the median of the ordered data to the left of the median.
- ▶ The 50<sup>th</sup> percentile is also called the median.
- ▶ The 75<sup>th</sup> percentile is also called the third quartile, **Q3**. It is computed as the median of the ordered data to the right of the median.

## Example: Odd sample size

Data: 2, 4, 5, 7, 10

- ▶ Sample size = 5 (odd)
- ▶ Median ( $Q_2$ ) = 5
- ▶ Lower half: 2, 4
- ▶ Upper half: 7, 10
- ▶  $Q_1 = \text{Median of Lower half} = \frac{2+4}{2} = 3$
- ▶  $Q_3 = \text{Median of Upper half} = \frac{7+10}{2} = 8.5$



## Example: Even sample size

Data: 1, 2, 4, 5, 7, 10

- ▶ Sample size = 6 (even)
- ▶ Median ( $Q_2$ ) = 4.5
- ▶ Lower half: 1, 2, 4
- ▶ Upper half: 5, 7, 10
- ▶  $Q_1$  = Median of Lower half = 2
- ▶  $Q_3$  = Median of Upper half = 7

# Mode

- ▶ The **mode** is the value that appears most frequently in a dataset.
- ▶ A dataset may have one mode (unimodal), more than one mode (multimodal), or no mode at all.
- ▶ The mode can be used for both numerical and categorical data.
- ▶ *Example:* The mode of  $\{4, 2, 2, 3, 5, 4, 6, 2\}$  is 2, because it occurs three times, which is more than any other number in the dataset.

# Range

- ▶ The range is the difference between the largest and smallest values in the dataset. It provides a measure of total spread.
- ▶ Mathematically, the range of  $n$  numbers  $x_1, \dots, x_n$  is defined to be the statistic

$$R = \max\{x_1, \dots, x_n\} - \min\{x_1, \dots, x_n\}.$$

- ▶ *Example:* The range of the 8 numbers 2, 5, 3, 8, 1, 6, 3, 2 is  $8 - 1 = 7$ .
- ▶ Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the **interquartile range**, or the **IQR**.

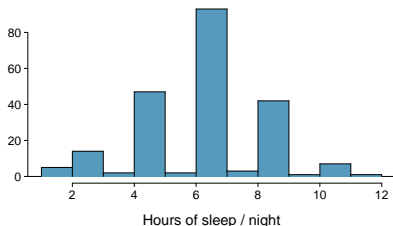
$$IQR = Q3 - Q1$$

# Variance

**Variance** is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- ▶ The sample mean is  $\bar{x} = 6.71$ , and the sample size is  $n = 217$ .
- ▶ The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5 - 6.71)^2 + (9 - 6.71)^2 + \dots + (7 - 6.71)^2}{217 - 1} = 4.11 \text{ hours}^2$$

## Different Idea

- ▶ Why use squares and not just take the average of the differences from the mean?
- ▶ In that case, we would always get

$$\begin{aligned}\frac{(x_1 - \bar{x}) + \cdots + (x_n - \bar{x})}{n} &= \frac{(x_1 + \cdots + x_n) - n\bar{x}}{n} \\ &= \frac{x_1 + \cdots + x_n}{n} - \bar{x} = 0,\end{aligned}$$

which is not informative.

# Standard deviation

The **standard deviation** is the square root of the variance, and has the same units as the data:

$$s = \sqrt{s^2}$$

- ▶ The standard deviation of the amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$

- ▶ We can see that all of the data are within 3 standard deviations of the mean.
- ▶ Same unit of measurement as that of the initial data.

# References

- ▶ Section 2.1 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 6

## Data Visualization

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 8, 2023



# Numerical vs. Categorical Data: What's the Difference?

## ▶ Numerical Data

- ▶ *Quantities*: Things you can **measure** or count.
- ▶ Examples: Age, height, salary, temperature.
- ▶ Operations: You can add, subtract, find averages, etc.

# Numerical vs. Categorical Data: What's the Difference?

## ▶ Numerical Data

- ▶ *Quantities*: Things you can **measure** or count.
- ▶ Examples: Age, height, salary, temperature.
- ▶ Operations: You can add, subtract, find averages, etc.

## ▶ Categorical Data

- ▶ *Qualities*: Things you can **categorize** but not measure.
- ▶ Examples: Colors, gender, types of cuisine, yes/no answers.
- ▶ Operations: You can sort, group, but can't perform arithmetic.

# Numerical vs. Categorical Data: What's the Difference?

## ▶ Numerical Data

- ▶ *Quantities*: Things you can **measure** or count.
- ▶ Examples: Age, height, salary, temperature.
- ▶ Operations: You can add, subtract, find averages, etc.

## ▶ Categorical Data

- ▶ *Qualities*: Things you can **categorize** but not measure.
- ▶ Examples: Colors, gender, types of cuisine, yes/no answers.
- ▶ Operations: You can sort, group, but can't perform arithmetic.

Overall:

- ▶ *Numerical* is about *numbers you can calculate with*
- ▶ *Categorical* is about *categories you can sort into*.

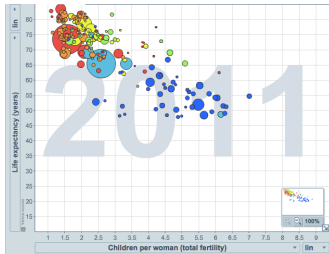
# Visualization of Numerical Data

Four main visualization objects:

- ▶ Scatterplot
- ▶ Dot plot
- ▶ Histogram
- ▶ Box-plot

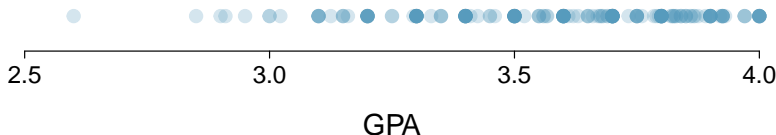
# Scatterplot

- ▶ Useful for visualizing the relationship between **two** numerical variables.
- ▶ They provide information on positive/negative association.
- ▶ *Example:* Fertility and life expectancy appear to be *linearly* and *negatively associated*: as fertility increases, life expectancy decreases.



## Dot plots

- ▶ Useful for visualizing **one** numerical variable. Darker colors represent areas where there are more observations.
- ▶ Darker colors represent areas where there are more observations.
- ▶ *Example:* Below is a dot plot of the GPA of the students in a class. We can see that most of the students have GPA higher than 3.5, quite some between 3 and 3.5, very few below 3, no one below 2.5, etc.

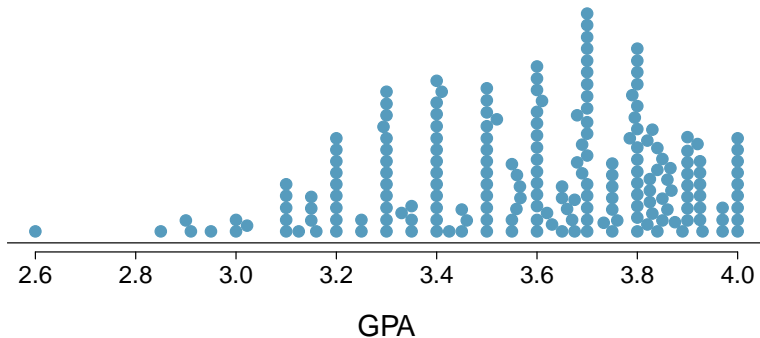


- ▶ Sometimes we indicate the mean of the data in the dot plot to give an idea about the center of the distribution.
- ▶ *Example:* The previous GPA dot plot with the mean (around 3.6).



# Stacked Dot Plot

- ▶ Higher bars represent areas where there are more observations.
- ▶ Makes it a little easier to judge the center and the shape of the distribution.



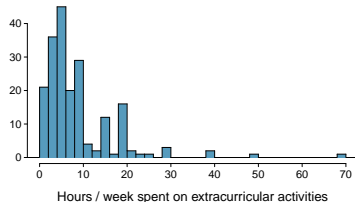
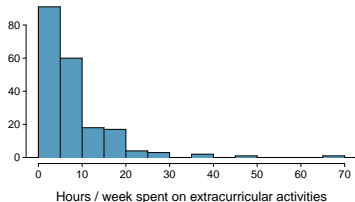
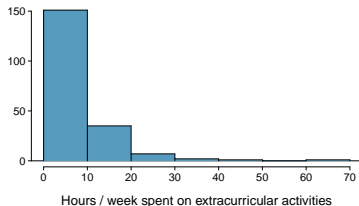
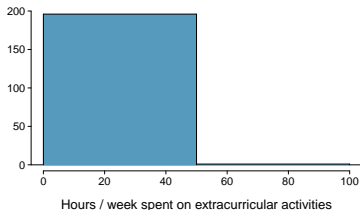


# Histograms

- ▶ Useful for visualizing **one** numerical variable.
- ▶ A histogram is a plot with rectangles whose area represents the **frequency** of data and width represents consecutive, non-overlapping intervals of a variable.
- ▶ The x-axis of a histogram represents the possible values that each data point can take. The y-axis represents the frequency of each of these values.
- ▶ The horizontal range of the histogram is divided into regular sub-ranges called **bins**. The number of data points that fall into each bin is represented by the height of the corresponding rectangle.
- ▶ Histograms provide a visual summary of large amounts of data, and are particularly useful for understanding the **skewness** and the **kurtosis** of the data distribution.

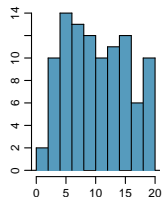
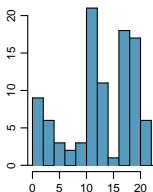
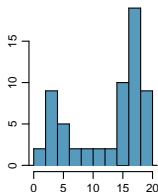
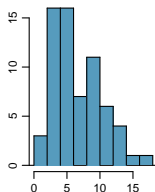
# Bin width

- ▶ The same dataset can lead to very different histograms based on the chosen bin width.
- ▶ Preferred: neither too many nor too few bins.



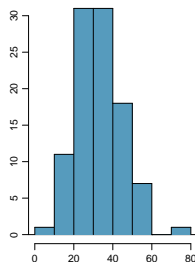
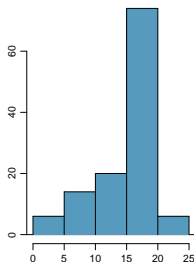
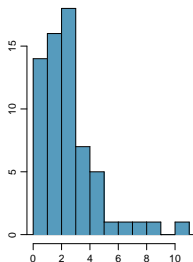
# Histograms and Modality

- ▶ The number of peaks in the histogram gives insight for the modality of the underlying data.
- ▶ *Example:* Below, from left to right, we have unimodal, bimodal, multimodal, uniform.



# Histograms and Skewness

- ▶ Skewness is a measure of **asymmetry**.
- ▶ A histogram can visually represent skewness.
- ▶ If the distribution of data is skewed to the left, the left tail is long and the mass of the distribution is concentrated on the right; if it's skewed to the right, the right tail is long and the mass of the distribution is concentrated on the left.



# Commonly observed shapes of distributions

- ▶ modality

# Commonly observed shapes of distributions

► modality

unimodal



# Commonly observed shapes of distributions

► modality

unimodal



bimodal



# Commonly observed shapes of distributions

► modality

unimodal



bimodal



multimodal





# Commonly observed shapes of distributions

► modality

unimodal



bimodal



multimodal



uniform



► skewness

# Commonly observed shapes of distributions

## ► modality

unimodal



bimodal



multimodal



uniform



## ► skewness

right skew



# Commonly observed shapes of distributions

## ► modality

unimodal



bimodal



multimodal



uniform



## ► skewness

right skew



left skew



# Commonly observed shapes of distributions

## ► modality

unimodal



bimodal



multimodal



uniform



## ► skewness

right skew



left skew

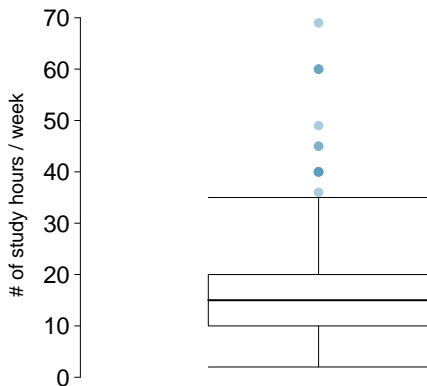


symmetric

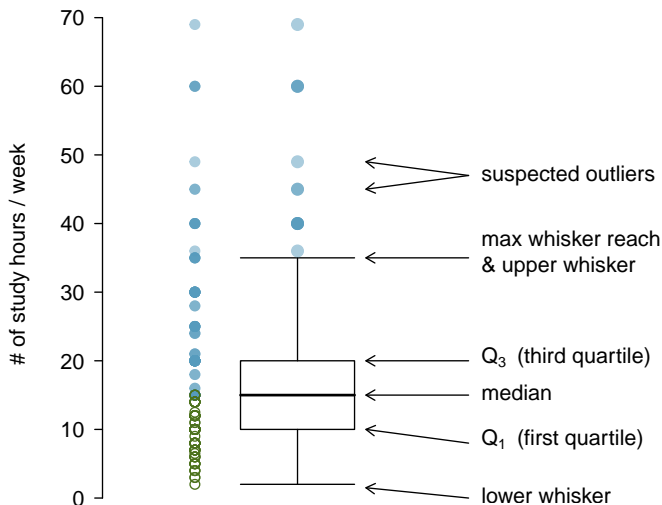


# Box plot

The box in a **box plot** represents the middle 50% of the data, and the thick line in the box is the median.



# Anatomy of a box plot



## Whiskers and outliers

- ▶ Define the **interquartile range** to be  $IQR = Q_3 - Q_1$ .
- ▶ **Whiskers** of a box plot can extend up to  $1.5 \times IQR$  away from the quartiles.

$$\text{max upper whisker reach} = Q_3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q_1 - 1.5 \times IQR$$

## Whiskers and outliers

- ▶ Define the **interquartile range** to be  $IQR = Q_3 - Q_1$ .
- ▶ **Whiskers** of a box plot can extend up to  $1.5 \times IQR$  away from the quartiles.

$$\text{max upper whisker reach} = Q_3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q_1 - 1.5 \times IQR$$

$$IQR : 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$



## Whiskers and outliers

- ▶ Define the **interquartile range** to be  $IQR = Q_3 - Q_1$ .
- ▶ **Whiskers** of a box plot can extend up to  $1.5 \times IQR$  away from the quartiles.

$$\text{max upper whisker reach} = Q_3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q_1 - 1.5 \times IQR$$

$$IQR : 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- ▶ A potential **outlier** is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

# Robust statistics

Median and IQR are more **robust** to skewness and outliers than mean and SD. Therefore,

- ▶ for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- ▶ for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

# Robust statistics

Median and IQR are more **robust** to skewness and outliers than mean and SD. Therefore,

- ▶ for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- ▶ for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

# Robust statistics

Median and IQR are more **robust** to skewness and outliers than mean and SD. Therefore,

- ▶ for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- ▶ for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

*Median*

# Categorical Data

- ▶ **Definition:** Categorical data represent *characteristics* such as a person's gender, marital status, hometown, or the types of movies they like.
- ▶ **Types:** Two types exist - **nominal** (no order) and **ordinal** (ordered).
- ▶ **Ubiquity and Usefulness:** Categorical data is common and valuable in many fields, from market research to health science. Its ability to classify individuals or items into different groups makes it ideal for comparison, identification of patterns, and decision-making.

# Visualization

5 main ways of visualizing categorical data:

- ▶ Contingency tables
- ▶ Bar plots
- ▶ Mosaic plots
- ▶ Pie charts
- ▶ Side-by-side box plots

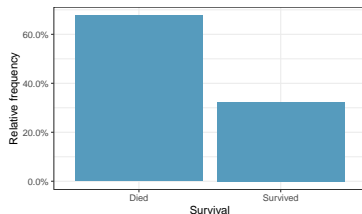
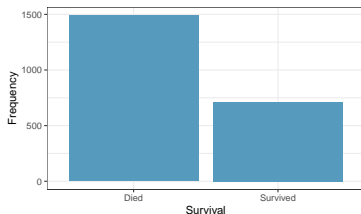
## Contingency tables

- ▶ A table that summarizes data for two categorical variables is called a **contingency table**.
- ▶ The contingency table below shows the distribution of survival and ages of passengers on the Titanic.

	Survival		Total
	Died	Survived	
Age	Adult	1438      654	2092
	Child	52        57	109
	Total	1490      711	2201

# Bar plots

- ▶ A **bar plot** is a common way to display a single categorical variable.
- ▶ A bar plot where proportions instead of frequencies are shown is called a **relative frequency bar plot**.



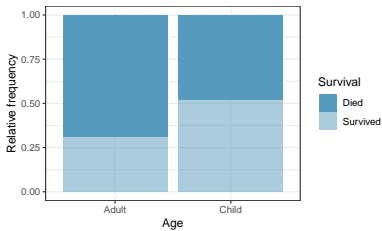
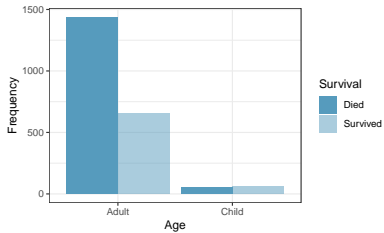
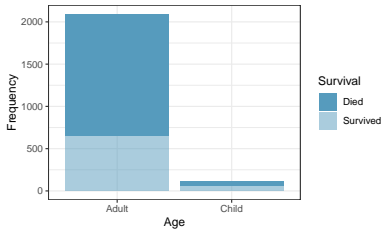


# Bar plots vs. Histograms

- ▶ Bar plots are used for displaying distributions of categorical variables.
- ▶ Histograms are used for numerical variables.
- ▶ The x-axis in a histogram is a number line, hence the order of the bars cannot be changed.
- ▶ In a bar plot, the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

## Bar plots with two variables

- ▶ **Stacked bar plot:** Graphical display of contingency table information, for counts.
- ▶ **Side-by-side bar plot:** Displays the same information by placing bars next to, instead of on top of, each other.
- ▶ **Standardized stacked bar plot:** Graphical display of contingency table information, for proportions.



# Mosaic Plots

- ▶ **Definition:** Mosaic plots are a graphical representation of the cell frequencies of a contingency table. They divide the area into rectangular regions corresponding to categories of variables, with the area of each rectangle proportional to the cell frequencies.
- ▶ **Visualizing Relationships:** Mosaic plots are useful for visualizing the relationships between categorical variables in a dataset. The distribution of categories in one variable can be compared across levels of other variables.
- ▶ **Coloring:** Colors or shading can be used in mosaic plots to show additional dimensions, such as the outcome of a statistical test, or to highlight particular aspects of the data.

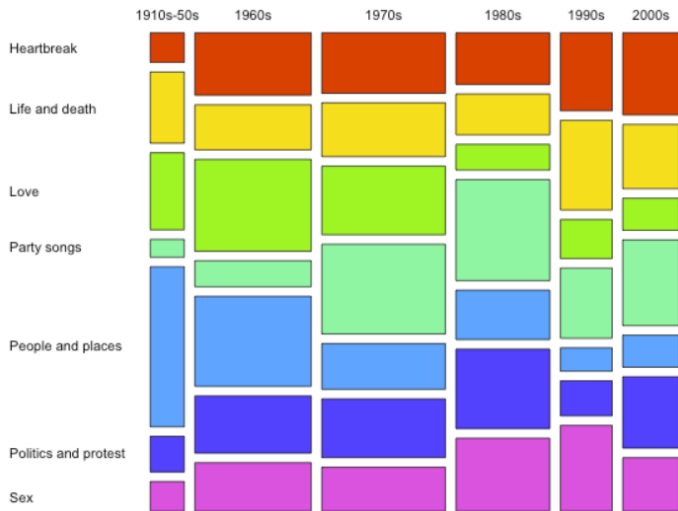


FIGURE: Source: Wikipedia

# References

- ▶ Sections 2.1 & 2.2 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 7

## Basic Probability

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 11, 2023

# Random Processes (informally)

- ▶ A **random process** is like a game of chance that keeps going over time. It's like rolling a dice, not just once, but repeatedly at specific intervals.
- ▶ The idea is that what happens at any given roll (or time point) doesn't necessarily depend on what happened before, and you can't predict exactly what will happen next. You might have some idea about the possibilities (like knowing the dice can land on any number from 1 to 6), but you can't say for sure.
- ▶ It's everywhere in real life, like checking the weather every day, watching the stock market, or even studying the habits of a bunch of ants!



## Random Processes (more formally)

- ▶ A **random process** (also known as stochastic process) is a collection of random variables indexed by time or space. Each random variable in the collection represents an event or outcome at a particular time or location.
- ▶ The randomness implies that the outcomes of the process (events or values it produces) are not fully predictable, although they may be governed by certain probabilistic laws.
- ▶ Depending on the properties of the process, different points in time may exhibit varying degrees of dependency or independence.
- ▶ Random processes are fundamental in the study of systems that evolve over time or space under uncertainty, such as queueing systems, stock prices, signal processing, and many natural phenomena.

# Examples

- ▶ **Coin Tossing:** Each toss of a fair coin is a simple random process. The outcome of each toss (head or tail) is random and does not depend on the outcome of any previous toss.
- ▶ **Temperature Measurements:** If you measure the temperature at a specific location every day at noon, the result is a random process. Each day's temperature is influenced by a variety of factors (weather, seasons, etc.), making the sequence of temperatures a random process.
- ▶ **Stock Prices:** The daily closing price of a stock is a random process. Various factors like market conditions, company performance, and global events contribute to the randomness of stock prices.

# Outcome and Sample Space

- ▶ An **outcome** is a possible result of a random process or experiment.
- ▶ The **sample space** is the set of all possible outcomes of an experiment.
- ▶ *Examples:*
  - ▶ A coin toss has sample space  $S = \{H, T\}$ .
  - ▶ A sequence of two coin tosses has sample space  $S = \{HH, HT, TH, TT\}$ .
  - ▶ When rolling a die, the sample space is  $S = \{1, 2, 3, 4, 5, 6\}$ , each number is an outcome.

# Event

- ▶ An **event** is a set of outcomes of an experiment (a subset of the sample space).
- ▶ *Examples:*
  - ▶ An event when rolling a die could be “the result is an even number”. This event includes is written  $\{2, 4, 6\}$ .
  - ▶ In a coin toss, “tails” is an event and we denote it as  $\{T\}$ .
  - ▶ In a sequence of two coin tosses, the event of getting “at least one head” is  $\{HH, HT, TH\}$ .
- ▶ The sample space itself is an event.

# Event Operations

- ▶ The **union** of events  $A$  and  $B$  is the event that either  $A$  or  $B$  (or both) occur. It's represented as  $A \cup B$ .
- ▶ The **intersection** of events  $A$  and  $B$  is the event that both  $A$  and  $B$  occur. It's represented as  $A \cap B$ .
- ▶ The **complement** of an event  $A$  is the event that  $A$  does not occur, denoted as  $A^c$ .
- ▶ The **set difference** of two events  $A$  and  $B$  consists of all elements of  $A$  which are not in  $B$ . It is denoted by  $A \setminus B$ .

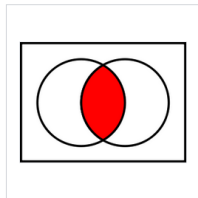
# Disjoint Events

- ▶ Events  $A$  and  $B$  are **disjoint** (or mutually exclusive) if they cannot both occur at the same time. In other words, if  $A$  occurs,  $B$  cannot, and vice versa.
- ▶ For disjoint events, the intersection of  $A$  and  $B$  is the empty set, denoted as  $A \cap B = \emptyset$ .
- ▶ *Examples:*
  - ▶ Suppose a dice is rolled. Then,  $S = \{1, 2, 3, 4, 5, 6\}$ . The events  $A = \{1\}$  and  $B = \{3\}$  are disjoint, since we cannot get a 1 and a 3 at the same dice roll.
  - ▶ For two coin tosses, consider the events  $A = \{HH, HT\}$ ,  $B = \{HT, TH\}$ ,  $C = \{TT\}$ . Then,  $A$  and  $C$  are disjoint,  $B$  and  $C$  are disjoint,  $A$  and  $B$  are NOT disjoint.

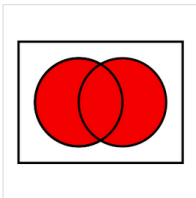
# Venn Diagrams

- ▶ A **Venn diagram** is a diagrammatic representation of events using circles or other shapes.
- ▶ Each circle represents an event, and the space inside the circle represents the elements of the event.
- ▶ Overlapping regions of circles represent common elements between events.
- ▶ The sample space is usually represented by a rectangle.
- ▶ Venn diagrams are very useful for visualizing operations between events.

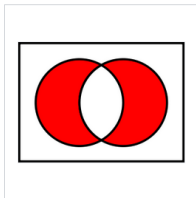
# Venn Diagrams



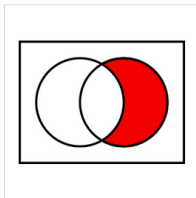
Intersection of two sets  $A \cap B$



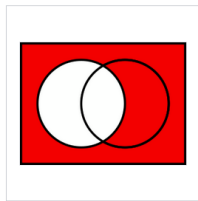
Union of two sets  $A \cup B$



Symmetric difference of two sets  
 $A \triangle B$



Relative complement of A (left) in B  
(right)  $A^c \cap B = B \setminus A$



Absolute complement of A in U  
 $A^c = U \setminus A$

FIGURE: Source: Wikipedia



## A cool Venn diagram

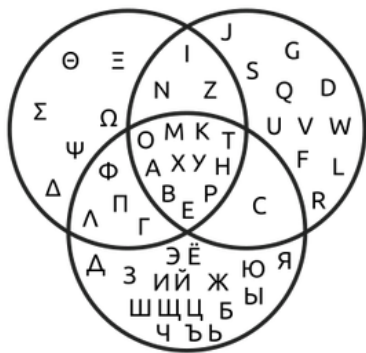


FIGURE: The uppercase letters shared by the Latin, Russian and Greek alphabets (source: Wikipedia)

## Example: Dice Rolling

**Problem:** Two six-sided dice are rolled.

- ▶ Event  $D$ : The sum of the numbers shown is 7.
- ▶ Event  $E$ : The first die shows a 4.

### Questions:

1. What is the sample space  $S$  for this experiment?
2. What is  $D \cap E$ ? Interpret its meaning.
3. If Event  $F$  is that the second die shows an even number, what is  $D \cap F$ ?
4. What is  $E \setminus D$ ? Interpret its meaning.

## Solution

1.  $S = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}.$
2.  $D \cap E = \{(4, 3)\}.$  The first die shows a 4 and the sum is 7.
3.  $D \cap F = \{(1, 6), (3, 4), (5, 2)\}.$  The sum is 7 and the second die shows an even number.
4.  $E \setminus D = \{(4, 1), (4, 2), (4, 4), (4, 5), (4, 6)\}.$  The first die shows a 4 but the sum is not 7.

## Example: Selecting Students

**Problem:** A class of 30 students is composed of 18 males and 12 females. Out of these, 10 males and 7 females have brown eyes.

- ▶ Event  $G$ : A randomly selected student is male.
- ▶ Event  $H$ : A randomly selected student has brown eyes.
- ▶ Event  $I$ : A randomly selected student is female.

### Questions:

1. What does the event  $G \cap H$  represent? How many elements does it have?
2. What does the event  $G^c$  represent? How many elements does it have?
3. What does the event  $H \setminus I$  represent, and how many elements does it have?

# Solution

## Solutions:

1.  $G \cap H$  represents males with brown eyes, which are 10.
2.  $G^c$  represents all females, which are 12.
3.  $H \setminus I$  represents all male students with brown eyes, which are  $17 - 12 = 5$ .

## Example: Music Preferences

**Problem:** A survey of 50 people asked about their preference for Jazz and Rock music.

- ▶ 20 people like Jazz
- ▶ 25 people like Rock
- ▶ 15 people like both Jazz and Rock
- ▶ Event  $J$ : A person likes Jazz.
- ▶ Event  $K$ : A person likes Rock.

### Questions:

1. How many people do not like Jazz or Rock?
2. What is  $J \cup K$ ? How many elements does it have?
3. How many people only like Rock music?
4. If Event  $L$  is that a person likes neither Jazz nor Rock, what is  $J^c \cap K^c$  and how many elements does it have?

## Solution

1. Total people liking Jazz or Rock or both  $= 20 + 25 - 15 = 30$ . So,  $50 - 30 = 20$  people do not like either.
2.  $|J \cup K| = 30$  people like either Jazz, Rock, or both.
3.  $|J \setminus K| = |J| - |J \cap K| = 25 - 15 = 10$  (in words, we are saying "Only Rock = Total liking Rock - Both liking Jazz and Rock  $= 25 - 15 = 10$ ".)
4.  $|J^c \cap K^c| = 20$  people like neither Jazz nor Rock.

# References

- ▶ Section 3.1 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.
- ▶ Wikipedia (Venn diagram)



# STOR 155 - Lecture 8

## Basic Probability

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 15, 2023

# Probability (informally)

- ▶ Informally, the probability of an event is the chance of occurrence of the event.
- ▶ **Frequentist interpretation:** The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
- ▶ **Bayesian interpretation:** A Bayesian interprets probability as a subjective degree of belief: for the same event, two separate people could have different viewpoints and so assign different probabilities.

# Probability (formally)

- ▶ We denote the probability of an event  $A$  by  $P(A)$ .
- ▶  $S$  denotes the sample space
- ▶ **Axioms** of probability:
  1.  $0 \leq P(A) \leq 1$
  2.  $P(S) = 1$
  3. **Union Law:** If  $A \cap B = \emptyset$ , then  $P(A \cup B) = P(A) + P(B)$ .

## Two useful rules

- ▶ General Union Law:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- ▶ Probability of the complement:

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

# Independence (informally)

- ▶ Two events  $A$  and  $B$  are independent if the occurrence of one does not give any information on the occurrence of the other.
- ▶ *Example:* If we toss a coin twice under the same conditions, the outcome of each coin (H or T) says nothing about the outcome of the other coin.

## Independence (formally)

- ▶ Two events  $A$  and  $B$  are independent, and we write  $A \perp\!\!\!\perp B$ , if

$$P(A \cap B) = P(A)P(B).$$

- ▶ *Example:* We toss a coin twice. Consider the events  $A = \{H \text{ in the first toss}\}$  and  $B = \{T \text{ in the second toss}\}$ . Then,

$$\begin{aligned} P(H \text{ in first}, T \text{ in second}) &= \frac{1}{4} \\ &= \frac{1}{2} \cdot \frac{1}{2} \\ &= P(H \text{ in first}) \cdot P(T \text{ in second}). \end{aligned}$$

Hence,  $A \perp\!\!\!\perp B$ .

## Independent $\nRightarrow$ Disjoint

- ▶ One might think that the concepts “independent” and “mutually exclusive (disjoint)” are similar, or even equivalent. This is wrong!
- ▶ **Independent:** the occurrence of  $A$  does not affect the occurrence of  $B$ , and vice versa. Mathematically,

$$P(A \cap B) = P(A)P(B).$$

*Example:* In tossing a coin twice, the outcomes of the first and the second toss define independent events.

- ▶ **Mutually exclusive:** the events  $A$  and  $B$  cannot happen simultaneously. Mathematically,

$$P(A \cap B) = 0.$$

*Example:* In tossing a coin twice, the events that we get tails twice and that we get heads twice are mutually exclusive.

# Law of Large Numbers

- ▶ As more observations are collected, the proportion of occurrences with a particular outcome,  $\hat{p}_n$ , converges to the probability of that outcome,  $p$ .
- ▶ *Examples:*
  1. If we keep tossing a coin, the proportion of times we observe Tails approximates  $1/2$ .
  2. If we keep throwing a die, the proportion of times we see the number 1 approximates  $1/6$ .



# Probability distributions

A **probability distribution** lists all possible events and the probabilities with which they occur.

- ▶ The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

# Probability distributions

A **probability distribution** lists all possible events and the probabilities with which they occur.

- ▶ The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

- ▶ Rules for probability distributions:
  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1

# Probability distributions

A **probability distribution** lists all possible events and the probabilities with which they occur.

- ▶ The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

- ▶ Rules for probability distributions:

1. The events listed must be disjoint
2. Each probability must be between 0 and 1
3. The probabilities must total 1

- ▶ The probability distribution for the genders of two kids:

Event	MM	FF	MF	FM
Probability	0.25	0.25	0.25	0.25

## Example 1: Tossing a Coin

**Problem:** You toss a fair coin. What is the probability of getting a heads (H)?

**Solution:**

- ▶ Sample space  $S = \{H, T\}$
- ▶ The event  $A = \text{Getting a heads} = \{H\}$
- ▶ Since the coin is fair,  $P(H) = \frac{1}{2}$

## Example 2: Rolling a Die





















































**Problem:** You roll a fair six-sided die. What is the probability of getting an even number?

**Solution:**

- ▶ Sample space  $S = \{1, 2, 3, 4, 5, 6\}$
- ▶ The event  $B = \text{Getting an even number} = \{2, 4, 6\}$
- ▶  $P(B) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes}} = \frac{3}{6} = \frac{1}{2}$

# Deck of Cards

- ▶ 52 cards
- ▶ Four suits: Clubs, Diamonds, Hearts, Spades
- ▶ Each suit has cards of 13 ranks: Ace, 2,3,4,5,6,7,8,9,10,Jack, Queen, King.
- ▶ Clubs and Spades are black suits; Diamonds and Hearts are red suits.
- ▶ The above means: there are 13 cards of each suit, 4 cards of each rank, 26 cards of each color.

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

## Example 3: Drawing a Card

**Problem:** From a standard deck of cards, you draw a card. What is the probability of getting a Queen or a King?

**Solution:**

- ▶ Sample space  $S = 52$  cards
- ▶ Event  $C =$  Getting a Queen  $= 4$  cards
- ▶ Event  $D =$  Getting a King  $= 4$  cards
- ▶  $P(C \cup D) = P(C) + P(D)$  [Using the Union Law]
- ▶  $P(C \cup D) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$

## Example 4: Independence

**Problem:** You toss a coin and roll a die. What is the probability of getting a tails (T) and a 5?

**Solution:**

- ▶ Probability of getting tails  $P(T) = \frac{1}{2}$
- ▶ Probability of getting a 5 on the die  $P(5) = \frac{1}{6}$
- ▶ Since the two events are independent:

$$P(T \cap 5) = P(T) \cdot P(5) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$



## Example 5

**Problem:** You flip a coin 6 times. What is the probability that we get at least one H?

**Solution:** Call  $A$  the event that we get at least one H. The idea is to think of the complement:

$$P(A) = 1 - P(A^c) = 1 - (1/2)^6 = \frac{63}{64}.$$

## Example 6: Law of Large Numbers

**Problem:** Imagine tossing a coin 10 times, 100 times, and 1000 times. What can we expect about the proportion of heads?

**Solution:**

- ▶ As the number of tosses increases, the proportion of heads should converge to the true probability of getting a head:  $\frac{1}{2}$ .
- ▶ For 10 tosses, you might see 7 heads or 3 tails, but for 1000 tosses, it's more likely you'll see closer to 500 heads.

# References

- ▶ Section 3.1 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 9

## Conditional Probability

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 18, 2023

# Motivation

- ▶ In real-life situations, the probability of an event can often depend on whether another event has occurred. This concept is known as **Conditional Probability**.
- ▶ *Examples:*
  1. Consider a medical test for a disease. The probability that a person has the disease, **given** a positive test result, depends on the overall prevalence of the disease in the population.
  2. Another example is weather prediction. The probability of rain tomorrow may depend on today's weather. If it's cloudy today, the chance of rain tomorrow could be higher.
- ▶ Conditional probability is a powerful tool that allows us to **update** probabilities based on new information.

## Example

- ▶ Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

- ▶ What is the probability that a patient relapsed?
- ▶ What is the probability that a patient took lithium?
- ▶ What is the probability that a patient relapsed given that they took placebo?

- ▶ Last row, 1st column:

$$P(\text{relapse}) = \frac{48}{72}$$

- ▶ 2nd row, last column:

$$P(\text{lithium}) = \frac{24}{72}$$

- ▶ 3rd row, 1st column:

$$P(\text{relapse}|\text{placebo}) = \frac{20}{24}$$

Note that

$$P(\text{relapse}|\text{placebo}) = \frac{P(\text{relapse} \cap \text{placebo})}{P(\text{placebo})}$$

# Conditional Probability

- ▶ **Definition:** the conditional probability of an event  $A$  given an event  $B$  is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- ▶ Essentially, by conditioning on the event  $B$ , we **shrink** the sample space.
- ▶ Note: in general,  $P(A|B) \neq P(B|A)$ .
- ▶ Multiplication Rules:

$$P(A \cap B) = P(A|B)P(B) \quad \text{and} \quad P(A \cap B) = P(B|A)P(A)$$



# Understanding Conditional Probability

- ▶  $P(A \cap B)$  means the probability of both  $A$  and  $B$  occurring.
- ▶  $P(B)$  is the probability of event  $B$ .
- ▶  $P(A|B)$  is the proportion of times  $A$  occurs when  $B$  has already occurred.
- ▶ The conditional probability formula informally reads:  
*“to find the probability of  $A$  given  $B$ , check what is the chance of  $B$  happening, and for each such occurrence what is the chance of  $A$  happening”.*

# Conditional Probability and Independence

- ▶ Recall:  $A$  and  $B$  are **independent** if

$$P(A \cap B) = P(A)P(B).$$

- ▶ *Alternative formulation:* If  $A$  and  $B$  are independent, then  $P(A|B) = P(A)$  and  $P(B|A) = P(B)$ .
- ▶ *Intuition:* the occurrence of  $B$  doesn't affect the probability of  $A$ , and vice-versa.
- ▶ If  $P(A|B) \neq P(A)$ ,  $A$  and  $B$  are not independent.

# Motivating Bayes' Theorem: A Medical Example

- ▶ Bayes' Theorem is a fundamental concept in probability that allows us to update our beliefs based on new evidence.
- ▶ Consider a medical test for a disease that affects 1% of the population. The test has a 5% false positive rate and no false negatives.
- ▶ You tested positive. What's the probability you have the disease? It's not 95%!
- ▶ Bayes' Theorem allows us to calculate this probability, incorporating both the test's false positive rate and the prevalence of the disease in the population.

# Bayes' Theorem

- ▶ The conditional probability formula we have seen so far is a special case of the Bayes' Theorem, which is applicable even when events have more than just two outcomes.
- ▶ Suppose a variable has  $k$  possible outcomes,  $A_1, \dots, A_k$ . Then, given that an event  $B$  has occurred, the conditional probability of  $A_1$  is given by

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + \dots + P(A_k)P(B|A_k)}$$

# Excellent Resource!

Great explanation of Bayes' Theorem by the YouTube channel 3blue1brown:

Bayes theorem, the geometry of changing beliefs

## Example

- ▶ A disease affects 1% of the population (Event D).
- ▶ A test is 99% accurate: it correctly identifies a sick person 99% of the time (true positive), and a healthy person 99% of the time (true negative).
- ▶ However, the test has a 1% false positive rate and a 1% false negative rate.
- ▶ You take the test and get a positive result (Event T). What's the probability you have the disease?

## Using Bayes' Theorem

- ▶ We want to find  $P(D|T)$ , the probability of having the disease given a positive test result.
- ▶ Bayes' Theorem states:  $P(D|T) = \frac{P(T|D) \cdot P(D)}{P(T)}$ .
- ▶ We know  $P(T|D)$  (the probability of a positive test given the disease) is 0.99.
- ▶  $P(D)$  (the probability of the disease) is 0.01.
- ▶ The hard part is usually the computation of the denominator  $P(T)$ .

## Calculating the Answer

- ▶  $P(T)$  (the probability of a positive test) is not just the false positive rate. It's the probability of a true positive or a false positive:

$$\begin{aligned}P(T) &= P(T \cap D) + P(T \cap D^c) \\ &= P(T|D)P(D) + P(T|D^c)P(D^c)\end{aligned}$$

- ▶ This gives us  $P(T) = 0.99 \cdot 0.01 + 0.01 \cdot 0.99 = 2 \cdot 0.99 \cdot 0.01$ .
- ▶ Plugging into Bayes' Theorem:  $P(D|T) = \frac{0.99 \cdot 0.01}{2 \cdot 0.99 \cdot 0.01} = 0.5$ .
- ▶ Even with a positive test, there's only a 50% chance you have the disease!



## Example

Suppose a student is enrolled in a particularly challenging statistics class. Based on records from previous years:

- ▶ 70% of students who study regularly for the class pass the midterm.
- ▶ 20% of students who do not study regularly for the class pass the midterm.
- ▶ Overall, 50% of students study regularly for the class.

If a randomly selected student passes the midterm, what is the probability that they studied regularly for the class?

# Solution

Let:

- ▶  $R$  be the event that a student studies regularly for the class.
- ▶  $P$  be the event that a student passes the midterm.

Given:

- ▶ Probability that a student passes given they study regularly:

$$P(P|R) = 0.7$$

- ▶ Probability that a student passes given they don't study regularly:

$$P(P|R^c) = 0.2$$

- ▶ Probability that a student studies regularly:

$$P(R) = 0.5$$

We want to find  $P(R|P)$  (Probability that a student studied regularly given they passed). Bayes' theorem:

$$P(R|P) = \frac{P(P|R) \cdot P(R)}{P(P)},$$

where

$$P(P) = P(R) \cdot P(P|R) + P(R^c) \cdot P(P|R^c).$$

Now, calculating step by step:

- ▶  $P(R^c) = 1 - P(R) = 0.5$  (Probability that a student doesn't study regularly)
- ▶  $P(P) = 0.5 \cdot 0.7 + 0.5 \cdot 0.2 = 0.45$
- ▶  $P(R|P) = \frac{0.7 \cdot 0.5}{0.45} = 0.7778$

Thus, if a student passes the midterm, the probability that they studied regularly is 77.78%.

# References

- ▶ Section 3.2 of the textbook
- ▶ 3blue1brown
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 10

## Random Variables

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 20, 2023

# Introduction to Random Variables

- ▶ We are often interested in numerical outcomes of random processes.
- ▶ For instance: The number of heads in 10 coin tosses.
- ▶ A **random variable** is a function that assigns a real number to each outcome in a sample space.
- ▶ This turns every possible outcome of a random process into a numerical value.
- ▶ We use
  - ▶ a capital letter  $X$  to denote a random variable
  - ▶ a lower-case letter  $x$  to denote a value of the random variable
  - ▶ we write  $P(X = x)$  to denote the probability of a random variable  $X$  taking the value  $x$

# Examples of Discrete Random Variables

- ▶ A Discrete Random Variable has a countable number of possible values.
- ▶ Examples:
  - ▶ The number of heads in 10 coin tosses.
  - ▶ The number of defective items in a batch.
  - ▶ The number of cars arriving at a toll booth in one hour.
- ▶ These random variables can take on a finite or countably infinite number of values.



# Examples of Continuous Random Variables

- ▶ A Continuous Random Variable can take on any value in an interval.
- ▶ Examples:
  - ▶ The amount of rain in a city over a year.
  - ▶ The time taken to run a marathon.
  - ▶ The weight of a randomly chosen adult.
- ▶ These random variables have uncountably many possible values.

# Expectation

- ▶ Suppose  $X$  is a discrete random variable taking values in the set  $\{x_1, \dots, x_k\}$ .
- ▶ We are often interested in the average outcome of a random variable.
- ▶ We call this the **expected value** (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^k x_i P(X = x_i)$$

## Example

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

## Example

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	$X$	$P(X)$	$X P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

# Variance and Standard Deviation

- ▶ Suppose  $X$  is a discrete random variable taking values in the set  $\{x_1, \dots, x_k\}$ .
- ▶ We are also often interested in the variability in the values of a random variable.
- ▶ **Variance:**

$$\sigma^2 = \text{Var}(X) = \sum_{i=1}^k (x_i - E(X))^2 P(X = x_i)$$

- ▶ **Standard Deviation:**

$$\sigma = \text{SD}(X) = \sqrt{\text{Var}(X)}$$

## Example

$X$	$P(X)$	$X P(X)$	$(X - E(X))^2$	$P(X) (X - E(X))^2$
1	$\frac{12}{52}$	$1 \cdot \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \cdot 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \cdot \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \cdot 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \cdot \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \cdot 84.4561 = 1.6242$
0	$\frac{35}{52}$	$0 \cdot \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \cdot 0.6561 = 0.4416$
		$E(X) = 0.81$		

## Example

$X$	$P(X)$	$X P(X)$	$(X - E(X))^2$	$P(X) (X - E(X))^2$
1	$\frac{12}{52}$	$1 \cdot \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \cdot 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \cdot \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \cdot 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \cdot \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \cdot 84.4561 = 1.6242$
0	$\frac{35}{52}$	$0 \cdot \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \cdot 0.6561 = 0.4416$
		$E(X) = 0.81$		

## Example

$X$	$P(X)$	$X P(X)$	$(X - E(X))^2$	$P(X) (X - E(X))^2$
1	$\frac{12}{52}$	$1 \cdot \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \cdot 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \cdot \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \cdot 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \cdot \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \cdot 84.4561 = 1.6242$
0	$\frac{35}{52}$	$0 \cdot \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \cdot 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$



## Example

$X$	$P(X)$	$X P(X)$	$(X - E(X))^2$	$P(X) (X - E(X))^2$
1	$\frac{12}{52}$	$1 \cdot \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \cdot 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \cdot \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \cdot 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \cdot \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \cdot 84.4561 = 1.6242$
0	$\frac{35}{52}$	$0 \cdot \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \cdot 0.6561 = 0.4416$
		$E(X) = 0.81$		$V(X) = 3.4246$ $SD(X) = \sqrt{3.4246} = 1.85$

# Linear combinations

- ▶ A **linear combination** of random variables  $X$  and  $Y$  is given by

$$aX + bY,$$

where  $a$  and  $b$  are some fixed numbers.

- ▶ The expected value of a linear combination of random variables is given by

$$E(aX + bY) = aE(X) + bE(Y)$$

## Example 1

- ▶ On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and chemistry homework for the week?

## Example 1

- ▶ On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and chemistry homework for the week?
- ▶ We can write

$$\begin{aligned} E(S + S + S + S + S + C + C + C + C) &= 5 \cdot E(S) + 4 \cdot E(C) \\ &= 5 \cdot 10 + 4 \cdot 15 \\ &= 50 + 60 \\ &= 110 \text{ min} \end{aligned}$$

## Example 2

- ▶ Let  $X$  and  $Y$  be two independent random variables, with  $E[X] = 2$ ,  $Var(X) = 4$ ,  $E[Y] = 3$  and  $Var(Y) = 9$ .
- ▶ Consider a new random variable  $Z = 3X - 2Y$ .
- ▶ Find the expected value and standard deviation of  $Z$ .

## Solution

- ▶ The expected value of  $Z$  is given by
$$E[Z] = 3E[X] - 2E[Y] = 3 \cdot 2 - 2 \cdot 3 = 0.$$
- ▶ To find the standard deviation of  $Z$ , we first compute its variance.
- ▶ The variance of  $Z$  is given by
$$\text{Var}(Z) = 3^2 \cdot \text{Var}(X) + (-2)^2 \cdot \text{Var}(Y) = 9 \cdot 4 + 4 \cdot 9 = 72.$$
- ▶ The standard deviation of  $Z$  is the square root of its variance, which equals  $\sqrt{72} \approx 8.49$ .

# References

- ▶ Section 3.4 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 11

## Normal Distribution

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 20, 2023

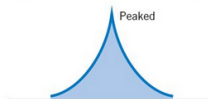
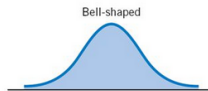
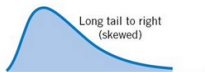
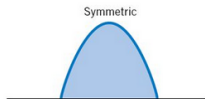
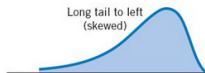


# Introduction to Continuous Distributions

- ▶ A continuous distribution describes the probabilities of the possible values of a continuous random variable.
- ▶ A **continuous** random variable is a random variable with a set of possible values (known as the range) that is infinite and **uncountable**.
- ▶ Examples include measurements like weight, height, and temperature.

# Density Curves

- ▶ A density curve is a graph that shows the probability of a given continuous outcome.
- ▶ The area under a density curve over an interval represents the probability that the variable falls within that interval.
- ▶ Density curves are always above the horizontal axis, and the total area under the curve is equal to 1.



# The Normal Distribution

- ▶ The most famous continuous distribution is the **Normal** (or Gaussian) **distribution**.
- ▶ It is characterized by its **bell-shaped** density curve.
- ▶ The curve is **symmetrical** around the mean, reflecting the fact that data near the mean are more frequent in occurrence than data far from the mean.

# Parameters of the Normal Distribution

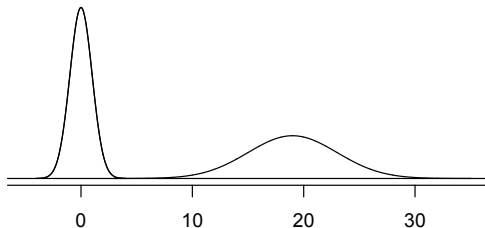
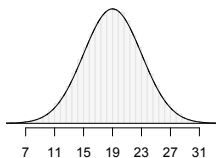
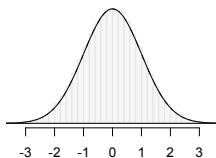
- ▶ The Normal distribution is defined by two parameters: the **mean** ( $\mu$ ) and the **standard deviation** ( $\sigma$ ).
- ▶ The mean determines the location of the center of the graph, and the standard deviation determines the height and width of the graph.
- ▶ When  $\mu = 0$  and  $\sigma = 1$ , it is known as the **Standard Normal Distribution**.

# Normal distributions with different parameters

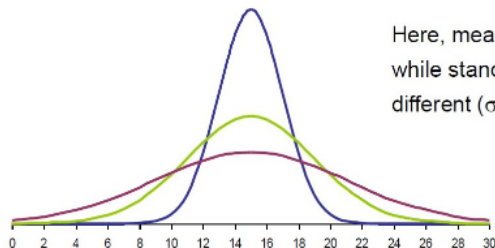
$\mu$ : mean,  $\sigma$ : standard deviation

$$N(\mu = 0, \sigma = 1)$$

$$N(\mu = 19, \sigma = 4)$$



# Normal Curves



Here, means are the same ( $\mu = 15$ ) while standard deviations are different ( $\sigma = 2, 4$ , and  $6$ ).

Here, means are different ( $\mu = 10, 15$ , and  $20$ ) while standard deviations are the same ( $\sigma = 3$ ).



# Standardizing with Z-scores

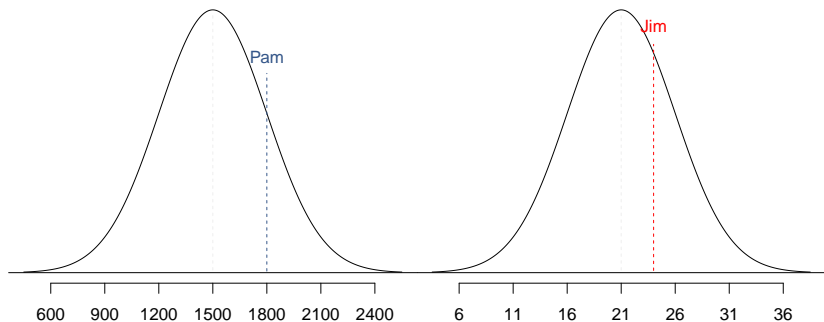
- ▶ **Z-score** of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{\text{observation} - \text{mean}}{SD}$$

- ▶ Z-scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- ▶ Observations that are more than 2 SD away from the mean ( $|Z| > 2$ ) are usually considered unusual.

## Example

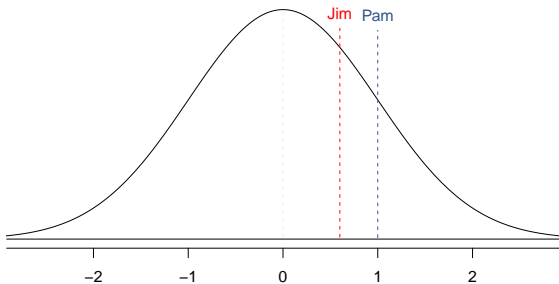
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?





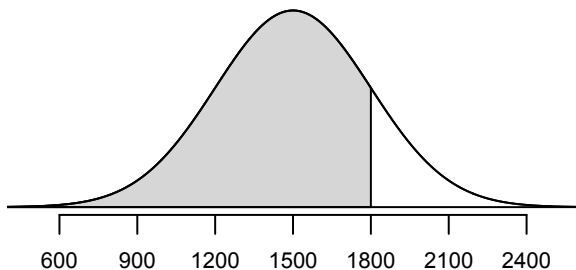
Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- ▶ Pam's score is  $\frac{1800-1500}{300} = 1$  standard deviation above the mean.
- ▶ Jim's score is  $\frac{24-21}{5} = 0.6$  standard deviations above the mean.



# Percentiles

- ▶ **Percentile** is the percentage of observations that fall below a given data point.
- ▶ Graphically, percentile is the area below the probability distribution curve to the **left** of that observation.



# Z-table

Z		Second decimal place of Z									
		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852	
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015	

## The 68-95-99.7 rule

- ▶ Approximately 68% of the data falls within one standard deviation from the mean on either side.
- ▶ Approximately 95% falls within two standard deviations.
- ▶ Almost all (about 99.7%) of the data falls within three standard deviations from the mean.

# Applications of the Normal Distribution

- ▶ The Normal distribution is widely used in the natural and social sciences to represent real-valued random variables whose distributions are not known.
- ▶ Its importance is largely due to the **Central Limit Theorem**.
- ▶ The Central Limit Theorem (CLT) is a fundamental concept in Statistics. It tells us that if we have a large enough sample size, the distribution of the sample means will approach a Normal distribution, regardless of the shape of the population distribution.

# References

- ▶ Section 4.1 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 12

## Geometric Distribution

Panagiotis Andreou

*based on notes by OpenIntro*

UNC Chapel Hill

September 29, 2023

# Motivation

- ▶ Imagine a series of repeated trials, like flipping a coin. What if we're interested in how many trials it takes until we get our first success?
- ▶ This is exactly what the **geometric distribution** models: the *number of trials needed to get the first success in repeated Bernoulli trials*.



# Bernoulli Trials

- ▶ A **Bernoulli trial** is an experiment that results in a success with probability  $p$  and failure with probability  $1 - p$ .
- ▶ Each trial is independent, meaning the outcome of one trial does not affect the outcome of another.
- ▶ *Examples:*
  - ▶ Toss of fair coin: success probability = 0.5, where we have to determine whether success means H or T.
  - ▶ Toss of biased coin: success probability could be any number in  $[0, 1]$ .

# PMF of the Geometric Distribution

- ▶ If  $X$  is a geometrically distributed random variable with probability  $p$ , we write  $X \sim \text{Geom}(p)$ .
- ▶ If  $X \sim \text{Geom}(p)$ , its probability mass function (PMF) is given by

$$P(X = k) = (1 - p)^{k-1}p$$

for  $k = 1, 2, 3, \dots$

- ▶ This formula gives the probability that we need  $k$  trials to get the first success.

# Mean and Variance

If  $X \sim \text{Geom}(p)$ , then

- ▶ its expectation is given by

$$E[X] = \frac{1}{p}$$

- ▶ its variance is given by

$$\text{Var}(X) = \frac{1-p}{p^2}$$

## Derivation

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} k \cdot P(X = k) \\ &= \sum_{k=1}^{\infty} k \cdot (1 - p)^{k-1} \cdot p \\ &= p \cdot \sum_{k=1}^{\infty} k \cdot (1 - p)^{k-1} \\ &= p \cdot (1 + 2(1 - p) + 3(1 - p)^2 + 4(1 - p)^3 + \dots) \\ &= p \cdot \frac{1}{p^2} \quad (\text{calculus fact}) \\ &= \frac{1}{p} \end{aligned}$$

# Problem 1

- ▶ Consider flipping a biased coin that comes up heads (success) with probability  $p = 0.3$ .
- ▶ What is the probability that the first head occurs on the third flip?

## Solution

- ▶ We can use the PMF of the geometric distribution to solve this.
- ▶  $P(X = 3) = (1 - 0.3)^{3-1} \cdot 0.3 \approx 0.147$
- ▶ So, the probability that the first head occurs on the third flip is approximately 0.147.

## Problem 2 - Binge Watching Dilemma

A popular streaming service claims that a user will find a new show they like 80% of the time. Assuming each suggestion is independent of the others:

- ▶ (a) What is the probability that a user will find a show they like on the  $k^{th}$  suggestion?
- ▶ (b) On average, how many suggestions does it take for a user to find a show they like?
- ▶ (c) What is the variability in the number of suggestions?

## Solution

- ▶ Let  $Y$  represent the number of suggestions it takes until the user finds a show they like.
- ▶  $Y \sim \text{Geom}(q)$ , where  $q = 0.8$  (probability of liking a show from a suggestion).
- ▶ (a) The pmf of  $Y$  is  $P(Y = k) = (1 - q)^{k-1}q$ , for  $k = 1, 2, \dots$
- ▶ For this problem,  $q = 0.8$ , so
$$P(Y = k) = (1 - 0.8)^{k-1} \cdot 0.8 = 0.2^{k-1} \cdot 0.8.$$



- ▶ (b) The expected (mean) number of suggestions from a geometric distribution is  $E[Y] = \frac{1}{q}$ .
- ▶ For our problem,  $E[Y] = \frac{1}{0.8} = 1.25$ . So, on average, it takes about 1.25 suggestions for a user to find a show they like.
- ▶ (c) The variance of a geometric distribution is  $Var(Y) = \frac{1-q}{q^2}$ .
- ▶ In our context,  $Var(Y) = \frac{1-0.8}{(0.8)^2} = \frac{0.2}{0.64} \approx 0.3125$ . So, there's a variance of approximately 0.3125 in the number of suggestions.

## Problem 3 - Malware detection

A software company develops a new antivirus software. They believe that each individual attempt to detect a certain type of malware with their software is independent and has a success probability of 0.7.

- ▶ (a) What is the probability of needing  $k$  attempts for a malware detection?
- ▶ (b) What is the expected number of attempts until detection?
- ▶ (c) What is the variance?

## Solution

- ▶ Let  $X$  represent the number of attempts it takes until the software detects this specific malware.
- ▶  $X \sim \text{Geom}(p)$ , with  $p = 0.7$
- ▶ (a) The pmf of  $X$  is  $P(X = k) = (1 - p)^{k-1}p$ , for  $k = 0, 1, 2, \dots$
- ▶ In this case,  $p = 0.7$ , so  
$$P(X = k) = (1 - 0.7)^{k-1} \cdot 0.7 = 0.3^{k-1} \cdot 0.7.$$

- ▶ (b) The expectation (mean) of a geometric distribution is  $E[X] = \frac{1}{p}$ .
- ▶ In this case,  $E[X] = \frac{1}{0.7} \approx 1.43$ . So, on average, it takes about 1.43 attempts to detect the malware.
- ▶ (c) The variance of a geometric distribution is  $Var(X) = \frac{1-p}{p^2}$ .
- ▶ In this case,  $Var(X) = \frac{1-0.7}{(0.7)^2} = \frac{0.3}{0.49} \approx 0.61$ . So, the variance in the number of attempts is approximately 0.61.

# References

- ▶ Section 4.2 of the textbook
- ▶ Sumit Kumar Kar's slides for STOR 151, Fall 2021.

# STOR 155 - Lecture 13

## Binomial Distribution

Panagiotis Andreou

UNC Chapel Hill

October 2, 2023

# Binomial Distribution: Introduction

- ▶ A binomial distribution is a probability distribution that describes the number of successes in a fixed number of Bernoulli trials.
- ▶ Let's denote  $X$  as a random variable that counts the number of successes in  $n$  Bernoulli trials.
- ▶ Each trial is independent, and the probability of success, denoted by  $p$ , remains constant.

# The Binomial Coefficient

- ▶ In how many ways can we select  $k$  out of  $n$  items?
- ▶ This is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

where  $n!$  is pronounced “ $n$  **factorial**” and is defined to be

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$



# Binomial Distribution: PMF

- ▶ The probability mass function (PMF) of a binomial distribution can be written as

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

for  $k = 0, 1, 2, \dots, n$ .

- ▶ The term  $p^k$  is the probability of getting  $k$  successes.
- ▶ The term  $(1 - p)^{n-k}$  is the probability of getting  $n - k$  failures.
- ▶ Multiplying these two terms together gives the probability of any specific order of  $k$  successes and  $n - k$  failures.
- ▶ We then multiply by  $\binom{n}{k}$  to account for all possible orders of these successes and failures.

# Mean

- ▶ The expected value or mean of a Binomial Distribution is given by  $E[X] = np$
- ▶ This is a straightforward result considering that  $X$  is the sum of  $n$  Bernoulli random variables each with expectation  $p$ .
- ▶ Hence, the mean is just the number of trials times the probability of success on each trial.

# Mean Derivation

- ▶ Suppose  $X_1, X_2, \dots, X_n$  are independent Bernoulli random variables, where each  $X_i$  equals 1 with probability  $p$  and 0 with probability  $1 - p$ .
- ▶  $X = X_1 + X_2 + \dots + X_n$  is a Binomial random variable.
- ▶ By the linearity of expectation,

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n].$$

- ▶ Since the expectation of each  $X_i$  is  $p$ , it follows that

$$E[X] = np.$$

# Variance

- ▶ The variance of a Binomial Distribution is given by

$$\text{Var}(X) = np(1 - p).$$

- ▶ The intuition behind this is that each Bernoulli trial contributes  $p(1 - p)$  to the variance and there are  $n$  such trials.
- ▶ Thus, the total variance is  $np(1 - p)$ .

## Example 1 - Flipping a Fair Coin

- ▶ You flip a fair coin 5 times. What is the probability of getting exactly 3 tails?
- ▶ What is the expected number of tails?

## Solution

- ▶ Let  $X$  be the number of tails.
- ▶  $X \sim \text{Binomial}(5, 0.5)$ .
- ▶ Using the PMF of the binomial distribution:

$$P(X = 3) = \binom{5}{3} \cdot 0.5^3 \cdot 0.5^2 = 10 \cdot 0.125 \cdot 0.25 = 0.3125$$

- ▶ So, there's a 31.25% chance of getting exactly 3 tails.
- ▶ The expected number of tails is  $np = 5 \cdot 0.5 = 2.5$

## Example 2 - Flipping a Biased Coin

Consider a biased coin, in which the probability of getting tails in one throw is  $p$ . You throw this coin 5 times.

- a. What is the probability that you get T, H, H, T, T?
- b. What is the probability of getting exactly 3 tails?
- c. How many tails do you expect to see?

## Solution

- a. The probability of observing T, H, H, T, T is

$$p \cdot (1 - p) \cdot (1 - p) \cdot p \cdot p = p^3(1 - p)^2$$

- b. For the other two, let  $X$  denote the number of tails.

- c. Then,  $X \sim \text{Bin}(5, p)$ .

- d. The probability of getting exactly 3 tails is

$$P(X = 3) = \binom{5}{3} p^3(1 - p)^2 = 10p^3(1 - p)^2$$

(10 times higher than the previous probability, since now the specific order doesn't matter!)

- e. The expected number of tails is  $E[X] = 5p$ .



## Example 3 - Quality Control at a Factory

- ▶ A factory produces light bulbs, 95% of which are perfect. You randomly select 10 light bulbs.
- ▶ What is the probability that at most 2 of them are defective?

## Solution

- ▶ Let  $X$  be the number of defective bulbs.
- ▶  $X \sim \text{Binomial}(10, 0.05)$ .
- ▶ Using the PMF of the binomial distribution:

$$P(X \leq 2) = \sum_{k=0}^2 \binom{10}{k} \cdot 0.05^k \cdot 0.95^{10-k}$$

- ▶ Calculating,  $P(X \leq 2) \approx 0.9884$ .
- ▶ So, there's a 98.84% chance that at most 2 out of the 10 bulbs are defective.

## Example 4 - Clinical Trial

- ▶ A new drug is believed to be effective in 60% of patients. In a clinical trial, it's administered to 7 patients.
- ▶ What is the probability that it is effective in at least 6 patients?
- ▶ What is the expected number of patients for which the drug is effective?

## Solution

- ▶ Let  $X$  be the number of patients for whom the drug is effective.
- ▶  $X \sim \text{Binomial}(7, 0.6)$ .
- ▶ Using the PMF of the binomial distribution:

$$P(X \geq 6) = \sum_{k=6}^7 \binom{7}{k} \cdot 0.6^k \cdot 0.4^{7-k} = 7 \cdot (0.6)^6 \cdot (0.4) + (0.6)^7$$

- ▶ Calculating,  $P(X \geq 6) \approx 0.2611$ .
- ▶ So, there's a 26.11% chance that the drug is effective in at least 6 out of the 7 patients.
- ▶ The expected number of patients for which the drug is effective, is  $np = 7 \cdot 0.6 = 4.2$ .

# References

- ▶ Section 4.3 of the textbook
- ▶ Great resource for the more curious:  
Binomial distributions (3blue1brown)

# STOR 155 - Lecture 14

## Practice Examples in Probability

Panagiotis Andreou

UNC Chapel Hill

October 4, 2023

## Example 1 - Expectation of Bernoulli

- ▶ Suppose that  $X$  is a random variable that follows Bernoulli distribution with success probability  $p$ . We write this as

$$X \sim \text{Bernoulli}(p).$$

- ▶ What is the expected value of  $X$ , namely  $E[X]$ ?

## Solution

- ▶ If we recall the definition of expectation, the first thing we need to determine is the set of values that  $X$  can take.
- ▶ Here,  $X$  follows a Bernoulli distribution, so by definition it can only take the values 0 and 1.
- ▶ The next step is to determine with what probability  $X$  takes each value.
- ▶  $X = 1$  with probability  $p$ , and  $X = 0$  with probability  $1 - p$ .
- ▶ Combining the above, we compute the expectation:

$$E[X] = 1 \cdot p + 0 \cdot (1 - p) = p.$$



## Example 2 - Hair Color Genetics

A couple both have black hair but carry genes that make it possible for their children to have:

- ▶ Black hair with probability 0.6,
- ▶ Blonde hair with probability 0.2,
- ▶ Red hair with probability 0.2.

(a) What is the probability the first red-haired child they have is their third child?

(b) On average, how many children would such a pair of parents have before having a red-haired child? What is the standard deviation of the number of children they would expect to have until the first red-haired child?

## Solution to (a)

- ▶ For the first red-haired child to be the third child, the first two children should not have red hair, while the third should.
- ▶ Probability that a child has red hair is  $p = 0.2$ .
- ▶ Let  $X$  denote the number of children the couple has before they have a red-haired child. Then,  $X \sim \text{Geom}(0.2)$ .
- ▶ Using the P.M.F. of the Geometric distribution, we get

$$\begin{aligned} P(\text{1st and 2nd not red, 3rd red}) &= P(X = 2) \\ &= (0.8)^2 \cdot 0.2 = 0.128. \end{aligned}$$

## Solution to (b)

- ▶ The expectation of a geometric distribution is  $E[X] = \frac{1}{p}$ .
- ▶ In this case,  $E[X] = \frac{1}{0.2} = 5$ . So, on average, they would need 5 children to have a red-haired child.
- ▶ The standard deviation of a geometric distribution is  $\sqrt{\frac{1-p}{p^2}}$ .
- ▶ Here,  $p = 0.2$ , so we compute

$$\text{st. dev.} = \sqrt{\frac{1 - 0.2}{(0.2)^2}} = \sqrt{\frac{0.8}{0.04}} = \sqrt{20} = 4.4721.$$

## Example 3 - Smartphone Sampling Problem

In the following situations, assume that 60% of the specified population has a smartphone, and the other 40% does not.

(a) Suppose you're sampling from a classroom with 10 students.

- ▶ What is the probability of sampling two smartphone users in a row when sampling with replacement?
- ▶ What is the probability when sampling without replacement?

(b) Now suppose you're sampling from a concert with 10,000 attendees.

- ▶ What is the probability of sampling two smartphone users in a row when sampling with replacement?
- ▶ What is the probability when sampling without replacement?

(c) We often treat individuals who are sampled from a large population as independent. Using your findings from parts a and b, explain whether or not this assumption is reasonable.

# Solution

(a) Classroom with 10 students:

- ▶ With replacement:  $P = \frac{6}{10} \cdot \frac{6}{10} = 0.36$  or 36%.
- ▶ Without replacement:  $P = \frac{6}{10} \cdot \frac{5}{9} = 0.3333$  or 33.33%.

(b) Concert with 10,000 attendees:

- ▶ With replacement:  $P = 0.6 \cdot 0.6 = 0.36$  or 36%.
- ▶ Without replacement:  $P = 0.6 \cdot \frac{5999}{9999} \approx 0.3600$  or 36%.

(c) Explanation:

- ▶ When the population size is small, the results of sampling with and without replacement are different.
- ▶ When the population is large, the two probabilities are very close.
- ▶ Therefore, the assumption of treating individuals from a large population as independent is reasonable.

## Example 4 - Tetrahedral Dice Problem

Consider a tetrahedral dice with four faces numbered 1, 2, 3, and 4. Each face is equally likely to come up in a single roll of the dice. Suppose you roll this dice four times. Calculate the probability of getting the following:

- (a) At least one '1'.
- (b) Exactly 2 rolls showing '2'.
- (c) Exactly 1 roll showing '3'.
- (d) At most 3 rolls showing '4'.

## Solution

(a)

- ▶ Let  $X$  denote the number of 1's.
- ▶ Then,  $X \sim \text{Bin}(4, 0.25)$ .
- ▶ Probability of getting at least one '1' is equal to

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{3}{4}\right)^4 \approx 0.6836.$$

(b)

- ▶ Let  $Y$  denote the number of 2's.
- ▶ Then,  $Y \sim \text{Bin}(4, 0.25)$ .
- ▶ The probability that we get exactly 2 rolls showing '2' is

$$P(Y = 2) = \binom{4}{2} \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^2 \approx 0.2109.$$

(c)

- ▶ Let  $W$  denote the number of 3's.
- ▶ Then,  $W \sim \text{Bin}(4, 0.25)$ .
- ▶ The probability that we get exactly 1 roll showing '3' is

$$P(W = 1) = \binom{4}{1} \cdot \frac{1}{4} \cdot \left(\frac{3}{4}\right)^3 \approx 0.4219.$$

(d)

- ▶ Let  $Z$  denote the number of 4's.
- ▶ Then,  $Z \sim \text{Bin}(4, 0.25)$ .
- ▶ The probability that we get at most 3 rolls showing '4' is

$$P(Z \leq 3) = 1 - P(Z = 4) = 1 - \left(\frac{1}{4}\right)^4 \approx 0.9961.$$



## Example 5 - Genetic Traits Problem

A couple, both having black hair, carry genes that make it possible for their offspring to have black hair (probability 0.70), blonde hair (0.20), or red hair (0.10).

- (a) What is the probability that their first child will have red hair and the second will not?
- (b) What is the probability that exactly one of their two children will have red hair?
- (c) If they have six children, what is the probability that exactly two will have red hair?
- (d) If they have six children, what is the probability that at least one will have red hair?
- (e) What is the probability that the first red-haired child will be the 5th child?
- (f) Would it be considered unusual if only 2 out of their six children had black hair?

## Solution: Parts (a) and (b)

(a) First child with red hair and second without:

►  $P = 0.10 \cdot 0.90 = 0.0900$ .

(b)

► Let  $X$  denote the number of children with red hair.

► Then,  $X \sim \text{Bin}(2, 0.1)$ .

► Thus, we get  $P(X = 1) = \binom{2}{1} \cdot 0.10 \cdot 0.90 = 0.1800$

## Solution: Parts (c) and (d)

(c) Working as in the previous part, the probability is given by

$$P(Y = 6) = \binom{6}{2} \cdot (0.10)^2 \cdot (0.90)^4 \approx 0.0886,$$

where  $Y \sim \text{Bin}(6, 0.1)$ .

(d) A common trick with “at least 1” questions is to think of the complementary event. Here, we get

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - (0.9)^6 \approx 0.4686,$$

where we wrote  $P(Y = 0) = (0.9)^6$  because we want all children to not have red hair, each children does not have red hair with probability 0.9, and they are independent with each other.

## Solution: Parts (e) and (f)

(e)

- ▶ The first four children don't have red hair. The fifth one does.
- ▶ That's the case of a Geometric distribution. Here,  
 $X \sim \text{Geom}(0.1)$ .
- ▶ Hence, the probability that the first red-haired child is the 5th child is

$$P(X = 4) = (0.90)^4 \cdot 0.10 \approx 0.0656.$$

(f) Only 2 out of their 6 children having black hair:

- ▶  $E[\text{black hair}] = 6 \cdot 0.70 = 4.2$  children on average.
- ▶  $\text{Var}(\text{black hair}) = 6 \cdot 0.70 \cdot 0.30 = 1.26$ .
- ▶  $\sigma(\text{black hair}) = \sqrt{1.26} \approx 1.12$ .
- ▶ Observing only 2 black-haired children is 2.2 standard deviations below the mean, thus it would be considered unusual.

## Example 6 - The Birthday Problem

- ▶ The STOR 155.007 class has 46 students.
- ▶ What is the probability that at least 2 students share the same birthday?

## Solution

- ▶ We can think of this as the complement of the probability that there are no matches among the 46 students.
- ▶ We compute

$$\begin{aligned} &P(\text{no matches among 46}) \\ &= 1 \cdot \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdot \dots \cdot \left(1 - \frac{45}{365}\right) \\ &= \frac{365 \cdot 364 \cdot \dots \cdot 320}{365^{46}} \\ &= \frac{365!}{365^{46} \cdot (365 - 46)!} \approx 0.0833 \end{aligned}$$

- ▶ Thus,

$$P(\text{at least 1 match among 46}) \approx 1 - 0.0833 = 91.67\%.$$