ECE750: Usable Security and Privacy

Study Structure

Dr. Kami Vaniea
Electrical and Computer Engineering
kami.vaniea@uwaterloo.ca

First, the news...

- First 5 minutes we talk about something interesting and recent
- You will not be tested on the news part of lecture
- You may use news as an example on tests
- Why do this?
    1. Some students show up late for various good reasons
    2. Reward students who show up on time
    3. Important to see real world examples

PLANNING A STUDY

Don’t panic! This is not a statistics class.

Could be on the exam

- Independent and dependent variables
- Correlation vs causation
- Between vs within subject design
- Study question design

Will <u>not</u> be on the exam

- Statistical test names
  - T-test, ANOVA, etc.
- When to use different tests
  - Chi Sq should be used with categorical dependent and independent variables
- P-values, distributions, confidence intervals or other outcomes from tests

What kind question are you asking?

- Attitudinal – User attitudes and opinions

vs.

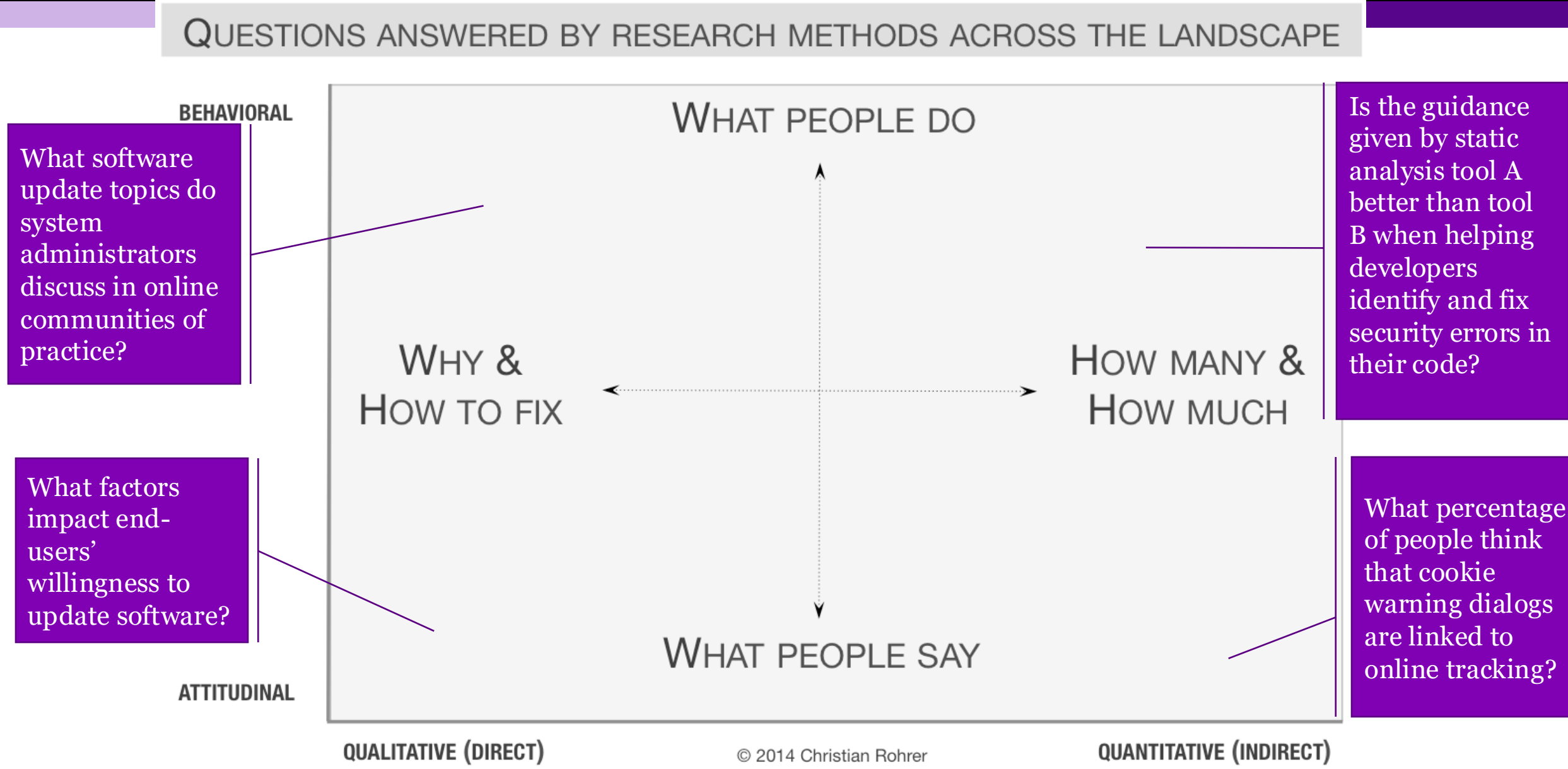
- Behavioral – What the user actually does or is capable of doing

- Qualitative – Unstructured data. Typically unstructured language data

vs.

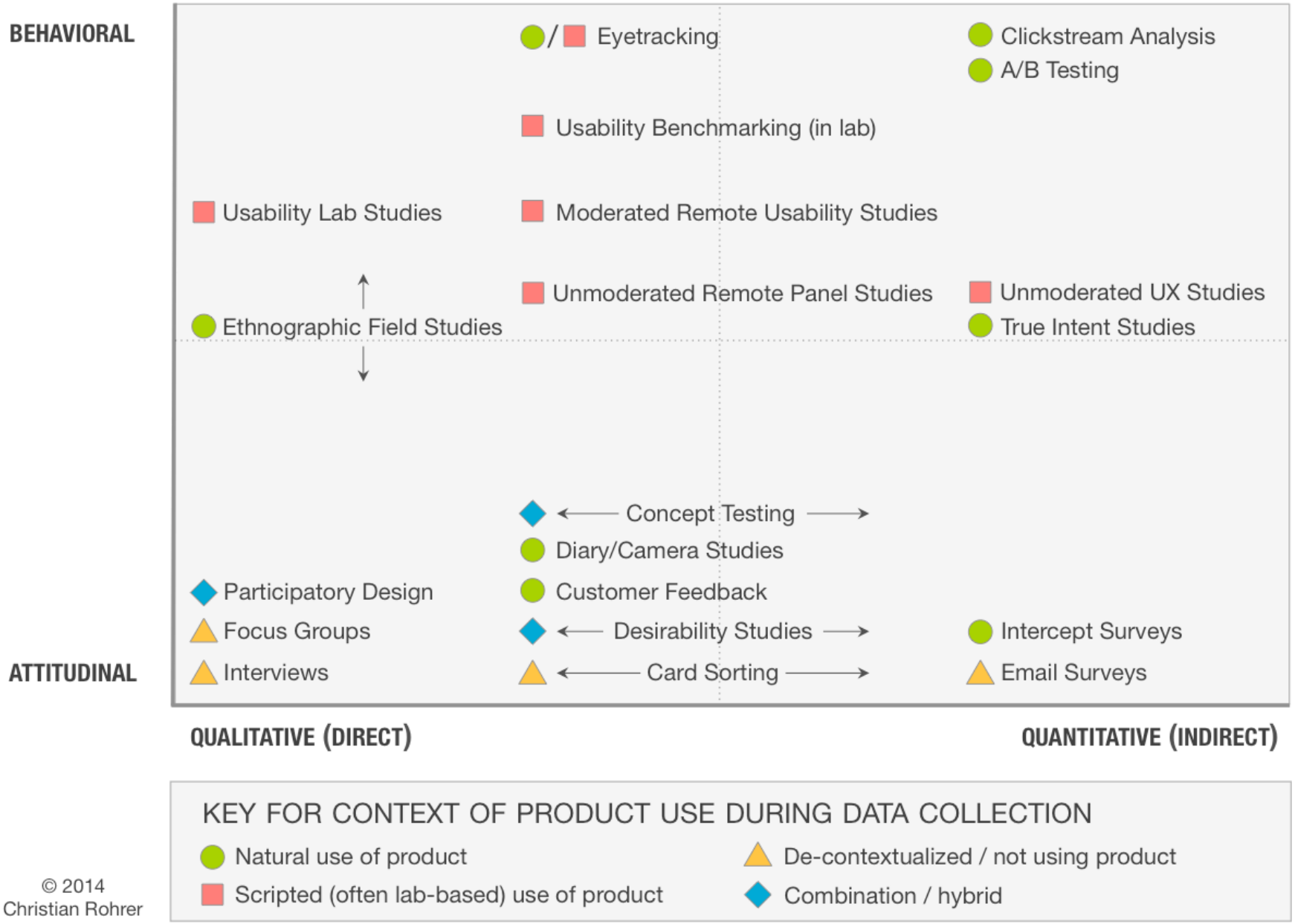
- Quantitative – Structured data. Typically numerical data that can be summed or counted

https://www.nngroup.com/articles/which-ux-research-methods/



https://www.nngroup.com/articles/which-ux-research-methods/

A Landscape of User Research Methods



https://www.nngroup.com/articles/which-ux-research-methods/

Planning a study

- Studies normally answer multiple research questions. With each research question tied to one or more aspects of the study, such as survey questions.
- Descriptive – learn something about the whole population.
  - How many people have heard of the term “phishing”?
  - What words do people use to describe cookie tracking?
- Testing for correlation or causation – show that two things are related or one thing causes the other thing.
  - If someone has been trained on phishing in the past, are they better at differentiating phishing emails?
  - We have three training options, each user goes through one training, which training causes people to identify phishing emails the best?

Descriptive Statistics

- Descriptive Questions – learn something about the whole population.
  - How many people have heard of the term “phishing”?
  - What words do people use to describe cookie tracking?
- Descriptive Numeric – fancy term for all the basic measures of numeric data: Mean, median, mode, standard deviation
  - What % of consumers are worried about privacy?
  - What % of people know the difference between behavioral advertising and cookies?
  - On average, how long does it take to decide if an email is phishing or not?
- Descriptive Qualitative – use data to learn about a studied population
  - What is the most common reason people avoid using ATMs?
  - Why do some people choose to not have a Google account?

Testing for correlation or causation

- Testing for correlation or causation – show that two things are related, or that one thing causes the other thing.
  - If someone has been trained on phishing in the past, are they better at differentiating phishing emails?
  - We have three training options, each user goes through one training, which training causes people to identify phishing emails the best?
- These tests require more complex statistics, such as:
  - T-test
  - ANOVA
  - Linear Models
  - CHI Squared

Topics Outline

- Descriptive questions vs testing a question
- **Correlation vs causation**
- Dependent vs independent variables
- Between and within subjects testing
- Numeric vs categorical data

Correlation vs. Causation

- Correlation
  - Two things tend to behave in a way that seems inter-related, where if one thing changes the other thing will also change in a related way.
  - For example, if the price of rice goes up at the same time as the price for beans.
- Causation
  - When one thing changes it causes the other thing to change.
  - For example, when the weather gets cold more people wear coats. Cold weather causes more people to wear coats.

Think-pair-share

For each of the following, is correlation or causation being measured?

- Lab study comparing two disk encryption tools. Users are randomly assigned to one of 2 tools then asked to encrypt everything in a specific folder, play a game, and then decrypt everything. The number of errors made are measured.
  - RQ: Which disk encryption tool leads to less errors?
- Interview study looking at social media non-use. Users are asked to select from a list of reasons they might choose to not use social media. They are also asked what percentage of their friends use social media.
  - RQ: Do people with high percentage of friends using social media choose to not use social media for different reasons than those with low percentages of friends using social media?
- Study of hard drives purchased off eBay. Researchers examined the drives to learn what operating system was installed and what percentage of each drive was encrypted.
  - RQ: Does operating system choice lead to more encryption?

Topics Outline

- Descriptive questions vs testing a question
- Correlation vs causation
- **Dependent vs independent variables**
- Between and within subjects testing
- Numeric vs categorical data

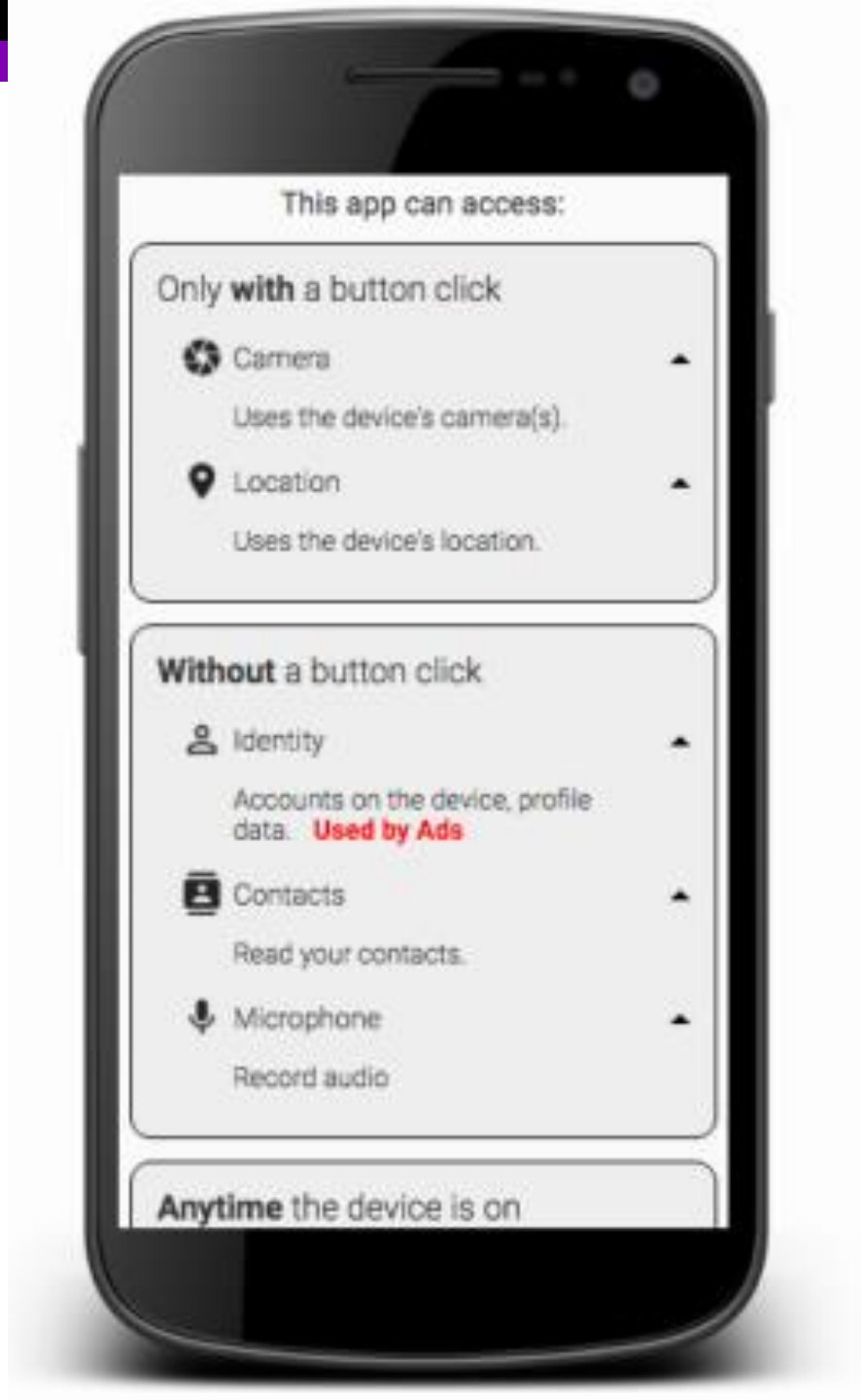
What are you going to measure?

- In statistics there are classically two types of measurements (variables): dependent and independent
- Dependent
  - Also known as the outcome variable
  - “Dependent” on the study
  - Measures the usability goal
- Independent
  - Anything you are directly manipulating
  - An element of the study which is under your control
  - A pre-existing feature of your participant

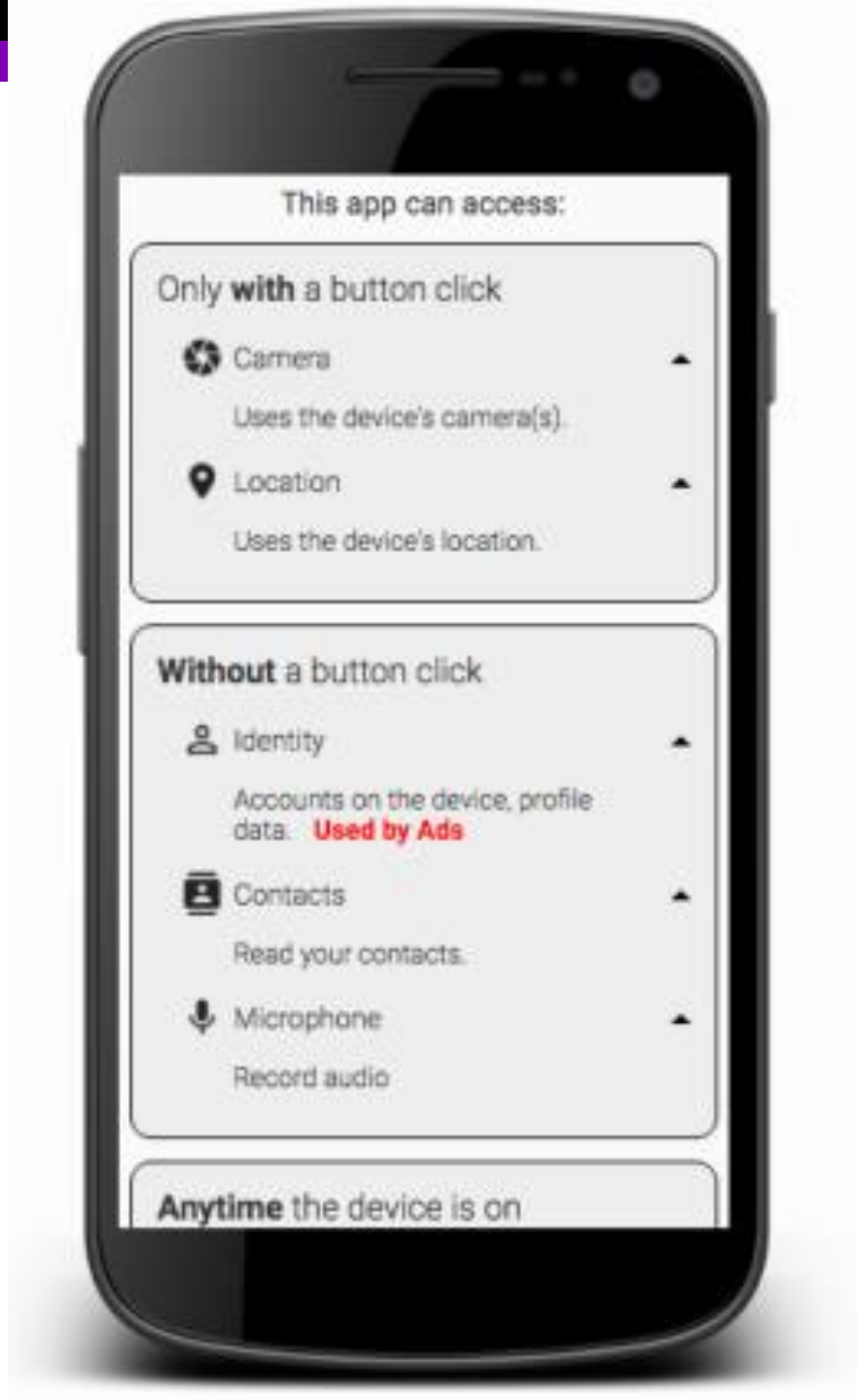
Some of my recent research questions:

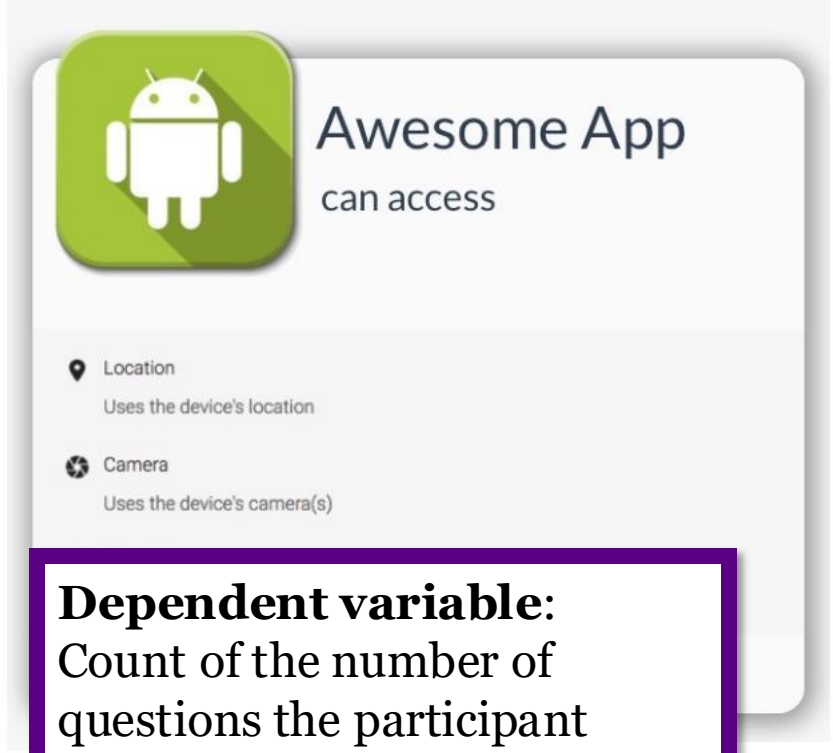
- Can people differentiate between a subdomain and a domain when reading a URL?
- Does [my new system] help people differentiate between malicious URLs and safe ones?
- Can users use [my new password manager] faster and with less errors than [the old password manager]?
- Does knowing how an app will use its permissions impact app installation decisions?
- Using [website], can users successfully opt-out of cookie tracking without forming inaccurate mental models?

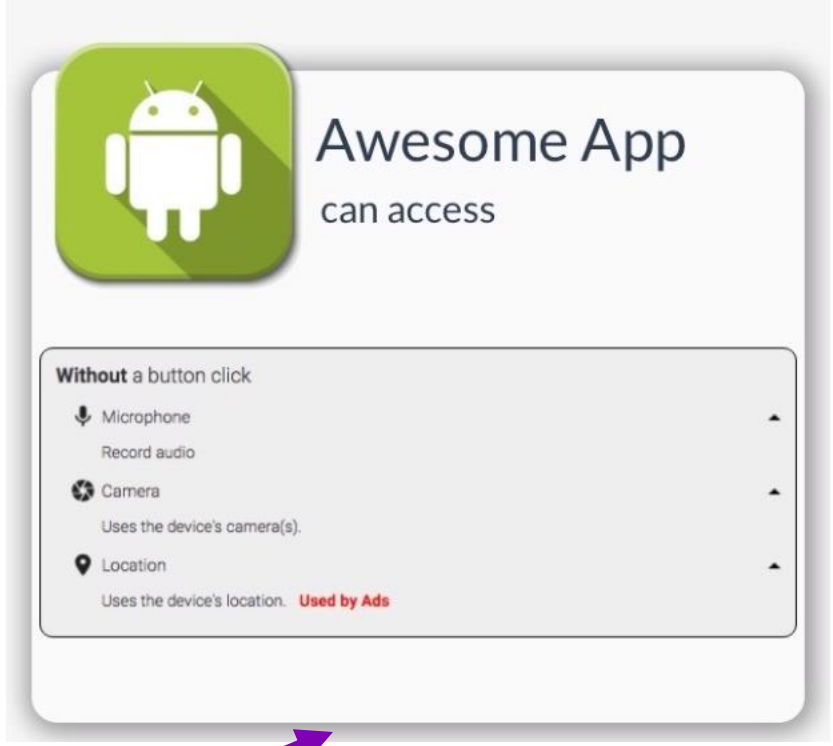
Lets use this study as an example



Research Question:
Can users reliably identify if an app can or cannot perform an action directly tied to a permission.







Dependent variable: Count of the number of questions the participant answered correctly

Independent variable: Which of the two interfaces the participant was shown

ing can this app do?

					Absolutely Possible
Charge purchases to your credit card at any time.					○
Get your location.	○	○	○	○	○
Allow ads to know your location.	○	○	○	○	○
Load ads.	○	○	○	○	○
Write on the SD card					

Variables that would make sense

- Research Question: Can users reliably identify if an app can or cannot perform an action directly tied to a permission?
- Dependent
  - Which permissions correctly/incorrectly read
  - Count of permissions correctly/incorrectly read
  - Time spent reading each permission screen
- Independent
  - Study group (which screen was shown)
  - If the permission was privacy sensitive or not
  - Order of the tasks
  - Time of day
  - Type of most used device (laptop, mobile, PC)
  - Demographics of the participants (gender, age, native language, …)

Common dependent things to measure

- Number of dangerous errors made
- Time to complete task
- Percent of task completed
- Percent of task completed per unit of time
- Ratio of successes to failures
- Time spent in errors
- Percent or number of errors
- Percent or number of competitors better than it
- Frequency of help and documentation use

Topics Outline

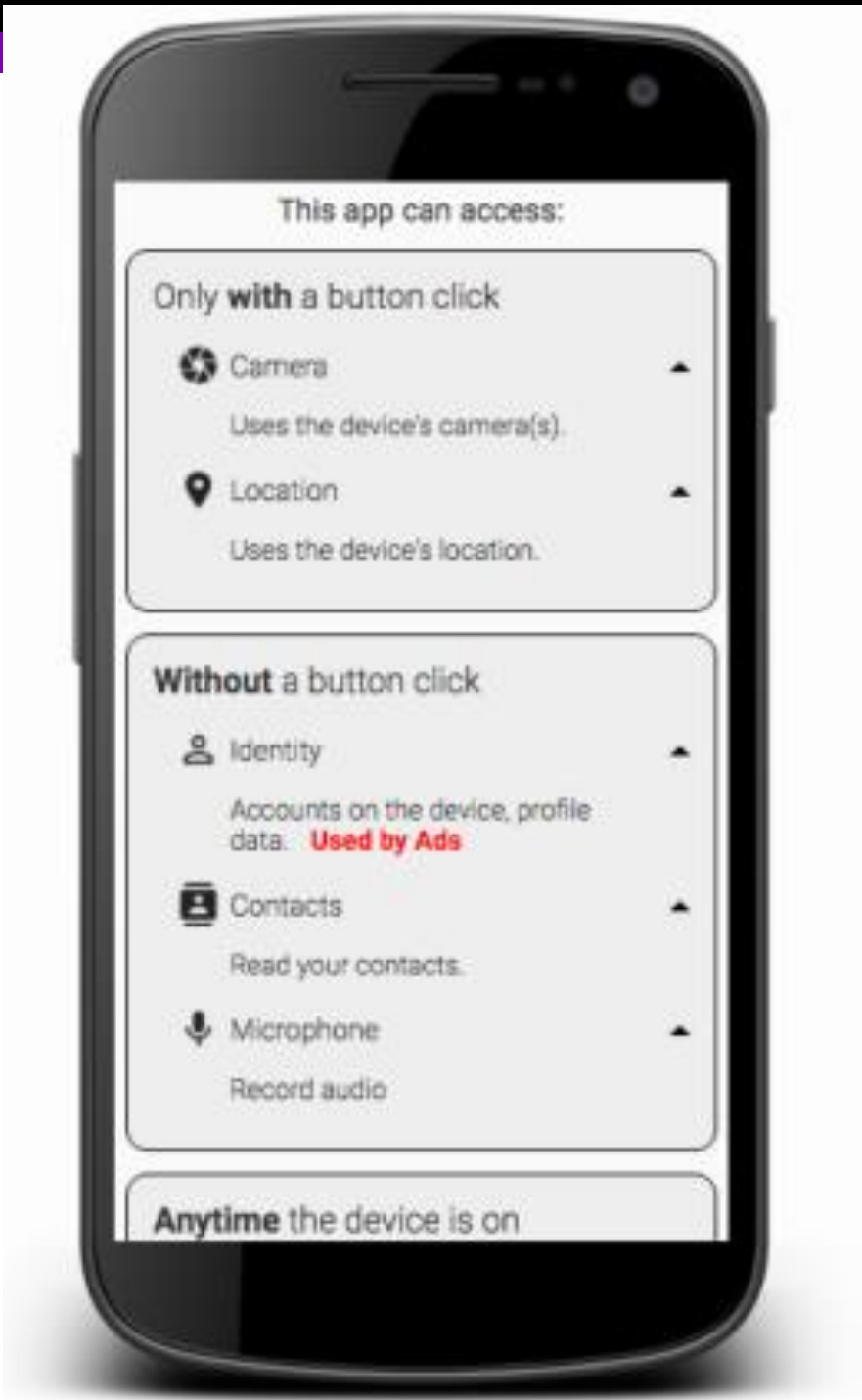
- Descriptive questions vs testing a question
- Correlation vs causation
- Dependent vs independent variables
- **Between and within subjects testing**
- Numeric vs categorical data

Between vs. Within subjects

- Between subjects
  - Your study only shows one interface to one person
  - You are measuring how well the people randomly assigned to the A interface did compared to the people randomly assigned to the B interface
  - Lots of variability with this method
- Within subjects
  - Your study shows all interfaces to all people
  - You are measuring the difference in how they do on the two interfaces
  - Less variability (same person) but more learning effects and priming

Study design

- RQ: Does [my new interface] enable people to accurately determine what permissions an app will use?
- A/B test between the existing and new interface
- Between subjects
- 10 Tasks shown in the same order to all participants
- Dependent variables
  - Accuracy on task
- Independent variables
  - Which interface (A or B)



Topics Outline

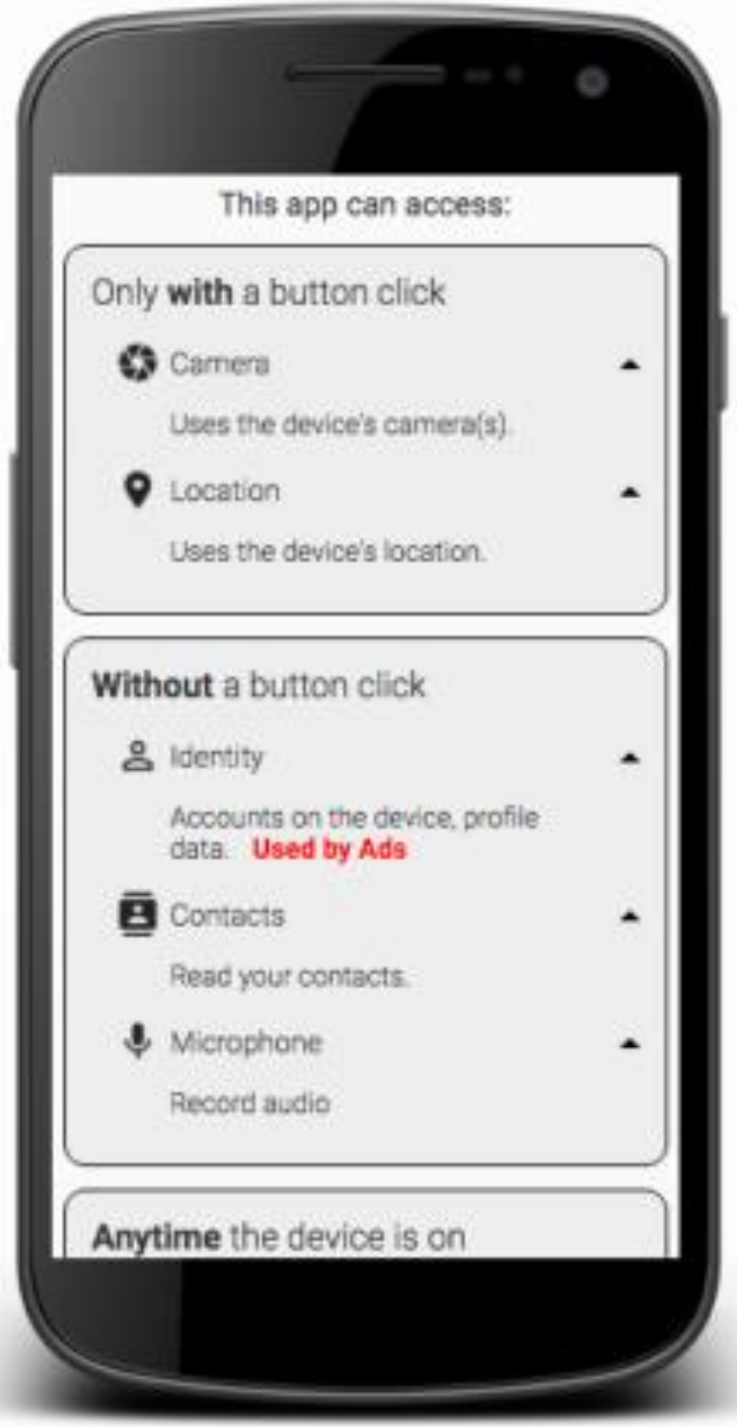
- Descriptive questions vs testing a question
- Correlation vs causation
- Dependent vs independent variables
- Between and within subjects testing
- **Numeric vs categorical data**

Types of data

- Numeric
  - **Continuous** – Any value on the range is possible including decimal (1-5)
  - **Discrete** – Only certain values on the range are possible (1,2,3,4,5)
  - **Interval** – Only certain values on the rage are possible and each has equal distance from its neighboring values (strongly agree, agree, neutral, disagree, strongly disagree)
- Categorical
  - **Binary** – Only two possibilities (true, false)
  - **Ordinal** – The values have an ordering (slow, medium, fast)
  - **Nominal** – The values have no ordering (apple, pear, kiwi, banana)

Study design

- Accuracy on all tasks
  - Discrete
- Which interface
  - Categorical binary



STATISTICAL TESTS

Comparing	Dependent	Independent	Parametric (Dependent variable is mostly normally distributed)	Non-parametric
The means of two independent groups	Continuous / scale	Categorical / nominal	Independent t-test	Mann-Whitney test
The means of 2 paired (matched) samples	Continuous / scale	Time variable (before/after)	Paired t-test	Wilcoxon signed rank test
The means of 3+ independent groups	Continuous / scale	Categorical / nominal	One-way ANOVA	Kruskal-Wallis test
3+ measurements on the same subject	Continuous / scale	Time variable	Repeated measures ANOVA	Friedman test
Relationship between 2 continuous variables	Continuous / scale	Continuous / scale	Pearson's Correlation Coefficient	Spearman's Correlation Co-efficient
Predicting the falue of one variable from the value of a predictor variable	Continuous / scale	Any	Simple Linear Regression	
Assessing the relationship between two categorical variables	Categorical / nominal	Categorical / nominal		Chi-squared test

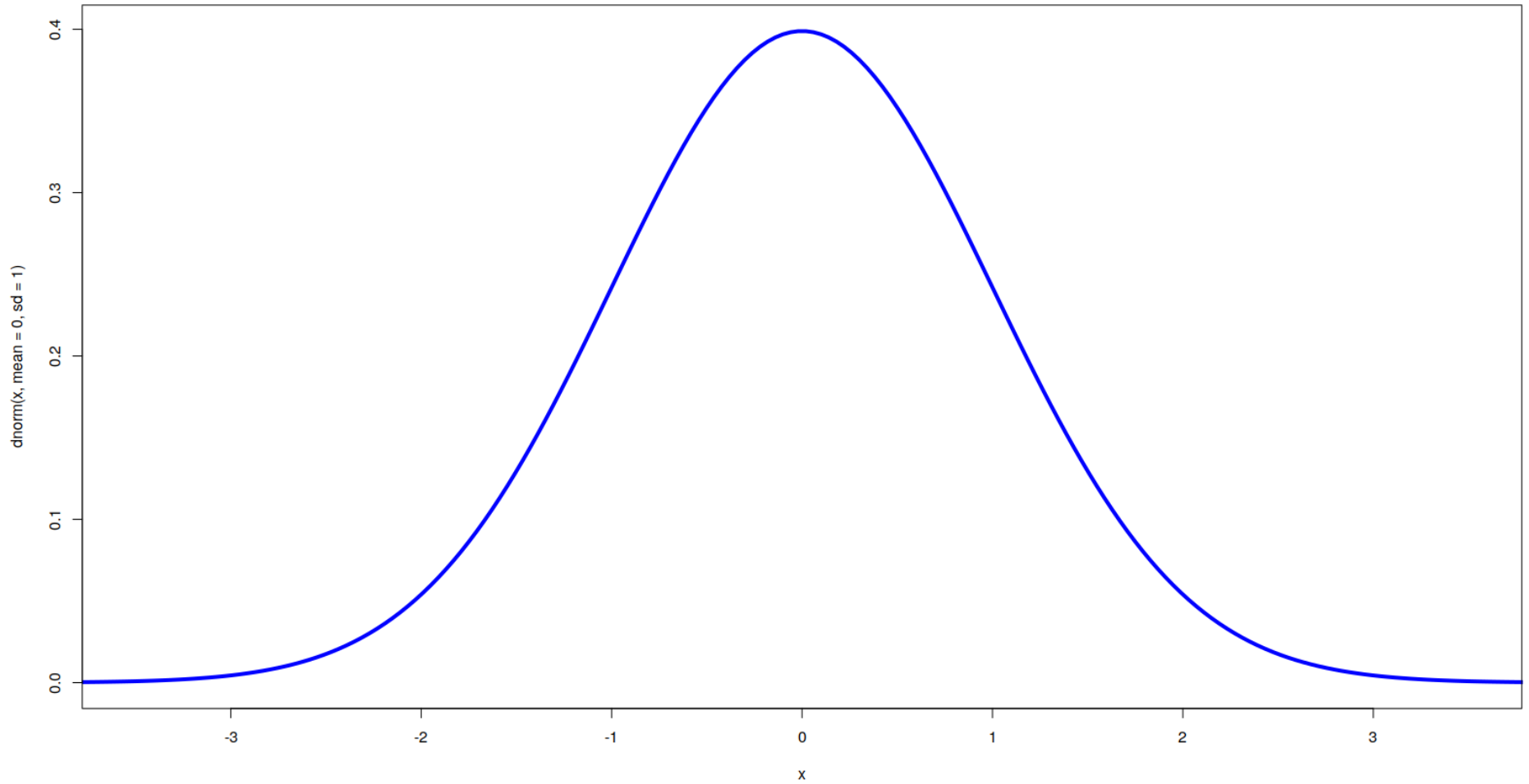
t-test: Test if two groups have the same mean (average)

T-test requires:

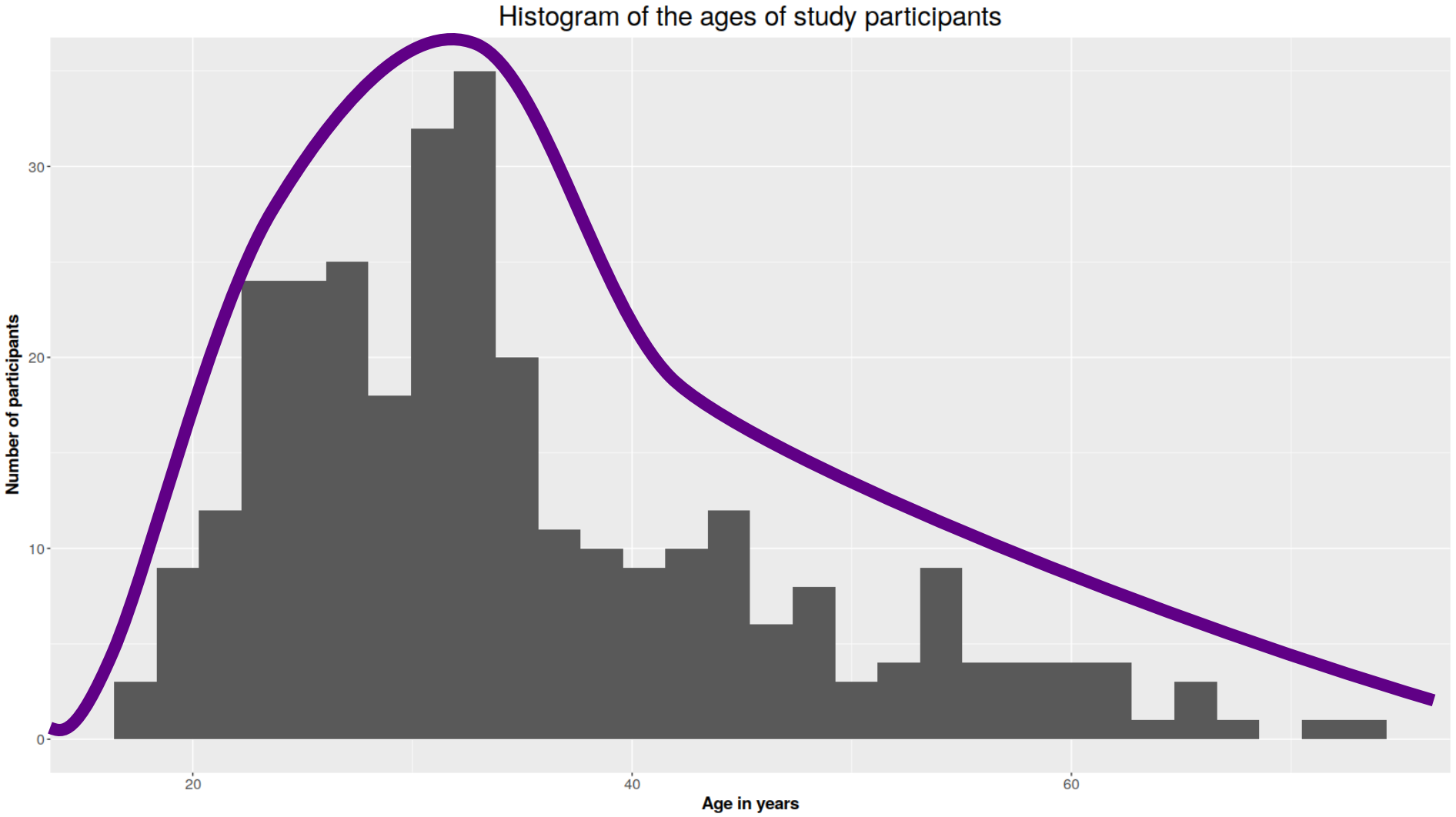
- Independent variable: categorical binary
- Dependent variable: numeric (continuous or discrete)

- Data must be normally distributed

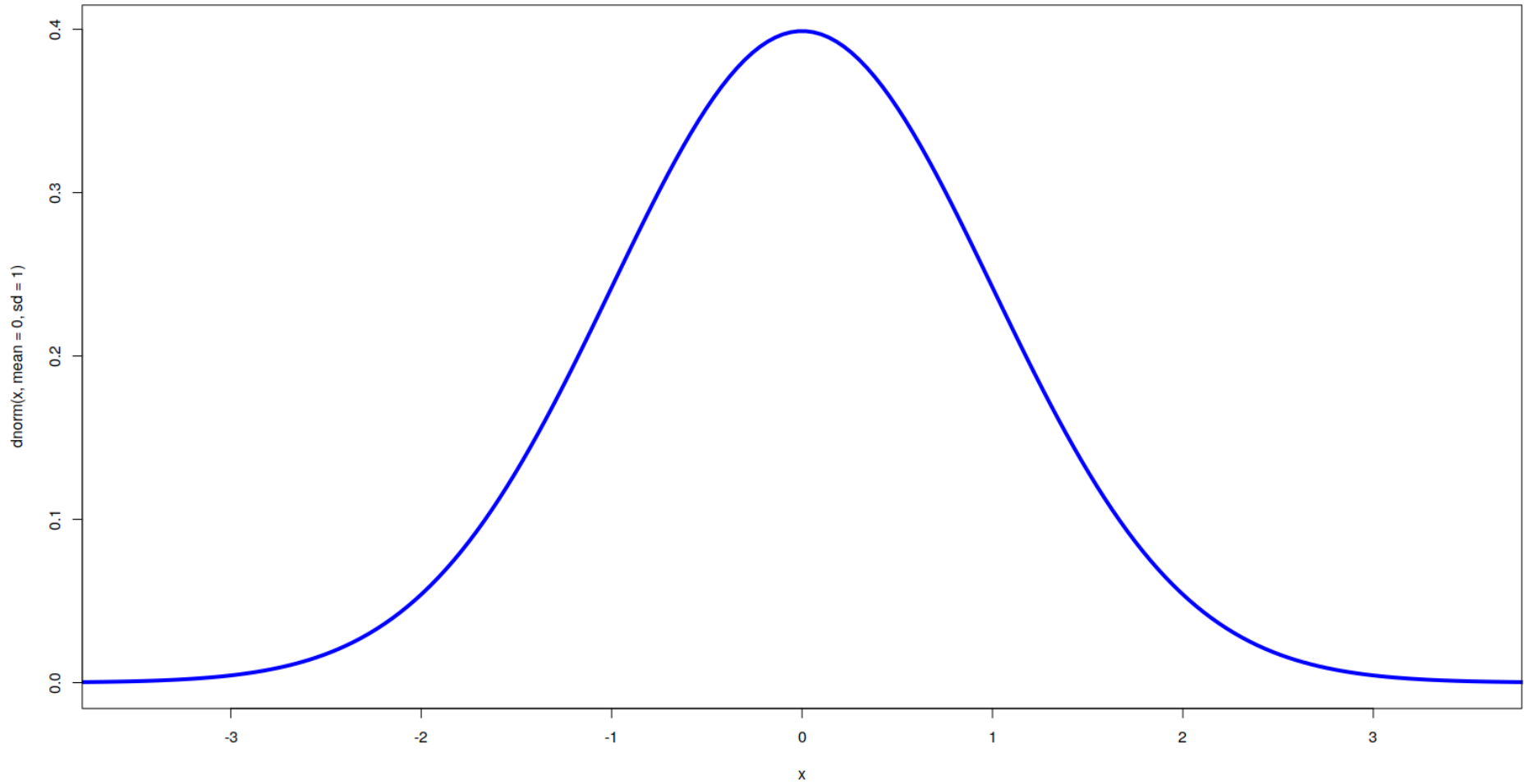
Normal distribution



Real data is messy

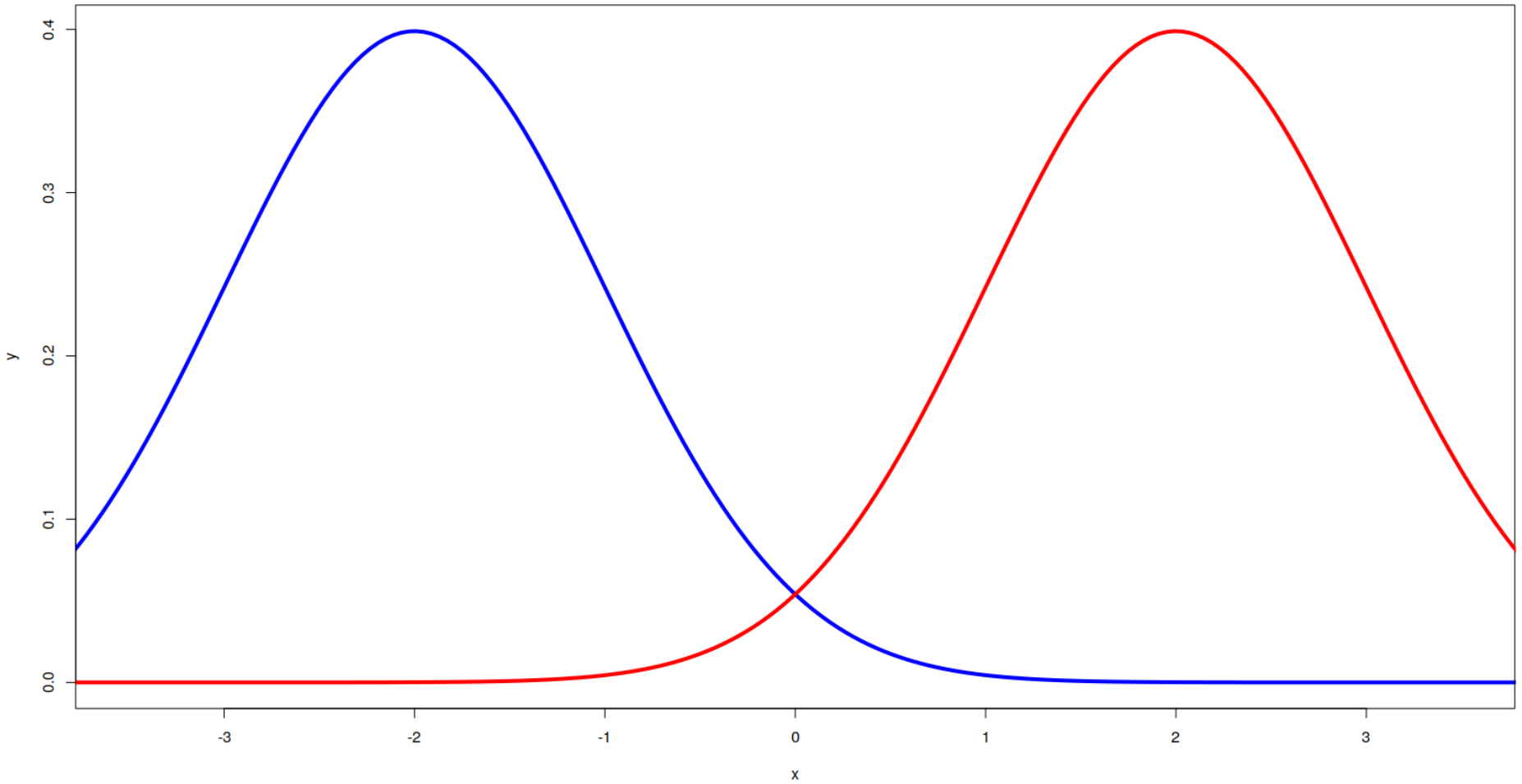


Normal distribution

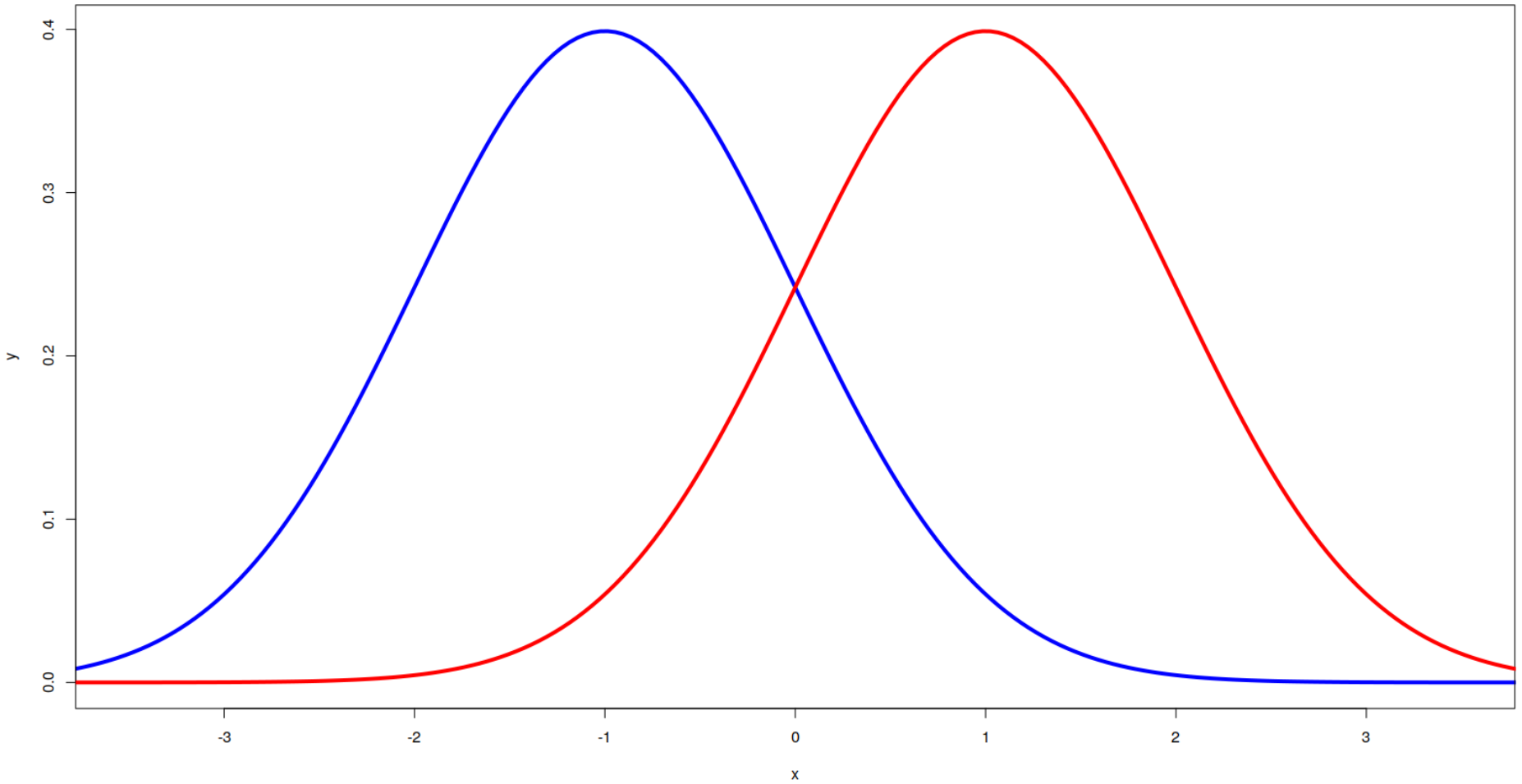


T-test: Do two populations have the same mean?

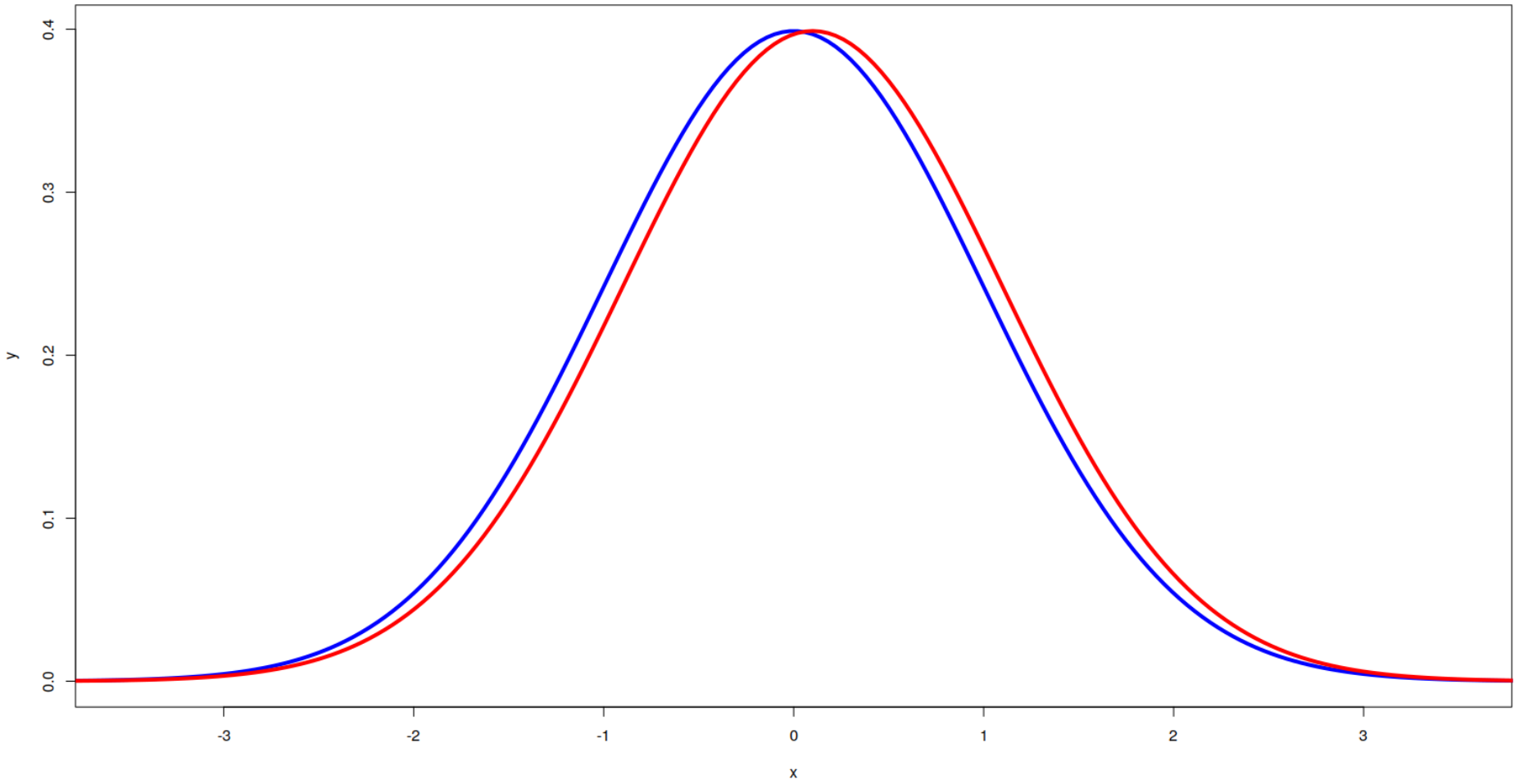
Different means



Maybe? different means



Likely true mean is the same



I showed participants 4 code samples and asked them what the code would do. I then asked them how confident they were in their answer.

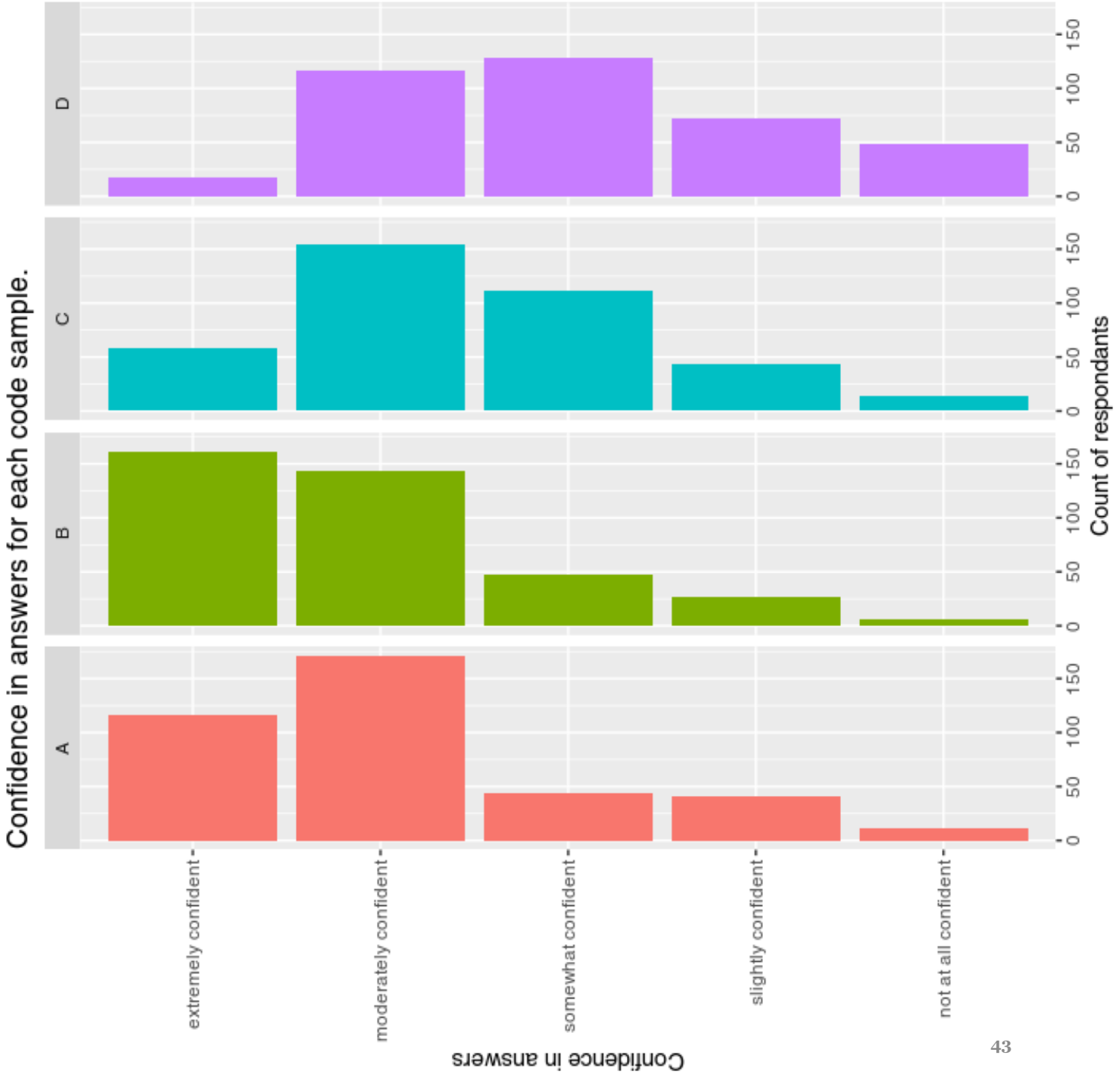
Research Question: Does the code sample shown impact confidence in their answer?

Research Question:
Does the code sample shown impact confidence in their answer?

Within-subjects

Independent:
Which code sample shown

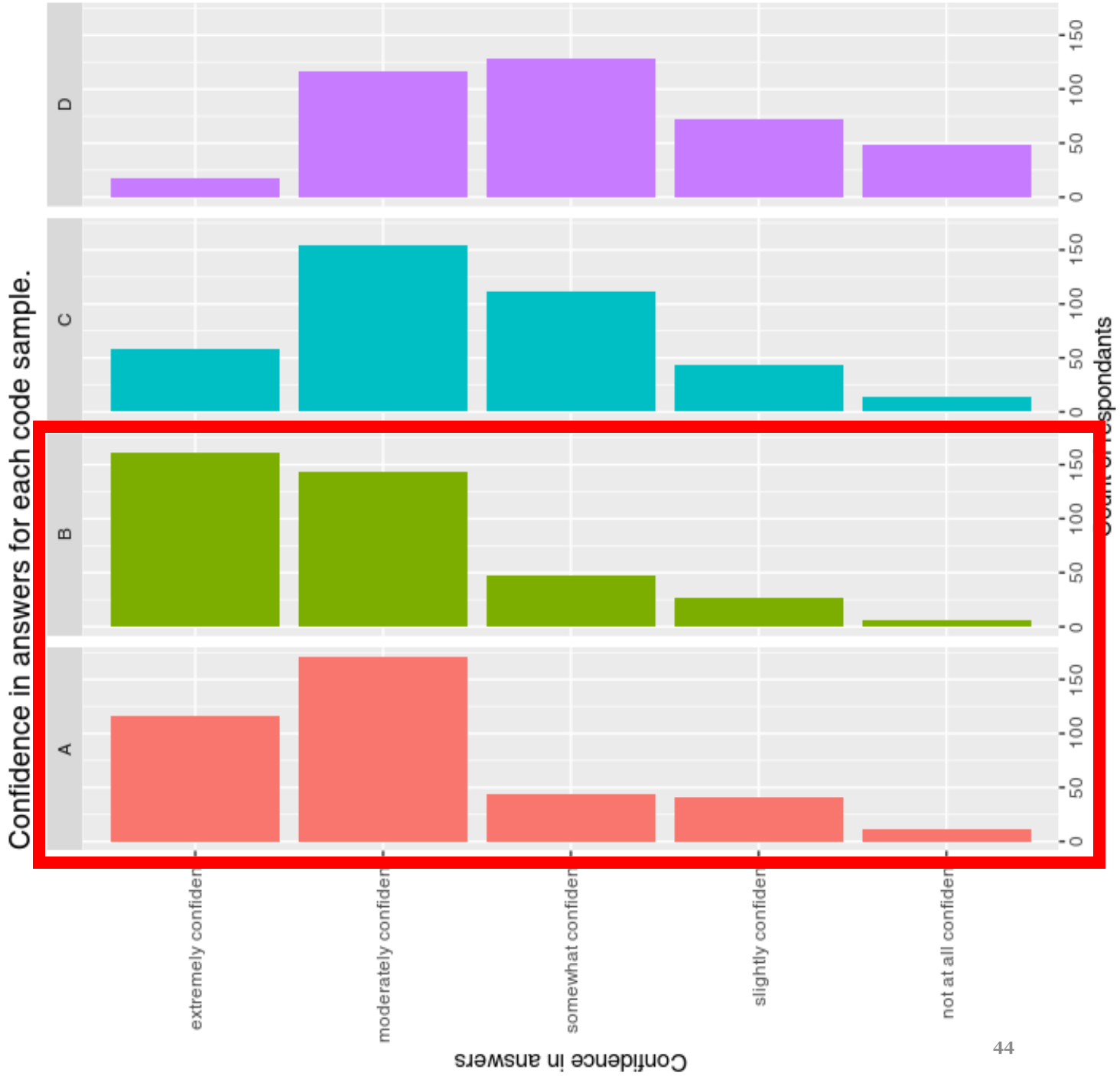
Dependent: Confidence



Problem: My categorical variable (code sample) is not binary, there are 4 levels.

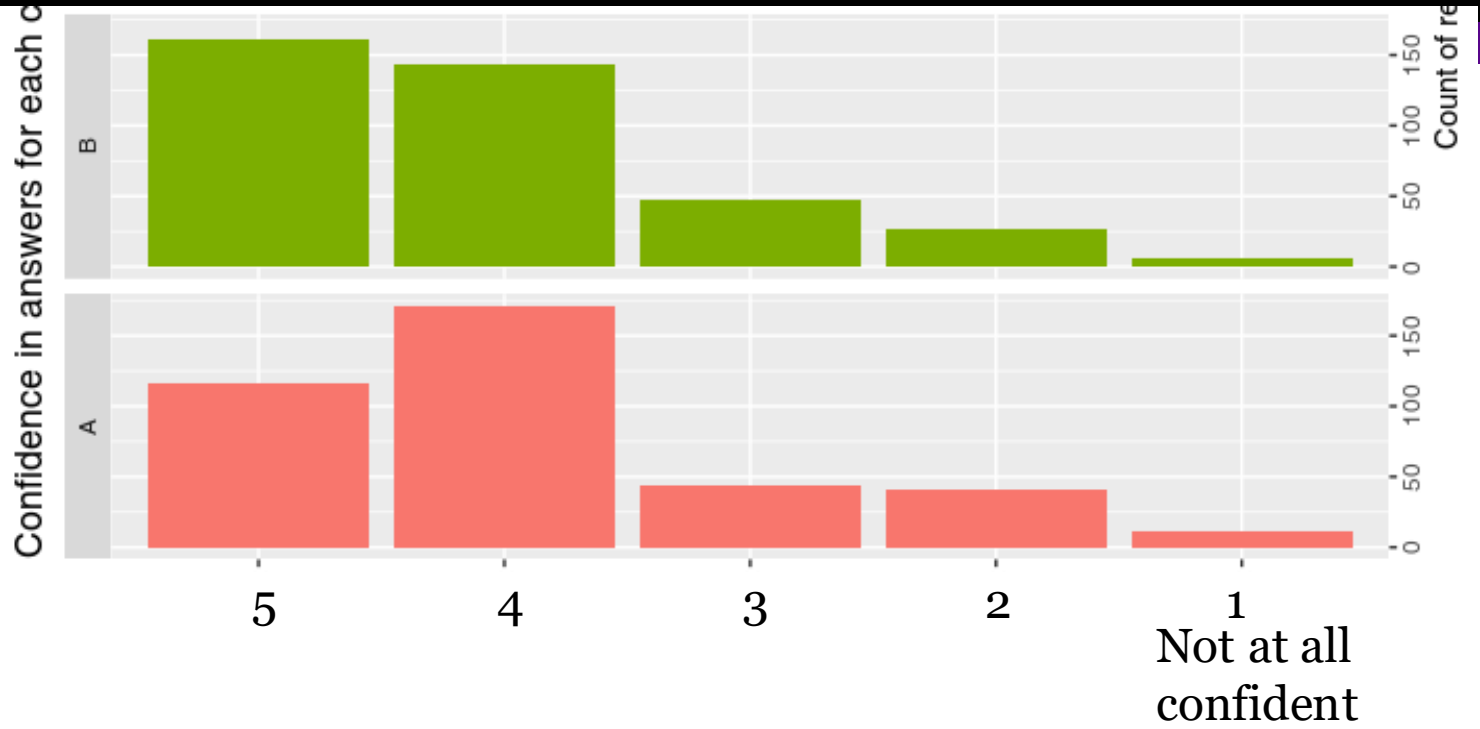
Solution: Run the t-test on each pair. So test A vs B, A vs C, …. C vs D.

Real solution: Use an ANOVA (not covered in this class)



Running the t-test

- This is a "**within** subjects" test where one person gave a confidence answer for both Code Sample A and Code Sample B
  - So we use a **Paired t-test**
- Create two arrays (or Excel columns) one with Code Sample A confidence, the other with Code Sample B confidence
- Two-sided (tailed)
  - For now, just do this. I don't have time to explain.
- Alpha of 0.05
  - p-value needs to be less than 0.05 to show that the two code samples produce different levels of confidence
  - Means that 5% of the time we will get the wrong answer from the statistical test

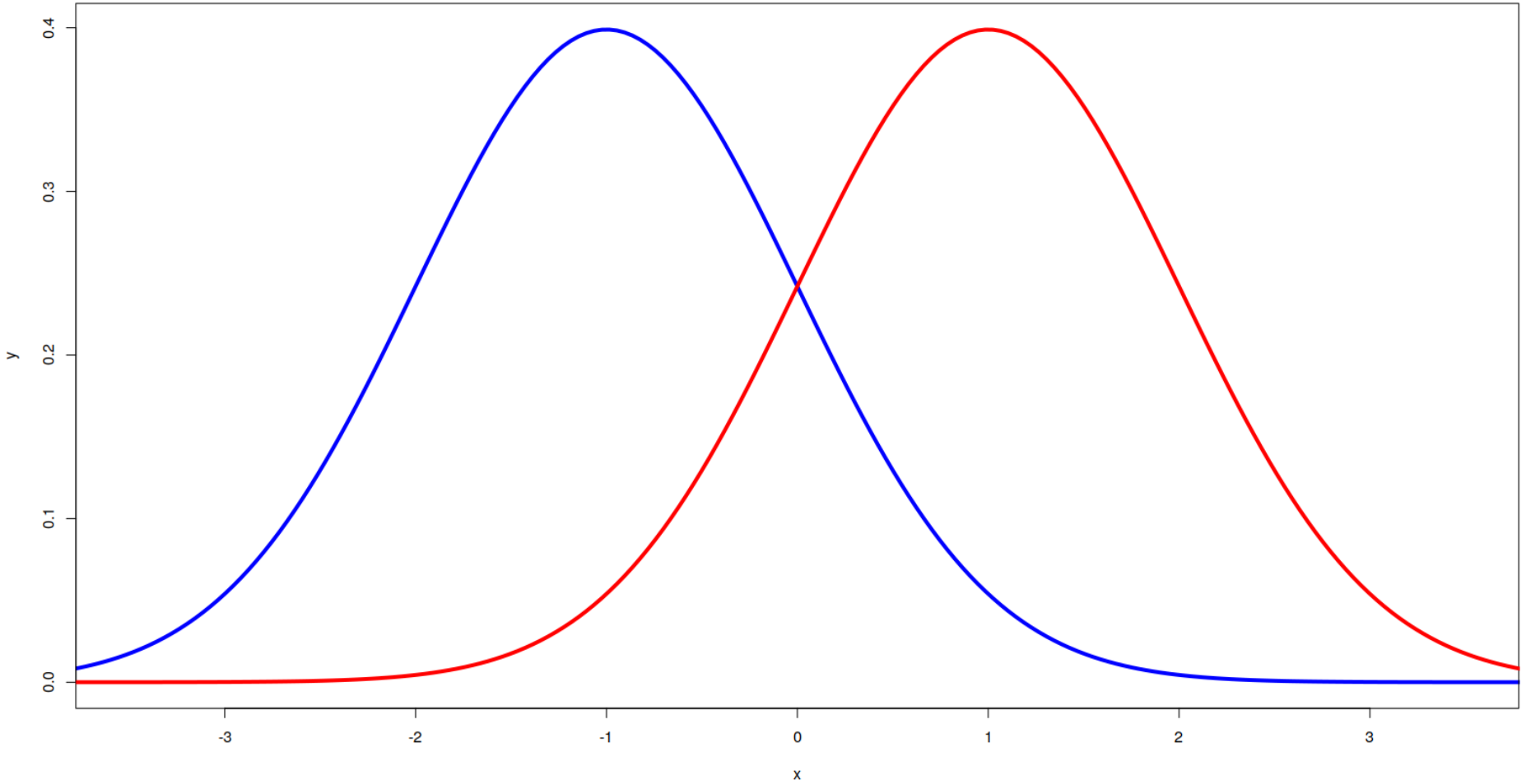



```
> t.test(a.confidence,b.confide

        Paired t-test

data:  a.confidence and b.confidence
t = -5.2699, df = 383, p-value = 2.285e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3218198 -0.1469302
sample estimates:
mean of the differences
                -0.234375
```

Different means, small difference



I ran a survey to learn about software update behaviors.

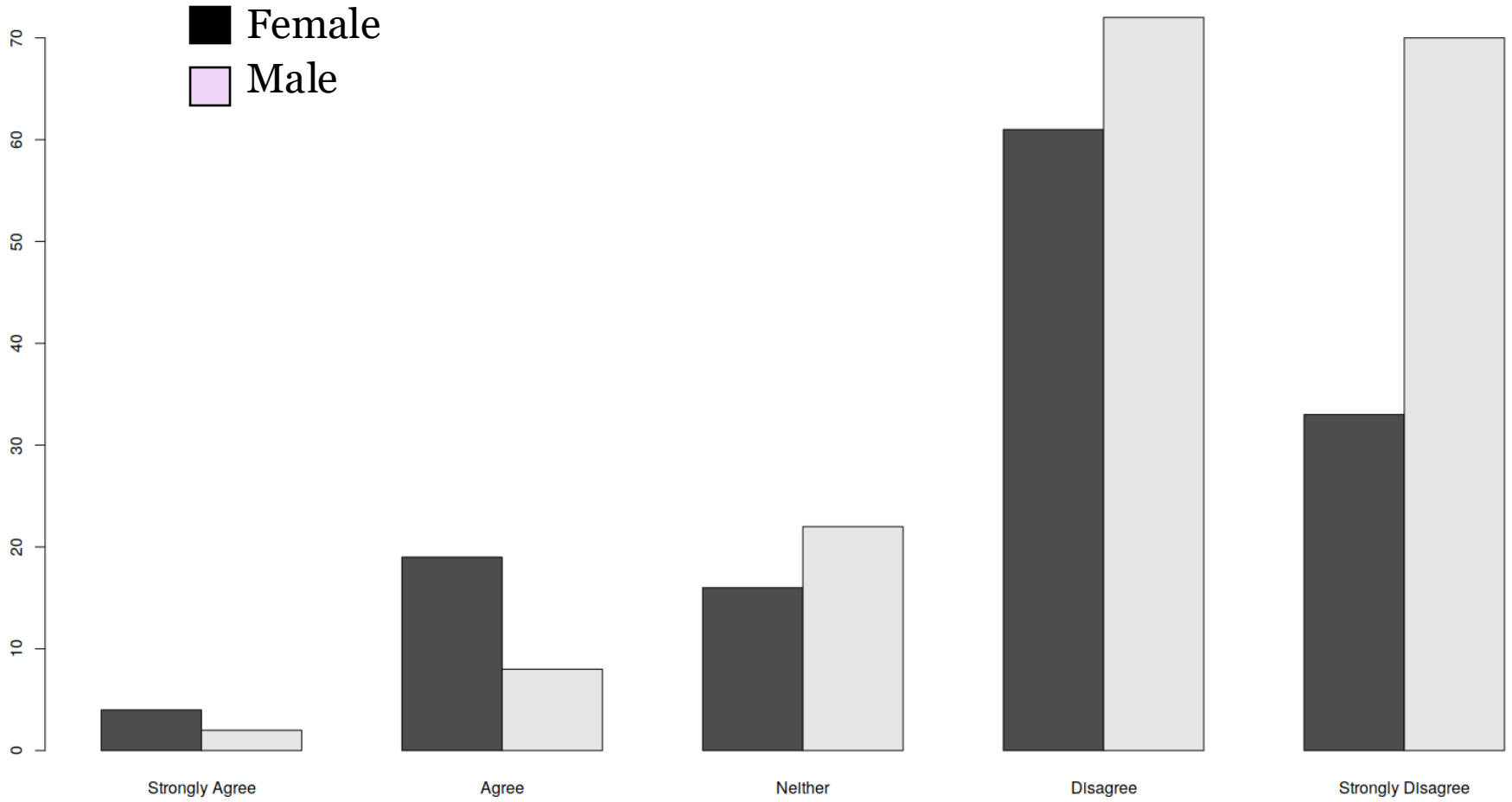
Research Question: Do women and men feel like they ask others for technical help with different frequency?

Research Question: Do women and men feel like they ask others for help with different frequency?

Between-subjects

Independent: Gender

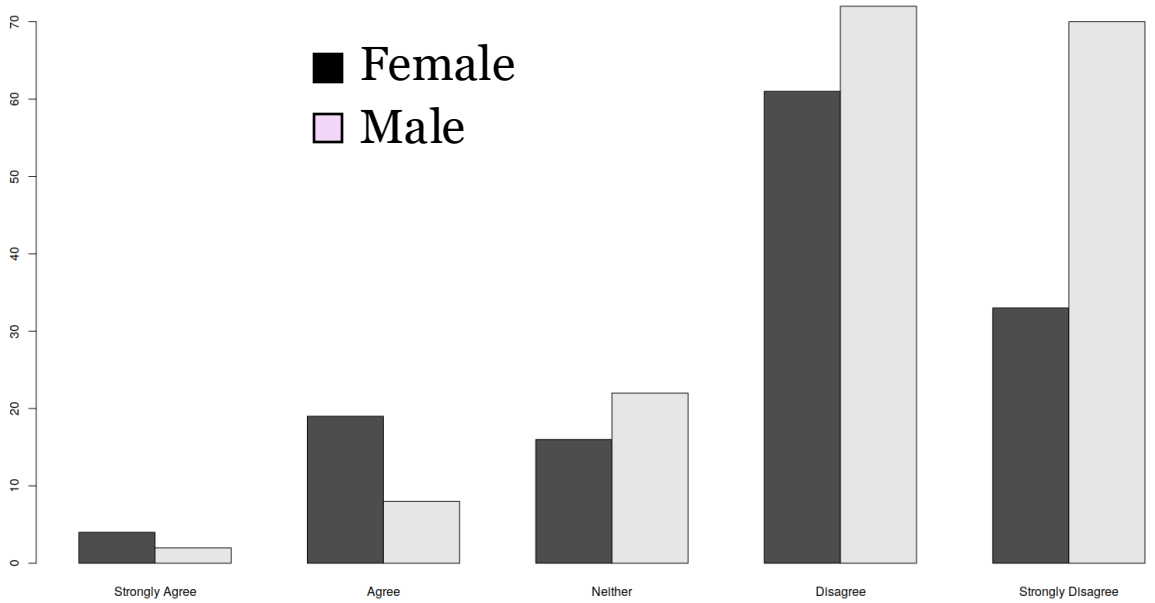
Dependent: Agreement



I often ask others for help with technical questions

Running the t-test

- This is a "**between** subjects" test where each person gave only one answer
  - So we use a **normal t-test** (not paired)
- Create two arrays one with women's responses, one with men's
- Two-sided (tailed)
  - For now, just do this. I don't have time to explain.
- Alpha of 0.05
  - p-value needs to be less than 0.05 to show that the two genders produce different levels of confidence
  - This choice means that 5% of the time we will get the wrong answer from the statistical test



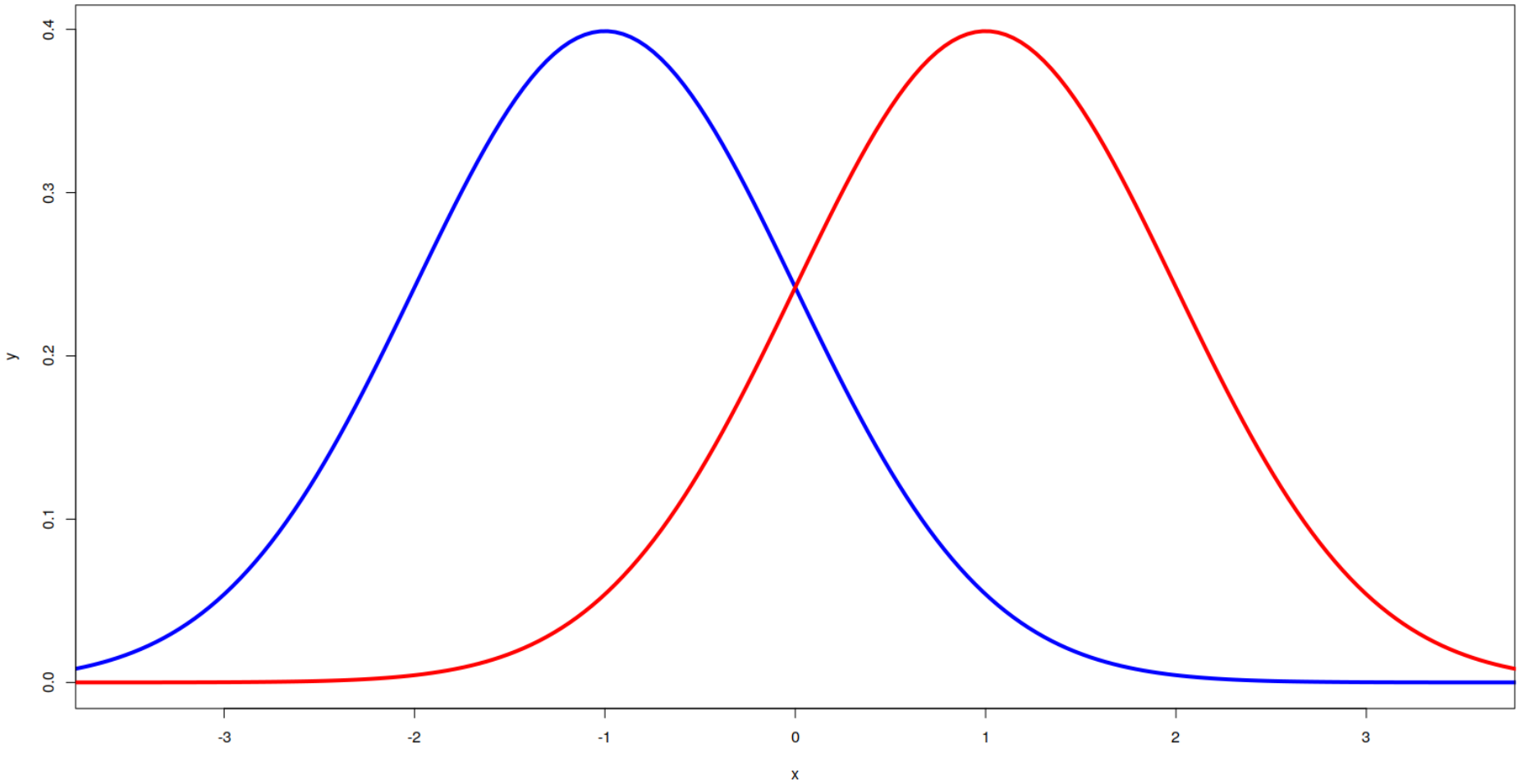
1 5

```
> t.test(as.numeric(d$i_ask_others_for_he

        Welch Two Sample t-test

data:  as.numeric(d$i_ask_others_for_help[d$gender == "Female"]) and
ender == "Male"])
t = -3.4481, df = 253.99, p-value = 0.0006606
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6245978 -0.1704934
sample estimates:
mean of x mean of y
 3.751880  4.149425
```

Maybe? different means

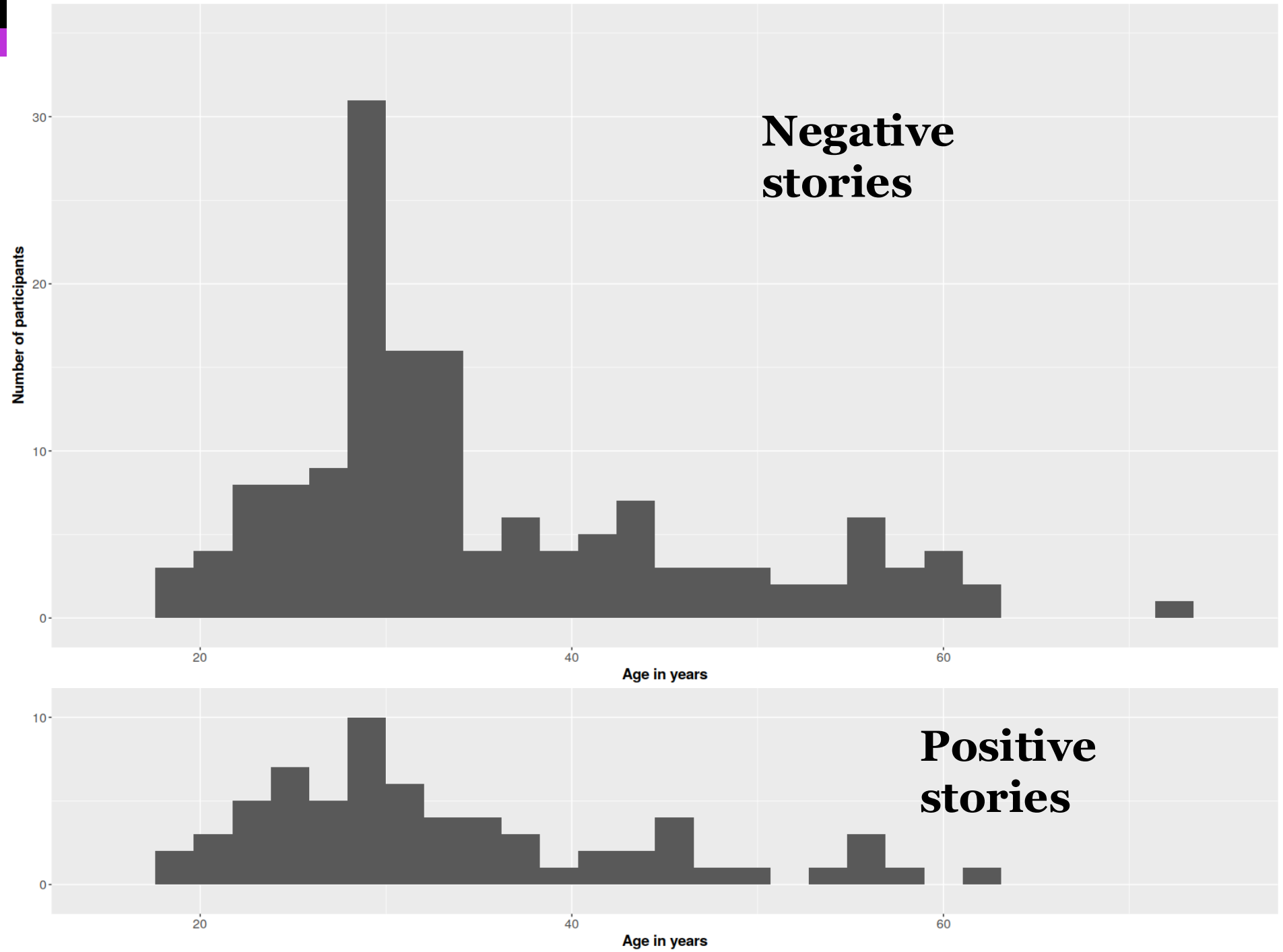


I asked participants to tell me a story about a prior software update.

Research Question: Are people who relate positive stories older or younger?

Research Question:
Are people who relate positive stories older or younger?

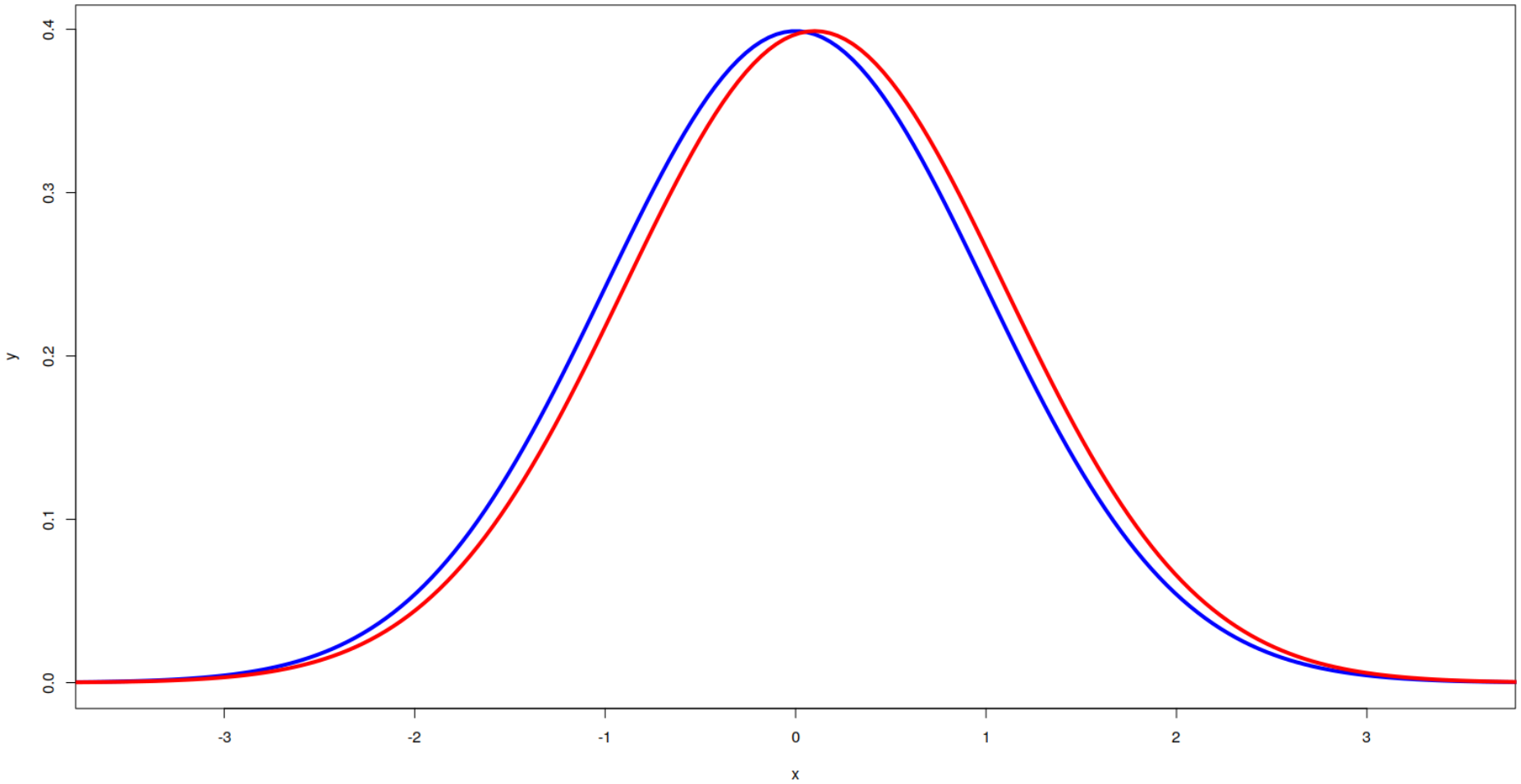
Between-subjects
Dependent:
- Age
- Numerical
Independent:
- Negative or Positive
- Binary



```
> t.test(s_neg$age, s_pos$age)

        Welch Two Sample t-test

data:  s_neg$age and s_pos$age
t = 0.75677, df = 123.07, p-value = 0.4506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.063833  4.618658
sample estimates:
mean of x mean of y
 35.42667  34.14925
```


0.4
0.3
0.2
0.1
0.0
y
-3
-2
-1
0
1
2
3
x

Questions