Survey and Analysis

INFR11158/11230 Usable Security and Privacy

Dr. Jingjie Li

07/02/2025

THE UNIVERSITY *of* EDINBURGH

Nice paper blog structure :)

The University of Edinburgh

Review of " Nod to Auth: Fluent AR/VR Authentication with User Head-Neck Modeling"[1]

Summary:

The reviewed paper introduces "Nod to Auth"[1], an authentication mechanism that is aimed towards making the verification process of shared AR/VR devices as seamless and time-efficient as the "Slide-to-Unlock" principle used in smartphones. The mechanisms relies on straightforward head gestures, such as nodding, to unlock the device in use and it achieves this goal by extending the use case of Inertial Measurement Unit (IMU) sensors which are already present in AR/VR headsets. It extracts biometric features from the head and neck of the users, which are then used for the identification sub-process via machine learning. With an impressive user identification accuracy of 97.1%, 97.7%, 98.5% and 99.1% in groups of five, four, three and two, this technique promises a swift and efficient authentication process in environments with frequent device sharing.

From where I stand, it is indeed the case that traditional methods used in AR/VR often do not benefit from such a good user experience as, for example, if one were to introduce a password this could lead to overhead in the hardware used (i.e. controllers for the input in the password field may be required). Alternative authentication techniques are also presented in the article, such as looking for objects in an encoded sequence, but one may be at a crossroad if the objects required are not part of the environment the user is currently in, therefore this work provides a fresh approach to a longstanding challenge in AR/VR: seamless, natural authentication.

Strengths:

- **Scalability:** It is often the case with emerging ideas that their implementation requires major changes in hardware to fit the new purpose for which they have been researched. However, this is not the case with "Nod to Auth"[1], because it makes use of the integrated IMU sensors and it just appends a new task to them through the use of additional software implies a module that is plug-and-play and easily updatable, thereby, through this flexibility facilitating a scalable solution
- **High Performing:** The performance of the module is remarkable, with the assumed setting of five users (i.e. a family) achieving an accuracy over 97% making it a strong candidate for real-world deployment once an experiment over a greater population reconfirms its performance. The detailed evaluation results, including the accuracy variation between different gestures, demonstrate that the system is not only reliable but also adaptable to different user behaviours.
- **User-centric design:** The aim of making the system feel natural is reflected through the minimal effort of the user, as head gestures are quick and easy to perform, thus leading to a more intuitive and fluid experience. The inspiration from the "Slide-to-Unlock" philosophy aligns perfectly with modern technology being expected to prioritise efficiency and ease of usability.

Weaknesses:

- **Confusion Among Similar Users:** As noted by the authors, one limitation exposed through the evaluation of the module is represented by users that share similar biometrics, such as those with comparable neck lengths and heights. Because of these homogenous groups, the module can lead to false positives in the classifications of their users, which in turn leads to lower identification accuracy. One could image the extreme example of this disadvantage by imagining twins being confused by the system or both being attributed the same identity by the system. As a result, the module ought to require a fallback mechanism when classes (i.e. users) possess such similar characteristics.
- **Limited Group Size Performance:** The evaluation only takes into account small groups of people (i.e. families or labs) which may lead to the system effectiveness being impacted if a larger group is encountered (i.e. an office). This is an indirect weakness that is related to the one above as it is often the case that when larger groups are involved, the variation of biometrical features reduces, thus the system may only confidently assign the right identities to outliers. Again, this limitation points to the need of the aforementioned fallback mechanism that would maintain accuracy even in cases where confusion amongst a larger group takes place.
- **Dependence on Gesture Type**: The performance of the system is influenced by the type of head gestures, with some of them yielding better results of the module than others (i.e. tilting), therefore, indicating that certain movements result in better biometric information derivation than others. Although this might first be seen as an advantage, the reliance of specific gestures might pose a challenge if users have difficulty performing them consistently or in the right manner (i.e. a disability).

Writing and Presentation:

Overall, I would argue that the paper is well written, having a concrete and easy to follow structure, with the technical details being presented in the right depth which make them easily digestible. The writing is precise, leaving no room for interpretation, and, coupled with the examples, they effectively illustrate the operational pipeline of the system. However, more information regarding user experience, such as a survey performed on the experimental group, would enhance the practical understanding of the system.

Credit to: Dan Stoicescu

Further suggestions:

- Could discuss strengths/weakness of the research methodology more
- Could provide more insights on potential research direction

Nice news blog structure :)

Haoyu Wang

January 2025

1 News Summary and Commentary

BBC recently reported that WhatsApp and several other messaging apps have expressed dissatisfaction and opposition toward government plans to monitor encrypted communications. These apps argue that end-to-end encryption (E2EE) is a critical technology for protecting user privacy. Government surveillance requirements may force these platforms to weaken encryption standards, increasing the risk of privacy breaches and potential harm to users. This stance underscores the importance these platforms place on user privacy, gaining support from privacy advocates but simultaneously sparking debates over national security and law enforcement needs.

2 Why I Find This Topic Interesting

These messaging apps focus on safeguarding user privacy, while governments prioritize security concerns. This issue lies at the heart of modern cybersecurity: the balance between privacy and security—a balance that is notoriously difficult to achieve. Often, one side must compromise for the other, whether willingly or reluctantly, as seen in challenges faced by TikTok in the U.S. Today, with tightening global regulatory environments, such issues are increasingly significant. Different stakeholders approach privacy and security from varying perspectives, and any choice inevitably leaves some unsatisfied. Understanding the motivations behind these decisions in different scenarios is a fascinating topic for study, as it also helps develop reasonable solutions for future conflicts of this nature. The ultimate question is how to ensure privacy while maintaining security.

3 Discussion and Future Research Directions

3.1 Research Motivation

The conflict between privacy and security has long coexisted in the realms of technology and policy. Governments argue that mandatory surveillance is essential for combating criminal activities, terrorism, and other security threats. However, privacy advocates disagree, warning that any backdoors or weakening of encryption could be exploited. Not only could law enforcement misuse such vulnerabilities, but in the worst-case scenario, leaked or forcibly broken information could be exploited by hackers or rogue states. The pressing question is whether it is possible to achieve lawful surveillance without compromising the integrity of encryption. This debate highlights the fundamental trade-off between individual rights and collective security, a core issue in modern cybersecurity.

3.2 Future Research Directions and Potential Solutions

A potential solution could involve designing a novel encryption protocol that balances privacy protection with law enforcement needs. For instance, exploring technologies like Zero-Knowledge Proofs (ZKPs) and Secure Multi-Party Computation (SMPC) may provide a way to verify communication legality without compromising user privacy.
In addition, Any technological solution must be accompanied by transparent policies and strong oversight mechanisms to ensure it is used ethically and appropriately. For example:

- Creating an independent oversight body that evaluates and authorizes surveillance requests to prevent misuse.
- Establishing clear criteria for when and how surveillance tools can be used, ensuring they are targeted and proportionate.

By combining technical innovations with clear policy guidelines, we may one day achieve a balance where privacy and security can coexist harmoniously.

Further suggestions:

- Deeper insights into what could be challenging (research-wise and practically speaking)
- Identify potential research agenda

Questionnaires

Questionnaires

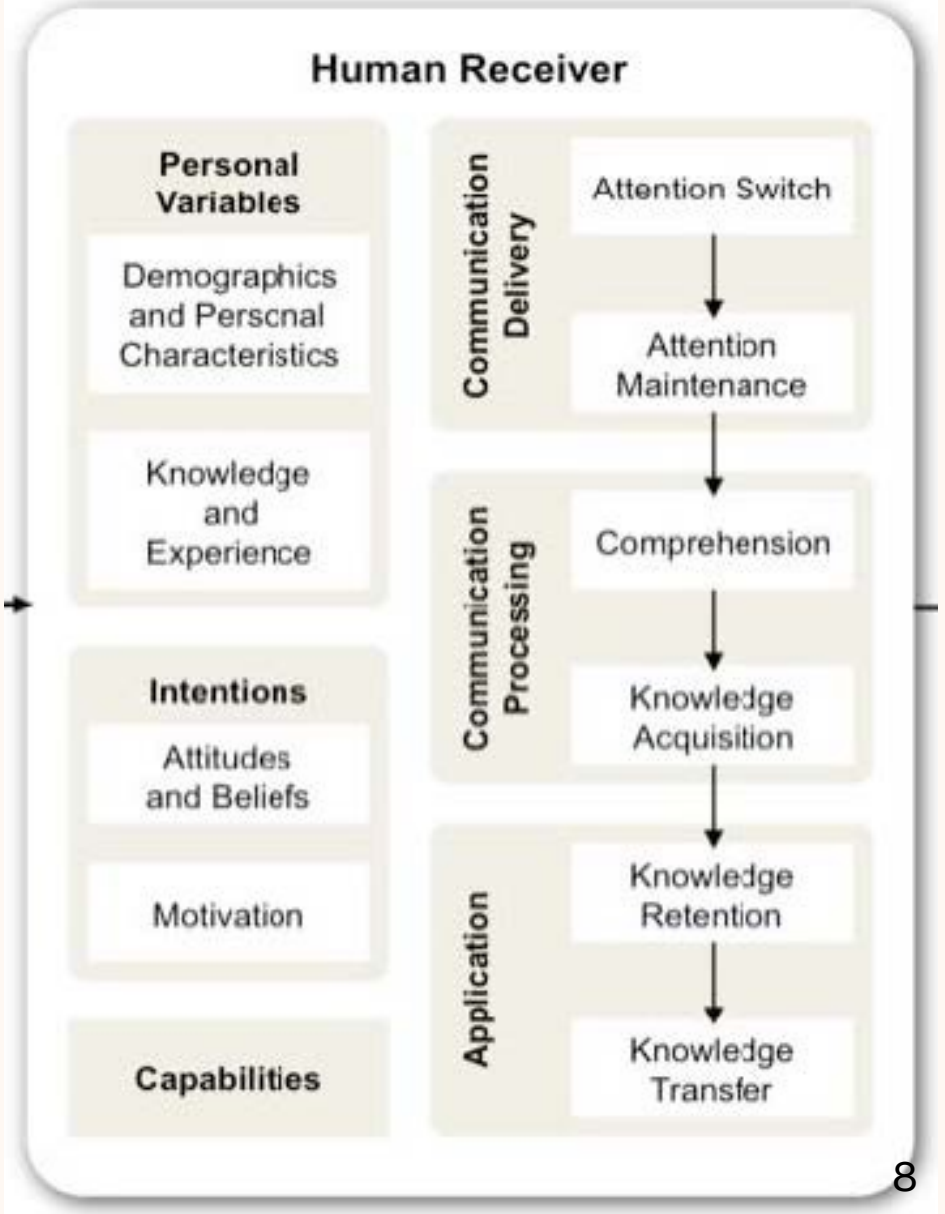
- Ask participants to answer a set of pre-defined questions.
- Pros:
  - gather data from a large number of people quickly
  - can determine how prevalent an issue or concern is
  - close-ended questions are easy to analyze
- Cons:
  - can only gather data you know about
  - careful planning is required before running a questionnaire
  - open-ended questions can take a lot of time to analyze and require careful setup

Questionnaires can be used at various points in the design process

- Understanding people
  - Understand the target population
  - Incorrect mental models
- Testing a theory
  - Are my assumptions correct?
  - Do people think that A==B?
- Testing a prototype design
  - How do people interpret my interface?
- Testing the final design
  - How are people actually using it?
  - What do people think after they use it?

What do you want to know?

- Attitudes
  - Do you like X?
  - Would using X work?
- Behaviors
  - How often do you use X?
  - Do you regularly do X?
- Knowledge
  - What is the best definition of X?
- Expectations
  - If the webpage did X what would you expect to happen?
- Capabilities
  - What is the result of adding 20 and 30?



Common survey elements

- Single and multiple choice checkboxes
- Matching
  - Rank the following from 1 to 5
- Rating scales
  - Likert Scales
    - 3, 5, 7 points scales
  - Semantic scales
- Open ended responses

OPEN ENDED

- Where does this URL go? What does it do?

Easier to write, harder to analyze

Harder to write, easier to analyze

CLOSE-ENDED

If you clicked on the link above, what web page would open?

- ◯ WWW3's main page
- ◯ National Geographic's main page
- ◯ World News's main page
- ◯ I will be taken to one of the sites above, but not their main page
- ◯ I will be taken to a website not listed above
- ◯ Other _____

Response Anchors

Psychologists have been working for quite some time to determine the least biased way to present a set of answers.

On the right are a set of response anchors that are known to work well.

Likert-Type Scale Response Anchors

Citation:
Vagias, Wade M. (2006). *Likert-type scale response anchors*. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University.

Level of Acceptability
- 1 – Totally unacceptable
- 2 – Unacceptable
- 3 – Slightly unacceptable
- 4 – Neutral
- 5 – Slightly acceptable
- 6 – Acceptable
- 7 – Perfectly Acceptable

Level of Appropriateness
- 1 – Absolutely inappropriate
- 2 – Inappropriate

My beliefs
- 1 – Very untrue of what I believe
- 2 – Untrue of what I believe
- 3 – Somewhat untrue of what I believe
- 4 – Neutral
- 5 – Somewhat true of what I believe
- 6 – True of what I believe
- 7 – Very true of what I believe

Priority:

Level of Support/Opposition
- 1 – Strongly oppose
- 2 – Somewhat oppose
- 3 – neutral
- 4 – Somewhat favor
- 5 – Strongly favor

Level of Probability
- 1 – Not probable
- 2 – Somewhat improbable
- 3 – Neutral
- 4 – Somewhat probable

Level of Acceptability
- 1 – Totally unacceptable
- 2 – Unacceptable
- 3 – Slightly unacceptable
- 4 – Neutral
- 5 – Slightly acceptable
- 6 – Acceptable
- 7 – Perfectly Acceptable

- 3 – Sometimes but

Affect on X

Frequency

Fill in the blank type question

Q2: What is your age? _____

Typical multiple choice question

Q8: What is the highest level of education you have achieved?

( ) High school or less

( ) Some College

( ) Bachelor's Degree

( ) Master's Degree

( ) Doctorate Degree

Scale where multiple questions are meant to be summed together

Q12: To what extent do you agree or disagree with each of the following statement

Please select one answer per row

	Strongly Disagree	Disagree	Neither Agree or Disagree	Agree	Strongly Agree
I often ask others for help with the computer	○	○	○	○	○
Others often ask me for help with the computer	○	○	○	○	○

Likert scale question using a pre-defined anchor

Q13: In terms of your Internet skills, do you consider yourself to be:

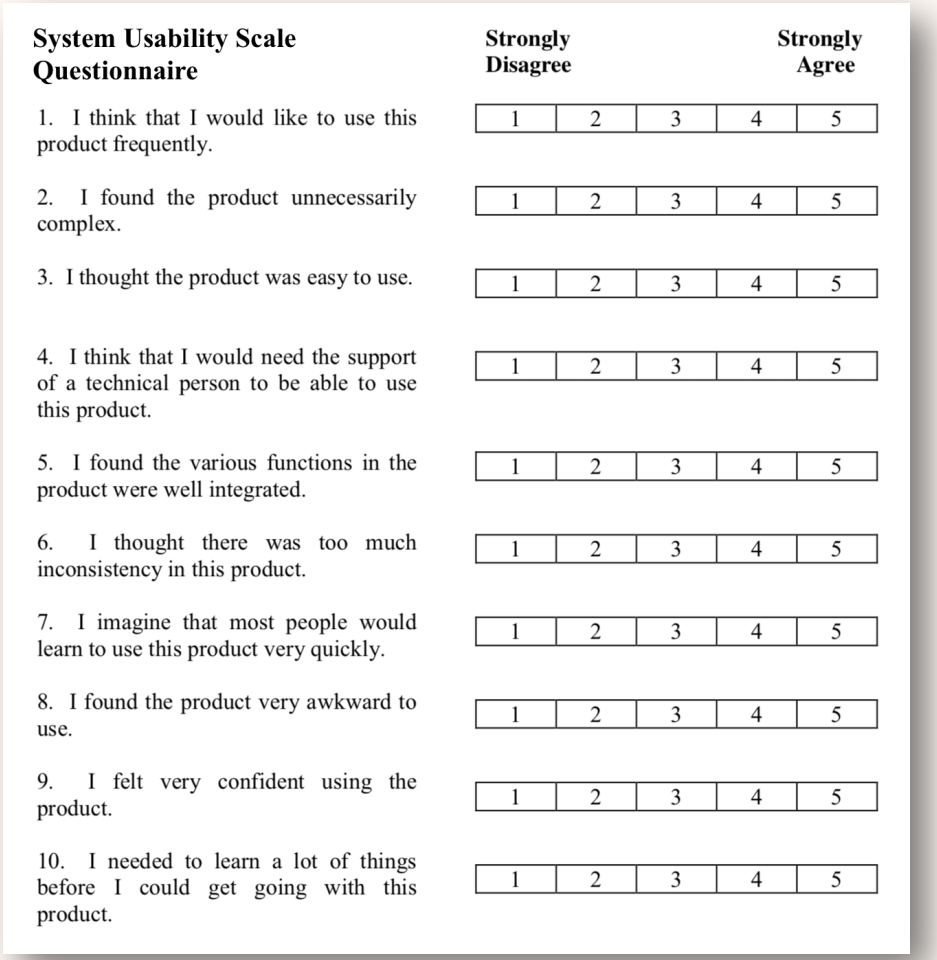
( ) Not at all skilled

( ) Not very skilled

( ) Fairly skilled

( ) Very skilled

( ) Expert



System Usability Scale Questionnaire

	Strongly Disagree				Strongly Agree
1. I think that I would like to use this product frequently.	1	2	3	4	5
2. I found the product unnecessarily complex.	1	2	3	4	5
3. I thought the product was easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this product.	1	2	3	4	5
5. I found the various functions in the product were well integrated.	1	2	3	4	5
6. I thought there was too much inconsistency in this product.	1	2	3	4	5
7. I imagine that most people would learn to use this product very quickly.	1	2	3	4	5
8. I found the product very awkward to use.	1	2	3	4	5
9. I felt very confident using the product.	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this product.	1	2	3	4	5

System Usability Scale

#	Question	N/A		μ	σ
A1	*I apply software updates as soon as my computer prompts me.*	5	(1.0%)	3.20	1.221
A2	*I am happy to use an older version of a program, as long as it meets my needs.*	5	(1.0%)	^{r}1.99	1.000
A3	*Whenever I step away from my computer, I lock the screen.*	5	(1.0%)	2.50	1.306
A4	*Others can access my smartphone or tablet without needing a PIN or passcode.*	21	(4.4%)	^{r}3.34	1.545
A5	*When I discover a computer security problem at work, I'm likely to promptly report it to my employer.*	64	(13.4%)	4.08	0.995
A6	*It's important to use a WiFi password to prevent unauthorized people from using my home network.*	11	(2.3%)	4.66	0.690
A7	*I frequently click links in email messages to see what they are, regardless of who sent the message.*	5	(1.0%)	^{r}4.51	0.922
A8	*It's important to run anti-virus software on my computer.*	7	(1.5%)	4.35	0.941
A9	*When browsing websites, I frequently mouseover links to see where they go, before clicking them.*	4	(0.8%)	4.13	0.977
A10	*When using public WiFi, I visit the same websites that I would visit when using the Internet at home.*	20	(4.2%)	^{r}2.93	1.266
A11	*I usually do not pay attention to where I'm downloading software from.*	2	(0.4%)	^{r}4.38	0.900
A12	*I frequently backup my computer.*	5	(1.0%)	3.07	1.165
A13	*I frequently visit websites even when my web browser warns me against it.*	8	(1.7%)	^{r}3.98	1.028
A14	*I circumvent my employer's computer usage policies when they prevent me from completing a task.*	86	(18.0%)	^{r}3.54	1.184
A15	*I am careful to never share confidential documents stored on my home or work computers.*	15	(3.1%)	4.36	0.757
A16	*Frequently checking the access control settings on social networking websites isn't worth the time it takes.*	18	(3.8%)	^{r}3.56	1.165
A17	*I always write down my passwords to help me remember them.*	6	(1.3%)	^{r}3.60	1.313
A18	*Creating strong passwords is not usually worth the effort.*	6	(1.3%)	^{r}4.05	1.047
A19	*I frequently check my financial accounts for fraudulent charges.*	10	(2.1%)	4.11	0.914
A20	*If I receive a suspicious email from a company that I do business with, I'll phone the company to make sure the email is accurate.*	22	(4.8%)	3.52	1.236
A21	*I never give out passwords over the phone.*	7	(1.5%)	4.53	0.787
A22	*I frequently purchase things that I see advertised in unsolicited emails.*	4	(8.8%)	^{r}4.51	0.840
A23	*I tend to ignore computer security stories in the news because they don't impact me.*	4	(8.8%)	^{r}3.83	1.050
A24	*I use encryption software to secure files or email messages.*	10	(2.1%)	2.74	1.225
A25	*Once I create a password, I tend to never change it.*	5	(1.0%)	^{r}3.30	1.182
A26	*I try to create a unique password for every account I have.*	5	(1.0%)	3.21	1.284
A27	*Rather than logging out of websites, I usually just navigate elsewhere or close the window when I'm done.*	7	(1.5%)	^{r}3.06	1.299
A28	*I always make sure that I'm at a secure website (e.g., SSL, "https://", a lock icon) when transmitting information online.*	4	(0.8%)	3.80	1.173
A29	*I frequently use privacy software, "private browsing" or "incognito" mode when I'm online.*	6	(1.3%)	3.17	1.247
A30	*I frequently let others use my computing devices (e.g., smartphone, tablet, laptop).*	3	(0.7%)	^{r}3.79	1.172

Table 1. Initial set of security questions evaluated on a 5-point Likert scale (from "strongly disagree" to "strongly agree") by 479 participants. Depicted are the questions, the rate of "N/A" responses, and the average responses and standard deviations after recoding negatively-phrased questions (^{r}).

Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS)

Planning a survey

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 NOT 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

Topics Outline

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

Planning a survey

- Surveys normally answer **multiple research questions**. With each research question tied to one or more 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 whole 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 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.

Does consuming chocolate increase the number of Nobel Laureates?

This is a correlation, not necessarily a causation.

Chocolate Consumption, Cognitive Function,
and Nobel Laureates
Franz H. Messerli, M.D.

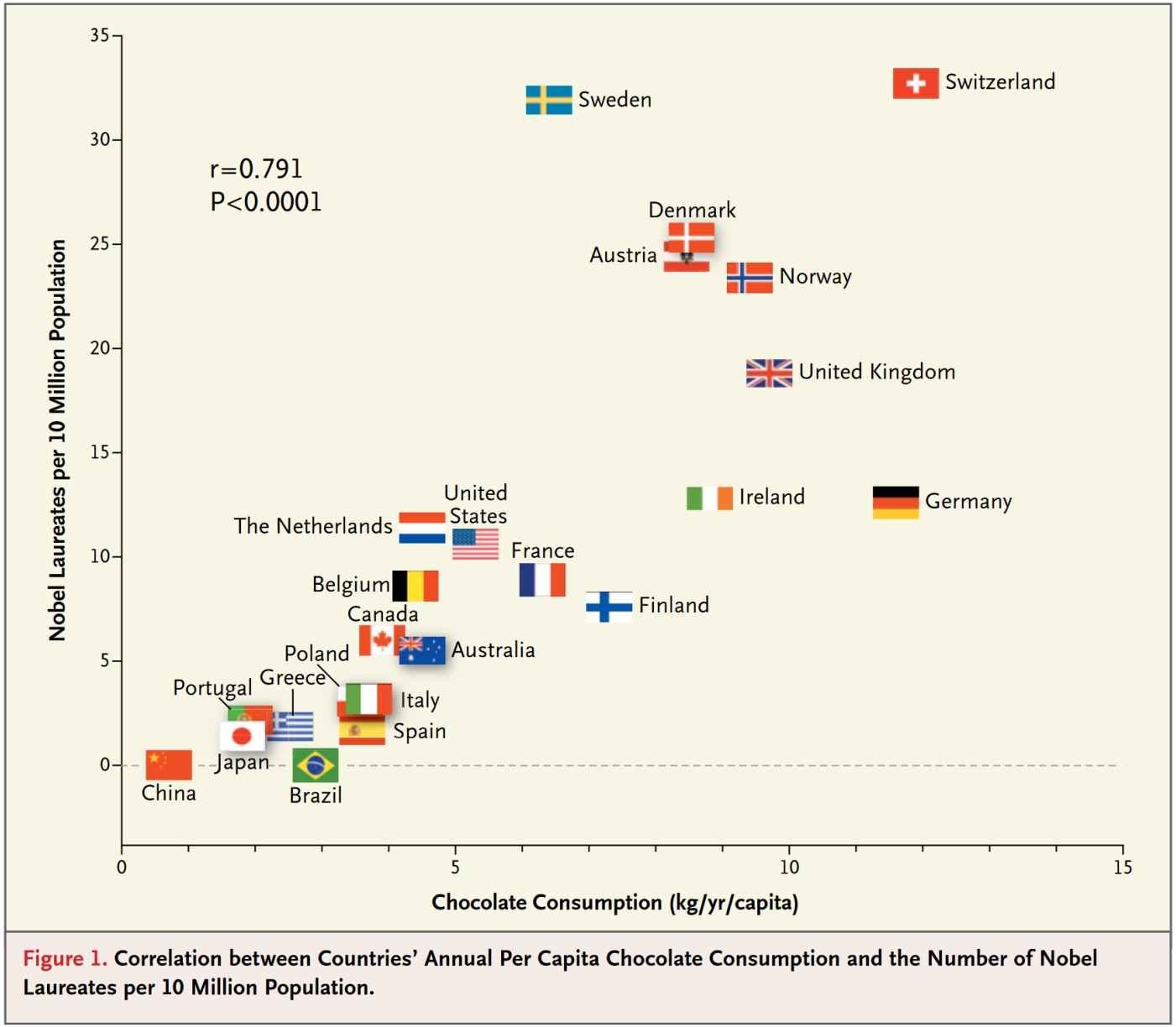
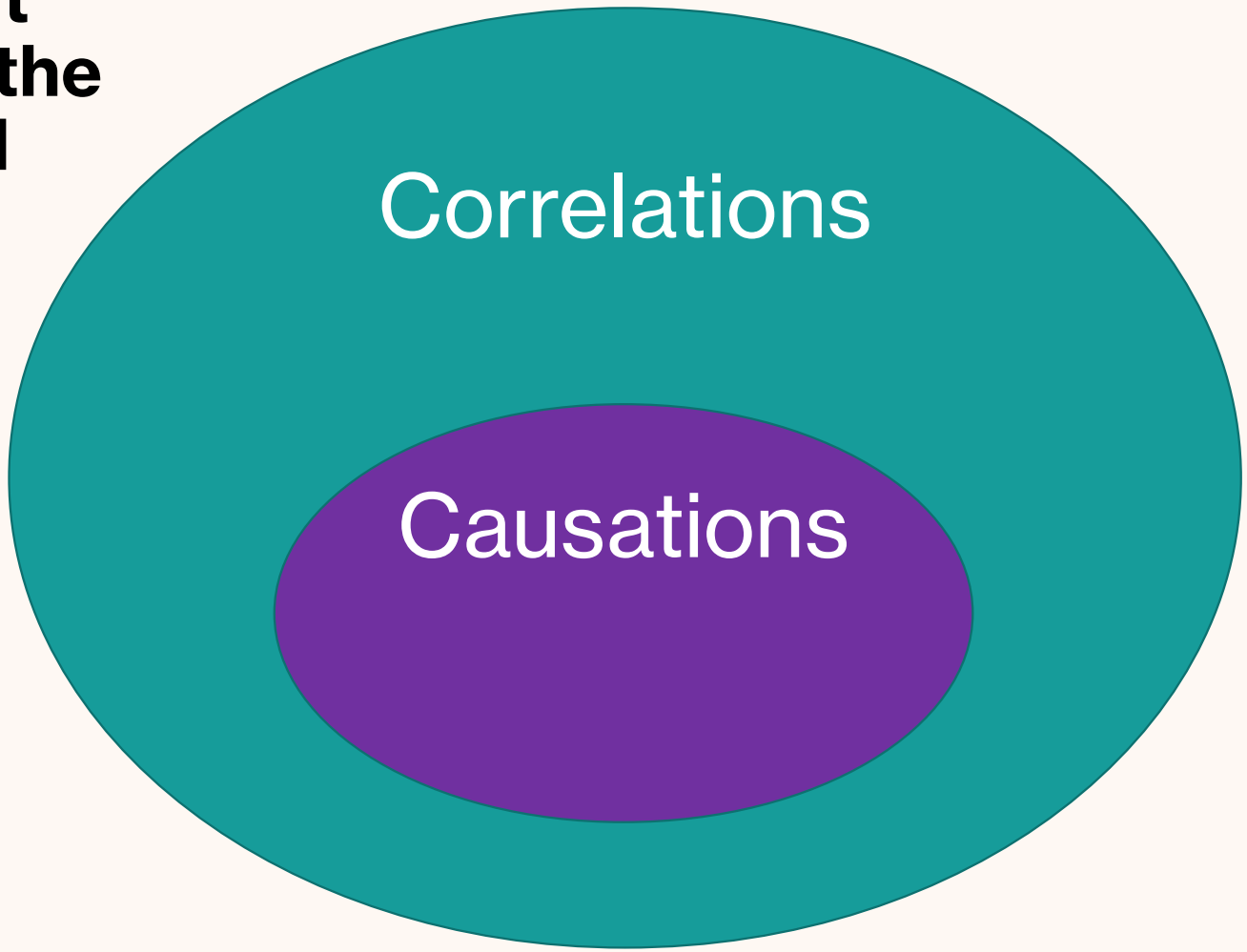


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Causations are Correlations, but not necessarily the other way round



History + CTR is a correlation.

How might you test if it is really a causation?

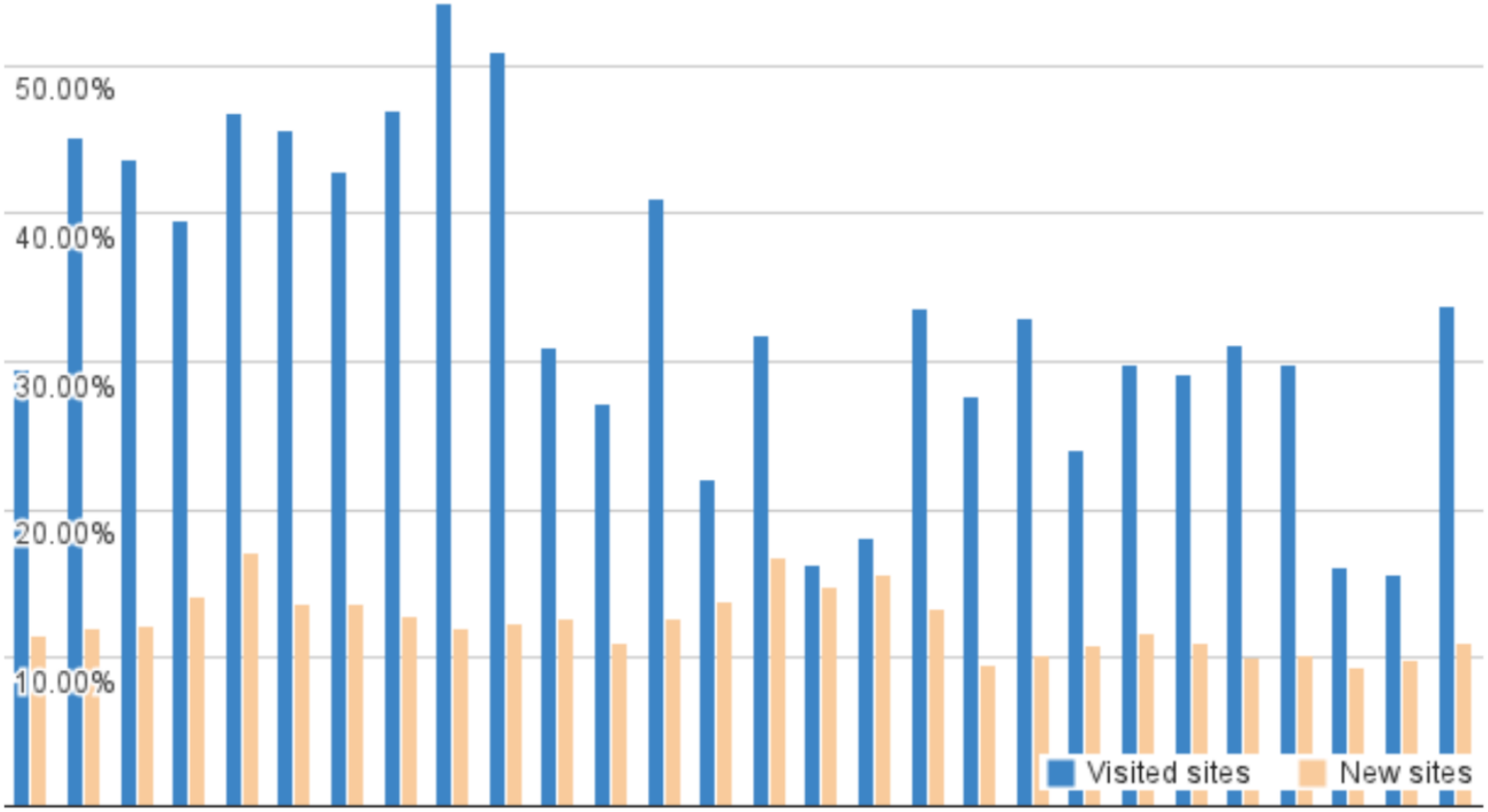


Figure 3: Daily CTR, separated by whether the website was already in the user's browsing history. For 28 days in January-February 2014.

Topics Outline

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

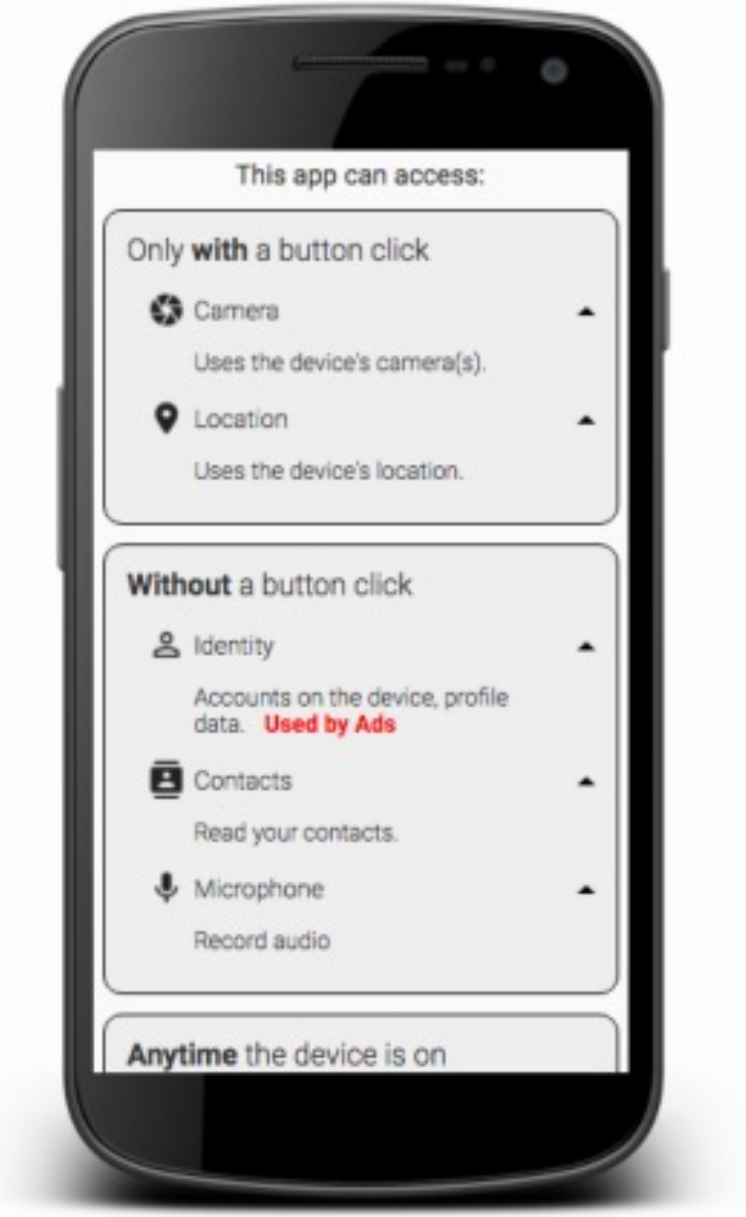
What are you going to measure?

- 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

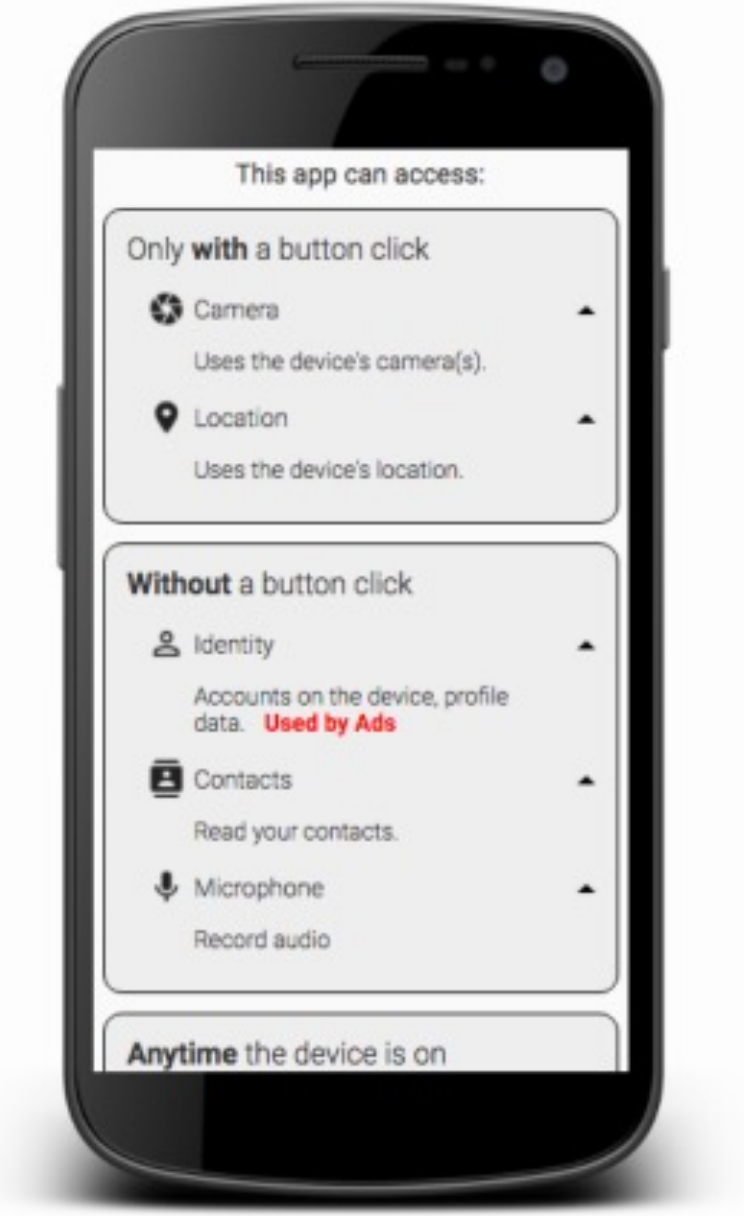
Some research questions:

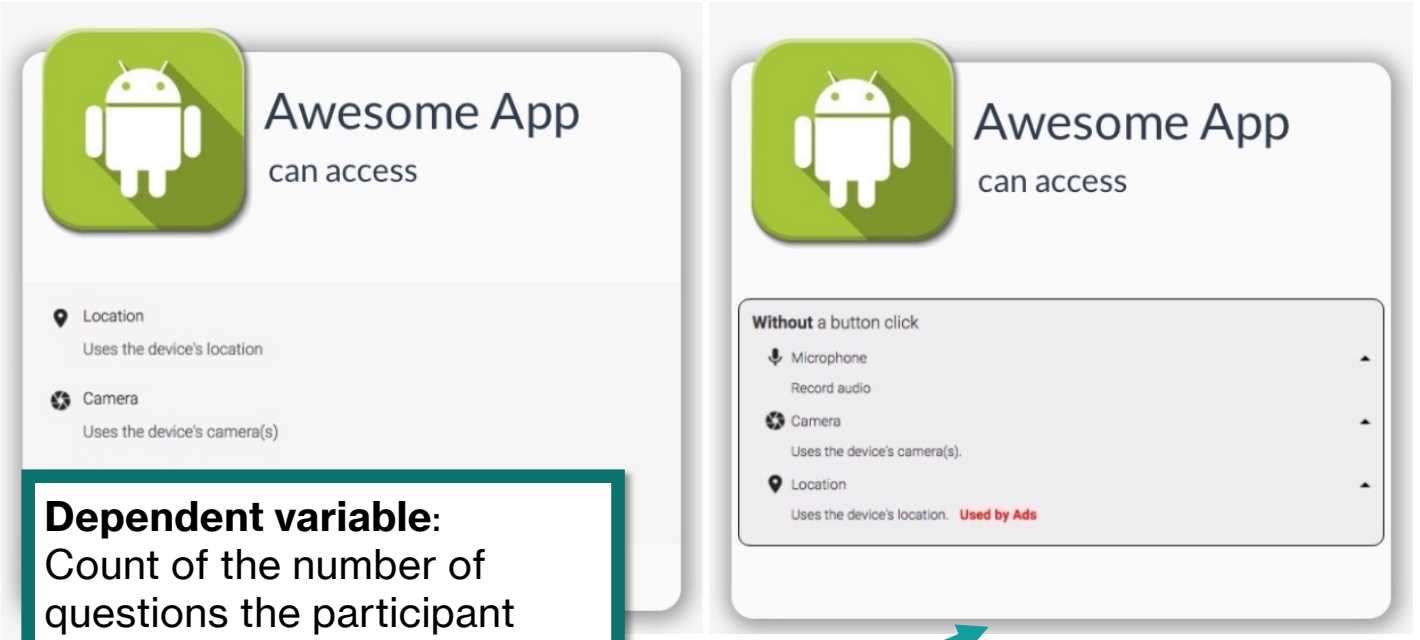
- Can people differentiate between a subdomain and a domain when reading a URL?
- Does [a new system] help people differentiate between malicious URLs and safe ones?
- Can users use [a new password manager] faster and with less errors than [the old password manager]?
- Does knowing how an app will use 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

g 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

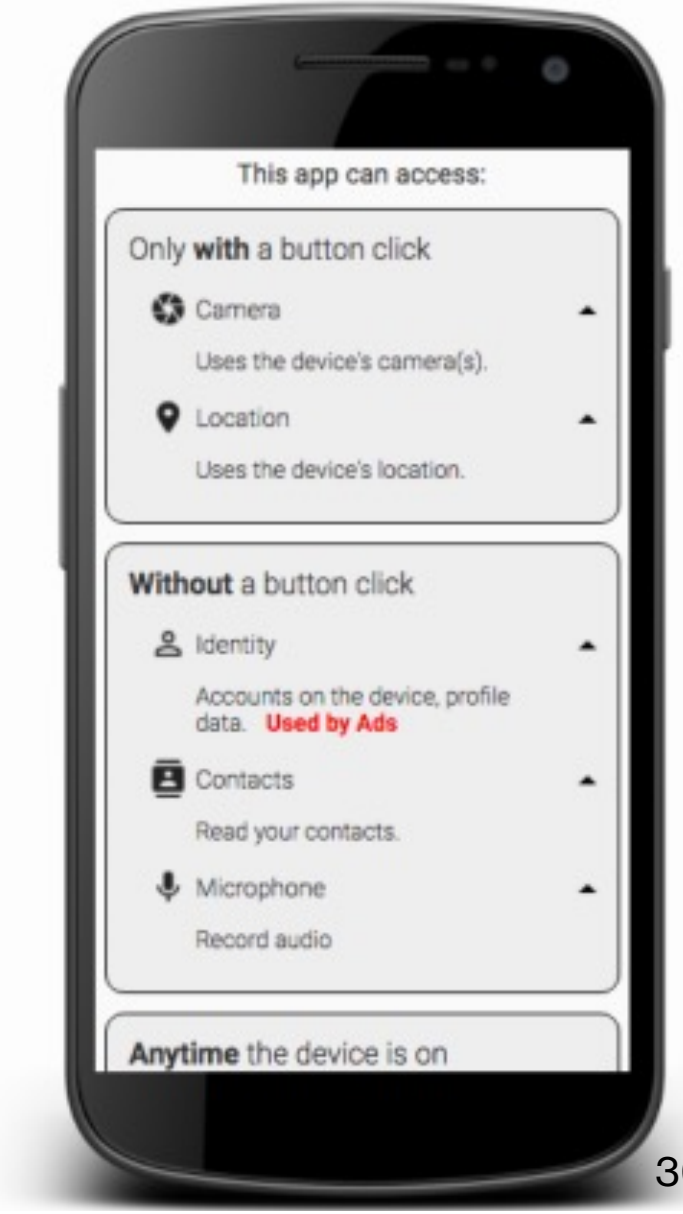
- Descriptive questions vs testing a question
- Correlation vs causation
- Dependent vs independent variables
- **Between and within subjects testing**
- Numeric vs categorical data

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

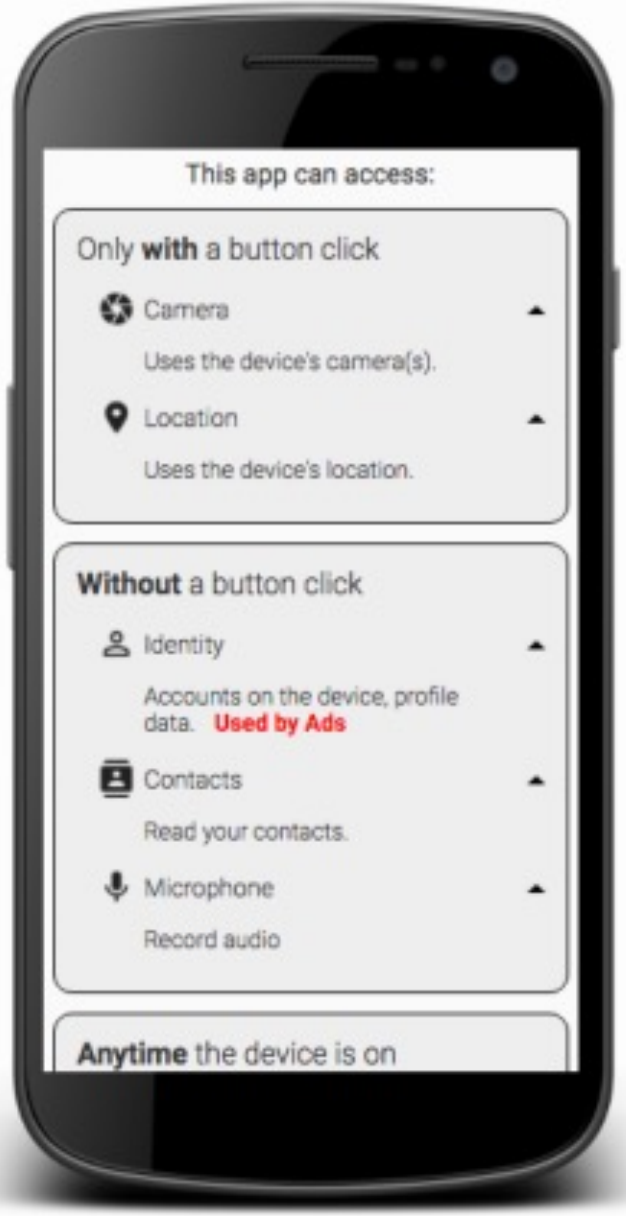
- Descriptive questions vs testing a question
- Correlation vs causation
- Dependent vs independent variables
- Between and within subjects testing
- **Numeric vs categorical data**

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

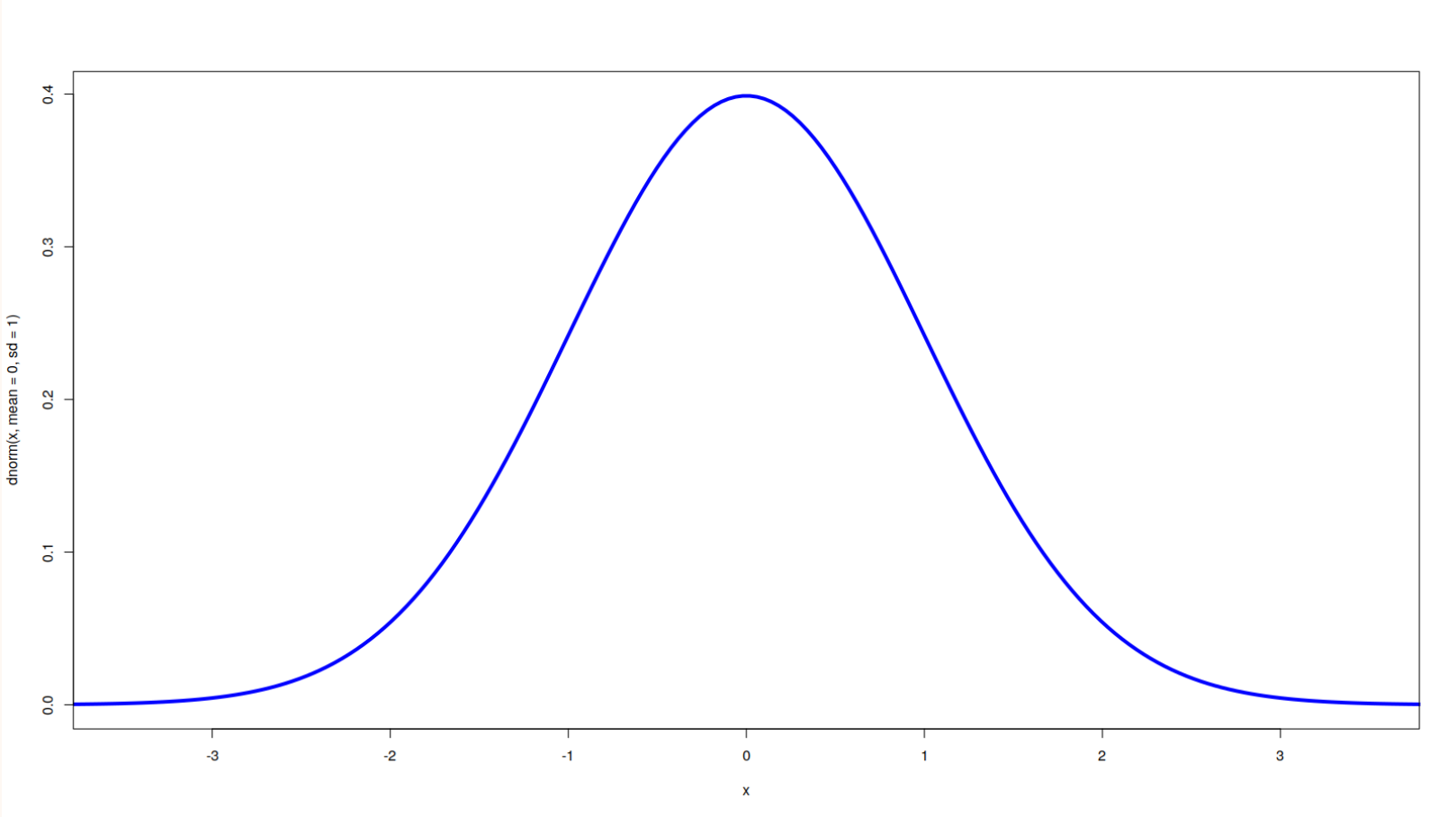
t-test: Test if two groups have the same mean (average)

T-test requires:

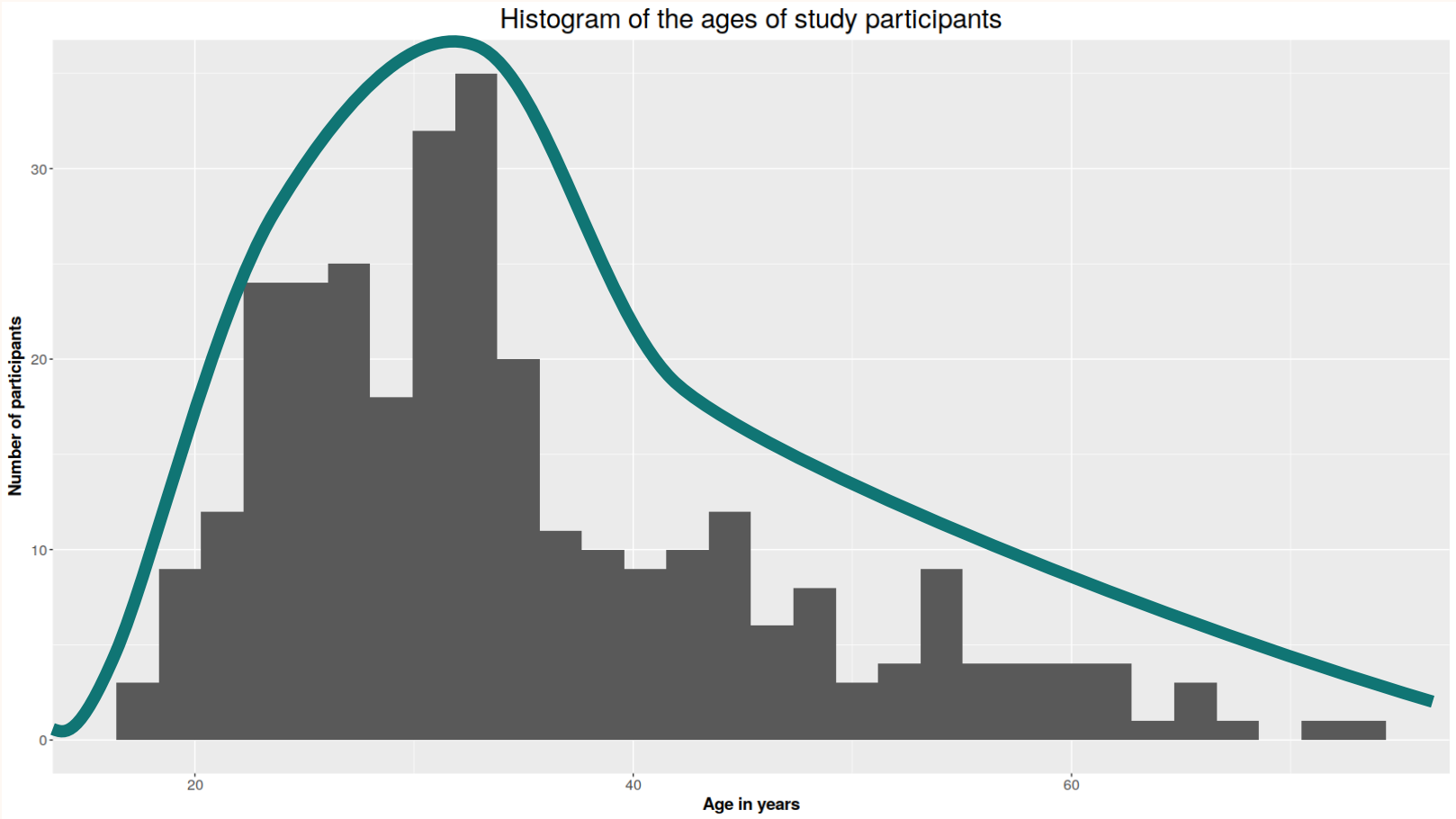
- Independent variable: categorical binary
- Dependent variable: numeric (continuous or discrete)

- Data must be **normally distributed**

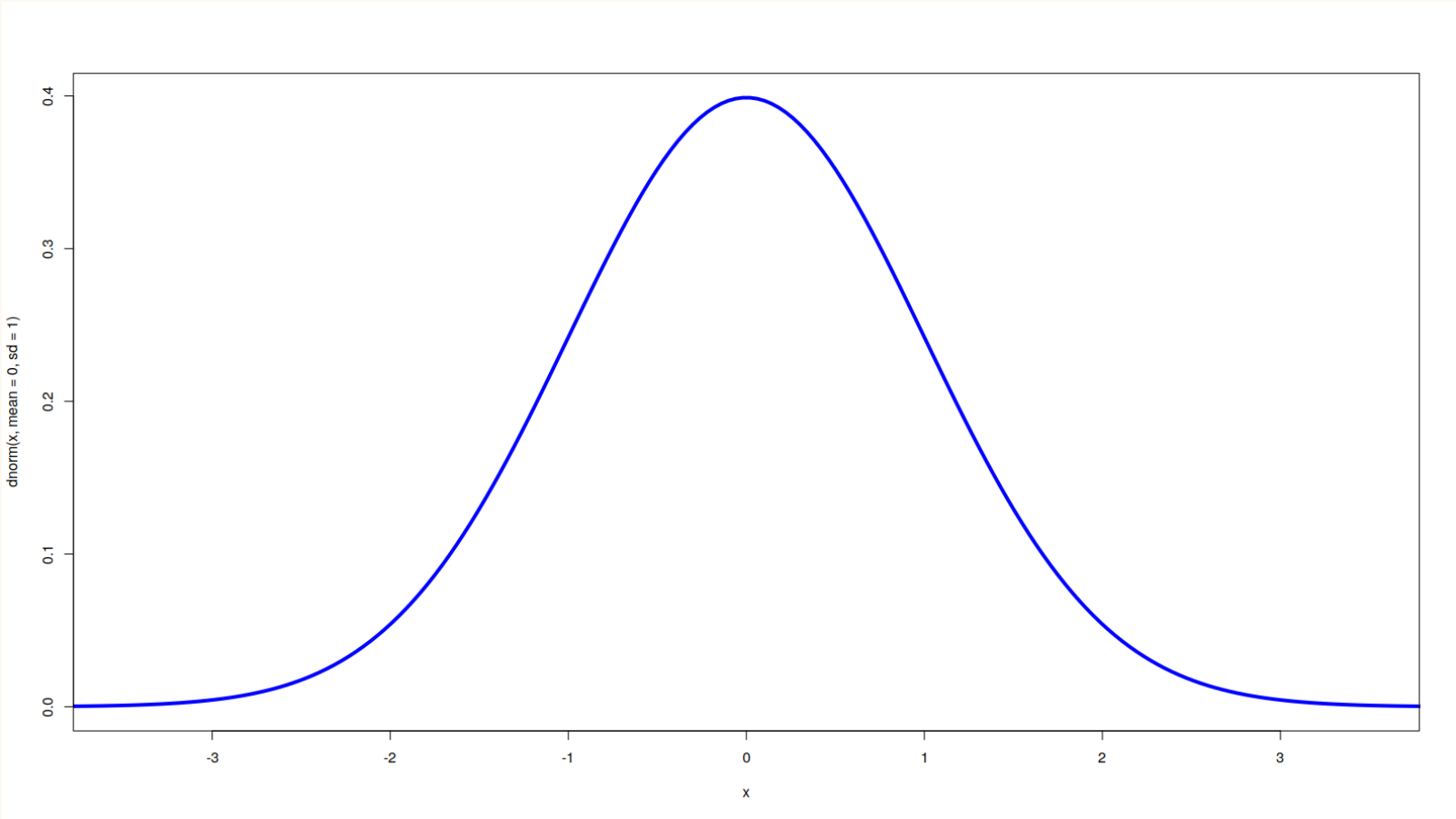
Normal distribution



Real data is messy

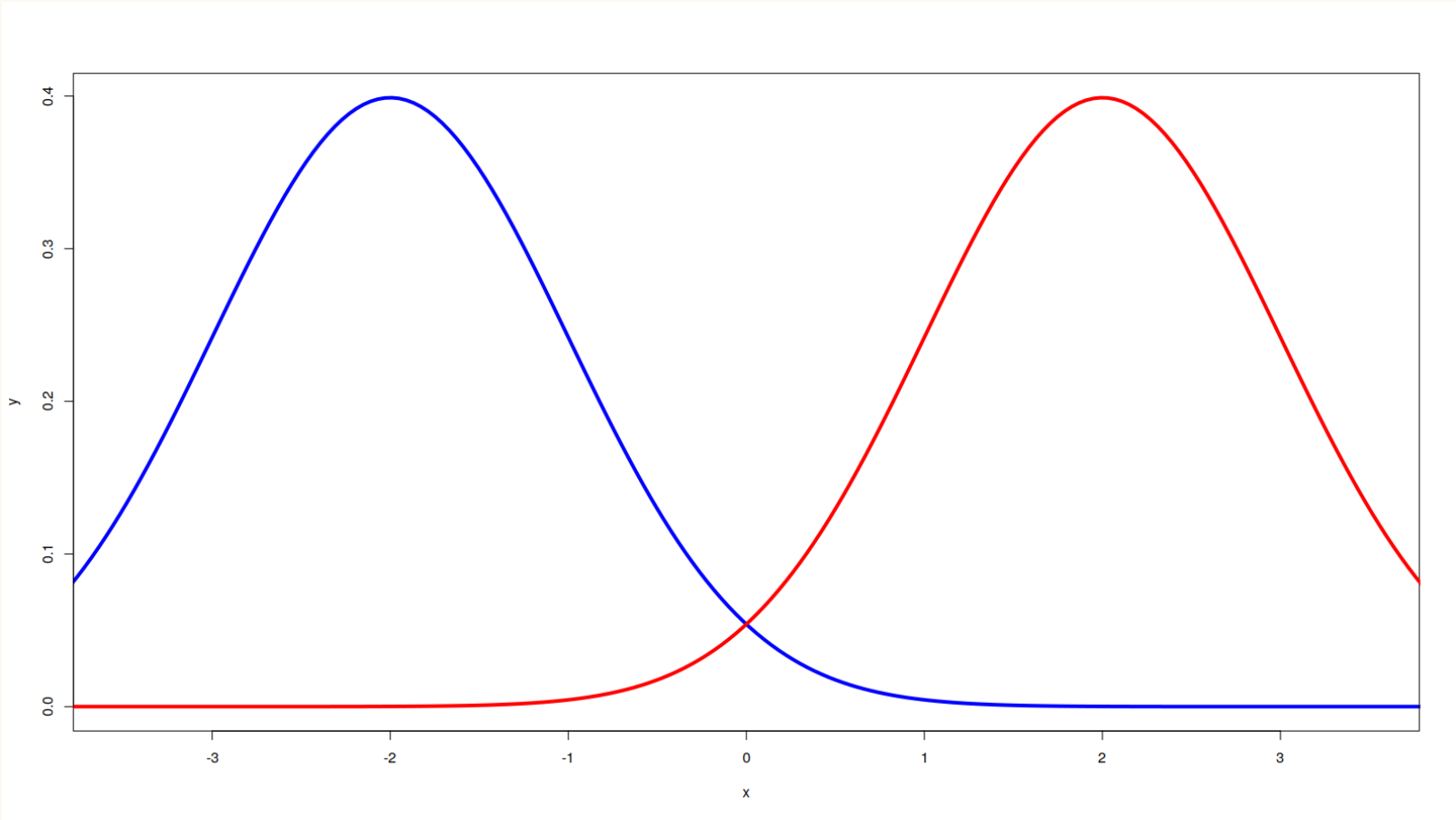


Normal distribution

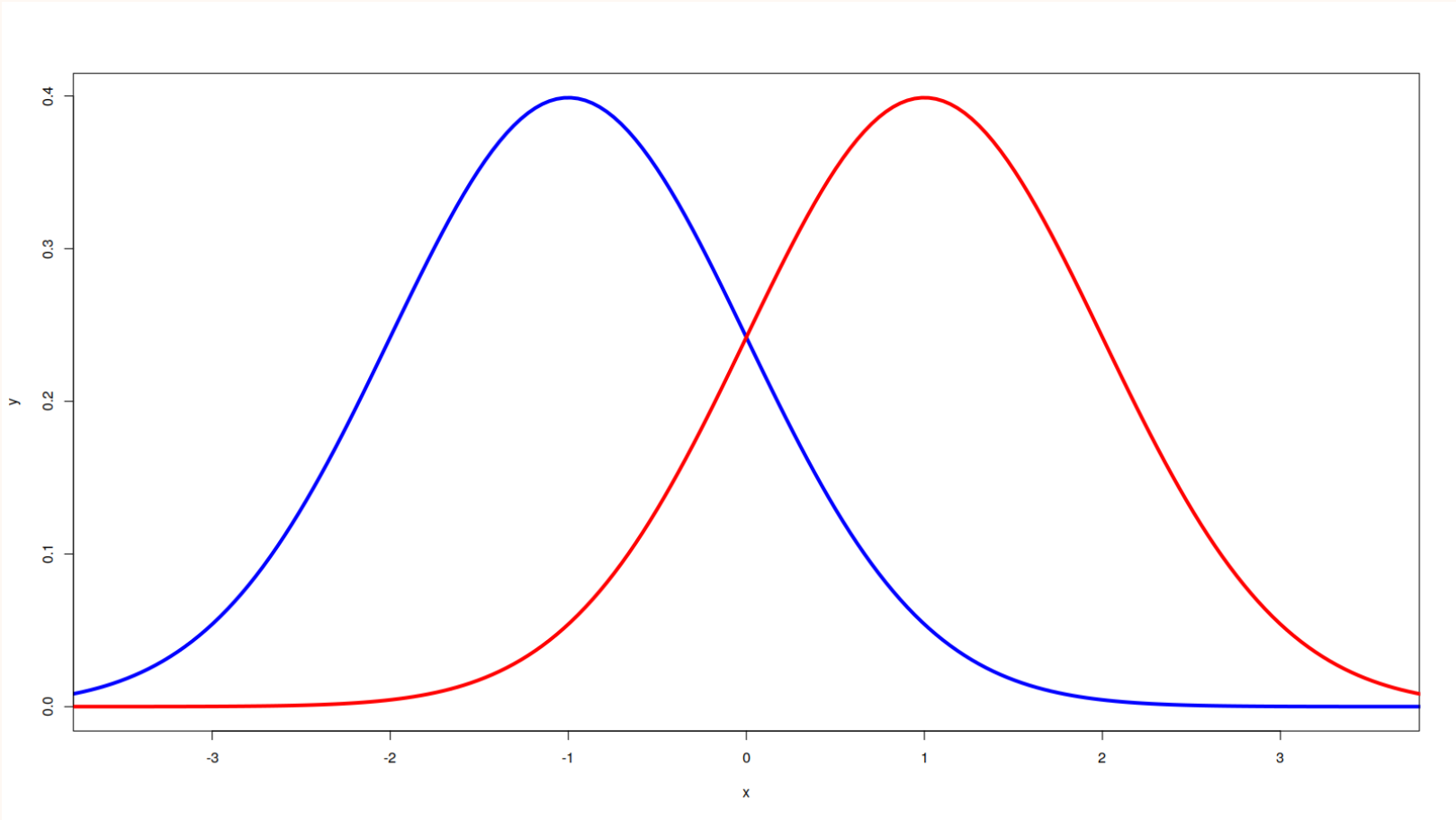


T-test: Do two populations have the same mean?

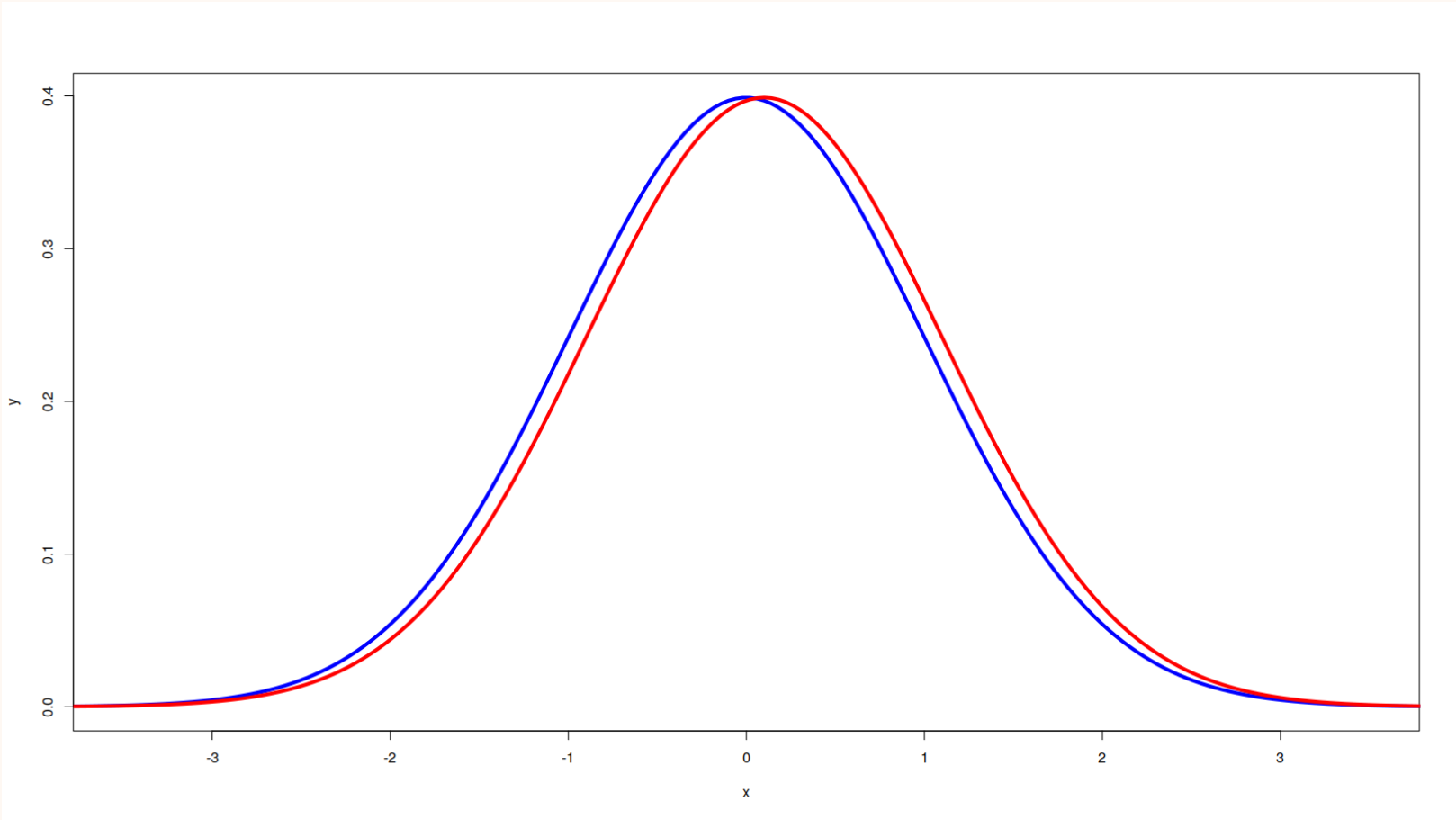
Different means



Maybe? different means



Likely not different means



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.

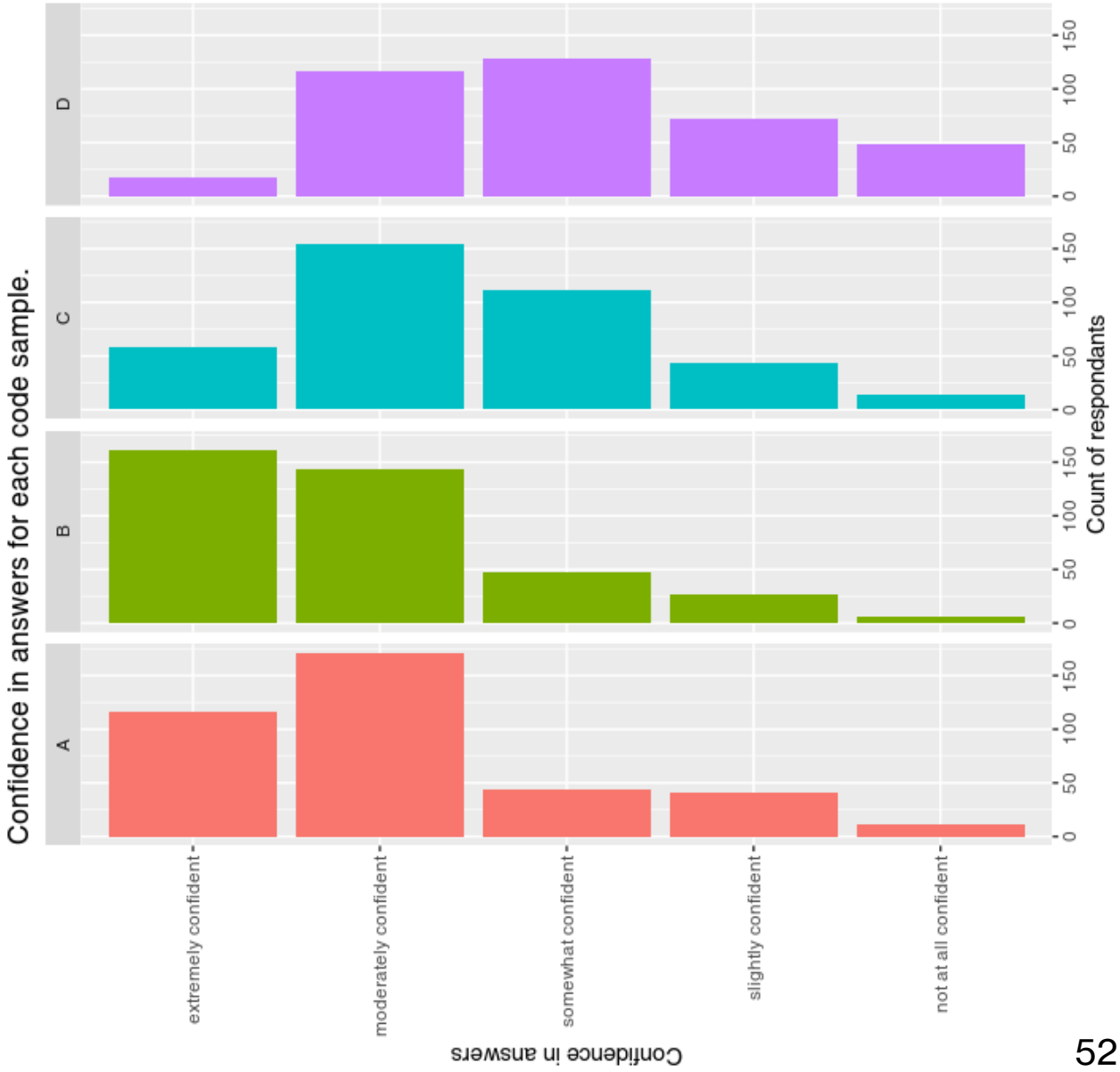
Research Question: Does the code sample shown impact confidence in their answer?

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

Within-subjects

Independent:
Which code sample shown

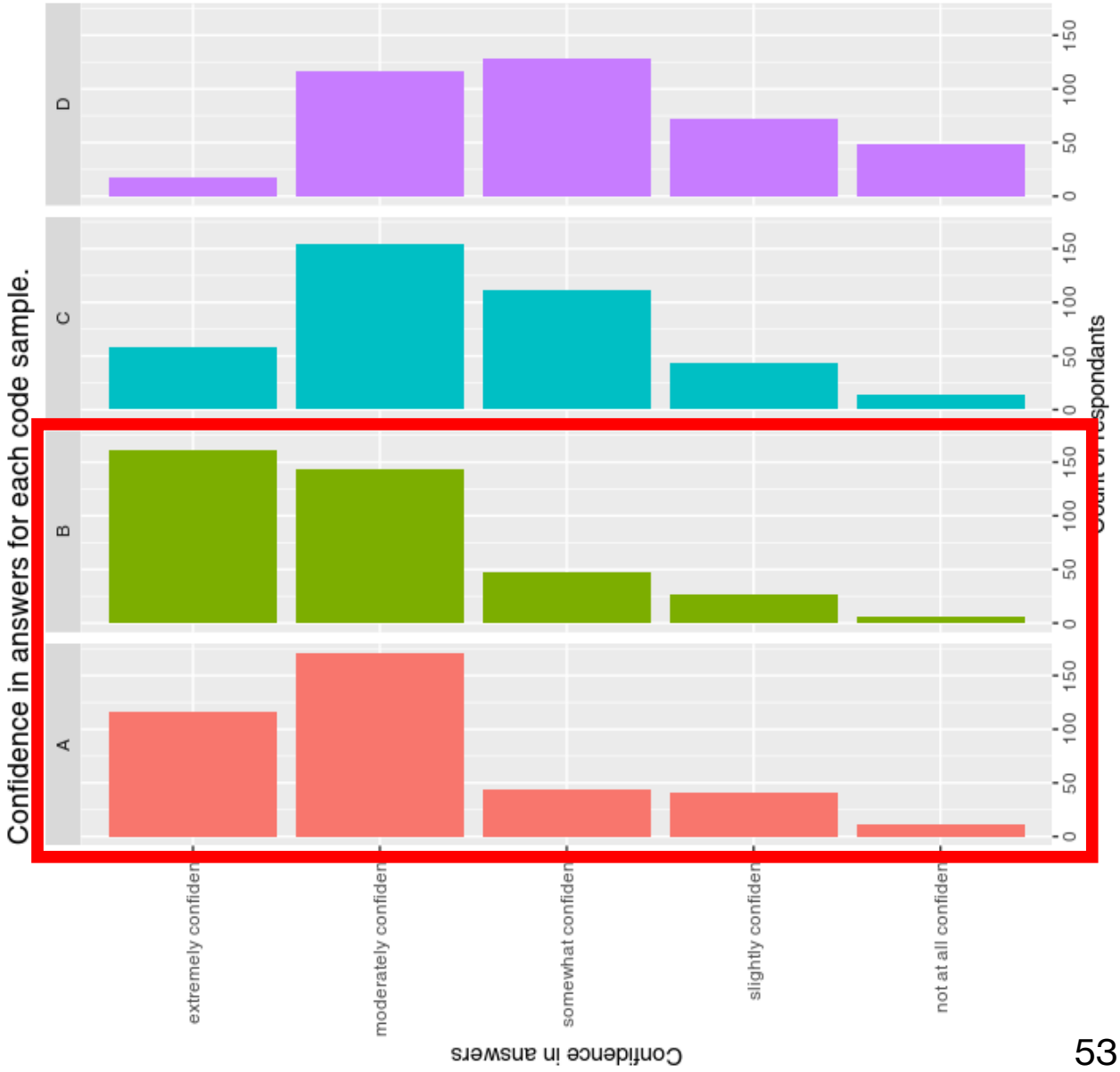
Dependent:
Confidence



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

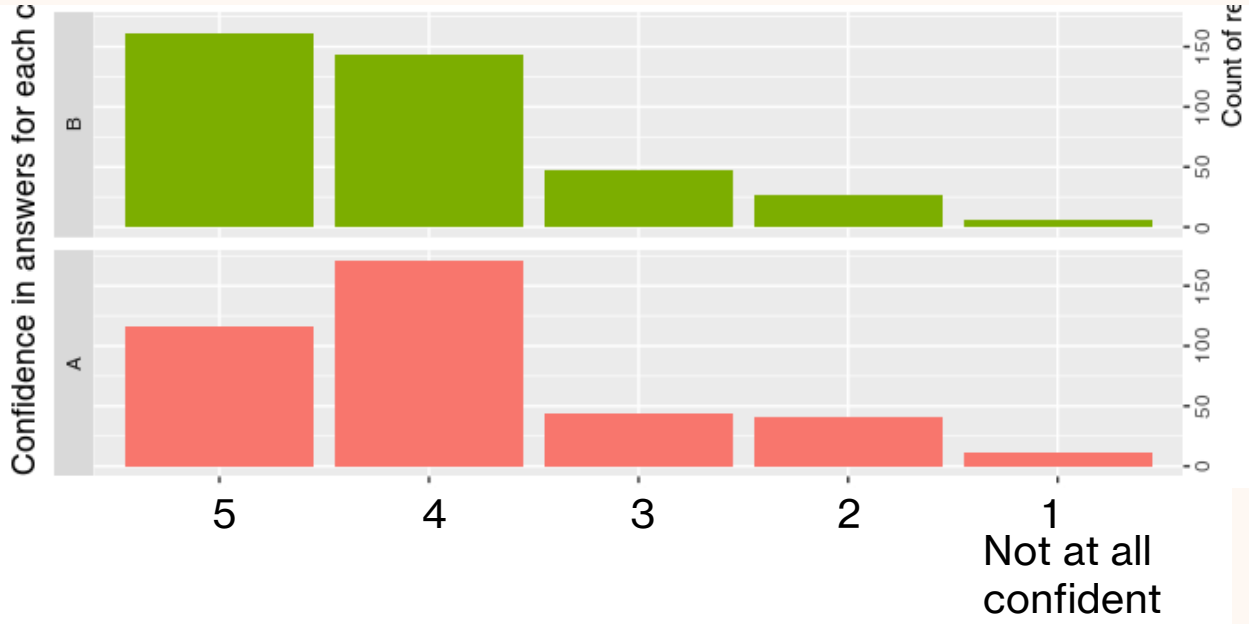
Solution: Run the t-test on each pair. So test A vs B, A vs C, …. C vs D.

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

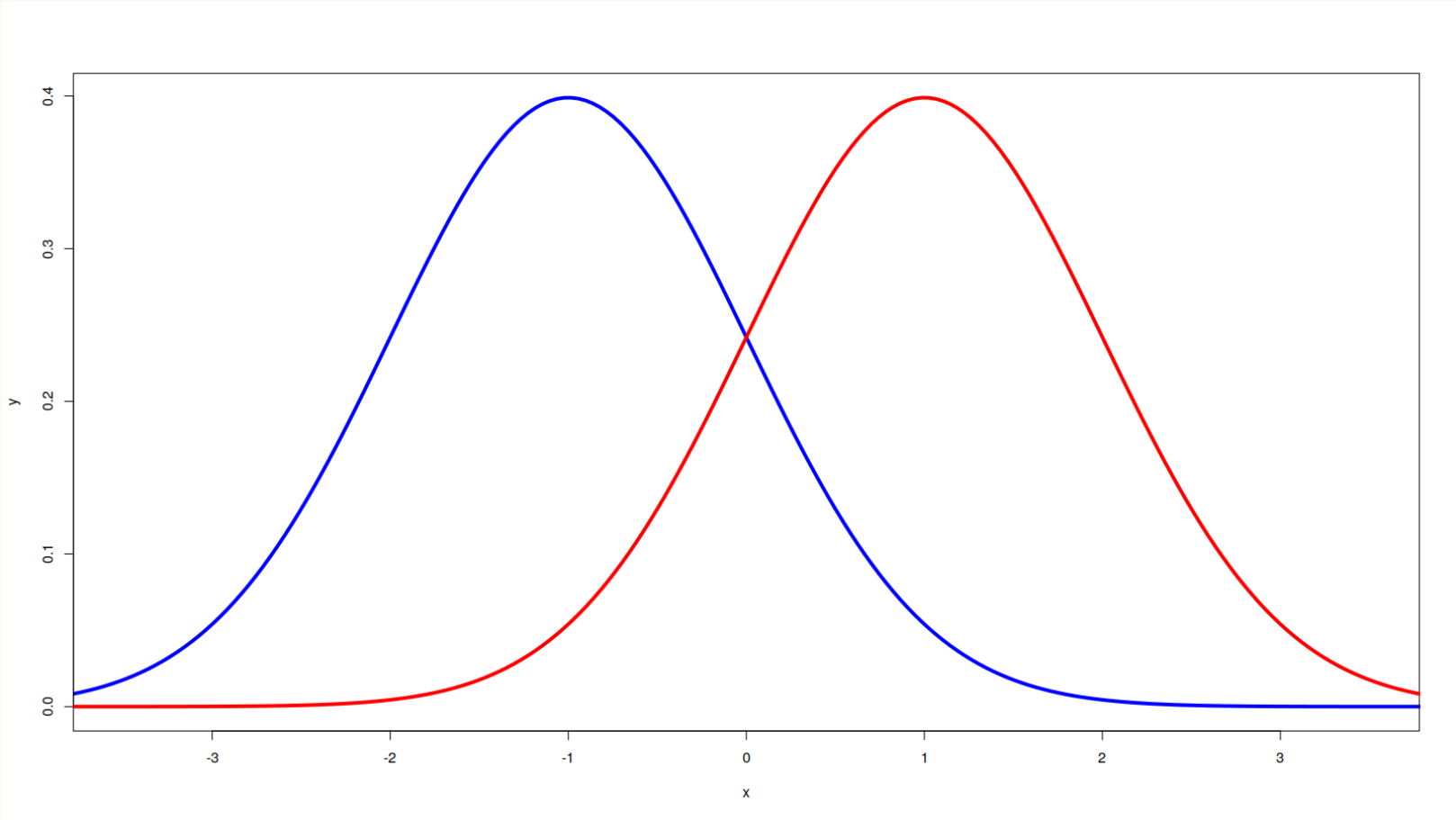



```
> t.test(a.confidence,b.confide

        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.

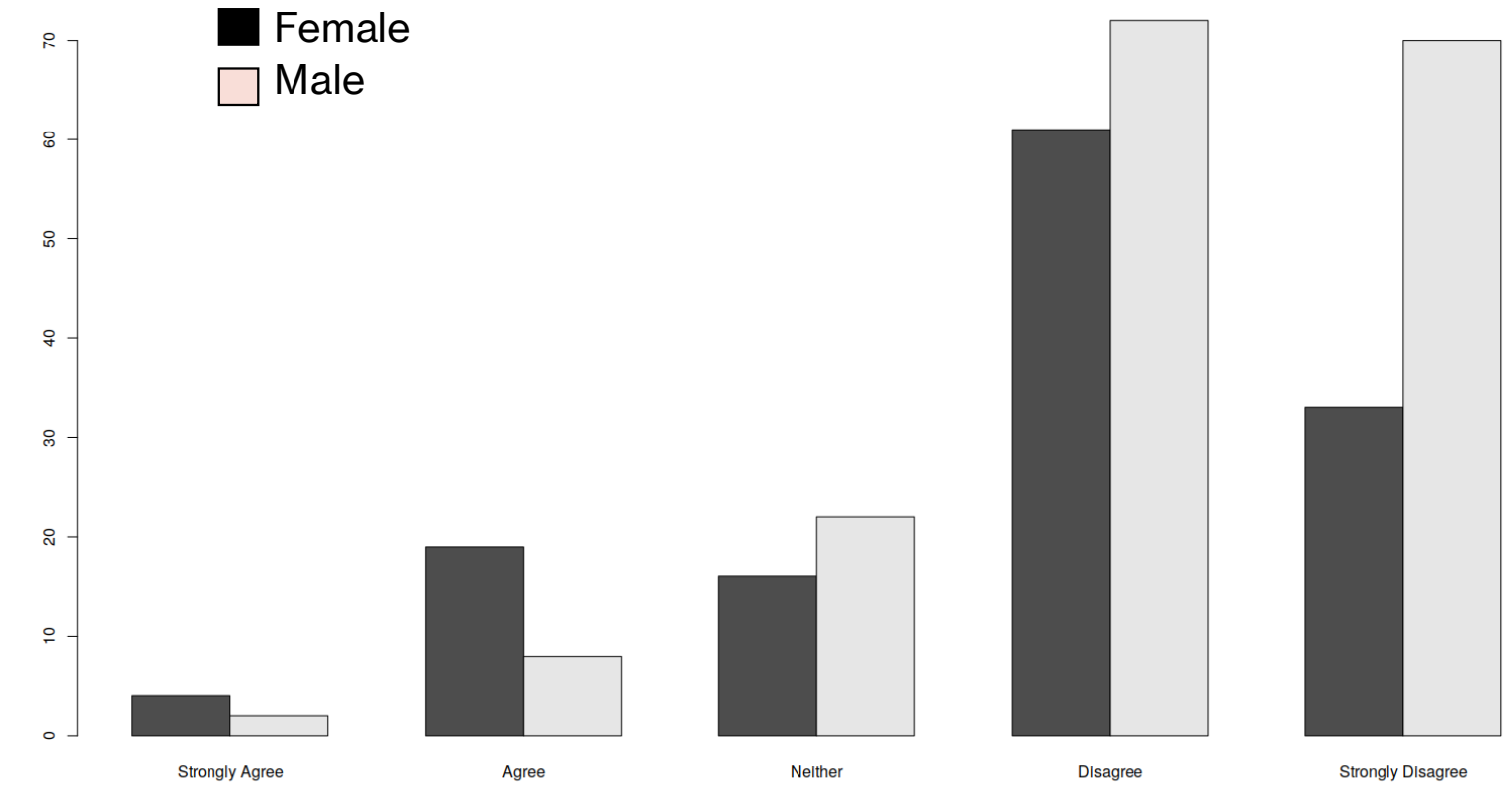
Research Question: Do women and men feel like they ask others for technical help with different frequency?

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

Between-subjects

Independent: Gender

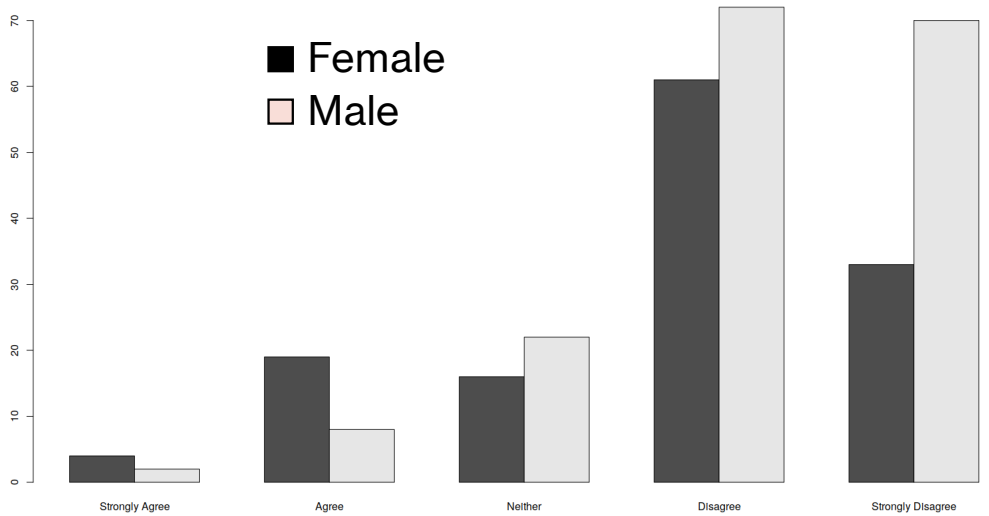
Dependent: Agreement



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

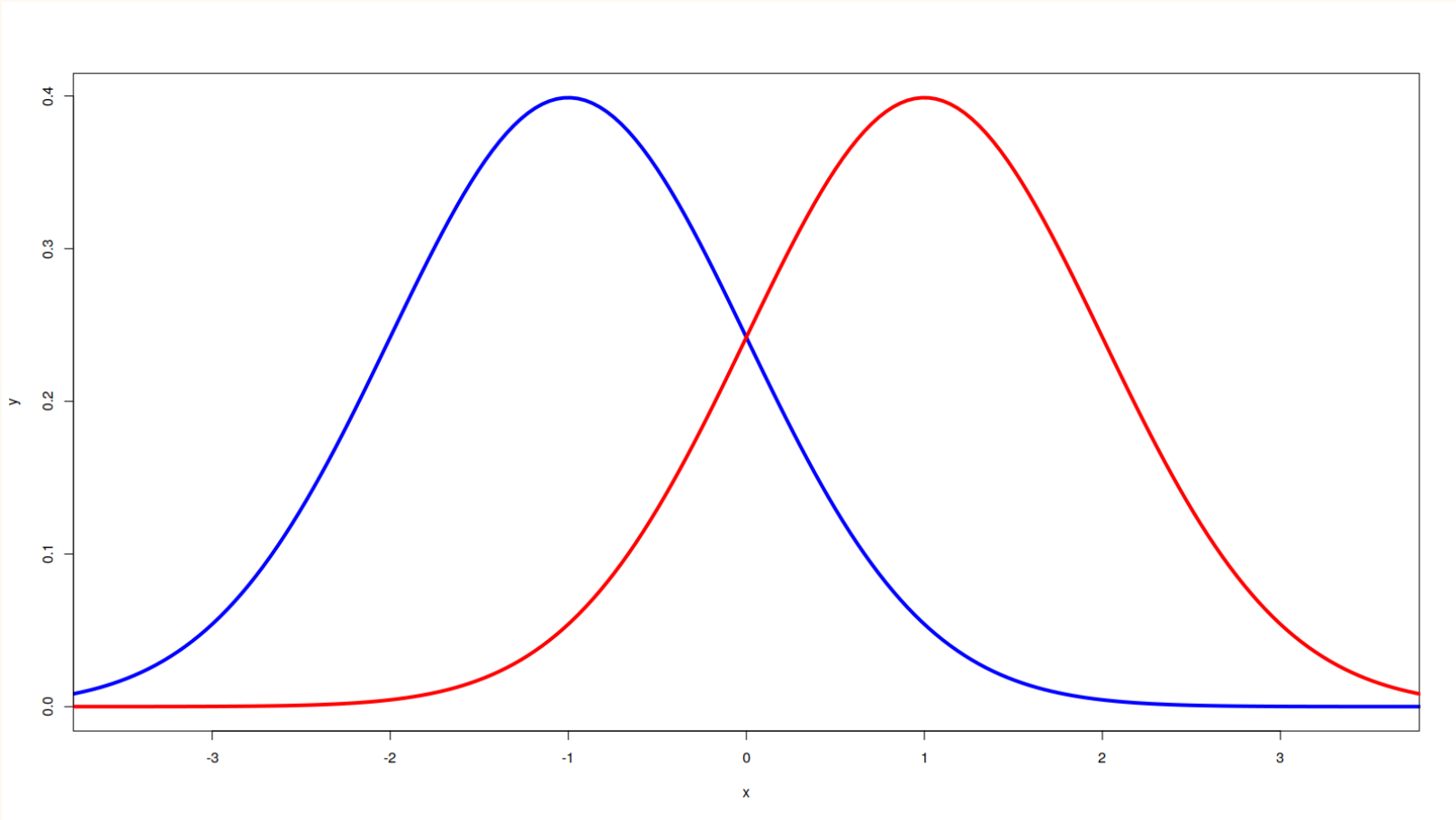



```
> t.test(as.numeric(d$i_ask_others_for_he

        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

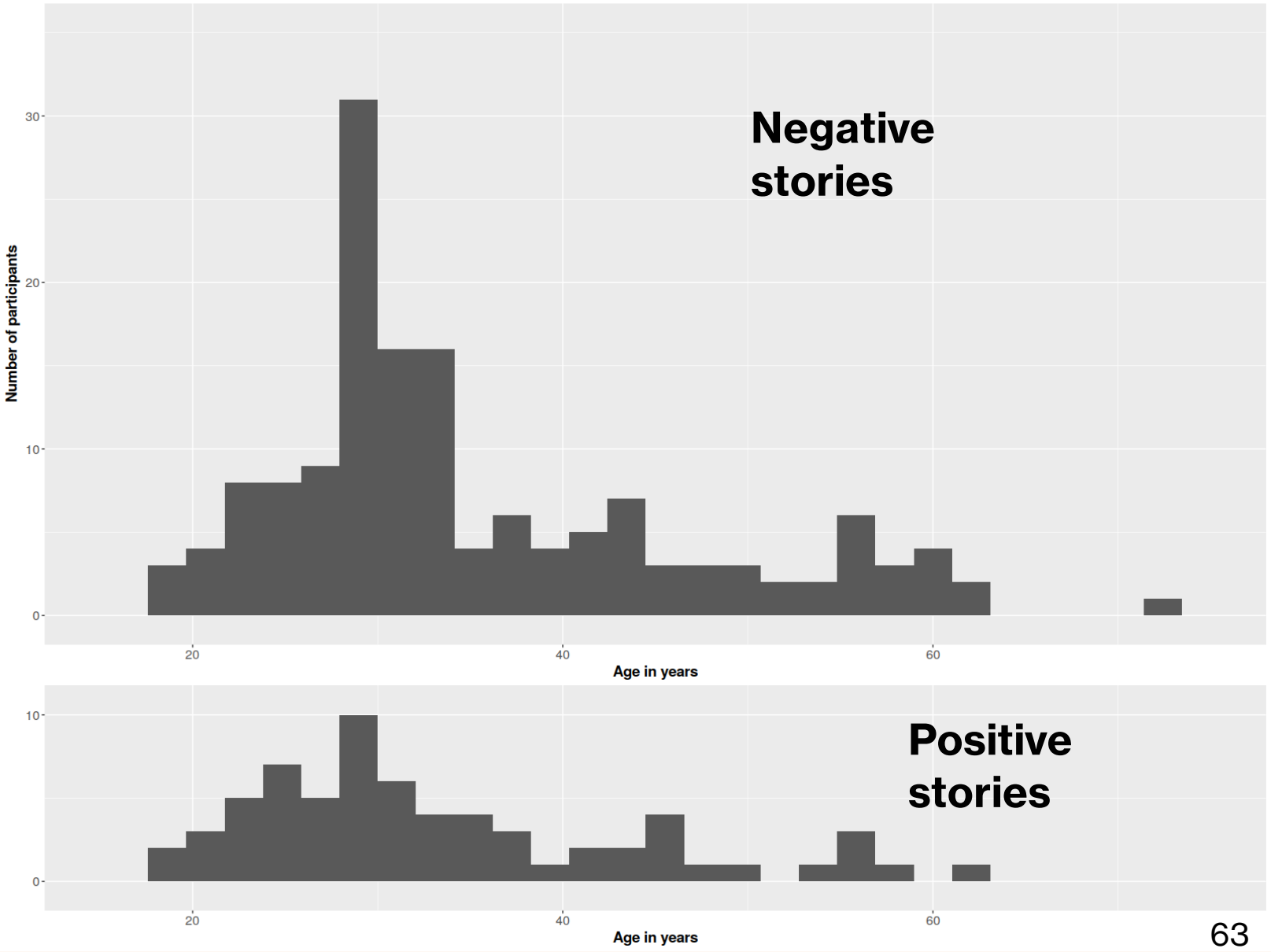


I asked participants to tell me a story about a prior software update.

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

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

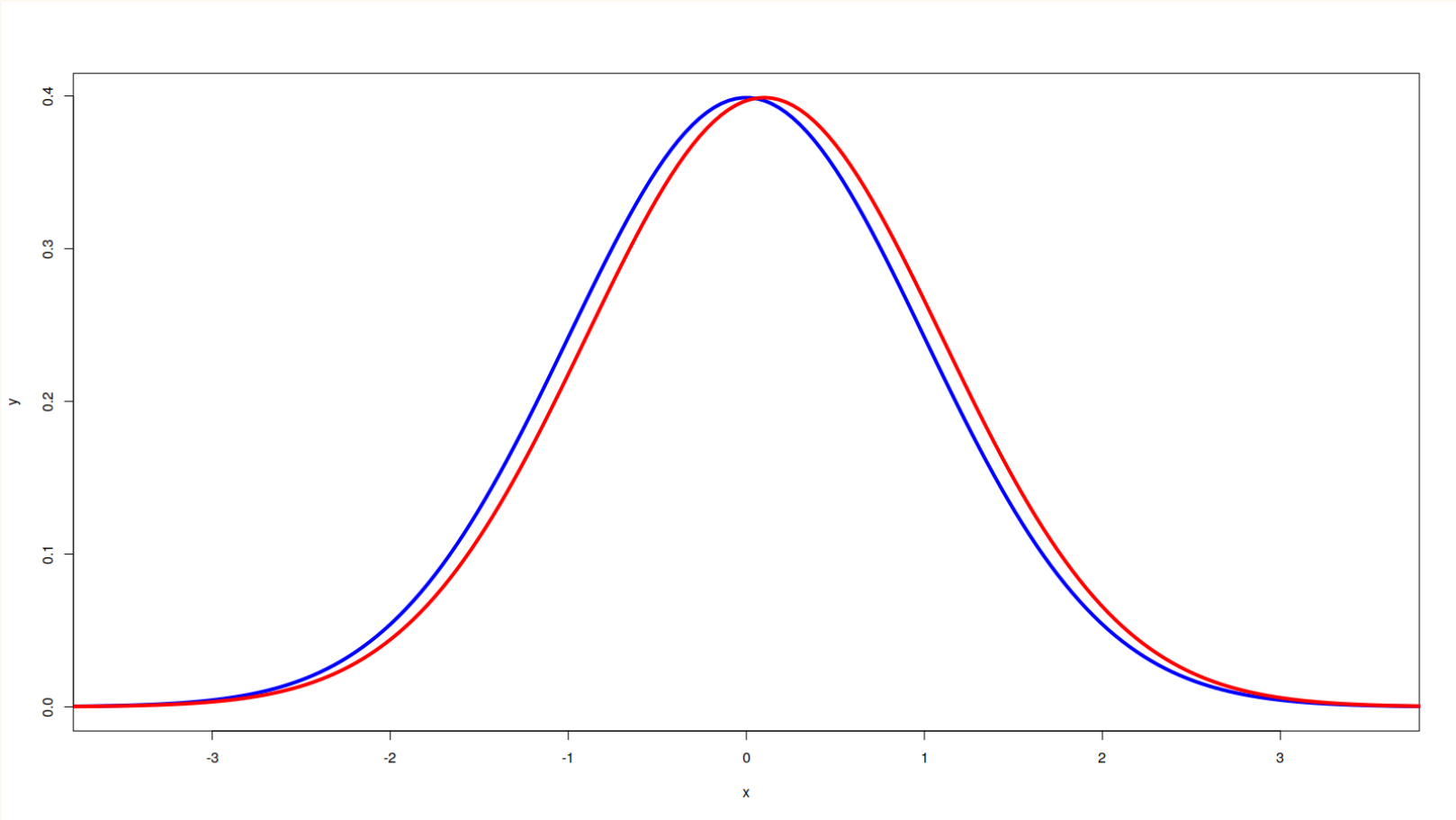
Between-subjects
Dependent:
- Age
- Numerical
Independent:
- Negative or Positive
- Binary



```
> 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

Take-home

- **(Blog)** Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M. and Holz, T., 2024, May. A representative study on human detection of artificially generated media across countries. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 55-73). IEEE.
- **(Blog)** Li, J., Sun, K., Huff, B.S., Bierley, A.M., Kim, Y., Schaub, F. and Fawaz, K., 2023, May. “It’s up to the Consumer to be Smart”: Understanding the Security and Privacy Attitudes of Smart Home Users on Reddit. In *2023 IEEE Symposium on Security and Privacy (SP)* (pp. 2850-2866). IEEE.