

From Idea to Conclusion: A Statistical Perspective

**The School for Science and Math at Vanderbilt
March 4, 2026**

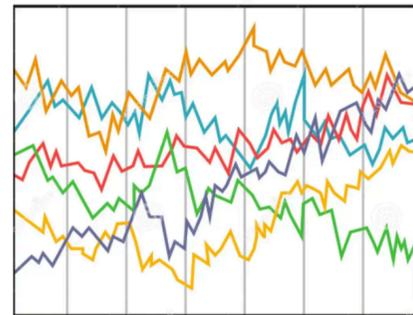
But first, who am I?



**Ashley Mullan, MS
PhD Student
Department of Biostatistics**

What's on the agenda for today?

1. Statistics & the Research Process: How do they fit together?
2. Your New Best Friend: $y = mx + b$
3. Interpretation: What do these numbers mean in context?
4. A Few Roadblocks & How to Get Around Them
5. Your Turn!

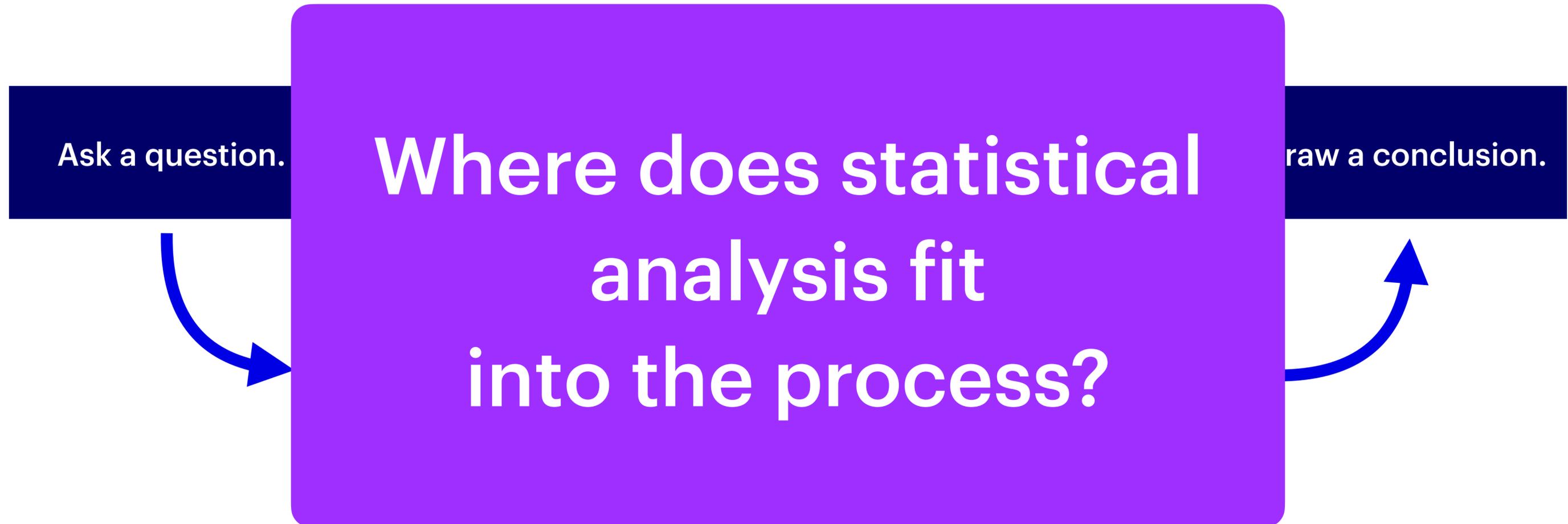


Statistics & the Research Process: How do they fit together?

Steps of the Research Process



Steps of the Research Process



the correct answer?

EVERYWHERE!

Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.

Does having fun make you do better in school?

Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.



what's wrong with this question?

Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.

Do high school students do better in school if there are memes in their lessons?

did we fix it?

Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.



Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.

Do high school students tend to have higher final grades if there are memes in their lessons?

how about now?

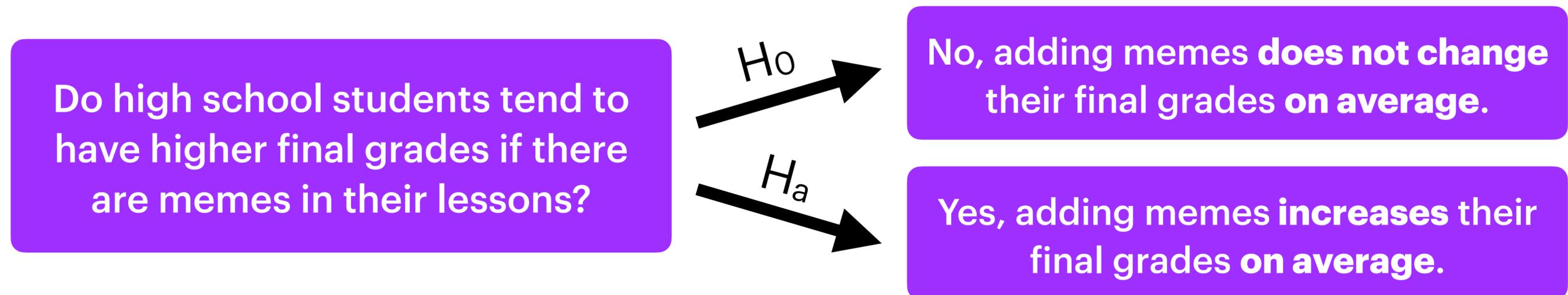
Step 1: Ask a question.

- Your question should be **specific** and **measurable**. Think about **who** we want to study, **what** we want to measure about them, and **how** we want to measure it.
- It's ok if it doesn't start out that way! We can always **refine** the question.



Step 2: Make a hypothesis.

- A **hypothesis** is your best guess about what the answer to your question may be.
- We usually make two hypotheses, the **null hypothesis** and the **alternative hypothesis**.
- The **null hypothesis (H_0)** is the skeptical answer to your question.
- You hope to find evidence for the **alternative hypothesis (H_a)**.



Step 3: Gather your data.

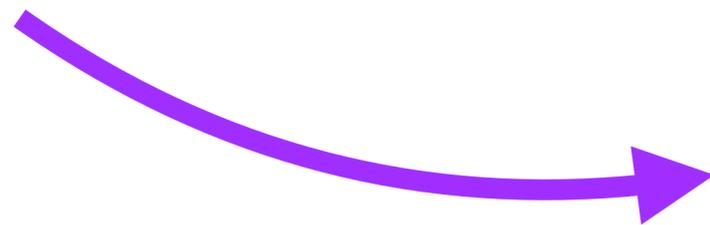
- There are **many statistical considerations** to think about when you gather data.
- Some questions you might want to ask are:
 - **Who** am I sampling? Do I have any **criteria** for my subjects?
 - **How** am I sampling? Do I pick randomly? Am I trying to balance any factors?
 - **What measurements** am I recording for each subject? What's my **unit**?
 - **When** am I sampling? Do I have to make observations more than once per subject?
 - How do I set up my comparison? Is there a **control** group?

Step 3: Gather your data.

- **Who** am I sampling? Do I have any **criteria** for my subjects?
- **How** am I sampling? Do I pick randomly? Am I trying to balance any factors?
- **What measurements** am I recording for each subject? What's my **unit**?
- **When** am I sampling? More than once per subject?
- How do I set up my comparison? Is there a **control** group?

Step 4: Analyze the data.

- There are many different ways to analyze data, and your method depends on your audience!
- Sometimes, you might want to make a **graph** or a **table** summarizing your variables.
- You can run a **hypothesis test** or a **statistical model** to generate **quantitative evidence** that can help you decide which hypothesis makes more sense.
- You can make the jump from information about your **sample** to the larger **population** by using a **confidence interval**.

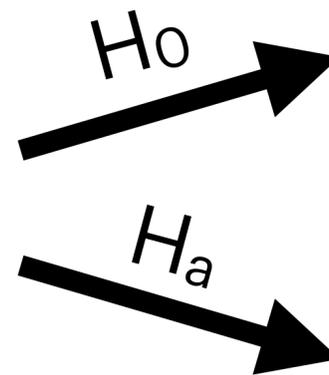


Don't worry if these are new, more on these later!

Step 5: Draw a conclusion.

- Once you've made your case, go back and use the **evidence** you've gathered to answer your **original question**!
- Ask yourself if what you've measured is a **useful** comparison.
- Suppose we noticed that the average final grade in the memes class was an **A-**, but the students without memes only averaged a **B+**. Which hypothesis makes more sense?

Do high school students tend to have higher final grades if there are memes in their lessons?



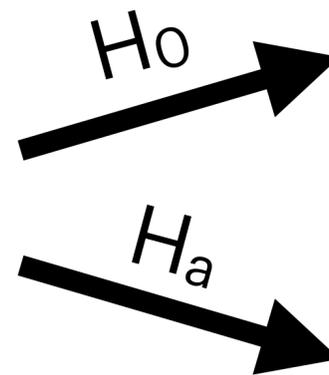
No, adding memes **does not change** their final grades **on average**.

Yes, adding memes **increases** their final grades **on average**.

Step 5: Draw a conclusion.

- Once you've made your case, go back and use the **evidence** you've gathered to answer your **original question!**
- Ask yourself if what you've measured is a **useful** comparison.
- Suppose we noticed that the average final grade in the memes class was an **A-**, but the students without memes only averaged a **B+**. Which hypothesis makes more sense?

Do high school students tend to have higher final grades if there are memes in their lessons?



No, adding memes **does not change** their final grades **on average**.

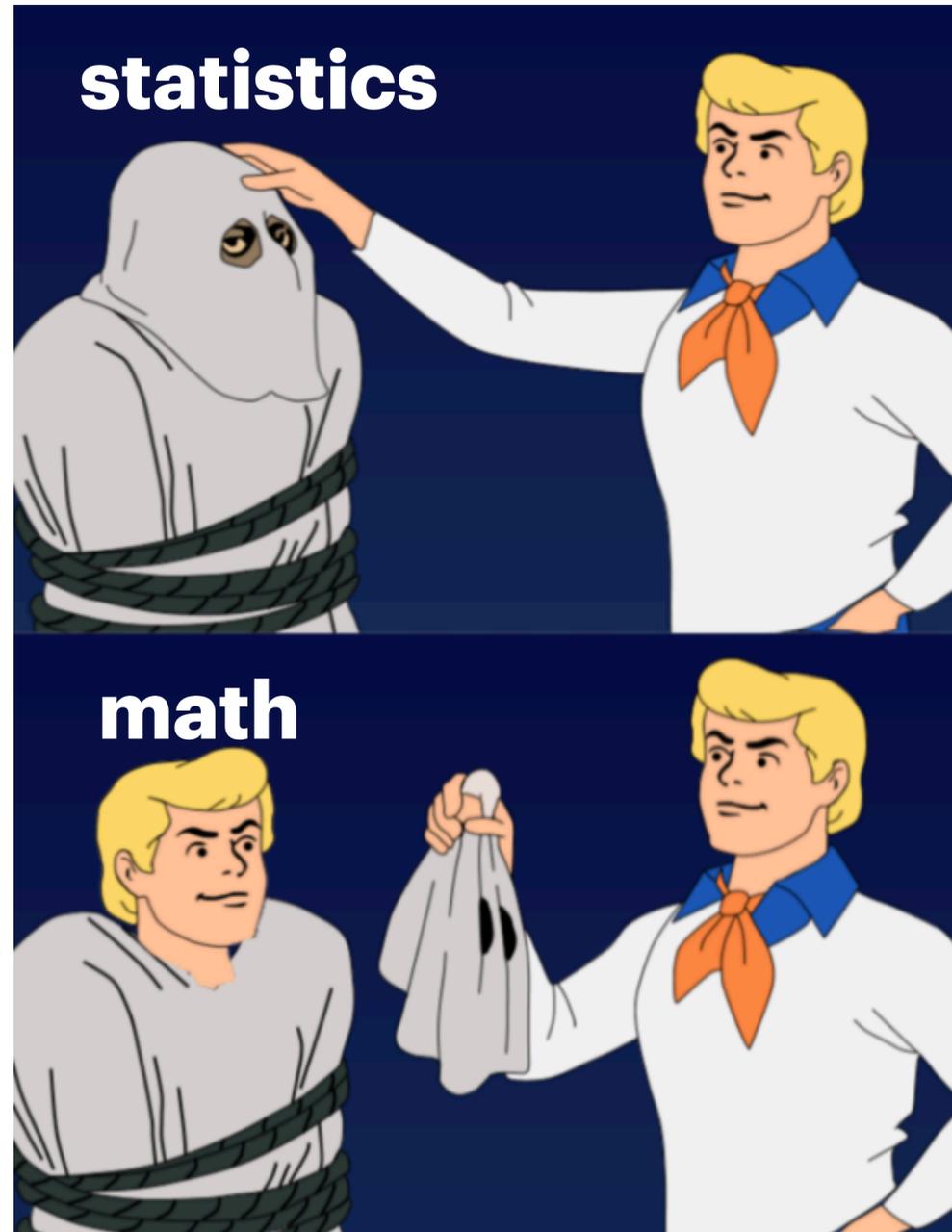
Yes, adding memes **increases** their final grades **on average**.



Your New Best Friend:

$$y = mx + b$$

But wait, wasn't that the line of best fit from my math classes?



Translators to the rescue!

**difference between
groups (slope)**

**independent
variable (predictor)**

$$y = mx + b$$

**dependent variable
(outcome)**

**baseline outcome
(intercept)**

Translators to the rescue!

**difference in final
average grades**

**for memes: $X = 1$
no memes: $X = 0$**

$$y = mx + b$$

average final grade

**average final grade for
the no meme group**

What if other factors play a role?

what adding memes
does to the grades

baseline grade (no
memes and no iPads)

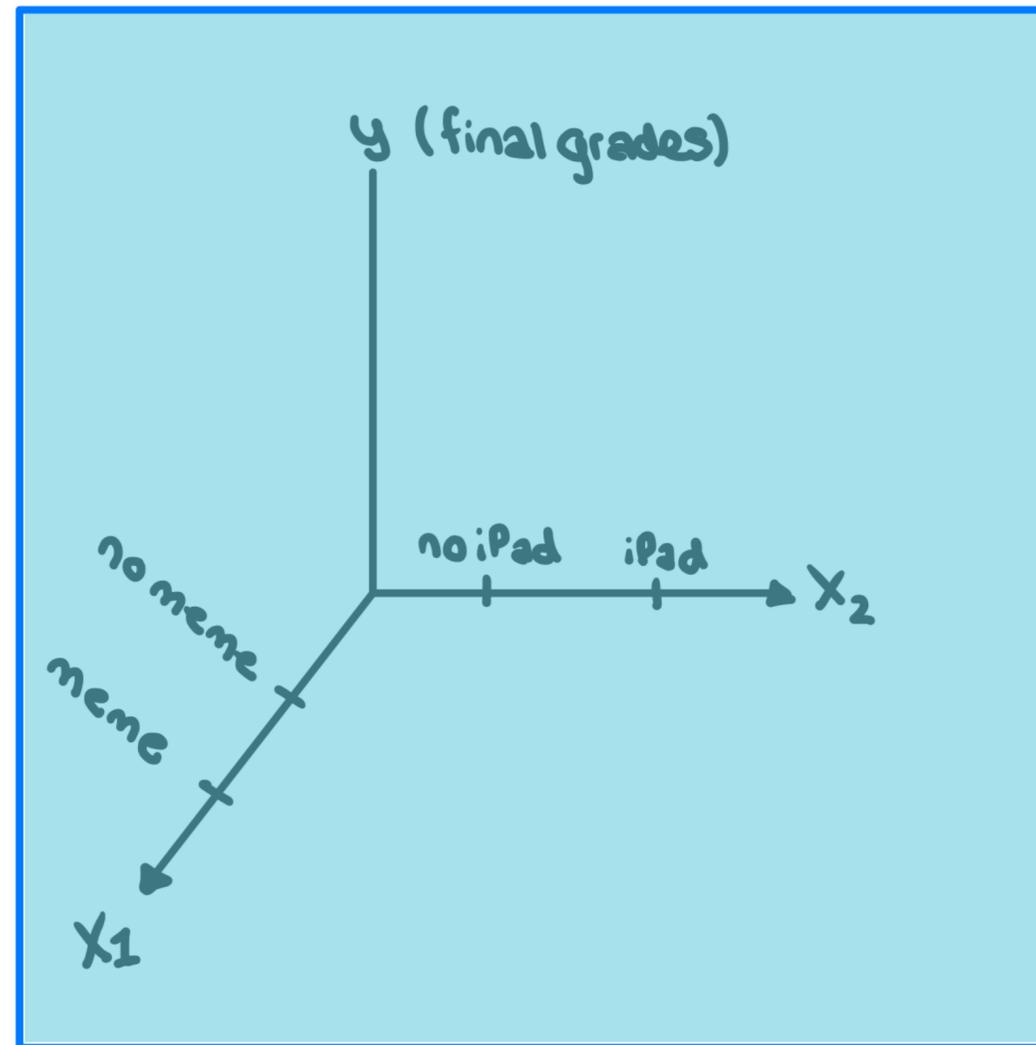
$$y = m_1x_1 + m_2x_2 + b$$

what using iPads does
to the grades

What if other factors play a role?

what adding
does to the g

$$y = m_1 x_1$$



baseline grade (no
meme and no iPads)

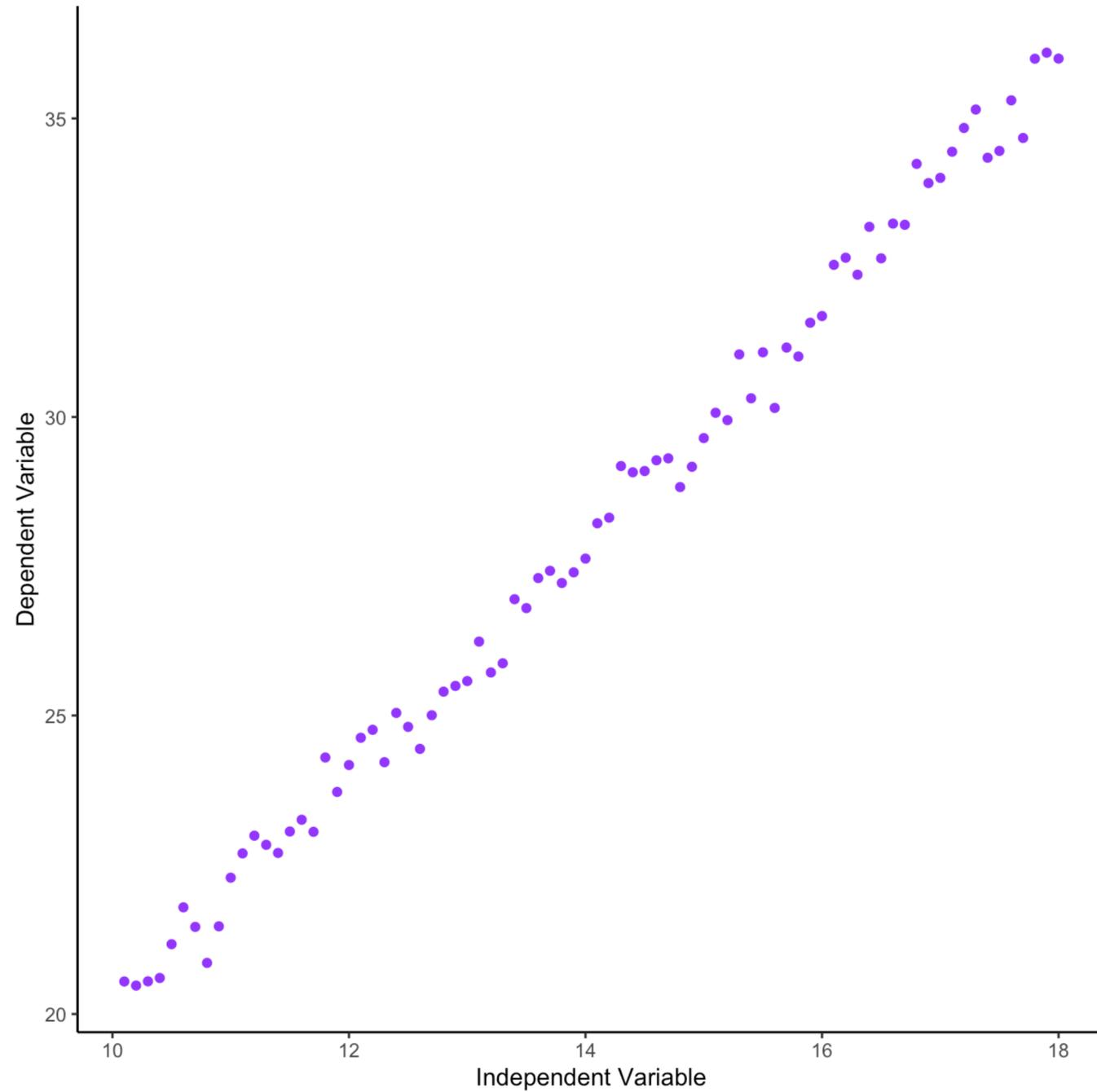
$$x_2 + b$$

s does

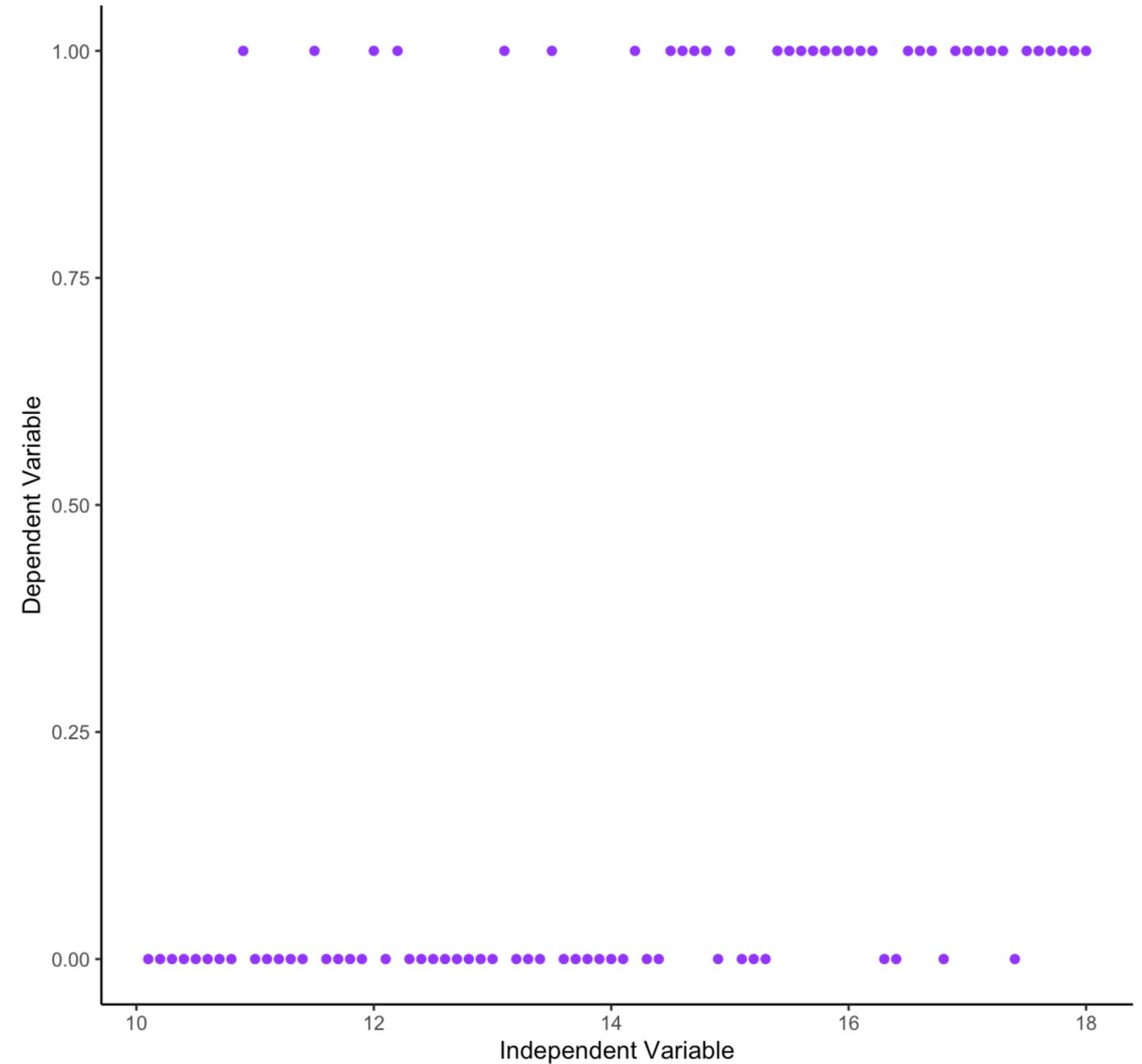
is

Is a line of best fit always the best idea?

Graph 1



Graph 2



 **New Question Alert:** 

Do high school students tend to pass their finals if there are memes in their lessons?

related to:

odds of passing with memes
odds of passing without memes

$$f(y) = mx + b$$

in this case:

probability of passing
probability of failing

But what are odds?

probability of outcome happening
probability of outcome not happening

Interpretation:
What do these numbers
mean in context?

Step 1: Peek* at your data.

- The **head()** command looks at the first few rows of your data.
- We tell it that the data is stored as **data_mg** and that we want to look at the first **6 rows**.
- This helps us understand how our data is **encoded**.
- How do we know that a student was in the meme class?

*we kind of already did this another way too when we looked at the plot

```
> head(data_mg, 6)
  memes grades
1     0    91
2     1    95
3     0    90
4     1    95
5     0    90
6     1    96
```

Step 2: Fit the model.

- We save the output of the **lm()** command (our model) as **linear_model**.
- We tell the **lm()** command which **formula** and **data** to use.
- Then, we can use **summary()** to print out some useful information about our model!
- Can you pick out what the model thinks “m” and “b” are?

```
> linear_model <- lm(formula = grades ~ memes, data = data_mg)
> summary(linear_model)

Call:
lm(formula = grades ~ memes, data = data_mg)

Residuals:
    Min       1Q   Median       3Q      Max
-2.86  -0.78   0.14   1.14   2.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.7800     0.1447  620.45  <2e-16 ***
memes         5.0800     0.2046   24.82  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.023 on 98 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8614
F-statistic: 616.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Step 2: Fit the model.

- We save the output of the **lm()** command (our model) as **linear_model**.
- We tell the **lm()** command which **formula** and **data** to use.
- Then, we can use **summary()** to print out some useful information about our model!
- Can you pick out what the model thinks “m” and “b” are?

```
> linear_model <- lm(formula = grades ~ memes, data = data_mg)
> summary(linear_model)

Call:
lm(formula = grades ~ memes, data = data_mg)

Residuals:
    Min       1Q   Median       3Q      Max
-2.86  -0.78   0.14   1.14   2.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.7800    0.1447   620.45  <2e-16 ***
memes         5.0800    0.2046   24.82  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

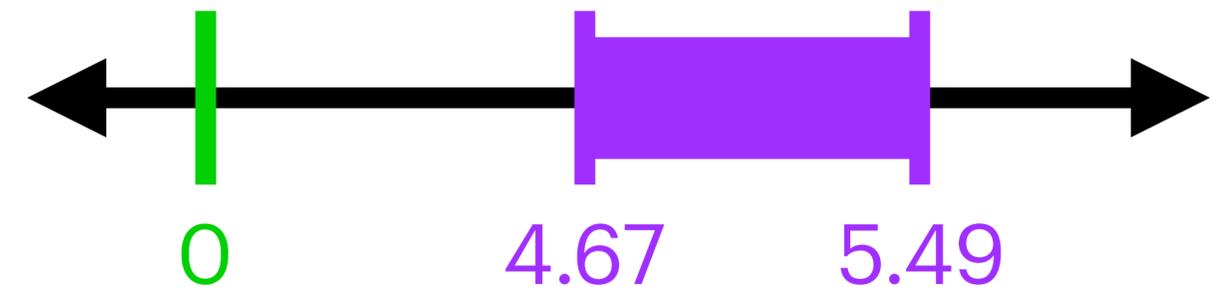
Residual standard error: 1.023 on 98 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8614
F-statistic: 616.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

Step 3: Use the model!

- The **confint()** command takes our saved model and spits out a **confidence interval** for each estimated **parameter**.
- A **confidence interval (CI)** describes values that would **make sense** for our parameters according to our data!
- Our parameters are **m** and **b**, but what we cared about (the one in our null hypothesis) was **m**!
- If adding memes **doesn't change** the average final grades, then m should be **zero**.

pro tip: always report **both** the parameter **AND** the CI!

```
> confint(linear_model)
              2.5 %      97.5 %
(Intercept) 89.492843 90.067157
memes        4.673898  5.486102
```



Does our **95% confidence interval** catch our **null hypothesis**?

Step 4: Make a conclusion.

*The alpha level is 100% - CI%.

- If our **null hypothesis** is **outside** the confidence interval, we **reject** it (at an **alpha level*** we picked in advance).
- If our **null hypothesis** is **inside** the confidence interval, we **fail to reject** it (at that alpha level).
- Remember that “fail to reject” and “accept” do not mean the same thing!
- In our example, we estimate **m = 5.08** and **reject our null hypothesis!** We have evidence to suggest that memes in the notes increase the average final grade.



Step 5: Why does your conclusion matter?

- We hope for **significant** results, but there are two kinds of significance!
- If our results are **statistically significant**, that means we can reject our null hypothesis. Our data sample gives us evidence for the **alternative!**
- Statistics alone can't help you determine if your results are **practically significant**. That requires **contextual knowledge!**
- In our example, we estimated that a meme group would do about **five points** better than a non meme group. This is our **effect size**.
- Pretend you're the teacher. What's the **minimum improvement** you'd want to see in your students' grades before taking the time to put memes in all of your slides?

A Few Roadblocks & How to Get Around Them

Watch out, your data are rarely perfect!

- It's always important to **peek** at your data to see if there's anything looking a little funny that you can easily fix.
- Look out for **missing values**! In R, something that's missing will show up as an **NA**. Check with your research team to make sure you handle missingness **consistently**!
- Plot your data to check for **outliers**, or data points that are very different from the others. Since they behave **differently**, they might mess with your trend. Check with your team on how to handle them **consistently**!

Watch out, your data are rarely perfect!

- Check and make sure your **encodings** match what you think they should be! For example, we could have encoded our glasses differently. If we had typed “Yes” instead of “1” in the spreadsheet, R might not **treat them the same way** in the model.
- Check for obvious **data errors**. For example, “yes” and “yasssss” might mean the same thing to you, but they don’t mean the same thing to a **computer**! You can always **manually fix** those before modeling.
- Always **keep a record** of any changes you make to your data!

Your Turn!

Let's turn this into a research question about vision!

enter the data!

Thanks for listening!