Overview of Hypothesis Testing

Statistics

Theoretical

Applied

Descriptive

Describe, organize and summarize information about an entire population
Overall monthly churn rate is 5%.

Measure of Central Tendency
Mean, Median, Mode

Measure of Variability
Range, Variance, Standard Deviation

Inferential

Used to generalize about a population based on a sample of data
E.g. 5% of customers churned last month —> 5% of customers will churn each month.

Hypothesis Testing
(While there are other types of inferential statistics)
1. Make inferences from the sample and generalize them to the population.
2. Compares, tests and predicts future outcomes
3. Final result is the probability scores.
4. Draws conclusions about the population that is beyond the data available.

Everything is Boring Except the Population

Population: Collection of all items **we are interested in**.

Anything else but the entire population is boring (and therefore we are not interested in it).

Sample: Any subset (or sub-collection) of our population.

One sample might be more representative than another. Is it the population? No. So it's **boring**.

Observation: Any single element of the sample.

Also not the population, therefore it is **boring**.

Cassie Kozyrkov, Chief Decision Scientist at Google

If we have a sample…

Is the population:

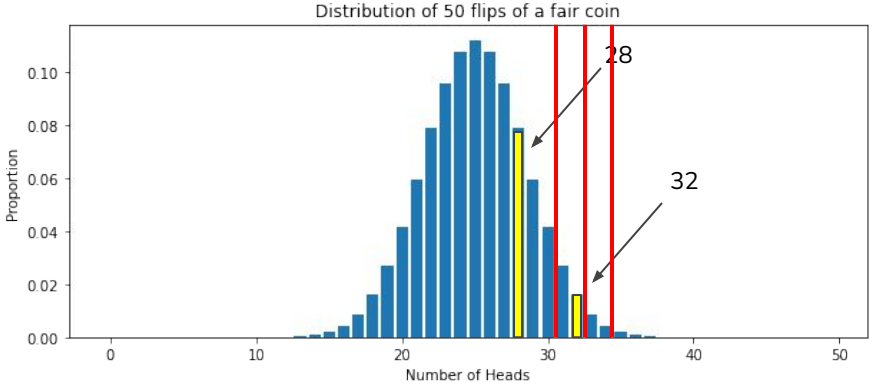
a) All statistics students?
b) All students?
c) All humans?
d) All mammals?
e) All living organisms?
f) All matter in the universe?

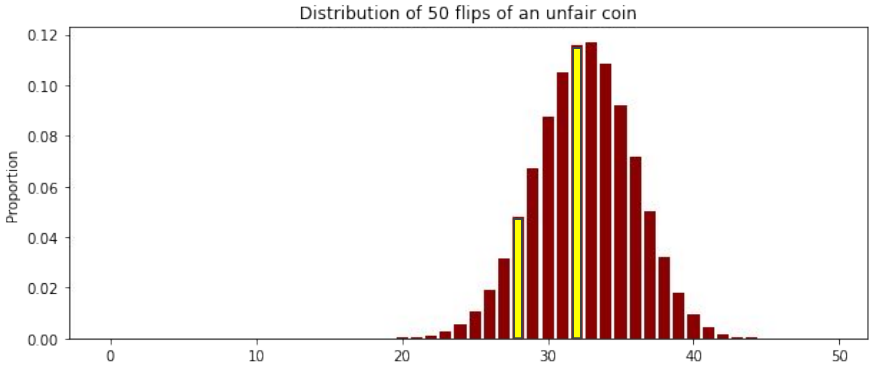
Answer: It's whatever you are interested in! But your sample is probably more **representative** of one of these answer choices than the others.

Let’s play a game…

= You give me $1

= I give you $1





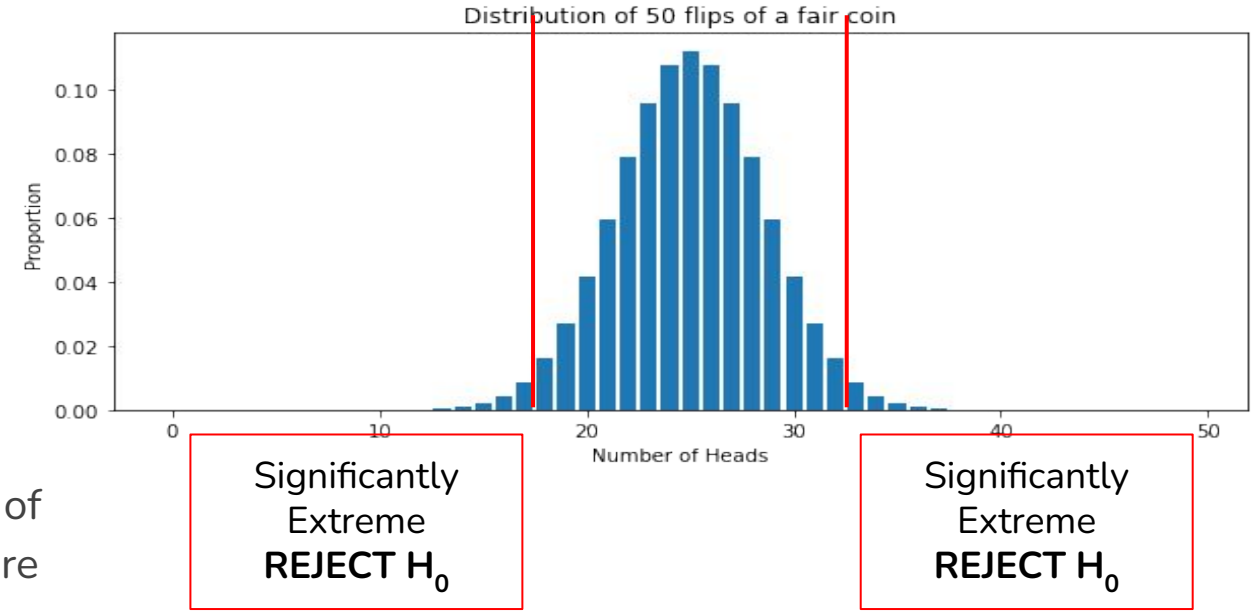
How confident are you about your claim?

Null hypothesis:

H_0 : The coin is fair

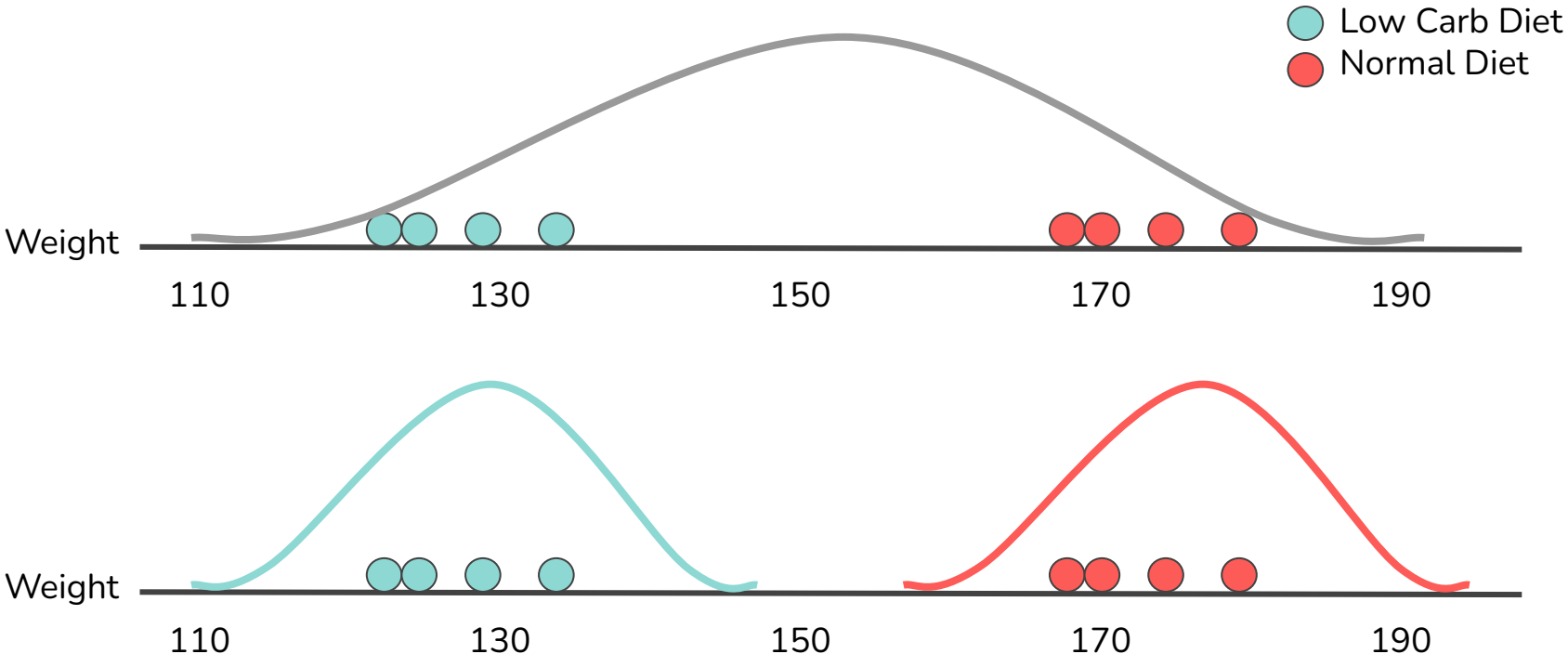
Alternative hypothesis:

H_A: The coin is not fair

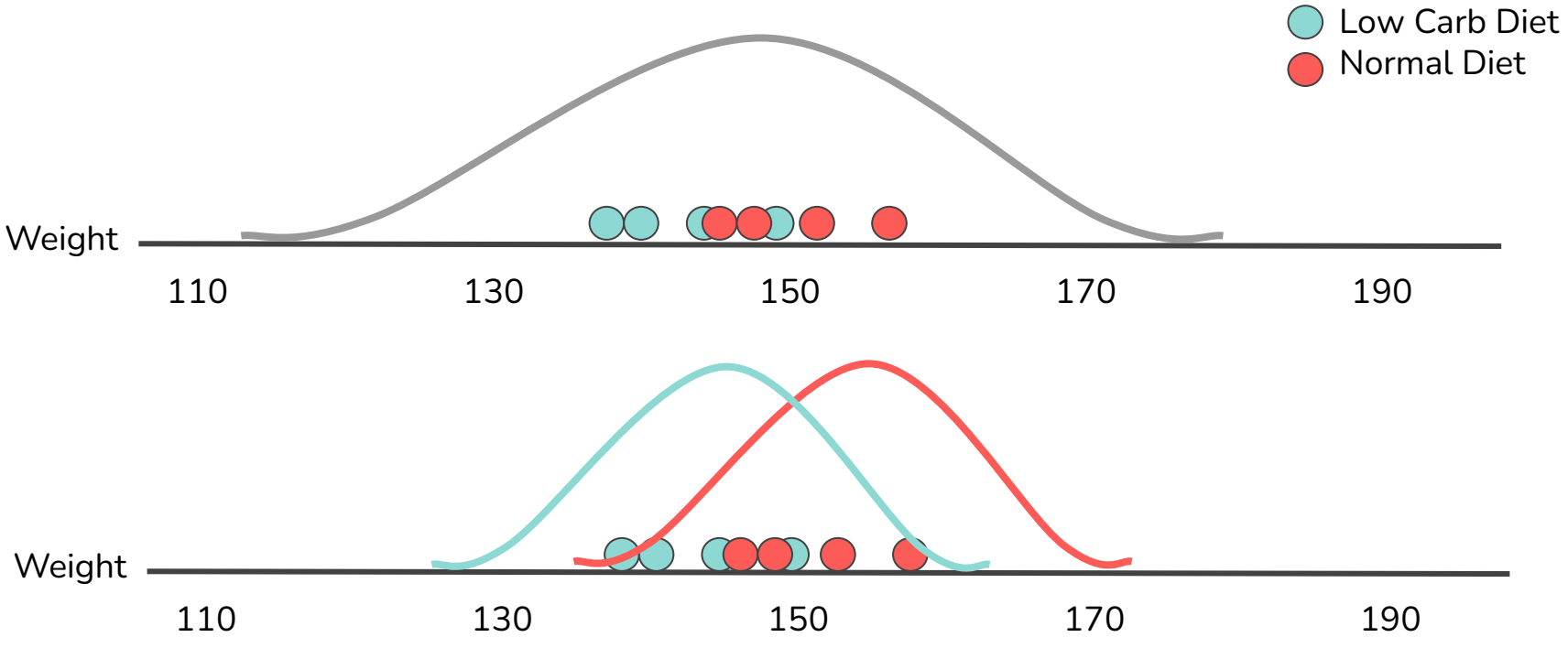


P-value: The probability of getting a result as or more extreme than our own

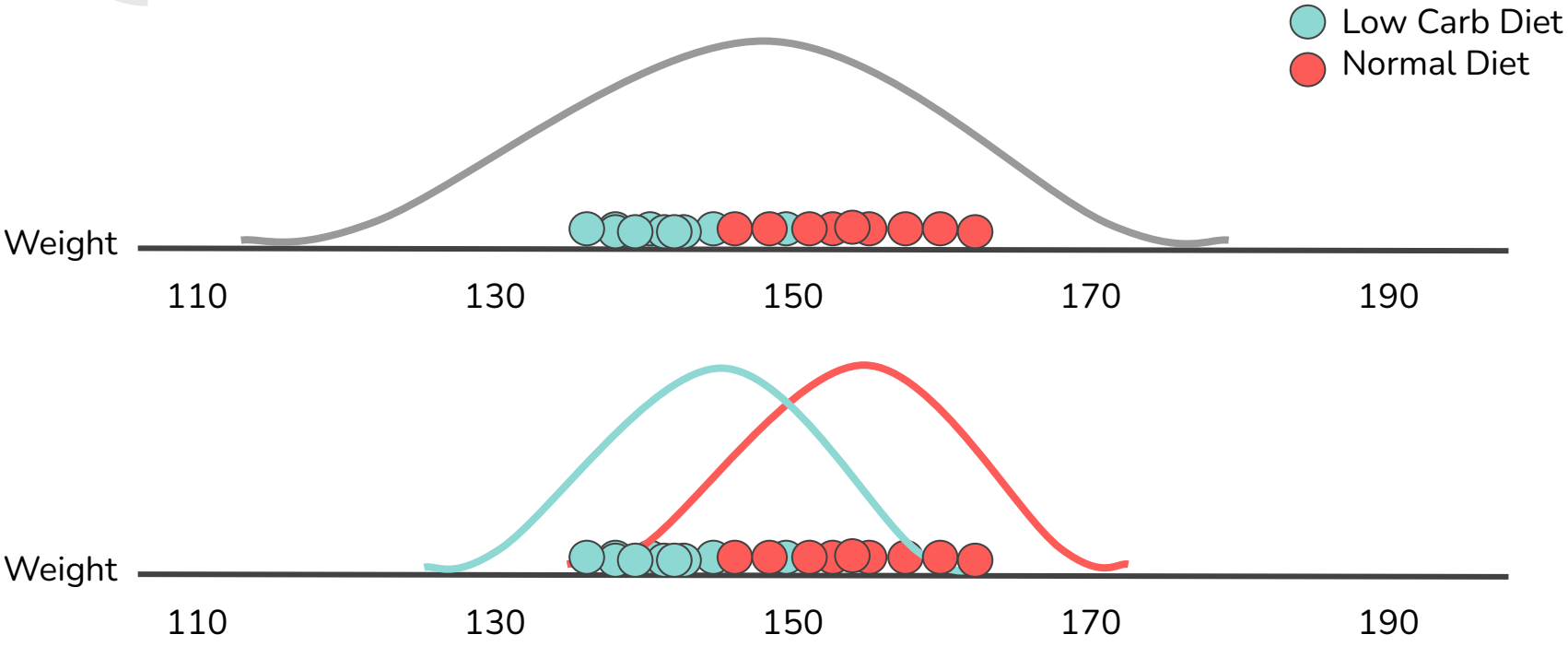
Crafting a Null Hypothesis



Crafting a Null Hypothesis

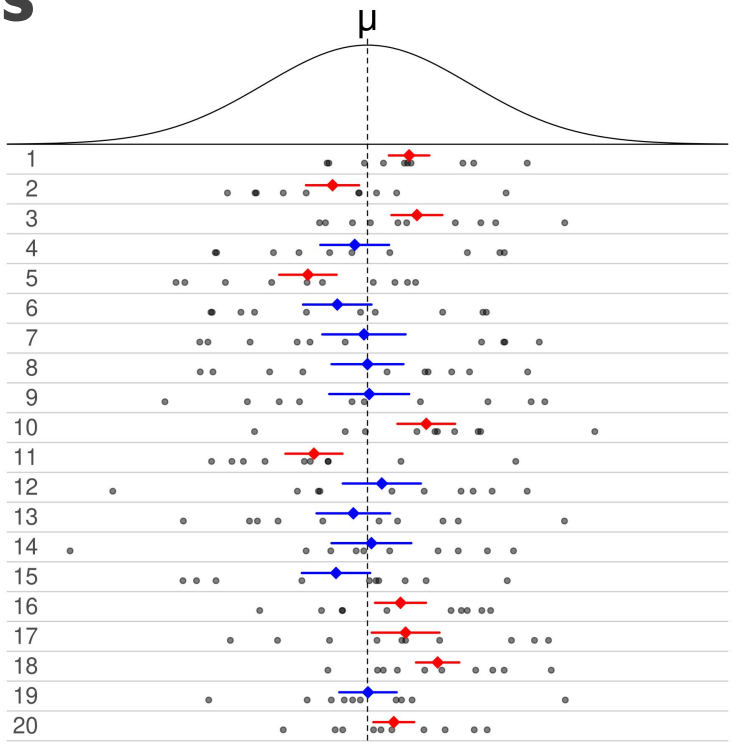


Crafting a Null Hypothesis



Confidence Intervals

“The latest ABC News-Washington Post poll showed 56 percent favored Peabody while 39 percent would vote for Flimflam. The ABC News-Washington Post telephone poll of 1,014 adults was conducted March 8-10 and had a *margin of error of plus or minus 3.5 percentage points.*”



What does it mean to reject the H_0?

- The alternative hypothesis is favored on the basis of probability. The test is not rejecting the null hypothesis, **YOU** are rejecting the null hypothesis and are using the results to provide support for your decision.

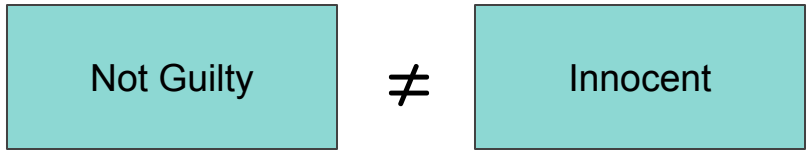
- Did we **prove** anything?

What does it mean to fail to reject the H_0?

We are not “proving” anything. We do not “accept” the null hypothesis.

“Absence of evidence is not evidence of absence” - DG Altman

“Failing to reject the H_0” is like declaring someone “not guilty” in a court of law.

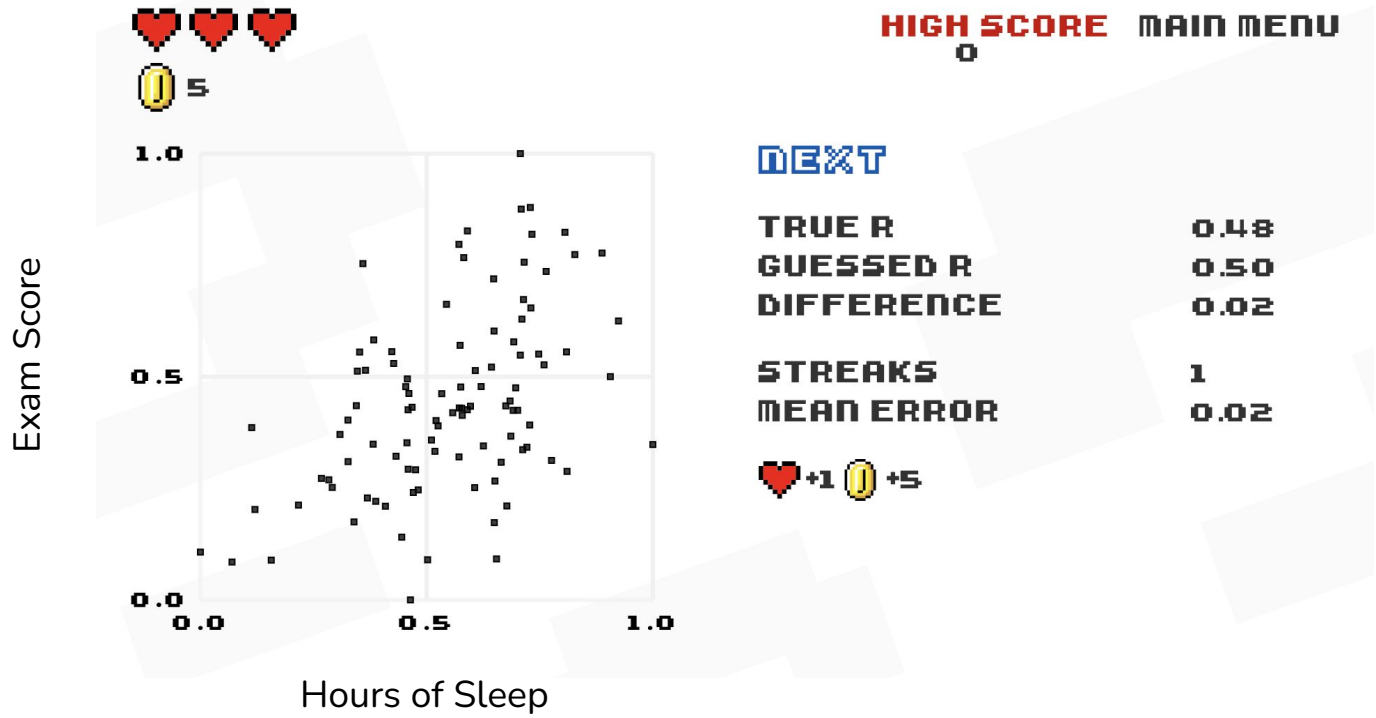


Another example: Sales Script Results

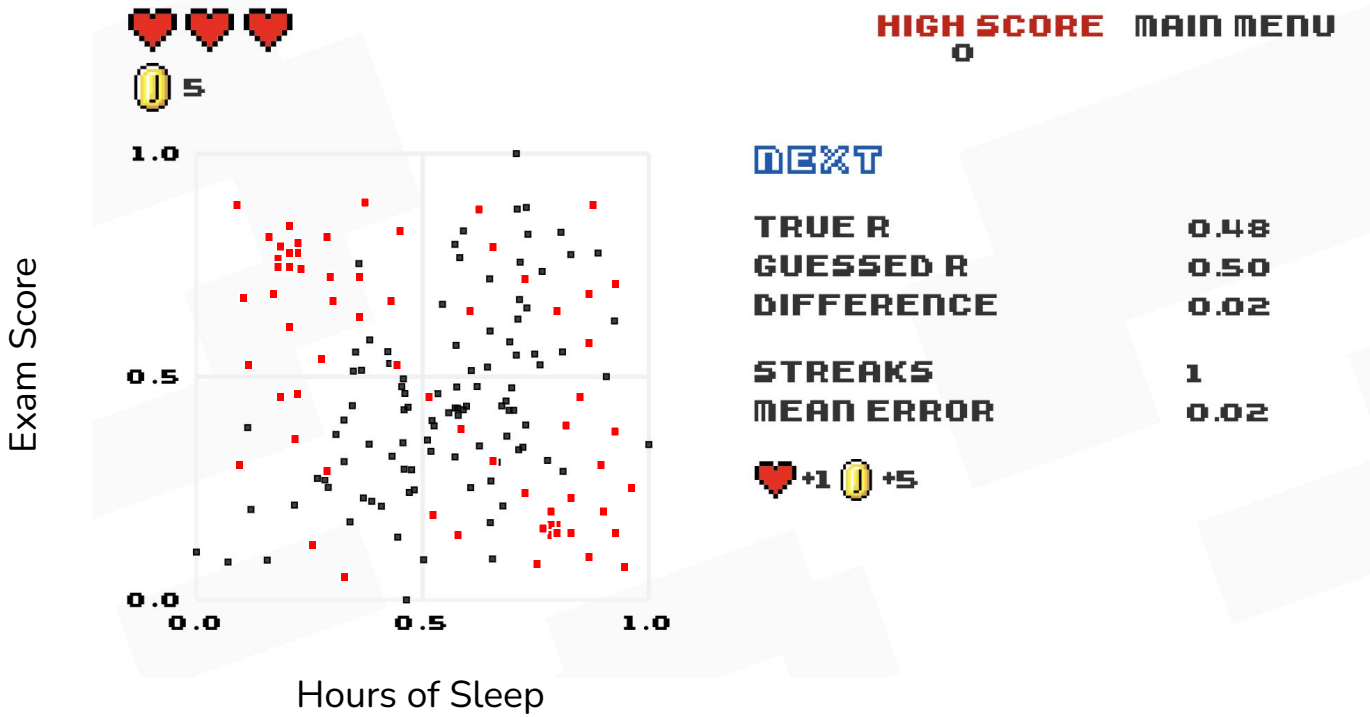
	Outcome		
Script	Sale	No Sale	Total
A	62	38	100
B	48	52	100
	110	90	200

If both groups are representative of the same population, then our samples should be similar

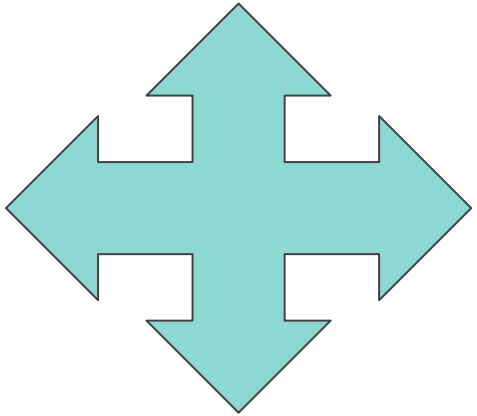
Another example: Relationship between several continuous variables



Another example: Relationship between several continuous variables



Test Statistic vs p-value



Test Statistic
Strength of test result, direction, and other details specific to the test used

p-value
Measure of extremeness

Errors in Hypothesis Testing

	H_0 is true	H_0 is false
Accept H_0	True Negative	False Negative (Type II Error)
Reject H_0	False Positive (Type I Error)	True Positive

Predicting Cat or Not Cat

Positive

IMAGE

CAT

PREDICTION

Correct? True

The prediction? Positive

TRUE POSITIVE

Predicting Cat or Not Cat

Positive

IMAGE

CAT

PREDICTION

Correct? False

The prediction? Positive

TYPE I ERROR

Predicting Cat or Not Cat

IMAGE

Positive

NOT CAT

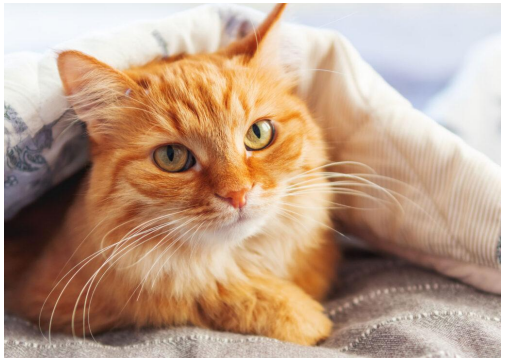
PREDICTION

Correct? True

The prediction? Negative

TRUE NEGATIVE

Predicting Cat or Not Cat



IMAGE

Positive

NOT CAT

PREDICTION

Correct? False

The prediction? Negative

TYPE II ERROR

Term	Formula / Symbol	Description
Null Hypothesis	H_0	The "default" hypothesis; usually no change, no effect, etc
Alternative Hypothesis	H_1 or H_a	
Significance Level, **False Positive Rate**	α	P(FP) = P(Type I Error)
Statistical Power	$1 - \beta$	P(Reject H_0 when H_0 is false)
False Negative Rate	β	P(FN) = P(Type II Error)
p-value	p	P(We observed this result due to chance | H_0 is true)

Hypothesis Testing Workflow

1. Choose the right type of test for your data / question
2. Set a desired confidence level and form hypothesis
3. Calculate the appropriate test statistics and p-value
4. Conclude based on the above statistics

Step 1: Choose the right type of test.

Narrow down broad questions to a single testable comparison:

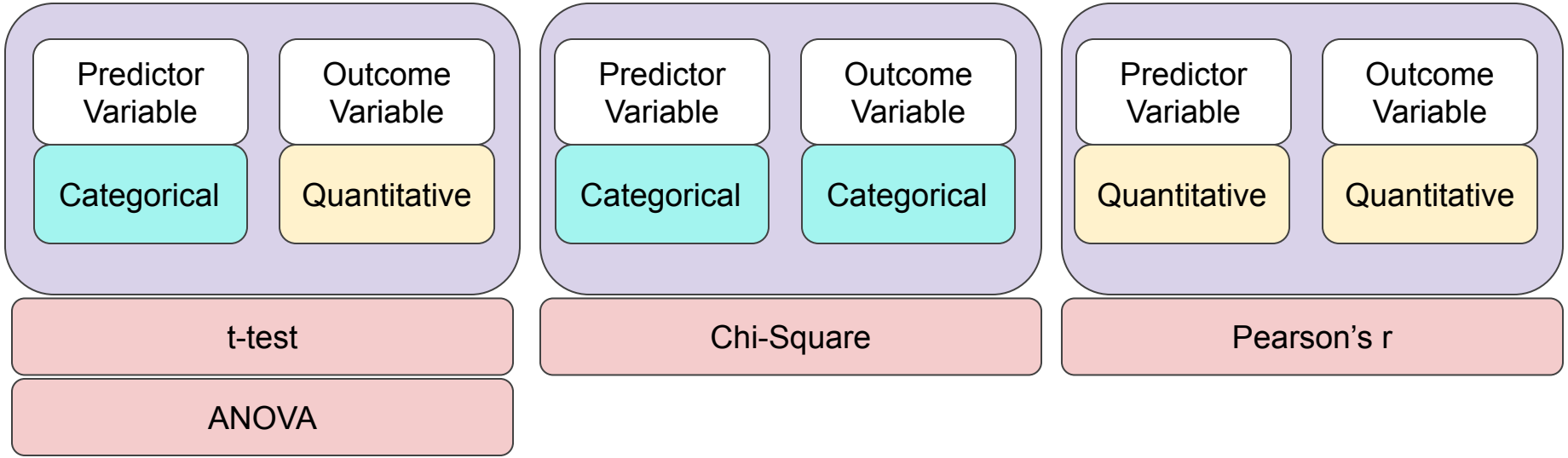
Do customers churn because their bills are too high?	→	Do those who churn spend more than those who don't churn?
Is their internet too slow? Maybe there's something wrong with certain internet options?	→	Are certain internet types more or less likely to churn?
Do customers get charged more the longer they have been with the company?	→	Is there a linear relationship between tenure and average monthly charges?

Step 1: Choose the right type of test
Check underlying assumptions

1. Sample Size?
2. Independent observations?
3. Similar variance between groups?
4. Normality of data (if quantitative)?

Step 1: Choose the right type of test

What type of variables are you working with?



Step 2: Set the desired confidence level

Depends on the consequences of being wrong:

Lenient*	Typical	Strict
α = 0.10	**α = 0.05**	**α = 0.01**

*rarely used, if ever

Step 3: Calculate test statistic and p-value

$$r_{xy} = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

```
x = df.hours_studied
y = df.exam_score

def stdev(x):
    variance = ((x - x.mean()) ** 2).sum() / n
    return sqrt(variance)

r_xy = (((x - x.mean()) * (y - y.mean())).sum() / n) / (stdev(x) * stdev(y))
r_xy
```

```
p = stats.t.sf(t, df=degf) * 2 # *2 for a two-tailed test
p
```

Step 4: Conclude based on results

- If the *p* is low, reject the H_0
- Just because something is statistically significant doesn't mean its useful
- Statistical results are not the same as controlled experimentation
- The p-value represents a **probability not certainty**: https://xkcd.com/882/

Step 4: Conclude based on results

The p-value is a function of sample size and variance. Caution is recommended.