Woefully Inadequate Intro to Stats for HCI

Griffin Dietz
CS 197 HCI Section

Adapted with permission from slides by Michael Bernstein and Tobi Gerstenberg

But first…administrivia

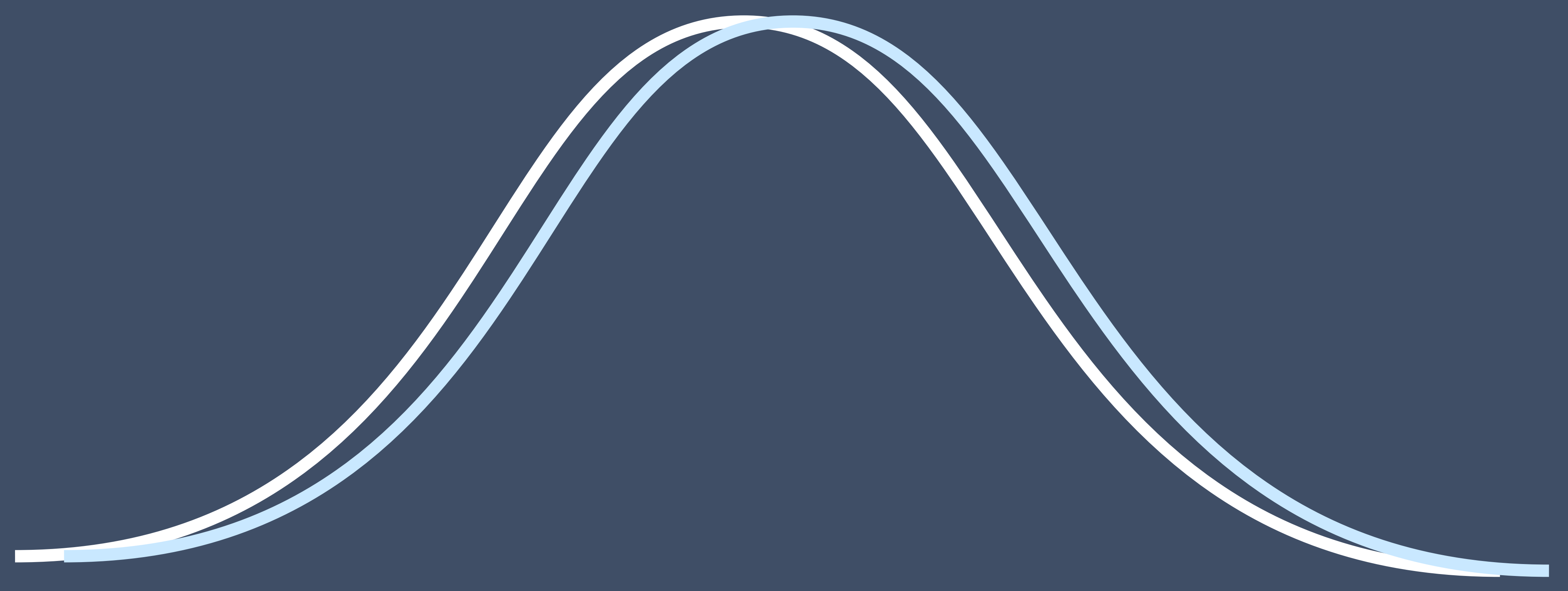
Feedback == more guidance needed —> “ambiguity challenge” and making the best use of office hours/section

Link to materials in project reports

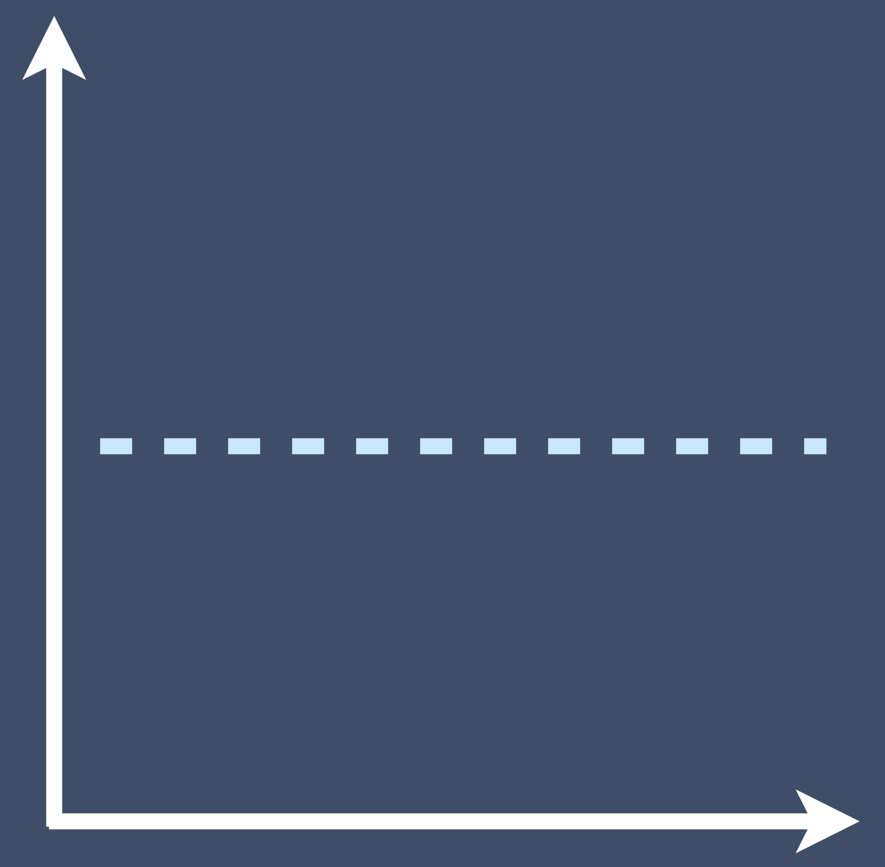
Evaluation assignment early release

Null Hypothesis

If your change/intervention had no effect what would the world look like?



No difference in means

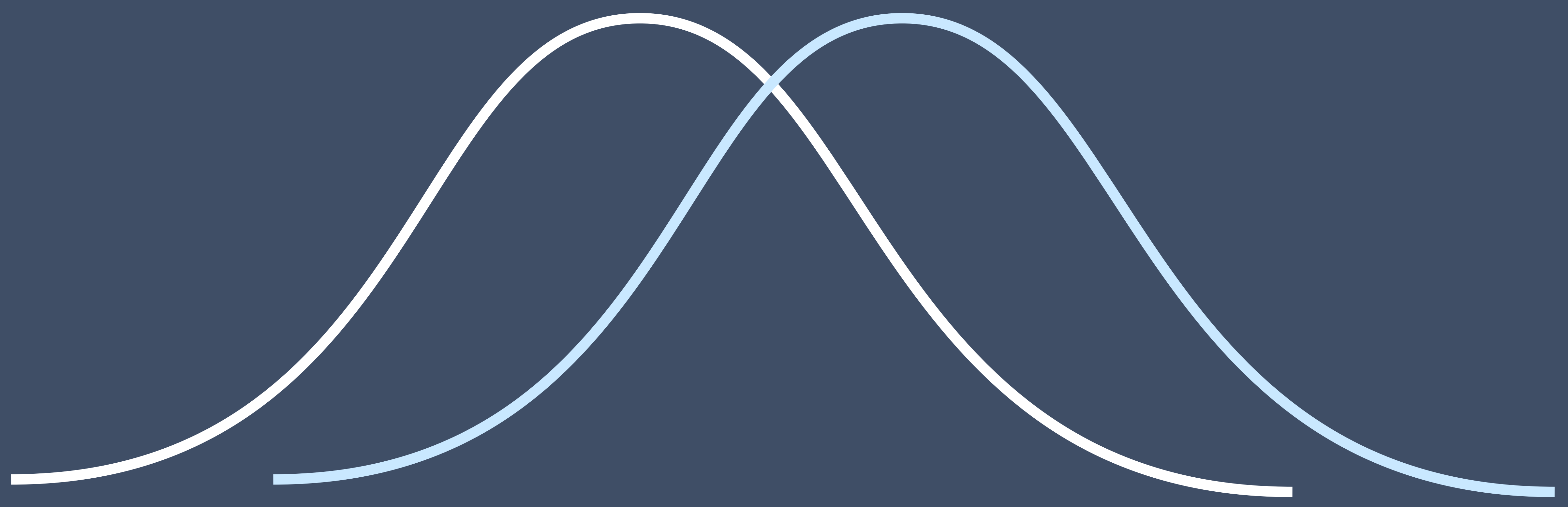


No slope in relationship

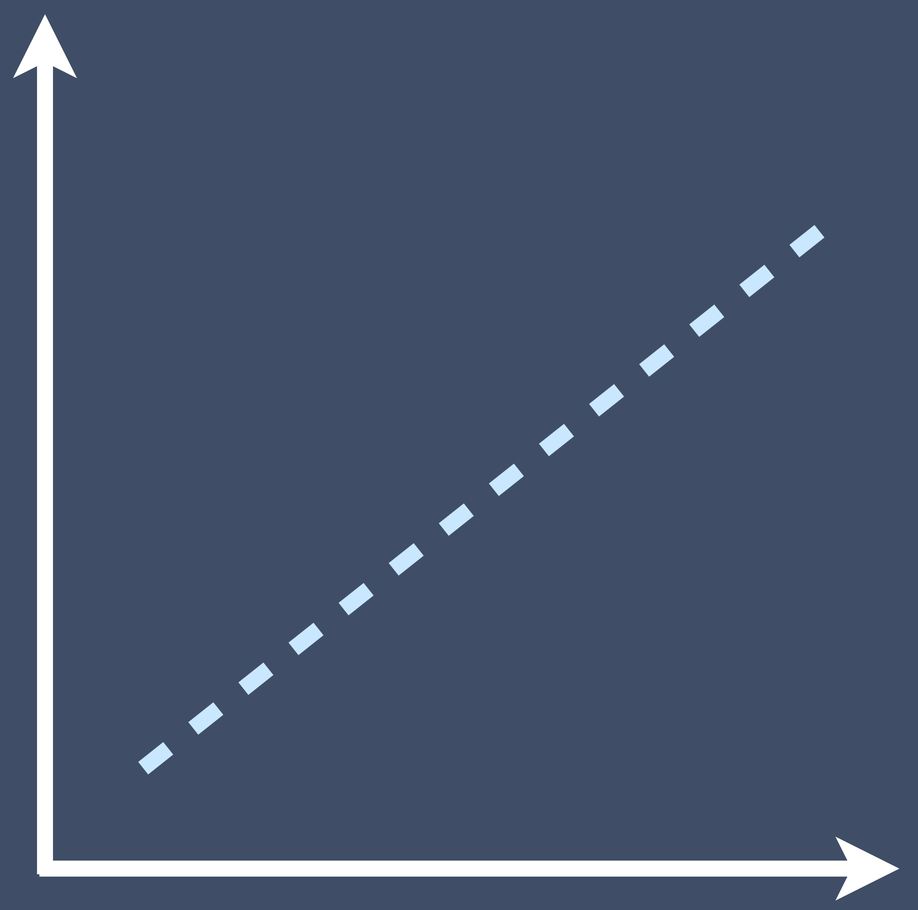
This is called the **null hypothesis**.

Null Hypothesis Significance Testing

Given the data you collected/difference you observed, how likely is it to have occurred by chance?



Probability of seeing a mean difference at least this large, by chance

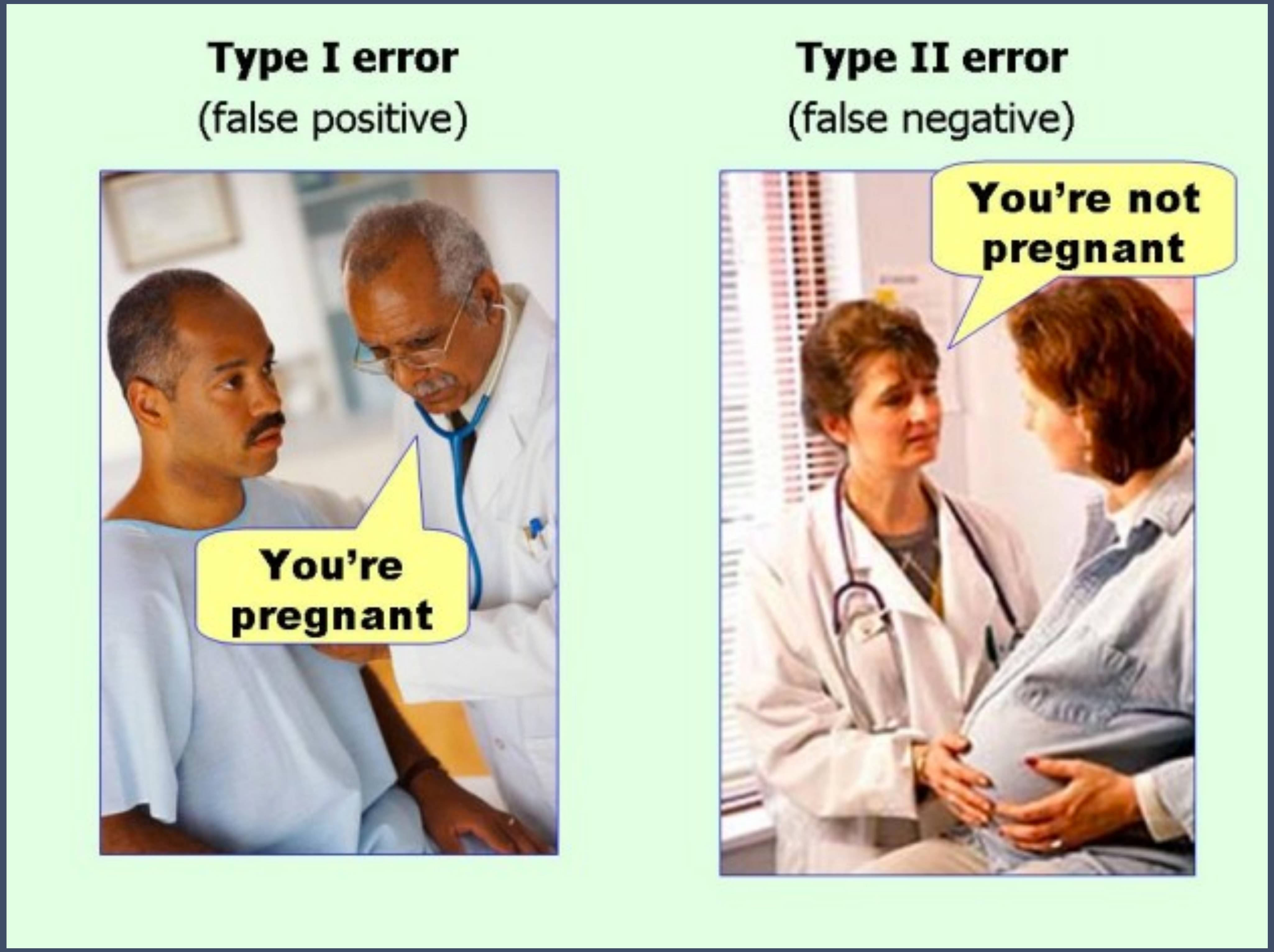


Probability of seeing a slope at least this large, by chance

Enter, *p*-values

P-value is the probability of seeing the observed data by chance (or, the probability of a Type I error)

Generally, $p < .05$ is accepted as "statistically significant" support for a condition difference



Types of Data

Continuous (e.g., duration)

Interval (e.g., exam scores)

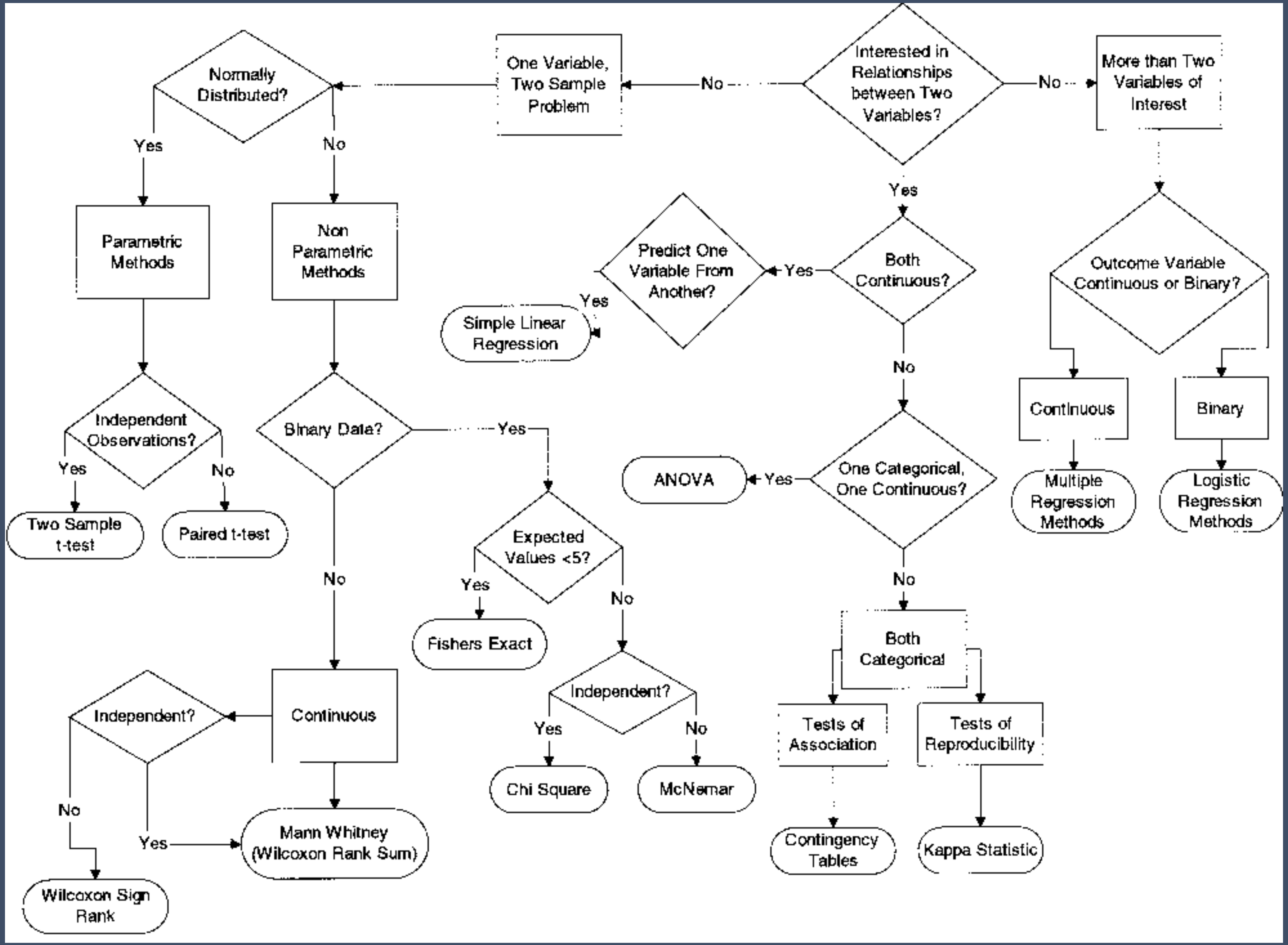
Ordinal (e.g., Likert scales)

Binary (e.g., success/failure)

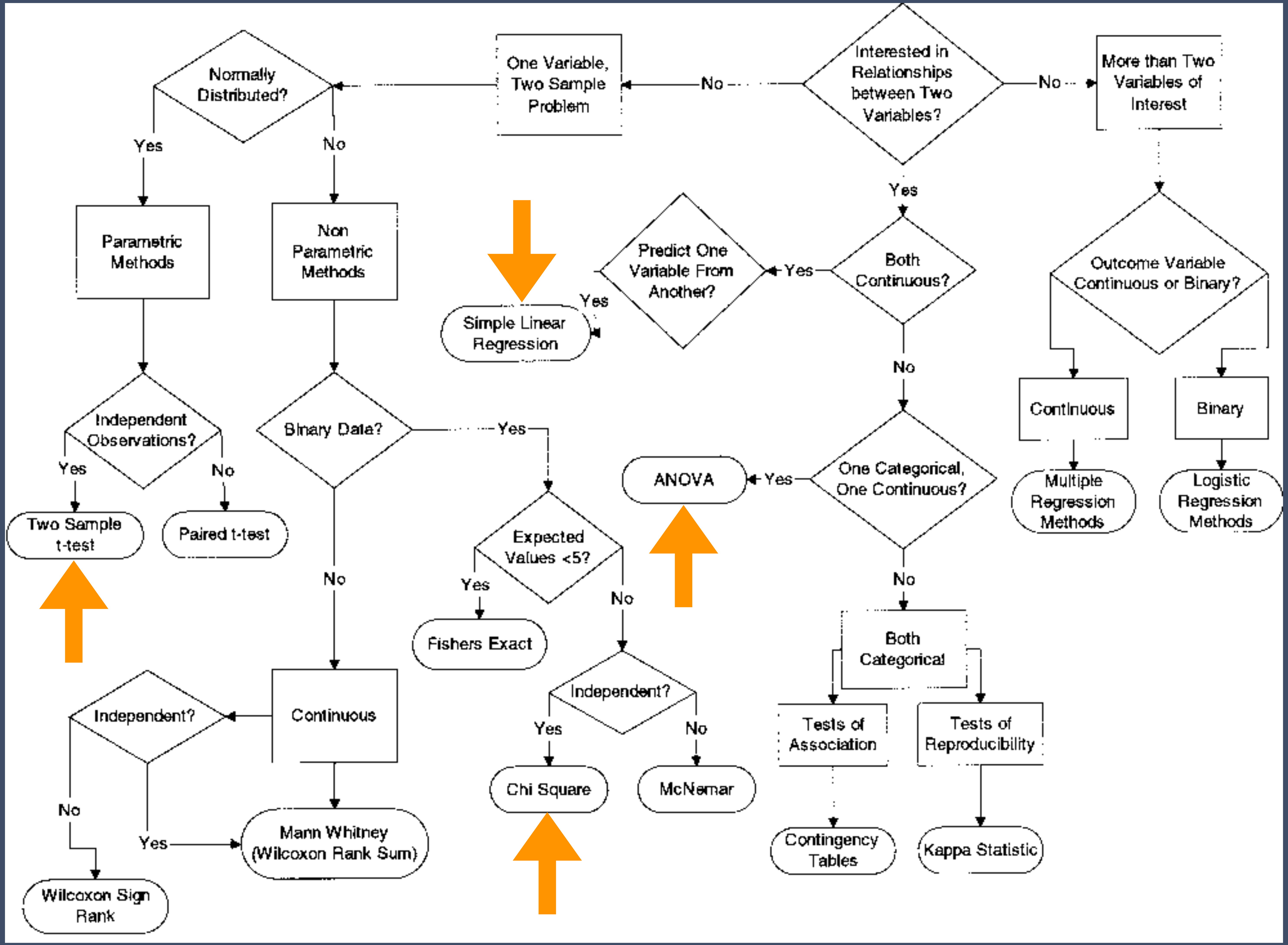
Categorical (e.g., ethnicity)

Type of data will change which statistical tests are appropriate.

A non-ideal method



A non-ideal method



Pearson’s Chi-Square

For Comparing Two Population Counts
(Binary Data)

Calculate Chi-Square

“Five people completed the trial with the control interface, and twenty two completed it with the augmented interface.”

	control	augmented
success	5	22
failure	35	18

Calculate Chi-Square

Determine the expected number of outcomes for each cell

	control	augmented	total
success	5	22	27
failure	35	18	53
total	40	40	80

Expected is (row total)*(column total) / overall total.

Upper left: expected is 27*40/80 = 13.5

Calculate Chi-Square

Expected values = (row total)*(column total) / overall total:

	control	augmented	total
success	13.5	13.5	27
failure	26.5	26.5	53
total	40	40	80

Calculate Chi-Square

Calculate a chi square statistics for each cell and sum over all cells

$$\chi^2 = \frac{(observed - expected)^2}{expected}$$

	control	augmented
success	5.35	5.35
failure	2.73	2.73

→ 5.35 + 5.35 + 2.73 + 2.73 = **16.16**

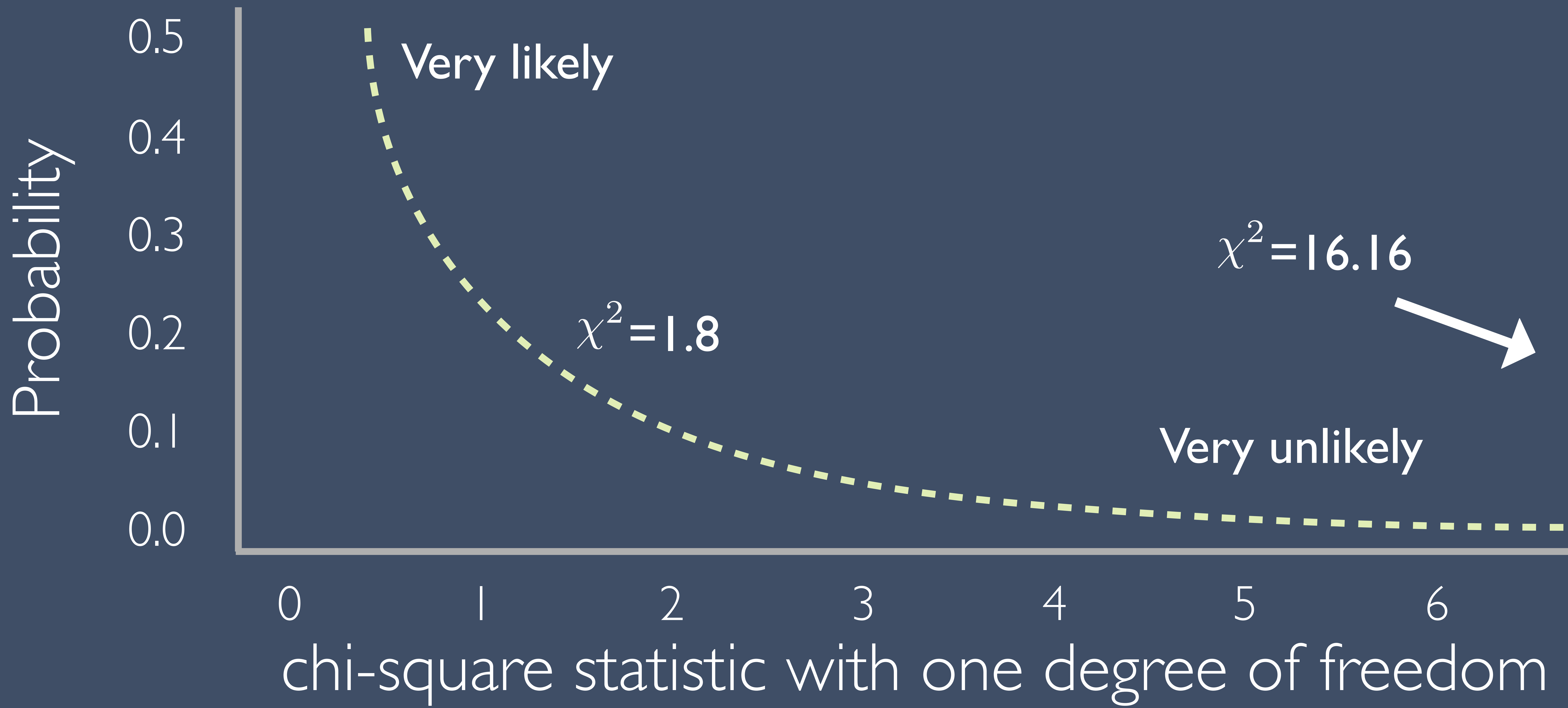
Calculate Degrees of Freedom

- If we know there are a total of 40 participants…

5	???
???	18

- We get (rows - 1) * (columns -1) degrees of freedom. So, if it's a two-by-two design, one degree of freedom.

Result: Chi-Square Distribution



Pearson's Chi-Square in R

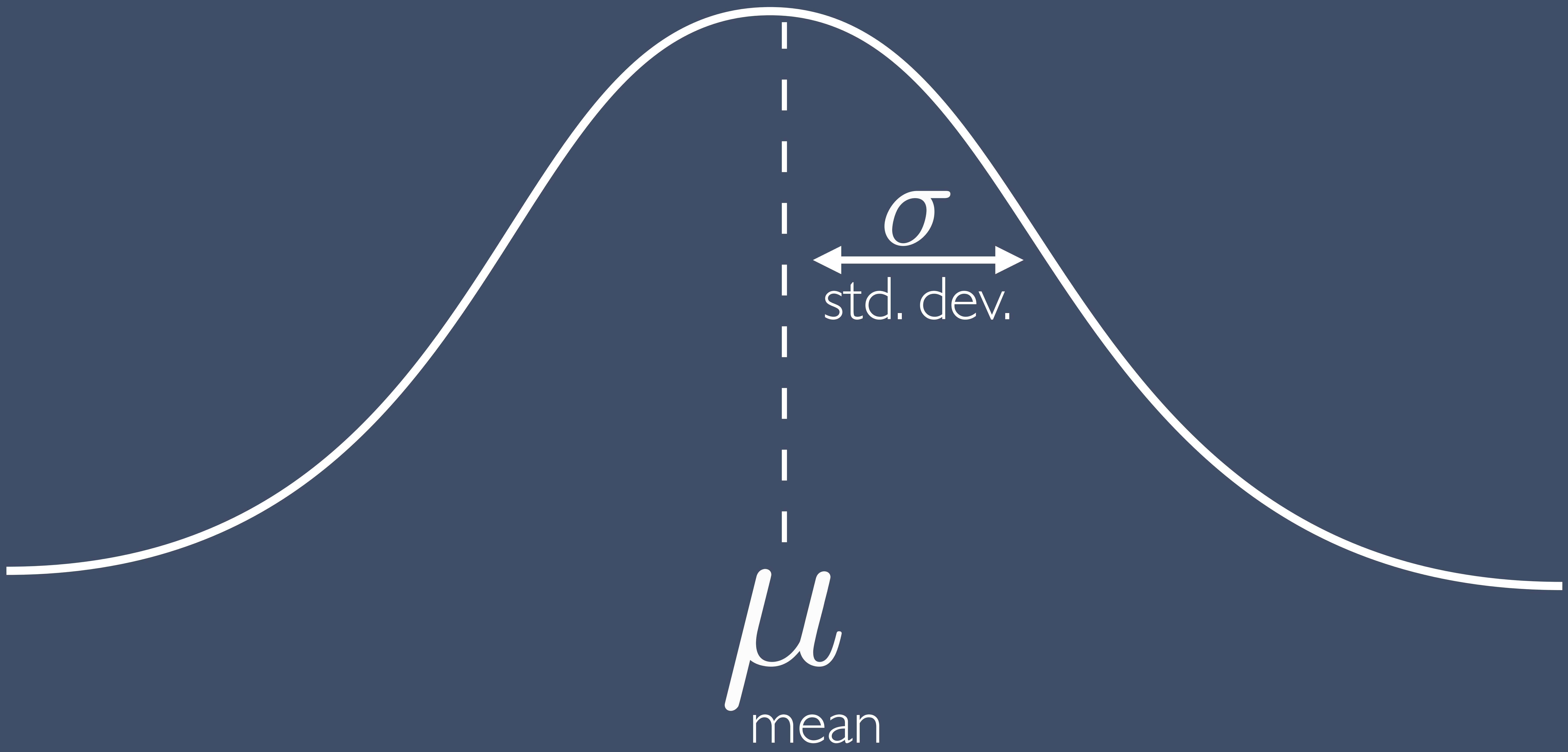
chisq.test (HCI R tutorial at http://yatani.jp/HCIstats/ChiSquare)

```
> data
     [,1] [,2]
[1,]    5   22
[2,]   35   18
> chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  data
X-squared = 14.3117, df = 1, p-value = 0.0001549
```

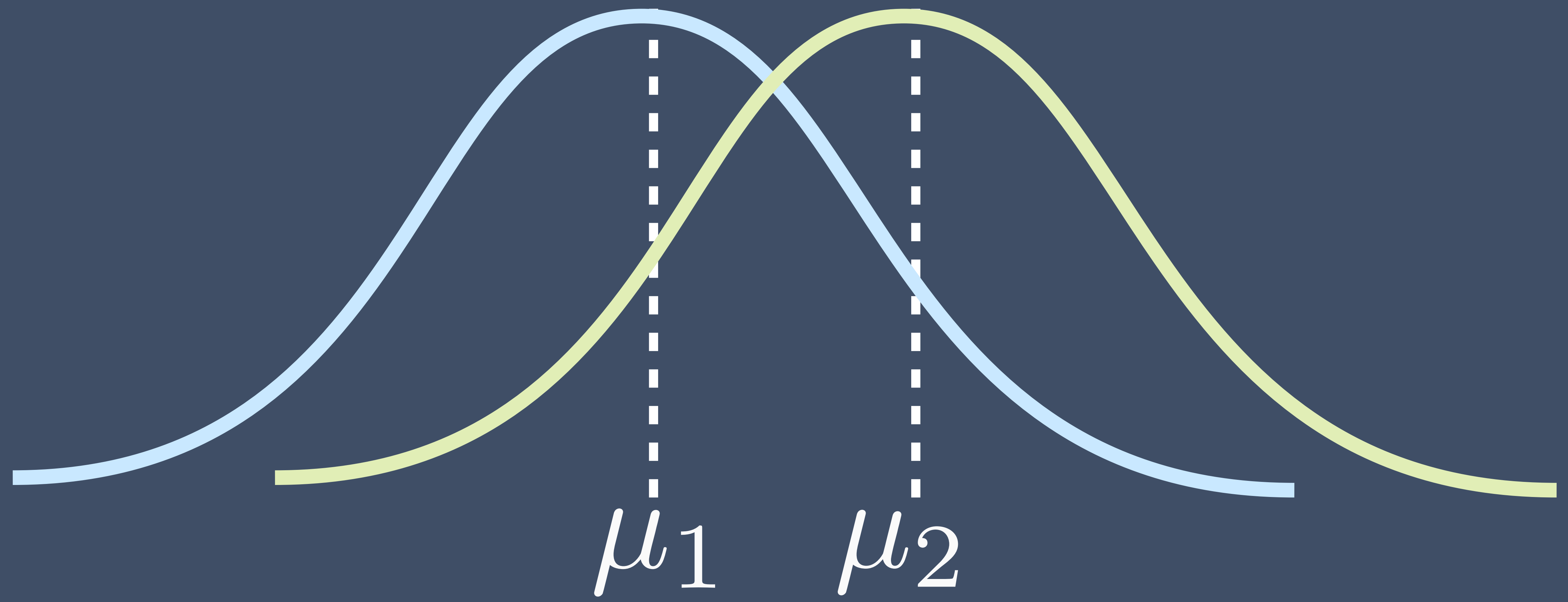
T-Test

For Comparing Two Population Means
(Continuous, Normally Distributed Data)

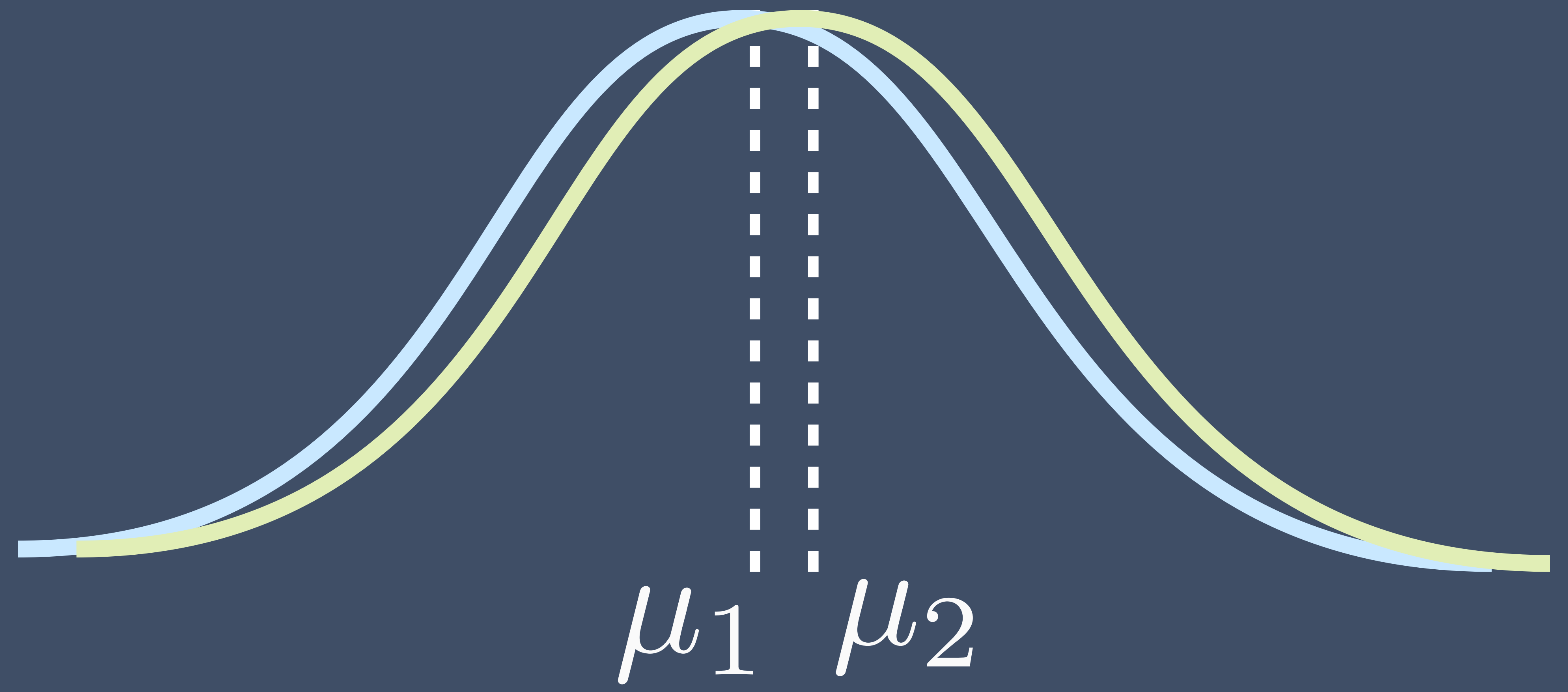
Normally Distributed Data



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



likely have different means



likely have the same mean
(null hypothesis)

Calculate the t-statistic

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Numbers that matter:

- Difference in means
  larger means more significant
- Variance in each group
  larger means less significant
- Number of samples
  larger means more significant

Calculate Degrees of Freedom

If we know the mean of N numbers, then only N-1 of those numbers can change.

Example: pick three numbers with a mean of ten (e.g., 8, 10, 12). Once you've picked the first two, the third is set.

We have two means, so a t-test has N-2 degrees of freedom.

Result: t-distribution

T-test in R

t.test (HCI R tutorial at http://yatani.jp/HCIstats/TTest)

```
> data
       group result
1    control      1
2    control      1
3    control      2
4    control      3
5    control      1
6    control      3
7    control      2
8    control      4
9    control      1
10   control      2
11 augmented      6
12 augmented      5
13 augmented      1
14 augmented      3
```

```
> t.test(data[data["group"] == "control", 2], data[data["group"]
 == "augmented", 2], var.equal=T)

        Two Sample t-test

data:  data[data["group"] == "control", 2] and data[data["group"]
] == "augmented", 2]
t = -2.2014, df = 18, p-value = 0.04099
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 -2.73610126 -0.06389874
sample estimates:
mean of x mean of y
      2.0       3.4
```

Paired t-test for within-subjects design

It can be easier to statistically detect a difference if the participants try both alternatives. Why?
A paired test controls for individual-level differences.

$$t = \frac{\mu - 0}{\sqrt{\frac{\sigma^2}{N}}}$$

Is the mean of that difference significantly different from zero?

Paired t-test in R

```
> t.test(data[data["group"] == "control", 2], data[data["group"]
 == "augmented", 2], paired=T)

	Paired t-test

data:  data[data["group"] == "control", 2] and data[data["group"
] == "augmented", 2]
t = -1.7685, df = 9, p-value = 0.1108
alternative hypothesis: true difference in means is not equal to
 0
95 percent confidence interval:
 -3.1907752  0.3907752
sample estimates:
mean of the differences
                   -1.4
```

Why no longer significant? (Hint: look at the degrees of freedom "df")

Ten participants. If we had twenty participants like before, much more likely.

ANOVA

For Comparing N>2 Population Means
(Continuous, Normally Distributed Data)

ANOVA: ANalysis Of VAriance

Use instead of a t-test when you have > 2 factor levels/conditions and a continuous DV

Example: the effect of phone vs. tablet vs. laptop on number of searches successfully performed

Very nice property: an ANOVA is just a regression with one predictor under the hood!

Linear Regression

For Comparing N>2 Population Means
(Continuous, Normally Distributed Data)

Linear Regression

Data = Model + Error

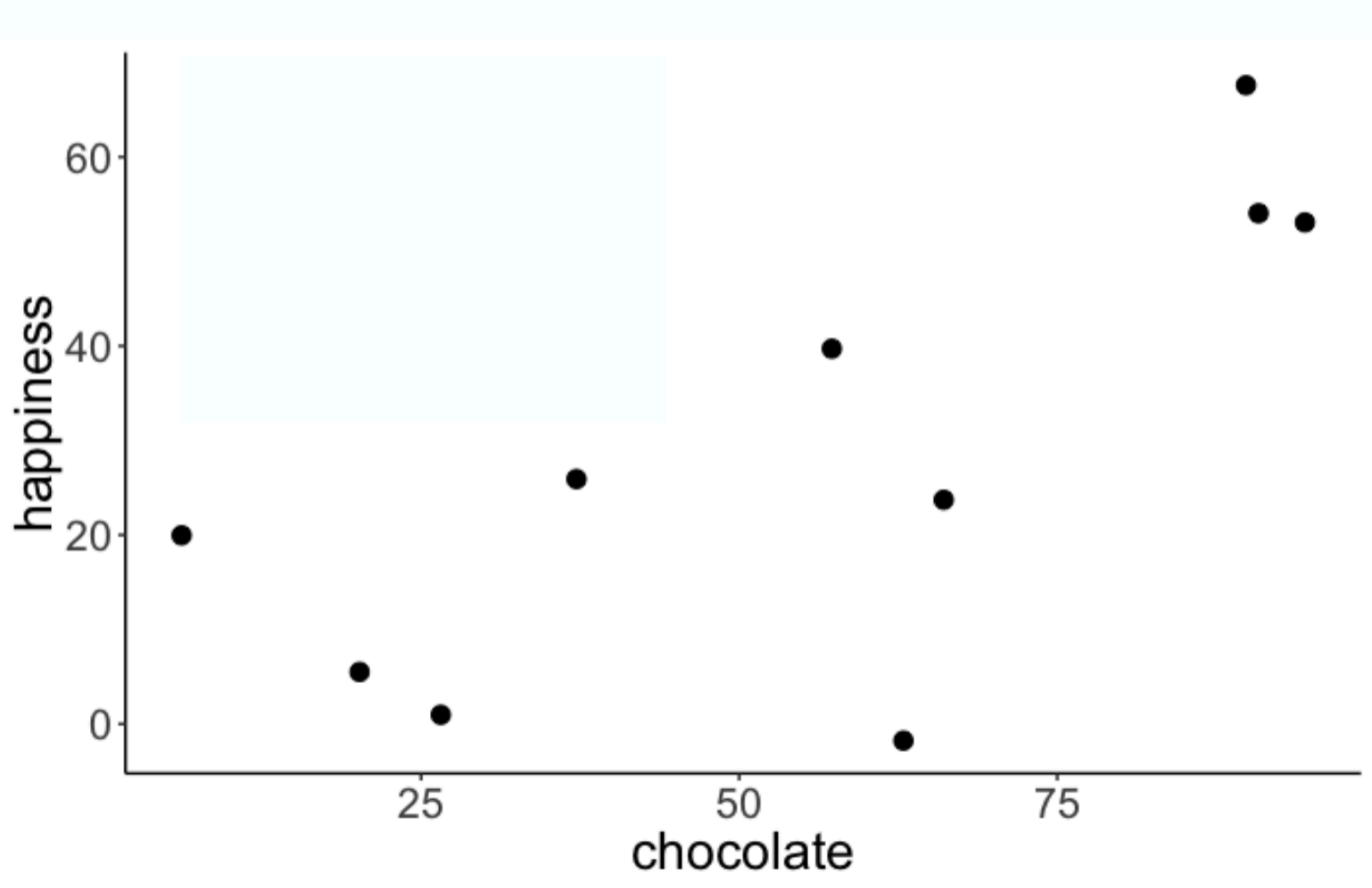
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_0$$

$$\downarrow$$

$$Y_i = \beta_0 + \beta_1 X_i$$

Model is a linear combination of predictors that minimizes error

Is there a relationship between chocolate and happiness?



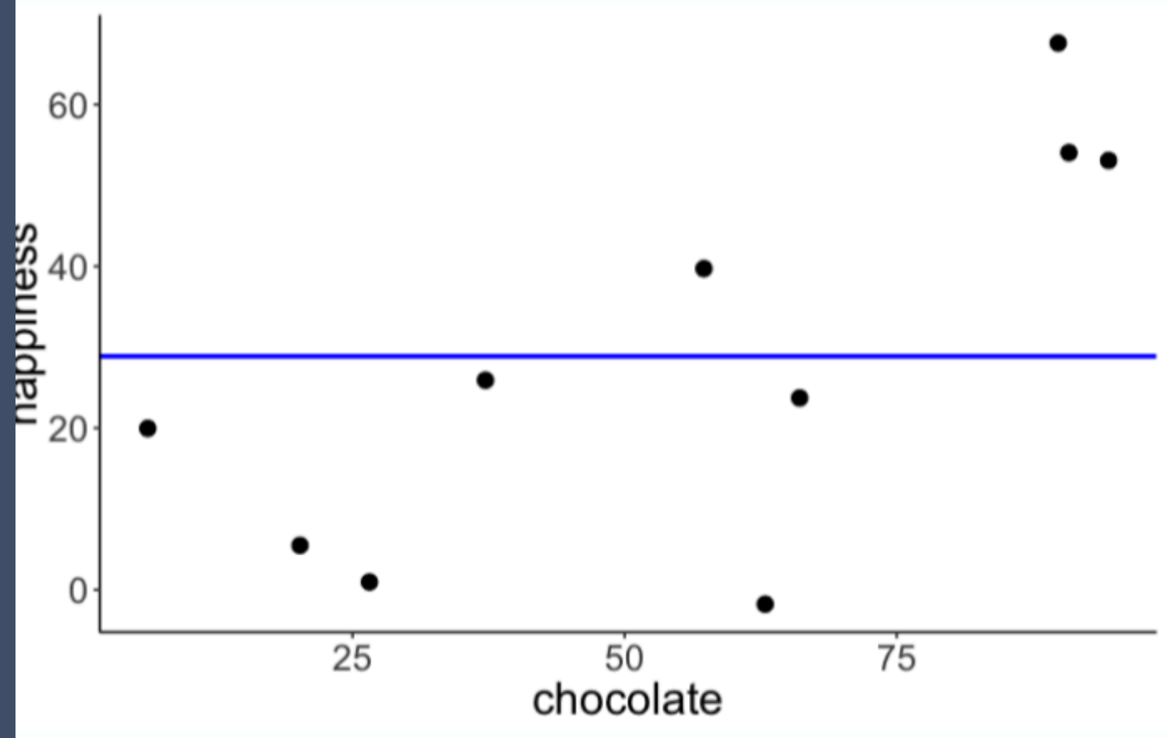
Create a model with chocolate as a predictor

H_0: Chocolate consumption and happiness are unrelated.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 28.88 + e_i$$

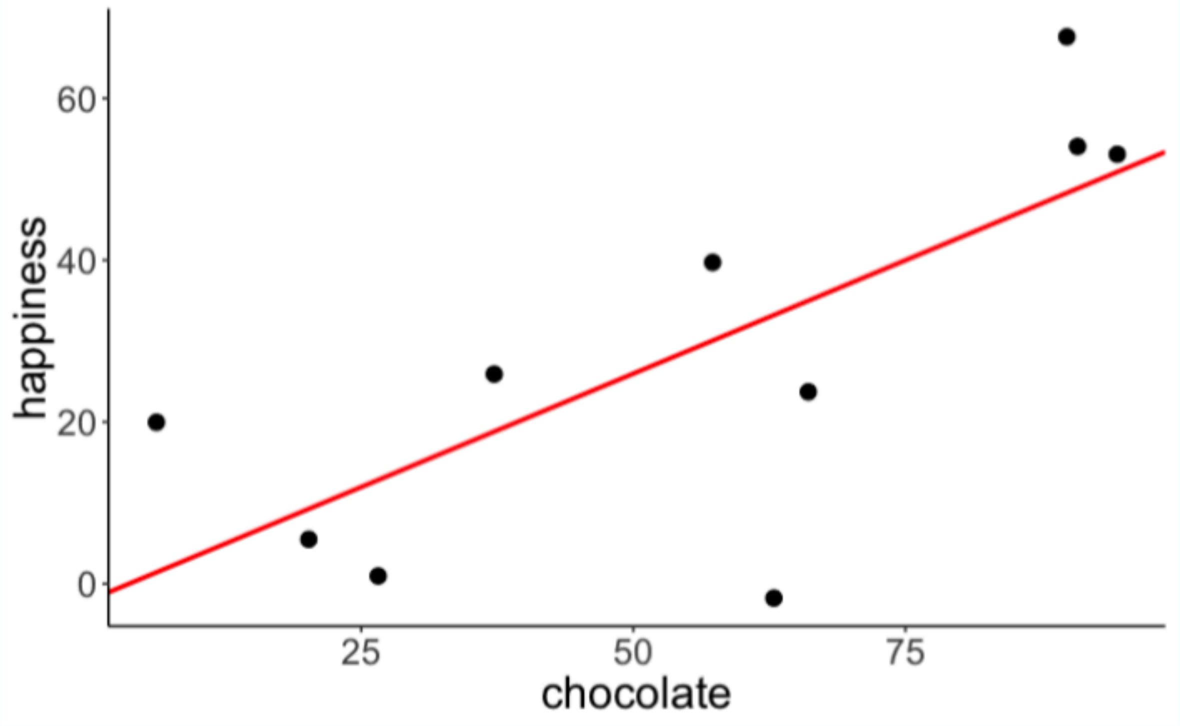
H_1: Chocolate consumption and happiness are related.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

chocolate consumption

Model prediction

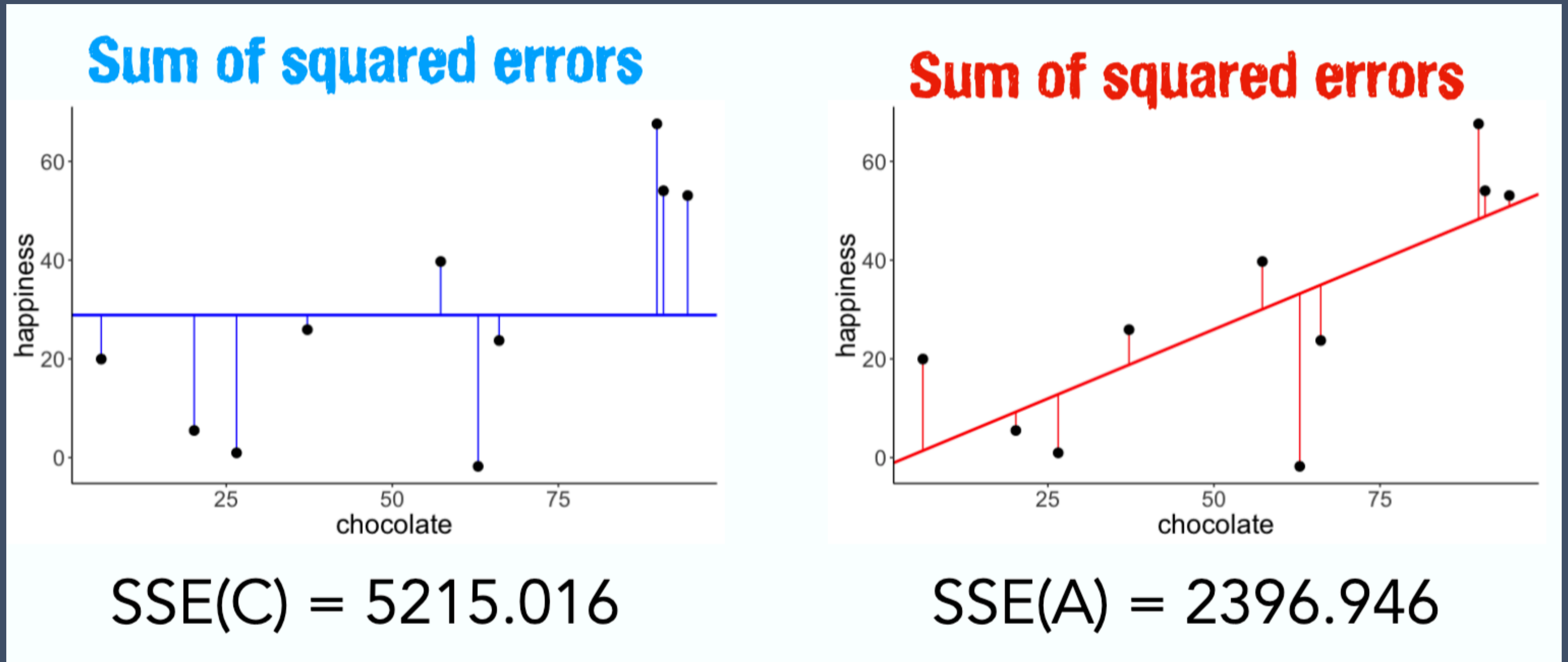


Fitted model

$$Y_i = -2.04 + 0.56 X_i + e_i$$

Is the model a better fit

Or, does the model decrease error?



Proportional Reduction in Error (PRE) = $1 - \dfrac{SSE(A)}{SSE(C)} = 1 - \dfrac{2396.946}{5215.016} \approx 0.54$

Model with chocolate as a predictor decreases error by about 54%.

Compute an F statistic

$$F = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} = \frac{0.54/(2 - 1)}{(1 - 0.54)/(10 - 2)} = 9.4$$

PRE = Proportional reduction in error

PA = number of parameters in Model C (PC) and Model A (PA)

n = number of observations

Result: F-distribution

Linear model in R

t.test (HCI R tutorial at http://yatani.jp/HCIstats/TTest)

```
> model <- lm(happiness ~ chocolate, data = df.regression)
> summary(model)

Call:
lm(formula = happiness ~ chocolate, data = df.regression)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.990  -9.400   3.671   9.009  19.276 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -2.0419    11.4713  -0.178   0.8631  
chocolate     0.5606     0.1828   3.067   0.0154 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.31 on 8 degrees of freedom
Multiple R-squared:  0.5404,	Adjusted R-squared:  0.4829 
F-statistic: 9.406 on 1 and 8 DF,  p-value: 0.01542
```

Impact of chocolate in model

When chocolate goes up one, happiness goes up .56 (p = .015)

Overall model fit