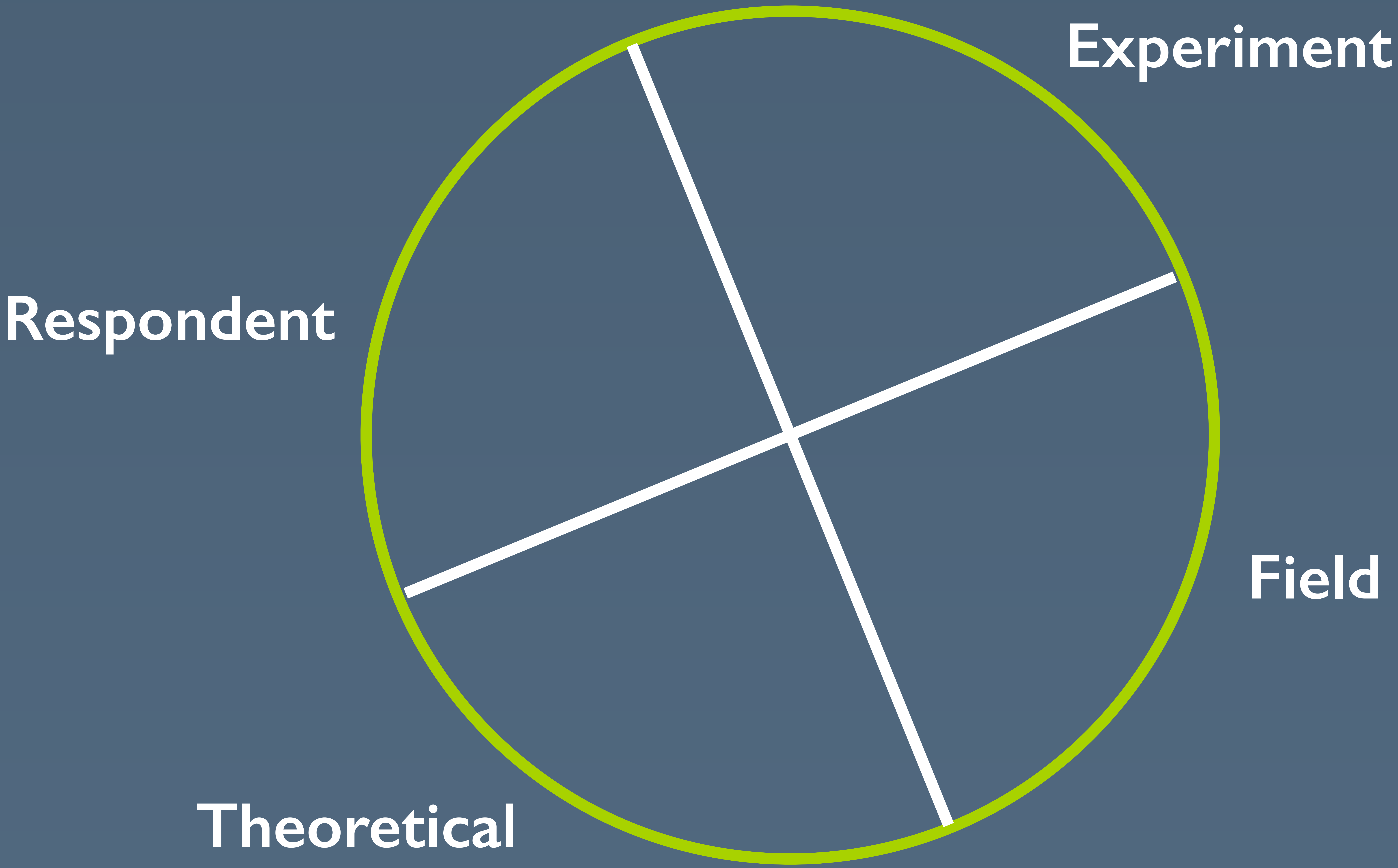
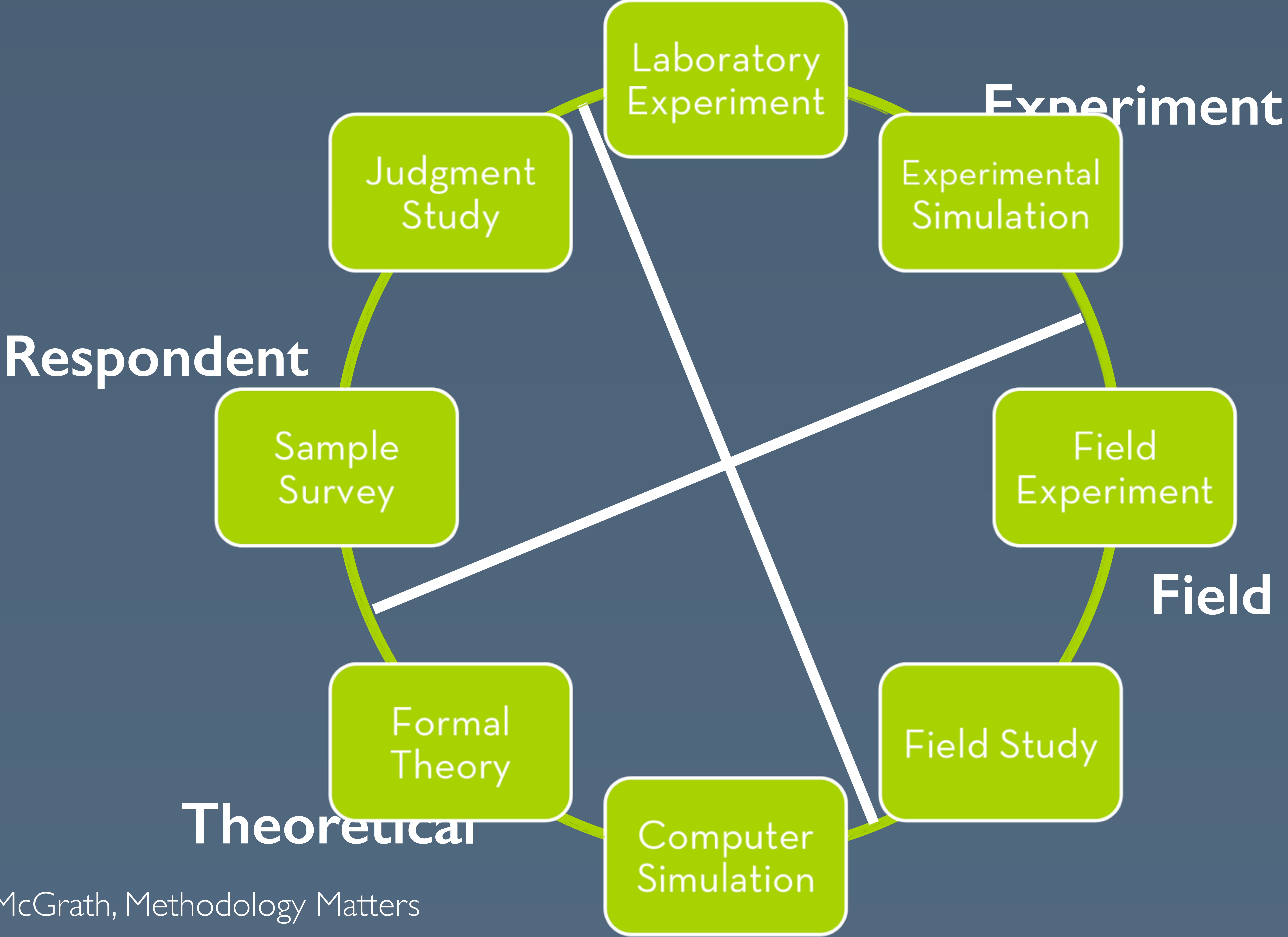
Research Methods I

MICHAEL BERNSTEIN
CS 376



From McGrath, Methodology Matters



From McGrath, Methodology Matters

Method triangulation

- All methods are flawed
- Thus, your argument becomes far stronger if you can demonstrate the same phenomenon using multiple methods
  - Complement your statistics with semi-structured interviews
  - Complement qualitative work with primary source evidence or log data

Objectivity in reporting

- Readers are more cynical if that paper is presenting a one-sided argument
- Which argument do you buy?
  - "Ellipsoidal windows were better for all tasks."
    vs.
    "Ellipsoidal windows were better for all tasks we measured. However, users found them to be confusing."

Framing an evaluation

- The difficulty: defining and isolating the construct that you are trying to maximize
- It is tempting to aim for something easy: time, task completion, number of clicks
- But, testing the easily quantifiable often misses the point.

Framing an evaluation

- Reflect on your implicit thesis about why your contribution is a good idea.
  - InForm is a good idea because…
  - Designing in parallel is a good idea because…
  - Soylent is a good idea because…
- This thesis can directly imply the claim that you need to test. (It may or may not be comparative in nature.)

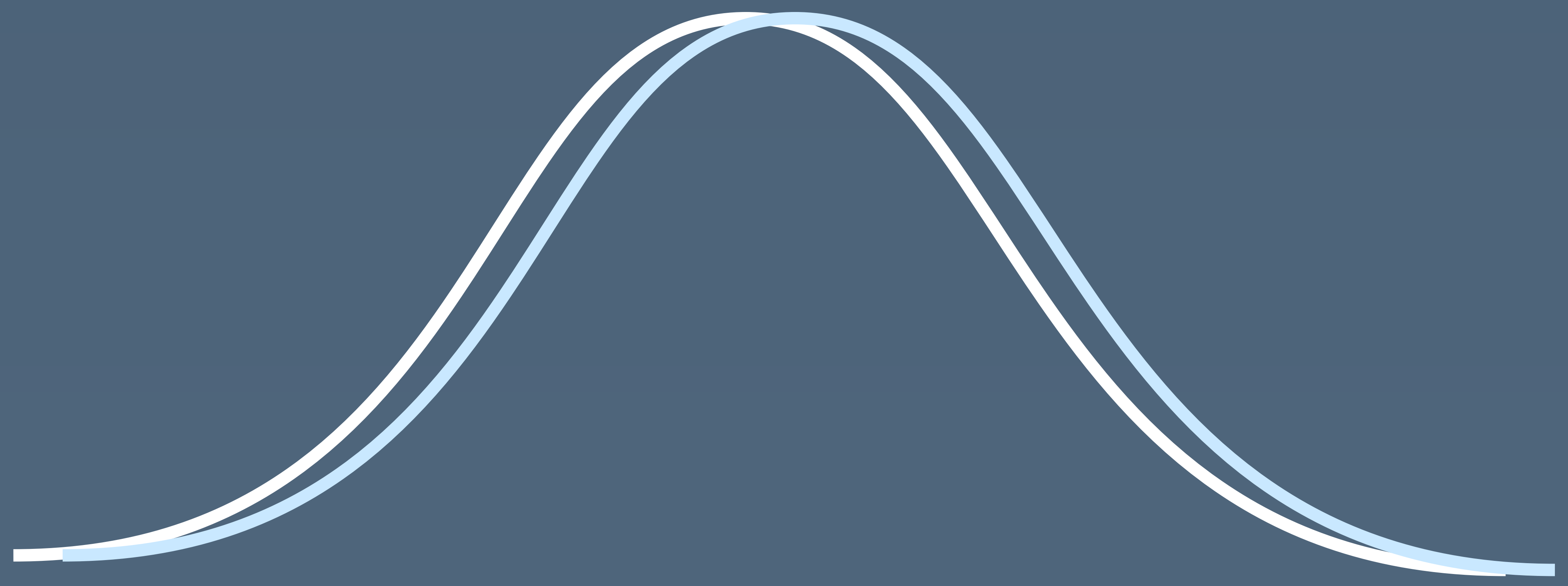
Example theses

- Enable previously difficult/impossible tasks
- Improve task performance or outcome
- Modify/influence behavior
- Improve ease-of-use, user satisfaction
- User experience

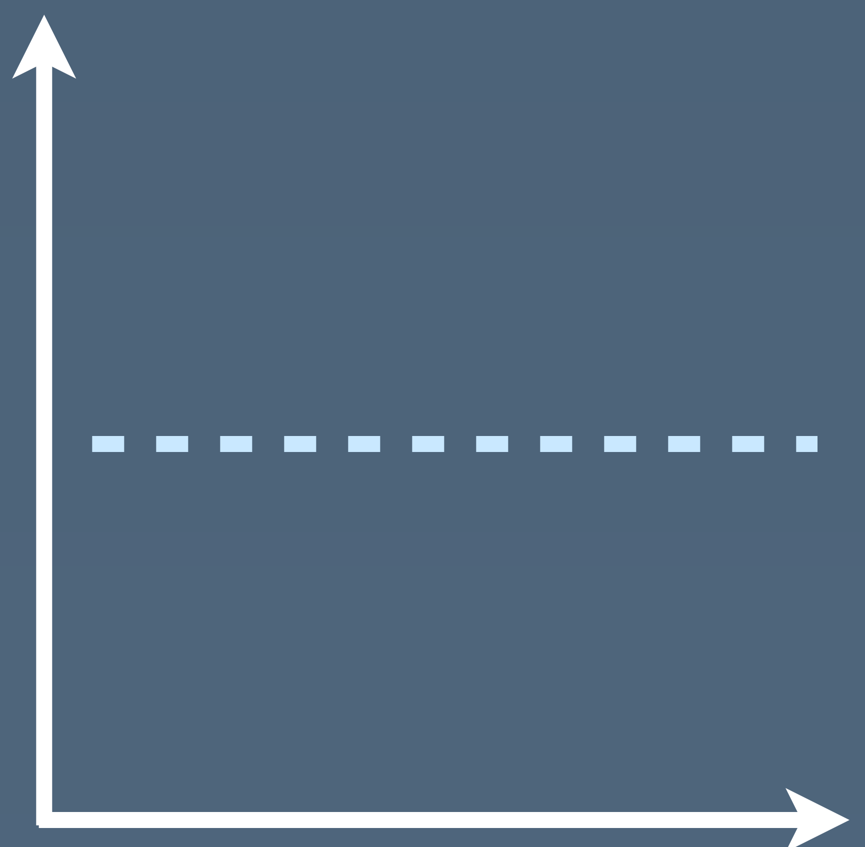
Hypothesis Testing

Anatomy of a statistical test

- If your change had no effect, what would the world look like?



No difference in means

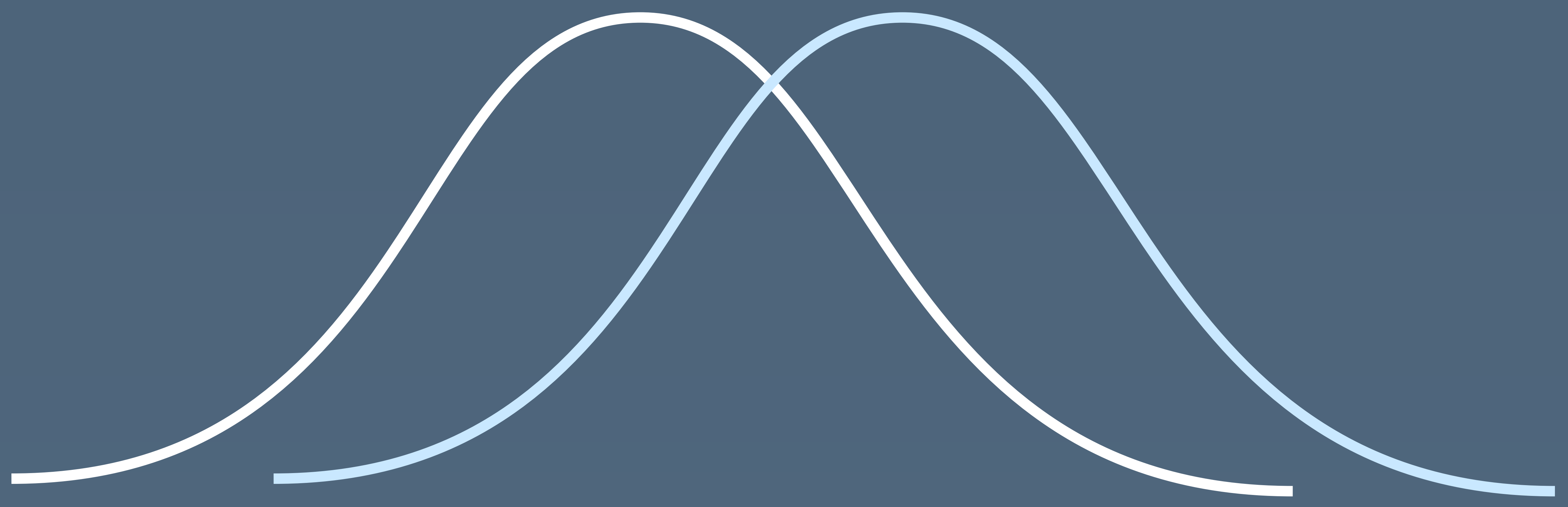


No slope in relationship

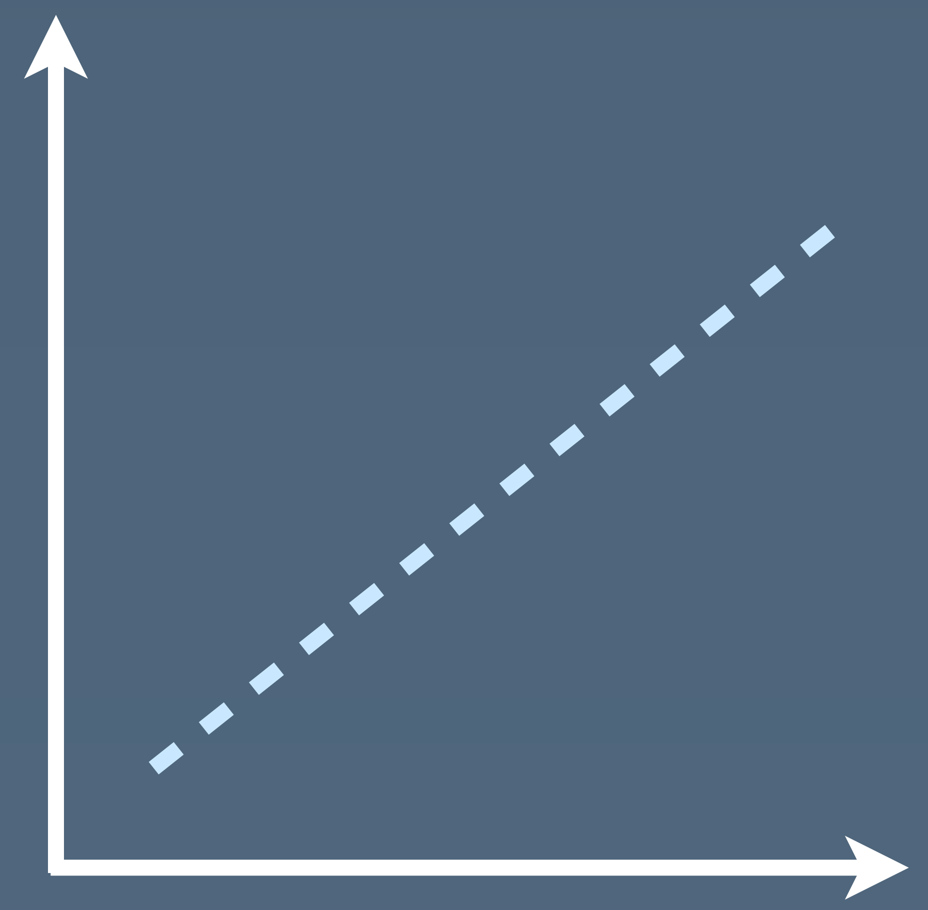
- This is known as the **null hypothesis**

Anatomy of a statistical test

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



Probability of seeing a mean difference at least this large, by chance, is 0.012

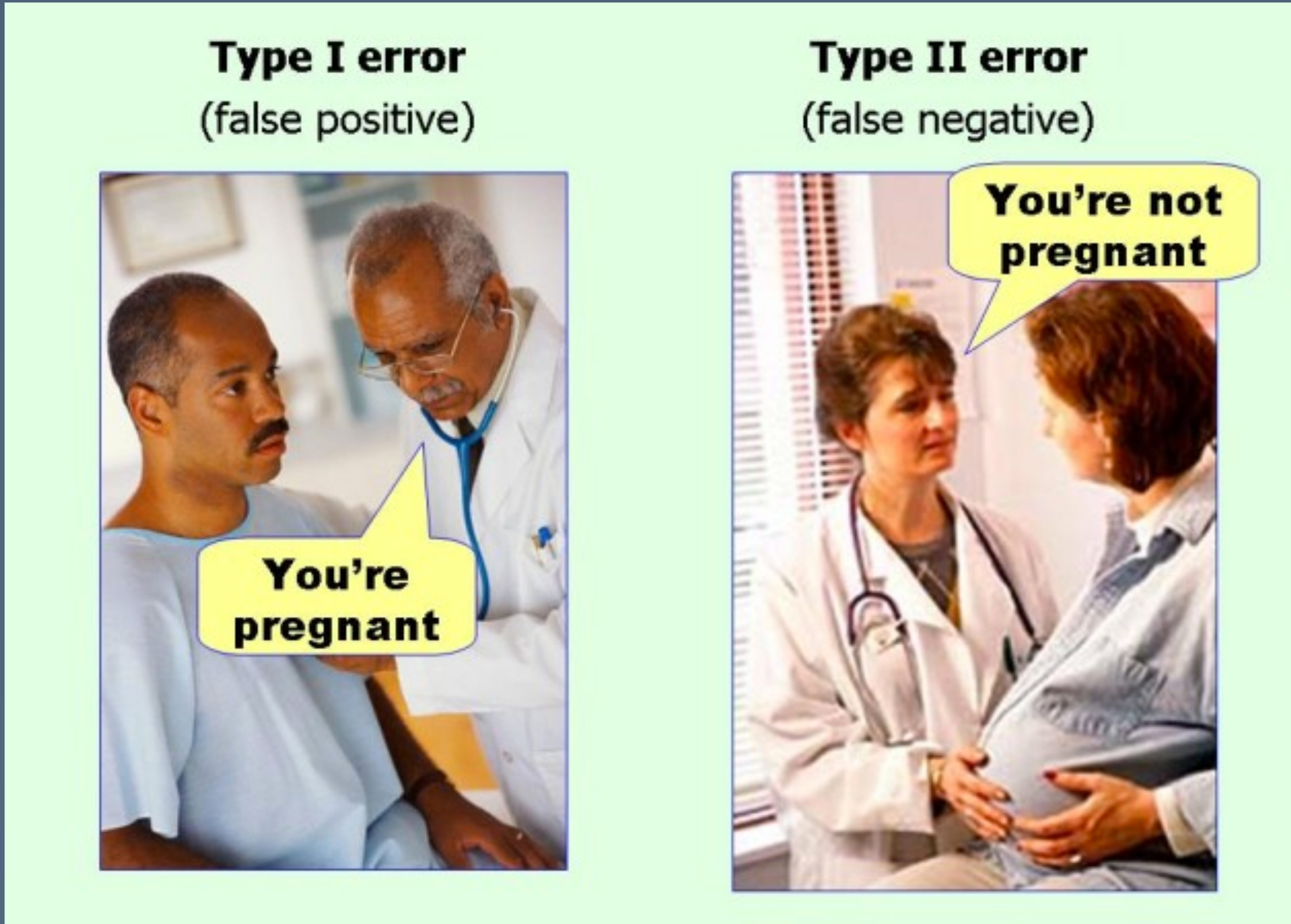


Probability of seeing a slope at least this large, by chance, is 0.012

Errors

Difference detected? / Difference exists?	Y	N
Y	True positive	Type 1 error publish false findings
N	Type 2 error get more data?	True negative

Errors


Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

p-value

- The probability of seeing the observed difference by chance
  - In other words, P(Type I error)
- Typically accepted levels: 0.05, 0.01, 0.001

Comparing two populations: counts

Count or occurrence data

- “Fifteen 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

Pearson's chi-square test for independence

- 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

Calculating a chi-square statistic

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

e.g., $(5\text{-}13.5)^2$ / 13.5 = 5.35

Sum this value over all possible outcomes

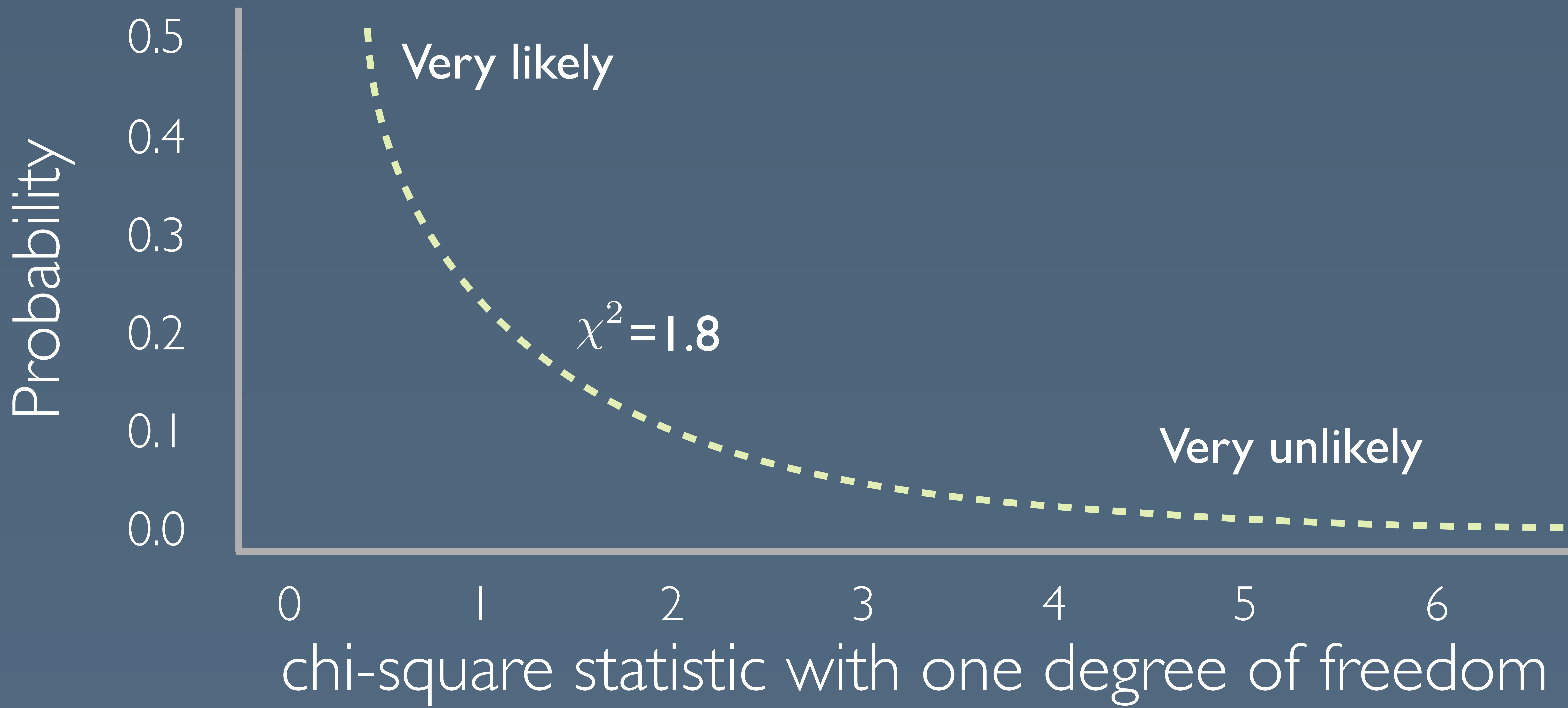
How many degrees of freedom?

- 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



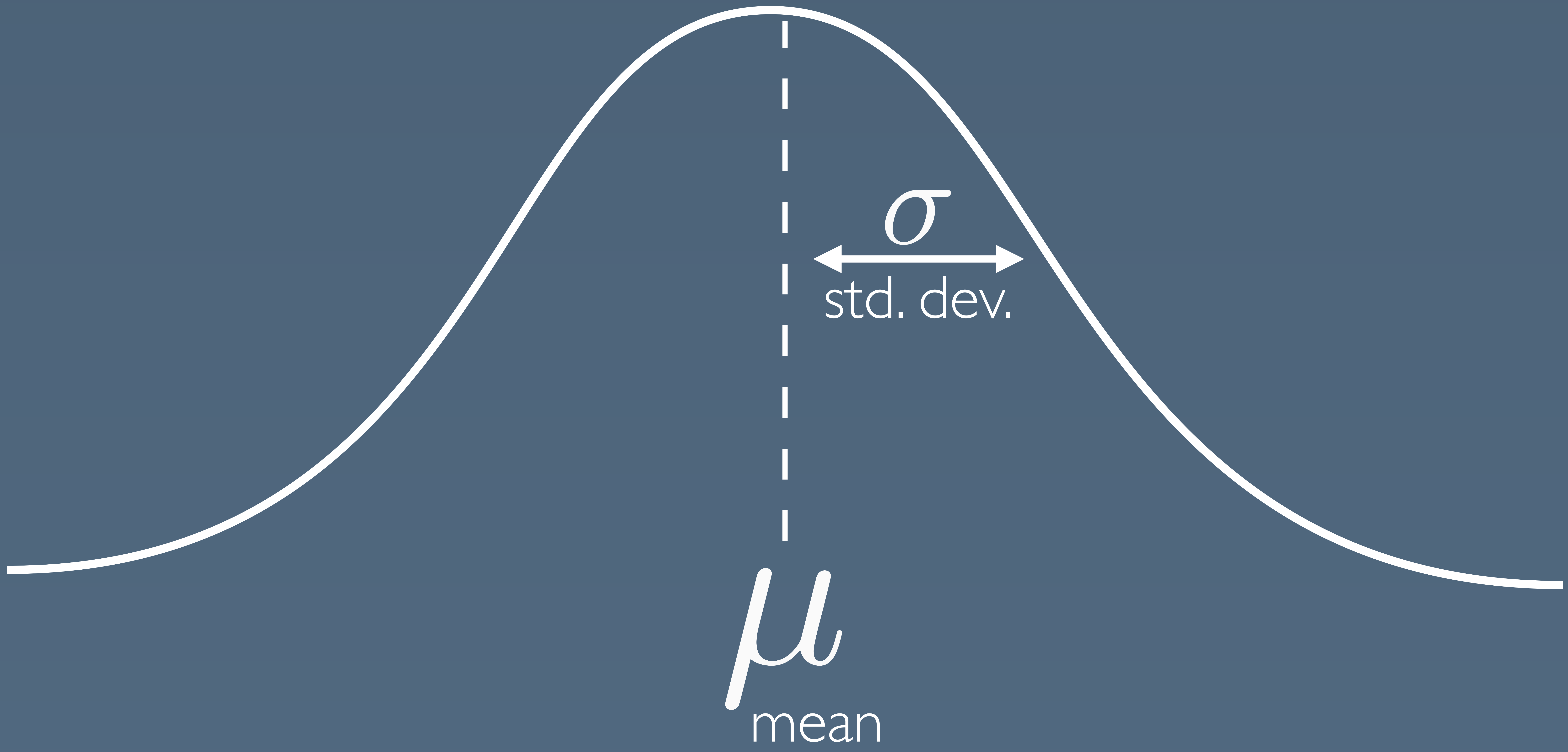
Pearson’s chi-square test for independence

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

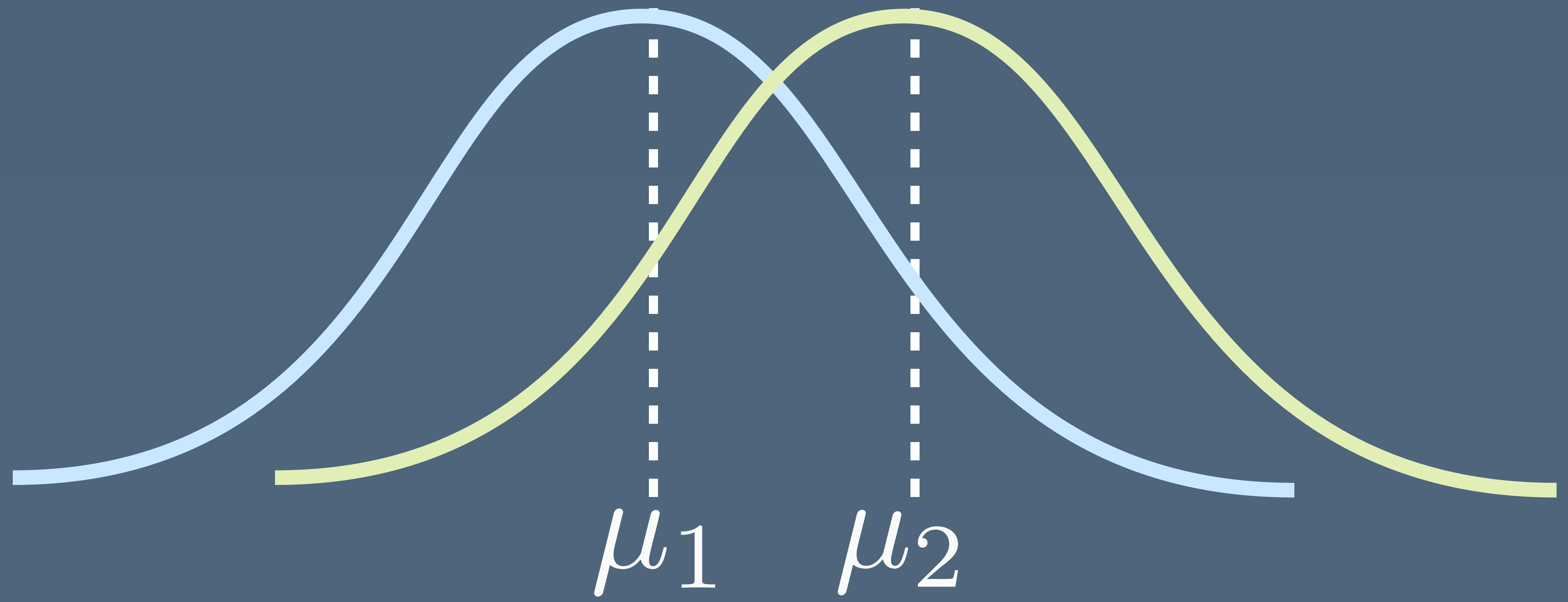
```
> data
     [,1] [,2]
[1,]    5   22
[2,]   35   18
> chisq.test(data)

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

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

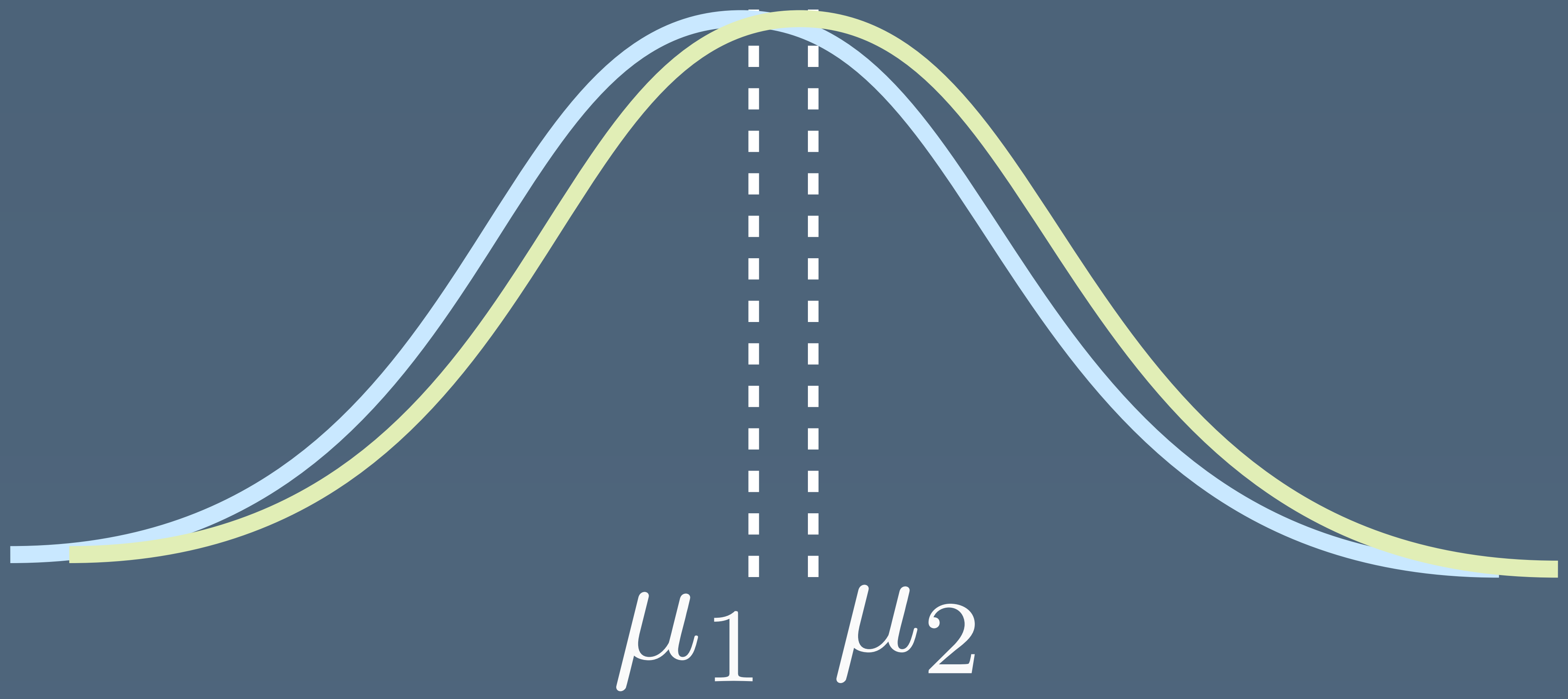
Comparing two populations: means

Normally distributed data



t-test: do they have the same mean?





likely have different means

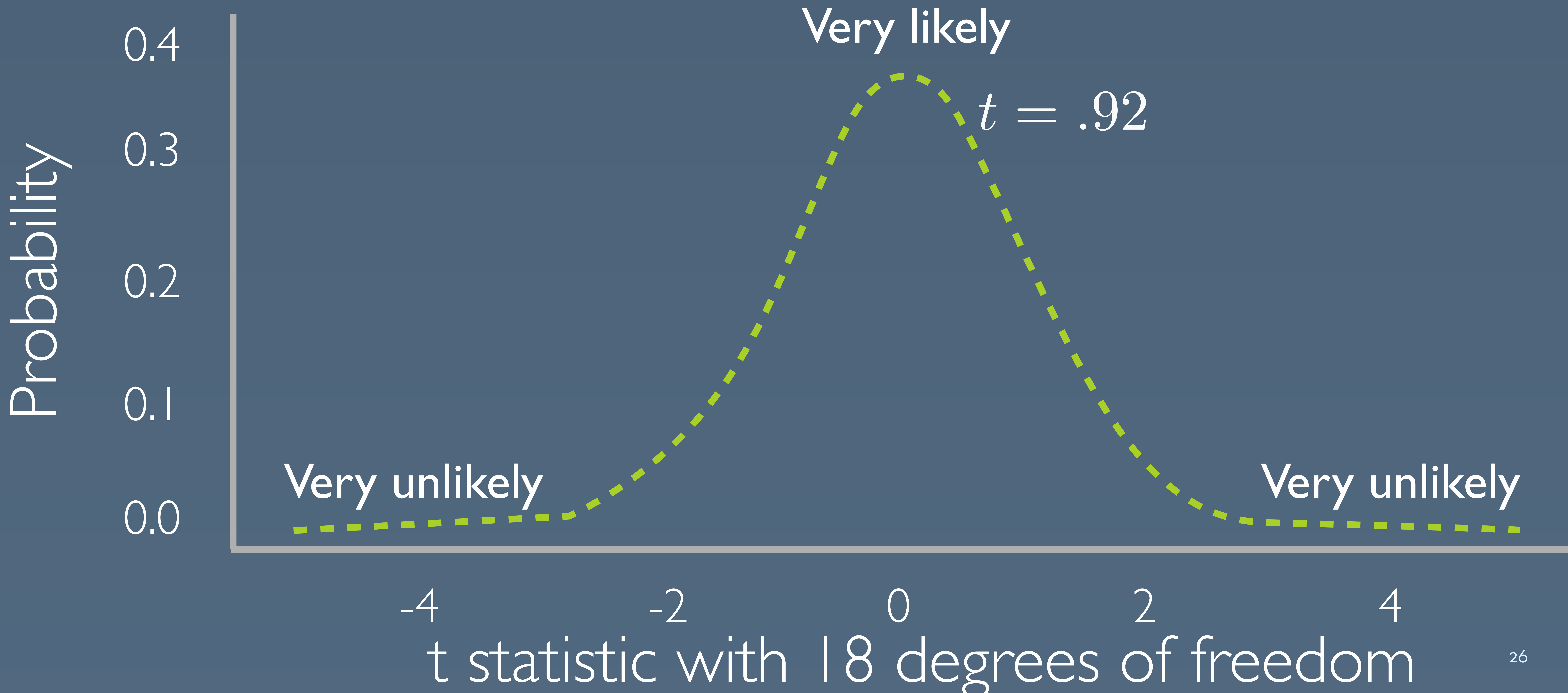
likely have the same mean
(null hypothesis)

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

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

Example t distribution



How many degrees of freedom?

- If we know the mean of N numbers, then only N-1 of those numbers can change.
- We have two means, so a t-test has N-2 degrees of freedom.

Running the test in R

- Use `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
```

Presenting the result

- "A t-test comparing the expert-rated scores of designs with the control (mean=2.0, std. dev=0.5) to the designs with the augmented condition (mean=3.4, std. dev=0.4) is significant ($t(18)=2.2$, $p<.05$)."

Within-subjects study designs

- It can be easier to statistically detect a difference if the participants try both alternatives.
- Why?

Paired t-test

Control	Augmented
1	6
1	5
2	1
3	3
1	5
3	1
2	2
4	3
1	3
2	4

A paired test controls for individual-level differences.

Paired t-test

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

- Is the mean of that difference significantly different from zero?

Running a 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 rows like before, much more likely.

Comparing two populations: nonparametrics

What if the data isn't normally distributed?

- Skewed data
- Timing data
- Rankings or any ordinal data
- Likert scales with too few options (e.g., only 1-3)

Parametric tests assume normally-distributed data.
Nonparametric tests do not.

Transform the data into ranks

Control	Augmented	Control	Augmented
11	64	rank 20	rank 1
13	55	19	3
23	15	12	17
35	34	7	9
17	59	16	2
33	18	10	15
25	21	11	13
43	35	4	7
14	37	18	6
21	43	13	4

Compare ranks

Control	Augmented
20	1
19	3
12	17
7	9
16	2
10	15
11	13
4	7
18	6
13	4

Intuition —

Control: average rank is 13

Augmented: average rank is 7.7

Mann-Whitney U

- Also known as the Wilcoxon rank sum test
  (Tutorial at http://yatani.jp/HCIstats/MannWhitney)

```
> wilcox.test(data[data["group"] == "control", 2], data[data["gr
oup"] == "augmented", 2])

        Wilcoxon rank sum test with continuity correction

data:  data[data["group"] == "control", 2] and data[data["group"
] == "augmented", 2]
W = 23.5, p-value = 0.04911
alternative hypothesis: true location shift is not equal to 0
```

- Also available: Wilcoxon signed rank test (for paired data)

Summary

- p-values encode our desired probability of a false positive
- Chi-square test compares count or rate data
- t-test compares means
- Paired t-test compares means within subjects
- Mann-Whitney U compares ranks for non-normal data