Hypothesis testing

Quantitative Methods for International Politics

IPOL 3270 • Fall 2026

November 8, 2026



Plan for today (and Tuesday)

- Evidentiary standards
- How to test any hypothesis
- Interpreting tests
- Other δs
- Math-based tests
- Problems with p-values

Quick review

Estimands

Parameters are the things we want to know about the population

Estimand	Parameter	Statistic	Example
Proportion	p	$\hat{p}$	Tunisians who have thought about emigrating
Mean	μ	$\bar{x}$	Average age of Tunisian adults
Difference in proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	Emigration: internet users vs. non-users
Difference in means	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	Age: urban vs. rural residents
Regression slope	β_1	b_1	Relationship between age and thinking about emigrating

- $\hat{p}$
- $\bar{x}$
- $\hat{p}_1 - \hat{p}_2$
- $\bar{x}_1 - \bar{x}_2$
- b_1

δ

(delta)

Your estimate

Whatever statistic answers your research question

How do we know if δ is any good?

Is it accurate?	Is it real?	Is it substantive?
Is δ a good guess?	Is δ real?	Does δ even matter?
Did it pick up the population parameter?	Could it maybe be zero?	
Confidence intervals	**Hypothesis tests and p-values**	**Your knowledge of the world—no stats!**

Do different groups of Tunisians think about emigrating at different rates?

Comparison	δ	95% CI
Women – men	0.1 points	-3.8 to 3.9
Internet users – non-users	26.1 points	22.3 to 29.9

- The gender interval includes 0, so “no difference” is plausible
- The internet interval is way above 0
- Is there a more formal way to ask “could it be zero”?

Evidentiary standards

Common law legal process

UK, India, Canada, Australia, US, etc.

- **The accused must be judged**
- **Presumption of innocence**
- **Accuser has burden of proving guilt**
- **Judge/jury decides guilt based on the amount of evidence**

We never prove innocence; we try (and fail) to reject innocence

Common law evidentiary standards

- **Preponderance of the evidence** (civil)
- **Clear and convincing evidence** (civil)
- **Beyond a reasonable doubt** (criminal)

Sharia court evidentiary standards

- **غلبة الظن / preponderance of assumption** (*ta'zir* and family matters)
- **اليقين / certainty** (*hudud* and *qisas* cases)

- **Why do we have all these levels?**
- **Why is the burden different for different types of cases?**
- **Why is it so high for criminal cases?**

We're afraid of punishing an innocent person!

"For if the Imam makes a mistake in forgiving it would be better than making mistake in punishment"

Jamiʿat-Tirmidhi 1424

فَإِنَّ الإِمَامَ أَنْ يُخْطِئَ فِي الْعَفْوِ خَيْرٌ مِنْ أَنْ يُخْطِئَ فِي الْعُقُوبَةِ

جامع الترمذي 1424


Actual truth
Guilty
Not guilty
Judge/jury decision
Guilty
Not guilty
Yay!
True positive
Oh no!
False positive (I)
Oh no!
False negative (II)
Yay!
True negative

Statistical "legal" system

- **Sample statistic δ must be judged**
- **Presumption of no effect** (null hypothesis, H_0)
- **You have the burden of proving δ is "real"**
- **You decide δ's "guilt" based on the amount of evidence**

We never prove that the null is true; we try (and fail) to reject the null

Two possible hypotheses

Is there a difference between men's and women's attitudes toward emigration in Lebanon?

Null hypothesis

$$H_0 : p_{\text{women}} - p_{\text{men}} = 0$$

No difference ("innocent")

Alternative hypothesis

$$H_A : p_{\text{women}} - p_{\text{men}} \neq 0$$

Some difference ("guilty")



α (alpha)

0.10

0.05

0.01

Your evidentiary standard: how often you're willing to make a Type I error

Statistical significance

There's enough evidence to safely declare that δ is "real" and not zero

"Rejecting the null hypothesis"

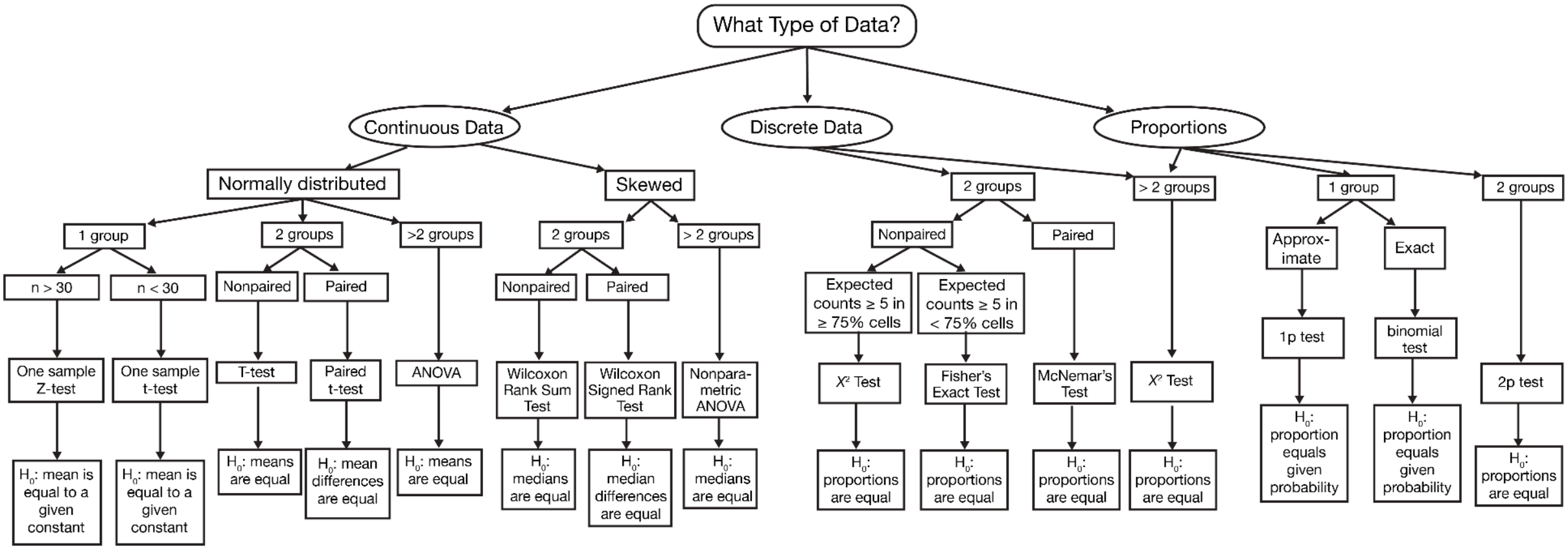
p-values

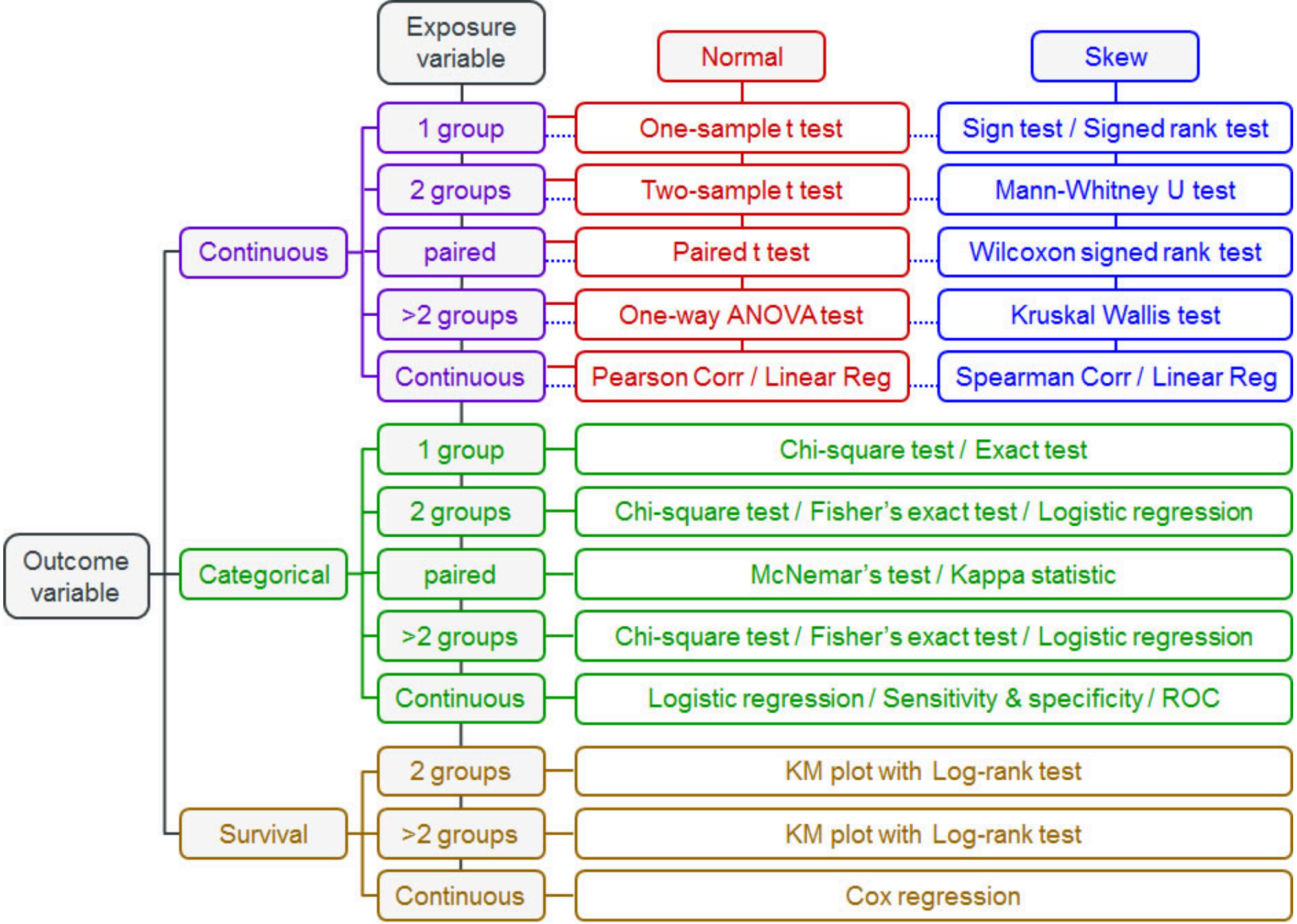
The probability of seeing a δ at least that extreme in a world where there's no effect

Wildly misunderstood—by everyone!

How to test any hypothesis

Flow chart: which test statistic should you use?

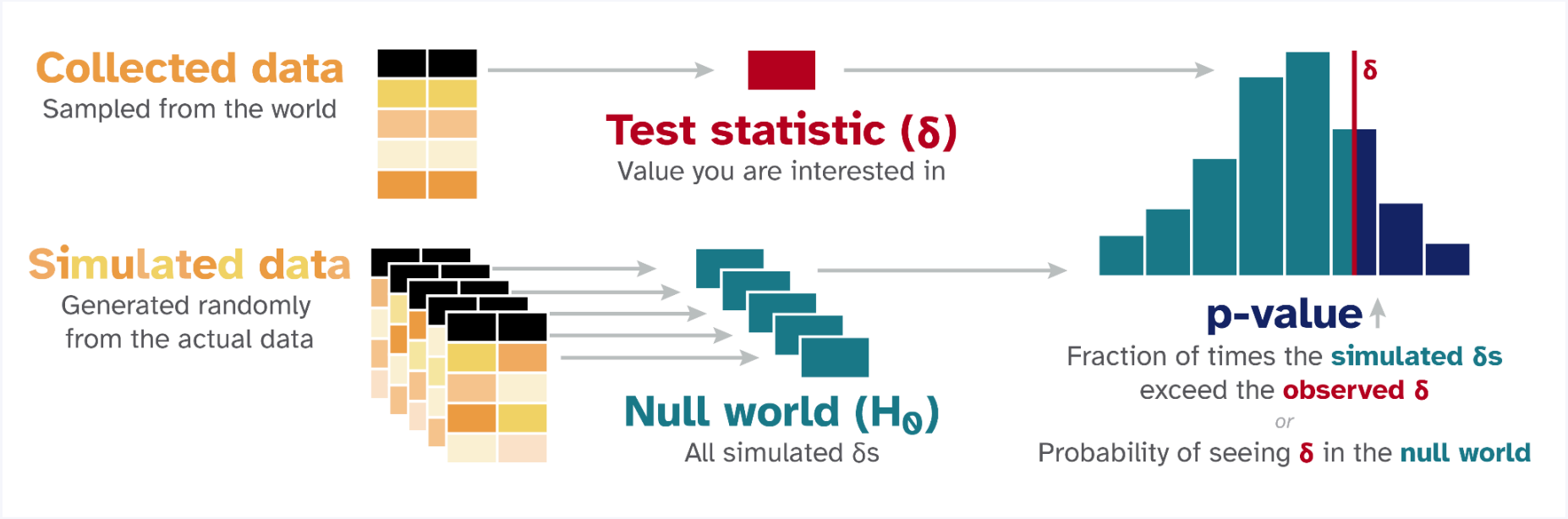


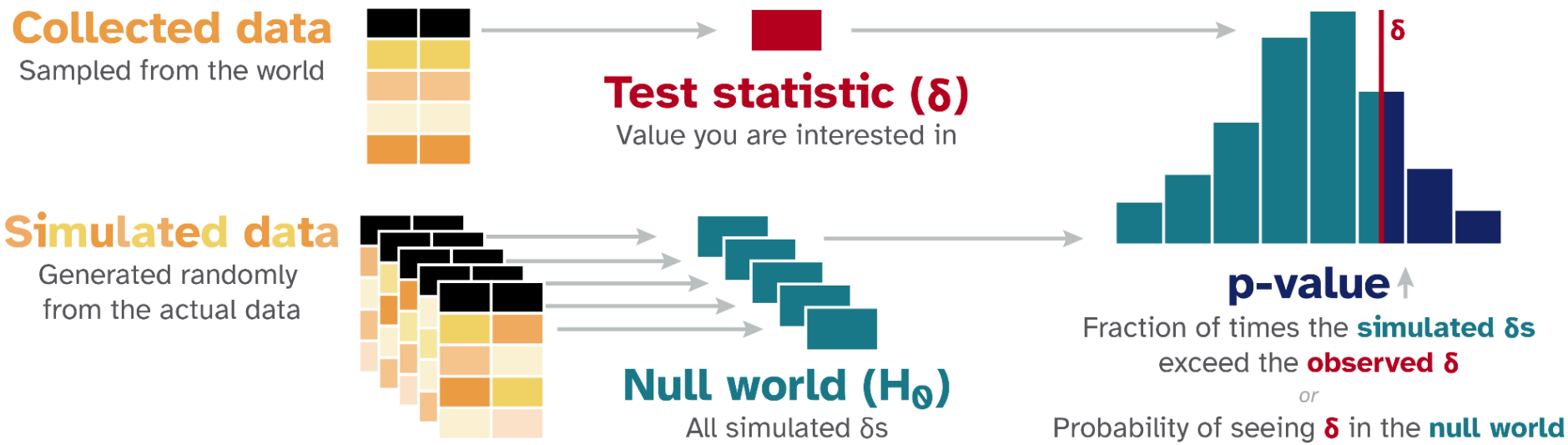

Outcome variable
Exposure variable
Normal
Skew
Continuous
1 group
One-sample t test
Sign test / Signed rank test
2 groups
Two-sample t test
Mann-Whitney U test
paired
Paired t test
Wilcoxon signed rank test
>2 groups
One-way ANOVA test
Kruskal Wallis test
Continuous
Pearson Corr / Linear Reg
Spearman Corr / Linear Reg
Categorical
1 group
Chi-square test / Exact test
2 groups
Chi-square test / Fisher's exact test / Logistic regression
paired
McNemar's test / Kappa statistic
>2 groups
Chi-square test / Fisher's exact test / Logistic regression
Continuous
Logistic regression / Sensitivity & specificity / ROC
Survival
2 groups
KM plot with Log-rank test
>2 groups
KM plot with Log-rank test
Continuous
Cox regression

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.262	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

There is only one test!

Simulations and null worlds




Collected data
Sampled from the world
Test statistic (δ)
Value you are interested in
δ
Simulated data
Generated randomly
from the actual data
Null world (H_0)
All simulated δs
p-value
Fraction of times the simulated δs
exceed the observed δ
or
Probability of seeing δ in the null world

1: Find δ

2: Invent a world where δ is null

3: Look at δ in the null world

4: Find the probability that δ could exist in the null world

5. Decide!

1: Calculate δ

Do Lebanese women and men think about emigrating at different rates?

- **Estimand**: $p_{\text{women}} - p_{\text{men}}$
- H_0: $p_{\text{women}} - p_{\text{men}} = 0$
- H_A: $p_{\text{women}} - p_{\text{men}} \neq 0$
- $\alpha = 0.05$

Gender	% yes	n
Women	32.3%	1,188
Men	36.1%	1,206

δ = -3.7 points

Is that real though? Could it be positive? Could it be zero?

2: Invent a world where δ is null

If gender doesn't matter, the categories "woman" and "man" are meaningless

- Shuffle them!
  - Keep everyone's answers, but randomly reassign who's a woman and who's a man
  - ("permutation shuffling" breaks any connection between the variables)
- Calculate δ for the shuffled data
- Repeat a bunch of times

Bootstrapping vs. permuting

Bootstrap

Resample **with** replacement

Shows how much δ bounces around

```
type = "bootstrap"
```

→ Confidence intervals

Permutation

Shuffle **without** replacement

Shows what δ looks like when there's no effect

```
type = "permute"
```

→ Hypothesis tests

```r
library(infer)

# δ from the actual data
obs_diff <- lebanon |>
  specify(
    thought_emigrating ~ gender,
    success = "Yes"
  ) |>
  calculate(
    stat = "diff in props",
    order = c("Female", "Male")
  )

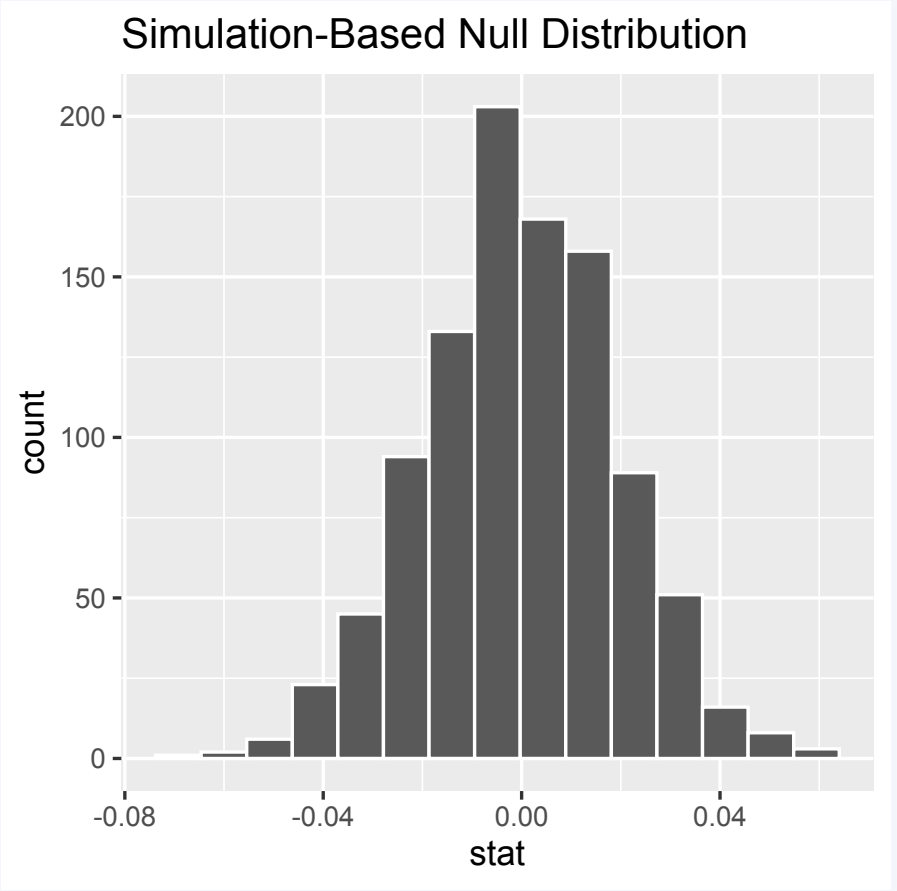
# Simulate null world
null_world <- lebanon |>
  specify(
    thought_emigrating ~ gender,
    success = "Yes"
  ) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(
    stat = "diff in props",
    order = c("Female", "Male")
  )
```

- `specify()`: What are the variables?
- `hypothesize()`: What does the null world look like? (`"independence"` = gender and emigration aren't related)
- `generate()`: Make the null world by shuffling (`"permute"`)
- `calculate()`: Find δ in each shuffled dataset

```
null_world
```

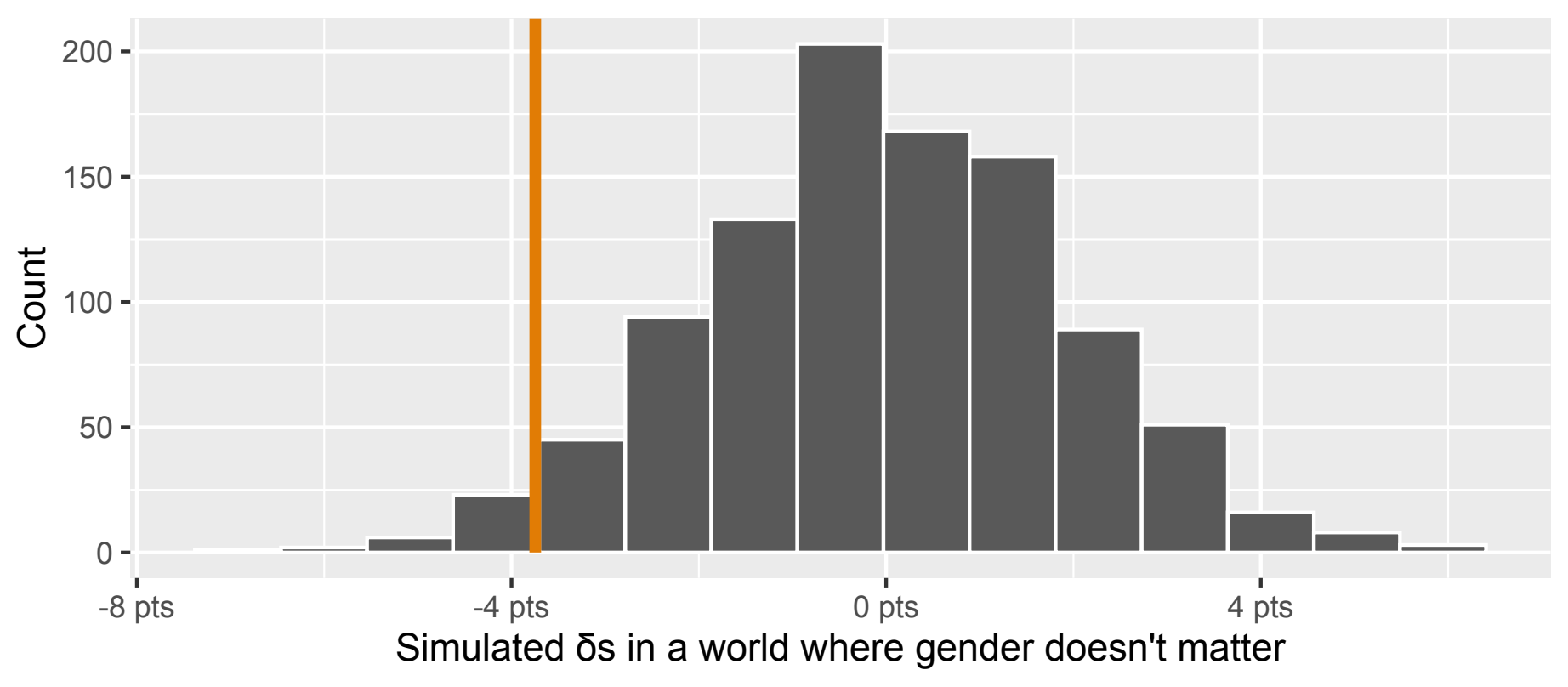
```
visualize(null_world)
```

```
Response: thought_emigrating
(factor)
Explanatory: gender (factor)
Null Hypot...
# A tibble: 1,000 × 2
   replicate     stat
       <int>    <dbl>
 1         1  0.0177
 2         2  0.0193
 3         3 -0.00906
```



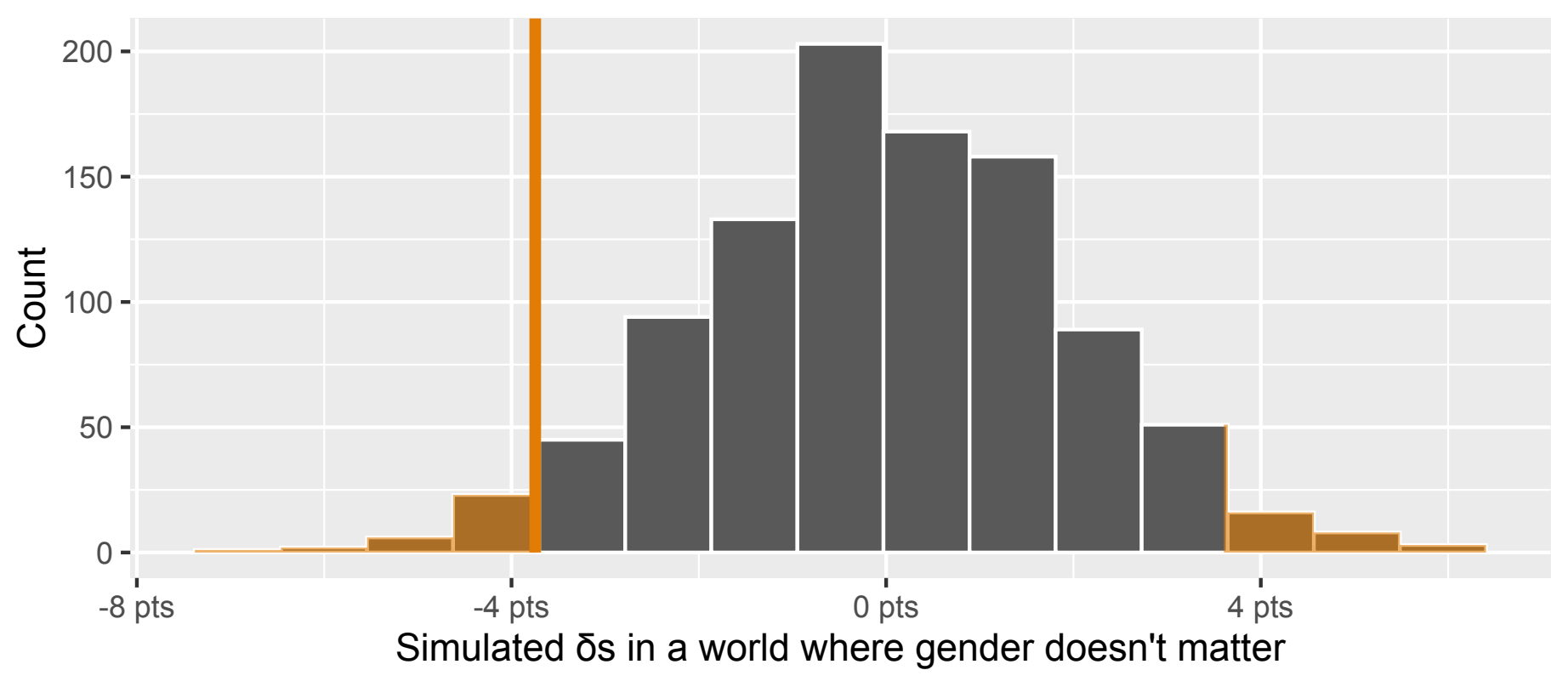
3: Look at δ in the null world

```
visualize(null_world) +
  shade_p_value(obs_stat = obs_diff, direction = NULL)
```



3: Look at δ in the null world

```
visualize(null_world) +
  shade_p_value(obs_stat = obs_diff, direction = "two-sided")
```



4: Find the p-value

Probability of seeing a δ at least as extreme as ours in this null world

```
null_world |>
  get_p_value(obs_stat = obs_diff, direction = "two-sided")
```

```
# A tibble: 1 × 1
  p_value
    <dbl>
1   0.064
```

In a world where Lebanese women and men think about emigrating at the same rate, there's a 6.4% chance that we'd see a difference at least as big as 3.7 percentage points

5: Decide!

$p = 0.064$

Do we have enough evidence to say that we don't live in the null world (i.e. to "convict", or to "reject the null hypothesis")?

- Is that below $\alpha = 0.05$?
- Is that below $\alpha = 0.10$?
- Is that below $\alpha = 0.01$?

andhs.co/nullworlds

p-values and confidence intervals

```
lebanon |>
  specify(
    thought_emigrating ~ gender,
    success = "Yes"
  ) |>
  generate(
    reps = 1000,
    type = "bootstrap"
  ) |>
  calculate(
    stat = "diff in props",
    order = c("Female", "Male")
  ) |>
  get_ci(level = 0.95)
```

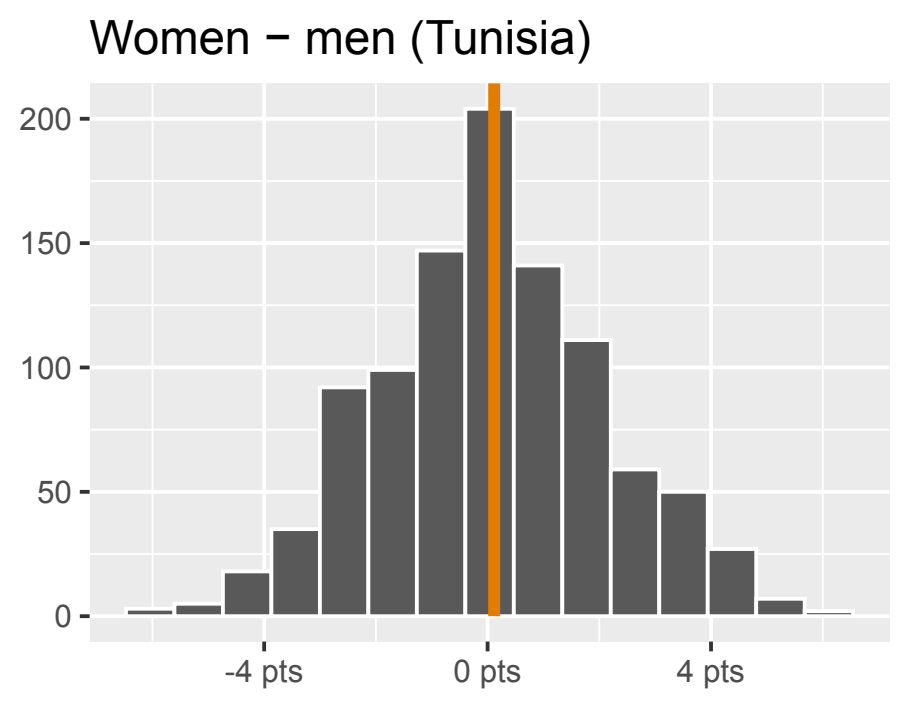
lower_ci	upper_ci
-0.076	0.0012

The 95% confidence interval includes 0

p is above 0.05

***KINDA SORTA SHORTCUT*: If a 95% CI includes 0, p > 0.05; if it excludes 0, p < 0.05**

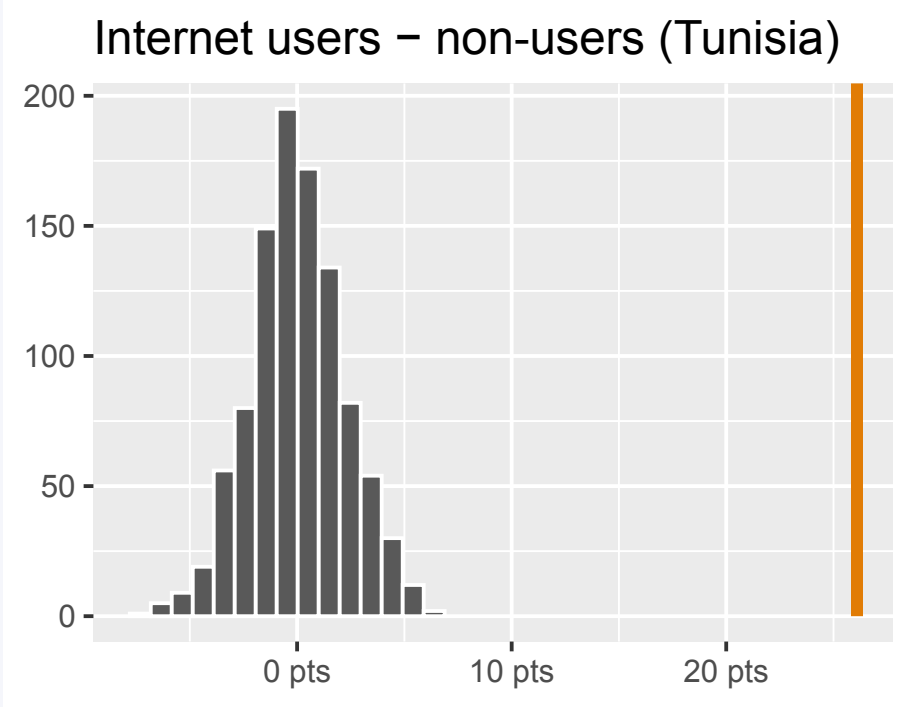
`thought_emigrating ~ gender`



δ = 0.1 points; **p ≈ 1**

Totally normal in the null world

`thought_emigrating ~ uses_internet`



δ = 26.1 points; **p < 0.001**

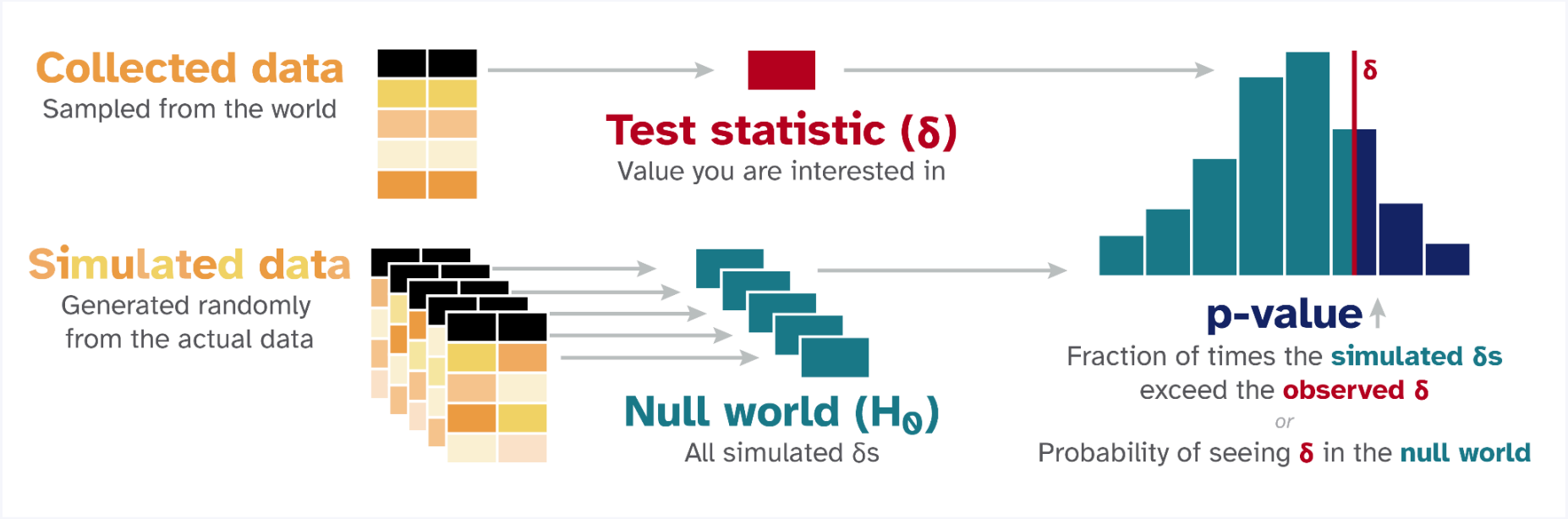
Way off the chart!

Our turn: test some hypotheses

Open `hypothesis-testing.qmd` in the project

1. Pick a country and a yes/no question and a comparison group
2. Write down H_0, H_A, and α *before* we look at the data
3. Go through all 5 steps
4. Interpret results

12:00



1. Find δ
2. Invent a world where δ is null
3. Look at δ in the null world
4. Find the probability that δ could exist in the null world
5. Decide!

Interpreting tests

Only two possible outcomes

Reject H_0

$p < \alpha$

δ doesn't fit comfortably in the null world, so we're pretty sure we're not in it, so δ is probably not 0

"Statistically significant"

Fail to reject H_0

$p \geq \alpha$

δ could plausibly show up in the null world

"Not statistically significant"

Temptations!

If $p \geq \alpha$ (like $p = 0.387$), you'll want to say there's no difference, or that δ is 0

No! Don't! Bad!

You never "accept" the null hypothesis!

~~"We find the defendant innocent"~~
"We find the defendant not guilty"

If p is really close to α (like $p = 0.052$), you'll want to say that δ is *almost* significant

No! Don't! Bad!

It's either "guilty" or not!

~~"We find the defendant almost guilty"~~
"We find the defendant not guilty"

The p-value template

In a world where there is no δ, the probability of seeing a δ at least as extreme as ours is p

Then you judge if p is below your evidentiary threshold (0.10, 0.05, 0.01, etc.)

Which is right?

A study finds that a 5-point difference has p = 0.032.

A. There's a 3.2% chance that the difference is just random

B. There's a 3.2% chance that our hypothesis is wrong

C. In a world where there's no difference, there's a 3.2% chance of seeing a difference at least as extreme as 5 points. That's below 0.05, so we say it's statistically significant and reject the null.

C! A p-value is about the data in a null world, not about whether your hypothesis is true

In Lebanon, women are 3.7 points less likely than men to have thought about emigrating (p = 0.064).

A. There's a 6.4% chance that the null hypothesis is true

B. If there were no gender gap among all Lebanese adults, we'd see a gap at least this big in about 6.4% of samples

C. Since p > 0.05, we've shown that Lebanese women and men think about emigrating at the same rate

D. With $\alpha = 0.10$, we'd reject the null hypothesis

B and D! C is "accepting" the null

Choosing a threshold

- Smaller α → fewer false positives, but more false negatives
- Bigger α → fewer false negatives, but more false positives
- Pick it before you see the data, based on how bad each mistake would be

Everyone uses `0.05`, but it's not magical!

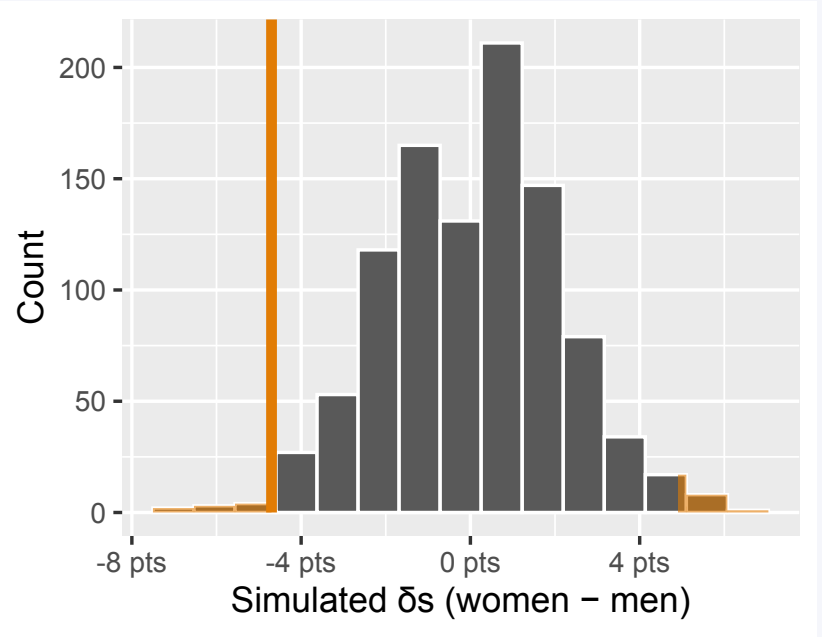
Other δs

Difference in proportions

Do Kuwaiti women and men think about emigrating at different rates?

$$H_0 : p_{\text{women}} - p_{\text{men}} = 0$$

```
null_world <- kuwait |>
  specify(
    thought_emigrating ~ gender,
    success = "Yes"
  ) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(
    stat = "diff in props",
    order = c("Female", "Male")
  )
```



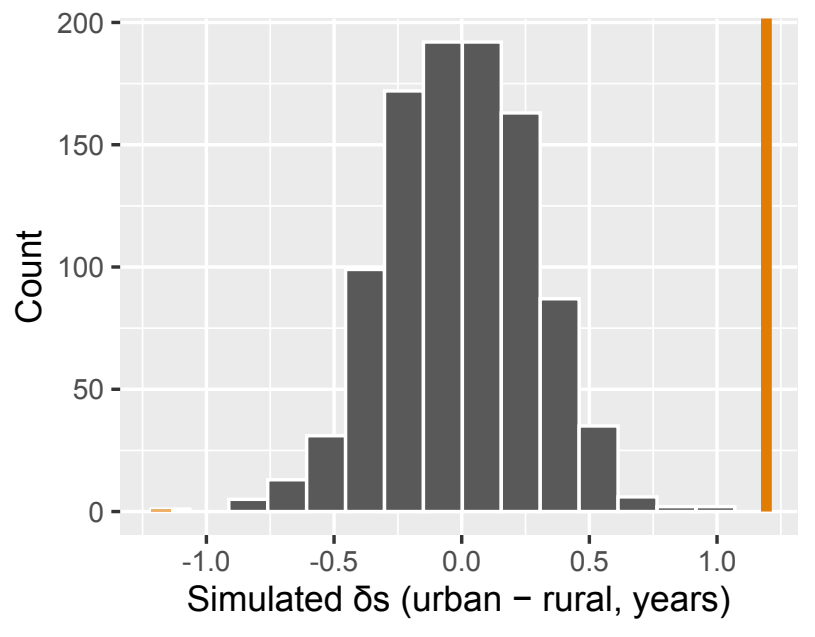
δ = -4.7 points; p = 0.018

Difference in means

Do people in cities and rural areas have different average ages?

$$H_0 : \mu_{\text{urban}} - \mu_{\text{rural}} = 0$$

```
null_world <- urban_rural |>
  specify(age ~ urban_rural) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(
    stat = "diff in means",
    order = c("Urban", "Rural")
  )
```



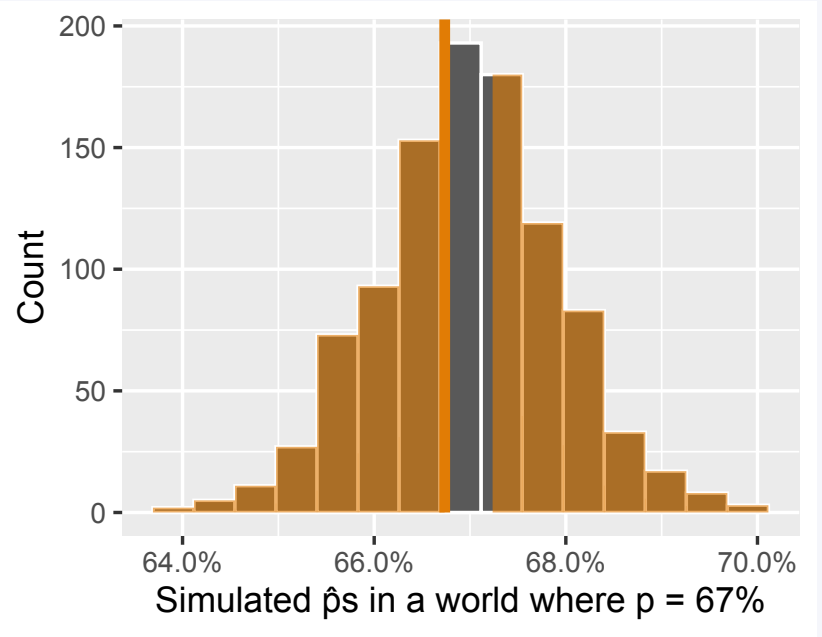
δ = 1.2 years; p < 0.001

One proportion

About 67% of the world used the internet in 2023. Is Tunisia different from the rest of the world?

$$H_0 : p_{\text{internet}} = 0.67$$

```
null_world <- tunisia |>
  drop_na(uses_internet) |>
  specify(
    response = uses_internet,
    success = "Yes"
  ) |>
  hypothesize(null = "point", p = 0.67) |>
  generate(reps = 1000, type = "draw") |>
  calculate(stat = "prop")
```



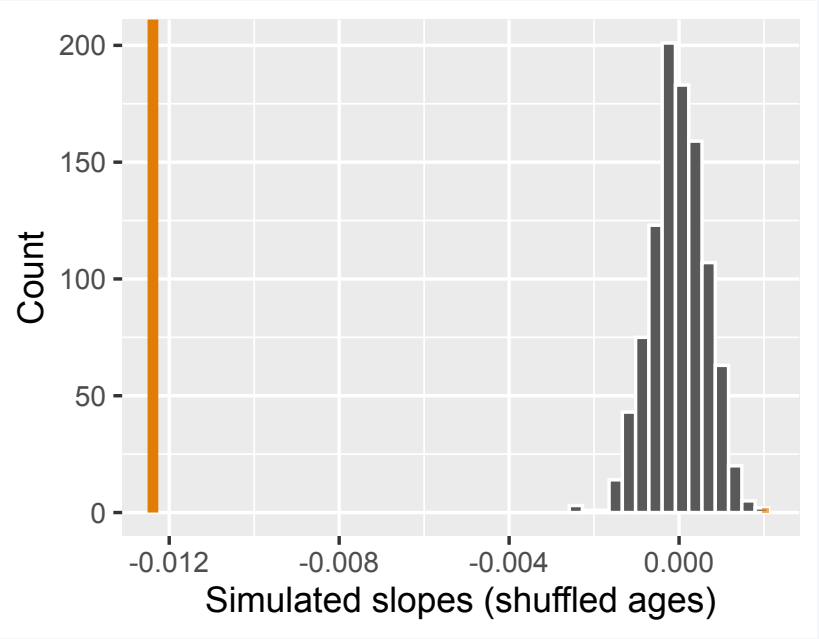
$\hat{p}$ = 66.7%; p = 0.790

Regression slope

Are older Tunisians less likely to have thought about emigrating?

$H_0 : \beta_1 = 0$

```
null_world <- tun_ages |>
  specify(yes_num ~ age) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "slope")
```



b_1 = -0.0124; p < 0.001

Same general code

	Difference in proportions	Difference in means	One proportion	One mean	Regression slope
`specify()`	`y ~ group, success = "Yes"`	`y ~ group`	`response = y, success = "Yes"`	`response = y`	`y ~ x`
`hypothesize()`	`null = "independence"`	`null = "independence"`	`null = "point", p = 0.67`	`null = "point", mu = 40`	`null = "independence"`
`generate()`	`type = "permute"`	`type = "permute"`	`type = "draw"`	`type = "bootstrap"`	`type = "permute"`
`calculate()`	`stat = "diff in props"`	`stat = "diff in means"`	`stat = "prop"`	`stat = "mean"`	`stat = "slope"`

Math-based tests

Do I have to simulate everything?

No!

- Math-based tests assume the null world has a known shape (normal, t, χ^2)
- (All those flowcharts and tables are for choosing the right flavor of math-based null world)

Same intuition:
how weird is δ in a world where it's 0?

Simulation vs. math

```
lebanon |>
  prop_test(
    thought_emigrating ~ gender,
    success = "Yes",
    order = c("Female", "Male")
  )
```

Method	p
Simulation (1,000 shuffles)	0.064
Math (χ^2 test)	0.059

Basically the same answer!

Difference in means with *t*-tests

Do people in cities and rural areas have different average ages?

```
urban_rural |>
  t_test(
    age ~ urban_rural,
    order = c("Urban", "Rural"),
    alternative = "two-sided"
  )
```

estimate	lower_ci	upper_ci	statistic	p_value
1.19	0.64	1.75	4.24	2.31e-05

- `statistic`: how many standard errors δ is away from 0
- `p_value`: the probability of seeing that in the null world
- `lower_ci` and `upper_ci`: the 95% confidence interval

Slopes

```
model <- lm(yes_num ~ age, data = tun_ages)
get_regression_table(model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.9755	0.0275	**35.5153**	**0**	0.9217	1.0294
age	-0.0124	0.0006	**-21.2940**	**0**	-0.0135	-0.0112

- Every row is a hypothesis test with $H_0 : \beta = 0$
- `statistic` is a *t* statistic, and `p_value` comes from a *t* distribution
- A `p_value` of 0 isn't really 0; it's just super tiny (report it as p < 0.001)
- Shuffling ages 1,000 times: none of the simulated slopes get anywhere near -0.0124

Build intuition

Even if you never simulate a p-value again, thinking about null worlds is valuable!

Every math-based p-value you see still measures the same thing: how surprising would this δ be in a world where it's actually 0?

Problems with p-values

Journal of Economic Perspectives—Volume 35, Number 3—Summer 2021—Pages 157–174

Statistical Significance, *p*-Values, and the Reporting of Uncertainty

Guido W. Imbens

Theodor Geisel, better known by the nom de plume Dr. Seuss, published *The Sneetches* in 1961. In this children's story, the Star-Belly Sneetches viewed themselves as superior to the Plain-Belly Sneetches. When the character of Sylvester McMonkey McBean arrives with a machine that can add or remove belly stars (for a modest fee), social upheaval results. In empirical work in economics, stars have long been attached to numbers in tables and figures to indicate the level of statistical significance: one star typically refers to an estimate that is statistically significant at a 10 percent level; two stars, the 5 percent level; and the coveted three stars, the 1 percent level. In the word of Dr. Seuss: "Those stars weren't so big. They were really so small./You might think such a thing wouldn't matter at all./But, because they had stars, all the Star-Belly Sneetches/Would brag, 'We're the best kind of Sneetch on the beaches.'" In empirical studies, estimates with one, two, or three stars are often viewed as superior to those without such adornments.

ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

3. Principles

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**

   A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

   Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

   Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. **Proper inference requires full reporting and transparency**

   P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the

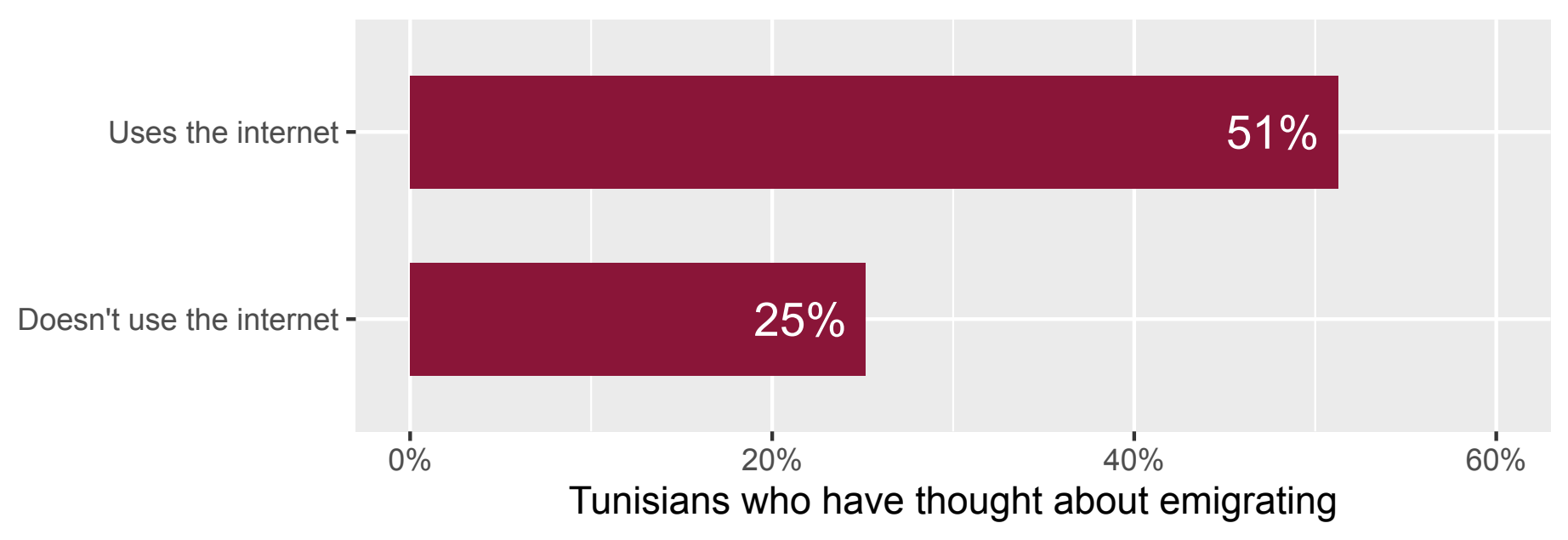
Significant ≠ substantive

Lives in	Average age	n
Rural	38.5	3,702
Urban	39.7	10,342

- δ = 1.2 years; $p < 0.001$
- Super statistically significant!
- But does a 1-year age gap matter for anything?
- With ~14,000 people, even tiny differences become “significant”

Is δ real? Use stats. Does δ matter? That’s up to you!

Do you even need a test?



- 2,400 people and a 26-point gap. Nobody needs a p-value for this!
- Look at your data before you run tests

p-hacking

Test enough things and something will be significant

xkcd #882, “Significant”

“Hack Your Way To Scientific Glory”

Decide on your hypotheses and α before you look at the data, and report everything you tested

Preregistration