Class 10: Hypothesis testing

Alison Presmanes Hill

General idea for hypothesis testing

- Partition the *parameter* sample space into two regions
- One region is defined by the **null hypothesis**: H_0
  - Usually what you wish to see evidence against
  - If the parameter estimate based on my sample is **null/dull**, then it will fall in this region (given some reasonable amount of random variation)
- Other region defined by the **alternative hypothesis**: H_1
  - Usually what you wish to see evidence in support of
  - If the parameter estimate based on my sample is **interesting**, then it will fall in this region (and thus, outside of the null/dull region)
- This is traditional null hypothesis significant testing (**NHST**)
  - Also called reject-support hypothesis testing

General idea for a test statistic

- A numerical summary used to collapse sample data into a single number
- When "big" or "extreme", suggests that the observed data is very unexpected under the null hypothesis H_0
- A p-value quantifies this incompatibility between the data and H_0 – specifically, it's a tail probability
- So, to get a p-value, you must know or approximate the probability distribution of the test statistic under the null H_0
- This means we need to know what the *null* sampling distribution looks like

General idea for an estimator

- Point estimate is your single best guess at the parameter
- Interval estimate, i.e. confidence interval, provides a set of possible values for that parameter value that are "compatible" with the data
- Construction of confidence interval requires knowledge of the estimator's distribution

Therefore…

- To complete a hypothesis test or convey the precision of a point estimate, we need the statistic's or estimator's *sampling distribution*
- "sampling" here should invoke "long-run", "hypothetical repeats of the experiment"
- For an estimator, the standard deviation of the sampling distribution is called the *standard error*

General idea for a p-value

- Probability under the null H_0 of observing a test statistic value as or more extreme than that computed from the observed data (your sample data)
- The p-value is **not** the probability that the null hypothesis is true
- Example in a two-sided test, i.e. when both very small and very large values of test stat are "extreme":

$$\text{p-value}(\text{obs. test statistic}) = P(|\text{test statistic rv}| \geq |\text{obs. test statistic}|)$$

"A *p*-value is a measure of how embarrassing the data are to the null hypothesis"

--Nicholas Maxwell

Maxwell (2004). Data Matters: Conceptual Statistics for a Random World. Emeryville, CA: Key College Publishing.

Cut-offs

- Whether we like to admit it or not, p-values ultimately are cut-offs:

p-value < **α**	p-value ≥ **α**
Hit	Not hit
Statistically significant	Not statistically significant
Fame and glory!	?
Reject H_0	Accept H_0 (cringe) Fail to reject H_0 (eye-roll)

P-value

- Standardized measurement of evidence
- Measure of probabilistic significance, not conceptual.
- **LOW P-VALUE:** low probability of sample looking like the null population
  - Decision → reject the null
- **HIGH P-VALUE:** high probability of sample looking unlike the null
  - Decision → do not reject null

Musing on p-values

- in some sense, it's supreme laziness (cleverness?) to work this way: easy on the analyst only need to characterize dist'n of the test stat under the null
- downside: an indirect, nonspecific measure of how interesting the data is
- just saying something is "not null" or "not boring" is not exactly equivalent to saying what's truly "exciting" about it
- another way in which we "work backwards" Bayesian critique

Process

1) Write down **null** and **alternative** hypotheses
2) Figure out good test statistic that estimates the parameter you are interested in (i.e., what numeric summary based on your sample captures what you want)
3) Work out null distribution for that statistic
4) Calculate p-value by comparing actual value to null distribution
5) Reject H_0 if p-value < $\boldsymbol{\alpha}$

Hypotheses

- Null hypothesis = H_0
- Alternative hypothesis = H_a / H_1

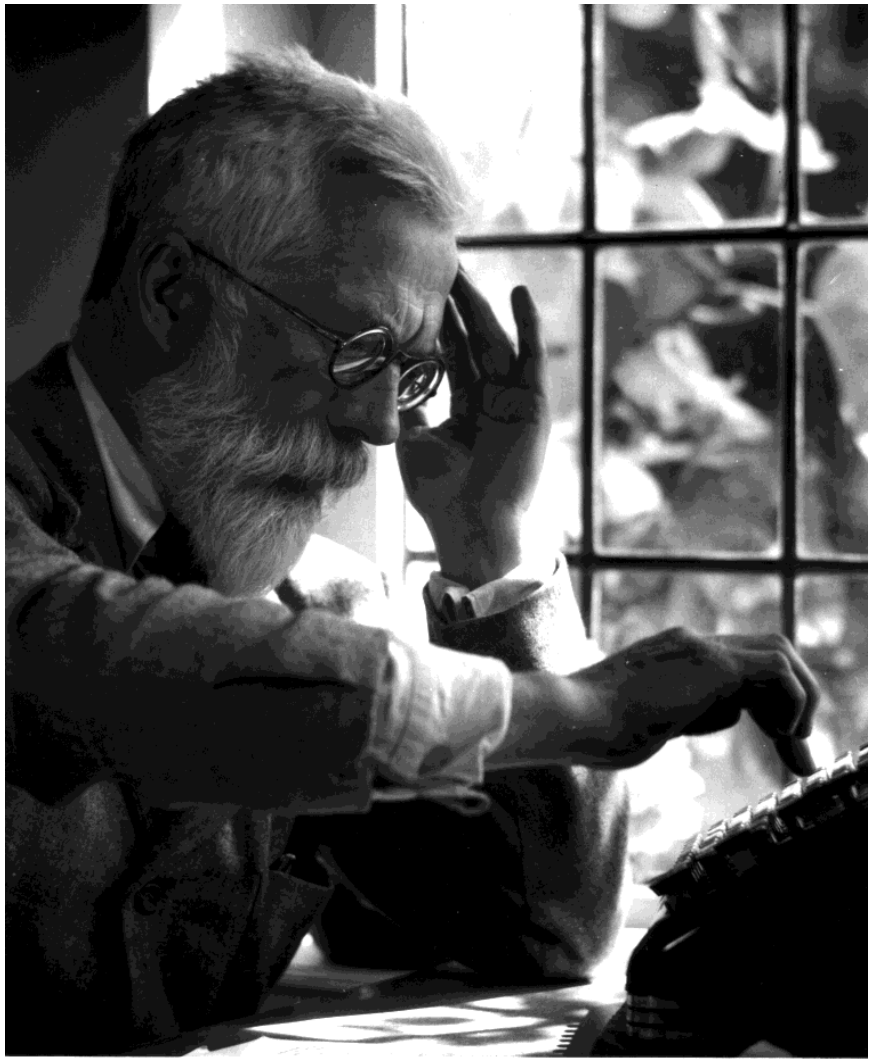
HINT: we have to know the null distribution, so H_0 will always be of the form:

Some parameter = some value as in…

$\mu = 0$

R. A. Fisher

- A experiment by R. A. Fisher (famous early statistician, 1890-1962)
- A lady at a tea party claims that she can tell the difference between putting the milk in first and second.
- How can we be sure?



The lady tasting tea

- Fisher set up an experiment
- The lady (Muriel Bristol) was presented with 8 cups of tea
  - 4 had tea with milk added second
  - 4 had milk with tea added second
- She was asked to select the 4 cups prepared by putting the milk in first and second.
- First, how many “draws” can Muriel take of the 8 cups?

How many ways can Muriel choose?

What is the sample space?

Right	Wrong	#	%
4	0	$\binom{4}{4}\binom{4}{0} = 1$	1
3	1	$\binom{4}{3}\binom{4}{1} = 16$	23
2	2	$\binom{4}{2}\binom{4}{2} = 36$	51
1	3	$\binom{4}{1}\binom{4}{3} = 16$	23
0	4	$\binom{4}{0}\binom{4}{4} = 1$	1
Total		$\binom{8}{4} = 70$	100

Hypergeometric distribution (discrete)

Arguments

- x, q vector of quantiles representing the number of white balls drawn *without replacement* from an urn which contains both black and white balls.
- m the number of white balls in the urn. [cups with milk → tea]
- n the number of black balls in the urn. [cups with tea → milk]
- k the number of balls drawn from the urn. [number of "draws"]
- p probability, it must be between 0 and 1.
- nn number of observations. If length(nn) > 1, the length

is taken to be the number required.

The lady tasting tea

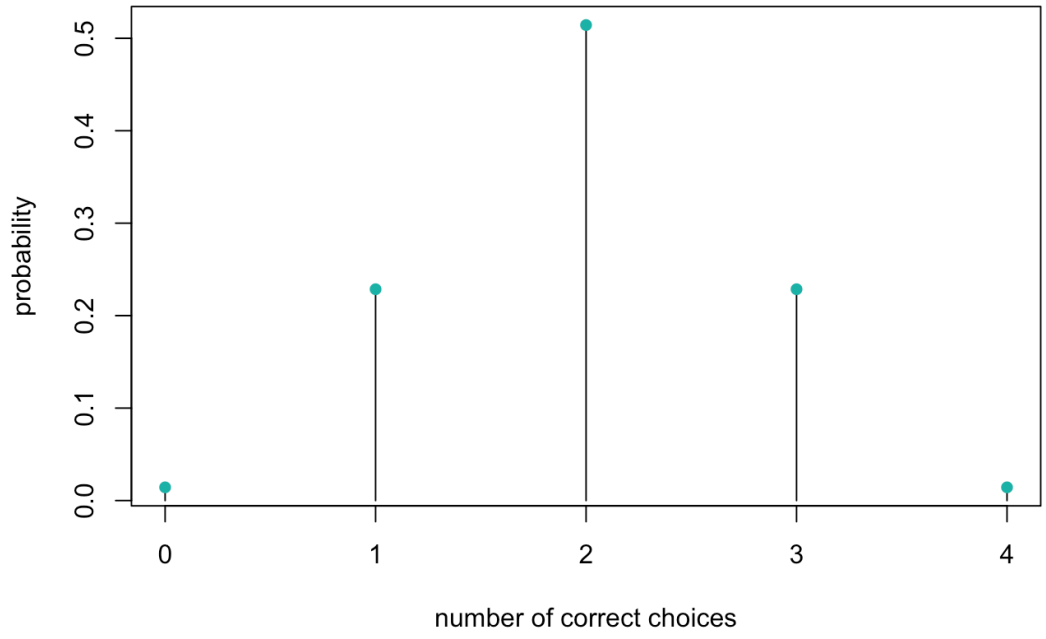
- As the story goes, the lady identified all 4 cups correctly. What were the chances (exactly!)?

```
> dhyper(x, m, n, k)
[1] 0.01428571
> 1 - phyper(q, m, n, k)
[1] 0.01428571
```

The lady tasting tea

- As the story goes, the lady identified all 4 cups correctly. What were the chances (exactly!)?

```
> dhyper(4, 4, 4, 4)
[1] 0.01428571
> 1 - phyper(3, 4, 4, 4)
[1] 0.01428571
```



The lady tasting tea

- The expected value of a hypergeometric variable is known to be:

```
k <- 4 # number of cups picked
m <- 4 # number of true "hits"
N <- 8 # total number of cups
k * (m/N) # population mean
[1] 2
```

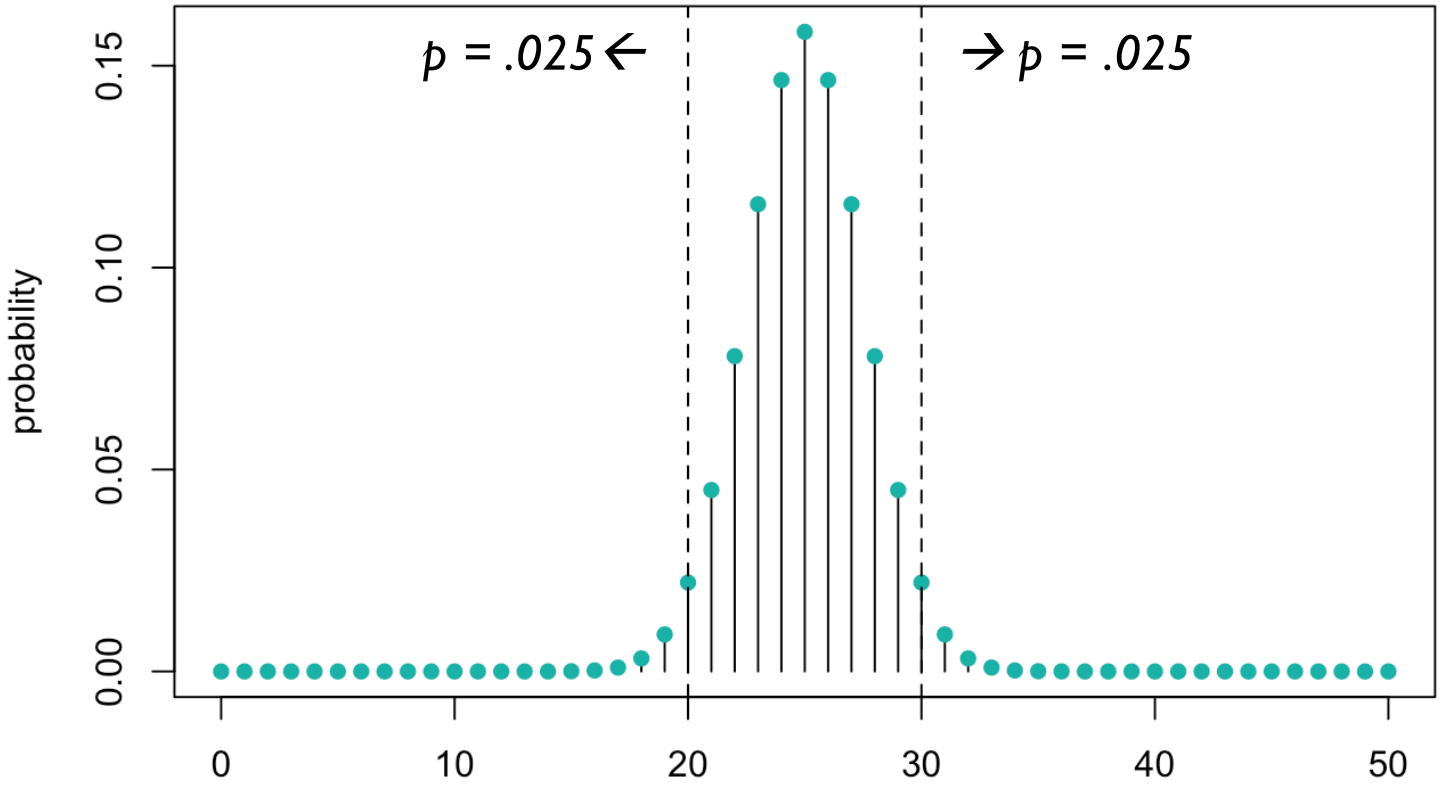
- What if we wanted to know whether Muriel was *better* than "average", with $\alpha = .05$?

What if we wanted to know whether Muriel was *better* than "average", with $\alpha = .05$?

- What is the null hypothesis?
- What is the alternative hypothesis?
- What if we ramp up to 100 cups of tea?
- What will accept as evidence of better than average?

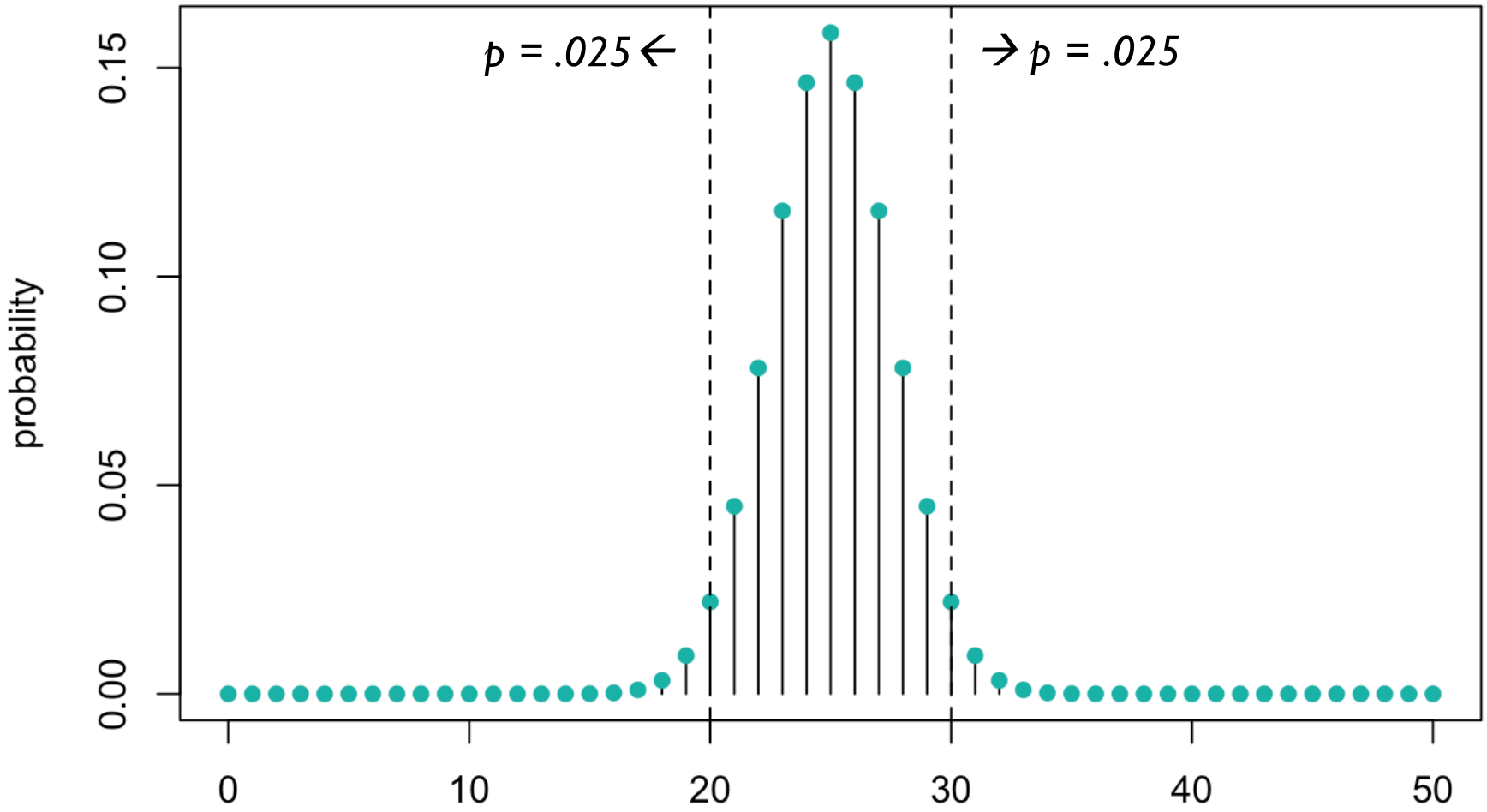
95% confidence interval for known μ with 100 cups

```
qhyper(.975, 50, 50, 50) # upper tail p = .025
[1] 30
qhyper(.025, 50, 50, 50) # lower tail p = .025
[1] 20
```



What if Muriel got exactly 31 correct?

```
dhyper(31, 50, 50, 50)
[1] 0.009163535
```



The null distribution

- If we know the population distribution, we can construct the null distribution: the distribution of a test statistic if the null hypothesis is true
  - Here, Fischer figured out that the null distribution was hypergeometric- you may not always be so lucky!
- If we don't know the population distribution, we can simulate it using resampling
- In the *two sample* scenario, this means we resample observations from our sample data without replacement, each time reshuffling the "sample" or "group" assignment.

Permutation testing

- **Idea:**
  Simulate the sampling distribution under the null hypothesis by shuffling the group labels repeatedly and computing the desired statistic.

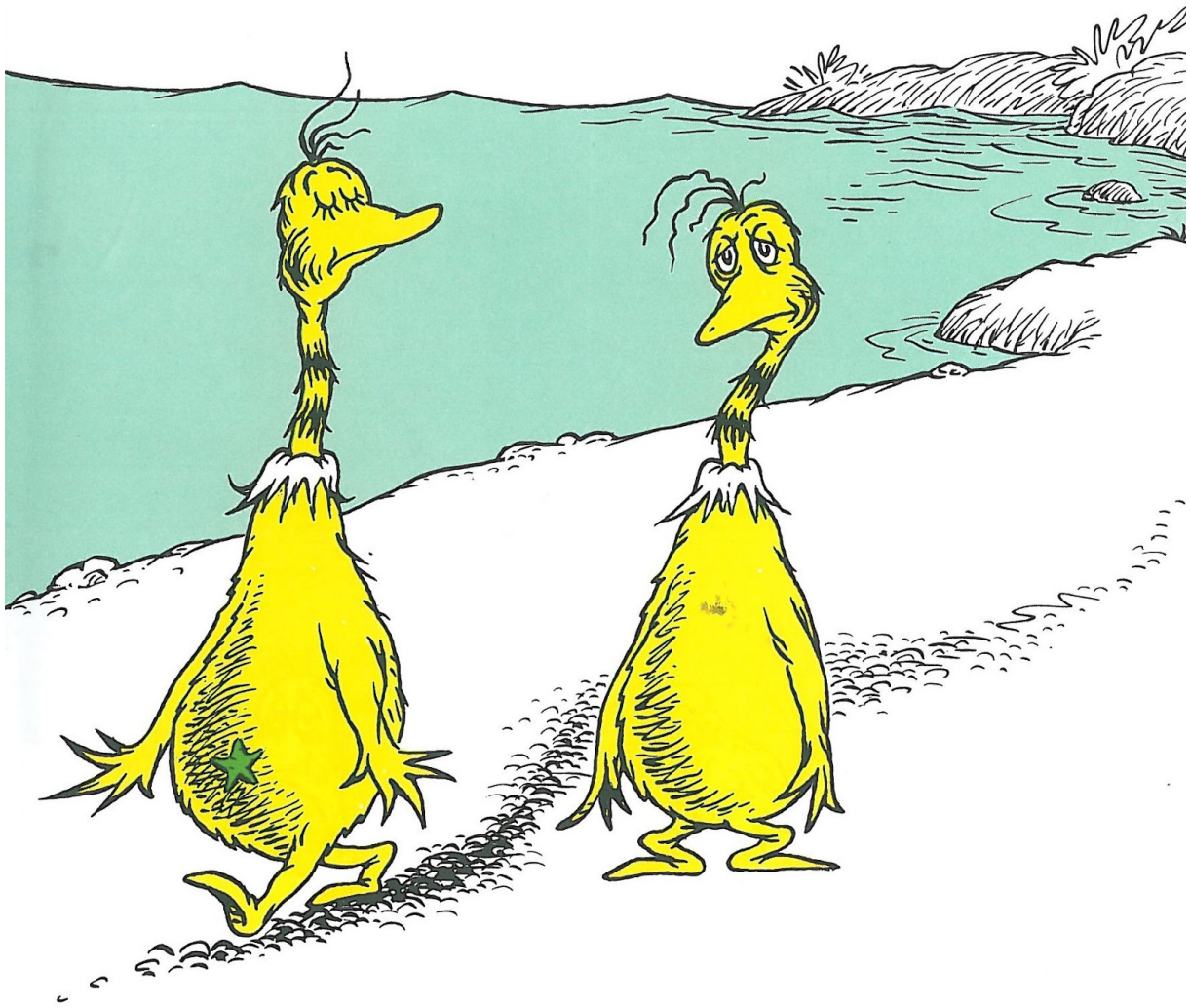
- **Motivation:**
  If the labels really don't matter, then switching them shouldn't change the result! We do this when we want to test whether observations from two groups follows the same distribution, without making any assumptions about the shape of that distribution (e.g., normality).

From Chihara & Hesterberg:

- Say that you have one sample with *m* observations, and a second sample with *n* observations

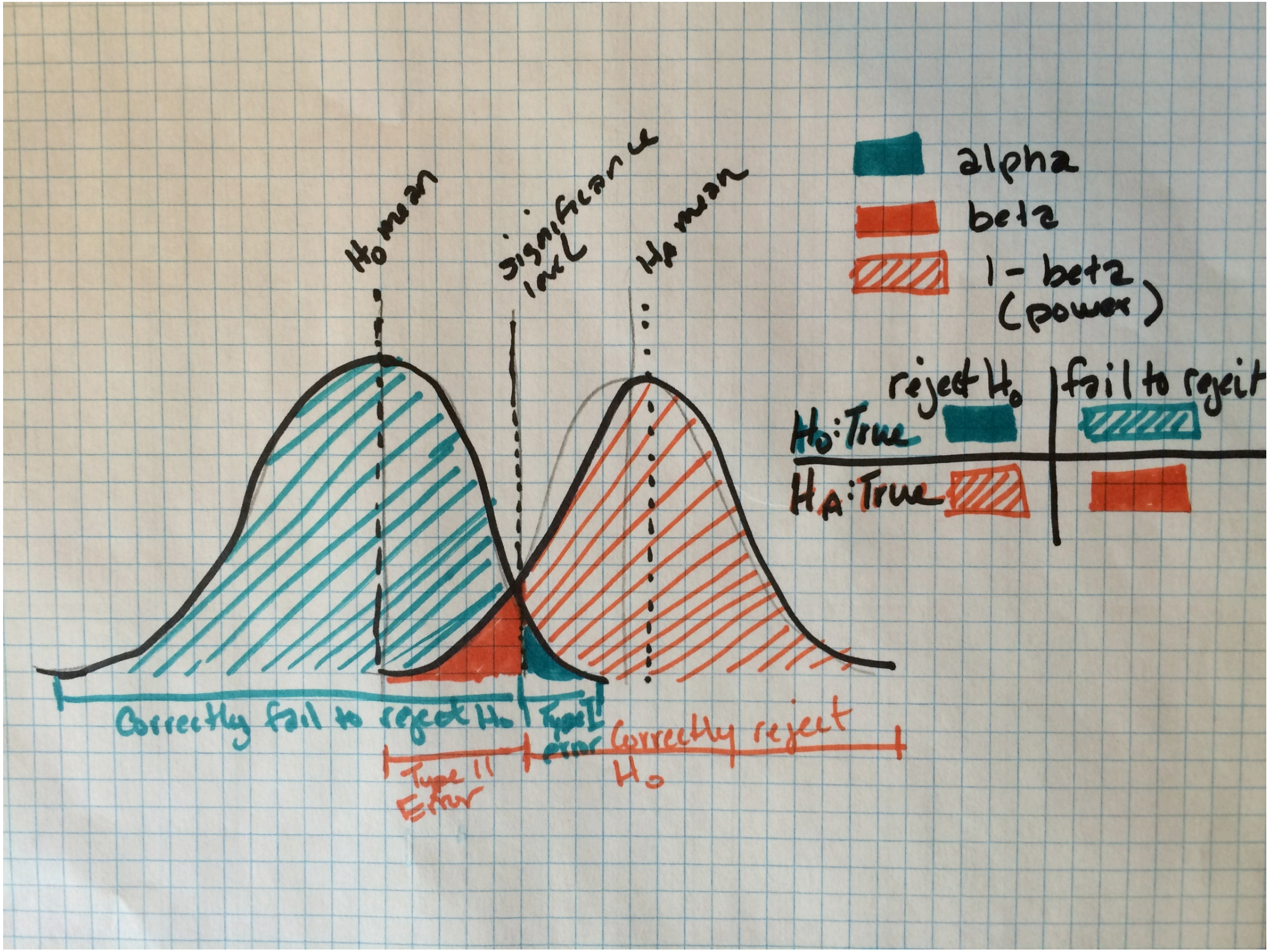
1. Pool the m + n values
2. **Repeat…**
   - Draw a resample of size m without replacement
   - Use the remaining n observations for the other sample
   - Calculate the difference in means or another statistic that compares samples
3. **…until you have enough samples**
4. Calculate the p-value as the fraction of times the random statistics exceed the original statistic (note: "Exceeds" generally means ≥ rather than >)
5. Multiple by 2 for a two-sided test, or use abs()

The star-bellied sneetches



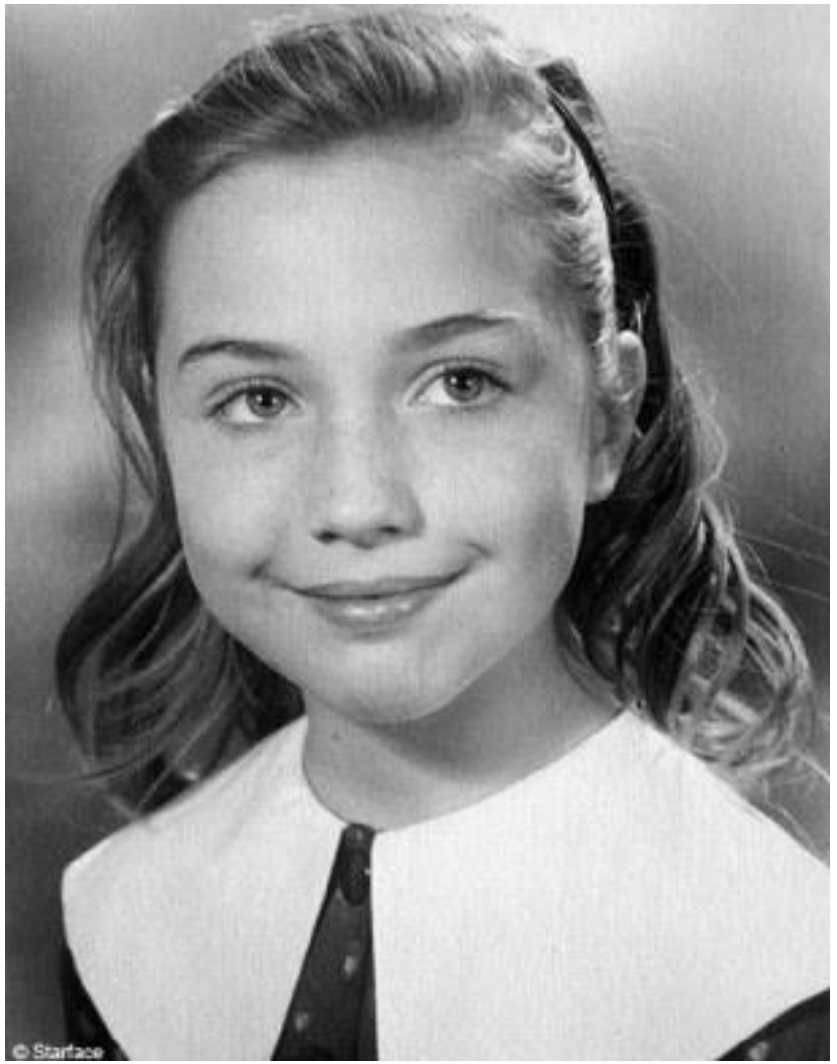
Confusion matrix

	Call based on observed data		
True state of the world	Fail to reject H_0	Reject H_0	
H_0	True negative $1 - \alpha$	False positive Type I error α	# true H_0's
H_1	False negative Type II error β	True positive $1 - \beta$	# true H_1's
		# rejected H_0's	# total tests


alpha
beta
1 - beta (power)
H_0 mean
significance level
H_A mean
reject H_0
fail to reject
H_0: True
H_A: True
Correctly fail to reject H_0
Type I error
Correctly reject H_0
Type II Error

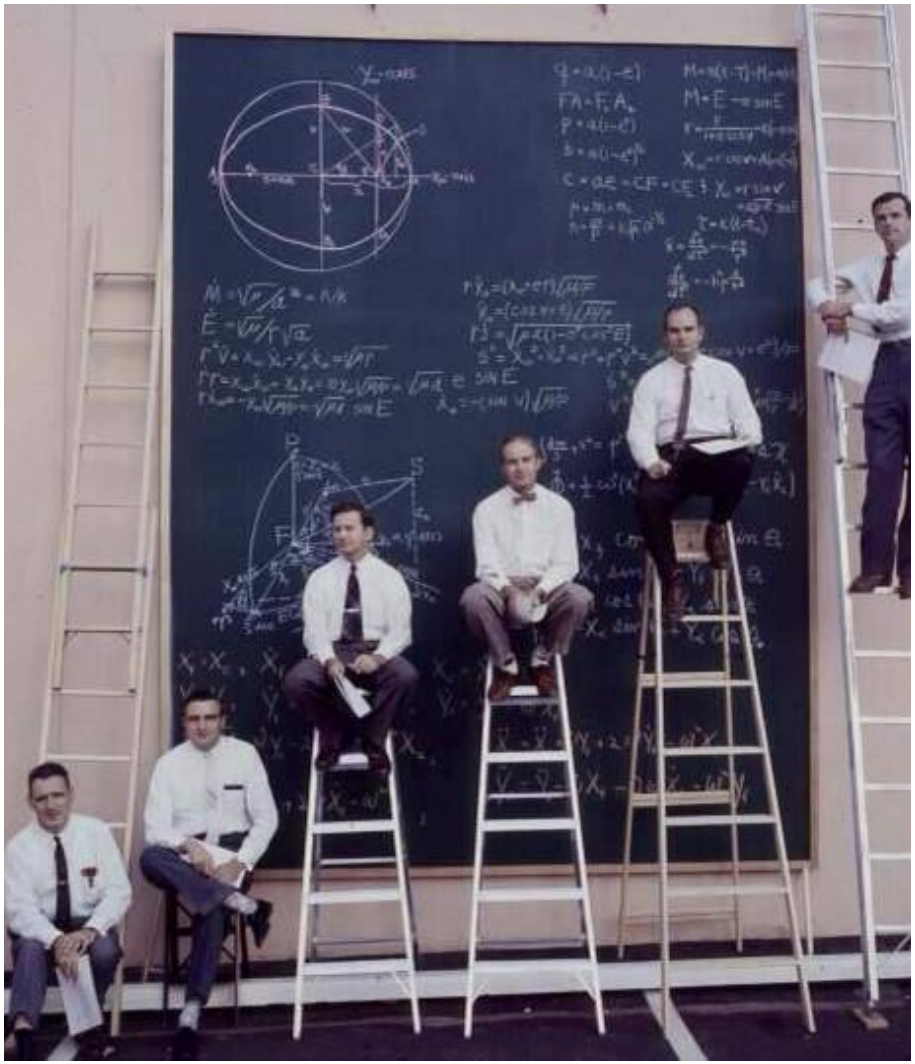
Aspiring astronauts

- Inspired by Alan Shepard, the first American to journey into space, a 14-year-old Hillary Rodham from suburban Chicago wrote a letter to NASA in 1961 asking what she needed to do to become an astronaut.



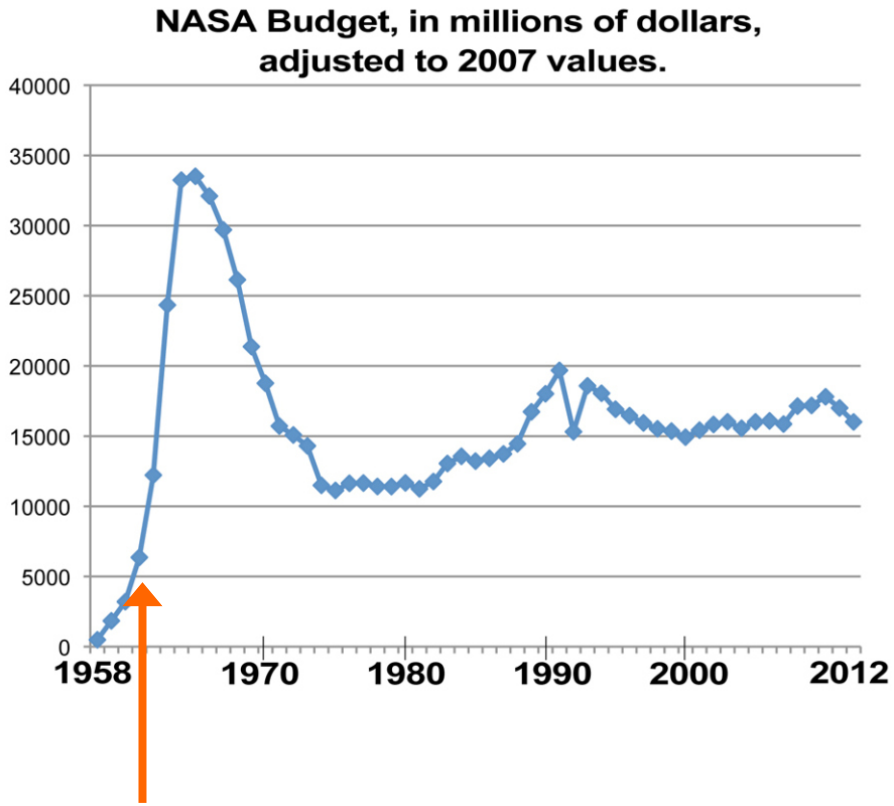
Let's pretend

- Upon receipt of the letter in 1961, NASA decided to conduct a study to test whether girls who are aspiring astronauts in high school have "above average" IQ



The (fictitious) study

- Unfortunately, the NASA budget in 1961 was pretty low
- So they studied only 25 high school girls, all of whom were aspiring astronauts (AA)
- Population (normally distributed):
  μ = 100; σ = 15



A (directional) alternative hypothesis

- H_0:

AA will have lower or comparable IQs;

$$\mu_{aa} \leq \mu_0$$

$$\mu_{aa} \leq 100$$

- H_1:

AA will have higher IQs;

$$\mu_{aa} > \mu_0$$

$$\mu_{aa} > 100$$

Rearrange…

- H_0:

AA will have lower or comparable IQs;

$$\mu_{aa} \leq \mu_0$$

$$\mu_{aa} - \mu_0 \leq 0$$

$$\mu_{aa} - 100 \leq 0$$

- H_1:

AA will have higher IQs;

$$\mu_{aa} > \mu_0$$

$$\mu_{aa} - \mu_0 > 0$$

$$\mu_{aa} - 100 > 0$$

Obtaining a test statistic

- Recall that we know the population mean and variance/standard deviation
- Remember our general formula for any test statistic about some generic parameter, θ:

$$\frac{\hat{\theta} - \theta_0}{SE_{\theta_0}}$$

Let's pretend we live in a crazy 1960's world, and NASA interns calculated just the sample mean and not the standard deviation before they "lost" the data. Luckily, we know μ *and* $\boldsymbol{\sigma}$.

What is the NULL distribution of this test statistic?

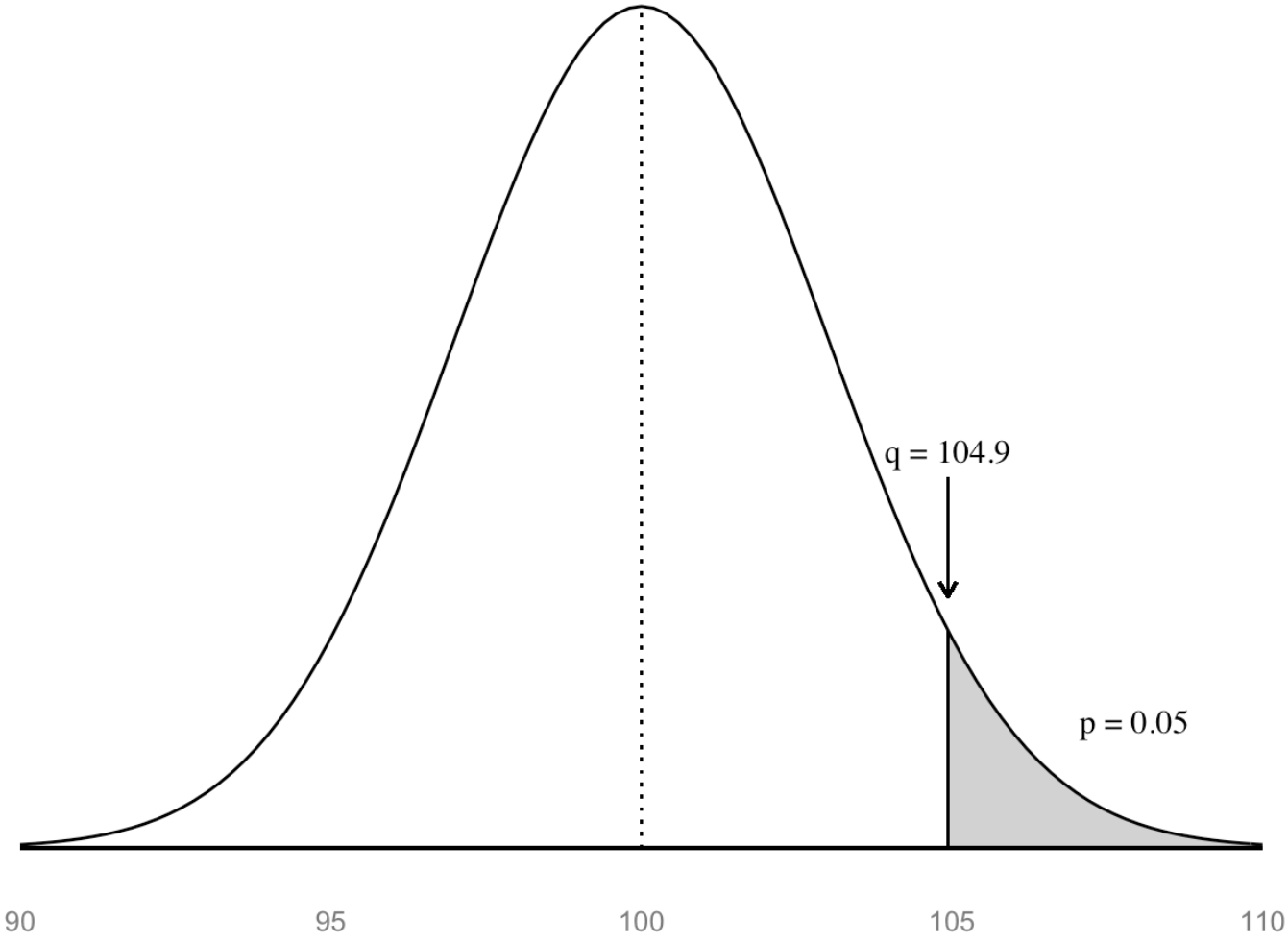
Before we even look at the data:
What is the mean and standard deviation (the "standard error") of the null distribution?

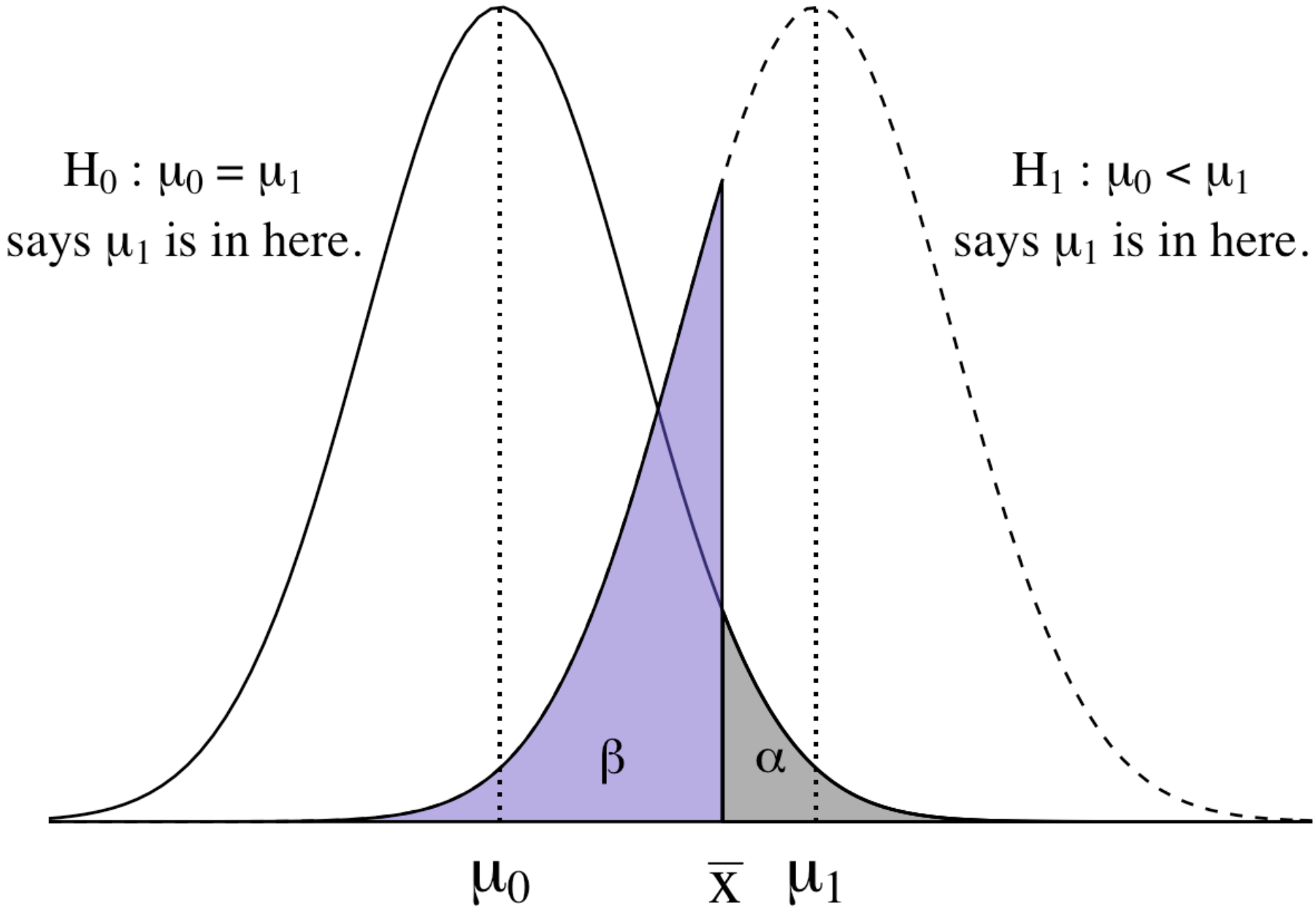
What shall we assume this distribution looks like? Draw it!

Let **α** = .05, 1- tailed: What value does our sample mean need to "beat" in order for us to conclude that it is higher than the population mean (given random variation present in our sample)? Draw this! What would you conclude if our sample mean is 104? What if it is 108?

Your Turn

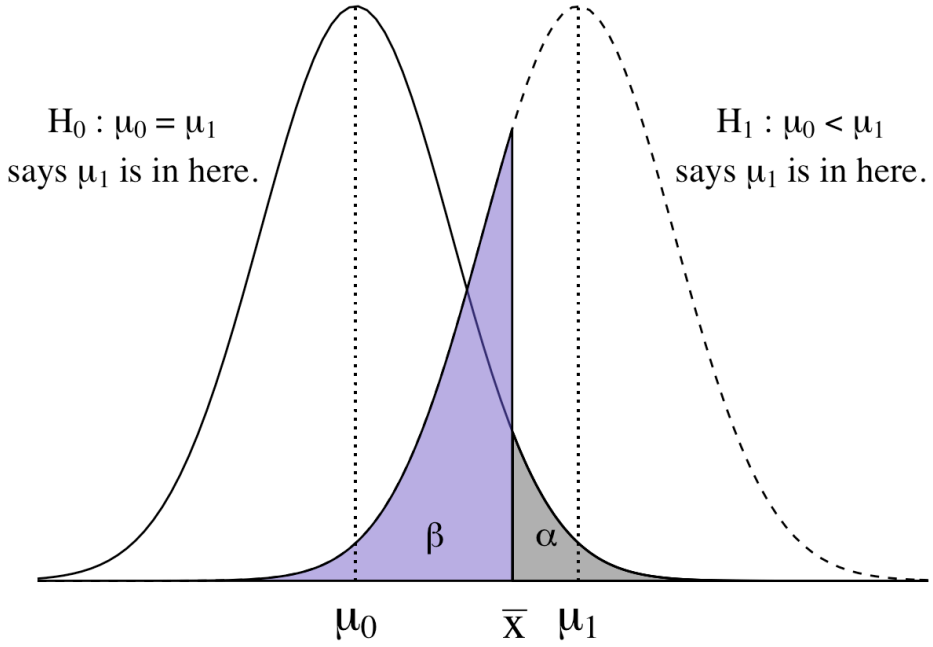
The null distribution for the z-statistic




$H_0 : \mu_0 = \mu_1$
says μ_1 is in here.
$H_1 : \mu_0 < \mu_1$
says μ_1 is in here.
β
α
μ_0
$\overline{x}$
μ_1

Dueling distributions

- In a one-tailed test, we have two dueling hypothetical distributions:
  - The null distribution, centered at μ_0
  - The alternative distribution, centered around some specific or unspecified other mean (higher or lower?) μ_1



A (non-directional) alternative hypothesis

- H_0:

  IQ scores for AA will not differ from population;

$$\mu_{aa} = \mu_0$$

$$\mu_{aa} = 100$$

$$\mu_{aa} - 100 = 0$$

- H_1:

  IQ scores for AA will be different from population

$$\mu_{aa} \neq \mu_0$$

$$\mu_{aa} \neq 100$$

$$\mu_{aa} - 100 \neq 0$$

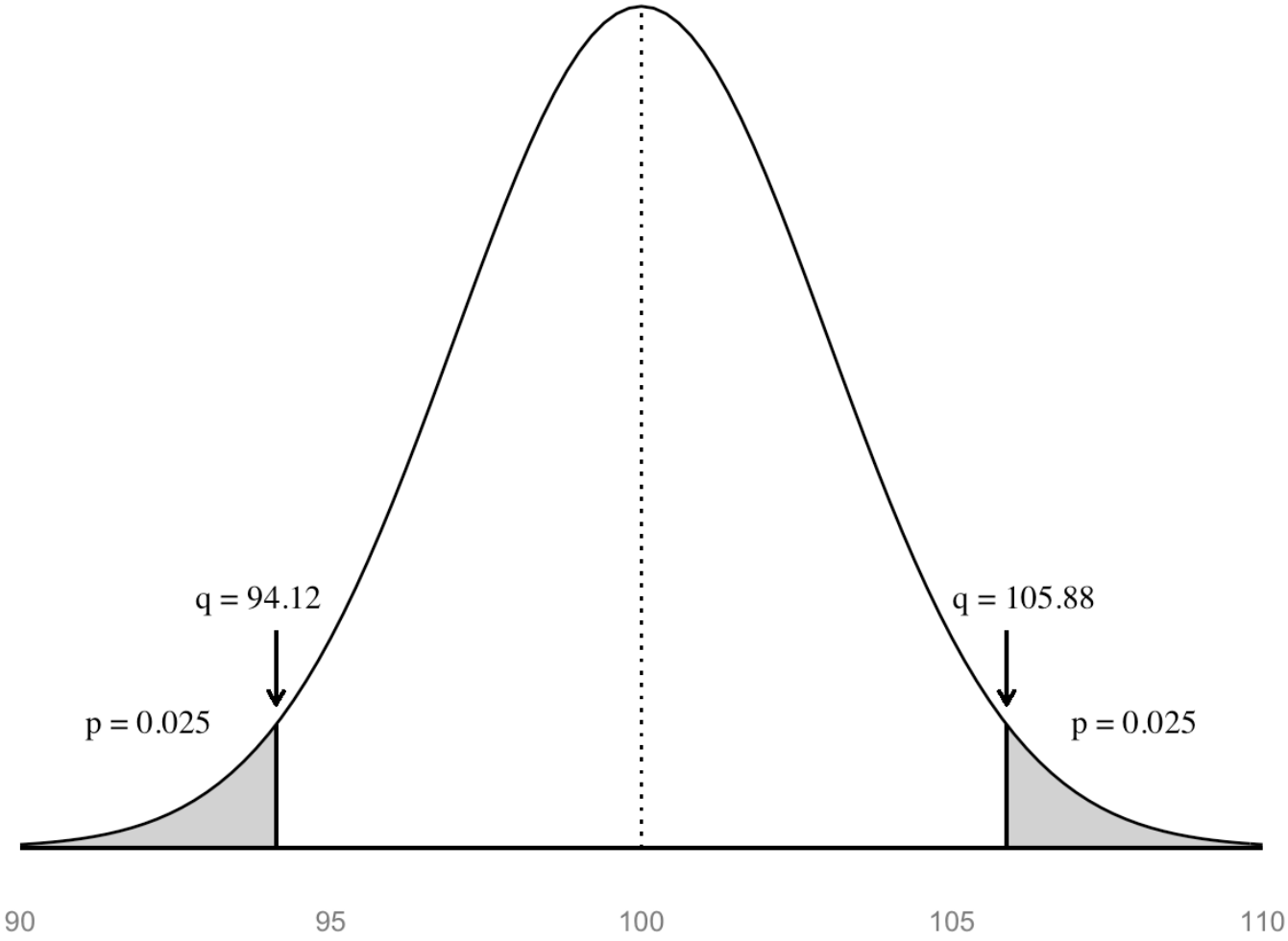
Before we even look at the data:
What is the mean and standard deviation (the "standard error") of the null distribution?

What shall we assume this distribution looks like? Draw it!

Let α = .05, 2-tailed: What value does our sample mean need to "beat" in order for us to conclude that it is different than the population mean (given random variation present in our sample)? Draw this! What would you conclude if our sample mean is 105?

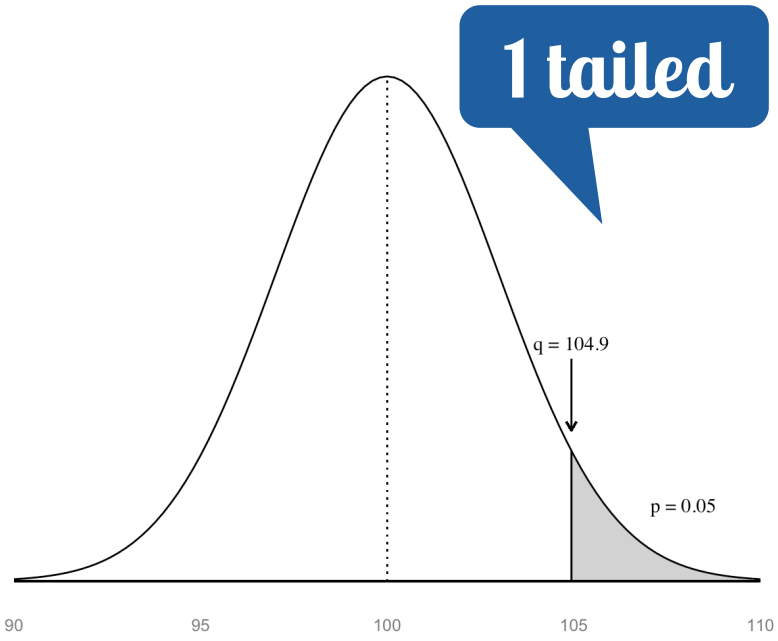
Your Turn

The null distribution for the z-statistic

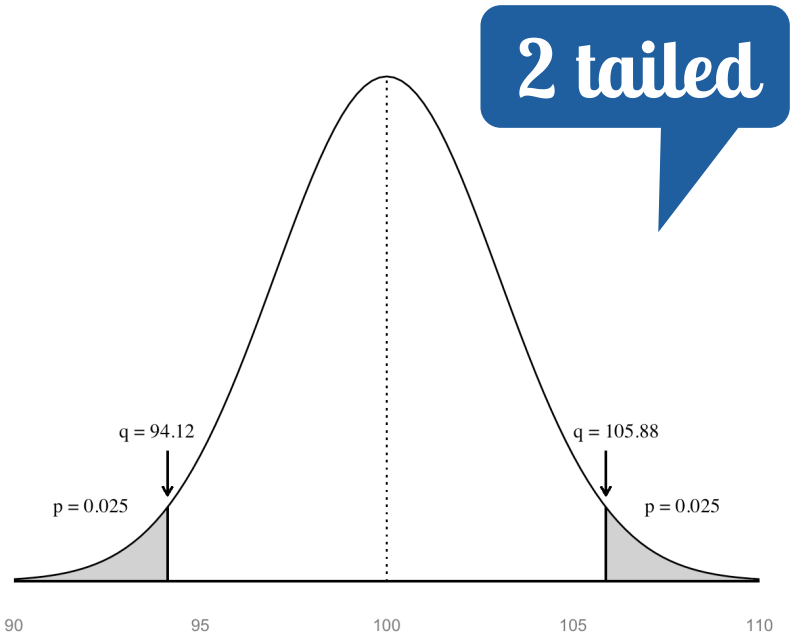


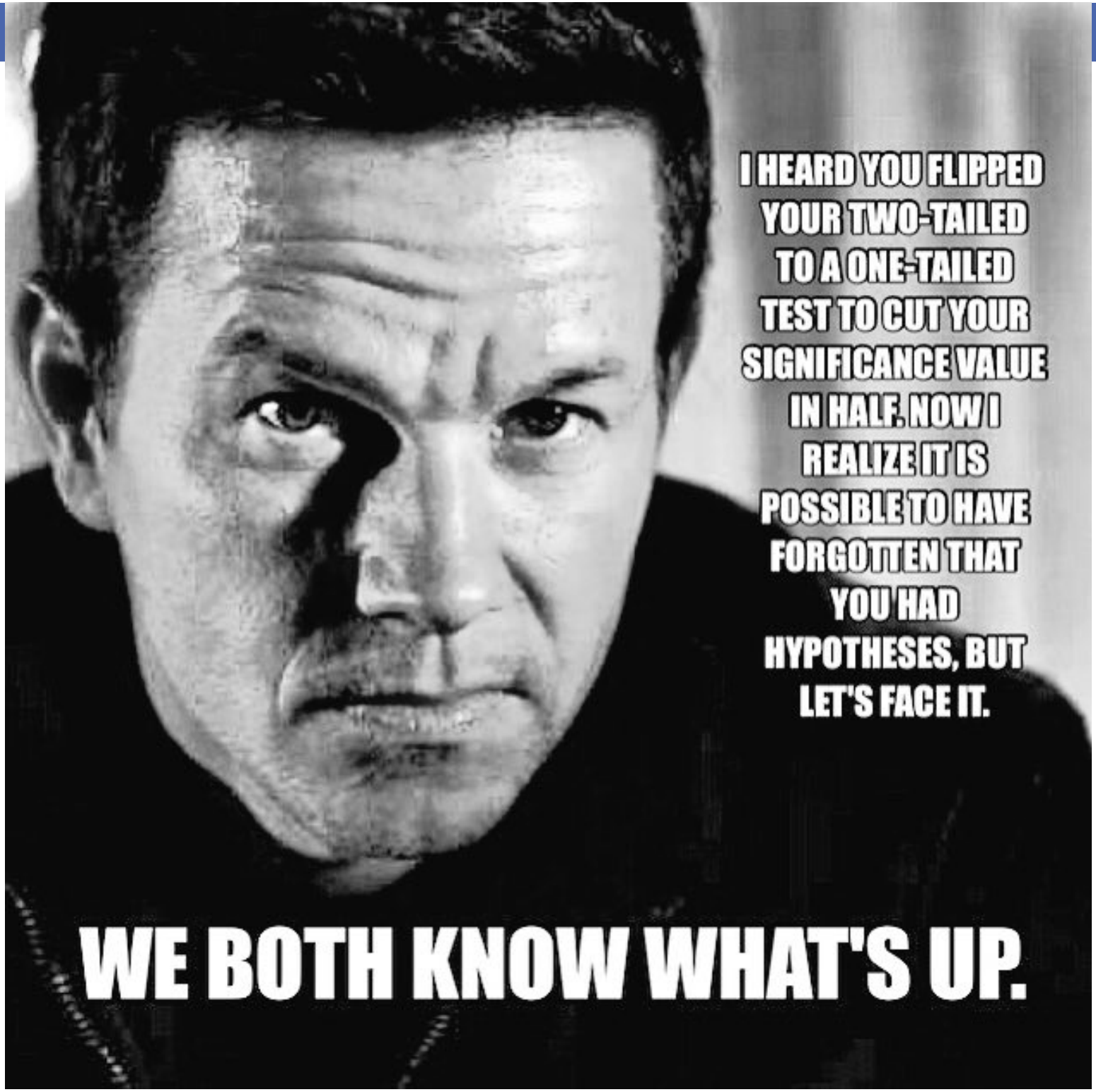
The alternative hypothesis

Directional H_1



Non-directional H_1




I HEARD YOU FLIPPED YOUR TWO-TAILED TO A ONE-TAILED TEST TO CUT YOUR SIGNIFICANCE VALUE IN HALF. NOW I REALIZE IT IS POSSIBLE TO HAVE FORGOTTEN THAT YOU HAD HYPOTHESES, BUT LET'S FACE IT.
WE BOTH KNOW WHAT'S UP.