Parametric hypothesis tests continued

Overview

Continuation of parametric

- The t-test

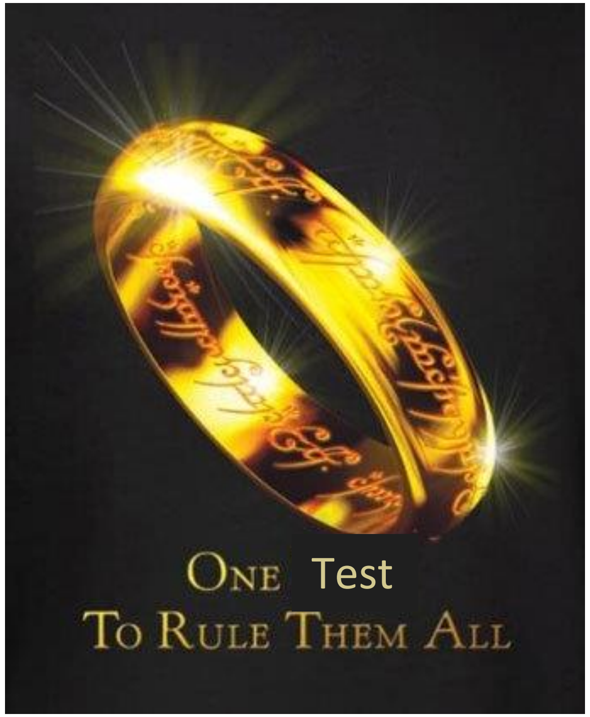
Theories of hypothesis testing

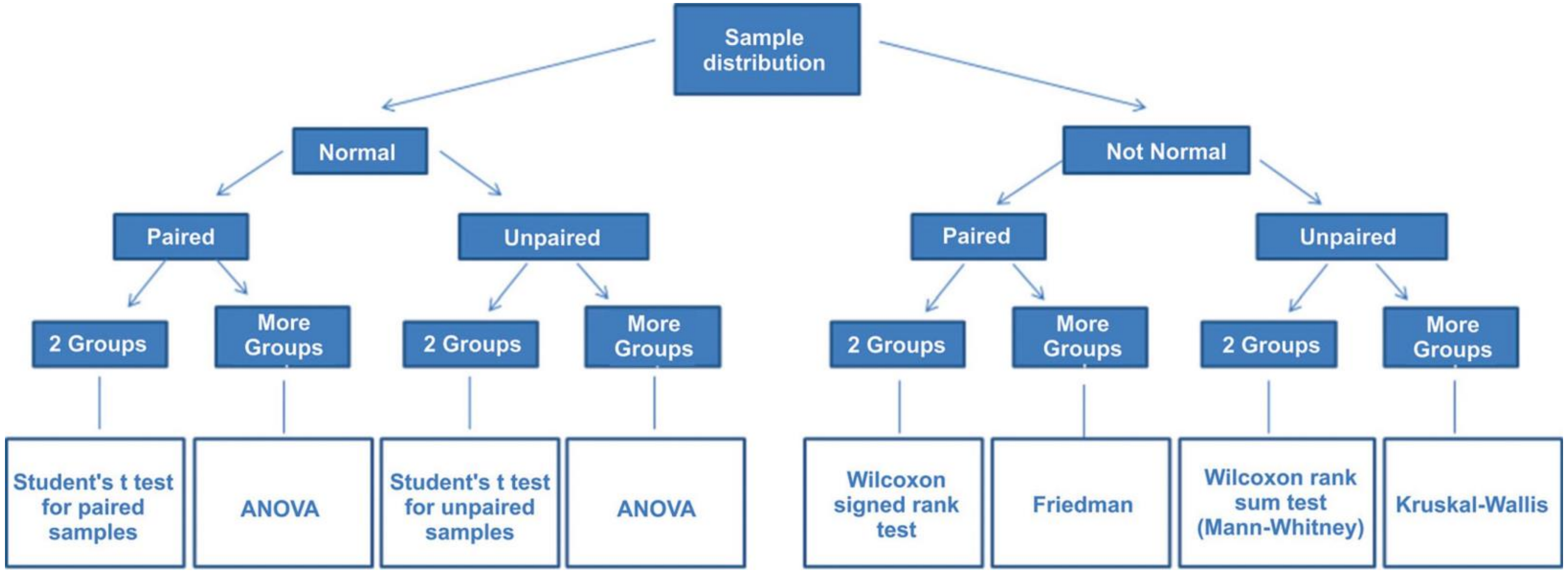
If there is time:

- Hypothesis tests for a single mean
- Connections between hypothesis tests and confidence intervals

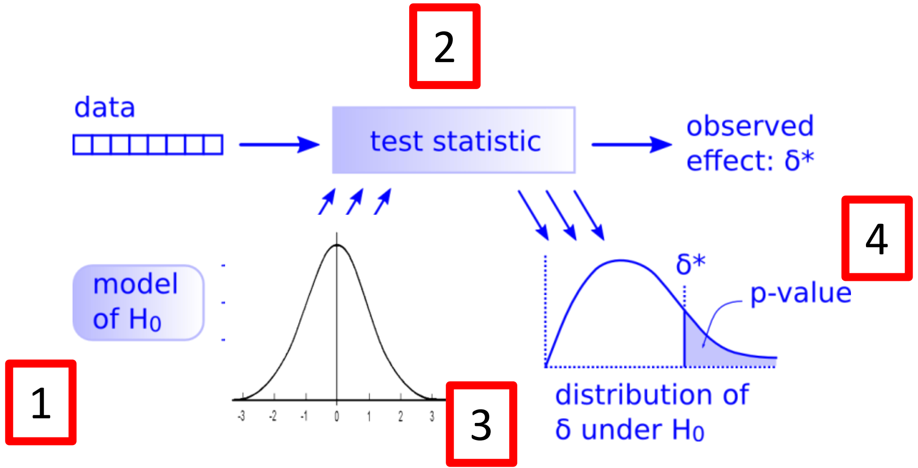
Review and continuation of parametric tests

The big picture: There is only one hypothesis test!





Just need to follow 5 steps!



Permutation/randomization tests

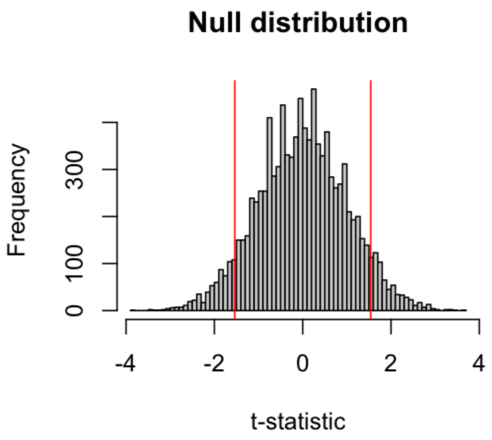
In **permutation/randomization hypothesis tests**, the null distribution is given by randomly shuffling the data

- i.e., in **step 3** of hypothesis testing we use randomization to get a null distribution…

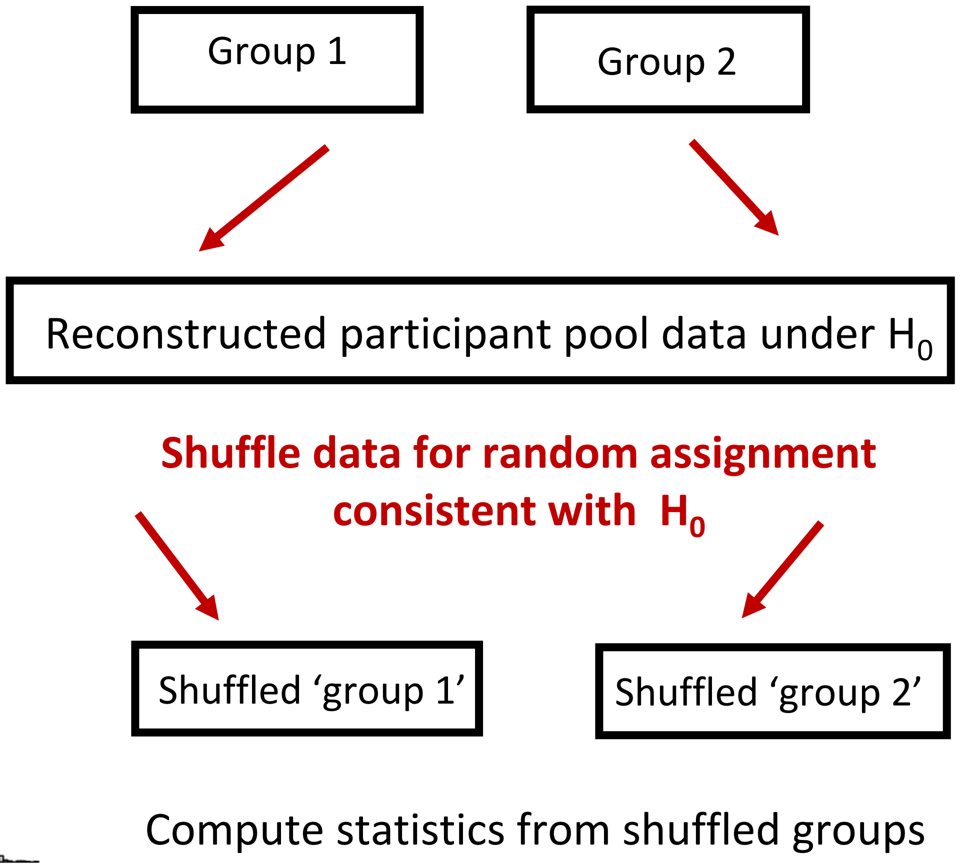
For example: drug study

- **Step 1**: H_0: $\mu_{Gingko} - \mu_{Placebo} = 0$
- **Step 2**:

$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$



Step 3



Parametric hypothesis tests

In **parametric hypothesis tests**, the null distribution is given by a density function

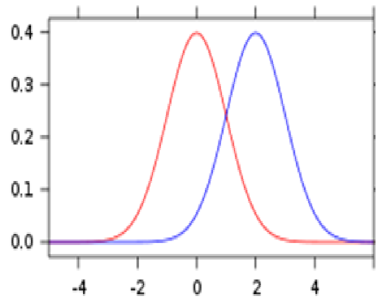
- i.e., in **step 3** of hypothesis testing we use parametric distribution as the null distribution

These density functions have a finite set of ***parameters*** that control the shape of these functions

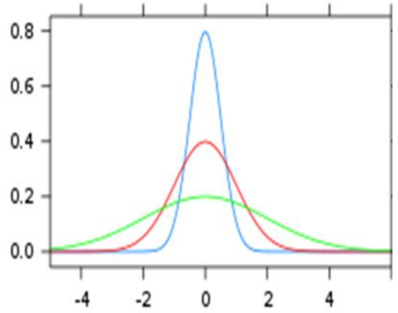
- Hence the name "parametric hypothesis tests"
- Example: the normal density function has two parameters: μ and ***σ***

Density curves…

Changing μ



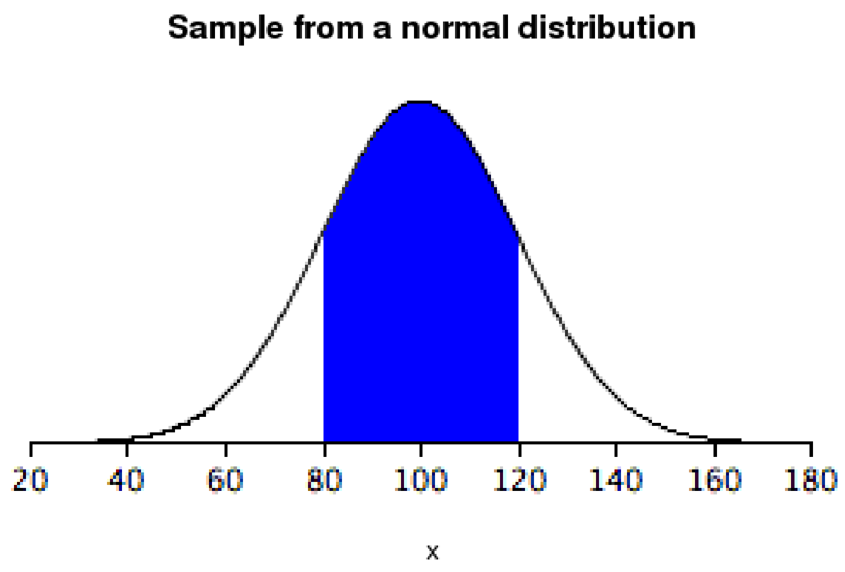
Changing σ



Quick review: probability functions

To **generate random data** we use functions that start with the letter ***r***

```
> rand_data <- rnorm(100)
> hist(rand_data)
```



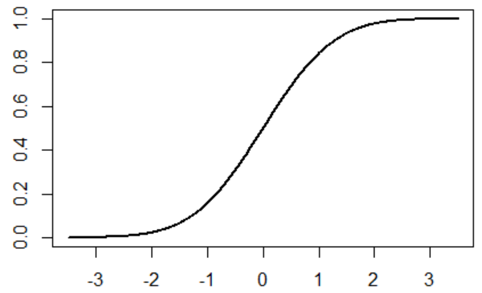
$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To **plot probability density functions** we use functions that start with the letter ***d***

```
> x <- seq(-3, 3, by = .001)
> y <- dnorm(x)
> plot(x, y, type = "l")
```

To **get the probability** that that a random number X is less than x, P(X ≤ x), we use functions that start with ***p***

```
> pnorm(2)
```



$$P(X \leq x)$$

$$= \int_{-\infty}^{x} f(x)dx$$

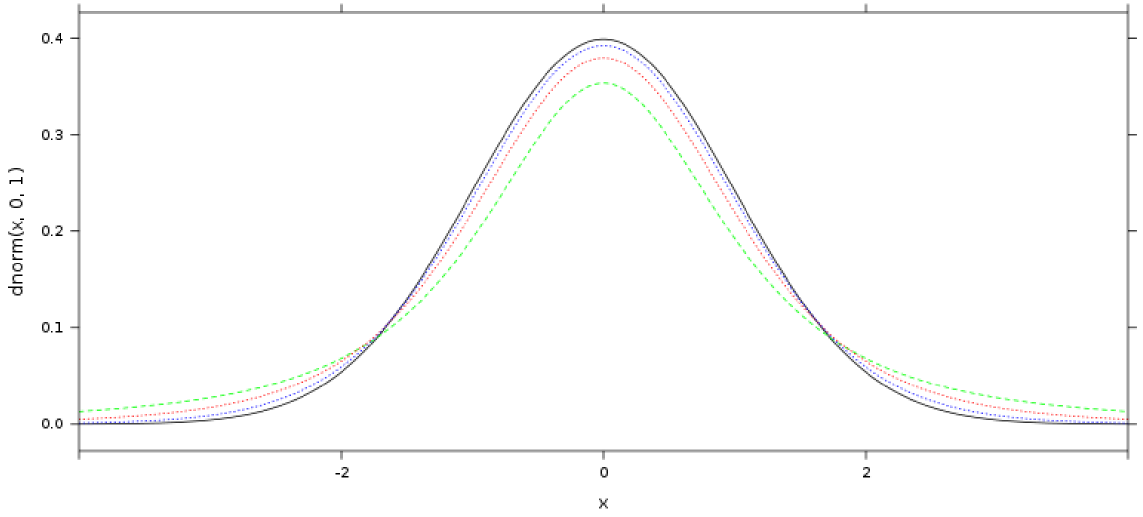
Parametric hypothesis tests

t-distributions

A commonly used density function (distribution) used for statistical inference is the t-distribution

- In R: rt(), dt(), pt() and qt()

t-distributions have one parameter called “degrees of freedom”



df = 2

df = 5

df = 15

N(0, 1)

t-distributions

When using t-distributions for statistical inference, each point in our t-distribution is a t-statistic

- i.e., we use t-distributions as null distributions for hypothesis tests and as sampling distributions when creating confidence intervals

t-statistics are a ratio of:

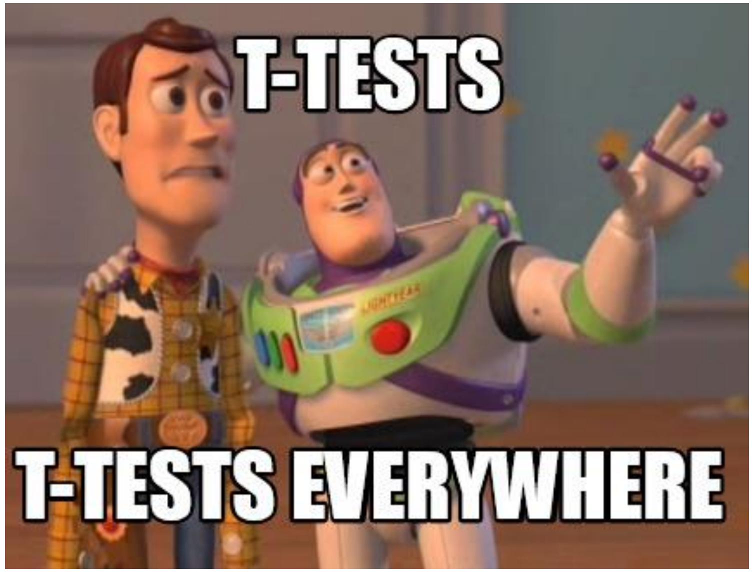
- The departure of an estimated value from a hypothesized parameter value
- Divided by an estimate of the standard error

$$t = \frac{estimate\ \text{-}\ param_0}{\hat{SE}}$$

If the SE did not involve in intermediate estimate (i.e., $\bar{x}$) this would be a "z-statistic" that comes from a standard normal distribution

t-tests

t-tests are parametric hypothesis tests where the null distribution is a density function called a t-distribution



t-tests can be used to test:

- If a mean is equal to a particular value: H_0: $\mu = 7$
- If two means are equal: H_0: $\mu_t = \mu_c$
- If a regression coefficient is equal to a particular value: H_0: $\beta = 2$
- etc.

t-tests for comparing two means

Let's examine t-tests for comparing **two means**

Step 1: what is the null hypotheses?

- H_0: $\mu_t - \mu_c = 0$

Step 2a: What is the numerator of the t-statistic?

$$t = \frac{\underbrace{estimate}_{(\bar{x}_t - \bar{x}_c)} - \underbrace{param_0}_{0}}{\hat{SE}} \longleftarrow = \frac{(\bar{x}_t - \bar{x}_c) - 0}{\hat{SE}} = \frac{\bar{x}_t - \bar{x}_c}{\hat{SE}}$$

t-tests for comparing two means

Step 2b: What is the denominator of the t-statistic?

$$t = \frac{stat\ \text{-}\ param_0}{\hat{SE}}$$

Students' t-test assumes the variance in each population is the same, and uses an SE estimate of:

$$\hat{SE}_{\bar{x}_t\text{-}\bar{x}_c} = s_p \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}} \qquad s_p = \sqrt{\frac{\sum_i^{n_t}(x_i - \bar{x}_t)^2 + \sum_j^{n_c}(x_j - \bar{x}_c)^2}{n_t + n_c\ \text{-}\ 2}}$$

Welch's t-test does **not** assume that the variance in each population is the same and uses an estimate of:

$$\hat{SE}_{\bar{x}_t\text{-}\bar{x}_c} = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$$

t-tests for comparing two means

Step 2b: What is the denominator of the t-statistic?

$$t = \frac{stat \text{ - } param_0}{\hat{SE}}$$

Students' t-test assumes the variance in each population is the same, and uses an SE estimate of:

$$t = \frac{\bar{x}_t \text{ - } \bar{x}_c}{s_p \cdot \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

$$s_p = \sqrt{\frac{\sum_i^{n_t}(x_i - \bar{x}_t)^2 + \sum_j^{n_c}(x_j - \bar{x}_c)^2}{n_t + n_c \text{ - } 2}}$$

Welch's t-test does **not** assume that the variance in each population is the same and uses an estimate of:

$$\hat{SE}_{\bar{x}_t \text{-} \bar{x}_c} = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$$

$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

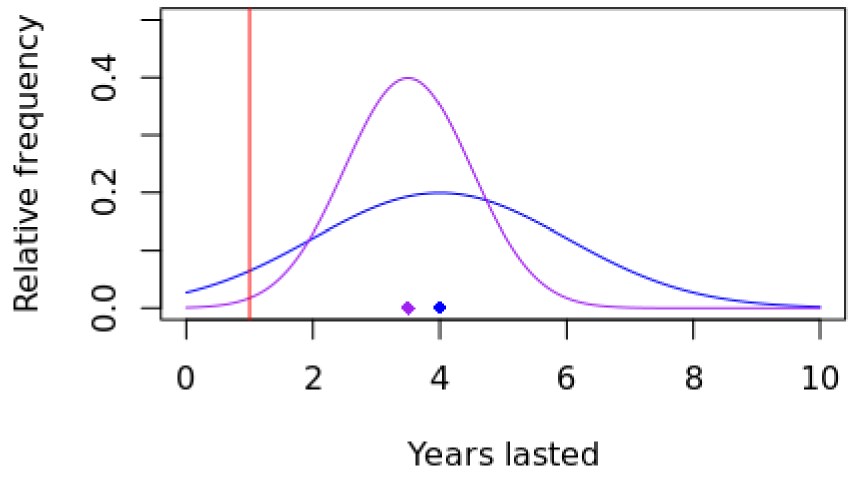
Side note: t-tests for comparing two means

Question: which statistic/test is better to use?

A: generally better to choose the "robust" test

- i.e., Welch's t-test is robust to unequal variances, so generally a better choice

However, we need to be careful with the decisions we make based on differences of means when there are unequal variances



E.g., Which car battery company produces better batteries in terms of how long they last?

- Company A: μ = 4 years, σ = 2 years
- Company B: μ = 3.5 years, σ = 1 years

- Company A: 7% fail within a year
- Company B: 0.6% fail with a year

Example: Does Gingko improve memory?

A double-blind randomized controlled experiment by Solomon et al (2002) investigated whether taking a Ginkgo supplement could improve memory

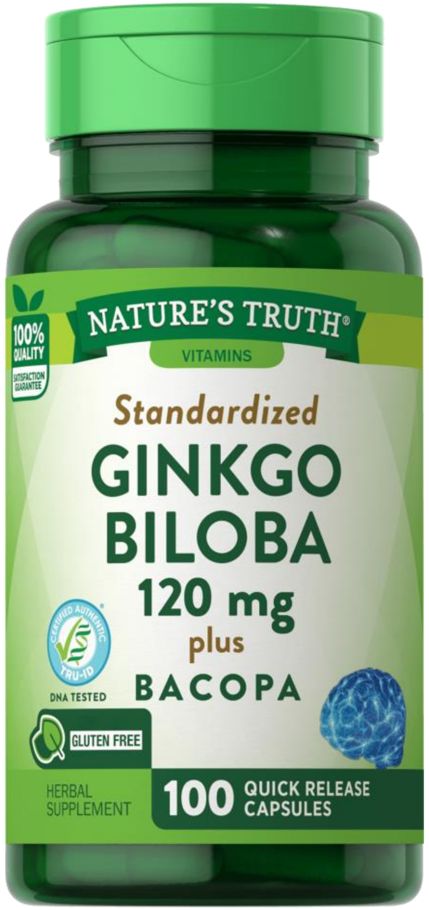
Let's try using a t-statistic!

- t = -1.53

$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

3. What is the null distribution?

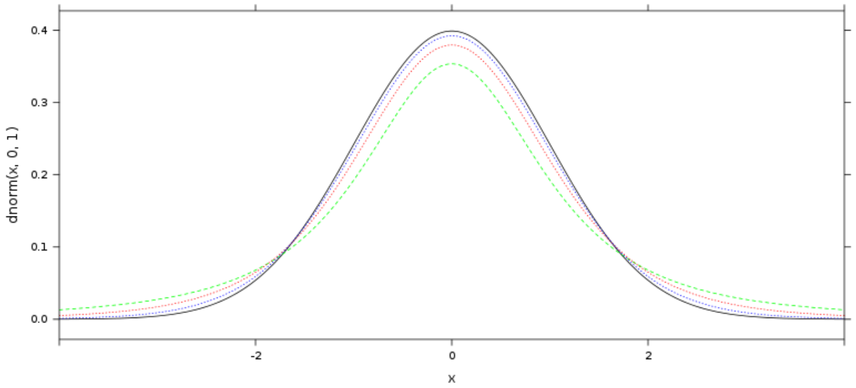
- What additional piece of information do we need to create it?



t-tests for comparing two means

When using a t-distribution to compare two means, a conservative estimate of the degrees of freedom is the minimum of the two samples sizes, n_t and n_c, minus 1

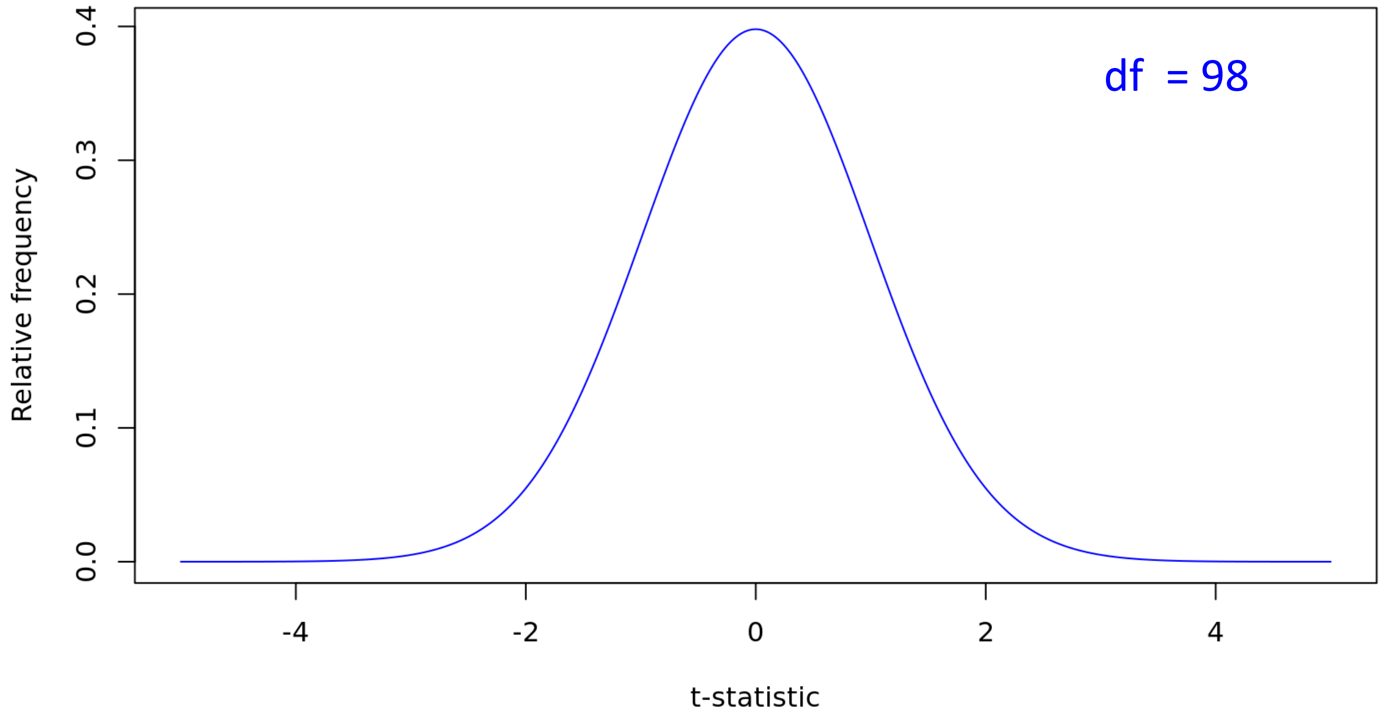
- df = $\min(n_t, n_c) - 1$



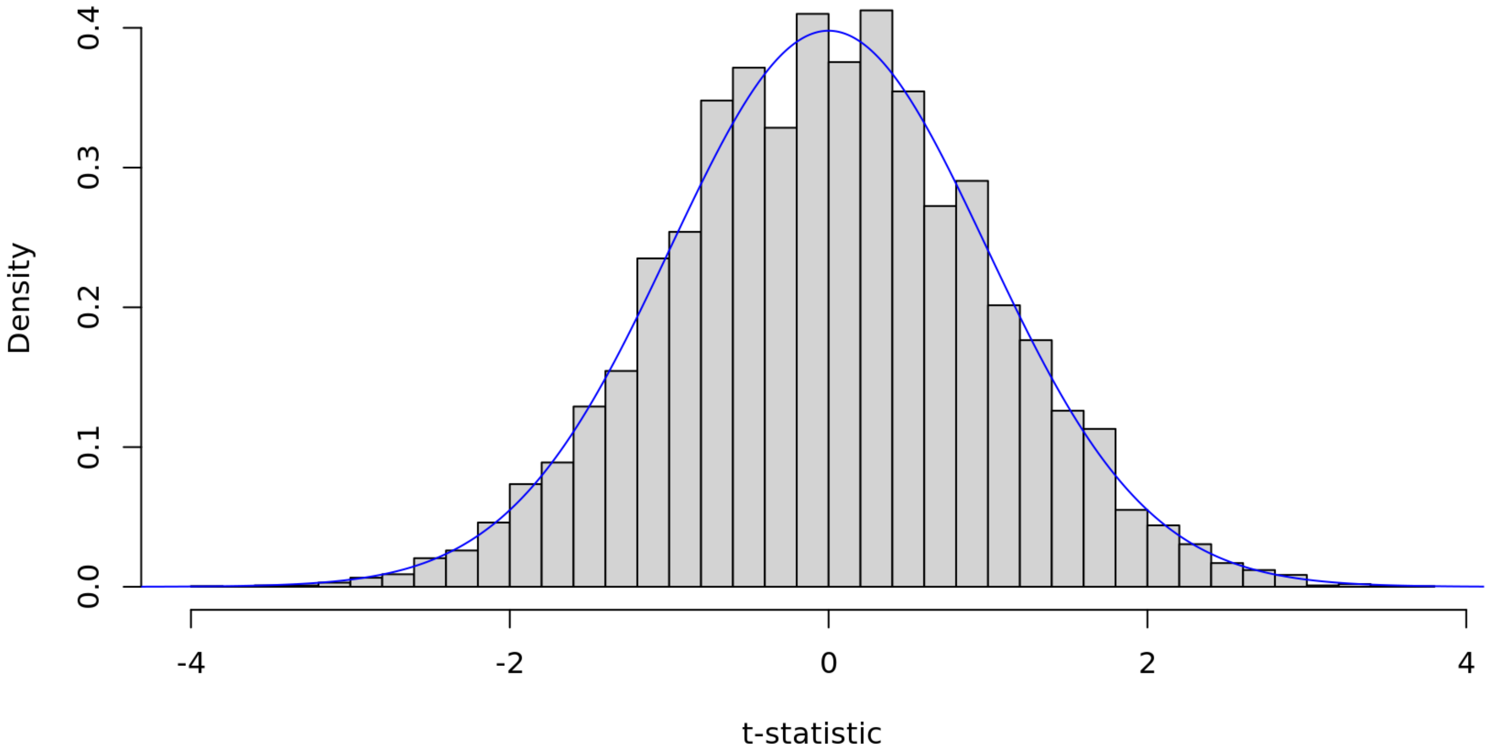
Q: For the Gingko study we had 104 people in the treatment group and 99 people in the control group so the degrees of freedom parameter is?

- 98

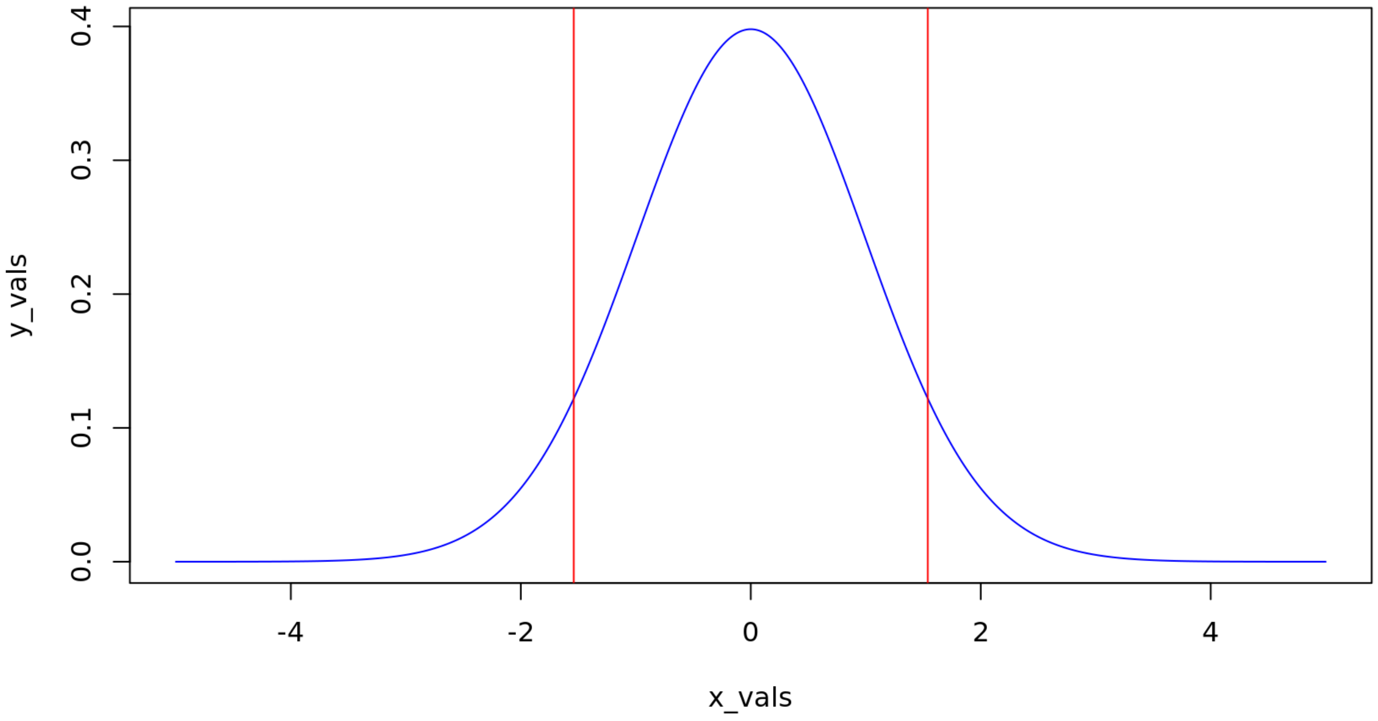
Step 3: Null t-distribution



Step 3: parametric vs. randomization distributions

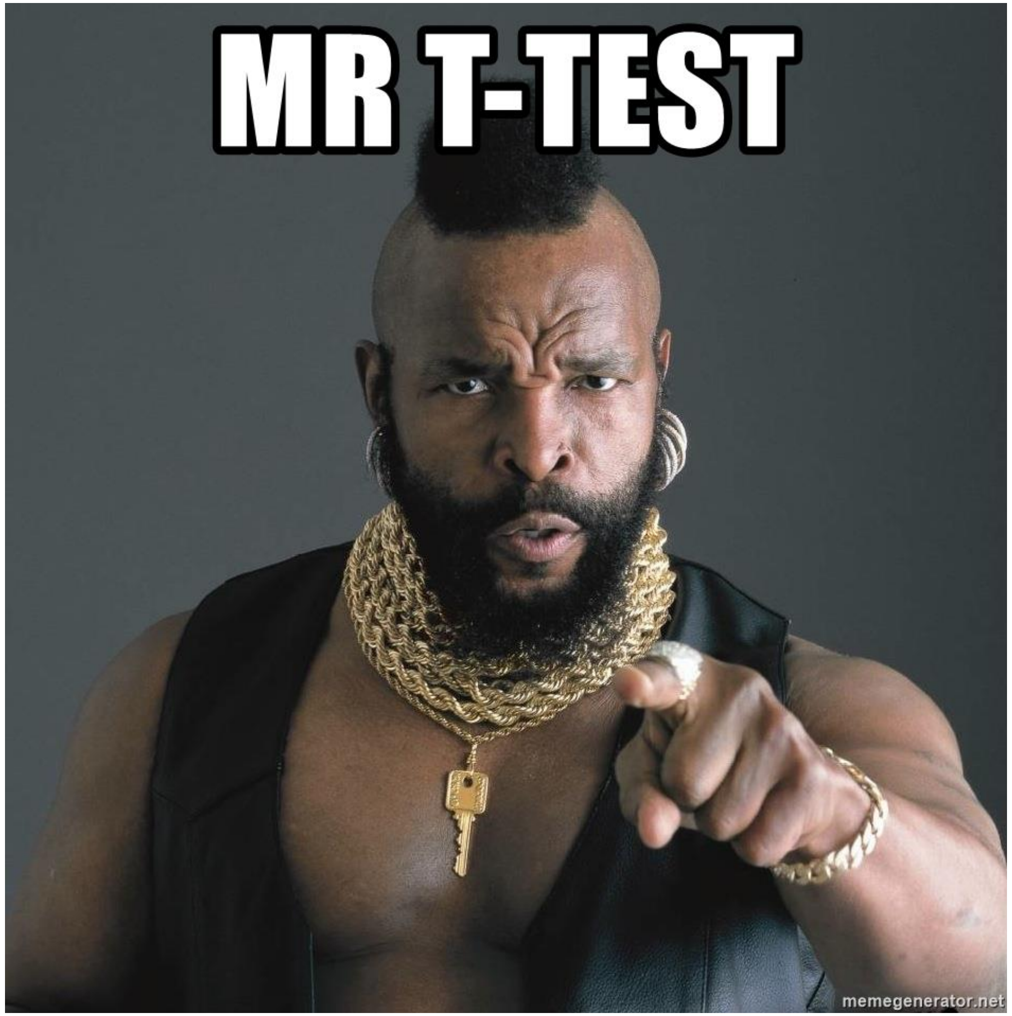


Step 4-5: p-value and conclusion



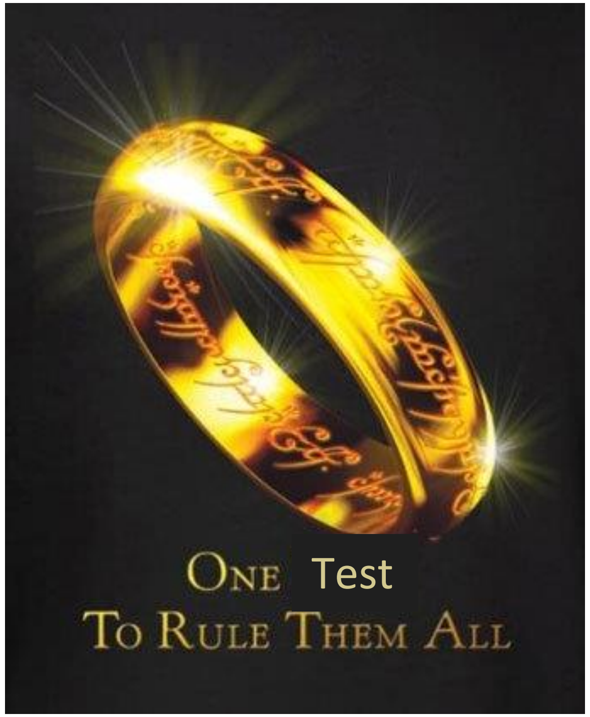
p-value = 0.127

Conclusion?

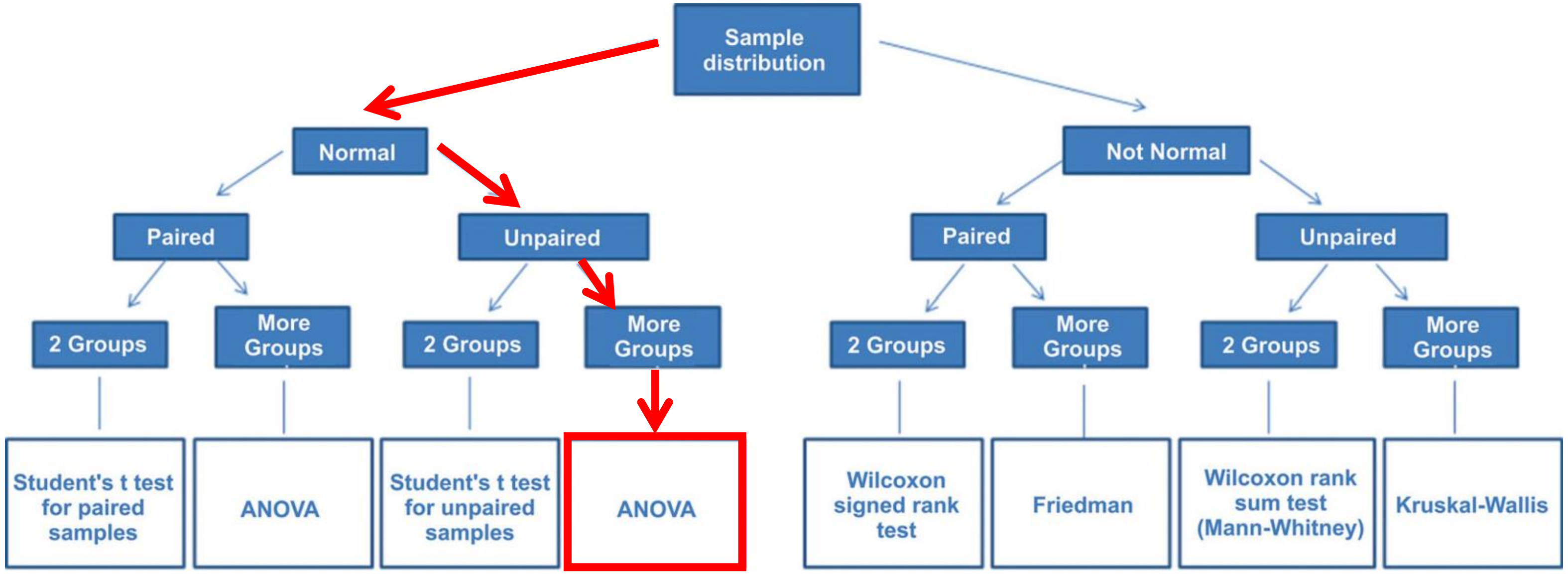


I pity the fool who doesn't want to try it in R!

Other parametric hypothesis test



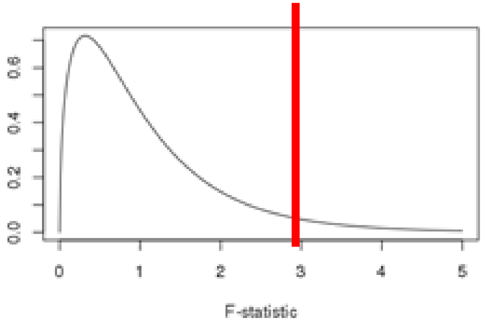
The hypothesis test zoo



We can run a large number of additional hypothesis tests by following the 5 steps!

ANOVA: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

$$F = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

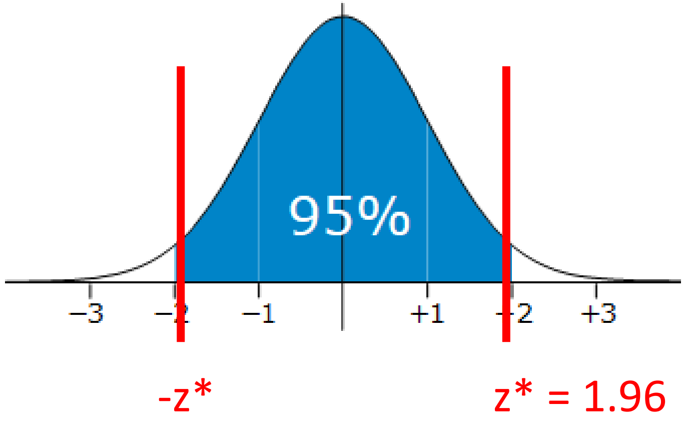


Confidence interval for the difference of two means

Confidence intervals for the bootstrap had the form:

$$CI_{95} \approx \text{stat} \pm 2 \cdot SE^*$$

qnorm(.975) = 1.96



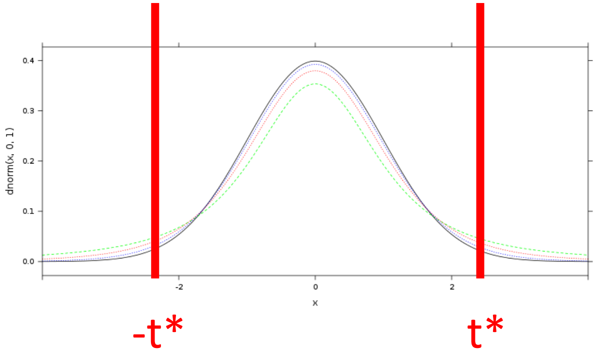
When creating confidence intervals based on a t-distribution we use:

$$CI_{95} \approx \text{statistic} \pm t^* \cdot \hat{SE}$$

df = min(n_t, n_c) - 1

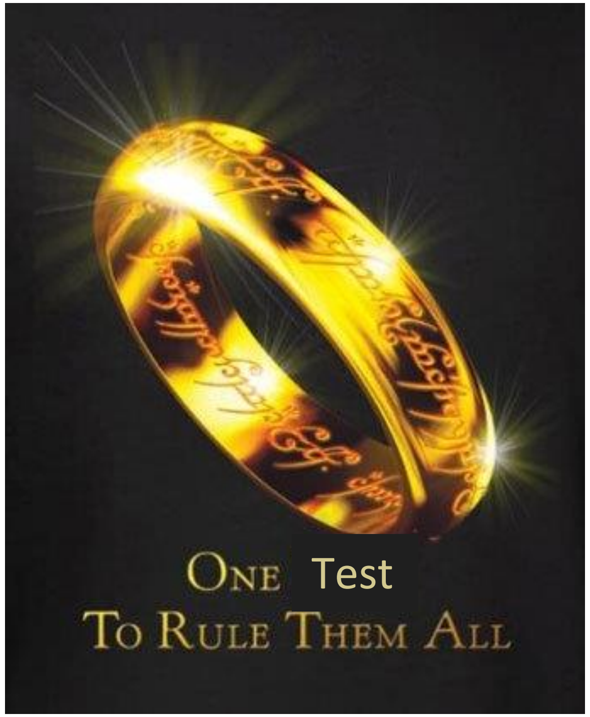
qt(.975, df)

$$\hat{SE}_{\bar{x}_t - \bar{x}_c} = \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$$

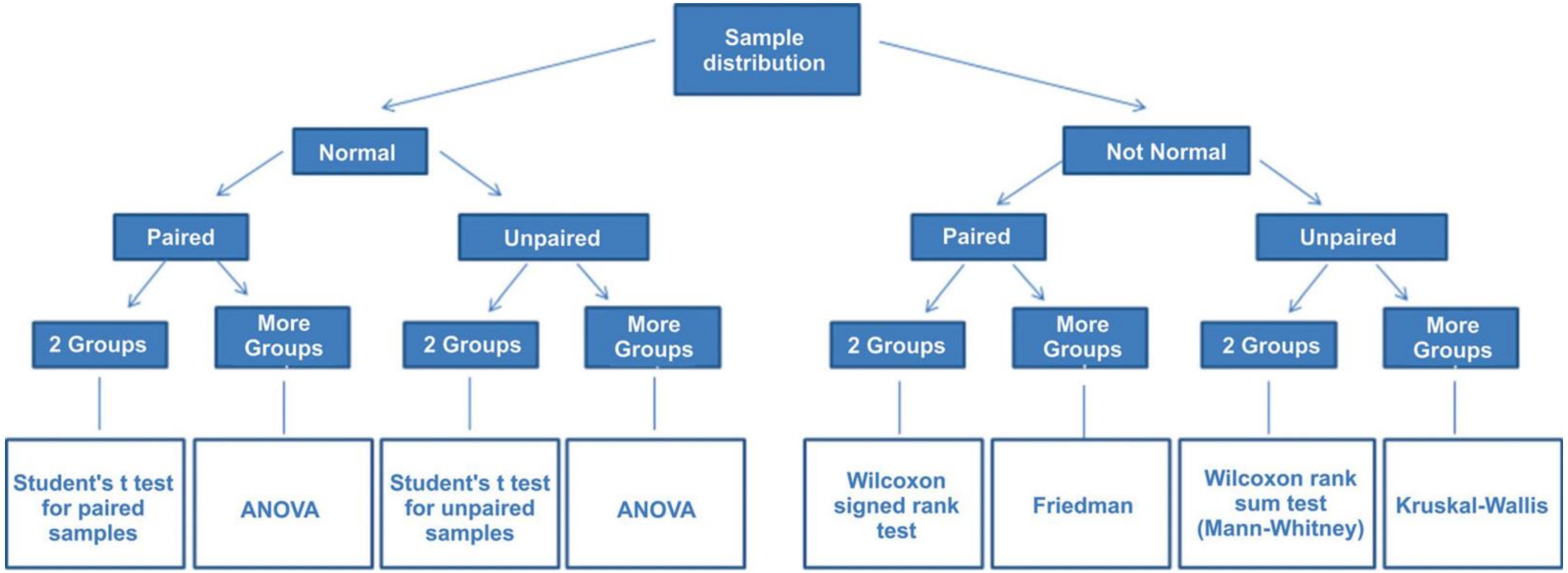


For a difference of means: $CI = (\bar{x}_t - \bar{x}_c) \pm t^* \cdot \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$

The big picture: There is only one hypothesis test!



The hypothesis test zoo



We can run a large number of additional hypothesis tests by following the 5 steps!

Nonparametric hypothesis tests

Brief mention: nonparametric hypothesis tests

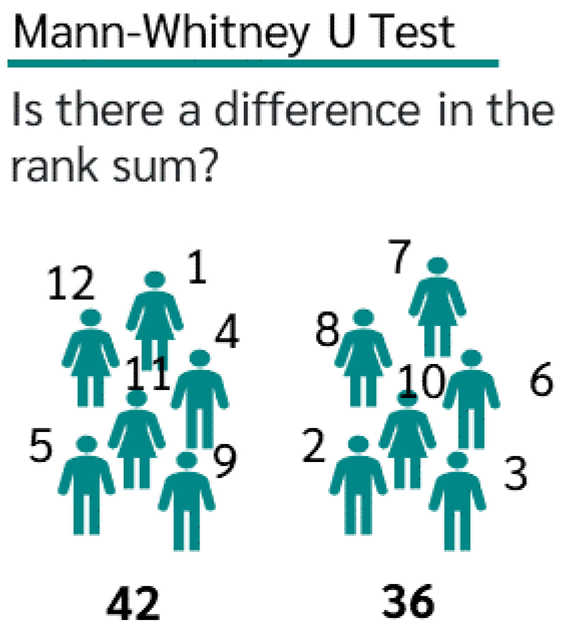
Nonparametric hypothesis tests use null distributions that do not have a small fixed set of parameters

Most nonparametric tests are based on converting the data to ranks

- E.g., Mann-Whitney U test/Wilcoxon rank-sum test
  - Tests whether the probability of X being greater than Y is equal to the probability of Y being greater than X.
    - (where X and Y come from two populations)

Nonparametric tests have fewer assumptions than parametric tests so they are potentially more robust

- e.g., they do not assume the data comes from a normal distribution, they are resistant to outliers, etc.

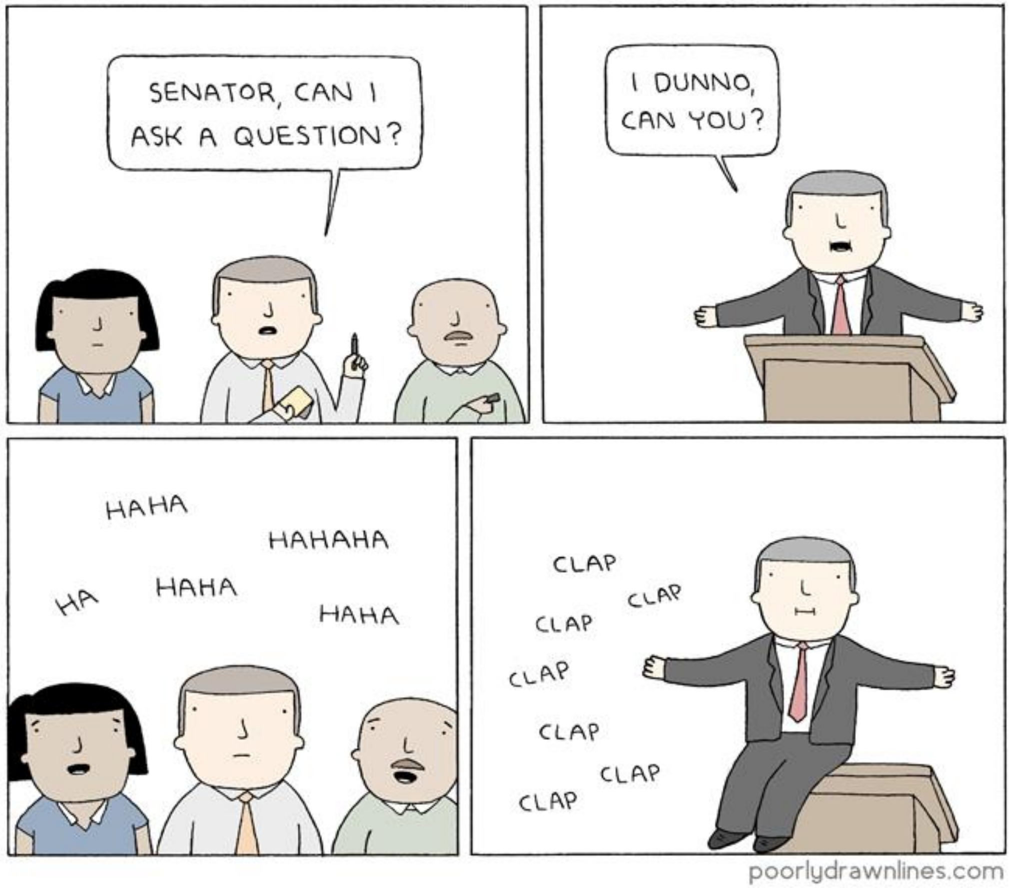


Questions?

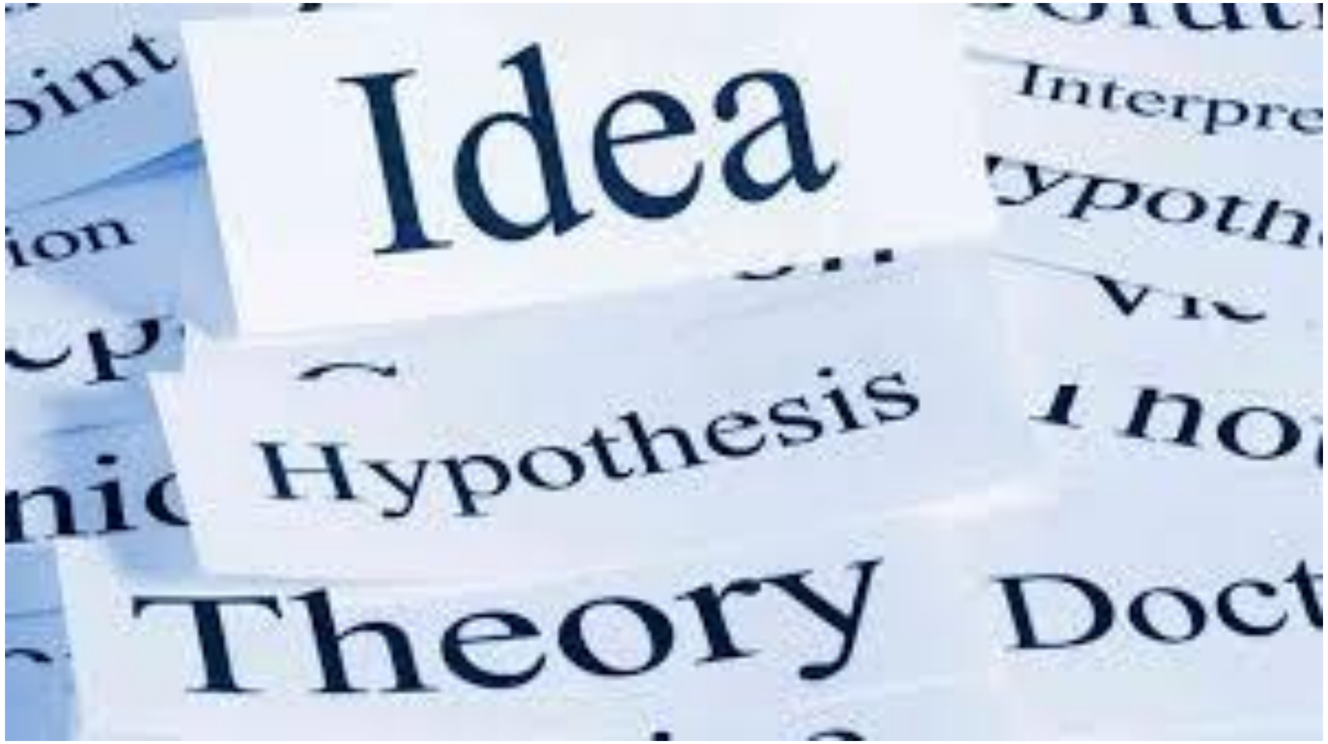
Question: When running a hypothesis test, is it better to…

1. Report the actual p-value

2. Just report if we reject/fail to reject the null hypothesis at the $\alpha = 0.05$ significance level?



Theories of hypothesis tests



Two theories of hypothesis testing

Null-hypothesis significance testing (NHST) is a hybrid of two theories:

1. Significance testing of Ronald Fisher

2. Hypothesis testing of Jezy Neyman and Egon Pearson

Fisher (1890-1962)

Neyman (1894-1981)

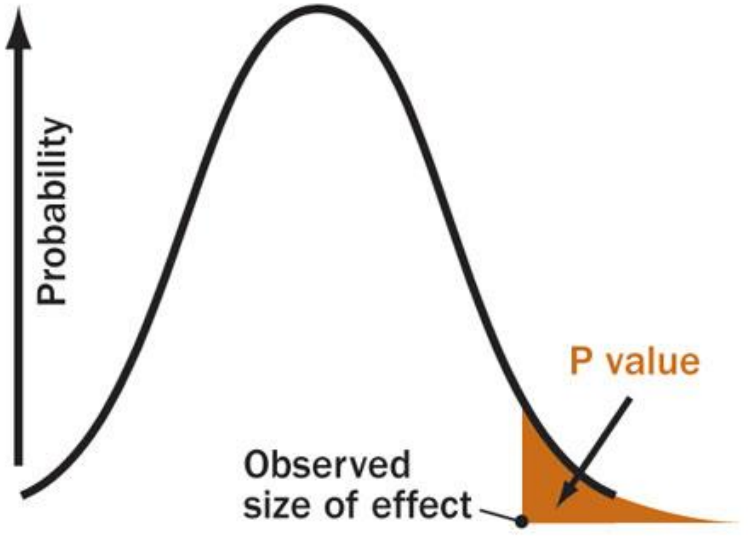
Pearson (1895-1980)

Ronald Fisher's significance testing

Views the p-value as strength of evidence against the null hypothesis

- p-values part of an on-going scientific process:

They tell the experimenter "what results to ignore"



Neyman-Pearson null hypothesis testing

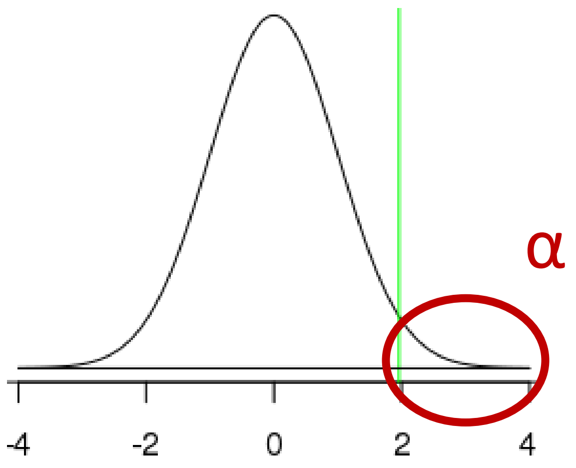
Makes ***a formal decision*** in statistical tests

Reject H_0: if the observed sample statistic is beyond a fixed value

- i.e., reject H_0 if the p-value is less than some predetermined **significance level** α

Do not reject H_0: if the observed sample statistic is not beyond a fixed value. This means the test is inconclusive.

Null distribution



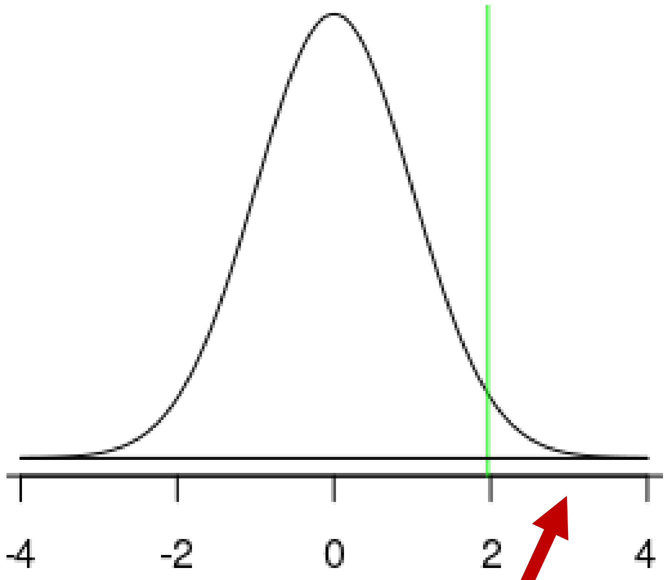
Neyman-Pearson frequentist logic

Type I error: incorrectly rejecting the null hypothesis when it is true

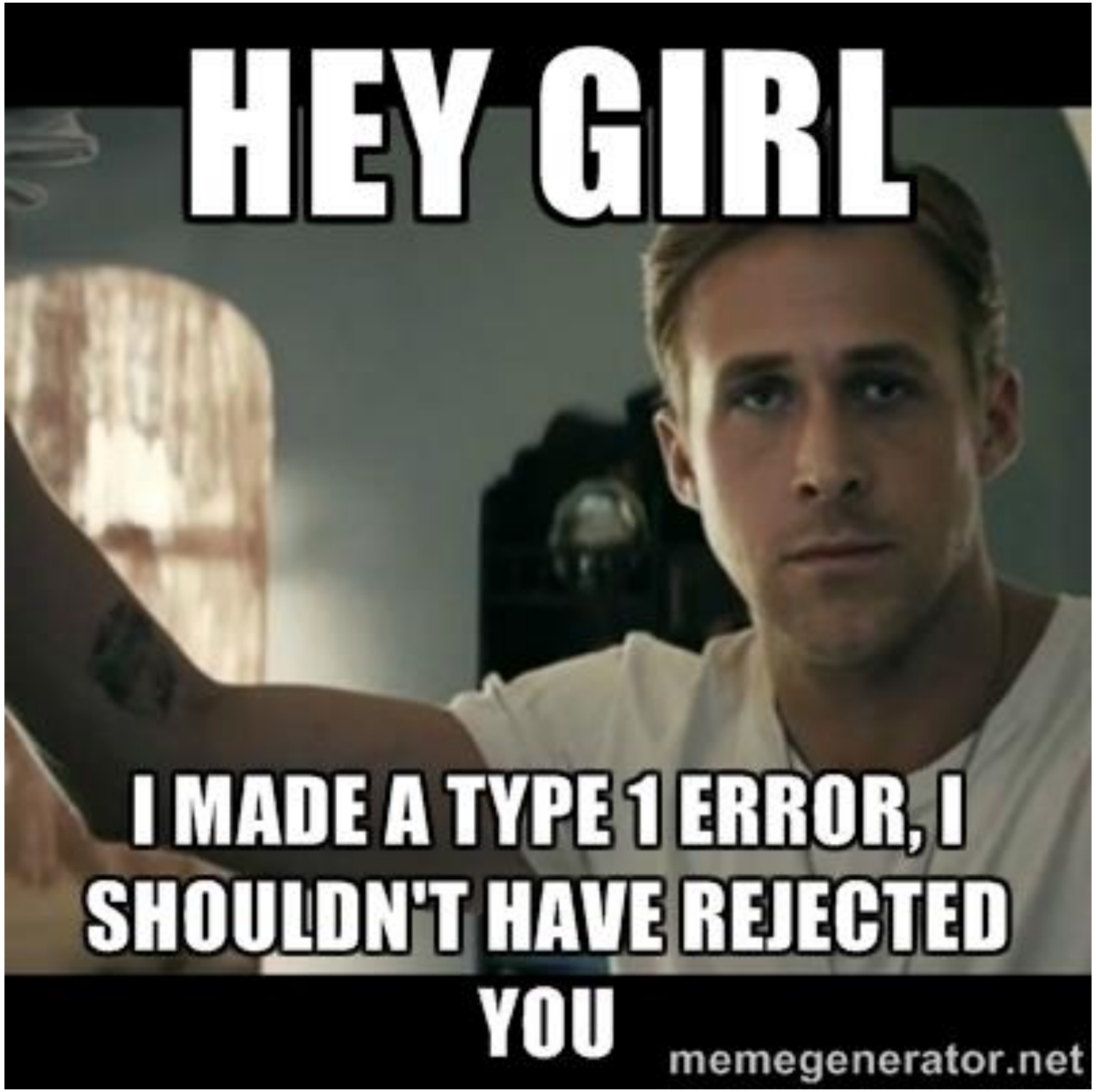
If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, and we were in a world where the null hypothesis was always true, then only ~5% of the time would we falsely report an effect (for $\alpha = 0.05$)

- i.e., we would only make type I errors 5% of the time

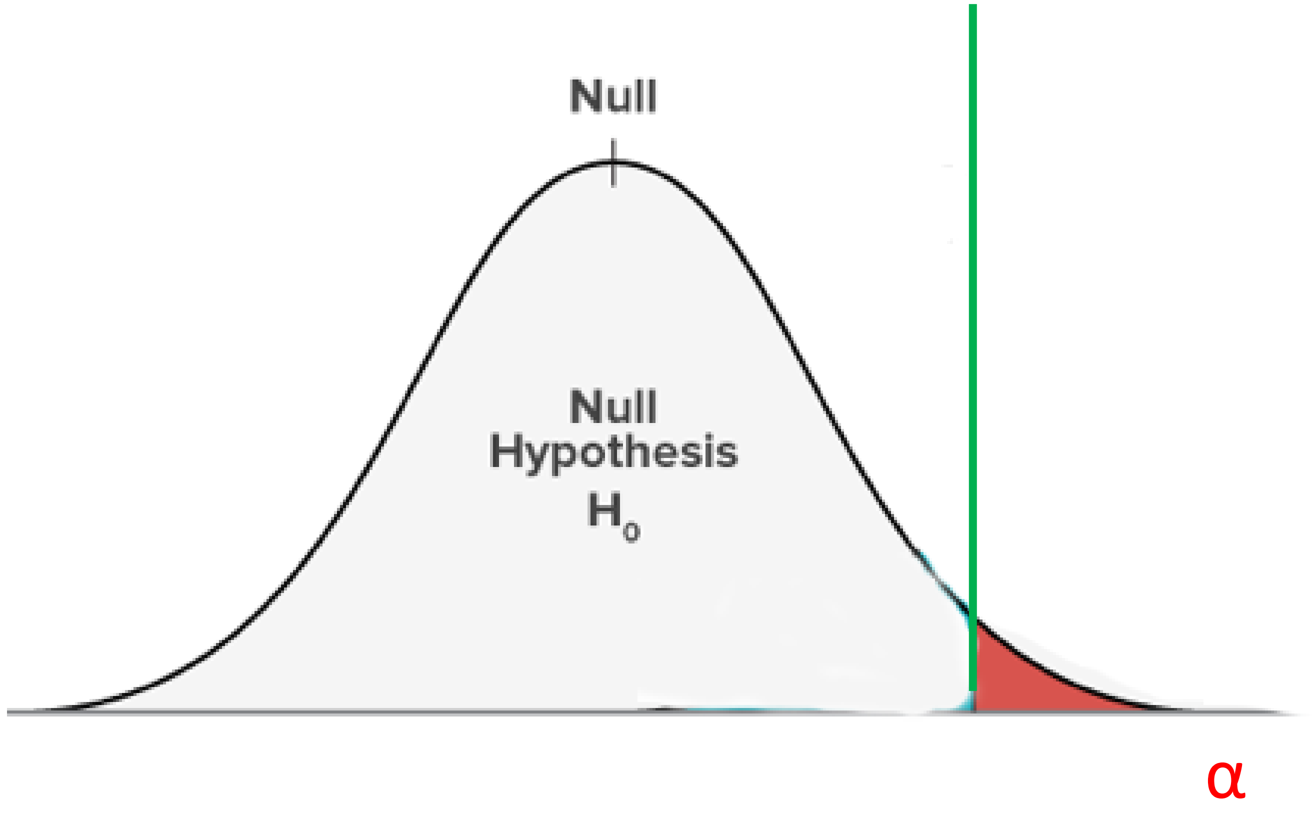
Null distribution



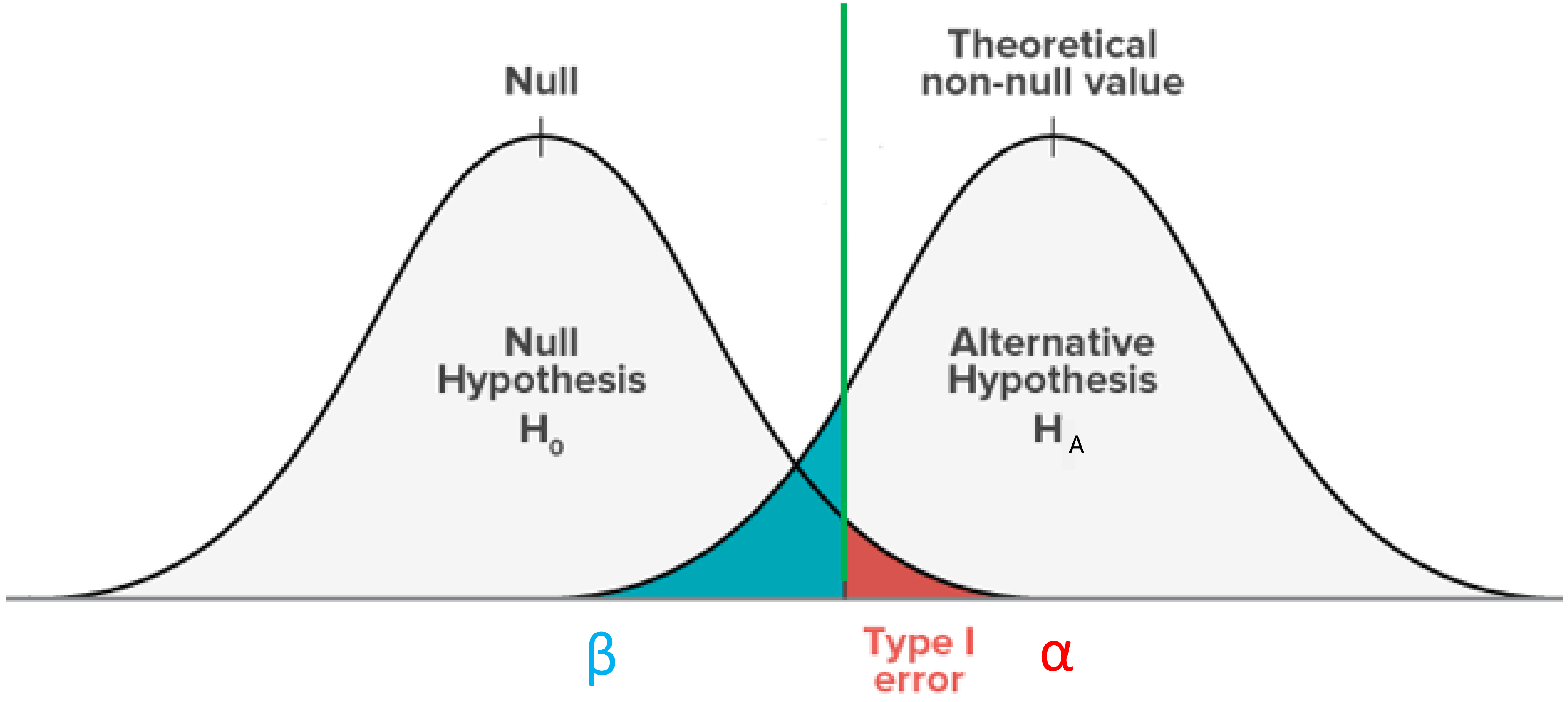
The null distribution is true but statistic landed here


HEY GIRL
I MADE A TYPE 1 ERROR, I SHOULDN'T HAVE REJECTED YOU
memegenerator.net

Neyman-Pearson Frequentist logic



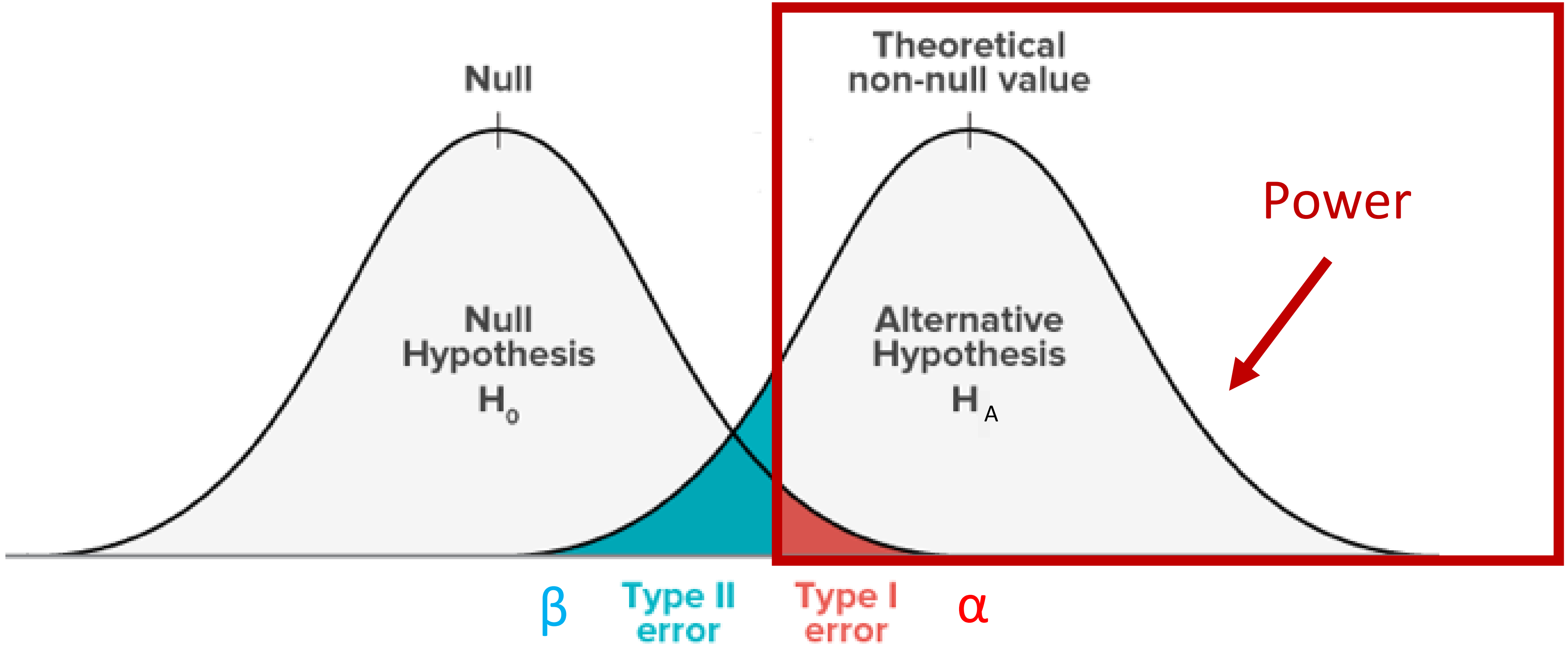
Neyman-Pearson Frequentist logic



Type II error: incorrectly rejecting failing to reject H_0 when it is false

- The rate at which we make type II errors is often denoted with the symbol β

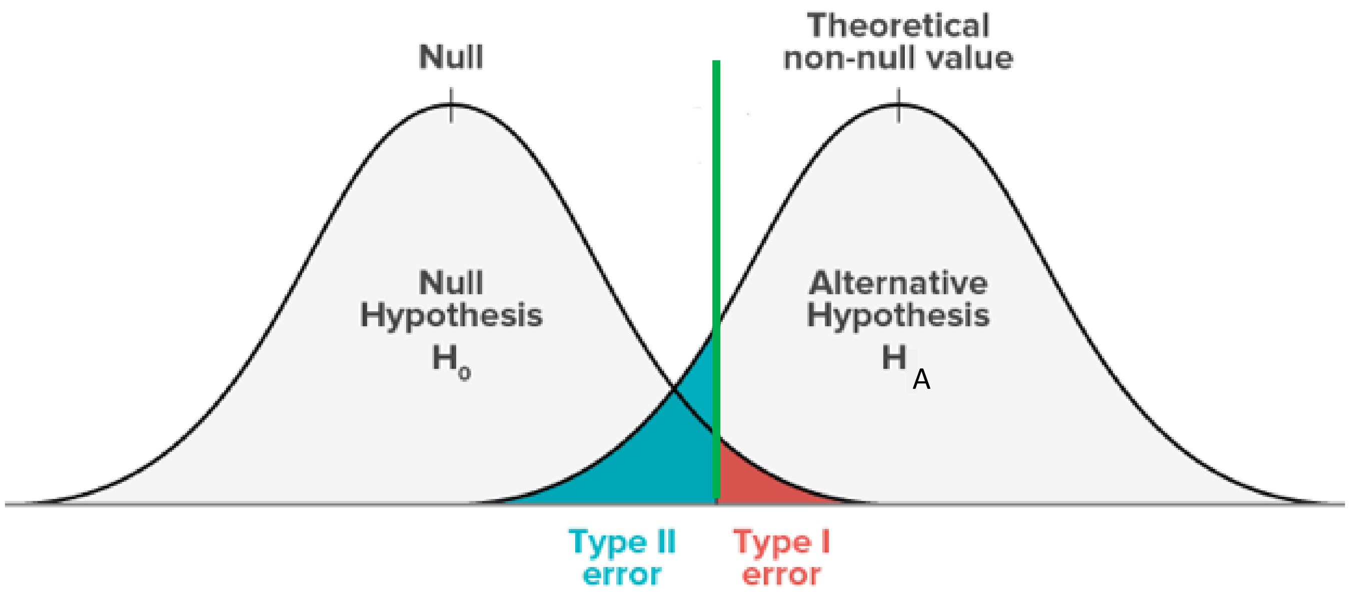
Neyman-Pearson Frequentist logic



The **power** of a test is the probability we reject the H_0 when it is **false**

- 1 - β
- For a fixed α level, it would be best to use the most powerful test

Type I and Type II Errors

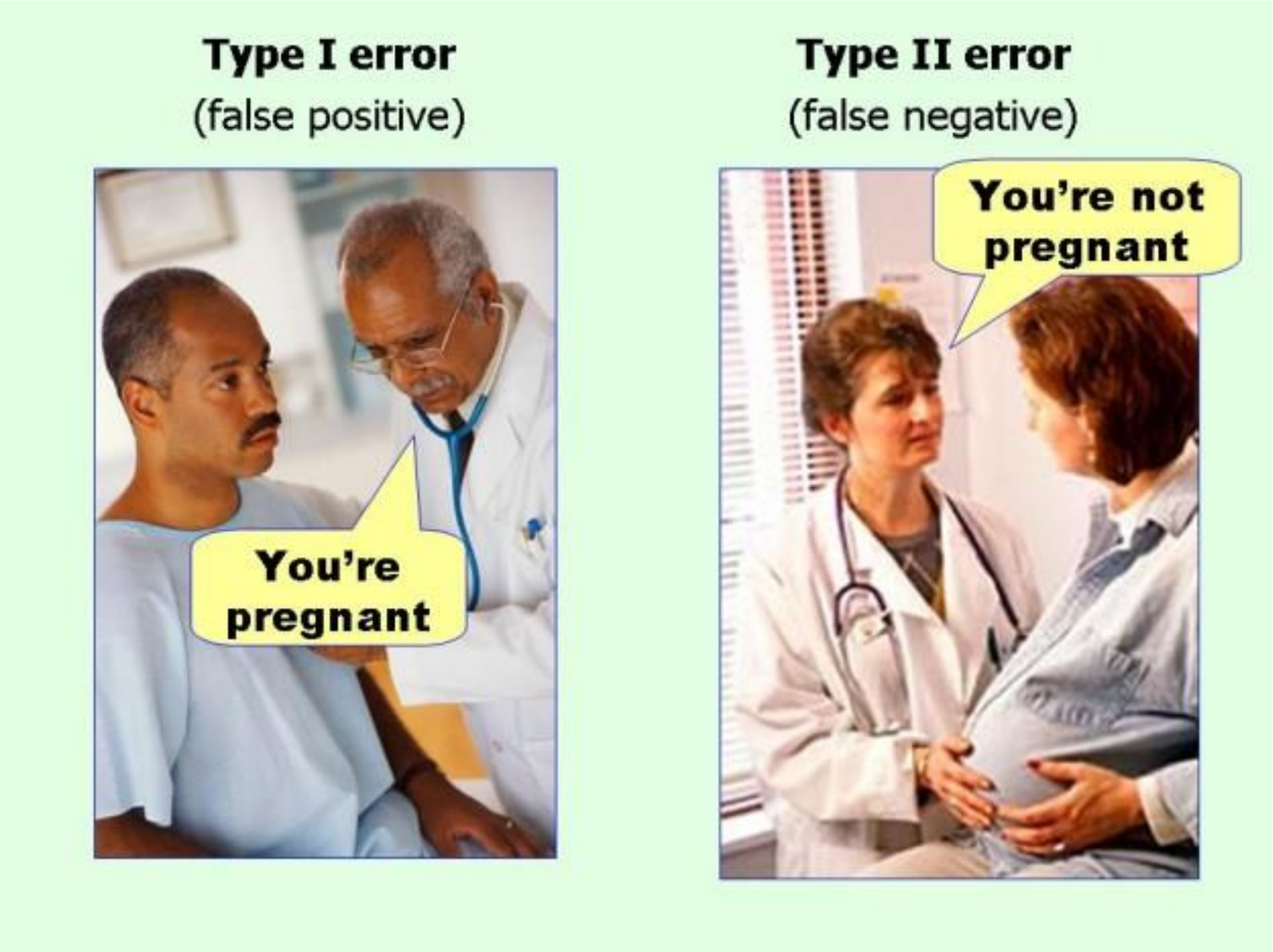


Decision

Truth

	Reject H_0	**Do not reject H_0**
H_0 is true	Type I error (α) (false positive)	No error

Type I and Type II Errors



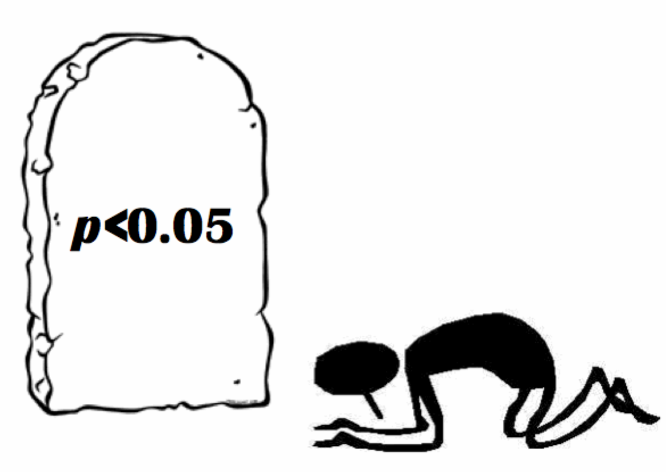
Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are false:
  - Joy can't smell Parkinson's disease, there is no difference in beer consumption across continents, Gingko has no benefits for your memory, …

Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0



Collectively Unconscious

News from the Frontiers of Science

ABOUT

NOVEMBER 3, 2012

New version SPSS will include 'celebratory fireworks' for significant results

An official press release has confirmed that the newest release of SPSS will be equipped with 'performance-rewarding features'. The new installment of the popular data-analysis package will light up with song, dance and fireworks whenever a statistical test is significant. 'We want to provide a package that is in line with the day-to-day experiences of researchers. We understand the pressure the publish, and the relief that is felt by many when those Stars of Significance appear in the results table. '

The level of significance will determine the abundance of the celebrations. If the *p*-value is below 0.05, researchers will automatically hear what is described as 'a cheerful tone', according to a company spokesman. "But if your *p*-value is below 0.01, the software package will play a series of congratulatory videos, complimenting your

SUBTITLE

RECENT POSTS

- Scientists may have 'sixth sense' for poor PSI research
- Matrix dimensions reach agreement at peace summit
- Controversial trial will provide free polymerase to junk DNA
- Animal rights activists outraged by infinite monkey experiment
- Scientists receive 12.6 million dollar grant to format references correctly

ARCHIVES

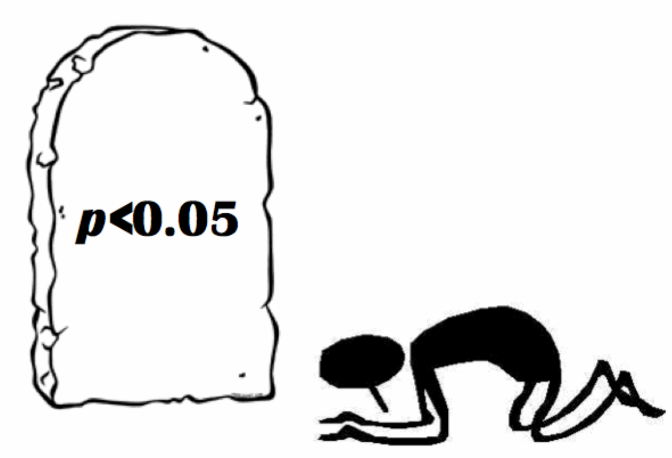
Problems with the NP hypothesis tests

<u>Problem 1</u>: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are false:
    - Joy can't smell Parkinson's disease, there is no difference in beer consumption across continents, Gingko has no benefits for your memory, …

<u>Problem 2</u>: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0?



<u>Problem 3</u>: running many tests can give rise to a high number of type I errors

Genes and leukemia example

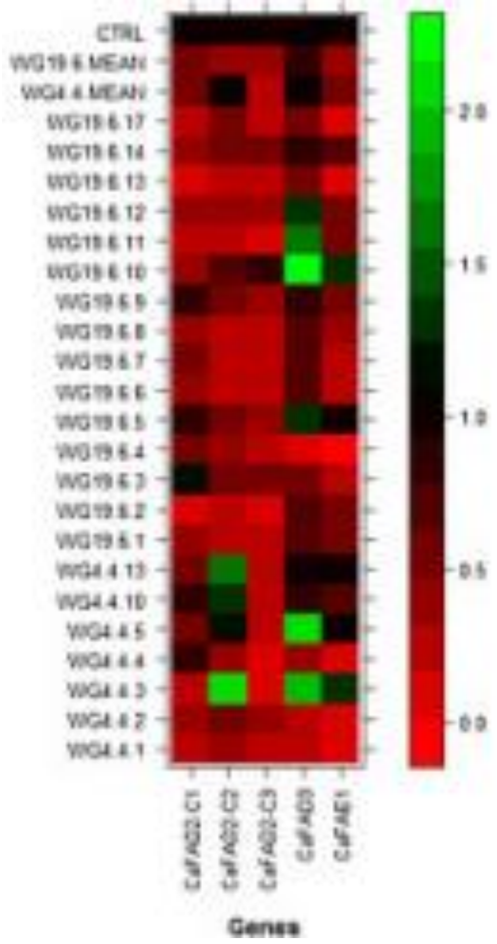
Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

- H_0: $\mu_{L1} = \mu_{L2}$ is true for all genes

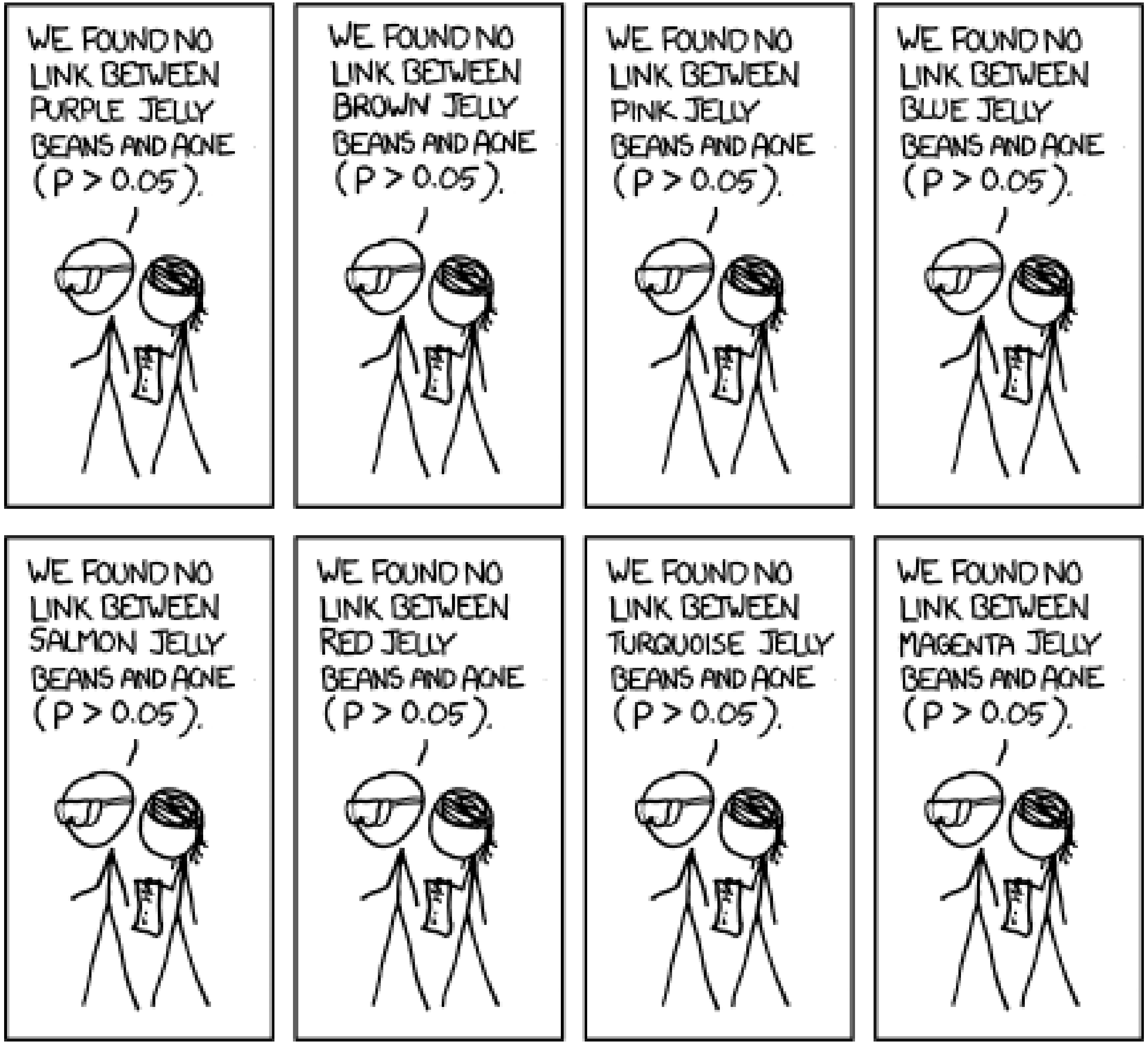
Q: If each gene was tested separately using a significance level of $\alpha = 0.05$, approximately how many type I errors would be expected?

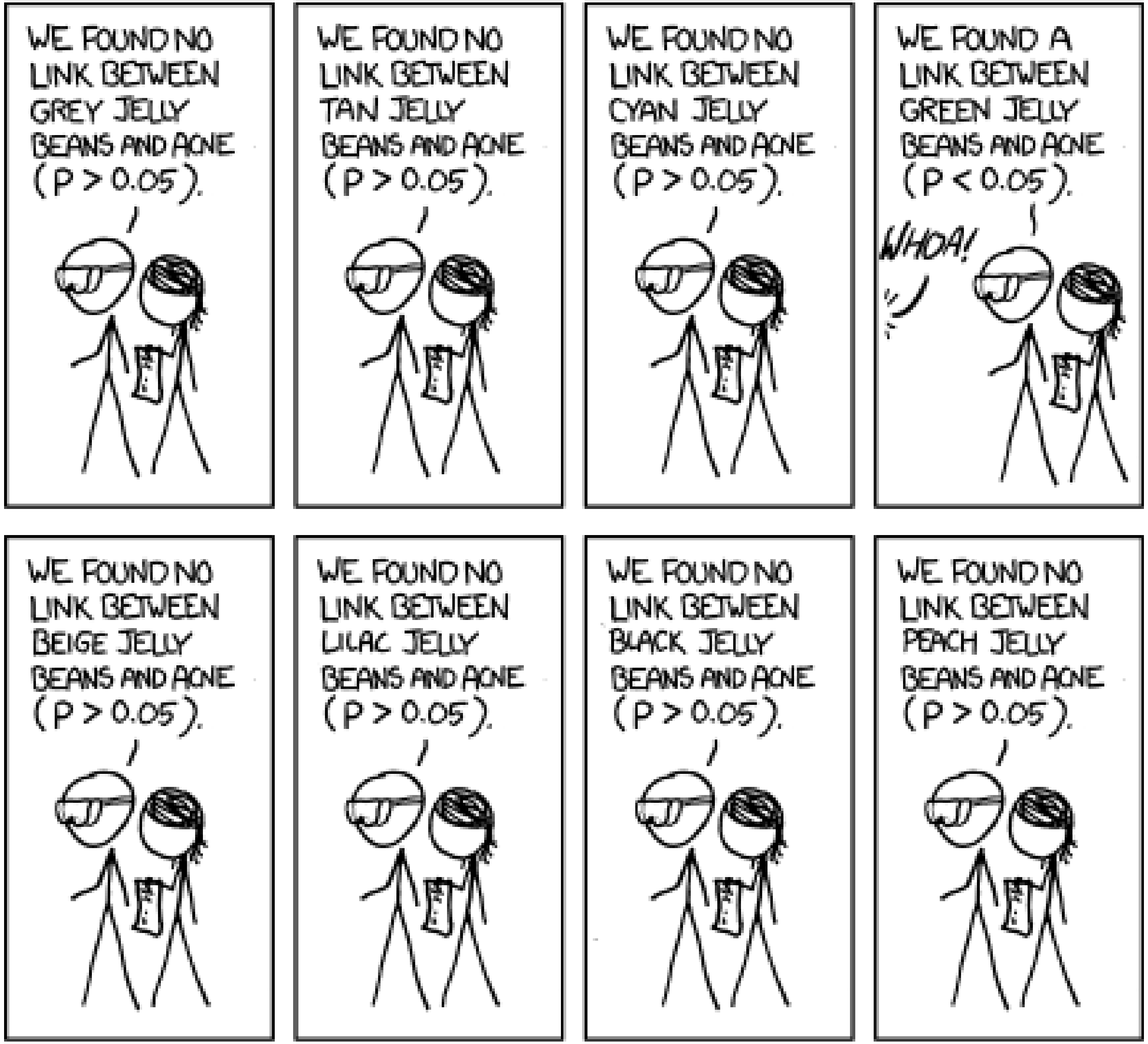
- A: 7129 x 0.05 = 356



Multiple hypothesis tests

http://xkcd.com/882/


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).

Don't ever do this!

http://xkcd.com/882/

The problem of multiple testing

For α = 0.05, when the null hypothesis is true, we should make type I errors 5% of the time

Publication bias (file drawer effect): Generally positive results are more likely to be published, so if you read the literature, the proportion of incorrect results could be greater than 5%

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The Earth Is Round ($p < .05$)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including

sure how to test H_0, chi-square with Yates's (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a

[American Statistical Association's Statement on p-values]

Some thoughts…

Better to have hypothesis tests than none at all. Just need to think carefully and use your judgment.

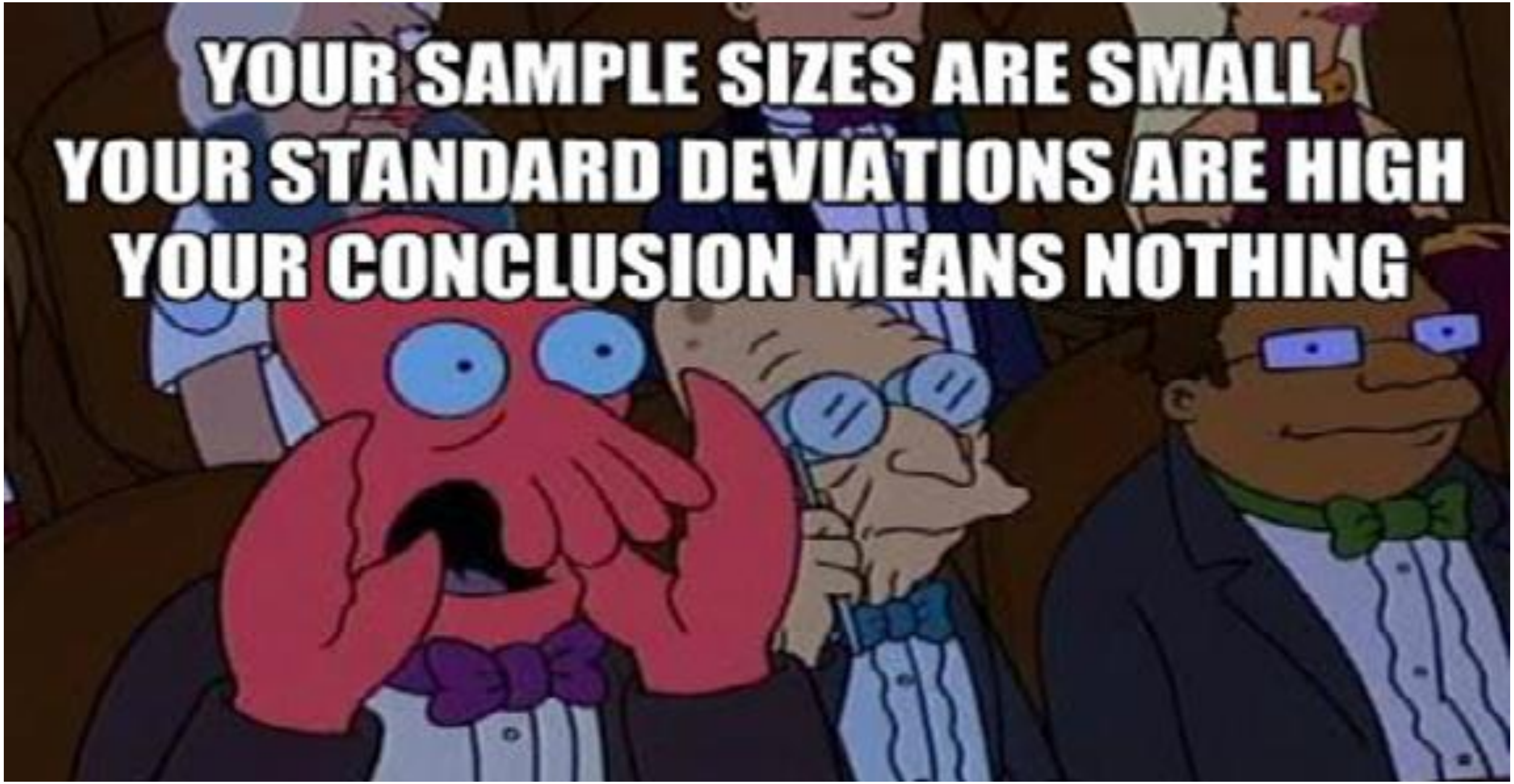
Report effect size in most cases – i.e., confidence intervals

Report the p-values rather than accept/reject H_0

- i.e., report $p = 0.23$ not $p < 0.05$

Replicate findings (perhaps in different contexts) to make sure you get the same results

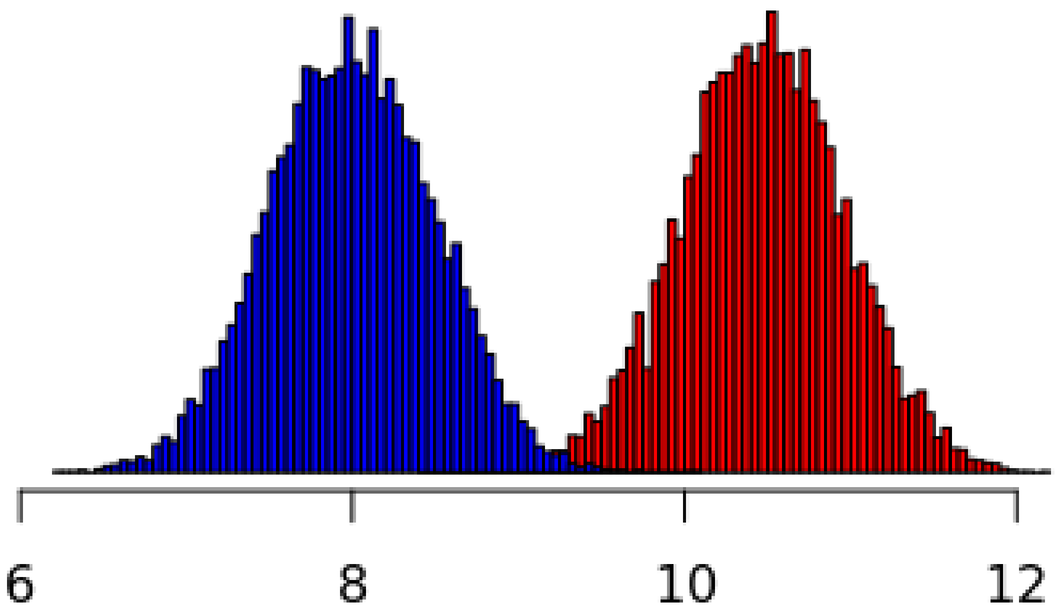
Be a good/honest scientists and try to get at the Truth!


YOUR SAMPLE SIZES ARE SMALL
YOUR STANDARD DEVIATIONS ARE HIGH
YOUR CONCLUSION MEANS NOTHING

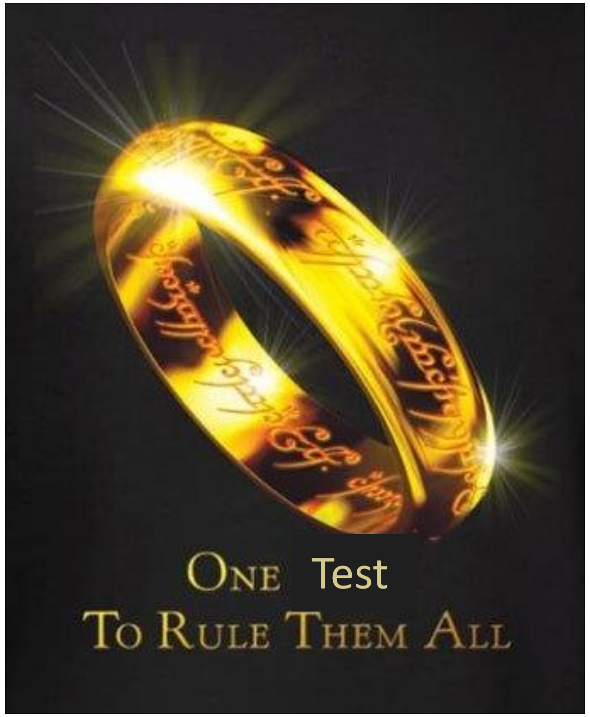
Questions?

Extra material if there is time

Connections between null, alternative and bootstrap distribution using test of a single mean

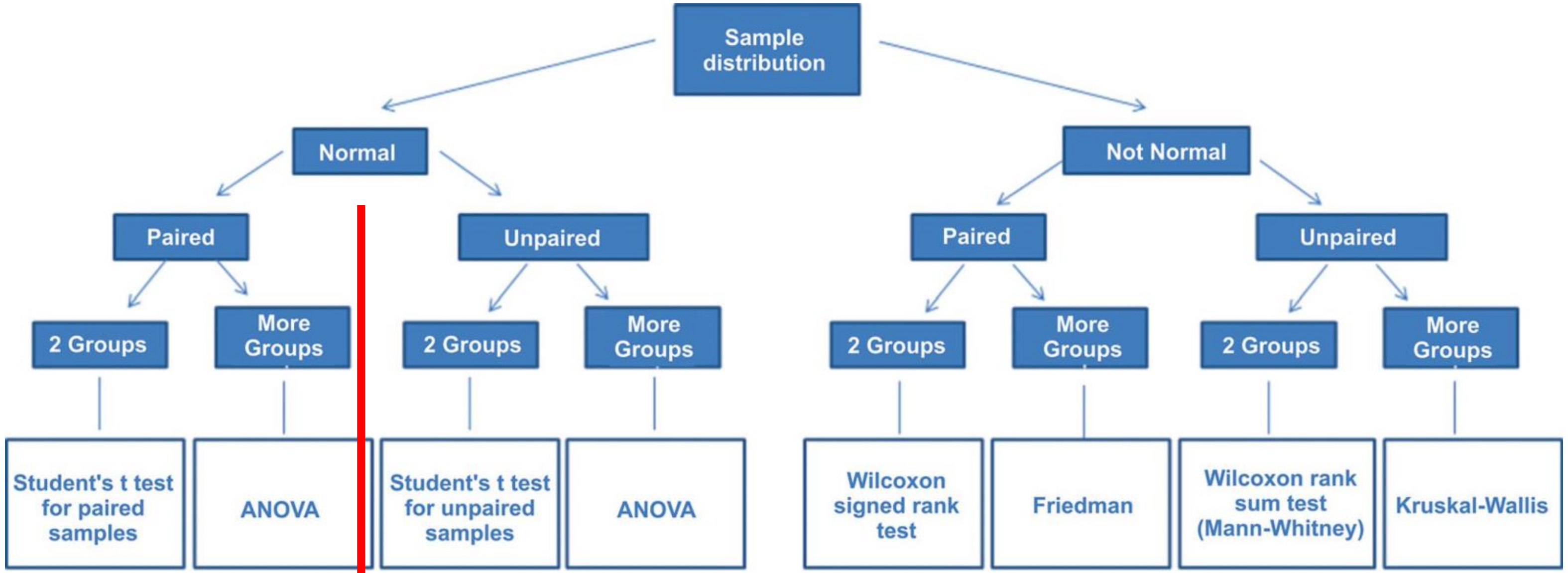


The big picture: There is only one hypothesis test!



We can run a large number of additional hypothesis tests by following the 5 steps!

The hypothesis test zoo



Single sample

Single sample t-test

Example: Do mammals on average sleep more than humans?

According to a data set that comes with the ggplot package, humans sleep 8 hours a day

- (I wish)

The data set also has the sleeping times of 82 other mammals

Let's test if the average sleep time of all mammals is different than 8 hours, based on the sample of 82 mammals.

- (warning: we obviously need to be careful drawing conclusions here because it's not clear whether this is a simple random sample of mammals)

Parametric hypothesis test for a single mean

Step 1: state the null hypothesis:

H_0: $\mu = 8$

Step 2: We can use a t-statistic:

$$t = \frac{estimate\ \text{-}\ param_0}{\hat{SE}} \qquad \hat{SE} = \frac{s}{\sqrt{n}}$$

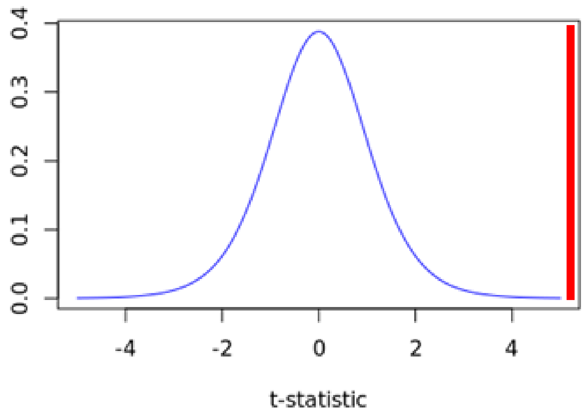
$$t = \frac{\bar{x}\ \text{-}\ 8}{\frac{s}{\sqrt{n}}}$$

$\bar{x}$ = 10.46 n = 82

s = 4.47 t = 4.99

Note: In a paired samples t-test we subtract the paired values in the two samples and run a one sample t-test on the differences.

Step 3: The null distribution is a t-distribution with n - 1 degrees of freedom



Step 4 and 5... ???

We can also get confidence intervals using:

$$CI = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

Randomization hypothesis test for a single mean

Step 1: Null hypothesis: H_0: $\mu = 8$

Step 2: We could use:

- The mean statistic $\bar{x}$
- A t-statistic

Step 3: Any ideas how to create one point in our null distribution?

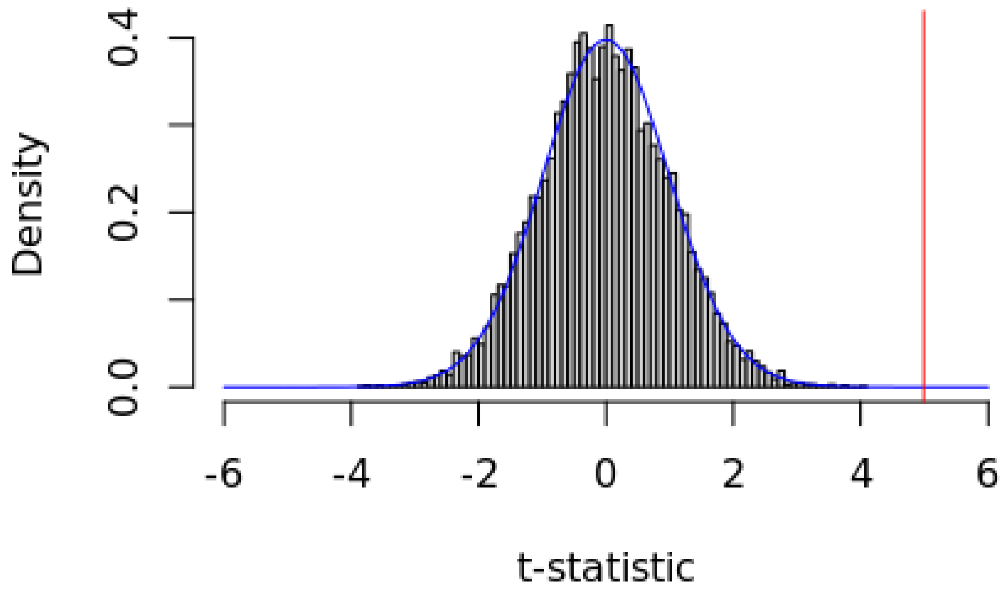
1. Modify the original sample by adding a constant to all data points to make the sample mean equal to the null hypothesis parameter value
    - > data_sample - mean(data_sample) + 8

2. Sample n points with replacement from the modified sample and calculate a statistic on this resampled data to get one statistic consistent with the null hypothesis

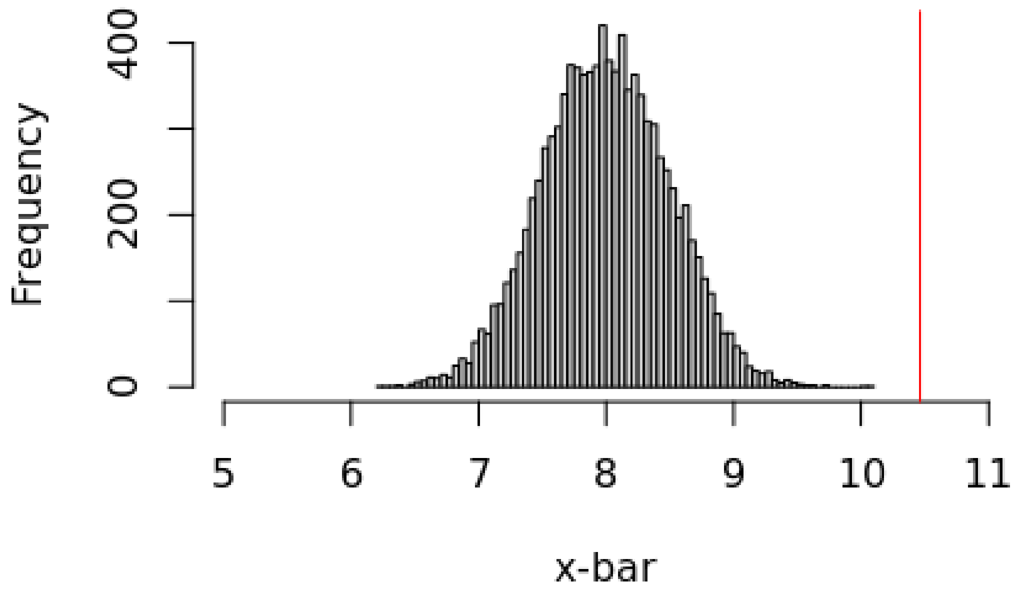
3. Repeat 10,000 times

Null distributions

Null distribution using a t-statistic

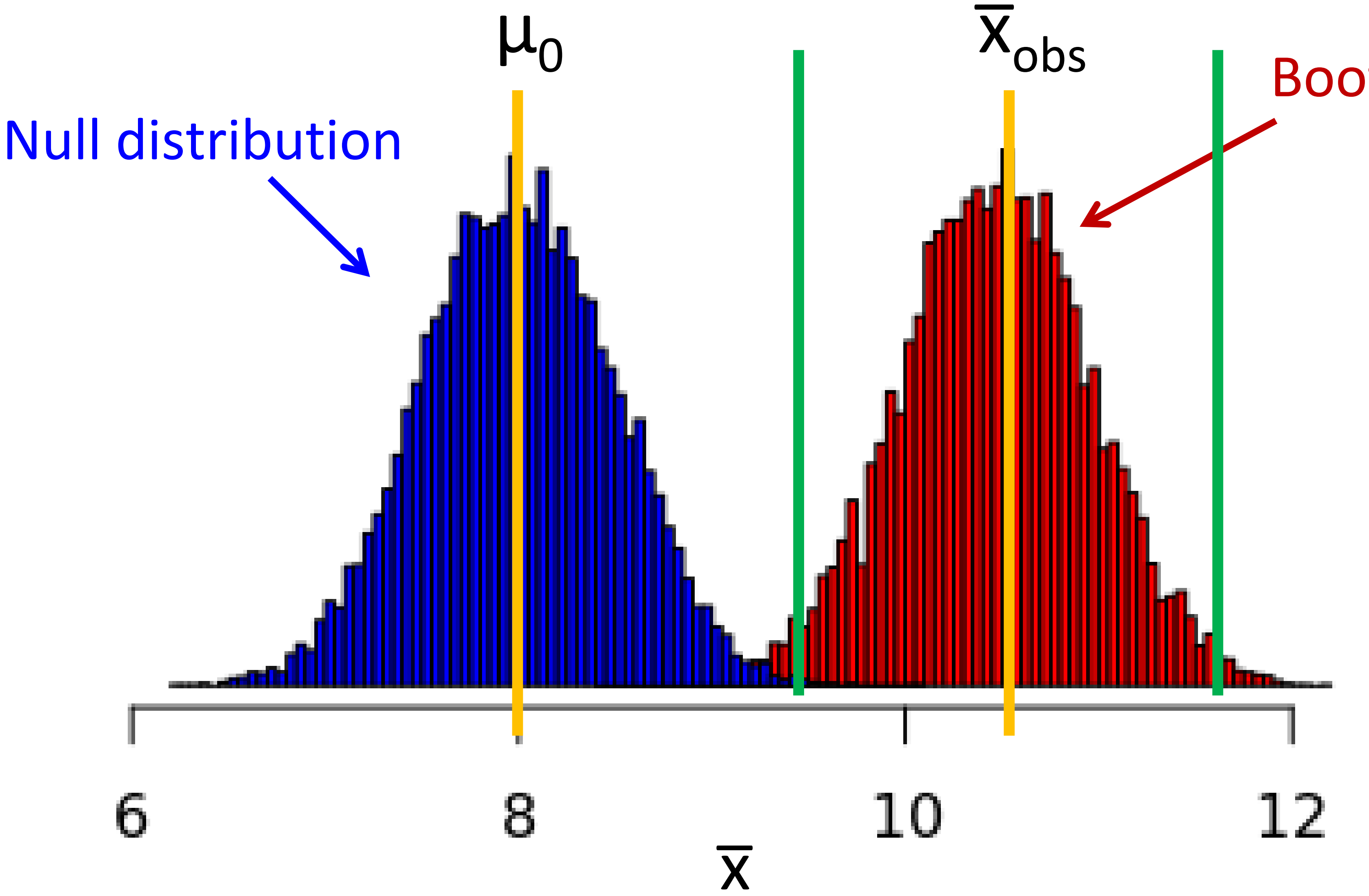


Null distribution using $\bar{x}$ statistic



The p-value in both cases is… 0

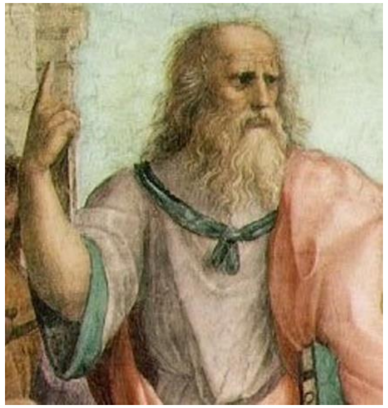
Relationship between null and bootstrap distributions

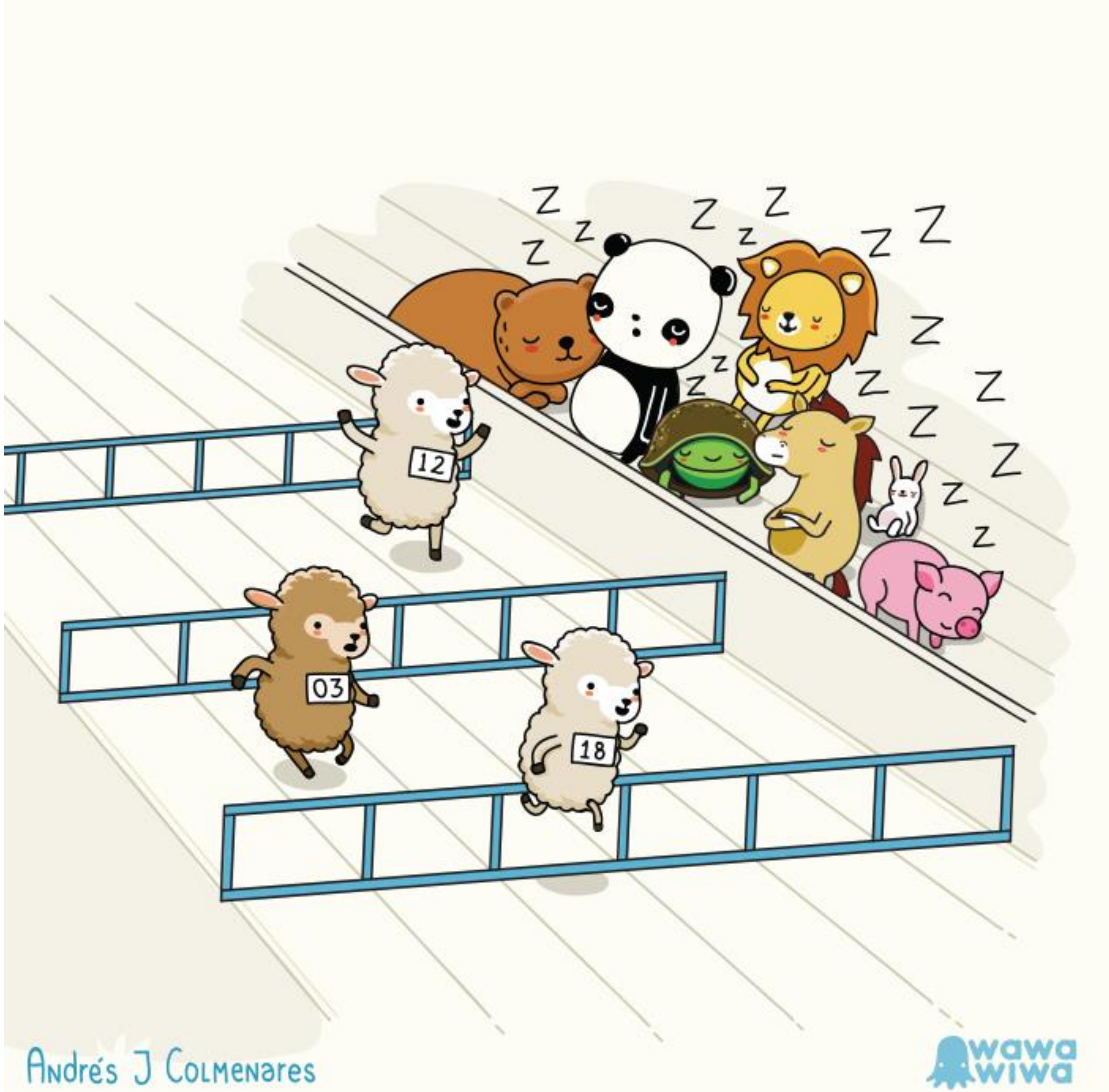


95% CI limits

Where is:

- 1. The true μ?
- 2. The true alternative sampling distribution?





12
03
18
Andrés J Colmenares
wawa wiwa

Next class: start on the tidyverse…