Parametric hypothesis tests continued

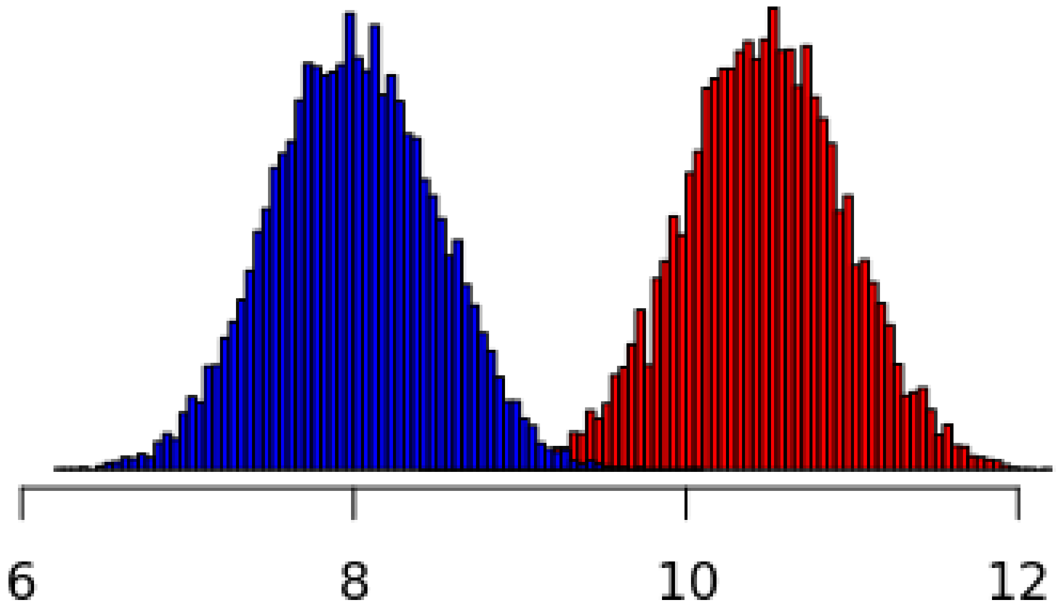
Overview

Quick review and extensions of parametric and nonparametric tests

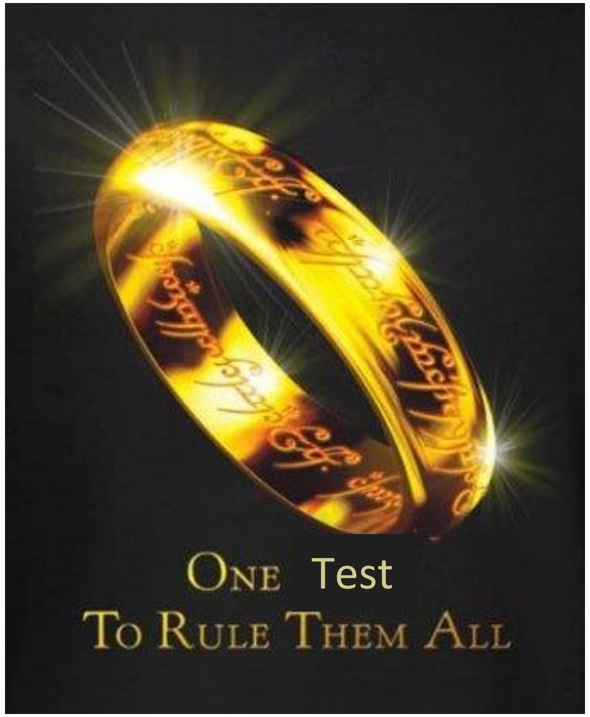
Theories of hypothesis testing

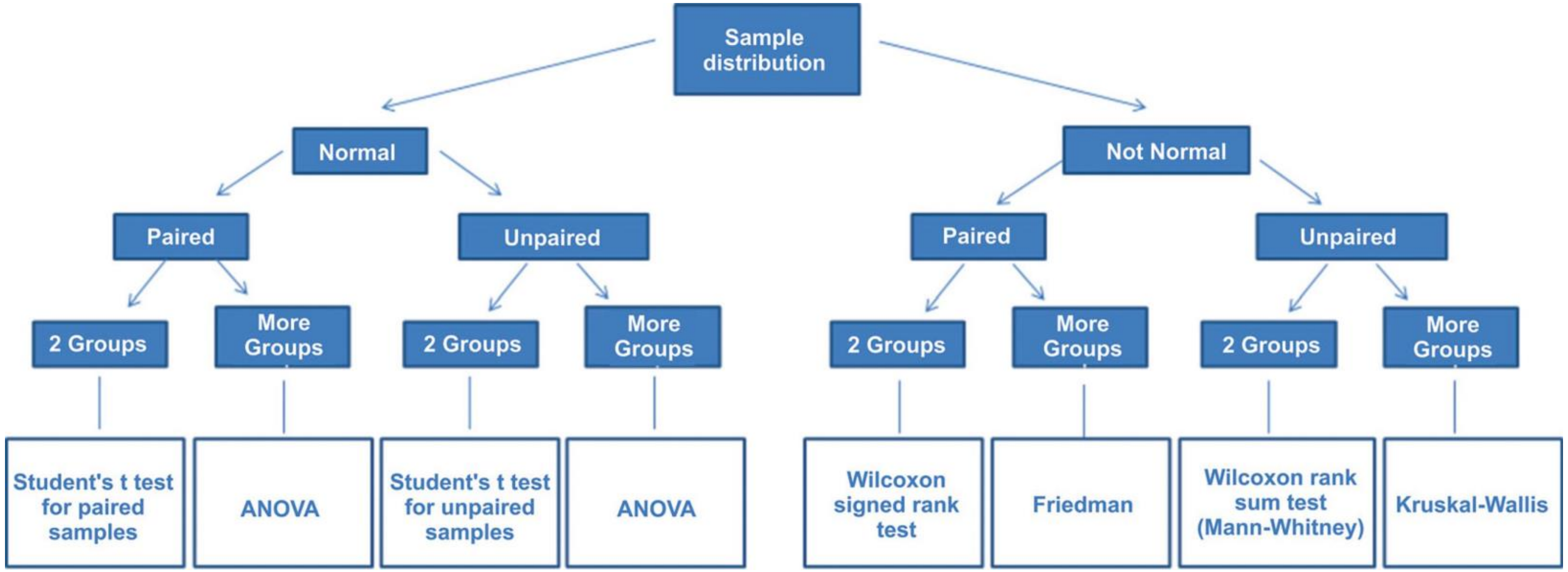
Hypothesis tests for a single mean and connections between hypothesis tests and confidence intervals

Review and extensions to hypothesis tests and CIs

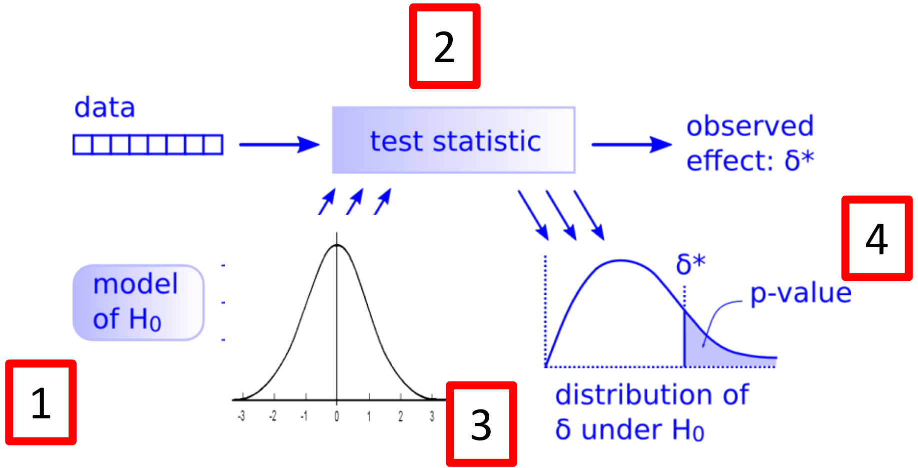


The big picture: There is only one hypothesis test!





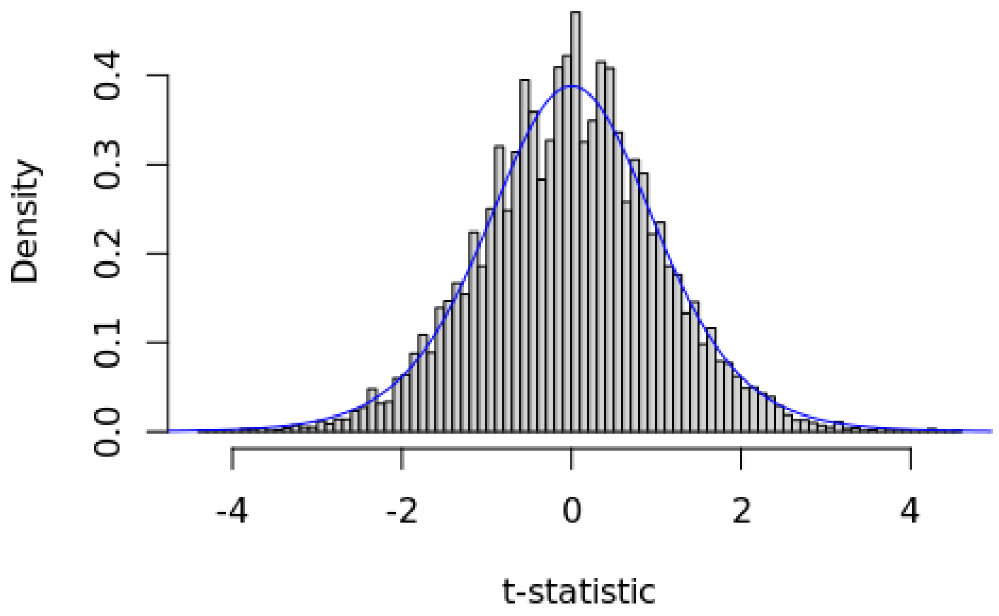
Just need to follow 5 steps!



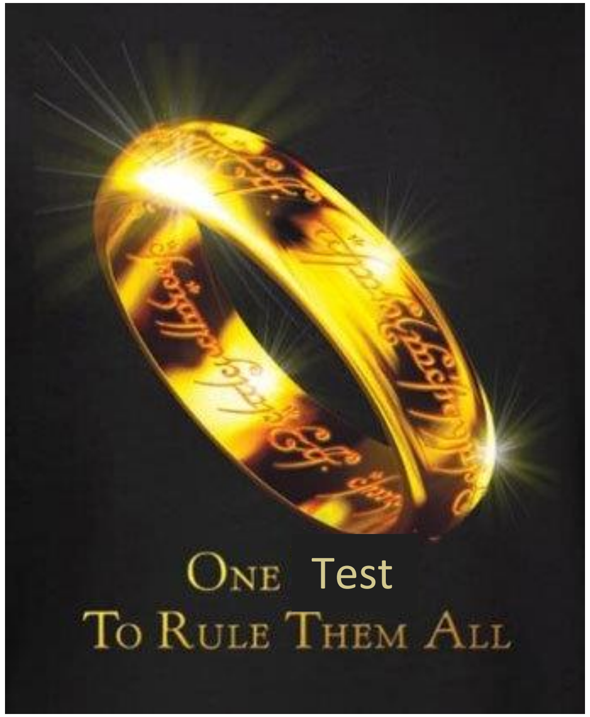
Review: Randomization and parametric hypothesis tests

The difference between randomization/permutation tests and parametric hypothesis tests is in how the null distribution is created (step 3):

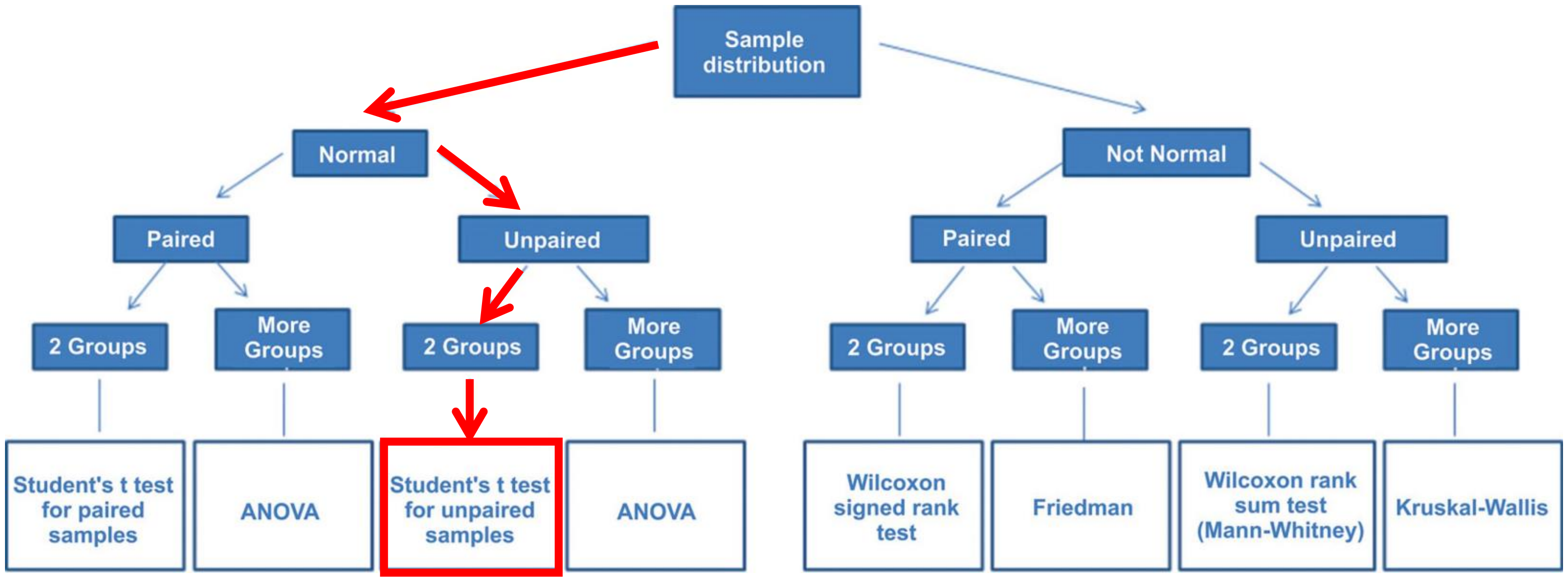
- Randomization/permutation tests the null distribution is created through computational simulations
- In parametric tests, the null distribution is created using a parametric probability distribution



Review: t-test for comparing two means

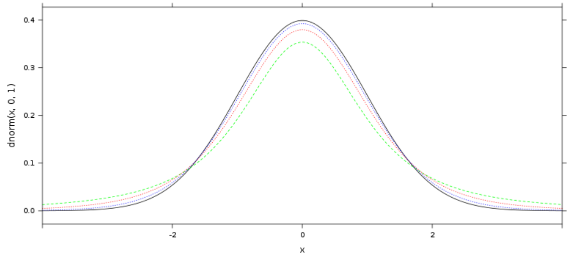


The hypothesis test zoo

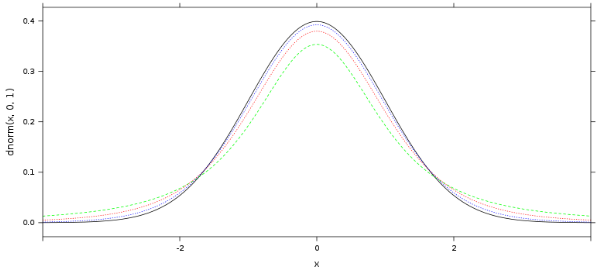


We can run a large number of additional hypothesis tests by following the 5 steps!

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



t-test for comparing two means



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

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

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

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

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

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

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

Since we have SE estimates, we can compute confidence intervals:

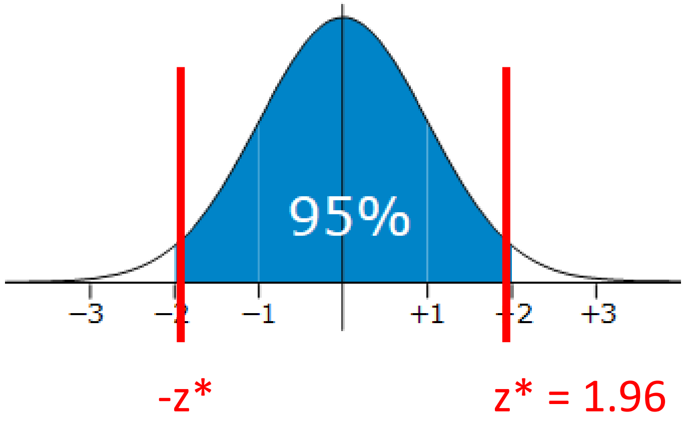
$CI_{95} \approx$ stat $\pm\ 2 \cdot$ SE

Confidence interval for the difference of two means

Confidence intervals for the bootstrap had the form:

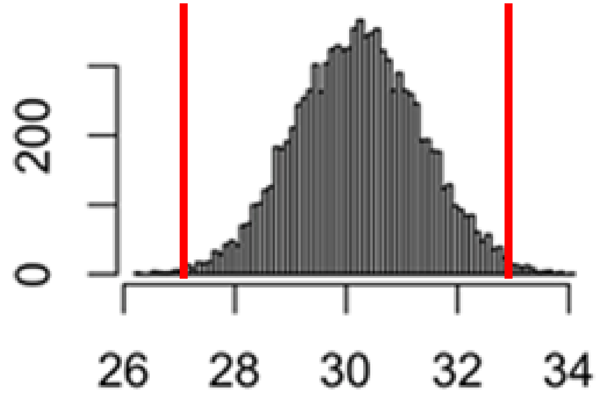
$$CI_{95} \approx \text{stat} \pm 2 \cdot SE^*$$

qnorm(.975) = 1.96



<u>Side note</u>: one can also calculate 95% bootstrap confidence intervals using:

quantile(bootstrap_distribution, c(.025, .975))

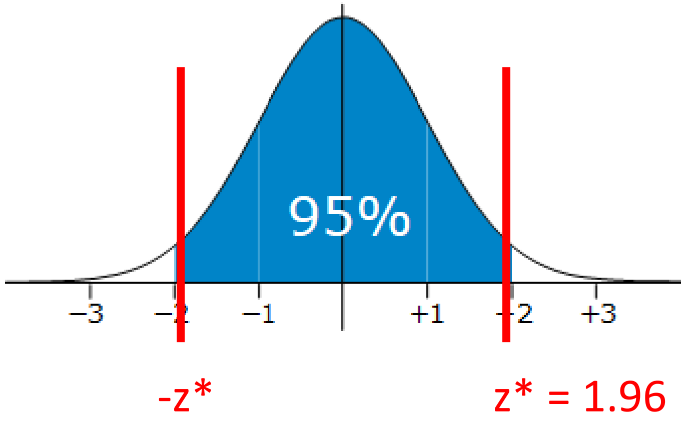


Confidence interval for the difference of two means

Confidence intervals for the bootstrap had the form:

$CI_{95} \approx$ stat $\pm$ 2 · SE*

qnorm(.975) = 1.96



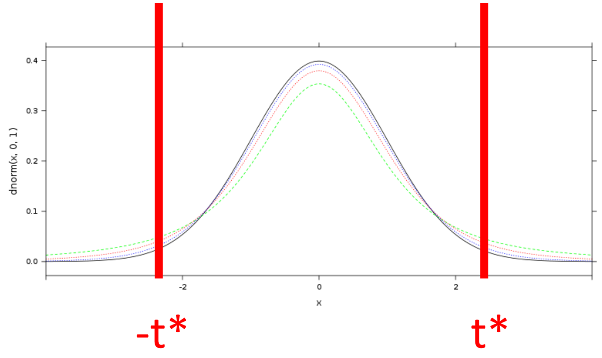
When creating confidence intervals based on a t-distribution we use:

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

df = min(n_t, n_c) - 1

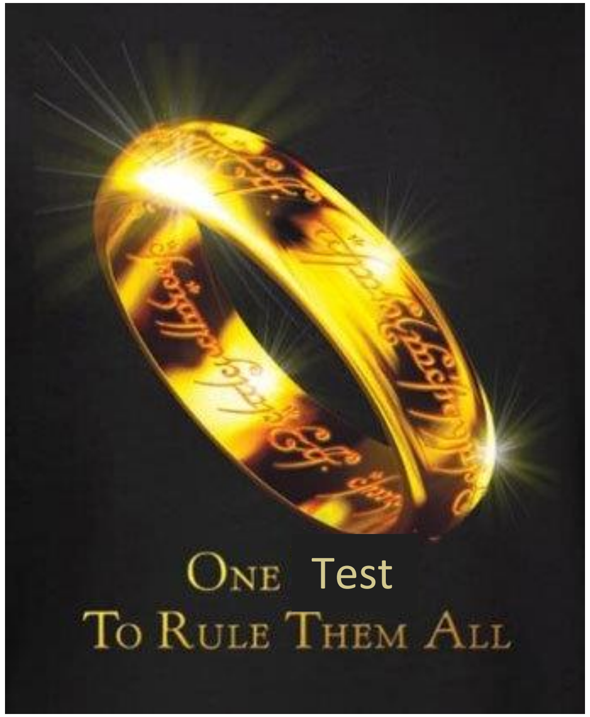
qt(.975, df)

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



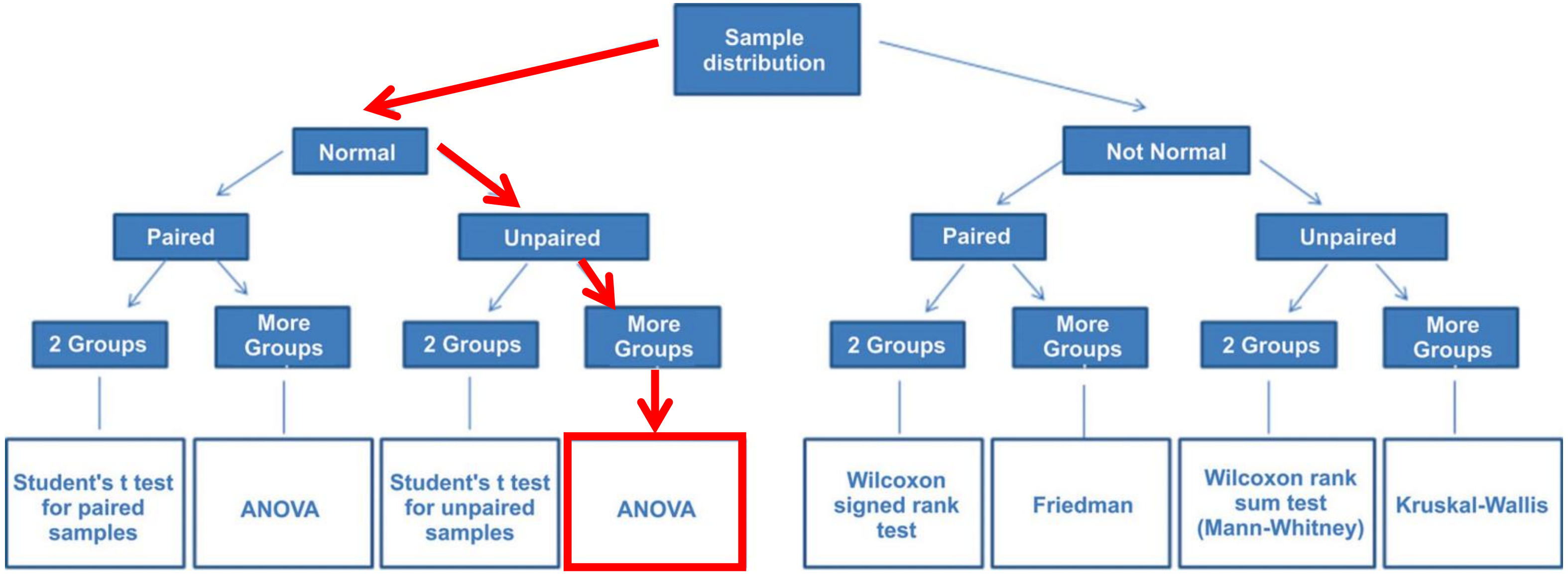
For a difference of means: $CI = (\bar{x}_t - \bar{x}_c) \pm t^* \cdot \sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}$

The big picture: There is only one hypothesis test!



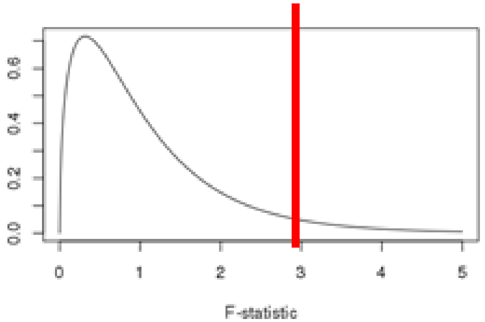
We can run a large number of additional hypothesis tests by following the 5 steps!

The hypothesis test zoo

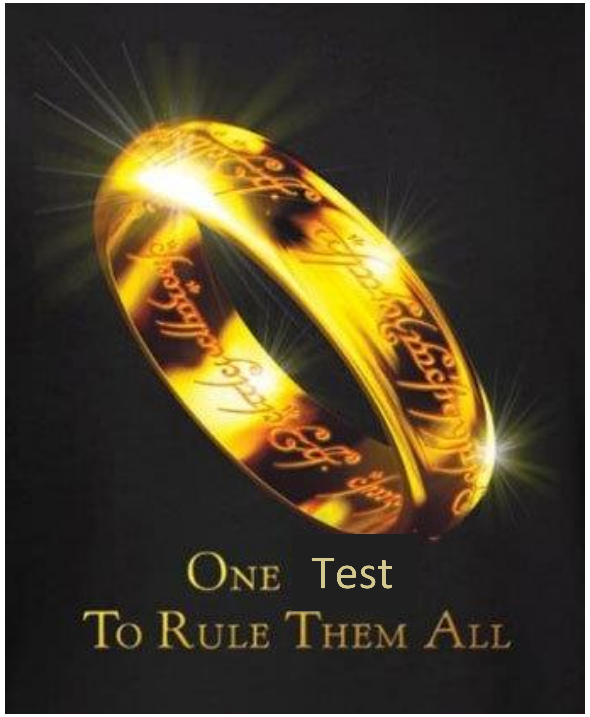


ANOVA: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

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

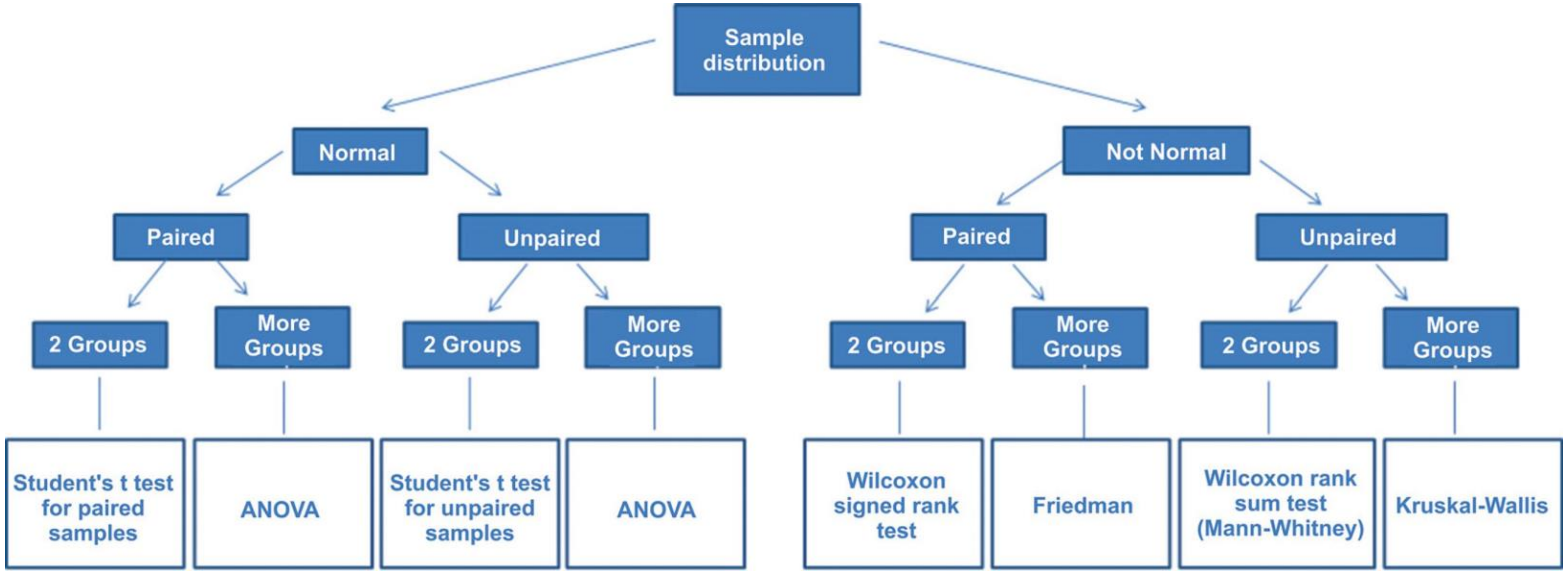


The big picture: There is only one hypothesis test!



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

The hypothesis test zoo



Nonparametric hypothesis tests

Brief mention: nonparametric hypothesis tests

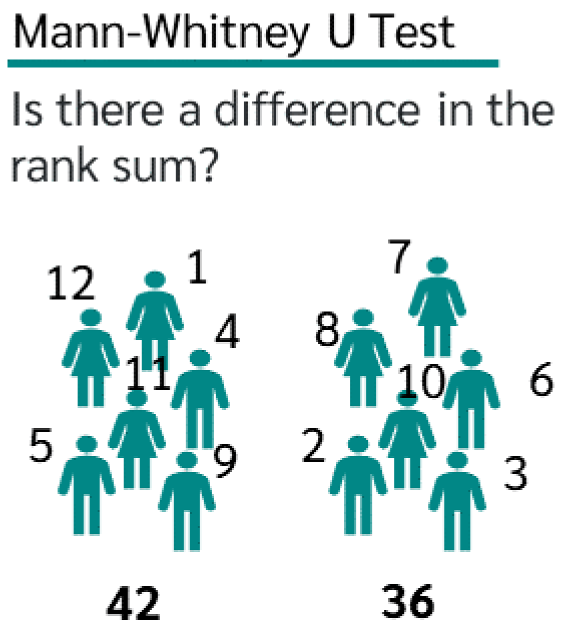
Nonparametric hypothesis tests use null distributions that do not have a small fixed set of parameters

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)

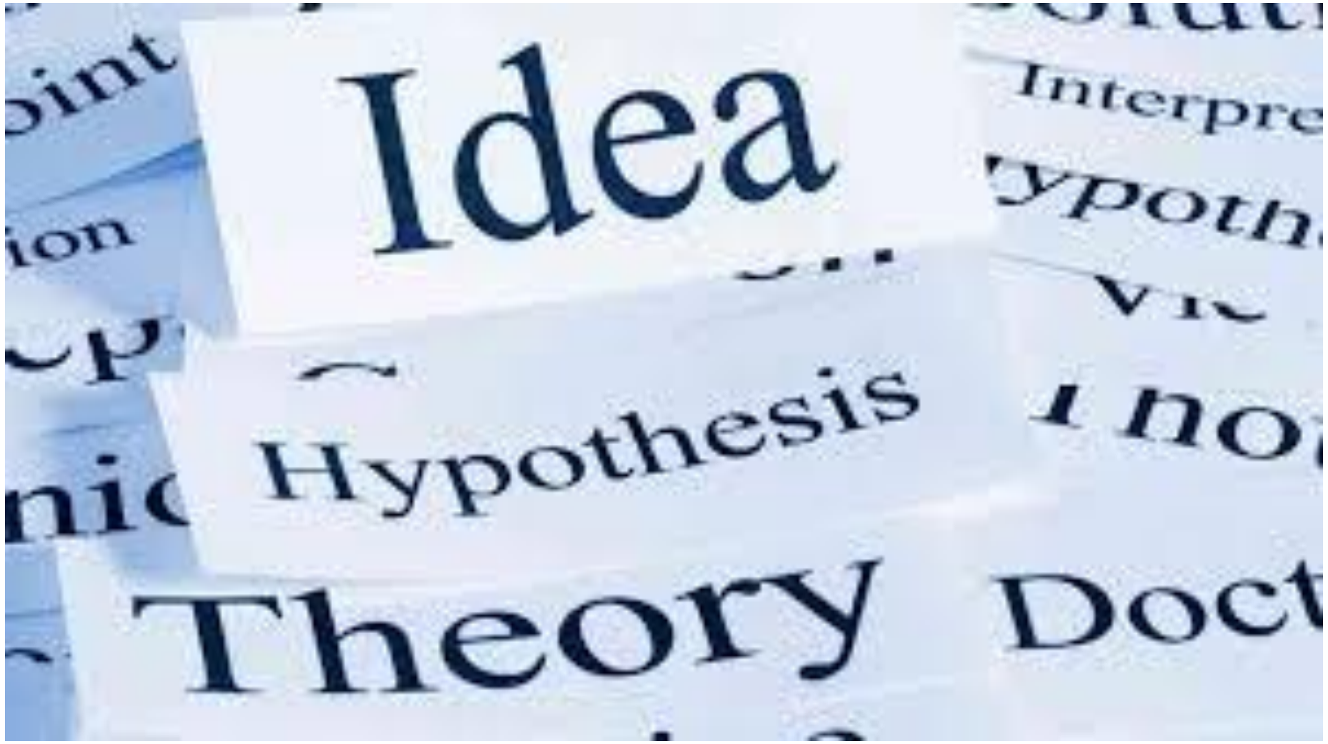
Nonparametric tests have fewer assumptions than parametric tests so they are potentially more robust

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



Questions?

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)

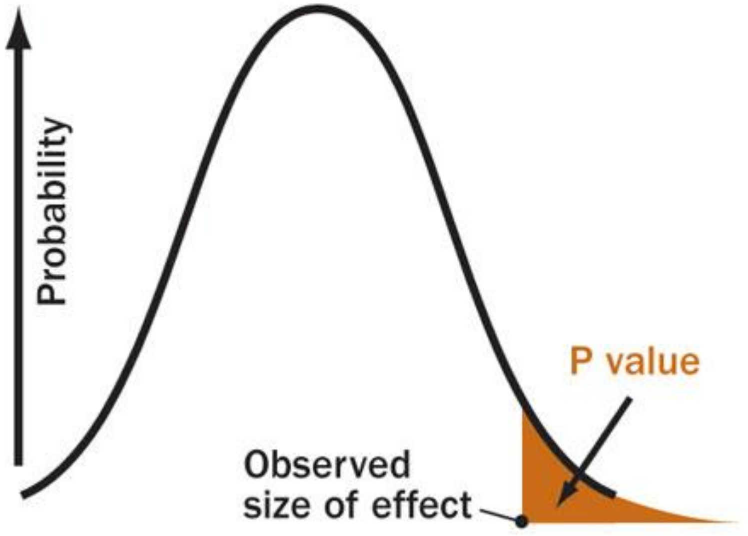
Pearson (1895-1980)

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”



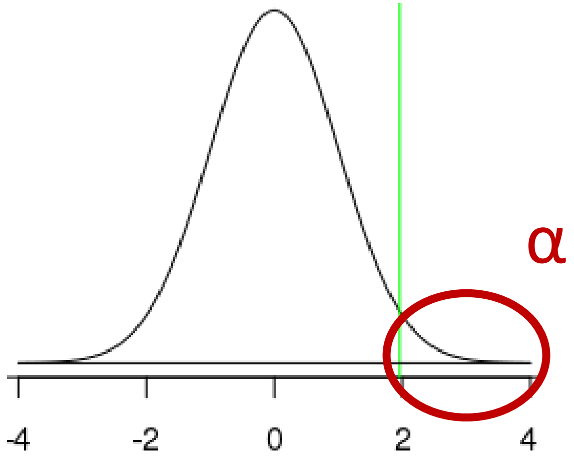
Neyman-Pearson null hypothesis testing

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

Null distribution



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

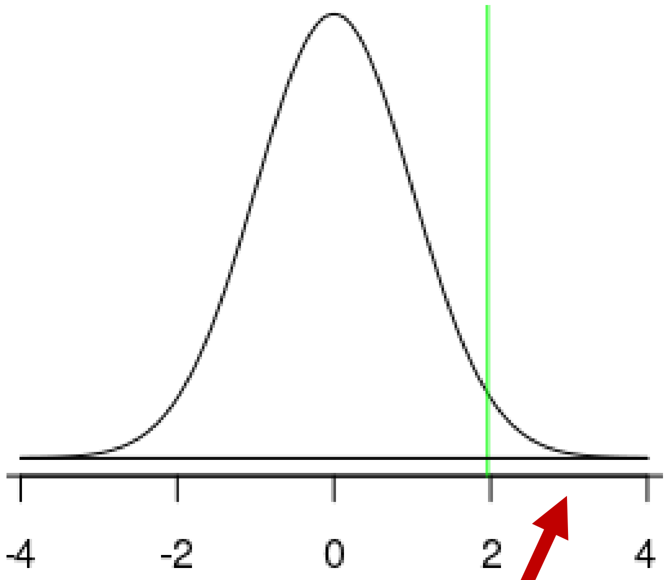
Neyman-Pearson frequentist logic

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 $\alpha = 0.05$)

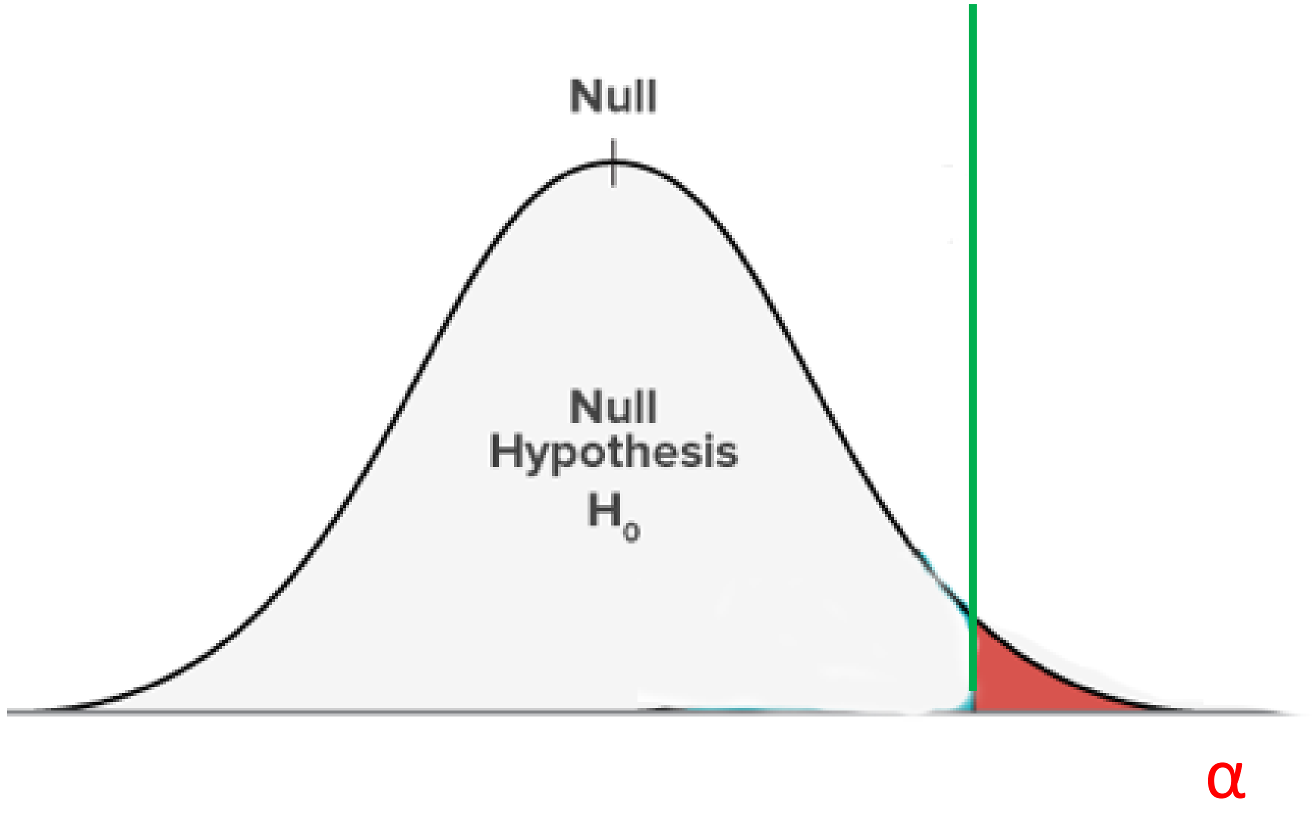
- i.e., we would only make type I errors 5% of the time

Null distribution

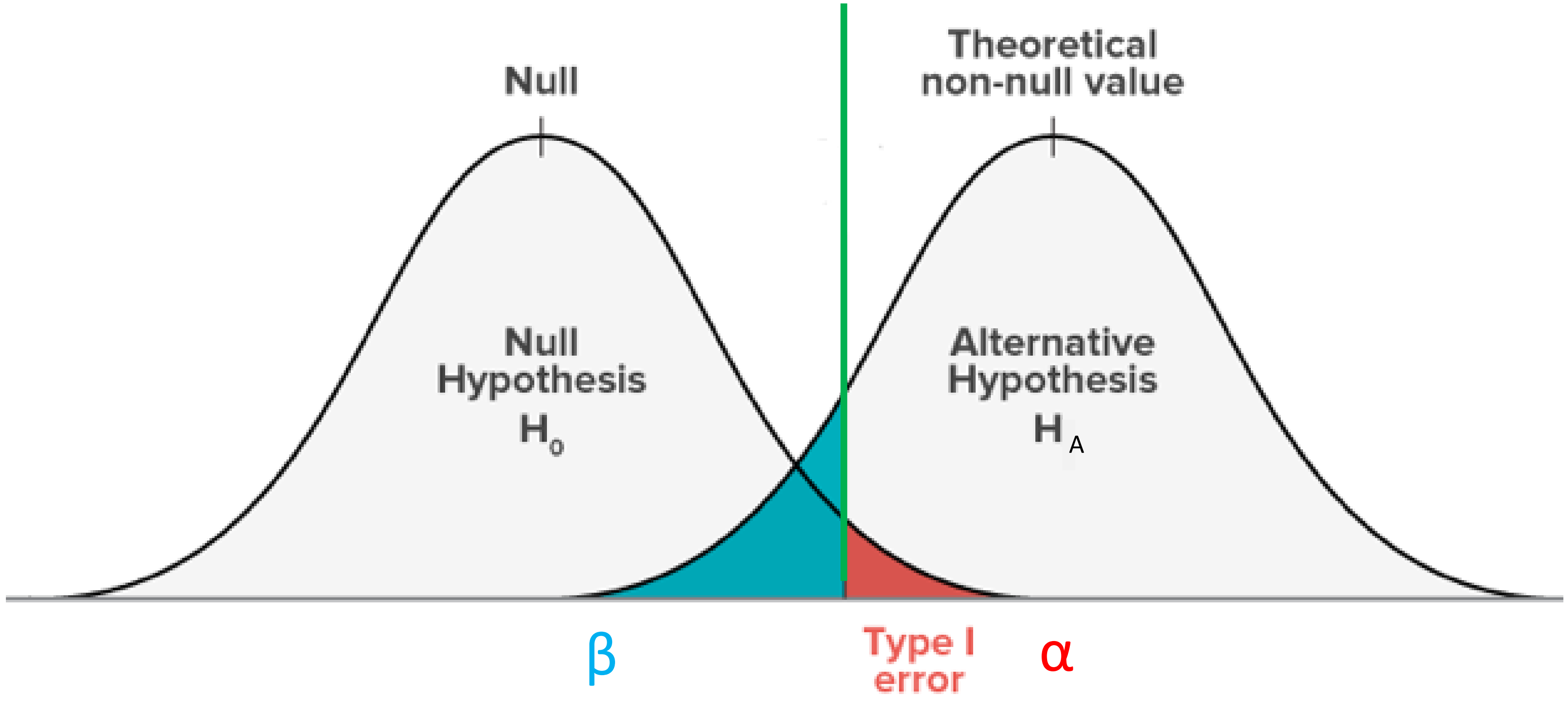


The null distribution is true but statistic landed here

Neyman-Pearson Frequentist logic



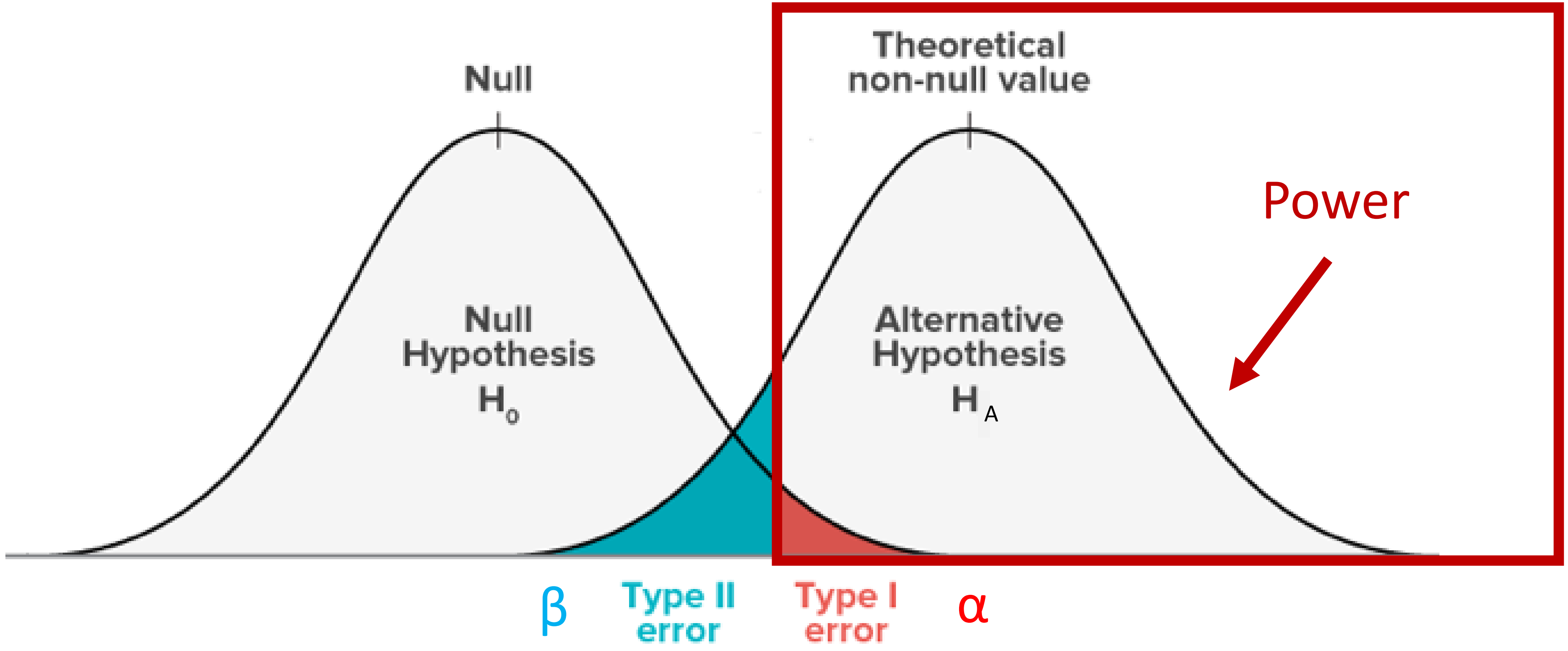
Neyman-Pearson Frequentist logic



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

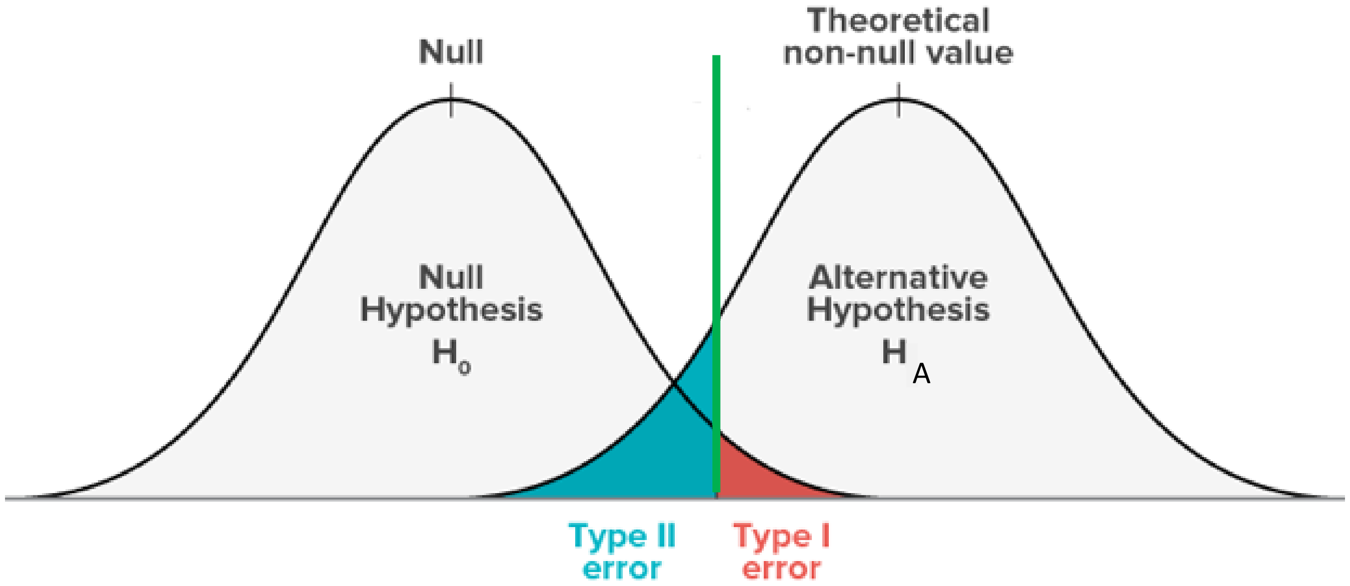
Neyman-Pearson Frequentist logic



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
H_A is true (H_o is false)	No error	Type II error (β) (false negative)

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

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?

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

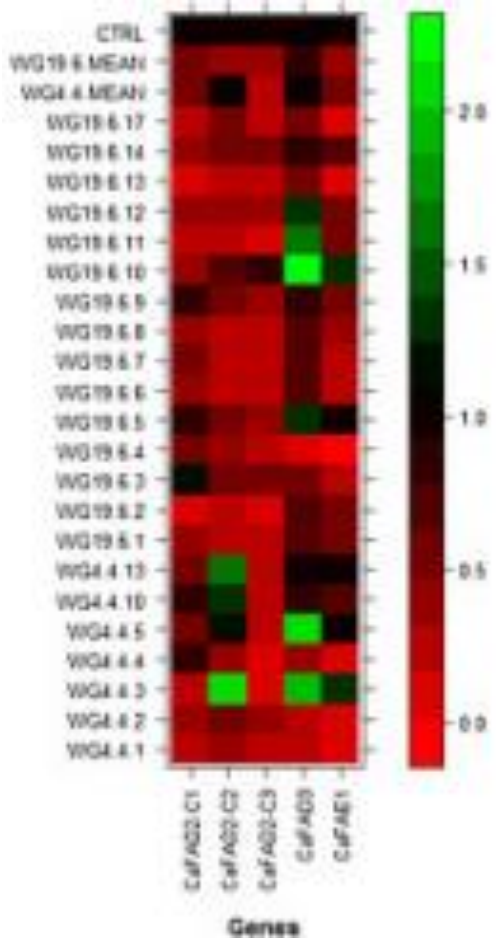
Genes and leukemia example

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?



The problem of multiple testing

For $\alpha = 0.05$, ~5% of all published research findings when the null hypothesis is true (i.e., H_0 is true), should incorrectly reject H_0

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

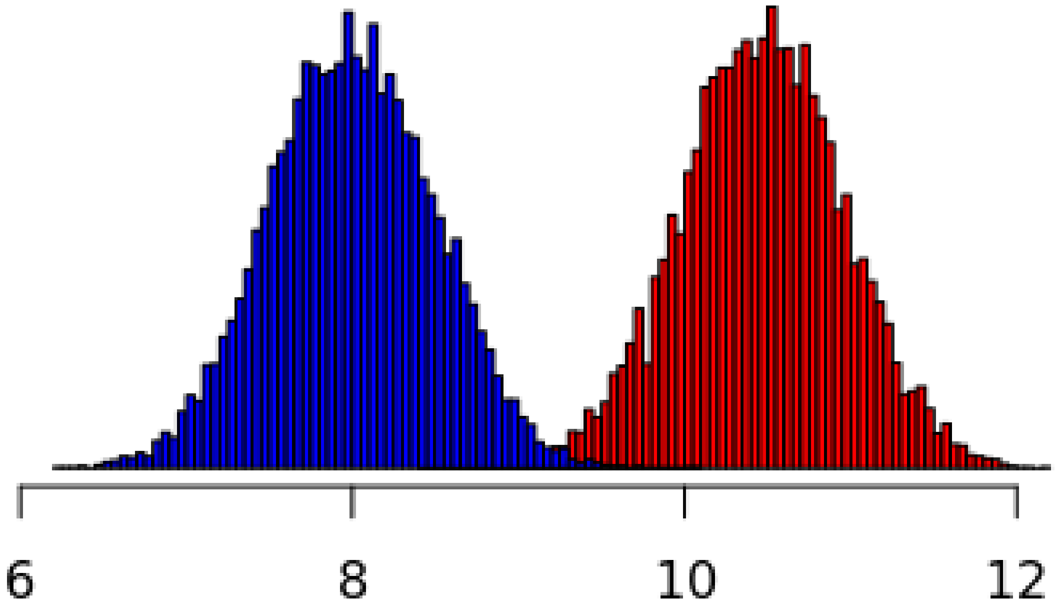
- i.e., report $p = 0.23$ not $p < 0.05$

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

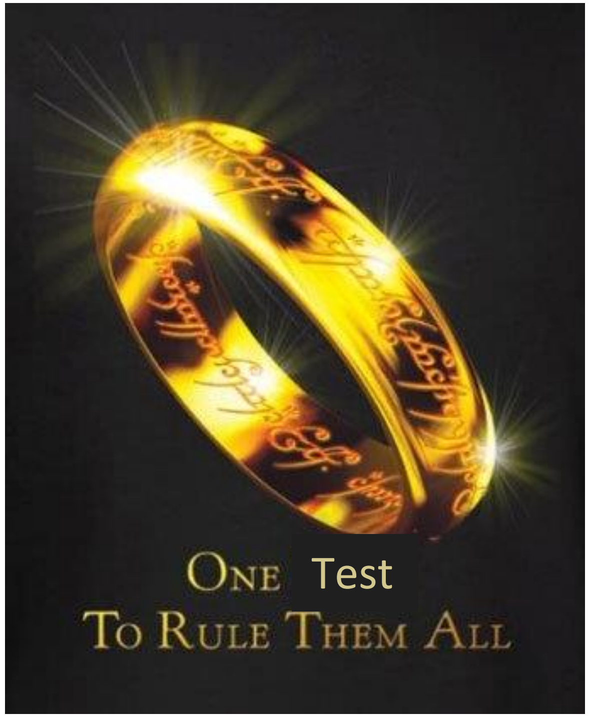
Be a good/honest scientists and try to get at the Truth!

Questions?

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

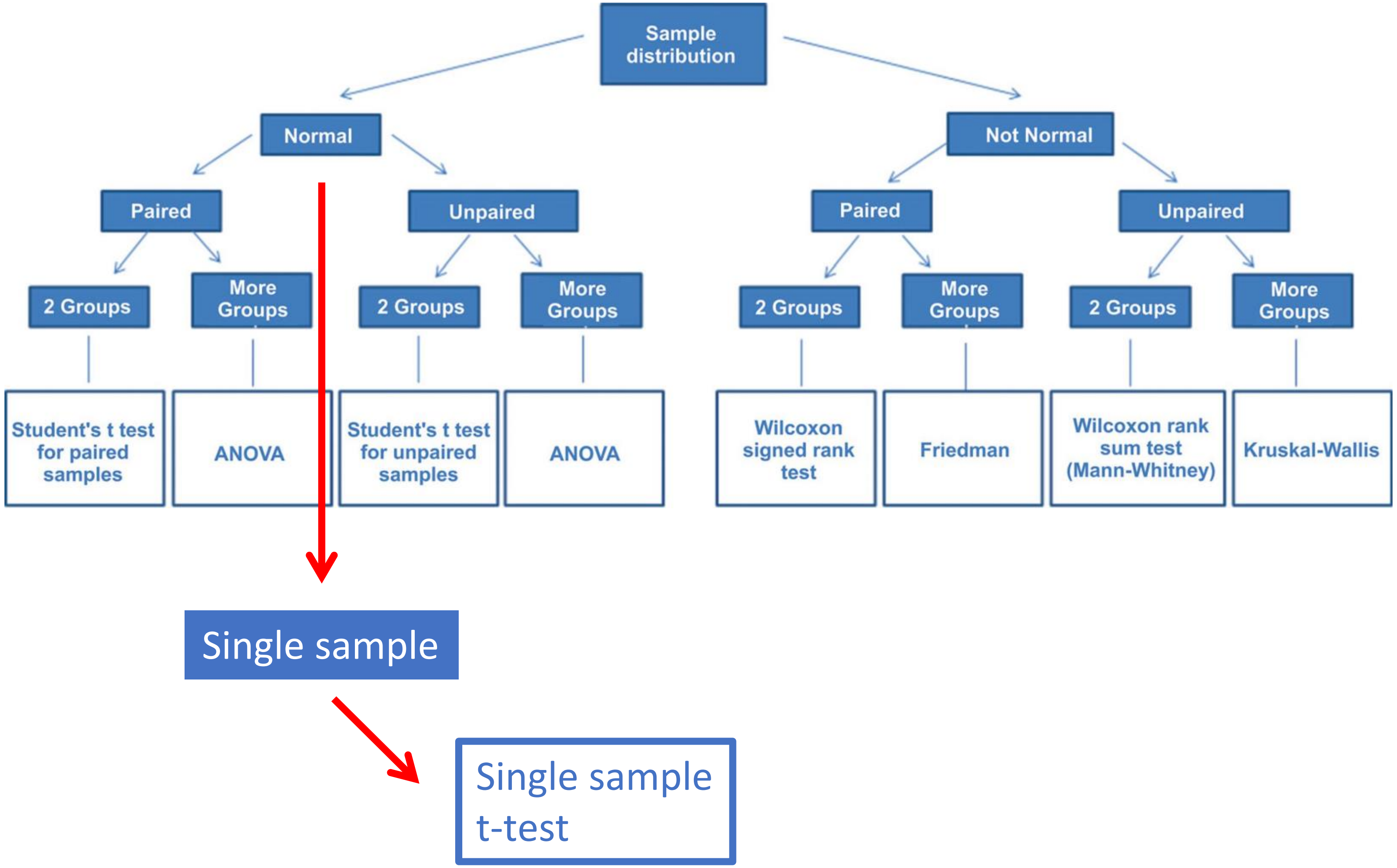


The big picture: There is only one hypothesis test!



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

The hypothesis test zoo



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

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:

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

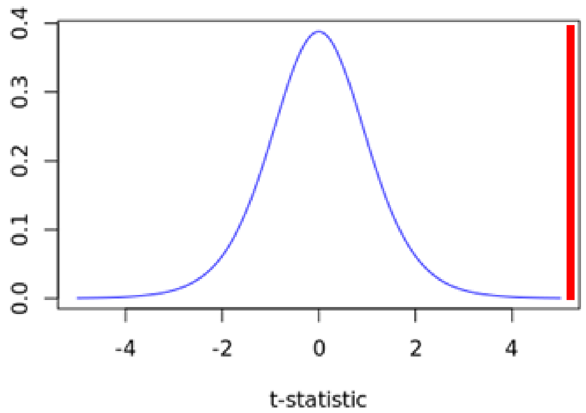
$$t = \frac{\bar{x}\ \text{-}\ 8}{\frac{s}{\sqrt{n}}}$$

$\bar{x} = 10.46$ 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?

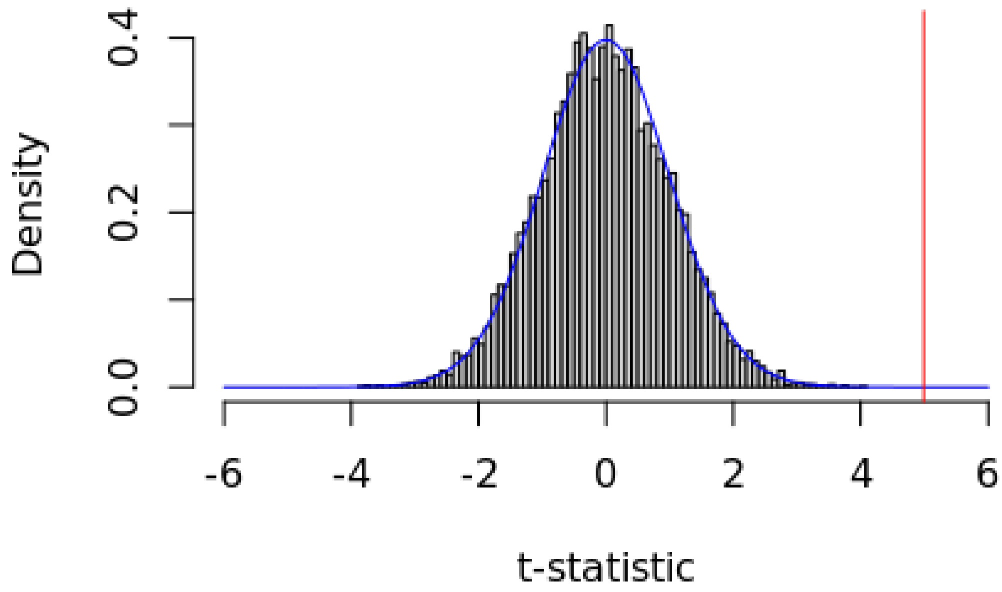
1. Modify the original sample by adding a constant to all data points to make the sample mean equal to the null hypothesis parameter value
   - > data_sample - mean(data_sample) + 8

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

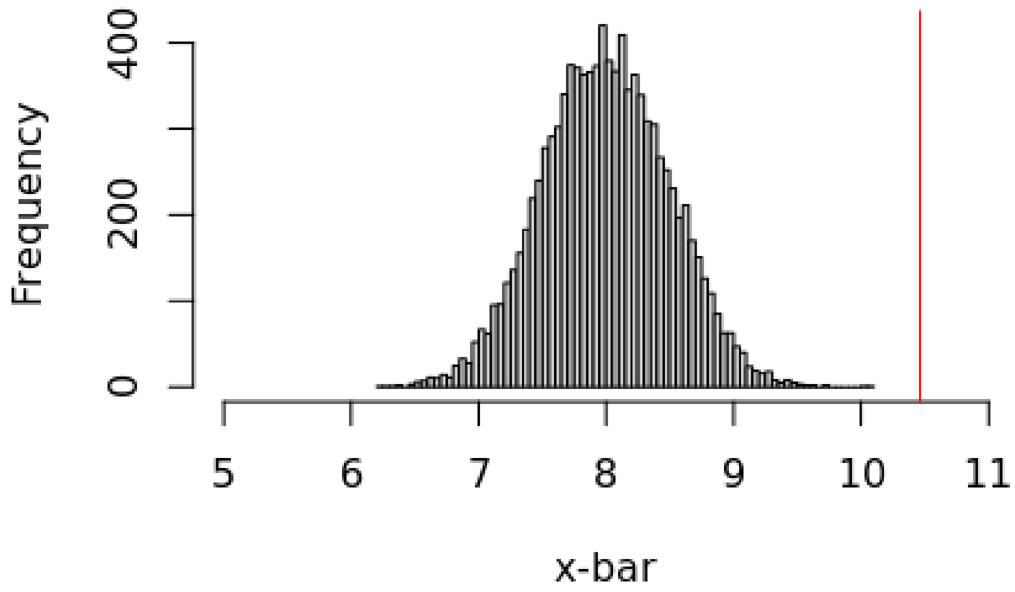
3. Repeat 10,000 times

Null distributions

Null distribution using a t-statistic

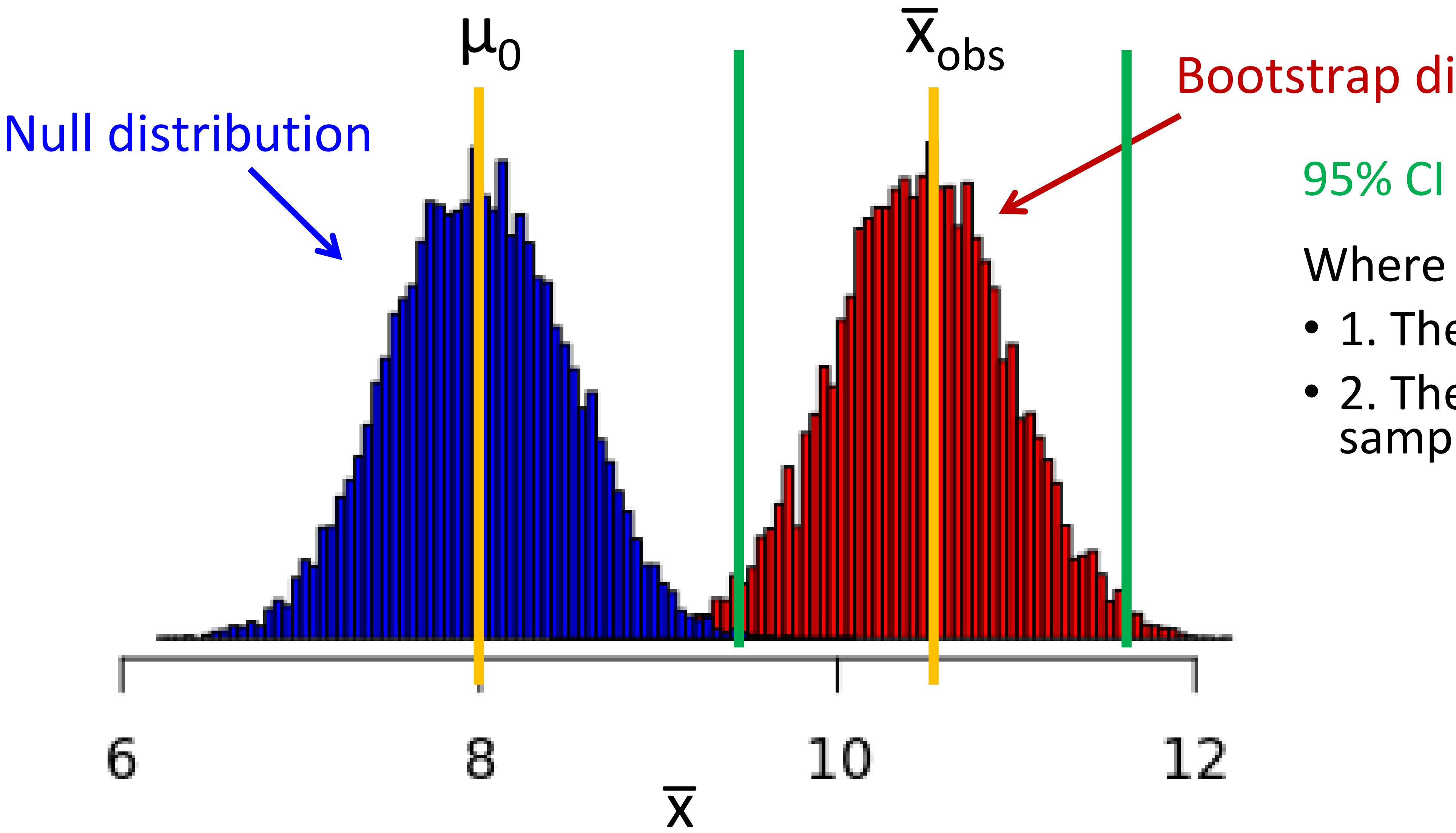


Null distribution using $\bar{x}$ statistic



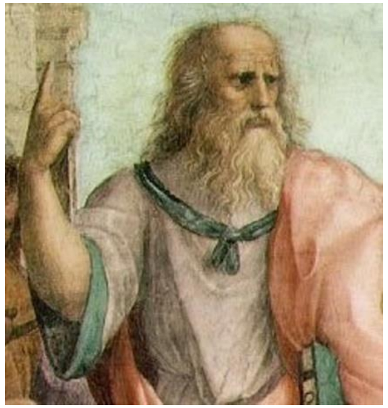
The p-value in both cases is…

Relationship between null and bootstrap distributions



Where is:

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



Next class: start on the tidyverse…