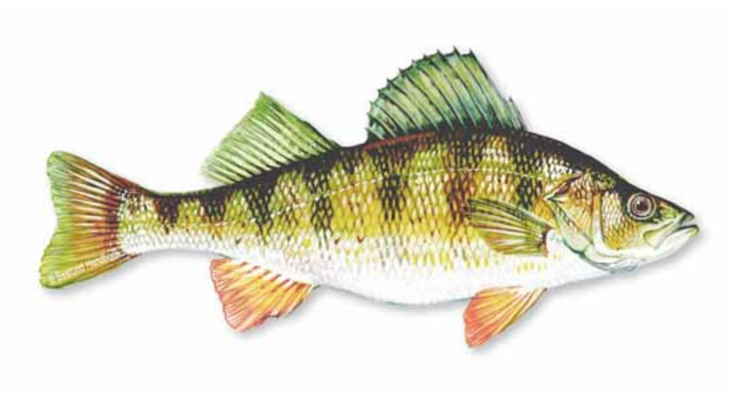
Theories of hypothesis tests and hypothesis tests for more than two means



Overview

Theories of hypothesis testing

Parametric test for more than two means: ANOVA

Conclusions

Announcements

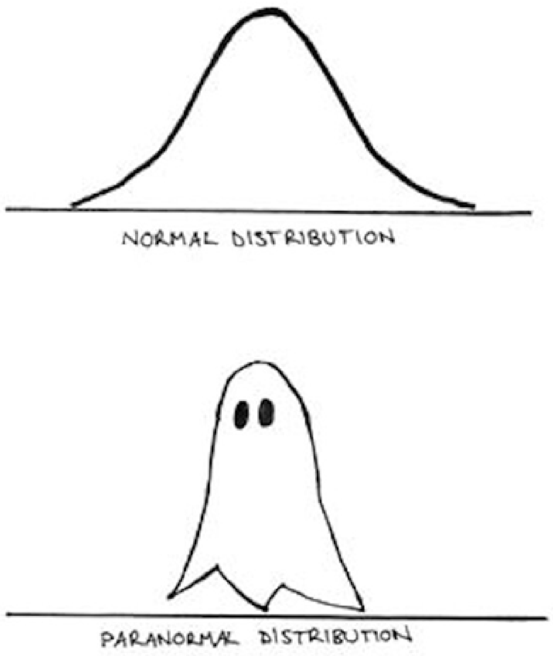
Taz

Review session

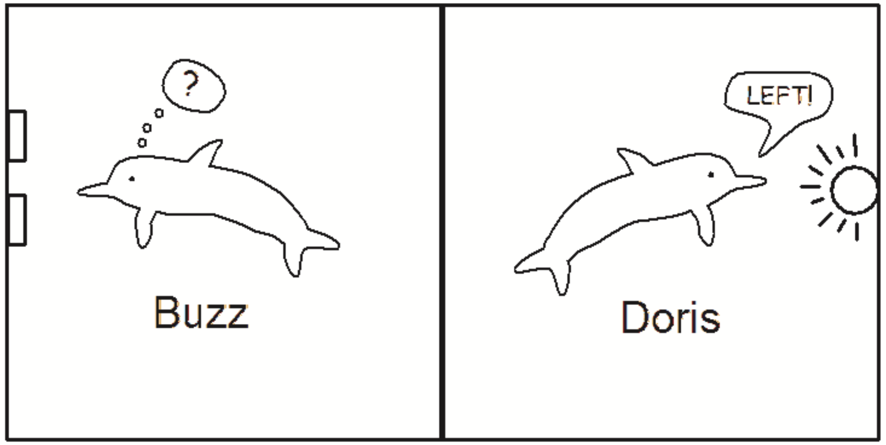
- Next Tuesday at 1:30pm

Final exam: 2pm on Saturday December 14th

In class we have done a number of hypothesis tests



Do more than 25% of Americans believe in ghosts?



Are dolphins capable of abstract communication?

Does eating at late lead to more weight gain?

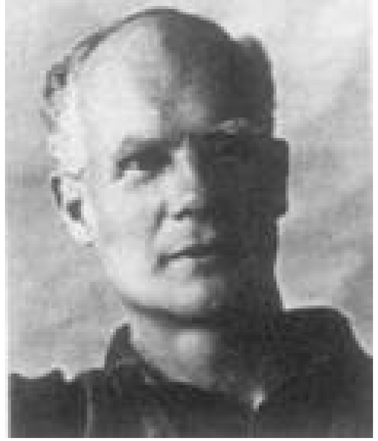
Two theories of hypothesis testing

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)

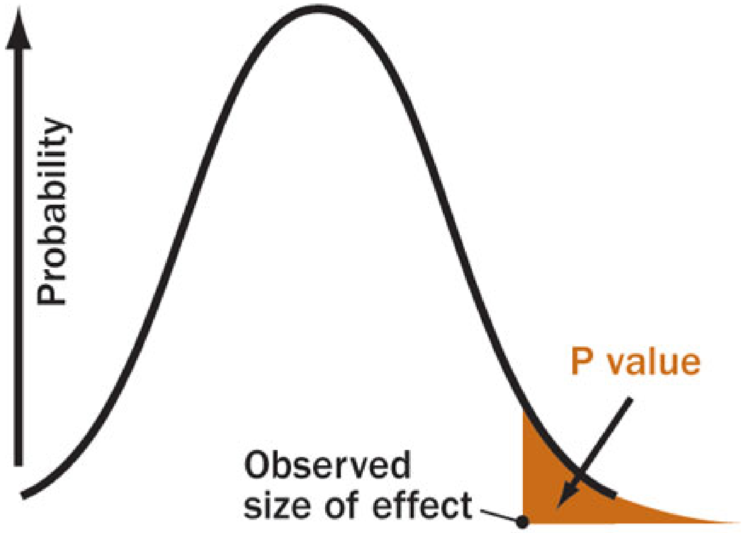


Pearson (1895-1980)

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: tells 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 so extreme is unlikely when H_0 is true

- i.e., reject H0 if the p-value is less than some predetermined **significance level** α

Do not reject H_0: if the statistic is not too extreme when H_0 is true. This means the test is inconclusive.

Frequentist logic

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

If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, then only ~5% of all published research findings should be wrong (for α = 0.05)

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

Frequentist logic

Type 2 error: incorrectly rejecting failing to reject H_0 when it is false

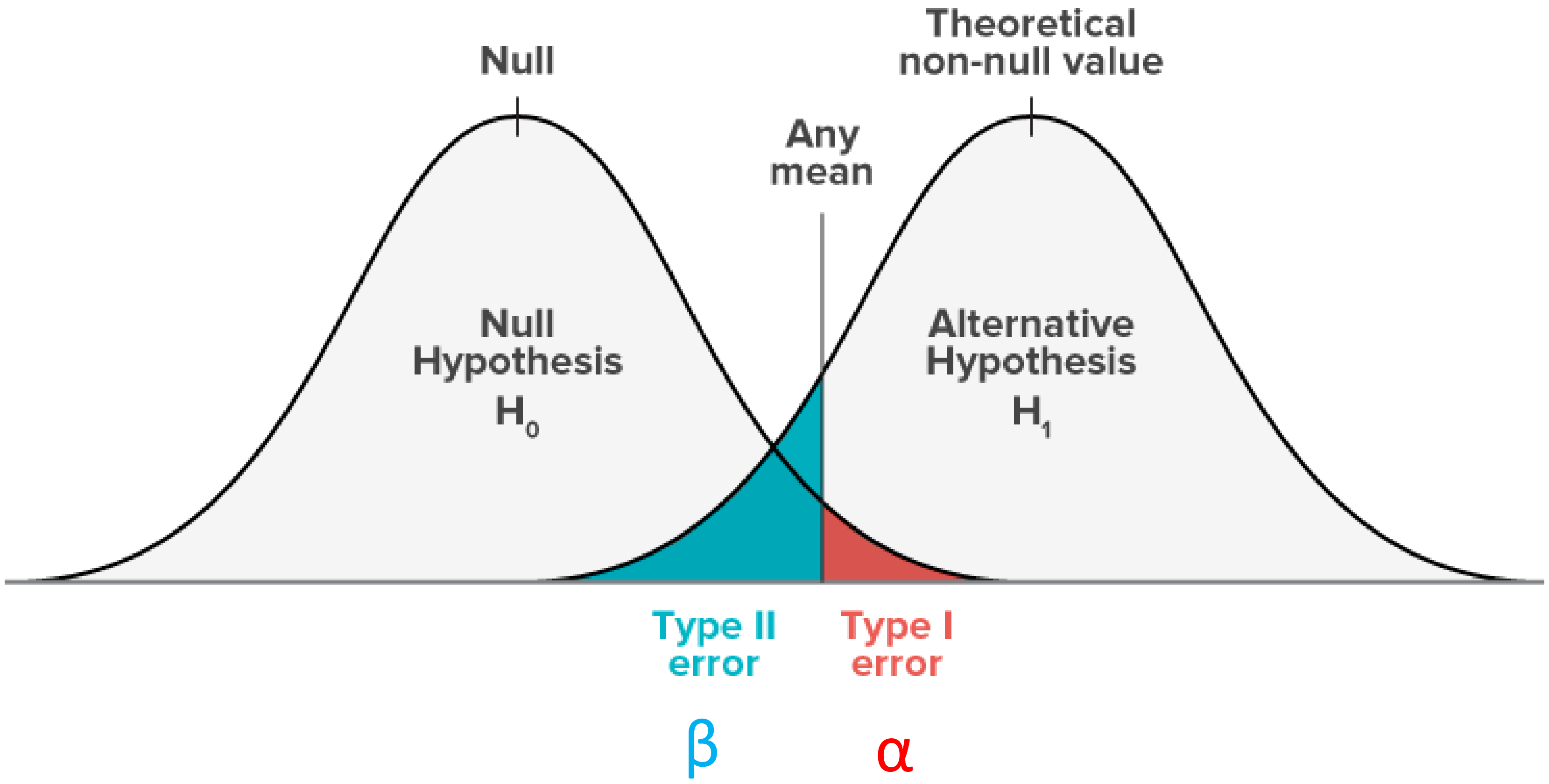
- The rate at which we make type 2 errors is often denoted with the symbol β

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

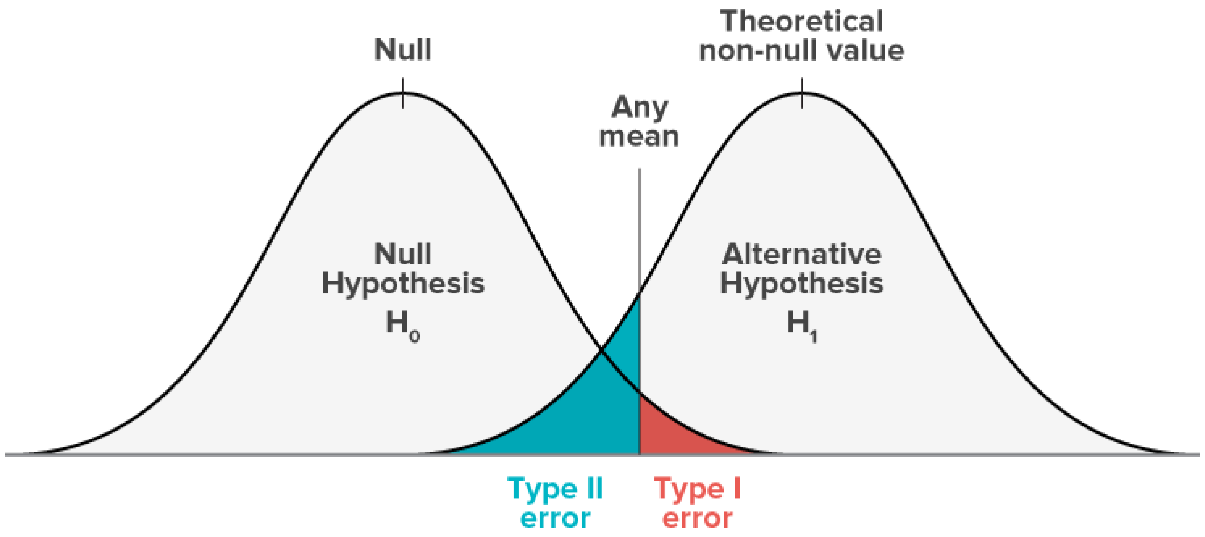
- 1 - β

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

Type I and Type II Errors

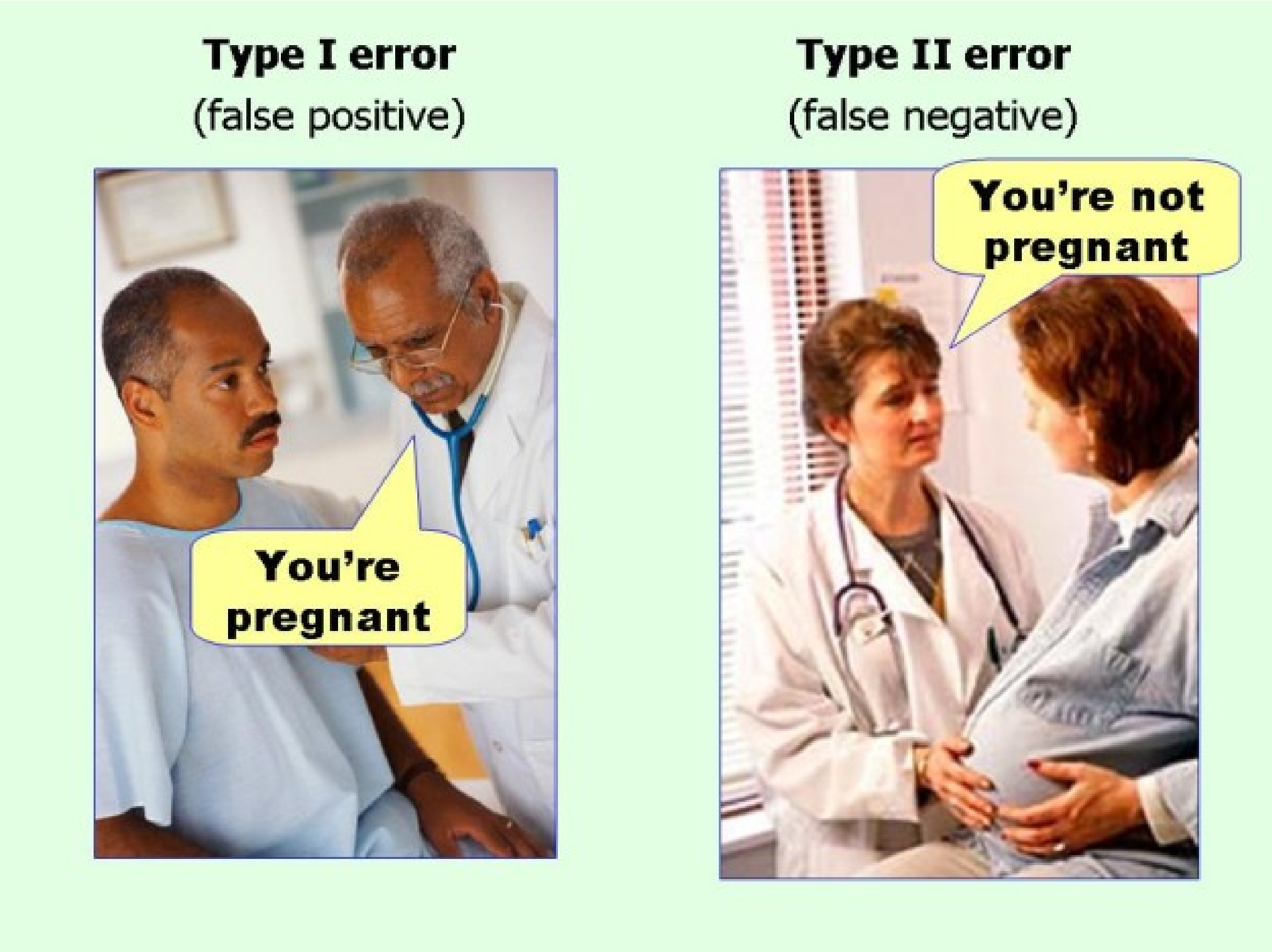


Type I and Type II Errors



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

Type I and Type II Errors



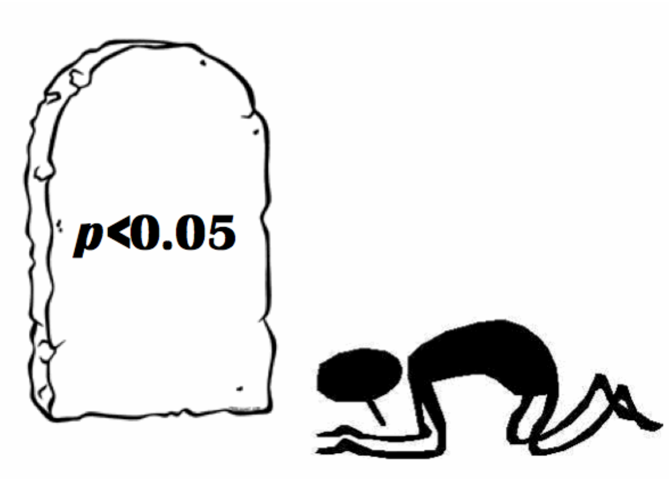
Problems with the NP hypothesis tests

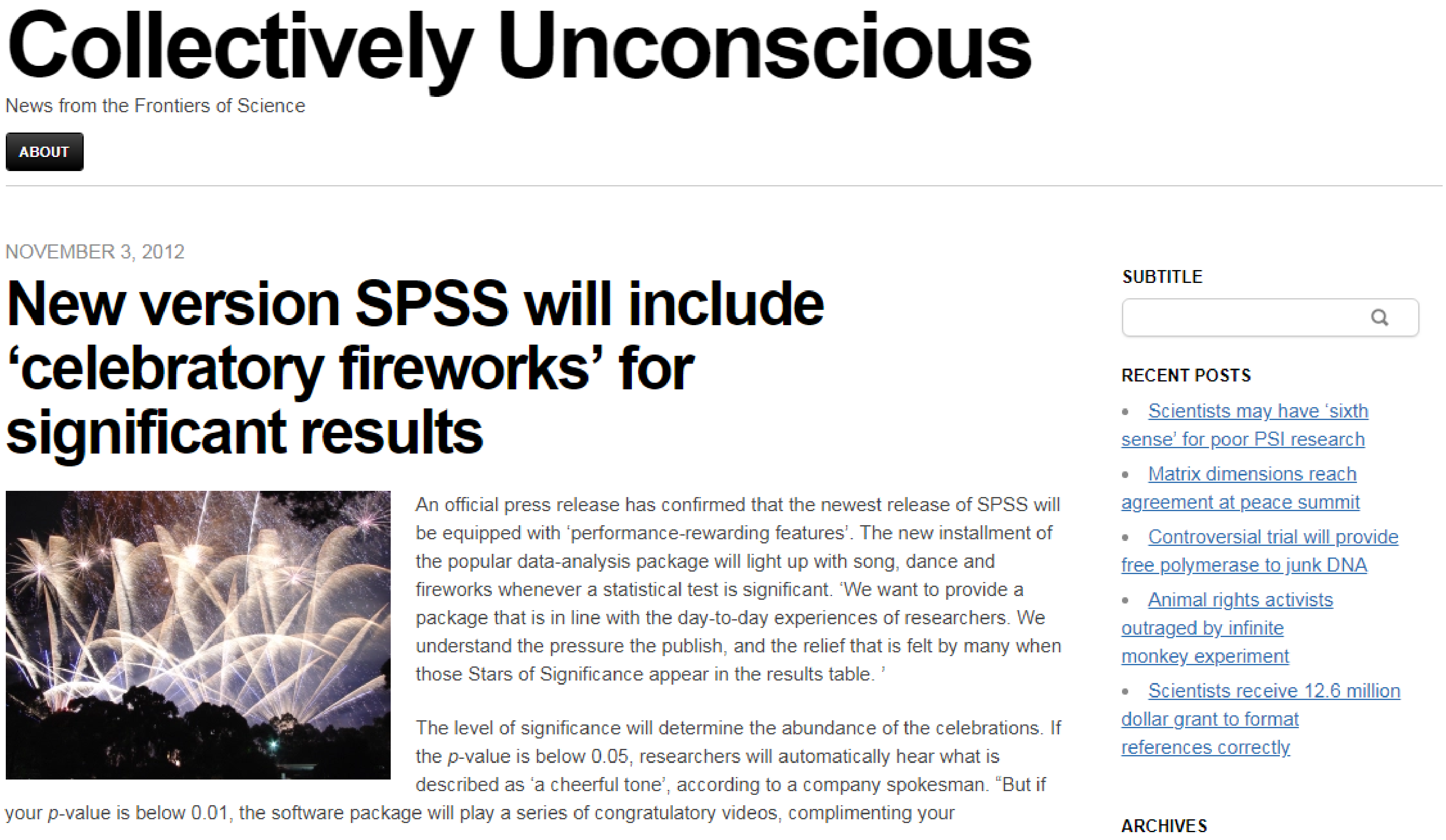
Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- E.g., 95% of these statements are true:
  - Calcium is good for your heart, Paul is psychic, Buzz and Doris can communicate, …

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

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

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

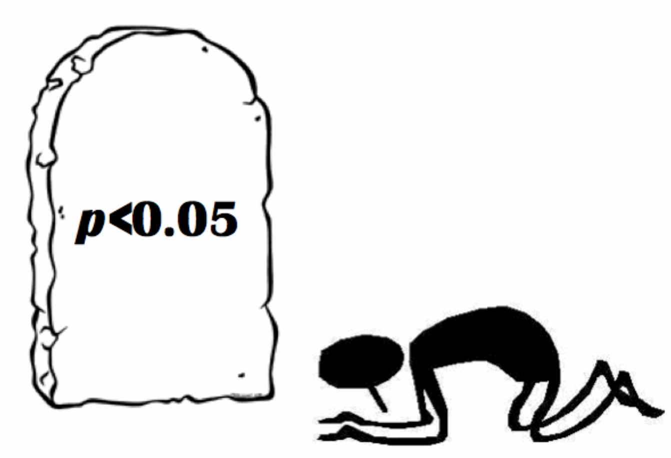
Problems with the NP hypothesis tests

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 true:
    - Calcium is good for your heart, Paul is psychic, Buzz and Doris can communicate, …

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 1 errors

Genes and leukemia example

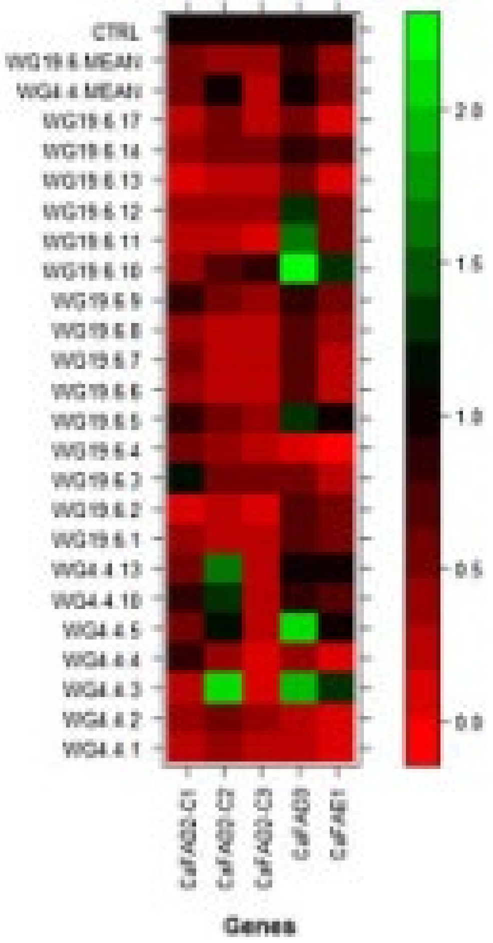
Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

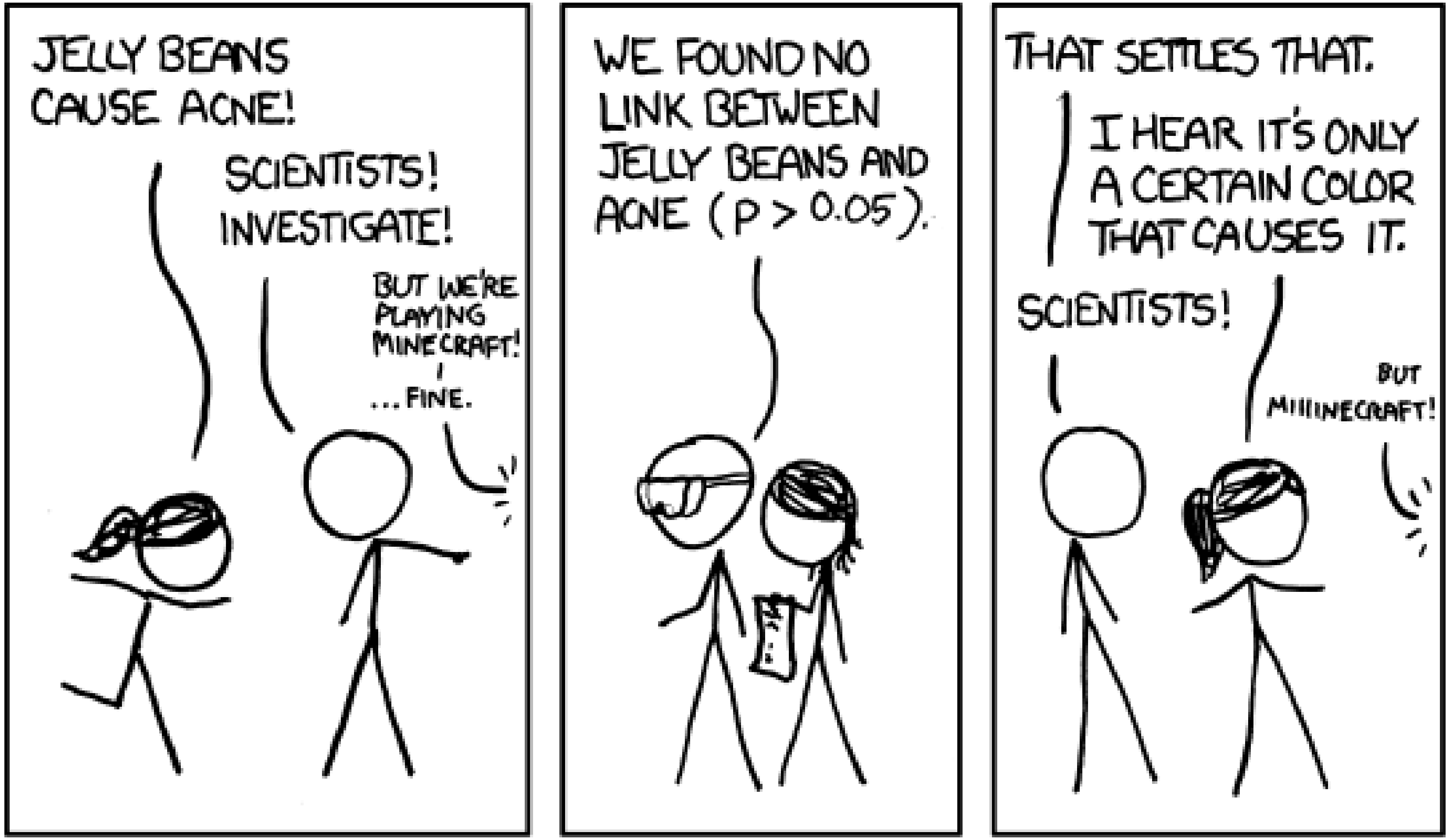
- H_0: $\mu_{L1} = \mu_{L2}$ is true for all genes

Q: If each gene was tested separately using a significance level of $\alpha = 0.05$, approximately how many type 1 errors would be expected?

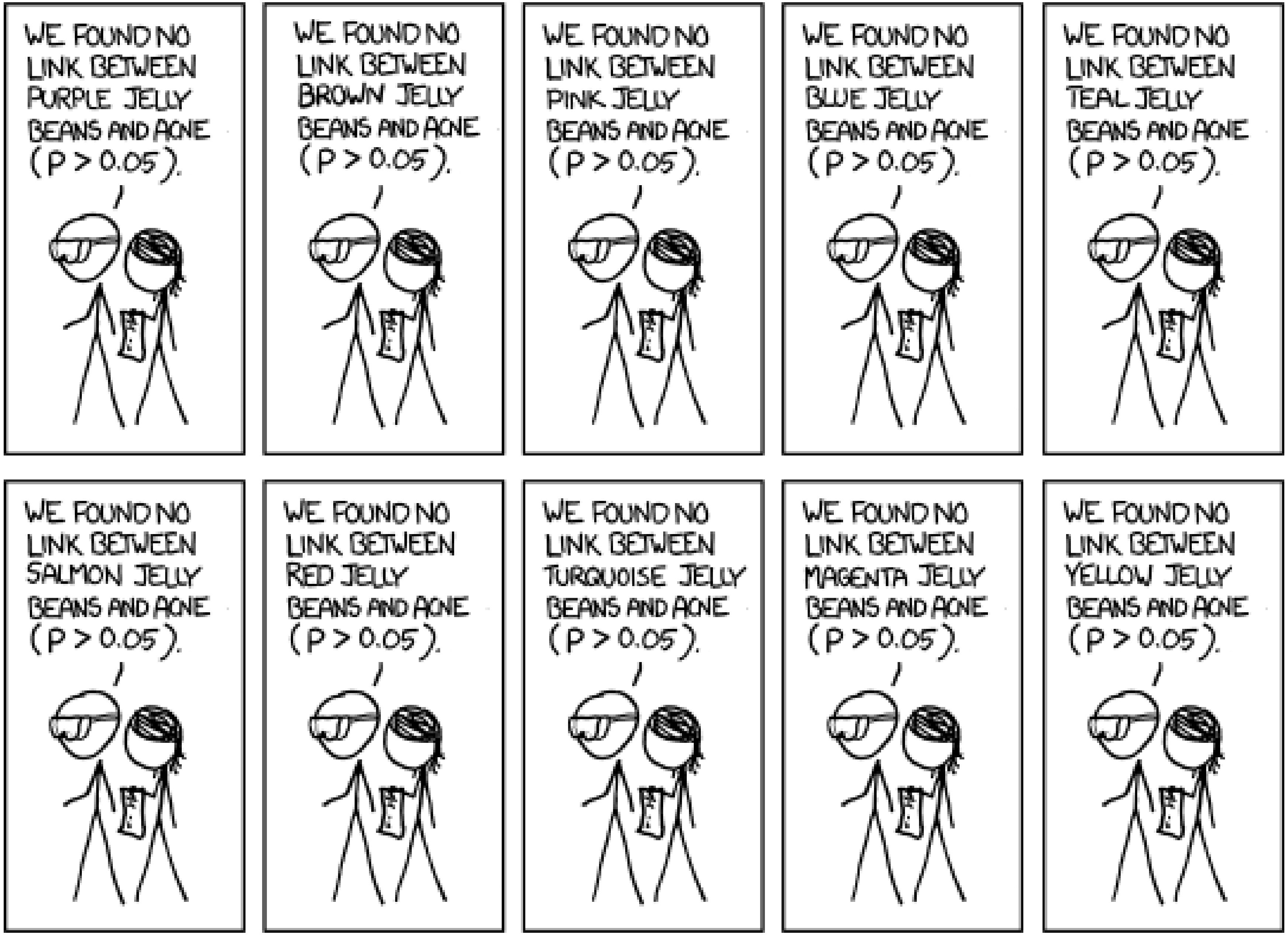
- A: 7129 x 0.05 = 356

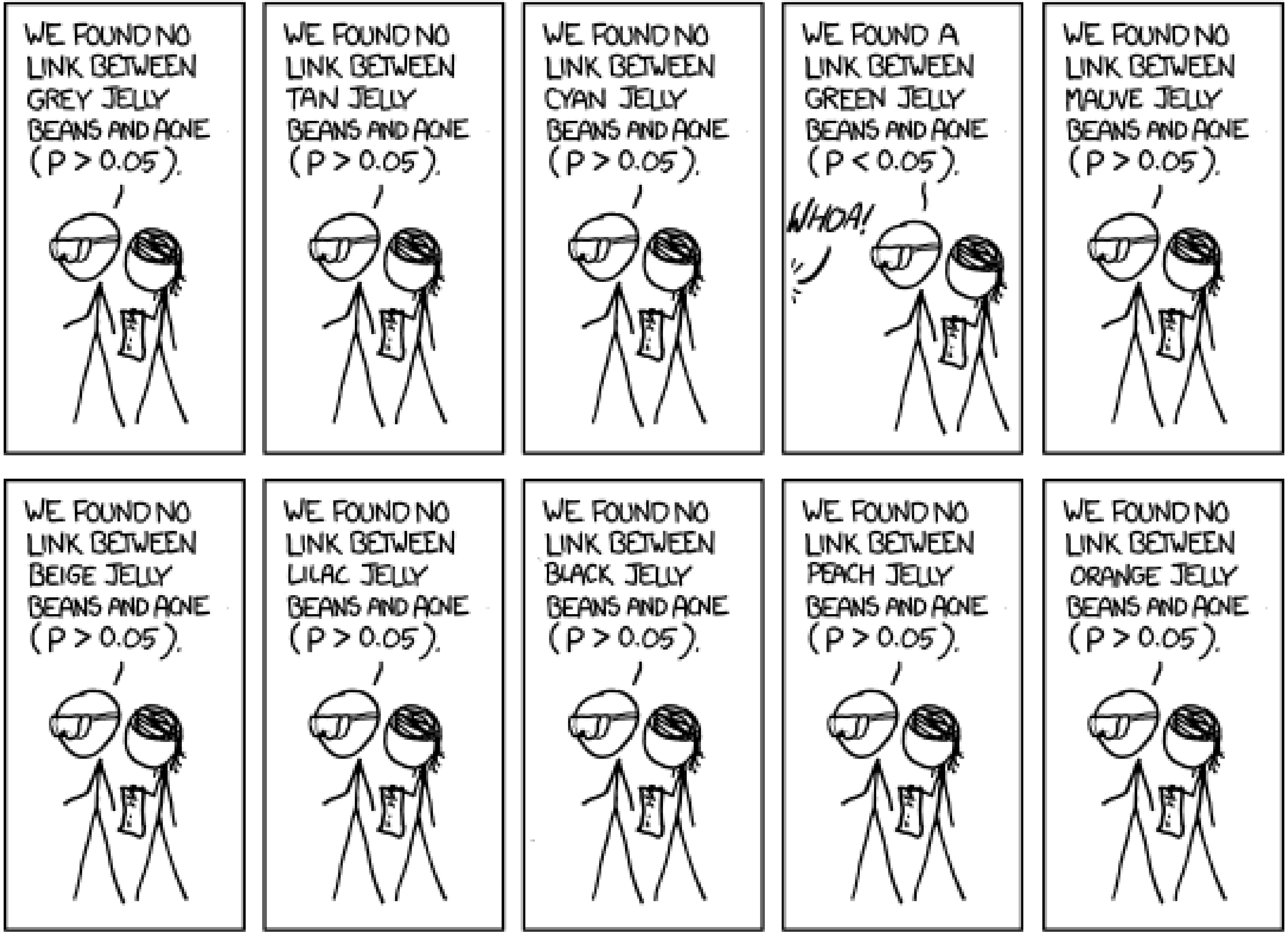


Multiple hypothesis tests


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!

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 TEAL 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 YELLOW 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 MAUVE JELLY BEANS AND ACNE (P > 0.05).
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).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

News
GREEN JELLY
S LINKED
95% CONF
ONLY 5% CHANCE OF COINCIDENCE!
SCIENT

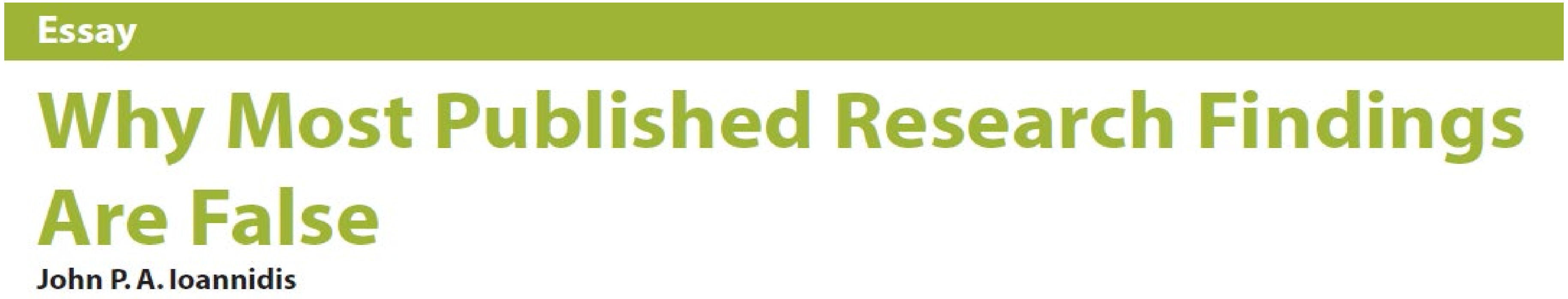
Don’t ever do this!

http://xkcd.com/882/

The problem of multiple testing

For α = 0.05, ~5% of all published research findings should be wrong

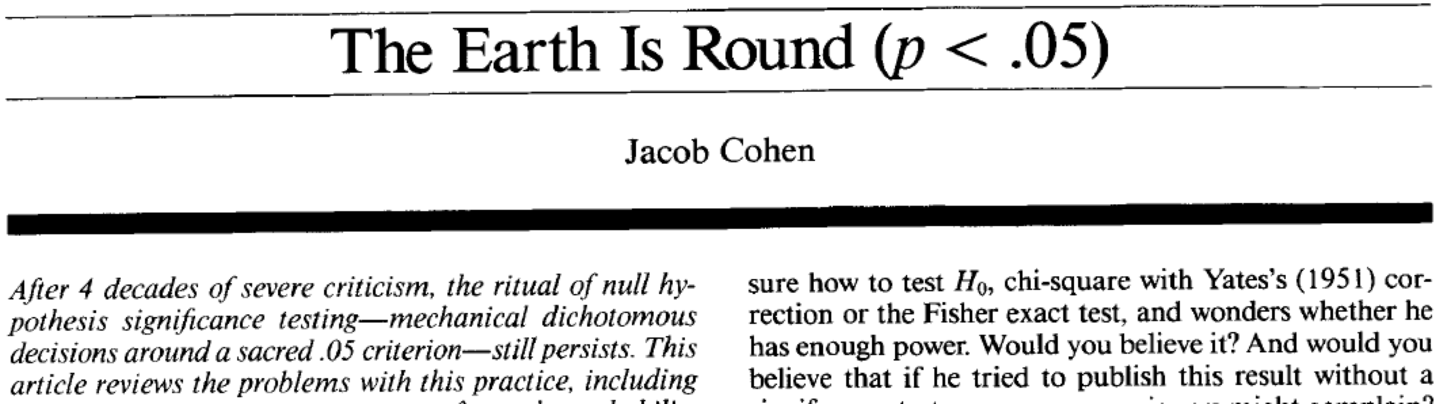
Publication bias (file drawer effect): Generally positive results are more likely to be published, so if you read the literature, the number of incorrect results (type 1 errors) will be greater than 5%.



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.

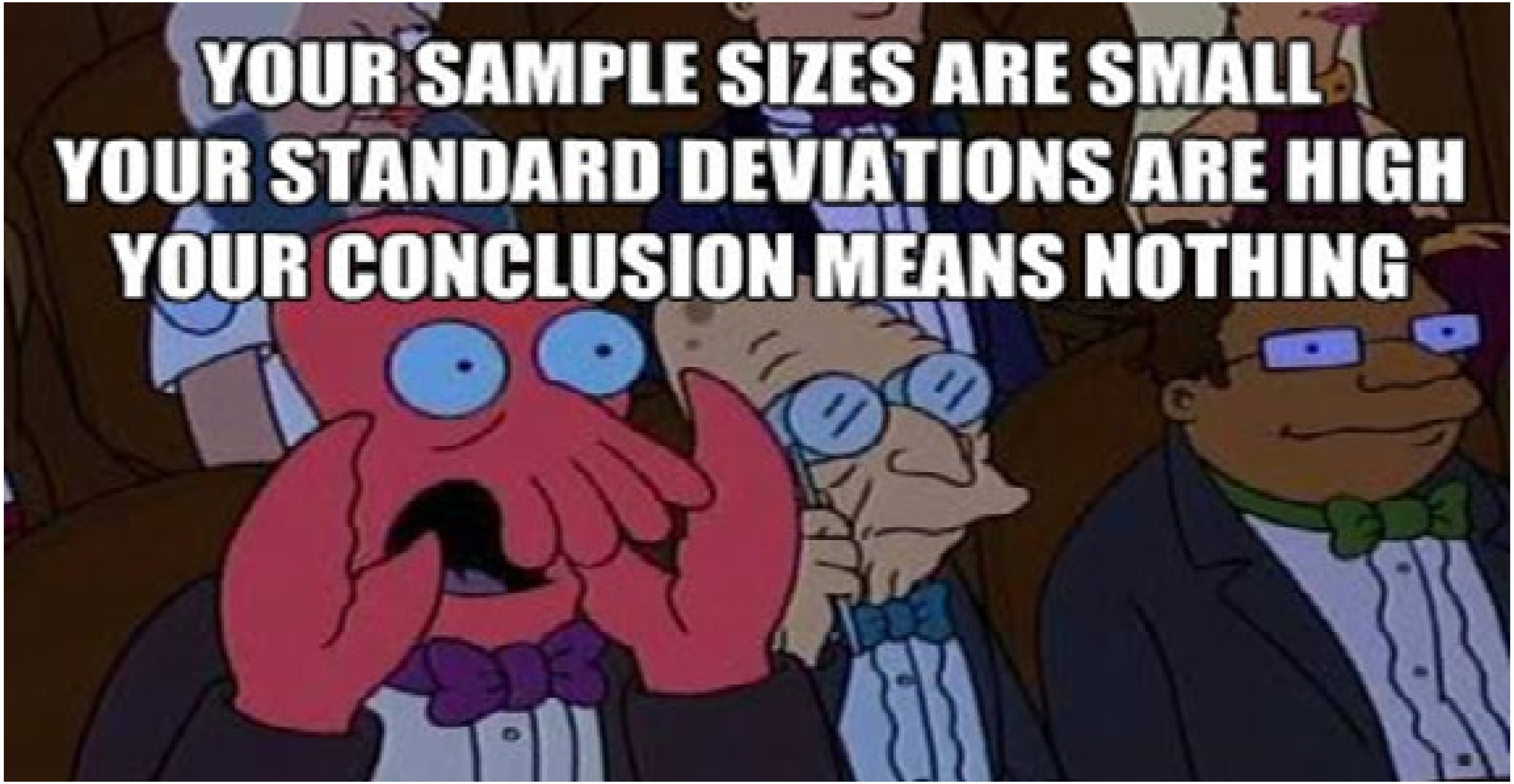
Report effect size in most cases – i.e., confidence intervals

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

Parametric test for comparing more than one mean: One-way ANOVA

An Analysis of Variance (ANOVA) is a test that can be used to examine if a set of means are all the same

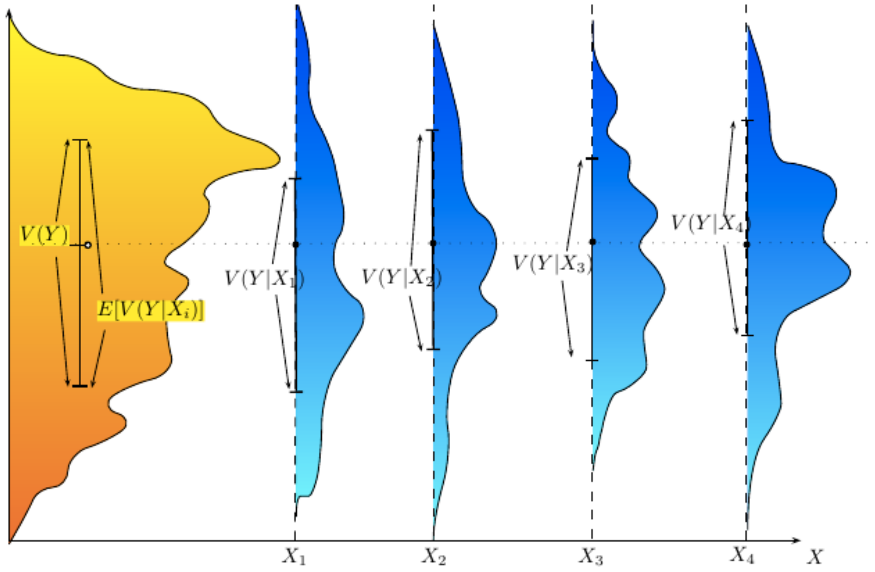
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- $\mu_i \neq \mu_j$ for some i, j

The statistic we use for a one-way ANOVA is the F-statistic

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

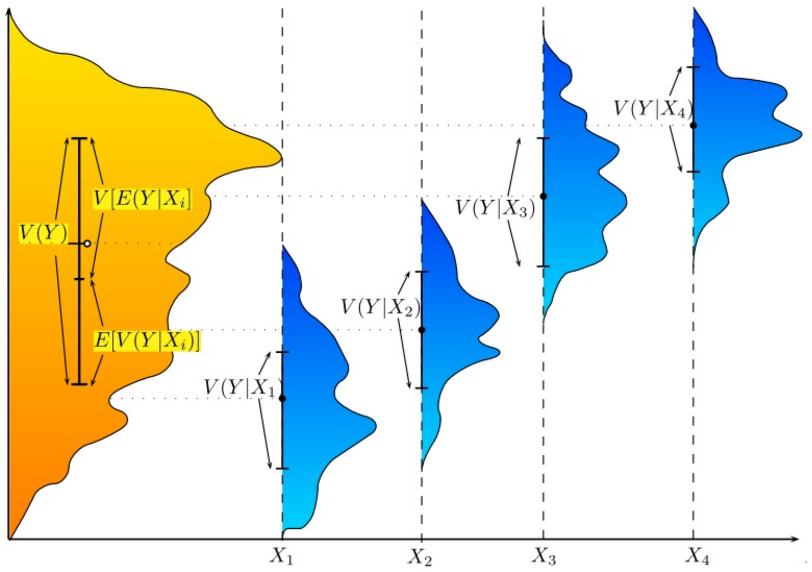
The F-Statistic

If data from all groups came from **the same distribution**



- Similar means $\bar{x}_i$
- Similar spread s_i

If data from all groups came from **different distributions**



- Different means $\bar{x}_i$
- Smaller spreads s_i

The F-statistic

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

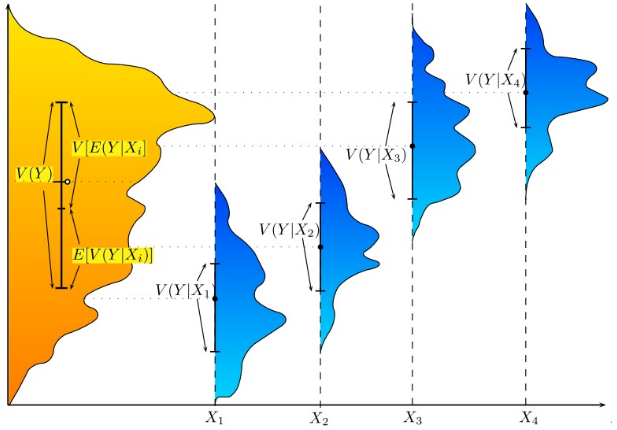
K: the number of groups

n_i: the number of points in group i

x_{ij} :the jth data point from group i

$\bar{x}_i$: the mean of group i

$\bar{x}_{tot}$: the mean across all the data



When the null hypothesis is true, F has a value around 1

- The numerator and denominator are both estimate of σ^2

Parametric test for comparing more than one mean: One-way ANOVA

The F-statistic comes from a F-distribution

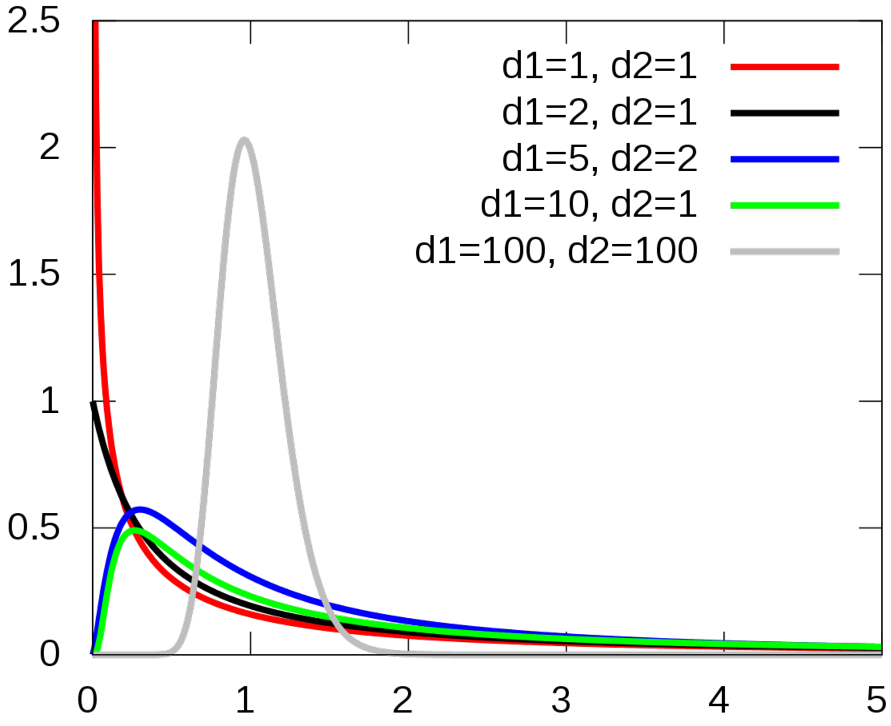
- df1 = K - 1
- df2 = N - K

Assumptions underlying a one-way ANOVA

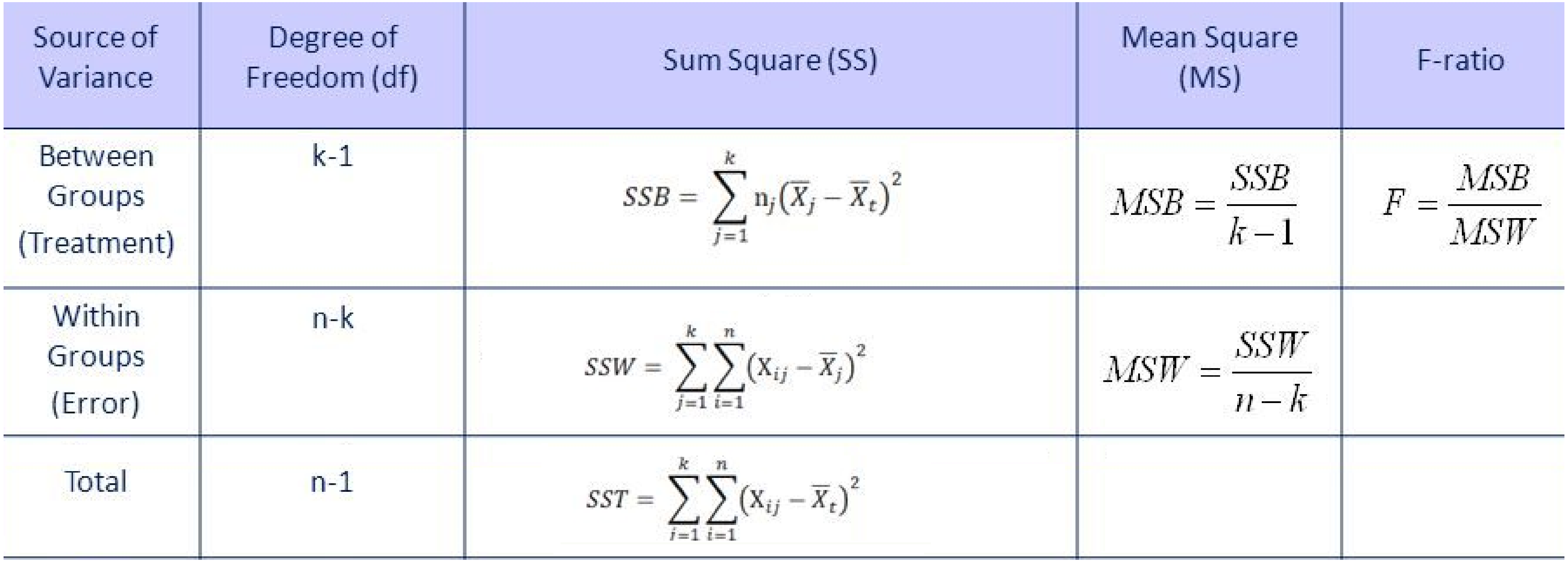
- Data in each group come from normal distributions
- Each group has equal variance (homoskedasticity)

Can check these assumptions by:

- Visualizing data in each group
- Seeing if the ratio of $s_{max}/s_{min} < 2$



One-way ANOVA table



Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k} \mathrm{n}_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \dfrac{SSB}{k-1}$	$F = \dfrac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(\mathrm{X}_{ij} - \overline{X}_j)^2$	$MSW = \dfrac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{i=1}^{k}\sum_{i=1}^{n}(\mathrm{X}_{ij} - \overline{X}_t)^2$		

k: number of groups

n: number of data points

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

One-way ANOVA in R

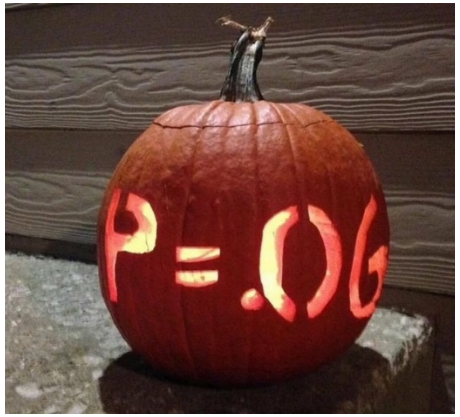
Let’s briefly look at how to use R’s built in functions to do a one-way ANOVA…

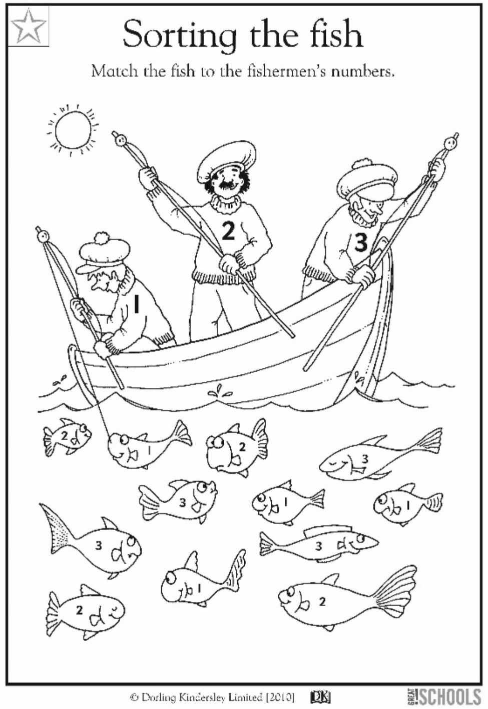
Also, check out this interactive tutorial on applying ANOVAs to neural data created by Brooke Fitzgerald

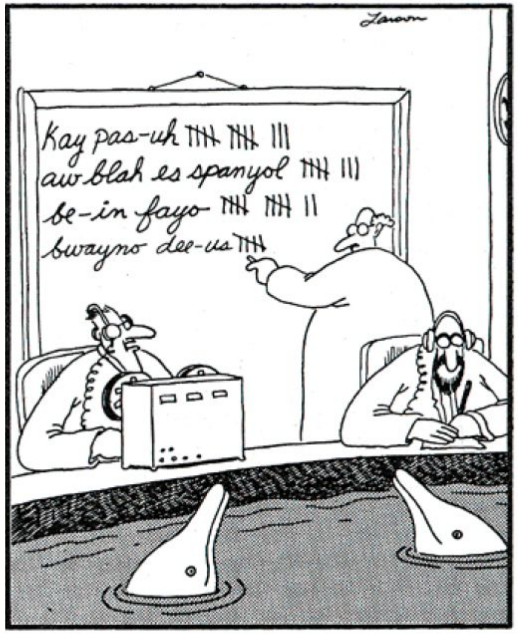
https://neuraldata.net/

One last question…

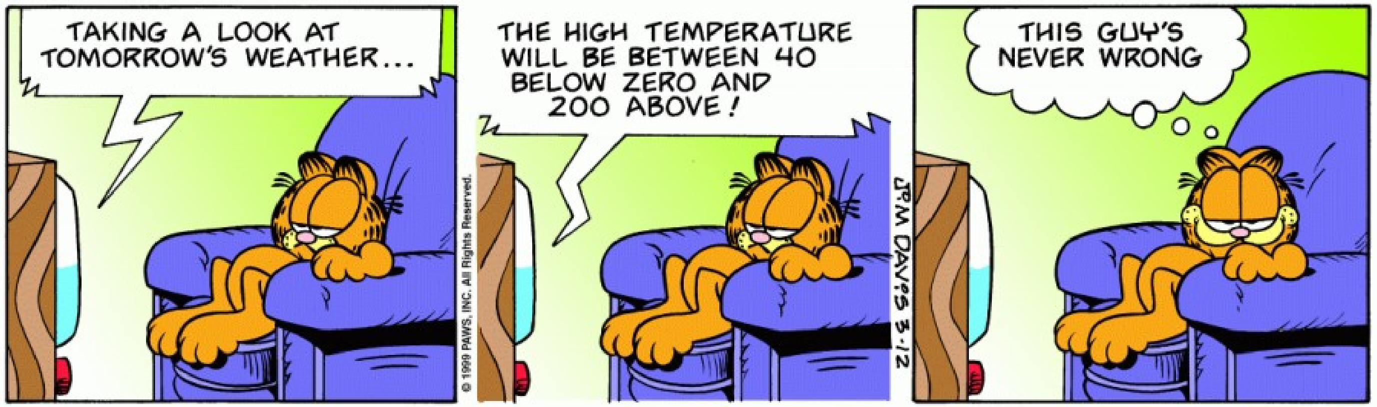
What was the worst joke of the semester?




Sorting the fish
Match the fish to the fishermen's numbers.
1
2
3
© Dorling Kindersley Limited [2010] DK
GREATSCHOOLS



"Matthews ... we're getting another one of those strange 'aw blah es span yol' sounds."


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

Good luck studying for the exam and have a good winter break!

- Final exam is on Saturday December 14th at 2pm