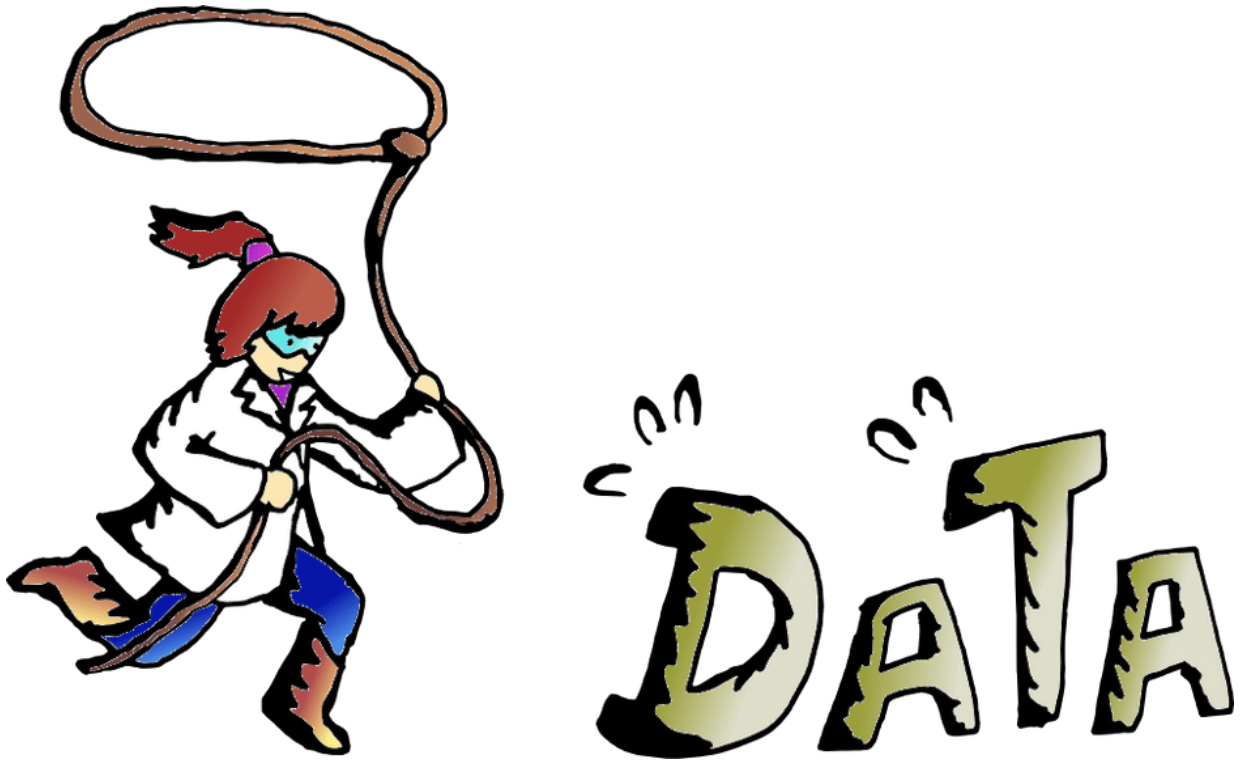
Data wrangling/manipulation



Overview

Review and continuation of theories of hypothesis tests

Data wrangling/manipulation with dplyr

Announcements

A practice midterm exam will be posted soon

- Midterm will be a written exam taken in class on Thursday 10/12

Homework 5 has been posted

- I strongly recommend you do the first two parts prior to next class

Plan for the semester

1	Sep 1	Course overview, introduction to R, descriptive statistics
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R
3	Sep 13-15	Confidence Intervals and the bootstrap
4	Sep 20-22	Review of hypothesis tests and permutation tests in R
5	Sep 27-29	Parametric, non-parametric and theories of hypothesis testing
6	Oct 4-6	Data manipulation and visualization
7	Oct 11-13	Review and midterm exam
8	Oct 18-22	Joining and mapping, October break

Analysis

R

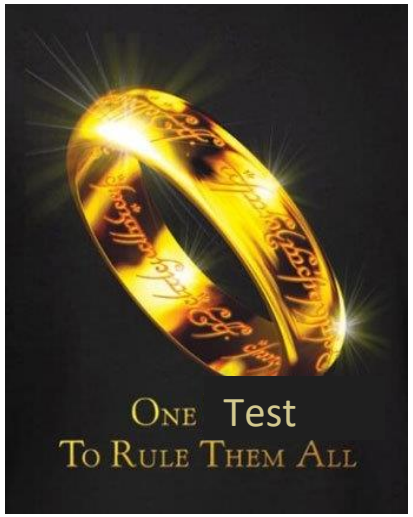
base R

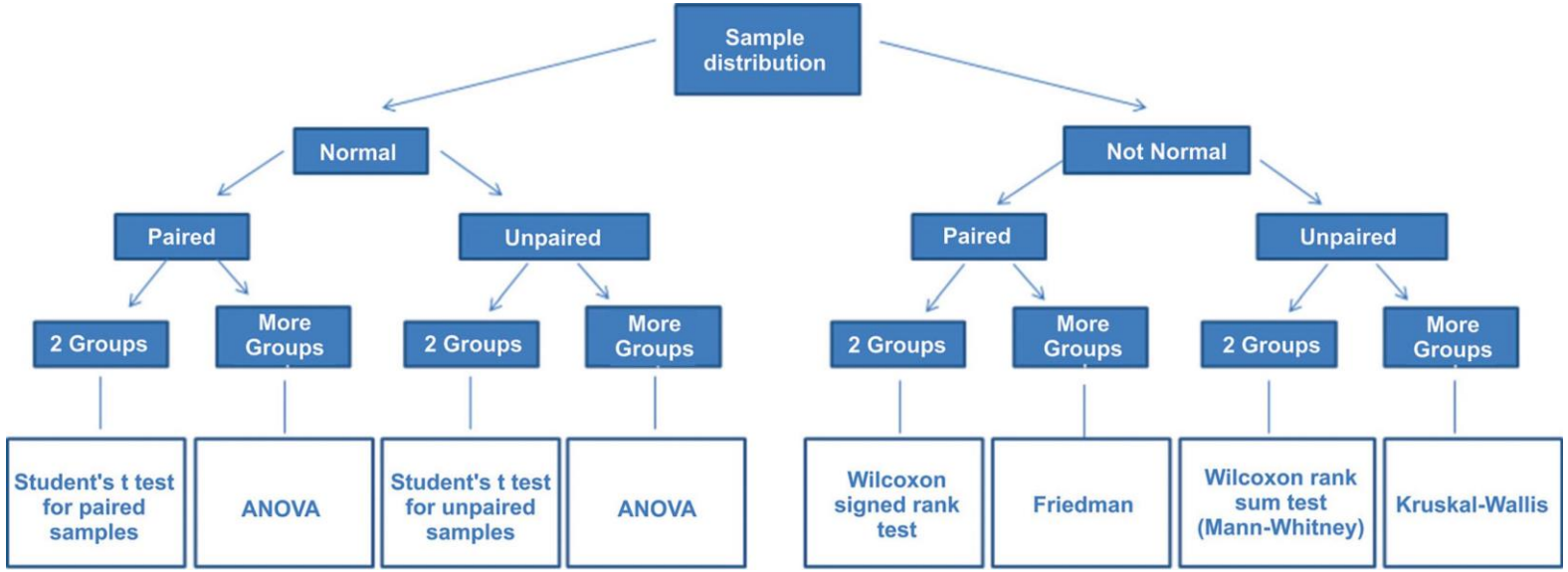
resampling methods

data wrangling visualization

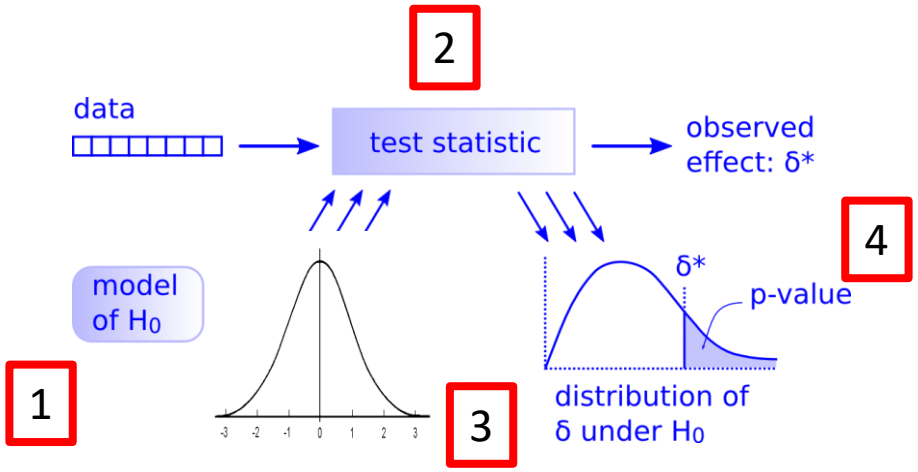
Questions about anything?

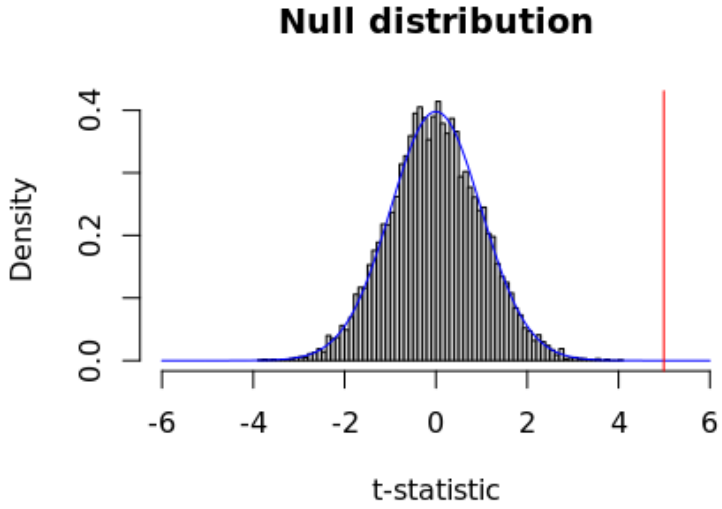
Very quick review



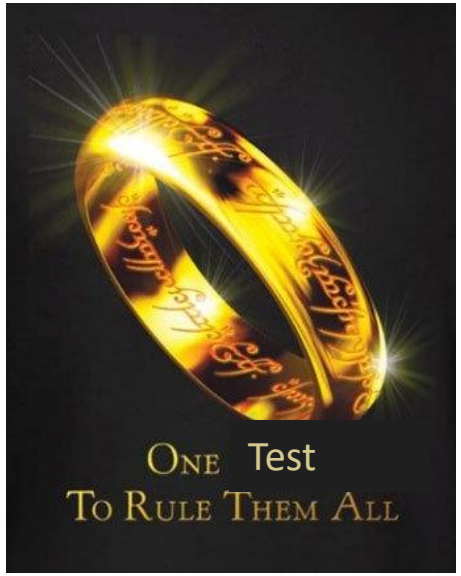


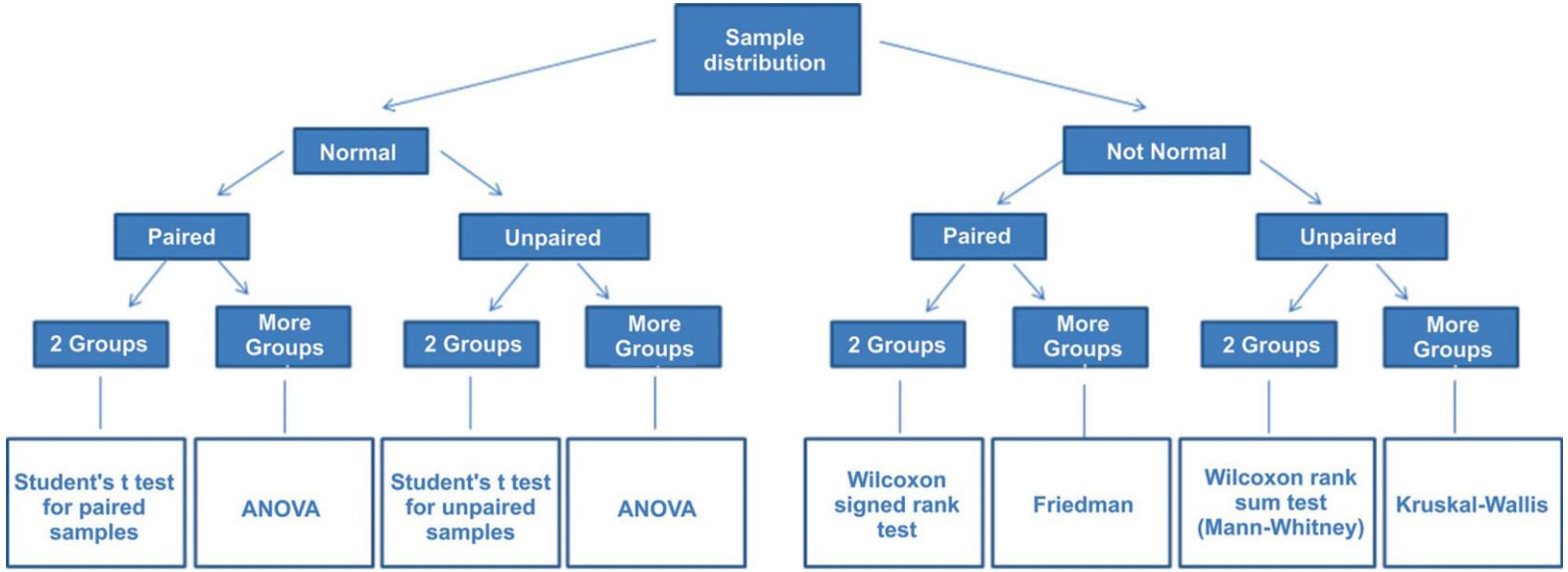
Just need to follow 5 steps!





Very quick review





To select the appropriate parametric test, focus on the parameters being tested in the null hypothesis

- E.g., $H_0: \pi = 0.5$ $H_0: \mu = 0.5$ $H_0: \mu_T = \mu_C$ $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

Parametric tests are derived from particular mathematical assumptions

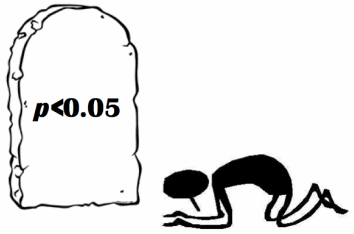
- E.g., data from the two samples comes from normal populations with the same variance
- Some hypothesis tests are "robust" to violations of these assumptions
  - The robustness can be evaluated this through computer simulations

Very quick review: theories of hypothesis testing

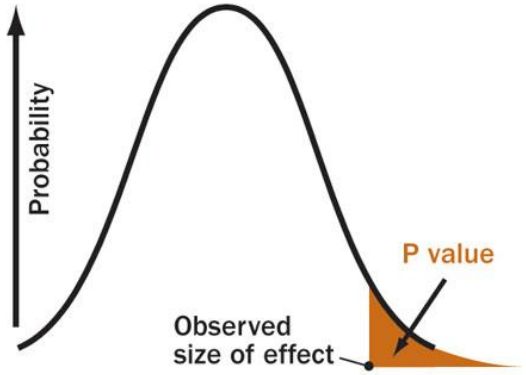
Fisher (1890-1962)

Neyman (1894-1981)

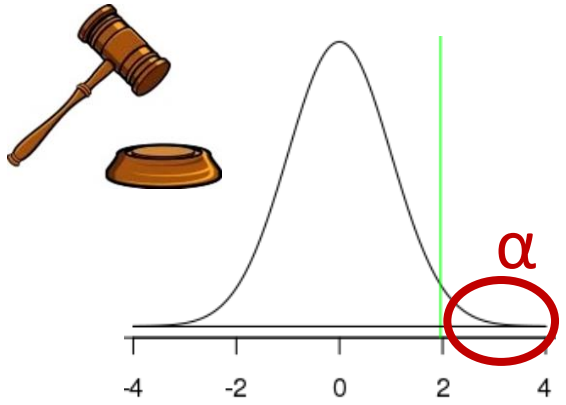
Pearson (1895-1980)

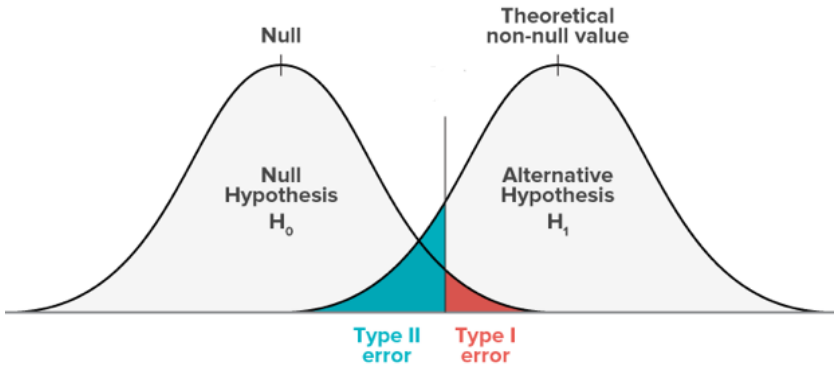


p-value a strength of evidence



Use p-value to make a decision





Neyman-Pearson frequentist logic

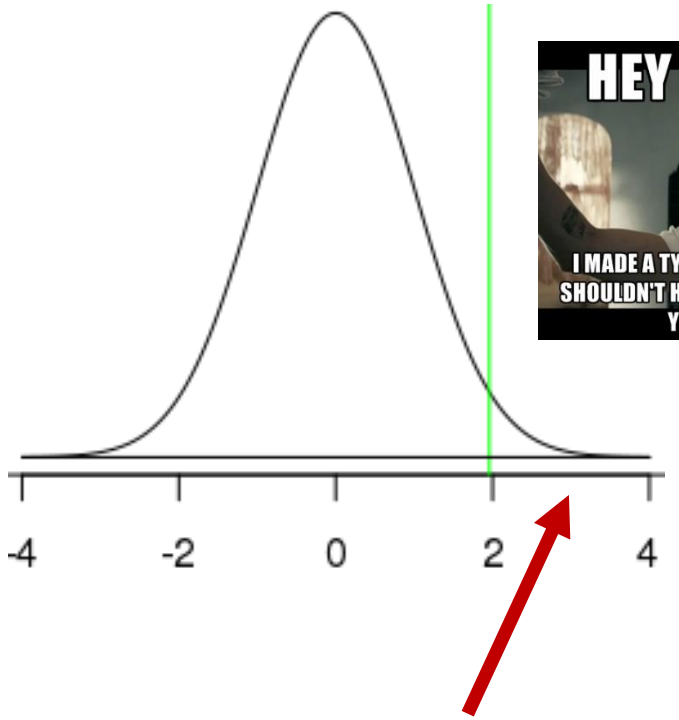
Type I error: incorrectly rejecting the null hypothesis when it is true

If 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

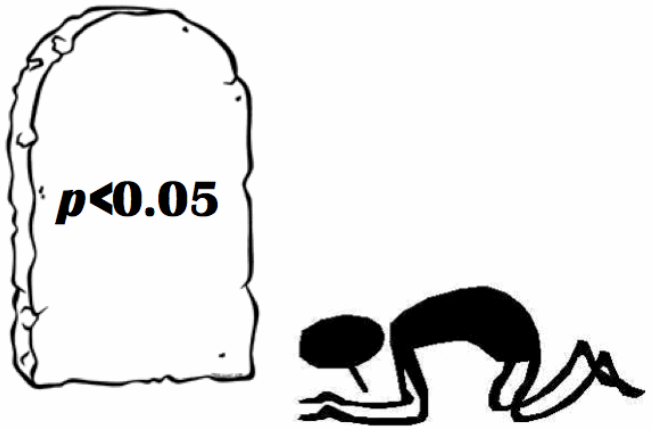
Problems with the NP hypothesis tests

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



Problems with the NP hypothesis tests

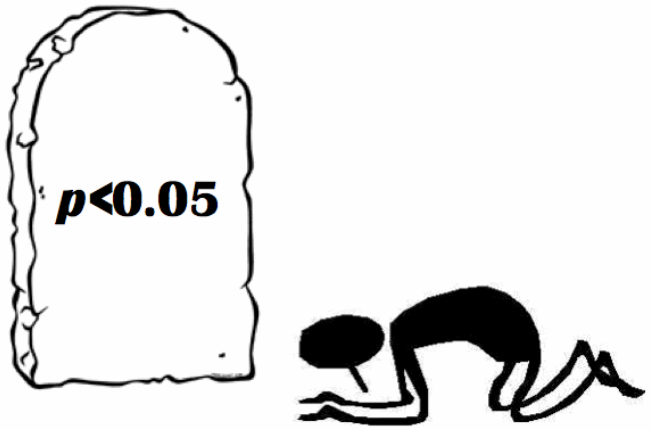
Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

- 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

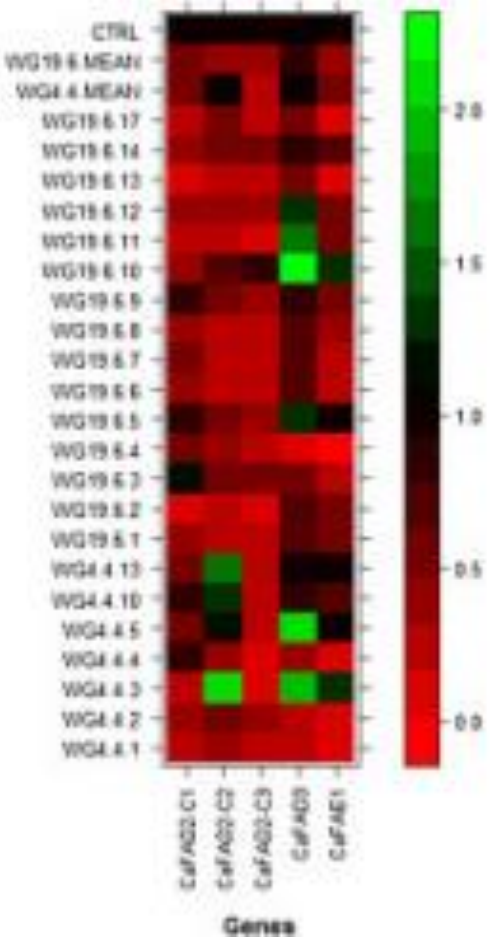
Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

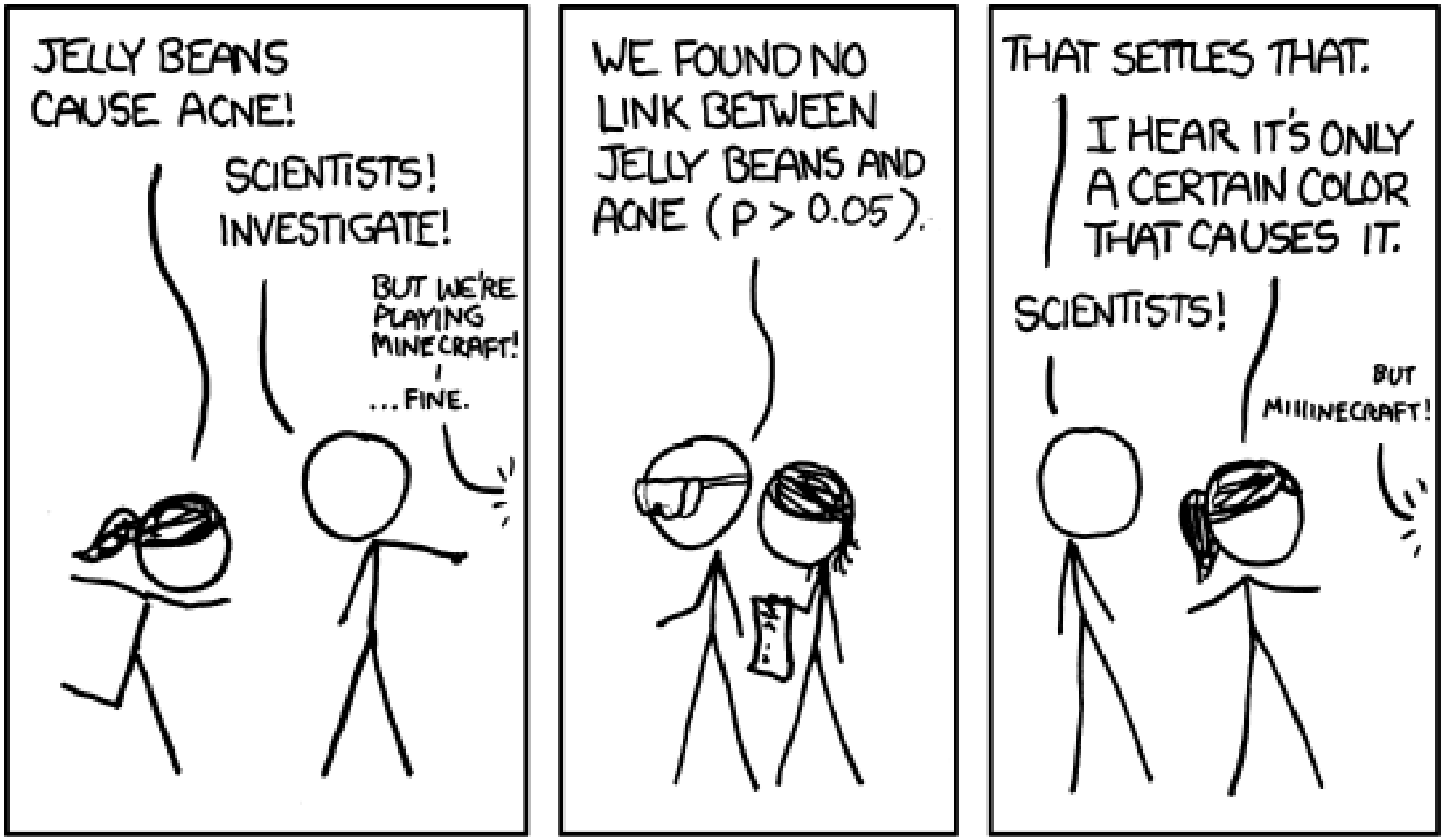
- H_0: $\mu_{L1} = \mu_{L2}$ is true for all genes

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

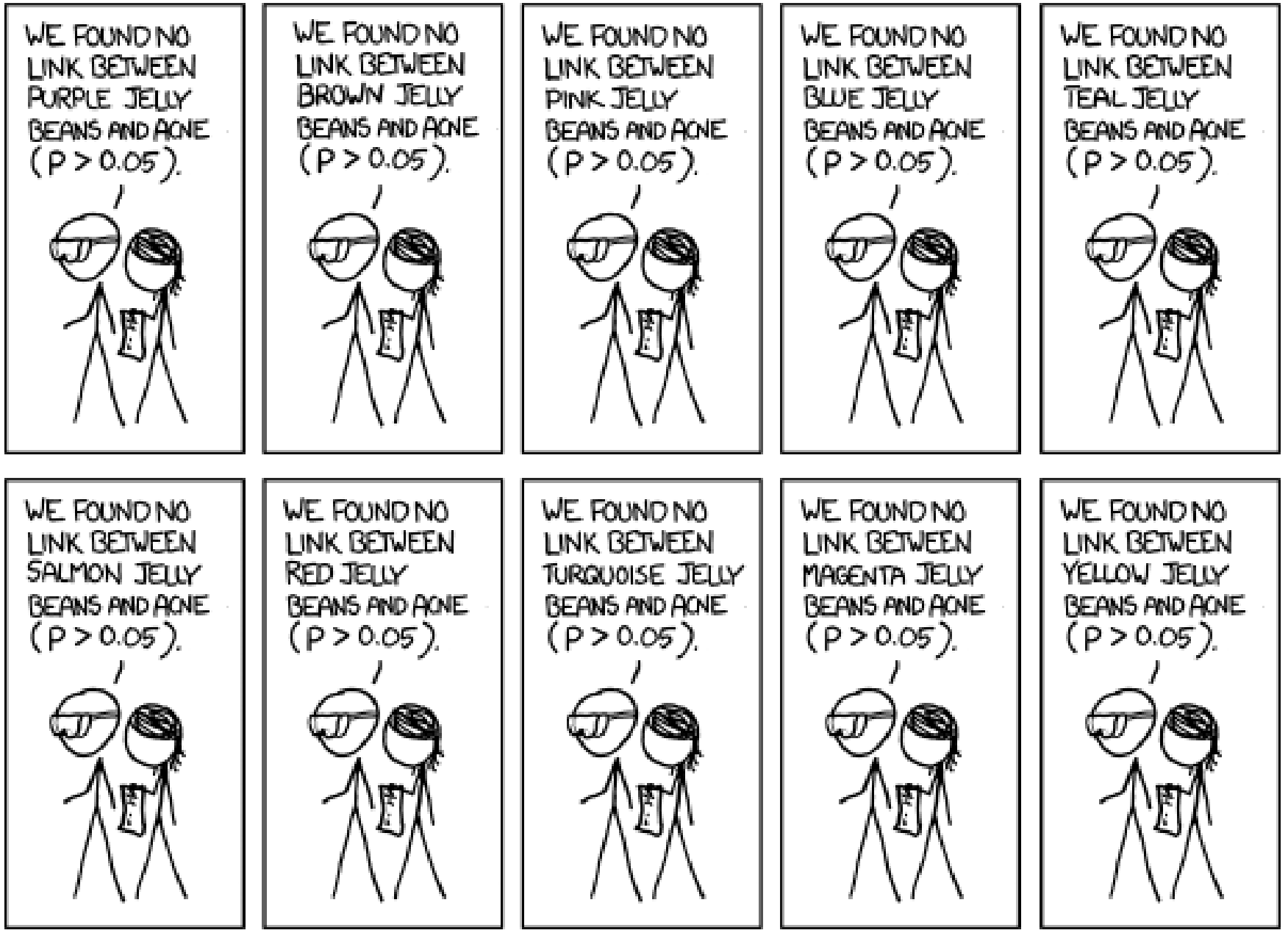
- A: 7129 x 0.05 = 356

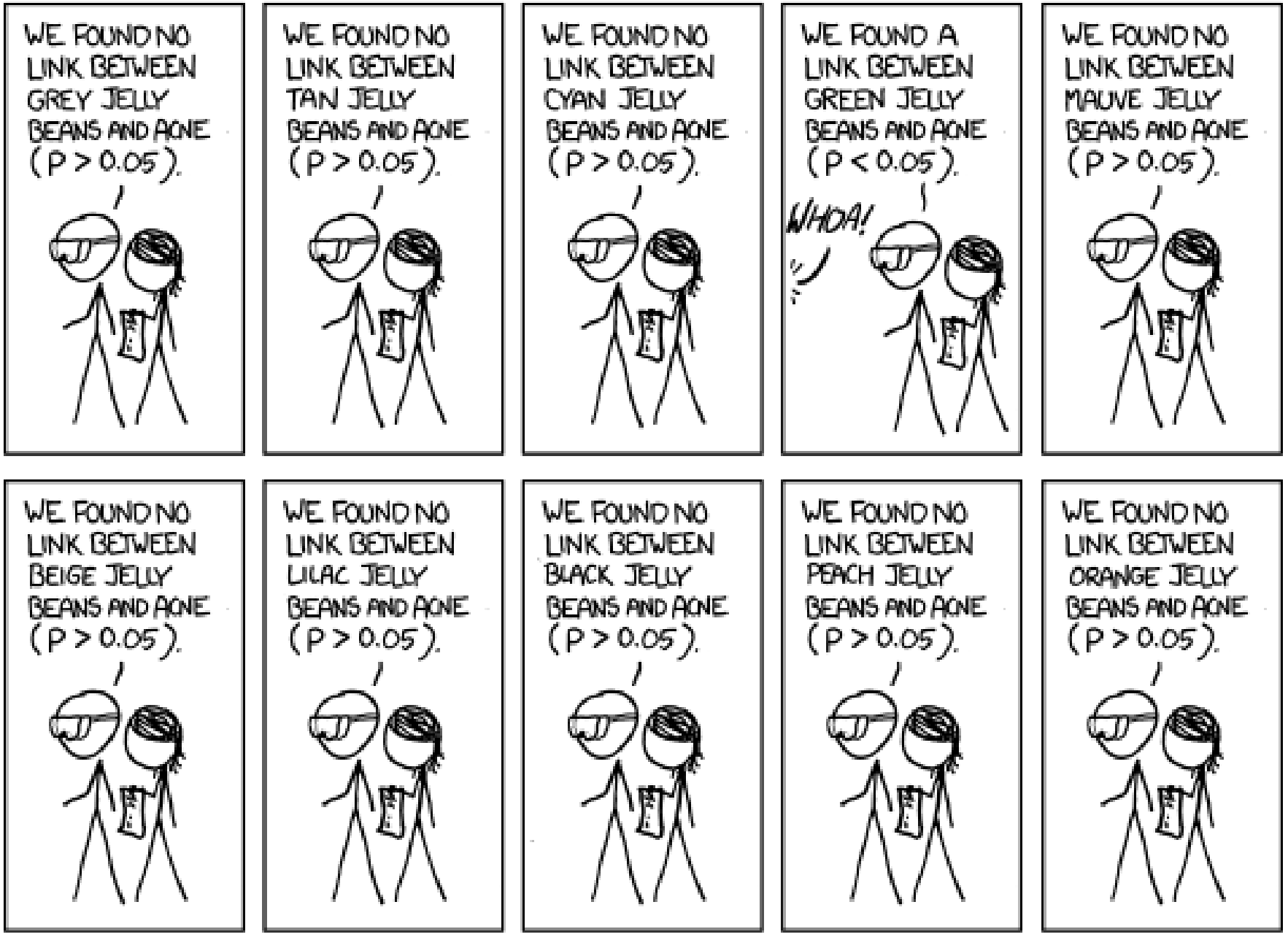


Multiple hypothesis tests



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 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).

Don't ever do this!

http://xkcd.com/882/

The problem of multiple testing

For $\alpha = 0.05$, 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.

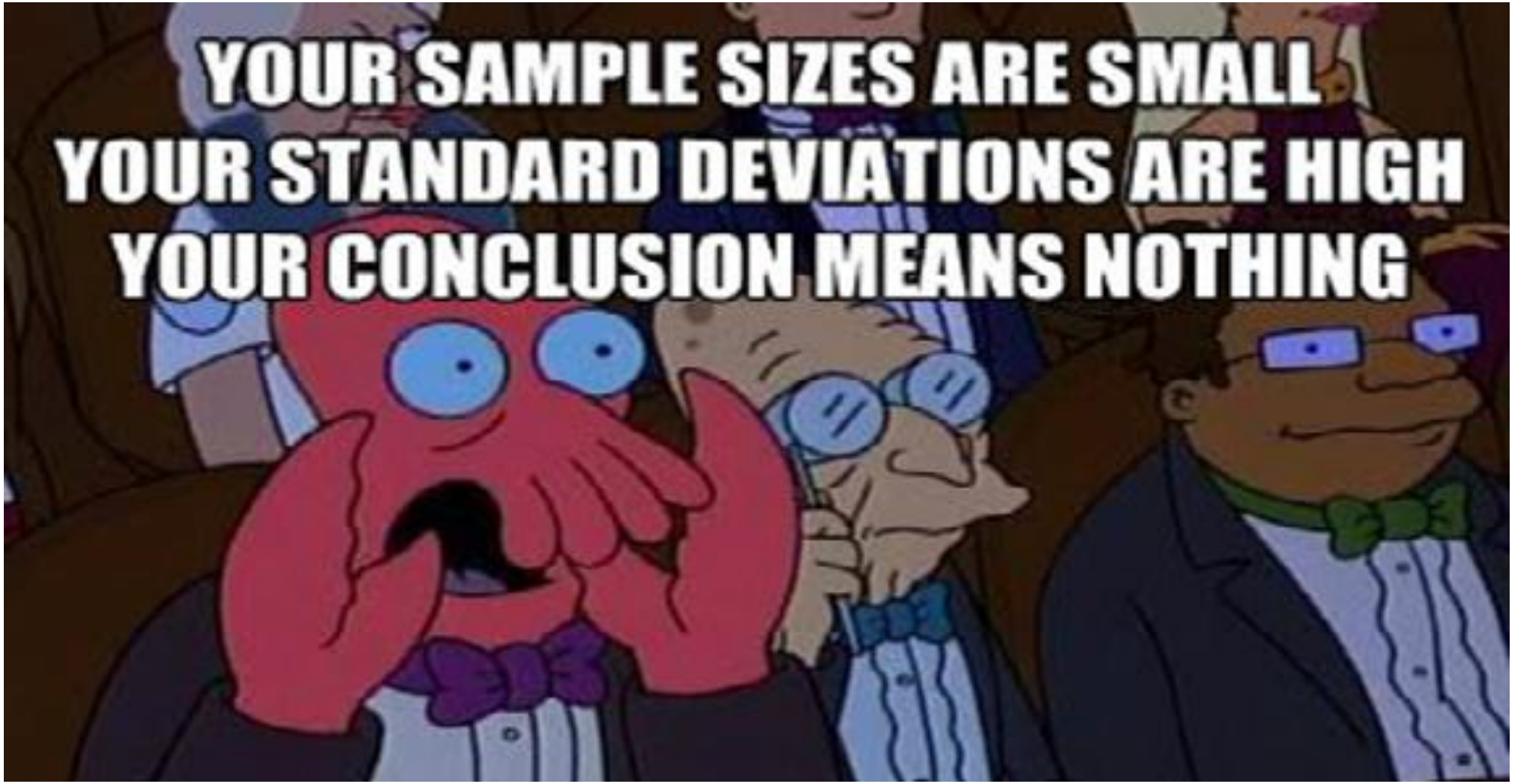
Report effect size in most cases – i.e., confidence intervals

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?

The tidyverse and dplyr

The ‘tidyverse’

The tidyverse is set of R packages that operate ‘tidy data’

- i.e., that operate on data frames (or tibbles)

Tidy data is data where:

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

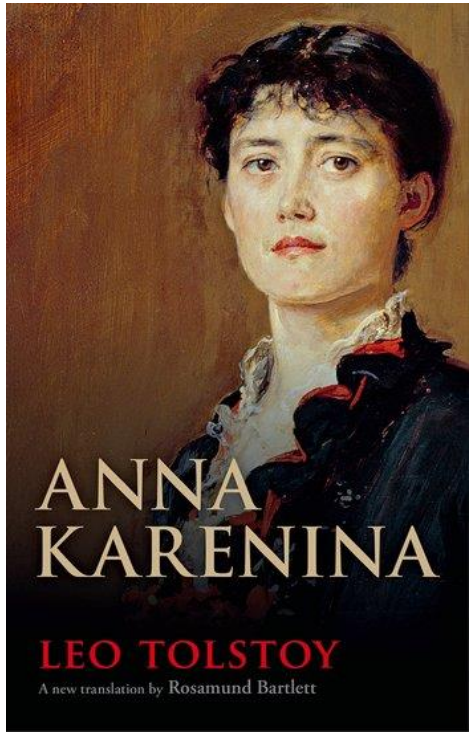
Messy data…

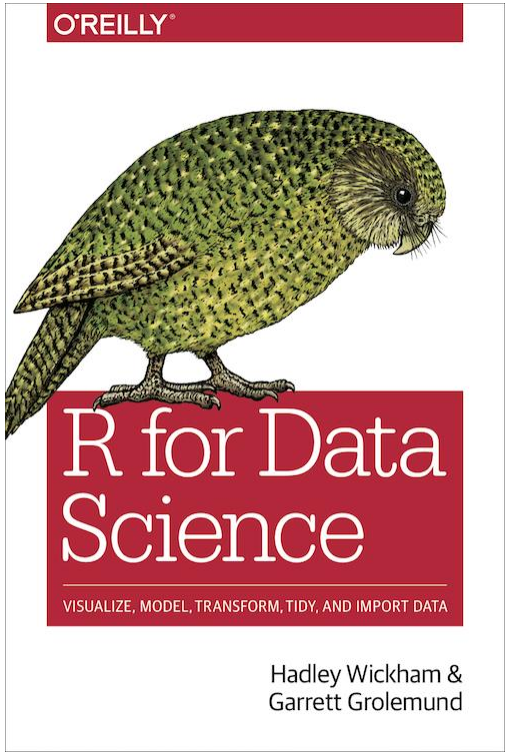
What would be an example of data that is not tidy?

Curve information - Curve quality data														
Name	Formula	Slope at	Intercept	ED-20	ED-50	ED-80	Correlatic	Forced through origo						
Standard	Calc 1: C	standarc	standarc	3792394	27752	0.2	0.5	0.8	1	No				
Plate information														
Plate	Repeat	Barcode	Measurec	Chamber	Chamber	Humidity	Humidity	Ambient	Ambient	Formula	Measurement date			
1	1		N/A	N/A	N/A	N/A	N/A	N/A	N/A	Calc 1: C	standarc	standarc	10.12.2013 10:23:33	
Background information														
Plate	Label	Result	Signal	Flashes/T	Measime	MeasInfo								
1	PicoGree	0	110307	10	0	De=1st Ex=Top Em=Top Wdw=N/A								
Calculate	standarc	standards on each plate) where Label: PicoGreenFilterTop(1) channel 1												
	1	2	3	4	5	6	7	8	9	10	11	12		
A	-0.0011	-0.0011	-0.001	-0.001	-0.0011	-0.0012	-0.0011	-0.0011	-0.0012	-0.0012	0.9973	1.0026		
B	0.0012	0.0014	0.0013	0.0012	0.0013	0.0012	0.0014	0.0003	-0.0011	-0.0011	0.0981	0.103		
C	0.0016	0.0013	0.0013	0.0011	0.0012	0.0015	0.0016	-0.0004	-0.0011	-0.0011	0.0104	0.0095		
D	0.0019	0.0024	0.0018	0.0015	-0.001	-0.001	-0.001	-0.001	-0.0011	-0.0011	0.0008	0.0009		
E	-0.001	-0.0011	-0.0011	-0.0011	-0.001	-0.0012	-0.0011	-0.001	-0.0009	-0.0011	-0.0001	-0.0002		
F	-0.001	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.0011	-0.0009	-0.001	-0.001	-0.0003	-0.0002		
G	-0.0011	-0.0011	-0.0011	-0.001	-0.001	-0.0012	-0.0011	-0.001	-0.001	-0.0011	-0.0002	0.0012		
H	-0.0011	-0.0012	-0.0011	-0.001	-0.0011	-0.0011	-0.0012	-0.0011	-0.0011	-0.001	-0.0003	-0.0003		

Messy data…

“Happy families are all alike; every unhappy family is unhappy in its own way.” –– Leo Tolstoy





“Tidy datasets are all alike, but every messy dataset is messy in its own way.” –– Hadley Wickham

Messy data…

Messy data can be difficult to deal with

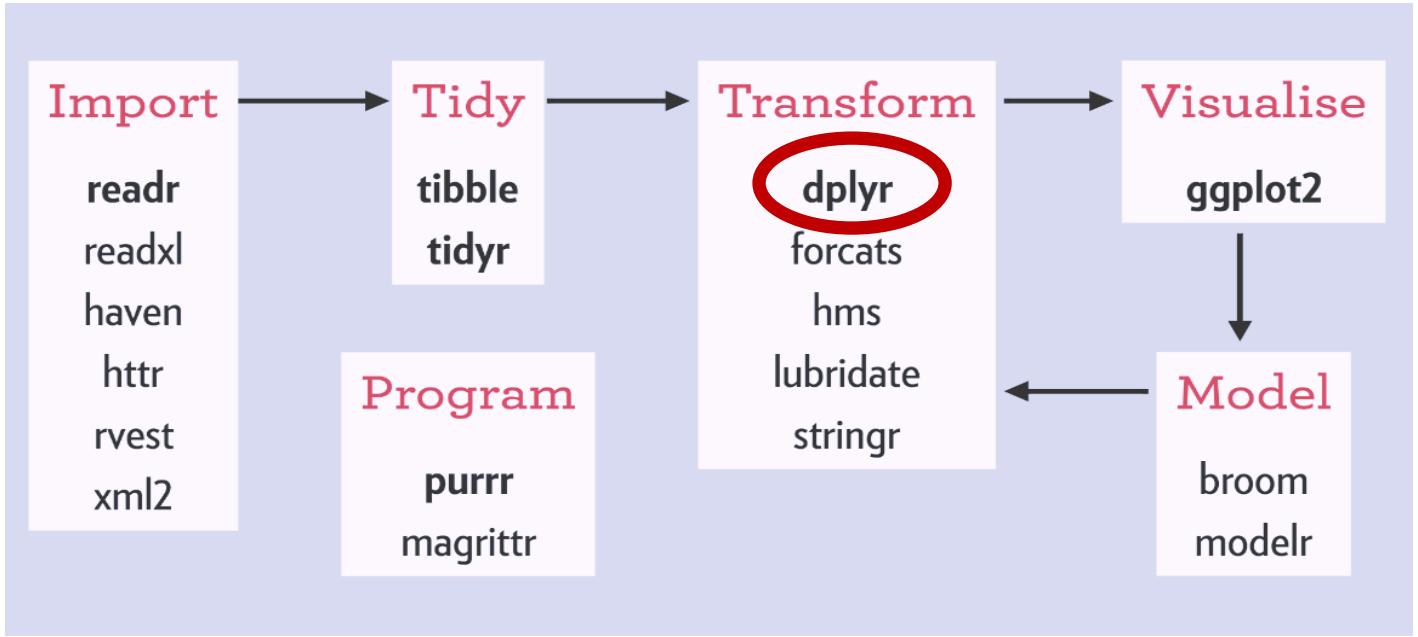
Curve information - Curve c		
Name	Formula	Slope at
Standard	Calc 1: C	standar
Plate information		
Plate	Repeat	Barcode
1	1	
Background information		
Plate	Label	Result
1	PicoGree	0
Calculate	standarc	standar
	1	2
A	-0.0011	-0.0011
B	0.0012	0.0014
C	0.0016	0.0013
D	0.0019	0.0024
E	-0.001	-0.0011
F	-0.001	-0.0011
G	-0.0011	-0.0011
H	-0.0011	-0.0012

te			
rc	10.12.2013 10:23:33		
.2			
6			
3			
5			
9			
2			
2			
.2			
3			

The ‘tidyverse’

The packages share a common design philosophy

- Most written by Hadley Wickham



dplyr: A grammar for data wrangling

Grammar: a set of components that can be combined to achieve a goal

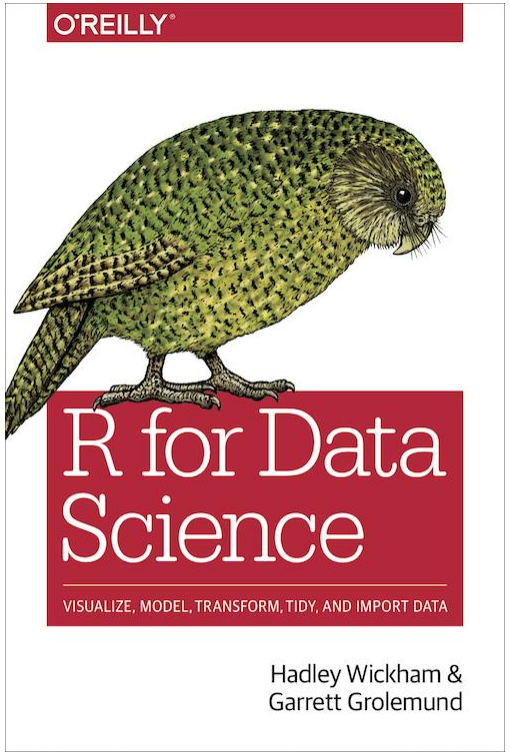
dplyr is a package that has a set of verbs that are useful for transformations data:

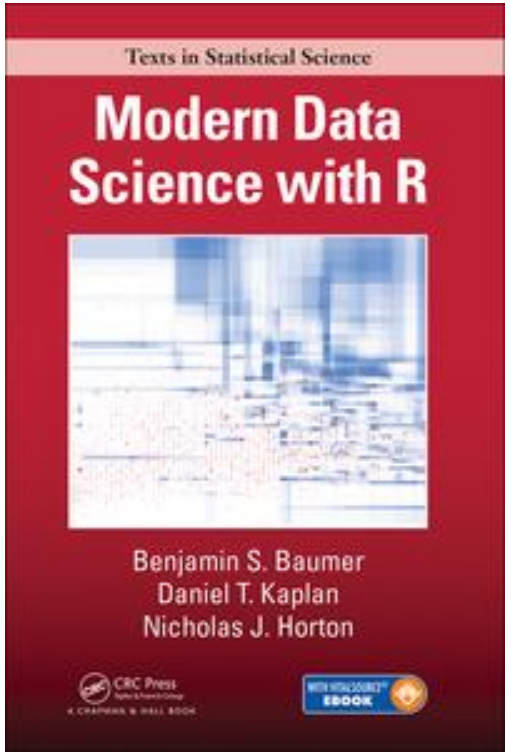
1. filter()
2. select()
3. mutate()
4. arrange()
5. group_by()
6. summarize()

All these function **take a data frame** and other arguments and **return a data frame**

```
> library(dplyr)   # load the dplyr package
```

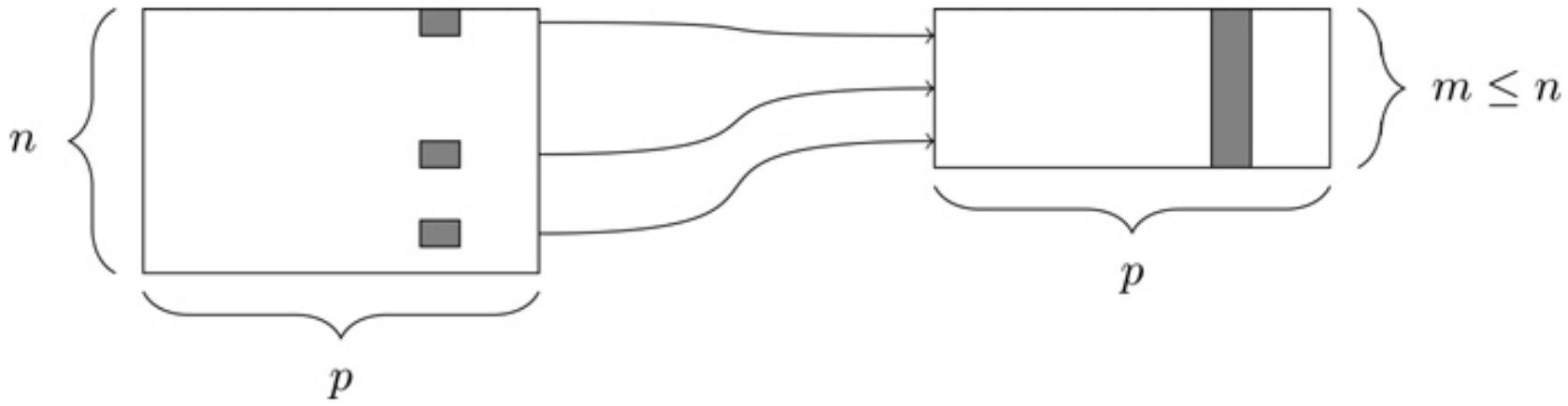
Quick overview of the dplyr functions





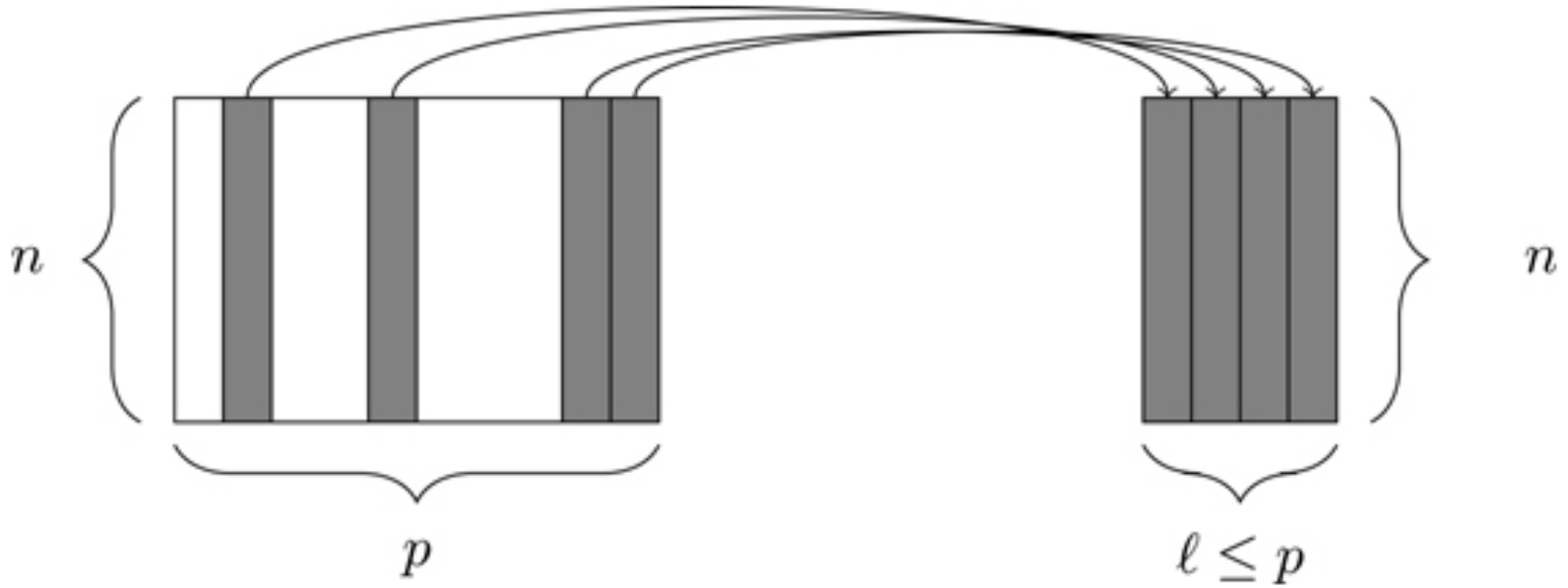
1. filter()

The filter() function allows you to select a subset of rows in data frame



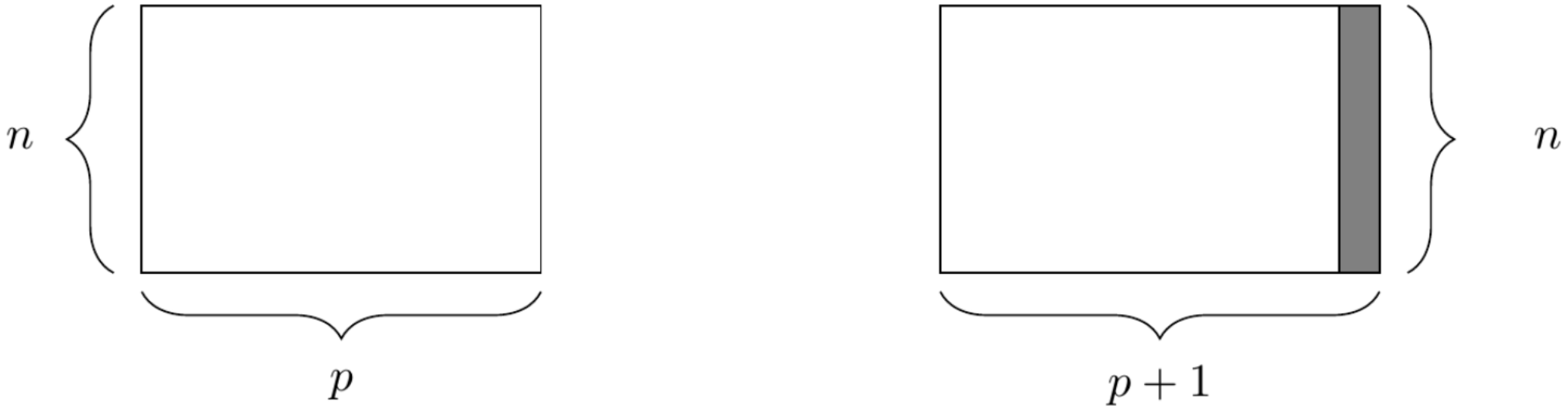
2. select()

The select() function allows you to select a subset of columns



3. mutate()

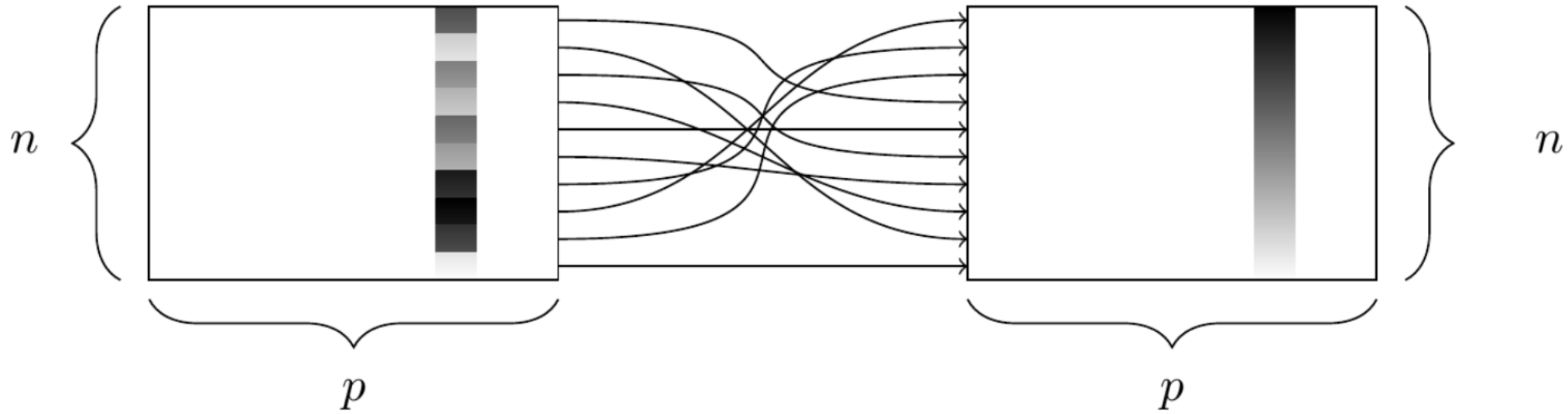
The mutate() function allows you to create new columns that are functions of existing columns



4. arrange()

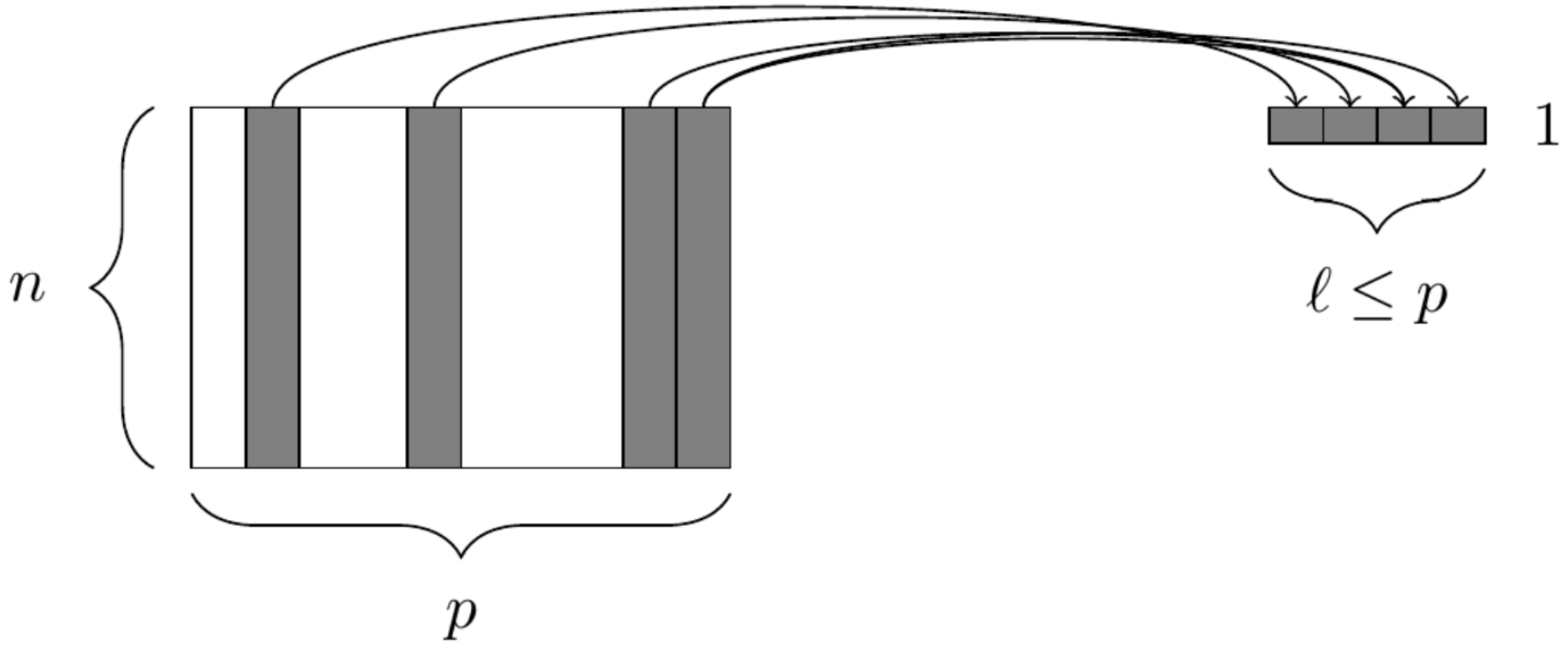
The arrange() function arranges the rows based values in a column

- arrange(desc()) arranges from largest to smallest



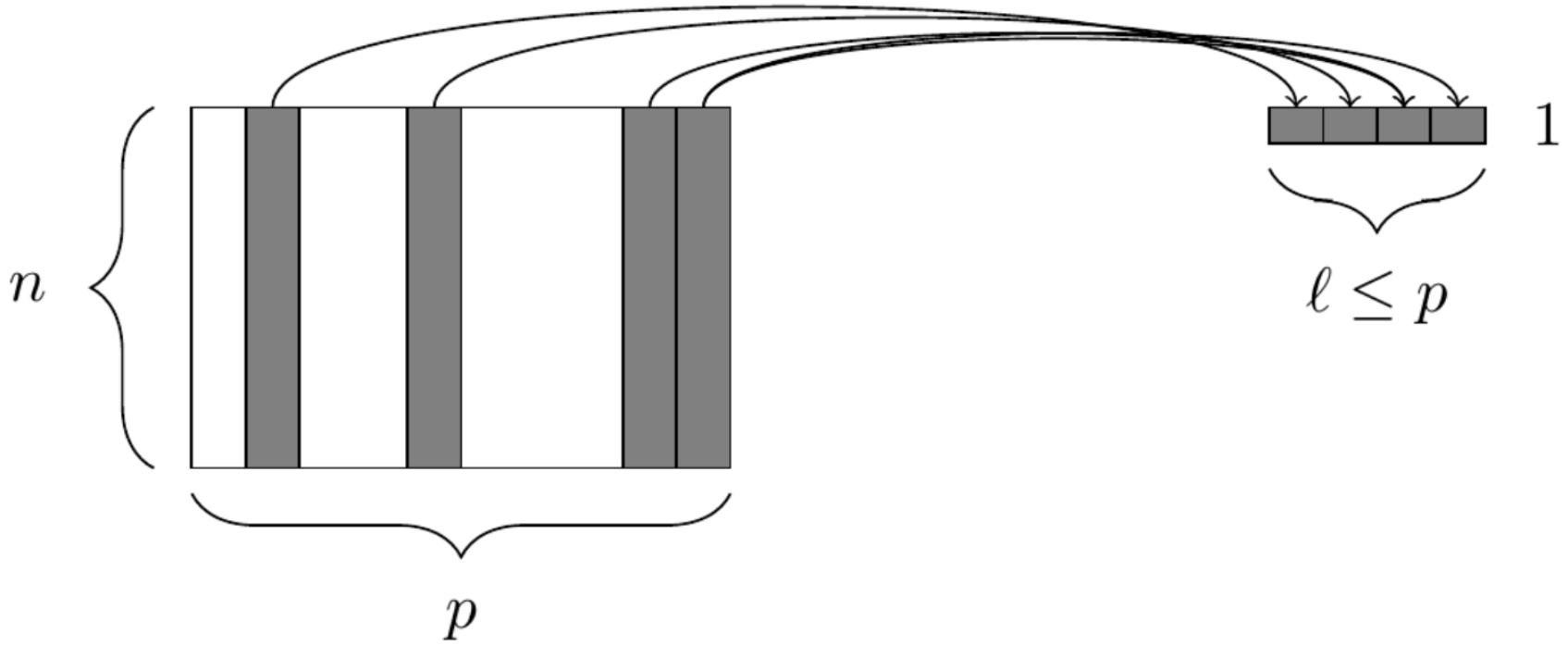
5. summarize()

The summarize() function reduces values in many rows into single values



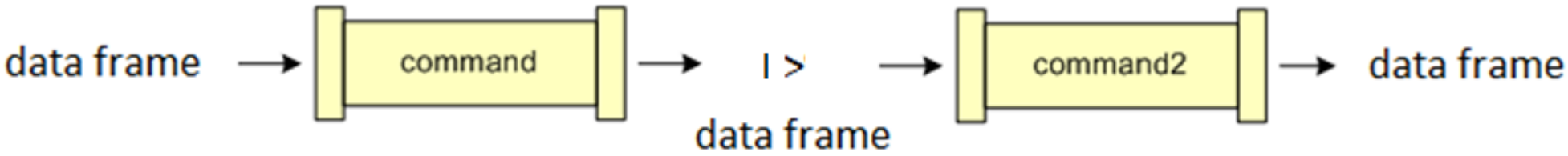
6. The group_by() function

The group_by() function groups variables for future operations



The pipe operator

The pipe operator |> allows us to chain commands together



Let’s try it out!

Homework 5: flight delays

Data set contains information about flights leaving NYC in 2013

```
> library("nycflights13")
> data(flights)
```

I recommend you get started soon

Next class: a grammar of graphics and ggplot