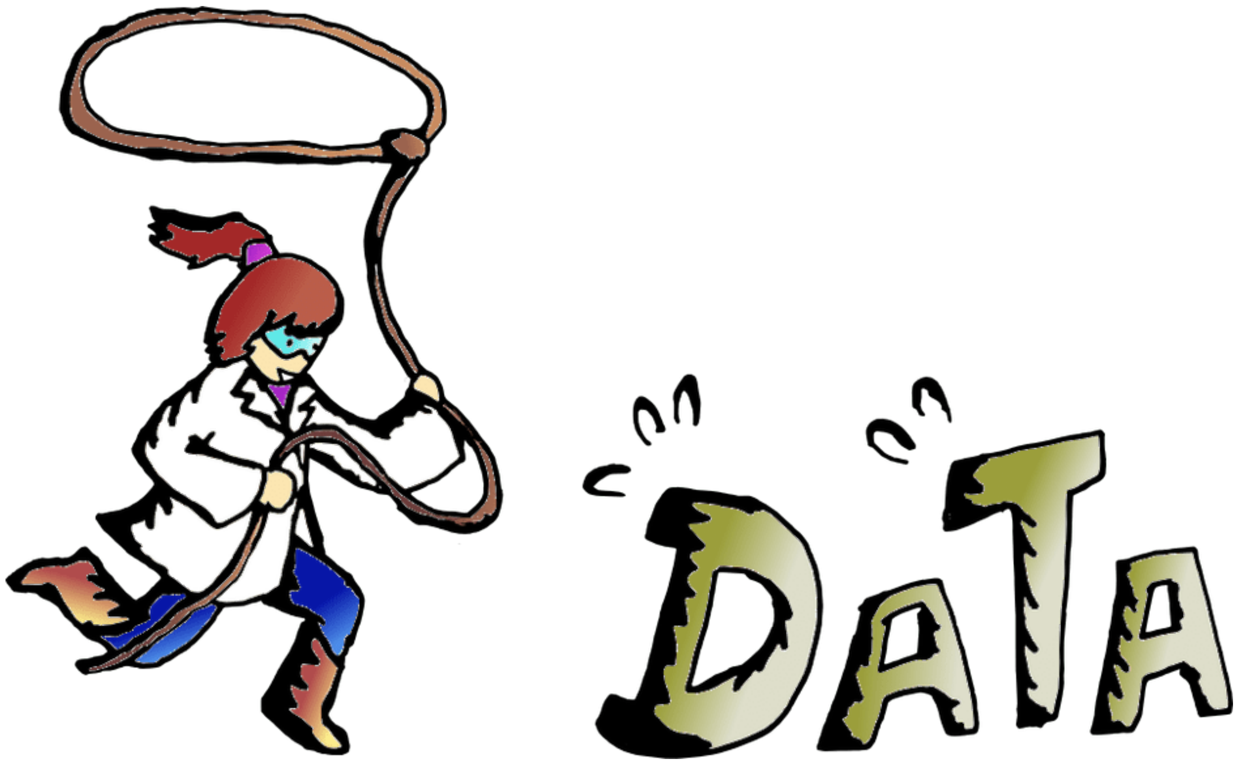
Theories, functions and data transformations



Overview

Theories of hypothesis tests

Functions and if statements in R

Data wrangling with dplyr

Week 5 survey

Don’t forget to fill out the week 5 survey!

Theories of hypothesis tests

Exploratory vs. confirmatory analyses

Confirmatory data analysis

- Confirming or falsifying **existing** hypotheses
- Need to state hypotheses before you collect the data
  - Pre-registered studies

Exploratory data analysis

- Visualize, describe and model data to find **potential** trends and to generate **new** hypotheses
- P-values are just descriptive statistics
  - Need to do confirmatory analyses to do real inference

Theories of hypothesis tests

Is it better to report the actual p-value or just whether we rejected the null hypothesis H_0?

Report the exact p-value	77 respondents	**91 %**
Just state whether to reject the null hypothesis	8 respondents	9 %

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)

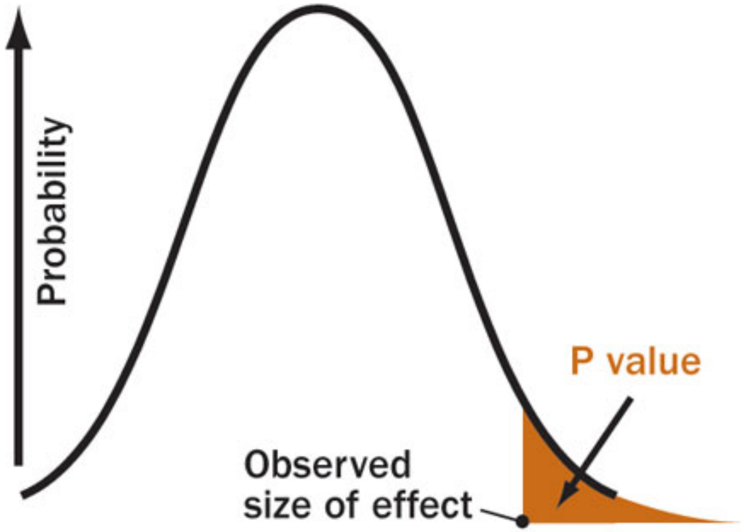
Neyman (1894-1981)

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”



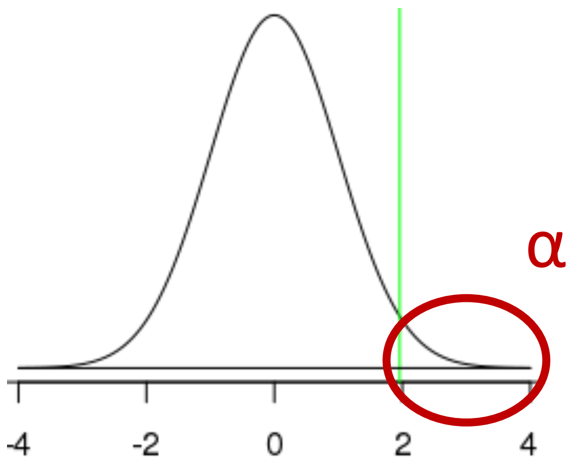
Neyman-Pearson null hypothesis testing

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

Reject H_0: if the observed sample statistic is 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.

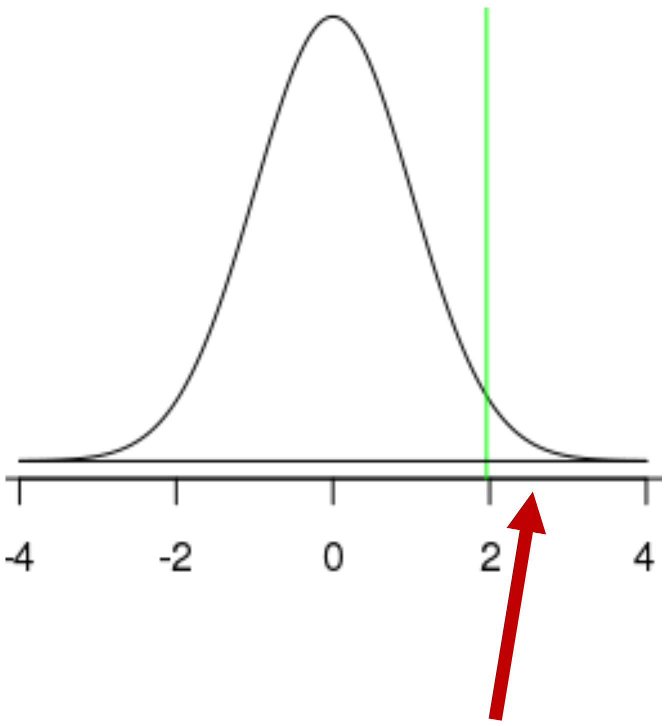
Neyman-Pearson frequentist logic

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

If Neyman-Pearson null hypothesis testing paradigm was followed perfectly, 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

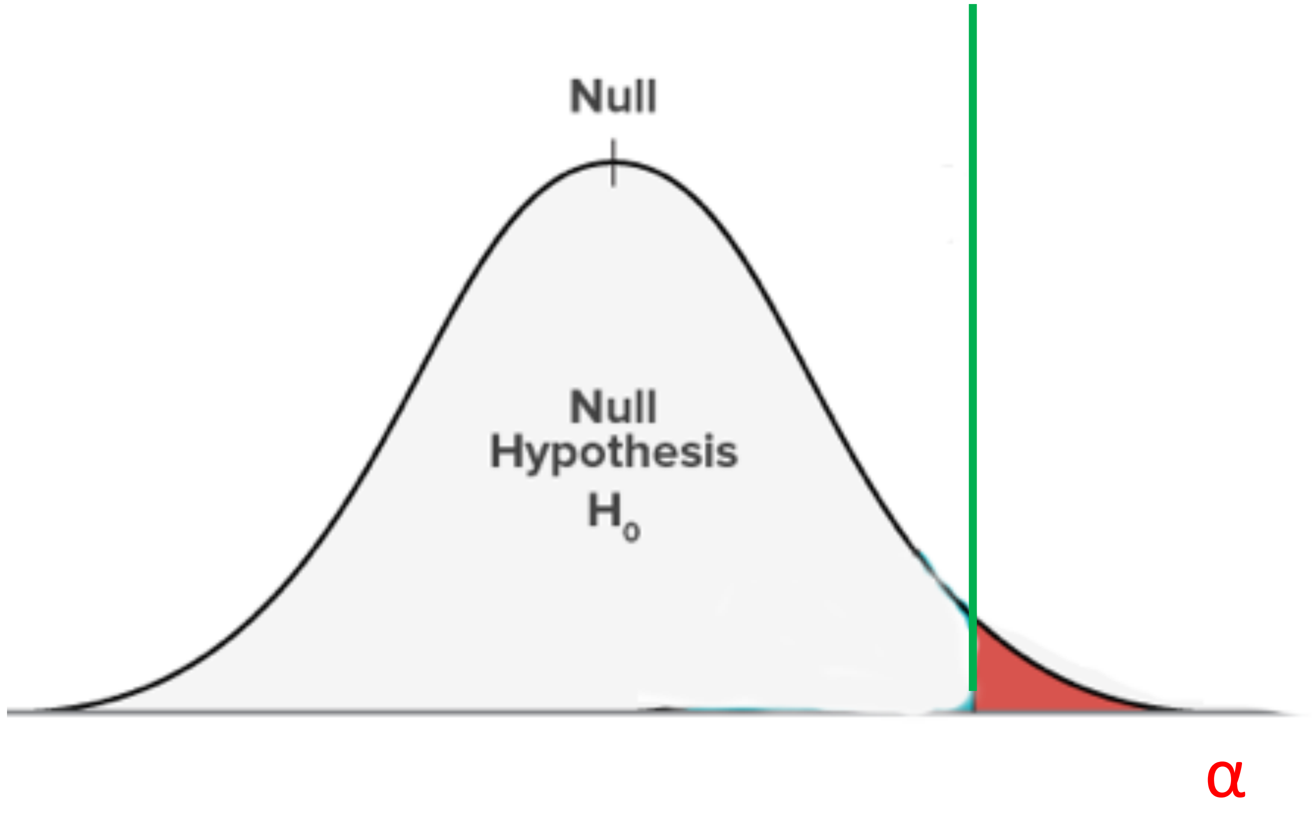
Null distribution



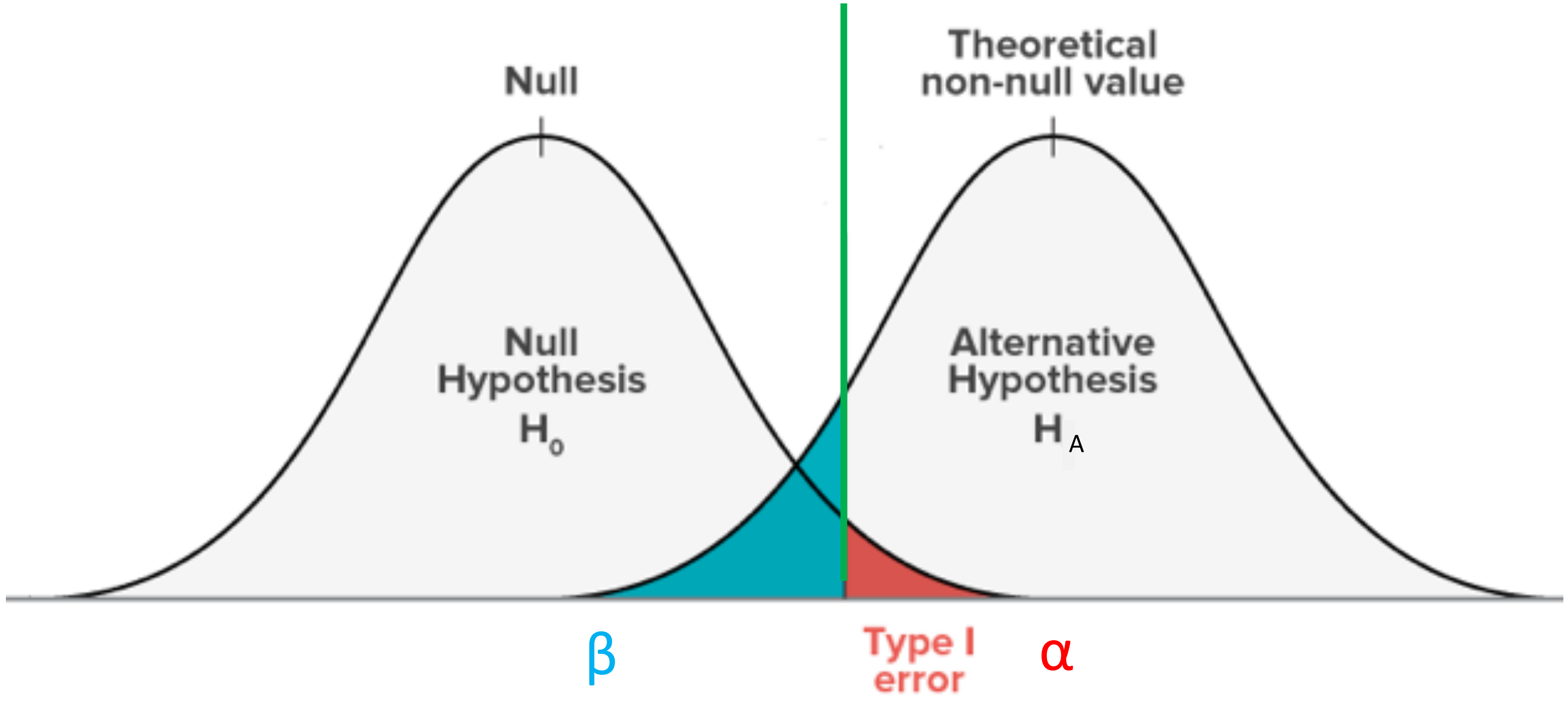
The null distribution is true but statistic landed here


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

Neyman-Pearson Frequentist logic



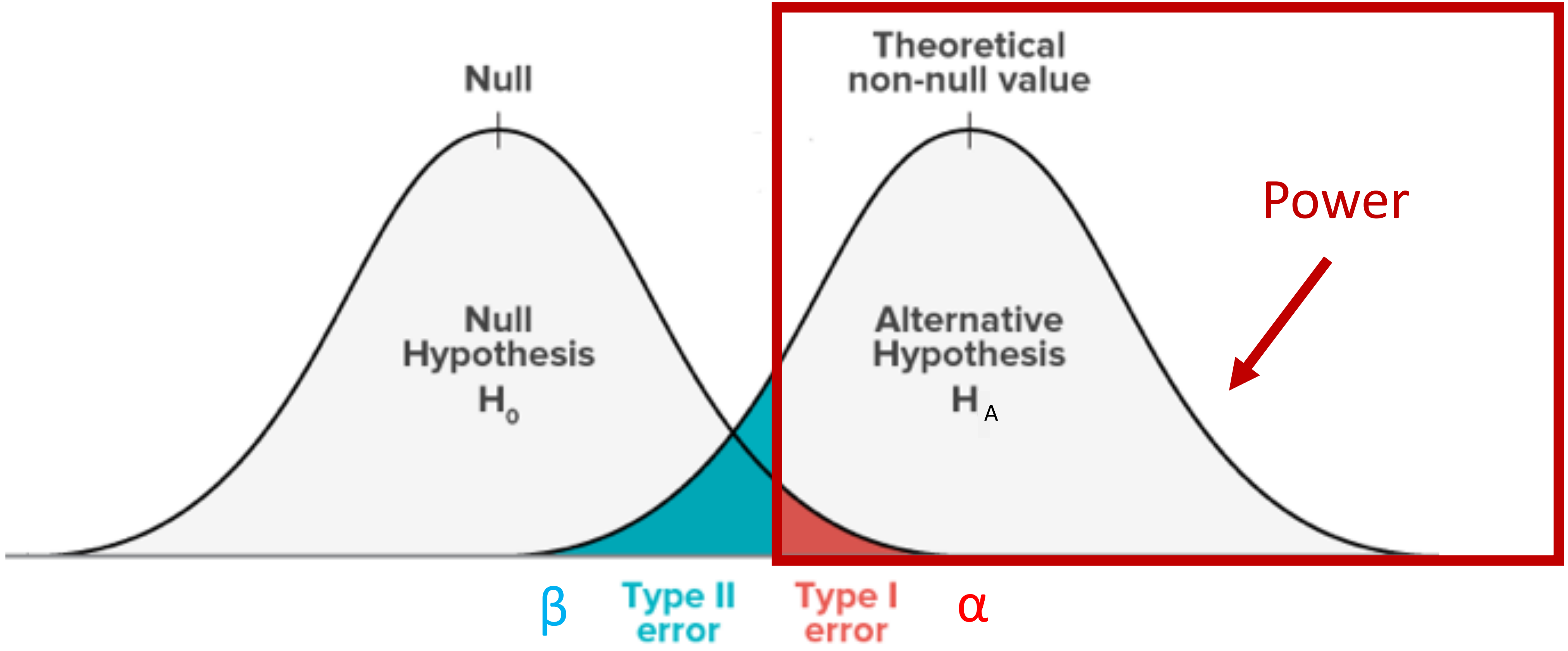
Neyman-Pearson Frequentist logic



Type 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

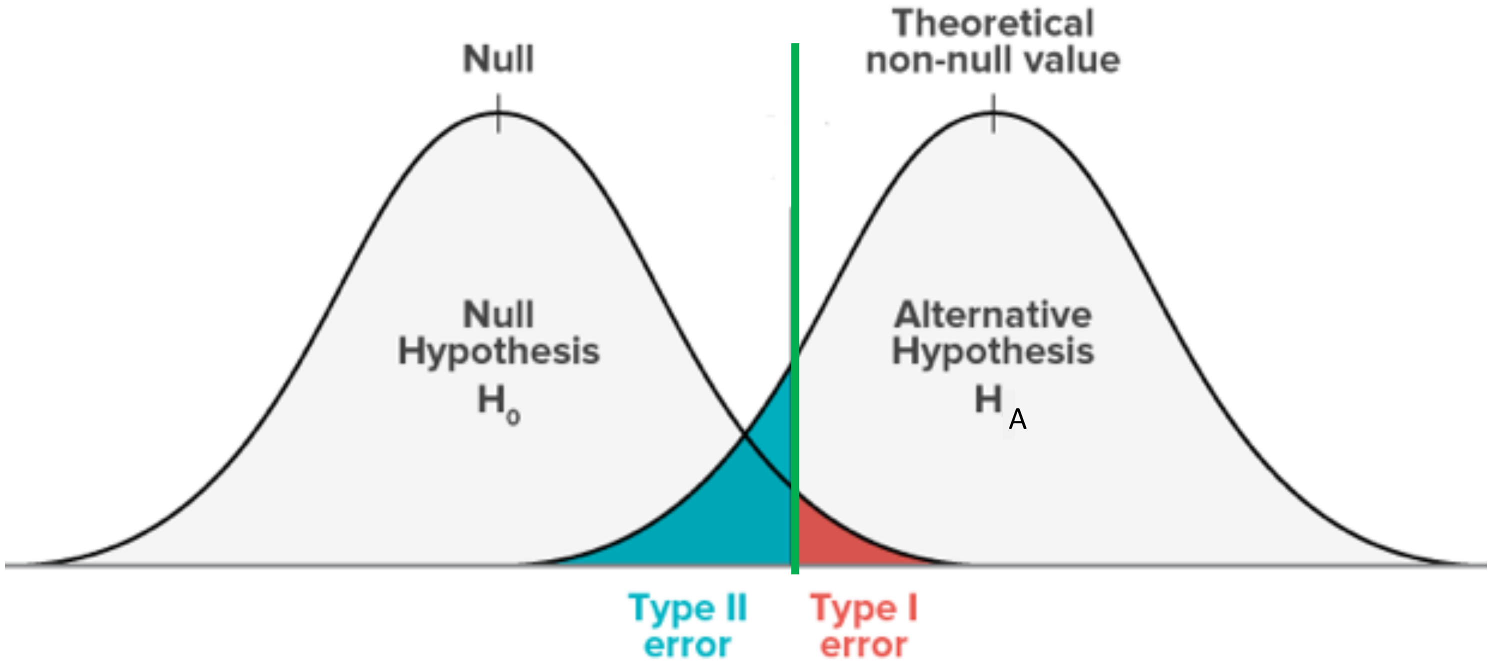
Neyman-Pearson Frequentist logic



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

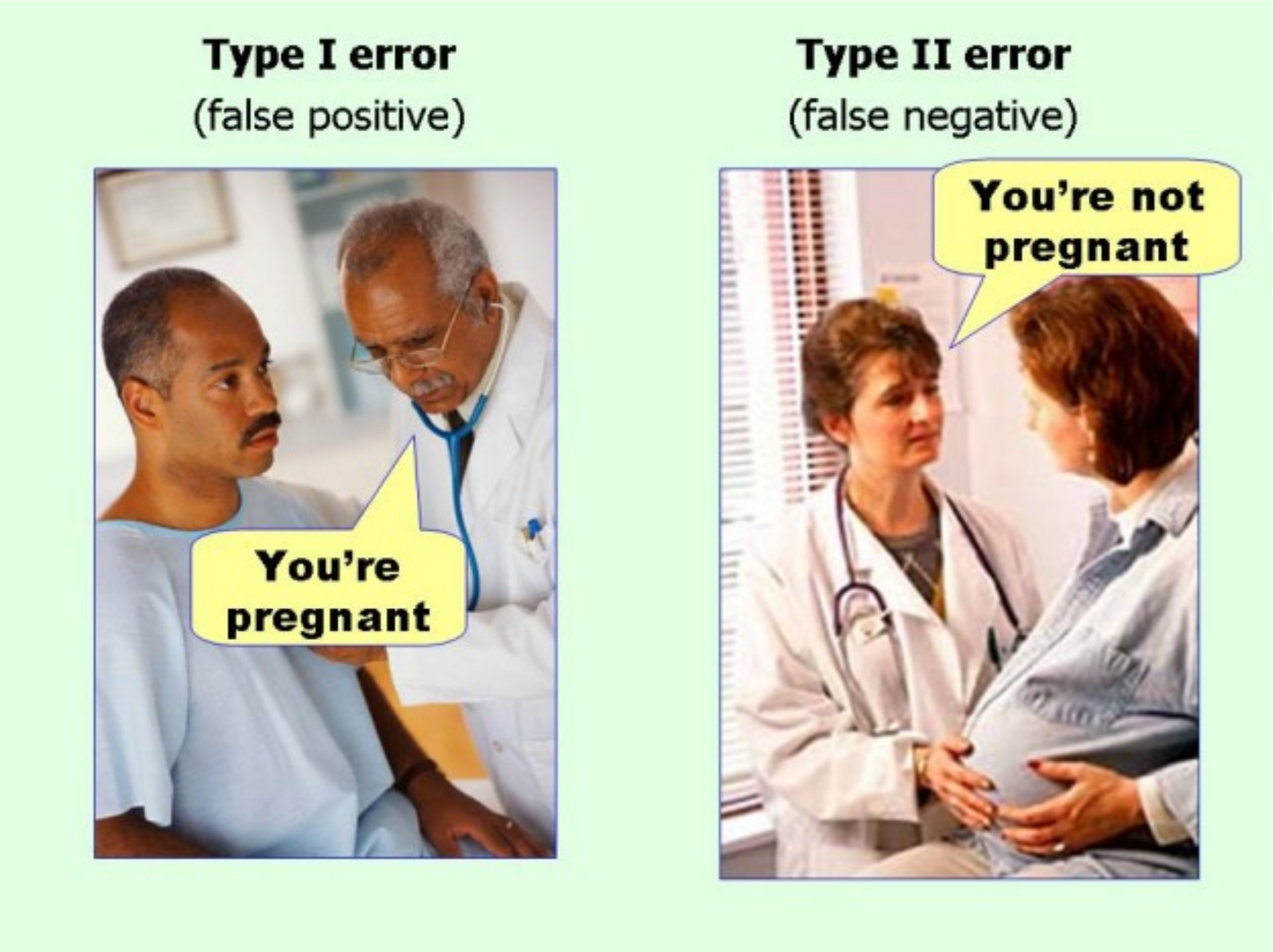
- 1 - β
- For a fixed α level, it would be best to use the most powerful test

Type I and Type II Errors



	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



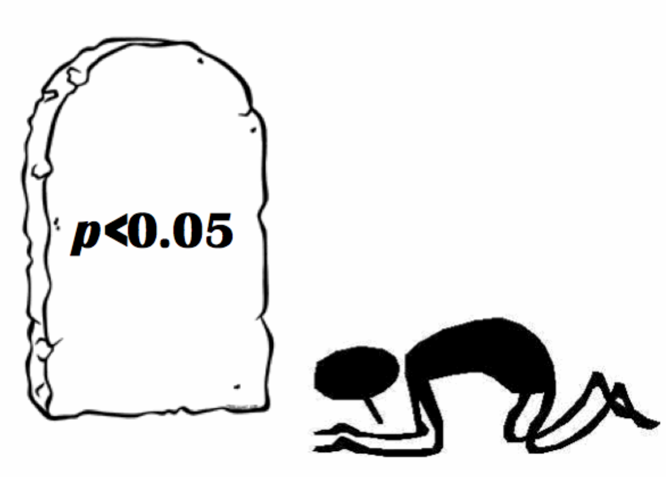
Problems with the NP hypothesis tests

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 true:
  - Joy can smell Parkinson's disease, calcium is good for your heart, organic and inorganic avocados have the same price, …

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

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

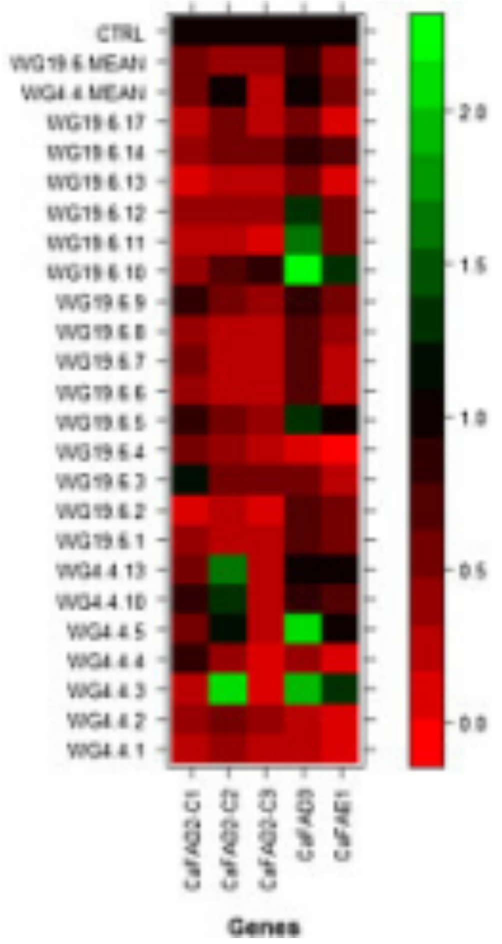
Scientists collected 7129 gene expression levels from 38 patients to find genetic differences between two types leukemia (L1 and L2)

Suppose there was no genetic differences between the types of leukemia

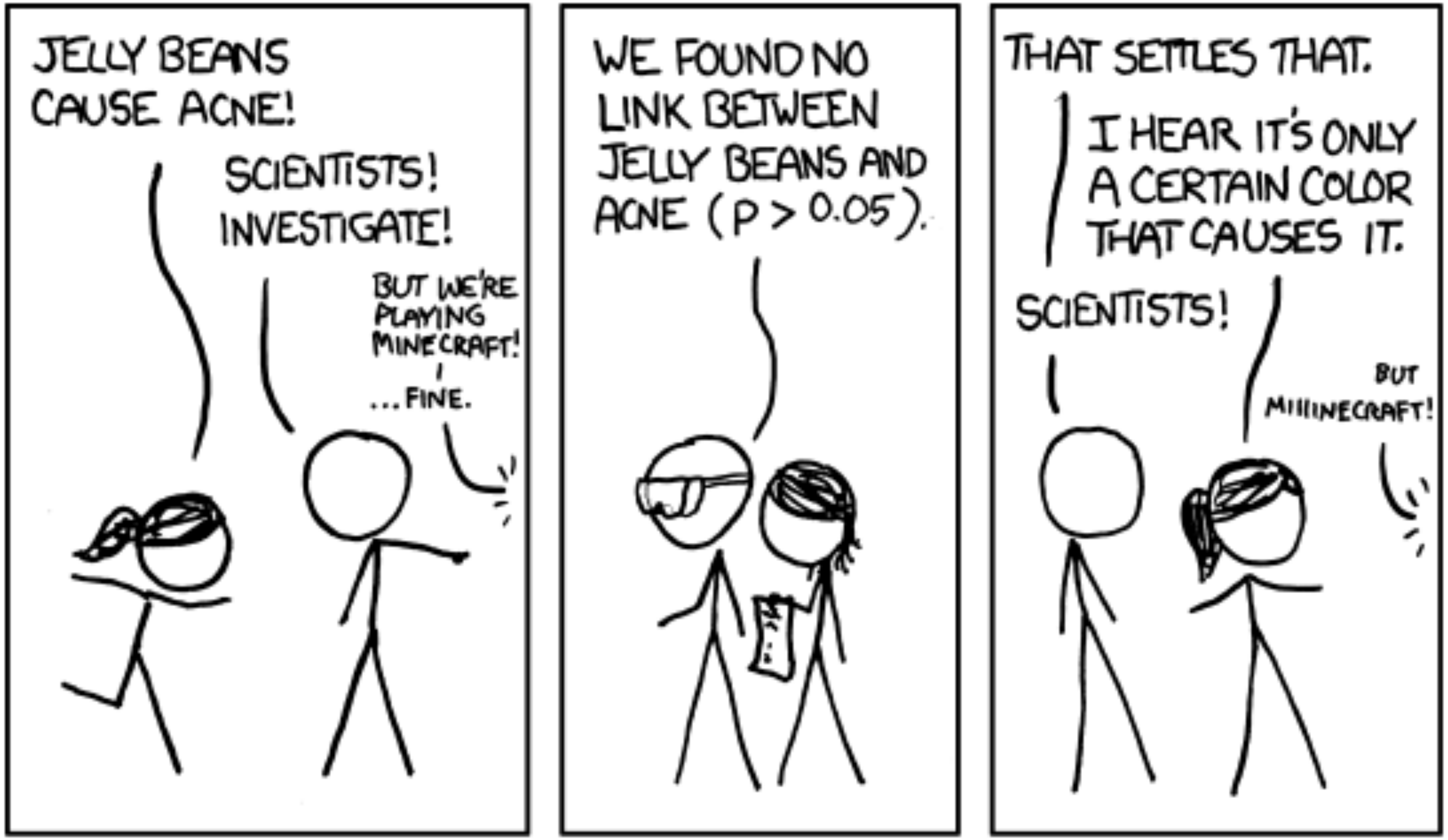
- H_0: $\mu_{L1} = \mu_{L2}$ is true for all genes

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

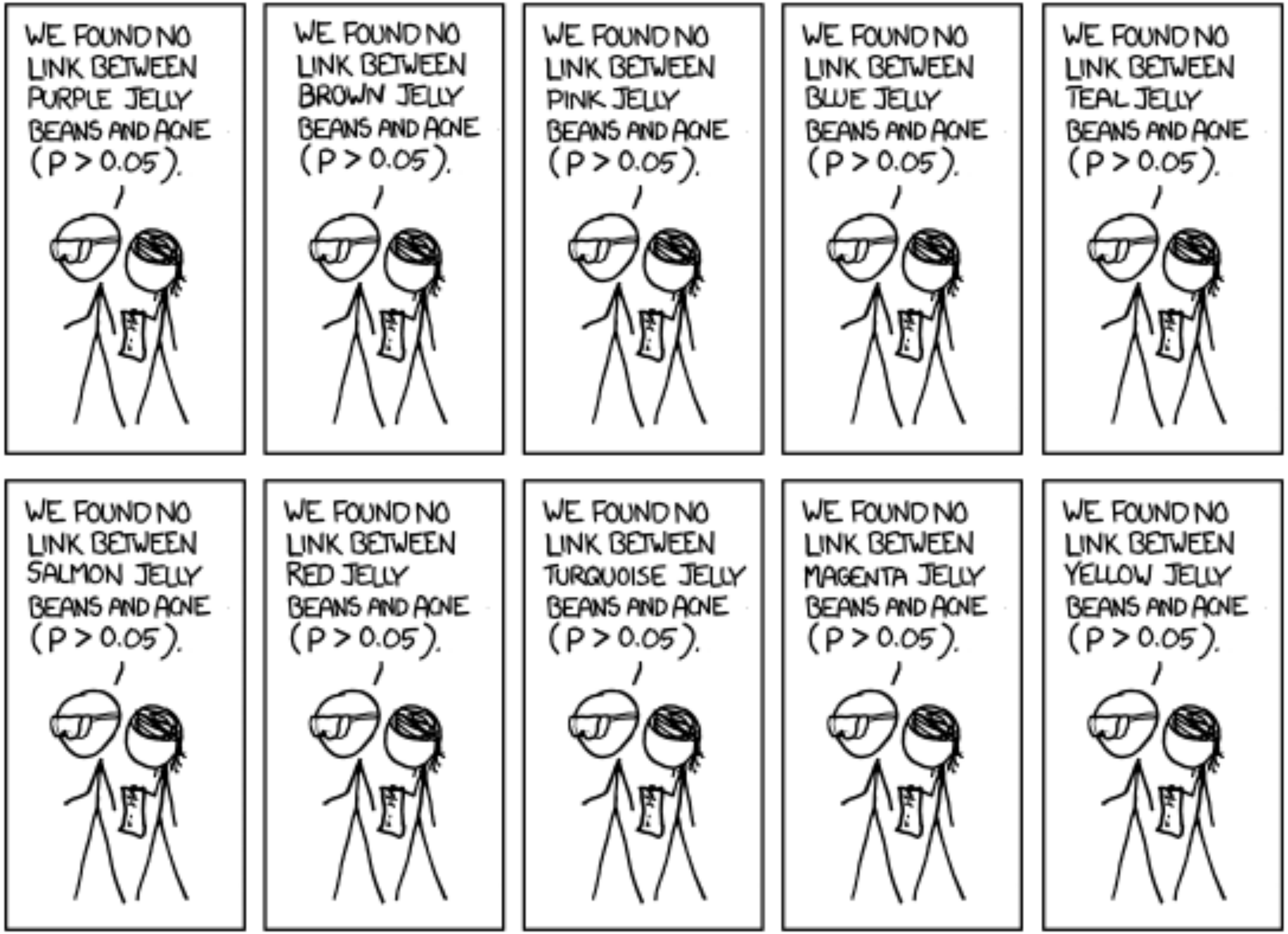
- A: 7129 x 0.05 = 356

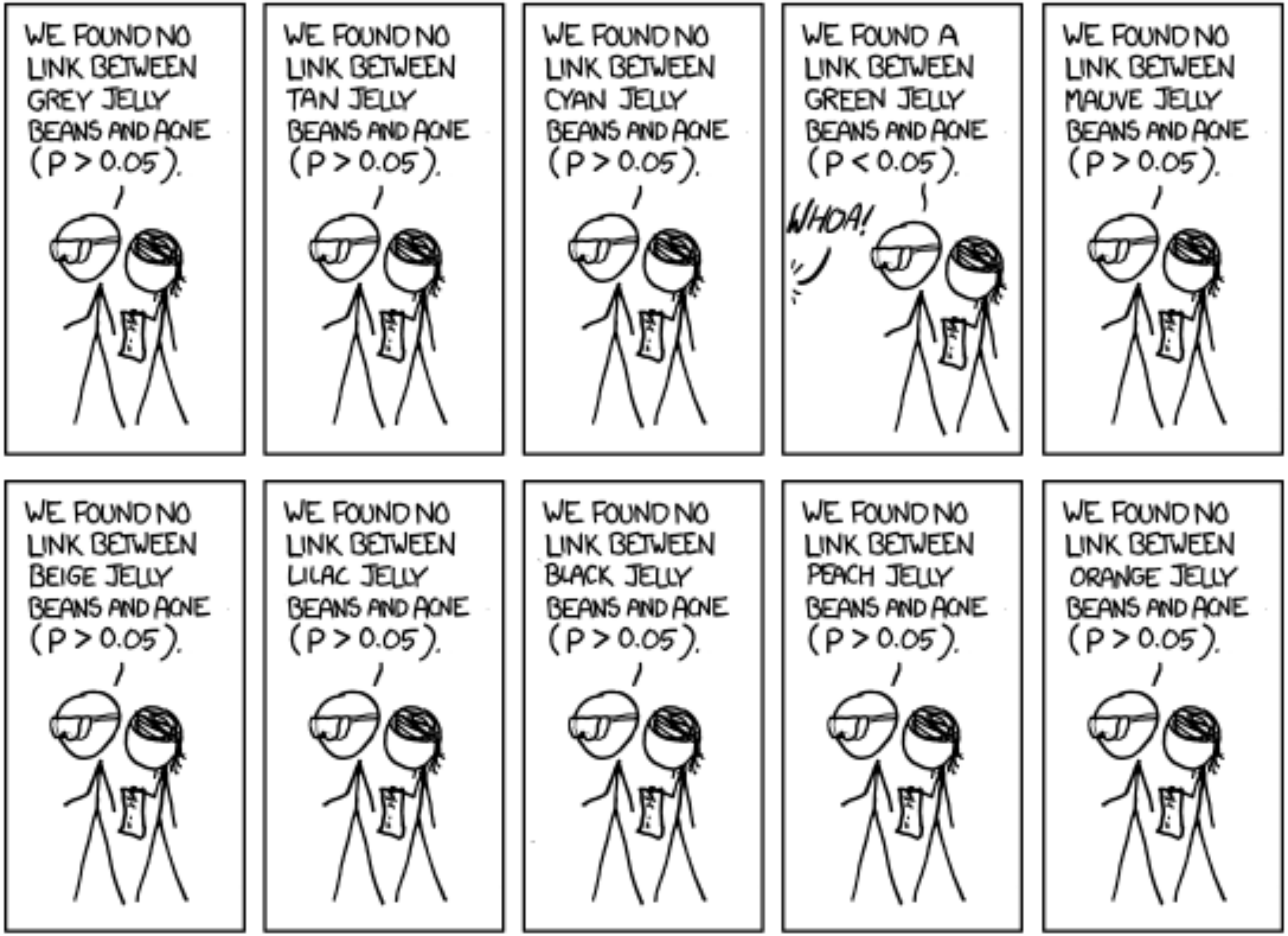


Multiple hypothesis tests



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!

(although can do many tests when doing exploratory data analysis)

http://xkcd.com/882/

Genes and leukemia example

There are methods that try to correct for running multiple hypothesis tests

The ***Bonferroni correction*** is one way that controls the probability of ***any*** hypothesis test giving a type 1 error

- i.e., controls the familywise error rate (no type 1 errors for any of the tests run)

It works by dividing the initial α level by the number of tests run

- E.g., $\alpha = 0.05/7129 = 0.000007$
- All p-values need to be below this level to be considered statistically significant
- This can lead to many type 2 errors (Type 2 error: failure to reject H_0 when it is false)

The problem of multiple testing

For $\alpha = 0.05$, ~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 on confirmatory data analyses

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

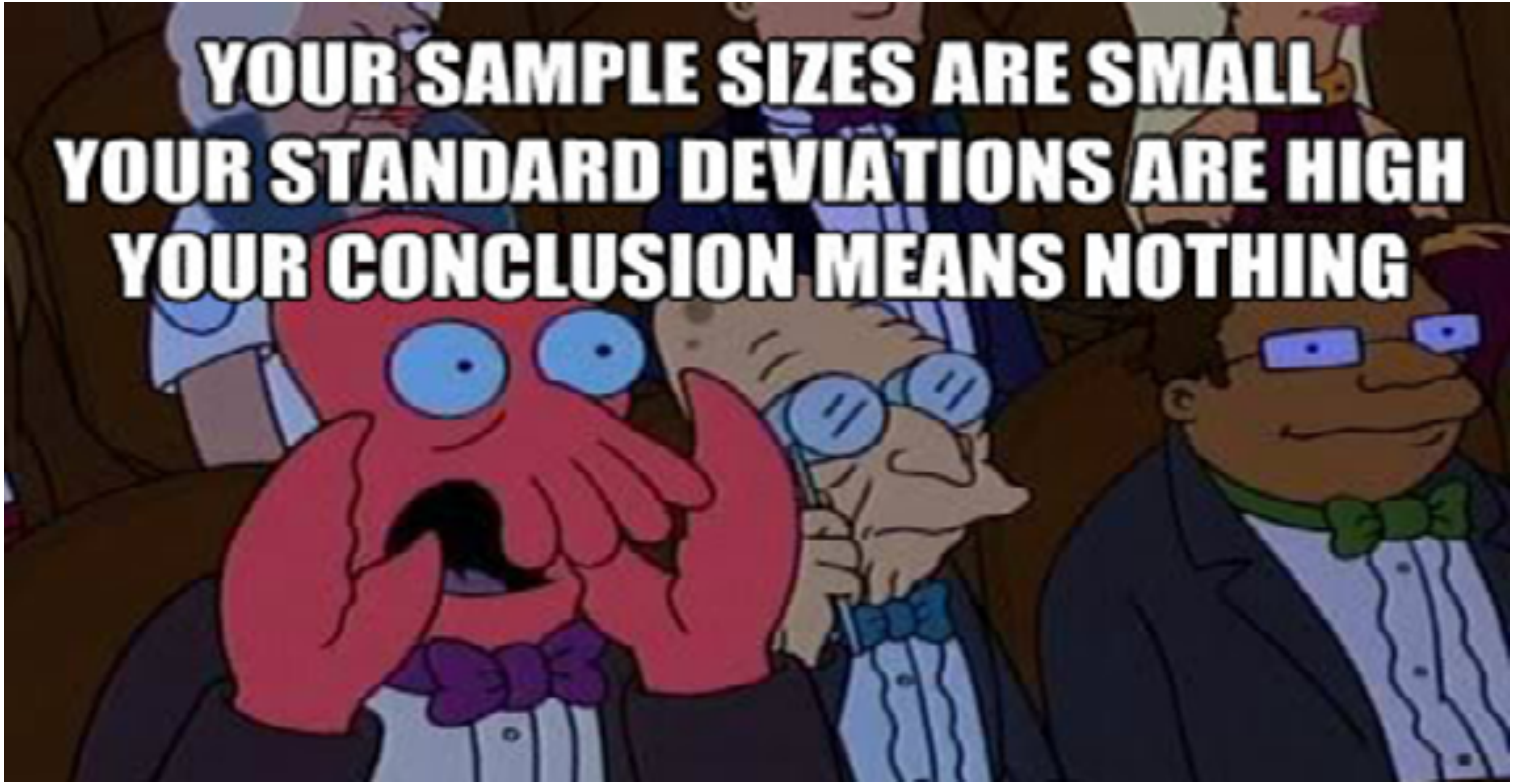
Report effect size in most cases – i.e., confidence intervals

Report the p-values rather than accept/reject H_0

- i.e., report $p = 0.023$ 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

Exploratory vs. confirmatory analyses

Confirmatory data analysis

- Confirming or falsifying **existing** hypotheses
- Need to state hypotheses before you collect the data
    - Pre-registered studies

Exploratory data analysis

- Visualize, describe and model data to find **potential** trends and to generate **new** hypotheses
- P-values are just descriptive statistics
    - Need to do confirmatory analyses to do real inference

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

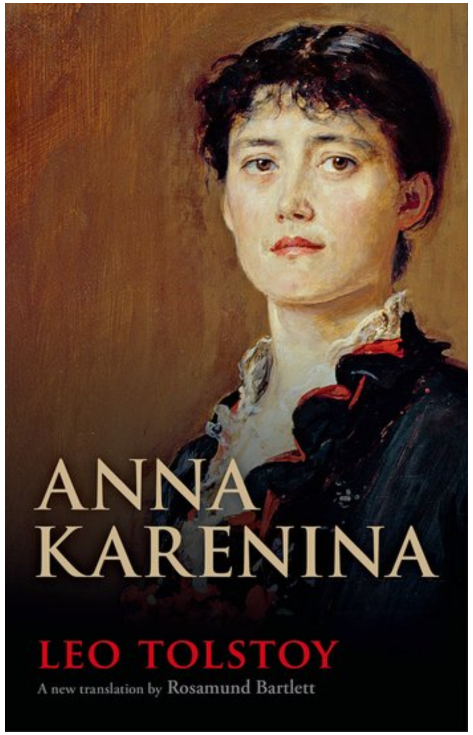
Messy data…

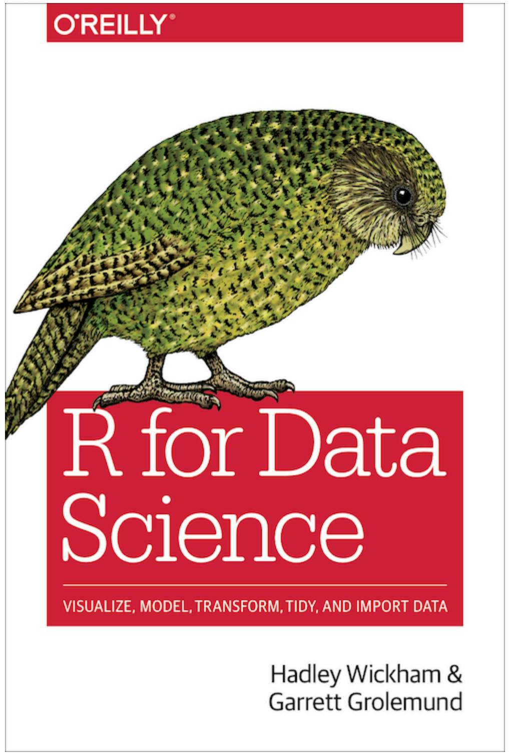
What would be an example of data that is not tidy?

Curve information - Curve quality data													
Name	Formula	Slope at	Intercept	ED-20	ED-50	ED-80	Correlati	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	Meastime	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

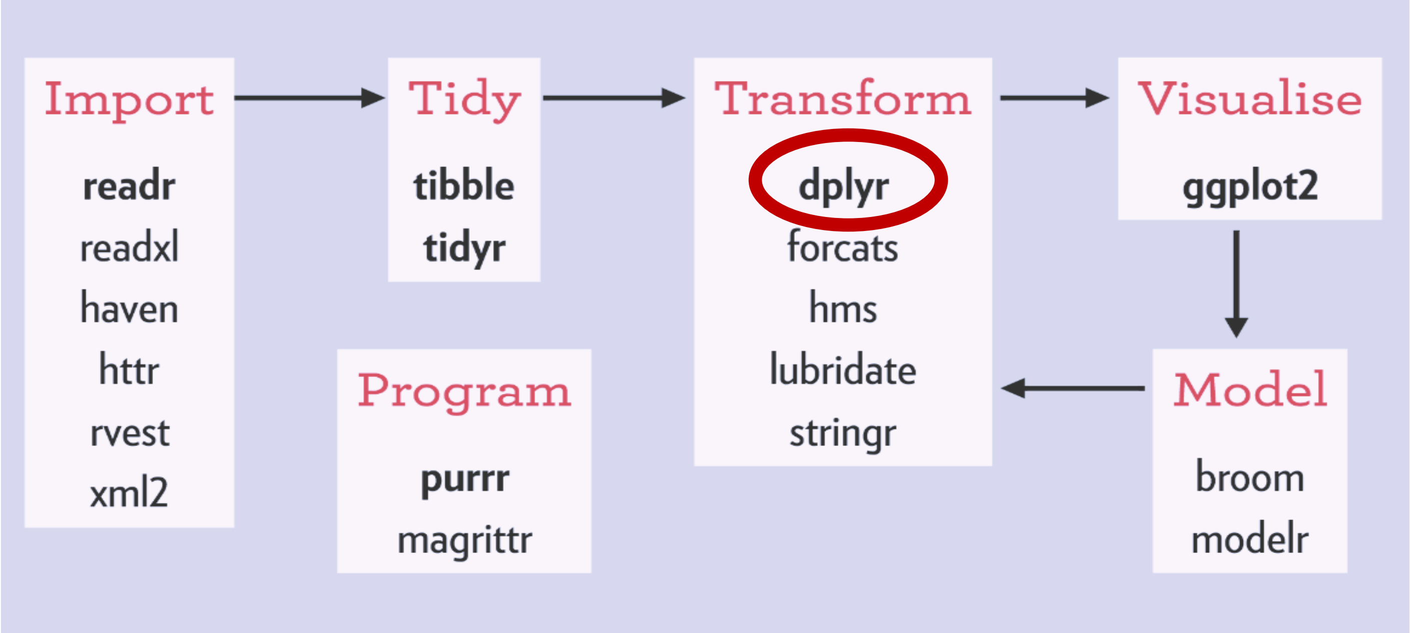
Messy data…

Messy data can be difficult to deal with

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

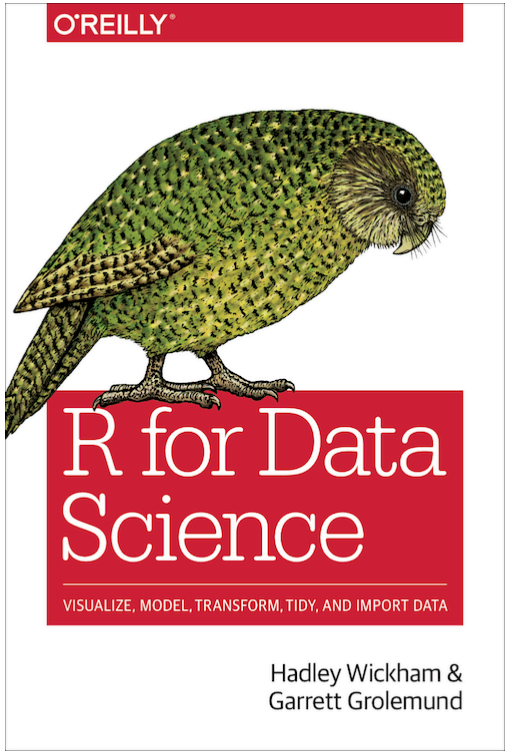
dplyr is a package that has a set of verbs that are useful for transformations data:

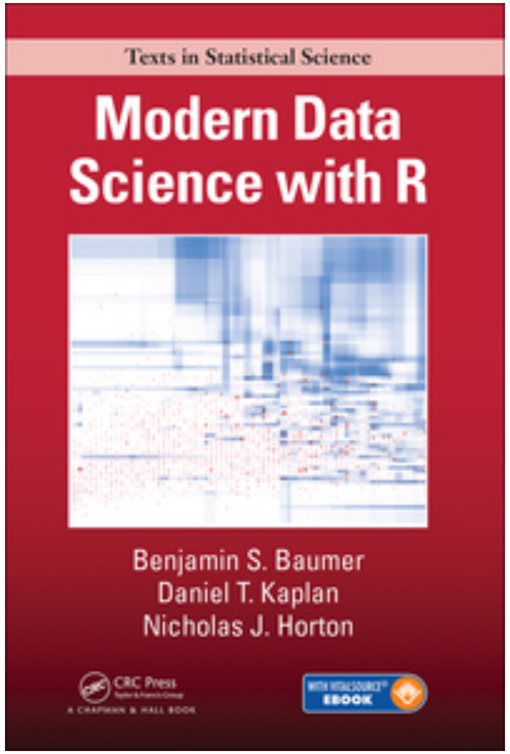
1. filter()
2. select()
3. mutate()
4. arrange()
5. summarize()
6. group_by()

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

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

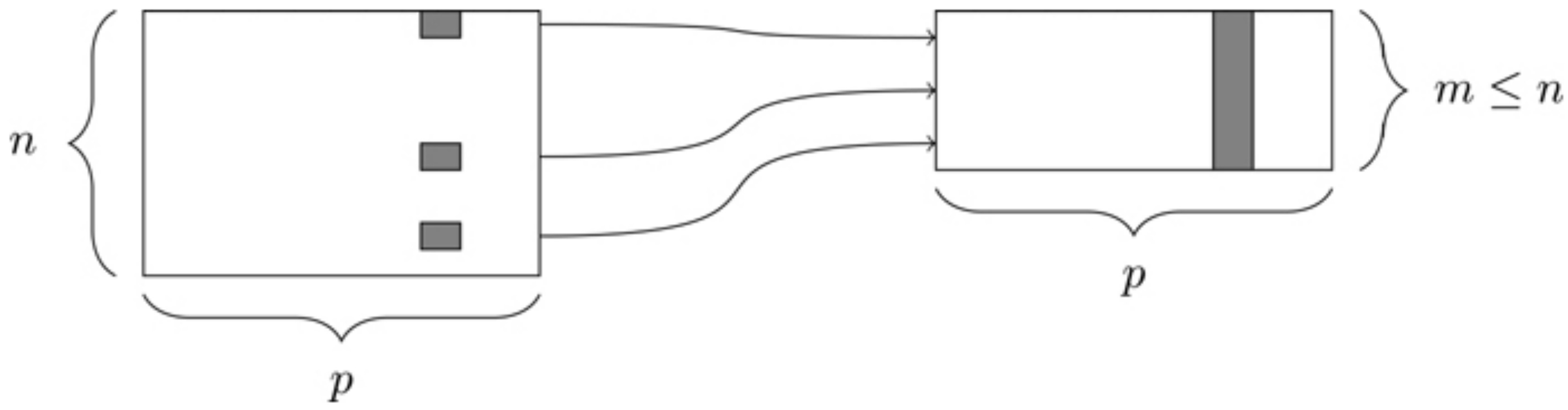
Quick overview of the dplyr functions





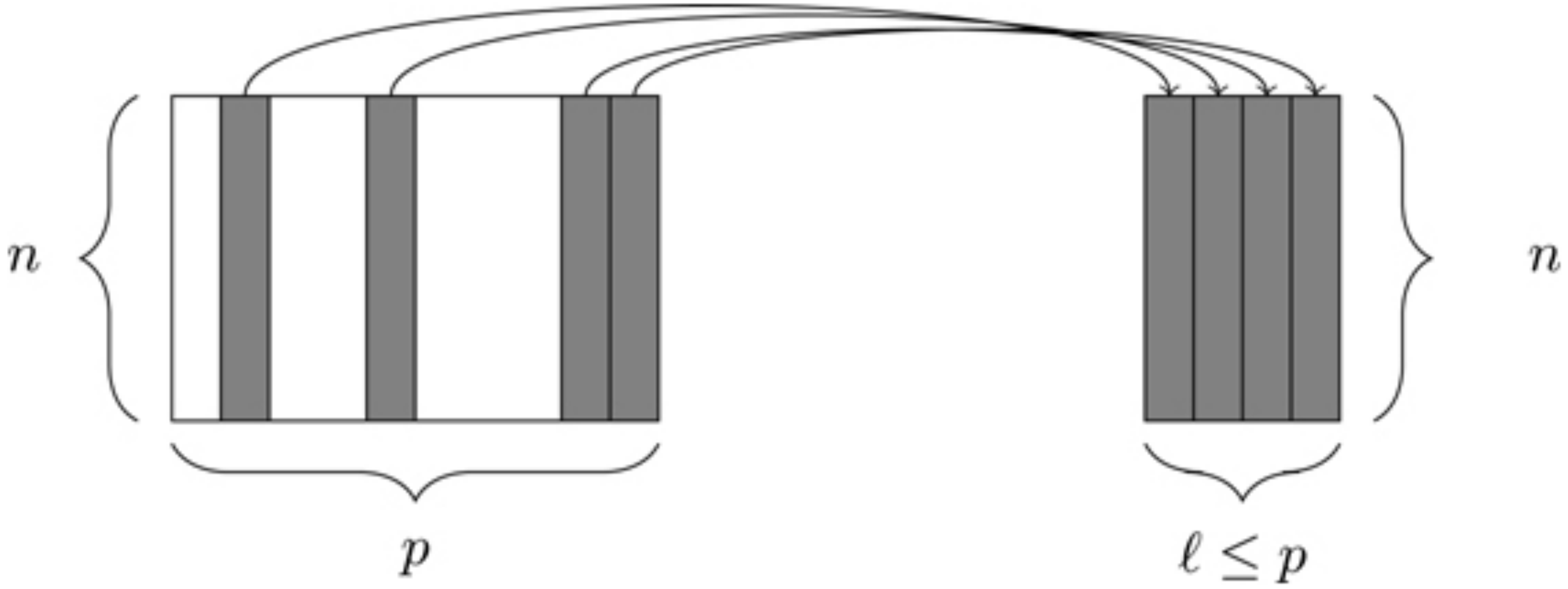
1. filter()

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



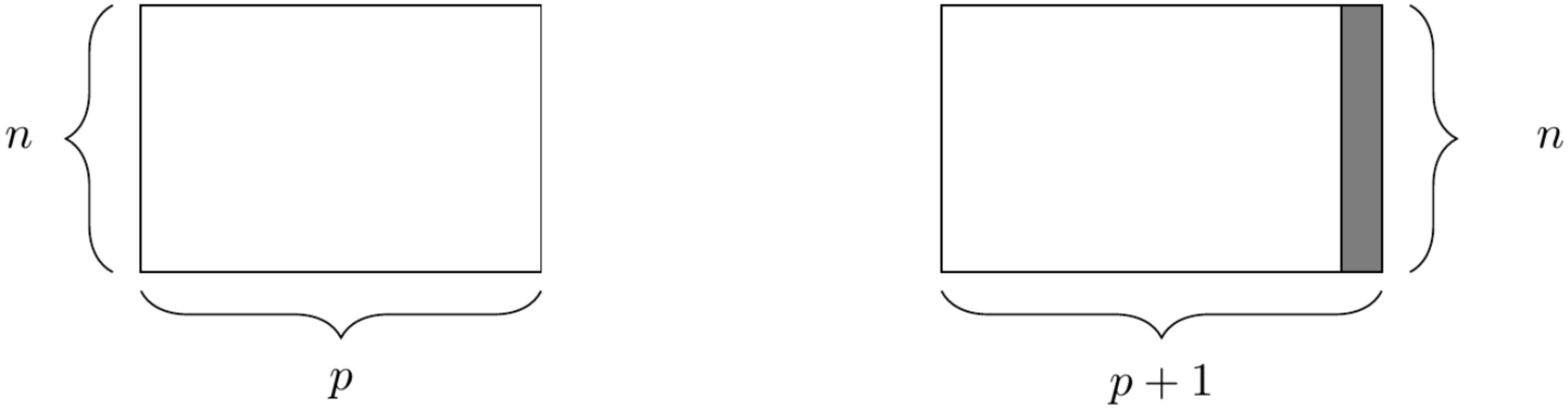
2. select()

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



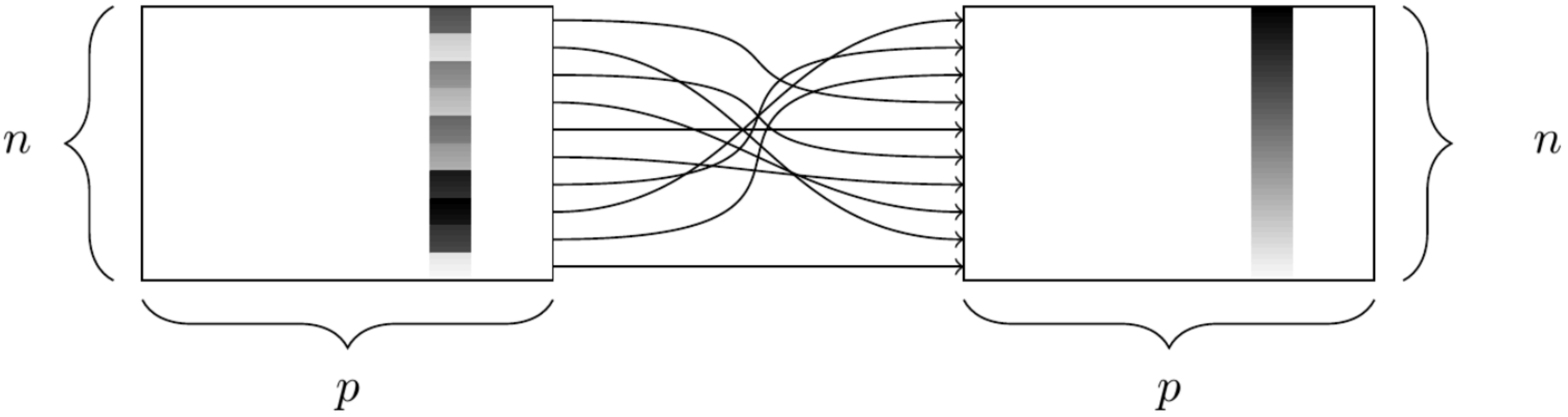
3. mutate()

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



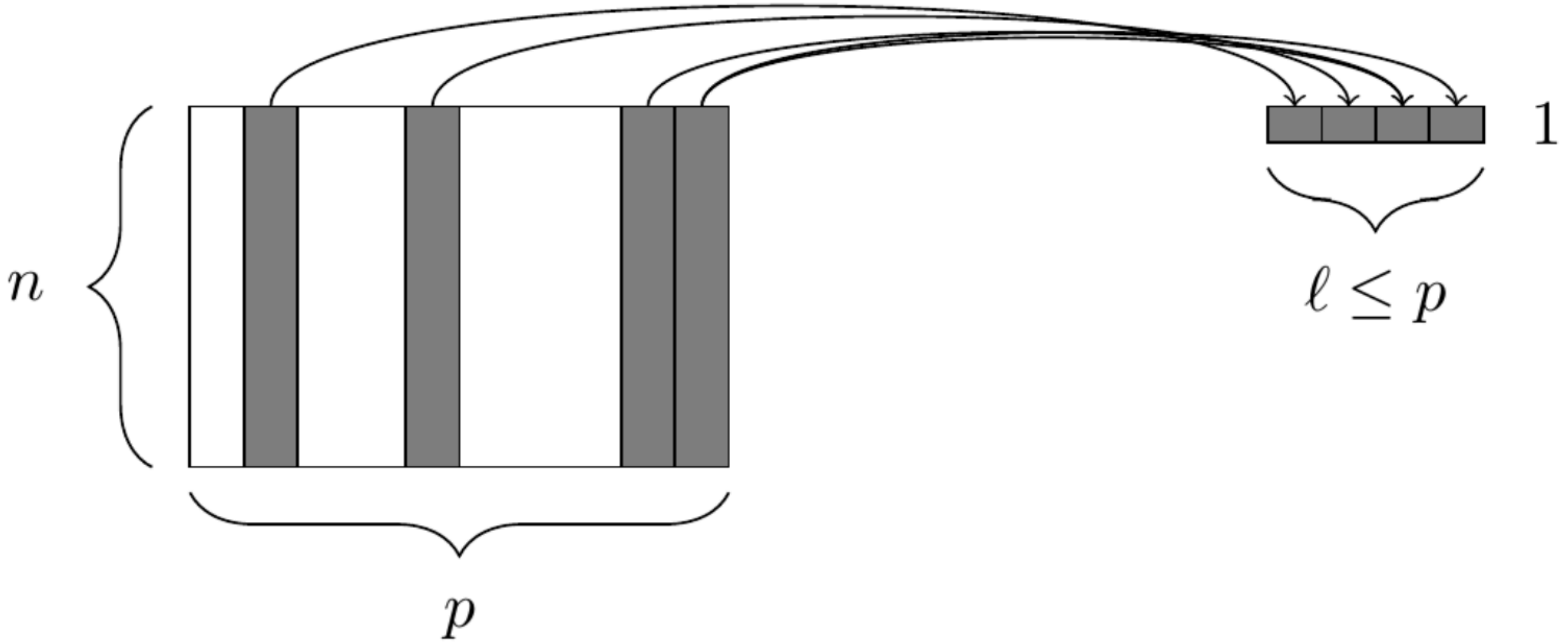
4. arrange()

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



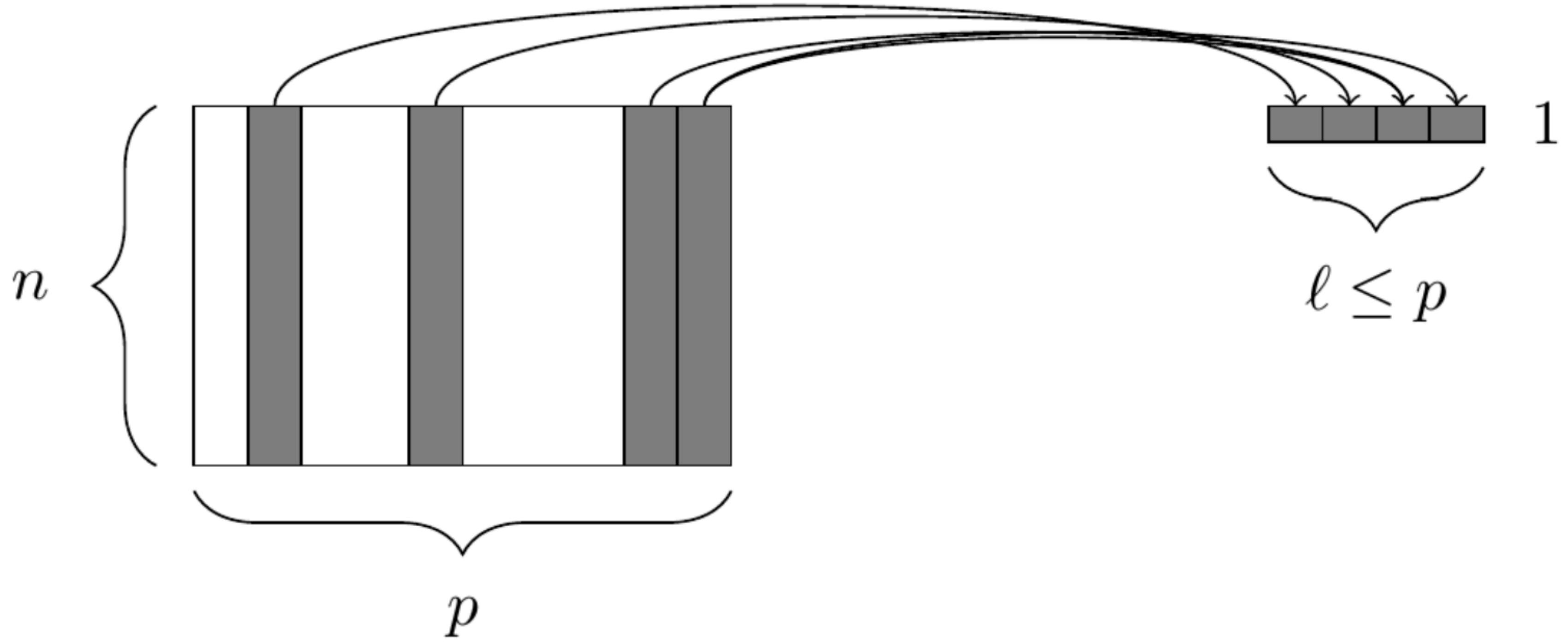
5. summarize()

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



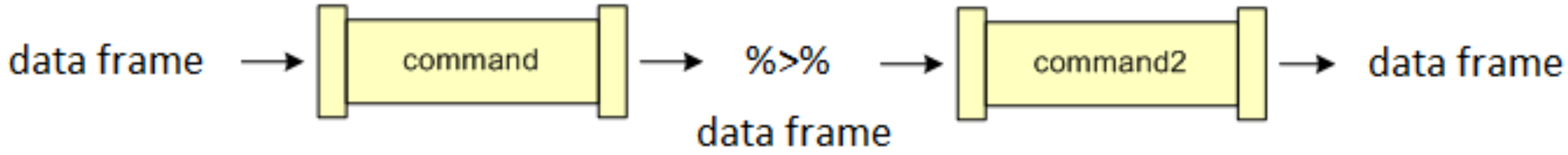
6. The group_by() function

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!

Also, don’t forget to fill out the week 5 survey