Before we start

Some organizational details

Sarah Kaspar

Biostatistical Basics 2021

EMBL

Course outline

Day	Title	Topics
1	**Summarizing and visualizing data**	How to work with data frames Use ggplot2 to create graphics Make graphs informative
2	**Statistical distributions**	what is sampling? what is a probability distribution? How can we fit data to a distribution?
3	**Statistical tests**	How statistical tests work Binomial test, T-test Non-parametric tests
4	**Categorical data + Multiple testing**	Contingency tables Test for independence Measures of association p-value adjustment + histogram

Each day:

- lecture
- demonstration in R
- tutored exercises
- discussion of solutions

Practical aspects

EMBL

Questions:

- allowed any time
    - unmute
    - raise hand
    - chat

Course homepage:

- slides
- demonstrations
- exercises

Exercises

- not all the functions needed are covered in the course
- take your time
- google
- ask your team mates
- ask your tutors
- solutions on the next course day
- help your collegues

Data exploration

Day 1

Sarah Kaspar

Biostatistical Basics 2021

EMBL

Tools

programming language

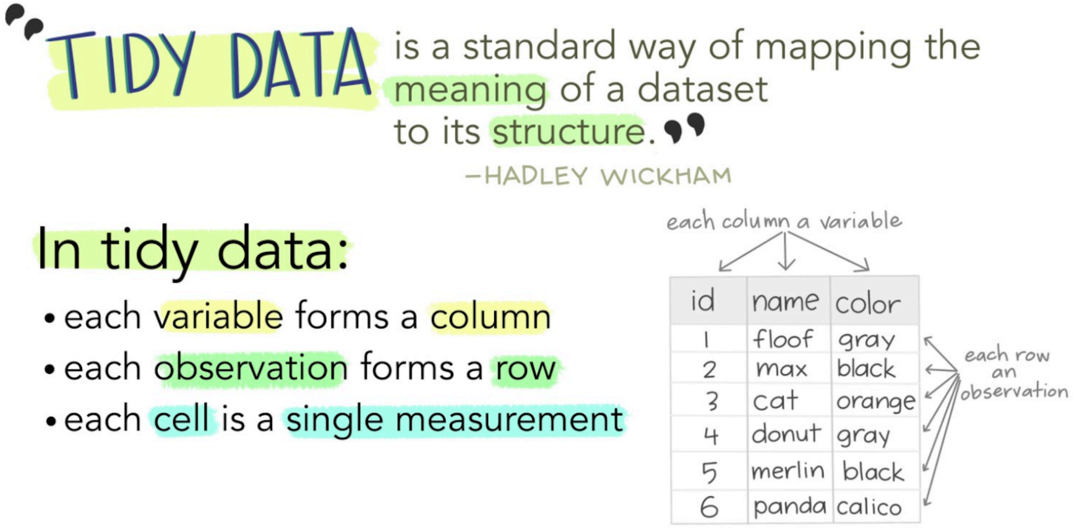
R packages that make data science user-friendly

Graphical user interface for R

data sets and software for analyzing biological data

Images:
https://www.r-project.org/logo/Rlogo.png
http://www.bioconductor.org/images/logo/jpg/bioconductor_logo_rgb.jpg
https://tidyverse.tidyverse.org/articles/tidyverse-logo.png
https://www.rstudio.com/wp-content/uploads/2018/10/RStudio-Logo.png

Tidy data



> "TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
>
> —HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst

Tidy data

Question: what should be the rows and columns in that table, if you want to tidy it up?

assessment	Billy	Suzy	Lionel	Jenny
quiz1	NA	F	B	A
quiz2	D	NA	C	A
test1	C	NA	B	B

name	quiz1	quiz2	test1
Billy	NA	D	C
Suzy	F	NA	NA
Lionel	B	C	B
Jenny	A	A	B

example from: https://tidyr.tidyverse.org/articles/tidy-data.html

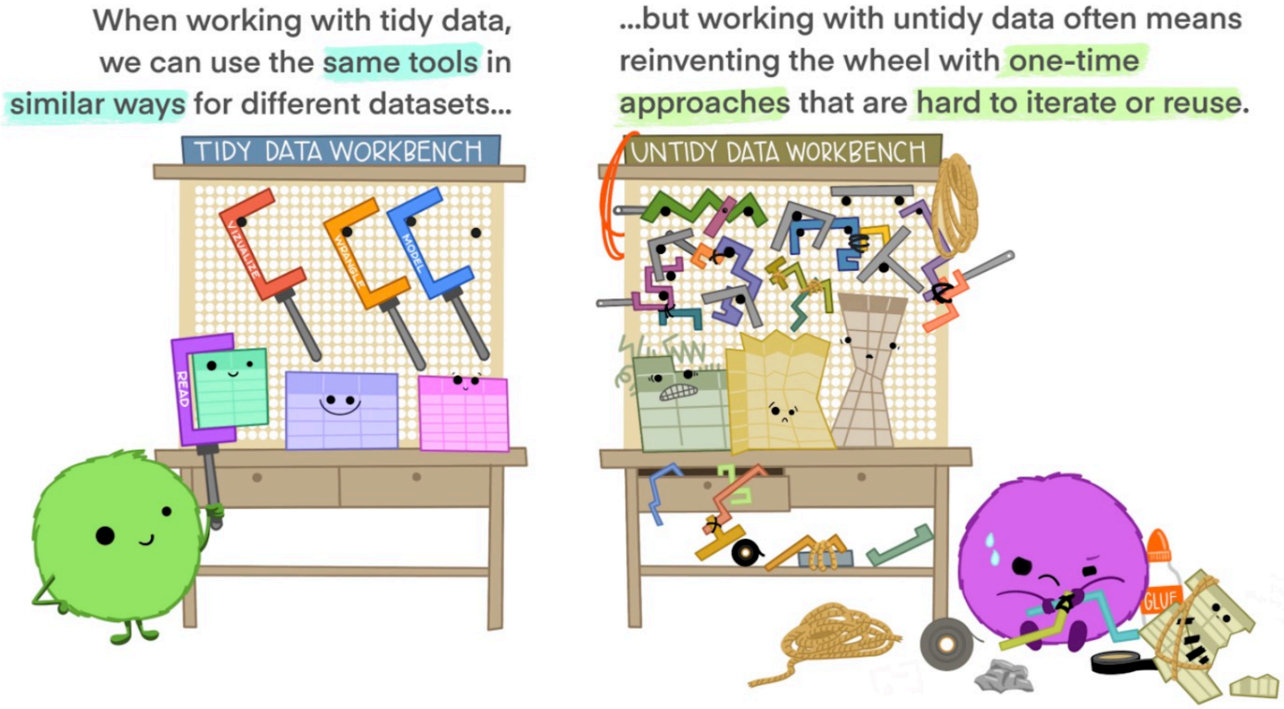
Tidy data

name	assessment	grade
Billy	quiz1	NA
Billy	quiz2	D
Billy	test1	C
Jenny	quiz1	A
Jenny	quiz2	A
...	...	...

- Every combinaton of name, assessment and grade is single observation.
- Every column is a variable (name, assessment, grade).
- Each cell is a single value.

example from: https://tidyr.tidyverse.org/articles/tidy-data.html

Tools for tidy data


When working with tidy data, we can use the same tools in similar ways for different datasets...
...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.
TIDY DATA WORKBENCH
VIZUALIZE
WRANGLE
MODEL
READ
UNTIDY DATA WORKBENCH
GLUE

Illustrations from the Openscapes blog *Tidy Data for reproducibility, efficiency, and collaboration* by Julia Lowndes and Allison Horst

Motivation

What is statistics?

A **summary statistic** “quantitatively describes or summarizes features from a collection” (Wikipedia)

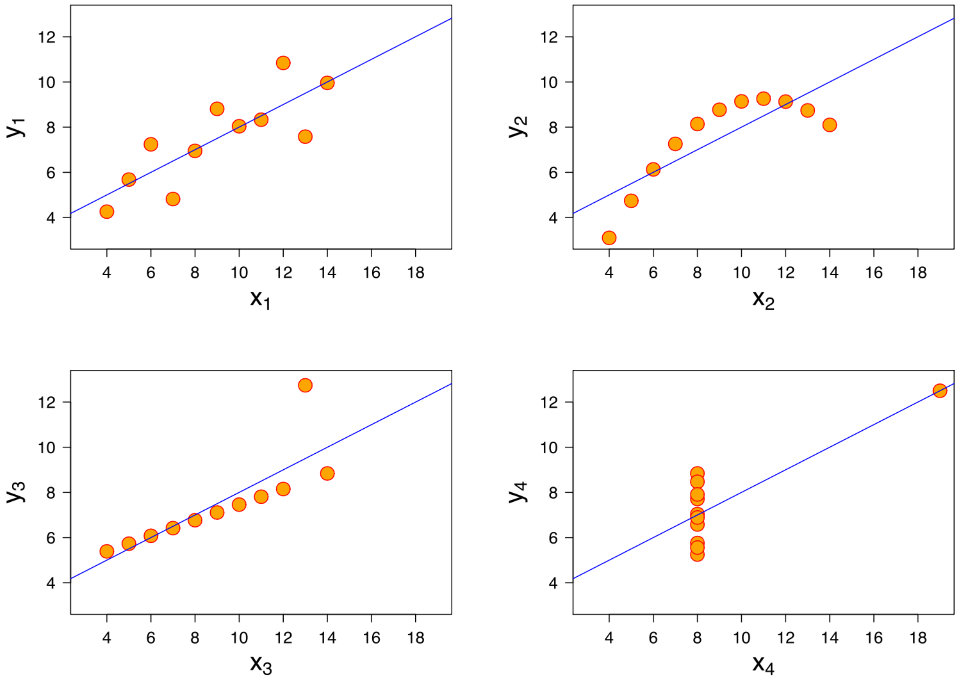
examples: mean, median, min, max,...

“**Inferential statistical analysis** infers properties of a population” (Wikipedia)

examples: hypothesis testing, t-test

Anscombe quartett

Why is a summary statistic not enough when exploring data?



All four data sets have the same mean, variance, correlation and regression line.

→ Whenever possible, plot the data points!

By Anscombe.svg: Schutz(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

Visual tools for data summary

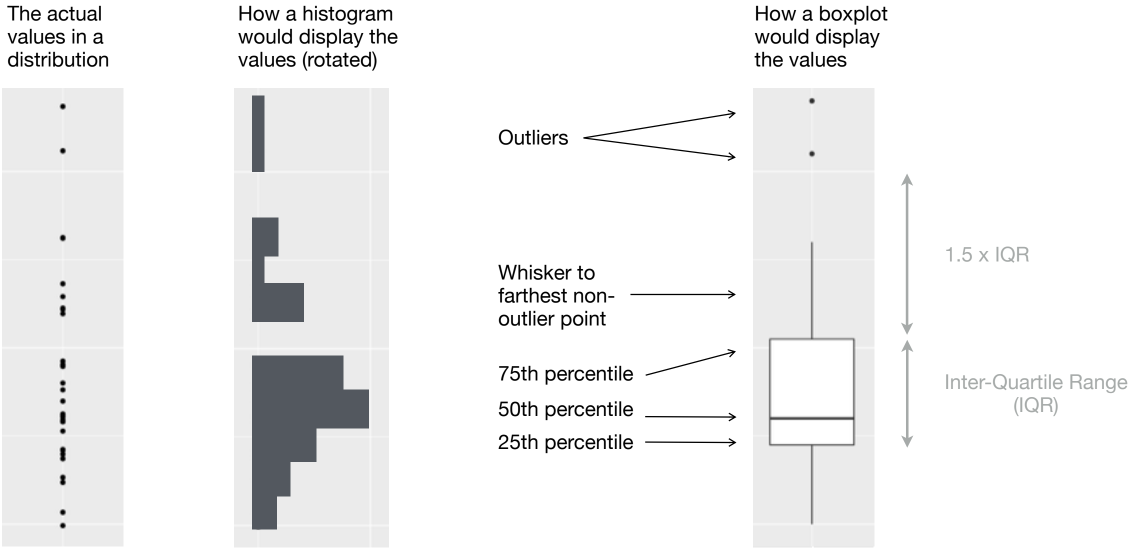

The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

Image from: https://github.com/hadley/r4ds/blob/master/images/EDA-boxplot.pdf