Improving your data visualisations in R

(No more bar plots:
How to visualise your data better!)

Dr Sophie Hardy

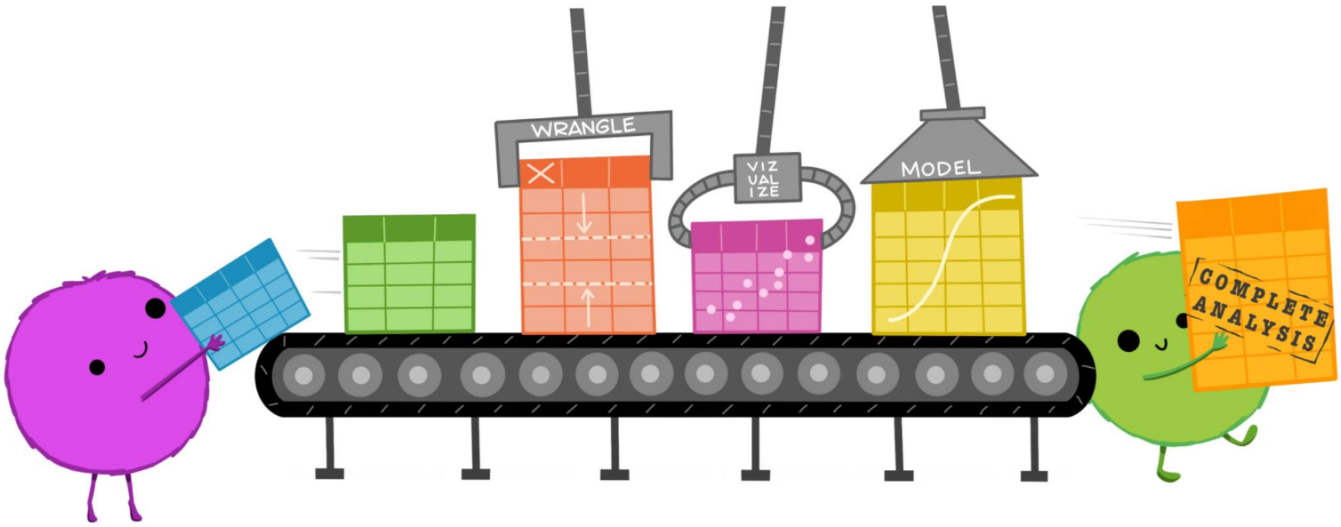
sophiehardy.co.uk

About me

@SophieMHardy

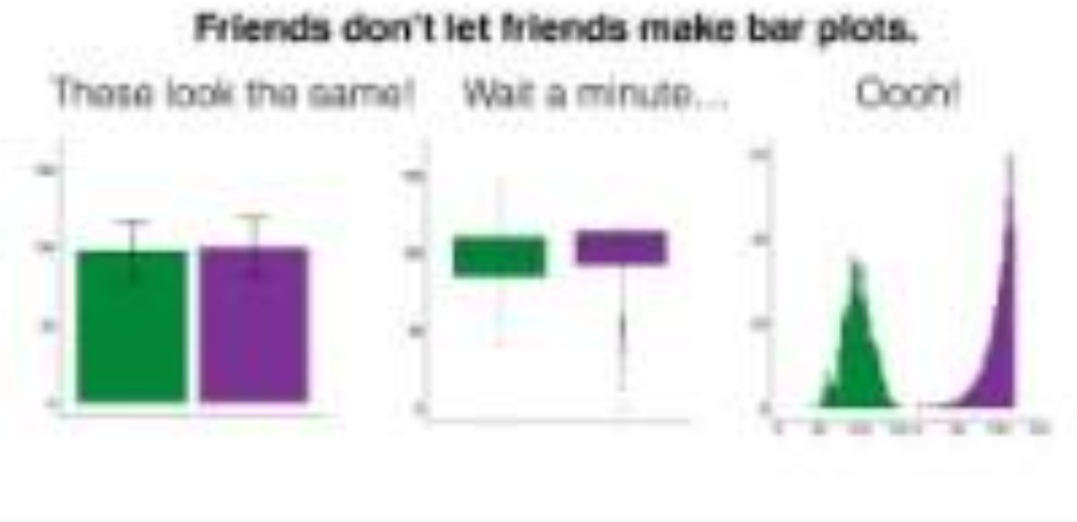
- PhD, University of Birmingham (UK)
- Postdoctoral Research Fellow, University of Warwick (UK)
- Research language processing across the lifespan
- Learned a lot (and still learning) about data visualisation from lots of different people (a lot on them on Twitter!)

The data conveyor belt



Allison Horst R illustrati

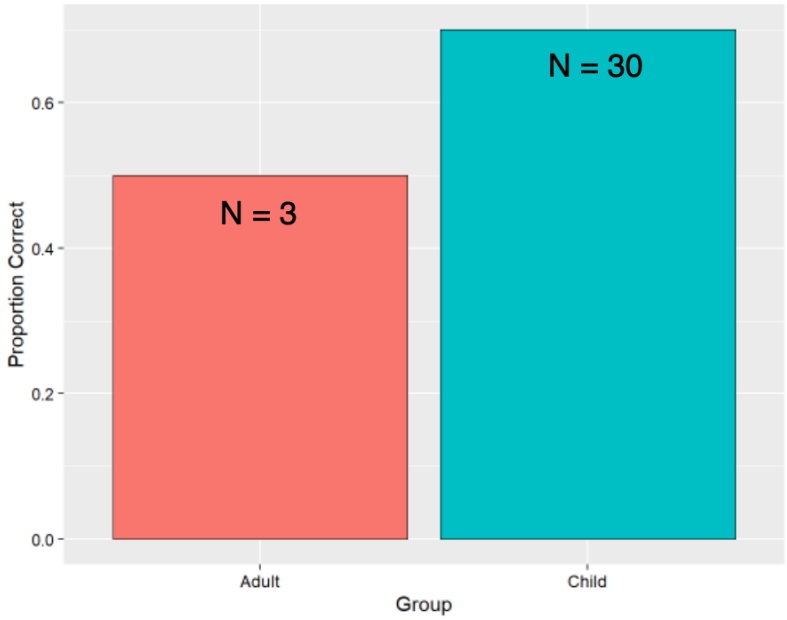
#barbarplots campaign (circa 2016)



http://barbarplots.github.io/index.htm

The issues with bar plots

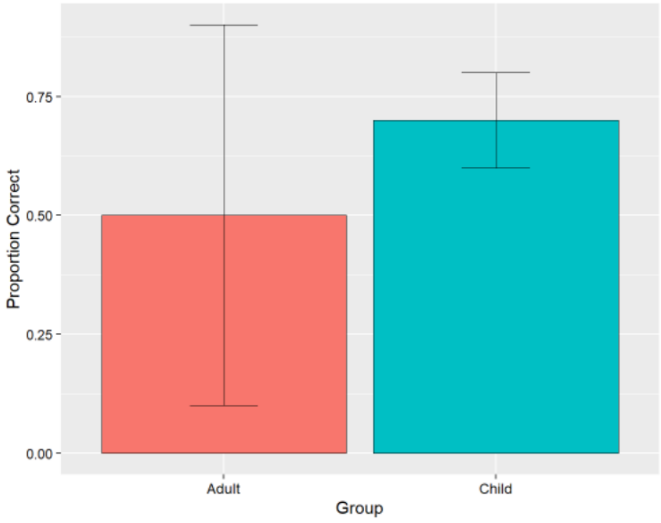
(1) Hide sample size

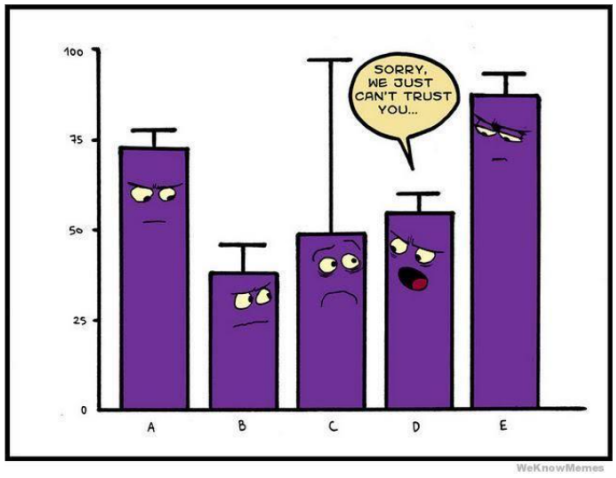


The issues with bar plots

But I use standard deviation / standard error bars?!

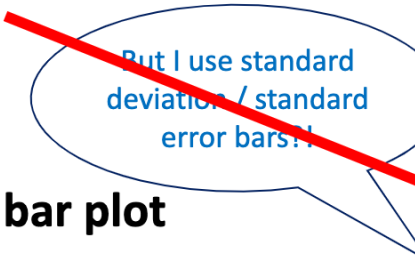
(1) Hide sample size



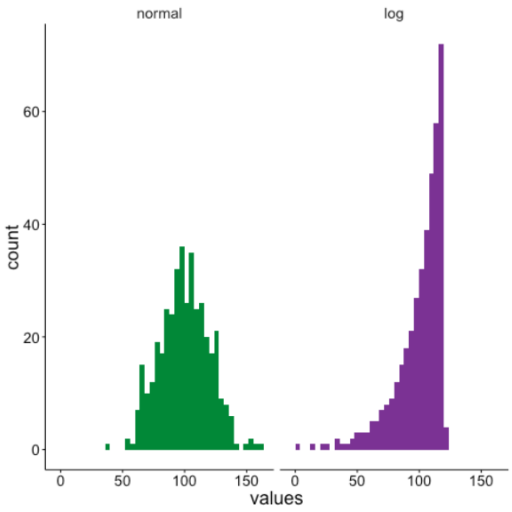


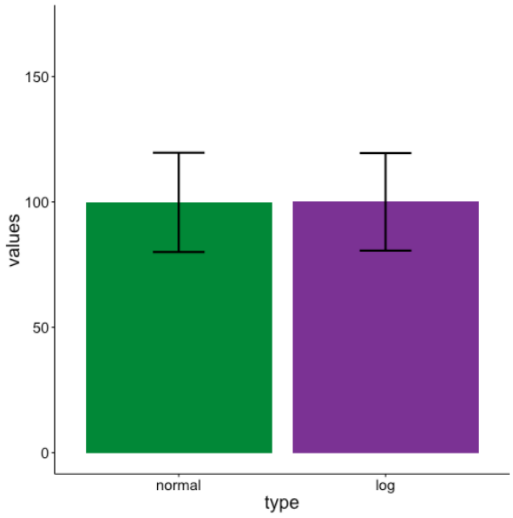
Trust none of them!

The issues with bar plots



(2) Many different distributions can lead to the same bar plot



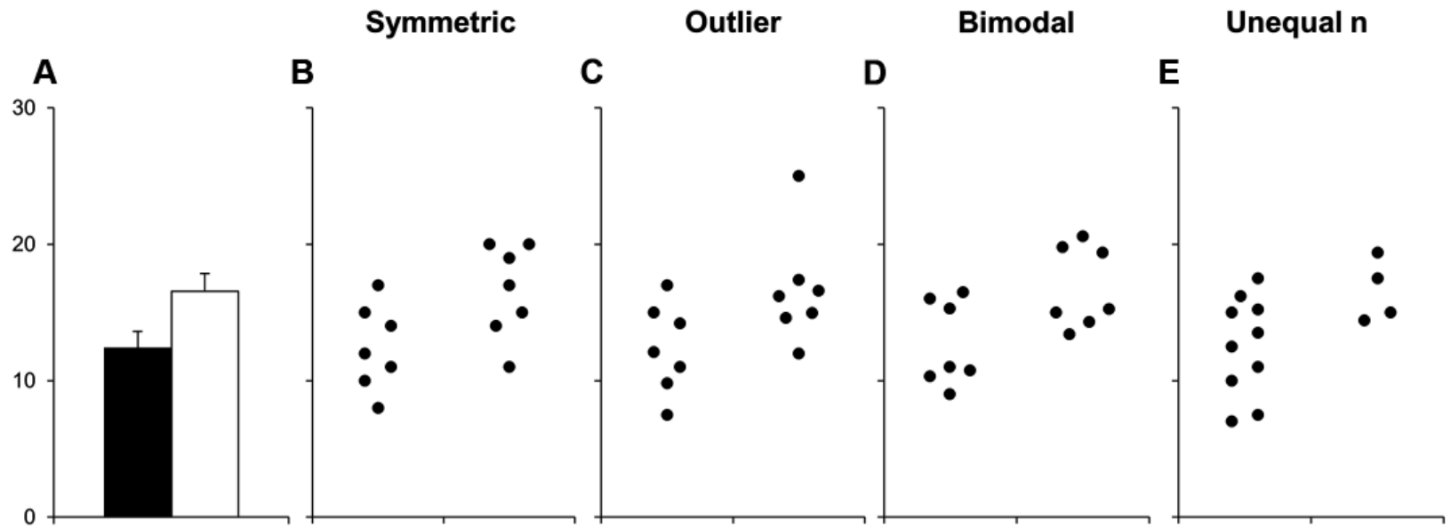


BOTH: Mean = 100; SD = 20

Piccinini (2016

The issues with bar plots

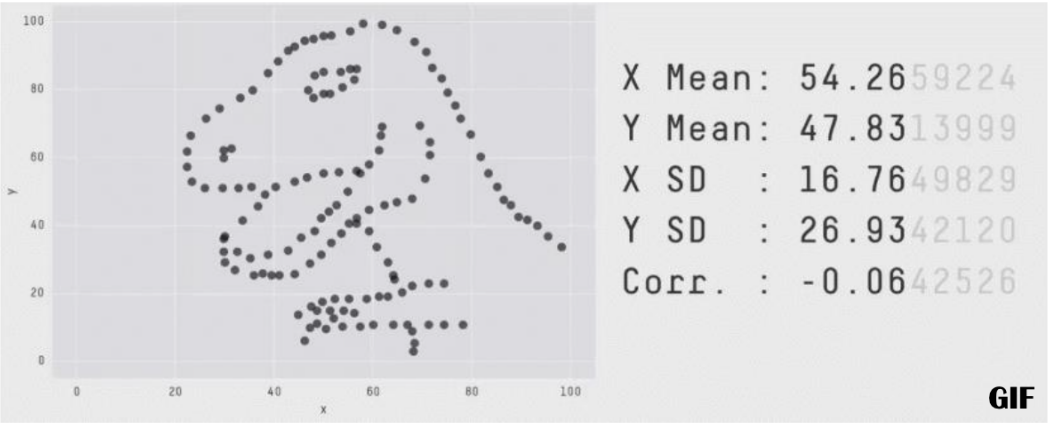
(2) Many different distributions can lead to the same bar plot

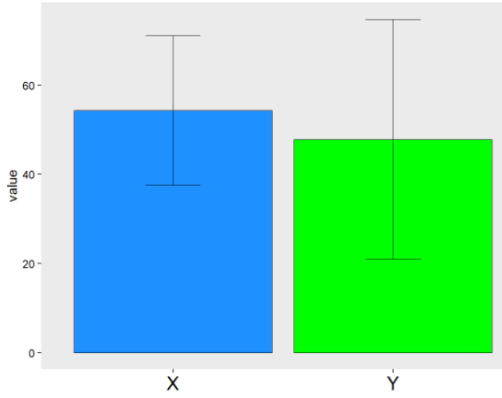


Weissgerber et al. (2015

The issues with bar plots

(2) Many different distributions can lead to the same bar plot

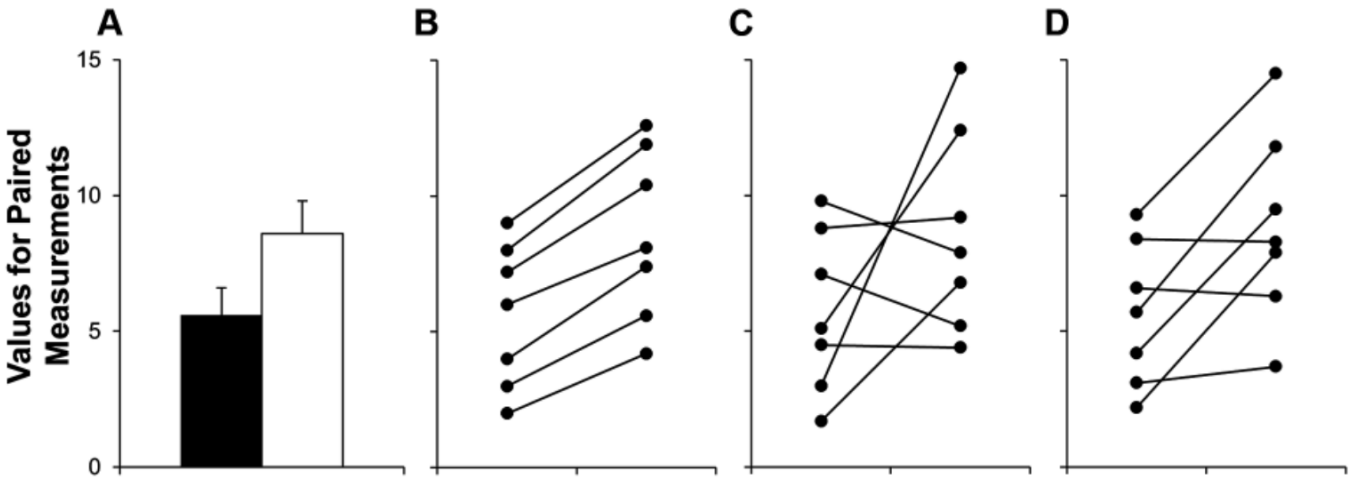




DATASAURUS S
Matejka and Fitzmaurice (201
(If in PDF, click link to see moving g

The issues with bar plots

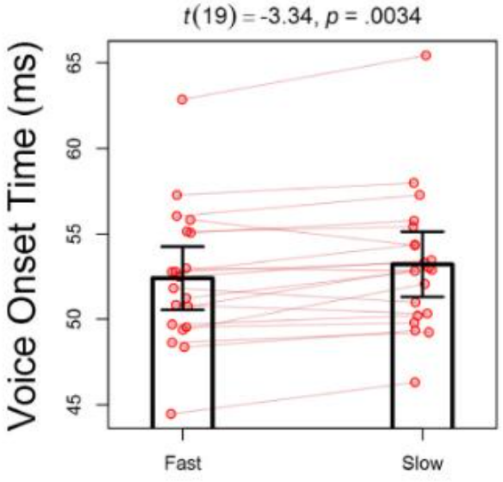
(3) Do not provide information about consistency across individuals

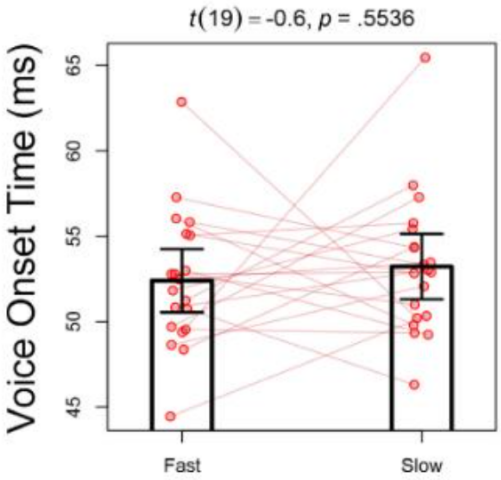


Weissgerber et al. (2015

The issues with bar plots

(3) Do not provide information about consistency across individuals

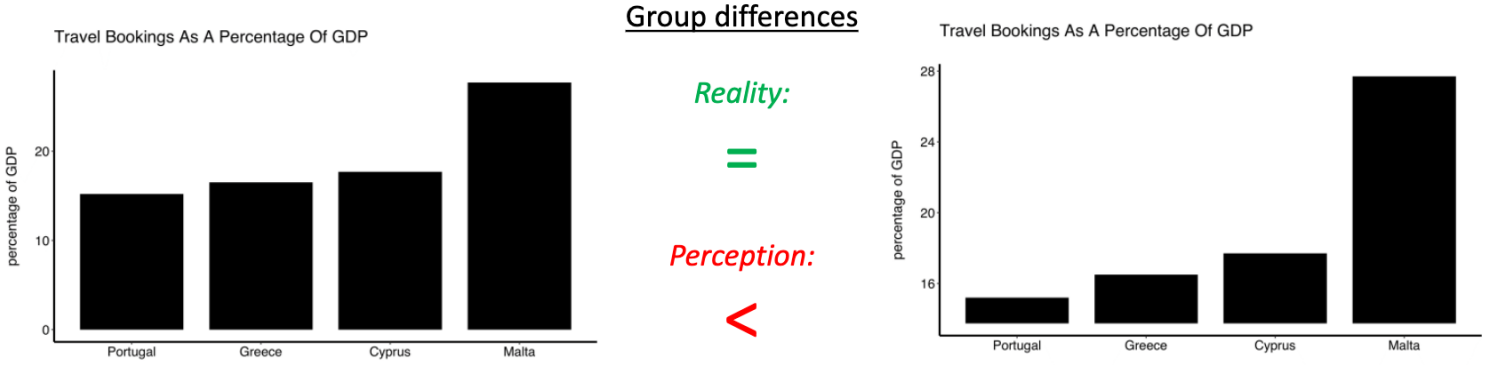




Politzer-Ahles and Piccinini (2018

The issues with bar plots

(4) Distort data interpretation: Truncation effect

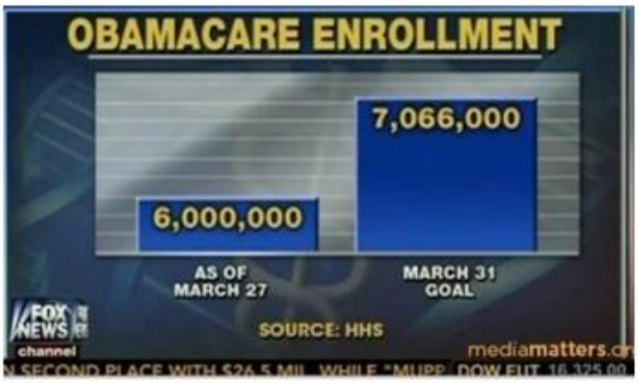


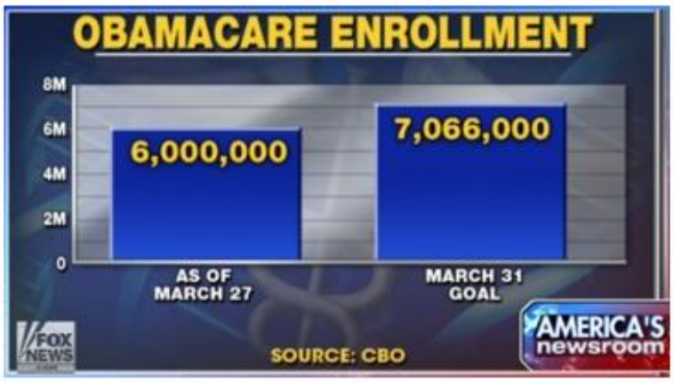
Rating: How different do you judge the values to be on this graph?

Yang et al. (2021

The issues with bar plots

(4) Distort data interpretation: Truncation effect

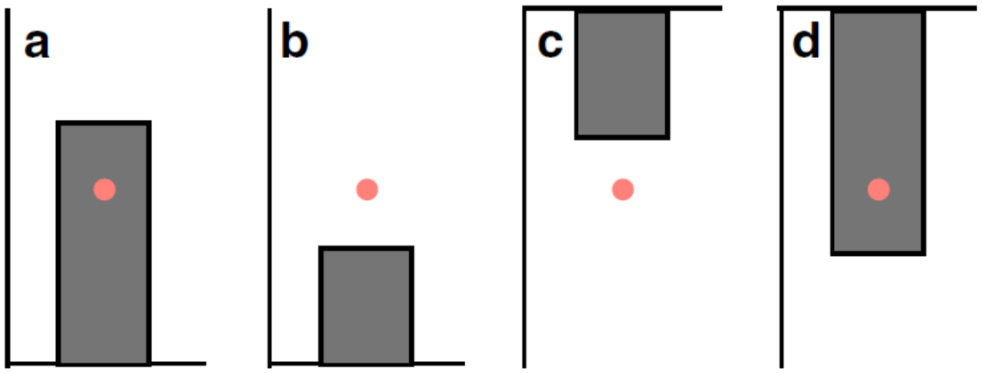




Yang et al. (2021

The issues with bar plots

(4) Distort data interpretation: Within-the-bar bias



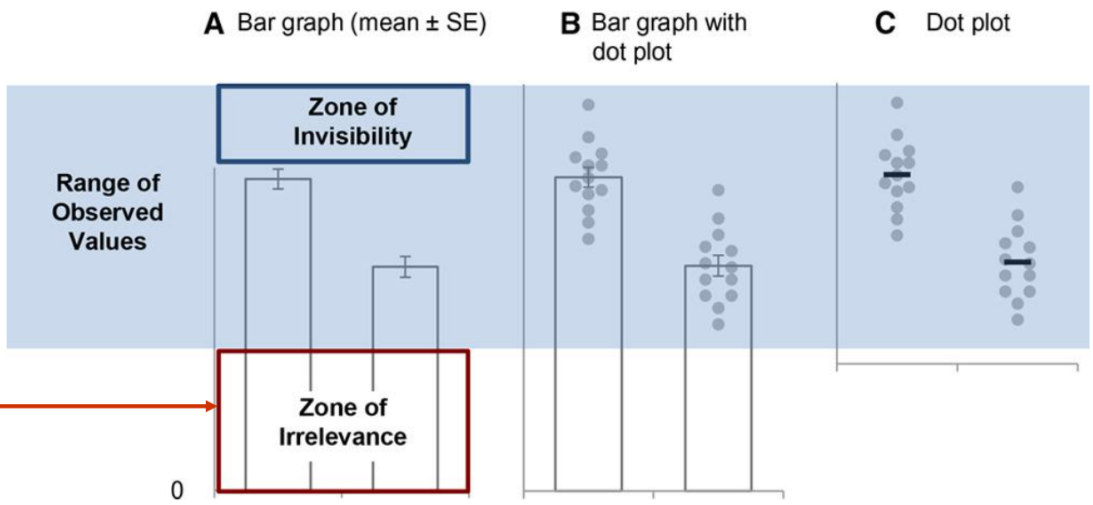
"When viewers are shown a bar depicting a mean value and are then asked to judge the likelihood of a particular data point being part of its underlying distribution, viewers judge points that fall within the bar as being more likely than points equidistant from the mean, but outside the bar—as if the bar somehow "contained" the relevant data."

Newman and Scholl (2012

The issues with bar plots

"Bar graphs arbitrarily assign importance to the height of the bar, rather than focusing attention o how the difference between means compares to t range of observed values."

(4) Distort data interpretation



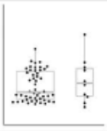
Yang et al (2021): *"Our recommendation to not truncate vertical axes is specific to bar graphs. Line graphs and dot plots, for example, do not represent numerical values as continuous visual areas and truncation may be appropriate in these cases."*

Weissgerber et al. (2019

Fine, bar plots aren't great… what should I do instead?

Allison Horst R illustrations

Fine, bar plots aren't great… what should I do instead?

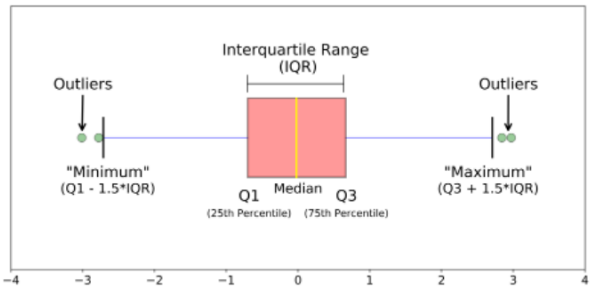
Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

Weissgerber et al. (2019

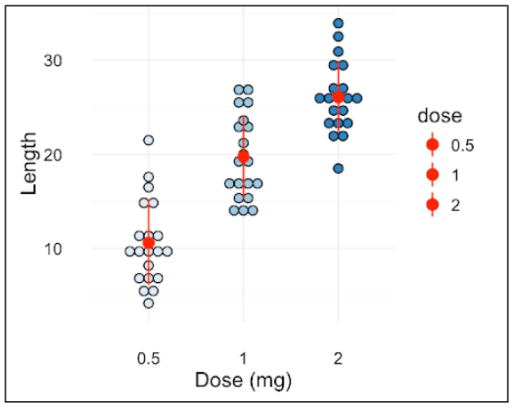
Fine, bar plots aren't great… what should I do instead?

Allen et al. (2019): *"A combined approach is most desirable as each of these visualization techniques have various advantages and trade-offs."*

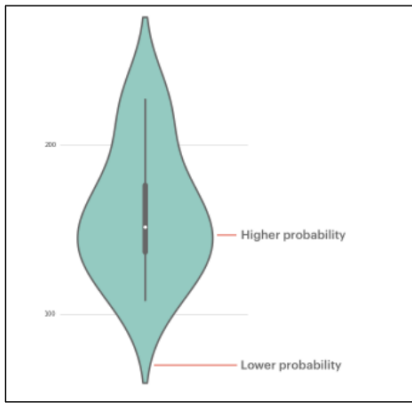
BOX PLOT



DOT PLOT

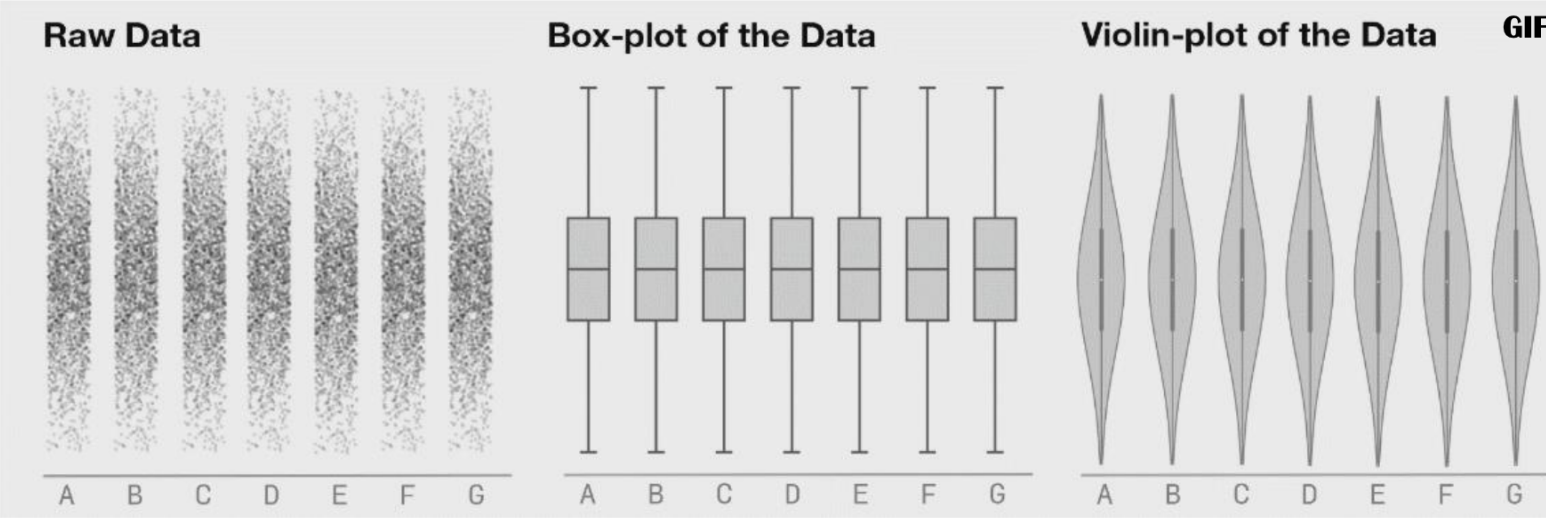


VIOLIN / BEAN PLOT



But should also not be used alone!

But don’t use alone…



Matejka and Fitzmaurice (201

(If in PDF, click link to see moving gif

Combine different approaches to show off your data in the best way!

BOX PLOT

VIOLIN / BEAN PLOT

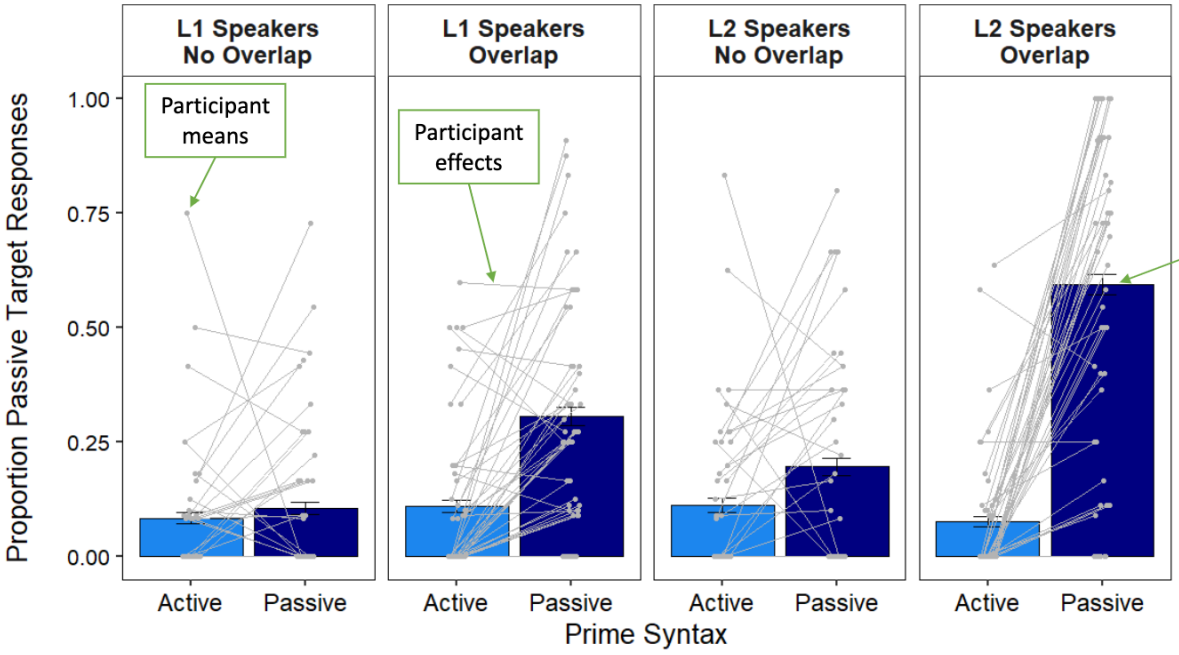
PARTICIPANT EFFECTS

ERROR BARS

DOT PLOT

BAR PLOT

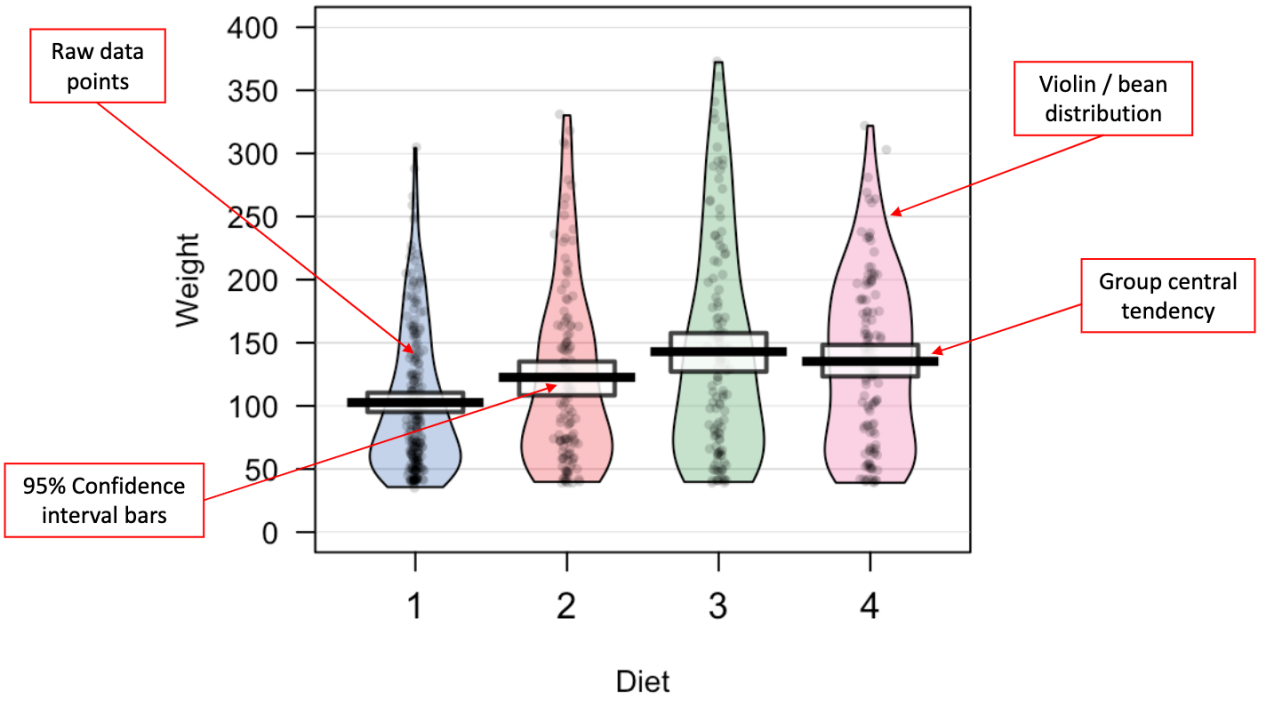
(1) JITTER PLOT



There is still a place for bar plots as long as you don't use them alone!

Coumel et al. (2020

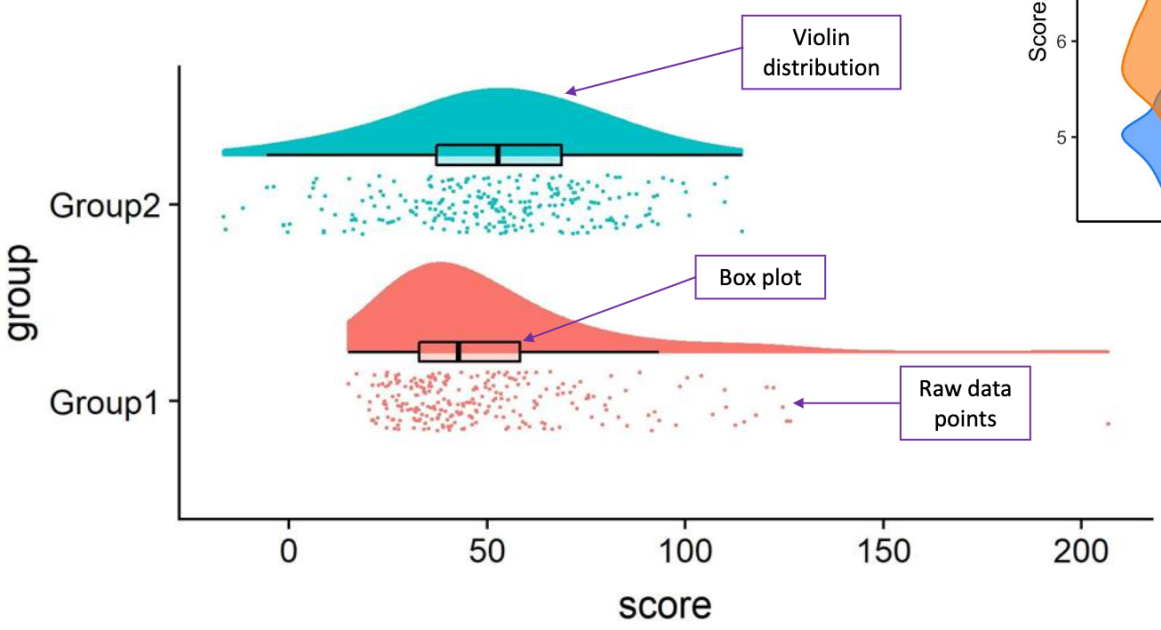
(2) PIRATE PLOT

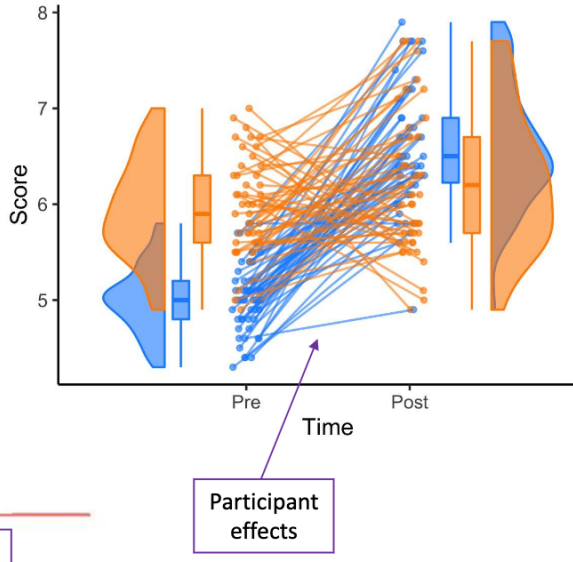


Central tendency:
- Mean
- Median
- Mode

Phillips (2018

(3) RAINCLOUD PLOT





Allen et al. (2019

Never wrong… just always improving

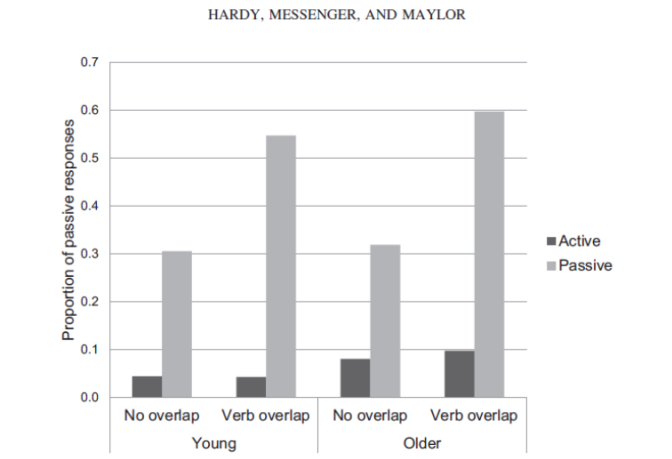


Figure 2. Mean proportion of passive responses produced by young and older adults following active and passive primes in the no-overlap and verb-overlap conditions.

Figure made in 2016

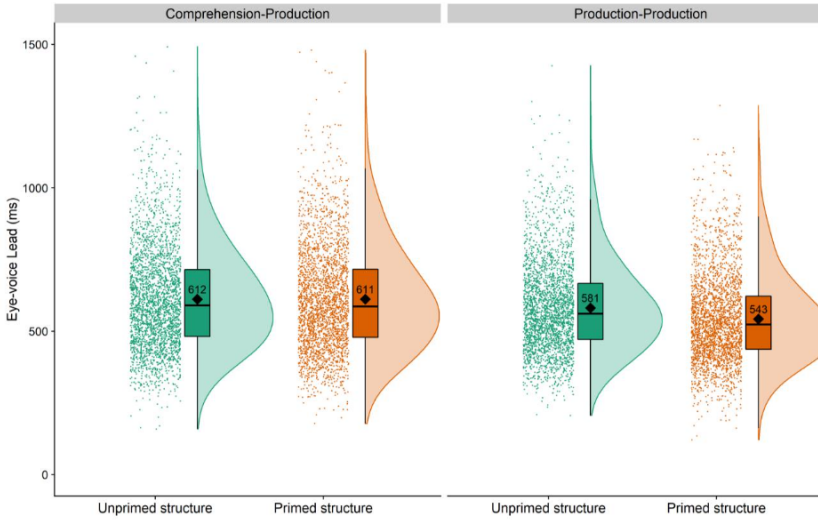


Figure made in 2020

Today's R session

(1) Jitter plot
- Categorical / discrete data
- Package: ggplot2

(2) Pirate plot
- Continuous data
- Package: yarrr

(3) Raincloud plot
- Continuous data
- Package: ggplot2 and raincloudplots

Allison Horst R illustrations

General tips for working in R

- You can always google the problem – R has the best online community and somebody will have most likely asked your question before!
- Always comment your code!
- You can assign variables using either <- or =
- Packages are your friend
- Watch out for your data types (string vs. integer vs. factor) – this can often be the cause of coding errors

Allison Horst R illustrations