Graphics

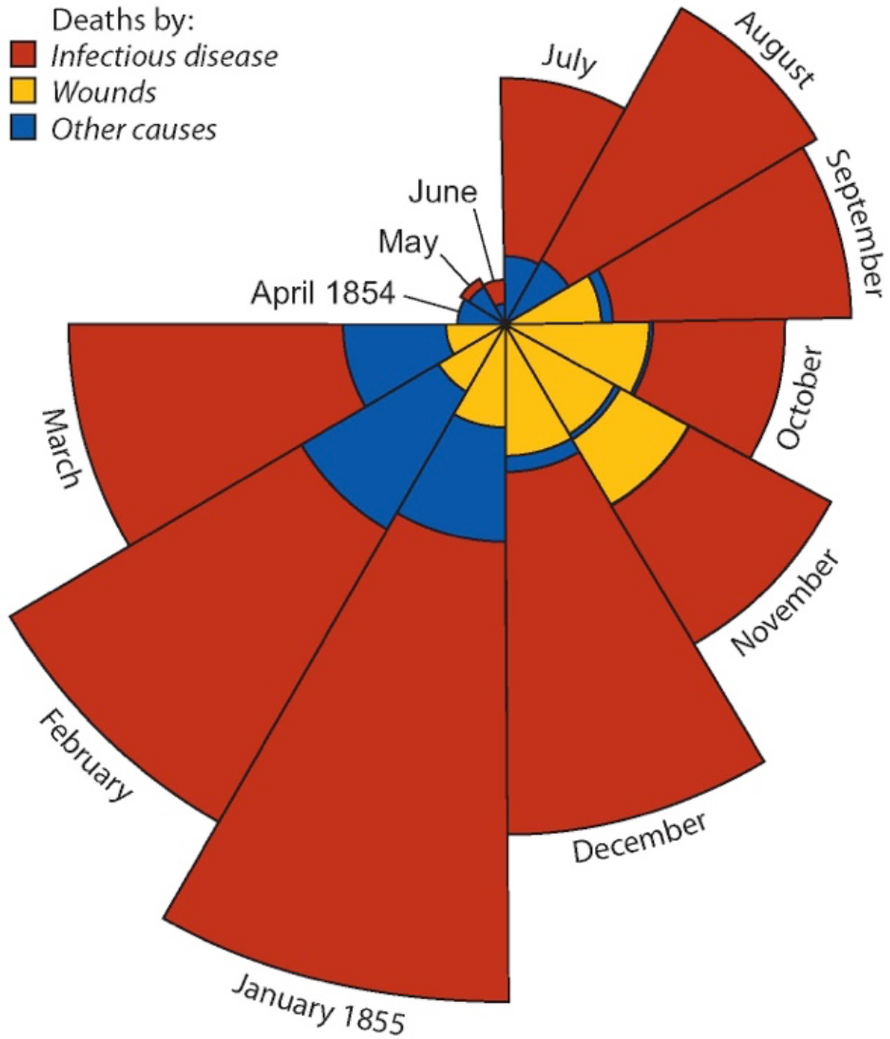
Outline for today

- Why make graphs?
- Principles of effective display
- Types of graphs to achieve these principles
- Why some graphs fail, and what can be done
- Some thoughts about tables
- Interactive graphs
- Motion

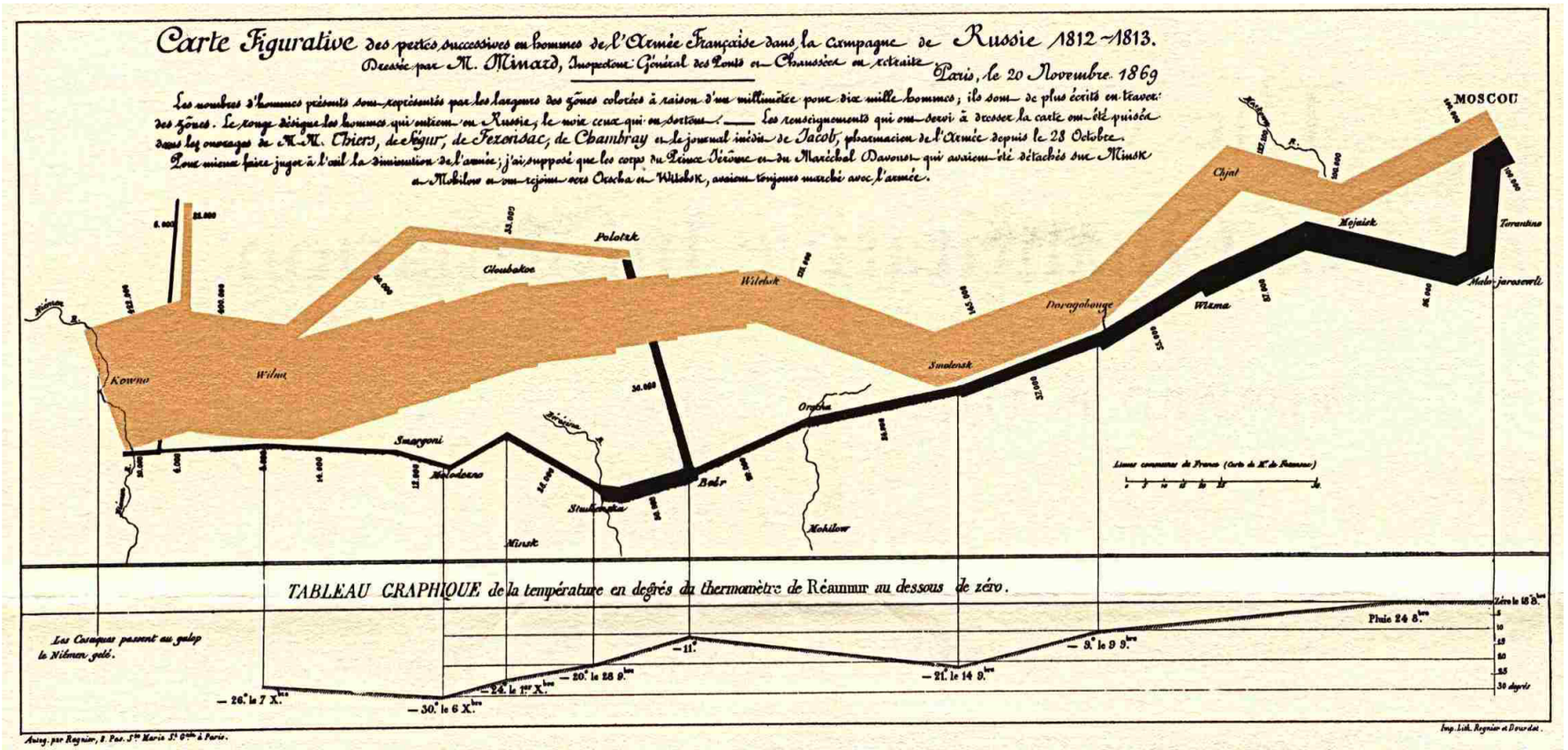
Why a graph?

- Because they can change the world!
- The human eye is a pattern detector.
- Graphs enable visual comparisons of measurements between groups and expose relationships between variables.
- They are the best method for communicating results to a wider audience

Causes of deaths in the British Army during the Crimean War (F. Nightingale 1858)
(area of pie = number of deaths)



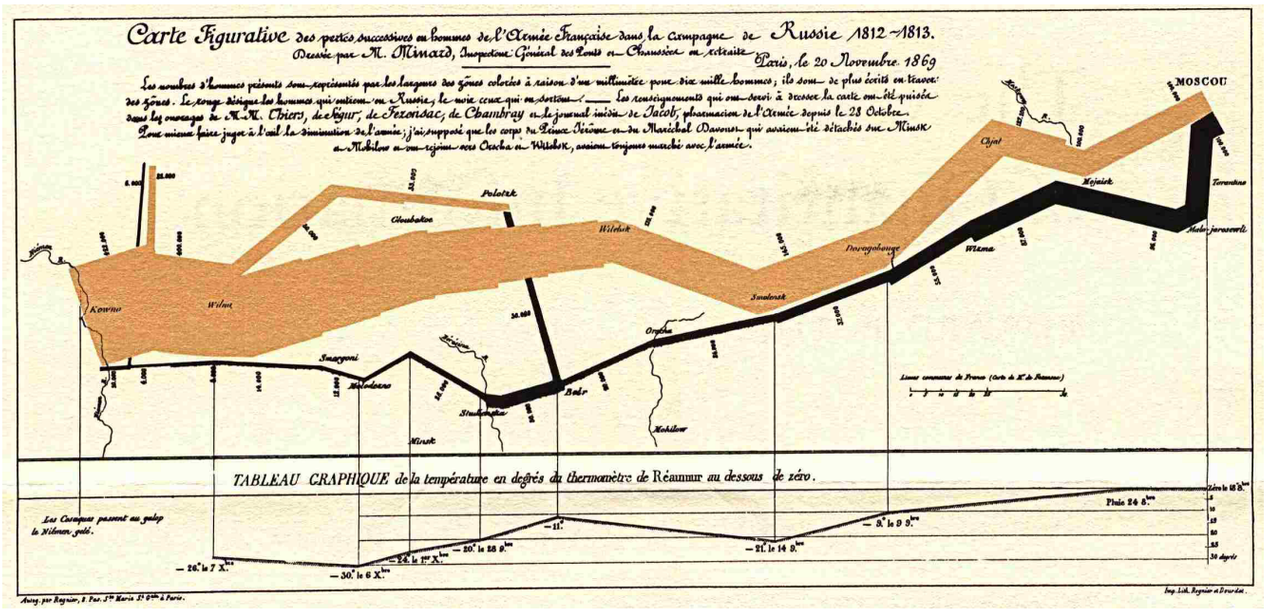
The best statistical graphic ever drawn (according to Edward Tufte)
This map by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the Polish-Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. This graph vividly depicts relative magnitudes in a way that the raw numbers alone cannot.



How to achieve graphical excellence?

Graphs should make the viewer goes “Oh!” and not “Huh?”

“Graphical excellence is that which gives to the viewer the greatest number of <u>ideas</u> in the shortest time with the least ink in the smallest space” – Tufte (1983)



When evaluating a graph, ask:

“What is it’s goal?” and then “Does it achieve it effectively?”

The goal of graphs

To reveal/show patterns in data.

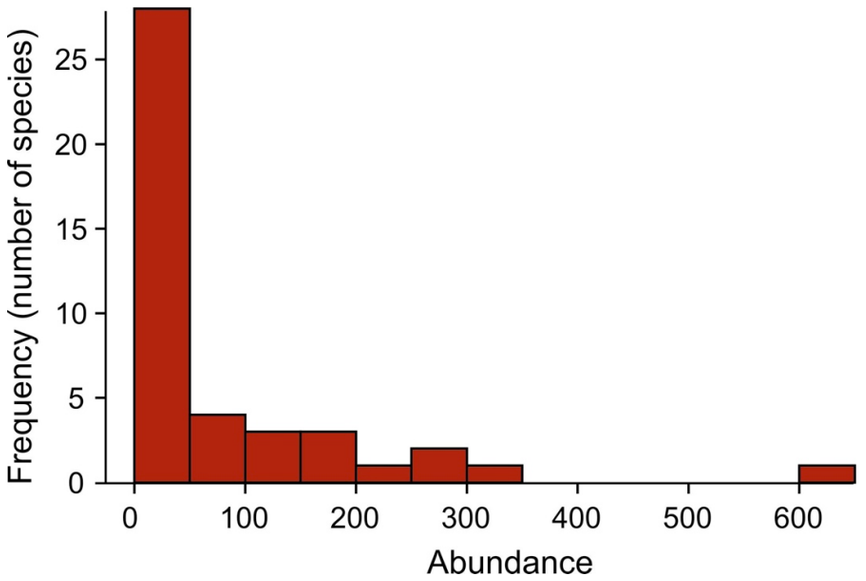
"...gives to the viewer the greatest number of ideas in the shortest time..."

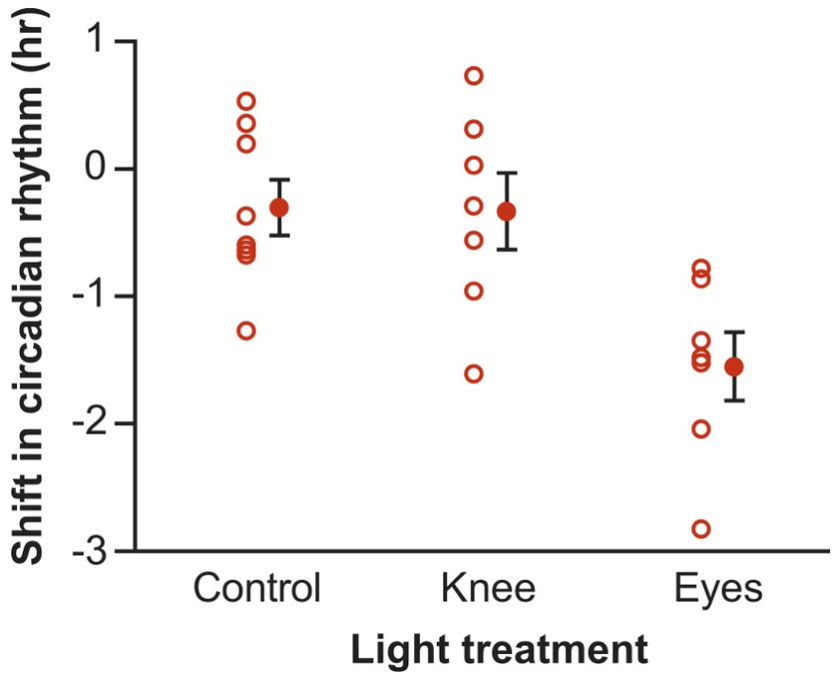
Frequency distributions

- The location, spread, shape of distribution

Associations between variables

- Relationships between variables
- Differences between groups





Best practices to achieve this goal

Four useful principles to increase the effectiveness of your graphs:

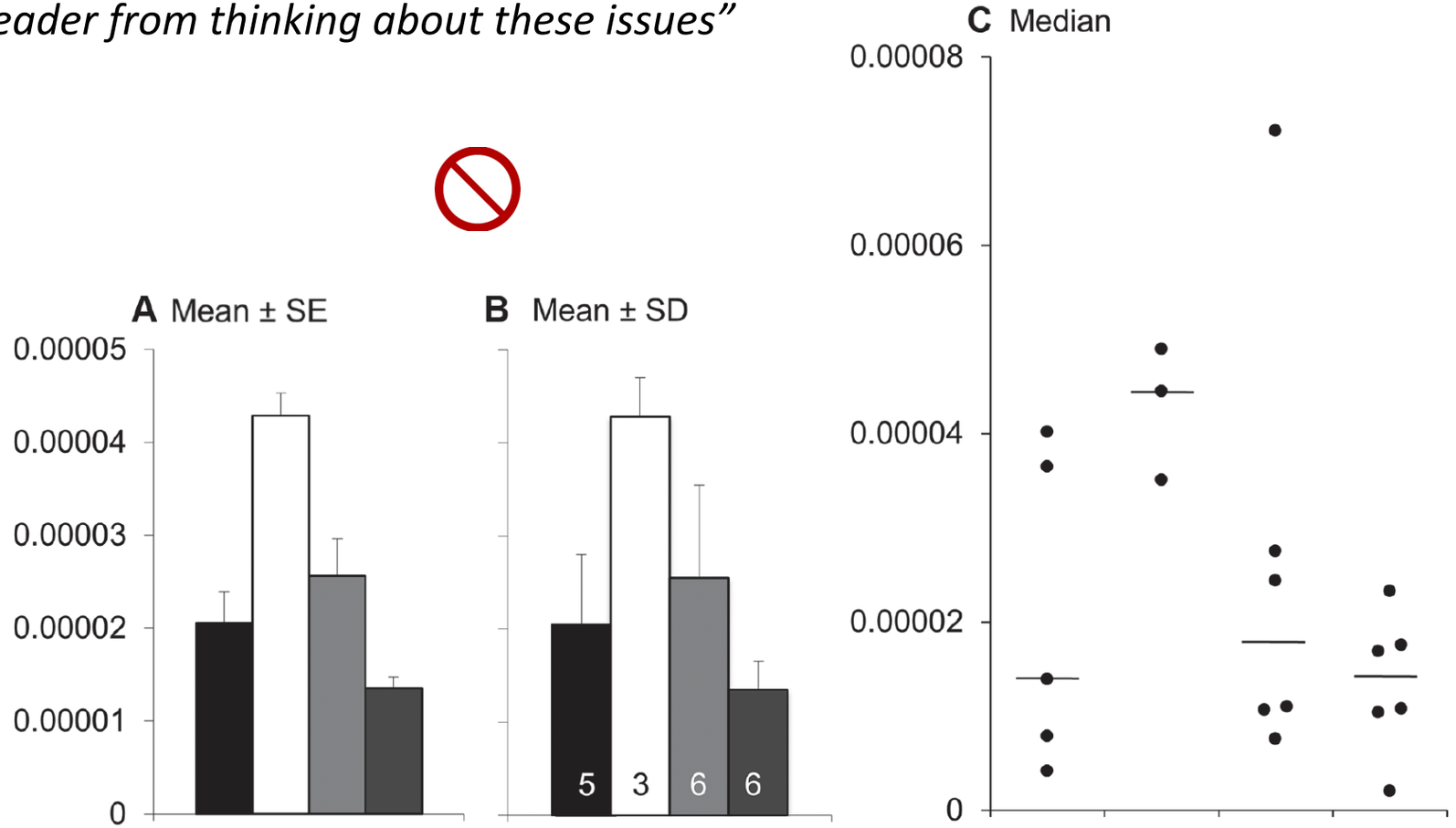
1. Show the data
2. Make patterns in the data easy to see
3. Represent magnitudes honestly
4. Draw graphical elements clearly, minimizing clutter

1. Show the data - *"Above all else show the data"* – Tufte (1983)

E.g., a strip chart reveals patterns that are hidden in the bar graph.

"While scatterplots prompt the reader to critically evaluate the statistical tests and the authors' interpretation of the data, bar graphs discourage the reader from thinking about these issues"

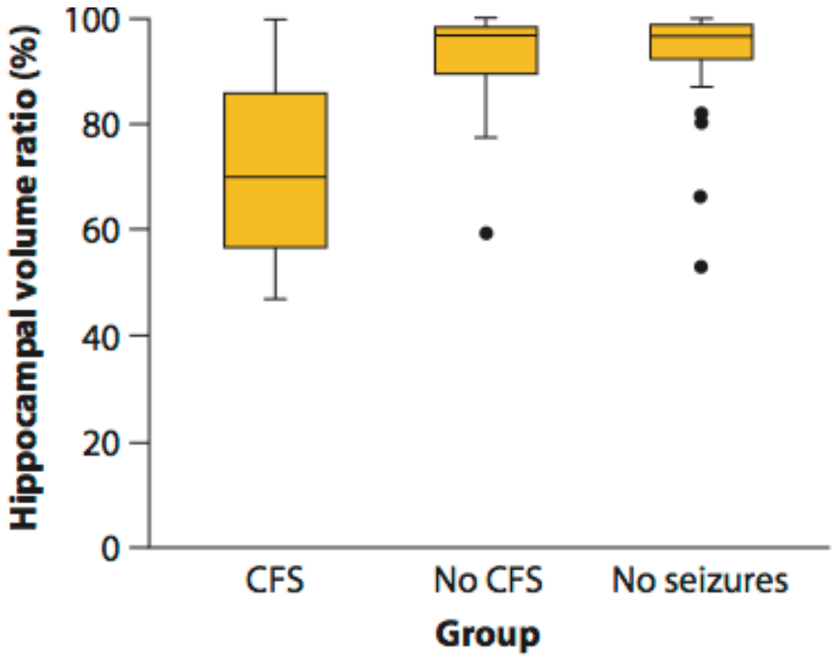
Weissgerber et al. (2015)
Beyond bar and line graphs: time for a new data presentation paradigm.
PLoS Biol.
DOI:10.1371/journal.pbio.1002128



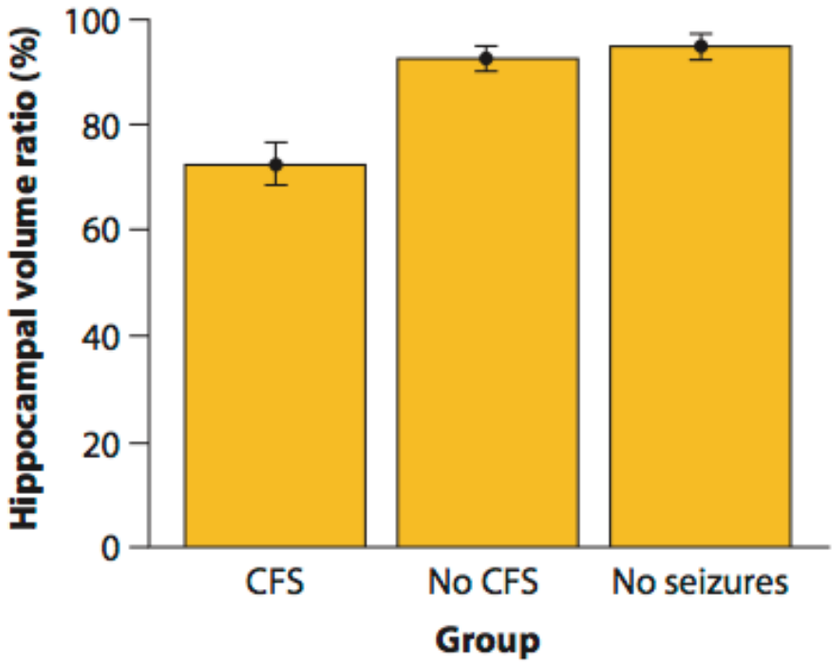
1. Show the data

Box plots can substitute for strip chart when there are many data points.

Which graph is more effective? Why?

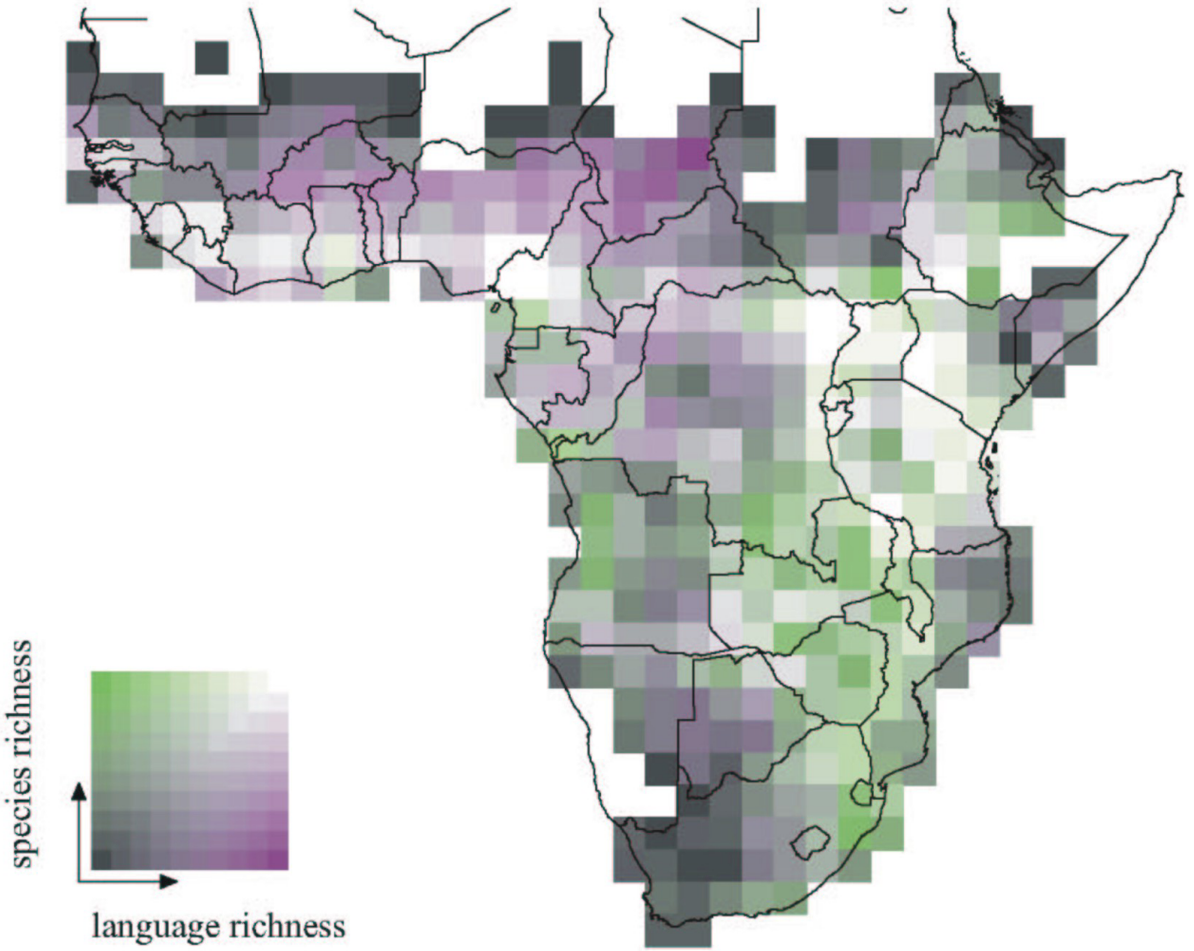


The graphs at the right are from a study investigating hippocampal volume loss in 107 patients with drug-resistant epilepsy (Cook et al. 1993). The graphs depict the association between hippocampal volume loss (measured using MRI as the volume of the smaller half of the hippocampus divided by the volume of the larger half, expressed as a percentage) and patient history. Patients were grouped on the basis of whether they had a record of childhood febrile seizures (CFS), childhood non-febrile seizures (no CFS) and no childhood seizures.



2. Make patterns in the data easy to see.

"Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency" – Tufte (1983)



Map displaying the number of bird species and the number of distinct human languages present in each square of a grid of continental Africa. Reproduced from Moore et al. (2002).

What is the pattern in these data?
How long did it take you to "see"?
Is it easy to appreciate how strong the relationship is between the variables?

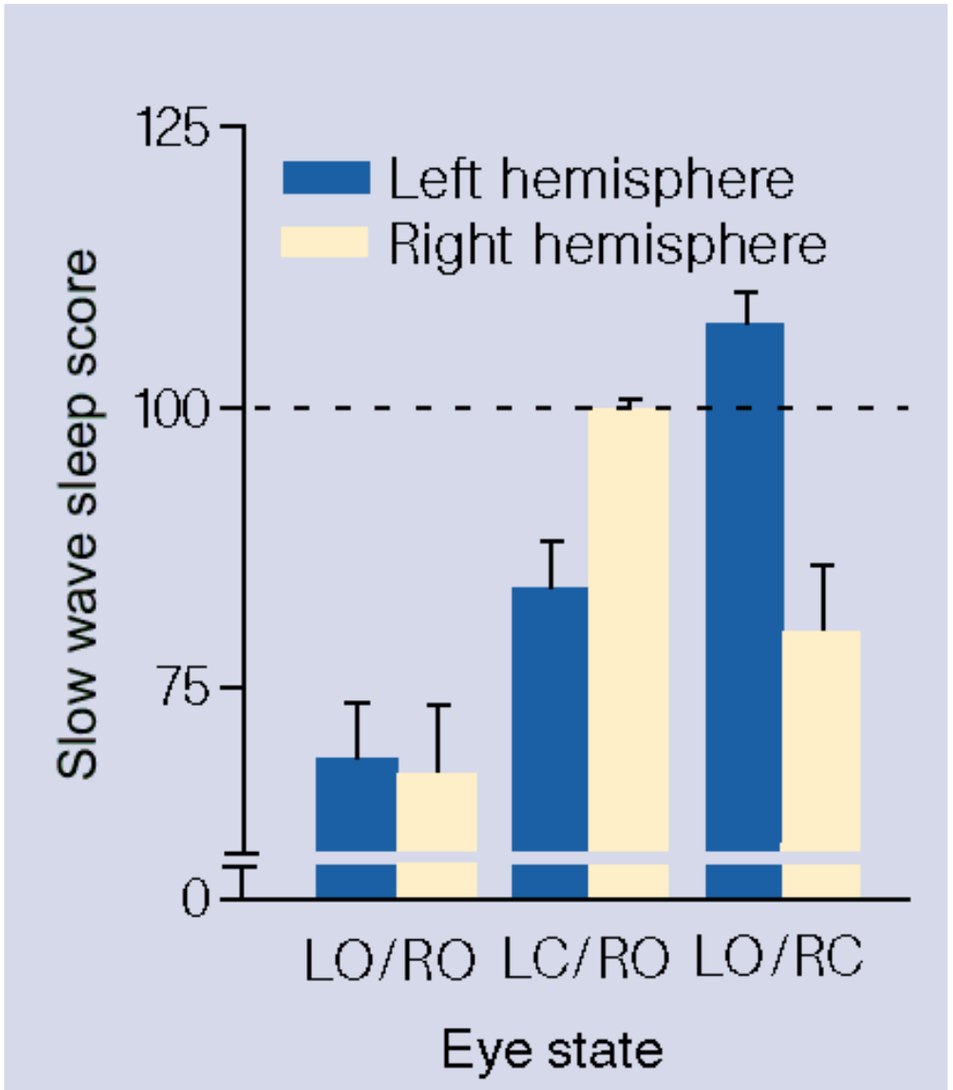
3. Represent magnitudes honestly

"A graphic does not distort if the visual representation of the data is consistent with the numerical representation" – Tufte (1983)

Are the bars "consistent with the numerical representation"?

Is 0 a reasonable baseline for evaluating sleep score?

Are there other issues with the graph?



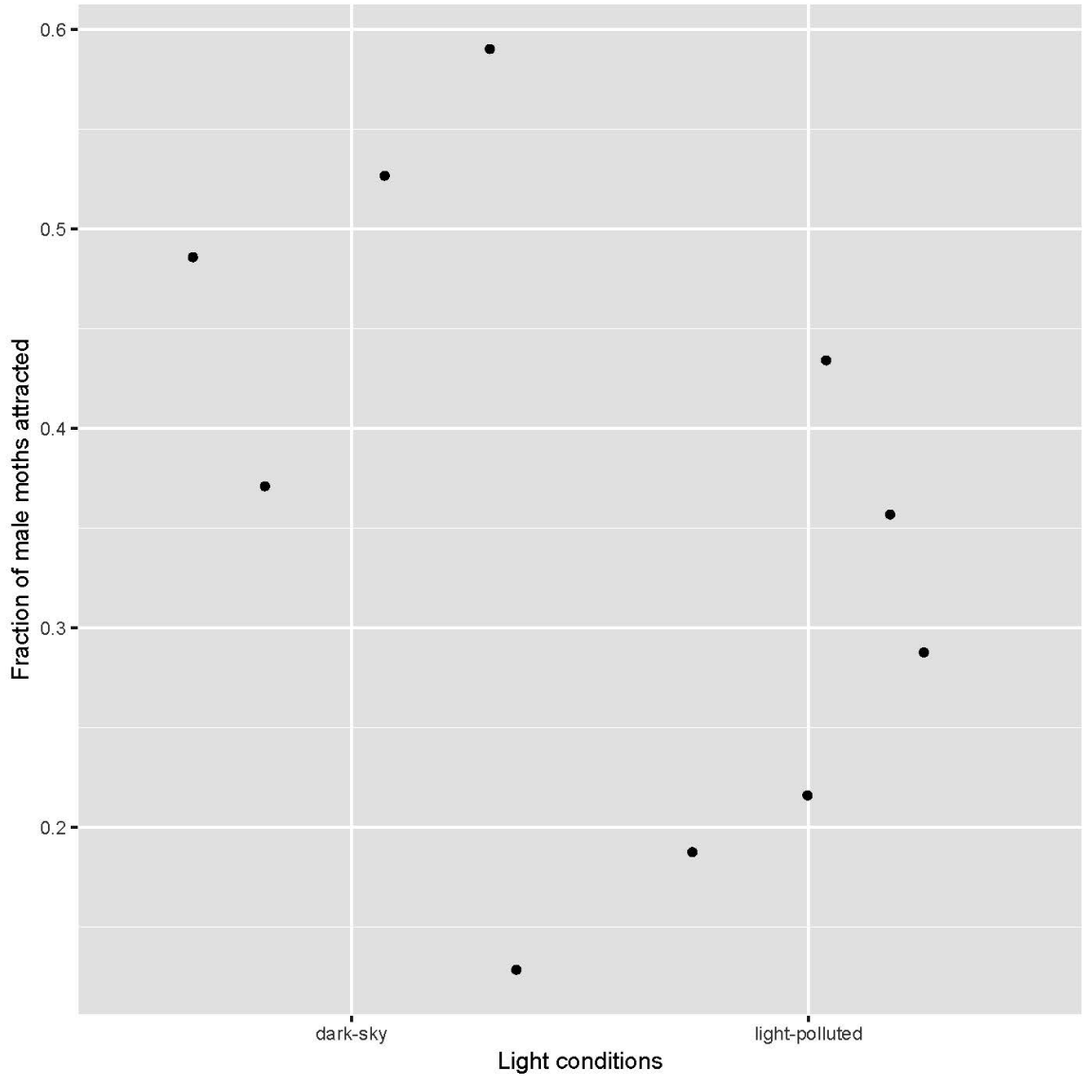
Slow wave sleep in the brain hemispheres of mallard ducks sleeping with one eye open. From Rattenborg et al. (1999) *Nature*.

4. Draw graphical elements clearly, minimizing clutter

"Maximize the data-ink ratio, within reason" – Tufte (1983)

Strip chart made with `ggplot()` using default graphical options.

Altermatt and Ebert (2016) measured light attraction of ermine moths (*Yponomeuta cagnagella*) from 10 different populations. Five of the populations were located in urban areas with plenty of human lights. The other 5 populations were located in pristine areas with no light pollution.



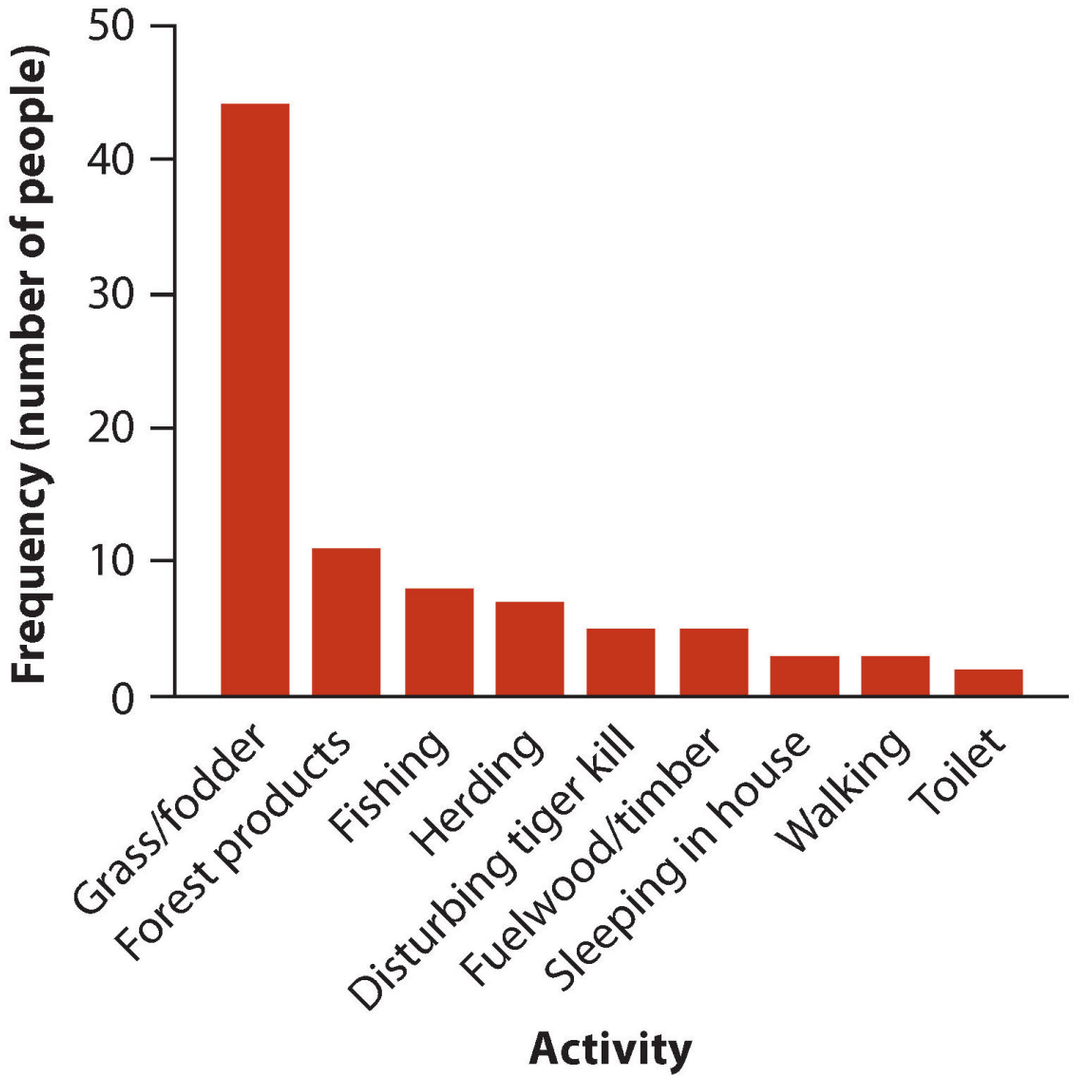
How to display frequencies for a categorical variable

Bar graph

Activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006.

Uses height of bars to display the frequency distribution of a categorical (grouping) variable

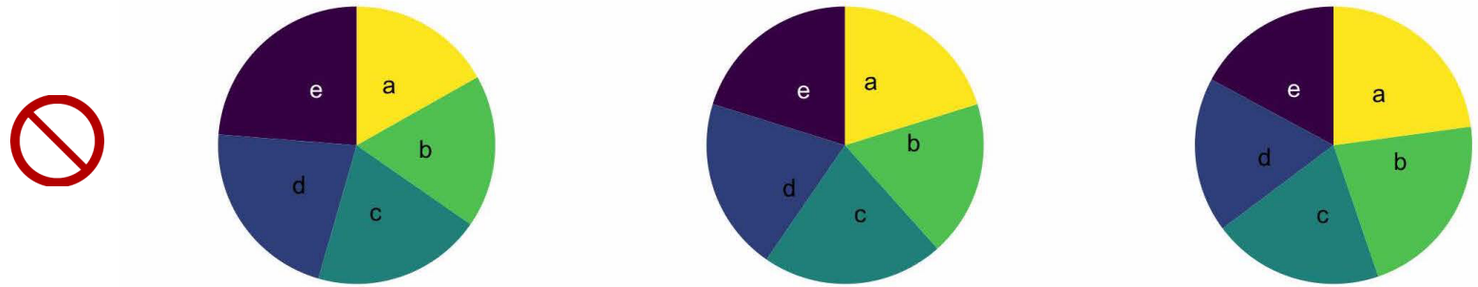
- Zero baseline
- Space between bars emphasize height
- Order of categories – most to least frequent is usually best

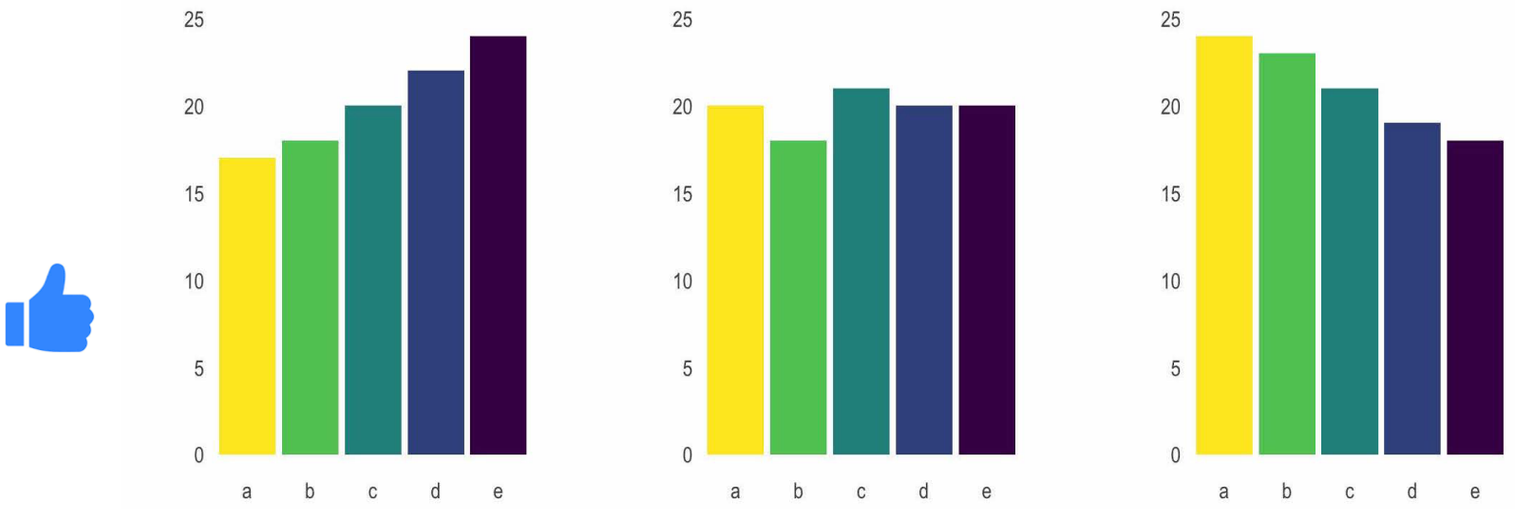


How to display frequencies for a categorical variable

Bar graph vs. Pie chart: which is more successful?

Humans are better at comparing relative areas in bar graphs than pie charts.





www.data-to-viz.com

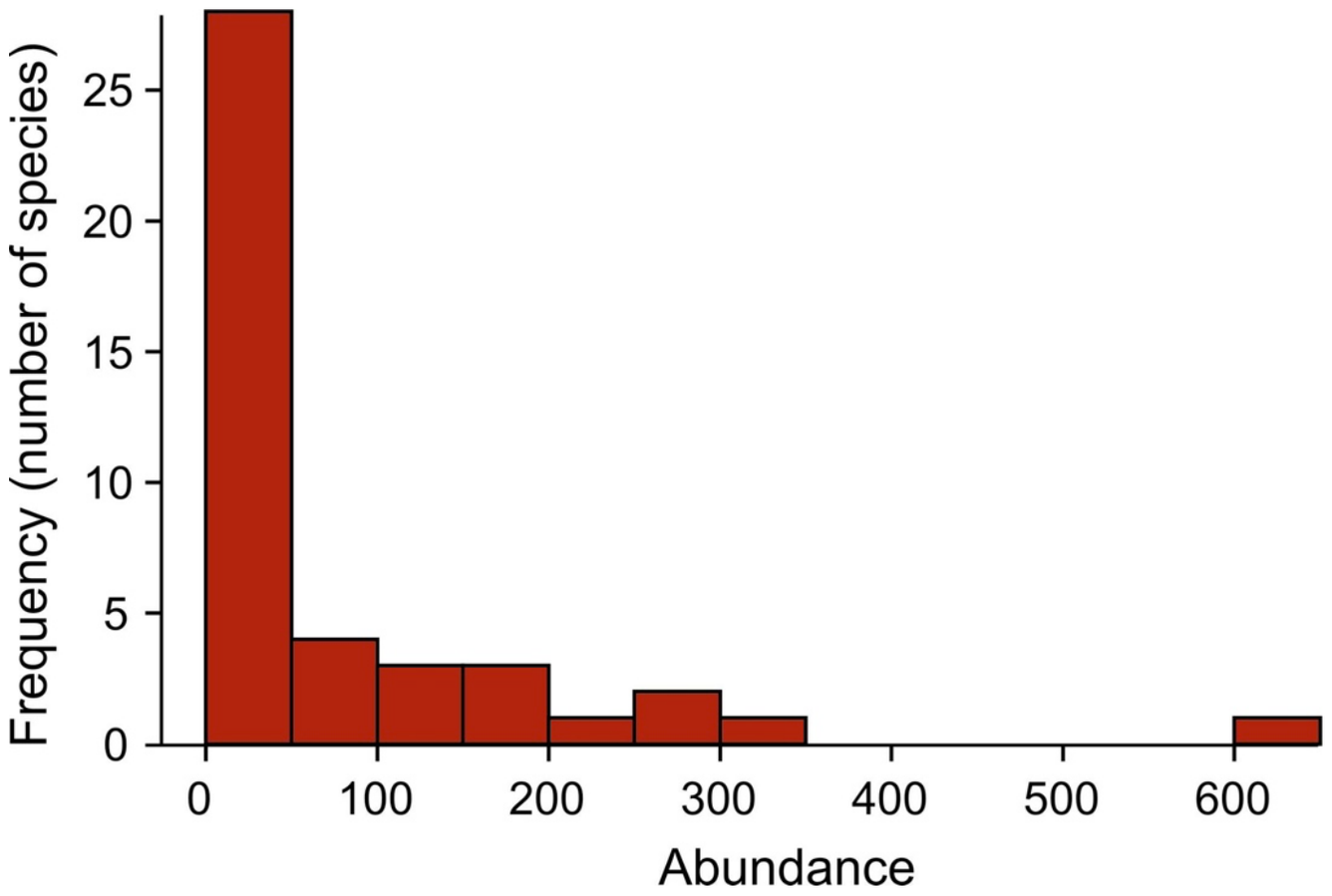
How to display frequency distribution for numeric variable

Histogram

Uses area of bars to display frequency distribution of a numerical variable

- Zero baseline
- No spaces between bars
- Choice of number of bins and bin width

The frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. *n* = 43 species

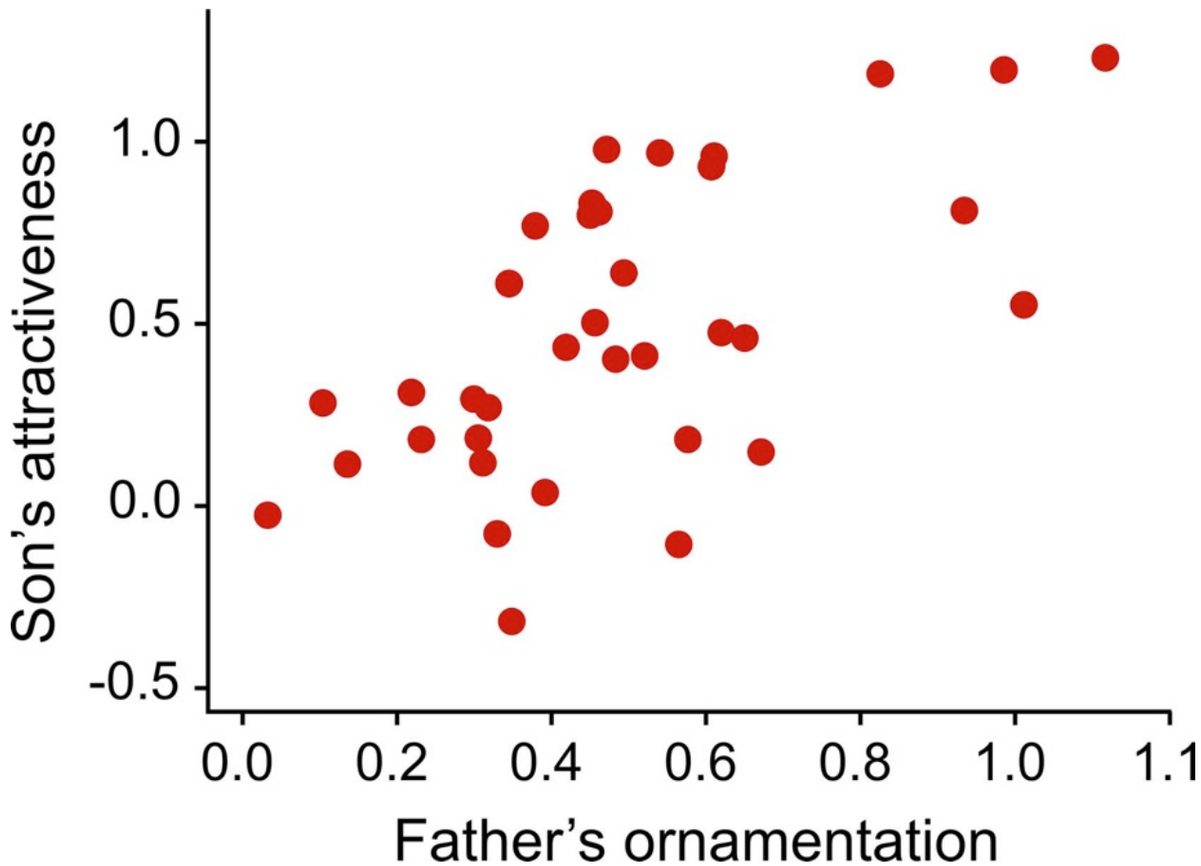


How to display association between two numerical variables

Scatter plot

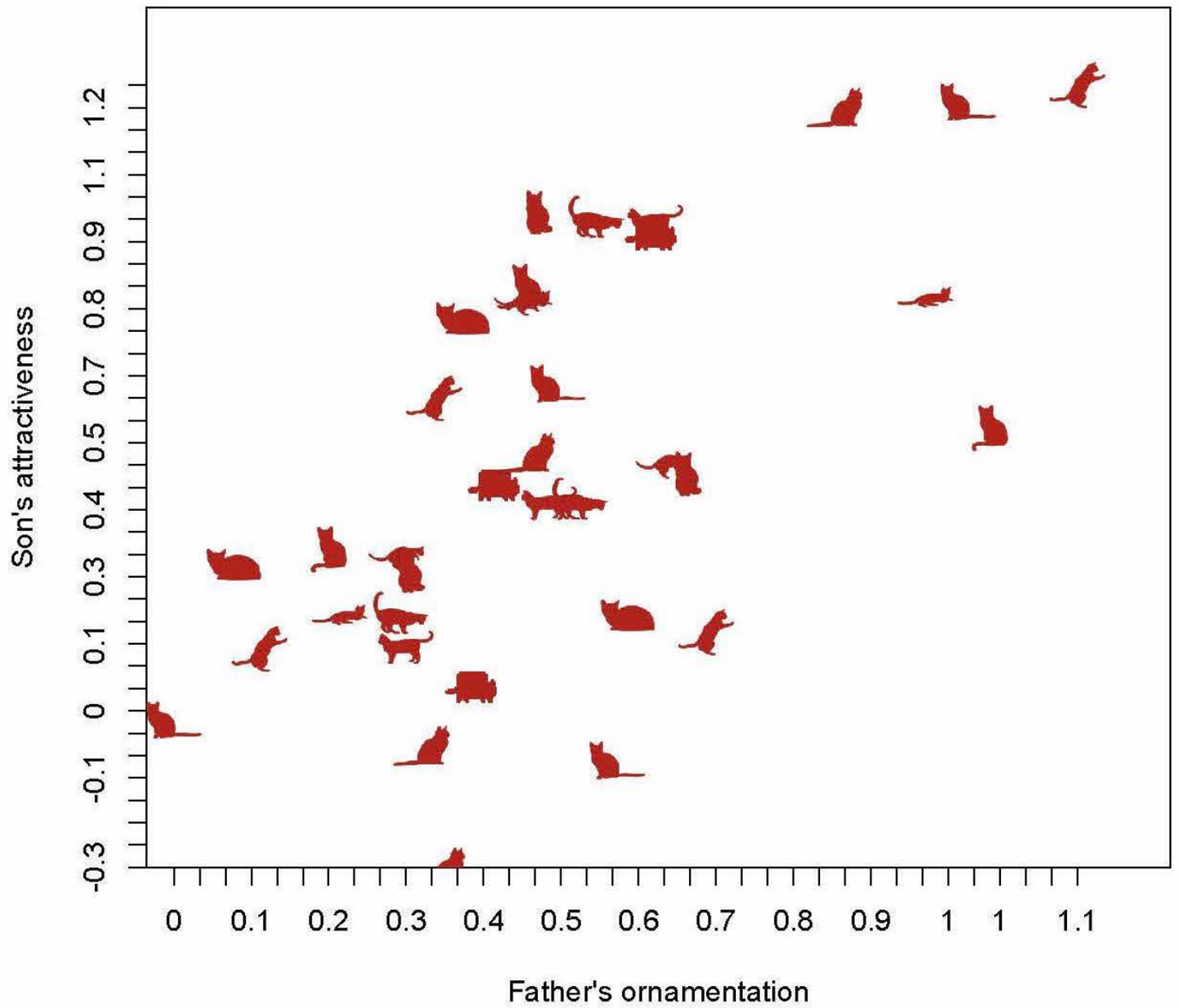
- Non-zero baseline often ok (goal is to show association, not height above 0)
- Points fill the space available

The relationship between the ornamentation of male guppies and the average attractiveness of their sons. *n* = 36 families.



Catterplot

Because we can



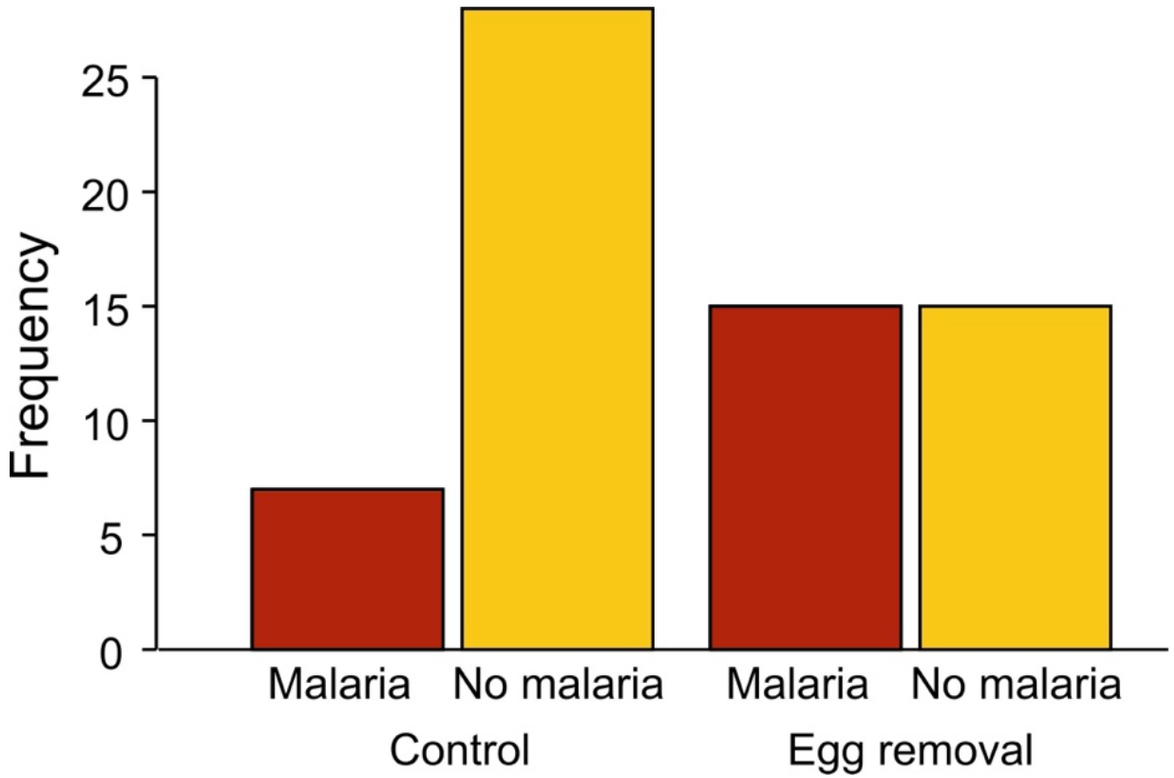
How to display association between categorical variables

Grouped bar graph

Uses height of bars to display association between two (or more) categorical variables.

- Explanatory variable = outer groups; response variable = inner groups
- Zero baseline (so that height is proportional to frequency)
- Spacing between bars wider between outer groups

Incidence of malaria in female great tits in relation to experimental treatment.
n = 65 birds.



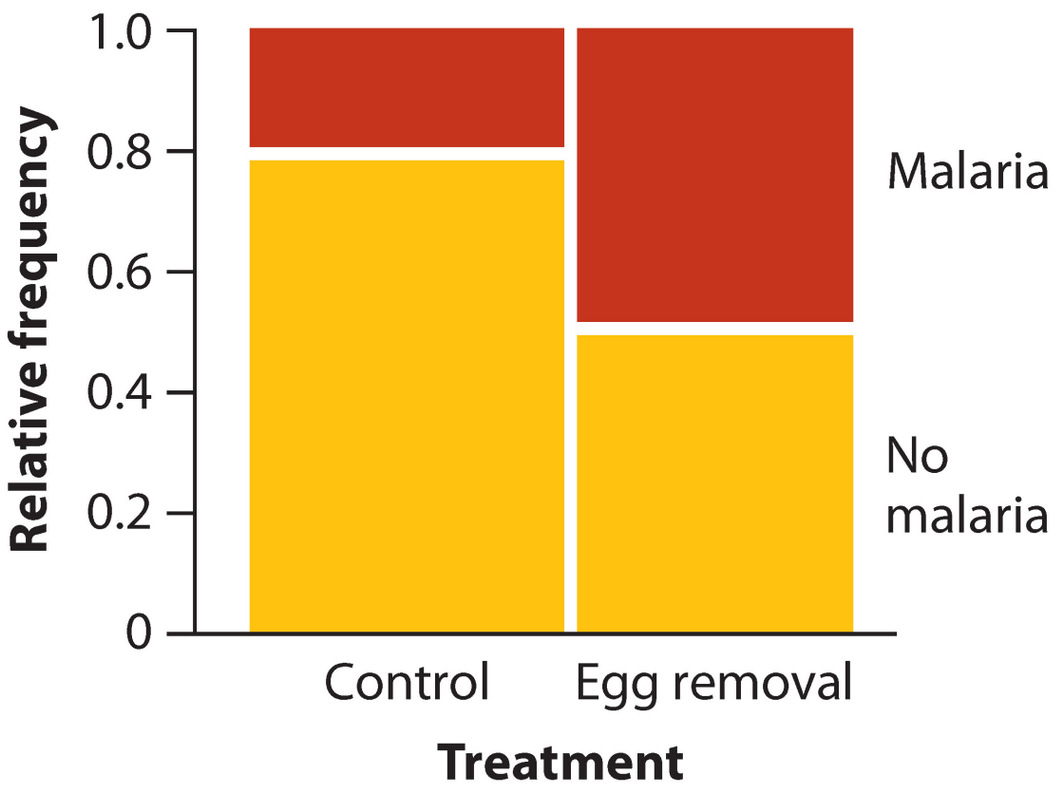
How to display association between categorical variables

Mosaic plot

Uses area of rectangles to display association between two (or more) categorical variables

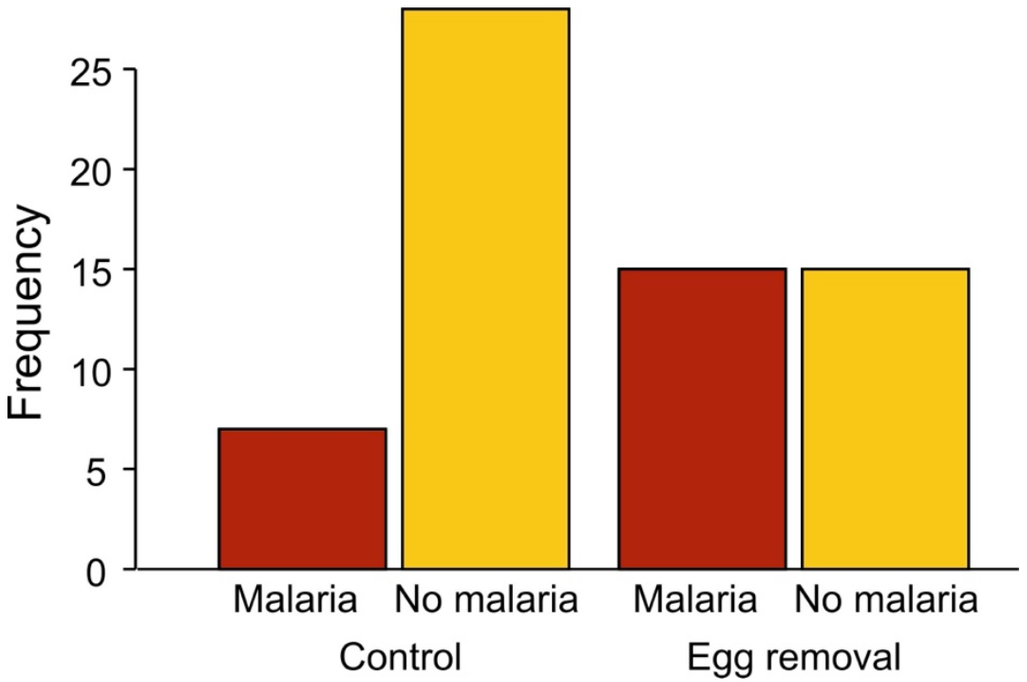
- Explanatory variable along horizontal axis; response variable stacked
- Area proportional to frequency
- Like a graphical representation of a contingency table

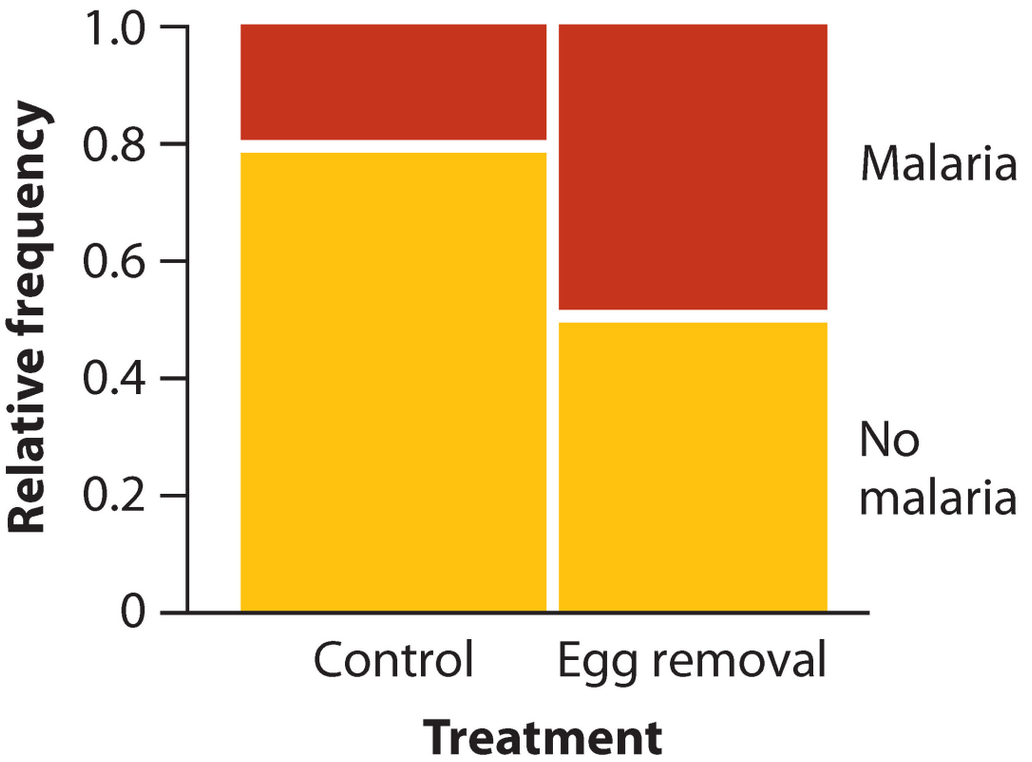
Incidence of malaria in female great tits in relation to experimental treatment.
n = 65 birds.



Grouped bar graph vs mosaic plot

Q: Which is more successful?





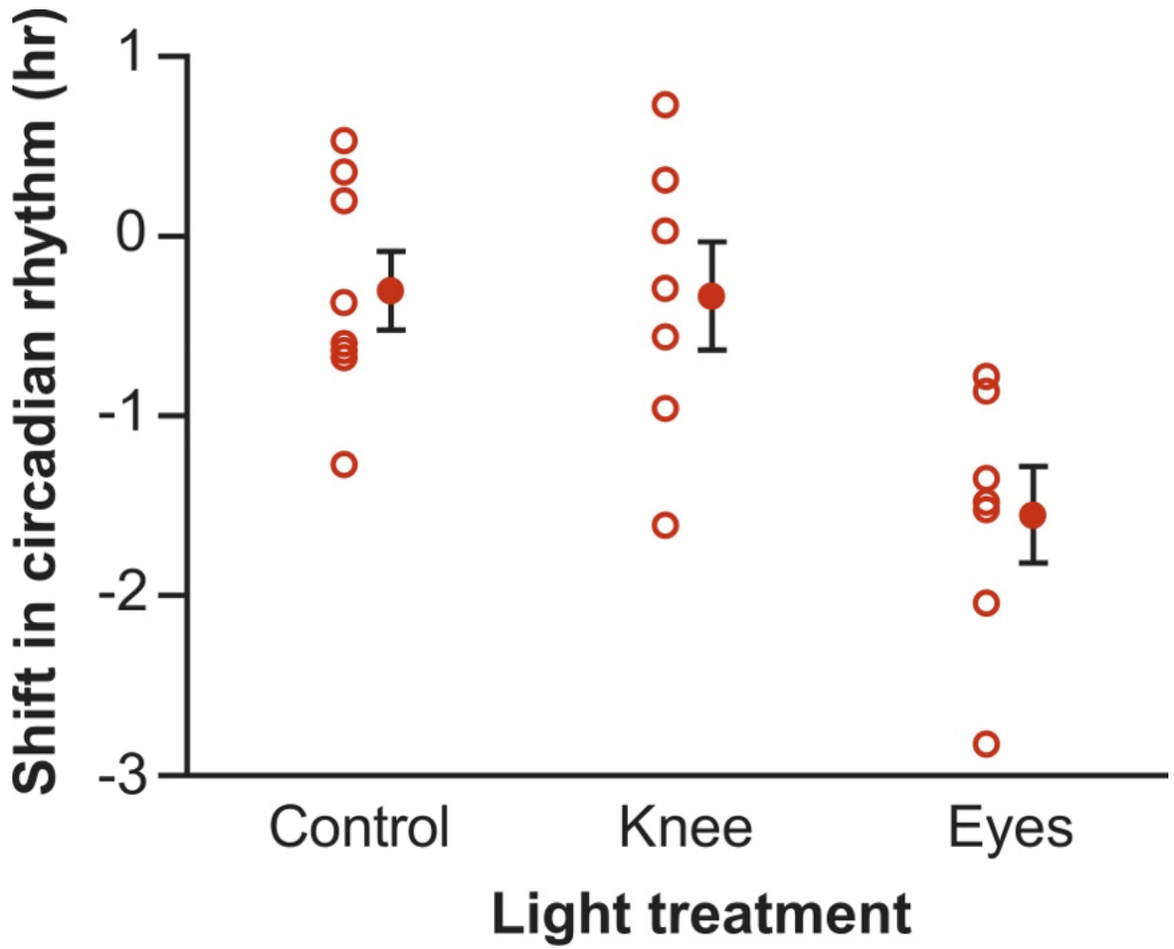
How to display association between numerical and categorical variable

Strip chart

Displays differences between groups

- Shows the data points
- Non-zero baseline often ok (goal is association not magnitude or frequency)
- Points fill the space available

Phase shift in the circadian rhythm of melatonin production in 22 subjects given alternative light treatments (open circles). Group means ± 1 SE also shown.



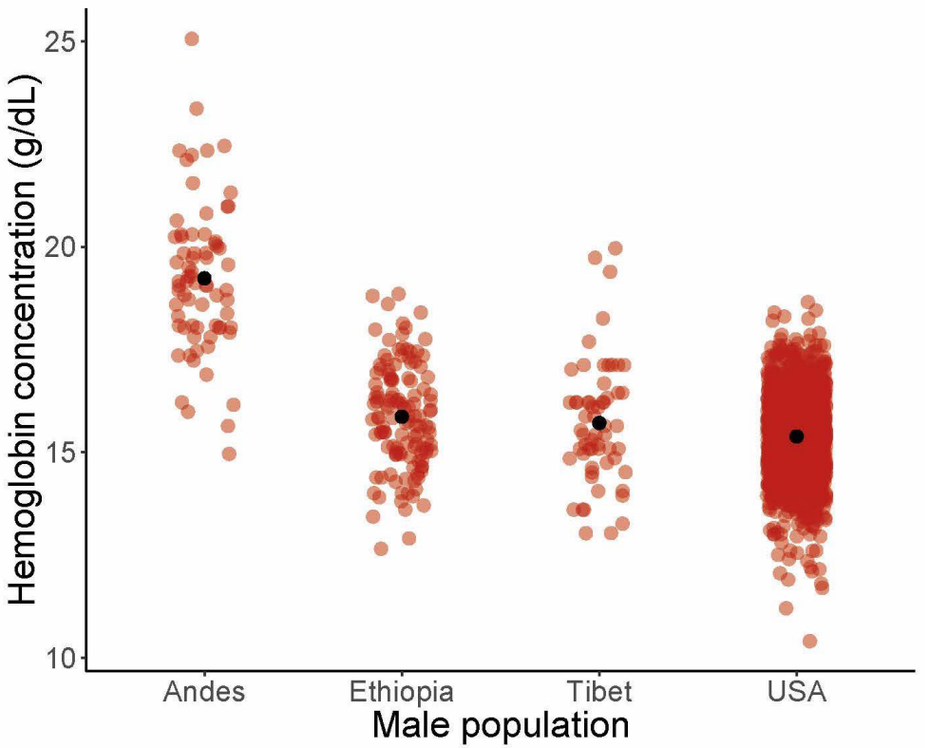
How to display association between numerical and categorical variable

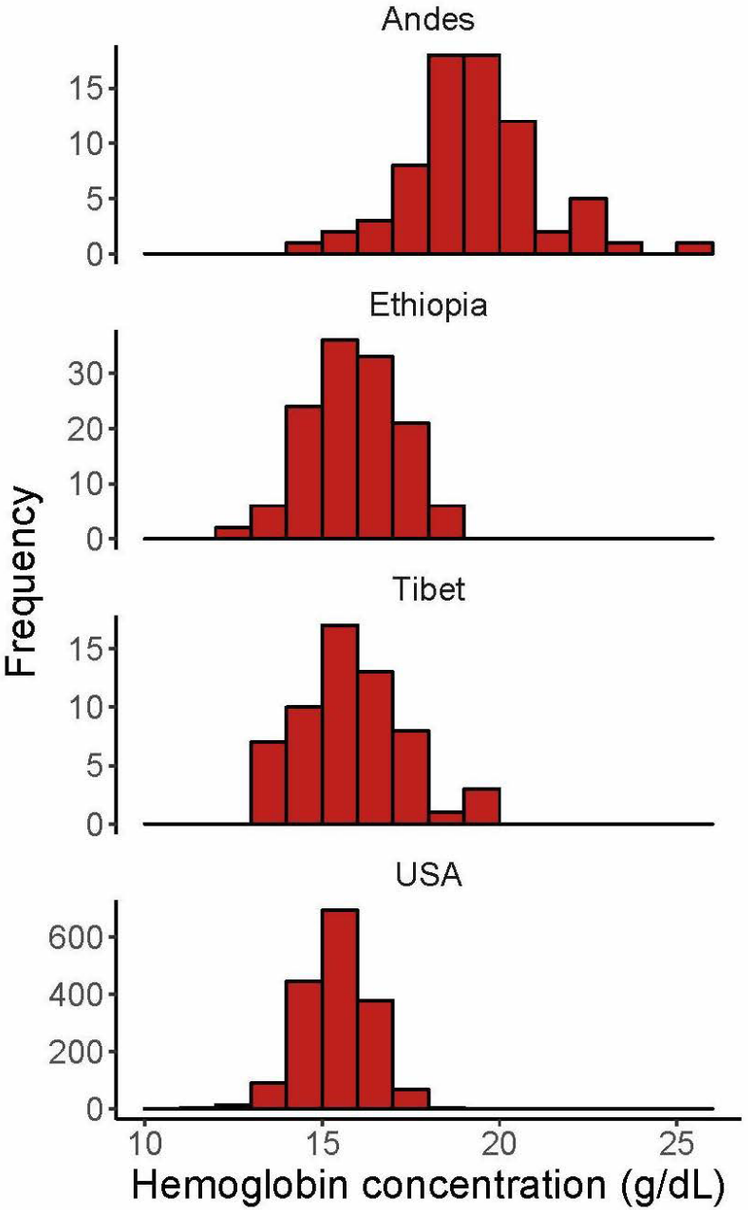
<u>Strip chart vs multiple histograms</u>

Too many data points for a strip chart.

Stack histograms vertically to best compare

Hemoglobin concentration in blood of males living at high elevation compared to sea-level USA control.



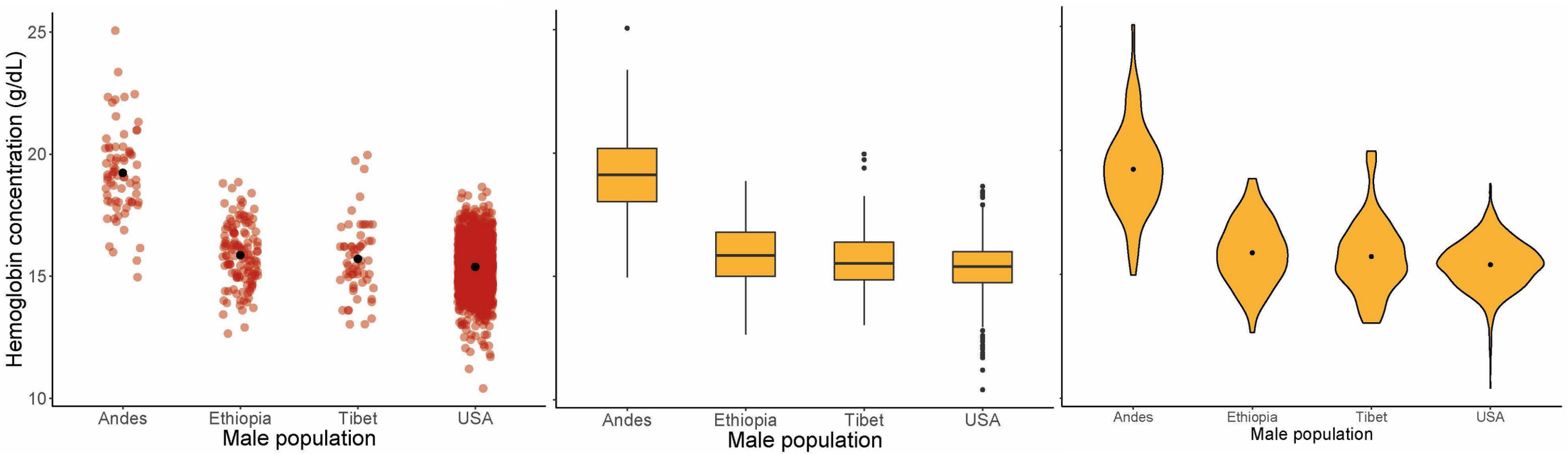


How to display association between numerical and categorical variable

Strip chart vs box plot vs violin plot

- Box plot displays median, first and third quartile, range, and extreme observations
- Violin plot estimates probability density of each group using “kernel smoothing”
- Non-zero baseline often ok (goal is to show differences not amounts)

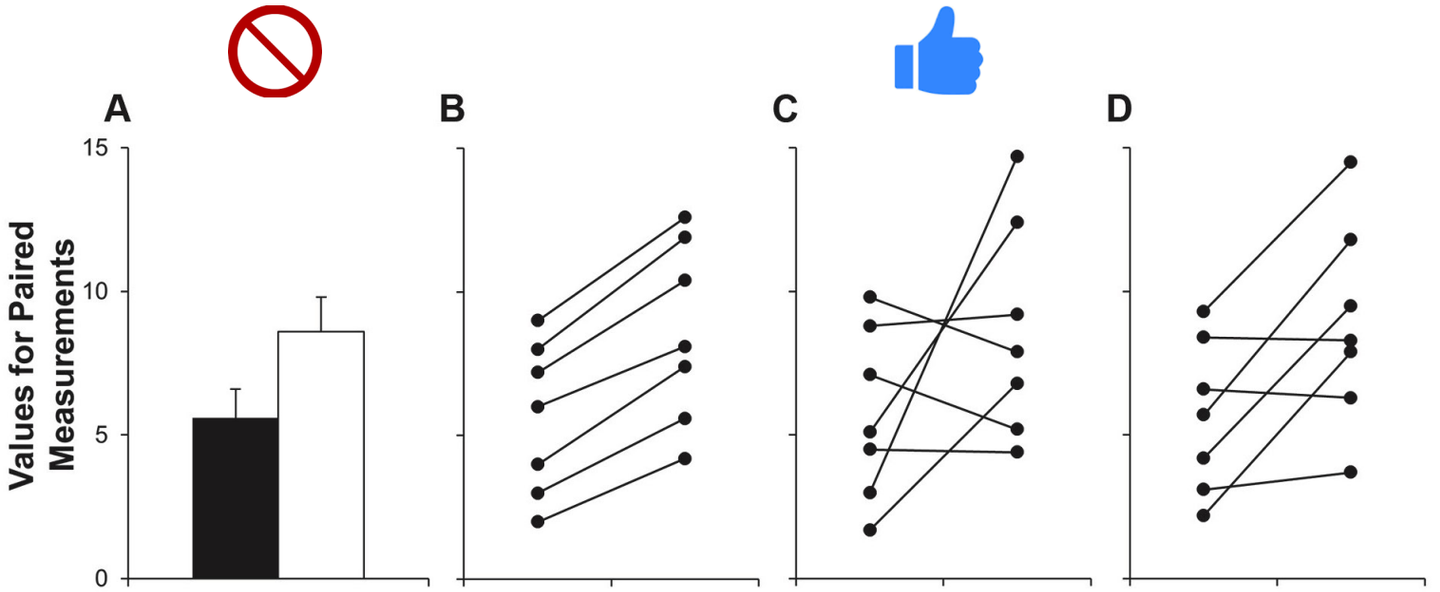
Q: which is more successful?



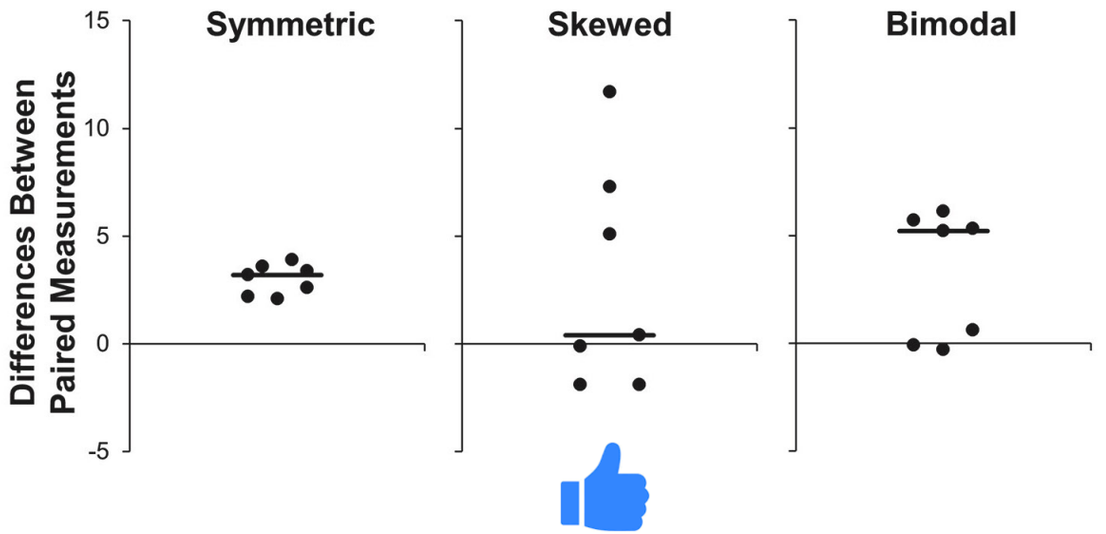
How to show repeated measures data (e.g., paired data)

Connect the dots to show pairing; or plot the differences.

Interaction plots



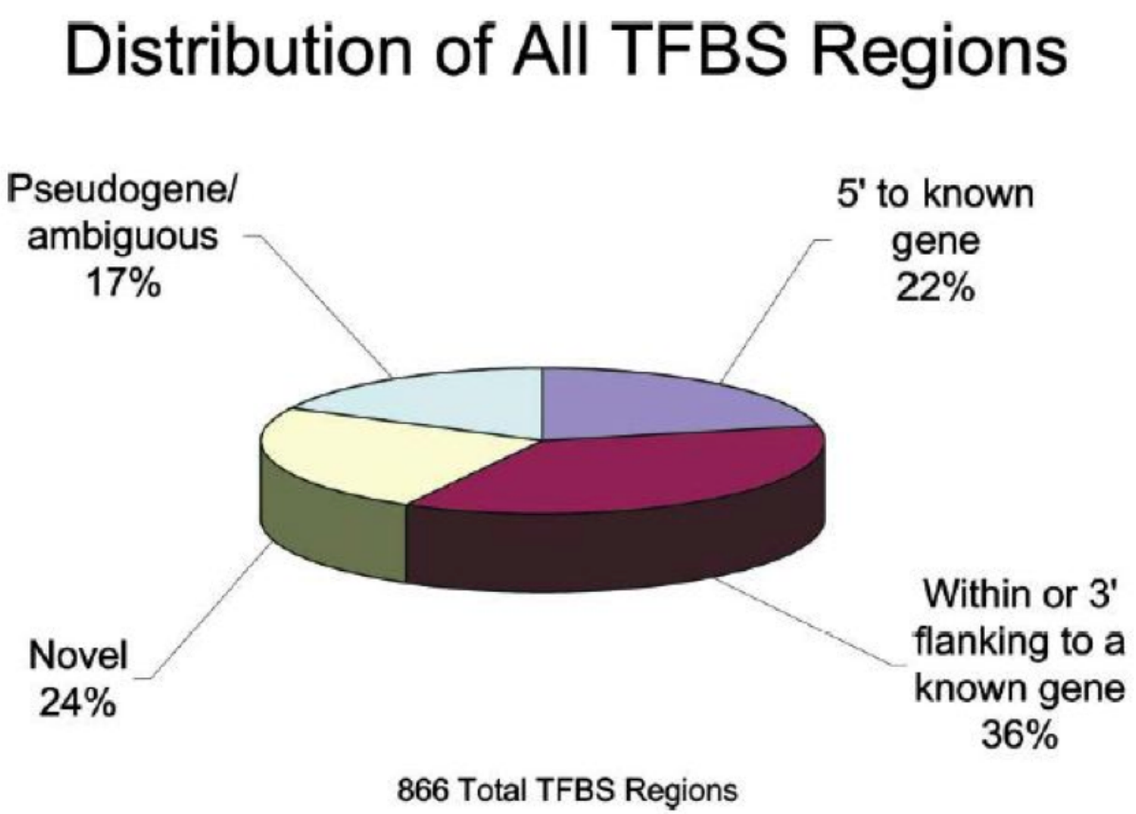
Strip charts of differences



Weissgerber et al. (2015) beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol*. DOI:10.1371/journal.pbio.1002128

Please don’t use 3D graphs to display frequencies

- 3D rendering is gratuitous
- It makes it more difficult to see area, and hence the frequencies
- A graph that is meaningful only with numbers added must be recognized as a failure.



Cawley S, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116:499-509, Figure 1

2D pies can also bewilder

The main patterns of interest were the change in NO_3 and SO_4 between winter and summer, and the consistency of change between geographic regions. This is not easy to see
("Huh?" not "Oh!")

Design a graph to show the change from summer to winter in NO_3 and SO_4, rather than try to display everything.

Bell ML, et al. (2007) Spatial and temporal variation in $PM_{2.5}$ chemical composition in the United States for health effects studies. Environmental Health Perspectives 115:989-995

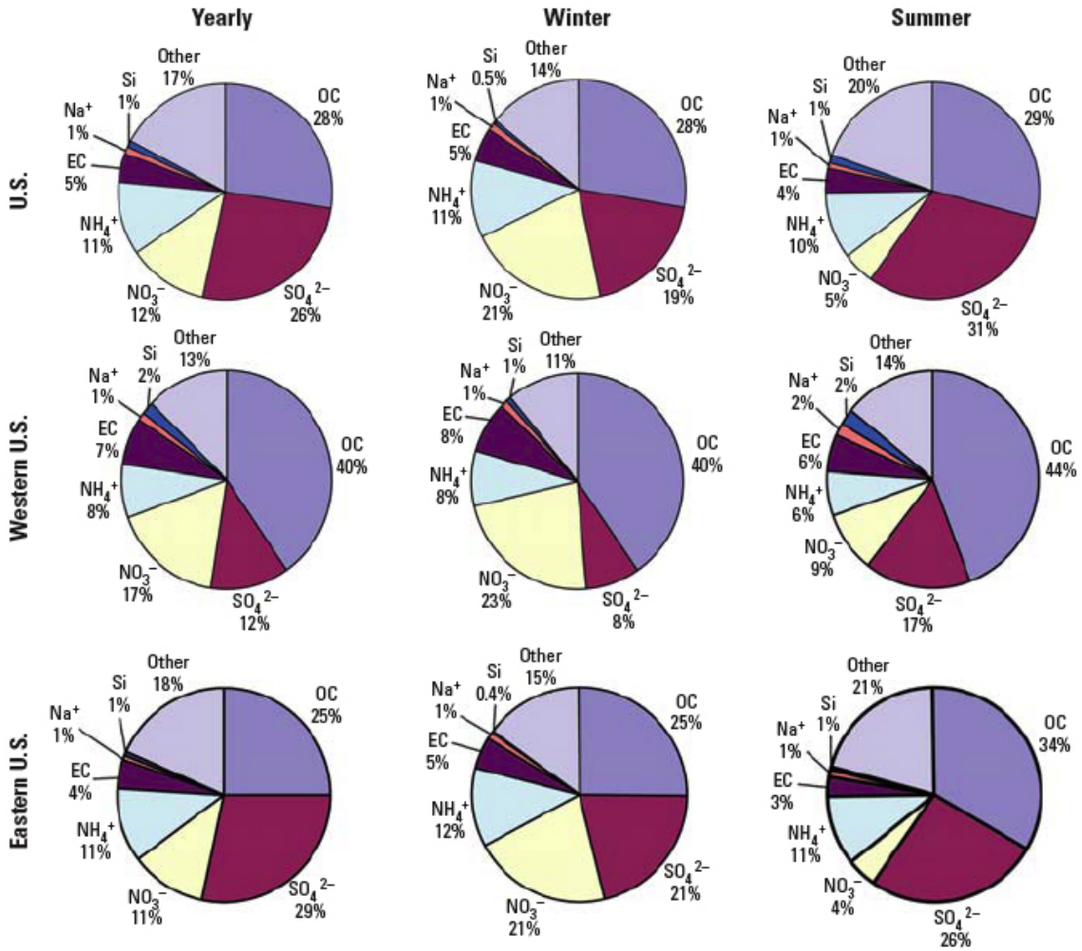
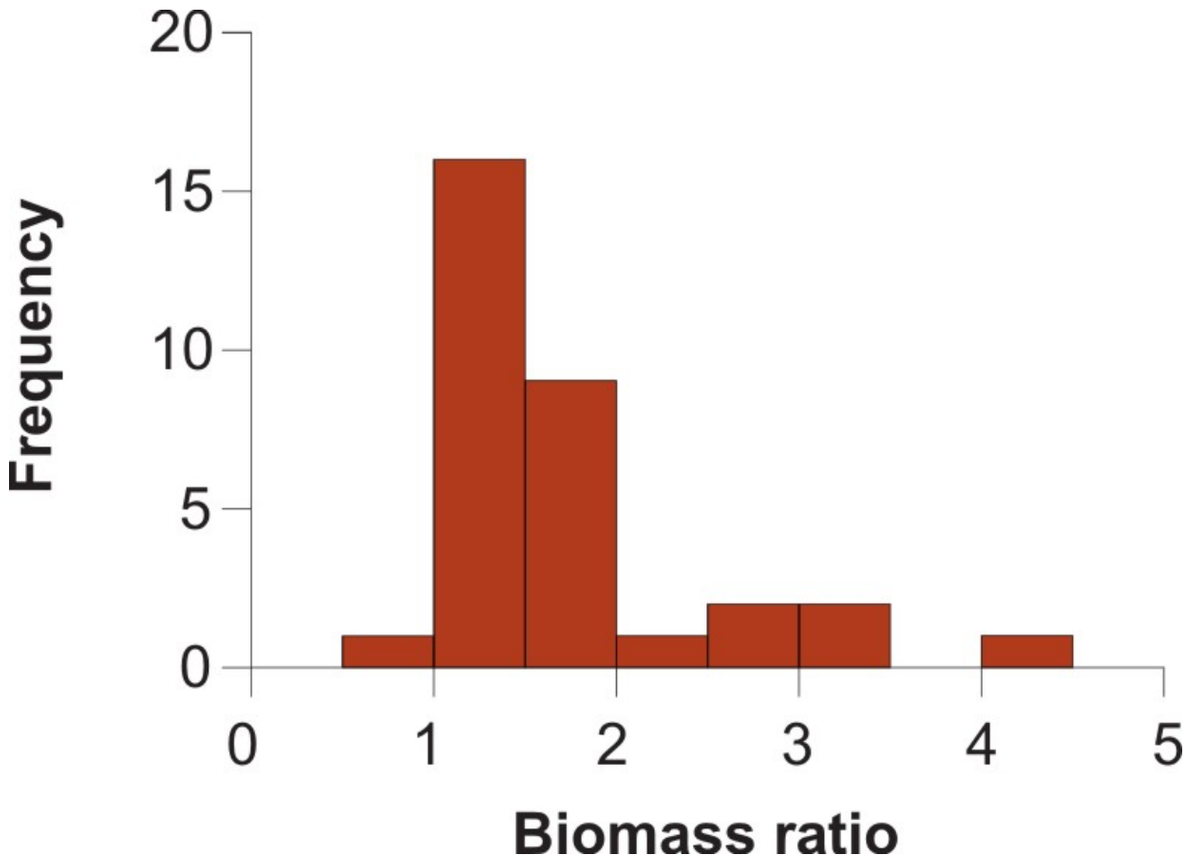


Figure 3. Percent of $PM_{2.5}$ composition by component for yearly, winter, and summer averages, by region.

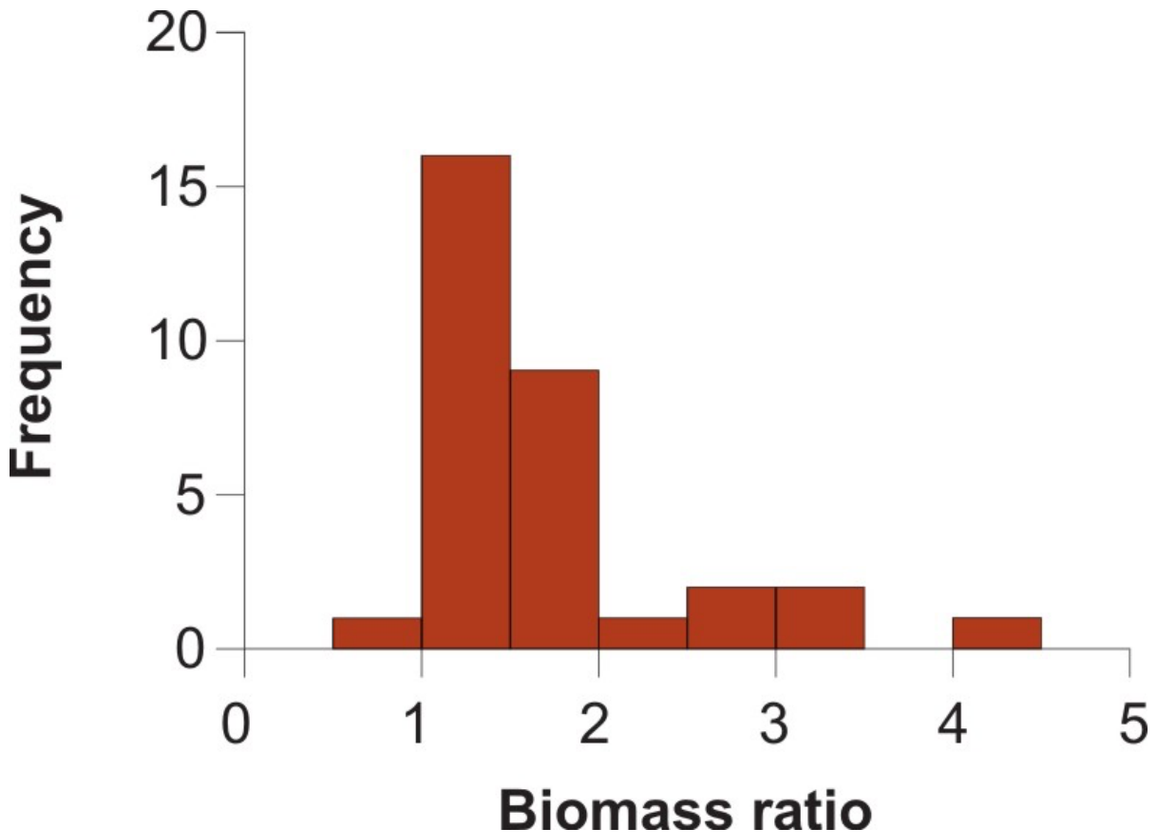
Handling ratio data



Biomass ratio is the total mass of all marine plants and animals per unit area of marine reserve divided by the same quantity in the unprotected control. N = 32 pairs (reserve and control). Data from Halpern (2003)

Whitlock, M. C. and D. Schluter. 2020. The analysis of biological data. 3rd Ed. Macmillan.

Handling ratio data



Problems:
Ratios less than 1 are sandwiched between 0 and 1, distorting magnitudes

Possible solutions:
Use log of ratio so that ratios above and below 1 have the same scale.

What about tables?

- Like graphs, tables are used to compare measurements between groups and expose relationships between variables.
- For some kinds of data, they may be the best way to communicate results to a wider audience.
- Use tables to illuminate patterns.

Make your tables so that they cause the viewer to go “Oh!” and not “Huh?”

Put tables for storing numbers into online Appendix or Supplement

Improving tables - example

Can you see a pattern?

Difficult to see a relationship between *F* and survival.

Uneven line spacing, the gaps break up patterns.

Too much empty space.

Too many decimals.

Table 2.5-1 Inbreeding coefficient (*F*) of Spanish Habsburg kings and queens and survival of their progeny.

King/Queen	*F*	Pregnancies	Miscarriages & stillbirths	Neonatal deaths	Later deaths	Survivors to age 10	Survival (total)	Survival (postnatal)
Ferdinand of Aragon								
Elizabeth of Castile	0.039	7	2	0	0	5	0.714	1.000
Philip I								
Joanna I	0.037	6	0	0	0	6	1.000	1.000
Charles I								
Isabella of Portugal	0.123	7	1	1	2	3	0.429	0.600
Philip II								
Elizabeth of Valois	0.008	4	1	1	0	2	0.500	1.000
Anna of Austria	0.218	6	1	0	4	1	0.167	0.200
Philip III								
Margaret of Austria	0.115	8	0	0	3	5	0.625	0.625
Philip IV								
Elizabeth of Bourbon	0.050	7	0	3	2	2	0.286	0.500
Mariana of Austria	0.254	6	0	1	3	2	0.333	0.400

Source: Data are from Alvarez et al. (2009).

Improving tables – example

Use vertical stacking of numbers you most want the eye/brain to compare; no gaps.

Put columns adjacent that you want to show associations between. Sort one of the columns.

Table 2.5-2 Inbreeding coefficient (*F*) of Spanish kings and queens and survival of their progeny. These data are extracted and reorganized from Table 2.5-1.

King/Queen	*F*	Survival (postnatal)	Survival (total)	Number of pregnancies
Philip II/Elizabeth of Valois	0.01	1.00	0.50	4
Philip I/Joanna I	0.04	1.00	1.00	6
Ferdinand/Elizabeth of Castile	0.04	1.00	0.71	7
Philip IV/Elizabeth of Bourbon	0.05	0.50	0.29	7
Philip III/Margaret of Austria	0.12	0.63	0.63	8
Charles I/Isabella of Portugal	0.12	0.60	0.43	7
Philip II/Anna of Austria	0.22	0.20	0.17	6
Philip IV/Mariana of Austria	0.25	0.40	0.33	6

Interactive plots

Graphs that allow the viewer to manipulate the image.

Enables user to probe for more information, but user interaction with plot shouldn't be required to recognize patterns in the data. I.e., hover etc doesn't further the main goal.

Line plot with basic "hover"
https://rstudio.github.io/dygraphs/

Interactive scatter plot:
https://cengel.github.io/R-data-viz/interactive-graphs.html#plotly

Data that moves

Video graphs for displaying data are becoming more common
http://www.r-graph-gallery.com/3-r-animated-cube/
https://science.ebird.org/en/status-and-trends/abundance-animations

Measles outbreak simulation
https://www.theguardian.com/society/ng-interactive/2015/feb/05/-sp-watch-how-measles-outbreak-spreads-when-kids-get-vaccinated

The Atlantic Slave Trade in Two Minutes
http://www.slate.com/articles/life/the_history_of_american_slavery/2015/06/animated_interactive_of_the_history_of_the_atlantic_slave_trade.html?wpsrc=sh_all_dt_fb_top

The fallen of WWII: https://vimeo.com/128373915

Discussion paper:

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187–211

Download from "**Handouts**" tab on course web site.

Presenters for next Tuesday, Jan 21: _____ & _____

Discussion moderators: _____ & _____

Homework assignment 1.

Due Friday, Feb 7, at 5 pm.

See instructions on "**Homework**" page on course web site

Basic idea:

Find a poor graph drawn from data and published by your thesis supervisor.

Analyze it: why does it not succeed?

Improve the graph using R.

Analyze it: why does it succeed?

Include R code.