Graphics

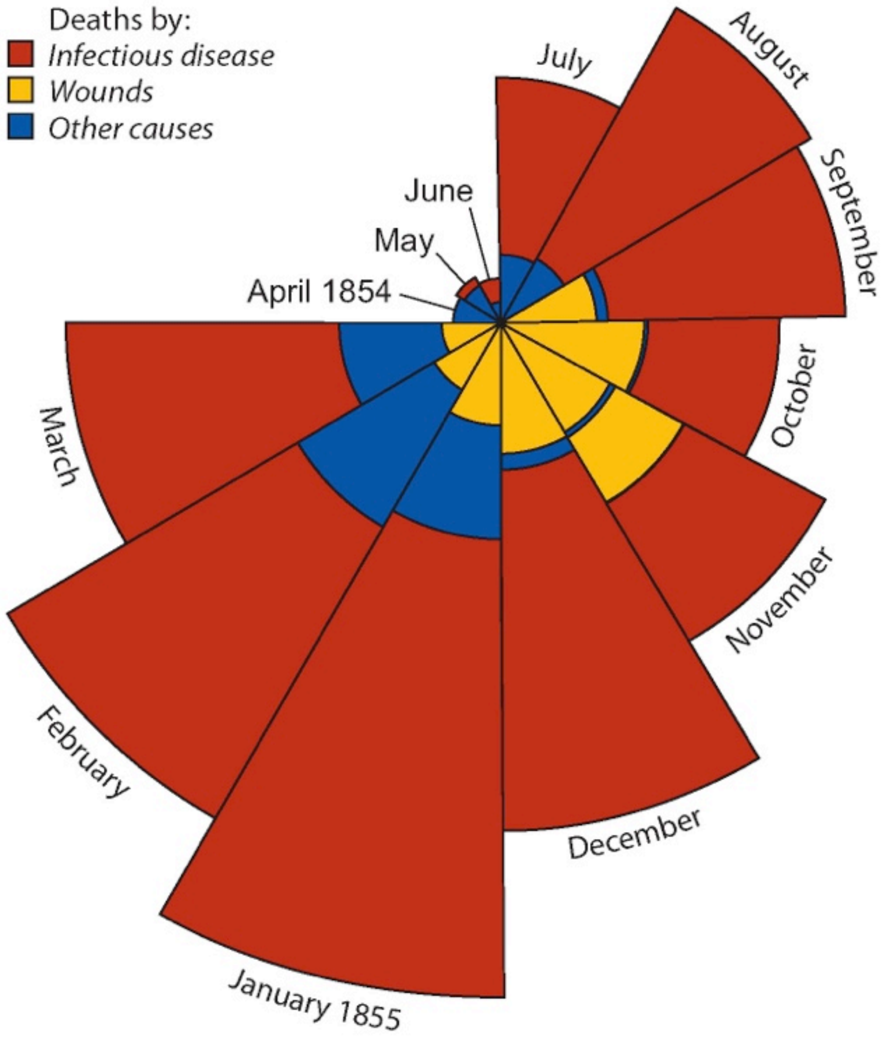
Outline for today

- The purpose of graphs
- Principles of effective display
- Types of graphs to achieve these principles
- How some graphs fail, and what can be done
- What about tables?

The purpose of graphs

- The human eye is a natural pattern detector, adept at spotting trends and exceptions.
- Graphs enable visual comparisons of measurements between groups and expose relationships between variables.
- They are the best method available for discovering patterns in your data.

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

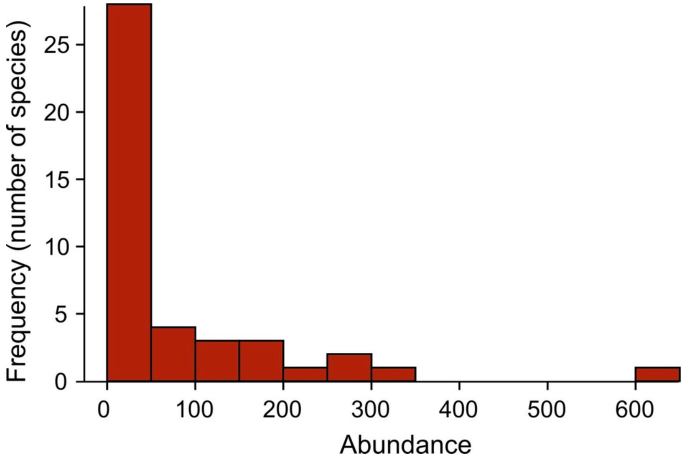


The purpose of graphs

- Graphs are the best method for communicating results to a wider audience

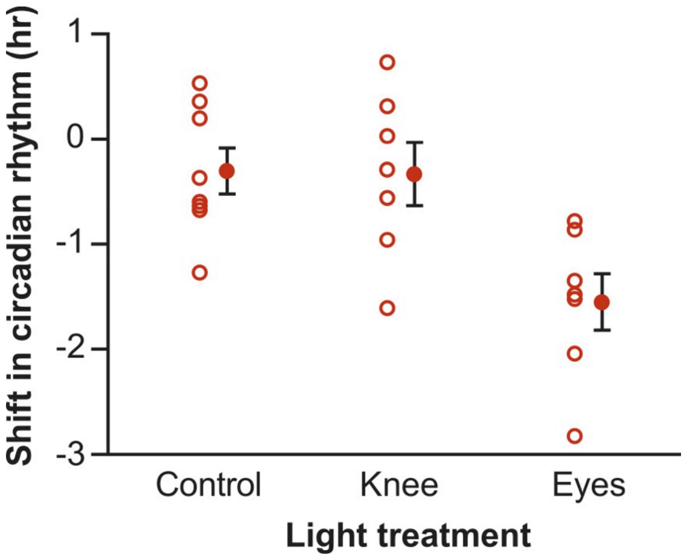
1. Frequency distributions

- The location, spread, shape of distribution



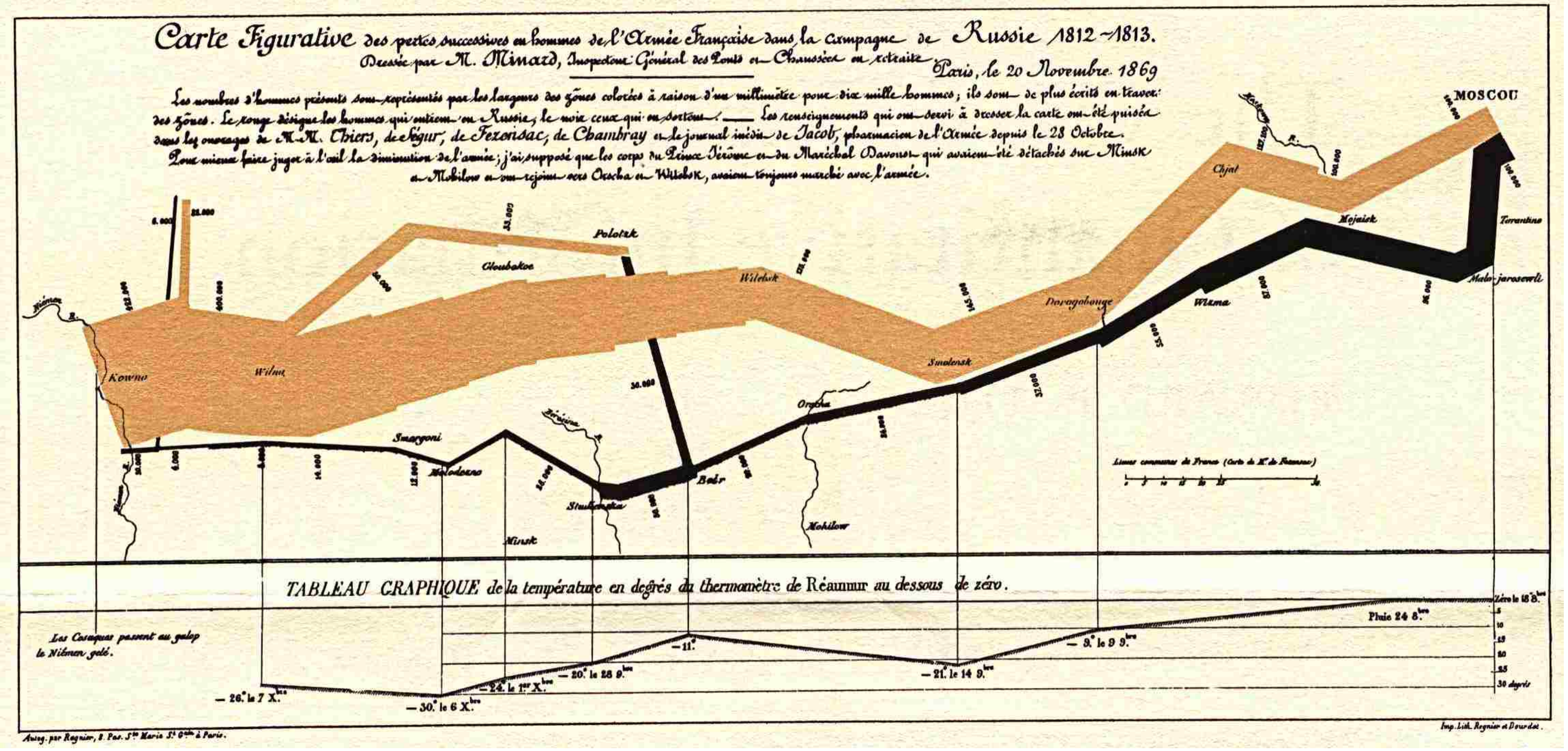
2. Associations between variables

- The relationship between two or more variables
- Differences between groups



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.


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wizma
Kowno
Wilna
Smorgoni
Molodezno
Minsk
Studienska
Bobr
Orscha
Mohilow
Niemen
6.000
22.000
422.000
400.000
33.000
60.000
30.000
175.000
145.000
100.000
127.000
96.000
87.000
55.000
37.000
24.000
28.000
50.000
20.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Zéro le 18 8bre
Pluie 24 8bre
– 9° le 9 9bre
– 21° le 14 9bre
– 11°
– 20° le 28 9bre
– 24° le 1er Xbre
– 30° le 6 Xbre
– 26° le 7 Xbre
30 degrés
Autog. par Regnier, 8 Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Principles of effective display

"Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" – Tufte (1983)

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

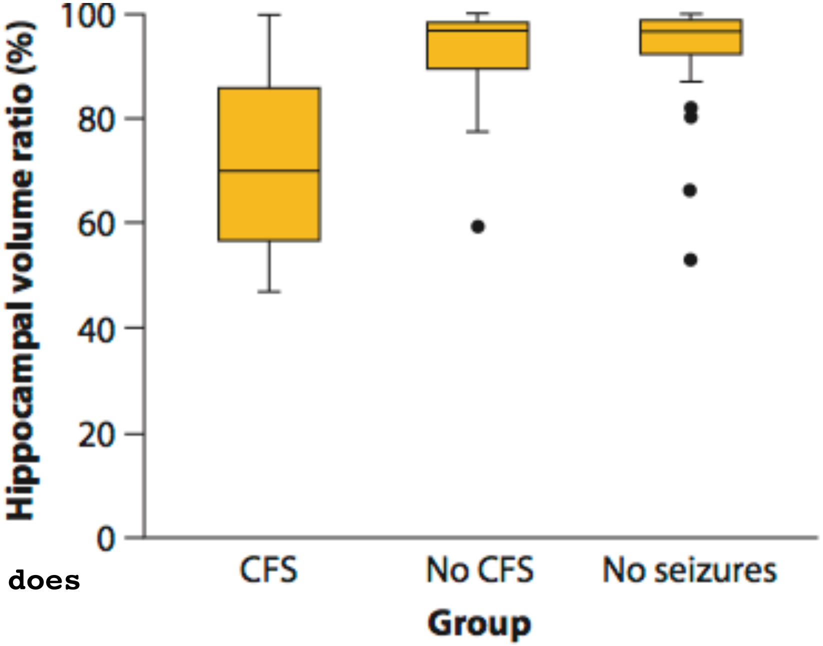
The following principles will help to increase the effectiveness of your graphs:

- Show the data
- Make patterns in the data easy to see
- Represent magnitudes honestly
- Draw graphical elements clearly, minimizing clutter

1. Show the data

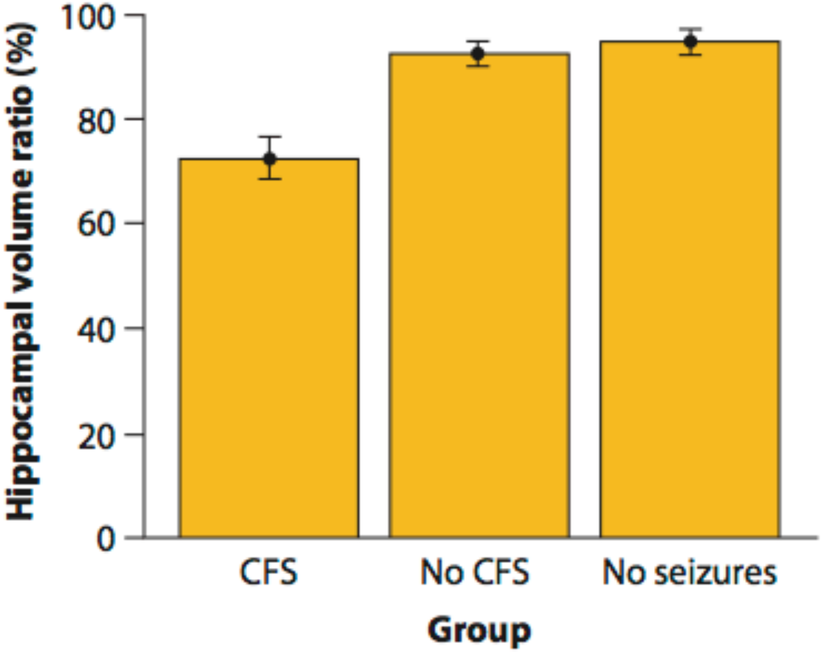
“Above all else show the data” – Tufte (1983)

Which graph is more effective? Why?



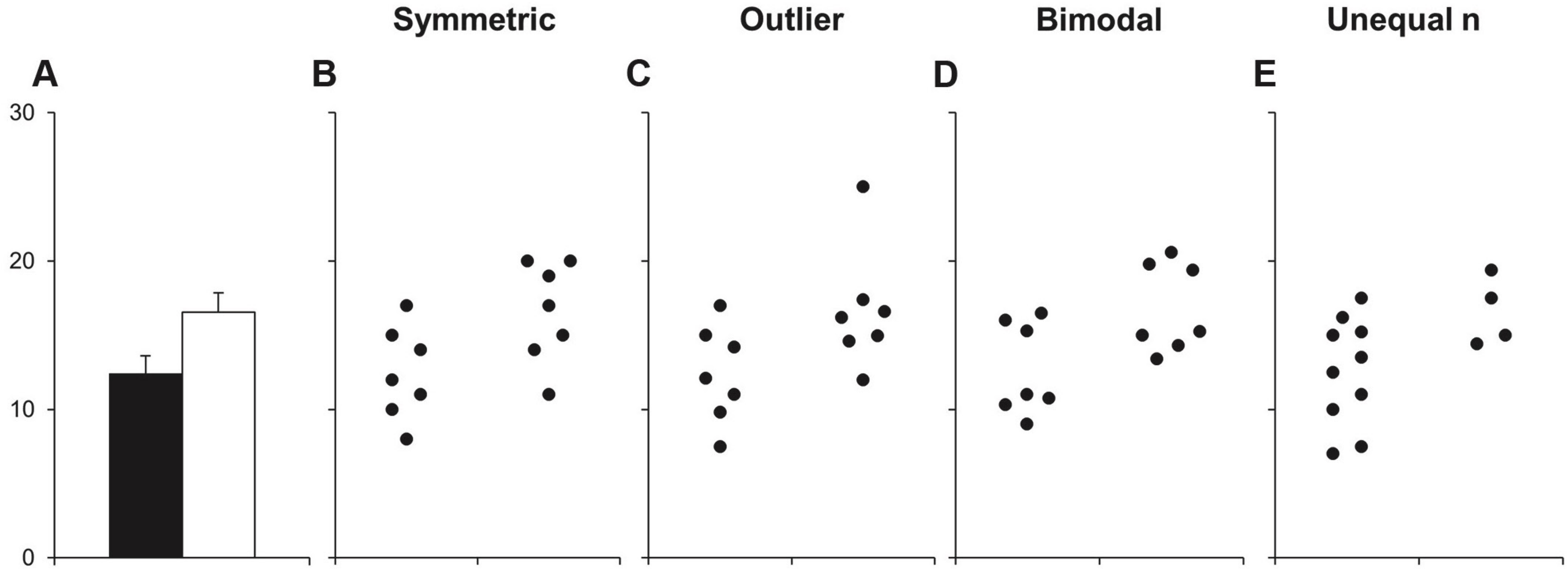
`box plot shows more about the distribution than the bar graph does`

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.



Why show the data?

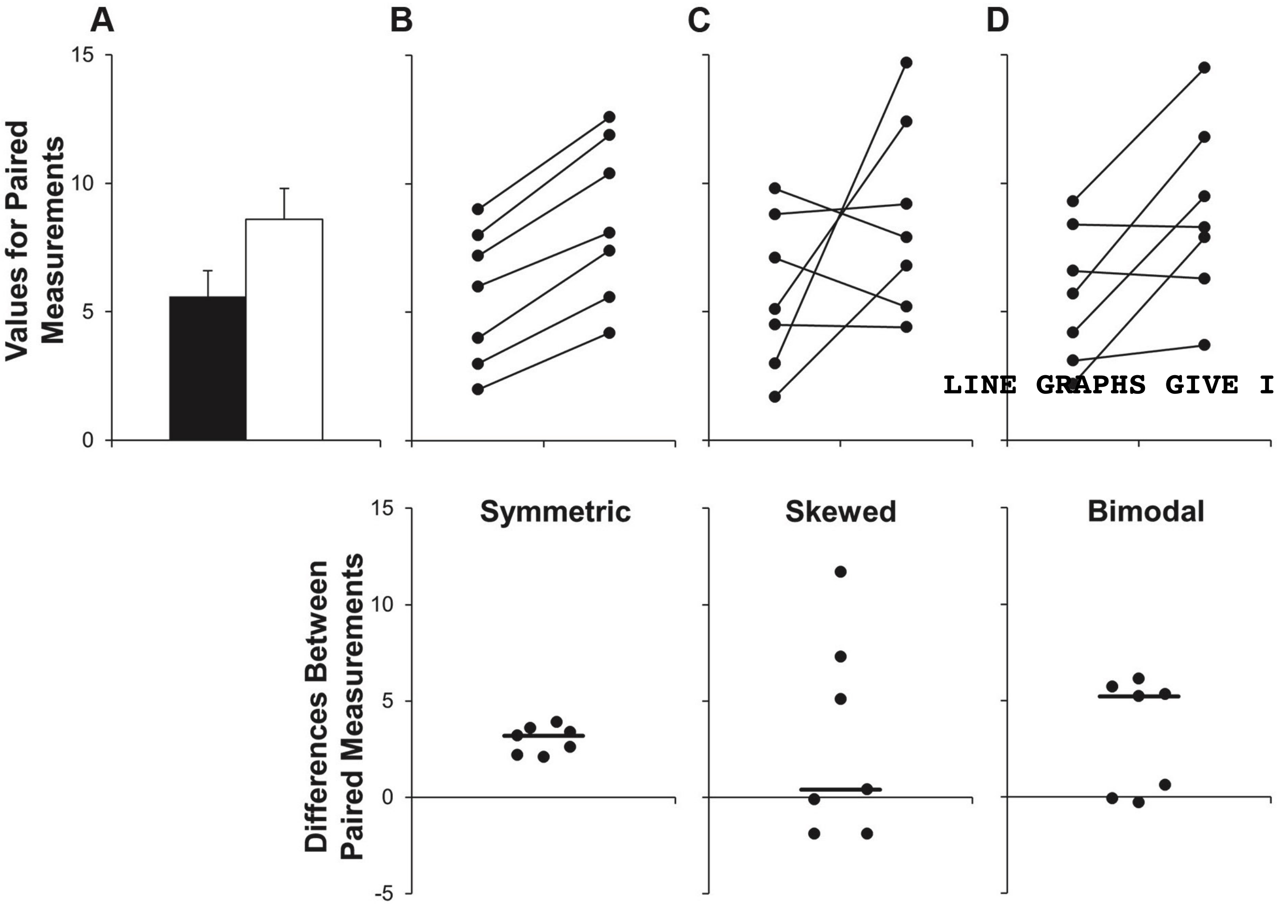
Many different data can generate the same bar graph:



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

Why show the data?

Paired data: additional problems:

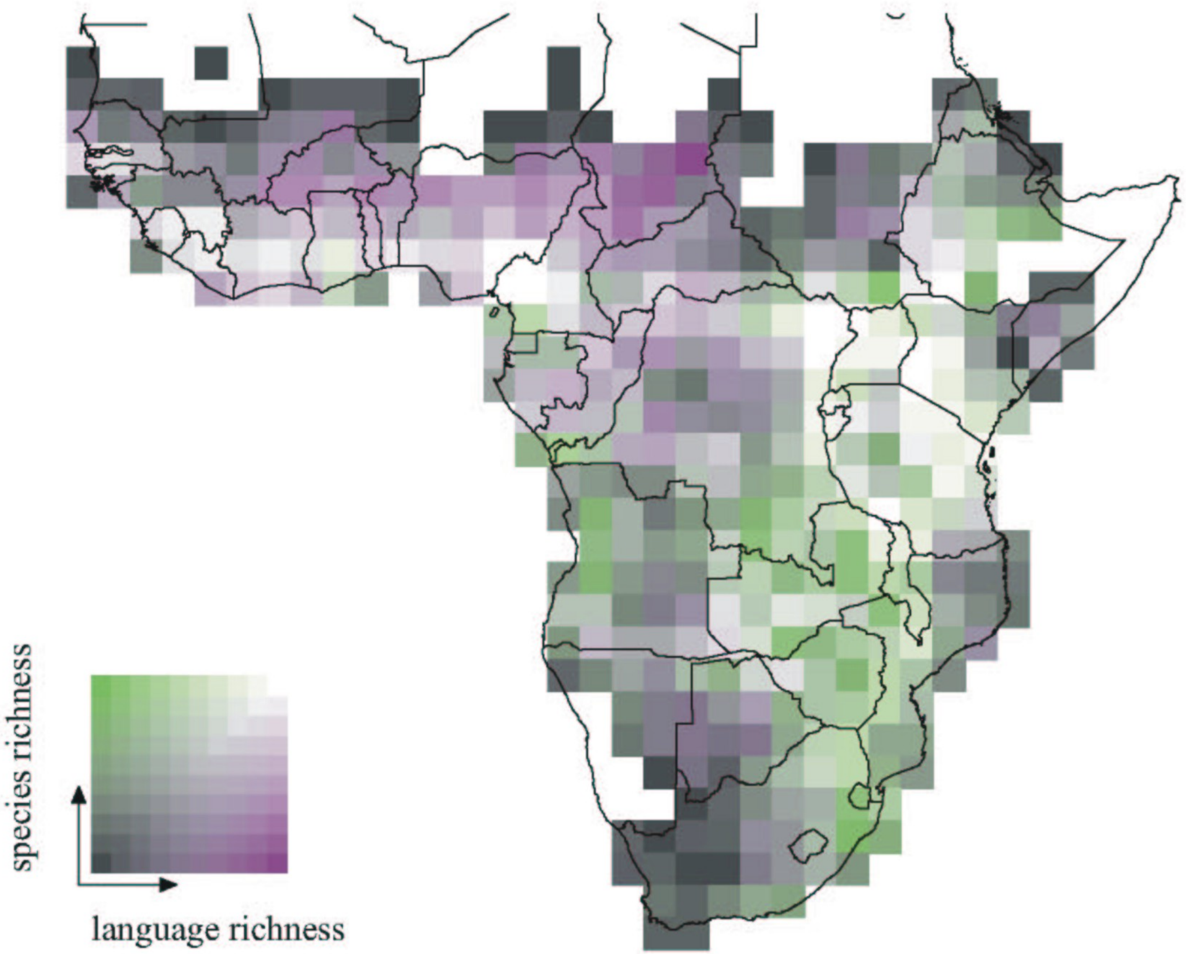


LINE GRAPHS GIVE IDEAS, BAR GRAPH HIDES INFORMATION

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

2. Make patterns in the data easy to see.

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



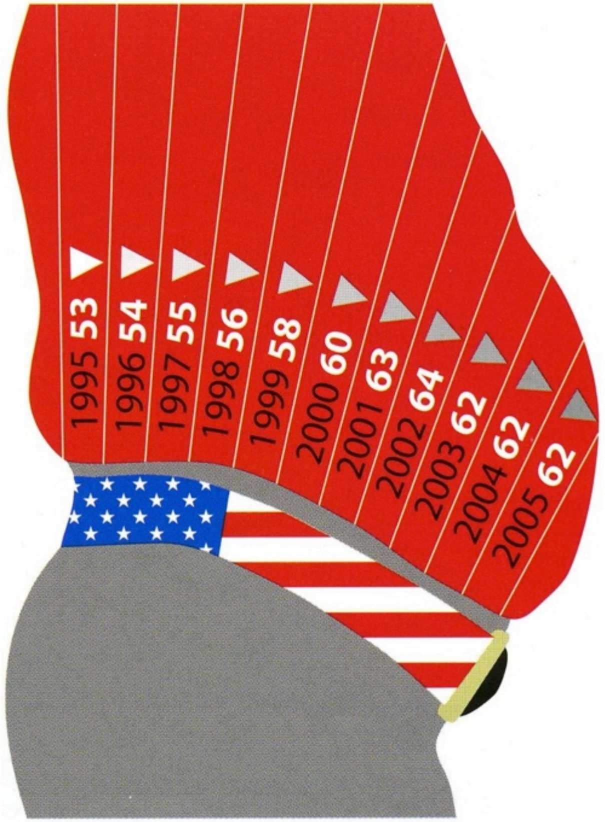
Map displaying the number of bird species and the number of distinct human languages present in each square of a grid of continental Africa. Reproduced from Moore et al. (2002).

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. Draw graphical elements clearly, minimizing clutter

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



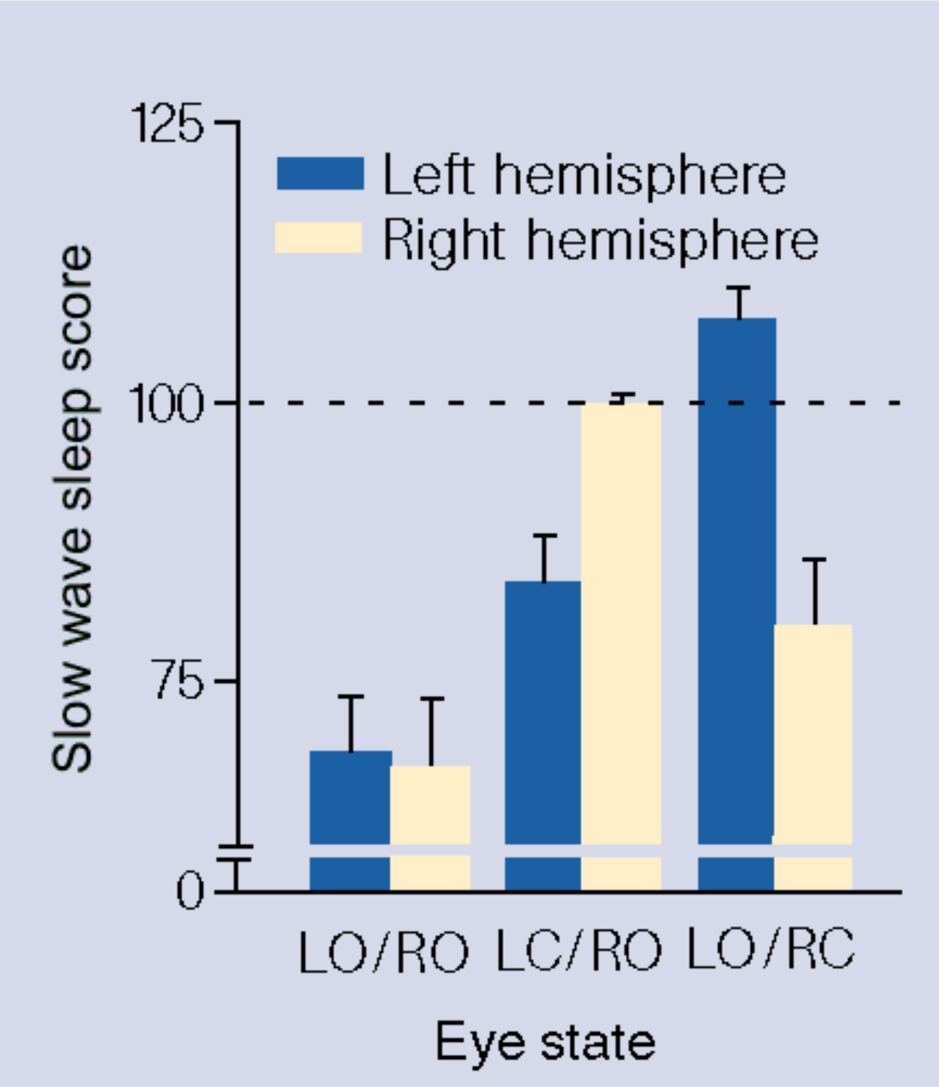
The percentage of adults over 18 with a "body mass index" greater than 25 in different years (The Economist 2006). Body mass index is a measure of weight relative to height.

What is the pattern in these data? Does the art help to show it?

What would be a better graphical method to show the pattern in the data?

4. Represent magnitudes honestly

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



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

Are the bars “consistent with the numerical representation”?

Is 0 a reasonable baseline for evaluating sleep score?

Are there other issues with the graph?

Basic graphs used to achieve these principles in ecology and evolution

1. Displaying *frequency distributions*:

- Bar graphs
- Histograms

2. Displaying *associations* between variables and differences between groups:

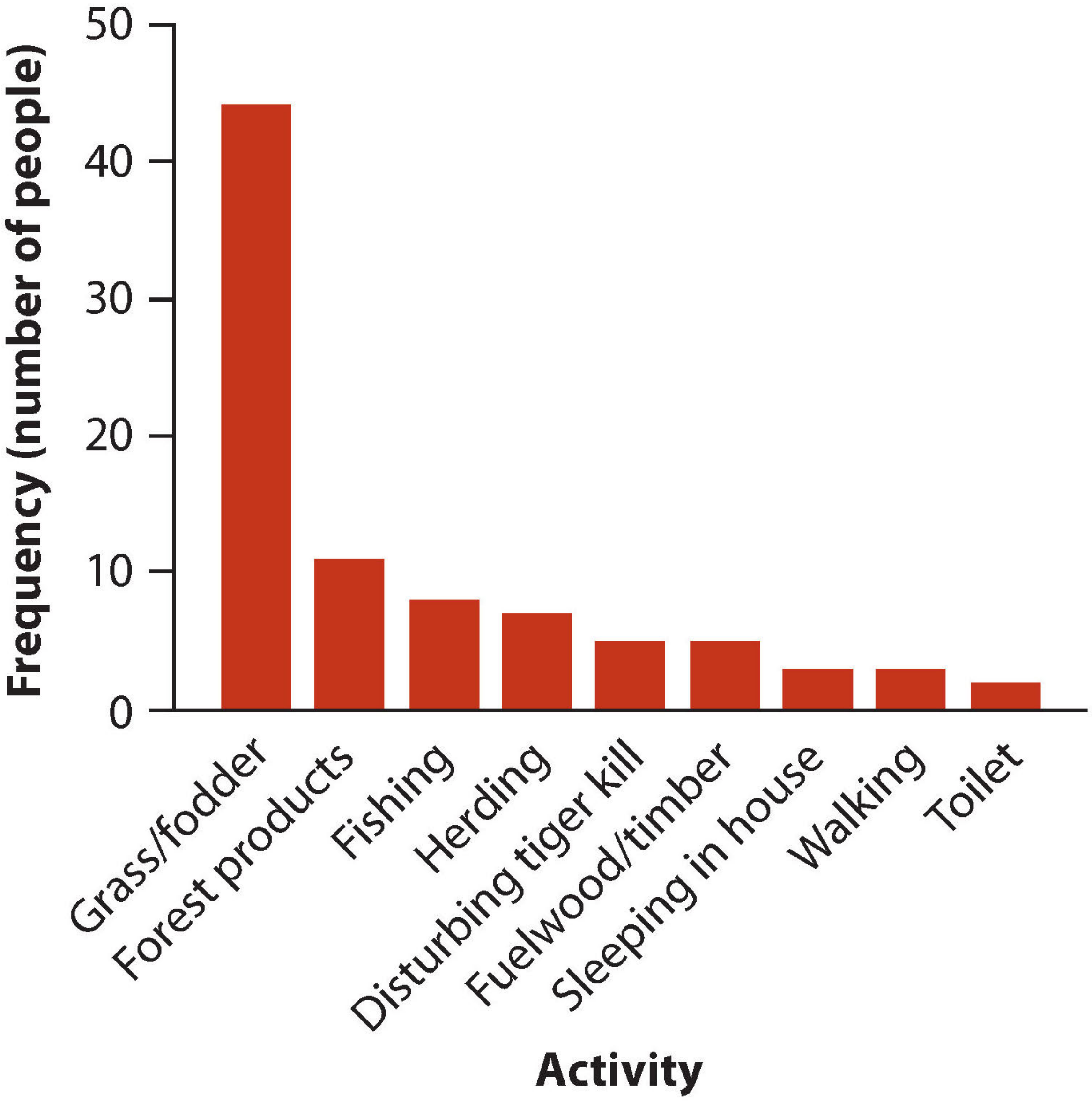
- Grouped bar graph
- Mosaic plot
- Box plot
- Scatter plot
- Strip chart

Bar graph

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

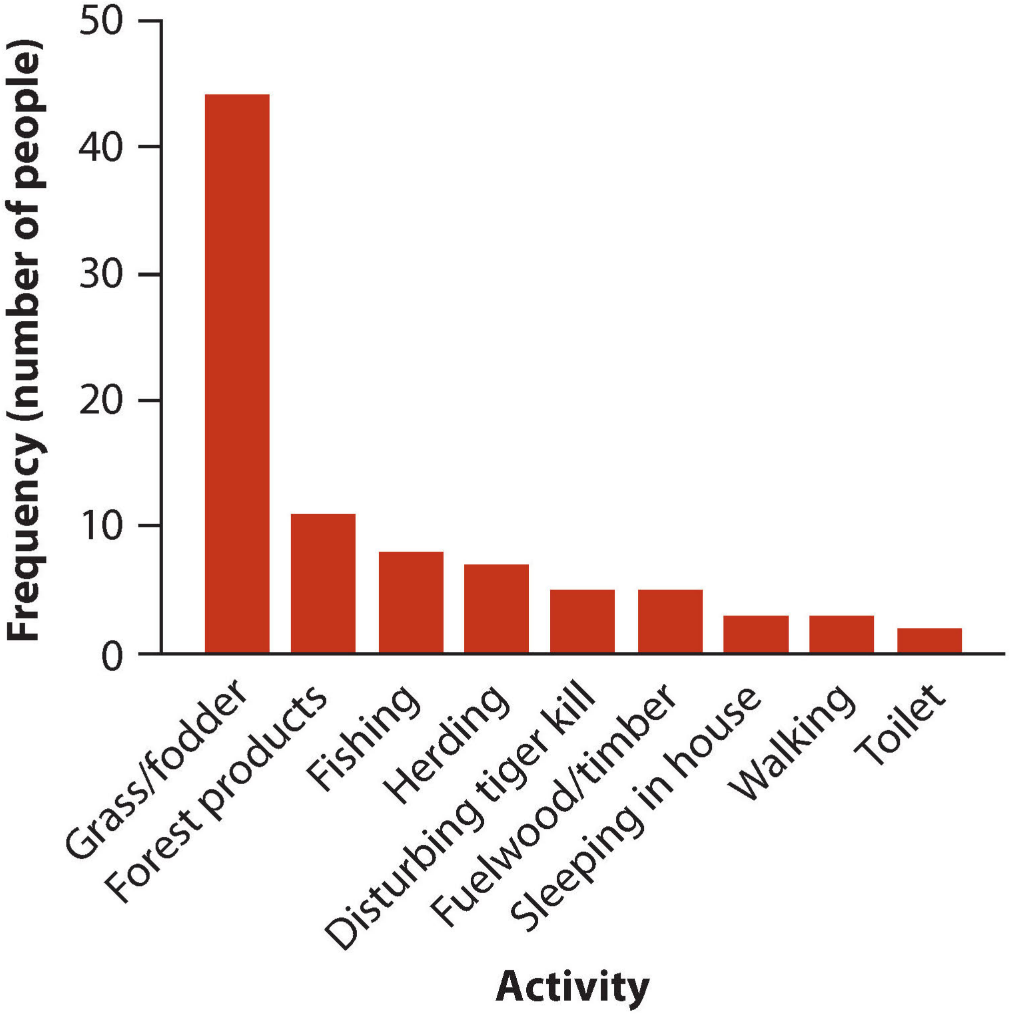
Activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006.

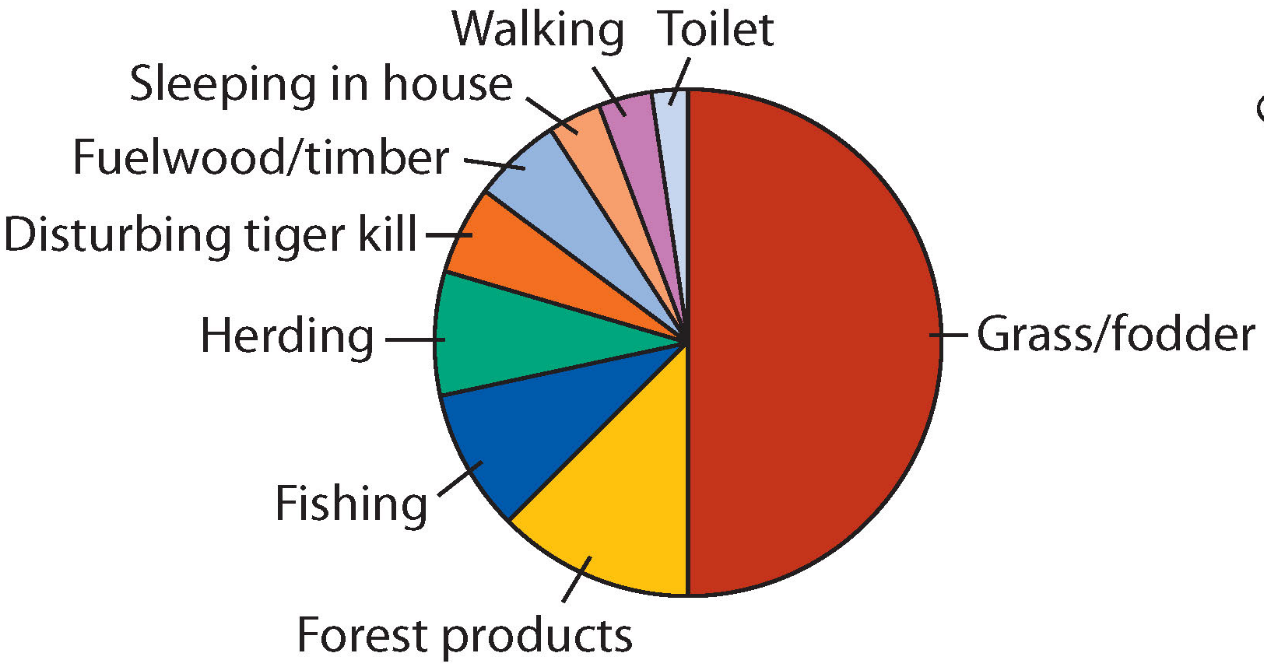


Bar graph vs. Pie chart

Q: which is more successful?

It is often said that humans are poor at comparing areas in pie charts



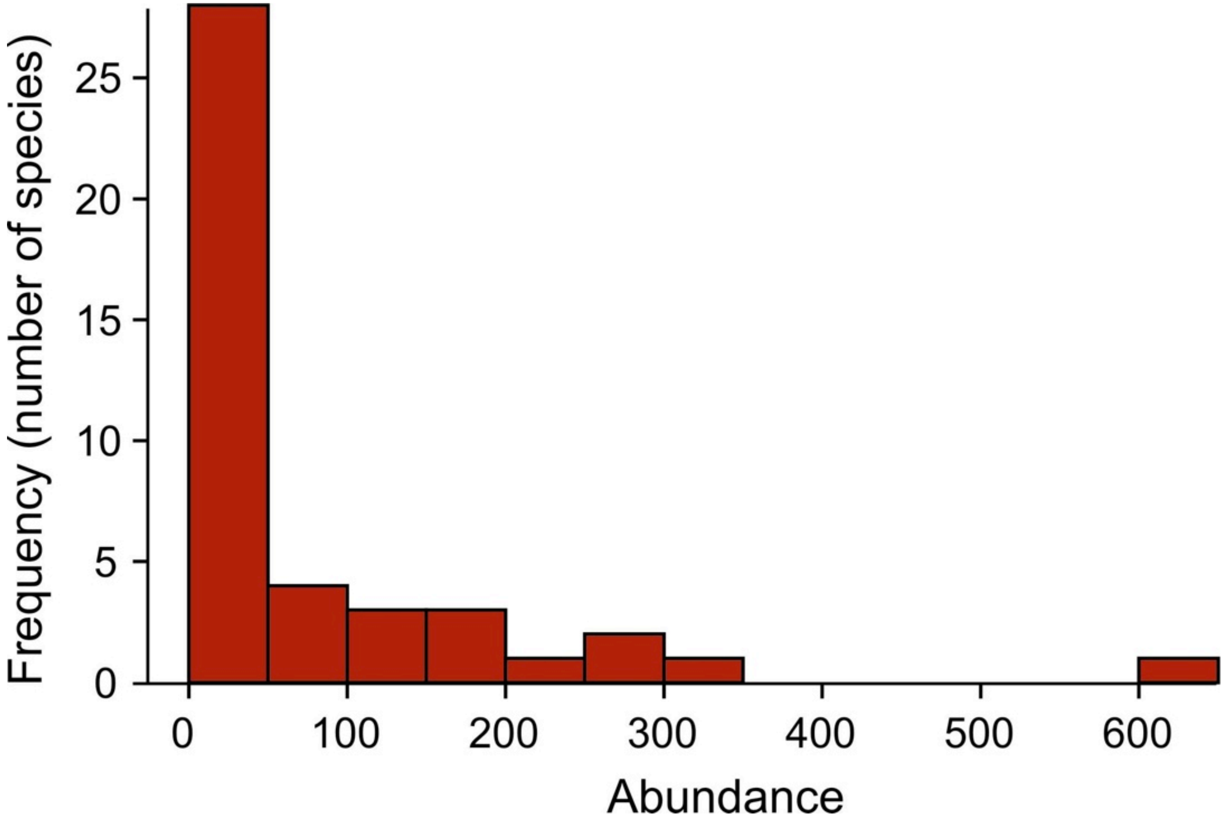


Histogram

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

continuous data



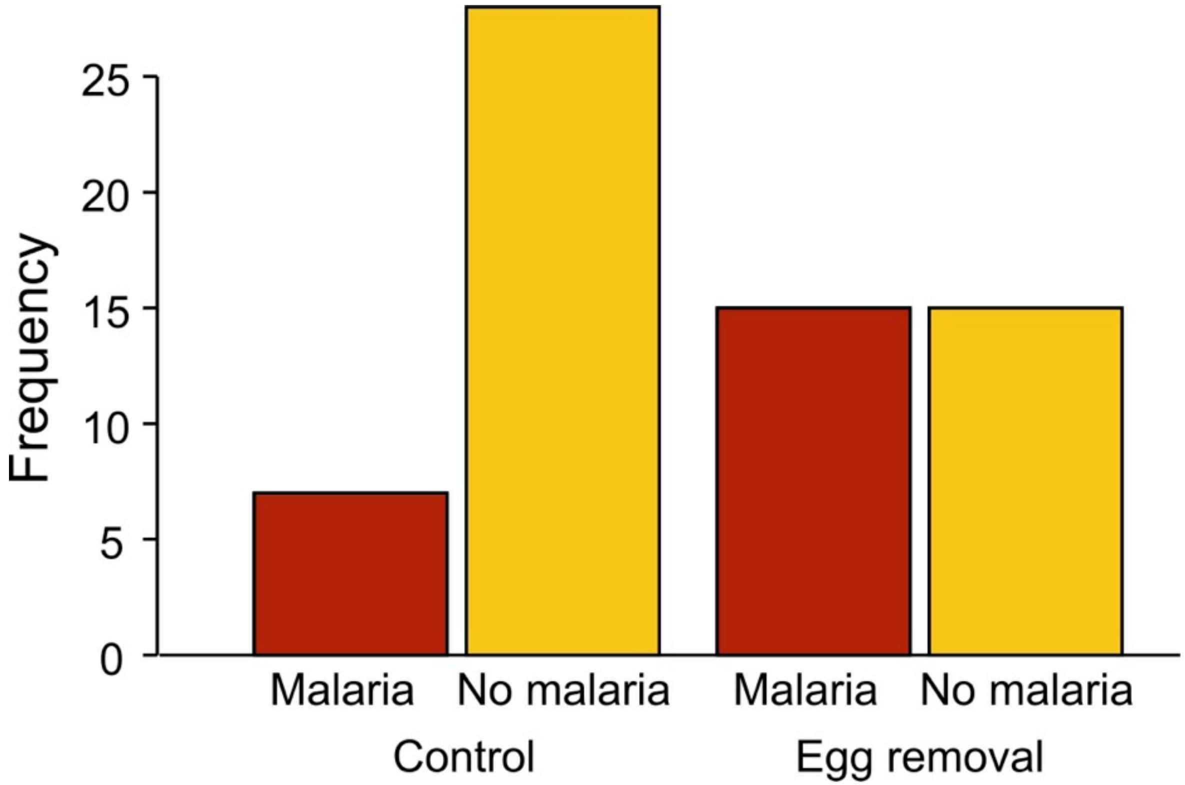
The frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. n = 43 species

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.



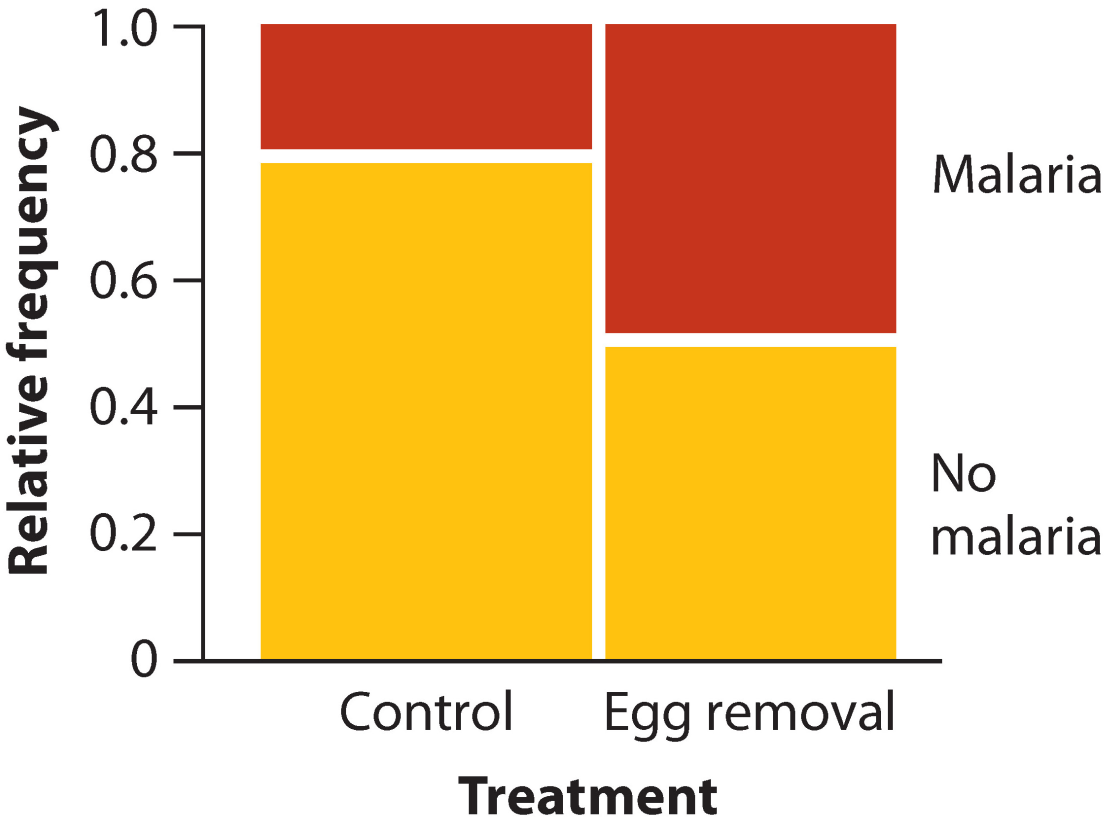
Mosaic plot

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

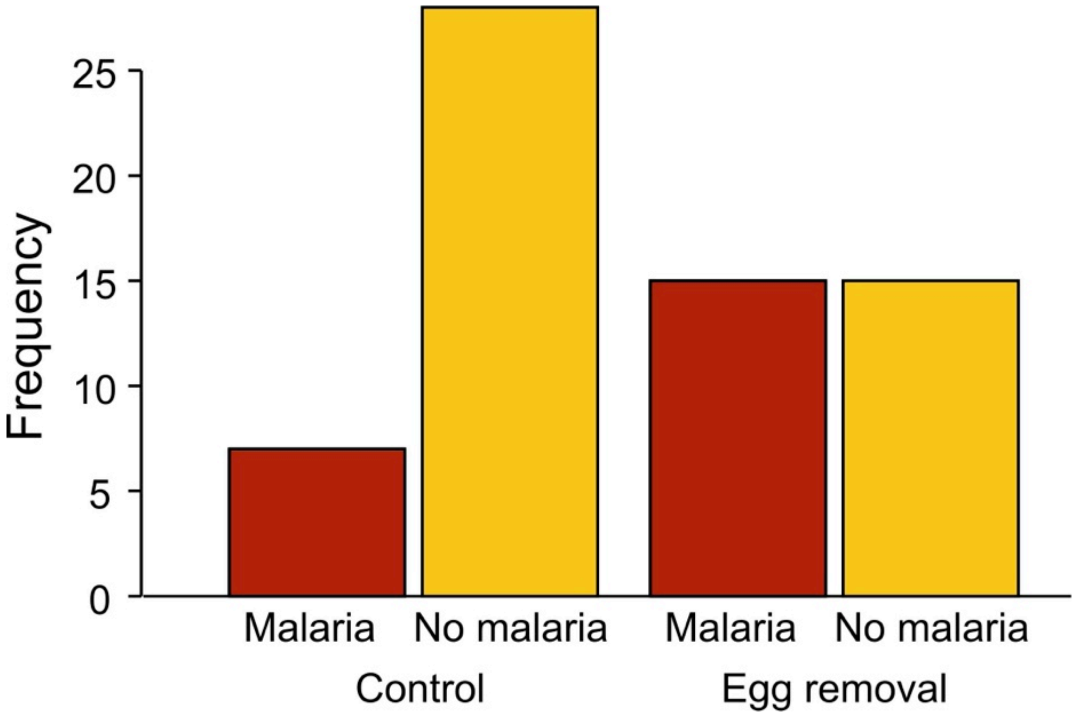
width of bar represents relative sample size

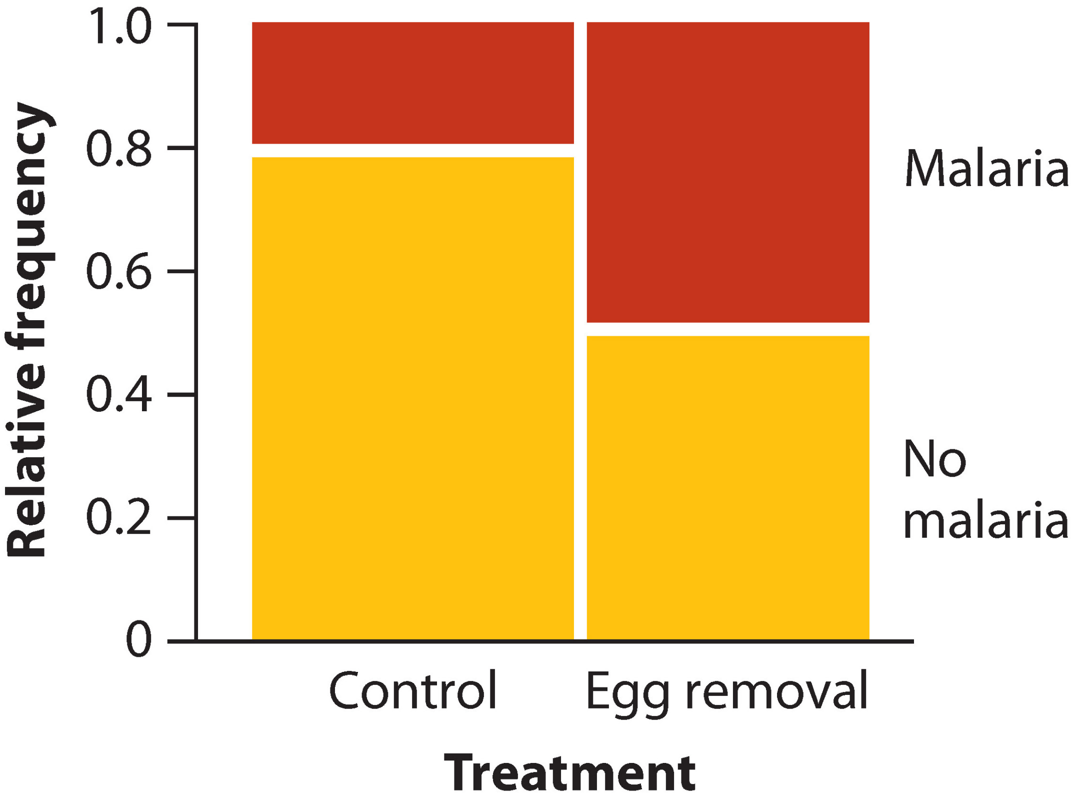
Incidence of malaria in female great tits in relation to experimental treatment. n = 65 birds.



Grouped bar graph vs mosaic plot

Q: which is more successful?



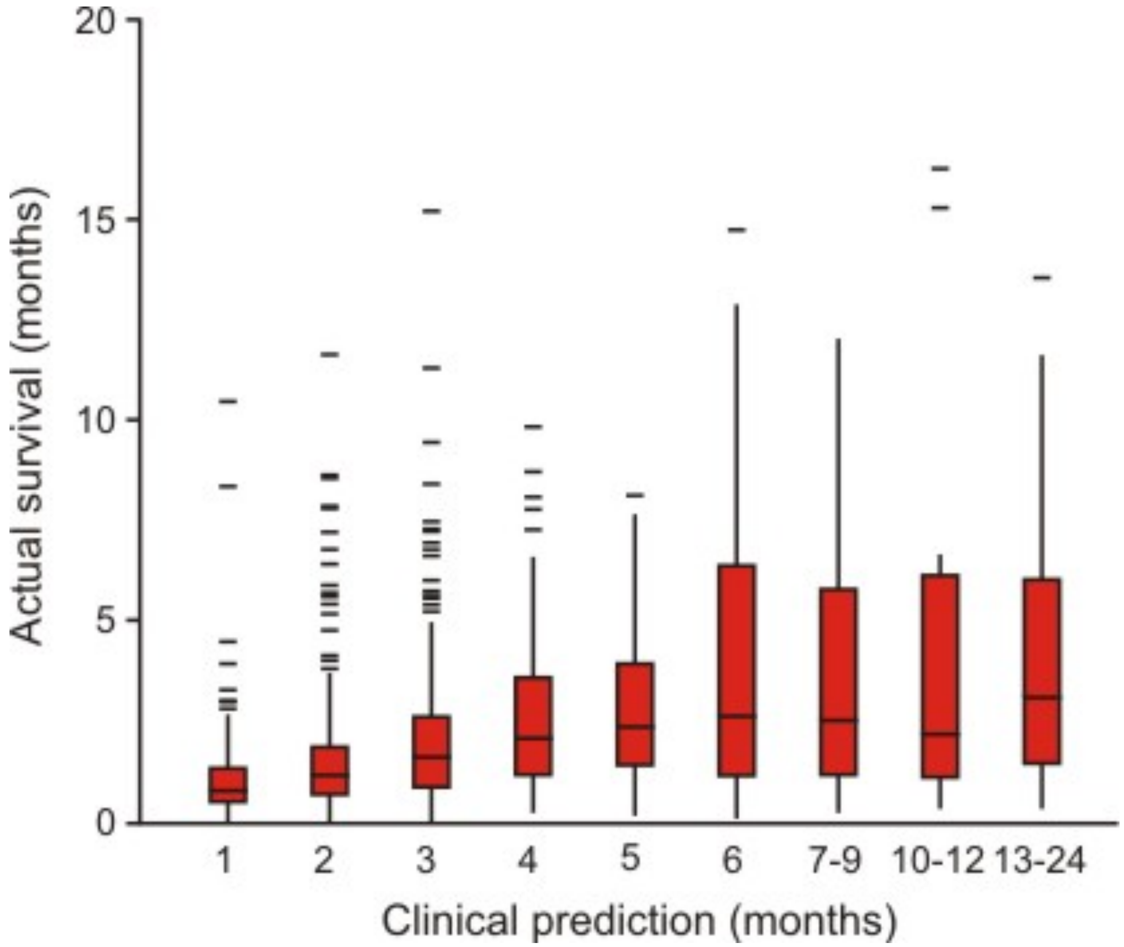


Box plot

Displays differences between groups in key features of frequency distributions

- Displays median, first and third quartile, range, and extreme observations
- More compact than plotting a separate histogram for each group
- Non-zero baseline often ok (goal is to show differences not amounts)

Survival times of terminally ill cancer patients with the clinical prediction of their survival times

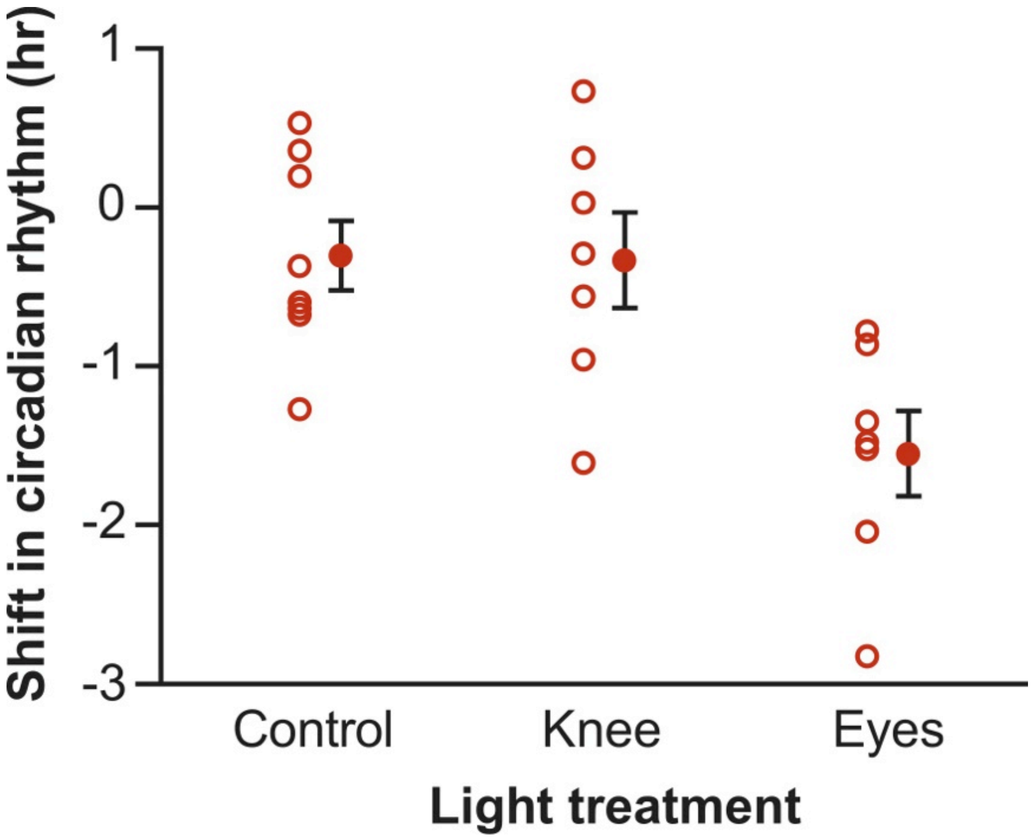


Strip chart

Displays differences between groups

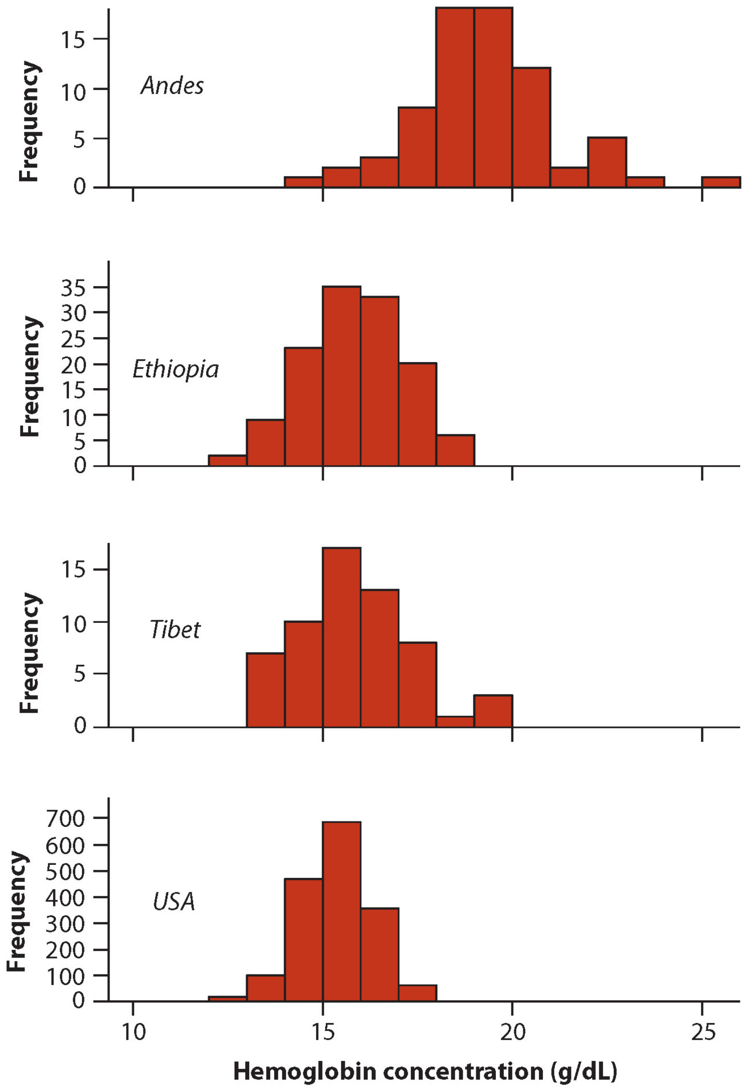
- Shows the data points
- Non-zero baseline often ok (goal is association not magnitude or frequency)
- Points fill the space available

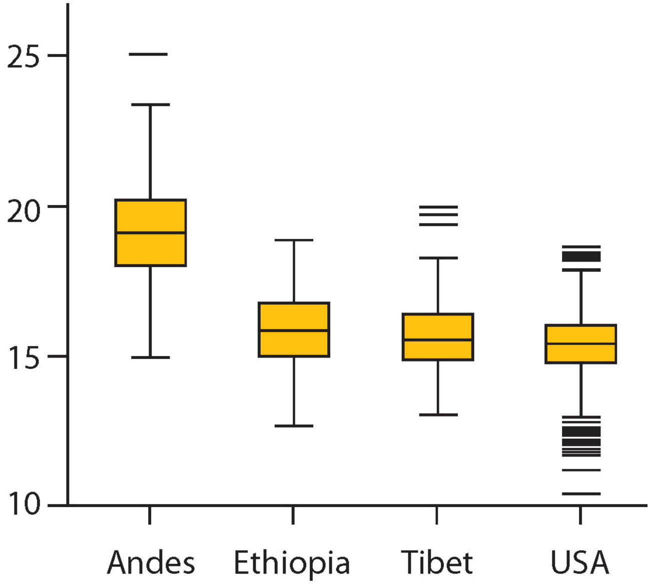
Phase shift in the circadian rhythm of melatonin production in 22 subjects given alternative light treatments (open circles). Group means ± 1 SE also shown.

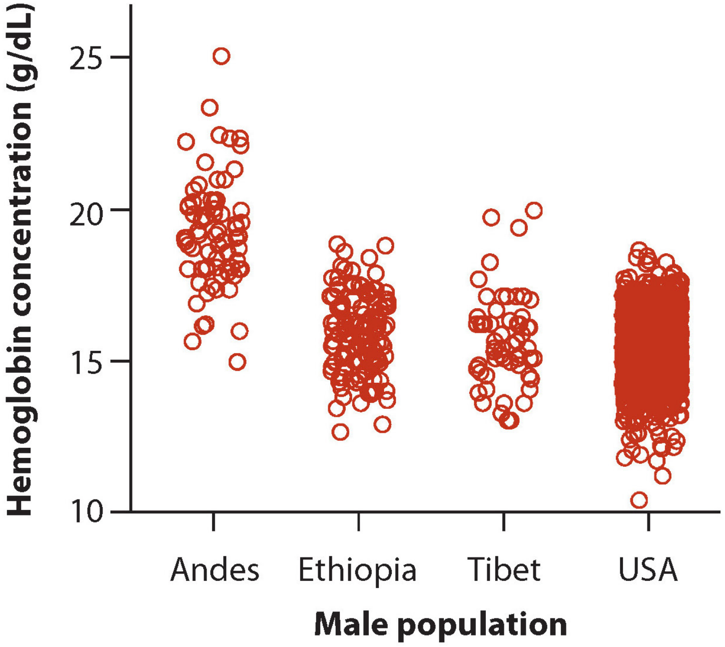


Multiple histograms, box plot, strip chart

Q: which is more successful?





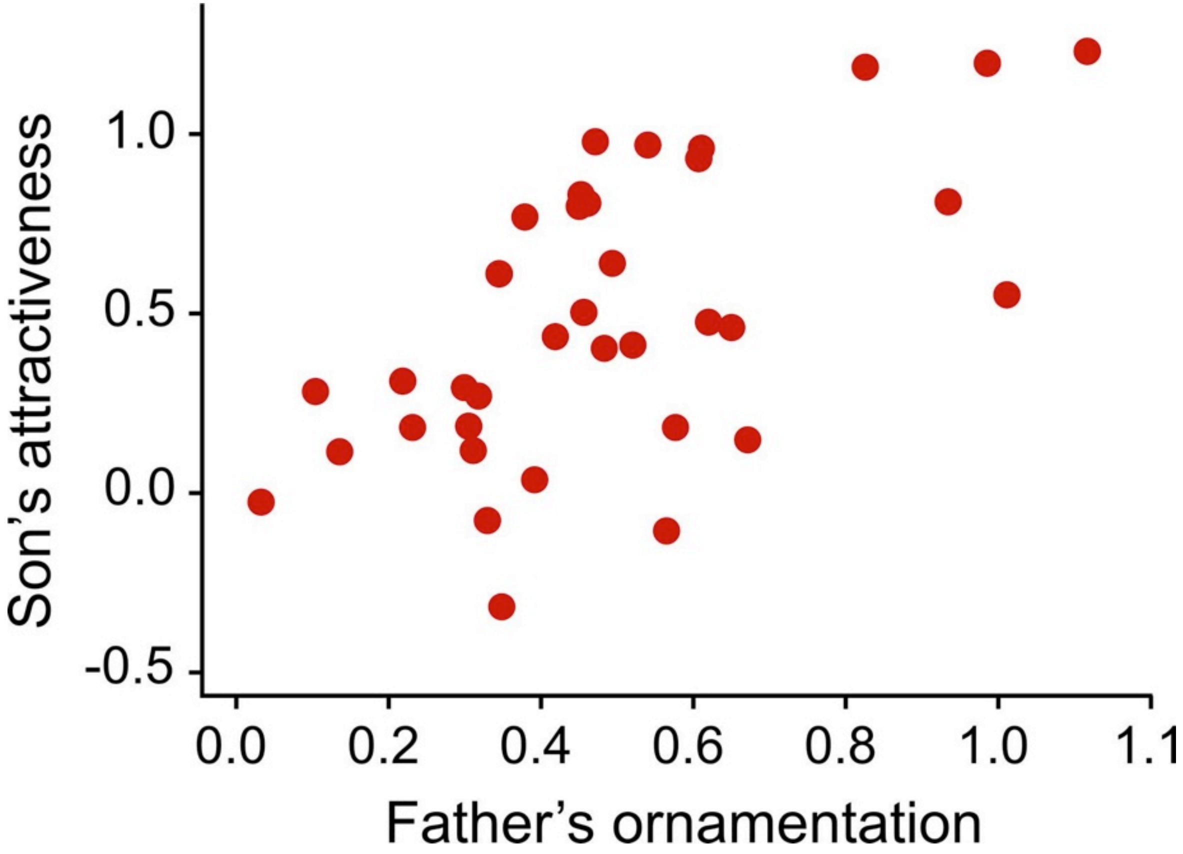


Scatter plot

Displays association between two numerical variables

- Non-zero baseline often ok (goal is to show association, not magnitude or frequency)
- Points fill the space available

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



Examples of bad graphs (how do they fail, and how to improve them)

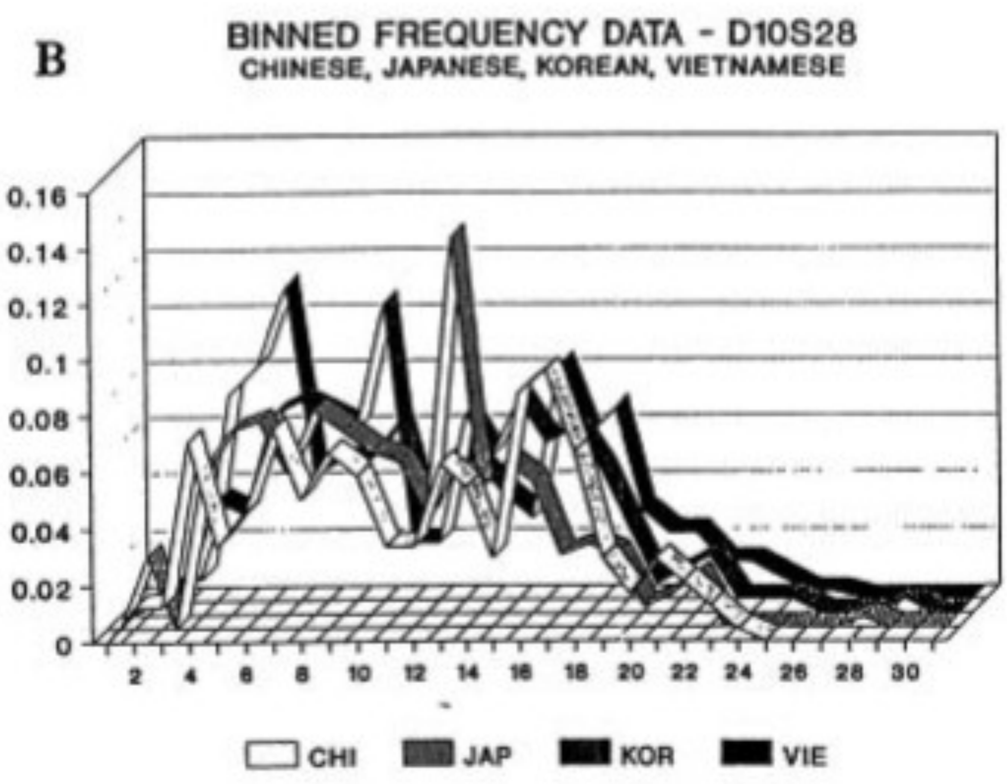


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for* D4S139 *and 120, 137, 100 and 193 for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

Some of the following were exposed by K.W. Broman ([www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)](www.biostat.wisc.edu/~kbroman/topten_worstgraphs/)

Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). Statistical Science 9:222-278, Figure 4

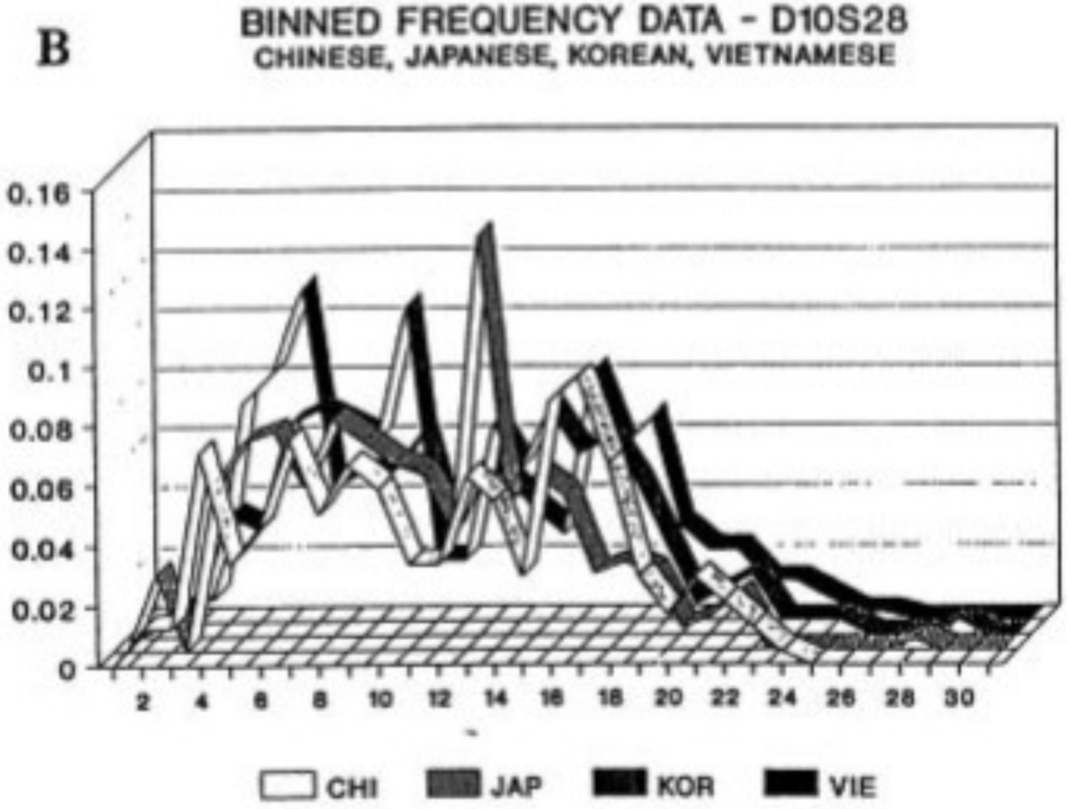


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

Problems:

What main pattern is the graph meant to show?

3-D rendering prevents us from seeing and comparing the frequency distributions.

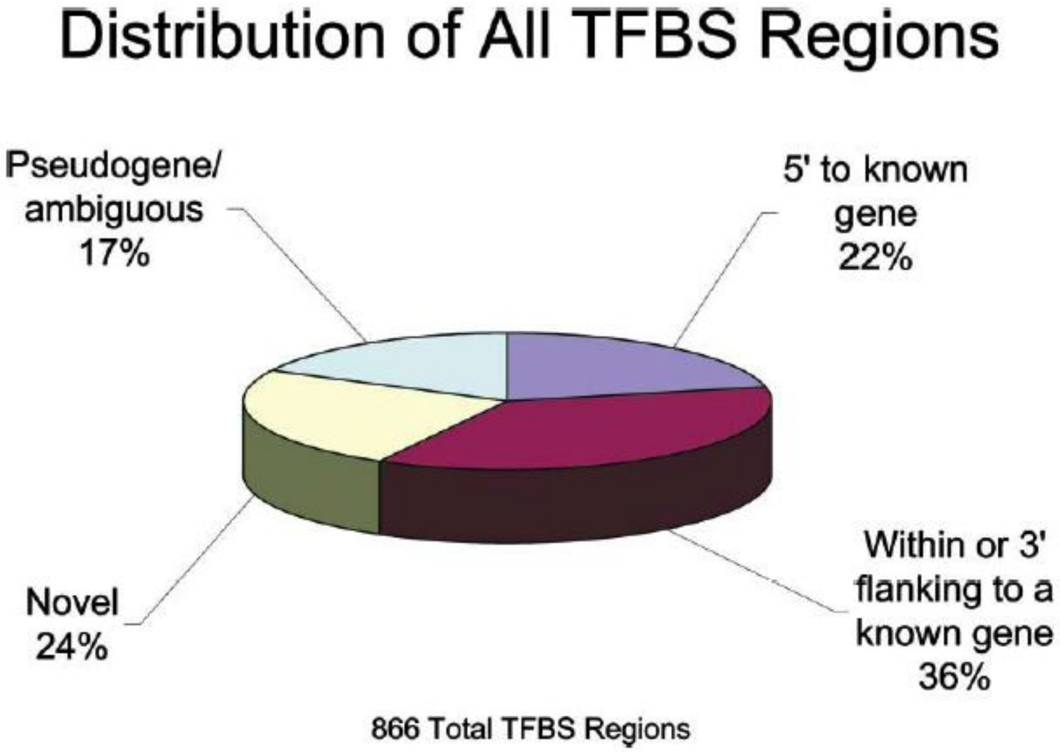
Line graph also not ideal for frequencies.

Possible solutions:

Multiple bar graphs (x-axis is not numeric, so histogram not appropriate), one per subpopulation.

3D pie chart!

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



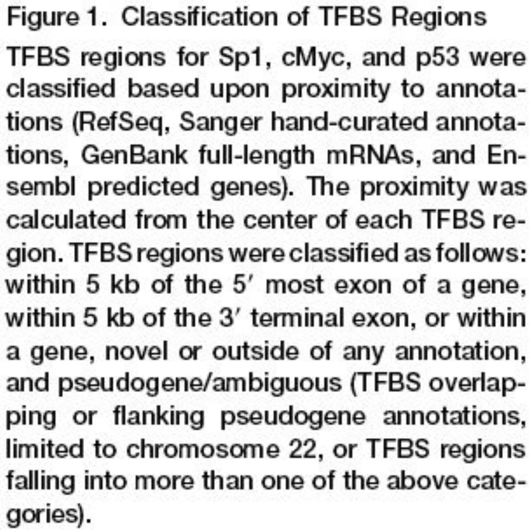
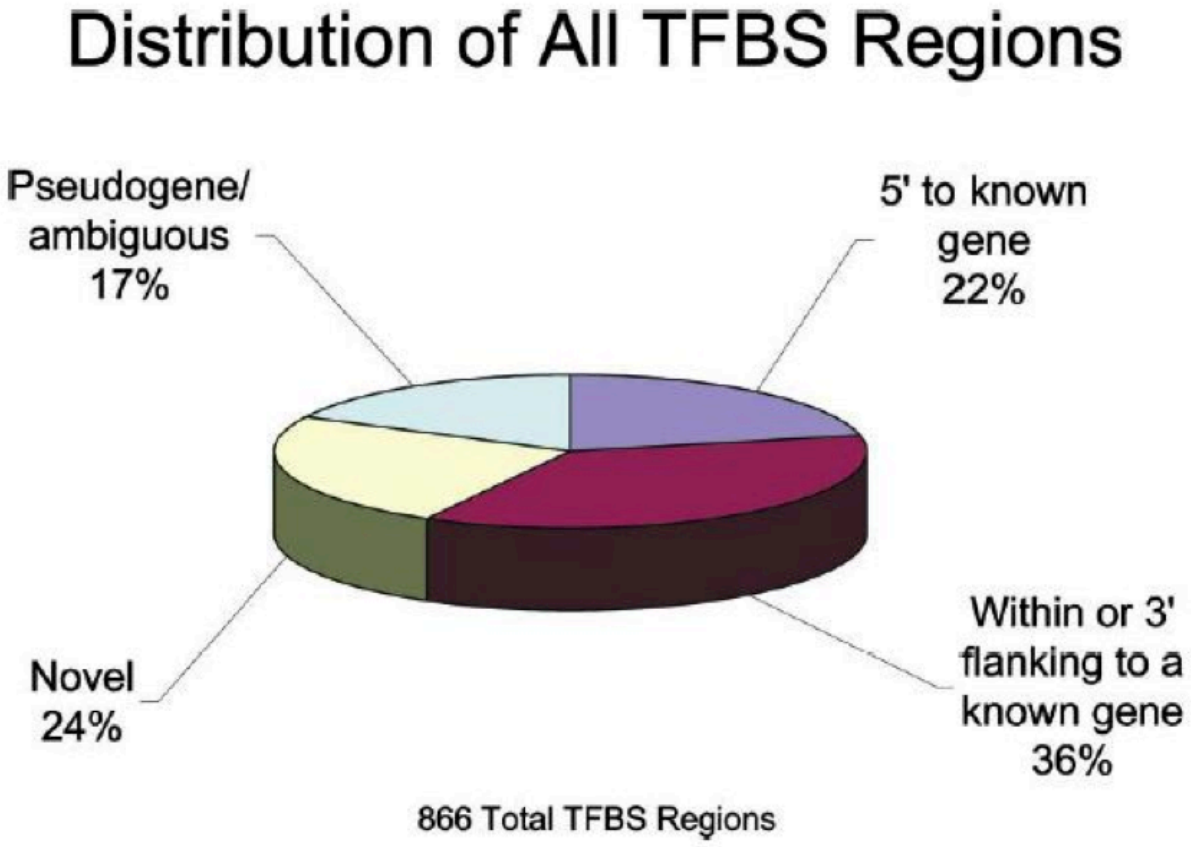


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).



Problems:

3D rendering is gratuitous and prevents eye from seeing areas of pie.

(Humans are poor at comparing areas in pie charts even in 2D.)

Any graph that is meaningful only with numbers added is a failure.

Possible solutions:

Use a 2D bar graph.

Many pies!

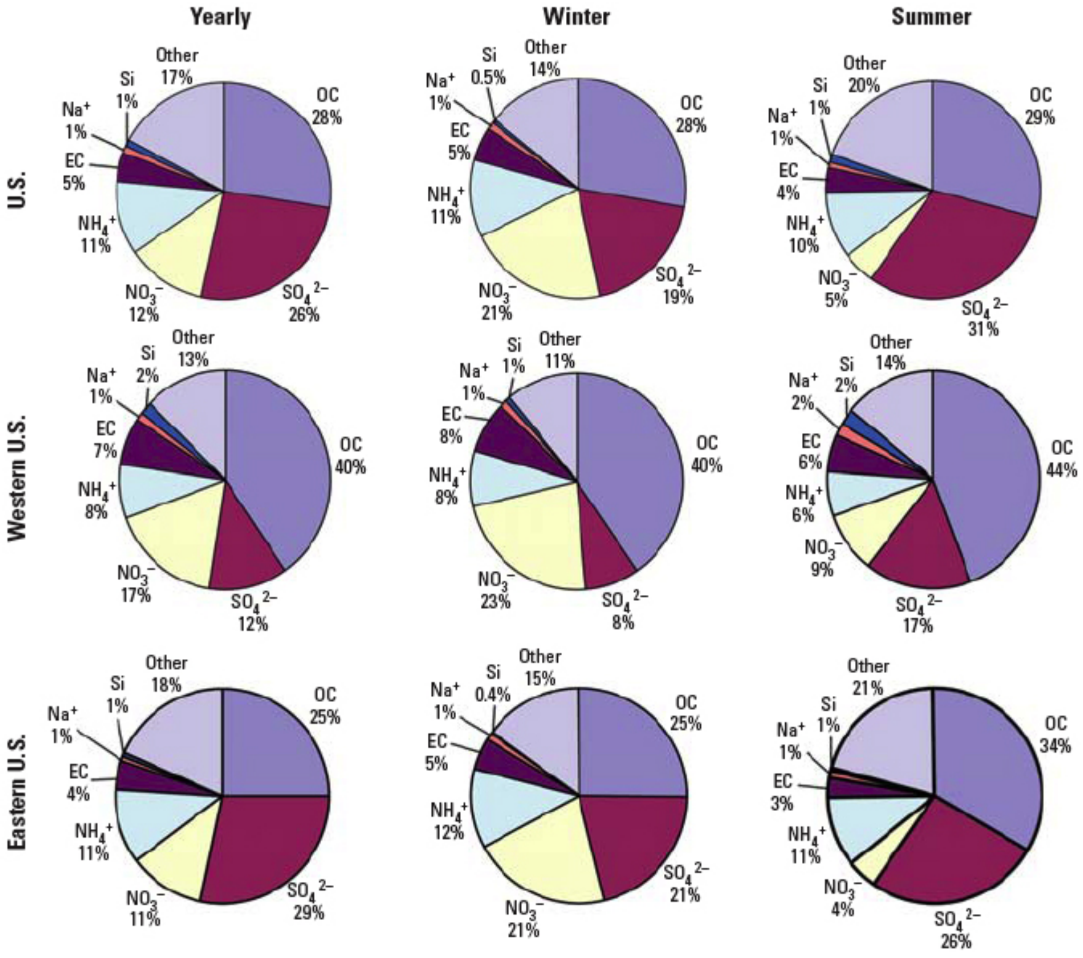


Figure 3. Percent of $PM_{2.5}$ composition by component for yearly, winter, and summer averages, by region.

The aim is to show changes in NO_3 and SO_4 between winter and summer, and consistency of change between geographic regions.

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

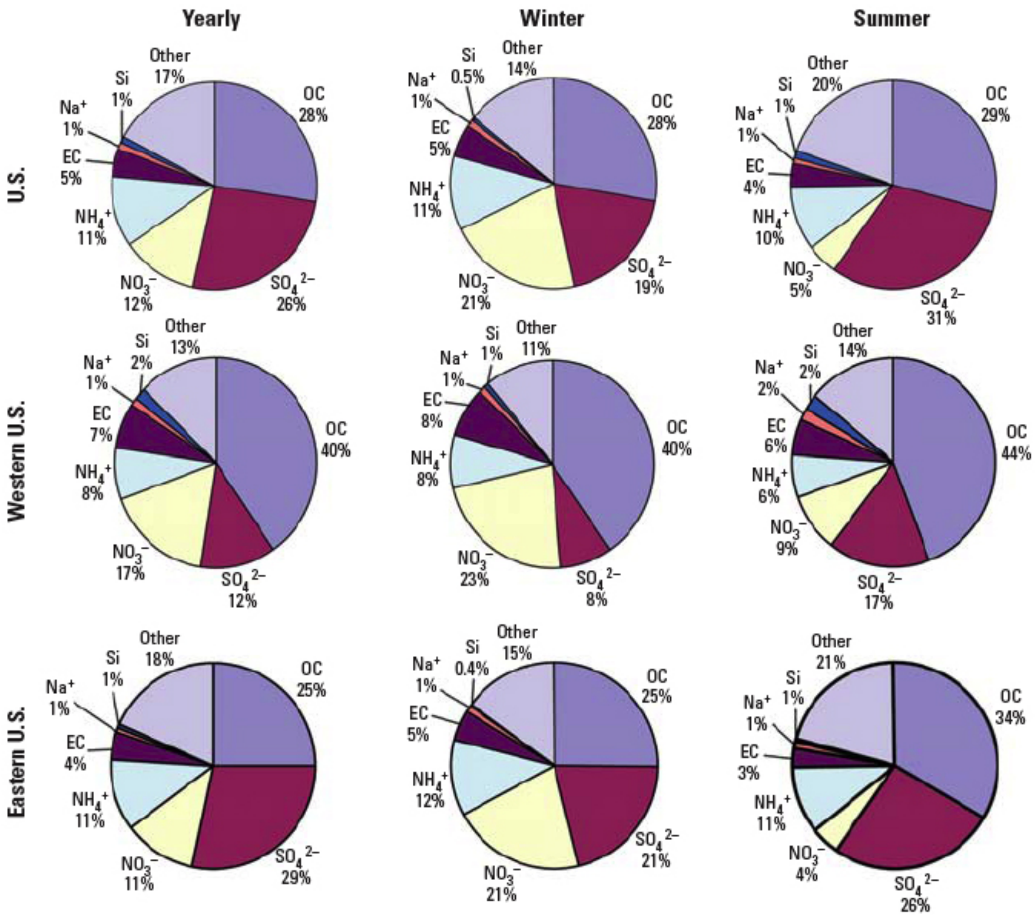


Figure 3. Percent of $PM_{2.5}$ composition by component for yearly, winter, and summer averages, by region.

Problems:

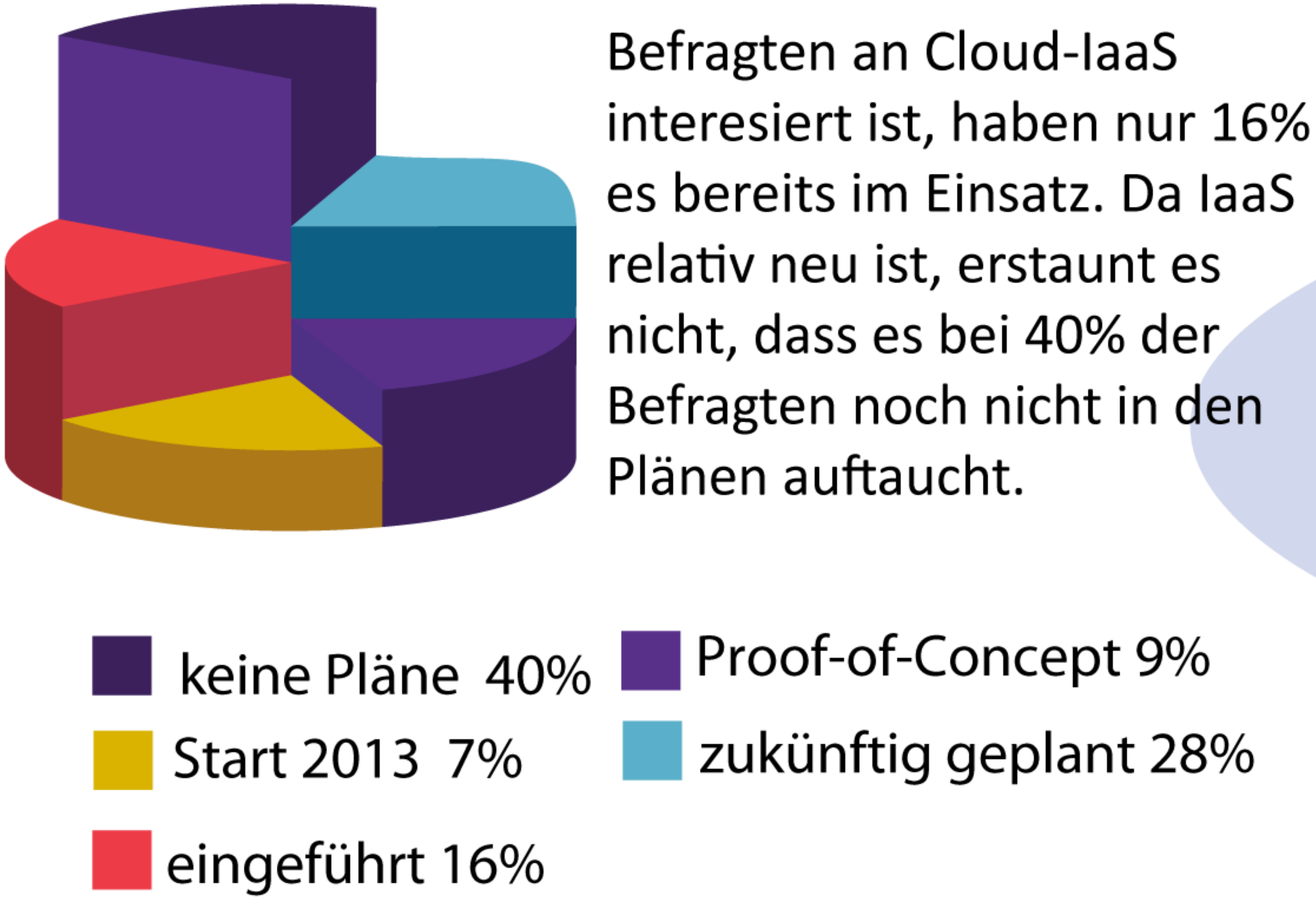
Pattern is not east to see

Possible solutions:

Use bar plots (response variable is a frequency).

If all that matters is the change from summer to winter in NO_3 and SO_4, focus the graph on this aspect rather than try to display everything.
]

3D pie



Während die Mehrheit der Befragten an Cloud-IaaS interesiert ist, haben nur 16% es bereits im Einsatz. Da IaaS relativ neu ist, erstaunt es nicht, dass es bei 40% der Befragten noch nicht in den Plänen auftaucht.

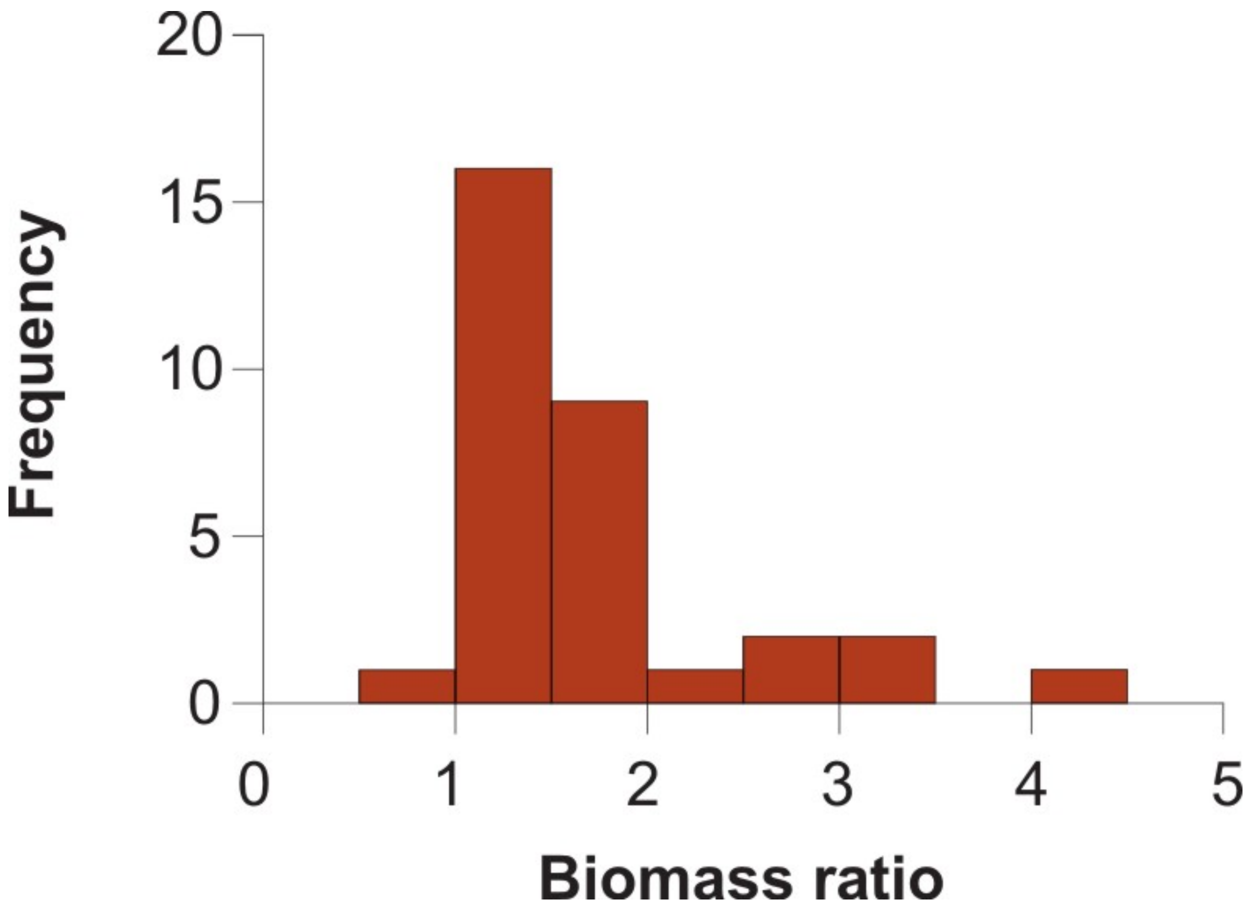
Where to begin?!

Height and area both compromised by 3D

A table alone would be vastly superior.

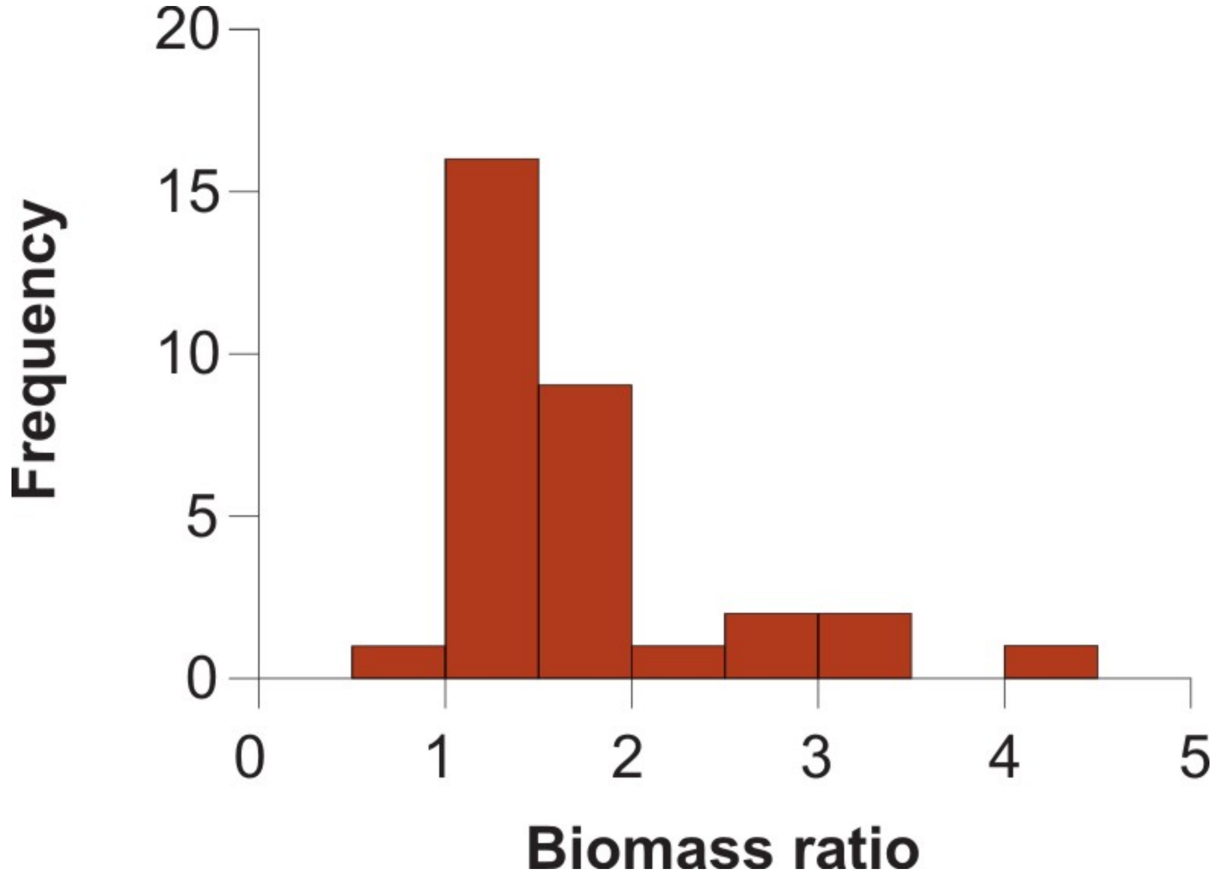
You probably would never even experience the temptation to draw something like this.

Ratio data



Whitlock, M. C. and D. Schluter. 2008. The analysis of biological data. Roberts & Company, Greenwood Village, CO, USA, Figure 13.1-4

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)



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?

Tables in main text are not for storing patterns but for illuminating patterns
(put storage tables into online Appendix or Supplement)

- 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.

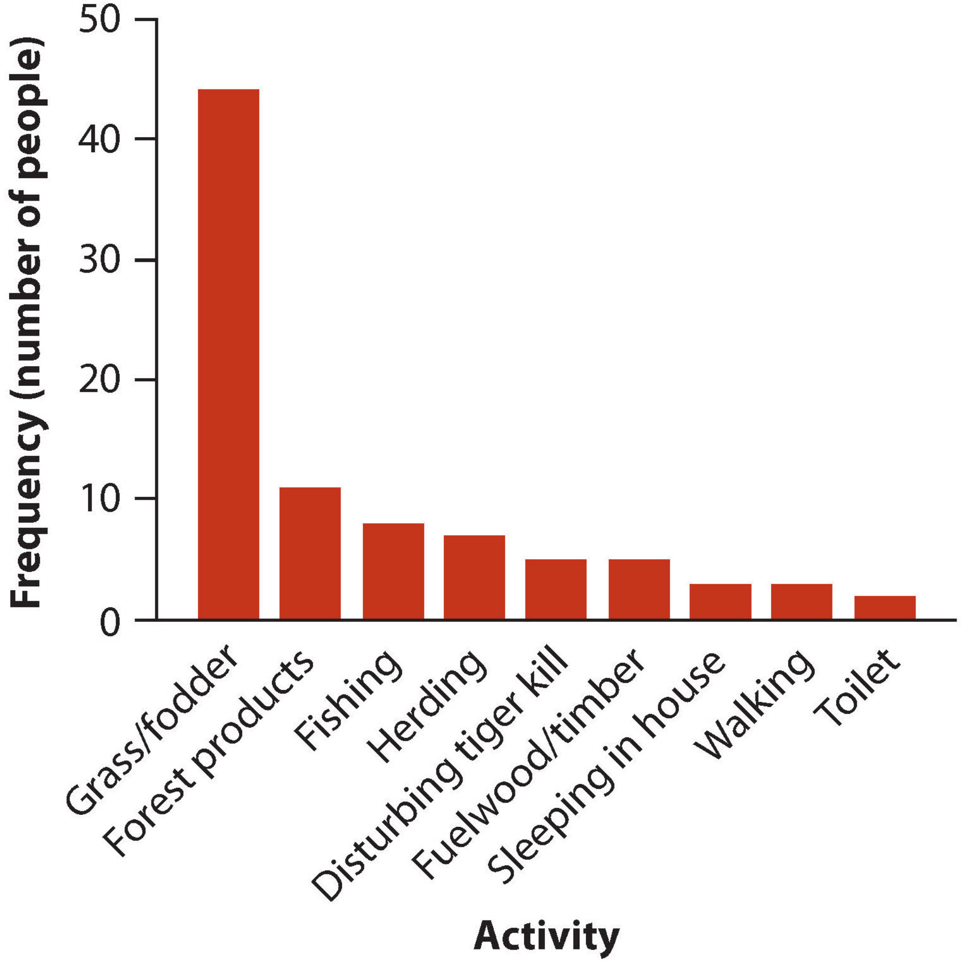
Tables should make the viewer goes “Oh!” and not “Huh?”

Frequency tables

Frequency table vs bar graph, which is preferred?

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

Activity	Frequency (number of people)
Collecting grass or fodder for livestock	44
Collecting non-timber forest products	11
Fishing	8
Herding livestock	7
Disturbing tiger at its kill	5
Collecting fuel wood or timber	5
Sleeping in a house	5
Walking in forest	3
Using an outside toilet	2
Total	88



Tables of measurements

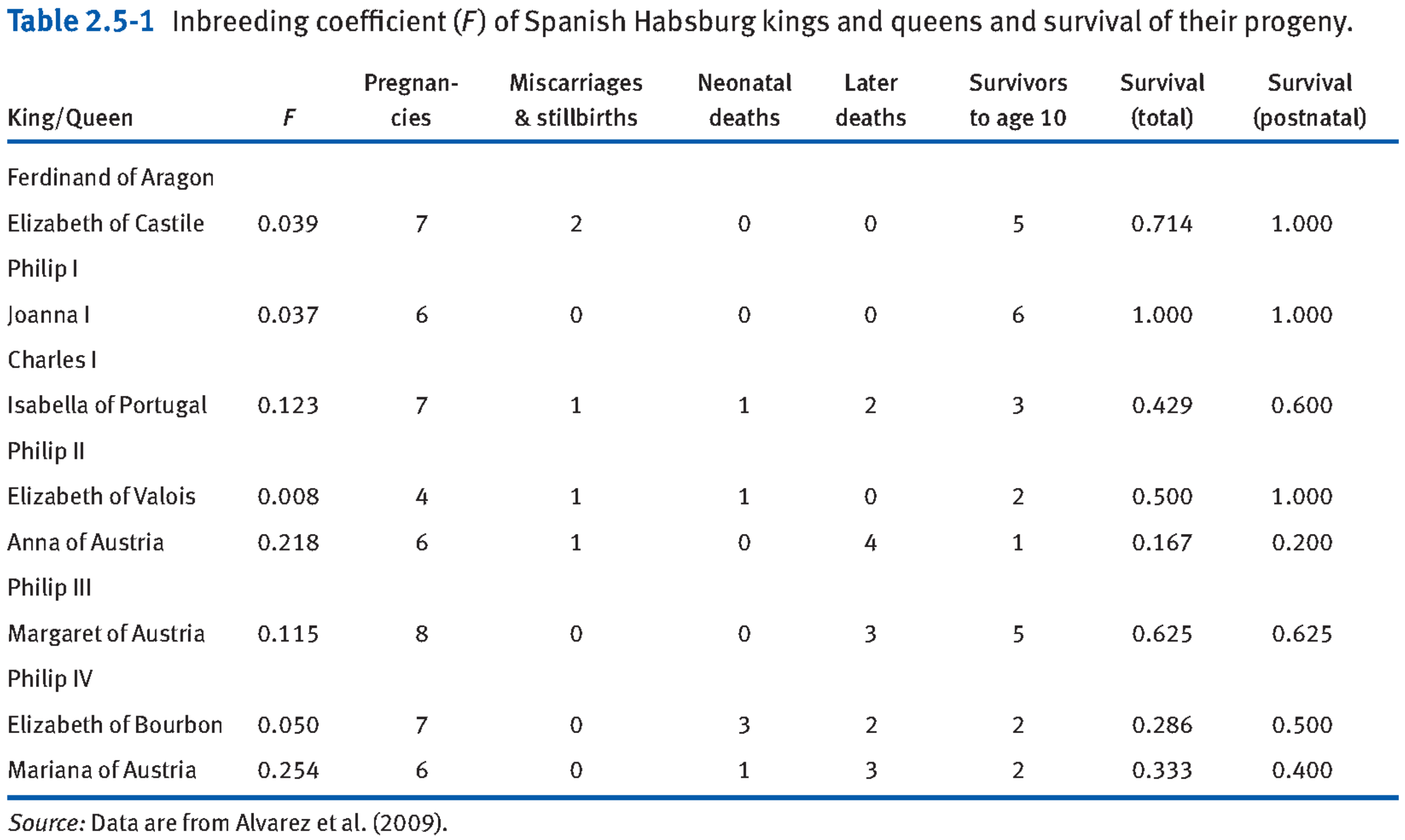


Table 2.5-1 Inbreeding coefficient (*F*) of Spanish Habsburg kings and queens and survival of their progeny.

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).

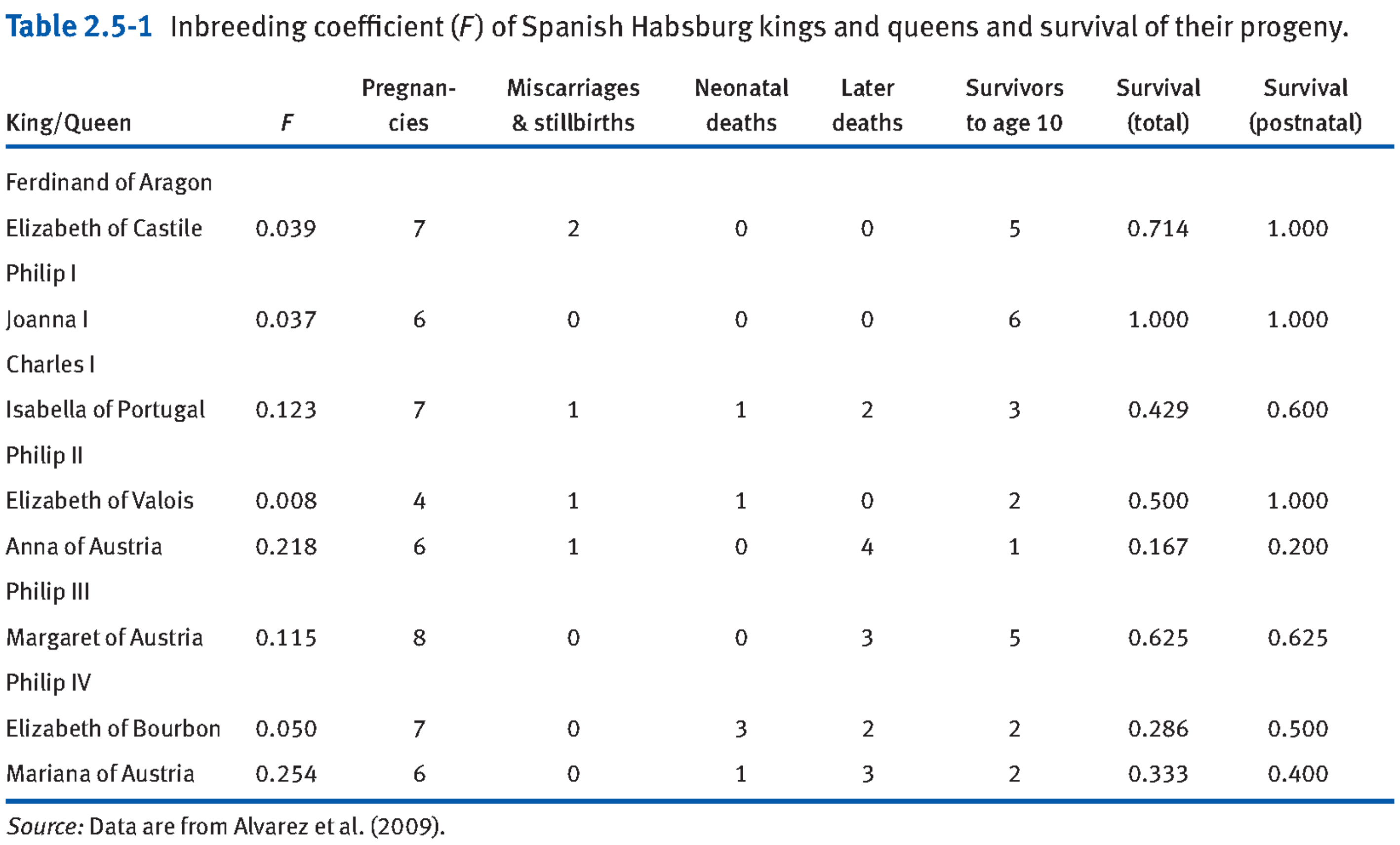


Table 2.5-1 Inbreeding coefficient (*F*) of Spanish Habsburg kings and queens and survival of their progeny.

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).

Problems:

Main comparison of interest is obscured. Can you see a relationship between *F* and survival?

Uneven line spacing, the gaps break up patterns.

Too many decimals.

Possible solutions:

Arrange the rows and columns to show association (or lack of) most clearly. Use fewer decimals. Remember, your goal should be to reveal a pattern, not simply store numbers.

Tables of measurements

Same data, revised table:

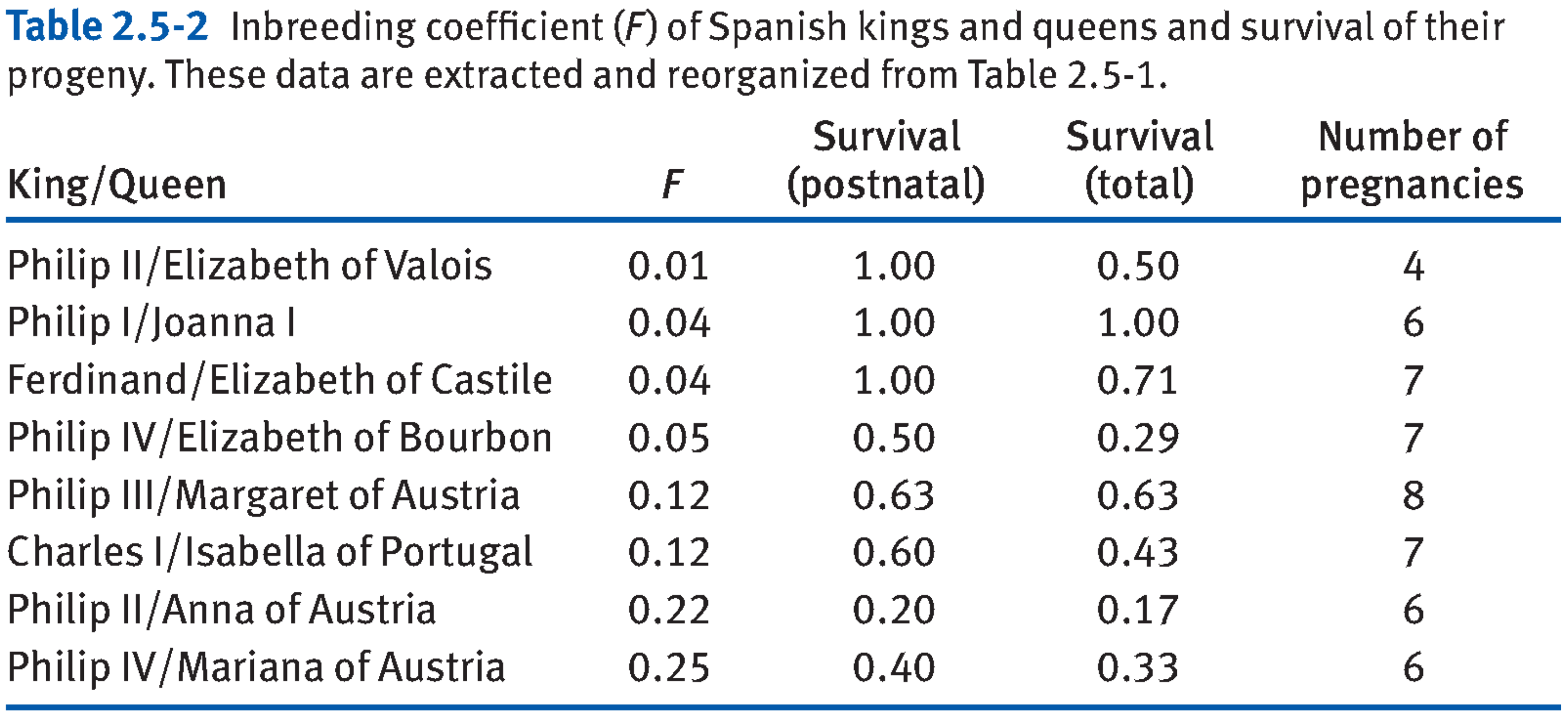


Table 2.5-2 Inbreeding coefficient (*F*) of Spanish kings and queens and survival of their progeny. These data are extracted and reorganized from Table 2.5-1.

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

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.

Need two presenters for next Tuesday, Sept 20. 15-20 minute presentation.

Need two discussion moderators.

Homework assignment 1.

Due Sept 30, 5pm.

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

Basic idea: Find a poor graph drawn from data and published by your thesis supervisor. Analyse and improve in R!