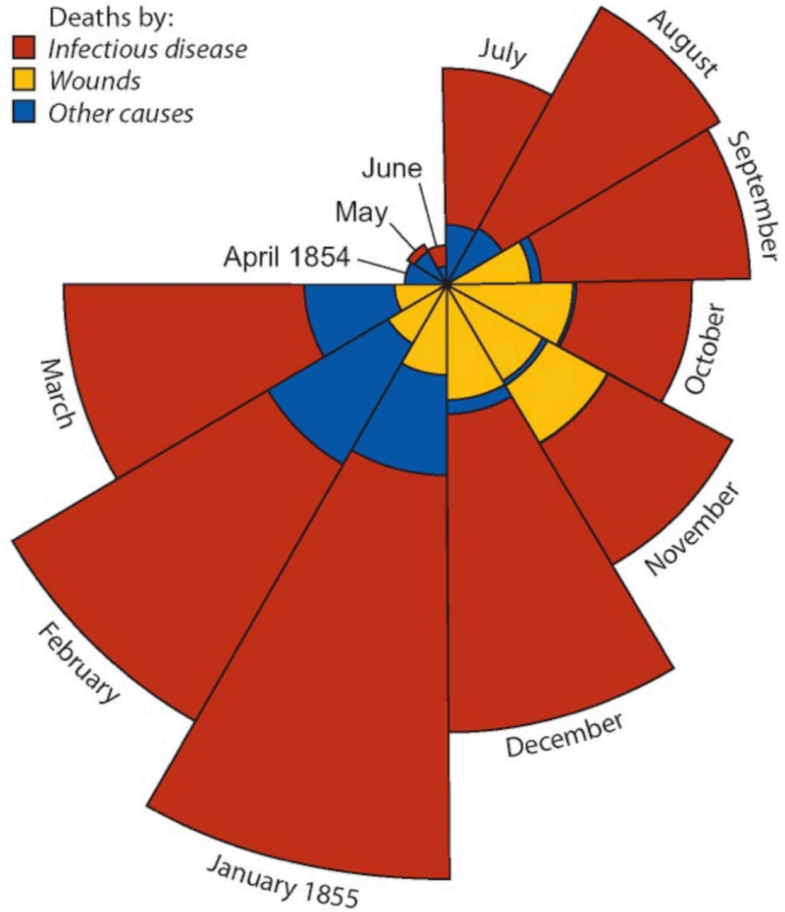
BISC-869, Graphics

Sept 19, 2024

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.



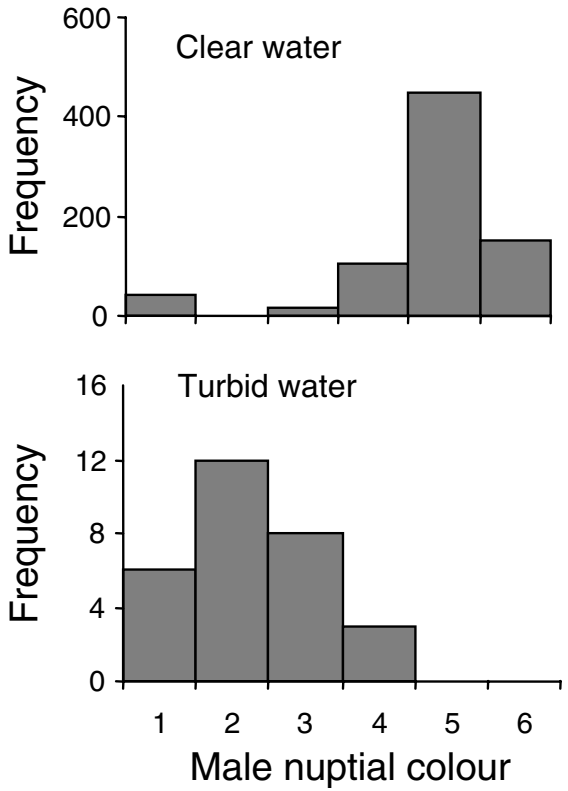
Causes of deaths in the British Army during the Crimean War (area of pie = number of deaths); F. Nightingale, 1858.

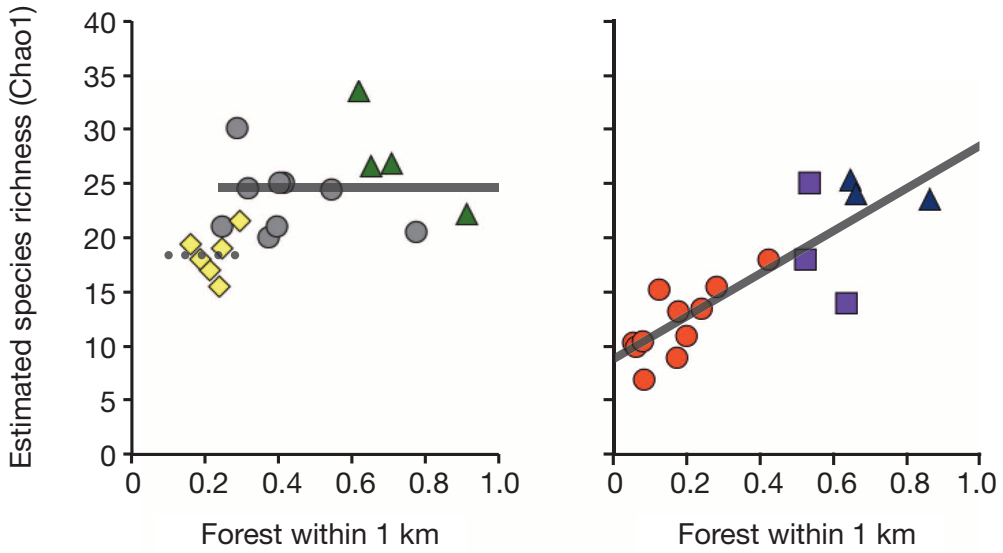
Graphs are the best method for communicating results.

1. Frequency distributions
   - The location, spread, shape of distribution

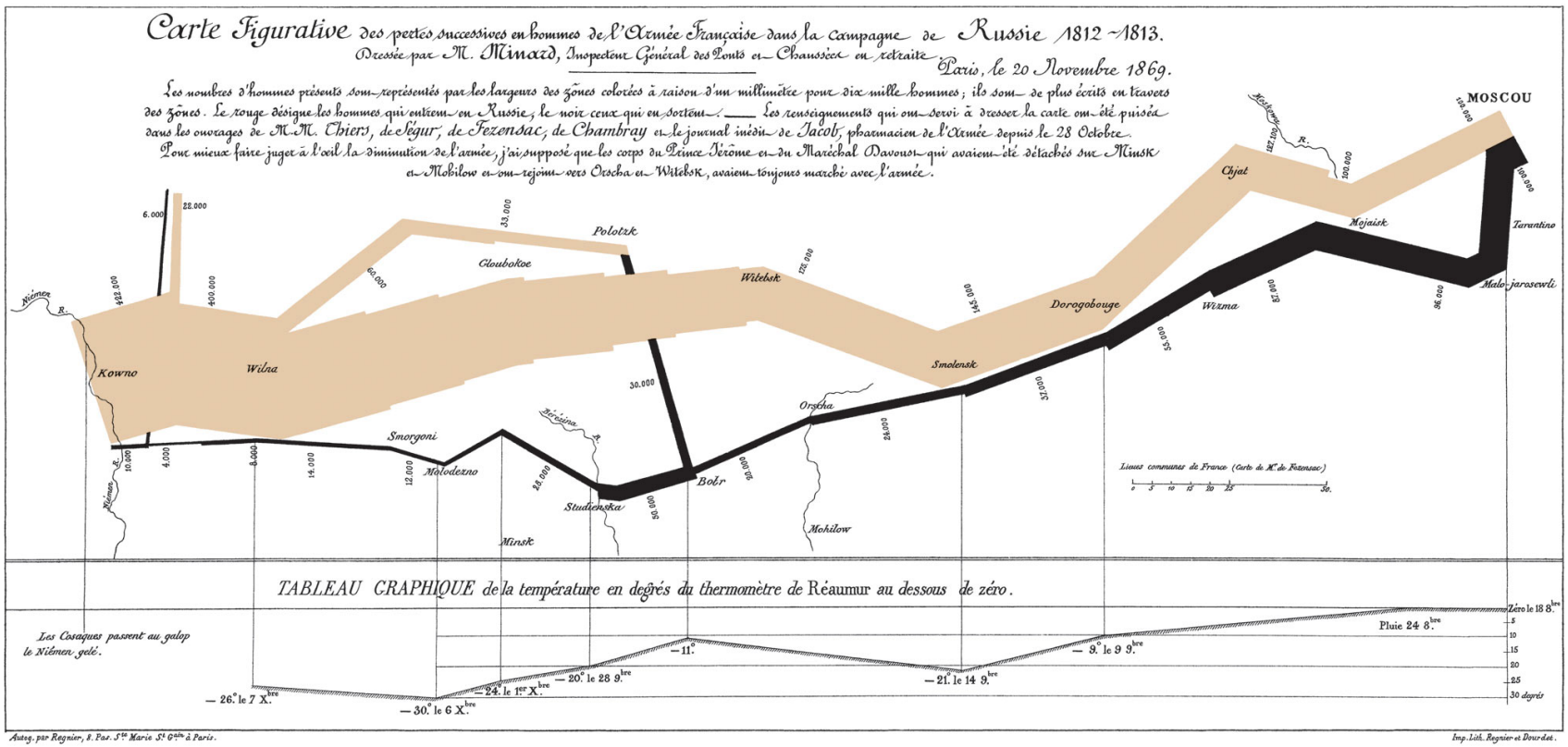
2. Associations between variables
   - The relationship between two or more variables
   - Differences between groups

G





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.



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

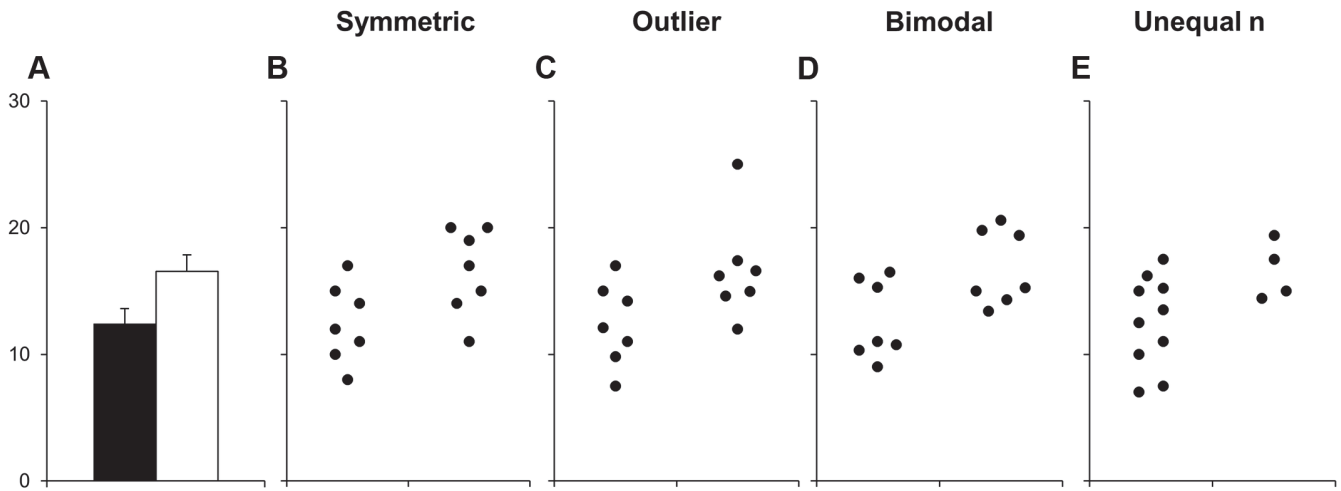
“Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space” - Tufte (1983)

Useful principles 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. *"Above all else show the data"* - Tufte (1983)

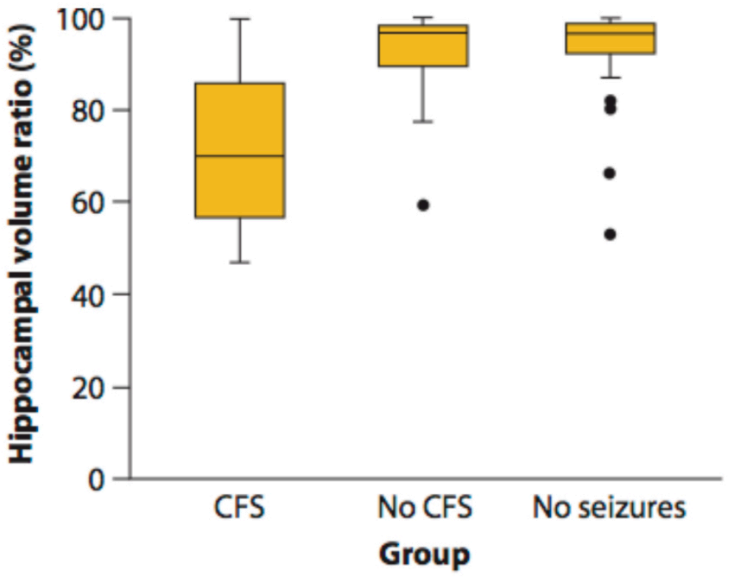
   For example, many different data values can generate the same mean and standard error. A strip chart reveals the pattern, whereas the bar graph hides it.

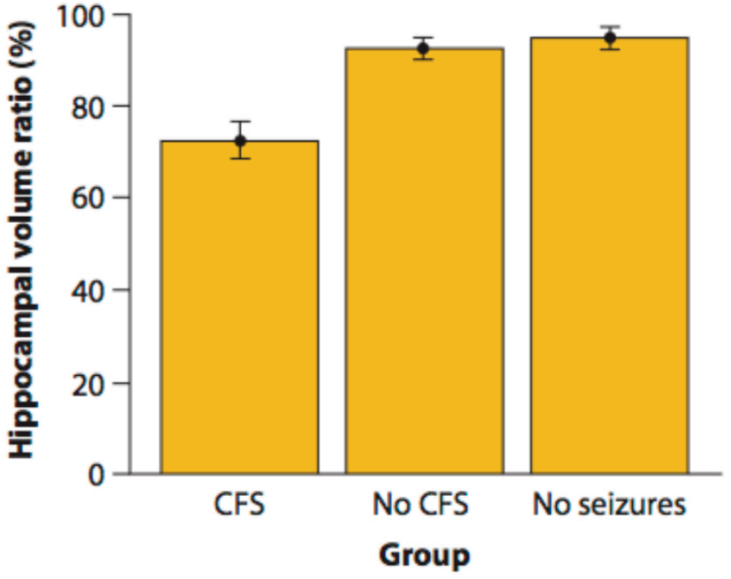


Weissgerber et al. (2015) Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biol. DOI:10.1371/journal.pbio.1002128

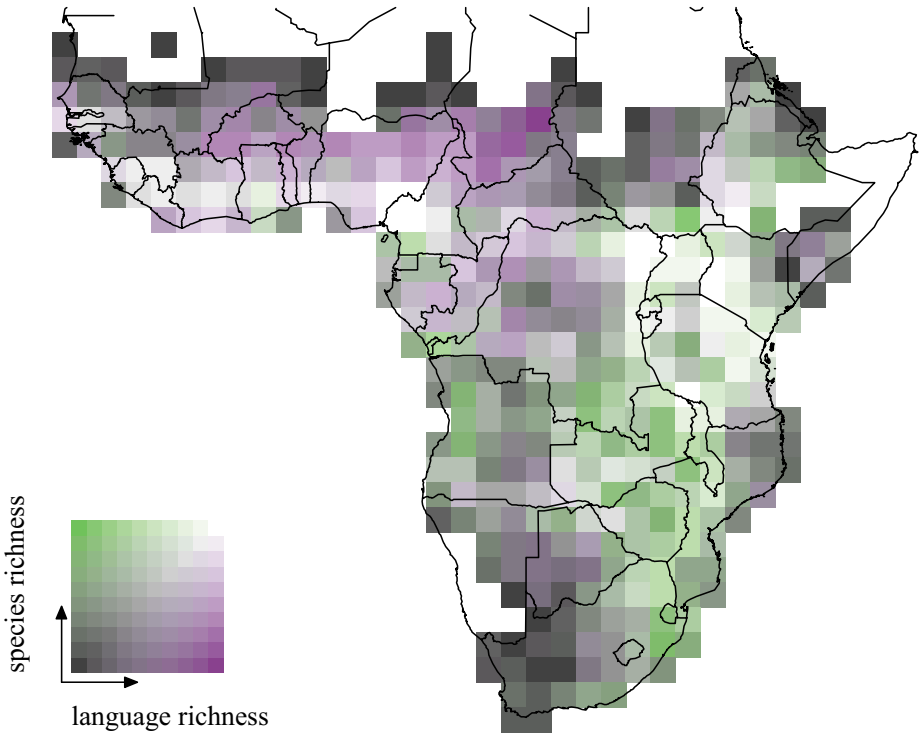
Why show the data?

Which graph is more effective? Why?





2. Make patterns in the data easy to see.
   Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency" - Tufte (1983)

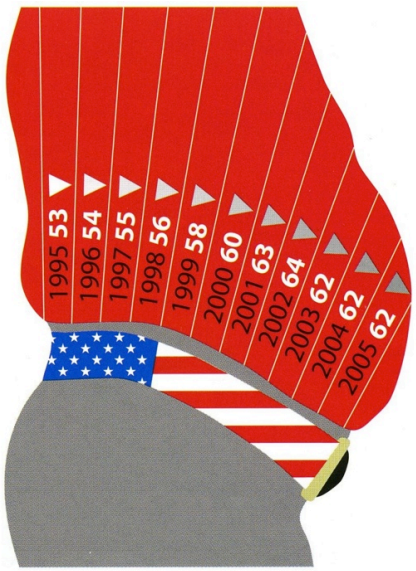


- Map displays the number of bird species and the number of distinct human languages present in each square of a grid of continental Africa. Reproduced from Moore et al. (2002).

- 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?

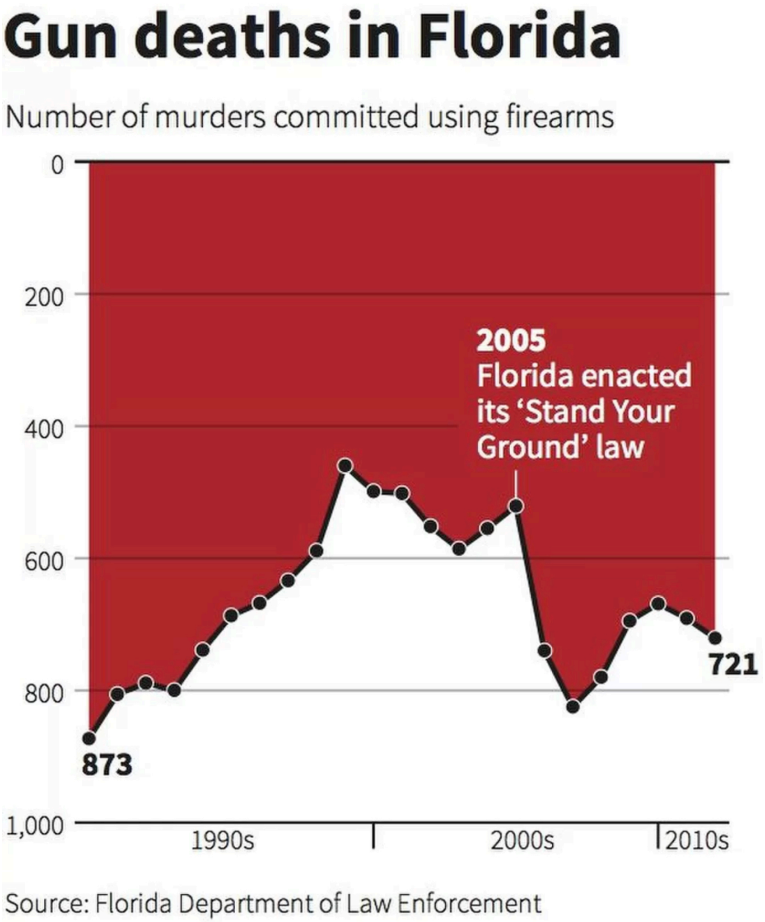
3. Draw graphical elements clearly, minimizing clutter
   Maximize the data-ink ratio, within reason - Tufte (1983)



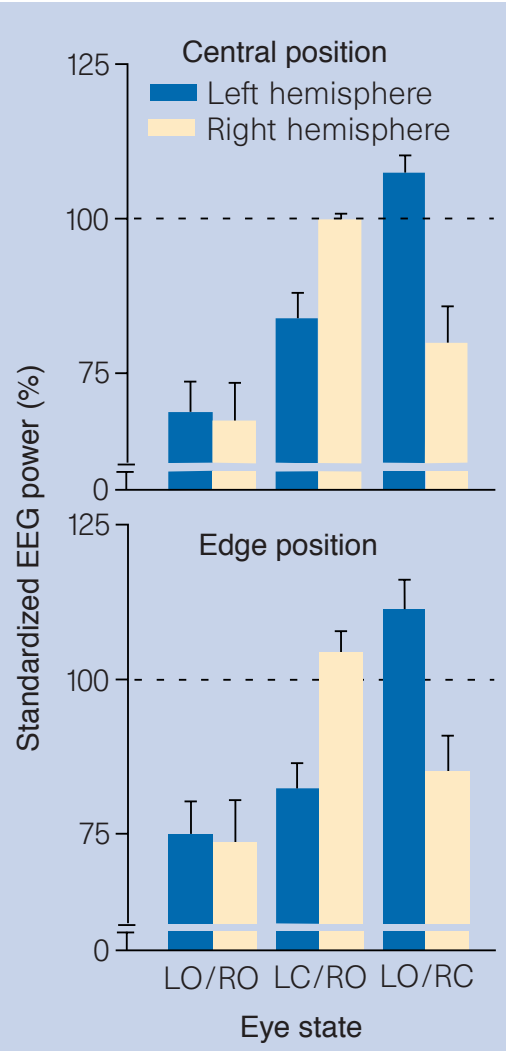
The percentage of adults over 18 with a "body mass index" greater than 25 in different years (The Economist 2006). Body mass index is a measure of weight relative to height.

- 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)



4. Represent magnitudes honestly
   A graphic does not distort if the visual representation of the data is consistent with the numerical representation - Tufte (1983)

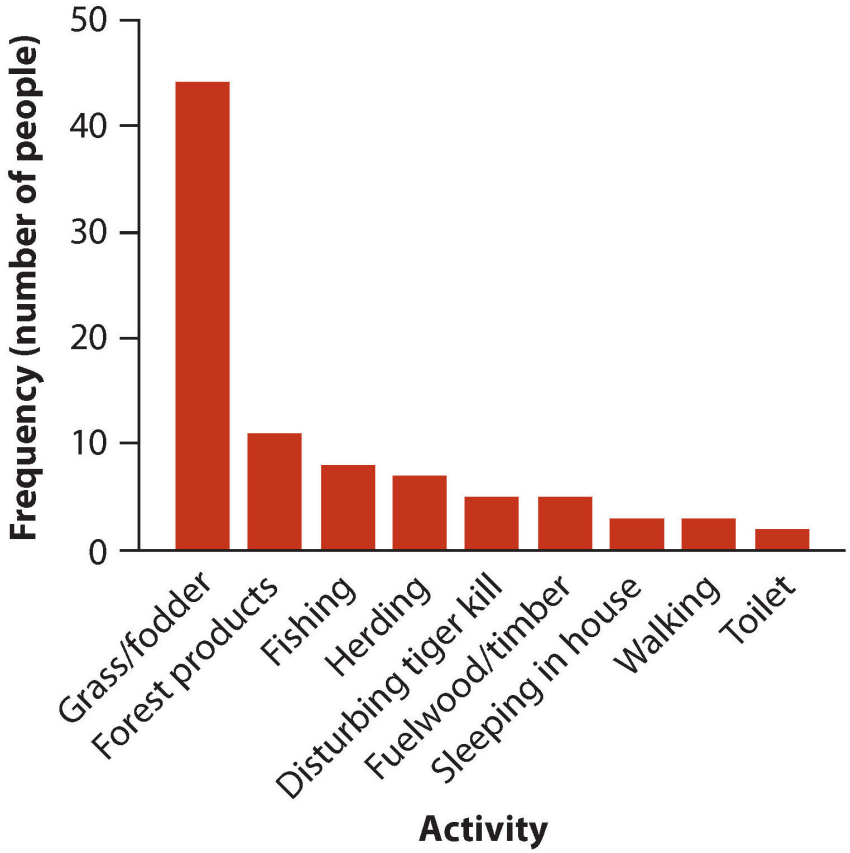


Slow wave sleep in the brain hemispheres of mallard ducks sleeping with one eye open. From Rattenborg et al. (1999) *Nature*.

- Are the bars "consistent with the numerical representation"?
- Is 0 a reasonable baseline for evaluating sleep score? Are there other issues with the graph? In the caption, they say: "EEG power was standardized as a percentage of the average power observed during bihemispheric slow-wave sleep for each bird's hemisphere, so the broken line at 100% indicates EEG power equivalent to that during bihemispheric slow-wave sleep."

Categorical frequencies: 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)

Categorical frequencies: Bar graph

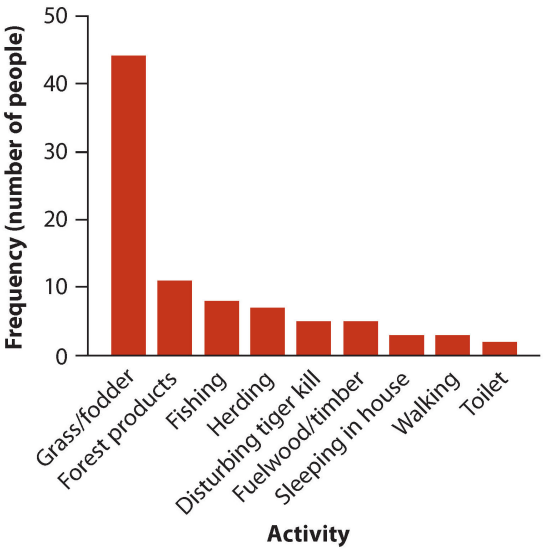
barplot.R

<functions>

Help search

```r
## create data-frame to plot
dd <- data.frame(animal=c('Zebra', 'Lion', 'Giraffe', 'Horse', 'Emu', 'Elephant', 'Hyena'),
                 count=c(102, 30, 77, 55, 62, 11, 29))
## barplot
barplot(height=dd$count, names.arg=dd$animal, col='red', las=1, ylab='Frequency')

## re-ordered by abundance
dd <- dd[order(dd$count, decreasing=TRUE),]
barplot(height=dd$count, names.arg=dd$animal, col='red', las=1, ylab='Frequency')
```

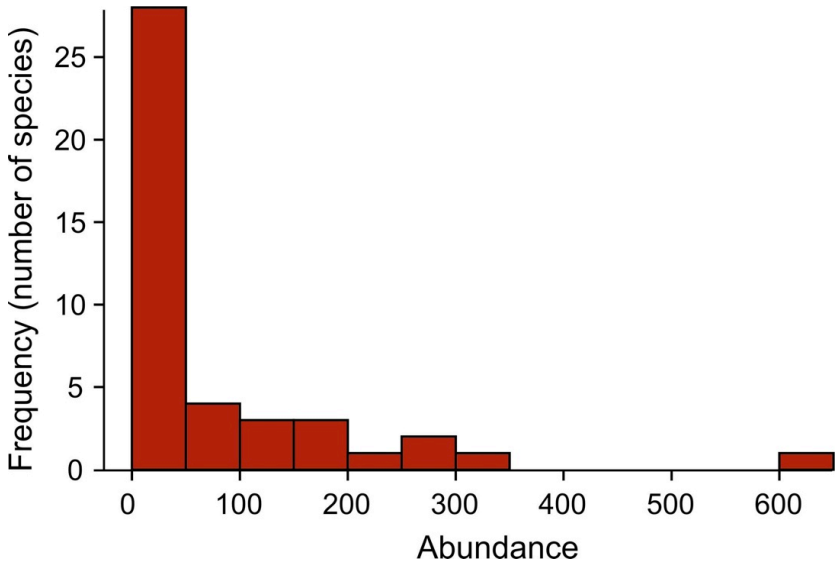
Categorical frequencies: Bar graph vs Pie Chart



Which is more successful?

Frequency distribution for a numeric variable: Histogram

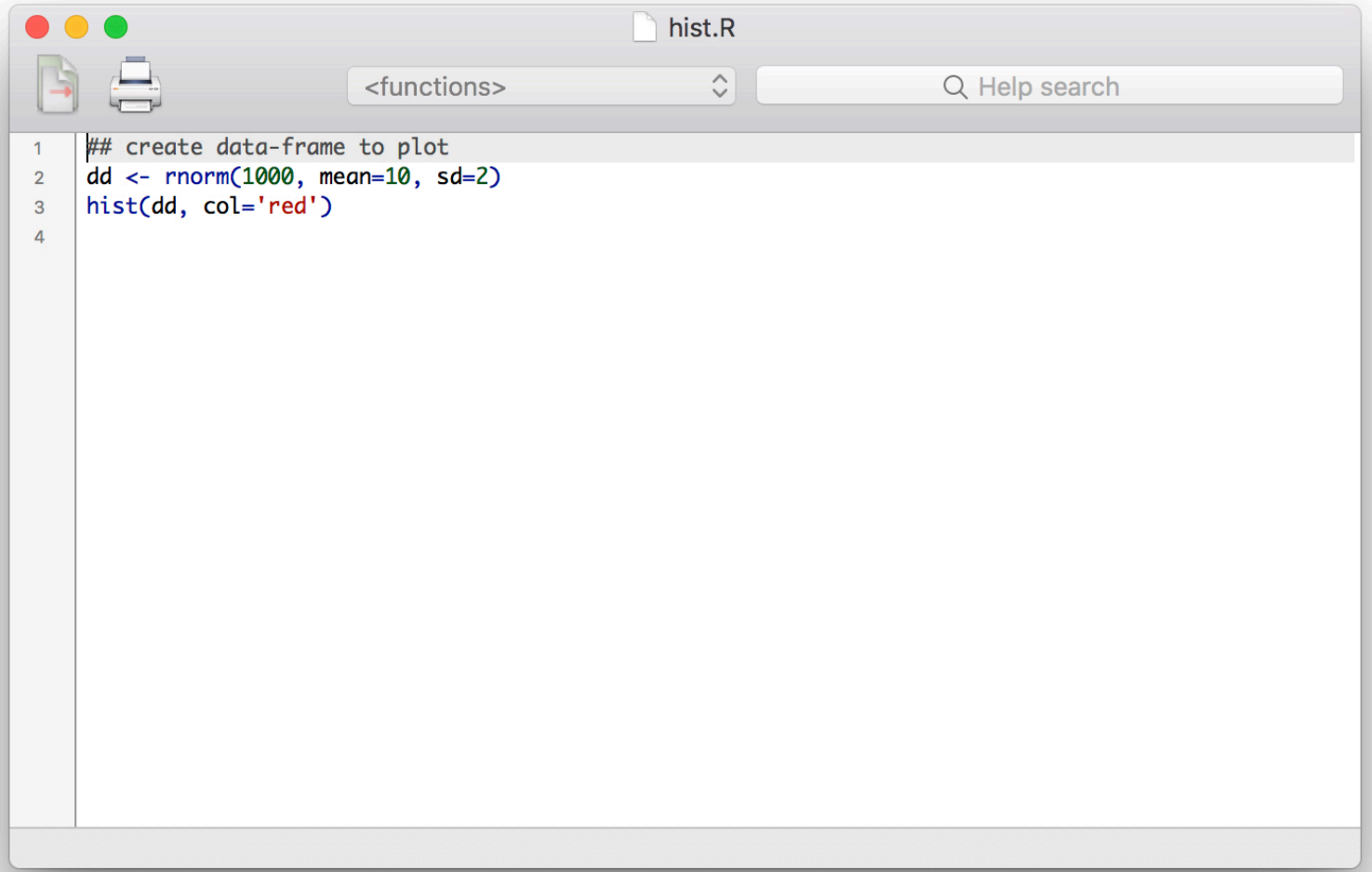
The frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. $n = 43$ species.



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

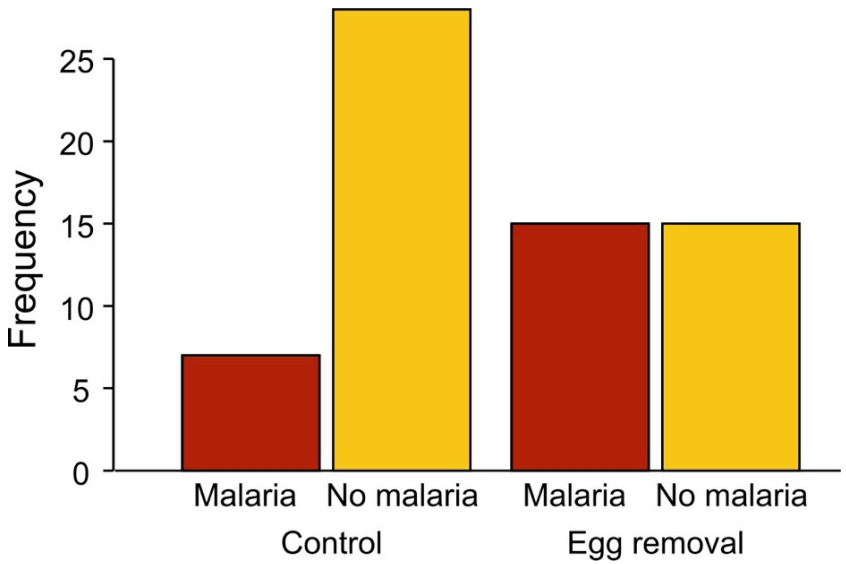
- Zero baseline
- No spaces between bars
- Choice of number of bins and bin width

Frequency distribution for a numeric variable: Histogram



Association between categorical variables: Grouped bar graph

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



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

Association between categorical variables: Grouped bar graph

grouped-barplot.R

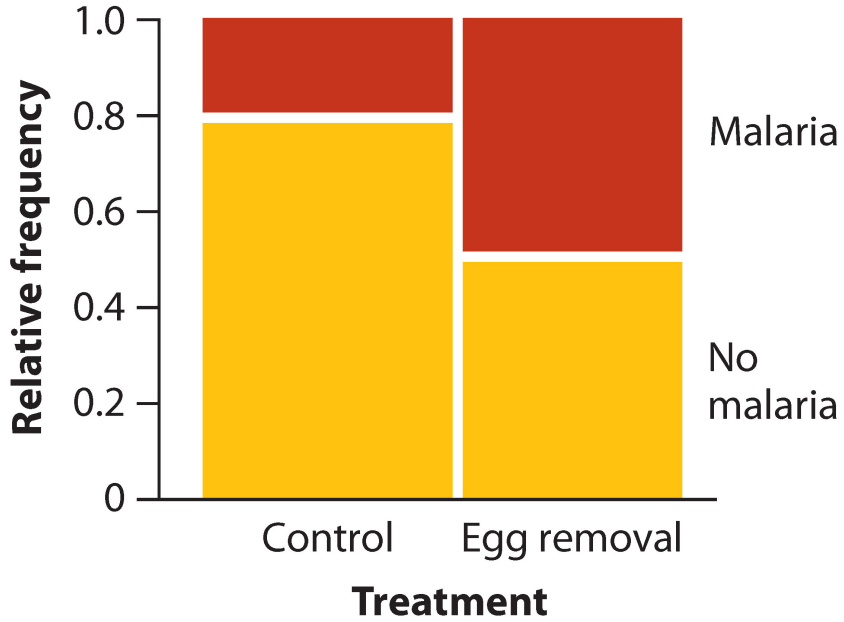
<functions>

Help search

```r
## R has many built in data-sets that are useful for examples.
##
## E.g., 1997 Motor Trends cars:
mtcars

## make data-set for grouped barplot
counts <- table(mtcars$vs, mtcars$gear)
rownames(counts) <- c('V', 'Straight')

barplot(counts,
        beside=TRUE,
        main='Car Distribution by Gears and Engine Type',
        xlab='Number of Gears',
        col=c('darkblue','red'),
        legend=rownames(counts),
        las=1)

## without the 'beside argument'
barplot(counts,
        main='Car Distribution by Gears and Engine Type',
        xlab='Number of Gears',
        col=c('darkblue','red'),
        legend=rownames(counts),
        las=1)

```

Association between categorical variables: Mosaic plot

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



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

Association between categorical variables: Mosaic plot

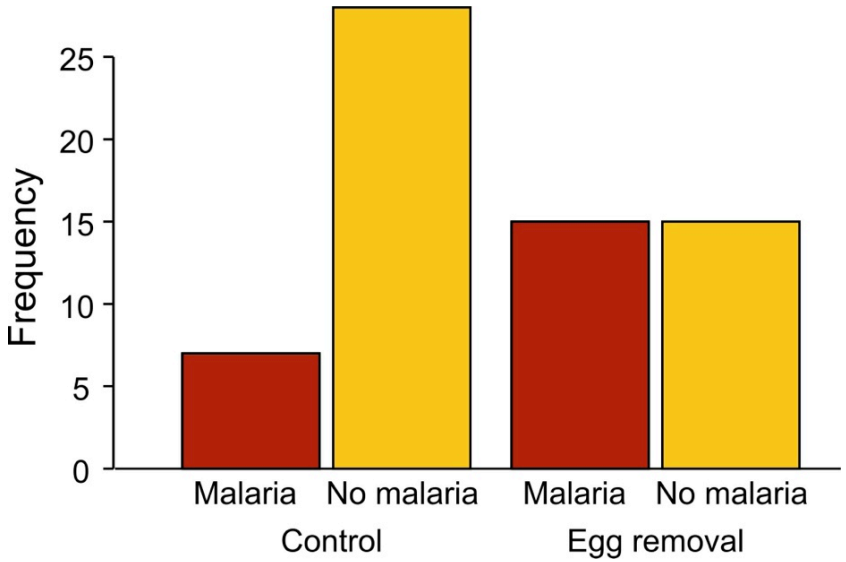
mosaic-plot.R

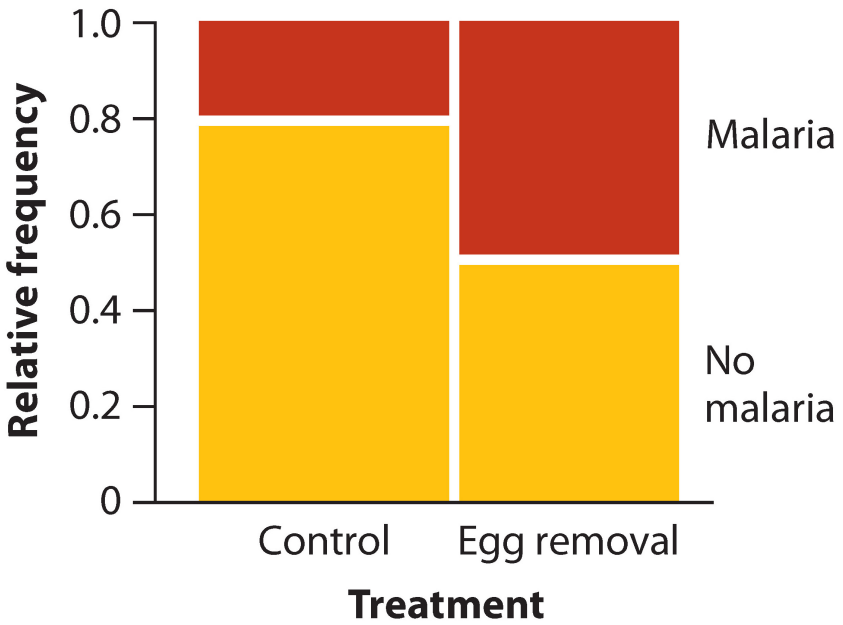
<functions>

Help search

```
1  mtcars
2
3  ## same data-set as for a grouped barplot
4  counts <- table(mtcars$vs, mtcars$gear)
5  rownames(counts) <- c('V', 'Straight')
6
7  mosaicplot(counts, main='Car Distribution by Gears and Engine Type',
8             color=c('blue', 'red', 'green4'),
9             off=c(20, 0), ## determines spacing between colours and bars
10            las=1,
11            cex.axis=1.2)
12
```

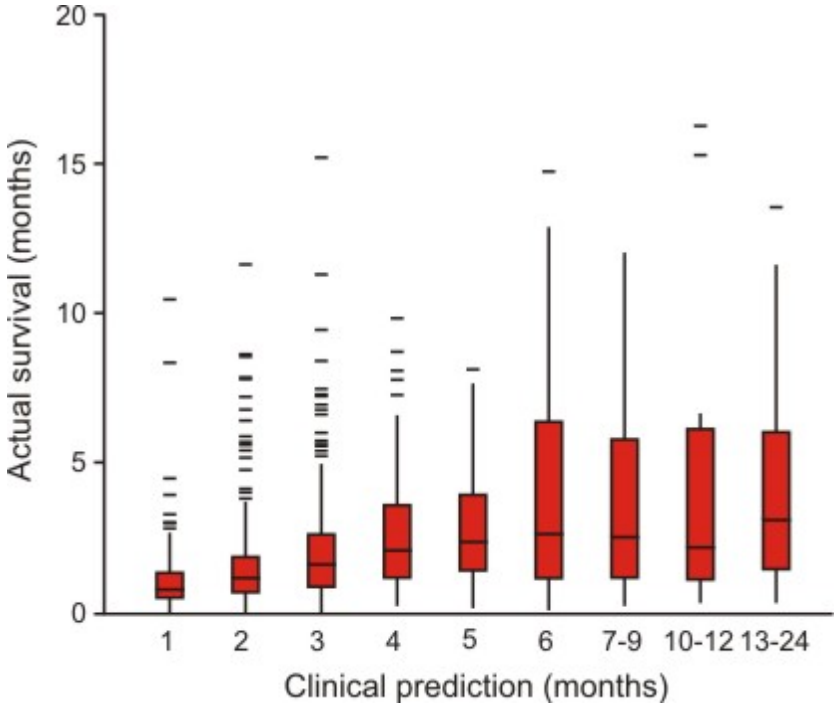
Which is more successful?





Association between numerical and categorical variable: Box plot

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



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)

Association between numerical and categorical variable: Box plot

box-plot.R

<functions>

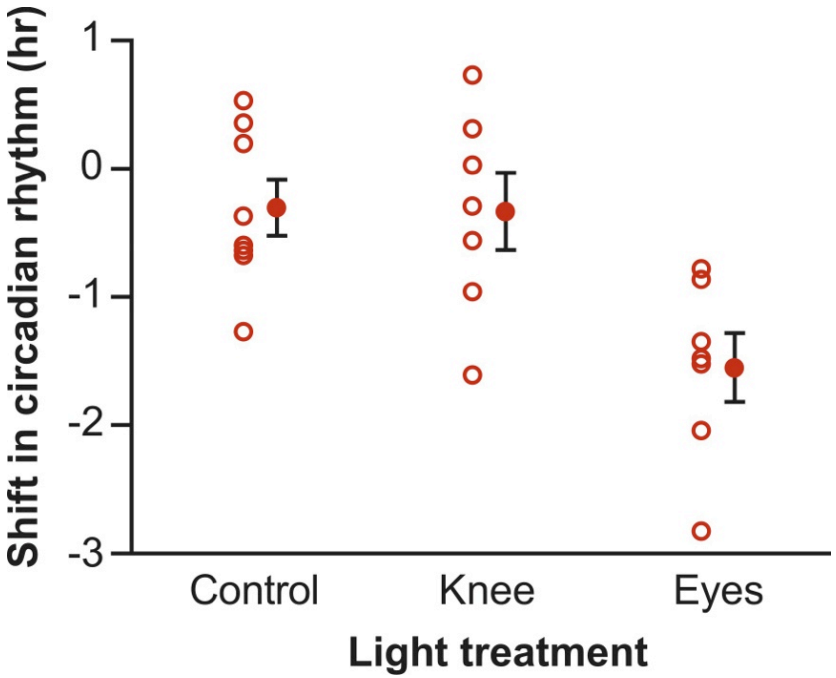
Help search

```
1  boxplot(mpg~cyl,
2          data=mtcars,
3          main='Car Milage Data',
4          xlab='Number of Cylinders',
5          ylab='Miles Per Gallon',
6          col='red',
7          las=1)
8
```

boxplot(x, ...)

Association between numerical and categorical variable: Strip chart

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



Displays differences between groups

- Shows the data points
- Points fill the space available
- Non-zero baseline often ok (goal is to show differences)

Association between numerical and categorical variable: Strip chart

strip-chart.R

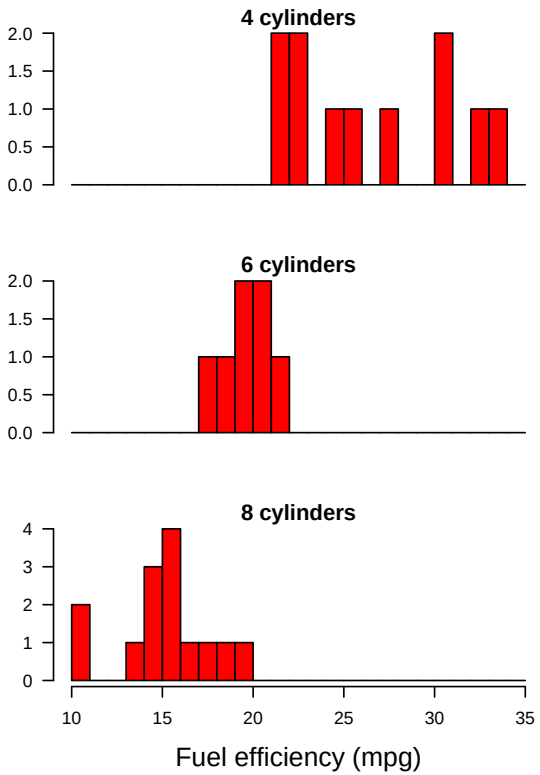
<functions>

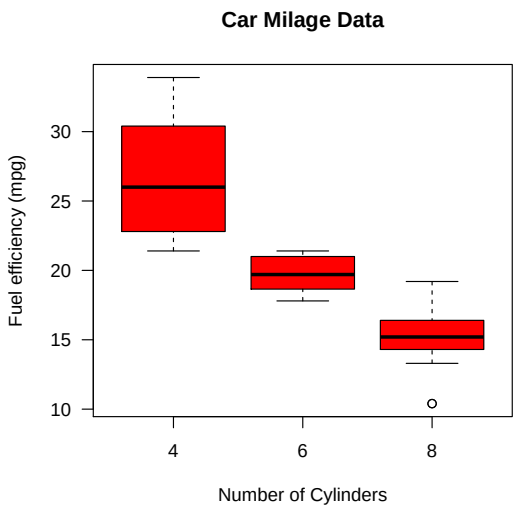
Help search

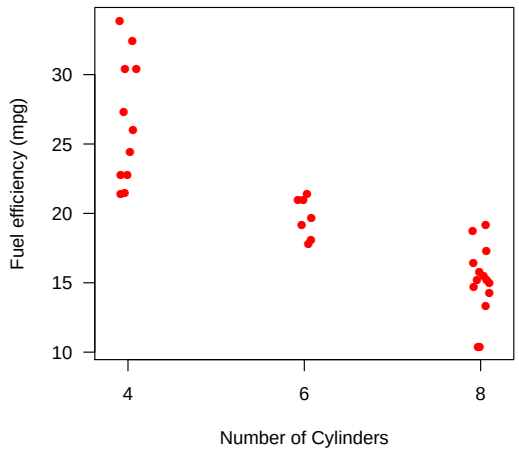
```r
stripchart(mpg ~ cyl,
           data=mtcars,
           xlab='Engine size (cylinders)',
           ylab='Fuel efficiency (mpg)',
           pch=16, las=1, vertical=TRUE)

## can add some "jitter" to the points, so that they are easier to see
stripchart(mpg ~ cyl,
           method='jitter',
           jitter=0.05, ## amount of jitter
           data=mtcars,
           xlab='Engine size (cylinders)',
           ylab='Fuel efficiency (mpg)',
           pch=16, las=1, vertical=TRUE)
```

Association between numerical and categorical variable

Multiple histograms, box plot, strip chart... Which is more successful?



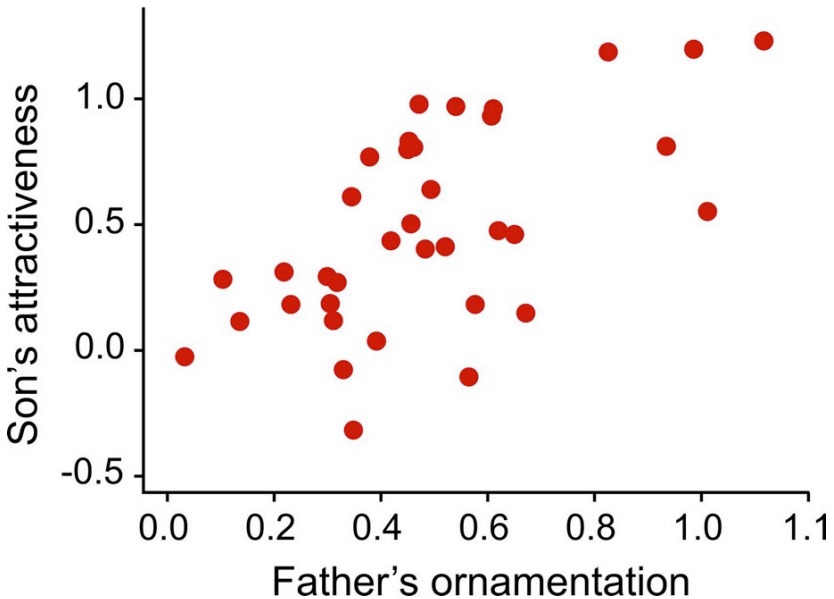




Note: Stacking histograms vertically makes it easier to compare distributions than presenting them in a row.

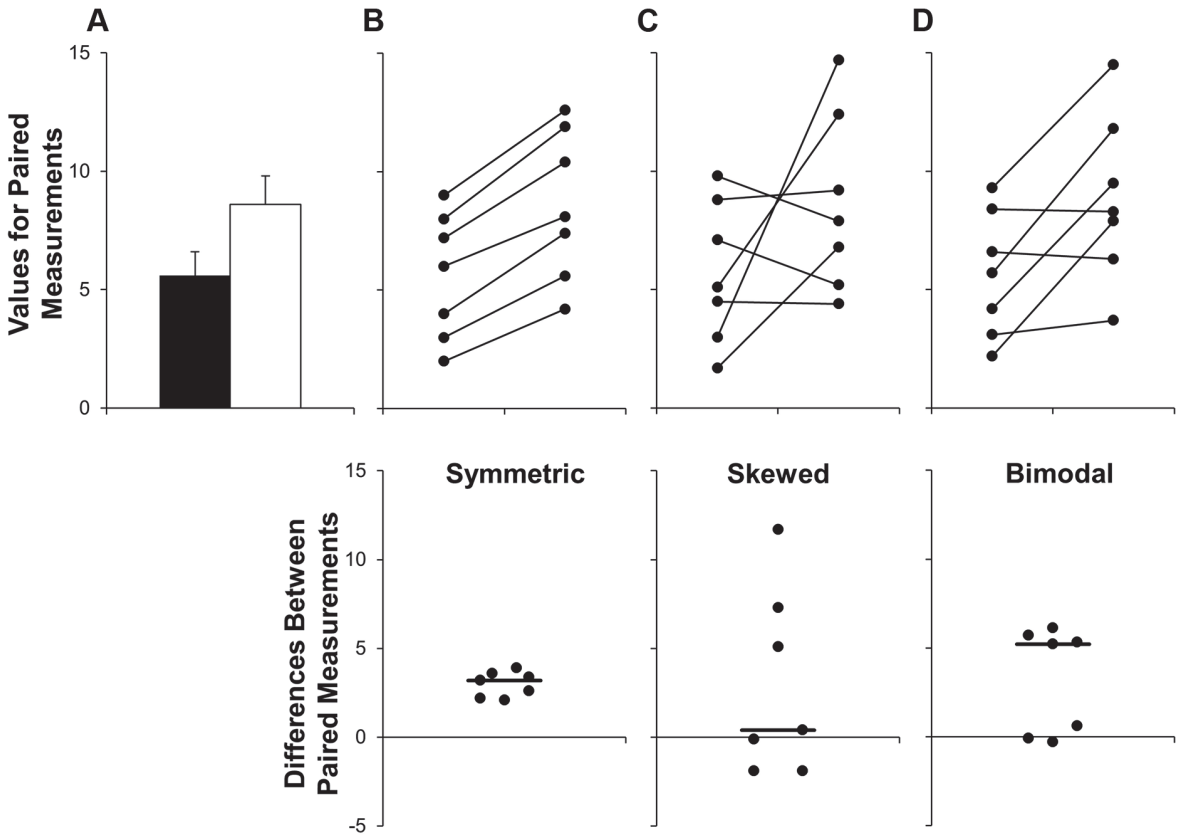
Association between two numerical variables: Scatter plot

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



- Points should fill the space available
- Non-zero baseline often ok (goal is to show association)

Paired data: Connect the dots to show pairing, or plot differences

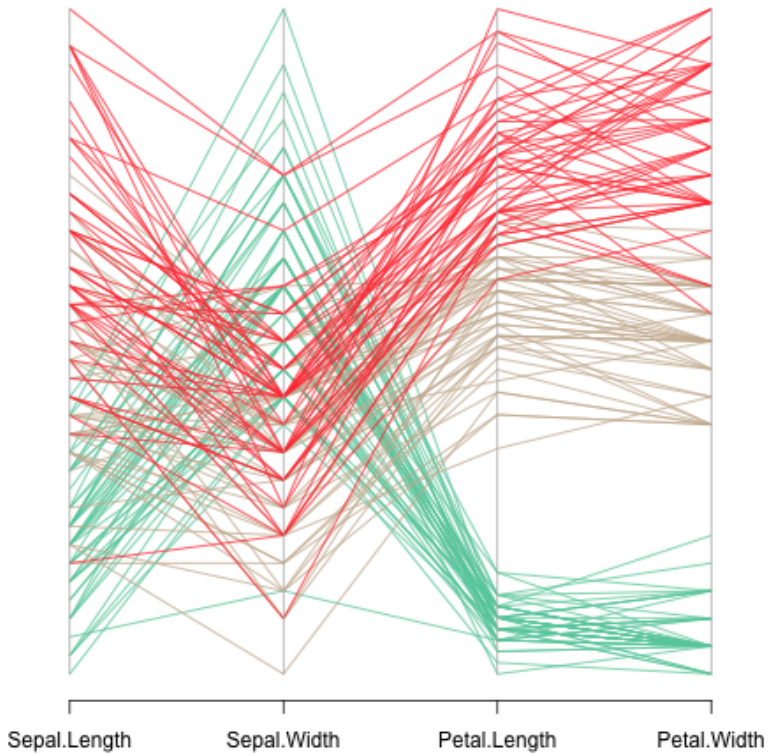


Interaction plots

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

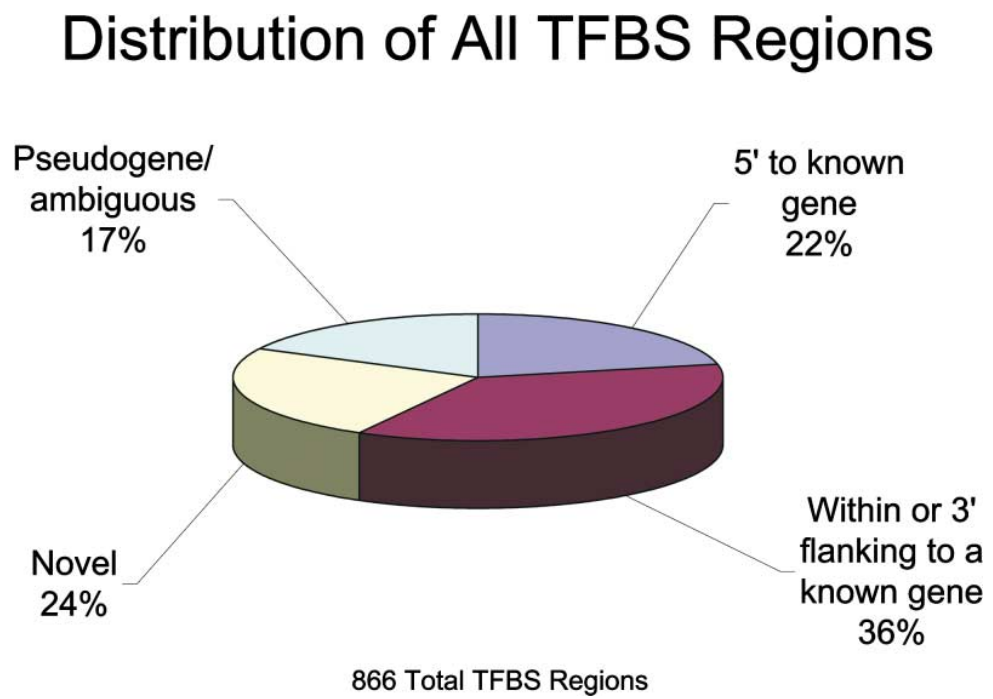
Grouped data (like paired, but >2 measurements)



http://www.r-graph-gallery.com/93-parrallel-plot/

Category frequencies: 3D graphs

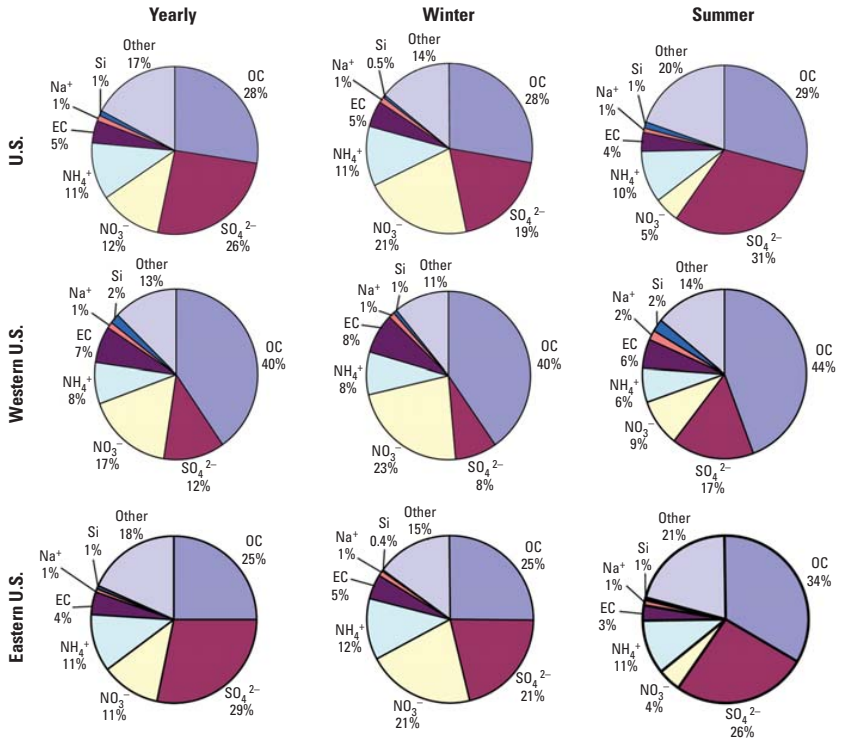
In addition, a graph that is meaningful only with numbers added is necessarily a failure.



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.

Category frequencies: Lots of 2D pies

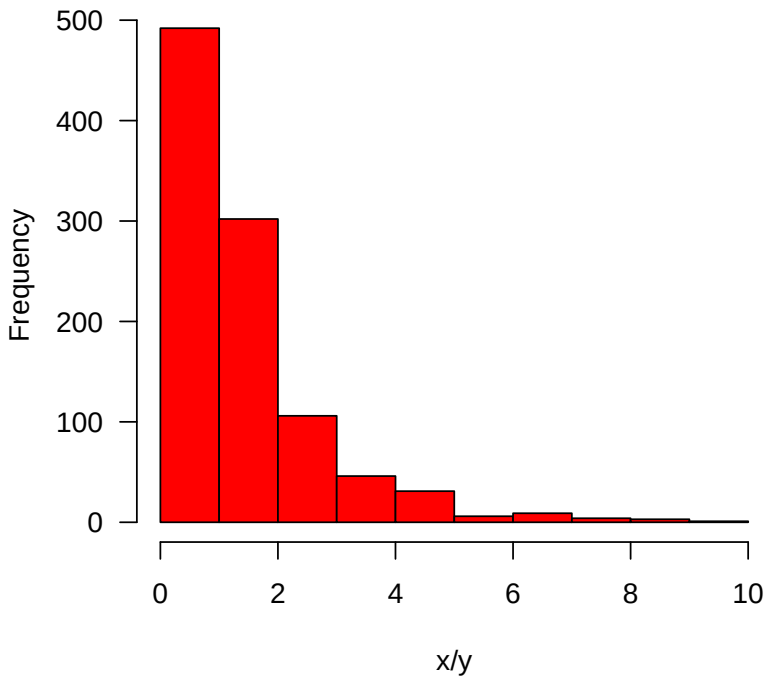
- The aim is to show changes in NO_3 and SO_4 between winter and summer, and consistency of change between geographic regions.
- This is not easy to see... ("Huh?" not "Oh!")
- Design a graph to show the change from summer to winter in NO_3 and SO_4, rather than try to display everything..



Bell ML, et al. (2007) Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies. Environmental Health Perspectives 115:989-995.

Ratio data

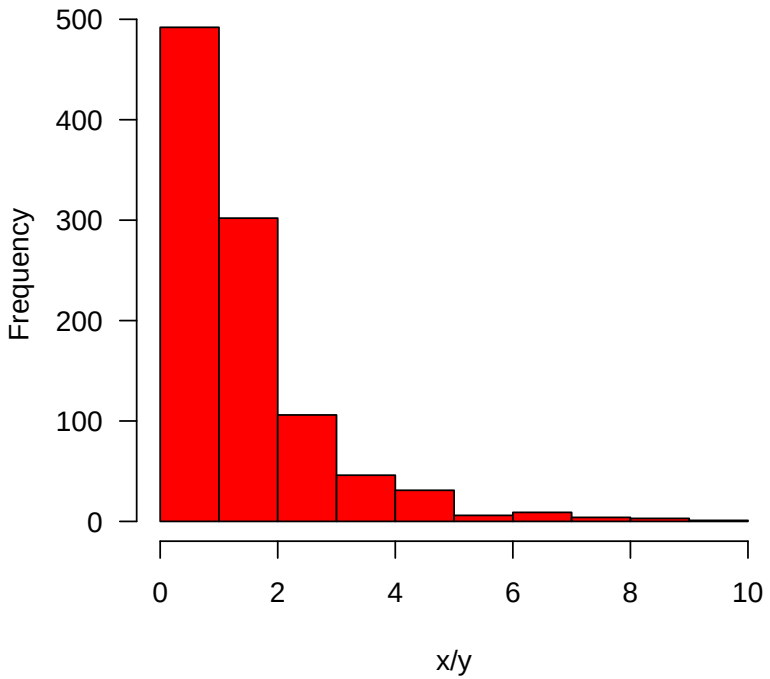
Generate two equivalent sets of 1000 random numbers (call them x and y). Then plot a histogram of the ratio x/y.

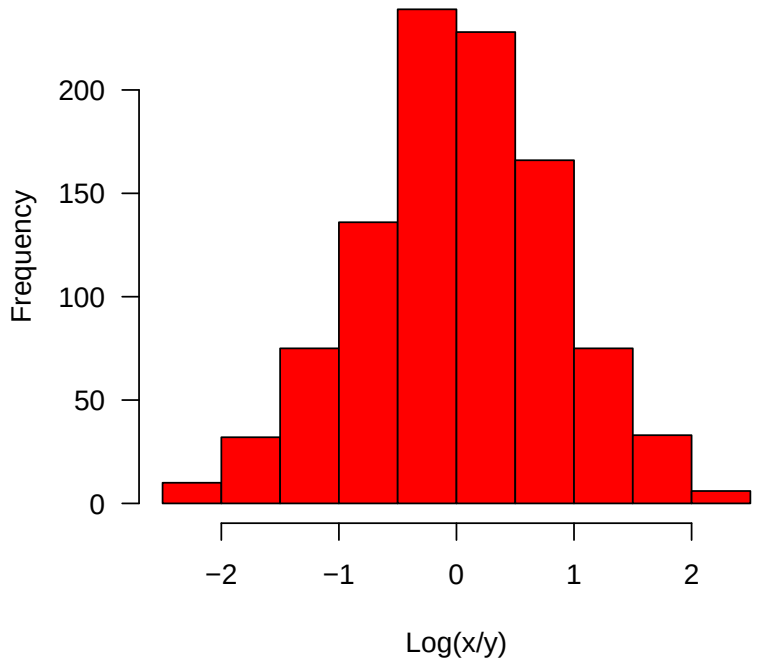


Any problems with this?

Ratio data

An alternative approach is to plot the Log of the ratio.





What is their purpose?

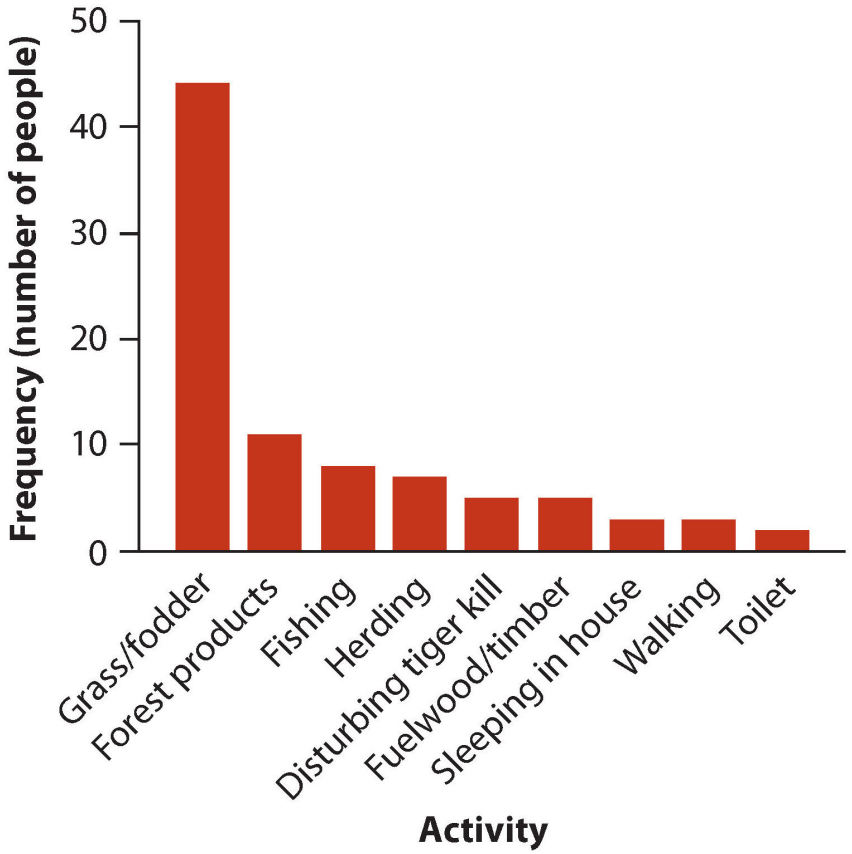
Tables in main text should also be used to illuminate patterns. Make your tables so that they cause the viewer to go "Oh!" and not "Huh?".

- 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.

Put tables for storing numbers into online Appendix or Supplement.

Frequency tables can also be used to display category frequencies

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

Which do you prefer?

Difficult to see a relationship between F and survival.
Uneven line spacing, the gaps break up patterns.
Too much empty space.
Too many decimals.

Improving tables

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

King/Queen	F	Pregnan-cies	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

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