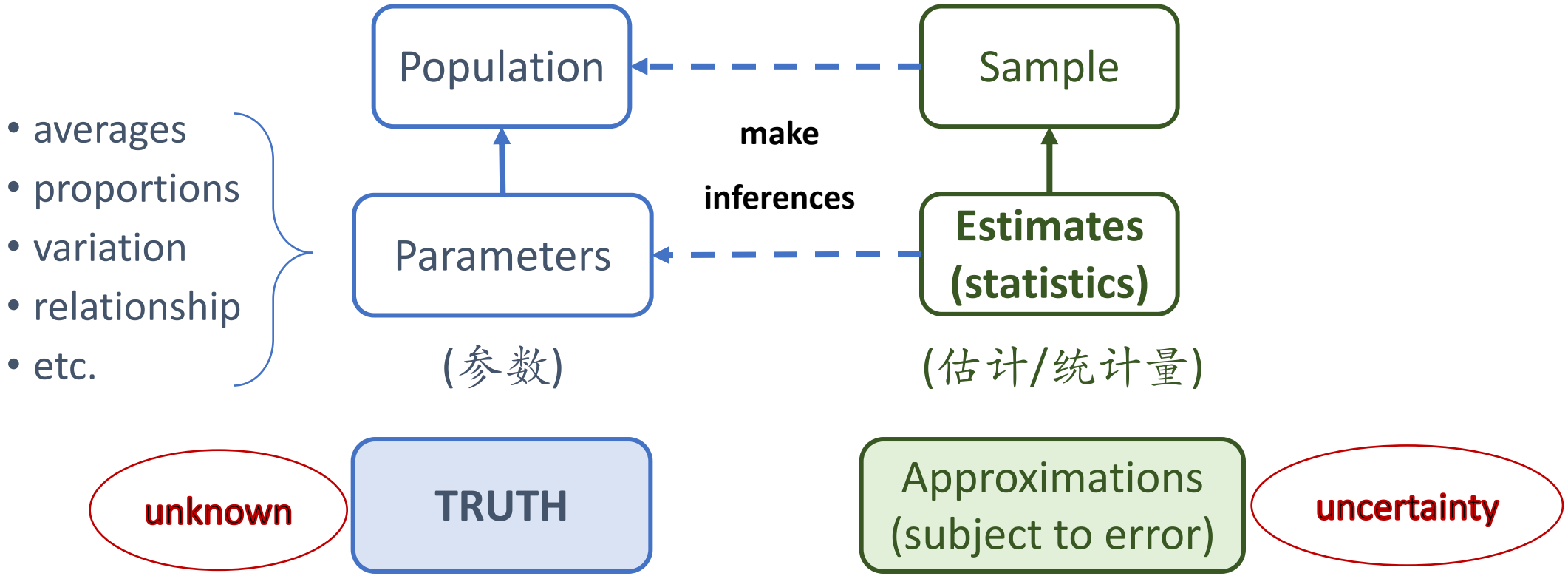
# Lecture 2 – Data Display (数据展示)

- Outline for today
  - Recall – Lecture 1
  - Data display – Graphics
    - Principles of effective display
    - Types of graphs to show patterns in data
  - Table
  - Summary
  - Discussion
  - R lab

# 1. Recall - What is statistics?

- About estimation of population (总体), with sample data (样本).



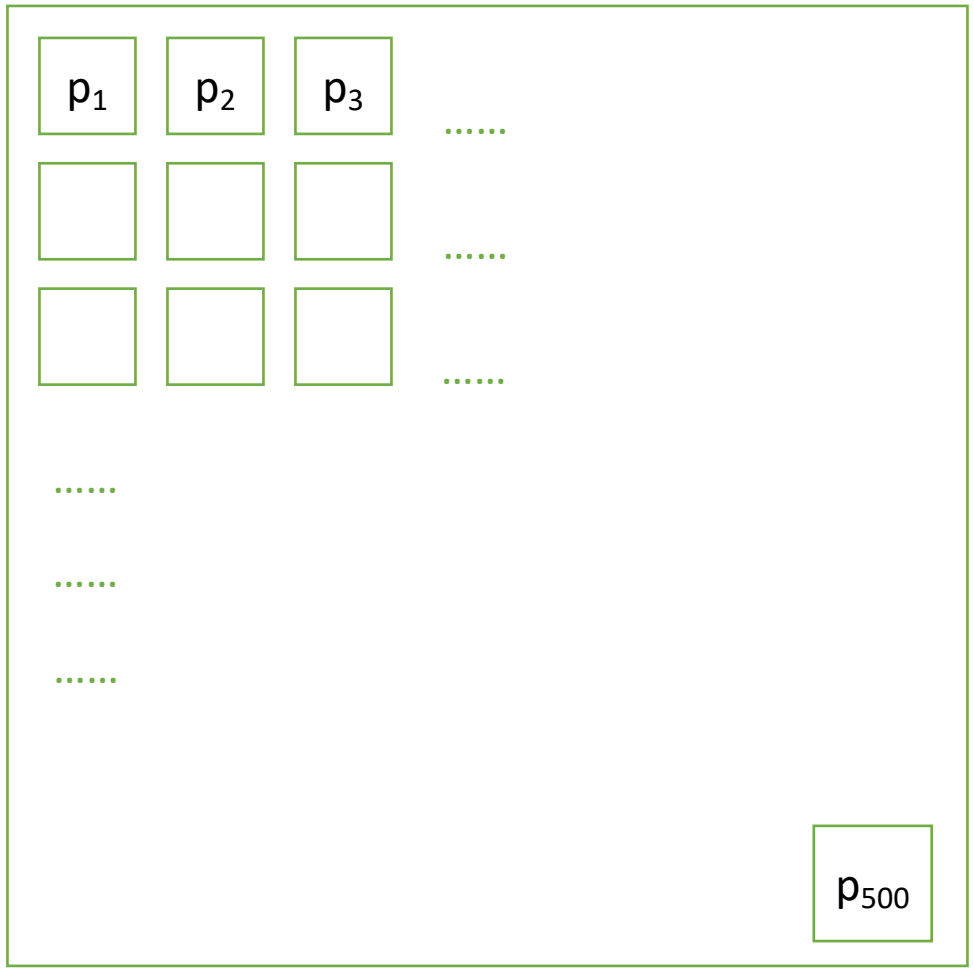
# 1. Recall - Types of data and variables

- Data type
  - Categorical variables (分类/类型变量)
    - Qualitative characteristics: describing membership in a category or group
    - Nominal (定性变量): variables have no inherent order;
    - Ordinal (定序变量) : variables can be ordered (but with unknown magnitude);
  - Numerical variables (数值变量)
    - Quantitative measurements that have magnitude on a numerical scale
    - Continuous (连续变量): can take on any real-number value within some range
    - Discrete (离散变量): come in indivisible units
- How to display, describe, and analyze your data?

# 1. Recall - Discussions

- 2. The average age of piñon/pinyon pine trees in the coast ranges of California was investigated by placing 500 10-hectare plots randomly on a distribution map of the species using a computer. Researchers then found the location of each random plot in the field, and they measured the age of every piñon pine tree within each of the 10-hectare plots. The average age within the plot was used as the unit measurement. These unit measurements were then used to estimate the average age of California piñon pines.
  - What is the population of interest in this study?
  - Why did the researchers take an average of the ages of trees within each plot as their unit measurement, rather than combine into a single sample the ages of all the trees from all the plots?

The average age of piñon/pinyon pine trees in the coast ranges of California
- investigated by placing 500 10-hectare plots




$$Age\_p1 = \frac{a_1+a_2+\dots a_{n\_p1}}{n\_p_1}$$

$$Age\_p2 = \frac{a_1+a_2+\dots a_{n\_p2}}{n\_p_2}$$

……

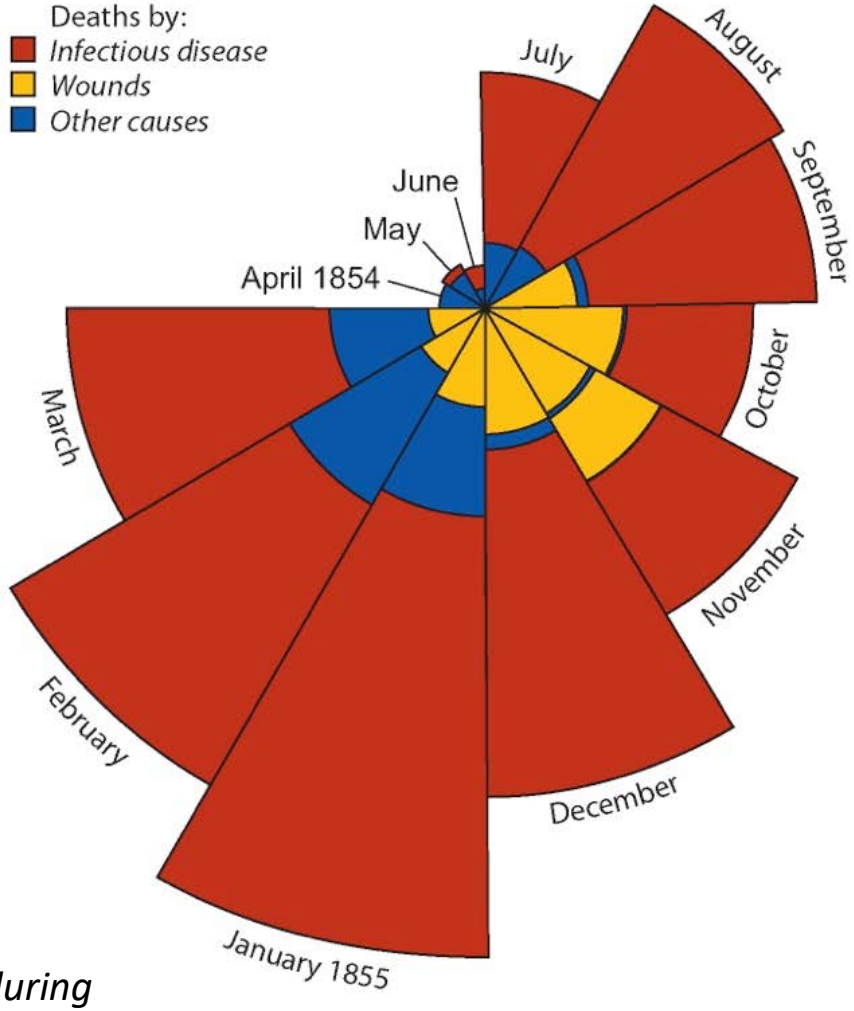
$$Age\_mean = \frac{Age\_p_1+Age\_p_2+\dots Age\_p_n}{500}$$

# 2. Display data (数据展示)

- Why make graphs?
- Principles of effective display
- Types of graphs to achieve these principles

# 2. Display data

- Graphics (图像/制图学)
  - Why make graphs?
    - 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)*

(Whitlock & Schluter 2020)

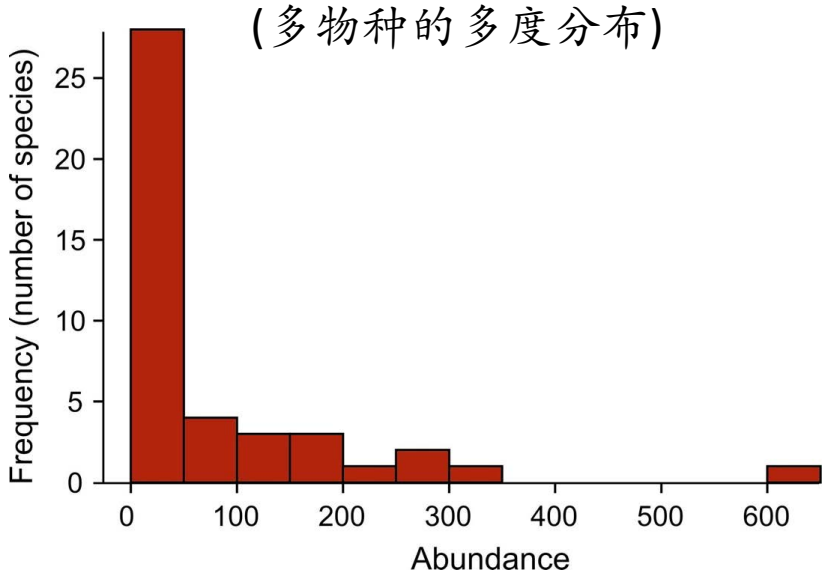
# 2. Display data

> *"...gives to the viewer the greatest number of ideas in the shortest time..."* – Tufte (1983)

- Why do we use graphs?
  - To reveal/show pattern in data.
  - The first step in any data analysis or statistical procedure is to graph the data and look at it, because analysis and presentation are largely coincident.
  - The top choices will depend on the type of data, numerical or categorical, and whether the goal is to show measurements of one variable or the association between two variables.

# 2. Display data

(多物种的多度分布)

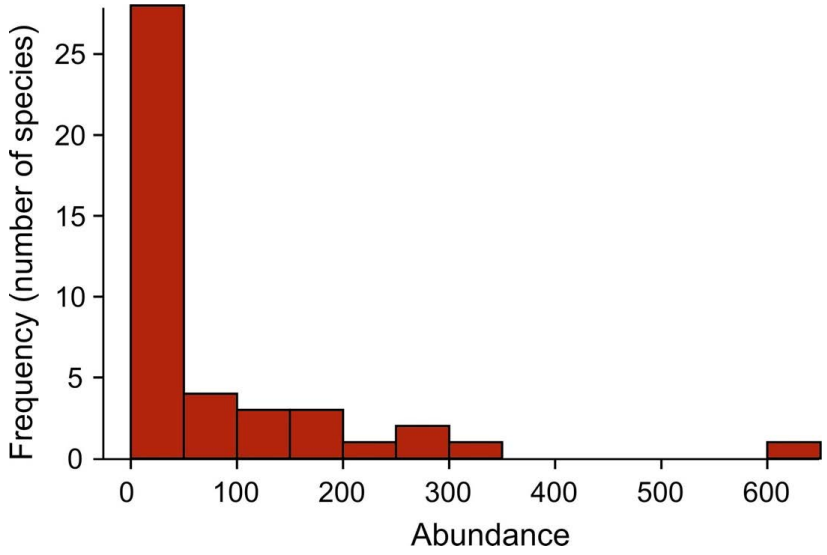


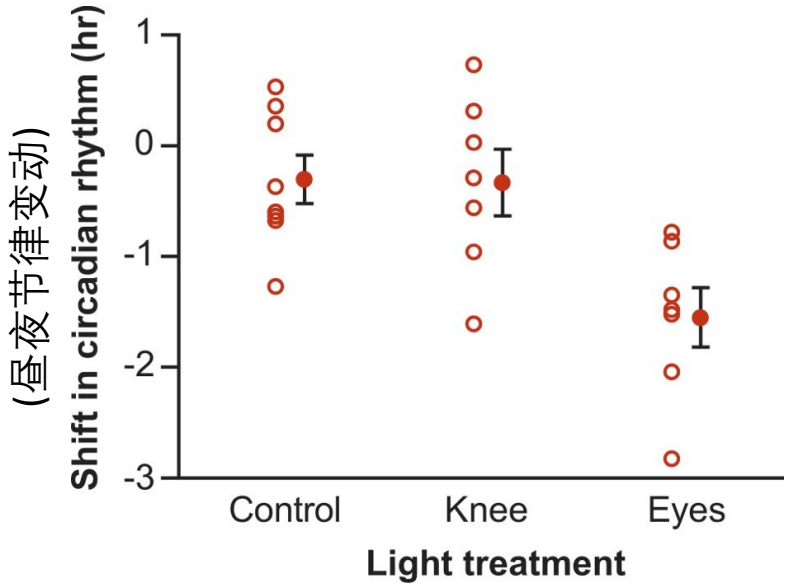

- Why do we use graphs?
  - To reveal/show pattern in data.
  - Frequency distributions
    - Frequency: the number of observations having a particular value of the measurement
    - The frequency distribution shows how often each value of the variable occurs in the sample.
    - Patterns: the location, spread, shape of distribution

(Whitlock & Schluter 2020)

# 2. Display data

- Why do we use graphs?
  - To reveal/show pattern in data.
  - Frequency distributions
    - The location, spread, shape of distribution
  - Associations between variables
    - The relationship between two or more variables
    - Differences between groups

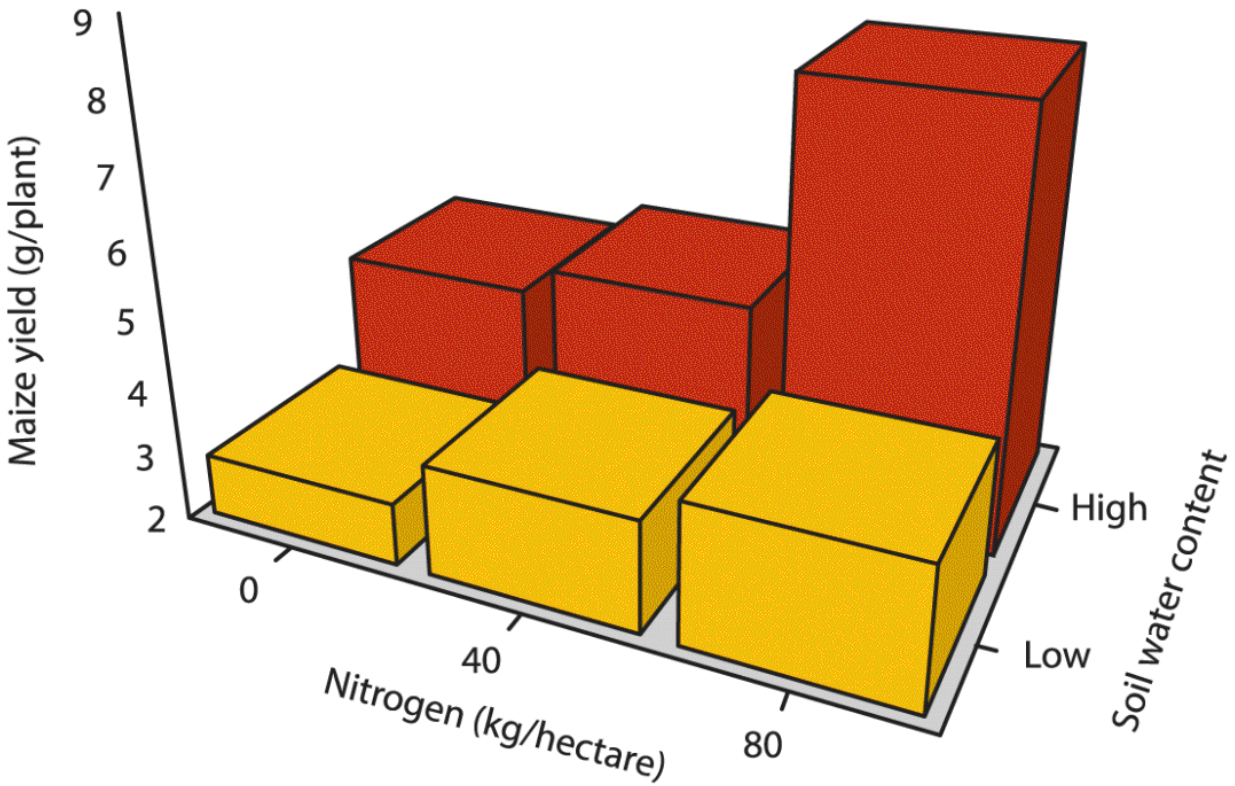



(Whitlock & Schluter 2020)

# 2. Display data

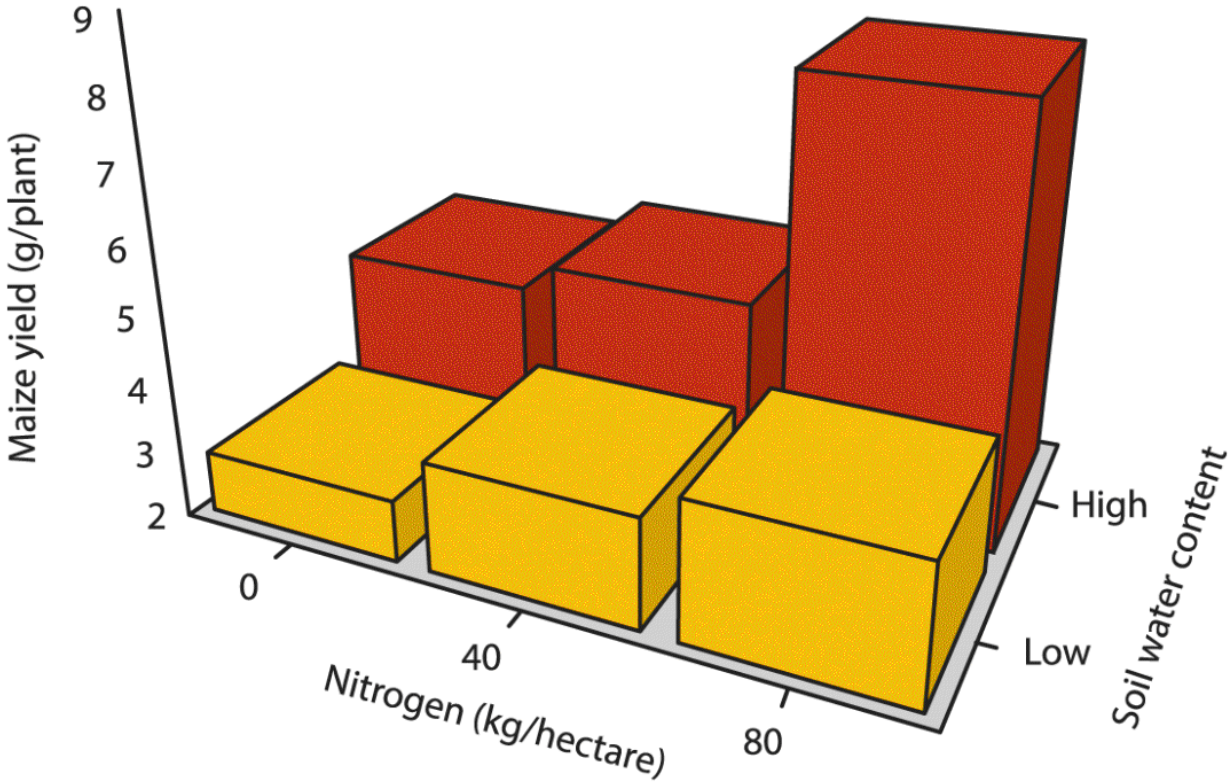
- A bad graph
  - The results of an experiment in which maize (玉米) plants were grown in pots under three nitrogen regimes and two soil water contents.
  - Height of bars represents the average maize yield.
- What mistakes can you tell?




Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

# 2. Display data

- A bad graph
- Mistakes
  - The graph hides the data.
  - Patterns in the data are difficult to see.
  - Magnitudes are distorted.
  - Graphical elements are unclear.

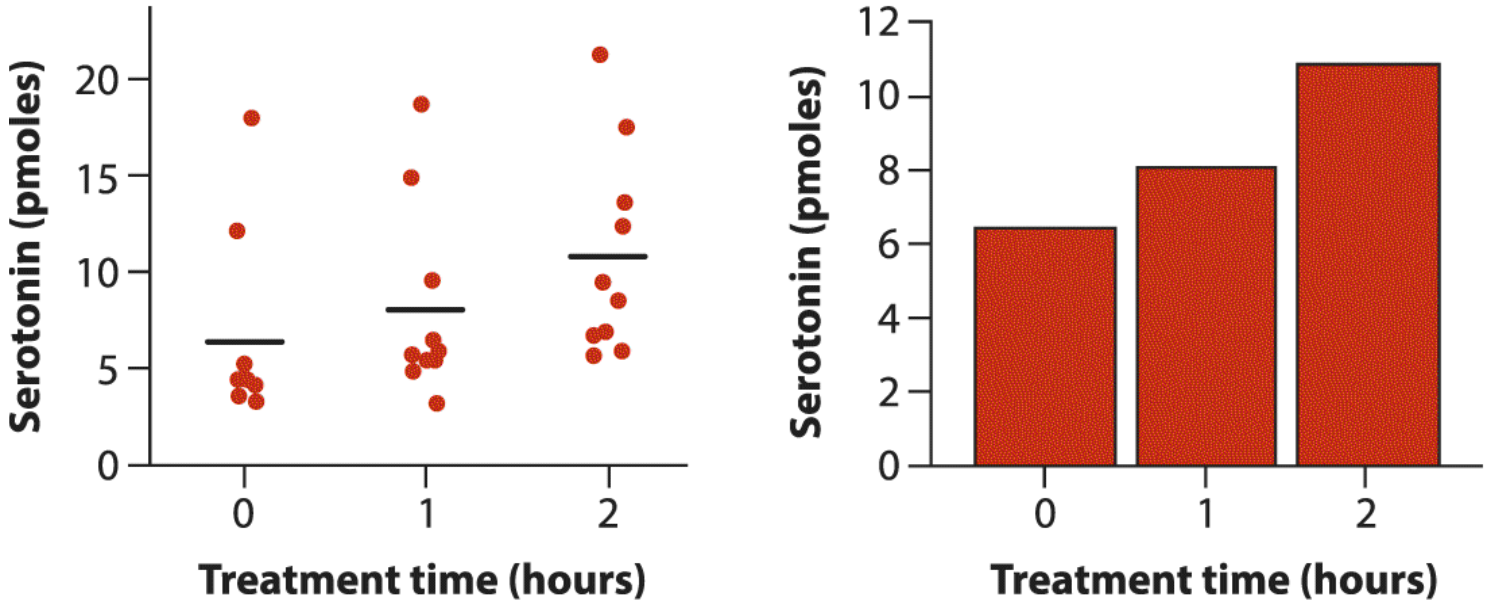



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

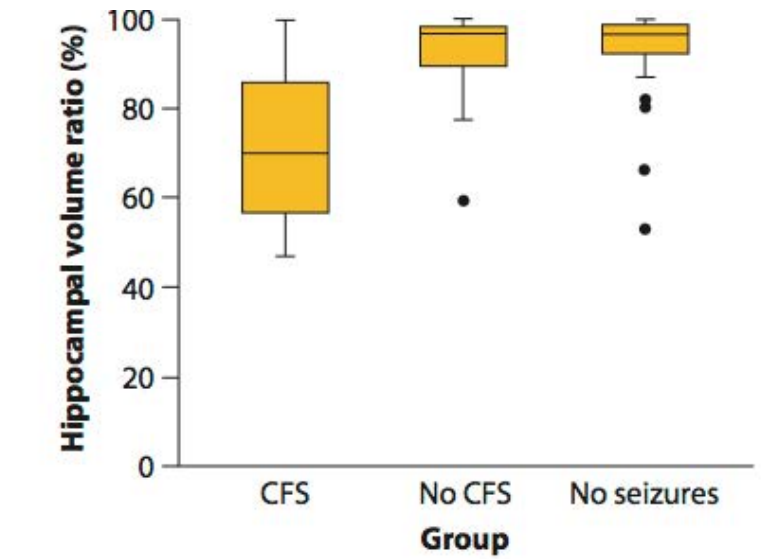
# 2. Display data

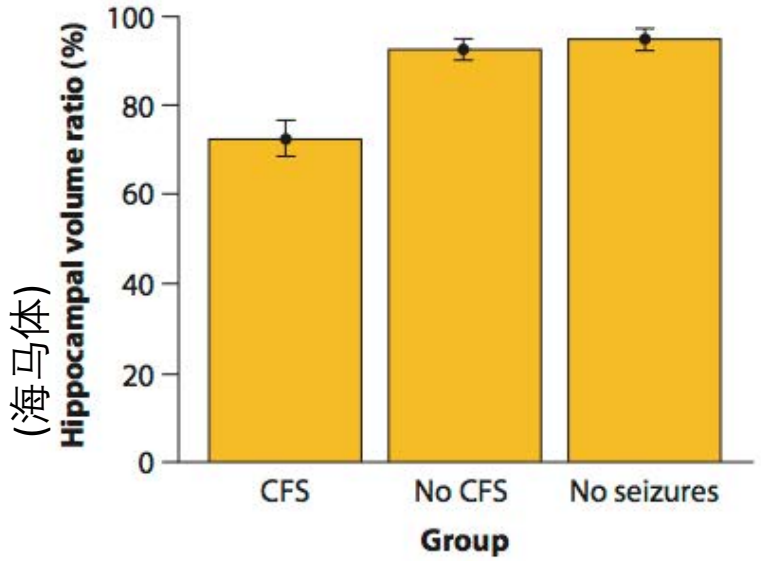
- Four useful principles (原则) to increase the effectiveness (有效性) of your graphs
  - Show the data
  - Make patterns in the data easy to see
  - Represent magnitudes honestly
  - Draw graphical elements clearly, minimizing clutter

- Principle 1: Show the data
  - Show the individual data points
    - E.g., a **scatter plot** (散点图) reveals patterns that are hidden in the **bar graph** (柱状图).
    - Each data point is the serotonin (血清素, 一种神经递质) level of one of 30 locusts (蝗虫) experimentally caged at high density for 0, 1, or 2 hours (0 as control).
      - A clear shift between treatment, but averages vs. extreme values?



- Principle 1: Show the data
  - Box plots (箱形图) can substitute for scatterplot when there are many data points.
  - Which graph is more effective? Why?



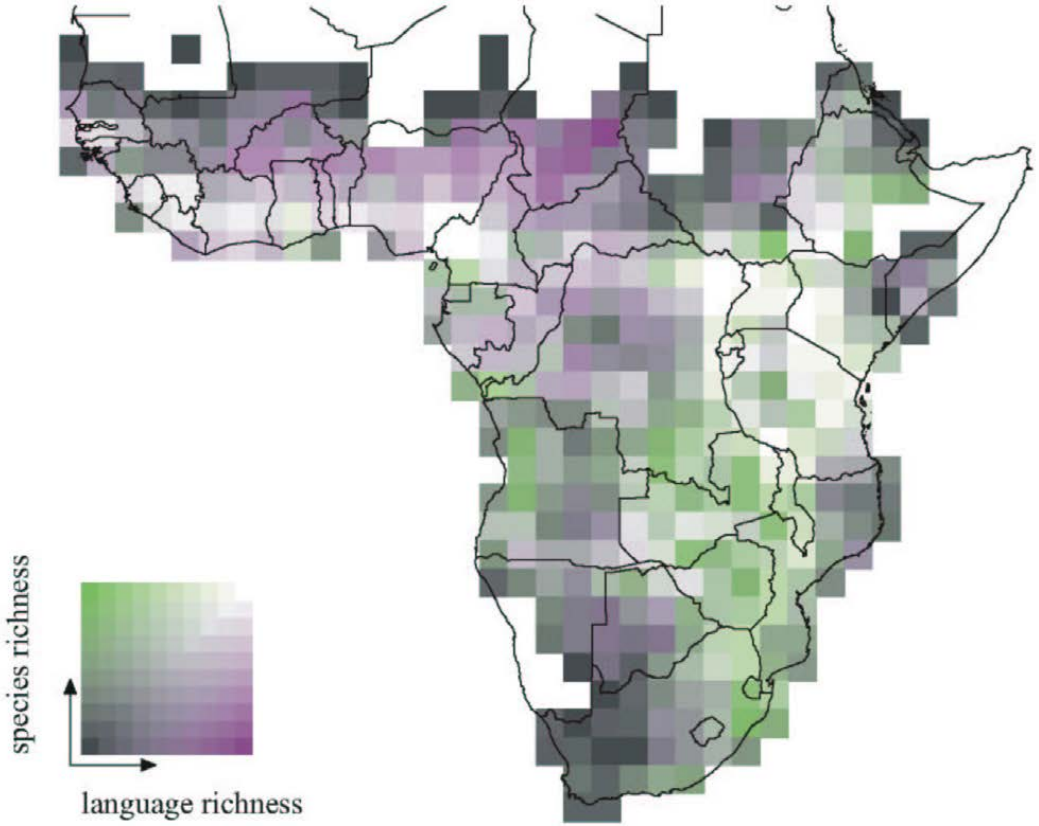

The graphs at the right are from a study investigating hippocampal volume loss in 107 patients with drug-resistant epilepsy (Cook et al. 1993). The graphs depict the association between hippocampal volume loss (measured using MRI as the volume of the smaller half of the hippocampus divided by the volume 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.

- Principle 2: Make patterns in the data easy to see

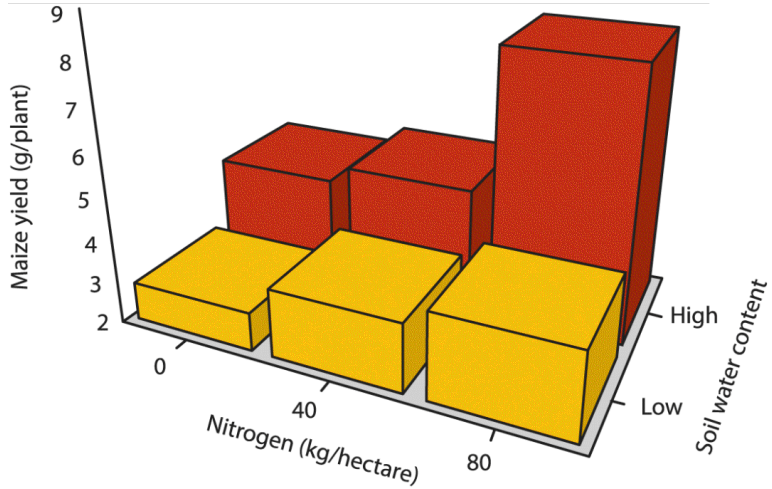
> *"Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency"* – Tufte (1983)

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?

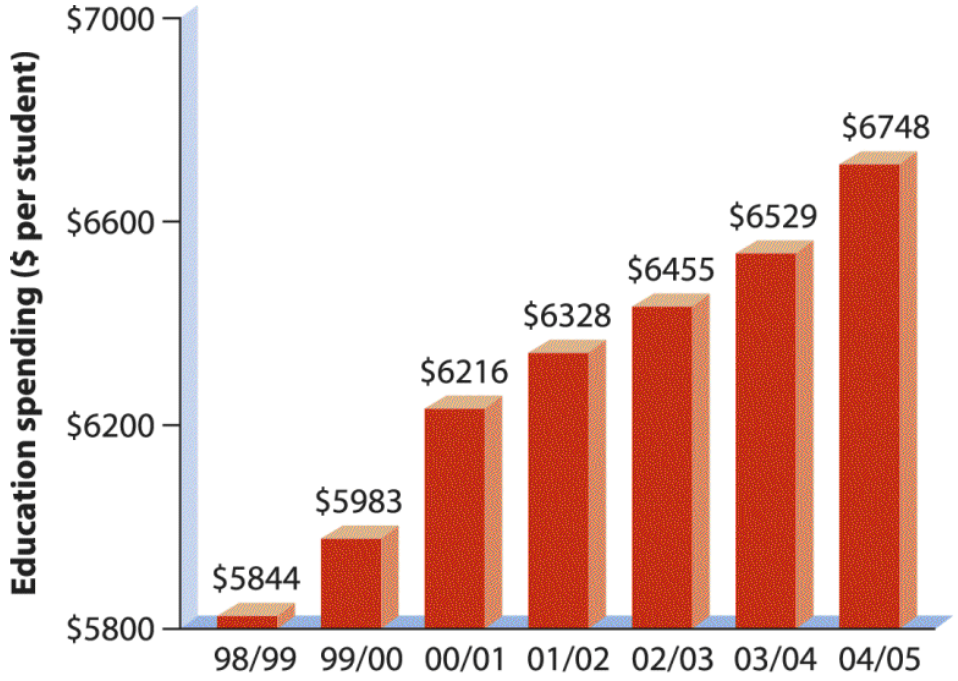


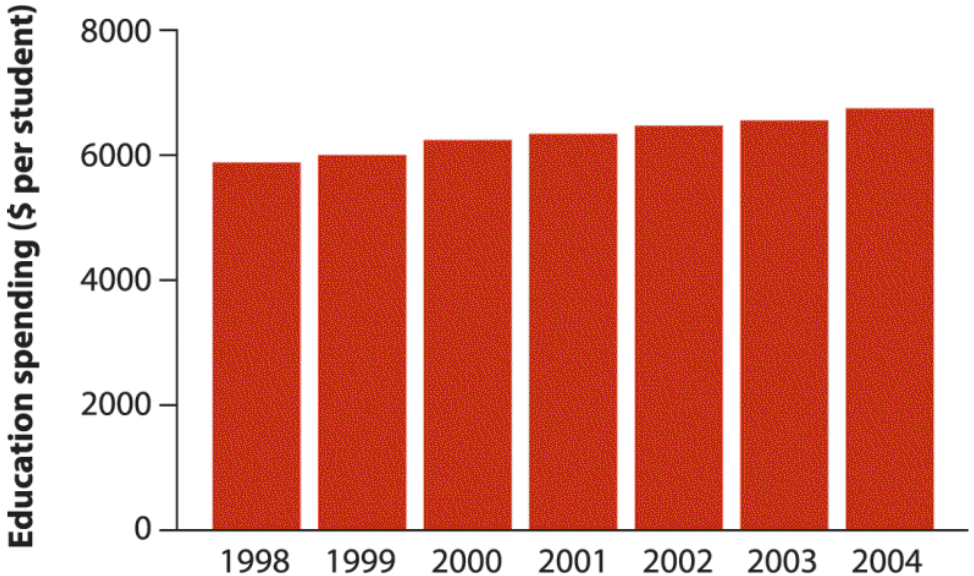
- Principle 2: Make patterns in the data easy to see
  - There are more than one way to show the same data, so, try displaying your data in different ways.
  - Stay away from 3-D effects and chartjunk, which tend to obscure the patterns in the data.
  - Is the main pattern in the data recognizable (可辨识的) right away?
  - Avoid putting too much information into one graph.




Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

- Principle 3: Represent magnitudes honestly
  - One of the most important decisions concerns the smallest value on the vertical axis of a graph (the “baseline”)(纵坐标最小值).
  - A bar graph must always have a baseline at zero, because the eye instinctively reads bar height and area as proportional to magnitude.



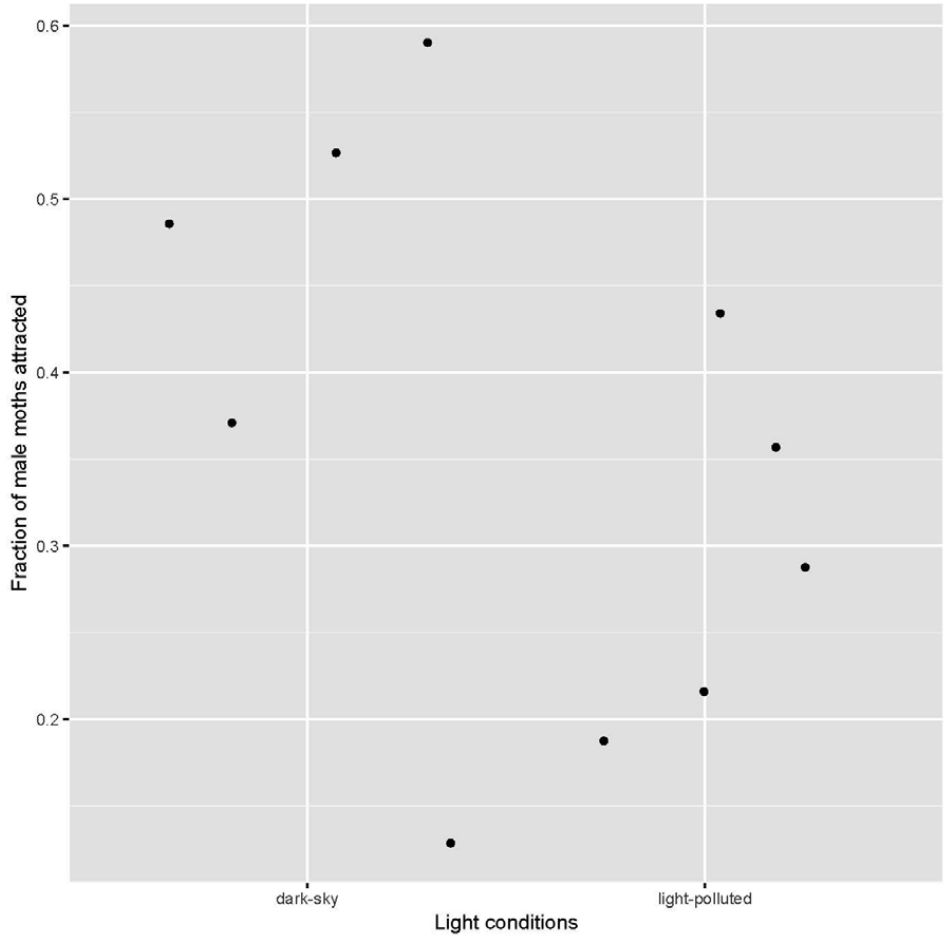
- Principle 3: Represent magnitudes honestly
  - One of the most important decisions concerns the smallest value on the vertical axis of a graph (the “baseline”)(纵坐标最小值).
  - A bar graph must always have a baseline at zero, because the eye instinctively reads bar height and area as proportional to magnitude.
  - Graphs without bars, such as scatterplots, don’t always need a zero baseline if the main goal is to show differences between treatments rather than proportional magnitudes.

- Principle 4: Draw graphical elements clearly
  - Clear axes and labels
  - Readable texts
  - The units of measurements
  - Clearly distinguishable graphical symbols
  - Distinguishable colors
    - E.g., for red-green color blind audience
  - Complete legend

- Principle 4: Draw graphical elements clearly
  - Axes, labels, texts, units, symbols, colors, legend

*How to improve this figure?*

*- 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.*



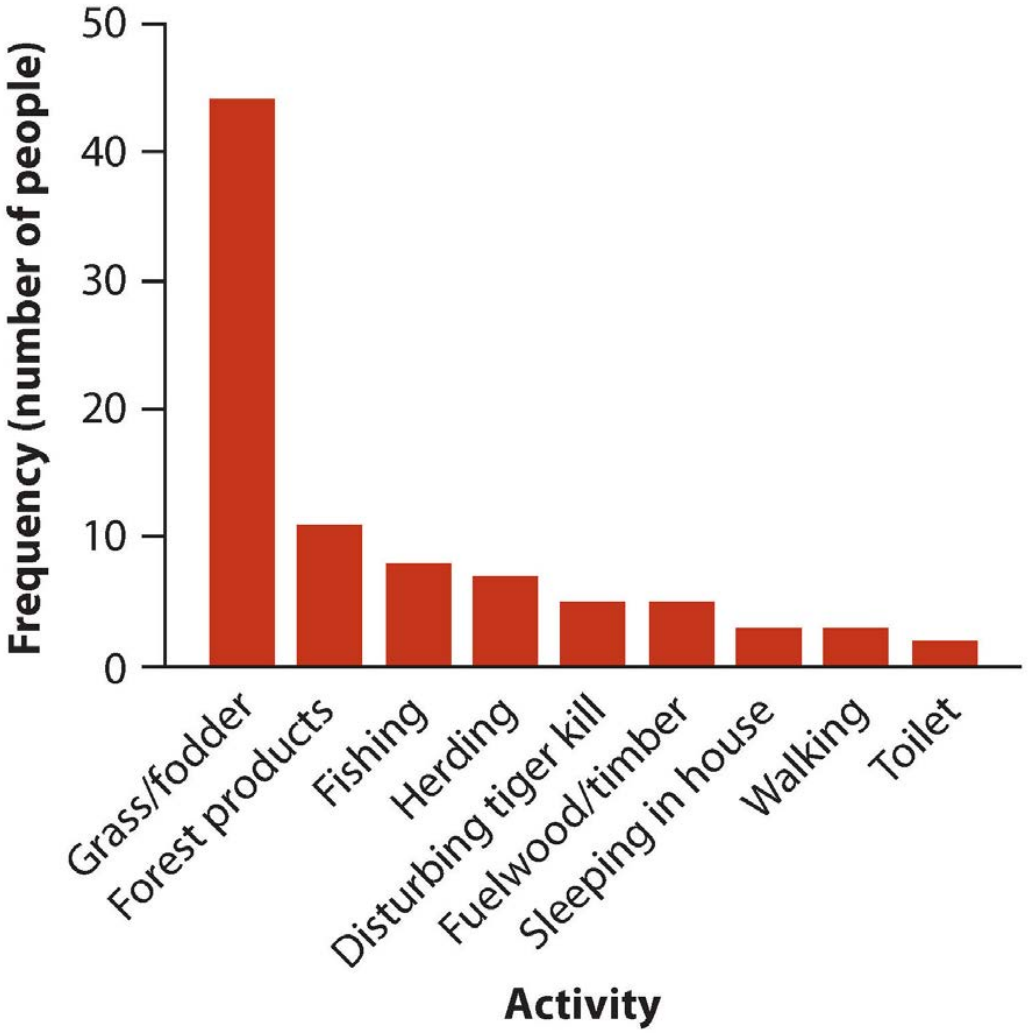
# 2. Display data

- Four useful principles (原则) to increase the effectiveness (有效性) of your graphs
  - Show the data
  - Make patterns in the data easy to see
  - Represent magnitudes honestly
  - Draw graphical elements clearly, minimizing clutter

# 3. Types of graphs

- Which types of graphs best show the data?
  - The top choices will depend on the type of data, numerical or categorical, and whether the goal is to show measurements of one variable or the association between two variables.

- Display category frequencies
  - Bar graph (柱状图)
  - 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

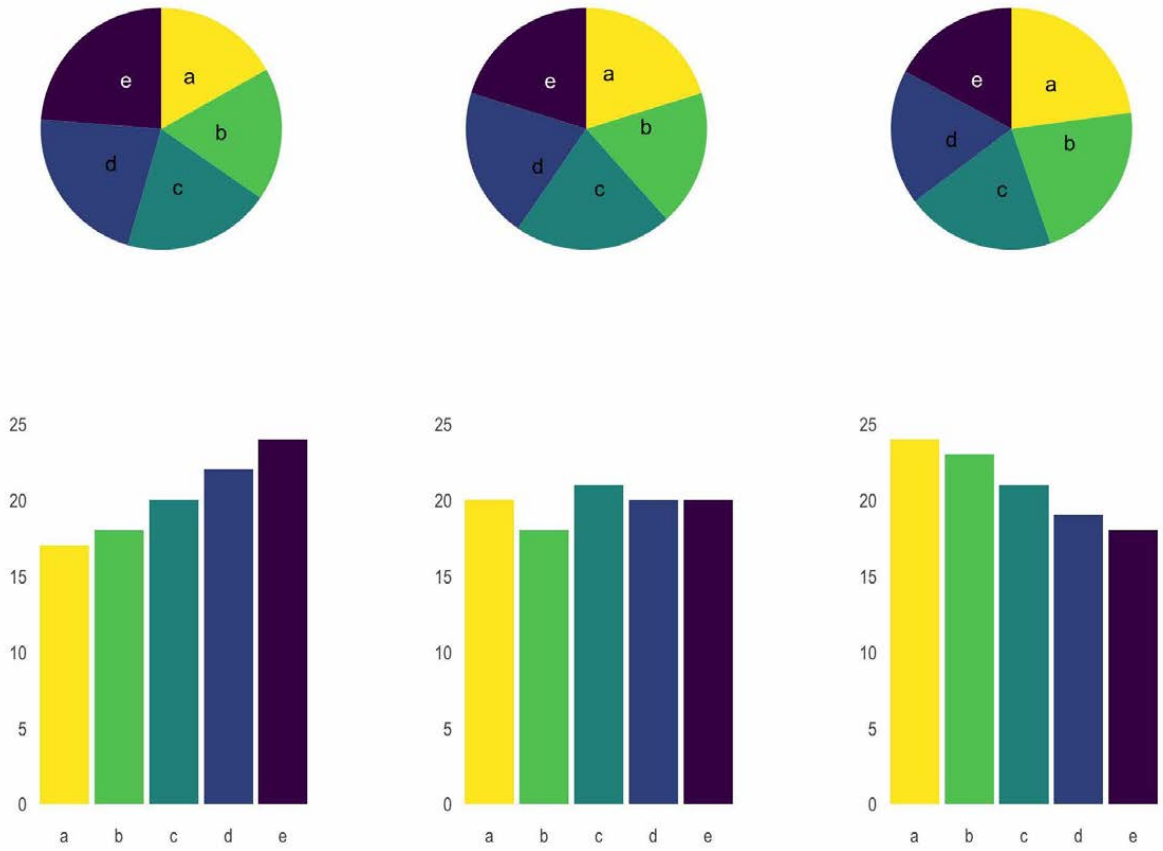



Activities of people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, between 1979 and 2006.

(Whitlock & Schluter 2020)

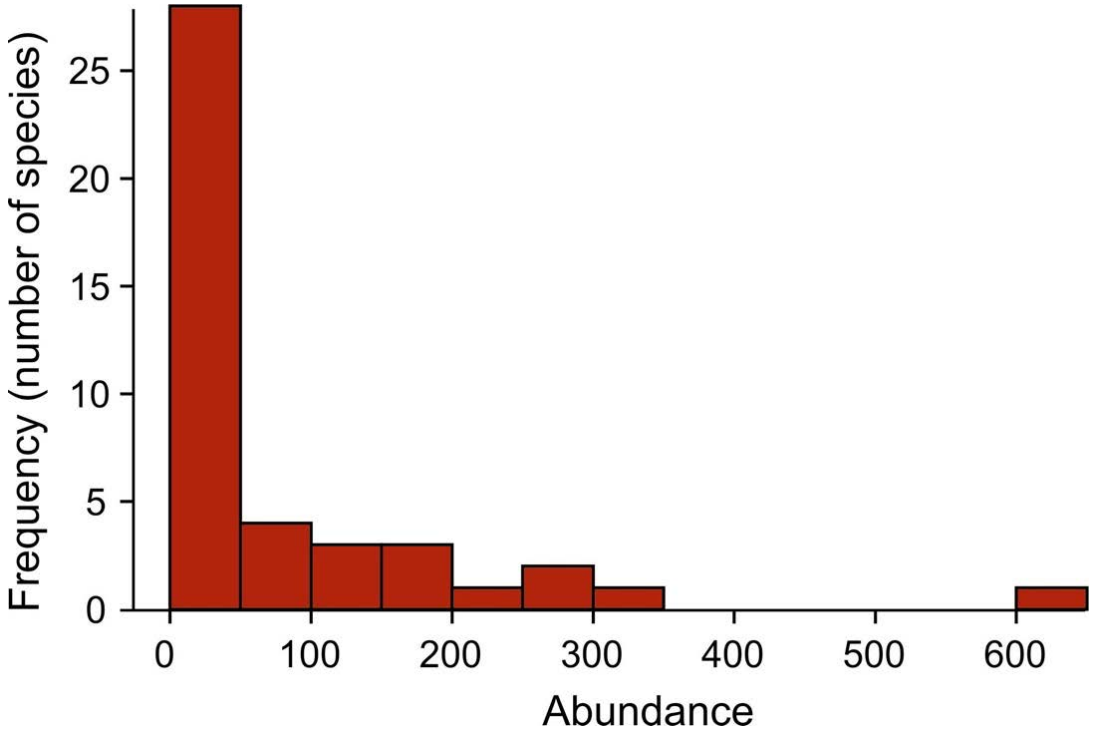
- Display category frequencies
  - Bar graph vs. Pie chart
  - Which is more successful?

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




(www.data-to-viz.com from Schluter 2022)

- Display frequency distribution for numeric variable
  - Histogram (柱状图)
  - 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

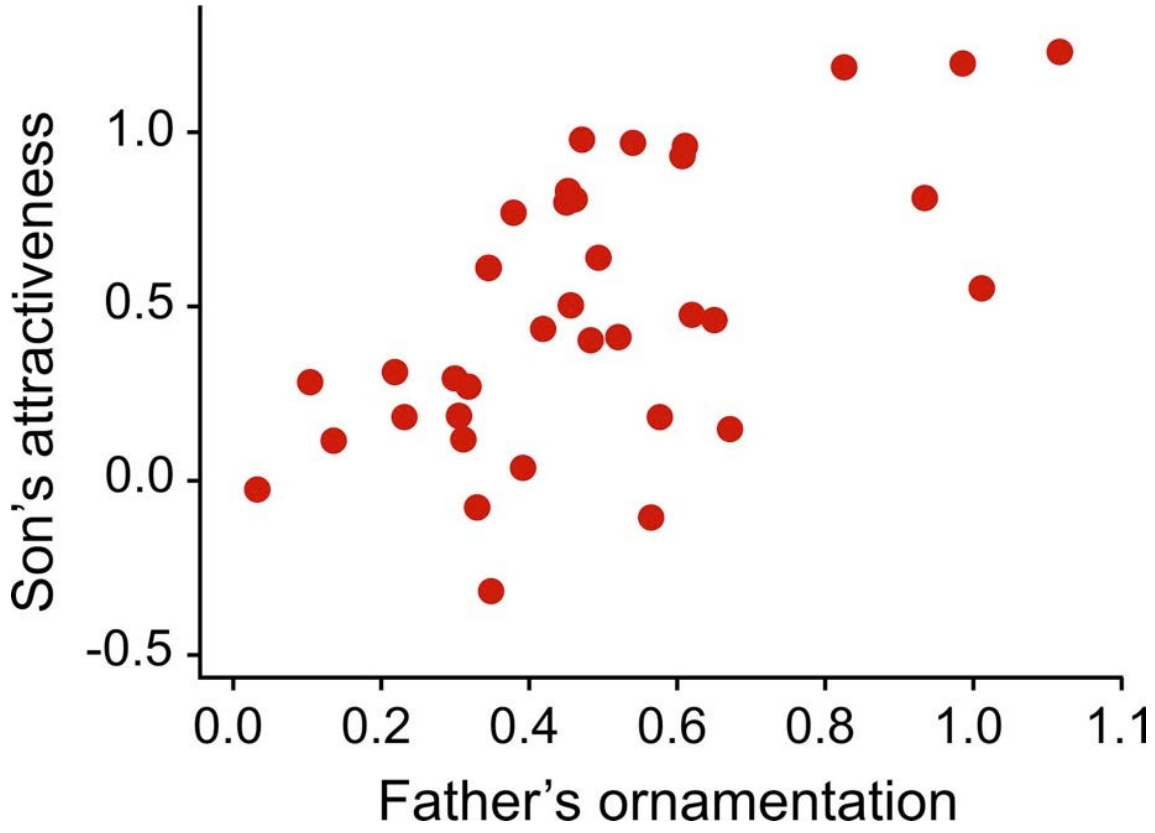



*The frequency distribution of bird species abundance at Organ Pipe Cactus National Monument. n = 43 species*

(Whitlock & Schluter 2020)

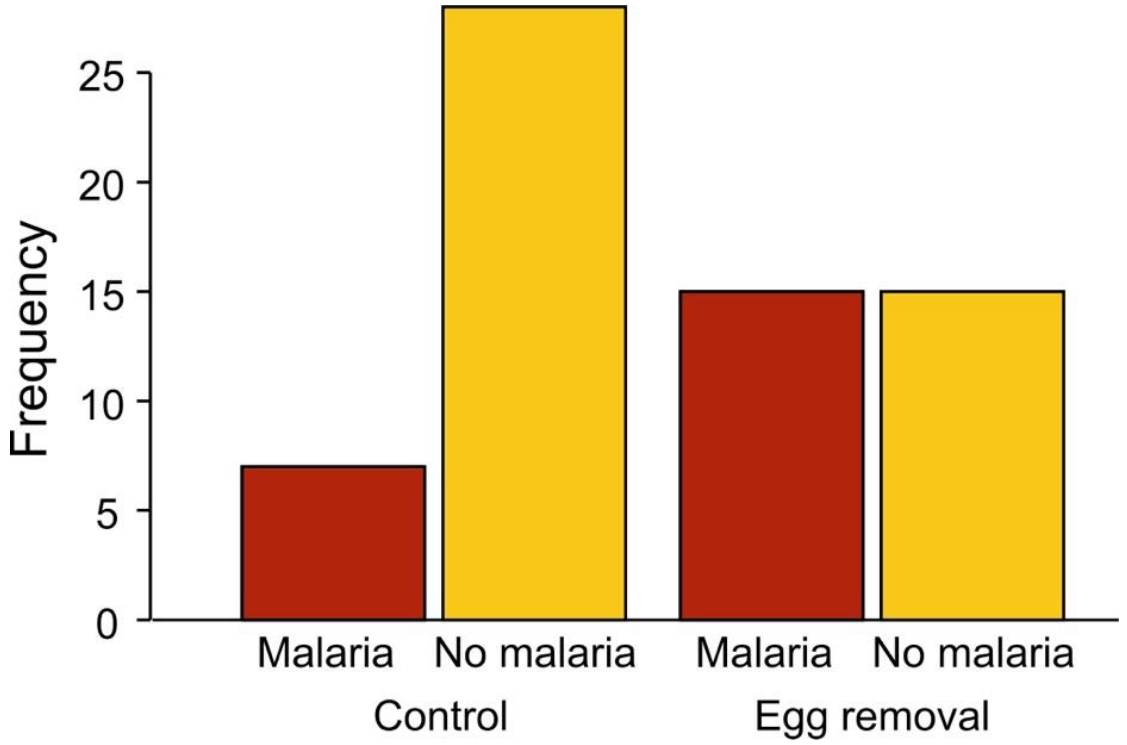
- Display association between two numerical variables
  - Scatter plot (散点图)
  - Non-zero baseline often ok
  - goal is to show association
    - not height above 0
  - Points fill the space available

*The relationship between the ornamentation (装饰) of male guppies (孔雀鱼) and the average attractiveness of their sons.*
*n = 36 families.*




(Whitlock & Schluter 2020)

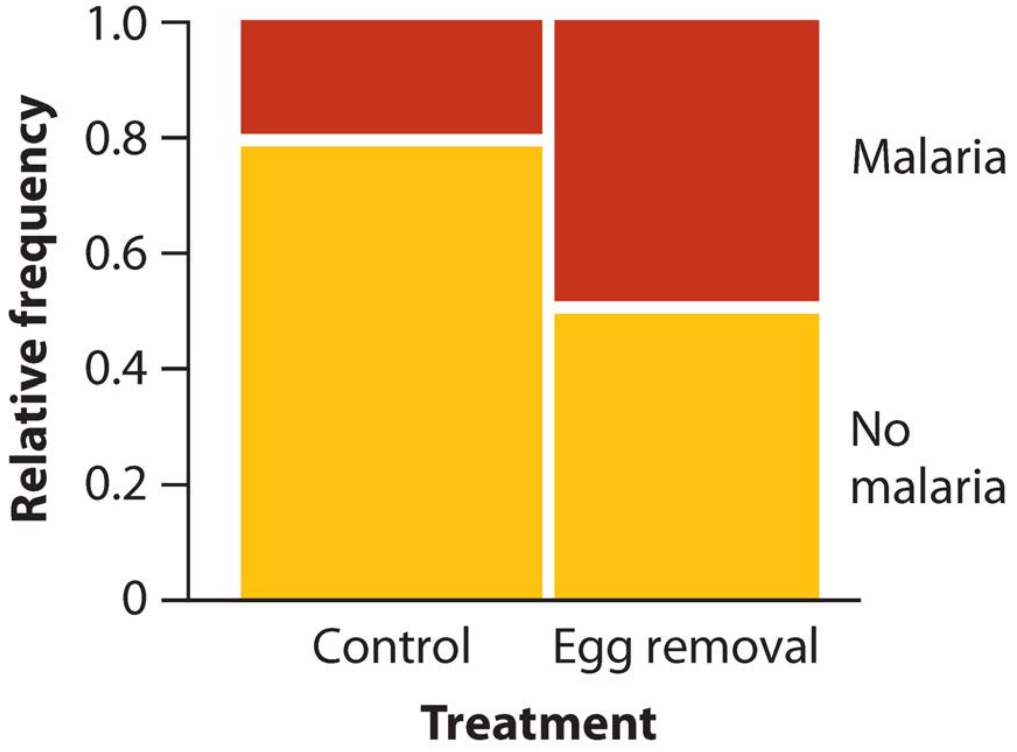
- Display association between categorical variables
  - Grouped bar graph (分组柱状图)
  - Uses height of bars to display association between two (or more) categorical variables.
    - Explanatory variable = outer groups;
    - response variable = inner groups
    - Zero baseline (so that height is proportional to frequency)
    - Spacing between bars wider between outer groups




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

(Whitlock & Schluter 2020)

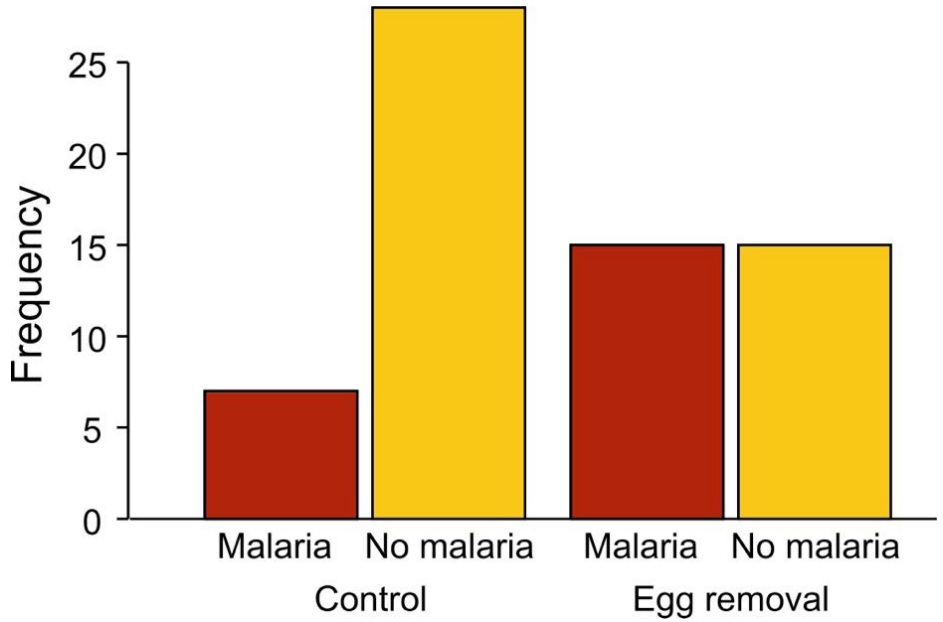
- Display association between categorical variables
  - Mosaic graph (马赛克图)
  - Uses area of rectangles to display association between two (or more) categorical variables.
    - Explanatory variable along horizontal axis; response variable stacked
    - Area proportional to frequency
    - Like a graphical representation of a contingency table

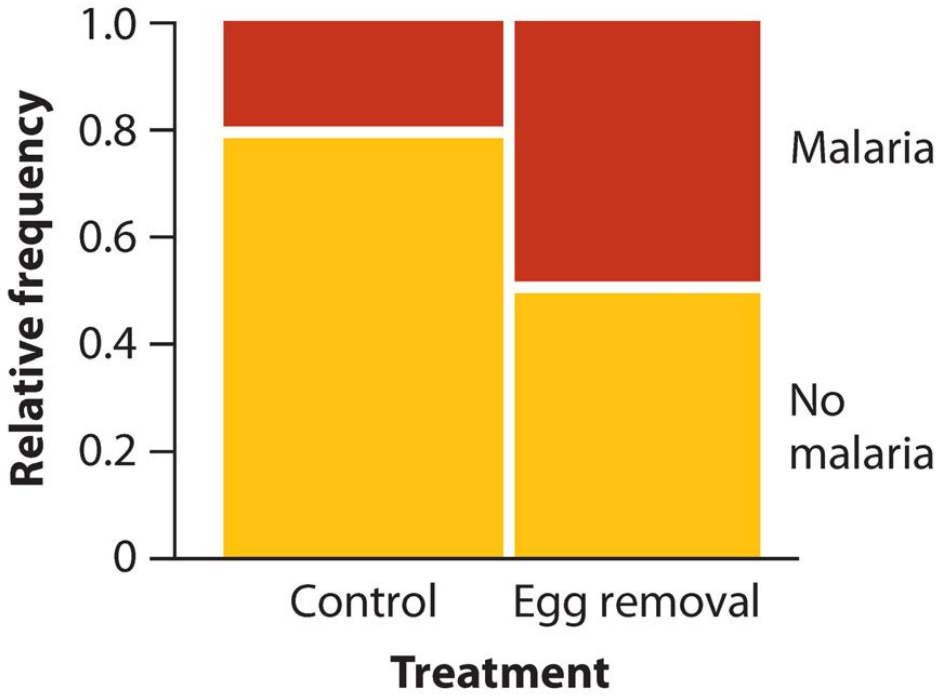



*Incidence of malaria in female great tits in relation to experimental treatment.*
*n = 65 birds*

(Whitlock & Schluter 2020)

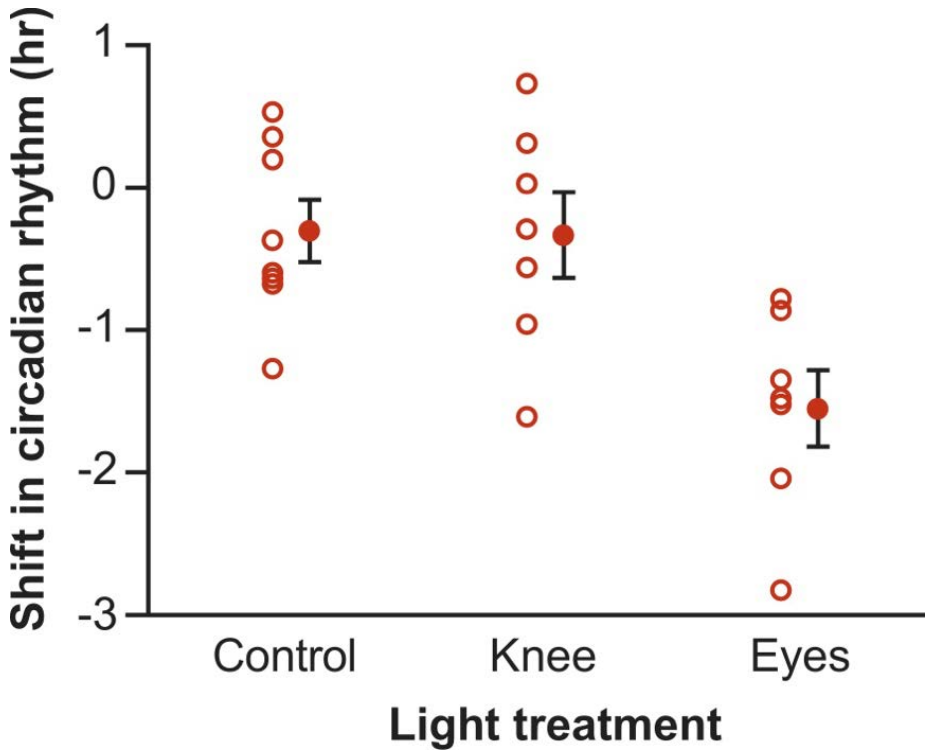
- Display association between categorical variables
  - Grouped bar graph vs . Mosaic graph
  - Which is more successful?




(Whitlock & Schluter 2020)

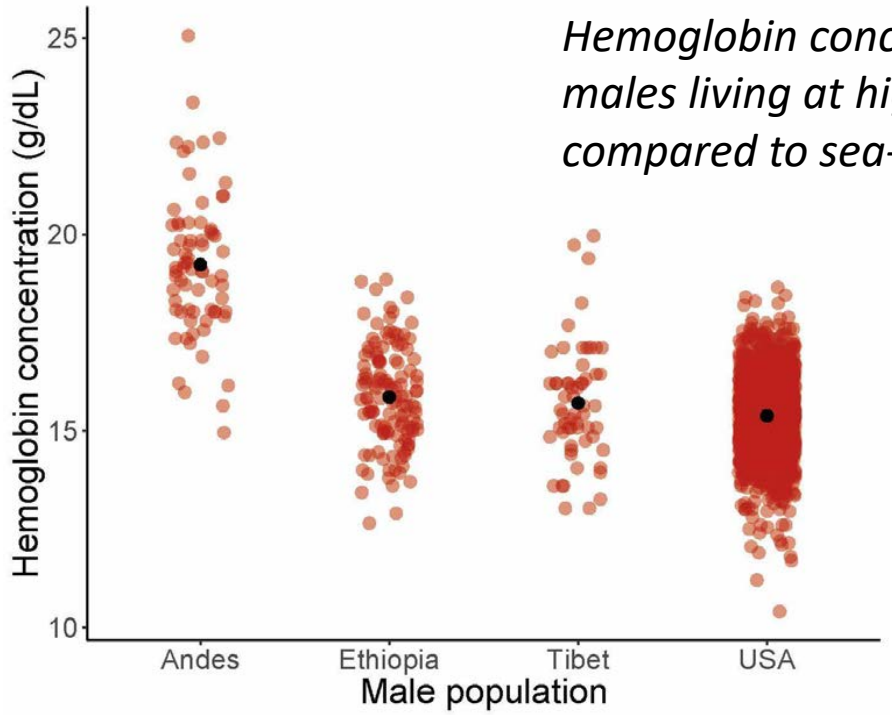
- Display association between numerical and categorical variable
  - Strip graph (条形图/带状图)
  - Displays differences between groups
    - Shows the data points
    - Non-zero baseline often ok
      - goal is association not magnitude or frequency
    - Points fill the space available




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

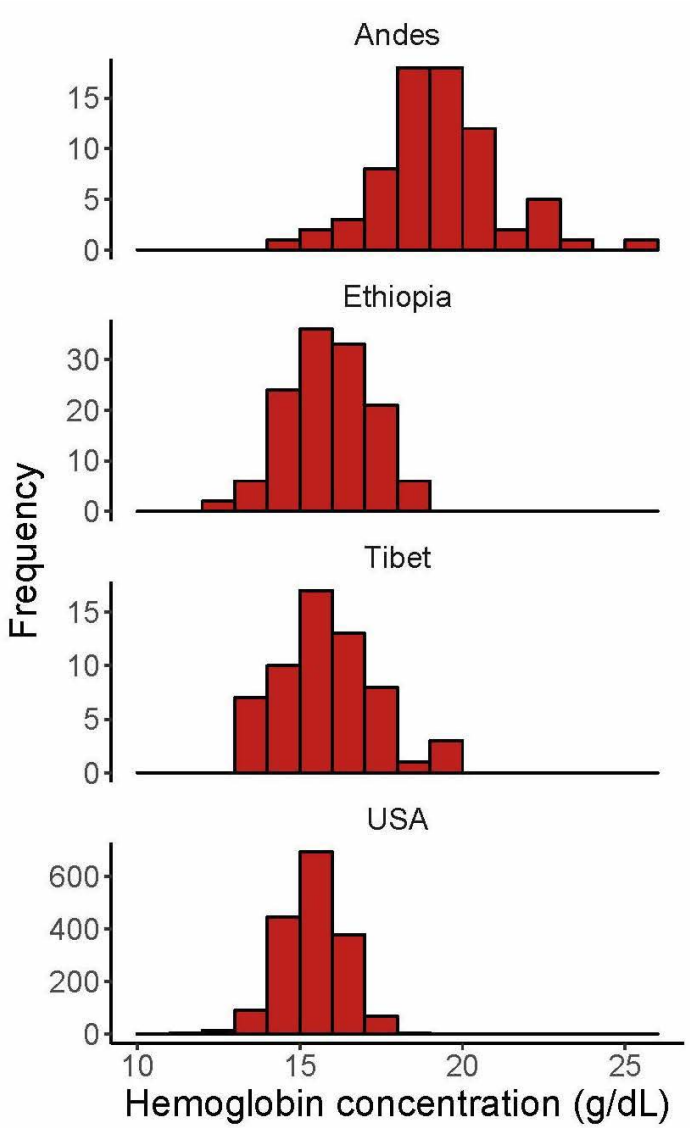
(Whitlock & Schluter 2020)

- Display association between numerical and categorical variable
  - Strip chart vs multiple histograms
  - Too many data points for a strip chart.
  - Stack histograms vertically to best compare

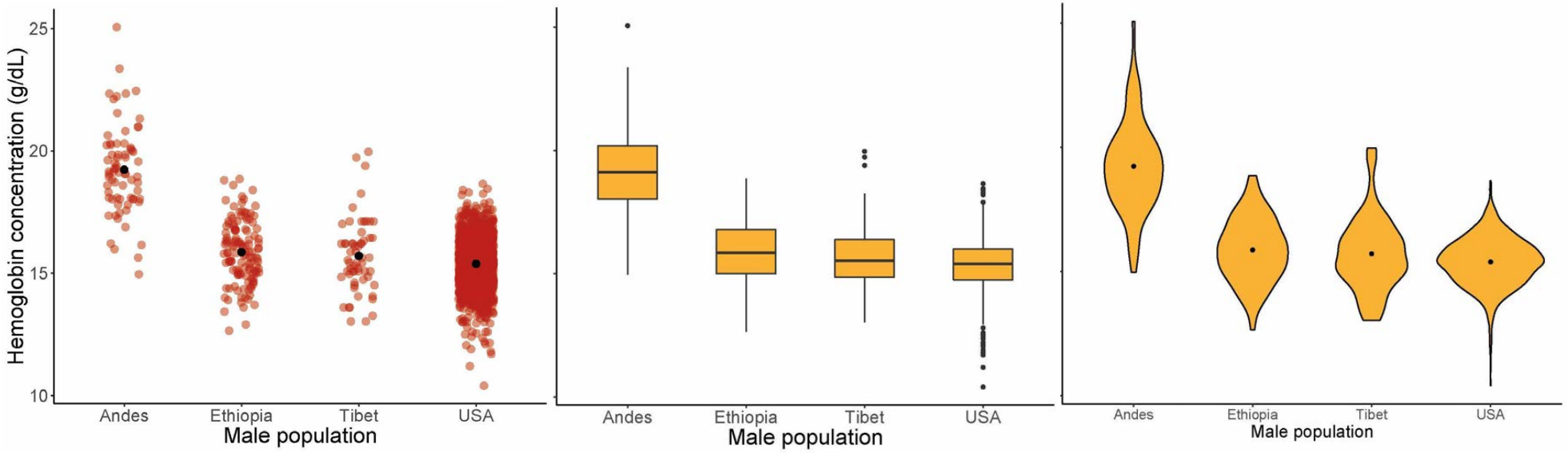



*Hemoglobin concentration in blood of males living at high elevation compared to sea-level USA control.*

(Whitlock & Schluter 2020)

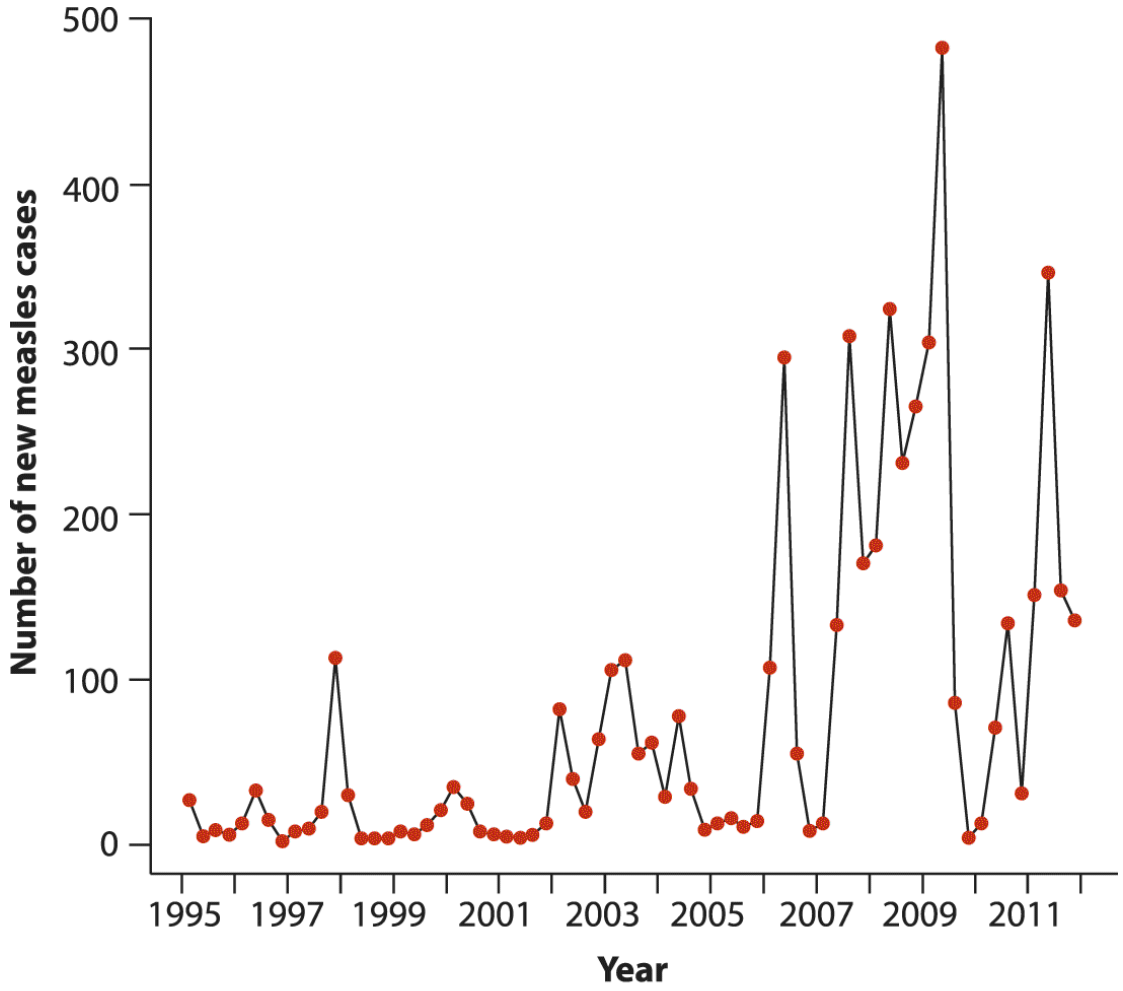


- Display association between numerical and categorical variable
  - Strip chart vs box plot vs violin plot (小提琴图)
    - Box plot displays median, first and third quartile, range, and extreme observations
    - Violin plot estimates probability density of each group using “kernel smoothing”
    - Non-zero baseline often ok (goal is to show differences not amounts)
  - Which is more successful?




(Whitlock & Schluter 2020)

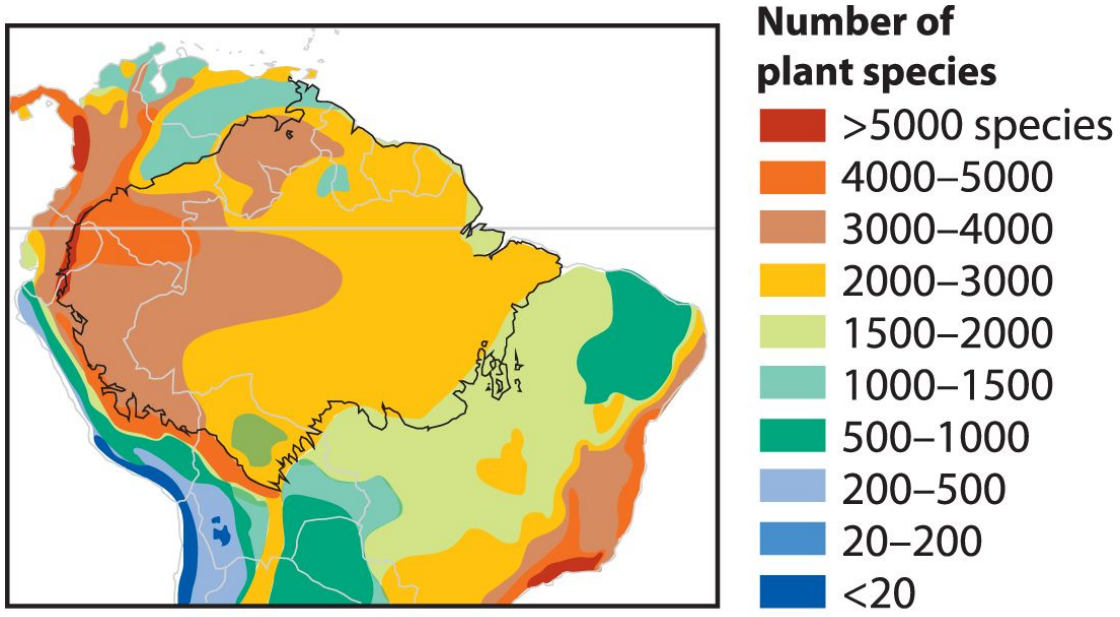
- Display trend in time
  - Line graph (线状图)
    - uses dots connected by line segments to display trends over time
    - the steepness of the line segments reflects the speed of change in the number.




*The number of cases of measles (麻疹) (the summary measurement) for quarterly time intervals between 1995 and 2011.*

(Whitlock & Schluter 2020)

- Display trend in space
  - Map (地图)
    - uses a color gradient to display a numerical response variable at multiple locations on a surface
    - the spatial equivalent of the line graph
    - the explanatory variable = location
      - a spatial grid or political boundaries
    - can be used to show measurements at locations on the surface of any two- or three-dimensional objects, such as the brain or the body.




*Numbers of plant species in northern South America, with each point consisting of an area 100 km×100 km.*

(Whitlock & Schluter 2020)

# 4. Tables (表格)

- 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

# 4. Tables (表格)

- 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).

# 4. Tables (表格)

- Improve the table!
  - 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 |

# 5. Summary

- Graphical displays must be clear, honest, and efficient.
- Strive to show the data, to make patterns in the data easy to see, to represent magnitudes honestly, and to draw graphical elements clearly (the same rules for tables).
- Bar graphs and histograms are recommended graphical methods for displaying frequency distributions of categorical and numerical variables.

# 5. Summary

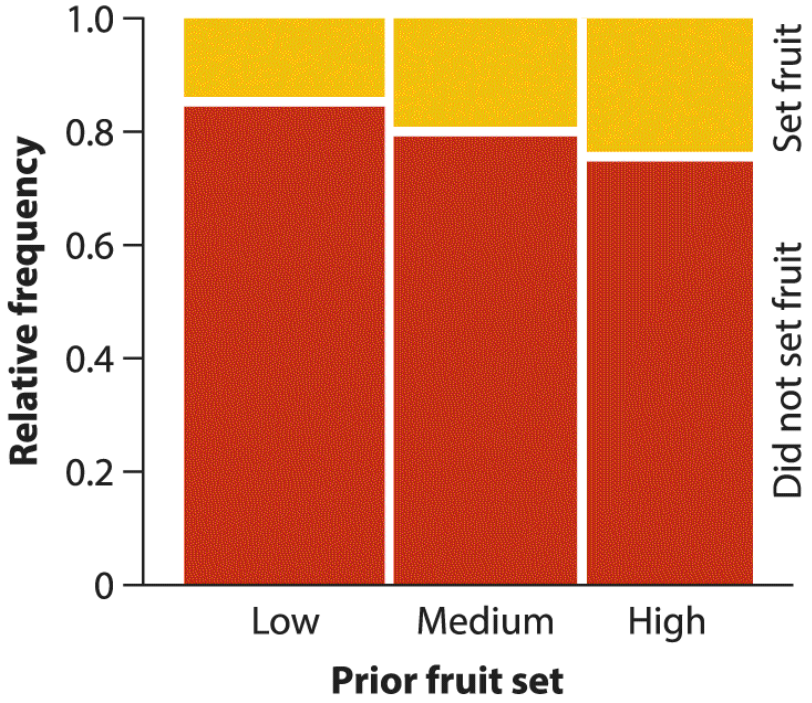
- Recommended graphical methods for displaying associations between variables and differences between groups include the following:

| **Types of data** | **Graphical method** |
|---|---|
| Two numerical variables | Scatter plot |
| Two categorical variables | Grouped bar graph<br>Mosaic plot |
| One numerical variable and one categorical variable | Strip chart<br>Violin plot<br>Multiple histograms<br>Cumulative frequency distributions (L03) |

# 6. Discussion

- 1. Each of the following graphs illustrates an association between two variables. For each graph, identify
  - (1) the type of graph,
  - (2) the explanatory and response variables,
  - (3) the type of data for each variable.
    - whether numerical or categorical

*a. Observed fruiting of individual plants in a population of Campanula americana (美国风铃草) according to the number of fruits produced previously (Richardson and Stephenson 1991):*

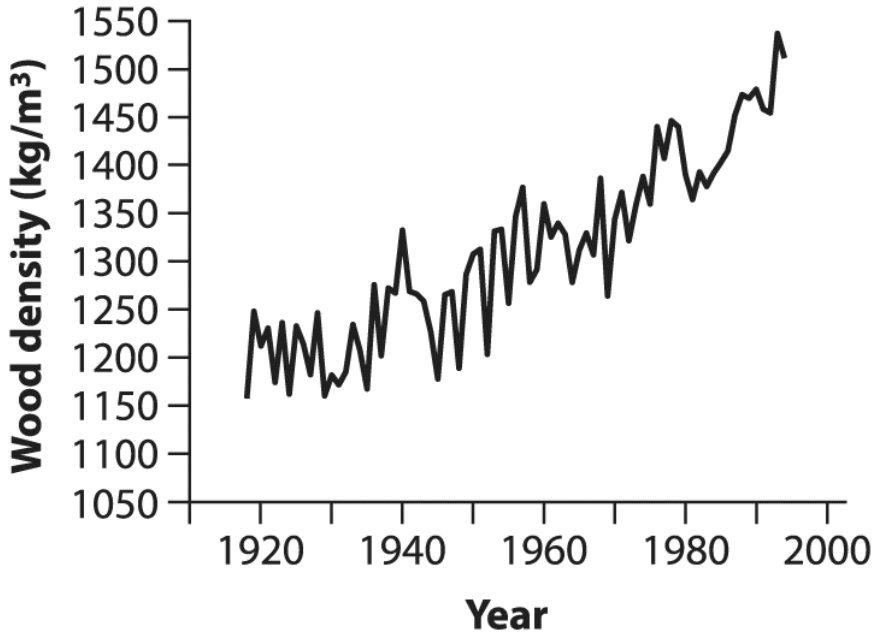



Whitlock & Schluter, *The Analysis of Biological Data*, 3e

# 6. Discussion

- 1. Each of the following graphs illustrates an association between two variables. For each graph, identify
  (1) the type of graph,
  (2) the explanatory and response variables,
  (3) the type of data for each variable.
    - whether numerical or categorical

*b. The maximum density of wood produced at the end of the growing season in white spruce trees (白云杉) in Alaska in different years (data from Barber et al. 2000):*




Whitlock & Schluter, *The Analysis of Biological Data*, 3e

# 6. Discussion

- 2. Draw scatter plots for invented data that illustrate the following patterns:

  a. Two numerical variables that are positively associated

  b. Two numerical variables that are negatively associated

  c. Two numerical variables whose relationship is nonlinear

  - which type of graph to use?