Chapter 2: Displaying data

1

Graphs

- Graphs are vital tools for analyzing data!
- …but many graphs fail to accurately convey the data

2

Can't see data:
Bars display average estimate, but fail to convey variation among samples

Patterns difficult to see due to 3-D and tilt:
e.g., hard to compare these two bars

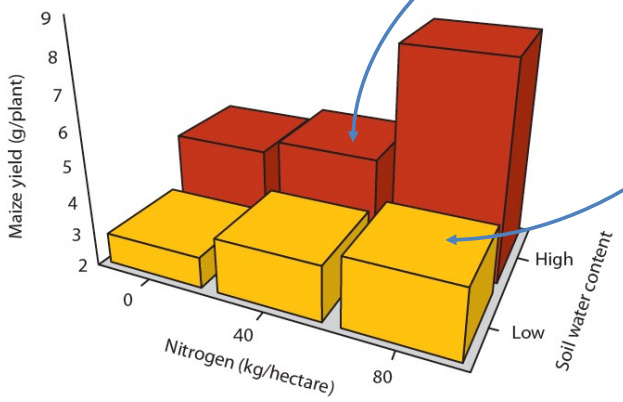


Fig 2.1-1

3

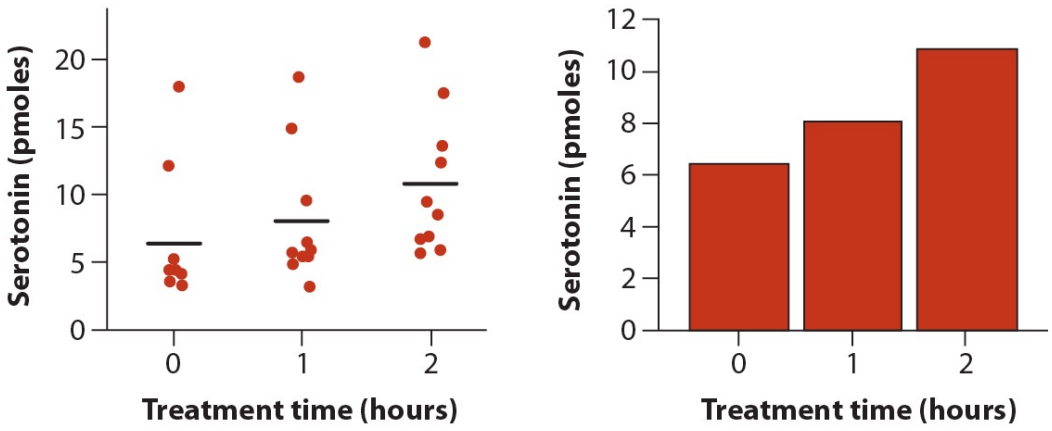
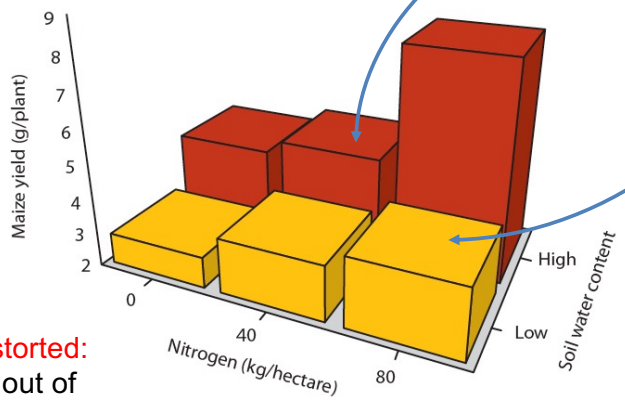


Fig 2.1-2

4

Can't see data:
Bars display average estimate, but fail to convey variation among samples

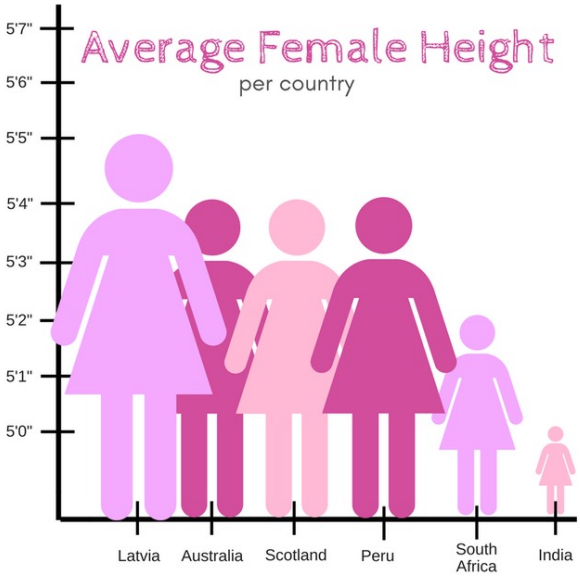
Patterns difficult to see due to 3-D and tilt:
e.g., hard to compare these two bars



Magnitudes are distorted:
Makes differences out of proportion

Fig 2.1-1

5



6

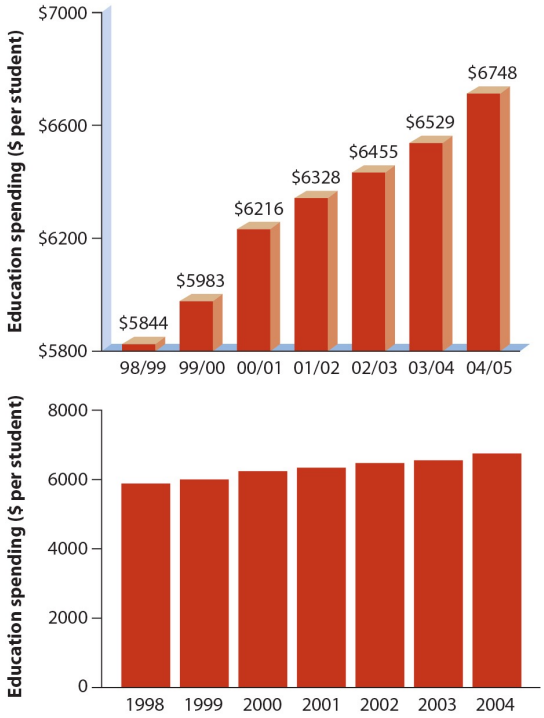


Fig 2.1-3

7

TABLE 2.2-1 Frequency table showing the activities of 88 people at the time they were attacked and killed by tigers near Chitwan National Park, Nepal, from 1979 to 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

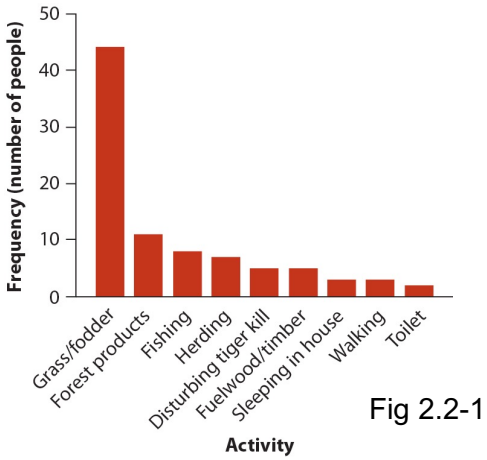


Fig 2.2-1

- Tables are good when you want to report/compare specific values with precision
- Graphs are good when you want to show trends or patterns in values

8

Bar graph

- A **bar graph** used the height of rectangular bars to display the frequency distribution (or relative frequency distribution) of a categorical variable

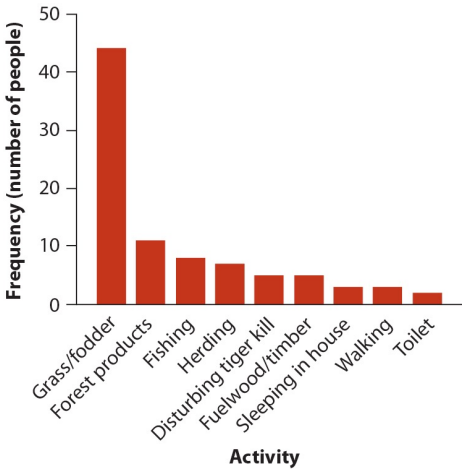


Fig 2.1-1

9

Grouped bar graph

- A **grouped bar graph** uses the height of rectangular bars to display the frequency distributions (or relative frequency distributions) of two or more categorical variables

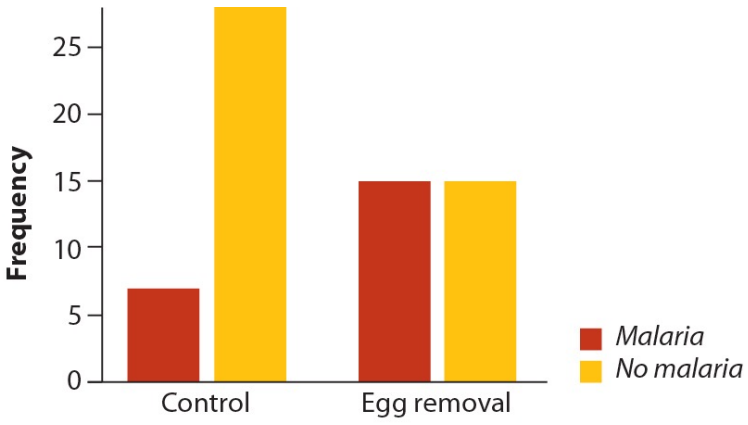


Fig 2.3-1

10

Bar graph or pie chart?

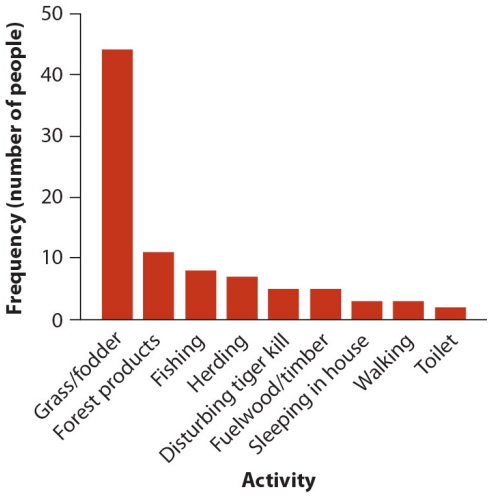


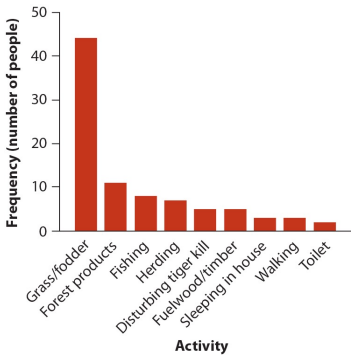
Fig 2.2-1

Fig 2.2-2

11

Bar graph vs histogram

- A **bar graph** used the height of rectangular bars to display the frequency distribution (or relative frequency distribution) of a **categorical** variable



- A **histogram** used the height of rectangular bars to display the frequency distribution (or relative frequency distribution) of a **numerical** variable

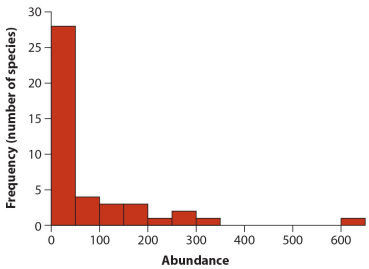
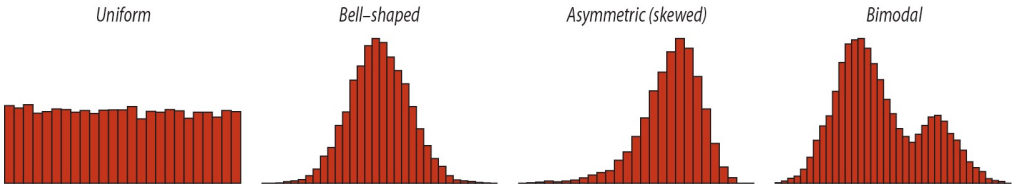


Fig 2.2-3

12

Shape of a histogram

- The **mode** is the interval corresponding to the highest peak in the frequency distribution



- **Skew** refers to asymmetry in the shape of a frequency distribution for a numerical variable

- An **outlier** is an observation well outside of the values of other observations in the data set

Fig 2.2-4

13

Some other types of graphs

14

Mosaic plot

- The **mosaic plot** uses the area of rectangles to display the relative frequency of occurrence of all combinations of two categorical variables

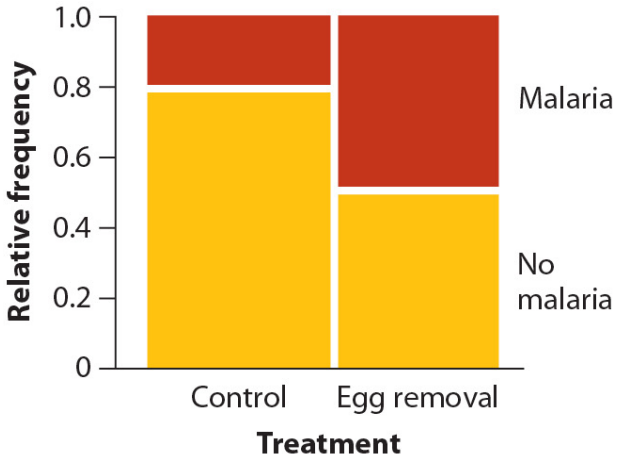


Fig 2.3-2

15

Scatter plot

- A **scatter plot** is a graphical display of two numerical variables in which each observation is represented as a point on a graph of two axes

- Shows association between two numerical variables

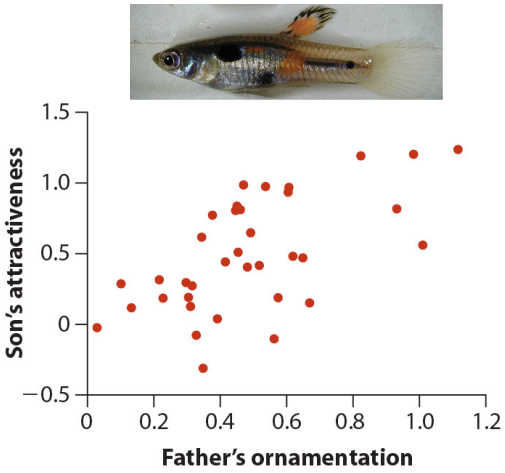


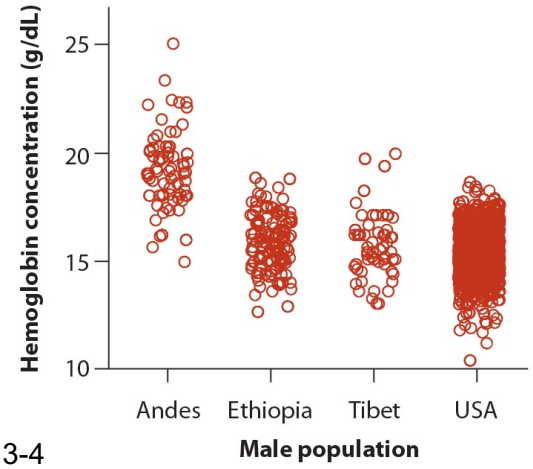
Fig 2.3-3

16

Scatter plot

- The **strip plot** is a graphical display of a numerical variable and a categorical variable in which each observation is represented as a dot

Note that when there are a lot of data points they are often "jittered" so that they don't overlap as much



https://youtu.be/elOn5ZYg5fc?t=112

Fig 2.3-4

17

Box plot

- A **box plot** is a graph that uses lines and a rectangle box to display the median, quartiles, range, and extreme measurements of the data

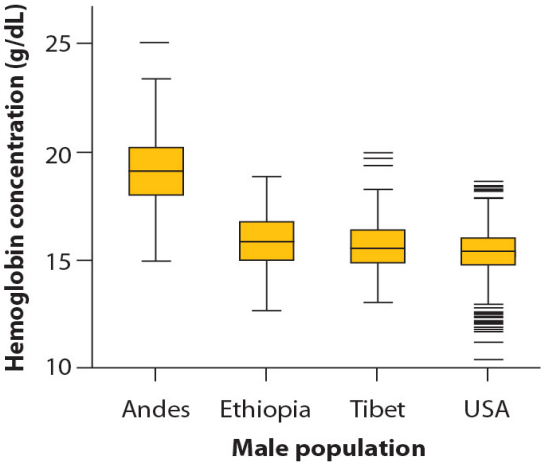


Fig 2.3-4

18

Violin plot

- A **violin plot** is a graph that shows an approximation of the frequency distribution of a numerical variable in each group and its mirror image

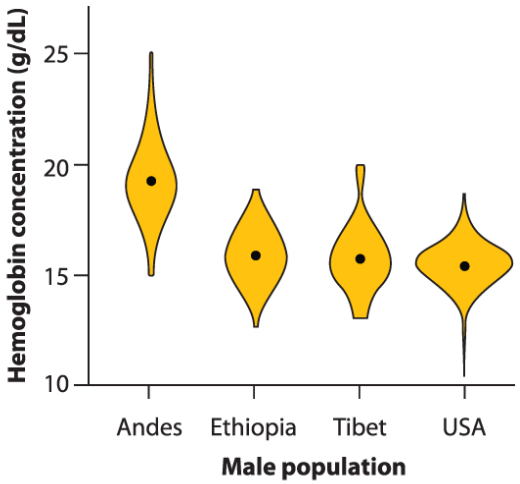


Fig 2.3-4

19

Multiple histograms

- Better to "stack" vertically rather than side-by-side because it is easier to compare groups
- Also, best to use same scale for x-axis

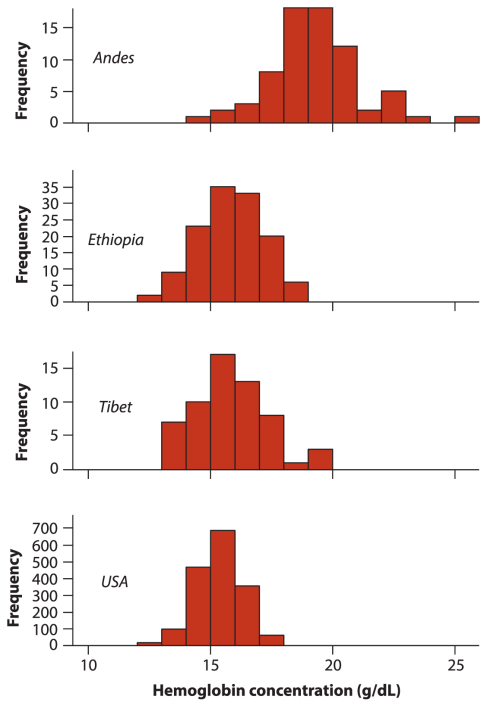


Fig 2.3-5

20

Line graph

- A **line graph** uses dots connected by line segments to display trends in a measurement over time or other ordered series

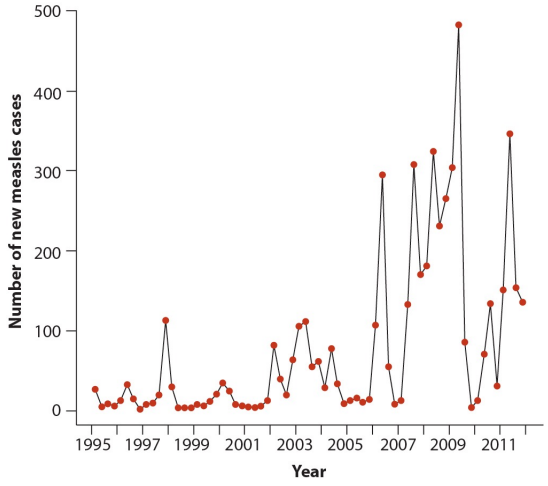


Fig 2.4-1

21

Map

- A **map** is a spatial equivalent of the line graph, using a color gradient to display a numerical response variable at multiple locations on a surface

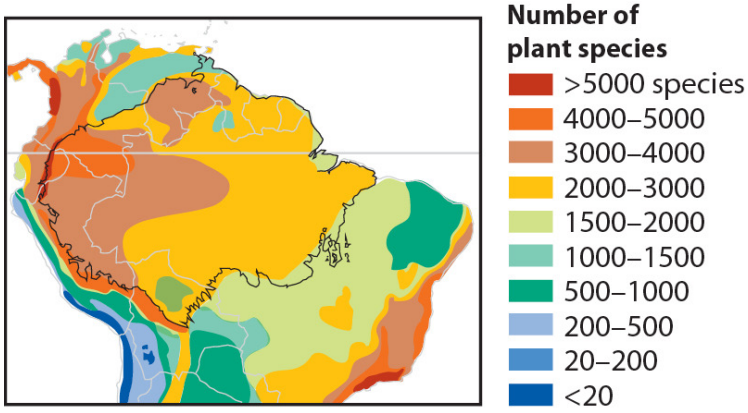
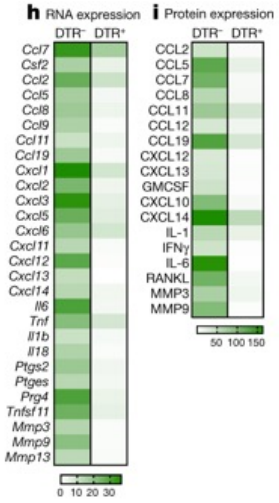


Fig 2.4-2

22

Map

- A **map** is a spatial equivalent of the line graph, using a color gradient to display a numerical response variable at multiple locations on a surface



Croft et al. 2019. Nature

23

Data files

24

Tips for making good data files

- Each row is an individual sampling unit, each column is a variable

	A	B	C	D	E	F
1	Patient	Treatment	HR	Patient	Treatment	HR
2	1	control	60	2	drug	65
3	3	control	64	4	drug	76
4	5	control	70	7	drug	83
5	6	control	74	8	drug	88
6	10	control	58	9	drug	73
7	11	control	61	12	drug	80

	A	B	C
1	Patient	Treatment	Heartrate_bmp
2	1	control	60
3	2	drug	65
4	3	control	64
5	4	drug	76
6	5	control	70
7	6	control	74
8	7	drug	83
9	8	drug	88
10	9	drug	73
11	10	control	58
12	11	control	61
13	12	drug	80

25

Tips for making good data files

- Each row is an individual sampling unit, each column is a variable
- Either leave empty cells blank or use a universal notation like “na”
- Avoid special characters
- Variable names
  - Avoid spaces
  - Make name intuitive and unambiguous
- Save as plain text
  - Most common formats are comma separated values (csv) and tab-delimited text (txt)

26