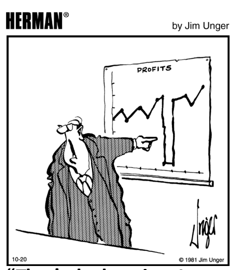
Graphs:
The art of designing information

"A picture tells a thousand words"

- Lake Blanche

1

Graphs are used to try to tell a story



"That's the last time I go on vacation"

…and to make a point

2

General definition of a graph

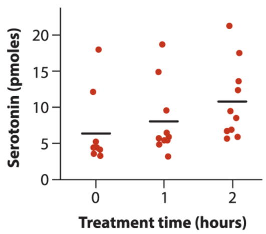
- Visual representation of a relationship between two or three variables (and more sometimes).
- Variables can be of any type (e.g., categorical or numerical).
- They commonly consist of two axes: x-axis (horizontal or abscissa) and y-axis (vertical or ordinate).

3

Average serotonin ("happy chemical") levels in the central nervous systems of desert locusts that were experimentally crowded for 0 (control group), 1 and 2 hours.

1 individual measured per cage of 30 individuals (i.e., control = 8 cages, 1 hour = 11 cages, 2 hours = 10 cages; total of (29 cages x 30) = 870 individuals were used for crowding but less individuals were measured as explained above).

Y-axis (also numerical)



X-axis (numerical)

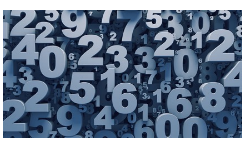
Dots represent averages per cage; horizontal bars within treatments represent average values (average of cage averages)

https://www.aljazeera.com/news/2020/02/south-sudan-hit-desert-locust-swarm-plague-spreads-200219050015190.html

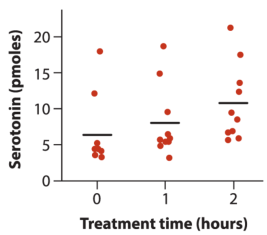
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

4

A graph tells a "thousand numbers"!



870 individual desert locusts



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

5

Why graphs?

- Powerful way of summarizing data that is easy to read (i.e., quick and direct).
- Highlight the most important information (i.e., facilitate communication).
- Facilitate (summarize) data understanding.
- Help convince others.
- Easy to remember (general trends).
- Aid in detecting unusual features in data.
- Tell stories.

6

Types of graphs

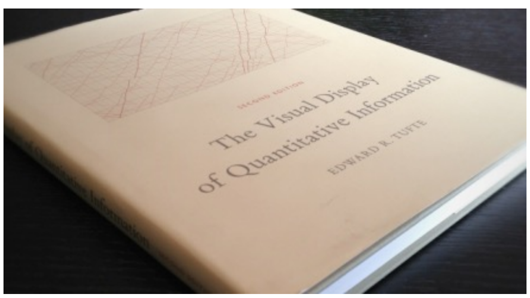
There are lots of types of graphs! The most commons (and covered in BIOL322) are:

TODAY
- Bar graph
- Pie chart
- Histogram
- Line graph
- Scatter plot
- Strip chart
- Graphs of data distributions (box plots, histograms, violin plot)

7

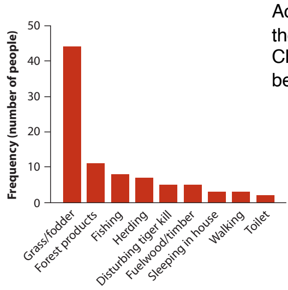
Types of graphs

There are a lots of types of graphs!



8

BAR GRAPH: Vertical or horizontal columns (bars) representing the distribution of a numerical variable against one or more categorical variable.



Activities of people at the time they were killed by tigers near Chitwan National Park (Nepal) between 1979-2006; n=88

Activity - categorical

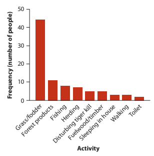
Frequency - numerical (discrete)

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

9

BAR GRAPHS are usually better than pie charts

Activities of people at the time they were killed by tigers near Chitwan National Park (Nepal) between 1979-2006; n=88



10

BAR GRAPH: Two categorical variables (often from a contingency table)

Is reproduction risky to health?

	Control group	Egg-removal group	Row total
Malaria	7	15	22
No Malaria	28	15	43
Column total	35	30	65

Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical
Frequency - numerical (discrete)

Parus major

Female birds put more energy in generating eggs to make up for those removed, thus reducing energy allocation towards immunocompetence.

11

BAR GRAPH: Two categorical variables (often from a contingency table)

Is reproduction (explanatory variable) risky to health (response variable)?

explanatory variable

		Control group	Egg-removal group	Row total
response variable	**Malaria**	7	15	22
	No Malaria	28	15	43
	Column total	35	30	65

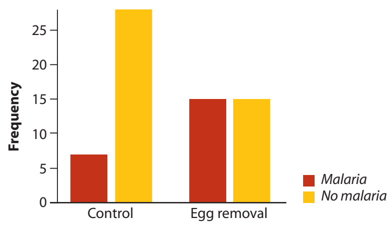
Parus major

Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical
Frequency - numerical (discrete)

egg removal forces females to produce additional eggs (i.e., increase reproduction)

12

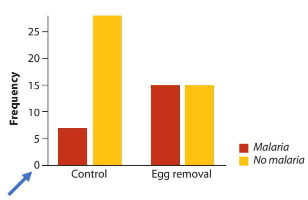
BAR GRAPH: Two categorical variables (often from a contingency table)



Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical
Frequency - numerical (discrete)

13

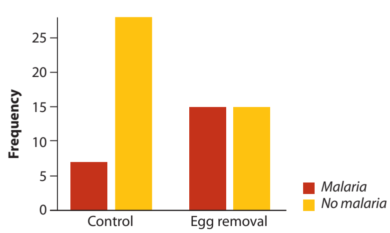
BAR GRAPH: Two categorical variables (often from a contingency table)



For bar graphs, it's generally recommended to start the measurement axis at zero to ensure the relative sizes of the bars accurately reflect the data (more on this issue at the end of this lecture).

14

Let's use this example to discuss the different types of studies & how cause and effect are established in biology.



Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical
Frequency - numerical (discrete)

15

Explanatory *versus* Response variables

- One major use of BioStatistics is to ***relate*** one variable to another, by examining associations between variables or differences between groups.

- When association between two variables is investigated, a common goal is to assess how well one of the variables, deemed the ***explanatory*** variable, *predicts* or *affects* (explain) the other variable, called the ***response*** variable.

16

Explanatory *versus* Response variables

- One major use of BioStatistics is to ***relate*** one variable to another, by examining associations between variables or differences between groups.

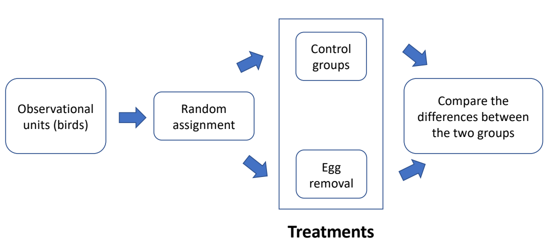
- When association between two variables is investigated, a common goal is to assess how well one of the variables, deemed the ***explanatory*** variable, *predicts* or *affects* (explain) the other variable, called the ***response*** variable.

"Assumed" explanatory power may depend on the type of study:
[1] experimental versus [2] observational studies

17

"Assumed" explanatory power may depend on the type of study

Experimental study - Researcher randomly assigns observational units (birds) to different groups (often called treatments), i.e., they control the treatments.



18

Explanatory and response variables (experiment)

When conducting an experiment (e.g., malaria study in the last slides), the treatment variable (the one manipulated by the researcher) is the ***explanatory*** variable, and the measured effect of the treatment is the ***response*** variable.

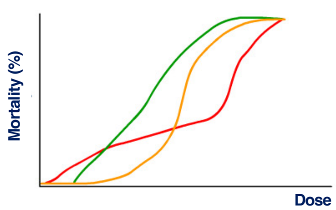
explanatory variable

		Control group	Egg-removal group	Row total
response variable	**Malaria**	7	15	22
	No Malaria	28	15	43
	Column total	35	30	65

19

Explanatory and response variables (experiment)

Another example of experiment: the administered dose of a toxin in a toxicology experiment would be the ***explanatory*** variable, and organism mortality would be the ***response*** variable.



Response to different agents (each one represented by a different color) may vary with increasing dose

https://toxlearn.nlm.nih.gov/htmlversion/module1.html

20

"Assumed" explanatory power may depend on the type of study

Observational study - Researchers have no control over which observational units fall into which treatment or values of the explanatory variable. Examples:

- Studies on the health consequences of cigarette smoking in humans (unethical to assign smoking and no-smoking treatments to observational units, i.e., people).
- Growth of fish in warm versus cold lakes (observational units, i.e., fish are already in lakes; the research has no control on which fish goes in which lake).

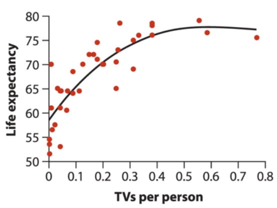
21

Let's take a break – 1 minute

22

Explanatory and response variables (observational study)

When neither variable is manipulated by the researcher (i.e., observational study; sample of convenience), their association might nevertheless be described by the "effect" of one of the variables (the explanatory) on the other (the response), even though the association itself is not direct evidence for causation.

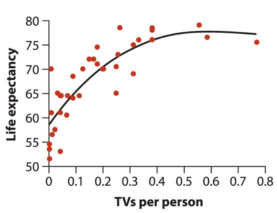


"The magic hilling powers of TV" in the US

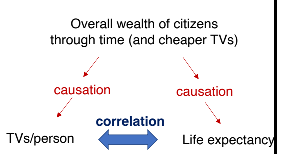
23

Explanatory and response variables (observational study)

When neither variable is manipulated by the researcher (i.e., observational study; sample of convenience), their association might nevertheless be described by the "effect" of one of the variables (the explanatory) on the other (the response), even though the association itself is not direct evidence for causation.



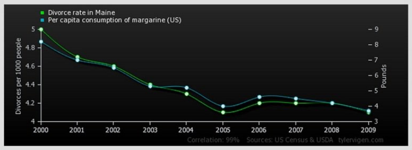
"The magic hilling powers of TV" in the US



24

Explanatory and response variables (observational study)

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Permalink - Mark as interesting - Not interesting   http://tylervigen.com/view_correlation?id=1703

25

Independent versus dependent variables = explanatory versus response variables, respectively

Strictly speaking, if one variable depends on the other, then neither is independent, so we rather say ***explanatory*** and ***response***.

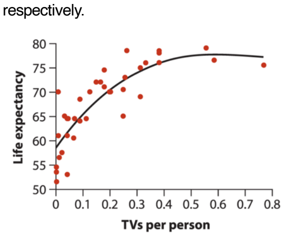
Sometimes you will hear variables referred to as "***independent***" and "***dependent***". These are the same as ***explanatory*** and ***response*** variables, respectively.

26

Independent versus dependent variables = Explanatory versus response variables, respectively

Strictly speaking, if one variable depends on the other, then neither is independent, so we rather say ***explanatory*** and ***response*** (e.g., in Whitlock and Schluter).

Sometimes you will hear variables referred to as "***independent***" and "***dependent***". These are the same as ***explanatory*** and ***response*** variables, respectively.



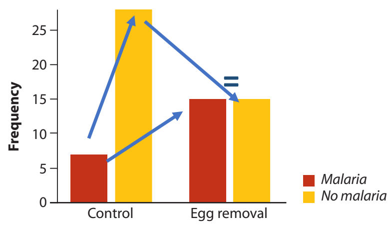
Regardless whether the association is causal, the expected explanatory variable goes in the X-axis and the expected response variable goes in the Y-axis.

27

Back to BAR GRAPHs: two categorical variables

Is reproduction risky to health?

Not so clear from this bar graph



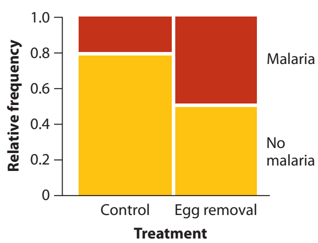
Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical

Frequency - numerical (discrete)

28

BAR GRAPHs (staked = mosaic graph): Two categorical variables

Is reproduction risky to health? Much clearer now!

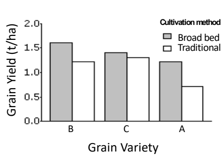


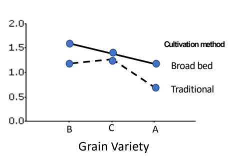
Treatment (egg removal/control) & outcome (malaria – yes/no) - categorical

Frequency - numerical (discrete)

29

BAR GRAPHS are not always the best way! (these graphs are based on the same data)





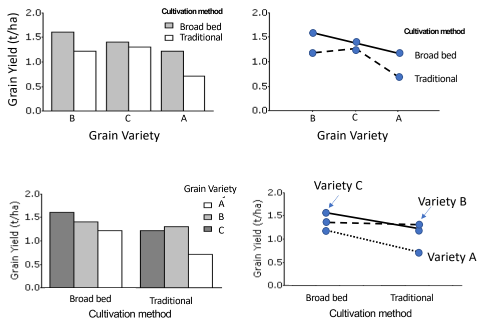
Traditional (continuous; non-spaced)

Broad bed (spaced)

- Source - Biometrics Advisory and Support Service to DFID

30

BAR GRAPHS are not always the best way!
which variable goes where? What message do you want to "sell"?



demonstrates better differences between cultivation methods

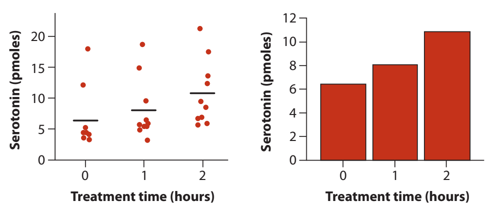
demonstrates better differences between grain varieties

- Source - Biometrics Advisory and Support Service to DFID

31

Different graphs produce different amounts of information

The scale of the response variable has been changed to better adapt to the data presentation



lines represent mean values
dots represent the data

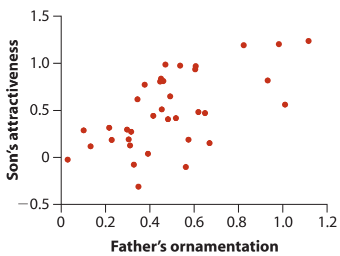
bar heights mean values

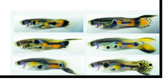
32

Let's take a break - 1 minute

33

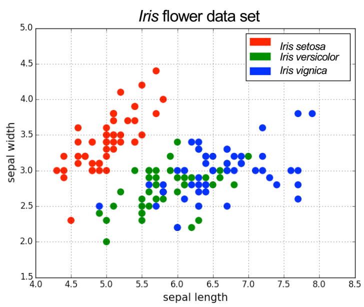
SCATTER PLOT: graphical display of two numerical variables in which each observation is represented as point on a graph with





34

SCATTER PLOT: multiple series – which species vary more from the others?

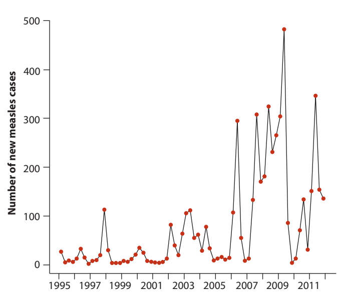


Sepal between petals

https://en.wikipedia.org/wiki/Iris_flower_data_set (collected by Edgar Anderson, Gaspé peninsula, data introduced by R. Fisher)

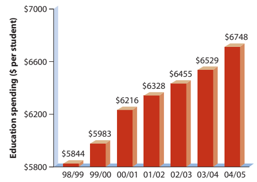
35

LINE GRAPH: uses dots connected by line segments to display trends in a measurement over time or other ordered states (e.g., size, etc).



36

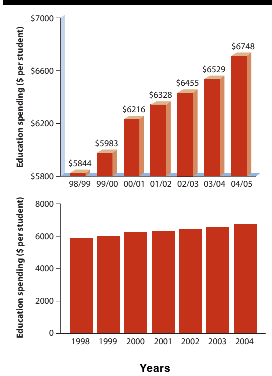
Graphs can be misleading! (both based on the same data): bad axis limits



Upper graph. A bar graph, taken from a British Columbia government brochure, indicating education spending per student in different years. *Lower graph:* A revised presentation of the same data, in which the magnitude of the spending is proportional to the height and area of bars. This revision also removed the 3-D effects and the numbers above bars to make the pattern easier to see. The upper graph is modified from British Columbia Ministry of Education (2004).

37

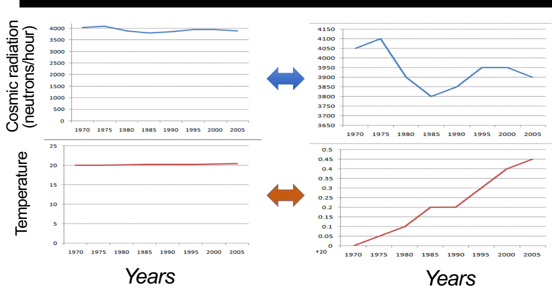
Graphs can be misleading! (both based on the same data): bad axis limits



Upper graph. A bar graph, taken from a British Columbia government brochure, indicating education spending per student in different years. *Lower graph:* A revised presentation of the same data, in which the magnitude of the spending is proportional to the height and area of bars. This revision also removed the 3-D effects and the numbers above bars to make the pattern easier to see. The upper graph is modified from British Columbia Ministry of Education (2004).

38

Report the "intended" interpretation!



Axis limits do not change the information but rather the interpretation that one wants to convey with the same data

The 4 graphs are from "Rick Ord, University of California San Diego, 2010, a slide from the course on Fluency with Information Technology".

39

Some “rules” of Data visualization

40

How to make a *bad* plot:

1. Hide the data.
2. Make patterns hard to see.
3. Display magnitudes dishonestly.
4. Draw graphics unclearly.

How to make a *good* plot:

1. Show the data.
2. Make patterns easy to see.
3. Display magnitudes honestly.
4. Draw graphics clearly.

41

Mistakes in displaying data

Mistake 1. Hide the data

42

Mistake 1: Hide the data

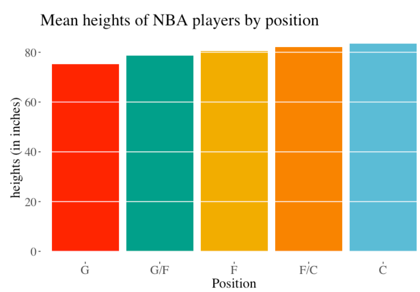
How to hide data:
- Provide only statistical summaries (e.g., means).

How to reveal data:
- Present all data points, while allowing all to be seen.

43

Not Showing Data, Just Summaries

This plot hides the variation within positions.

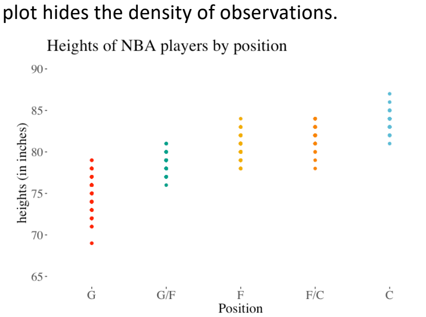


center, forward, guard

44

Not Showing Data, Over-Plotting

This plot hides the density of observations.

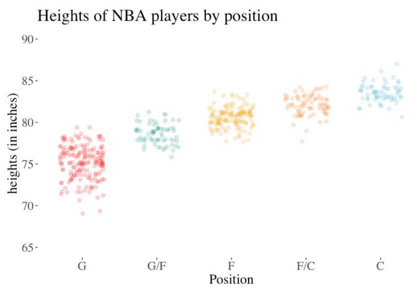


center, forward, guard

45

Showing Data, Jittering

This plot shows all the observations.



center, forward, guard

46

Mistakes in displaying data

Mistake 2. Making patterns hard to see

47

Mistake 2: Making Patterns Hard to See

How to hide patterns:

- Make one plot and call it good.
- Use unreasonable scales.
- Arrange factors nonsensically.

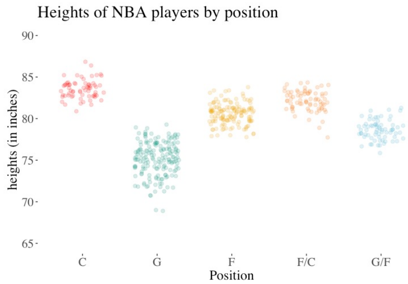
How to reveal patterns:

- Explore multiple potential plots.
- Use appropriate scales.
- Arrange factors meaningfully.
  Arrange in order for ordinal, by mean for nominal.

48

Nonsensical Order Hides Patterns

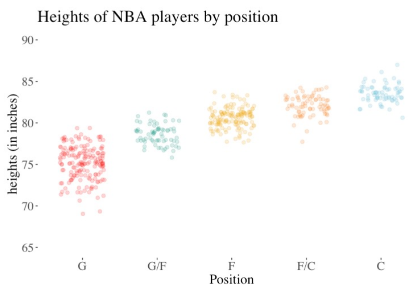
Non-intuitive ordering of factors hides patterns.



49

Nonsensical Order Hides Patterns

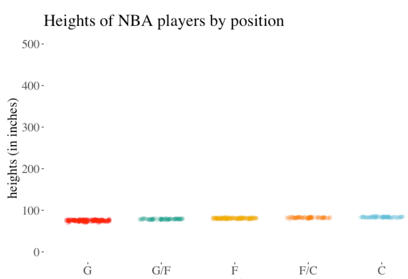
Intuitive ordering of factors make patterns more evident



50

Bad Axis-Limits Hide Patterns

In this plot, the large scale (limits of the Y-axis) hides the pattern.



51

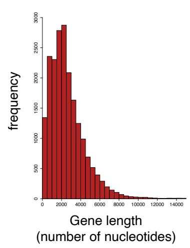
Graphs:
The art of designing information

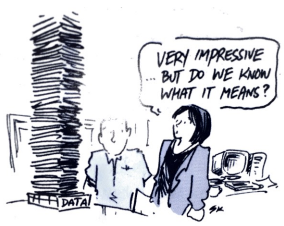
"A picture tells a thousand words"

- Lake Blanche

52

Next lecture: How to build frequency distributions and introduction to descriptive (or summary) statistics





53