Introduction to Data Science

Statistical Charts

Components of Statistics

- A general process of investigation:
  - 1. Identify a question or problem.
  - 2. Collect relevant data on the topic.
  - 3. Analyze the data.
  - 4. Form a conclusion.

- Statistics is the study of how best to collect, analyze, and draw conclusions from data (stages 2-4).
  - How best can we collect data?
  - How should it be analyzed?
  - What can we infer from the analysis?

Data Matrix

Variable
↓

email	spam	num_char	line_breaks	format	number
1	No	21705	551	html	small
2	No	7011	183	html	big
3	Yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	No	15829	242	html	small

⟵ *Observation (case)*

Types of Variables

Variables

Numerical

(quantitative)

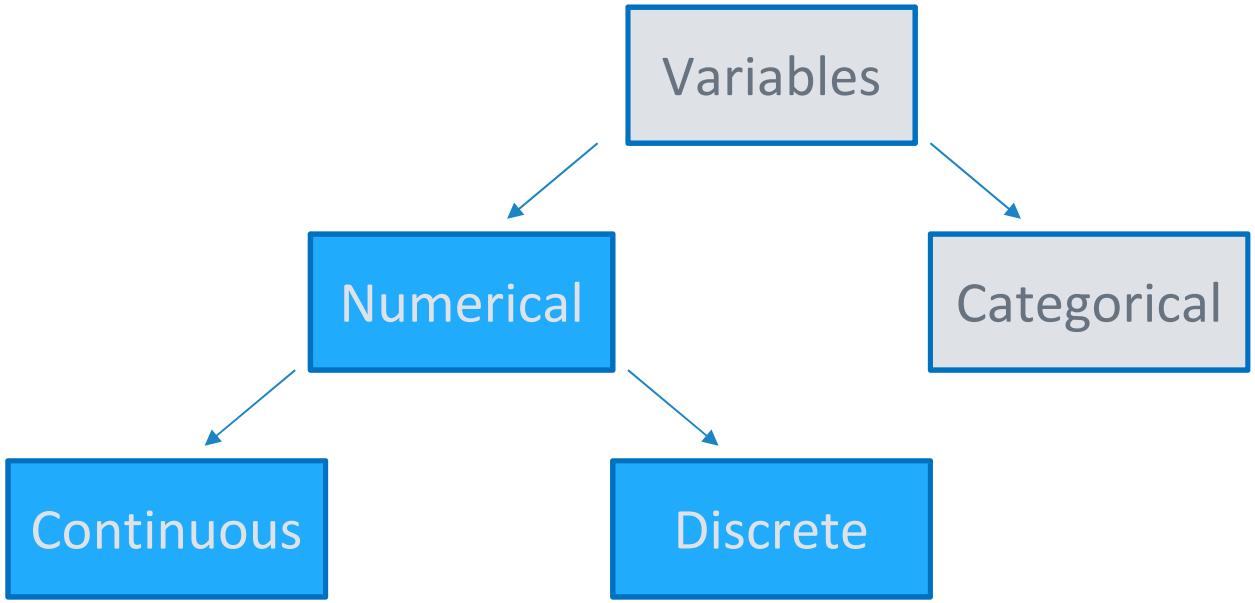
- take on numerical values
- sensible to add, subtract, take averages, etc. with these values

Categorical

(qualitative)

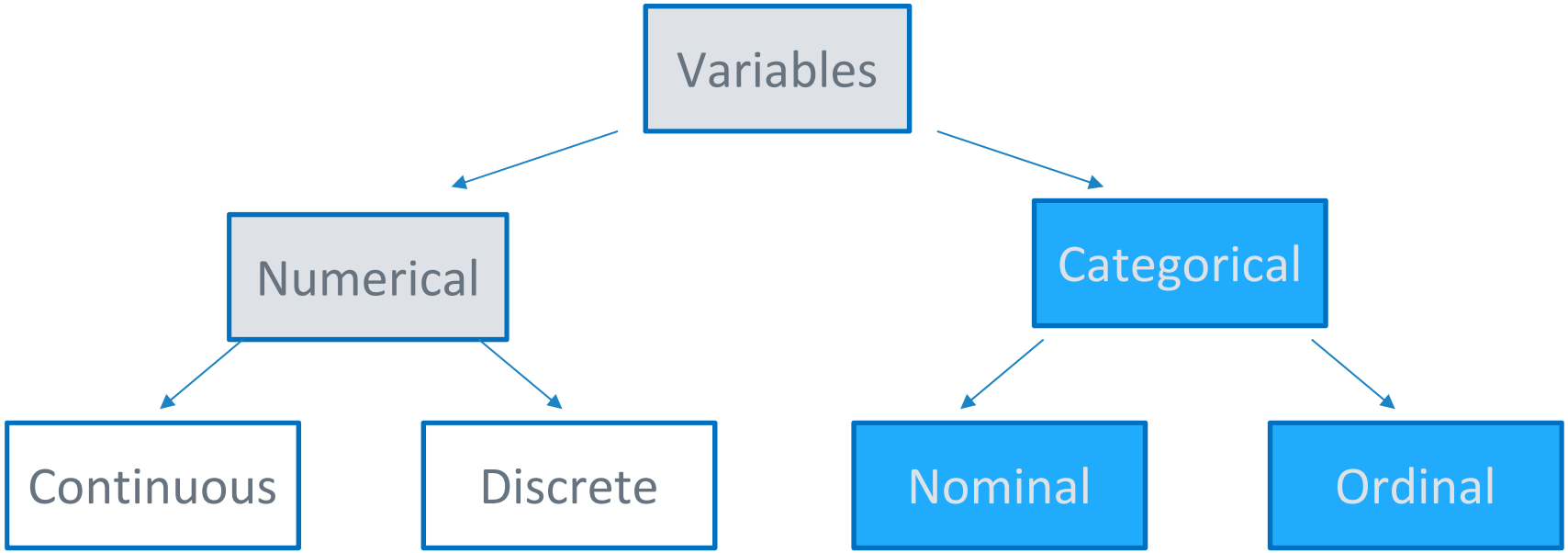
- take on a limited number of distinct categories.
- categories can be identified with numbers, but not sensible to do arithmetic operations

Numerical Variables



- Take on any of an infinite number of values within a given range
- Take on one of a specific set of numeric values

Categorical Variable



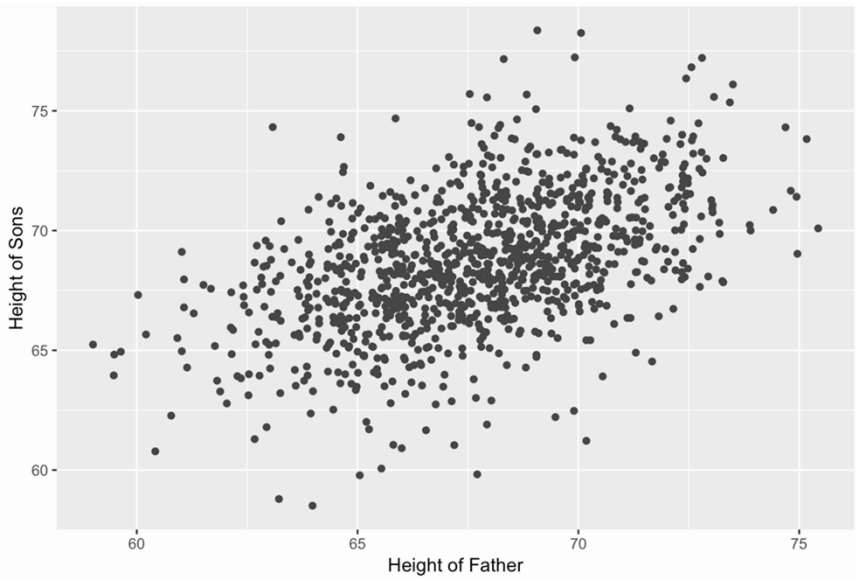
- Levels have an inherent ordering

Example

email	spam	num_char	line_breaks	format	number
1	No	21705	551	html	small
2	No	7011	183	html	big
3	Yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	No	15829	242	html	small
↓ Identity	↓ Nominal Categorical	↓ Discrete Numerical	↓ Discrete Numerical	↓ Nominal Categorical	↓ Ordinal Categorical

Relationships between variables

- Two variables that show some connection with one another are called associated.
- Association can be further described as positive or negative.
- If two variables are not associated, they are said to be independent.



Height of fathers and sons

Population and Sample

Population

- Each research question refers to a target population.

- Example:
  - *Research question:* Can adult men become better, more efficient runners on their own, merely by running?
  - *Population of interest: All men over 18*

- Often it is too expensive to collect data for every case in a population.

Census

- Census: collect data from *everyone* in the population.

رئیس مرکز آمار ایران:

هزینه سرشماری سال 1395 پنج هزار میلیارد ریال است/آغاز سرشماری نفوس از سوم مهر

Sampling

- A sample represents a subset of the cases and is often a small fraction of the population.
- Think about sampling something you are cooking: you taste a small part of what you're cooking to get an idea about the dish as a whole.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.

Anecdotal Evidence

- Consider the following statements:
  - My uncle smokes three packs a day and he's in perfectly good health, so smoking doesn't affect your health.

- The conclusion is based on data, but there are two problems:
  - First, the data only represent one or two cases.
  - Second, it is unclear whether these cases are actually representative of the population.

- Data collected in this haphazard fashion are called anecdotal evidence.

Sampling Bias

Some Sources of Sampling Bias

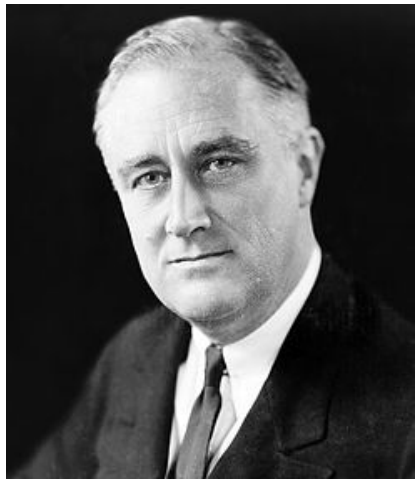
- *Non-response:* If only a *non-random* fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.
- *Convenience sample:* Individuals who are easily accessible are more likely to be included in the sample.

Sampling Bias Example

- A historical example of a biased sample yielding misleading results:

Alf Landon

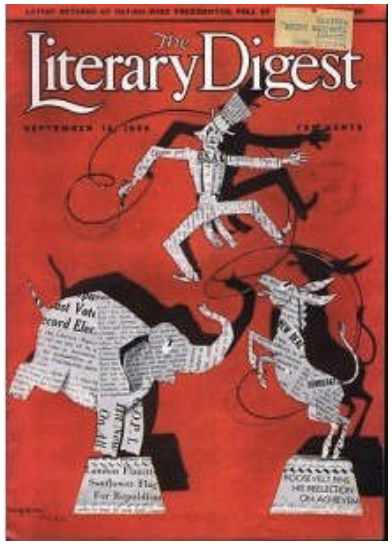
- In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



Franklin D. Roosevelt

The Literary Digest Poll

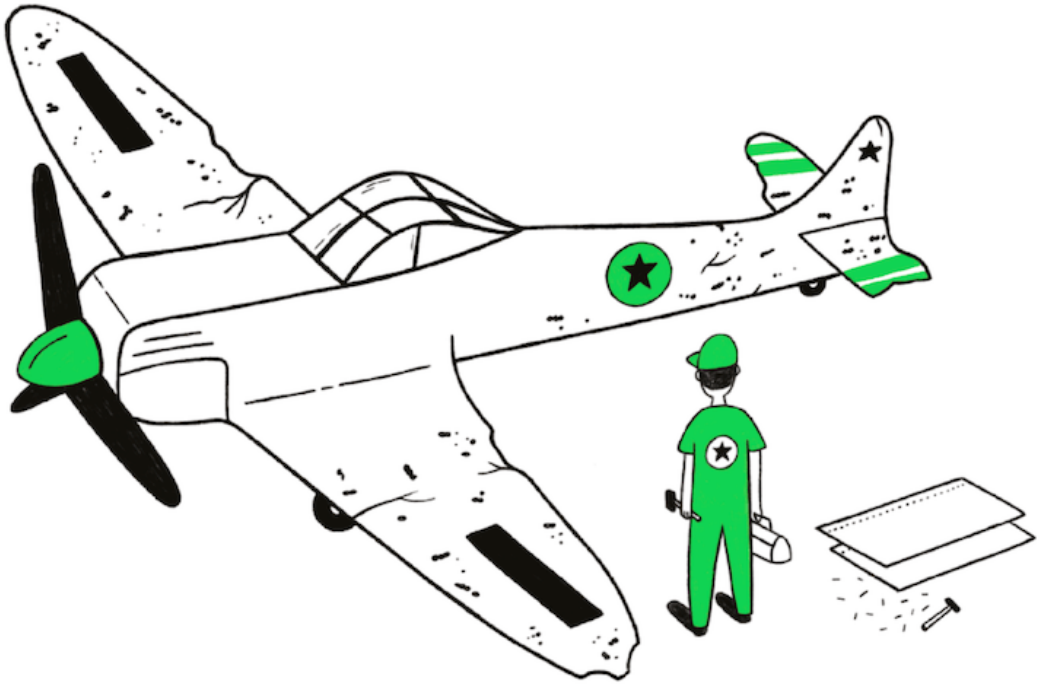
- The Literary Digest polled 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the winner and FDR would get 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.



What went wrong?

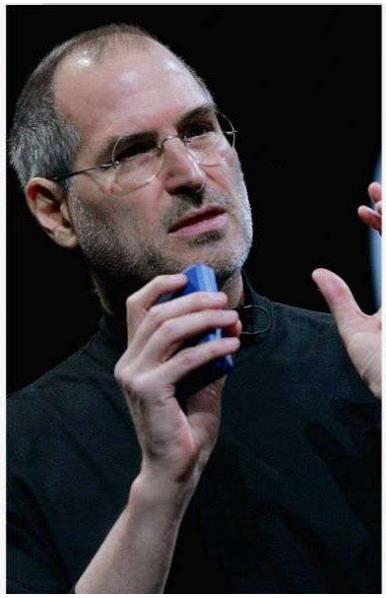
- The magazine had surveyed:
  - its own readers
  - registered automobile owners, and registered telephone users
- These groups had incomes well above the national average of the day which resulted in lists of voters far more likely to support Republicans than a truly *typical* voter of the time.
- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was *biased*, the sample did not yield an accurate prediction.

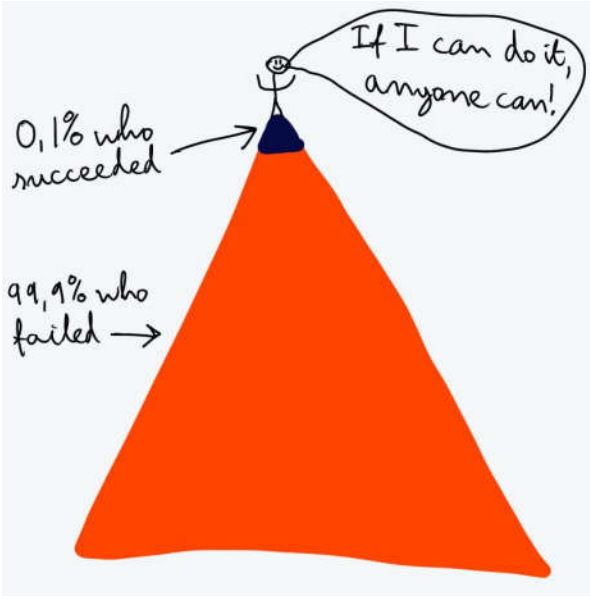
Survivorship Bias



Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

Survivorship Bias



Always ask: **"What data are we not seeing?"**

Type of Studies

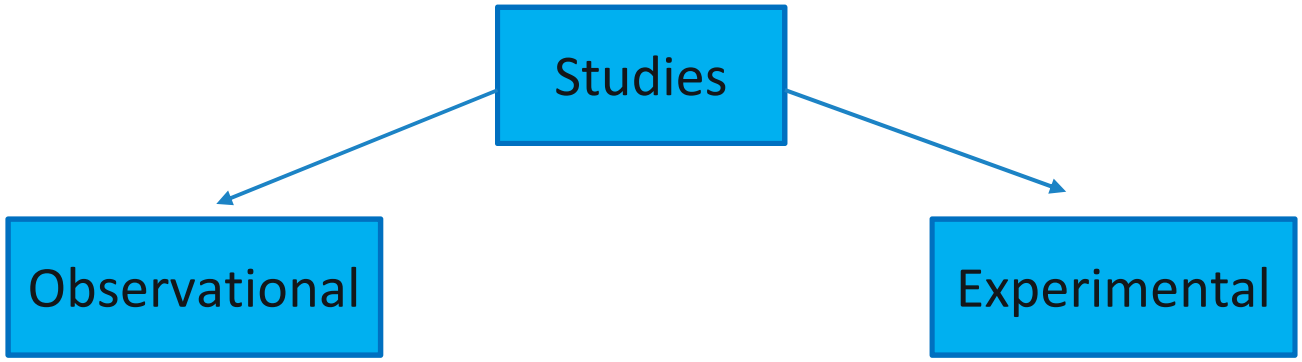
Explanatory and Response Variables

- To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{Explanatory variable} \xrightarrow{\text{might affect}} \text{Response variable}$$

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is a high correlation between the two variables.
- We use these labels only to keep track of which variable we suspect affects the other.

Observational Studies & Experiments

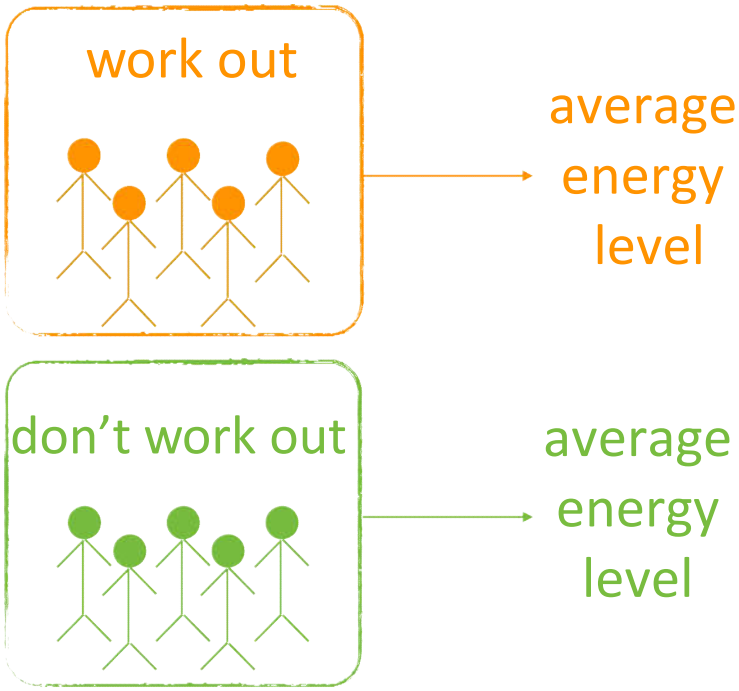


- collect data in a way that does not directly interfere with how the data arise ("observe")
- only establish an association
- retrospective: uses past data
- prospective: data are collected throughout the study

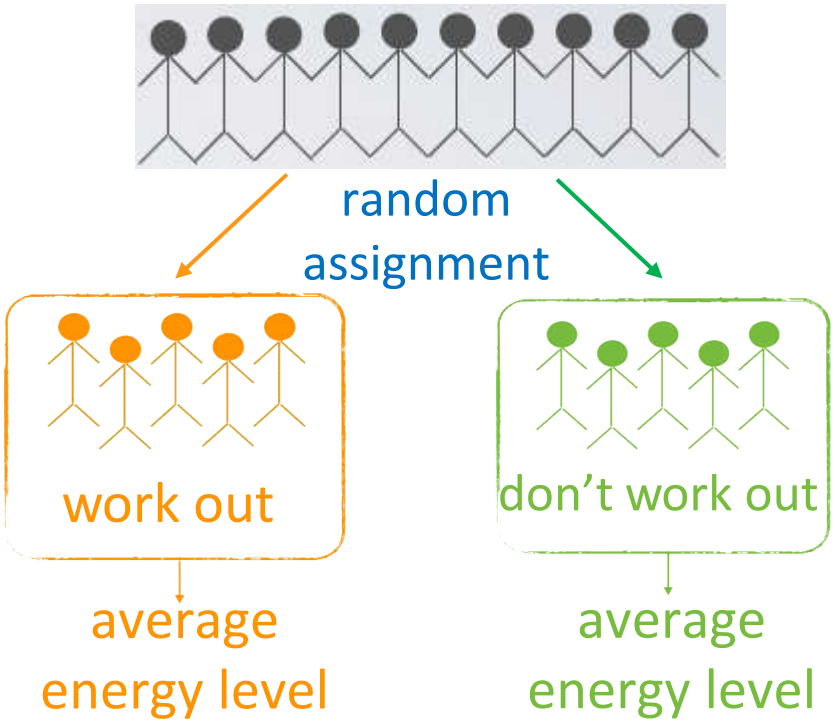
- randomly assign subjects to treatments
- establish causal connections between explanatory and response variables.

Observational vs. Experimental Studies

Observational Study

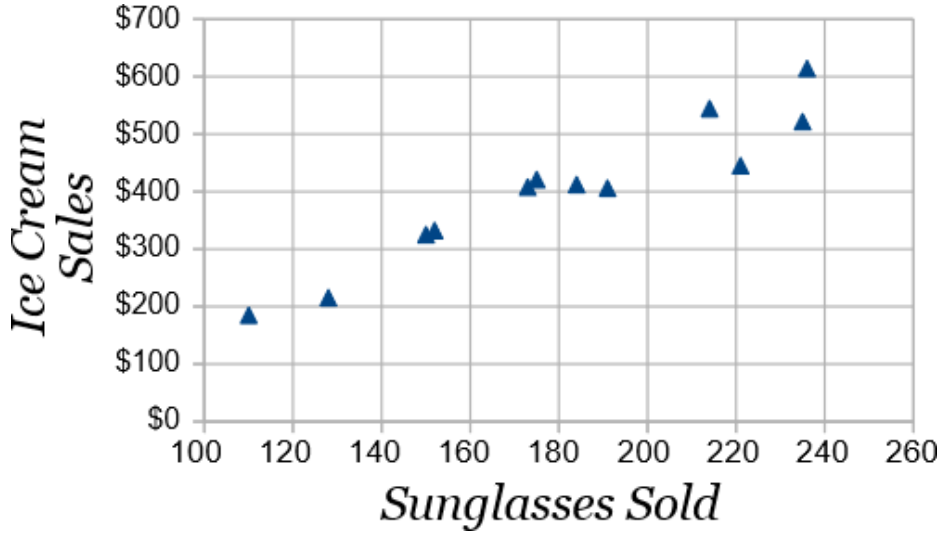


Experiment



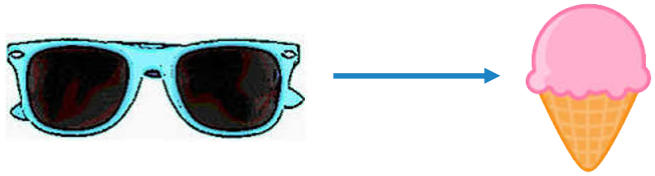
Correlation does not imply causation

- The local ice cream shop keeps track of how much ice cream they sell.
- The ice cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales.

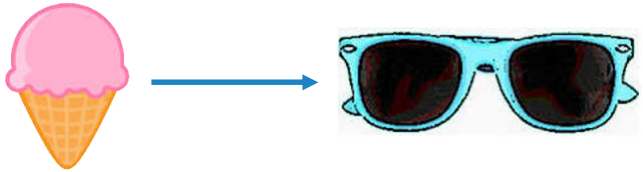


Three possible explanations

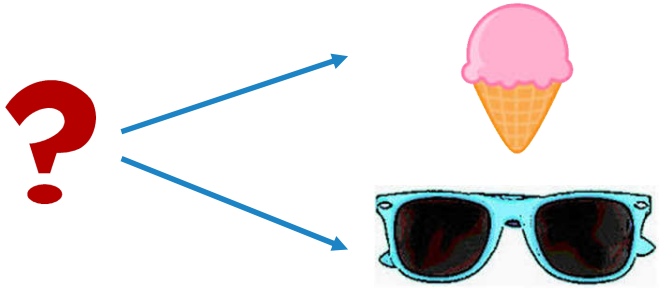
1. Sunglasses make people want ice cream!



2. Eating ice cream makes people buy sunglasses!

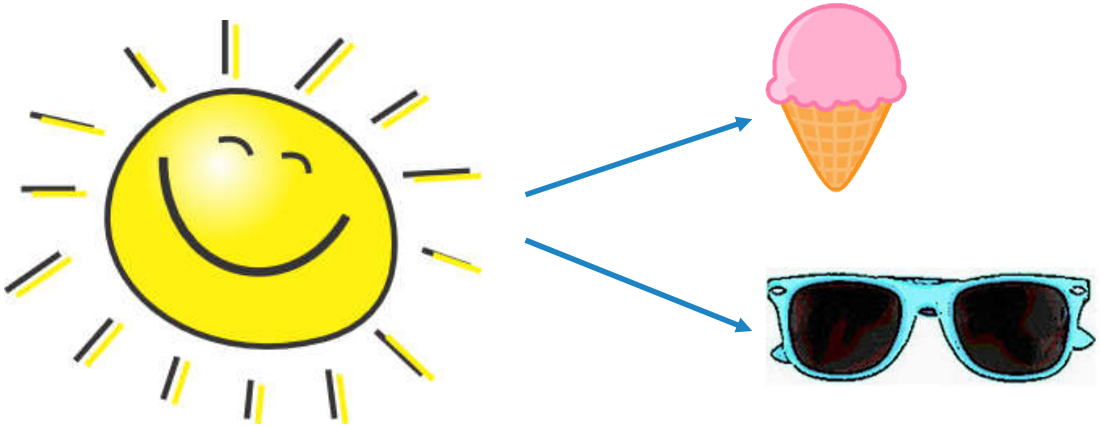


3. A third variable is responsible for both.



Confounding Variable

- An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called confounders or confounding variables.



MMR Vaccination and Autism

THE LANCET

Log in Register Subscribe Claim

EARLY REPORT | VOLUME 351, ISSUE 9103, P637-641, FEBRUARY 28, 1998

PDF [942 KB] Figures Save Share Reprints Request

RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Dr AJ Wakefield, FRCS • SH Murch, MB • A Anthony, MB • J Linnell, PhD • DM Casson, MRCP • M Malik, MRCP • et al.
Show all authors

Published: February 28, 1998 • DOI: https://doi.org/10.1016/S0140-6736(97)11096-0

PlumX Metrics

Summary
Introduction
Patients and methods
Results
Discussion
References
Article Info
Figures

Request Your Institutional Access

Summary

Background

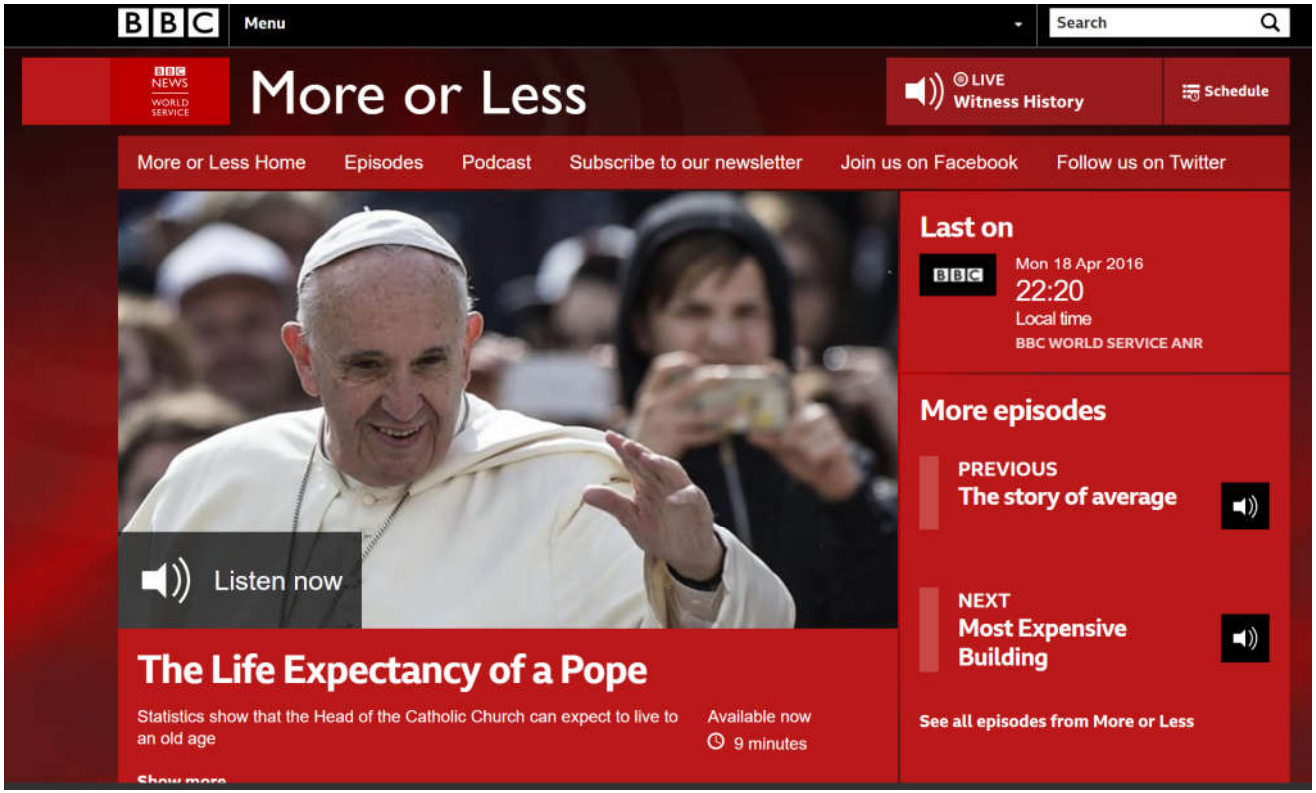
We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods

12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent

RETRACTED

Do popes live longer?



Left-handedness and Life Expectancy

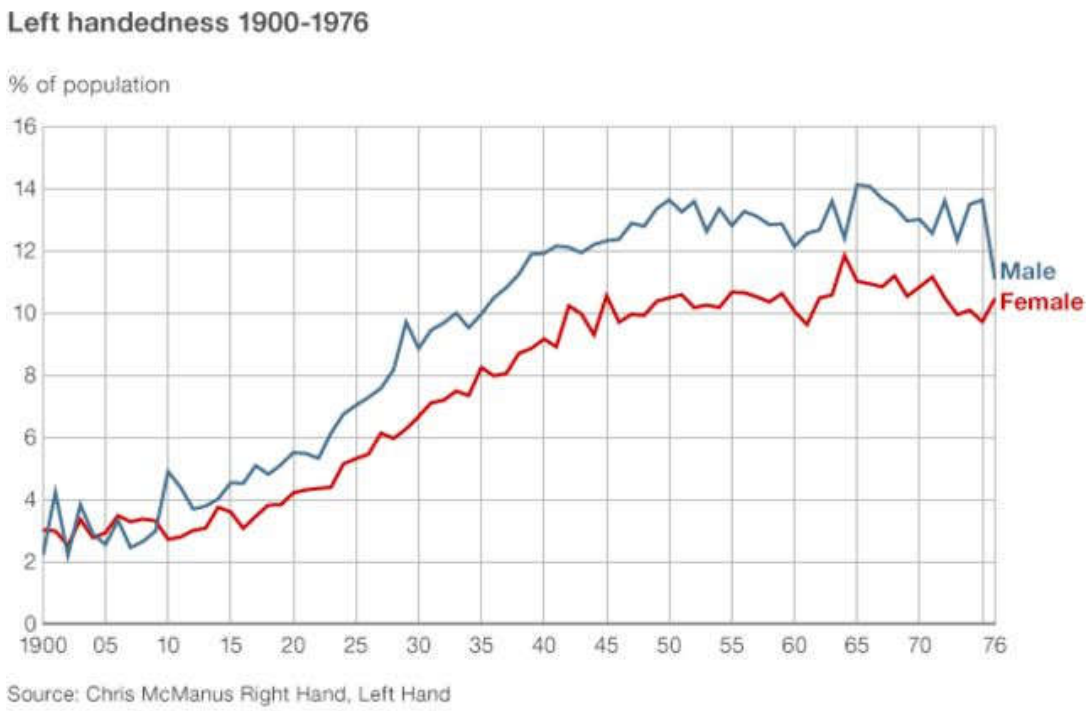
The New York Times

Being Left-Handed May Be Dangerous To Life, Study Says

Reuters

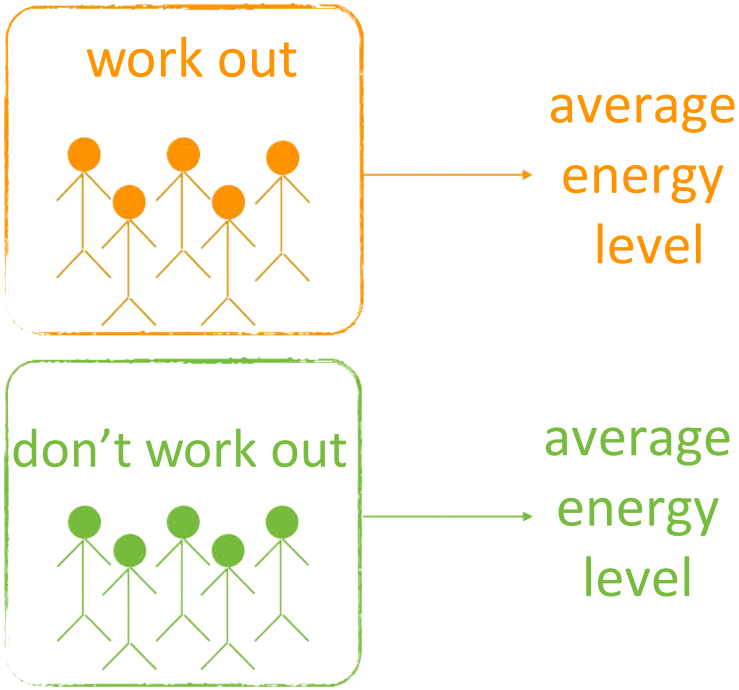
April 4, 1991

Left-handedness and Life Expectancy

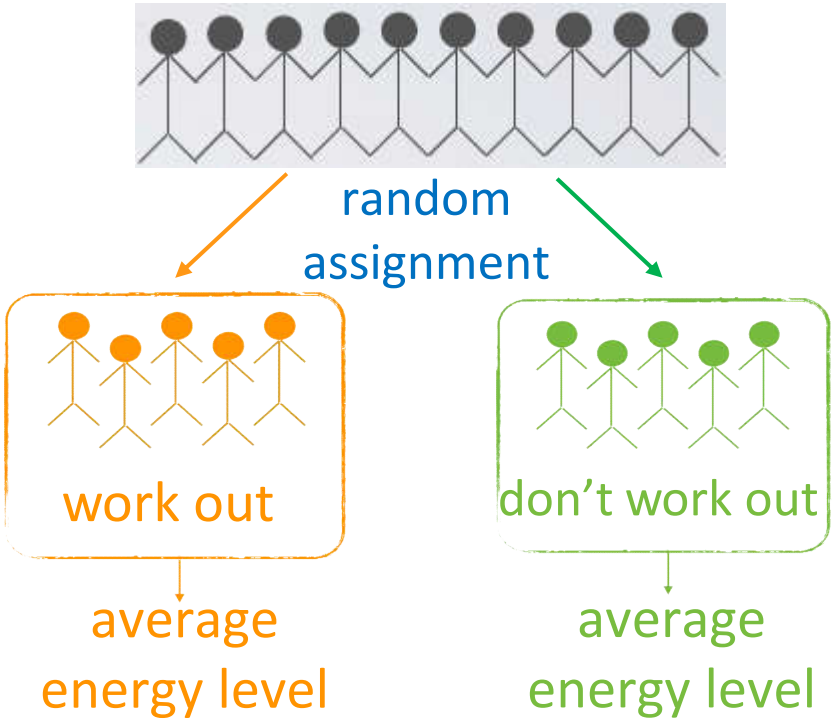


Observational vs. Experimental Studies

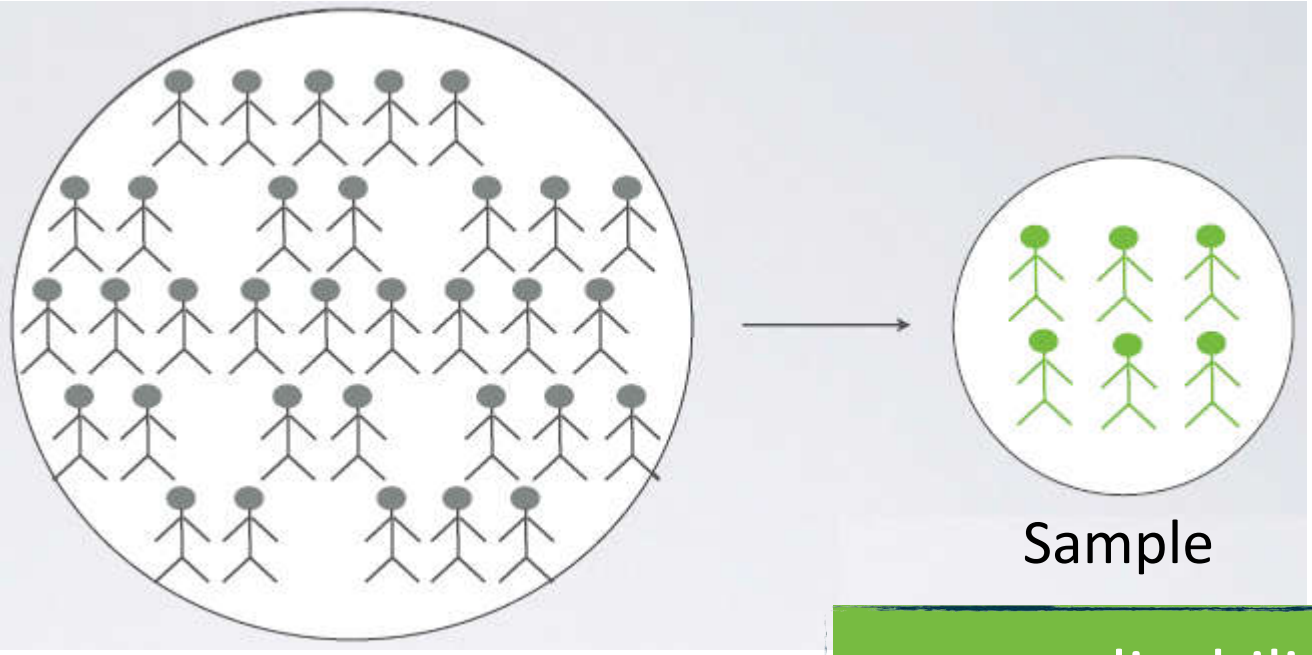
Observational Study



Experiment

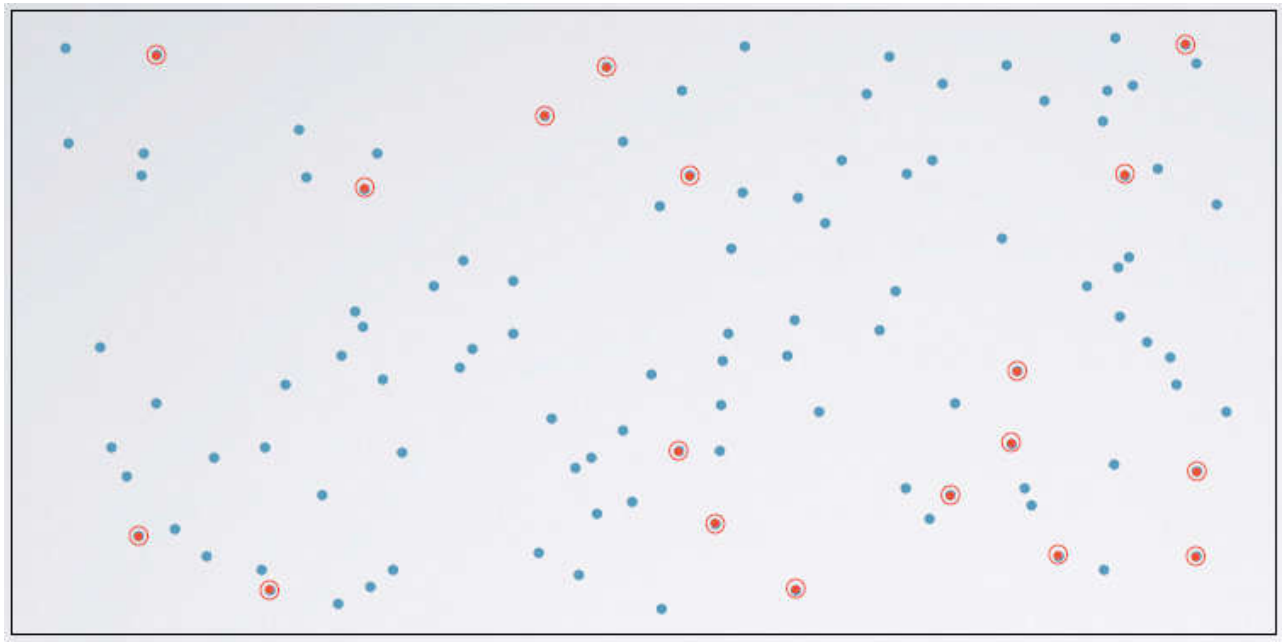


Random Sampling



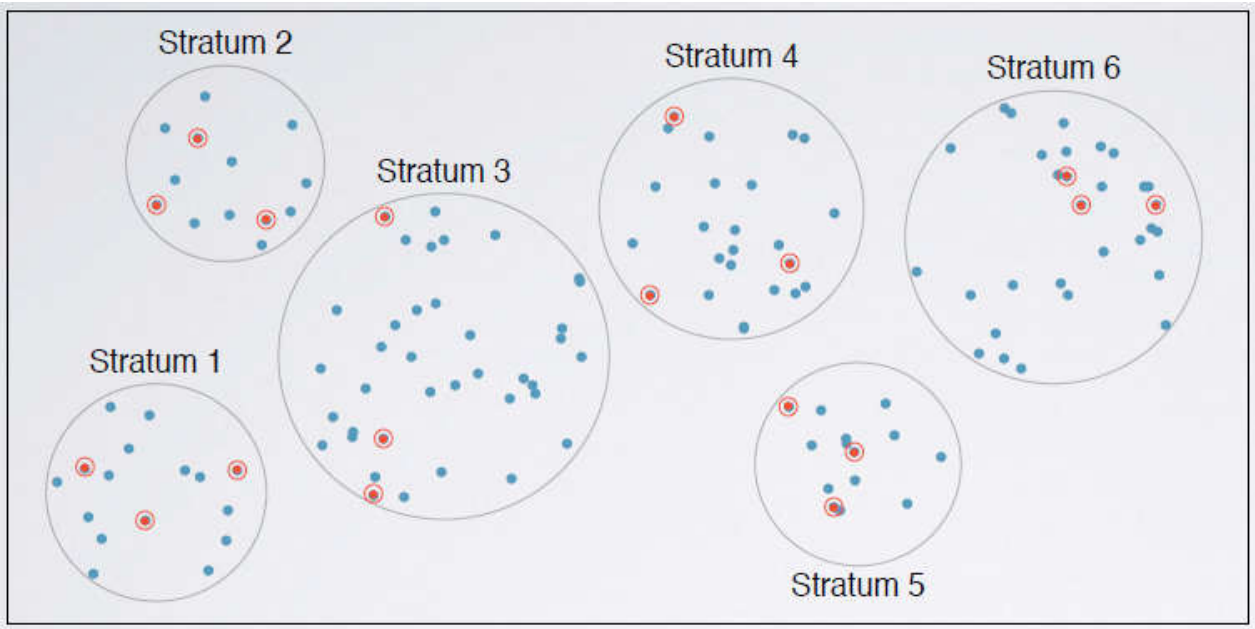
generalizability

Simple Random Sampling (SRS)



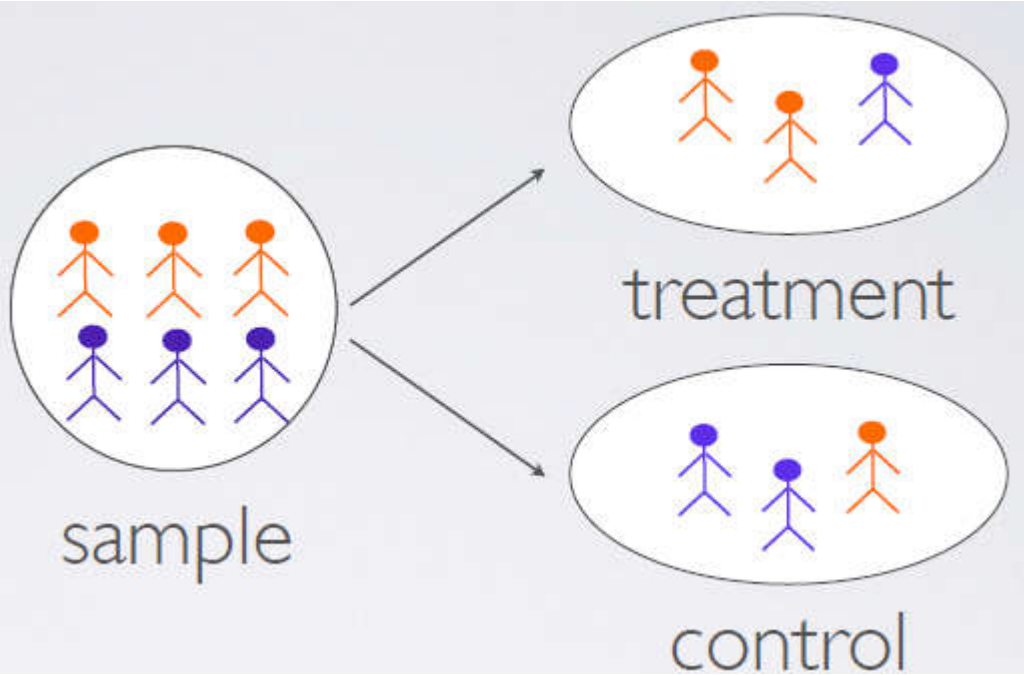
- Each case is equally likely to be selected.

Stratified Sampling



- Divide the population into homogenous strata, then randomly sample from within each stratum.

Random Assignment



Causality

Principles of Experimental Design

- *Control:* Compare treatment of interest to a control group.
- *Randomize:* Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
- *Replicate:* Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
- *Block:* If there are variables that are known or suspected to affect the response variable, first group subjects into *blocks* based on these variables, and then randomize cases within each block to treatment groups.

Random Assignment vs. Random Sampling

ideal experiment	Random assignment	No random assignment	most observational studies
Random sampling	Causal conclusion, generalized to the whole population.	No causal conclusion, correlation statement generalized to the whole population.	Generalizability
No random sampling	Causal conclusion, only for the sample.	No causal conclusion, correlation statement only for the sample.	No generalizability
most experiments	Causation	Correlation	bad observational studies

A/B Testing for US Presidential Campaign

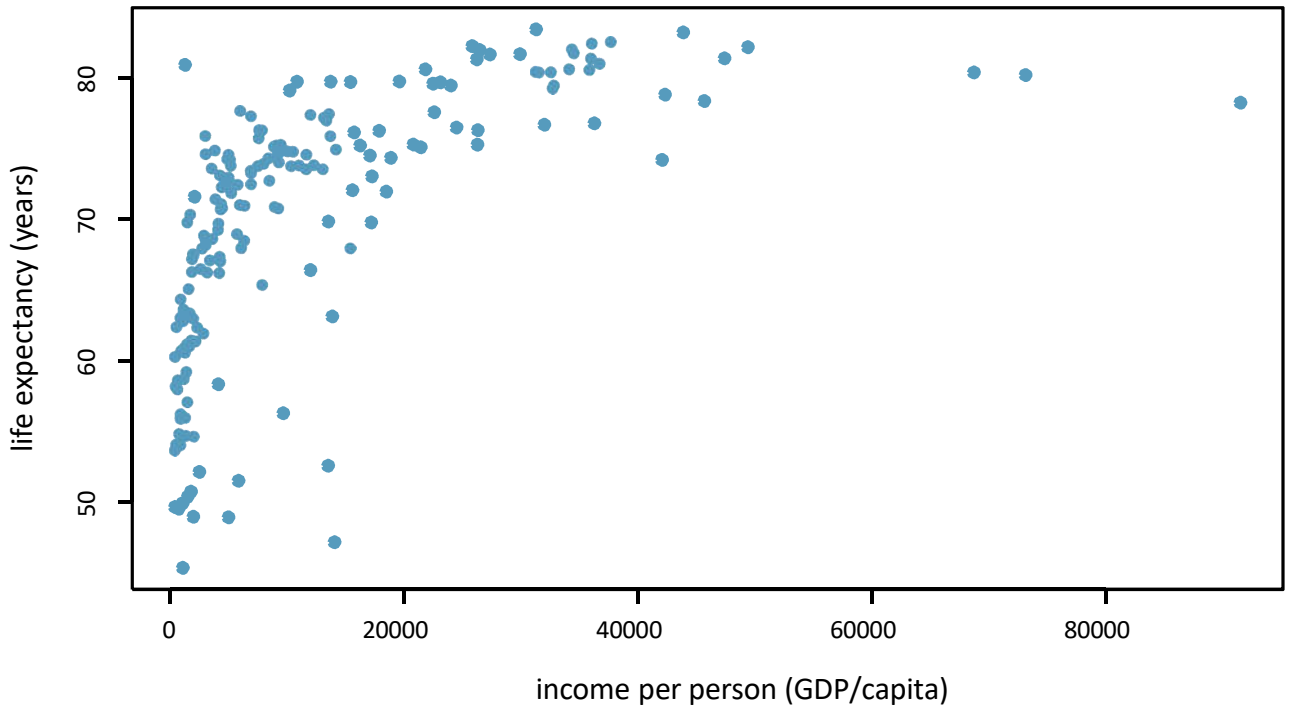
The Winner


OBAMA'08
CHANGE
WE CAN BELIEVE IN
JOIN THE MOVEMENT
Email Address
Zip Code
LEARN MORE
PAID FOR BY OBAMA FOR AMERICA
CONTINUE to WEBSITE

Visualizing Numerical Data

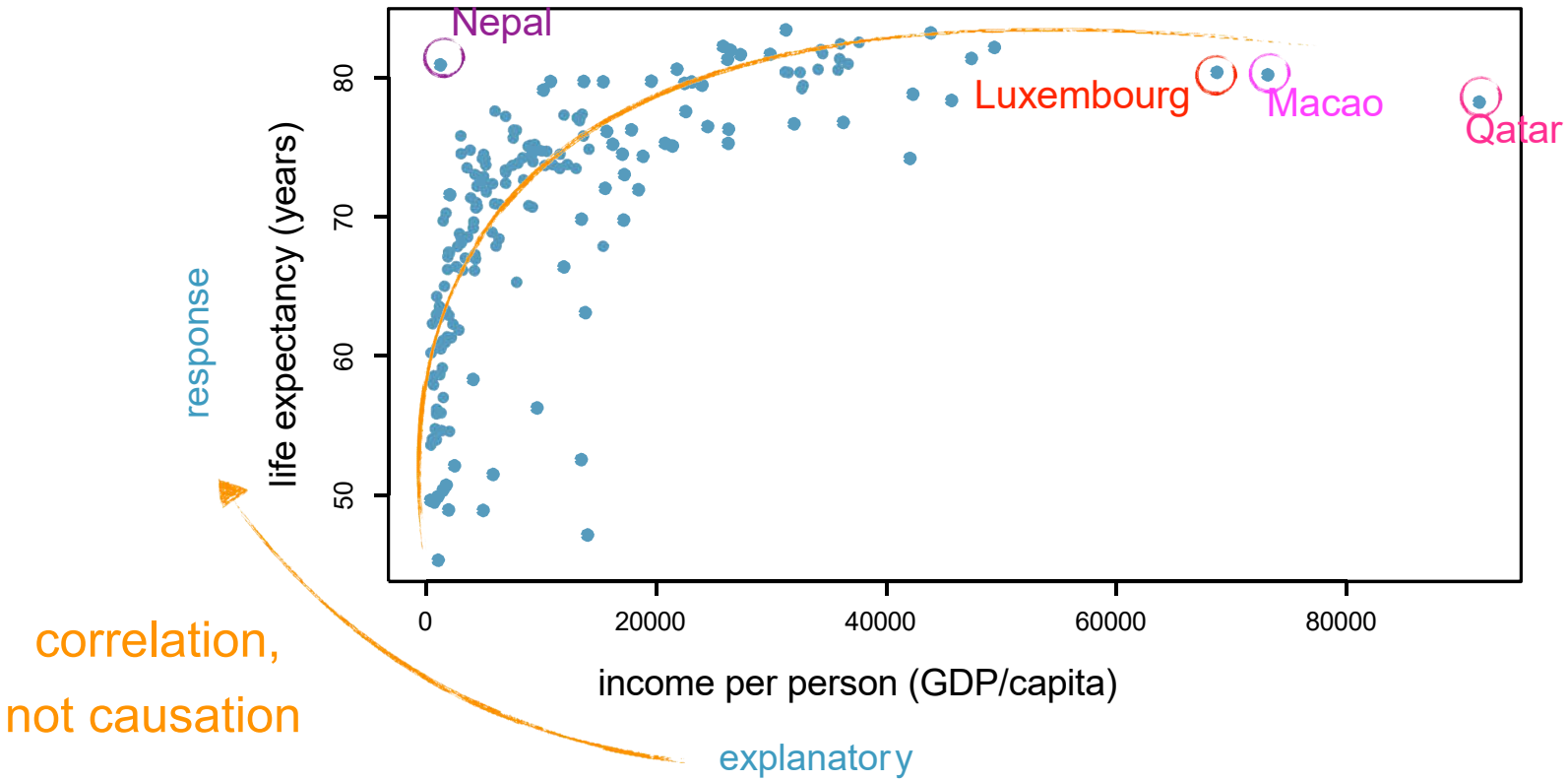
Scatterplot

data	income /person	life expectancy
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
⋮	⋮	⋮
Zimbabwe	545.3	58.142



- *Scatterplots* are useful for visualizing the relationship between two numerical variables.

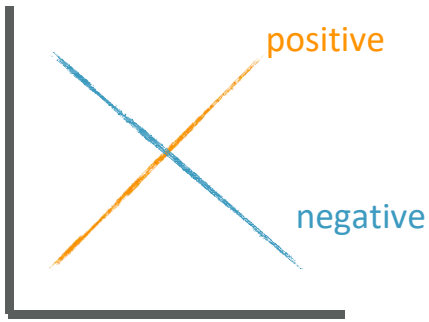
Scatterplot



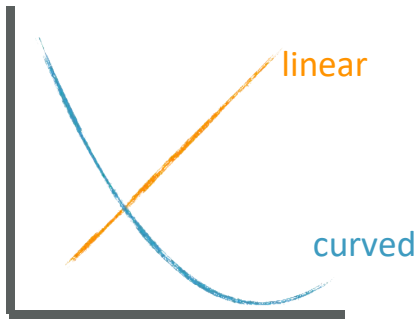
correlation,
not causation

Evaluating the relationship

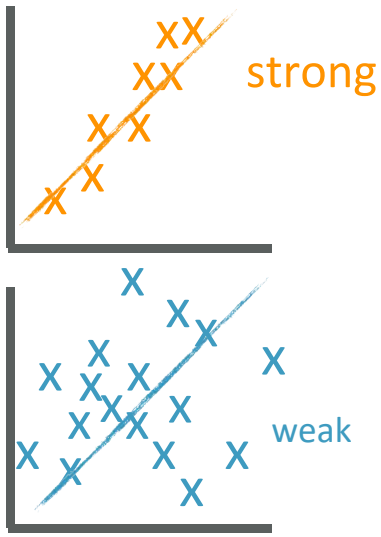
direction



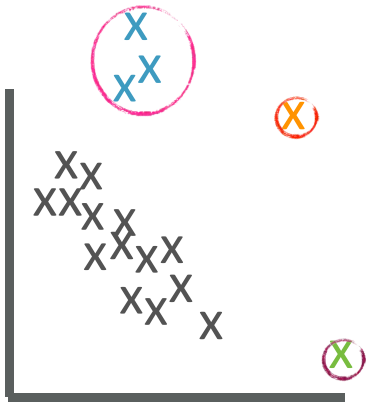
shape



strength

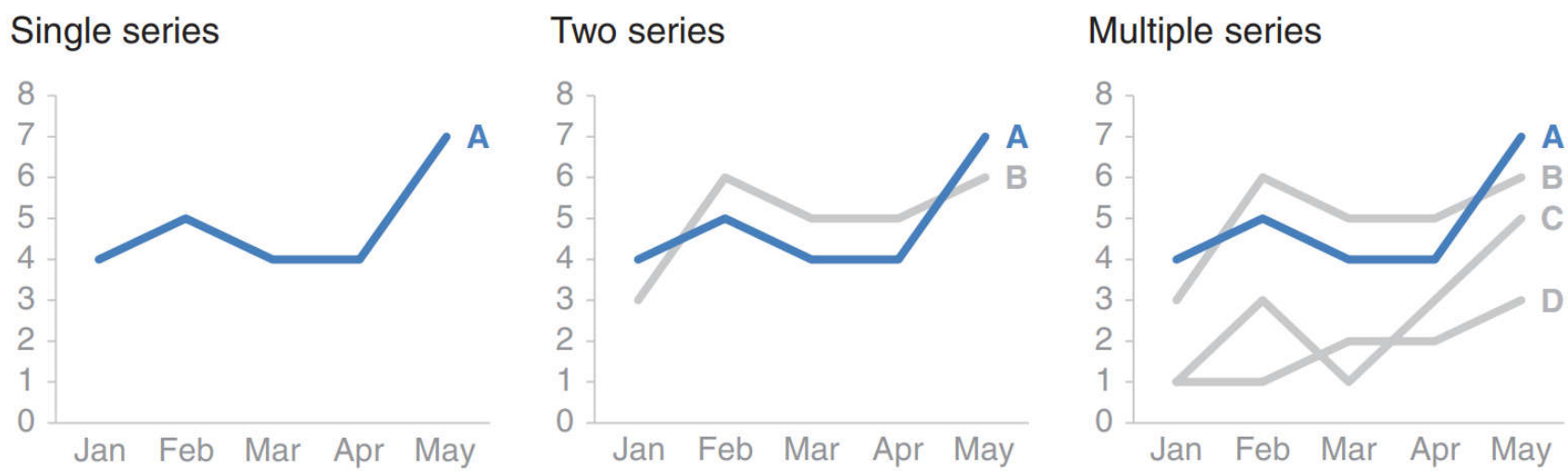


outliers

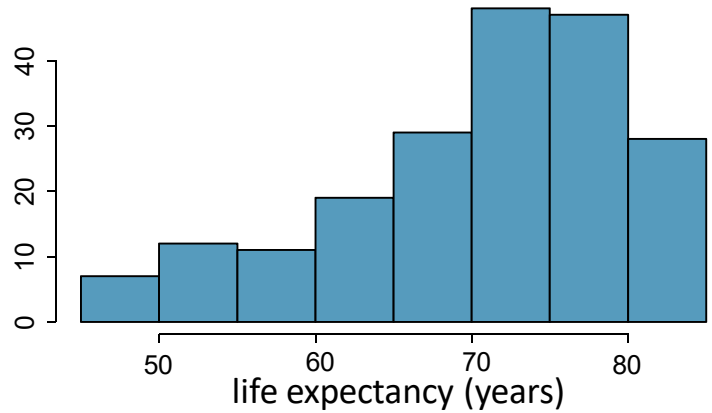


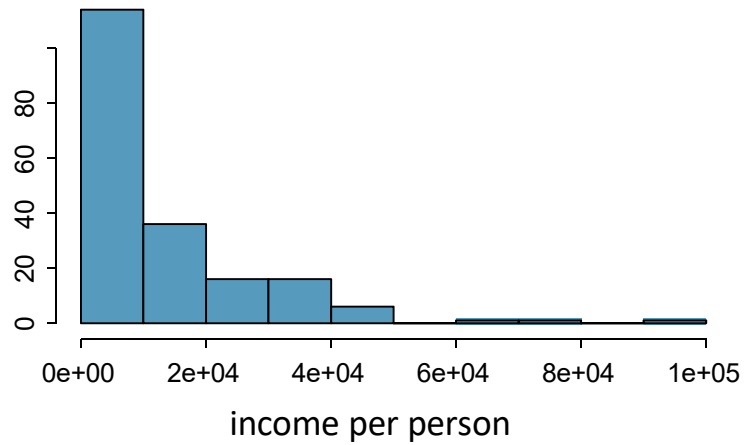
Line Graph

- Line graphs are used to plot continuous data often in some unit of time.



Histogram

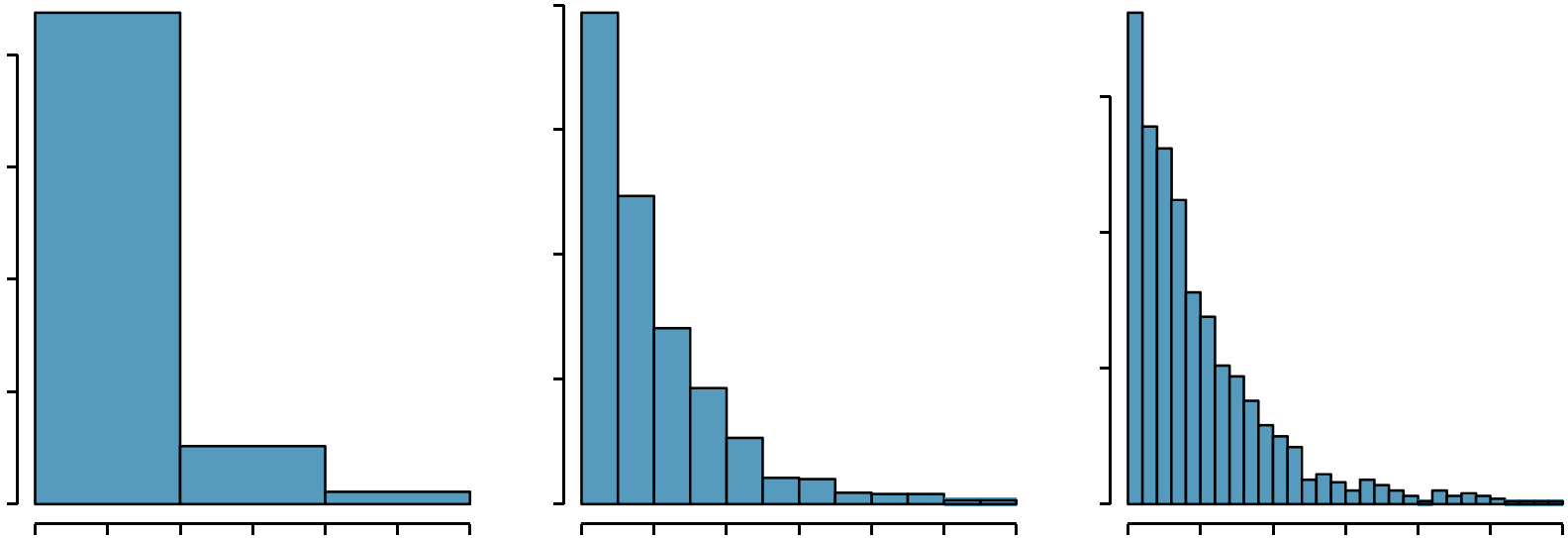




- Histograms provide a view of the data density.
- Histograms are especially convenient for describing the shape of the data distribution.
- The chosen bin width can alter the story the histogram is telling

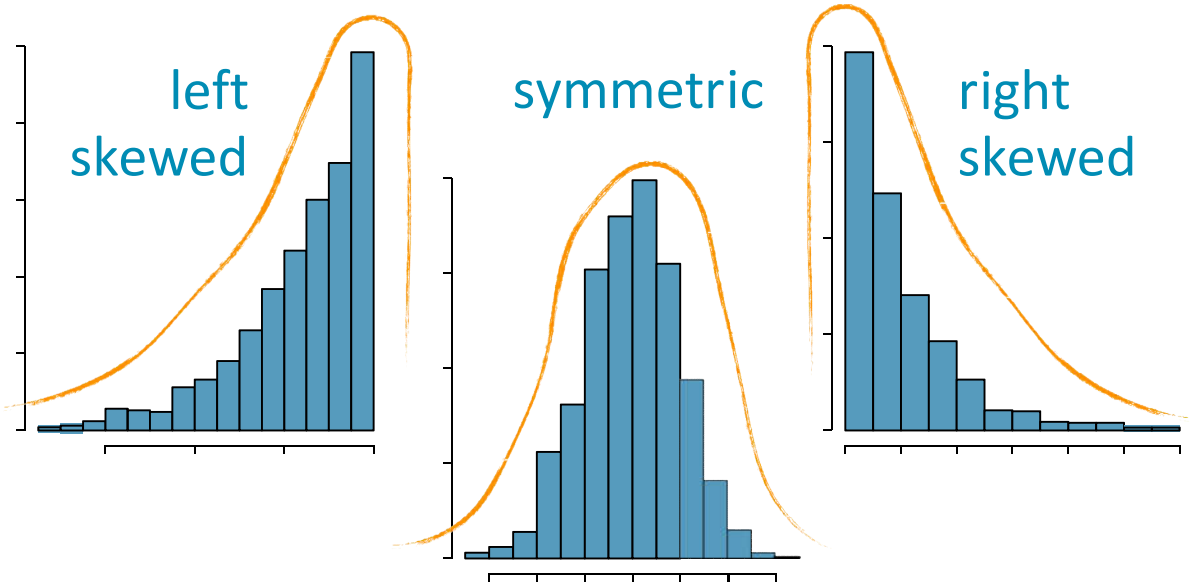
Bin Width

- When the bin width is too wide, we might lose interesting details.
- When the bin width is too narrow, it might be difficult to get an overall picture of the distribution.

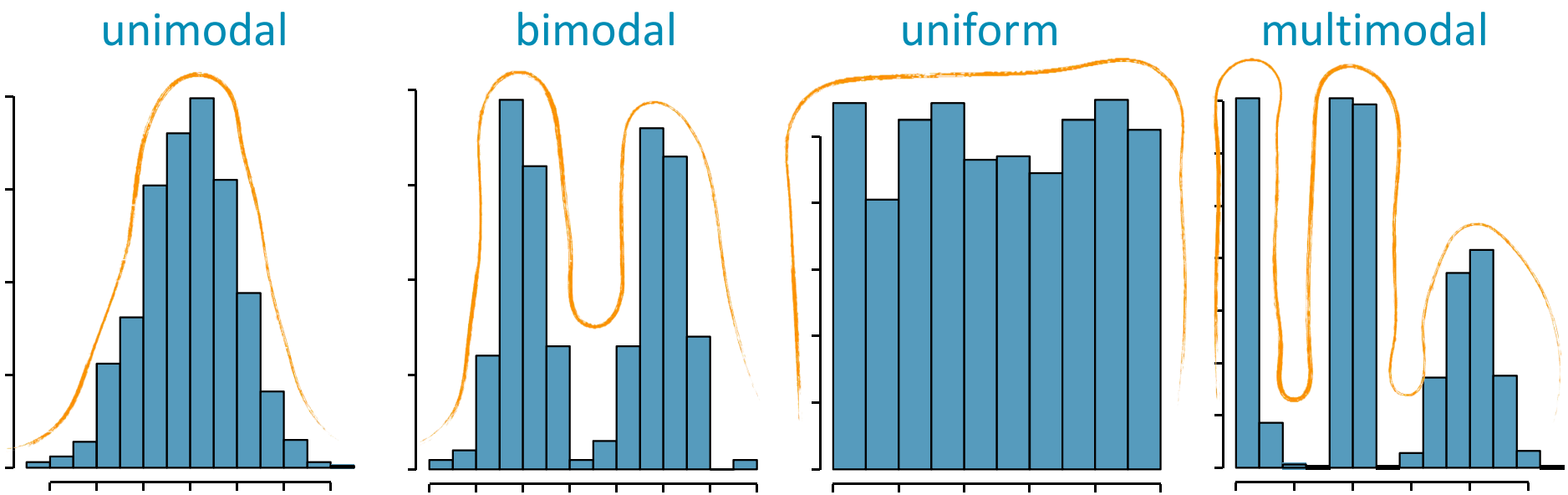


Skewness

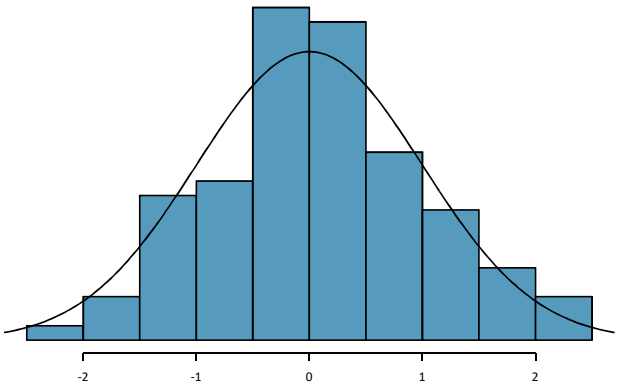
- Distributions are skewed to the side of the long tail



Modality

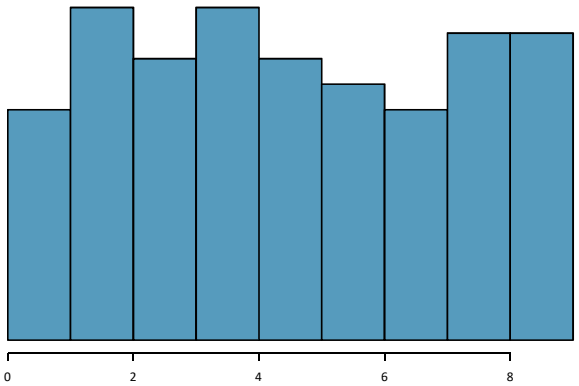


Modality



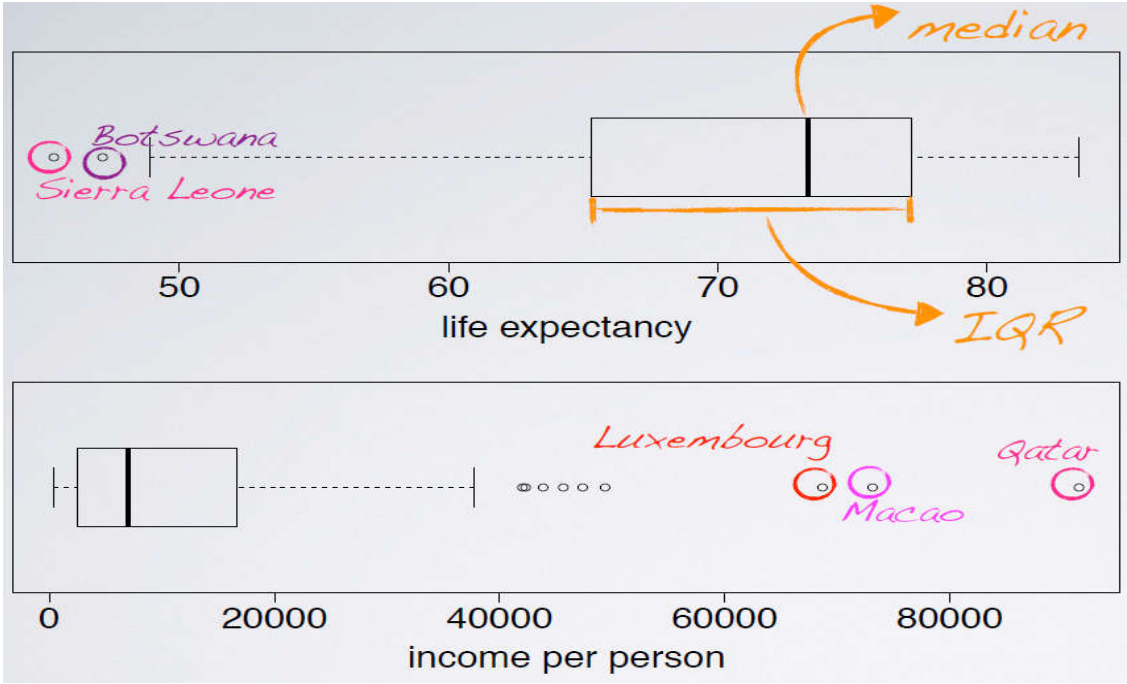
normal

heights at elementary school



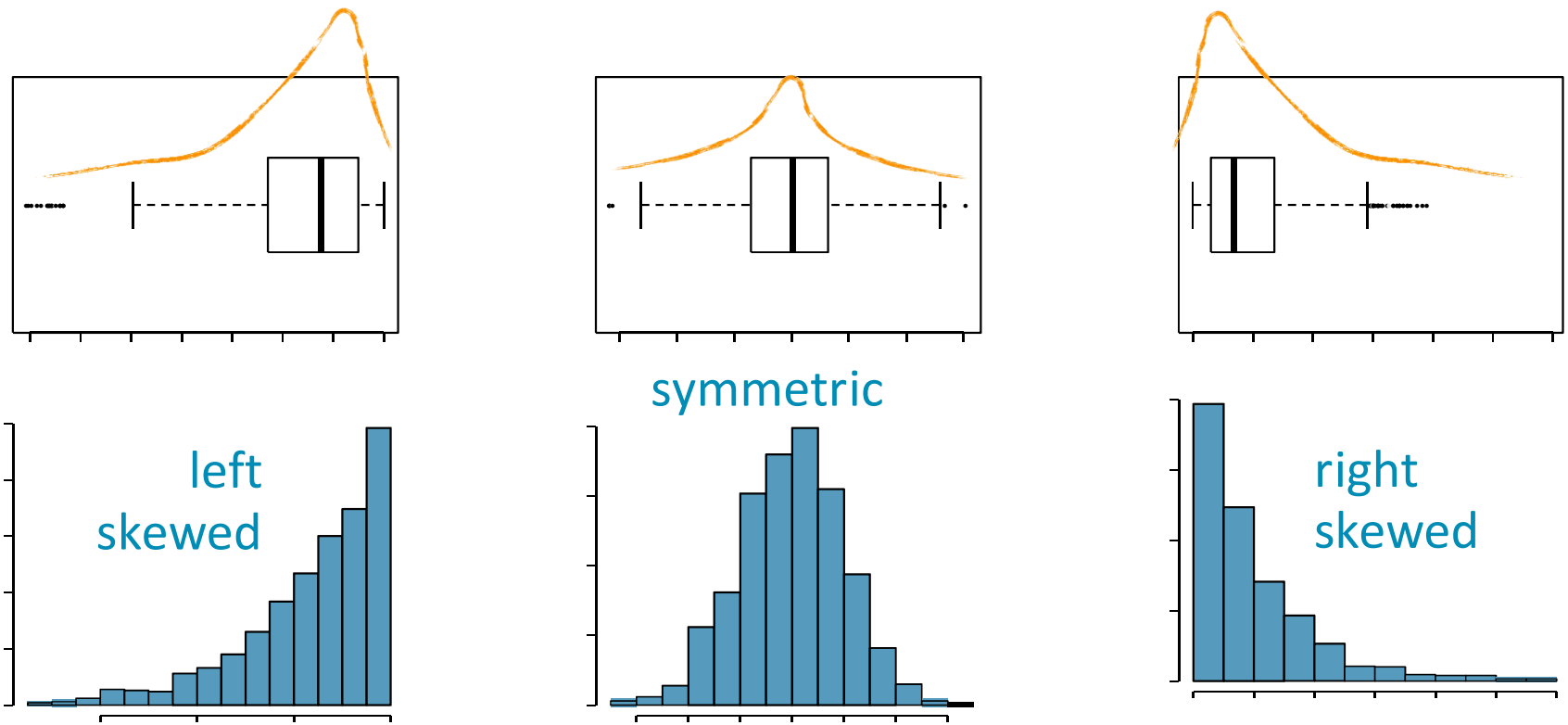
last digit of national ID

Box plot

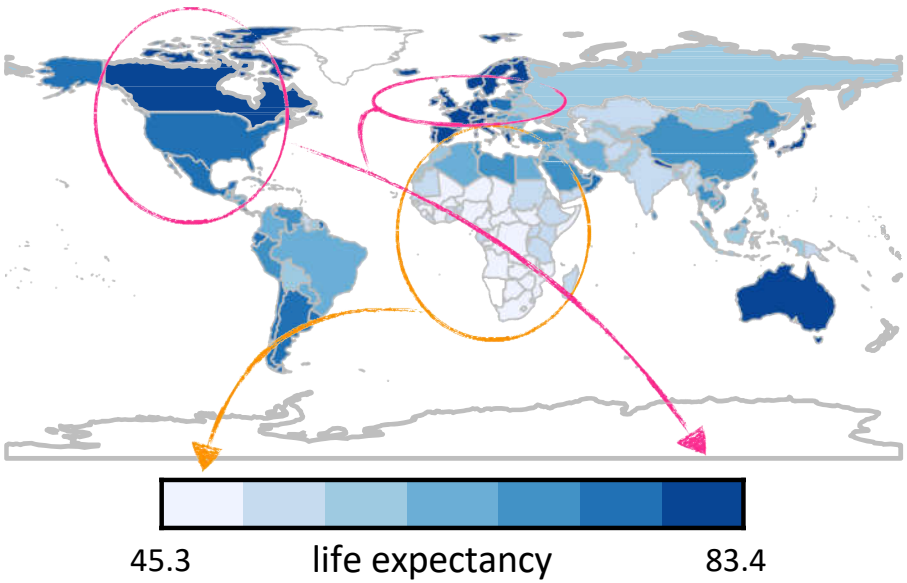


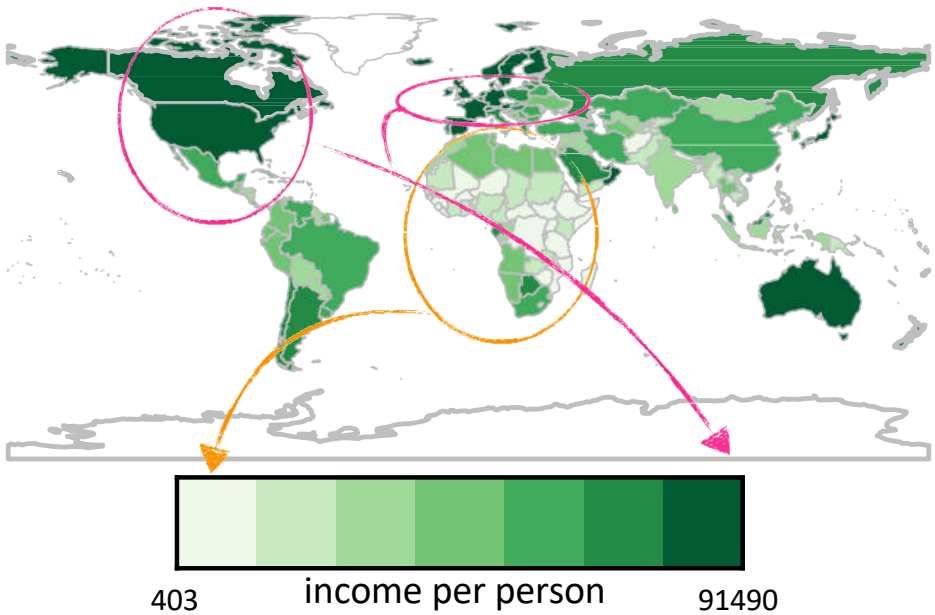
- Useful for highlighting outliers, median, IQR.

Determining the skewness from a box plot



Intensity Map

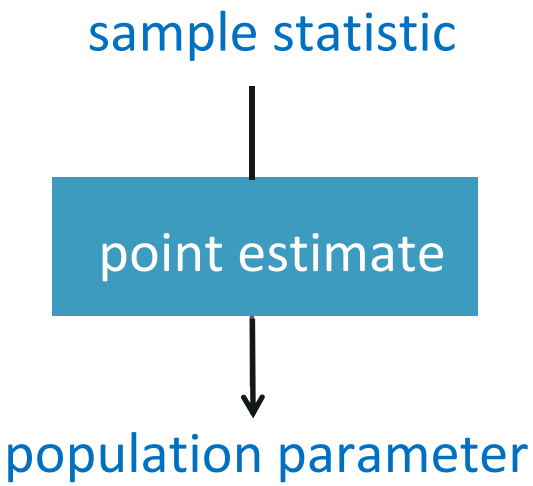




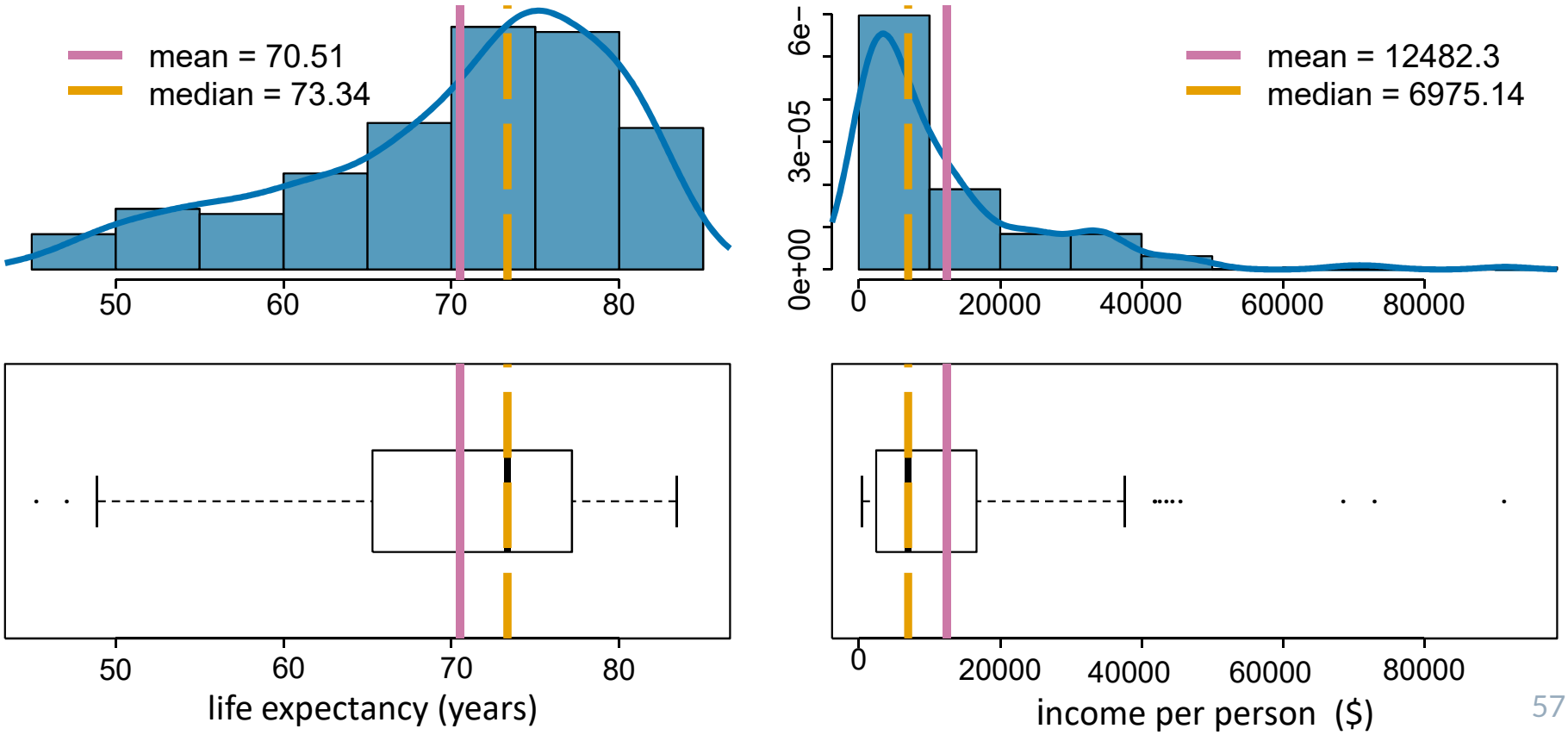
- Useful for highlighting the spatial distribution.

Measures of Center

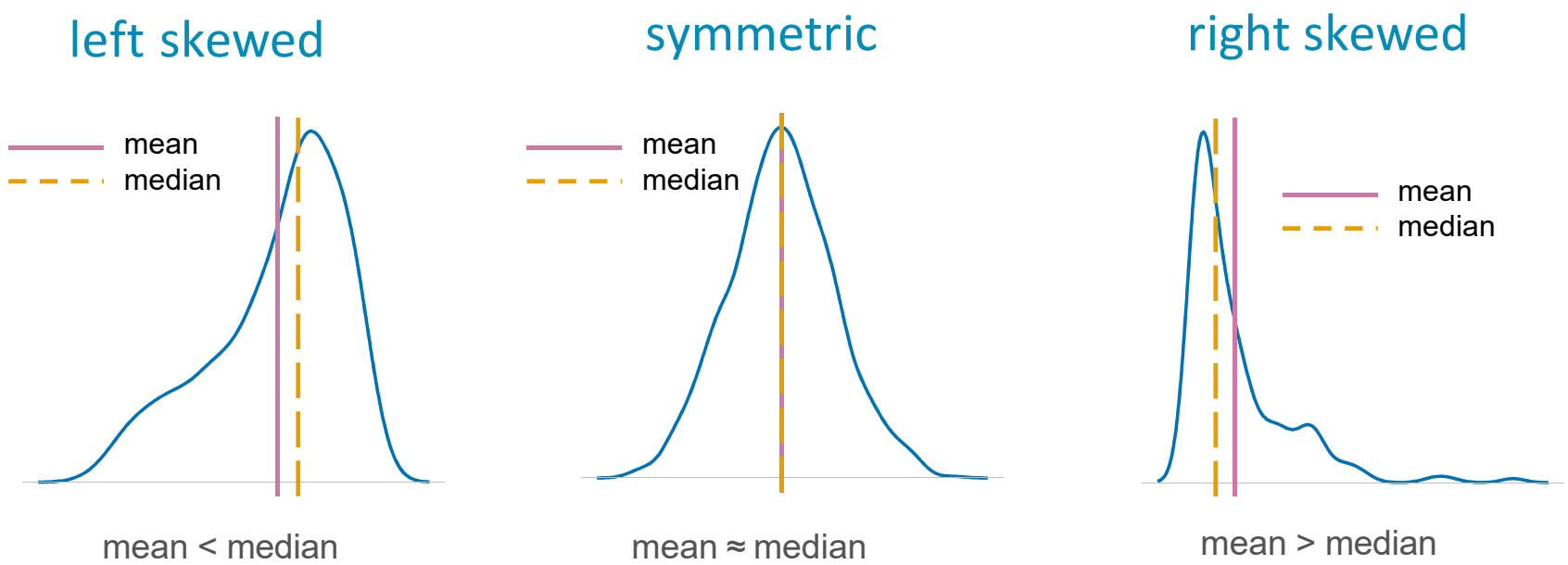
- Mean: arithmetic average
  - Sample mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
  - Population mean: μ
- Median: midpoint of the distribution
  - 50th percentile
- Mode: most frequent observation



Relation between Mean and Median

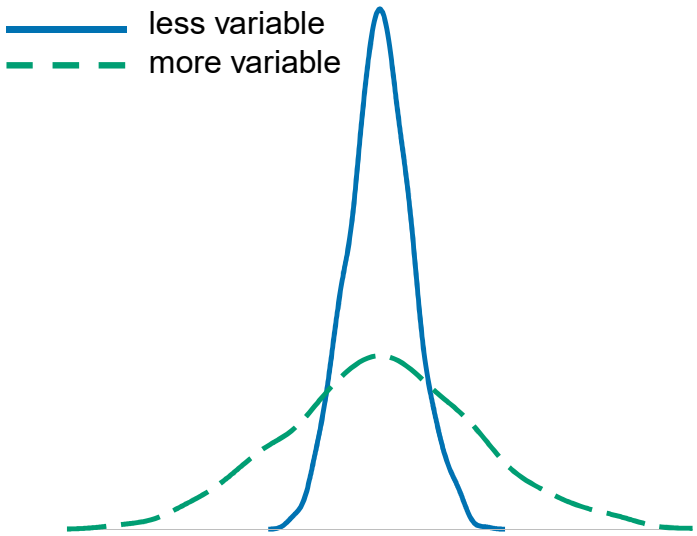


Skewness vs. Measures of Center



Measures of Spread

- In other words, statistics that tell us about the variability in the data:
  - Range = (*max* – *min*)
  - Variance
  - Standard deviation
  - Inter-quartile range



Variance

- Variance: roughly the average squared deviation from the mean
  - Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - Population variance: σ^2

- Example: Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

$$s^2 = \frac{(60.3 - 70.5)^2 + (77.2 - 70.5)^2 + \ \ldots + (58.1 - 70.5)^2}{201 - 1}$$

$$= 83.06 \ \text{years}^2$$

	data	life expectancy
1	Afghanistan	60.254
2	Albania	77.185
3	Algeria	70.874
⋮	⋮	⋮
201	Zimbabwe	58.142

Standard Deviation

- Standard deviation: roughly the average deviation from the mean that has the same units as the data
  - Sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

square root of the variance

  - Population standard deviation: σ

- Example: Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

$$s = \sqrt{83.06} = 9.11 \text{ years}$$

Interquartile Range

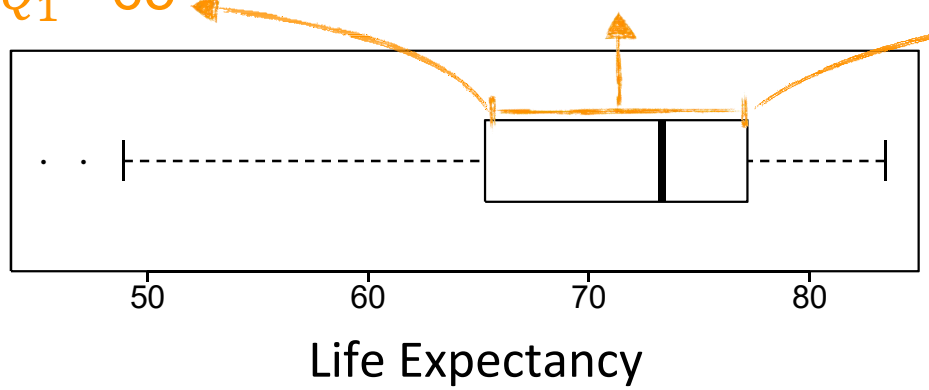
- Range of the middle 50% of the data, distance between the first quartile (25th percentile) and third quartile (75th percentile):

$$IQR = Q_3 - Q_1$$

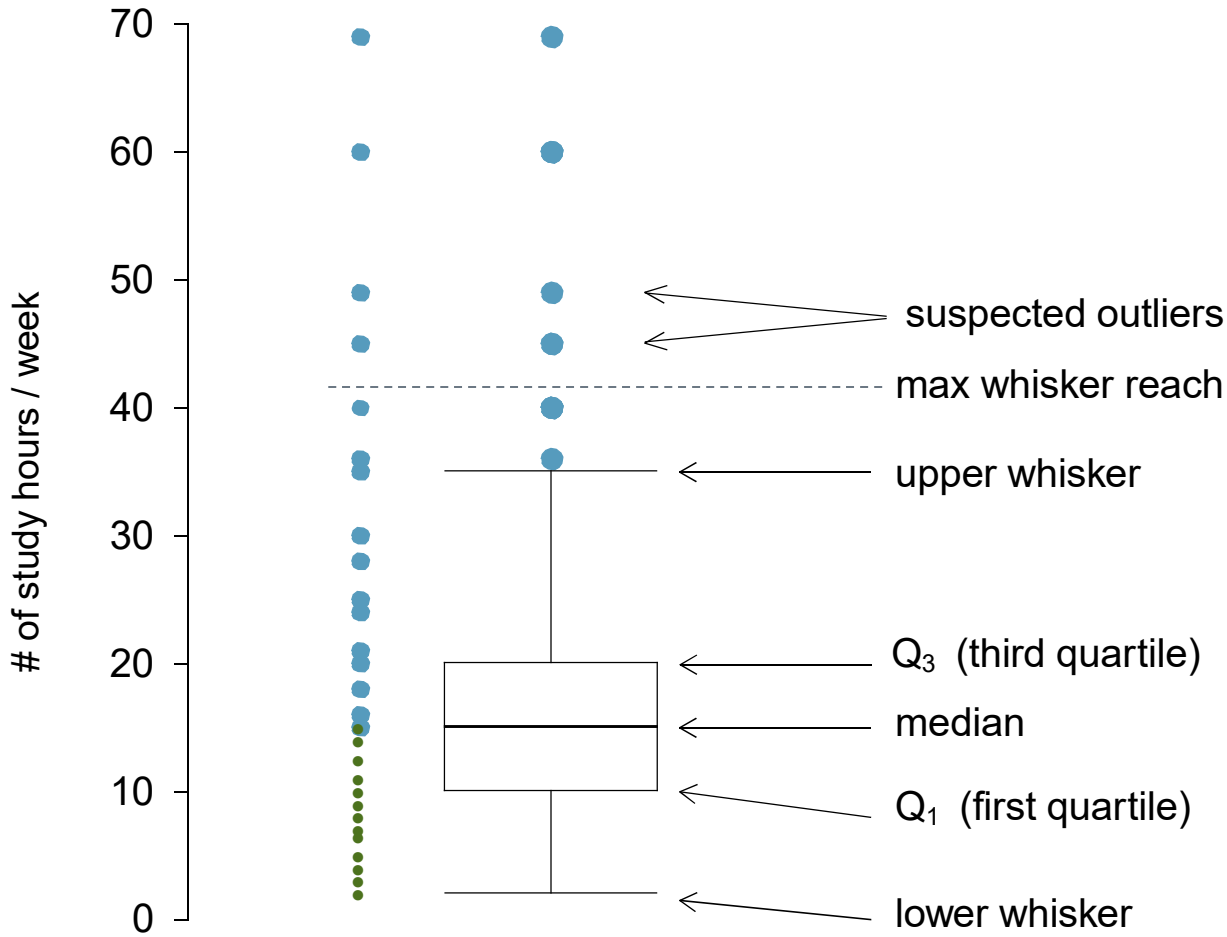
$$IQR = 77 - 65 = 12$$

$Q_1 \approx 65$

$Q_3 \approx 77$



Boxplot


70
60
50
40
30
20
10
0
# of study hours / week
suspected outliers
max whisker reach
upper whisker
Q_3 (third quartile)
median
Q_1 (first quartile)
lower whisker

Whiskers

- The whiskers attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times \mathrm{IQR}$:

$$\text{max upper whisker reach} = Q_3 + 1.5 \times \text{IQR}$$

$$\text{max lower whisker reach} = Q_1 - 1.5 \times \text{IQR}$$

- Example:

$$\text{IQR} : 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential outlier is defined as an observation beyond the maximum reach of the whiskers.
  - An observation that appears extreme relative to the rest of the data.

Outliers

- Why it is important to look for outliers?

- Examination of data for possible outliers serves many useful purposes, including:
  1. Identifying strong skew in the distribution.
  2. Identifying data collection or entry errors.
  3. Providing insight into interesting properties of the data.

Robust Statistics

- We define robust statistics as measures on which extreme observations have little effect.
- Example:

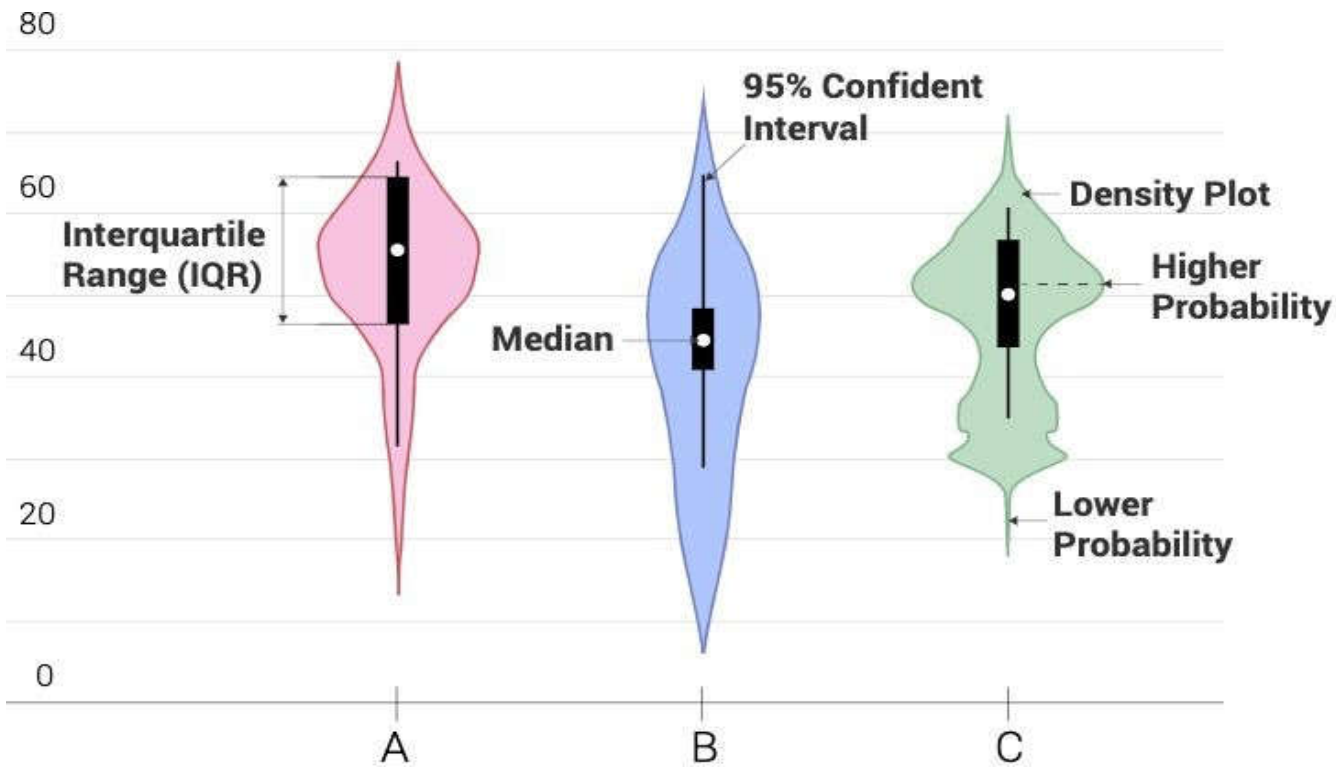
Data	Mean	Median
1, 2, 3, 4, 5, 6	3.5	3.5
1, 2, 3, 4, 5, 1000	169	3.5

	robust	non-robust
center	median	mean
spread	IQR	SD, range

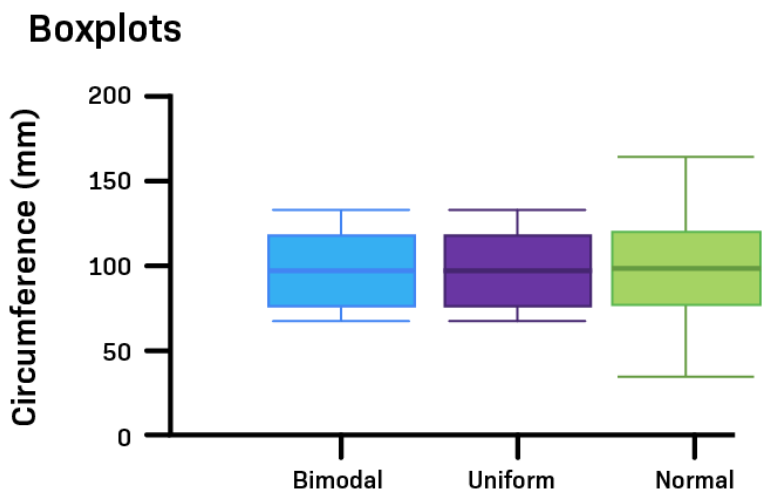
skewed, with extreme observations

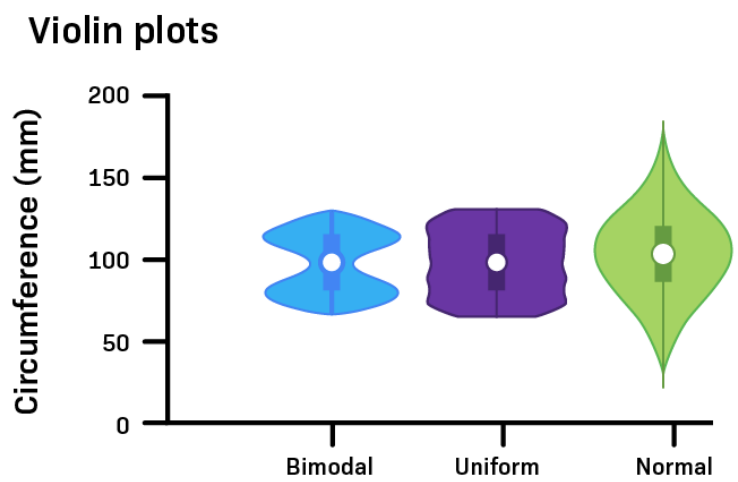
symmetric

Violin Plot


80
60
40
20
0
Interquartile Range (IQR)
95% Confident Interval
Median
Density Plot
Higher Probability
Lower Probability
A
B
C

Violin Plot vs. Box Plot

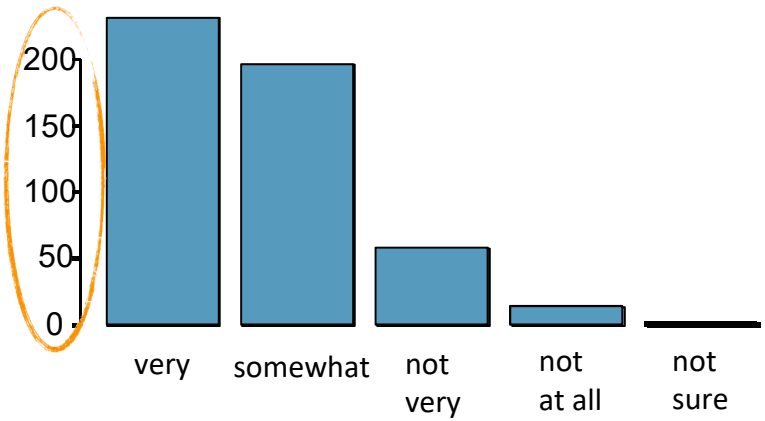




Describing Categorical Variables

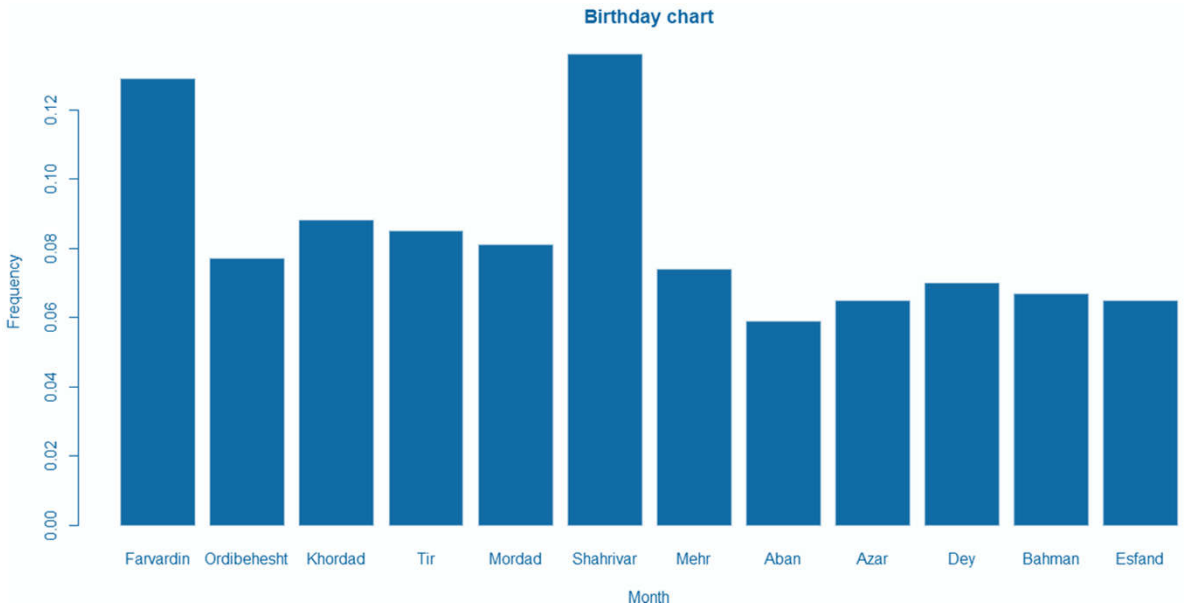
Frequency Table & Bar Plot

Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%

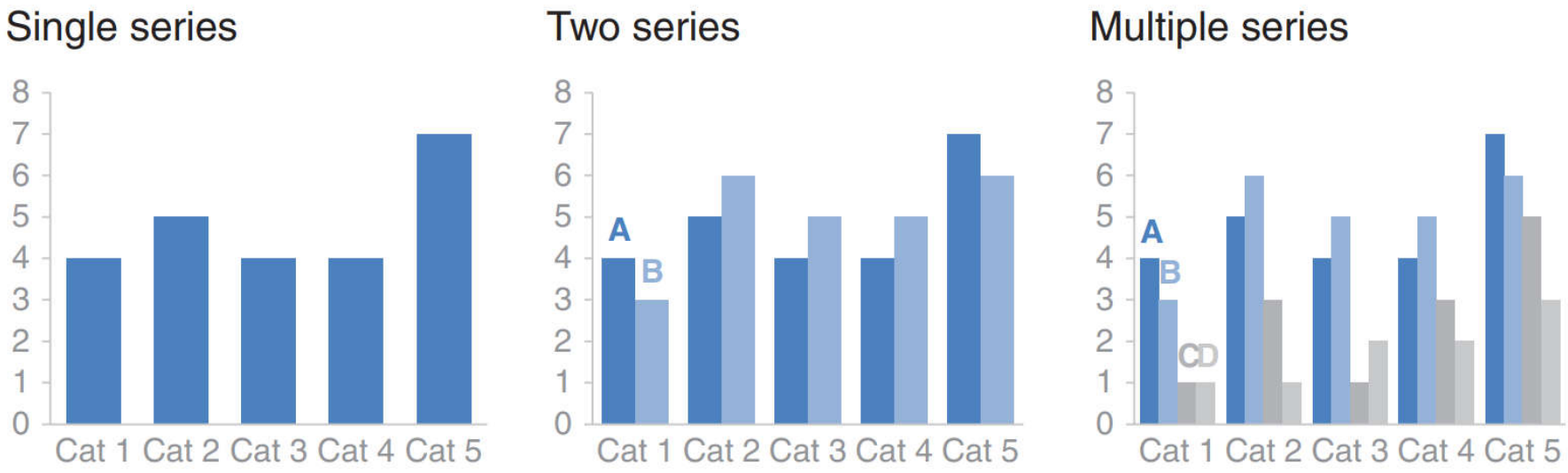


Birthdays in Iran

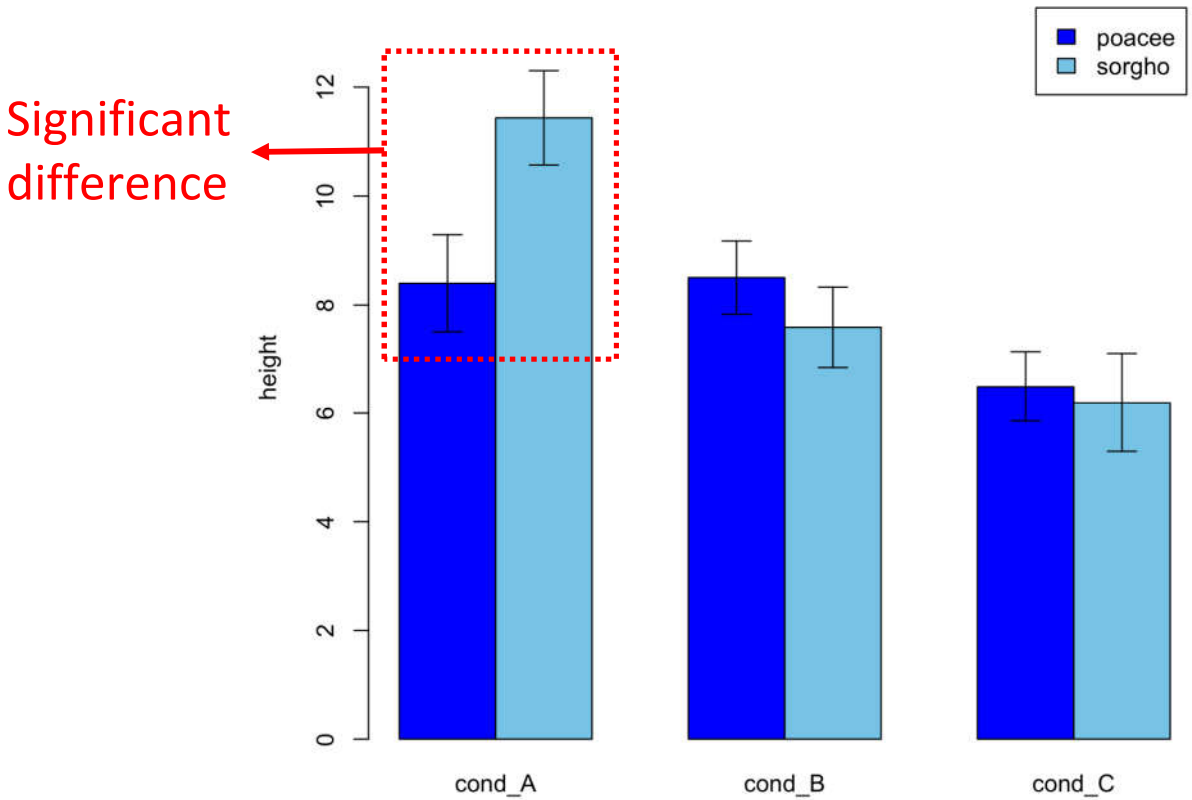
- Based on 1395 Census (A sample of 1,048,575 individuals)
  - Total number of valid data with Persian calendar: 1,000,222



Grouped Bar Chart

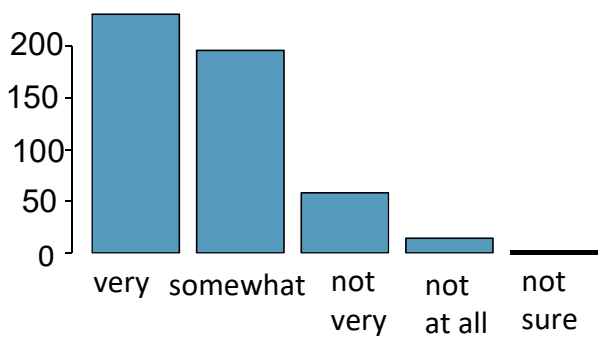


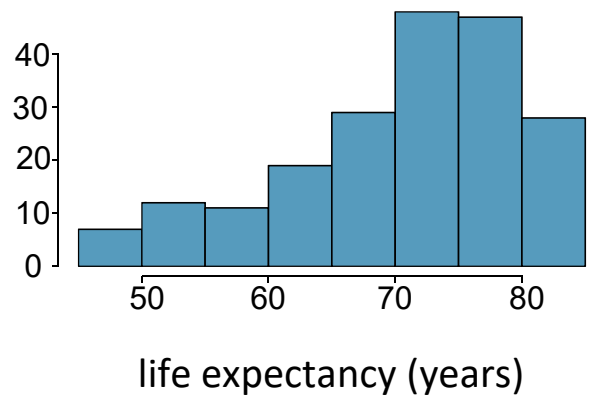
Bar Plot + Error Bar



Bar Plots vs. Histograms

- Barplots for categorical variables, but histograms for numerical variables.
- x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable.

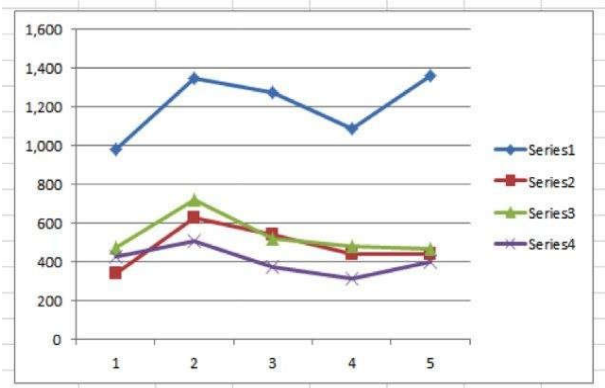


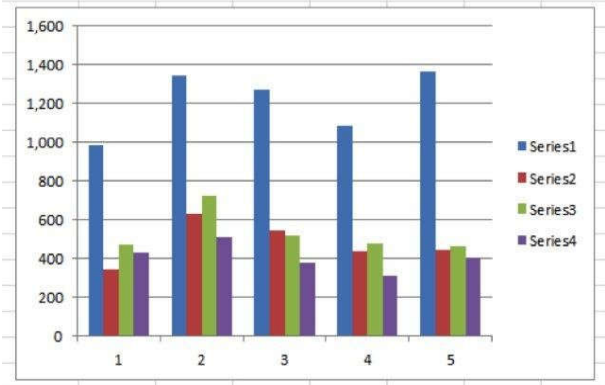


Bar Plots vs. Line Charts

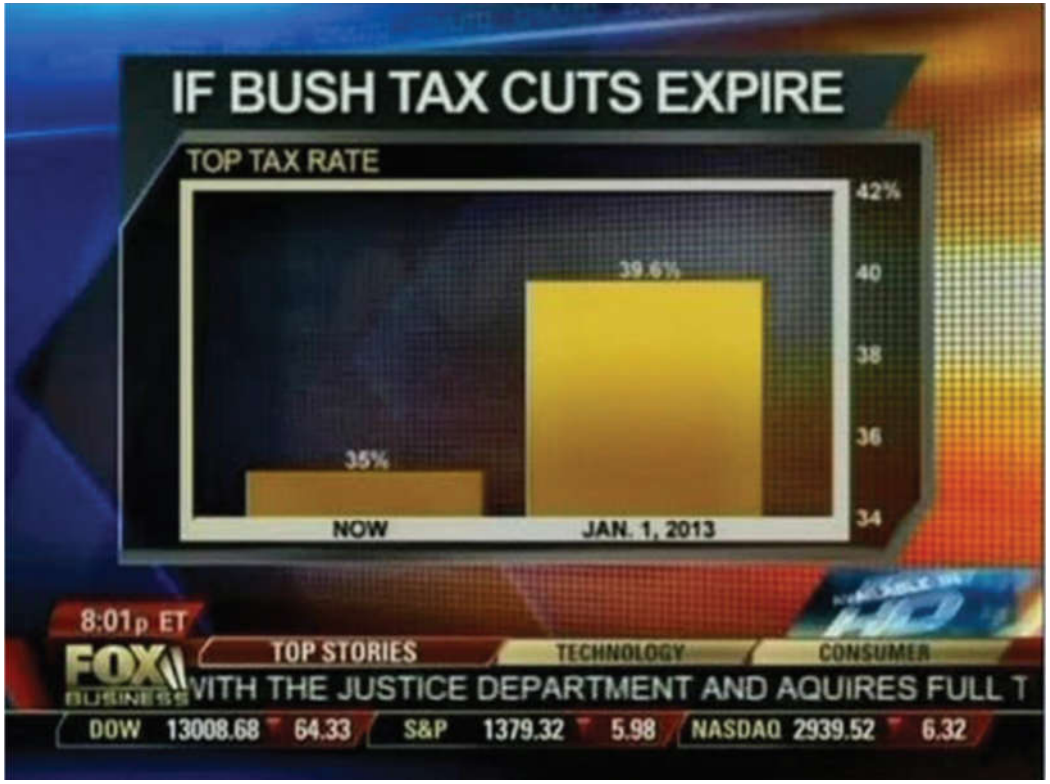
Continuous values
e.g., time series

Discrete values
e.g., countries



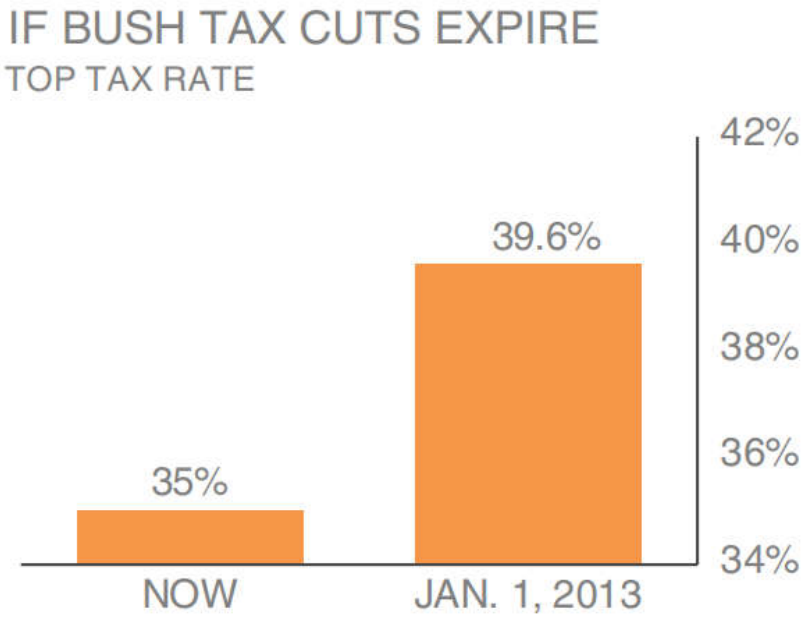


Bar Plot Abuse

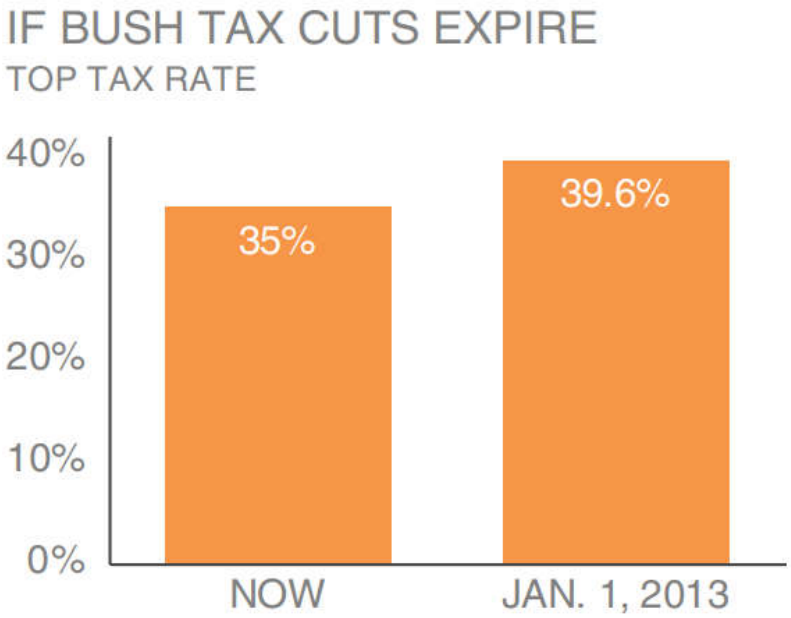


Bar Plot Abuse

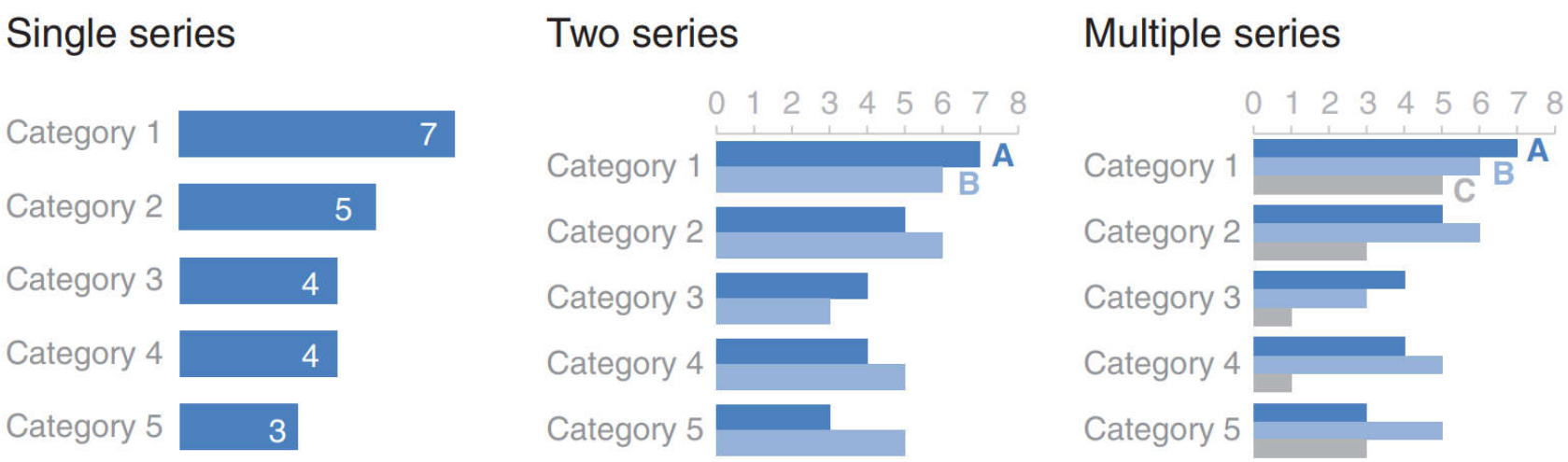
Non-zero baseline: as originally graphed



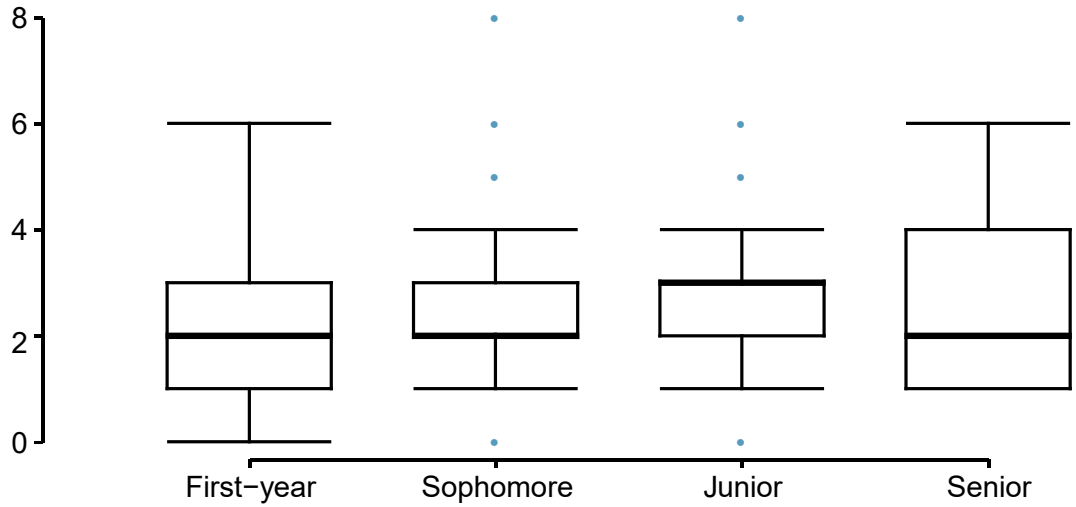
Zero baseline: as it should be graphed



Horizontal Bar Plot

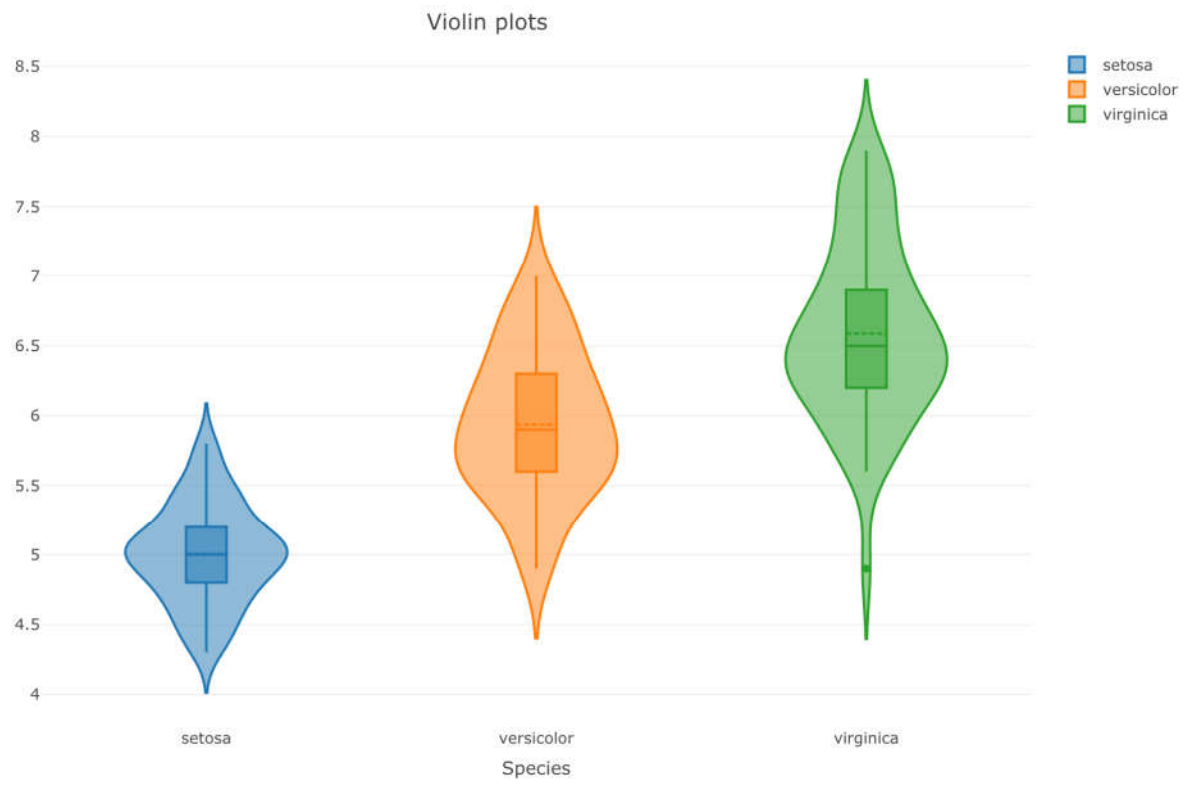


Side-by-side box plots

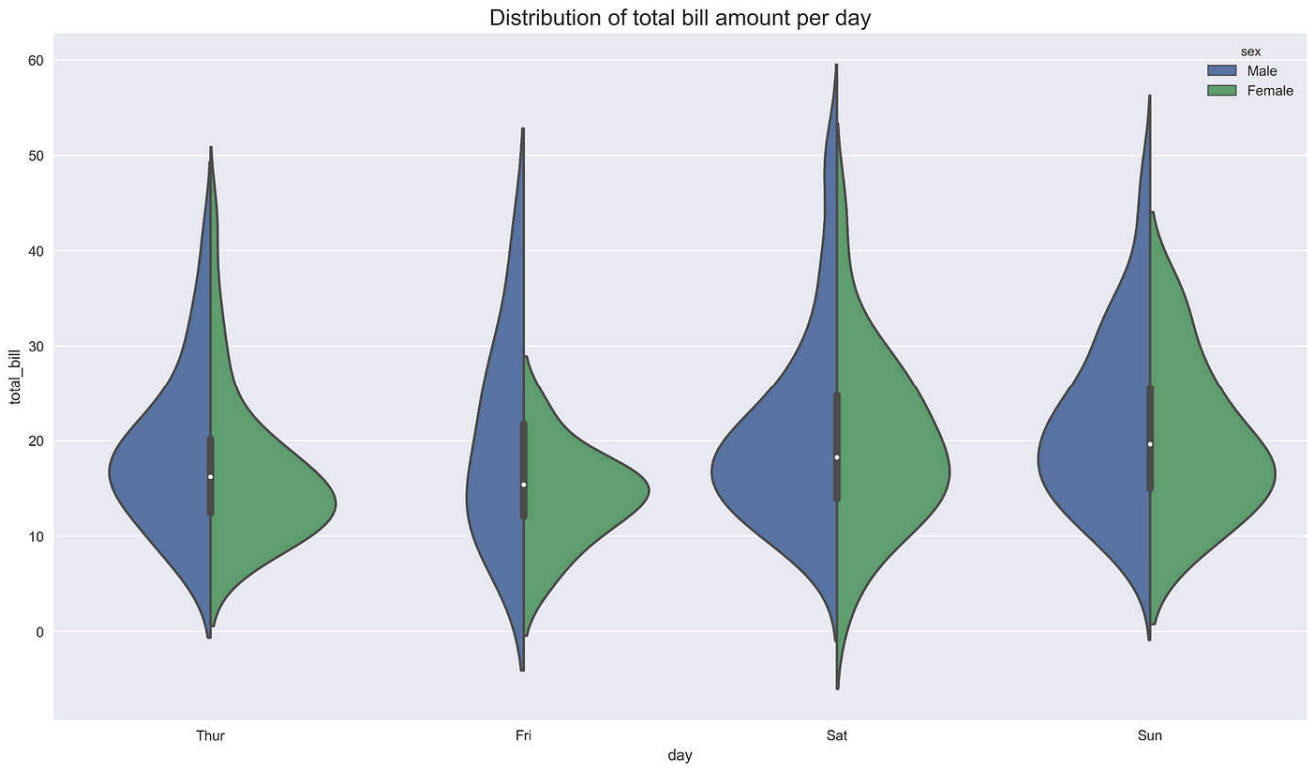


- Does there appear to be a relationship between class year and number of societies students are in?

Side-by-side Violin Plot



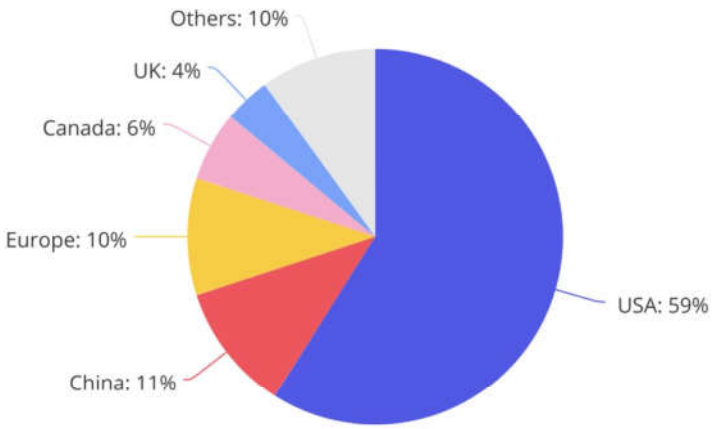
Violin Plots for Comparison



To Be Avoided

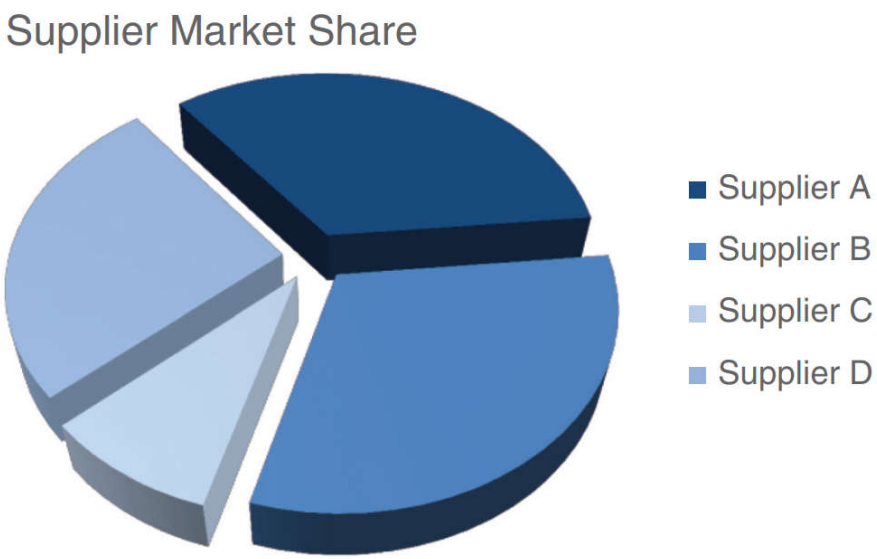
~~Pie Chart?~~ NO!

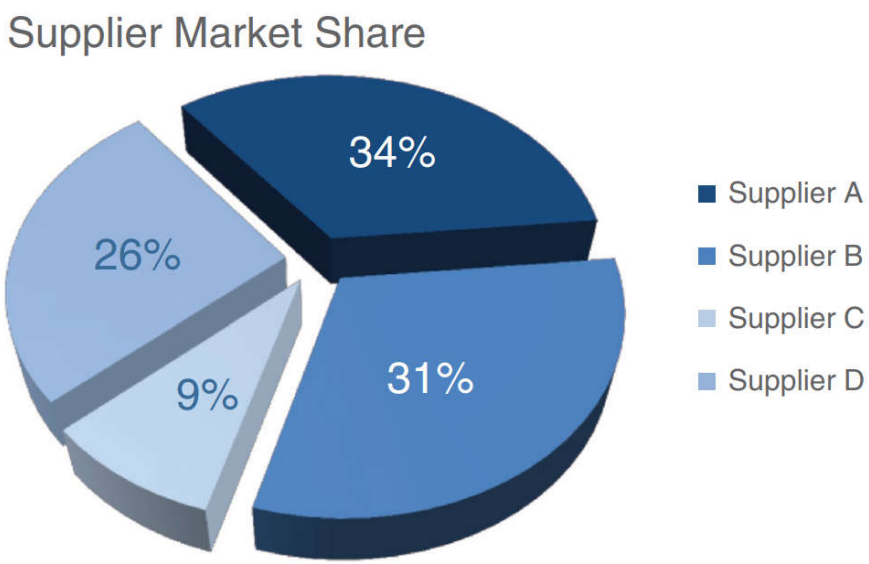
Where do top-tier AI researchers work today?



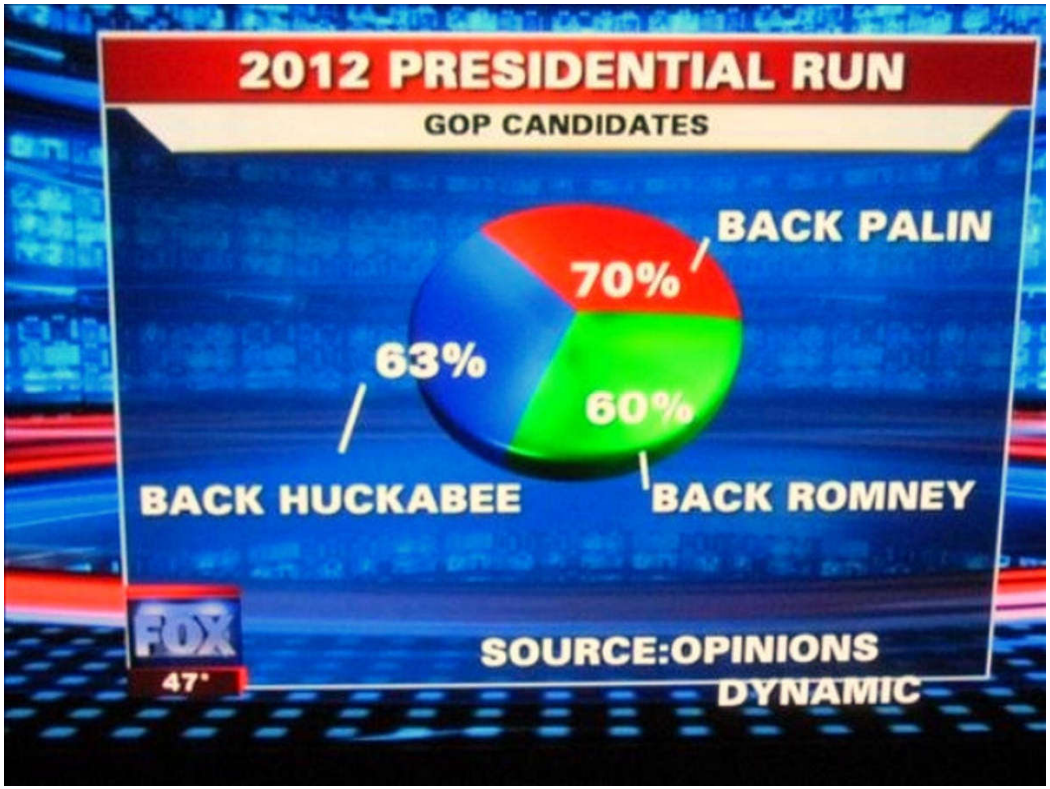
Country affiliations are based on the headquarters of institutions in which the researchers currently work.

3D Pie Charts

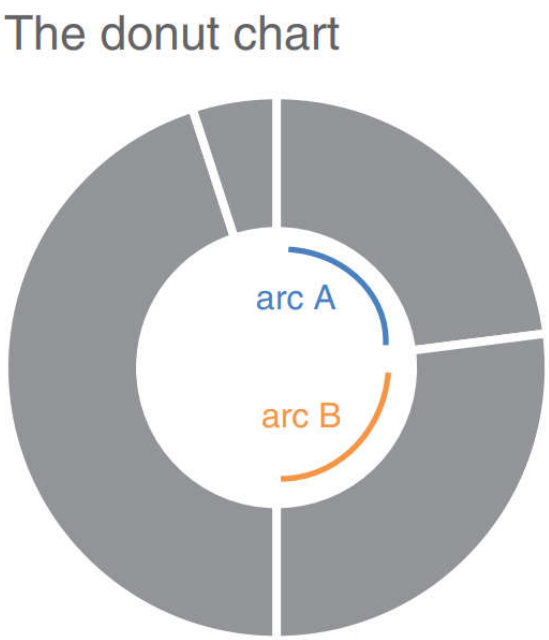




Terrible Pie Chart

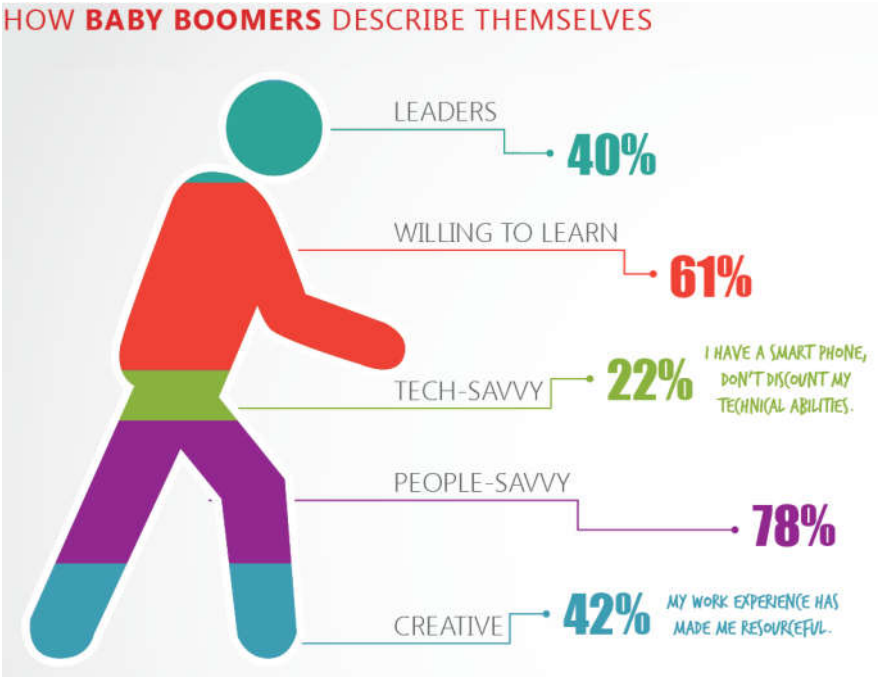


Donut Chart

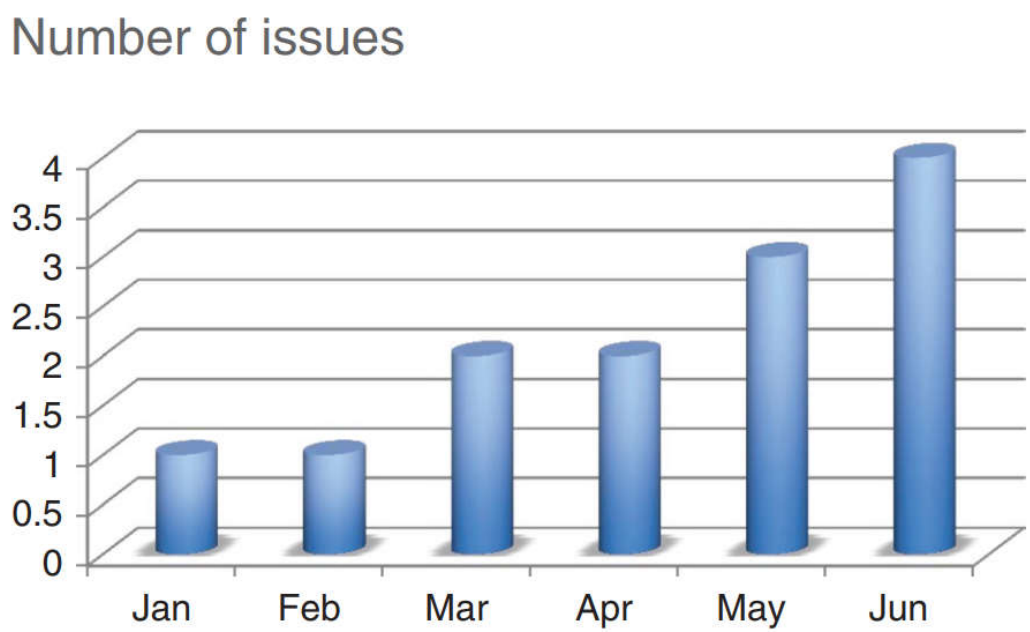


Area Graphs

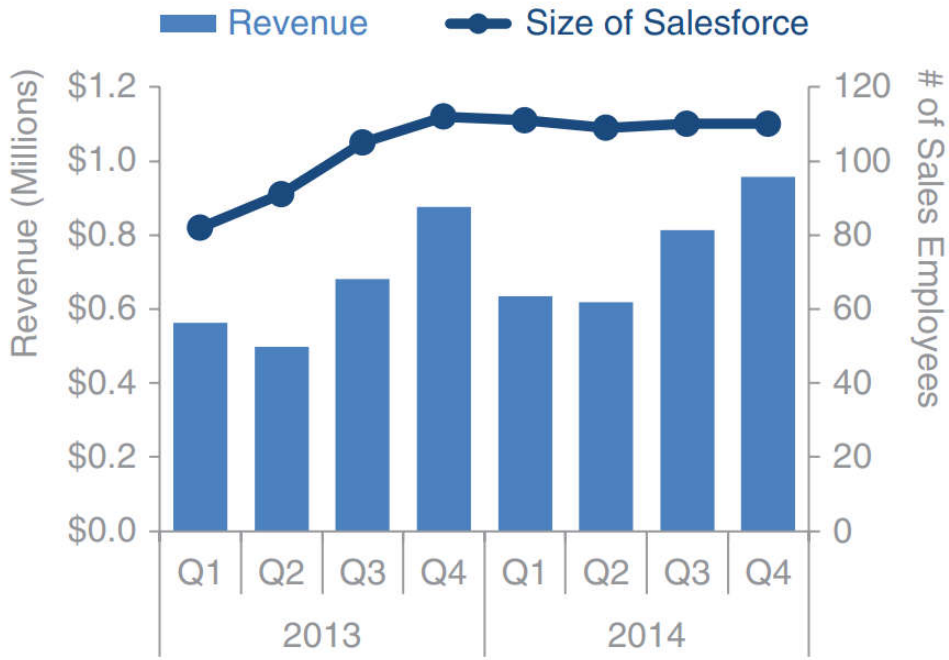
- Humans' eyes don't do a great job of attributing quantitative value to two-dimensional space.



Never use 3D

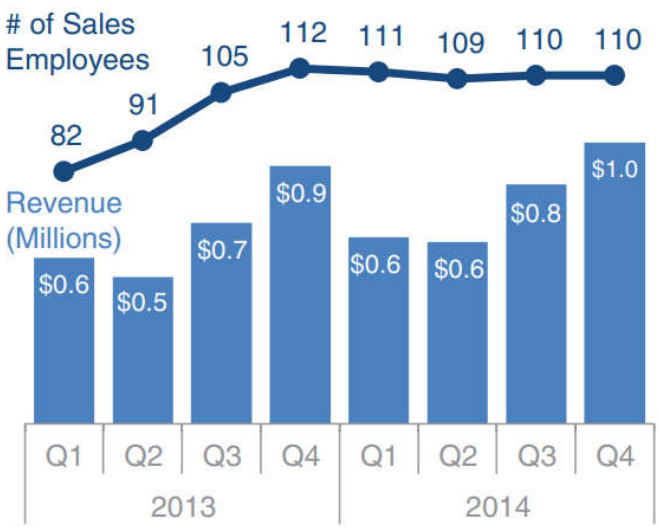


Secondary *y*-axis

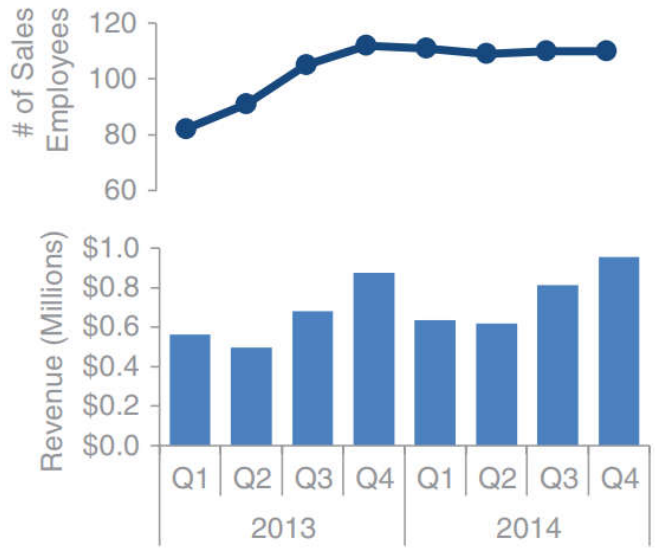


Alternatives for Secondary *y*-axis

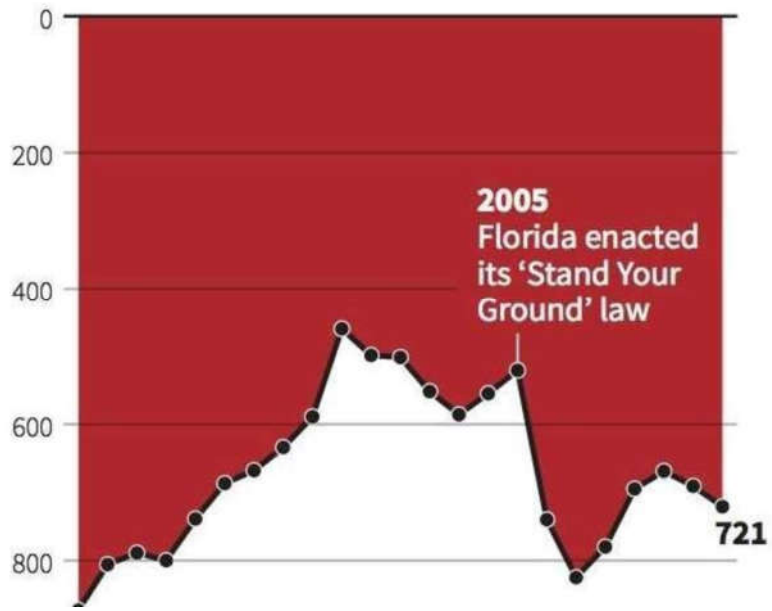
Alternative 1: **label directly**

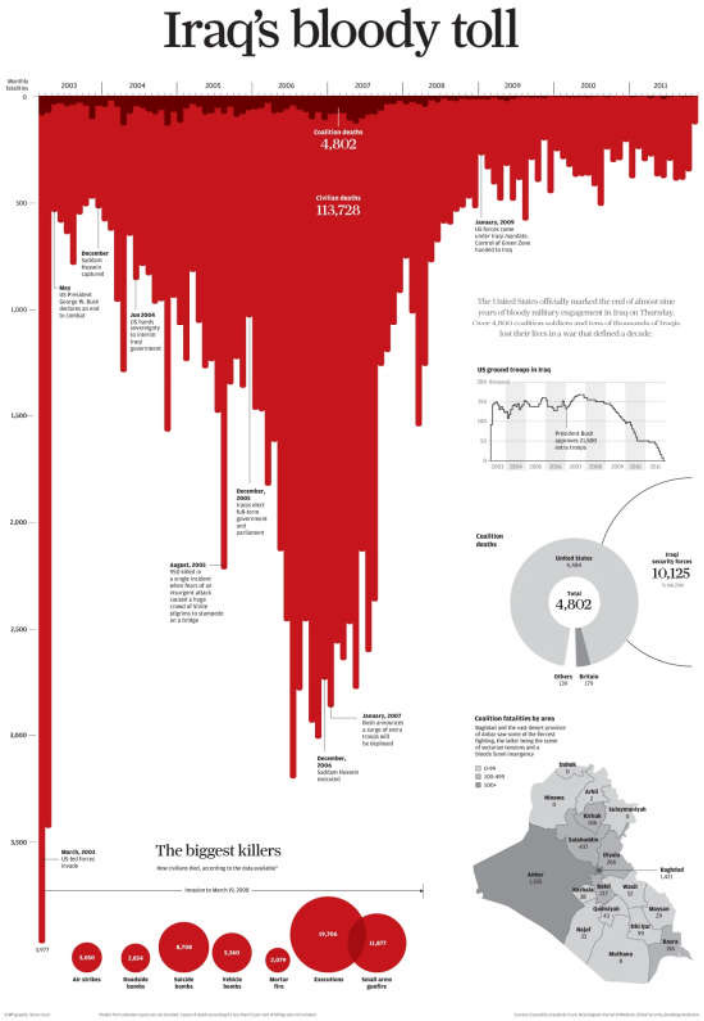


Alternative 2: **pull apart vertically**



Inverse Charts





Cumulative Charts

Cumulative Charts