Statistical Inference

Introduction to Data

Behnam Bahrak

Spring 2024

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

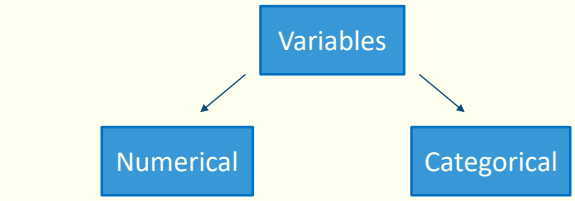
Variable ↓

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)*

The data matrix of the **email50** data set.

Types of Variables



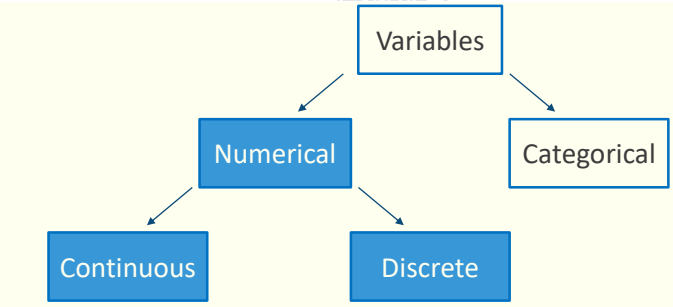
(quantitative)

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

(qualitative)

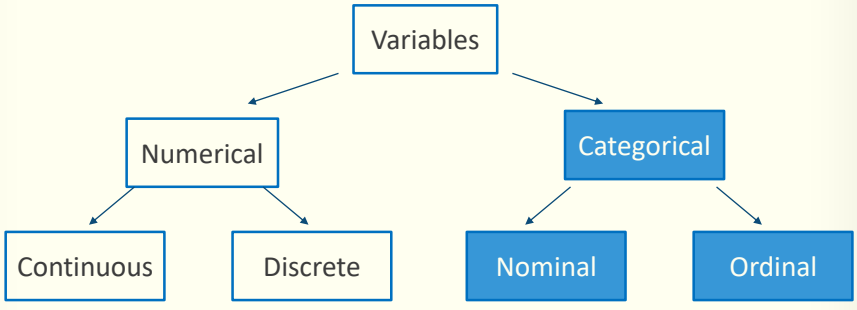
- take on a limited number of distinct categories.
- categories can be identified with numbers, but not sensible to do arithmetic operations

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

Question?

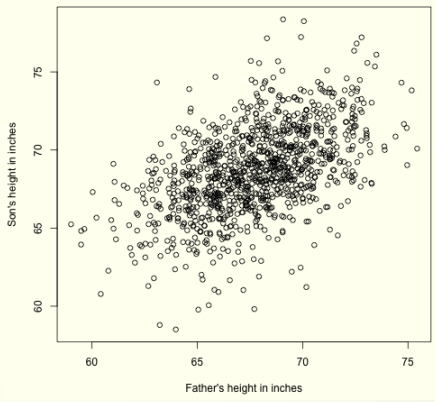
What type of variable is a telephone area code?

(a) numerical, continuous
(b) numerical, discrete
(c) categorical, nominal ✓
(d) categorical, ordinal

Statistical Inference Behnam Bahrak

Relationships between variables

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



Populations and Samples

- Each research question refers to a target population.
  - Often times, it is too expensive to collect data for every case in a population.
- A sample represents a subset of the cases and is often a small fraction of the population.

Research question: Can people become better, more efficient runners on their own, merely by running?
Population of interest: *All people*
Sample: Group of adult men who recently joined a running group
Population to which results can be generalized: *Adult men, if the data are randomly sampled*

Sampling from a population

- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).

Anecdotal Evidence

- Consider the following statements:

  1. My uncle smokes three packs a day and he's in perfectly good health, so smoking doesn't affect your health.

  2. I met two students who took more than 7 years to graduate from UT, so it must take longer to graduate at UT than at many other colleges.

- Each 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.

Census

- Wouldn't it be better to just include everyone and "sample" the entire population?
  - This is called a *census*.
- There are problems with taking a census:
  - It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. *And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.*
  - Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
  - Taking a census may be more complex than sampling.

Sampling Bias

Some Sources of Sampling Bias

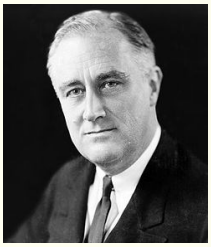
- *Non-response:* If only a *non-random* fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- *Voluntary response:* Occurs when the sample consists of people who volunteer to respond because they have strong opinions on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

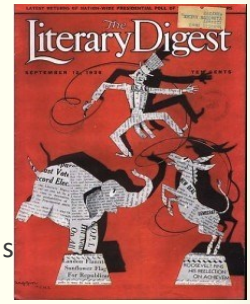
- In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



Franklin D. Roosevelt

The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 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.

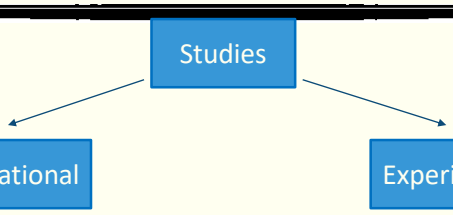
Explanatory and Response Variables

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

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

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

Observational Studies & Experiments

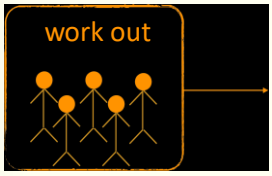


- collect data in a way that does not directly interfere with how the data arise ("observe")
- only establish an association
- retrospective: uses past data
- prospective: data are collected throughout the study

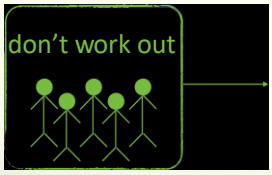
- randomly assign subjects to treatments
- establish causal connections between explanatory and response variables.

Example

Observational Study

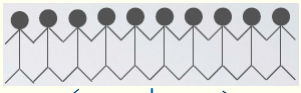


average energy level

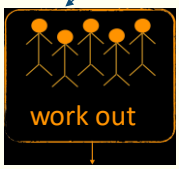


average energy level

Experiment



random assignment

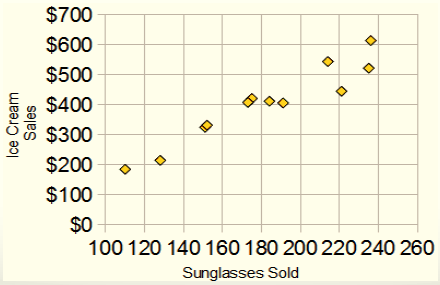


average energy level

average energy level

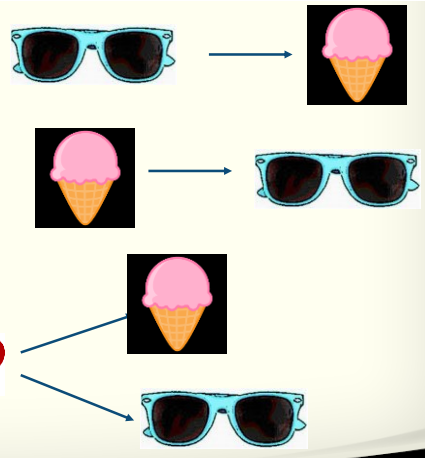
Correlation does not imply causation

- 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. The correlation between Sunglasses and Ice Cream sales is high. Does this mean that sunglasses make people want ice cream?



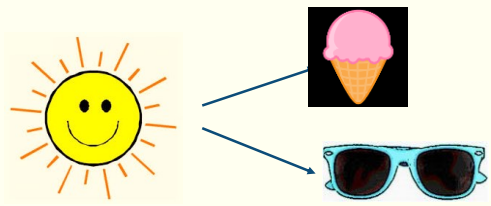
Three possible explanations

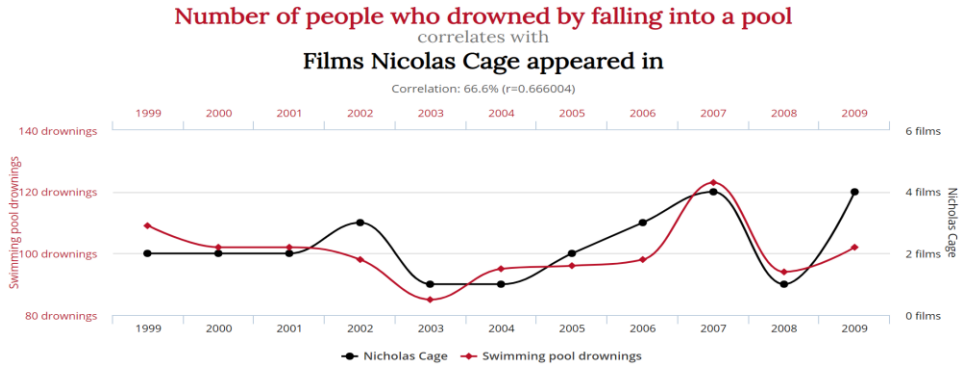
1. Sunglasses make people want ice cream!
2. Eating ice cream makes people buy sunglasses!
3. A third variable is responsible for both.

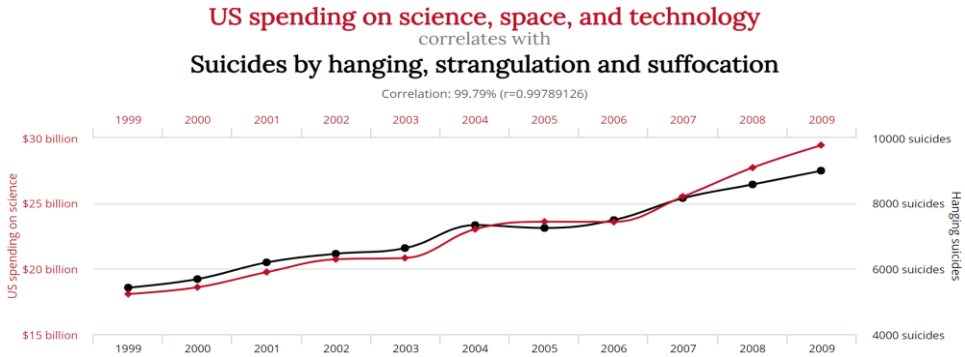


Confounding Variable

➢ 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 confounding variables.







Left-handedness and life expectancy

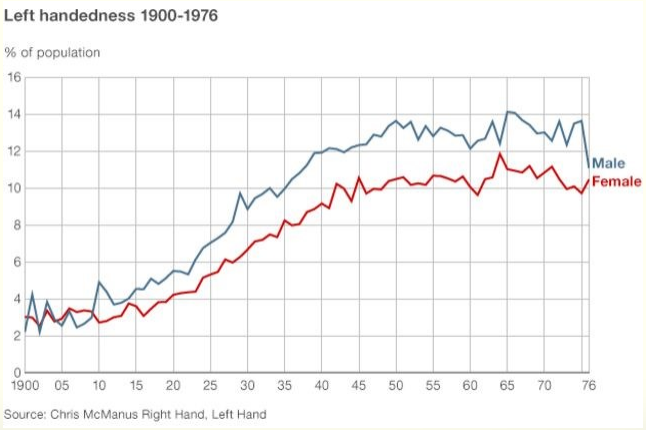
The New York Times

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

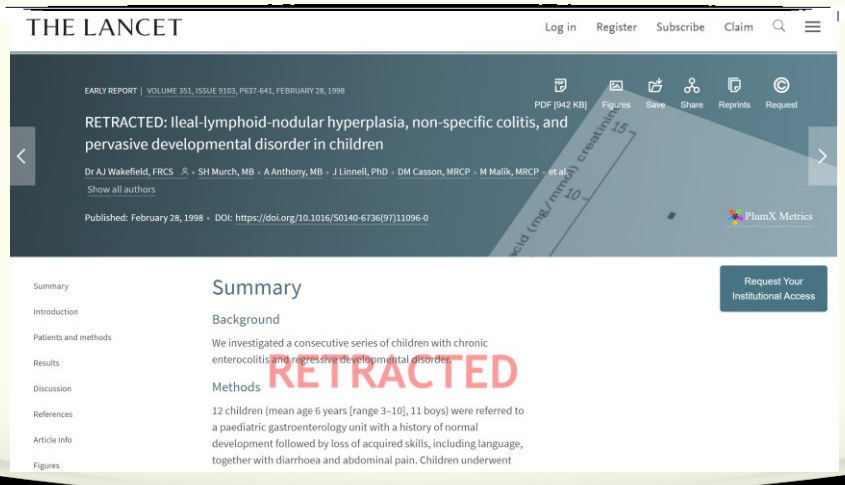
Reuters

April 4, 1991

Left handedness



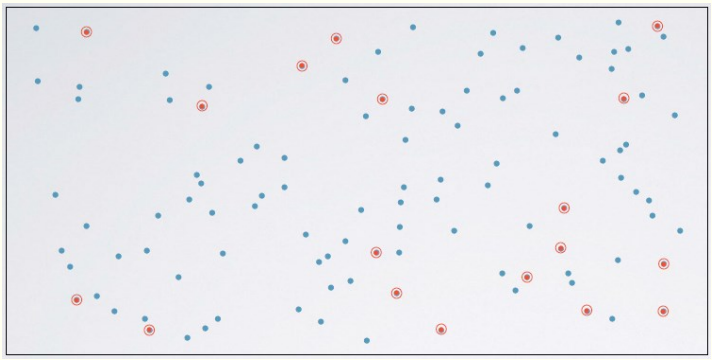
MMR Vaccination and Autism



Sampling Strategies

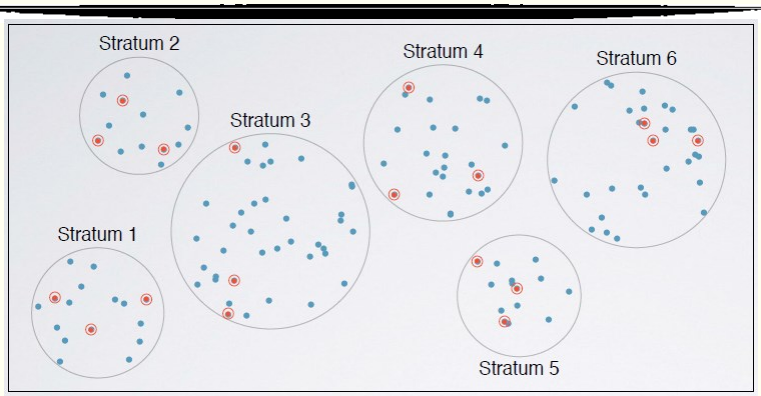
- Almost all statistical methods are based on the notion of implied randomness.
- If observational data are not collected in a random framework from a population, these statistical methods (the estimates and errors associated with the estimates) are not reliable.
- Most commonly used random sampling techniques are:
  - *Simple Random Sampling (SRS)*
  - *Stratified Sampling*
  - *Cluster Sampling*
  - *Multistage Sampling*

Simple Random Sampling (SRS)



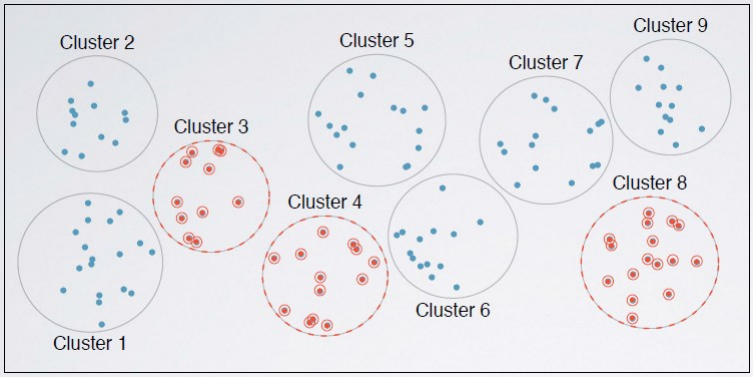
- Each case is equally likely to be selected.

Stratified Sampling



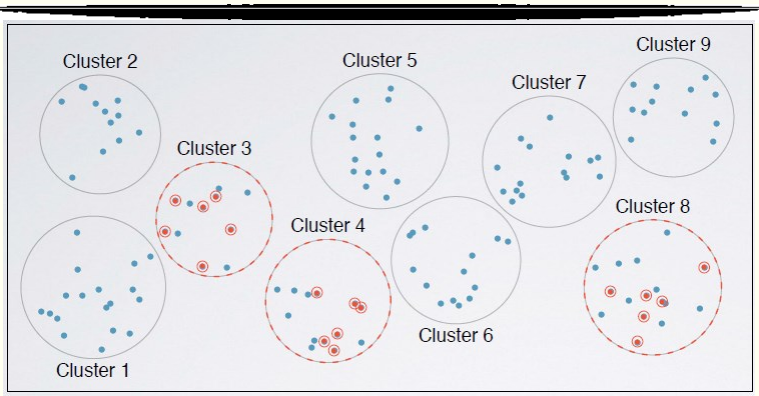
- Divide the population into homogenous strata, then randomly sample from within each stratum.

Cluster Sampling



- Divide the population to clusters, randomly sample a few clusters, then sample all observations within these clusters

Multistage Sampling



- Divide the population clusters, randomly sample a few clusters, then randomly sample within these clusters

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.

More on Blocking

- We would like to design an experiment to investigate if energy gels make you run faster:
  - Treatment: energy gel
  - Control: no energy gel

- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

Blocking vs. Explanatory Variables

- Explanatory variables (factors): conditions we can impose on experimental units
- Blocking variables: characteristics that the experimental units come with, that we would like to control for
- Blocking is like stratifying:
  - blocking during random assignment
  - stratifying during random sampling

Question

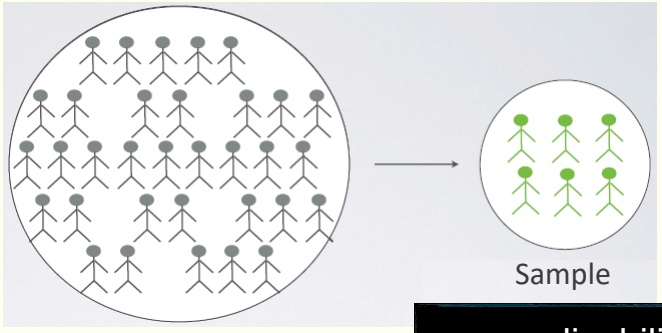
A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Determine explanatory, response, and blocking variables.

Answer: There are 2 explanatory variables *(light and noise)*, 1 blocking variable *(gender)*, and 1 response variable *(exam performance)*.

Experimental Design Terminology

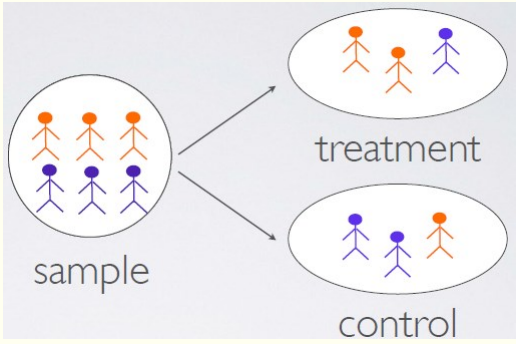
- *Placebo:* fake treatment, often used as the control group for medical studies
- *Placebo effect:* experimental units showing improvement simply because they believe they are receiving a special treatment
- *Blinding:* when experimental units do not know whether they are in the control or treatment group
- *Double-blind:* when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Random Sampling



generalizability

Random Assignment



Causality

Random Assignment vs. Random Sampling

	Random assignment	No random assignment	
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
	Causation	Correlation	

ideal experiment

most observational studies

most experiments

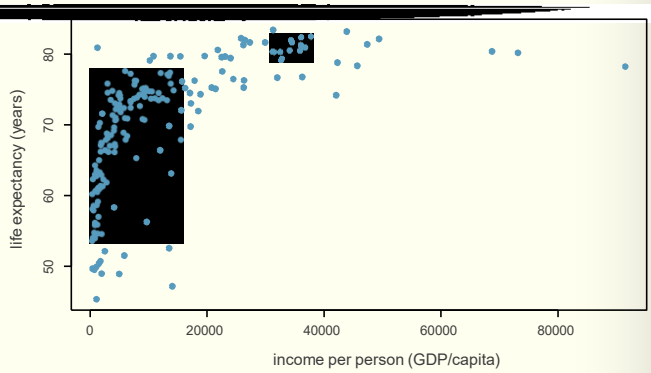
bad observational studies

Visualizing Numerical Data

- Scatterplot
- Histogram
- Dot plot
- Box plot
- Intensity map

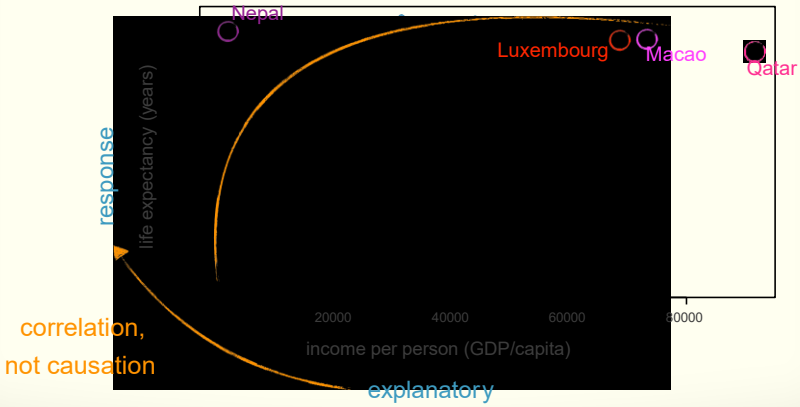
Scatterplot

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

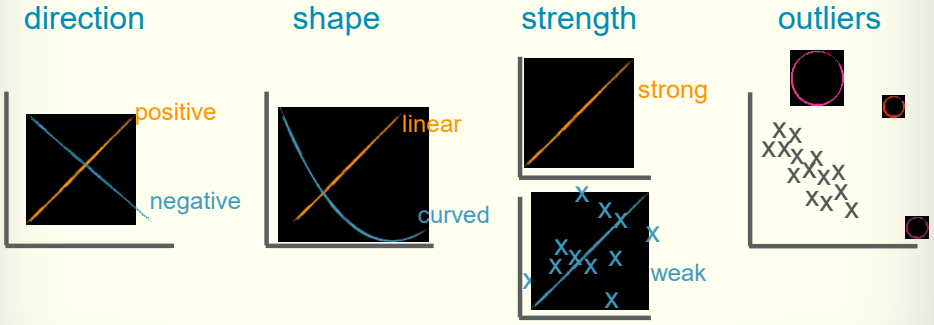


- *Scatterplots* are useful for visualizing the relationship between two numerical variables.

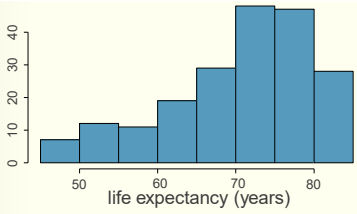
Scatterplot

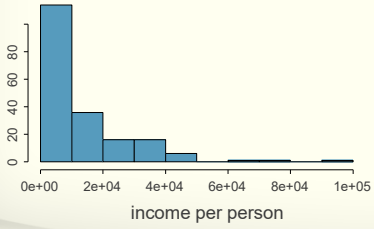


Evaluating the relationship



Histogram

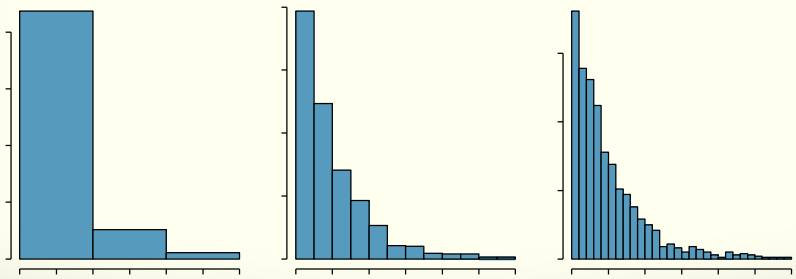




- Histograms provide a view of the data density.
- Histograms are especially convenient for describing the shape of the data distribution.
- The chosen bin width can alter the story the histogram is telling

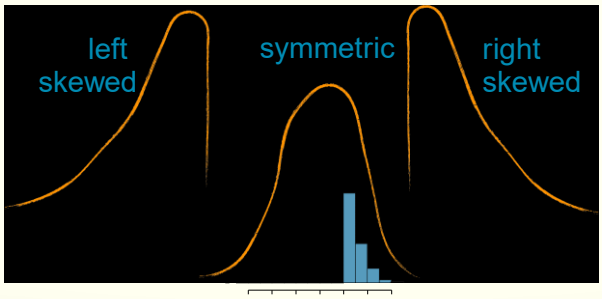
Bin Width

- 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.

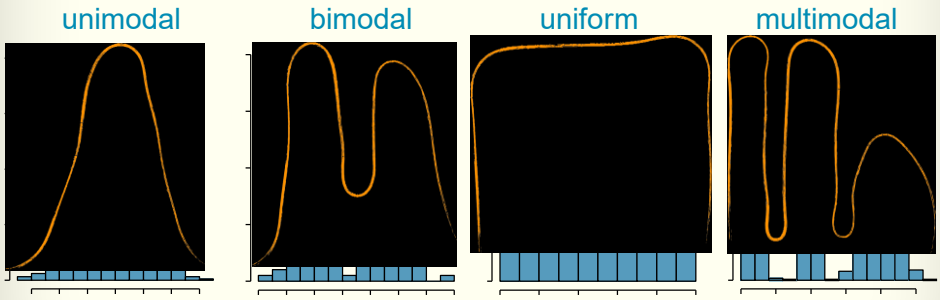


Skewness

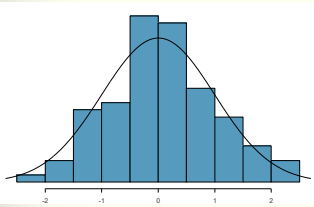
- Distributions are skewed to the side of the long tail



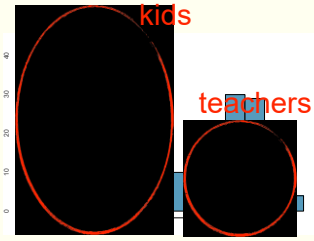
Modality



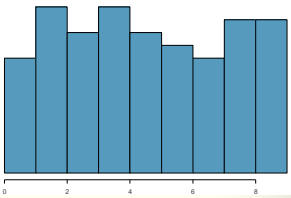
Modality



normal

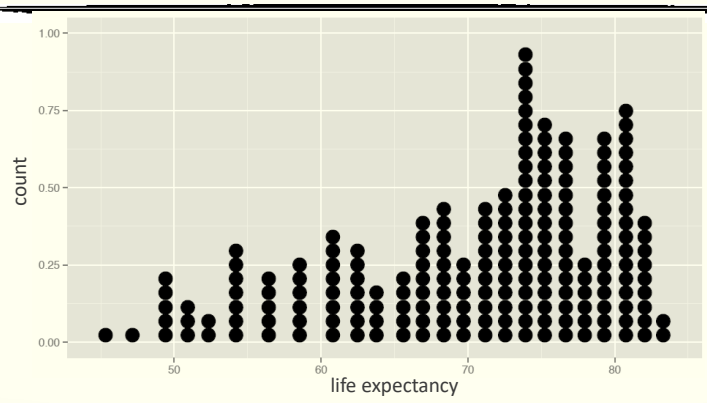


heights at elementary school



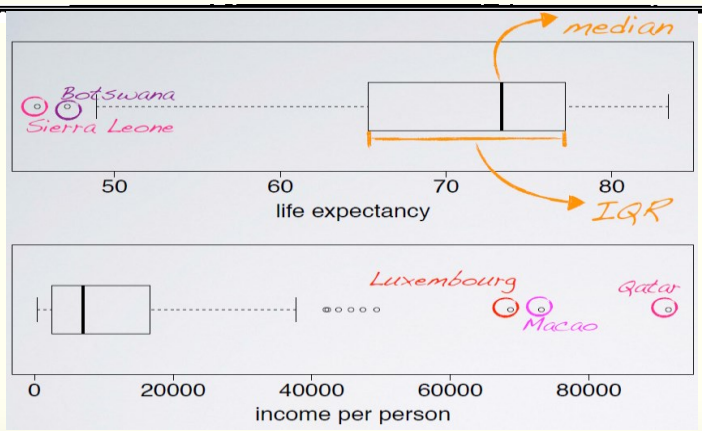
last digit of national ID

Dotplot



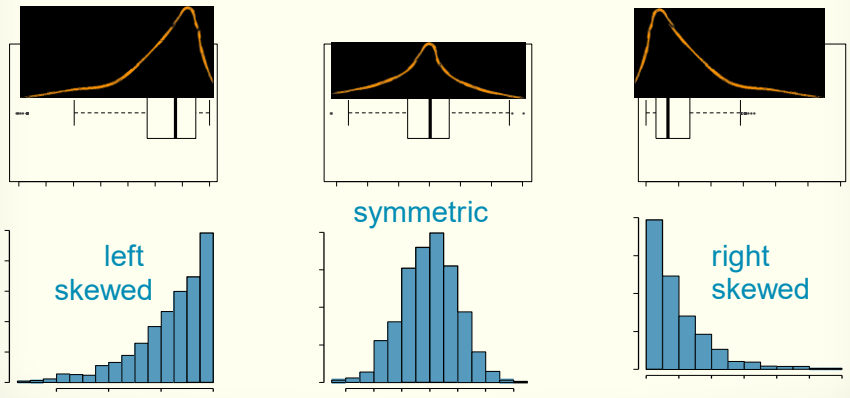
- Useful when individual values are of interest
- Can get busy as the sample size increases

Box plot

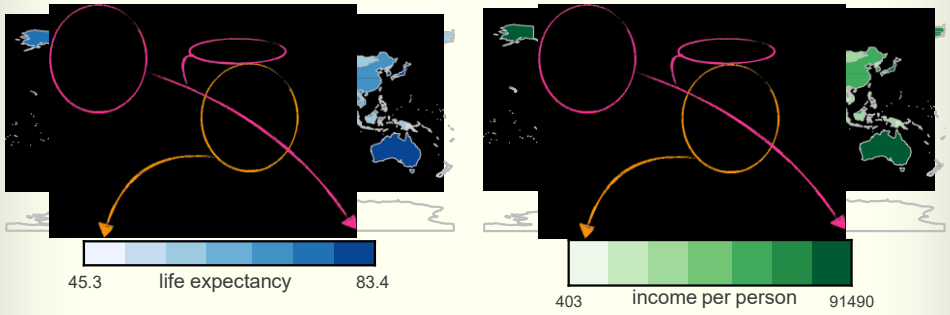


- Useful for highlighting outliers, median, IQR.

Determining the skewness from a box plot

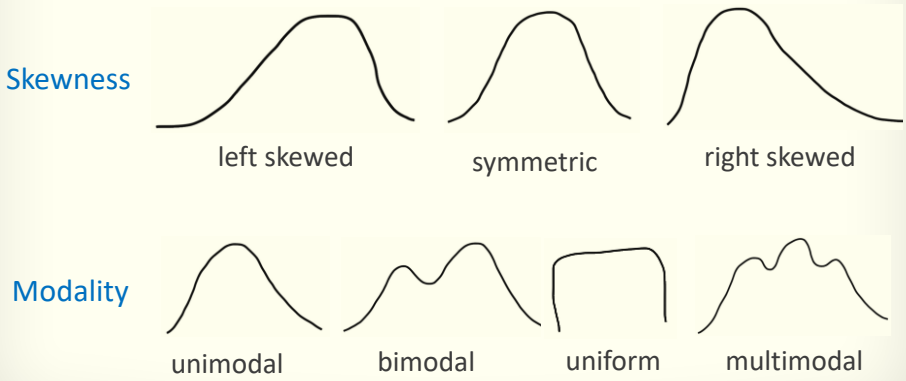


Intensity Map



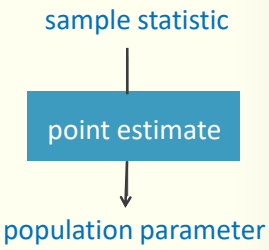
- Useful for highlighting the spatial distribution.

Shapes of Numerical Distributions



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



Example

- Nine students exam score:

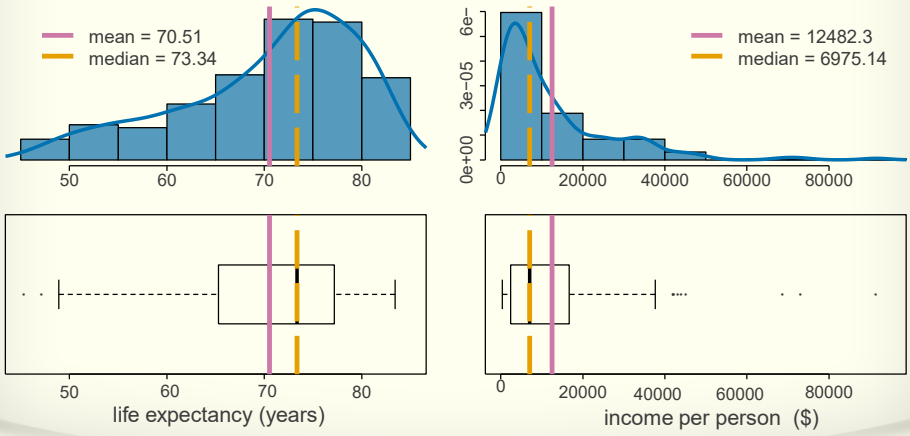
75, 69, 88, 93, 95, 54, 87, 88, 27

mean: $\frac{75+69+88+93+95+54+87+88+27}{9} = 75.11$

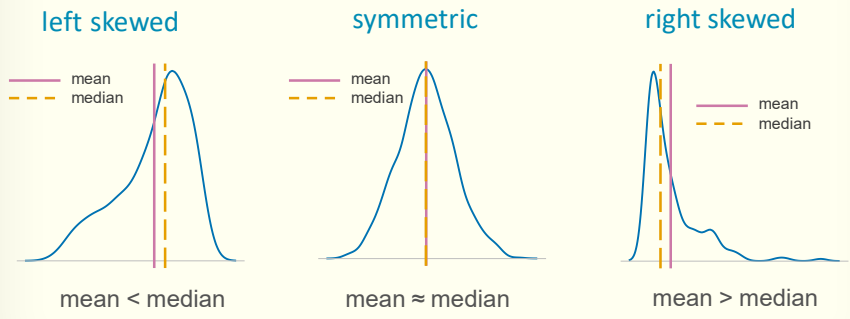
mode: 88

median: 27, 54, 69, 75, 88, 88, 93, 95

Relation between Mean and Median

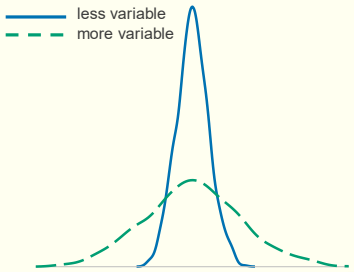


Skewness vs. Measures of Center



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

Strandard 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

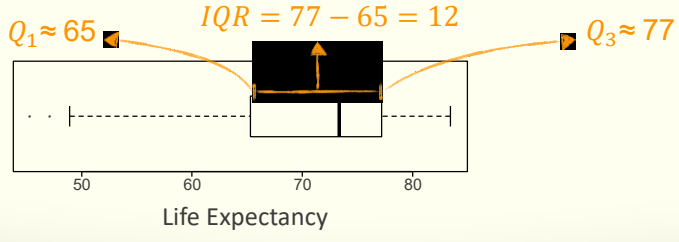
- Example: Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

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

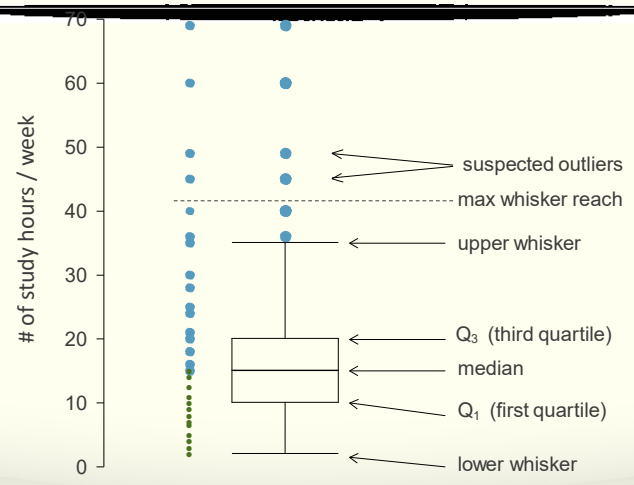
Interquartile Range

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

$$IQR = Q_3 - Q_1$$



Anatomy of a Boxplot



Whiskers

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

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

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

- Example:

  IQR : 20 − 10 = 10

  max upper whisker reach = 20 + 1.5 × 10 = 35

  max lower whisker reach = 10 − 1.5 × 10 = −5

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

Outliers

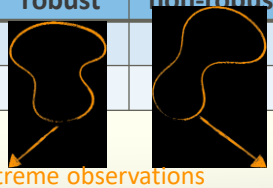
- Why it is important to look for outliers?
- Examination of data for possible outliers serves many useful purposes, including:
  1. Identifying strong skew in the distribution.
  2. Identifying data collection or entry errors.
  3. Providing insight into interesting properties of the data.

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		
spread		



skewed, with extreme observations

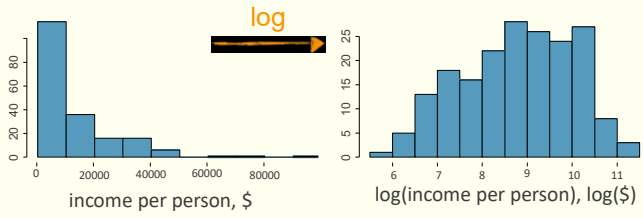
symmetric

Data Transformation

- A transformation is a rescaling of the data using a function.
  - Log transformation
  - Square root transformation
  - Inverse transformation
- When data are very strongly skewed, we sometimes transform them so they are easier to model.

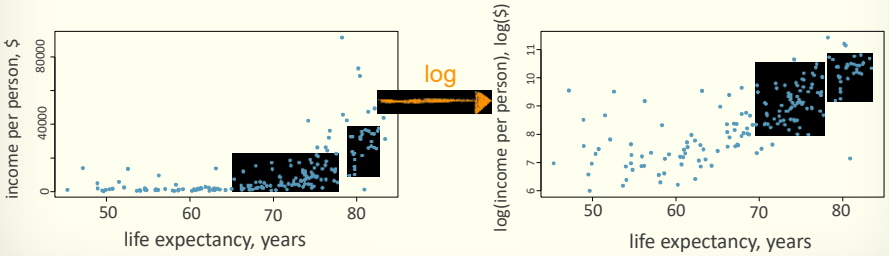
(Natural) Log Transformation

- Often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive.

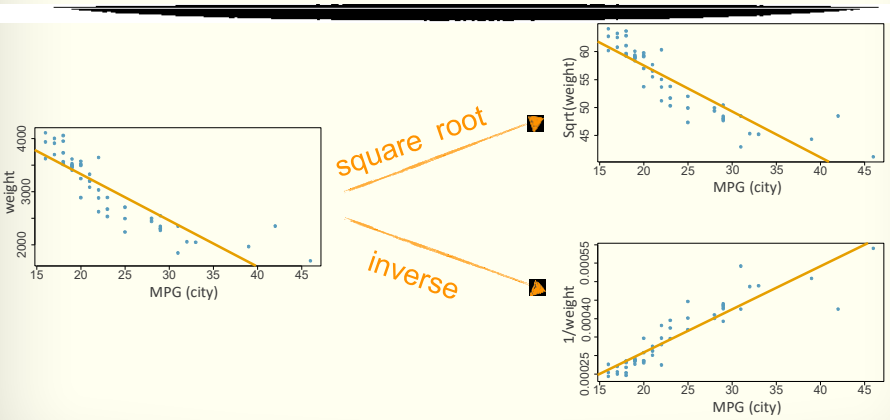


Log Transformation

- To make the relationship between the variables more linear, and hence easier to model with simple methods



Other Transformations



Goals of Transformation

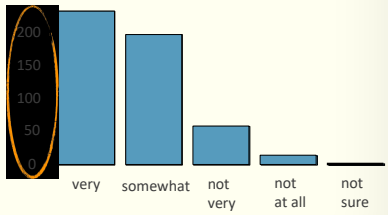
- To see the data structure differently
- To reduce skew and assist in modeling
- To straighten a nonlinear relationship in a scatterplot
- To model the relationship with simpler methods

Describing Categorical Variables

- Contingency tables
- Bar plots
- Segmented bar
- Mosaic plots
- Pie charts

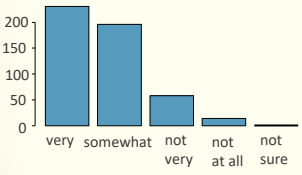
Frequency Table & Bar Plot

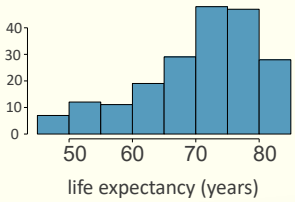
Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%



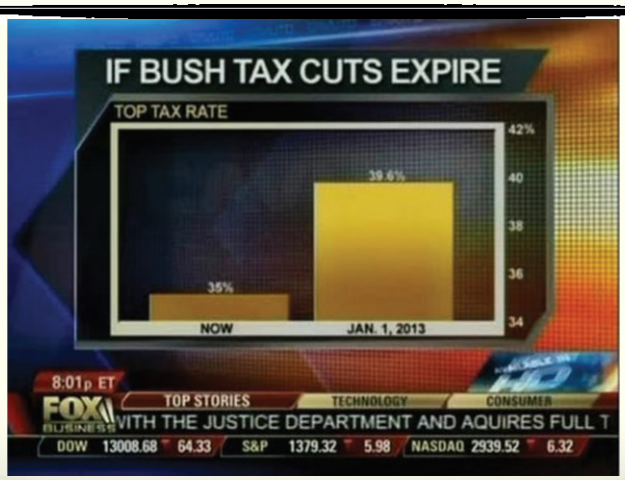
Bar Plots vs. Histograms

- 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



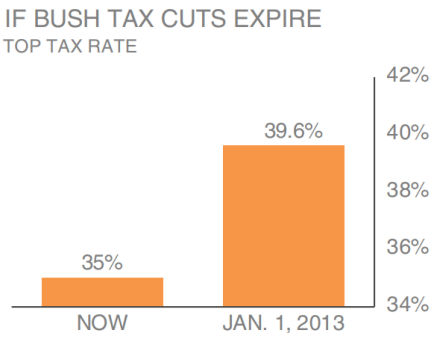


Bar Plot Abuse

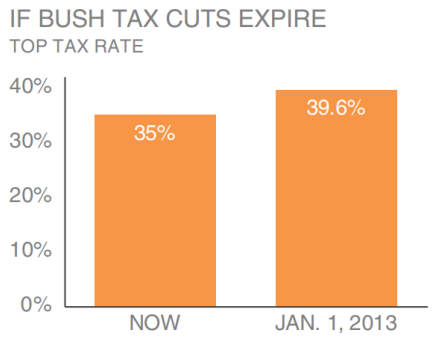


Bar Plot Abuse

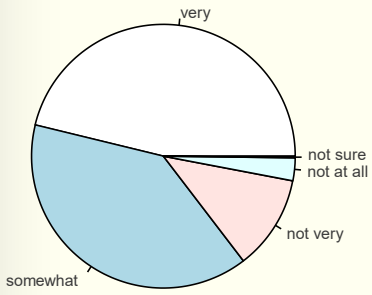
Non-zero baseline: as originally graphed

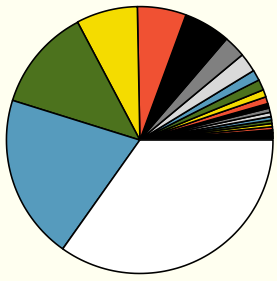


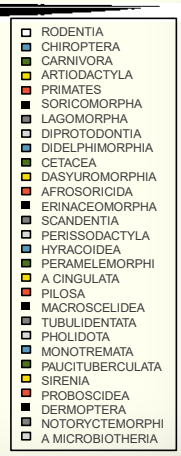
Zero baseline: as it should be graphed



~~Pie Chart?~~ NO!

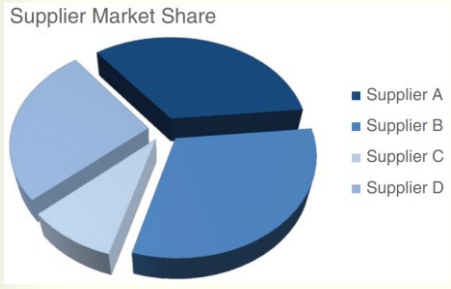


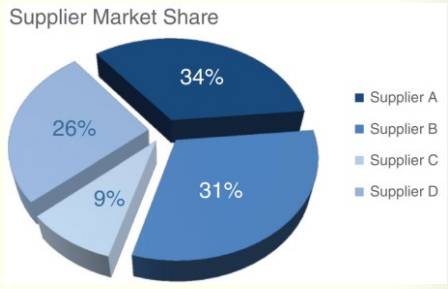




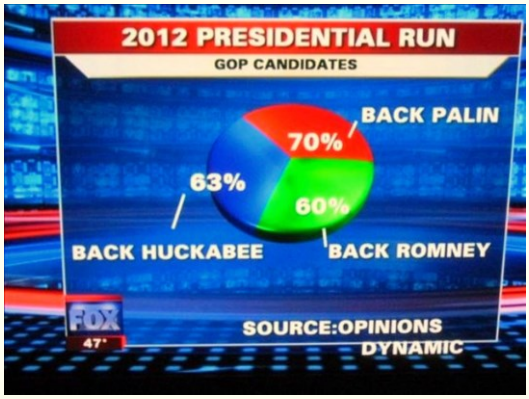
- A pie chart is actually much less informative than a bar plot.

To be avoided: Pie Charts





Terrible Pie Chart



Contingency Table

		Income				
		< $40K	$40-80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

- A table that summarizes data for two categorical variables is called a contingency table.

Relative Frequency

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

< $40K: 128/202 = 63% find it very difficult to save

$40K-$80K: 63/148 = 43%

$80K: 31/124 = 25%

Refused: 9/26 = 35%

⇒ feelings about difficulty of saving money and income are associated (dependent)

Contingency Table

➢ In January 1971, a Gallup Poll asked, "A proposal has been made in Congress to require the US government to bring home all US troops before the end of the year. Would you like to have your congressman vote for or against this proposal?" Guess the results, for respondents in each education category.

	Elementary Education	High School Education	College Education	Total
For Withdrawal				73%
Against Withdrawal				27%
Total	100%	100%	100%	100%

Contingency Table

- In January 1971, a Gallup Poll asked, "A proposal has been made in Congress to require the US government to bring home all US troops before the end of the year. Would you like to have your congressman vote for or against this proposal?" Guess the results, for respondents in each education category.

	Elementary Education	High School Education	College Education	Total
For Withdrawal	80%	75%	60%	73%
Against Withdrawal	20%	25%	40%	27%
Total	100%	100%	100%	100%

Simpson's Paradox

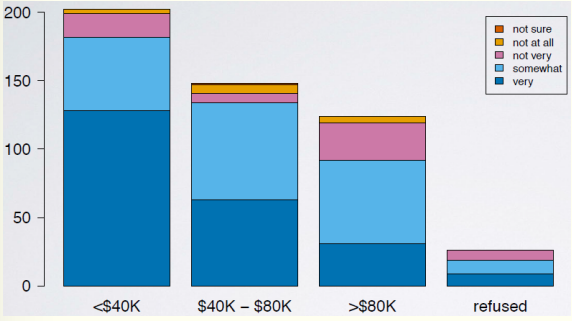
- A phenomenon in which a trend appears in different groups of data but disappears or reverses when the groups are combined.

Major	Women acceptance rate	Men acceptance rate
Computer science	27%	25%
Economics	26%	22%
Engineering	32%	26%
Medicine	24%	24%
Veterinary Medicine	16%	12%
Total	**23%**	**24%**

Simpson's Paradox

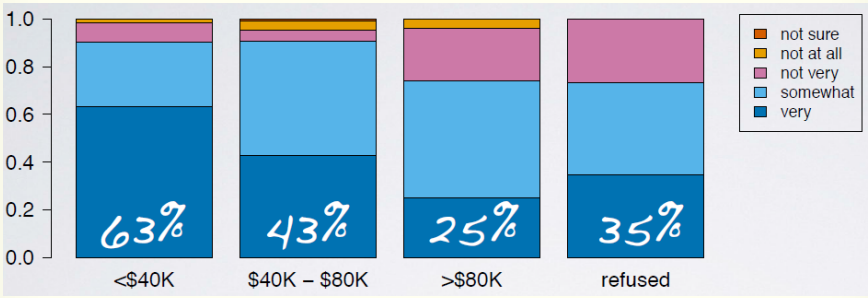
	Male	Female
Major A	$\frac{50}{150} = 30\%$	$\frac{180}{360} = 50\%$
Major B	$\frac{200}{250} = 80\%$	$\frac{36}{40} = 90\%$
Combined	$\frac{250}{400} = 62.5\%$	$\frac{216}{400} = 54\%$

Segmented Bar Plot

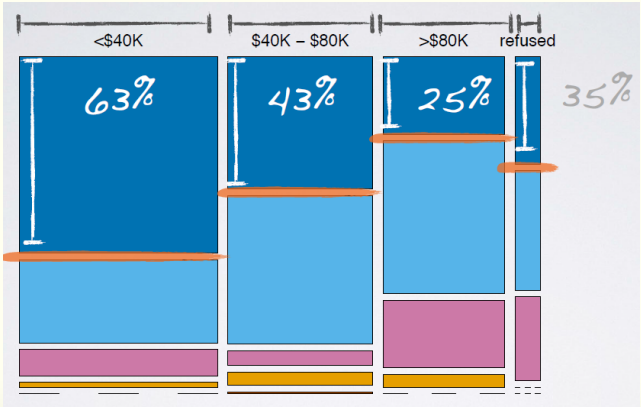


- Useful for visualizing conditional frequency distributions
- Compare relative frequencies to explore the relationship between the variables

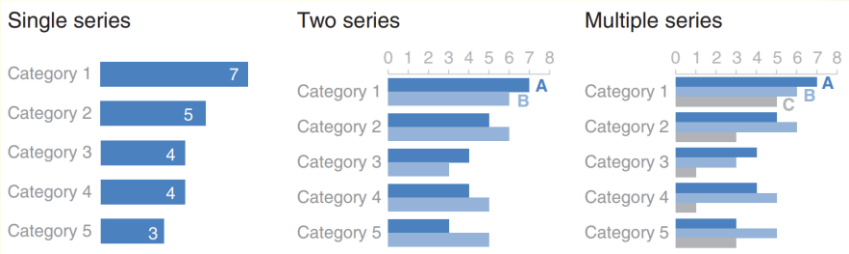
Relative Frequency Segmented Bar Plot



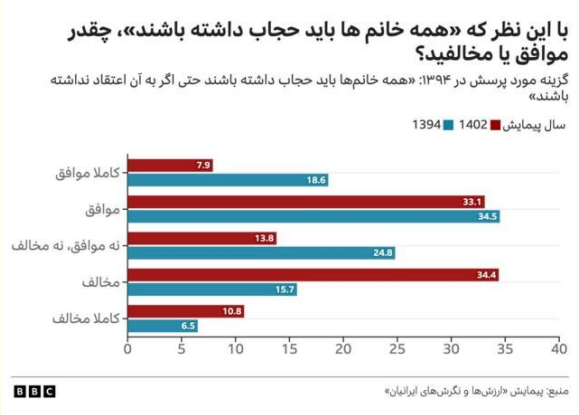
Mosaic Plot

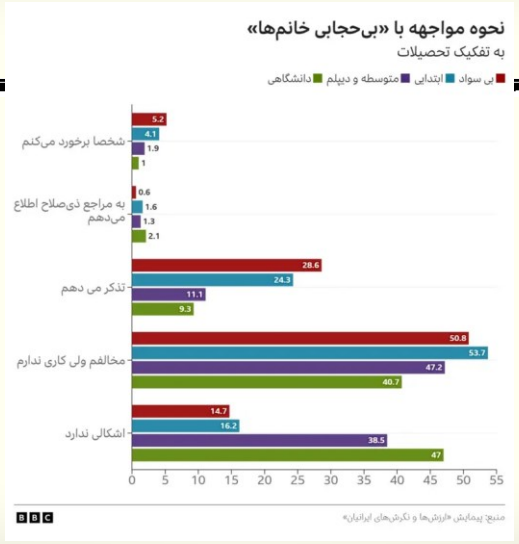


Horizontal Bar Plot

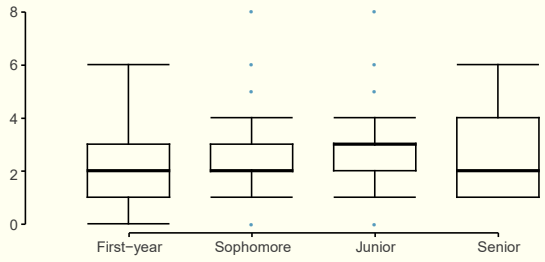


Relative Frequency Bar Plot





side-by-side box plots



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

Case Study: Gender Discrimination

- 48 male bank supervisors given the same personnel file, asked to judge whether the person should be promoted.
- Files were identical, except for gender of applicant
- Random assignment
- 35 / 48 promoted
- Are females are unfairly discriminated against?
  - Is this an observational study or an experiment?

Case Study: Gender Discrimination

		Promotion		
		Promoted	Not Promoted	total
Gender	Male	21	3	24
	Female	14	10	24
	total	35	13	48

% of males promoted: 21/24 ≈ 88%

% of females promoted: 14/24 ≈ 58%

Two Competing Claims

Null Hypothesis	Alternative Hypothesis
"There is nothing going on"	"There is something going on"
promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance	promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance.
H_0	H_A

A trial as a hypothesis test

- Hypothesis testing is very much like a court trial.
  - H_0: Defendant is innocent
  - H_A: Defendant is guilty
- We then present the evidence
  - Collect data.
- Then we judge the evidence:
  "Could these data plausibly have happened by chance if the null hypothesis were true?"
  - YES: Fail to reject H_0
  - NO: Reject H_0

Hypothesis Testing Framework

- Start with a null hypothesis (H_0) that represents the status quo
- Set an alternative hypothesis (H_A) that represents the research question, i.e. what we're testing for.
- Conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods
  - if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis
  - if they do, then reject the null hypothesis in favor of the alternative

Simulation-based Inference

- Suppose the bankers' decisions were independent of gender. Then, if we conducted the experiment again with a different random arrangement of files, differences in promotion rates would be based only on random fluctuation.
- We can actually perform this randomization, which simulates what would have happened if the bankers' decisions had been independent of gender but we had distributed the files differently.

Simulation Scheme

- We thoroughly shuffle 48 personnel files, 24 labeled male-sim and 24 labeled female-sim, and deal these files into two stacks.
- We will deal 35 files into the first stack, which will represent the 35 supervisors who recommended promotion.
- The second stack will have 13 files, and it will represent the 13 supervisors who recommended against promotion.

		Promotion		
		Promoted	Not Promoted	total
Gender-sim	Male-sim	18	6	24
	Female-sim	17	7	24
	total	35	13	48

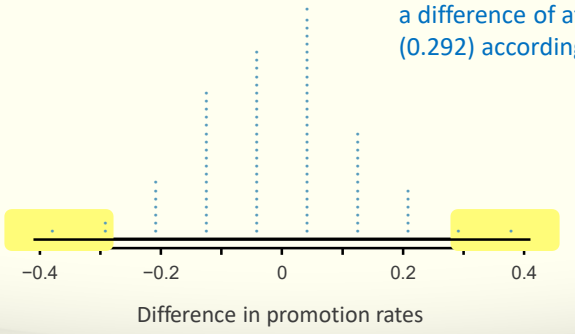
difference: $\frac{18}{24} - \frac{17}{24} = \frac{1}{24} = 0.042$

Making a Decision

- Results from the simulations look like the data → the difference between the proportions of promoted files between males and females was due to chance
  - promotion and gender are independent
- Results from the simulations do not look like the data → the difference between the proportions of promoted files between males and females was not due to chance, but due to an actual effect of gender
  - promotion and gender are dependent

Distribution of Differences from Chance

- Repeat the simulation 100 times:
  - 1st time: difference = 0.042
  - 2nd time: difference = 0.208
  - …
- How often would you observe a difference of at least 29.2% (0.292) according to this plot?



Simulation-based Inference

- Set a null and an alternative hypothesis
- Simulate the experiment assuming that the null hypothesis is true
- Evaluate the probability of observing an outcome at least as extreme as the one observed in the original data (p-value)
- If this probability is low, reject the null hypothesis in favor of the alternative