Introduction to Statistics

DCOVA framework

Define	Collect	Organize	Visualize	Analyze
• Define problem/opportunity • Define variable • Define data - set of observations or cases • Statistics - the methods that transform data into useful information for decision making	• Internal data sources • External data sources • Questionnaires • Interviews • Observations • Experiments	• Using MS-excel, database • Data entry • Data tabulation	• Column chats • Bar charts • Pie charts • Line charts • Scatter	• Descriptive statistics • Inferential statistics

Classical statistics vs EDA

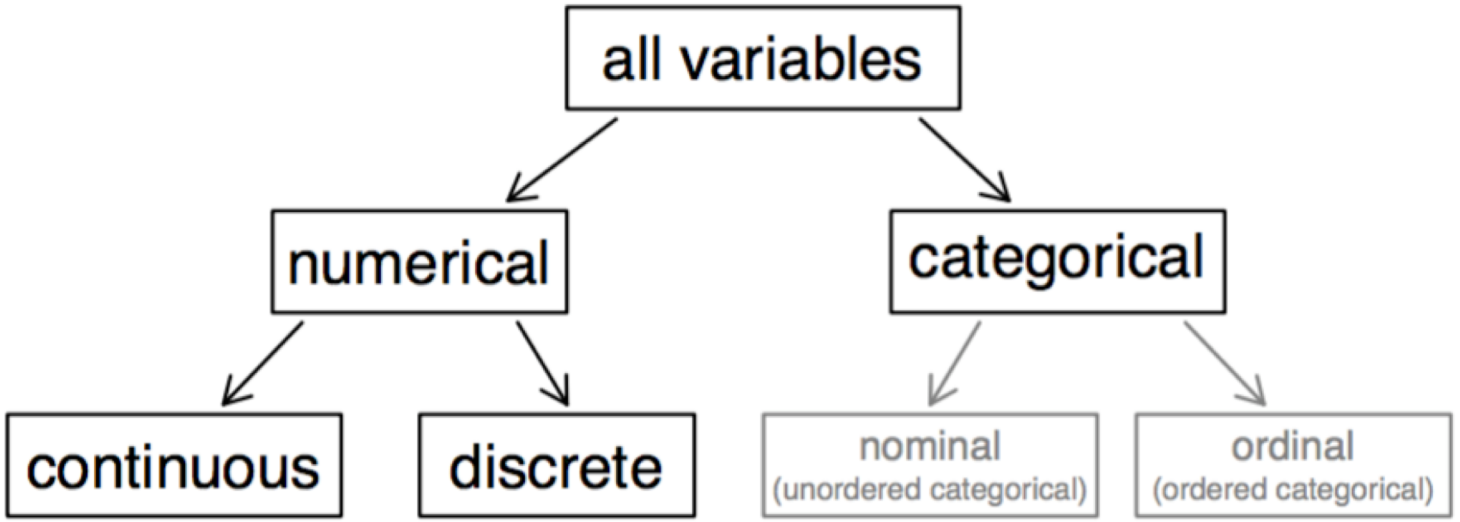
Classical Statistics

- ANNOVA
- t-test,
- chi-squared tests,
- F test

Exploratory data analysis

- Scatter plots
- charter plots
- box plot
- Histogram
- bi-histogram
- probability plots
- residual plots
- mean plots

Variable Types

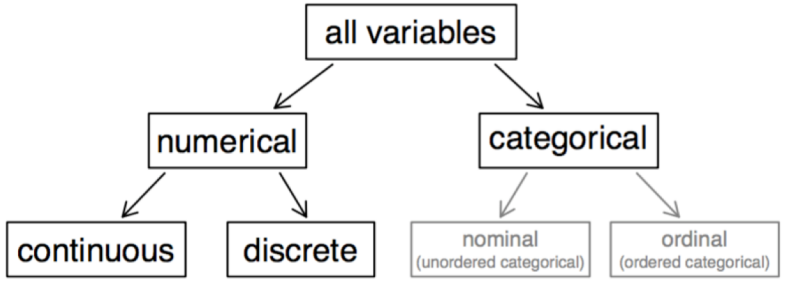


Height	Rank	Gender	Grade in school
Salary	Year of birth	Country	Rank
BMI		Occupation	

Introduction to Data

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Look at each column and identify which type of variable it is.



Visualization Strategy

Single Variable

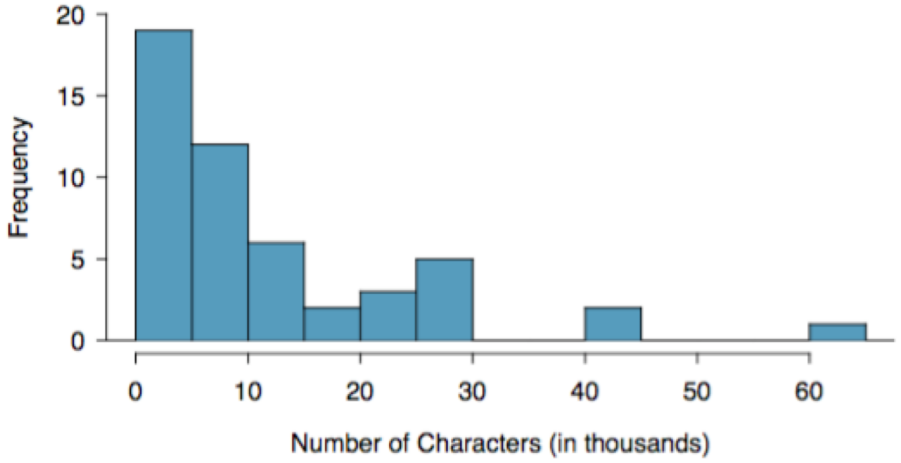
- Continuous
  - Distribution - histogram
  - Outliers - boxplot
- Categorical
  - Frequency plot

Two variable association

- Continuous vs Continuous - scatter plot
- Continuous vs Categorical - vertical barplot
- Categorical Continuous - horizontal bar plot
- Categorical vs Categorical - heatmap

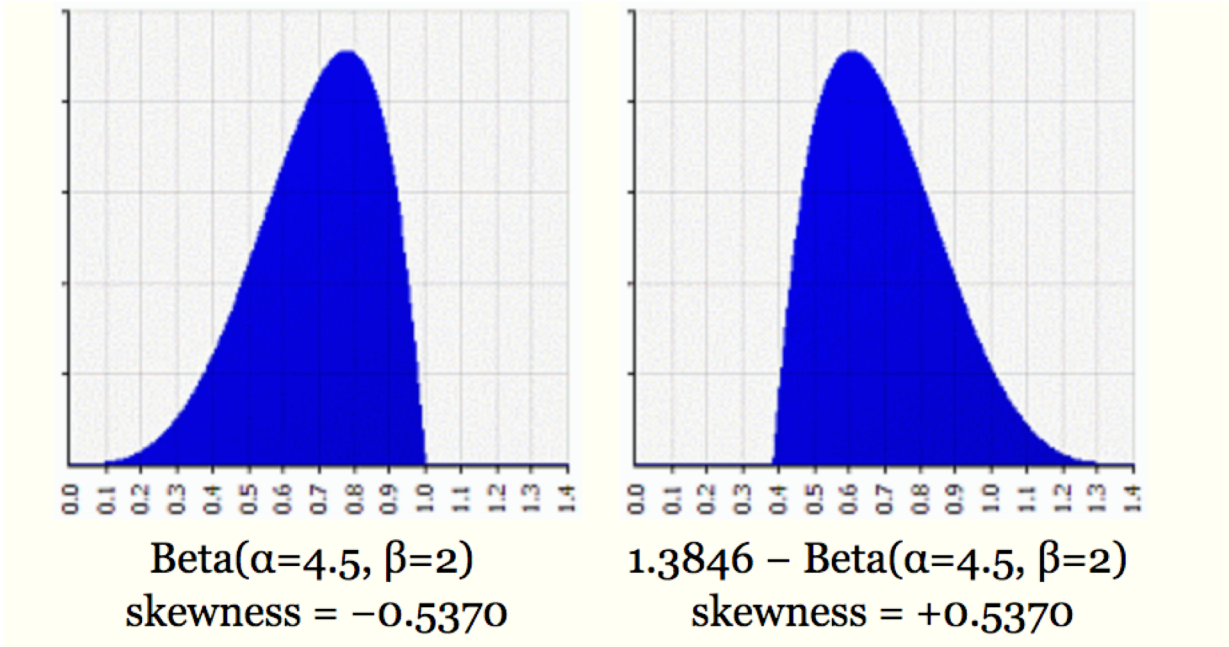
Histogram

- Histograms provide a view of the data density. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the shape of the data distribution.
- Data sets with the reverse characteristic – a long, thin tail to the left – are said to be left skewed.
- Data sets that show roughly equal trailing on in both directions are called symmetric.
- In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes - unimodal, bimodal, and multimodal



What to observe from the histogram

- The distribution's shape - one mode (peak) or more than one?
  - If it's unimodal (has just one peak), like most data sets,
- Whether it's symmetric or skewed to one side.
  - **Skewed Right:** If the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed
  - **Skewed Left:** If the peak is toward the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed



Skewed Left

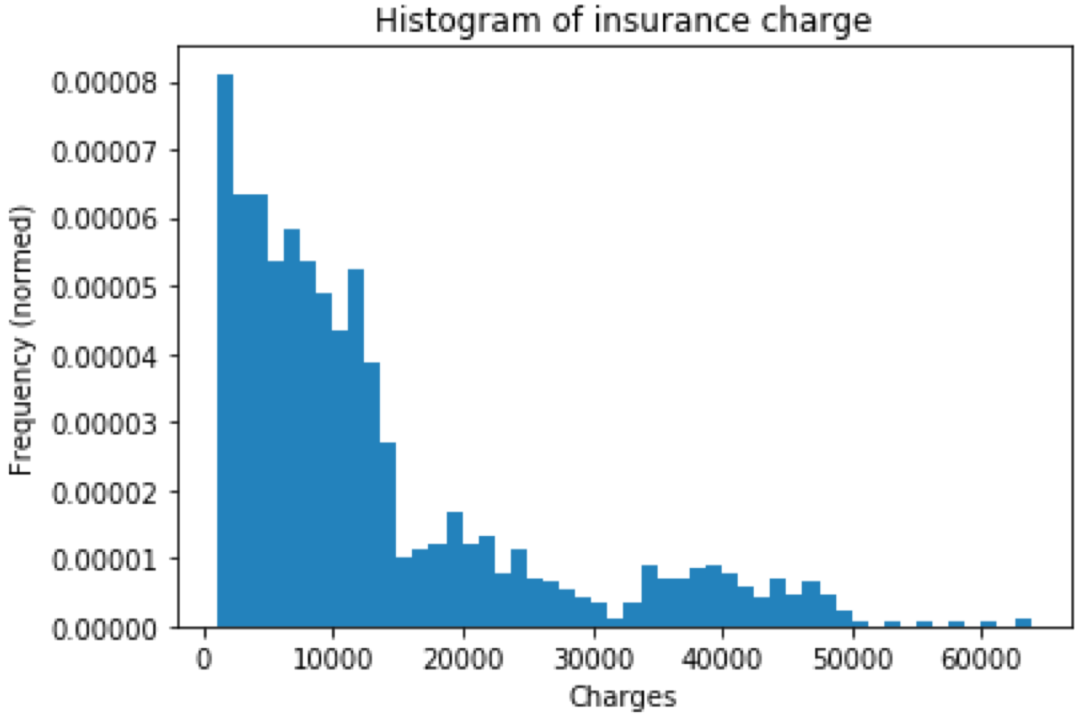
Skewed Right

Skewness

$$Z_{score} = \frac{x - \bar{x}}{SD_x}$$

$$Skeness = \frac{1}{n}\sum_{i=1}^{n} {z_i}^3$$

Negative value of skewness indicates left skewed data and positive value indicates right skewed data



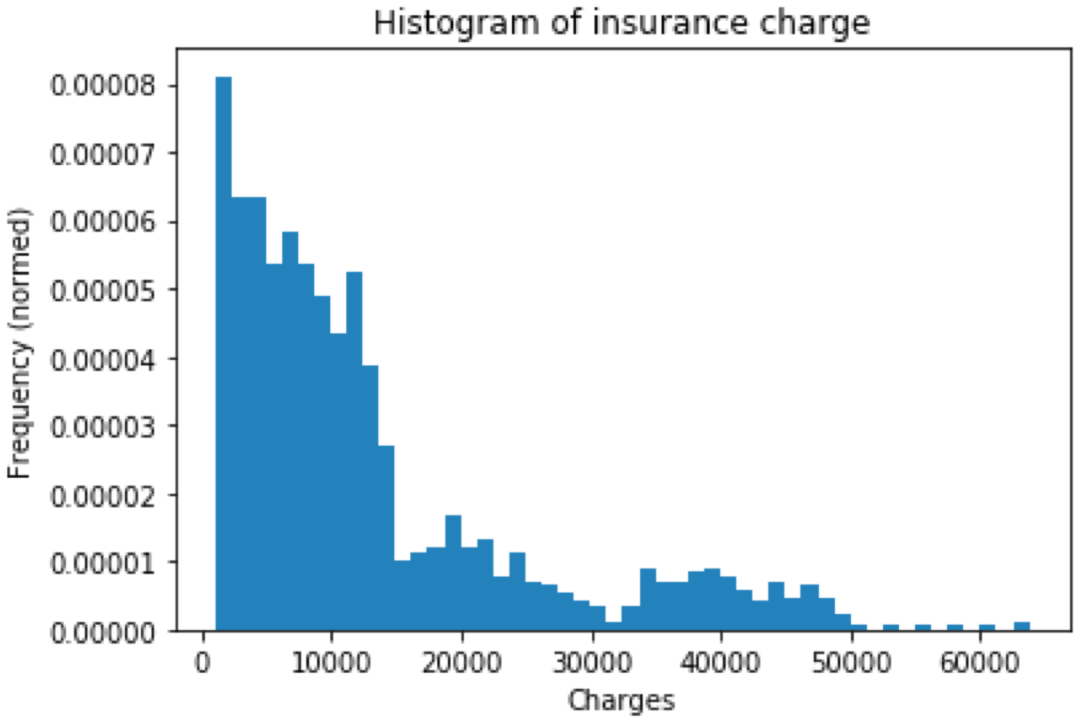
Skewness: 1.514

Kurtosis

$$Skeness = \frac{1}{n}\sum_{i=1}^{n} {z_i}^4$$

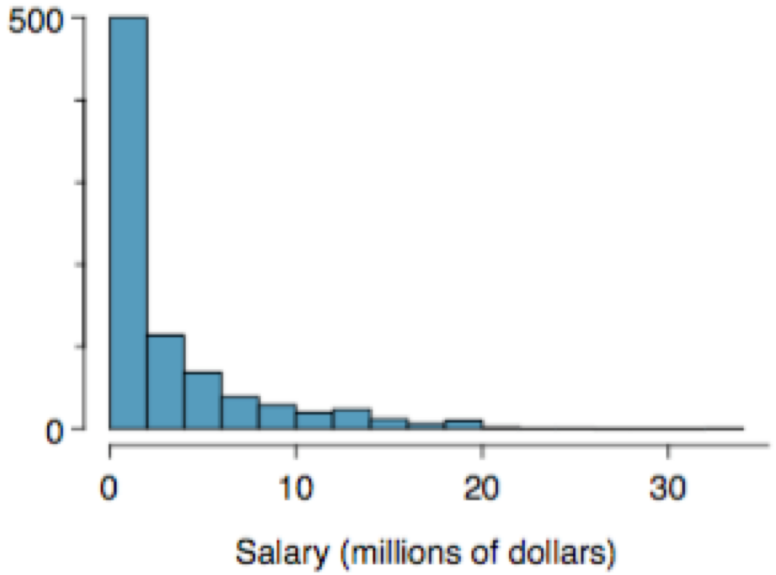
Interpretation:

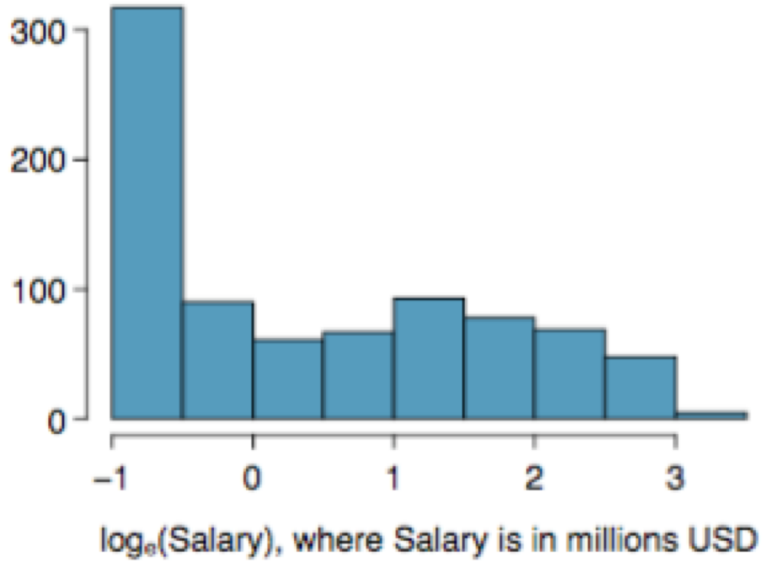
- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called mesokurtic.
- A distribution with kurtosis <3 (excess kurtosis <0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
- A distribution with kurtosis >3 (excess kurtosis >0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.



Kurtosis: 1.59

Transform long tail data





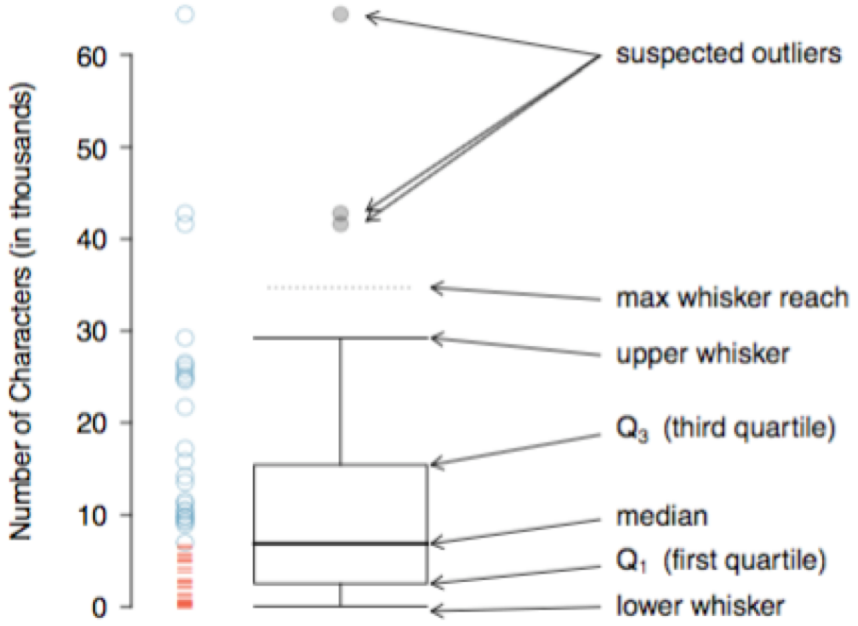
Transformed data are

- Less skewed
- Outliers are usually less extreme

Box plots, quartiles, and the median

- A box plot summarizes a data set using five statistics while also plotting unusual observations
- Any observation that lies beyond the whiskers labeled with a dot are called outliers

IQR: The total length of the box
Upper whisker: min(Highest point, Q3 +1.5 * IQR)
Lower whisker: max(Lowest point, Q1 - 1.5 * IQR)



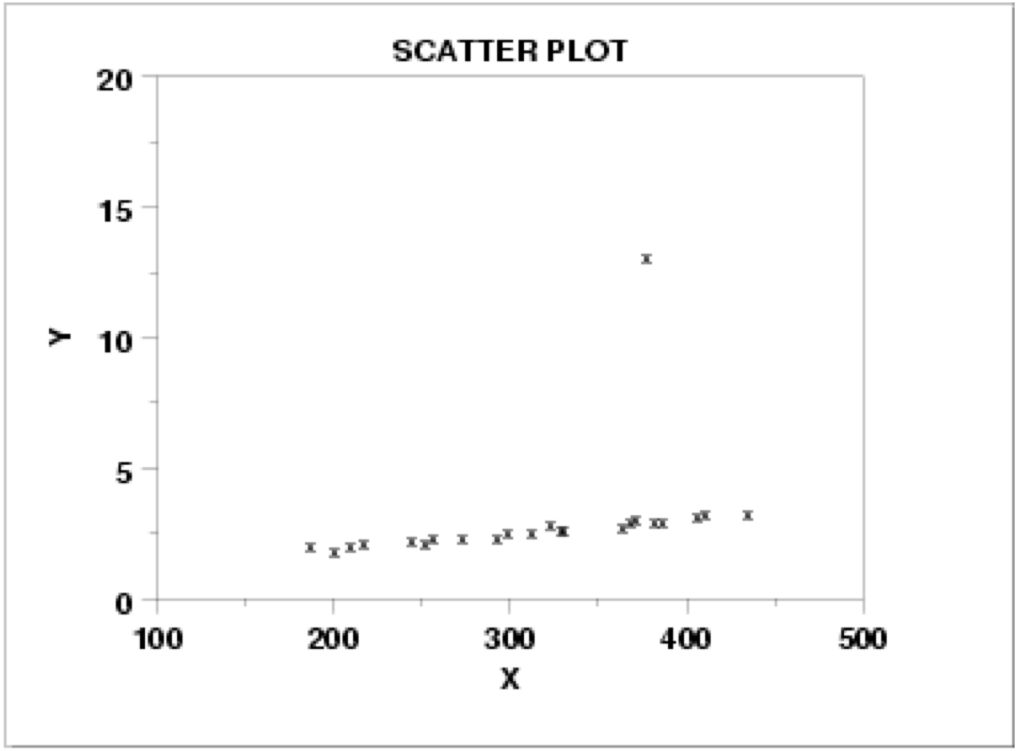
classic box-and-whisker plot

Outlier detection

- Box plot is a commonly used for outlier detection
- You can use Mean +/- 3 * SD as boundary to detect outliers. Although, mean itself is not a robust statistic. So, this is not an effective technique

Outlier

- What outliers reveal
  - Identifying strong skew in the distribution.
  - Identifying data collection or entry errors.
  - Providing insight into interesting properties of the data.
- While modeling - the outliers should be excluded from such model fitting.
  - For example, if all the data here are included in a linear regression, then the fitted model will be poor virtually everywhere. If the outlier is omitted from the fitting process, then the resulting fit will be excellent almost everywhere (for all points except the outlying point). Related: RANSAC models.



Outlier treatment

- Suspect the outlier variable and give null treatment if required
  - Example, you have year built of a house in house price prediction 2200
- Cap the values
- Take log scale/min-max scale to minimize the impact of the extreme values

Robust Statistic

- Median and IQR are called robust estimates because extreme observations have little effect on their values.
- The mean and standard deviation are much more affected by changes in extreme observations.

	robust		**not robust**	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 64,401 observation	6,768	11,702	10,521	10,798
move 64,401 to 150,000	6,890	12,875	13,310	22,434

Mean, Median, Mode, Range

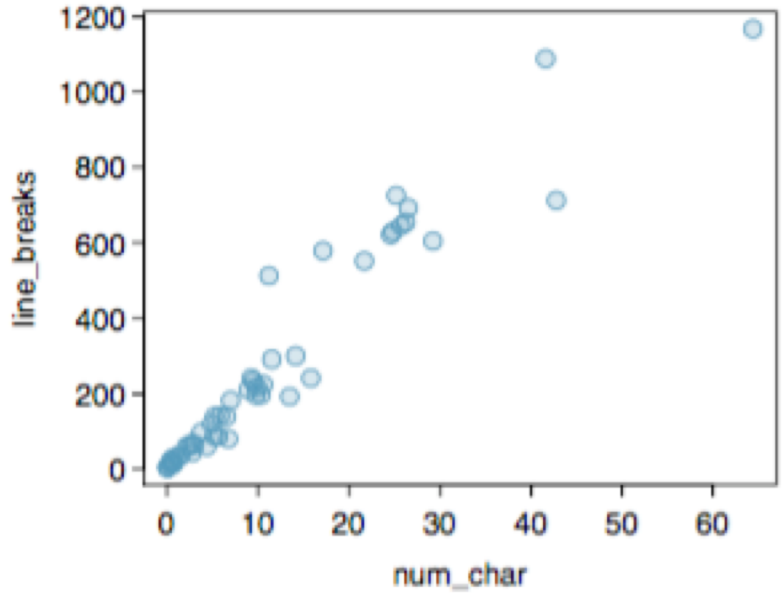
Mean: the mean is the average of all numbers $\frac{\sum_{i=1}^{m} x^{(i)}}{m}$

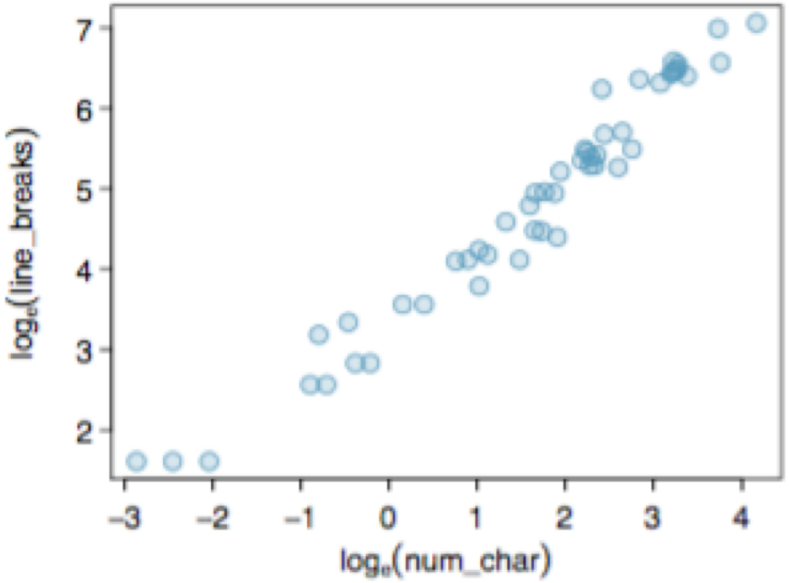
Median: the median is the middle number in a sequence of numbers. To find the median, organize each number in order by size; the number in the middle is the median.

Mode: the mode is the number that occurs most often within a set of numbers.

Range: The range is the difference between the highest and lowest values within a set of numbers

Scatter Plot





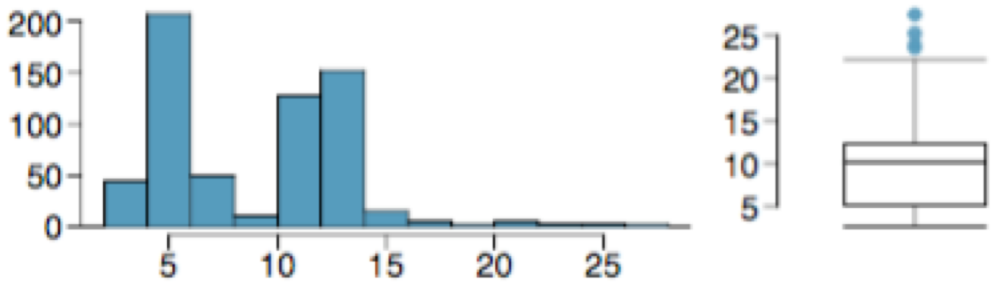
Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

Common goals in transforming data

- To see the data structure differently
- Reduce skew
- Assist in modeling
- Straighten a nonlinear relationship in a scatterplot

Exercise - Histogram and boxplot

Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?



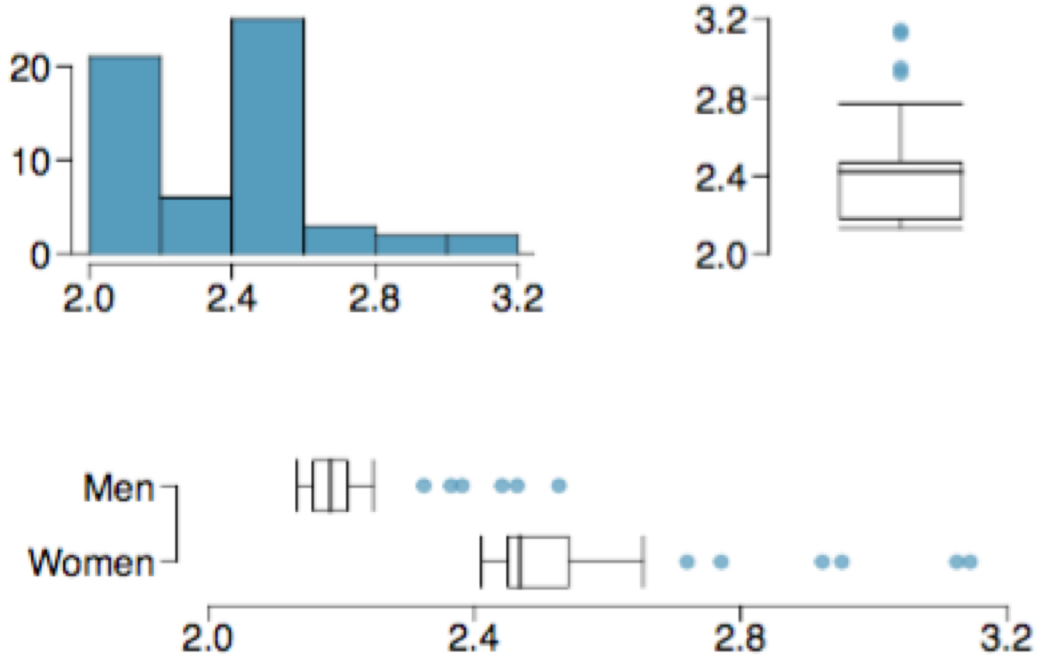
Exercise - Histogram and boxplot

The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.

(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown



Relationship measure between numeric variables

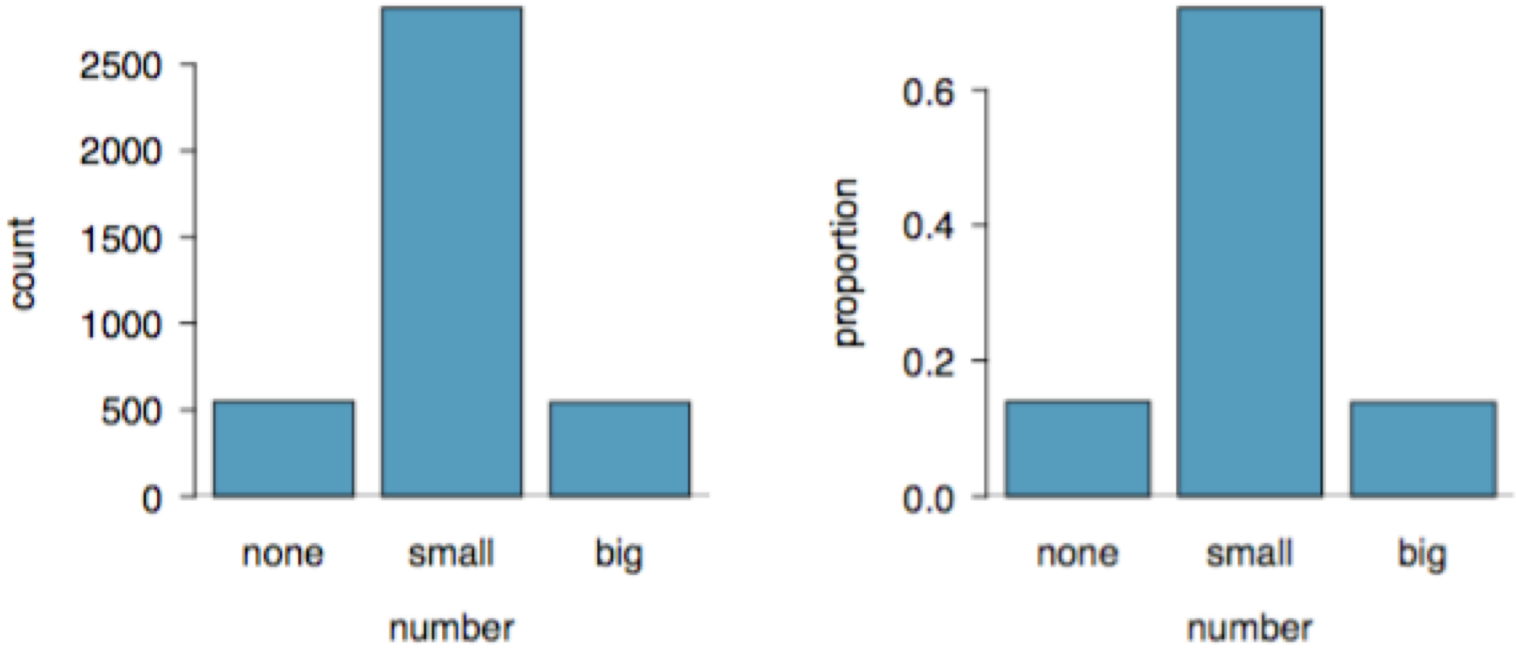
Covariance: tells whether two variables are positively or negatively (inversely) related.

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x - \bar{x})\,(y - \bar{y})$$

Correlation: in addition to whether two variables are positively or negatively related, it tells degree to which two variables move together.

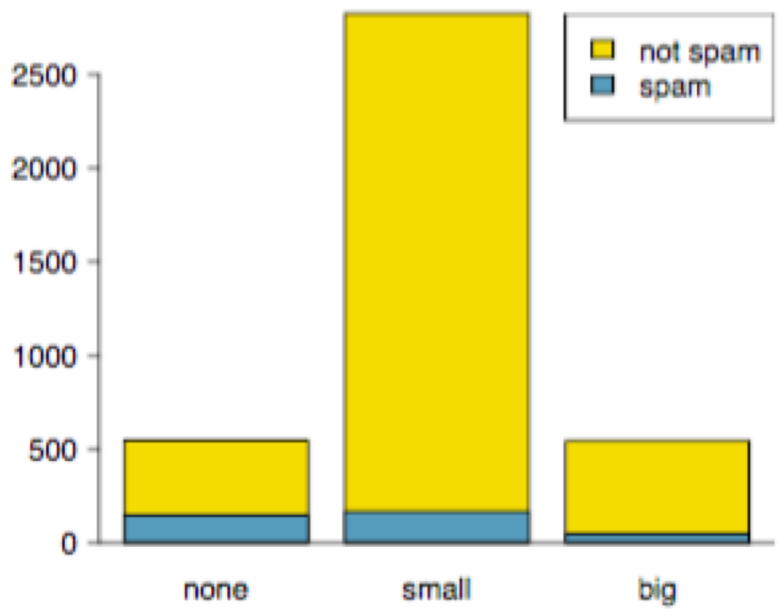
$$corr(x, y) = \frac{cov(x, y)}{SD_x SD_y}$$

Visualize categorical column

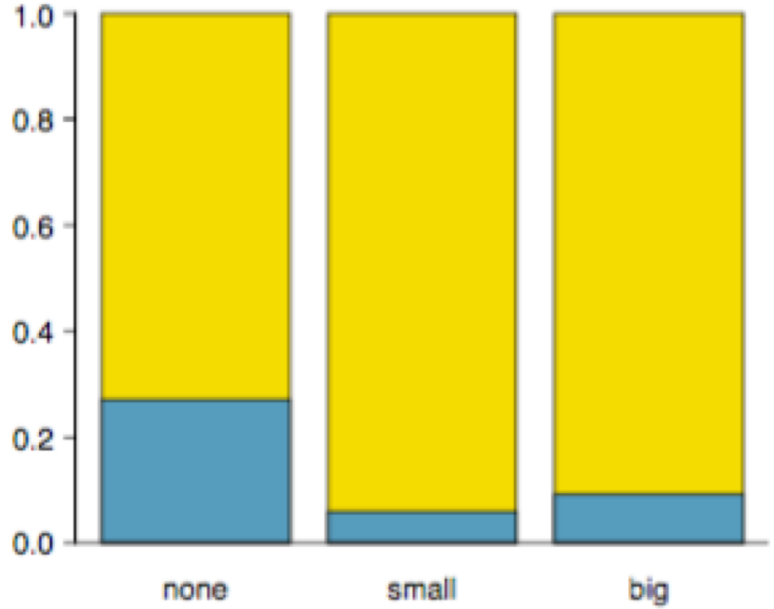


Two bar plots of number. The left panel shows the counts, and the right panel shows the proportions in each group.

Segmented Bar



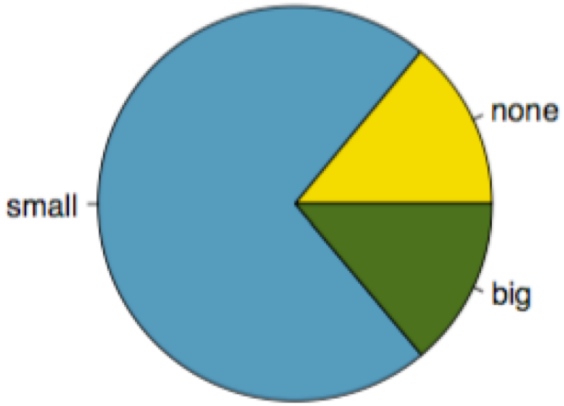
Segmented bar plot for numbers found in emails, where the counts have been further broken down by spam.

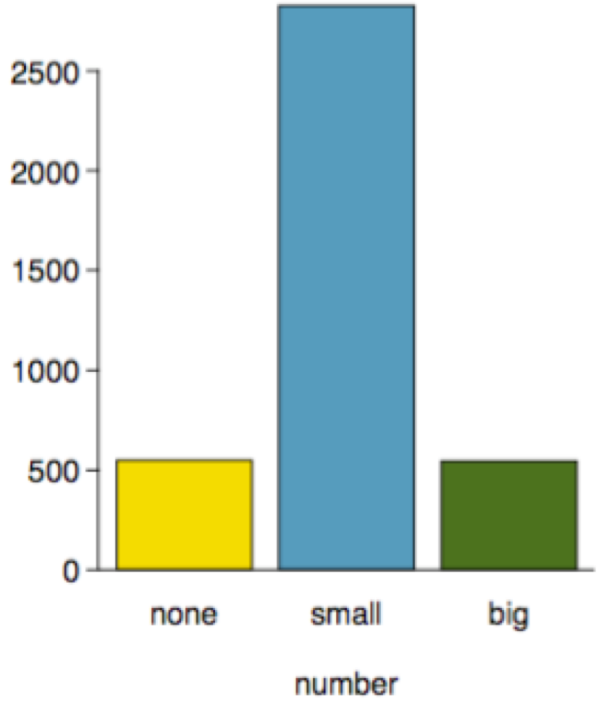


Standardized version of the figure on left

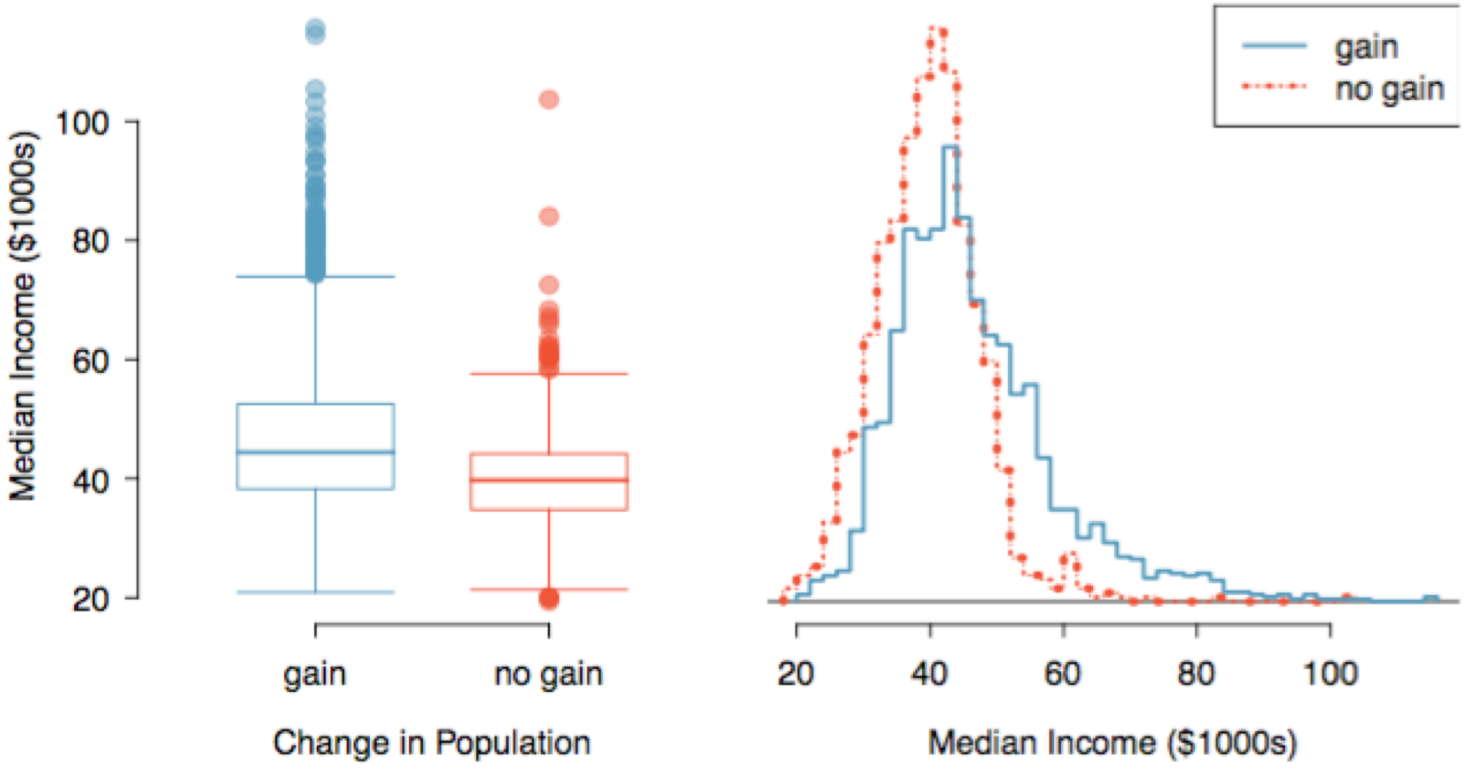
Pie chart

While pie charts are well known, they are not typically as useful as other charts in a data analysis. It is generally more difficult to compare group sizes in a pie chart than in a bar plot, especially when categories have nearly identical counts or proportions. Bar chart is more useful for displaying summary for categorical values since it easier to visually compare the values.





Compare numeric feature of 2 sets



Probability Distribution of Variables

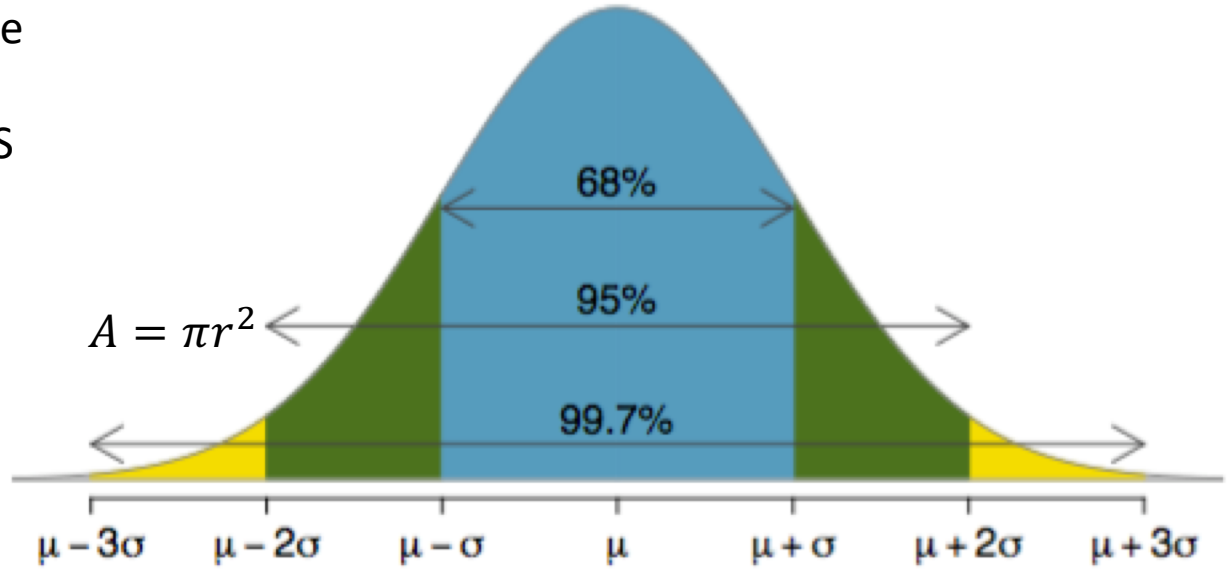
Normal Distribution

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\boldsymbol{\pi}\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

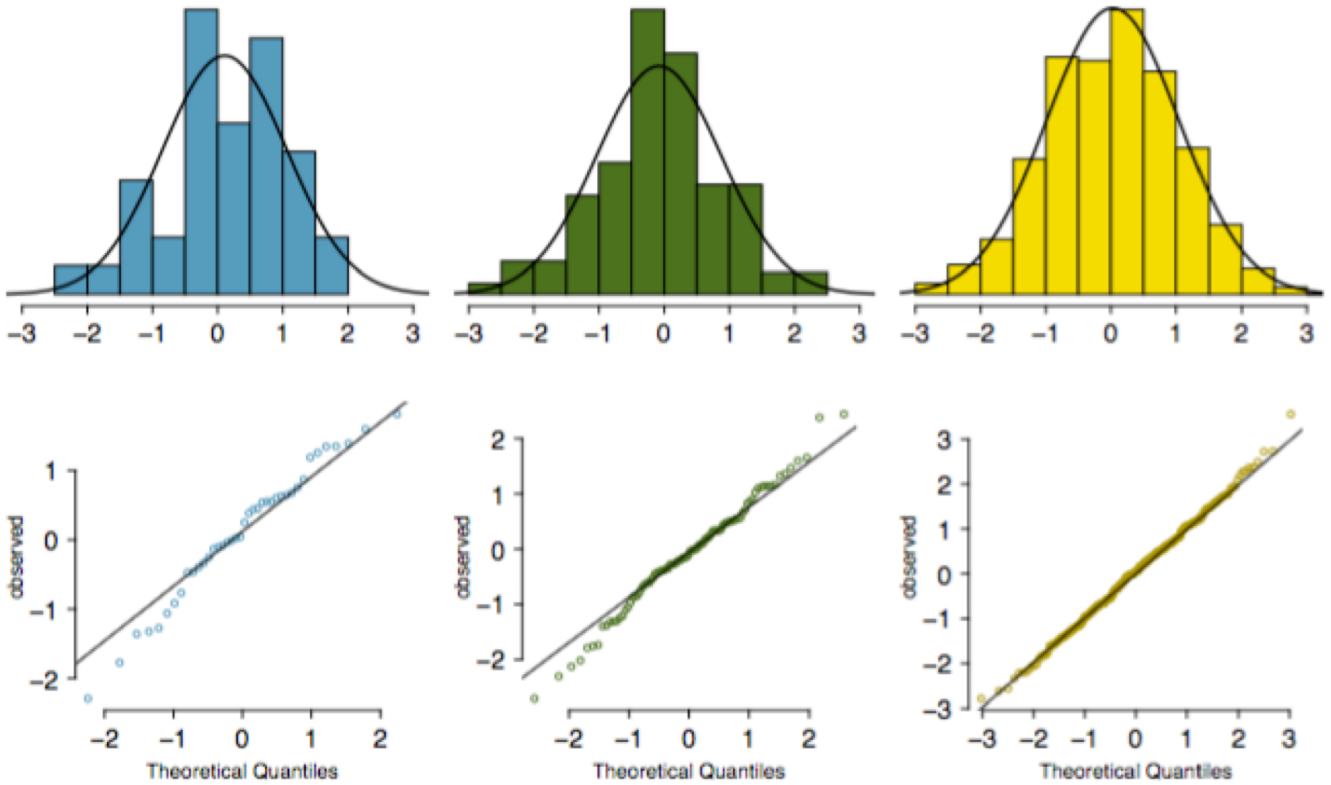
z-score

μ *is mean*, σ^2 *is variance*



Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

Visually test normality



Histogram plots

Normal probability plot or quantile-quantile plots. Closer to the straight line, closer to normality

(R code)

```
require(car)
qqPlot(x)
```

https://stats.stackexchange.com/questions/245396/how-to-read-the-x-axis-of-this-qqplot

Bernoulli Distribution

- When an individual trial only has two possible outcomes, it is called a Bernoulli random variable.
- If X is a random variable that takes value 1 with probability of success p and 0 with probability 1 - p, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p, \qquad \sigma = \sqrt{p(1-p)}$$

Question - Bernoulli Distribution

You tossed a fair coin 4 times that turned up head. What is probability the 5^{th} toss will turn up head?

Binomial Distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$$\mu = np \qquad \sigma^2 = np(1-p)$$

Is it Binomial? 4 conditions to check

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a success or failure.
(4) The probability of a success, p, is the same for each trial.

For large enough n, such that np > 10 and n(1-p) >10, binomial distribution can be approximated by normal distribution

Negative binomial distribution

The negative binomial distribution describes the probability of observing the kth success on the nth trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Poisson distribution

The Poisson distribution is often useful for estimating the number of events in a large population over a unit of time. For instance, consider each of the following events:

- No of people having a heart attack per day
- No of people getting married on Sun
- No of cars crossing a bridge per minute on average

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate . Then,

$$P(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and k! represents k-factorial. The letter e ≈ 2.718 is the base of the natural logarithm. The mean and standard deviation of this distribution is λ and $\sqrt{\lambda}$, respectively.

Exercise - Poisson Distribution

A coffee shop serves an average of 75 customers per hour during the morning rush.

(a) Which distribution we have studied is most appropriate for calculating the probability of a given number of customers arriving within one hour during this time of day?

(b) What are the mean and the standard deviation of the number of customers this coffee shop serves in one hour during this time of day?

(c) Would it be considered unusually low if only 60 customers showed up to this coffee shop in one hour during this time of day?

(d) Calculate the probability that this coffee shop serves 70 customers in one hour during this time of day?

Sampling Theory

Population vs Samples

Population statistics

- What is the average mercury content in swordfish in the Atlantic Ocean?
- Over the last 5 years, what is the average time to complete a degree for Duke under- graduate students?
- Does a new drug reduce the number of deaths in patients with severe heart disease?

Anecdotal Evidence

Data collected in a haphazard fashion are called anecdotal evidence.

- A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
- I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
- My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Problems with anecdotal evidence

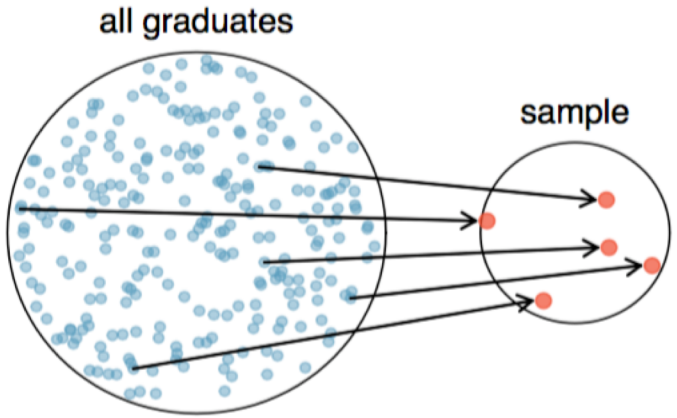
Problems with anecdotal evidence

- First, the data only represent one or two cases.
- Second, and more importantly, it is unclear whether these cases are actually representative of the population.

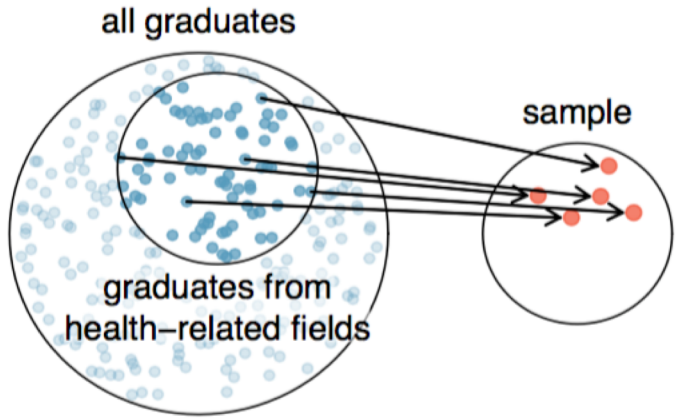
Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Random sample from a population

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?



[Random Sample] Five graduates are randomly selected from the population to be included in the sample.



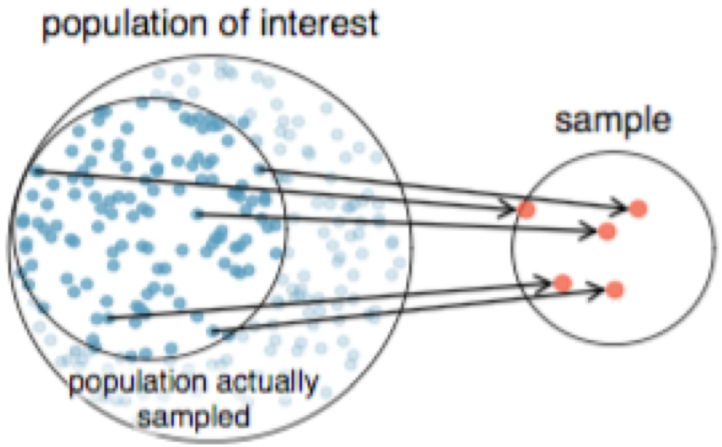
[Sample with bias] Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors dis- proportionately often.

Simple Random Sample

- Simple random sample, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.
- The act of taking a simple random sample helps minimize bias. However, bias can crop up in other ways.

Non-response bias

Even when people are picked at random, e.g. for surveys, caution must be exercised if the non-response is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are representative of the entire population.



Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Exercise: sampling

Q: We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?

A: Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

Convenience sample

- Another common downfall is a convenience sample, where individuals who are easily accessible are more likely to be included in the sample.
- For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City.

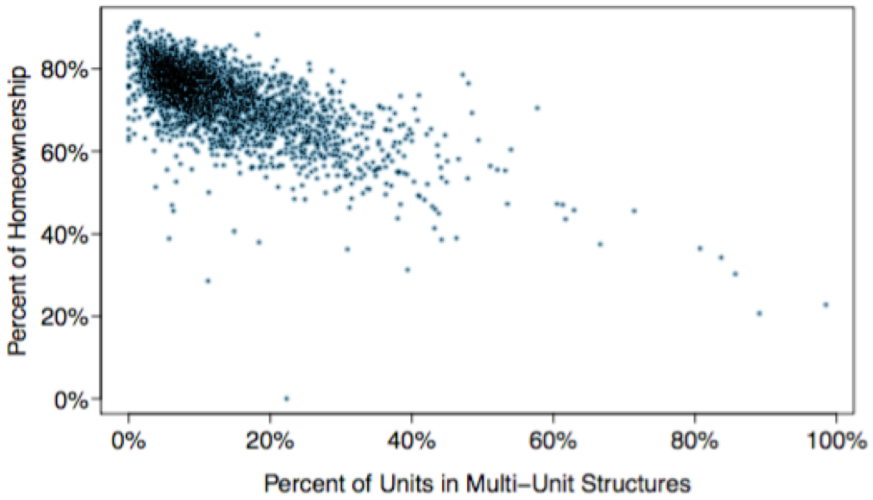
Explanatory and response variables

Is federal spending, on average, higher or lower in counties with high rates of poverty?

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

If we suspect poverty might affect spending in a county, then poverty is the explanatory variable and federal spending is the response variable in the relationship.

Find Association



A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties.

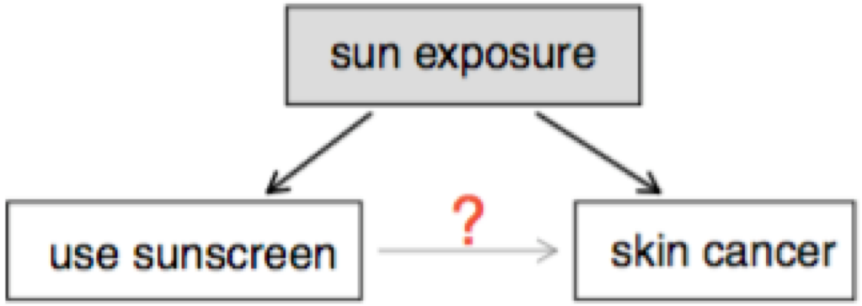
Caution: association does not imply causation

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

Exercise: correlation does not mean causation

Scenario: Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

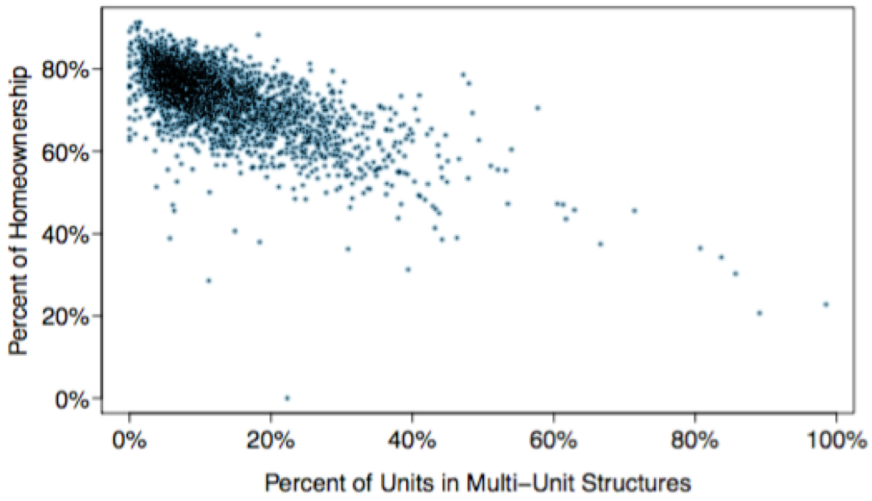
Possible Explanation: Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen and more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a **confounding variable or lurking variable or confouder**, which is a variable that is correlated with both the explanatory and response variables.

Example: correlation does not mean causation

Figure shows a negative association between the home-ownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in



One Possible Answer: Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

Explanatory or response variables are not always clear

In some cases, there is no explanatory or response variable. For example: If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

It is difficult to decide which of these variables should be considered the explanatory and response variable

Data collection

Two primary types of data collection we will look further

A. Observational study
B. Experiment

Observational study

- Researchers perform an observational study when they collect data in a way that does not directly interfere with how the data arise.
- For instance, researchers may collect information via surveys, review medical or company records, or follow a cohort of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

Experiment

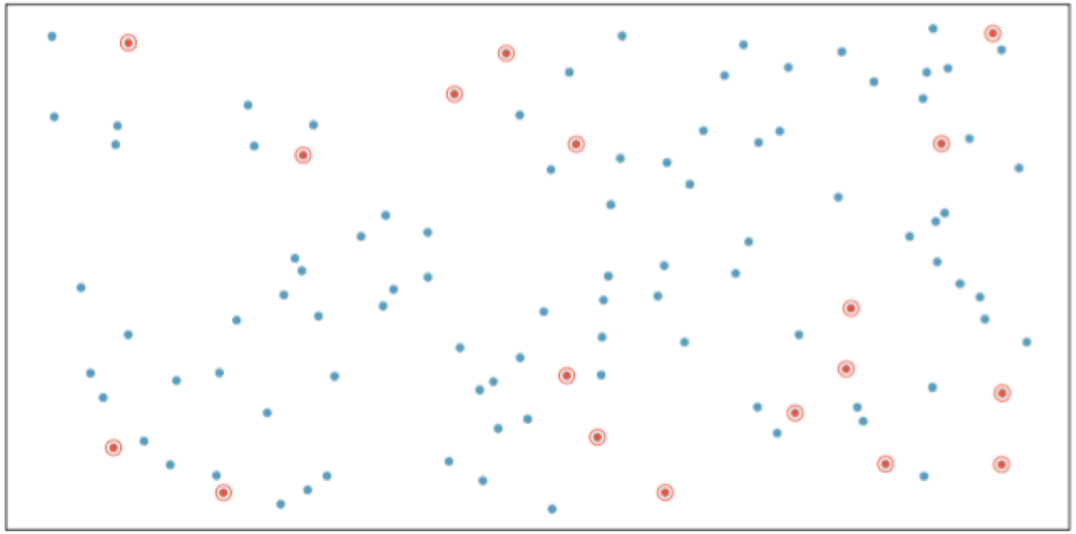
- When researchers want to investigate the possibility of a causal connection, they conduct an experiment. Usually there will be both an explanatory and a response variable.
- For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups.

Sampling methods

- 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.
- Four random sampling techniques:
  - Simple
  - Stratified
  - Cluster
  - Multistage sampling

Simple random sampling

Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's. 828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players.

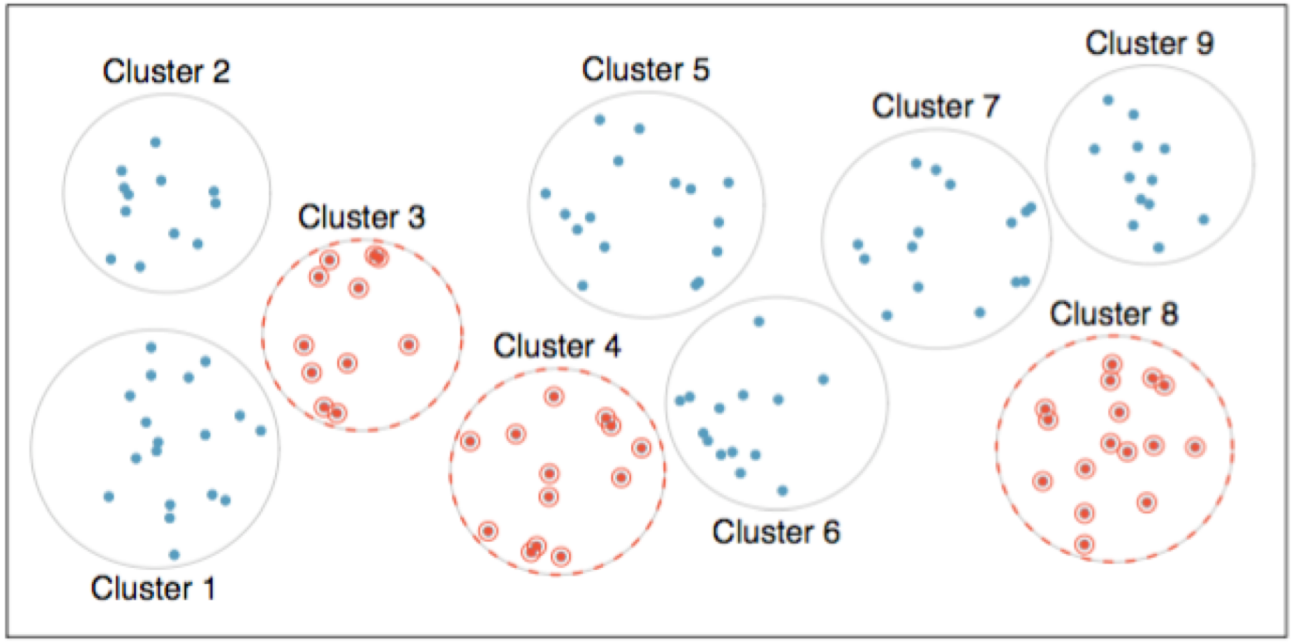


Stratified sampling

- Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum.
- Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest.
- In the baseball salary example, the teams could represent the strata, since some teams have a lot more money (up to 4 times as much!). Then we might randomly sample 4 players from each team for a total of 120 players.

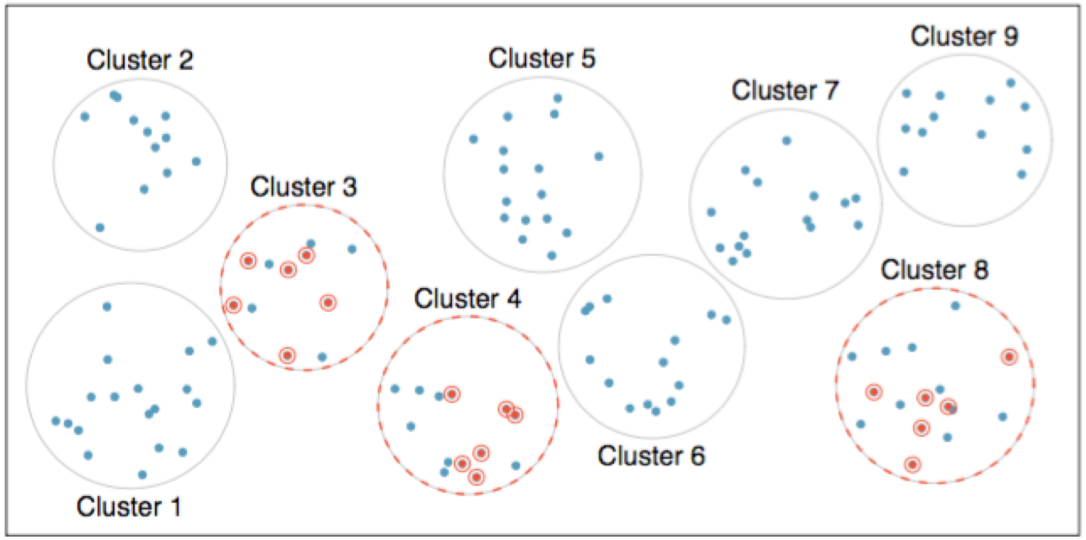
Cluster sample

In a cluster sample, we break up the population into many groups, called clusters. Then we sample **a fixed number of clusters** and include **all observations** from each of those clusters in the sample.



Multi stage sample

A multistage sample is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.



Cluster and multi stage sampling

- Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques.
- Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don’t look very different from one another.
- For example, if neighborhoods represented clusters, then cluster or multistage sampling work best when the neighborhoods are very diverse. A downside of these methods is that more advanced analysis techniques are typically required.

Example - sampling method

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and this approach would still give us reliable information.

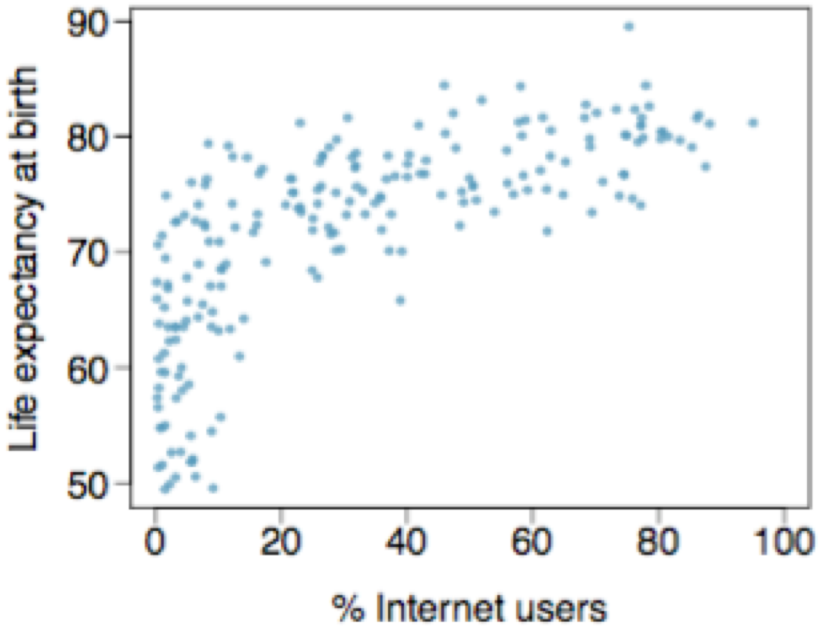
Exercise - causal relationship

The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available

A. Describe the relationship between life expectancy and percentage of internet users.

B. What type of study is this?

C. State a possible confounding variable



Dataset from:
https://www.cia.gov/library/publications/the-world-factbook/rankorder/rankorderguide.html

Point Estimate

- Estimate the population parameters like mean, standard deviation, median based on the sample
- Estimates generally vary from one sample to another- this sampling variation suggests our estimate may be close, but it will not be exactly equal to the parameter
- The sample mean tends to approach the true population average as more data become available

Standard error

- The standard deviation associated with an estimate is called the standard error. It describes the typical error or uncertainty associated with the estimate.
- Given n independent observations from a population with standard deviation , the standard error of the sample mean is equal to

$$SE_{\tilde{x}} = \sigma_{\tilde{x}} = \frac{\sigma_x}{\sqrt{n}}, n: sample\ size$$

- A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

Question: Point Estimate

Questions

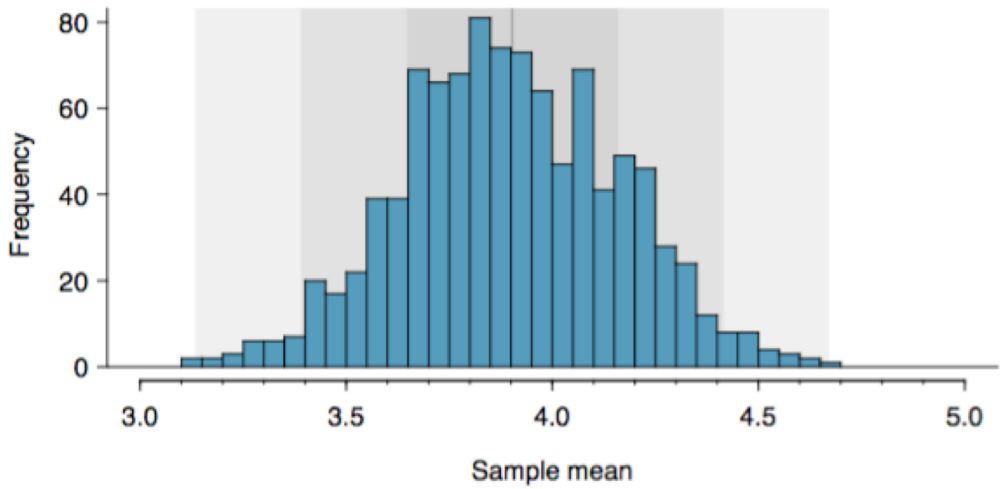
- (a) Would you rather use a small sample or a large sample when estimating a parameter? Why?
- (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard error than a point estimate based on a larger sample?

Answers

- *(a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard error.*

Sampling distribution

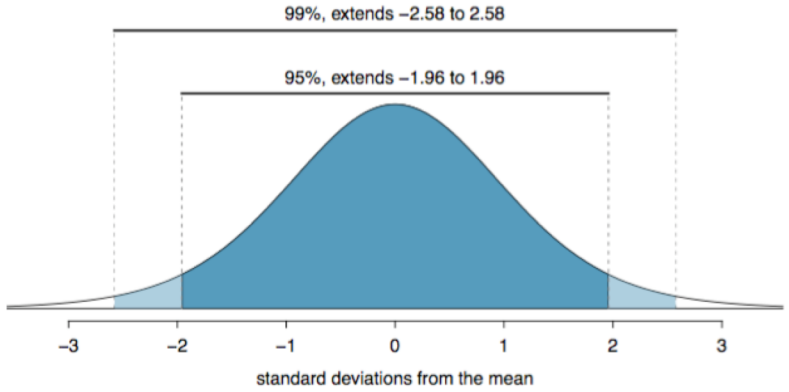
- The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population.
- It is useful to think of a particular point estimate as being drawn from such a distribution.



A histogram of 1000 sample means for number of days physically active per week, where the samples are of size n = 100.

Confidence Interval

- A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible range of values for the parameter - which is called confidence interval
- If the interval spreads out 2 standard errors from the point estimate, we can be roughly 95% confident that we have captured the true parameter: point estimate ± 2 * SE



Interpretation: We are XX% confident that the population parameter is between...

Central Limit Theorem

The sample mean is well approximated by a normal model.

Conditions:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a good rule of thumb.
- The population distribution is not strongly skewed. This condition can be difficult to evaluate, so just use your best judgement.

What are independent observations

If the observations are from a simple random sample and consist of fewer than 10% of the population, then they are independent.

Checking for strong skew

- Checking for strong skew usually means checking for obvious outliers
- When there are prominent outliers present, the sample should contain at least 100 observations, and in some cases, much more.

Hypothesis Testing

Are students lifting weights or performing other strength training exercises more or less often than they have in the past? [YRBSS dataset]

- *Null hypothesis H0:* The average days per week that YRBSS students lifted weights was the same for 2011 and 2013.
- *Alternate hypothesis HA:* The average days per week that YRBSS students lifted weights was different for 2013 than in 2011.

Null and alternative hypotheses

- The null hypothesis (H_0) often represents either a skeptical perspective or a claim to be tested. The alternative hypothesis (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.
- Do not reject the null hypothesis (H0), unless the evidence in favor of the alternative hypothesis (HA) is strong.

Example

- H_0: $\mu_{13} = 3.09$
- $H_{A:}$ $\mu_{13} \neq 3.09$

Example of null and alternate hypothesis

A US court considers two possible claims about a defendant: she is either innocent or guilty. If defendant pleas innocence, which would be the null hypothesis and which the alternative in a hypothesis framework?

H0: defendant is innocent

HA: defended is not innocent

The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

Example: Hypothesis Testing

In the sample of 100 students from the 2013 YRBSS survey, the average number of days per week that students lifted weights was 2.78 with a standard deviation of 2.56 (coincidentally nearly the same as days active). Compute a 95% confidence interval for the average for all students from the 2013 YRBSS survey. You can assume the conditions for the normal model are met.

Example: Hypothesis Testing

- The general formula for the confidence interval based on the normal distribution is $\tilde{x} \pm z * SE_{\tilde{x}}$
- $\tilde{x}$ = 2.78, z= 1.96 since we want to track 96% confidence interval.
- $SE_{\tilde{x}} = \frac{s_{13}}{\sqrt{n}} = \frac{2.56}{\sqrt{100}} = 0.256$
- Confidence interval 2.78 $\pm$ 1.96 * 0.256 = (2.27, 3.29)
- Because the average of all students from the 2011 YRBSS survey is 3.09, which falls within the range of plausible values from the confidence interval, we cannot say the null hypothesis is implausible. That is, we fail to reject the null hypothesis, H0.

Confidence level	Z-score
0.90	1.645
0.95	1.96
0.99	2.575

Exercise: hypothesis testing …

Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was $650 per month. What are the null and alternative hypotheses to test whether this claim is accurate?

Answer:

H_0: The average cost is $650 per month, μ = $650.
H_A: The average cost is different than $650 per month, μ = $650.

Exercise: hypothesis testing

The sample (175 observations) mean for student housing is $616.91 and the sample standard deviation is $128.65. Construct a 95% confidence interval for the population mean and evaluate the hypotheses.

You can compute 95% confidence interval, which turns out

$$616.91 \pm 1.96 * \frac{128.65}{\sqrt{175}} = (597.84, 635.98)$$

Because the null value $650 is not in the confidence interval, a true mean of $650 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that the actual average housing expense is less than $650 per month.

Decision errors in hypothesis testing

Hypothesis tests are not flawless, since we can make a wrong decision in statistical hypothesis tests based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. However, the difference is that in statistical hypothesis tests, we have the tools necessary to quantify how often we make such errors.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

p-value

- The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the p-value is a conditional probability.
- The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.
- The smaller the p-value, the stronger the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

Example: p-value

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than 7 hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate skeptical position for this research?

H_0: $\mu = 7$
H_A: $\mu > 7$ one-sided hypothesis test

Note: H_0 must be stated as an equality, where as H_A can be one sided or two sided inequality

One sided vs two sided test

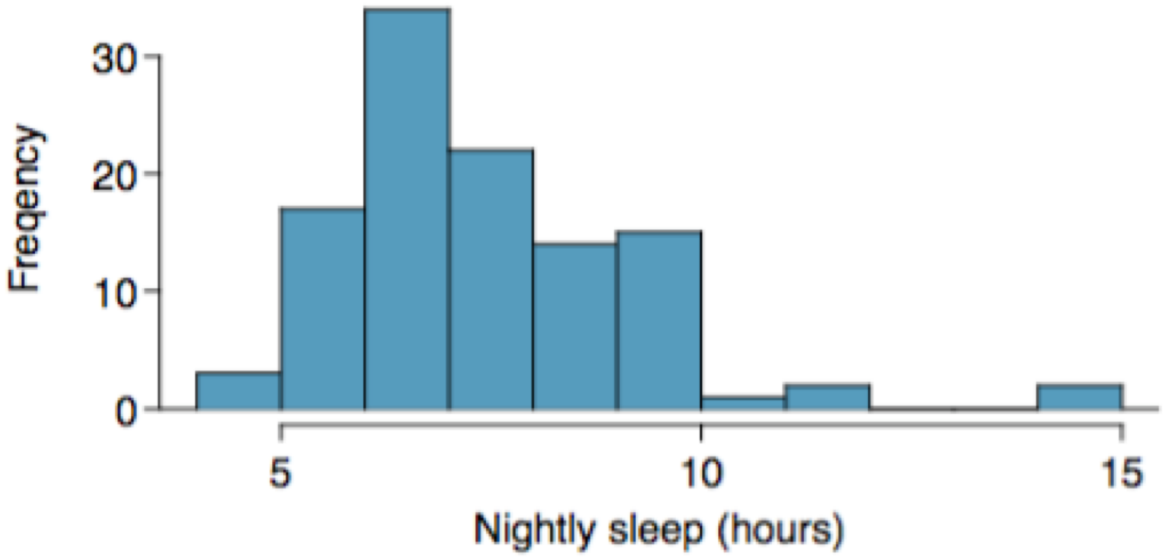
When you are interested in checking for an increase or a decrease, but not both, use a one-sided test. When you are interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

Normal model is applicable?

A. Because this is a simple random sample from less than 10% of the student body, the observations are independent.

B. The sample size in the sleep study is sufficiently large since it is greater than 30.

C. The histogram shows skew and the presence of a couple of outliers. This skew and the outliers are acceptable for a sample size of n = 110. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and we can reasonably calculate the standard error.



The researchers at the rural school conducted a simple random sample of n = 110 students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours.

Summarize hypothesis test with p-value

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_A.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

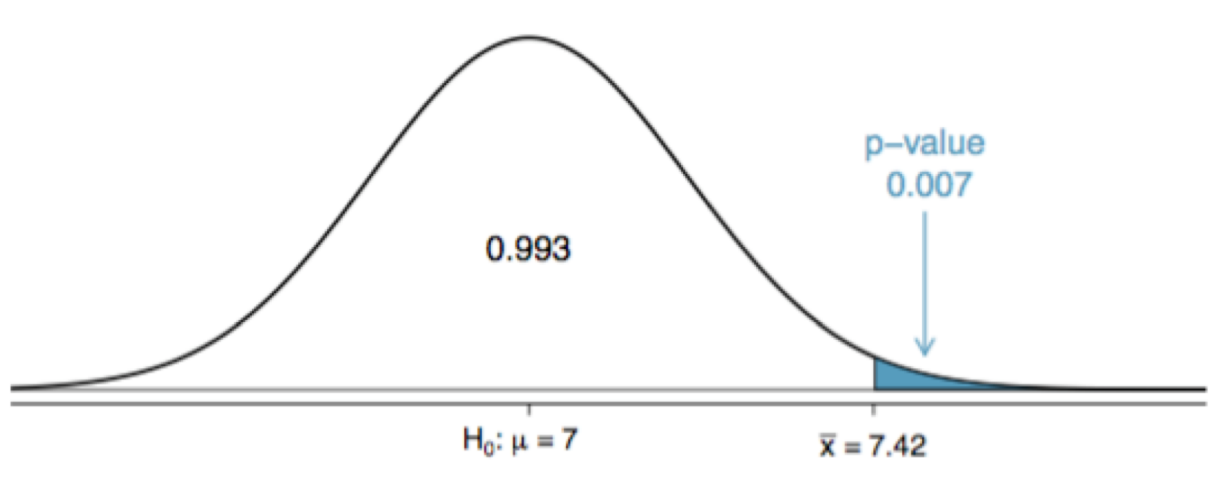
p-value calculation

$$SE_x = \frac{1.75}{\sqrt{110}} = 0.166856 \approx 0.17$$

$$Z = \frac{\tilde{x} - null\ value}{SE_{\tilde{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Area under normal curve (right)

= 1 - pnorm(2.47) = 0.007



We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value = 0.007 < 0.05 = $\boldsymbol{\alpha}$), we reject the H_0 in favor to the H_A.

Python:

from scipy.stats import norm

1- norm.cdf(2.47)

Interpretation of -p-value

- We reject the null hypothesis whenever p-value < α
- If the null hypothesis is true, the data only has a 5% (assuming $\alpha = 0.05$) chance of being in the 5% of data most favorable to H_A.

Exercise: hypothesis test

Ebay might be interested in showing that buyers on its site tend to pay less than they would for the corresponding new item on Amazon. We'll research this topic for one particular product: a video game called Mario Kart for the Nintendo Wii. During early October 2009, Amazon sold this game for $46.99. Set up an appropriate (one-sided!) hypothesis test to check the claim that Ebay buyers pay less during auctions at this same time.

H_0: $\mu_{ebay} = 46.99$

H_A: $\mu_{ebay} < 46.99$

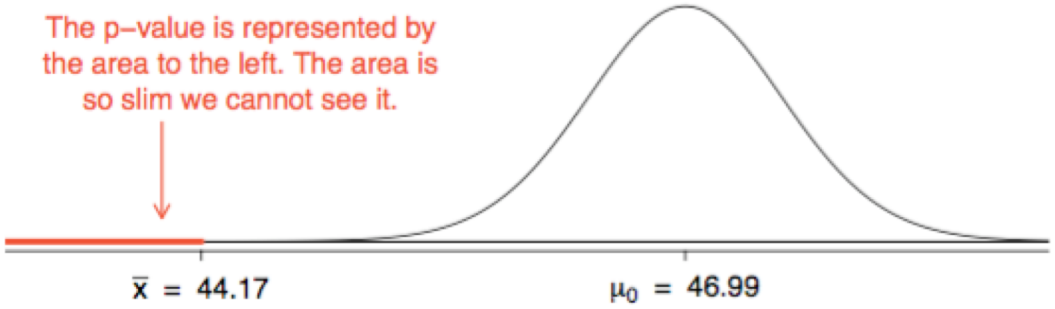
Exercise: hypothesis test using p-value

The average sale price of the 52 Ebay auctions for Wii Mario Kart was $44.17 with a standard deviation of $4.15. Does this provide sufficient evidence to reject the null hypothesis? Use a significance level of $\boldsymbol{\alpha}$ = 0.01.

$$SE = \frac{s}{\sqrt{n}} = \frac{4.15}{\sqrt{52}} = 0.5755$$

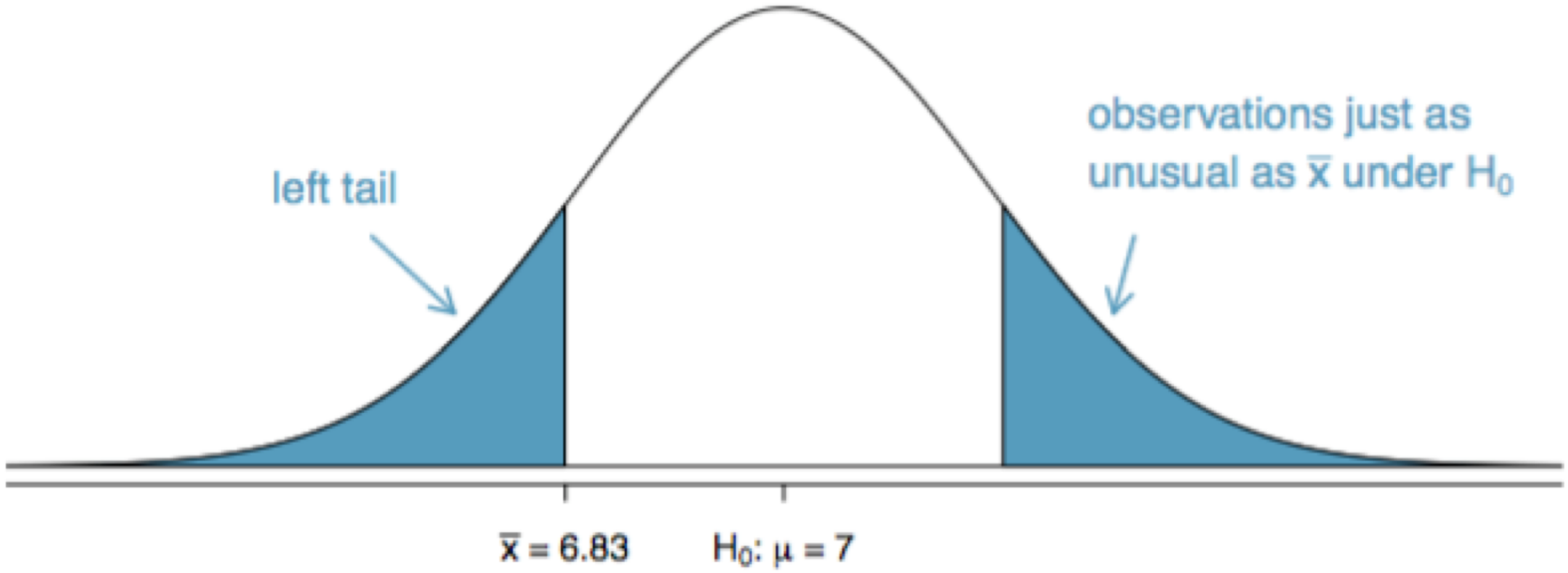
$$Z = \frac{\tilde{x} \ - null\ value}{SE} = \frac{44.17 \ - 46.99}{0.5755} = -4.9$$

$$pnorm(-4.9) = 4.791833e - 07$$



Because the p-value is so small – specifically, smaller than $\boldsymbol{\alpha}$ = 0.05 – this provides sufficiently strong evidence to reject the null hypothesis in favor of the alternative.

Two sided p-value



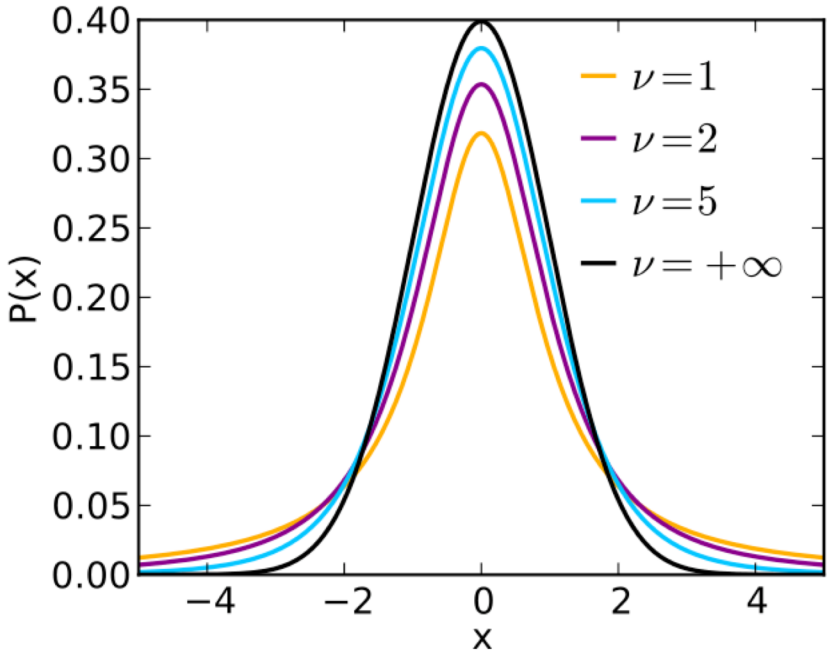
H_A is two-sided, so both tails must be counted for the p-value.

p-value = left tail + right tail = 2 * (left tail)

t-distribution

- When sample size is small, to calculate the standard error, it will be useful to rely on the t-distribution rather than normal distribution
- The t-distribution, always centered at zero, has a single parameter: degrees of freedom.
- The degrees of freedom (v) describe the precise form of the bell-shaped t-distribution.

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$



Degree of freedom

- The degrees of freedom describe the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.
- When the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution.

Confidence interval using t-distribution

```
a = 5
s = 2
n = 20
error = qt(0.975, df = n - 1) * s / sqrt(n)
left = a - error  #4.063971
right = a + error #5.936029
```

(R code)