MSBA Boot Camp

Statistics I

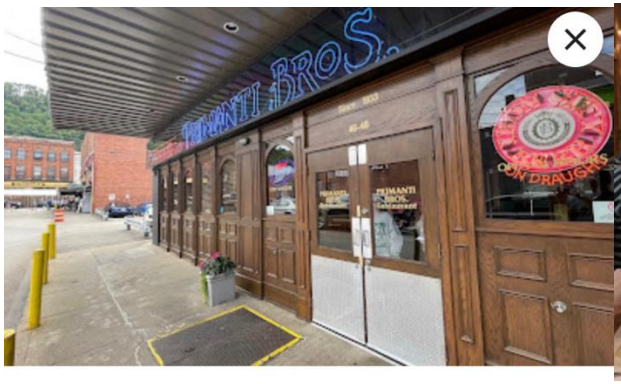
Zhanrui Cai

Assistant Professor in Analytics and Innovation
Faculty of Business and Economics
University of Hong Kong

Business Analytics

- Turn data into information/value.
  - Business managers need to make decisions.
  - They need to make the most informed decisions that they can, and generate value.
- Decision making under uncertainty.
  - Most of the decisions are based on guesses, rather than “facts.”
  - How to make the “best” guess possible as well as how to measure the accuracy of their guesses.

Find the best restaurant?



Primanti Bros. Restaurant and Bar Strip District

4.5 ★★★★½ (6,159) · $

Sports bar

Overview | Reviews | About

Directions | Save | Nearby | Send to phone | Share

Long-running Pittsburgh-born chain known for its sandwiches piled high with coleslaw & french fries.

✓ Dine-in · ✓ Kerbside pickup

✓ No-contact delivery

46 18th St, Pittsburgh, PA 15222, United States

Open · Closes 2 am

Confirmed by this business 3 weeks ago

See more hours

Dorido's Restaurant

4.6 ★★★★½ (1,376) · $$

American restaurant

Overview | Reviews | About

Directions | Save | Nearby | Send to phone | Share

Long-running haunt offering casual seafood dishes, sandwiches, martinis & lots of beers on tap.

✓ Dine-in · ✓ Takeaway · ✗ Delivery

6408 Brownsville Rd, Pittsburgh, PA 15236, United States

Open · Closes 2 am

See more hours

EVIA Greek Restaurant

4.7 ★★★★½ (167)

Greek restaurant

Overview | Reviews | About

Directions | Save | Nearby | Send to phone | Share

✓ Dine-in · ✓ Kerbside pickup

✓ No-contact delivery

564 Lincoln Ave, Bellevue, PA 15202, United States

Located in: Untethered Therapy

Closed · Opens 11 am Sat

Place an order

Data and Statistics

"Data don't make any sense,
we will have to resort to statistics."

Statistics: Discovery through Data

- Statistics: the science of collecting, organizing, and interpreting *data*.

- Data: a collection of numbers, labels, or symbols, and the context of those values.
  - Often, a subset of a larger group (a sample from a population)
    - Performance of Class 2020 MSBA students
  - Often, a sequence of measurements on a process (a time series)
    - The closing price of a stock every day
    - The exchange rate between RMB and USD every minute

- Statistic: any numerical summary of data (average, …)

Data and Variable

- Variable: a characteristic that may assume more than one set of values.
  - Purchase, Amount ($), Host, Region-USA
  - The value of the variable can "vary" from one entity to another
- Data: a collection of numbers, labels, or symbols with context
- Data table: a rectangular arrangement of data with rows and columns
  - Observation/cases/subject: row
  - Variable/attribute: column

Visitor ID	Purchase	Amount ($)	Host	Region-USA
A	No	0	yahoo.com	Northeast
B	No	0	google.com	West
C	Yes	10.00	cnn.com	South
D	Yes	125.13	twitter.com	South
…	…	…	…	…

Type of Variables

- Numerical variable (a.k.a. Quantitative)
  - Continuous or discrete
  - Describe numerical properties of cases
  - Can be expressed with numbers (quantity) and have measurement units
  - Example: “Amount” in Amazon Example.

- Categorical variable （a.k.a. Qualitative）
  - Nominal or ordinal
  - Identify group membership
  - More than 1 possible outcome
  - Example: “Purchase”, “Host”, “Region” in Amazon Example.

Characterization of Variation in Data

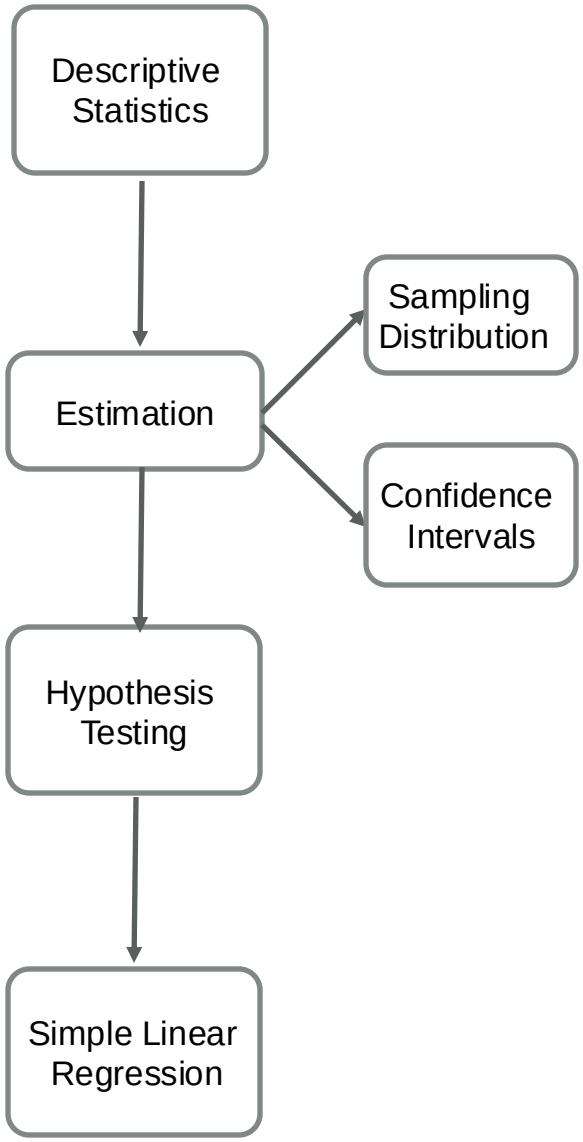
- Variation
  - A common feature **-** virtually all data exhibit variation.
  - A principal goal of statistics - describe and understand the implications of variation.

- Discovery with Statistical Tools

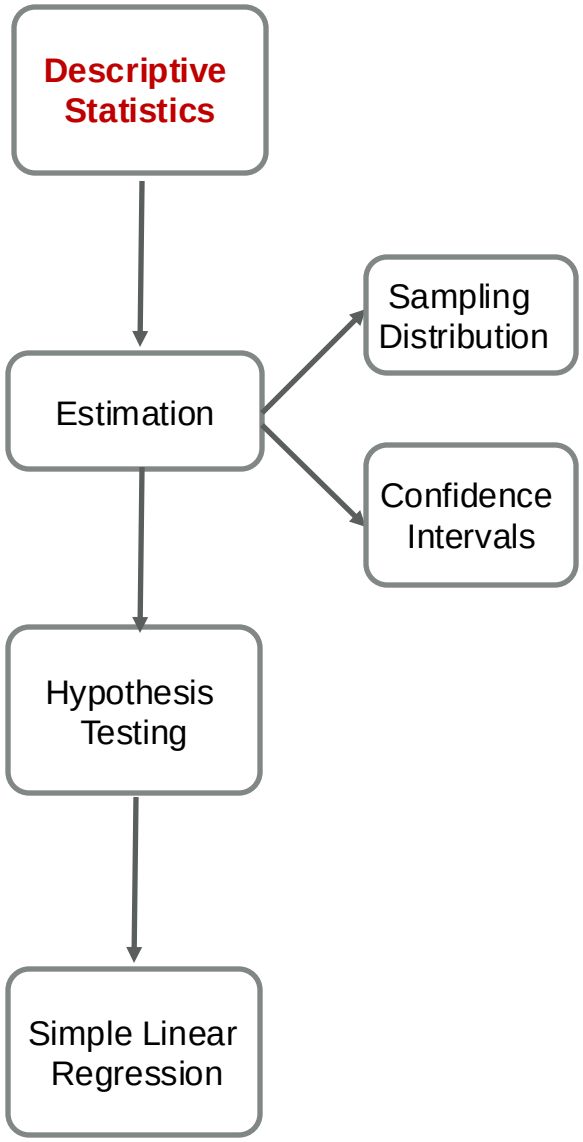
  - Finding a revealing view of the data: key to effective statistical analysis.
    - Different types of displays: graphical and numerical.
    - Goal is to focus attention on essential features of data.
    - Separate reproducible patterns from random, coincidental features.

  - Relevance for decision making
    - Summarization can be very useful when using data to form decisions.
    - Avoid distraction from extraneous features of data.

Overview


Descriptive Statistics
Estimation
Sampling Distribution
Confidence Intervals
Hypothesis Testing
Simple Linear Regression

Overview


Descriptive Statistics
Estimation
Sampling Distribution
Confidence Intervals
Hypothesis Testing
Simple Linear Regression

Descriptive Statistics- Graphical Methods

Graphical Methods

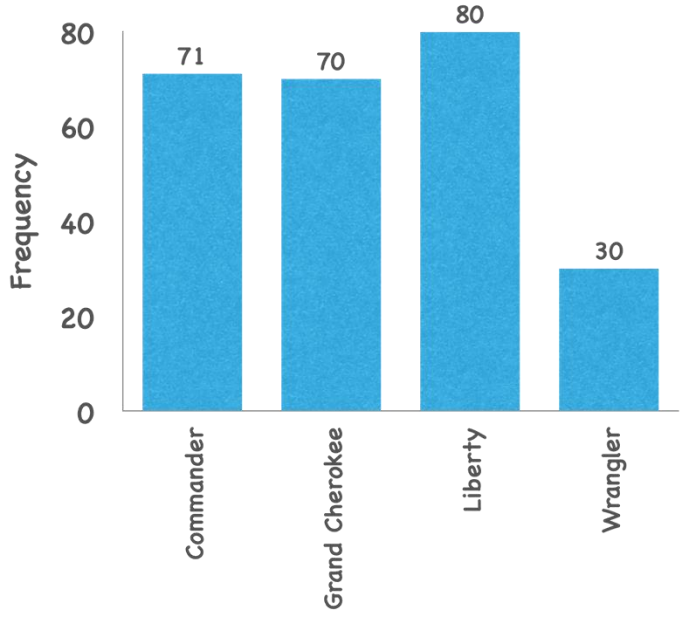
- Categorical: Bar chart; Pie chart
- Numerical: Histogram; Density Curve; Scatter Plot

Categorical Data - Graphing Methods

Bar Chart

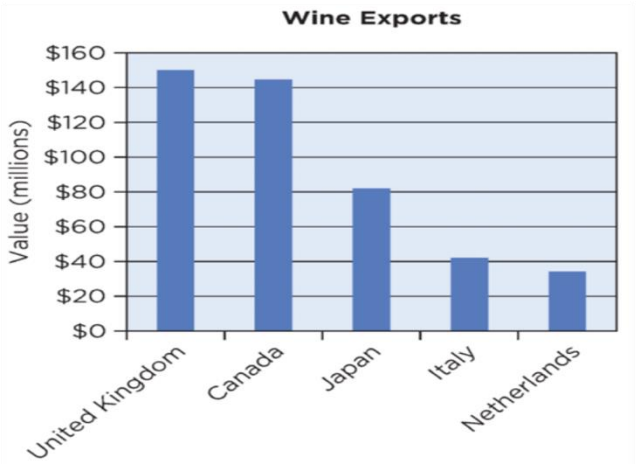
- Bar Chart: a vertical or horizontal rectangle represents the frequency for each category
- Height can be frequency, or percent frequency
- General bar chart: x-axis as categories, y-axis represents values for each category.

Jeep Model	Frequency
Commander	71
Grand Cherokee	70
Liberty	80
Wrangler	30



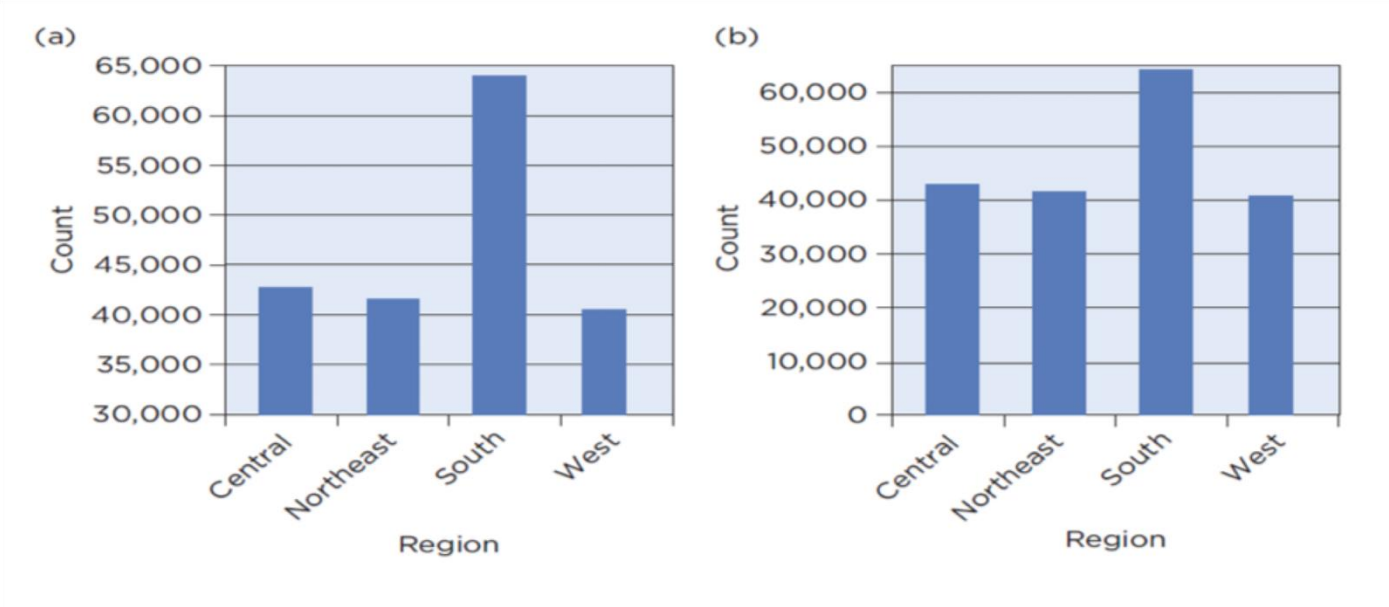
Area Principal

- The area of a plot that shows data should be proportional to the amount of data
- Violation Case I: decoration



Area Principal

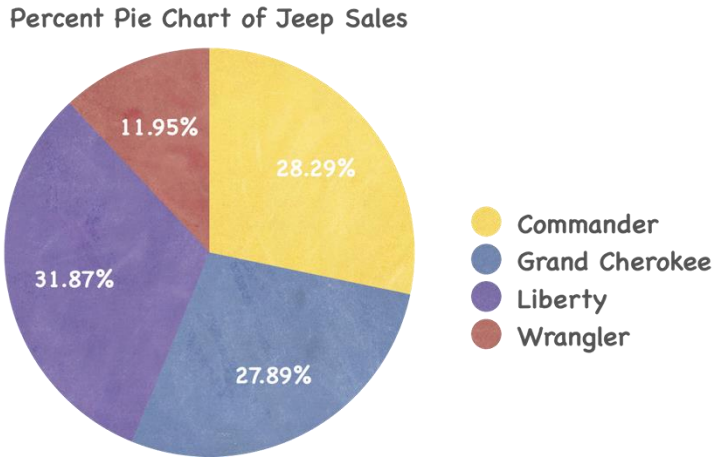
- Violation Case II: baseline of a bar chart is not zero.
- Example: a data from the US - the regional distribution of the 188,996 Amazon visitors.



Pie Chart

- Pie Chart: a circle divided into slices where the size of each slice represents its relative frequency or percent frequency.
- If you want to compare actual counts, will you use pie chart or bar chart?

Jeep Model	Relative Frequency
Commander	28.29%
Grand Cherokee	27.89%
Liberty	31.87%
Wrangler	11.95%



Numerical Data - Graphing Methods

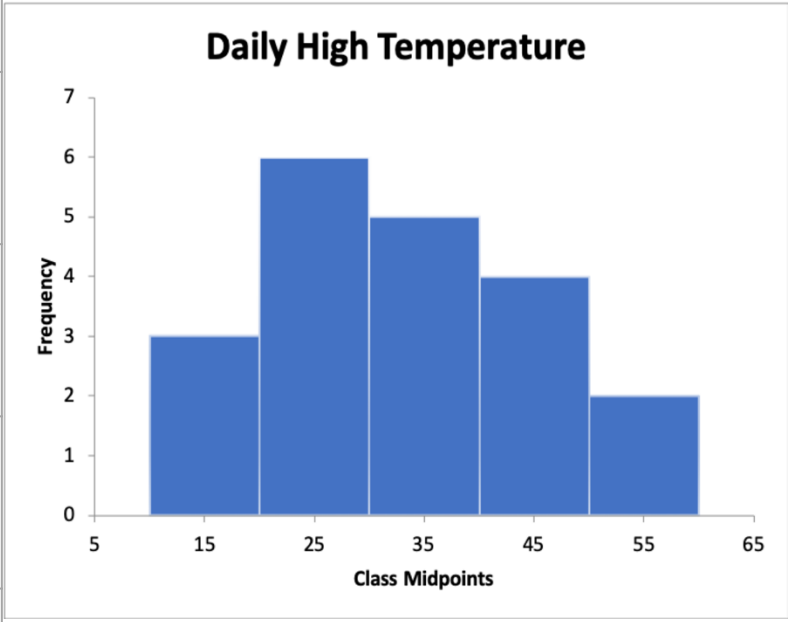
Histogram: Temperature Example

- A weather report center randomly selects 20 winter days and records the daily high temperature:

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Histogram: Temperature Example

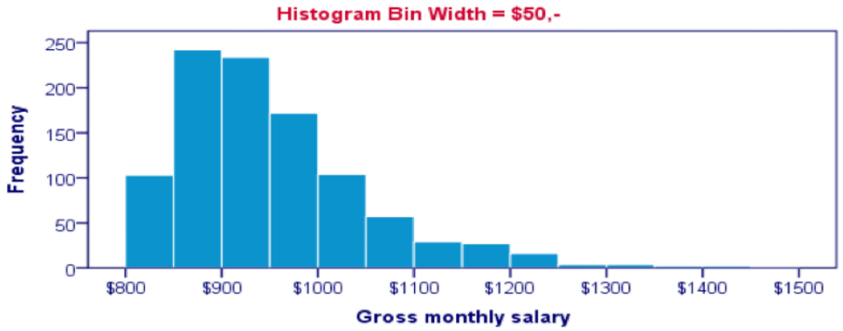
Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

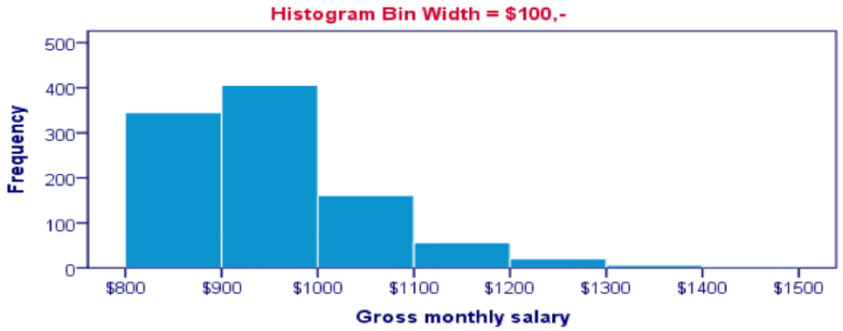


Draw a Histogram

- Create a frequency table: decide the number of bins, bin width and bin limits;
- In each bin, compute the number of observations;
- Over each bin, plot a rectangle with height equal to:
    - Frequency of observations in the bin
    - Relative frequency of observations in the bin
    - Percentage frequency

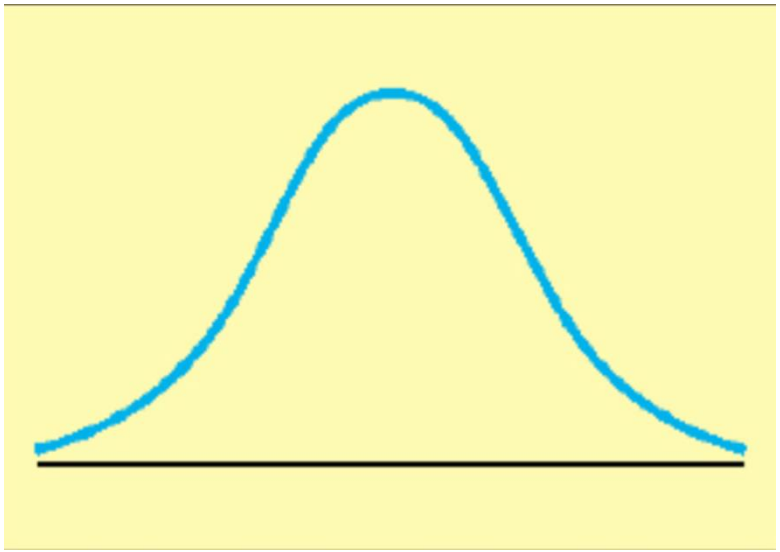
Varying Bin Width





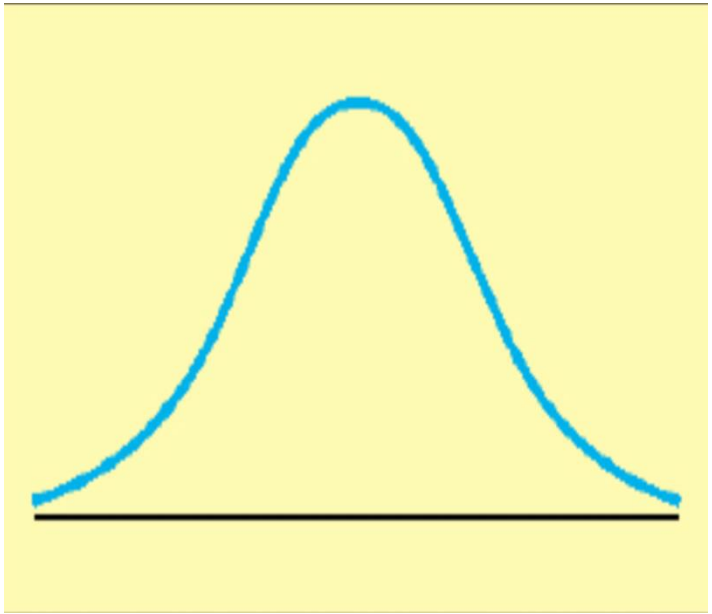
Histogram to Curve

- When the amount of data go to infinity, and as the bin sizes shrink to very small, the histogram can look like a smooth curve
- One special curve that is observed most often is what we called the Normal curve



The Normal Curve

- Bell shape
- Symmetric
- The height of the curve over any point represents the relative proportion of values near that point



What are Curves?

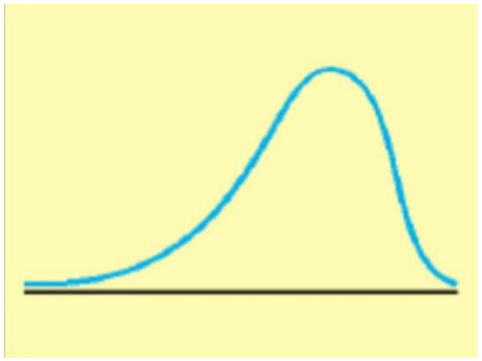
- Sometimes, we call those curves *distributions*.
- *Distribution* of a variable: what possible values the variable takes and how frequently it takes those values.
- There are many methods to describe and display distributions. Curves are one of them to describe numerical variables.

Words that describe distributions

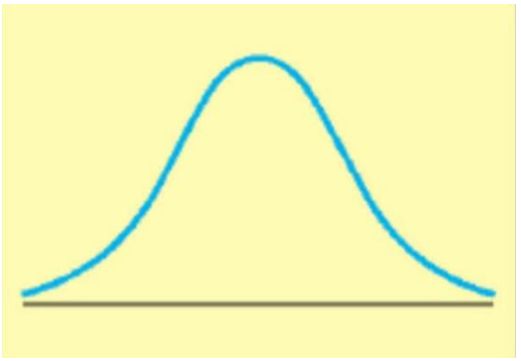
- (looking at a histogram or curve)
- **Unimodal**: has one major peak
- **Bimodal**: has two major peaks
- **Symmetric**: there is a symmetry with respect to the middle point
- **Skewed to the right**: when the right tail (larger values) is much longer than the left tail (smaller values)
- **Skewed to the left: …**

Skewness

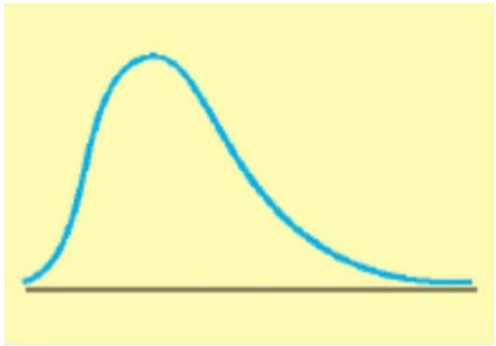
Left Skewed



Symmetric

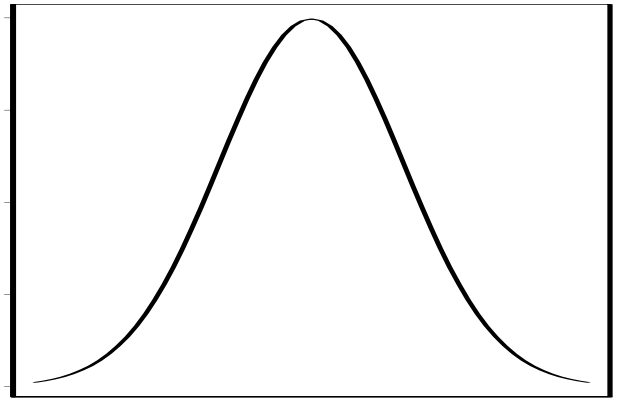


Right Skewed

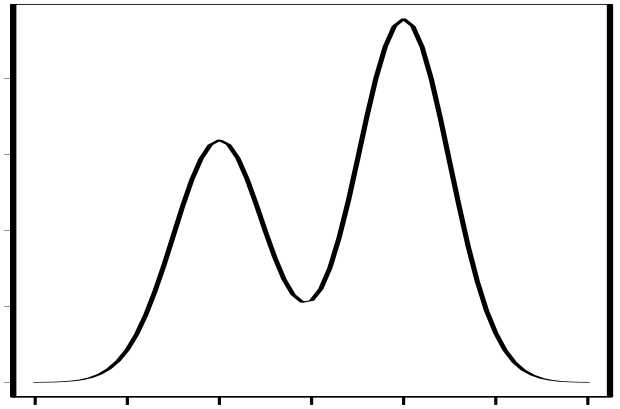


Modal

Unimodal



Bimodal



Difference Between Histogram and Bar Chart

- Histogram: numerical data, construct bins and compute frequency; no gap between bars
- Bar chart: categorical data, compute frequency directly; gap between bars
- You will lose information if you use Bar chart to represent numerical data due to truncation

Descriptive Statistics- Numerical Methods

The Mean

E.g. if data is 6, 9, 8, 3, 3, 1

$$\text{Mean} = \frac{6+9+8+3+3+1}{6} = 5$$

For a variable *x* with *n* observed values $x_1, x_2, \ldots, x_n$ the mean of *x* is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

notation: “x-bar”

The Mean

- The most common measure of central tendency:
  - n is the sample size
  - x_i are the observed values
  - Easily affected by extreme values

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The Median = 50th percentile

Arrange data in order.

Median Md = 50th percentile = “middle observation”

[if number of observations is even, average the middle two.]

E.g. for data 1, 3, 3, 6, 8 $M_d = 3$

E.g. for data 1, 3, 3, 6, 8, 9 $M_d = (3 + 6) / 2 = 4.5$

“Robust” or “resistant”

- Robust = insensitive to a few extreme observations

  (imagine a typo of adding several zeros to a number)

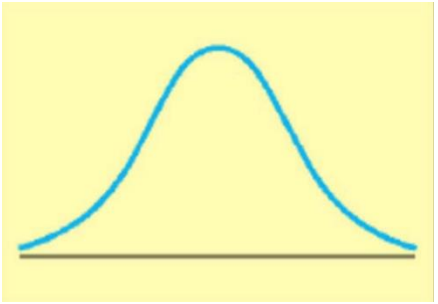
- Which is more robust: mean or median ?

Compare	1, 3, 3, 6, 8
to	1, 3, 3, 6, 8000000

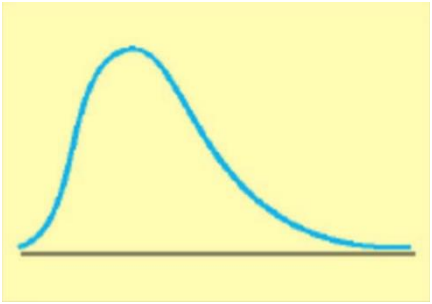
Q&A:

- What can you tell if you learn the average yearly income of a certain college class is $1 million?
- What do you suspect the histogram of the income of this college class look like?
- What do you think “is more impressive”, the median or the mean of the class is $1 million?

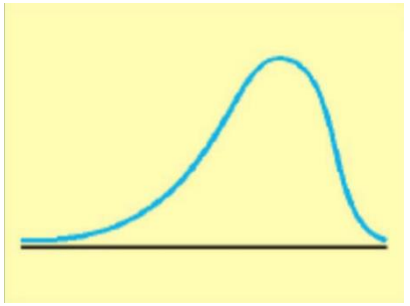
Relationships Between Mean and Median



Symmetrical
Mean = Median



Right-skewed
Mean > Median



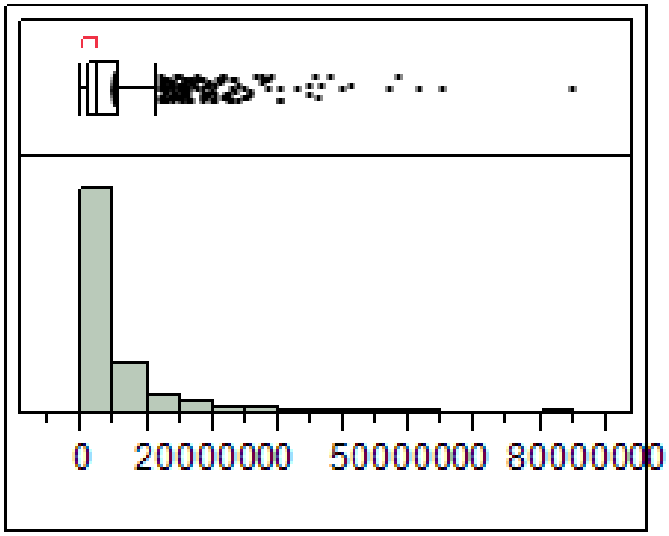
Left-skewed
Mean < Median

Executive Compensation Data

- Not all data is symmetric. Some data is not even close.
- Data: the annual total compensation of 1,495 executives in 2003.
- The summary of TotalComp (total compensation) is

Distributions

TotalComp



Quantiles

100.0%	maximum	7.48e+7
99.5%		3.87e+7
97.5%		2.23e+7
90.0%		1.06e+7
75.0%	quartile	5477262
50.0%	median	2532583
25.0%	quartile	1254480
10.0%		697351
2.5%		340924
0.5%		55670.8
0.0%	minimum	0

Summary Statistics

Mean	4628354.7
Std Dev	6231619.2
Std Err Mean	161168.55
Upper 95% Mean	4944495.4
Lower 95% Mean	4312214.1
N	1495

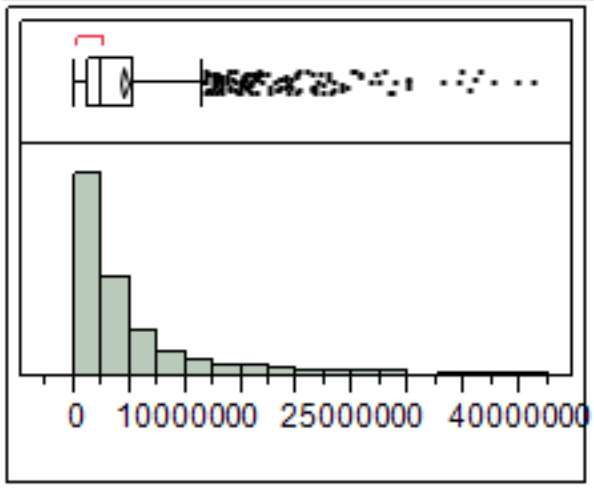
- Someone made nearly $75,000,000! Who? (label the data)

Outliers

- Extreme values, often called *outliers*, dominate the output and make it difficult to see the rest of the distribution.
- Let's exclude the five largest values and see what happens.

Distributions

TotalComp



Quantiles

100.0%	maximum	4.11e+7
99.5%		3.45e+7
97.5%		2.15e+7
90.0%		1.03e+7
75.0%	quartile	5403590
50.0%	median	2525643
25.0%	quartile	1251064
10.0%		695533
2.5%		340516
0.5%		55440.6
0.0%	minimum	0

Summary Statistics

Mean	4458977.3
Std Dev	5479958.2
Std Err Mean	141965.92
Upper 95% Mean	4737451.7
Lower 95% Mean	4180502.8
N	1490

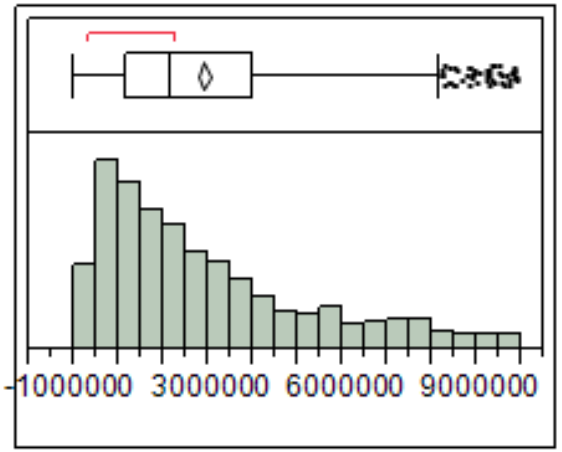
- That doesn't help much. The compensations of a relatively small number of executives continue to dominate the plot.

Effect of Outliers on Summary Statistics

- The effects of dropping the 5 highest salaries
- What about removing salaries>$10 million?
- What's going on?

Distributions

TotalComp



Quantiles		
100.0%	maximum	9887704
99.5%		9821930
97.5%		8837706
90.0%		6730893
75.0%	quartile	3978805
50.0%	median	2174400
25.0%	quartile	1154878
10.0%		654995
2.5%		331691
0.5%		50655
0.0%	minimum	0

Summary Statistics	
Mean	2907107.1
Std Dev	2299002.4
Std Err Mean	63015.938
Upper 95% Mean	3030728.6
Lower 95% Mean	2783485.6
N	1331

Transforming Data

- An alternative way of looking at the big picture is to transform the data to a different scale. For these data, let's consider the log:

$$y = \log_{10} x$$

- If we use base 10, then the logs of the compensations essentially count the number of zeros (for example, $\log_{10} 100 = 2$).

- Recall $\log(y/x) = \log(y) - \log(x)$, so that multiplicative increases in original scale correspond to additive increases in the log scale.

- Thus. percentage changes in the original scale become additive changes in the log scale.

- Question: Which is larger
  - $\log\$11{,}000 - \log\$10{,}000$ vs. $\log\$11{,}000{,}000 - \log\$10{,}000{,}000$?

CEO Compensation

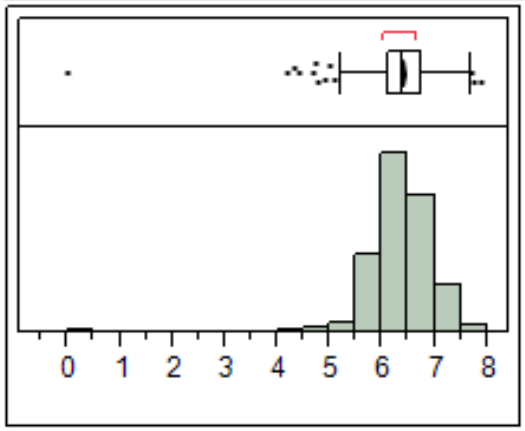
- For the CEO_Comp_03 data, let's create the new variable

$$\text{Log10TotalComp} = \log_{10}(TotalComp)$$

- A summary of the logged compensation values is revealing:

Distributions

Log10 TotalComp



Quantiles		
100.0%	maximum	7.87361
99.5%		7.58779
97.5%		7.34867
90.0%		7.02522
75.0%	quartile	6.73888
50.0%	median	6.40416
25.0%	quartile	6.09964
10.0%		5.84712
2.5%		5.53621
0.5%		4.84393
0.0%	minimum	0

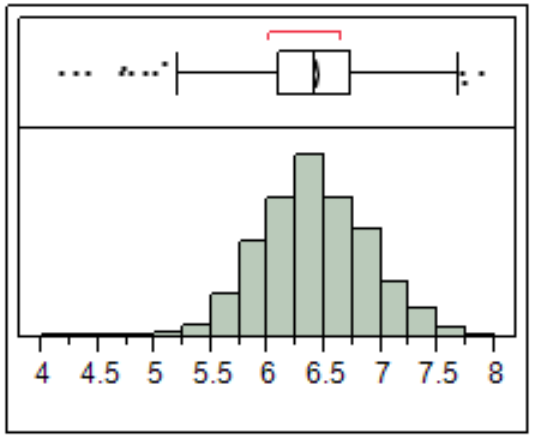
Summary Statistics	
Mean	6.4106808
Std Dev	0.5006985
Std Err Mean	0.0129539
Upper 95% Mean	6.4360906
Lower 95% Mean	6.385271
N	1494

- Wow! The transformation has revealed an unusually small extreme value. Who is it?

Exclude the $1 CEO …

Distributions

Log10 TotalComp



Quantiles		
100.0%	maximum	7.87361
99.5%		7.58791
97.5%		7.34874
90.0%		7.02541
75.0%	quartile	6.73919
50.0%	median	6.40475
25.0%	quartile	6.10016
10.0%		5.84999
2.5%		5.53861
0.5%		4.95387
0.0%	minimum	4.15863

Summary Statistics	
Mean	6.4149746
Std Dev	0.4725503
Std Err Mean	0.0122298
Upper 95% Mean	6.438964
Lower 95% Mean	6.3909852
N	1493

- By transforming, can see the variation that distinguishes the compensation of the majority of the executives rather than just singling out those that make much more than the rest.
- The log trans. has done this by pulling in large values and stretched out the small values. It has transformed right skewed data into more normal data.
- Is any information lost by transforming the data?
  - Is the average of the logs equal to the log of the average?
  - Given one, can you find the other?

Three Wins for The Log Transformation

Empirically

- The log transformation pulls in outliers in right skewed distributions resulting in data more amenable to analysis.

Practically

- Differences on a log scale are naturally interpreted as percent changes.

Theoretically

- Log transformations have the potential to pick up interesting relationships such as *diminishing returns to scale*.

Quartiles

- Define **first quartile** to be the median of the observations below the median
- Define **third quartile** to be the median of the observations above the median

$$1, 3, 3, \mid 6, 8, 9$$

M=4.5

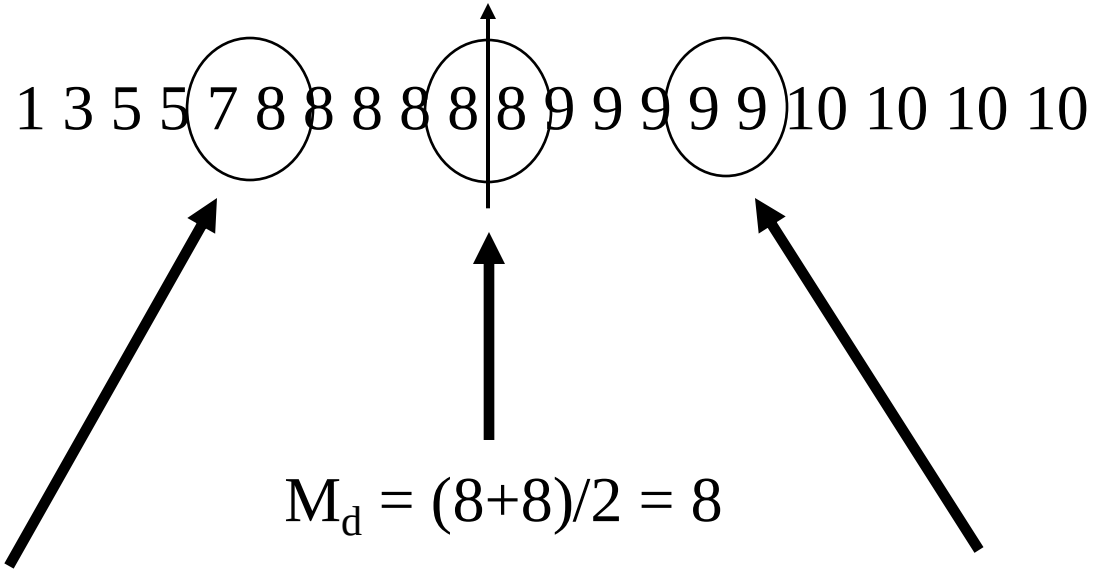
Q1=3 Q3=8

- The **interquartile range**, **IQR**, is Q3 - Q1

Example: customer satisfaction ratings

20 measurements on the 10 point scale:

9, 8, 3, 8, 10, 9, 8, 9, 5, 8, 1, 10, 8, 10, 7, 8, 9, 10, 5, 9



$M_d = (8+8)/2 = 8$

$Q_1 = (7+8)/2 = 7.5$

$Q_3 = (9+9)/2 = 9$

$IQR = Q_3 - Q_1 = 9 - 7.5 = 1.5$

More measures of spread

Range:

Largest minus smallest measurement: *max* - *min*

Variance:

The "average" of the squared deviations of all the measurements from the mean (*details to follow*)

Standard Deviation:

The square root of the variance

Variance and Standard Deviation (SD)

- SD is the most common measure of spread (variability)
- Relationship: $\mathrm{SD} = \sqrt{Variance}$

Notation:

- $Variance = s^2$, $\mathrm{SD} = s = \sqrt{s^2}$

Idea of variance and SD

- How far away are the observations, on average, from the mean?

- Based on the **deviations**

e.g. $x_i - \bar{x}$ = deviation from the mean for the *i*-th observation

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

"average" squared deviation

Five Number Summary

The smallest observation, the first quartile, the median, the third quartile, and the largest observation.

Min Q1 M_d Q3 Max

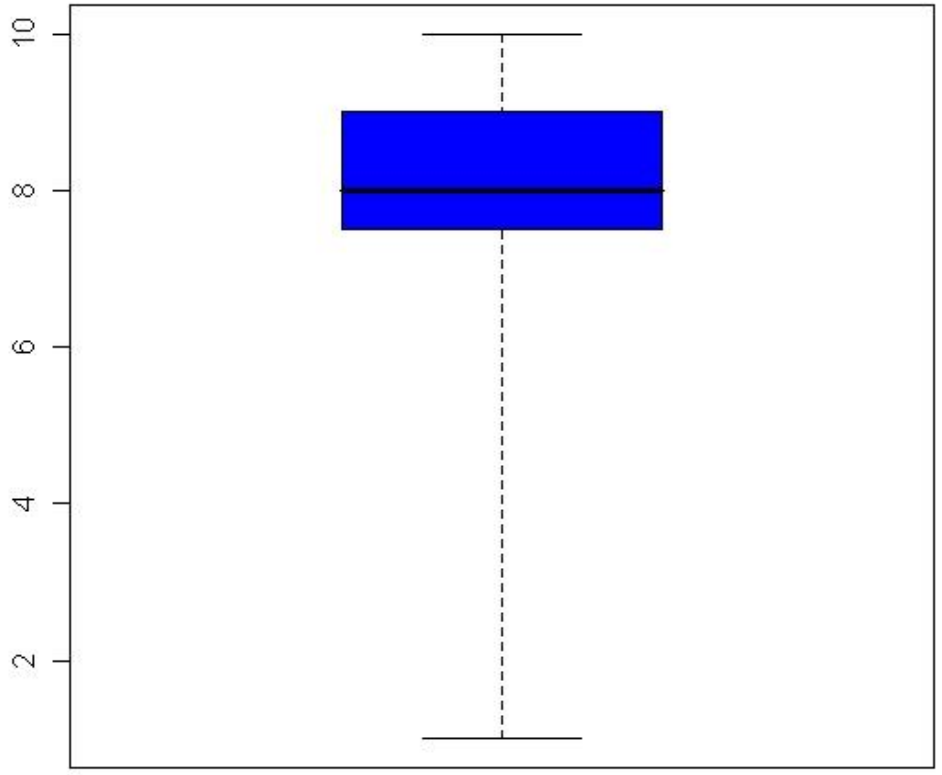
Customer satisfaction data

Minimum	Q1	Median	Q3	Maximum
1	7.5	8	9	10

Can plot this summary using a **boxplot**

Boxplot (simplest form)

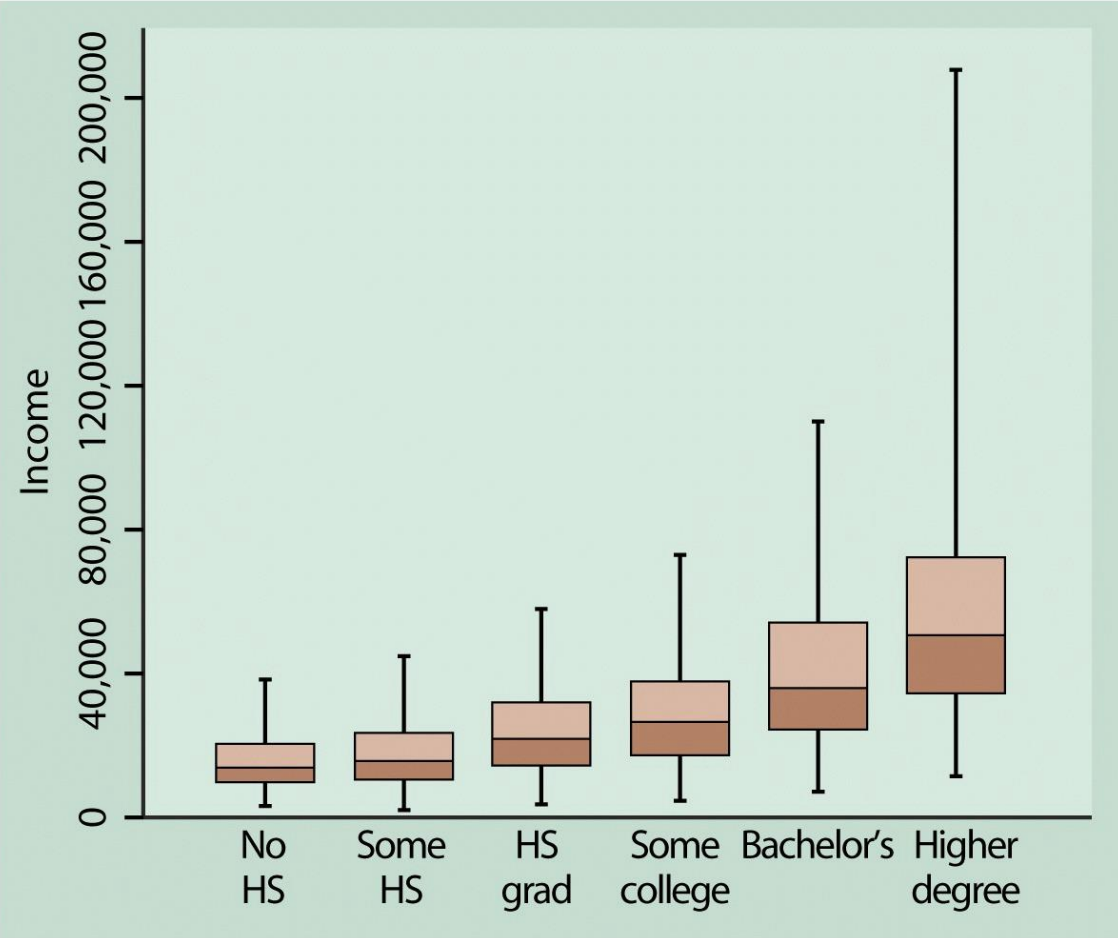
- The central box spans the quartiles Q1 and Q3.
- The line in the box marks the median M.
- The whiskers extend out to the smallest and largest observations



Customer satisfaction data

More on Boxplots

Box plots are especially good for showing differences between distributions across groups.



Choosing a Summary

- For a skewed distribution or a distribution with strong outliers the five number summary is usually better than mean and SD
- Use SD for the spread when you use the mean for the center

WARNING: Do not use only boxplots and numerical summaries to describe the shape of a distribution. Add a histogram.

Normality of Data

- Using the refined histogram interpretation, the relative area under the curve over an interval can be associated with the
  - Relative frequency of values in the interval, and the
  - Probability that a randomly drawn observation falls within the interval.

- Normality can be used to describe how data concentrate near the mean.

$$(\bar{x} - \mathrm{s}, \bar{x} + \mathrm{s}): about\ 68\%\ of\ the\ data$$

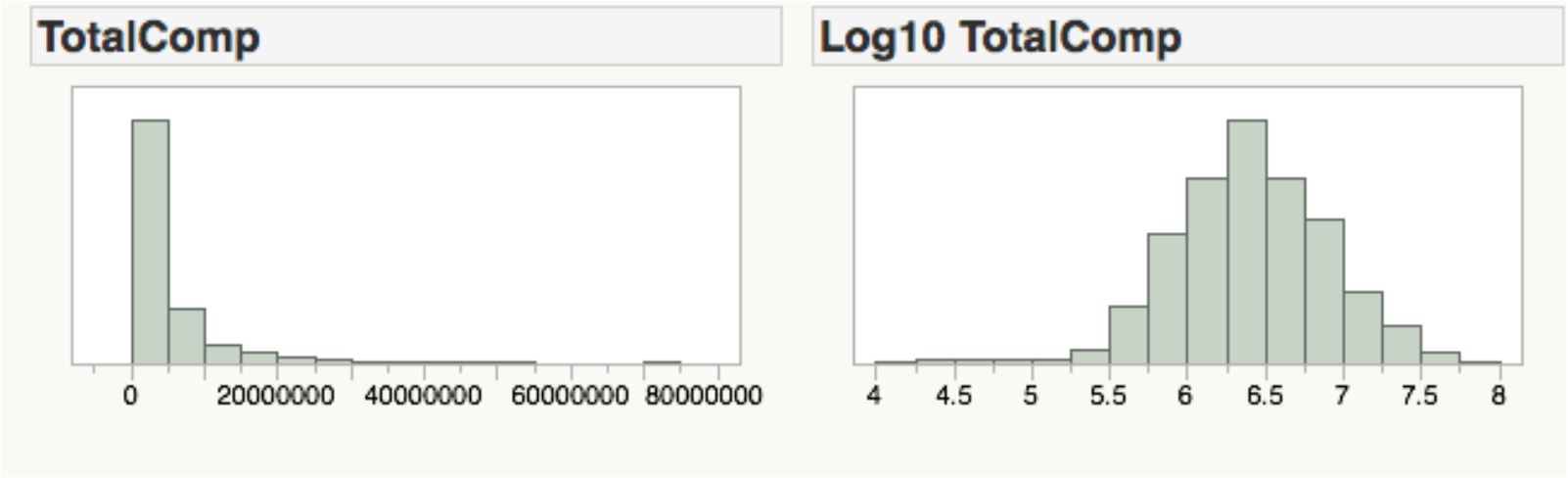
$$(\bar{x} - 2\mathrm{s}, \bar{x} + 2\mathrm{s}): about\ 95\%\ of\ the\ data$$

$$(\bar{x} - 3\mathrm{s}, \bar{x} + 3\mathrm{s}): about\ 99.7\%\ of\ the\ data$$

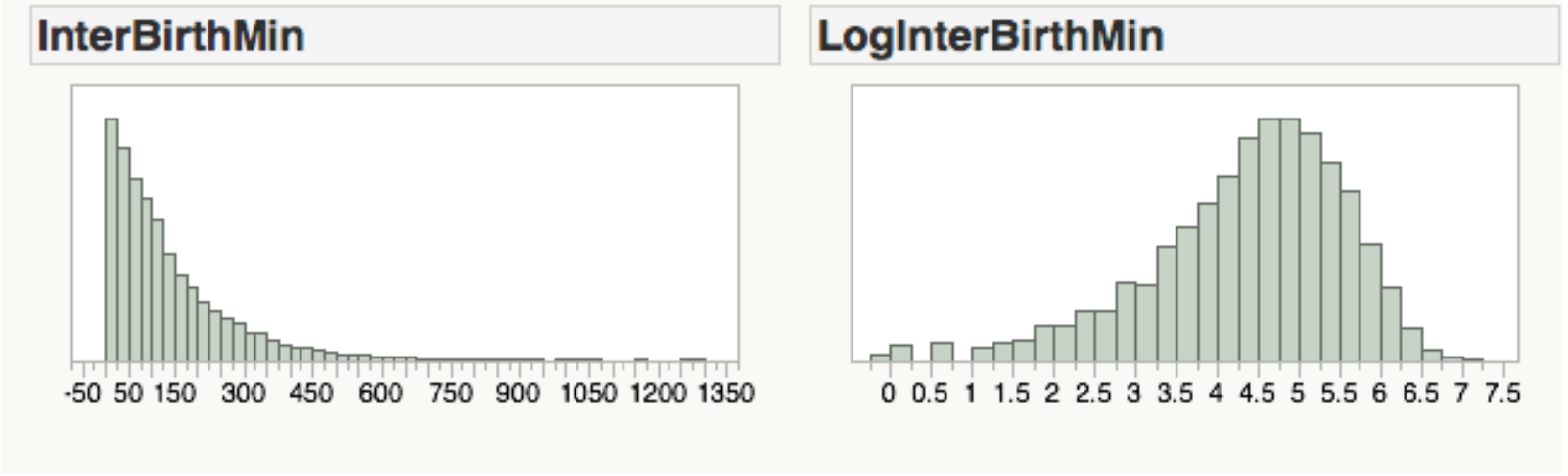
The Normal Quantile Plot

- Normal distributions: nice models for a lot of data.
- Nice calculation can be done if assuming normality.
- Normality is not everywhere!!!
  - Economic variables: income, gross sales of business
  - Financial variables: stock/option price
  - Other variables: conversation time
- Dangerous to assume normality without testing it.
- The normal quantile plot is a graphical diagnostic tool
  - to decide whether the data come from a normal distribution
- https://www.youtube.com/watch?v=okjYjClSjOg

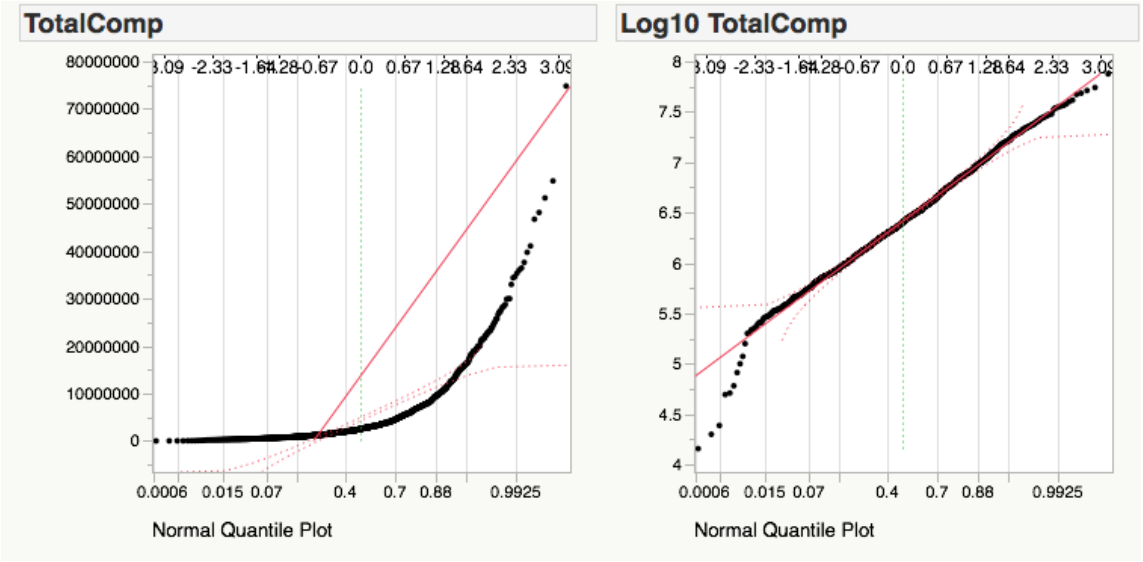
Histogram: Compensation, InterBirthMin

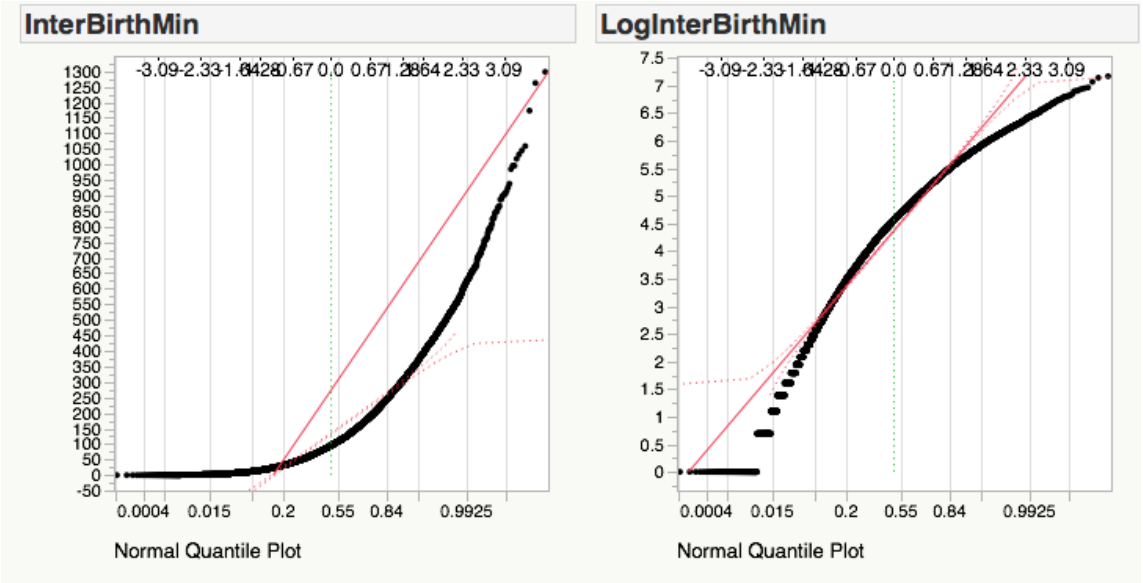


- Time (in minutes) between two consecutive newborns at a hospital (barring twins)



Normal Quantile Plot: Compensation, InterBirthMin

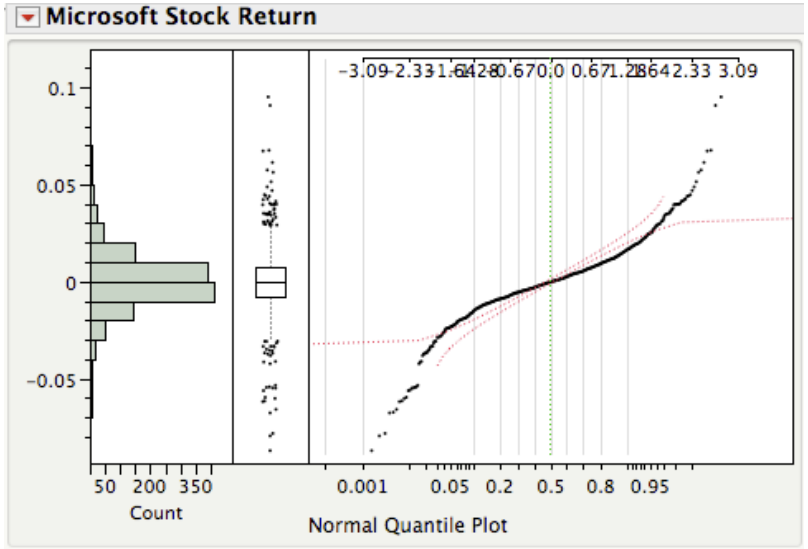




Use of Normal Quantile Plots

- If the points on a normal quantile plot lie close to a straight line, the plot indicates the data are normal.
- Systematic deviations from a straight line indicate a non-normal distribution.
- Outliers appear as points that are far away from overall pattern of the plot.
- Two dashed lines gauge amount of acceptable sample to sample variation
  - 95% of the time, a random sample from a normal distribution will lie between the two dashed lines

Microsoft Stock Return and Defects



- Key points
  - Many statistical inferences rely on normal distributions.
  - Use a normal quantile plot to check for normality rather than assume normality.

Association between Two Variables

Association between Two Categorical Variables

Outline

- So far we have focused more on individual variables
- Now we will focus on relationship between two variables

Two-way tables

- Two-way tables are used describe the relationship between two categorical variables. The tables contain counts or proportions (percentages)
- note: textbook uses the term “contingency tables”

Two-way tables

E.g. Cross-classification of a sample of 980 Americans by gender and party identification

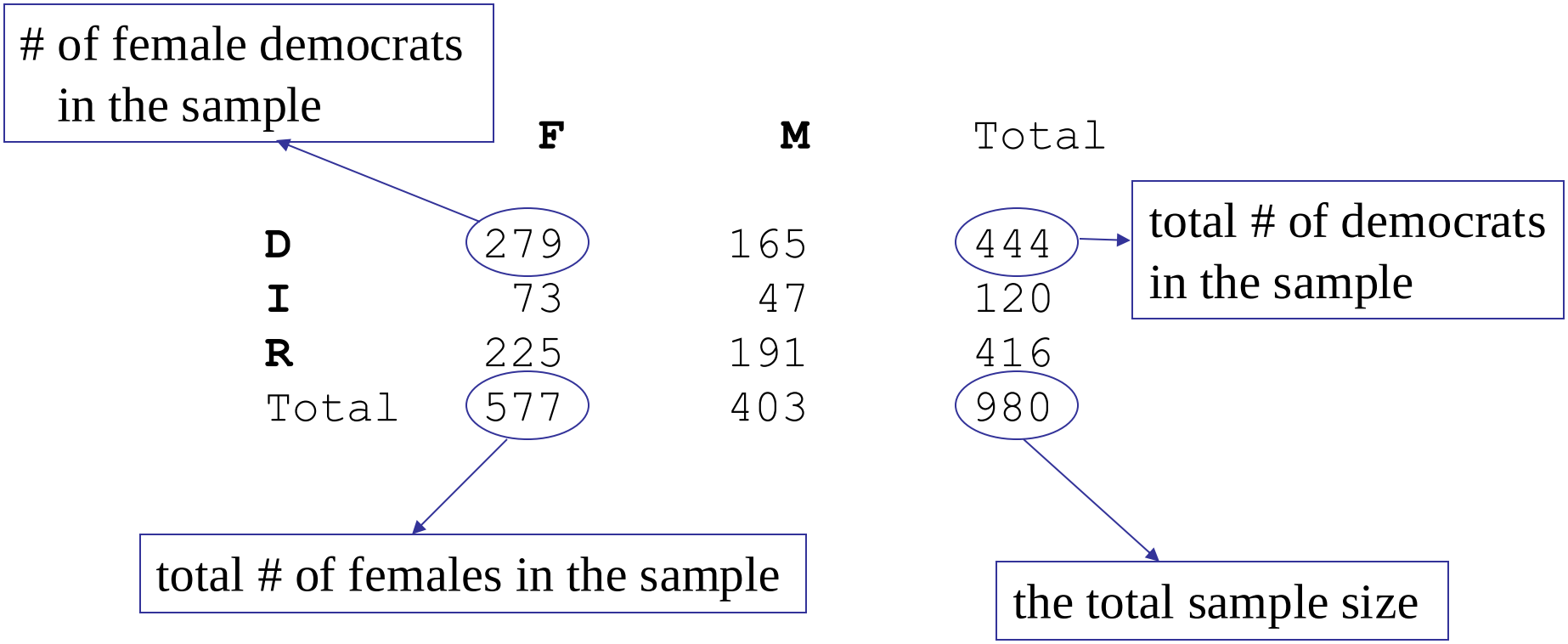
rows: Party (D,I,R) *columns: Gender (F,M)*

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Two-way tables

E.g. Cross-classification of a sample of 980 Americans by gender and party identification

rows: Party (D,I,R) *columns: Gender (F,M)*



Notation

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Party = **row variable**

Gender = **column variable**

Each combination of values of the two variables = **cell**

What is the total # of cells in the above table**?**

Joint distribution

A two-way table with proportions (or percentages) describes the **joint distribution** of the two variables.

Each cell gives the proportion of the total sample size

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Joint distribution

28.5% of the sample are female democrats

	F	M	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

45.3% of the sample are democrats

41.1% of the sample are males

Marginal distribution

Distribution of a single variable in a two-way table = **marginal distribution**

rows: "**Party**" columns: "**Gender**"

	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Marginal distribution

Distribution of a single variable in a two-way table = **marginal distribution**

	"Party"		"Gender"
D	45.3%	**F**	58.9%
I	12.2%	**M**	41.1%
R	42.5%		

marginal distribution of "**Party**"

rows: "**Party**" columns: "**Gender**"

	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Conditional distribution

distribution of one variable after we condition on (i.e. restrict our attention to) the value of the other variable = **conditional distribution**

E.g. What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Conditional distribution

…What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

D	279/577
I	73/577
R	225/577

D	48.4%
I	12.6%
R	39.0%

More definitions

- **Lurking Variable**: a variable that has an important effect but was overlooked

- **Simpson's Paradox**: a change in the direction of association between two variables when data are separated into groups defined by a third variable

Example

Two-way table:

	Hospital A	Hospital B
Died	300	50
Survived	2700	950

Death status distributions are specified by the % died:

Hospital A: 300/3000 = 10%

Hospital B: 50/1000 = 5%

Suppose the data on each patient's condition is also available

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Simpson's paradox:

Association between two variables has a different direction from the association conditional on a third variable (lurking variable)

What is the lurking variable in our example?

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Response and explanatory variables

A **response** (dependent) variable is the variable of interest (measures the outcome of a study)

An **explanatory** (independent) variable explains or causes changes in the response variable

e.g. income and education (in years)

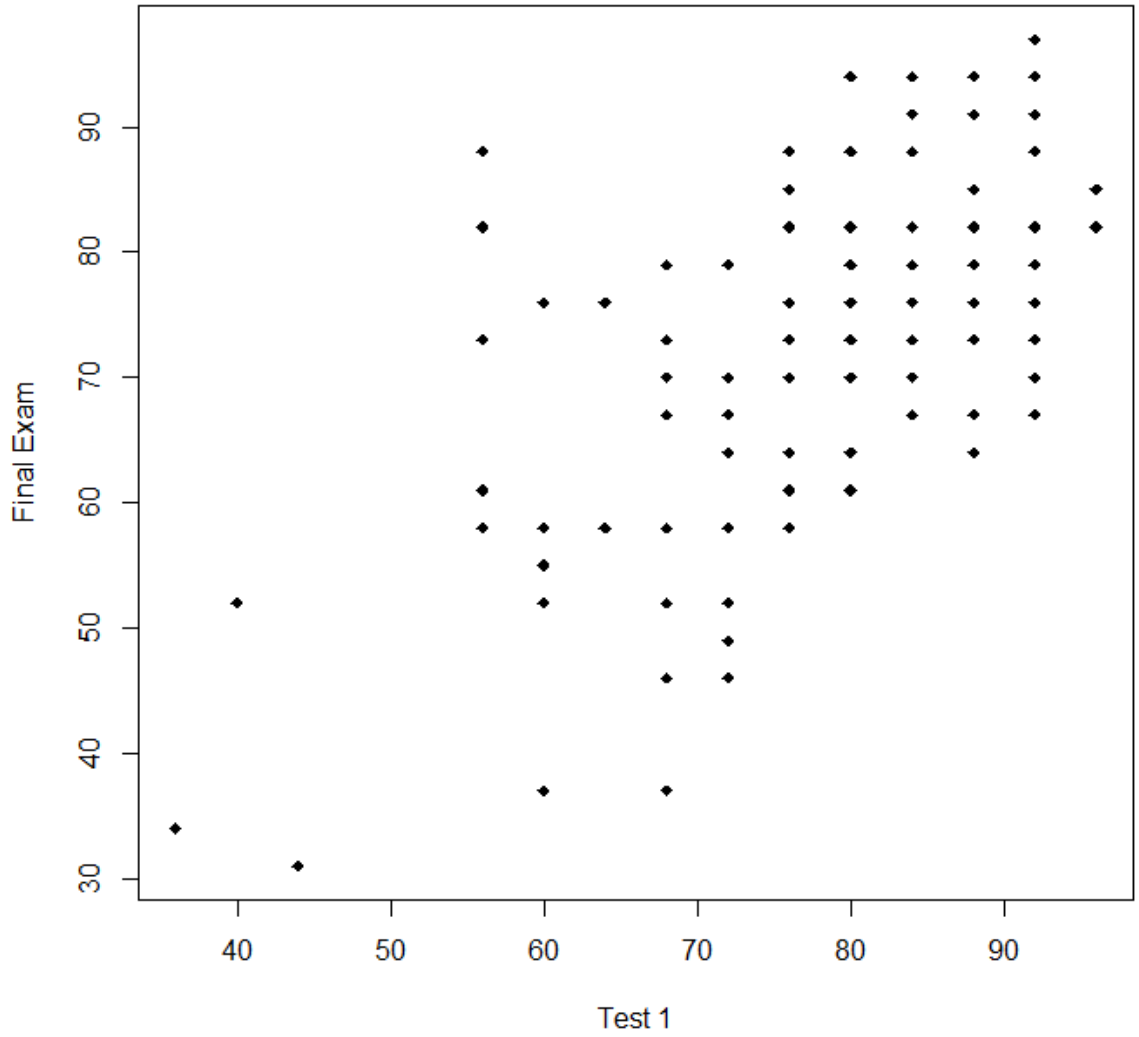
Note: definition works for both quantitative and qualitative variables

Association between Two Numerical Variables

Scatterplot

- Shows the relationship between two quantitative variables measured on the same individuals

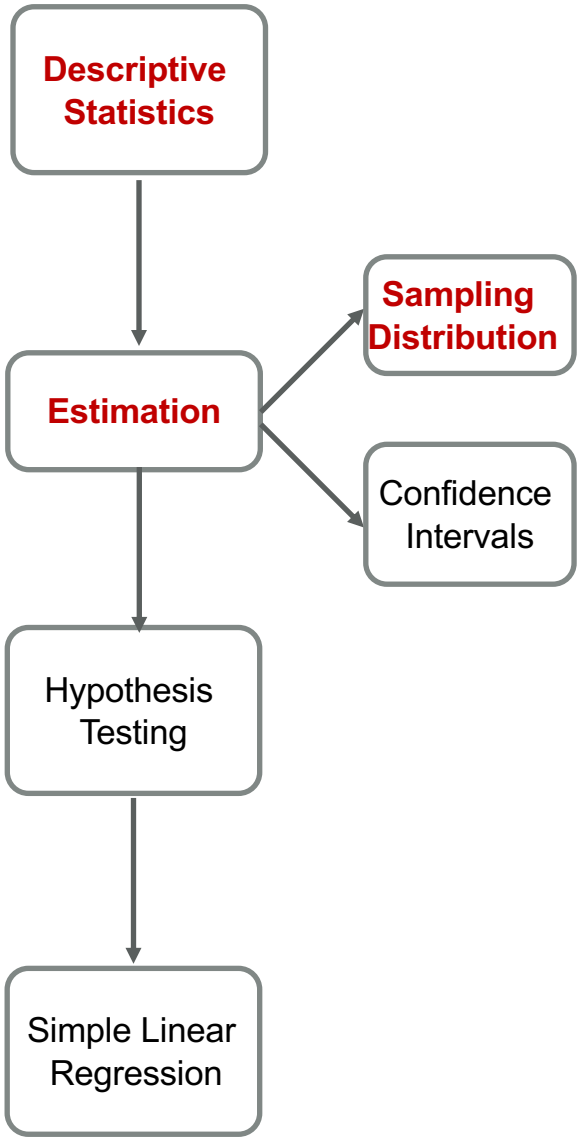
- The values of one variable -> horizontal axis
- The values of the other variable -> vertical axis
- Each individual appears as a point in the plot
- Explanatory variable (if there is one) -> horizontal axis, Response -> vertical axis


Final Exam
30
40
50
60
70
80
90
40
50
60
70
80
90
Test 1

Example: Scatterplot

- How to model the relationship between Y and X?
- To be taught in Stat II…

Overview


Descriptive Statistics
Estimation
Sampling Distribution
Confidence Intervals
Hypothesis Testing
Simple Linear Regression

Motivating Example: Case Double E (EE)

- EE is one of the largest retailers of consumer electronics in USA. However, of late, EE's profits have been declining. The primary reasons for this are suspected to be falling quality of service and growing competition
- Managers suspect that a number of customers use EE to learn about a product but do not buy from EE (called pseudo customers). Such customers add nothing to EE's revenues and reduce the quality of service for true customer

- EE management, based on the costs and industry benchmarks, has concluded that if < 20% of a salesperson' s day (≈1 hour and 36 minutes of an 8-hour day) is spent with pseudo customers, then the drain on service personnel by pseudo customers will not be considered a serious problem

- Otherwise, EE must change policy to cut down pseudo customers

- The management team at EE would like to know the average time a salesperson spends on pseudo customers
- However, all it has is the information of 100 salespersons (a sample)
- What is the best way to use the sample to estimate the population (or "true") mean?

The sample mean, $\overline{x}$

- Sample size, $n = 100$.
- Sample mean, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 4800.03$ seconds.
- The threshold set by management

  = 1hr 36 minutes = 5760 seconds

- The sample mean is below it!

- Is the time spent by every salesperson with pseudo customers the same as the population mean?

  -No, it fluctuates!

  -We will use a distribution to summarize it

  -Are a few spending a long time while the others are spending a short time? Is there a serious fluctuation of the time spent?

- Need to estimate the distribution's standard deviation

- Based on a sample, the best we can use is the sample standard deviation s, where

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

- So $s = 2610.62$ seconds

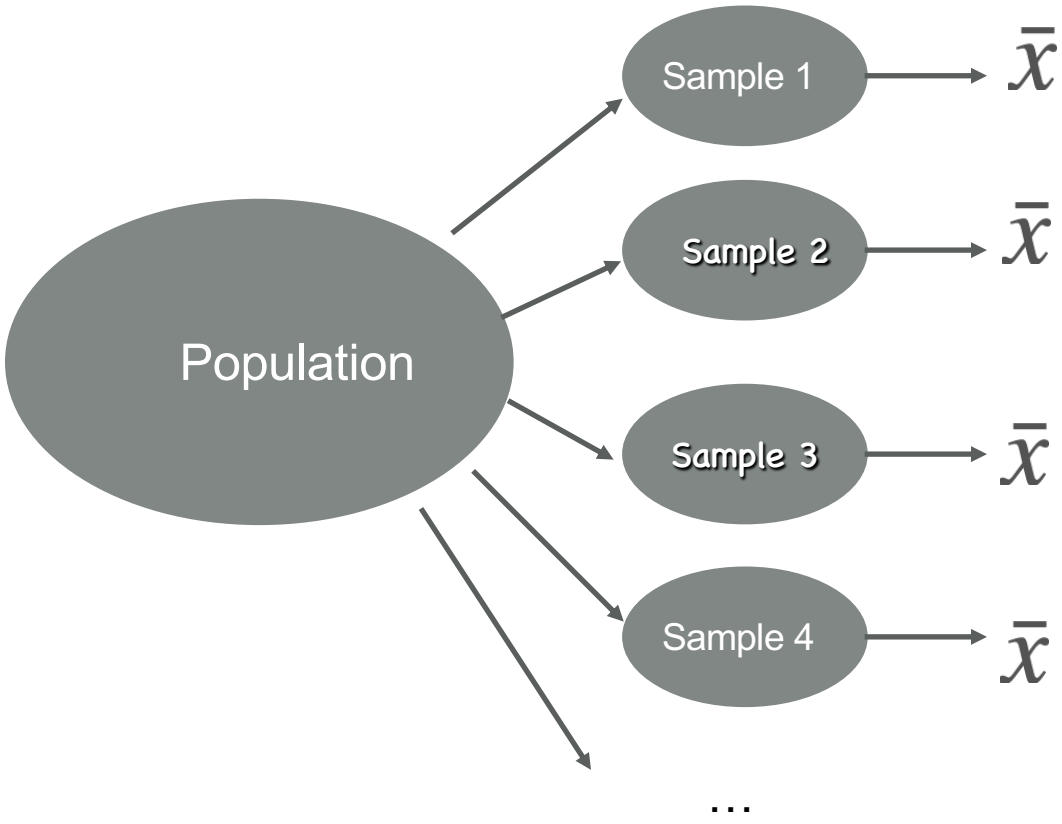
Sample Variation

- The value of the sample mean varies from sample to sample
- The source of the variation in the values of the sample mean is the potential variation in the sample drawn from the population
- We can view the sample mean as a random variable having a probability distribution. This distribution is called the sampling distribution of the sample mean

Sampling Distribution of the Sample Mean

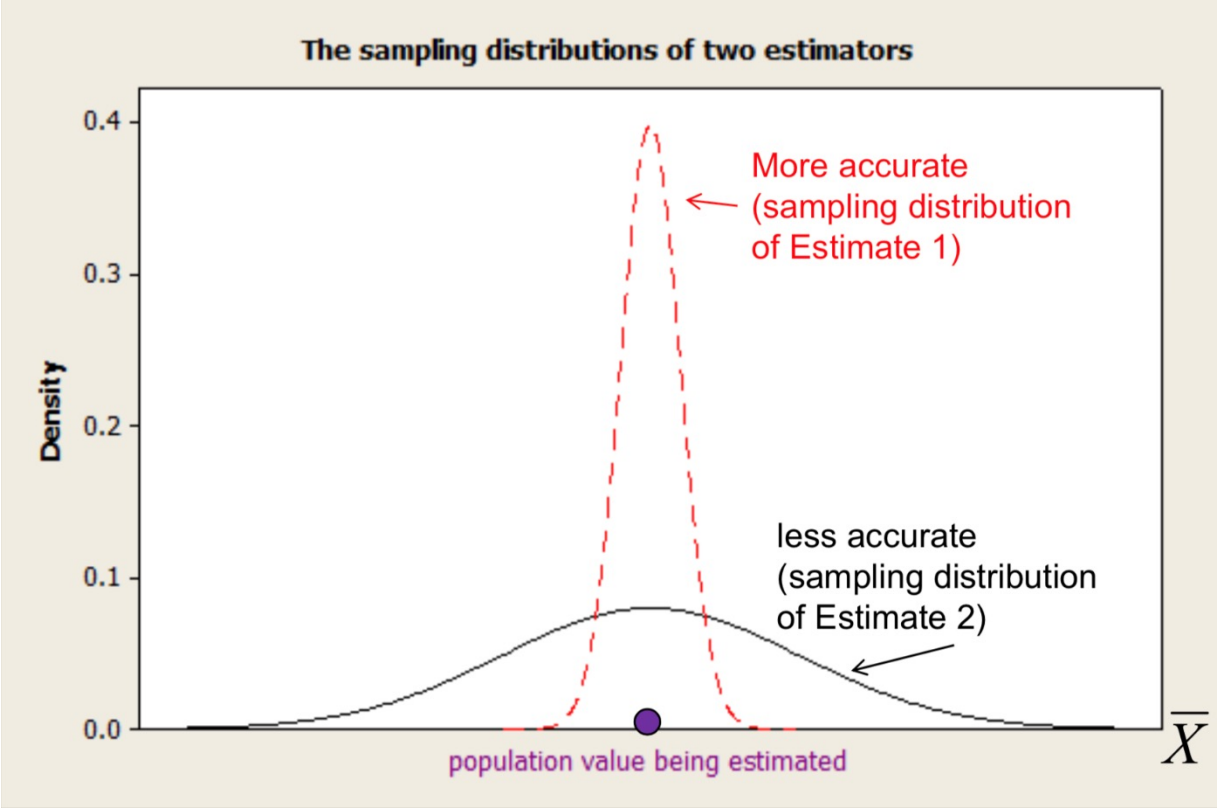
- Sampling distribution: calculate the sample statistic for every possible sample (of size n). The distribution of these sample statistics is the sampling distribution
- Sampling distribution of the sample mean $\bar{X}$ is the probability distribution of the population of the sample means obtainable from all possible samples of size n from a population of size N.

Draw samples of size n from the population



Sampling distribution is constructed using all these $\bar{x}$' s.
It centers at the population mean, i.e.
$E(\bar{X}) = \mu$

- A sampling distribution tightly concentrated around the mean tells us that the estimator is likely to be much more accurate (i.e., closer to the true value and has a smaller standard deviation) than one that has a sampling distribution widely dispersed around the average (i.e., has a larger standard deviation)

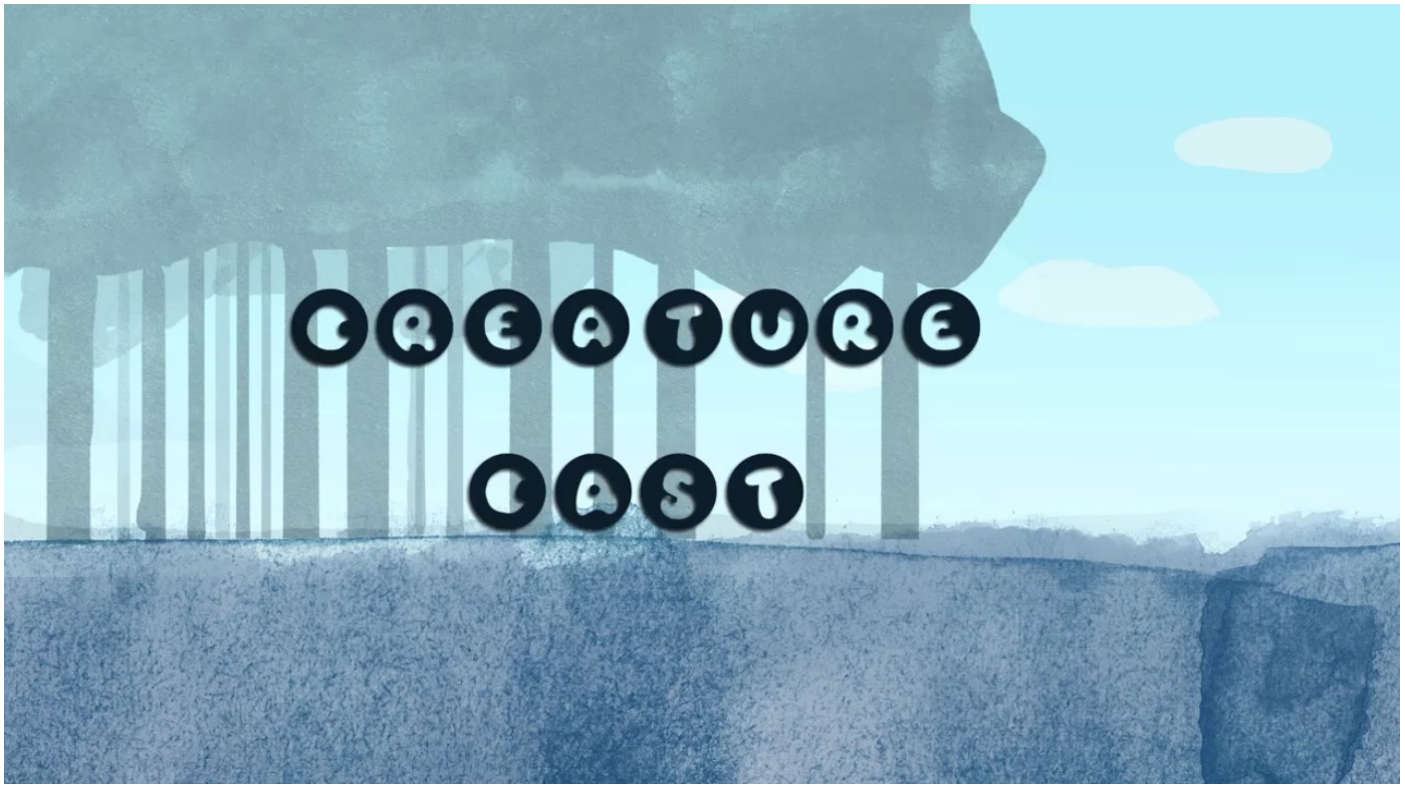

The sampling distributions of two estimators
Density
0.4
0.3
0.2
0.1
0.0
More accurate
(sampling distribution
of Estimate 1)
less accurate
(sampling distribution
of Estimate 2)
population value being estimated
$\overline{X}$

- How accurate an estimator is the sample mean?
- What is the sampling distribution of the sample mean?

Central Limit Theorem

- As the sample size increases, the sampling distribution of the sample mean approaches the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, i.e.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$



https://www.youtube.com/watch?v=jvoxEYmQHNM

Example

- Cola bottles are filled so that contents X have a normal distribution with $\mu = 298ml$ and standard deviation $\sigma = 3ml$

- What proportion of bottles have less than 295ml?

$$X \sim N(298, 3^2)$$

$$P(X < 295) = P(\frac{X - 298}{3} < \frac{295 - 298}{3}) = P(Z < -1) = 0.1586$$

  - In the first equation, we used: if $X \sim N(\mu, \sigma^2)$, then
    - $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$
  - In the last equation, we used the Z table.

Example Continued

- What is the probability of the average of a six pack of bottles being less than 295ml?

$$\bar{X} \sim N\left(298, \left(\frac{3}{\sqrt{6}}\right)^2\right)$$

$$P(\bar{X} < 295) = P\left(\frac{\bar{X} - 298}{3/\sqrt{6}} < \frac{295 - 298}{3/\sqrt{6}}\right)$$

$$= P(Z < -2.45) = 0.0071$$

Example Continued

- $P(X < 295) = 0.1586$, but $P(\bar{X}<295) = 0.0071$.
- Averages have less variation than individual observations
- As the sample size increases, the variation in the distribution decreases so that a value like 295ml is very difficult and rare to occur in an average of a six pack or more of bottles, but could quite easily occur in a single bottle

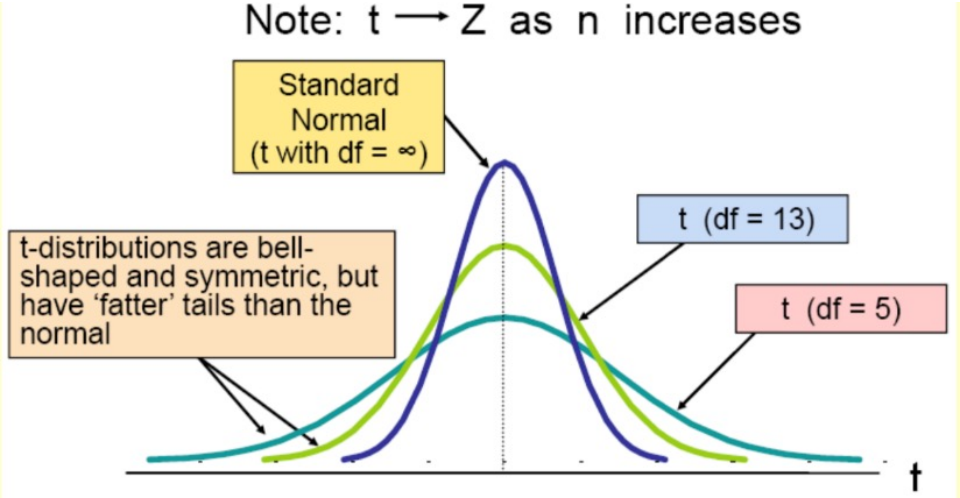
The Sampling Distribution of the Standardized Sample Mean

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

- s is only an estimate based on the sample.
- It introduces some additional sampling error into our calculations.
- The estimation part (s) is reflected in the fatter tails of the t-distribution, compared to the standard normal.

Properties of t Distributions

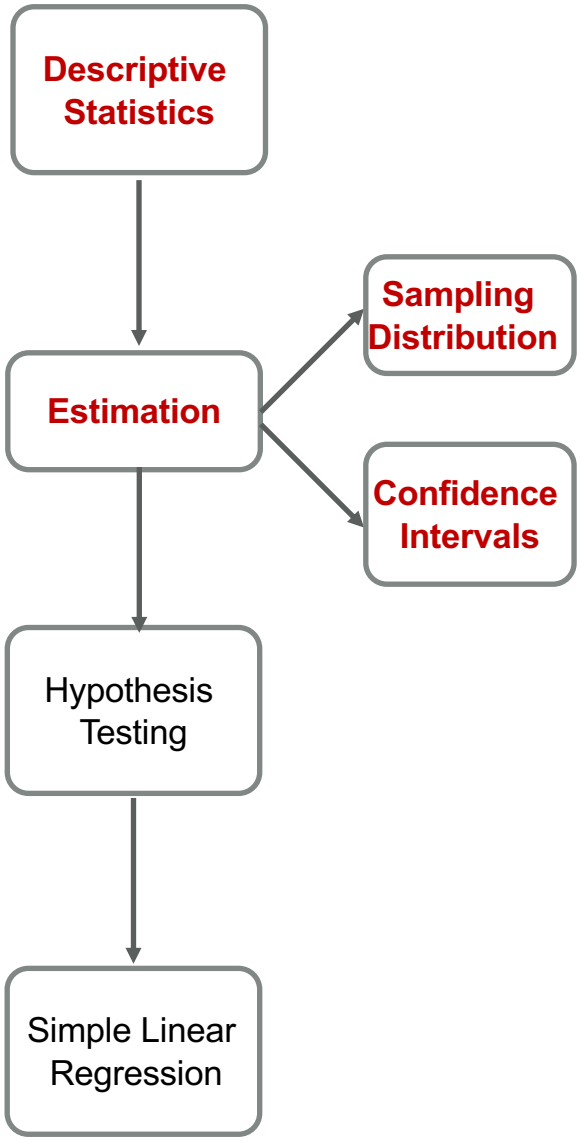
- t is a family of continuous distributions
- It is indexed by the degree of freedom (df)
- df = n-1, where n is the sample size



A Quick Summary

Known σ; normal population or large sample ($n \geq 30$)	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
Unknown σ; normal population and small sample	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
Unknown σ; large sample ($n \geq 30$)	$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$

Overview


Descriptive Statistics
Estimation
Sampling Distribution
Confidence Intervals
Hypothesis Testing
Simple Linear Regression

Confidence Intervals

Why a point estimate is not enough?

- Drawbacks:
  1. It is almost certain that the estimate will be wrong based on a single sample
  2. What if we want to know how close this estimator is to the true parameter μ (mean) and p (proportion)?
  3. Intuitively, larger samples will produce more accurate results, but point estimators alone does not fully reflect the effect of large sample size

Interval Estimate

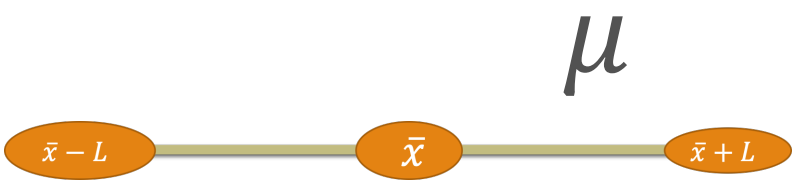
- Imagine you try to capture a butterfly in the dark. Is it better to use
  - A dart?
  - Or a large net?

– Instead of using a point estimate (using a single value to estimate the population parameter), we can use an interval estimate, where we use a range of values to estimate the population parameter (we say that the population parameter is within our interval estimate)

Confidence Interval

- Most often, it is very informative to say
  - I don't know exactly what the mean is, but I am fairly confident that it is between XXX and YYY.
  - For example, I don't really know what the mean diameter is, but I am almost certain that it is between 814.90 and 815.08.

- This is a *Confidence Interval*
  - Much more informative and realistic than just stating that ``I estimate the mean to be 814.99."

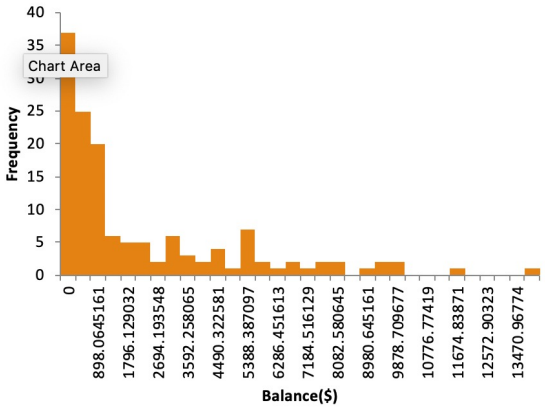


Motivating Example: Launch a New Credit Card

- To launch an affinity credit card, the contemplated launch process proposes sending pre-approved application to N= 100,000 alumni of a large university (population)
- Two parameters of the population determine whether the card will be profitable:
  - p, the proportion who will return the application
  - μ, the average monthly balance carried by those who accept the card

- To estimate the parameters, the credit card issuer sent pre-approved application to a sample of 1,000 alumni. Of these, 140 accepted the offer and received a card

Number of offers	1000
Number accepted	140
Proportion who accepted	$\hat{p} = 0.14$
Average balance	$\overline{x} = 1990.5$
SD of balance	$s = 2833.33$



Computing the Interval

- The CLT suggests,

$$P(\mu - 2SE(\bar{X}) \leq \bar{X} \leq \mu + 2SE(\bar{X})) \approx 0.95$$

- Alternatively,

$$P(\bar{X} - 2SE(\bar{X}) \leq \mu \leq \bar{X} + 2SE(\bar{X})) \approx 0.95$$

- Suppose we take a sample and compute $\bar{X}$ and $SE(\bar{X})$. Then, we can state that μ is somewhere between

$$\bar{X} - 2SE(\bar{X}) \text{ and } \bar{X} + 2SE(\bar{X}),$$

- and be (approximately) 95% sure that we are correct.

General Confidence Interval

- In general, the confidence interval will be

$$[\bar{X} - z * SE(\bar{X}),\ \bar{X} + z * SE(\bar{X})]$$

- where z determines how sure we are of being correct.
- Some common choices for z:
  - 90% interval: z=1.645
  - 95% interval: z=1.96
  - 99% interval: z=2.57
- Note the trade-off between the size of interval and probability of being correct
  - Margin of error: $z * SE(\bar{X})$, half width of the interval

Confidence Intervals: σ Unknown

- Use s to estimate σ
- OK to use $[\bar{X} - z * s/\sqrt{n},\ \bar{X} + z * s/\sqrt{n}]$?
- No! This ignores the extra uncertainty of estimating σ.
  - As a result, the interval is narrower than it is supposed to be.
- When σ is unknown, use s to estimate σ and use the T distribution (with d.f. n-1) in place of the normal.

$$[\bar{X} - t * s/\sqrt{n},\ \bar{X} + t * s/\sqrt{n}]$$

- The t intervals are wider than the z intervals.
- For n>30, very similar results.

General Formula for $1-\alpha$ Confidence Interval

Point Estimate $\pm$ (Critical Value)*(Standard Error)

Known σ; normal population or large sample ($n \geq 30$)	$\overline{X}$	$z_{\alpha/2}$	$\frac{\sigma}{\sqrt{n}}$
Unknown σ; normal population and small sample	$\overline{X}$	$t_{\alpha/2,n-1}$	$\frac{s}{\sqrt{n}}$
Unknown σ; large sample	$\overline{X}$	$z_{\alpha/2}$	$\frac{s}{\sqrt{n}}$
Unknown p; $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$	$\hat{p}$	$z_{\alpha/2}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

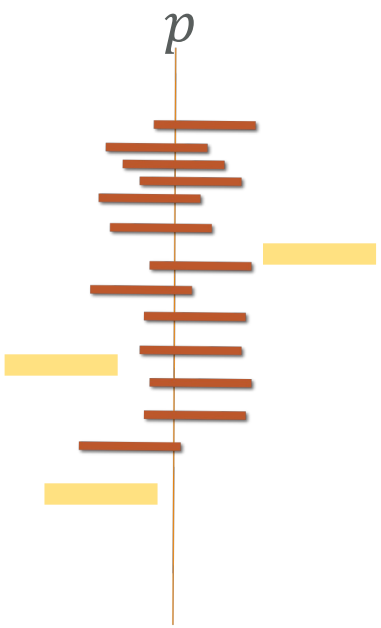
For all the other cases, we cannot apply the z or t intervals

Back to the Credit Card Example

- In the data, the estimated standard error is
  - $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.14(1-0.14)}{1000}} \approx 0.011$
- The 95% CI for p is
  - $[0.14 - 1.96 * 0.011, 0.14 + 1.96 * 0.011] = [11.84\%, 16.16\%]$
- Interpretation to non-expert: we are 95% confident that the population proportion that will accept this offer is between about 12% and 16%
- Is it correct to say $P(11.84\% \leq p \leq 16.16\%) = 95\%$?

Statistical Interpretation

- We say "we are 95% confident that…"
- But NEVER say "the probability that the true proportion p is between 12% and 16% is 0.95"
  - For a realized confidence interval, the true mean p is either in there or not
- Statistically, it really means: if you line up the 95% confidence intervals from many, many samples, 95% of these intervals would cover the population parameter p



Back to the Credit Card Example

- The monthly balance for the $n = 140$ customers:
  - $\bar{x} = 1990.5$ and $s = 2833.33$
- Since n>30, we may treat $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$
- The 95% CI for the balance is
  - $[1990.5 - 1.96 * (2833.33/\sqrt{140}), 1990.5 + 1.96 * (2833.33/\sqrt{140})]$ $= [1523.553, 2457.447]$

Margin of Error

- $L = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, or $t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}}$, or $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$,…
- Three factors affects the margin of error
  - Confidence level: 90%, 95%
  - Sample size: n
  - The population variation or the sample variation

Sample Size Requirement

- To control the Margin Error to be less than some L, there are some requirements on sample sizes.
- For proportions
  - with unknown p: $L = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \rightarrow n \geq \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{L^2}$, we stick to this formula for our class.
  - Sometimes, since $\sqrt{a(1-a)} \leq 1/2$ for any a, people may simply use $n \geq \frac{z_{\alpha/2}^2 \times 0.25}{L^2}$, which is greater than $\frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{L^2}$ or $\frac{z_{\alpha/2}^2 p(1-p)}{L^2}$

Example - Sample Size for p

- What sample size is needed to estimate the true proportion of defective bulbs to be within $\pm 5\%$ with 90% confidence? Out of a population of 1000 bulbs, we randomly select 100 of which, and 30 among them are defective
- $\hat{p} = 0.3,\ n = 100$
- $n \geq \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{L^2} = \frac{1.645^2 \times 0.3 \times 0.7}{0.05^2} = 227.3$
- So the minimum sample size requirement is 228

Sample Size Requirement

- To control the Margin of Error to be less than some L, there are some requirements on sample sizes.
- For mean μ
  - with known σ: $L = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n \geq \frac{z_{\alpha/2}^2\sigma^2}{L^2}$
  - with unknown σ: $n \geq \frac{z_{\alpha/2}^2 s^2}{L^2}$

Example - Sample Size for μ

- A lumber company has acquired the rights to a large tract of land containing thousands of trees
- A lumber company needs to estimate the amount of trees it can harvest in a tract of land to determine whether the effort will be profitable
- To do so, it must estimate the mean diameter of the trees
- It decides to estimate that parameter to within ±1 inch with 95% confidence
- A forester familiar with the territory guesses that the diameters are normally distributed with a standard deviation of 6 inches

- Question 1: determine the number of trees he should sample

- Question 2: After the sample is taken, the forester discovered that the sample mean is 25 and the sample standard deviation is 12. He decided to use the sample standard deviation instead of his old guess. What is the 90% confidence interval then?

Q & As

- For the same sample, the width of a confidence interval will be
  - A. Narrower for 99% confidence than 95% confidence
  - B. Wider for a sample size of 100 than for a sample size of 50
  - C. Narrower for 90% confidence than 95% confidence

Q & As

- With the same confidence level, as standard deviation increases, samples size need to _____ to achieve a specified margin of error

  A. Increase

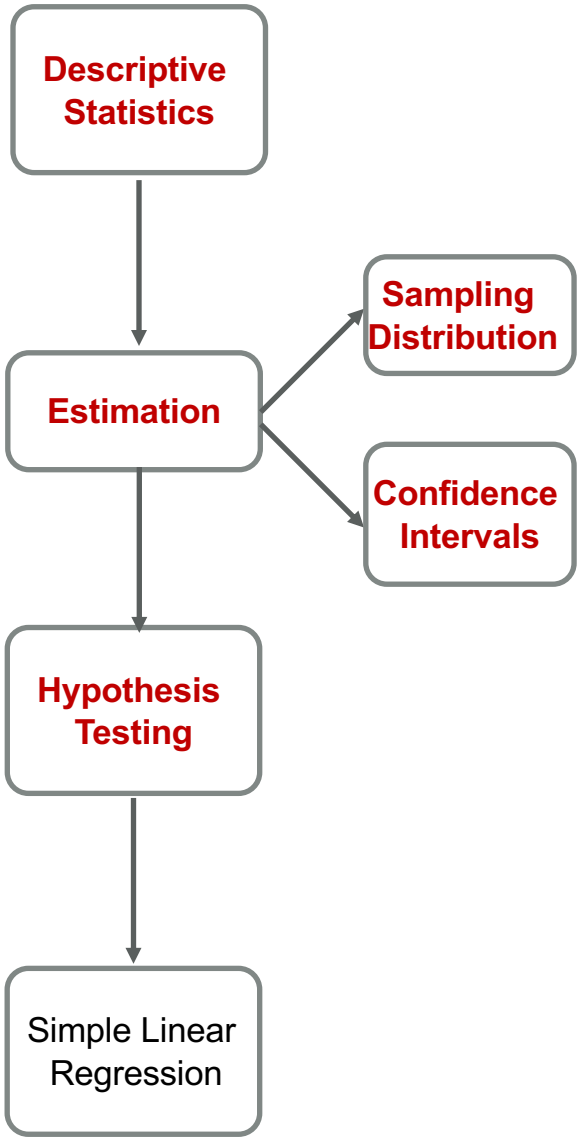
  B. Decrease

  C. Remains the same

Confidence Interval

- Provides range of values based on observations from one sample
- Stated in terms of confidence level
- Examples: 90% CI for the μ, 95% CI for the p

Overview


Descriptive Statistics
Estimation
Sampling Distribution
Confidence Intervals
Hypothesis Testing
Simple Linear Regression

Hypothesis Testing

Procedures for Statistical Inferences

- Point estimation
  - ``I think the population parameter is XXX."
- Confidence interval
  - ``I am 95% confident that the population mean is between XXX and YYY."
- Hypothesis testing
  - ``I think your claim that no more than 10% of the clicks are fraudulent is not valid."
  - Using statistics to decide which of two possibilities is the truth given imperfect information
  - Hypothesis: a statement/claim/belief about the parameter
    - Two hypotheses: two possibilities (complement of each other)

Realty Agency Expansion

A realty agency manages rental properties, and is considering expanding into the Denver metropolitan area.

To justify the costs of opening a new office, the agency needs rents in the area to be more than $500 per month.

Lower rental generates smaller fees that would make the office unprofitable.

Are rents in Denver high enough to justify the cost of the move?

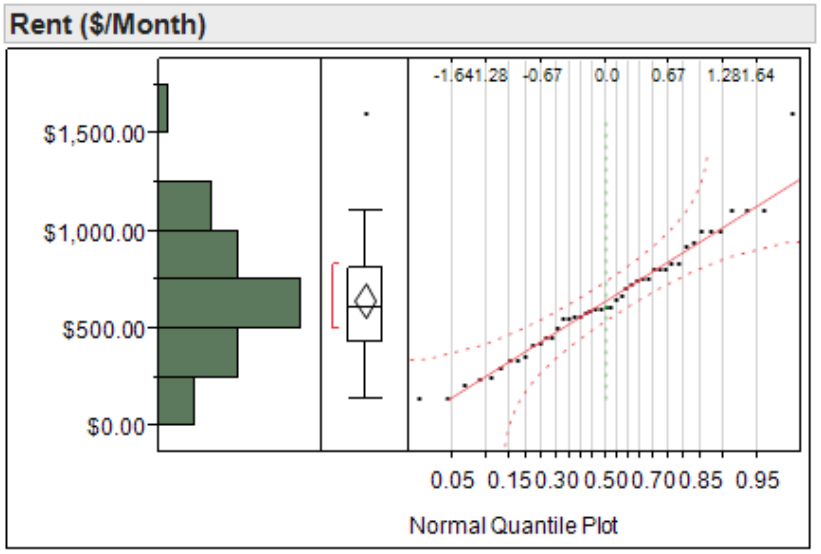
Realty Agency Expansion

Managerial decision: Expand vs. Don't expand

Population: all rental properties in the area, mean rent μ

- Is $\mu > 500$?

Data collection: 45 houses, sample mean $\bar{x}$ = \$647.33



The data suggests that the mean rent μ is above \$500.

Hence Expand! (Wait! is the evidence strong enough?)

Concepts of Hypothesis Testing

- Two hypotheses:
  - H_a: the alternative hypothesis
    - The statement we hope or suspect is true.

    $H_a : \mu > \$500$ (one-sided alternative)
  - H_0: the null hypothesis
    - The statement of "no effect" or "no difference".
    - The statement we try to find evidence against.

    $H_0 : \mu = \$500$
- Usually one would decide on H_a first
  - Sometimes, make sense to consider $H_a : \mu \neq \$500$ (two-sided alternative)

Basic Ideas of Hypothesis Testing

To “prove” or “establish” some claim statistically based on the data collected in a sample

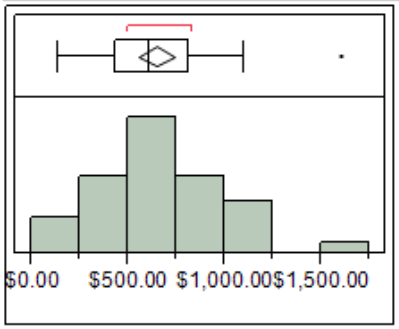
- Prove by contradiction
- Assume that the “opposite” is true
  - This “opposite” is your Null Hypothesis H_0
- Given that H_0 is true, calculate the probability for you to see what you ``saw” (i.e. the data)
  - This probability is called the ``p-value,” which is the probability of seeing ``data this unusual” due to random chance alone.
- If the p-value is very small, then “probably” what you assume – H_0 – is incorrect.
  - Then, “reject” H_0 and conclude that what you claim is true.
  - Otherwise, “can not reject” H_0 and the data do not support your claim.

Denver_Rent: One-sample T Test

- Test H_0: $\mu = 500$ versus H_a: $\mu > 500$

Distributions

Rent ($/Month)

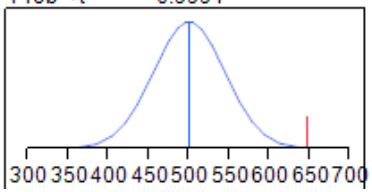


Quantiles		
100.0%	maximum	1600
99.5%		1600
97.5%		1525
90.0%		1040
75.0%	quartile	815
50.0%	median	610
25.0%	quartile	435
10.0%		246
2.5%		140
0.5%		140
0.0%	minimum	140

Summary Statistics	
Mean	647.33333
Std Dev	298.76564
Std Err Mean	44.537352
Upper 95% Mean	737.09247
Lower 95% Mean	557.5742
N	45

Test Mean	
Hypothesized Value	500
Actual Estimate	647.333
DF	44
Std Dev	298.766

t Test	
Test Statistic	3.3081
Prob > |t|	0.0019*
Prob > t	0.0009*
Prob < t	0.9991



- The p-value is 0.0009 or (1 out of 1111)!
  - In other words, if the mean is not above $500, you would need to collect 1111 such samples, before you can see the sample mean this high above $500!
  - In units of standard errors, $\bar{x} = \$647$ lies 3.3 standard errors from the hypothesized value $\mu_0 = \$500$.
- We are convinced that $\mu > 500$! (Are we 100% correct?)

P-value

- The probability of observing ``the data at least this unusual as what we saw", assuming that H_0 is true.
  - In the direction of the alternative
- The amount of statistical evidence that supports H_0
  - The smaller a *p-value*, the less evidence for H_0, or more evidence for H_a

❑ Small p-values indicate:

- false H_0 or something rare happened; statistical practice is to believe in the former, hence reject H_0

Rent: P(observing $\bar{x} > \$647$) = .0009

- If μ were \$500, observing $\bar{x} > \$647$ would occur in only 1 out of 1111 samples!
- So we reject H_0. (There is risk of making Type I Error! But Prob<0.0009)

Test Statistic

- A test is based on a statistic, which estimates the parameter that appears in the hypotheses
  - Point estimate

- Values of the estimate far from the parameter value in H_0 give evidence against H_0.
- H_a determines which direction will be counted as "far from the parameter value".

- Commonly, the test statistic has the form

  T=(estimate-hypothesized value)/(standard deviation of the estimate)

One-Sample T Test: Test Statistic

- Parameter μ with hypothesized value μ_0
- Estimate $\bar{X}$ with observed value $\bar{x}$, and estimated standard deviation $s/\sqrt{n}$

- Test statistics

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

with observed value

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

One-Sample T Test: p-value

- State null and alternative hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \begin{array}{l} \mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0 \end{array}$$

- p-value equals, assuming H_0 holds

$$2\mathrm{P}(T \geq |t|)$$

$$\mathrm{P}(T \geq t)$$

$$\mathrm{P}(T \leq t)$$

Legal System: Type I and Type II Errors

		Decision	
		Acquit	*Convict*
Truth	*Innocent* H_0	Correct	Type I error
	Guilty H_a	Type II error	Correct

- Given evidence (partial information), the jury is always at risk of making a mistake.

- Type I error (wrongly sentence an innocent person)
  - Safe strategy to remove Type I error is to let everybody go free
- Type II error (wrongly let a guilty person go free)
  - Safe strategy to remove this error is to convict everybody

- Tradeoff between the two errors
  - Unless with perfect information

Hypothesis Testing: Type I and Type II Errors

		Decision	
		H_0 true	H_a true
Truth	H_0	Correct	Type I error
	H_a	Type II error	Correct

- Denver Rent Example:
  - Type I error: Reject H_0 and claim profitable when it's not
  - Type II error: Fail to reject H_0 and miss opportunity
  - Which error has the higher expected cost?

- Analogy to Legal System
  - A hypothesis testing between being innocent and guilty
  - H_0: innocent, H_a: guilty
  - Type I error: more severe

Common Practice in Hypothesis Testing

- Limit the chance of a Type I Error to a chosen level α
  - referred to as *significance level*
  - upper bound on Type I error
  - commonly set at 5%

- Reject H_0 when the p-value <= α

- If so, we claim that the data support the alternative H_a at level α, or
  - The data are statistically significant at level α

α and P-value

- P-value and significance level α :
  - Reject H_0 if p-value <= α
  - Do not reject H_0 if p-value > α.
  - Denver Rent: p-value=*0.0009*< *0.05*= α; hence reject H_0: $\mu = 500$

- When is the evidence against H_0 stronger?
  - *Large P-value or small P-value?*
  - The smaller the *P-value*, the stronger the evidence against H_0 and in favor of the alternative H_a.

- When is it easier to reject H_0?
  - *Large α or small α ?*
  - We need a lot more evidence to reject H_0 for small α than for large α.

Summary: The One-Sample T Test

- Consider an iid sample $x_1, \ldots, x_n$ from a population with unknown mean μ
- State null and alternative hypothesis

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_a: \begin{array}{l} \mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0 \end{array}$$

- t Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$

  - Measure the distance between the observed (data) and the believe (null hypothesis)

- The larger, the more evidence against H_0.

Example: Click Fraud

A specialty retailer pays a hosting site for each click on an ad that brings customers to its web site. Recently, however, the retailer suspects that many of these clicks have been generated by automated systems designed to imitate real customers.

The on-line host has promised that no more than 10% of the clicks are imitations.

To learn more, the retailer hired a service to identify fraudulent clicks.

In a sample of 1,200 clicks, the service identified 175 computer-generated fraudulent clicks. The file *Click_Fraud* summarizes these counts.

Click Fraud: One-Sample T Test

- p : proportion of fraudulent clicks
- $H_0: p \leq 0.1$ versus $H_a: p > 0.1$

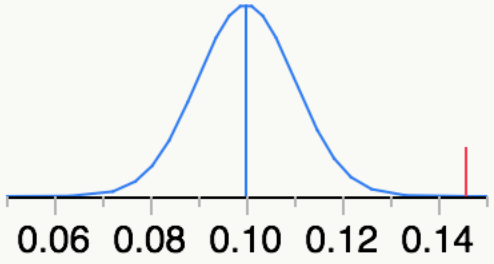
- Since the p-value is less than 0.05, we can reject the claim that $p \leq 0.1$

Test Mean

Hypothesized Value	0.1
Actual Estimate	0.14583
DF	1199
Std Dev	0.35309

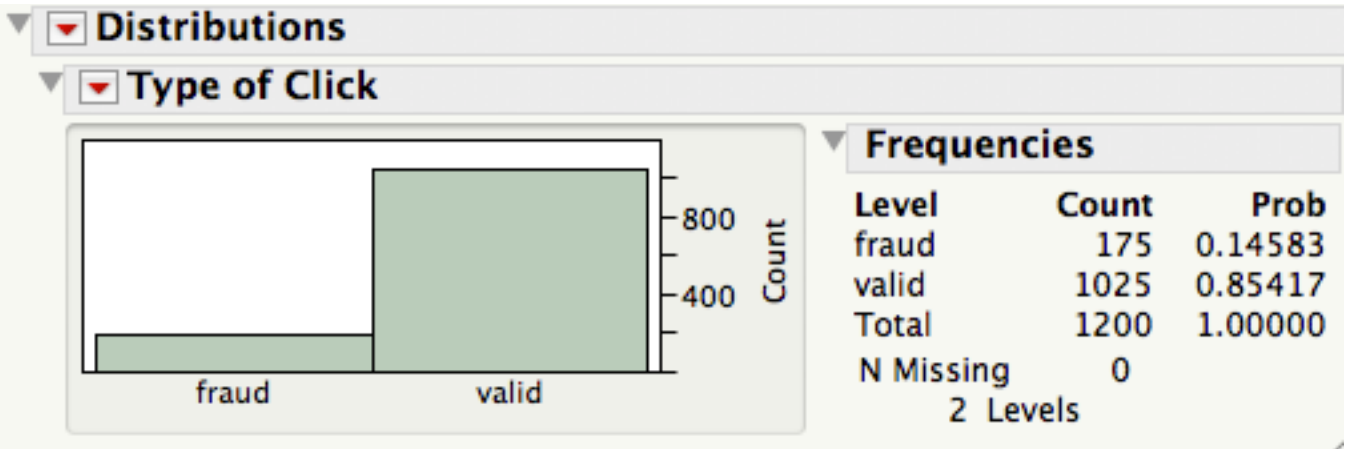
t Test

Test Statistic	4.4967
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000



Click Fraud

- p : proportion of fraudulent clicks
- Goal: verify whether p <=10%



- Solution: 95% confidence interval for p
  - 95% CI: [0.127, 0.167], which is above 0.10.
  - The host's claim is not valid, which is consistent with the earlier testing conclusion.

Summary: Testing Population Proportion

- State null and alternative hypotheses

$$H_0: p = p_0 \quad \text{vs.} \quad H_a: \begin{cases} p \neq p_0 \\ p > p_0 \\ p < p_0 \end{cases}$$

- Test statistic $$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

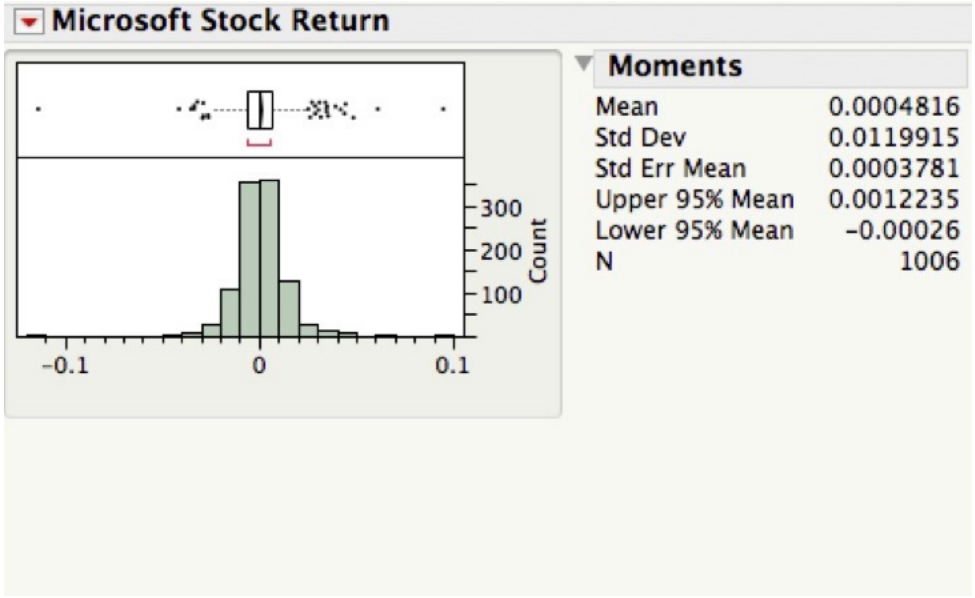
- Under $H_0: p \leq p_0$ and some conditions, the test statistic z is approximately standard normal.

Two-sided Hypothesis Tests

A two-sided hypothesis test detects a deviation in *either* direction from a claimed specific value for the population parameter. Confidence intervals provide an alternate method that can be used to test such hypotheses.

2004-2007 Microsoft Returns: mean return μ

H_0: $\mu = 0$ versus H_a: $\mu \neq 0$



Can we reject the null under 5% significance level?

Two-sided Hypothesis Tests

A two-sided hypothesis test detects a deviation in *either* direction from a claimed specific value for the population parameter. Confidence intervals provide an alternate method that can be used to test such hypotheses.

2004-2007 Microsoft Returns: mean return μ

Compute the 95% CI:

$$\left[0.0004816 - 1.96 * \frac{0.0119915}{\sqrt{1006}}, 0.0004816 + 1.96 * \frac{0.0119915}{\sqrt{1006}}\right]$$
$$= [-0.00026, 0.00122]$$

Two observations from the Stock Returns example:

- Under 5% level, one can not reject H_0: $\mu = 0$.
  - It is possible that $\mu = 0$
- The 95% CI for μ is [-0.00026, 0.0012], which includes 0.
  - So, it is possible that $\mu = 0$

Equivalence between CI and 2-Sided Test!

Equivalence between CI and 2-Sided Tests

H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$

A level α 2-sided test

- Rejects H_0 when the value μ_0 falls outside a level 1 - α confidence interval for μ.
- Can't reject H_0 when the value μ_0 falls inside the CI.

CI can be used to test 2-sided hypotheses:

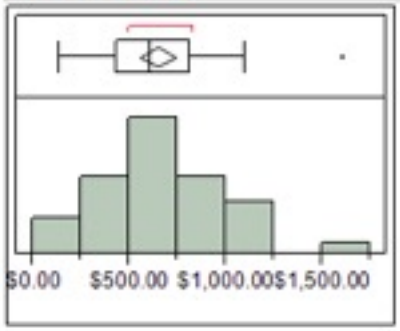
- Calculate the 1 - α level confidence interval
- Then
  - if μ_0 falls outside the interval, reject the null hypothesis;
  - otherwise, can't reject the null hypothesis.

Example: Denver_Rent

- Test H_0: $\mu = \$500$ versus H_a: $\mu \neq \$500$

Distributions

Rent ($/Month)

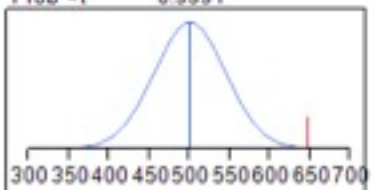


Quantiles		
100.0%	maximum	1600
99.5%		1600
97.5%		1525
90.0%		1040
75.0%	quartile	815
50.0%	median	610
25.0%	quartile	435
10.0%		246
2.5%		140
0.5%		140
0.0%	minimum	140

Summary Statistics	
Mean	647.33333
Std Dev	298.76564
Std Err Mean	44.537352
Upper 95% Mean	737.09247
Lower 95% Mean	557.5742
N	45

Test Mean	
Hypothesized Value	500
Actual Estimate	647.333
DF	44
Std Dev	298.766

t Test	
Test Statistic	3.3081
Prob > |t|	0.0019*
Prob > t	0.0009*
Prob < t	0.9991



- What about H_a: $\mu \neq \$600$?
- What about H_a: $\mu \neq \$750$?
- What is the range of values for μ that H_0 can not be rejected?

Parts of a Hypothesis Test

- All hypothesis tests work the same way
  - State the null and alternative hypotheses
  - Compute the p-value
  - Interpret the results

- The differences are only in how the hypotheses are stated and the p-value computed.

- The p-value measures how much evidence there is against the null. A small p-value says the data are inconsistent with H_0, so that we should reject it.

- How small p-value needs to be depends on the consequence of making a mistake.