MSBA Boot Camp

Statistics I

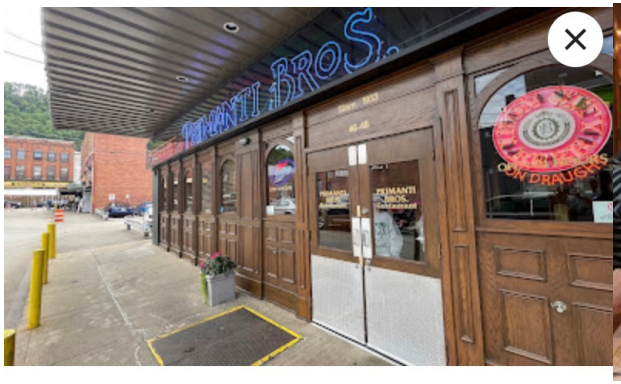
Zhanrui Cai

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

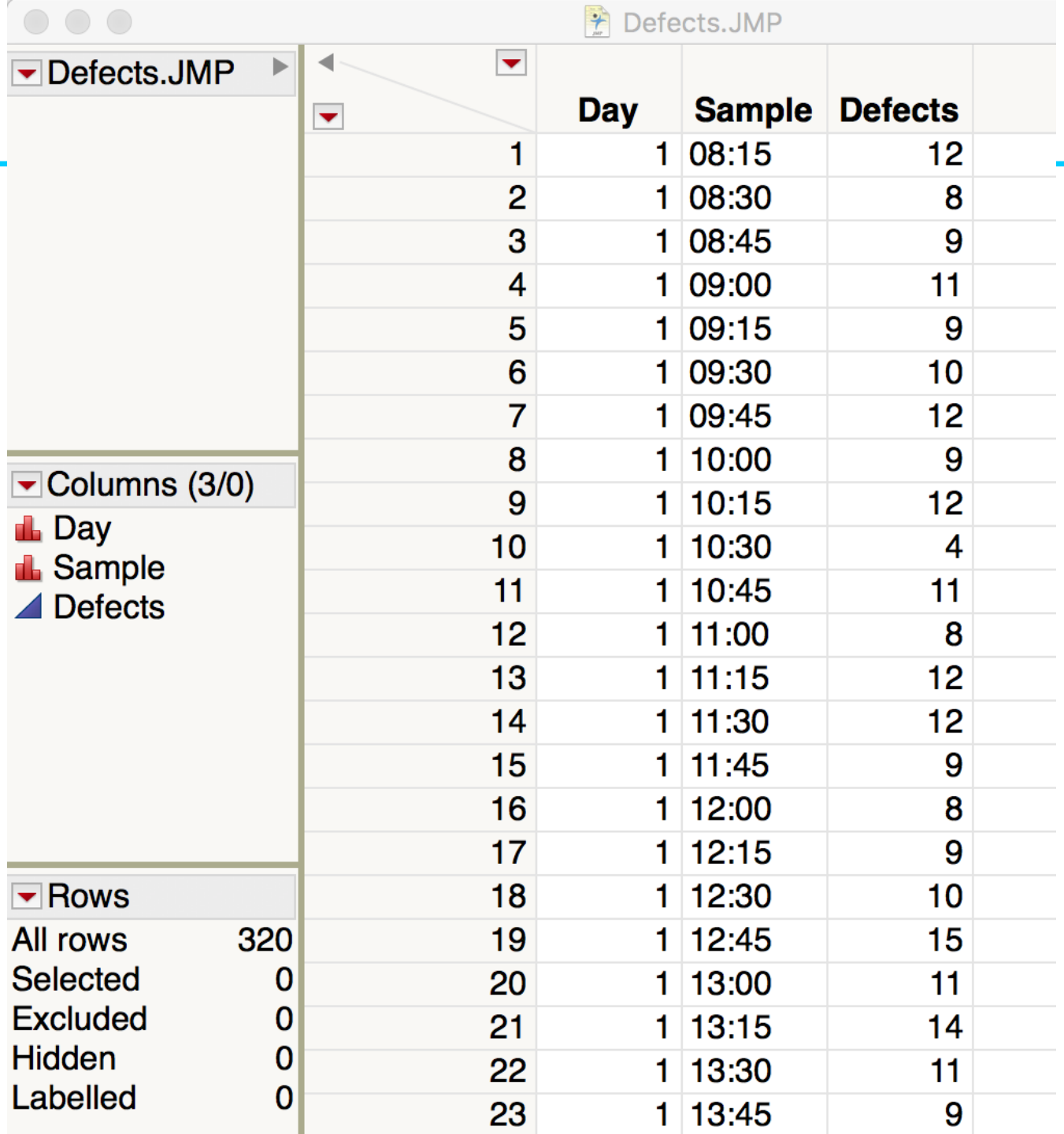
- Statistics: the science of collecting, organizing, and interpreting *data.*

- 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, …)

Defect Sampling

- Electronic manufacturer
- Randomly inspect production line every 15 minutes
  - 8-hour shift, 10 working days
- Variables
  - Day: day of the test 1-10
  - Sample: sampling time
  - Defects: number of minor defects on the sampled item


Defects.JMP
Defects.JMP
Columns (3/0)
Day
Sample
Defects
Rows
All rows 320
Selected 0
Excluded 0
Hidden 0
Labelled 0
Day Sample Defects
1 1 08:15 12
2 1 08:30 8
3 1 08:45 9
4 1 09:00 11
5 1 09:15 9
6 1 09:30 10
7 1 09:45 12
8 1 10:00 9
9 1 10:15 12
10 1 10:30 4
11 1 10:45 11
12 1 11:00 8
13 1 11:15 12
14 1 11:30 12
15 1 11:45 9
16 1 12:00 8
17 1 12:15 9
18 1 12:30 10
19 1 12:45 15
20 1 13:00 11
21 1 13:15 14
22 1 13:30 11
23 1 13:45 9

Data Table

- The *rows* in the table go by various names such as observations, cases, or even subjects.
  - The table collects values of various attributes of the observations.

- The *columns* in the table are called "variables."
  - Each column holds the values of some attribute of the observations that comprise the rows of the table.
  - Type
    - Quantitative or numerical variable  : Defects
      - Continuous or discrete
    - Categorical variable: Day, Sample
      - Nominal  or ordinal 

Fundamental Concepts

- Population: the entire group of individuals that we want information about.
  - All manufactured items
- Sample: a part of the population that we actually examine in order to gather information.
  - Sampled items
- Sample size: number of individuals in a sample.
  - 320
- Statistical inference: make inference about a population based on information in a sample.
  - Based on the data in the sample, to study characteristics of the population.

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.

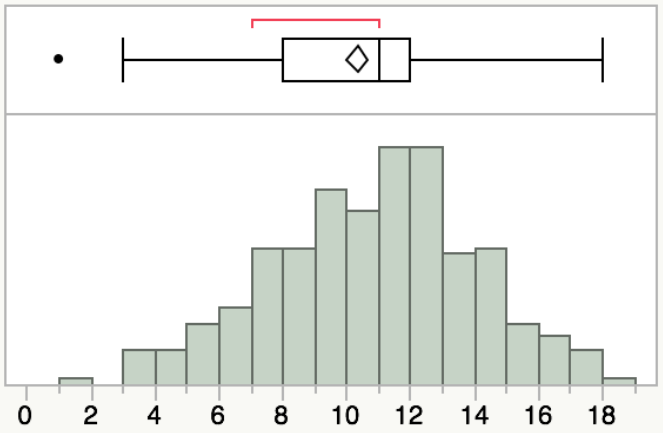
  - Relevance for decision making
    - Summarization can be very useful when using data to form decisions.
    - Avoid distraction from extraneous features of data.

Graphical and Numerical Summaries

- Graphical and numerical summaries of the 320 observations (rows) of the variable Defects (a column).

Distributions

Defects



Quantiles

100.0%	maximum	18
99.5%		17.395
97.5%		16
90.0%		14
75.0%	quartile	12
50.0%	median	11
25.0%	quartile	8
10.0%		6
2.5%		4
0.5%		2.21
0.0%	minimum	1

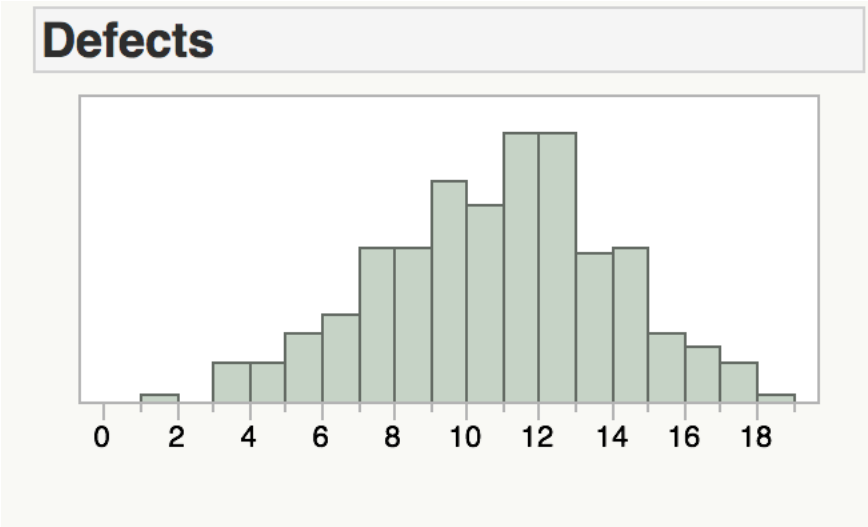
Summary Statistics

Mean	10.325
Std Dev	3.1722741
Std Err Mean	0.1773355
Upper 95% Mean	10.673895
Lower 95% Mean	9.9761051
N	320

- Graphical summaries: histogram, boxplot
- Numerical summaries: mean, median, quantiles, …

Histogram

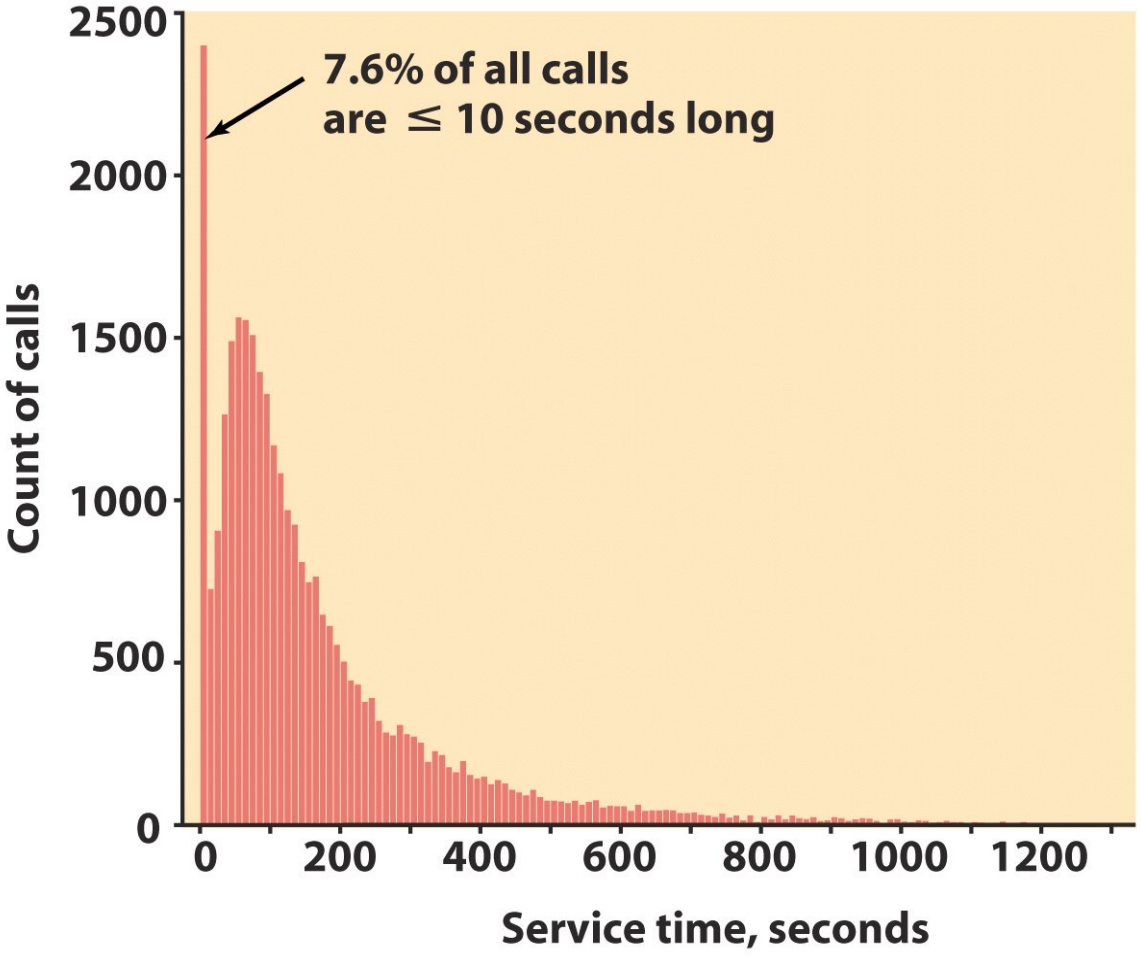
- A histogram is a visual display that reveals the location, dispersion, and shape of the data distribution. This distribution is vaguely *symmetric*, with "tails" reaching to the left and right away from the center.



- Shape: uni-modal (with one mode), symmetric, left-skewed or right-skewed.
  - Mode: the most frequent value in the data.

- When a distribution is bell-shaped and symmetric we call it "approximately normally distributed".

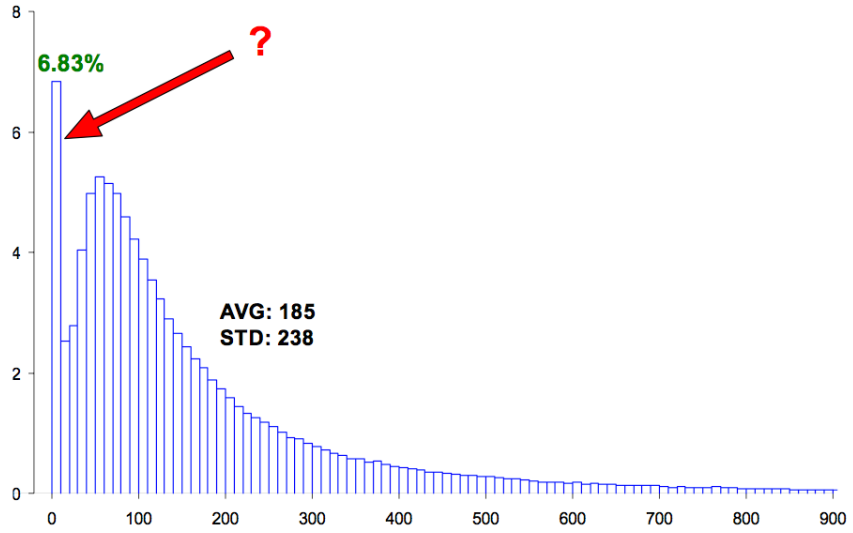
Customer Service Call Center Puzzle



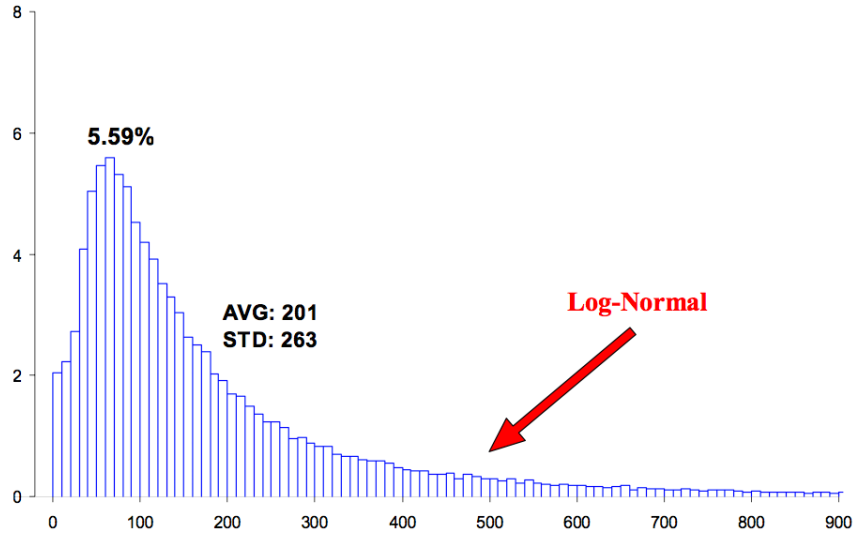
Service time: conversation time between a customer and a service rep

Call Center Service Time

January-October



November-December

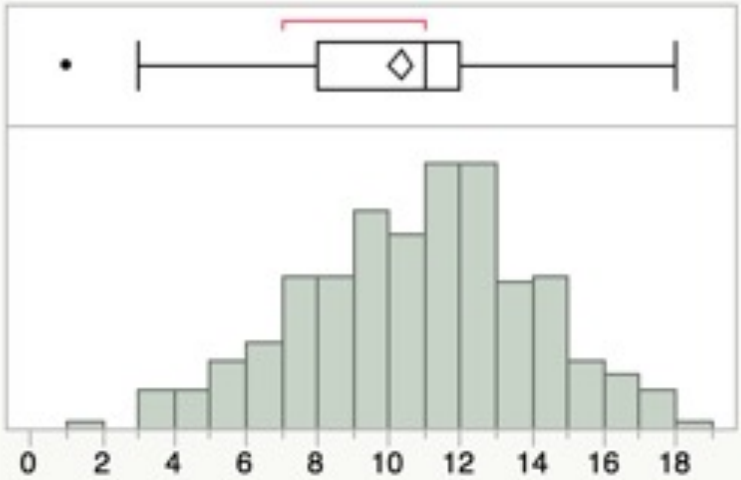


Log-normal, the log of which is normal

Defects Sampling

Distributions

Defects



Quantiles

100.0%	maximum	18
99.5%		17.395
97.5%		16
90.0%		14
75.0%	quartile	12
50.0%	median	11
25.0%	quartile	8
10.0%		6
2.5%		4
0.5%		2.21
0.0%	minimum	1

Summary Statistics

Mean	10.325
Std Dev	3.1722741
Std Err Mean	0.1773355
Upper 95% Mean	10.673895
Lower 95% Mean	9.9761051
N	320

Center?

Variability, deviation from the center?

Sample Mean $\bar{X}$

- We have observations $X_1, X_2, \ldots, X_n$
- The (sample) mean, $\bar{X}$, is average of the data:

$$\bar{X} = \frac{X_1 + X_2 + \ \ldots, + X_n}{n}$$

- $\bar{X}$: measure of the center of the data, a ``typical value"

Sample Variance S^2

- Deviation from mean: the difference between an observation and the sample mean:

$$X_i - \bar{X}$$

- Sample variance S^2: the average of the squares of the deviations from the mean

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n - 1}$$

- Sample standard deviation S: square root of S^2
  - the size of a "typical" deviation from the mean. I.e. volatility in finance.

Percentiles or Quantiles

- Percentiles (Quantiles) are derived from the ordered data values.

- The *p*th percentile is the value such that *p* percent of the observations fall at or below it.

- The three quartiles
  - The median M = 50th percentile
  - The first quartile Q1 = 25^{th} percentile
  - The third quartile Q3 = 75^{th} percentile

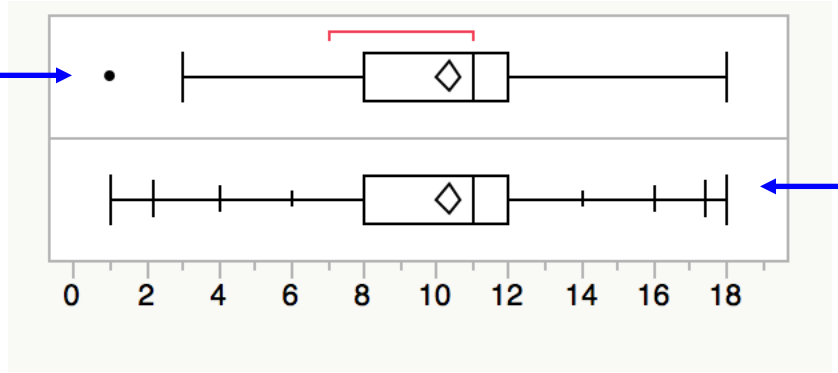
- How do the quartiles divide up the data?

Five-number Summary and Boxplots

- To get a quick summary of both center and spread, use the following five-number summary:

  Minimum Q1 *M* Q3 Maximum

- Boxplot: visual rep. of the five-number summary.
  - A central box spans the quartiles Q1 and Q3.
  - A line inside the box marks the median M.
  - Lines extend from the box out to minimum and maximum

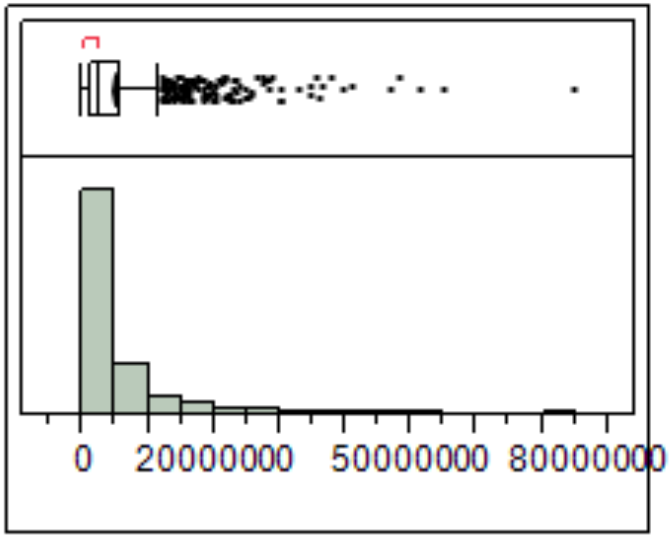


Executive Compensation Data

- 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

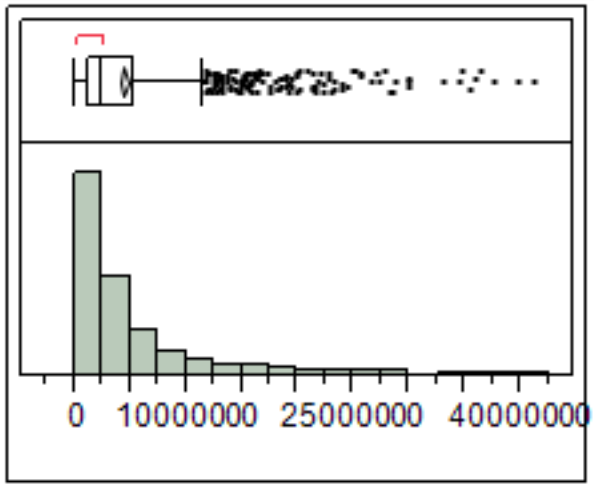
- Someone made nearly $75,000,000! Who? (label the data)

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

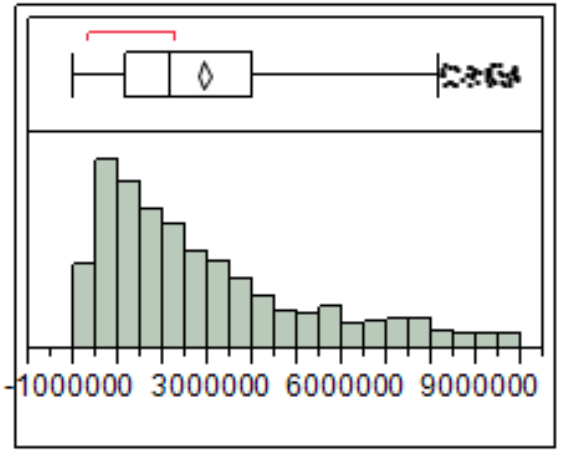
- That doesn't help much. The compensations of a relatively small number of executives continue to dominate the plot.

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

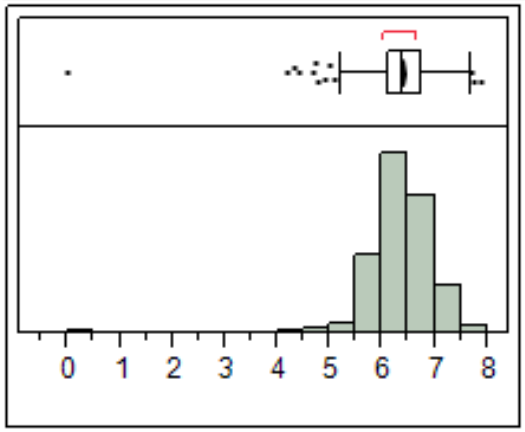
- For the CEO_Comp_03 data, let's create the new variable

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

- 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

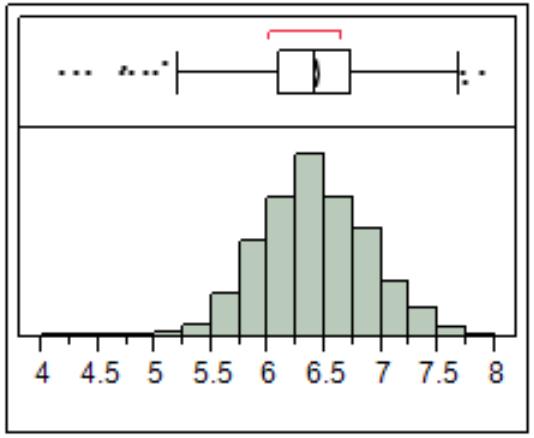
Summary Statistics	
Mean	6.4106808
Std Dev	0.5006985
Std Err Mean	0.0129539
Upper 95% Mean	6.4360906
Lower 95% Mean	6.385271
N	1494

- 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.

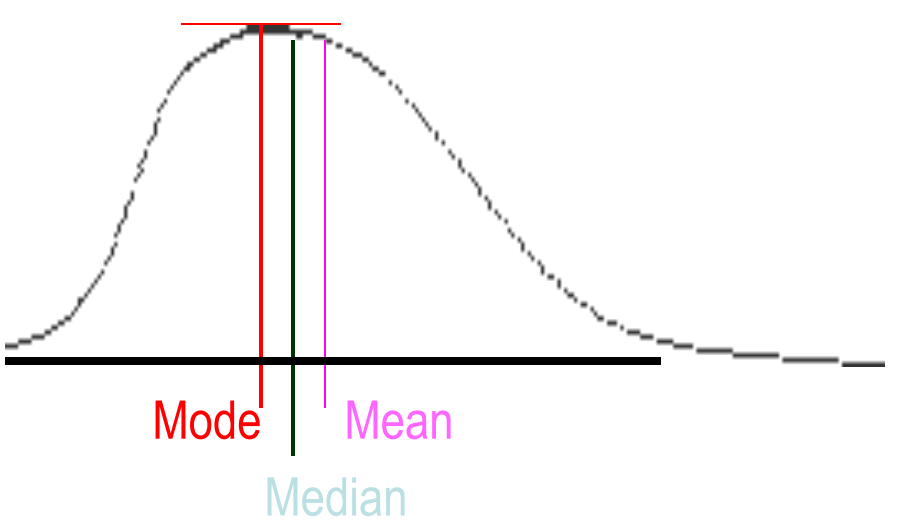
Theoretically

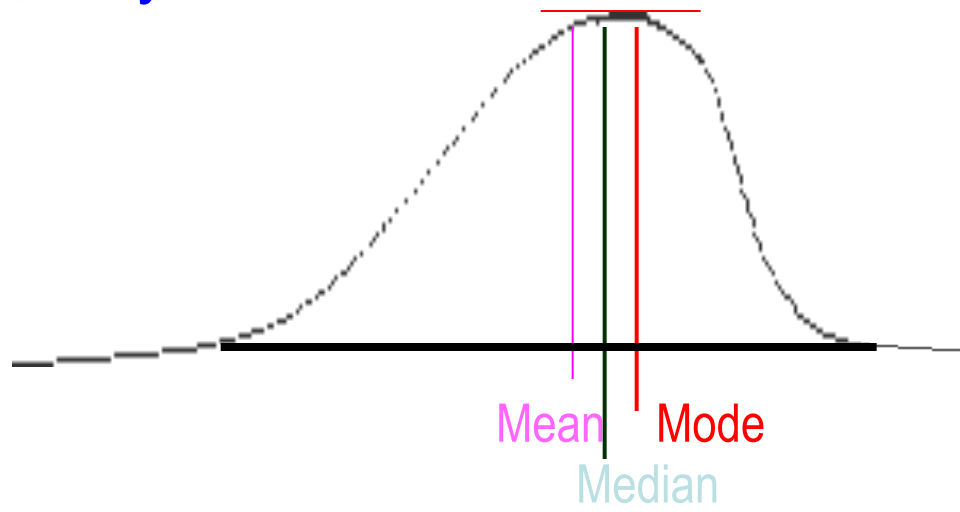
- Log transformations have the potential to pick up interesting relationships such as *diminishing returns to scale*.

Mean, Median and Mode

- If the distribution is exactly symmetric and unimodal, the mean, the median and the mode are the same.

- If the distribution is skewed, they differ.

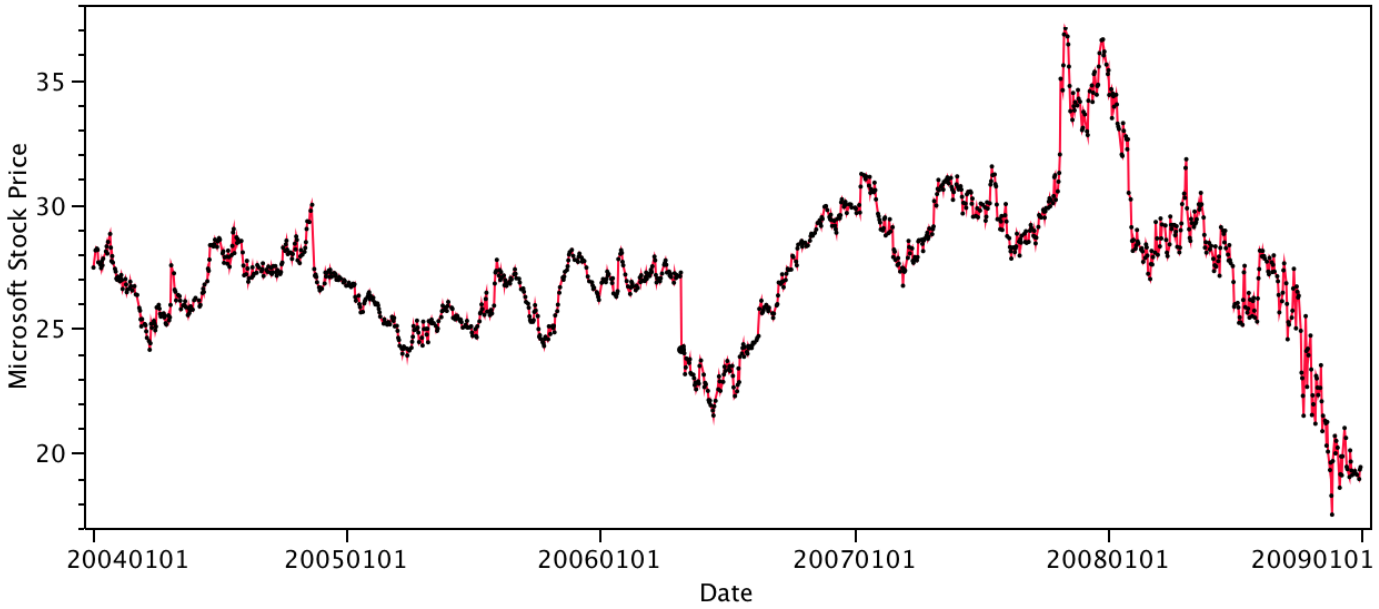




- Two stocks with mean return 0%, one symmetric, one right-skewed.
  - Which one is more likely to lose money?

Prices of Microsoft Stock

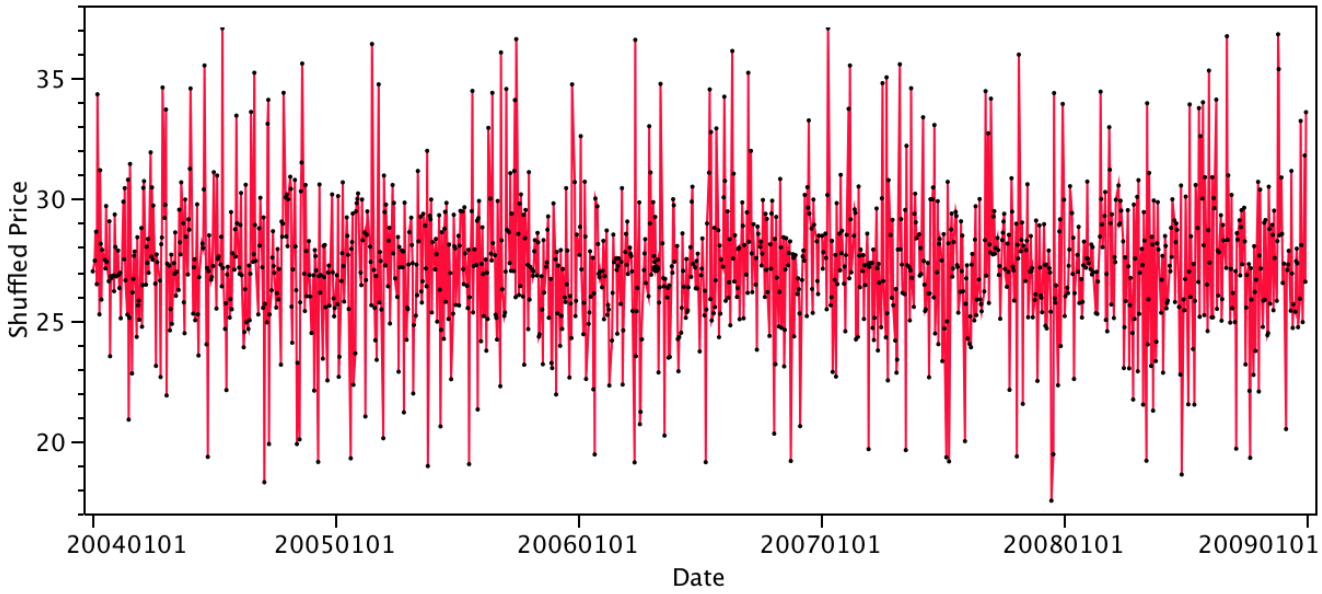
- Question: What is the risk associated with owning Microsoft stock?
- The file *Microsoft* contains daily closing prices from 2004 through 2008 on a share of Microsoft stock. Because the prices form a time series (*i.e.*, a sequence of observations that are ordered in time), we begin by plotting the prices over time.



- Note the meandering behavior of this time series.

Exhibit Sequential Dependence

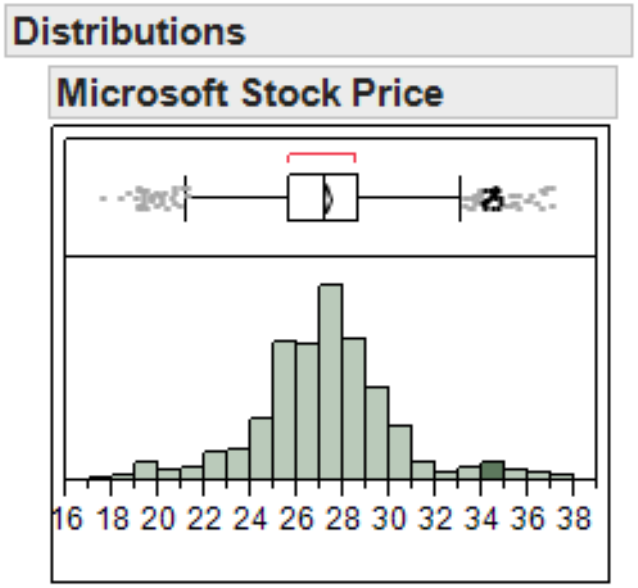
- Successive draws do not stray far from each other. This is a result of sequential *dependence*.
- The price each day *depends* on the price of the previous day. To show what these data would look like without sequential dependence, let's randomly shuffle the order of Price. A time series plot of Shuffled Price looks like



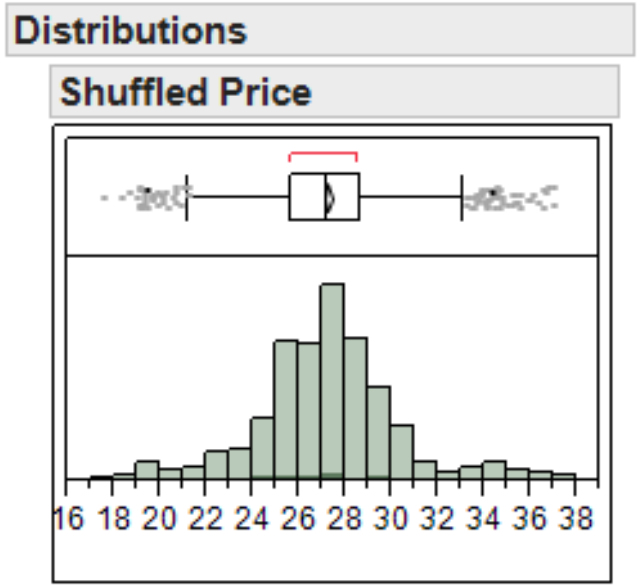
- No more meandering - a sequence of *independent* random draws from a population (i.e., a random sample).
  - Lack of meandering, stable level, stable variation

Drawbacks of Histogram

- Original Price



Shuffled Price



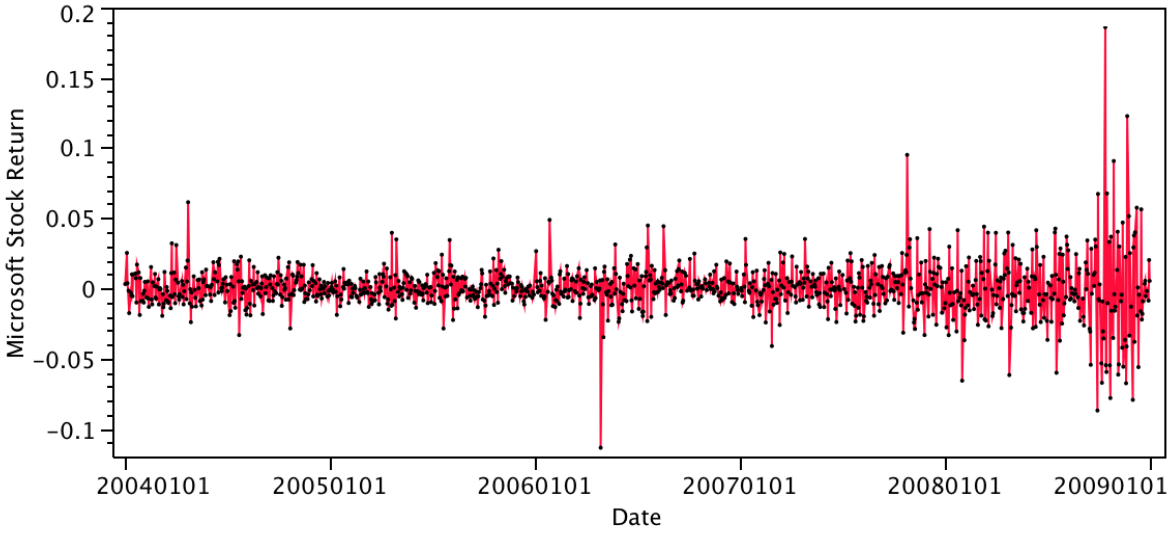
- Why identical?
- Histogram only reveals marginal information, not dependence.
  - Not useful for dependent time series
- Prediction
  - Histogram
  - Time series plot
  - Which one more precious?

Stock Return

- Letting p_t denote the price of Microsoft at time t, a different perspective is obtained by focusing on the successive daily relative changes or returns.

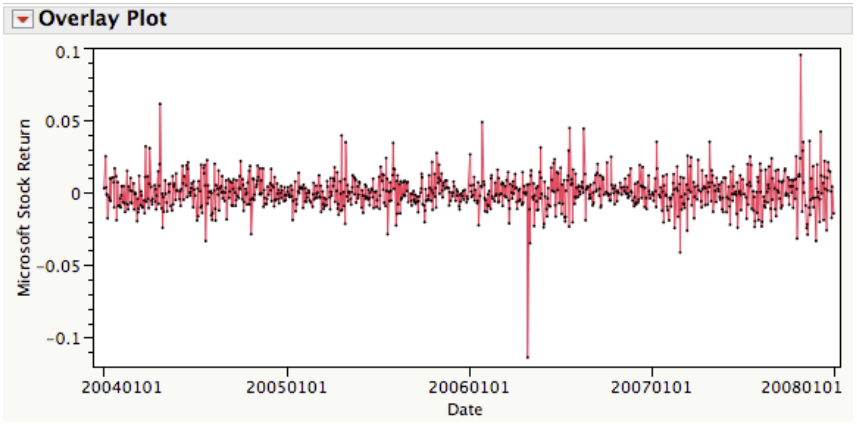
$$Return_t = \frac{p_t - p_{t-1}}{p_{t-1}}$$

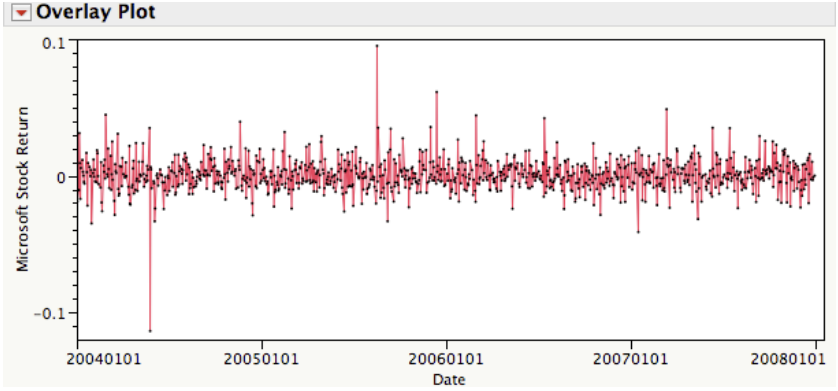
- The sequence plot of *Return* resembles a sequence of independent draws from a population. There's no evident meandering pattern.
- We instead see an increase in volatility (variation) in 2008 and several conspicuous outliers.



Stock Returns are Independent

- Let's focus for the moment on the period from 2004 through 2007, the period of stable variation. One of these plots shows the actual returns, and the other shows the returns after shuffling the order.
- Except for the outliers, can you tell which is which?



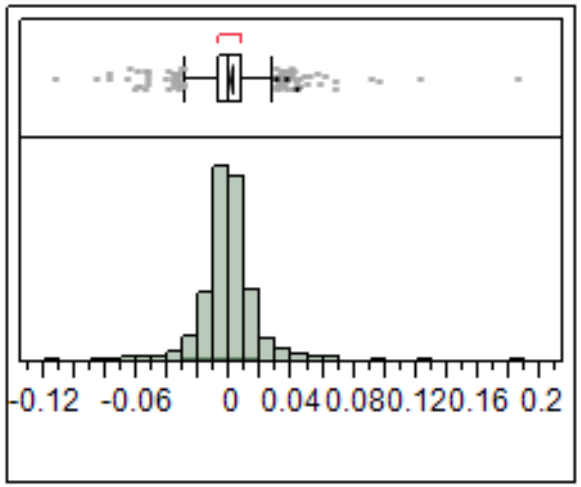


Histogram of Return

- Other than hiding the timing of outlying events, the histogram of this time series provides a good summary of the variation in the time series.

Distributions

Microsoft Stock Return



Quantiles		
100.0%	maximum	0.18605
99.5%		0.0654
97.5%		0.03506
90.0%		0.01651
75.0%	quartile	0.00704
50.0%	median	-0.0004
25.0%	quartile	-0.0075
10.0%		-0.0159
2.5%		-0.0335
0.5%		-0.067
0.0%	minimum	-0.1138

Summary Statistics	
Mean	0.0000113
Std Dev	0.0173987
Std Err Mean	0.0004903
Upper 95% Mean	0.0009733
Lower 95% Mean	-0.000951
N	1259

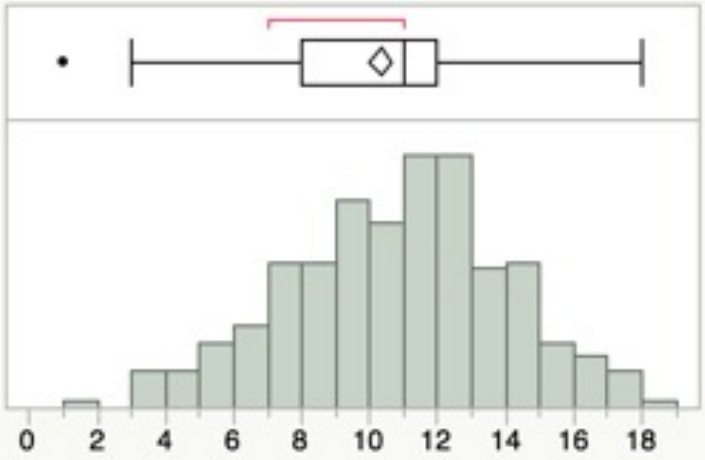
- For independent samples, the timing is not that important.
- The histogram of the returns during these years shows that the data have a “bell-shaped” distribution with a scattering of outliers.
- This bell-shaped tendency is what you should expect when sampling from a *normal population*.

Histogram of Defects Sampling

- This bell-shaped tendency is also evident in the histogram of number of defects and many other phenomena.

Distributions

Defects



Quantiles		
100.0%	maximum	18
99.5%		17.395
97.5%		16
90.0%		14
75.0%	quartile	12
50.0%	median	11
25.0%	quartile	8
10.0%		6
2.5%		4
0.5%		2.21
0.0%	minimum	1

Summary Statistics	
Mean	10.325
Std Dev	3.1722741
Std Err Mean	0.1773355
Upper 95% Mean	10.673895
Lower 95% Mean	9.9761051
N	320

- When the data is a random sample from a population, the shape of the histogram tends to mimic the shape of the population distribution.
  - Especially when the sample size is large

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. Let's see how well a normal distribution describes the returns on Microsoft during 2004 through 2007. Now that we are dealing with data, we use our sample mean and sample standard deviation notation. For this data, $\bar{x} \approx .00048$ and $s \approx .01199$ so that:

$$(\bar{x} - \mathrm{s}, \bar{x} + \mathrm{s}) = (-0.01151, 0.01247)$$

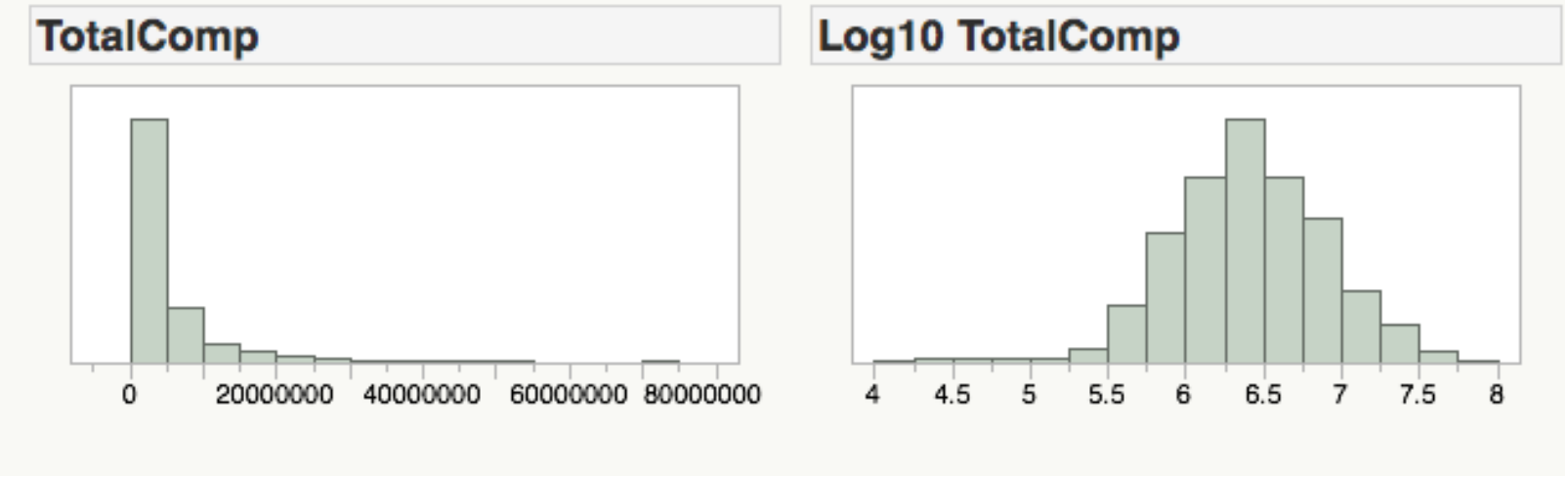
$$(\bar{x} - 2\mathrm{s}, \bar{x} + 2\mathrm{s}) = (-0.02350, 0.02446)$$

$$(\bar{x} - 3\mathrm{s}, \bar{x} + 3\mathrm{s}) = (-0.03549, 0.03645)$$

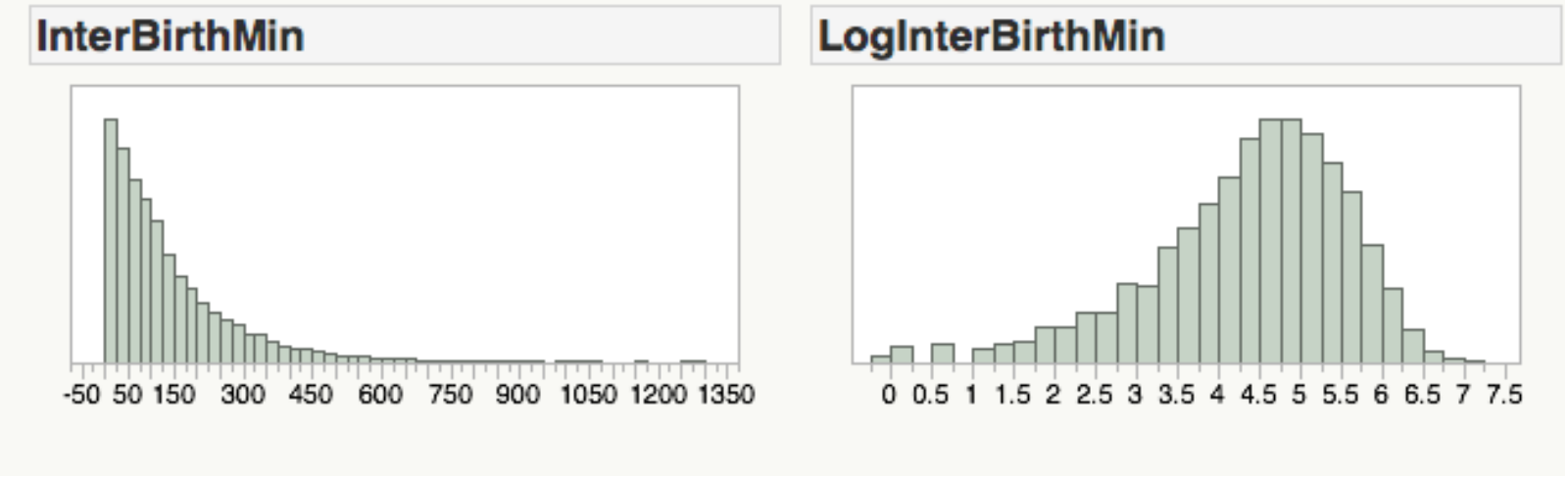
The Normal Quantile Plot

- 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=okjYjCISjOg

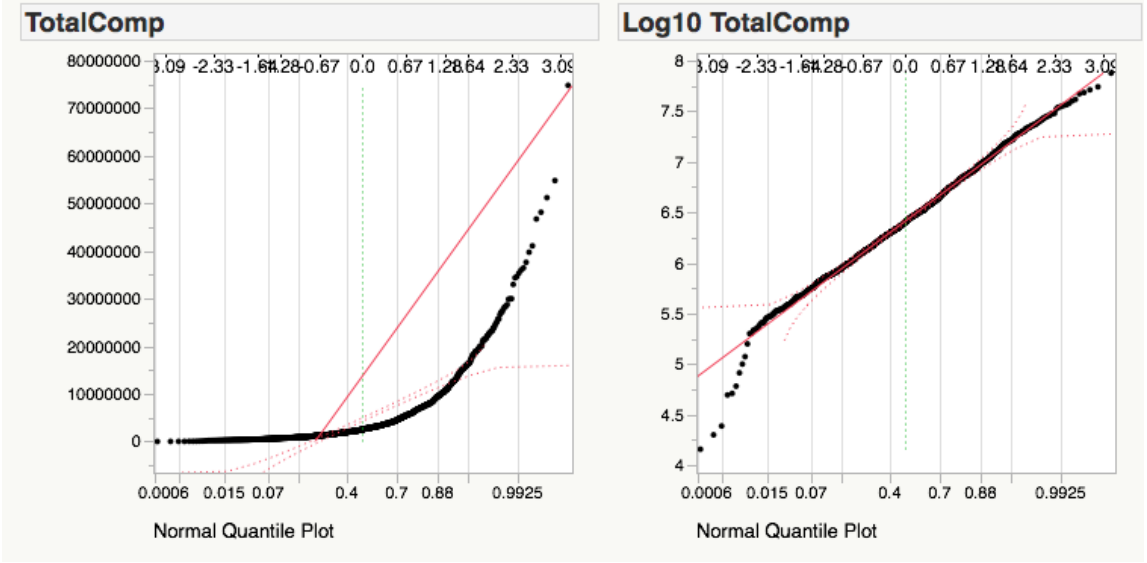
Histogram: Compensation, InterBirthMin

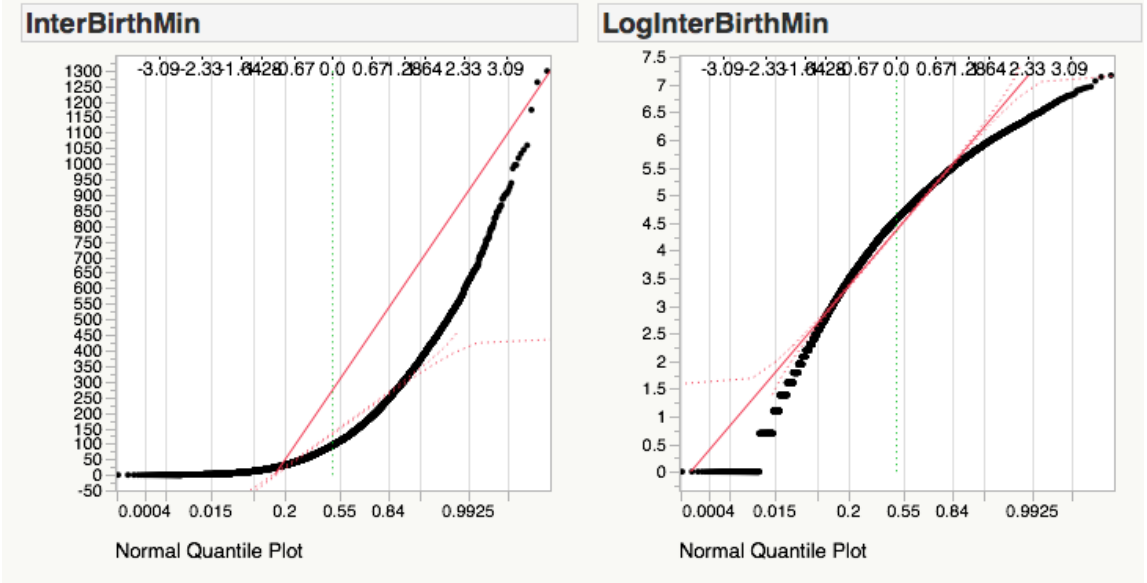


- Time (in minutes) between two consecutive newborns at a hospital (barring twins)



Normal Quantile Plot: Compensation, InterBirthMin

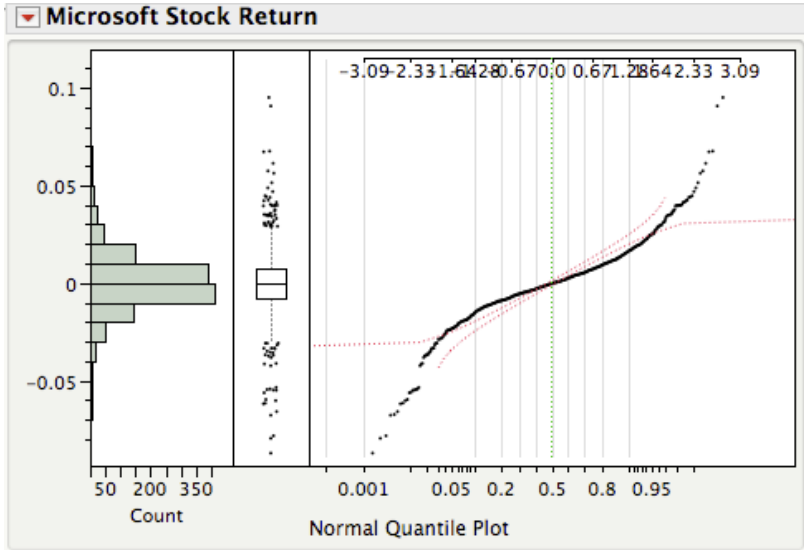


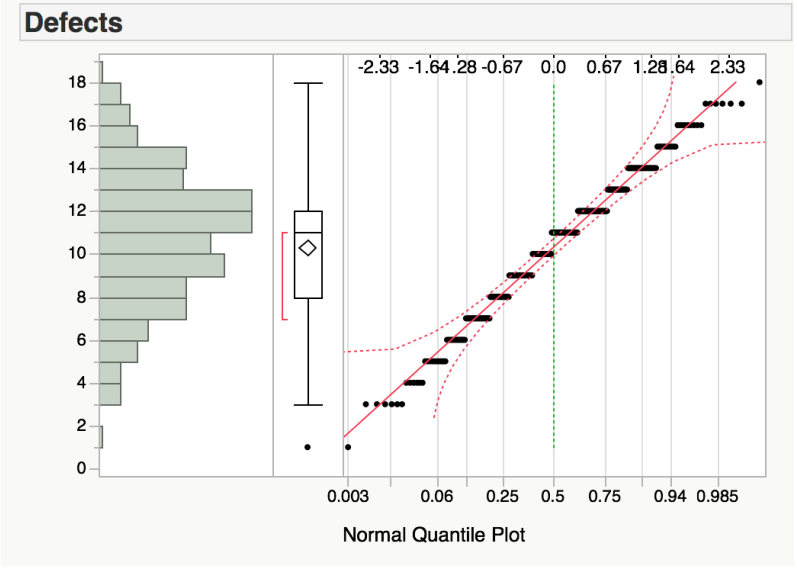


Use of Normal Quantile Plots

- 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



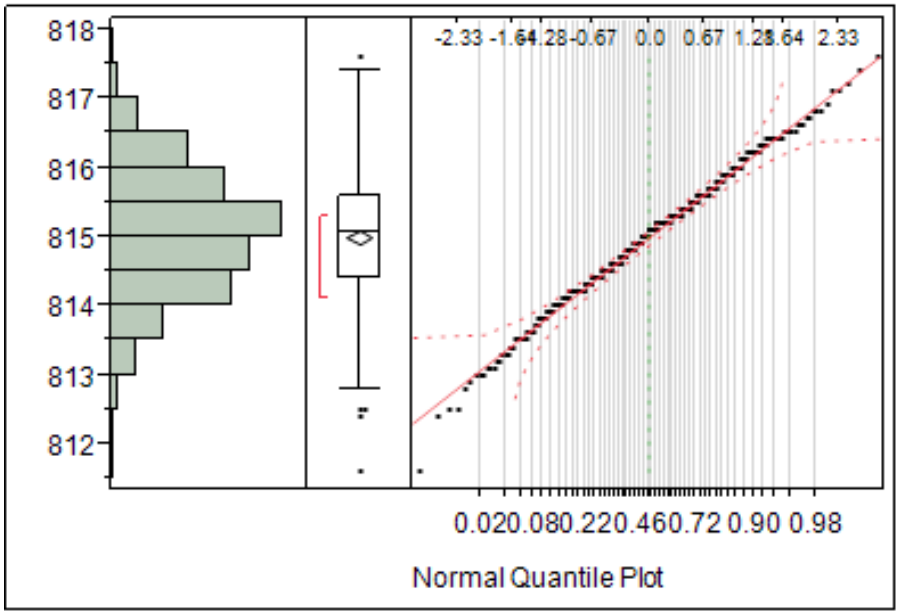


- What's causing those flat ``steps?"
- Key points
  - Many statistical inferences rely on normal distributions.
  - Use a normal quantile plot to check for normality rather than assume normality.

Motor Shaft Data

- Consider a manufacturing plant that produces motor shafts
- Quality control: inspect the diameters of the shafts produced
- Population: all the shafts produced and to be produced
- Parameter: mean diameter μ
- Data: *Shaft* contains diameters (in thousands of an inch) of 400 motor shafts produced here
  - Five observations were taken per day for 16 weeks

Example: Motor Shaft



Summary Statistics	
Mean	814.99175
Std Dev	0.9298486
Std Err Mean	0.0464924
Upper 95% Mean	815.08315
Lower 95% Mean	814.90035
N	400

Questions:

1) What is your best guess of the mean diameter of the motor shaft population?
2) Can you place a range of possible values for your best guess?
3) Suppose the manufacture claims the population mean is 810. Do the data provide strong evidence for that claim?

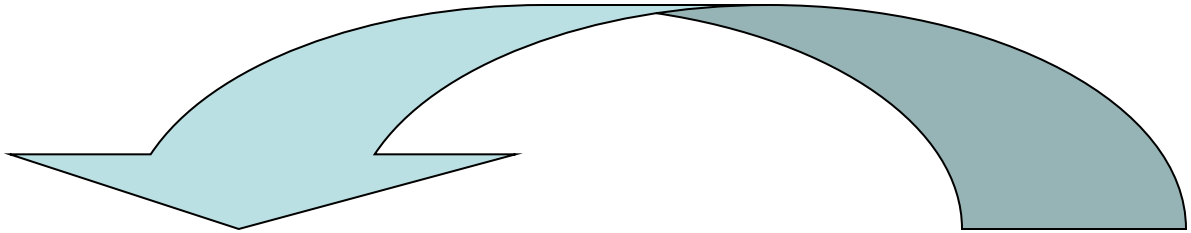
Guessing the Mean of a Population

- Population: all the shafts produced and to be produced
- Parameter: mean diameter μ
- How can we know what μ (the population mean) is?
- We can never know μ exactly, but a natural guess is the average of our data (sample): the diameters of the 400 motor shafts produced, $\bar{x}$ or the sample mean.
- However, our guess is almost certainly not going to be exactly correct, so a critical question is

``How sure can we be of our guess?"

The Population-Sample Paradigm

Statistical Inference



Population

All shafts

Sample

400 shafts measured

Fixed

- Population summary
  - Pop. Mean (μ)
  - Pop. SD (σ)
  - Pop. Proportion (p)
- Parameters

- Sample summary
  - Sample Mean ($\bar{x}$)
  - Sample SD (s)
  - Sample Proportion ($\hat{p}$)
- Statistics

Random

Sample statistics are guesses for population parameters, but they are unlikely to be exactly correct, so we need to quantify the ``error".

We will start by looking at μ and $\bar{x}$.

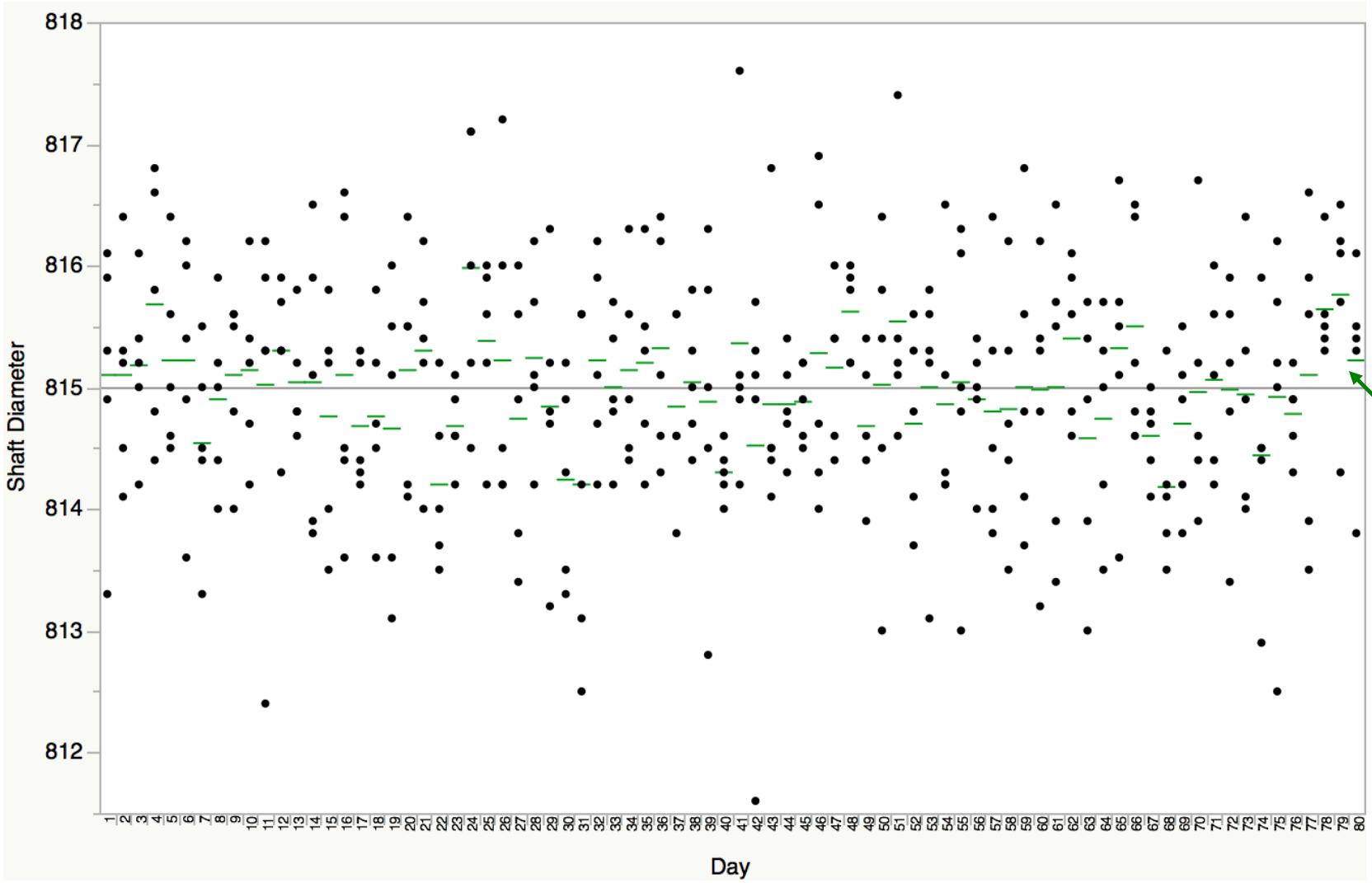
IID Sampling

- We are especially interested in simple random samples obtained by sampling from a population of a conceptually infinite size.

- Real populations have finite size, but often reasonable to treat them as infinite when the size of the sample is small relative to the size of the population.

- In this case, the data x_1, x_2, ..., x_n can be thought of as n independent draws from the same population, called *iid* samples:
  - *iid* = independent and identically distributed

Sample Means are Random!



Properties of Sample Mean $\bar{X}$

1. $E(\bar{X}) = \mu$ i.e. will average out to the population mean (the thing we want to estimate).
   - Unbiasedness

2. $SD(\bar{X}) = SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ so $\bar{X}$ gets more accurate for larger sample sizes.
   - Most efficient estimator (i.e. with the smallest variance out of all the possible linear estimators)

3. $\bar{X}$ is the least squares estimator

Mean and Variance of $\bar{X}$

Given independence,

- $E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n)$
- $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var(X_1) + \cdots + \frac{1}{n^2}Var(X_n)$

Given random sampling, we have

$$E(X_i) = \mu$$
$$Var(X_i) = \sigma^2$$

Mean and Variance of $\bar{X}$

Hence

- $E(\bar{X}) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) = \mu$
- $Var(\bar{X}) = \frac{1}{n^2}Var(X_1) + \cdots + \frac{1}{n^2}Var(X_n)$

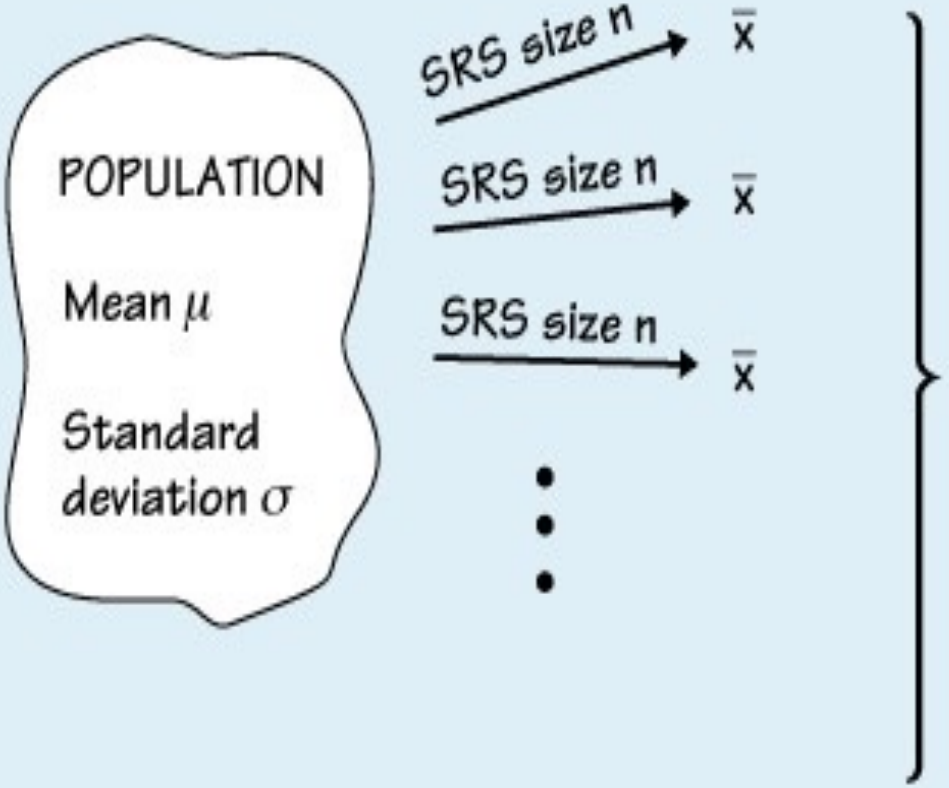
$$= \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2$$

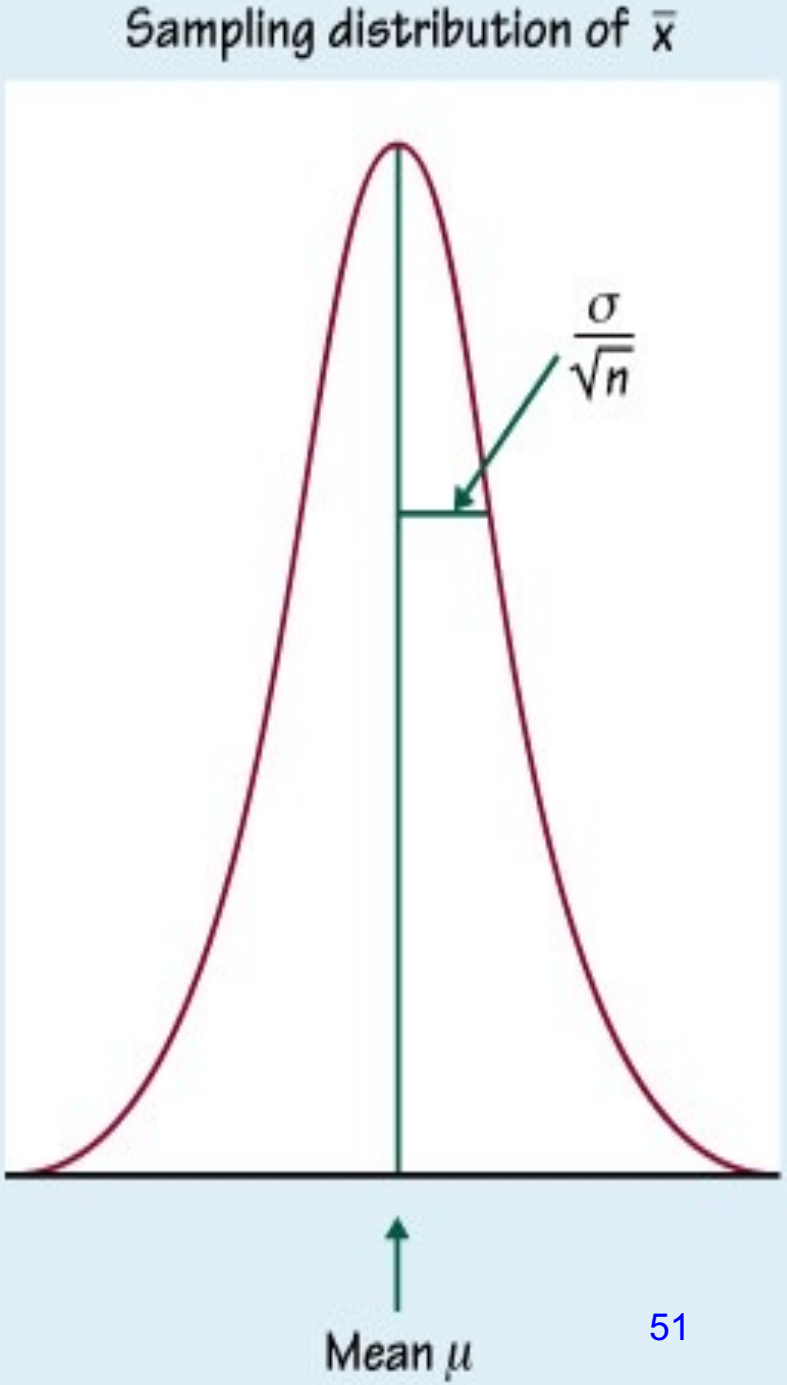
$$= n \times \frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

- $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Sampling Distribution of $\bar{X}$

- Sample means are random variables.
- Hence, they follow a certain distribution, i.e. sampling distribution of $\bar{X}$.
- This distribution has a mean and a variance.
- The mean and the variance turn out to be very important.
- So is the shape of the distribution.





SRS: simple random sample

- every possible subset of a given size has an equal chance of being drawn

Sampling Distribution of $\overline{X}$

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\overline{x}$ is approximately normal:

$$\overline{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

If the population distribution is normal, then the sampling dist of the sample mean is exactly normal;

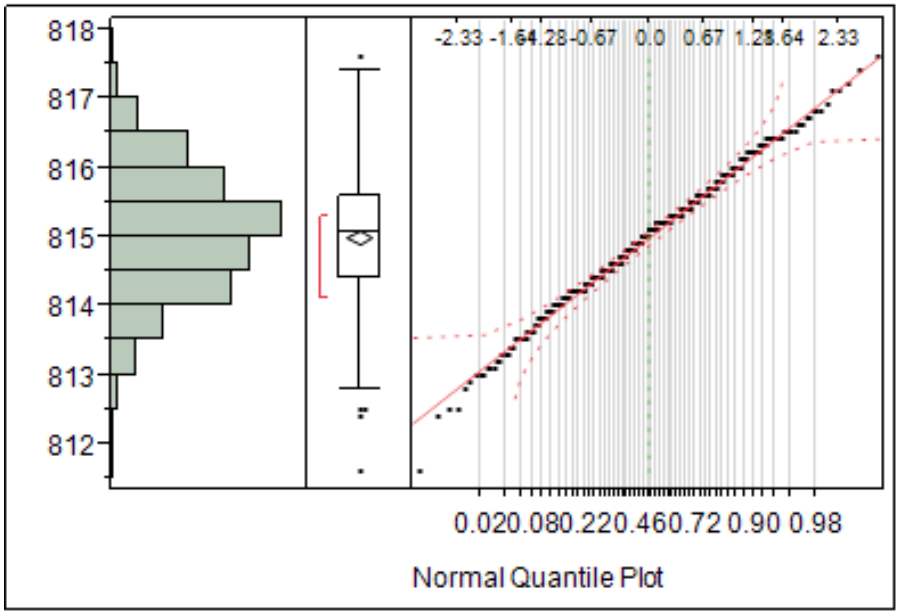
Otherwise, the sample mean is approximately normal.

The benefit of being normal: Empirical Rule.

Standard Error of Sample Mean $\bar{X}$

- The sample mean $\bar{X}$ is random, with the sampling distribution as its distribution.
- The distribution variation measured by $\sigma/\sqrt{n}$, reflecting two factors
  - Population variability
  - Sample size
- In practice, use s to estimate σ
- (Estimated) standard error of the mean: $s/\sqrt{n}$
  - An estimate of the standard deviation of the sampling dist.
  - Used in confidence interval, hypothesis testing

Example: Motor Shaft



Summary Statistics	
Mean	814.99175
Std Dev	0.9298486
Std Err Mean	0.0464924
Upper 95% Mean	815.08315
Lower 95% Mean	814.90035
N	400

Standard error mean: 0.0465.

What does it tell us?

- If we were to repeat the experiment many times and compute a new sample mean each time, we would expect it differ from the population mean by 0.0465 on average.

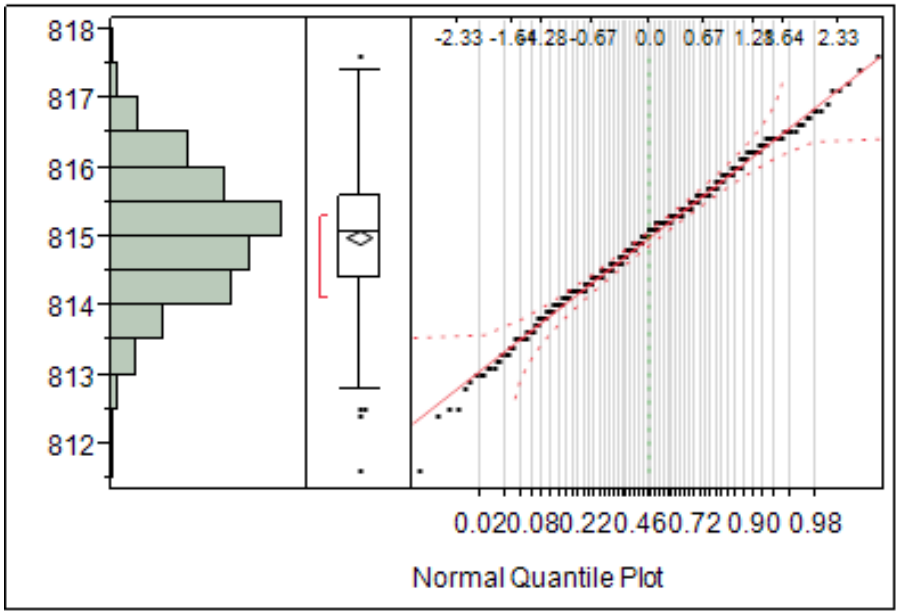
Summary of $\bar{X}$

1. $\bar{X}$ is a random variable (each time we take a random sample from a population and compute the mean we will get a different answer).
2. $E(\bar{X}) = \mu$ i.e. will average out to the population mean (the thing we want to estimate).
3. $SD(\bar{X}) = SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ so $\bar{X}$ gets more accurate for larger sample sizes.
4. The CLT tells us that $\bar{X}$ will have approximately a normal distribution which allows us to compute

$$P\left(\mu - zSE(\bar{X}) \leq \bar{X} \leq \mu + zSE(\bar{X})\right)$$

for any value of z.

Example: Motor Shaft



Summary Statistics	
Mean	814.99175
Std Dev	0.9298486
Std Err Mean	0.0464924
Upper 95% Mean	815.08315
Lower 95% Mean	814.90035
N	400

Questions:

1) What is your best guess of the mean diameter of the motor shaft population?
2) Can you place a range of possible values for your best guess?
3) Suppose the manufacture claims the population mean is 810. Do the data provide strong evidence for that claim?

Point Estimation

- A point estimator draws inference about a population by estimating the value of an unknown parameter using a single value or a point.

- For example, sample mean estimates pop. mean.
  - The mean diameter: 814.99175

- Drawbacks:
  - How reliable is our estimate? – it is bound to change for another sample

  - How close is the estimate from the true parameter?

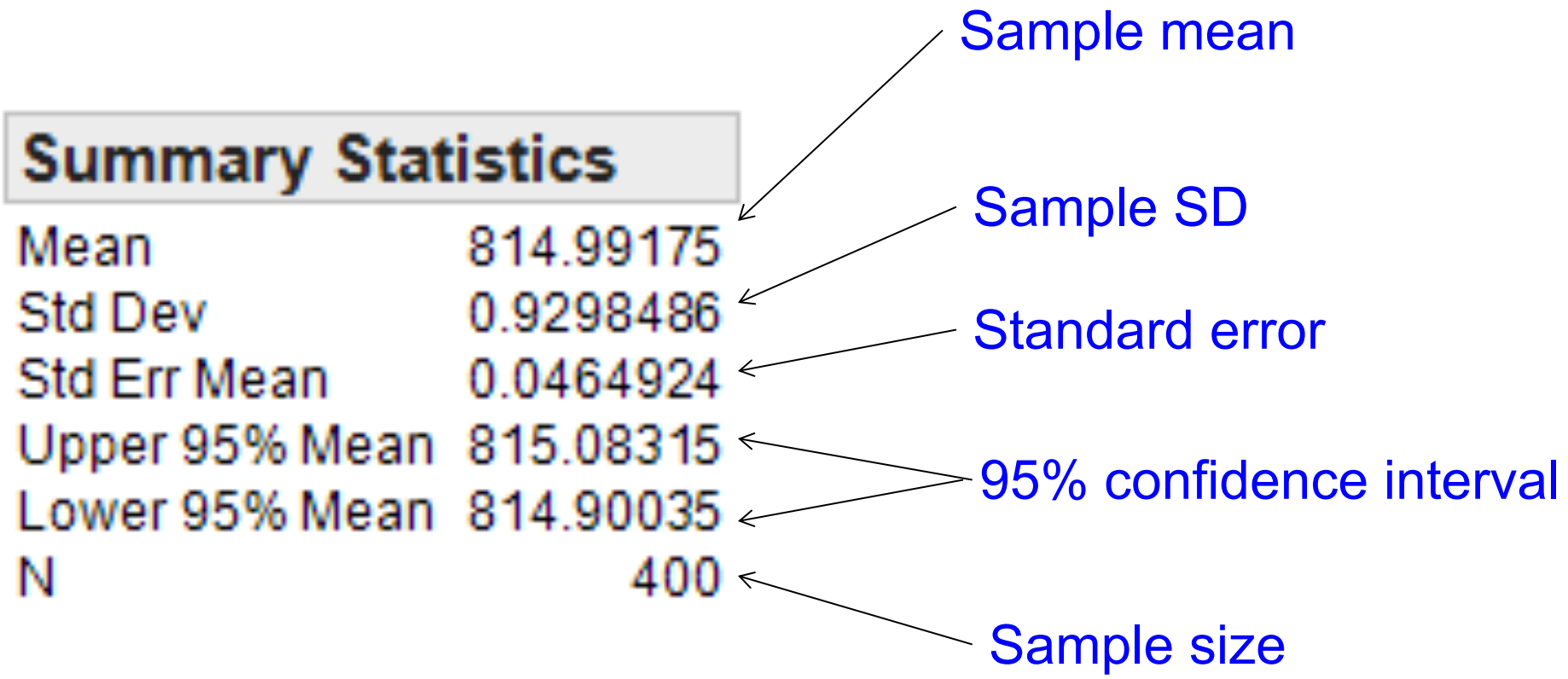
Confidence Interval

- Instead of giving just one guess, let's place a range of possible values on our best guess for the population mean (parameter)

- 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."

Example: Motor Shaft



Computing the Interval

- The CLT suggests,

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

- Alternatively,

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

- 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

- 95% confidence intervals are common, but we can compute any level of confidence that we like.
- In general, our 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 commonly used values of 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

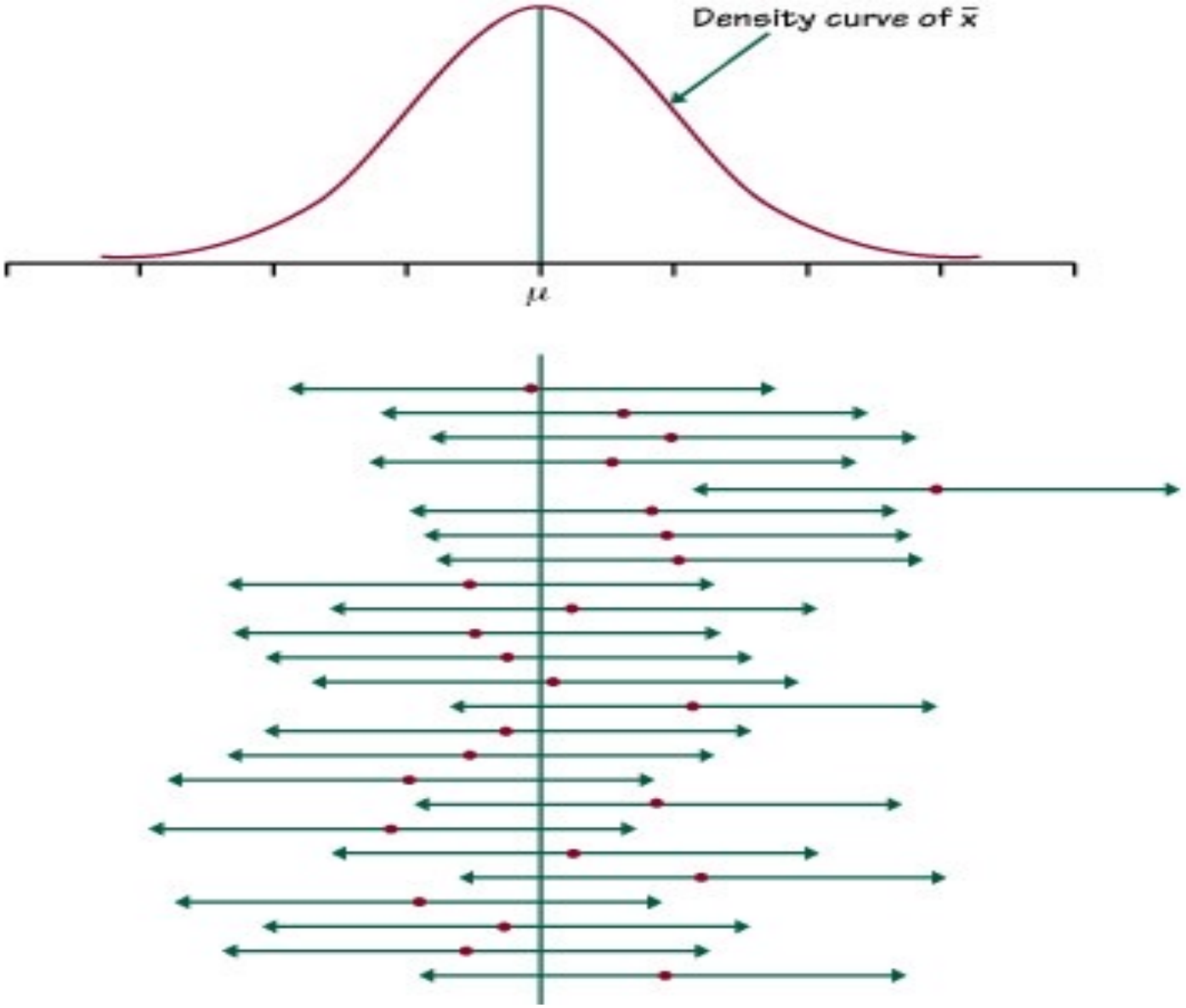
Confidence Intervals: σ Unknown

- 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.

How to Understand Confidence Intervals?



How to Understand Confidence Intervals?

- A 95% confidence interval (CI) means that, consider all possible samples, the obtained CIs cover the true value 95% of the time.
- Consider 100 samples.
- For each sample, a 95% CI can be derived.
- Approximately there will be 95 CIs that will cover the true mean.

Case: 2-04PriceQuotes

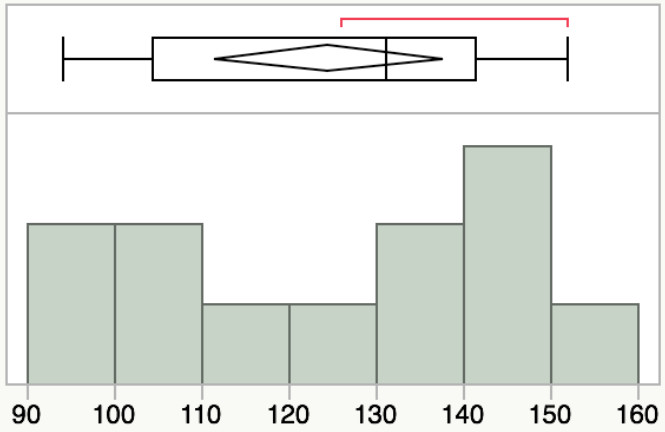
- A manufacturer supplies price quotes when receiving an order.
- Quotes are given by pricing experts on an order-by-order basis.
- Sales manager concerns about too much variability in the quoted prices, since the pricing process is too complex.
- Problem: is there variability between pricing experts?
- Data:
  - 12 randomly selected orders
  - 2 randomly selected pricing experts: Mary and Barry

Data: PriceQuotes

	Order Number	Barry Price	Mary Price
1	1	126	114
2	2	110	118
3	3	138	114
4	4	142	111
5	5	146	129
6	6	136	119
7	7	94	97
8	8	103	104
9	9	140	127
10	10	152	133
11	11	108	103
12	12	97	108

Price Summary

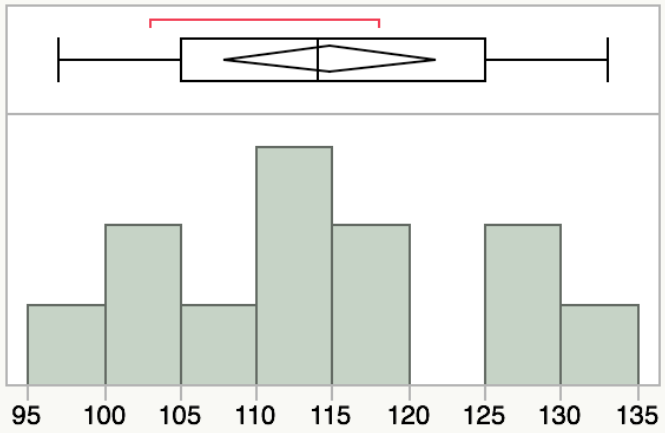
Barry Price



Summary Statistics

Mean	124.33333
Std Dev	20.698412
Std Err Mean	5.9751168
Upper 95% Mean	137.48448
Lower 95% Mean	111.18219
N	12

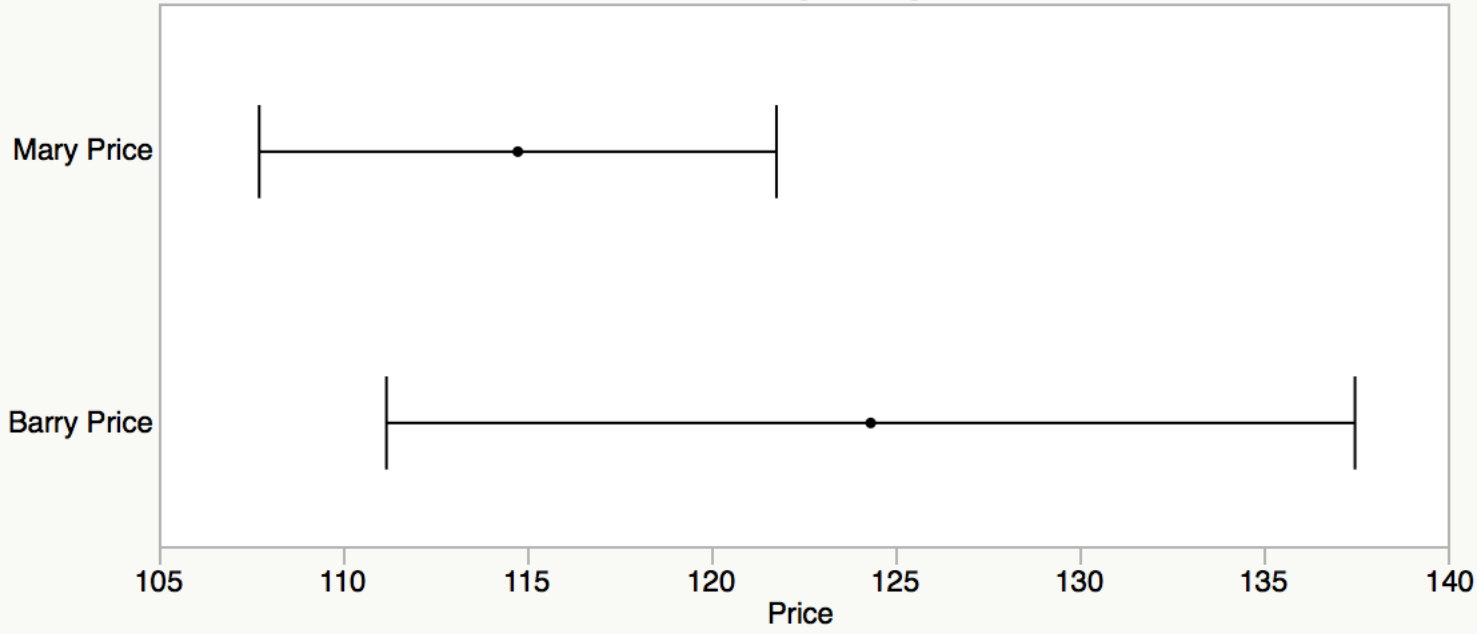
Mary Price



Summary Statistics

Mean	114.75
Std Dev	11.054616
Std Err Mean	3.1911929
Upper 95% Mean	121.77377
Lower 95% Mean	107.72623
N	12

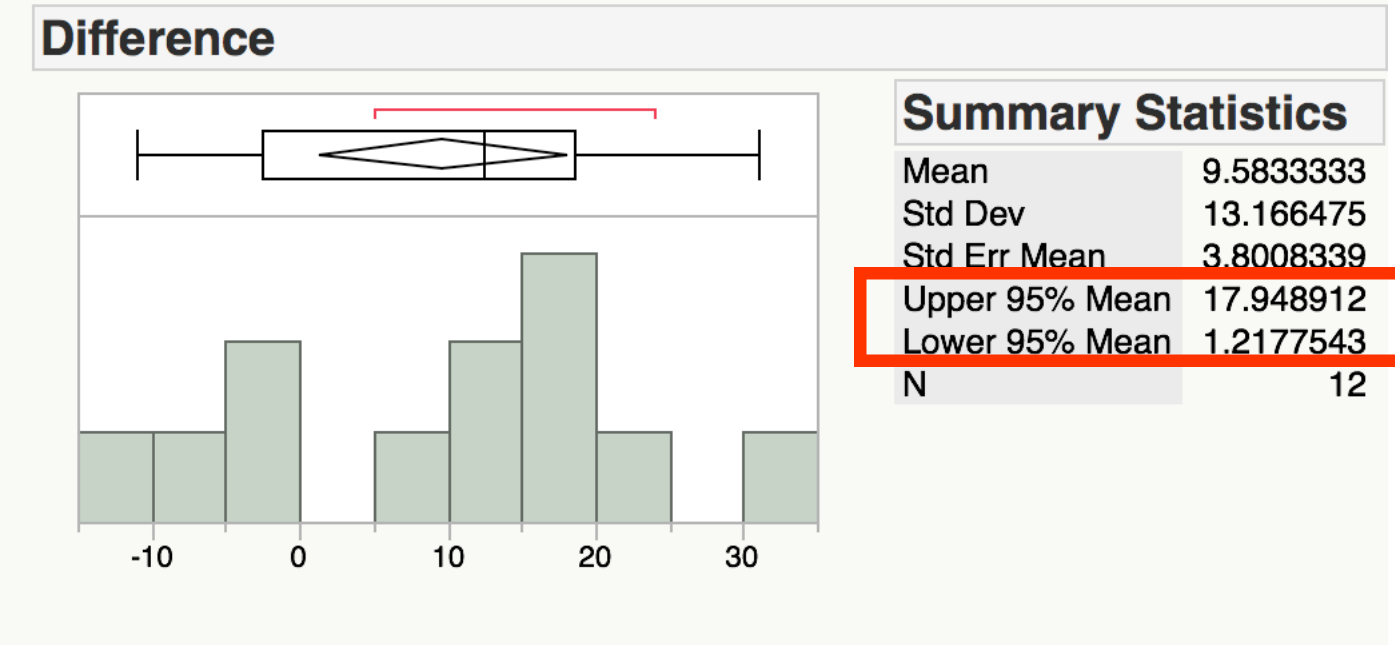
Comparison of Confidence Interval



- The two confidence intervals overlap – no statistical difference between Mary and Barry!

Look at Price Difference

- Difference=Barry Price – Mary Price



- Now the CI is above 0, meaning that Barry quotes higher than Mary on average!
  - Differencing removes order-to-order variation
  - More powerful

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.

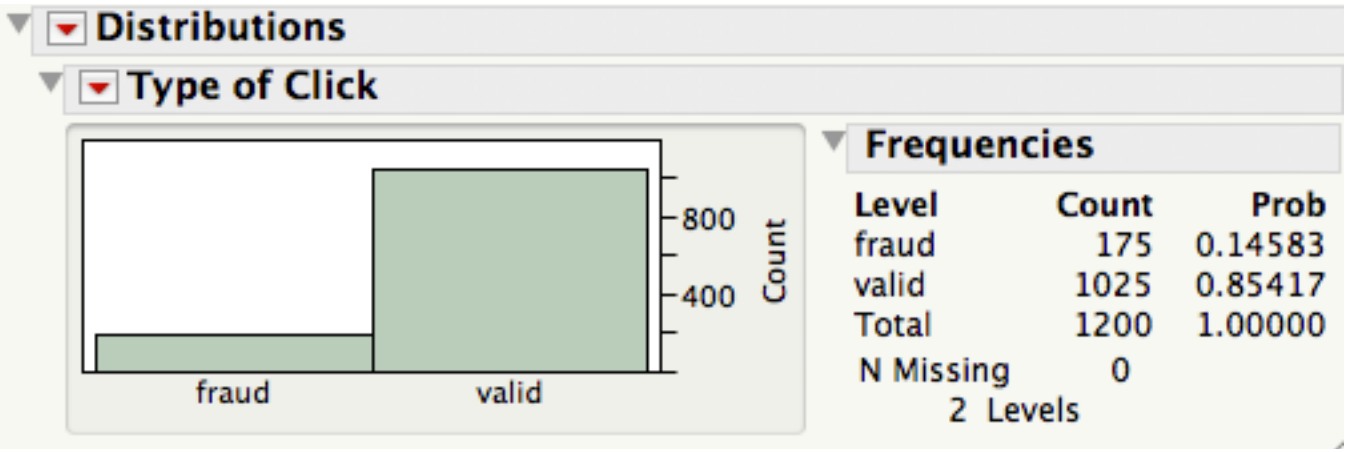
The on-line host has promised that no more than 10% of the clicks are imitations.

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

- 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.

Population Proportion

- Let p denote the proportion of "successes" in a population.
- A random sample of size n is selected
- X is the number of successes in the sample.

Binomial Random Variable

A Binomial random variable counts the number of successes in a specified number of Bernoulli trials.

It is identified by two parameters:

- *n*, the number of Bernoulli trials in the sequence
- *p*, the probability of success for each Bernoulli trial

Suppose $B_1, B_2, \ldots, B_n$ form a sequence of *n* Bernoulli trials, each with probability *p* of success. Then

$$X = B_1 + B_2 + \ldots + B_n$$

counts the number of successes in the *n* trials.

The random variable *X* is a *binomial random variable* with parameters *n* and *p*, denoted Bi(*n*, *p*).

Population Proportion

- Suppose *n* is small relative to the population size (less than its 1/20), *X* can be regarded as Binomial(n,p) with

$$E(X) = np, \text{ and } \sigma_X = \sqrt{np(1-p)}.$$

- A natural estimator of *p* is the sample proportion

$$\hat{p} = X / n,$$

which has the following property:

$$E(\hat{p}) = p, \text{ and } \sigma_{\hat{p}} = \sqrt{p(1-p)/n}.$$

Confidence Interval for Population Proportion

- Conditions: both $np \geq 10$ and $n(1-p) \geq 10$.
- Under the above conditions, $\hat{p}$ is approximately normally distributed.

- So, $$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard normal.

- A 100(1-α)% normal confidence interval for p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

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?

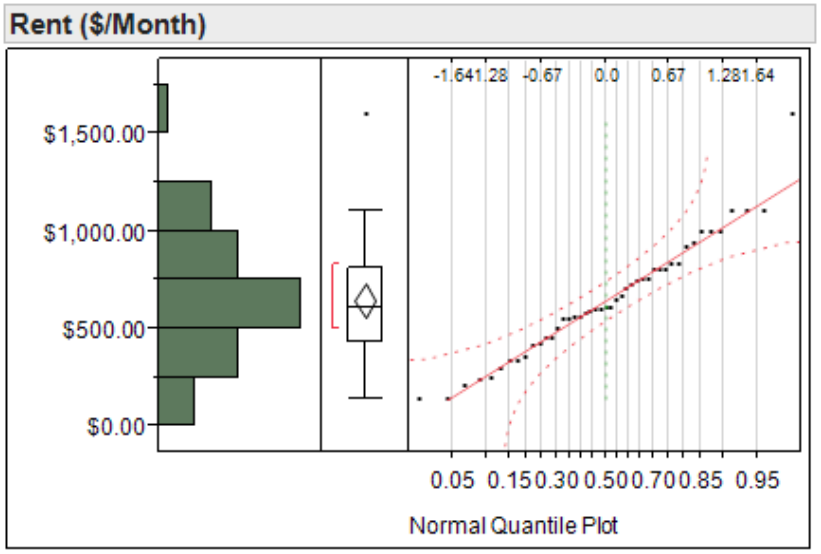
Realty Agency Expansion

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

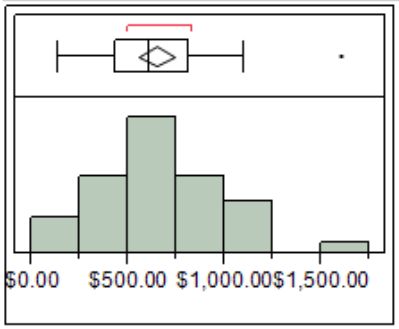
- Prove by contradiction
- Assume that the “opposite” is true
  - This “opposite” is your Null Hypothesis H_0
- Given that H_0 is true, calculate the probability for you to see what you ``saw” (i.e. the data)
  - This probability is called the ``p-value,” which is the probability of seeing ``data this 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)

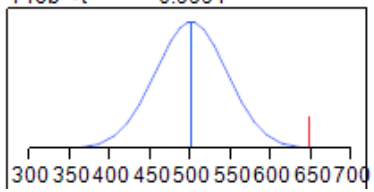


Quantiles		
100.0%	maximum	1600
99.5%		1600
97.5%		1525
90.0%		1040
75.0%	quartile	815
50.0%	median	610
25.0%	quartile	435
10.0%		246
2.5%		140
0.5%		140
0.0%	minimum	140

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

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

- 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 $H_a: p > 0.1$

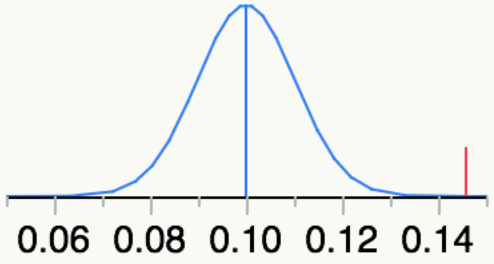
- Since the p-value is less than 0.05, we can reject the claim that $p \leq 0.1$

Test Mean

Hypothesized Value	0.1
Actual Estimate	0.14583
DF	1199
Std Dev	0.35309

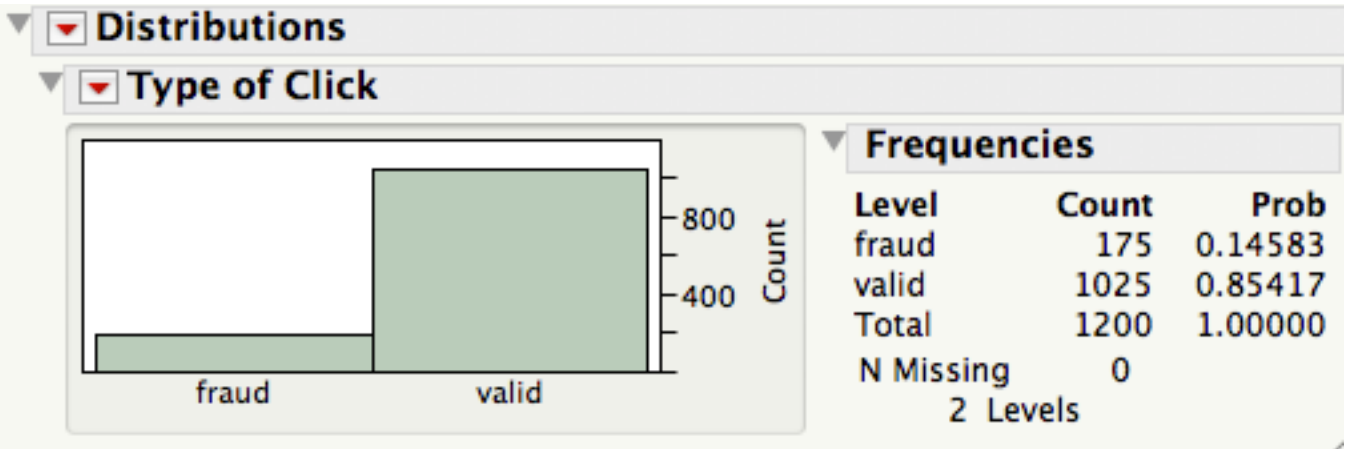
t Test

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

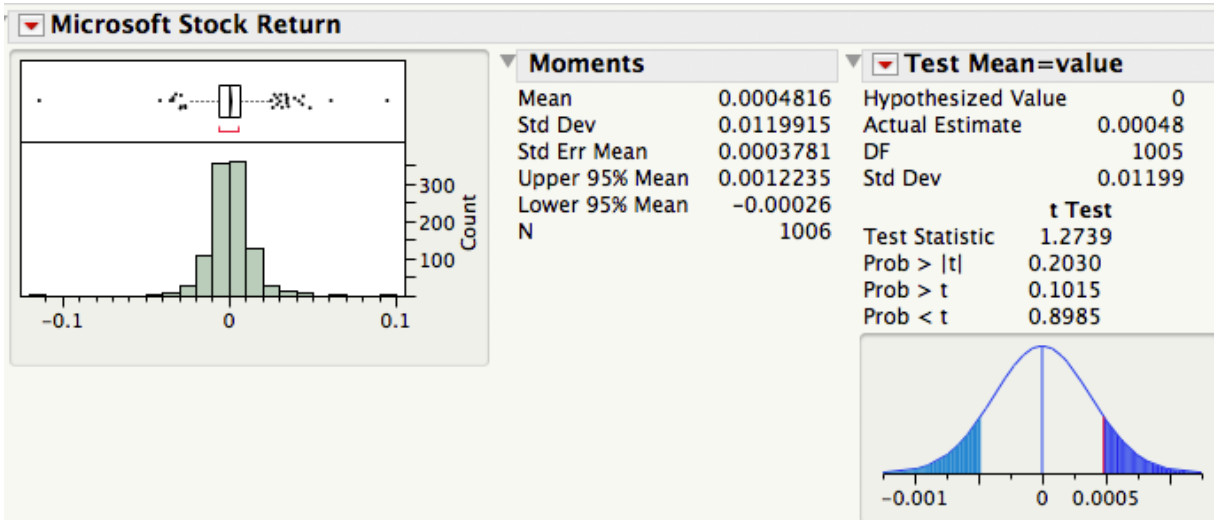
- Under $H_0: p \leq p_0$ and some conditions, the test statistic z is approximately standard normal.

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 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 [-0.0003, 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:

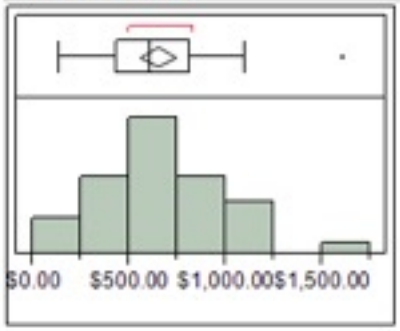
- Calculate the 1 - α level confidence interval
- Then
  - if μ_0 falls outside the interval, reject the null hypothesis;
  - otherwise, can't reject the null hypothesis.

Example: Denver_Rent

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

Distributions

Rent ($/Month)

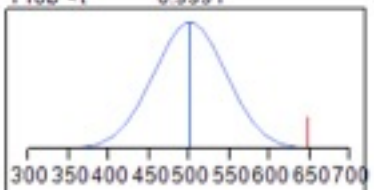


Quantiles		
100.0%	maximum	1600
99.5%		1600
97.5%		1525
90.0%		1040
75.0%	quartile	815
50.0%	median	610
25.0%	quartile	435
10.0%		246
2.5%		140
0.5%		140
0.0%	minimum	140

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?

Comparisons Between Two Groups

- One of the key business tasks faced by analysts is making comparisons:
  - between the achievements of two regional sales forces
  - between two competing hedge funds
  - between conversion rates between a new format for a retail web site and an old one.

- Comparisons are often the quantitative basis for activities such as “benchmarking” and “best practices”.

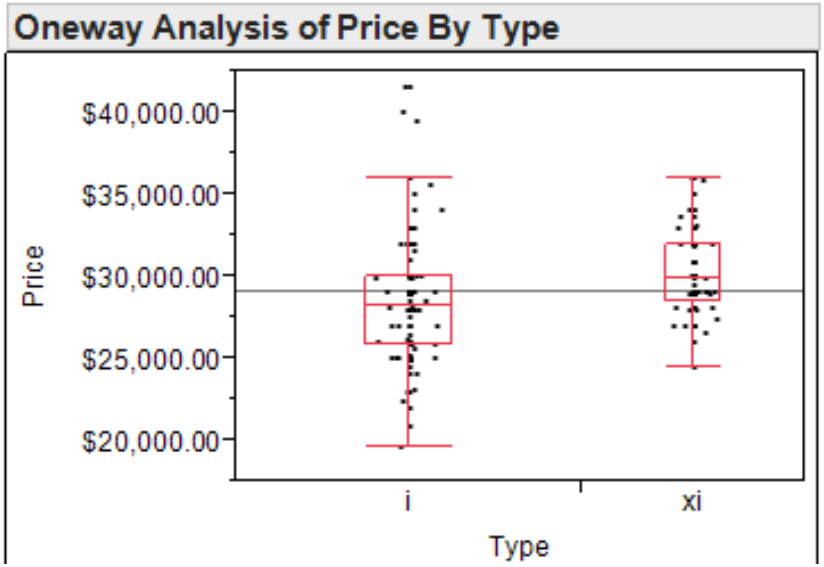
- This lecture considers comparisons between two groups. Later in the semester we will also compare many groups when we cover regression analysis.

Example: UsedCars

A car manufacturer uses the price of used cars to determine the initial cost charged to customers who lease its automobiles.

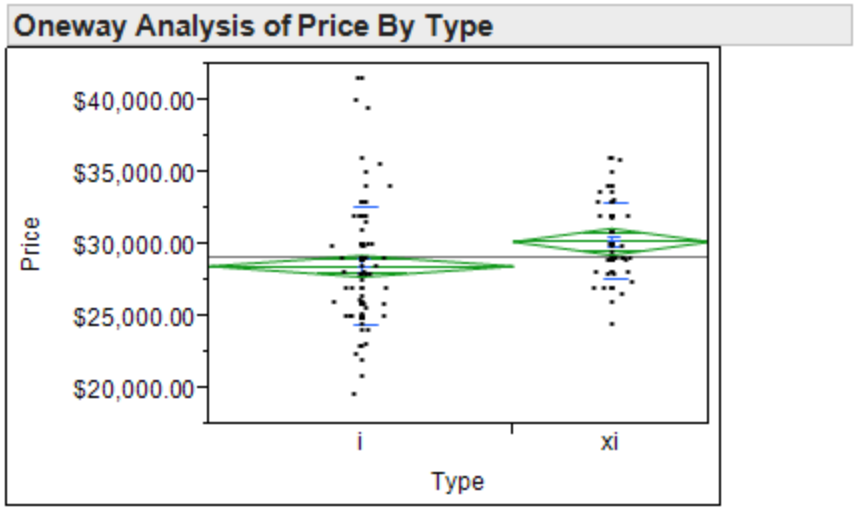
The file *UsedCars* contains the prices of 155 used BMW automobiles. Some are the *xi* model (4-wheel drive) and the others are the standard *i* model.

Boxplots of the prices show considerable overlap for the two types of cars.



Example: UsedCars

Compare the 95% confidence intervals for the two means.



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
i	94	28534.5	4158.98	428.97	27683	29386
xi	61	30233.4	2648.65	339.12	29555	30912

These intervals *do not* overlap, then the difference between the means is significantly different from 0.

If they overlap, you need to be more careful. (Example: Price Quotes)

UsedCars: Two-Sample T Test

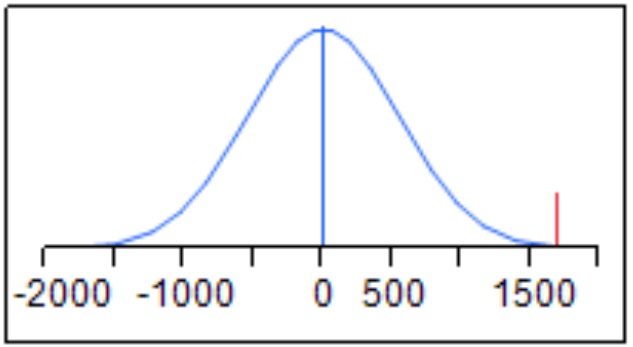
Oneway Analysis of Price By Type

t Test

xi-i

Assuming unequal variances

Difference	1698.95	t Ratio	3.106932
Std Err Dif	546.82	DF	152.9635
Upper CL Dif	2779.25	Prob > |t|	0.0023*
Lower CL Dif	618.64	Prob > t	0.0011*
Confidence	0.95	Prob < t	0.9989



This result is statistically significant because

1. The p-value=0.0023 is less than 0.05.
2. The 95% CI is [619,2779], not covering 0.

Summary: The Two-Sample T Test

- Consider two independent samples
  - $x_1, \ldots, x_m$ from population with mean μ_x
  - $y_1, \ldots, y_n$ from population with mean μ_y

- $H_0: \mu_x = \mu_y$ versus $H_a: \mu_x \neq \mu_y$

- Two-sample t statistic

$$t = \frac{\bar{x} - \bar{y}}{se(\bar{x} - \bar{y})} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x{}^2}{m} + \frac{S_y{}^2}{n}}}$$

  - It measures the number of standard errors between the observed statistic $\bar{x} - \bar{y}$ and 0.

Math Behind the Two-Sample Test

- According to CLT,

$$\bar{x}: \quad N(\mu_x, \sigma_x / \sqrt{m})$$
$$\bar{y}: \quad N(\mu_y, \sigma_y / \sqrt{n})$$

- Under $H_0: \mu_x = \mu_y$,

$$\bar{x} - \bar{y} \sim N(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}) = N(0, \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}})$$

- Standard error:

$$s_{\bar{x}-\bar{y}} = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}$$

Example: Comparing Sales Force Performance

Two pharmaceutical companies, GL and BW, are recently merged. Management of the merged company needs to reduce its sales force.

The management wants to know if the sales force from the two divisions differ.

Pharmasal-split

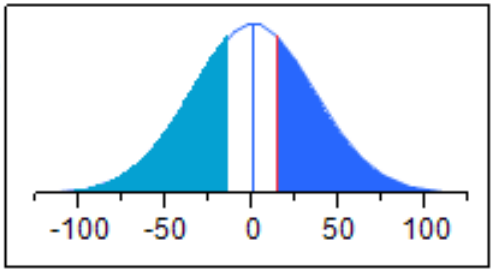
Oneway Analysis of Sale By BW/GL

t Test

GL-BW

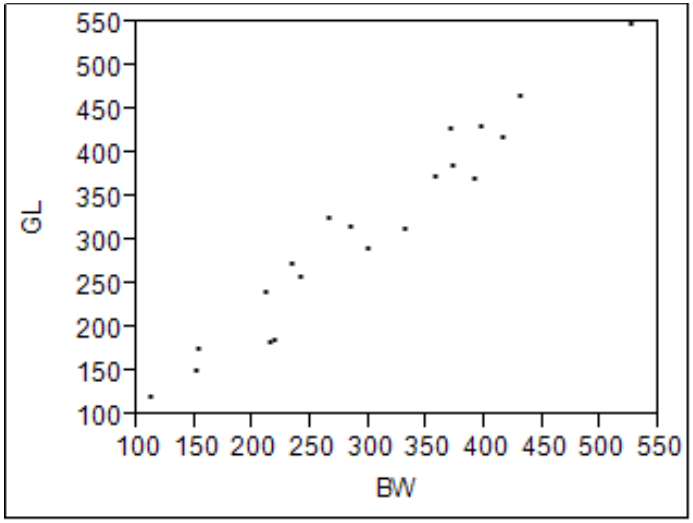
Assuming unequal variances

Difference	13.500	t Ratio	0.38045
Std Err Dif	35.484	DF	37.82717
Upper CL Dif	85.345	Prob > |t|	0.7057
Lower CL Dif	-58.345	Prob > t	0.3529
Confidence	0.95	Prob < t	0.6471



Example: Comparing Sales Force Performance

Both sales forces cover the same 20 sales district.



This pairing is a powerful step in the analysis, because there will be other sources of variability between the sales districts that has nothing to do with the performance between the two divisions.

- the size and demographics of the territory would be confounding variables.

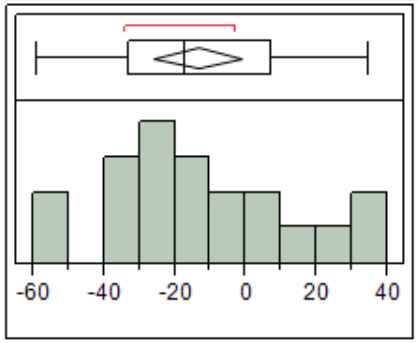
Pairing the sales districts mitigates this unwanted variability.

Comparing Sales Force: Paired T Test

- Differences=BW-GL
- One sample of differences
- See whether the mean is negative, or positive, or zero
- Paired t Test (i.e. one-sample t test on the difference)

Distributions

Differences

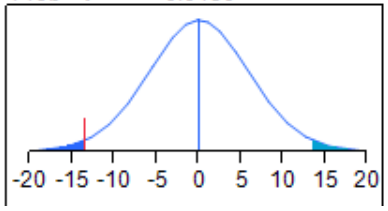


Quantiles		
100.0%	maximum	35
99.5%		35
97.5%		35
90.0%		30
75.0%	quartile	7.25
50.0%	median	-17
25.0%	quartile	-33
10.0%		-56.1
2.5%		-59
0.5%		-59
0.0%	minimum	-59

Summary Statistics	
Mean	-13.5
Std Dev	26.747356
Std Err Mean	5.9808906
Upper 95% Mean	-0.981852
Lower 95% Mean	-26.01815
N	20

Test Mean	
Hypothesized Value	0
Actual Estimate	-13.5
DF	19
Std Dev	26.7474

t Test	
Test Statistic	-2.2572
Prob > |t|	0.0360*
Prob > t	0.9820
Prob < t	0.0180*



The Paired T Test

In some situations the collected data naturally occurs in pairs,

- the weight of a person before and after a diet
- Subtracting one measurement from the other for the same person gives an interpretable quantity relating to the individual: the weight that they lost or gained.

Medical studies view ``identical twins" as a gold standard, because identical twins control for genetic variability, a confounding factor.

It is the controlling for *unwanted variability* that makes the paired tests we are about to discuss potentially very powerful.

Summary: The Paired *t* Test

Begin with two *matched* random samples:

- $x_1, \ldots, x_n$ from a population with unknown mean μ_x
- $y_1, \ldots, y_n$ on the same sample with unknown mean μ_y

Treat the differences $d_i = (x_i - y_i)$ as a random sample from a population with unknown mean $\mu_d = (\mu_x - \mu_y)$

- $H_0: \mu_x = \mu_y$ versus $H_a: \mu_x \neq \mu_y$
- $H_0: \mu_d = 0$ versus $H_a: \mu_d \neq 0$
- Test using the previous one-sample t test based on d_i

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}.$$

Two-Sample t Test vs. Paired t Test

- With appropriate pairing, the differencing between the pair members should provide a directly interpretable quantity.

Cape Cod: Covid Vaccine Really Works ✌️

capecod covid.jmp

	Treatment	Outcome	Counts
1	Vaccine	Covid	346
2	Vaccine	NonCovid	56654
3	NonVaccine	Covid	123
4	NonVaccine	NonCovid	2877

Contingency Table

Treatment	Count / Row %	Covid	NonCovid	Total
NonVaccine	123	2877	3000	
	4.10	95.90		
Vaccine	346	56654	57000	
	0.61	99.39		
Total	469	59531	60000	

Outcome

Tests

N	DF	-LogLike	RSquare (U)
60000	1	118.12750	0.0431

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	236.255	<.0001*
Pearson	448.357	<.0001*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	1.0000	Prob(Outcome=NonCovid) is greater for Treatment=NonVaccine than Vaccine
Right	<.0001*	Prob(Outcome=NonCovid) is greater for Treatment=Vaccine than NonVaccine
2-Tail	<.0001*	Prob(Outcome=NonCovid) is different across Treatment

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.

Impact of Sample Size on Inference

- Case: 2-101Kerrich
- John Kerrich, English mathematician, flipped a coin 10,000 times, while being jailed during WWII, and got 5,067 heads and 4,933 tails.
- Is the coin fair?
- p *:* the probability of head
- $H_0: p = 0.5$ versus $H_a: p \neq 0.5$

- Consider a small sample of 50, with 31 heads.
- And a large sample of 500,000, with 251,000 heads.

Impact of Sample Size

- For Kerrich sample, heads occurred 50.7% of the time.
  - The data are consistent with what we would expect from a fair coin.
  - However, we CAN'T state that the coin is fair.
- For the small sample, heads occurred 62% of the time.
  - But, we can not reject that the coin is fair!
  - The sample size is too small to detect the difference even though it appears so.
- For the large sample, heads occurred only 50.2% of the time.
  - However, inference says that the coin is not fair with a p-value of 0.0047!
  - Impact of large samples: large power to detect very small deviations from the null with large samples.
  - In fact, with large samples, rejection of null may be routine; but it may not have any practical significance.

Statistical Significance and Practical Significance

- When drawing conclusions from a hypothesis test, it is important to keep in mind the difference between Statistical and Practical Significance.

  - **Statistical Significance** : We can be sure that H_0 is false i.e. the difference from the hypothesized value is too large to be attributed to chance. Statistics can answer this question.

  - **Practical Significance** : Is the difference large enough that in practice we care? Statistics can not answer this one!

- Coin flipping – who takes out the trash? vs. betting in a casino.
- Marketing – a product appeals to 35.6% of females vs. 35.7% of males.

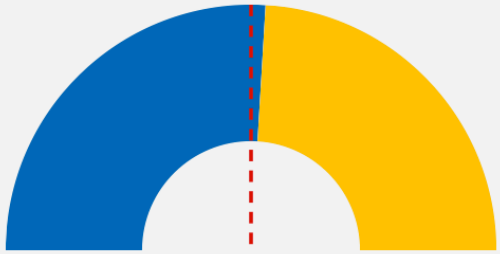
Brexit: 33,551,983 Final Votes Counted

EU REFERENDUM

Results

UK votes to **LEAVE** the EU

Leave
51.9%
17,410,742 VOTES



Remain
48.1%
16,141,241 VOTES

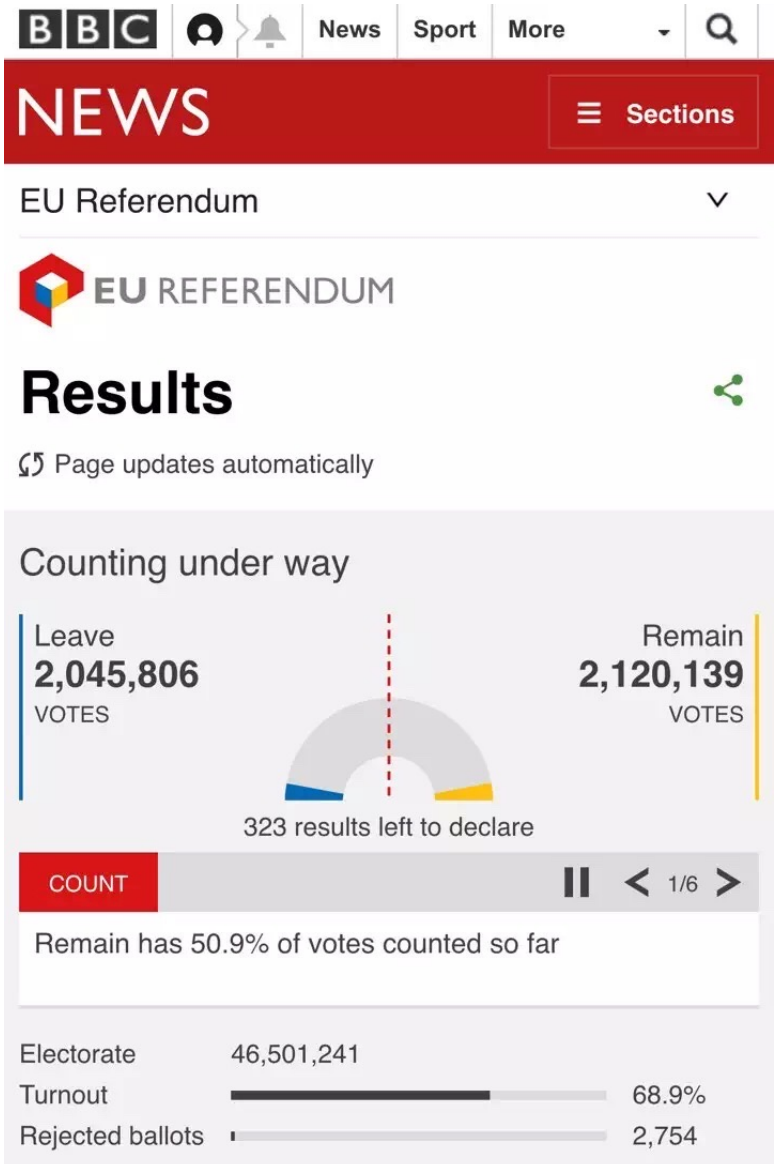
0 results left to declare

LEAVE | UK votes to **LEAVE** the EU

Electorate	46,501,241	
Turnout		72.2%
Rejected ballots		26,033

How results are calculated

Hindsight: 4,165,945 Votes Counted



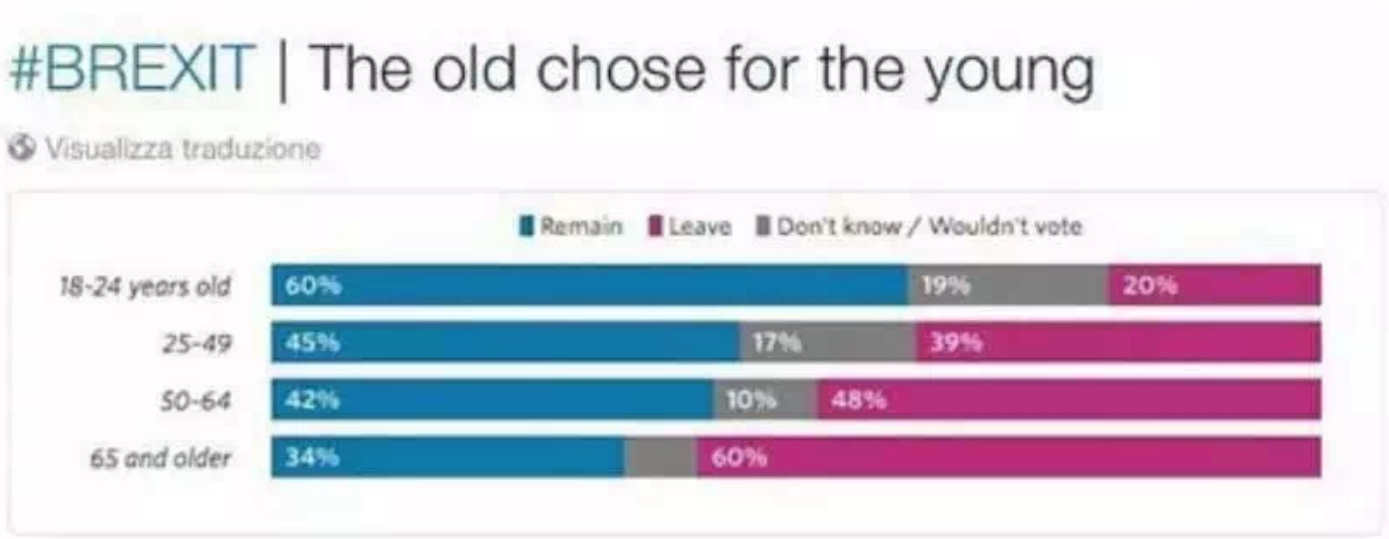
Hindsight: 21,062,344 Votes Counted

Age: The Old Chose for The Young

Age Group	Median Age	Remain	Leave	Life Expectancy	Average number of years they have to live with the decision
18-24	21	64%	24%	90	69
25-49	37	45%	39%	89	52
50-64	57	35%	49%	88	31
65+	73	33%	58%	89	16

Polling Data = YouGov. 1652 people. 17-19th June 2016
Life Expectancy based on ONS pension planner life expectancy estimator
Average 65+ year old was estimated to be 73 using ONS age distribution data
Those who were undecided or wouldn't say have been excluded

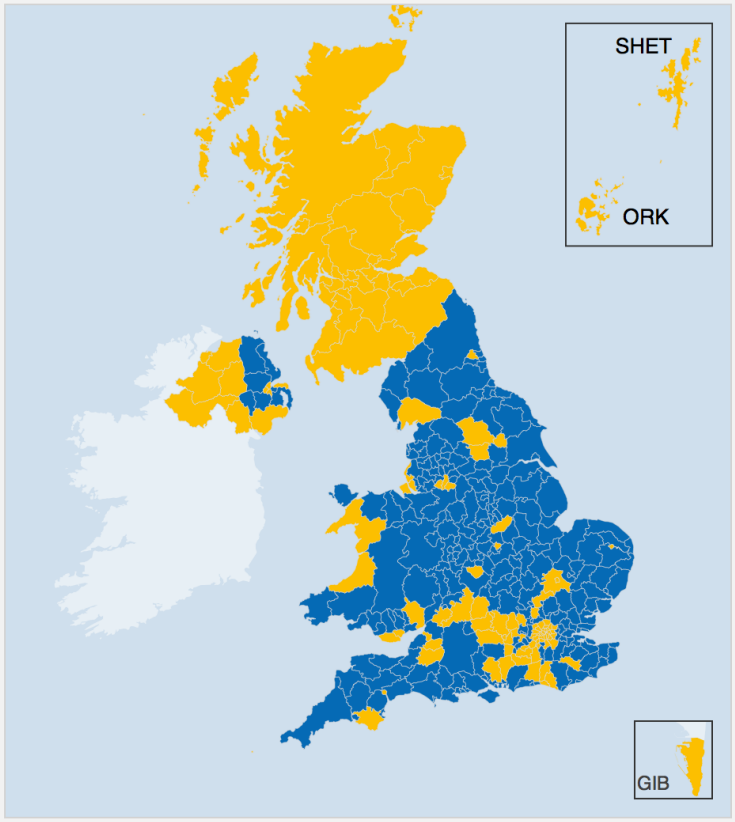
Those who must live with result of the EU referendum the longest want to remain.



Region: North vs. South

Find local results

Enter a postcode, council name or NI constituency



Key:
Majority leave
Majority remain
Tie
Undeclared

Nation results

England

Leave **53.4%**
15,188,406 VOTES

Remain **46.6%**
13,266,996 VOTES

Counting complete

Turnout: 73.0%

Northern Ireland

Leave **44.2%**
349,442 VOTES

Remain **55.8%**
440,437 VOTES

Counting complete

Turnout: 62.9%

Scotland

Leave **38.0%**
1,018,322 VOTES

Remain **62.0%**
1,661,191 VOTES

Counting complete

Turnout: 67.2%

Wales

Leave **52.5%**
854,572 VOTES

Remain **47.5%**
772,347 VOTES

Counting complete

Turnout: 71.7%

Relevance for Big Data

- Variation
- Uncertainty
- Factors to explain variation
- Confounding factors
- The bigger the data, the more valuable?
  - Representative sampling
- How big is enough?

Target Population Versus Sampled Population

Sample	Target Population	Sampled Population
Political Poll	Actual Voters	Registered Voters
20 Volunteer Tasters	Potential Consumers	Hungry Volunteers

Sampling bias is a mismatch between the target population and the sampled population.

Typical causes of sampling bias:

- self-selection, non-response, incentives to answer, interviewer characteristics, formulation of questions, and wording of questions.

The Digest Announcement, August 22, 1936

- The Poll represents thirty years' constant evolution and perfection. Based on "commercial sampling" methods used for more than a century by publishing houses to push book sales, the present mailing list is drawn from every telephone book in the United States, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail order and occupational data.

- Once again, THE DIGEST was asking more than ten million voters—one out of four, representing every county in the United States—to settle November's election in October.

- Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be *triple-checked*, verified, *five times* cross-classified and totaled. When the last figure has been totted and checked, if past experience is a criterion, the country will know *to within a fraction of 1 percent* the actual popular vote of forty millions.

The Literary Digest Poll

- In the 1936 US presidential election, *The Literary Digest* mailed questionnaires to 10 million people (25% of voters). 2.4 million people responded. (BIG DATA!)

- *The Digest* predicted an overwhelming victory of Alfred Landon (R) over Franklin Roosevelt (D):

  57% to 43%.

- Landon lost the election by a landslide

  38% to 62%.

- What went wrong?

Problems in the Literary Digest Poll

- Under-coverage bias:
  - telephone books, club memberships, mail order lists, automobile ownership lists.

- Nonresponse bias:
  - only 24% responded (and these were biased toward the Republicans)

- Gallup Poll:
  - surveyed 50,000 people and correctly predicted Roosevelt's victory
  - Gallup accurately predicted the *Digest* would declare Landon the winner, 56 percent to 44 percent, by sending postcards to a list of only 3,000 persons chosen at random from the same lists used by the *Digest*.
  - The *Digest* went bankrupt soon after the election.

The Bigger The Data, The More Valuable?

- Blood test: a boy, a grown-up
- Wine tasting: a sip, a glass, a bottle
- Marketing research about prospect of a new product:
  - In US, sufficient to sample 1,000 people
  - How about China?
    - Population size: China = 4 times US
- Sampling accuracy: population variability, sample size

Survivor Bias

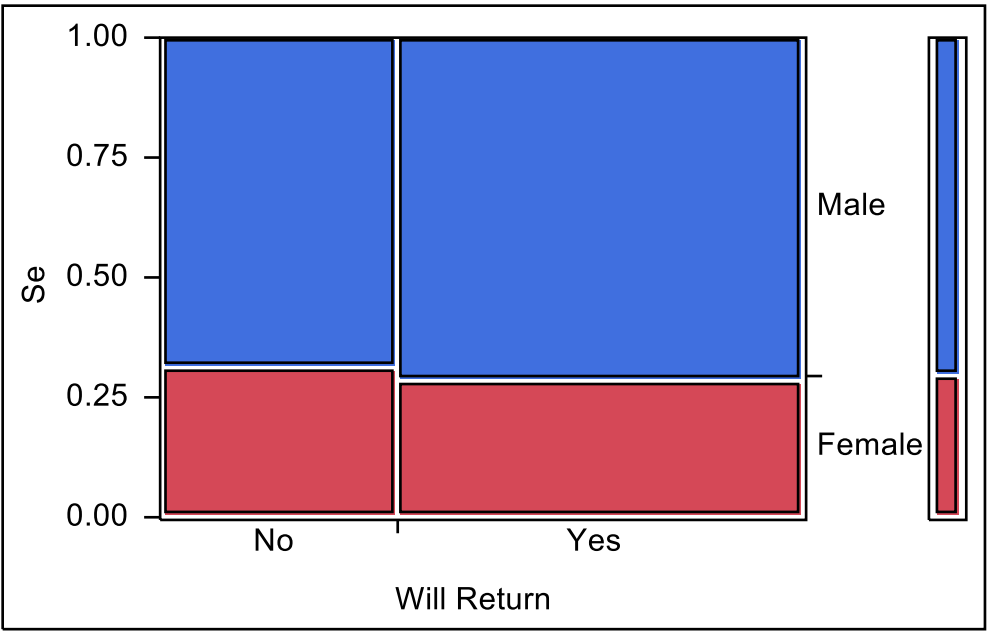
A regional hotel chain was studying its customer base. One question of interest was what percentage of guests plan to return to hotels in the chain.

Voluntary customer satisfaction surveys had been left in the rooms, but only 10% of the guests were completing them. This resulted in strong selection bias—only those who were very pleased and those who were very displeased were filling out the questionnaires.

A different approach decided that the population of interest consists of "primary guests," those who sign the register and make payment. A decision was made to interview every primary guest present 15 June 1994.

Management was able to interview 97 percent of the primary guests staying in the chain on that day.

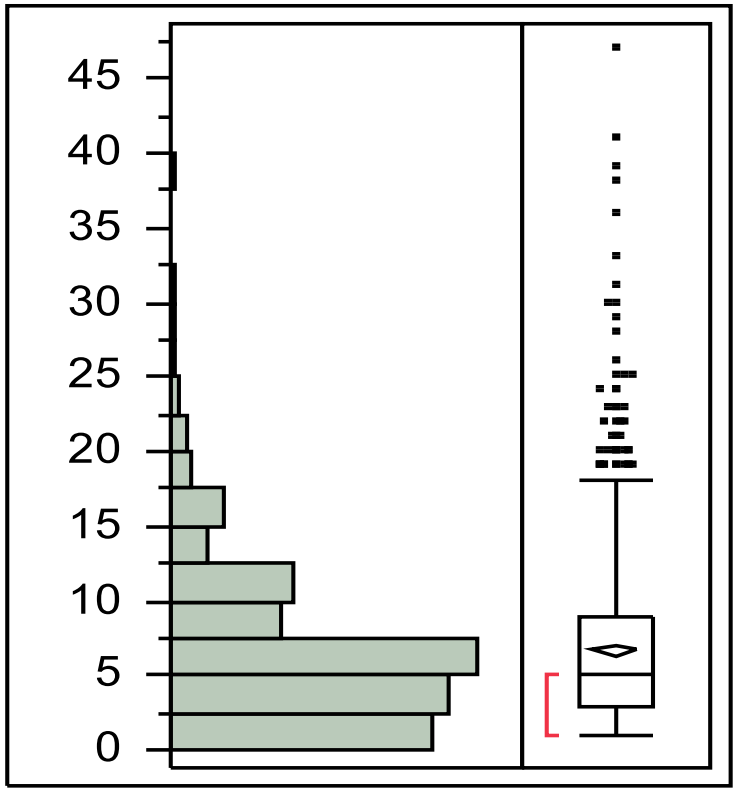
Survey1



Count Total % Col % Row %	Female	Male	
No	130 11.57 39.27 31.25	286 25.44 36.07 68.75	416 37.01
Yes	201 17.88 60.73 28.39	507 45.11 63.93 71.61	708 62.99
	331 29.45	793 70.55	1124

plans to return are not associated with gender.

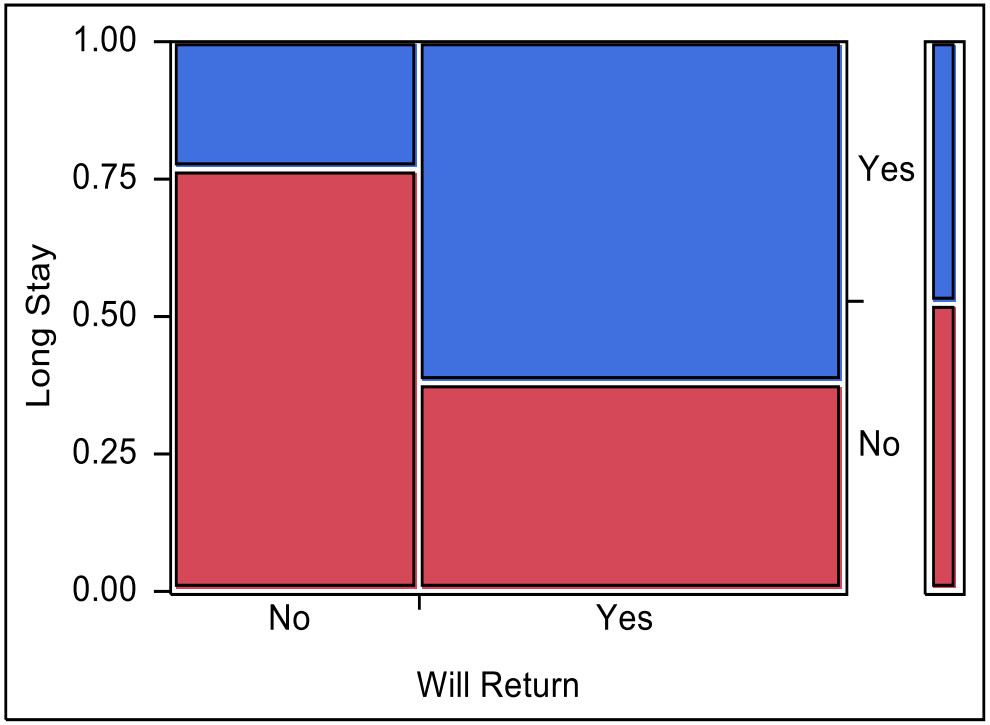
Length of Stay (in days)



100.0%	maximum	47
90.0%		14
75.0%	quartile	9
50.0%	median	5
25.0%	quartile	3
10.0%		2
0.0%	minimum	1

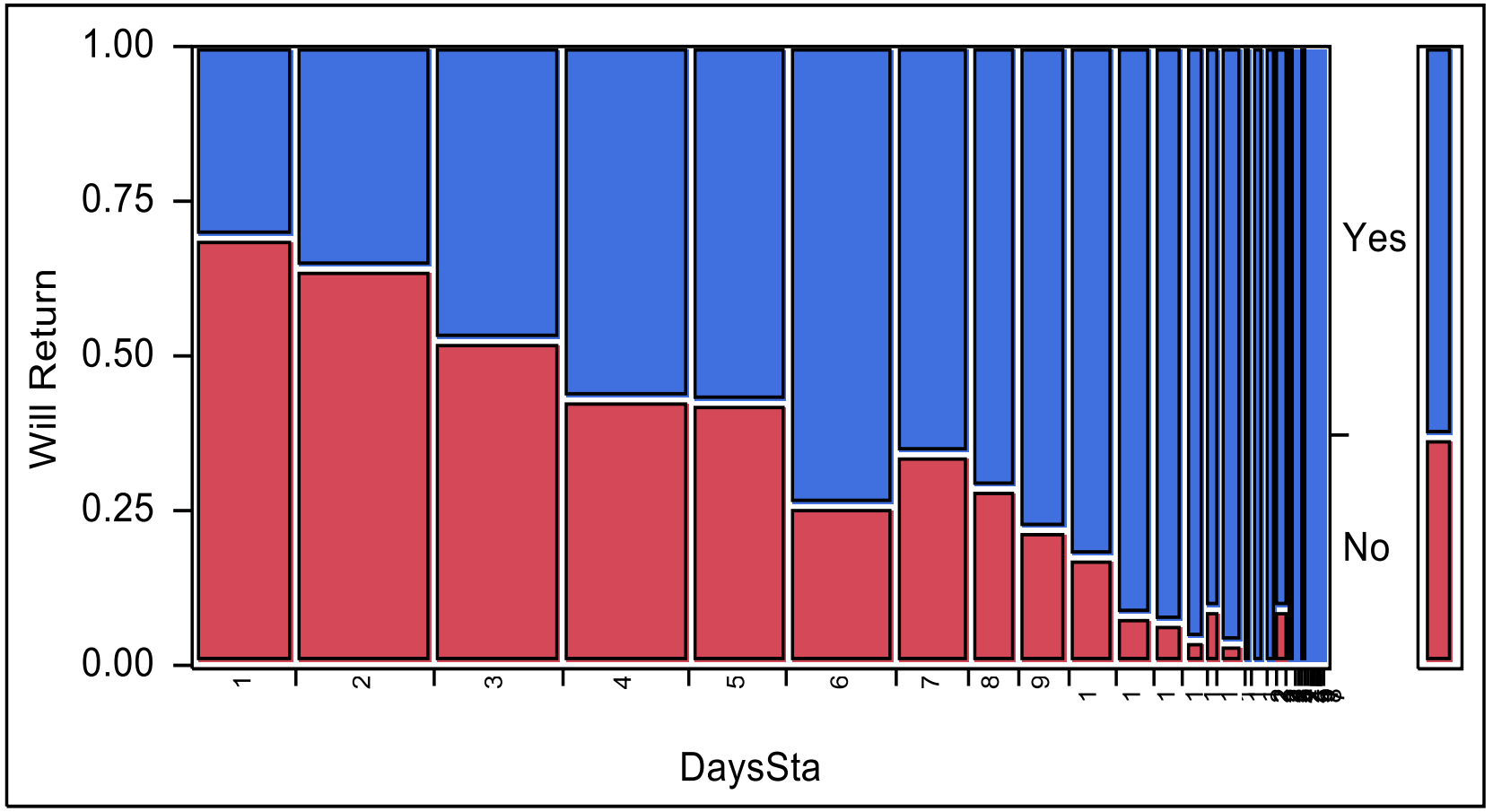
Mean	6.701
Std Dev	5.614
N	1124

LongStay?=Yes, if length of stay >= 6 days



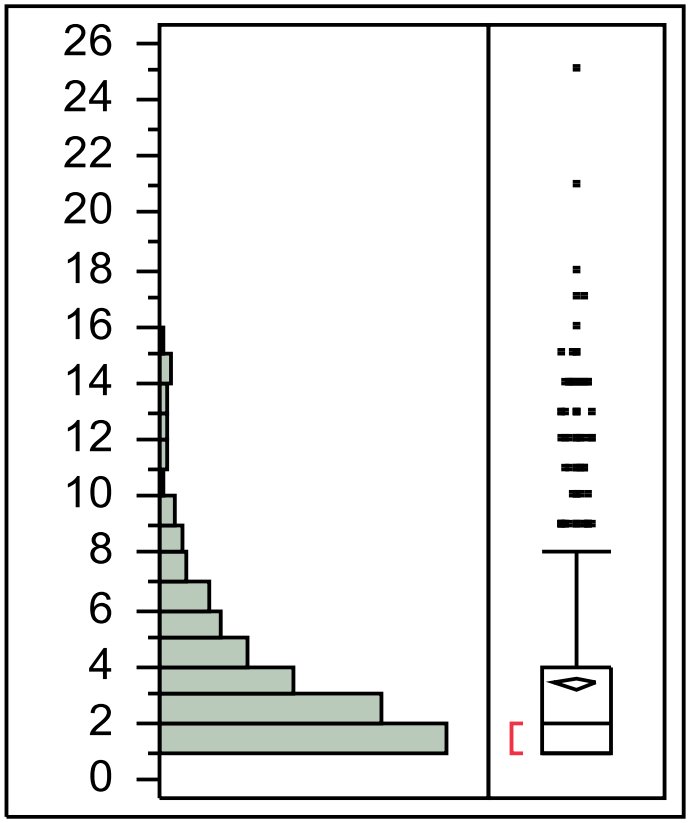
Clearly, the longer stays form a more satisfied group.

Association between DaysStay and Willingness to Return

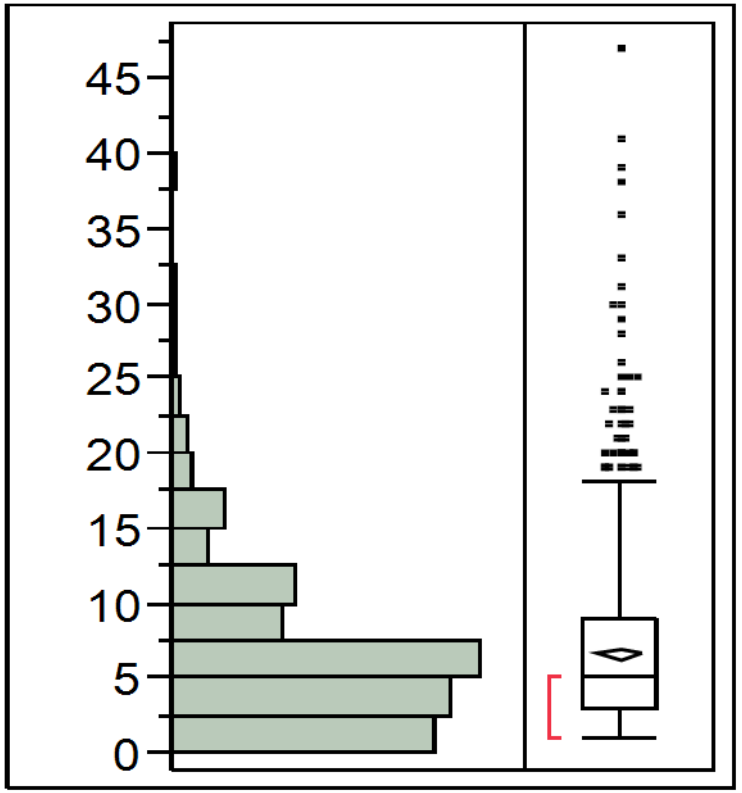


Survey2

1000 stays from the past year



One-day survey



Size-based Sampling Bias

Last year data

100.0%	maximum	25
90.0%		7
75.0%	quartile	4
50.0%	median	2
25.0%	quartile	1
10.0%		1
0.0%	minimum	1

Mean	3.397
Std Dev	3.0654
N	1000

One-day survey

100.0%	maximum	47
90.0%		14
75.0%	quartile	9
50.0%	median	5
25.0%	quartile	3
10.0%		2
0.0%	minimum	1

Mean	6.701
Std Dev	5.614
N	1124

The one-day survey has selection bias. By interviewing all the guests present on one day, the chain oversampled those with long lengths of stay, who are more likely to return. Hence, overestimate the intent-to-return.

How to Fix the Sampling? - Weighting

Weight the observations, with less weight given to those who stayed long, in order to compensate for the selection bias. Choose weights equal to the reciprocal of the length of stay and perform a weighted analysis.

	unweighted
Mean	6.701
Std Dev	5.614
N	1124

	weighted
Mean	3.473
Std Dev	1.798
N	1124

	last-year
Mean	3.397
Std Dev	3.065
N	1000

- The weighting analysis estimates 47% intent-to-return.
- The unweighting analysis: 63% intent-to-return.