

Math 321

Applied Statistical Methods

Spring 2021

Dr. Gerri Dunnigan

Introduction: Statistics

- The discipline of statistics teaches us how to make intelligent and informed decisions in the presence of uncertainty and variation.
- Statistics is the science of analyzing data.
- Statistics is the mathematical formalization of using data for estimation and decision making.

Suppose you are the quality control officer at a soft drink bottling plant. The bottles being filled are labeled as holding 16 ounces of soft drink, which is the target fill for each bottle. Due to variation in the filling process, the exact amount of soft drink in each bottle will likely deviate from the target of 16 ounces. Not wanting to severely over- or under-fill the bottles, you want the amount of soft drink in the bottles to be $16 \pm .25$ ounces. The goal is to achieve at least 90% conformity with these specifications.

On any given day, 1600 bottles are filled per minute at the plant. To determine if specifications are being met on a given day, you consider the first 1600 bottles filled. You do not have the time or resources to measure the amount of fill in all of the bottles to determine the number of them meeting specifications, so you randomly select 100 of the bottles, measure the amount of soft drink in each bottle, and determine the number that meet specifications.

Suppose 92 of the 100 bottles, or 92% of the sampled bottles meet specifications.

The proportion of bottles filled during the first minute of production that meet specifications is unlikely to be 92% as the sample of 100 is unlikely to represent perfectly the entire population of bottles filled during that time.

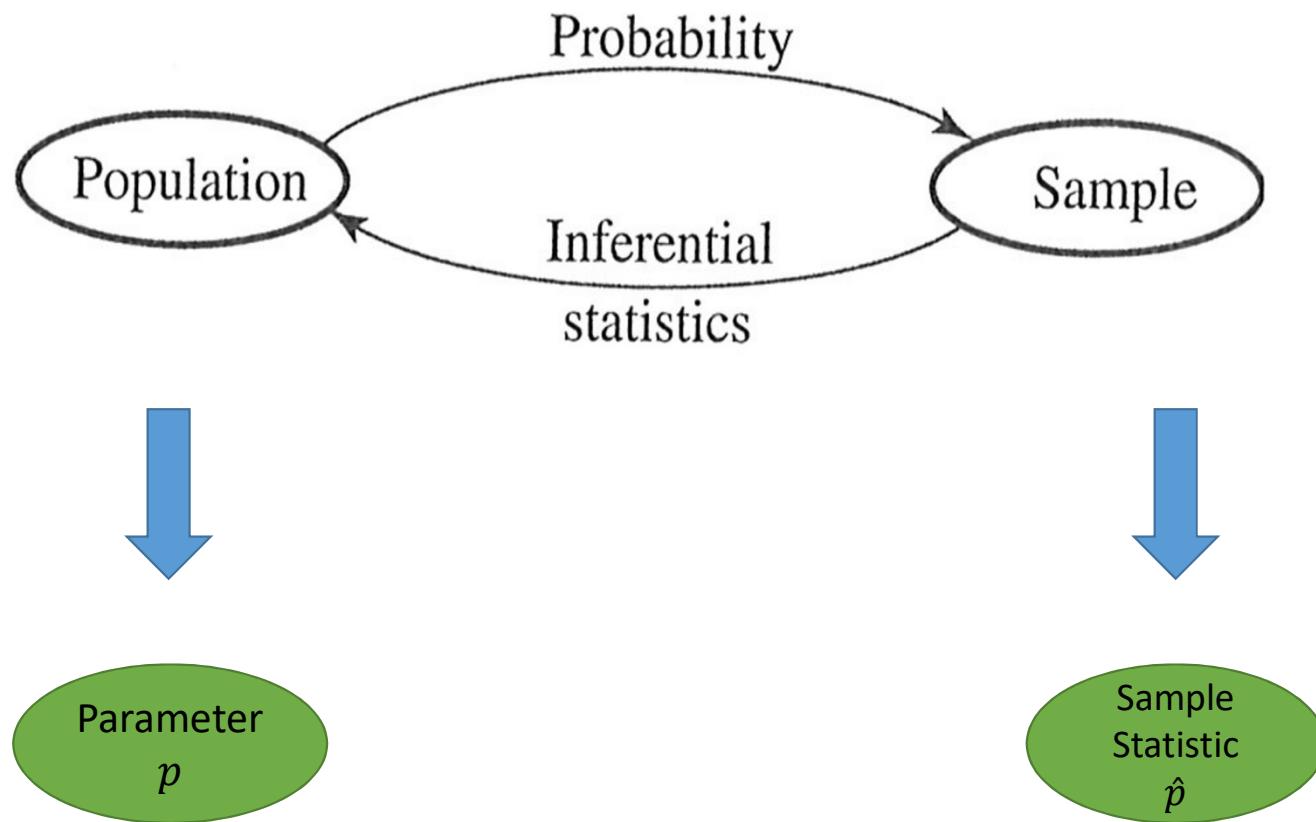
One question that you may want to answer is: What proportion of bottles filled during the first minute of operation meet specifications?

Using the sample proportion of 92%, you want to determine an interval of the form $92\% \pm x\%$ for which you can be reasonably certain contains the true value of the proportion of bottles filled during the first minute of operation that meet specifications.

This is an example of what is called a **confidence interval** which is one of the statistical inference tools that we will be studying.

You may want to address the specific question: Is the proportion of bottles filled during the first minute of production that meet specifications at least 90% as desired?

Answering this question involves another tool of statistical inference called a **hypothesis test**.



Statistics is used in almost all disciplines.

Medicine

Engineering

Economics

Biology

Business

https://en.wikipedia.org/wiki/List_of_fields_of_application_of_statistics

List of fields of application ...

File Edit View Favorites Tools Help

Google iomics Search More Sign In

Create account Log in

WIKIPEDIA The Free Encyclopedia

Article Talk Read Edit View history Search

List of fields of application of statistics

From Wikipedia, the free encyclopedia

Statistics is the mathematical science involving the collection, analysis and interpretation of data. A number of specialties have evolved to apply statistical theory and methods to various disciplines. Certain topics have "statistical" in their name but relate to manipulations of probability distributions rather than to statistical analysis.

- **Actuarial science** is the discipline that applies mathematical and statistical methods to assess risk in the [insurance](#) and [finance](#) industries.
- **Astrostatistics** is the discipline that applies statistical analysis to the understanding of astronomical data.
- **Biostatistics** is a branch of [biology](#) that studies biological phenomena and observations by means of statistical analysis, and includes [medical statistics](#).
- **Business analytics** is a rapidly developing business process that applies statistical methods to data sets (often very large) to develop new insights and understanding of business performance & opportunities.
- **Chemometrics** is the science of relating measurements made on a [chemical](#) system or process to the state of the system via application of mathematical or statistical methods.
- **Demography** is the statistical study of all [populations](#). It can be a very general science that can be applied to any kind of dynamic population, that is, one that changes over time or space.
- **Econometrics** is a branch of [economics](#) that applies statistical methods to the empirical study of economic theories and relationships.
- **Environmental statistics** is the application of statistical methods to [environmental science](#). Weather, climate, air and water quality are included, as are studies of plant and animal populations.
- **Epidemiology** is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine.
- **Geostatistics** is a branch of [geography](#) that deals with the analysis of data from disciplines such as [petroleum geology](#), [hydrogeology](#), [hydrology](#), [meteorology](#), [oceanography](#), [geochemistry](#), [geography](#).
- **Operations research** (or Operational Research) is an interdisciplinary branch of applied mathematics and formal science that uses methods such as mathematical modeling, statistics, and algorithms to arrive at optimal or near optimal solutions to complex problems.
- **Population ecology** is a sub-field of [ecology](#) that deals with the dynamics of species populations and how these populations interact with the [environment](#).
- **Psychometric** is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits.
- **Quality control** reviews the factors involved in manufacturing and production; it can make use of [statistical sampling](#) of product items to aid decisions in [process control](#) or in accepting deliveries.
- **Quantitative psychology** is the science of statistically explaining and changing mental processes and behaviors in humans.
- **Reliability Engineering** is the study of the ability of a system or component to perform its required functions under stated conditions for a specified period of time.
- **Statistical finance**, an area of [econophysics](#), is an empirical attempt to shift finance from its [normative](#) roots to a [positivist](#) framework using exemplars from statistical physics with an emphasis on emergent or collective properties of financial markets.
- **Statistical mechanics** is the application of [probability](#) theory, which includes mathematical tools for dealing with large populations, to the field of [mechanics](#), which is concerned with the motion of particles or objects when subjected to a force.
- **Statistical physics** is one of the fundamental theories of [physics](#), and uses methods of [probability](#) theory in solving physical problems.
- **Statistical thermodynamics** is the study of the microscopic behaviors of [thermodynamic](#) systems using probability theory and provides a molecular level interpretation of thermodynamic quantities such as [work](#), [heat](#), [free energy](#), and [entropy](#).

Statistics

Outline · Statisticians · Glossary · Notation · Journals · Lists of topics · Articles · Portal · Category

V · T · E

See also [edit]

- List of statistics topics

Statistics

Outline · Index

10 [hide]

Basic Definitions

Population: a well-defined collection of objects of interest

- **Tangible (or concrete):** the population is a finite collection of objects which can be enumerated
- **Conceptual (or hypothetical):** the collection of possible values or outcomes that might have been observed under identical experimental conditions

Sample: a subset of the population used to make decisions or inferences concerning the population

Parameter: a characteristic or piece of information about a population that is of interest

Sample statistic: a value calculated from the sample used to make inferences about a parameter

Examples:

Tangible Population: The GPAs of all students currently enrolled at UND.

- **Parameter:** Average GPA (μ) “Greek letter mu”
- **Sample:** Randomly select 100 students
- **Statistic:** Sample average GPA of the 100 students (\bar{x})

Conceptual Population: The strength of welds produced by a specific welding process

- **Parameter:** Average strength of welds using this process (μ)
- **Sample:** Produce 50 welds using the process
- **Statistics:** Sample average strength of the 50 welds (\bar{x})

Sampling Strategies

- **Simple random sampling:** (For tangible populations.) Label items in a population 1 through N (where N represents the number of items in the population), select a sample of size n by generating n random numbers and selecting the corresponding items from the population as the sample. This way each sample of size n will have the same chance of occurring as any other sample of size n .
- equivalent to drawing names out of a hat
- no guarantee the sample will be representative of the desired population

- **Stratified random sample:** subdivide the population into non-overlapping groups and randomly select a sample from each group and combine them to form the sample
- Controls for under or overrepresentation of any portion of the population

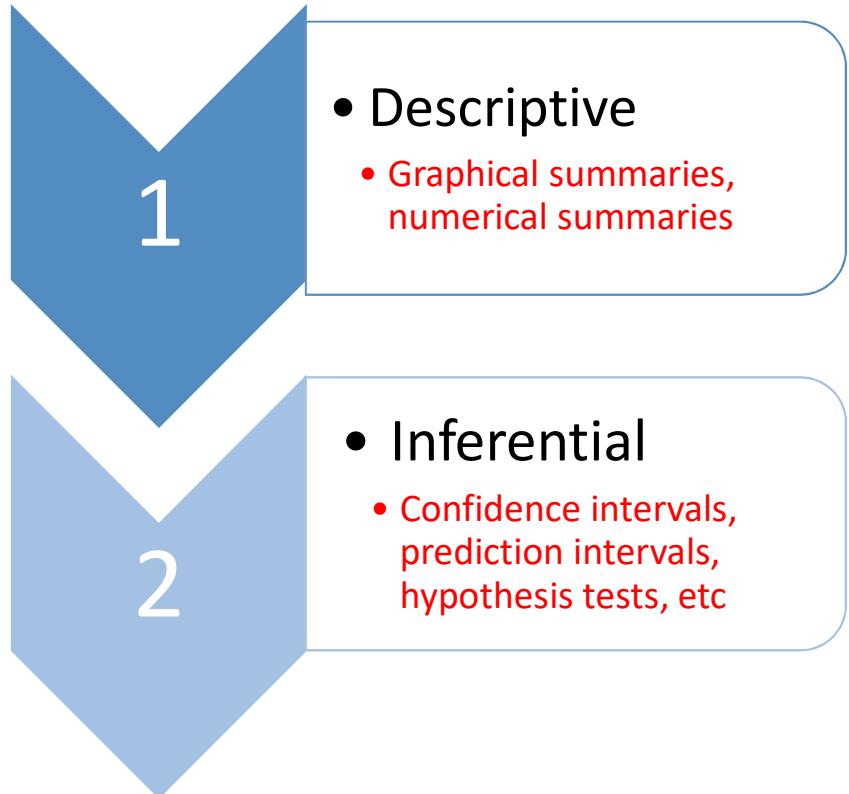
- **Samples of convenience:** choosing a sample without any attempt at randomization
- Very likely to not be representative of the population
- Not recommended for statistical inference!!

Types of Data Sets

- **Univariate:** measurements or observations made on a single variable of interest
- **Bivariate:** measurements or observations made on two variables of interest (**simple linear regression**)
- **Multivariate:** measurements or observations made on multiple variables of interest

- **Categorical or qualitative:** data values represent membership in a collection of categories corresponding to a particular attribute
 - Hair color: Brown, Blonde, Black, Red
 - Gender: Male, female
- **Numerical or quantitative:** measurement or count data
 - Lifetime of an electrical component
 - The number of diseased ash trees in a given section of a forest

Two Branches of Statistics



HOW TO KNOW IT'S TIME TO STOP DRIVING!!!
A picture is worth a thousand words!



Graphical Displays

- Stem-and-leaf plots
- Dotplots
- Histograms
- Boxplots
- Probability plots
- Scatterplots

Stem-and-leaf Plots

Do running times of American movies differ somehow from French movies? The author investigated this question by randomly selecting 25 recent movies of each type, resulting in the following running times:

Am: 94 90 95 93 128 95 125 91 104 116 162 102
90 110 92 113 116 90 97 103 95 120 109 91 138

Fr: 123 116 90 158 122 119 125 90 96 94 137 102
105 106 95 125 122 103 96 111 81 113 128 93 92

Am: 94 90 95 93 128 95 125 91 104 116 162 102
90 110 92 113 116 90 97 103 95 120 109 91 138

Smallest value: 90 Largest Value: 162

Leaf Unit: 1 minute

Fr: 123 116 90 158 122 119 125 90 96 94 137 102
105 106 95 125 122 103 96 111 81 113 128 93 92

Largest: 158 Smallest: 81

8	1								
9	0	0	6	4	5	6	3	2	
10	2	5	6	2					
11	6	9	1	3					
12	3	2	5	5	2	8			
13	7								
14									
15	8								

Leaf Unit: 1 minute

French Film Lengths

American Film Lengths

										1	8																
				6	6	5	4	3	2	0	0	9	0	0	0	1	1	2	3	4	5	5	5	7			
						6	5	2	2	2	10	2	3	4	9												
						9	6	3	1		11	0	3	6	6												
						8	5	5	3	2	2	12	0	5	8												
										7	13	8															
											14																
										8	15																
											16	2															

Leaf Unit: 1 minute

Minitab 17 Stem-and-leaf Plots

Always include an indicator of the leaf unit.

Stem-and-leaf of French N = 25
Leaf Unit = 1.0

1 8 1
9 9 00234566
(4) 10 2356
12 11 1369
8 12 223558
2 13 7
1 14
1 15 8

Stem-and-leaf of American N = 25
Leaf Unit = 1.0

12 9 000112345557
(4) 10 2349
9 11 0366
5 12 058
2 13 8
1 14
1 15

Minitab - Untitled

File Edit Data Calc Stat Graph Editor Tools Window Help Assistant

Session

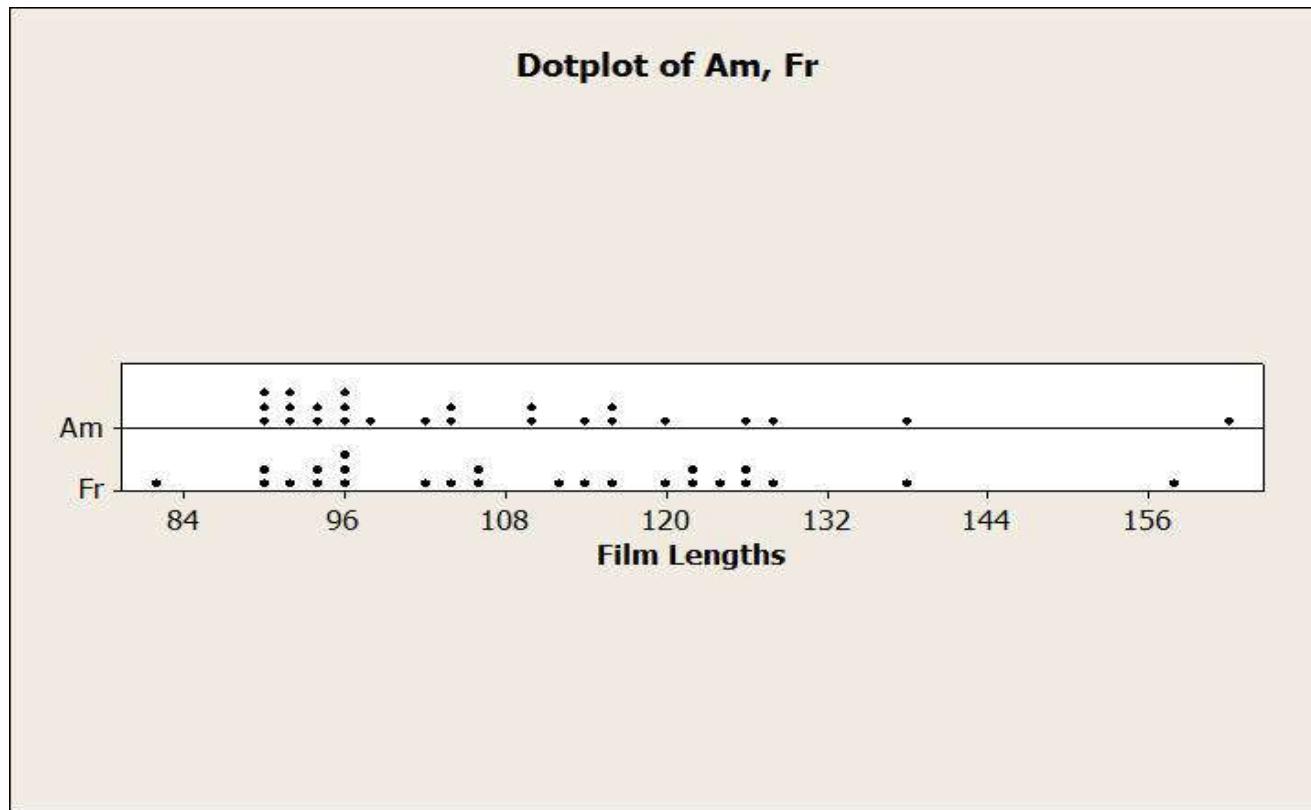
6/15/2016 8:56:08 AM

Welcome to Minitab, press F1 for help.

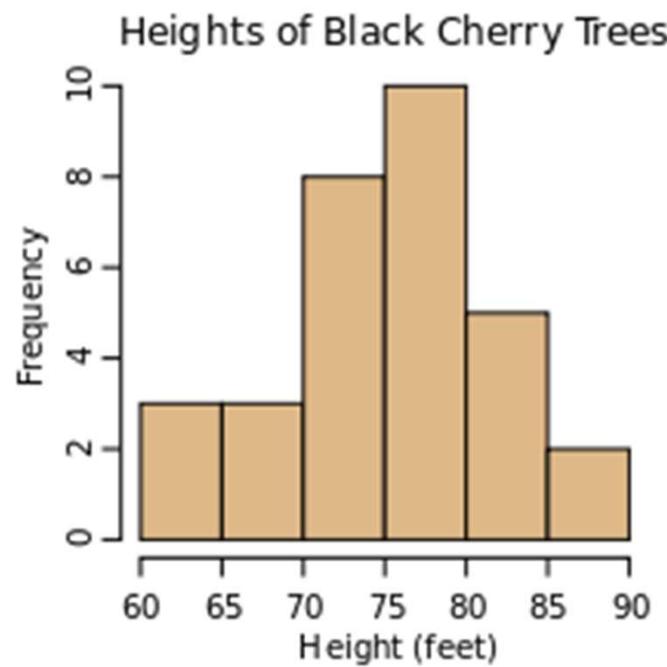
FilmData.MTW **

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	
French	American																							
1	123	94																						
2	116	90																						
3	90	95																						
4	158	93																						
5	122	128																						
6	119	95																						
7	125	125																						
8	90	91																						
9	96	104																						
10	94	116																						
11	137	162																						

Dotplots



Histograms



Constructing a histogram

1. Determine the minimum and maximum values in the data set.
2. Divide the interval formed by these values into subintervals or “classes”.
 - Between 5-20 classes is usually a good starting point, with the larger data sets requiring a greater number of classes.
 - A reasonable rule of thumb is
$$\text{number of classes} \approx \sqrt{\text{sample size}}$$

- Once the number of classes is determined, choose non-overlapping classes of equal width that cover the entire range of values in the data set.
- Next we will construct a frequency distribution that keeps a tally of the number of data values in each class, n_i .
 - Once this is done, we will construct a graph using the Cartesian coordinate system with the horizontal axis representing the classes and the vertical axis representing frequency (n_i) or relative frequency $f_i = n_i/n$, where n = the sample size.
 - A rectangle will be constructed above each class with height equal to the frequency or relative frequency of observations in that class.

Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from Wisconsin Power and Light determined the energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. An adjusted consumption value was calculated as follows:

$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather, in degree days})(\text{house area})}$$

This resulted in the following set of data.

7.87	11.12	8.58	12.91	12.28	14.24	11.62	7.73	7.15
9.43	13.43	8.00	10.35	7.23	11.43	11.21	8.37	12.69
7.16	9.07	5.98	9.60	2.97	10.28	10.95	7.29	13.38
8.67	6.94	15.24	9.58	8.81	13.60	7.62	10.49	13.11
12.31	10.28	8.54	9.83	9.27	5.94	10.40	8.69	10.50
9.84	9.37	11.09	9.52	11.29	10.36	12.92	8.26	14.35
16.90	7.93	11.70	18.26	8.29	6.85	15.12	7.69	13.42
10.04	13.96	12.71	10.64	9.96	6.72	13.47	12.19	6.35
12.62	6.8	6.78	6.62	10.30	10.21	8.47	5.56	9.83
7.62	4.00	9.82	5.20	16.06	8.61	11.7	9.76	12.16

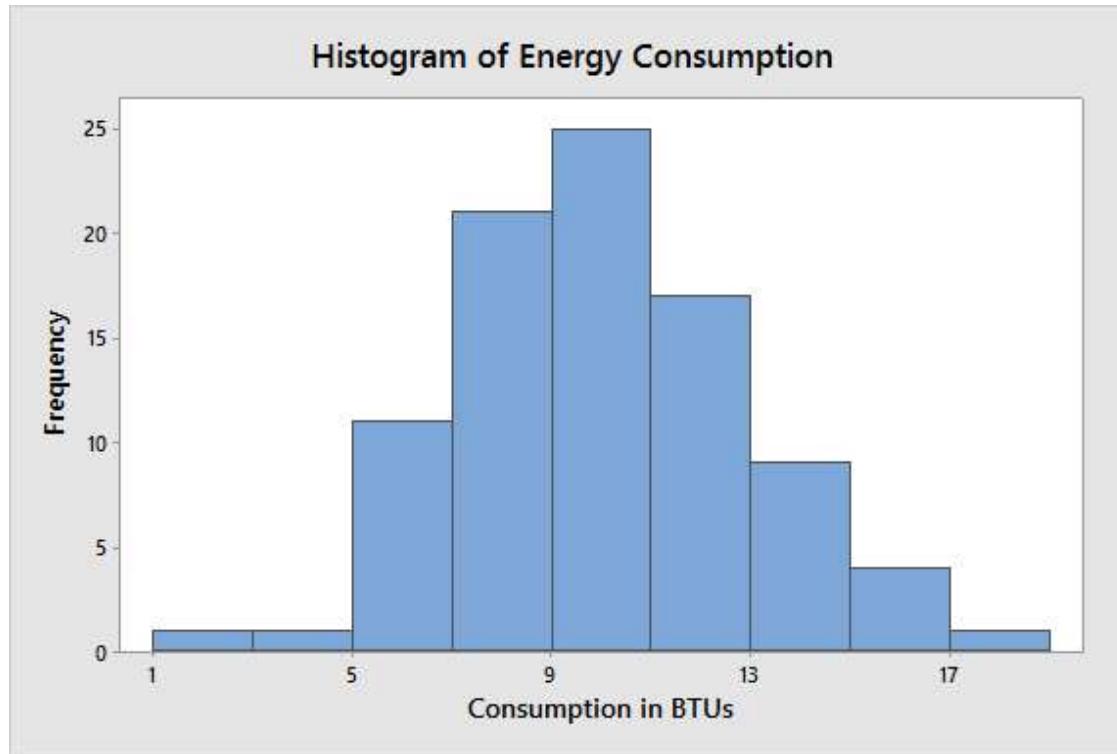
We have 90 data values so we should consider approximately $\sqrt{90} \approx 9.5$ classes. So around 9 or 10 classes should be appropriate.

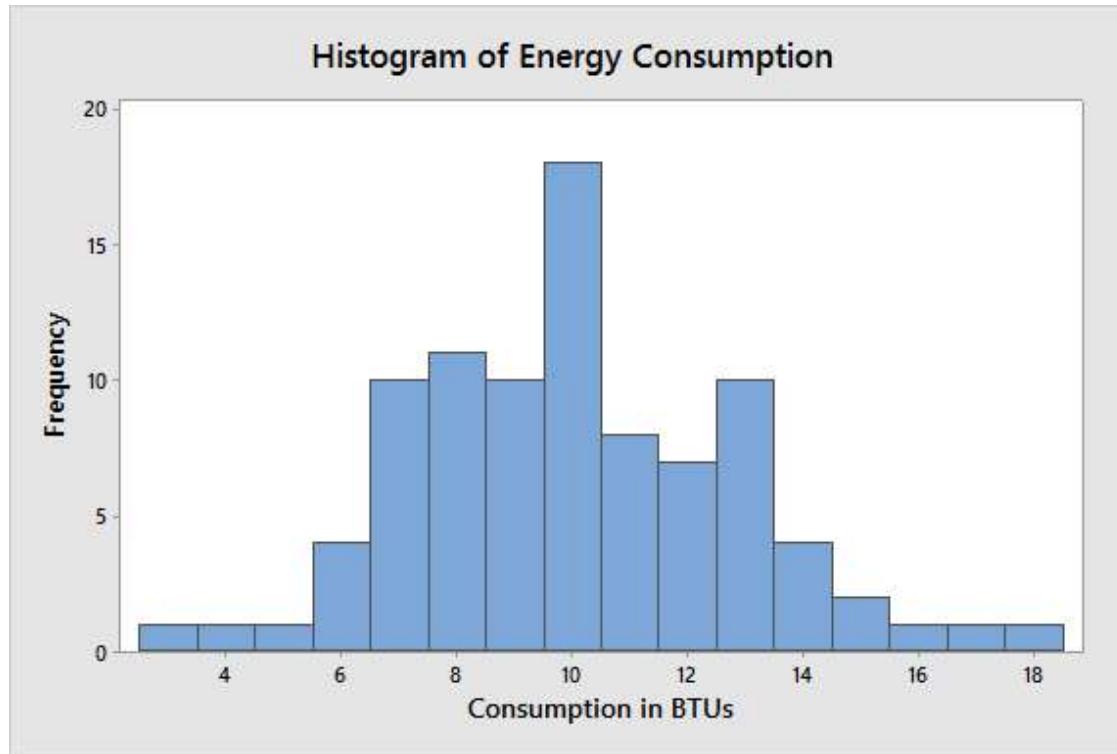
Let's use 9 classes and start the first class at 1 and end the last class at 19, with equal class lengths of 2. The classes will be closed on the left, open on the right.

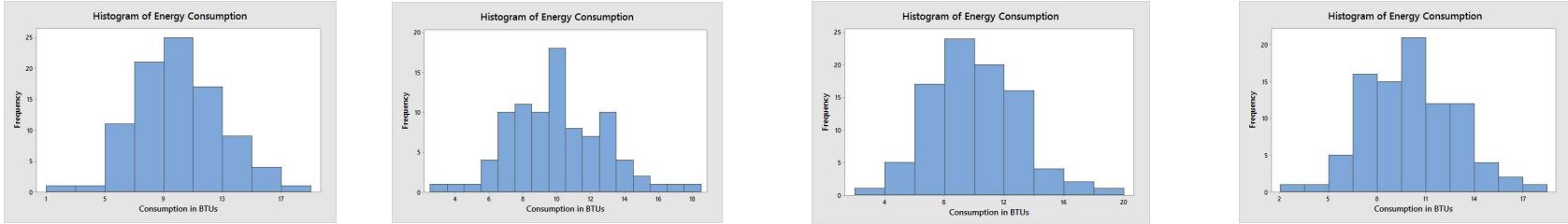
Frequency Distribution

Class	Tally	Frequency	Relative Frequency
1-<3			
3-<5			
5-<7			
7-<9			
9-<11			
11-<13			
13-<15			
15-<17			
17-<19			

Class	Tally	Frequency	Relative Frequency
1-<3		1	.0111
3-<5		1	.0111
5-<7		11	.1222
7-<9		21	.2333
9-<11		25	.2778
11-<13		17	.1889
13-<15		9	.1000
15-<17		4	.0444
17-<19		1	.0111







It's good practice to look at several histograms of a data set using different numbers of classes and different class intervals. Important features of the data set should appear in many of the histograms. Features that only appear in some of the graphs are less likely to be important.

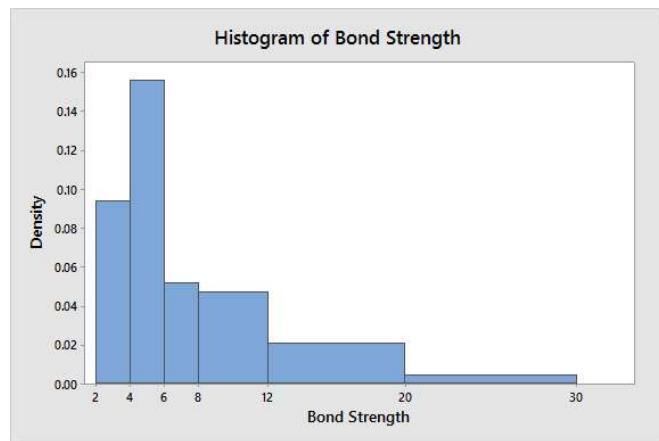
Data sets that have some regions with high concentrations of data and other regions where data is sparse, equal class widths may not be a sensible choice.

If we use classes of unequal width, we do not want to use the frequency or relative frequency histograms as they tend to result in incorrect interpretations.

In these cases we should use a density histogram where the area of each bar represents the relative frequency of observations in the given class.

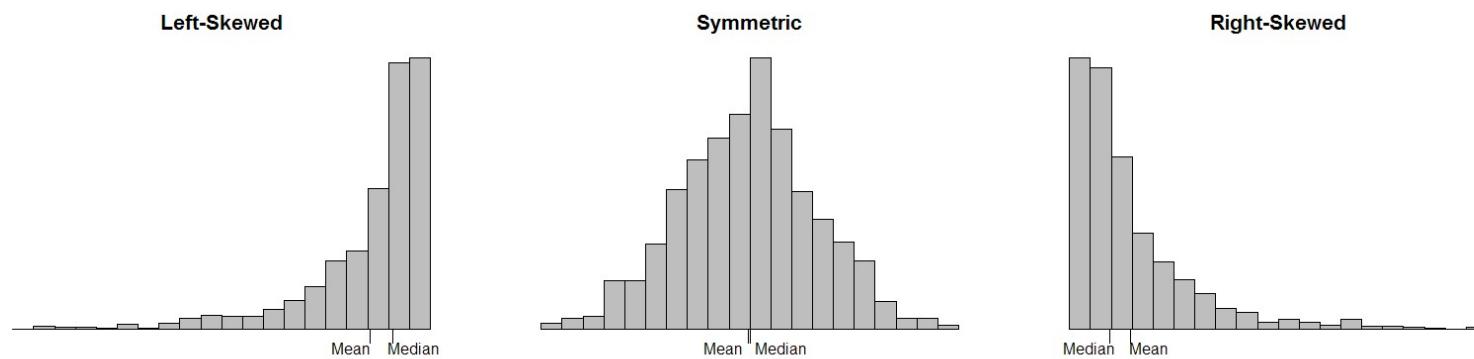
Corrosion of reinforcing steel is a serious problem in concrete structures located in environments affected by severe weather conditions. For this reason, researchers have been investigating the use of reinforcing bars made of composite material. One study was carried out to develop guidelines for bonding glass-fiber-reinforced plastic rebars to concrete. Consider the following 48 observations on measured bond strength:

11.5	3.4	9.3	10.7	4.1	4.2	3.6	3.8
5.7	5.5	5.1	12.6	7.0	8.2	9.3	7.6
3.6	9.9	25.5	4.8	8.5	3.7	3.9	3.6
5.2	5.2	13.8	6.6	8.9	17.1	14.2	5.0
12.1	20.6	5.2	15.2	3.8	4.0	3.6	7.8
5.4	5.1	6.2	13.1	13.4	10.7	5.6	4.9

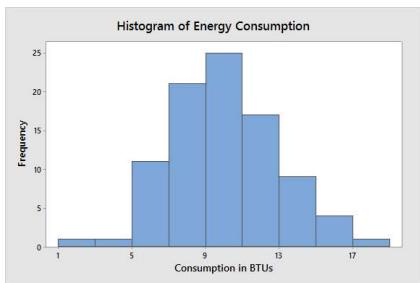


Interpreting Graphical Displays

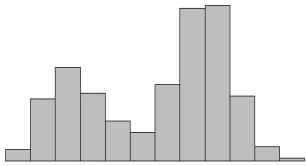
1. Shape



2. Modality (major peaks in the graph)



Is the graph unimodal (one peak)?



Bimodal (two peaks)?

Multimodal (three or more peaks)?

3. Gaps and/or outliers?

Gaps in the tails of the graph suggest possible outliers in the data set.

Gaps in the center of the graph are also of interest and the reason for unrepresented values in the center of the data should be investigated. One possible explanation is that the data represents more than one population and a gap indicates the separation of the sampled values according to the populations sampled.

4. Central (representative) value

We usually look for a number that represents the center of the data. Such a value is often used as a representative value of the entire data set.

Look for a value at which the display would “balance”.

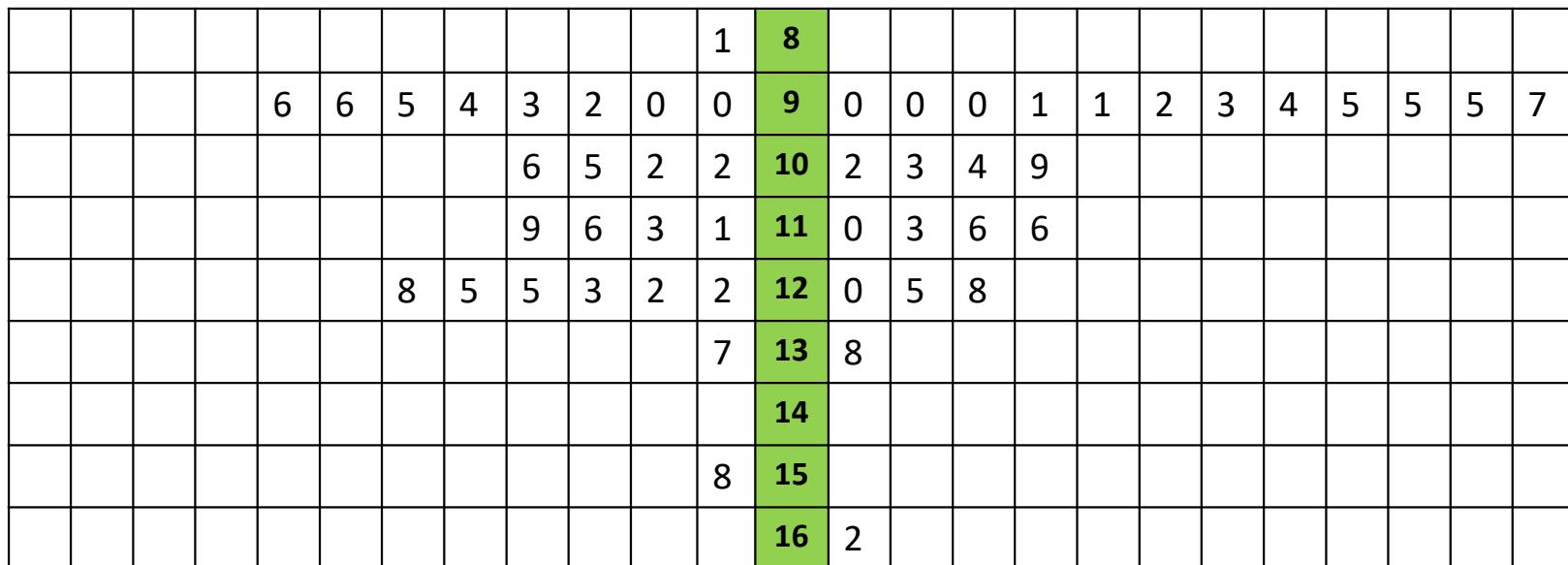
5. Measure of variability (or spread)

We will be looking at values computed from the data that can be used to indicate how spread out the data values are, but by looking at a graphical display, we should be able to get a general idea of the degree of spread in the data. When comparing two or more sets of data, the graphical displays will tell us at a glance which sets are more variable than the others.

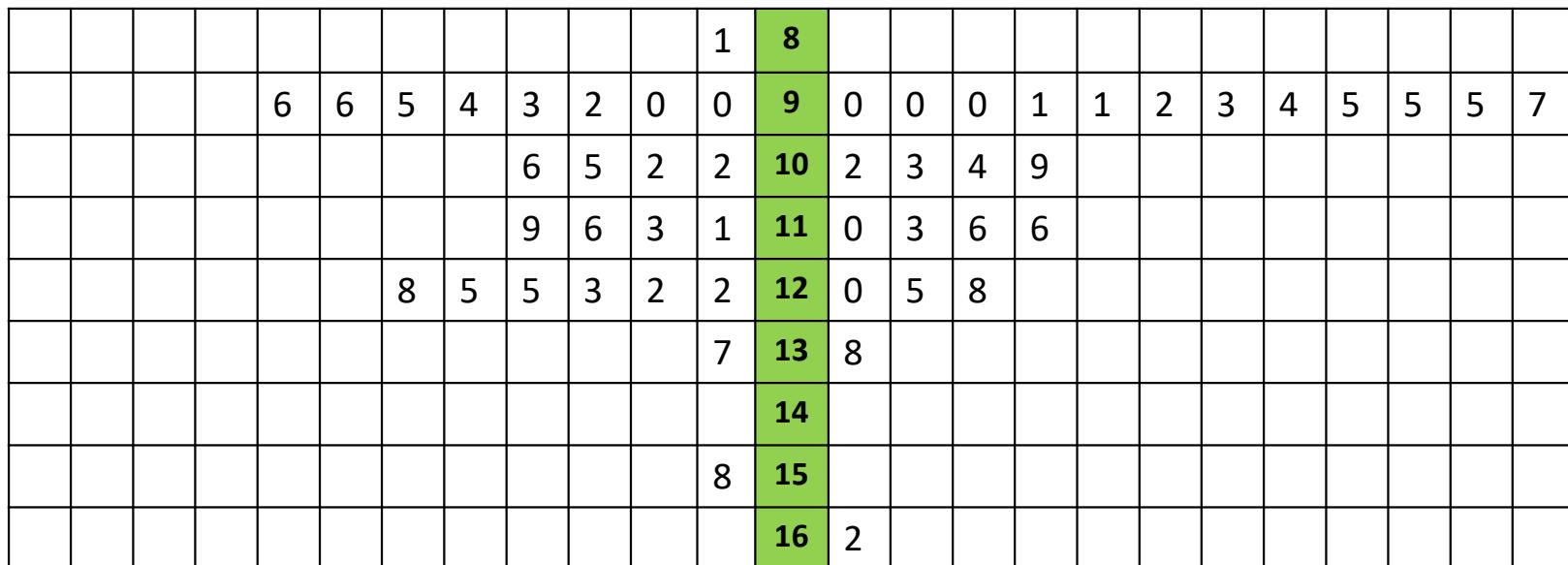
Let's revisit some of our graphical displays and interpret them keeping these five points of interest in mind.

We should use the terminology discussed in these points of interest in our interpretations.

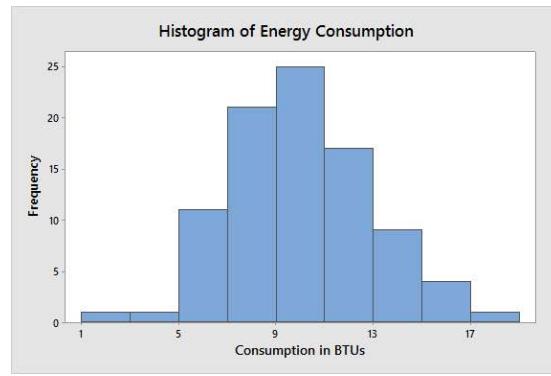
French Film Lengths



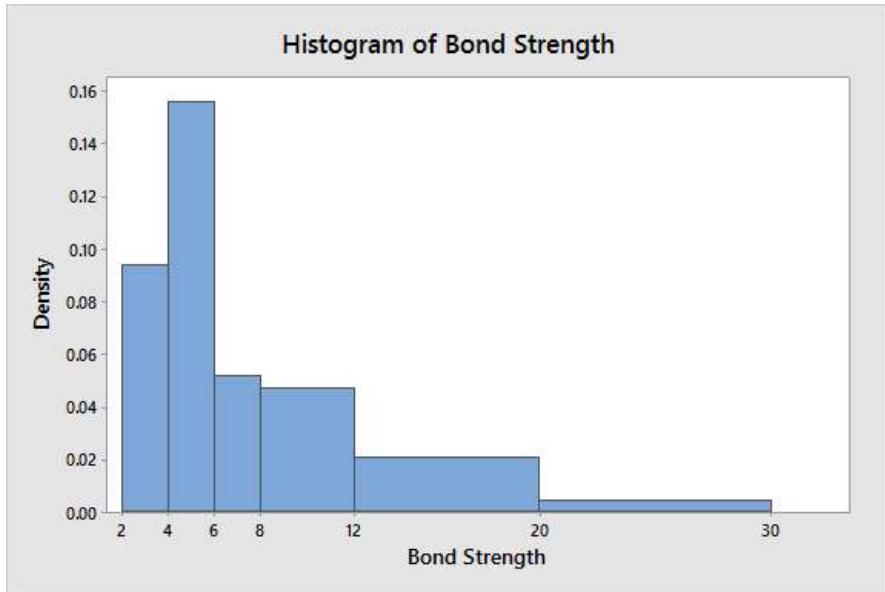
American Film Lengths



French Data: This graph looks reasonably symmetric, perhaps slightly right-skewed, and bimodal. The length of 158 minutes may be an outlier. A central or representative value appears to be around 106 minutes. The spread of the data appears moderate. **American Data:** This graph is definitely right-skewed and unimodal. The length of 162 minutes looks like an outlier. A central value appears also to be around 106 minutes. The spread of the data appears moderate. **Comparison:** Without the possible outliers, the film lengths seem comparable between the two countries.



This graph is symmetric and unimodal. No gaps or outliers are present. The variability seems low to moderate with a central or representative value of around 10 BTUs.



This data set is right-skewed, unimodal, with no gaps or outliers. A central value looks to be around 9 or 10 and there appears to be moderate to high variability.

Numerical Measures for Summarizing a Data Set

Let X_1, X_2, \dots, X_n represent the n (numerical) values in the data set.

A *statistic* is anything which may be calculated from a dataset.

A *sample statistic* simply makes clear that it is derived from a sample.

Sample statistics provide numerical measures of center and variability as well as suggest certain shape characteristics of the data.

Sample Mean (\bar{X})

The most important feature of a dataset to describe is generally its location, or the location of its **center**.

The most commonly used statistic for center is the familiar average, or *sample mean*.

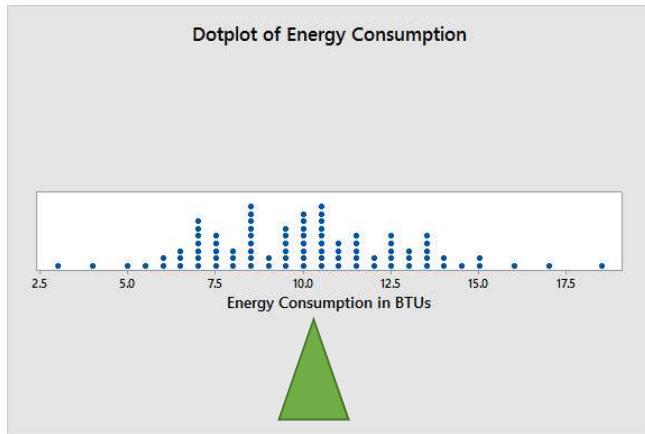
Definition: The *sample mean* of the data X_1, X_2, \dots, X_n is

$$\bar{X} = (\sum_{i=1}^n X_i) / n$$

7.87	11.12	8.58	12.91	12.28	14.24	11.62	7.73	7.15
9.43	13.43	8.00	10.35	7.23	11.43	11.21	8.37	12.69
7.16	9.07	5.98	9.60	2.97	10.28	10.95	7.29	13.38
8.67	6.94	15.24	9.58	8.81	13.60	7.62	10.49	13.11
12.31	10.28	8.54	9.83	9.27	5.94	10.40	8.69	10.50
9.84	9.37	11.09	9.52	11.29	10.36	12.92	8.26	14.35
16.90	7.93	11.70	18.26	8.29	6.85	15.12	7.69	13.42
10.04	13.96	12.71	10.64	9.96	6.72	13.47	12.19	6.35
12.62	6.8	6.78	6.62	10.30	10.21	8.47	5.56	9.83
7.62	4.00	9.82	5.20	16.06	8.61	11.7	9.76	12.16

$$\sum_{i=1}^{90} x_i = 903.460 \quad \text{and} \quad n = 90$$

$$\text{So } \bar{x} = \frac{903.460}{90} = 10.038$$



To understand how the mean works, suppose we were to take a very thin yardstick or similarly marked board, and place a small (equal) weight at the mark for each observation's value.

The mean may be thought of as the point where this system would balance.

Outliers

An *outlier* is an observation which is very different from the rest of the sample. For univariate data, this means it is much larger or much smaller than the rest.

Outliers should be carefully examined. Often they are the result of measurement or recording errors.

If so, they should be fixed or deleted. Correct but unusual values, however, should be kept.

Definition: A sample statistic is said to be **robust** if it is resistant to outliers, which means an outlier does not seriously impact the value of the statistic if at all.

The sample mean is *not* a robust measure of center.

Suppose the energy consumption of 18.26 were incorrectly entered as 1826. The sample mean would now be 30.1.

The Sample Median, a **robust** measure of center

Definition: The i^{th} **order statistic**, $X_{(i)}$, is the i^{th} smallest value when the X 's are sorted. The minimum is $X_{(1)}$, the second smallest $X_{(2)}$, and so on up to the maximum, $X_{(n)}$.

The **sample median**, \tilde{X} , is the middle of the sorted data.

- If n is odd, $\tilde{X} = X_{\left(\frac{n+1}{2}\right)}$.
- If n is even, $\tilde{X} = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2}$

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72
6.78	6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62
7.62	7.69	7.73	7.87	7.93	8.00	8.26	8.29	8.37
8.47	8.54	8.58	8.61	8.67	8.69	8.81	9.07	9.27
9.37	9.43	9.52	9.58	9.60	9.76	9.82	9.83	9.83
9.84	9.96	10.04	10.21	10.28	10.28	10.30	10.35	10.36
10.40	10.49	10.50	10.64	10.95	11.09	11.12	11.21	11.29
11.43	11.62	11.70	11.70	12.16	12.19	12.28	12.31	12.62
12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43	13.47
13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

$$\tilde{x} = [x_{(45)} + x_{(46)}]/2 = (9.83+9.84)/2 = 9.835$$

The sample median divides a data set into a lower 50% of the data and an upper 50%.

The sample median is very robust; changing one or a few observations won't change it much, if at all.

If 18.26 had been entered as 1826, the median would not be affected since the two "middle" observations are the same.

Trimmed mean (compromise between \bar{X} and \tilde{X})

To compute a trimmed mean, we choose a trimming percentage, $p\%$. We remove $p\%$ of the smaller values in the data set and $p\%$ of the larger values in the data set and then compute the mean of the remaining data values. We denote this trimmed mean as $\bar{x}_{tr(p)}$.

Am: 94 90 95 93 128 95 125 91 104 116 162
102 90 110 92 113 116 90 97 103 95 120 109
91 138

Ordered: 90 90 90 91 91 92 93 94 95 95 95 97
102 103 104 109 110 113 116 116 120 125 128
138 162

Consider a 4% trimmed mean, $\bar{x}_{tr(4)}$. Since 4% of 25 is 1, we remove the smallest and largest values from the data set and compute the sample mean of the remaining values which gives $\bar{x}_{tr(4)} = 104.65$. This is smaller than $\bar{x} = 106.36$ because we have compensated for the effect of the outlier 162.

Measures of Variability

After center, the second-most-used feature to describe a sample is its *variability*, or *spread*.

The simplest measure of variability is the *range*, the difference between the maximum and minimum values

$$R = \max(X) - \min(X)$$

Unfortunately, the range both wastes most of the data, and is maximally non-robust, using only the two extreme data points, so it is rarely used.

A better solution looks at the *deviations from the mean*, $X_i - \bar{X}$. This removes the effect of the mean (location), and looks only at the variability around the mean.

One option: Look at the average deviation from the mean.

Problem: Positive deviations cancel out negative ones, and the average deviation from the mean is always 0.

We could take absolute values of the deviations, and compute an average absolute deviation from the mean, but for a few theoretical reasons, it's better to look at the squared deviations instead, and compute a (pseudo) average squared deviation from the mean.

The *sample variance*, s^2 , measures the spread of a dataset.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Definition: The *sample standard deviation*, s , is the square root of the sample variance.

Use of the definition formula is tedious, as it requires finding and squaring each of the n deviations from the mean. It is usually simpler to calculate s^2 using the following computation formula.

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 \right)$$

Let's return to the energy consumption data set.

$$\sum_{i=1}^{90} x_i^2 = 9801.39$$

and

$$\left(\sum_{i=1}^{90} x_i \right)^2 = 903.460^2 = 816,239.9716$$

So

$$\begin{aligned}s^2 &= \frac{1}{89} \left(9801.39 - \frac{1}{90} (816,239.9716) \right) \\&\approx \frac{1}{89} (9801.39 - 9069.3330) \\&= \frac{1}{89} (732.0570) \\&\approx 8.2254\end{aligned}$$

and

$$s = \sqrt{s^2} \approx 2.868$$

The sample variance and standard deviation are measures of the spread of a dataset. They will also be used to **estimate** the variance and standard deviation of the underlying population or distribution.

Like the sample mean, they are **not** robust statistics. They are not resistant to outliers.

If the 18.26 BTU value had been incorrectly entered as 1826 we would have $s^2 = 36,652.5$ and $s = 191.4$.

Suppose $Y_i = a + bX_i$ for $i = 1, \dots, n$. Then

- $\bar{Y} = a + b\bar{X}$
- $S_Y^2 = b^2 S_X^2$

and

- $S_Y = |b|S_X$

Am: 94 90 95 93 128 95 125 91 104 116 162 102
90 110 92 113 116 90 97 103 95 120 109 91 138

Let these film lengths be our X's. The film lengths are given in minutes. Suppose we wanted to instead give their lengths in hours. Then we would be looking at

$$Y = \frac{1}{60}X$$

Since $\bar{x} = 106.36 \text{ min}$, $s_x^2 = 319.66 \text{ sq min}$, and $s_x = 17.88 \text{ min}$, it follows that $\bar{y} = 1.773 \text{ hr}$, $s_y^2 = .089 \text{ sq hr}$, and $s_y = .298 \text{ hr}$.

Percentiles or Quantiles:

Let $0 < p < 1$. The $(p \times 100)$ th percentile of a set of data is the number $q(p)$ (not necessarily in the set) for which $(p \times 100)\%$ of the observations in the data set are no larger than $q(p)$ and $((1-p) \times 100)\%$ are no smaller than $q(p)$.

This value simply divides the data set into a lower $(p \times 100)\%$ portion and an upper $((1-p) \times 100)\%$ portion.

To compute $q(p)$: (Three Step Procedure)

Step 1: Order the observations from smallest to largest.

Step 2: Compute $p \times (n + 1) = a.b$, where n is the number of observations in the data set.

Step 3: $q(p) = x_{(a)} + .b(x_{(a+1)} - x_{(a)})$ where $x_{(a)}$ and $x_{(a+1)}$ are the a^{th} and $(a + 1)^{st}$, respectively, observations in the ordered data set.

Example : Suppose we have the following data:

117 124 132 138 148 157 170 182 199 211

$n = 10$

Let $p = .39$. To find the 39th percentile:

Step 1: Data is ordered.

Step 2: $.39 \times (10 + 1) = .39(11) = 4.29$.

Step 3: We have $a = 4$ and $.b = .29$. The fourth and fifth observations in the ordered data set are $x_{(4)} = 138$ and $x_{(5)} = 148$. So $q(.39) = 138 + .29(148 - 138) = 140.9$.

Our robust measure of center will involve the computation of the **quartiles (25th and 75th percentiles)**. Note: The median would correspond to the 50th percentile.

Quartiles:

First Quartile=25th Percentile=q(.25)= Q_1

Third Quartile=75th Percentile=q(.75)= Q_3

Sorted energy consumption data

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72
6.78	6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62
7.62	7.69	7.73	7.87	7.93	8.00	8.26	8.29	8.37
8.47	8.54	8.58	8.61	8.67	8.69	8.81	9.07	9.27
9.37	9.43	9.52	9.58	9.60	9.76	9.82	9.83	9.83
9.84	9.96	10.04	10.21	10.28	10.28	10.30	10.35	10.36
10.40	10.49	10.50	10.64	10.95	11.09	11.12	11.21	11.29
11.43	11.62	11.70	11.70	12.16	12.19	12.28	12.31	12.62
12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43	13.47
13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

.25(91) = 22.75 So a=22 and .b=.75 and $Q_1 = 7.87 + .75(7.93 - 7.87) = 7.915$

.75(91) = 68.25 So a=68 and .b=.25 and $Q_3 = 12.16 + .25(12.19 - 12.16) = 12.1675$

Definition: The *sample interquartile range* is a robust measure of spread found as the difference between the sample quartiles, $IQR = Q_3 - Q_1$.

For the energy consumption data we would have

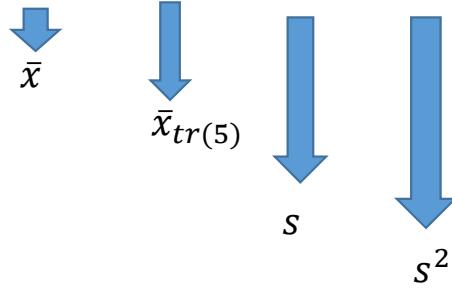
$$IQR = 12.1675 - 7.915 = 4.2525 \approx 4.253$$

If the 18.26 data point were incorrectly entered as 1826, the IQR would not be affected.

Minitab 17 Computed Values

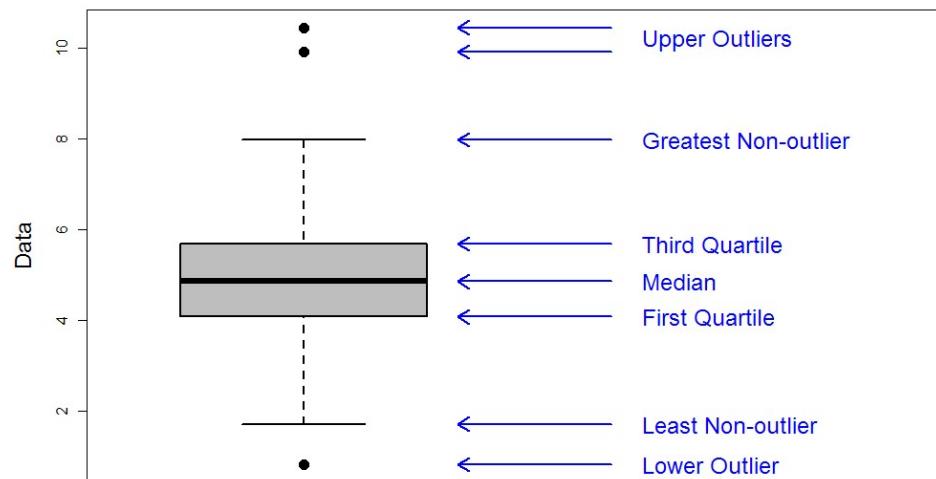
Descriptive Statistics: BTU

Variable	N	Mean	TrMean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum	IQR
BTU.In	90	10.038	9.978	2.868	8.225	2.970	7.915	9.835	12.168	18.260	4.253



Boxplots

Definition: A *boxplot* (or box and whisker plot) is another graphical tool for displaying a sample.



For boxplots, outliers are usually defined as any values below

$$Q_1 - 1.5IQR$$

or above

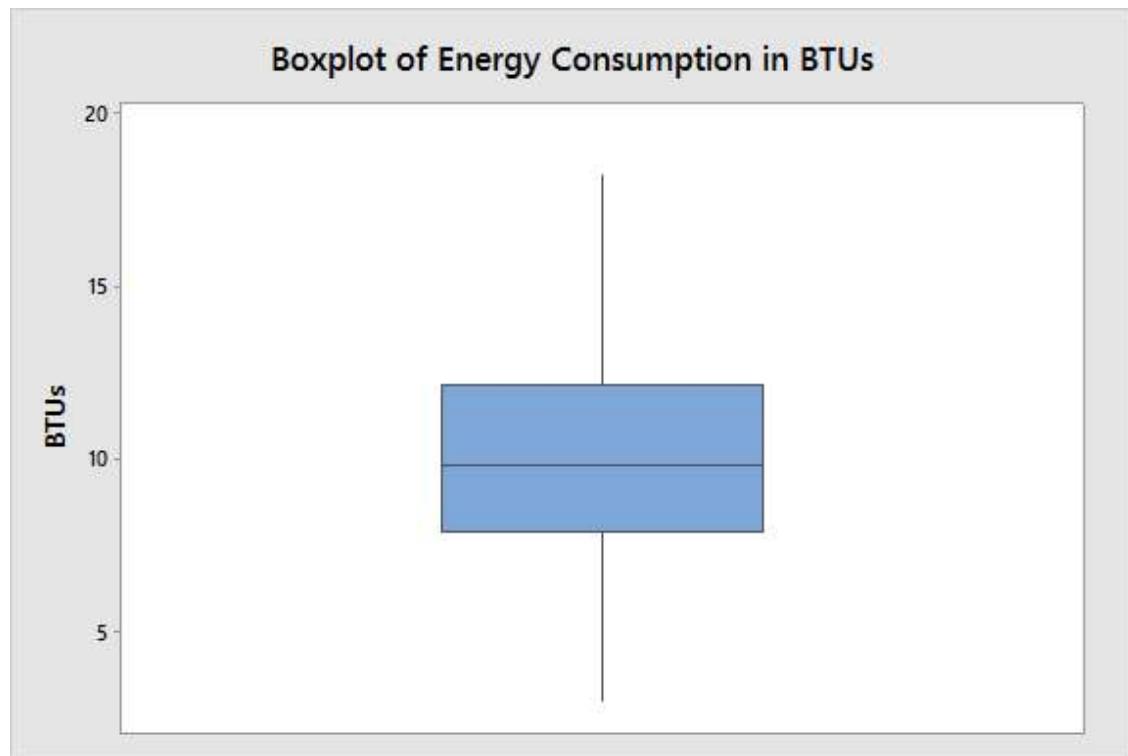
$$Q_3 + 1.5IQR$$

These values are referred to as the **inner fences**. Replacing the 1.5 with 3 in these formulas yields the **outer fences**.

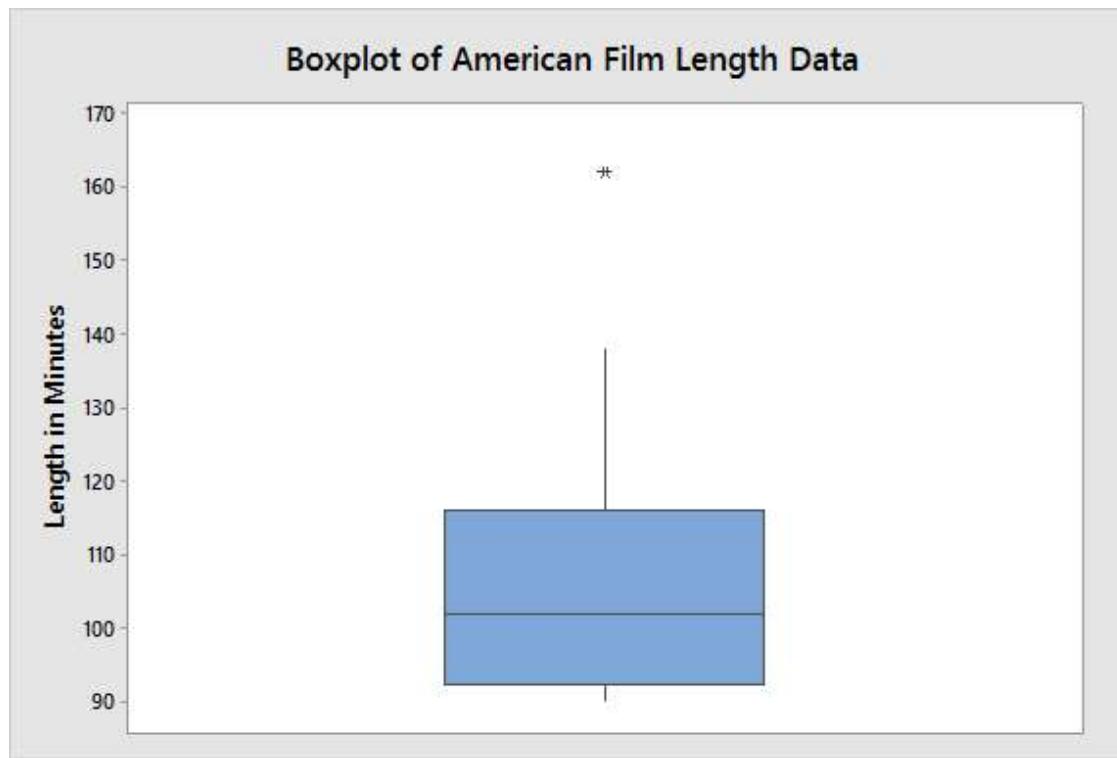
Mild (**extreme**) outliers are **between 1.5 IQR and 3 IQR** (**more than 3 IQR**) below Q1 or above Q3. So mild outliers are between the inner and outer fences and extreme outliers are beyond the outer fences.

The **whiskers** go from the quartiles to the least and greatest values among the non-outliers.

Energy Consumption Boxplot Constructed in Minitab 17

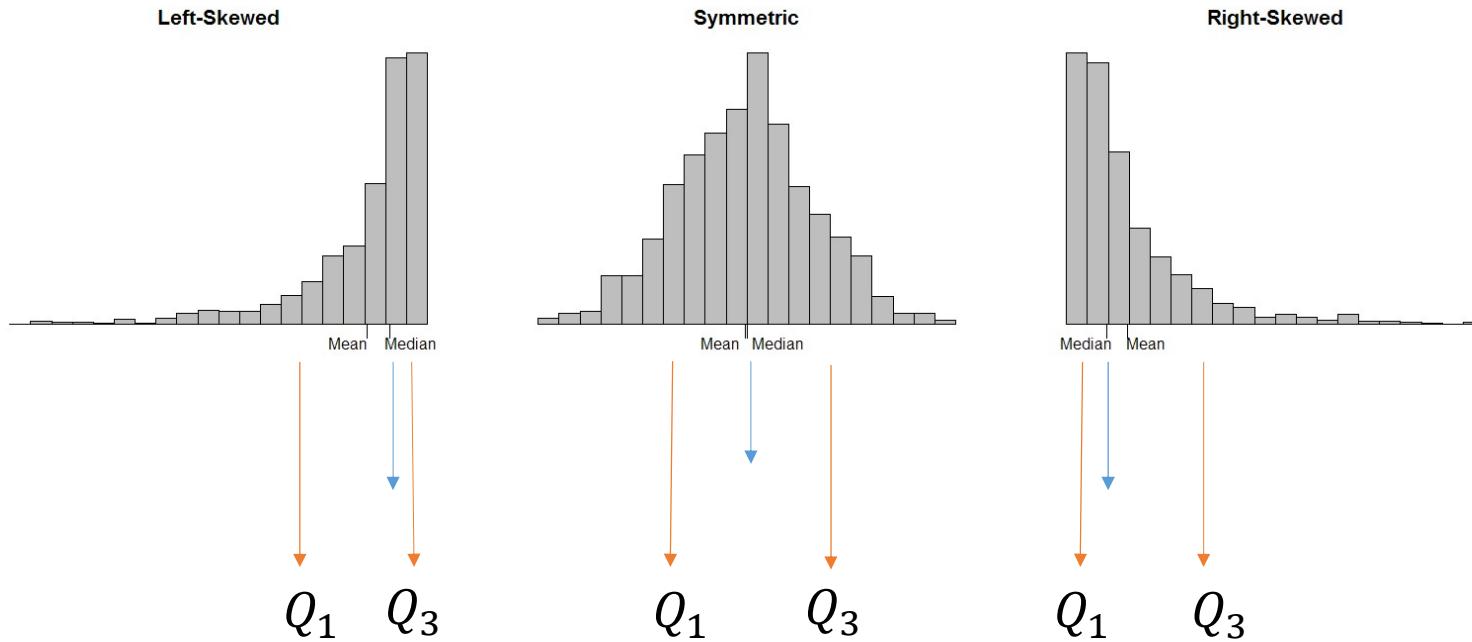


Boxplot of American Film Length Data Set



Interpreting Boxplots

1. Skewness may be indicated when the line inside the box, which is placed at the location of the median, is much closer to the top of the box than the bottom or much closer to the bottom than the top. Skewness may also be indicated when one whisker is much longer than the other.



2. Outliers are indicated with asterisks. A quick check using the guidelines for classifying outliers can determine if they are mild or severe.
3. The median can be used as a measure of center for the data set.
4. Variability or spread can be assessed by looking at the lengths of the box and whiskers.

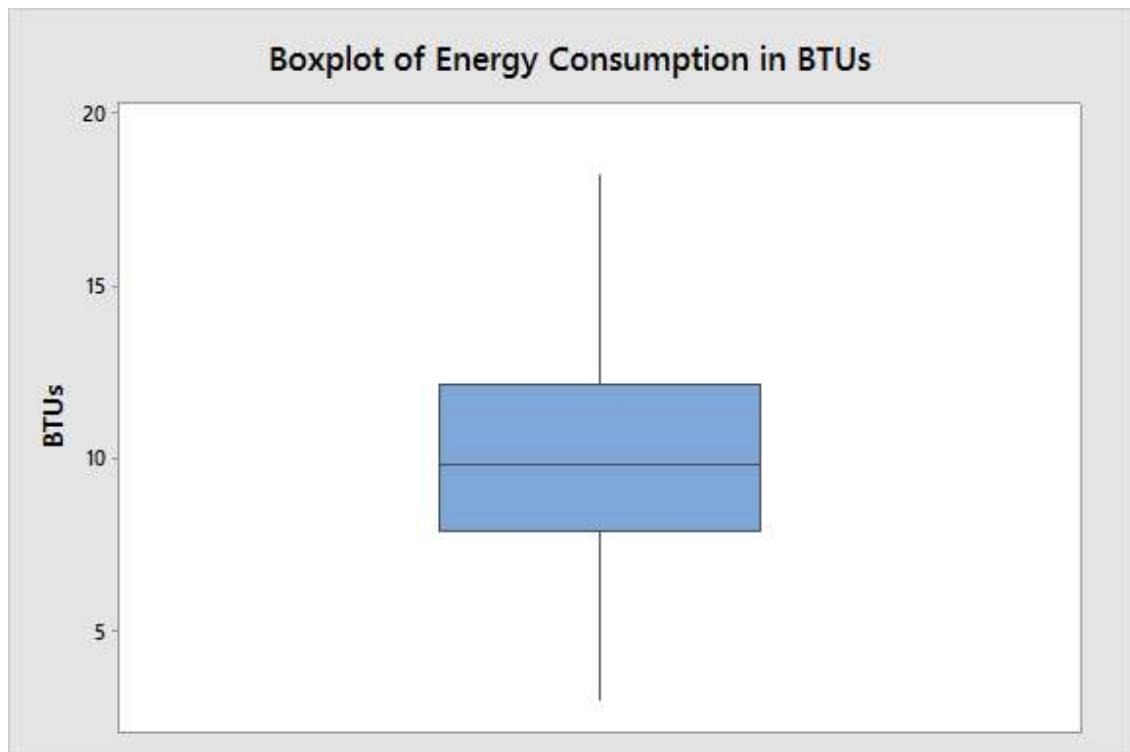
Some properties of the data set cannot be determined from a boxplot.

1. There is no way to assess modality with a boxplot.
2. If the line at the location of the median is near the center of the box, that suggests a symmetric distribution. It does not necessarily mean that the distribution is bell-shaped or *normal*.

When reporting suitable measures of center and variability for a set of data I recommend the following:

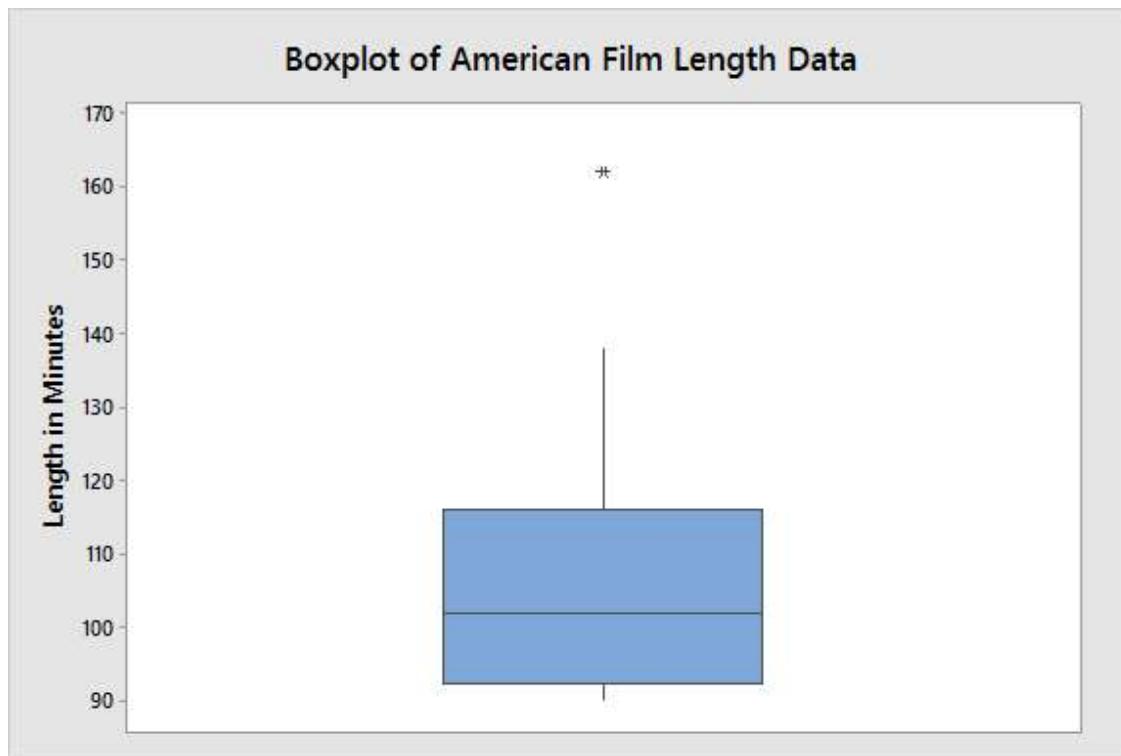
- For data sets that have outliers and/or display strong skewness, use the robust measures of the sample median and interquartile range.
- For data sets that are reasonably symmetric, use the sample mean and sample variance (or sample standard deviation).

Energy Consumption Boxplot Constructed in Minitab 17

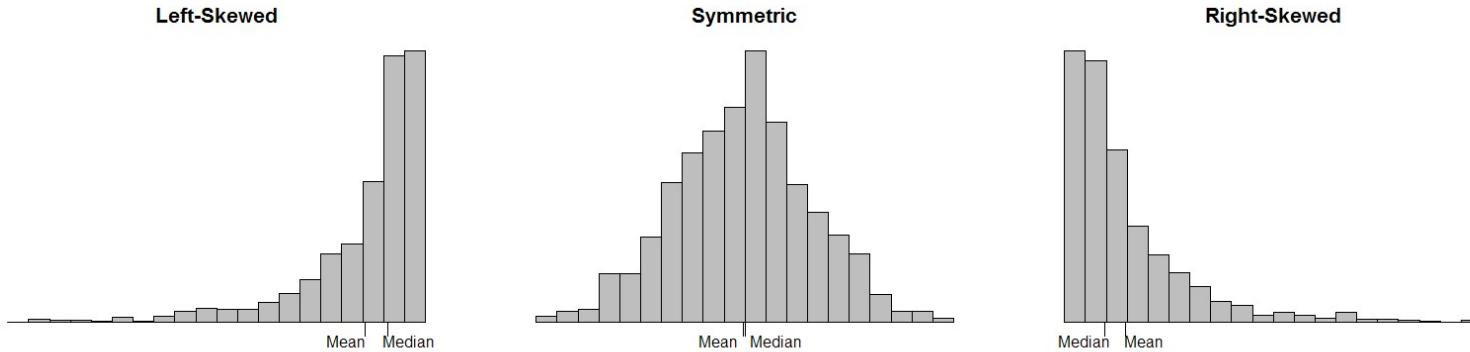


This boxplot shows a symmetric distribution with no outliers. A good measure of center would be the mean of 10.038 and a suitable measure of spread would be the standard deviation of 2.868.

Boxplot of American Film Length Data Set



With the median being closer to the first quartile (bottom of box) and the long upper and short lower whisker, I would classify this as a right-skewed distribution. The outlier 162 is between the inner and outer fences and would therefore be mild. Suitable measures of center and spread would be the sample median and IQR, 102 and 23.5, respectively.



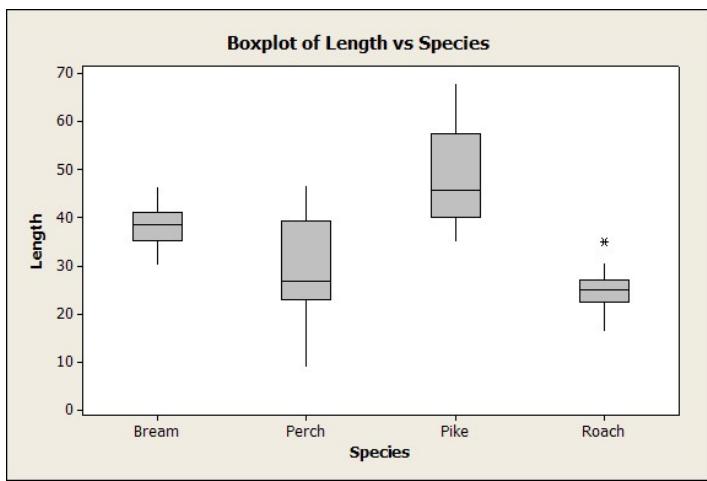
Skewness is also a possibility if the **median** is much smaller or much larger than the **mean**. The mean will be pulled in the direction of the skewness as illustrated in the histograms above. Symmetric distributions will have means and medians that are approximately the same.

Boxplots are much less informative than histograms for a single distribution, so the histogram is usually preferable.

On the other hand, comparing histograms is difficult, while comparing boxplots is easy.

Use boxplots to compare 2-20 (or more) distributions.

Samples from four species of fish were collected and the fish lengths were recorded. Side-by-side boxplots were constructed for the four sets of data.



Summary interpretation: Bream and Pike seem to have comparable central values but the Pike lengths are much more variable than Bream. Likewise, Perch and Roach have comparable central values, but Perch is much more variable. Perch and Pike display right skewness in the central 50% of the data whereas Roach and Bream are much more symmetric.

General Statement of Order of Magnitude: Pike, Bream, Perch, Roach

Probability

Definition: *Probability* is the branch of mathematics dealing with chance, randomness, and uncertainty.

Probability provides most of the mathematical foundation for inferential statistics.



"I wish we hadn't learned probability 'cause I don't think our odds are good."

Definition: A situation for which the outcome cannot be determined in advance is called an *experiment*.

Examples:

- The roll of a die.
- The draw of a card.
- The lifetime of an electronic component.
- The strength of a weld.

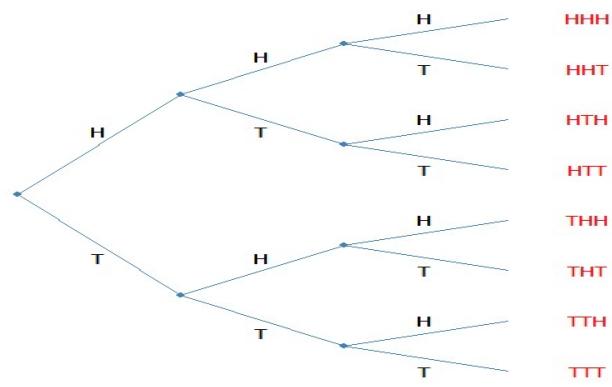
Definition: The *sample space*, S , of an experiment is the set of all possible outcomes.

Examples:

- Die: $S = \{1, 2, 3, 4, 5, 6\}$
- Card: $S = \{\text{AH}, \text{2H}, \dots, \text{KH}, \text{AS}, \text{2S}, \dots, \text{KS}, \text{AD}, \text{2D}, \dots, \text{KD}, \text{AC}, \text{2C}, \dots, \text{KC}\}$
- Component and Weld Strength: $S = [0, \infty)$ (units of measure would differ)

An experiment with several steps can be visually represented by a *tree diagram*:

Example: Toss a coin three times:



$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Definition: Set A is a *subset* of set B ($A \subset B$) if every element of A is also in B .

Example: $S = \{1, 2, 3, 4, 5, 6\}$

$$A = \{1, 3, 5\} \subset S$$

$$B = \{1, 2, 6, 7\} \not\subset S$$

Every set is a subset of itself.

The empty set, \emptyset , consisting of no elements, is a subset of every set.

Definition: Any “interesting” subset of the sample space can be called an *event*.

Examples:

- Die: $A = \text{“odd numbers”} = \{1, 3, 5\}$
- Card: $B = \text{“drawn card is an ace”} = \{\text{AH}, \text{AS}, \text{AD}, \text{AC}\}$
- Component: $C = \text{“component exceeds 1 year warranty”} = (1, \infty)$
- Coin tosses: $D = \text{“all three tosses are the same”} = \{\text{TTT}, \text{HHH}\}$

The individual outcomes which make up S are sometimes called *simple events*. Events consisting of more than one outcome are called *compound events*.

Creating New Events Using Set Arithmetic

For subsets of S , A and B ($A \subset S, B \subset S$):

1) The **union** of A and B ($A \cup B$) is the set consisting of all elements found in A , B , or both.

Keyword: **or**

Example: $S = \{1, 2, 3, 4, 5, 6\}$

- $A = \{1, 3, 5\} \subset S$
- $B = \{1, 2, 3\} \subset S$
- $A \cup B = \{1, 2, 3, 5\}$

2) The *intersection* of A and B ($A \cap B$) is the set consisting of all elements found in *both* A and B .

Keywords: **and, both**

Example: $S = \{1, 2, 3, 4, 5, 6\}$

- $A = \{1, 3, 5\}$
- $B = \{1, 2, 3\}$
- $A \cap B = \{1, 3\}$

3) The *complement* of A (A^c) is the set consisting of all elements of S *not* found in A .

Keyword: **not**

Example: $S = \{1, 2, 3, 4, 5, 6\}$

$$A = \{1, 3, 5\}$$

$$A^c = \{2, 4, 6\}$$

4) Sets A and B are said to be *mutually exclusive* or *disjoint* if there are no common elements in both A and B . That is, if $A \cap B = \emptyset$ (the empty set).

Example: $S = \{1, 2, 3, 4, 5, 6\}$

- $A = \{1, 3, 5\}$
- $C = \{4, 6\}$
- $A \cap C = \emptyset$, so A and C are mutually exclusive.

Example: Three coin tosses.

$$S = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}$$

Let A = “First toss is a head” = $\{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}\}$

Let B = “Last toss is a head” = $\{\text{HHH}, \text{HTH}, \text{THH}, \text{TTH}\}$

What simple events make up the event A and B ?

A or B ?

Not A ?

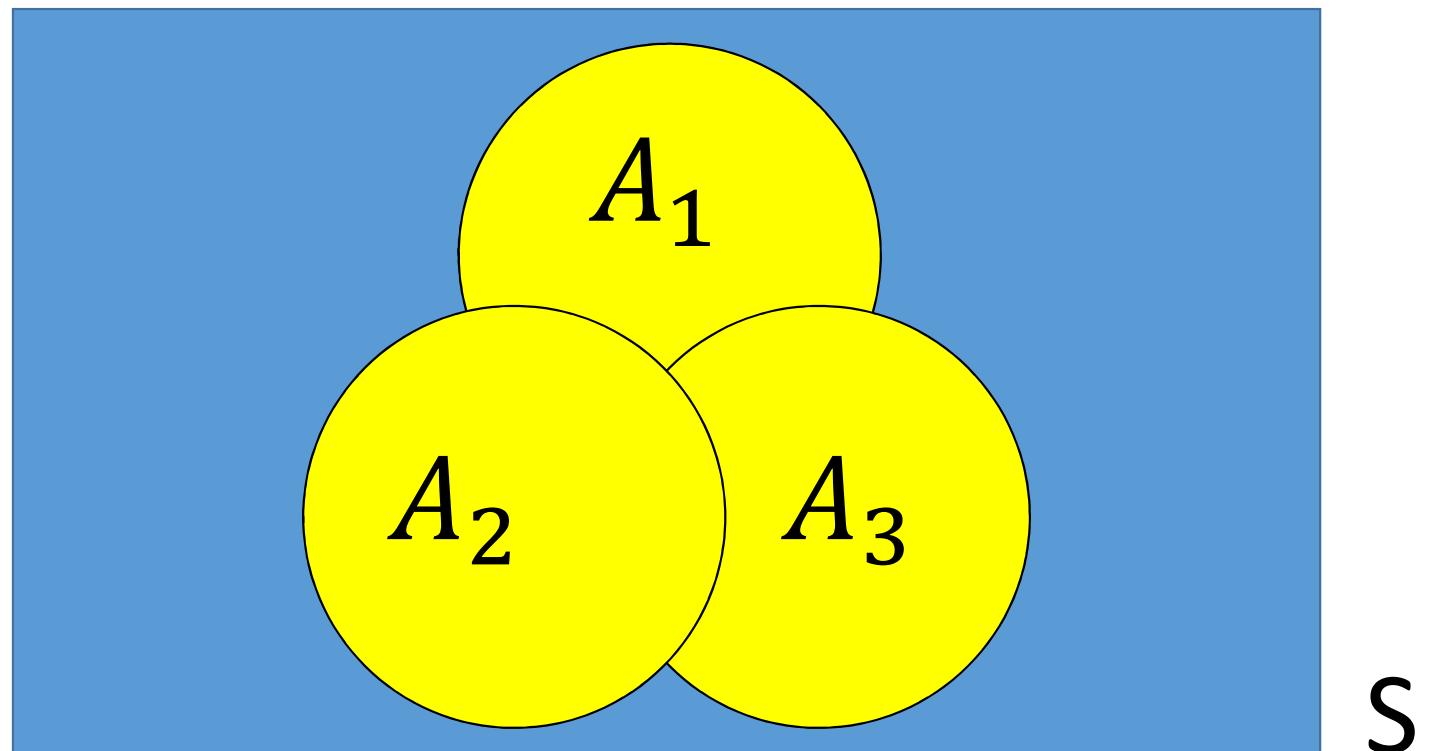
Are A and B mutually exclusive?

Example:

An engineering construction firm is currently working on power plants at three different sites. Let A_i denote the event that the plant at site i is completed by the contract date. Use the operations of union, intersection, and complementation to describe each of the following events in terms of A_1 , A_2 , and A_3 , draw a Venn diagram, and shade the region corresponding to each one.

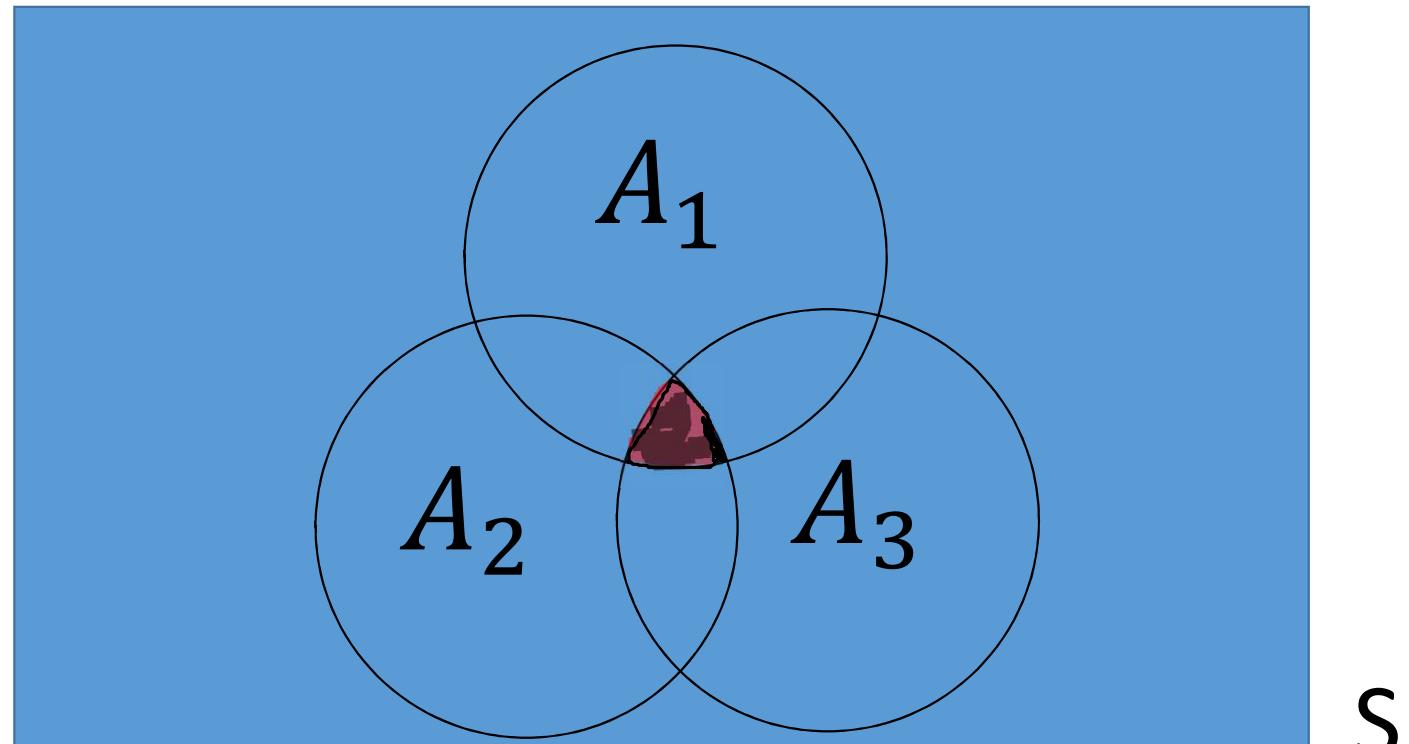
a.) At least one plant is completed by the contract date.

$$A_1 \cup A_2 \cup A_3$$



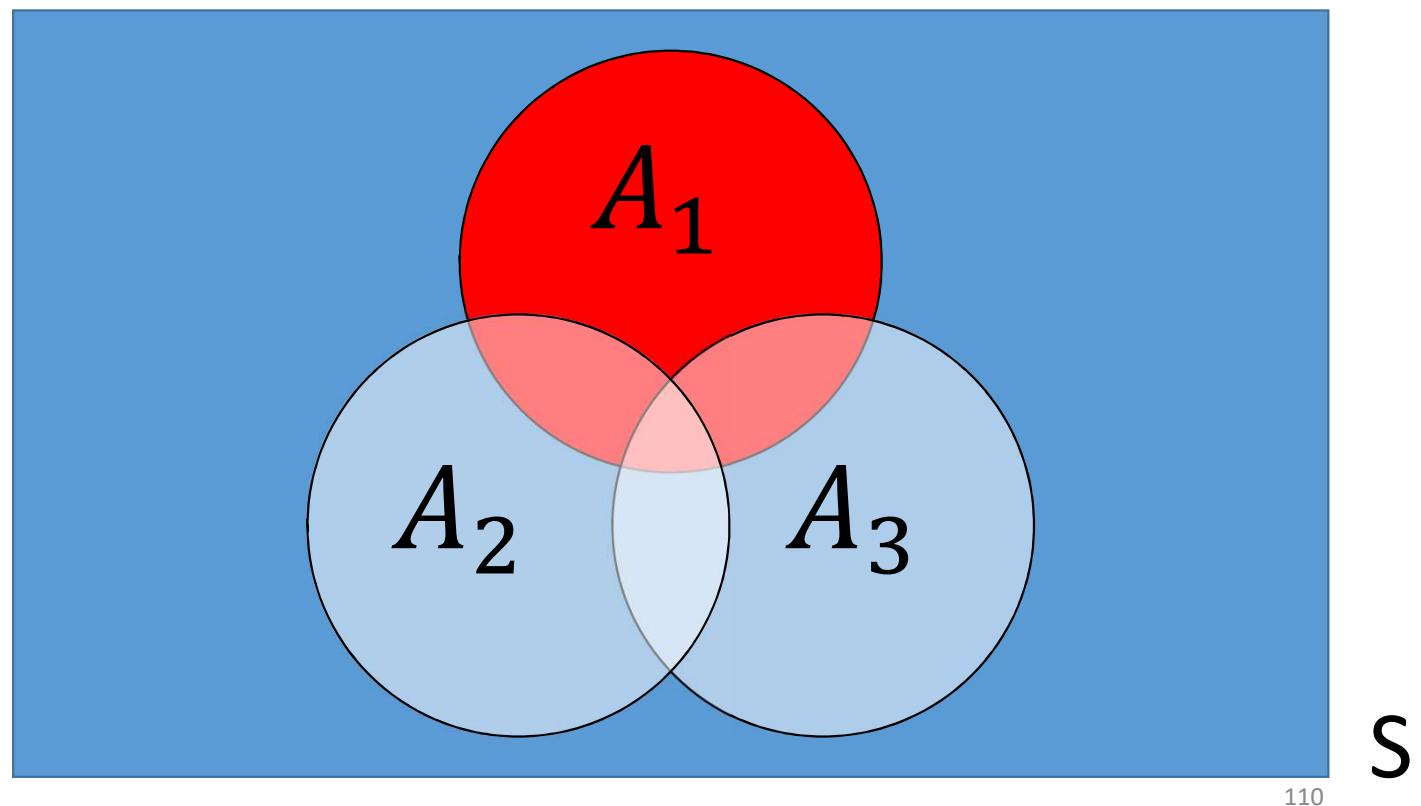
b.) All plants are completed by the contract date.

$$A_1 \cap A_2 \cap A_3$$



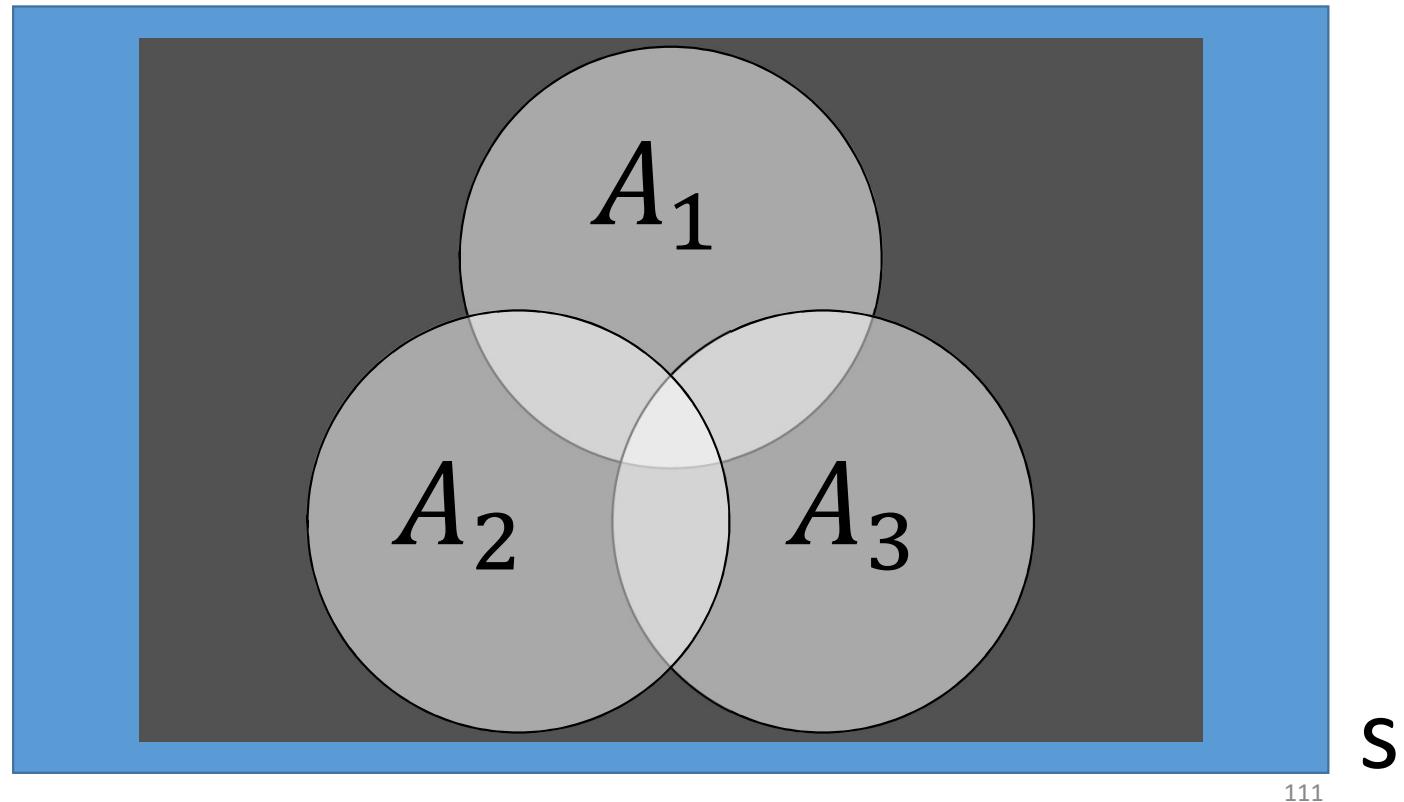
c.) Only the plant at site 1 is completed by the contract date.

$$A_1 \cap A_2^c \cap A_3^c$$



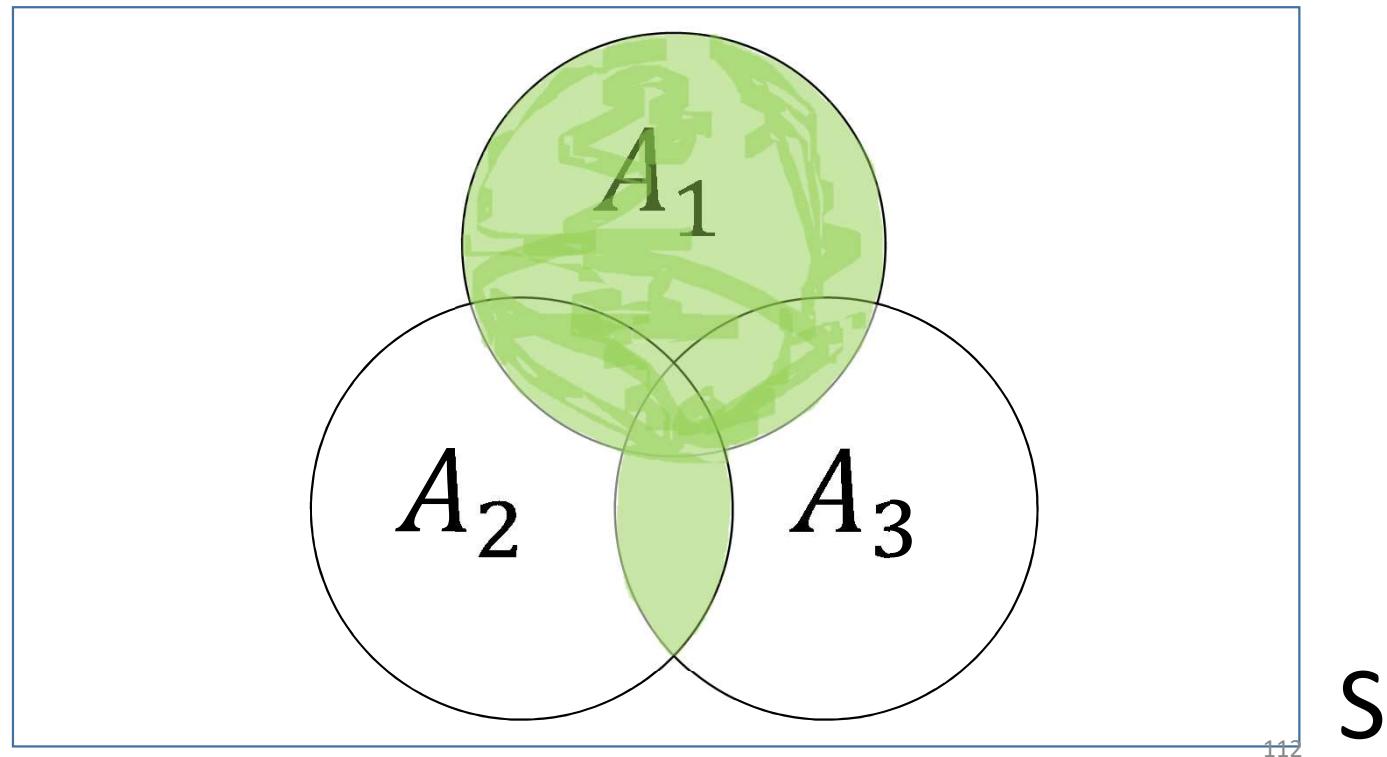
d.) Exactly one plant is completed by the contract date.

$$\begin{aligned} & [A_1 \cap A_2^c \cap A_3^c] \\ & \cup [A_1^c \cap A_2 \cap A_3^c] \\ & \cup [A_1^c \cap A_2^c \cap A_3] \end{aligned}$$



e.) Either the plant at site 1 or both of the other two plants are completed by the contract date.

$$A_1 \cup (A_2 \cap A_3)$$



Axioms of Probability

Definition: A *probability function* $P(\cdot)$ is a function from the subsets of S (events) to the real numbers which satisfies the following *axioms of probability*:

1. $P(S) = 1$.
2. $0 \leq P(A) \leq 1$ for any event A .
3. If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$. Or, more generally, $P(A_1 \cup A_2 \cup A_3 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$ for any collection of mutually exclusive events A_1, A_2, A_3, \dots .

Experiment: Roll a fair die.

$$P(1) = 1/6, P(2) = 1/6, P(3) = 1/6, P(4) = 1/6, P(5) = 1/6, P(6) = 1/6.$$

Probabilities of bigger events are found by axiom 3:



Equally Likely Outcomes

If S consists of N equally likely outcomes, and event A consists of k of them, then $P(A) = k/N$.

Example: Draw a card at random from a standard deck (52 cards, 13 spades). What is the probability of drawing a spade?

Example: A shipment of 1000 hard drives contains 6 which do not work. If we draw one at random, what is the probability of selecting a defective drive?

$$P(1) = 1/4, P(2) = 1/4, P(3) = 1/8, P(4) = 1/8, P(5) = 1/8, P(6) = 1/8.$$

So



$P(A)$ measures the likelihood associated with the event A. One way to think about this is in terms of long-term likelihood.

If the experiment is repeated many times, the proportion of times that A will occur or be observed is roughly $P(A)$.

Example: Toss a possibly biased coin 10000 times and count the number of heads that occur. Suppose this number is 8000. Then it is safe to assume that $P(H) \approx 8000/10000 = .8$.

Additional Properties of Probability

The axioms of probability imply some additional properties:

1. For any event A , $P(A^c) = 1 - P(A)$.

This is sometimes called the “complementary events rule”, or the “opposites rule”.

Show:

2) For any events A and B ,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This is sometimes called the “general addition rule”.

Show:

Example: Roll a fair die.

$$P(1) = 1/6, P(2) = 1/6, P(3) = 1/6, P(4) = 1/6, P(5) = 1/6, P(6) = 1/6$$

$$A = \{1, 3, 5\}, P(A) = 3/6 = 1/2, \quad B = \{1, 2\}, \quad P(B) = 2/6 = 1/3$$

$$P(A^c) =$$

$$A \cap B = \{1\}, \quad P(A \cap B) = 1/6$$

$$P(A \cup B) =$$

We don't need to know the entire probability function to use these.

Example: Lifetime of a component (T) in days. Suppose we know:

$$P(A) = P(T \leq 60) = .47 \quad P(B) = P(40 \leq T \leq 80) = .34$$

$$P(A \cap B) = P(40 \leq T \leq 60) = .26$$

Then:

$$P(T > 60)$$

$$P(\text{lifetime no more than } 80) = P(T \leq 80)$$

Example: Suppose the probability that an integrated circuit chip has defective etching is 0.12. The probability that the chip has a crack defect is 0.29. And the probability of both defects is 0.07.

Let C = “chip has a crack defect” and E = “chip has defective etching”

$$P(C) = .29, \quad P(E) = .12, \quad \text{and} \quad P(C \cap E) = .07$$

What is the probability the chip does not have defective etching?

E^c = “chip does not have defective etching”

What is the probability it has at least one defect?

$C \cup E$ = “chip has at least one defect”

What is the probability it has neither defect?

$C^c \cap E^c$ = “chip has neither of the defects” = $(C \cup E)^c$

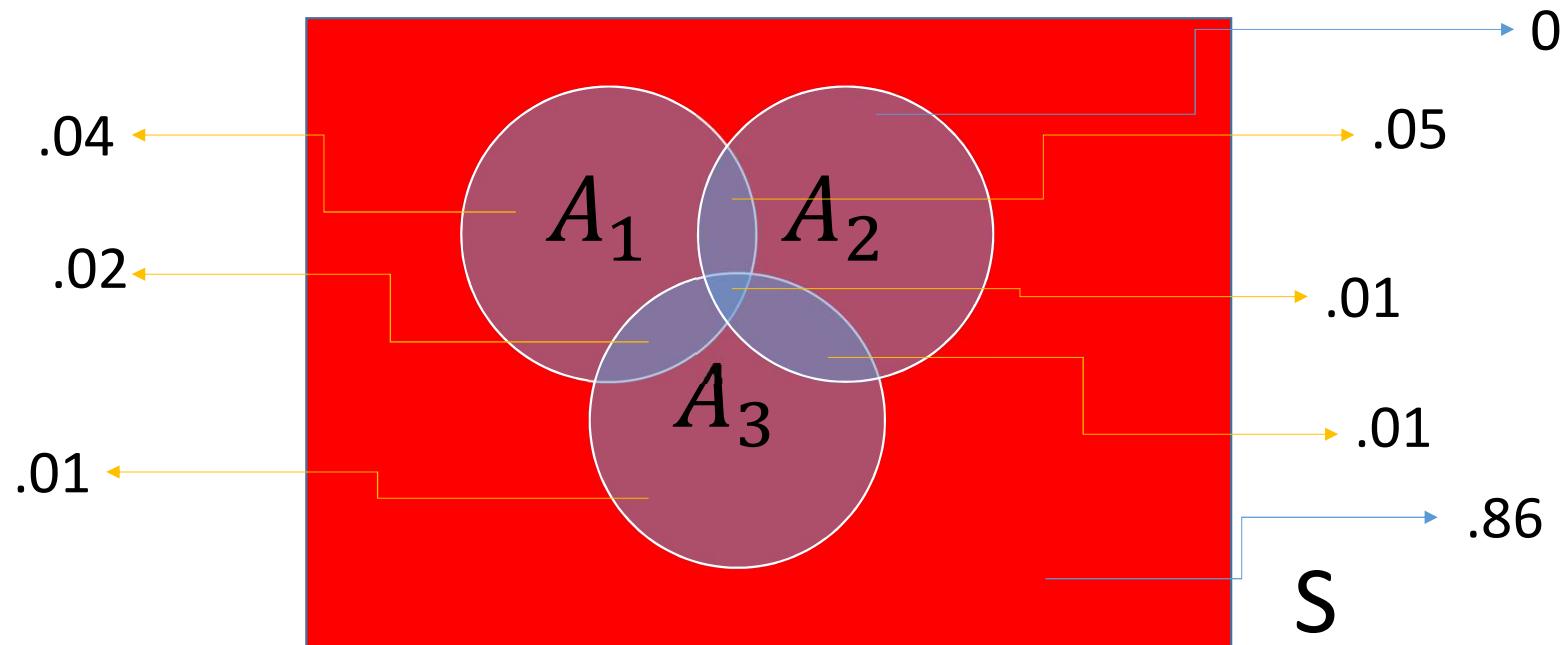
Note that using the formula

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

and solving for $P(A \cap B)$ we get

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

A certain system can experience three different types of defects. Let A_i ($i = 1, 2, 3$) denote the event that the system has a defect of type i . Suppose that $P(A_1) = .12$, $P(A_2) = .07$, $P(A_3) = .05$,
 $P(A_1 \cup A_2) = .13$, $P(A_1 \cup A_3) = .14$, $P(A_2 \cup A_3) = .10$,
 $P(A_1 \cap A_2 \cap A_3) = .01$



a.) What is the probability that the system does not have a type 1 defect?

b.) What is the probability that the system has both type 1 and type 2 defects?

c.) What is the probability that the system has both type 1 and type 2 defects but not a type 3 defect?

Venn diagram

d.) What is the probability that the system has at most two of these defects?

Conditional Probability

Suppose we have partial information about the outcome of an experiment. In particular, suppose we know that the event B has occurred.

We may use this information to revise the probability of another event, A .

We call the revised probability a *conditional probability*, as it depends on the condition of B being true.

Example: Roll a fair die. Let

$$A = \{1, 3, 5\} \quad P(A) = 3/6 = 1/2$$

$$B = \{1, 2, 3\} \quad P(B) = 3/6 = 1/2$$

$$P(A \cap B) = P(\{1, 3\}) = 2/6 = 1/3$$

If I roll the die and, without showing you, tell you event B has occurred (I rolled no greater than 3), now what is the probability of event A ?

Since B has occurred, the sample space reduces to $B : \{1, 2, 3\}$.

Two of the three possibilities are odd (in A), and the chances are still equal. So

Once we know the roll is 3 or less, the probability increases to $2/3$ that it's odd.

Definition: The *conditional probability of A given B* is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(undefined if $P(B) = 0$).

This is the probability, given that event B has occurred, that event A has also occurred.

Die Example:

Example from earlier slide:

$$P(E) = P(\text{defective etching}) = 0.12$$

$$P(C) = P(\text{crack defect}) = 0.29$$

$$P(C \cap E) = P(\text{etching and crack defects}) = 0.07$$

If a chip has a crack defect, what is the (conditional) probability that it also has defective etching?

What is the probability that a chip has a crack defect but satisfactory etching?

If a chip has a crack defect, what is the probability that it has satisfactory etching?

Note: $P(\cdot | B)$ is a valid probability function. In particular, for any event A,

$$P(A | B) = 1 - P(A^c | B),$$

just like

$$P(A) = 1 - P(A^c).$$

So in the previous example,

If a chip has defective etching, what is the probability that it also has a crack defect?

Note: There is no relationship between $P(C|E) = .583$ and $P(E|C) = .241$ which is generally the case.

No relationship between $P(A|B)$, $P(B|A)$.

Recall,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

provided $P(B) > 0$.

Therefore,

$$P(A \cap B) = P(A|B) P(B).$$

We have used several times the “trick” of expressing an event B as the union of mutually exclusive events in the following way

$$B = (A \cap B) \cup (A^c \cap B)$$

The sets A and A^c form a very simple **partition** of the sample space S .

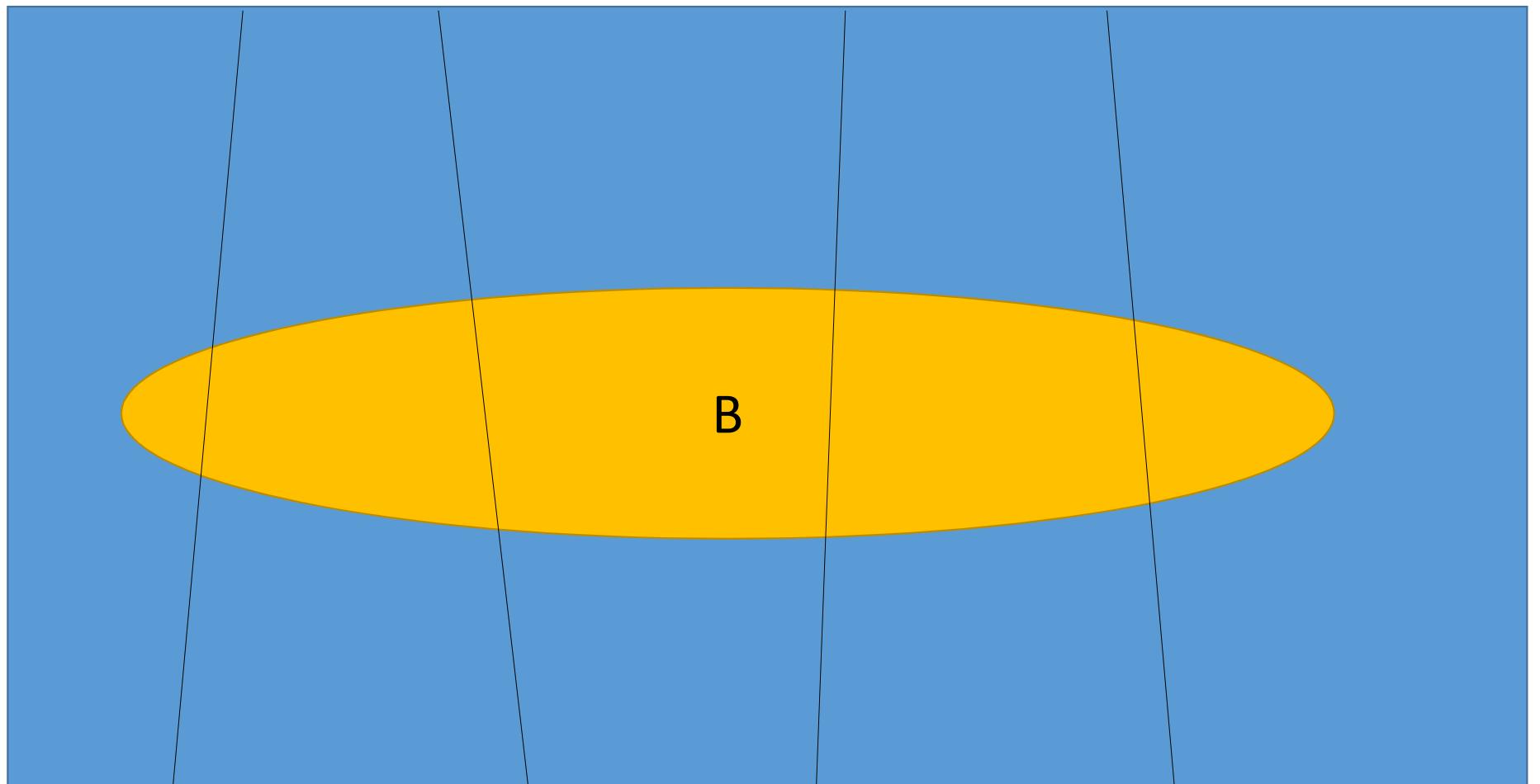
A **partition** of S is a finite collection of mutually exclusive events whose union is all of S .

Law of Total Probability

For any collection of mutually exclusive events $A_1, A_2, A_3, \dots, A_k$ such that $A_1 \cup A_2 \cup A_3 \cup \dots \cup A_k = S$

$$P(B) = \sum_{i=1}^k P(A_i \cap B) = \sum_{i=1}^k P(B|A_i)P(A_i)$$

for any event B .

 A_1 A_2 A_3

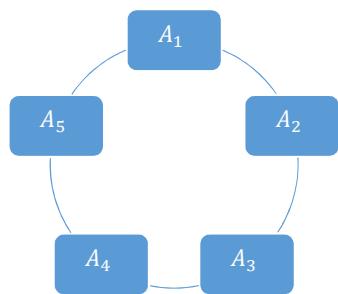
ooo

 A_k^{140}

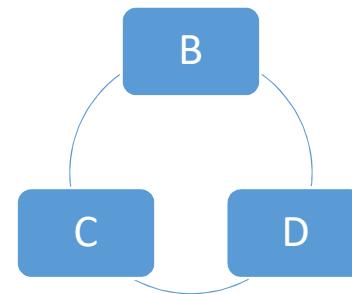
S

B

Stage 1



Stage 2



$P(A_i)$ is called a prior probability

Suppose we want to compute $P(A_i|B)$,
which is an example of a posterior probability.
Bayes' Theorem tells us how to compute
posterior probabilities.

It seems logical to condition on the outcome from stage 1 to compute the probability of an event associated with stage 2, as in the Law of Total Probability.

Bayes' Theorem

For any collection of mutually exclusive events $A_1, A_2, A_3, \dots, A_k$ such that $A_1 \cup A_2 \cup A_3 \cup \dots \cup A_k = S$ and any event B ,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^k P(B|A_i)P(A_i)}$$

Example: Blue Cab operates 15% of the taxis in a certain city, and Green Cab operates the other 85%. After a nighttime hit-and-run accident involving a taxi, an eyewitness said the vehicle was blue. Suppose, though, that under night vision conditions, only 80% of individuals can correctly distinguish between a blue and a green vehicle. What is the (posterior) probability that the taxi at fault was blue?

Let B = “cab in accident was blue” so that B^c = “cab in accident was green”.

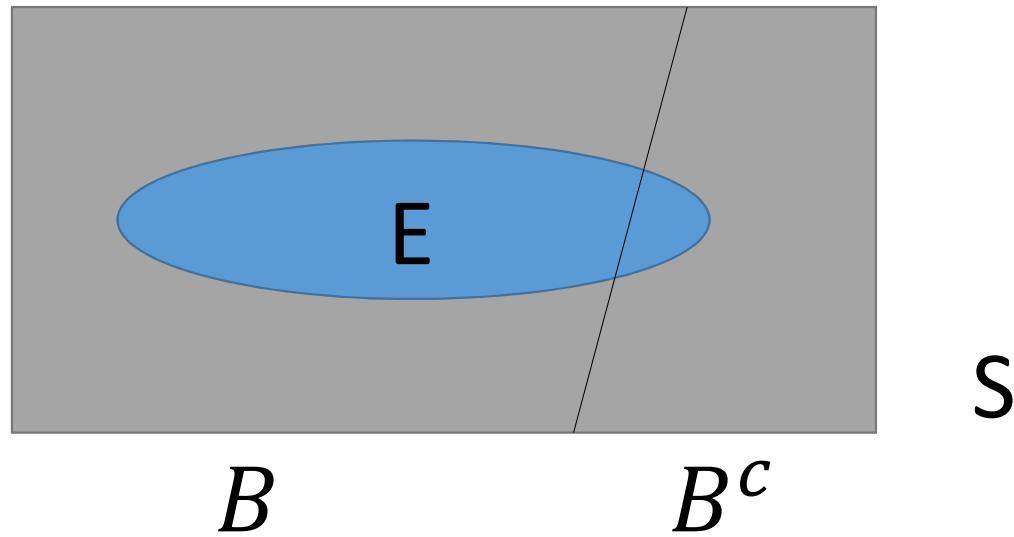
Let E = “eye witness says cab was blue”.

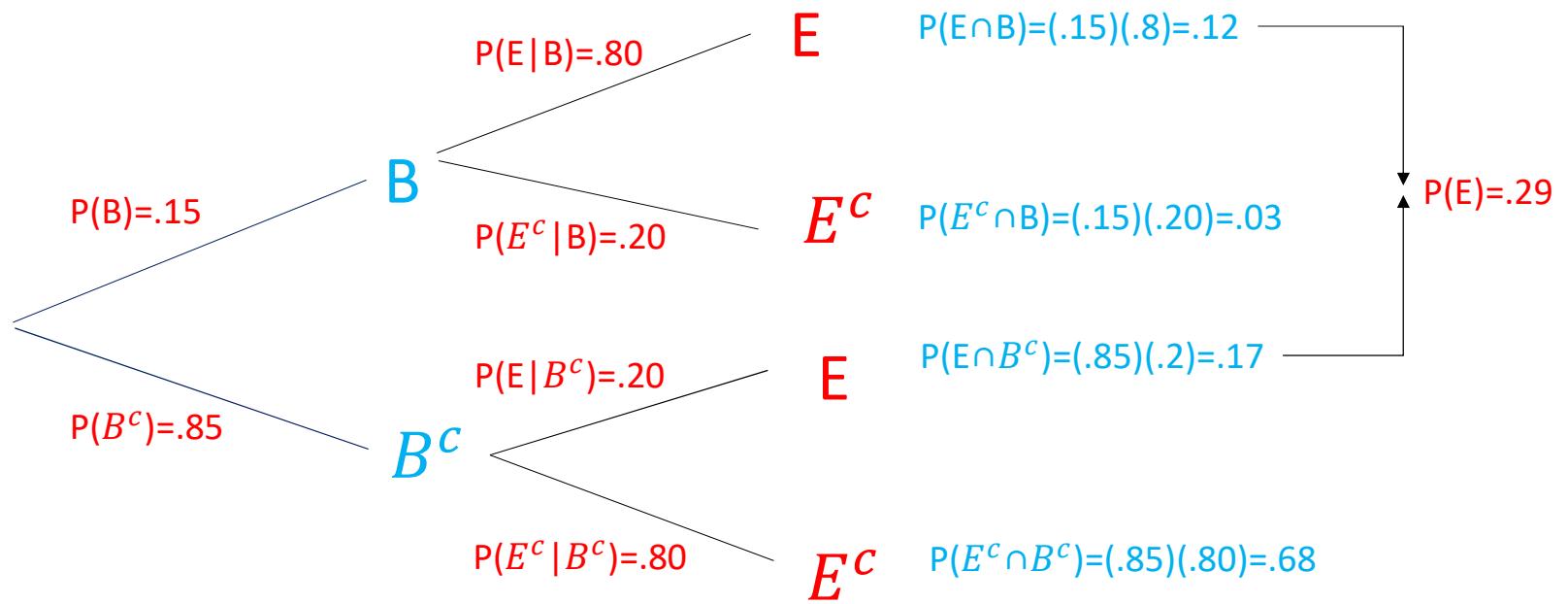
$P(B) = .15$, $P(B^c) = .85$, $P(E|B) = .80$, and $P(E|B^c) = .20$

Note: B and B^c form the partition of S in this problem.

We want to compute

$$P(B|E) = \frac{P(B \cap E)}{P(E)}$$





We have been using the following result quite often in previous slides:

$$P(A_1 \cap A_2) = P(A_1) P(A_2 | A_1)$$

We refer to this as the *general multiplication law*.

Example: Suppose we draw two cards at random without replacement from a standard deck. What is the probability both cards are spades?

Let A_1 = “first card drawn is a spade” and A_2 = “second card drawn is a spade”

The general multiplication law extends to more than two events. Just condition each new event on *all* of the previous events.

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) P(A_2 | A_1) P(A_3 | A_1 \cap A_2) \cdots P(A_n | A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Suppose four cards are drawn from the deck without replacement, what is the probability that all are spades?

Let A_i = “ i^{th} card drawn is a spade” for $i = 1, 2, 3, 4$.

Then

Independence

Definition: If $P(A|B) = P(A)$, or equivalently $P(B|A) = P(B)$, we say A and B are *independent*.

If A and B are independent, $P(A)>0$, and $P(B)>0$, then

$$P(A \cap B) = P(A|B) P(B) \implies P(A \cap B) = P(A) \cdot P(B)$$

which is an alternative definition for independence.



Warning: $P(A \cap B) = P(A) \cdot P(B)$ is **only** true if A and B are **independent**.

Unless you are certain of the independence of A and B, use $P(A \cap B) = P(A | B) P(B)$

Assuming $P(A) > 0$, $P(B) > 0$, any one of

- $P(A \cap B) = P(A) P(B)$
- $P(A | B) = P(A)$
- $P(B | A) = P(B)$

proves independence and the other two.

Example: Draw one card at random from a well-shuffled deck. Define:
 A = “draw a club”, B = “draw an ace”, C = “draw a red card”

Are A and B independent?

A and C ?

B and C ?

Example: A population of 200 ballpoint pens manufactured by a company are classified as coming from one of two assembly lines (A1, A2) and as being Defective-Trash (DT), Defective-To be Fixed (DF), and Nondefective (N). The results are given in the following table.

	Assembly Line 1	Assembly Line 2	Total
Defective-Trash	8	2	10
Defective-To be Fixed	13	27	40
Nondefective	59	91	150
Total	80	120	200

- a.) Suppose a pen is chosen at random during an inspection. Express the event that “the chosen pen was produced by Assembly Line 1 or is Defective-trash” using the definitions given above and the appropriate set operations (unions, intersections, and complements) **then** compute its probability.
- b.) Compute the probability that the chosen item is Nondefective given that it was produced by Assembly Line 2.
- c.) Are the events N and A₂ independent or dependent? Explain.

We'll often *assume* independence, when it is reasonable to do so, to calculate probabilities of intersections.

Example: Roll a red die and a black die. Assume the dice are fair.

A = “red 6”

B = “black 6”

$P(A) = 1/6 = P(B)$

The outcome on one die shouldn't influence the other, so we assume independence.

This extends to more than 2 events.

The *multiplication law for independent events* says that if events A_1, A_2, \dots, A_n are independent (that is, knowledge of any combination of the A_i 's does not change the probabilities of the remainder), then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) P(A_2) \cdots P(A_n).$$

Note: this is the probability that *all* n events occur.

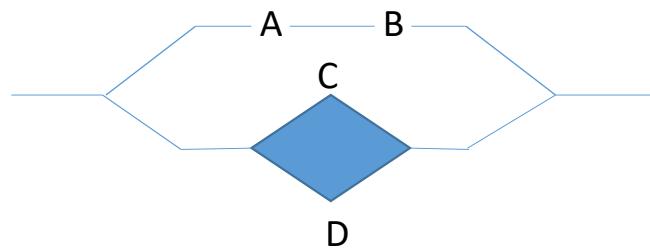
Example: Flip a fair coin 4 times.

Let A_i = “Flip i is a head”.

$$P(A_i) = 1/2, \quad i = 1, 2, 3, 4$$

Separate flips may be considered independent.

Example: A system consists of 4 components connected as shown in the diagram below. Suppose all components work independently and correctly with probability .9. Find the probability that the system functions.



Let S_1 = “the series system of A and B functions”,
 S_2 = “the parallel system of C and D functions”, and
 S = “the parallel system of S_1 and S_2 functions”.

Random Variables

Definition: A *random variable* is a rule that assigns a number to each outcome in a sample space S . It is a function from S into the real numbers.

Example: Roll a die. The number rolled is a random variable.

We can think of the function assigning real values to the outcomes in S as the identity function.

Example: Flip a coin 5 times. Is the sequence of heads and tails a random variable (Example: HHTHT)?

Some random variables we could generate from 5 coin flips:

$$X = \# H$$

$$Y = \# H \times \# T$$

$$Z = \# H \text{ before first } T$$

We usually denote random variables by capital letters from the end of the alphabet.

Example: Select a kitten at random from a large litter. What are some possible random variables?

- X = length of tail
- Y = gender
- W = weight
- V = number of white paws

Definition: A **quantitative variable** represents a measurement or count. The values assumed by the variable are associated with an ordered scale which indicates relative size.

Definition: A **qualitative (or categorical)** variable assumes numerical values that indicate membership in a collection of classes or categories associated with the outcomes in a sample space. The assigned values are generally arbitrary and are not representative of an ordered scale.

Definition: The **support** of a random variable is the set of all possible values that the random variable can assume.

Kitten Example

- $X = \text{length of tail}$: quantitative variable, support $(0, \infty)$ inches
- $Y = \text{gender}$: qualitative variable, support $\{0,1\}$ where $0 = \text{Male}$, $1 = \text{Female}$
- $W = \text{weight}$: quantitative variable, support $(0, \infty)$ ounces
- $V = \text{number of white paws}$: quantitative, support $\{0,1,2,3,4\}$

5 coin toss example (all quantitative)

$X = \# H$: support $\{0, 1, 2, 3, 4, 5\}$

$Y = \# H \times \# T$: support $\{0, 4, 6\}$

$Z = \# H$ before first T : support $\{0, 1, 2, 3, 4, 5\}$
(where Z=5 if all tosses are heads)

There are two main types of random variables: **discrete** and **continuous**.

Definition: A **discrete** random variable can only take on a specified (countable) list of values. There is a gap between any two elements in its sample space.

In practice, these are usually counts of some sort, and thus whole numbers.

Definition: A **continuous** random variable may take any real number in some (set of) interval(s).

Kitten Example

- $X = \text{length of tail}$: quantitative variable, support $(0, \infty)$ inches : continuous
- $W = \text{weight}$: quantitative variable, support $(0, \infty)$ ounces : continuous
- $V = \text{number of white paws}$: support $\{0,1,2,3,4\}$: discrete

5 Coin Toss Example

- $X = \# H$: support $\{0, 1, 2, 3, 4, 5\}$: discrete
- $Y = \# H \times \# T$: support $\{0, 4, 6\}$: discrete
- $Z = \# H \text{ before first } T$: support $\{0, 1, 2, 3, 4, 5\}$: discrete
(where $Z=5$ if all tosses are heads)

We will need to deal differently with discrete and continuous random variables when computing their associated probabilities.

We will begin by discussing discrete random variables.

Definition: The *probability mass function* (p.m.f.) of a discrete random variable X is a function $p(\cdot)$ from the support of X to the real numbers, where

$$p(x) = P(X = x).$$

Notation:

X : capital letter, indicates a random variable

x : lowercase letter, indicates a specific value

Example: Let X be the roll of a fair die with support $\{1, 2, 3, 4, 5, 6\}$.

$$p(1) = P(X = 1) = 1/6$$

$$p(2) = P(X = 2) = 1/6$$

and so on.

We might write

$$p(x) = 1/6 \quad x \in \{1, 2, 3, 4, 5, 6\}$$

or in table form

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

Example: An industrial plant has 3 machines. The probability that X are operating at a given random time may be found from

x	0	1	2	3
p(x)	0.12	0.27	0.46	0.15

- (a) Find the probability that more than 2 machines are operating at any given time.

(b) Find the probability that at most 2 machines are operating at any given time.

or, alternatively,

(c) Find the probability that between 1 and 2 (inclusive) machines are running at any given time.

(d) Find the probability that at least 1 machine is running at any given time.

or, equivalently,

The laws of probability tell us that:

- 1) $0 \leq p(x) \leq 1$ for all $x \in S$ (we will use S to denote the support of X)
- 2) $\sum_{x \in S} p(x) = 1$

Ex: Suppose

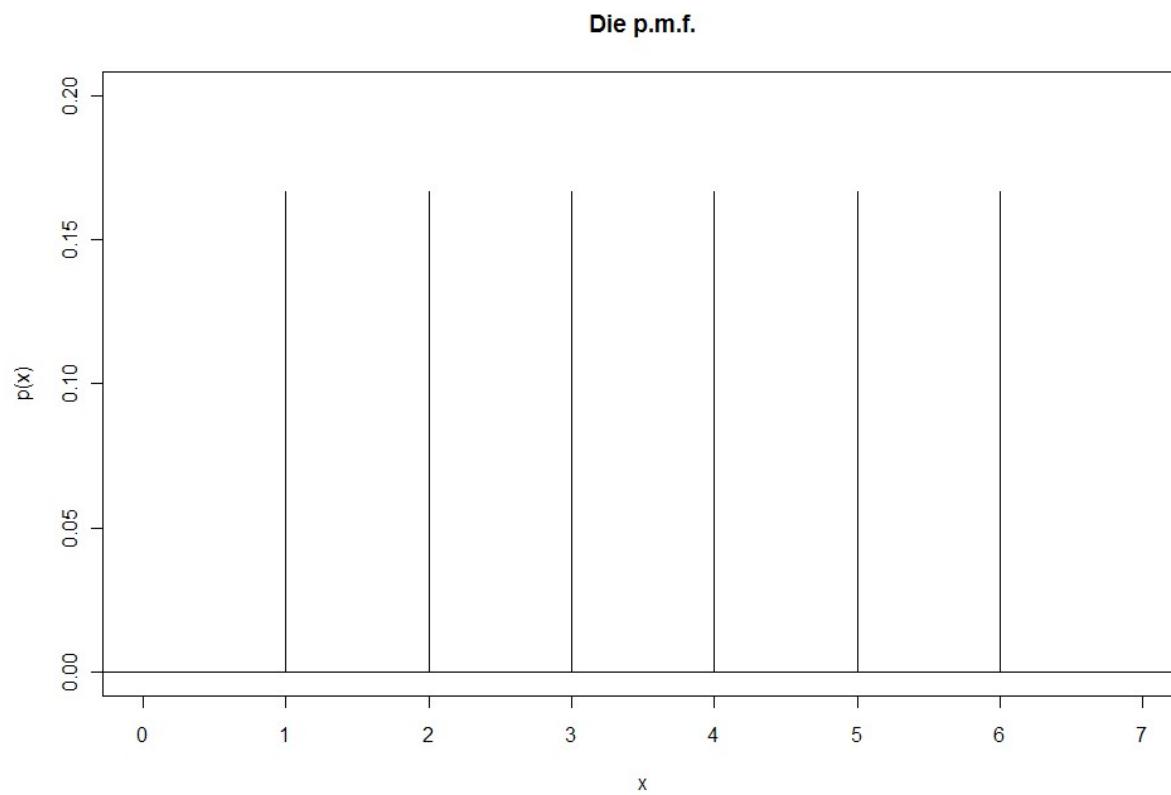
$$p(x) = k \left(\frac{1}{2}\right)^x \text{ for } x=1,2,\dots$$

(a) Find k.

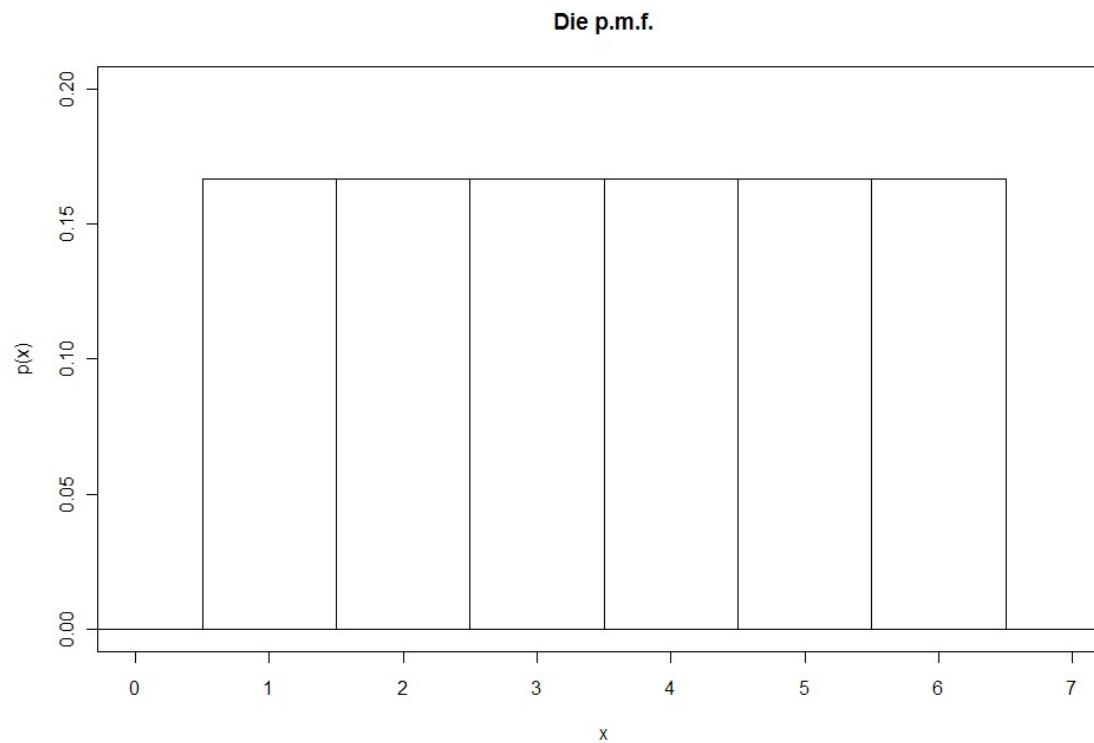
The main requirement for choosing k is to be sure that $\sum_{x=1}^{\infty} p(x) = 1$.

(b) Compute $P(X \geq 5)$.

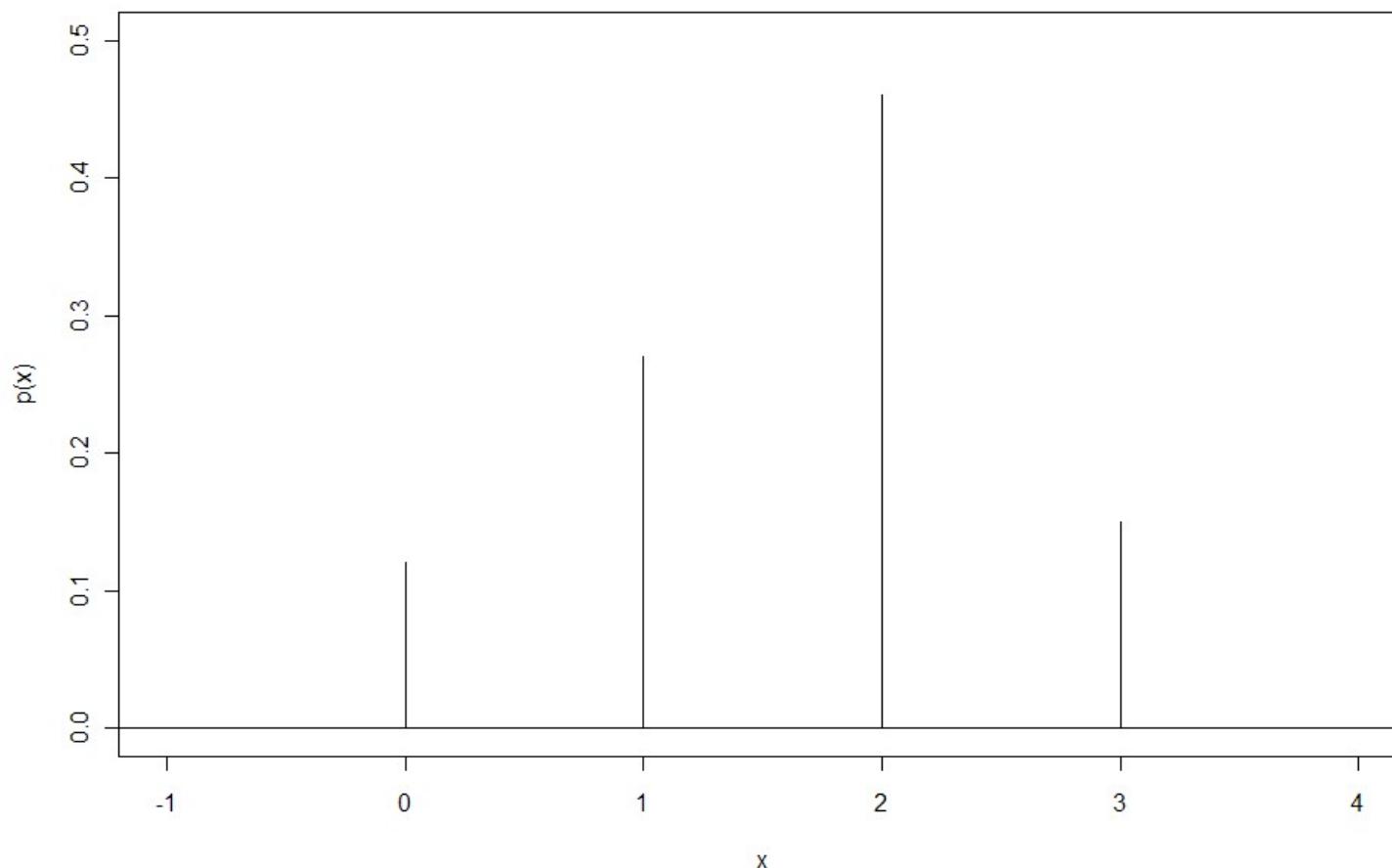
A p.m.f. is plotted as spikes:



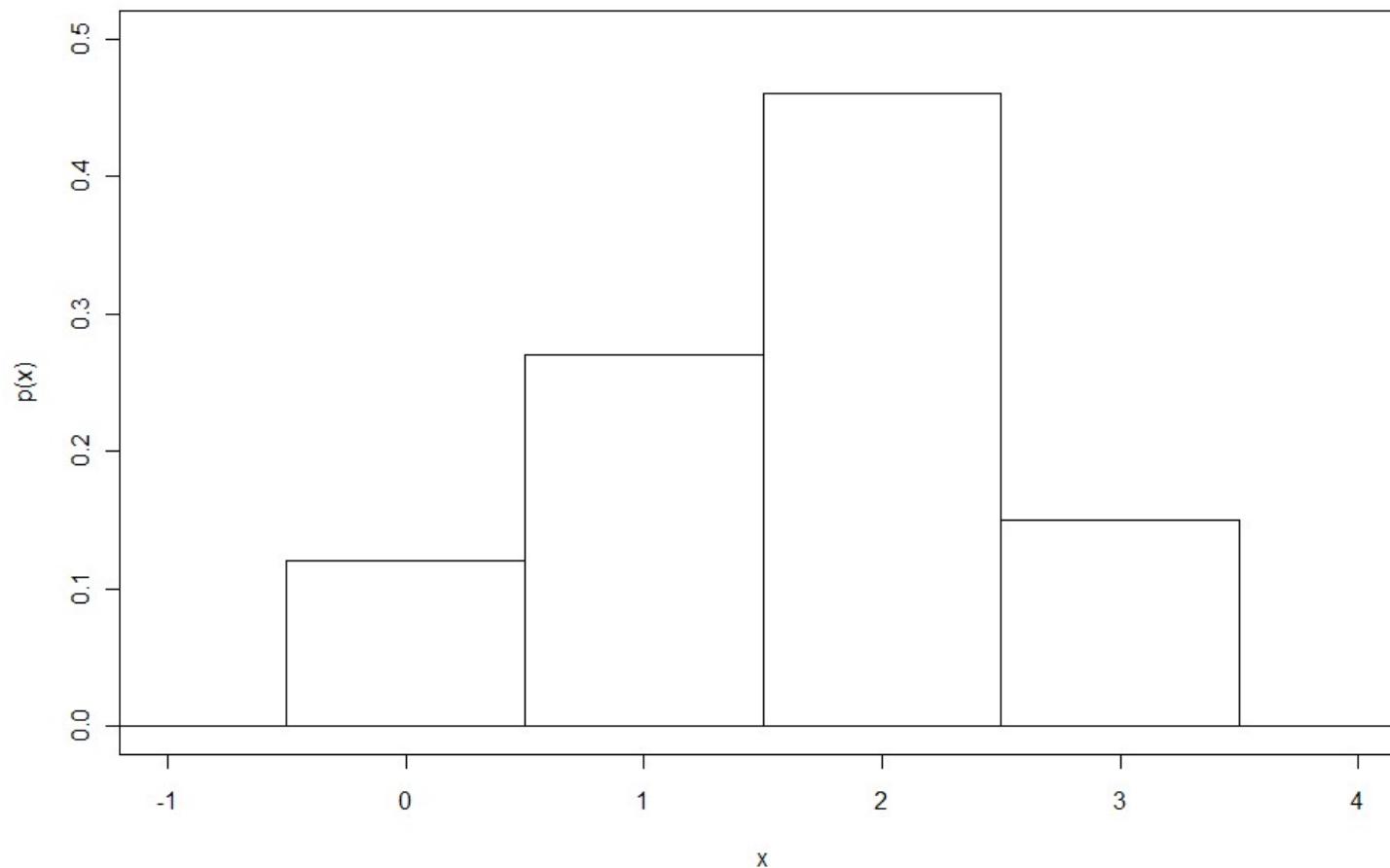
Or as a *probability histogram*, with areas equal to probabilities:



Machines p.m.f.



Machines p.m.f.



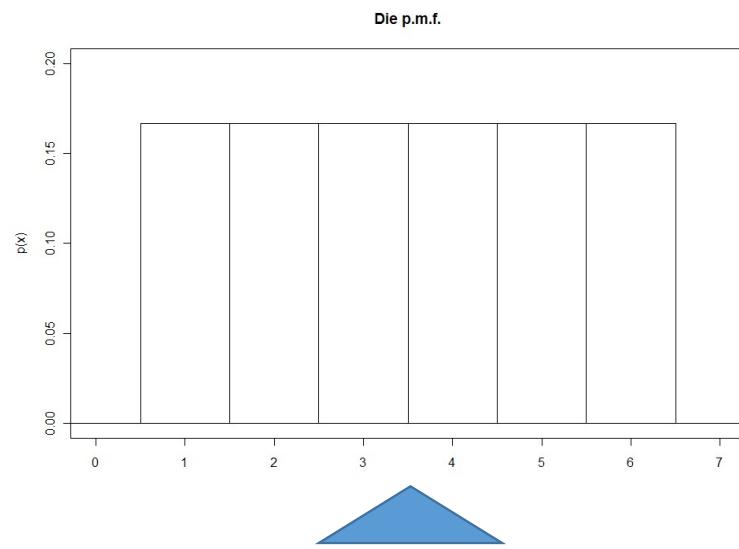
The Population Mean

Definition: The *population mean* (*expectation, expected value*) of a **discrete** random variable X is

$$\mu = E(X) = \sum_{x \in S} xp(x)$$

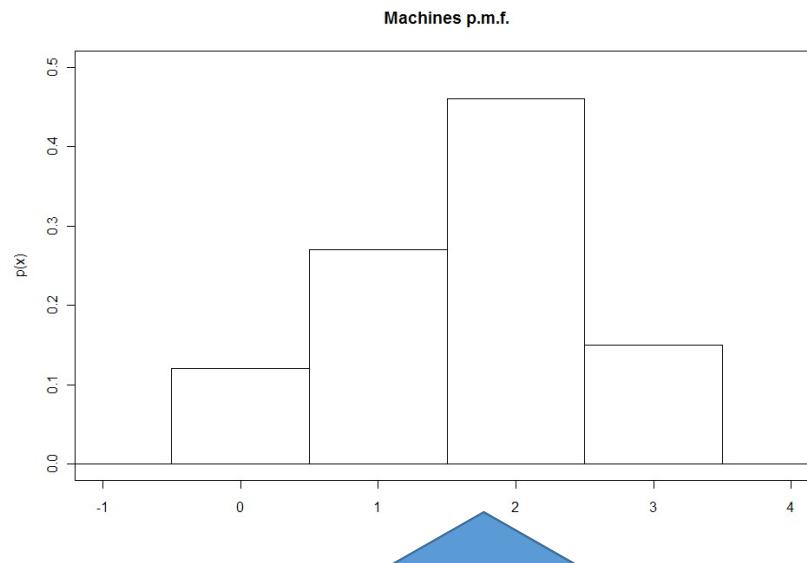
It can be thought of as the **long-term average of X** , or the location of the **balance point** of the system of probabilities located at the points in the support of X .

Example: Die roll with $p(x) = 1/6$ for $x \in \{1, 2, \dots, 6\}$



Example: Machines

x	0	1	2	3
p(x)	0.12	0.27	0.46	0.15



Expectations of Functions of Random Variables

Given a random variable, X , suppose we are really interested in a function, $h(X)$.

The expected value of $h(X)$ is

$$E(h(X)) = \sum_{x \in S} h(x)p(x)$$

if X is discrete.

Example: Suppose you manage a movie theater and you have a probability mass function for X = the number of adult tickets sold on any given Saturday. One of your main interests would be the revenue produced by the ticket sales. If an adult ticket costs \$10, you would be interested in the random variable $h(X) = 10X$.

Example: $X \sim p(x) = \frac{1}{2}$, $x = 1, 2$. “ \sim ” represents “is distributed as”

What is $E(X^2)$?

Is $E(X^2) = [E(X)]^2$?

Note: In general, $E[h(X)] \neq h[E(X)]$.

The Population Variance and Standard Deviation

Just as we have a population mean to measure the center of a distribution, the *population variance* and *standard deviation* measure a distribution's spread.

Definition: Let X be a random variable with mean μ . Then the *population variance of X* , σ^2 , is

$$\sigma^2 = V(X) = E(X - \mu)^2 = E(X^2) - \mu^2$$

Definition: The *population standard deviation*, σ , of a random variable X is the square root of the variance of X .

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}$$

Example: Die roll with $p(x) = 1/6$ $x \in \{1, 2, \dots, 6\}$. Recall $\mu = 3.5$.

Example: Machines

x	0	1	2	3
p(x)	0.12	0.27	0.46	0.15

Recall: $\mu = 1.64$

Continuous Random Variables

Recall, a continuous random variable may take any value in some real interval.

Continuous random variables are typically measurements (length, weight, lifetime, etc.).

With continuous random variables, we can't use a p.m.f. to find probabilities. Instead:

Definition: A *probability density function (density, p.d.f.)*, $f(x)$, is a function which determines the probability properties of a continuous random variable. If $X \sim f(x)$, then

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

If $f(x)$ is a p.d.f., then

1) $f(x) \geq 0$ for all x , and

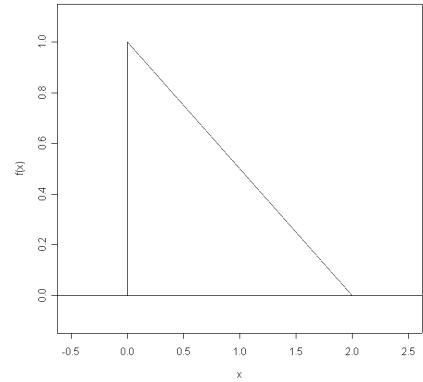
2) $\int_{-\infty}^{\infty} f(x) dx = 1$

Note: For a continuous random variable,

$$\int_a^a f(x) dx = 0$$

Example: A continuous random variable has p.d.f.

$$f(x) = \begin{cases} (1 - \frac{x}{2}), & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$



Is this a valid p.d.f?

What is the probability that X will be between 0.5 and 1.0?

$$P(2.5 \leq X \leq 3.0) = ?$$

$$P(0.2 \leq X \leq 0.2) = ?$$

$$P(X < 1.0) = ?$$

Definition: The *cumulative distribution function (c.d.f.)*, $F(x)$, of a random variable is defined as

$$F(x) = P(X \leq x).$$

If X is continuous,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

and

$$\frac{d}{dx} F(x) = f(x)$$

Properties of continuous c.d.f.'s:

$$1) \lim_{x \rightarrow -\infty} F(x) = 0$$

$$2) \lim_{x \rightarrow \infty} F(x) = 1$$

3) F is nondecreasing (if $x < y$, $F(x) \leq F(y)$).

$$\begin{aligned} 4) P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a). \end{aligned}$$

Example: A continuous random variable has p.d.f.

$$f(x) = \begin{cases} (1 - \frac{x}{2}), & 0 \leq x \leq 2 \\ 0 & otherwise \end{cases}$$

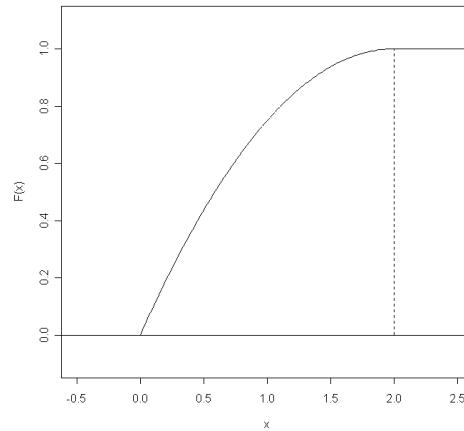
If $x < 0$,

If $0 \leq x < 2$,

If $x > 2$,

So

$$F(x) = \begin{cases} 0 & x < 0 \\ x - \frac{x^2}{4} & 0 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$



Definition: The *population mean (expectation, expected value)* of a continuous random variable X is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

and the *population variance* of X , is

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2\end{aligned}$$

Example: A continuous random variable has p.d.f.

$$f(x) = \begin{cases} (1 - \frac{x}{2}), & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Example: Particles are a major component of air pollution in many areas. It is of interest to study the sizes of contaminating particles. Let X represent the diameter, in micrometers, of a randomly chosen particle. Assume that in a certain area, the probability density function of X is inversely proportional to the volume of the particle; that is, assume that $f(x) = \frac{c}{x^3}$ for $x \geq 1$ and $f(x) = 0$ for $x < 1$ where c is a constant.

(a) Find c so that f is a valid pdf.

We need $\int_1^\infty \frac{c}{x^3} dx = 1$.

(b) Find the mean (and variance?) of X .

(c) Find the cdf of X .

(d). The term PM_{10} refers to particles 10 Mm or less. Find the proportion of particles that are PM_{10} .

(e) Find the proportion of particles that are $PM_{2.5}$.

(f) What proportion of PM_{10} particles are $PM_{2.5}$?

Linear Functions of Random Variables

A *linear function* (or *linear combination*) of variables x_1, x_2, \dots, x_n , is a function of the form

$$f(x_1, x_2, \dots, x_n) = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$$

where b and all of the a_i 's are fixed constants

Given any random variables X_1, X_2, \dots, X_n and known constants a_1, a_2, \dots, a_n , and b , then

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \dots + a_nX_n + b) &= \\ a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) + b. \end{aligned}$$

To find the expectation of a linear combination of random variables, we need only know the constants and the expectation of each random variable individually.

Example: Let X be a random temperature measured in degrees Celsius, with $E(X) = 10$. Let Y be the same temperature in degrees Fahrenheit, $Y = 9/5 X + 32$. What is $E(Y)$?

Example: The expectation of the roll of a fair die is 3.5. What is the expectation of the sum of four such rolls?

Independent Random Variables

Recall, events are said to be independent if knowledge of one does not affect the probability of the other.

Likewise, random variables X and Y are independent if knowing the value of X does not affect probabilities of Y , no matter what value X takes (and vice-versa).

If X and Y are independent, any event involving X alone will be independent from any event involving Y alone.

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B)$$

for any A and B .

Draws with replacement are independent.

Draws in a simple random sample are not independent, but may be treated as though they are if the sample size is much smaller than the population size.

If the random variables are independent, then

$$V(a_1X_1 + a_2X_2 + \dots + a_nX_n + b) = \\ a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n).$$

Notes:

- The shift b does not affect the variance.
- The coefficients a_i are squared.
- Dependent random variables require a more complex formula.

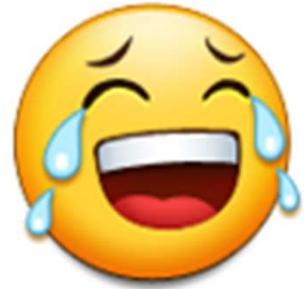
Example: Let the variance of the Celsius temperature X be $V(X) = 25$.

- What is the standard deviation of X ?
- What is the variance of $Y = 9/5 X + 32$?
- What is the standard deviation of Y ?

Example: The variance of the roll of a fair die is $35/12$. What is the variance of the sum of four such rolls?

The four rolls are independent. So

Suppose X and Y each have mean 10 and variance 4. What are the mean and variance of $Z = X - Y$? (Assume X and Y are independent.)



*“Two random variables were talking in a bar.
They thought they were being discrete, but I
heard their chatter continuously.”*

Families of Random Variables

Discrete

Binomial

Geometric

Poisson

Negative
Binomial

Discrete
Uniform

Gamma

Normal

Continuous

Weibull

Beta

Uniform

Binomial Random Variables

The binomial distribution is the most important common named family of discrete distributions.

Recall, a discrete distribution is described by a probability mass function $p(\cdot)$, where

$$p(0) = P(X = 0)$$

$$p(1) = P(X = 1)$$

and so on.

Suppose our experiment consists of “trials” with only two possible outcomes.

One outcome – called a “success” – occurs with probability p .

The other outcome is called a “failure”, and occurs with probability $(1 - p)$.

Such a process is called a *Bernoulli trial* (after 17th-century probabilist James Bernoulli).

The binomial distribution looks at a fixed number of independent identical Bernoulli trials, and counts the number of successes.

Example: Suppose silicon computer chips are made in pairs, and that 30% of all chips produced are defective.

Also assume that the chips in a pair are independent of each other.

Out of pairs in which the first chip is good, the second is defective in 30% of pairs. This remains true for pairs in which the first chip is defective.

Out of all pairs, 70% will have a good first chip. Out of *those*, 70% will also have a good second chip. Overall, 70% of 70%, or 49% ($.7 \times .7 = .49$) will have two good chips.

Likewise, 30% of that 70%, or 21% overall ($.7 \times .3 = .21$) will have a good first chip and a defective second chip.

By the same reasoning, 30% will have a defective first chip, and 70% of those (21% overall) will have a good second chip.

Finally, 30% of 30%, or 9% will have both chips defective.

If we let the letter S (for success) represent a good chip, and F (for failure) represent a defective one, we can summarize as:

$$P(SS) = .7 * .7 = .49$$

$$P(SF) = .7 * .3 = .21$$

$$P(FS) = .3 * .7 = .21$$

$$P(FF) = .3 * .3 = .09$$

Now let X be the number of good chips produced in a pair.

Then X can take the values 0, 1, or 2.

From the above,

$$p(0) = P(X = 0) = P(\text{FF}) = .09$$

$$p(2) = P(X = 2) = P(\text{SS}) = .49$$

$$p(1) = P(X = 1) = P(\text{SF or FS}) = .21 + .21 = .42$$

What if the chips are produced in sets of 4?

If we want the probability of a set consisting of 2 good and 2 defective chips, we can think about the case of SSFF – the first and second chips are good, while the third and fourth are defective.

The probability of this particular outcome will be
 $.7 * .7 * .3 * .3 = .0441$ or 4.41%.

But there are other ways we can have two successes and two failures – 5 other ways, in this case:

$$P(SSFF) = .7 \cdot .7 \cdot .3 \cdot .3 = .0441$$

$$P(SFSF) = .7 \cdot .3 \cdot .7 \cdot .3 = .0441$$

$$P(SFFS) = .7 \cdot .3 \cdot .3 \cdot .7 = .0441$$

$$P(FSSF) = .3 \cdot .7 \cdot .7 \cdot .3 = .0441$$

$$P(FSFS) = .3 \cdot .7 \cdot .3 \cdot .7 = .0441$$

$$P(FFSS) = .3 \cdot .3 \cdot .7 \cdot .7 = .0441$$

Overall, $p(2) = P(X = 2) = 6 \cdot .0441 = .2646$.

In general, suppose we have an experiment consisting of n independent Bernoulli trials.

Those trials which satisfy the condition we wish to count are called successes, and occur with probability p .

The remaining trials are called failures; these occur with probability $(1 - p)$.

Let X be the number of successes in the full experiment.

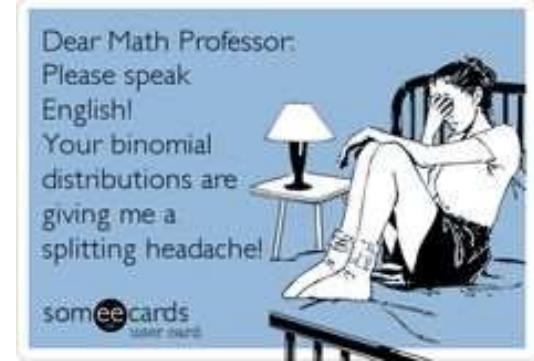
If these conditions are true, we say that X , the number of successes in the experiment, has a binomial distribution with parameters n and p .

$$X \sim \text{Binomial}(n, p) \text{ or } X \sim \text{Bin}(n, p).$$

The mass function for X is:

$$p(x) = P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for $x = 0, 1, 2, \dots, n$.



Note: the exclamation mark is pronounced “factorial”.

Given n items, the number of arrangements of all n is

$$n! \equiv n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1.$$

Since there is one (empty) way to arrange 0 objects, we define $0! = 1$.

Example: Chips (30% defective) are produced in batches of 4. Let X be the number of good chips in a batch.

What distribution does X follow?

What is $p(2)$?

What proportion of batches contain no more than one good chip?

Example: In a genetics study, a second-generation cross of pure green peas with pure yellow peas leads to pods where $p = P(\text{yellow}) = \frac{3}{4}$.

If pods contain 8 seeds, what is the probability that a random pod will contain exactly 6 yellow seeds?

What is the probability that a random pod will contain *at least* 6 yellow seeds?

Minitab has built in probability tables and there are many probability tables available online. One of our options is the Stat Trek site.

Example: Draw 15 times *with replacement* from a standard deck, and let X = number of spades drawn.

Then $X \sim Bin(15, .25)$.

Find $P(X > 6)$. (.0566)

Statistics Tables

Distributions

Binomial
Chi-Square
 f Distribution
Hypergeometric
Multinomial
Negative Binomial
Normal
Poisson
 t Distribution

Browse

Tutorials

AP Statistics
Statistics and Probability
Matrix Algebra

AP Statistics

Test Preparation
Practice Exam
Study Guide Review
Approved Calculators

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial

Number of trials

Number of successes (x)

Binomial Probability: $P(X = 6)$

Cumulative Probability: $P(X < 6)$

Cumulative Probability: $P(X \leq 6)$

Cumulative Probability: $P(X > 6)$

Cumulative Probability: $P(X \geq 6)$

Calculate

With standard distributions, the mean and variance may generally be found as a function of the parameters.

If $X \sim \text{Binomial}(n, p)$, then $\mu = np$.

Example: If 75% of all seeds are yellow, and each pod contains 8 seeds, what is the mean number of yellow seeds per pod?

Example: If we have 4 fair coins which we flip as a batch, what is the mean number of heads?

Additionally, if $X \sim \text{Bin}(n, p)$, then $\sigma^2 = np(1 - p)$.

Example: $X = \# \text{ yellow seeds} \sim \text{Bin}(8, .75)$. What are the variance and standard deviation of X ?

Example: $X = \# \text{ heads in 4 flips} \sim \text{Bin}(4, .5)$. What are the variance and standard deviation of X ?

Example: A k out of n system is one in which there is a group of n components, and the system will function if at least k of the components function. Assume the components function independently of one another.

- A. In a 3 out of 5 system, each component has probability .9 of functioning. What is the probability that the system will function?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial

Number of trials

Number of successes (x)

Binomial Probability: $P(X = 3)$

Cumulative Probability: $P(X < 3)$

Cumulative Probability: $P(X \leq 3)$

Cumulative Probability: $P(X > 3)$

Cumulative Probability: $P(X \geq 3)$

Calculate

B.) In a 3 out of n system, in which each component has probability .9 of functioning, what is the smallest value of n so that the probability that the system functions is at least .9?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial	<input type="text" value=".9"/>
Number of trials	<input type="text" value="4"/>
Number of successes (x)	<input type="text" value="3"/>
Binomial Probability: $P(X = 3)$	<input type="text" value="0.2916"/>
Cumulative Probability: $P(X < 3)$	<input type="text" value="0.0523"/>
Cumulative Probability: $P(X \leq 3)$	<input type="text" value="0.3439"/>
Cumulative Probability: $P(X > 3)$	<input type="text" value="0.6561"/>
Cumulative Probability: $P(X \geq 3)$	<input type="text" value="0.9477"/>

Calculate

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial	<input type="text" value=".9"/>
Number of trials	<input type="text" value="3"/>
Number of successes (x)	<input type="text" value="3"/>
Binomial Probability: $P(X = 3)$	<input type="text" value="0.729"/>
Cumulative Probability: $P(X < 3)$	<input type="text" value="0.271"/>
Cumulative Probability: $P(X \leq 3)$	<input type="text" value="> 0.999999"/>
Cumulative Probability: $P(X > 3)$	<input type="text" value="< 0.000001"/>
Cumulative Probability: $P(X \geq 3)$	<input type="text" value="0.729"/>

Calculate

For a certain 4 out of 6 system, assume that on a rainy day each component has probability .7 of functioning and that on a nonrainy day each component has a .9 probability of functioning.

A.) What is the probability that the system functions on a rainy day?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial

Number of trials

Number of successes (x)

Binomial Probability: $P(X = 4)$

Cumulative Probability: $P(X < 4)$

Cumulative Probability: $P(X \leq 4)$

Cumulative Probability: $P(X > 4)$

Cumulative Probability: $P(X \geq 4)$

Calculate

B.) What is the probability that the system functions on a nonrainy day?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial	<input type="text" value="0.9"/>
Number of trials	<input type="text" value="6"/>
Number of successes (x)	<input type="text" value="4"/>
Binomial Probability: $P(X = 4)$	<input type="text" value="0.098415"/>
Cumulative Probability: $P(X < 4)$	<input type="text" value="0.01585"/>
Cumulative Probability: $P(X \leq 4)$	<input type="text" value="0.114265"/>
Cumulative Probability: $P(X > 4)$	<input type="text" value="0.885735"/>
Cumulative Probability: $P(X \geq 4)$	<input type="text" value="0.98415"/>
<input type="button" value="Calculate"/>	

C.) Assume that the probability of rain tomorrow is .2. What is the probability that the system will function tomorrow?

Recall, draws without replacement (simple random samples) are not independent.

However, we may do calculations as though they *are* independent (including binomial calculations) as long as the sample size is small (less than 5%) compared to the population size.

Example: A lot of several thousand components contains 7% defective. We sample 8 at random.

What is the probability of no defective components in our sample?

What is the probability of at least one defective?

What is the expected number of defectives in our sample?

The Normal Distribution

The continuous *normal* (or *Gaussian*) distribution has two parameters, μ and σ^2 .

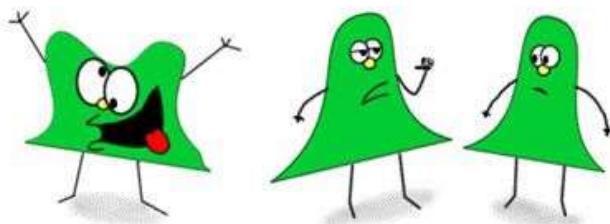
If $X \sim N(\mu, \sigma^2)$,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

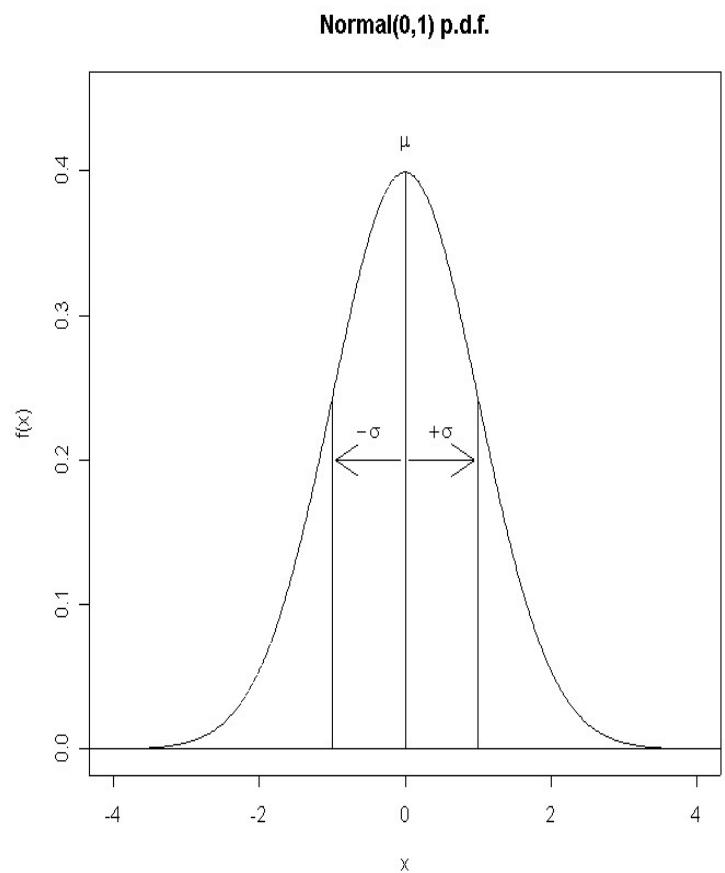
This distribution is often seen in practice, and is also very important theoretically.

The normal p.d.f. is a bell-shaped curve, symmetric around, and with its peak at, μ . $E(X) = \mu$.

Its width is determined by σ^2 ; large values of σ^2 imply a wide, low curve, while small values imply a narrow, tall one. $V(X) = \sigma^2$.



"KEEP YOUR EYE ON THAT GUY, TOM. HE'S NOT, YOU KNOW...NORMAL!"



An important special case is the **standard normal distribution**, with $\mu = 0$ and $\sigma^2 = 1$.

We usually identify standard normal variables with the letter **Z**.

If Z is standard normal, $Z \sim N(0,1)$ and the density of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad -\infty < z < \infty$$

There is no closed-form integral for the normal probability density function, so we can't find probabilities that way. There are tables that can be used for computing probabilities or we can use Minitab or online sources such as Stat Trek for normal probability calculators.

Examples:

$$P(Z \leq 1.00) = ?$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	<input type="text" value="1"/>
Cumulative probability: $P(Z \leq 1)$	<input type="text" value="0.84134"/>
Mean	<input type="text" value="0"/>
Standard deviation	<input type="text" value="1"/>

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

$$P(Z > 1.00) = ?$$

Since normal random variables are continuous, endpoints are not important. So, for example,

$$P(Z \leq 1) = P(Z < 1)$$

and

$$P(Z > 1) = P(Z \geq 1)$$

$$P(-2.00 \leq Z \leq 0.75) = ?$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	<input type="text" value=".75"/>
Cumulative probability: $P(Z \leq .75)$	<input type="text" value="0.77337"/>
Mean	<input type="text" value="0"/>
Standard deviation	<input type="text" value="1"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	<input type="text" value="-2"/>
Cumulative probability: $P(Z \leq -2)$	<input type="text" value="0.02275"/>
Mean	<input type="text" value="0"/>
Standard deviation	<input type="text" value="1"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Examples: Let $X \sim N(3, 4)$.

$$P(X \leq 6.00) = ?$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="6"/>
Cumulative probability: $P(X \leq 6)$	<input type="text" value="0.93319"/>
Mean	<input type="text" value="3"/>
Standard deviation	<input type="text" value="2"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

$$P(X > 4.00) = ?$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="4"/>
Cumulative probability: $P(X \leq 4)$	<input type="text" value="0.69146"/>
Mean	<input type="text" value="3"/>
Standard deviation	<input type="text" value="2"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Normal Percentiles

Just as for samples, the p^{th} percentile of a distribution has $p\%$ of the probability below it, and $(100 - p)\%$ above.

Example: $Z \sim N(0,1)$. What is the 70th percentile of Z?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	0.524
Cumulative probability: $P(Z \leq 0.524)$	0.7
Mean	0
Standard deviation	1

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Example: What is the 25th percentile of Z?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

■ Enter a value in three of the four text boxes.
■ Leave the fourth text box blank.
■ Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	-0.674
Cumulative probability: $P(Z \leq -0.674)$	0.25
Mean	0
Standard deviation	1

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Example: $X \sim N(10, 25)$. What is the 95th percentile of X ?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="18.224"/>
Cumulative probability: $P(X \leq 18.224)$	<input type="text" value=".95"/>
Mean	<input type="text" value="10"/>
Standard deviation	<input type="text" value="5"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Example: Weights of female cats of a certain breed are normally distributed with mean 4.1 kg and standard deviation 0.6 kg. ($\mu = 4.1$ and $\sigma = 0.6$)

(a) What proportion of female cats have weights between 3.7 and 4.4 kg?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	4.4
Cumulative probability: $P(X \leq 4.4)$	0.69146
Mean	4.1
Standard deviation	0.6

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	3.7
Cumulative probability: $P(X \leq 3.7)$	0.25249
Mean	4.1
Standard deviation	0.6

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

(b) A certain female cat has a weight that is 0.5 standard deviations above the mean. What proportion of female cats are heavier than this one?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="4.4"/>
Cumulative probability: $P(X \leq 4.4)$	<input type="text" value="0.69146"/>
Mean	<input type="text" value="4.1"/>
Standard deviation	<input type="text" value="0.6"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

(c) How heavy is a female cat whose weight is on the 80th percentile?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="4.605"/>
Cumulative probability: $P(X \leq 4.605)$	<input type="text" value=".8"/>
Mean	<input type="text" value="4.1"/>
Standard deviation	<input type="text" value="0.6"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

(d) A female cat is chosen at random. What is the probability that she weighs more than 4.5 kg?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="4.5"/>
Cumulative probability: $P(X \leq 4.5)$	<input type="text" value="0.74751"/>
Mean	<input type="text" value="4.1"/>
Standard deviation	<input type="text" value="0.6"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

(e) Six female cats are chosen at random. What is the probability that exactly one of them weighs more than 4.5 kg?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial

Number of trials

Number of successes (x)

Binomial Probability: $P(X = 1)$

Cumulative Probability: $P(X < 1)$

Cumulative Probability: $P(X \leq 1)$

Cumulative Probability: $P(X > 1)$

Cumulative Probability: $P(X \geq 1)$

Calculate

(f) At least 2?

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute Binomial and Cumulative Probabilities.

Probability of success on a single trial

Number of trials

Number of successes (x)

Binomial Probability: $P(X = 2)$

Cumulative Probability: $P(X < 2)$

Cumulative Probability: $P(X \leq 2)$

Cumulative Probability: $P(X > 2)$

Cumulative Probability: $P(X \geq 2)$

Calculate

The Normal Approximation to the Binomial Distribution

Recall, if $X \sim B(n, p)$, then $E(X) = np$ and $V(X) = np(1-p)$.

If the particular values of n and p lead to a binomial distribution which is not very skewed, the $N(\mu = np, \sigma^2 = np(1 - p))$ distribution can be a good approximation to the $B(n, p)$ distribution.

We usually require that $np \geq 10$ and $n(1-p) \geq 10$.

Example: Roll a die 120 times and count the number of 6's rolled (X).

What distribution does X follow?

What are $E(X)$ and $V(X)$?

and

What is $P(X \geq 25)$?



Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	25
Cumulative probability: $P(X \leq 25)$	0.88966
Mean	20
Standard deviation	4.08248

Binomial Calculator: Online Statistical Table

Use the Binomial Calculator to compute individual and cumulative binomial probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the binomial distribution, go to Stat Trek's [tutorial on the binomial distribution](#).

▪ Enter a value in each of the first three text boxes (the unshaded boxes).
▪ Click the **Calculate** button.
▪ The Calculator will compute Binomial and Cumulative Probabilities.

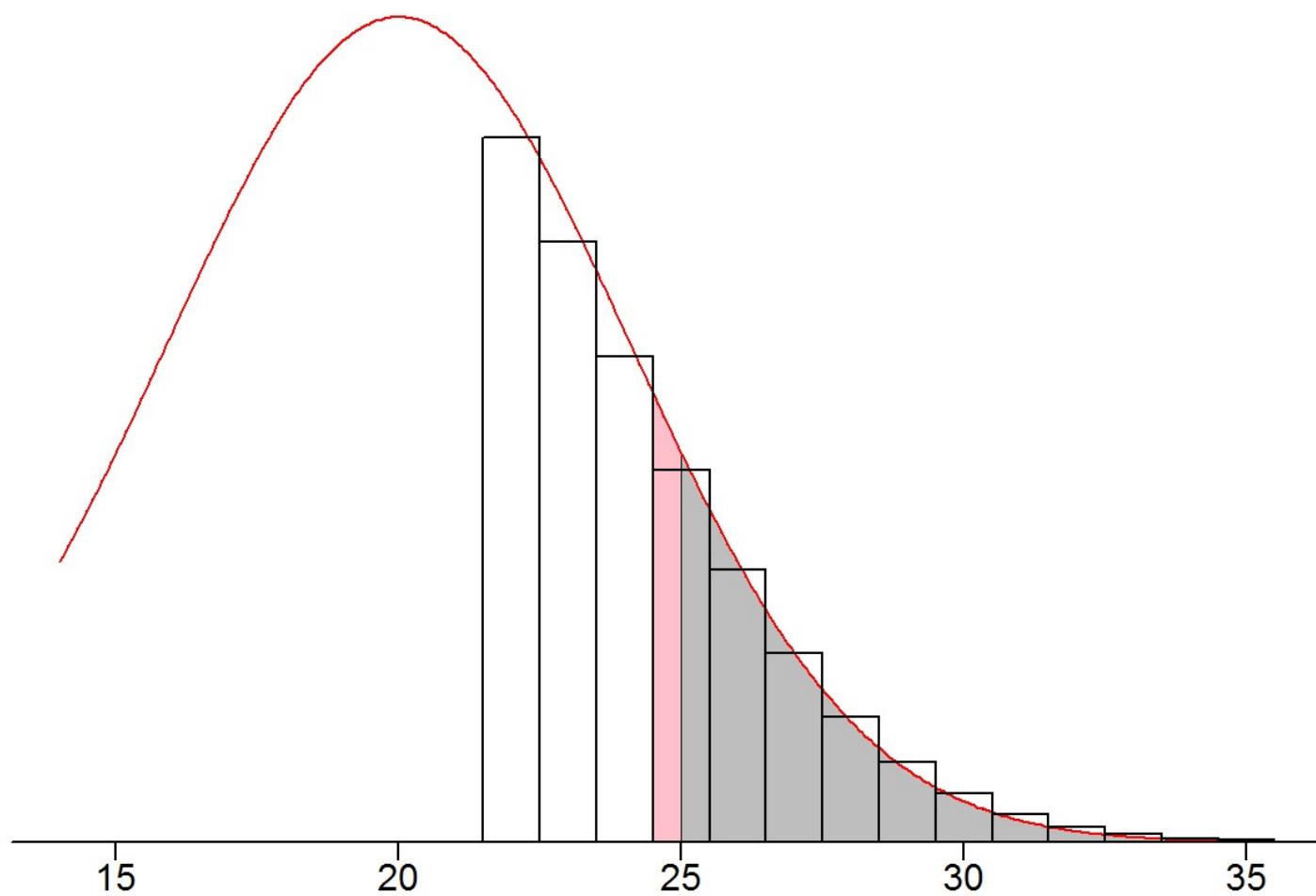
Probability of success on a single trial	.1666667
Number of trials	120
Number of successes (x)	25
Binomial Probability: $P(X = 25)$	0.0441214359514684
Cumulative Probability: $P(X < 25)$	0.86392808337988
Cumulative Probability: $P(X \leq 25)$	0.908049519331348
Cumulative Probability: $P(X > 25)$	0.091950480668652
Cumulative Probability: $P(X \geq 25)$	0.13607191662012

Calculate

The true binomial probability is (approximately) .136. We're pretty close but we can do better.

Binomial probabilities are located entirely on the integers, but normal probabilities are “smeared out” over the whole real line (remember the probability histogram).

We'll get a better approximation if we use a *continuity correction*, by taking the normal probability from $(x - .5)$ to $(x + .5)$ to approximate the binomial $P(X = x)$.



So, for $X \sim B(120, 1/6)$,

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="24.5"/>
Cumulative probability: $P(X \leq 24.5)$	<input type="text" value="0.86483"/>
Mean	<input type="text" value="20"/>
Standard deviation	<input type="text" value="4.08248"/>

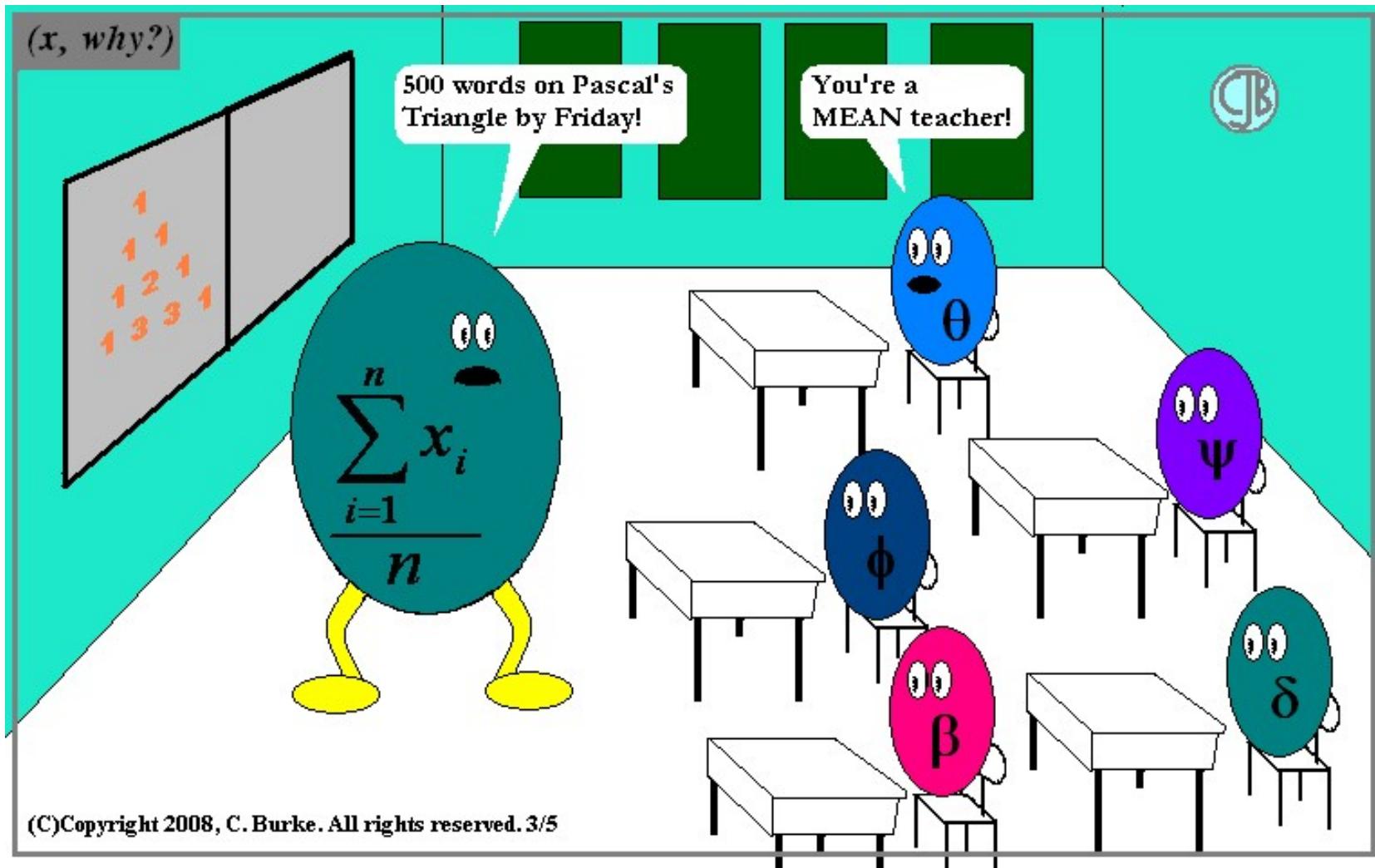
Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Example: If $X \sim Bin(120, 1/6)$, use the normal approximation to estimate

$$P(15 < X < 25).$$

This relationship between the binomial and normal distributions is a good demonstration of the Central Limit Theorem that we will be discussing soon.



Sampling Distributions

Suppose random variable X is drawn from some distribution with distribution function f . ($X \sim f$)

Now suppose we generate n of these random variables, X_1, \dots, X_n , independently from f .

We say that X_1, \dots, X_n is a *random sample* from f .

Sometimes we say that X_1, \dots, X_n are i.i.d. (independent and identically distributed) from f .

Since the X 's make a sample, we can compute sample statistics such as the mean, \bar{X} .

Recall linear combinations. Since the X 's are random, so is \bar{X} and since it assumes real values, \bar{X} is itself a random variable with a distribution.

This distribution is referred to as the *sampling distribution* of \bar{X} and plays a large role in inferential statistics.

Example: Let $p(x) = 1/3$, $x = 1, 2, 3$, be the p.m.f of a discrete distribution. Suppose X_1 and X_2 are independent and follow this distribution.

Now let $\bar{X} = (X_1 + X_2)/2$ be the average of X_1 and X_2 .

Note that \bar{X} is also a discrete random variable, and therefore has a probability mass function.

What is the mass function (sampling distribution) of \bar{X} ?

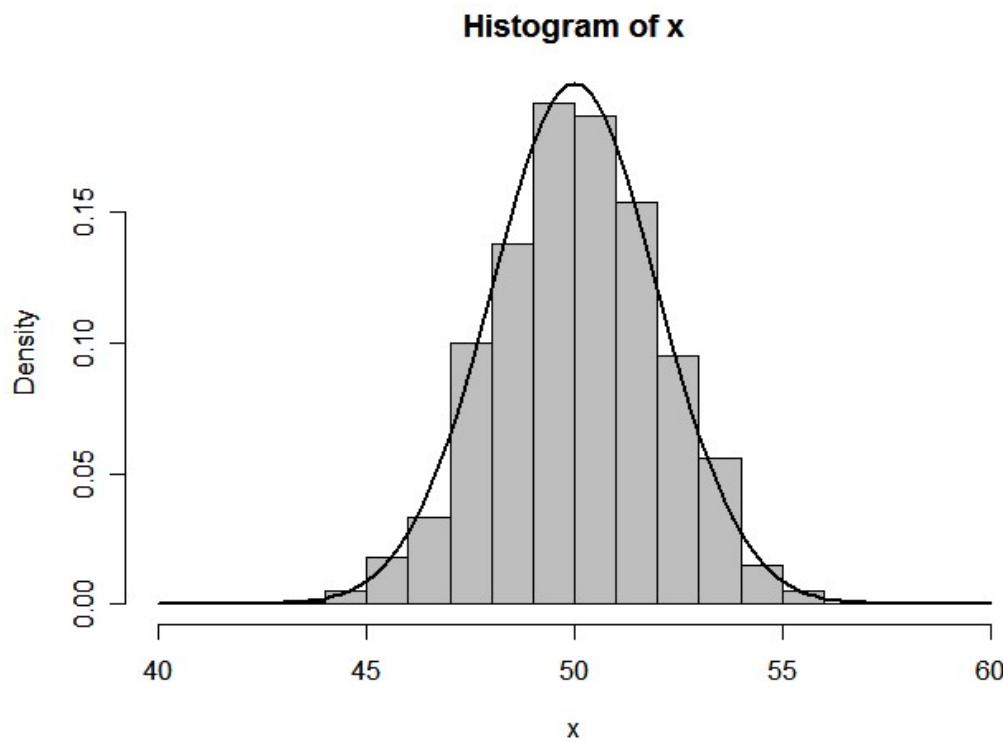
X_1

	1	2	3
1	$\bar{X}=1$ $1/9$	$\bar{X}=1.5$ $1/9$	$\bar{X}=2$ $1/9$
2	$\bar{X}=1.5$ $1/9$	$\bar{X}=2$ $1/9$	$\bar{X}=2.5$ $1/9$
3	$\bar{X}=2$ $1/9$	$\bar{X}=2.5$ $1/9$	$\bar{X}=3$ $1/9$

There is one ordered pair, $(1,1)$, for which $\bar{x}=1$ and one ordered pair, $(3,3)$, for which $\bar{x}=3$. There are two ordered pairs, $(1,2)$ and $(2,1)$, for which $\bar{x}=1.5$, three ordered pairs, $(1,3)$, $(3,1)$ and $(2,2)$, for which $\bar{x}=2$, and two ordered pairs, $(2,3)$ and $(3,2)$, for which $\bar{x}=2.5$.

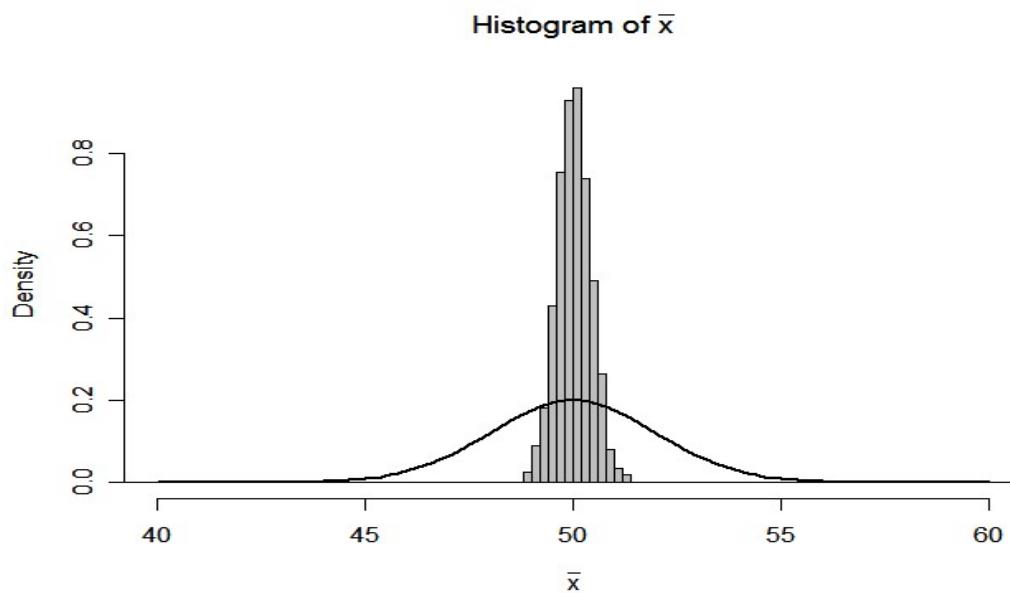
\bar{x}	1	1.5	2	2.5	3
$p(\bar{x})$	1/9	2/9	3/9	2/9	1/9

Example: Suppose $X \sim N(50, 4)$. A histogram of 1000 X 's looks like this:



Sample 25 X 's and compute \bar{X} .

If we repeat this process 1000 times, we get a histogram such as this:



Note that \bar{X} has a distribution that:

- Is centered on 50 (μ);
- Is narrower than the solid normal curve for the individual X 's – the variance and standard deviation of \bar{X} are smaller than those of X .
- Remains bell-shaped and (roughly?) normal.

Understanding the distributions of sample statistics and their relationships to the associated population parameters is the basis of most of inferential statistics.

In general, if a sample statistic is used to estimate a population parameter we desire

- The sampling distribution of the statistic is centered on (or at least near) the parameter.
- The spread of the sampling distribution will decrease as the sample size gets larger.
- As the sample size gets larger, the shape of the sampling distribution will usually get more and more bell-shaped (normal).

Sampling Distribution of the Sample Mean

Let \bar{X} be the sample mean of a random sample X_1, X_2, \dots, X_n , from a population or process with mean μ and standard deviation σ . Then:

The **mean of the sampling distribution of \bar{X}** , $\mu_{\bar{X}}$, is μ , the population mean, regardless of the sample size n .

The **standard deviation of the sampling distribution of \bar{X}** , $\sigma_{\bar{X}}$, is σ/\sqrt{n} , the population standard deviation divided by the square root of the sample size.

$$\mu_{\bar{X}} = \mu,$$

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

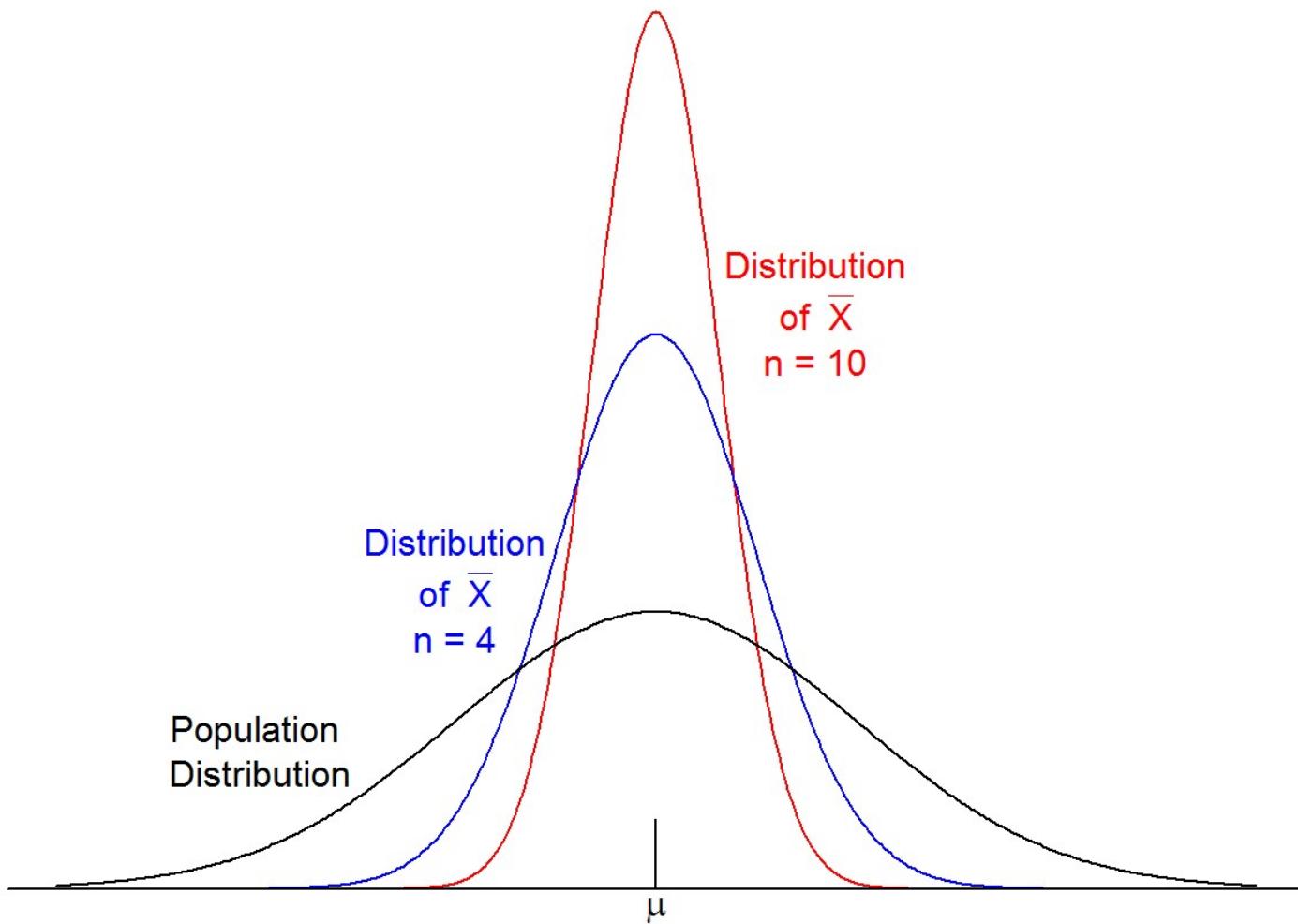
The standard deviation of the sample mean, \bar{X} , is often called the *standard error* of the sample mean.

This emphasizes that it describes a sampling distribution, not a population.

As the sample size gets larger, we have more information and can make better estimates, so the standard error decreases.

(Note, however, that the square root means we have diminishing returns; each new observation provides less new information than the previous one.)

The larger the sample, the closer \bar{X} is likely to be to μ .



If our original population has a normal distribution, the sampling distribution of \bar{X} is also normal, regardless of sample size.

Example: An automated filling machine fills soft drink cans with a volume that has a normal distribution with $\sigma = 0.05$ ounces. If we sample 4 cans and take the sample mean, what is the probability that \bar{X} will be within 0.04 ounces of the population mean μ ?

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{.05}{\sqrt{4}} = .025$ and $\bar{X} \sim N(\mu, (.025)^2)$ where μ is the mean of the sampled distribution

What is $P(\mu - .04 \leq \bar{X} \leq \mu + .04)$?

Since we are not given the value of μ , we cannot use the normal calculator on Stat Trek to compute this probability.

One nice property of normal distributions, is that a linear function of normal random variable is also a normal random variable.

Transformation Formula

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Note: Since $\bar{X} \sim N(\mu, (.025)^2)$, we would have $Z = \frac{\bar{X}-\mu}{.025} \sim N(0,1)$.

The Central Limit Theorem

The Central Limit Theorem is the most important theorem in statistics.

It shows the importance of the normal distribution, and provides the justification of many of the most fundamental statistical methods.

If we know that a population or process has a normal distribution, we know that the sampling distribution of \bar{X} will also be normal. This allows us to compute useful probabilities.

Unfortunately, we often do not know the population distribution (or perhaps we know that it is *not* normal).

Fortunately, this is not always required.

The sample mean (or sum) of a **large** number of independent random variables has a sampling distribution which is approximately normal, no matter what distribution the original random variables come from.

This important result is the *Central Limit Theorem*.

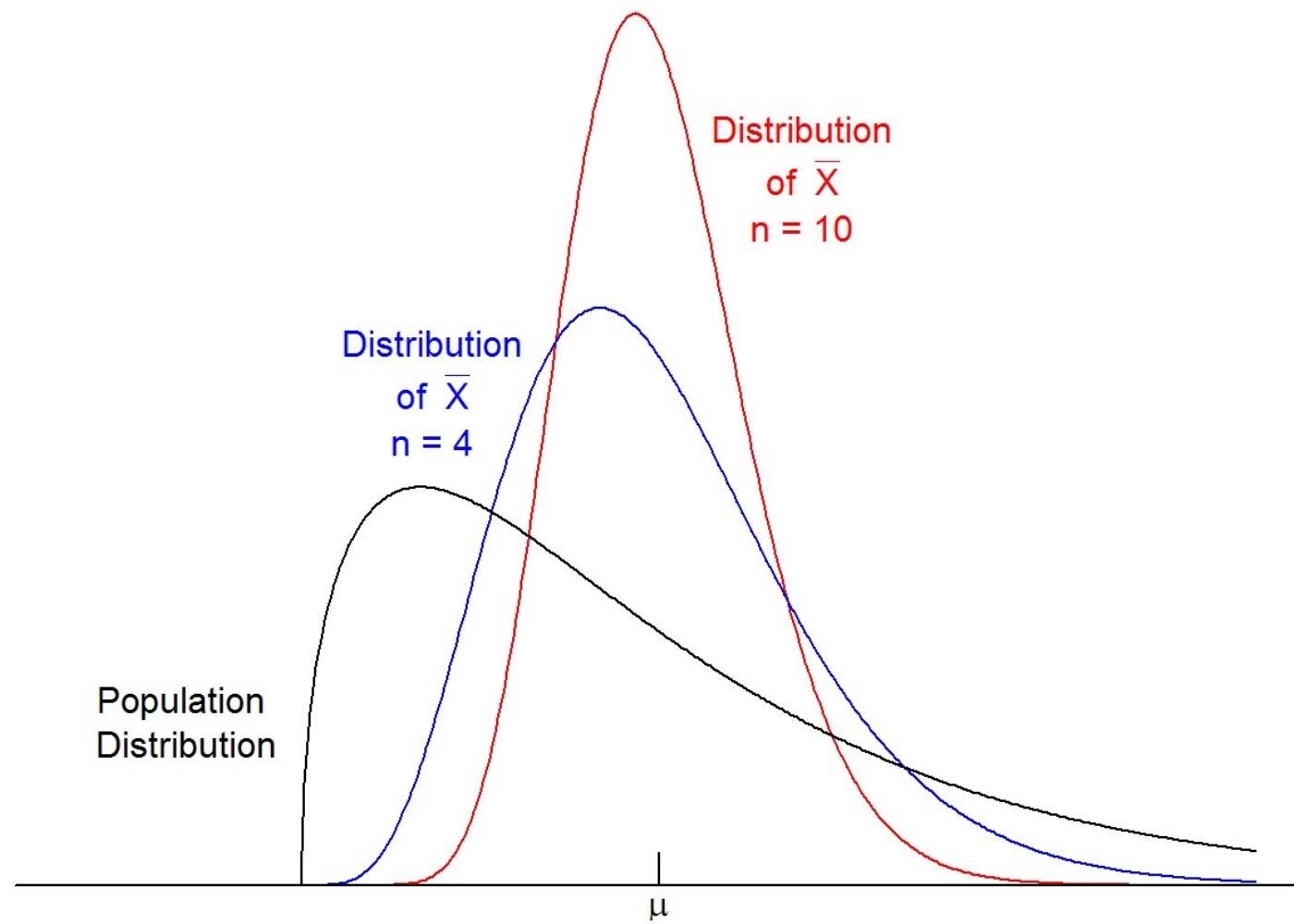
Theorem (*Central Limit Theorem*): If X_1, X_2, \dots, X_n are independent random variables, from a population or process with mean μ and standard deviation σ , then as long as n is sufficiently large

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



and

$$T = \sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$$



Example: The (population) mean time required for maintenance on an air-conditioning unit is 1 hour, and the standard deviation is also 1 hour. A company operates 50 such units.

Could we find the probability that the maintenance on a single unit requires more than 2 hours from the information given?

What is the probability that the average time for the maintenance of the company's 50 AC units will be more than 75 minutes?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	1.25
Cumulative probability: $P(X \leq 1.25)$	0.96145
Mean	1
Standard deviation	.14142

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

What is the probability that the total time for maintenance will be less than 40 hours?

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="40"/>
Cumulative probability: $P(X \leq 40)$	<input type="text" value="0.07865"/>
Mean	<input type="text" value="50"/>
Standard deviation	<input type="text" value="7.0711"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

Example: The amount of warpage in a type of wafer used in the manufacture of integrated circuits has mean 1.3 mm and standard deviation 0.1 mm. A random sample of 200 wafers is drawn.

A.) What is the probability that the sample mean warpage exceeds 1.305 mm? ($\mu = 1.3$, $\sigma = 0.1$, $\sigma^2 = (.1)^2$, $n = 200$)

B.) Find the 25th percentile of the sample mean.

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

▪ Enter a value in three of the four text boxes.
▪ Leave the fourth text box blank.
▪ Click the **Calculate** button to compute a value for the blank text box.

Normal random variable (x)	<input type="text" value="1.295"/>
Cumulative probability: P(X \leq 1.295)	<input type="text" value=".25"/>
Mean	<input type="text" value="1.3"/>
Standard deviation	<input type="text" value=".00707"/>

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

C.) How many wafers must be sampled so that the probability is 0.05 that the sample mean exceeds 1.305?

In this exercise we will use the

Reverse Transformation Formula:

If $X \sim N(\mu, \sigma^2)$, then since $Z = \frac{X-\mu}{\sigma}$, it follows that

$$X = \sigma Z + \mu.$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	1.645
Cumulative probability: $P(Z \leq 1.645)$.95
Mean	0
Standard deviation	1

Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

How large is “large”?

As a general rule, $n \geq 30$ is usually large enough that the Central Limit Theorem is reasonable.

Symmetric populations can get by with much less, often as few as 10, or even fewer.

Highly skewed populations require more. 50 or more should be fairly safe in all but the worst cases, where the population deviates greatly from a normal distribution.

Data analysis

THE GATHERING, DISPLAY, AND SUMMARY OF DATA.

Probability

THE LAWS OF CHANCE, IN AND OUT OF THE CASINO.

Statistical inference

THE SCIENCE OF DRAWING STATISTICAL CONCLUSIONS FROM SPECIFIC DATA, USING A KNOWLEDGE OF PROBABILITY.



Point Estimation

The remainder of the course will focus on **inferential statistics**.

Recall: A **parameter** is an unknown quantity related to a population or distribution.

A **statistic** is a known quantity which can be calculated from a dataset.

Estimation uses a statistic (what we know) to tell us something about an unknown parameter (what we wish we knew).

Definition: A *point estimator* of a parameter θ , is a statistic, $\hat{\theta}$, which provides a “best guess” for θ .

Example: We have an unknown distribution, $X \sim f(x)$, and we wish to know the unknown parameter $\mu = E(X)$. We take a sample X_1, X_2, \dots, X_n , and estimate μ with the known statistic $\hat{\mu} = \bar{X}$.

Other common point estimators:

Estimate $V(X) = \sigma^2$ with $\hat{\sigma}^2 = S^2$, the sample variance discussed at the beginning of the semester.

If $X \sim \text{Binomial}(n, p)$ (n known, p unknown), estimate p with $\hat{p} = \frac{X}{n} = \frac{\# \text{ of successes}}{n}$, the **sample proportion**.

All of our standard sample statistics (median, quartiles, etc.) are good estimators of the corresponding population or distribution parameters.

Properties of Estimators

There are a few properties that we like to see in a point estimator.

- On average (over many samples), an estimator should give the correct value for the parameter. If the mean of the sampling distribution of our estimator is the parameter we are estimating, that is, $E(\hat{\theta}) = \theta$ we say that $\hat{\theta}$ is an *unbiased* estimator of θ .

Example: We know that $E(\bar{X}) = \mu$ so \bar{X} is an unbiased estimator of μ .

Also

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} (np) = p$$

$\Rightarrow \hat{p}$ is an unbiased estimator of p

It can also be shown that

$$E(S^2) = \sigma^2$$

⇒

the sample variance, S^2 , is an unbiased estimator of σ^2

This is one reason why we used $n - 1$ rather than n in the denominator of the definition of S^2 , to get an *unbiased* estimator for σ^2 .

Unfortunately, $E(S) \neq \sigma$, so S is not an unbiased estimator of σ .

Fortunately, the **bias** (defined as $E(S) - \sigma$, or more generally, $E(\hat{\theta}) - \theta$) is small, especially as n gets large.

Note that just because an estimator is unbiased, does not guarantee that it will give you the exact value of the parameter for any given sample.

Example: $X \sim \text{Binomial}(n = 25, p = 0.3)$. Even though \hat{p} is unbiased for p , there is no value of $X (\in \{0, 1, \dots, 25\})$ that will give $\frac{X}{25} = 0.3$

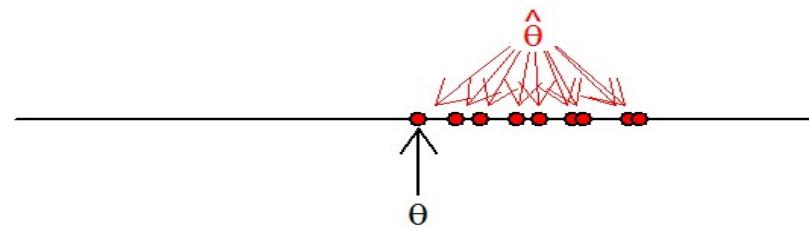
Remember our sampling distributions; an unbiased estimator's distribution will be centered correctly, but it will still have some spread.

- The **variance of the sampling distribution of our estimator** measures that spread and is also important in measuring how well it performs. We prefer low variance in the sampling distribution of our estimator, which results in more precise estimates.

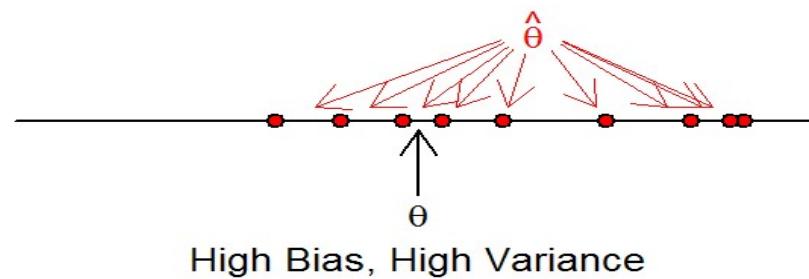


Low Bias, Low Variance

Low Bias, High Variance

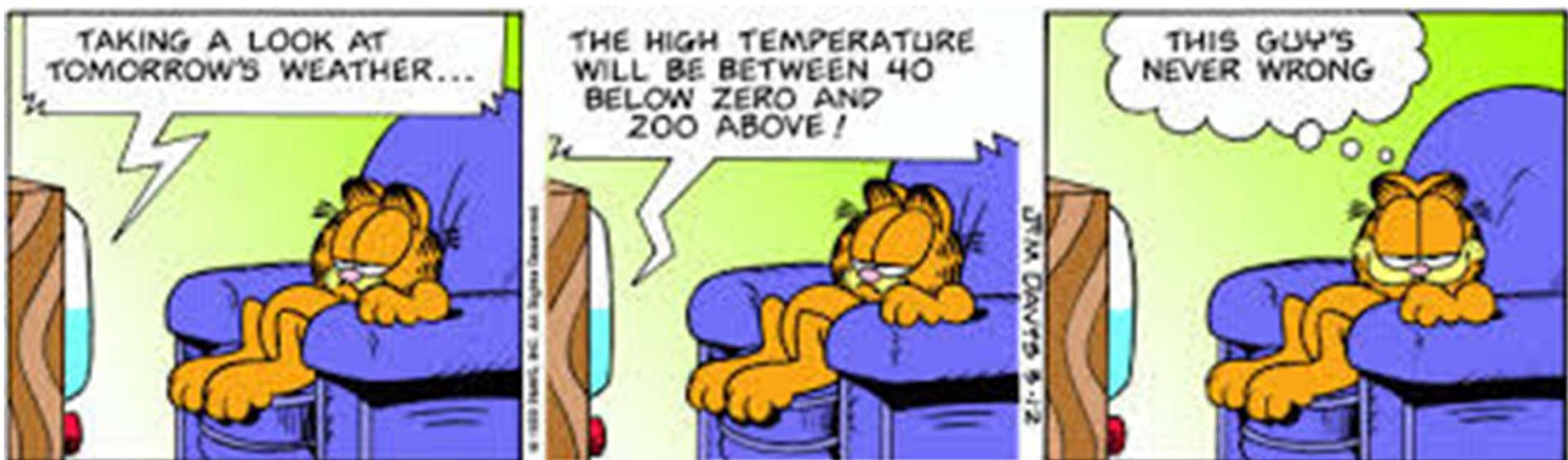


High Bias, Low Variance



High Bias, High Variance

One Sample Confidence Intervals



Having a good estimate is a good first step in learning about a population parameter.

We should also be interested in how close our estimate is likely to be to the parameter.

One approach is to calculate the standard error, remembering that we will usually be within 2-3 standard errors of the parameter (if we use an unbiased estimator).

Another way to look at this issue is that we know our estimate is likely not equal to our parameter. (We just don't know by exactly how much it differs from the parameter's true value.)

We can improve this situation by expanding our point estimate to an *interval estimate*, providing a range of plausible values for θ .

Done carefully, we can identify how likely it is that our interval includes θ .

Let's begin by finding a confidence interval formula for the mean, μ , of a given population based on a **large** sample size, n . Since n is **large**, we can use the Central Limit Theorem to give us the following.

$$\begin{aligned}
 \bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right) \\
 \Rightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}} &\approx \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \\
 \Rightarrow P\left(1.96 \geq \frac{\bar{X}-\mu}{S/\sqrt{n}} \geq -1.96\right) &= 0.95 \\
 \Rightarrow P\left(-\bar{X} + 1.96 \frac{S}{\sqrt{n}} \geq -\mu \geq -\bar{X} - 1.96 \frac{S}{\sqrt{n}}\right) &= 0.95 \\
 \Rightarrow P\left(\bar{X} - 1.96 \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{S}{\sqrt{n}}\right) &= 0.95
 \end{aligned}$$

Therefore, the interval

$$(\bar{X} - 1.96 \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \frac{s}{\sqrt{n}})$$

is a random interval which covers the population mean μ with probability 0.95.

Replacing \bar{X} and S with \bar{x} and s computed from a specific sample we have what is called a **95% confidence interval** for μ .



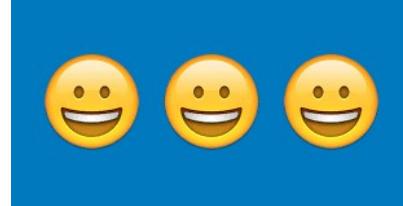
This represents a set of plausible values of μ that are consistent with the data with confidence level .95.

Example: A random sample of 80 auto body shops was selected and each was asked for the cost to repair a particular kind of damage. The mean cost of repair for the 80 shops was \$472.36 with standard deviation \$62.35.

What is the 95% confidence interval for the mean of this population (this is the mean cost of repair for this type of damage over a population of body shops)?

Note: $\bar{x} = \$472.36$ and $s = \$62.35$

Correct interpretation of the interval:



$(\$458.70, \$486.02)$ is a set of plausible (reasonable) values for the mean cost of repair for this particular kind of auto damage which is consistent with our data.

We are 95% confident that $\$458.70 < \mu < \486.02 .

Is it correct to say

$$P(\$458.70 < \mu < \$486.02) = 0.95 ?$$



No! Nothing inside the probability statement is random.

Recall: $P\left(\bar{X} - 1.96 \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{S}{\sqrt{n}}\right) = 0.95$

The random parts are the sample statistics.

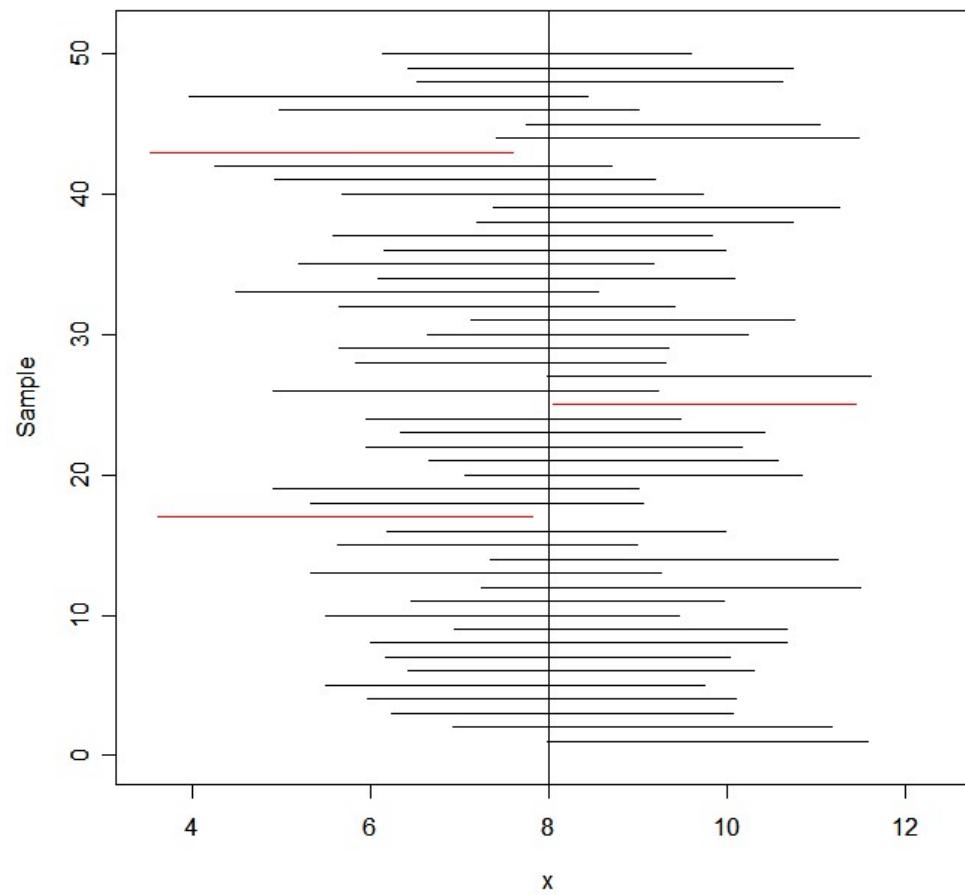
The *interval* is random, not the population parameter, μ .

If we constructed many 95% confidence intervals from independent datasets, we'd get many different sample means and sample standard deviations, and each would lead to a different confidence interval.

In the long run, about 95% of these different confidence intervals would contain the true parameter μ .

Remember, randomness is in the sample and the interval, not in the parameter!

50 95% Confidence Intervals (True Mean = 8, True S.D. = 10)



We call the value 95% the *confidence level*. We say we are 95% confident that the population mean μ lies within the computed interval.

We can select other confidence levels if desired, by replacing the *critical value* 1.96 with the Z-percentile that gives the appropriate center probability.

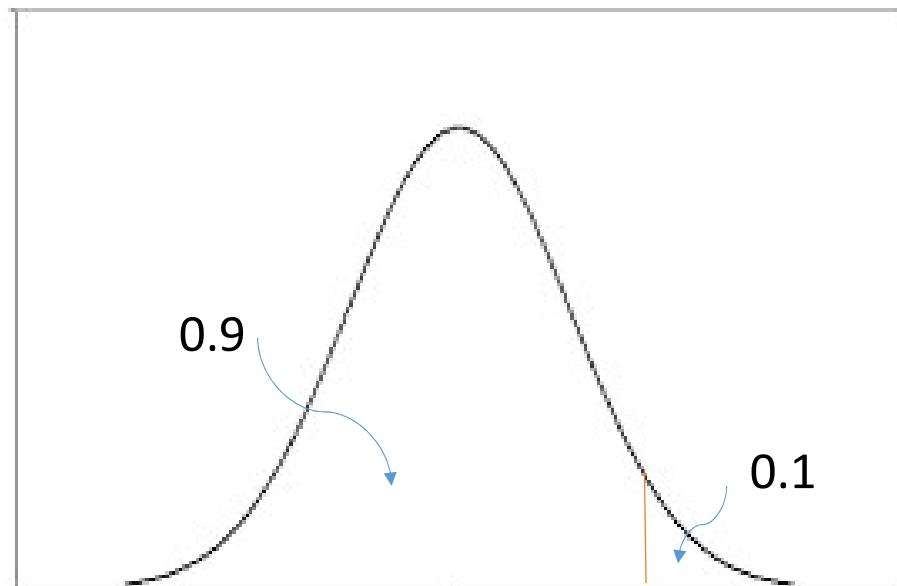
A confidence level of **95% (1.96)** is most common, but levels of **90% (1.645)** and **99% (2.575)** are also often used.

In general, define z_p to be the value, above which there is probability p in the tail of the standard normal distribution.

Then z_p will be the $100(1 - p)^{th}$ percentile of the standard normal distribution.

For a $100(1-\alpha)\%$ confidence interval, we use the critical value $z_{\alpha/2}$.

Example: What critical value would we use for an 80% confidence interval?



$$z_{0.1} = 1.282$$

Normal Distribution Calculator: Online Statistical Table

The Normal Distribution Calculator makes it easy to compute cumulative probability, given a normal random variable; and vice versa. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the normal distribution, go to Stat Trek's [tutorial on the normal distribution](#).

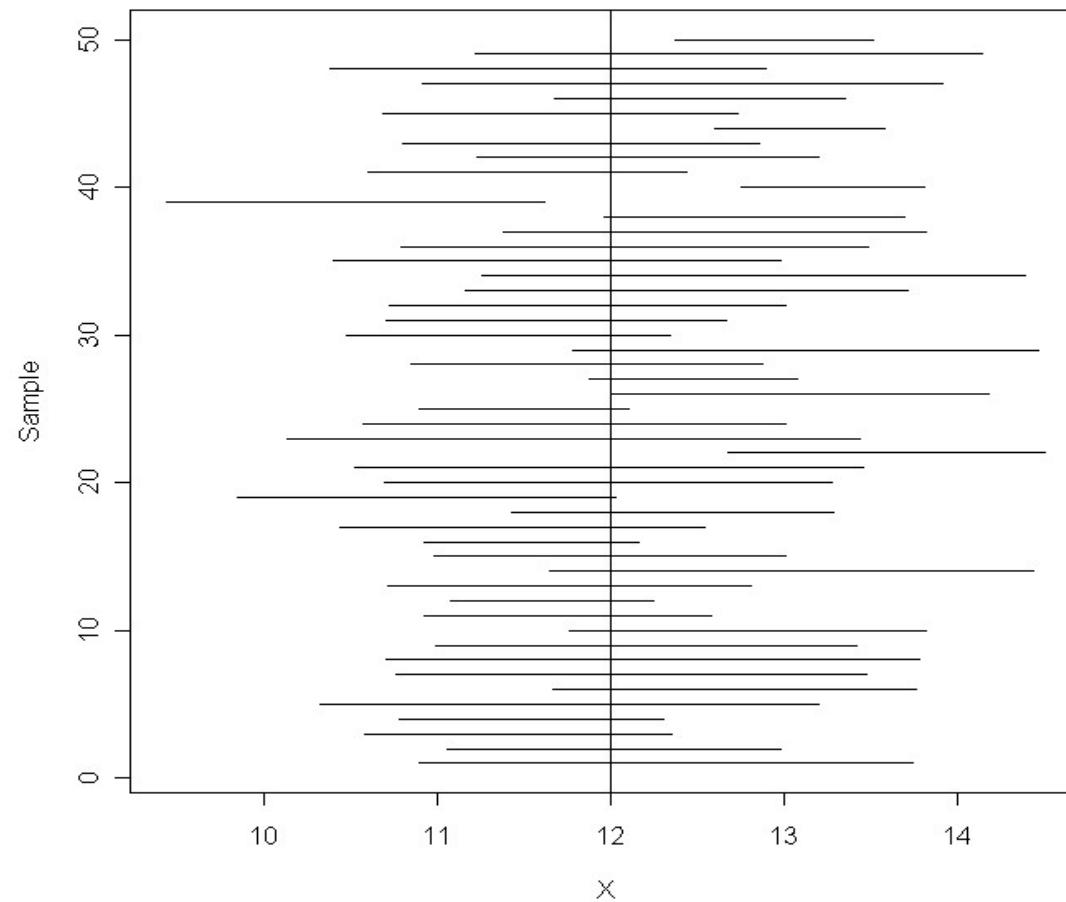
- Enter a value in three of the four text boxes.
- Leave the fourth text box blank.
- Click the **Calculate** button to compute a value for the blank text box.

Standard score (z)	<input type="text" value="1.282"/>
Cumulative probability: $P(Z \leq 1.282)$	<input type="text" value=".9"/>
Mean	<input type="text" value="0"/>
Standard deviation	<input type="text" value="1"/>

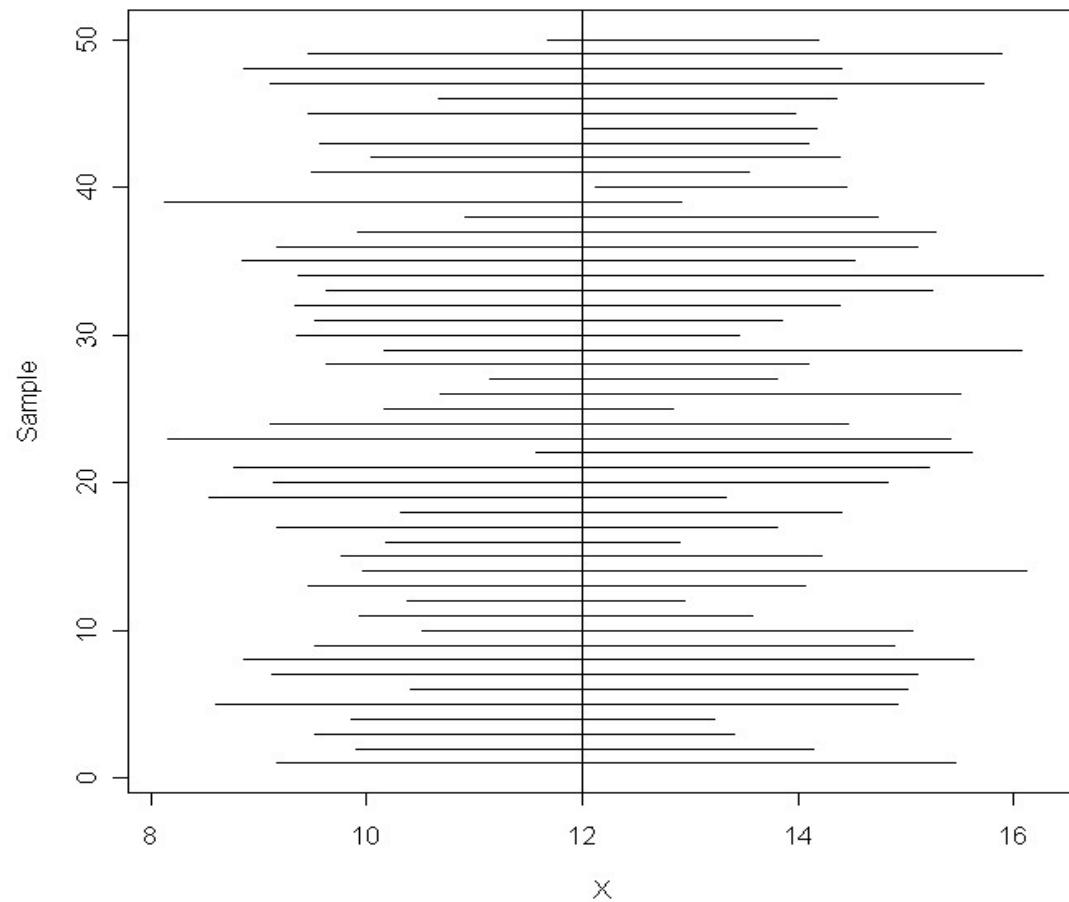
Calculate

Note: The normal distribution table, found in the appendix of most statistics texts, is based on the [standard normal distribution](#), which has a mean of 0 and a standard deviation of 1. To produce outputs from a standard normal distribution with this calculator, set the mean equal to 0 and the standard deviation equal to 1.

50 90% Confidence Intervals (True Mean = 12)



50 99% Confidence Intervals (True Mean = 12)



What factors affect the length (precision) of the confidence interval?

$$\left(\bar{X} - 1.96 \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \frac{s}{\sqrt{n}} \right)$$

- s – If s is bigger, \bar{X} is less accurate, and the interval must be wider.
- **Confidence level** – To be more confident of including the true value, we must make the interval wider.
- n – as n gets bigger, the standard error of \bar{X} gets smaller, and the interval gets narrower.

If we require a 95% confidence interval of error (interval half-width) no more than w , we can compute a (rough) minimum sample size if we have an estimate or upper bound for s .

$$\text{Since } w = 1.96 \frac{s}{\sqrt{n}}, n = \left(\frac{1.96s}{w} \right)^2.$$

Of course, we can substitute the appropriate Z critical value to find sample sizes for other confidence levels.

Example: Milk fill amounts: $n = 50$, $\bar{x} = 2.0727$, $s = 0.0711$
Find a 95% confidence interval for μ .

If we require $w \leq 0.01$, how big should n be?

Example: One step in the manufacture of a certain metal clamp involves the drilling of four holes. In a sample of 150 clamps, the average time needed to complete this step was 72 seconds and the standard deviation was 10 seconds.

- (a) Find a 95% confidence interval for the mean time needed to complete the step.

(b) Find a 99.5% confidence interval for the mean time needed to complete this step.

(c) What is the confidence level of the interval (71, 73)?

(d) How many clamps must be sampled so that a 95% confidence interval specifies the mean to within ± 1.5 seconds.

(e) Repeat part (d) with 99.5% confidence.

Upper and Lower Confidence Bounds

Sometimes, we only wish to know a lower (or upper) bound for μ .

We can generate *one-sided confidence intervals*, also called **confidence bounds**, in a similar way to the usual two-sided case.

If we have a large sample, then:

- A $(1 - \alpha) \times 100\%$ lower confidence bound for μ is:

$$LCB = \bar{x} - z_\alpha \left(\frac{s}{\sqrt{n}} \right)$$

(We are $(1 - \alpha) \times 100\%$ confident that $\mu > \bar{x} - z_\alpha \left(\frac{s}{\sqrt{n}} \right)$.)

- A $(1 - \alpha) \times 100\%$ upper confidence bound for μ is:

$$UCB = \bar{x} + z_\alpha \left(\frac{s}{\sqrt{n}} \right)$$

(We are $(1 - \alpha) \times 100\%$ confident that $\mu < \bar{x} + z_\alpha \left(\frac{s}{\sqrt{n}} \right)$.)

If we wanted to compute a 95% LCB or UCB for μ we would use $z_\alpha = z_{.05} = 1.645$ for our critical value. Note that this amounts to replacing the 1.96 critical value used for a 95% confidence interval with 1.645.

To get 90%, 99%, or $100(1-\alpha)\%$ bounds, replace 1.645 with 1.28, 2.575 with 2.33, or $z_{\frac{\alpha}{2}}$ with z_α , respectively.

Example: A sample of 48 Shear strength measurements give a mean of 17.17 N/mm^2 and a standard deviation of 3.28 N/mm^2 .

If we only care that the population mean shear strength is great enough, find a 95% lower bound on μ .

Example: One step in the manufacture of a certain metal clamp involves the drilling of four holes. In a sample of 150 clamps, the average time needed to complete this step was 72 seconds and the standard deviation was 10 seconds.

- (a) Find a 98% lower confidence bound for the mean time to complete the step.

(b) An efficiency expert says that the mean time to complete this step is greater than 70 seconds. With what level of confidence can this statement be made?

For our normal-based confidence interval and level to be valid, we must know (or at least assume) that:

- The sample is a random draw from the population.
- The sample size n is large enough that the sample mean is approximately normally distributed and that s is a good estimate of σ .

Confidence Intervals for Population Proportion

We have:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right),$$

and

$$P\left(1.96 \geq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq -1.96\right) \approx 0.95.$$

Isolating p in the probability statement is trickier than in the case for μ , because it appears in both the numerator and the denominator. It can be done, but the interval is tedious and we will use an alternative method.

In the past, people usually replaced the unknown p 's in the standard error with \hat{p} , so the 95% confidence interval for p would be

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

This “traditional” interval has the same format as the μ -interval:

$$\hat{\theta} \pm \text{critical-value} \times \text{standard error of } \hat{\theta}.$$

It works well for large n , but is no longer recommended.



Unfortunately, the probability of containing p in the interval can be well below 95% for smaller n .

It turns out this can be corrected by adding 2 successes and 2 failures to our counts:

$$\tilde{n} = n + 4, \quad \tilde{p} = \frac{X + 2}{\tilde{n}}.$$

Then a $100(1-\alpha)\%$ confidence interval for p will be

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

for all n .



Example: A sample of 125 building tiles finds 10 cracked. Find a 99% confidence interval for p , the probability that an individual tile is cracked.

Example: Suppose we flip a (possibly biased) coin. Let p be the probability of a head. If 100 tosses result in 45 heads, find a 95% confidence interval for p . Is it plausible that our coin could be fair?

If we have an idea of the value of p (such as a value \tilde{p} from a pilot sample), and we require a 95% confidence interval with an error bound (interval half-width) no more than w , we can compute a minimum sample size needed to achieve this.

$$\text{Since } w = 1.96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}},$$

$$\tilde{n} = \tilde{p}(1-\tilde{p}) \left(\frac{1.96}{w} \right)^2, \Rightarrow n = \tilde{n} - 4.$$

Of course, we can substitute the appropriate z critical value to find sample sizes for other confidence levels.

If p is unknown, we should replace it with the conservative value 0.5, and require a minimum sample size of

$$\tilde{n} = 0.5(1-0.5)\left(\frac{1.96}{w}\right)^2 = \frac{0.9604}{w^2}, \quad \Rightarrow \quad n = \tilde{n} - 4.$$

The associated error bound of

$$w = 1.96 \sqrt{\frac{0.5(1-0.5)}{\tilde{n}}} = \frac{0.98}{\sqrt{\tilde{n}}}$$

is the “margin of error” that surveys and polls generally report.

Example: If we take a survey and want a 2% margin of error ($w = 0.02$), how big a sample must we take?

$$\tilde{n} = 0.5(1 - 0.5) \left(\frac{1.96}{.02} \right)^2 = \frac{0.9604}{.0004}$$

What if we're willing to settle for a 3% margin of error?

$$\tilde{n} = 0.5(1 - 0.5) \left(\frac{1.96}{.03} \right)^2 = \frac{0.9604}{.0009}$$

Small sample (t) confidence intervals for μ

- Strictly speaking, the confidence intervals for means we've looked at require that we know the standard deviation, σ , for our population.
- In practice, for large samples, s is a good enough estimate for σ that we can use s without harming our interval coverage severely.
- If n is small, s may be far off from σ , and we require an adjustment to our intervals that takes into account this uncertainty.

Since in practice, both μ and σ are usually unknown, when n is small ($n < 30$) we often use the following result:

If $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ are independent, then

$$T = \frac{(\bar{X} - \mu)}{S / \sqrt{n}}$$

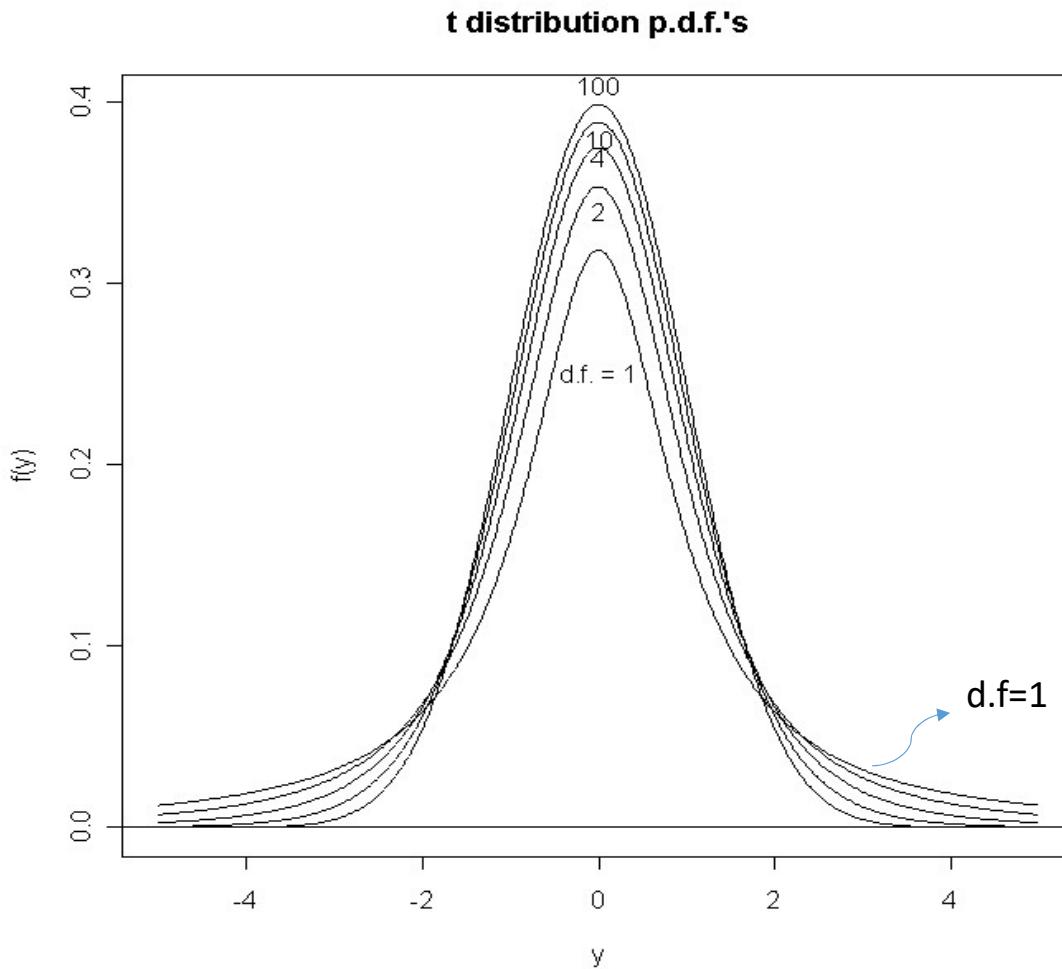
has a *t distribution with $n-1$ degrees of freedom.*

The t_v distribution has a bell-shaped curve with heavier tails than the standard normal distribution.

It is always centered on, and symmetric around, 0.

The t_1 curve is most spread out (has the heaviest tails). As v gets larger, the tails get lighter and the curve gets less spread out.

As $v \rightarrow \infty$, the t_v distribution approaches the standard normal distribution.



Example: $T \sim t_{12}$

$$P(T \geq 2.681)$$

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

■ In the dropdown box, describe the random variable.

■ Enter a value for degrees of freedom.

■ Enter a value for all but one of the remaining text boxes.

■ Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable	<input type="text" value="t score"/> ✓
Degrees of freedom	<input type="text" value="12"/>
t score	<input type="text" value="2.681"/>
Cumulative probability: $P(T \leq 2.681)$	<input type="text" value="0.9900"/>
<input type="button" value="Calculate"/>	

Example: $T \sim t_9$

$$P(T \leq 1.833) = .95$$

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable

Degrees of freedom

t score

Cumulative probability: $P(T \leq 1.833)$

$$P(-2.170 < T < 2.262)$$

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable	<input type="text" value="t score"/>
Degrees of freedom	<input type="text" value="9"/>
t score	<input type="text" value="2.262"/>
Cumulative probability: P(T ≤ 2.262)	<input type="text" value="0.9750"/>
<input type="button" value="Calculate"/>	

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable	<input type="text" value="t score"/>
Degrees of freedom	<input type="text" value="9"/>
t score	<input type="text" value="-2.170"/>
Cumulative probability: P(T ≤ -2.170)	<input type="text" value="0.0291"/>
<input type="button" value="Calculate"/>	

Find $t_{9,.025}$.

T Distribution Calculator: Online Statistical Table

The t distribution calculator makes it easy to compute cumulative probabilities, based on t statistics; or to compute t statistics, based on cumulative probabilities. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about Student's t distribution, go to Stat Trek's [tutorial on the t distribution](#).

- In the dropdown box, describe the random variable.
- Enter a value for degrees of freedom.
- Enter a value for all but one of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Describe the random variable

Degrees of freedom

t score

Cumulative probability: $P(T \leq t)$

Let \bar{x} and s be the sample mean and standard deviation from a sample of size n from a **normal population** or process. Then a confidence interval for the population mean μ has the form

$$\bar{x} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}.$$

Very Important!

Note this is the same form as for the usual z -interval. The only difference is the replacement of the usual normal (z) critical value (such as 1.96) from one found on the t -table or using Stat Trek with $(n-1)$ degrees of freedom.

One-sided intervals or confidence bounds may be found by taking only the appropriate endpoint (+ or -) and choosing the one-sided t critical value ($t_{n-1,\alpha}$).

Example: An experiment is conducted to investigate the mileage to end-of-life of tires made using a new rubber compound. A sample of size 10 finds a mean of 61,492 miles and a standard deviation of 3,035 miles. Assume a normal model is appropriate.

- (a) Find a 95% confidence interval for the population mean tire life.

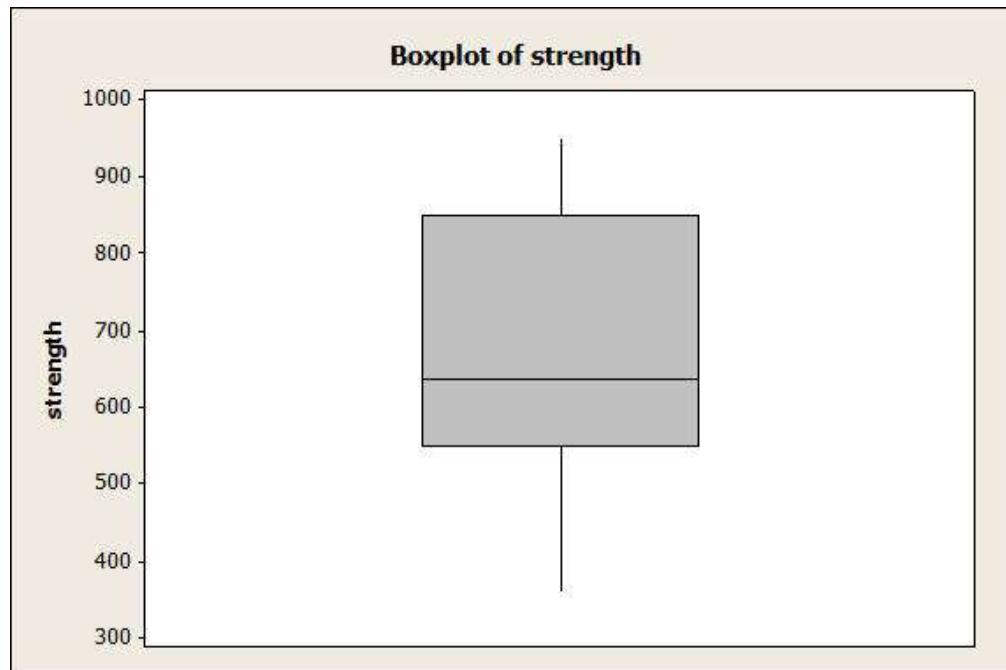
(b) If we only care about a minimum, find a 95% lower bound for population mean tire life.

Example: A brand of margarine was analyzed to determine the level of polyunsaturated fatty acid. In 6 samples, the mean percent is 16.98%, and the s.d. is 0.32%. A normal distribution is reasonable for this variable. ($n = 6$, $\bar{x} = 16.98$, $s = .32$)

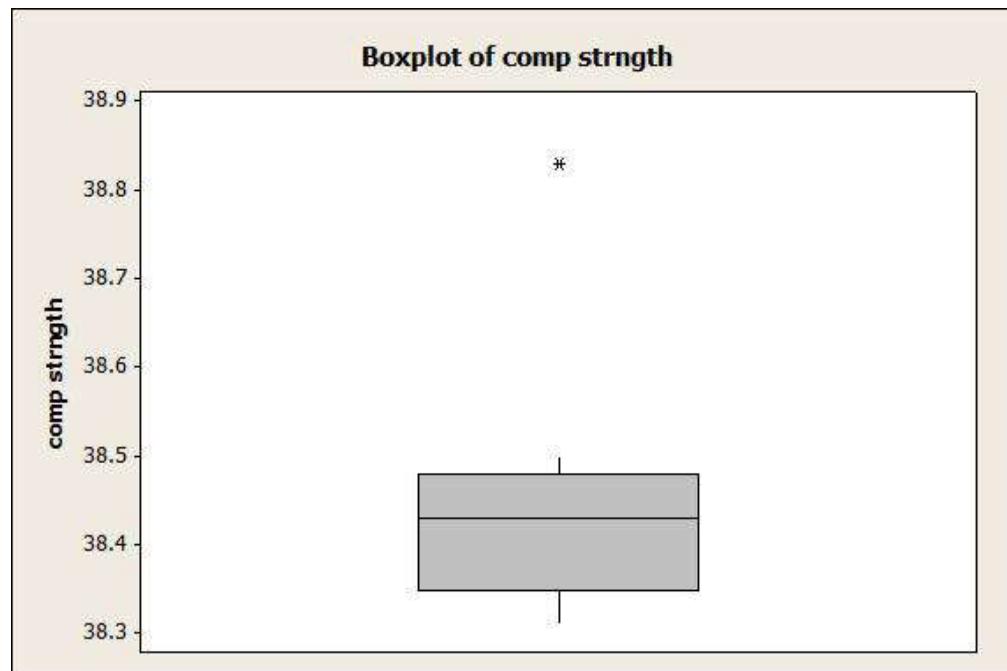
Find a 99% confidence interval for population mean percent pfa.

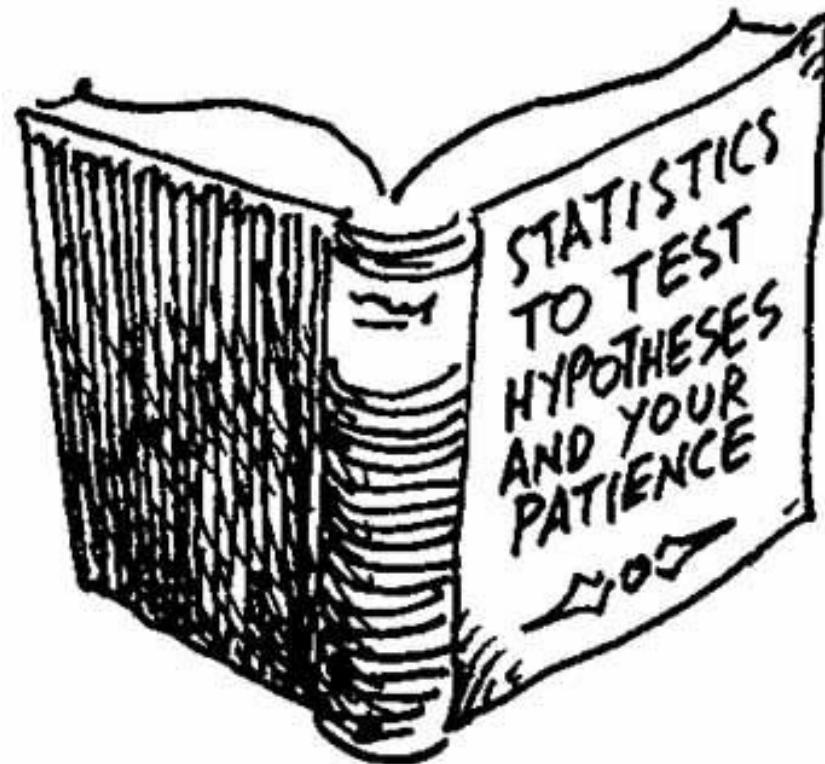
Find a 90% upper bound for population mean percent pfa.

Example: A sample of the nominal shear strength of 15 prestressed concrete beams is collected. Suppose we want to compute a confidence interval for the mean nominal shear strength of the population of such beams. Should a t based CI be used if a boxplot of the data were as follows?



A sample of the cylindrical compressive strength (in MPa) for 11 beams is collected. Suppose we want to construct a CI for the mean cylindrical compressive strength for the population of such beams. Should a t-based CI be used given the following boxplot?





Hypothesis Tests

Estimation (both point and interval) is useful for providing an idea of the value of a population parameter.

Frequently, we may wish to investigate a more specific question about a parameter. For this purpose, we use the other major branch of inferential statistics, *hypothesis testing*.

One Sample Z-test

Example: (Milk data) Suppose our bottle-filling machine is set to dispense 2.04 L of milk. Recall, a sample of size 50 gave $\bar{x} = 2.0727$, $s = 0.0711$. Does the machine need to be recalibrated?

To answer this, let's assume that the machine is working properly, and see how likely we are to get a sample mean as far or further from the expected value as the sample mean we actually saw (2.0727).

More formally, we choose a *null hypothesis*, H_0 .
This is a statement about a population parameter (say, μ),
generally that it is equal to the value of interest (denoted μ_0).

Usually, the null hypothesis means everything is as it should be, or nothing interesting is happening.

Here:

$$H_0: \mu = 2.04$$

We also choose an *alternative hypothesis*, H_1 , that the null is incorrect.

For our null hypothesis:

$$H_1: \mu \neq 2.04$$

The alternative is literally simply that the null is incorrect, but this is often the more interesting or important result.

Next, we compute a *test statistic*, under the assumption that H_0 is correct.

For large-sample tests on the population mean, μ , we usually use the z-statistic:

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

In our example:

$$z = \frac{2.0727 - 2.04}{\frac{.0711}{\sqrt{50}}} = 3.252$$

If H_0 is true,

$$Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim N(0, 1)$$

Is $z = 3.252$ a “typical” value from a $N(0, 1)$ distribution?

Since nearly 100% of the standard normal values are within 3 standard deviations of the mean, i.e., $P(-3 < Z < 3) \approx 1$, $z = 3.252$ seems too large to be a “typical” value of Z .

Formally, we find a *P-value*, the probability that a sample from the null distribution would give a test statistic as unusual as (or more unusual than) the one we just saw.

Since $H_1: \mu \neq 2.04$, we use a *two-sided*

P-value: $P = P(|Z| \geq 3.252)$ ($Z \sim N(0,1)$).

From our Stat Trek table, if $Z \sim N(0,1)$,

$$\begin{aligned}P(|Z| \geq 3.252) &= P(Z \leq -3.252) + P(Z \geq 3.252) \\&= 2P(Z \leq -3.252) = .00114.\end{aligned}$$

So we have two possibilities:

- 1) H_0 is correct, $\mu = 2.04$, and we were very unlucky to get the (roughly) 1 in 850 chance of getting $\bar{X} \geq 2.0727$ (or the equally unusual $\bar{X} \leq 2.0037$), or
- 2) H_0 is wrong.

Which seems more reasonable to believe?

Since P is so small, we *reject* H_0 and decide that the filling machine does require recalibration.

All hypothesis tests follow this general pattern:

1. We observe some difference in a sample and wish to decide if it reflects a true difference in the population.
2. Identify the null and alternative hypotheses.
3. Compute a test statistic which has a known distribution when the null hypothesis is true.
4. Find a P -value: the probability of a statistic as or more unusual than the one we observed, when the null hypothesis is true.
5. If P is small, reject the null hypothesis. Otherwise, fail to reject it.

This basic pattern holds for many different tests on different parameters with different assumptions.

For questions about the population mean for a single population, we often use the *one sample z-test* demonstrated in the previous example.

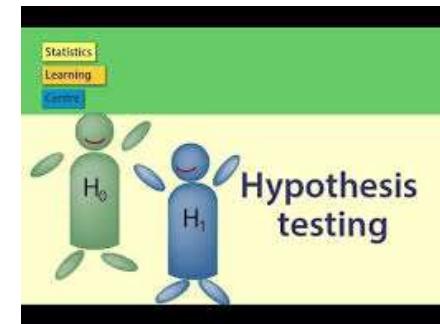
Details on the one-sample z-test:

1) We have a single population, and a specific value, μ_0 , we wish to consider for the population mean.

- This may be a known population mean for some related population (see next example).
- Or it may be a desired or target population mean (example: milk data).
- A sample from the population will likely give a sample mean different from μ_0 , even if that is the actual population mean.

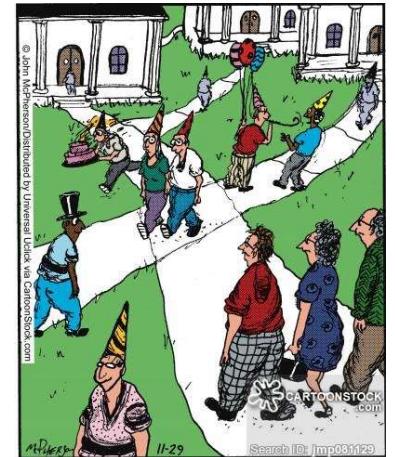
2) Identify H_0 and H_1 .

- H_1 is a statement that something interesting is going on. It is usually what we wish to prove.
- We should decide if we care about a one-sided or two-sided alternative, ideally before we ever see data.
- Two-sided: $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$.
- One-sided: $H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$
- or: $H_0: \mu \geq \mu_0$ vs. $H_1: \mu < \mu_0$
- We always compute z and P assuming $\mu = \mu_0$, so $\mu = \mu_0$ is always part of H_0 .



Example: A newspaper article says that college freshmen average 7.5 hours per week at parties. We suspect the number is lower at our college.

$$H_0 : ? \quad H_1 : ?$$



As they toured Burfman College, Brad and his parents could tell it was a big party school.

3) Compute the test statistic.

- For a one-sample z-test, use the z statistic with μ_0 :

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}.$$

- If σ is unknown, use s instead.

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

Example (cont): Interview 100 freshmen. The average reported time spent at parties is 6.6 hours, and the standard deviation is 9 hours.

4) Find the P-value.

- Under H_0 , $Z \sim N(0,1)$. Use probabilities $Z \sim N(0,1)$, depending on H_1 :

H_1	P-value
$\mu \neq \mu_0$	$2P(Z \geq z) = 2 P(Z \leq - z)$
$\mu > \mu_0$	$P(Z \geq z) = P(Z \leq -z)$
$\mu < \mu_0$	$P(Z \leq z)$

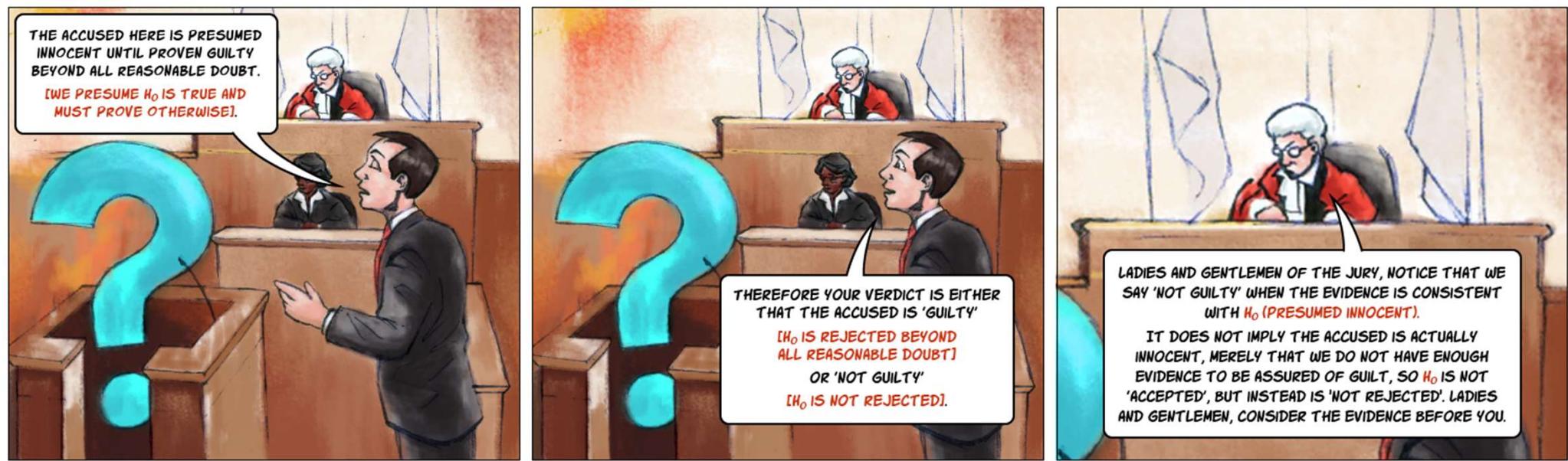
Example (cont):

5) Reject H_0 for small *P-values*.

Choose a small *significance level, α* . Values of 0.05 or 0.01 are most commonly used.

- If $P \leq \alpha$, the evidence is pretty strong against H_0 , and we say we *reject H_0* (at the α significance level). We have strong evidence in support of H_1 .
- If $P > \alpha$, our test statistic is pretty reasonable under H_0 , so we say we *fail to reject H_0* .

- Note: A large P-value is not proof of H_0 ; many other hypotheses may also be reasonable. This is why we do not say that we *accept* H_0 .



- Many people call any result with $P < 0.05$ *statistically significant*, and any with $P < 0.01$ *highly (statistically) significant*.
- Note that this is very artificial. A P -value of 0.049 is only slightly stronger than one of 0.051, yet we treat them very differently.
- We should always report the P -value, to provide full information.

Example (Partying freshmen test, cont): What does our P -value suggest about our hypotheses?

You should always explain in words what your conclusion implies for the situation.

What does this say about the party habits of freshmen at our university?

Since $P = 0.1587$, is it correct to say $P(\mu \geq 7.5) = .1587$?

In general, with P-value P , can we say $P(H_0 \text{ is true}) = P$?

Example: A machine that produces metal cylinders is set to make cylinders with diameter 50 mm. A random sample of 60 cylinders has $\bar{x} = 49.9865$ and $s = 0.0524$. Is the machine calibrated correctly?

Note that *practical significance* is not the same as *statistical significance*.

Even though we found statistical significance in support of the machine's recalibration, it may be that the difference between 50 mm and 49.9865 mm is too small to justify the expense of recalibration.

Large samples are particularly prone to indicate statistical significance despite a difference too small to be important.

Conversely, small samples may come with large standard errors, so that a difference which might be very important if confirmed cannot be shown to be statistically significant.

We should supplement our test with a confidence interval, which will do a much better job of indicating the size and therefore importance of a potential difference.

Example: Cylinder: $n = 60$, $\bar{x} = 49.9865$ and $s = 0.0524$.
What is a 95% confidence interval for μ ?

Note that a $100(1 - \alpha)\%$ confidence interval for μ will include (exclude) μ_0 exactly whenever a two-sided test of

$H_0: \mu = \mu_0$ fails to reject (rejects) at the α level.

Similarly, a $100(1 - \alpha)\%$ lower bound will fall below (above) μ_0 exactly whenever a one-sided test of $H_0: \mu \leq \mu_0$ fails to reject (rejects) at the α level.

Likewise for upper bounds and tests on $H_0: \mu \geq \mu_0$.

Cylinder Example

Partying Freshmen Example

Z-tests for Proportions

Recall that if $X \sim \text{Binomial}(n, p)$ and

$$\hat{p} = \frac{X}{n}, \quad \text{then} \quad \mu_{\hat{p}} = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

If $np \geq 10$ and $n(1 - p) \geq 10$, then n is large enough that the Central Limit Theorem tells us that \hat{p} is approximately normal as well.

We can use this sampling distribution to do hypothesis tests on p . Since we use the normal table, these will be z-tests.

1) We have a single population and a specific value, p_0 , we wish to consider for the population probability of success or population proportion of successes.

A sample from the population will likely give a sample proportion different from p_0 , even if that is the actual population proportion.

Example: We have a possibly biased coin. We wish to test whether or not

$$p = P(\text{Heads}) = 0.5 = p_0.$$

2) Identify H_0 and H_1 .

- Set up H_1 as the statement that something interesting is going on, or what we wish to prove.
- Choose a one-sided or two-sided alternative, depending on our purpose.
- Two-sided: $H_0: p = p_0$ vs. $H_1: p \neq p_0$.
- One-sided: $H_0: p \leq p_0$ vs. $H_1: p > p_0$
or: $H_0: p \geq p_0$ vs. $H_1: p < p_0$

Example: Coin: $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$.

3) Compute the test statistic.

- Just as with tests on μ , we find the z test statistic assuming H_0 is true.

$$z = \frac{\text{point estimate} - \text{null value}}{\text{standard error under } H_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Example: 400 coin flips, 176 heads.

4) Find the P-value.

- This is still a z-test. Under H_0 , $Z \sim N(0,1)$. Use probabilities on $Z \sim N(0,1)$, depending on H_1 :

H_1	P-value
$p \neq p_0$	$2P(Z \geq z) = 2 P(Z \leq - z)$
$p > p_0$	$P(Z \geq z) = P(Z \leq -z)$
$p < p_0$	$P(Z \leq z)$

Example: $z = -2.4$.

5) Choose a small α , and reject H_0 for $P < \alpha$. We have strong evidence against the null hypothesis.

Otherwise, fail to reject H_0 . The null hypothesis is plausible.

If $\alpha = 0.05$, what should we conclude?

Minitab Output for Z-test

Test and CI for One Proportion

Test of $p = 0.5$ vs $p \neq 0.5$

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	176	400	0.440000	(0.391355, 0.488645)	-2.40	0.016

Minitab will do the same test using exact binomial probabilities for a slightly more accurate P-value.

Test and CI for One Proportion

Test of $p = 0.5$ vs $p \neq 0.5$

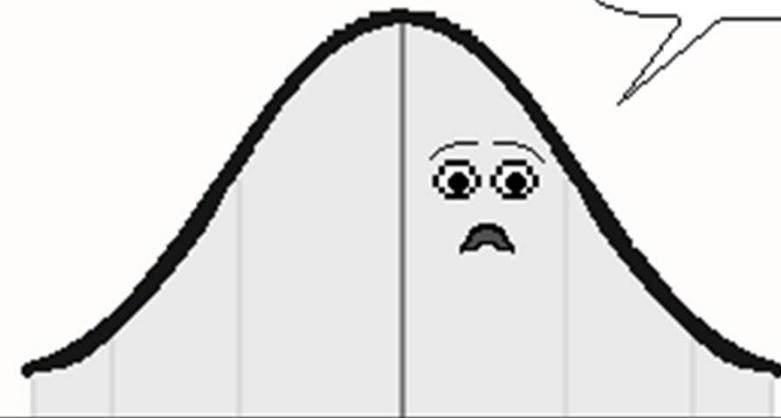
Sample	X	N	Sample p	95% CI	Exact P-Value
1	176	400	0.440000	(0.390707, 0.490187)	0.019

Example: 600 students, 217 knew the author of *The Canterbury Tales*. Should we believe that more than 1/3 of all students know this?

(x, why?)



Mr. Whiskers, at school
today, I failed my t-test!



(C)Copyright 2014, C. Burke 6/3

t-Tests

We should **use the *t* distribution** to conduct hypothesis tests when n is small, and σ is unknown.

This is especially important when $n < 30$, but can be used for any sample size.

We will continue to require the **assumption of a normal population**.

The process of conducting a t -test is identical to conducting a z -test, except for step 4, computation of the P-value.

1. We have a single population, and a specific value, μ_0 , we wish to consider for the population mean.
2. Identify H_0 and H_1 just as for a z -test.
3. Compute the test statistic

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

4) Find the P-value.

- Under H_0 , $T \sim t_{n-1}$. Use probabilities on $T \sim t_{n-1}$, depending on H_1 :

H_1	P-value
$\mu \neq \mu_0$	$2P(T \geq t) = 2 P(T \leq - t)$
$\mu > \mu_0$	$P(T \geq t) = P(T \leq -t)$
$\mu < \mu_0$	$P(T \leq t)$

5) Reject H_0 for small P .

Example: A car manufacturer claims a model gets 35 mpg. A consumer group wishes to test this claim. We measure 14 cars, find $\bar{x} = 34.271$ mpg and $s = 2.915$ mpg.

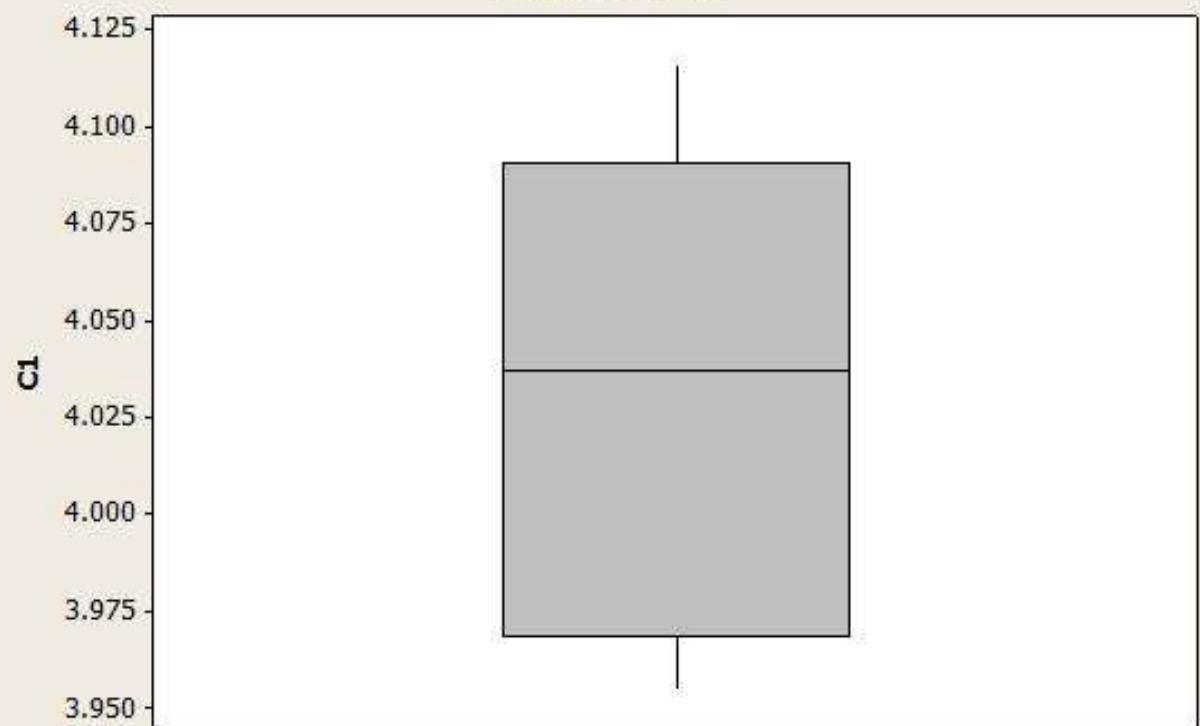
Example: Specs call for the wall thickness of two-liter polycarbonate bottles to average 4.0 mils. A control engineer samples 7 two-liter poly bottles from a large batch and measures the wall thickness for each. It is desired to test

$$H_0: \mu = 4.0 \text{ vs } H_1: \mu \neq 4.0$$

Construct a boxplot for the data.

3.999, 4.037, 4.116, 4.063, 3.969, 3.955, 4.091

Boxplot of C1



No reason to doubt normality assumption

Using Minitab to conduct the test:

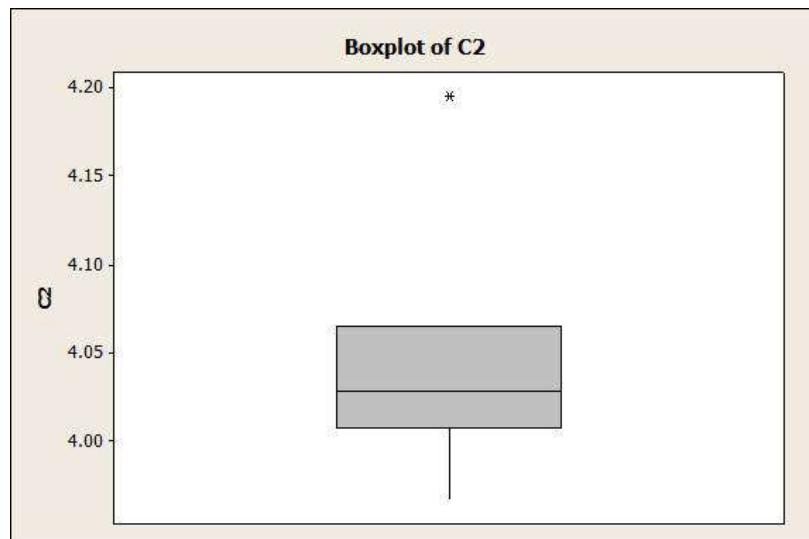
One-Sample T: Wall Thickness (C1)

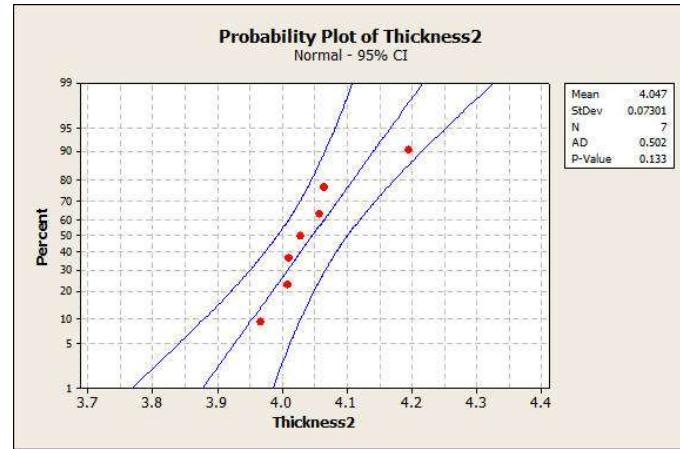
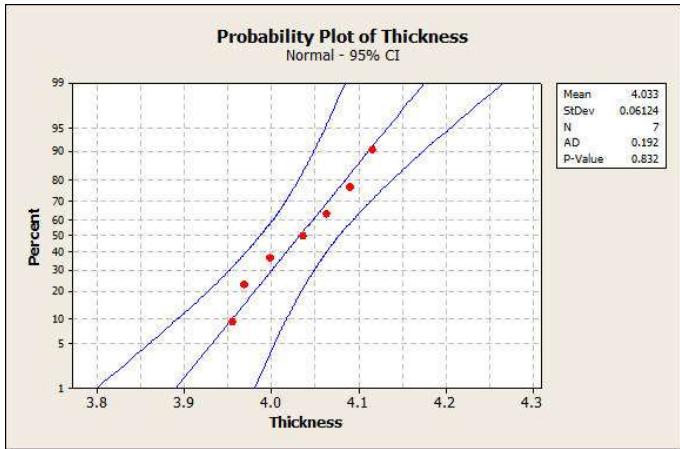
Test of $\mu = 4$ vs not = 4

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
C1	7	4.03286	0.06124	0.02315	(3.97622, 4.08950)	1.42	0.206

Another sample gave the measurements

4.065, 3.967, 4.028, 4.008, 4.195, 4.057, 4.010





Normal probability plots can also be used to assess whether the sample came from a normal distribution. Look for the plotted points to be roughly linear. Any major deviations from a linear pattern would constitute a red flag for the normal assumption. The results of a formal hypothesis test that the data came from a normal distribution are given in the small box in the upper right corner of the plot. If the p-value is small, that would suggest that the normal assumption is not satisfied.

For these two plots, the first plot is reasonably linear and the p-value=.832. Thus there is no reason to question the normal assumption. In the second plot, there is one point that causes the points to deviate from a linear pattern and the p-value=.133. The p-value isn't particularly "small" enough to indicate a violation of the normal assumption, but it is greatly reduced from the first sample. For the second sample, it would be best to err on the side of caution and avoid the t-test.

Power of a Test



After years of worrying sick about the worried well, some days Rhona felt almost cynical

Example: The installation of a radon abatement device is recommended in any home where the mean radon concentration is 4.0 picocuries per liter or more, because it is thought that long-term exposure to sufficiently high doses of radon can increase the risk of cancer. Seventy-five measurements are made in a particular home. The mean concentration was 3.72 pCi/L, and the standard deviation was 1.93.

- a) The home inspector who performed the test says that since the mean measurement is less than 4.0, radon abatement is not necessary. Explain why this reasoning is incorrect.

b.) Because of health concerns, radon abatement is recommended whenever it is plausible that the mean radon concentration may be 4.0 pCi/L or more. State the appropriate H_0 and H_1 for determining if abatement is appropriate.

c.) Compute the p-value. Would you recommend abatement?
Explain. (Recall: $\bar{x} = 3.72$, $s = 1.93$, and $n = 75$.)

Type I and Type II Errors

		Truth	
		H_0 True	H_1 True
Decision	Reject H_0	Type I Error	Correct Decision
	Fail to Reject H_0	Correct Decision	Type II Error

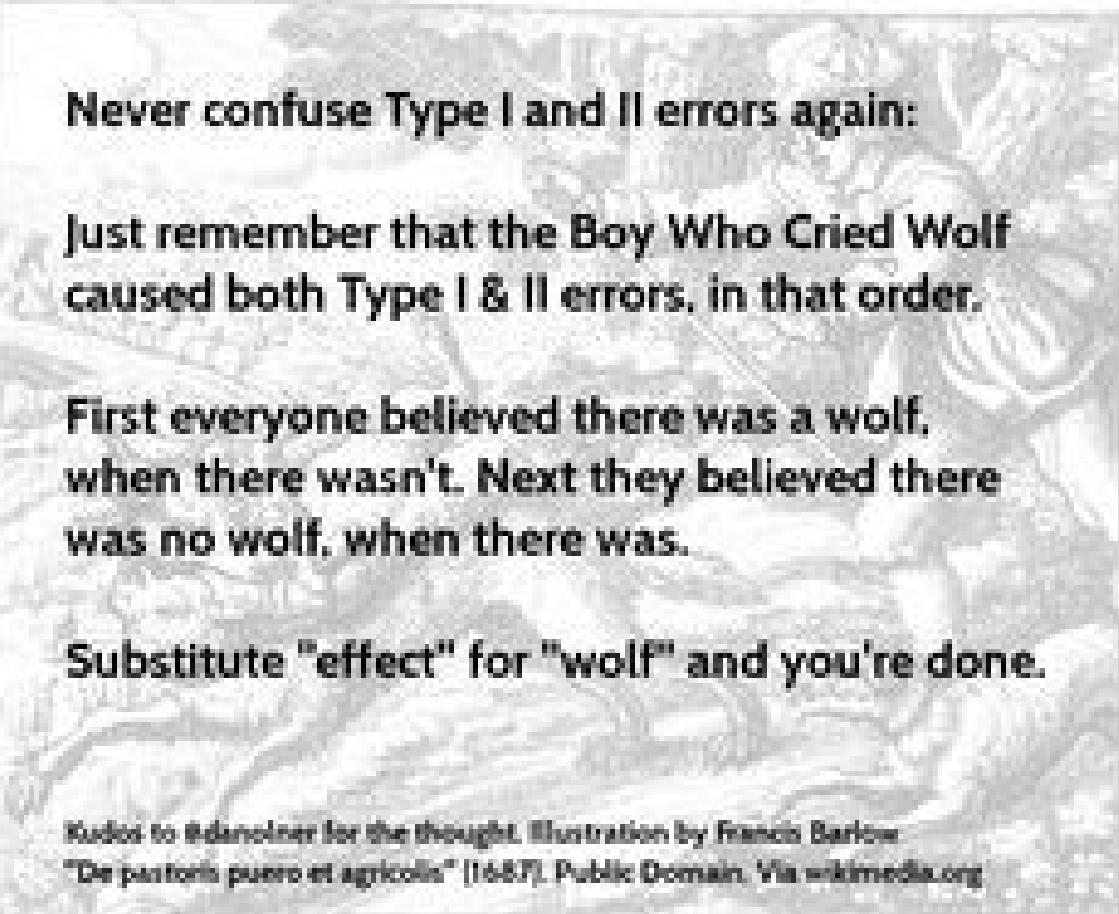
A Type I Error is generally considered more serious. This may influence the choice of H_0 and H_1 .

Our significance level, α , is the probability of Type I error we are willing to accept (when the null hypothesis is true).

Note that this does not control Type II error at all. The probability of Type II error may be very large, especially for small n .

It might help to think about the possible consequences of the different errors, by considering the (usually nonstatistical) example of a jury trial.

		Truth	
		H_0 True (Defendant Innocent)	H_1 True (Defendant Guilty)
Decision	Reject H_0 (Convict)	Type I Error	Correct Decision
	Fail to Reject H_0 (Acquit)	Correct Decision	Type II Error



Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danielner for the thought. Illustration by Francis Barlow
"De partoris pueri et agricola" (1687). Public Domain. Via [wikimedia.org](https://commons.wikimedia.org)

Consider the Radon Abatement Example

b.) Because of health concerns, radon abatement is recommended whenever it is plausible that the mean radon concentration may be 4.0 pCi/L or more. State the appropriate H_0 and H_1 for determining if abatement is appropriate.

Type I error: Decide abatement is not necessary when in fact it is.

Risks: Dangerous exposure to radon which includes serious health consequences.

Type II error: Decide abatement is necessary when it is not.

Risks: Spend money to perform unnecessary intervention.

The risks associated with Type I error appear much more serious than for Type II. We can control the probability of making a Type I error by choosing a small α level.

If we had used the hypotheses

$$H_0: \mu \leq 4.0 \text{ and } H_1: \mu > 4.0$$

which would mean that we would recommend abatement only if strong evidence indicates that the radon level is too high, the Type I and Type II errors would be the reverse of what they were in the previous example. So the Type II error would have the more severe associated risks. We would not be able to control the probability of Type II error with a small α .

Power

Definition: The *power* of a significance test with significance level α is

$$\begin{aligned}\text{power} &= P(\text{reject } H_0 \mid H_0 \text{ false}) \\ &= 1 - P(\text{Type II Error} \mid H_0 \text{ false}).\end{aligned}$$

If there is more than one value of μ associated with H_1 , power will generally be computed for a specific value of μ .

Large power is good, and power will increase as the sample size increases.

There is no firm rule about power, but it is desirable to have a power of at least 0.8 or 0.9 for a difference which is big enough to be important.

Power should be computed prior to conducting an experiment whenever possible, to verify that the experiment will probably show results if the difference you desire or anticipate exists.

To compute power (in a z-test for μ), we must:

1. Decide on a significance level, α .
2. Compute the rejection region, which is the set of possible values of \bar{X} which would lead to rejecting H_0 .
3. Compute the probability of finding \bar{X} in the rejection region, given a specific value of μ (a value in H_1).

Example: A tire company claims that the lifetimes of its tires average 50000 miles (μ_0). The standard deviation of tire lifetimes is known to be 5000 miles (σ). You sample 100 tires (n) and will test the hypothesis that the mean tire lifetime is at least 50000 miles against the alternative that it is less. Assume, in fact, that the true mean lifetime is 49500 miles (μ_1).

- a. State the null and alternative hypotheses. Which hypothesis is true?

b. It is decided to reject H_0 if the sample mean is less than 49400. Find the level and power of this test.

c. If the test is made at the 5% level ($\alpha = .05$), what is the power?

d. At what level should the test be conducted so that the power is .80?

Now we want power = .80. So

The corresponding value of α would be

e. You are given the opportunity to sample more tires. How many tires should be sampled in total so that the power is .80 if the test is made at the 5% level?

Statistical software such as Minitab can compute the power for a specified test, or the sample size necessary to achieve a given power. It can do the computations for power and sample size for t-tests as well, which are quite formidable.

Power and Sample Size

1-Sample Z Test (Part b of previous example)

Testing mean = null (versus < null)

Calculating power for mean = null + difference

Alpha = 0.1151 Assumed standard deviation = 5000

Sample

Difference	Size	Power
-500	100	0.420801

Power and Sample Size

1-Sample Z Test (Part e previous example)

Testing mean = null (versus < null)

Calculating power for mean = null + difference

Alpha = 0.05 Assumed standard deviation = 5000

Difference	Sample	Target	Actual Power
	Size	Power	
-500	619	0.8	0.800419



Two-Sample Inference

In many situations, we may not care about the specific value of the mean of a population, so much as comparing the means of two separate, but related, populations.

We use *two-sample inference* to investigate these sorts of questions.

Example: Sexual discrimination?

μ_M = population mean of male salaries

μ_F = population mean of female salaries

Test $H_0: \mu_M \leq \mu_F$ vs. $H_1: \mu_M > \mu_F$

Example: New fertilizer for corn

μ_1 = average yield for the new treatment

μ_0 = average yield for a common current treatment

Test $H_0: \mu_1 \leq \mu_0$ vs. $H_1: \mu_1 > \mu_0$

Confidence interval on $(\mu_1 - \mu_0)$

Suppose we have two populations or processes.

Population 1 (X) has mean μ_x and standard deviation σ_x .

Likewise, population 2 (Y) has mean μ_y and standard deviation σ_y .

We will compare the individual means by looking at the difference, $\mu_x - \mu_y$.

We usually do this by collecting independent samples from each population (of sizes n_X and n_Y , which may or may not be the same) and computing the sample means (\bar{X} and \bar{Y}) and standard deviations (S_X and S_Y).

We will **estimate** the difference of population means, $\mu_X - \mu_Y$, by the difference of the sample means, $\bar{X} - \bar{Y}$.

To do inference, we need to know about the sampling distribution of $\bar{X} - \bar{Y}$.

Recall the following results from our discussion of linear combinations of random variables:

- 1) For any X and Y , $\mu_{X-Y} = \mu_X - \mu_Y$. (The mean of the difference is the difference of the means.)
- 2) For any independent X and Y , $\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$. (The variance of the difference is the *sum* of the variances.)
- 3) If X and Y are (approximately) normal, so is the difference.

Those results, together with what we already know about the sampling distributions of \bar{X} and \bar{Y} give us:

1) $\bar{X} - \bar{Y}$ is an unbiased estimator of $\mu_X - \mu_Y$, that is

$$\mu_{\bar{X}-\bar{Y}} = \mu_X - \mu_Y$$

2) The standard error of $\bar{X} - \bar{Y}$ is $\sqrt{\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)}$.

- 3) If both populations are normal, so is the sampling distribution of $\bar{X} - \bar{Y}$.
- 4) If both sample sizes are large, the Central Limit Theorem tells us that the sampling distribution of $\bar{X} - \bar{Y}$ will be approximately normal no matter what shapes the population distributions have.
- 5) When standardizing using sample standard deviations from small samples from normal populations, we should continue to use a t -distribution.

Two-Sample z-Tests

If we have two independent samples, we construct a *two-sample test* in much the same way as the one-sample version.

If *both samples are large*, we may use the *normal distribution* as we do in the single-sample case.

Remember our five steps of hypothesis testing.

1) We have two populations whose means we wish to compare, and a sample from each population.

- Population 1 (X) has mean μ_X and standard deviation σ_X .
- Population 2 (Y) has mean μ_Y and standard deviation σ_Y .
- The sample means \bar{X} and \bar{Y} will likely be different, even if the population means μ_X and μ_Y are the same.
- Are \bar{X} and \bar{Y} different enough to provide strong evidence that μ_X and μ_Y are different as well?

Example: A psychological test (the “Chapin Social Insight Test”) is given to a large number of college students, with a desire to see if there is a difference in how men and women score.

$$\text{Men : } n_X = 133, \quad \bar{X} = 25.34, \quad s_X = 5.05$$

$$\text{Women : } n_Y = 162, \quad \bar{Y} = 24.94, \quad s_Y = 5.44$$

Does a test suggest that the populations of college men and women have different means on this test?

2) Identify H_0 and H_1 . We generally state our hypotheses using statements about a difference in the population means.

- Two-sided:

$$H_0: \mu_x - \mu_y = \Delta \quad \text{vs.} \quad H_1: \mu_x - \mu_y \neq \Delta.$$

- One-sided:

$$H_0: \mu_x - \mu_y \leq \Delta \quad \text{vs.} \quad H_1: \mu_x - \mu_y > \Delta.$$

or:

$$H_0: \mu_x - \mu_y \geq \Delta \quad \text{vs.} \quad H_1: \mu_x - \mu_y < \Delta.$$

- Usually, $\Delta = 0$, so that $\mu_x = \mu_y$ is part of H_0 .

Example: (Students, continued):

We are interested in showing any difference between the men's population mean (μ_x) and the women's population mean (μ_y), so use a two-sided alternative.

$$H_0: \mu_x - \mu_y = 0 \text{ vs. } H_1: \mu_x - \mu_y \neq 0.$$

3) Compute the test statistic.

- A z (or t) statistic always has the form:

$$z = \frac{\text{point estimate} - \text{null value}}{\text{standard error}}.$$

- In this case, our parameter is $\mu_X - \mu_Y$.
- Our point estimator is $\bar{X} - \bar{Y}$.
- Our null value is Δ .
- The standard error of $\bar{X} - \bar{Y}$ is $\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}$

Our z statistic is therefore

$$z = \frac{(\bar{X} - \bar{Y}) - \Delta}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}.$$

Example: Students:

Men : $n_X = 133$, $\bar{X} = 25.34$, $s_X = 5.05$

Women : $n_Y = 162$, $\bar{Y} = 24.94$, $s_Y = 5.44$

4) Find the P-value.

- This is still a z-test. Under H_0 , $Z \sim N(0,1)$. Use probabilities on $Z \sim N(0,1)$, depending on H_1 :

H_1	P-value
$\mu_X - \mu_Y \neq \Delta$	$2P(Z \geq z) = 2 P(Z \leq - z)$
$\mu_X - \mu_Y > \Delta$	$P(Z \geq z) = P(Z \leq -z)$
$\mu_X - \mu_Y < \Delta$	$P(Z \leq z)$

Example: $z = 0.654$.

5) Reject H_0 for small P .

- The interpretation of P is exactly the same as for any other hypothesis test.
- If $P \leq \alpha$, the evidence is pretty strong against H_0 , and we say we *reject* H_0 (at the α level). We have strong evidence of H_1 .
- If $P > \alpha$, our test statistic is pretty reasonable under H_0 , so we *fail to reject* H_0 . H_0 is plausible.

Example: What does our P say about the mean test scores for the populations of student men and women?

Example: It is claimed that (on average) tensile strength should be at least 8 N/mm^2 greater for 12mm-diameter steel rods than for 10mm-diameter rods. Samples of size 50 give:

$$12\text{mm}: n_X = 50, \quad \bar{x} = 562, \quad s_X = 18$$

$$10\text{mm}: n_Y = 50, \quad \bar{y} = 545, \quad s_Y = 24$$

Can we confirm the claim?

Two-Sample Confidence Intervals

Instead of, or in addition to, a two-sample test, we may desire a confidence interval for the difference of the population means, $\mu_x - \mu_y$.

The same results that allow us to conduct a test also allow computation of this interval.

A $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ will be

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}.$$

Interpretation of the confidence interval is exactly the same as in the one-sample case.

Example: Students:

$$\text{Men : } n_X = 133, \quad \bar{X} = 25.34, \quad s_X = 5.05$$

$$\text{Women : } n_Y = 162, \quad \bar{Y} = 24.94, \quad s_Y = 5.44$$

Find a 95% confidence interval for the difference in population means between men and women.

Example: Steel rods:

$$12\text{mm} : n_X = 50, \quad \bar{X} = 562, \quad s_X = 18$$

$$10\text{mm} : n_Y = 50, \quad \bar{Y} = 545, \quad s_Y = 24$$

Find a 99% lower bound on the difference in population means.

Two-Proportion Inference

Just as we can use hypothesis tests and confidence intervals to compare means of two related populations, we may also use them to compare two related binomial probabilities or population proportions.

Suppose $X \sim Bin(n_X, p_X)$, and $Y \sim Bin(n_Y, p_Y)$.

We estimate the probabilities with

$$\hat{p}_X = \frac{X}{n_X}, \quad \hat{p}_Y = \frac{Y}{n_Y}.$$

If we want a confidence interval on the difference between two proportions, we can use the fact that for independent X and Y with large n_X and n_Y ,

$$\hat{p}_X - \hat{p}_Y \stackrel{\sim}{\sim} N\left(p_X - p_Y, \frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}\right).$$

Recall that for a modern confidence interval on a single proportion, we used an alternative estimator of p , adding two successes and two failures to our counts:

$$\tilde{p} = \frac{X + 2}{n + 4}.$$

With two samples, we distribute the extras between the two estimates:

$$\tilde{n}_X = n_X + 2, \tilde{p}_X = \frac{X + 1}{\tilde{n}_X}, \tilde{n}_Y = n_Y + 2, \tilde{p}_Y = \frac{Y + 1}{\tilde{n}_Y}.$$

Our $100(1 - \alpha)\%$ confidence interval for $p_X - p_Y$ is

$$\tilde{p}_X - \tilde{p}_Y \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_X(1 - \tilde{p}_X)}{\tilde{n}_X} + \frac{\tilde{p}_Y(1 - \tilde{p}_Y)}{\tilde{n}_Y}}.$$

Example: Are rural households more likely to use a natural Christmas tree than urban ones?

Rural: $n_X = 160$, $X = 64$

Urban: $n_Y = 261$, $Y = 89$

Find a 95% confidence interval for $p_X - p_Y$.

To test $H_0: p_X = p_Y (= p)$ vs. $H_1: p_X \neq p_Y$ we must estimate the common null proportion p with the *pooled proportion*:

$$\hat{p} = \frac{X+Y}{n_X+n_Y}.$$

Then our z statistic may be found as

$$z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}.$$

Compute P-values as for any other z test.

Note: $H_0: p_X - p_Y = 0$ vs $H_1: p_X - p_Y \neq 0$ are equivalent to the null and alternative hypotheses given above.

Example: Natural Christmas trees

Rural: $n_X = 160, X = 64$

Urban: $n_Y = 261, Y = 89$

Can we conclude that rural households are more likely to use a natural Christmas tree than urban ones?

Example: Are male college students (Y) more likely to be “frequent binge drinkers” than female students (X)?

Of 9916 college women surveyed, 1684 were classified as frequent binge drinkers.

Of 7180 college men surveyed, 1630 were considered frequent binge drinkers.

Test and CI for Two Proportions

Sample	X	N	Sample p
1	1684	9916	0.169827
2	1630	7180	0.227019

Difference = $p_1 - p_2$

Estimate for difference: -0.0571930

95% upper bound for difference: -0.0469659

Test for difference = 0 (vs < 0): Z = -9.34 P-Value = 0.000



$$H_1: p_X - p_Y < 0$$

$$H_0: p_X - p_Y \geq 0$$

Since the p-value is so small, we reject the null hypothesis and conclude that male college students are more likely to binge drink than female college students.

Two-Sample t -Tests and Intervals

Just as in the one-sample case, if at least one of the sample sizes is small, we run into the same dangers for estimating the standard error from the sample as we do in the single-sample case.

We again use a t -table to compensate.

A two-sample t -test is conducted exactly as a two-sample z -test, but uses t probabilities for P-values.

We find P-values from the t-table in the text or Stat Trek or a computer package such as Minitab, just as with the one-sample t-test.

The degrees of freedom to be used can be calculated as

$$v = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \right)^2}{\frac{(s_X^2 / n_X)^2}{(n_X - 1)} + \frac{(s_Y^2 / n_Y)^2}{(n_Y - 1)}} \quad (\text{rounded down}).$$

Confidence intervals use critical values $t_{v,\alpha/2}$.

Example: Researchers are investigating using kudzu as an alternative to wood pulp in paper production. They wish to determine if adding the chemical anthraquinone increases the pulp yield.

Treatment: 25 experiments with

$$\bar{x} = 44.17, \quad s_X = 3.994$$

Control: 20 experiments without

$$\bar{y} = 38.55, \quad s_Y = 3.627$$

Is the anthraquinone increasing the population mean yield?
Conduct a two-sample *t*-test.

Construct a 95% lower confidence bound for mean improvement.

- Example Minitab output:
Two-sample tests, intervals

```
Results for: kudzu.txt
Two-Sample T-Test and CI: With, Without

Two-sample T for With vs Without
      N    Mean   StDev   SE Mean
With     25  44.18    3.99     0.80
Without  20  38.56    3.63     0.81

Difference = mu (With) - mu (Without)
Estimate for difference:  5.62500
95% lower bound for difference:  3.71000
T-Test of difference = 0 (vs >): T-Value = 4.94  P-Value = 0.000  DF = 42
```

Inference with Paired Data

Sometimes, we can improve an estimate of population differences by arranging to collect the data in a *paired* fashion.

Each observation from population 1 (X) should be paired with an observation from population 2 (Y).

This will be effective if the pairing is such that the pairs tend to be correlated.

Example: Compare two drugs (old and new) for effect on heart rate reduction.

For samples of size 20, we could get 40 volunteers and divide them at random into two groups to get independent samples.

If drug response varies substantially from subject to subject, it may be better to give both drugs to each subject (on different occasions, in random order). This reduces the effect of subject variability, and is probably cheaper and easier as well!

Other examples:

Test tire wear for two brands by putting Brand A on the left front wheel and Brand B on the right front wheel (or vice-versa, at random) on the *same* cars.

Compare airplane deicing procedures by using one method on the left wing and the other on the right wing (at random) of the *same* planes.

Dealing with paired data (X_i, Y_i) is actually simpler than dealing with two independent samples.

For each observation, compute the difference $D_i = X_i - Y_i$, and then conduct a one-sample z- or t-test or construct a one-sample z- or t-confidence interval on the differences D_i , depending on the number of pairs.

This will give us a test or interval for $\mu_D = \mu_X - \mu_Y$.

Example: (Heart rate data, continued):

Let X_i be the percent rate reduction from the standard drug for subject i , and let Y_i be the same for the new drug. Let $D_i = X_i - Y_i$.

Data:

Patient	X_i	Y_i	D_i
1	28.5	34.8	-6.3
2	26.6	37.3	-10.7
:	:	:	:
40	40.1	40.8	-0.7

Example (Heart rate, continued):

$$n = 40, \quad \bar{d} = -2.655, \quad s_D = 3.730$$

Is there enough evidence to conclude that the new drug is better than the old one at reducing heart rate on average?

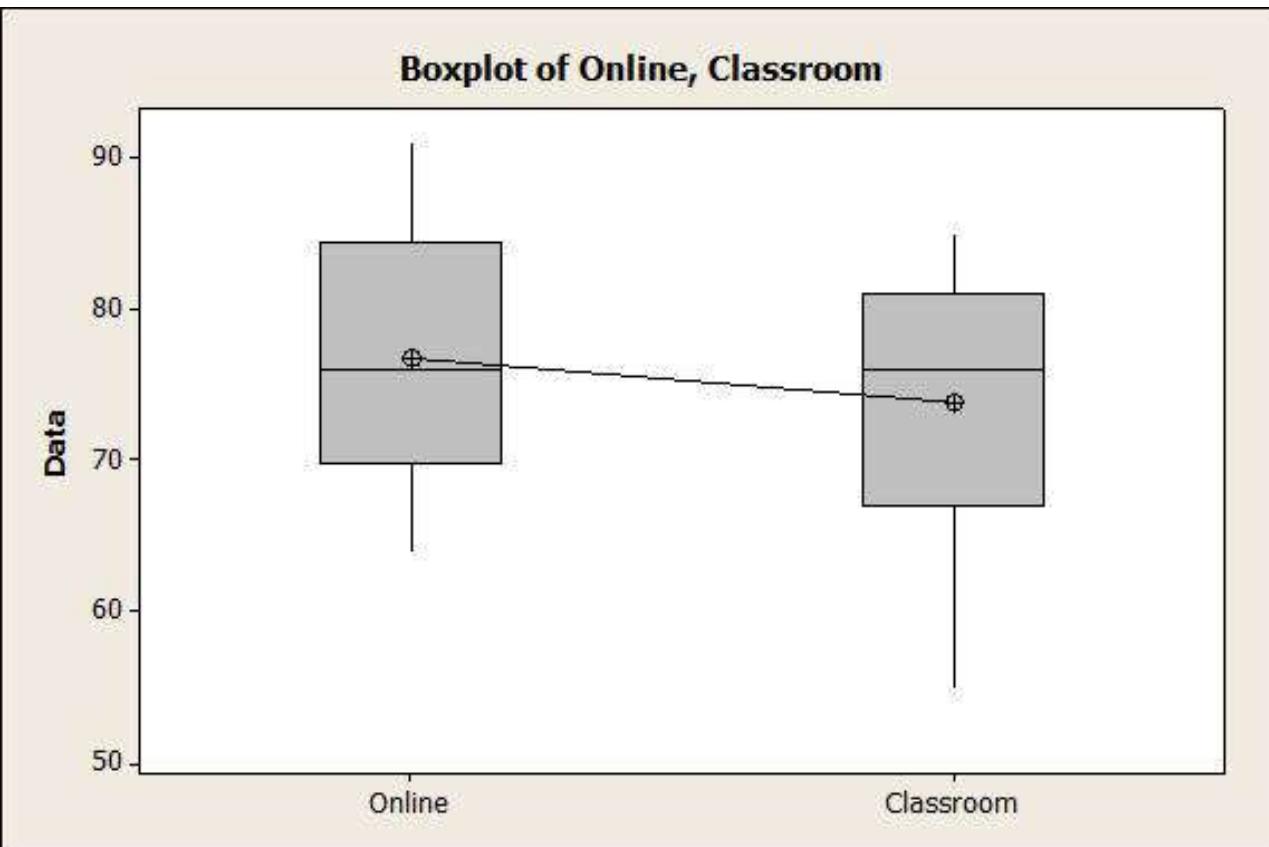
Construct a 99% upper confidence bound for $\mu_D = \mu_1 - \mu_2$.

Example: In a comparison of the effectiveness of distance learning with traditional classroom instruction, 12 students took a business administration course online, while 14 students took it in a classroom. The final exam scores were as follows:

Online: 64, 66, 74, 69, 75, 72, 77, 83, 77, 91, 85, 88

Classroom: 80, 77, 74, 64, 71, 80, 68, 85, 83, 59, 55, 75, 81, 81

Can you conclude that the mean score differs between the two types of course?



- **Two-Sample T-Test and CI: Online, Classroom**

- Two-sample T for Online vs Classroom

	N	Mean	StDev	SE Mean
• Online	12	76.75	8.57	2.5
• Classroom	14	73.79	9.25	2.5

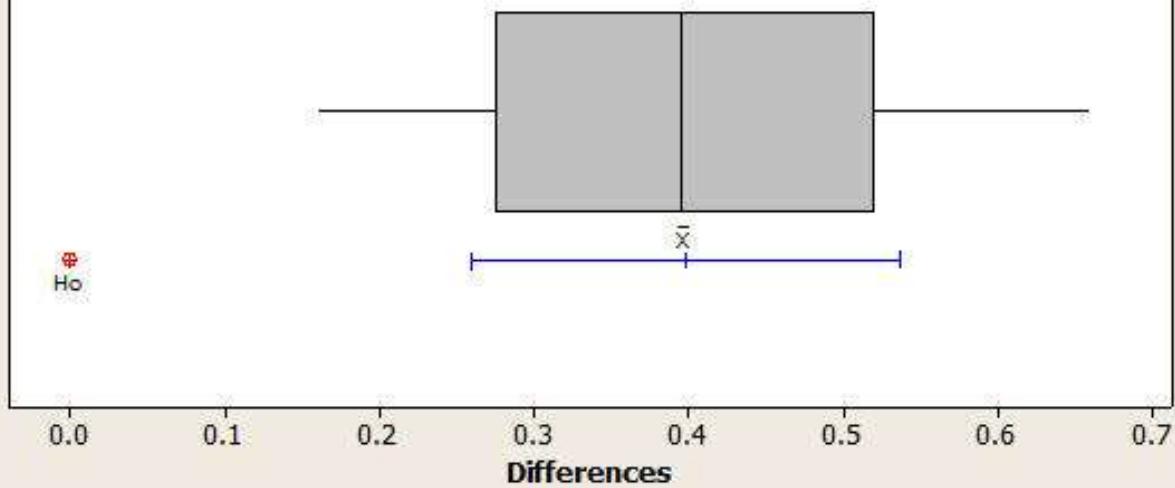
- Difference = mu (Online) - mu (Classroom)
- Estimate for difference: 2.96429
- 95% CI for difference: (-4.27157, 10.20014)
- T-Test of difference = 0 (vs not =): T-Value = 0.85 P-Value = 0.405 DF = 23

Example: A sample of 10 diesel trucks were run both hot and cold to estimate the difference in fuel economy. The results in mpg are presented below:

Truck	Hot	Cold
1	4.56	4.26
2	4.46	4.08
3	6.49	5.83
4	5.37	4.96
5	6.25	5.87
6	5.90	5.32
7	4.12	3.92
8	3.85	3.69
9	4.15	3.74
10	4.69	4.19

Find a 98% confidence interval for the difference in mean fuel mileage between hot and cold engines and determine whether we can conclude that fuel economy differs between diesel engines run with hot and cold engines.

Boxplot of Differences
(with H_0 and 98% t-confidence interval for the mean)



- **Paired T-Test and CI: Hot, Cold**

- Paired T for Hot - Cold

	N	Mean	StDev	SE Mean
• Hot	10	4.98400	0.95030	0.30051
• Cold	10	4.58600	0.84037	0.26575
• Difference	10	0.398000	0.155835	0.049279

- 98% CI for mean difference: (0.258962, 0.537038)
- T-Test of mean difference = 0 (vs not = 0): T-Value = 8.08 P-Value = 0.000

Simple Linear Regression

Statistics is most powerful when looking at relationships *between* variables.

In the simplest case, this involves looking at pairs of measurements made on the same subjects, (x, y) .

Recall, such data is called *bivariate* (two variables).

Examples:

- Heights and weights of a group of people.
- ACT score and Freshman GPA for college students.
- January and April average temperatures for many years at a specified location.
- January and February inflows of the Nile river at a location.

We usually picture our variables in a cause-and-effect relationship, which may or may not be the case.

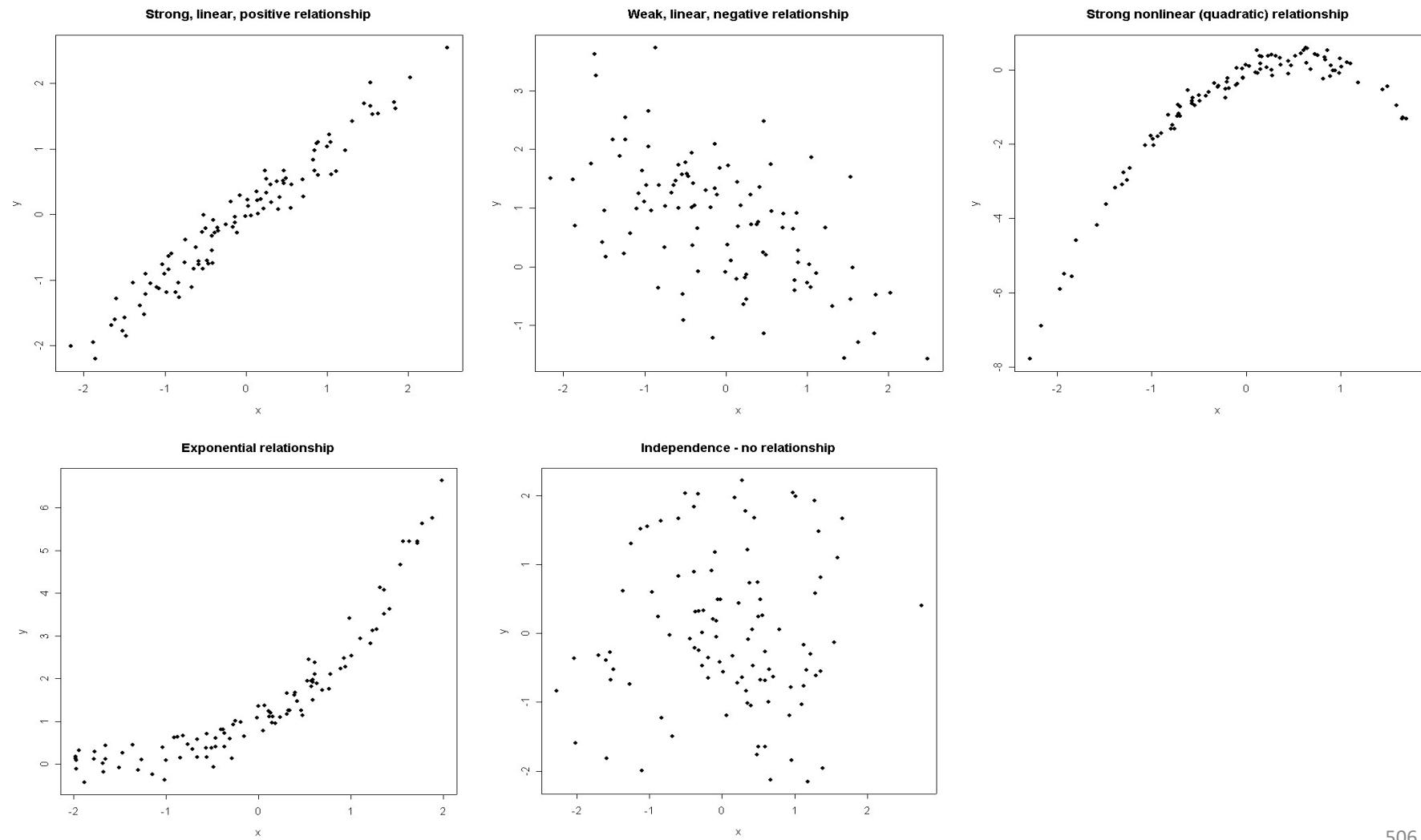
The *explanatory* (independent, predictor) variable, X, is assumed to play some role in determining the value of the *response* (dependent) variable, Y.

$$x \rightarrow y$$

Scatterplots

Definition: A *scatterplot* is the most common graph for displaying bivariate data. It consists of plotting each point at (x_i, y_i) , on a standard x - y graph.

The pattern formed by the points describes the relationship between the variables.



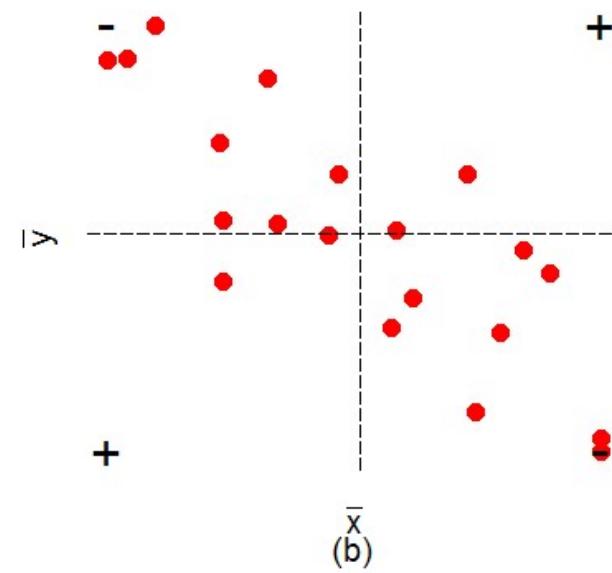
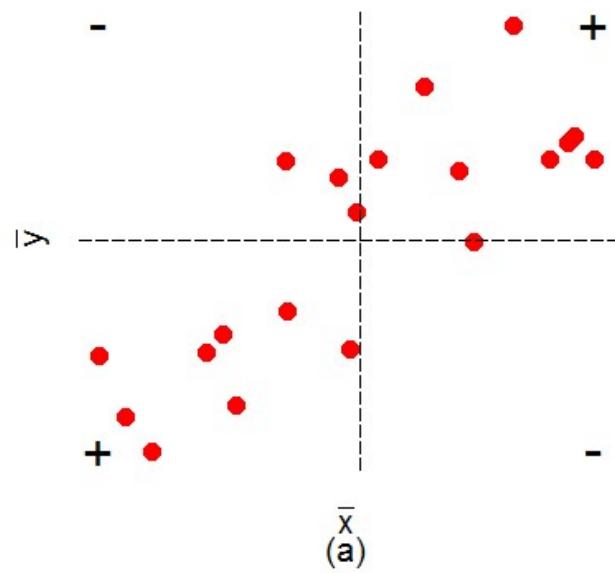
Correlation

Suppose we have a sample of (x, y) pairs and compute the sample means, \bar{x} and \bar{y} .

For each observation (x_i, y_i) , compute the product of the two deviations from the means.

Dividing the scatterplot at the means results in two quadrants where the product is positive, and two where it is negative.

Sign of $(x_i - \bar{x})(y_i - \bar{y})$.



For a scatterplot with a positive relationship, most of the products will have a positive sign, and the sum will be positive.

Likewise, if the picture shows a negative relationship, the sum of the products will be negative.

Unfortunately, the exact value of the sum depends on the units and spread (as measured by standard deviation) of the variables.

Dividing by measures of spread for x and y solves these issues.

Let $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 = (n-1)s_x^2$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = (n-1)s_y^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right).$$

Then $r = S_{xy} / \sqrt{S_{xx} S_{yy}}$ is a good, unitless measure of the linear relationship between x and y called the **correlation coefficient**.

Example: Nile flow data: $n=115$

$$\text{Jan}(x): \sum_{i=1}^n x_i = 462.9, \quad \sum_{i=1}^n x_i^2 = 1982.5264$$

$$\bar{x} = 4.0252, \quad s_x = 1.0228$$

$$\text{Feb}(y): \sum_{i=1}^n y_i = 333.04, \quad \sum_{i=1}^n y_i^2 = 1060.2076$$

$$\bar{y} = 2.8960, \quad s_y = 0.9163$$

$$\sum_{i=1}^n x_i y_i = 1440.2743$$

What is r ?

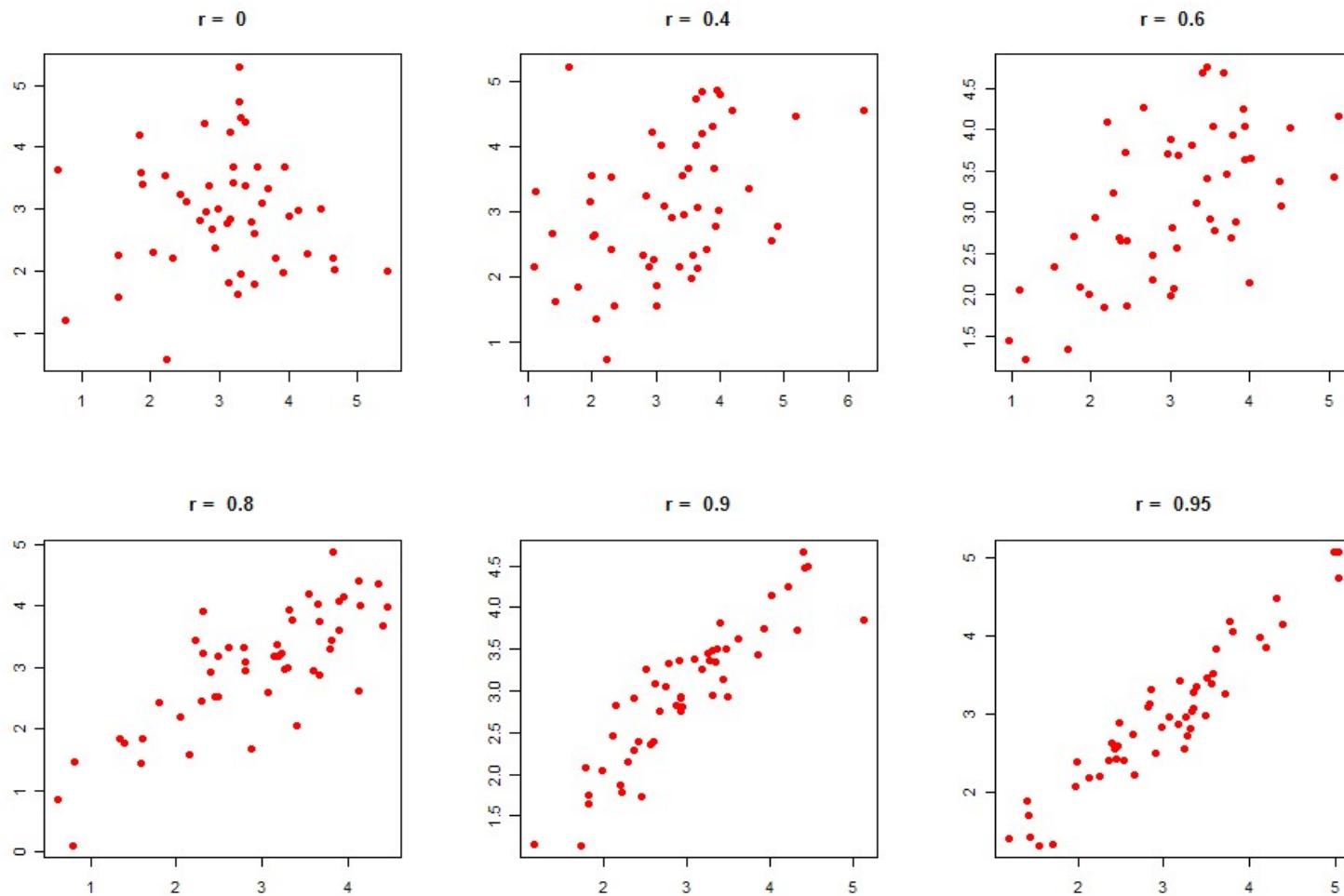
$$r=S_{xy}\,/\,\sqrt{S_{xx}S_{yy}}$$

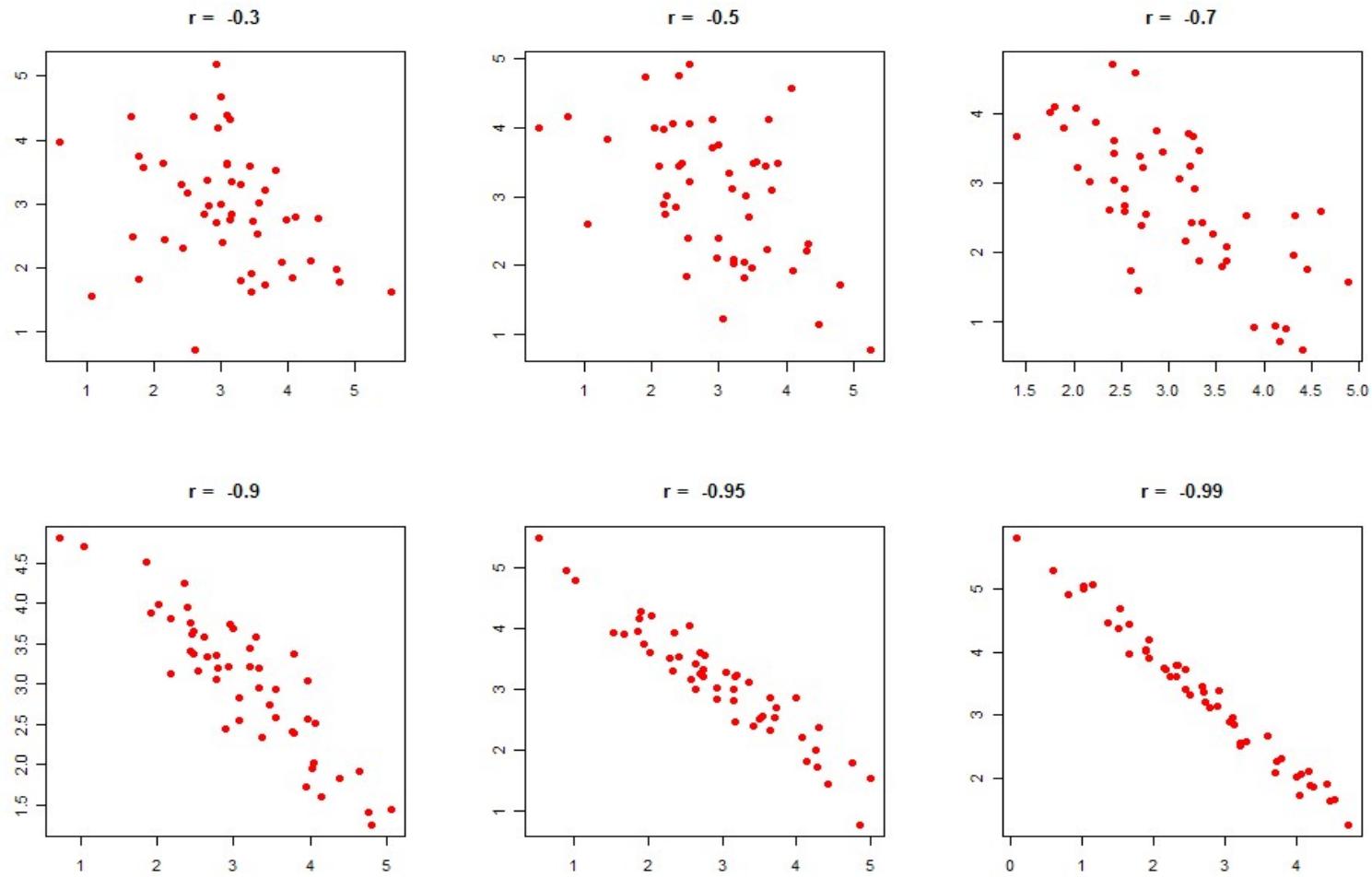
Properties of r

1. The value of r does not depend on the units of x or y . We will not change r if we multiply all x 's, all y 's, or both by a positive constant or if we add any constant to all x 's, all y 's, or both.
2. The value of r does not depend on which variable is labeled x .
3. Correlation is always between -1 and +1.
4. The sign of r shows whether the relationship between x and y is positive or negative.

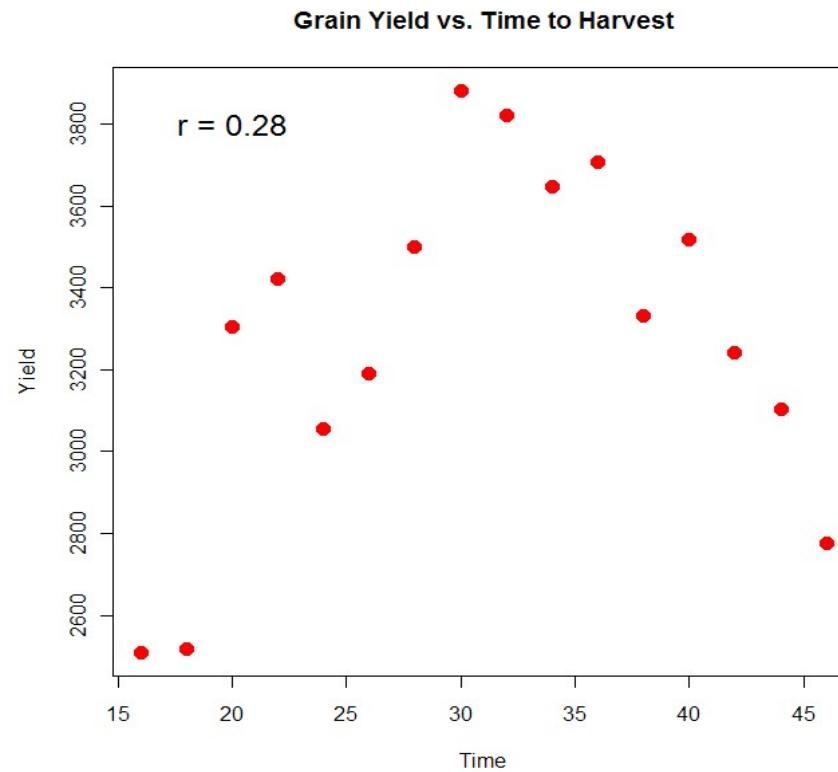
5.The absolute value of r measures the strength of the *linear* association between x and y .

- a.If $|r| < 0.5$, the association (if any) is weak.
- b.If $0.5 < |r| < 0.8$, the association is moderate.
- c.If $0.8 < |r| < 1.0$, the association is strong.
- d.If $|r| = 1.0$, the association is perfect. This occurs only when all (x, y) points fall in a perfect line.

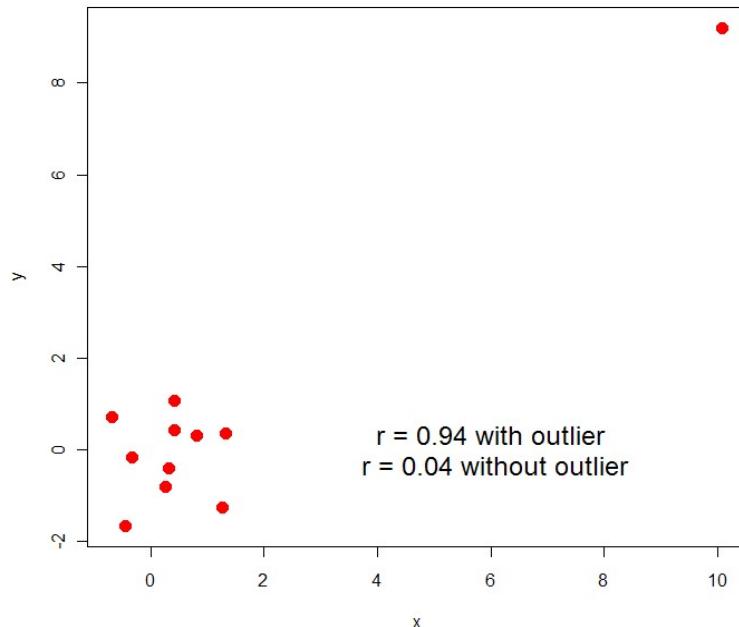
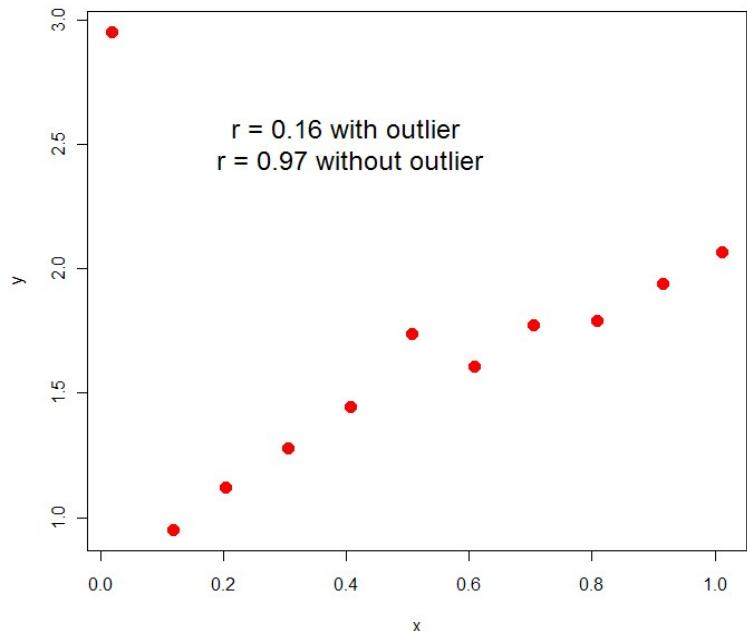




6. The correlation coefficient cannot measure the strength of a *nonlinear* (curved) relationship.



7. Outliers can lead to an inappropriate r value - in either direction!



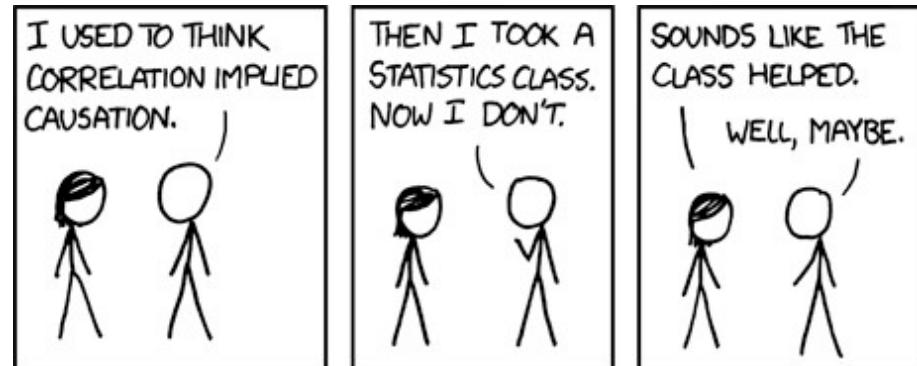
High correlation indicates strong association, *not* necessarily causality.

If $|r|$ is large, there are at least 3 possible explanations:

1) x determines y

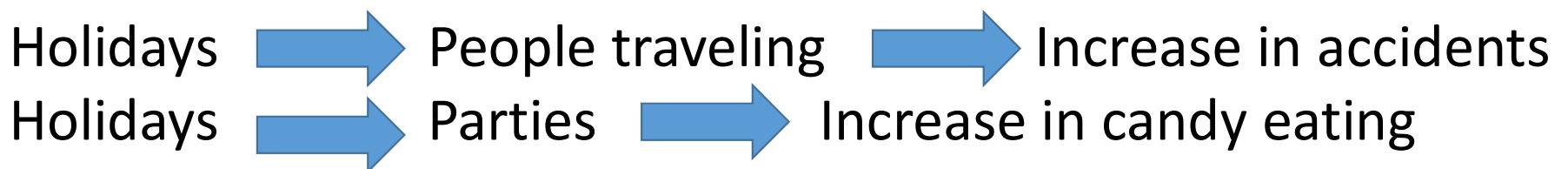
2) y determines x

3) Some third variable, z , (called a *confounding factor*) determines both x and y .

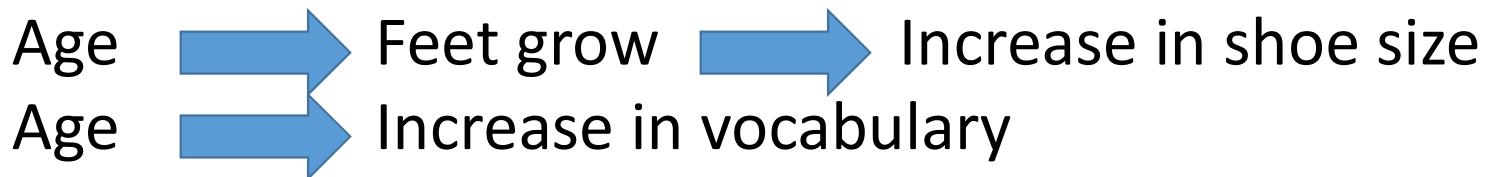


Example: Weekly surveys show that per capita chocolate consumption is strongly correlated with traffic fatalities.

- Should “driving under the influence of chocolate” be outlawed?
- Do people eat a lot of chocolate at funerals?
- Is there a third explanation that makes more sense?



Example: Children's shoe size is correlated with size of vocabulary.
What is the causal relationship?



One advantage of well-designed or randomized, controlled experiments is that potential confounding factors should be (roughly) balanced between levels of the independent variable we are investigating, so should be much less likely to produce a spurious correlation.

Linear Regression

Definition: *Regression* involves modeling and predicting the values of one response variable, based on the observed values of one or more explanatory variables.

We'll focus on the case of *simple linear regression*, where a straight line is fit to a scatterplot of x and y .

We want an equation for a line of the form

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The most common way to obtain the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ uses the *least squares* fit, minimizing

$$S = \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2.$$

This leads to the *least squares estimates*,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r \frac{s_y}{s_x},$$

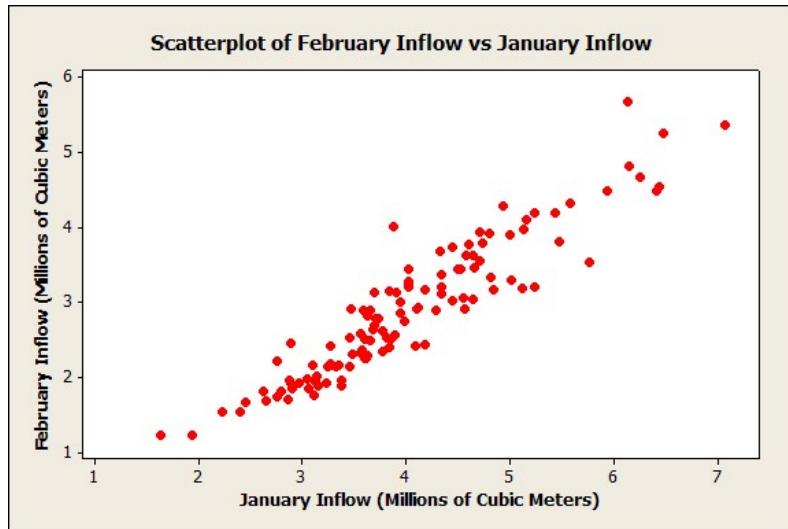
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Example: Nile flow data

Jan(x): $\bar{x} = 4.0252$, $s_x = 1.0228$

Feb(y): $\bar{y} = 2.8960$, $s_y = 0.9163$, $r = 0.933$

What is the least-squares line for this data, and what should we predict the flow for February to be if January's was 3?



Regression analysis

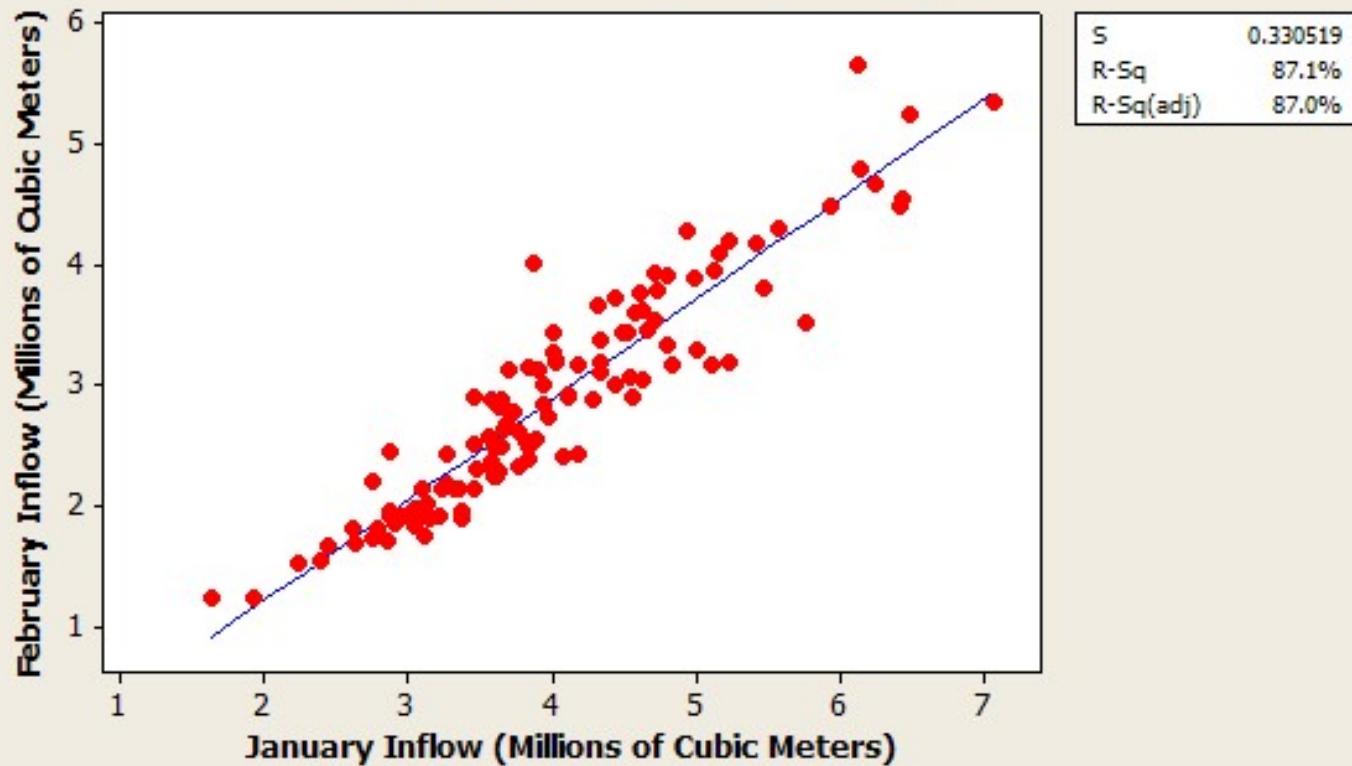
FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE INDEPENDENT OR PREDICTOR VARIABLE, AND y IS THE DEPENDENT OR RESPONSE VARIABLE. THE REGRESSION OR PREDICTION LINE HAS THE FORM

$$y = a + bx$$



Fitted Line Plot

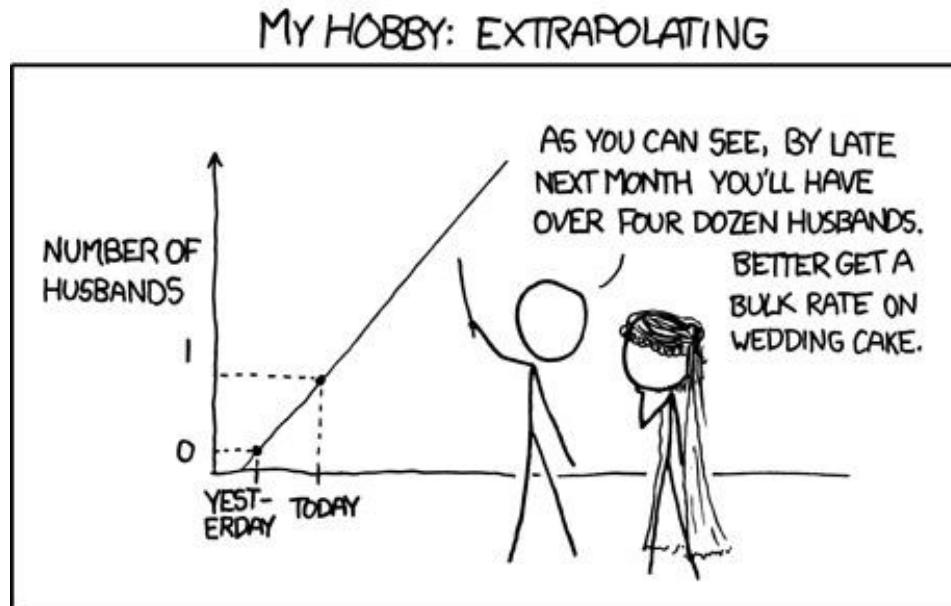
February Inflow = $-0.4698 + 0.8362 \text{ January Inflow}$



What would we predict for February from a January value of 10?

Is this likely to be a valid prediction? (Recall, January's mean is about 4, and its standard deviation is about 1.)

Extrapolation outside the range of the data is dangerous.



Residuals and Goodness-of-Fit

Definition: Given a data set (x_i, y_i) and an associated fitted regression model, the *fitted value* for observation i is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Definition: The *residual* for observation i is

$$e_i = y_i - \hat{y}_i.$$

The smaller the residuals, the better x and the regression line are at predicting y .

The *error sum of squares* (SSE) is

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

SSE is usually compared to the *total sum of squares*, SST:

$$\text{SST} = S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

and the *regression sum of squares*, SSR:

$$\text{SSR} = \hat{\beta}_1 S_{xy} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

To avoid having to calculate all the residuals, we may use the computing formula:

$$\text{SSE} = \text{SST} - \text{SSR}$$

The *coefficient of determination*, r^2 , measures the proportion of the total variation of y which is explained by x :

$$r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

The closer r^2 is to 1, the more successful the relationship is at explaining the variation in y .

As the notation suggests, the coefficient of determination is the square of the correlation coefficient.

Example: Nile flow data:

$$S_{xx} = 119.2533, \quad S_{yy} = 95.7238, \quad S_{xy} = 99.7159$$

$$y = -.4698 + .8362x$$

Find SST, SSR, SSE, and r^2 .

The coefficient of determination r^2 is found as “R-Sq” in Minitab output. The sums of squares may be found in the SS column of the *Analysis of Variance* table.

The regression equation is

$$\text{February Inflow} = -0.4698 + 0.8362 \text{ January Inflow}$$

$$S = 0.330519 \quad \text{R-Sq} = 87.1\% \quad \text{R-Sq(adj)} = 87.0\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	83.3794	83.3794	763.25	0.000
Error	113	12.3444	0.1092		
Total	114	95.7238			

Inference in Simple Linear Regression

The *simple linear regression* model fits a straight line to a set of paired data observations.

Formally:

- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- β_0 and β_1 are (unknown) constants
- $\varepsilon_1, \dots, \varepsilon_n$ are assumed to be independent draws from a $N(0, \sigma^2)$ distribution.
- $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$
- $E(Y_i) = \beta_0 + \beta_1 x_i$

The most common way to estimate β_0 and β_1 uses the *least squares* fit, minimizing

$$S = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

This leads to the *least squares estimates*,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{s_y}{s_x}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}.\end{aligned}$$

Recall: Given a data set (x_i, y_i) and an associated fitted regression model, the *fitted value* for observation i is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The *residual* for i is

$$e_i = y_i - \hat{y}_i.$$

The *error sum of squares* (SSE) is found as

$$\text{SSE} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

An *estimate of σ* , the standard deviation around the regression line is

$$s = \sqrt{\frac{\text{SSE}}{n-2}}.$$

Inference in simple linear regression usually focuses on $\hat{\beta}_1$, the estimator of the slope parameter β_1 , which measures how much y changes for a one-unit change in x .

Just like other sample statistics, $\hat{\beta}_1$ has a sampling distribution.

Under our model, this distribution is known and may be used to construct confidence intervals and hypothesis tests involving β_1 .

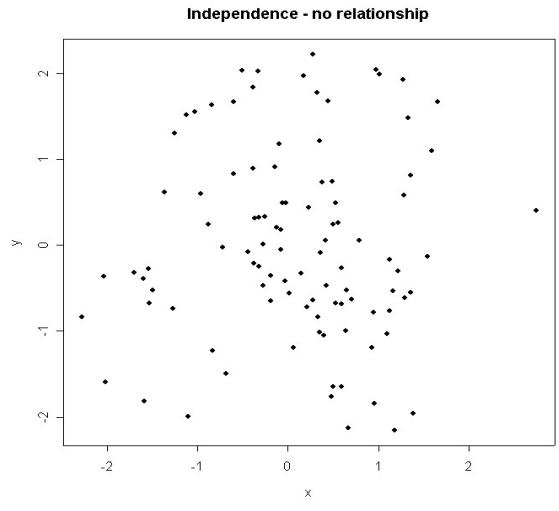
Under the formal regression model

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

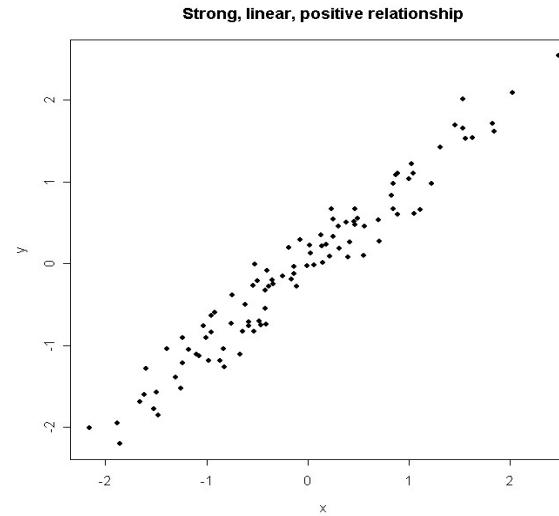

and

$$t = \frac{\hat{\beta}_1 - \beta_1}{s / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}.$$
S_{xx}

We may test $H_0: \beta_1 = \beta_{10}$ using test statistic t with β_{10} in place of β_1 and using a t table with $v=n-2$ degrees of freedom to find a P-value. Most commonly, $\beta_{10} = 0$.



$$\beta_1 = 0$$



$$\beta_1 \neq 0
(significantly so)$$

Example: (Nile data) Test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.

The regression equation is

$$\text{February} = -0.470 + 0.836 \text{ January}$$

Predictor	Coef	SE Coef	T	P
Constant	-0.4698	0.1257	-3.74	0.000
January	0.83617	0.03027	27.63	0.000

Can we conclude that the population slope β_1 is greater than 0.8?

We can use the distribution on t to construct a confidence interval for β_1 as

$$\hat{\beta}_1 \pm t_{v,\alpha/2} \times s_{\hat{\beta}_1}.$$

Example: Nile data (95% c.i. for β_1):

$(1-\alpha) \times 100\%$ CI for $E(Y|X = x) = \mu_{Y|X=x}$:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2, \alpha/2} s_{\hat{y}}$$

where

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

$(1-\alpha) \times 100\%$ Prediction Interval for $Y|X=x$:

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2, \alpha/2} s_{pred}$$

where

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

Nile River example:

Compute a 95% confidence interval for the mean inflow in February when the January inflow is 3 million cubic meters.

Compute a 95% prediction interval for the February inflow if the January inflow is 3 million cubic meters.

In a study of reaction times, the time to respond to a visual stimulus (X) and the time to respond to an auditory stimulus (Y) were recorded for 10 subjects. Times were measured in ms.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

- a.) Compute the least-squares line for predicting auditory response time based on visual response time.

Use Minitab.

Minitab Output

Regression Analysis: y versus x

The regression equation is

$$y = 22.4 + 0.864 x$$

Predictor	Coef	SE Coef	T	P
Constant	22.36	42.95	0.52	0.617
x	0.8638	0.2164	3.99	0.004

B.) Compute the error standard deviation s .

$$S = 15.1005 \quad R-Sq = 66.6\% \quad R-Sq(\text{adj}) = 62.4\%$$

C.) Compute a 95% CI for the slope.

Predictor	Coef	SE Coef	T	P
Constant	22.36	42.95	0.52	0.617
x	0.8638	0.2164	3.99	0.004

d.) Find a 95% CI for the mean auditory response time for subjects with a visual response time of 200 ms.

Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	195.12	4.81	(184.02, 206.22)	(158.57, 231.67)

Values of Predictors for New Observations

New

Obs	x
1	200

e.) Can you conclude that the mean auditory response time for subjects with a visual response time of 200 ms is greater than 190 ms? Perform a hypothesis test and report the P-value.

Predicted Values for New Observations

New
Obs Fit SE Fit 95% CI 95% PI
1 195.12 4.81 (184.02, 206.22) (158.57, 231.67)

Values of Predictors for New Observations

New
Obs x
1 200

f.) Find a 95% prediction interval for the auditory response time for a particular subject whose visual response time is 200 ms.

Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	195.12	4.81	(184.02, 206.22)	(158.57, 231.67)

Values of Predictors for New Observations

New

Obs	x
1	200

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3631.9	3631.9	15.93	0.004
Residual Error	8	1824.2	228.0		
Total	9	5456.1			

Predictor	Coef	SE Coef	T	P
Constant	22.36	42.95	0.52	0.617
x	0.8638	0.2164	3.99	0.004

$$F=t^2$$

Same p-values

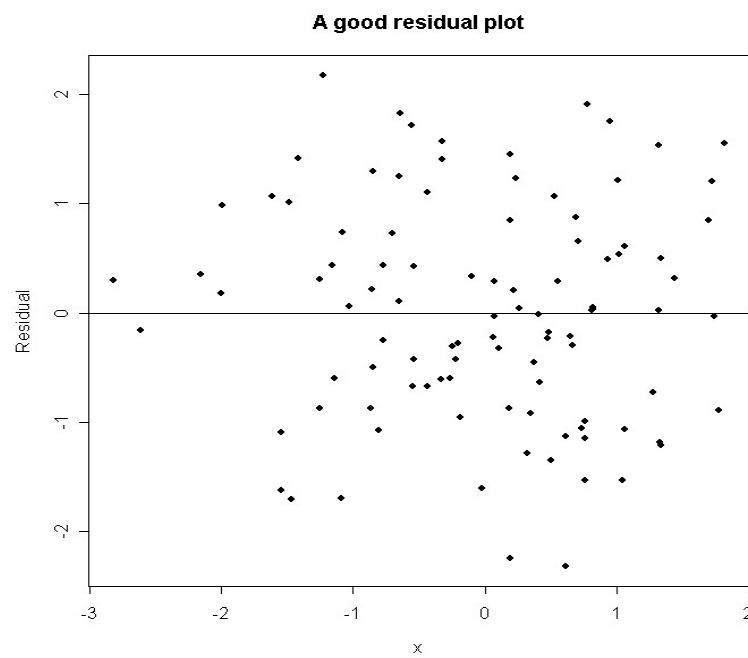
Note: These are equivalent tests for the overall significance of the linear model.

Residual Plots (looking for model assumption violations or lack of fit issues)

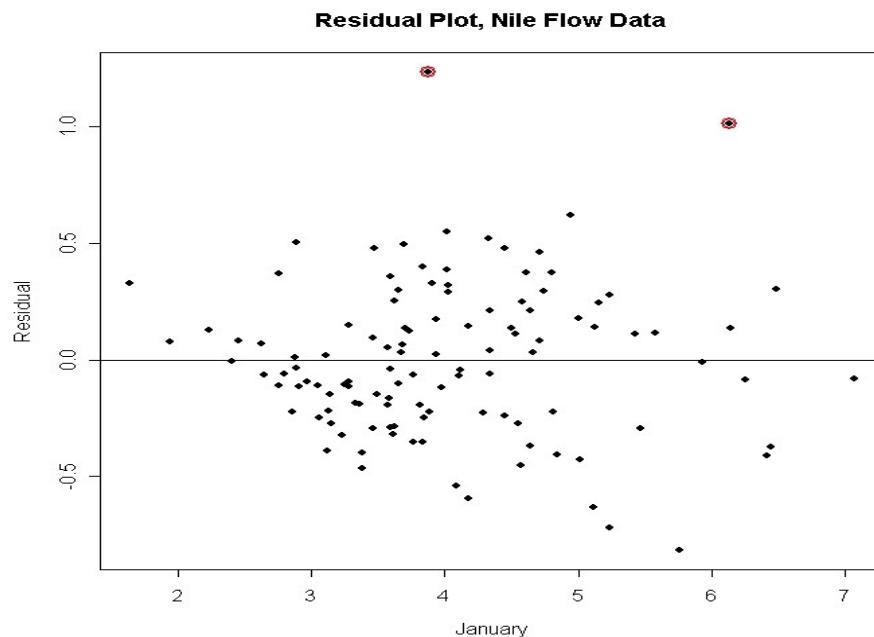
Examination of residuals is an important check on a regression analysis.

Generally, we look at a *residual plot* of x_i (or \hat{y}_i) versus e_i .

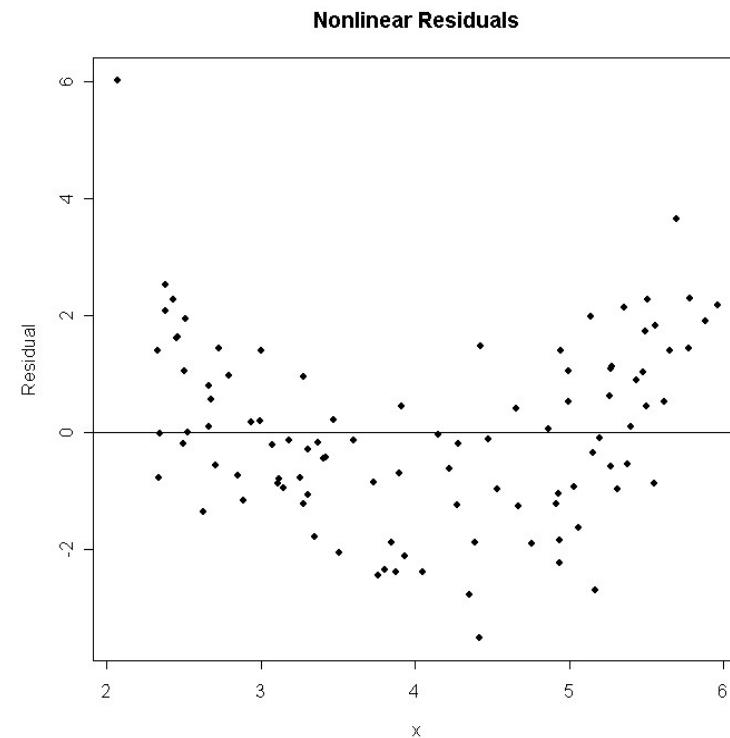
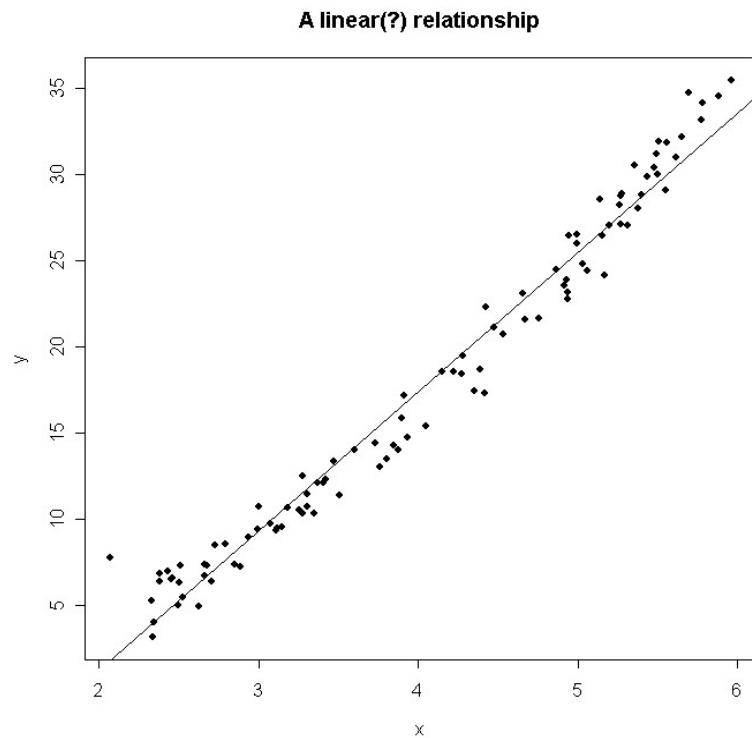
If the residual plot is flat, even, and appears random around $e = 0$, everything is probably fine.



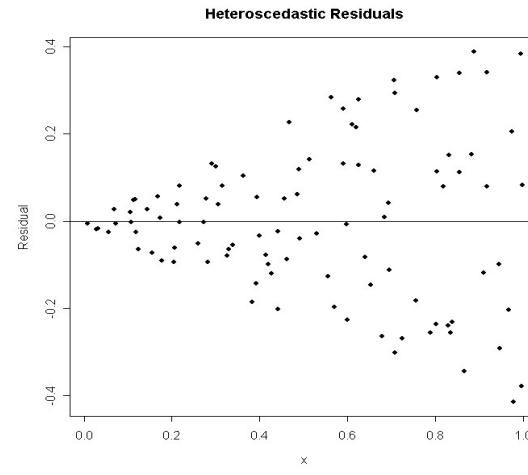
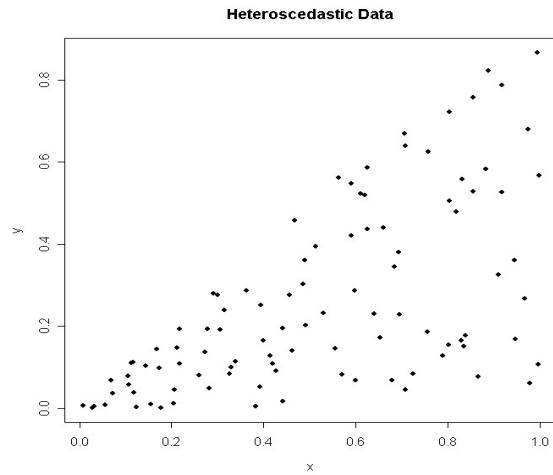
Residual outliers indicate points which are not well fit by the model. Check for explanations, and possibly remove those points.



Nonlinear patterns in the residual plot suggest a linear fit is inappropriate.



Funnel-shaped plots mean your data is *heteroscedastic* (“different scatter”), meaning the standard deviation of y is not constant – it depends on x . Fitted values may still be reasonable, but r^2 and s may not mean much.



Power Transformations

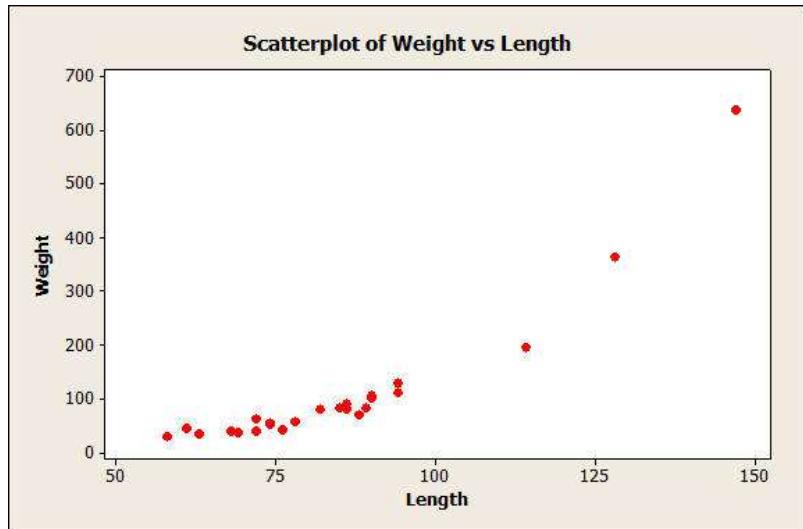
Skewness (including outliers), curvature, and heteroscedasticity can often all be improved by the use of nonlinear transformations on y , on x , or on both.

Such transformations include such options as logs, square roots, and reciprocals ($1/x$).

Apply the transformation to all observations on this variable.

Example: The Florida Game and Freshwater Fish Commission is interested in developing a model that will allow the accurate prediction of the weight of an alligator from more easily observed data on length. Length (inches) and weight (pounds) data for a sample of 25 alligators is used to fit a model for prediction.



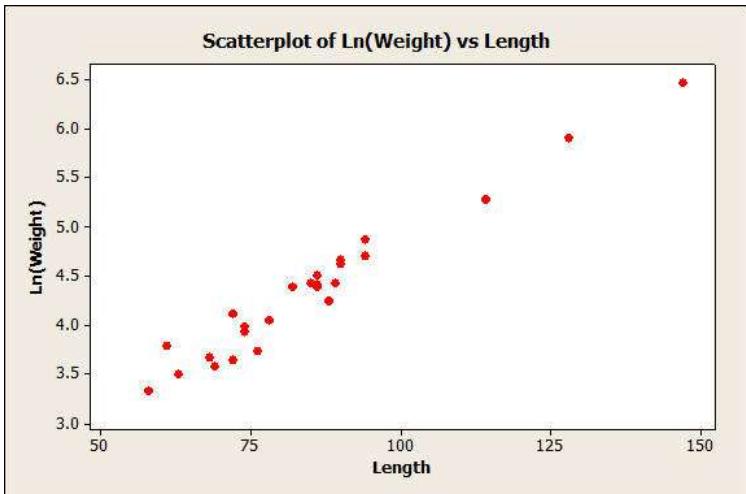


Minitab Output:

Type equation here. The regression equation is
Weight = - 393 + 5.90 Length

Predictor	Coef	SE Coef	T	P
Constant	-393.26	47.53	-8.27	0.000
Length	5.9024	0.5448	10.83	0.000

S = 54.0115 R-Sq = 83.6% R-Sq(adj) = 82.9%
(r=.914)

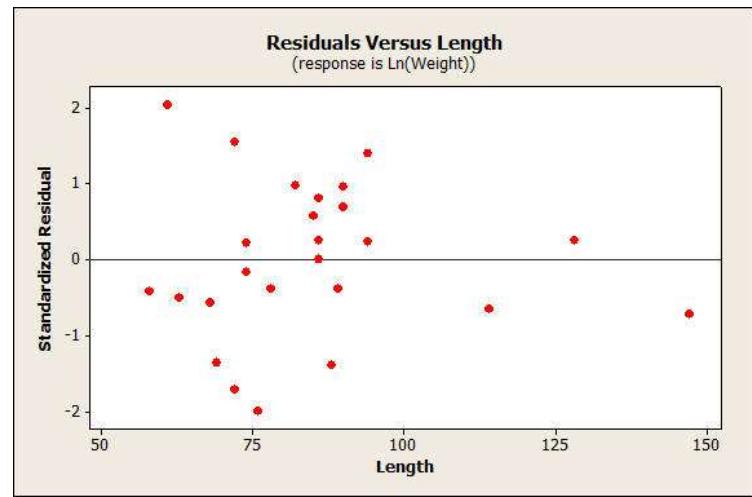
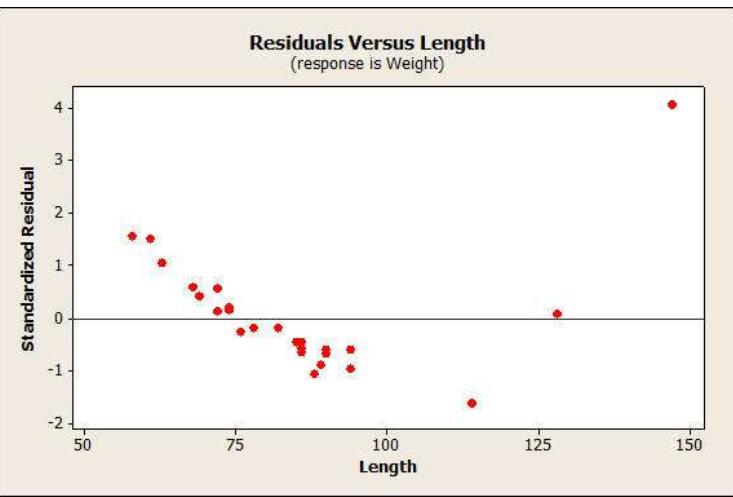


Minitab Output:

The regression equation is

$$\text{Ln(Weight)} = 1.34 + 0.0354 \text{ Length}$$

Predictor	Coef	SE Coef	T	P
Constant	1.3353	0.1314	10.16	0.000
Length	0.035416	0.001506	23.52	0.000
S	0.149299	R-Sq = 96.0%	R-Sq(adj) = 95.8%	
(r = .9798)				



Suppose we wanted to predict the weight of an alligator that is 80 inches long using the better-fitting transformed model.

One Way Analysis of Variance

Earlier we looked at comparing the means of samples from two populations.

Suppose we have samples from more than two populations, and we wish to test whether all of the populations have the same mean.

We use an extension of two-sample tests called **Analysis of Variance (ANOVA)**.

ANOVA generates P-values using another important distribution, the F distribution.

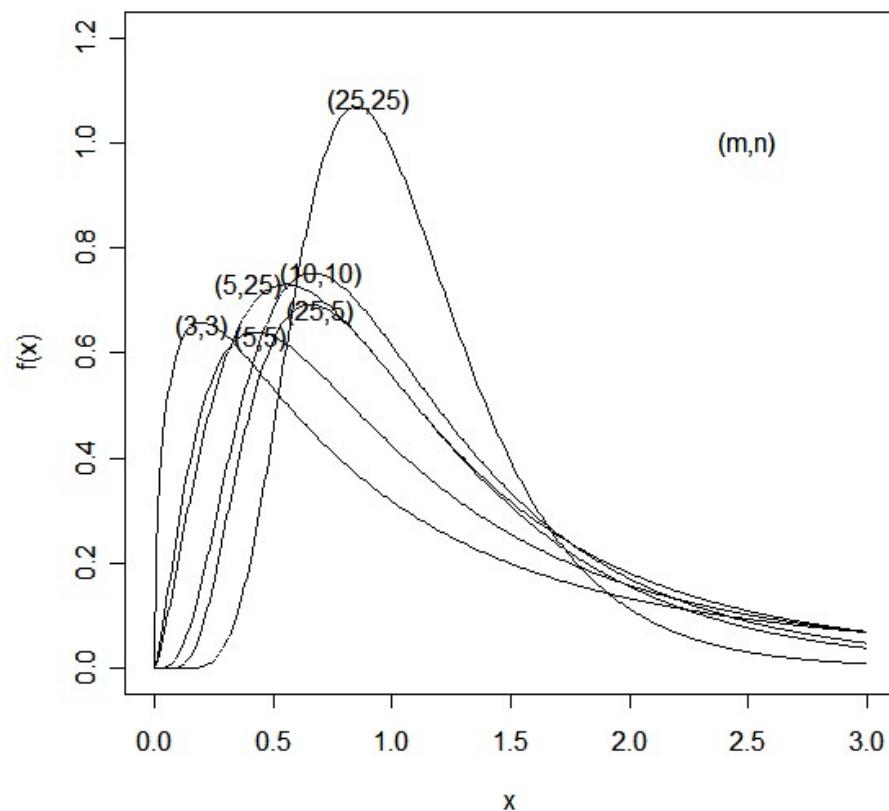
The F distribution

The **F distribution** is another distribution commonly used in hypothesis tests.

It is a skewed distribution with support on the positive real line.

It has two degrees of freedom parameters, v_1 and v_2 , and we use the notation $X \sim F(v_1, v_2)$ to denote that “ X has an F distribution with v_1 and v_2 degrees of freedom”.

F distribution p.d.f.'s



Stat Trek or Minitab can compute p-values and determine critical values for a given F distribution.

Example: $X \sim F(5,7)$

For what c is $P(X \geq c) = .10$?

F Distribution Calculator: Online Statistical Table

The F distribution calculator makes it easy to find the cumulative probability associated with an F value. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the F distribution, read Stat Trek's [tutorial on the F distribution](#).

▪ Enter values for degrees of freedom.
▪ Enter a value for one, and only one, of the remaining text boxes.
▪ Click the **Calculate** button to compute a value for the blank text box.

Degrees of freedom (v_1)	<input type="text" value="5"/>
Degrees of freedom (v_2)	<input type="text" value="7"/>
Cumulative probability: $P(F \leq 2.88)$	<input type="text" value=".9"/>
f value	<input type="text" value="2.88"/>

Calculate

For what c is $P(X \geq c) = .01$?

F Distribution Calculator: Online Statistical Table

The F distribution calculator makes it easy to find the cumulative probability associated with an f value. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the F distribution, read Stat Trek's [tutorial on the F distribution](#).

▪ Enter values for degrees of freedom.
▪ Enter a value for one, and only one, of the remaining text boxes.
▪ Click the **Calculate** button to compute a value for the blank text box.

Degrees of freedom (v_1)	<input type="text" value="5"/>
Degrees of freedom (v_2)	<input type="text" value="7"/>
Cumulative probability: $P(F \leq 7.46)$	<input type="text" value=".99"/>
f value	<input type="text" value="7.46"/>

Calculate

If an F statistic is 10 (with degrees of freedom 5 and 7), what can we say about an upper-tailed P-value?

F Distribution Calculator: Online Statistical Table

The F distribution calculator makes it easy to find the cumulative probability associated with an f value. For help in using the calculator, read the [Frequently-Asked Questions](#) or review the [Sample Problems](#).

To learn more about the F distribution, read Stat Trek's [tutorial on the F distribution](#).

- Enter values for degrees of freedom.
- Enter a value for one, and only one, of the remaining text boxes.
- Click the **Calculate** button to compute a value for the blank text box.

Degrees of freedom (v_1)	5
Degrees of freedom (v_2)	7
Cumulative probability: $P(F \leq 10)$	0.996
f value	10

Calculate

Single Factor Analysis of Variance

Suppose we have I populations ($I \geq 3$). The populations are often called *levels*.

The variable identifying the levels is called a *factor*.

A factor variable may be categorical, but it also may identify different treatment groups in a controlled experiment.

From each population i we take an independent sample of size J_i .

$$X_{i1}, \dots, X_{iJ_i} \sim N(\mu_i, \sigma^2)$$

Note: All variances and standard deviations are assumed to be equal.

The total sample size is $N = J_1 + J_2 + \dots + J_l$.

We wish to test

$H_0: \mu_1 = \mu_2 = \dots = \mu_I$ vs. $H_1:$ Two or more of the μ_i are different.

We begin by estimating the individual population means as

$$\hat{\mu}_i = \bar{X}_{i\bullet} = \frac{1}{J_i} \sum_{j=1}^{J_i} X_{ij}$$

and the common, or *grand*, mean (if H_0 is true) as

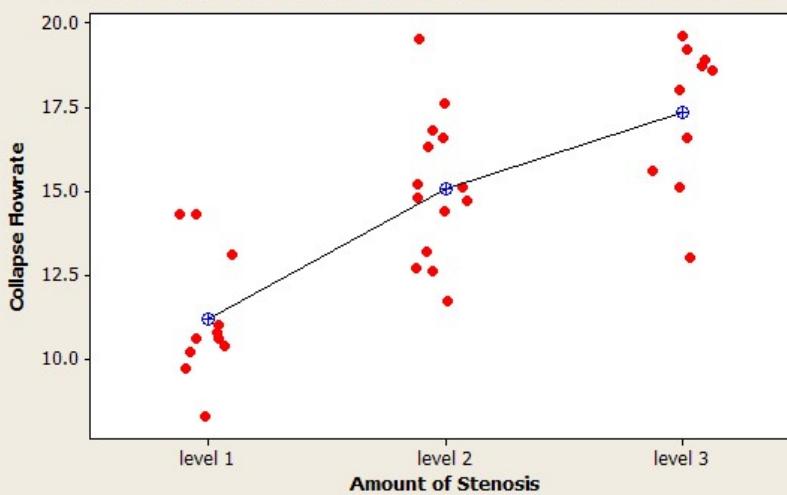
$$\hat{\mu} = \bar{X}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{J_i} X_{ij} = \frac{1}{N} \sum_{i=1}^I J_i \bar{X}_{i\bullet} .$$

Example: Artery data (measure flow rate until artificial arteries collapse).

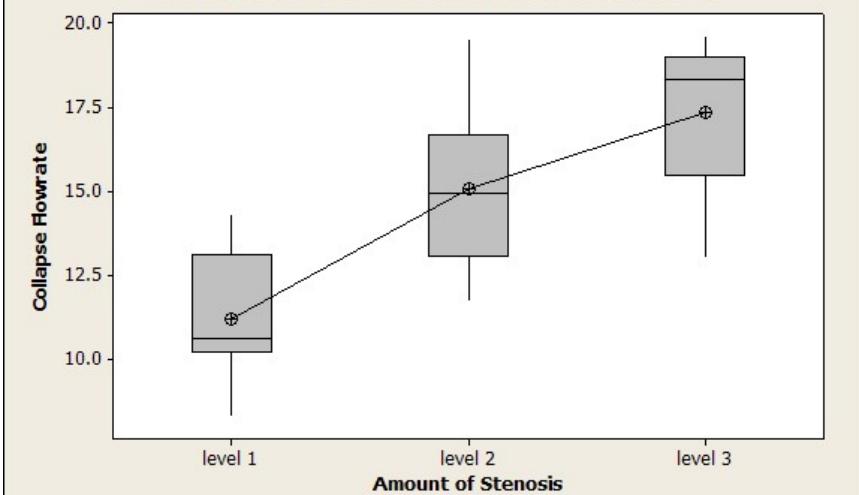
	Stenosis		
	Level 1	Level 2	Level 3
Flowrate (ml/s) at collapse	10.6 9.7 :	11.7 12.7 :	19.6 15.1 :
	8.3	17.6	16.6
J_i	11	14	10
$\bar{X}_{i\bullet}$	11.209	15.086	17.330

$$N = 11 + 14 + 10 = 35$$

Individual Value Plot of Collapse Flowrate vs Amount of Stenosis



Boxplot of Collapse Flowrate by Amount of Stenosis



The variability *between* levels is estimated by the *treatment sum of squares*.

$$\text{SSTr} = \sum_{i=1}^I J_i (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^I J_i \bar{X}_{i\bullet}^2 - N \bar{X}_{\bullet\bullet}^2.$$

The variability *within* levels is estimated by the *error sum of squares*.

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{i\bullet})^2 = \sum_{i=1}^I \sum_{j=1}^{J_i} X_{ij}^2 - \sum_{i=1}^I J_i \bar{X}_{i\bullet}^2 \\ &= \sum_{i=1}^I (J_i - 1) s_i^2. \end{aligned}$$

The *total* variability of the dataset is found as the *total sum of squares*.

$$\begin{aligned} \text{SST} &= \sum_{i=1}^I \sum_{j=1}^{J_i} (X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^I \sum_{j=1}^{J_i} X_{ij}^2 - N \bar{X}_{\bullet\bullet}^2 \\ &= (N-1)s^2. \end{aligned}$$

where s^2 is the sample variance of the full dataset.

Note that $\text{SST} = \text{SSTr} + \text{SSE}$.

Example (Artery data):

$$\bar{X}_{1\bullet} = 11.209, \quad \bar{X}_{2\bullet} = 15.086, \quad \bar{X}_{3\bullet} = 17.330$$

$$J_1 = 11, \quad J_2 = 14, \quad J_3 = 10, \quad N = 35$$

$$\sum_{i=1}^3 \sum_{j=1}^{J_i} X_{ij}^2 = 10.6^2 + 11.7^2 + \cdots + 16.6^2 = 7710.39$$

To test our hypotheses, we compare these two measures of variability in an F-statistic

$$F = \frac{SSTr/(I-1)}{SSE/(N-I)}.$$

The numerator and denominator of this statistic are called the mean square for treatments and the mean square error, respectively.

If the null hypothesis is true, then the mean square for treatments and the mean square error are both estimates of the common variance, σ^2 , and

$$F \sim F_{I-1, N-I}.$$

Large differences between the level means will lead to large values of F , so the P-value is defined as $P = P(F \geq f)$, where f is the observed test statistic value.

We summarize our computations in an ANOVA table.

Source	Sum of Squares	df	Mean Squares	F	P-value
Treatment	203.863	2	101.932	23.497	0.00
Error	138.819	32	4.338		
Total	342.682	34			

Since the p-value is so small, we reject the null hypothesis of equal means and conclude that two or more means differ from one another. Note: $p - value = P(F \geq 23.497) \approx 0$ (Stat Trek)

Minitab Output:

Results for: arteries.txt

One-way ANOVA: Collapse Flowrate versus Amount of Stenosis

Source	DF	SS	MS	F	P
Amount of Stenos	2	204.02	102.01	23.57	0.000
Error	32	138.47	4.33		
Total	34	342.49			

S = 2.080 R-Sq = 59.57% R-Sq(adj) = 57.04%

Pairwise Comparisons

If an ANOVA test rejects H_0 , it suggests that at least some of the level means are different from one another, but does not automatically identify which ones.

We can use the Tukey-Kramer multiple comparisons procedure which controls the overall Type I error level by adjusting for the number of tests.

Such a procedure will have probability α of one or more Type I errors (false significant differences) out of the full set.

The more tests conducted, the smaller the individual probabilities of Type I error (and the more conservative the individual tests) must be.

In addition to the level means and sample sizes, we need an estimate of the common variance, σ^2 .

The mean square error,

$$MSE = \frac{SSE}{N - I} = \hat{\sigma}^2$$

is an estimate of σ^2 , so we use that.

We also need a critical value, $q_{I,N-I,\alpha}$ from the Studentized range distribution, which adjusts for the multiple comparisons.

A $100(1-\alpha)\%$ confidence interval for the difference between level means μ_i and μ_j is

$$\bar{X}_{i\bullet} - \bar{X}_{j\bullet} \pm q_{I,N-I,\alpha} \sqrt{\frac{MSE}{2} \left(\frac{1}{J_i} + \frac{1}{J_j} \right)}.$$

Then μ_i and μ_j are considered significantly different at level α if this interval does not include 0.

Example: (Artery data)

$$MSE = \frac{138.467}{35 - 3} = 4.327$$

$$q_{3,32,.05} \approx 3.49$$

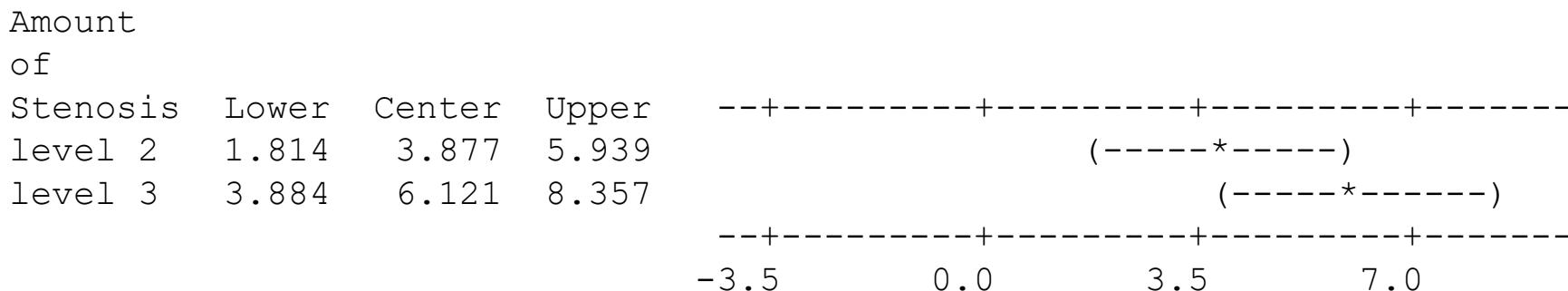
Find 95% simultaneous c.i.s for:

Which levels are significantly different?

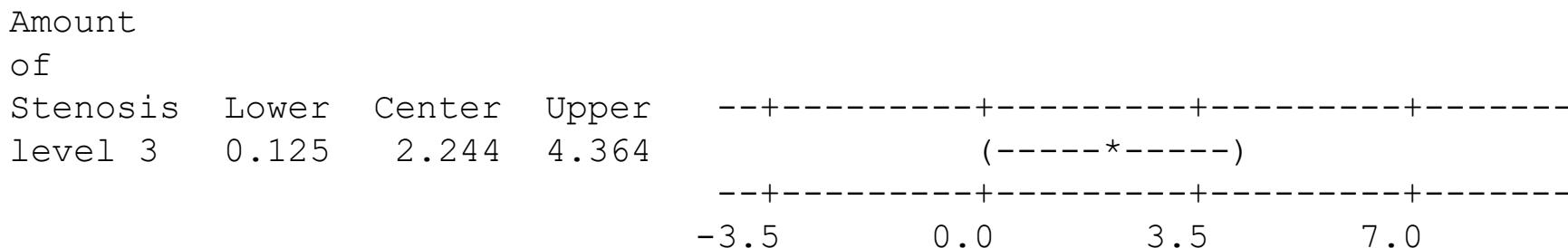
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Amount of Stenosis

Individual confidence level = 98.06%

Amount of Stenosis = level 1 subtracted from:



Amount of Stenosis = level 2 subtracted from:



Example: Wear tests were performed on metal artificial hip joints. The data presented in the following table on head roughness are consistent with the means and standard deviations reported in the article.

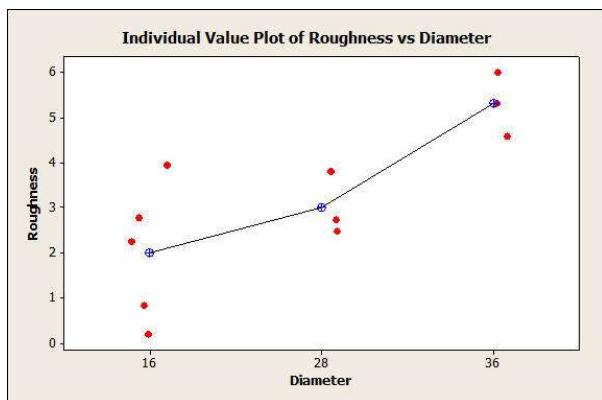
Diameter (mm)	Roughness (nm)
16	.83, 2.25, .20, 2.78, 3.93
28	2.72, 2.48, 3.80
36	5.99, 5.32, 4.59

Descriptive Statistics: Roughness

Total					
Variable	Diameter	Count	Mean	StDev	Variance
Roughness	16	5	1.998	1.500	2.251
	28	3	3.000	0.703	0.494
	36	3	5.300	0.700	0.490

Descriptive Statistics: Roughness

Total				
Variable	Count	Mean	StDev	Variance
Roughness	11	3.172	1.776	3.154



One-way ANOVA: Roughness versus Diameter

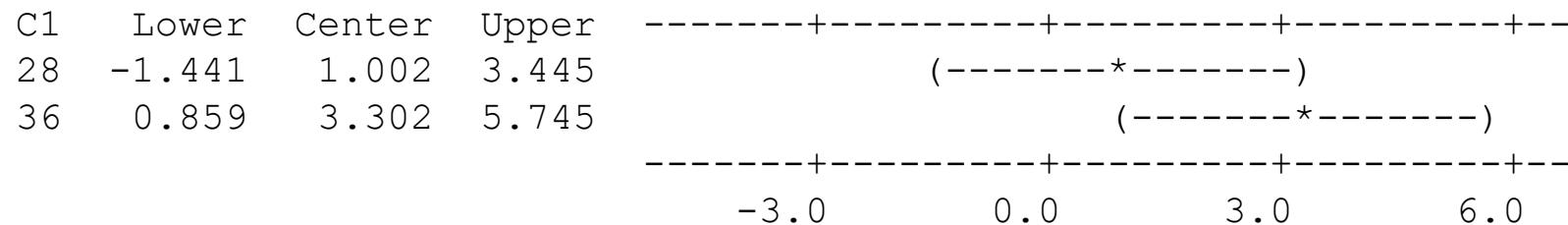
Source	DF	SS	MS	F	P
Diameter	2	20.57	10.28	7.50	0.015
Error	8	10.97	1.37		
Total	10	31.54			

S = 1.171 R-Sq = 65.21% R-Sq(adj) = 56.51%

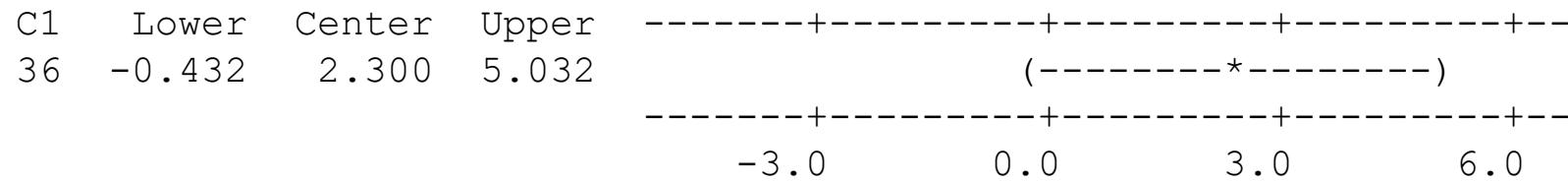
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of C1

Individual confidence level = 97.87%

C1 = 16 subtracted from:



C1 = 28 subtracted from:



Model Assumptions

Remember that ANOVA is based on a model of independent draws from normal populations with a common variance.

Minor deviations from normality or a common variance will not have a strong effect, but large deviations will require other techniques.

Histograms, boxplots, and prior experience can all be useful guides.