Welcome to FES 510(e) Introduction to Statistics in the Environmental Sciences

Syllabus Overview

- On CANVAS under web page and syllabus. Updated periodically.

Videos

To Flip or Not to Flip . . .

Why I'm here . . .

Your Questions . . .

http://www.sciencedirect.com/science/article/pii/S0006320716303044

Things to Do BEFORE TUESDAY

- Make Sure you have FES 510 or FES 510e on your list of classes on CANVAS – you should be added automatically if you've registered on Yale SIS : http://www.yale.edu/sis.
- Sign up for MINITAB intro session if you like (don't bother if you're doing flipped) : www.reuningscherer.net/MINITAB
- PRINT OUT NOTES if you want hardcopy – available in resources folder online
- Get a clicker/card at Bass Library Circulation Desk

What is Statistics?

- Like dreams, statistics are a form of wish fulfillment *– Jean Baudrillard*

"As you can see, we now have conclusive proof that smoking is one of the leading causes of statistics" – Fletcher Knebel

More charitable view :

- Statistics is the art of stating in precise terms that which one does not know. *– William Kruskal*

In other words :

- Statistics is about quantifying **VARIATION** *(De Veaux et all)*

*Statistics helps us address **hard questions**.*

- Will Hillary win? *(Two-sample test of proportions / logistic regression)*
- Does Roundup™ really degrade into harmless substances after the specified amount of time? *(ANOVA)*
- Does street cannabis impair brain function in MS patients? *(two-sample t-test)*
- Is global warming happening in CT? *(regression)*
- Does the probability of a woman washing her hands in the Kmart bathroom change if she knows she's being watched? *(Hypothesis Testing for Binomial)*

Statistics are also important in separating fact from fiction (despite what you might think . . .)

- Statistics turned medicine into a field that (by 1910) had a better than even chance of helping you rather than hurting you (no more leeches . . .)

Example *: which animals kill the most humans?*

What will we cover?

Another Definition of Statistics :

"The science of **collecting**, **organizing**, and **interpreting** numerical facts, which we call data." (Moore and McCabe)

I **Organizing** – how to look at the data!

- distributions, graphical displays *(histograms, boxplots,...)*
- numerical summaries *(mean, median, standard deviation,...),*
- Normal distributions
- more than one variable: correlation, regression

II **Collecting** (and producing) data : Sampling, bias, design of experiments, randomization

Probability *(need for Interpretation) Random variables , rules of probability, c probability, Bayes' rule, binomial & Normal distributions, Central Limit Theorem*

III **Interpreting** data -- <u>Statistical inference</u>

- Confidence intervals
- Hypothesis testing
- Advanced Techniques *(mostly in sections)*
  - o Regression
  - o ANOVA
  - o Multiple Regression

Statistical Inference *(what's in the bag . . ?)*

Usual research situation:

- We have a question about a ***population.***
- We take a ***sample*** of individuals from the population.
- Quantify our question with a ***parameter***, <u>a number describing the **population**.</u>
- Using our sample, calculate a ***statistic***: <u>a number describing the **sample**.</u>

*Note : the **P**'s and **S**'s go together*

***P**opulation*	←→	***P**arameter*
***S**ample*	←→	***S**tatistic*

SO : If we design our statistics wisely . . . . .

- We **INFER** that what we observe about our sample is true of the population
- That is, we infer that our **sample statistic** approximates the **population parameter**.

Goal of Statistical INFERence : quantify the success of our approximation!!

Example : Online Article. *Discussed in class*

Population, parameter, sample, statistic?

Data – the Statistician's Raw Material (Cartoon Guide)

Data consists of values of some ***variables*** measured on ***individuals*** from a ***population***.

Population	Individual – a <u>noun</u>	Variable – a <u>characteristic</u> of the individual
All adults in 18 countries http://news.nationalgeographic.com/news/2014/09/140926-greendex-national-geographic-survey-environmental-attitudes/	Person	Greendex Score (level of sustainable consumer behavior)
A frog egg mass	Hatched tadpole	Death temperature when boiled
Past 10 years	An hour of trading on the futures market	Oil spot price

Quantitative variable (*continuous*) **–** a *numeric* valued variable, where math on the variable makes sense. Has associated **units** (don't forget about these!!)

Categorical Variables (*factor, discrete variable, dichotomous variable (two levels))*– places an individual into one of several groups or *categories*

Individual	Categorical Variable	Quantitative Variable
Person http://news.nationalgeographic.com/news/2014/09/140926-greendex-national-geographic-survey-environmental-attitudes/	Greendex Score Up or down	Greendex Score (scale of 1-100)
Tadpole	Dead/Alive	Weight (mg)
Hour on Futures Market	Oil spot price Up or Down (BINARY variable)	Oil Price (dollars)

Sometimes a variable can be made categorical or quantitative

***Example** : Person's age in years, or a person's age category (0-20 years, 21-40, 41-60, 60+)*

Categorical Variables are further subdivided into

- **ORDinal** – there is a natural ORDering to the categories (Low, Medium, High or Agree, Disagree, etc)
- **NOMinal** – categories are an unordered group of NAMes (red, green, blue)

Who cares?!?

Different statistical tools have been developed for different kinds of data. What tools you use depends on the kind of data you collect and what you want to know

Let's suppose you have some data – now what do you do?!?!?

Displaying Data

Before you try fancy statistical analyses, <u>always</u> make some data displays

(A picture is worth a thousand data points!)

Statistical Graphics :

- Reveal patterns that numbers do not
- Show important **<u>patterns</u>** and **<u>relationships</u>** in your data (more on this in regression!!!)
- A concise, effective way to tell others about your data.

Example : Crimean War*.* *March 1854, Russia defeats a Turkish fleet in the Black Sea, threatening British and French shipping who join the war against Russia. They fight on the Crimean Peninsula through 1856, ultimately defeating Russia at a cost on all sides of some 300,000 lives.*

Florence Nightingale was one of the nurses with the British army during the conflict. She began to keep meticulous records on her patients. Her records probably looked something like this :

Soldier	Regiment	Date of Death	Cause of Death
Frank Butler	13th Light Cavalry	10/15/1855	Cholera
George Parson	7th Infantry	10/15/1855	Cholera
Michael Summers	13th Light Cavalry	10/15/1855	Horse Kick
Matthew Green	14th Infantry	10/16/1855	Gunshot
……			

In total, Nightingale records nearly 18,000 deaths in army hospitals during a two year period.

Making Statistical Graphs Step 1 : Make piles

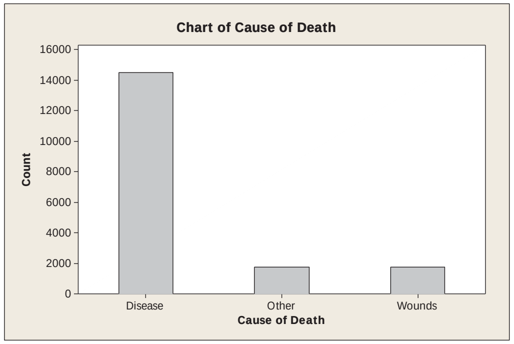
Frequency Table – gives number of individuals at each level of a variable.

Nightingale examined deaths in her hospital between April, 1854 and March, 1856. Three categories : deaths as due to battle wounds, deaths due to preventable disease, and other causes (various) :

Cause of Death	Number of Deaths
Wounds	1758
Disease	14476
Other	1748
Total	17982

Making Statistical Graphs Step 2 : Turn piles into pictures

Bar Chart : AREA = RELATIVE FREQUENCY



MINITAB: use `Graph` → `Bar Chart` and choose `SIMPLE`. You can enter raw data or summarized data (for summarized data, click on `DATA OPTIONS`, choose `FREQUENCY`, and enter the variable with the summarized counts).

SPSS: SPSS has a Chart Builder which lets you drag variables. This works fine. You can also use `Graphs` → `Legacy Dialogs` → `Bar.` Choose `Simple`. For the Nightingale data, choose Bars Represent `OTHER` and then `SUM(Number_Dead).` Category Axis is `Death_Cause`

- Bar heights gives the count of observations in each death category. This is called the <u>distribution</u> of the variable cause of death

A Distribution gives

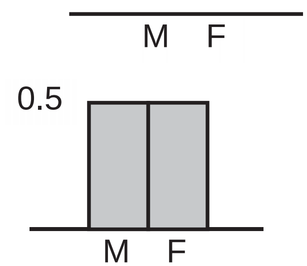
- **A list of all possible values of a variable** *(i.e. list of all possible causes of death)*
- **The frequency with which each value of the variable occurs** *(i.e. how many deaths from disease, how many from wounds, how many from other causes).*

Sampling Distribution : distribution of a SAMPLE (known and measured)

Population/True Distribution : distribution of the entire population (unknown and estimated with sampling distribution!!)

Example *:* ***Gender*** *: Two values : M, F*

<u>Sampling Distribution</u> : Sample of size 10 yields ?– draw a ***histogram***

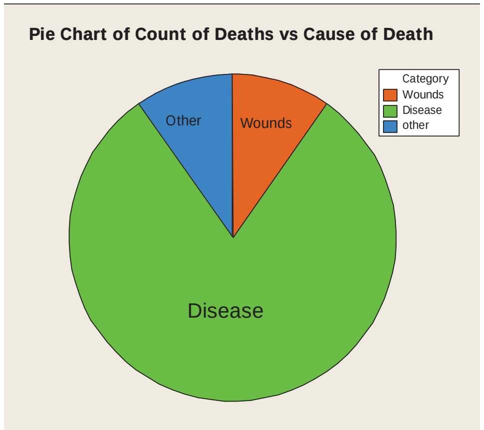


<u>True distribution</u> : (about .5, .5)

Height : Range of possible values?

Pie Chart – advantage is that it clearly shows relative proportion in each category.

AREA = RELATIVE FREQUENCY



MINITAB : use `Graph → Pie Chart.` For this example, choose `Chart Values from a Table.` Enter variables in the appropriate boxes.

SPSS: `Graphs → Legacy Dialogs → Pie.` Choose `Simple.` For the Nightingale data, choose `Summaries for groups of cases. Define Slices` by `Death_Cause`, `Slices Represent Number_Dead.`

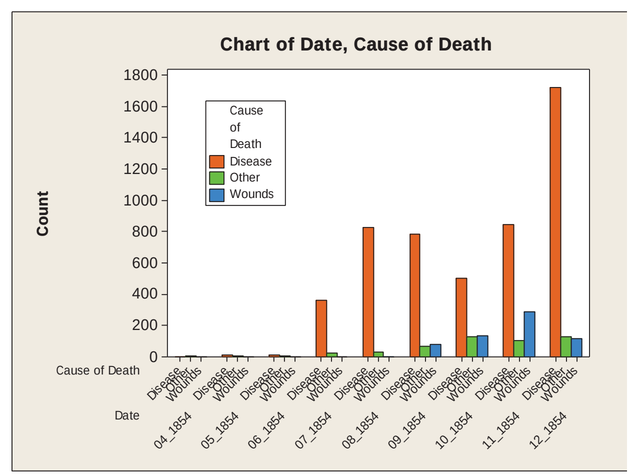
Sometimes, want to consider **two categorical variables at the same time**. Use a **two-way** or **contingency table** to display this information : each **cell** contains the count of individuals who had a combination of categorical characteristics.

***Example** : Crimean War. Nightingale recorded deaths in each category for each month over a two year period :*

Date	Died of Wounds	Died of Disease	Died of Other Causes	Total Deaths	% Disease Deaths
Apr_1854	0	1	5	6	0.17
May 1854	0	12	9	21	0.57
Jun 1854	0	11	6	17	0.65
Jul 1854	0	359	23	382	0.94
Aug 1854	1	828	30	859	0.96
Sep 1854	81	788	70	939	0.84
Oct 1854	132	503	128	763	0.66
Nov 1854	287	844	106	1237	0.68
Dec 1854	114	1725	131	1970	0.88
Jan 1855	83	2761	324	3168	0.87
Feb 1855	42	2120	361	2523	0.84
Mar 1855	32	1205	172	1409	0.86
Apr 1855	48	477	57	582	0.82
May 1855	49	508	37	594	0.86
Jun 1855	209	802	31	1042	0.77
Jul 1855	134	382	33	549	0.7
Aug 1855	164	483	25	672	0.72
Sep 1855	276	189	20	485	0.39
Oct 1855	53	128	18	199	0.64
Nov 1855	33	178	32	243	0.73
Dec 1855	18	91	28	137	0.66
Jan 1856	2	42	48	92	0.46
Feb 1856	0	24	19	43	0.56
Mar 1856	0	15	35	50	0.3
Total	1758	14476	1748	17982	

Bar Chart – can include frequencies for two categorical variables.

Here is bar chart for first 9 months of Crimean War hospital data :



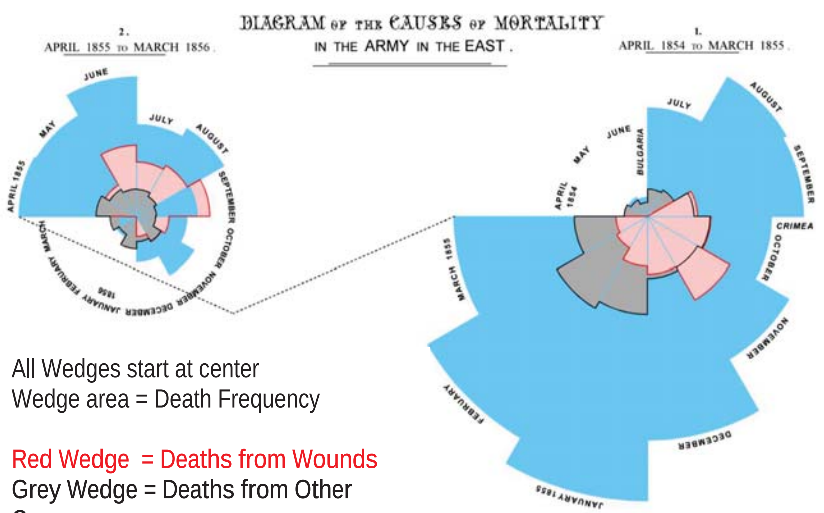
MINITAB : use `Graph → Bar Chart` and choose `Cluster`. You can enter raw data or summarized data (for summarized data, click on `DATA OPTIONS`, choose `FREQUENCY`, and enter the variable with the summarized counts).

SPSS: Use `Graphs → Legacy Dialogs → Bar.` Choose `Clustered.` Then proceed as before, but `Define Clusters` by Death_Cause, `Catogory Axis` as Month_Year.

*Here is how F. Nightingale displayed frequencies for two categorical variables (time and death category) : this is a **Polar Graph** (first one!!)*

Features of Graph :

- *Clearly shows frequencies in three categories of death over time.*
- *Allows comparisons of years – note change between winter of 1854 and winter 1855 (after Nightingale presented year one figures to army brass and got major improvements in sanitation*



All Wedges start at center
Wedge area = Death Frequency

Red Wedge = Deaths from Wounds
Grey Wedge = Deaths from Other Causes

Nightingale saved thousands of lives using statistical graphics!

Graphical methods for displaying *quantitative* sampling distributions

***Example** : 2013 World Poverty Data (data from World Bank). For a sample of 232 countries, examine the average number of children a woman will have over her lifetime.*

Data :

1.7, 4.9, 5.9, 1.8, 3.2, 1.8, 2.2, 1.7, 2.1, 1.9, 1.4, 2.0, 6.0, 1.8, 4.8, 5.6, 2.2, 1.5, 2.1, 1.9, 1.3, 1.6, 2.7, 1.6, 3.2, 1.8, 1.8, 2.0, 2.2, 2.6, 4.4, 1.6, 1.4, 1.5, 1.5, 1.8, 1.7, 4.9, 4.8, 5.0, 2.3, 4.7, 2.3, 1.8, 2.2, 1.4, 1.5, 1.5, 1.4, 3.4, 1.7, 2.5, 2.8, 1.9, 1.8, 2.0, 1.7, 2.6, 2.8, 1.5, 4.7, 1.3, 1.6, 4.5, 1.6, 4.3, 1.8, 2.6, 2.0, 3.3, 4.1, 1.9, 1.8, 3.9, 4.9, 5.8, 4.9, 4.8, 1.3, 2.2, 2.1, 3.8, 2.4, 2.5, 1.7, 1.1, 3.0, 5.0, 1.5, 3.1, 1.3, 2.3, 2.5, 2.0, 1.9, 4.0, 2.0, 3.0, 1.4, 2.3, 3.2, 1.4, 2.6, 4.4, 3.2, 2.9, 3.0, 1.2, 2.2, 2.6, 2.2, 3.0, 1.5, 4.8, 2.4, 1.9, 2.2, 4.2, 4.8, 1.5, 2.3, 2.8, 2.6, 3.0, 1.6, 1.6, 1.4, 1.1, 1.8, 2.7, 1.5, 4.5, 2.3, 2.7, 2.2, 2.4, 1.4, 6.8, 1.4, 1.9, 2.8, 1.7, 2.4, 5.2, 4.7, 1.4, 5.4, 2.0, 1.8, 3.1, 2.3, 7.6, 6.0, 2.5, 1.7, 1.9, 1.9, 2.3, 2.0, 1.7, 1.7, 2.9, 3.5, 3.2, 2.5, 2.4, 3.0, 3.8, 1.3, 1.6, 2.0, 1.3, 2.9, 3.3, 2.1, 2.0, 1.5, 1.7, 4.5, 2.6, 2.6, 4.4, 4.9, 1.2, 4.0, 4.7, 2.2, 6.6, 1.5, 5.0, 4.9, 5.0, 3.2, 4.1, 2.3, 1.3, 1.6, 1.9, 3.3, 2.4, 3.0, 6.3, 4.6, 1.4, 3.8, 2.3, 5.2, 3.8, 1.8, 2.3, 2.0, 5.2, 5.9, 1.5, 1.9, 2.0, 1.9, 2.2, 2.0, 2.4, 1.8, 1.7, 3.4, 4.0, 2.5, 4.1, 4.1, 2.4, 5.9, 5.7, 3.5

Making Statistical Graphs Step 1 : Make piles

1) Sort data from low to high :

1.1, 1.1, 1.2, 1.2, 1.3, 1.3, 1.3, 1.3, 1.3, 1.3, 1.3, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.7, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.8, 1.9, 1.9, etc.

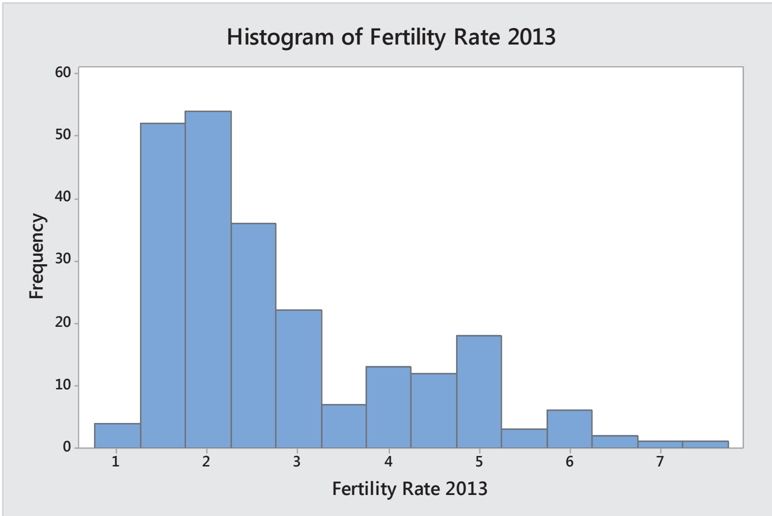
2) Make equal sized data ranges and count the number of observations in each range :

Range	0.75 to 1.25	1.25 to 1.75	1.75 to 2.25	Etc.
Count	4	52	54	

Making Statistical Graphs Step 2 : Make a picture

Histogram – basically a bar chart where the range bins are the grouping variable. Once again,

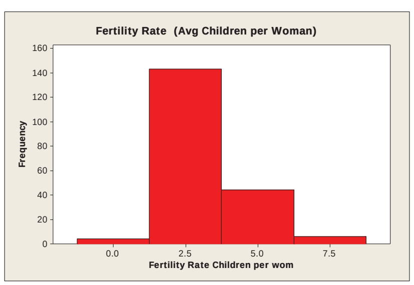
AREA = RELATIVE FREQUENCY

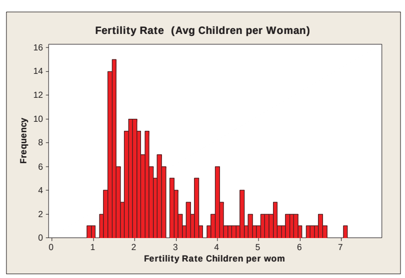


MINITAB : use `Graph → Histogram.` Enter the variable with the raw data : you can let MINITAB choose the bins.

SPSS notes : use `Graphs → Legacy Dialogs → Histogram` or `Graphs → Chart Builder.`

Note : the shape of a histogram is highly dependent on the number of bins – in general, let the computer choose!

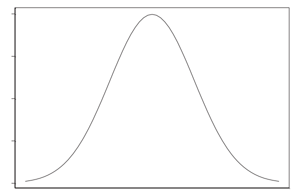




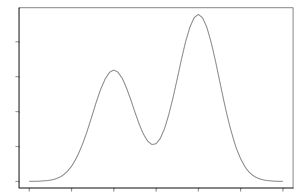
Words that describe quantitative distributions

Symmetric : one half is a mirror image of the other half
Asymmetric : not symmetric *(AAAHH!)*

Unimodal : one mode
Bi-modal : two modes
(you get the idea . . . .)



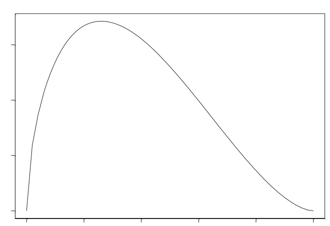
Symmetric, unimodal



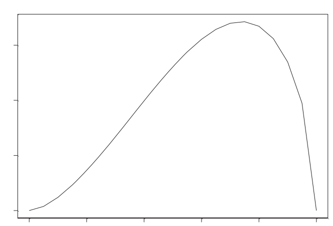
Not symmetric, bimodal

A **Mode** of a distribution is
- A local maximum (math definition)
- A peak
- The most common value (and the second most common value, etc)

Skewness :



Skewed to the right



Skewed to the left

Numerical Descriptions Of Sample Distributions

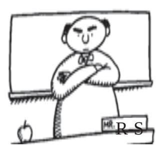
The CENTER and the SPREAD

– the CENTER

Two primary measures :

- **Sample Mean** = average
- **Sample Median** = Middle value = 50th percentile

MEAN



Example *(Journal of Forensic and Legal Medicine, 2010) :*
Brain weights of six men in Tehran (grams) :

Sample Mean $= \dfrac{1290+1306+1285+1279+1243+1317}{6} = 1287g$

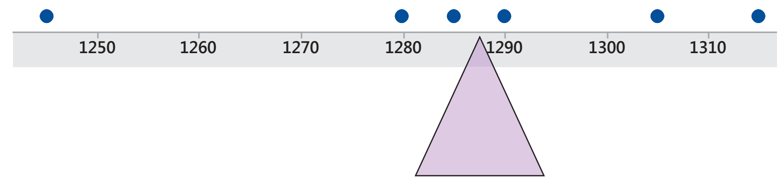
For a variable x with n observed values $x_1, x_2, \ldots, x_n$, the **sample mean** of x (called 'x-bar') is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ex : *brain weights of people, number of observations is* $n=6$

"Physical" interpretation of Mean :

Mean is the Center of mass – balance point of a distribution



$\bar{x} = 1287$

MEDIAN

Sample Median
= Middle observation
= 50th percentile
= the value such that 50% of the data is less than that value

Calculating Sample Median *M* –
first, sort data from low to high. Then

For **ODD** number of observations, sample median is middle observation

Ex : 5 test scores :	*Data :*	*72, 57, 88, 76, 93*
	Sorted :	*57, 72, **76**, 88, 93*
	Median :	***76***

For **EVEN** number of observations, sample median is average of middle observations

Example: *Height of six students in class in cm*

Data :	*178, 163, 168, 167, 170, 150*
Sorted :	*150, 163, **167, 168**, 170, 178*
Median :	***167.5***

QUARTILES $\frac{1}{4}, \frac{3}{4}$

First quartile Q_1 is

- the sample median of the observations below the median
- the 25th percentile
- the value such that 25% of the data is below this value

Third quartile Q_3 is

- the sample median of the observations above the median
- the 75th percentile
- the value such that 75% of the data is below this value

(So the second quartile would be the ????)

Ex : *people heights*

*150, 163, **167, | 168**, 170, 178*

$$M = 167.5$$

$$Q_1 = 163 \qquad Q_3 = 170$$

Note : some programs/books will include the median when calculating *quartiles* :

*150, 163, **167, (167.5) 168**, 170, 178*

$$M = 167.5$$

$$Q_1 = 165 \qquad Q_3 = 169$$

Both answers are fine!

PERCENTILE : the value below which a particular percentage of the observed data fall. *(there are several ways to calculate this number, so don't be surprised if your calculator/computer gives a different answer)*

Find k th percentile	***Ex :Find percentile in Height Data***
1) Sort data	*150, 163, 167, 168, 170, 178*
2) Multiply sample size n by the percentile. Call this T.	*Get 37th percentile :* $T = 6 * 0.37 = 2.22$ *Get 50th percentile :* $T = 6 * 0.5 = 3$
3) If T is NOT an integer, round up. The corresponding observation in the data is the k th percentile.	*37th percentile : round up 2.2 to 3, third data point is 167cm, the 37th percentile.*
4) If T IS an integer, percentile is the average of the T th and $(T+1)$st observations	*50th percentile : average of third and fourth data points is 167.5cm*

OUTLIER : An unusually large or small observation

Example *: One year, on the first day of one class, I asked students to give the probability they would actually take the class. Here are some of the values.*

1.0, 0.9, 0.99, 1.0, 0.3, 0.95, 1.0, 0.5, 7.0, 1.0

ROBUST or **RESISTANT**

A statistic is **robust** (or **resistant**) if it is not sensitive to outliers.

FACT – The MEDIAN is more robust than the MEAN

Example *: probabilities of taking this class :*

1.0, 0.9, 0.99, 1.0, 0.3, 0.95, 1.0, 0.5, 7.0, 1.0

Mean = 1.46 (impossible), Median = 0.99

Assuming that 7.0 was supposed to be 0.7,

Mean = 0.83, Median = 0.97

Median changes very little, Mean changes quite a bit!

Mean vs. Median

- Robustness is nice, protects against outliers
- However, MEAN is often still of more interest, say, where **$Money$** is concerned.

Example *: Damage from hurricanes 1991-2005*

Year	Damage (billions of 2002 $)	Year	Damage (billions of 2002 $)
2005	Wilma ($10)	**1999**	Floyd ($6)
2005	Rita ($10)	**1998**	Georges ($6)
2005	Katrina ($100)	**1998**	Bonnie ($1)
2005	Denis ($2)	**1996**	Fran ($5)
2004	Jeanne ($6.9)	**1995**	Opal ($3)
2004	Ivan ($14)	**1995**	Marilyn ($2)
2004	Frances ($9)	**1992**	Iniki ($2)
2004	Charley ($15)	**1992**	Andrew ($27)
2003	Isabel ($5)	**1991**	Bob ($2)

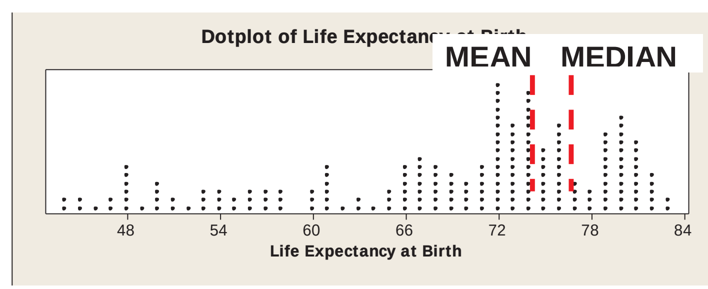
Mean = 12.5 B Median = 6 B, but we pay for mean!

Mean, Median, and Skewness

(and the **dotplot** – like a histogram)

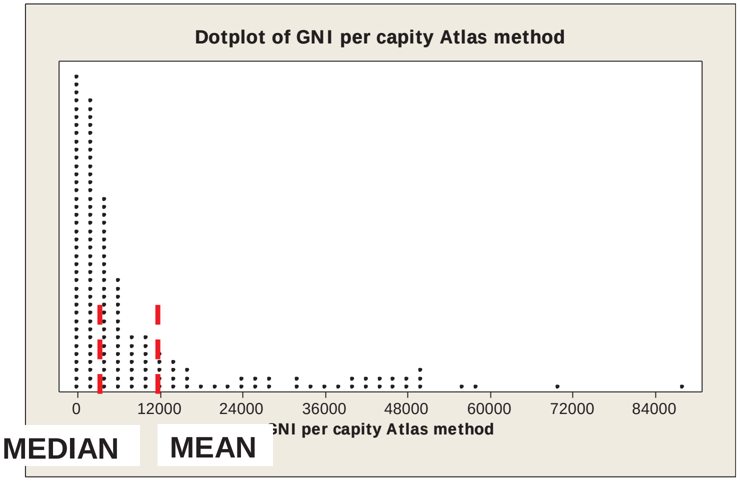
In a **Left Skewed** distribution, **MEAN < MEDIAN**

Example : Poverty Data. *Life Expectancy for 198 Countries. Mean=69, Median=72*



In a **Right Skewed** distribution, **MEAN > MEDIAN**

Example : Poverty Data. *GNI per capita for 198 Countries around the world. Mean=10637, Median=3715*

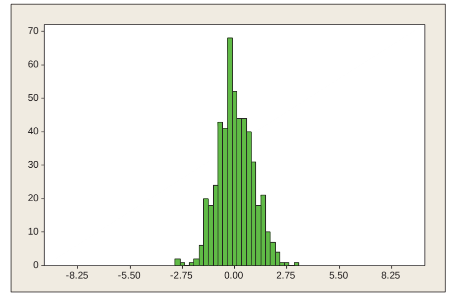


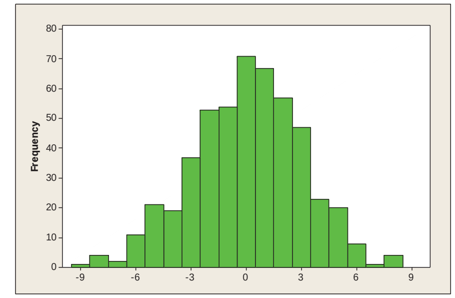
In a **Symmetric** distribution,
MEAN = MEDIAN

MINITAB Dotplot : Use `Graph` → `Dotplot.` Choose `Simple`. Enter variable with data.

SPSS: use `Graphs` → `Legacy Dialogs` → `Scatter/Dot` and then choose `Simple Dot.` Enter variable with data.

NOW : Consider the two plots below (both roughly symmetric so mean=median). The centers in both plots are about zero. So what's different between the plots ?





Numerical descriptions of sample distributions

– the SPREAD or VARIABILITY

FIRST – we discuss spread around a MEDIAN

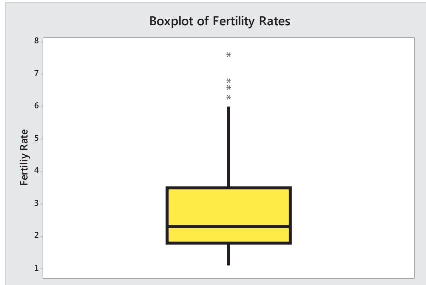
BOXPLOTS (and the Interquartile range IQR)

- **Range** = maximum – minimum
- **Interquartile Range (IQR)** : Q3 - Q1. This gives the width of the middle 50% of the data
- **Five Number Summary** :

{min, Q1, median, Q3, maximum}

Boxplot – a graph based on the five number summary that helps detect outliers

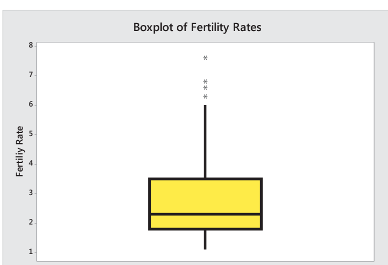
Example : *Fertility Rates in 232 Countries*

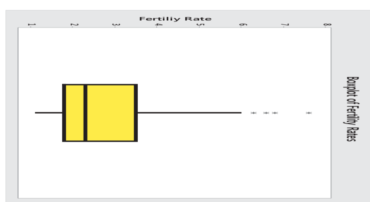


<u>Making a Boxplot</u>

- Central box spans Quartiles
- Middle line marks median
- Observations more than 1.5*IQR above Q_3 or below Q_1 are **possible outliers**
- **Explicitly :** Potential Outliers are
  - o Observations greater than Q_3 + 1.5*IQR
  - o Observations smaller than Q_1 - 1.5*IQR
  - o Marked by an Asterisk *

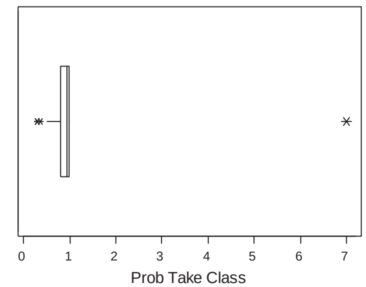
- 'Whiskers' extend to **largest and smallest observations that are not suspected outliers**
- Doesn't matter if the graph is vertical or horizontal – it's a one dimensional graph! (and width of the box means nothing . . .)

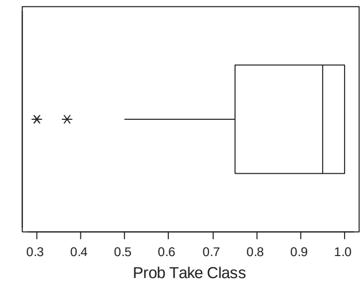




- *Why 1.5? Because John Tukey who invented the boxplot in 1977 choose this as a reasonable cutoff for outliers, and it's still used today. More on this later . . .*

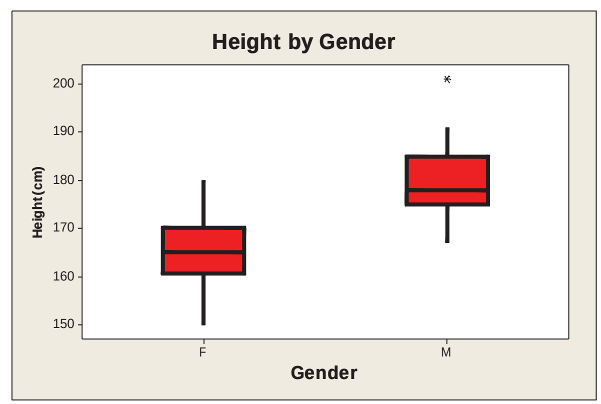
Example *: Probability of taking my class – boxplot of <u>all</u> responses, then with 7.0 changed to 0.7*





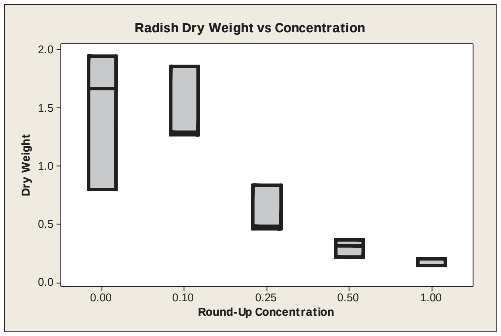
Boxplots are very useful for comparing the distribution of two or more groups

Example *: Height of students in a class by gender*



Example : Roundup™. *Two FES students examined the claim that Roundup degrades into harmless substances after a specified period. They examined 5 levels of Roundup – full strength, down to plain water. The dry weight of 3 different radish plants was measured at each Roundup level.*

Boxplot of Results : where are the whiskers???



MINITAB Boxplot : To make a boxplot, use `Graph` → `Boxplot.` Enter variables. Double-click on graph to change colors, etc.
SPSS: use `Graphs` → `Legacy Dialogs` → `Boxplot` and then choose `Simple.` Enter variable.

Example (let's work it out!) *: Here are values for % urban population for 12 countries (based on year 2000 estimates). Make a boxplot : what is IQR? Any outliers?*

Rwanda	*6*	*Brazil*	*81*	*Australia*	*91*
Chad	*24*	*Denmark*	*85*	*Kuwait*	*96*
Greece	*60*	*Chile*	*86*	*Belgium*	*97*
France	*75*	*Lebanon*	*90*	*Singapore*	*100*

VARIANCE and STANDARD DEVIATION (SD)

- **Most common and useful measure of SPREAD of a distribution. Measures spread around the MEAN**
- Relationship: $\text{Standard Deviation} = \sqrt{\text{Variance}}$
- Notation :

  Sample Variance = s^2,
  Standard Deviation = s

  The **S**ample Variance s^2 is a **S**tatistic.

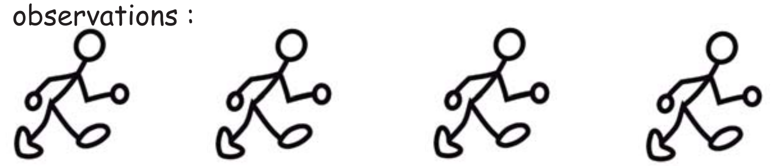
  This is a number we can calculate

- Idea of variance
  - How far away are the observations, on average, from the mean?
  - This calculation involves the <u>DEVIATIONS</u>

- A **deviation** is defined as the difference between an observation and the mean :

$$x_i - \bar{x}$$

Aside : i is an **index** that keeps track of our particular observations. If we collect the heights of 4 people, then our sample size is n = 4 and i is an **index** that keeps track of the observations :



$i = 1$	$i = 2$	$i = 3$	$i = 4$
x_1= 178cm	x_2= 156cm	x_3= 182cm	x_4= 167cm

- Formula for **Sample Variance** : in words, **<u>the average of the squared deviations</u>**

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

Example *: Heights of six people in class (cm) :*

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

x_i *(Data)*	$\bar{x}$ *(Mean)*	$x_i - \bar{x}$ *(Deviations)*	$(x_i - \bar{x})^2$ *(Squared Deviations)*
178	166	12	144
163	166	-3	9
168	166	2	4
167	166	1	1
170	166	4	16
150	166	-16	256

Sum = $\sum_{i=1}^{6}(x_i - \bar{x})^2$ = 430

$$n-1=5$$

Sample Variance :

$$s^2 = \frac{430}{5} = 86$$

Sample Standard Deviation :

$$s = \sqrt{86} = 9.3$$

Why *squared* Deviations?

- Sum of deviations is just 0. Squaring the deviations converts the negative deviations to positive numbers
- *Could use average absolute value of the deviations – called '<u>mean absolute deviation</u>'*
- Summing squares is a natural operation – (think Pythagoras)
- **Real Reason** – Squared Deviations are best way to capture spread in a normal distribution (Fisher, 1920) **AND** they are a more accurate measure of dispersion. See nice discussion here: http://www.separatinghyperplanes.com/2014/04/why-do-statisticians-use-standard.html and also here : http://www.leeds.ac.uk/educol/documents/00003759.htm

Example *: In class . . .*

Why *divide by n-1*?

- If n=1, you shouldn't be calculating a variance!
- If n is big, it doesn't matter anyway
- Real Reason – it makes the estimate **unbiased** (more on this later) – **<u>for large sample sizes, the sample statistic will approach the correct true value!</u>**

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

More on VARIANCE and STANDARD DEVIATION (SD)

- SD of 3, 3, 3, 3, 3, 3 is zero (no variation)
- **Robustness** *(recall that a statistic that is not sensitive to outliers is **robust**)*

IQR is robust; SD is not

***Example** : probabilities of taking my class :*

1.0, 0.9, 0.99, 1.0, 0.3, 0.95, 1.0, 0.5, 7.0, 1.0

SD =2.0 , IQR=0.2

Change 7.0 to 0.7 ⟶ *SD =0.25, IQR=0.35*

IQR changes little, SD changes quite a bit

MINITAB Summary Stats : To get <u>all</u> of the summary statistics discussed so far (mean, median, IQR, SD), use `Stat → Basic Statistics → Display Descriptive Statistics.`

SPSS: use `Analyze → Descriptive Statistics → Descriptives`

Results for 6 heights :

```
Variable  N    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Height    6  166.00     3.79   9.27   150.00  159.75  167.50  172.00

Variable  Maximum
Height     178.00
```

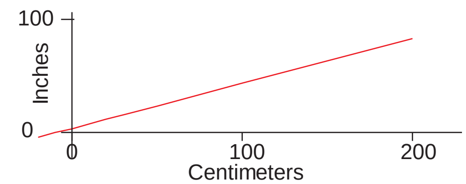

Rules for LINEAR TRANSFORMATIONS of Means and Variances

(this looks boring, but will be useful later – hint hint!!!)

***Example** : Height conversion. Suppose I measured student height in cm and then calculated a mean and standard deviation. I actually want the mean and standard deviation of height in inches while wearing shoes with ½ inch soles. Could convert each data point and recalculate mean/std dev. OR . . . use brain!*

A conversion formula from centimeters (x) to inches (y) :

$$y_i = \frac{1}{2.54} x_i + 0.5$$



*This is called a **linear transformation.***

Suppose we have some data with sample mean $\bar{x}$ and sample variance s_x^2. Make a **linear transformation** (multiply by a, add b) to create $y_i = ax_i + b.$

Rules for Linear Transformations :

- $\bar{y} = a\bar{x} + b$

  (mean of y is a times mean of x times plus b)

- $s_y^2 = a^2 s_x^2$ (no change due to b) i.e. $s_y = |a| s_x$

  (variance of y is the variance of x times a squared)

  (SD of y is the SD of x times the absolute value of a)

Example : *Mean height in inches with shoes :*

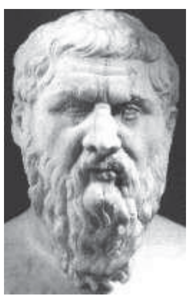
$$\bar{y} = \frac{1}{2.54}166 + 0.5 = 65.9 \textit{ inches}$$

Variance and SD of height in inches with shoes :

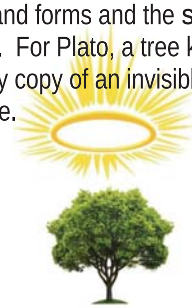
$$s_y^2 = \left(\frac{1}{2.54}\right)^2 * 86 = 13.3 \textit{ (inches }^2\textit{)}$$

$$s_y = \frac{1}{2.54} * 9.3 = \sqrt{13.3} = 3.65 \textit{ inches}$$

Now : A bit of Philosophy . . . .



PLATO : (427-347 BCE). Plato divided existence into two realms (the 'Doctrine of Forms') : The **intelligible realm** of perfect, eternal, invisible ideas and forms and the **sensible realm** of concrete, familiar objects. For Plato, a tree known through the senses was a shadowy copy of an invisible, unchangeable idea of a perfect tree.



What this has to do with statistics . . .

Sample data we collect is **observable** *(it's sensible)*

> Sample → Statistic
>
> Population → Parameter

- We calculate various properties of our sample data (**S**tatistics**)**
  - **Sample Mean**
  - **Sample Variance**
  - **Histogram** – the observed frequency of data in various ranges
- However, our sample data is just an observable reflection of the true, unknown, ideal **Population** *(intelligible)*.
- The population has various unknown, true, fixed properties (**P**arameters):
  - **True Mean** height of people
  - **True Variance** in the height of people
  - **True Histogram** – true relative frequency of observations over various ranges (i.e. 22% of men are between 5'6" and 6'0"). This 'True Histogram' is called a **Probability Density Curve** or **Probability Density Function.**

Notation :

- Things we can observe and calculate are generally notated in **Latin Script**
- Things we cannot know (true parameters) are usually notated in **Greek**

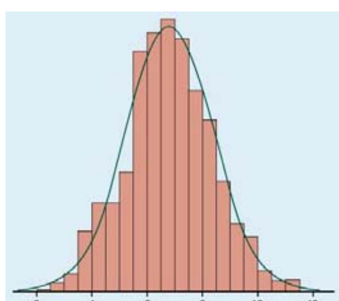
Remember this :

Only the Gods know true properties of populations (parameters), and the Gods speak GREEK!

What we see (mortals)	Invisible Truth (gods)
Sample	Population
Statistic	Parameter
Sample mean $\overline{X}$	True Mean μ
Sample Standard Deviation S	True Standard Deviation σ
. . . . and	
Sample Histogram	True Density Curve

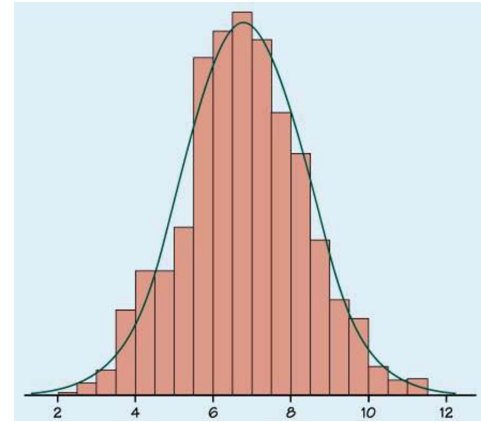
Density curves

- Learned how to describe the center and spread of <u>sample</u> data
- Often, a large number of observations follows a regular pattern that can be described by a smooth curve. We can describe the data with a **mathematical model** called a **density curve**



Density Curve is

- Idealized, smoothed histogram. Limit of large population (sample size $\rightarrow \infty$ *(infinity)*)
- A positive valued curve
- A mathematical curve with area **EXACTLY=1** underneath it



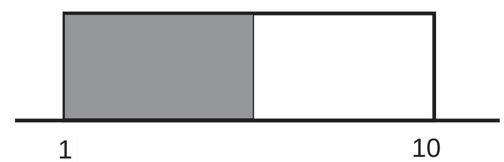
(Aside : for those who admitted knowing calculus, this means that the integral of the function over the entire real line is equal to ***one****!)*

Properties of Density Curves –

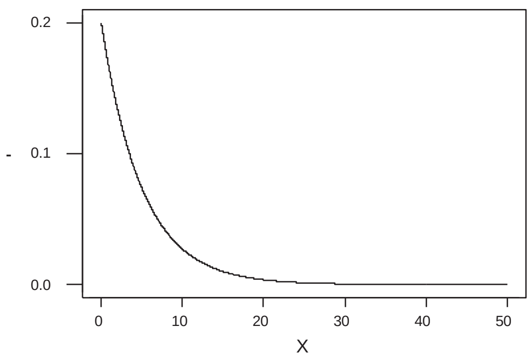
- **Areas under the curve correspond to proportions of population**
- Density curves have means, medians, and standard deviations (IQR's too, but not really useful).
  - o Median of a Density is the point with 50% of area above and below
  - o Mean of a Density is the center of balance – notation μ
  - o Standard deviation measures average spread - notation σ
  - o Use calculus to calculate μ and σ

Examples *of Density Curves :*

Uniform Density *(i.e. flat) on the interval [1,10]. Proportion of the population between 1 and 5 is 0.454545.*



Exponential Density *with mean 5 (used in failure times – like the life of light bulbs)*



Example : Dissolved Oxygen in CT waters*(Robin Kriesberg). Water samples were collected at 12 CT shoreline locations at different times during the day on the surface and 10 meters below the surface. Weather measurements and water chemistry measurements were taken.*

We look at 66 observations of surface of waters in Bridgeport harbor in summer of 2000 and examine the dissolved oxygen levels (DO in mg/liter)

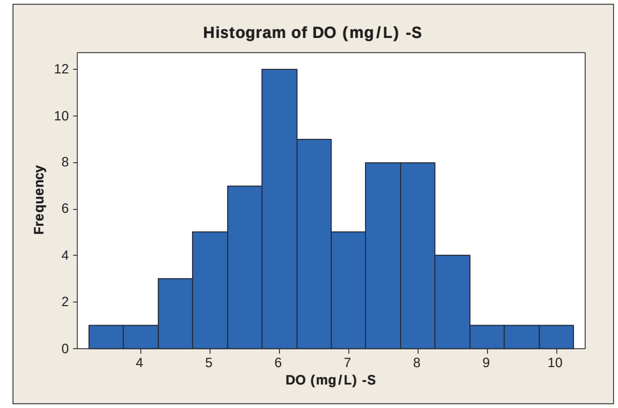
Sample Mean : $\overline{X} = 6.6$ *mg/l*

i.e. the average surface water in Bridgeport had a DO of 6.6 mg/l

Sample standard deviation : $S = 1.4$ *mg/l*

i.e. a typical DO measurement s is 1.4 mg/l away from 6.6

Histogram of Sample Data :



What is the shape of the underlying density? Maybe . . .

NORMAL DENSITY CURVE

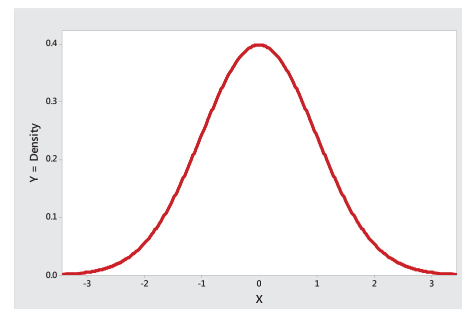
Most important Density Curve

Standard Normal Density

Mean = $\mu = 0$

standard deviation = $\sigma = 1$

Equation is
$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



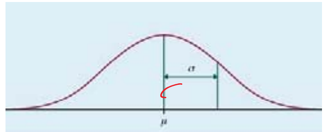
In general, the equation is
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(don't need to memorize these)

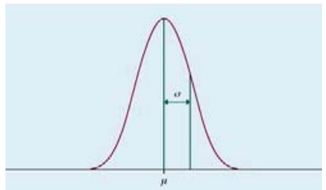
General Normal Densities
Underlying Shape is Always the Same!

Remember This!

σ larger



σ smaller

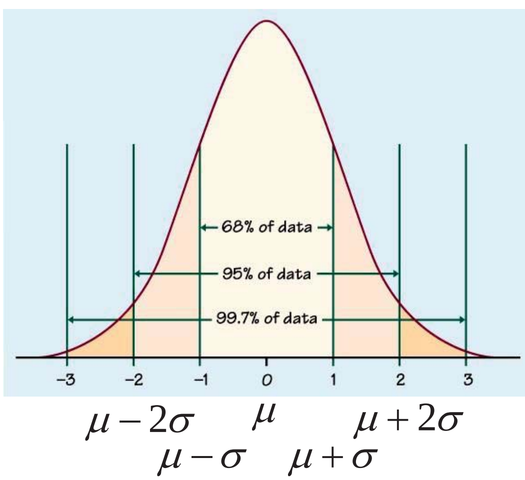


Note that the distance σ away from the mean μ indicates the inflection point of the curve – the place where it goes from concave down to concave up.

Notation : A normal density with mean μ and standard deviation σ is written as $N(\mu, \sigma)$

"68, 95, 99.7 rule"

- 68% of the population is within 1 SD of the mean (i.e. between $\mu - \sigma$ and $\mu + \sigma$)
- 95% of the population is within 2 SD's of the mean (i.e. between $\mu - 2\sigma$ and $\mu + 2\sigma$)
- 99.7% of the population is within 3 SD's of the mean



Example : Dissolved Oxygen

DO content in Bridgeport harbor has an approximately normal distribution with mean 6.6 mg/l and standard deviation 1.4 mg/l. (i.e., data is N(6.6,1.4)).

Let's pretend that the ***TRUE VALUES – the parameters*** *(think gods) are μ=6.6, σ=1.4*

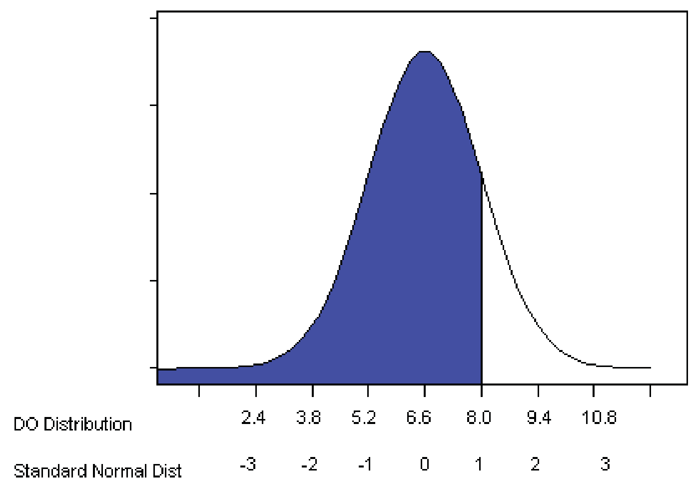
What percentile is a DO level of 8?

- *That is, what percent of the probability density function is below 8?*
- *That is, is we took SAMPLE data, about what percent of the data should have a value of 8 or less?*

Draw a picture :

8.0 is 1 standard deviation ***above*** *the mean. Want the shaded area.*

Think about the same area in a ***standard normal distribution***



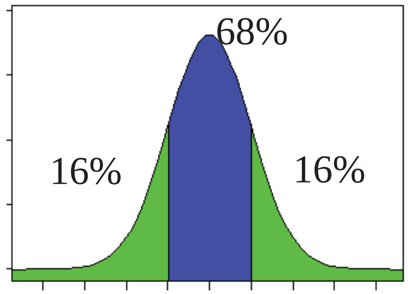
Remember : all normal distributions are equivalent – the shape stays the same, only the **units change!**

Use Picture :

16% + 68% = 84%

<u>Answer</u> : 8.0 is the 84th percentile.

That is, about 84% of DO levels observed in Bridgeport Harbor will have a value of 8.0 or less



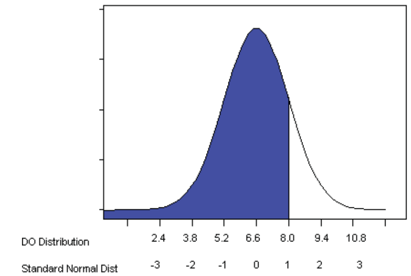
MINITAB Normal Probabilities : Use `Calc → Probability Distributions → Normal.` Fill in 6.6 for mean and 1.4 for SD. Do a cumulative probability for an input constant of 8.0

SPSS: use `Transform → Compute Variable.` In `Target Variable`, enter some variable name (like `normquant`) and then in `Numeric Expression` enter `CDF.NORMAL(8,6.6,1.4)`

```
Cumulative Distribution Function
Normal with mean = 6.60 and standard deviation = 1.40
      x     P( X <= x)
  8.0000        0.8413
```


Example *: (DO levels). What percentage of DO levels observed in Bridgeport harbor will have a value of 5.0 mg/l or less?*

Not so simple now (5.0 is not a 'nice' number of standard deviations away from the mean – i.e. not an easy use of the '68, 95, 99.7' rule)



1) Use MINITAB / SPSS - same procedure as above, just change 8.0 to 5.0

```
       x     P( X <= x)
  5.0000         0.1265
```
← about 13%

This is what you should do. However, in the old days . . . .

2) Use ***Normal Score*** *(also called the* ***z-score****, or the* ***Standardized value****). How many standard deviations below 6.6 mg/l is 5.0?*

$$\frac{5.0-6.6}{1.4} = -1.14.$$

In General : If x has a normal distribution with mean μ and standard deviation σ, then the normal score is

$$z = \frac{x-\mu}{\sigma}$$

Furthermore, $z \sim N(0,1)$ (standard normal distribution)

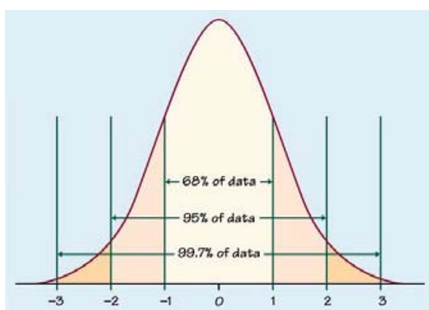
Look up the *z*-score in a Table of Standard Normal Probabilities *(Appendix of any stat book)*

So, why are we doing this!?!

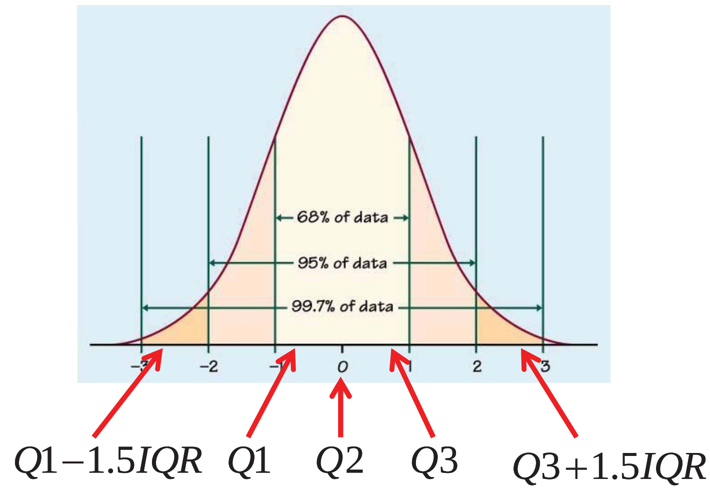
In a few weeks, we'll find that in most situations, our search for the true mean depends on understanding normal distrituions and z-scores.

Aside: Remember the 1.5*IQR rule for boxplots? For a Standard Normal Distribution:

- Q1 =-0.67
- Q3 = 0.67
- 1.5*IQR =2
- Q1-1.5*IQR = -2.7
- Q3+1.5*IQR = 2.7



That is, we're calling anything 2.7 standard deviations away from the mean an outlier! This is equivalent to keeping the middle 99.3% of the distribution – sort of like the three standard deviation rule!



Assessing Normality - Normal Quantile Plots

(Normal Probability Plots)

Our Story –

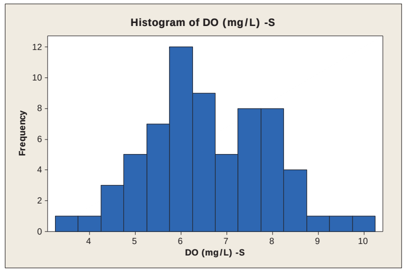
- Have some data on a variable
- Does this data come from a normal distribution?
- If not, how is it different from a normal distribution?

The idea –

- Hard to look at a histogram or a dotplot and guess if it has the right shape
- Instead, make a plot where we judge how well points fall **<u>along a straight line</u>**

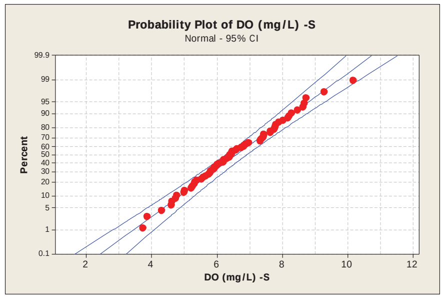
***Example** : DO in Bridgeport harbor, summer 2000.*

Is this normal?

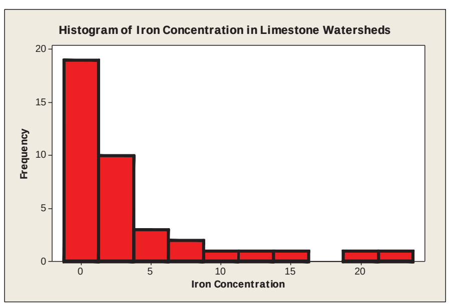


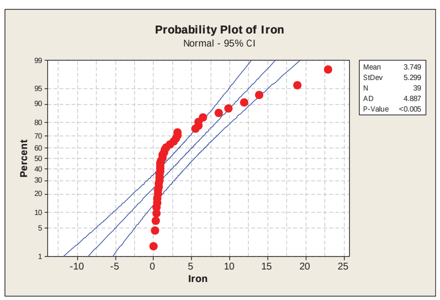
Normal Quantile Plot

Do the observations seem to fall on a straight line?



***Example** : Iron concentration levels from 39 Limestone watersheds.*





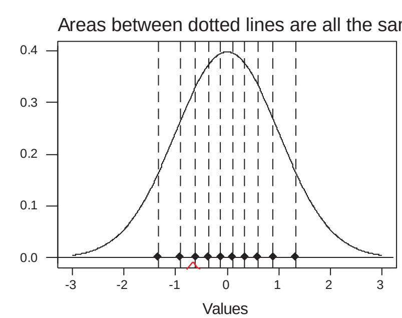
How to construct quantile plots

- Plot the observed data vs. "where we would expect them to be if they came from a (standard) Normal distribution." *(since every normal distribution has the same shape, it doesn't matter which one we use)*
- Quantile plots use percentiles of the Normal distribution.
- Roughly, plot ith largest observation vs. $(i/n)^{th}$ percentile of a N(0,1) distribution

(there are actually several ways of doing this (MINITAB and SPSS have different options), but I'll describe one way below . . .)

***Example :** Damage from hurricanes from 1991-2004 (billions of $2004) ordered from smallest to largest: {3, 4, 4, 5, 6, 7, 9, 14, 15, 44}*

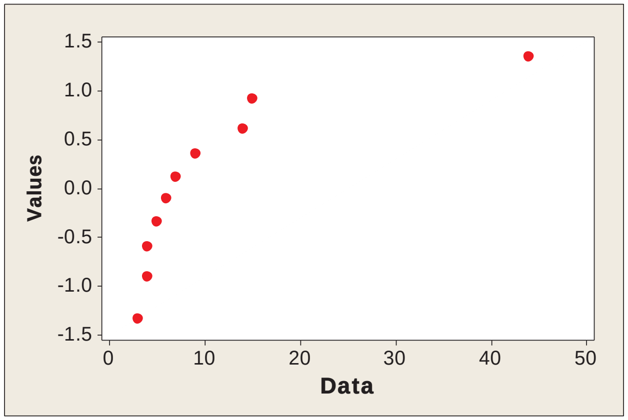
- *Divide a Standard normal distribution into **11 equal areas** (i.e. one more than the number of data points).*
- *Write down the values in a Standard normal distribution that correspond to this division (i.e. the dots in the picture below : these are the **QUANTILES!!!)***



{-1.34, -0.91, -0.6, -0.35, -0.11, 0.11, 0.35, 0.6, 0.91, 1.34}

- *Plot the* ***values*** *in the standard normal distribution vs. the* ***data***

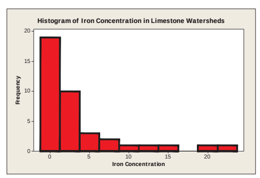
{-1.34, -0.91, -0.6, -0.35, -0.11, 0.11, 0.35, 0.6, 0.91, 1.34}



{3, 4, 4, 5, 6, 7, 9, 14, 15, 44}

Normal Quantile Plots by Hand (to get a rough idea)

In class

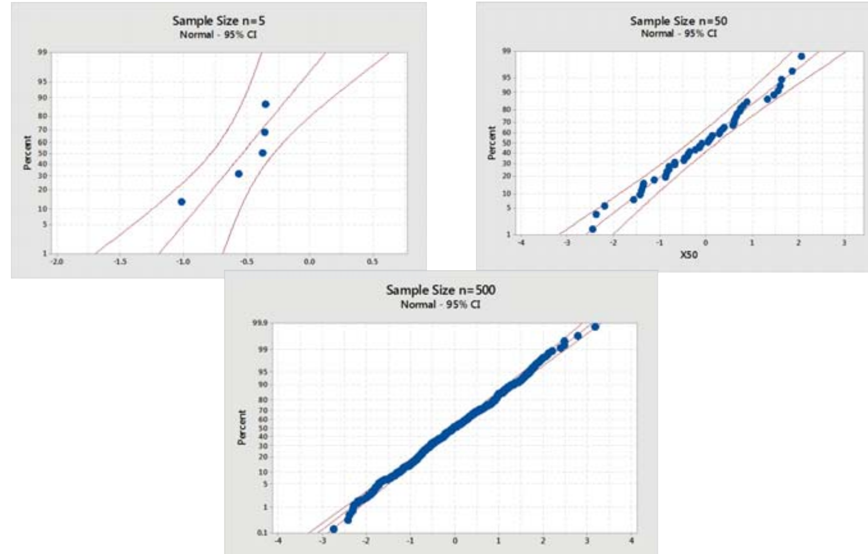


MINITAB Normal Quantile Plot : Use `Graph` → `Probability Plot` and choose `Simple.` Put variables on interest in the `Graph Variables` dialogue box.

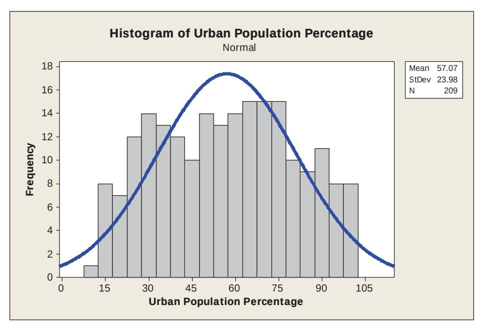
SPSS: use `Analyze` → `Descriptive Statistics` → `Q-Q Plots.` Enter variable(s). Make sure the the Test Distribution is `Normal`!

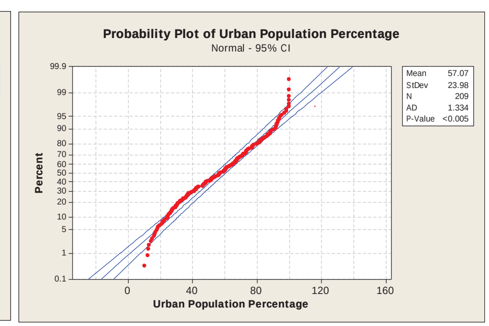
Note – The larger the dataset sample size, the straighter the line will be if the data is really normal

Example *: Data generated from a N(5,10) distribution with different sample sizes*



Example *:* ***World Poverty 2013.*** *Does data on percent population in urban areas have an approximately normal distribution? (this is called a* ***truncated*** *distribution)*





Data Relationships

Up Next: describing relationship between two **quantitative** variables :

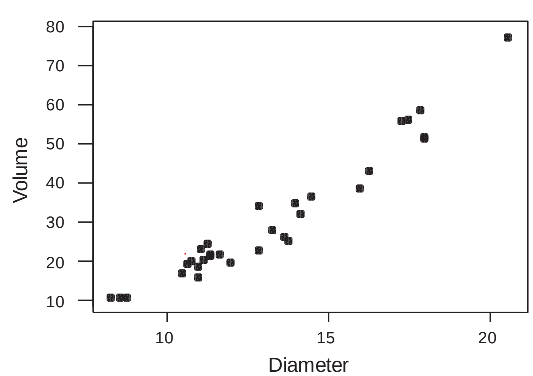
- Scatterplots
- Association and Correlation
- Regression

Visualizing Relationships : Scatterplots

- Plot two variables simultaneously
- Put one variable on horizontal axis, other variable on vertical axis
- Plot data pairs – for each observation, plot the value of one variable vs. the value of the other variable

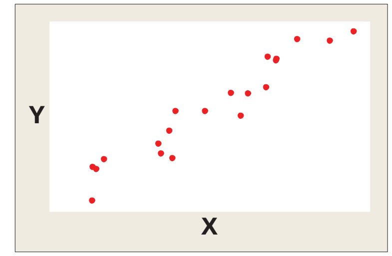
Example *: Height and Diameter of 31 Black Cherry Trees in Alleghany Forest. Here is a sample of the data and a Scatterplot :*

Diameter	Height	Volume
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7



Scatterplot Notation :

- The Horizontal axis is **ALWAYS** called the **X axis**.
- The Vertical axis is **ALWAYS** called the **Y axis.**



Scatterplot in MINITAB : `Stat → Graph → Scatterplot.`
Choose simple. Enter the variables of interest (can be more than two)

SPSS: use `Graph → Legacy Dialoges → Scatter/Dot.`
Choose simple. Enter variables (can be more than two)

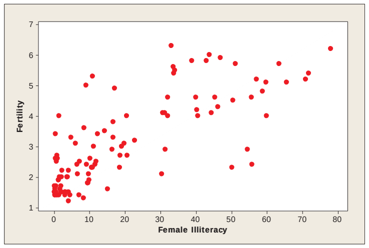
Association between Variables

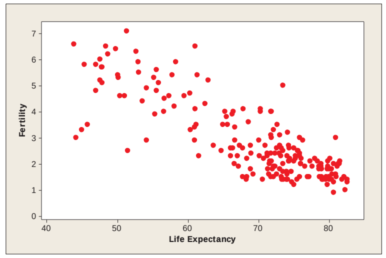
- Some values of the first variable seem associated with particular values of the second variable.
- **Does not imply linear!**

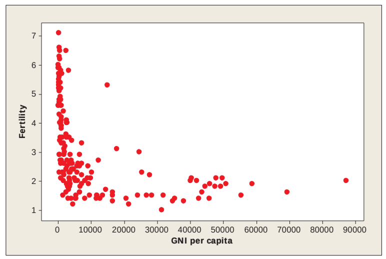
Example *: World Poverty Data : Factors associated with Fertility Rate (2008 data)*

Negative Association

Positive Association

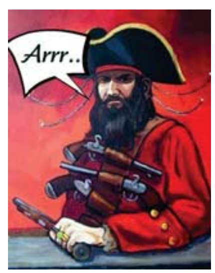






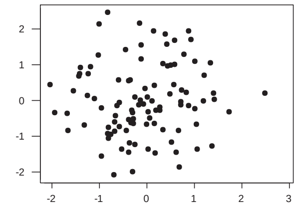
Note : Association talks about <u>direction</u> of relationship, not the <u>nature</u> of the relationship

Sample Correlation

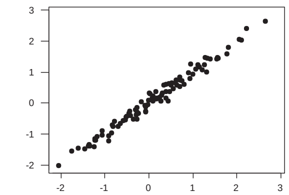


- Measures the **<u>strength of the linear relationship</u>** between two variables.
- Denoted by r
- Value is between –1 and +1
  - Zero indicates no correlation (random scatter)
  - +1 indicates all points are on a line with positive slope
  - -1 indicates all points are on a line with negative slope

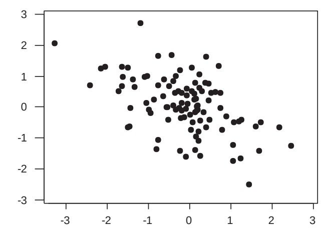
Weak **positive** correlation (near zero) $r = 0.06$



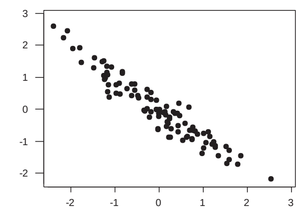
Strong **positive** correlation (near one) $r = 0.99$



Moderate **negative** correlation r = - 0.52



Strong **negative** correlation r = - 0.96



Definition of correlation

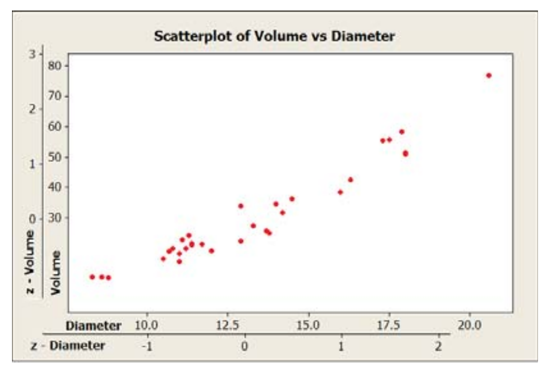
- First standardize variables: use Z-scores!

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x} \quad \text{and} \quad z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

i.e. How many SD's is each observation above or below the mean for each variable?

This is just a **change of units** – same picture!

Original and z-scores for the cherry tree data :



- Second, multiply and average :

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

Algebra shows this is the same thing :

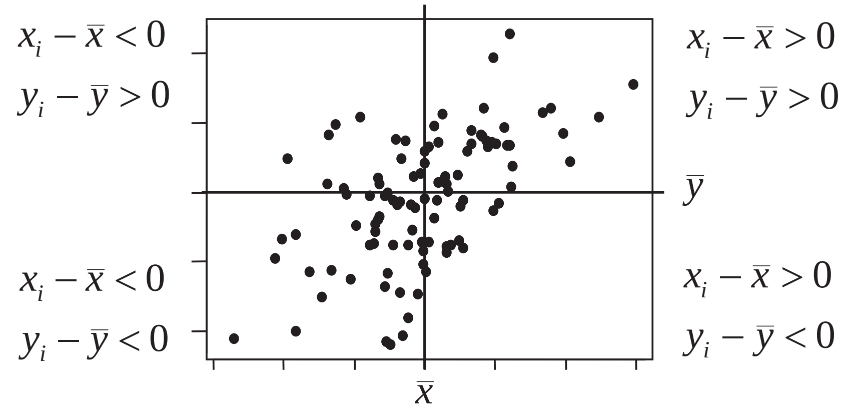
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Algebra also shows this is a **dimensionless** number **between –1 and +1** *(try this if you like!)*

Idea of Correlation

Formula $\quad r = \frac{1}{(n-1)} \sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$

What is the value of $\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$?



$$\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \qquad r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Sample Correlation in MINITAB : `Stat → Basic Statistics → Correlation.` Enter the variables of interest (can be more than two)

SPSS: use `Analyze → Correlate → Bivariate.` Enter variables (can be more than two)

Correlation is often confused with **Association.** This happens frequently in the media, and not so infrequently in journals and scientific papers!

Correlation should only be used when

- Quantifying the relationship between two **QUANTITATIVE** variables.
- The relationship between these variables is **LINEAR**
- There is no evidence of **LARGE OUTLIERS**!

Most problems can be avoided by always making a scatterplot <u>before</u> calculating correlation!

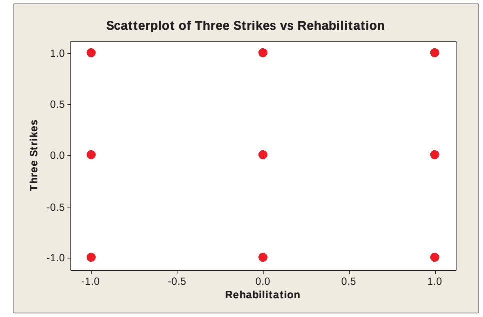
***Example** : **Crime**. A survey asked 550 people about ways of reducing crime. Respondents could answer 'Agree', 'No opinion', or 'Disagree', coded as 1, 0, -1.*

The questions were

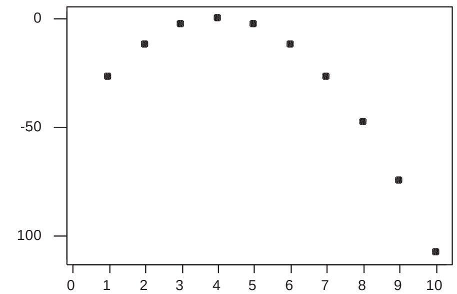
- *Violent criminals who commit three crimes should always receive a mandatory life sentence without parole.*
- *Underage violent offenders should be rehabilitated rather than incarcerated.*

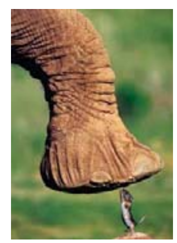
The sample correlation is r = -0.73

***However** – here is a scatterplot of the data : these variables are really **categorical** not quantitative. Correlation is NOT appropriate!*

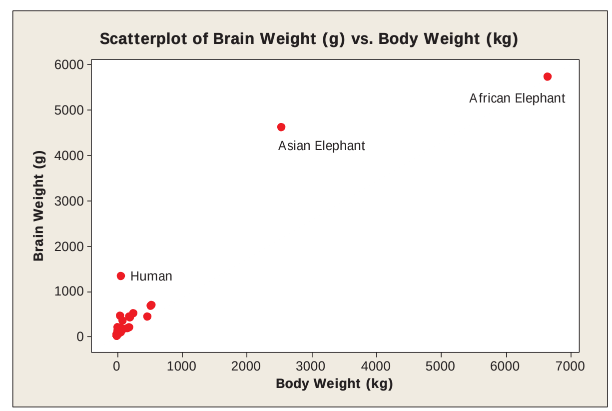


Example** : correlation is – 0.72, but correlation is clearly not appropriate!! This is negative **ASSOCIATION**, not negative **CORRELATION.





***Example** : Relationship between Body Weight and Brain Weight* (extracted from *"Sleep in Mammals: Ecological and Constitutional Correlates" by Allison, T. and Cicchetti, D. (1976)*). *This data records the average body and brain weight of 62 species of mammals. Correlation is r = 0.93 – seems great! However, here is the scatterplot :*



In this case, the elephants are Large Outliers – they are driving the large correlation. Remove elephants – correlation is only 0.651 (and then human looks like an outlier . . .)

It seems we should be able to do something about this . . .

Transformations

- Sometimes, it is necessary to transform one or more variables before calculating correlations.
- The hope is that while the original variables do not have a linear relationship, the transformed variables **WILL** have a linear relationship.

Common transformations include

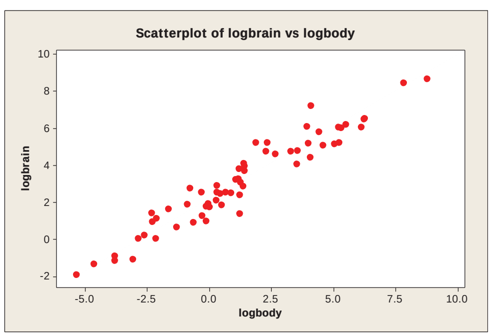
- Logarithms (natural)
- Square root
- Exponential
- Arcsine (sq rt)

Knowing what transformation to use is a matter of **experience** (that is, experience looking at data relationships and knowing what transformations work), and **knowledge of systems** (i.e. knowledge of a natural system).

Example *: Relationship between Body Weight and Brain Weight. The elephant is an outlier in both body weight and brain weight. Experience (mine) suggests trying to take the natural logs of both variables.*

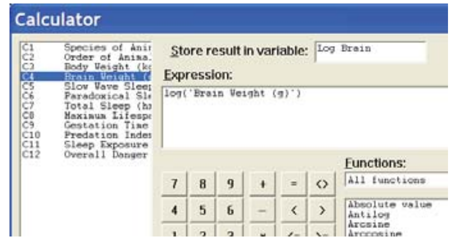
A plot of the relationship between the transformed variables :

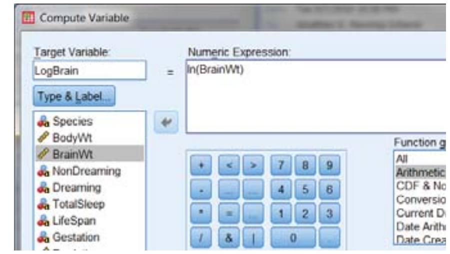
This relationship is linear ***AND*** *there are now* ***no outliers.*** *Correlation is 0.96, very strong!*



Transformations in MINITAB : `Calc → Calculator.` Make a new variable name (e.g 'Log Brain') and write the formula for the new variable in the `Expression` box.

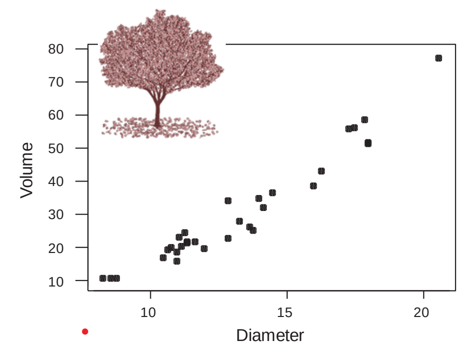
Transformations in SPSS : `Transform → Compute Variable.` Make a new variable name (e.g 'LogBrain') and write the formula for the new variable in the `Numeric Expression` box.





Sometimes, not even a picture will help you – sometimes you just have to think!

Example *: Diameter and Volume of 31 Black Cherry Trees in Allegheny Forest. Look at Scatterplot – seems mostly linear, correlation is $r = 0.967$ – quite high.*

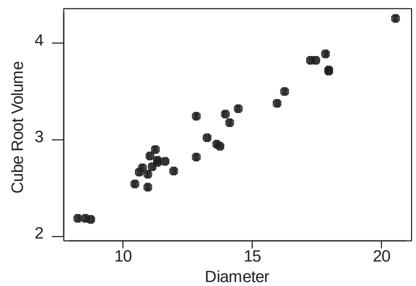


Use brain – pretend a tree is a box. Let d be the length of a side. Volume of Tree is $V = d^3$. So diameter is proportional to the cube-root of volume ($d = \sqrt[3]{V}$).

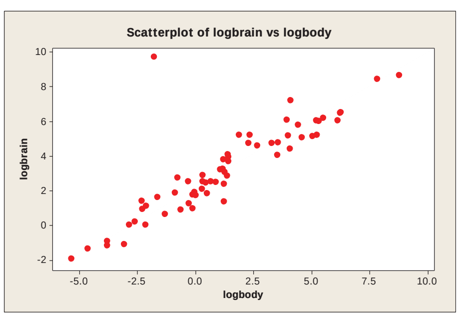
*That is, there is a **linear relationship** between the diameter and the cube-root of the volume.*

SO : Make a cube-root transformation of volume, *then compute the correlation.*

Picture is slightly more linear, and correlation improves : Correlation is now $r = 0.98$.



Let's think about the transformed brain/body weight data. Suppose we have a new mammal. We can easily measure its weight, but measuring brain weight is somewhat more difficult. It seems like we might come up with a model based on known data to relate body and brain weight. This would allow us to estimate the brain weight of the new animal without performing crude brain surgery. This idea is called . . .



Regression

Aside – at the moment, we're doing regression 'lite'. The idea is to pique your interest. In a few weeks, we'll do 'serious' regression.

The simplest case :

- Take two **quantitative / continuous** variables.
- You think one variable can be used to predict the values of the other variable.
- You think this predictive relationship would be well described by a line.
- Find the 'best' line.

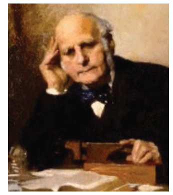
This is process is called Simple Linear Regression

Notation Convention

Explanatory Variable – the variable doing the predicting, is **ALWAYS** plotted on the **horizontal** or X axis.

Response Variable – the variable being predicted is **ALWAYS** plotted on the **vertical** or Y axis.

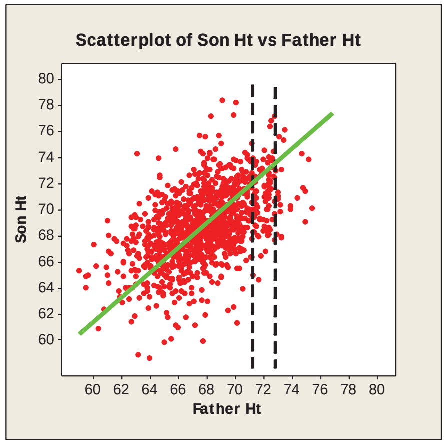
Some Regression History



Example *: Sir Francis Galton : cousin of Darwin, invented eugenics, the weather map, correlation, the idea of surveys about human communities, AND methods for classifying fingerprints!*

Galton collected data which measured the heights of fathers and sons. How do fathers' heights predict sons' heights?

Specifically, if a father is 72" tall, what's our guess of the son's height?

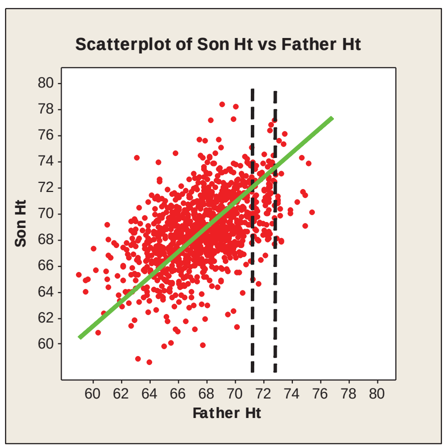


Summary statistics :

	Mean	Standard Deviation
Fathers	68"	3"
Sons	69"	3"

Correlation r = 0.5

Since 72" is 4/3 SD above the father's mean, natural guess for son is 4/3 above the sons mean, i.e. 73". This seems a bit high in the picture.



'Best' guess depends on correlation!

'Guess that son will be,
not 4/3 SD's above mean, but

correlation * 4/3 = 2/3 SD's *above mean, that is 71".*

The equation for all the best guesses :

$$\frac{\hat{y}-69}{3} = r\left(\frac{x-68}{3}\right)$$

Standardized y value *Standardized x value*

($\hat{y}$, read 'y hat', symbolizes our guess of y for a given x)

In general :

$$\frac{\hat{y}-\bar{y}}{s_y} = r\left(\frac{x-\bar{x}}{s_x}\right)$$

Rearranging, we get

The equation of the Least Squares Linear Regression Line

$$\hat{y} = b_0 + b_1 x$$

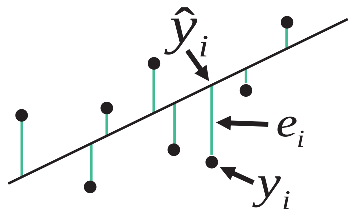
with slope $b_1 = r\dfrac{s_y}{s_x}$ and intercept $b_0 = \bar{y} - b_1\bar{x}$

Let's Review : Least Squares Linear Regression

- There frequently exists a linear relationship between two variables.
- If this linear relationship exists, use the value of one variable to predict the value of another
- Use regression to find the 'best' line that describes this relationship

What's 'best'?

For each data point we observe, look at the **vertical** difference between the observed Y value (y_i) and the corresponding point on the regression line ($\hat{y}_i$ - the **fitted value**). This difference is called the **residual** or **error**: $e_i = y_i - \hat{y}_i$



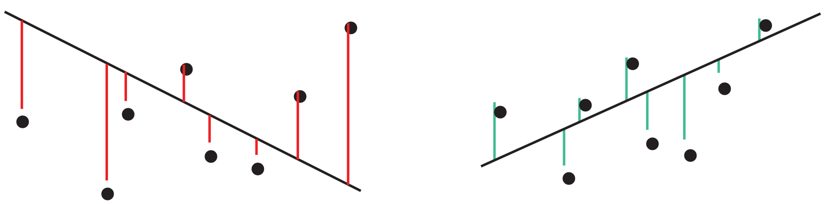
So what's 'Best'?

The regression line minimizes the sum of the squared residuals :

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Bad Fit

Better Fit

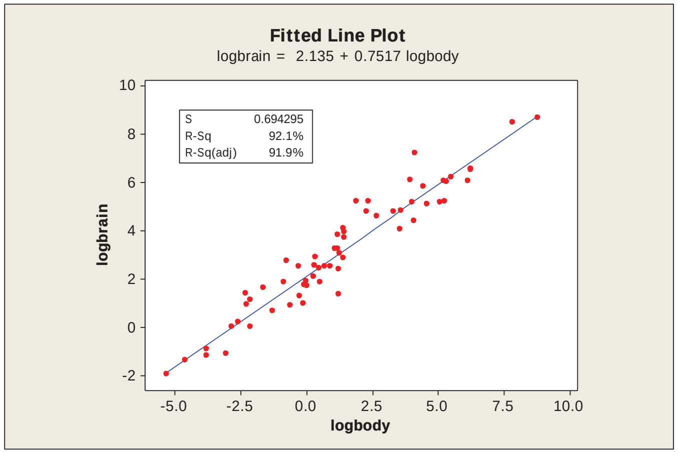


***Example** : Brain/Body weight relationship. Let's use regression to find a 'best' line that predicts log brain weight based on log body weight.*

Regression in MINITAB – use `Stat → Regression → Fitted Line Plot.` Later, we'll use `Stat → Regression → Regression.` This gives lots of information we haven't discussed yet.

SPSS: for now, use `Analyze → Regression → Curve Estimation.` Enter variables and make sure that under models you've checked `Linear.` Later, you'll use `Analyze → Regression → Linear.`

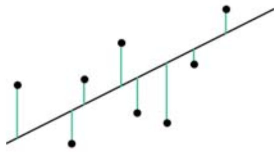
MINITAB produces a plot that is a scatterplot of the data, the 'best' fitted line, and the **estimated** regression equation (slope and intercept) calculated by minimizing the sum of the squared residuals.

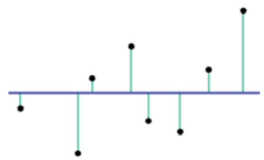


Relationship of Regression to Correlation :

Notice that $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad < \quad \sum_{i=1}^{n}(y_i-\bar{y})^2$

Sum of Squared Residuals < Sum of Squared Deviations





In fact, algebra shows that

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

In Words : Variance of y's around fitted values + Variance of fitted values (around mean) = Variance of y's.

NOW : plugging in the definition of correlation, more algebra shows

$$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-r^2$$

which is equivalent to

$$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=r^2$$

How far is fitted line from mean line

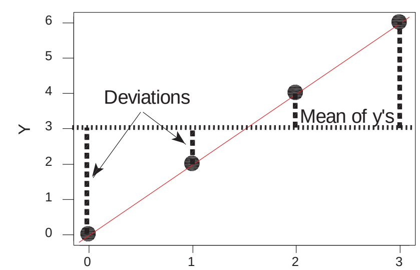
How far are y data values from mean

Correlation measures the improvement of a sloped line to a flat-line. In other words (take home message – forget formulas)

r^2 measures the proportion of the variance of the y's explained by the regression

Example : *Consider 4 points (x,y) : (0,0) (1,2) (2,4) (3,6).* $\bar{y}$ = 3. *Correlation is 1, i.e.*

$r^2=1$

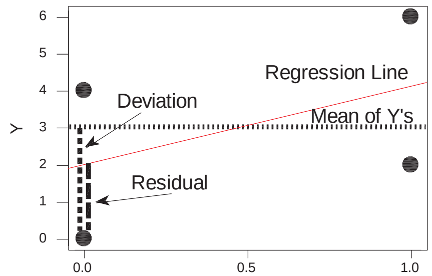


Now :

- *Variance(Y) = 6.6 (based on DEVIATIONS)*
- *Regression residuals are all 0, i.e. Variance (residuals) = 0*
- *SO : Regression explains all of variation in the Y's*

Example : *Consider 4 points*
(x,y) : (0,0) (0,4) (1,2) (1,6)
(same y's!!!!). $\bar{y} = 3$.

Correlation is 0.45, i.e. $r^2 = 0.2$



Now :

- *Variance(Y) = 6.6*
- *Variance (residuals) = 5.3*

SO : Amount of variation of y's explained by regression is $\frac{6.6 - 5.3}{6.6} = .2$

Regression – when things go bad

(Not bad really, but things that can be problematic . . .)

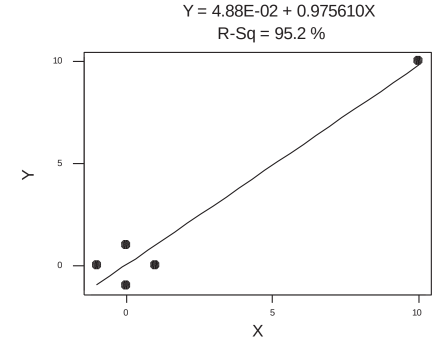
- Least-squares regression is not robust (resistant) to 'unusual' points.

- Two kinds of interesting points:
  - o **Outlier** : a point with a <u>large residual.</u> Note this is **<u>NOT</u> the same thing as being an outlier from the data** *(for example, a point might be an outlier in the X and/or Y direction, but might not have a large residual!)*
  - o **Influential point** : if removed, causes a large change in the regression line. These points often (but not always) have large X values.

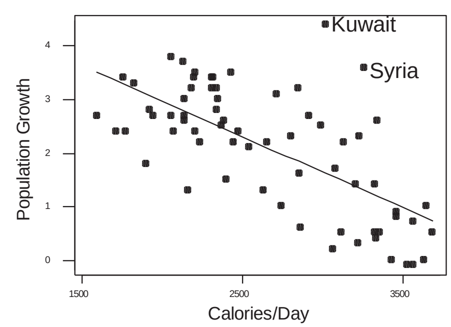
A point can be influential, an outlier, or BOTH!

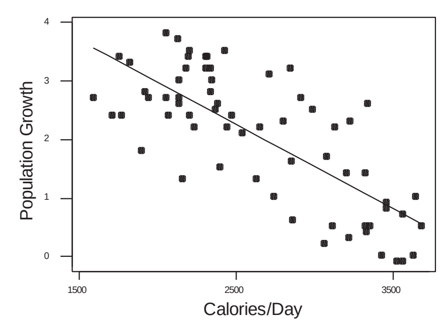
Example : *The point (10, 10) is . . .? (what happens without this point?)*

X	Y
1	0
0	1
-1	0
0	-1
10	10



Example : Poverty Data *(year 2000 data). Try predicting population growth rate (percent growth) with per capita calories per day. Two 'unusual points'. These are (outliers, influential?). Notice that the regression line does not move much when they are removed.*





Evaluating Regression models : Residual Plots

Here is our simple linear regression model : this describes the **predicted** value of y, denoted by y 'hat'.

$$\hat{y} = b_0 + b_1 x$$

The more complete model is that **y is a linear function of x, with added errors (denoted by ε_i)**

$$y = b_0 + b_1 x + \varepsilon_i$$

For reasons we'll discuss later, we assume the errors come from a normal distribution with mean zero and some standard deviation sigma – in 'stat' notation :

$$\varepsilon_i \sim N(0, \sigma)$$

SO : what does this mean?

Simple Linear Least-Squares regression assumes that

- The relationship between explanatory (X) and response variables(Y) variables is **linear.**
- The explanatory variable ‘explains’ **all of the predictable variation** in the Y’s
- The residuals have **no discernible pattern** – they should be random noise due a variety of sources.
- In fact, it is assumed that the **residuals have a Random Normal Distribution.**

To test these assumptions, we look at **residual plots**

- To see if residuals have a normal distribution, make a **Normal Quantile** (or Probability) plot of the residuals.
- To see if there are discernible patterns in the residuals, make a plot of the **fitted values** ($\hat{y}$: on the horizontal axis) vs. the residuals (vertical axis). **There should be no discernable patterns in this plot!**

Essentially, this plot removes the linear trend from your data and displays any trends that remain!

Residual Plots in MINITAB : use `Stat → Regression → Fitted Line Plot.` Click on the `Graphs` button, choose `Normal Plot of Residuals` and `Residuals versus Fits.`

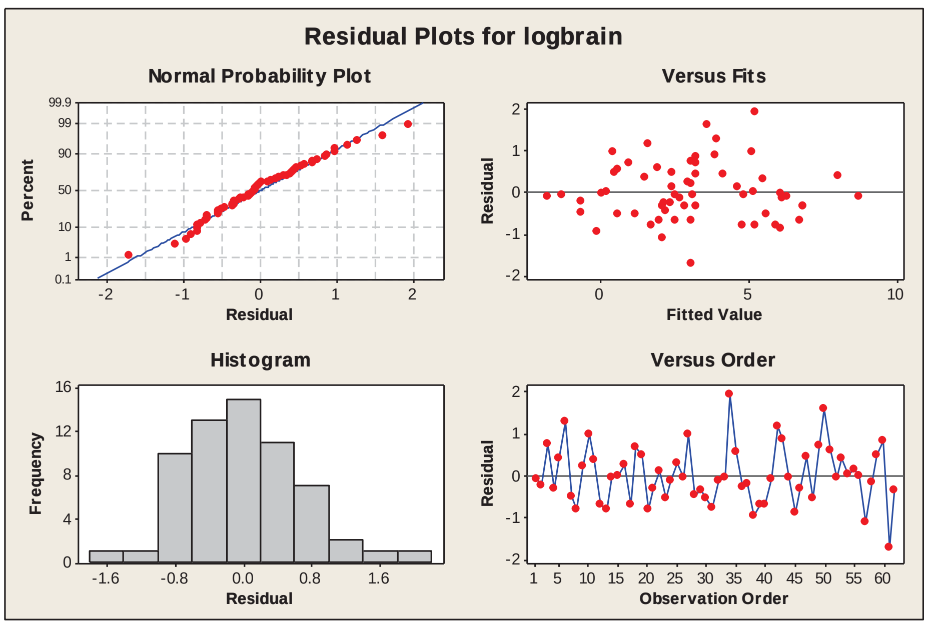
SPSS: use `Analyze → Regression → Linear.` Enter dependent and independent variables. Click on `Plots`. Choose `Normal probability plot`, and then under Y enter `ZRESID` and for X put `DEPENDNT.`

Example *: Brain/Body weight relationship in mammals. We used regression to predict log(brain wt) based on log(body st).*

Here are plots of residuals :

A normal quantile plot of the residuals is approximately linear – this is good!

A plot of residuals vs. fitted values shows no discernible pattern, and there are no obvious outliers. This is good!

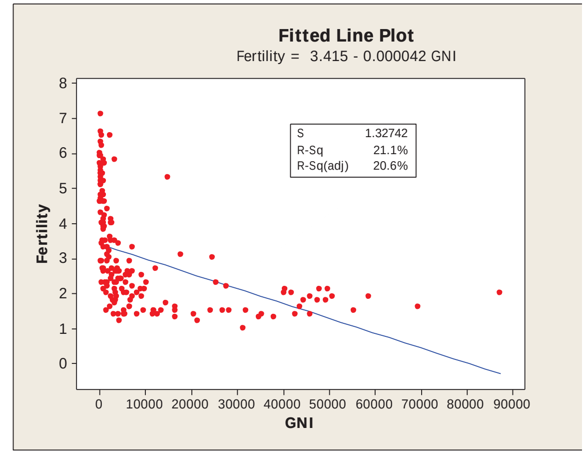


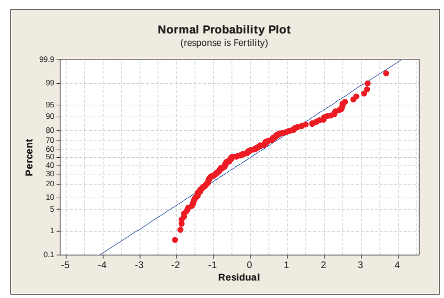
Example – Poverty Data. *Try to use GNI per capita to predict fertility rates.*

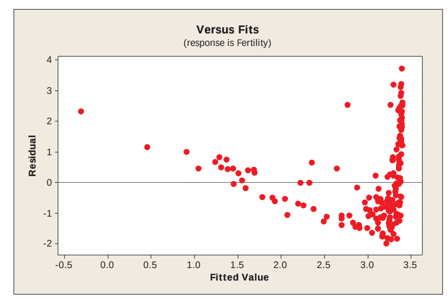
Here is the fitted line plot : <u>NOT</u> linear.

Here are residual plots :

Residuals are somewhat curved, but plot of fitted vs. residuals shows a definite pattern!

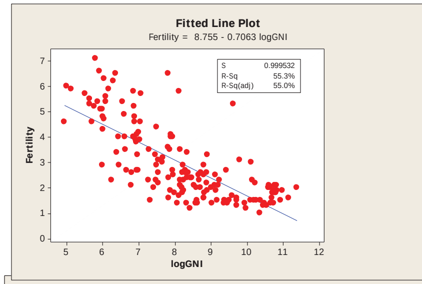


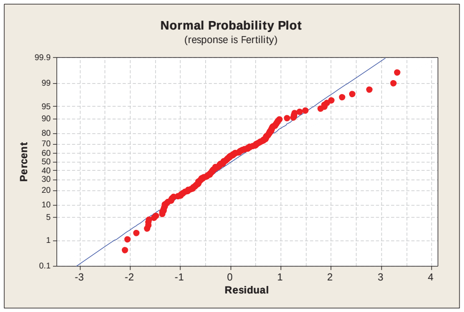


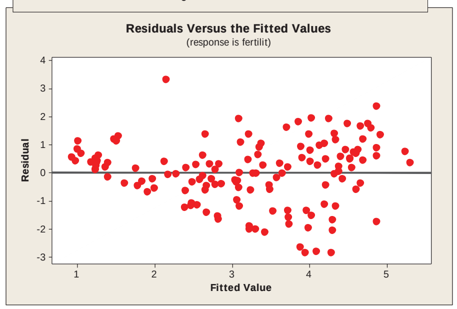


So : what to do – try a *Transformation – take logs of GNIPC.*
Fit model again.

Shows improvement – notice that relationship is more linear, R^2 increases (i.e. more of overall variation is explained by the model). However, might still require further tweaking.







How Big should R^2 be?

- R^2 measures the amount of variability in the response variable explained by the regression model.
- Must be between 0 and 1 (since it's just the correlation squared).

There is no threshold value for what constitutes good vs bad R^2 - a 'good' fit depends on the situation. When modeling chemistry relationships, might expect an R^2 of 0.99. When modeling social relationships, might be happy with an R^2 of 0.12.

Lurking Variables in Regression (things hidden . .)

A variable that has an important effect but was overlooked.

DANGER – Confounding! This is when we think an effect is due to one variable, but it is really due to another, lurking variable.

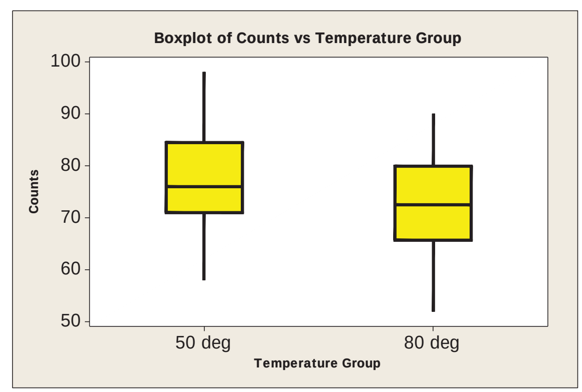
Example *: A 1970 study showed coffee drinkers had a higher incidence of bladder cancer. However, a 1993 study showed that, if you also considered smoking, there was no evidence of a link between coffee and bladder cancer. (i.e. people who drink lots of coffee also smoke!).*

Example *: There is strong association between GNP/capita and Fertility Rates. This does **not** mean that getting paid less CAUSES women to have more babies!*

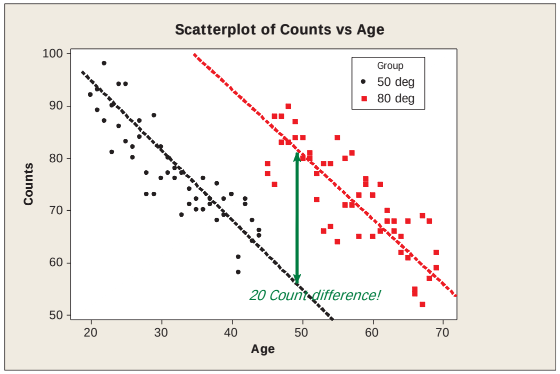
Lurking variables can actually cause a reversal in the apparent magnitude of an effect

Example *: Examine how many times people can click a counter at two different temperatures. At what temperature do people get more counts?*

Boxplot of Results : 50 deg seems to have higher counts than 80 degrees.



However *: look at plot of counts by temperature group with* ***age of subject included*** *(age is the* ***covariate****)*



50 deg actually has lower counts 80 deg. BUT Age has a strong impact on Response Time, and the average age is quite different in each treatment group ($\bar{x}_{1.} = 32$, $\bar{x}_{2.} = 57$)*, so* ***age effect becomes confounded with temperature effect****.*

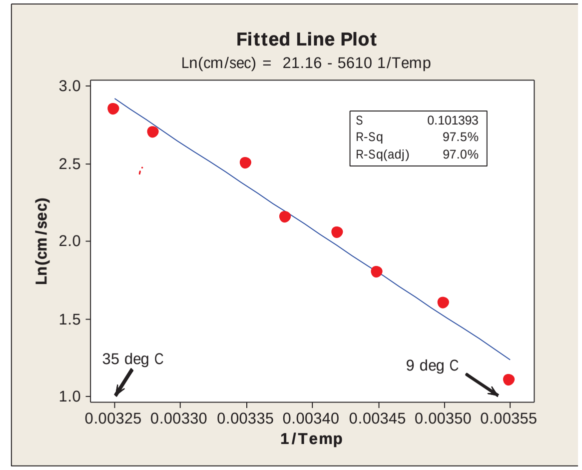
One Last Regression Warning : Regression estimates are valid **ONLY** over the region of explanatory variables **where you have data!**

Example *: Speed of Ants.* *(see http://pubs.acs.org/doi/pdf/10.1021/ed077p183)* *Ants are cold-blooded and their speed is temperature dependent. Experiments have shown that in fact ant speed can be modeled according to known facts about the rate of chemical reactions :*

$$\ln(Speed) = b_0 + b_1\left(\frac{1}{Temp}\right)$$

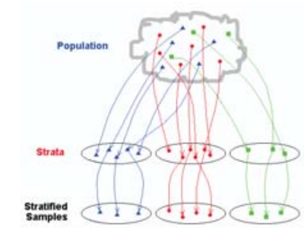
That is, the natural log of ant speed is linearly related to the inverse of the Temperature.

Here is a plot of some experimental data. Now – think about what happens as temperature increases or decreases – is the linear estimate still valid?



Sampling and Experiments

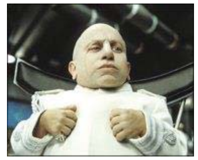
Research Design : Some General Advice



- **Decide What you Want to Know :**
  - Explicitly define the parameters of your study.
  - Make sure the things you measure will answer the question you're asking.
  - Have an analysis plan BEFORE you begin!

- **Design Carefully** : Consider
  - Feasibility of obtaining measurements
  - Cost and time associated with obtaining measurements
  - The number of measurements needed to obtain ‘good’ results (sample size calculation if possible)
  - Possible sources of variability, bias, error

- **Run a pilot study :** Mini- study to test your procedures and identify problems
  - How long will data collection take?
  - Will your measurement procedure work?
  - Will people answer your survey?
  - Will people lie?
  - Will your plants die?



- **If in doubt, consult with a Statistician :** you can save immense heartache, loss of resources, etc. by checking with an expert first!

Ronald Fisher (1890 - 1962)

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of. (Indian Statistical Conference, 1938).

In 1926, Fisher Published ‘Arrangement of Field Experiments” which outlined <u>three components for successful studies:</u>

Local Control, Replication, Randomization

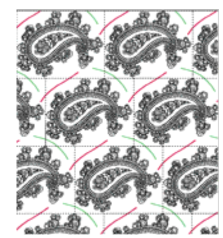
Local Control

- Methods used to control experimental / observational error, increase accuracy of observations, and allow for inference regarding treatment factors in an experiment.

Examples of Local Control :

- Have a placebo group to use as an experimental baseline
- For oral surveys, ask questions the same way every time
- Make sure your equipment is properly calibrated (thermometers, scales, etc.)
- Make ground plots the same size and soil composition

Replication



- Implies measuring the same treatment levels on several independent units to estimate error variance.
- Increases the precision of statistical estimates
- Allows us to distinguish between true relationships and chance occurrence

Randomization

- Prevents Bias (researcher or system)
- Fisher (1935) showed that randomization provides the proper reference population for statistical inference using normal theory *(more on this later . . .)*

Example : Scurvy

- *Disease caused by vitamin C deficiency*
- *Causes the bodies connective tissues to degenerate (teeth fall out, wounds open).*
- *Killed an estimated 2 million sailors between 1500 and 1800 – number one cause of death among sailors in this period*

One of the first experiments (every) was conducted by Dr. James Lind who in 1753 published 'A Treatise of the Scurvy".

Here's an excerpt - look for local control, replication, randomization :

On the 20th May, 1747, I took twelve patients in the scurvy on board the *Salisbury* at sea. Their cases were as similar as I could have them. . . . They lay together in one place . . and had one diet in common to all. Two of these were ordered each a quart of cyder a day. Two others took elixir vitriol three times a day. Two others took two spoonfuls of vinegar three times a day upon an empty stomach. Two of the worst patients, with the tendons in the ham rigid (a symptom none the rest had) were put under a course of sea water. Two others had each two oranges and one lemon given them every day. The two remaining patients took the bigness of a nutmeg three times a day of an electuary recommended by an hospital surgeon.

The consequence was that the most sudden and visible good effects were perceived from the use of the oranges and lemons; one of those who had taken them being at the end of six days fit four duty. The other was the best recovered of any in his condition, and being now deemed pretty well was appointed nurse to the rest of the sick.'

This study exhibits

- ***Local Control*** *(common diet, similar patients, control group)*
- ***Randomization*** *(some problem here . . .)*
- ***Replication*** *(2 per treatment group, comparative treatment groups)*

Randomization

Randomization is the process by which

- Experimental units are assigned to treatment groups
- Observational units are selected

Why Randomize?

- Randomization prevents **BIAS**
  We say a study is **biased** if it favors certain outcomes

Examples :

- *You want to compare treatments for dealing with woolly adelgid on hemlock trees. You assign the most severe trees to get pruning and spraying while the less severe trees get spraying only. You compare one year survival rates.*
- *You need to buy a new toaster. You go to epinion.com and see what other consumers liked.*

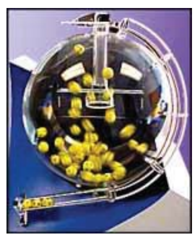
How do we Randomize?

Worst

- Machines (drums of paper, numbers in a hat, lotto balls)
- Computers – most common way to achieve randomization.
- Tables of random digits

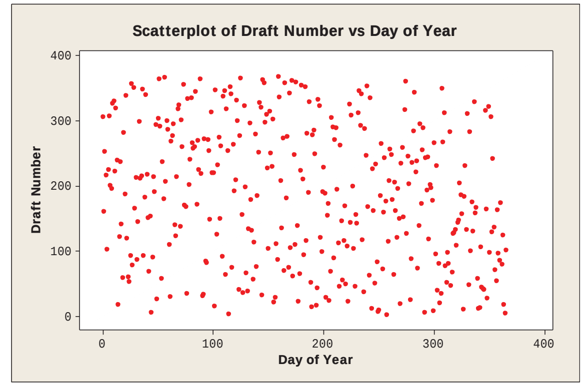
Best

True Randomization is difficult to achieve!



Machines – often cool, but not always random!

Example : Randomization Failure in the 1969 Draft. *During the Vietnam War, men over 18 year old were drafted in groups of identical birthdays. The Selective Service needed to assign a random draft order to days of the year. To do this, 366 capsules were put into a shoebox (??) and then dumped into a glass jar. Each capsule contained a piece of paper containing one birthday of the year. The head of the Selective Service pulled out capsules one at a time, assigning each successive capsule the next draft number (i.e. September 14 got drafted first, April 24 was drafted second, etc). Here is a plot of draft order vs. day of year : Does it seem random?*



Turned out that the correlation of Day of Year *and* Draft Number *is -0.22, a* ***significantly negative correlation.*** *Process was NOT random (challenged in court, but judge ruled that was 'random enough'.)*
http://en.wikipedia.org/wiki/Draft_lottery_(1969)

Computer Randomization

- Most common way to achieve randomization.
- Really pseudo-randomization (uses an algorithm, so it is necessary a predictable process)
- Adequate for most randomization needs (this is what pharmaceutical companies use, for example)

There are **many** ways to do this – the method depends on the situation.

Example *(in MINITAB) : Suppose you want to :*

- *Give a survey to 100 people.*
- *There are 3 versions of the survey (A, B, and C)*
- *40 people should receive survey A, and 30 people each should receive surveys B and C.*

Solution : in MINITAB

1. Use `Calc → Make Patterned Data → Simple Set of Numbers.` Choose numbers from 1 to 100 in steps of 1, store in a new column.
2. Use `Calc → Random Data → Choose from Columns.` Sample from the column above WITHOUT replacement. Save in a new column. This gives a random ordering of the original column.
3. Assign the first 40 people in the new column to receive survey A, the next 30 receive survey B, the final 30 receive survey C.

Tables of random digits

- These are the most accurate method of randomization
- Generated by measuring radioactive decay times or following other natural, random processes
- Available in stat books or online
  http://www.nist.gov/pml/wmd/pubs/upload/AppenB-HB133-05-Z.pdf

Example *: Previous survey of three types given to 100 people. First, we give people numbers from 0 to 99 rather than 1 to 100.*

Random Digit Table	Person
00	1
01	2
. . . .	. . . .
99	100

The first row of a random digit table –

19223 95034 05756 28713 96299 07196 98642

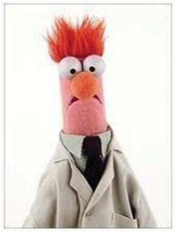
Assign survey type A first : give to persons 19, 22, 39, 50, 34, 5, 75, 62, 87, 13, 96, 29, 90, 71, (ignore 96 – repeat), 98, 64, etc., until 40 people are assigned. Then continue to assign people to survey types B and C.

19223 95034 05756 28713 96299 071~~96~~ 98642

What if we only want to assign 30 people, ten to each survey type?

Use the same process, ignore all numbers above 29.

Experiments vs. Observational Studies



Experiments :

- Deliberately vary factors in order to see what happens
- Can often assign causal relationships between responses and treatments (i.e. cause and effect)
- Can be balanced when there are multiple factors to ensure the effects can be separated to establish causation.
- Usually performed under controlled conditions (often in a lab or an experimental plot).
- Randomize individuals (people, plants) to particular treatment groups (which can include doing nothing)
- Are often very expensive or highly immoral to perform

Observational Studies :

- Simply observe people or situations with levels of various factors to draw inferences
- Cannot establish causal relationships between responses and treatments.
- Are performed in situations where we would like to perform an experiment
- Are prone to **BIAS.**
- Are often the main study type available to researchers.

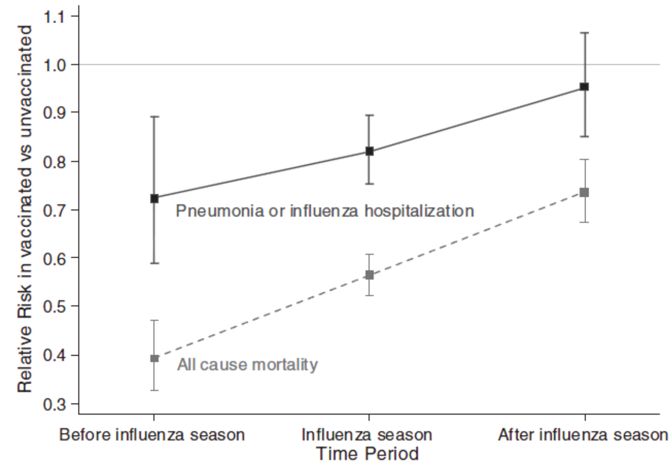
Example : Smoking and lung cancer . . . . .

http://www.economist.com/node/10952815

Example : Flu Vaccines. *Numerous observational studies have shown a 50% lower risk of dying among vaccinated persons vs. nonvaccinated people. In each case, studies identify self-selecting groups of vaccinated and unvaccinated people.*

However, a study from 2006 looked at individuals ***BEFORE*** *and* ***AFTER*** *they had the flu shot! Note that a Relative Risk of 1 indicates no difference between groups. A Relative Risk of less than one indicates that vaccinated people are less likely to have a particular outcome (death / pneumonia). i.e. a Relative Risk of 0.333 means Vaccinated people are 1/3 as likely to die relatively to unvaccinated people. This study suggests the likelihood of dying/getting pneumonia is greatest* ***BEFORE*** *they ever receive a vaccine – i.e. healthy people are getting the vaccine!*

Journal of Epidemiology 2006;35:337–344 Evidence of bias in estimates of influenza vaccine effectiveness in seniors Lisa A Jackson,1,2* Michael L Jackson,1,2 Jennifer C Nelson,1,3 Kathleen M Neuzil4 and Noel S Weiss2



Sampling Designs

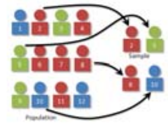
(Several courses around campus are offered in Sampling Design (FES) and Survey Design (Political Science)

Why Sample ?

- Hope to learn about a population.
- Usually unfeasible to sample EVERY INDIVIDUAL from a population.
- Try to take a sample that reflects the population, and make inferences about the population based on the sample.

Aside – incidentally, a sample of EVERY individual from a population is called a **census**. A census is often less accurate than a sample since it almost always misses some members of the population (the **undercount**!)

It is tempting to try to make a sample 'representative'. However, it is usually best just to take a



Simple Random Sample : <u>Unbiased and Independent</u>

<u>Unbiased :</u> Every sample of size *n* has equal chance of being selected.

<u>Independent :</u> selection of one unit has no influence on the selection of other units

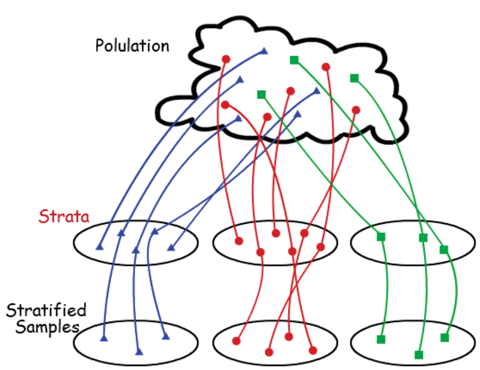
A simple random sample requires a

Sampling Frame : a list of every possible unit (person/tree/day) from which a simple random sample can be made.

Example *: Measure attitudes about the requirement of taking an intro stats course among undergraduates. The sampling frame would be provided by the . .*

You may want to insure that certain sub-populations are represented in your sample. In this case, use a

Stratified Sample : First <u>stratify</u> the population into known homogenous sub-groups and then sample within subgroups.



Examples : *Candidate preferences : might want to see how political views change according to gender, age, ethnicity. Stratify by these variables first, then sample within each sub-group. Or for forest measurements, stratify by species (say maple, oak, birch), then sample within each species.*

Note – to make population level inferences, you need to sample according to the sub-population sizes.

Example : *In a class with 60% women, if you stratify by gender and then sample 100 individuals, you need to sample 60 women and 40 men in order to make accurate inferences about the entire class!*

If goal is to be **COMPARATIVE**, choose equal sample sizes in each strata

(i.e. 50 women, 50 men)

If goal is to be **REPRESENTATIVE** of the entire population, choose strata sample sizes according their prevalence in the entire population

(i.e. 60 women, 40 men)

Simple Random Sample vs. Stratified Sample

- For both a Simple Random Sample and a Stratified Random Sample (if chosen proportionately), EACH individual is equally likely to be chosen.
- However, the requirement for a Simple Random Sample is a bit more stringent – each sample of size n is equally likely to be chosen.

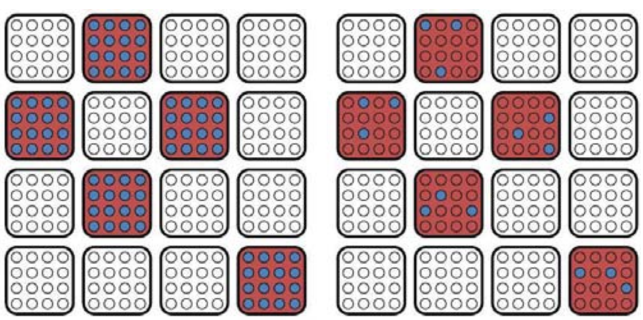
Example : Population with 5 women, 5 men. Choose a sample of size 4, SRS [any 4] vs. Stratified [2 women, 2 men].

F M
F M
F M
F M
F M

Another problem that often arises in sampling is that individuals are too far apart, or that there is no complete list of individuals. In this case, use

Cluster Sampling :

- Subdivide population into clusters
- Sample from within each cluster OR sample all individuals inside chosen clusters.



***Example** : **New Haven household survey.** Instead of driving all over New Haven, divide the city into blocks, choose a sub-sample of blocks, and then sample individuals **ONLY** within the sub-sample of blocks.*

***Example : Forest Surveys** – divide forest into blocks, choose a few blocks at random, sample trees ONLY inside chosen blocks. Similarly, choose a few random starting locations and then use a TRANSECT at random starting locations.*

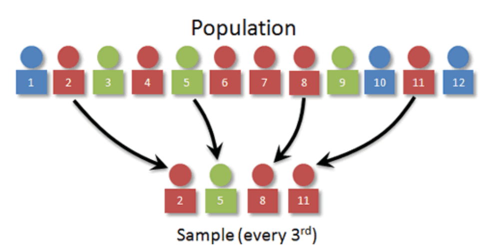
***Example : Class Surveys** – choose a few random Yale classes, then survey students ONLY from the chosen classes.*

Stratified Sample vs. Clustered Sample

Example . . ..

Sometimes, the total population size is unknown, but can be observed at regular intervals. In this case, use a

Systematic Sample : take every *i*th unit that comes along (every 10th person to leave a grocery store, every 100th item in a production line, every 10th day in a year).



Finally, you can mix-and-match sampling techniques to achieve a

Multi-Stage Sample : **ANY combination of the sampling techniques mentioned above**

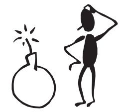
- **Stratified – Clustered**
- **Clustered – Stratified**
- **Stratified – Systematic**
- **Etc.**

Example : A Stratified, Clustered survey *(Cheryl Margoluis, FES). Cheryl conducted a survey to measure relationship between land tenure (stability) and land. These were conducted on villages in a protected region in Guatemala.*

Villages were placed into one of three strata :

- *Had land title*
- *Had limited agreement with government*
- *No land agreement (i.e. squatters)*

There were 4-8 villages in each strata. Two villages were chosen from each strata and ALL members of each village were sampled (this is the clustered part). This was necessary because of opportunity sampling otherwise created by social structures (i.e., any time a survey would be conducted, those standing around would also want to be sampled!)



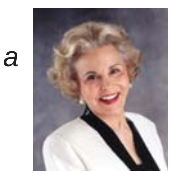
PITFALLS IN SAMPLING

- **Response Bias** : the interviewer treats groups differently.
- **Undercoverage** : subjects/units are left out of a sample by design

Example : *people without phones in phone surveys*
Example : *big fish in deep waters in animal capture/recapture surveys*

- **Non-response and Volunteer Sampling** : people who don't respond may well represent an opinion/state of nature that is otherwise not represented. Similarly, people who volunteer to be selected probably represent a common viewpoint

Example : Ann Landers.



http://www.stats.uwo.ca/faculty/bellhouse/stat353annlanders.pdf In a 1976 survey, Ann Landers asked her (primarily women) readers 'If you had to do it again, would you have children?" Of the 10,000 responses she received, 70% said 'NO!' Shortly afterward, Good Housekeeping asked their readers the same question. An astonishing 95% of readers responded 'YES!'.

This is an excellent example of bias caused by

- Volunteer sampling
- Different populations (of readers)
- Different question contexts
  - *Ann Landers question was preceded by letter from woman detailing many couples who were happier without children*
  - *Good Housekeeping question was preceded by request to readers to essentially repudiate Ann Landers survey results*

- **Measurement Error** – caused by machine, human, data entry, bad luck. Best way to avoid is to do a PILOT STUDY – a practice study to work out the bugs, calculate sample size, and make sure study is feasible!

- **Survey Design** – a few of the dozens of things that can go wrong :
  - People don't respond
  - Survey is to long
  - Questions are biased – *"Wouldn't you agree that statistics is the most useful course you've ever taken?"*
  - Questions are unclear / poorly worded

***Survey Design** – lots to say here, you can take another entire class on this subject . . .*

http://lap.umd.edu/survey_design/index.html

DESIGN of EXPERIMENTS

Semester long classes are offered in Experimental Design and Analysis in several schools around campus.

A Few Definitions :

- **Comparative Experiment** : implies that an experiment will compare two or more sets of circumstances to draw inferences
- **Treatments** (Treatment Factors) : any substance or item or procedure whose effect on the data is to be studied
- **Levels** : the particular types or amounts or levels taken on by the Treatment Fact ors.
- **Experimental Unit** : the material to which levels of the treatment factor(s) are applied. THIS IS SOMETIMES HARD TO DETERMINE. *We'll discuss in class.*
- **Experimental error** : the variation among identically and independently treated experimental units

Recall the principles of good study design :

Randomization, Local Control, Replication

Randomization

- Make sure that experimental units are assigned randomly to treatment groups.
- Randomization helps prevents **bias!**

Experimenter bias (i.e. you or your assistants' bias).

- *You give the growth factor to the healthier plants*
- *You tell the placebo patients "You're on placebo"*
- *Your assistant knows what outcome you want and makes up the results (Fisher suggested this happened to Mendel, although recent research suggests otherwise)* http://www.genetics.org/cgi/content/full/175/3/975
- *Bias is often unconscious – see recent study on teachers' gender bias in math grading* http://www.nber.org/papers/w20909

- Part of eliminating experimenter bias is to make a study **blinded** :
  - If the subject is unaware of the treatment they receive, a study is called **single blinded.**
  - If both the subject and the experimenter are unaware of the treatment being applied, a study is called **double blinded.**

Systematic Bias

- Field experiments : one treatment is applied to only the upper portion of a field. Soil type changes across the field (also edge effect)

- Time experiments : conditions change from day to day. Also, if subjects receive multiple treatments, there may be an accumulation affect if treatment order is not random

Local Control for Experiments

- Describes methods used to control experimental error, increase accuracy of observations, and allow for inference regarding treatment factors.
- Includes things like
  - Placebo groups – are affects due to thought of being treated
  - Measurement accuracy – scale calibration, consistency of research assistants, etc.
  - Making treatment groups as similar as possible to control for confounding factors
  - May require additional design to reduce between subject variability (matching or blocking)

Replication

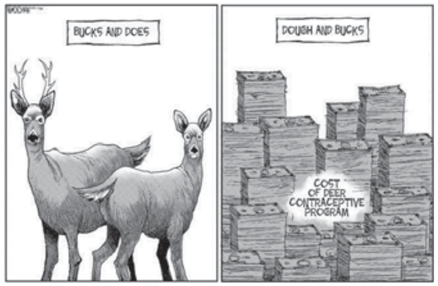
- Implies measuring the same treatment levels on several independent units to estimate the experimental error variance.
- Ensures results should be reproducible
- Allows us to distinguish real effects from random chance
- Requires **sample size calculations!**

The experimental counterpart of the simple random sample is the

Completely Randomized Design : Experimental units are randomly assigned to treatment groups. Every experimental unit has the same chance of being assigned to a particular treatment group.

Balanced Design – the most common type of completely randomized design – equal number of experimental units in each treatment group. When in doubt, use a balanced design!

Example** : **Deer Contraceptives.



As an alternative to hunting, deer contraceptives are sometimes used to control deer populations. A study examined the effect of Norgestomet at 0mg (placebo), 14mg, 21mg, 28mg, or 42mg. Twenty does were randomly assigned to each treatment group for a total of 100 does. For each doe, it was recorded whether or not she had a fawn during the next mating season.

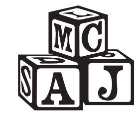
*This is a **balanced completely randomized design** with*

- *Five treatment groups,*
- *20 experimental units per treatment group where a doe was an experimental unit*
- *100 observations total*

My Grandpa's Farm . . .

One of the most common methods for achieving local control (i.e. increasing ability to observe differences between treatment groups) is to use a

Block Design :



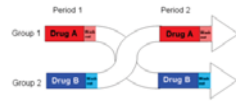
- Intention of blocking is to divide experimental units according to factors that are thought to have an effect on the response variable.
- Experimental units known to be similar are divided into blocks. Each block receives ALL treatment groups.
- Analogous to a stratified sample (a block is like a strata)

Example : Round-Up™. *Blocked by Species – anticipated that effect of Round-Up would be species dependent. Three experimental units (one plant in one pot) were assigned to each combination of treatment group and blocking factor.*

Round-Up Concentration

Species	**0**	**0.1**	**0.25**	**0.5**	**1.0**
Rye Grass	3 pots	3 pots	3 pots	3 pots	3 pots
Radish	3 pots	3 pots	3 pots	3 pots	3 pots

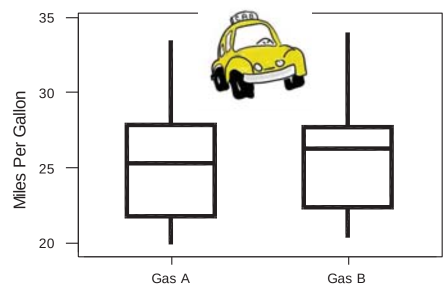
Crossover Design



Often, variability within subject/unit is greater than the variability between treatments/groups. In this case, the usual experimental design will not detect treatment differences. However, a **Crossover Design** looks at the treatment difference WITHIN subjects.

Example *: (Cartoon Guide) : Compare ethanol vs. regular gas in 10 cabs to test for differences in gas mileage.*

First experiment *: Completely Randomized Design. Assign 5 cabs to each group.*



Boxplots show little difference between gas treatments. Trouble is, lots of variability between cars (causes ? ?). Treatment variability is small relative to Error (other variability)

Second Experiment *:* ***Crossover Design*** *: Randomly assign 5 cars to each treatment group.*

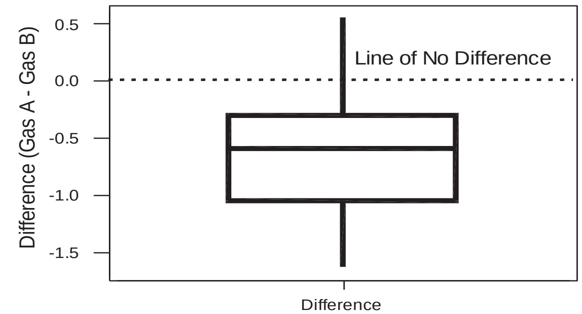
- *On Day 1, Group 1 gets gas A and Group 2 gets Gas B.*
- *On Day 2, Group 1 gets gas B and Group 2 gets Gas A.*

Look at the difference in gas mileage ***within each car.***

	Group 1	**Group 2**
Day 1	Gas A	Gas B
Day 2	Gas B	Gas A

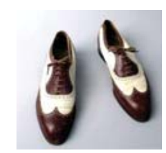
(Why do we need both orders Gas A-B, Gas B-A ???)

Boxplot of differences now shows strong evidence of a difference.

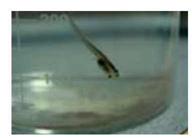


In this case, we ***reduced the experimental variability relative to the treatment variability.***

Matched Pair Design



Sometimes, it isn't possible to use the same individual for multiple treatments :



Example : *You investigate the effects of two incubation temperatures on tadpoles. After incubation, tadpoles are boiled and death temperature is noted.*

Example *: you want to compare job skills after a 25 period between people who had pre-school experience and people who don't have pre-school experience.*

In this case, it is better to use a **matched pair** of very similar individuals and give each one a different treatment. Then, you take the **difference** in their responses to treatment.

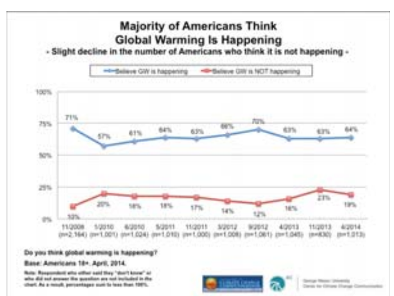
Other Experimental Designs

- **Factorial Design** – two or more treatment factors (we'll discuss this when we do Two-Way ANOVA)
- **Analysis of Covariance** – take out effect of a continuous factor (we'll cover later in the semester)
- **Latin Squares, Unbalanced Designs, Random Effects Models** (take experimental design class), **. . . . . .**

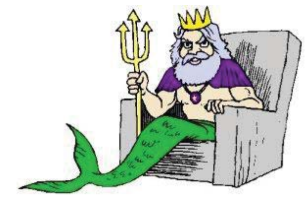
Toward Statistical Inference

Inference – use information about a sample to draw an inference about the population

Example *: A Yale poll of 1000 people reveal that in 2014, about 64% of Americans believed that Global Warming was actually happening (compare to 22% in 1991, 71% in 2008). We turn the <u>fact</u> that 64% of the sample have this opinion into an <u>estimate</u> that 64% of all Americans feel this way.*



Remember :



Parameter – a fixed number that describes a **population** (i.e., μ = true population mean height). We don't know this number (Gods only)

Statistic – a number that describes a **sample** (i.e., $\overline{X}$ = sample mean height). We know this number, <u>but</u> the number can (and usually does) **change from sample to sample**. Use the statistic to estimate the unknown parameter!

Sampling Variability – If we repeated our sampling procedure 'many' times, the same way each time, how much would our statistics change from one sample to the next?

Sampling Distribution of a statistic – the distribution of values of a sample statistic in all possible samples of the same size from a fixed population.

Example : *Flip a coin 10 times*

Example *: Let p be the **true proportion** of the population that believes in global warming (the **PARAMETER).***

*Suppose a **TOTAL of FOUR** people live in the U.S. (this is the **population**). I.e., just this once, we know the entire population. Here are their opinions (known to Gods, who are letting us know just this once . . .)*

- *We want to estimate p using a **STATISTIC** $\hat{p}$, the sample proportion that believes in global warming.*

Individual	***Attitude***
1	*Believe*
2	*Believe*
3	*Don't Believe*
4	*Don't Believe*

*In this population, p=0.5 (the **parameter,** i.e. the **true** proportion of the population that believes in global warming).*

<u>NOW :</u> *Pretend we don't know p=0.5, so we take a sample of size n=2.*

*List **all possible** Simple Random Samples (SRS) of size 2 from this population, and record the sample proportion for each sample (the **statistic**, $\hat{p}$)*

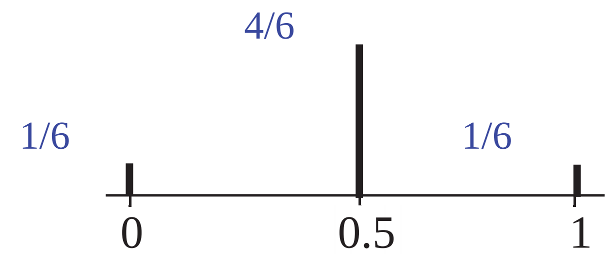
POPULATION

Individual	***Attitude***
1	Believe
2	Believe
3	Don't Believe
4	Don't Believe

POSSIBLE SAMPLES

Individuals in SRS	***Attitude***	$\hat{p}$
1 2	*B B*	1
1 3	*B DB*	0.5
1 4	*B DB*	0.5
2 3	*B DB*	0.5
2 4	*B DB*	0.5
3 4	*DB DB*	0

*In terms of probability, this is the **sampling distribution** for $\hat{p}$ for samples of size two :*

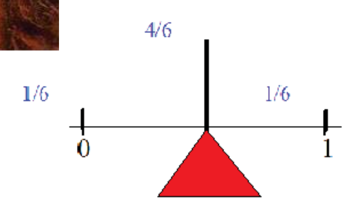


The sampling distribution gives

- All possible values of $\hat{p}$
- The **proportion** of times $\hat{p}$ takes on each of these values, or the **probability** that $\hat{p}$ takes each of these values.

Now : what is the average value of the sample proportion, $\hat{p}$?. This is called the **Mean of a sampling distribution.**

Mean of a distribution =



sampling balancing point

In this case, the *Mean of sampling distribution of $\hat{p}$ = 0.5*

This is also the value of the parameter p, the <u>true</u> proportion of the population that believes in global warming.

If the mean of the sampling distribution of a statistic equals the true value of a parameter, the statistic is said to be an **UNBIASED ESTIMATOR** of the parameter

Now : what is the variability of the sampling distribution of $\hat{p}$?

We'll see formulas for calculating this later. Suffice it to say that the standard deviation of $\hat{p}$ is about 0.32. Trust me.

***NOW** : Suppose our budget is so small we can only afford a sample of size n=1. How does the sampling distribution of $\hat{p}$ change?*

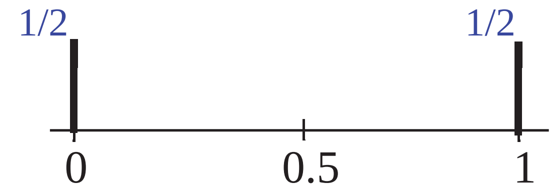
POPULATION

Individual	*Attitude*
1	*Believe*
2	*Believe*
3	*Don't Believe*
4	*Don't Believe*

POSSIBLE SAMPLES

Individuals in SRS	*Attitude*	$\hat{p}$
1	*Believe*	1
2	*Believe*	1
3	*Don't Believe*	0
4	*Don't Believe*	0

Sampling Distribution of $\hat{p}$ for n=1 :



$\hat{p}$ *is still **unbiased**: Mean of sampling distribution = 0.5*

= true proportion that believe in global warming, p

***Variability** of the sampling distribution of $\hat{p}$ is clearly 0.5 in this case.*

<u>SO :</u> *Mean of sampling distribution of $\hat{p}$ = .5 for samples of size 1 or 2.*

However :

***Standard Dev.** of $\hat{p}$ with samples of size 2 $\approx$ 0.3*

***Standard Dev.** of $\hat{p}$ with samples of size 1 $\approx$ 0.5*

*Taking samples of size 2 gives us estimates that are **less variable!***

BIAS and VARIABILITY

Bias of an estimator

= (mean of sampling distrib.) - (true value of parameter)

Statistic is **unbiased** if bias = 0.

Taking a random sample guarantees that a statistic will be unbiased.

Variability of an estimator

= (Standard Deviation of sampling distrib.)

Variability is a function of sample size – larger sample size = less variability in estimator

To see this, we simulate (make up) data. . . . .

Example : *Suppose the true proportion of people who believe in global warming is 0.80, or 80%.*

- *We now assume that the population is large, and we just happen to know the true proportion of believers (i.e. p = 0.80)*
- *If we take a sample of size n=10, what sort of values for $\hat{p}$ (the sample proportion) might we see? How often will be see these particular values? This is the* ***sampling distribution****.*
- *To estimate the sampling distribution, let's simulate taking many samples of size n=10 (how about 1000 such samples), and make a histogram of how often we see each value of $\hat{p}$.*

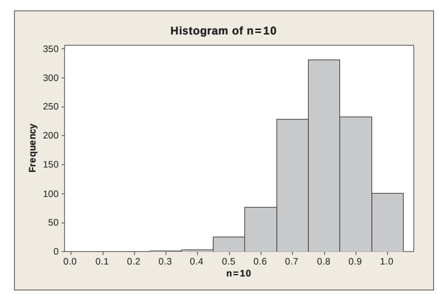
Random Binomial Data in MINITAB : use `Calc → Random Data → Binomial` (more on the binomial distribution later). The number of trials is 10, the probability of success is 0.8, we generate 1000 rows of data.

SPSS: This is what I can figure out : start in excel, make a spreadsheet with the numbers 1 though 1000 in one column. Import into the SPSS. Then use `Transform → Compute.` Use the function `RV.BINOM(10,.8)` (first number is n, second is p, probability)

Note *: this is just an* ***EXERCISE*** *– you would never actually take many samples of size 10, only* ***ONE*** *sample of size 10. We are doing this to see what values (and with what frequency) our sample proportion $\hat{p}$ might take! Another way to think about it : about how far can $\hat{p}$ be from the true value 0.80 just by chance?*

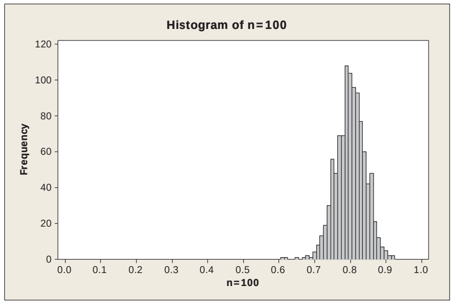
Here is the estimated sampling distribution of $\hat{p}$ for samples of size

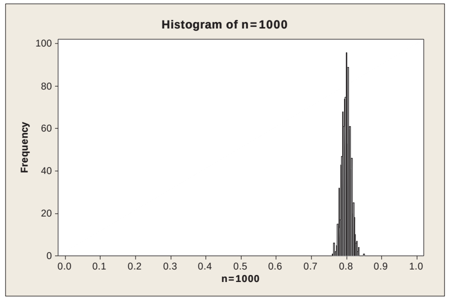
n = 10 :

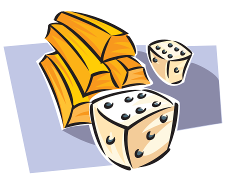


Now, suppose we took samples of size n=100. How does the estimated sampling distribution of $\hat{p}$ change?

Now, suppose we took samples of size n=1000. How does the estimated sampling distribution of $\hat{p}$ change?







PROBABILITY

Chapter 3 in Cartoon Guide –

STRONGLY RECOMMENDED

Probability is crucial to **statistical inference**

- Inferences are always expressed in terms of probability

  (i.e. a "95% Confidence Interval". 0.95 is the probability of something . . .)

A survey – Suppose exactly 80% of people believe in global warming.

- We take a random sample of size 100
- Expect to see about 80 believe in global warming.
- How likely are we to observe more than 90 who believe in global warming?

Probability Models – We're modeling some random phenomenon. A probability model consists of

- A List of possible outcomes
- A probability for each outcome

Sample space = S = set of all possible outcomes

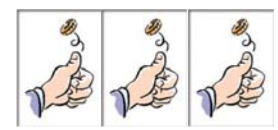
Examples :

Discrete

Toss a coin: S = *{H,T}.*

Watch a tree for a year and see if it dies :
S = *{Dead, Alive}.*

Toss a coin 3 times:



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

Continuous

Record the gas mileage for a car :
S = *{a positive number between 0 and ? (300mpg)?}*

An **event** : set of *some* possible outcomes :

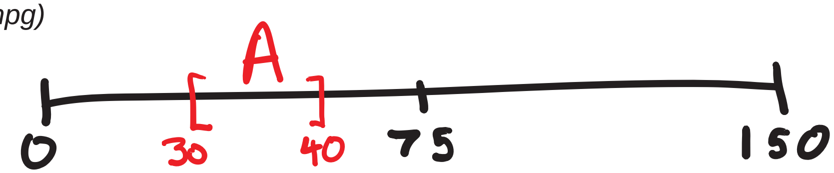
- An event is a <u>subset</u> of outcomes in S
- Denoted by A (or B or C)

Example *: Let the event* A *= (get one head in 3 tosses)*

{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

= {HTT, THT, TTH}

Example *: Let the event* A *= (Gas mileage between 30 and 40 mpg)*



Probability measure : a function (satisfying certain conditions) that assigns a ***probability*** (a number between 0 and 1) to each event.

If A is an event, $P(A)$ denotes the probability of A.

SO : What does probability mean, and how do we assign probabilities to outcomes?!? (i.e., how do we define a probability measure?)

Three approaches :

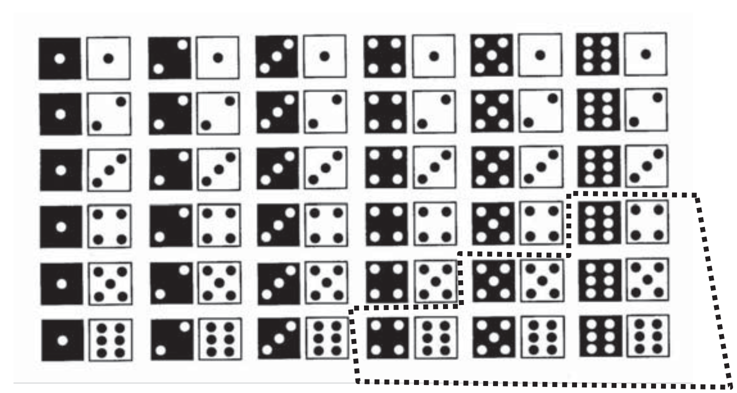
1) *Classical*
2) **Relative Frequency**
3) Personal/Subjective Probability (Baysian)

1) *Classical* (based on gambling): Sometimes, we believe all possible outcomes are equally likely (i.e. the game is fair!). In this case

$$P(A) = \frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S}$$

Example : *Toss a coin once,* $S=\{H,T\}$.
If $A=\{H\}$, then $P(A)=0.5$

Example : *Roll Two Dice* $S=\{\text{see picture}\}$



If $A=\{$*Sum of Dice at least 10*$\}$, then

$$P(A) = \frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S} = \frac{6}{36} = 0.167$$

2) **Relative Frequency (Long Run Frequency):** When an experiment can be repeated, the probability of an event is the proportion of times the event occurs in the long run

Example : *What's the probability a radish seed will grow in soil treated with RoundUp? Plant many, many seeds, count the number of times the seed germinates.*

3) Personal/Subjective Probability (Baysian): Most events in life aren't repeatable. We assign probabilities all the time :

- What's the probability I'll take statistics?
- What's the probability this guy will ask me out on a date?
- What's the probability a huge body of fresh water will halt the gulf stream and lead to an ice age within a century? (some people think high . .)

Think in terms of betting . . . .

This may seem arbitrary, but it's actually a good description of how we assign probabilities. Baysians assign a probability based on the information at hand, **and update their probability as more data becomes available.**

Regardless of how you choose to assign probabilities, the following two rules apply :

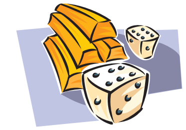
1) For any event A, $0 \leq P(A) \leq 1$

(Probabilities are always between zero and one)

2) $P(S) = 1$

(Something must happen)

Now a bit of gambling history . . . . .



A rich Frenchman, Antoine Gombaud, known as the **'Chevalier de mere'** liked to gamble. However, he was confused by certain experiences at the gambling tables. He posed the following question to his mathematician friend, **Blaise Pascal** in 1654 :

Which is more likely :

1) *At least one six in four rolls of a single die*

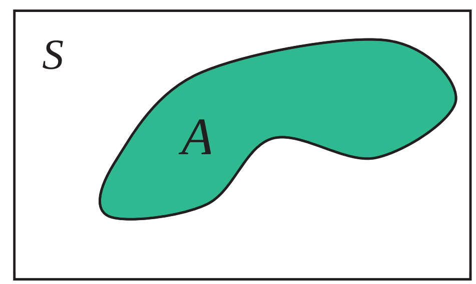
2) *At least one double-six in 24 rolls of a pair of dice*

Pascal with his friend **Pierre de Fermat** soon worked out the Chevalier's problem, and in the process developed the algebraic basis of probability theory.

Here is the theory they worked out . . .

Venn Diagrams : Represent Sample Spaces and Events with **pictures!**

- Think about S as a car windshield
- A is an area in the windshield.
- It's about to start raining.
- Let A be the event that the first drop lands in the area A.
- Rain is equally likely to fall anywhere on the windshield.



Probability measure for this picture : **Probability = Area**

$$P(A) = \frac{\text{area of } A}{\text{area of } S}$$

For convenience assume (area of windshield S) = 1.

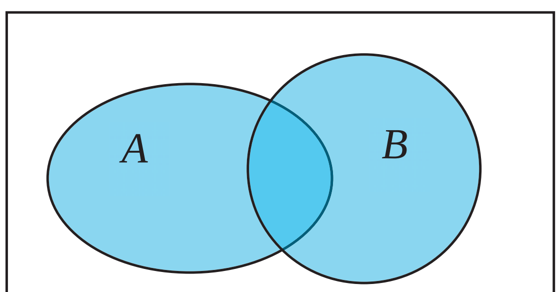
So $P(A)$ = area of A.

Remember : $0 \leq P(A) \leq 1$ and $P(S) = 1$

Make New Events from Old Events

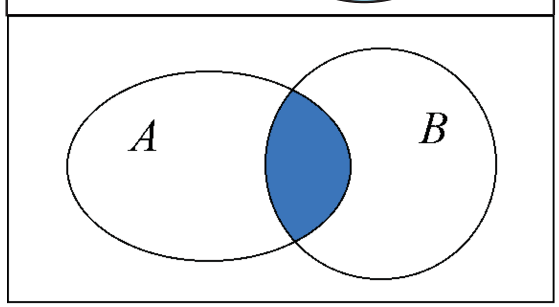
A **or** B

(raindrop falls in A or B)

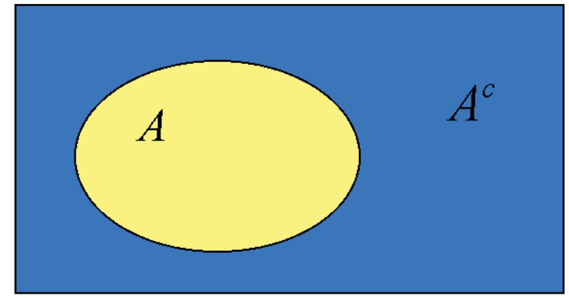


A **and** B

(raindrop falls in A and B)



Complement of A
(raindrop falls in 'not A')

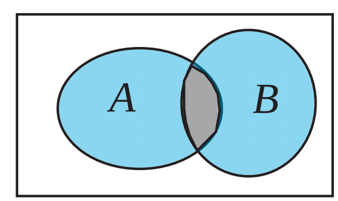


Axioms of Probability
(properties of Probability Measures)

- For each event A, $\quad 0 \leq P(A) \leq 1$
- $P(S) = 1$, where S is the whole sample space.
- **Addition Rule :**

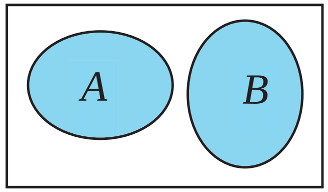
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

(Middle shaded area gets counted twice, so we have to subtract this area, which is A *and* B*)*



- **Disjoint Events :** If A and B are disjoint, then

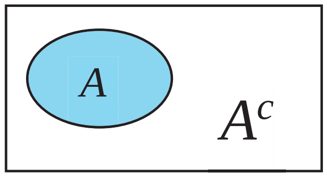
$$P(A \text{ or } B) = P(A) + P(B)$$



i.e. $P(A \text{ and } B) = 0$

Complement rule

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



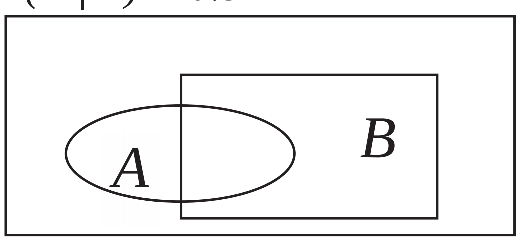
Conditional Probability

<u>Notation</u> : $P(B \mid A)$

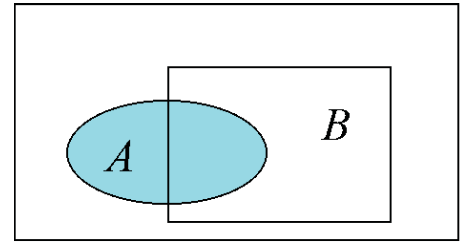
Read as " Probability of B given A "

<u>Meaning</u> : Given that A has already happened, what is the probability of B?

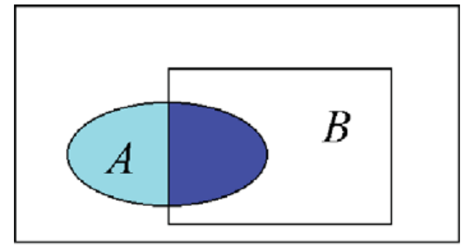
By eyeball : $P(B \mid A) = 0.5$



Given that the raindrop fell in A, we restrict our attention to the set A. The drop is equally likely to fall anywhere within A.



Given A, the event B also occurs when the drop falls in the darker region, i.e., the event (A and B).

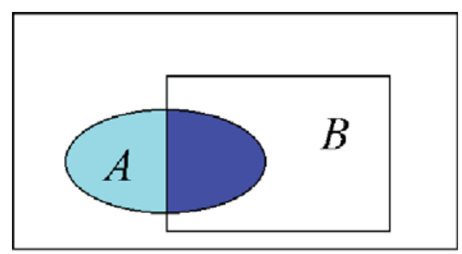


This gives the formal definition for Conditional Probability :

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Or, equivalently

$$P(A \text{ and } B) = P(B \mid A) P(A)$$



Independence

Two events A and B are **independent** if being told that A occurred has no effect on the probability that B also occurs.

Formal Definition :

A and B are independent if

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

Equivalently,

$$\frac{P(A \text{ and } B)}{P(A)} = P(B)$$

Equivalently,

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

<u>WARNING :</u>

Don't confuse <u>Independent</u> with <u>Disjoint</u>

Independent Events : $P(B \mid A) = P(B)$

Disjoint Events : $P(A \text{ and } B) = 0$ or equivalently $P(B \mid A) = 0$

If events are disjoint, they cannot be independent. If events are independent, they cannot be disjoint.

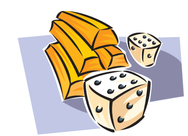
Example *: Toss two coins. Given that the first coin is heads, what is the probability that the second coin is also heads?*

Coins are independent so

P(Heads (Toss 2) | Heads (Toss 1)) = P(Heads (Toss 2))

*That is, these events are **independent**.*

Example *: Roll a pair of dice. Let A be the event that the first die equals 5. Let B be the event that the sum of the dice equals 4.*



$$P(B) = \frac{3}{36}$$

(3&1, 1&3, or 2&2; 36 possible outcomes)

$$P(B \mid A) = 0$$

So

$$P(A \text{ and } B) = P(A)*P(B|A) = 1/6 *0 = 0$$

SO : *A and B are **disjoint*** (can't roll a 5 and have sum be 4)

*However, they are **not independent** since* $P(B \mid A) = P(B)$

Back to the Chevalier's problem . . . . .

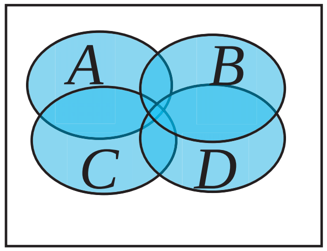
Which is more likely :

1) *At least one six in four rolls of a single die*
2) *At least one double-six in 24 rolls of a pair of dice*

1) Let the events A, B, C, D be the events of getting a six on roll 1, 2, 3, or 4, respectively.

We want

$$P(A \text{ or } B \text{ or } C \text{ or } D)$$
$$= P(A) + P(B) + P(C) + P(D)$$
$$- \quad (\text{probability of overlaps})$$



Hmm. Overlaps look hard.

Better idea : USE COMPLEMENT RULE :
Get area of region outside the discs.

$$P(\text{at least one six}) = 1 - P(\text{no sixes})$$
$$= 1 - P(A^c \text{ and } B^c \text{ and } C^c \text{ and } D^c)$$
$$= 1 - P(A^c)P(B^c)P(C^c)P(D^c)$$
$$= 1 - \left(\frac{5}{6}\right)^4 = 0.518$$

*Because dice rolls are **independent** – the value of one die throw does not change the likelihood of outcomes on subsequent*

2) *At least one double-six in 24 rolls of a pair of dice*

Similar reasoning (use complement rule), gives that

$$P(\text{at least one double six}) = 1 - P(\text{no double sixes})$$
$$= 1 - \left(\frac{35}{36}\right)^{24} = 0.491$$

SO : more likely to get one six in four throws of the dice!

Probability in Practice

(The trick is knowing when to apply which probability rule!)

Suggestions :

- Read the Textbooks. I like them.
- Do problems in textbooks.
- Make a picture. This helps to clarify sample spaces.
- Do some more problems.

Example : Who will you vote for? *The Real Clear Politics Average on 8/30/16 for likely voters had the following distribution for a four way race:*

- *Clinton* *42%*
- *Trump* *38%*
- *Johnson* *8%*
- *Stein* *3%*

Suggestion : Check to see if probabilities satisfy requirements :
Let A be the event that a person has a particular opinion about 'who is to blame'.

1) For any event A, $0 \le P(A) \le 1$ *(all fine here)*
2) $P(S) = 1$ *(umm . . .)*

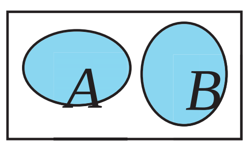
Probabilities of events so far only add to 91% - we need another category (none/other : 9%)

***Question :** what is the probability that a person picked at random would vote for Johnson or Stein?*

Help : 'or' means $P(A \text{ or } B)$. Use addition rule and then think about whether events are disjoint.

$$P(J \text{ or } S) = P(J) + P(S) - P(J \text{ and } S)$$

*Last probability is zero since can't vote for two people– i.e., these events are <u>disjoint</u> **:***



$P(J \text{ and } S) = 0$. So :

$$P(J \text{ or } S) = P(J) + P(S) = .08 + .03 = .11$$

***Question :** what is the probability that three randomly chosen people all plan to vote for Trump?*

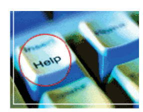
Help : Probabilities multiply <u>**ONLY**</u> if events are <u>**independent**</u>. Think carefully about if this is true!

In this case, it is reasonable to assume people are independent (we picked them at random!) so

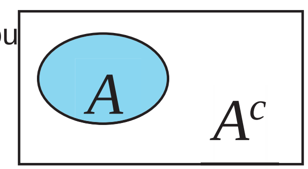
$$P(A \text{ and } B) = P(A)P(B)$$

$$P(\text{"T" and "T" and "T"}) = P(\text{"T"})^3 = .38^3 = .055$$

***Question :** what is the probability that if we choose four people at random, at least one will vote for Clinton?*



Help : When you see ***'AT LEAST'*** you should think ***'Complement Rule'*** :
$P(A) = 1 - P(A^c)$



$$\begin{aligned} &P(\text{At least one Clinton}) \\ &= 1 - P(\text{none of four choose Clinton}) \\ &= 1 - P(\text{one does not choose Clinton})^4 \\ &= 1 - (.58)^4 \\ &= 0.89 \end{aligned}$$

Counting and Probability

***Example** : In 5 card draw, a straight is a sequence of cards in numerical sequence without regard to suit. An ace may be the low card or the high card in a straight. Show how to compute the probability of getting a straight.*

Probability problems can be solved in two ways . . .

1. Counting – based on the CLASSICAL definition of probability :

$$P(A) = \frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S}$$

In our case :

$$P(Straight) = \frac{\#\, ways \text{ to get a straight}}{\#\, ways \text{ to pick 5 cards from 52}}$$

Let's start with the denominator :

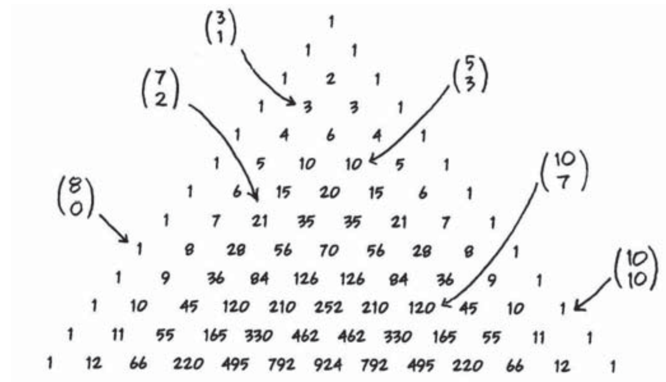
Number of Ways to Choose *k* of *n* things

For k in $0, 1, 2, \; . . . , n,$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad n! = 1*2*...*(n-1)*n$$

(read as "n choose k")

You can get these values from Pascal's Triangle = Each entry is the sum of the two numbers just above it



***SO :** number of ways to choose 5 from 52 cards is*

$$\binom{52}{5} = \frac{52!}{5!(47)!} = 2598960$$

NOW : the numerator. How many ways can we get a straight? Here are the relevant facts :

1. There are **10** possible starting cards for the straight : {A,2,3,4,5,6,7,8,9,10}
2. There are **4** possible suits for **each card** in the straight. That is, given a starting value, there are 4^5 possible straights starting on that value.

SO : Number of ways to get a straight is $10*4^5$.
THUS : Probability of a Straight is

$$P(Straight) = \frac{10*4^5}{\binom{52}{5}} = \frac{10240}{2598960} = .0039$$

2. Probability (with a bit of counting) –

Let's calculate

1) The probability of getting dealt a straight in order
2) The number of orders for five cards

The first card can be any card up to 10 in value :

$$\Pr(\text{first card} < 10) = 40/52$$

Next cards must be one of the four cards that are one higher in value than the previous card :

$$\Pr(\text{second card}) = 4/51 \qquad \Pr(\text{third card}) = 4/50$$
$$\Pr(\text{fourth card}) = 4/49 \qquad \Pr(\text{fifth card}) = 4/48$$

$$P(Straightinorder) = \frac{40}{52}\frac{4}{51}\frac{4}{50}\frac{4}{49}\frac{4}{48}$$

NOW : There are 120=5*4*3*2*1 ways to get dealt five particular cards. **SO :**

$$P(Straight) = 120 * \frac{40}{52} * \frac{4}{51} * \frac{4}{50} * \frac{4}{49} * \frac{4}{48} = .0039$$

Bayes Theorem

Example : Uganda, after an extensive safe-sex campaign, has reduced the rate of HIV infection from almost 15% (early 1990's) of adults to 7.1% (2015, estimated 10% in cities). (click here to learn more about AIDS in Uganda : http://www.avert.org/aidsuganda.htm)

An oral HIV test is given at random to an adult in the capital Kampala. If a given person has a positive test result, what is the conditional probability that the person indeed has the virus?

Information :

- 10% of the population has the virus
- Test has a false positive rate of 1%
- Test has a false negative rate of .03%

A = {HIV virus in blood} → A^c = {HIV **NOT** in blood}

B = {Blood test **positive**} → B^c = {Blood test **negative**}

We Want $P(A|B)$ = P (Have HIV | Positive Test)

$P(A)$ = 0.1 so $P(A^c)$=0.9

$P(B|A^c)$ = 0.01 so $P(B^c|A^c)$ = 0.99

$P(B^c|A)$ = 0.0003 so $P(B|A)$= 0.9997

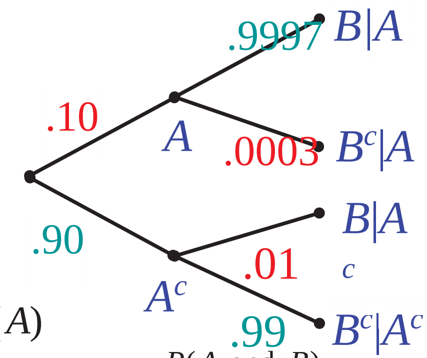
Want $P(Have\ HIV \mid Positive\ Test) =$

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

Draw a Probability Tree!

Get numerator :

Conditional Probability
(the other direction)



$$P(A \text{ and } B) = P(B \mid A) * P(A)$$
$$= 0.9997 * 0.1$$
$$= 0.099$$

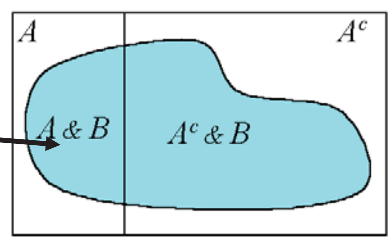
$$P(A \text{ and } B) = P(B \mid A) * P(A) = P(A \mid B) * P(B)$$

<u>Note that we multiply probabilities along the tree</u>

Get denominator :

$$P(B) = P(A \& B) + P(A^c \& B)$$
$$= (.1)(.9997) + (.9)(.01) = 0.1089$$

We want P(B), the shaded area!!



$$P(A \mid B) = \frac{P(A \& B)}{P(B)} = \frac{.099}{.1089} = .91$$

SO :

That is : a persons risk of having HIV after a positive tests increases from about 10% to 91% *(still a 9% chance of not having HIV).*

BAYES THEOREM

(What we just did)

Given $P(A)$, $P(B \mid A)$, and $P(B \mid A^c)$.
Want to find a "turned-around" probability like $P(A \mid B)$.

$$P(B) = P(A \& B) + P(A^c \& B)$$
$$= P(A)P(B \mid A) + P(A^c)P(B \mid A^C)$$

$$\boxed{P(A \mid B) = \frac{P(A \& B)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)}}$$

[Know this, don't memorize this . . .]