Statistical Analysis I

Discrete Random Variables

Terminology

A **RANDOM VARIABLE** is a function that assigns a numerical value to each outcome of a sample space.

Examples: Political Affiliation (1 = Democrat, 2 = Republican, 3 = Other) for
All U.S. citizens of voting age **DISCRETE**

Product Status (1 = Shippable, ie, meets all specs, 0 = Not Shippable, ie, does not meet >= 1 spec) for
All product produced from a specific manufacturing line **DISCRETE**

Complaint Response Time for **CONTINUOUS**
All complaints made about airline travel

Number of "matches" (0, 1, 2, or 4) for **DISCRETE**
All possible ways 4 babies could be randomly assigned to 4 mothers

NOTE: The second example above is a special kind of Random Variable called a ***Bernoulli Trial***.
A *Bernoulli Trial* is a Random Variable with only 2 outcomes – in the example, either Product is Shippable or it is not – other examples include the following:

Examples: The face-up after tossing a coin (1 = Heads, 2 = Tails)
The answer to a True/False question on an exam (1 = True, 2 = False)
The color of an M&M is Brown (1 = Brown, 2 = Not Brown)
The gender of a medical patient (1 = Female, 2 = Male)
ALL DISCRETE

NOTE: Three of the four initial examples above have a countable number of outcomes, but Complaint Response Time does not – this time could be any value > 0.

A **DISCRETE RANDOM VARIABLE** is one involving a countable set of outcomes

A **CONTINUOUS RANDOM VARIABLE** is one involving an uncountable set of outcomes

Discrete Random Variables

Probability Mass Functions

A Discrete **PROBABILITY MASS FUNCTION** (pmf) is a function f(x) such that

$$f_X(x) = P[X = x],$$

where x is a specific value (outcome) for Discrete Random Variable X.
Simply a function defining the probability of the specific outcome x.

Example: For tossing a coin, the Random Variable X = 0 if Tails & 1 if Heads has pmf given by X(0) = ½ and X(1) = ½, or in chart form as:

x	0	1
$f_X(x)$	½	½

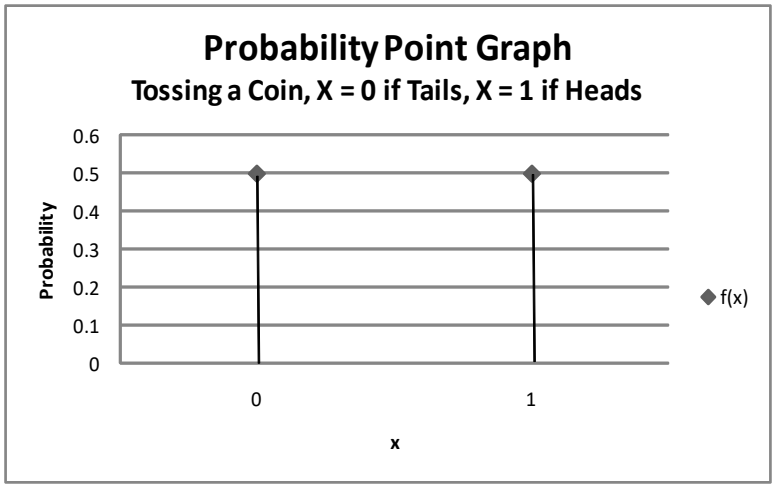
Example: For the exact matching of a random arrangement of the numbers 1, 2, 3, & 4, the Random Variable X = # Matches has pmf given by X(0) = 9/24, X(1) = 8/24, X(2) = 6/24, and X(4) = 1/24, or in chart form as:

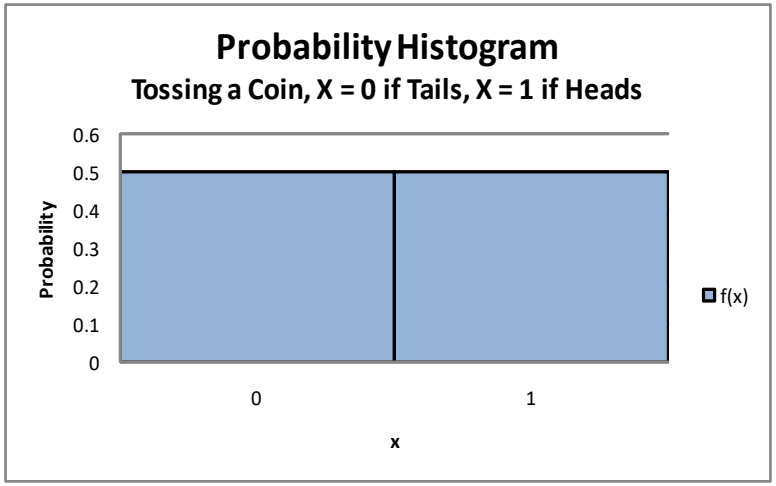
x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24

Discrete Random Variables

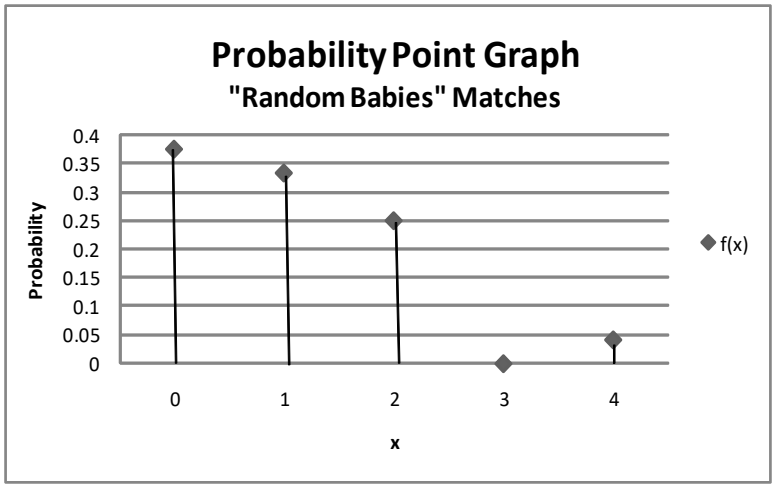
Point Graphs and Histograms

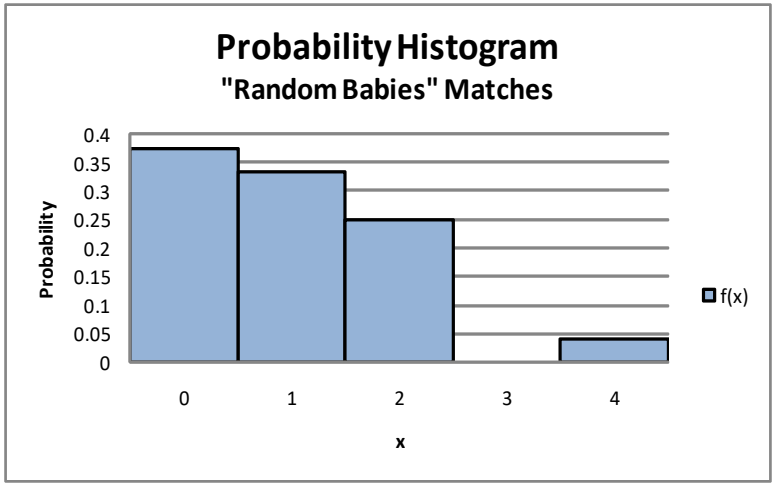
Example: Tossing a Coin





Example: "Random 1,2,3,4" Matches





Discrete Random Variables

Probability Mass Function Axioms and Summary

For any Discrete Probability Mass Function (pmf) given by $f_X(x)$:

1) $0 <= f_X(x) <= 1$ for each value of x

2) $\text{Sum}_{\text{All values of x}} \{ f_X(x) \} = 1$

Example: Tossing a Coin

x	0	1
$f_X(x)$	½	½

Sum = 1

Example: "Random 1,2,3,4" Matches

x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24

Sum = 1

In summary, a discrete probability mass function consists of two elements:

1) some type of listing of all the possible values of the random variable, and

2) the corresponding probabilities for each value.

Discrete Random Variables

Expected Value = Mean

The **EXPECTED VALUE** (or Mean) of a Discrete Random Variable X is given by:

$$\mu = E[X] = \text{Sum}_{\text{All values of } x}\{ x*f_X(x) \}$$

Example: "Random 1,2,3,4" Matches

x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24
$x*f_X(x)$	0	8/24	12/24	4/24

μ = Sum of Bottom Row
= (8 + 12 + 4)/24
= 24/24
= 1

Expected Value = Mean is simply Sum of each of:

1) Outcome x (ie, "times")
2) That Outcome's Probability of Occurring

Example: Sum when Rolling 2 Die

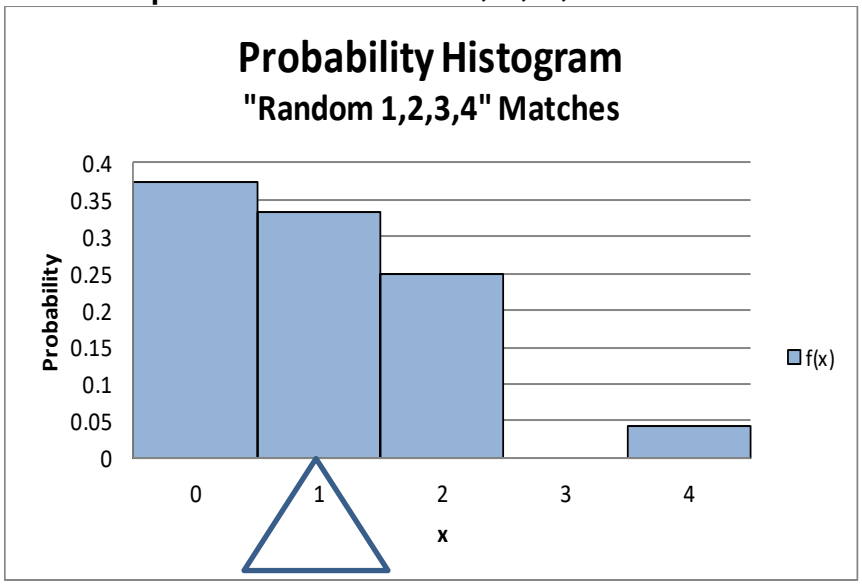
Result of Interest: **Sum of Both Die**			
Outcome	n(A)	P[A]	Outcome*P[A]
2	1	1/36 = 0.0278	2/36 = 0.0556
3	2	2/36 = 0.0556	6/36 = 0.1667
4	3	3/36 = 0.0833	12/36 = 0.3333
5	4	4/36 = 0.1111	20/36 = 0.5556
6	5	5/36 = 0.1389	30/36 = 0.8333
7	6	6/36 = 0.1667	42/36 = 1.1667
8	5	5/36 = 0.1389	40/36 = 1.1111
9	4	4/36 = 0.1111	36/36 = 1
10	3	3/36 = 0.0833	30/36 = 0.8333
11	2	2/36 = 0.0556	22/36 = 0.6111
12	1	1/36 = 0.0278	12/36 = 0.3333
All	36	36/36 = 1	252/36 = 7

μ = Sum of Last Column
= 252/36
= 7

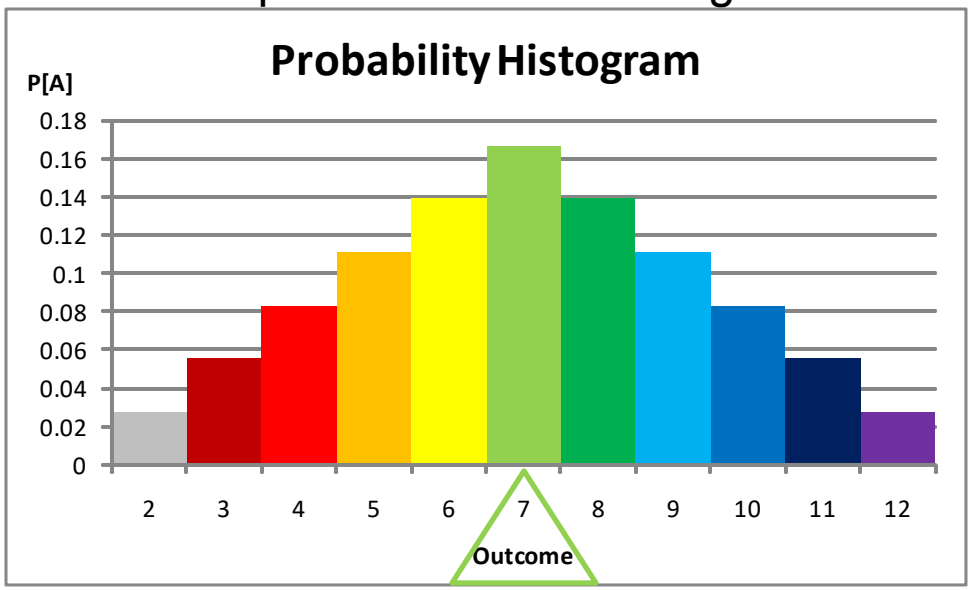
Discrete Random Variables

Expected Value = Center of Mass of Distribution

Example: “Random 1,2,3,4” Matches



Example: Sum when Rolling 2 Die



The EXPECTED VALUE is the point at which the PMF would be expected to “balance” as if on a fulcrum.

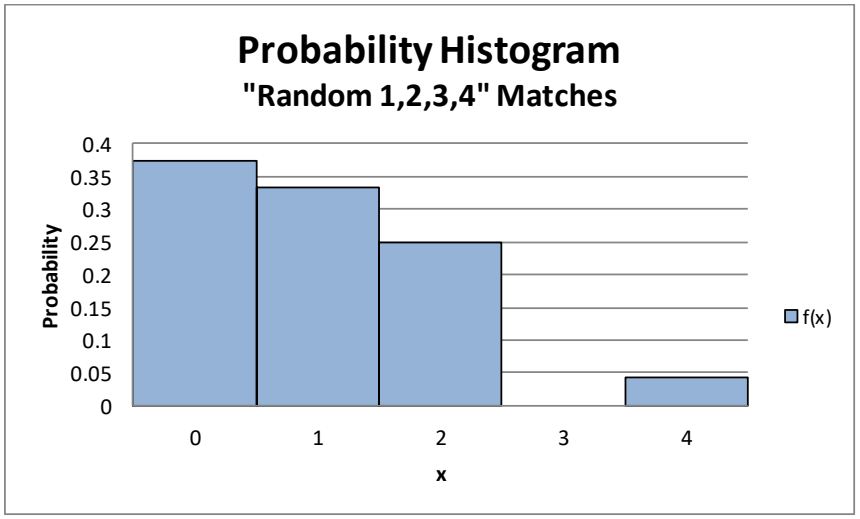
The MEAN as the “balancing” point is more obvious for this example, but is the Mean sufficient to describe a probability distribution?

Discrete Random Variables

Measure of Spread = Variance

The **VARIANCE** of a Discrete Random Variable X is given by:

$$\sigma^2 = Var(X) = E[(x-\mu)^2]$$
$$= Sum_{All\ values\ of\ x}\{ (x-\mu)^2 * f_X(x) \}$$



Alternative formula for Variance:

$$\sigma^2 = Var(X)$$
$$= Sum_{All\ values\ of\ x}\{ x^2 * f_X(x) \} - \mu^2$$

Example: "Random 1,2,3,4" Matches

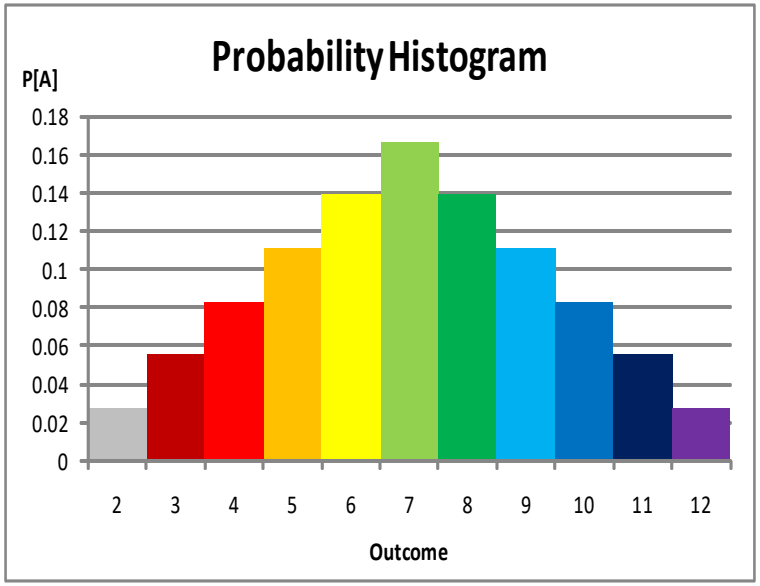
x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24
$x*f_X(x)$	0	8/24	12/24	4/24
Expected Value = E[X] = μ = 1				
$(x-\mu)$	-1	0	1	3
$(x-\mu)^2$	1	0	1	9
$(x-\mu)^2 f_X(x)$	9/24	0	6/24	9/24
Variance = Var(X) = 9/24 + 6/24 + 9/24 = 1 Standard Deviation = Sqrt[Var(X)] = 1				

x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24
x^2	0	1	4	16
$x^2 * f_X(x)$	0	8/24	24/24	16/24
Variance = 8/24 + 24/24 + 16/24 – $(1)^2$ Var(X) = 48/24 – 1 = 2 – 1 = 1 (same as above)				

Discrete Random Variables

Measure of Spread = Variance

Example: Sum when Rolling 2 Die



Result of Interest: **Sum of Both Die**					
Outcome	n(A)	P[A]	Outcome*P[A]	Outcome^2	Outcome^2*P[A]
2	1	1/36 = 0.0278	2/36 = 0.0556	4	4/36 = 0.1111
3	2	2/36 = 0.0556	6/36 = 0.1667	9	18/36 = 0.5
4	3	3/36 = 0.0833	12/36 = 0.3333	16	48/36 = 1.3333
5	4	4/36 = 0.1111	20/36 = 0.5556	25	100/36 = 2.7778
6	5	5/36 = 0.1389	30/36 = 0.8333	36	180/36 = 5
7	6	6/36 = 0.1667	42/36 = 1.1667	49	294/36 = 8.1667
8	5	5/36 = 0.1389	40/36 = 1.1111	64	320/36 = 8.8889
9	4	4/36 = 0.1111	36/36 = 1	81	324/36 = 9
10	3	3/36 = 0.0833	30/36 = 0.8333	100	300/36 = 8.3333
11	2	2/36 = 0.0556	22/36 = 0.6111	121	242/36 = 6.7222
12	1	1/36 = 0.0278	12/36 = 0.3333	144	144/36 = 4
All	36	36/36 = 1	252/36 = 7	-	1974/36 = 54.8333

Expected Value = μ = 252/36 = 7

Variance = σ^2 = 1974/36 – $(7)^2$ = 54.8333 – 49 = 5.8333

Standard Deviation = σ = Sqrt(5.8333) = 2.4152

Using Standard Deviation

Chebyshev's Theorem

For any random variable, X, with mean μ and standard deviation σ, then the probability that X lies within k standard deviations of the mean is at least $(1 - 1/k^2)$, where $k > 0$, that is

$$P[\mu - k\sigma <= X <= \mu + k\sigma] >= 1 - 1/k^2, \quad k>0.$$

For k = 1, Chebyshev indicates:

$$P[\mu - 1\sigma <= X <= \mu + 1\sigma] >= 1 - 1/1^2 = 0$$

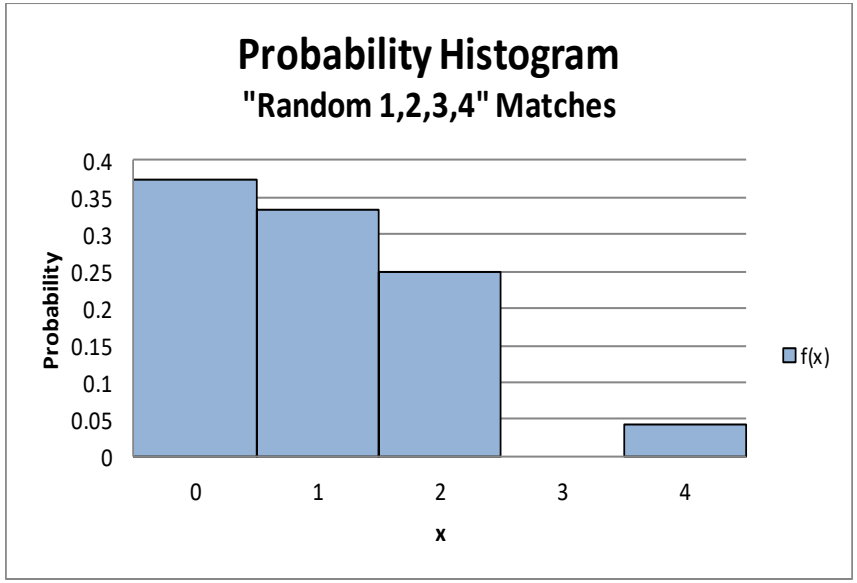
So … no real information here, but for k>1:

k	$1-k^{-2}$	Information about Distribution
1.5	0.556	At least 55.6% of Distribution within +/-1.5 Standard Deviations of the Mean
2	0.75	At least 75% of Distribution within +/-2 Standard Deviations of the Mean
3	0.889	At least 88.9% of Distribution within +/-3 Standard Deviations of the Mean
4	0.938	At least 93.8% of Distribution within +/-4 Standard Deviations of the Mean

Using Standard Deviation

Chebyshev's Theorem

Example: "Random 1,2,3,4" Matches



x	0	1	2	4
$f_X(x)$	9/24	8/24	6/24	1/24
Mean = μ = 1				
Standard Deviation = σ = 1				

P[μ - 1σ <= X <= μ + 1σ] =
P[1 -1 <= X <= 1 + 1] =
P[0 <= X <= 2] =
9/24 + 8/24 + 6/24 =
23/24 = 0.9583 >= 0

P[μ - 2σ <= X <= μ + 2σ] =
P[1 -2 <= X <= 1 + 2] =
P[-1 <= X <= 3] =
9/24 + 8/24 + 6/24 =
23/24 = 0.9583 >= 0.75

P[μ - 3σ <= X <= μ + 3σ] =
P[1 -3 <= X <= 1 + 3] =
P[-2 <= X <= 4] =
9/24 + 8/24 + 6/24 + 1/24 =
24/24 = 1 >= 0.889

Using Standard Deviation

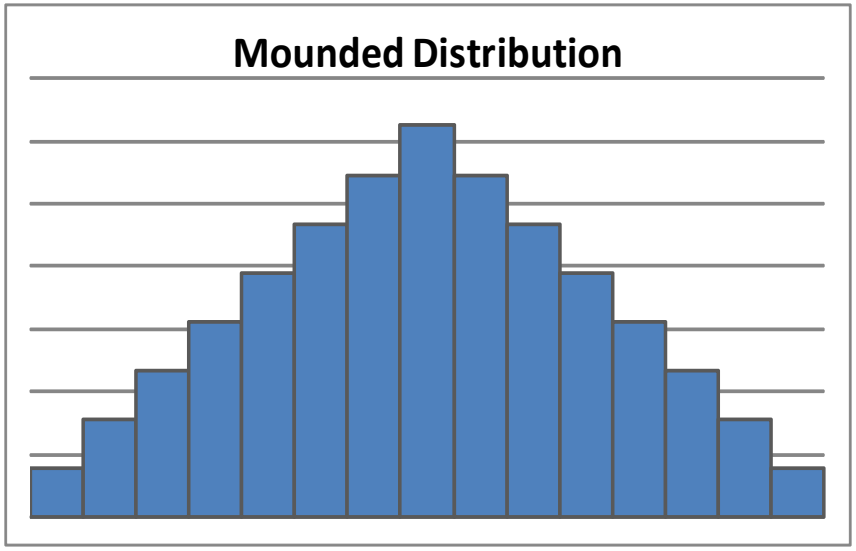
Empirical Rule

For a “mounded” distribution,

$$P[\mu - 1\sigma <= X <= \mu + 1\sigma] \approx 68\%$$

$$P[\mu - 2\sigma <= X <= \mu + 2\sigma] \approx 95\%$$

$$P[\mu - 3\sigma <= X <= \mu + 3\sigma] \approx 100\%$$

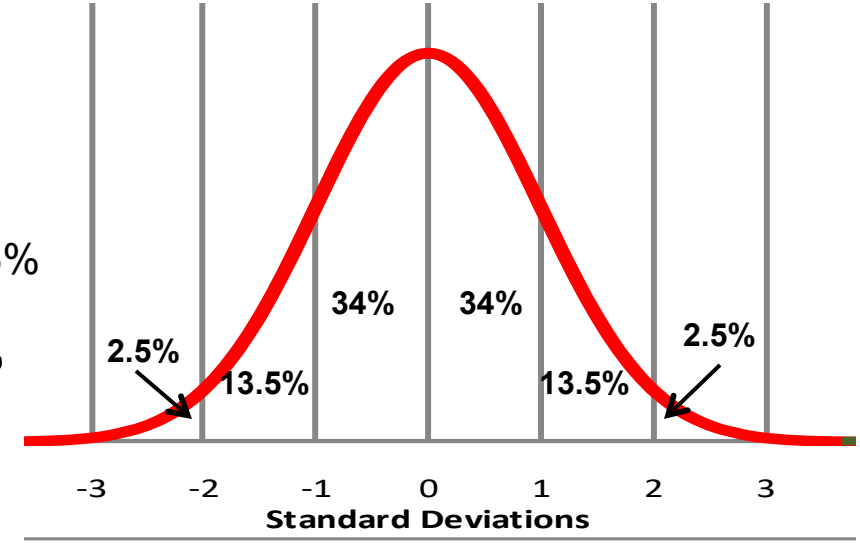


Since “mounded” distributions are Symmetric about their Mean:

$$P[\mu <= X <= \mu + \sigma] = P[\mu - \sigma <= X <= \mu] \approx 34\%$$

$$P[\mu <= X <= \mu + 2\sigma] = P[\mu - 2\sigma <= X <= \mu] \approx 47.5\%$$

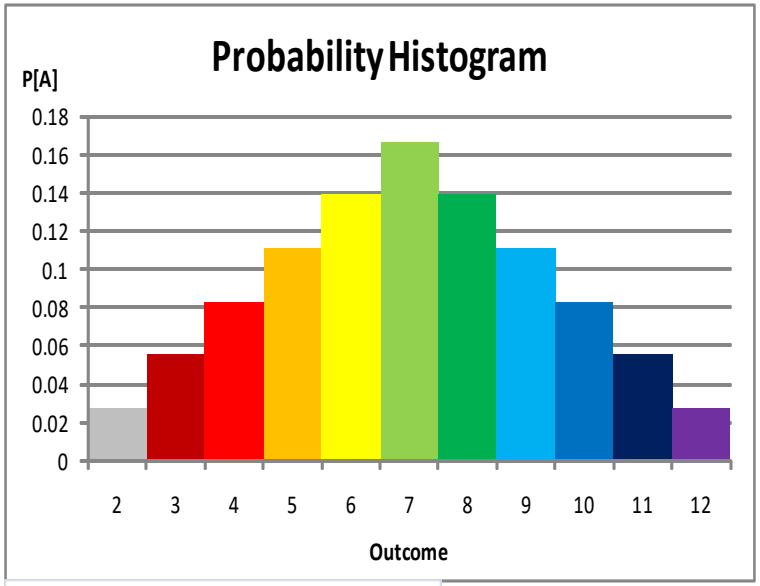
$$P[\mu <= X <= \mu + 3\sigma] = P[\mu - 3\sigma <= X <= \mu] \approx 50\%$$



Using Standard Deviation

Empirical Rule

Example: Sum when Rolling 2 Die



Result of Interest: **Sum of Both Die**

Outcome	n(A)	P[A]
2	1	1/36 = 0.0278
3	2	2/36 = 0.0556
4	3	3/36 = 0.0833
5	4	4/36 = 0.1111
6	5	5/36 = 0.1389
7	6	6/36 = 0.1667
8	5	5/36 = 0.1389
9	4	4/36 = 0.1111
10	3	3/36 = 0.0833
11	2	2/36 = 0.0556
12	1	1/36 = 0.0278

$\mu = 7$
$\sigma = 2.145$

$P[\ \mu - 1\sigma <= X <= \mu + 1\sigma] =$
$P[\ 7 - 2.145 <= X <= 7 + 2.145] =$
$P[\ 4.855 <= X <= 9.145] =$
$4/36 + 5/36 + 6/36 + 5/36 + 4/36 =$
$24/36 = 0.6667 \approx 0.68$

$P[\ \mu - 2\sigma <= X <= \mu + 2\sigma] =$
$P[\ 7 - 4.29 <= X <= 7 + 4.29] =$
$P[\ 2.71 <= X <= 11.29] =$
$2/36 + 3/36 + 24/36 + 3/36 + 2/36 =$
$34/36 = 0.9444 \approx 0.95$

$P[\ \mu - 3\sigma <= X <= \mu + 3\sigma] =$
$P[\ 7 - 6.435 <= X <= 7 + 6.435] =$
$P[\ 0.565 <= X <= 13.435] =$
$1/36 + 34/36 + 1/36 =$
$36/36 = 1 = 1$

Binomial Probability Distribution

Let a random variable Y be the number of "successes" in n independent Bernoulli trials each with P["success"] = p, then Y has a **Binomial Probability Distribution**, and the pmf can be expressed as

$$f_Y(y) = P[Y=y] = \binom{n}{y} p^y (1-p)^{n-y},$$

where $\binom{n}{y}$ is the combination of n things taken y at a time = $\frac{n!}{y!(n-y)!}$.

Example: Consider a baseball player who has a lifetime on-base percentage of 0.300. This means he is "successful" in reaching at least 1st base in 30% of his "plate appearances" (ie, "PA's = "trials").

In each "plate appearance" he is either "successful" in getting "on-base", or not, these are Bernoulli trials (recall – a Bernoulli random variable is one with only 2 outcomes – "on-base" or "not on-base" = "out" in this case).

So in 10 "plate appearances" (ie, "trials") what is the probability this player reaches at least 1st base (ie, a "success") exactly 3 times?

P[Exactly 3 "successes" in 10 "PA"s] = P[3 "successes (s)" and 7 "outs (f)" in 10 "PA"s]

P["on-base" = "success" on a single "PA"] = 0.3 — Occurs 3 times, so P[S = 3 in 3] = $(0.3)^3$

P["out" = "failure" on a single "PA"] = 1 – 0.3 = 0.7 — Occurs 7 times, so P[F= 7 in 7] = $(0.7)^7$

Number of ways to arrange 3 Ss & 7 Fs is 10!/(3!*7!) = $\binom{10}{3}$, so …

P[Exactly 3 "successes" in 10 "PA"s] = P[3 in 10] = $\binom{10}{3} * (0.3)^3 * (0.7)^7$

≈ 0.2668

Binomial Probability Distribution

Example: The chips operating communications satellites can fail with sufficient exposure to radiation present in space (eg, due to solar flares, etc). Consequently, these satellites are designed to carry multiple redundant chip sets (chips will only fail if in operation when radiation threshold exceeded).

If the probability of a chip set failure over the lifetime of a satellite is 20%, and the designers have included 5 redundant chip sets in the satellite, what is the probability it will operate for its full lifetime?

P[≥1 Remain Functional] = P[1 Functional] + … + P[5 Functional]

$$= \binom{5}{1}(.8)^1(.2)^4 + \binom{5}{2}(.8)^2(.2)^3 + \ldots + \binom{5}{5}(.8)^5(.2)^0$$

$$= 5*.8*.2^4 + 10*.8^2*.2^3 + 10*.8^3.2^2 + 5*.8^4*.2 + .8^5$$

$$= 0.99968$$

What if due to desire to cut costs, management edicts a design with only 4 chip sets, then what does the probability above become?

P[≥1 Remain Functional] = 1 – P[All Fail]

$$= 1 - .2^4$$

$$= 0.99840$$

Binomial Probability Distribution

Let the random variable X have pmf $f_X(x)$, then for any value x, the **Cumulative Distribution Function** (cdf) is the total probability of all the potential outcomes for X that are less than x. The cdf for a discrete random variable, can be expressed as:

$$P[X \le x] = \sum_{All\ x_i \le x} P[X = xi]$$

Example: What is the probability the baseball player (noted previously) gets "on-base" **<u>at most</u>** 3 times in 10 "plate appearances"?

P[≤ 3 "successes" in 10 "trials"] = P[S ≤ 3 in 10]

= P[S=0 in 10] + P[S=1 in 10] + P[S=2 in 10] + P[S=3 in 10]

$$= \binom{10}{0} *.3^0 *.7^{10} + \binom{10}{1} *.3^1 *.7^9 + \binom{10}{2} *.3^2 *.7^8 + \binom{10}{3} *.3^3 *.7^7$$

$$= .7^{10} + 10*.3*.7^9 + 45*.3^2*.7^8 + 120*.3^3*.7^7$$

$$\approx 0.0282 + 0.1211 + 0.2335 + 0.2668$$

$$\approx 0.650$$

Note: There is an Excel function that calculates Binomial probabilities:

=BINOMDIST(x, n, p, False/True),

where x = number of successes of interest,
n = number of trials considered,
p = probability of success involved, with
False = P[Exactly x successes in n trials] (pmf) and
True = P[At Most x successes in n trials] (cdf)

Binomial Probability Distribution

OR … we could just use an Excel utility to do this for us:

Binomial Probabilities				
	Exact	N	10	Input
	Number	n(Success)	3	Input
	of	P[Success]	0.3	Input
	Successes	P[3 in 10]	0.266828	Output
	At Least	N	10	Input
	Number	n(Success)	4	Input
	of	P[Success]	0.3	Input
	Successes	P[>=4 in 10]	0.350389	Output
	At Most	N	10	Input
	Number	n(Success)	3	Input
	of	P[Success]	0.3	Input
	Successes	P[<=3 in 10]	0.649611	Output

Requires 3 Inputs:

1) N = Number of "Trials" (ie, n)
2) n(Success) = Number of "Successes" (ie, x)
3) P[Success] = Probability of "Success" (ie, p)

Note requires Input for the specific Probability of Interest:

1) **Exact** (ie, P[3 in 10]),
2) **At Least** (ie, P[>=4 in 10], or
3) **At Most** (ie. P[<=3 in 10]

Example: X ~ Bin(n=15, p = 0.6)

a) P[X = 8] = ?
b) P[X < 8] = ?
c) P[X >= 8] = ?

Binomial Probabiilities				
	Exact	N	15	Input
	Number	n(Success)	8	Input
	of	P[Success]	0.6	Input
	Successes	P[8 in 15]	0.177084	Output
	At Least	N	15	Input
	Number	n(Success)	8	Input
	of	P[Success]	0.6	Input
	Successes	P[>=8 in 15]	0.786897	Output
	At Most	N	15	Input
	Number	n(Success)	7	Input
	of	P[Success]	0.6	Input
	Successes	P[<=7 in 15]	0.213103	Output

Binomial Probability Distribution

Mean, Variance, & Standard Deviation

If $X \sim Bin(n, p)$, then

$E[X] = \mu_X = n*p$,

$Var(X) = \sigma_X^2 = n*p*(1-p)$, &

Std Dev(X) = $\sigma_X = \sqrt{n*p*(1-p)}$

Example: The number of times the ballplayer previously described would be ***expected*** to reach at least 1st base in ten "plate appearances" would be …

10*0.3 = 3

Example: The ***expected*** number of chip set failures on the newly designed satellite previously described would be …

4*0.2 = 0.8 (actually less than the old design: 5*0.2 = 1)

The variance of failures would be …

4*0.2*0.8 = 0.64

The standard deviation of failures would be …

$\sqrt{0.64}$ = 0.8

Hypergeometric Probability Distribution

Let a random variable Y be the number of "successes" in n dependent Bernoulli trials, where there are N possible outcomes, s of which are "successes" and (N-s) are not, then Y has a **Hypergeometric Probability Distribution**, and the pmf can be expressed as

$$f_Y(y) = P[Y=y] = \frac{\binom{s}{y}\binom{N-s}{n-y}}{\binom{N}{n}}, \quad 0 \le y \le min(n, s),$$

where $\binom{a}{b}$ is the combination of a things taken b at a time $= \frac{a!}{b!(a-b)!}, \quad 0 \le b \le a,$

Example: You are throwing a party and need to purchase 5 dozen eggs. There are 50 cartons each with a dozen eggs at the store, but unbeknownst to you, 5 of these have at least one broken egg. You are "successful" if you choose a carton with no broken eggs, but your P["success"] changes after each choice you make (ie, put a carton in your cart).

Each "choice" you are either "successful" in getting no broken eggs, or not, so these are Bernoulli trials (recall – a Bernoulli random variable is one with only 2 outcomes – "0 broken" or ">=1 broken" in this case), but they are NOT independent since the P["0 broken"] changes with each choice/trial.

So in choosing 5 cartons (ie, trials") what is the probability you end up with exactly 2 cartons with some broken eggs?

Number of ways to choose 2 damaged cartons from the 5 available $\binom{5}{2}$ = 5!/(2!*3!)

Number of ways to choose 3 cartons from the 45 undamaged ones $\binom{45}{3}$ = 45!/(3!*42!)

Number of ways to choose any 5 cartons from the 50 available = $\binom{50}{5}$ = 50!/(5!*45!), so …

P[Exactly 3 "successes" in 5 "choices"] = P[3 in 5] = $\binom{45}{3}\binom{5}{2} / \binom{50}{5}$

≈ 0.067

Hypergeometric Probability Distribution

Example: At the end of a production line for video gaming equipment, lots of 100 units are sampled for testing prior to shipment. If no issues are discovered with the sampled units, then the entire lot is shipped on to distributors. However, if any sampled unit is found to have a performance issue, then the entire lot is tested before any units are shipped.

If the current test sample size is 3 units from each lot, then what is the probability a lot with 5 defective units will ship with the defectives remaining undetected and included?

P[Lot Ships with 5 defective units] = P[All 3 Sampled Units OK]

= (# ways to get 3 Good)*(# ways to get 0 Bad)/
(# ways to sample 3)

$$= \frac{\binom{95}{3}\binom{5}{0}}{\binom{100}{3}}$$

$$\approx 0.8560$$

What if due to customer complaints management edicts an increase in sampling to 10 units per lot, but due to a concern about costs, also moves the number of defectives required to fully test the lot to two or more problem units in the sample. Now what is the probability a lot with 5 defective units will ship with 5 or 4 defectives?

P[Lot Ships with 5 or 4 defectives] = P[All 10 Pass] + P[Exactly 1 Fails]

$$= \frac{\binom{95}{10}\binom{5}{0}}{\binom{100}{10}} + \frac{\binom{95}{9}\binom{5}{1}}{\binom{100}{10}}$$

Hypergeometric Probability Distribution

So is management … misguided?

Well … lets look a little more closely …

Lets consider the following table:

Approach	Number Defectives Found				Approach	Number Defectives Found			
Sample = 3	0	1	>1	Expected	Sample = 10	0	1	>1	Expected
Prob	0.856	0.138	0.006	Value	Prob	0.584	0.339	0.077	Value
# Def Shipped	5	0	0	4.2800	# Def Shipped	5	4	0	4.2763
# Units Tested	3	100	100	16.9681	# Units Tested	10	10	100	16.9171
% Def Shipped	5%	0%	0%	4.280%	% Def Shipped	5%	4.040%	0%	4.290%

So … slightly less expected defectives per lot

And … slightly less expected units tested per lot

But … slightly higher expected percentage of defectives shipped

So … change is really just about a “wash” (ie, not really different than before),

But … management looks like it is doing something ☺

Hypergeometric Probability Distribution

Mean, Variance, & Standard Deviation

If X ~ Hypgeom(s,f, n), then

$E[X] = \mu_X = n*[s/(s + f)] = n*(s/N)$, where $N = s + f$

$Var(X) = \sigma_X^2 = n*(s/N)*(f/N)*[(N-n)/(N-1)]$, &

Std Dev(X) = $\sigma_X = \sqrt{n*(s/N)*(f/N)*[(N-n)/(N-1)]}$

Example: The ***expected*** number of damaged cartons chosen when choosing 5 from a group of 50 including 5 damaged cartons would be …

5*(5/50) = 0.5 (Note: Expected number of undamaged = 5*(45/50) = 4.5)

Example: The ***expected*** number of defective units chosen when choosing 3 from a lot of 100 including 5 defectives would be …

3*0.05 = 0.15 (Note: Expected number of working units = 3*0.95 = 2.85)

The variance of number of defectives would be …

3*0.05*0.95*(97/99) ≈ 0.1396

The standard deviation of failures would be …

$\sqrt{0.1396}$ ≈ 0.3737

Poisson Distribution

One of the stories surrounding the development of the Poisson distribution is that it was developed by a Frenchman (Poisson) as part of a study of the frequency of horses killing soldiers in the process of re-shoeing them during the Napoleonic wars.

Recall that $e^x = 1 + x + x^2/2! + x^3/3! + \ldots = \sum_{k=0 \text{ to } \infty} x^k/k!$,

so $p(x) = \lambda^x e^{-\lambda}/x!$, $x = 0, 1, \ldots$ defines a pmf for a random variable X, and

such a pmf is called a ***Poisson*** pmf with parameter λ.

Random variables that have Poisson pmfs not only include the number of soldiers kicked in the head by their horses. Such random variables tend to appear in other situations as well:

- number of accidents occurring in a given space and time,
- number of defectives produced for a given process over a given amount of time,
- number of defects on a given product,
- number of calls received at a given telephone number over a given period of time,
- number of customers entering a place of business for a given period of time.

There are generally three postulates that give rise to Poisson processes:

1. The probability of an occurrence of interest actually occurring in a fixed time interval of length t is approximately proportional to the length of the interval (ie, $\approx \lambda t$, where λt is the parameter of the respective Poisson pmf)
2. The probability of 2 or more occurrences in this interval is essentially zero.
3. Occurrences in non-overlapping intervals are mutually independent of each other.

Poisson Distribution

A property of a Poisson random variable X is that E[X] = Var(X), which is fairly restrictive.

However, because of this property, the parameter of a Poisson pmf is often denoted as μ, the common symbol used for the mean of any random variable.

As with the sum of independent Binomial random variables, it can be shown that the sum of independent Poisson random variables, $Y = \sum_{k=1}^{n} Xk$, $X_k \sim \text{Poisson}(\mu_k)$ is also $\text{Poisson}(\sum_{k=1}^{n} \mu k)$.

Example: Assume that the number of passes thrown by Patrick Mahomes in a specific game has a Poisson pmf. Then the pmf can be expressed as:

$$p(x) = e^{-\mu}\mu^x/x!,\ x = 0, 1, \ldots$$

As of 9/20/23, Mahomes had thrown 3,596 passes in 96 regular season games, an average of ~34 per game.

So, setting μ = 34 (large for most Poisson applications), an estimate of the probability that Mahomes will throw 20 passes in his next game would be given as

$$P[X = 20] = e^{-34}*(34)^{20}/20! \approx 0.003$$

Clearly, some of these values are large (eg, $20! \approx 2.433 \times 10^{18}$), so such probabilities are generally obtained via computer. As with the Binomial distribution, there is also an Excel function for the Poisson distribution: =POISSON(x, μ, T/F) where x is the number of occurrences of interest, μ is the applicable Poisson parameter, and T/F, as for the Binomial, is FALSE for the pmf values and TRUE for the cdf results.

Poisson Distribution

Example (continued): So the probability Mahomes throws 20 passes in his next game is approximately 3 in 1000. A more interesting probability might be

$$P[\text{Mahomes throws fewer than 20 passes in his next game}] = P[X < 20] \approx 0.00675$$

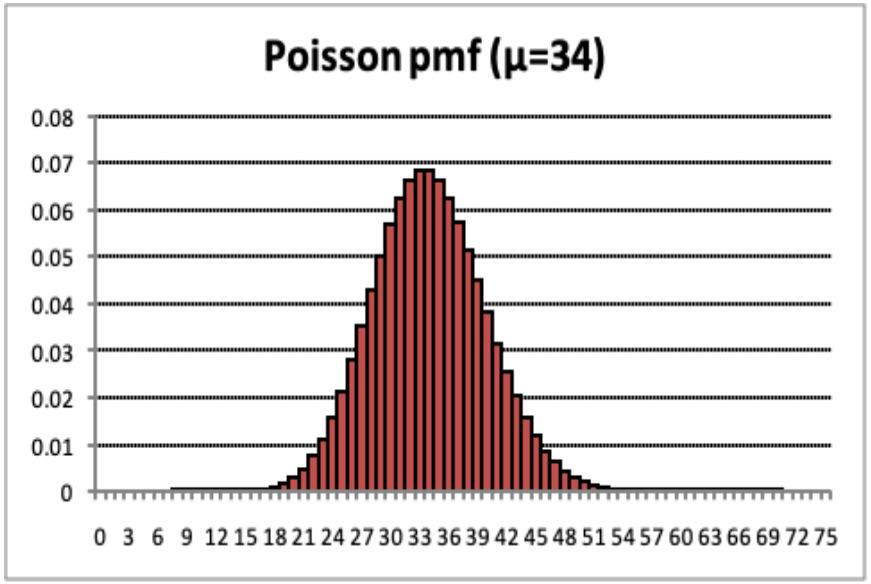
This is still small, and the probability of real interest to a bettor and more important to his bookie would be the number of passes k where $P[X < k] \approx P[X > k] \approx 0.5$, as the value of k represents a reasonable value at which to set a standard over-under bet.

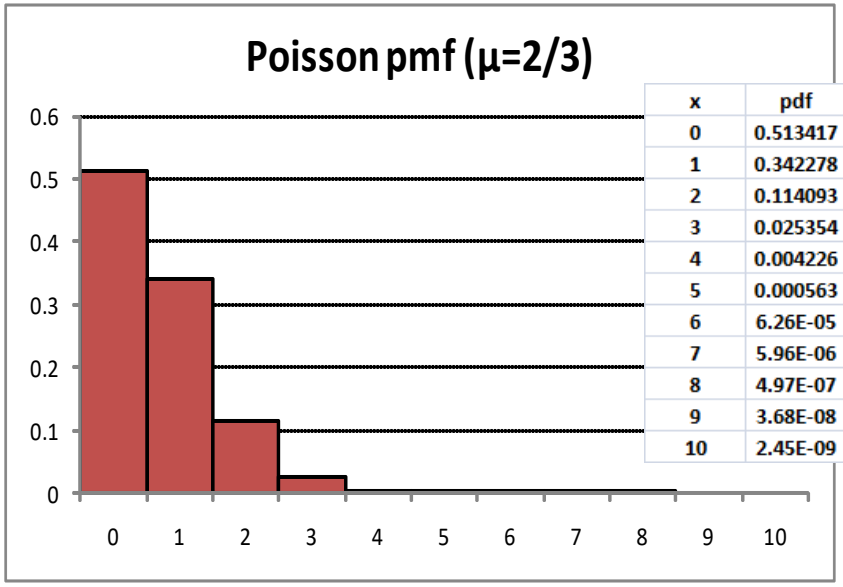
It turns out for the assumptions made here, $k = 34$ results in $P[X < 34] \approx 0.477$ & $P[X > 34] \approx 0.455$.

Does it make sense that $k = 34$ is near the median of this distribution?

Would it be expected that the median of any Poisson(μ) random variable would be = μ?

Why might it be expected in this case?





x	pdf	cdf
0	0.513417	0.513417
1	0.342278	0.855695
2	0.114093	0.969788
3	0.025354	0.995142
4	0.004226	0.999367
5	0.000563	0.999931
6	6.26E-05	0.999994
7	5.96E-06	0.999999
8	4.97E-07	1
9	3.68E-08	1
10	2.45E-09	1

Poisson Distribution

Example (continued): So one might be interested in the probability Mahomes will throw at least one touchdown pass in his next game. If we assume the number of passes thrown in the game (X) has a Poisson(μ=34) pmf and then, of these, the number of touchdown passes thrown in the game (Y) has a b(n=X, p=0.065) pmf, then

$$P[Y > 1] = 1 - P[Y = 0] = 1 - \sum_{x=0}^{\infty} e^{-34}\frac{34^x}{x!}\binom{x}{0}(0.065)^0(0.935)^x$$

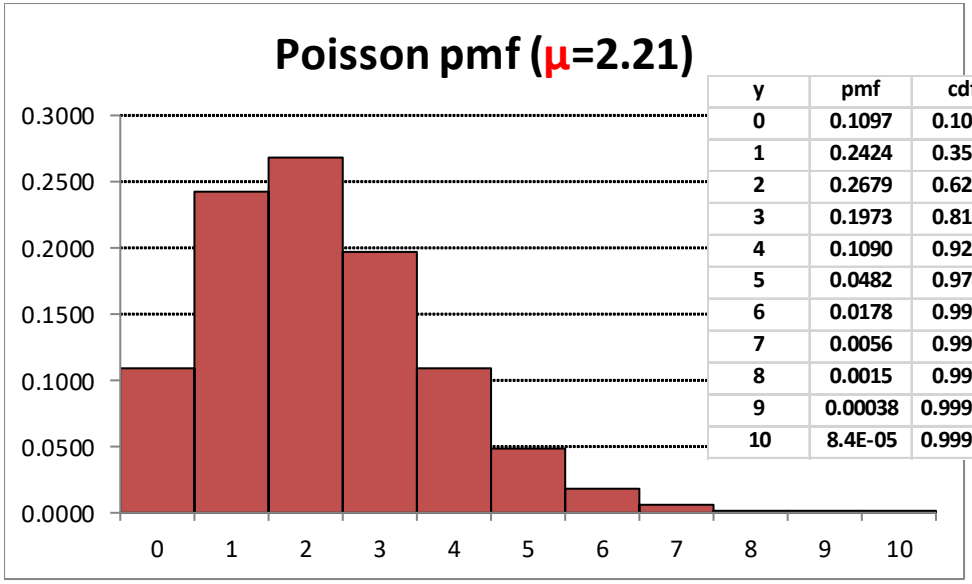
$$= 1 - e^{-34(0.065)}\sum_{x=0}^{\infty} e^{-34(0.935)}\frac{[34(0.935)]^x}{x!}$$

$$= 1 - e^{-2.21} \approx 0.89$$

Again, to set a reasonable value for an over-under bet on the number of touchdown passes Mahomes will throw in his next game, it would be desirable to know k where

$$P[Y < k] \approx P[Y > k] \approx 0.5$$

The marginal pmf for Y is Poisson(μ=34*0.065=2.21), and the pmf for Y is as seen below:



y	pmf	cdf
0	0.1097	0.1097
1	0.2424	0.3521
2	0.2679	0.6200
3	0.1973	0.8174
4	0.1090	0.9264
5	0.0482	0.9746
6	0.0178	0.9924
7	0.0056	0.9980
8	0.0015	0.9995
9	0.00038	0.999895
10	8.4E-05	0.999979

So where does the bookie set the over-under value?

If set at Y=2, then if Mahomes throws 2 TD passes, no one wins. The chances of this occurring appear to be ~ 1 in 4.

However, the bookie's exposure is approximately the same on either side of the bet, as P[Under Wins] = 0.3521 and P[Over Wins] = 0.38.

Mahomes threw 33 passes of which 3 were touchdown passes in his next game in win vs Bears ("Over" wins).

Continuous Probability Density Function

While Discrete Random Variables have a Probability Mass Function (pmf), the differences between countable and uncountable sets (recall – discrete random variables are defined on countable sets, continuous on uncountable sets) no longer allow the use of a pmf.

Instead, for Continuous Random Variables, we utilize a Probability Density Function (pdf). Let X be a continuous random variable with pdf $f_X(x)$, then

1) $f_X(x) >= 0$ for all applicable values x,

2) P[a <= X <= b] is the area under the graph of $f_X(x)$ for the interval [a, b], and

3) the total area under the graph of $f_X(x) = 1$.

STAT 5340
Statistical Analysis I

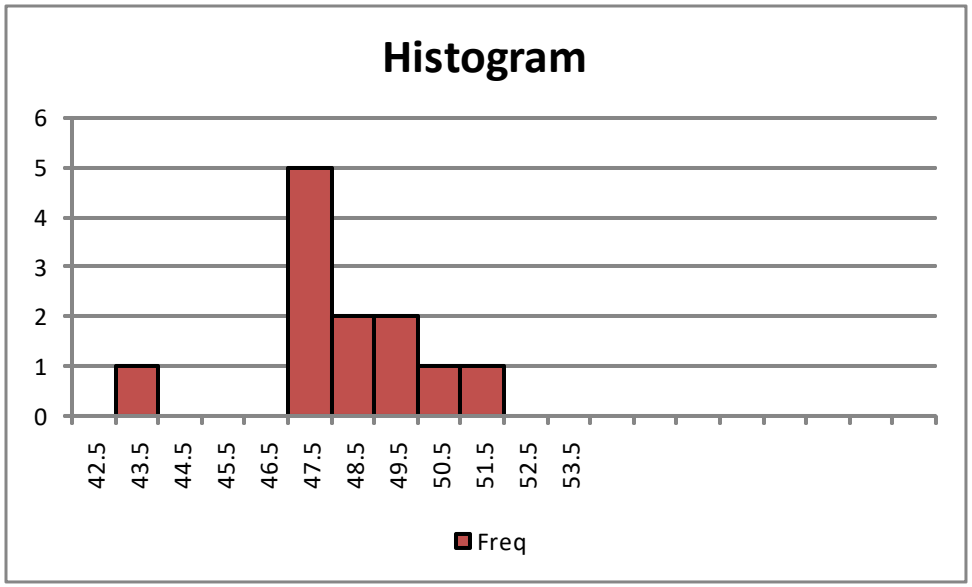
Graphical & Numerical Summaries
Of Data

Plot/Graph Your Data!

The Human Eye-Brain Combination is perhaps the best Pattern Recognition Processor in the known Universe

Consider the example of Hours Worked by Java Developers by Location

Region	Hours Worked
U.S.	48
Northeast	47
Mid-Atlantic	49
South	47
Midwest	47
Central Mt	51
California	50
Pacific NW	47
Canada	43
Europe	48
Asia	47
South America and Africa	49



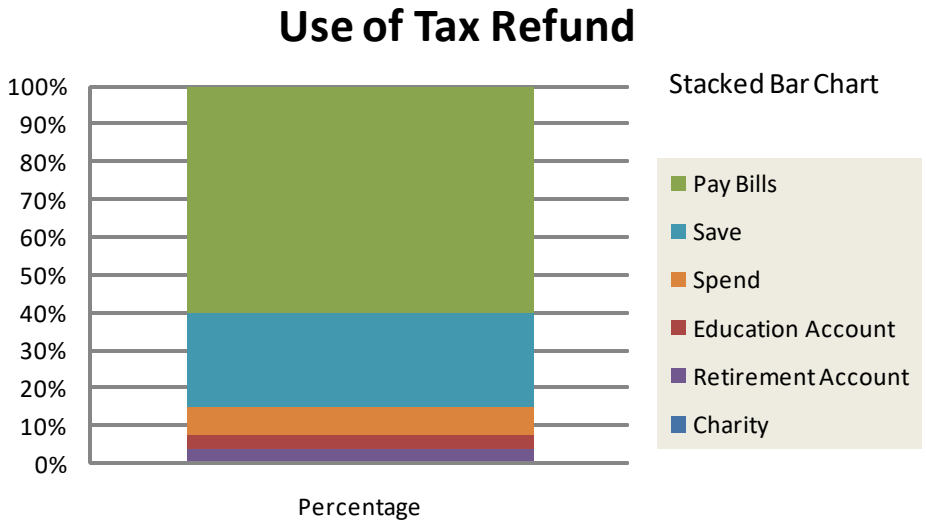
The low value for Canada is not as Obvious in the Table as it is in the Plot

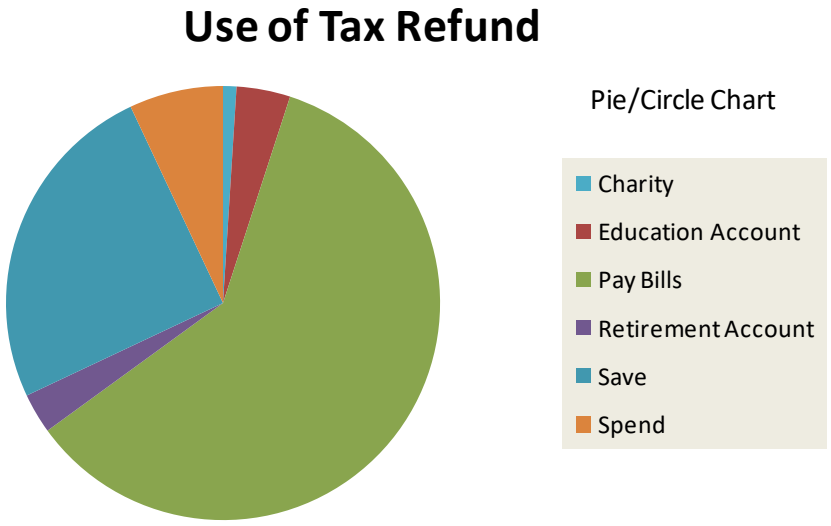
Plots & Graphs are generally of more value as data sets get larger

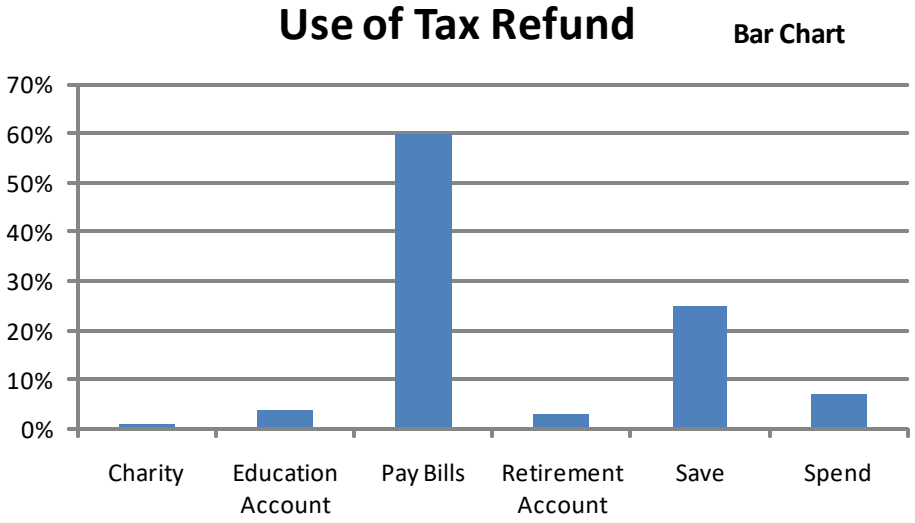
Graphs for Qualitative Data

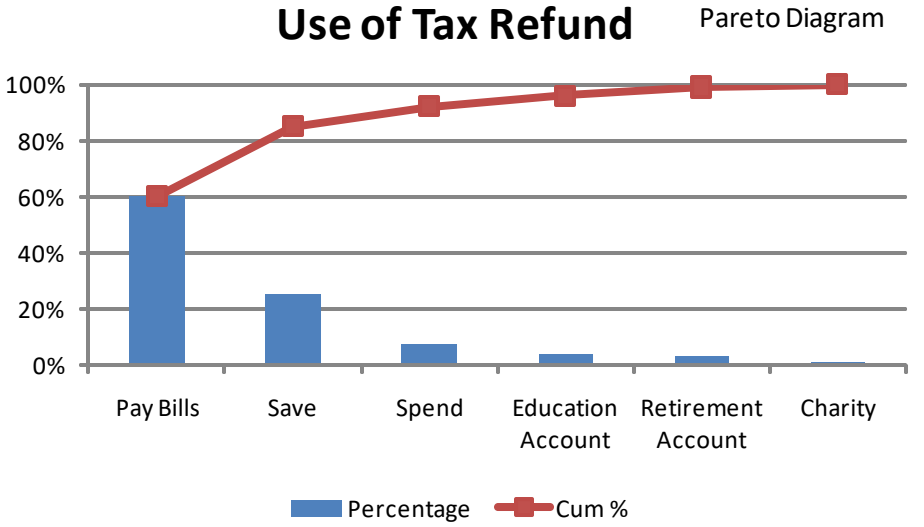
Graphs/Plots for Qualitative Data are generally more limited

Recall Qualitative Data carries Less Information than Quantitative Data





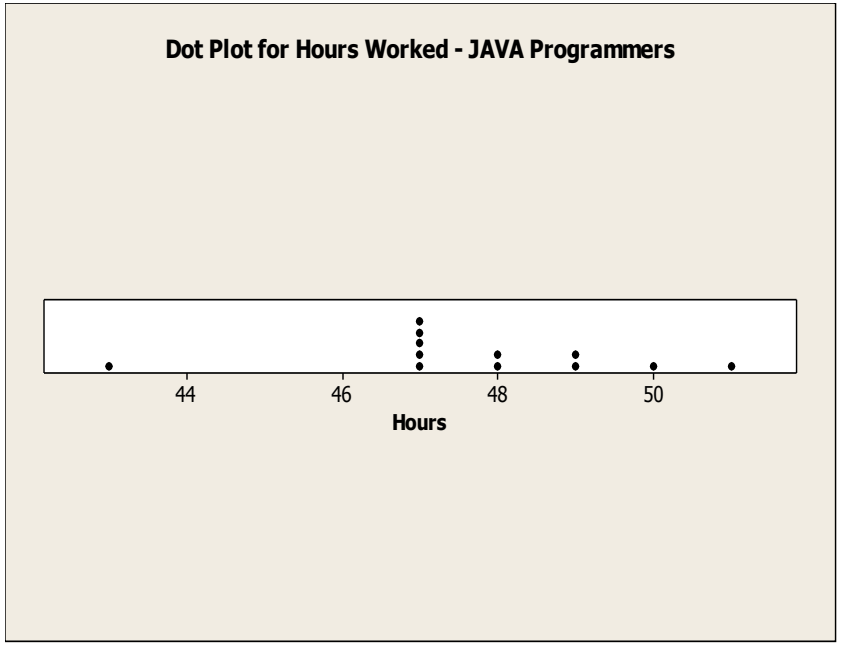




Graphs for Quantitative Data

Smaller Data Sets (say N < 40)

Dot Plots – Display All the Data



Stem & Leaf Diagrams – Display All the Data Both Graphically and Numerically

Stem-and-Leaf Display: Hours

```
Stem-and-leaf of Hours
N   = 12
Leaf Unit = 0.10

 1   43   0
 1   44
 1   45
 1   46
 6   47   00000
 6   48   00
 4   49   00
 2   50   0
 1   51   0
```

Graphs for Quantitative Data

For Small Data Sets (say N<40)

- Dot Plots – display all data values
- Stem-and-Leaf Displays – display all data values graphically and numerically
- Special Cases of Histograms

For Larger Data Sets

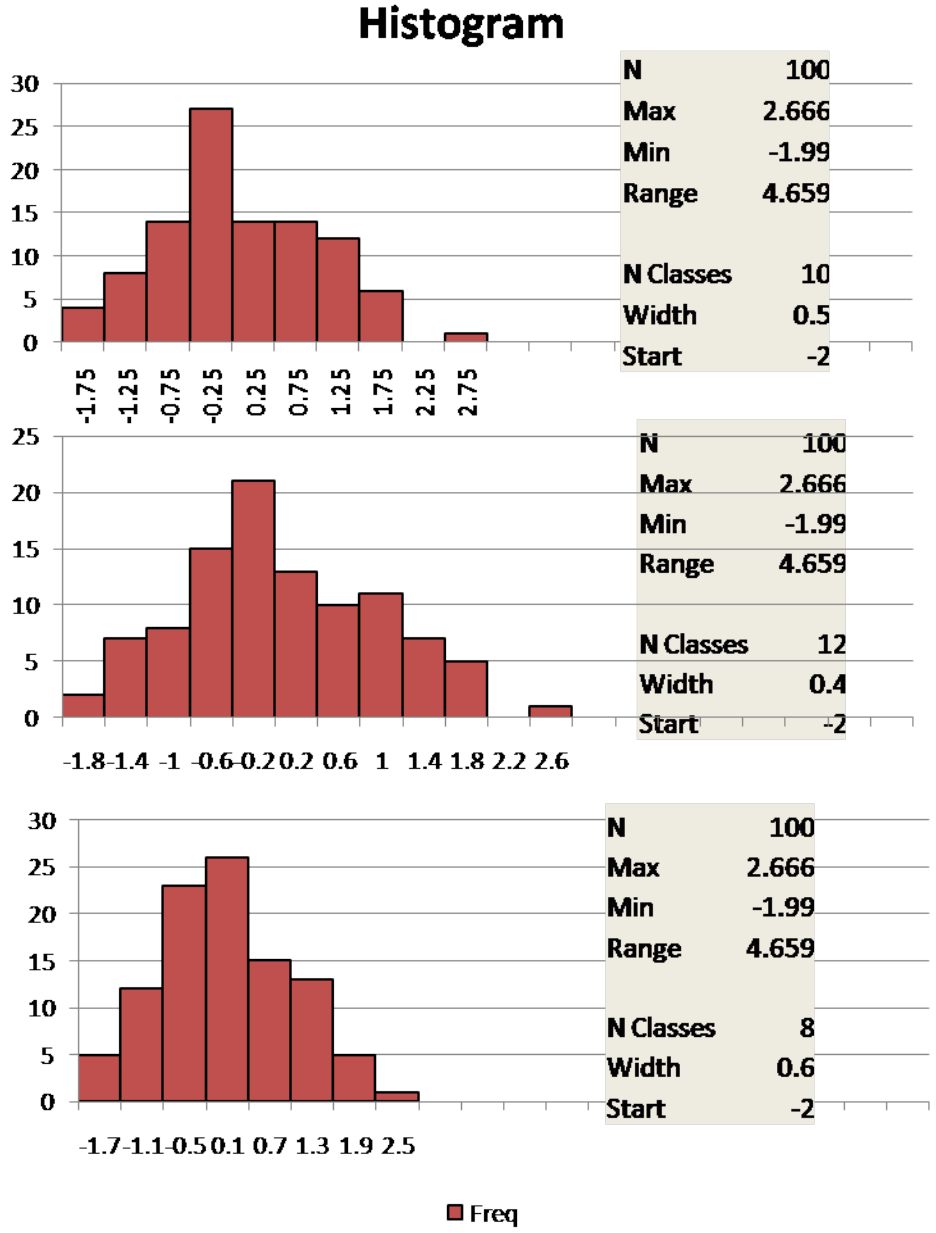
- Histograms
- Cumulative Frequency Plots (Ogives)

Constructing Histograms

1) Find N, Max, Min, Range
2) Choose
    1) Number of Classes (m)
        1) m~ Sqrt(N)
        2) Usually <=20
    2) Class Width (c)
        1) m*c > Range
        2) "Even" width if possible
    3) Start Point (< Min)

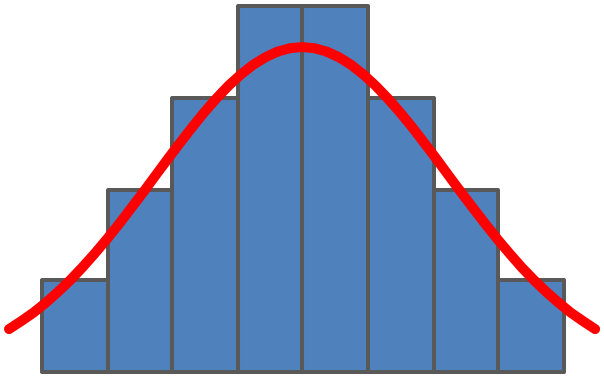
Guidelines

1) Classes all same width
2) No Overlap
3) Data Values Belong to Only 1 Class

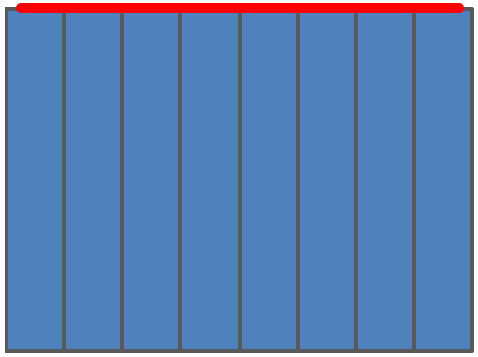


Histograms Suggest Population Distributions

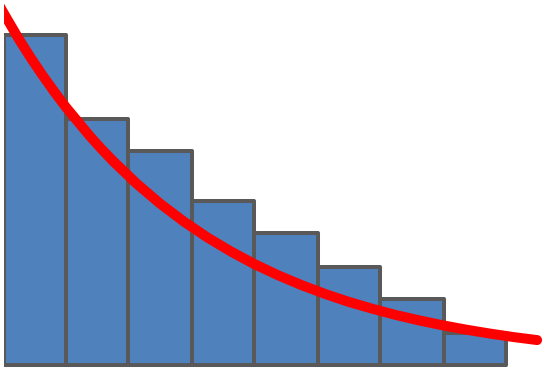
Normal



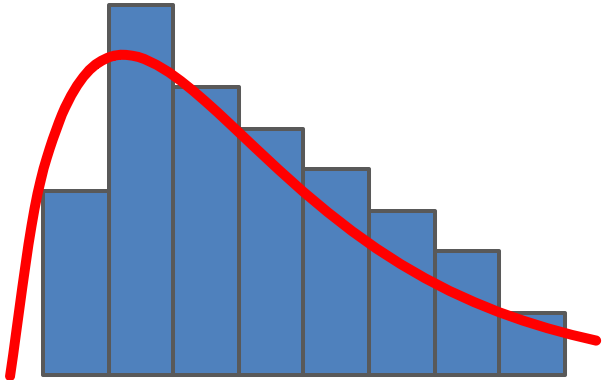
Uniform



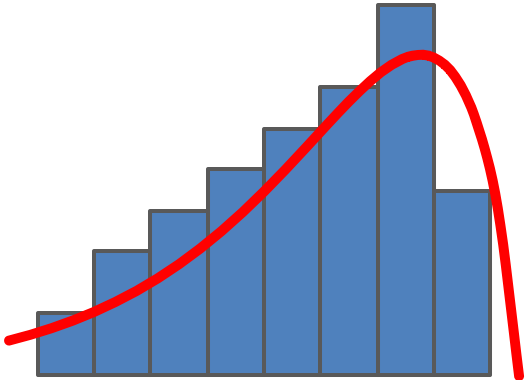
J-Shaped



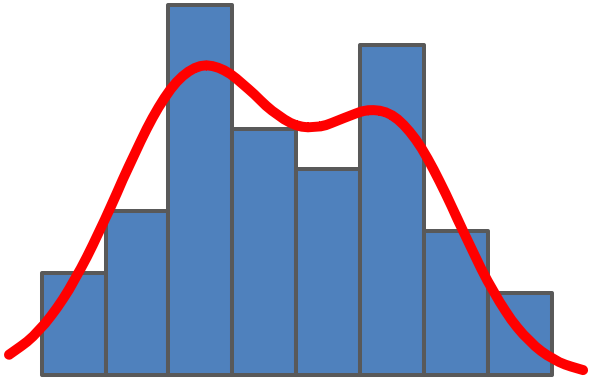
Skewed Right



Skewed Left



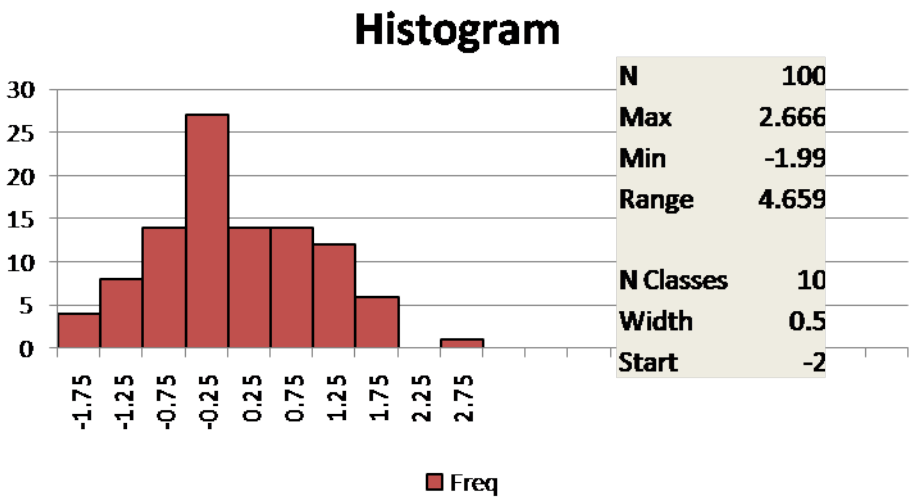
Bi-Modal

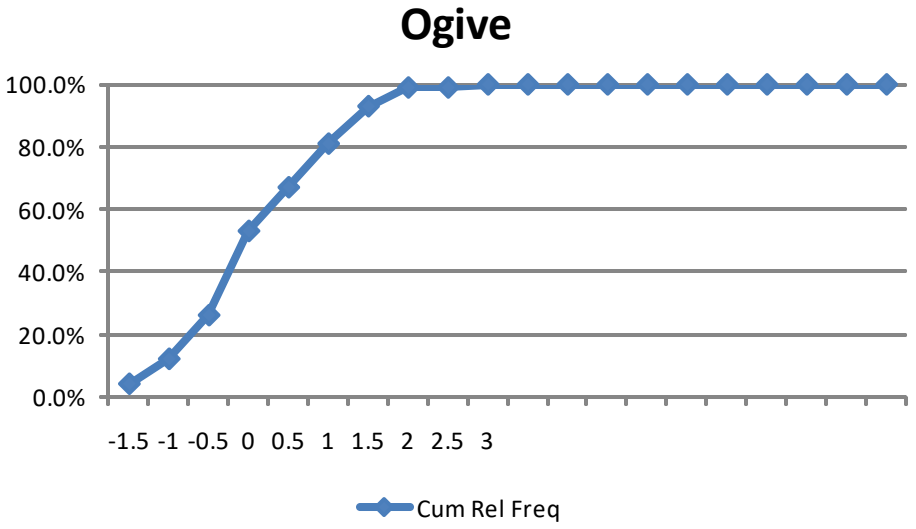


Histograms & Ogives

Class Information

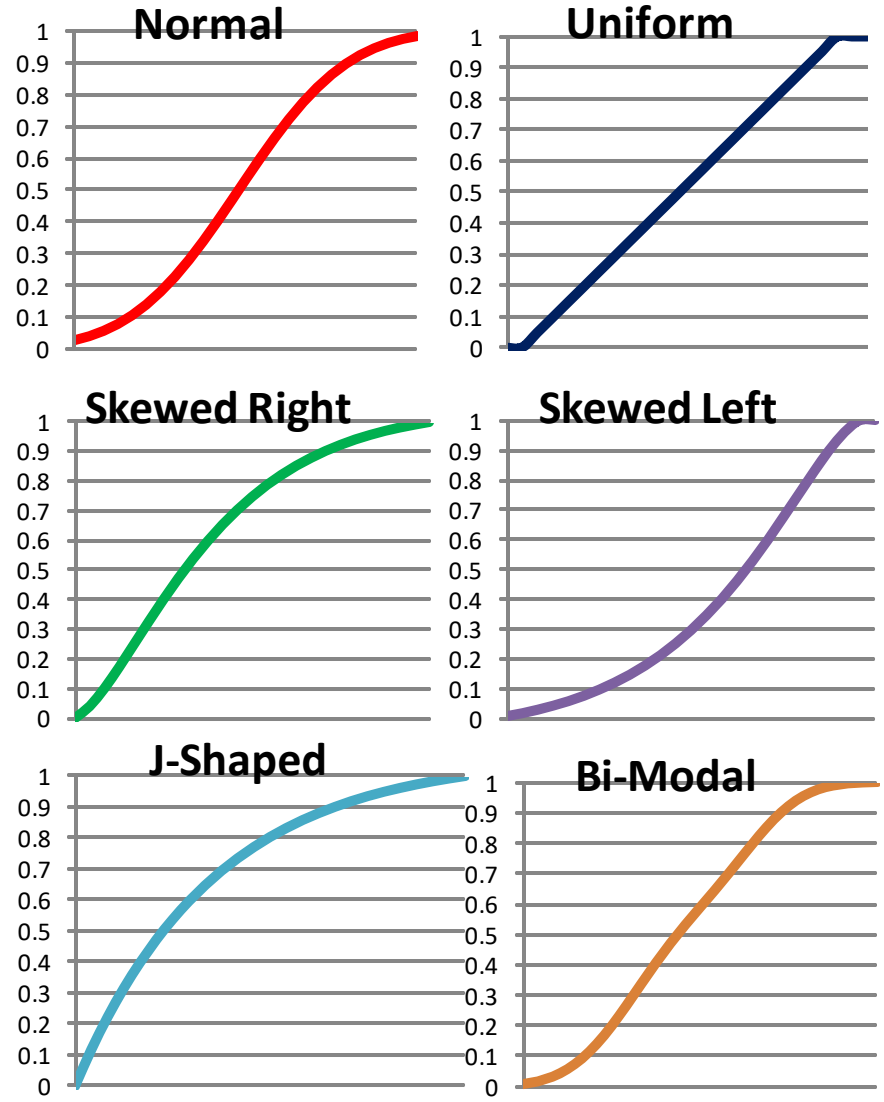
Boundary Lo	Mark	Boundary Hi	Freq	Cum Freq	Cum Rel Freq
-2	-1.75	-1.5	4	4	4.0%
-1.5	-1.25	-1	8	12	12.0%
-1	-0.75	-0.5	14	26	26.0%
-0.5	-0.25	0	27	53	53.0%
0	0.25	0.5	14	67	67.0%
0.5	0.75	1	14	81	81.0%
1	1.25	1.5	12	93	93.0%
1.5	1.75	2	6	99	99.0%
2	2.25	2.5	0	99	99.0%
2.5	2.75	3	1	100	100.0%





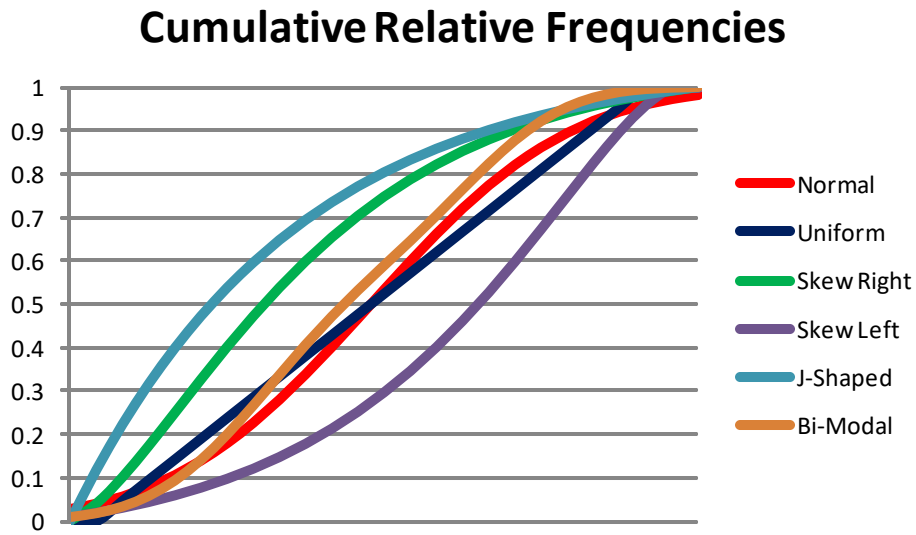
An Ogive is more commonly Called a Cumulative Frequency or Cumulative Relative Frequency Plot

Cumulative Relative Frequency Plots



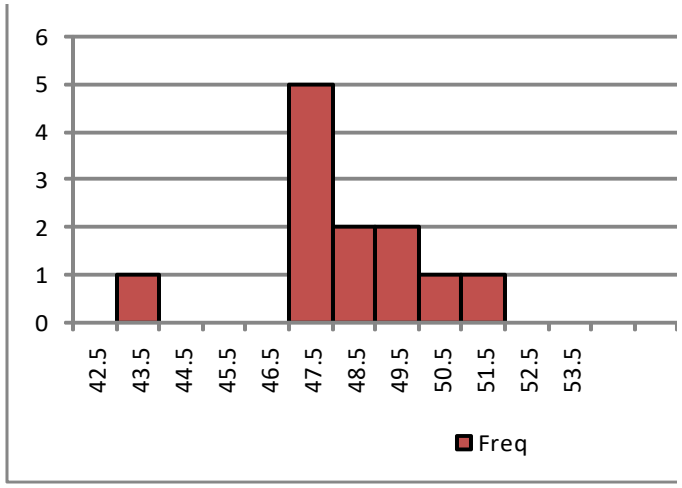
Normal – S-shaped
Uniform – Straight Line
Skewed – Curved

Multiple Distributions can be plotted on a single chart for Comparison



Measures of Central Tendency

Region	Hours Worked
U.S.	48
Northeast	47
Mid-Atlantic	49
South	47
Midwest	47
Central Mt	51
California	50
Pacific NW	47
Canada	43
Europe	48
Asia	47
South America and Africa	49



Mean = Average
- Add all the values = 573
- Divide by the number of values = 12
- Average = 573/12 = 47.75

Median = Middle Value
- Order the data lowest to highest
  43,47,47,47,47,47,48,48,49,49,50,51
- Odd number of data points – choose middle value
- Even number of data points – average middle two
  12 data points so average 6th & 7th value
  (47+48)/2 = 47.5

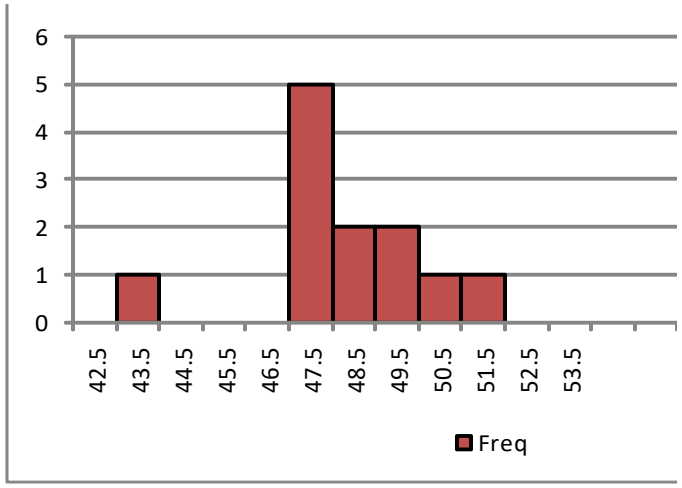
Mode = Most Frequently Occurring Value
- Self-explanatory, there are 5 47's, so 47

Mid-Range = Midpoint of the Extremes
- Average of smallest and largest data values
  (43 + 51)/2 = 47

Mean (X-bar) is most commonly used, Median is next most common

Measures of Dispersion

Region	Hours Worked
U.S.	48
Northeast	47
Mid-Atlantic	49
South	47
Midwest	47
Central Mt	51
California	50
Pacific NW	47
Canada	43
Europe	48
Asia	47
South America and Africa	49



Range = Largest – Smallest Data Value
- Range = 51 -43 = 8

Mean Absolute Deviation from the Mean
- Subtract the Mean from each data value, and
- Take Absolute Value of result (make positive)
  .25,.25,.75,.75,.75,.75,.75,1.25,1.25,2.25,3.25,4.75
- Find Mean of resultant data values (1.4167)

Median Absolute Deviation from the Median (MAD)
- Subtract the Median from each data value, and
- Take Absolute Value of result (make positive)
  .5,.5,.5,.5,.5,.5,.5,1.5,1.5,2.5,3.5,4.5
- Find Median of result (0.5)
- Divide by 0.6745, scalar to ~1 standard deviation (0.7143)

Variance
- Sum Squared Deviations from the Mean = 44.25
- Divide by Number of data values – 1 = 12 -1 = 11
- Variance = 44.25/11 = 4.022727

NOTE: Can also be expressed as:

{Sum of Squares – $n(Mean)^2$}/(n-1)

Standard Deviation = Square Root of the Variance
- Sqrt(Variance) = Sqrt(4.022727) = 2.005674

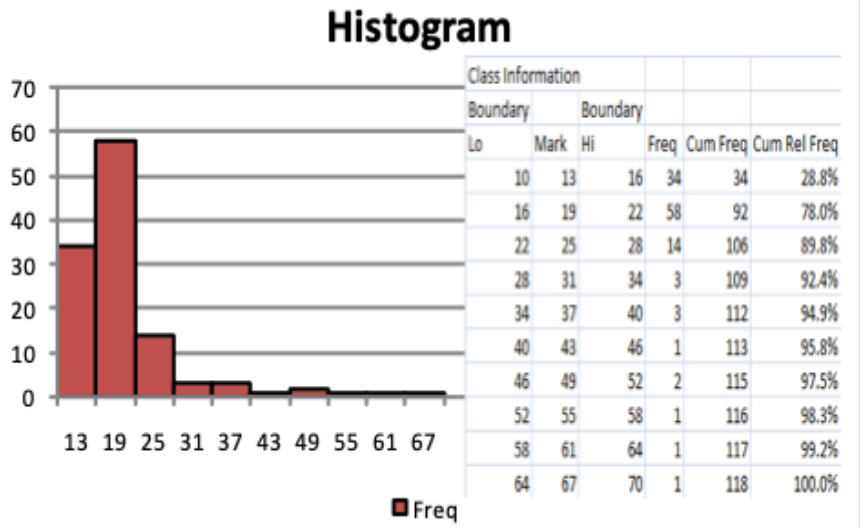
Standard Deviation (S) is Most Commonly Used, Range also Common

Statisticians Like Variances (S^2) because they are additive, S is not

Measures of Position

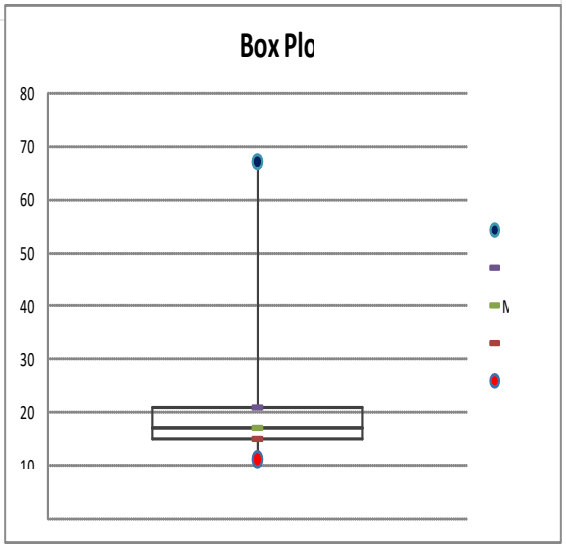
Auto Theft Offenders Garden City, Michigan

Age	Z-Scores
11	-0.975
12	-0.867
13	-0.759
13	-0.759
13	-0.759
13	-0.759
13	-0.759
13	-0.759
14	-0.651
14	-0.651
⋮	⋮
31	1.183
34	1.507
36	1.723
39	2.047
43	2.478
46	2.802
50	3.234
54	3.666
59	4.205
67	5.069



Class Information					
Boundary		Boundary			
Lo	Mark	Hi	Freq	Cum Freq	Cum Rel Freq
10	13	16	34	34	28.8%
16	19	22	58	92	78.0%
22	25	28	14	106	89.8%
28	31	34	3	109	92.4%
34	37	40	3	112	94.9%
40	43	46	1	113	95.8%
46	49	52	2	115	97.5%
52	55	58	1	116	98.3%
58	61	64	1	117	99.2%
64	67	70	1	118	100.0%

Mean	20.0339
Median	17
Mode	16
Q1	15
Q3	20.75
P10	14
P95	39.6
Variance	85.86209
Std Dev	9.26618
Range	56
Mean Abs Dev	5.762425
MAD	2.965159
IQR	5.75



Quartiles-

- 1st Quartile (Q1)
  - Value with 25% of the Data below it & 75% above
- 2nd Quartile (Q2) = Median
  - Half data below & half above
- 3rd Quartile (Q3)
  - Value with 75% of the Data below it & 25% above
- Inter-Quartile Range (IQR)
  - Q3-Q1
  - Includes middle 50% of Data
  - Can be divided by 1.35 to approximate 1 standard deviation

Percentiles-

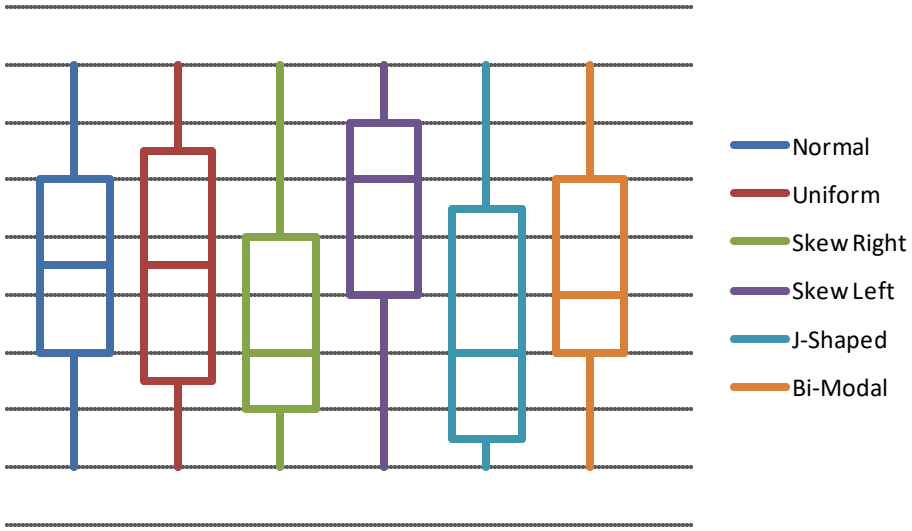
- Analogous to Quartiles
- k^{th} Percentile (P_k)
  - Value with k% of the Data below it & (1-k)% above

Z-Scores-

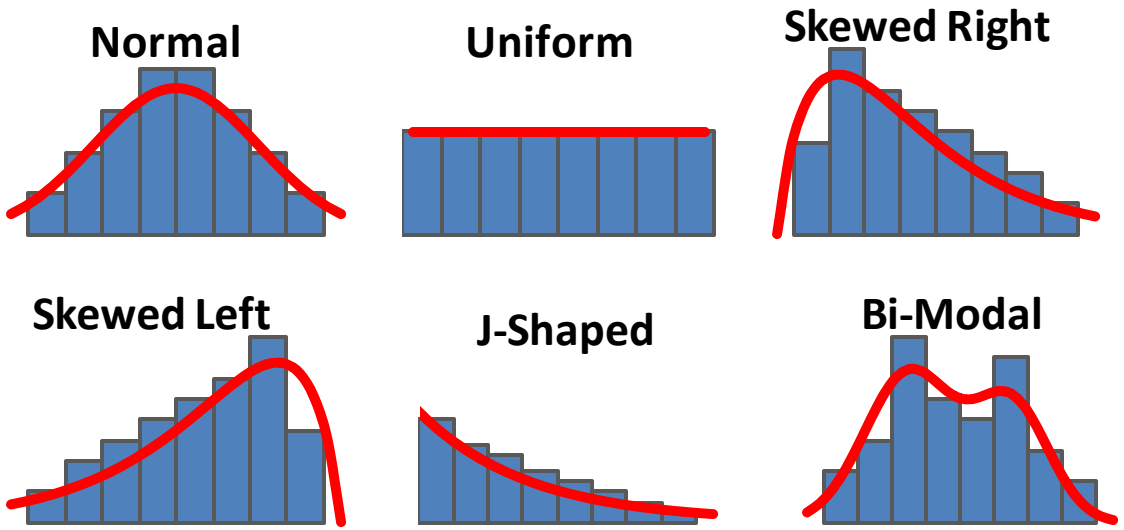
- Number of Standard Deviations from Mean
  - (Value – X-bar)/S

Measures of Position

5-Value Summary – Box Plots



Box (& Whisker) Plots formed using
5 values from the sample data
- Box – top = Q3, bottom = Q1
- Cross-bar in Box = Median
- Whiskers – top = Max, bottom = Min



Box Plots do not suffer
From same subjectivity
As Histograms, Can be
Plotted with fewer Data
Values, and Many Can be
Plotted on the same Chart

Using Measures of Position to Identify Outliers

Outliers = Fliers tend to show up in most data sets

- Do not want them to overly influence the summary and conclusions for the majority of the results
- May carry important information, so source for unusual results should be investigated

Often, flier limits are established using multiples of the Inter-Quartile Range (IQR)

One common set of flier limits is given by:

- UFL = Q3 + 2*IQR
- LFL = Q1 – 2*IQR

Should erroneously throw out a valid result only about 7 in 10,000 data results (~0.07%)

If outside these limits, then value is likely to be different than the majority of the data

Bivariate Data

Bivariate Data is comprised of two results where there is some common bond or link between the results, eg:

- HDL and LDL cholesterol from the same individual
- Revenue and Earnings per Share for the same company
- Tumor size before and after a specific treatment protocol for a specific patient
- Throughput and Yield for a specific manufacturing line
- Soil acidity and moisture levels for a specific plot of land
- Strikeouts and Earned Run Average for a specific pitcher
- Many others

Generally interested in the relationship between the two variables comprising a bivariate data set.

Summaries by Types of Variables

Two Qualitative Variables

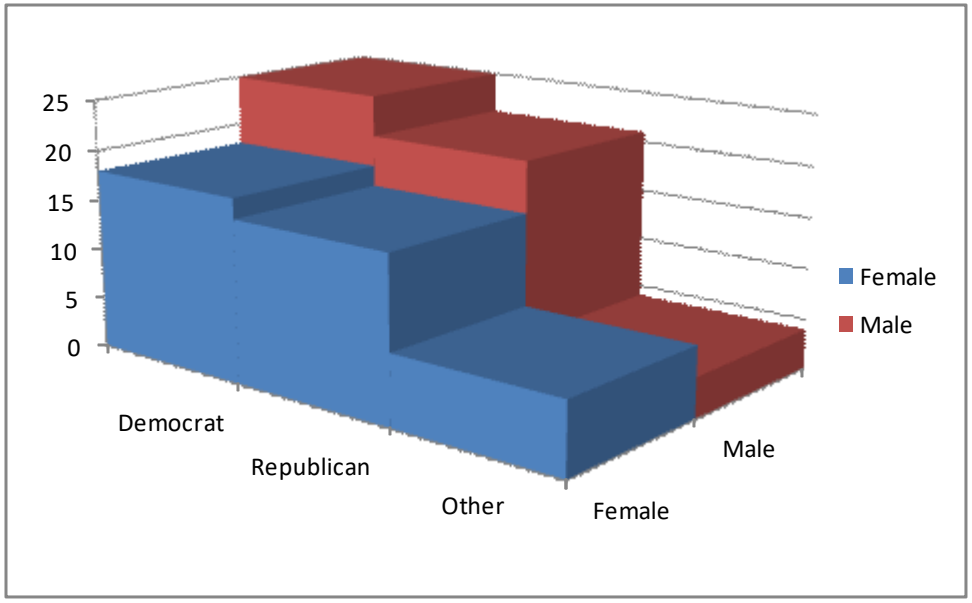
Cross-Tabulation Tables

Political Affiliation	Gender		
	Female	Male	All
Democrat	18	25	43
Republican	16	21	37
Other	7	4	11
All	41	50	91

By Count

Political Affiliation	Gender		
	Female	Male	All
Democrat	19.8%	27.5%	47.3%
Republican	17.6%	23.1%	40.7%
Other	7.7%	4.4%	12.1%
All	45.1%	54.9%	100.0%

By Percentage

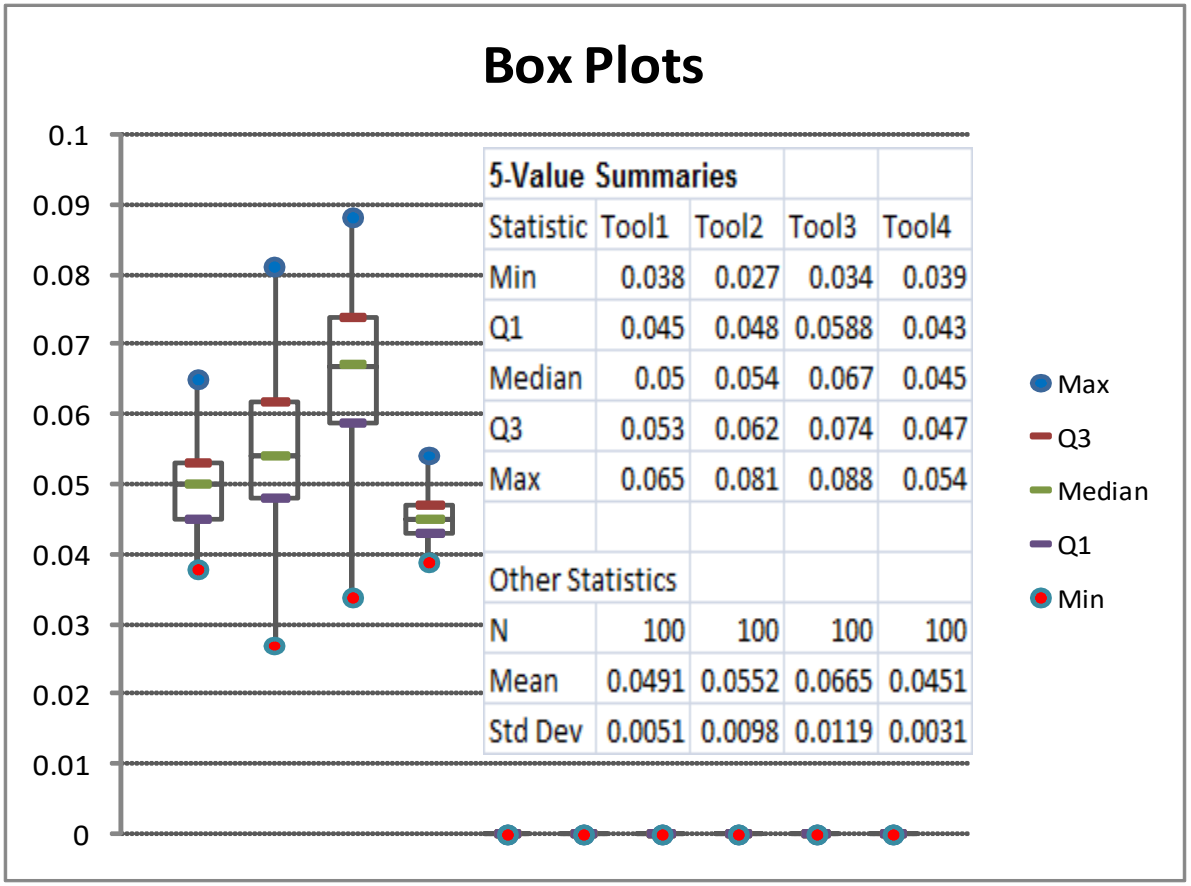


Bivariate Bar Chart

Summaries by Types of Variables

One Qualitative & One Quantitative Variable

Side-by-Side Box Plots



5-Value Summaries				
Statistic	Tool1	Tool2	Tool3	Tool4
Min	0.038	0.027	0.034	0.039
Q1	0.045	0.048	0.0588	0.043
Median	0.05	0.054	0.067	0.045
Q3	0.053	0.062	0.074	0.047
Max	0.065	0.081	0.088	0.054
Other Statistics				
N	100	100	100	100
Mean	0.0491	0.0552	0.0665	0.0451
Std Dev	0.0051	0.0098	0.0119	0.0031

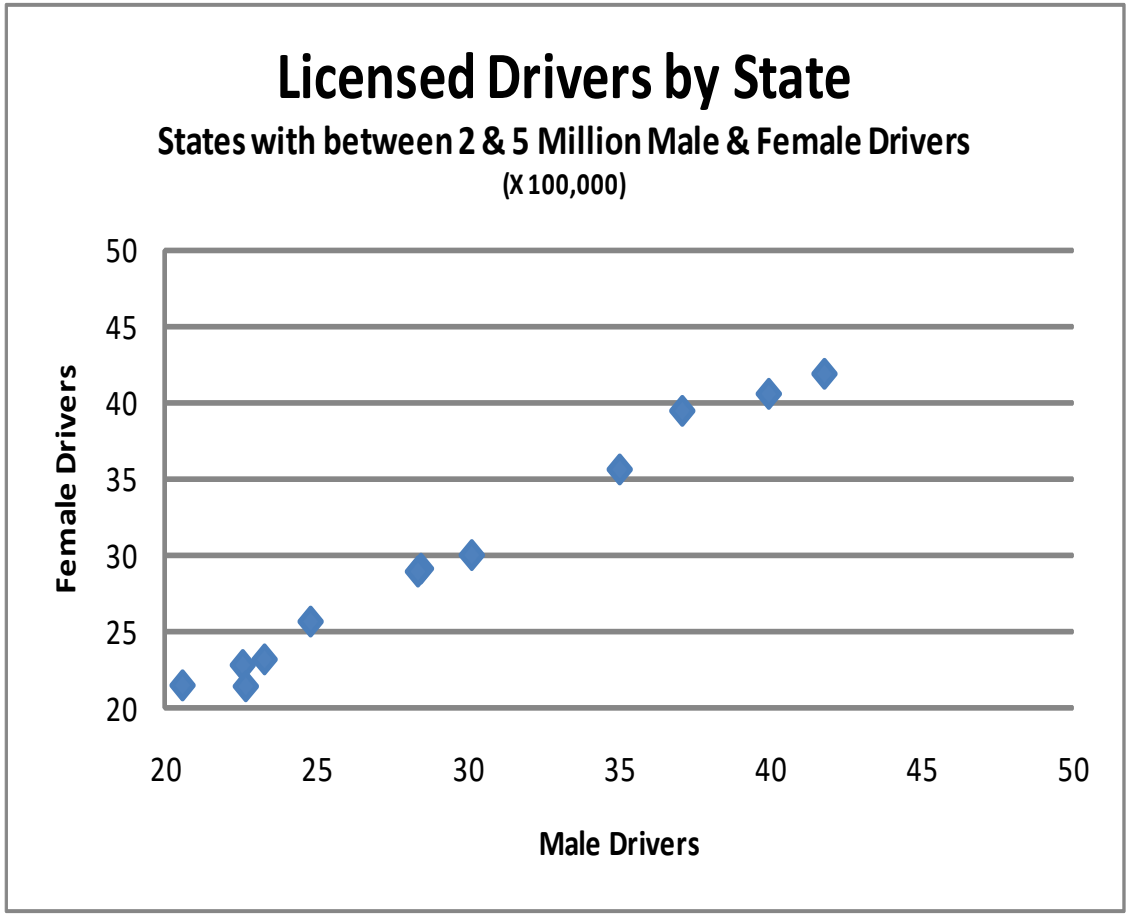
Example

- 4 Different Tools Producing Same Product (Qualitative)
- Measurements of Critical Dimension on Product (Quantitative)

Summaries by Types of Variables

Two Quantitative Variables

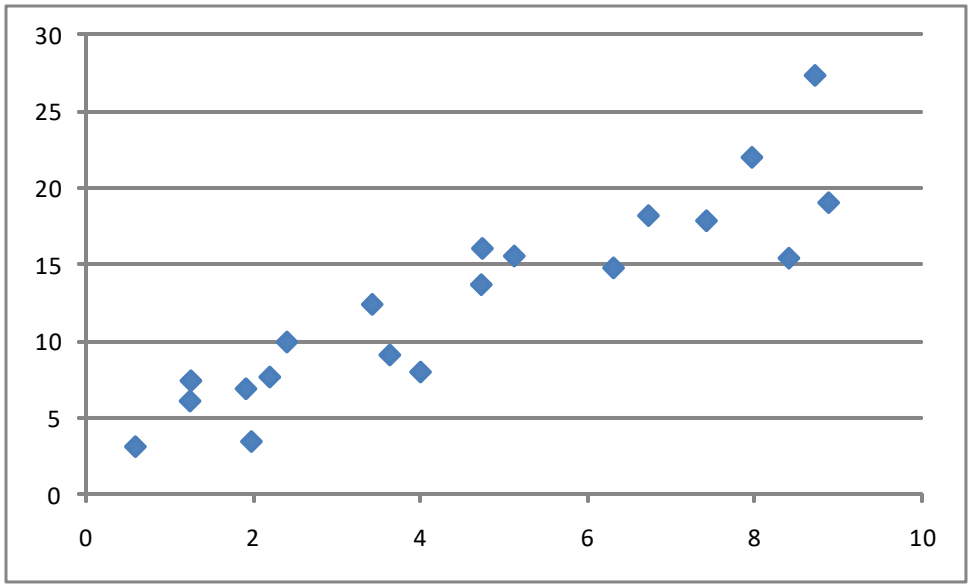
Scatter Diagrams



Licensed Drivers (x100,000)		
State	Male	Female
Georgia	28.448	29.131
Illinois	39.965	40.578
Indiana	22.550	22.812
Massachusetts	23.270	23.187
Michigan	35.028	35.626
New Jersey	28.348	28.941
North Carolina	30.131	30.017
Ohio	37.100	39.463
Pennsylvania	41.806	41.889
Tennessee	20.556	21.486
Virginia	24.795	25.664
Washington	22.651	21.422

Plot the Data

X	Y
7.41	17.82
2.40	9.92
8.87	19.00
3.63	9.08
6.30	14.75
4.72	13.66
1.97	3.44
3.42	12.38
1.25	7.40
3.99	7.97
8.40	15.38
2.19	7.64
6.72	18.16
8.71	27.29
1.24	6.08
5.11	15.53
4.73	16.02
7.95	21.95
1.91	6.88
0.58	3.11



Scatter plots provide a picture of the data

"A picture is worth a thousand words." - Confucius

Here we see the variables are generally linearly related with Y increasing as X increases

There are also some statistics that can be generated to assess the nature of the relationship between two variables:

- Sample Covariance
- Sample Correlation Coefficient

Sample Covariance

Sample Covariance is given as

$$s_{xy} = \text{Sum}\{(x_i - x_{Avg})*(y_i - y_{Avg})\}/(n-1),$$

where

(x_i, y_i), i = 1, …, n form a bivariate data set with

x_{Avg} = Sum(x_i)/n & y_{Avg} = Sum(y_i)/n, and all Summations from i = 1 to n.

X	Y	$X-X_{Avg}$	$Y-Y_{Avg}$	$(X-X_{Avg})*(Y-Y_{Avg})$
7.41	17.82	2.84	5.15	14.60
2.40	9.92	-2.18	-2.75	5.99
8.87	19.00	4.30	6.33	27.19
3.63	9.08	-0.95	-3.59	3.41
6.30	14.75	1.72	2.08	3.59
4.72	13.66	0.14	0.98	0.14
1.97	3.44	-2.60	-9.23	24.03
3.42	12.38	-1.16	-0.30	0.34
1.25	7.40	-3.33	-5.27	17.53
3.99	7.97	-0.58	-4.70	2.74
8.40	15.38	3.82	2.71	10.35
2.19	7.64	-2.38	-5.03	11.98
6.72	18.16	2.14	5.48	11.75
8.71	27.29	4.13	14.62	60.40
1.24	6.08	-3.34	-6.59	21.99
5.11	15.53	0.54	2.85	1.54
4.73	16.02	0.16	3.34	0.53
7.95	21.95	3.38	9.28	31.35
1.91	6.88	-2.67	-5.80	15.46
0.58	3.11	-3.99	-9.57	38.16
X_{Avg}	Y_{Avg}			Covariance
4.57	12.67			15.95

$s_{xy} > 0$

Positively Related

As x increases, y increases

$s_{xy} < 0$

Negatively Related

As x increases, y decreases

Range for s_{xy}

$-\infty$ to $+\infty$

So what can the sample covariance tell us about the relationship between X and Y?

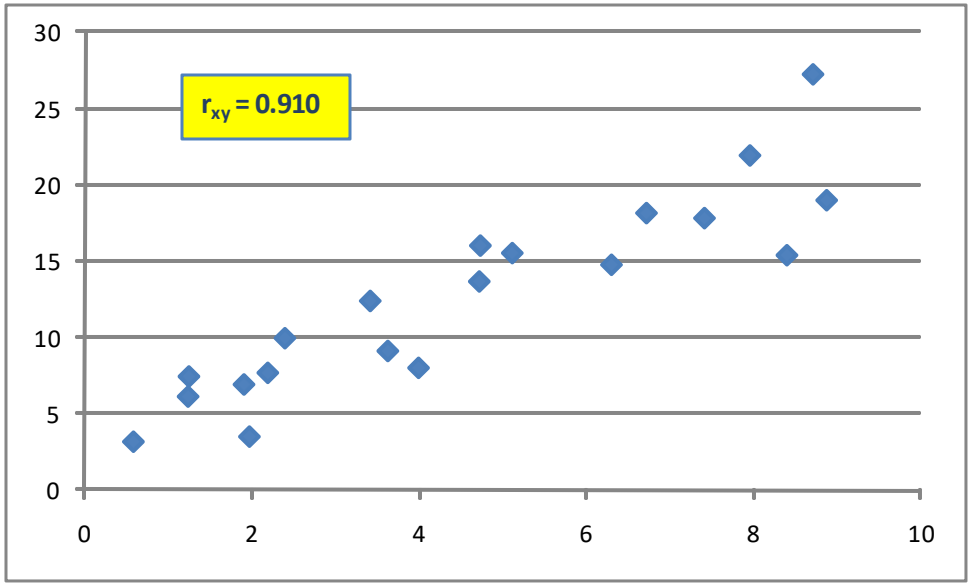
What does the sample covariance fail to tell us about the relationship between X and Y?

Linear Correlation Coefficient

If we normalize the Sample Covariance through division by the respective sample standard deviations of the variables X and Y, we obtain the Linear Correlation Coefficient, given as

$$r_{xy} = s_{xy}/(s_x * s_y),$$

where $s_x^2 = \text{Sum}\{(x_i - x_{Avg})^2\}/(n-1)$, $s_y^2 = \text{Sum}\{(y_i - y_{Avg})^2\}/(n-1)$,
with (x_i, y_i), $i = 1, \ldots, n$ form a bivariate data set,
$x_{Avg} = \text{Sum}(x_i)/n$ & $y_{Avg} = \text{Sum}(y_i)/n$, and all Summations from i = 1 to n.



Range for r_{xy}: -1 to 1

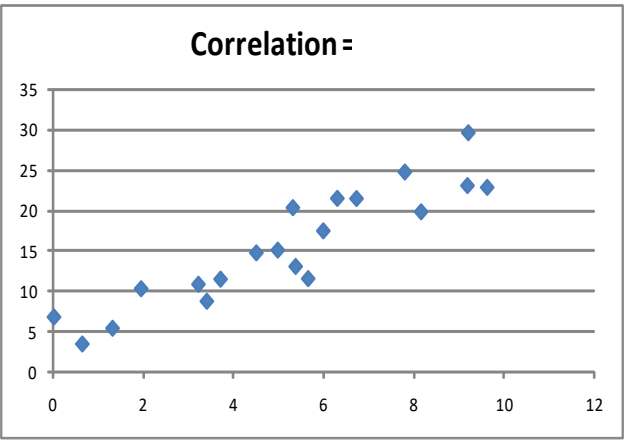
Same interpretation of sign as for Covariance

Advantage for Linear Correlation Coefficient over Covariance:
Can also measure relative "strength" of linear relationship

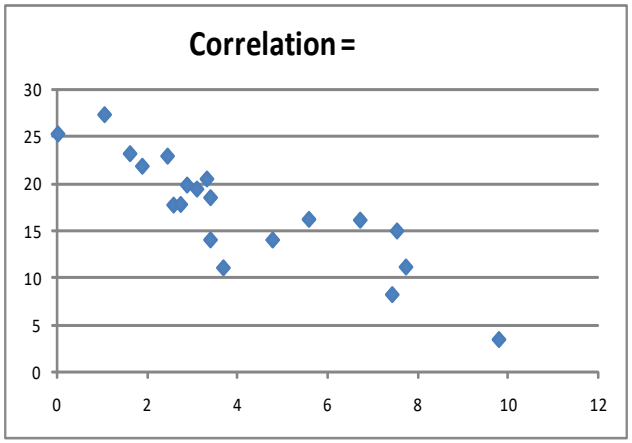
Linear Correlation

Measures LINEARITY of Relationship

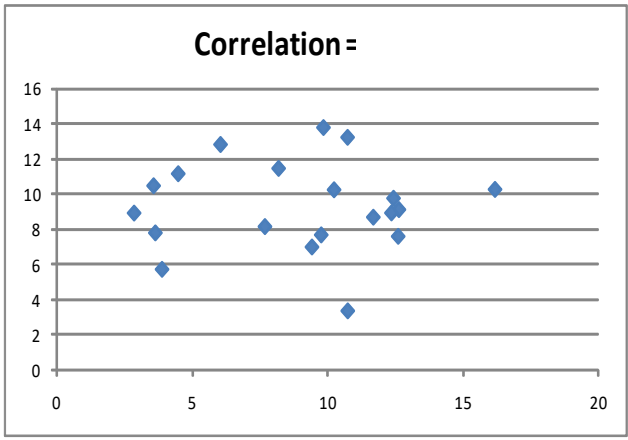
Positive Linearity

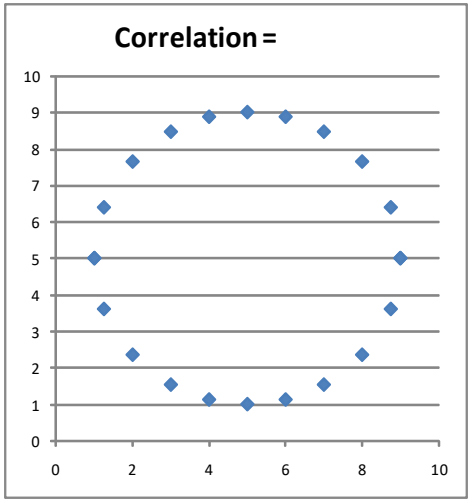


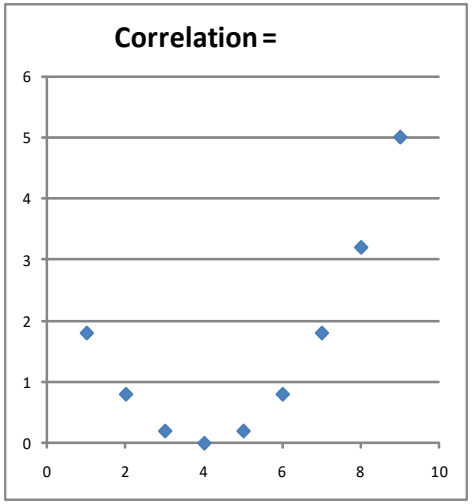
Negative Linearity

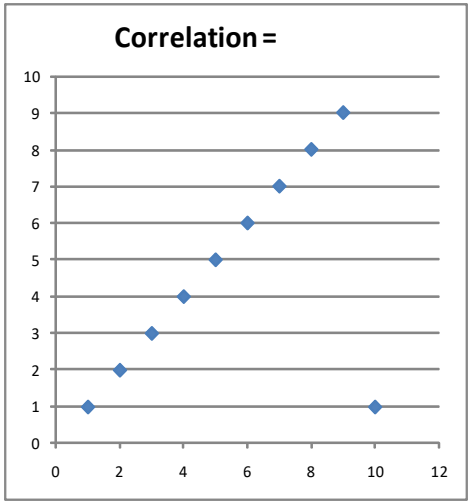


No Linearity





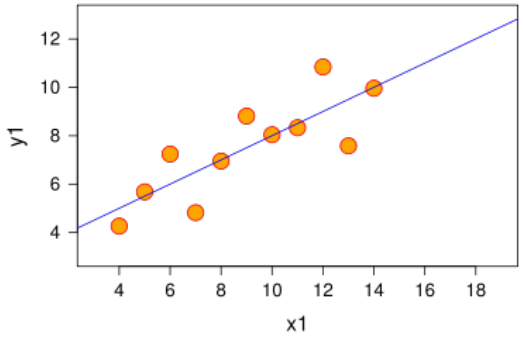




Plot Your Data!

Do Not Rely on r Values Alone To Assess the Relationship

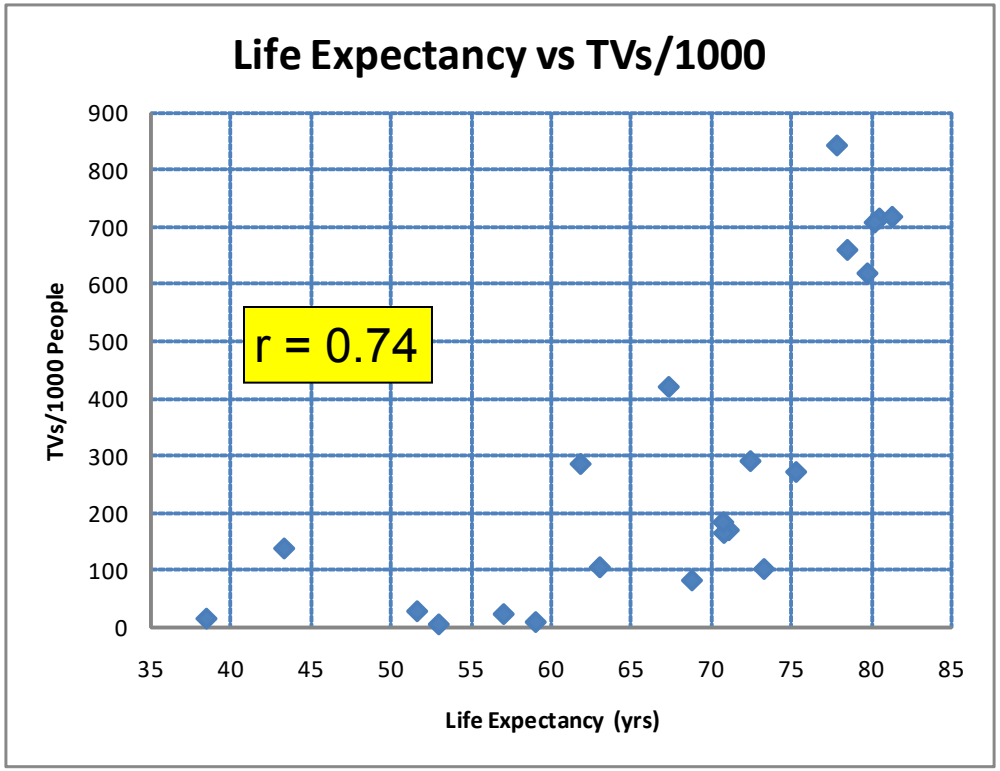
Linear Correlation



Linear Correlation

Does Not Necessarily Imply CAUSATION

Country	Life Exp	TVs/1000
Angola	38.45	15
Australia	80.45	716
Cambodia	59	9
Canada	80.15	709
China	72.4	291
Egypt	71.05	170
France	79.7	620
Haiti	52.95	5
Iraq	68.75	82
Japan	81.25	719
Madagascar	57	23
Mexico	75.25	272
Morocco	70.75	165
Pakistan	63	105
Russia	67.3	421
South Africa	43.3	138
Sri Lanka	73.25	102
Uganda	51.6	28
United Kingdom	78.45	661
United States	77.8	844
Vietnam	70.7	184
Yemen	61.8	286



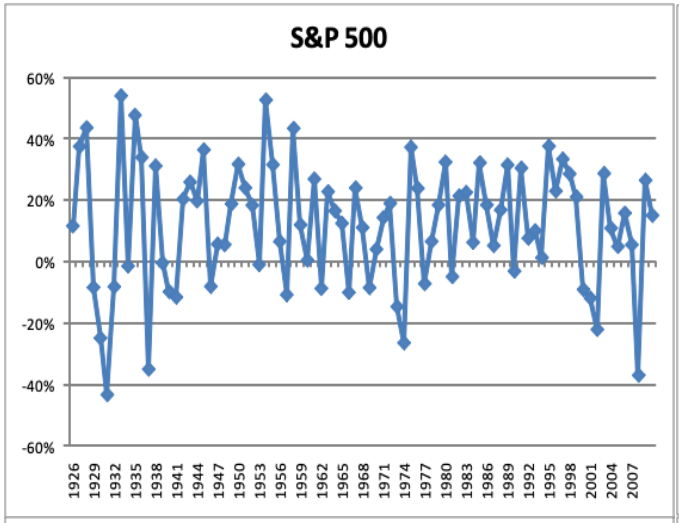
Positive Correlation between Life Expectancy, TVs/1000
Would sending more TVs to Countries Increase Life Expectancy?

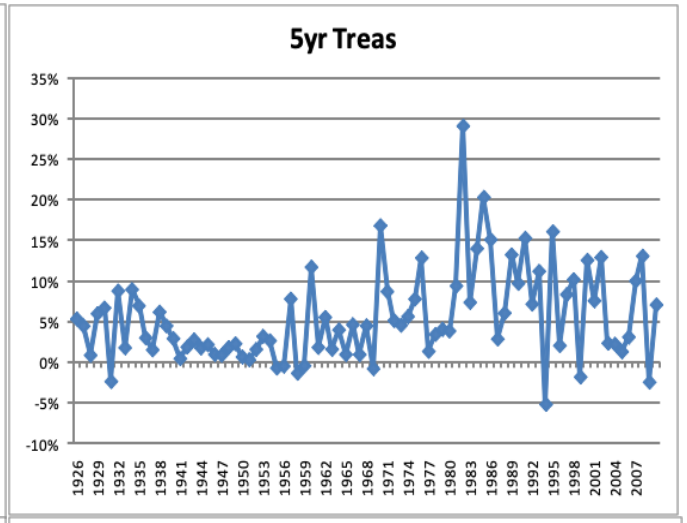
Both Driven More by Wealth or Income per Capita

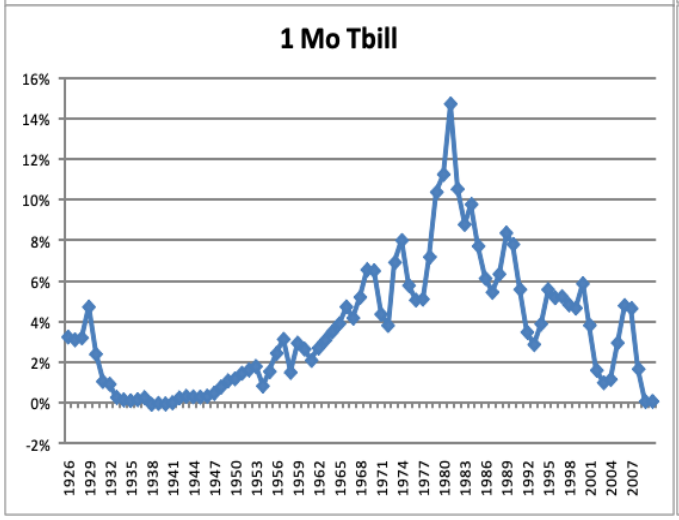
Strong Correlation Does NOT Imply Causation; however,
Causation Generally Results in Some Degree of Correlation

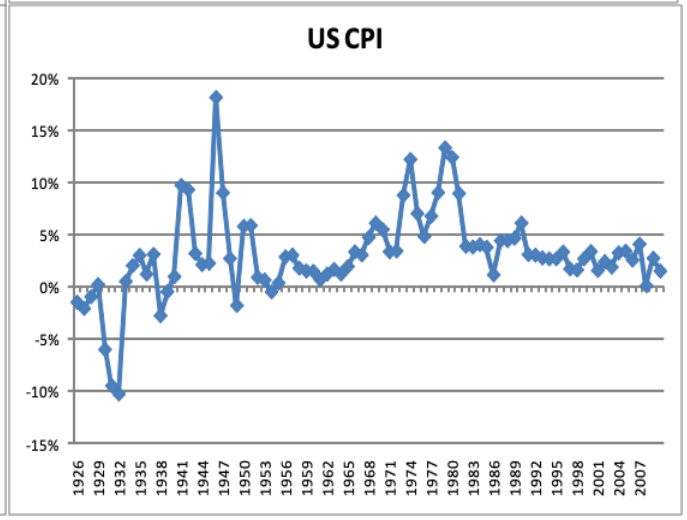
Sequence Plots

Univariate data that is collected in sequence (usually, time) is essentially bivariate data with the sequence as the second variable.









All these time series show different behavior, and while plots are always of value, there is an entire course dedicated to time series modeling.

Date	S&P 500	5yr Treas	1 Mo Tbill	US CPI
1926	12%	5%	3%	-1%
1927	37%	5%	3%	-2%
1928	44%	1%	3%	-1%
1929	-8%	6%	5%	0%
1930	-25%	7%	2%	-6%
1931	-43%	-2%	1%	-10%
1932	-8%	9%	1%	-10%
1933	54%	2%	0%	1%
1934	-1%	9%	0%	2%
1935	48%	7%	0%	3%
1936	34%	3%	0%	1%
1937	-35%	2%	0%	3%
1938	31%	6%	0%	-3%
1939	0%	5%	0%	0%
1940	-10%	3%	0%	1%
1941	-12%	0%	0%	10%
1942	20%	2%	0%	9%
1943	26%	3%	0%	3%
1944	20%	2%	0%	2%
1945	36%	2%	0%	2%
1946	-8%	1%	0%	18%
1947	6%	1%	1%	9%
1948	6%	2%	1%	3%
1949	19%	2%	1%	-2%
1950	32%	1%	1%	6%
1951	24%	0%	1%	6%
1952	18%	2%	2%	1%
1953	-1%	3%	2%	1%
1954	53%	3%	1%	-1%
1955	32%	-1%	2%	0%
1956	7%	0%	2%	3%
1957	-11%	8%	3%	3%
1958	43%	-1%	2%	2%
1959	12%	0%	3%	2%
1960	0%	12%	3%	1%
1961	27%	2%	2%	1%
1962	-9%	6%	3%	1%
1963	23%	2%	3%	2%
1964	17%	4%	4%	1%
1965	12%	1%	4%	2%
1966	-10%	5%	5%	3%
1967	24%	1%	4%	3%
1968	11%	5%	5%	5%
1969	-8%	-1%	7%	6%
1970	4%	17%	7%	5%
1971	14%	9%	4%	3%
1972	19%	5%	4%	3%
1973	-15%	5%	7%	9%
1974	-26%	6%	8%	12%
1975	37%	8%	6%	7%
1976	24%	13%	5%	5%
1977	-7%	1%	5%	7%
1978	7%	3%	7%	9%
1979	18%	4%	10%	13%
1980	32%	4%	11%	12%
1981	-5%	9%	15%	9%
1982	21%	29%	11%	4%
1983	23%	7%	9%	4%
1984	6%	14%	10%	4%
1985	32%	20%	8%	4%
1986	18%	15%	6%	1%
1987	5%	3%	5%	4%
1988	17%	6%	6%	4%
1989	31%	13%	8%	5%
1990	-3%	10%	8%	6%
1991	30%	15%	6%	3%
1992	8%	7%	4%	3%
1993	10%	11%	3%	3%
1994	1%	-5%	4%	3%
1995	38%	16%	6%	3%
1996	23%	2%	5%	3%
1997	33%	8%	5%	2%
1998	29%	10%	5%	2%
1999	21%	-2%	5%	3%
2000	-9%	13%	6%	3%
2001	-12%	8%	4%	2%
2002	-22%	13%	2%	2%
2003	29%	2%	1%	2%
2004	11%	2%	1%	3%
2005	5%	1%	3%	3%
2006	16%	3%	5%	3%
2007	5%	10%	5%	4%
2008	-37%	13%	2%	0%
2009	26%	-2%	0%	3%
2010	15%	7%	0%	1%

STAT 5340
Statistical Analysis I

Inference for Multiple Samples

Two Sample Problems

Consider 2 Types of Two Sample Problems:

1) **Paired (Dependent) Samples**
   - These samples are ***"linked "*** in some way by observation
     - Results for same individual, same location, same object, etc.
   - Samples will *Always* be Same Size

2) **Independent Samples**
   - No obvious ***"link"*** across samples
   - Each group comprised of results for entirely different individuals, locations, objects, etc.
   - Samples *Not Necessarily* of Same Size

Paired Sample Data

Suppose we work for an advertising firm, and have prepared a new marketing campaign for one of our national clients.

Before running the campaign nationwide, we want to know if it will have sufficient impact to justify such an expense.

Consequently, we run the campaign in 12 selected test markets around the country.

The data we have is Sales data for the two weeks prior to running the new ads in each market, and Sales data for the two weeks after running the ads.

Sales Data (Unitsx1000)			
Market	Before	After	Delta
Atlanta	195.8	233.9	38.1
Boston	237.4	239.0	1.6
Chicago	323.8	377.4	53.6
Dallas	205.5	220.3	14.8
Los Angeles	353.5	414.5	61.0
Miami	92.3	109.0	16.7
Nashville	22.2	50.7	28.5
New Orleans	52.2	32.9	-19.3
New York	577.7	602.7	25.1
Philadelphia	237.9	271.1	33.1
San Francisco	145.6	160.6	15.0
Seattle	82.0	108.7	26.7
Total	2525.8	2820.9	295.1

With paired data such as this, we generally consider the ***difference*** in results for each observation.

"Link" is same Market

Now the **Research Hypothesis** becomes:

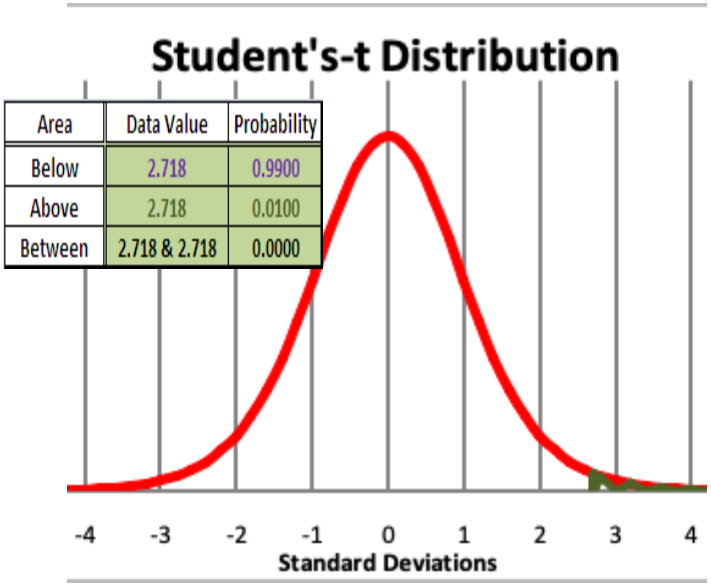
$$H_1: \mu_{Delta} > 0$$

and the corresponding **Null Hypothesis** is:

$$H_0: \mu_{Delta} = 0$$

Paired Sample Data

Sales Data (Unitsx1000)			
Market	Before	After	Delta
Atlanta	195.8	233.9	38.1
Boston	237.4	239.0	1.6
Chicago	323.8	377.4	53.6
Dallas	205.5	220.3	14.8
Los Angeles	353.5	414.5	61.0
Miami	92.3	109.0	16.7
Nashville	22.2	50.7	28.5
New Orleans	52.2	32.9	-19.3
New York	577.7	602.7	25.1
Philadelphia	237.9	271.1	33.1
San Francisco	145.6	160.6	15.0
Seattle	82.0	108.7	26.7
Total	2525.8	2820.9	295.1



Test Statistic: $T = (\bar{X}_{Delta} - \mu_{Delta})/[S_{Delta}/\sqrt{n}]$

n = 12 < 30 (small)
σ Unknown, so T

Null Distribution: $T \sim t_{(n-1)} = t_{(11)}$

Decision Rule:

Type I Error: Reject H_0 when H_0 TRUE

Conclude Positive Impact when Really None

Potentially Pursue Campaign when No Added Value

Type II Error: Fail to Reject H_0 when H_0 FALSE

Fail to Recognize Positive Impact of Campaign

Potentially Lose Revenue if Campaign Not Implemented

Suggest Small $\alpha = 0.01$

Reject H_0 if $T > 2.718 = t_{(11,0.99)}$

Decision: $\bar{X}_{Delta} = 24.6$, $S_{Delta} = 21.63$, so $T = 3.94$, and we Reject H_0 since $T = 3.94 > 2.718$

Conclude: Marketing Campaign had a positive impact on Unit Sales at 0.01 Significance Level (p-Value = Observed Sig Level = 0.0012)

Paired Sample Data

So it appears the Marketing Campaign Has the potential to Increase Unit Sales, but by how much?

Confidence Interval: Best Point Estimate ± $M_{(conf)}$*(Std Dev of Best Pt. Est.)

$\bar{X}_{Delta}$ ± $M_{(conf)}$*[$S_{Delta}/\sqrt{n}$]

n = 12 < 30 (small)
σ Unknown, so T

$\bar{X}_{Delta}$ ± $t_{(n-1,1-\alpha/2)}$*[$S_{Delta}/\sqrt{n}$]

α = 0.01, α/2 = 0.005,
so 1 – α/2 = 0.995

$\bar{X}_{Delta}$ ± $t_{(11,0.995)}$*[$S_{Delta}/\sqrt{n}$]

24.6 ± 3.106 * [21.63 / $\sqrt{12}$]

(5.2 to 44.0)

So, with 99% Confidence, the Expected Increase in Unit Sales is between 5,200 and 44,000 Units per Market.
Does this tell us all we need to know to make a decision here?

For the 12 Markets tested, the Increases above represent a ~2.5% to ~20.9% in Unit Sales, but if the Marketing Campaign needs to Increase Sales by 25% in order to provide a return greater then the firm's cost of capital, then it is likely the money for the campaign could be better invested elsewhere.

Important Point: Statistical Significance <u>Does NOT</u> Imply Practical Significance

Independent Samples Data

In many situations, we do not have, or are unable to obtain paired results.
Consider a Sleep Deprivation study.
There were two sample groups involved:

Group 1: Sleep Deprived & Group 2: Unrestricted Sleep

Group 1 had 11 individuals & Group 2 had 10 individuals
No clear, obvious "link" between individuals across groups, so considered

Independent

Response Time Improvement (Negative Values Indicate Worse Performance on Post-Test)	
Unrestricted	Deprived
-7	-14.7
11.6	-10.7
12.1	-10.7
12.6	2.2
14.5	2.4
18.6	4.5
25.2	7.2
30.5	9.6
34.5	10
45.6	21.3
	21.8

The measurement of interest was
Response Time Improvement after 3 days
- Group 1 was Deprived of Sleep on 1st day after initial test
- Group 2 was allowed Unrestricted Sleep all 3 days

Question of Interest is Whether or Not Sleep Deprivation Effects Linger for Multiple Days, so

Research Hypothesis: $\mu_{Unrestricted} - \mu_{Deprived} > 0$

Improvement Greater for Unrestricted Sleep Group

and

Null Hypothesis: $\mu_{Unrestricted} - \mu_{Deprived} = 0$

No Difference in Improvement Between Groups

Independent Samples Data

Response Time Improvement (Negative Values Indicate Worse Performance on Post-Test)	
Unrestricted	Deprived
-7	-14.7
11.6	-10.7
12.1	-10.7
12.6	2.2
14.5	2.4
18.6	4.5
25.2	7.2
30.5	9.6
34.5	10
45.6	21.3
	21.8

Function of the Difference In sample Averages

Test Statistic: ~~U-bar – D-bar~~ → T as given below

Difference in Sample Averages

Null Distribution: ?

n is small for both groups, as well as combined, and there is no information about the value of σ, for either group, or collectively – suggests we need something similar to the T statistic we use for single sample hypothesis testing.

There is an analogous T statistic for two groups …
If we assume that the data is reasonably normal and that the two populations (Deprived & Unrestricted) only differ in their mean values (ie, they have a common standard deviation, σ – still unknown, however), then

NOTE: Will be Zero Under Null Model

$$T = \{ (\bar{U} - \bar{D}) - (\mu_{Unrestricted} - \mu_{Deprived}) \} / S_{pooled} * \sqrt{(1/n_U) + (1/n_D)}$$

$$\sim t_{(nU + nD - 2)}$$

$$S_{pooled} = \sqrt{[(n_U - 1)SU2 + (n_D - 1)SD2] / (n_U + n_D - 2)}$$

Independent Samples Data

Response Time Improvement (Negative Values Indicate Worse Performance on Post-Test)	
Unrestricted	Deprived
-7	-14.7
11.6	-10.7
12.1	-10.7
12.6	2.2
14.5	2.4
18.6	4.5
25.2	7.2
30.5	9.6
34.5	10
45.6	21.3
	21.8

Decision Rule:

Type I Error: Reject H_0 when H_0 TRUE

Conclude there are lingering effects of Sleep Deprivation when there are none

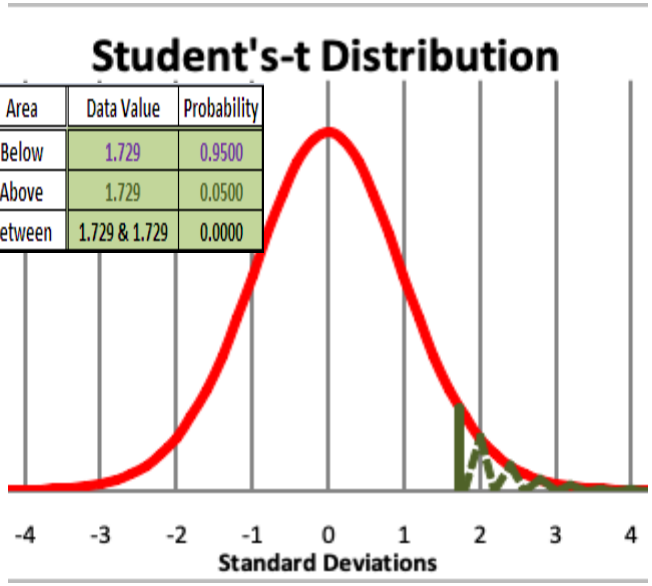
Type II Error: Fail to Reject H_0 when H_0 FALSE

Fail to recognize real lingering Sleep Deprivation effects

Suggest $\alpha = 0.05$

Reject if $T > 1.729 = t_{(19,0.95)}$

NOTE: $n_U + n_D - 2 = 10 + 11 - 2 = 19$



Decision:

$T = (\text{U-bar} - \text{D-bar})/\{S_{pooled}\sqrt{(1/n_U)+(1/n_D)}\}$

U-bar = 19.82, D-bar = 3.90, $n_U = 10$, $n_D = 11$

$S_U = 14.73$, $S_D = 12.17$, $S_{pooled} = 13.44$

$T = 2.711 > 1.729$, so Reject H_0

Conclude: There are lingering effects of Sleep Deprivation even 3 days later (p-Value = 0.0069)

Independent Samples Data

Similar to all the other Hypothesis Tests, this one also has a corresponding Confidence Interval. A (1-α)% Confidence Interval for $\mu_{Unrestricted} - \mu_{Deprived}$ is given by:

Best Point Estimate ± M_{conf} * Standard Deviation of Best Point Estimate

U-bar – D-bar ± $M_{conf} * S_{pooled} * \sqrt{(1/n_U) + (1/n_D)}$

U-bar – D-bar ± $t_{(nU + nD -2,\ 0.975)} * S_{pooled} * \sqrt{(1/n_U) + (1/n_D)}$

19.82 – 3.90 ± 2.093 * 13.44 * $\sqrt{(1/10) + (1/11)}$

15.92 ± 12.29

(3.63, 28.21)

So, we can state that with 95% confidence, $\mu_{Unrestricted} - \mu_{Deprived}$ is between 3.63 and 28.21 milli-seconds. Note that this interval does ***not*** include zero, which is consistent with the rejection of H_0: $\mu_{Unrestricted} - \mu_{Deprived} = 0$.

Conclusion is that for the 18-25 year-old age group, there are apparently some effects of sleep deprivation that are still present two days after a sleep deprived night.

Comparison of Two Means

Consider a company that annually distributes bonuses to its employees, but uses a rather involved method to determine the size of the bonus (as a percentage of each employee's regular salary) for each individual.

The primary factor in the method is the evaluation of the employee's direct supervisor, and the Human Resources department is concerned that male employees are being routinely rated higher than female employees with a resultant difference in annual bonus pay.

To assess this concern, the HR group randomly sampled some recent bonus pay percentages for a number of employees of each gender.

What might be a reasonable first step?

Female	Male
9.2	10.4
7.7	8.9
11.9	11.7
6.2	12
9	8.7
8.4	9.4
6.9	9.8
7.6	9
7.4	9.2
8	9.7
9.9	9.1
6.7	8.8
8.4	7.9
9.3	9.9
9.1	10
8.7	10.1
9.2	9
9.1	11.4
8.4	8.7
9.6	9.6
7.7	9.2
9	9.7
9	8.9
8.4	9.2
	9.4
	9.7
	8.9
	9.3
	10.4
	11.9
	9
	12
	9.6
	9.2
	9.9
	9

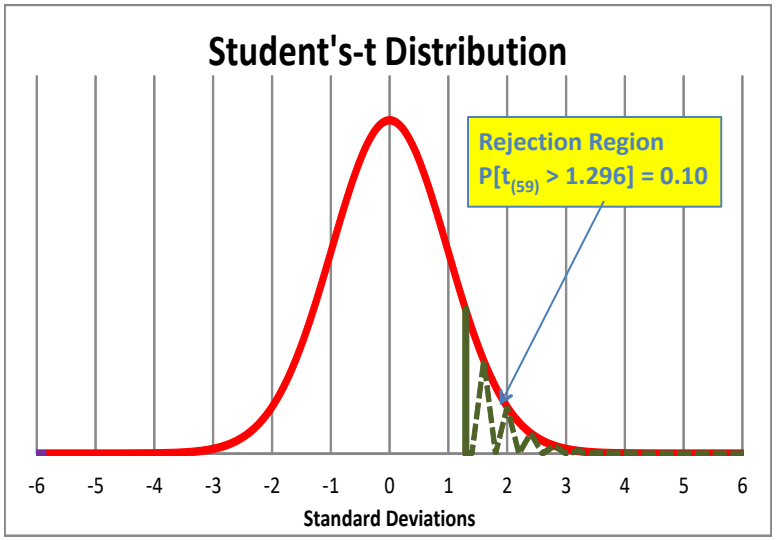
Comparison of Two Means

A more formal analysis might include a hypothesis test:

1) Research Hypothesis: H_1: $\mu_M > \mu_F$, or $\mu_M - \mu_F > 0$
2) Null Hypothesis: H_0: $\mu_M = \mu_F$, or $\mu_M - \mu_F = 0$
3) Test Statistic:
   1) $T = (\bar{X}_M - \bar{X}_F)/\sqrt{SPooled2(1/n_M + 1/n_F)}$
   2) with $S_{Pooled}^2 = [(n_M-1)S_M^2 + (n_F-1)S_F^2]/(n_M+n_F-2)$
4) Null Distribution: $t_{(n_M+n_F-2)}$
5) Decision Rule:
   1) Type I Error: Conclude higher bonus % for Males when really no difference – Adjust unnecessarily creating higher bonus % for Females
   2) Type II Error: Fail to realize higher bonus % for Males – Potential Class Action Suit
   3) Set $\alpha = 0.10$, Reject H_0 if $T > t_{(n_M+n_F-2,\ 0.90)} = 1.296$
6) Decision:
   1) $n_M = 36$, $n_F = 24$, $\bar{X}_M = 9.68$, $\bar{X}_F = 8.53$, $S_{Pooled} = 1.081$, so
   2) $T = 4.037 > 1.296$; hence, Reject H_0
7) Conclusion: Males at the company receive a higher bonus percentage, on average, than the female employees (p-Value ≈ 0.0001).

Clearly, this result is unsatisfactory, we would want to know how much higher.

NOTE: There is an implicit assumption being made here that the variance for the Male and Female bonus percentages are the same value (ie, $\sigma_M^2 = \sigma_F^2$). This may or may not be reasonable. If not, then the test needs to be modified.



A 90% Confidence Interval for the Difference is obtained as

$$\bar{X}_M - \bar{X}_F \pm t_{(n_M+n_F-2,\ 0.95)} * S_{Difference},$$

where

$$S_{Difference} = S_{Pooled} * \sqrt{1/n_M + 1/n_F}.$$

With the available data, this result is

1.15 ± 0.48 = (0.67 to 1.62),

so the answer to "How Much?" is that we have 90% confidence that bonus percentages are 0.67% to 1.62% higher for male employees.

Comparison of Two Means

If the variances of the two populations of interest are not equal, then pooling the data across the two samples (to obtain S_{Pooled}) is no longer appropriate.

The test statistic, T, is modified as follows:

$$T = (\bar{X}_M - \bar{X}_F)/\sqrt{SM2/n_M + SF2/n_F}$$

However, an issue arises in determining the sampling distribution for this statistic since when no longer pooling the samples, we no longer have fully $n_M + n_F - 2$ degrees of freedom with which to estimate the standard deviation of the difference in the sample means.

So … while T can still be approximated by a student's t distribution, what degrees of freedom (df) are appropriate?

One common, and easy approach is to use

$$df = \min(n_M - 1, n_F - 1)$$

Since we know there are $n_M - 1$ and $n_F - 1$ df for S_M^2 and S_F^2, respectively, choosing the minimum of these two values provides a conservative approach (recall, lower df results in heavier tails, so critical values and p-values will be larger in magnitude, and rejection of H_0 less likely).

A more accurate approach is to use

$$df = \mathbf{[}(A + B)^2/\{A^2/(n_M-1) + B^2/(n_F-1)\}\mathbf{]}, \text{ where } A = S_M^2/n_M,\ B = S_F^2/n_F, \text{ and } \mathbf{[}x\mathbf{]} = \text{greatest integer in } x$$

This will result in a value between $\min(n_M - 1, n_F - 1)$ and $n_M + n_F - 2$.

Comparison of Two Means

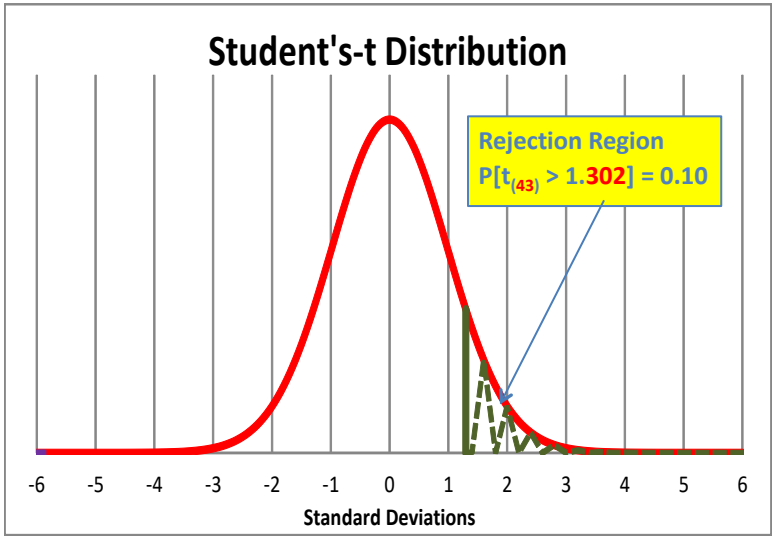
The hypothesis test procedure is modified as in red:

1) Research Hypothesis: H_1: $\mu_M > \mu_F$, or $\mu_M - \mu_F > 0$
2) Null Hypothesis: H_0: $\mu_M = \mu_F$, or $\mu_M - \mu_F = 0$
3) Test Statistic: If unwilling to assume $\sigma_M = \sigma_F$, we use
    1) $T = (\bar{X}_M - \bar{X}_F)/\sqrt{SM2/n_M + S2F/n_F}$
4) Null Distribution: $t_{(df)}$
    1) where $df = [(A+B)^2/(A^2/(n_M-1) + B^2/(n_F-1))]$,
    2) with $A = S_M^2/n_M$, $B = S_F^2/n_F$, and
    3) [x] = greatest integer in x.
5) Decision Rule:
    1) With $\alpha = 0.10$, Reject H_0 if $T > t_{(df=43,\ 0.90)} = 1.302$
6) Decision:
    1) $n_M = 36$, $n_F = 24$, $\bar{X}_M = 9.68$, $\bar{X}_F = 8.53$, $S_M = 1.00$ and $S_F = 1.19$, so
    2) $T = 3.901 > 1.302$; hence, Reject H_0
7) Conclusion: Males at the company receive a higher bonus percentage, on average, than the female employees (p-Value ≈ 0.000166).

"To borrow from a famous bard, is this 'much ado about nothing'?

Well, in this case, this is close to valid, as there is not much difference in variation across the groups.

So, most analysts simply assume unequal variances across groups, because if TRULY same, results will be very similar if had assumed same.



A 90% Confidence Interval for the Difference is obtained as

$$\bar{X}_M - \bar{X}_F \pm t_{(df,\ 0.95)} * S_{Difference},$$

where

$$S_{Difference} = \sqrt{SM2/n_M + SF2/n_F}.$$

With the available data, this result is

$1.15 \pm 0.50 = (0.65 \text{ to } 1.65)$,

and we have 90% confidence that bonus percentages are 0.65% to 1.65% higher for male employees.

Comparison of Two Means

HR Data:

R Code:

```
> hr_data = read.csv("HR Data.csv")
> hr_data
```

	Female	Male
1	9.2	10.4
2	7.7	8.9
3	11.9	11.7
4	6.2	12.0
5	9.0	8.7
6	8.4	9.4
7	6.9	9.8
8	7.6	9.0
9	7.4	9.2
10	8.0	9.7
11	9.9	9.1
12	6.7	8.8
13	8.4	7.9
14	9.3	9.9
15	9.1	10.0
16	8.7	10.1
17	9.2	9.0
18	9.1	11.4
19	8.4	8.7
20	9.6	9.6
21	7.7	9.2
22	9.0	9.7
23	9.0	8.9
24	8.4	9.2
25	NA	9.4
26	NA	9.7
27	NA	8.9
28	NA	9.3
29	NA	10.4
30	NA	11.9
31	NA	9.0
32	NA	12.0
33	NA	9.6
34	NA	9.2
35.	NA	9.9
36.	NA	9.0

Assuming Equal Variances:

```
> female = hr_data[,1]
> male = hr_data[,2]
>
> t.test(male, female, alternative = "greater", var.equal = TRUE)

        Two Sample t-test

data:  male and female
t = 4.0367, df = 58, p-value = 8.044e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6738034       Inf
sample estimates:
mean of x mean of y
 9.683333  8.533333
```

NOTE: One-sided options (eg, "greater") produce one-sided confidence bounds. To obtain two-sided confidence bounds, use: "two.sided" (default).

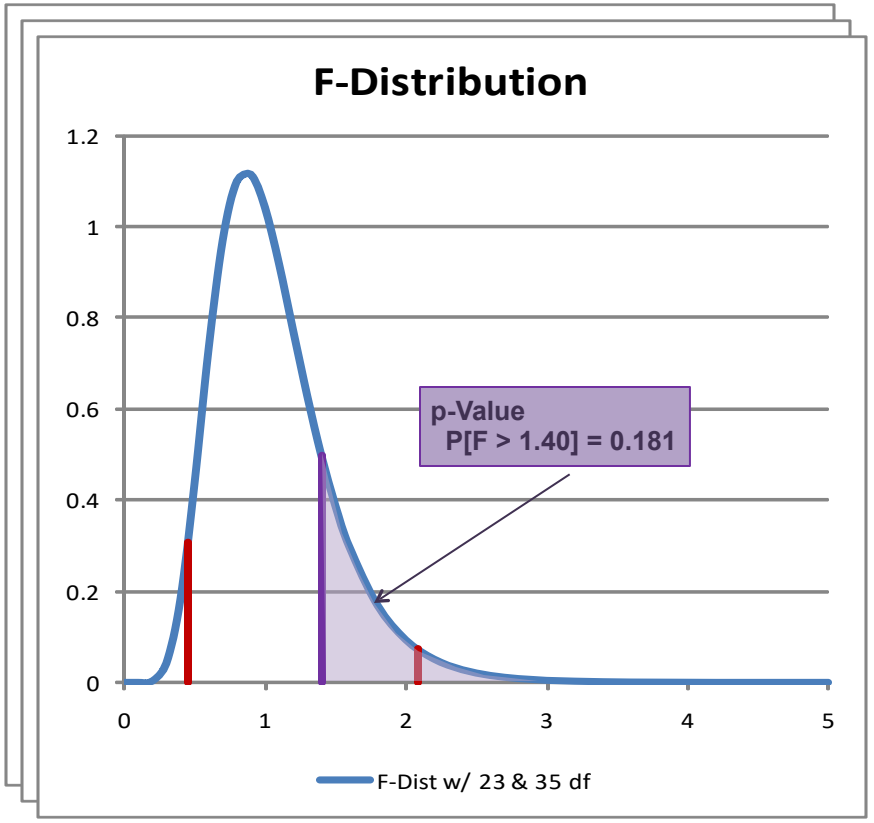
NOT Assuming Equal Variances:

```
> t.test(male, female, alternative = "greater")

        Welch Two Sample t-test

data:  male and female
t = 3.9013, df = 43.587, p-value = 0.0001635
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6546071       Inf
sample estimates:
mean of x mean of y
 9.683333  8.533333
```

Comparison of Two Variances

So, what about a statistical comparison of two population variances (or standard deviations)? Can we test to see if the assumption of equal variances is reasonable or not?

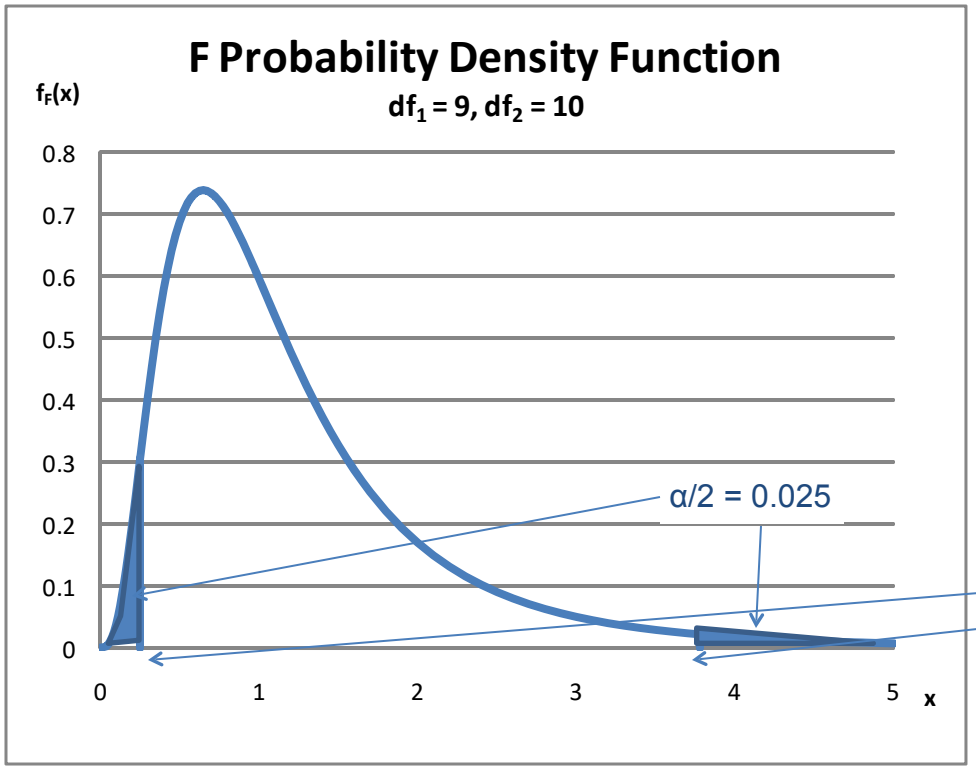
1) Research Hypothesis: $H_1: \sigma_M^2 \neq \sigma_F^2$
2) Null Hypothesis: $H_0: \sigma_M^2 = \sigma_F^2$
3) Test Statistic: ?
    1) $F = S_F^2/S_M^2$, ratio of Sample Variances
    2) S_F^2 = Sample Variance for Females
    3) S_M^2 = Sample Variance for Males
    4) Generally, put Larger Sample Variance in Numerator, but not necessary since this is a two-tailed hypothesis test.
4) Null Distribution: ?
    1) Under H_0, $F \sim F_{(n_F-1,\ n_M-1)}$
    2) F distributions indexed by 2 degrees of freedom values (numerator & denominator)
    3) Requires samples to be
        1) Independent
        2) Random
        3) From a **Normal Distribution**
5) Decision Rule:
    1) w/ α=0.05, Reject H_0 if
    2) $F < 0.45$ or $F > 2.07$
6) Decision:
    1) $S_F^2 = 1.41$, $S_M^2 = 1.01$
    2) $F = 1.40$, so Fail to Reject H_0



NOTE: Frequently, for two-sided cases, software packages will report twice the value above, or a p-Value = 0.362. However, in this case, a more appropriate consideration of the lower tail would be to calculate P[F < 1/1.40] = 0.200, then add this to the above to get a p-Value = 0.381.

Differences Between Two Standard Deviations

In the Sleep Deprivation study, when evaluating the means of the two groups, we pooled the information on variation by calculating a pooled standard deviation. The inherent assumption here was that the two samples came from populations with the same (or at least very nearly the same) variance. We can check this assumption:



Research Hypothesis: $H_1: \sigma_1 \neq \sigma_2$

Null Hypothesis: $H_0: \sigma_1 = \sigma_2$

Test Statistic: $F = [S_1^2/\sigma_1^2]/[S_2^2/\sigma_2^2]$

Null Distribution: $F_{(n_1-1,\ n_2-1)}$

Provided both populations follow normal distributions

Decision Rule: Reject H_0 if

$F < F_{(n_1-1,\ n_2-1,\ \alpha/2)}$ or $F > F_{(n_1-1,\ n_2-1,\ 1-\alpha/2)}$,

with $\alpha=0.05$, $n_1 = 10$, $n_2 = 11$:

$F < 0.25$ or $F > 3.78$

Decision: with $S_1 = 14.73$ & $S_2 = 12.17$

Sleep Deprivation Study data

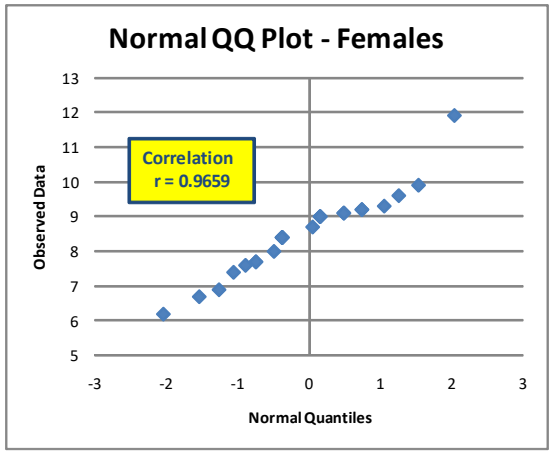
$F = (14.73/12.17)^2 = 1.465$

Conclusion: Data do not present sufficient evidence to conclude different variances for the two groups (p-Value ≈ 0.280).

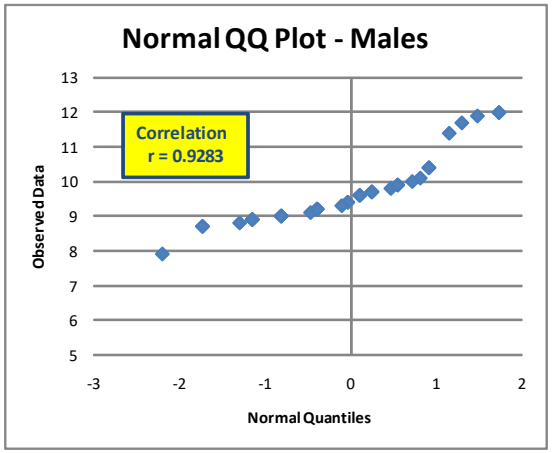
Comparison of Two Variances

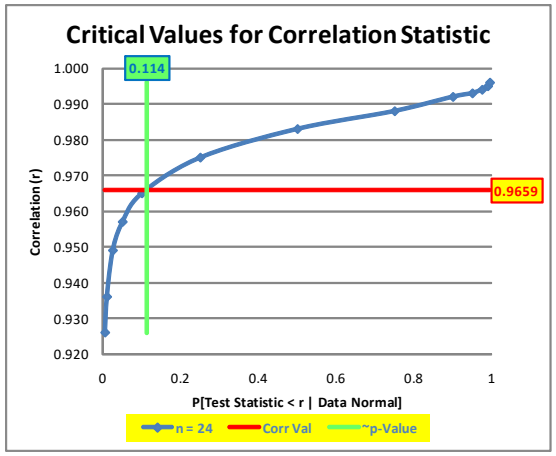
The assumption of Normally distributed populations being involved is an important one for use of the F-Distribution to make inferences about the two respective variances.
Consequently, it is generally desirable to evaluate this assumption prior to using this approach, or to use a different approach.

Evaluation for HR Data:

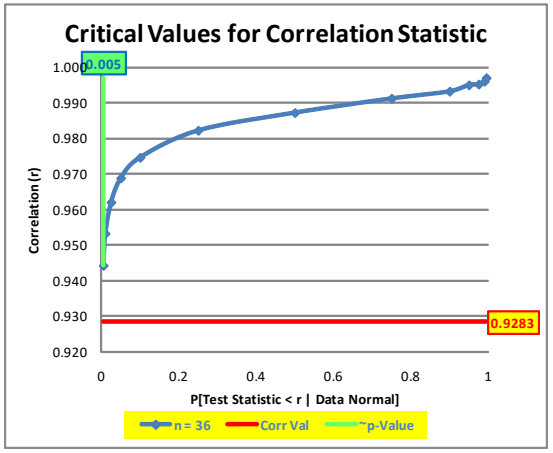


Conclusion:
Insufficient evidence to indicate Female results are not normally distributed. (So assumption of normality considered reasonable.)





Conclusion:
Evidence suggests Male results are not normally distributed. (So assumption of normality considered unreasonable.)



F-Test may not have generated valid conclusion.

Comparison of Two Variances

So … what do we do now?

Consider a different approach that does not depend on Normality.

One relatively easy test is described as follows:

1) Calculate the absolute deviations from the sample mean (ie, $|x_i - \bar{x}|$) for each respective sample
2) Order these within each sample from largest to smallest
3) Let c_1 = number of results for sample 1 greater than the largest result for sample 2
4) Let c_2 = number of results for sample 2 greater than the largest result for sample 1

Note: At least one of c_1 or c_2 must be zero.

5) If c_1 > Critical Value, then conclude $\sigma_1 > \sigma_2$
6) If c_2 > Critical Value, then conclude $\sigma_2 > \sigma_1$

Note: Critical Value = $\ln(\alpha/2)/\ln[n_1/(n_1+n_2)]$

7) Otherwise, conclude Insufficient evidence to Reject $\sigma_1=\sigma_2$

For the HR Data …

$c_1 = c_F = 2$ (and $c_2 = c_M = 0$)

Critical Value = 4.02 (w/ α=0.05)

Conclude: Insufficient evidence to indicate $\sigma_F \neq \sigma_M$.

NOTE: This is consistent with the F-Test result.

Data

Female	Male
9.2	10.4
7.7	8.9
11.9	11.7
6.2	12
9	8.7
8.4	9.4
6.9	9.8
7.6	9
7.4	9.2
8	9.7
9.9	9.1
6.7	8.8
8.4	7.9
9.3	9.9
9.1	10
8.7	10.1
9.2	9
9.1	11.4
8.4	8.7
9.6	9.6
7.7	9.2
9	9.7
9	8.9
8.4	9.2
	9.4
	9.7
	8.9
	9.3
	10.4
	11.9
	9
	12
	9.6
	9.2
	9.9
	9

Means	
Female	Male
8.53	9.68

Mean Absolute Deviations

MnAD - F	MnAD - M
0.6667	0.7167
0.8333	0.7833
3.3667	2.0167
2.3333	2.3167
0.4667	0.9833
0.1333	0.2833
1.6333	0.1167
0.9333	0.6833
1.1333	0.4833
0.5333	0.0167
1.3667	0.5833
1.8333	0.8833
0.1333	1.7833
0.7667	0.2167
0.5667	0.3167
0.1667	0.4167
0.6667	0.6833
0.5667	1.7167
0.1333	0.9833
1.0667	0.0833
0.8333	0.4833
0.4667	0.0167
0.4667	0.7833
0.1333	0.4833
	0.2833
	0.0167
	0.7833
	0.3833
	0.7167
	2.2167
	0.6833
	2.3167
	0.0833
	0.4833
	0.2167
	0.6833

Sorted Mean Absolute Deviations

MnAD - F	MnAD - M
3.3667	2.3167
2.3333	2.3167
1.8333	2.2167
1.6333	2.0167
1.3667	1.7833
1.1333	1.7167
1.0667	0.9833
0.9333	0.9833
0.8333	0.8833
0.8333	0.7833
0.7667	0.7833
0.6667	0.7833
0.6667	0.7167
0.5667	0.7167
0.5667	0.6833
0.5333	0.6833
0.4667	0.6833
0.4667	0.6833
0.4667	0.5833
0.1667	0.4833
0.1333	0.4833
0.1333	0.4833
0.1333	0.4833
0.1333	0.4167
	0.3833
	0.3167
	0.2833
	0.2833
	0.2167
	0.2167
	0.1167
	0.0833
	0.0833
	0.0167
	0.0167
	0.0167

Comparison of Two Variances

Another commonly used alternative approach that does not require the assumption of Normality is the Levene-Brown-Forsythe test.

The Levene test can be described as follows:

1) Calculate the absolute deviations from the sample median (ie, $|x_i - \tilde{x}|$) for each respective sample
2) Test the hypothesis of equal means for these absolute deviations from the median treating them as independent samples with potentially different variances

1) Research Hypothesis: H_1: $\mu_{MdAD-F} \neq \mu_{MdAD-M}$, or $\mu_{MdAD-F} - \mu_{MdAD-M} \neq 0$
2) Null Hypothesis: H_0: $\mu_{MdAD-F} = \mu_{MdAD-M}$, or $\mu_{MdAD-F} - \mu_{MdAD-M} = 0$
3) Test Statistic:
    1) $T = (\bar{X}_{MdAM-F} - \bar{X}_{MdAD-M})/\sqrt{SMdAD{-}F2/n_F + S2MdAD{-}M/n_M}$
4) Null Distribution: $t_{(df)}$
    1) where $df = \mathbf{[}(A+B)^2/(A^2/(n_M-1) + B^2/(n_F-1))\mathbf{]}$,
    2) with $A = S_{MdAD-M}^2/n_M$, $B = S_{MdAD-F}^2/n_F$, and
    3) **[**x**]** = greatest integer in x.
5) Decision Rule:
    1) With $\alpha = 0.05$, Reject H_0 if
    2) $T > t_{(df=48,\ 0.975)} = 2.011$ or $T < t_{(df=48,\ 0.025)} = -2.011$
6) Decision:
    1) $n_M = 36$, $n_F = 24$, $\bar{X}_{MdAD-F} = 0.88$, $\bar{X}_{MdAD-M} = 0.72$, $S_{MdAF-F} = 0.77$ and $S_{MdAD-M} = 0.75$, so
    2) $T = 0.827$, between ±2.011; hence, Fail to Reject H_0
7) Conclusion: Insufficient evidence to indicate difference in mean absolute deviations from medians (ie, variances) across populations

NOTE: The conclusion is still consistent with the other approaches. (p-Value ≈ 0.206, note, since 2-sided, some software doubles to 0.412).

Data

Female	Male
9.2	10.4
7.7	8.9
11.9	11.7
6.2	12
9	8.7
8.4	9.4
6.9	9.8
7.6	9
7.4	9.2
8	9.7
9.9	9.1
6.7	8.8
8.4	7.9
9.3	9.9
9.1	10
8.7	10.1
9.2	9
9.1	11.4
8.4	8.7
9.6	9.6
7.7	9.2
9	9.7
9	8.9
8.4	9.2
	9.4
	9.7
	8.9
	9.3
	10.4
	11.9
	9
	12
	9.6
	9.2
	9.9
	9

Medians	
Female	Male
8.55	9.4

Median Absolute Deviations

MdAD - F	MdAD - M
0.65	1
0.85	0.5
3.35	2.3
2.35	2.6
0.45	0.7
0.15	0
1.65	0.4
0.95	0.4
1.15	0.2
0.55	0.3
1.35	0.3
1.85	0.6
0.15	1.5
0.75	0.5
0.55	0.6
0.15	0.7
0.65	0.4
0.55	2
0.15	0.7
1.05	0.2
0.85	0.2
0.45	0.3
0.45	0.5
0.15	0.2
	0
	0.3
	0.5
	0.1
	1
	2.5
	0.4
	2.6
	0.2
	0.2
	0.5
	0.4

Comparison of Two Variances

HR Data: Test for Equal Variances

MINITAB: Stat → Basic Statistics → 2 Variances

Test for Equal Variances: Female, Male

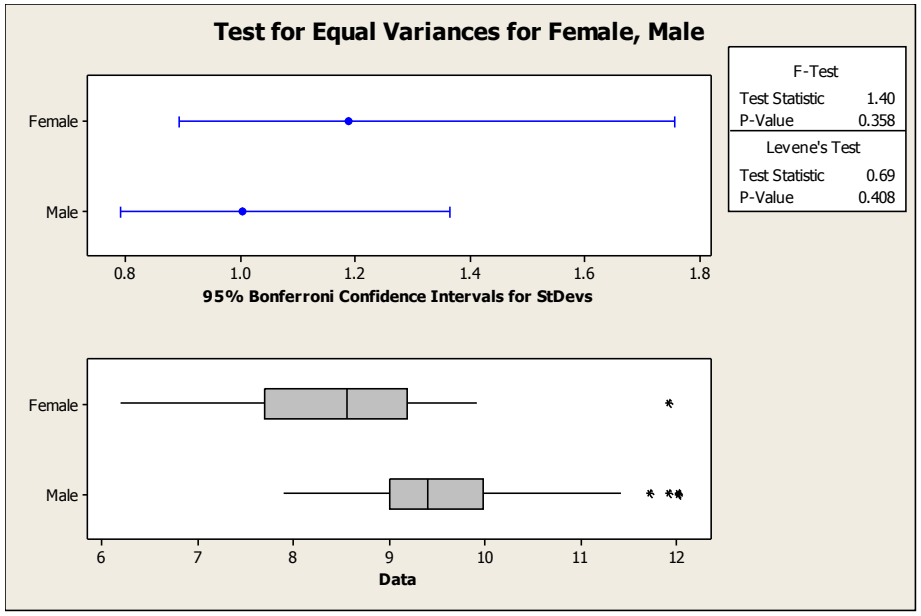
95% Bonferroni confidence intervals for standard deviations

	N	Lower	StDev	Upper
Female	24	0.892753	1.18896	1.75726
Male	36	0.791041	1.00385	1.36343

F-Test (Normal Distribution)
Test statistic = 1.40, p-value = 0.358

Levene's Test (Any Continuous Distribution)
Test statistic = 0.69, p-value = 0.408

Test for Equal Variances for Female, Male



What are "Bonferroni" confidence intervals?

What is a confidence interval?

Recall: A (1-α)% Confidence Interval implies the parameter of interest will be in the interval (1-α)% of the time such an interval is obtained.

Hence, α% of the time such intervals will **fail** to include the parameter of interest.

Now, two (1-α)% confidence intervals obtained from independent samples from two different populations will each **fail** to include their respective parameters of interest α% of the time.

Let A = Event that the 1st Interval Does **Not** Include its respective Parameter, then P[A] = α
Let B = Event that the 2nd Interval Does **Not** Include its respective Parameter, then P[B] = α

The probability that at least one of the intervals **fails** to include its respective parameter is P[A **or** B], so by the Compliment Rule, the probability that both intervals **INCLUDE** their respective parameters is 1 – P[A **or** B], but we know from the Additive Rule that

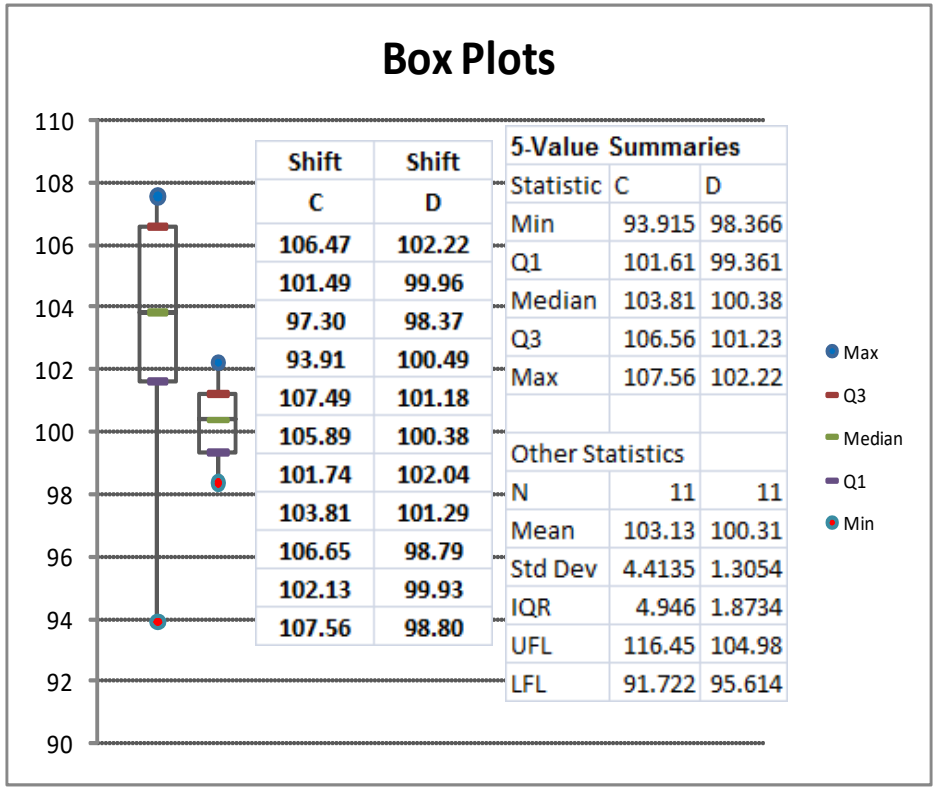
$$\begin{aligned} 1 - P[A \textbf{ or } B] &= 1 - \{P[A] + P[B] - P[A \textbf{ and } B]\} \\ &= 1 - P[A] - P[B] + P[A \textbf{ and } B] \\ &> 1 - P[A] - P[B] = 1 - 2\alpha, \text{ Bonferroni's Inequality.} \end{aligned}$$

Hence, with α=0.05, two independent 95% confidence intervals have at least a 90% probability of both including their respective parameters, and would be referred to as 90% "Bonferroni" confidence intervals.

The 95% "Bonferroni" confidence intervals at left are obtained by calculating individual 97.5% confidence intervals for each of the respective parameters (in this case, σ_F and σ_M).

Independent Samples Data
Variances within Groups Unequal

Example: In Workshop 1, it was observed that in addition to having more variable gate widths, C Shift also appeared to have widths averaging higher than the other shifts. Consider comparing the averages observed on C & D Shifts (since D Shift was near target = 100nm).



Shift C	Shift D
106.47	102.22
101.49	99.96
97.30	98.37
93.91	100.49
107.49	101.18
105.89	100.38
101.74	102.04
103.81	101.29
106.65	98.79
102.13	99.93
107.56	98.80

5-Value Summaries		
Statistic	C	D
Min	93.915	98.366
Q1	101.61	99.361
Median	103.81	100.38
Q3	106.56	101.23
Max	107.56	102.22
Other Statistics		
N	11	11
Mean	103.13	100.31
Std Dev	4.4135	1.3054
IQR	4.946	1.8734
UFL	116.45	104.98
LFL	91.722	95.614

Fairly obvious that variances within shifts are different, but $F = S_C^2/S_D^2 = 11.43$ (p-Value≈0.0003); hence, using S_{Pooled} would be inappropriate when testing if $\mu_C > \mu_D$.

Research Hypothesis: $\mu_C > \mu_D$.

Null Hypothesis: $\mu_C = \mu_D$.

Test Statistic: $T = (\text{X-bar}_C - \text{X-bar}_D)/\text{sqrt}(S_C^2/n_C + S_D^2/n_D)$

Null Distribution: $t_{(12)}$, where df from Welch

Decision Rule: Reject H_0 if $T > t_{(12,\alpha=0.05)} = 1.782$

Decision: T = 2.031, so Reject H_0, and

Conclude: C Shift Gate Widths are Larger than those produced on D Shift (p-Value ≈ 0.0325).

A 90% confidence interval for how much larger is determined as:

$(\text{X-bar}_C - \text{X-bar}_D) \pm 1.782*\text{sqrt}(S_C^2/n_C + S_D^2/n_D) = 2.818 \pm 1.782*1.388 =$ (0.35nm to 5.29nm) ,

so with 90% confidence, C Shift gate widths are between 0.35nm & 5.29nm larger than those produced on D Shift

Independent Samples Data
Differences Between Two Proportions

Suppose we work for a manufacturer of textiles, and suspect that one of the machines producing fabric is producing more bolts with defects than the others.

For a given time period, we sample 100 bolts from this machine and find 12 with defects.

During the same period, we also sample 200 bolts across the other machines in operation and find 15 with defects.

Question here is whether or not the machine of interest is producing a higher rate of defective bolts than the other machines, so ...

Research Hypothesis: $\pi_{Defective,Machine} - \pi_{Defective,Others} > 0$

Defect Rate Greater for Machine of Concern

and

Null Hypothesis: $\pi_{Defective,Machine} - \pi_{Defective,Others} = 0$

No Difference in Defect Rates Between Machine of Concern and Others in Operation

Independent Samples Data
Differences Between Two Proportions

Test Statistic: $p_{Machine} - p_{Others}$

Bolts	Machine of Concern	Other Machines	Total
Defective	12	15	27
Pass	88	185	273
Total	100	200	300

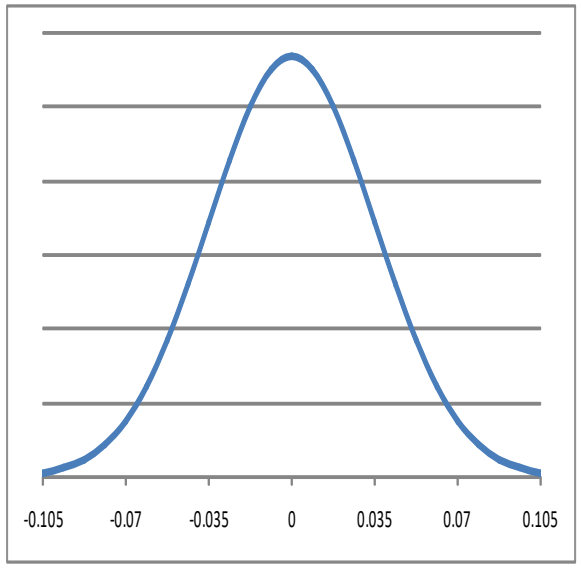
Difference in Sample Proportions Defective

Null Distribution: ?

Recall that sample proportions are averages in disguise; hence, we can make use of the Central Limit Theorem, and

Under the Null Model, Defective Rates Same = π

$$p_{Machine} - p_{Others} \sim N(0, \sqrt{\pi*(1-\pi)[1/n_M + 1/n_O]})$$



However, we do not know what under the null model would be a common defect rate, π, so we will need to estimate it from the available sample data.

Given the common defect rate under the null model, the best estimate of π would be:

$$p_{Defect} = (n_{Defect,M} + n_{Defect,O}) / (n_M + n_O)$$
$$= (12 + 15) / (100 + 200)$$
$$= 27 / 300$$
$$= 0.09$$

So … $p_{Machine} - p_{Others} \sim N(0, \sqrt{0.09*(1-0.09)[1/100 + 1/200]})$
$\sim N(0, 0.035)$,
where π estimated by $p_{Defect} = 0.09$

Independent Samples Data

Differences Between Two Proportions

Bolts	Machine of Concern	Other Machines	Total
Defective	12	15	27
Pass	88	185	273
Total	100	200	300

Decision Rule:

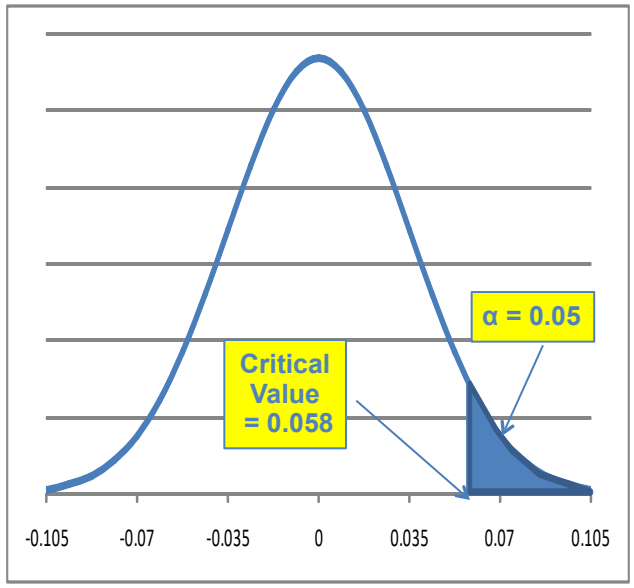
Type I Error: Reject H_0 when H_0 TRUE

Conclude the Machine is producing a higher rate of defectives when it is not

Type II Error: Fail to Reject H_0 when H_0 FALSE

Fail to recognize Machine is actually producing higher rate of defectives

Suggest α = 0.05



Reject if $p_{Machine} - p_{Others} > z_{(0.95)}^{1.645} * 0.035 = 0.058$

Decision:

$p_{Machine} - p_{Others} = (n_{Defect,M}/n_M) - (n_{Defect,O}/n_O)$

$= 12/100 - 15/200$

$= 0.12 - 0.075$

$= 0.045 < 0.058$, so

Fail to Reject H_0

Conclude: There is insufficient evidence to conclude Machine is producing higher defect rate than the Other machines in operation.

(p-Value = 0.0996)

Independent Samples Data
Differences Between Two Proportions

Again, just as with all the other Hypothesis Tests, this one also has a corresponding Confidence Interval. A (1-α)% Confidence Interval for $\pi_{Defective,Machine} - \pi_{Defective,Others}$ is given by:

Best Point Estimate $\pm M_{conf}$ * Standard Deviation of Best Point Estimate

$$p_{Machine} - p_{Others} \pm M_{(1-\alpha/2)} * \text{sqrt}[p_{Defect}*(1-p_{Defect})*(1/n_M) + (1/n_O)]$$

$$p_{Machine} - p_{Others} \pm z_{(0.975)} * \text{sqrt}[p_{Defect}*(1-p_{Defect})*(1/n_M) + (1/n_O)]$$

$$0.12 - 0.075 \pm 1.96 * \text{sqrt}[0.09*(1-0.09)*(1/100) + (1/200)]$$

$$0.045 \pm 0.069$$

$$(-0.024, 0.114)$$

So, we can state that with 95% confidence, $\pi_{Defective,Machine} - \pi_{Defective,Others}$ is between -2.4% and 11.4%. Note that this interval ***does*** include zero, which is consistent with the failure to reject H_0: $\pi_{Defective,Machine} - \pi_{Defective,Others} = 0$.

Conclusion is still that there is insufficient evidence in the sample data to conclude that the defect rate for the suspect machine is higher than that for the other machines in operation.

Independent Samples Data
Differences Between Two Proportions

What if we had Rejected H_0? Then using p_{Defect} in the calculation of the Standard Deviation of the Best Point Estimate (which amounts to "pooling" data) is no longer a valid approach – we have evidence the two proportions are different; hence their respective standard deviations are also different. In this case, a (1-α)% Confidence Interval for $\pi_{Defective,Machine} - \pi_{Defective,Others}$ is given by:

Best Point Estimate $\pm M_{conf}$ * Standard Deviation of Best Point Estimate

$$p_{Machine} - p_{Others} \pm z_{(1-\alpha/2)} * \text{sqrt}[p_{Machine}*(1-p_{Machine})/n_M + p_{Others}*(1-p_{Others})/n_O]$$

Suppose we observed 16 Defective Bolts in the sample of 100 from the suspect Machine, then $p_{Machine} = 0.16$, and would lead to rejection of H_0, then 95% CI would be:

$$p_{Machine} - p_{Others} \pm z_{(0.975)} * \text{sqrt}[p_{Machine}*(1-p_{Machine})/n_M + p_{Others}*(1-p_{Others})/n_O]$$

$$0.16 - 0.075 \pm 1.96 * \text{sqrt}[0.16*(1-0.16)/100) + 0.075*(1-0.075)/200)]$$

$$0.085 \pm 0.081$$

$$(0.004, 0.166)$$

Then we could state that with 95% confidence, $\pi_{Defective,Machine} - \pi_{Defective,Others}$ is between 0.4% and 16.6%. Note that this interval ***does not*** include zero, which is consistent with rejection of H_0: $\pi_{Defective,Machine} - \pi_{Defective,Others} = 0$.

Conclusion would now be that the defect rate for the suspect machine is 0.4% to 16.6% higher (with 95% confidence) than that for the other machines in operation.

Comparison of Two Proportions

There is a requirement for the normal approximation to be valid in the comparison of proportions: the expected number of results within each of the table "cells" should be larger than 5.

This is equivalent to $\min[n_1, n_2]*\min[p_{Tot}, (1-p_{Tot})] > 5$.

For the textile machine problem, the value of $\min[n_M, n_O] = 100$ is sufficiently large that this not an issue [$\min(p_{Tot}, 1-p_{Tot}) = 0.09$, and $0.09*100 = 9$].

However, consider a clinical trial of two drug therapies for leukemia: P and PV, where 21 patients were assigned to P and 42 were assigned to drug PV, with the following results:

	Drug		
Statistic	P	PV	Both
Number of Patients	21	42	63
Number Successful	14	38	52
Proportion Successful	0.667	0.905	0.825

Here,

$$\min[n_P, n_{PV}] = 21$$

and

$$\min[p_{Tot}, 1-p_{Tot}] = 0.175,$$

so

$$21*0.175 \approx 3.67 < 5$$

So, how can we test whether PV is in fact a more successful treatment than drug P?

Comparison of Two Proportions

Commonly, **Fisher's Exact Test** is utilized.

The idea is that if both drugs are equally successful in treating leukemia, then the total number of successes (here, this is 52) should be distributed across the drugs reasonably close to proportionally to the respective numbers of patients treated with each of the drugs.

	Drug		
Statistic	P	PV	Both
Number of Patients	21	42	63
Number Successful	14	38	52
Proportion Successful	0.667	0.905	0.825

So, as the number of successes increases for the group with the highest number of successes, this would suggest that the true probability of success for this drug (here, π_{PV}) is greater than that for the other (ie, π_P).

We can use the hypergeometric distribution to assess the probability of observing at least as many PV drug successes as were actually observed if the probability of successful treatments is the same (ie, $\pi_{PV} = \pi_P$).

The test statistic is $X = \max(p_P, p_{PV})*n_{\max(p_P,p_{PV})}$ = number of successes for group with highest success rate (X=38 here). Under H_0: $\pi_{PV} = \pi_P$, the p-Value for this test is given by:

$$\sum_{j=X \text{ to } \min(n_{PV},n_S)} P[j \text{ Successes of } n_{PV} \text{ Patients} \mid p_P n_P + p_{PV} n_{PV} = n_S \text{ Total Successes}] =$$

$$\sum_{j=X \text{ to } \min(n_{PV},n_S)} C(n_{PV},j)*C(n_P, n_S-j)/C(n_{Tot}, n_S) = 0.0254, \text{ where } C(a, b) = a!/[b!(a-b)!],\ a \geq b$$

Hence, if $\alpha > 0.0254$, then there is sufficient evidence to reject H_0, and conclude that the PV drug therapy has a higher success rate in treating leukemia than the drug P.

Comparison of Two Proportions

Textile Machines

Minitab: Stat→ Basic Statistics → 2 Proportions

Test and CI for Two Proportions

Sample	X	N	Sample p
1	12	100	0.120000
2	15	200	0.075000

Difference = p (1) - p (2)
Estimate for difference: 0.045
95% CI for difference: (-0.0284104, 0.118410)
Test for difference = 0 (vs not = 0): Z = 1.20 P-Value = 0.230

Fisher's exact test: P-Value = 0.206

Test and CI for Two Proportions

Sample	X	N	Sample p
1	16	100	0.160000
2	15	200	0.075000

Difference = p (1) - p (2)
Estimate for difference: 0.085
95% CI for difference: (0.00440579, 0.165594)
Test for difference = 0 (vs not = 0): Z = 2.07 P-Value = 0.039

Fisher's exact test: P-Value = 0.027

Drug Trials

Minitab: Stat→ Basic Statistics → 2 Proportions

Test and CI for Two Proportions

Sample	X	N	Sample p
1	38	42	0.904762
2	14	21	0.666667

Difference = p (1) - p (2)
Estimate for difference: 0.238095
95% lower bound for difference: 0.0532147
Test for difference = 0 (vs > 0): Z = 2.35 P-Value = 0.009

Fisher's exact test: P-Value = 0.025

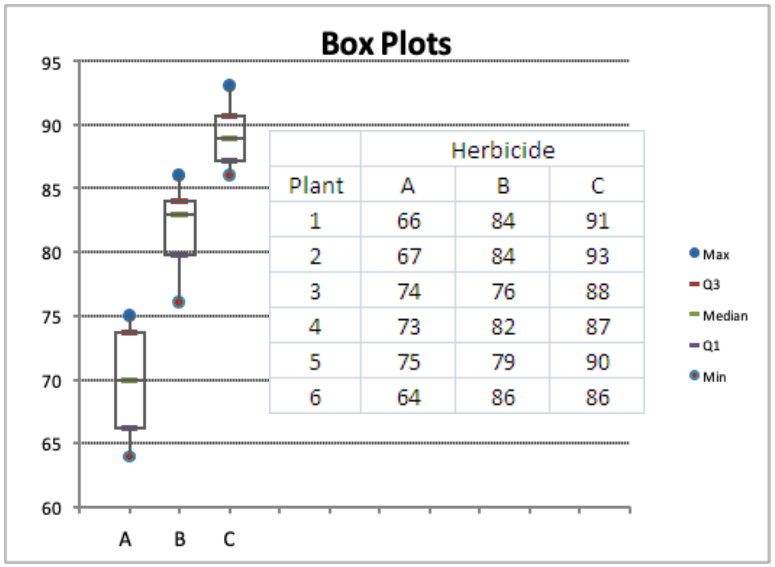
* NOTE * The normal approximation may be inaccurate for small samples.

Sample Data From 3 or More Groups

Frequently, more than a single separation of a relevant data set is involved resulting in comparisons across more than two groups.

When we move into multiple comparisons for more than two groups, the approaches we have considered to this point need to be modified to be successfully extended.

Consider an example where several different herbicide mixtures (A, B, & C) are to be evaluated for their effect on plant growth. Six plants were assigned at random to receive one of these herbicides during the study period. The data are the additional growth for each plant over the study period (in cm).



	Herbicide		
Plant	A	B	C
1	66	84	91
2	67	84	93
3	74	76	88
4	73	82	87
5	75	79	90
6	64	86	86

Research Hypothesis: H_1: $\mu_i \neq \mu_j$, for some $i \neq j$,
$i,j = A, B, C$

Null Hypothesis: H_0: $\mu_A = \mu_B = \mu_C$

Test Statistic: $F = S^2_{Between}/S^2_{Within}$

Provided data are normally distributed & variances within groups are equal

Null Distribution: $F_{(n_G-1,\ n-n_G)}$ (where n_G=3, the number of groups)

Decision Rule: Reject H_0 if $F > F_{(n_G-1,\ n-n_G,\ 1-\alpha)}$

Decision: $F = S^2_{Between}/S^2_{Within} >$
$= 571.6/14.3$
$= 39.97 > F_{(2,15,\ 0.95)} = 3.68$,
so Reject H_0 in Favor of H_1.

Conclude: The average additional growth is different with respect to the different herbicide applied.

So why this test statistic and how is it specifically determined from the data?

Sample Data From 3 or More Groups

When more than two groups are involved and we are interested in determining if the respective populations from which the samples within each have been acquired have different population mean values, then the approach most often utilized is ***Analysis of Variance***.

The approach, often abbreviated as ***ANOVA***, simply compares the variation between group averages to the variation of the results within groups. Hence, the name refers to "variance" when it is actually an evaluation of the group means.

The concept is that if there are no differences between the population means from which the respective sample groups have been obtained, then all the data is from essentially the same population, and the group averages* would not be expected to vary any more than the individual data values. The "*" is included to note that these group averages need to be properly scaled … why?

The CLT tells us that Var(X-bar) = σ^2/n, where σ^2 is the variance of individual results and n is the number of

The most popular means of outlining an Analysis of Variance is through use of an ANOVA table:

Herbicide Data

ANOVA					
Source	df	SS	MS	F	p-Value
Herbicide	2	1143.111	571.556	39.969	9.8E-07
Error	15	214.5	14.3		
Total	17	1357.611			

Note: MSB = $n_W * S_B^2$, where S_B^2 = Variance of the Group Averages, and MSW = S_W^2, where S_W^2 = Average of the Within Group Variances

If no difference in group means, these are both estimators of σ^2, the population variance.

In general,

Source: Identifies the ***Factors*** involved, last two rows are virtually always "Error" and "Total"

df: Respective ***Degrees of Freedom*** for each Source of Variation, for a 1-way ANOVA $df_{Factor} = n_G - 1$, where n_G = number of groups

SS: ***Sums of Squares*** for each Source of Variation, for a balanced 1-way ANOVA, these are:

Factor: $n_W \sum_{i=1 \text{ to } n_G} (Y_i\text{-bar} - Y\text{-bar})^2$, n_W = number within each group

Error: $\sum_{i=1 \text{ to } n_G} \sum_{j=1 \text{ to } n_W} (Y_{ij} - Y_i\text{-bar})^2$

Total: $\sum_{i=1 \text{ to } n_G} \sum_{j=1 \text{ to } n_W} (Y_{ij} - Y\text{-bar})^2$, SS_{Total} = Sum of all rows above

MS: ***Mean Squares*** for each Source of Variation, $MS_i = SS_i/df_i$, for all rows of the table, except the "Total" row.

F: ***F statistic(s),*** generally MS_i/MSE, where MSE = MS for "Error" row

p-Value: Respective ***p-Values*** for each F statistic

Sample Data From 3 or More Groups

Note that the implied statistical model involved here is a little more involved than the simple Mean + Error model outlined previously. The model implied in the simple, balanced 1-way ANOVA situation of the previous examples is:

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1, \ldots, g$ (the number of groups) and $j = 1, \ldots, m$ (number of results for each group)

where Y_{ij} = the jth result from the ith group,

μ = an overall mean level for the data,

α_i = an offset for the mean level of the ith group from the overall mean (so $\mu_i = \mu + \alpha_i$),

ε_{ij} = random error term

The assumptions for this model include $\Sigma_{i=1 \text{ to } g} \alpha_i = 0$ and $\varepsilon_{ij} \sim$ NID$(0, \sigma^2)$.

So … when we reject H_0: all μ_i equal in favor of H_1: $\mu_i \neq \mu_j$ for some $i \neq j$, the result is still unsatisfactory. Why?

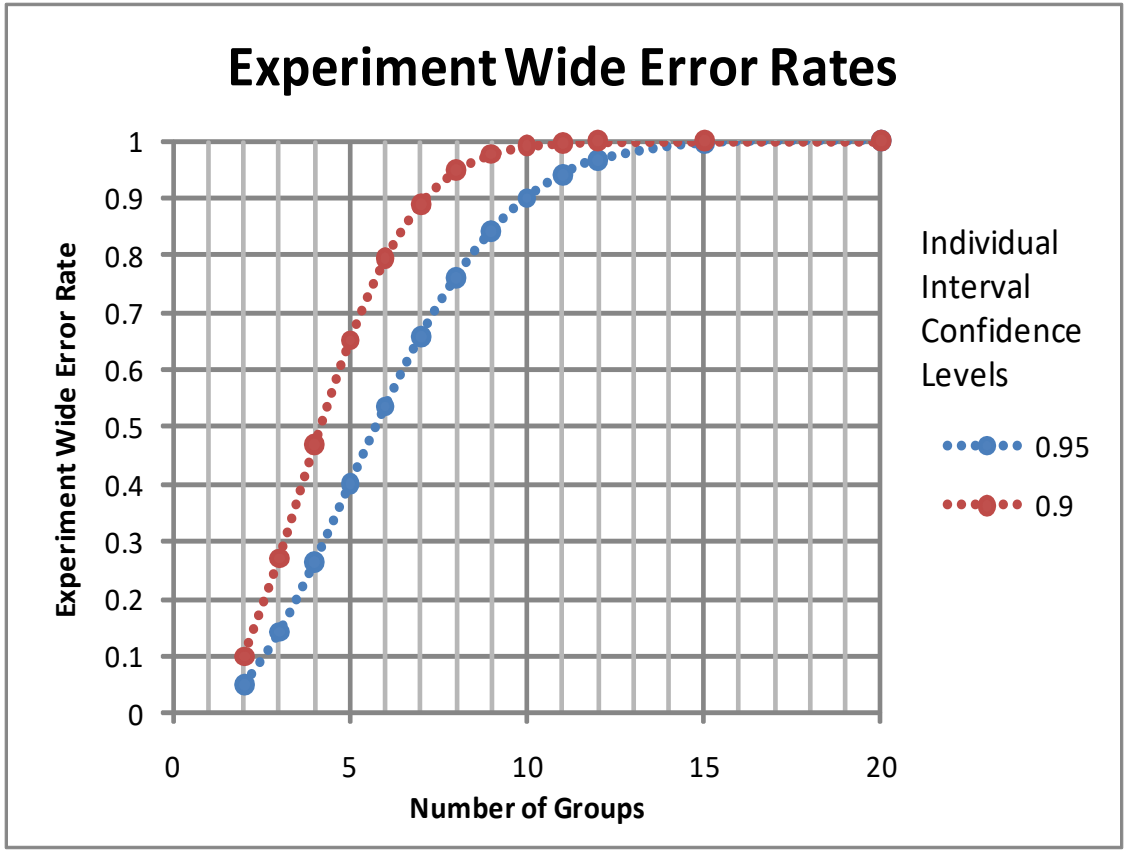
For single and two-sample problems, the tool used to address the "How Much?" question was a confidence interval. A similar approach can be used with 3 or more groups, but there is a need to be careful. Why?

Sample Data From 3 or More Groups

For g groups, there are effectively g*(g-1)/2 comparisons to be made.

As more intervals are required and the likelihood of all intervals containing their respective parameters can be much lower than the stated confidence level for any one interval.

For g = 3, there are only 3 comparisons to be made, but if we use 90% confidence level for all the relevant intervals, then the probability of one not including the true $\mu_i - \mu_j$ value is relatively large (~27% = 1 - $.9^3$).



Sample Data From 3 or More Groups

There are a variety of approaches to address this issue and these are all described in some detail in the text (see Chapter 9). Effectively, they all are of the form:

$$\bar{Y}_i - \bar{Y}_j \pm M(\alpha) * \text{Std Deviation of } (\bar{Y}_i - \bar{Y}_j)$$

where the Std Deviation$(\bar{Y}_i - \bar{Y}_j) = \text{Sqrt}(2*MSE/m)$, with MSE = Mean Square Error from the associated ANOVA table, and m = number of results for each group.

The differences between the most widely used approaches involve the multiplier $M(\alpha)$.

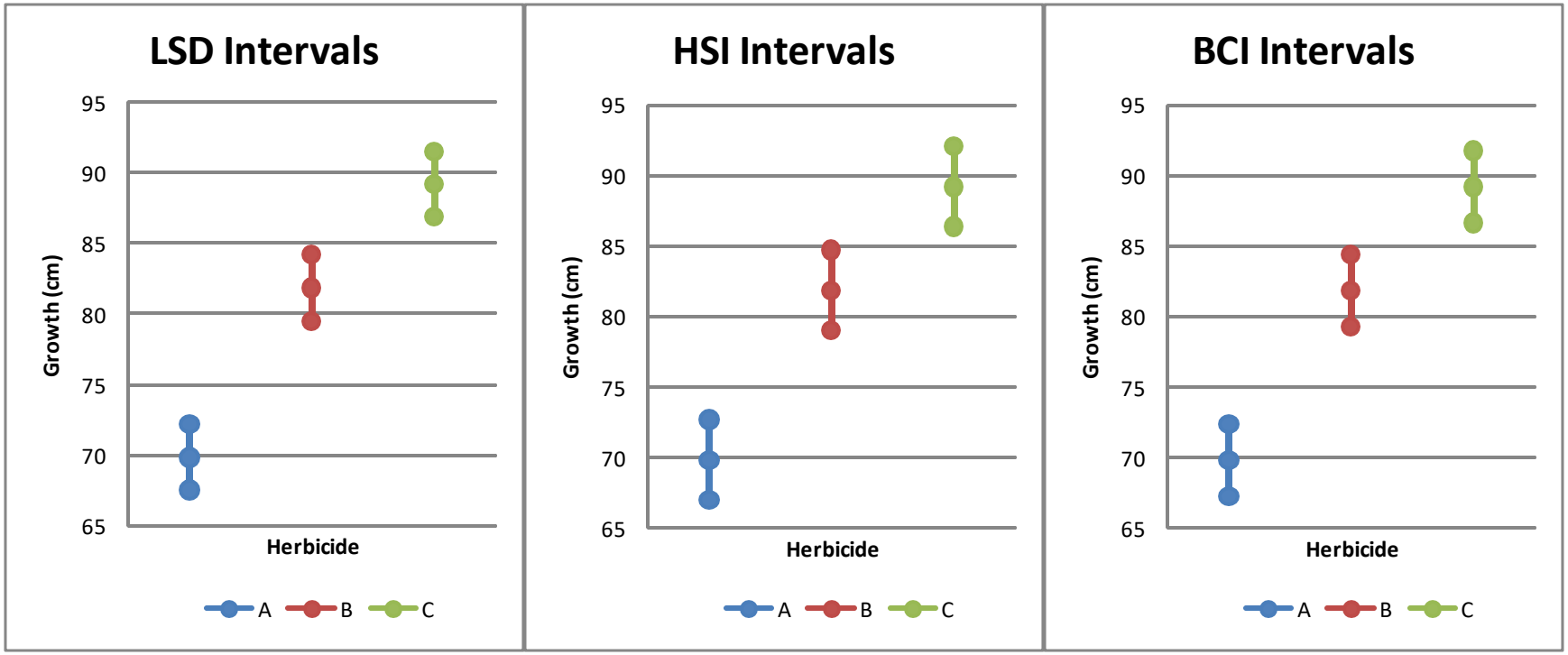
Sample Data From 3 or More Groups

Regardless of which approach is used, graphically displaying the results can become complicated, especially as the number of groups (g) increases.

One way to overcome this is to plot each group mean ($\bar{Y}_i$) ± half the width of the respective interval being used. On such a plot, and intervals that do not overlap indicate differences between those group means.

Be careful not to use these intervals as confidence intervals for the specific means. The intervals being plotted will be narrower than intervals with the same confidence for each specific group mean.

For the herbicide example, the respective half-width intervals appear below ($\alpha = 0.05$):



Clearly, in this case, all the group means are different, with Herbicide C providing the most growth and Herbicide A the least. Actually, Herbicide A in this instance was a control group (ie, no herbicide).

Sample Data From 3 or More Groups

What about groups with unequal variances?

Recall that the 1-way ANOVA procedure assumes similar variation for each group. Can we test to see if this is valid?

One test is to simply consider the ratio of the largest within group variance to that of the smallest (Hartley's F_{max} Test):

$$F_{max} = S_{max}^2/S_{min}^2$$

Critical values for this test appear in Table 12, for $\alpha = 0.05$ and 0.01. Each table is entered with the number of groups involved (Table columns) & the common group sample size less 1 (ie, $n_W - 1$, for Table rows). Note this test was developed assuming equal sample sizes for each group, but if they are close, then can use something like a harmonic mean of the n_i values.

Another test for equal variances would be the Brown-Forsythe-Levine (BFL) test. This test essentially involves a 1-way ANOVA of a newly constructed variable $v_{ij} = |y_{ij} - m_i|$, where m_i = median of the results for the i^{th} group. This can involve some computations, but is available in many software packages.

If no such package is available, then the BFL test statistic is calculated as:

$$L = \{\sum_{i=1 \text{ to } g} n_i(\text{V-bar}_i - \text{V-bar})^2/(g - 1)\} / \{\sum_{i=1 \text{ to } g} \sum_{j=1 \text{ to } n_i} (v_{ij} - \text{V-bar}_i)^2/(N - g)\},$$

where g = number of groups and $N = \sum_{i=1 \text{ to } g} n_i$; which is simply MSB_v/MSE_v, so the critical value is determined from the upper tail of an $F_{g-1,N-g}$ distribution. Note L can be evaluated for unequal group sizes.

This test is preferred also because it is less sensitive to departures from normality. Although the Hartley test is more powerful when the data is indeed normal, it is usually more desirable to keep the Type I error rate low in these tests, and this can become large for the F_{max} test as conditions depart from the basis for its development (ie, equal group sizes and normally distributed data). Also, a special table of critical values is unnecessary for the BFL test.

Sample Data From 3 or More Groups

Herbicide example:

Raw Data

	Herbicide		
Plant	A	B	C
1	66	84	91
2	67	84	93
3	74	76	88
4	73	82	87
5	75	79	90
6	64	86	86
Median	70	83	89

Absolute Deviations from Median

	Herbicide		
Plant	A	B	C
1	4	1	2
2	3	1	4
3	4	7	1
4	3	1	2
5	5	4	1
6	6	3	3

ANOVA for above

Source	df	SS	MS	F	p-Value
Herbicide	2	12.4444	6.22222	2.1961	0.14571
Error	15	42.5	2.83333		
Total	17	54.9444			

Research Hypothesis: H_1: Population variances not all equal.
Null Hypothesis: H_0: Population variances all equal.

Test Statistic: L
Null Distribution: $F_{g-1,N-g}$

Decision Rule: Reject H_0 if $L > F_{(2,15,0.95)} = 3.682$

Decision: $L = 2.1961 < 3.682$; hence, Fail to Reject H_0

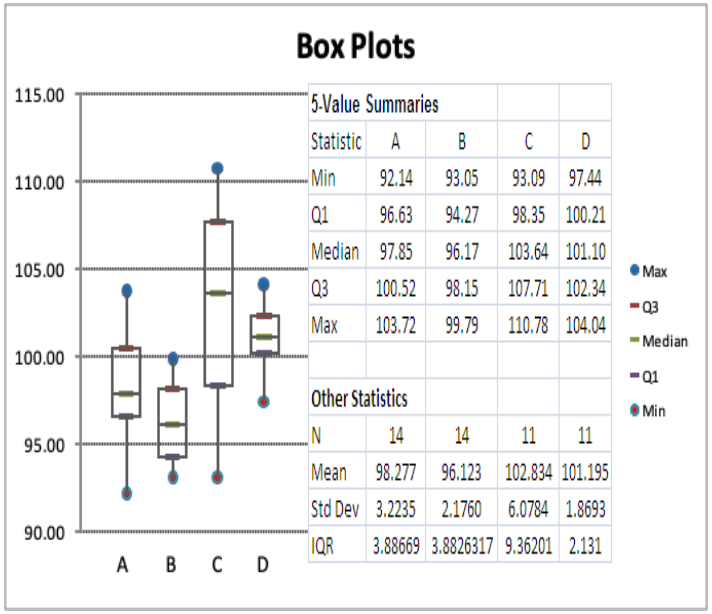
Conclusion: Insufficient evidence in the data to indicate population variances are not all the same, or at least nearly the same (at significance level 0.05, p-Value ≈ 0.146).

Note that $F_{max} = S_A^2/S_C^2 = 3.182$,

which is less than the corresponding Table 12 critical value for 3 groups, group size = 6, and $\alpha = 0.05$ of 10.8. Hence, this result is consistent with the BFL test result.

Sample Data From 3 or More Groups

Workshop 1 Example:

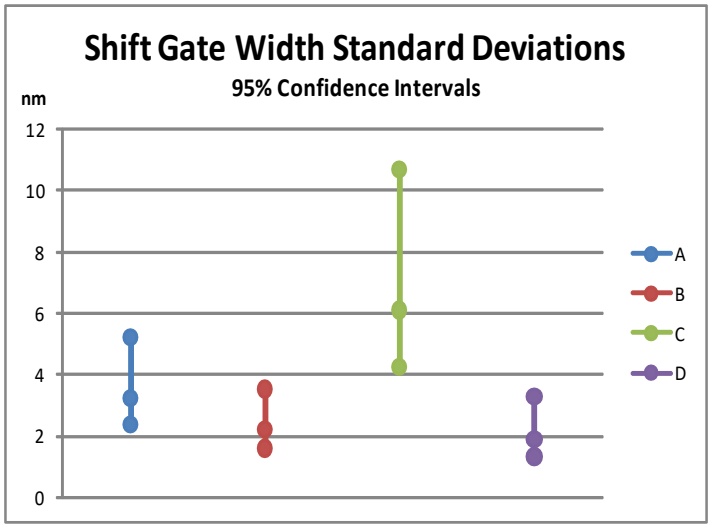


5-Value Summaries

Statistic	A	B	C	D
Min	92.14	93.05	93.09	97.44
Q1	96.63	94.27	98.35	100.21
Median	97.85	96.17	103.64	101.10
Q3	100.52	98.15	107.71	102.34
Max	103.72	99.79	110.78	104.04

Other Statistics

N	14	14	11	11
Mean	98.277	96.123	102.834	101.195
Std Dev	3.2235	2.1760	6.0784	1.8693
IQR	3.88669	3.8826317	9.36201	2.131



Research Hypothesis: H_1: Population variances not all equal.
Null Hypothesis: H_0: Population variances all equal.

Test Statistic: L
Null Distribution: $F_{g-1,N-g}$

Decision Rule: Reject H_0 if $L > F_{(3,46,0.95)} = 2.807$

Decision: $L = 8.893 > 2.807$; hence, Reject H_0

Conclusion: The population variances are not all the same (at significance level 0.05, p-Value ≈ 0.000093).

So, the σ_i^2 are not all equal, but how are they different?

There are several possible ways to handle this.
Probably the most technically correct would be to compare all six possible ratios of pairs of S_i using Bonferroni-type confidence intervals based on F-distributions.
However, providing simple confidence intervals for each group might also suffice and be easier to communicate.

The real concern is what if we are most interested in differences in group means? The standard ANOVA procedure assumes equal variances within groups, but if evidence to contrary, then …

1) Could consider nonparametric approach
2) Could "normalize" the data (eg, divide by S_i within each group)
3) Perhaps a transformation of some kind (eg, ln(y))

Sample Data From 3 or More Groups

MINITAB: Stat → ANOVA → One-Way (Unstacked)

Comparisons button: Chose Tukey & Fisher

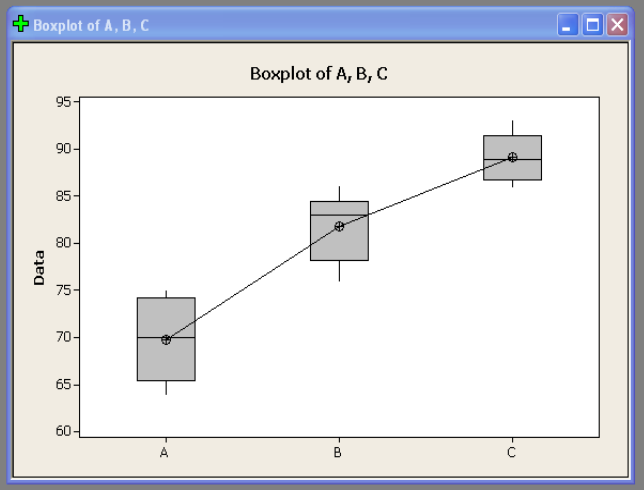
Graphics button: Chose Box Plots & 3 in 1 Residual Plots

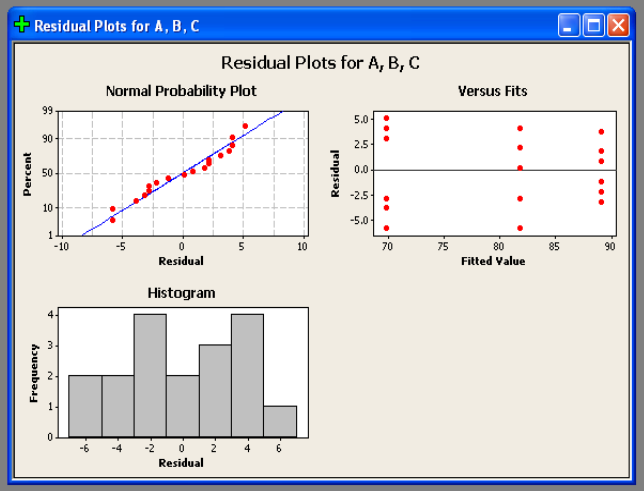
Session

```
One-way ANOVA: A, B, C

Source  DF      SS     MS      F      P
Factor   2  1143.1  571.6  39.97  0.000
Error   15   214.5   14.3
Total   17  1357.6

S = 3.782   R-Sq = 84.20%   R-Sq(adj) = 82.09%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  -----+---------+---------+---------+----
A      6  69.833  4.708  (----*---)
B      6  81.833  3.710                   (----*----)
C      6  89.167  2.639                             (---*----)
                          -----+---------+---------+---------+----
                            70.0      77.0      84.0      91.0
Pooled StDev = 3.782

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 97.97%

A subtracted from:
    Lower  Center  Upper  ---+---------+---------+---------+------
B   6.334  12.000  17.666                (-----*-----)
C  13.668  19.333  24.999                          (-----*-----)
                          ---+---------+---------+---------+------
                           -10         0        10        20
B subtracted from:
   Lower  Center  Upper  ---+---------+---------+---------+------
C  1.668   7.333  12.999             (-----*-----)
                          ---+---------+---------+---------+------
                           -10         0        10        20

Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons

Simultaneous confidence level = 88.31%

A subtracted from:
Lower  Center  Upper  --+---------+---------+---------+-------
B   7.346  12.000  16.654                (----*----)
C  14.680  19.333  23.987                          (----*----)
                          --+---------+---------+---------+-------
                           -10         0        10        20

B subtracted from:
   Lower  Center  Upper  --+---------+---------+---------+-------
C  2.680   7.333  11.987             (----*----)
                          --+---------+---------+---------+-------
                          -10         0        10        20
```

Worksheet 1 ***

	C1	C2	C3	C4	C5
	Plant	A	B	C	
1	1	66	84	91	
2	2	67	84	93	
3	3	74	76	88	
4	4	73	82	87	
5	5	75	79	90	
6	6	64	86	86	
7					
8					





Two-Way Analysis of Variance

The Herbicide data was analyzed with what if referred to as a One-Way Analysis of Variance.
It is only a One-Way analysis because the groups are only identified in a single way (ie, by the type of Herbicide used).
Often, groups are identified in more than a single way. Consider, for example, the evaluation of Crop Yield for plots treated with different amounts of Nitrogen and Phosphorus:

Plot	Nitrogen (lbs)	Phosphorus (lbs)	Crop Yield (bu/ac)
1	60	20	183
2	60	20	178
3	60	10	147
4	60	10	143
5	40	20	162
6	40	20	156
7	40	10	122
8	40	10	127

Note that each plot is now classified two ways: Amount of Nitrogen applied & Amount of Phosphorus applied.

Now we have two factors or treatments to consider rather than just one.

Note that the implied statistical model involved here is a little more involved than the One-Way ANOVA model.
The model implied in the simple, balanced 2-way ANOVA situation present in this example is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

with $i = 1, \ldots, a$ (the number of levels for Factor A),
$j = 1, \ldots, b$ (the number of levels for Factor B),
and $k = 1, \ldots, m$ (number of results for each group)

where Y_{ijk} = the kth result from group (i, j),
μ = an overall mean level for the data,
α_i = an offset for the mean level when Factor A = i from the overall mean
β_j = an offset for the mean level when Factor B = j from the overall mean (so $\mu_{ij} = \mu + \alpha_i + \beta_j$),
ε_{ijk} = random error term

The assumptions for this model include $\Sigma_{i=1 \text{ to } a} \alpha_i = 0$, $\Sigma_{j=1 \text{ to } b} \beta_j = 0$, and $\varepsilon_{ijk} \sim$ NID$(0, \sigma^2)$.

Two-Way Analysis of Variance

A formal hypothesis test is generally built around the relevant ANOVA Table.

Research Hypothesis: H_1: $\mu_{ij} \neq \mu_{i'j'}$, for some $i \neq i'$ and/or $j \neq j'$, $i,i' = 1,\ldots,a$, $j,j' = 1, \ldots, b$

Null Hypothesis: H_0: $\mu_{ij} = \mu_{i'j'}$, for all $i,i' = 1,\ldots,a$ and $j,j' = 1,\ldots,b$

Could have single Test Statistic:
$F_{All} = [(SSA + SSB)/(a+b-2)]/MSE$

Test Statistics: $F_A = MSA/MSE$ & $F_B = MSB/MSE$

Provided data within groups are normally distributed & variances within groups are equal

Null Distributions: $F_{(a-1,\ abm-a-b+1)}$ & $F_{(b-1,\ abm-a-b+1)}$ (where a=2 & b=2, the number of levels for Factors A &B, respectively)

Null Distribution:
$F_{All} \sim F_{(a+b-2,abm-a-b+1)}$

ANOVA Table

Source	df	SS	MS	F
Factor A	a-1	SSA = $bm\sum_a(Y_i\text{-bar} - Y\text{-bar})^2$	MSA = SSA/(a-1)	F_A = MSA/MSE
Factor B	b-1	SSB = $am\sum_b(Y_j\text{-bar} - Y\text{-bar})^2$	MSB = SSB/(b-1)	F_B = MSB/MSE
Error	abm-a-b+1	SSE = $\sum_a\sum_b\sum_m(Y_{ijk} - Y_i\text{-bar} - Y_j\text{-bar} + Y\text{-bar})^2$	MSE = SSE/(abm-a-b+1)	
Total	abm-1	TSS =		

Decision Rule: With $\alpha = 0.05$,

Reject if either F statistic is larger than its corresponding 95th percentile point

Decision Rule: Reject if
$F_{All} > F_{(a+b-2,abm-a-b+1,1-\alpha)}$

ANOVA					
Source	df	SS	MS	F	p-Value
Nitrogen	1	882	882	85.631	0.000248
Phosphorus	1	2450	2450	237.86	2.08E-05
Error	5	51.5	10.3		
Total	7	3383.5			

Decision: Reject H_0 since both F_A & $F_B > 6.608$

Decision: Reject since $F_{All} = 161.75 > F_{(a+b-2,abm-a-b+1,0.95)} = 5.786$

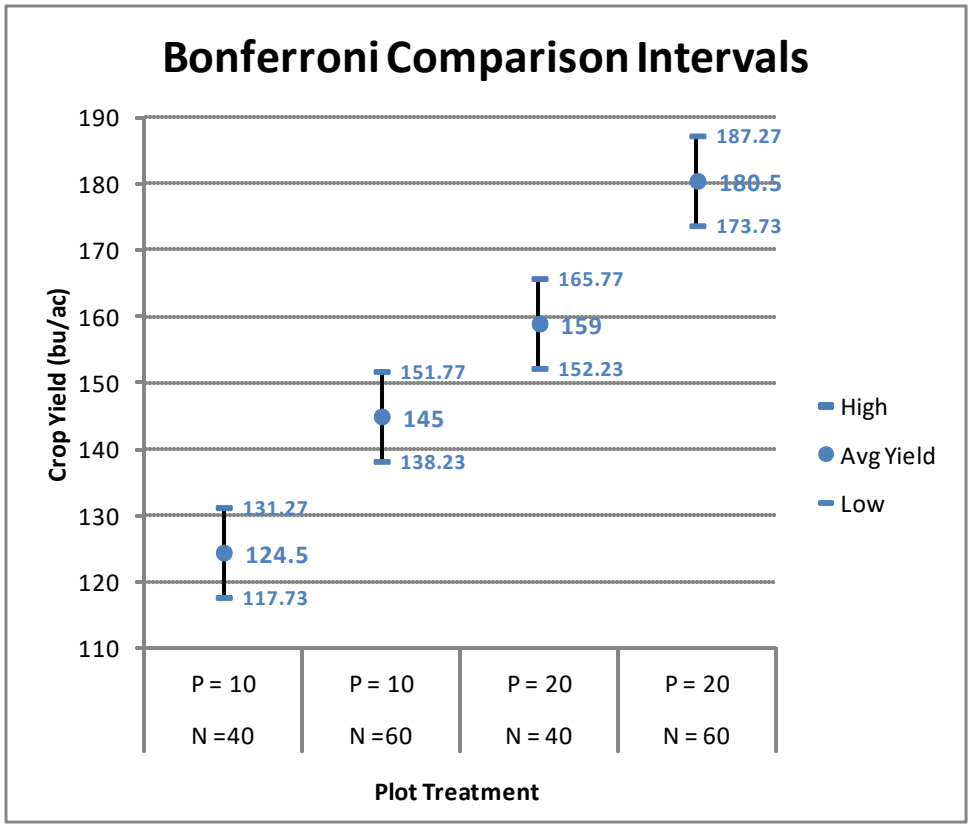
Conclusion: Application of Nitrogen and Phosphorus effects Crop Yield (p-Value < 0.00025)

Conclusion: Same (p-Value = 0.000029)

Two-Way Analysis of Variance

Similar to most hypothesis tests, a reject decision provides a conclusion that leaves unanswered questions. For the Crop Yield data, the conclusion is that application of Nitrogen and Phosphorus impacts Crop Yield; however, it is not clear how they do so.

Again, we resort to a suitable multiple comparison technique to compare the group averages. Note that intervals below are half-width confidence intervals so those that do not overlap suggest significant differences in corresponding population means.



These intervals were obtained by first generating the actual Bonferroni confidence intervals for the difference between any two of the respective group means:

$$(Y_{ij}\text{-bar} - Y_{i'j'}\text{-bar}) \pm t_{(abm-a-b+1,\ 1-\alpha/[ab*(ab-1)])} * \text{sqrt}(2MSE/m)$$

Then take the half-width of these intervals (ie, the second row above divided by 2) and add and subtract it from each of the respective group means:

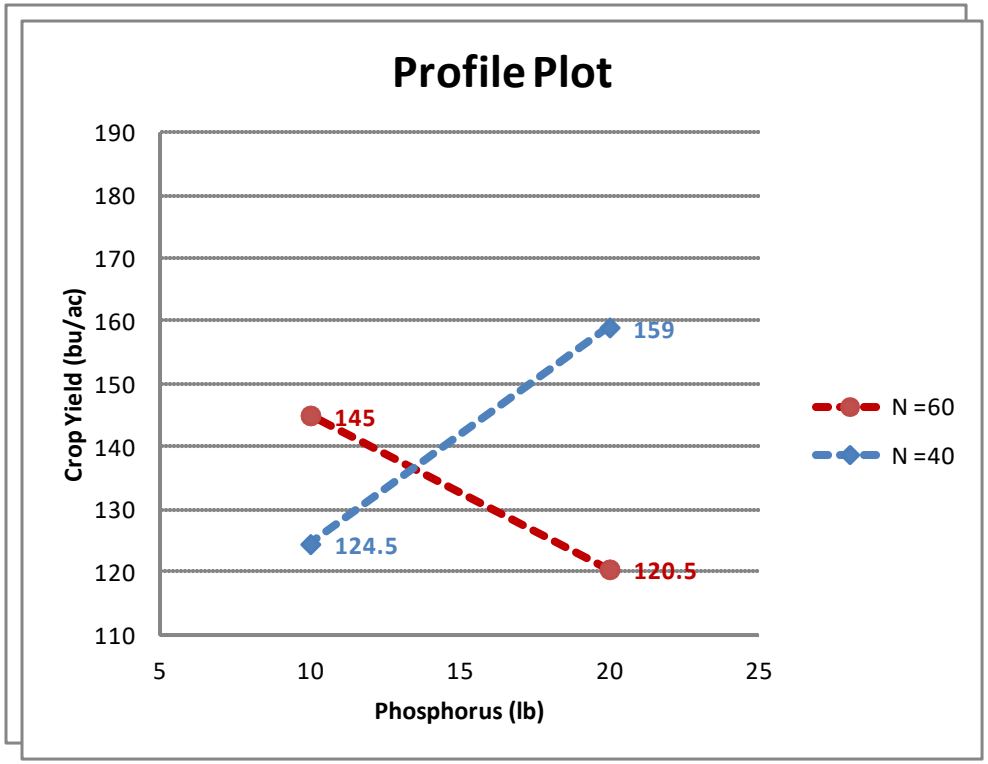
$$Y_{ij}\text{-bar} \pm (1/2) * t_{(abm-a-b+1,\ 1-\alpha/(ab*[ab-1])} * \text{sqrt}(2MSE/m)$$

Note that this is ~70% (sqrt(2)/2) of the width of a Bonferroni confidence interval for μ_{ij} alone.

It appears Yield increases with application of both chemicals (~1bu/ac per 1lb N & ~3.5bu/ac per 1lb P).

Two-Way Analysis of Variance

Another type of plot often used in Analyses of Variance is the Profile Plot, especially when at least one of the group identifiers is continuous in nature. Both of these identifiers (Amt of N & Amt of P) are continuous in this case.



Clearly, the profile plot changes significantly. What about the ANOVA?

Note that this plot displays the effect of applying additional Phosphorus (P) for each specific application of Nitrogen (N).

The form of the relationship between Crop Yield and Phosphorus is depicted as linear in nature since only two levels of Phosphorous application were evaluated.

The lines being virtually parallel is an indication that the purely additive two-way model evaluated is appropriate.

However, what happens if we arbitrarily subtract 60 bu/ac Yield from the observed values for the plots receiving 60 lbs of Nitrogen and 20 lbs of Phosphorus?

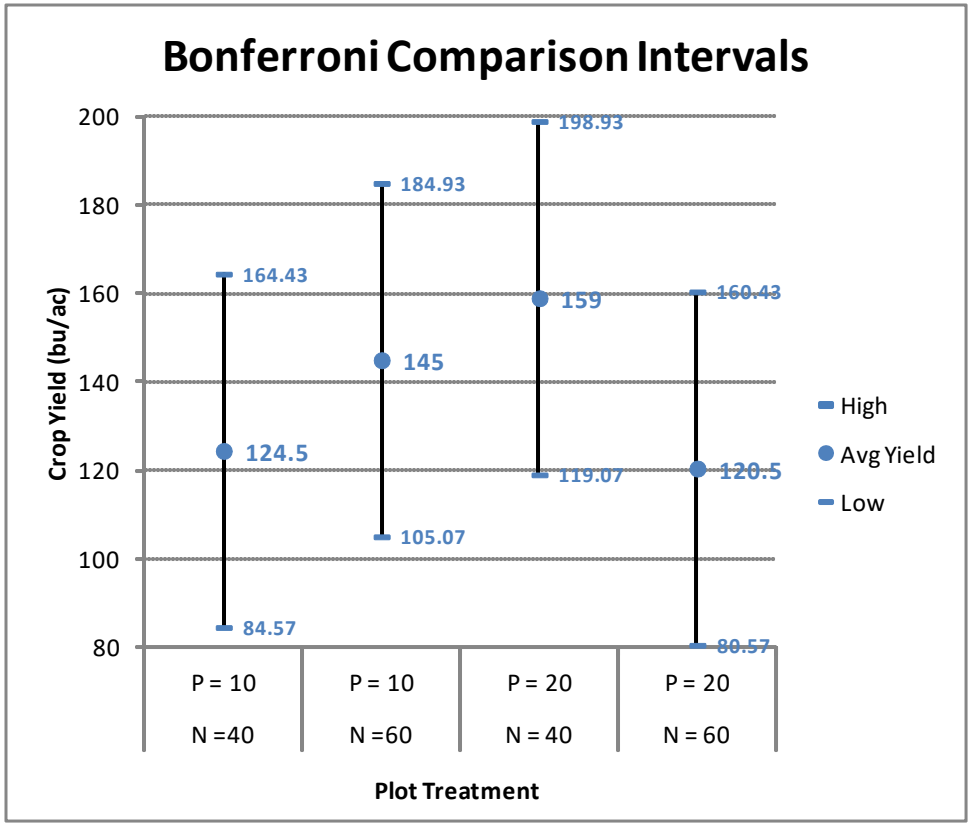
Two-Way Analysis of Variance

Data

Plot	Nitrogen (lbs)	Phosphorus (lbs)	Crop Yield (bu/ac)
1	60	20	123
2	60	20	118
3	60	10	147
4	60	10	143
5	40	20	162
6	40	20	156
7	40	10	122
8	40	10	127

Analysis of Variance Table

ANOVA						
Source	df	SS	MS	F	p-Value	F-Critical
Nitrogen	1	162	162	0.452	0.531	6.608
Phosphorus	1	50	50	0.140	0.724	6.608
Error	5	1791.5	358.3			
Total	7	2003.5				



What happened?!?

Nothing appears significant any longer,

1) Both F statistics are < 1, with p-Values > 0.5,
2) All multiple comparison intervals overlap, and
3) The MSE ~35X larger than before (which is the root of the problem),

but we did not change the within group variation at all!

Two-Way Analysis of Variance

The problem here is that the simple two-way model, which is purely additive in nature is no longer appropriate. The departure from parallel for the lines in the Profile Plot indicate the need for a multiplicative term in the model. This is the role of what is called the Interaction term which needs to be added to the original simple additive model.

The new, and more general two-way, balanced ANOVA model including Interaction is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk},$$

with $i = 1, \ldots, a$ (the number of levels for Factor A),
$j = 1, \ldots, b$ (the number of levels for Factor B),
and $k = 1, \ldots, m$ (number of results for each group)

where Y_{ijk} = the kth result from group (i, j),
μ = an overall mean level for the data,
α_i = an offset for the mean level when Factor A = i from the overall mean
β_j = an offset for the mean level when Factor B = j from the overall mean
$\alpha\beta_{ij}$ = a multiplicative effect for Factors A and B for group (i, j) (so, now $\mu_{ij} = \mu+\alpha_i+\beta_j+\alpha\beta_{ij}$),
ε_{ijk} = random error term

The assumptions for this model include $\Sigma_{i=1 \text{ to } a} \alpha_i = 0$, $\Sigma_{j=1 \text{ to } b} \beta_j = 0$, $\Sigma_{i=1 \text{ to } a}\Sigma_{j=1 \text{ to } b} \alpha\beta_{ij} = 0$, and $\varepsilon_{ijk} \sim$ NID$(0, \sigma^2)$.

Two-Way Analysis of Variance

The New ANOVA Table including the accounting for Interaction is given as:

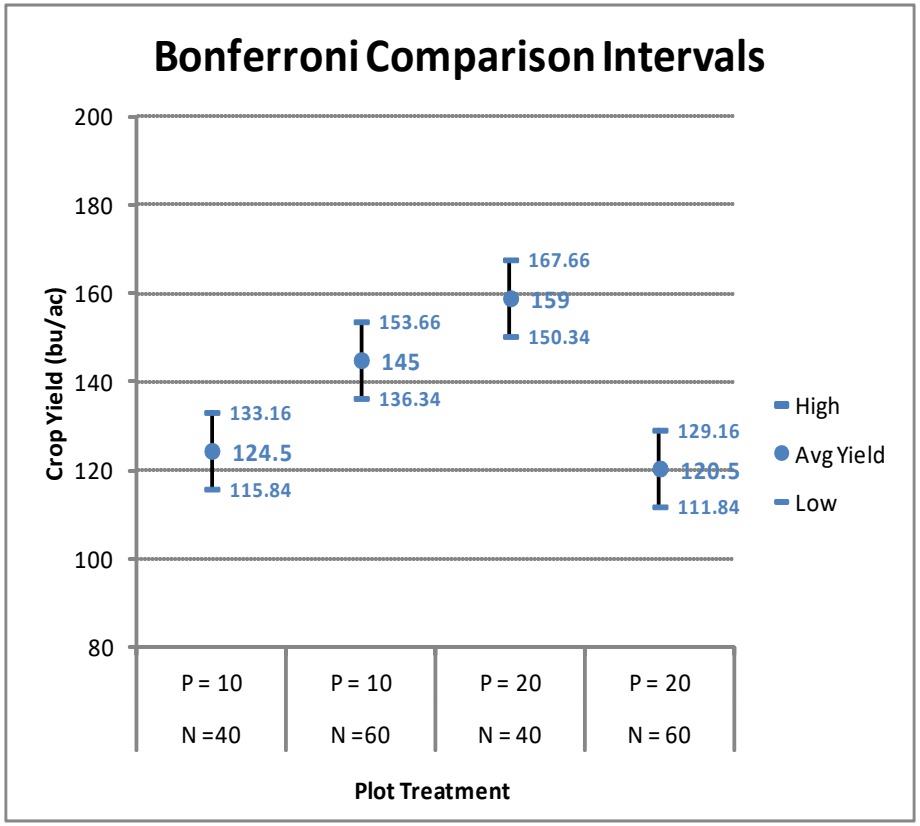
Source	df	SS	MS	F
Factor A	a-1	SSA = $bm\sum_a(Y_i\text{-bar} - Y\text{-bar})^2$	MSA = SSA/(a-1)	F_A = MSA/MSE
Factor B	b-1	SSB = $am\sum_b(Y_j\text{-bar} - Y\text{-bar})^2$	MSB = SSB/(b-1)	F_B = MSB/MSE
Interaction	(a-1)(b-1)	SSAB = $m\sum_a\sum_b(Y_{ij}\text{-bar} - Y_i\text{-bar} - Y_j\text{-bar} + Y\text{-bar})^2$	MSAB = SSAB/[(a-1)(b-1)]	F_{AB} = MSAB/MSE
Error	ab(m-1)	SSE = $\sum_a\sum_b\sum_m(Y_{ijk} - Y_{ij}\text{-bar})^2$	MSE = SSE/[ab(m-1)]	
Total	abm-1	TSS = $\sum_a\sum_b\sum_m (Y_{ijk} - Y\text{-bar})^2$		

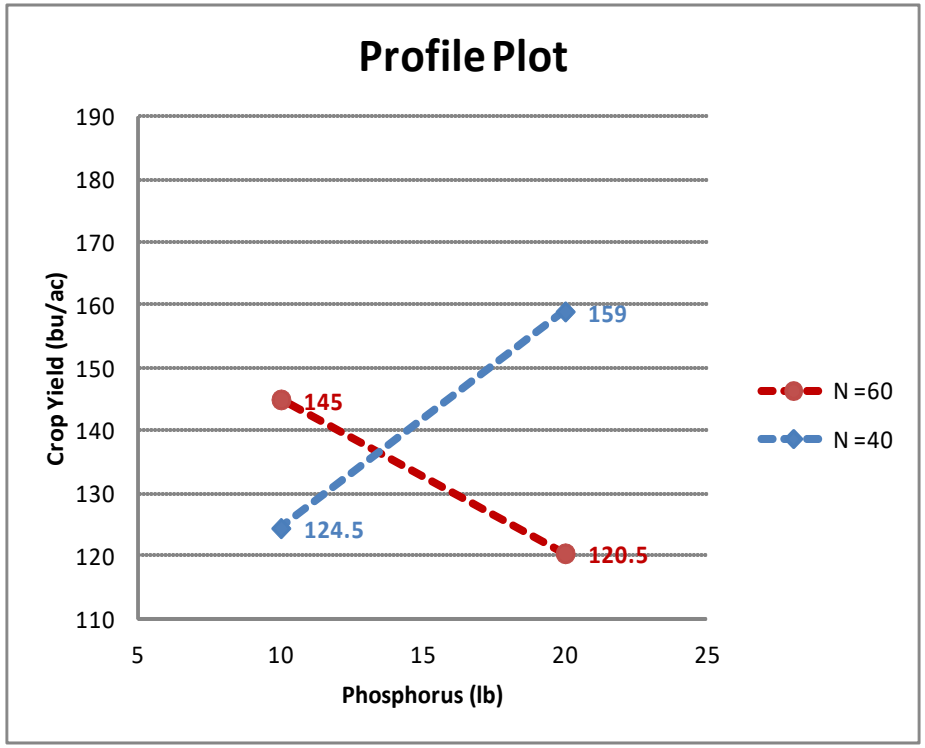
The Error df and SS are separated into two parts, one of which captures the effect of the Interaction element of the model.

ANOVA					
Source	df	SS	MS	F	p-Value
Nitrogen	1	162	162	12.706	0.023
Phosphorus	1	50	50	3.922	0.119
Interaction	1	1740.5	1740.5	136.510	0.000
Error	4	51	12.75		
Total	7	2003.5			

Note that testing in mutli-way ANOVA situations proceeds with evaluation of the highest order interaction.
If it is significant, indicating a multiplicative aspect to the model, then the main effects associated with this interaction are not evaluated, but are considered necessary to properly support the significant interaction.

Two-Way Analysis of Variance





Bonferroni comparisons indicate a significant difference between the plots with both N and P at their lowest & highest values and the plots where one is high and the other low.

The profile plot was the origin of this evaluation of interaction, which is highly significant here since the profiles actually trend in different directions. However, any departure from parallel suggests the potential for significant interaction to be involved.

Two-Way Analysis of Variance

MINITAB: Stat → ANOVA → Two-Way

Rows: Nitrogen, Columns: Phosphorus

Graphs button: Chose 4 in 1 & Box Plots

Session

Two-way ANOVA: Crop Yield (bu/ac) versus Nitrogen (lbs), Phosphorus (lbs)

Source	DF	SS	MS	F	P
Nitrogen (lbs)	1	882.0	882.00	69.18	0.001
Phosphorus (lbs)	1	2450.0	2450.00	192.16	0.000
Interaction	1	0.5	0.50	0.04	0.853
Error	4	51.0	12.75		
Total	7	3383.5			

S = 3.571 R-Sq = 98.49% R-Sq(adj) = 97.36%

Individual 95% CIs For Mean Based on Pooled StDev

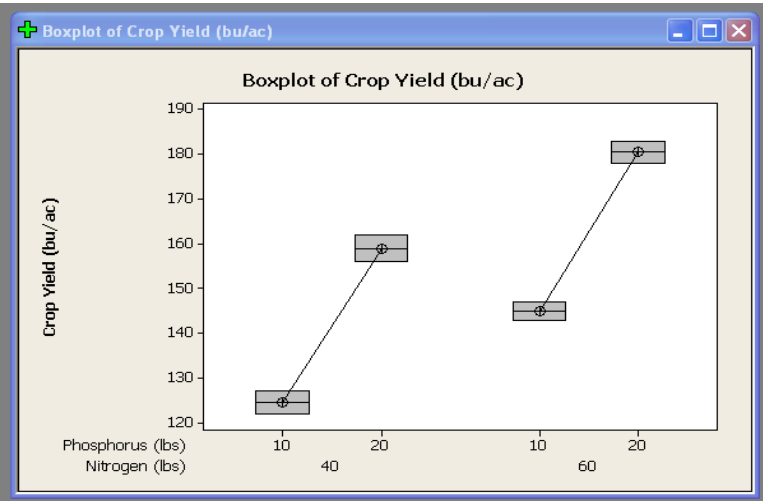
Nitrogen (lbs)	Mean
40	141.75
60	162.75

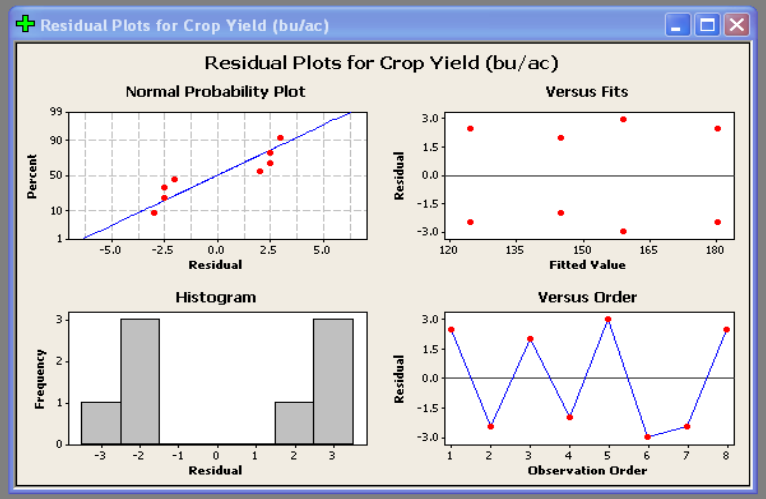
Individual 95% CIs For Mean Based on Pooled StDev

Phosphorus (lbs)	Mean
10	134.75
20	169.75

Boxplot of Crop Yield (bu/ac)

Residual Plots for Crop Yield (bu/ac)





Worksheet 1 ***

	C1	C2	C3	C4	C5
	Plot	Nitrogen (lbs)	Phosphorus (lbs)	Crop Yield (bu/ac)	
1	1	60	20	183	
2	2	60	20	178	
3	3	60	10	147	
4	4	60	10	143	
5	5	40	20	162	
6	6	40	20	156	
7	7	40	10	122	
8	8	40	10	127	
9					
10					

NOTE: Two-Way automatically includes an Interaction Effect in the Analysis

Two-Way Analysis of Variance

MINITAB: Stat → ANOVA → General Linear Model

Response: Crop Yield
Model: Nitrogen Phosphorus
Factor Plots button: Chose Main Effects (N & P)

Session

General Linear Model: Crop Yield (versus Nitrogen (lb, Phosphorus (

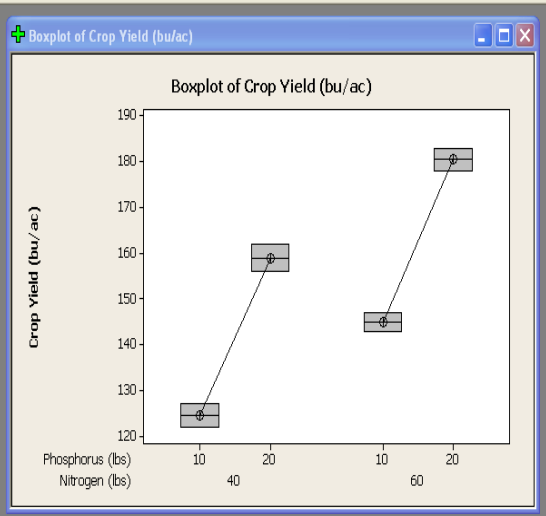
```
Factor           Type   Levels  Values
Nitrogen (lbs)   fixed       2  40, 60
Phosphorus (lbs) fixed       2  10, 20

Analysis of Variance for Crop Yield (bu/ac), using Adjusted SS for Tests

Source           DF  Seq SS  Adj SS  Adj MS      F      P
Nitrogen (lbs)    1   882.0   882.0   882.0  85.63  0.000
Phosphorus (lbs)  1  2450.0  2450.0  2450.0 237.86  0.000
Error             5    51.5    51.5    10.3
Total             7  3383.5

S = 3.20936   R-Sq = 98.48%   R-Sq(adj) = 97.87%
```

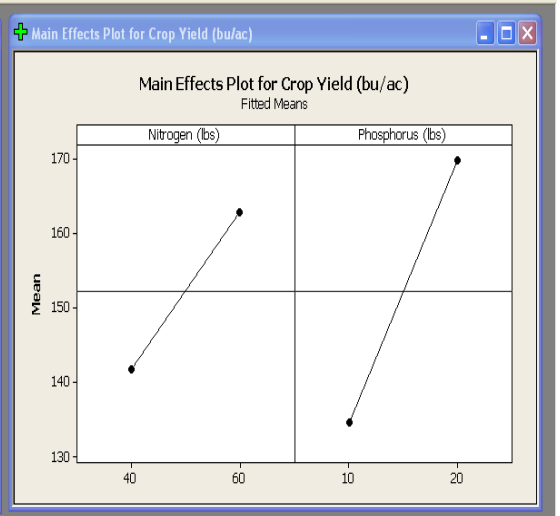
Residual Plots for Crop Yield (bu/ac)

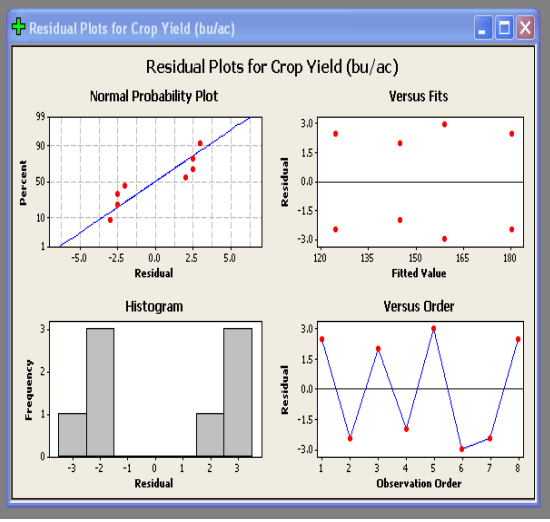
Main Effects Plot for Crop Yield (bu/ac)

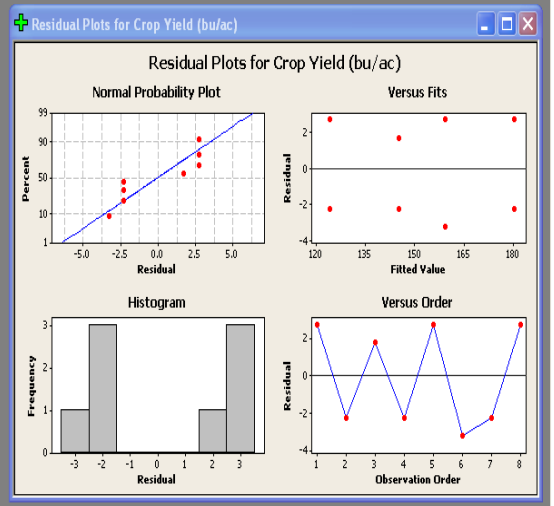
Worksheet 1 ***

	C1	C2	C3	C4	C5
	Plot	Nitrogen (lbs)	Phosphorus (lbs)	Crop Yield (bu/ac)	
1	1	60	20	183	
2	2	60	20	178	
3	3	60	10	147	
4	4	60	10	143	
5	5	40	20	162	
6	6	40	20	156	
7	7	40	10	122	
8	8	40	10	127	
9					
10					









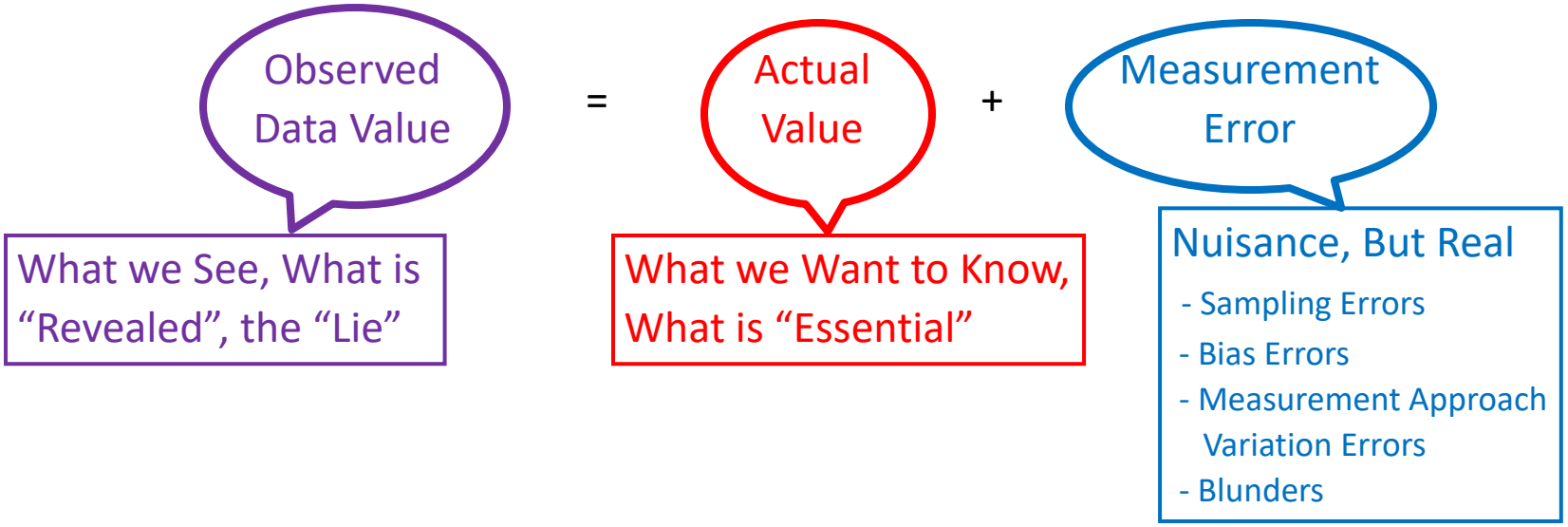
STAT 5340
Statistical Analysis I

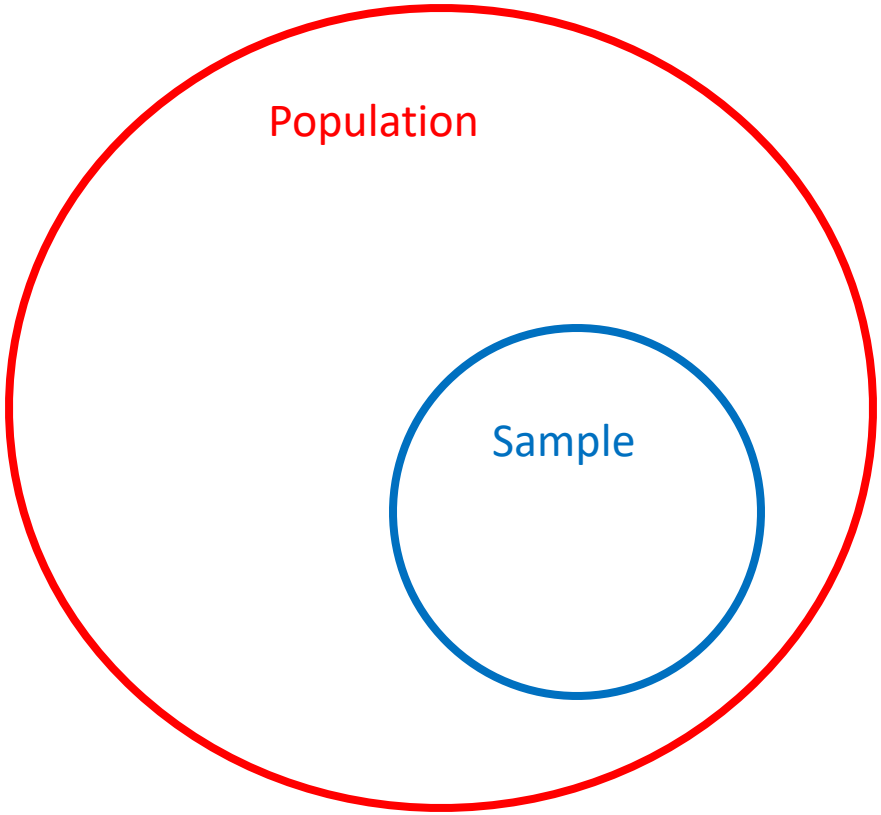
Introduction

In God we trust, all others bring data.

There are lies, damned lies, and statistics.

Statistics are like a string bikini, what they reveal can be very interesting, but what they hide is often essential.





Population is the group of individuals, objects, or events that we want to know something about

- All U.S. Citizens
- All Product from a Specific Manufacturing Line
- All Customer Complaints about a Specific Service

Sample is a subset of the Population, and is comprised of the individuals, objects, or events we can observe

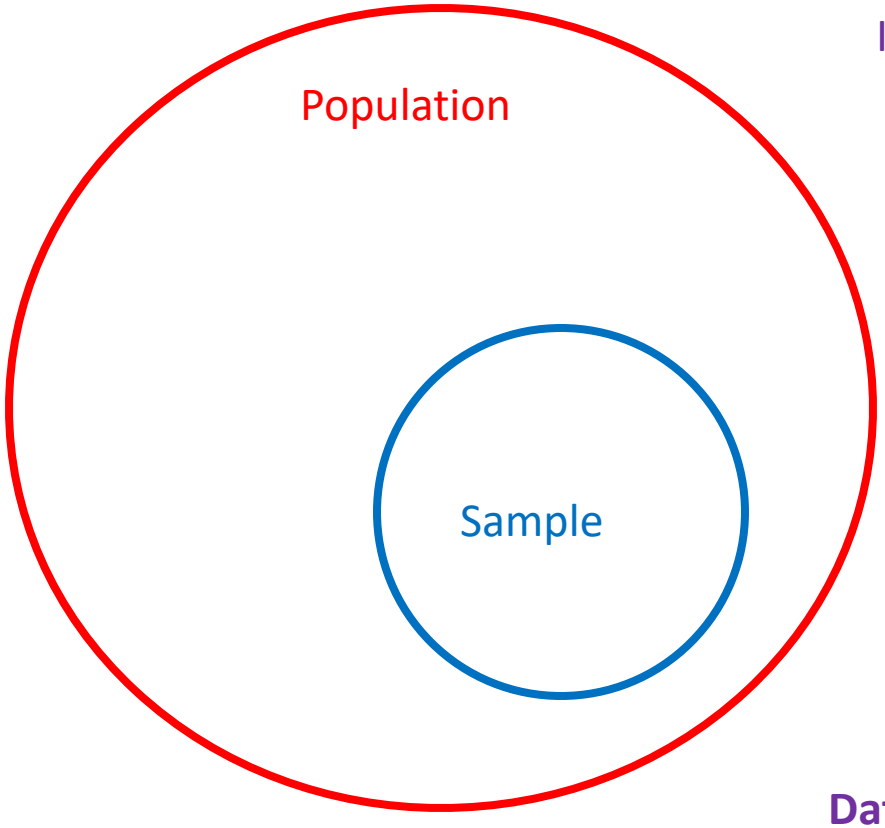
- A poll of U.S. Citizens
- Product from a Specific Lot
- Complaints from a Specific Time Period

Types of Variables	Other Descriptors	Additional Breakdown	Examples
Qualitative	Attribute, Categorical	Nominal (No Inherent Order)	Demographic - Race, Gender, Religion, etc
			Political Affiliations, Job Types, etc
		Ordinal (Inherent Order)	Non-numerical Ratings (eg, Good, OK, Bad)
			Age Groupings (eg, Under 21, 21-40, Over 40)
Quantitative	Numerical	Discrete (Finite Number of Possible Results)	Counts (eg, Particles, Defects, etc)
			Numerical Ratings (eg, 1 to 10 integers only)
		Continuous (Theoretically Infinite Results Possible)	Measured Values (eg, Distance, Weight, etc)
			Time

Variable is a Characteristic of Interest, Generally can be Measured for any Item in the Population or Sample

- Voting Preference/Intent
- A Specific Quality Characteristic (eg, Size, Strength, etc)
- Specific Nature of a Complaint

More Information is Carried in the Data as we go Down the Table, So "Best" Data is Quantitative, Continuous, Which is the Type we will Use Most



Variable is a Characteristic of Interest, Generally can be Measured for any Item in the Population or Sample

- Voting Preference/Intent
- A Specific Quality Characteristic (eg, Size, Strength, etc)
- Specific Nature of a Complaint

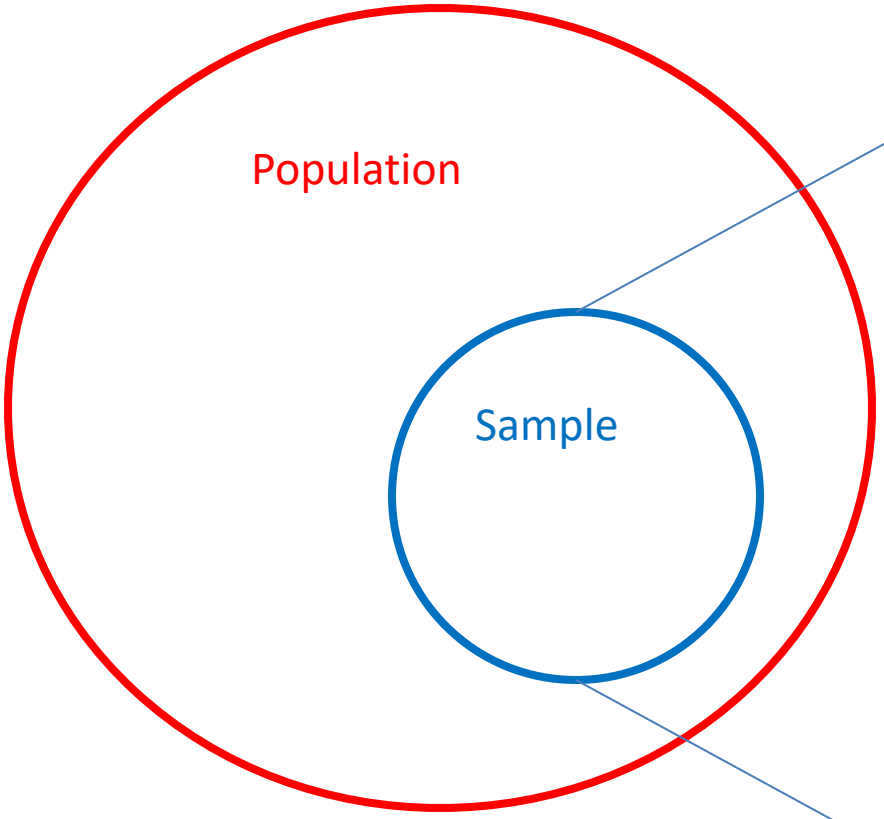
Data Value is a quantity or description associated with a specific Variable for One element of the Population or Sample

- Democrat
- 0.011 microns
- Lost luggage

Data is the group of Data Values for a specific Variable for the items in the Sample

Generic	Party	Microns	Complaint
x_1	Democrat	0.014	Poor In-Flight Service
x_2	Republican	0.015	Late Flight
x_3	Republican	0.006	Late Flight
x_4	Democrat	0.012	Late Flight

Data is the group of Data Values for a specific Variable for the items in the Sample



Generic	Party	Microns	Complaint
x_1	Democrat	0.014	Poor In-Flight Service
x_2	Republican	0.015	Late Flight
x_3	Republican	0.006	Late Flight
x_4	Democrat	0.012	Late Flight
x_5	Other	0.010	Late Flight
x_6	Republican	0.007	Late Flight
x_7	Republican	0.015	Late Flight
x_8	Democrat	0.003	Poor In-Flight Service
x_9	Democrat	0.014	Late Flight
x_{10}	Democrat	0.006	Late Flight
x_{11}	Democrat	0.014	Lost Luggage
x_{12}	Republican	0.014	Lost Luggage
x_{13}	Democrat	0.010	Lost Luggage
x_{14}	Democrat	0.016	Poor In-Flight Service
x_{15}	Democrat	0.012	Poor In-Flight Service
x_{16}	Democrat	0.015	Late Flight
x_{17}	Democrat	0.014	Late Flight
x_{18}	Democrat	0.012	Other
x_{19}	Republican	0.015	Late Flight
x_{20}	Democrat	0.011	Late Flight

Statistic is a numerical value summarizing the Sample Data

Party	Number	Percent
Democrat	13	65.0%
Republican	6	30.0%
Other	1	5.0%
All	20	100.0%

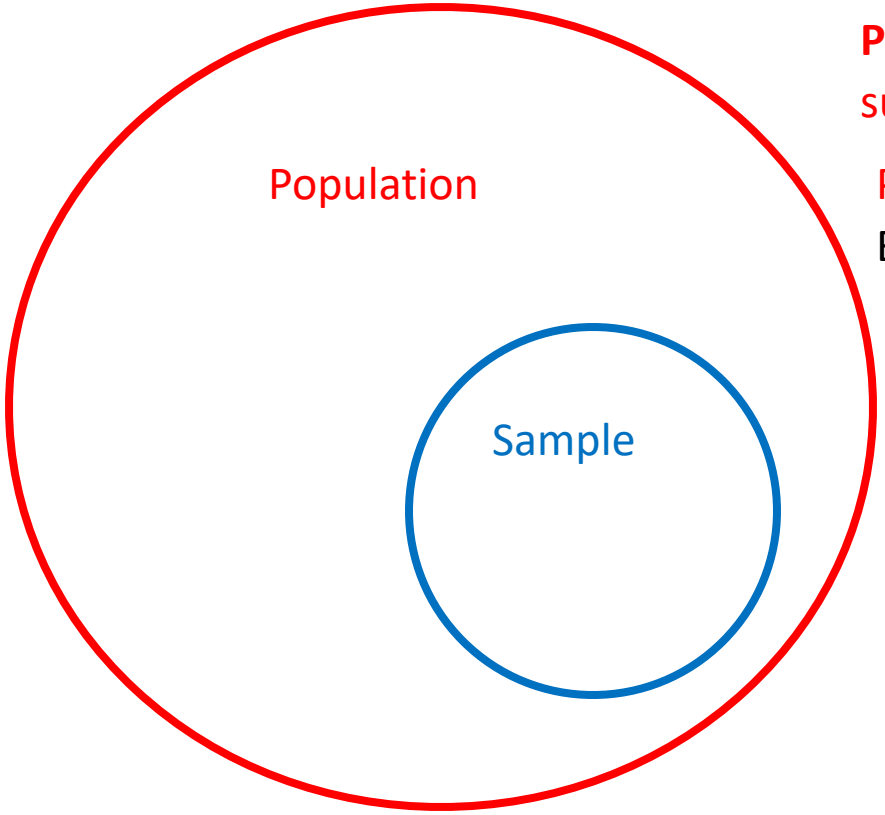
Statistic	Microns
Average	0.0118
Std Dev	0.0037

Complaint	Number	Percent
Late Flight	12	60.0%
Lost Luggage	3	15.0%
Poor In-Flight Service	4	20.0%
Other	1	5.0%
All	20	100.0%

But Are the Statistics What we Want to Know?

Statistic is a numerical value summarizing the Sample Data

Parameter is a corresponding numerical value summarizing the Population Data



Parameters are generally what we want to know,
But Statistics are all we generally have available

Sample Statistics

Party	Number	Percent
Democrat	13	65.0%
Republican	6	30.0%
Other	1	5.0%
All	20	100.0%

Statistic	Microns
Average	0.0118
Std Dev	0.0037

Population Parameters

Party	Percent
Democrat	52.0%
Republican	43.2%
Other	4.8%
All	100.0%

Parameter	Microns
Average	0.0120
Std Dev	0.0030

Inferential Statistics is using the Descriptive Statistics of the Sample to Draw Conclusions about the Population

Differences Between Sample Statistics and Population Parameters are **Sampling Errors**

Complaint	Number	Percent
Late Flight	12	60.0%
Lost Luggage	3	15.0%
Poor In-Flight Service	4	20.0%
Other	1	5.0%
All	20	100.0%

Complaint	Percent
Late Flight	60.0%
Lost Luggage	15.0%
Poor In-Flight Service	20.0%
Other	5.0%
All	100.0%

Sampling Methods
Data Collection Approaches

Data Collection Process

1) Define the Objective
2) Define the Population & Variable(s) of Interest
3) Define the Data Collection & Measurement Approaches
4) Collect the Sample Data
5) Review the Process vs the Plan

Some Guidelines

1) Keep it Manageable
2) Most Frequent Error is Attempting to Do Too Much
3) Still Strive for Representative Sample & Keep Analysis in Mind – Obtain Most Informative Data Reasonably Possible
4) Take care in Data Collection – Avoid Blunders
5) Review Process to Improve Future Efforts

Common Sampling Methods

Sample Type	Stages	Sample Name
Probability	Mutli	Proportional Stratified
		Stratified Random
		Cluster
		Multi-Stage Random
	Single	Simple Random
		Systematic
Non-Probability	Single	Judgment
		Volunteer
		Convenience

Some Observations

- Sampling is Generally Driven by the Budget
- Sampling Approach should consider the Objective
- Compromises in Sampling can Increase Likelihood of Non-Representative Samples
- Going Down Table Above, Generally
  - Less Expensive
  - Less Information Required Up Front
  - More Chance for Non-Representative Samples
- **Bias Errors** are the Likely Result of Non-Representative Samples

Election Polling

In 1936, *Literary Digest* ran their by then traditional presidential election poll of more than10 million Americans.

Similar polls they had run correctly predicted the outcome each time they had been conducted – 1916, 1920, 1924, 1928, and 1932.

Their results are displayed at right and effectively predict a landslide for the Republican candidate Alf Landon.

- 57% of the voters
- 370 Electoral College votes

State	Electoral Vote	Landon 1936 Total Vote For State	Roosevelt 1936 Total Vote For State	State	Electoral Vote	Landon 1936 Total Vote For State	Roosevelt 1936 Total Vote For State
Ala.	11	3,060	10,082	Nebr.	7	18,280	11,770
Ariz.	3	2,337	1,975	Nev.	3	1,003	955
Ark.	9	2,724	7,608	N.H.	16	9,207	2,737
Calif.	22	89,516	7,608	N.J.	16	58,677	27,631
Colo.	6	15,949	10,025	N.M.	3	1,625	1,662
Conn.	8	28,809	13,413	N.Y.	47	162,260	139,277
Del.	3	2,918	2,048	N.C.	13	6,113	16,324
Fla.	7	6,087	8,620	N. Dak.	4	4,250	3,666
Ga.	12	3,948	12,915	Ohio	26	77,896	50,778
Idaho	4	3,653	2,611	Okla.	11	14,442	15,075
Ill.	29	123,297	79,035	Ore.	5	11,747	10,951
Ind.	14	42,805	26,663	Pa.	36	119,086	81,114
Iowa	11	31,871	18,614	R.I.	4	10,401	3,489
Kans.	9	35,408	20,254	S.C.	8	1,247	7,105
Ky.	11	13,365	16,592	S.Dak.	4	8,483	4,507
La.	10	3,686	7,902	Tenn.	11	9,883	19,829
Maine	5	3,686	7,902	Texas	23	15,341	37,501
Md.	8	17,463	18,341	Utah	4	4,067	5,318
Mass.	17	87,449	25,965	Vt.	3	7,241	2,458
Mich.	19	51,478	25,686	Va.	11	10,223	16,783
Minn.	11	30,762	20,733	Wash.	8	21,370	15,300
Mis.	9	848	6,080	W.Va.	8	13,660	10,235
Mo.	15	50,022	8,267	Wis.	12	33,796	20,781
Mont.	4	4,490	3,562	Wyo.	3	2,526	1,533
State Unknown	7	1586	545				
Total	531	1,293,669	972,897				

Source: *Literary Digest*, 31 October 1936.

Election Polling

So what was the outcome?

Roosevelt won, but was it at least close?

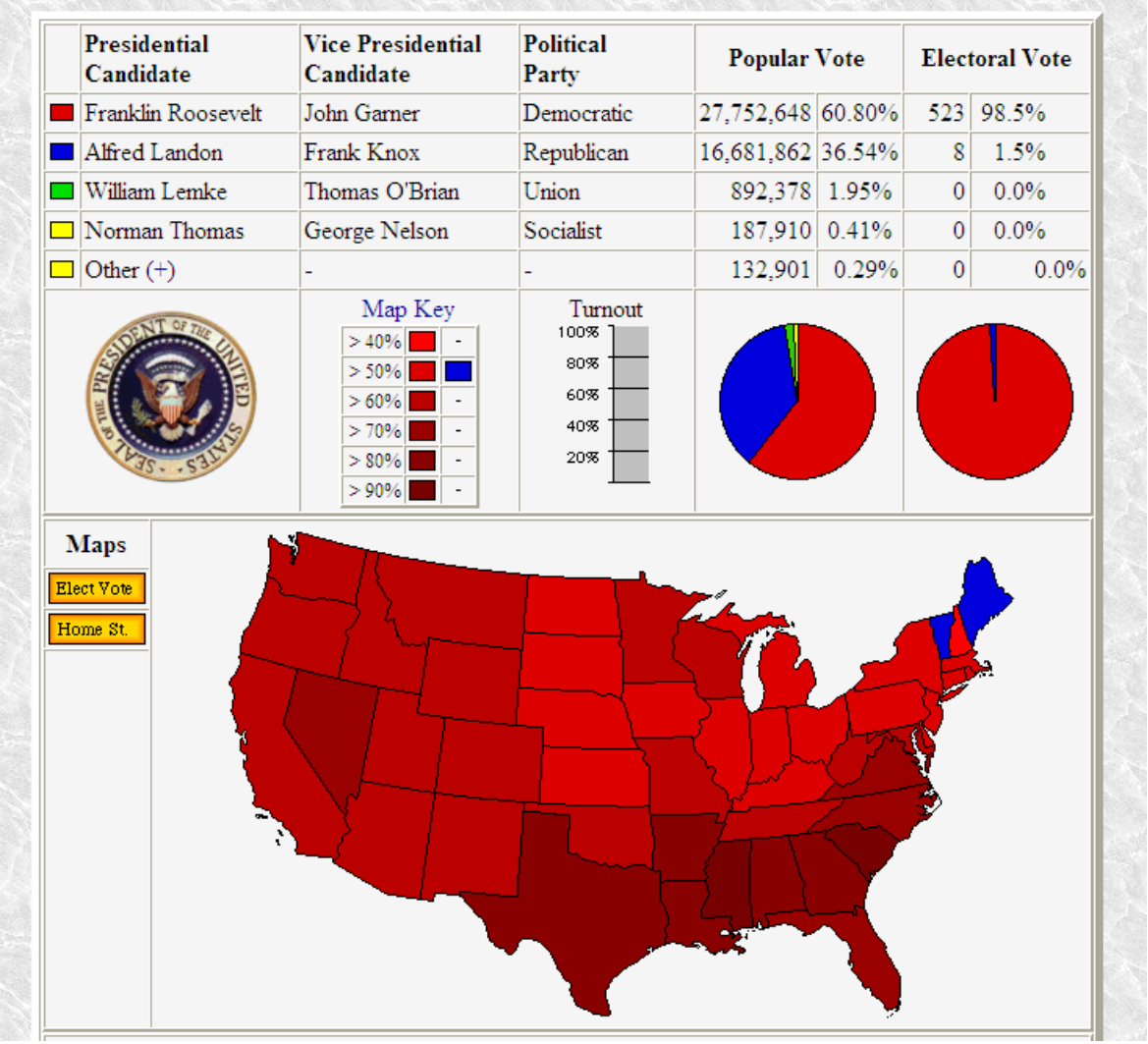
- Well, no

Roosevelt won >60% of the vote and all but two states (ie, 523 of the 531 Electoral College votes)

So what happened?

Sample Bias Error

- Self-selected sample
- Bias in mailings



	Presidential Candidate	Vice Presidential Candidate	Political Party	Popular Vote		Electoral Vote	
	Franklin Roosevelt	John Garner	Democratic	27,752,648	60.80%	523	98.5%
	Alfred Landon	Frank Knox	Republican	16,681,862	36.54%	8	1.5%
	William Lemke	Thomas O'Brian	Union	892,378	1.95%	0	0.0%
	Norman Thomas	George Nelson	Socialist	187,910	0.41%	0	0.0%
	Other (+)	-	-	132,901	0.29%	0	0.0%

http://uselectionatlas.org/RESULTS/national.php?year=1936

Infant Mortality Data

So where does the US rank among the most populous countries in the world in terms of Infant Mortality Rate (deaths/1000 live births)?

- Lowest Rate (Rank 1)
- Lowest Five (Ranks 2-5)
- Lowest Ten (Ranks 6-10)
- Outside Lowest Ten (Ranks > 10)

- Over/Under 20?

List by the United Nations Population Division

Rank	Country or territory	Infant mortality rate (deaths/1,000 live births)
1	Iceland	2.9
2	Singapore	3.0
3	Japan	3.2
4	Sweden	3.2
5	Norway	3.3
6	Hong Kong	3.7
7	Finland	3.7
8	Czech Republic	3.8
9	Switzerland	4.1
10	South Korea	4.1
11	Belgium	4.2
12	France	4.2
13	Spain	4.2
14	Germany	4.3
15	Denmark	4.4
16	Austria	4.4
17	Australia	4.4
18	Luxembourg	4.5
19	Netherlands	4.7
20	Israel	4.7
21	Slovenia	4.8
22	United Kingdom	4.8
23	Canada	4.8
24	Ireland	4.9
25	Italy	5.0
26	Portugal	5.0
27	New Zealand	5.0
28	Cuba	5.1
29	Channel Islands (Jersey and Guernsey)	5.2
30	Brunei	5.5
31	Cyprus	5.9
32	New Caledonia	6.1
33	United States	6.3

http://en.wikipedia.org/wiki/List_of_countries_by_infant_mortality_rate

Here's a general overview for you. When you compare infant mortality statistics you need to look for the definitions. What, for instance, constitutes a live birth? In the United States any infant exhibiting any sign of life is considered to be alive. It doesn't matter how small, how premature or how much it weights. In countries like France, the Netherlands and Ireland they don't count the birth as a live birth unless the infant weighs more than 500 grams or the mother was at least 22 [weeks] along in the pregnancy. Other countries won't count the birth as being a live birth unless the infant survives for a specified period of time.

http://boortz.com/nealz_nuze/2009/11/those-phony-infant-mortality-s.html

So measurement approach is relevant, and differences in approach (ie, ***Measurement Errors***) will work to hide the actual information that is really of interest.

Measurement Errors

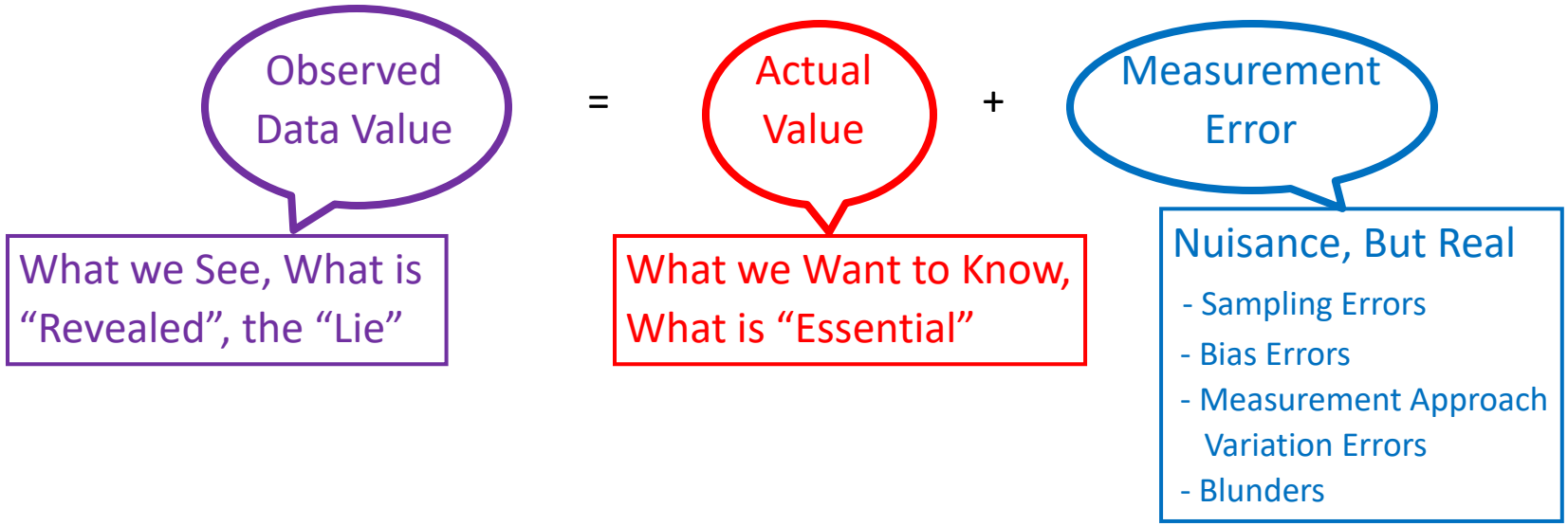
Measurement Accuracy = No Bias Errors = Well-Calibrated Device

Measurement Precision = Good Repeatability & Reproducibility

Virtually Always Present in Quantitative, Continuous Data – Gauge Studies

Can be Present in Survey Data – Same Question Phrased Slightly Differently

Signal + Noise Model



The Basic Statistical Model reflected above is:

$Y_i = \mu + \varepsilon_i$

Observed Data Value

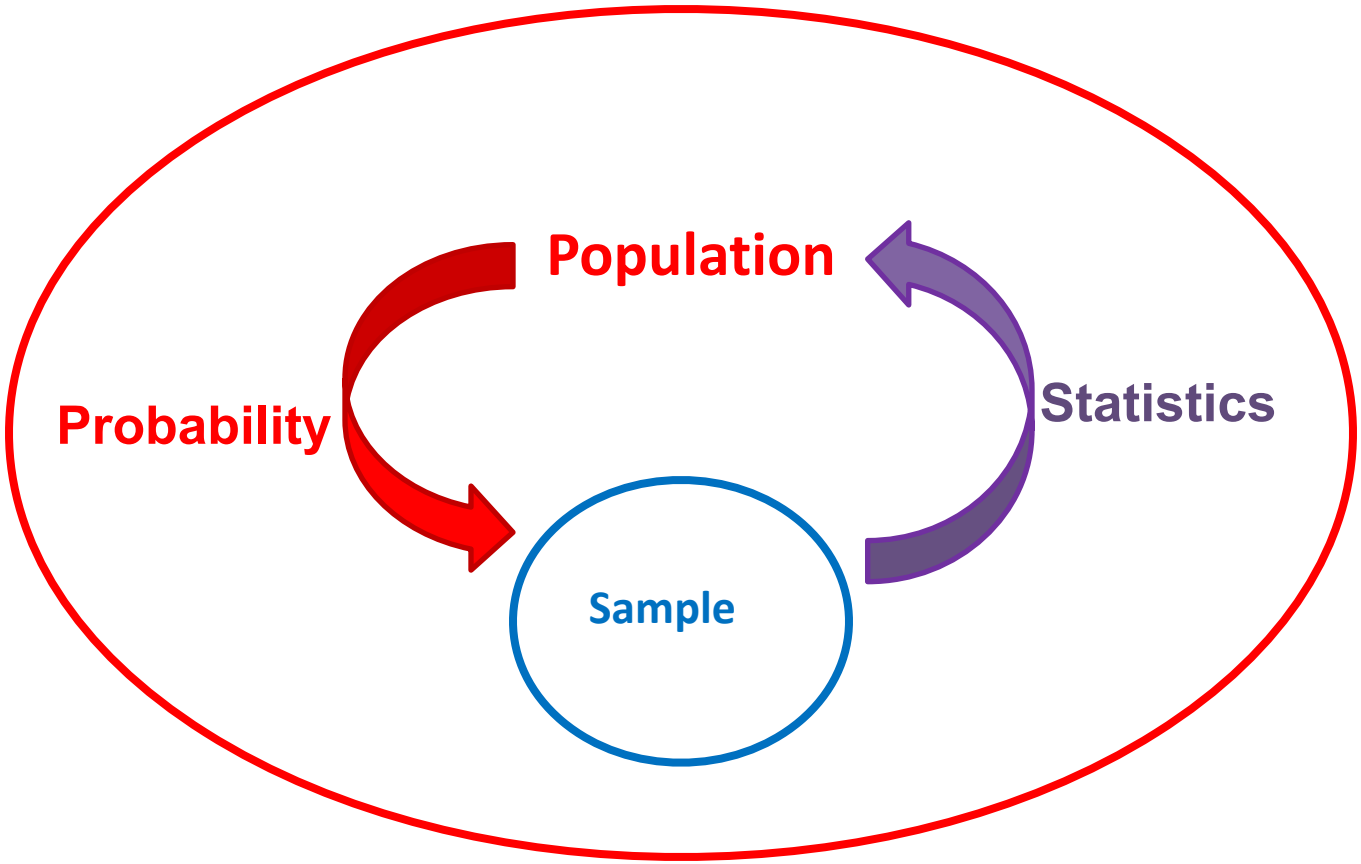
Signal = Parameter We Really Want to Know

Noise = Errors due to Sampling, Bias, Measurement, Mistakes, etc.

How much of what is **Observed** is **Signal** and how much is **Noise**?

Probability and Statistics

- Probability attempts to evaluate the chance that specific events will occur given knowledge of a Population
- Statistics attempts to evaluate what a Population looks like given the occurrence of an event (ie, a Sample)



STAT 5340
Statistical Analysis I

Probability

Terminology

A SAMPLE SPACE is a well-defined collection of items, eg, people places, things, objects, numbers, etc. "Well-defined" implies that membership in a given set or not is clear, obvious.

Examples: All U.S. citizens of voting age
All product produced from a specific manufacturing line
All complaints made about airline travel

An EVENT is a collection of items that all belong to a given SAMPLE SPACE.

Examples: U.S. citizens of voting age that provided responses to a specific opinion poll
Lots of product sampled for specific evaluation from a specific manufacturing line
Complaints made about airline travel over a specific time frame

A SAMPLE POINT is a specific Item in a SAMPLE SPACE.

Examples: A specific U.S. citizen of voting age
A specific lot of production from a specific manufacturing line
A specific complaint made about airline travel

Probability

Classical Definition

If a SAMPLE SPACE is comprised of N SAMPLE POINTS $\{s_i\}$, each of which is **equally likely** to occur, be observed, or comprise an EVENT, then the

Probability of any individual SAMPLE POINT = $P[s_i] = 1/N$, for i = 1 to N

If an EVENT, A, is comprised of M **equally likely** SAMPLE POINTS from a SAMPLE SPACE of N (>= M) **equally likely** SAMPLE POINTS, then

Probability of EVENT A = P[A] = M/N.

This is the Classical concept of Probability, which largely arose from the study of games of chance (ie, those involving coins, dice, and cards)

Probability

Relative Frequency Definition

Another approach to determining probabilities is the Relative Frequency concept.

This approach suggests that if we repeat a data acquisition activity (eg, an experiment, a survey, etc) a large number of times and each time observe when an event of interest (eg, A) occurs, say N_A, then with a sufficiently large number of repeats, say N

$$P[A] \approx N_A/N$$

Consider the activity of rolling 12 fair dice with the event of interest (ie, A) being observing at least two "6" results among the 12 dice.

Now imagine repeating this activity N=100 times and counting how many of the 100 rolls satisfy event A (ie, N_A).

At each roll, r = 1, 2, …, 100, we could calculate the relative frequency of event A as $N_A(r)/r$, and plot the ordered pairs $(r, N_A(r)/r)$.

The resultant plot should converge on P[A].

For this activity, the exact P[A] can be obtained, and with 100 repeats, we got as close as possible with our Relative Frequency estimate.

Probability

Definition

The Probability of an Event equals the number of Items in the Event (either calculated or observed) divided by the number of Items in the Sample Space (ie, P[A] = n(A) / n(S)).

Example: Rolling a pair of dice

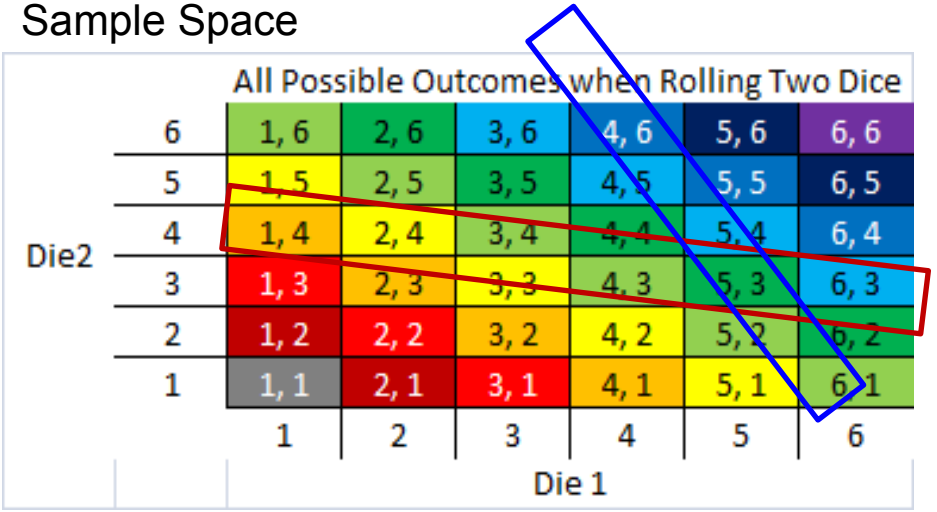
Event: Roll Doubles (ie, same number on both die)

n(Doubles) = 6, n(S) = 36
P[Doubles] = 6/36 = 1/6

Event: Sum of Die = 9

n(Sum=9) = 4, n(S) = 36
P[Sum=9] = 4/36 = 1/9

Sample Space



So … Probability is 2 Counting Results and a Division Step

1) Count the number of items in the Sample Space – n(S)
2) Count the number of items in the Event – n(A)
3) Divide the result in 2) by that in 1) – P[A] = n(A)/n(S)

Probability

Example – Birthdays

There are approximately 30 people in a general undergrad class, what do think the probability is that there are at least 2 people with the same birth date (not necessarily year) in the class?

1) Count the number in the Sample Space (assume no 2/29 birthdays)

How many possible birth dates for 1st person? 365
How many for 2nd person? 365
…
How many for 30th person? 365

So, the number of possible lists of birth dates for n = 30 people is $n(S) = 365^{30}$

2) Count the number in the Event (at least 1 matched set of birth dates)

(Easier to count number with NO matches and subtract from above)

How many possible birth dates for 1st person? 365
How many for 2nd person? 364
…
How many for 30th person? 336

So, the number of possible lists of birth dates for n = 30 people with at least 1 match is $n(A) = 365^{30} – P(365, 30)$

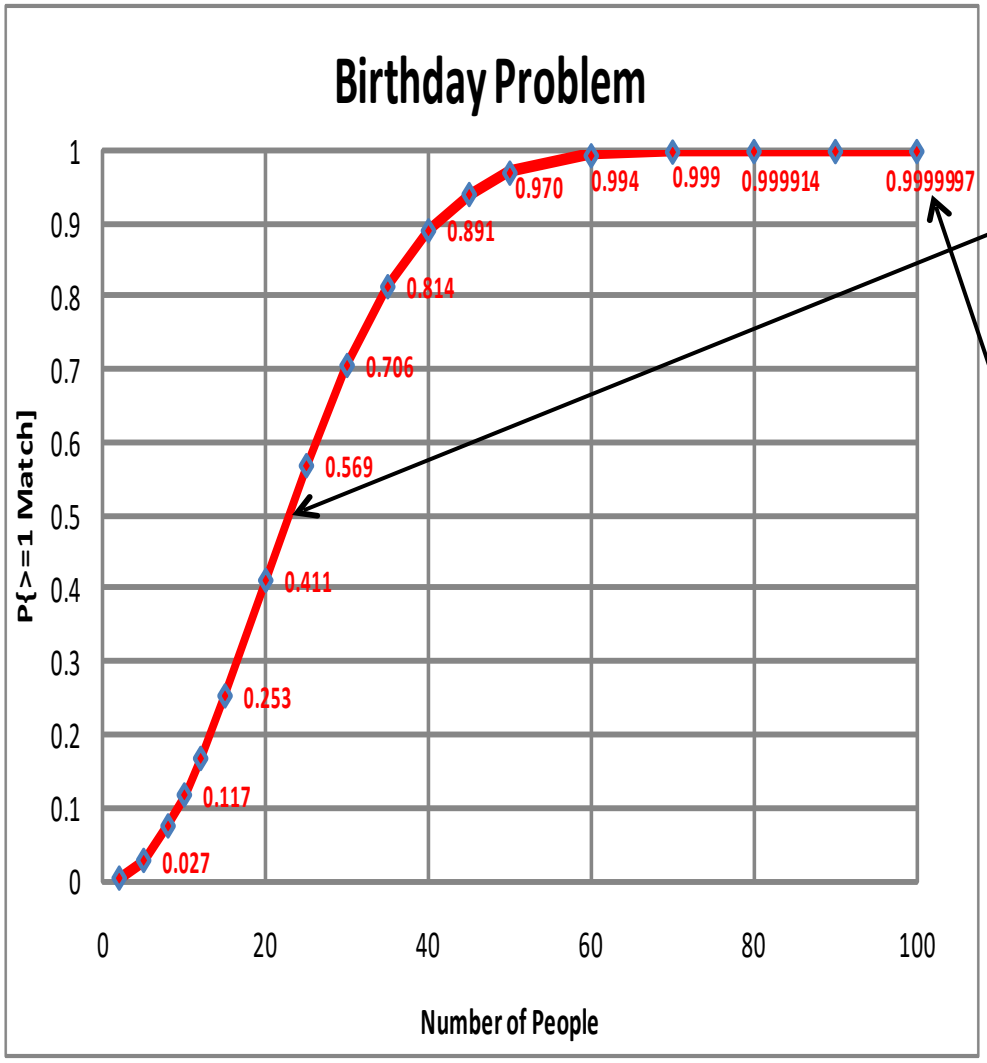
3) Divide result in 2) by that for 1)

$$P[>=1 \text{ match}] = [365^{30} – P(365, 30)] / 365^{30}$$
$$= 1 – 0.294$$
$$= \sim 0.706$$

So, about a 7 in 10 chance that there will be at least one match in the room

Probability

Example – Birthdays



Probability of at least one match in a group of 23 people is about 50%.

In a group of 100 people, it is a virtual certainty that at least two people will share the same birth date.

If my wife and I are both in the room, then it is a certainty that at least two people will share the same birth date, because we do … which is great for me, as I never forget her birthday. ☺

Sets

A SET is a well-defined collection of items, eg, people places, things, objects, numbers, etc. "Well-defined" implies that membership in a given set or not is clear, obvious.

Examples: All U.S. citizens of voting age
All product produced from a specific manufacturing line
All complaints made about airline travel

A SUBSET is a collection of items that all belong to a given SET.

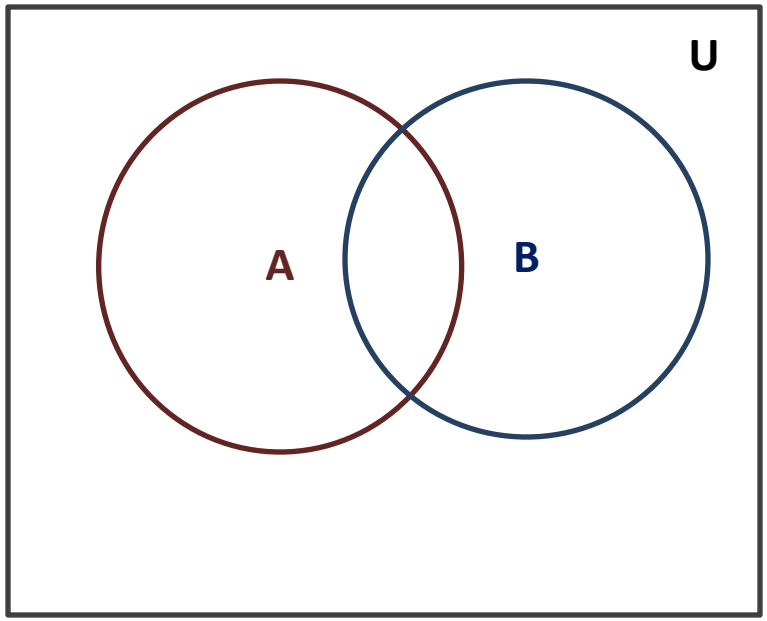
Examples: U.S. citizens of voting age that provided responses to a specific opinion poll
Lots of product sampled for specific evaluation from a specific manufacturing line
Complaints made about airline travel over a specific time frame

Every SET is a SUBSET of itself, and the EMPTY SET (ie, the set with no items) is a SUBSET of every SET

Two SETs are EQUAL if and only if they contain exactly the same items

Sets

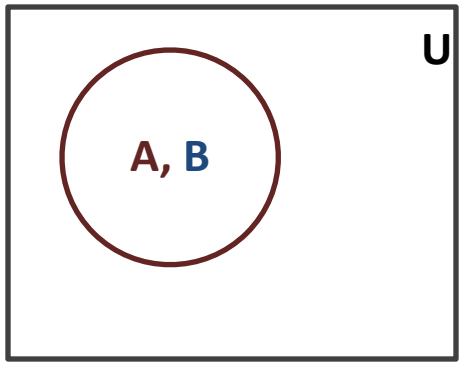
Venn Diagrams



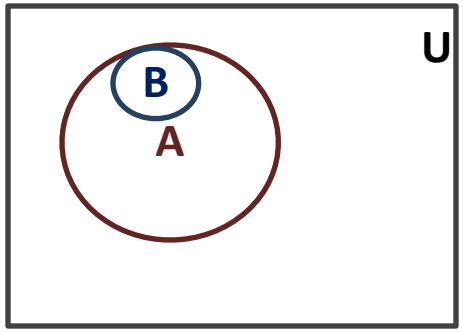
The Rectangle represents the Universal Set = Set of All (“Relevant”) Items

Set A and Set B share some common Items, but are not Subsets of each other

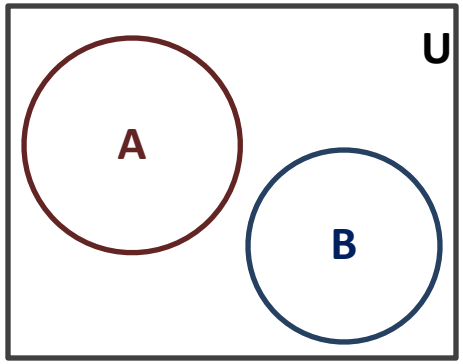
A = B



B is Subset of A



A and B Mutually Exclusive Sets (No Common Items)



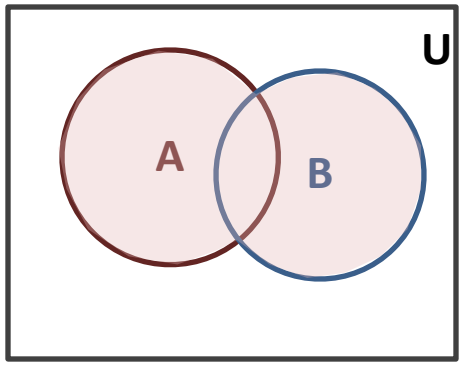
Sets

Set Operations

UNION

A U B = A or B

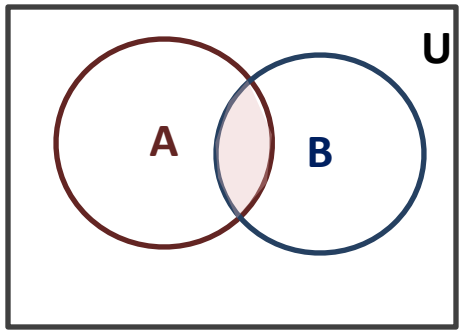
The Set of All Items in Set A, in Set B or in Both Sets A and B



INTERSECTION

A ∩ B = A and B

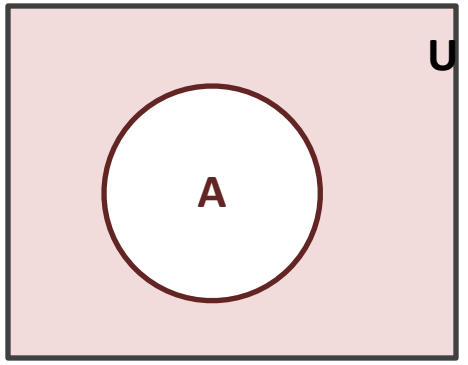
The Set of All Items in Both Set A and Set B



COMPLEMENT

A' = not A

The Set of All Items in the Universal Set, U, But Not in A



Probability

Rules of Probability

Rule 1: A probability is always a number between 0 and 1, inclusive.

$0 <= P[\text{Any Event}] <= 1$

Note: Usually, probabilities are expressed in decimal form, but can also be expressed as percentages, eg 1 = 100%

Note: P[A] = 0 indicates outcome A can not occur

Note: P[A] = 1 indicates outcome A is the only possible outcome, and always occurs

Rule 2: The sum of probabilities for all possible outcomes is equal to one.

$\text{Sum}(\text{All } P[A_i]) = 1$

NOTE: Here A_i is a partitioning of the entire Sample Space so that all A_i are mutually exclusive of each other

Rule 3 (Law of Large Numbers): If an experiment is repeated many times, the empirically observed probability of a given outcome will tend to approach the theoretical probability of that outcome for a single experimental trial.

Probability

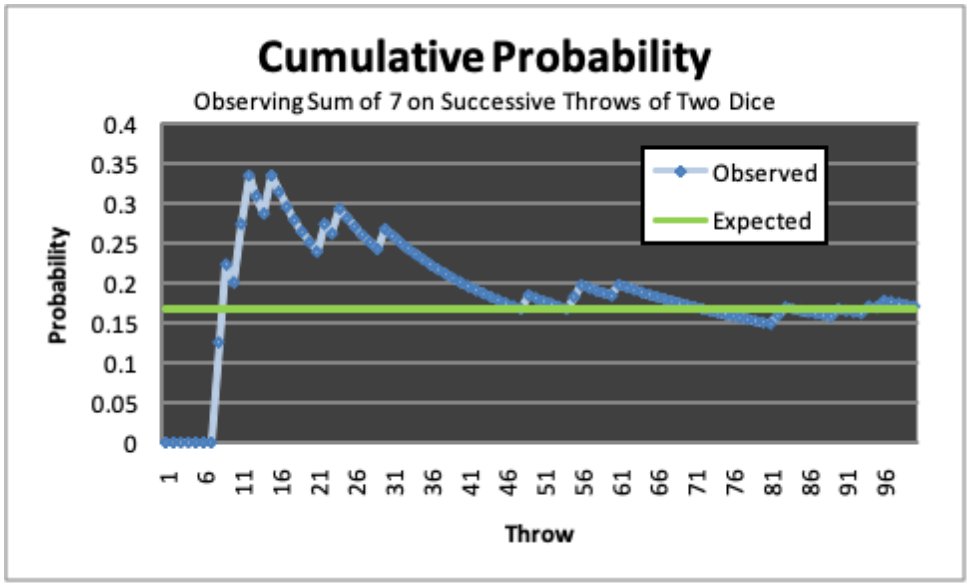
Rules - Example

All Possible Outcomes when Rolling Two Dice

Die2 \ Die 1	1	2	3	4	5	6
6	1, 6	2, 6	3, 6	4, 6	5, 6	6, 6
5	1, 5	2, 5	3, 5	4, 5	5, 5	6, 5
4	1, 4	2, 4	3, 4	4, 4	5, 4	6, 4
3	1, 3	2, 3	3, 3	4, 3	5, 3	6, 3
2	1, 2	2, 2	3, 2	4, 2	5, 2	6, 2
1	1, 1	2, 1	3, 1	4, 1	5, 1	6, 1

Result of Interest: **Sum of Both Die**

Outcome	n(A)	P[A]
2	1	1/36 = 0.0278
3	2	2/36 = 0.0556
4	3	3/36 = 0.0833
5	4	4/36 = 0.1111
6	5	5/36 = 0.1389
7	6	6/36 = 0.1667
8	5	5/36 = 0.1389
9	4	4/36 = 0.1111
10	3	3/36 = 0.0833
11	2	2/36 = 0.0556
12	1	1/36 = 0.0278
All	36	36/36 = 1



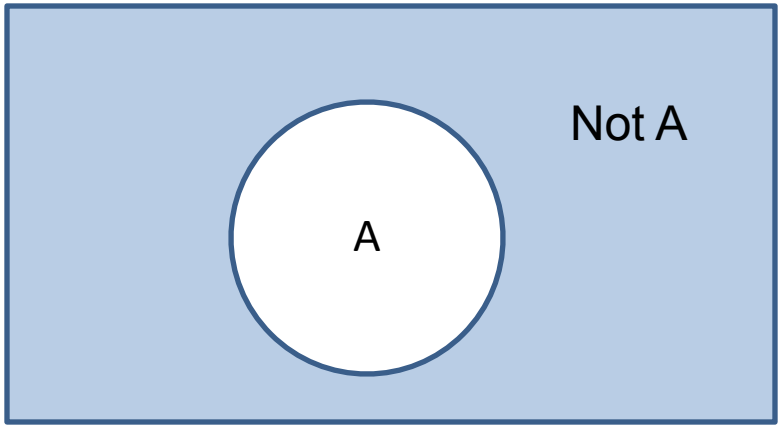
Rule 1: All P[A] between 0 and 1
Rule 2: Sum of P[A] for All Outcomes = 1
Rule 3: P'[7] Observed approaches
P[7] Expected with Repeated Throws

Probability

Complementary and Mutually Exclusive Events

The Complement of Event A is the Event "Not A"
P[Not A] = P[A'] = 1 – P[A]

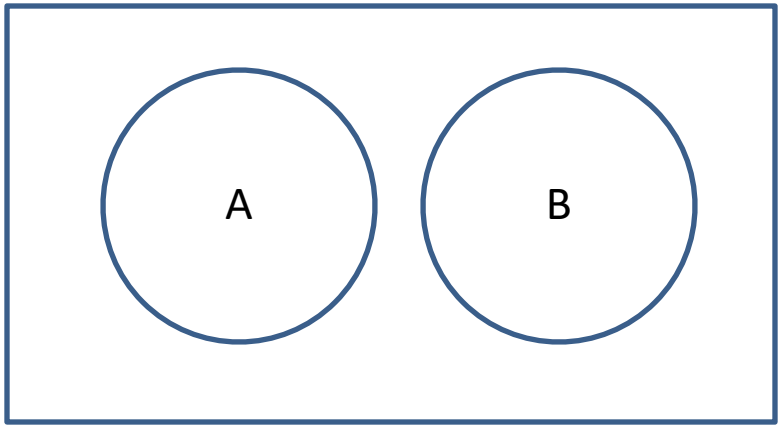
The Complement of an Event is All Area Outside The Circle for that Event, But Still within the Venn Diagram



Mutually Exclusive Events, A and B, are Events that Can NOT Occur Simultaneously
P[A and B] = P[A ∩ B] = 0

NOTE: A and Not A are mutually exclusive
If A & B are mutually exclusive, then
P[A or B] = P[A U B] = P[A] + P[B]

Mutually exclusive Events Have no Overlapping Area within the Venn Diagram



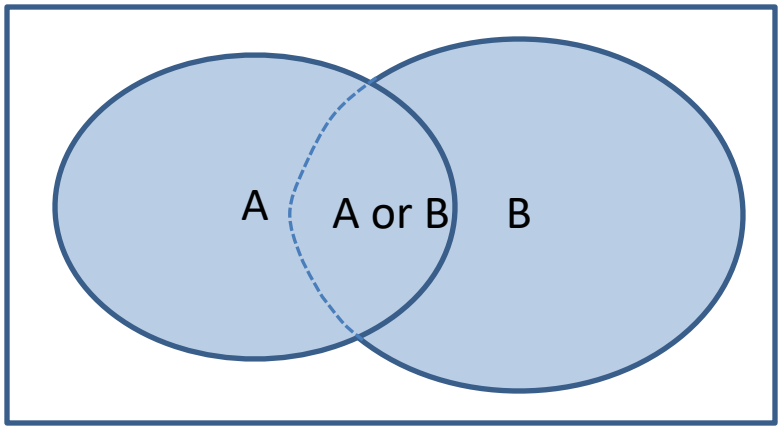
Probability

"Additive Rule"

P[A or B] = P[A U B] = Probability that either A alone, B alone, or both A and B occur
= P[A] + P[B] – P[A ∩ B]
= P[A] + P[B] – P[A and B]

NOTE: P[A] includes all of A, and P[B] includes all of B, so P[A & B] added in twice in P[A] + P[B], so needs to be subtracted out once; hence, the formula above

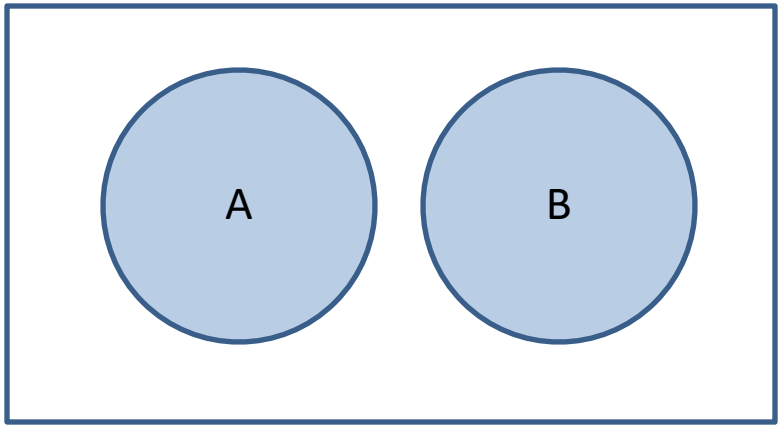
Event A or B includes all the Area included by both Circles in Venn Diagram



<u>If A & B are Mutually Exclusive Events</u>, then
P[A or B] = P[A U B] = Probability that A alone, B alone, or both A & B occur
= P[A] + P[B]

NOTE: This is the same as the result above since for Mutually Exclusive events P[A & B] = P[A ∩ B] = 0.

Event A or B is area including both Circles in Venn Diagram



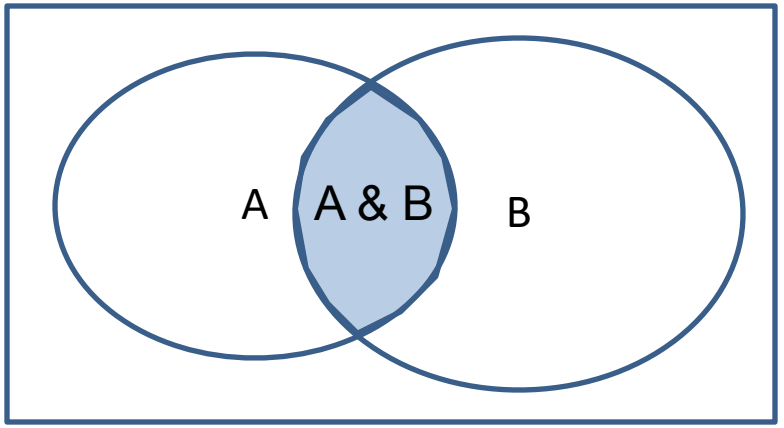
Probability

"Conditional Probability"

P[A | B] = Probability that event A occurs given that event B has occurred
= P[A and B] / P[B], or
= P[A ∩ B] / P[B]

NOTE: P[B] ≠ 0

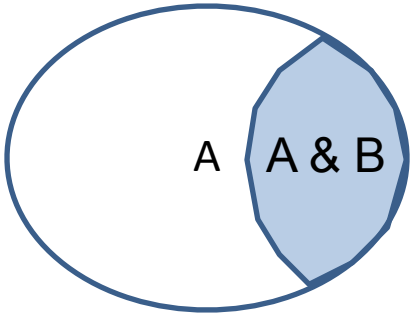
Event A | B is where Circles Overlap Restricted to the Circle **B**



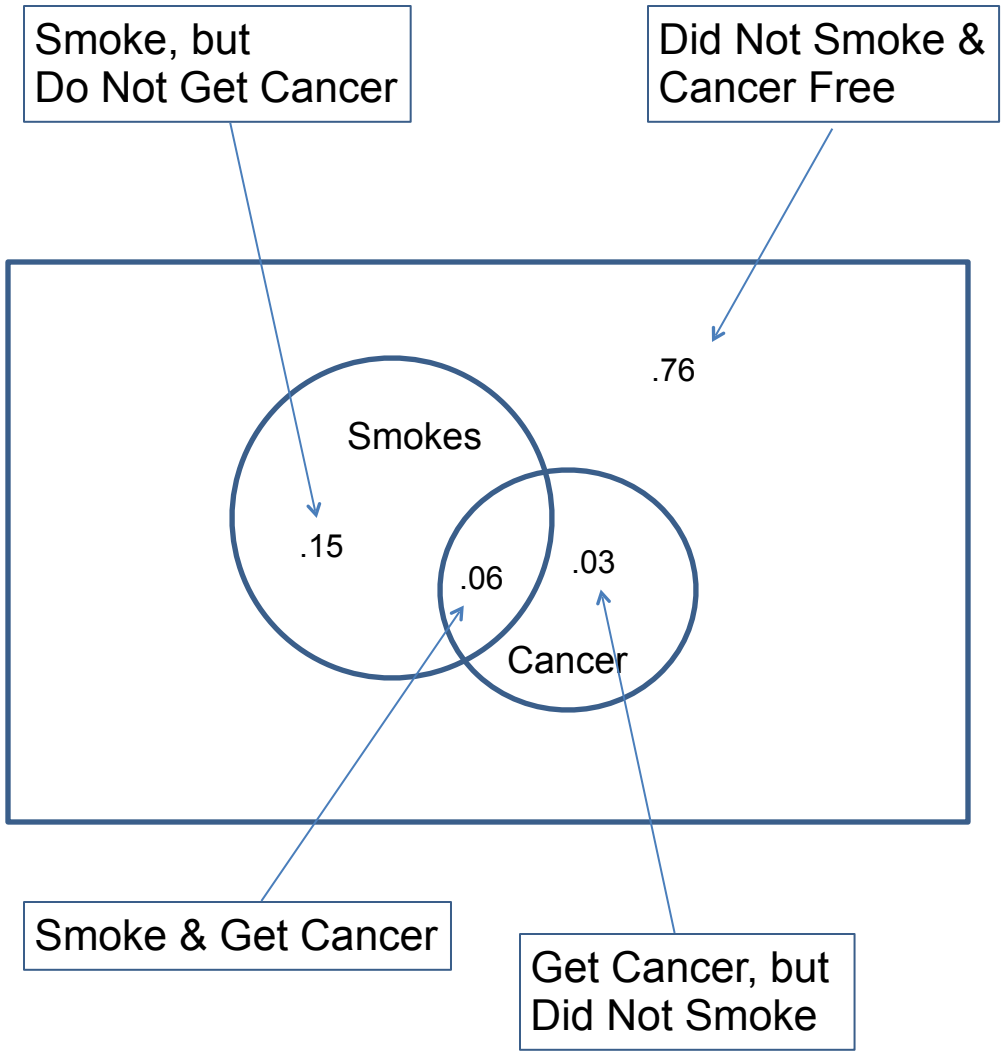
P[B | A] = Probability that event B occurs given that event A has occurred
= P[A and B] / P[A], or
= P[A ∩ B] / P[A]

NOTE: P[A] ≠ 0

Event B | A is where Circles Overlap Restricted to the Circle **A**



Conditional Probability



Consider the following information:
9% of Individuals get Cancer
21% of Individuals Smoke
76% of Individuals Neither Smoke nor get Cancer

What is the probability someone who smokes will get cancer?

Now answer is straightforward:

P[Cancer|Smokes] = .06/.21 ≈ 0.286

What is the probability someone who has cancer smokes?

P[Smokes|Cancer] = .06/.09 ≈ 0.667

	Cancer		
Smokes	Yes	No	Total
Yes	.06	.15	.21
No	.03	.76	.79
Total	.09	.91	1

Conditional Probability

Let $C_1, C_2, \ldots, C_k$ be mutually exclusive, but exhaustive events (ie, a partition) of sample space *C*, with $P[C_i] > 0$, for all i = 1, 2, …, k. Note these events do not need to be equally likely.

Let C be another event in *C*, then

$$C = C \cap (C_1 \cup C_2 \cup \ldots \cup C_k) = (C \cap C_1) \cup (C \cap C_2) \cup \ldots \cup (C \cap C_k)$$

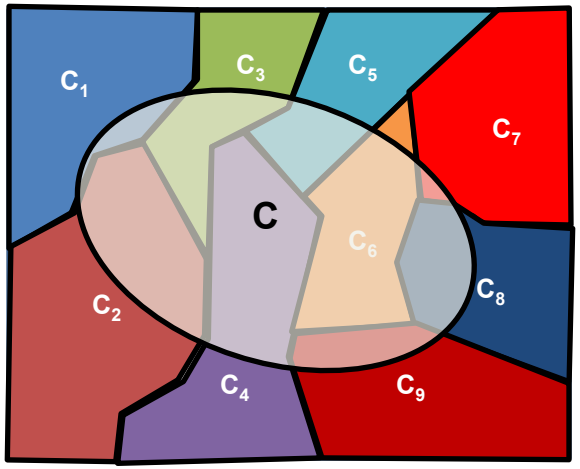
and, since $C \cap C_i$, i = 1, 2, …, k are all mutually exclusive,

$$P[C] = P[(C \cap C_1)] + P[(C \cap C_2)] + \ldots + P[(C \cap C_k)].$$

However, $P[(C \cap C_i)] = P[C_i]*P[C|C_i]$, i = 1, 2, …, k; so

$$P[C] = P[C_1]*P[C|C_1] + P[C_2]*P[C|C_2] + \ldots + P[C_k]*P[C|C_k]$$
$$= \sum_{i = 1 \text{ to } k}\{P[C_i]*P[C|C_i]\},$$

which is often recognized as the ***law of total probability***.



When P[C] > 0, from the definition of conditional probability, and using the law of total probability:

$$P[C_j \mid C] = P[(C_j \cap C)]/P[C] = P[C_j]*P[C \mid C_j]/(\sum_{i = 1 \text{ to } k}\{P[C_i]*P[C|C_i]\}),$$

which is the well-known ***Bayes' Theorem***.

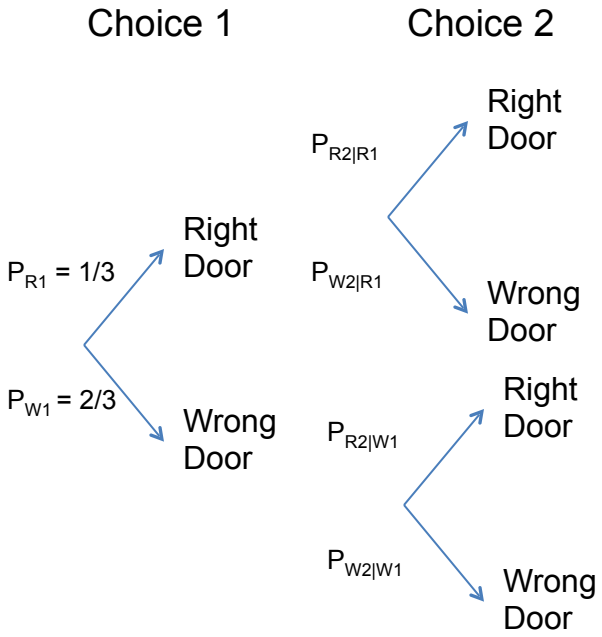
Conditional Probability

Example – Let's Make a Deal

Premise: 3 Hidden Doors, Behind One is a Prize (Chosen Randomly)
Choice 1: Choose a Door at Random
Information: Of the Two Doors Not Selected, One w/o Prize is Identified
Choice 2: Stay with First Choice, or Change to Remaining Hidden Door

What Choice 2 Strategy Maximizes Chances of Getting Prize?

- Always Stay with Initial Choice
- Always Change to Other Door
- Flip a Coin to Choose (Heads = Stay, Tails = Change)



P[Win Prize] = P[Right Door]
= P[R1 and R2] + P[W1 and R2]
= P[R1] * P[R2|R1] + P[W1] * P[R2|W1]

Strategy	Probabilities						
	R2| R1	W2|R1	R2|W1	W2|W1	R1 & R2	W1 & R2	WIN
STAY	**1**	**0**	**0**	**1**	**1/3**	**0**	**1/3**
CHANGE	**0**	**1**	**1**	**0**	**0**	**2/3**	**2/3**
FLIP COIN	**1/2**	**1/2**	**1/2**	**1/2**	**1/6**	**1/3**	**1/2**

Strategy of Always Changing to Other Door Maximizes Chances of Winning the Prize

Conditional Probability

Example

Situation:

A cab sideswipes a car late on a winter night.
A witness testifies that he saw a Blue cab commit the offense.
There are only two cab companies in town: Blue and Green.
Green has 85% of the cabs on the road, Blue the remaining 15%.
Independent studies indicate that the witness will be correct 80% of the time.

Question:

What is the probability the offending cab was indeed Blue given the Testimony of the witness that it was Blue?

What do we know?

P[Cab was Blue] = .15 and P[Cab was Green] = .85

What do we want to know?

P[Cab was Blue | Testimony is Blue] = 0.8?

What is 0.8?

P[Testimony is Blue | Cab was Blue] = 0.8

Cab was … / Testimony is …	Blue	Green	All
Blue	0.12	0.17	0.29
Not Blue	0.03	0.68	0.71
All	0.15	0.85	1

P[Cab was Blue & Testimony is Blue] =
P[Cab was Blue] * P[Testimony is Blue | Cab was Blue] =
.15 * .80 = .12

P[Cab was Blue | Testimony is Blue] =
P[Cab was Blue & Testimony is Blue] / P[Testimony is Blue] =
.12 / .29 = .4138

P[Cab was Green & Testimony is Blue] =
P[Cab was Green] * P[Testimony is Blue | Cab was Green] =
.85 * .20 = .17

Actually, More Likely cab was Green, even with Testimony that it was Blue

Probability

Independent Events

Two events, C_1 and C_2, are independent if the occurrence (or nonoccurrence) of one gives no information about the likeliness of occurrence for the other.
$P[C_2] = P[C_2|C_1] = P[C_2|C_1{}^C]$

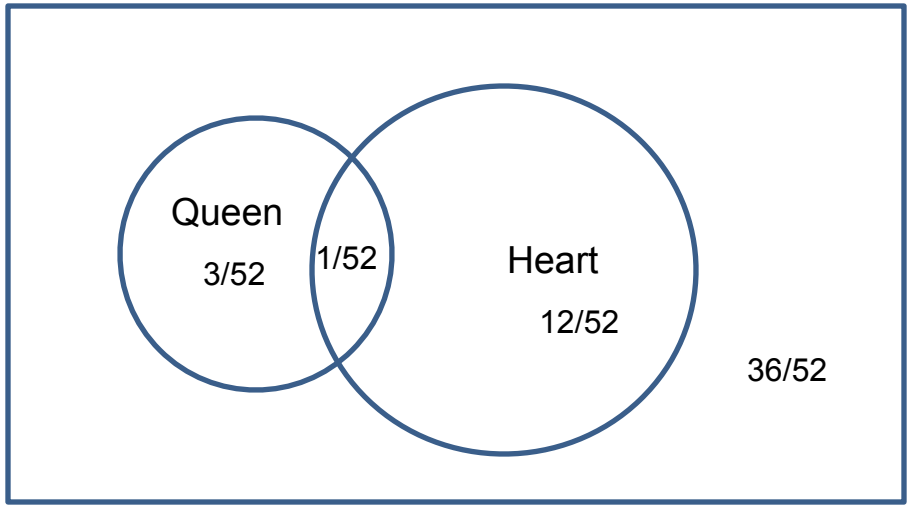
Consider Drawing a Single Card
From a Well-Shuffled Card Deck

Let Q = Card is a Queen
Let H = Card is a Heart

P[Q] = 4/52 = 1/13
P[Q|H] = 1/13
P[Q|Not H] = 3/39 = 1/13

P[H] = 13/52 = 1/4
P[H|Q] = 1/4
P[H|Not Q] = 12/48 = 1/4

Hence, Q and H are Independent



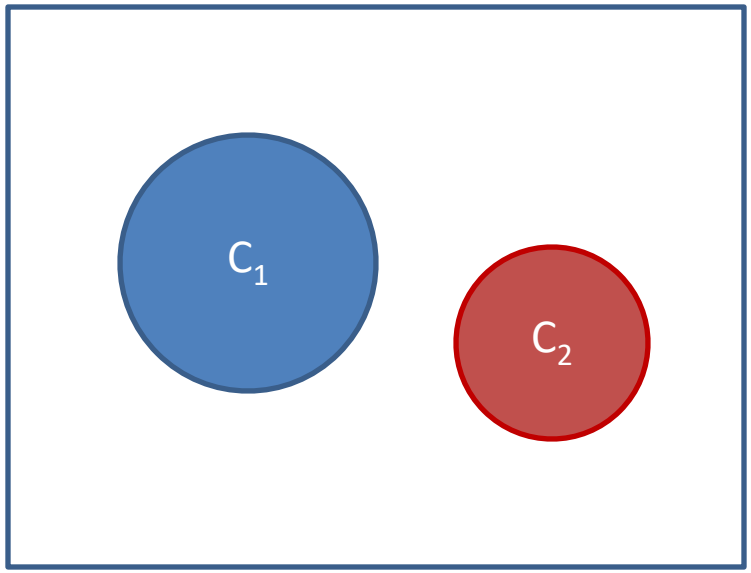
NOTE: If events C_1 and C_2 are Independent, then $P[C_1 \cap C_2] = P[C_1] * P[C_2]$

P[Q & H] = P[Q] * P[H] = 1/13 * 1/4 = 1/52

Probability

Independent Events

Question: Are C_1 and C_2 Independent?



Mutually Independent Events $C_1, C_2, \ldots, C_k$ implies for every collection of n events ($2 \le n \le k$):

$$P[\, C_{d_1} \cap C_{d_2} \cap \ldots \cap C_{d_n} \,] = P[C_{d_1}] * P[C_{d_2}] * \ldots * P[C_{d_n}]$$

with $\{d_1, d_2, \ldots, d_n\}$ being n distinct integers from 1, 2, …, k.

Probability

Summary

Work very carefully with the information given and the definitions of the concepts involved.

Where possible, try to draw a picture of the situation described in the problem (eg, a Venn Diagram, a Tree Diagram, a Table etc).

Often a first “off-the-top” response to a probability question turns out to be incorrect.

Use the knowledge you have of

- Rules of Probability (eg, all between 0 & 1, sum of all possibilities = 1)
- P[A or B] = P[A] + P[B] – P[A & B]
- P[A & B] = P[A] * P[B|A] = P[B] * P[A|B]
- P[Not A] = 1 – P[A]
- P[A | B] = P[A & B] / P[B] and P[B | A] = P[A & B] / P[A] for P[B]>0, P[A]>0
- If A & B Mutually Exclusive, P[A & B] = 0 and P[A or B] = P[A] + P[B]
- If A & B Independent, P[A] = P[A|B] = P[A|Not B] & P[B] = P[B|A] = P[B|Not A]
- If A & B Independent, P[A & B] = P[A] * P[B]

Statistical Analysis I

Statistical Models, Sampling Distributions, and Basic Inference Procedures

Statistical Models

"All models are wrong ,

but some are more useful than others." - George Box

One model that has proved "useful" for many years to describe outcomes X_i, i = 1, …, n of random processes is given as

$$X_i = \mu + \epsilon_i,$$

where μ = an expected value or mean value for each X_i, and

ϵ_i = a random deviation from this mean value for each X_i, i = 1, …, n.

What makes the model "statistical" is the use of a distributional model for the random deviations ϵ_i.

Far and away the most commonly applied distributional model is the Normal distribution, including the assumptions of

1) a mean (expected value) of zero for each of the deviations,
2) a common amount of variation for each of the deviations, and
3) mutual independence among the deviations, denoted as

$\epsilon_i \sim NID(0, \sigma^2)$, where $\sigma^2 = Var(\epsilon_i)$, i = 1, …, n.

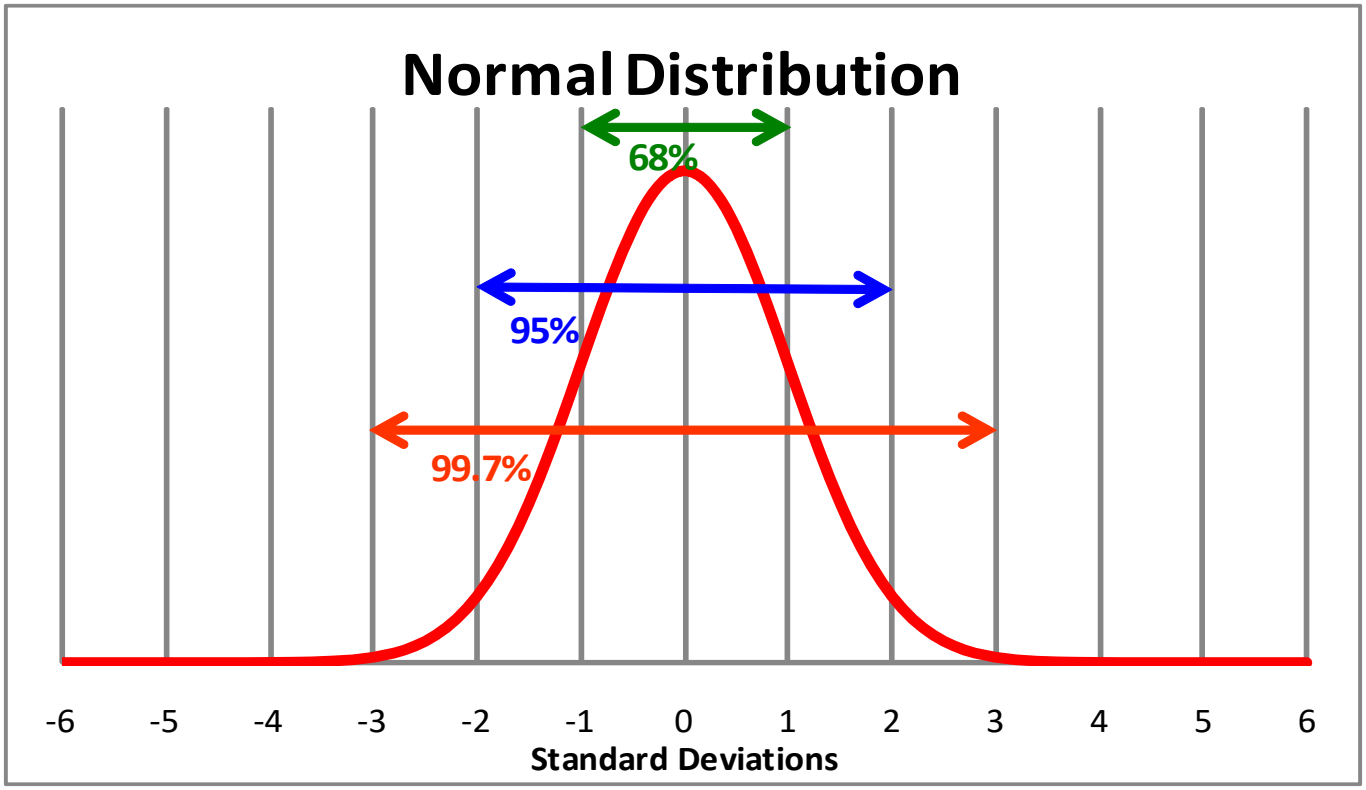
Normal Probability Density Function

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

Mean = μ

Standard Deviation = σ

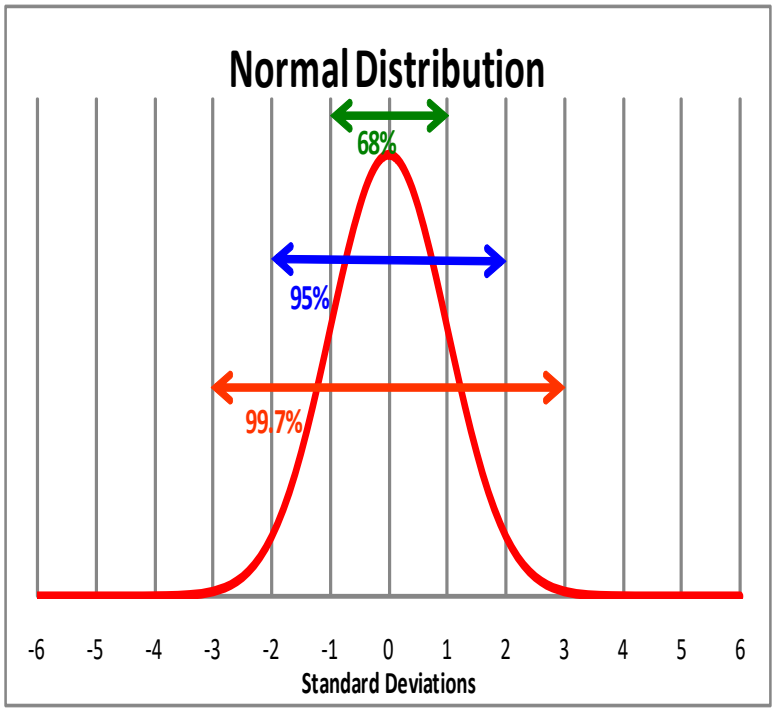
$P[a < x < b] = {}_a\int^b f(x)\,dx$



Note: As with all distributions, total area under the curve = 1

Standard Normal Distribution

A Standard Normal Distribution is when
the mean $\mu=0$ and
the standard deviation $\sigma=1$



Properties of the Standard Normal Distribution

1. The total area under the curve is equal to 1
2. The distribution is “bell-shaped”
    1. Symmetrical
    2. Mounded (Unimodal)
    3. Extends infinitely in both directions
3. Has Mean = 0 and Standard Deviation = 1
4. The Mean divides the area under the curve in half
    1. 0.5 Below the Mean
    2. 0.5 Above the Mean
5. Nearly All the area (99.7% = 0.9970) lies between
    1. a = -3 and
    2. b = +3
6. The probability of any specific value is Zero (ie, P[x=a] = 0)

Normal Distributions

Any Normal Distribution can be Converted to a Standard Normal Distribution by
- Subtracting its Mean from the Value(s) of Interest, and then
- Dividing this by its Standard Deviation
- This is commonly referred to as generating a "Z-score"

Consider the information that the Heights of Kindergarten children (ie, X) is Normally distributed with a Mean = 39 inches, and a Standard Deviation = 2 inches

In order to find the Probability an Individual Kindergartener will be between 38 and 40 inches tall, we would convert these values to their Z-scores:

$X = 38 \Rightarrow Z = (38-39)/2 = -1/2$ and $X = 40 \Rightarrow Z = (40-39)/2 = 1/2$, and then

Compute the desired probability using the Standard Normal Distribution result
For the $P[38 < X < 40] = P[-1/2 < Z < 1/2] = 0.3829$

This can be done using a table

OR

By using the following Excel commands:
=NORMSDIST(0.5) – NORMSDIST(-0.5) **R** command: pnorm(0.5) – pnorm(-0.5)

OR

No need to convert to z-scores if you use the following Excel commands:
=NORMDIST(40,39,2,TRUE) – NORMDIST(38,39,2,TRUE)

OR

R command: pnorm(40,39,2) – pnorm(38,39,2)

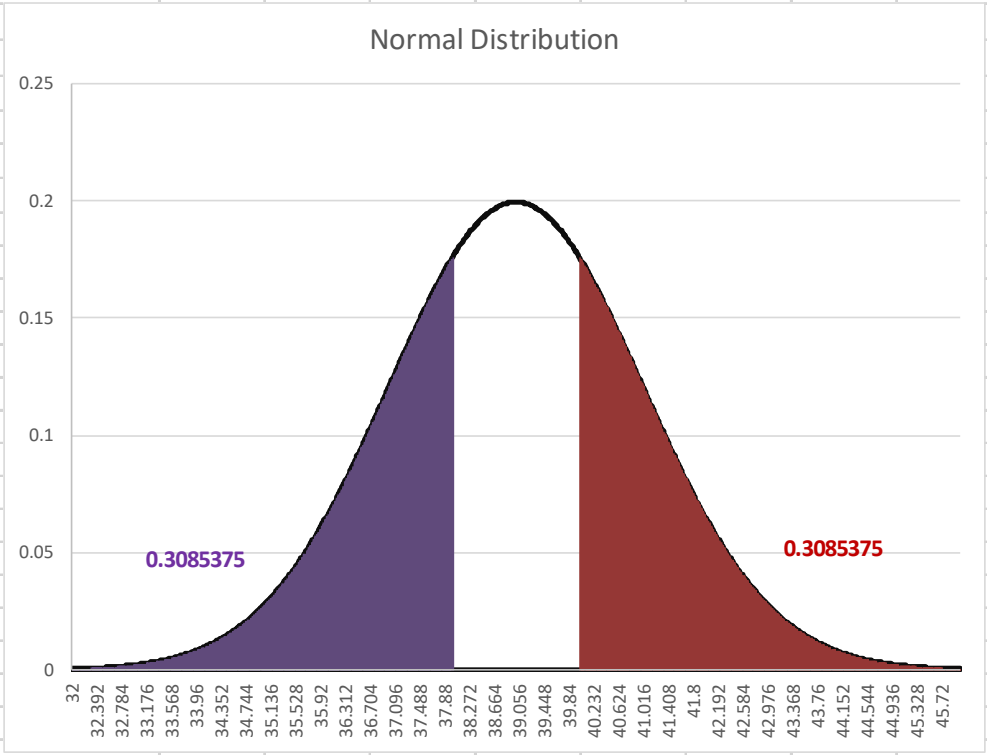
Normal Distributions

Or we could just use an Excel Utility:

Input Information					
Mean	39	No Entry will Default to 0			
Standard Deviation	2	No Entry or Non-positive Entry will Default to 1			
Data Value	38	OR	Probability		Will Default to Data Value if Both Entered
Optional Additional					Note: Any entered Probabilities MUST be between 0 and 1, EXCLUSIVE
Data Value	40	OR	Probability		Will Default to Data Value if Both Entered

Output Results

Area	Data Value	Probability
Below	38	0.3085
Above	40	0.3085
Between	38 & 40	0.3829



Normal Distributions

This Utility can also be used to answer such questions as:

What Height would we only expect 5% of Kindergartners to exceed?

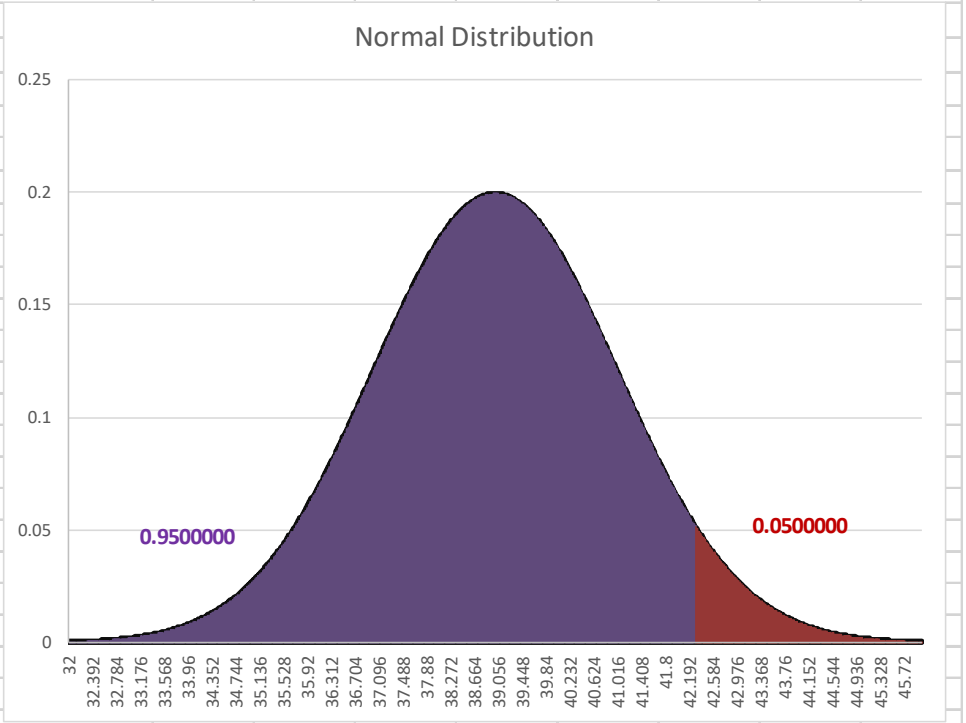
In other words, to find x_0 where $P[X < x_0] = P[Z < (x_0 - \mu)/\sigma] = 0.95$.

Input Information					
Mean	39	No Entry will Default to 0			
Standard Deviation	2	No Entry or Non-positive Entry will Default to 1			
Data Value		OR	Probability	0.95	Will Default to Data Value if Both Entered
Optional Additional Data Value		OR	Probability		Will Default to Data Value if Both Entered

Note: Any entered Probabilities MUST be between 0 and 1, EXCLUSIVE

Output Results

Area	Data Value	Probability
Below	42.29	0.9500
Above	42.29	0.0500
Between	42.29 & 42.29	0.0000

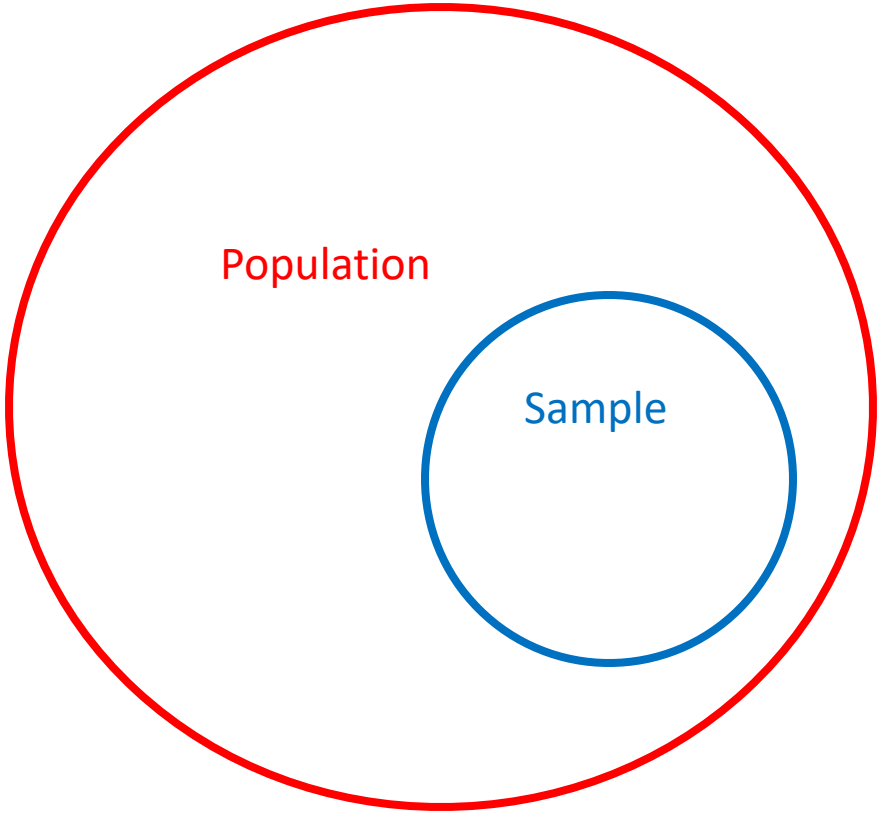


Excel command: norm.inv(0.95,39,2)

R Command: qnorm(0.95,39,2)

Point Estimates

A **sample statistic** is used to estimate a **population parameter**. Recall, what we generally want to *know* is a **population parameter**, but all we have available is a subset of the **population** – a **sample**, from which we can generate **sample statistics**.



Suppose we are interested in Knowing the Mean of a Specific Quality Characteristic Being Produced by a Manufacturing Process

Population: All parts being produced

Sample: A few parts selected for Measurement of the Specific Quality Characteristic (eg, Size)

Parameter: Population Mean, μ
Mean Size being produced by entire process

Statistic: Sample Mean, $\bar{X}$
Sample Average Size

Sample Average Size, $\bar{X}$, is **Point Estimate** of Production Process Mean Size, μ.

$\bar{X}$ is also a random variable, and the model for each X_i of the Sample will extend to $\bar{X}$.

Central Limit Theorem

Most Powerful Theorem in Statistics

If repeated samples of size n are obtained from any population,
The sampling distribution of the sample means will

1. Have the same mean value as the original population
2. Have a smaller standard deviation than the original population by a factor $1/\sqrt{n}$
3. Have a distribution that tends to be normal, more so as n increases in size

Sampling Distribution of Sample Averages ($\bar{X}$) can be described as follows:

1. $\mu_{Avg} = \mu$
2. $\sigma_{Avg} = \sigma/\sqrt{n}$

NOTE: $E[\bar{X}] = \mu_{Avg} = \mu$, so $\bar{X}$ **unbiased** estimator of μ

NOTE: Standard Deviation of $\bar{X}$ = σ_{Avg} is **SMALLER** than the Standard Deviation of Individual Results = σ by a factor of $1/\sqrt{n}$

3. Tends towards a N

NOTE: Distribution is **EXACTLY** Normal if distribution of Individual Results is Normal. Only need n ~10 for a "mounded" distribution, n > 30 usually sufficient for most other distributions

Central Limit Theorem

Example

A common practice in manufacturing is to sample product as produced by the manufacturing process. Often several parts/packages/units are sampled at a given time and evaluated.

Imagine you work for a company that produces touch screen panels for smart phones, and at the end of each shift 5 screens are sampled from the end of the line and measured for flatness.

The specifications for flatness results is 0 ± 1 micron and currently, the process appears to be running near its target of 0 with a standard deviation of 0.35 microns. In addition, the flatness results can be reasonably described using a normal distribution.

1) What is the probability that an individual screen will be out-of-spec?

Let F = Flatness Measurement on a single screen

$F \sim N(0, 0.35)$

Looking for $1 - Pr[-1 < F < 1] =$

$1 - Pr[(-1 - 0)/0.35 < Z < (1 - 0)/0.35] =$

$1 - Pr[-2.857 < Z < 2.857] = 0.0043$

2) What is the probability that a sample average of 5 screens will be out-of-spec?

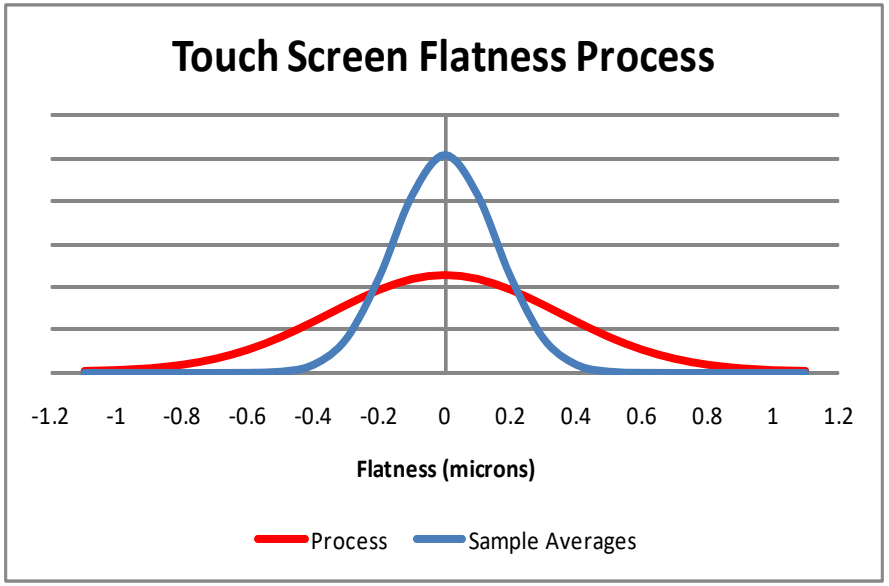
Let $\bar{F}$ = Average Flatness of 5 screens

$\bar{F} \sim N(0, 0.35/(\sqrt{5})) = N(0, 0.1565)$

Looking for $1 - Pr[-1 < \bar{F} < 1] =$

$1 - Pr[(-1 - 0)/0.1565 < Z < (1 - 0)/0.1565] =$

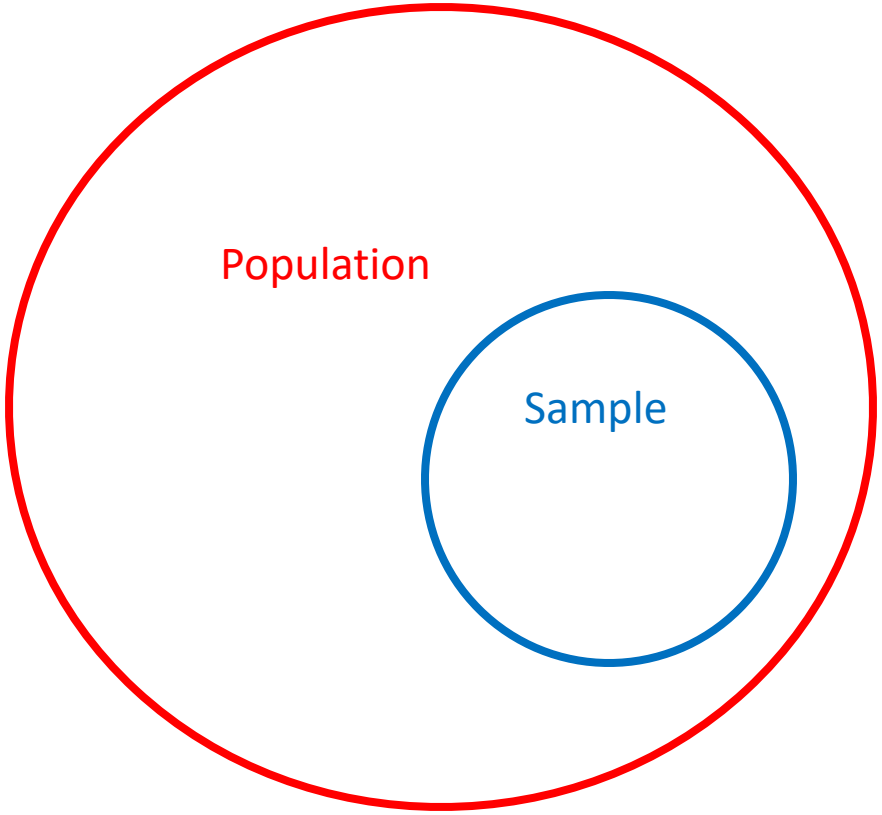
$1 - Pr[-6.389 < Z < 6.389] = 0.000000000167$



So what do you think of the practice of releasing screens for shipment if the sample average is within specifications?

Point Estimates

A sample proportion, **p**, is also a **sample statistic** used to estimate a **population parameter**. The statistical model involved is not quite the same as the simple mean + error model, but it is still relatively simple: $X_i = 1$ with probability **π** and $X_i = 0$ with probability (1-**π**). The parameter **π** represents the probability of the outcome of interest. [Bernoulli Trials]



Suppose we are interested in Knowing the True Proportion of Orange Reese's Pieces Being Produced by Hershey

Population: All Reese's Pieces Being Produced by Hershey

Sample: A group of Reese's Pieces selected at random from a Bag of Reese's Pieces (eg, samples of 25)

Parameter: Population Proportion, π, of Orange RPs being produced by Hershey

Statistic: Sample Proportion, p
Number of Orange in Sample (X) / Sample Size (n), so p = X/n

Sample Proportion, p = X/n, is **Point Estimate** of Hershey's Process Proportion, π.

Sample Proportion

A sample proportion is really a sample average in disguise

If we imagine each sample point as a Bernoulli trial, and
Let the random variable

$$X_i = \begin{cases} 1 \text{ if that trial is a “success”} \\ 0 \text{ if that trial is a “failure”} \end{cases}$$

Then the average of the X_is = $\bar{X}$ = [$\text{Sum}_{i=1 \text{ to } n}(X_i)$]/n = p,
Since [$\text{Sum}_{i=1 \text{ to } n}(X_i)$] is a count of the number of “successes” in the sample

Recall the sample proportion p is a **Point Estimate** of the corresponding population parameter π, and since p is really an average statistic, by the **Central Limit Theorem**, the sampling distribution for p has

Mean = π,
Standard Deviation = Std Dev(X_i)/($\sqrt{n}$) = $\sqrt{\pi(1-\pi)/n}$, and
Will tend to be Normally Distributed as n increases

How large does n need to be for distribution to reasonably approximate a normal?
n > 9*max[(1-π)/π, π/(1-π)]
(ie, if π=0.5, then n>9)

Sample Proportion

Example: Suppose we want to estimate the proportion of Orange Reese's Pieces produced by Hershey, call this population parameter: π

The information we have is a sample of $n = 25$ Reese's Pieces

If we imagine each Reese's Piece (ie, sample point) as a Bernoulli trial, and Let the random variable

$$X_i = \begin{cases} 1 \text{ if that piece is Orange (ie, trial is a "success")} \\ 0 \text{ if that piece is not Orange (ie, trial is a "failure")} \end{cases}$$

Then the $[\text{Sum}_{i=1 \text{ to } 25}(X_i)]$ = Count of Orange Reese's Pieces in the sample,
Say this Count = 10, then $p = 10/25 = 0.4$ (ie, $[\text{Sum}_{i=1 \text{ to } 25}(X_i)]/n$)

Now, suppose Hershey claims that the proportion of Orange Reese's Pieces it routinely produces is 50% (ie, $\pi = 0.5$), then the sampling distribution for p should have

Mean = $\pi = 0.5$,
Standard Deviation = $\sqrt{\pi(1-\pi)/n} = \sqrt{0.5(1-0.5)/25} = 0.1$, and
Will tend to be Normally Distributed as n increases ($n=25>9$, so reasonably normal)

So … is the proportion $p = 0.4$ in our sample of 25 unusual?

Can we use the information above to help answer this question?

Sample Proportion

Let's use the information about the sampling distribution for p from the CLT:

Mean = $\pi = 0.5$,
Standard Deviation = $\sqrt{\pi(1-\pi)/n} = \sqrt{0.5(1-0.5)/25} = 0.1$, and
Will tend to be Normally Distributed as n increases (n=25>9, so reasonably normal)

To estimate the probability that we would observe a proportion p <= 0.4 in a sample of 25 if, in fact, Hershey is making 50% of its Reese's Pieces candies Orange:

$$P[\, p <= 0.4 \mid \pi = 0.5 \,] = P\left[\, (p - 0.5)/0.1 <= (0.4 - 0.5)/0.1 \,\right]$$
$$= P[\, Z <= -1 \,]$$
$$\approx 0.1587$$

So … is the proportion p = 0.4 in our sample of 25 unusual? What do you think?

What if our sample size was 100 and we only observed 40 Orange RPs (ie, p still 0.4)?

The sample proportion is still the same (ie, p = 0.4), but we have more information n = 100 vs n = 25, and the sampling distribution for p becomes more narrow.
The CLT tells us how much more narrow as Standard Deviation was 0.1 for n = 25, but now is $\sqrt{0.5(1-0.5)/100} = 0.05$, for n = 100, and

$$P[\, p <= 0.4 \mid \pi = 0.5 \,] = P\left[\, (p - 0.5)/0.05 <= (0.4 - 0.5)/0.05 \,\right]$$
$$= P[\, Z <= -2 \,]$$
$$\approx 0.0228$$

So … now what do you think?

Elements of a Test of Hypothesis

Suppose we work for a pharmaceutical company, and have a new drug for the treatment heart murmurs.
The competition has an already established drug for this on the market that is known to work successfully with 65% of patients to which it is administered.
We want to know if the success rate for our new drug is higher than the 65% success rate for the currently available competitive offering.

The hypothesis we would like to validate is that the success rate for our drug (call this π) is greater than 65% (ie, 0.65).

This is the **RESEARCH HYPOTHESIS**

(also often called the **ALTERNATIVE HYPOTHESIS**)

It can be expressed in abbreviated form as: $\mathbf{H_1: \pi > 0.65}$

The **RESEARCH HYPTOHESIS** is the one we would like to validate.

Elements of a Test of Hypothesis

In order to establish the validity of the **RESEARCH HYPOTHESIS**, we need take an indirect approach and assume the opposite is true.

We assume that the success rate of the new drug, π, is 65% (or less).

This is the **NULL HYPOTHESIS**

(also often called the **NULL MODEL**)

It can be expressed in abbreviated form as: $\mathbf{H_0: \pi = 0.65}$

(NOTE: Being less than 65% is implied from the form of the **RESEARCH HYPOTHESIS**)

The **NULL HYPTOHESIS** is the OPPOSITE of the one we would like to validate.

The actual test is focused on evaluating if the **NULL HYPOTHESIS** is valid, or reasonable, hoping for sufficient evidence in the test to contradict this, and then conclude that the **RESEARCH HYPOTHESIS** must be valid instead.

Elements of a Test of Hypothesis

So now we have two Hypotheses:

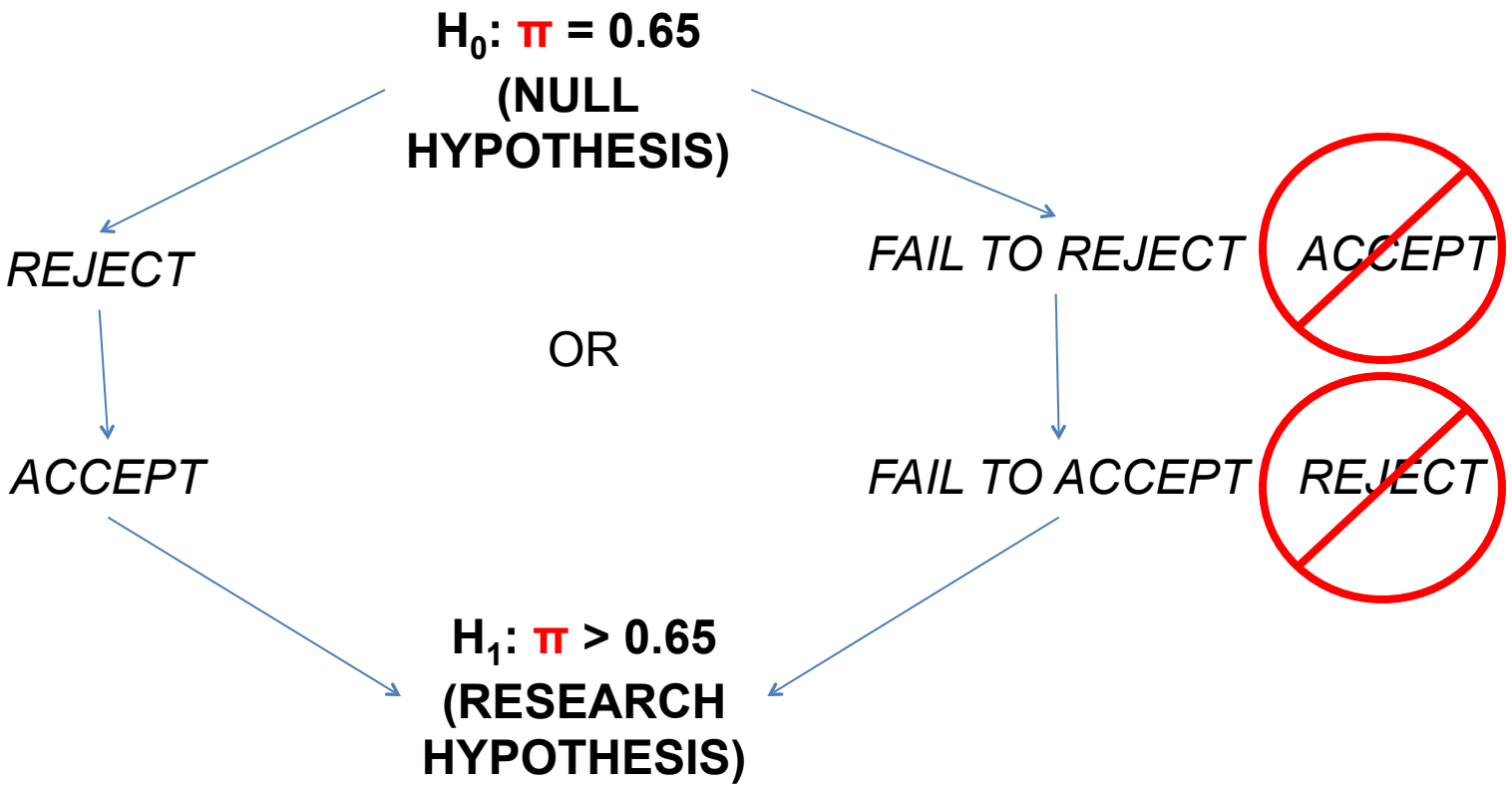
H_0: $\pi = 0.65$
(NULL HYPOTHESIS)
H_1: $\pi > 0.65$
(RESEARCH HYPOTHESIS)

We obtain test information (ie, data from a sample of some nature).
In the drug example, this would ideally be treatment results of patients administered the new drug.

We then evaluate the test (ie, sample) information in hopes that it will provide "sufficient evidence" to contradict (ie, *REJECT*) the **NULL HYPOTHESIS** and conclude the **RESEARCH HYPOTHESIS** is indeed true.

However, this sample data might not provide such evidence, and we will be left to conclude that the **NULL HYPOTHESIS** could still be viable (ie, *FAIL TO REJECT*) and fail to determine that the **RESEARCH HYPOTHESIS** is true.

Elements of a Test of Hypothesis



So … NEVER *ACCEPT* the **NULL HYPOTHESIS**, Only *REJECT* or *FAIL TO REJECT*
and NEVER *REJECT* the **RESEARCH HYPOTHESIS**, Only *ACCEPT* or *FAIL TO ACCEPT*

Elements of a Test of Hypothesis

Once a **NULL MODEL** has been established, then we need a way to use the sample data to evaluate (ie, test) it.
This is the role of the **TEST STATISTIC**, which is generally a function of the sample statistic corresponding to the population parameter of interest.
The **TEST STATISTIC** has a sampling distribution that can be specified under the **NULL MODEL**, this is called the **NULL DISTRIBUTION**.

For the drug example, the **TEST STATISTIC** will be the sample proportion
p = number of patients tested where drug is success (n_S) / total number tested (n)

Under the **NULL HYPOTHESIS** (**H_0: π=0.65**), the sampling distribution of **p** is

$p \sim Bin(\pi, n)$, which for $n > 9*.65/.35 = 16.7$ can be approximated by

$p \sim N(\mu = \pi, \sigma = \sqrt{\pi*(1-\pi)/n}) = N(\mu=0.65, \sigma \approx 0.477/\sqrt{n})$

Suppose n = 50, then $p \sim N(\mu=0.65, \sigma \approx 0.067)$

The **NULL DISTRBUTION** is the sampling distribution of the **TEST STATISTIC** under the **NULL MODEL** (ie, assuming the **NULL HYPOTHESIS** is true).

Elements of a Test of Hypothesis

Once we have a **TEST STATISTIC** and its **NULL DISTRIBUTION**, then we need a **DECISION RULE** to determine the validity of the **NULL MODEL**.

In any decision, we run a risk of making an incorrect choice.
It is no different in testing hypotheses.

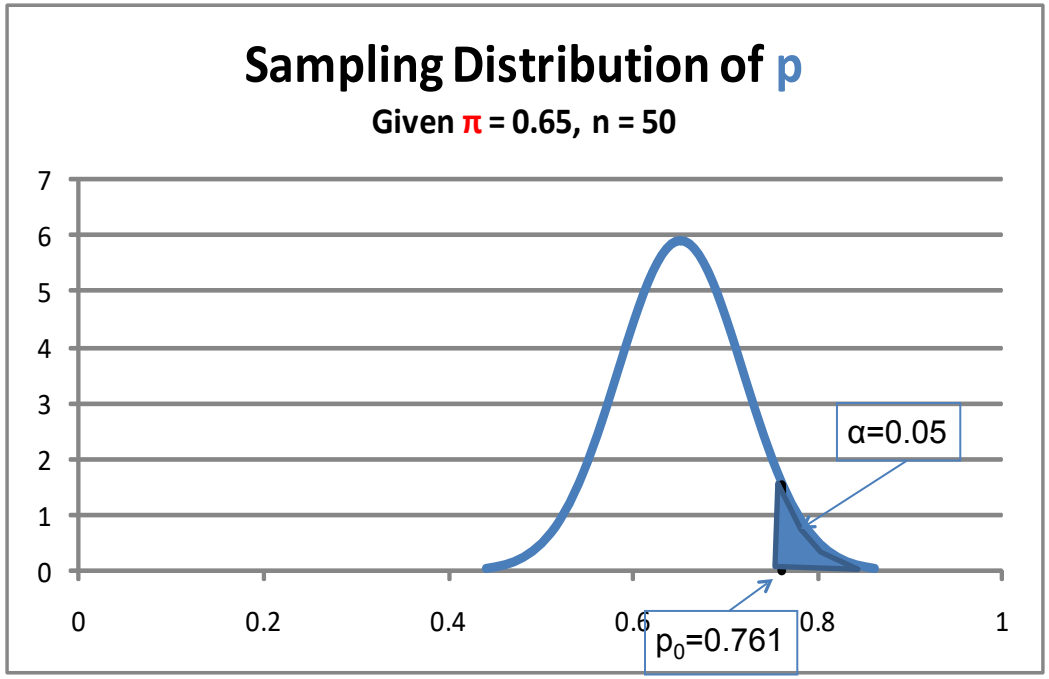
	NULL MODEL is Actually …	
Based on Sample **TEST STATISTIC** …	**True**	**False**
REJECT **NULL MODEL**	Type I Error P[Type I Error] = α α = Significance Level Determines Critical Region for the **TEST STATISTIC**	Correct
FAIL TO REJECT **NULL MODEL**	Correct	Type II Error P[Type II Error] = β 1- β = Power Function of Δ Between **NULL MODEL** & Actual

The **DECISION RULE** is established by choosing a Significance Level = α
(ie, P[Type I Error] = risk of incorrectly rejecting the **NULL HYPOTHESIS**)

Elements of a Test of Hypothesis

For our drug example, if we set the significance level $\alpha = 0.05$, then given the **NULL DISTRIBUTION** of our **TEST STATISTIC**, we would define our **DECISION RULE** as any $p > p_0$ where

$$P[\, p > p_0 \mid \pi = 0.65, n = 50] = 0.05 \text{ (ie, } \alpha)$$



Since the **NULL DISTRIBUTION** of p is its Sampling Distribution Given $\pi = 0.65$ & $n = 50$ is

$p \sim N(\mu = 0.65, \sigma \approx 0.067)$,

$p_0 = 0.761$ (ie, 1.645*0.067)

So … **DECISION RULE** is

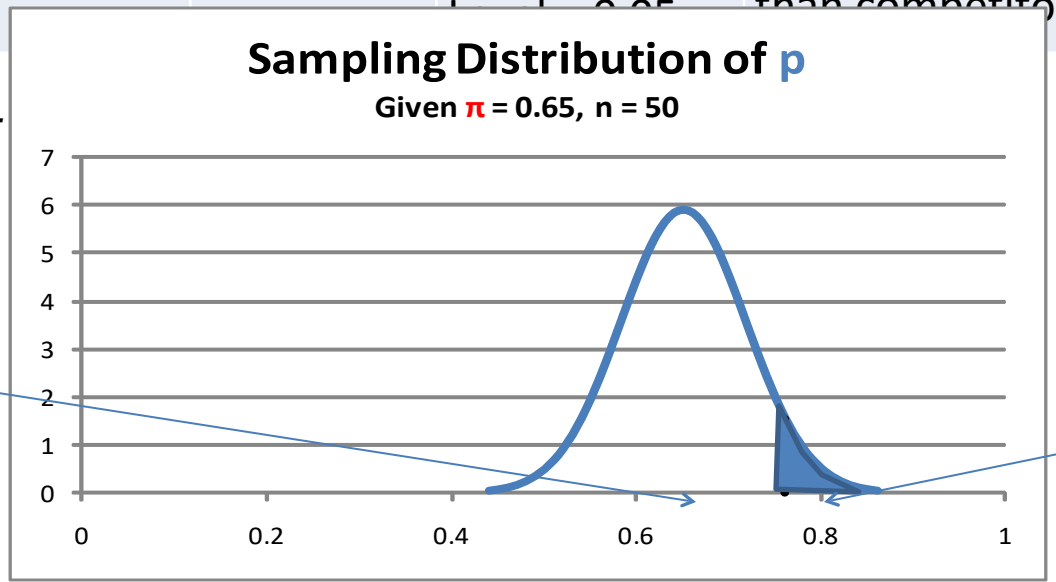
REJECT $\mathbf{H_0}$**:** $\boldsymbol{\pi}$**=0.65** if p **> 0.761**

Elements of a Test of Hypothesis

The final elements are a **DECISION** and a **CONCLUSION**.

Observed Successful Results	TEST STATISTIC	DECISION	CONCLUSION
$n_S = 40$	$p = 0.80$	REJECT H_0 at Significance Level = 0.05	New Drug has success rate higher than competitor
$n_S = 35$	$p = 0.70$	FAIL TO REJECT H_0 at Significance	Insufficient Evidence to conclude success rate for New Drug is higher than competitor

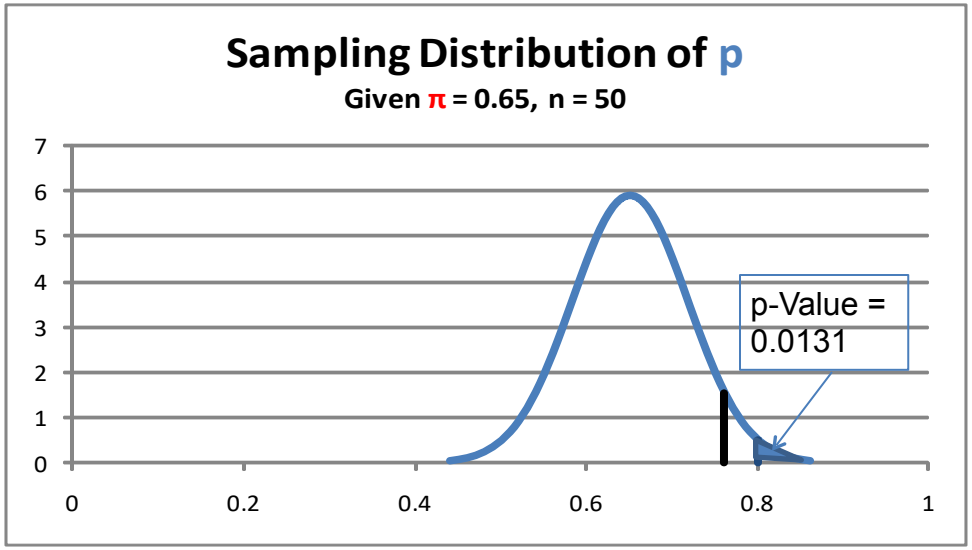
Not Large Enough to Fall in Critical/*REJECT* Region



Large Enough to Fall in Critical/*REJECT* Region

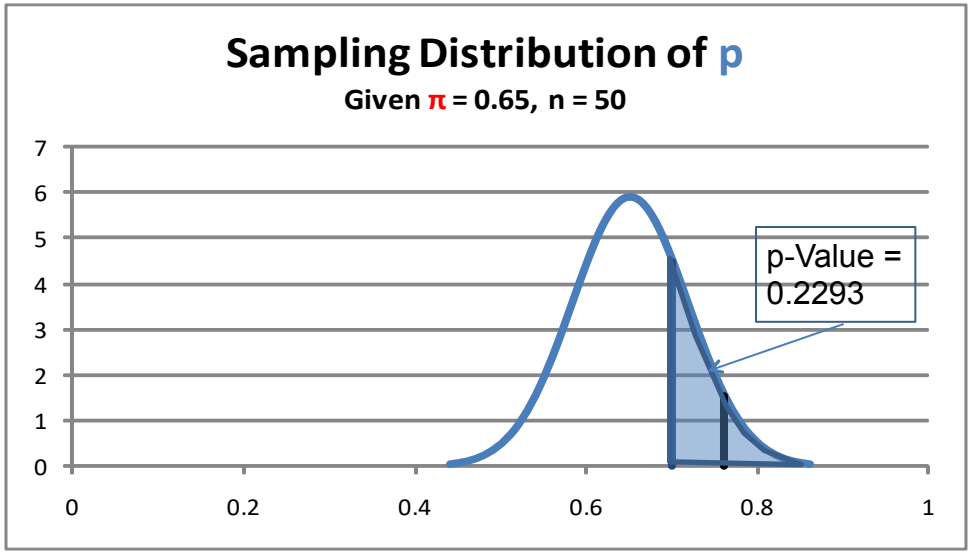
Elements of a Test of Hypothesis

Observed Significance Level = p-Value



If we observe p = 0.8, then
P[p > 0.8 | π = 0.65, n = 50] ≈ 0.0131
is the observed significance level, and is known as the p-Value for this test.

A p-Value < Significance Level = α indicates that H_0 will be *REJECT*ed.



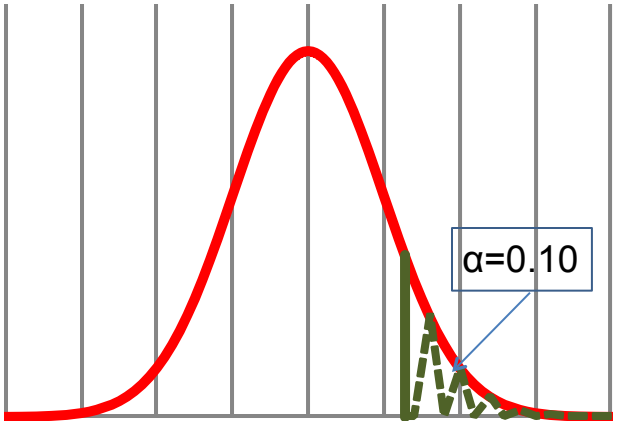
If we observe p = 0.7, then
P[p > 0.7 | π = 0.65, n = 50] ≈ 0.2293
is the observed significance level, and is known as the p-Value for this test.

A p-Value > Significance Level = α indicates that we will *FAIL TO REJECT* H_0

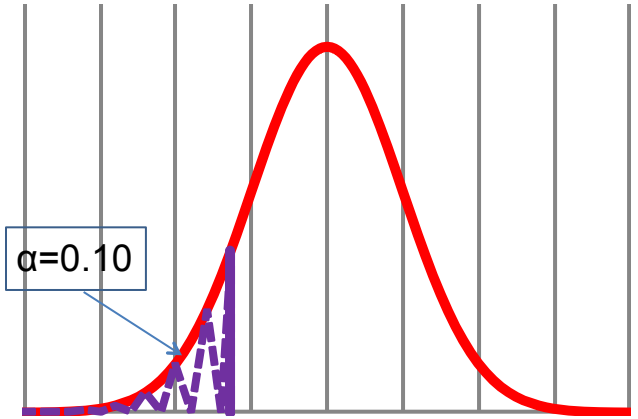
Types of Research Hypotheses

For Significance Level α = 0.10

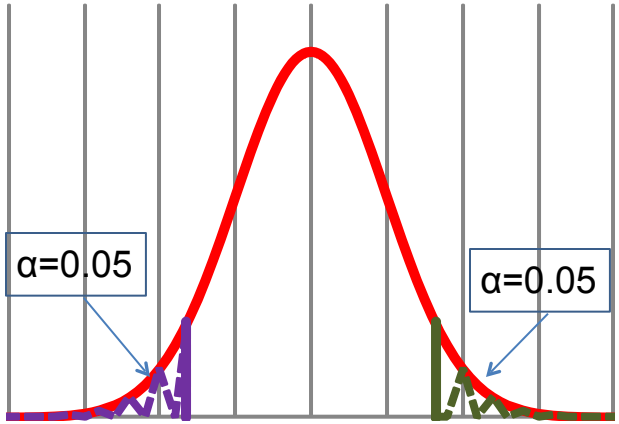
Right Tail Research Hypothesis: H_1: $\pi > \pi_0$
Rejection Region all in Right Tail



Left Tail Research Hypothesis: H_1: $\pi < \pi_0$
Rejection Region all in Left Tail

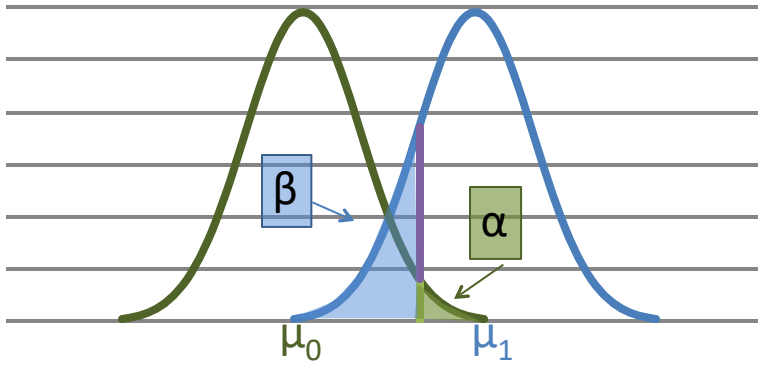


Two-Tail Research Hypothesis: H_1: $\pi \neq \pi_0$
Rejection Region Split into Both Tails

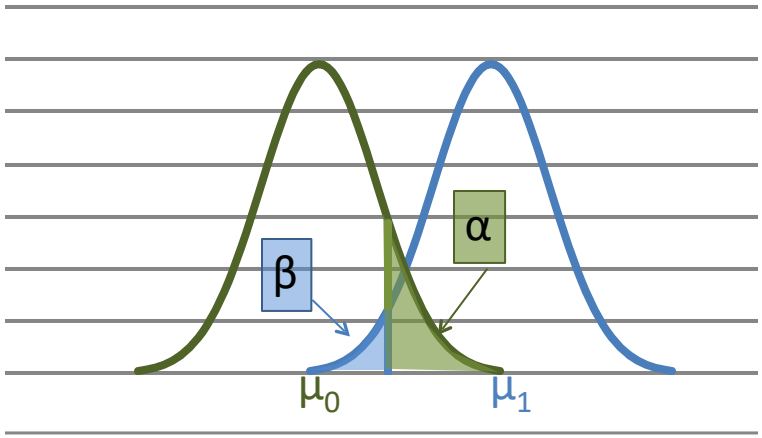


Type I and Type II Errors

Small α, Large β



Large α, Small β



As we seek smaller Type I Error rates,
We necessarily increase Type II Error rates
(and vice versa).

So … we need to assess the relative consequences of making each type of error, and choose the Significance Level for the test accordingly.

Drug Example:
Type I Error – Determining Better when Really Not Better
Type II Error – Failing to Recognize Better when Really is Better

Can reduce risk of Type II Error by Increasing Risk of Type I Error

A rule-of-thumb approach:
Type II Error more Severe
- Set α from 0.05 to 0.10

Type I Error more Severe
- Set α from 0 to 0.05

~Same Consequences
- Use α ≈ 0.05

Confidence Intervals

In introducing hypothesis testing, the question of interest was whether or not a newly developed drug for heart murmurs performed better than the competitor's currently available offering.

Recall that there were 50 patients tested with the new drug, and if 40 responded positively, then we would have a sample proportion p = 0.8.

This was greater than the critical value of 0.76, so we Rejected H_0 (our new drug no better than the competitor's 65% success rate) and concluded that our drug is better than the competitor's drug.

There is still an unanswered question here. The test suggests it is better, but how much better is it?

Is the answer simply 15% better? Do we simply take the observed sample proportion 0.80 and subtract the 0.65 success rate of the competitive offering?

This is certainly a reasonable response, as the sample statistic here does provide our best *point* estimate of the population parameter we want to advertise, but there is a problem

$$P[\pi_{New} = p = 0.8] = 0$$

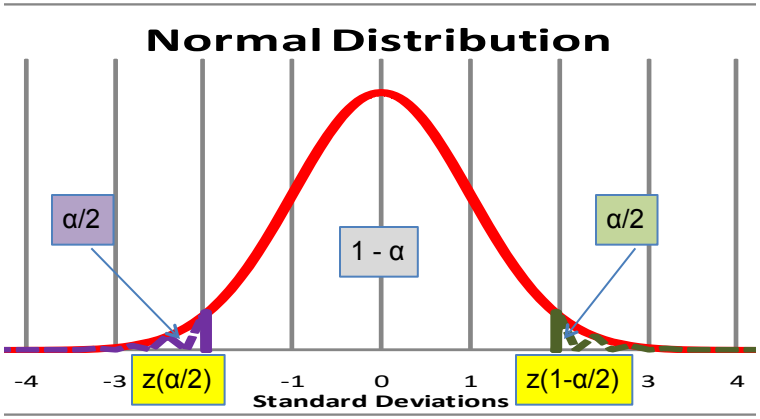
Interval Estimates

However, perhaps we can find an Interval that has a positive probability of including the Population Parameter we want to know.

Recall: A proportion is just a mean (average) in disguise

eg, P[Sample Average – Δ(ξ) < Population Mean < Sample Average + Δ(ξ)] = ξ

where Δ(ξ) > 0 and 0 < ξ < 1



Well … we know that

P[z(α/2) < Z < z(1-α/2)] = 1-α

and that

Z = (Sample Average – Population Mean) / Standard Error

Note: Standard Error is $\sigma/\sqrt{n}$, for a proportion this is $\sqrt{\pi(1-\pi)/n}$

P[z(α/2) < (Sample Average – Population Mean) / Standard Error < z(1-α/2)] = 1 – α

P[z(α/2) * Standard Error < Sample Average – Population Mean < z(1-α/2) * Standard Error] = 1 – α

P[z(α/2)*Std Error – Sample Average < - Population Mean < z(1-α/2)*Std Error – Sample Average] = 1 – α

P[Sample Average – z(1-α/2)*Std Error < Population Mean < Sample Average – z(α/2)*Std Error] = 1 - α

Population Mean between
Sample Average ± z(1-α/2)*Standard Error
with Probability 1-α

Confidence Intervals

Basic format of Confidence Intervals:

Best Point Estimate ± M(Confidence) * Standard Deviation of Point Estimate

where M(Confidence) is a Multiplier based on the Desired Level of Confidence that the Interval Contains the True Value of the respective Population Parameter of interest

For example, a 95% Confidence Interval for the success rate of the new drug, π_{New}, is given by

$$p \pm z(0.975) * \sqrt{p(1-p)/n}$$

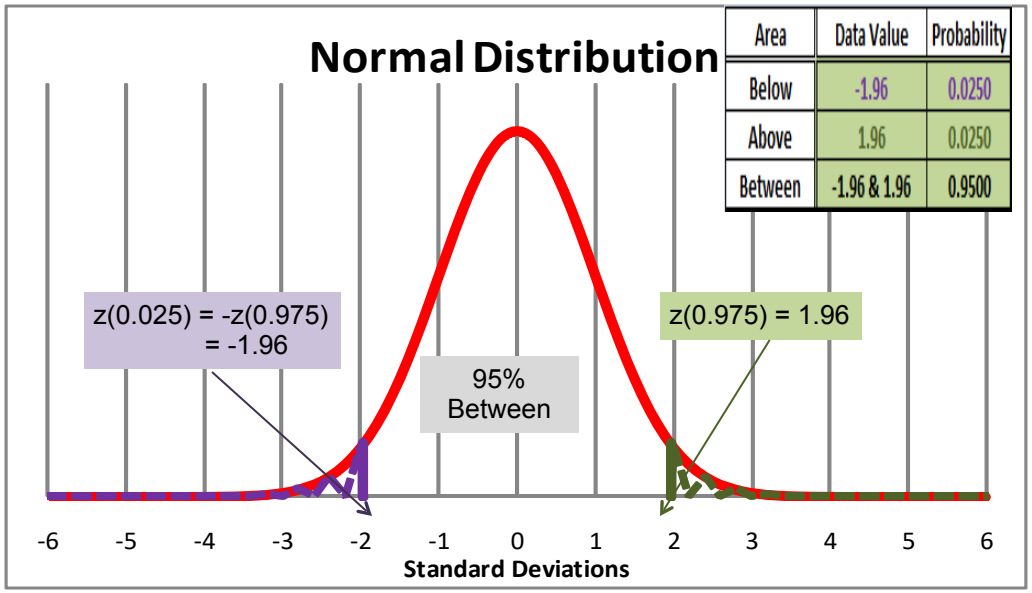
With p = 0.80, we can say with 95% Confidence that

π_{New} is in the interval defined by

$$0.8 \pm 1.96 * \sqrt{0.8*0.2/50}$$

or (0.689 to 0.911)

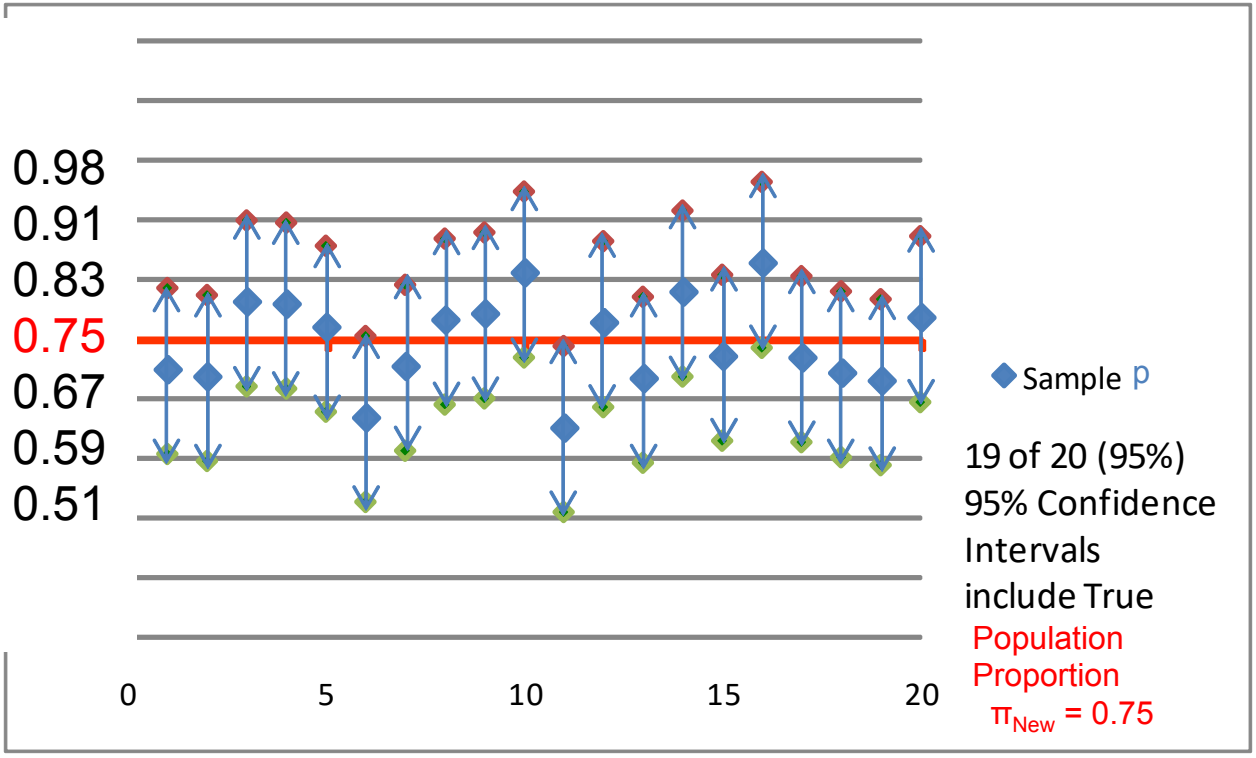
So, with 95% Confidence, the success rate for the new Drug is between ~69% and ~91%



Confidence Intervals

Concept: 95 of 100 95% Confidence Intervals will contain the True Population Parameter

Suppose that success rate for new drug, π_{New} = 0.75, then if we had 20 samples, and constructed 20 95% Confidence Intervals, we would expect 19 of the 20 to include 0.75.



Confidence Intervals = Hypothesis Tests

A Test of H_0: $\pi = \pi_0$ (or $\mu = \mu_0$) at a Significance Level of α

Is **Equivalent** to

A 1-α% Confidence Interval

If the Interval Does Not Contains π_0 (or μ_0), then Reject H_0

Otherwise, Conclude Insufficient Evidence to Reject H_0

In the example, $\pi_0 = 0.65$ is below the 95% Confidence Interval (0.689 to 0.911);

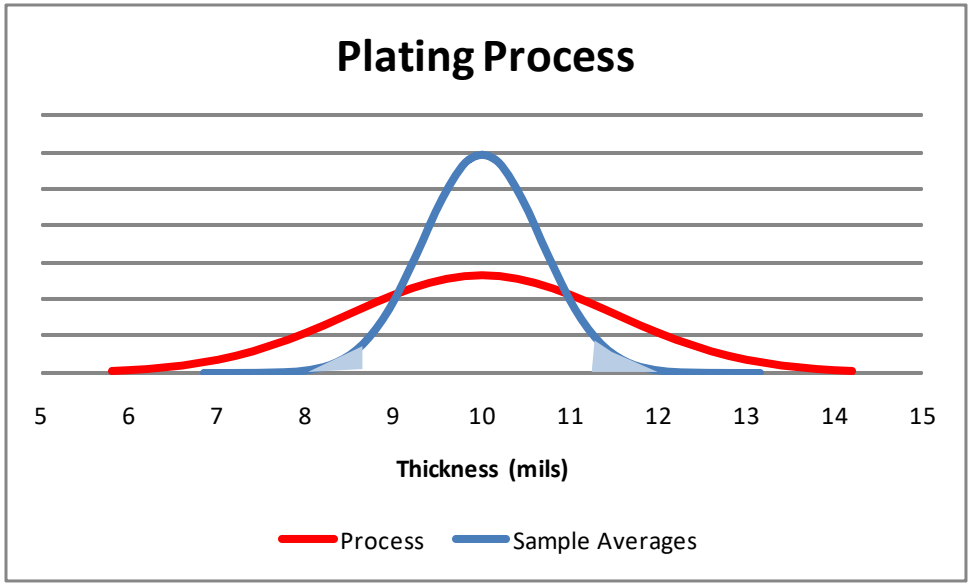
Hence, we would Reject H_0: $\pi = 0.65$ and Conclude that H_1: $\pi > 0.65$ is True,

at the 0.05 Level of Significance.

Inference on μ

σ Known

Suppose we have responsibility for managing the gold plating step in the production of specific electronic connector.
The process is routinely sampled with 5 connectors being evaluated for plating thickness.
It is important to keep the plating process on target since too little gold compromises functionality and too much gold significantly increases manufacturing costs.
The process target is 10 mils and the process has a known standard deviation, σ = 1.5 mil.
At each sampling opportunity, we need to decide if process is on-target or not.
Thickness results are normally distributed.



Research Hypothesis: H_1: μ ≠ 10
Null Hypothesis: H_0: μ = 10

Test Statistic: $\bar{X}$ (Sample Avg)
Null Distribution: $N(10, 1.5/\sqrt{5})$
$N(10, 0.67)$

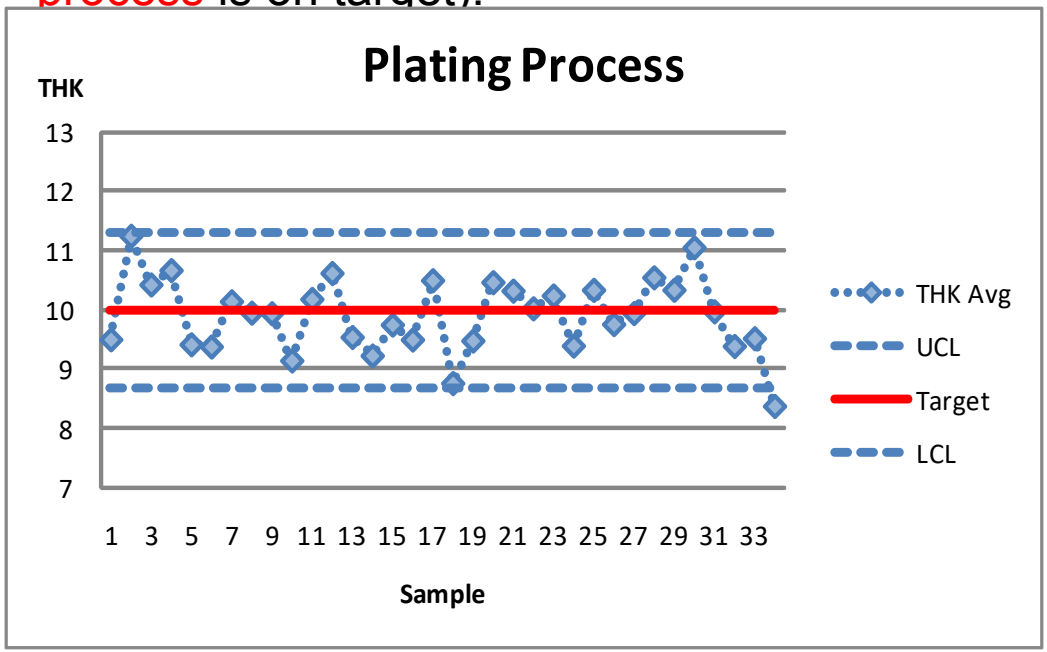
Decision Rule: Reject H_0 if
$\bar{X} < 8.69$ or
$\bar{X} > 11.31$

Values = 10 +/- 1.96 * 0.67
w/ P[Z > 1.96] = 0.025 (so α=0.05)

Confidence Intervals for μ

σ Known

Recall the gold plating process example, where with each sample we essentially perform a test of the Null Hypothesis (H_0: μ = 10) (ie, a test of whether or not the process is on target).



For Samples 1-33, $8.69 < \bar{X} < 11.31$, so Conclude Not Enough Evidence to Reject H_0, and Let Process Run with No Other Action

At Sample 34, $\bar{X} = 8.38$ We Reject H_0 and Conclude Process Off Target Low, and Take Action to Get Process Back On Target

At Sample 34, a 95% Confidence Interval for the Current Process Mean would be

Point Estimate = $\bar{X}$

Standard Error = $\sigma/\sqrt{n} = 1.5/\sqrt{5}$

$8.38 +/- 1.96*0.67 = (7.09 \text{ to } 9.69)$

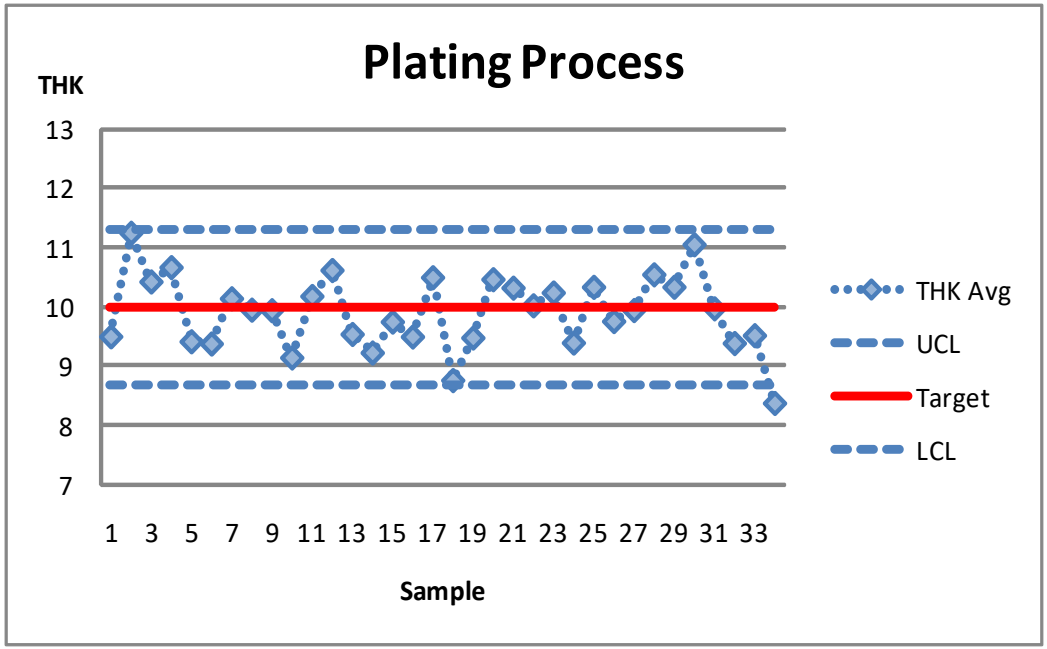
Multiplier = z(0.975)

Note that the interval does not include the Target = 10, Confidence Intervals for the previous 33 Samples will include Target = 10.

Inference on μ

σ Known

With each sample we essentially perform a test of the Null Hypothesis (H_0: μ = 10) (ie, a test of whether or not the process is on target).



For Samples 1-33,
8.69 < $\bar{X}$ < 11.31, so
Conclude Not Enough Evidence to Reject H_0, and
Let Process Run with No Other Action

At Sample 34, $\bar{X}$ = 8.38
We Reject H_0 and Conclude Process Off Target Low, and
Take Action to Get Process Back On Target

Is there a potential problem with managing the process this way?
How Often Can we Expect False Signals (ie, Type I Errors)? α = 0.05, so 1 in 20
What if the sampling frequency is 2X per day? Expect a False Signal once every 10 days
Similarly, we would expect 1 in 20 confidence intervals to not include μ = 10.
For this reason most organizations set limits at $\pm 3\sigma/\sqrt{n}$ (α = 0.003)

Student's t Distribution

By the CLT, we know

$$Z = (\bar{X} - \mu)/[\sigma/\sqrt{n}]$$

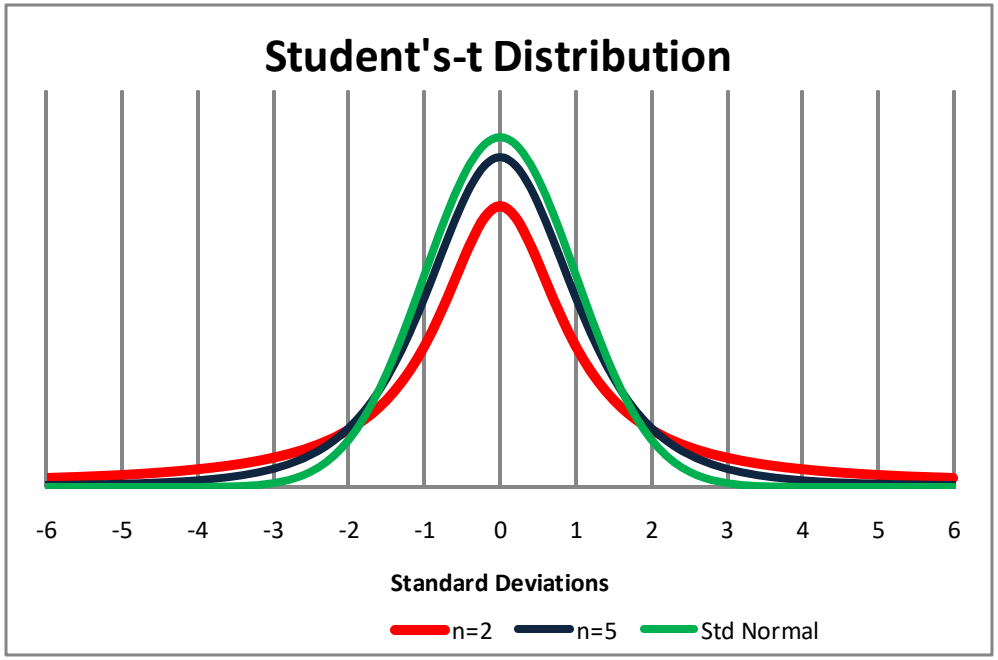
has a Standard Normal Distribution

For many years it was thought that since σ is rarely, if ever, known that simply replacing σ with an estimate from the sample (ie, S) still left

$$T = (\bar{X} - \mu)/[S/\sqrt{n}]$$

with a Standard Normal Sampling Distribution. However, …

About 100 years ago William Gosset working at Guiness discovered that for Smaller sample sizes, the sampling distribution of T had heavier tails than would be expected for a Standard Normal distribution. He published this work as "Student"; hence, for T we have



Student's t Distribution

So … if $X_1, X_2, \ldots, X_n$ is a random sample of size n drawn from a normal pdf with mean μ and variance σ^2, ie, each $X_i \sim N(\mu, \sigma)$, then the sampling distribution of

$$T = (\bar{X} - \mu)/[S/\sqrt{n}]$$

follows a Student's t Distribution with **n-1** degrees of freedom.

While the Standard Normal is only one distribution, Student's t is actually a family of distributions, and the members are identified by their degrees of freedom – for most problems we will have this will be the sample size minus one (ie, n-1).

As the degrees of freedom increase (as the sample size increases), the Student's t Distribution becomes increasingly close to a Standard Normal Distribution. For n ~100 or more, the distributions are virtually identical.

Student's t Distribution

Example: Suppose a random sample of size 10 is obtained from a normal pdf, and T = $(\bar{X} - \mu)/[S/\sqrt{n}]$ is calculated, what is

P[T < 3]?

Could use a t table, and find the value 3 on the row n-1 =9
Not likely to be in the table, but could interpolate between 2.821 & 3.250, and find 3 is 41.69% of the distance between these two values, the estimate P[T > 3] by taking the value 41.69% of the way between 0.010 and 0.005 to get 0.0079, so P[T < 3] ≈ 1 – 0.0079 = 0.9921

OR

Degrees of Freedom (ie, n – 1 = 10 – 1 = 9)

Could use Excel command: = 1 – TDIST(3, 9, 1) = 0.9925

Value

Tails

R command: pt(3, 9)

OR ….

Student's t Distribution

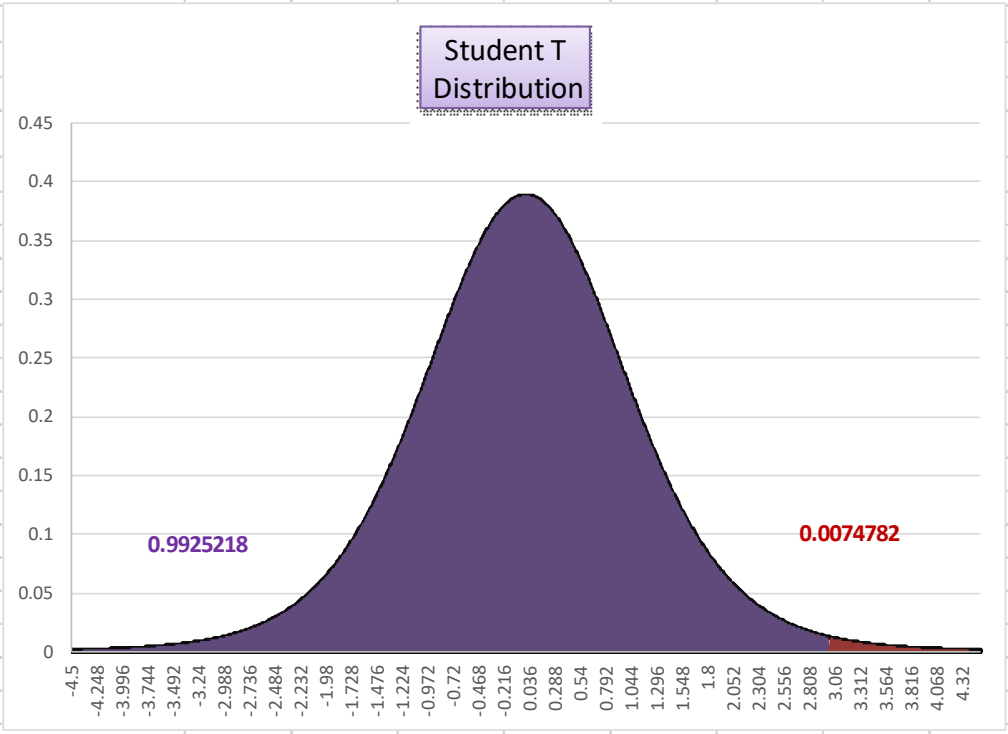
Could use Excel utility:

Input Information

Degrees of Freedom	9				
Low T Score	3	OR	Probability		Will Default to Data Value if Both Entered
				Note: Any entered Probabilities MUST be between 0 and 1, EXCLUSIVE	
Upper T Score		OR	Probability		Will Default to Data Value if Both Entered

Output Results

Area	Data Value	Probability
Below	3	0.9925
Above	3	0.0075
Between	3 & 3	0.0000



Student's t Distribution

Example: Find the value t_0 such that $P[-t_0 < T < t_0] = 0.95$ (again, for T with 9 df)

Could use a t table and find entry in row n – 1 = 9, and column 0.025 = 2.262

OR

Could use Excel command: TINV(0.05,9) or **R** command: qt(0.975, 9)

OR

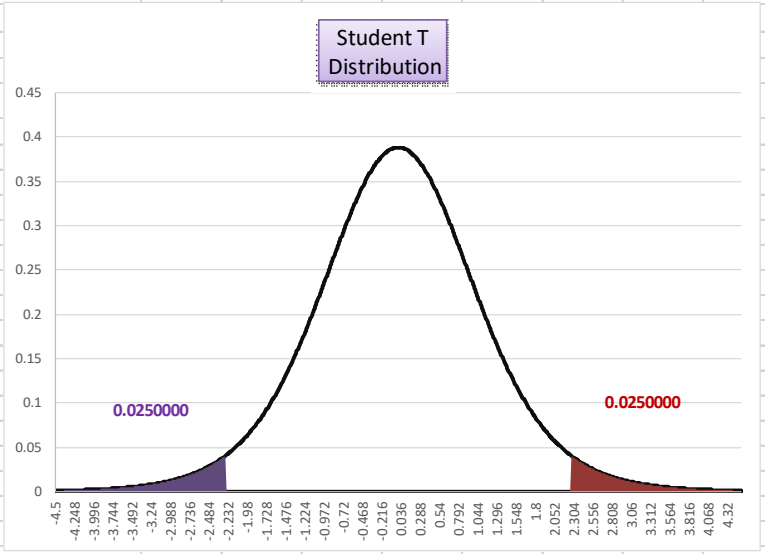
Could use Excel utility:

Input Information

Degrees of Freedom	9				
Low T Score		OR	Probability	0.025	Will Default to Data Value if Both Entered
					Note: Any entered Probabilities MUST be between 0 and 1, EXCLUSIVE
Upper T Score		OR	Probability	0.975	Will Default to Data Value if Both Entered

Output Results

Area	Data Value	Probability
Below	-2.262	0.0250
Above	2.262	0.0250
Between	-2.262 & 2.262	0.9500



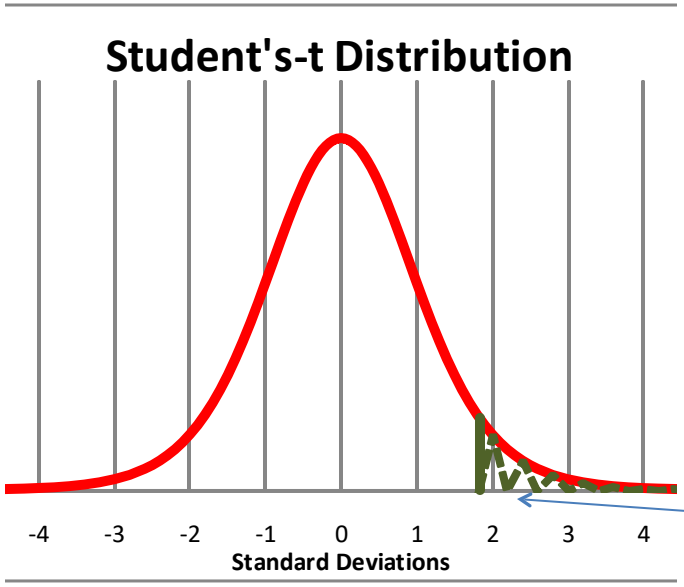
Inference on μ

σ Unknown

In many inference situations, the population standard deviation, σ, is not known.
Consequently, it is necessary to estimate it using the sample data.
Generally, it is estimated with the sample standard deviation, S.

Example: In order to test a new teaching methodology, a school district randomly selects 10 classes in the district to experience the new approach. A standardized AP test for the subject is given as the final exam to these classes. The national average score on this test is 500, and scores are normally distributed. Data we have is the average AP test result for each of the 10 classes.

Question: Will the mean AP scores for students receiving instruction by the new method be above the national average?



Research Hypothesis: H_1: μ > 500

Null Hypothesis: H_0: μ = 500

Test Statistic: $T = (\bar{X} - 500)/[S/\sqrt{10}]$

Null Distribution: $T \sim t_{(9)}$

Decision Rule: Reject H_0 if $T > 1.833$ ($\alpha = 0.05$)

With $\bar{X} = 518$ and $S = 22.2$,

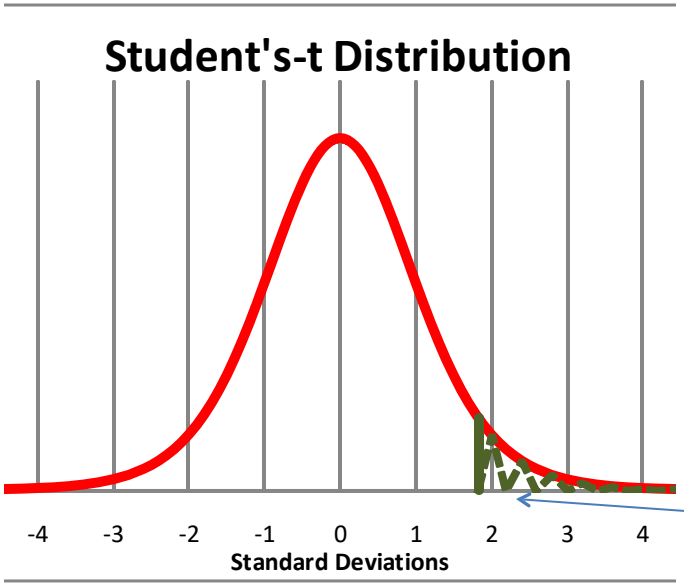
$T = 2.56 > 1.833$, so Reject H_0 and Conclude Average AP test scores higher with new method.

Confidence Intervals for μ

σ Unknown

Example: In order to test a new teaching methodology, a school district randomly selects 10 classes in the district to experience the new approach. A standardized AP test for the subject is given as the final exam to these classes. The national average score on this test is 500, and scores are normally distributed. Data we have is the average AP test result for each of the 10 classes.

Question: Will the mean AP scores for students receiving instruction by the new method be above the national average?



Research Hypothesis: H_1: μ > 500

Null Hypothesis: H_0: μ = 500

Test Statistic: $T = (\bar{X} - 500)/[S/\sqrt{10}]$

Null Distribution: $T \sim t_{(9)}$

Decision Rule: Reject H_0 if T > 1.833 (α=0.05)

With $\bar{X}$ = 518 and S = 22.2,

T = 2.56 > 1.833, so Reject H_0 and Conclude Average AP test scores higher with new method.

A 95% Confidence Interval for μ_{New} is given by

Point Estimate = $\bar{X}$

Estimated Standard Error = $S/\sqrt{n} = 22.2/\sqrt{10}$

518 +/- 2.262*7.02 or (502 to 534)

Multiplier = t(9, 0.975)

NOTE: Does not include 500

Inference on σ

Note that σ is also a population parameter, is usually unknown, and also often needs to be estimated from available sample data. The estimator we have been using is the sample standard deviation, S. Like $\bar{X}$, S is a sample statistic and also has a sampling distribution.

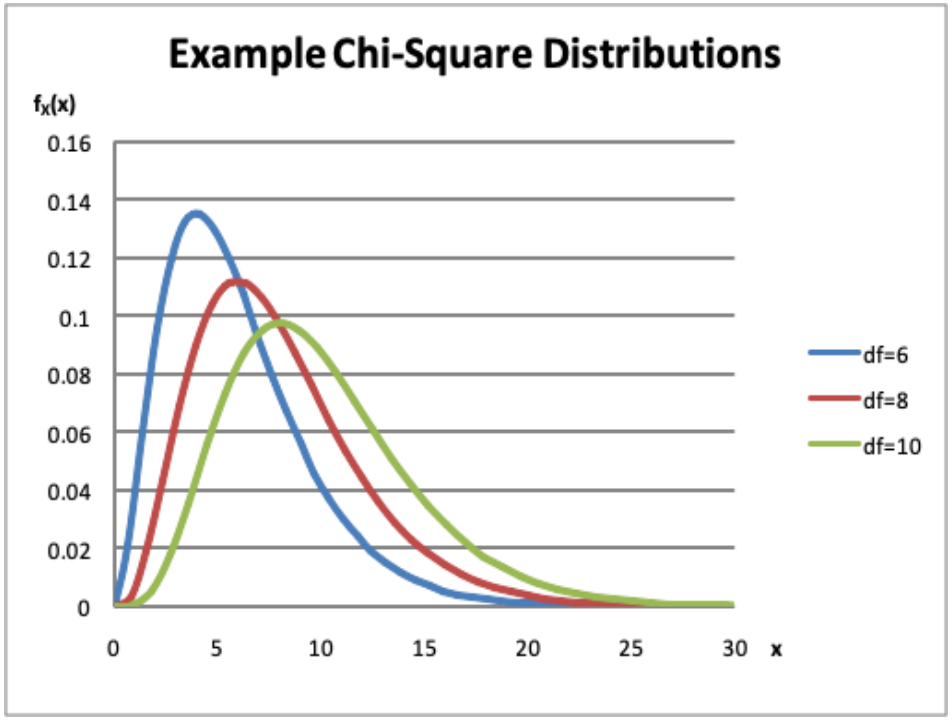
While the CLT helps us with the sampling distribution of $\bar{X}$ statistics, acquiring the sampling distribution of S is more involved. It can be shown that when the distribution of X is normal that

$$(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)},$$

where $\chi^2_{(n-1)}$, is notation for a ***Chi-Square Distribution*** with n-1 degrees of freedom.

Chi-square distributions do not, in general have the nice symmetrical bell-shape of a normal distribution (although as n-1 gets large, they do tend to take on this shape).

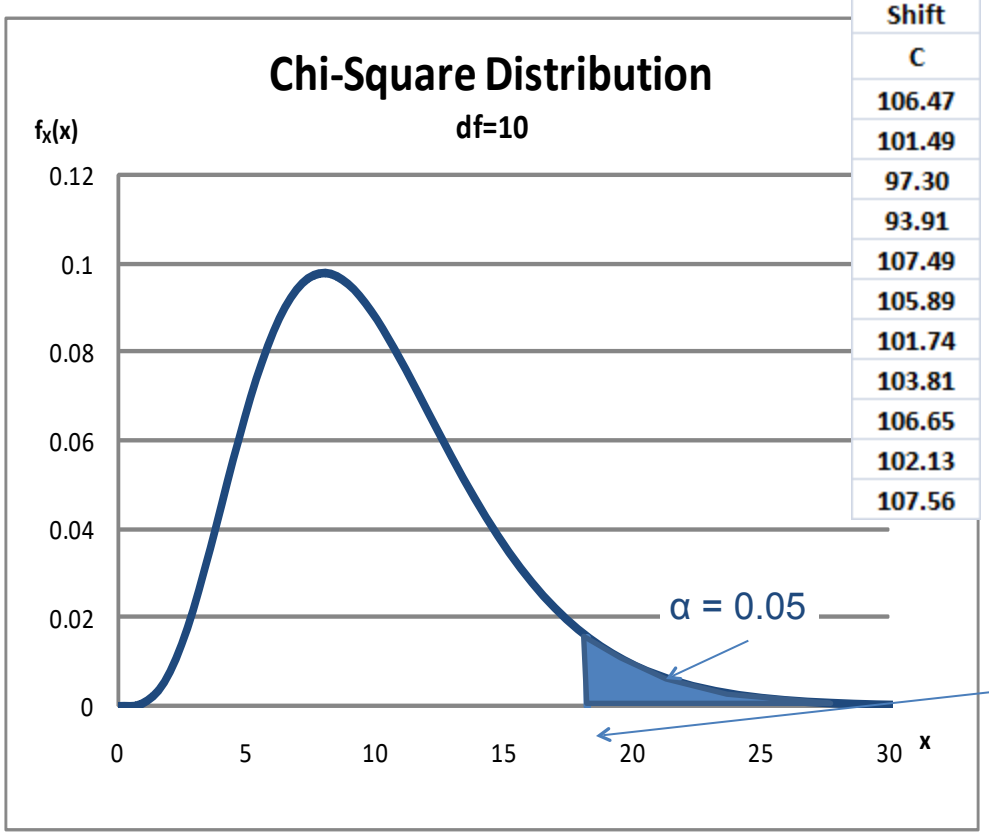
However, this distribution can be utilized to conduct tests of hypotheses for σ, as well as construct confidence intervals for this population parameter.



Chi-square probabilities can be found with the Excel command =CHIDIST(x, n-1) which returns P[>x], or the **R** command pchisq(x, n-1) = P[<x] and Chi Square percentiles can be found with the Excel command =CHIINV(p, n-1) which returns x | P[>x] = p, or the **R** command qchisq(p, n-1) returning x | P[<x] = p. In each case, n-1 is the related degrees of freedom.

Inference on σ

In Workshop 1, it was discovered that shift C appeared to have more variable gate width results than the other shifts. If expected gate width standard deviation for the routinely obtained results is 1.5nm, then assessing if the results obtained on shift C are more variable than this is a simple hypothesis test:



Shift
C
106.47
101.49
97.30
93.91
107.49
105.89
101.74
103.81
106.65
102.13
107.56

Research Hypothesis: H_1: σ > 1.5nm

Null Hypothesis: H_0: σ = 1.5nm

Test Statistic: $X^2 = (n-1)S^2/\sigma^2$

Null Distribution: $\chi^2_{(n-1)}$

Decision Rule: With α = .05,

Reject H_0 if $X^2 > \chi^2_{(n-1,\ 1-\alpha)}$

Decision: n = 11, S = 4.14, so

$X^2 = 95.23 > \chi^2_{(10,\ 0.95)} = 18.31$

Conclusion: Variation observed on Shift C is far in excess of the expected amount (p-Value ≈ $5*10^{-16}$)

OK, so variation on C Shift is greater than expected, but by how much?

Inference on σ

To address the "how much?" question, we again resort to a confidence interval.

Since $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$ when σ^2 is the respective population variance:

$$P[\chi^2_{(n-1,\alpha/2)} < (n-1)S^2/\sigma^2 < \chi^2_{(n-1,1-\alpha/2)}] = 1 - \alpha$$

$$P[1/\chi^2_{(n-1,1-\alpha/2)} < \sigma^2/[(n-1)S^2] < 1/\chi^2_{(n-1,\alpha/2)}] = 1 - \alpha$$

$$P[(n-1)S^2/\chi^2_{(n-1,1-\alpha/2)} < \sigma^2 < (n-1)S^2/\chi^2_{(n-1,\alpha/2)}] = 1 - \alpha$$

$$P[S^*\sqrt{(n-1)/\chi 2(n-1,1-\alpha/2)} < \sigma < S^*\sqrt{(n-1)/\chi 2(n-1,\alpha/2)}] = 1 - \alpha$$

Hence, a (1-α)% confidence interval for a population standard deviation is of the form:

[Best Point Estimate * m(α/2)] to [Best Point Estimate * M(α/2)],

where m(α/2) < 1 and M(α/2) > 1.

For the C Shift data, to generate a 90% confidence interval for $\sigma_{C\ Shift}$, the limits would be:

$$S * \sqrt{(n-1)/\chi 2(n-1,0.95)} \quad \text{to} \quad S * \sqrt{(n-1)/\chi 2(n-1,0.05)}$$

$$4.14 * \sqrt{10/18.31} \quad \text{to} \quad 4.14 * \sqrt{10/3.94}$$

$$3.26 \quad \text{to} \quad 7.03$$

Note that this interval does not include the expected result of 1.5, and the range it covers is from just over 2X to just less than 5X this value. C Shift has much more variable results than would usually be expected, again consistent with the test of hypothesis, but more informative.