Fundamental Statistical Concepts and Methods Needed in a Test-and-Evaluator's Toolkit

Mark Kiemele
Air Academy Associates
mkiemele@airacad.com

ITEA 2010 Symposium
Glendale, AZ
13 September 2010

INTRODUCTIONS

- **Name**
- **Organization**
- **Job Title/Duties**
- **Experience in using statistical concepts in T&E**

AGENDA

- Some Key Terms and Concepts
- Sampling Distribution of the Mean
- Confidence Intervals
  - Estimating a population mean (continuous data)
  - Estimating a population proportion (binary data)
- Determining Sample Size
  - For estimating a population mean (continuous data)
  - For estimating a population proportion (binary data)
- Drawing Conclusions When Comparing Data Sets (Hypothesis Testing)
  - Comparing means
  - Comparing standard deviations
  - Simple Rules of Thumb
  - Formal tests
  - Controlling both the Alpha and Beta risks
  - Comparing proportions
  - Critical Thinking

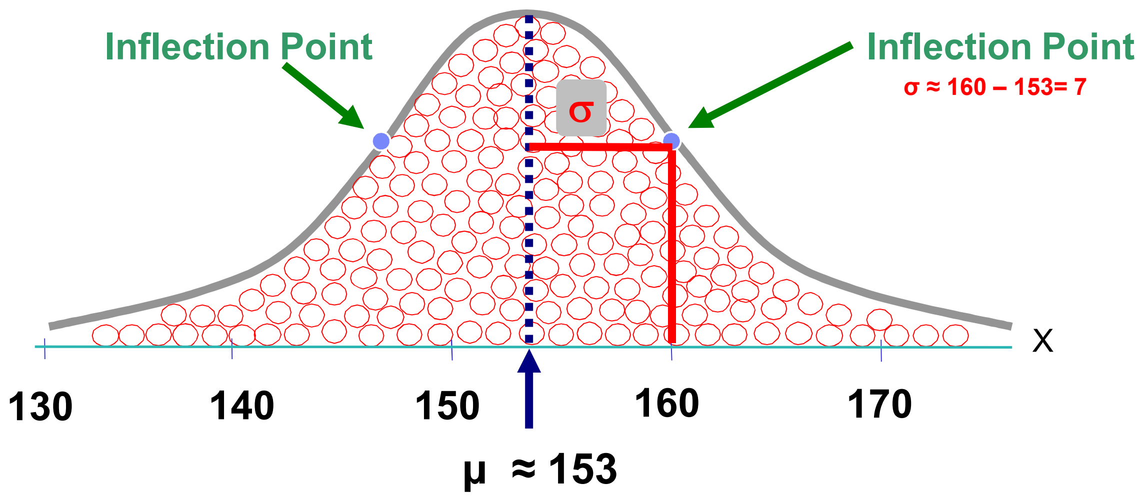
SOME KEY TERMS

- **Population (Parameters)-use Greek letters (μ, σ)**
- **Sample (Statistics)-use Latin letters ($\bar{x}$, s)**
- **Random Sampling**
- **Parent and Child Distributions**
- **Point estimate**
- **Interval estimate**
- **Confidence Interval**
- **- Upper confidence limit**
- **- Lower confidence limit**
- **Half-interval width**
- **Determining Sample Size**
- **Hypothesis Testing**
- **- Type I error (alpha risk)**
- **- Type II error (beta risk)**
- **Significance and Power**

GRAPHICAL MEANING OF μ and σ

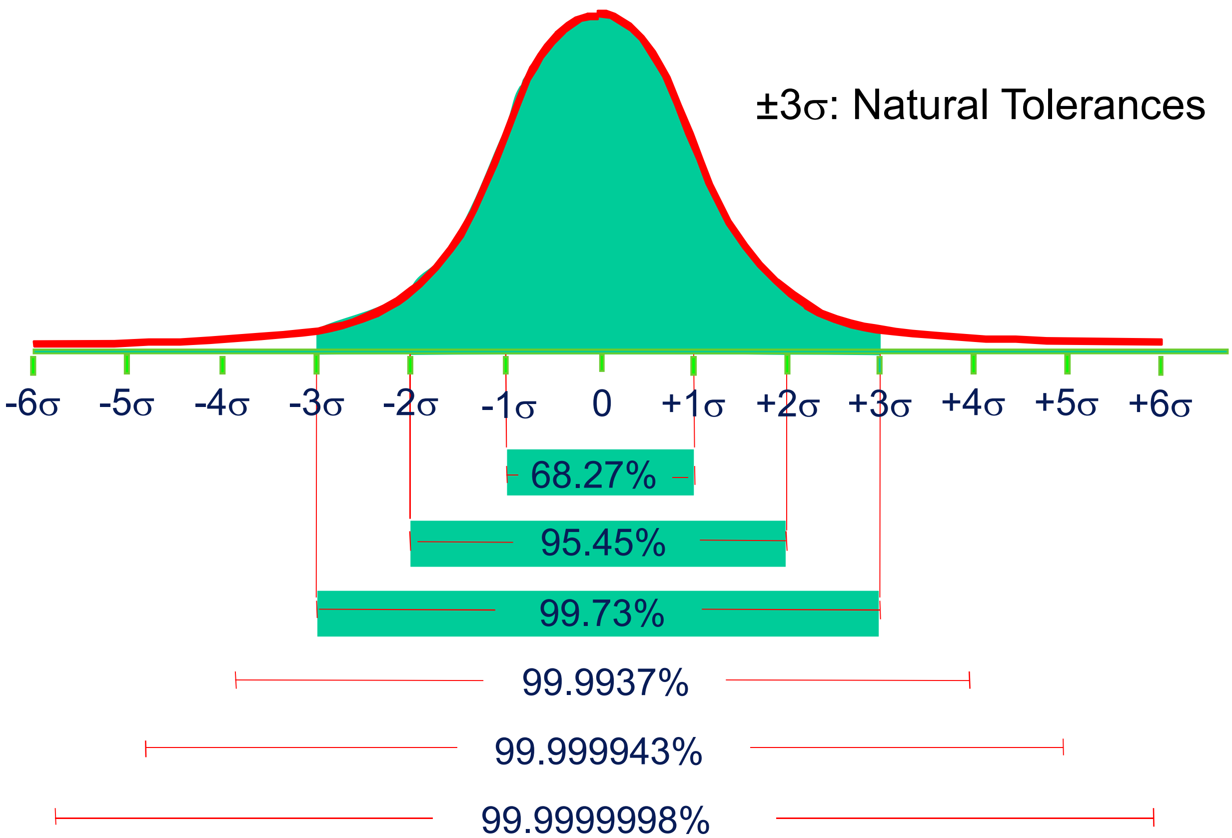
μ = Average = Mean = Balance Point

σ = Standard Deviation = Measure of Variation



σ ≈ average distance of points from the centerline

GRAPHICAL VIEW OF VARIATION



Typical Areas under the Normal Curve

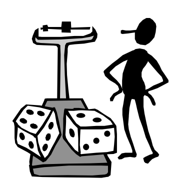
DATA GATHERING, SAMPLING, AND MEASUREMENTS

- Data gathered should be representative of the process being studied.

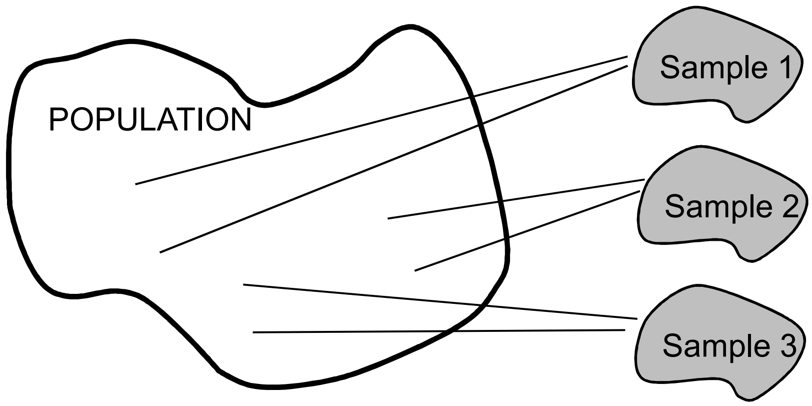
- Poor approach to sampling
  - convenience sampling

- Better approaches to sampling
  - random sampling
  - systematic sampling

- Before gathering the data, think about the measurement system being used to generate the data
  - is it accurate? precise? repeatable?

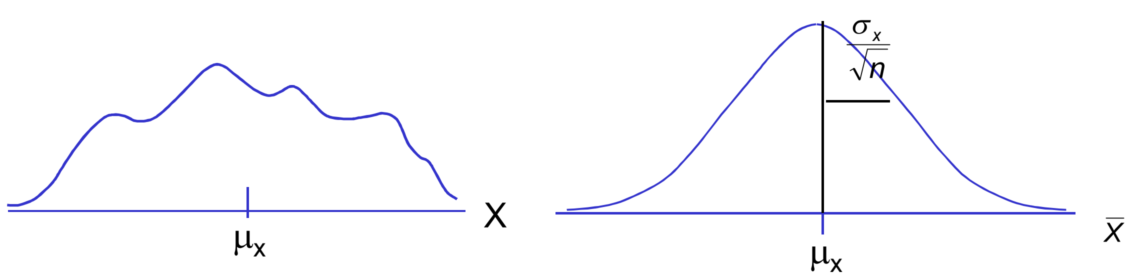


PARENT AND CHILD DISTRIBUTIONS



- We refer to the distribution of individual values from the population as the "**parent**" distribution. The population may be normal, but it could also follow some other distribution.

- We refer to the distribution of sample averages as the "**child**" distribution.

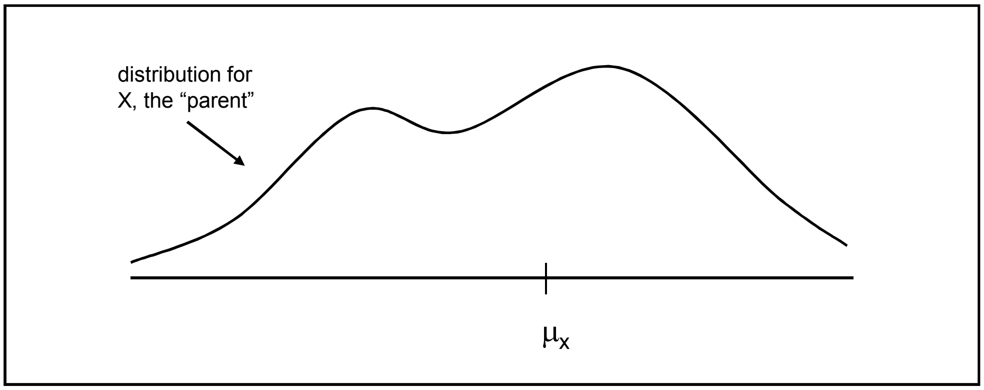


- There is an important theorem in statistics, called the Central Limit Theorem (CLT), which says that no matter what the parent distribution looks like, as long as the sample size is big enough, the child distribution (of sample averages) will be approximately normal.

SAMPLING DISTRIBUTION OF THE MEAN

DEFINITION:

The **sampling distribution of the mean (sometimes called the $\bar{X}$ or "child" distribution)** is the distribution of all means (or averages) obtained from all possible samples of a fixed size (say n) taken from some "parent" population.



NOTATION:

μ_x	=	center (or mean) of the "parent" or X distribution
σ_x	=	standard deviation of the "parent" or X distribution
$\mu_{\bar{x}}$	=	center (or mean) of the "child" or $\bar{X}$ distribution
$\sigma_{\bar{x}}$	=	standard deviation of the "child" or $\bar{X}$ distribution

IMPORTANT RESULTS:

(1) The center of the "child" ($\bar{X}$) distribution is the same as the center of the "parent" (X) distribution.

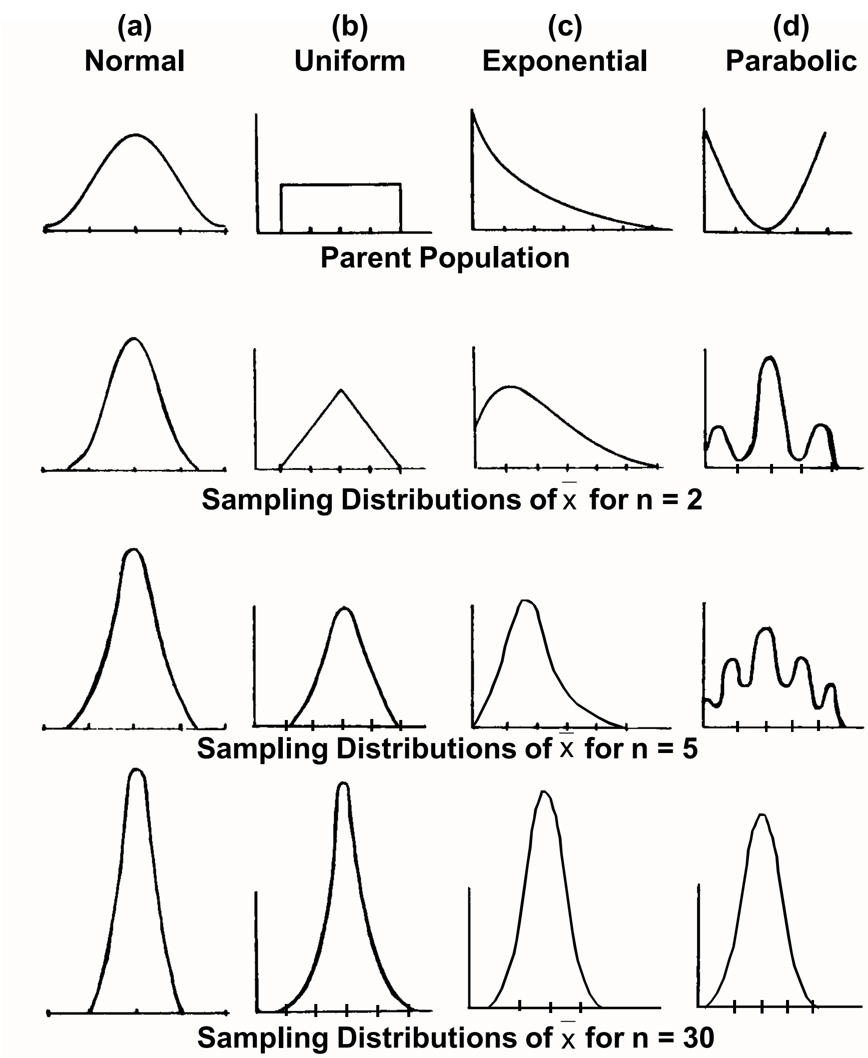
$$\mu_{\bar{x}} = \mu_x$$

(2) The standard deviation of the "child" ($\bar{X}$) distribution is smaller than the standard deviation of the "parent" (X) distribution.

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

CENTRAL LIMIT THEOREM

For almost all populations, the sampling distribution of the mean can be approximated closely by a normal distribution, provided the sample size is sufficiently large.



Sampling Distributions of $\overline{X}$ for Various Sample Sizes

ILLUSTRATING PARENT AND CHILD DISTRIBUTIONS EXERCISE (Optional)

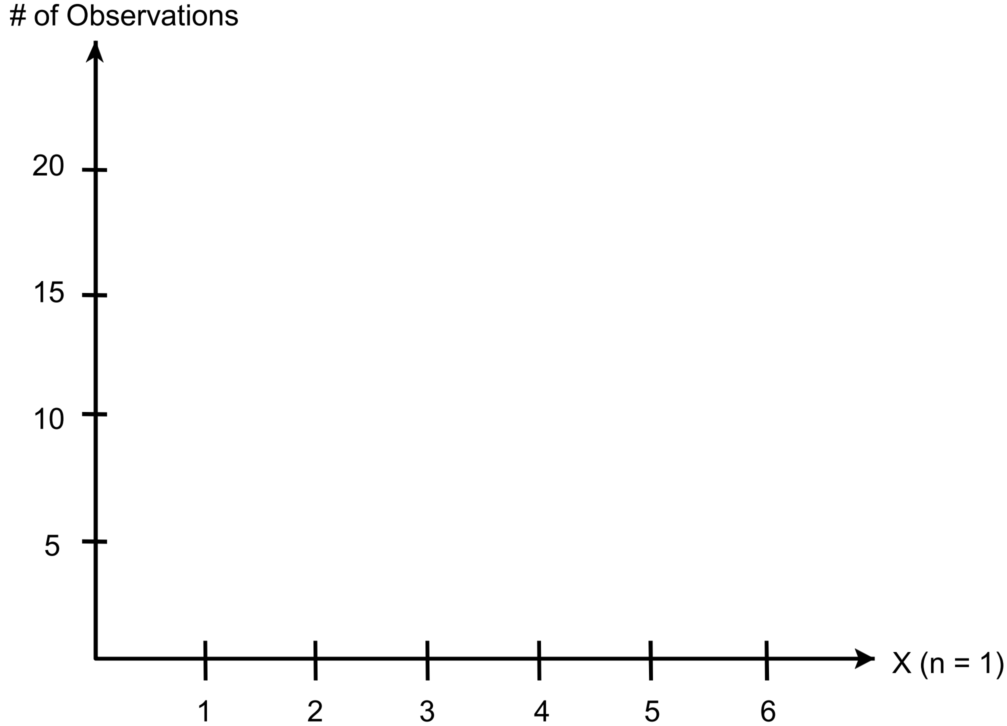
Parent (single roll of 1 die)

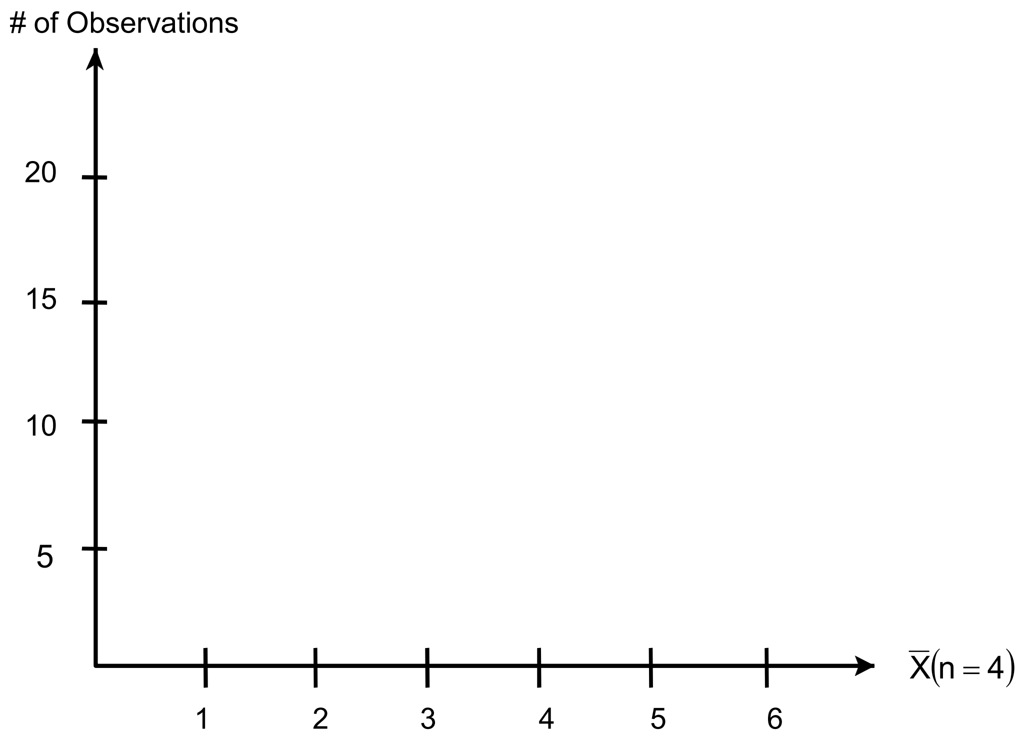
Number Rolled	# of Observations
1	
2	
3	
4	
5	
6	

Child (average of 4 rolls)

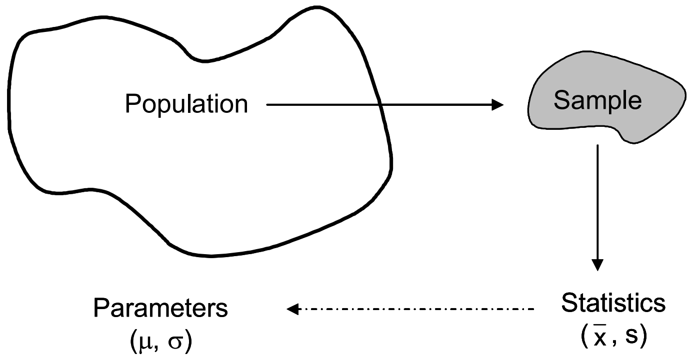
Total On Dice	Average	# of Observations
4	1.00	
5	1.25	
6	1.50	
7	1.75	
8	2.00	
9	2.25	
10	2.50	
11	2.75	
12	3.00	
13	3.25	
14	3.50	
15	3.75	
16	4.00	
17	4.25	
18	4.50	
19	4.75	
20	5.00	
21	5.25	
22	5.50	
23	5.75	
24	6.00	

ILLUSTRATING PARENT AND CHILD DISTRIBUTIONS (Optional)





SAMPLING AND CONFIDENCE INTERVALS



- Suppose we measure the time it takes for a customer service representative to answer a call.

- We use the data from our sample to estimate the average call time. How good is our estimate?

- Confidence intervals provide error bounds or an estimate of uncertainty for a population parameter based on our sample data.

Confidence Interval =
Point Estimate ± Margin of Error

CONFIDENCE INTERVAL FOR POPULATION MEAN, μ (Continuous Data)

$$\begin{pmatrix} U \\ L \end{pmatrix} = \bar{x} \pm Z\left(\frac{s}{\sqrt{n}}\right)$$

Where

U	**=**	**upper confidence limit**
L	**=**	**lower confidence limit**
$\bar{X}$	**=**	**sample average**
Z	**=**	**2 (for 95% confidence) or**
		3 (for 99% confidence)
s	**=**	**sample standard deviation**
n	**=**	**sample size**

Computational Template for Confidence Limits

EXAMPLE:

Over the course of a week, we randomly select and time 16 customer service calls. We find that the average time is 15.6 minutes with a standard deviation of 2.1 minutes. What is a 99% confidence interval for the true average service time?

USING SPC XL FOR CONFIDENCE INTERVALS FOR POPULATION MEAN, μ (Continuous Data)

- SPC XL will produce a more exact interval. Rather than using Z = 2 or 3 from our 68/95/99 rule of thumb, it uses the exact values from a t-distribution (similar to Z, but adjusted slightly to account for the fact that we're using a small sample to estimate σ)

- SPC XL > Analysis Tools > Confidence Interval > Normal

Normal Confidence Interval (Mean)	
User defined parameters	
Sample Size (n)	16
Sample Avg	15.6
Sample Standard Dev	2.1
Confidence Level	99.00%
Confidence Interval	
Lower Limit	Upper Limit
14.05296861	17.14703139

- **Exercise:** Suppose we sampled 64 calls and got the same average time and sample standard deviation. How does this affect our confidence interval?

- **Exercise:** Suppose we're studying the average cycle time to pay invoices. We randomly sample 30 invoices, and find the average is 7.8 days with a standard deviation of 1.4 days. Using SPC XL, construct a 95% confidence interval for the true average cycle time for paying these invoices.

USING SPC XL TO DETERMINE SAMPLE SIZE FOR POPULATION MEAN, μ (Continuous Data)

- Confidence interval calculation can be worked backwards to determine an appropriate sample size

- Suppose in the previous example, we wanted to estimate the true average service call time to within ±1 minutes with 99% confidence

- Steps to determine sample size:

    1. Decide on the level of confidence you want, typically 95% or 99% (in our example, 99%)

    2. Specify the desired width of the confidence interval (Upper bound – Lower bound = 2h, where h = half interval width) (in our example, h = 1)

    3. Find an approximation for the population standard deviation from historical data, small pre-sample, etc. (in our example, use the previous sample standard deviation of 2.1 minutes)

USING SPC XL TO DETERMINE SAMPLE SIZE FOR POPULATION MEAN, μ (Continuous Data)

Select <u>SPC XL > Analysis Tools > Sample Size > Normal</u>

Sample Size to Estimate the Mean of a Normal Distribution	
User defined parameters	
Estimated Standard Dev	2.1
Half Interval Width	1
Confidence Level	99.00%
Results	
Estimated Sample Size (n)	29

CONFIDENCE INTERVAL FOR POPULATION PROPORTION, π (Binary Data)

$$\begin{pmatrix} U \\ L \end{pmatrix} = p \pm Z\sqrt{\frac{pq}{n}}$$

Where

U	**=**	**upper confidence limit**
L	**=**	**lower confidence limit**
p	**=**	**proportion of "defectives" (or category of interest) in the sample**
q	**=**	**1 – p (q is the proportion of "non-defectives")**
Z	**=**	**2 (for 95% confidence) or 3 (for 99% confidence)**
n	**=**	**sample size**

Computational Template for Confidence Limits

EXAMPLE:

You work in a finance office and are in charge of processing travel vouchers submitted by several different organizations. You sample 100 vouchers and find 8 to have discrepancies or errors. Find a 95% confidence interval for the true but unknown proportion of vouchers containing errors.

USING SPC XL FOR CONFIDENCE INTERVALS FOR POPULATION PROPORTION, π (Binary Data)

- Again, we can use SPC XL to produce a more exact answer.

- Select SPC XL > Analysis Tools > Confidence Interval > Proportion (Binomial)

Binomial Confidence Interval (Proportion)		
User defined parameters		
Sample Size (n)	100	
Number Defective(x)	8	
Confidence Level	95.00%	
	Confidence Interval	
Lower Limit	< p <	Upper Limit
0.035171509	0.08	0.151557446

- **Exercise:** Suppose in the previous example we had sampled 1,000 travel vouchers and found 80 to have discrepancies or errors. Find a 95% confidence interval for the true but unknown proportion of vouchers containing error.

EXERCISE

Using the data from <u>your</u> sample of M&M's, find the 95% confidence limits for the true (but unknown) proportion of M&M's that Mars Corporation produces for each color.

	Brown	Yellow	Red	Orange	Blue	Green
X						
n						
$p = \frac{X}{n}$						
Lower Bound						
Upper Bound						

What do you notice about the width of your confidence intervals?

If you wanted to estimate the true percentage of blue M&M's to within ± 8% with 95% confidence, what sample size would be required? (note: you can use the proportion (p) of blue above for your estimated proportion of blue M&M's)

DETERMINING SAMPLE SIZE FOR PROPORTIONS (Binary Data)

- Confidence interval calculation can be worked backwards to determine an appropriate sample size

- Suppose in the previous example, we wanted to estimate the true proportion of travel vouchers with discrepancies to within ± 1% with 95% confidence. Also, assume that the historical proportion of vouchers with discrepancies is no more than 2%.

- Steps to determine sample size:

  1. Decide on the level of confidence you want, typically 95% or 99% (in our example, 95%)

  2. Specify the desired width of the confidence interval (Upper bound – Lower bound = 2h, where h = half interval width) (in our example, h = .01 or 1%)

  3. Find an approximation for the proportion of interest, p. It can come from historical data. If no estimates are available, then use p=.50 to provide a worst case (conservative) sample size estimate. (in our example, use 2% (.02))

DETERMINING SAMPLE SIZE FOR PROPORTIONS (Binary Data) (continued)

Select <u>SPC XL > Analysis Tools > Sample Size > Binomial Conf. Interval (proportion)</u>

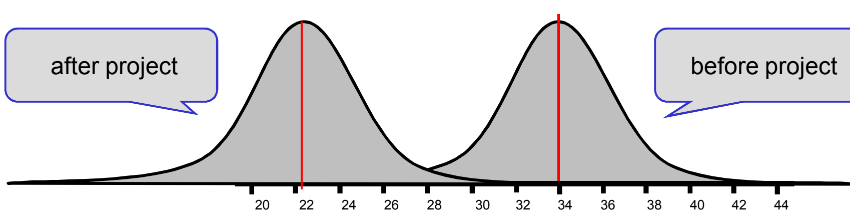
Binomial Sample Size	
User defined parameters	
Proportion defectives (p)	0.02
Half Interval Width	0.01
Confidence Level	95.00%
Results	
Estimated Sample Size (n)	753

HYPOTHESIS TESTING

- A method for looking at data and comparing results
  - Method A vs. Method B
  - Material 1 vs. Material 2
  - Before vs. After Project results

- Helps us make good decisions and not get fooled by random variation:
  - "Is a difference we see REAL, or is it just random variation and no real difference exists at all?"

- We set up 2 hypotheses
  - H_0 is called the null hypothesis (no change, no difference)
  - H_1 is called the alternate hypothesis
  - Example: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$



  - Based on the data we collect, we must decide in favor of either H_0 or H_1. Which does the evidence support?

NATURE OF HYPOTHESIS TESTING

H_0: Defendant is Innocent

H_1: Defendant is Guilty

Since verdicts are arrived at with less than 100% certainty, either conclusion has some probability of error. Consider the following table.

		True State of Nature	
		H_0	H_1
Conclusion Drawn	H_0	Conclusion is Correct	Conclusion results in a Type II error
	H_1	Conclusion results in a Type I error	Conclusion is Correct

Type I or II Error Occurs if Conclusion Not Correct

The probability of committing a Type I error is defined as α ($0 \leq \alpha \leq 1$) and the probability of committing a Type II error is β ($0 \leq \beta \leq 1$). The most critical decision error is usually a Type I error.

2-SAMPLE HYPOTHESIS TEST

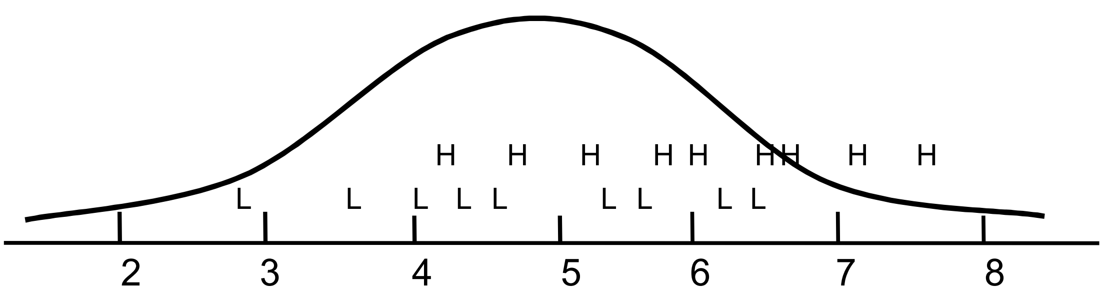
Strength Measurements												
Temp	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	$\bar{y}$	s^2	s
(L) 200°	2.8	3.6	6.1	4.2	5.2	4.0	6.3	5.5	4.5	4.6889	1.3761	1.17
(H) 300°	7.0	4.1	5.7	6.4	7.3	4.7	6.6	5.9	5.1	5.8667	1.1575	1.08

Composite Material Data (Low=200°, High=300°)

The graphical interpretation of the hypotheses to be tested are:

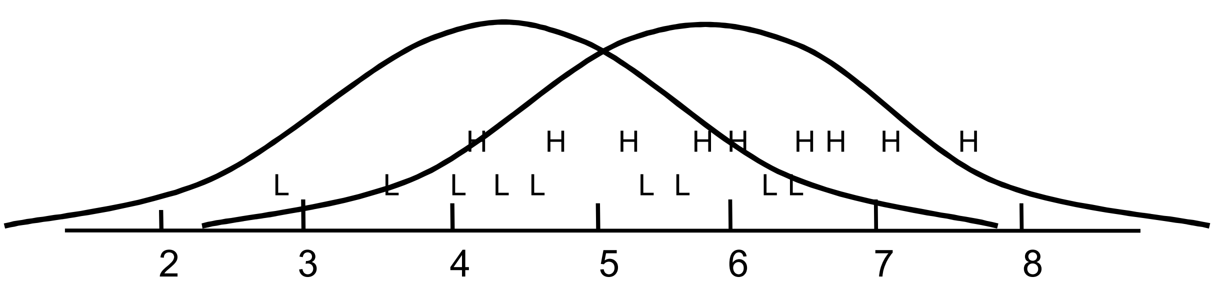
$H_0 : \mu_L = \mu_H$

$H_1 : \mu_L \neq \mu_H$



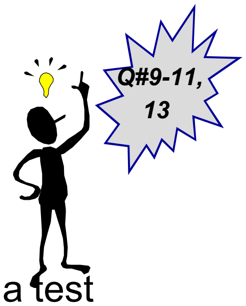
H_0: $\mu_L = \mu_H$

versus



H_1: $\mu_L \neq \mu_H$

RULE OF THUMB



Tukey <u>Quick Test</u> for Detecting a Significant Shift in Average

To determine if a significant shift in average has occurred, a test developed by John Tukey in 1959, (*Technometrics*, 1, 31-48) and popularized by Dorian Shainin, (*World Class Quality*, Bhote, K.R., 1988) is called the Tukey Quick Test or End Count Technique. To perform this test:

1. Arrange all of the data on a scale such that each of the two groups is represented by a different symbol. Refer to the previous example, on page 9-5, where L = low temperature and H = high temperature.

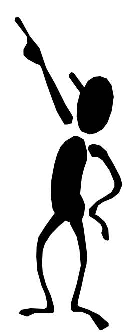
2. Starting from the left, *count* the number of similar symbols until an opposite symbol is encountered.

3. Likewise, starting from the right, *count* the number of similar symbols until an opposite symbol is encountered.

4. Summing the two counts yields the End Count.

Note: If the leftmost and rightmost symbols are the same, then the End Count is zero. The significance associated with a given End Count can be found using the following table:

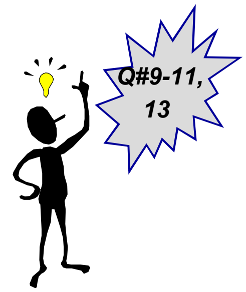
End Count	Significance	Confidence there exists a significant shift in average
6	.10	.90
7	.05	.95
10	.01	.99
13	.001	.999



The previous example illustrates an End Count of 7 (3 on left and 4 on right), giving approximately 95% confidence in concluding H_1: $\mu_L \neq \mu_H$.

FORMAL STATISTICAL TEST
(2-SAMPLE t-TEST)

For Detecting a Significant Shift in Average



Step 1: $H_0: \mu_L = \mu_H$ or $H_0: \mu_H - \mu_L = 0$

$H_1: \mu_L \neq \mu_H$ $H_1: \mu_H - \mu_L \neq 0$

Step 2: $\alpha = .05$

Step 3: $\bar{y}_L = 4.6889$ $\bar{y}_H = 5.8667$

$s_L^2 = 1.3761$ $s_H^2 = 1.1575$

$n_L = 9$ $n_H = 9$

$$t_0 = \frac{\bar{y}_H - \bar{y}_L}{s_p\sqrt{\frac{1}{n_H} + \frac{1}{n_L}}} \quad , \quad \text{where} \quad s_p = \sqrt{\frac{(n_H - 1)s_H^2 + (n_L - 1)s_L^2}{n_H + n_L - 2}}$$

represents the pooled standard deviation of the two samples. It is just the square root of a weighted average of the two variances.

Therefore,

$$t_0 = \frac{5.8667 - 4.6889}{1.126\sqrt{\frac{1}{9} + \frac{1}{9}}} = 2.22,$$

$$\text{where} \quad s_p = \sqrt{\frac{8(1.1575) + 8(1.3761)}{9 + 9 - 2}} = 1.126$$

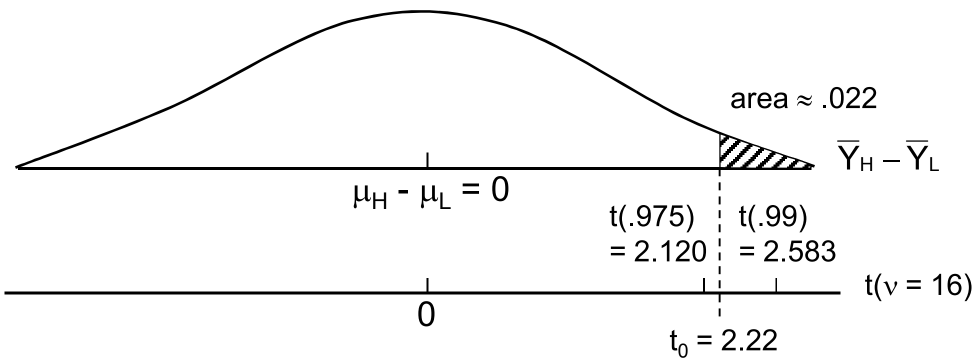
The standardized test criteria will be from a t distribution with:

$$\nu = (n_H - 1) + (n_L - 1) = 8 + 8 = 16 \text{ degrees of freedom.}$$

FORMAL STATISTICAL TEST
(2-SAMPLE t-TEST) (cont.)

For Detecting a Significant Shift in Average

Step 4:



Step 5: Based on the graphic in Step 4, P ≈ 2(.022) = .044.

Step 6: Since [P ≈ .044] < [α = .05], we reject H_0. In this case, we can be (1 - P)100% = 95.6% confident in the H_1 conclusion, i.e., the change in temperature shifts the average strength.

SUMMARY OF A TWO-SAMPLE TEST

The following table summarizes a two-sample t-test:

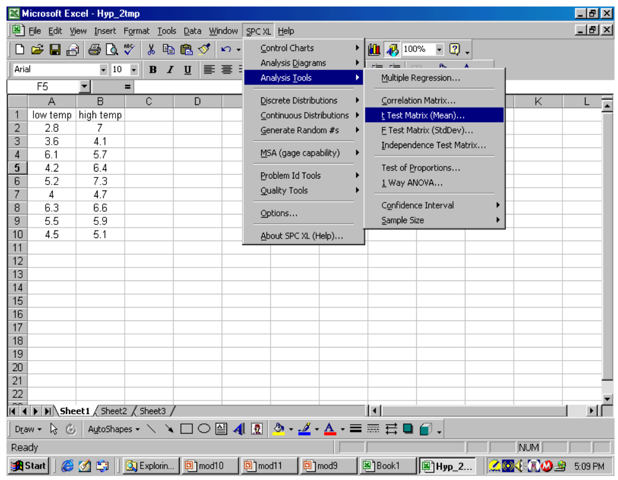
Step 1:	State the hypotheses to be tested. Two-Tail: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$ Upper-Tail: $H_0: \mu_1 \leq \mu_2$, $H_1: \mu_1 > \mu_2$ Lower-Tail: $H_0: \mu_1 \geq \mu_2$, $H_1: \mu_1 < \mu_2$
Step 2:	Select a pre-planned α = .05, .01, or .001.
Step 3:	Obtain information from two samples. For example, n_1, $\bar{y}_1$, s_1^2; n_2, $\bar{y}_2$, s_2^2; and compute the test statistic. $t_o = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
Step 4:	Use percentiles of the t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom to estimate the area in the tail beyond $\lvert t_0 \rvert$. Note: If $(n_1 + n_2) \geq 30$, use the Z distribution.
Step 5:	For a two-tail test, P = 2 times the area in Step 4. For a one-tail test, P = area in Step 4.
Step 6:	If $P < \alpha$, conclude H_1 with (1 - P)100% confidence. If $P \geq \alpha$, fail to reject H_0.

Summary of a Two-Sample t-Test

t-TEST USING SPC XL

- SPC XL will give us p-values

- <u>SPC XL > Analysis Tools > t Test matrix (mean)</u>



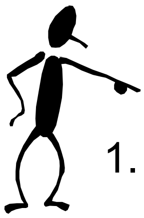
The results below represent the p-values from a 2 sample t-test. This means that the probability of falsely concluding the alternative hypothesis is the value shown (where the alternate hypothesis is that the means are not equal). Another way of interpreting this result is that you can have (1-pvalue)*100% confidence that the means are not equal.

t Test Analysis (Mean)
P-value = 0.041

- Rule of Thumb:
  - If p-value < .05, highly significant difference
  - If .05 < p-value < .10, moderately significant difference
  - (1 – p-value) • 100% is our percent confidence that there is a significant difference.

RULES OF THUMB

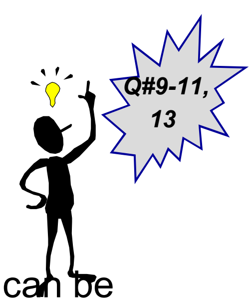
Quick Test for Detecting a Significant Shift in Standard Deviation



1. If $\frac{s_{max}}{s_{min}} > 2.72$ then there is a significant shift in standard deviation.

A “gray zone” occurs when $1.5 \leq \frac{s_{max}}{s_{min}} \leq 2.72$. If this is the case, then ROT #2 should be applied.

2. If $\frac{s^2_{max}}{s^2_{min}} \sqrt{\frac{n_1 + n_2}{2}} > 10,$ then there is a significant shift in standard deviation. This is a more discerning test than ROT #1 above since it involves the sample sizes, n_1 and n_2. However, neither n_1 nor n_2 should be greater than 60, and the smaller n should not be less than 70% of the larger n.



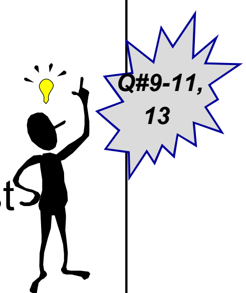
Note: If the sample sizes are the same, s (standard deviation) can be replaced by R (range).

FORMAL STATISTICAL TEST (F test) For Detecting a Significant Shift in Standard Deviation

Step 1: State the hypotheses to be tested.

Two-Tail	Upper-Tail	Lower-Tail
$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 \leq \sigma_2^2$	$H_0: \sigma_1^2 \geq \sigma_2^2$
$H_1: \sigma_1^2 \neq \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$	$H_1: \sigma_1^2 < \sigma_2^2$

Step 2: Select a pre-planned α = .05, .01, or .001.



Step 3: Obtain information from two samples. For example, n_1, s_1^2; n_2, s_2^2; and compute the test statistic.

$$F_0 = \frac{s_1^2}{s_2^2}$$

Step 4: Use percentiles of the F distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom to estimate the area in the tail. (see Appendix F in the Basic Statistics text for details on the F distribution)

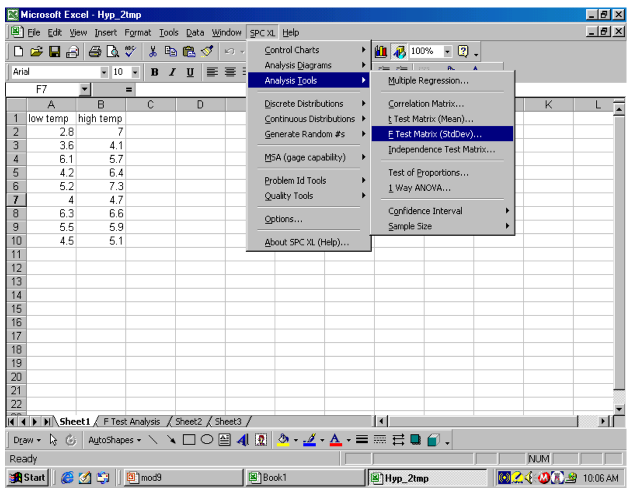
Step 5: For a two-tail test, P = 2 times the area in Step 4. For a one-tail test, P = area in Step 4.

Step 6: If $P < \alpha$, conclude H_1 with (1 - P)100% confidence. If $P \geq \alpha$, fail to reject H_0.

Summary of a Two-Sample F-Test For $\sigma_1^2 = \sigma_2^2$

F-TEST USING SPC XL

- SPC XL will give us p-values
- SPC XL > Analysis Tools > F Test matrix (StdDev)



The results below represent the p-values from a 2 sample F-test. This means the probability of falsely concluding the alternative hypothesis is the value shown (where the alternate hypothesis is that the variances are NOT equal). Another way of interpreting this result is that you can have (1-pvalue)*100% confidence that the variances are not equal.

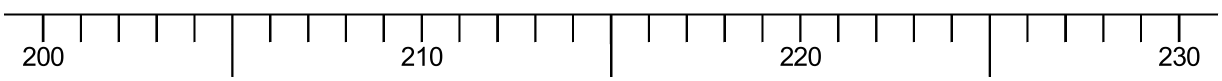
F Test Analysis (Std Dev)
P-value = 0.813

- Rule of Thumb:
  - If p-value < .05, highly significant difference
  - If .05 < p-value < .10, moderately significant difference
  - (1 – p-value) • 100% is our percent confidence that there is a significant difference.

EXERCISE

Using the Rules of Thumb on the tabulated data below, conduct tests for different averages and different standard deviations.

Factor A	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$\overline{Y}$	S
Lo	201	209	215	221	211	213	217	205	218	208	203	214	212	215	211.6	5.8
Hi	218	225	217	222	223	220	222	216	221	224	224	221	220	219	220.9	2.7

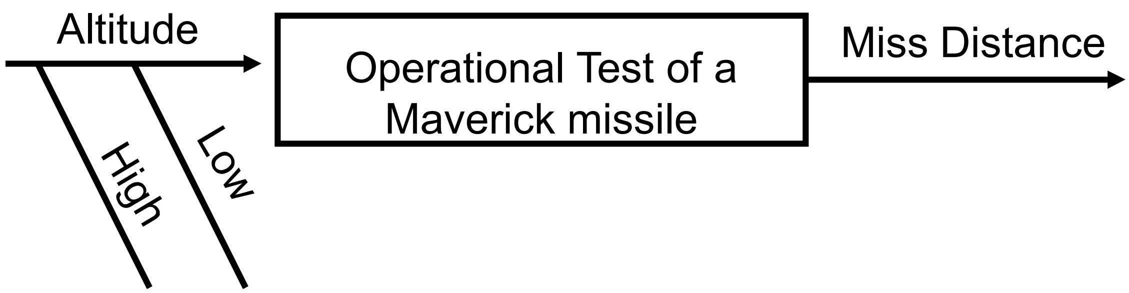


EXERCISE

Using SPC XL > Analysis Tools > t-Test (and F-Test), conduct the formal t-test and F-test for the two samples on the previous page to find the corresponding p-values. Compare these results with the results you obtained using the Rules of Thumb.

CONTROLLING BOTH ALPHA AND BETA RISKS

Suppose that in an operational test of a Maverick air-to-ground missile, we test at two different altitudes. The measure of performance is miss distance.



Test Requirements:

1. Alpha = .05 (Confidence = 95%)

2. Beta = .20 (Power = 80%)

3. Be able to detect an average miss distance of at least 10 feet between the two altitudes

Known: Standard Deviation of Miss Distance for each altitude is 20 feet (estimated from previous testing or simulation).

What is the required sample size to satisfy these requirements?

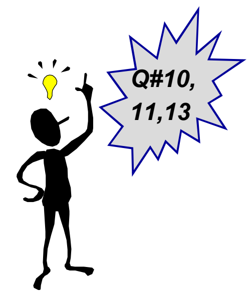
HYPOTHESIS TEST FOR THE EQUALITY OF TWO PROPORTIONS

1. H_0: $\pi_1 = \pi_2$
   H_1: $\pi_1 \neq \pi_2$

2. Select α = .05, .01, or .001

3. Compute the test statistic Z_0 as:

$$Z_0 = \frac{|p_1 - p_2|}{\sqrt{\frac{(x_1 + x_2)}{(n_1 + n_2)}\left(1 - \frac{x_1 + x_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



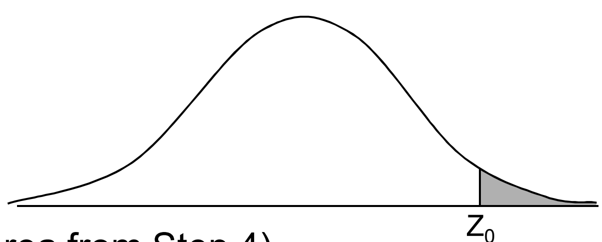
where

n_1 = size of sample #1
n_2 = size of sample #2
x_1 = number of elements in sample #1 in category of interest
x_2 = number of elements in sample #2 in category of interest

$$p_1 = \frac{x_1}{n_1}$$

$$p_2 = \frac{x_2}{n_2}$$

4. Find the area in the tail beyond Z_0, as shown here:



5. Let P = 2•(Area from Step 4)

6. If P < α, conclude H_1 with (1 - P)100% confidence
   If P $\geq \alpha$, fail to reject H_0

EXAMPLE

Two different radar systems are tested to determine their capability to detect a particular target under specific conditions.   Radar System 1 failed to detect the target in 5 of 60 tests, while Radar System 2 failed to detect the target in 11 of 65 tests. Are the proportion of detection failures produced by the two radar systems significantly different? Use α = .05.

Hint: use SPC XL > Analysis Tools > Test for Proportions

EXAMPLE OF CRITICAL THINKING

- There are two kinds of treatment for kidney stones.

- It is known that Treatment B (83%) is more effective than Treatment A (78%), as shown in the following test of proportions that turns out to be significant at the p=.042 level. Sample sizes are equal and sufficiently large (n=600 for each treatment) to detect significance.

Test of Proportions	
User defined parameters	
Number of Successes for Trt A	468
Size of Sample #1 (n_1)	600
Number of Successes for Trt B	496
Size of Sample #2 (n_2)	600
Results	
Proportion Sample #1 (p_1)	0.78000
Proportion Sample #2 (p_2)	0.82667
p-value	*0.04200*

(1 – pValue)*100%
is your percent confidence that
the proportions are not equal

EXAMPLE OF CRITICAL THINKING (cont)

- **Suppose that you visit your physician after the advent of a kidney stone attack, and you are presented with two alternative treatments along with the data shown on the previous page. And your physician asks which procedure you would prefer.**

- **What are some of the questions you might ask to help you select the best treatment for you?**

EXAMPLE OF CRITICAL THINKING (cont)

- One such question might be: Are my kidney stones large or small and does the size of the stone impact the success rate of the two treatments?

- Your doctor searches his computer database for more information on kidney stone treatments.   He then informs you that Treatment A is better than Treatment B for small size kidney stones, and Treatment A is also better than Treatment B for large size kidney stones.

- Now you are confused. How can this be? Just a minute ago, your doctor told you that Treatment B was better than Treatment A and even showed you the data and test of proportions. Your doctor then quotes his computer database by saying that for small stones, Treatment A has a 93% success rate while Treatment B has a 87% success rate.   He goes on to state that for large stones, Treatment A has a 73% success rate while Treatment B has a 67% success rate.   Your doc is now admittedly confused as well. But fortunately, you have learned to think statistically, and you ask for the complete set of data, including all sample sizes.   This is shown on the next page.

EXAMPLE OF CRITICAL THINKING (cont)

	Small Stones	Large Stones		Total
Treatment A	140/150	328/450		468/600
	93%	73%		78%
Treatment B	402/460	94/140		496/600
	87%	67%		83%
Total	542/610	422/590		964/1200
	89%	72%		80%

Note the inequity in sample sizes between size of stone and the treatment. Treatment A was performed much more frequently on Large Stones, while Treatment B was performed much more frequently on Small Stones, for which the overall success rate is much better. In this case, Stone Size is a lurking variable which confounds the overall result. This phenomenon of percentage reversal is called Simpson's Paradox. This is just one more reason why we need DOE and why we need to look at all potential variables before we test.

For more information please contact

Kathi Swagerty or Mark Kiemele

Air Academy Associates, LLC

1650 Telstar Drive, Ste 110

Colorado Springs, CO 80920

Toll Free: (800) 748-1277 or (719) 531-0777

Facsimile: (719) 531-0778

Email: aaa@airacad.com

Website: www.airacad.com