

Statistics for Software Process Improvement

September 23, 2000 – last refined November 2, 2006

[Glenn Booker](#) (Drexel University)

Table of Contents

Introduction.....	1
Basic Concepts.....	1
Measurement Scales	1
Parametric and Dependent.....	2
Population versus Sample Measures.....	2
Basic Measurements	3
Common Errors	4
Hypothesis Testing Model	5
Required Inputs.....	5
The CANDOALL Model.....	5
Types of Statistical Analysis.....	6
Statistical Test Selection.....	8
Contingency Table.....	9
Terminology.....	9
General Definitions.....	9
Hypothesis Testing Terminology.....	10
Statistical Test Method Terminology	11
Reporting Statistical Significance.....	12
References.....	12
Statistical Q&A	13
Curve Fit Formulas for SPSS	15
Evaluating Statistical Test Results	17
Evaluate ‘t’ or ‘z’ Values.....	17
Evaluate Confidence Interval	18
Evaluate Significance Value.....	18

Introduction

This report summarizes basic statistical concepts, particularly as they apply to software development and measuring process improvement using SPSS. It is based in part on the tutorial “Statistical Orientation Support” from <http://www.hitcorp.com/SOS.htm> (link and domain no longer available). Vocabulary used in this report is defined in the section, Terminology.

Basic Concepts

Measurement Scales

Data may take many forms, which will determine the types of analysis possible on them. One consideration is the type of scale used to measure something.

There are four types of measurement scales, as shown in Table 1 in order of increasing usefulness. Notice that a plain text field doesn’t correspond to any of these scales – a text field would only work as a statistical variable if there are a limited set of possible values, such as the status of a problem (Open, Assigned, Testing, Closed, etc.), or the name of a code module.

Table 1. Measurement Scales

Scale	Description	Examples
Nominal (names)	A nominal scale divides the subject into categories which have no numeric scale, or relative order.	[male, female] [chocolate, vanilla, strawberry] [a list of CSCI elements] [a list of COTS products]
Ordinal (order)	A scale in which categories have a sequential order, but no quantifiable difference between steps	[CR priority: low, medium, high] [CMM levels: 1 to 5] [grades: A, B, C, D, F]
Interval	A scale in which categories have a sequential order, and a quantifiable difference between steps (i.e. there are consistent data <i>intervals</i>), but <i>no meaningful zero point</i> . Addition and subtraction are allowed on Interval scale data.	[degrees F or C] A single date [April 2, 2003]
Ratio	A scale in which categories have a sequential order, quantifiable difference between steps, and <i>has a meaningful zero point</i> . Addition, subtraction, multiplication, and division are allowed on Ratio scale data.	[degrees K or R] [defect rate] [turnover rate] [most integer or real number measurements]

In Table 1, ‘CR’ refers to a Change Request (a document to propose a change to a system), a CSCI is a software configuration item (a large portion of the overall system; akin to a

subsystem), the CMM is the [Capability Maturity Model](#), and COTS refers to commercially available software (something you could buy Off The Shelf).

Parametric and Dependent

Two key considerations when choosing a statistical model or test are whether the variables of interest are parametric (or non-parametric) and dependent (or independent).

Statistical tests are divided into two major types, *parametric* and *non-parametric*.

- A parametric test involves variables which use Interval or Ratio scales.
- A non-parametric test involves variables which use Nominal or Ordinal scales.

Most studies related to software will be parametric, but if you wish to measure the influence of gender on programming productivity, for example, a non-parametric testing method will have to be chosen since gender is a nominal variable.

Dependency identifies whether one set of data depends on another for its value. When drawing several lottery balls from one bin, each ball selected is dependent on the ones drawn before them. So if the first ball is '8,' all later draws may not be an '8.' Drawing socks from a drawer describes dependent events, because the socks available to be drawn depend on what was already removed.

Most events during software development are dependent upon each other, so this may have to be taken into consideration when choosing testing methods to be applied.

The classic example of independent events is flipping a 'fair' or unbiased coin. Each time the coin is flipped, it is in no way influenced by the previous flips. Hence each flip is independent.

Another expression of dependency is whether the population is replaced between samples.

- If the population size goes down with each sample drawn from it (no replacement), then the sample is dependent.
- If the population is replaced after each sample, then each sample is independent.

Population versus Sample Measures

Statistics is based on looking at a small fraction (a sample) of the total subject of interest (population), and trying to find out meaningful information based on that sample. In the context of software development, sampling may be largely irrelevant, since we will often "sample" the entire "population." For example, if we want to measure the defect rate for a software component, we will measure the total number of defects found to date, and the total number of lines of code – not just a sample of each.

Sampling is more important when a large number of things need to be evaluated. Suppose you just manufactured a million paper clips, and want to know if they are good enough to ship to a customer. A sample size is determined – say, a hundred paper clips – and based on measuring

the quality of that sample, you decide whether to accept the entire batch. Statistical sampling can tell you 1) how big that sample needs to be, and 2) how many paper clips could fail quality inspection for the overall batch to still be acceptable.

We look at sampling primarily in the context of sampling to measure customer satisfaction. We might want to know how many customers to poll to determine if they are happy with our product.

It is important to distinguish between sample and population measures, so the scope of the measurement is clear.

- A population measure might be the ‘average closure time for urgent problems’ – it’s based on the set of closure times for every urgent problem ever recorded.
- A sample measure might be the ‘average customer satisfaction rating’ – based on some kind of sampling method (cluster, stratified, etc.), a sample of customers is chosen, they are surveyed for their satisfaction, and that sample is the basis for the measure. By choosing an unbiased sample, we intend to understand the population satisfaction rating by measuring the sample.

Basic Measurements

Once a variable has been selected, there are several basic measurements which will often be needed before further analysis is possible. They are presented in Table 2, with the types of measurement scales from which may be calculated. For example, you can’t find the “average” gender of a group of people, or the average car color, because those are Nominal variables.

Table 2. Basic Measurements

Measure	Definition	Measurement Scales	Examples
Count	Number of that data element present	Any	Number of CRs Number of milestones completed on time
Mean	Average	Interval or Ratio	Mean CR close time
Percentage	A proportion expressed as a percentage	Any	% of CRs which are closed per month
Proportion	Compares count of two measures. If the measures are real numbers, it’s called a fraction instead of a proportion.	Any	Ratio of open to closed CRs
Standard Deviation (s.d.)	Measures the distribution of measurements – how consistent they are	Interval or Ratio	s.d. of CR close time
Variance	Square of standard deviation; measures data dispersion	Interval or Ratio	Variance of CR close time

The standard deviation and variance measurements are rarely used in a software development context. They are more likely to apply to an engineering environment where, for example, there may be requirements on the consistency of output from a device (e.g. an electrical signal). The closest software corollaries may be to measure the variance in traffic in a network, or in processing speed for a test case (hence measuring non-functional requirements like throughput or performance).

Common Errors

Table 3 describes ten mistakes which are commonly made in statistics. Look for them while watching commercials, Congress, or documentaries. ☺

Table 3. Ten Common Statistical Mistakes

Error	Description
1. Selective sampling	<p>Statistical sampling needs to be random, and take into account human response tendencies. A typical example of non-randomness is to take a small, poorly selected sample, and extrapolate to the rest of the universe from it.</p> <p>A typical violation of the latter is to ask people a sensitive question, and not give a way to hide their responses.</p> <p>A neat trick around this is to ask the person to flip a coin before giving their response; if the coin is heads, give the “safe” answer regardless of the truth. If the coin is tails, answer honestly. Then double the number of non-safe answers.</p>
2. Jump to conclusions	<p>This is commonly done by taking two unrelated data sets, and trying to reach a third conclusion from them. Look for faulty logic or alternative conclusions which could be reached from the same data.</p>
3. Spurious accuracy	<p>Realize the limits of accuracy in reporting data (i.e. significant digits). If a river was 3,000,000 years old twenty years ago, it is still 3,000,000 years old today – not 3,000,020! Similarly, if a result is 1000 +/- 5%, don’t report the results with three decimal places, like 1003.563 +/- 50.409.</p>
4. Faulty Comparison	<p>Be sure that comparisons are made across comparable sets of data. E.g. don’t compare LOC productivity in Ada with that for Visual Basic.</p>
5. Convenient Averages	<p>Beware of the differences among mean, median, mode, and mid-range. Depending on the distribution of data, the one selected may tilt the result substantially.</p>
6. Faulty Definitions	<p>Make sure the scope of definitions are clear when comparing data sets. Is everyone using the same definition for LOC?</p>
7. Convenient Definitions	<p>Changing the definition of a measure doesn’t change the reality hidden behind it. Often seen in crime reporting, or budget analysis.</p>

Error	Description
8. Wild Estimates	Make sure estimates are clearly described and documented. Consider the reliability of data sources. Wild numbers can result from speculation by an “expert” source, such as a cop guessing the total value of bribes in a year.
9. Confusing coincidence with probability	Strange-but-true events can be due to hidden causes, psychological factors, and seemingly rare events simply being more likely than expected.
10. Fuzzy Cause and Effect	Beware of casual associations between events – they are not the same as cause and effect! E.g. “If 99% of mass murderers drank milk as a child, then drinking milk make people kill.”

Hypothesis Testing Model

Required Inputs

In order to apply statistical modeling through hypothesis testing, one needs three pieces of information:

- 1) A Substantial Hypothesis – a statement or question of what you are trying to prove.
- 2) The Sample Size must be defined.
- 3) What Statistic is being challenged? This may be a mean, standard deviation, percentage, etc. as discussed in the Basic Measurements section earlier.

The CANDOALL Model

Given those inputs, one method for applying statistical testing to a set of data is the CANDOALL model. CANDOALL is an acronym for each step in the model, as shown in Table 4.

Table 4. The CANDOALL Model

Step	Description	Example
C – Claim a substantial hypothesis	Establish what you want to prove.	Is the mean defect rate for JCALS software is less than 20 defects per 1000 LOC?
A – Define the Alternative hypothesis	Define the opposite of the initial hypothesis.	The average defect rate for JCALS software is greater than or equal to 20 defects per 1000 LOC.

Step	Description	Example
N- Select the Null hypothesis	Determine which hypothesis is the null hypothesis, H0. The other hypothesis is labeled H1.	H0 = defect rate <20 H1 = defect rate >= 20
D – Decide on the desired confidence interval	Choose the level of confidence you want in the results; generally picked from 80 to 99%.	Alpha = 0.05 means 95% confidence in the results of the analysis.
O – Order a test	Choose appropriate tests to apply from the next section	Hmmm, this sounds like “Testing a Claim About a Mean”
A – Use Arithmetic to apply the test	Execute the test. Determine the test statistic, the critical value, and the critical region.	This may be done with Excel, SPSS, SAS, Minitab, etc.
L – Look to accept the null hypothesis	Examine the test output and use it to determine if the null hypothesis is accepted.	Decide if the data passed or failed.
L – In Layman’s terms, describe what happened	Translate the results into relatively normal English.	Express the results in ‘managerese.’ May want to focus on the actions needed as a result of the test results.

Types of Statistical Analysis

There are seven major types of statistical analysis for testing hypotheses. Table 5 provides a summary of them, with a sample of the specific methods which may be used to test them, and guidelines how they may be applied to software development and maintenance environment.

Note that many comparisons which one may make based on simpler data can be made more meaningful by applying statistical analysis techniques. For example, if one wants to know if the mean defect rate is less than a defined performance standard, the natural tendency for most people is to just compare the mean measured defect rate to the standard. If the standard is 20 defects/KLOC, and we measured 11.5 defects/KLOC, then we would conclude the measured defect rate is below the goal. However, by applying the techniques shown here, we can add more weight to the results presented by giving a confidence level in the results.

Table 5. Types of Statistical Tests

JCALs is a project name, CSC is an organization, SOW is a Statement of Work

Major Types and Specific Methods	Description	Software Applicability
Testing a Claim About Variation Z-Test	Prove a product or process is more consistent than a standard.	Prove that JCALS performance (e.g. response time) has a lower variation than that required by the SOW.
Testing a Claim About a Mean Chi square (X^2) Test	Prove a product or process has higher characteristic than a standard	Prove the (defect rate, response time, etc.) of the system is higher than the value stated in the SOW. <i>or</i> Prove the productivity of the JCALS team is higher than the average for CSC.
Testing Claims About Two Means – Dependent Student T-Test	Determine whether one product or process was affected by a change	Prove if a process improvement effort resulted in a real improvement in some variable (e.g. defect rate, productivity, etc.)
Testing Claims About Two Means – Independent F-test of variances, then T-test	Determine whether two unrelated products have similar characteristics	Prove if two projects have comparable defect rate, productivity, etc.
Testing Claims About Several Means Analysis of Variation ANOVA	Prove a product or process is affected similarly by three or more different changes; <i>or</i> Prove three processes are similarly affected by the same change.	See if the same process improvement had the same effect on three different projects.
Testing Claims Using Nonparametric Methods Sign Test	Prove two groups have the same trait; <i>or</i> Prove two groups have the same median	See if the staff for two projects have the same median level of programmer experience.
Spearman’s Rank Order Correlation Coefficient Spearman Test	Prove if two projects have the same characteristic; <i>or</i> Prove whether two characteristics are related for one project Finds correlation on a scatter plot, where r^2 is the % of variation explained by this regression.	See if there is a relationship between programmer experience level and defect rate for one project.

Statistical Test Selection

More detailed information on how to select a statistical test is given in Table 6. Notice that we don't use many of these tests, but it's good to know they are available.

Table 6. Common Statistical Test Selection Options

Subject	Hypothesis	Assumption	Parametric Test	Nonparametric Test
A mean	Class A has a higher than average IQ	N>30, or sd known N<30, or sd unknown	Z T	Sign Test Wilcos/Mann Whitney (U)
A proportion	75% of voters prefer Fred	Np>5 and Nq>5	Z	Not applicable
A standard deviation	Instrument X has fewer errors than Instrument Y	Normal population	X ²	Kruskal Wallis (H)
Two means	EZ diet is more effective than TY diet	Dependent Independent	T T or Z	Sign Test (U)
Two proportions	Drugs A and B result in the same percent of cures		Low T or U means they are similar	Not applicable
Two standard deviations	The ages of group A are more homogeneous than group B		F	(H)
Two proportions	There are more Democrats in Denver than in LA		Z	Sign Test
Relationship between 2 variables	Smoking is related to cancer		Pearson R*	Spearman R*
Dependence between 2 variables	The answer to question 1 will affect the answer to question 2		F	(H)
How close do variables match expected curve?	See contingency table below		X ² (Chi squared)	Not applicable
3 or more means	Drug A is more effective than drugs B, C, or D		ANOVA, F	Not applicable

* if R is close to zero, there is no relationship

Contingency Table

A contingency table is a table relating the frequency of two variables, such as that seen in Table 7. It can be used to show a dependency between the variables, but does not indicate their relationship. Compute X^2 ; if large, there is a dependency between the variables.

Table 7. Contingency Table

	Study of the deaths of 1000 males		
	Cancer	Heart Disease	Other
Smoking	135	310	205
Non-smoking	55	155	140

Terminology

General Definitions

Table 8 presents definitions of common statistical terms, mostly related to the normal distribution.

Table 8. Definitions

Term	Definition
Central Limit Theorem	For large values of N, the distribution about the mean becomes a Normal curve.
Confidence Interval	The level of confidence that the answer is correct – generally from 80 to 99%. A.k.a. confidence level.
Histogram	A bar chart of frequency distribution
Mean	The average of a set of data; equal to the sum of their values, divided by the number of data points.
Median	In a set of data, the median is the value occurring nearest the middle of the set.
Mid-range	The average of the maximum and minimum of a set of data.
Mode	The most frequently occurring value in a set of data.
Normal distribution	The symmetric bell-shaped curve representing the distribution due to normal chance. A.k.a. the Gaussian curve.
Population	The entire set of subjects of interest, such as the total number of customers.
Probability	The likelihood that a given event will occur, from 0 to 1.
Range	The difference between the highest and lowest measured data.
Regression line	The line of best fit through a set of data on a scatter plot.
Sample	The subset of the population, N, used to determine the characteristics of the entire population.
Sampling error	Error due to the sampling process itself (not random or due to differences in the variables being measured)
Standard deviation	The weighted average amount by which values deviate from the mean.

Term	Definition
Standard normal distribution	A normal distribution with mean=0 and standard deviation=1.
Statistic	A measured characteristic of a sample.
Variance	The standard deviation squared; a measure of dispersion.
Z Score	The number of standard deviations from the mean.

Hypothesis Testing Terminology

Table 9 presents definitions of terms related to hypothesis testing.

Table 9. Hypothesis Testing Definitions

Term	Definition
Alternative Hypothesis	The hypothesis which is the opposite of the null hypothesis.
Correlation Coefficient	A real number in the range from -1 to 1 describing the relationship between two variables. A value of -1 means a negative correlation; namely if x goes up, y goes down. A value of +1 means a positive correlation; if x goes up, y goes up. A zero correlation means there is no relationship between them; if x goes up, y could go up or down.
Critical Region	The portion of a distribution not due to random chance, generally equal to 1 to 20% of the total distribution (per confidence interval).
Critical Statistic	The answer you got on a test, such as the value of Z, chi squared, etc.
Critical Value	The value of a distribution between random chance and not.
Degrees of Freedom	The number of uncontrolled variables in a problem.
Goodness of Fit	The degree to which actual data match the expected values.
Hypothesis	A claim that some characteristic of a population is true.
Hypothesis Test	A method for testing claims about populations. A.k.a. test of significance
Level of Significance	The probability at which the null hypothesis is rejected. A.k.a. alpha (α)
Null Hypothesis	A statement of the assumption of no change; differences observed are due only to random chance. The equation with an equals sign is generally the null hypothesis. The purpose of a statistical test is to accept or reject the null hypothesis.
Test statistic	The sample statistic based on sample data; i.e. the calculated statistic

Statistical Test Method Terminology

Table 10 defines terminology related to several types of common statistical tests.

Table 10. Test Method Terminology

Term	Definition
Analysis of Variance (ANOVA)	A method for determining the significance of differences among a set of sample means.
Chi-square Test	A method for testing hypotheses about variation
Dependent samples	Samples which depend on each other for their characteristics; e.g. sampling before and after a change in process
F Test	A distribution used to test two variances
Independent samples	Samples from groups which are physically or logically separate
Kruskal-Wallis Test	A nonparametric hypothesis test to compare three or more independent samples.
Linear Regression	A regression analysis for continuous-valued (real number) data.
Logistic Regression	A regression analysis for two-valued (dichotomous, T/F, or Y/N) data.
Multiple Regression	A study of linear relationships among three or more variables.
Nonparametric Methods	Tests for data which do not have a normal distribution, and/or pertain to data not using an interval or ratio scale.
One-Tailed Tests (Left or Right)	The test for a parameter being strictly above (Right) or below (Left) some value.
Parametric methods	Hypothesis testing method for parameters on interval or ratio scales. Normal distribution is assumed.
P-Value Test	Indicates the probability a test statistic was due to random chance.
Sign Test	A nonparametric test for comparing samples from two populations.
Spearman's Rank Order Correlation Coefficient	A measure of the strength of the relationship between two variables.
Student T Test	See T Test
T Test	A bell-shaped test for small populations of experimental data. A.k.a. Student T Test
Two-tailed Test of Significance	A test in which a parameter is proposed to equal to some value; hence the alternative hypothesis is both tails of the distribution.

Reporting Statistical Significance

There are four possible answers frequently used for expressing the statistical significance of a test's results. The purpose of these limited answers is to bracket the approximate level of significance of the test output, instead of literally reporting, for example, Sig. = 0.083.

The same rules for choosing the correct answer are used, regardless of the type of test (T test, F test, Chi square, ANOVA, etc.). To choose the correct answer, use the "Sig." value reported by SPSS with Table 11.

Table 11. Reporting Significance

If the value of "Sig." is	Report the significance as	Interpretation
Sig. > 0.050	n.s.	(not significant) Results are significant to less than 95% level of confidence.
0.010 < Sig. ≤ 0.050	p<0.05	Results are significant to at least 95% level of confidence (but less than 99%).
0.001 < Sig. ≤ 0.010	p<0.01	Results are significant to at least 99% level of confidence (but less than 99.9%).
0.000 ≤ Sig. ≤ 0.001	p<0.001	Results are significant to at least 99.9% level of confidence.

References

- The Schaum outlines for Beginning Statistics (ISBN 0070612595, 1998) or Statistics (ISBN 0070602816, 1998) are good for more of the math behind the terms.
- *Measuring the Software Process*, by Florac and Carleton, 1999, ISBN 0201604442 focuses mostly on using control charts to manage software development
- ISO publishes a two-volume set, "[ISO STANDARDS HANDBOOK: STATISTICAL METHODS FOR QUALITY CONTROL](#)," for the most authoritative source

Statistical Q&A

Why don't we just use Excel instead of SPSS?

Microsoft Excel is adequate for simple arithmetic, but has been shown to be unreliable for more complex math, such as regression analyses.

See test data sets by the NIST for sample cases (<http://www.nist.gov/itl/div898/strd/> - **warning**, the site contains the most painful word art ever seen on the planet). SPSS successfully passes such tests, has a much wider range of statistical tests available, and produces an inexpensive student version.

Why do we do regression analysis?

Regression analysis is done to find a mathematical equation which accurately describes the relationship between two or more measures (variables), such as project schedule versus size, or defect rate versus time, etc. In brief, regression analysis is best-guess curve fitting using specific types of possible equations.

The dependent variable (Y) must use a ratio or interval scale, and the independent variable (X) may use a ratio, interval, or ordinal scale.

What kind of equations can we use for regression analysis?

See "Curve Fit Formulas for SPSS" below. We mostly use linear, logarithmic, and power equations. The SPSS regression analysis finds the best coefficients (constants) to match the data provided, and expresses them in terms of a value (B) and its standard error (SE B).

What is a "good" curve fit?

The goodness of fit of a curve fit is summarized in the R^2 (R Square) value. A perfect curve fit, with all data points on the regression curve, has R Square of unity (exactly one). The worst possible case, with no relationship between the variables, gives an R Square of zero. Generally, an R Square above 0.90 is very good, and above 0.98 is excellent.

Can I believe the coefficients which come from a regression analysis?

SPSS will generate coefficients for almost any set of data. One way to check the validity of coefficients (a sanity check, if you will) is to make sure the coefficients are, to a 95% level of confidence (LOC), not zero. This is where we apply the T test. SPSS gives the T values in the output:

$$T = B / (SE B)$$

T is a coefficient's value (B) divided by that coefficient's standard error (SE B).

If $|T| > 2$ (the absolute value of T is greater than two), then that coefficient is plausible; more precisely, that coefficient is not zero to at least a 95% level of confidence.

If $|T| \leq 2$ for any of the coefficients, then that curve fit is meaningless.

Another way to check is to use the significance level of T, "Sig T". For 95% LOC, we want Sig T ≤ 0.050 for the coefficient to be usable.

How can I tell two distributions of data apart?

A normal distribution of data may be expressed using the form " $a \pm b$ ". We know that 95% of the data lies between ± 2 times the standard error (b) from the mean (a). So that set of data will, 95% of the time, range from a value of $(a - 2*b)$ to $(a + 2*b)$. If a second data set is given by " $c \pm d$ ", then its 95% probable range will be from $(c - 2*d)$ to $(c + 2*d)$. If those two ranges of values overlap, *then there is no statistically significant difference between those data sets.*

Example: Let's say you want to compare test results for two classes. Is a data set with a distribution of 5.2 ± 0.8 (mean of 5.2, standard error of 0.8) significantly different from one which is 6.5 ± 1.1 ?

Answer: The 95% range for the first data set is $(5.2 - 2*0.8 = 3.6)$ to $(5.2 + 2*0.8 = 6.8)$. The 95% range for the second data set is $(6.5 - 2*1.1 = 4.3)$ to $(6.5 + 2*1.1 = 8.7)$. Since the range $3.6 - 6.8$ overlaps $4.3 - 8.7$, the data sets are not significantly different at 95% level of confidence.

NOTE: Where we use "2" for the 95% interval, to be more precise you should use 1.96. For most purposes here, 2 is close enough, and easier to remember.

How can I tell if a range of data is likely to include one number?

The 95% range can also be used to determine if a distribution of data might include a single value. For example, if a distribution of data is given by " $a \pm b$ ", and you might want to determine if is likely to contain a point " c ", then find the 95% range for the distribution (i.e. $(a - 2*b)$ to $(a + 2*b)$), and see if " c " is within that range. If " c " is in the 95% range, then that range could contain " c ".

Example: Is a data set with mean of 7.1 and a standard error of 0.2 agree with a predicted value of 7.4?

Answer: The 95% confidence interval for the data set is $(7.1 - 2*0.2 = 6.7)$ to $(7.1 + 2*0.2 = 7.5)$. The range $6.7 - 7.5$ does include the predicted value of 7.4, so this data set does agree with the predicted value to 95% confidence.

Can't we let SPSS do the T test for us?

SPSS can do the T test several different ways, if you are given the sets of individual data points to compare them (e.g. two sets of test scores). The method used here just shows that the same type of comparison can be done manually if you only have summary data available (mean and standard error).

Curve Fit Formulas for SPSS

The regression output from SPSS is for an equation of the type $Y = f(X)$ (dependent variable Y on the vertical axis is some function of independent variable X on the horizontal axis). For example, load the Employee Data file that comes with SPSS (typically under C:\Program Files\SPSS\)) and generate a linear regression analysis (Analyze > Regression > Curve Estimation) of current salary versus beginning salary (variables salary (dependent) vs salbegin (independent)). The end of the analysis includes the following:

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Beginning Salary	1.909	.047	.880	40.276	.000
(Constant)	1928.206	888.680		2.170	.031

Of this, we are concerned with the significance of the variables under B, which are the actual curve fit coefficients, their standard errors, Std. Error, the 't' values, and the significance of T (Sig.). Disregard the Standardized Coefficients.

SPSS refers to the coefficients under B as b_i , where the letter i is a non-negative integer: 0, 1, 2, etc. The actual value of these coefficients can be any real number.

X^{**2}	$\frac{B}{b2}$	(b2 and b3 only exist for some polynomial curve fits)
X	b1	
(Constant)	b0	

*Notice that the label of the independent variable is used in place of "X", so if the independent variable is "Beginning Salary", then the output would show "Beginning Salary" (like seen above) and "Beginning Salary**2" instead of "X" and "X**2".*

So in the above example, $b_0 = 1928.206$ with a standard error of 888.680, and $b_1 = 1.909$ with a standard error of .047. The coefficient b_0 has a T value of 2.170, with a significance of .031, and b_1 has a T value of 40.276, with a significance of .000. In this case, b_2 and b_3 don't exist, since there are only two parameters in a linear curve fit. Since $|T| > 2$ for both parameters b_0 and b_1 , and $\text{Sig} < 0.050$ for both parameters, we can conclude this regression is statistically significant to at least 95% level of confidence (i.e. the trend observed is statistically likely not to be due to random error).

The meaning of b_0 and b_1 depends on the type of regression which was performed. The equations and their usage of these coefficients are shown in Table 12, where 'e' is the base of natural logarithms (2.71828...), 'ln' is the natural logarithm, and '**' means 'to the power of' (which is done in Excel using '^').

In this case, we did a linear regression, so $Y = b_0 + b_1 * X$. More correctly, we can show the standard errors of b_0 , $s.e.(b_0)$, and of b_1 , $s.e.(b_1)$, in the equation like this:

$$Y = (b_0 +/- s.e.(b_0)) + (b_1 +/- s.e.(b_1)) * X$$

That gives us

$$\text{Current Salary} = (1928.206 +/- 888.680) + (1.909 +/- .047) * (\text{Beginning Salary})$$

To better reflect the number of significant digits, and because +/- 888.680 looks more accurate than our data should imply, I'd round the standard errors to two significant digits, then round the values of b_0 and b_1 to match that number of decimals. Hence our conclusion is that there is a significant linear relationship between these variables, and it is given, with the units added, by:

$$\text{Current Salary} = (\$1930 +/- \$890) + (1.909 +/- .047) * (\text{Beginning Salary})$$

Table 12. SPSS Curve Estimation Equations

Name	Formula	Comment
Linear	$Y = b_0 + b_1 * X$	A straight line
Logarithmic	$Y = b_0 + b_1 * \ln(X)$	Natural logarithm
Inverse	$Y = b_0 + b_1 / X$	Similar to the form $X * Y = \text{constant}$
Quadratic	$Y = b_0 + b_1 * X + b_2 * X^2$	A parabola, or second order polynomial
Cubic	$Y = b_0 + b_1 * X + b_2 * X^2 + b_3 * X^3$	Third order polynomial
Power	$Y = b_0 * (X^{b_1}) = b_0 * (X^{b_1})$	
Compound	$Y = b_0 * (b_1^{**} X) = b_0 * (b_1^X)$	Like exponential, but with base 'b1'
S	$Y = e^{**}(b_0 + b_1 / X) = e^{(b_0 + b_1 / X)}$	An inverse exponential?
Logistic	$Y = 1 / (1 / u + b_0 * (b_1^{**} X))$ $= 1 / (1 / u + b_0 * (b_1^X))$	Where 'u' is a given upper bound, larger than any value of Y.
Growth	$Y = e^{**}(b_0 + b_1 * X) = e^{(b_0 + b_1 * X)}$	
Exponential	$Y = b_0 * e^{**}(b_1 * X) = b_0 * e^{(b_1 * X)}$	

If you choose not to “Include constant in equation”, then there will be no ‘ b_0 ’ value calculated, and you will get only b_1 (and maybe b_2 and b_3 , as appropriate) in the “Coefficient” output. The implied value of ‘ b_0 ’ will be zero or one, as appropriate.

- The difference between these cases is, in the case of a linear regression, the difference between [finding an arbitrary straight line if you **do** include the constant], *versus* [finding a straight line which must pass through the origin ($X=Y=0$) if you **do not** include the constant, since the equation becomes $Y = b_1 * X$]. We generally choose to include the constant to get the best possible curve fit (i.e. check the box “Include constant in equation”).

For regression to an arbitrary equation, such as the Rayleigh model, use the SPSS Regression Module.

Evaluating Statistical Test Results

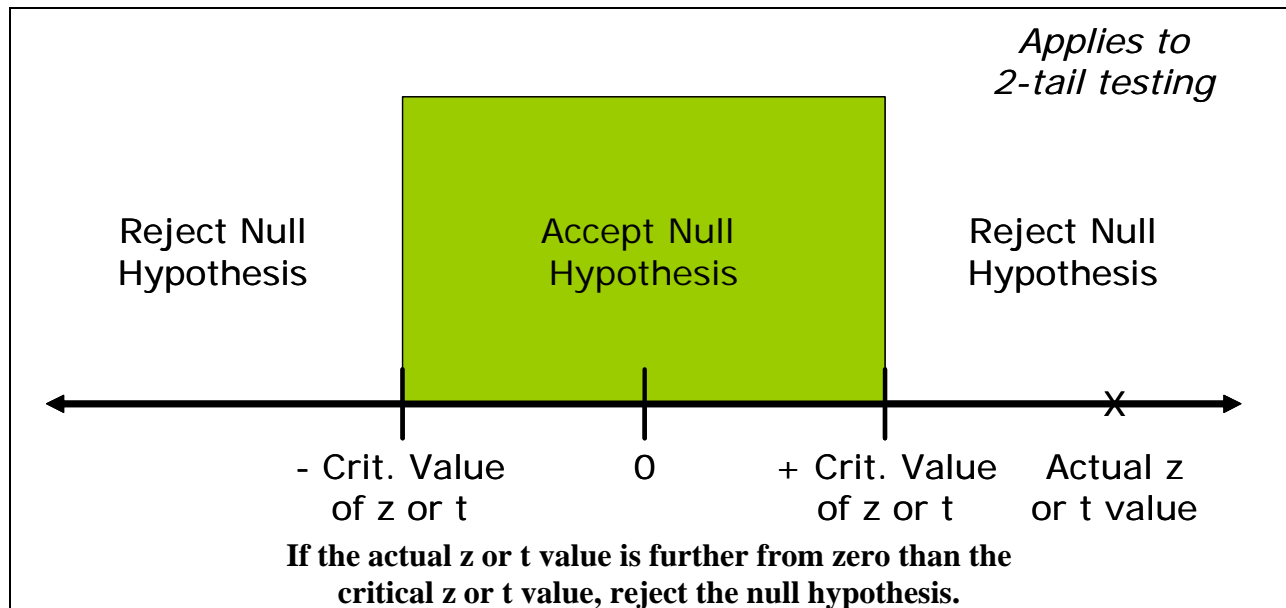
There are typically up to three outputs from statistical tests that can be used to determine the outcome. The outcome is one of two possibilities:

- Accept the null hypothesis (there is no statistically significant difference)
- Reject the null hypothesis (there is a statistically significant difference)

The types of outputs that can be evaluated are 1) 't' or 'z' values, such as from T tests or Z scores, 2) confidence intervals, mainly from T tests, and/or 3) Significance of the test result. In general, all three results must agree on the test's outcome, if all are present.

Evaluate 't' or 'z' Values

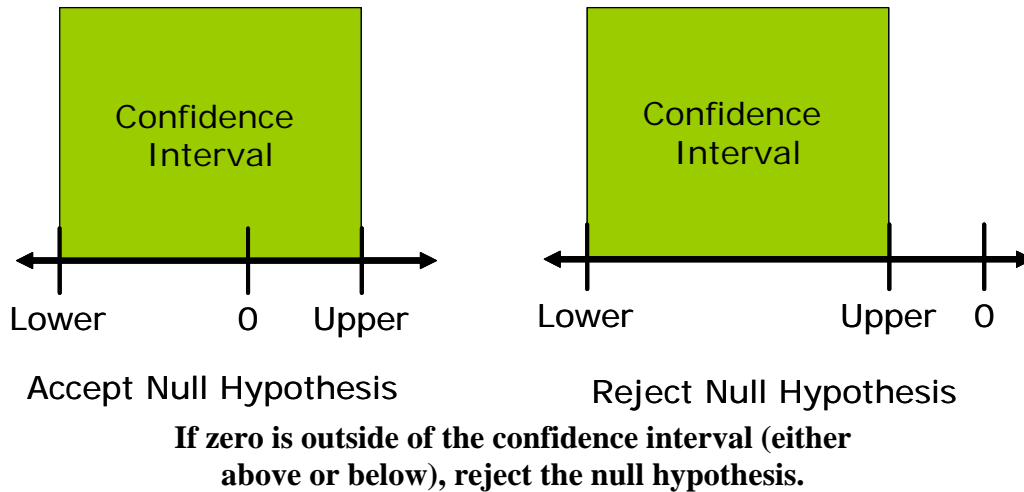
When a T test produces a 't' value for your data, it must be compared to the critical t value for the given confidence level desired (generally 95% or 99%), sample size (n, which determines the degrees of freedom, $df = n - 1$), and whether the test is using 1- or 2-tailed results (the latter is the default, for reasons explained in INFO 515, lecture 5).



This tends to be the hardest kind of output to interpret, since the critical t value (or for ANOVA, the critical F value) has to be looked up. Critical z values are based solely on the confidence level of the test; $z_{crit} = 1.96$ for 95% confidence, $z_{crit} = 2.57$ for 99% confidence.

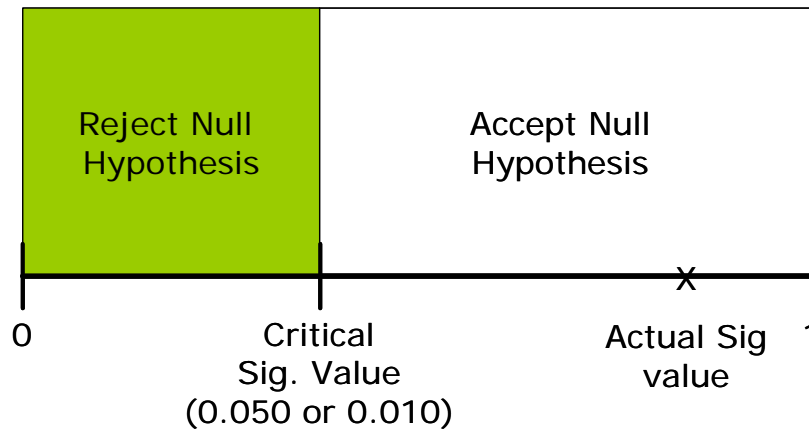
Evaluate Confidence Interval

In order to judge whether a value is statistically different from zero, its confidence interval can be measured. For examples, the value could be the difference between two sample means (independent or dependent T tests), or the difference between a sample mean and a fixed value (one sample T test). If the confidence interval of that difference includes zero, accept the null hypothesis. If zero is outside of the confidence interval (i.e. both Upper and Lower limits of the confidence interval are the same sign), then reject the null hypothesis.



Evaluate Significance Value

The Significance of a test result is the most universal output to assess – many tests don't produce readily evaluated F, z or t values, and many don't have confidence intervals, but all of them produce a Significance, usually called Sig. *The critical value of significance is one minus the desired confidence level in the results*, so for 95% confidence level, use a critical Sig of 0.050. For a 99% confidence level, use a critical Sig of 0.010. If the actual value of Sig is less than the critical value, reject the null hypothesis.



If the actual value of Sig is less than the critical value, reject the null hypothesis.