Reasons to evaluate the relationships

- To describe relationship

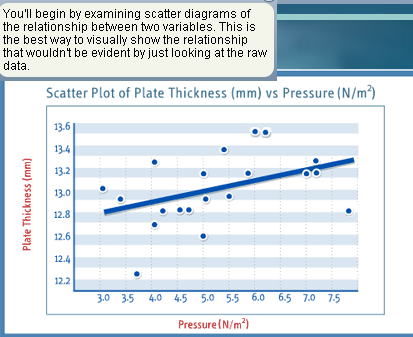
- To change or control the value of variables

- To make estimates and predictions

Tools to explore correlation

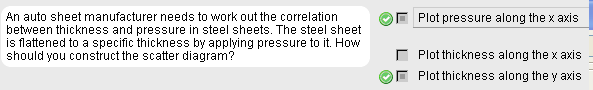
* Scatter diagrams
* Correlation
* Regression

**1. SCATTER DIAGRAMS**



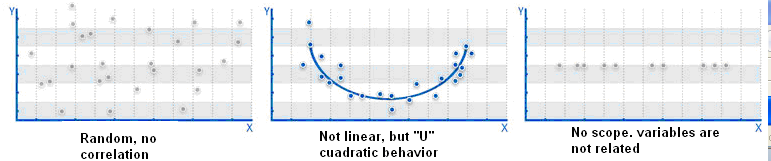
Y: Dependent variable

X: Independent variable

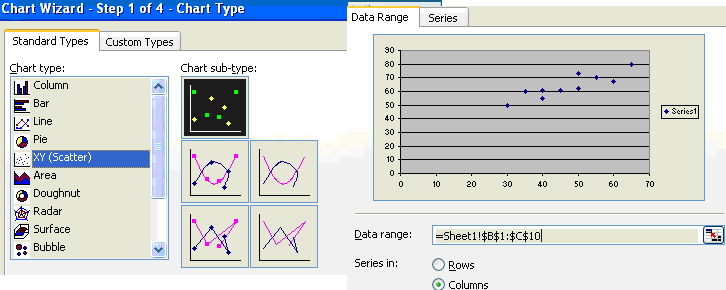


Interpretation Rules:

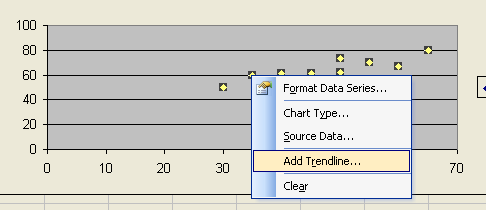
* Direction Positive – negative correlated , if line goes up or down
* Strength Strong – weak correlated : how close to tendency or loose
* Form Making line - random, (correlation or no correlation)



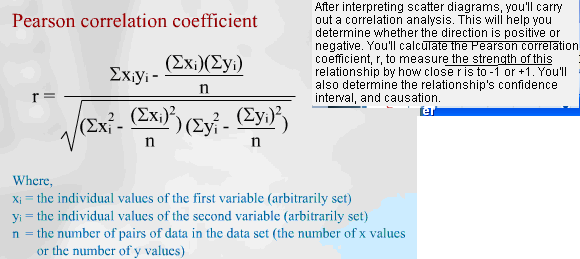
**Using Excel: Insert chart to explore**



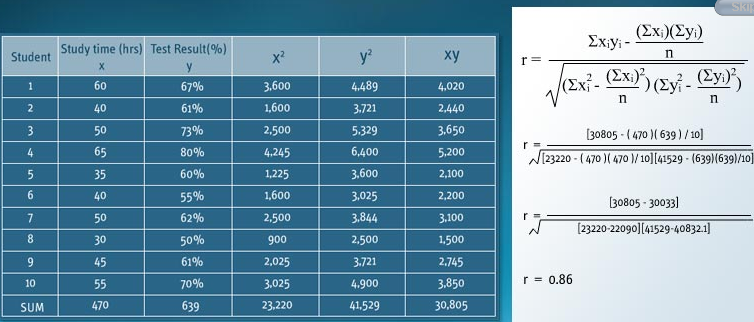
Click in any point to add trend line

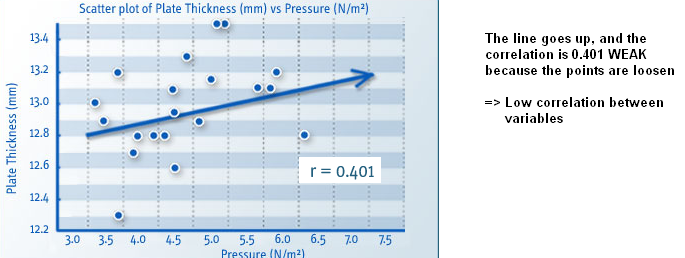
**2. CORRELATION Pearson Correlation coefficient**

To assess how much is the strength



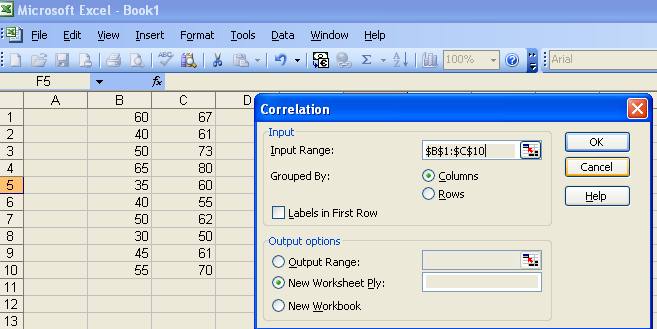
(Neg. correlation) -1 < r < +1 (Pos. correlation) **r : the SLOPE**



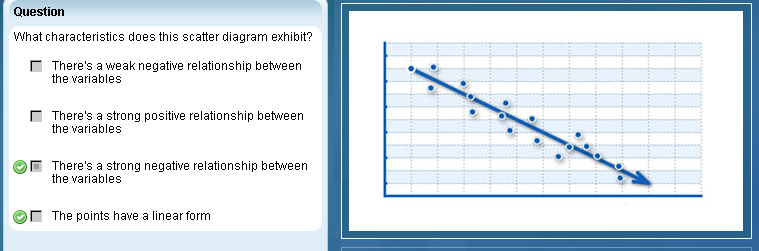


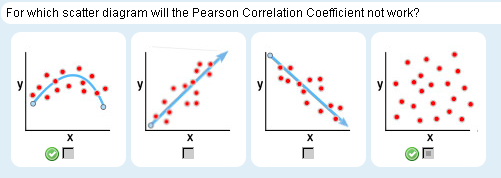
**Using Excel**

Tools - Data Analysis - Correlation

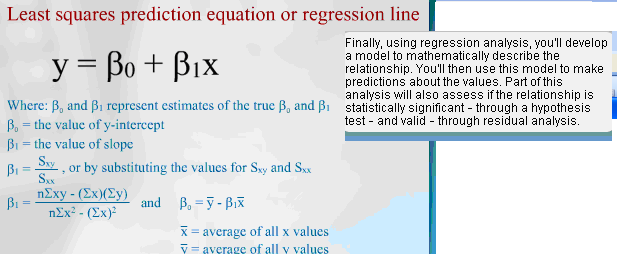


|  |  |  |
| --- | --- | --- |
|  | *Column 1* | *Column 2* |
| Column 1 | 1 |  |
| Column 2 | 0.877749 | 1 |



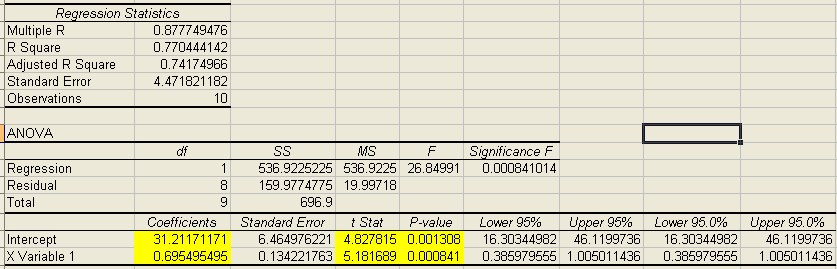


Pearson is ONLY for linear**3. REGRESSION. CREATE PATTERN, FORMULA**



Using excel

Tools - Data Analysis - Regression



**Y = 31.21 + 0.695495 X**

Remember

- Correlation is not causation. There could be a pure coincidence

- Context is critical

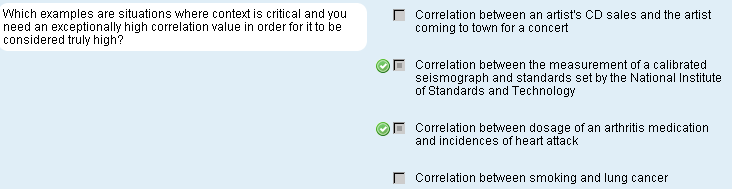
- Pearson correlation works ONLY FOR LINEAR RELATIONSHIP !!!!

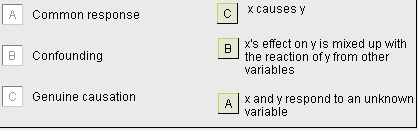
- Correlation coefficient is highly sensitive to outliers

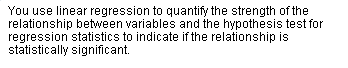
Correlation may help to find the real causation. The relationship between variables could be affected by a hidden reason or a third variable.

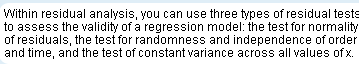
Context set the relevance of correlation to set if it is high enough, depending industry or importance.

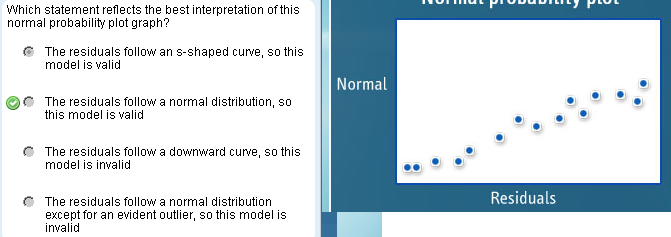
Outliers may affect the correlation, check if they should be taken away. Investigate the cause of outlier

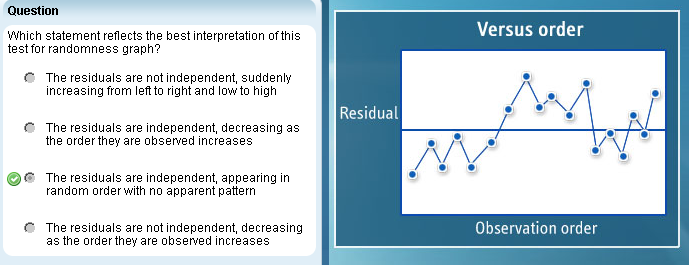


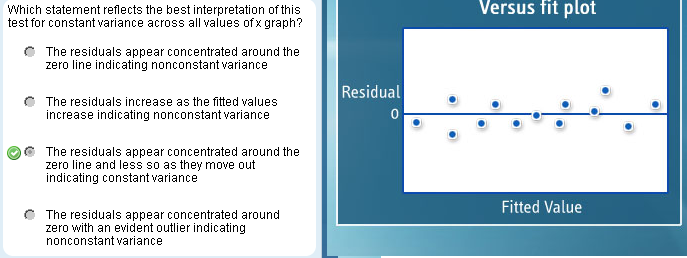


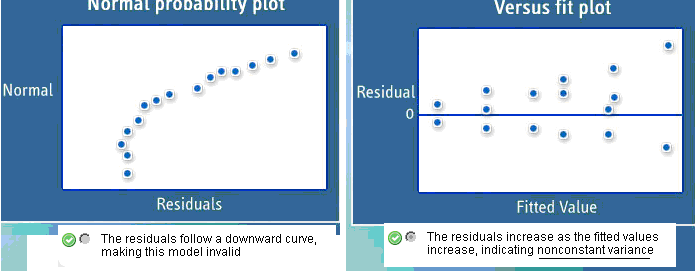


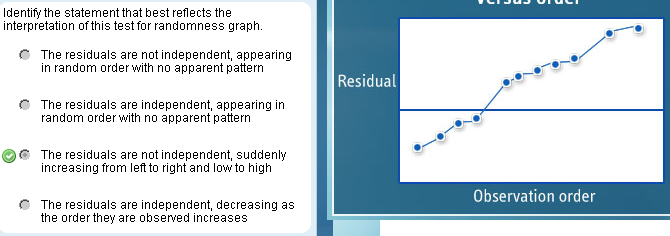


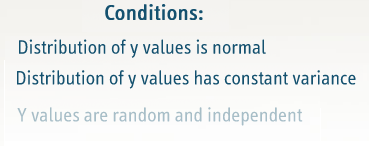












### Key Concepts in Hypothesis Testing

A hypothesis is a theory or opinion about the relationship between variables. Statistical analysis can show whether observed differences between samples are true differences, or if they are due to random variation or chance.

To begin a hypothesis test, you must define two complementary hypotheses:

* The **null hypothesis** (designated H0) expresses the status quo. It assumes that any observed differences or variances between two or more populations, or between a sample and a specified value, are due to chance or random variation.
* The **alternative hypothesis** (designated Ha) assumes that the observed differences or relationships between two populations are real and not the result of chance or random variation.

The two hypotheses are mutually exclusive. If one is true, the other must be false. Your overall strategy is to use statistical processes to reject the null hypothesis by proving the alternative hypothesis. If you can't prove your alternative hypothesis, the null hypothesis stands.

In classic statistics, results are always stated in terms of the null hypothesis. That is, the result of the hypothesis test is to "reject" or "fail to reject" the null hypothesis. You never speak in terms of accepting or rejecting the alternative hypothesis. You also never accept the null hypothesis. However, non-statisticians who use hypothesis testing often do speak of accepting the alternative hypothesis. This is what they hope to achieve, and it's a more natural way of speaking.

#### Risk and significance in testing

All hypothesis tests involve testing a sample of a larger population. Samples can contain unrepresentative data, and sample sizes can be too small or too large, both of which can result in one of these flawed outcomes:

* Rejecting a null hypothesis that is actually correct is a **type I error**, which is akin to mistakenly rejecting a part that has no defects. This may result in you making unnecessary changes to your process.
* Failing to reject a null hypothesis that is actually incorrect is a **type II error**, which is similar to mistakenly failing to reject a defective part. This may result in you failing to make necessary changes to your process.

The probability of committing a type I error is called the **alpha risk**, or the producer's risk, since it's a risk the producer runs of discarding a product that is, in fact, good. The alpha is the maximum risk you are willing to take of rejecting a null hypothesis when it is true. The more risk averse you are, the smaller you want the alpha risk to be, and the larger your sample size will need to be. In hypothesis testing, the alpha is used as the basis for rejecting or not rejecting the null hypothesis.

The risk of committing a type II error is called the **beta risk**, or the consumer's risk, because it's a risk that the producer will fail to reject flawed goods, and the consumer will suffer. Ideally, you want a low beta risk, meaning that there is a low probability of failing to reject the null hypothesis that is false.

In hypothesis testing, alpha has a specific and important use: it represents the level of **statistical significance** that is required in order to reject the null hypothesis. Ideally, your hypothesis test calculations will produce a statistically significant result. This means that the change or relationship being observed is unlikely to have occurred by chance. Statistical significance does not mean that the result is important or that it has any decision-making utility. A result may be statistically significant but not practically significant, which means the result is reliable but isn't worth the time or expense to implement.

It's very important not to reject a null hypothesis on the basis of a very small statistical significance, unless your intent is to detect very small differences. To prevent this from happening, a cut-off point is established. The cut-off point is equal to the alpha value (also called the critical alpha):

* if your result is greater than the critical alpha, you reject the null hypothesis
* if your result is less than critical alpha, you do not reject the null hypothesis

**P-value** is the probability of getting a value of the sample test statistic that is at least as extreme as the one found from the sample data, assuming the null hypothesis is true. The p-value is compared to the critical alpha level of the test. If the p-value is smaller than alpha, the result is considered conclusive or significant.

Small p-values suggest that the null hypothesis is unlikely to be true. The smaller the p-value, the more convincing the rejection is of the null hypothesis. P-value indicates the strength of evidence for rejecting the null hypothesis, rather than simply concluding to reject or not reject it.

The **power** of a hypothesis test is the probability that the test will give you the correct result, and that you will reject the null hypothesis when it is actually false. You want the power of your test to be high.

Beta is the probability of committing a type II error; the probability of not committing a type II error is one minus beta. The maximum power a test can have is 1; the minimum is 0.

The probability of making type I and type II errors figures into the decision whether to reject the null hypothesis, and into the calculation of the **test statistic**, which is calculated using values from the data and a formula that is different for each type of test (t-test, z-test, or F-test). The test statistic is compared with the **critical value** that is found using degree of freedom and test significance from an appropriate table. The way you look up the critical value is also different for each test:

* if the test statistic value is less than the t-critical value, you fail to reject the null hypothesis
* if the test statistic value is greater than the t-critical value, you reject the null hypothesis

An important concept in hypothesis testing is the decision of whether you are going to use a one-tailed or two-tailed test of significance:

* A **one-tailed test** is used when you want to know whether your sample mean is significantly larger or smaller than your population mean. The test is in one direction – to the right (more) or to the left (less). The critical region for a one-tailed test is the set of values less than the critical value (significance level) of the test, or the set of values greater than the critical value of the test. Note that the critical value (alpha) is usually set at 5% (0.05). A one-tailed test indicates whether your sample mean is significantly more or significantly less than your population mean.
* A **two-tailed test** is used if you want to know whether the sample is significantly different from your null hypothesis. A two-tailed hypothesis states only that a difference exists; it does not specify the direction. With a two-tailed test, the critical value (alpha) of 0.05 is halved; it is 0.025 in either direction, which means you're willing to take a 2.5% risk of rejecting a null hypothesis that is actually true. In a two-tailed test, the null hypothesis is rejected if the observed sample statistic is more extreme than the critical value in either direction (higher than the positive critical value or lower than the negative critical value).

Hypothesis testing is an important Six Sigma tool. It gives you a straightforward method for determining whether an observed change is statistically significant or the result of chance. The success of your test depends on choosing the correct test, setting up your test parameters correctly, choosing an appropriate sample size, and selecting appropriate risk levels.

You can test a hypothesis in many ways, including the z-test, t-test, F-test, chi square, and ANOVA, to name a few. The appropriate test is chosen on the basis of what the Six Sigma team wants to know and on what form the data is in. Regardless of the test that is chosen, the steps in the testing process are the same.

Hypothesis testing is similar to other testing processes with which you are familiar: determine what you want to know, gather data, analyze data to reach a conclusion, and decide the best action to take, based on that conclusion. The only difference between hypothesis testing and other testing processes is the application of statistics. With hypothesis testing, you use statistics to factor in the probabilities of errors in the testing process itself, and errors in the sample data. Thus, hypothesis testing removes some of the guesswork and gives you more objective evidence on which to base a business decision.

#### 1. Define the business problem

The first step in the hypothesis testing process is to define the business problem – that is, determine what you want the test to tell you. In the first two phases of the DMAIC process, you likely used a variety of tools, such as a Pareto chart, to define and measure potential inputs and outputs of your process.

#### 2. Establish the hypotheses

Knowing the question you need answered, you set up your two hypotheses. Your question is answered, in opposite ways, by the null hypothesis and the alternative hypothesis:

* the null hypothesis (H0) is traditionally the hypothesis of no change – it assumes any change results from random variation or noise
* the alternative hypothesis (Ha) is the hypothesis of change – it assumes that observed change is real

Three conditions are possible for a null hypothesis: equal to, less than or equal to, or greater than or equal to a value. The null hypothesis always has an equal sign. The alternative hypothesis can be less than, not equal to, or greater than the value given in the null hypothesis.

#### 3. Determine the test parameters

After establishing your hypotheses, you decide how to conduct your hypothesis test by determining your test parameters. There are a number of considerations, including

* selecting the appropriate statistical test
* identifying key test considerations
* conducting sampling
* doing checks

Choosing the appropriate test can be tricky. To begin, you must first decide between two main families of tests:

* Parametric tests are based on the assumption that your data is sampled from a normal distribution. The data fits the bell-shaped curve. There are many parametric tests from which to choose: t-test, ANOVA, chi square, logistic analysis, and regression analysis, for example. The major deciding factor is the kind of data – continuous or discrete – that you are working with.
* Nonparametric tests make no assumptions about the population distribution of the data.

In selecting the type of test, you must also address other key considerations:

* What kind of comparison are you making? Are you comparing a sample to a larger population? Or to another sample? Or to a standard value? This consideration affects what kind of test you use, as well as the sampling plan.
* What is the direction of comparison? Should you use a one-tailed or two-tailed test? Here is a helpful rule of thumb: if your alternative has a not equal sign, use a two-tailed test. If the alternative hypothesis has a less than sign, use a one-tailed test to the left. If the alternative has a greater than sign, use a one-tailed test to the right.
* What is the degree of freedom? Every test has a different way to calculate the statistic.
* What significance level (alpha) is needed? Many tests use 0.05 as the alpha; some use 0.01; some use 0.001. Determine the required alpha level for your test. Note that your organization may set the alpha level.

Sampling is a critical aspect of test planning. Sampling must be random to ensure that the sample population is as close as possible to being representative of the population. Sampling has three parts:

* determine the sample size – Determining the number of measurements you need is critical. If your sample size is too large, you may waste time, money, and resources; if the sample size is too small, you may get inaccurate results. You normally use software to calculate sample size. However, if you're using a small sample, you can easily calculate sample size yourself. The sample size (n) equals the confidence level (Z) squared multiplied by the standard deviation (sigma) squared, all divided by the margin of error (E) squared.
* create a sampling plan – The plan is a detailed outline that describes how you intend to gather your sample data. It should describe which measurements to take at what times, in what manner, and by whom. Your plan must be designed so the data that is collected will be representative of your population.
* collect sample data – After calculating your sample size and creating the sampling plan, you conduct the data-collection process.

All parametric tests are based on data being normal. Periodically, it's important to do checks to ensure that you're getting an appropriate reference distribution. Also, check the independence of data, as needed.

#### 4. Calculate the test statistic

The test statistic indicates whether your sample data is consistent with the population mean, or whether it's less than or greater than the mean. This will be the basis for rejecting or not rejecting the null hypothesis. Each type of hypothesis test has a formula for the test statistic. Some of the formulas use inputs that come from sample data.

In addition to calculating the test statistic, you must also determine the critical value or the p-value. These values are the cut-off points to which you compare your test statistic in order to determine whether to reject or not reject the null hypothesis. The critical value is obtained from tables, using the degree of freedom and significance level (alpha). The p-value is found using the test statistic, degree of freedom, and significance level (alpha).

#### 5. Interpret the results

The last step is to interpret the result of your calculations. Compare your test statistic to the critical value found in the appropriate table. Two outcomes are possible:

* If the test statistic is higher than the critical value, it means that your finding is significant and the null hypothesis is likely to be false. Also, the probability is small that the difference or relationship occurred by chance.
* If the test statistic is lower than the critical value, it means that your finding is not significant and you do not have enough evidence to reject the null hypothesis. The probability is high that the difference or relationship expressed in the alternative hypothesis occurred by chance.

Is statistical significance by itself sufficient to reject a null hypothesis? Not always. You must also evaluate the result in practical terms. A small statistical significance might not justify the expense of improving the inputs. Only a practical assessment will tell you whether this is the case.

Test 2 tail test 0.05 is checked with 0.025 for each side

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pressure (N/m^2)** | **Plate Thickness (mm)** | **x2** | **y2** | **xy** |
| 4.5 | 12.9 | 20.25 | 166.41 | 58.05 |
| 5 | 13.1 | 25 | 171.61 | 65.5 |
| 5.1 | 13.5 | 26.01 | 182.25 | 68.85 |
| 6.3 | 12.8 | 39.69 | 163.84 | 80.64 |
| 3.7 | 12.3 | 13.69 | 151.29 | 45.51 |
| 3.5 | 12.9 | 12.25 | 166.41 | 45.15 |
| 4 | 12.8 | 16 | 163.84 | 51.2 |
| 3.3 | 13 | 10.89 | 169 | 42.9 |
| 5.7 | 13.1 | 32.49 | 171.61 | 74.67 |
| 5.8 | 13.2 | 33.64 | 174.24 | 76.56 |
| 4.3 | 12.8 | 18.49 | 163.84 | 55.04 |
| 3.9 | 12.7 | 15.21 | 161.29 | 49.53 |
| 4.8 | 12.9 | 23.04 | 166.41 | 61.92 |
| 5.8 | 13.1 | 33.64 | 171.61 | 75.98 |
| 5.2 | 13.5 | 27.04 | 182.25 | 70.2 |
| 4.5 | 12.6 | 20.25 | 158.76 | 56.7 |
| 4.2 | 12.8 | 17.64 | 163.84 | 53.76 |
| 3.9 | 13.2 | 15.21 | 174.24 | 51.48 |
| 4.7 | 13.3 | 22.09 | 176.89 | 62.51 |
| 4.5 | 13.1 | 20.25 | 171.61 | 58.95 |
| **∑x = 92.7** | **∑y = 259.6** | **∑x2 = 442.77** | **∑y2 = 3371.24** | **∑xy = 1205.1** |
|

The regression hypothesis test statistic (t) is calculated by dividing the slope of the regression line (represented by the Greek lowercase letter Beta subscript one with a caret accent, pronounced as small beta subscript one hat) by the result of dividing the standard deviation of errors in the sample (Se or S Subscript e) by the square root of the sum of squared deviations of the x observations (Sxx or S subscript xx). This can be evaluated using the following formula: t equals small beta subscript one hat divided by the result of S subscript e divided by the square root of S subscript xx.

The value of standard deviation of error (Se or S subscript e) is calculated by taking the square root of the difference between the sum of squared deviations of the y observations and the product of slope of the regression line times the sum of cross products of x deviations with the y deviations (Sxy or S subscript xy). This can be evaluated using the following formula: S subscript e equals the square root of S subscript yy minus the result of the product of small beta subscript one hat and S subscript xy.

The following calculations handle the component expressions of the standard deviation of error formula.

Sxx is calculated by subtracting the result of the sum of x and squared divided by n, from the sum of all x squared, where ∑ (pronounced as sigma or summation of) represents sum of and n represents the number of points. The formula to calculate this is: Sxx equals ∑x2 minus ∑ x raised to the power of two divided by n.

Sxy is calculated by subtracting the product of the sum of x times the sum of y divided by n from the sum of all x times y where ∑ (pronounced as sigma or summation of) represents the sum of and n represents the number of points. The formula to calculate this is: Sxy equals ∑ x times y minus the product of ∑x times ∑y divided by n. 

Syy is calculated by subtracting the result of the sum of y and squared divided by n, from the sum of all y squared, where ∑ (pronounced as sigma or summation of) represents sum of and n represents the number of points. The formula to calculate this is: Syy equals ∑y2 minus ∑ y raised to the power of two divided by n.

The regression equation is stated as: Plate Thickness (mm) = 12.3 + 0.141 Pressure (N/m^2)

Sxx = 13.11

Sxy = 1.85

Syy = 1.63

n = 20 => df = 19

Se = Sqr( 1.63 - 0.141x 1.85) = 1.17

Beta one hat = 0.141

t = 0.141 x sqr(13.11) / 1.17 = 0.44

H0 = b1 = 0

Ha = b1 ≠ 0

The degree of freedom (n - 1) is:

df = 19

Alpha is 0.05

| **t-Distribution Table** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **df** | **.40** | **.30** | **.20** | **.10** | **.050** | **.025** | **.010** | **.005** | **.001** | **.0005** |
| 1 | .325 | .727 | 1.376 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.3 | 636.6 |
| 2 | .289 | .617 | 1.061 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.33 | 31.60 |
| 3 | .277 | .584 | .978 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.22 | 12.94 |
| 4 | .271 | .569 | .941 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 8.610 |
| 5 | .267 | .559 | .920 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 | 6.859 |
| 6 | .265 | .553 | .906 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 5.959 |
| 7 | .263 | .549 | .896 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 5.405 |
| 8 | .262 | .546 | .889 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 5.041 |
| 9 | .261 | .543 | .883 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 4.781 |
| 10 | .260 | .542 | .879 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| 11 | .260 | .540 | .876 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 4.437 |
| 12 | .259 | .539 | .873 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 4.318 |
| 13 | .259 | .538 | .870 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 4.221 |
| 14 | .258 | .537 | .868 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 4.140 |
| 15 | .258 | .536 | .866 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| 16 | .258 | .535 | .865 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| 17 | .257 | .534 | .863 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| 18 | .257 | .534 | .862 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.611 | 3.922 |
| 19 | .257 | .533 | .861 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| 20 | .257 | .533 | .860 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| 21 | .257 | .532 | .859 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| 22 | .256 | .532 | .858 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| 23 | .256 | .532 | .858 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 3.767 |
| 24 | .256 | .531 | .857 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| 25 | .256 | .531 | .856 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |
| 26 | .256 | .531 | .856 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 3.707 |
| 27 | .256 | .531 | .855 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 3.690 |
| 28 | .256 | .530 | .855 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 3.674 |
| 29 | .256 | .530 | .854 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 3.659 |
| 30 | .256 | .530 | .854 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 3.646 |

Test 2 tail test 0.05 is checked with 0.025 for each side

Significant level of alpha

Using the p-value method and t-statistic of 0.44 for df = 19 p value is 2.093 > 0.44 p > t-stat

* You fail to reject the null hypothesis
* The independent variable IS NOT useful for predicting the dependent variable

**Other sample**

Sxx = 13.41

Sxy = -94.8

Syy = 780

n = 15 => df = 14 => p = 2.145

Se = 10.48

Beta hat one = -7.07

t-statistic is -2.47

Using the p-value method and t-statistic of -2.47 for df = 14 p value is 2.145 < Abs( 2.47 ) p < t-stat

* You reject the null hypothesis
* The independent variable IS useful for predicting the dependent variable

If your p-value is smaller than alpha, the test result is considered significant and the null hypothesis is rejected.

