2014

Analyzing Relations between data set-II

Author: Naseha Sameen

About: Statistical Terms used in any Analysis especially for BPO/KPO, their interpretation and when to use what….

Terms to Understand

LAB NOTEBOOK

NASEHA SAMEEN

Terms to Understand

Table of Contents

Terms to Understand

Terms to familiarize before we move ahead:

Linear and Non Linear Relationships:

- If all the other variables increases by x times when we increase one variable by x, it is considered to be a linear equation.
- In non-linear equation, the other variables will not increase by x times. They can be fitted into polynomials, logarithm
- Examples:
- The side of a square and its area are not linear it is a quadratic relationship. If you double the side of a square, its area will increase 4 times.
- While charging a capacitor, the amount of charge and time are non-linearly dependent. Thus the capacitor is not twice as charged after 2 seconds as it was after 1 second. This is an exponential relationship.

The term R-squared refers to the fraction of variance explained by a model

R Square:

- R Square: **It is the percent of the variance in the dependent variable, explained by the independent variables**. It is also called the coefficient of multiple determination. I**t is the square of the co-efficient of correlation** between y and x. R-squared value indicates the proportion of variation in the y-variable that is due to variation in the x-variables.

  The term R-squared refers to the fraction of variance explained by a model. R-squared calculated relative to the variance of the original series (deflated if appropriate), and this number may be the more important number for purposes of characterizing the predictive power of your model.

  If R-Squared is 1.0 for the value of one term, we can perfectly predict the value of another term. If R-Squared is 0.0, then knowing one term doesn't not help you know the other term at all.

  If the main goal is to determine which predictors are statistically significant and how changes in the predictors relate to changes in the response variable, R-squared is almost totally irrelevant.

  A low R square doesn't negate a significant predictor or change the meaning of its coefficient. R-squared doesn't need to be any particular value to allow for a valid interpretation. Look at Result fits, rational for the model, & P values

  Formulae: R^2 =SSR / SST = 1 - SSE/SST
  Where - SSR, SSE and SST are regression, residual (error) and total sum of squares

Terms to Understand

Multiple R:

- Multiple R: **The correlation coefficient between the observed and predicted values. It ranges in value from 0 to 1.** A small value indicates that there is little or no linear relationship between the dependent variable and the independent variables.

  This is the multiple correlation coefficient which shows the strength of the relationship between the dependent variable and the independent variable(s). It would be analogous to the correlation coefficient "r" which shows the strength between two variables. It is square root of R^2, which is noted by correlation

  Multiple R is always between 0 and 1, and closer to 1 indicates stronger relation. In other words, if the value or "Multiple R" in the "Regression Statistics" is close to 1, then the least-square regression line indeed fits the data points pretty well, and there is a linear (positive or negative) relation between the two variables. It does not indicate which way the relation goes, as the correlation coefficient does.

  Multiple R is the correlation between Y and Y Mean. It ranges between 0 and 1 (it won't be negative.)

Adjusted R Square:

- Adjusted R square: The Adjusted R^2 is attempting to account for statistical shrinkage. The 'Adjusted R Square' is computed because the 'R Square' tends to somewhat over-estimate the success of the model when applied to the real world. **The adjusted R^2"penalizes" you for adding the extra predictor variables that don't improve the existing model.** It can be helpful in model selection. Adjusted R^2 will equal R^2 for one predictor variable. As you add variables, it will be smaller than R^2. R^2 increases even when you add variables which are not related to the dependent variable, but adjusted R^2 take care of that as it decreases whenever you add variables that are not related to the dependent variable. Adjusted R square gives idea of the goodness of fit measure, here it says 98.62% of Y is determined by X

  Adjusted R Square = 1 - (Total df / Residual df)(Residual SS / Total SS)

Residual:

- Residual – The residuals show you how far away the actual data points are from the predicted data points. They are the difference between the true value and the predicted value. e = y (true value) – y predicted

  Random pattern of residuals indicates a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.
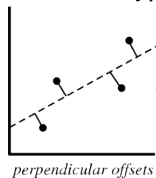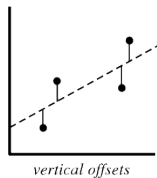
  Regression Residuals: Unless the R^2 is 100%, there will be some amount of variation in y

which remains unexplained by x. The unexplained variation is the error component of the regression equation. That error is the sum of the differences between each observed value and its value as predicted by the regression equation.

Multiple Regression is nothing but attempts to fit a line through a set of plotted points that minimizes the residuals to give us a line that is called line of best fit. This line is fit in such a way that the sum of the squared residuals is minimized.

- Note: There are two type of offsets to calculate the least distance. In practice, the vertical offsets from a line (polynomial, surface, hyperplane, etc.) are almost always minimized instead of the perpendicular offsets. This provides a fitting function for the independent variable X that estimates y for a given x (most often what an experimenter wants), allows uncertainties of the data points along the x- and y-axes to be incorporated simply, and also provides a much simpler analytic form for the fitting parameters than would be obtained using a fit based on perpendicular offsets.

vertical offsets   *perpendicular offsets*

Degree of Freedom (Df):

For a moment suppose that we know the mean of data is 25 and that the values are 20,10, 50, and one unknown value. To find the mean of a list of data, we add all of the data and divide by the total number of values. This gives us the formula $(20 + 10 + 50 + x)/4 = 25$, where x denotes the unknown. Despite calling this unknown, we can use some algebra to determine that $x = 20$.

Let's alter this scenario slightly. Instead we suppose that we know the mean of a data set is 25, with values 20, 10, and two unknown values. These unknowns could be different, so we use two different variables, x and y to denote this. The resulting formula is $(20 + 10 + x + y)/4 = 25$. With some algebra we obtain $y = 70 - x$. The formula is written in this form to show that once we choose a value for x, the value for y is determined. This shows that there is one degree of freedom.

- Df: Degree of freedom is the number of values in the final calculation that are free to vary. The number of independent ways by which a dynamic system can move without violating any constraint imposed on it. This refers to a positive whole number that indicates the lack of restrictions in our calculations. The degree of freedom is the number of values in a calculation that we can vary.

If you have a sample population of N random values then the equation has N degrees of freedom. If your data set required you to subtract the mean from each data point--as in a chi-squared test--then you will have N-1 degrees of freedom. Knowing the degrees of freedom for a population or sample does not give you much insight in of itself. Rather, the correct degrees of freedom and your chosen alpha together give you a critical value. This value allows you to determine the statistical significance of your results.

In terms of keeping score, degrees of freedom can be counted or found out as

Terms to Understand

- o A single sample: There are n observations. There's one parameter (the mean) that needs to be estimated. That leaves n-1 degrees of freedom for estimating variability.
- o Two samples: There are n1+n2 observations. There are two means to be estimated. That leaves n1+n2-2 degrees of freedom for estimating variability.
- o One-way ANOVA with g groups: There are n1+..+ng observations. There are g means to be estimated. That leaves n1+..+ng-g degrees of freedom for estimating variability. This accounts for the denominator degrees of freedom for the F statistic.
- o There are n observations with p+1 parameters to be estimated--one regression co-efficient for each of the predictors plus the intercept. This leaves n-p-1 degrees of freedom for error, which accounts for the error degrees of freedom in the ANOVA table.
- o We differentiate between these distributions by use of degrees of freedom. Here the probability distribution that we use depends upon the size of our sample. If our sample size is n, then the number of degrees of freedom is n - 1. For instance, a sample size of 22 would require us to use the row of the t-score table with 21 degrees of freedom.

Standard Error

Standard Error or Standard Error of Measurement

- If a single student were to take the same test repeatedly (with no new learning taking place between testings and no memory of question effects), the standard deviation of his/her repeated test scores is denoted as the standard error of measurement.

- It follows the concept unknown, as the value of true score can only be obtained by infinite test with no memory or new learning to affect the result of the test. It estimates how repeated measures of a person on the same instrument tend to be distributed around his or her "true" score.

- SEm is directly related to the reliability of a test; that is, the larger the SEm, the lower the reliability of the test and the less precision there is in the measures taken and scores obtained. Since all measurement contains some error, it is highly unlikely that any test will yield the same scores for a given person each time they are retested.

- We assume that each student's test score is our best estimate of the true score, but we recognize that there are sampling errors in that estimate, just as there were for estimating the population mean. Those sampling errors are normally distributed and have a standard deviation called the standard error of measurement.

Terms to Understand

- PN: the difference between the "standard deviation of scores on a test" and the "standard error of measurement on a test" – is that standard deviation is a measure of the "spread" of scores between students while It is a measure of the "spread" of scores within a student had the student been tested repeatedly.

- From the test scores of a population of students on a single exam, one can easily compute estimates of the test score mean, test score standard deviation, and the test score reliability (notated as x, s_x, and r_{xx}, respectively). From these estimates, an estimate of the standard error of measurement, s_E, is computed using the following formula:

$$s_E = s_x \sqrt{1 - r_{xx}}$$

- Where S_x is standard deviation of the test, and r_{xx} the reliability of the test
- Reliability Test is extensive, for details please follow the link
- Here we use Kuder-Richardson 21 formula for calculating reliability because it is easy to compute, relying only on the number of test items, and the average and variance of the test scores.
- The formula is KR-21 = n/n-1[1-{(X-X2/n)/S2}], where n is the number of test items, X is the average score, and S the standard deviation.

Standard Error of Estimate (SE):

- This statistic takes into account regression toward the mean and the fact that scores at the extreme end (very high or very low scores) of the distribution are more prone to error than scores near the average. Because of this fact, the standard error of estimation is not equivalent around the obtained score.

- Standard error of the estimate. The standard error of the estimate (se) indicates how close the actual observations fall to the predicted values on the regression line. S_{yx} =square Root ($\sum_1^n (y - yest)^2$/n)

  Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is N-2 rather than N. The reason N-2 is used rather than N-1 is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares.

Standard Error of Mean

Thus the standard error of the mean is the standard deviation for the distribution of errors or random fluctuations that are likely to occur in estimating the population mean from sample means in a particular situation.

However, if you were to select random samples of 50 students one after the other until you had 100 such random samples and administered the same test to all the students, you would find that the 100 samples would have many different means. If you were to

Terms to Understand

plot the means for the 100 random samples you would find that a histogram of those means would probably be normal in distribution.

It shows that the means themselves would have a mean. This mean of means, by the way, would probably be the best estimate of the population mean.

The standard deviation of such a distribution of means is referred to as the standard error of the mean because it represents the distribution of errors (or random fluctuations) in estimating the population mean.

- SM = S/ √N

Where S = Standard Deviation, N = Number of scores in test

Here it is 50, although to derive the mean we have taken 100 samples but essentially the test are 50.

So, if you had 50 students (N) on a test with a standard deviation (S) of 4.89, the standard error of the mean (SM) would turn out to be as follows:

$4.89/\sqrt{50} = 6.917$

At a Glance

| Standard deviation	Standard error of the mean	Standard error of measurement	Standard error of estimate
Helps to estimate the dispersion in a given distribution	Estimate the dispersion of sampling errors when you are trying to estimate the population mean from a sample mean	Estimate the dispersion of the measurement errors from the closest true score	Estimates dispersion of prediction errors when you are trying to predict Y values from X values in a regression analysis
	Here we use mean of the mean	Here we assure that true scores can never be found	

Normal Distribution

- These are symmetric in nature and peak at the mean, with the probability distribution decreasing away before and after this mean smoothly, as shown in the figure below.

Terms to Understand

Statistics Involving Z

Confidence Interval

- A confidence interval is how much uncertainty there is with any particular statistic. a 95% confidence interval covers 95% of the normal curve -- the probability of observing a value outside of this area is less than 0.05 (the decimal difference between 100% and 95%). This difference is commonly referred to as α (alpha).

- As normal curve is symmetric, half of the area is in the left tail of the curve, and the other half of the area is in the right tail of the curve. For a confidence interval with a Confidence level C, the area in each tail of the curve is equal to (1 - C)/2 or α /2.

- Therefore for a 95% confidence level, the area in each tail is equal to 100%-95% = 5% = (5/100)/2 = 0.05 / 2 = 0.025.

- Since there are two "tails", the central area is always 1 - 2(tail area), and the tail area is always 0.5(1 – central area).

- This means that all the area to the left must be 1– 0.025 or 0.975. This value is used to determine $z_{\alpha/2}$.
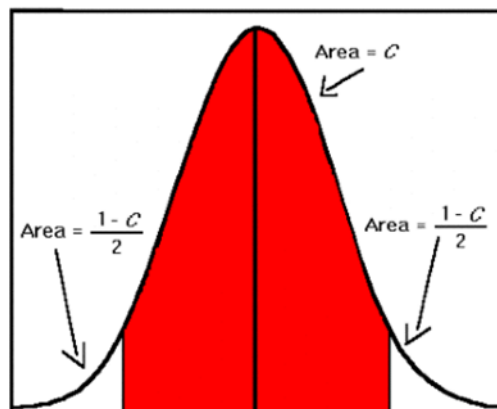
Confidence Level:

- The selection of a confidence level for an interval determines the probability that the confidence interval produced will contain the true parameter value.

- In survey with 95% Confidence level, the age of students in class 7 in between 10 to 12. It means that 95% of the students in class 7 belong are in the age range of 10 to 12. Confidence level or intervals are often used with a margin of error.

Terms to Understand

Z-Score

- The standard score is the number of standard deviations an observation is above or below the mean.

- The Z value provides an assessment of how off-target a process is operating. And what would be the percentage of the population inside the accepted level or Confidence Level

- It is obtained by **subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation**. This conversion process is called **standardizing or normalizing** (though normalizing may have other methods too). These are called z-values, z-scores, normal scores, and standardized variables; the use of "Z" is because the normal distribution is also known as the "Z distribution".

    - Z score = (sample value - sample mean)/standard deviation.

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

μ is the mean of the population;

σ is the standard deviation of the population.

- For example if we want to know what would be the percentage of population that would fall under the curve, here is the method

- Calculate z, Please refer to the Z Score Table
- If z=1.67, then you would split this number into 1.67 = 1.6 + .07. The number located in the 1.6 row and .07 column is .953. Thus 95.3% of the area under the bell curve is to the left of z=1.67.

Terms to Understand

Normal, Bell-shaped Curve

Percentage of cases in 8 portions of the curve	.13%	2.14%	13.59%	34.13%	34.13%	13.59%	2.14%	.13%	
Standard Deviations	-4σ	-3σ	-2σ	-1σ	0	+1σ	+2σ	+3σ	+4σ
Cumulative Percentages		0.1%	2.3%	15.9%	50%	84.1%	97.7%	99.9%	
Percentiles			1	5 10 20 30 40 50 60 70 80 90 95			99		
Z scores	-4.0	-3.0	-2.0	-1.0	0	+1.0	+2.0	+3.0	+4.0
T scores		20	30	40	50	60	70	80	
Standard Nine (Stanines)		1	2	3 4	5	6	7 8	9	
Percentage in Stanine		4%	7%	12% 17%	20%	17%	12% 7%	4%	
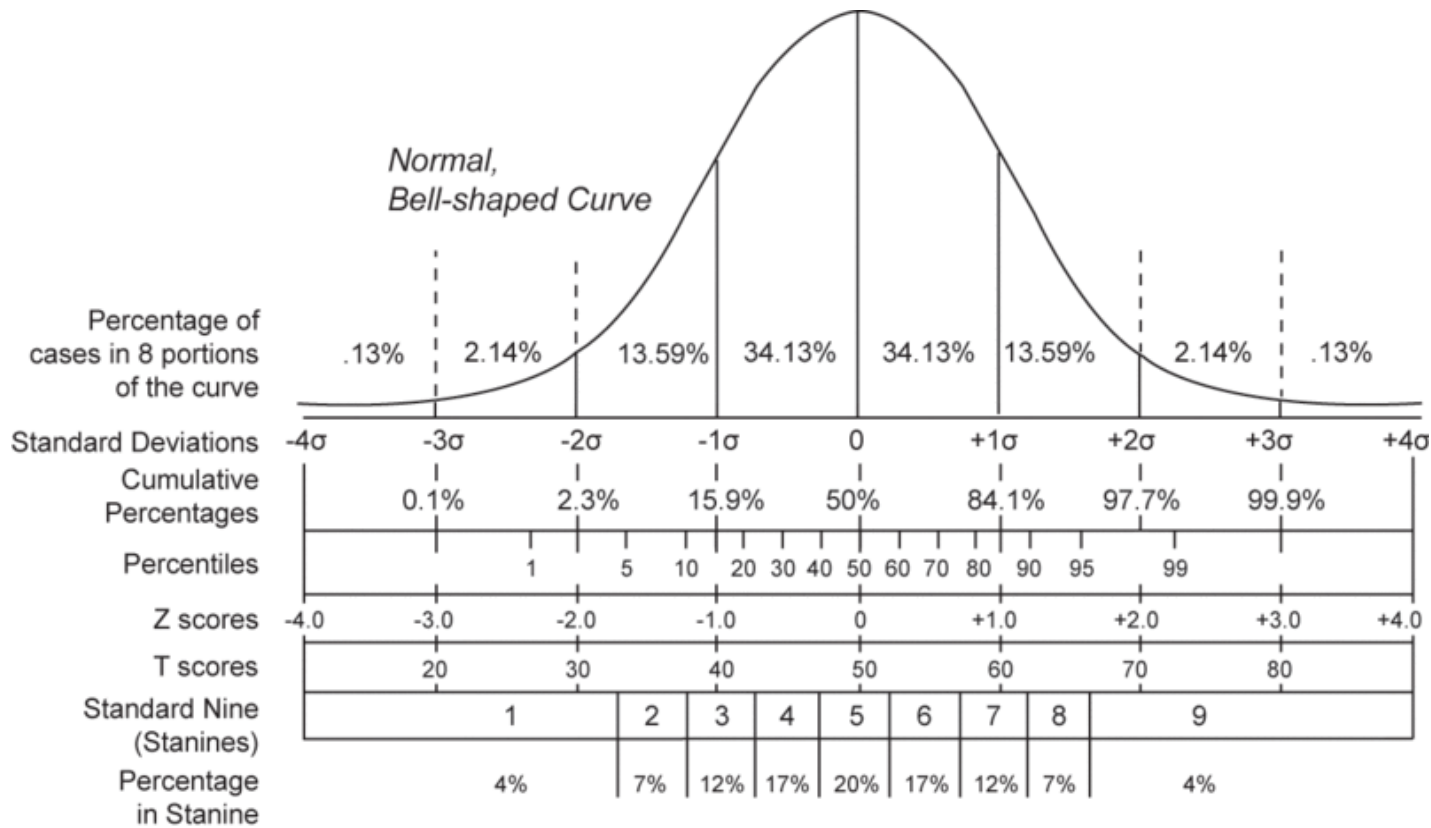
Critical Value

The critical value(s) for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected.

The critical value for any hypothesis test depends on the significance level at which the test is carried out, and whether the test is one-sided or two-sided.

To test hypotheses, cut off number is decided in advance. This number is sometimes called the **critical** or **tabled value** because it is looked up in a table, associated with significant level. If the computed test statistic has a smaller probability than that of the critical value, the null hypothesis will be rejected.

The critical region CR, or rejection region RR, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test.

That is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H_0, the other will not. If the observed value of the test statistic is a member of the critical region, we conclude "Reject H_0"; if it is not a member of the critical region then we conclude "Do not reject H_0".

Terms to Understand
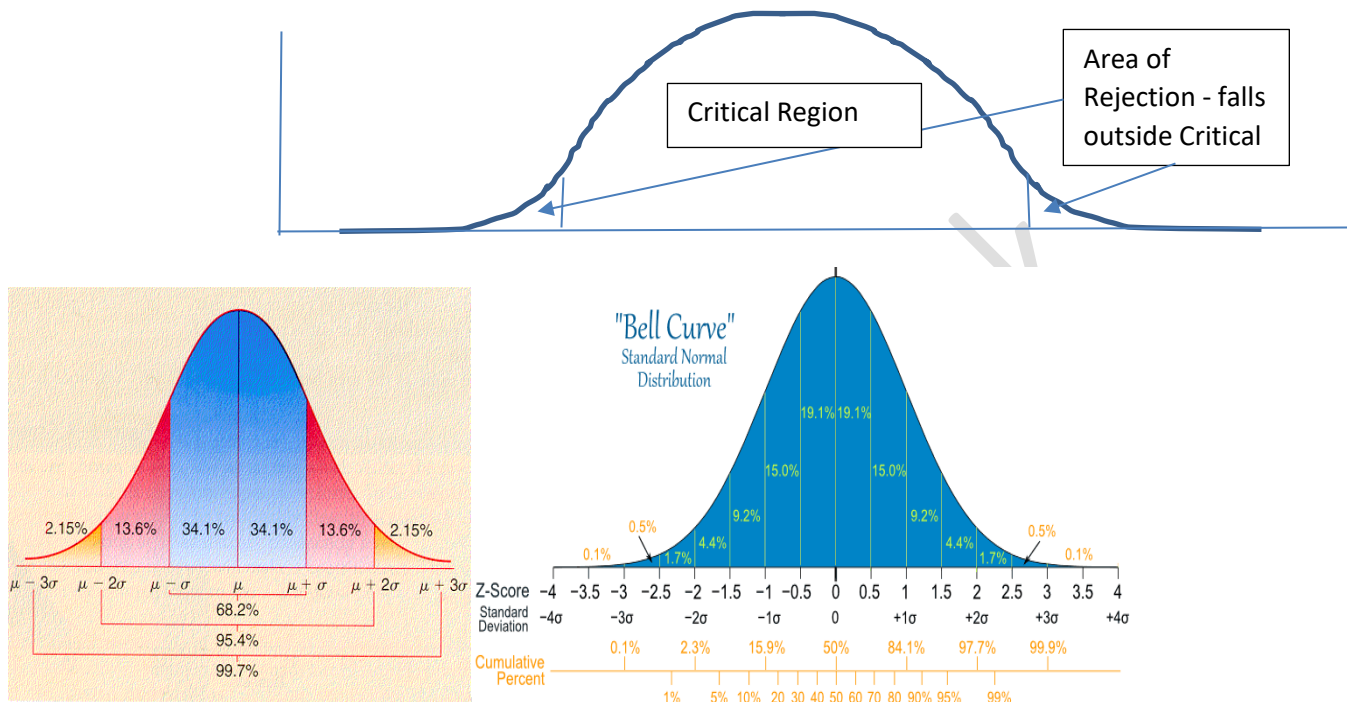
Critical Z

- A(z) is the integral of the standardized normal distribution from −∞ to z. It gives the probability of a normal random variable not being more than z standard deviations above its mean.

- Find out Critical Value of Z
    - For Two Tailed Test for 95% confidence level => α = 100%-95% = 5/100 = 0.05.
        - Subtract α from 1 => 1- α => 1-0.05 = 0.95
        - Since it is two tailed – both the tails fall on either side, so the number is to be divided by 2
        - 0.95/2 => .475
        - From the table below search the value 0.475
        - The addition of Row and Column gives value of Critical Z
        - Here the Row that contains 0.475 is 1.9 and Column is 0.06 = so Critical Z is - 1.96 and +1.96
    - For Left One Tailed Test for an alpha level of 0.012,
        - Confidence Level is (100 – Alpha) = (100 – 0.012)= 99.98%
        - Sanity Check = 100% - 99.98% = CI
        - Subtract α from 0.5 => 0.5 - α => 0.5 - 0.012 = 0.488 (it is 1 tailed it is NOT divided by 2)
        - The addition of Row and Column gives value of Critical Z,
        - In absence of exact match take the closest towards left

- Here the Row that contains 0.4921 is 2.2 and Column is 0.06 = so Critical Z is – 2.26
    - For Right One Tailed Test for an alpha level of 0.0079,
        - Confidence Level is (100 – Alpha) = (100 – 0.0079)= 99.9921%
        - Sanity Check = 100% - 99.9921% = CI
        - Subtract α from 0.5 => 0.5 - α => 0.5 - 0.0079 = 0.4921 (1 tailed so it is NOT divided by 2)
        - The addition of Row and Column gives value of Critical Z,
        - In absence of exact match take the closest towards right
        - Here the Row that contains 0.4921 is 2.4 and Column is 0.02 = so Critical Z is 2.42, Please refer to Z Critical Table here

Z Test

Z-test is not the same as a Z-score. The Z-test is where you take the Z-score and compare it to a critical value to determine if the null hypothesis will be rejected or fail to be rejected.

- This statistic is a measure of the likelihood that the actual value of the parameter is not zero. The larger the absolute value of t, the less likely that the actual value of the parameter could be zero.

- These three approximations yield three different z tests of the hypothesis that $\mu = \mu_0$ at approximate significance level a:

    - Reject the null hypothesis whenever $Z < z_a$ (left-tail z test)
    - Reject the null hypothesis whenever $Z > z_{1-a}$ (right-tail z test)
    - Reject the null hypothesis whenever $|Z| > z_{1-a/2}$ (two-tail z test)

T-Stats

We have been using a test statistic, $Z_{\bar{x}}$, that allows us to make an inference about whether our sample is different from some population mean. In order to use $Z_{\bar{x}}$, we had to know four things,

- the population mean
- and standard deviation,
- our sample mean, and
- our sample size.

As you know, however, we typically do not know the **population standard deviation. Instead, we pick a sample from the population and test whether the sample mean is significantly different from some other value**. To conduct this sort of test we compute what we refer to as the one-sample **t statistic**, or *t*-test. The *t* is pretty similar to the *z*, in fact, it's almost identical.

The only difference is that we have to estimate the population standard deviation, σ. Remember, if you know σ, then use the *z*-test; if you don't know σ, then estimate σ (find S)

Terms to Understand

The formula is
$$s = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$$.

Our estimate uses $n - 1$ in the denominator. To calculate the t-test, we calculate the standard error of

the estimate, $s_{\overline{X}} = \frac{s}{\sqrt{n}}$, and use the formula $t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$.

Similarity between the z-test and the t-test. The only difference is that in the z-test we use σ, and in the t-test we use s.

- The t statistic allows researchers to use sample data to test hypotheses about an unknown population mean. The particular advantage of the t statistic, is that the t statistic does not require any knowledge of the population standard deviation.

- The t statistic can be used to test hypotheses about a completely unknown population; that is, both μ and σ are unknown, and the only available information about the population comes from the sample.

- In statistics, the t-statistic is a ratio of the departure of an estimated parameter from its notional value and its standard error.

- The larger the value of the t-stat, the further the distance, or number of standard deviations. It helps to determine if the null hypothesis (that sample observations result purely from chance) were true.

It is used in hypothesis testing. The t statistic is the coefficient divided by its standard error. Thus, it measures "how many standard deviations from zero" the estimated coefficient is, and it is used to test the hypothesis that the true value of the coefficient is non-zero. Non-zero is checked in order to confirm that the independent variable really belongs in the model. The coefficient being referred here is correlation coefficient ρ for the population. This is explained by the following example:

Suppose we want to evaluate whether or not a linear relationship exists between a husband's age and his wife's age? Correlation Coefficient is 0.939. We determine existence of a linear relationship by using the t-test for testing the population correlation coefficient $H0$: $\rho = 0$. We follow standard hypothesis test procedures in conducting a hypothesis test for the population correlation coefficient ρ. First, we specify the null and alternative hypotheses:

Null hypothesis H_0: $\rho = 0$ (corresponding items are not significant)
Alternative hypothesis H_A: $\rho \neq 0$ or H_A: $\rho < 0$ or H_A: $\rho > 0$

This value is the probability of obtaining the estimated value of the parameter if the actual parameter value is zero. The smaller the value of t, the more significant the parameter and the less likely that the actual parameter value is zero.

Terms to Understand

$$t = \frac{\text{obtained difference}}{\text{standard error of Mean}} = \frac{M - \mu}{sM}$$

- The top of the ratio contains the obtained difference between the sample mean and the hypothesized population mean. The bottom of the ratio is the standard error which measures how much difference is expected by chance.

- For example, assume the estimated value of a parameter is 1.0 and its standard error is 0.7. Then the t value would be 1.43 (1.0/0.7). If the computed t value was 0.05 then this indicates that there is only a 0.05 (5%) chance that the actual value of the parameter could be zero.

- If t was 0.001 this indicates there is only 1 chance in 1000 that the parameter could be zero. If t was 0.92 this indicates that there is a 92% probability that the actual value of the parameter could be zero; this implies that the term of the regression equation containing the parameter can be eliminated without significantly affecting the accuracy of the regression.

t statistic as an "estimated z-score." The estimation comes from the fact that we are using the sample variance to estimate the unknown population variance. With a large sample, the estimation is very good and the t statistic will be very similar to a z-score. With small samples, however, the t statistic will provide a relatively poor estimate of z.

- The value of **degrees of freedom**, df = n - 1, is used to describe how well the t statistic represents a z-score. Also, the value of df will determine how well the distribution of t approximates a normal distribution.
- For large values of df, the **t distribution** will be nearly normal, but with small values for df, the t distribution will be flatter and more spread out than a normal distribution.

To evaluate the t statistic from a hypothesis test, you must select an α level, find the value of df for the t statistic, and consult the t distribution table.

- If the obtained t statistic is larger than the critical value from the table, you can reject the null hypothesis.
- The T statistic tests the hypothesis that a population regression coefficient is 0, ie when all the predictors are in the model. It is the ratio of the sample regression coefficient to its standard error.
- The t-test produces a single value, t, which grows larger as the difference between the means of two samples grows larger;
- t does not cover a fixed range such as 0 to 1 like probabilities do
- You can convert a t-value into a probability, called a p-value
- The t value is compared with Critical T for the degree of freedom to accept or reject a Hypothesis

Terms to Understand

Critical T

A T-critical value is used in statistics to construct a confidence interval, a range of values that is likely to include a parameter being studied. To find a critical value, you need to know the alpha value and degrees of freedom.

- Compare tcalc to the tabulated t-value, for the appropriate significance level and degree of freedom.

- If tcalc > tcritical, we reject the null hypothesis and accept the alternate hypothesis. Otherwise, we accept the null hypothesis.

- PN: Degree of Freedom = Sample Size – 1

- Please refer to the Critical T table

- If your t-value is greater than or equal to the critical value in the table, then you can conclude that your p-value is less than the significance level you have chosen. The procedure is simple. Once you have chosen your significance level you look to see whether your t-value is larger than the critical t-value shown in the column relating to your chosen significance level. If it is, then you can say that p is less than your chosen significance level. The p-value for a two tailed test is twice what it would be for a one tailed test.

Effect Size

It is recommended that a measure of effect size be computed along with the hypothesis test. With a t test it is also possible to measure effect size by computing the **percentage of variance accounted.**

- **W**e obtain a measure of the size of the treatment effect. For the t statistic hypothesis test,

$$\text{percentage of variance accounted for } = r^2 = \frac{t^2}{t^2 + df}$$

Margin of Error:

- The value or the data out of the Confidence level or below and above the sample statistic in a confidence interval is called the margin of error.

  Margin of Error Percentage is percentage the results will differ from the real population value. The Margin of Error can be calculated in two ways:

  o Margin of error = Critical value x Standard deviation of the statistic

Terms to Understand

o Margin of error = Critical value x Standard error of the statistic

Sample Size

Sample size is important because too large may waste time, resources and money, while samples that are too small may lead to inaccurate results.

Sample Size = (Z-value ^2 * (percentage picking a choice)*(1- percentage picking a choice)) / (Margin of error)^2

Sum of Square

- SS - Sum of Squares represents variation from several sources. It is used for Total, Residual or Error and Regression. The basic regression line concept, DATA = FIT + RESIDUAL, is rewritten as follows:

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i).$$

The first term is the total variation in the response y, the second term is the variation in mean response, and the third term is the residual value. Squaring each of these terms and adding over all of the n observations gives the equation

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2.$$

This equation may also be written as SST = SSM + SSE, where SS is notation for sum of squares and T, M, and E are notation for total, model, and error, respectively.

Total SS

- Total SS is the **variation within the values of y**. It is the sum of the squares of the differences between values of y and the average y.

Formulae = (n-1)*(standard deviation of y)^2
Total SS = Regression SS + Residual SS

Residual SS or Error SS

- Residual SS = sum of squares of the **differences between the values of y predicted by equation and the actual values of y.** In other words, it is derived from the cumulative addition of the square of each residual, where a residual is the distance of a data point above or below the fitted line

If the data exactly fit equation 1, then Residual SS would be 0 and R^2 would be 1.

Terms to Understand

Sum of Sq of diff between Predicted & Actual y
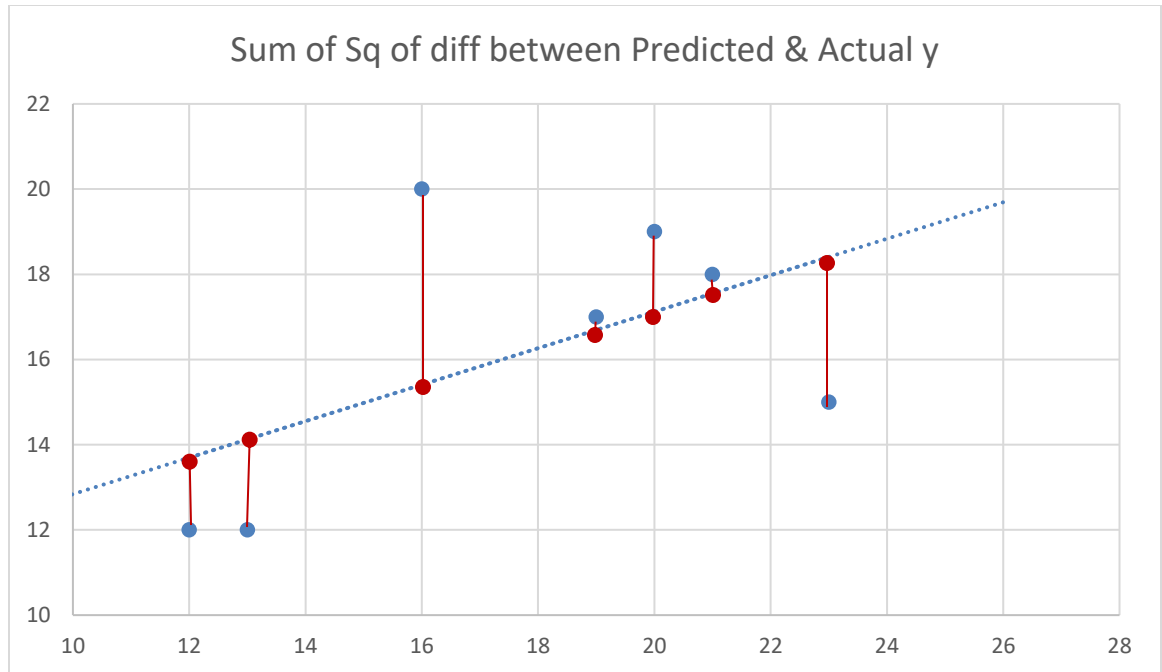
Regression SS

- Regression SS describes the variation within the fitted values of y, and is the sum of the squared difference between each **fitted value of y and the mean of y**. The squares are taken to 'remove' the sign (+ or -) from the residual values to make the calculation easier.

  Regression SS = Total SS - Residual SS

Sum of Sq of diff between Predicted & Mean of Actual y
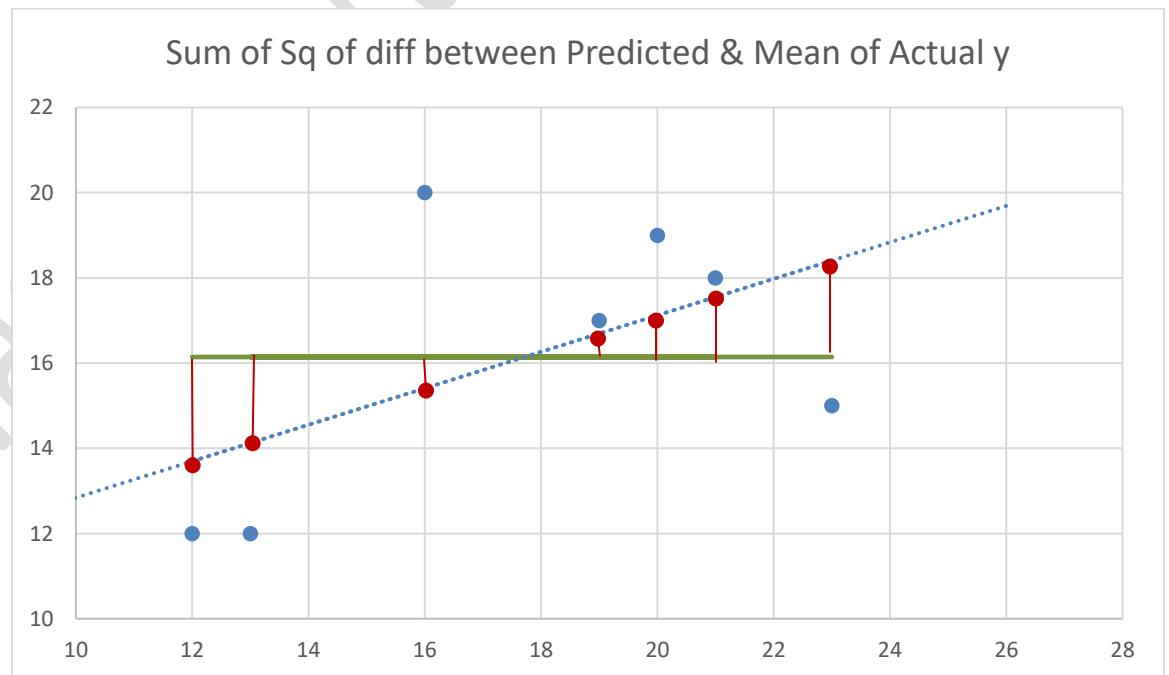
  o
  o   Mathematical formulae

Terms to Understand

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

Total sum of squares

Regression (explained) sum of squares

Residual (unexplained) sum of squares

Computation of sums of squares	
Case	*Formula(s)*
General case:	$SST = \sum_{j=1}^{N}(y_j - \bar{y})^2 = SSR + SSE$ $SSR = \sum_{j=1}^{N}(\hat{y}_j - \bar{y})^2 = SST - SSE$ $SSE = \sum_{j=1}^{N}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{N}e_j^2 = SST - SSR$

Mean Square

- MS: Mean Square – It is some estimate of the variance based on certain sources of variation available to us in our experiment.

  Reason for analysis of variance is to compare the ratio of between group variance to within group variance. The variance is linked to degree of freedom as Df is nothing but the number of values a data set can vary.

  The total variation (not variance) is comprised the sum of the squares of the differences of each mean with the grand mean. There is also variation the between group variation and the within group variation. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom. Within groups. Each sample is considered independently, no interaction between samples is involved. The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k.

Types of MS

- Total
- Between
- Within

Terms to Understand

MS can be estimate of the variance one of 3 ways:

- ○ from the deviation from the grand mean (SS total); It is also called total variation. MS total=SS/N-1
- ○ from the deviation from group means from the grand mean (SS between); MS between=SS/k-1
- ○ from the pooled variation of S scores around their respective group means (SS within); MS within=SS/N-k

Mean square is simply some particular sum of squares divided by its appropriate degrees of freedom and is used for both residual and Regression.

Regression MS

Regression MS= Regression SS / Regression df and

Residual MS

Residual MS = mean square error = Residual SS / Residual df

The residual mean square is a measure of how poorly or how well the regression line fits the actual data points. A large residual mean square indicates poor fit. It has to be checked with F test to derive any conclusion on Hypothesis testing.

F Test

An F test is used to determine if the relationship can be generalized to the population represented by the sample.

Another method of determining the best model for prediction is to test the significance of adding one or more variables to the model using the *partial F-test*. In general, the *partial F-test* is similar to the *F-test* used in analysis of variance. It assesses the statistical significance of the difference between values for R^2 derived from 2 or more prediction models using a subset of the variables from the original equation.

Significant F

There is also a **significance level for the model as a whole**. This measures the likelihood that the model as a whole describes a relationship that emerged at random, rather than a real relationship.

- **The lower the significance F value, the greater the chance that the relationships in the model are real.**

- This indicates the probability that the Regression output could have been obtained by chance. A small Significance of F confirms the validity of the Regression output. For example, if Significance of F = 0.030, there is only a 3% chance that the Regression output was merely a chance occurrence.

Terms to Understand

This tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable.

Significance F gives us the probability at which the F statistic becomes 'critical', ie below which the regression is no longer 'significant'.

This is calculated (as explained in the text above) as =FDIST(F-statistic, 1, T-2), where T is the sample size. In this case, =FDIST(9.126559714795,1,8) = 0.0165338014602297

If F has a value of 0.01000 then there is 1 chance in 100 that all of the regression parameters are zero. This low a value would imply that at least some of the regression parameters are nonzero and that the regression equation does have some validity in fitting the data (i.e. the independent variables are not purely random with respect to the dependent variable

F Value or Ratio

The F Value or **F ratio is the test used to decide whether the model as a whole has statistically significant predictive capability**. F value tests the full model against a model with no variables and with the estimate of the dependent variable being the mean of the values of the dependent variable. It is a part of Test statics and F Test

Under the null hypothesis that the model has no predictive capability, which means that all the results are purely due to chance and they cannot be predicted. **The null hypothesis is rejected if the F ratio is large**. High F ratio or Value means predictions can be done.

F ratio tests whether the regression SS is big enough, considering the number of variables needed to achieve it. Its value will range from zero to an arbitrarily large number.

- The F value = mean regression sum of squares/ mean error sum of squares.

Larger an F-value indicates a consistent pattern that is unlikely due to chance. It also indicates that the model is robust and has significant effect (the denominator which has error is lesser). The simple rule in most research situations is: the higher the F value, the better…

- F = Effect Variance (or "Treatment Variance") / Error Variance Or,
- F = Between-group Variance / Within-group Variance
- F = Explained Variance / Unexplained Variance

F value is used with p value for interpretation.

- Larger R^2 produce bigger values of F. That is, the stronger the relationship is between the DV and the IVs, the bigger F will be.

- Larger sample sizes also tend to produce bigger values of F. The larger the sample, the less uncertainty there is whether population parameters actually differ from 0. Conversely, in a small sample, even large effects may not be statistically significant.

If additional variables do not produce large enough increases in R2, then putting them in the model can actually decrease F. It may be difficult to detect important effects. The F statistic

Terms to Understand

does not tell you which effects are significant, only that at least one of them is. FR could be also used to compare two models describing the same experimental data: the higher FR the more adequate the corresponding model.

A very large F-value means that the between-group variance (the effect variance) exceeds the within-group variance (the error variance) by a substantial amount. In such cases p-value then just gives a number to how likely a particular F-value is going to occur, with lower p-values indicating that the probability of obtaining that particular F-value is pretty low. This is actually why the degrees of freedom influence the F distribution.

We will cover more of this when we create model and interpret results.

F-Critical Value

First, use the Critical Significance Level (α: alpha) chosen in Step 2 and the between treatment (df_1= number of treatments – 1) and within treatment (df_2= total sample size – number of treatments) degrees of freedom calculated in Step 3 to find the Critical Value of F ($F_{critical}$) using a Critical Value Table such as the one below e.g., if α = 0.05, df1 = 2 and df2 = 12 then $F_{critical}$ = 5.096.

Table of Critical Values for a Critical Significance Level (α: alpha) = 0.05 for the F statistic where df_1 is the degrees of freedom between treatments (the numerator; the number of treatments - 1) and df_2 is the degrees of freedom within treatments (the denominator; the total sample size – number of treatments) for Anova.

- F-critical value is the F value above which 100% of the null sampling distribution occurs. This type of logic is used in many other types of statistical tests, which compare averages and other measures, instead of variances.

  Note that this value can be obtained from a computer before the experiment is run, as long as we know how many subjects will be studied and how many levels the explanatory variable has. Then when the experiment is run, we can calculate the observed F-statistic and compare it to F-critical. If the statistic is smaller than the critical value, we retain the null hypothesis because the p-value must be bigger than alpha, and if the statistic is equal to or bigger than the critical value, we reject the null hypothesis because the p-value must be equal to or smaller than alpha

  If Fcalculated > Fcritical, H0 is rejected.
  If Fcalculated < Fcritical, H0 cannot be rejected.

- **Compare your f-value with your f-critical value. If the f-critical value is smaller than the f-value, you should reject the null hypothesis.**

- Please refer to F Critical Table

P - Value

- The p-value is a percentage. It tells you how likely it is that the coefficient for that independent variable emerged by chance and **does not** describe a real relationship.

Terms to Understand

- It varies between 0 and 1 and can never be negative. A p-value of .05 means that there is a 5% chance that the relationship emerged randomly and a 95% chance that the relationship is real.

- It is generally accepted practice to consider variables with a p-value of less than .1 as significant, though the only basis for this cutoff is convention.

- A P value can be defined as the actual probability associated with the obtained value of Z. It is a measure of how unusual or rare our obtained statistic is compared with what is stated in our null hypothesis. **The smaller the P value, the more evidence we have that the null hypothesis should be rejected in favor of the research hypothesis.**

- P value calculate the probability that a sample statistic takes values left (in left tailed test) or right to the observed value (in right tailed test), but both tails in two tailed test.

- Therefore in two-tailed test, the P-value is twice of the value found in one tailed test.

- The difference between P and alpha is that P is the actual probability associated with the obtained value of Z, whereas alpha is the level of probability determined in advance at which the null hypothesis is rejected. The null hypothesis is rejected when $P \leq a$.

- Small p values provide evidence against the null hypothesis because they say the observed data are unlikely when the null hypothesis is true. We apply the following conventions:

  - When p value > .10 → the observed difference is "not significant"
  - When p value ≤ .10 → the observed difference is "marginally significant"
  - When p value ≤ .05 → the observed difference is "significant"
  - When p value ≤ .01 → the observed difference is "highly significant"

- Use of "significant" in this context means "the observed difference is not likely due to chance." It does not mean of "important" or "meaningful.

- Based on the P value, we can also make a statement regarding the "significance" of the results. If the P value is equal to or less than our alpha level, our obtained Z statistic is considered statistically significant—that is to say, it is very unlikely to have occurred by random chance or sampling error.

- There are more than 1 method to find P-value like Second Method: Fisher's Exact, Chi-Squared Test and Using Z-Score. The easiest method is using Z-Score.

- If a P-value is greater than the applied level of significance, and the null hypothesis should not just be blindly accepted. Other tests should be performed in order to determine the true relationship between the variables which are being tested.

- While α =0.05 is standard, a p-value of 0.06 should be examined for clinical importance.

- P-values are computed based on the assumption that the null hypothesis is true. The p-value is the probability that the data could deviate from the null hypothesis as much as

Terms to Understand

they did or more. Consequently, the p-value measures the compatibility of the data with the null hypothesis, not the probability that the null hypothesis is correct

- The method for finding the P-Value is actually rather simple. First calculate the z-score and then look up its corresponding p-value using the standard normal table.

- Please Refer to P Value Table

Relation between P – Value And Level of Significance

- If Z score is 1.13, the p-value is 0.87076

- Example – with Alpha = 0.05, Confidence level of 95%,

- Population Mean is 2.86, Say our Sample Mean is 3.11, Sample Size is 100, Standard Deviation = 0.17,
  If Z = (3.11-2.86)/0.17 = 1.47

  - For Left Tailed Test Z value is +1.47 then p-value is .92922
  - For Right Tailed Test Z value is -1.47 then p-value is 0.0778
  - For Two Tailed Test Z = Value = +1.47 to -1.47 = 0.92922 – 0.0778 = 0.85142
    - Subtract this value from 1 to get p value = 1-0.85142 = 0.14858
- The null hypothesis is rejected when P ≤ a. Here a is 0.05
  - Since in any of the Test (Left, Right or Two Tailed as per the requirement) P>a, so we cannot reject the null hypothesis.

Upper and Lower Limit 95%

This evaluating the Fitness of the Model Using Confidence Intervals. This helps us to determine how accurate our model is.

- To find the accuracy of the model, we find confidence interval for the slope of our regression line.

Here it explains (at say x level of confidence), we are confident with x% that that the value of predictor will be within lower and upper bound depicted.
In Excel, generally it is taken to be 95%

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Up
Intercept	80.27609682	27.30507654	2.9399697	0.0322578	10.08616307	150.
This is the label attached to the x variable	This is the slope of the regression line or b_1.	This is the standard error of the coefficient which we've called s_{b1}	t statistic is $b_1 \div s_{b1}$. It has n-2 degrees of freedom	This is the p-value of the hypothesis test H_0: $\beta_1 = 0$. To reject it is to conclude that there is a significant relationship between x and y. Note that it is also the p-value of the test for the correlation coefficient H_0: $\rho = 0$.	These numbers are the a confidence interval f slope of the regression are 95% confidence th $\leq \beta_1 \leq 150.47$.	

- If the confidence interval contains 0, then we have significant evidence to believe that there is no linear relationship between x and y
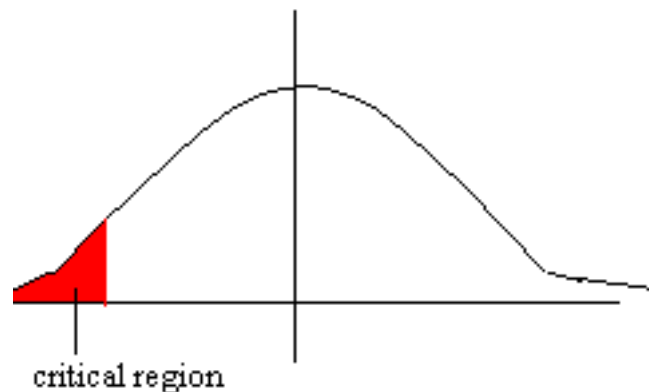
Terms to Understand

- The (1-α) confidence interval is calculated using: (point estimate) ± (t-critical value)(standard error)

- These ranges allow us to judge whether the values of the coefficients are different from zero at the given level of confidence.
- Calculation:
  - Upper value = (co-efficient) + (t-critical value)(standard error)
  - Lower value = (co-efficient) - (t-critical value)(standard error)
  - Generally here two tailed test is used with alpha of 0.05%, and for fdegree of freedom use n-2 instead of n-1

Type of Test

- The type of test is determined by the *Alternative Hypothesis* (H_1)

One-Sided Test

- A one-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution. The critical region for a one-sided test is the set of values less than the critical value of the test, or the set of values greater than the critical value of the test.
- It uses an alternate hypothesis that states either H_1: $\mu > \mu_0$ **OR** H_1: $\mu < \mu_0$, but not both.
- If our sample value lies in this region, we reject the null hypothesis in favour of the alternative.
- Suppose we are looking for a definite decrease. Then the critical region will be to the left. Note, however, that in the one-tailed test the value of the parameter can be as high as you
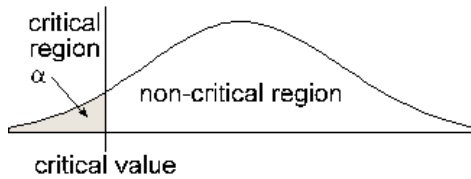

critical region

like.

Left Tailed Test

H_1: parameter **<** some specific value
The inequality points to the left, Decision Rule: Reject H_0 if t.s. < critical value
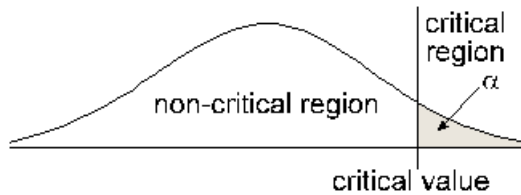
Terms to Understand


Right Tailed Test

H_1: parameter > some specific value
The inequality points to the left, Decision Rule: Reject H_0 if t.s. > critical value


Two-Sided Test

- A two-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located in both tails of the probability distribution. The critical region for a two-sided test is the set of values less than a first critical value of the test and the set of values greater than a second critical value of the test.
- It uses an alternate hypothesis that states either H_1: $\mu <= \mu_0$ **OR** H_1: $\mu >= \mu_0$.
- H_1: parameter <> value
  The inequality points to the left, Decision Rule: Reject H_0 if t.s. > critical value or Reject H_0 if t.s. < critical value

- Therefore, the critical values for α = 0.05 in one-tailed test are the same as that for α = 0.1 in two-tailed test.
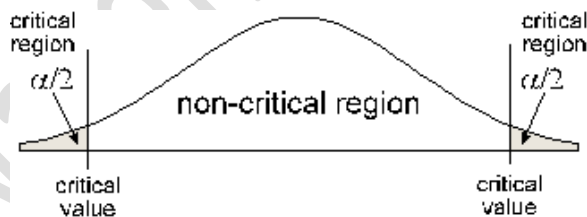
Determine which test to use

- If the question referring to the test statistic has a significant decrease from the expected it is a left tail question.
- If the question referring to the test statistic has a significant increase from the expected it is a right tail question.
- If the question referring to the test statistic has a change either more or less from the expected it is a two tail question

Terms to Understand
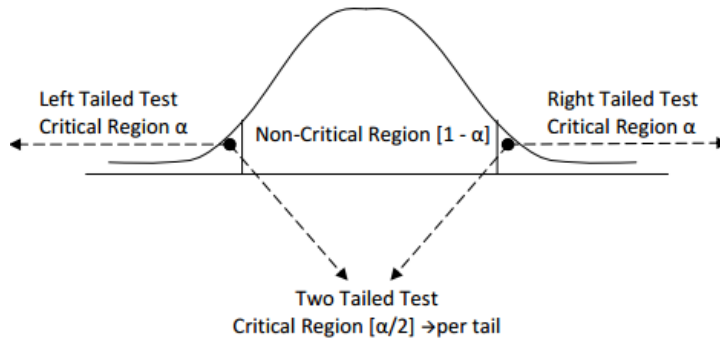
Here we come across new terms, Critical Region, Z score, Z Critical and Alpha or significance level for calculation.

The z-score needed to reject H_0 is called the critical value for significance. The critical value depends on the significance level, which we state as α. Each type of alternative hypothesis has its own critical values:

For example: The price of petrol in California is same as that of the nation. The real question we are asking is that Is it possible that in the population there is no real difference between the mean price of gas in California and the mean price of gas in the nation which is $2.86

Our H_0 = Petrol Prices in California = Nation mean = $c.
H_0 = 2.86
We can formulate our H_1 on basis of our requirement.

- If we want to test if the prices in California is higher than nation mean, we formulate
    - H_1 = California Petrol Prices > $2.86. This is Right Tail Question
- If we want to test if the prices in California is lower than nation mean, we formulate
    - H_1 = California Petrol Prices < 50. This is Left Tail Question
- If we want to test if the prices in California not equal to nation mean, we formulate
    - H_1 = California Petrol Prices <> 50. This is Two Tail Question
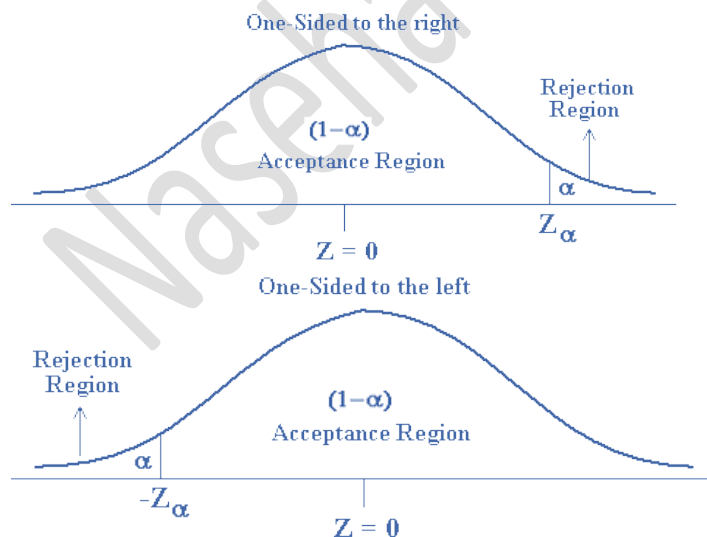
Terms to Understand

Using Tables for One- and Two-Tailed Tests

- Some tables of critical t-values only give you the values for either a one- or two-tailed test, but not both. Because of this, you will need to know how to use one-tailed tables for two-tailed tests, and vice versa. The conversion is actually quite simple:

Table you have	Operation	To get ...
One-tailed	Divide P by 2	Two-tailed test for $P/2$
Two-tailed	Multiply P by 2	One-tailed test for $2P$

- For example, assume you have a table to a one-tail test at the 98% confidence level and want to perform a two-tailed test. For the 98% confidence level, $P = 0.02$. Divide P by 2 to get 0.01, which is a 99% confidence level. So you would compare $t$calc to the value from the 98% one-tailed table, and it would be equivalent to a two-tailed test at the 99% confidence level. Will come back to this when we look at Critical Values.

Hypothesis Testing

- Refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. Hypothesis testing could be differentiated in terms of range.

Simple Hypothesis

- A simple hypothesis is a hypothesis which specifies the population distribution completely.

Composite Hypothesis

- A Composite hypothesis is a hypothesis which does not specifies the population distribution completely.

Hypothesis Steps Tests

Formal process to determine whether to reject a null hypothesis, based on sample data is called hypothesis testing. There are 4 steps of Hypothesis testing.

- Making Assumptions – Ground work
- State the hypotheses.
- Set the criteria for a decision.
- Compute the test statistic.
- Make a decision.

Terms to Understand

Making Assumption – Ground Work

- Statistical hypothesis testing involves making several assumptions regarding the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size.
- We have already covered How to determine Sample Size, Sampling Method, Critical Value, Type of Test,

State the hypotheses

In other words, hypothesis testing is a theory has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. The question is simplified into two competing hypotheses H_0 & H_1.

- **Null hypothesis.** The null hypothesis, denoted by H0, is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis**. The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Special consideration is given to the null hypothesis. The experiment has been carried out in an attempt to disprove or reject null hypothesis, thus we give that one priority so it cannot be rejected unless the evidence against it is sufficiently strong. The only reason we are testing the null hypothesis is because we think it is wrong.

In a courtroom, since the defendant is assumed to be innocent (this is the null hypothesis so to speak), the burden is on a prosecutor to conduct a trial to show evidence that the defendant is not innocent. In a similar way, we assume the null hypothesis is true, placing the burden on the researcher to conduct a study to show evidence that the null hypothesis is unlikely to be true.

Hypothesis test can have one of two outcomes: Accept the null hypothesis or Reject the null hypothesis. Acceptance implies that the null hypothesis is true. Failure to reject implies that the data are not sufficiently persuasive for us to prefer the alternative hypothesis over the null hypothesis.

Null Hypothesis:

The null hypothesis, H_0, represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.

H_0: there is no difference between the two drugs on average.

Null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if / when the null is rejected. The final conclusion is

stated with respect to the null hypothesis. We either "Reject H_0 in favor of H_1" or "Do not reject H_0"; we never conclude "Reject H_1", or even "Accept H_1".

If we conclude "Do not reject H_0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_0 in favor of H1. Rejecting the null hypothesis then, suggests that the alternative hypothesis **may** be true.

Alternative Hypothesis

The alternative hypothesis, H_1, is a statement of what a statistical hypothesis test is set up to establish by disproving the null hypothesis. It directly contradicts a null hypothesis. The alternative hypothesis states what we think is wrong about the null Hypothesis. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write

- H_0: there is no difference between the two drugs on average.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write

- H_1: the new drug is better than the current drug, on average.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

$$H_0: P = 0.5$$
$$H_a: P \neq 0.5$$

Set the criteria for a decision

Significance level

In hypothesis testing, data is collected to show that the null hypothesis is not true. It is based on the likelihood of finding is the criterion. We can say the Significance level is the [Confidence level](#) that we need to accept or reject a null hypothesis.

The likelihood or level of significance is typically set at 5%. When the probability of obtaining a sample mean is less than 5% - the null hypothesis were true, then we reject the value stated in the null hypothesis.

- Level of significance, or significance level, refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value

stated in the null hypothesis were true.

- **The significance level as small as possible in order to protect the null hypothesis** and to prevent, as far as possible, the investigator from inadvertently making false claims.

When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we conclude that the sample we selected is too unlikely (it was one off random purely happened by chance event) and so we reject the null hypothesis.

- H_0: The children in the United States watch an average of 3 hours of TV/week.

We presume it is true but we hope to find that it is not true. We can look for the probability of obtaining a sample where possible values of the sample mean are not normally distributed. That means less than 5% of the sample fall within the range of standard deviation. And then we reject the null hypothesis.

Once we reject the null hypothesis, we can state if we have rejected the null hypothesis in favor of alternate hypothesis or the data was not conclusive

Test statistic

These test gives us a direction to accept or reject a Null hypothesis. These test guide our decision to either accept or reject a hypothesis. Test tells us how far, or how many standard deviations, a sample mean is from the population mean. The larger the value of the test statistic, the further the distance, or number of standard deviations, a sample mean is from the population mean stated in the null hypothesis. Its value is used to decide whether or not the null hypothesis should be rejected in our hypothesis test.

- Reject the null hypothesis. The sample mean is associated with a low proba-bility of occurrence when the null hypothesis is true.

- Retain the null hypothesis. The sample mean is associated with a high proba-bility of occurrence when the null hypothesis is true.

Perform the test below on same lines as we had covered earlier:

- Z Statistics: Give idea about % of population under the curve
  - o Z Score:
  - o Z Test:
- F Test: Guides if the relationship can be generalized.
  - o Significant F:
  - o F-Value:
  - o Critical F:
- T Test:
  - o T Critical:

Terms to Understand

Make a decision

The value of the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true.

Here we have already been guided by the test.

- Z Statistics: Give idea about % of population under the curve
  - Z Score: we find the area or percentage of population that falls under the curve
  - Z Test: Guides us to accept or reject Null Hypothesis.
    - If $Z_{Critical}$ < Alpha – Reject null hypothesis
    - $Z_{Critical}$ > Alpha – Consider accepting null hypothesis.
- F Test: Guides if the relationship can be generalized.
  - Significant F: Indicates % of the model that has occurred by pure chance, which is not explained by the model.
  - F-Value: Is compared with Critical F to accept or reject a hypothesis
  - Critical F:
    - If Fcalculated > Fcritical, H0 is rejected.
    - If Fcalculated < Fcritical, H0 cannot be rejected.
- T Test: Guides to accept hypothesis if we do not know population standard deviation. Helpful to test whether the sample mean is significantly different from some other value. It is Compared with tcritical for the appropriate significance level and degree of freedom.
  - T Critical: If tcalc > tcritical, we reject the null hypothesis and accept the alternate hypothesis. Otherwise, we accept the null hypothesis.
  - Effect of the T Test: Check for percentage of variance accounted for. Higher the value better
- P Value: The percentage the model **does not** describe a real relationship. This is compared to the P value for level of significance.
  - P Value wrt level of Significance: Null hypothesis is rejected when $P \leq a$.
- Region of acceptance: The **region of acceptance** is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected.
- Region of rejection: The set of values outside the region of acceptance. If the test statistic falls within the region of rejection, the null hypothesis is rejected.

Decision Errors

There are two types of errors that can be committed.

Type I Error – rejecting the null when it is true - predetermined by the significance level

Type II Error – failing to reject the null when it is false - The probability of a Type II Error cannot generally be computed because it depends on the population mean which is unknown. It can be computed at, however, for given values of µ, 2 σ, and n .

Terms to Understand

Type I error:

A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level or as an error of the first kind**. This probability is also called **alpha**, and is often denoted by α.

- In the first step of the hypothesis test, we select a level of significance, α, and α= P(Type I error). Because we purposely select a small value for α, **we control the probability of committing a Type I error**.

- For example, if we select α=0.05, and our test tells us to reject H0, then there is a 5% probability that we commit a Type I error.

- A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error.

The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0.

If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).

- For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

Type II error:

A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β. The probability of *not* committing a Type II error is called the **Power** of the test.

- A type II error is frequently due to sample sizes being too small.

- The probability of a type II error is generally unknown, but is symbolized by β and written

  - P(type II error) = β

- A type II error can also be referred to as an error of the second kind.

- The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision.

- In other words, the power of a hypothesis test is the probability of not committing a type II error. It is calculated by subtracting the probability of a type II error from 1, usually expressed as: Power = 1 - P(type II error) = $(1 - \beta)$

- The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1.

- Since Power is the probability of correctly rejecting a false null hypothesis, It is to our best interest to increase power. Ways of increasing Power

Terms to Understand

- o   Make alpha larger
- o   Use one-tailed rather than two tailed test
- o   Decrease variance
  - ▪ Increase sample size
  - ▪ Better measures
  - ▪ Increase effect size
- o   Magnitude of Power
  - ▪ Strong effect - 0.00-0.15 or 0.85-1.00
  - ▪ Moderate effect - 0.15-0.35 or 0.65-0.85
  - ▪ Weak effect - 0.35-0.49 or 0.51-0.65

Steps to calculate Power

- •   In short it is finding probability area of rejection on both the tails, adding them both and then then subtracting the result from 1.

Suppose an engines run for an average of 295 minutes, with a standard deviation of 20 minutes.

The inventor tests the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes, using a 0.05 level of significance.

Region of acceptance for this problem consists of the values between 294.45 and 305.55 as per region of acceptance

Critical parameter value is 290. Another way to express the critical parameter value is through effect size. The effect size is equal to the critical parameter value minus the hypothesized value. Thus, effect size is equal to 290 - 300 or -10.

Since the region of acceptance is 294.45 to 305.55, the null hypothesis will be rejected when the sampled run time is less than 294.45 or greater than 305.55.

Therefore, we need to compute the probability that the sampled run time will be less than 294.45 or greater than 305.55. To do this, we make the following assumptions:

We enter the following values into the calculator: value = 294.45; mean = 290; and standard deviation = 2.83. Given these inputs, we find that the cumulative probability is 0.94207551. This means the probability that the sample mean will be less than 294.45 is 0.942.

We enter the following values into the calculator: normal random variable = 305.55; mean = 290; and standard deviation = 2.83. Given these inputs, we find that the probability that the sample mean is less than 305.55 (i.e., the cumulative probability) is 0.99999998. Thus, the probability that the sample mean is greater than 305.55 is 1 - 0.99999998 or 0.00000002.

The power of the test is the sum of these probabilities: 0.94207551 + 0.00000002 = 0.94207553. This means that if the true average run time of the new engine were 290 minutes, we would correctly reject the hypothesis that the run time was 300 minutes 94.2 percent of the time. Hence, the probability of a Type II error would be very small.

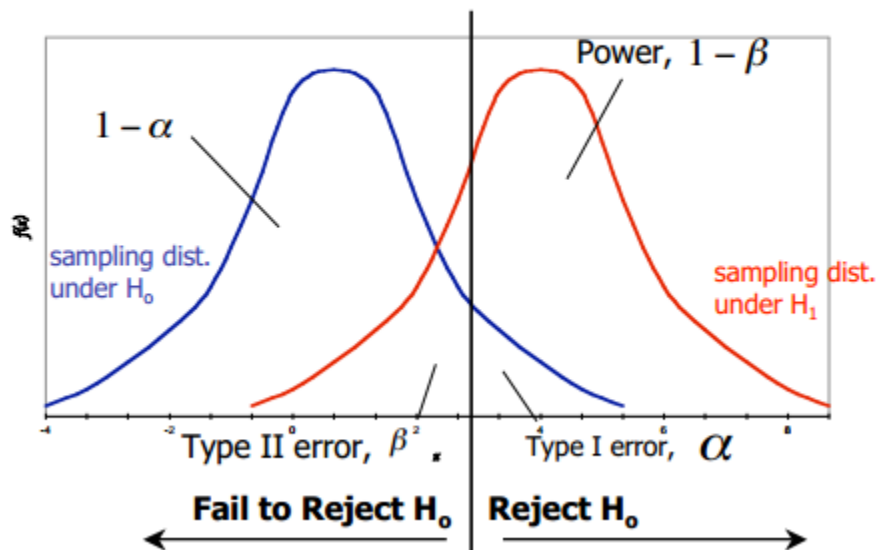Specifically, it would be 1 minus 0.942 or 0.058.

Terms to Understand

Consequences of Type I and Type II Error

- Normally, no adverse consequences occur when we correctly fail to reject a null hypothesis
  - Declaring not guilty an innocent man → he is free to go
- Type I and II errors are mistakes we do not want to make
  - – Letting a criminal go free (Type II)
  - – Or worse, sending to jail an innocent man (Type I)
  - That's why we set alpha to 0.05
- On the other hand, the ability to convict a guilty person is essential to our justice system
  - Reject Ho, when Ho is false

	Conclusion in Test of Hypothesis	
	Do Not Reject H_0	Reject H_0
H_0 is True – Not Guilty. Do not Reject H_0	Correct Decision $p = 1 - \alpha$	Type I Error $p = \beta$
H_0 is False – Guilty Reject H_0	Type II Error $p = \alpha$	Correct Decision $p = 1 - \beta$

Power & Errors

Naseha Sameen