- **F-statistic;** to compare the variability explained by the model to the unexplained variability of the error we first need to adjust each by an appropriate degrees of freedom. We divide each sum of squares by its degrees of freedom to get a mean square for each source of variability;
  - $Mean\ square\ for\ the\ model = MSModel = \frac{SSModel}{1}$
  - $Mean\ square\ error = MSE = \frac{SSE}{n-2}$
- To compare the two mean squares, we use their ratio to obtain the F-statistic;
  - $F = \frac{MSModel}{MSE}$
- The formal hypothesis being tested are:
  - Ho: the model is ineffective (or equivelantly B1=0)
  - Ha: the model is effective (or equivelantly B1 not equal to 0)
- **ANOVA to test a simple linear model;** to test for the effectiveness of a regression model we partition the variability to construct a ANOVA table for regression.

Source	df	Sum of Sq	Mean Square	F-statistic	p-value
Model	1	SSModel	$\frac{SSModel}{1}$	$F = \frac{MSModel}{MSE}$	$F_{1,n-2}$
Error	n-2	SSE	$\frac{SSE}{n-2}$		
Total	n-1	SSTotal			

- **Standard deviation of the error, Se;** for a simple linear model, we estimate the standard deviation of the error term with;
  - $s_\epsilon = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$
  - Where SSE is obtained from the ANOVA table.
- **Confidence interval based on normal distribution (to find a confidence interval);** we find a confidence interval by using sample statistic +/- z x SE. where z Is chosen so that the area between -z and +z in the standard normal distribution is the desired confidence level.

Confidence Interval	80%	90%	95%	98%	99%
Z*	1.282	1.645	1.960 (use 2)	2.326	2.575 (use 2.5)

- **Standard error and central limit theorem for sample proportions;** when choosing random samples of size n from a population with proportion p, the distribution has the following characteristics;
  - Center – The mean is equal to the population parameter, p.
  - Spread – The standard Error is calculated by $SE = \sqrt{\frac{P(1-p)}{n}}$.
  - Shape – If the sample size is sufficiently large, the distribution is reasonably normal.
  - The larger the sample size n the smaller the standard error (SE) and the more like a normal distribution it becomes. A normal distribution is a good approximation as long as
  - np >/= 10 and n(1-p)>/=10.
  - Using our notation for a normal distribution, this means that if n is sufficiently large $\hat{p} \sim N\left(P\sqrt{\frac{P(1-p)}{n}}\right)$
  - Margin of error is largest when p=0.5
- **Confidence Interval for a single proportion;** provided the sample size is large enough so that np >/= 10 and n(1-p)>/=10, a confidence interval for a population parameter can be computed based on a random sample using;
  - $\hat{p} \pm z^* x \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Where $\hat{p}$ is the sample proportion and z^* is a standard normal endpoint (z score for the appropriate amount of SDs) to give the desired level of confidence.
- **Determination of sample size to estimate a proportion;** if we want to estimate a population proportion to within a desired Margin of Error, with a given level of confidence, we should select the sample size of;
  - $n = \left(\frac{Z^*}{ME}\right)^2 \tilde{P}(1-\tilde{P})$

- Where we use $\tilde{P} = 0.5$ or if available, some other estimate of p.
- The question of how large of sample should be collected? Is solved by answering how accurate do we want the estimate to be? How much confidence do we want to have in the interval? What sort of proportion do we expect to see?

- **Test for a single proportion;** to test Ho: p=p0 vs Ha: p≠ $p0$ (or a one tailed alternative), we use the standardized test statistic;
  - $Z = \frac{\hat{P}_{-P_0}}{\sqrt{\frac{P_0(1-P_0)}{n}}}$
  - Where $\hat{P}$ is the proportion in a random sample of size n. Provided the sample size is reasonably large (using test above), the p value of the test is computed using the standard normal distribution.
- **Distribution of a sample mean;** when choosing random samples of size n from a population with mean u and standard deviation σ the standard error of the sample means is;
  - $SE = \frac{\sigma}{\sqrt{n}}$
  - The larger the sample size n the smaller the standard error (SE).
- **Degrees of freedom;** The t-distribution is characterized by its degrees of freedom (df). Degrees of freedom are calculated based on the sample size. The higher the degrees of freedom, the closer the t-distribution is to the standard normal.
  - If a population with mean μ_0 is approximately normal or if *n* is large (*n* ≥ 30), the standardized statistic for a mean using the sample *s* follows a *t*-distribution with *n* – 1 degrees of freedom:
  - $$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$
- **Central limit theorem for sample means;** when choosing random samples of size n from a population with mean u and standard deviation σ, the distribution of the sample means is reasonably normal if the sample size is sufficiently large ($n \geq 30$), (meaning even if skewed data is expanded over 30 samples it will resemble a normal distribution) with mean u and standard error SE= $\frac{\sigma}{\sqrt{n}}$.
- **The distribution of sample means using the same standard deviation;** when choosing random samples of size n from a population with mean u, the distribution of sample means has the following characteristics;
  - Center – The mean is equal to the population mean, u.
  - Spread – The standard Error is estimated using $SE = \frac{s}{\sqrt{n}}$. Where s is the standard deviation of the sample.
  - Shape – the standardized sample means approximately follow a t-distribution with n-1 degrees of freedom (df).
  - For small sample sizes ($n < 30$), the t-distribution is only a good approximation if the underlying population has a distribution that is approximately normal.
- **Confidence interval for a single mean;** a confidence interval for a population mean u can be computed based on a random sample of size n using;
  - $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
  - Where $\bar{x}$ and s are the mean and standard deviation from the sample and t^* is an endpoint chosen from the t-distribution with n-1df to give the desired level of confidence.
  - The t-distribution is appropriate if the distribution of the population is approximately normal or the sample size is large ($n \geq 30$).
- **T-test for single mean;** to test Ho: u=u0 vs Ha: u≠ $u0$ (or a one tailed alternative), we use the t-statistic;
  - $t = \frac{\bar{x} - u_o}{s/\sqrt{n}}$
  - Where $\bar{x}$ is the mean and s is the standard deviation in a random sample of size n. provided the underlying population is reasonably normal (or the sample size is large), the p value of the test is computed using the appropriate tail(s) of a t-distribution with n-1 degrees of freedom.
- **Exclude outliers if no longer a valid member of population of interest or errors.**