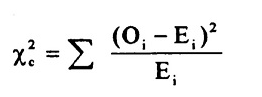
What is a Chi Square Test?

There are **two types of chi-square tests**. Both use the chi-square statistic and distribution for different purposes:

* A**chi-square goodness of fit test** determines if a sample data matches a population. For more details on this type, see: [*Goodness of Fit Test*](http://www.statisticshowto.com/goodness-of-fit-test/).
* A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of [categorical variables](http://www.statisticshowto.com/what-is-a-categorical-variable/) differ from each another.
  + A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
  + A **very large chi square test statistic**means that the data does not fit very well. In other words, there isn’t a relationship.

**A chi-square statistic is one way to show a relationship between two**[**categorical variables**](http://www.statisticshowto.com/what-is-a-categorical-variable/). In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. **The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.**



**A low value for chi-square means there is a high correlation between your two sets of data.**

### **Uses**

The chi-squared distribution has many uses in statistics, including:

* [Confidence interval](http://www.statisticshowto.com/probability-and-statistics/confidence-interval/)estimation for a population [standard deviation](http://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of a normal distribution from a sample standard deviation.
* Independence of two criteria of classification of [qualitative variables](http://www.statisticshowto.com/qualitative-variable/).
* Relationships between [categorical variables](http://www.statisticshowto.com/what-is-a-categorical-variable/) ([contingency tables](http://www.statisticshowto.com/what-is-a-contingency-table/)).
* [Sample variance](http://www.statisticshowto.com/sample-variance/) study when the underlying distribution is normal.
* Tests of deviations of differences between expected and observed frequencies (one-way tables).
* The chi-square test (a [goodness of fit](http://www.statisticshowto.com/goodness-of-fit-test/)test).

## Fstat :-variance among groups

## F Value in Regression

**The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero.**In other words, the model has no predictive capability. Basically, the f-test compares your model with zero predictor variables (the intercept only model), and decides whether your added coefficients improved the model. If you get a significant result, then whatever coefficients you included in your model improved the model’s fit.

The F value in [one way ANOVA](http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/#OneWayANOVA) is a tool to help you answer the question “Is the [variance](http://www.statisticshowto.com/probability-and-statistics/variance/)between the [means](http://www.statisticshowto.com/mean/)of two populations [significantly different](http://www.statisticshowto.com/what-is-statistical-significance/)?” The F value in the [ANOVA](http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/)test also determines the [P value](http://www.statisticshowto.com/p-value/); The P value is the probability of getting a result at least as extreme as the one that was actually observed, given that the [null hypothesis](http://www.statisticshowto.com/probability-and-statistics/null-hypothesis/)is true.

* **MEAN is more influenced by Oulier**
* **Median is not influenced by Outlier**

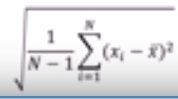
**MEASURES OF CENTRAL TENDENCY:-**

Mean /Median /Mode

**MEASURES OF DISPERSION**

Range

SD



Why do we use Squares??

There will be deviations which will be + and some with –ve. So it might tend to 0 which will not capture the variance we want to calculate. So we use square.

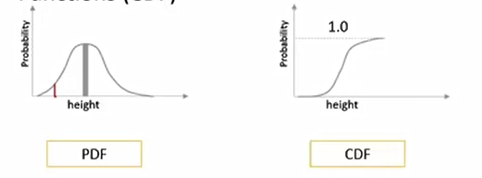
Why do we use SD and not Variance

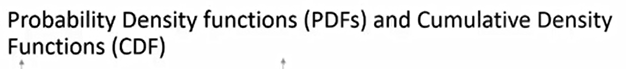
Variance will not be expressive in units as it is unit square as well

N-1 🡺 degree of freedom. We can move rest variables freely and still maintain the final statistic

Implications are that the effect of square

IQR 75th percentile – 25 percentile

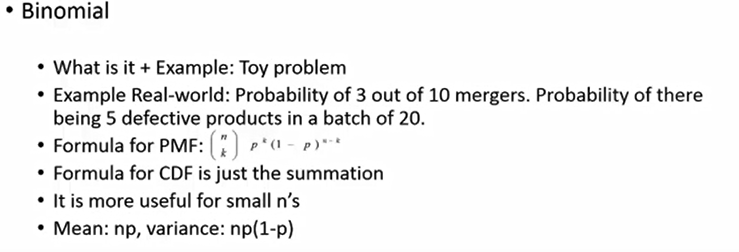




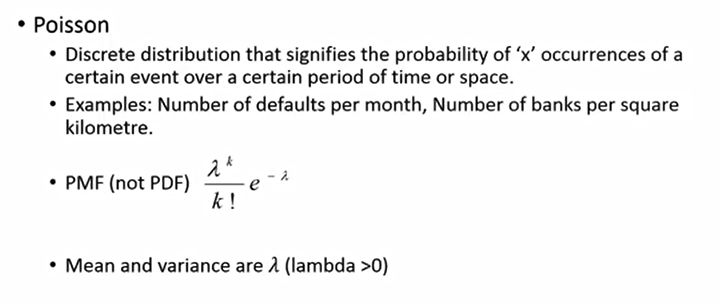
BErnouli Dist:-

One trial and probability of first success /failure

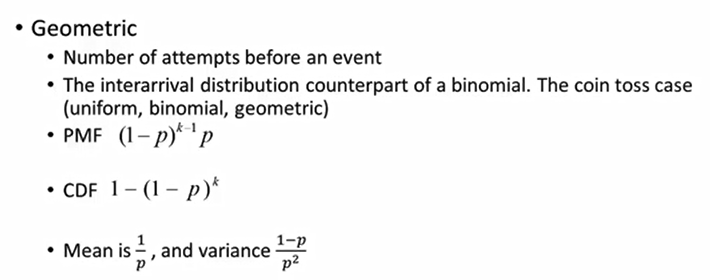
CDF is cumulative



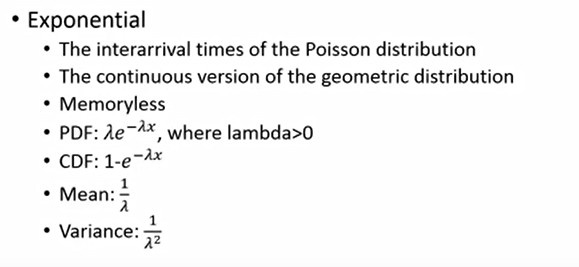
EX :- Toss a coin n times, probability of k heads or k tails in ntrials



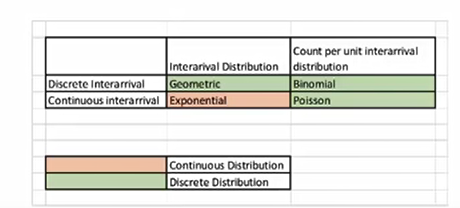
PMF is probability mask function

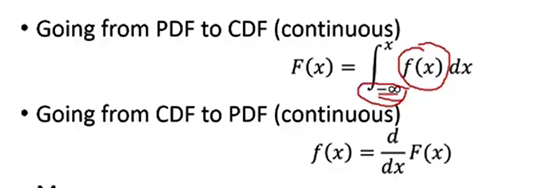


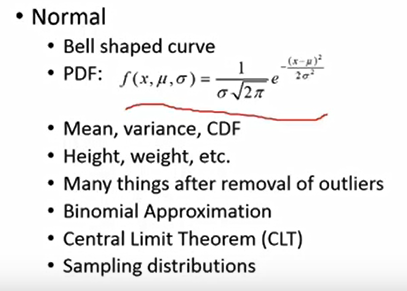
How many times I need to toss to get my first head/tail

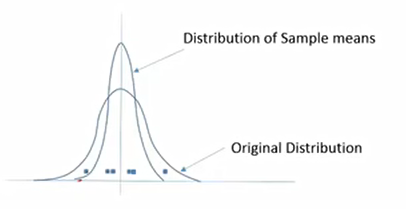
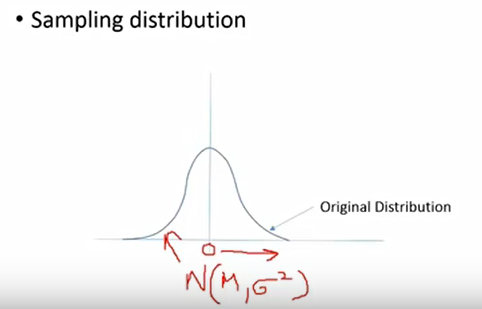


Example :-How long to wait for next person to arrive at toll





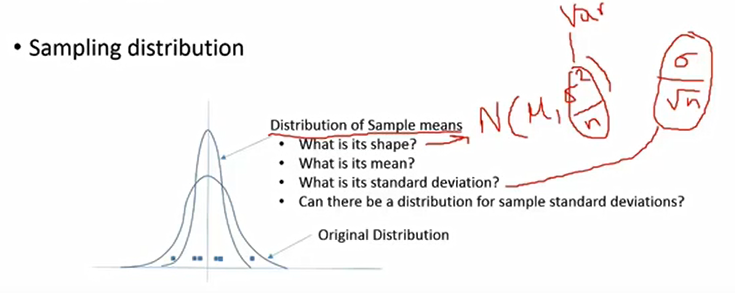




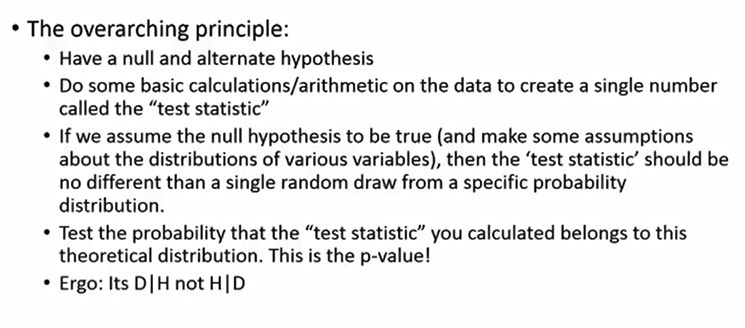
**Central limit theorem**

From original distribution take out sample and calculate Mean. For example I take random sample of each from set . The mean of all samples in aggreagation will be equal to overall mean of the distribution.

As we keep on increasing the sample the mean will tend to come close to Mean of the distribution. That is called Central Limit Distribution

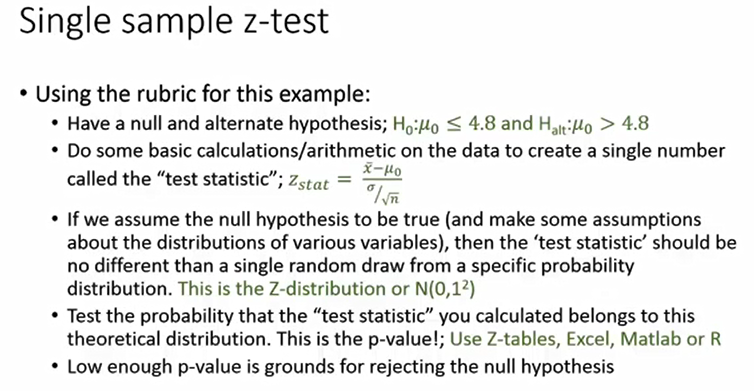


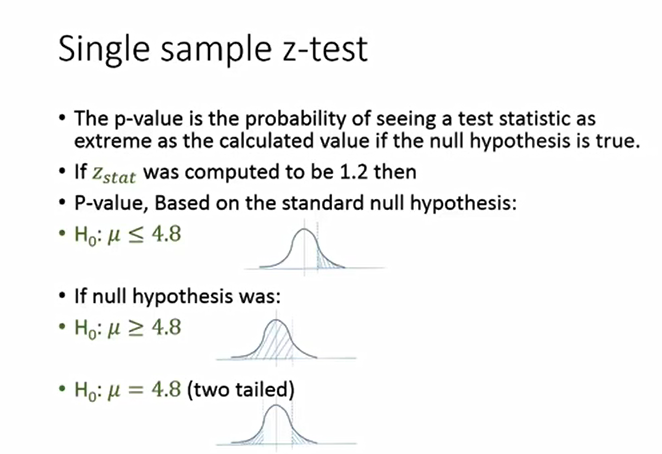
Introduction to Inferential Statistics

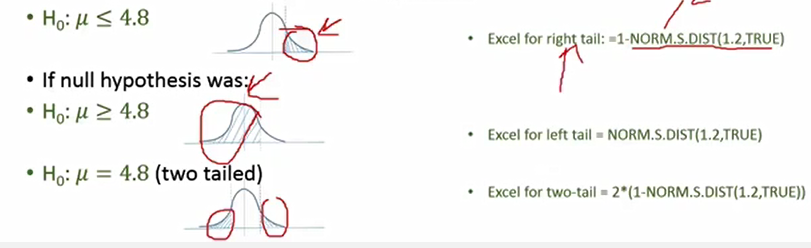


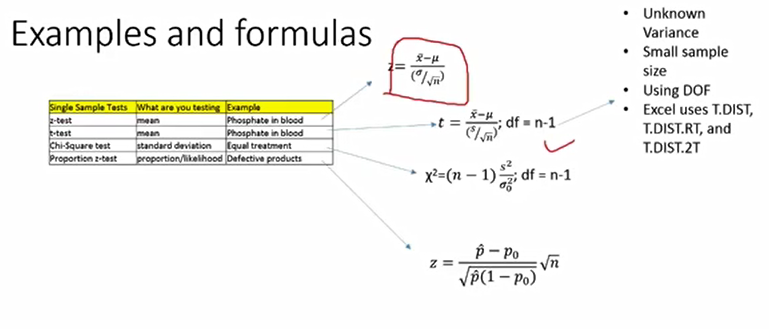
A **t**-**test** is used for **testing** the mean of one population against a standard or comparing the means of two populations if **you do not know the populations' standard deviation** and

when you have a limited sample (n < 30). If **you know the populations' standard deviation**, you may use a **z**-**test**



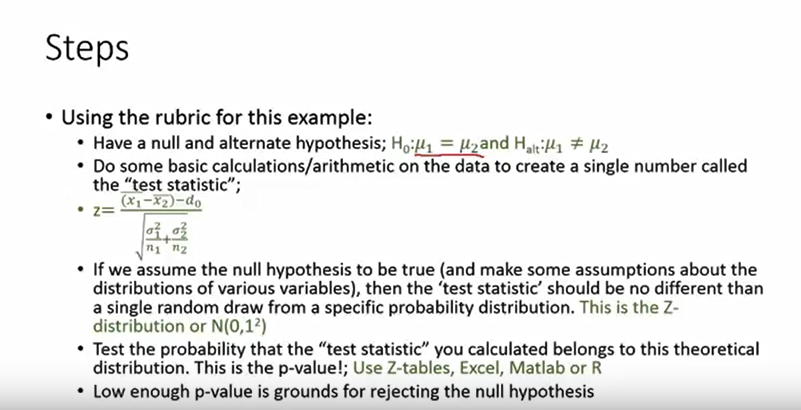


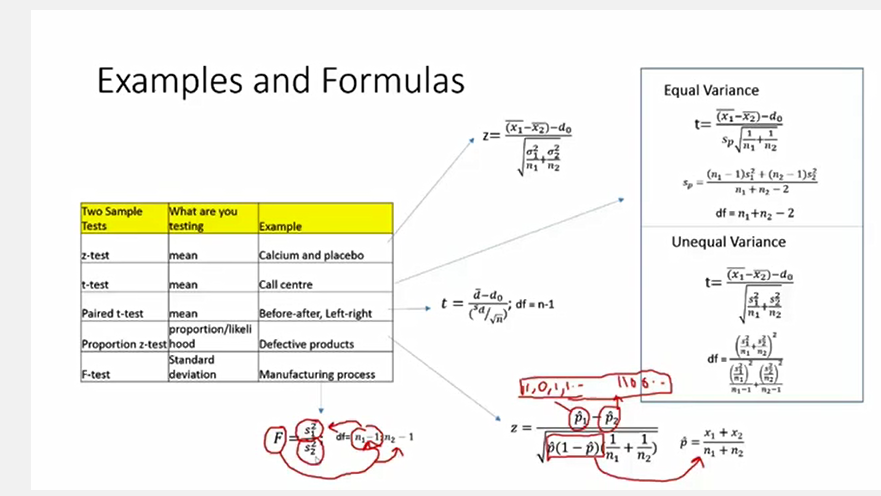


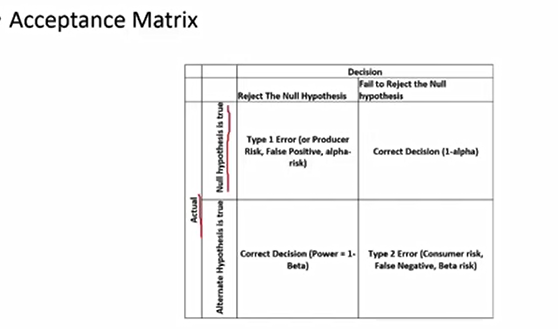


**T distribution :-Mean is 0,SD =1**

**Two Sample Tests**



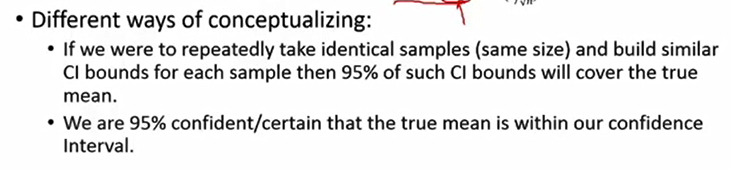




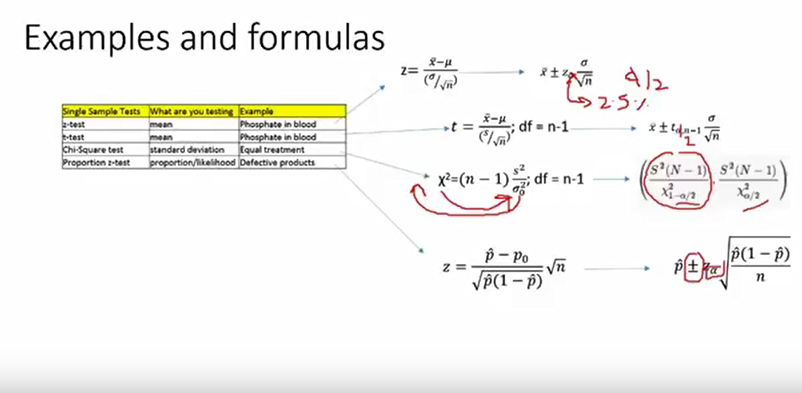
Pvalue is ideally type 1 error (alpha)

Type 2 error is ideally dependent on Alpha

CONFIDENCE INTERVAL:-



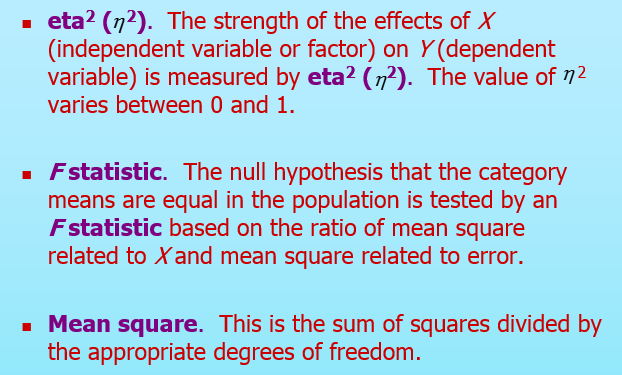
Identify range when we don’t have any hypothesis in place.



ANOVA

* **Analysis of variance (ANOVA) is used as a test of means for two or more populations. The null hypothesis, typically, is that all means are equal.**

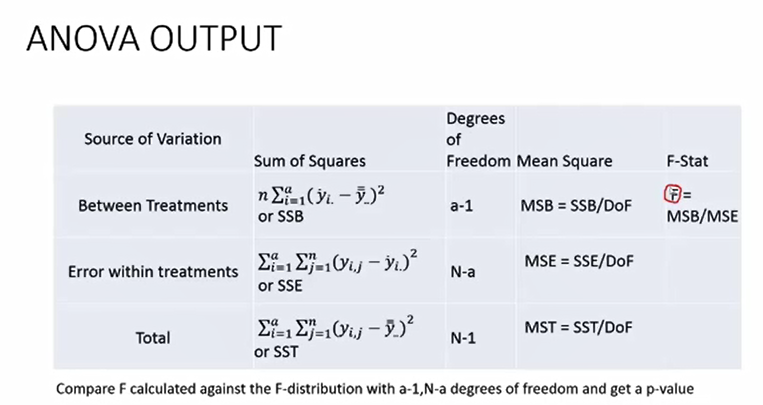
**Statistics Associated with One-Way   
Analysis of Variance**



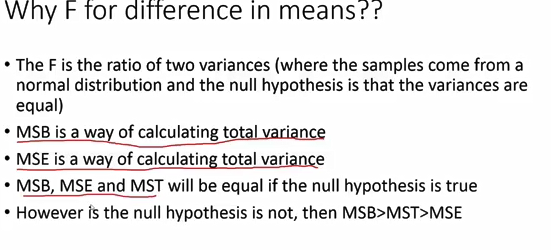
**Fstat = Mean Sqaure related to X/Mean square related to error**

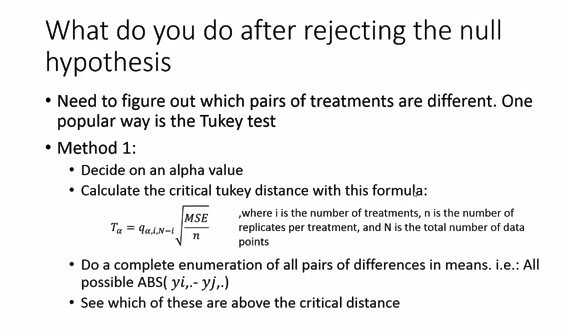
**Y**

***SSy* = *SSbetween* + *SSwithin***

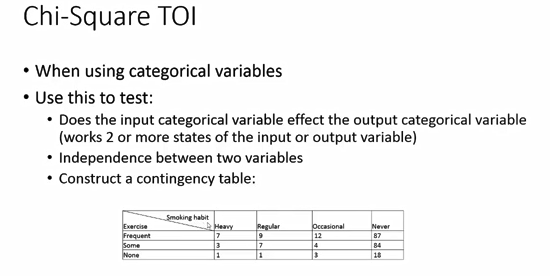


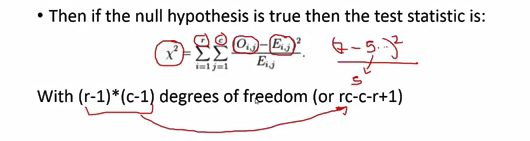
**Its basically standard deviation within and between groups**





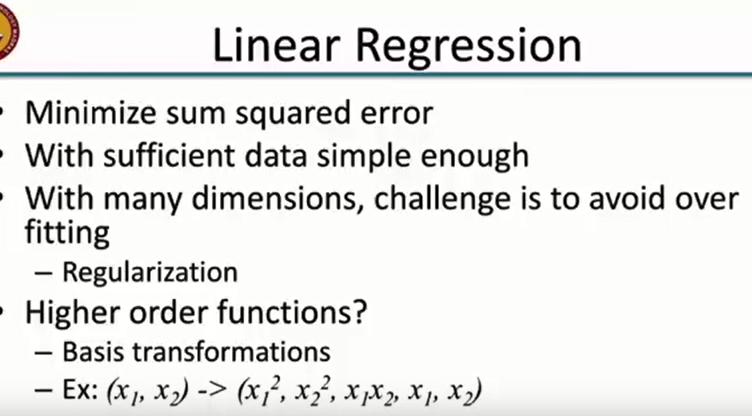
**Tukey Distance** calculates the pairs which are different.

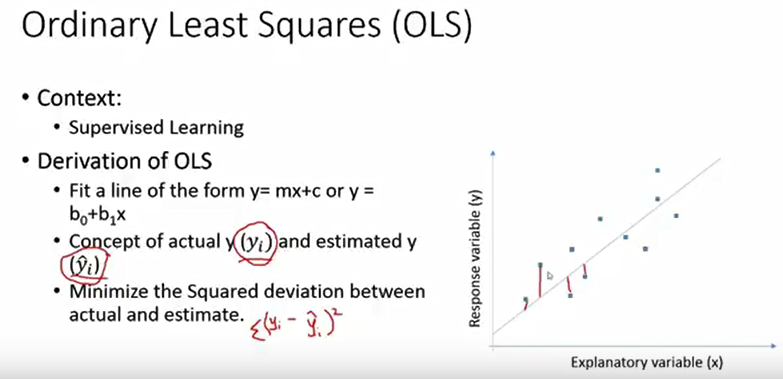


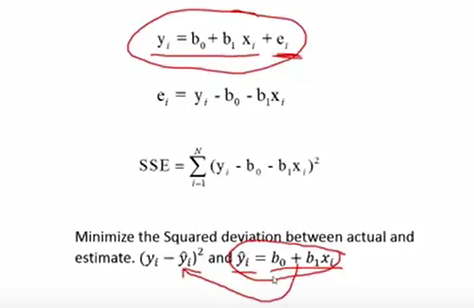


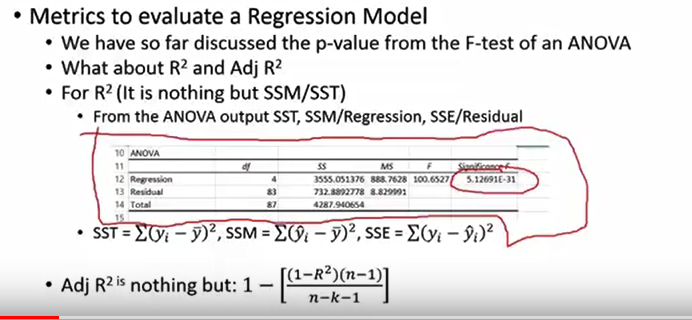
Cross sectional :- not a function of time

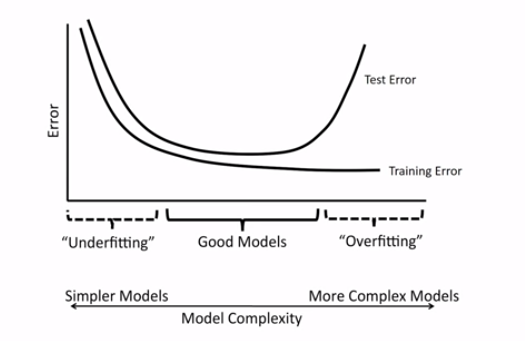
Time-series : previous time data impact current data. So the data attributes are function of time.

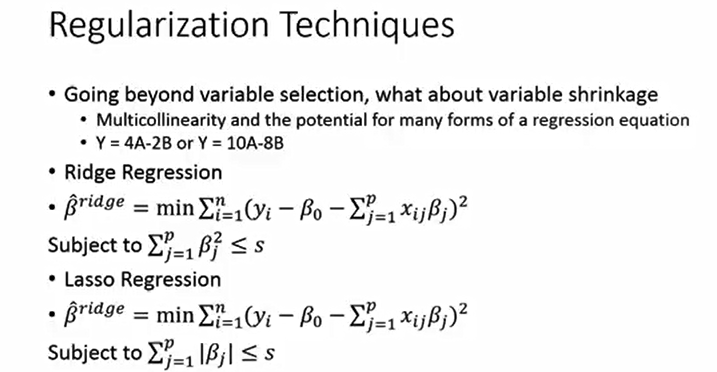


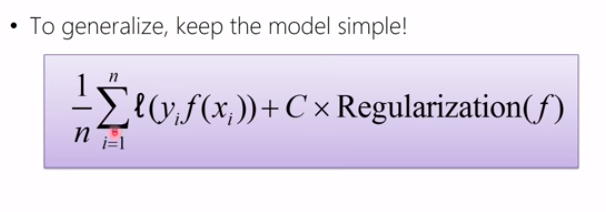


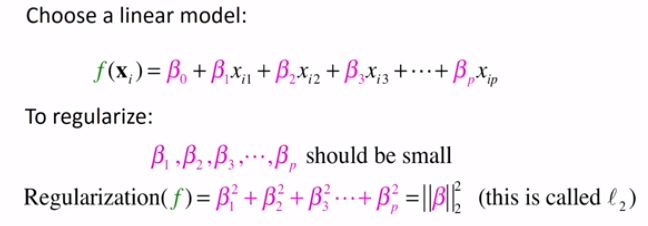












Regularization reduces the slope of the line

 Now L2 regularization tends intuitively to make all of the coefficients

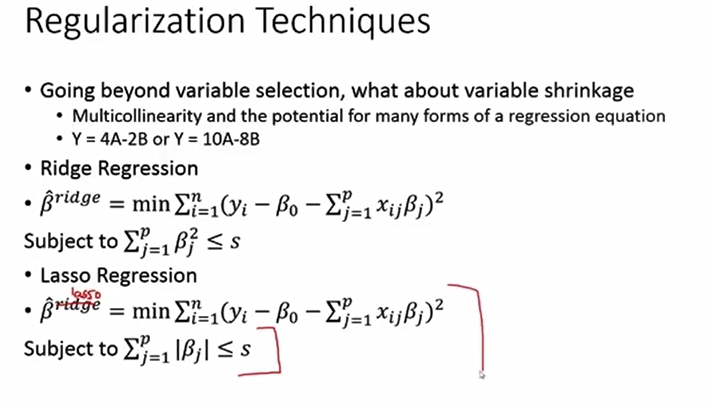
kind of smaller, and L1 regularization – it’s a bit different. That one’s useful for making

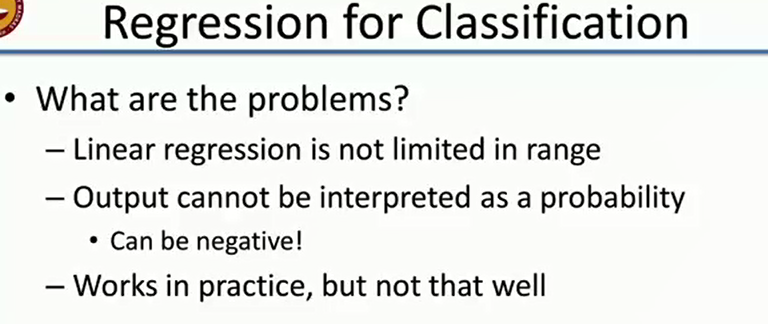
sparse solutions, it sends a bunch of the coefficients to 0.

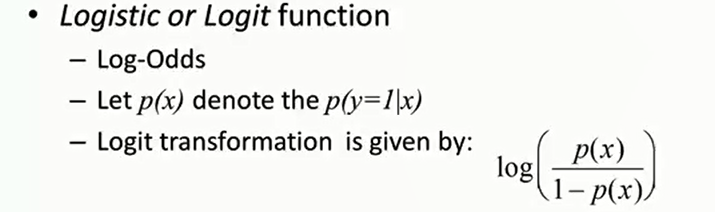
L1==Lasso

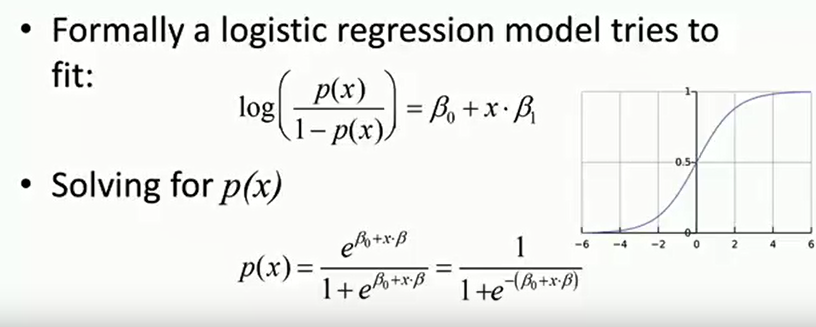
L2==Ridge

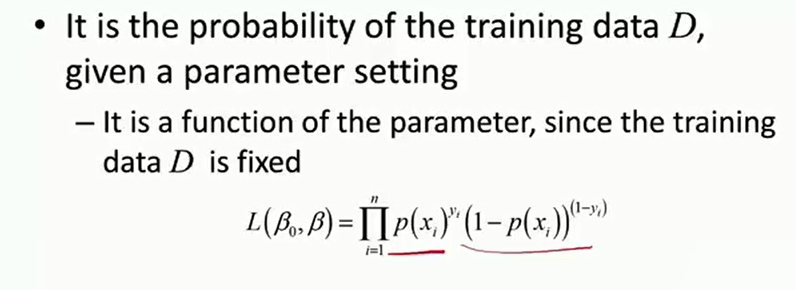
Regularization is also called shrinkage











Feature Selection :-

Greedy Backward Selection

**X**