1. **(10 pts.) Explain what an F-test is and how it differs from a t-test.**

**Ans:**

An F-test is any statistical test in which the test statistics has an F distribution under the null hypothesis. Common uses of F-test are: F-test for the quality of two variances, The F-test in one-way analysis of variance is used to assess whether the expected values of a quantitative variable(means) within several pre-defined groups differ from each other. The overall F-test for regression tests the null hypothesis that β1 = β2 = ... = βj = 0.

A *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t* distribution if the null hypothesis is supported. It is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. Common uses of t-test are: A test of whether the slope of a [regression line](http://en.wikipedia.org/wiki/Linear_regression) differs [significantly](http://en.wikipedia.org/wiki/Statistical_significance) from 0. A one-sample [location test](http://en.wikipedia.org/wiki/Location_test) of whether the mean of a normally distributed population has a value specified in a [null hypothesis](http://en.wikipedia.org/wiki/Null_hypothesis). A two sample location test of the null hypothesis that the [means](http://en.wikipedia.org/wiki/Expected_value) of two [normally distributed](http://en.wikipedia.org/wiki/Normal_distribution) populations are equal.

1. **(10 pts.) Explain what cross-validation is and describe some of its specific implementations.   How it used to validate a regression model?**

**ANS:**

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. Cross-validation is primarily a way of measuring the predictive performance of a statistical model. Every statistician knows that the model fit statistics are not a good guide to how well a model will predict: high R^2 does not necessarily mean a good model. The possibility of overfitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model. Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.

**Steps of cross-validation:**

One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

## Cross-validation for linear models

While cross-validation can be computationally expensive in general, it is very easy and fast to compute LOOCV for linear models. A linear model can be written as

\[ \mathbf{Y} = \mathbf{X}\mbox{\boldmath$\beta$} + \mathbf{e}. \]

Then

\[ \hat{\mbox{\boldmath$\beta$}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \]

and the fitted values can be calculated using

\[ \mathbf{\hat{Y}} = \mathbf{X}\hat{\mbox{\boldmath$\beta$}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}, \]

where \mathbf{H} =  \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' is known as the “hat-matrix” because it is used to compute \mathbf{\hat{Y}} (“Y-hat”).

If the diagonal values of \mathbf{H} are denoted by h_{1},\dots,h_{n}, then the cross-validation statistic can be computed using

\[ \text{CV} = \frac{1}{n}\sum_{i=1}^n [e_{i}/(1-h_{i})]^2, \]

where e_{i} is the residual obtained from fitting the model to all n observations. Thus, it is not necessary to actually fit nseparate models when computing the CV statistic for linear models. This remarkable result allows cross-validation to be used while only fitting the model once to all available observations

**Referral: www.wikipedia.com**