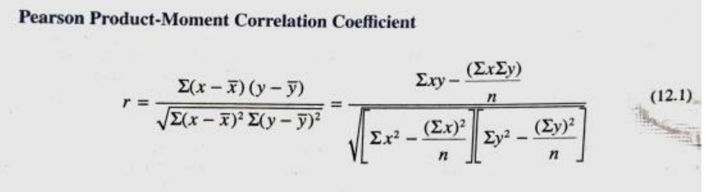
Week 8 - Review

Wednesday, February 28, 2018

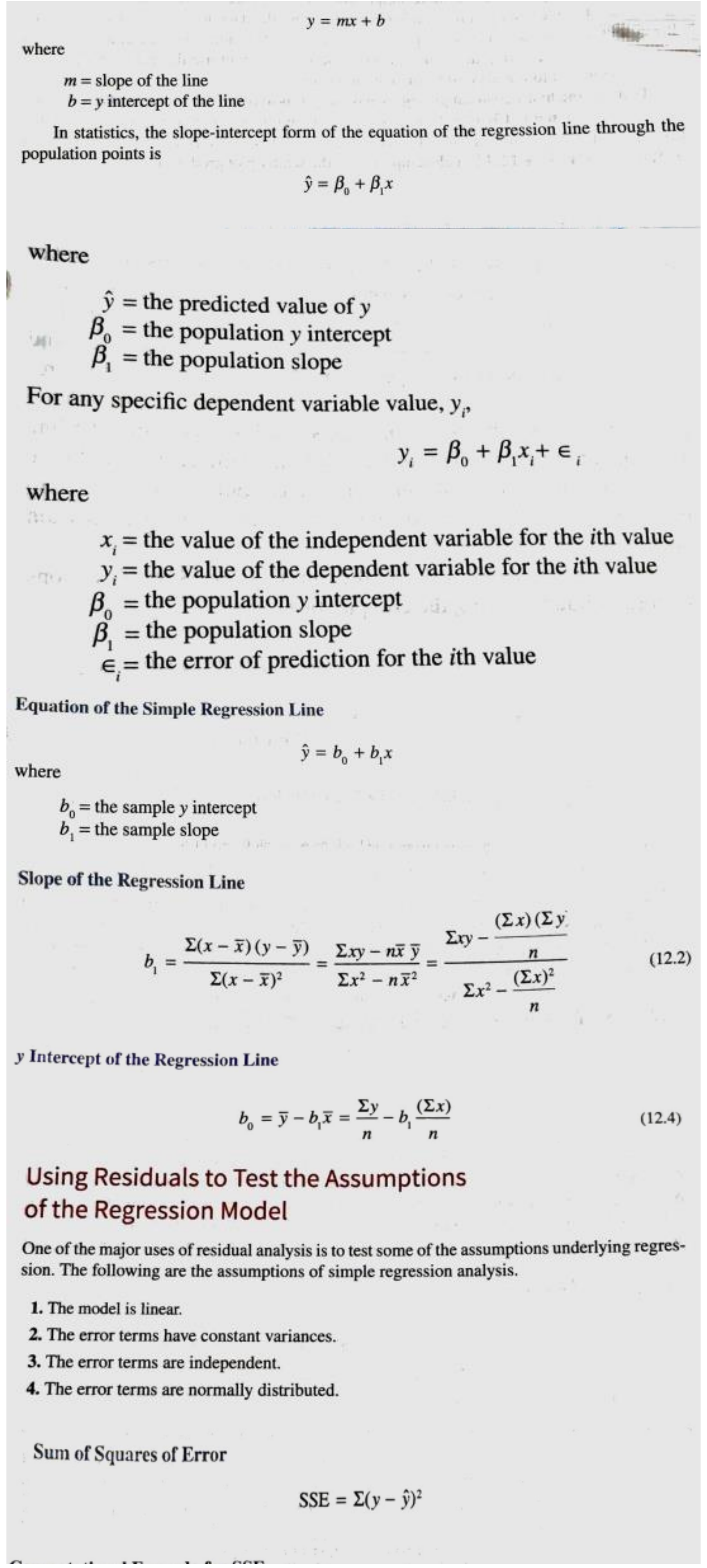
3:53 AM

- ## Design of Experiments
  - **Experimental design** is a plan and a structure to test hypotheses in which the researcher either controls or manipulates one or more variables.
  - **Independent variable (aka factors)** may be either a treatment variable or a classification variable.
  - **Treatment variable** is a variable the experimenter controls or modifies in the experiment
  - **Classification variable** is one characteristic of the experiment subject that was present prior to the experiment and is not a results of the experiment's manipulations or control.
  - **Dependent variable** is the response to the different levels of the independent variables
  - **Levels or Classifications**, of independent variables are the subcategories of the independent variable used by the researcher in the experimental design.
  - **Analysis of Variance (ANOVA)**
    - This concept begins with the notion that dependent variable response (measurement, data) are not all the same in a given study.
- ## The Completely Randomized Design (One-Way ANOVA)
  - Completely randomized design - subjects are assigned randomly to treatment.
    - This design will only contain one independent variable, with two or more treatment levels, or classifications.
  - ANOVA is computed with the three sums of squares:
    - Total, Treatment (columns) and error
  - Assumptions for Analysis of Variance
    - Observations are drawn from normally distributed populations
    - Observations represent random samples from the populations
    - Variances of the population are equal
  - **F Value** - a ratio of the treatment variance to the error variance.
  - **F Distribution** - is made up of two unique degrees of freedom values:
    - Numerator (df$_c$) and Denominator (df$_e$)
- ## Multiple Comparison Tests
  - **Multiple comparisons** are to be used only when an overall significant difference between groups has been obtained by using the F value of the analysis of variance.
    - Some of these techniques protect more for Type I errors and others protect more for Type II errors.
    - Some multiple comparison tests require equal sample sizes.
  - **A posteriori** or **post hoc** pairwise comparisons are made after the experiment when the researcher decides to test for any significant differences in the samples based on a significant overall F value.
  - **A priori** comparisons are made when the researcher determines before the experiment which comparisons are to be made.
- ## Tukey's Honestly Significant Difference (HSD) Test
  - Tukey's HSD Test is a popular test for pairwise a posteriori multiple comparisons.
  - It is limited by the fact that it requires equal sample sizes.
  - The Test takes the following into consideration:
    - Number of treatment levels
    - The value of mean square error
    - The sample size
  - The results computed determines the critical difference necessary between the means of any two treatment levels for the means to be significantly different.
- ## Tukey-Kramer Procedure
  - The Tukey-Kramer Procedure is a modified version of HSD.
  - The formula for computing the significant difference with this procedure is similar to that for the equal sample sizes, with the exception that the mean square error is divided in half and weighted



Pearson Product-Moment Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} = \frac{\Sigma xy - \dfrac{(\Sigma x \Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right]}} \qquad (12.1)$$



$$y = mx + b$$

where

m = slope of the line
b = y intercept of the line

In statistics, the slope-intercept form of the equation of the regression line through the population points is

$$\hat{y} = \beta_0 + \beta_1 x$$

where

$\hat{y}$ = the predicted value of y
β_0 = the population y intercept
β_1 = the population slope

For any specific dependent variable value, y_i,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

x_i = the value of the independent variable for the ith value
y_i = the value of the dependent variable for the ith value
β_0 = the population y intercept
β_1 = the population slope
ϵ_i = the error of prediction for the ith value

Equation of the Simple Regression Line

$$\hat{y} = b_0 + b_1 x$$

where

b_0 = the sample y intercept
b_1 = the sample slope

Slope of the Regression Line

$$b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} \qquad (12.2)$$

***y* Intercept of the Regression Line**

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\Sigma y}{n} - b_1\frac{(\Sigma x)}{n} \qquad (12.4)$$

Using Residuals to Test the Assumptions of the Regression Model

One of the major uses of residual analysis is to test some of the assumptions underlying regression. The following are the assumptions of simple regression analysis.

1. The model is linear.
2. The error terms have constant variances.
3. The error terms are independent.
4. The error terms are normally distributed.

Sum of Squares of Error

$$SSE = \Sigma(y - \hat{y})^2$$

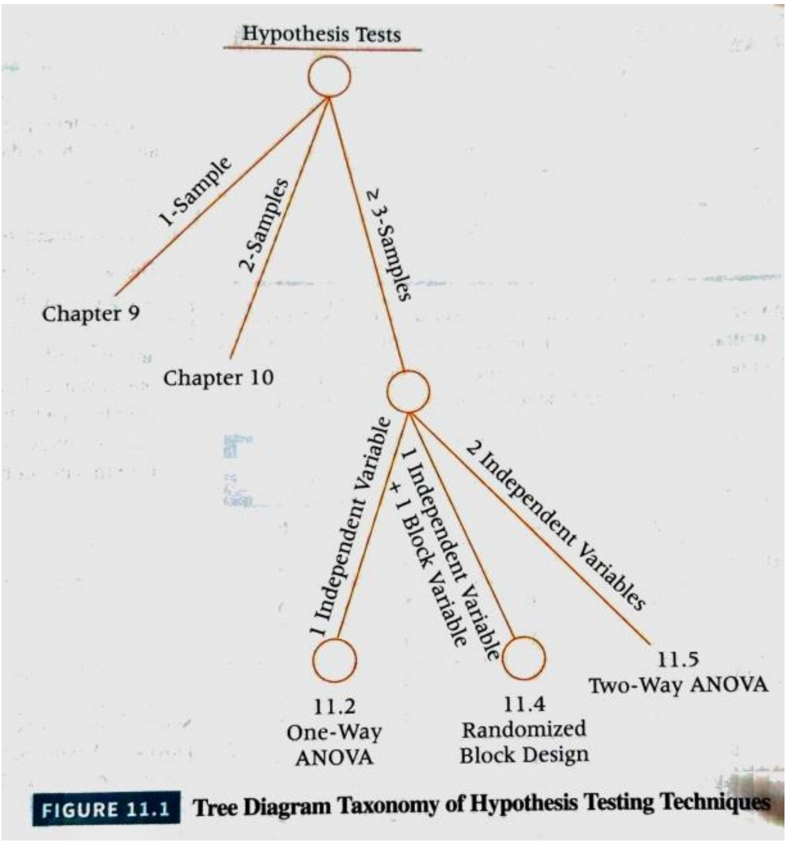
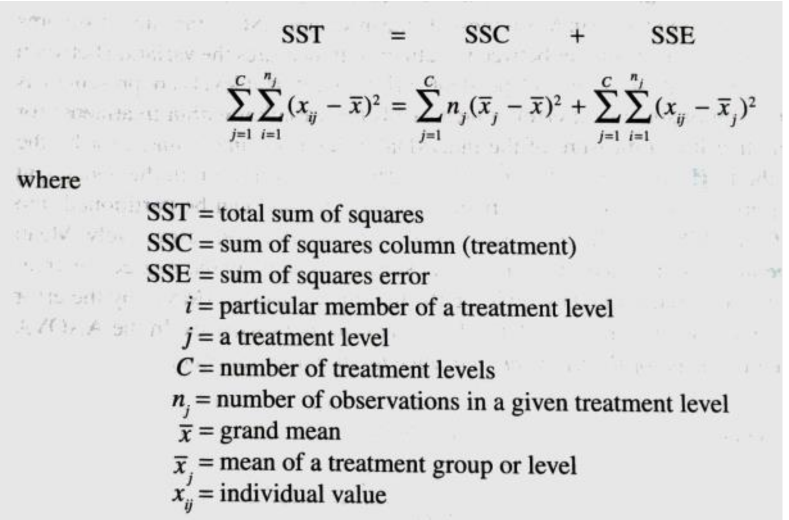


FIGURE 11.1 **Tree Diagram Taxonomy of Hypothesis Testing Techniques**

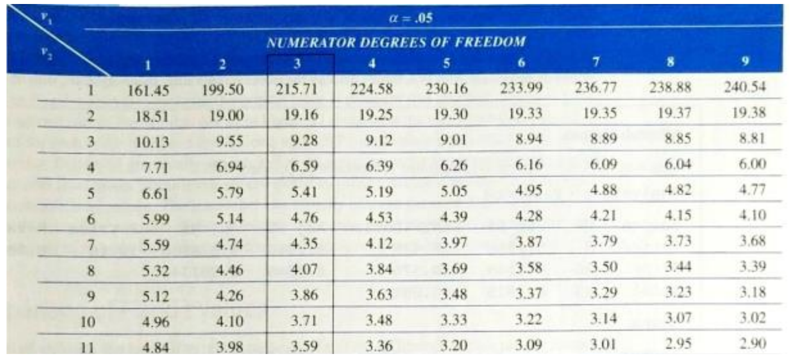


SST = SSC + SSE

$$\sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{C} n_j(\bar{x}_j-\bar{x})^2 + \sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

where

SST = total sum of squares
SSC = sum of squares column (treatment)
SSE = sum of squares error
i = particular member of a treatment level
j = a treatment level
C = number of treatment levels
n_j = number of observations in a given treatment level
$\bar{x}$ = grand mean
$\bar{x}_j$ = mean of a treatment group or level
x_{ij} = individual value

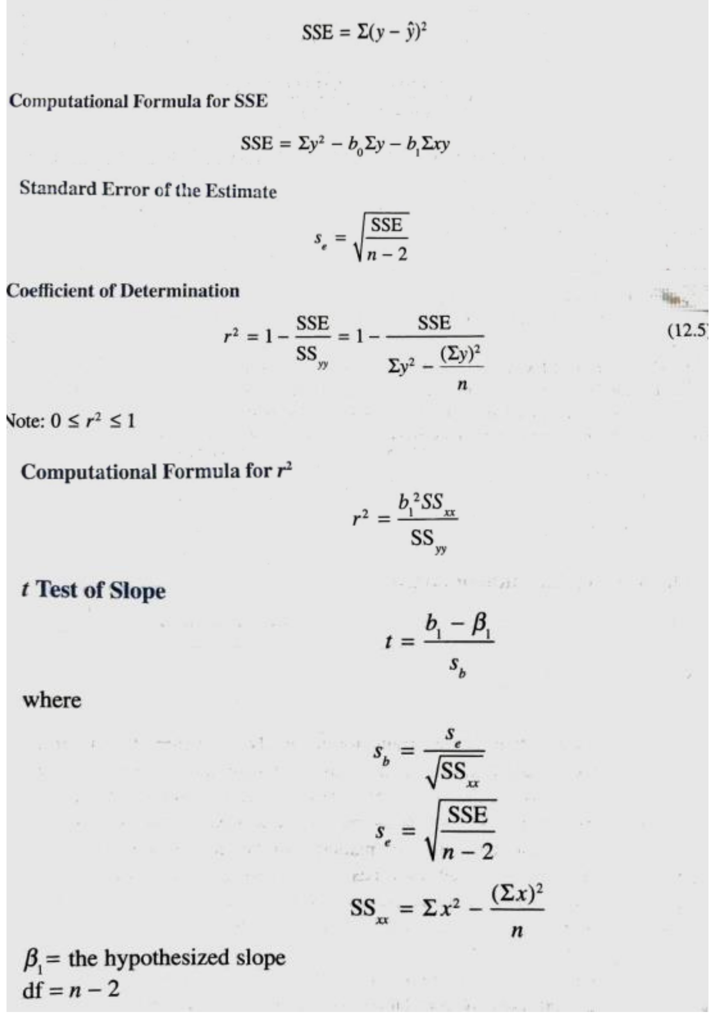


v_1 / v_2	α = .05								
	NUMERATOR DEGREES OF FREEDOM								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90

- The Tukey-Kramer Procedure is a modified version of HSD.
  - The formula for computing the significant difference with this procedure is similar to that for the equal sample sizes, with the exception that the mean square error is divided in half and weighted by the sum of the inverses of the sample size under the root sign.
- The Randomized Block Design
  - The randomized block design is similar to the completely randomized design in that it focuses on one independent variable of interest.
  - The randomized block design also includes a second variable, referred to as a blocking variable, that can be used to control for confounding or concomitant variables.
  - **Confounding variables or concomitant variables** - are variables that are not being controlled by the researcher in the experiment but can have an effect on the outcome of the treatment being studied.
  - **Blocking variable -** is a variable that the researcher wants to control but is no the treatment variable of interest.
  - **Repeated measures design -** is a randomized block design in which each block level is an individual item or person, and that person or item is measured across all treatments.
- A Factorial Design (Two-Way ANOVA)
  - **Factorial Design** – every level of each treatment is studied under the conditions of every level of all other treatments.
  - **Two-Way Analysis of Variance** – the process used to statistically test the effects of variables in factorial designs with two independent variables.
  - **Interaction** – when the effects of one treatment vary according to the levels of treatment of the other effect.
    - Interaction occurs when the pattern of cell means in one row (going across columns) varies from the pattern of cell means in the other rows. This variation indicates that the difference in column effects depend on which row is being examined.
    - When interaction effects are significant, the main effects are confounded and should not be analyzed in the usual manner.

Simple Regression Analysis and Correlation

- Correlation
  - **Correlation** is a measure of the degree of relatedness of variables.
  - **R –** a measure of the linear correlation of two variables.
  - In correlation analysis, it does not matter which variable is designated x and which is designated y.
- Regression Analysis
  - The process of constructing a mathematical or function that can be used to predict or determine one variable by another variable or other variables.
    - **Simple Regression (or Bivariate Regression)** – involving two variables in which one variable is predicted by another variable. In a simple regression, only a straight-line relationship between two variables is examined.
      - **Dependent variable** – the variable to be predicted
      - **Independent variable** – the predictor or explanatory variable.
  - Equation of the Regression Line
  - ***Deterministic models*** – mathematical models that produce an exact output for a given input.
    - Math course: $y = mx + b$
      - ***m*** = slope of the line; ***b*** = y intercept of the line
    - Statistics: $\hat{y} = \beta_0 + \beta_1 x$
      - $\hat{y} = the\ predicted\ value\ of\ y$
      - $\beta_0 = the\ population\ y\ intercept$
      - $\beta_1 = the\ population\ slope$
  - ***Probabilistic model*** – A model that includes an error term that allows for various values of output to occur for a given value of input.
    - Specific dependent variable value: $y_i = \beta_0 + \beta_1 x_i + \in_i$
      - $x_i = the\ value\ of\ the\ independent\ variable\ for\ the\ \boldsymbol{i}th\ value$



$$\text{SSE} = \Sigma(y - \hat{y})^2$$

Computational Formula for SSE

$$\text{SSE} = \Sigma y^2 - b_0\Sigma y - b_1\Sigma xy$$

Standard Error of the Estimate

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

Coefficient of Determination

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{\text{SSE}}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} \qquad (12.5$$

Note: $0 \le r^2 \le 1$

Computational Formula for r^2

$$r^2 = \frac{b_1^{\,2} SS_{xx}}{\text{SS}_{yy}}$$

***t* Test of Slope**

$$t = \frac{b_1 - \beta_1}{s_b}$$

where

$$s_b = \frac{s_e}{\sqrt{\text{SS}_{xx}}}$$

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

$$\text{SS}_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

β_1 = the hypothesized slope
df = $n - 2$

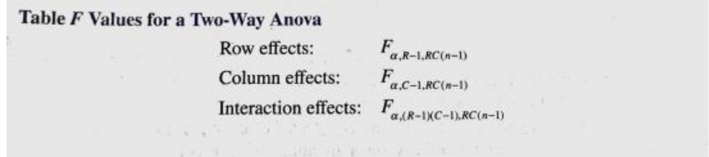
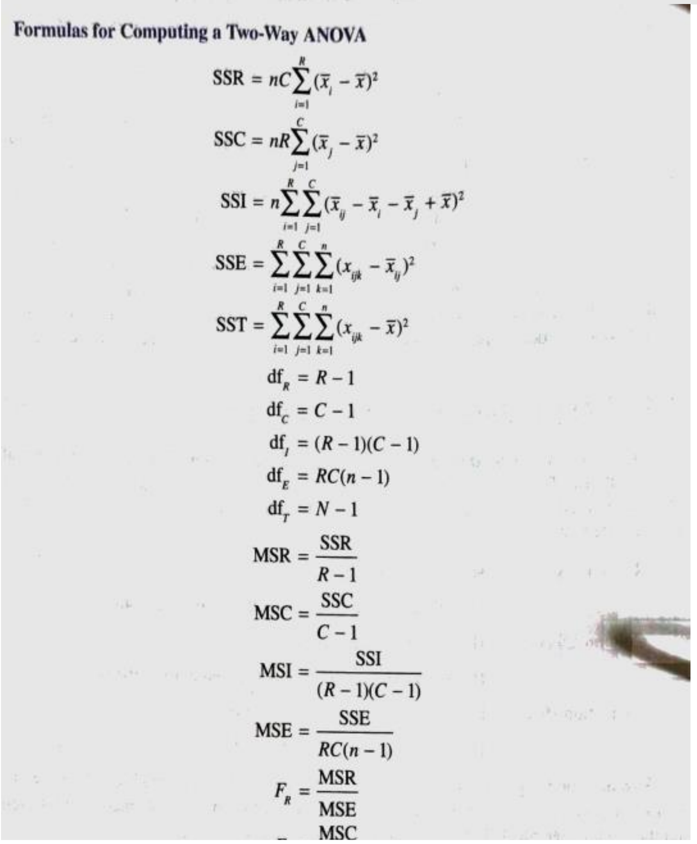


Table *F* Values for a Two-Way Anova

Row effects: $F_{\alpha,R-1,RC(n-1)}$
Column effects: $F_{\alpha,C-1,RC(n-1)}$
Interaction effects: $F_{\alpha,(R-1)(C-1),RC(n-1)}$



Formulas for Computing a Two-Way ANOVA

$$\text{SSR} = nC\sum_{i=1}^{R}(\bar{x}_i - \bar{x})^2$$

$$\text{SSC} = nR\sum_{j=1}^{C}(\bar{x}_j - \bar{x})^2$$

$$\text{SSI} = n\sum_{i=1}^{R}\sum_{j=1}^{C}(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

$$\text{SSE} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij})^2$$

$$\text{SST} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$$

$$\text{df}_R = R - 1$$

$$\text{df}_C = C - 1$$

$$\text{df}_I = (R-1)(C-1)$$

$$\text{df}_E = RC(n-1)$$

$$\text{df}_T = N - 1$$

$$\text{MSR} = \frac{\text{SSR}}{R-1}$$

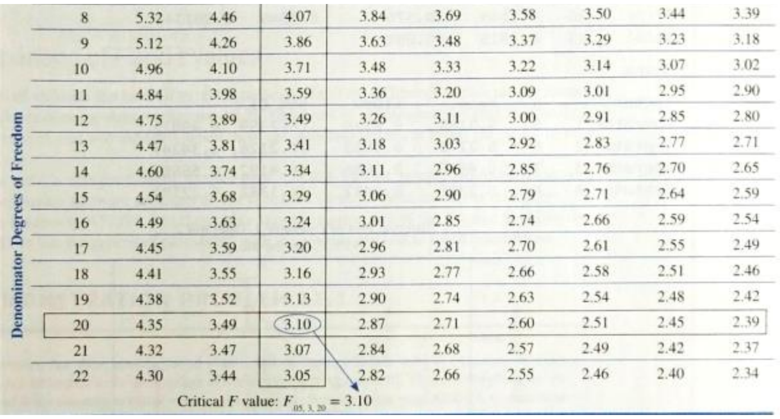
$$\text{MSC} = \frac{\text{SSC}}{C-1}$$

$$\text{MSI} = \frac{\text{SSI}}{(R-1)(C-1)}$$

$$\text{MSE} = \frac{\text{SSE}}{RC(n-1)}$$

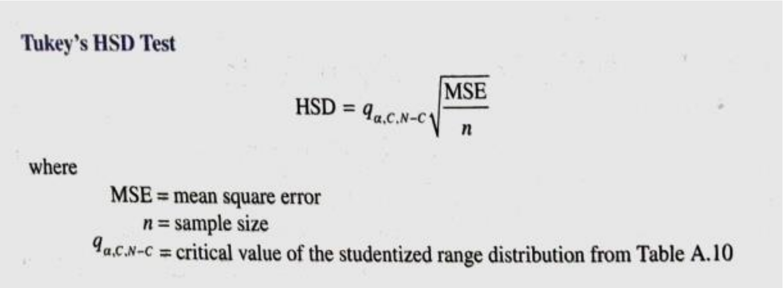
$$F_R = \frac{\text{MSR}}{\text{MSE}}$$

MSC



Denominator Degrees of Freedom									
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.47	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34

Critical *F* value: $F_{.05,3,20} = 3.10$

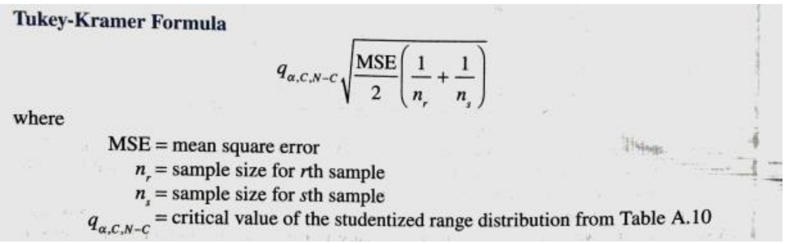


Tukey's HSD Test

$$\text{HSD} = q_{\alpha,C,N-C}\sqrt{\frac{\text{MSE}}{n}}$$

where

MSE = mean square error
n = sample size
$q_{\alpha,C,N-C}$ = critical value of the studentized range distribution from Table A.10

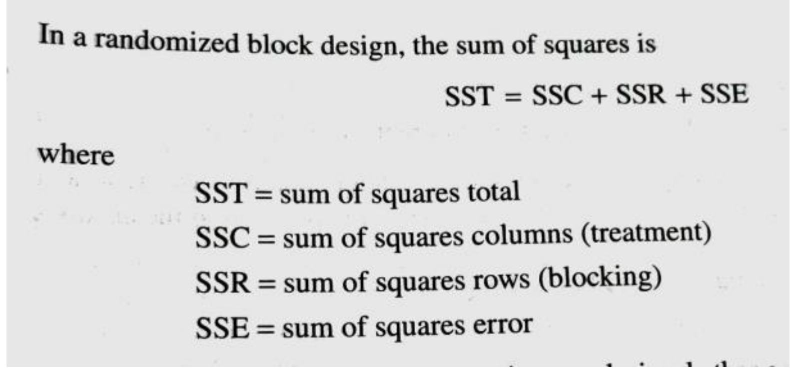


Tukey-Kramer Formula

$$q_{\alpha,C,N-C}\sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{n_r} + \frac{1}{n_s}\right)}$$

where

MSE = mean square error
n_r = sample size for *r*th sample
n_s = sample size for *s*th sample
$q_{\alpha,C,N-C}$ = critical value of the studentized range distribution from Table A.10

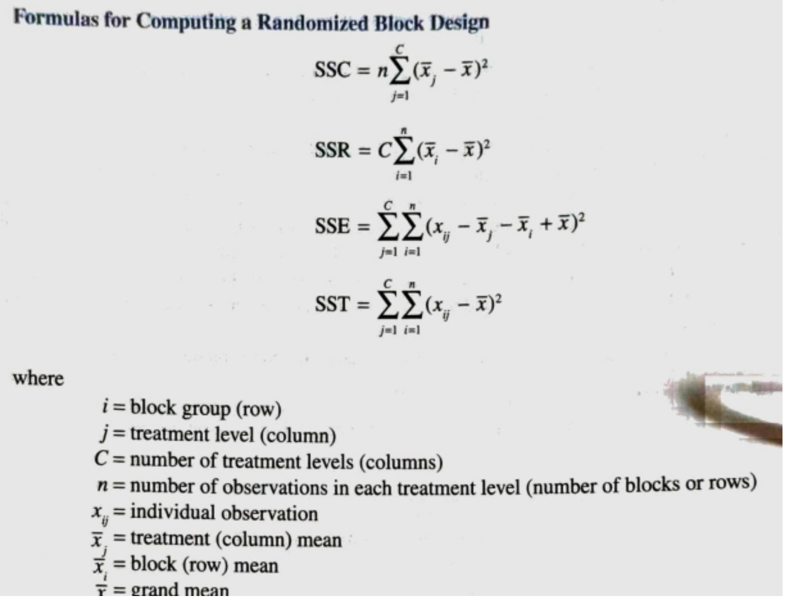


In a randomized block design, the sum of squares is

$$\text{SST} = \text{SSC} + \text{SSR} + \text{SSE}$$

where

SST = sum of squares total
SSC = sum of squares columns (treatment)
SSR = sum of squares rows (blocking)
SSE = sum of squares error



Formulas for Computing a Randomized Block Design

$$\text{SSC} = n\sum_{j=1}^{C}(\bar{x}_j - \bar{x})^2$$

$$\text{SSR} = C\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2$$

$$\text{SSE} = \sum_{j=1}^{C}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x})^2$$

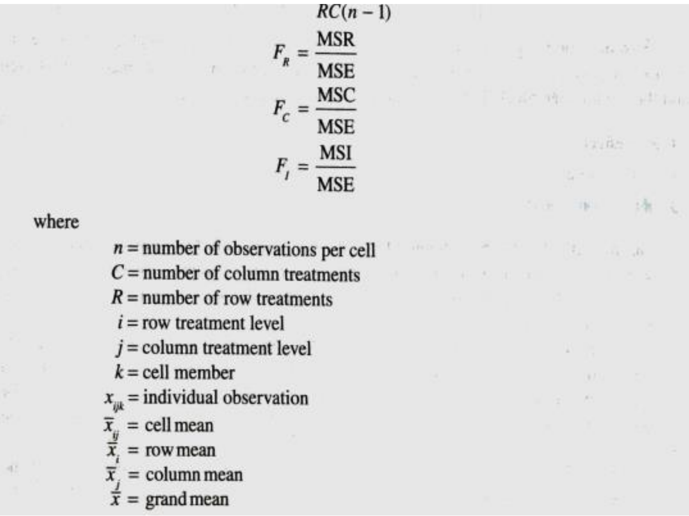
$$\text{SST} = \sum_{j=1}^{C}\sum_{i=1}^{n}(x_{ij} - \bar{x})^2$$

where

i = block group (row)
j = treatment level (column)
C = number of treatment levels (columns)
n = number of observations in each treatment level (number of blocks or rows)
x_{ij} = individual observation
$\bar{x}_j$ = treatment (column) mean
$\bar{x}_i$ = block (row) mean
$\bar{x}$ = grand mean

to occur for a given value of input.

- Specific dependent variable value: $y_i = \beta_0 + \beta_1 x_i + \in_i$
  - $x_i = the\ value\ of\ the\ independent\ variable\ for\ the\ \boldsymbol{i}th\ value$
  - $y_i = the\ value\ of\ the\ dependent\ varible\ for\ the\ \boldsymbol{i}th\ value$
  - $\beta_0 = the\ population\ y\ intercept$
  - $\beta_1 = the\ population\ slope$
  - $\in_i = the\ error\ of\ prediction\ for\ the\ \boldsymbol{i}th\ value$
- ***Least Squares Analysis***
  - The process by which a regression model is developed based on calculus techniques that attempt to produce a minimum sum of the squared error values.
  - The least squares regression line is the regression line that results in the smallest sum of errors squared.



$RC(n-1)$

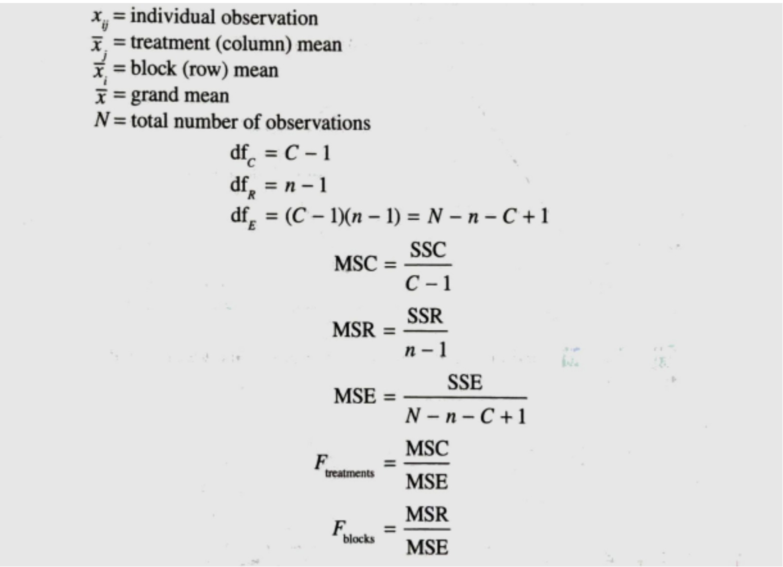
$$F_R = \frac{\text{MSR}}{\text{MSE}}$$

$$F_C = \frac{\text{MSC}}{\text{MSE}}$$

$$F_I = \frac{\text{MSI}}{\text{MSE}}$$

where

n = number of observations per cell
C = number of column treatments
R = number of row treatments
i = row treatment level
j = column treatment level
k = cell member
x_{ijk} = individual observation
$\bar{x}_{ij}$ = cell mean
$\bar{x}_i$ = row mean
$\bar{x}_j$ = column mean
$\bar{\bar{x}}$ = grand mean



x_{ij} = individual observation
$\bar{x}_j$ = treatment (column) mean
$\bar{x}_i$ = block (row) mean
$\bar{\bar{x}}$ = grand mean
N = total number of observations

$$\text{df}_C = C - 1$$

$$\text{df}_R = n - 1$$

$$\text{df}_E = (C-1)(n-1) = N - n - C + 1$$

$$\text{MSC} = \frac{\text{SSC}}{C-1}$$

$$\text{MSR} = \frac{\text{SSR}}{n-1}$$

$$\text{MSE} = \frac{\text{SSE}}{N-n-C+1}$$

$$F_{\text{treatments}} = \frac{\text{MSC}}{\text{MSE}}$$

$$F_{\text{blocks}} = \frac{\text{MSR}}{\text{MSE}}$$