PARTE I – PLANEJAMENTO DE EXPERIMENTOS

ANOVA 1 FATOR

Prof. Anna Carla Araujo
COPPE/UFRJ

CEFCON
CENTRO DE ESTUDOS EM FABRICAÇÃO E COMANDO NUMÉRICO

COPPE UFRJ

Engenharia Mecânica
Politécnica ▪ COPPE UFRJ

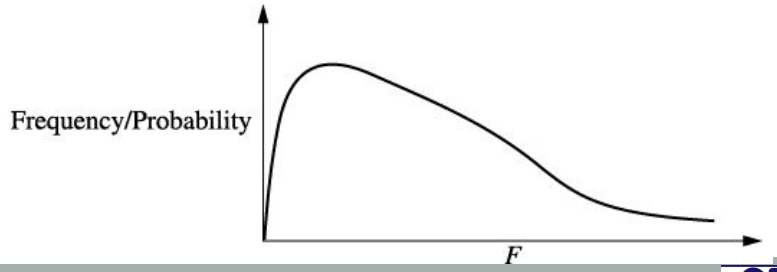
Limitações do teste com amostras não pareadas

- O número de grupos está limitado a dois
- A interação entre os grupos não pode ser testada
- Não há como relacionar as variáveis pareadas

Distribuição F

+ Avalia a razão entre a variância de duas variáveis independentes com variâncias S1 e S2 e número de amostras n1 e n2.

$$F = \frac{[(n_1 - 1)S_1^2/\sigma_1^2]/(n_1 - 1)}{[(n_2 - 1)S_2^2/\sigma_2^2]/(n_2 - 1)}$$



Introdução a Análise de Variância

Table 2.2 Sales Data of Salesmen

		Salesman j						
		1	2	3	.	.	j	.
	1	Y_{11}	Y_{12}	Y_{13}	.	.	Y_{1j}	.
	2	Y_{21}	Y_{22}	Y_{23}	.	.	Y_{2j}	.
	3	Y_{31}	Y_{32}	Y_{33}	.	.	Y_{3j}	.
Replication i	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	i	Y_{i1}	Y_{i2}	Y_{i3}	.	.	Y_{ij}	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	n	Y_{n1}	Y_{n2}	Y_{n3}	.	.	Y_{nj}	.

- Y_{ij} is the ith sales data (ith replication) of the jth salesman.
- Salesman is the factor which has effect on the response variable Y_{ij}. Let it be factor A.
- a is the number of salesmen of the company. This is also known as the number of levels/treatments of factor A.
- n is the number of sales data under each salesman. This is also known as the number of replications under each level of factor A.

Modelo:

$$Y_{ij} = \mathrm{n} + A_j + e_{ij}$$

where

n is the overall mean of the sales revenue.

A_j is the effect of the jth treatment of factor A (salesman) on the response, and

e_{ij} is the random error associated with the ith replication of the jth treatment of factor A.

The decomposition of total variability into its component parts is called analysis of variance (ANOVA). The partitioning of the total variability is presented below.

The total corrected sum of squares is given by

$$SS\text{Total} = \sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{..})^2$$

The above equation is re-written by adding and subtracting the term $\overline{Y}_{.j}$ as per the following presentation:

$$\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}[(\overline{Y}_{.j} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{.j})]^2$$

$$\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}[(\overline{Y}_{.j} - \overline{Y}_{..})^2 + \sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{.j})^2 + 2\left(\sum_{i=1}^{n}\sum_{j=1}^{a}(\overline{Y}_{.j} - \overline{Y}_{..})\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{.j})\right)$$

The following part of the last term on the right hand side of the equation is equal to 0 because of the following:

$$\sum_{i=1}^{n}(Y_{ij} - \overline{Y}_{.j}) = (Y_{.j} - n\overline{Y}_{.j}) = \left[Y_{.j} - n\left(\frac{Y_{.j}}{n}\right)\right] = 0$$

Therefore, the total sum of squares corrected is reduced to the following:

$$\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{a}(\overline{Y}_{.j} - \overline{Y}_{..})^2 + \sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{.j})^2$$

$$\sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{..})^2 = n\sum_{j=1}^{a}(\overline{Y}_{.j} - \overline{Y}_{..})^2 + \sum_{i=1}^{n}\sum_{j=1}^{a}(Y_{ij} - \overline{Y}_{.j})^2$$

☐ Total sum of squares = Sum of squares of treatment + Sum of squares of error

$$SS\text{Total} = SS_A + SS\text{Error}$$

☐ Total sum of squares = Sum of squares of treatment + Sum of squares of error

$$SS_{\text{Total}} = SS_A + SS_{\text{Error}}$$

Mean sum of squares of treatment = $\dfrac{\text{Sum of squares of treatment}}{\text{Degrees of freedom of treatment}} = \dfrac{SS_A}{(a-1)}$

Mean sum of squares of error = $\dfrac{\text{Sum of squares of error}}{\text{Degrees of freedom of error}}$

$$= \frac{SS_{\text{Error}}}{a(n-1)} = \frac{SS_{\text{Error}}}{(N-a)}$$

where N is the total number of observations in the experiment (an).

The hypotheses of the model are given as:

Null hypothesis, H_0: $\overline{Y}_1 = \overline{Y}_2 = \overline{Y}_3 = \overline{Y}_4 = \cdots = \overline{y}_a$

Alternate hypothesis, H_1: Treatment means are not equal for at least one pair of the treatment means. The generalized results are summarized in Table 2.3.

Table 2.3 Generalized Results of ANOVA with Single Factor

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	*F*-ratio
Between treatments	$a - 1$	$SS_{\text{treatment}}$	$SS_{\text{treatment}}/(a-1)$	$MSS_{\text{treatment}}/MSS_{\text{error}}$
Within treatments	$a(n-1)$	SS_{error}	$SS_{\text{error}}/[a(n-1)]$	
Total	$N - 1$	SS_{total}		

Teste da média para amostras pareadas

Hipótese H0: médias são iguais, logo a diferença é zero.
Hipótese H1: médias são diferentes, logo a diferença não é zero.

Variância Total:

$$\bar{\bar{y}}_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{a} y_{ij}/N$$

$$SStotal = \sum_{i=1}^{n}\sum_{j=1}^{a}(Yij - \bar{\bar{y}}_{ij})^2$$

Variância entre as m amostras:

Média da amostra j
j: [1,a]

$$\bar{y}_i = \frac{1}{a}\sum_{j=1}^{a} y_{ij}$$

$$SStreat = \sum_{i=1}^{n} n_j(\bar{y}_i - \bar{\bar{y}}_{ij})^2$$

Variância dentro das amostras:

$$SSerro = \sum_{i=1}^{n}\sum_{j=1}^{a}(Yij - \bar{y}_i)^2$$

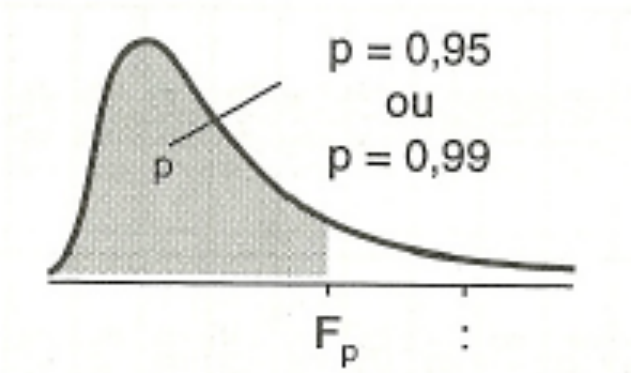
Tabela de Análise de Variância

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrados Médios	Variavel de teste F
Entre amostras	S_entre	a-1	$\frac{SStreat}{a-1}$	$\frac{MSStreat}{MSSerro}$
Dentro das amostras	S_erro	N-a	$\frac{SSerro}{N-a}$	
Total	S_total	N-1	$\frac{S_{total}}{(N-1)}$	

Teste de Hipótese da ANOVA

Hipótese H0: médias são iguais, logo a diferença é zero.
Hipótese H1: médias são diferentes, logo a diferença não é zero.

Uma vez fixado o nível de significância α do teste Ho deve ser rejeitado se F(observações) > F (1-α)



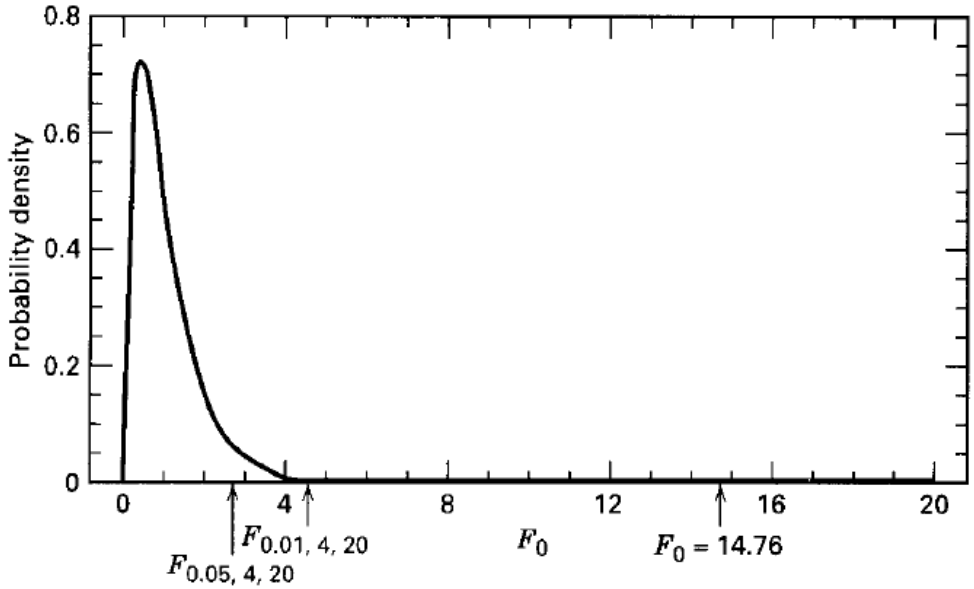


Figure 3-3 The reference distribution ($F_{4,20}$) for the test statistic F_0 in Example 3-1.

Exemplo

Fornece os quantis $F_{0,95}$ (em cima) e $F_{0,99}$ (embaixo) em função do nº de g.l. numerador ν_1 (coluna) e do nº de g.l. denominador ν_2 (linha)
F tem distribuição F com ν_1 g.l. no numerador e ν_2 g.l. no denominador $P(F \leq F_{0,95}) = 0,95$ e $P(F \leq F_{0,99}) = 0,99$

ν_2 \ ν_1	1	2
1	161,45	199,50
	4052,18	4999,50
2	18,51	19,00
	98,50	99,00
3	10,13	9,55
	34,12	30,82
4	7,71	6,94
	21,20	18,00
5	6,61	5,79
	16,26	13,27
6	5,99	5,14
	13,75	10,92
7	5,59	4,74
	12,25	9,55
8	5,32	4,46
	11,26	8,65
9	5,12	4,26
	[illegible]	8,02
10	4,96	4,10
	10,04	7,56
20	4,35	3,49
	8,10	5,85
40	[illegible]	3,23

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Quadrados Médios	F
Entre amostras	48.05	1	48.05	1.9581
Dentro das amostras	441.7	18	24.54	
Total	489.75	19	25.78	

Uma vez fixado o nível de significância α do teste, Ho (médias iguais) deve ser rejeitado se F(observações) > F_teste (1-α)

F_obs= 1.96 < F_teste=4.5

Ho é aceito, as médias são iguais!!!!

Exemplo 2 (Montgomery)

Weight Percentage of Cotton	Observed Tensile Strength (lb/in^2)					Totals $y_{i.}$	Averages $\bar{y}_{i.}$
	1	2	3	4	5		
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						$y_{..} = 376$	$\bar{y}_{..} = 15.04$

We will use the analysis of variance to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against the alternative H_1: some means are different. The sums of squares required are computed as follows:

Weight Percentage of Cotton	Observed Tensile Strength (lb/in^2) 1	2	3	4	5	Totals $y_{i.}$	Averages $\bar{y}_{i.}$
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						$y_{..} = 376$	$\bar{y}_{..} = 15.04$

Table 3-4 Analysis of Variance for the Tensile Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Cotton weight percentage	475.76	4	118.94	$F_0 = 14.76$	<0.01
Error	161.20	20	8.06		
Total	636.96	24			

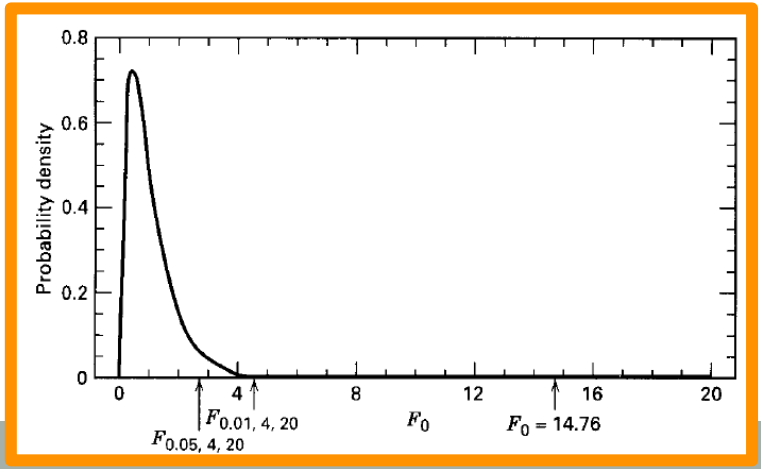
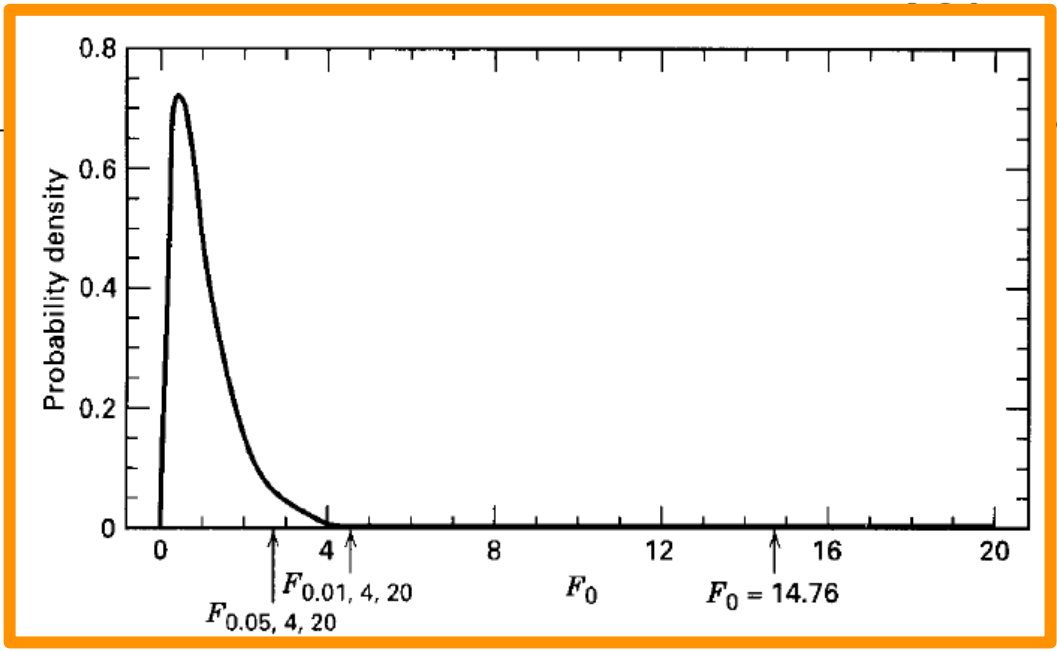


Table 3-4 Analysis of Variance for the Tensile Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Cotton weight percentage	475.76	4	118.94	$F_0 = 14.76$	<0.01
Error					
Total					



The analysis of variance is summarized in Table 3-4. Note that the between-treatment mean square (118.94) is many times larger than the within-treatment or error mean square (8.06). This indicates that it is unlikely that the treatment means are equal. More formally, we can compute the F ratio $F_0 = 118.94/8.06 = 14.76$ and compare this to an appropriate upper-tail percentage point of the $F_{4,20}$ distribution. Suppose that the experimenter has selected $\alpha = 0.05$. From Appendix Table IV we find that $F_{0.05,4,20} = 2.87$. Because $F_0 = 14.76 > 2.87$, we reject H_0 and conclude that the treatment means differ; that is,

Relação entre ANOVA e ajuste de Modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, N,$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Cap 1 – Wu e Hamada

The unknown parameters in the model are the regression coefficients $\boldsymbol{\beta}$ and the error variance σ^2. Thus, the purpose for collecting the data is to estimate and make inferences about these parameters. For estimating $\boldsymbol{\beta}$, the least squares criterion is used; i.e., the least squares estimators (LSEs), denoted by $\hat{\boldsymbol{\beta}}$, minimize the following quantity:

$$\sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)\right)^2 \tag{1.5}$$

which in matrix notation is

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.6}$$

The solution to this equation is the **least squares estimate** which is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.10}$$

In fitting the model, one wants to know if any of the variables (regressors, predictors, covariates) has explanatory power. None of them has explanatory power if the null hypothesis

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \tag{1.11}$$

In fitting the model, one wants to know if any of the variables (regressors, predictors, covariates) has explanatory power. None of them has explanatory power if the null hypothesis

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \tag{1.11}$$

holds. In order to test this null hypothesis, one needs to assess how much of the total variation in the response data can be explained by the model relative to the remaining variation after fitting the model, which is contained in the residuals.

Table 1.1 ANOVA Table for General Linear Model

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	p	$\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2$	$(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - N\bar{y}^2)/p$
residual	$N-p-1$	$(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$	$(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/(N-p-1)$
total (corrected)	$N-1$	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

If the null hypothesis (1.11) holds, the F statistic

$$\frac{\left(\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} - N\bar{\mathbf{y}}^2\right)/p}{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)/(N-p-1)} \tag{1.15}$$

(the regression mean square divided by the residual mean square) has an F distribution with parameters p and $N-p-1$, which are the degrees of freedom of its numerator and denominator, respectively. The p value is calculated by evaluating

$$Prob\left(F_{p,N-p-1} > F_{obs}\right), \tag{1.16}$$

where $Prob(\cdot)$ denotes the probability of an event, $F_{p,N-p-1}$ has an F distribution with parameters p and $N-p-1$, and F_{obs} is the observed value of the F statistic. The F critical values can be found in Appendix D. The p value in (1.16) can be obtained from certain pocket calculators or by interpolating the values given in Appendix D. An example of an F distribution is given in Figure 1.4 along with its critical values.

Note that the **p value** gives the probability under the null hypothesis that the F statistic value for an experiment conducted in comparable conditions will exceed the observed value F_{obs}. The smaller the p value, the stronger is the evidence that the null hypothesis does not hold. Therefore it provides a quantitative measure of the significance of effects in the experiment under study. The same interpretation can be applied when other test statistics and null hypotheses are considered.

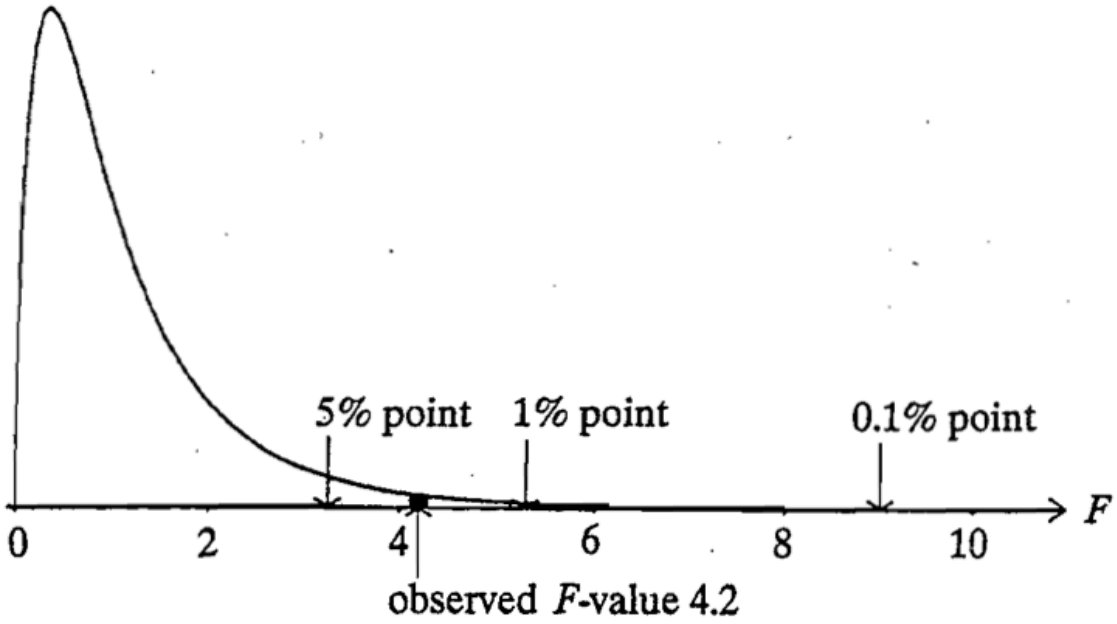


Figure 1.4. Observed F Value of 4.20 in Relation to an F Distribution With 3 and 16