ANOVA
(ANalysis Of VAriance)

Prof. Lorí Viali, Dr.
http://www. ufrgs.br/~viali/
viali@mat.ufrgs.br

Conceitos Básicos

Objetivos

A Análise de variância (ANOVA) É utilizada para mostrar os efeitos principais de variáveis categóricas independentes (denominadas de fatores) sobre uma variável quantitativa dependente.

O Modelo Linear Geral (GLM - General Linear Model) suporta, também, variáveis categóricas dependentes.

Um "*efeito principal*" é um efeito direto de uma variável independente sobre a variável dependente. Um "*efeito de interação*" é o efeito de duas ou mais variáveis independentes sobre a variável dependente.

Os modelos de regressão não podem manejar interações a menos que um termo de produto cruzado seja explicitamente adicicionado. A ANOVA mostra efeitos de interação como resultado da própria técnica.

Existe uma variante para a utilização de variáveis de controle quantitativas denominada de ANCOVA (Analysis of Covariance). Existe, também, para o caso de múltiplas variáveis dependentes a MANOVA (Multiple analysis of Variance) e finalmente existe uma combinação das duas denominada de MANCOVA (MANOVA + ANCOVA).

A estatística teste na ANOVA é a F (de Snedcor) que testa a diferença entre as médias grupos. Ela testa se as médias dos grupos formados pelos valores da variável independente (ou combinação de valores para as múltiplas variáveis independentes) pode ter ocorrido por acaso.

Se as médias dos grupos não diferem significativamente então pode-se assumir que a variável independente não tem efeito sobre a variável dependente.

One-Way ANOVA

Testa a diferença entre uma única variável quantitativa dependente contra dois, três ou mais grupos formados pelas categorias de uma uma única variável categórica independente. É também conhecida cono ANOVA univariada, ANOVA de classificação simples ou ANOVA de um fator.

Two-Way ANOVA

A Two-way ANOVA ou Análise de Variância de dupla classificação analisa uma variável quantitativa dependente em termos de categorias (grupos) de duas variáveis qualitativas independentes, uma das quais pode ser considerada como variável de controle.

n-Way ANOVA ou MANOVA

Generaliza a ANOVA lidando com “n” variáveis independentes. Note-se que o número de interações cresce neste caso. Duas variáveis independentes apresentam uma única interação de primeira ordem (AB). Três variáveis independentes apresentam três interações de primeira ordem (AB,AC,BC) e uma de segunda-ordem (ABC), ou seja, quatro no total.

Quatro variáveis independentes apresentam seis interações de primeira ordem (AB, AC, AD, BC, BC, CD), três de segunda-ordem (ABC, ACD, BCD) e uma de terceira ordem (ABCD). A medida que o número de interações aumenta torna-se extremamente difícil interpretar o modelo.

Projetos (Designs)

A ANOVA e a ANCOVA apresentam vários projetos experimentais. O tipo de projeto (desenho) afeta o cálculo da razão F. Independente do projeto a tabela de saída é interpretada de forma semelhante - a significância da razão F indica a significância de cada efeito principal e de cada efeito de interação (e o efeito de cada covariável ANCOVA).

ANOVA entre Grupos

Quando a variável dependente é medida em grupos independentes de amostras membros,onde cada grupo é exposto a uma condição diferente, o conjunto de condições é denominado de fatores entre conteúdos. Os grupos correspondem a condições que são categorias da variável independente. As condições são atribuídas aleatoriamente pelo pesquisador.

Delineamentos (Designs)

Este é o desenho usual de Análise de Variância. Existe um conjunto de observações, os "grupos" que se referem a subconjuntos associados com cada categoria da variável independente (One-Way ANOVA) ou com cada célula formada por múltiplas variáveis categóricas independentes (na ANOVA Multivarida).

Após as mensurações de cada grupo, as variâncias da variável dependente entre grupos e dentro dos grupos é calculada. Se apenas o acaso está intervindo pode-se esperar que as duas variâncias sejam semelhantes. Se a variância entre grupos é maior que a dentro dos grupos medida pela razão F, então pode-se concluir que o fator de agrupamento (a variável independente apresenta efeito significativo).

Delineamento Completamente Casualizados

Um Delineamento Completamente Casualizado é uma ANOVA entre-grupos. A Aleatorização é um esforço para controlar todos os fatores não mensurados. Se existe uma razão para supor que alguma variável independente adicional é importante, esta variável pode ser controlada explicitamente por blocos casualizados se categórica ou pela ANCOVA se for uma variável contínua.

Delineamento em Blocos Completos Casualizados

Este delineamento é um desenho experimental na qual as observações são emparelhados em alguma variável de controle. As observações são divididas em grupos com base nesta variável (às vezes chamada de "variável ruído").

Delineamento Quadrado Latino

O delineamento do Quadrado Latino extende a lógica do design de blocos para controlar duas variáveis categóricas. Este tipo de delineamento, também, reduz o número de observações necessárias para computar a ANOVA.

Este delineamento requer que se assuma que todos os efeitos de interação sejam zero. Normalmente, se existirem três variáveis com cada uma assumindo quatro valores então serão necessários $4^3 = 64$ observações apenas para se ter uma única observação para cada possível observada. Com o delineamento Quadrado Latino, no entanto, o número necessário de observações é reduzido para $4^2 = 16$.

Delineamento Quadrado-Grego-Latino

Um delineamento Quadrado Greco-Latino estende o design de blocos para controlar três variáveis categóricas.

ANOVA Fatorial

Uma ANOVA Factorial é para mais de um fator (mais do que uma variável independente, isto é, para a análise two-way ou acima) e é utilizada para acessar a importância relativa das várias combinações das variáveis independnetes.

Em um projeto fatorial, todas as possíveis combinações dos níveis das variáveis independentes são representadas como grupos na análise. Com tal desenho a ANOVA não é uma forma separada de design mas uma forma de combinar os delineamentos.

Uma tabela matricial do projeto mostra as interseções das categorias das variáveis independentes. Uma tabela ANOVA correspondente é construída onde as colunas são as várias covariáveis (na ANCOVA) e os efeitos principais e de interação.

Os fatores são variáveis categóricas independentes. As categorias de um fator são seus grupos ou níveis. Utilizando terminologia da ANOVA 2 x 3 ("two-by-three") delineamento fatorial significa que existem dois fatores com o primeiro tendo duas categorias e o segundo três, para um total de seis grupos (níveis).

Um delineamento fatorial 2x2x2 apresenta três fatores, cada um com duas categorias. A ordem dos fatores não faz diferença. Se eles forem multiplicados tem-se o úmero de grupos (algumas vezes "grupos de tratamento") formados por todas as independentes coletivamente.

Assim um delineamento 2x3 tem seis grupos e um 2x2x2 tem 8 grupos. Na pesquisa experimental um número igual de observações são atribuídas para cada grupo aleatoriamente.

GLM (Modelo Linear Generalizado)

O GLM (General Linear Model) Modelo Linear Geral da ANOVA é um substituto da ANOVA fatorial a partir da versão do SPSS 8. A abordagem GLM é mais geral e suporta o uso de variáveis dependentes categóricas.

Modelos de Efeitos Aleatórios

Algumas modelos de ANOVA são modelos efeitos fixos ("Model I"), isto é, os dados são coletados em todas as categorias das variáveis independentes. Nos modelos de Efeitos Aleatórios ("Model II"), os dados são coletados somente para uma amostra das categorais.

One-Way ANOVA

One-Way ANOVA

O Modelo: $Y_{ij} = u_i + \mathcal{U}_{ij}$

Cada valor obsevado da variável quantitativa dependente Y_{ij} é dado pela soma da média (μ_i) da população de onde este valor foi retirado mais um erro aleatório ($\mathcal{U}_{ij}$).

Suposições:

1) Os erros são variáveis aleatórias com média zero, isto é, $E(U_{ij}) = 0$, para $i = 1, 2, ..., k$ e $j = 1, 2, ..., n$;

2) Os erros são variáveis aleatórias independentes, isto é, $E(U_{ij}.U_{hl}) = 0$, se $i \neq h$ e $j \neq l$;

3) Os erros apresentam variância constante, isto é, $E(U_{ij}^2) = \sigma^2$, para $i = 1, 2, ..., k$ e $j = 1, 2, ..., n_i$;

4) Os termos erro U_{ij} seguem uma normal.

Resumindo:

Supõem-se que os valores Y_{ij} são valores que resultam da adição de um valor médio μ_i com um termo erro U_{ij} que são variáveis aleatórias independentes com distribuição normal de média zero e variância constante igual a σ^2.

Metodologia:

Fazendo $\mu_i = \mu + \alpha_i$, *onde os* $\alpha_\imath$ *são os efeitos dos tratamentos, o modelo fica:*

$$Y_{ij} = \mu + \alpha_i + \mathcal{U}_{ij}$$

Os $\alpha_\imath$, *estão sujeitos a restrição* $\Sigma n_i \alpha_i = 0$.

Então de $\mu_i = \mu + \alpha_i$, *segue que:* $\mu = \frac{1}{n}\sum_i n_i \mu_i$.

Fazendo m_i *indicar as estimativas de* μ_i $(i = 1, 2, ..., k)$. *Tem-se que:*

$$Y_{ij} = m_i + E_{ij}$$

Onde E_{ij} *é o desvio da j=ésima observação em relação a estimativa da média do tratamento i.*

Dados os valores Y_{ij} com $i = 1, 2, ..., k$ e $j = 1,2, ..., n_i$, de acordo com o Método dos Mínimos Quadrados, as estimativas de m_i são os valores que minimizam a soma dos quadrados dos desvios ou soma residual, dada por:

$$Q = S.Q.R = \sum_{i}^{k} \sum_{j}^{n_i} E_{ij}^2 = \sum_{i}^{k} \sum_{j}^{n_i} (Y_{ij} - m_i)^2$$

Derivando e igualando a zero, tem-se:

$$\frac{\partial Q}{\partial m_i} = 2 \sum_{j}^{n_i} (Y_{ij} - m_i)(-1) = 0$$

Segue, então: $$n_i m_i = \sum_{j}^{n_i} Y_{ij}$$

Ou

$$m_i = \frac{\sum_{j}^{n_i} Y_{ij}}{n_i} = \overline{Y}_i$$

Isto é, o estimador de Mínimos Quadrados para a média do i-ésimo tratamento é a média aritmética das observações deste tratamento.

Indicando por $\mathcal{A}_i$ o total do i-ésimo tratamento, isto é, fazendo: $\mathcal{A}_i = \sum_{j=1}^{n_i} Y_{ij}$

Tem-se: $$m_i = \overline{Y}_i = \frac{\mathcal{A}_i}{n_i}$$

As Somas dos Quadrados

$$S.Q.R = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - m_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

Elevando o binômio ao quadrado, segue:

$$S.Q.R = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - 2\sum_{i=1}^{k}\overline{Y}_i \sum_{j=1}^{n_i} Y_{ij} + \sum_{i=1}^{k} n_i \overline{Y}_i^2$$

Substituindo as expresssões anteriores

e simplificando, tem-se:

$$S.Q.R = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}\frac{A_i^2}{n_i}$$

Pela definição, tem-se:

$$S.Q.Total = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$$

Onde: $$\bar{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

Pode-se verificar, também:

$$S.Q.Total = \sum_{i}^{k} \sum_{j}^{n_i} Y_{ij}^{2} - \frac{G^{2}}{n}$$

Onde: $$G = \sum_{i}^{k} \sum_{j}^{n_i} Y_{ij} = \sum_{i}^{k} A_i$$

Pela definição, a soma de quadrados dos tratamentos é:

$$S.Q.Trat. = \sum_{i}^{k} n_i(\overline{Y}_i - \overline{Y})^2$$

Lembrando que:

$$m_i = \overline{Y}_i = \frac{A_i}{n_i} \quad e \quad \sum_{i=1}^{k} \alpha_i = 0$$

Tem-se: $$S.Q.Trat. = \sum_{i}^{k} \frac{A_i^2}{n_i} - \frac{G^2}{n}$$

Juntando os resultados, segue que:

$$S.Q.Res. = S.Q.Total - S.Q.Trat. \quad \text{ou}$$

$$S.Q.Total = S.Q.Trat. + S.Q.Res.$$

Esta expressão mostra que a Soma dos Quadrados Totais é composta de duas parcelas:

A Soma dos Quadrados dos Tratamentos (variação **entre** tratamentos) e a Soma dos Quadrados dos Resíduos (variações **dentro** de tratamentos).

Exemplo:

Considere-se os valores Y_{ij} de três amostras supostamente independentes:

Am. 1	Am. 1	Am. 1
10	15	20
09	13	19
12	12	17
11	17	15
13	14	16
55	16	18
	87	14
		119

Tem-se:

$$n_1 = 5,\ n_2 = 6 \ \text{e}\ n_3 = 7 \quad n = 18 \ \text{e}\ k = 3$$

$$A_1 = 55,\ A_2 = 87 \ \text{e}\ A_3 = 119$$

$$\bar{Y}_1 = 11;\ \bar{Y}_2 = 14,5\ ;\ \bar{Y}_3 = 17 \ \text{e}\ \bar{Y} = 14,5$$

$$S.Q.Total = \sum_{i}^{k}\sum_{j}^{n_i} Y_{ij}^2 - \frac{G^2}{n} = 160,50$$

$$S.Q.Trat. = \sum_{i}^{k}\frac{A_i^2}{n_i} - \frac{G^2}{n} = 3889,50 - 3784,5 = 105$$

A soma dos resíduos vale:

$$S.Q.R = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}\frac{A_i^2}{n_i} =$$

$$= 3945 - 3889{,}50 = 55{,}50$$

$$S.Q.Total = S.Q.Trat. + S.Q.Res.$$

Assim: $160{,}50 = 105 + 55{,}50$

Espectância da Somas de Quadrados

Temos: $E(Y_{ij}) = \mu_i$ *e* $E(Y_{ij}^2) = \mu_i^2 + \sigma^2$

Então: $$E\left(\sum_{i}^{k}\sum_{j}^{n_i} Y_{ij}^2\right) = \sum_{i=1}^{k} n_i \mu_i^2 + n\sigma^2$$

Como: $$A_i = \sum_{j=1}^{n_i} Y_{ij}$$

De acordo com o modelo, tem-se:

$$\mathcal{A}_i = n_i \mu_i + \sum_{j=1}^{n_i} \mu_{ij}$$

Segue, então:

$$\mathcal{A}_i^2 = n_i^2 \mu_i^2 + 2 n_i v_i \sum_{j=1}^{n_i} \mu_{ij} + \left(\sum_{j=1}^{n_i} \mu_{ij} \right)^2$$

Mas:

$$\mathcal{E}\left(\sum_{j=1}^{n_i} \mu_{ij} \right) = 0$$

Com l ≠ j, segue:

$$E\left(\sum_{j=1}^{n_i} \mu_{ij}\right)^2 = E\left(\sum_{j=1}^{n_i} \mu_{ij}^2 + \sum_{l}^{n_i} \mu_{ij}\mu_{il}\right) = n_i\sigma^2$$

Daí: $E(A_i^2) = n_i^2\mu_i^2 + n_i\sigma^2$

Como: $G = \sum_{i=1}^{k} n_i\mu_i + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \mu_{ij}$

Segue:

$$G^2 = \left(\sum_{i=1}^{k} n_i \mu_i\right)^2 + 2.\left(\sum_{i=1}^{k} n_i \mu_i\right)\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mu_{ij}\right) + \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mu_{ij}\right)^2$$

Mas:

$$E\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} \mu_{ij}\right) = 0$$

E com $h \neq i$ e/ou $l \neq k$, segue

$$E\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}\right)^2 = E\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mu_{ij}^2 + \sum_{i=1}^{k}\sum_{h=1}^{k}\sum_{l=1}^{n_i}\sum_{j=1}^{n_i}\mu_{ij}\mu_{hl}\right) = n\sigma^2$$

Assim: $$E(G^2) = \left(\sum_{i=1}^{k} n_i\mu_i\right)^2 + n\sigma^2$$

Portanto:

$$E(S.Q.Total) = \sum_{i=1}^{k} n_i\mu_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i\mu_i\right)^2 + (n-1)\sigma^2$$

Ou: $E(S.Q.Total) = W + (n-1)\sigma^2$

$$W = \sum_{i=1}^{k} n_i \mu_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \mu_i\right)^2$$

Ou: $W = \sum_{i=1}^{k} n_i (\mu_i - \mu)^2$

A expressão mostra que W = 0, apenas se

$\mu_1 = \mu_2 = \ldots = \mu_k = \mu$

Para os tratamentos, tem-se:

$$E(S.Q.Trat.) = \sum_{i=1}^{k}(n_i\mu_i^2 + \sigma^2) - \frac{1}{n}\left(\sum_{i=1}^{k} n_i\mu_i\right)^2 - \sigma^2 =$$

$$= \sum_{i=1}^{k} n_i\mu_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i\mu_i\right)^2 + (k-1)\sigma^2$$

Ou

$$E(S.Q.Trat.) = W + (k-1)\sigma^2$$

Causa da Variação	Soma dos Quadrados	Espec. da Soma	Esp. da Soma (sob H_0)
Tratamentos	$\sum_{i}^{k} \frac{A_i^2}{n_i} - \frac{G^2}{n}$	$W + (k - 1)\sigma^2$	$(k - 1)\sigma^2$
Resíduo	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} \frac{A_i^2}{n_i}$	$(n - k)\sigma^2$	$(n - k)\sigma^2$
Total	$\sum_{i}^{k} \sum_{j}^{n_i} Y_{ij}^2 - \frac{G^2}{n}$	$(n - 1)\sigma^2$	$(n - 1)\sigma^2$

Os Quadrados Médios

Consideremos a hipótese de nulidade:

$\mathcal{H}_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$

Isto é, consideremos a hipótese de que as médias das "k" populações sob análise sejam idênticas.

Sob esta hipótese, o valor W, definido anteriormente é igual a zero. Então, tem-se:

$E(S.Q.Total) = (n - 1)\sigma^2$ e ainda que;

$E(S.Q.Trat.) = (k - 1)\sigma^2$

Pode-se mostrar que se, os μ_{ij} são variáveis aleatórias independentes com distribuição normal de média "zero" e variância σ^2 então:

$(S.Q.Res)/\sigma^2$ tem uma distribuição Qui-Quadrado com "$n - k$" graus de liberdade.

Além disso, pode-se demonstrar que sob H_0: $(S.Q.Trat.)/\sigma^2$ tem uma distribuição Qui-Quadrado com "$k - 1$" graus de liberdade e $(S.Q.Total)/\sigma^2$ tem uma distribuição Qui-Quadrado com "$n - 1$" graus de liberdade e as três distribuições são independentes entre si.

Por definição o Quadrado Médio é o quociente entre a Soma dos Quadrados pelo respectivo "Número de Graus de Liberdade". Desta forma, o Quadrado Médio dos Tratamentos é:

$$Q.M.Trat = (S.Q.Trat.)/(k - 1)$$

$$Q.M.Res. = (S.Q.Res.)/(n - k)$$

Substituindo alguns resultados anteriores, tem-se:

$$E(Q.M.Trat.) = \sigma^2 + W/(k-1) \text{ e}$$

$$E(Q.M.Res.) = \sigma^2$$

A tabela, seguinte, resume alguns resultados.

Causa da Variação	Grau de Liberdade (G.L.)	Soma dos Quadrados (S. Q.)	Espec. Do Quadrado Médio	Esp. da Soma (sob H_0)
Tratamentos	$k - 1$	$\sum_{i}^{k} \frac{A_i^2}{n_i} - \frac{G^2}{n}$	$\frac{W}{k-1} + \sigma^2$	σ^2
Resíduos	$n - k$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} \frac{A_i^2}{n_i}$	σ^2	σ^2
Total	$n - 1$	$\sum_{i}^{k} \sum_{j}^{n_i} Y_{ij}^2 - \frac{G^2}{n}$		

O Teste F

Pode-se mostrar que se X_1 e X_2 são variáveis aleatórias independentes com distribuições Qui-Quadrado de g_1 e g_2 graus de liberdade, respectivamente, então a variável resultante do quociente: $(X_1/g_1)/(X_2/g_2)$ apresenta uma distribuição F com g_1 e g_2 graus de liberdade. Anota-se $F(g_1; g_2)$.

A Distribuição F (de Snedecor)

Uma variável aleatória X tem uma distribuição "F" ou de Snedecor se sua fdp for do tipo:

$$f(x)=\begin{cases} \dfrac{\Gamma\left(\dfrac{m+n}{2}\right) m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (n+mx)^{-\frac{m+n}{2}}}{\Gamma\left(\dfrac{m}{2}\right)\Gamma\left(\dfrac{n}{2}\right)} & se\ x>0 \\ 0 & se\ x \leq 0 \end{cases}$$

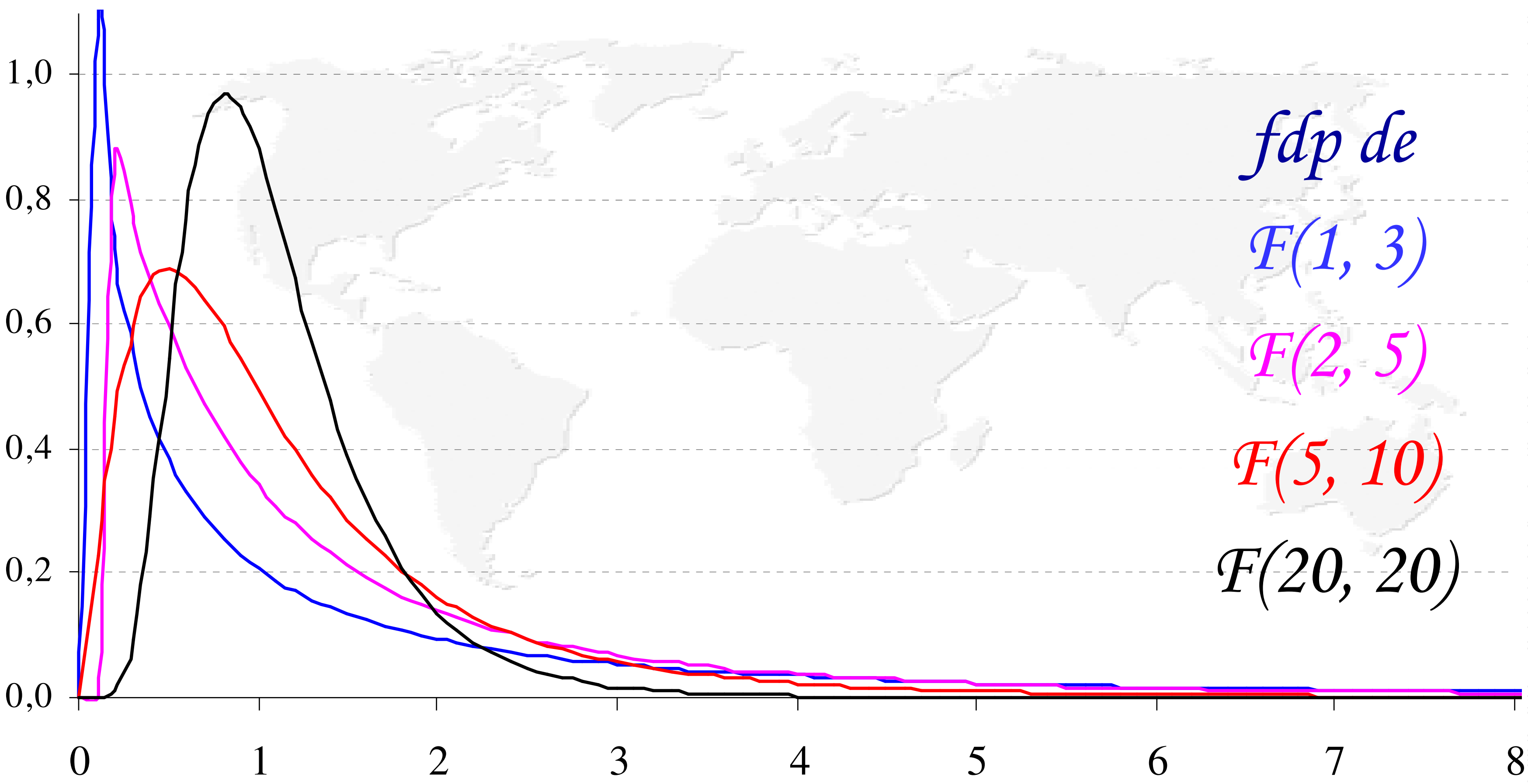
Expectância e Variância

$$E(X) = \frac{m}{m-2}$$

$$Var(X) = \frac{2(m+n-2)-m^2}{m(n-2)(n-4)}$$

m é o grau de liberdade do numerador e **n** do denominador

Diagramas



Planilha

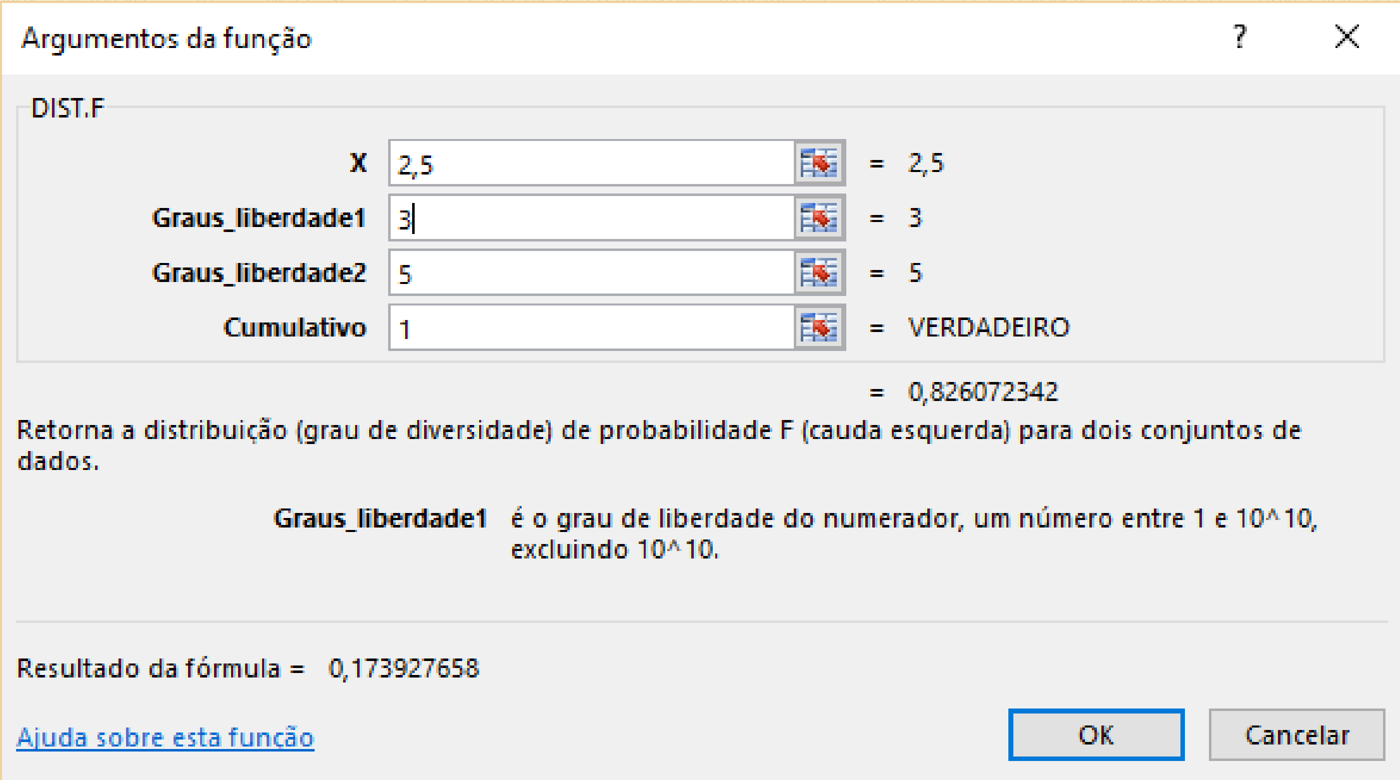
O que é tabelado é a área à direita de cada curva (função direta), isto é, dado um certo valor de "f", tem-se: $P[F(m, n) \geq f] = \alpha$, *ou dado uma área à direita* α *pode-se determinar o valor "f" que satisfaça* $P[F(m, n) \geq f] = \alpha$ *(função inversa).*

Exemplo

(a) Dada uma distribuição F com parâmetros g.l. do numerador = 3 e g.l. do denominador igual a 5, determinar $P(F \geq 2{,}5)$

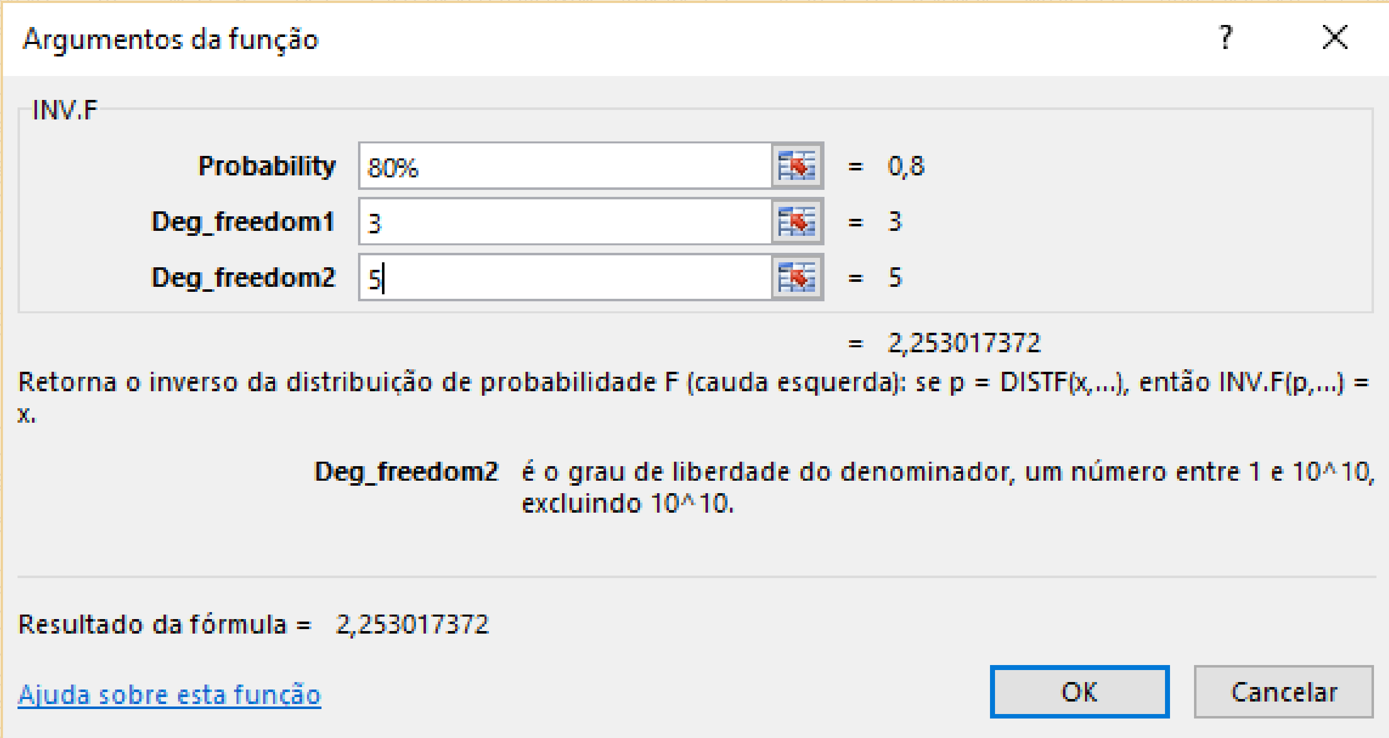
(b) O valor de "f" tal que $P(F \leq f) = 80\%$.

Item (a)



Então $P(F \geq 2,5) = 17,39\%$.

Item (b)



Então, o valor de "f" tal que,

$P(F \leq f) = 80\%$ é $f = 2{,}25.$

Exercício

Com base em um teste preliminar um grupo de alunos foi classificado de acordo com o desempenho em: Ótimo, Bom, Regular e Fraco. Para verificar se este teste era útil como previsor da média final dos alunos, amostras de cada grupo foram selecionadas. Teste se existe diferença entre as médias dos grupos ao nível de 1% de significância.

Dados

Ótimo	Bom	Regular	Fraco
9,4	7,5	7,0	6,8
9,0	6,8	7,3	7,0
8,5	7,7	7,6	7,2
8,0	8,3	7,8	6,5
	8,8	8,0	7,4
		6,8	6,5
		6,5	