

# ANOVA (ANalysis Of VAriance)

Leonardo Araújo

11 de julho de 2008

## 1 Introdução

Suponha que tenhamos  $M$  grupos, como no caso em questão, grupo referentes ao tempo de reação à palavras de baixa frequência e abstratas, grupo para as palavras de alta frequência e concretas, etc. Cada grupo é provavelmente diferente, possuindo pequenas diferenças de máximos e mínimos e é bem provável que cada grupo possua um valor médio para o tempo de reação diferente. Apesar de evidenciarmos as diferenças em cada grupo queremos saber se esta diferença no tempo médio de reação dos grupos é de fato uma evidência para mostrar que os grupos são diferentes e que talvez o efeito de frequência e concretude cause esta diferença. Note que, mesmo que não exista tal efeito da frequência e concretude no tempo de reação (hipótese nula), os grupos provavelmente terão valores médios diferentes para o tempo de reação.

A análise de variância teste a hipótese nula de que as médias de todas as populações são iguais:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_M \quad (1)$$

através da comparação entre duas estimativas de variância ( $\sigma^2$ ). Uma das estimativas é o Erro Médio Quadrático (MSE, Mean Square Error) que é baseada na variância entre amostras. O MSE é uma estimativa de  $\sigma^2$  quer ou não a hipótese nula seja verdadeira. A outra estimativa é MSB (Mean Square Between) baseada na variância da média das amostras. O MSB só será uma estimativa de  $\sigma^2$  se a hipótese nula for verdadeira. Se a hipótese nula for falsa, então a estimativa dada pela MSB será maior do que  $\sigma^2$ . A lógica utilizada na análise de variância é a seguinte: se a hipótese nula for verdadeira, então as estimativas dadas por MSE e MSB deverão ser aproximadamente as mesmas, já que ambas são estimativas de  $\sigma^2$ ; no entanto, se a hipótese nula foi falsa, então espera-se que MSB seja maior do que MSE, já que MSB estará dando uma estimativa maior do que  $\sigma^2$ .

### 1.1 Estimando $\sigma^2$ através de MSE

Para simplificar, vamos assumir que o número de amostras em cada grupo seja igual. Para estimar  $\sigma^2$  devemos então tomar o valor médio  $\mu$  e então calcular

$$MSE = \frac{\sum_{i=1}^N (s_{i,j} - \mu_j)^2}{\mu_j} \quad (2)$$

onde  $s_{i,j}$  é o valor da  $i$ -ésima amostra de um grupo  $j$ ,  $\mu_j$  a média do grupo  $j$  e  $N$  o número de amostras no grupo.

## 1.2 Estimando $\sigma^2$ através de MSB

O primeiro passo nesta estimativa consiste em estimar a variância da distribuição das médias das amostras ( $\sigma_M^2$ ). Num experimento existem  $M$  médias, uma para cada grupo. A variância dessas  $M$  médias é utilizada para estimar  $\sigma_M^2$ . Teremos então

$$d_M^2 = \frac{\sum_{j=1}^M (\mu_j - \bar{\mu})^2}{M - 1} \quad (3)$$

onde a média total é dada pela média das médias de cada grupo

$$\bar{\mu} = \frac{\sum_{j=1}^M \mu_j}{M} \quad (4)$$

$d_M^2$  calculado será uma estimativa de  $\sigma_M^2$ . Mas queríamos não uma estimativa de  $\sigma_M^2$ , mas sim uma estimativa de  $\sigma^2$ . Felizmente existe uma relação simples entre  $\sigma^2$  e  $\sigma_M^2$ . Como a distribuição das médias possui um desvio padrão de  $\sigma_M = \frac{\sigma}{\sqrt{N}}$ , então podemos obter a estimativa de  $\sigma^2$  com base na estimativa de  $\sigma_M^2$  bastando fazer

$$\sigma^2 = N \sigma_M^2 \quad (5)$$

onde  $N$  é o número de amostras em cada grupo. Para calcular MSB basta então fazer

$$MSB = N d_M^2 \quad (6)$$

onde  $N$  é o número de amostras em cada grupo e  $d_M^2$  é a variância das médias.

O teste de significância associado à análise de variância é baseado na razão entre MSB e MSE. Se a razão for muito grande, então a hipótese nula pode ser rejeitada.

## 1.3 O teste de significância no ANOVA

Se a hipótese nula for verdadeira, então ambos MSB e MSE estimam a mesma quantidade  $\sigma^2$ , resultando assim numa razão  $F = MSB/MSE$  igual a um. Se, por outro lado, a hipótese nula for falsa, então MSB estima uma quantidade  $\sigma_M^2$  maior do que  $\sigma^2$ , e assim  $F$  será maior do que um. Quanto mais distante a razão  $F$  for de um, mais provável será de que a hipótese nula seja falsa. Para realizar um teste de significância é necessário conhecer a função de densidade de probabilidade (PDF) de  $F$  dado que a hipótese nula é verdadeira. Através da PDF podemos determinar a probabilidade de se obter um  $F$  igual ou maior ao valor calculado. Se este valor for menor do que o nível de significância, então a hipótese nula é rejeitada. Os estudos sobre a PDF de  $F$  foram feitos pelo estatístico R. A. Fisher e é chamada distribuição de  $F$  em sua homenagem.

Um teste de significância é realizado para determinar se um valor observado de uma determinada estatística difere suficientemente do valor hipotético de um parâmetro para levar à inferência de que o valor hipotético do parâmetro não é verdadeiro. O valor hipotético do parâmetro é chamado 'hipótese nula'. O teste de significância consiste em calcular a probabilidade de obter-se uma estatística tão diferente ou mais diferente da hipótese nula (tomando a hipótese nula como correta) do que a estatística obtida nas amostras. Se esta probabilidade for suficientemente baixa, então a diferença entre o parâmetro e a estatística é dita 'estatisticamente significativa'.

No teste de hipótese, o nível de significância é o critério adotado para rejeitar a hipótese nula. Para se realizar o teste primeiro calcula-se a diferença entre os resultados do experimento e da hipótese nula. Depois, assumindo que a hipótese nula é verdadeira, a probabilidade de uma diferença igual ou maior é computada. Por último, a probabilidade é comparada com o nível de significância. Se a probabilidade for menor ou igual ao nível de significância, então a hipótese nula é rejeitada e o resultado é dito ser estatisticamente significativo. Tradicionalmente, tem-se utilizado valores de 0.05 (chamado de nível de 5%) ou valores de 0.01 (nível de 1%), embora a escolha do nível seja muito subjetiva. Quanto mais baixo o nível de significância, mais os dados devem divergir da hipótese nula para serem significantes. Então o nível 0.01 é mais conservador do que o nível 0.05.

A distribuição  $F$  é a distribuição da razão de duas estimativas de variância. Ela é utilizada para computar os valores de probabilidade na análise de variância. A distribuição  $F$  possui dois parâmetros: graus de liberdade do numerador ( $dfn$ ) e grau de liberdade do denominador ( $dfd$ ).

$$dfn = M - 1 \tag{7}$$

$$dfd = N_T - M \tag{8}$$

Onde  $M$  é o número de grupos e  $N_T$  o número total de amostras no experimento. A forma da distribuição  $F$  depende de  $dfn$  e  $dfd$ . Na figura abaixo são ilustradas as distribuições de  $F$  quando  $dfn = 4$  e  $dfd = 12$  (primeira figura) e quando  $dfn = 10$  e  $dfd = 100$  (segunda figura).

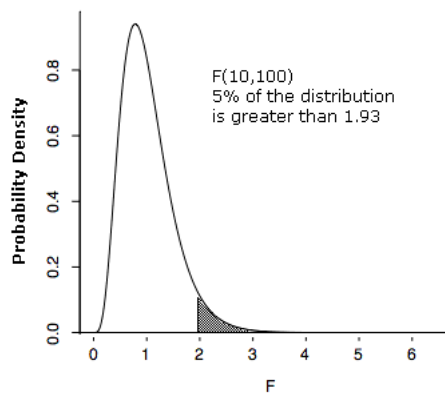
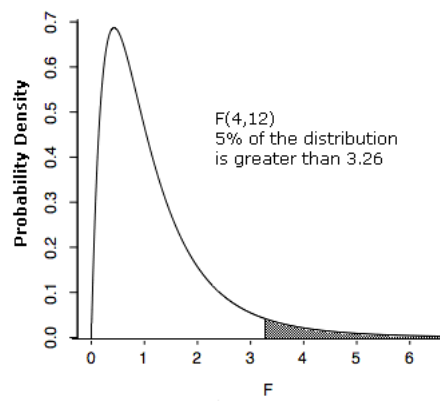


Figura 1: Distribuição F