Ponto 12 (Teórica) - Análise de Variância

Ben Dêivide

14 de Novembro de 2016

A Análise de variância (ANAVA) introduzida por R. A. Fisher na década de 1930, embora o termo "análise de variância" tenha sido chamado depois por John Wilder Tukey, é um método sistemático baseado em distribuições amostrais. Embora essa técnica tenha sido desenvolvida inicialmente para desenhos de análises de experimentos em que avaliava se o efeito dos tratamentos não era por mero acaso, esses desenhos não passam de casos especiais de um modelo linear geral, hoje existe diversos tipos de análises de variâncias com finalidades diversas. Vejamos alguns propósitos da ANAVA:

- Comparar médias de várias populações;
- Testar os parâmetros de um determinado modelo linear;
- Estimar os componentes da variância,

dentre outros. Iremos nos restringir aos dois primeiros para demonstrar o desenvolvimento da análise de variância.

Nessa dissertação, trataremos do procedimento matemático de como Fisher desenvolveu a ANAVA, sem necessariamente darmos detalhes sobre princípios de experimentação ou outros enfoques, uma vez que estes foram meios para atender as pressuposições necessárias para utilizar a metodologia desse método.

Formalmente, definimos

Definição 1 (Análise Variância - ANAVA)**.** *Uma família de procedimentos que usam o teste F para verificar o ajuste global de um modelo linear a um conjunto de dados.*

Esse ajuste global é verificado particionando a variação total presente nos dados em partes que refletem a variabilidade explicada pelo modelo linear e a variabilidade devido ao acaso.

A base da análise de variância está relacionada com o teste F que depende da distribuição F, cujo nome da distribuição foi em homenagem ao próprio Fisher. Uma variável aleatória contínua W tem distribuição F se sua função densidade é expressa por:

$$f_W(w) = \frac{\Gamma[(m+q)/2]}{\Gamma(m/2)\Gamma(q)} \left(\frac{m}{q}\right)^{m/2} \frac{w^{(m-2)/2}}{[1+(m/q)w]^{(m+q)/2}}, \quad w > 0, m > 0, n > 0. \tag{1}$$

Em notação, temos $W \sim F_{m,q}$, isto é, W tem distribuição F com m e q graus de liberdade.

Pode-se mostrar que se as variáveis aleatórias $U \sim \chi^2_m$ e $V \sim \chi^2_q$ independentes, então

$$W = \frac{U/m}{V/q} \sim F_{m,q}. \tag{2}$$

Contudo, a origem de uma variável aleatória com distribuição de quiquadrado, advém de uma amostra aleatória Y_1, Y_2, ..., Y_n, tal que $Y \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Sabemos que $Z = (Y - \mu)/\sigma \sim N(0, 1)$. Assim, apresentaremos o seguinte teorema,

Teorema 1. *Seja as v.a. iid com* $Y_i \sim N(\mu; \sigma^2)$*, então a v.a.* $U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2$ *tem distribuição qui-quadrado com n graus de liberdade.*

Demonstração. Seja $Z_i = \frac{Y_i - \mu}{\sigma}$. Logo $Z_i \sim N(0, 1)$. Vamos utilizar a técnica da função geradora de momentos para provar a distribuição de U. Assim,

$$m_U(t) = E\left[e^{tU}\right] = E\left[e^{t\sum_{i=1}^{n} Z_i^2}\right] = E\left[\prod_{i=1}^{n} e^{tZ_i^2}\right] = \prod_{i=1}^{n} E\left[e^{tZ_i^2}\right].$$

Mas,

$$\begin{aligned}
E\left[e^{tZ^2}\right] &= \int_{-\infty}^{\infty} e^{tZ^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\left(t - \frac{1}{2}\right)Z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz = \frac{\sqrt{1-2t}}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)Z^2} dz = \frac{1}{\sqrt{1-2t}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \frac{1}{(1-2t)}}} e^{-\frac{1}{2}\left(\frac{Z}{\sqrt{\frac{1}{1-2t}}}\right)^2} dz}_{=1} \\
&= \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}} = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{1}{2}},
\end{aligned}$$

em que a integral resulta em 1, pois é a densidade de uma Normal padrão com média 0 e variância $\frac{1}{1-2t}$. Portanto,

$$m_U(t) = \prod_{i=1}^{n} \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{1}{2}} = \left[\left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{1}{2}}\right]^n = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^{\frac{n}{2}}, \; t < \frac{1}{2},$$

que é f.g.m. de uma distribuição qui-quadrado com n graus de liberdade. □

O próximo teorema também é necessário.

Teorema 2. *Seja* $Z_1, \ldots, Z_n$ *uma a.a. e* $Z_i \sim N(0, 1)$*. Considere que* $\bar{Z}$ *e* $\sum_{i=1}^{n} (Z_i - \bar{Z})^2$ *sejam independentes. Seja* $Q = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$*. Então* $Q \sim \chi^2(n-1)$*.*

Demonstração. Note que

$$\begin{aligned}
\sum Z_i^2 &= \sum \left(Z_i - \bar{Z} + \bar{Z}\right)^2 = \sum \left[\left(Z_i - \bar{Z}\right)^2 + 2\left(Z_i - \bar{Z}\right)\bar{Z} + \bar{Z}^2\right] \\
&= \sum \left(Z_i - \bar{Z}\right)^2 + 2\bar{Z} \sum \left(Z_i - \bar{Z}\right) + n\bar{Z}^2 = \sum \left(Z_i - \bar{Z}\right)^2 + n\bar{Z}^2.
\end{aligned}$$

Como $\sum \left(Z_i - \bar{Z}\right)^2$ e $\bar{Z}$ são independentes , temos que

$$m_{\sum Z_i^2}(t) = m_{\sum \left(Z_i - \bar{Z}\right)^2}(t)\, m_{n\bar{Z}^2}(t).$$

Daí,

$$m_{\sum \left(Z_i - \bar{Z}\right)^2}(t) = \frac{m_{\sum Z_i^2}(t)}{m_{n\bar{Z}^2}(t)} = \frac{\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{n}{2}}}{\left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{1}{2}}} = \left(\frac{\frac{1}{2}}{\frac{1}{2}-t}\right)^{\frac{(n-1)}{2}}, \; t < \frac{1}{2},$$

que é a f.g.m. de uma $\chi^2(n-1)$. Observe ainda que se $Z_i \sim N(0,1)$, $Z_1 + \ldots + Z_n \sim N(0,n)$, então $\bar{Z}_n \sim N\left(0, \frac{1}{n}\right)$. Logo

$$\sqrt{n}\bar{Z}_n \sim N(0,1) \Rightarrow \left(\sqrt{n}\bar{Z}_n\right)^2 = n\bar{Z}_n^2 \sim \chi^2(1).$$

□

Como consequência dos teoremas vistos:

(i) Como $\sum \left(Z_i - \bar{Z}\right)^2$ e $\bar{Z}$ são independentes, implica que $\bar{Y}$ e $\sum (Y_i - \bar{Y})^2$ também o são.

(ii) Como $\sum \left(Z_i - \bar{Z}\right)^2 \sim \chi^2(n-1)$, implica que $\sum \frac{\left(Y_i - \bar{Y}\right)^2}{\sigma^2}$ também tem distribuição $\chi^2(n-1)$.

(iii) $\sum \left(Z_i - \bar{Z}\right)^2 = \sum \left(\frac{Y_i - \mu}{\sigma} - \frac{\bar{Y} - \mu}{\sigma}\right)^2 = \sum \left(\frac{Y_i - \bar{Y}}{\sigma}\right)^2$.

Com base nessas informações, as pressuposições para a utilização da distribuição F, são:

- A amostra aleatória $Y_1, Y_2, \ldots, Y_n$ seja normalmente distribuída com média μ_i para $i = 1, 2, \ldots, n$, e variância constante σ^2. Essas médias podem ser consideradas iguais;
- O fato da amostra ser aleatória, supõe-se independência entre as variáveis aleatórias;
- A razão entre as variáveis aleatórias com distribuição de qui-quadrado, com seus respectivos graus de liberdade, têm que ser independentes.

Dessa forma, apresentamos o primeiro tipo de finalidade da ANAVA que é a comparação de médias em um delineamento experimental. Escolheremos um modelo mais simples para apresentar a ideia que servirá de base para qualquer outro tipo de delineamento ou esquema fatorial.

Considere o seguinte modelo:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{3}$$

em que:

- Y_{ij} representa a observação j no i-ésimo tratamento, para $i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, n$;;
- μ representa a média geral do experimento, sob a restrição $\sum_i^t \tau_i = 0$;

- τ_i é o efeito do tratamento i sobre a observação Y_{ij}, para $i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, n$;
- considere μ e τ_i parâmetros fixos;
- ϵ_{ij} é o erro experimental para a observação Y_{ij}, para $i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, n$.

Como consequência, dizemos que Y_{ij} são independentes, tal que $Y_{ij} \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$, em que $\mu_i = \mu + \tau_i$. Portanto, o objetivo da ANAVA nessa situação é verificar se as médias μ_i são estatisticamente iguais ou não, que é equivalente a testar o efeito dos $\tau_i's$ sob a hipótese:

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_t = 0, \tag{4}$$

contra a hipótese alternativa

$$H_1 : \text{Pelo menos um } \tau_i \neq 0. \tag{5}$$

Assim, utilizando o método dos mínimos quadrados, sob a restrição $\sum_{i=1}^{t} \tau_i = 0$, chegamos aos estimadores de μ e τ_i, de tal modo que a variabilidade dos Y_{ij} serão apresentados por meio de somas de quadrados. A variabilidade total é expressa por:

$$SQT = \sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2, \tag{6}$$

em que SQT pode ser particionado em:

$$SQT = \sum_{i=1}^{t} n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2 \tag{7}$$

$$= SQTrat + SQErro, \tag{8}$$

sendo:

- $\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$ é a média do i-ésimo tratamento;
- $\bar{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n} Y_{ij}}{tn}$ é a média geral do experimento.

Dizemos que SQT representa a soma de quadrado total e refere a soma global do quadrado dos desvios entre as observações e a média geral do experimento. O SQTrat se refere a soma de quadrado entre tratamentos e refere a variabilidade da média dos tratamentos e a média geral do experimento. E por fim, o SQErro, é a soma de quadrados dentro dos tratamentos, que reflete a variabilidade das observações de cada tratamento à sua média.

Podemos mostrar pelos Teoremas 1 e 1 podemos mostrar que

$$\frac{SQTrat}{\sigma^2} \sim \chi^2_{t-1,\lambda}, \tag{9}$$

isto é, $\frac{SQTrat}{\sigma^2}$ tem distribuição qui-quadrado não central, com $t-1$ graus de liberdade, e parâmetro de não centralidade dado por $\lambda = \frac{1}{2\sigma^2}\sum_{i=1}^{t} n(\tau_i - \bar{\tau})^2$, sendo $\bar{\tau} = \sum_{i=1}^{t} \tau_i/t$, e

$$\frac{SQErro}{\sigma^2} \sim \chi^2_{tn-t}, \tag{10}$$

isto é, $\frac{SQErro}{\sigma^2}$ tem distribuição quiquadrado central com $tn-t$ graus de liberdade.

Pode-se mostrar que as esperanças de SQTrat e SQRes são, respectivamente,

$$E[SQTrat] = (t-1)\sigma^2 + n\sum_{i=1}^{t}\tau_i^2, \tag{11}$$

e

$$E[SQRes] = (nt-t)\sigma^2. \tag{12}$$

Assim, uma variável aleatória com base nessas duas somas de quadrado, tem distribuição F não central da seguinte forma:

$$W = \frac{\frac{SQtrat}{\sigma^2(t-1)}}{\frac{SQres}{\sigma^2(tn-t)}} = \frac{\frac{SQtrat}{(t-1)}}{\frac{SQres}{(tn-t)}} = \frac{QMTrat}{QMErro} \sim F_{t-1,tn-t,\lambda}, \tag{13}$$

sendo $QMTrat = \frac{SQtrat}{(t-1)}$ representando o quadrado médio dos tratamentos, $QMErro = \frac{SQErro}{(tn-t)}$ representando o quadrado médio do erro, e $F_{t-1,tn-t,\lambda}$ a distribuição F não central com $t-1$ e $tn-t$ graus de liberdade e parâmetro de não centralidade dado por $\lambda = \frac{1}{2\sigma^2}\sum_{i=1}^{t} n(\tau_i - \bar{\tau})^2$, sendo $\bar{\tau} = \sum_{i=1}^{t}\tau_i/t$.

Agora, a genialidade do Fisher para a análise de variância foi baseada nos resultados a seguir. Observe as esperanças dos quadrados médios são:

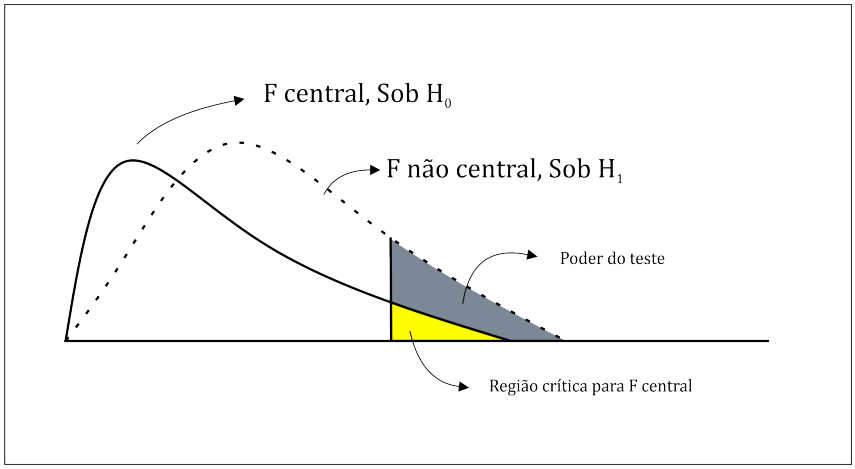
$$E[QMTrat] = \sigma^2 + \frac{n\sum_{i=1}^{t}\tau_i^2}{t-1} \tag{14}$$

e

$$E[QMErro] = \sigma^2. \tag{15}$$

Estas estimam σ^2, a menos da primeira que apresenta $\frac{n\sum_{i=1}^{t}\tau_i^2}{t-1}$.

Entretanto, sob H_0 expresso em (4), os estimadores $QMTrat$ e $QMErro$ de σ^2 são não-viesados, e portanto, se espera que $W \approx 1$. Dessa forma, se ao menos um $\tau_i \neq 0$, se espera que $W > 1$ e aí teremos evidências para rejeitar a hipótese H_0. Outro fato interessante é que sob H_0 a distribuição de W é F central, uma vez que o parâmetro de não centralidade é $\lambda = 0$. Pela Figura abaixo, percebemos que mais W se afasta de 1, maior é o poder do teste, e consequentemente, mais evidências teremos para rejeitar a hipótese H_0



Abaixo, segue o quadro da análise de variância:

Tabela 1: Quadro da ANAVA para o DIC.

FV	GL	SQ	QM	teste F	Valor-p
Tratamentos	$t-1$	SQTrat	QMTrat	W	$P(W > w)$
Erro	$tn-t$	SQErro	QMErro	-	-
Total	$tn-1$	SQTot	-	-	-