

Testes de hipótese e p-valores

Muitas vezes é difícil atribuir inequivocamente certos resultados a fatores causais. Pode ser que algo tenha ocorrido, mas seu efeito não é pronunciado o suficiente para ser distinguido de possíveis erros de medida ou fatores ambientais ou de outra natureza não controláveis.

Testes de hipótese são a ferramenta estatística para abordar esse tipo de problema. Para ilustrar essa abordagem, considere novamente o caso de uma variável do tipo Bernoulli

$$\phi(x) = p^x (1-p)^{1-x} \quad x \in \{0, 1\}$$

cujo parâmetro $p=\theta$ associando a uma sequência de valores $\{x_1, x_2, \dots, x_n\} = \bar{x}_n$ não é conhecido. Numa linguagem estatística definimos a hipótese nula

$$H_0: \theta < 1/2$$

e a hipótese alternativa

$$H_1: \theta > 1/2.$$

Um teste de hipótese vai ajudar a decidir qual hipótese é mais consistente com os dados \bar{x}_n . Para isso, vamos definir uma função G que depende de \bar{x}_n , de modo que se $G(\bar{x}_n)$ exceder um limiar c , vamos considerar H_1 e rejeitar H_0 (costuma-se dizer que a hipótese nula foi rejeitada). Por outro lado, caso $G(\bar{x}_n)$ não exceda o limiar c , então não podemos rejeitar H_0 . Resumindo:

$G(\bar{x}_n) < c \Rightarrow H_0$ (hipótese nula não pode ser rejeitada)

$G(\bar{x}_n) > c \Rightarrow H_1$ (podemos rejeitar a hipótese nula)

(1)

Independentemente do teste G , teremos dois tipos de erros: falso-negativo (erro tipo-II) & falso-positivo (erro tipoI)

A tabela abaixo ilustra os dois casos:

	H_0 é verdadeira	H_0 é falsa
H_0 é verdadeira	α → significância $1 - \alpha$ → confiança	
Teste diz aceitar H_0	Aceitago verdadeira Sensibilidade $\text{True positive rate}$ TPR	Falso-negativo Aceitago falsa Erro tipo II <u>falso-negativo rate FNR</u>
Teste diz rejeitar H_0	Falso-positivo Rejeigo falsa Erro tipo I <u>falso-positive rate FPR</u>	Rejeigo verdadeira Poder do teste Especificidade <u>True negative rate TNR</u>

Assim, quando o teste recomenda a rejeição de H_0 , mas H_0 é verdadeira, temos um falso-positivo.

Por outro lado, quando o teste recomenda aceitar H_0 , mas H_0 é falsa, temos um falso-negativo.

No caso do teste anterior a probabilidade associada aos falso-positivos é

$$P_{FA} = P_{FPR} = P(G(\bar{x}_n) > c \mid \theta \leq \frac{1}{2}) \\ = P(G(\bar{x}_n) > c \mid H_0)$$

Similarmente, podemos definir a probabilidade associada aos falso-negativos como

$$P_{FN} = P_{FNR} = P(G(\bar{x}_n) < c \mid H_1)$$

Ainda podemos definir a chamada probabilidade de detecção (TPR) como

$$P_D = P_{TPR} = 1 - P_{FA} = P(G(\bar{x}_n) > c \mid H_1) \quad (2)$$

Voltando ao exemplo da variável de Bernoulli, podemos pensar que no lugar de estimar θ usando, por exemplo o estimador de verossimilhança

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

estamos fazendo uma pergunta mais branda sobre o valor de θ , isto é,

$$H_0: \theta \leq 1/2$$

$$H_1: \theta > 1/2$$

Suponha que tenhamos 5 observações e que nossa estatística do teste seja

$$G(\bar{x}_5) = \sum_{i=1}^5 x_i$$

Ou seja, o número vezes que $x_i = 1$. Devemos definir ainda um limiar c a partir do qual rejeitamos H_0 em favor de H_1 . Suponha $c = 5$. Desse modo, rejeitamos H_0 apenas se os cinco valores de x_i forem iguais a 1. Assumindo que os valores de x_i são independentes, sabemos que a probabilidade de ocorrerem cinco $x_i = 1$ é θ^5 . Assim, a probabilidade de rejeitar H_0 é θ^5 e, portanto, a probabilidade de falso-positivo é

$$P_{FA} = P(G(\bar{x}_5) = 5 | H_0)$$

$$= P(\theta^5 | H_0) = \theta^5 \quad \theta < 1/2$$

→ Aceitar H_1
quando H_0 é
verdadeira

Observe que P_{FA} depende de θ . Para sermos conservadores, vamos considerar o máximo de θ^5 , o chamado tamanho do teste (size),

$$\alpha = \sup_{\theta \in \Theta} \theta^5$$

(seria a pior probabilidade de Falso-positivo ou rejeição falsa ou erro tipo I)
pior FPR

sendo $\Theta = \{\theta < 1/2\}$, temos

$$\alpha = \sup_{\theta < 1/2} \theta^5 = \left(\frac{1}{2}\right)^5 \approx 0,03$$

Por outro lado, a probabilidade de detecção é
(sensibilidade)
 $P_D = P(\theta^5 | H_1) = \theta^5 \quad \theta > 1/2$
Aceritação verd.
TPR

que também é uma função de θ . O problema desse teste é que P_D é muito baixa para a maior parte dos valores de θ . Por exemplo, valores de $P_D \approx 90\%$ só acontecem para $\theta > 0,98$. Idealmente, devemos provar por um teste que seja zero para o domínio de H_0 ($\theta < 1/2$) e 1 caso contrário. Entretanto, esse teste não permite isso, mesmo que o número de observações aumente muito.

Teste do voto da maioria

Nesse caso, vamos propor rejeitar H_0 se a maioria das observações forem do tipo $x_i = 1$. Nesse caso, a probabilidade de rejeitar H_0 (power function) fica

$$\beta(\theta) = \sum_{k=3}^5 \binom{5}{k} \theta^k (1-\theta)^{5-k}$$

(4)

Cada termo na soma representa a probabilidade de obter $K \geq 3$ valores de $x_i = 1$. Podemos calcular o tamanho do teste

$$\alpha = \sup_{\theta < \frac{1}{2}} \beta(\theta) = \frac{1}{2} \quad (\text{pior FPR})$$

Como no caso anterior, a probabilidade de detecção P_D

$$P_D = P(\beta(\theta) | H_1) \quad (\text{TPR})$$

é muito baixa. Para $P_D \approx 90\%$, $\theta > 0,75$. Podemos verificar o que ocorre com o aumento de n . Suponha $n=100$ e que rejeitamos H_0 se o número de $x_i = 1$ superar 60, ou seja,

$$\beta(\theta) = \sum_{K=60}^{100} \binom{100}{K} \theta^K (1-\theta)^{100-K}$$

Nesse caso, encontramos $\alpha = 0,018$ e

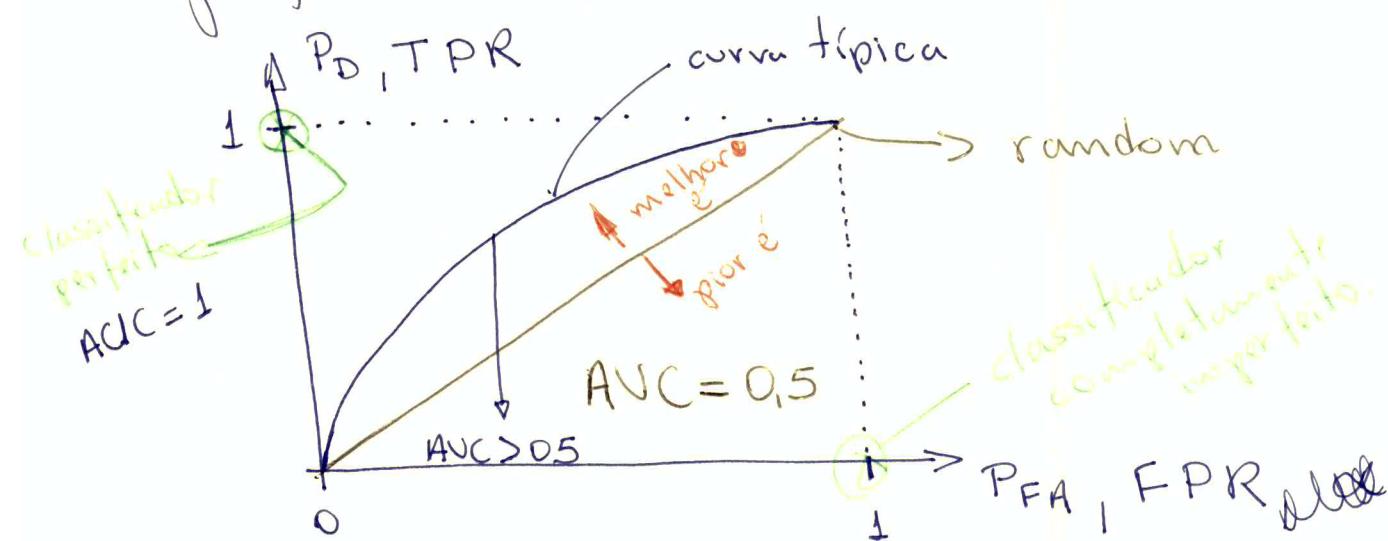
$$P_D = P(\beta(\theta) | H_1) \Big|_{\theta=0,70} \approx 0,98$$

Ou seja, se rejeitarmos H_0 ao observar 60 $x_i = 1$, estaremos errando $\approx 1,8\%$ das vezes. Ao passo que se $P > 0,70$, o teste rejeita H_0 em $\approx 98\%$ das vezes.

Exemplo notebook

Característica de Operação do Receptor

A chamada curva ROC (Receiver Operating Characteristic) é um gráfico para quantificar a performance de um classificador bimário. A curva ROC é um gráfico P_D versus P_{FA} (TPR vs. FPR)



Vejamos um exemplo, no contexto de processamento de sinal:

$$H_0: X = \epsilon$$

$$H_1: X = \mu + \epsilon$$

sendo $\epsilon \sim N(0, \sigma^2)$. Assim, dado um conjunto de valores \bar{X} devemos decidir se a média de \bar{X} difere de zero ou não. Devemos escolher um limiar c , tal que,

$X > c \Rightarrow$ rejeitar H_0 em favor de H_1 ,

$X < c \Rightarrow$ não rejeitar H_0 .

Discutir as figuras no notebook

Naturalmente ao aumentar os valores de C , P_{FA} e P_D variam, de modo a produzirem a curva ROC. Nessa curva é comum incluir uma reta ($P_{FA} = P_D$) que representa o caso de uma decisão aleatória. Também é comum resumir a curva ROC pela sua área (AVC), a qual varia de zero a um. Sendo que quanto mais próximo de um, melhor o classificador / teste. Os valores ideais para P_D e P_{FA} dependem do tipo de aplicação. Por exemplo, suponha

H_0 : tem uma doença fatal

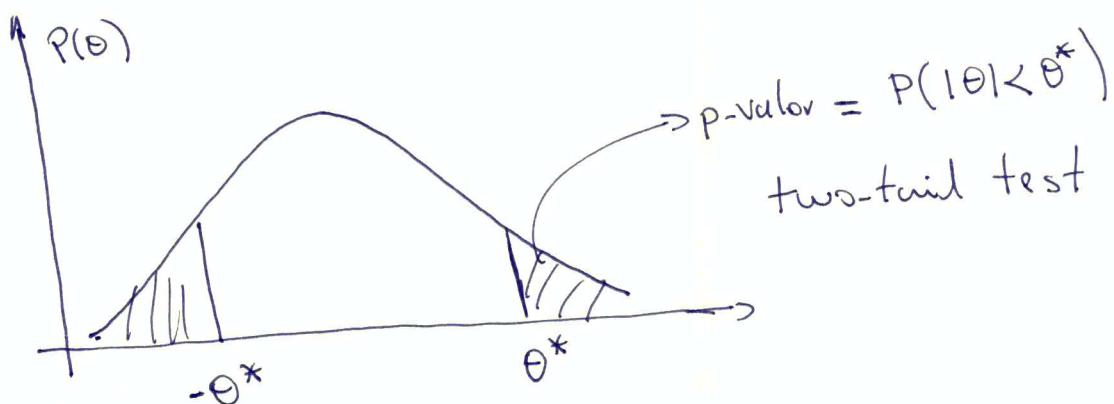
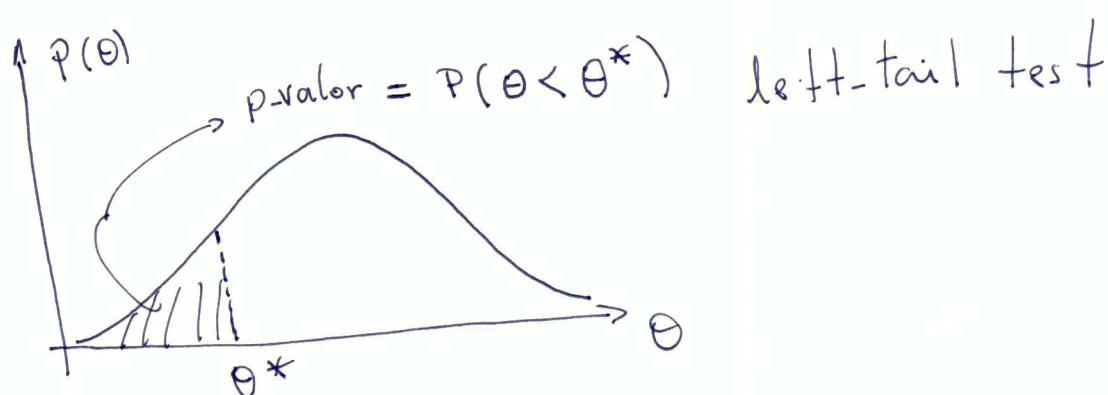
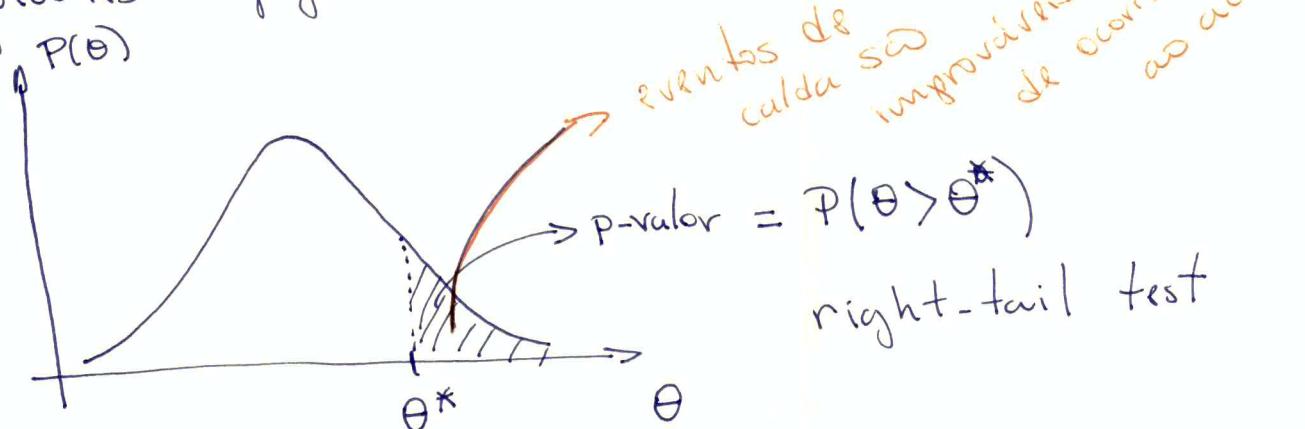
H_1 : não tem

Nesse caso, se o teste for barato e fácil de aplicar, podemos aceitar P_{FA} (FPR) alta para garantir uma P_D (TPR) também alta. Por outro lado, se os custos envolvidos em tratar uma pessoa saudável for alta, um alta P_{FA} (FPR) não costuma ser desejável e pode ser preferível um P_D (TPR) mais baixa para que P_{FA} seja também mais baixa. Existe um trade-off natural.

P - Valores

Como deve estar claro, existem muitas variáveis associadas à performance de um teste. O chamado p-valores são uma tentativa de consolidar as diferentes medidas de performance.

O p-valor é a probabilidade (supondo H_0 verdadeira) de que a estatística do teste seja ~~maior ou igual~~^{mais extrema} que de fato é observada. Assim, quanto menor for o p-valor ~~ou considerada incompatível com o dado~~, H_0 deve ser rejeitada. Se θ for a estatística do teste e θ^* for o valor observado, então temos as seguintes opções:



Podemos ^{pensar} que o p-valor é a probabilidade de rejeitar H_0 ao acaso. Um valor baixo indica que isso é improvável, enquanto valores altos indicam uma maior probabilidade disso ocorrer.

Podemos ainda escrever que

$$P(\text{rejeitar } H_0 \mid H_0) = P(p \leq \alpha \mid H_0) = \alpha = FPR$$

sendo α o chamado nível de significância.

Usualmente $\alpha \in (0,05; 0,01, 0,005, 0,001)$. Costuma-se definir também o nível de confiança como $1-\alpha$, que também representa o TPR.

Por exemplo, no caso do teste do voto da maioria,

$$\begin{aligned} P(\text{rejeitar } H_0 \mid H_0) &= \sup_{\theta \in \Theta} \sum_{K=3}^5 \binom{5}{K} \theta^K (1-\theta)^{5-K} \\ &= 0,5 \end{aligned}$$

Já no teste de todos iguais a m

$$P(\text{rejeitar } H_0 \mid H_0) = \sup_{\theta \in \Theta} \theta^5 = 0,03$$

Veja que olhando apenas para o p-valor somos inclinados a pensar que o teste de todos iguais é melhor. Entretanto, que o TPR desse teste é baixa.

Estatísticas de testes

Como vimos, é difícil encontrar uma boa estatística de teste sem um procedimento sistemático. O lema de Neyman-Pearson ou teste de Neyman-Pearson mostra como

obter uma estatística que maximiza o poder do teste (TNR) fixando a taxa de falso positivo ou nível de significância (α). A estatística do teste é definida como:

$$L(\bar{x}) = \frac{f_{X|H_1}(\bar{x})}{f_{X|H_0}(\bar{x})} \begin{cases} > \gamma \\ H_1 \\ H_0 \end{cases}$$

sendo γ definido a partir de

$$\int_{x: L(x) > \gamma} f_{X|H_0}(\bar{x}) d\bar{x} = \alpha$$

Probabilidade H_0
ser verdadeira quando
que $L > \gamma$, ou seja,
 $P(\text{rejeitar } H_0 | H_0)$.

ou ainda

$$\alpha = P(L(\bar{x}) \leq \gamma | H_0)$$

Aqui $f_{X|H_0}(\bar{x})$ é a verossimilhança de H_0
(probabilidade de H_0 ser verdadeira
dado que observamos \bar{x})

$f_{X|H_1}(\bar{x})$ é a verossimilhança de H_1
(probabilidade de H_1 ser verdadeira
dado que observamos \bar{x})

Note se $L(\bar{x}) < 1$, então $f_{X|H_0} > f_{X|H_1}$, ou seja,
 H_0 é mais provável que H_1 . Por outro lado, se
 $L(\bar{x}) > 1$, então $f_{X|H_1} > f_{X|H_0}$, ou seja, H_1 é
mais provável. O limiar γ serve para
impor o nível de significância α .

Exemplo: suponha que desejamos distinguir entre um ruído e um sinal, isto é,

$$H_0: X \sim N(0, 1)$$

$$H_1: X \sim N(1, 1)$$

Vemos, portanto, que o efeito do sinal é aumentar a média de X . Para dados gaussianos, a verossimilhança é

$$\begin{aligned} L(\mu, \sigma | \bar{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \end{aligned}$$

Sendo assim, a estatística do teste fica

$$L(\bar{x}) = \frac{\exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - 1)^2\right\}}{\exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i)^2\right\}}$$

No caso de uma única observação ($n=1$)

$$L(\bar{x}) = \frac{e^{-\frac{1}{2}(x^2 - 2x + 1)}}{e^{-\frac{1}{2}x^2}}$$

$$L(x) = e^{x-1/2}$$

Assim,

$$e^{x-1/2} \begin{cases} H_1 \\ H_0 \end{cases} \gamma$$

$$x - \frac{1}{2} \geq \frac{\ln \gamma}{H_0}$$

$$x \geq \frac{H_1}{H_0} \frac{1}{2} + \ln \gamma$$

Para encontrar γ dado o valor de α , temos

$$\int_{\frac{1}{2} + \ln \gamma}^{\infty} f_{X|H_0}(x) dx = \alpha$$

$$\int_{\frac{1}{2} + \ln \gamma}^{\infty} \frac{1}{(2\pi)^{-1/2}} e^{-x^2/2} dx = \alpha$$

Usando a definição de função erro

$$Erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

e da função erro complementar

$$Erfc(x) = 1 - Erf(x)$$

$$= \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt$$

Temos

$$(2\pi)^{-1/2} \left(\frac{\pi}{2}\right)^{1/2} Erfc\left(\frac{\frac{1}{2} + \ln \gamma}{\sqrt{2}}\right) = \alpha$$

$$\frac{1}{2} Erfc\left(\frac{\frac{1}{2} + \ln \gamma}{\sqrt{2}}\right) = \alpha$$

$$\left(\frac{1}{2} + \ln \gamma\right)/\sqrt{2} = Erf^{-1}(2\alpha)$$

$$\frac{1}{2} + \ln \gamma = \sqrt{2} \operatorname{Erfc}^{-1}(2\alpha)$$

$$\ln \gamma = \sqrt{2} \operatorname{Erfc}^{-1}(2\alpha) - \frac{1}{2}$$

$$\gamma = \exp \left\{ \sqrt{2} \operatorname{Erfc}^{-1}(2\alpha) - \frac{1}{2} \right\}$$

Para $\alpha = 0,01$, temos $\gamma = 6,21116$, e o teste fica

$$X \begin{cases} > \\ \leq \end{cases}_{H_0} \frac{1}{2} + \ln \gamma$$

$$X \begin{cases} > \\ \leq \end{cases}_{H_0} 2,3235$$

Assim, para termos um nível de significância de 0.01 (contrança de 99%), o valor de X para ser considerado um sinal, deve superar 2,33.

* Exemplo notebook

Teste da razão de Verossimilhança generalizada

A estatística desse teste escreve-se

$$\Lambda(\bar{x}) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

sendo que $\hat{\theta}_0$ maximiza $L(\theta)$ para $\theta \in \Theta_0$ e $\hat{\theta}$ é o estimador de verossimilhança para um parâmetro θ do modelo. Note que essa estatística é razão entre a máxima verossimilhança $L(\hat{\theta})$ e o máximo de $L(\theta)$ restrito ao domínio Θ_0 .

Sendo assim, $\Lambda(\bar{x}) \leq 1$. Desse modo, quando menor for $\Lambda(\bar{x})$, maior será a ~~estimador~~ verossimilhança em todo o domínio (Θ) e, portanto, mais seguro é rejeitar H_0 .

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \notin \Theta_0 \end{cases}$$

A dificuldade está em encontrar a distribuição de Λ . Porém, o teorema de Wilks garante que para n grande a distribuição de $-2\log\Lambda$ se aproxima da distribuição chi-quadrado com $r - r_0$ graus de liberdade, sendo r o número de parâmetros em Θ e r_0 o número de parâmetros em Θ_0 .

A distribuição chi-quadrado é

$$p(x) = \frac{1}{2^{K/2} \Gamma(K/2)} x^{K/2 - 1} e^{-x^2/2} \quad (x > 0)$$

sendo K o número de graus de liberdade. Ela surge ao somar o quadrado de variáveis normais, isto é

$$x = \sum_{i=1}^K z_i^2$$

$$z_i \sim \text{Normal}$$

então $x \sim \text{chi-quadrado com } K \text{ graus de liberdade}$

Assim, a partir da estatística do teste, temos que

$$-2 \log \Lambda > \chi^2_{r-r_0}(\alpha) \Rightarrow \text{rejeitar } H_0$$

$$-2 \log \Lambda < \chi^2_{r-r_0}(\alpha) \Rightarrow \text{não rejeitar } H_0$$

ou ainda

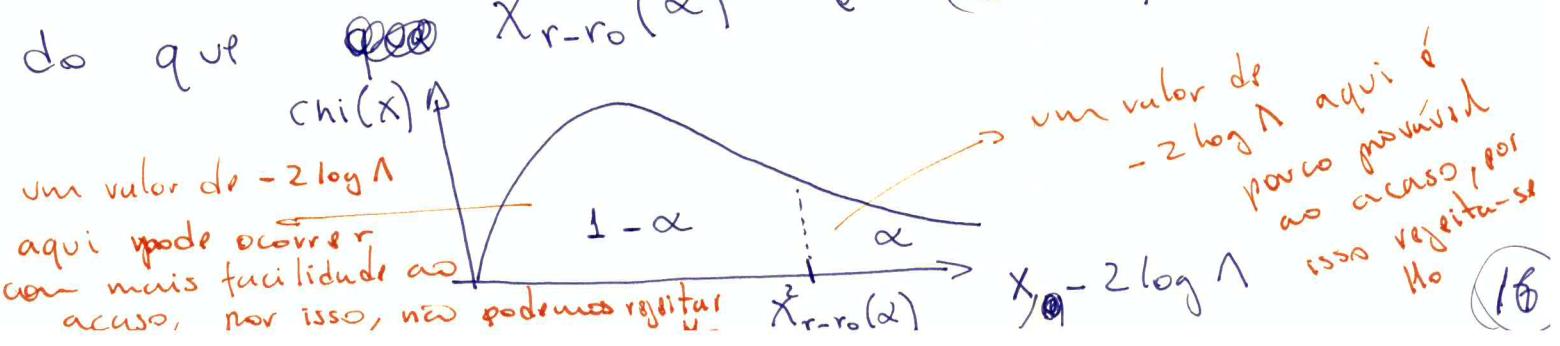
$$-2 \log \Lambda \stackrel{H_1}{\geq} \chi^2_{r-r_0}(\alpha)$$

sendo $\chi^2_{r-r_0}(\alpha)$ o quantil $(1-\alpha)$ da distribuição chi-quadrado. O quantil de uma distribuição X_θ é definido como

$$\int_{-\infty}^{x_\theta} P(x) dx = q$$

ou seja, é a probabilidade de encontrar um número em X_θ menor ou igual a x_θ . No caso, $\chi^2_{r-r_0}(\alpha)$ é a probabilidade de encontrar um valor de chi-quadrado,

Ou seja, a probabilidade de $X < X_\theta$ é igual a q . No caso de $\chi^2_{r-r_0}(\alpha)$, ~~que representa~~ a probabilidade de encontrar um valor de X (distribuído como uma chi-quadrado) menor do que $\chi^2_{r-r_0}(\alpha)$ é $(1-\alpha)$.



Exemplo: três binomiais

$$\text{binom}(K; n, p) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k}$$

Podemos imaginar esse exemplo como o resultado do lançamento de três moedas (n_1, n_2 e n_3) vezes, produzindo (K_1, K_2, K_3) caras. Queremos testar se todas tem a mesma probabilidade de produzir caras, isto é,

$$H_0: p = p_1 = p_2 = p_3$$

$$H_1: p_1 \neq p_2 \text{ ou } p_1 \neq p_3 \text{ ou } p_2 \neq p_3 \text{ ou } p_1 \neq p_2 \neq p_3$$

Vamos definir a estrutura do teste como:

$$\Lambda(\bar{x}) = \frac{L(\theta_0)}{L(\hat{\theta})}$$

Assumindo H_0 verdadeira, temos que a verossimilhança

$$L(p_1, p_2, p_3) = \text{binom}(K_1, n_1, p_1) \text{binom}(K_2, n_2, p_2) \text{binom}(K_3, n_3, p_3)$$

fica

$$L(\hat{p}_0, \hat{p}_0, \hat{p}_0) = \text{binom}(K_1, n_1, \hat{p}_0) \text{binom}(K_2, n_2, \hat{p}_0) \text{binom}(K_3, n_3, \hat{p}_0)$$

sendo

$$\hat{p}_0 = \frac{K_1 + K_2 + K_3}{n_1 + n_2 + n_3} = \frac{K}{n}$$

o estimador de verossimilhança.

O denominador de Λ , por outro lado, fica

$$L(\hat{p}_1, \hat{p}_2, \hat{p}_3) = \text{binom}(K_1, n_1, \hat{p}_1) \text{binom}(K_2, n_2, \hat{p}_2) \text{binom}(K_3, n_3, \hat{p}_3)$$

sendo

$$\hat{p}_i = \frac{K_i}{n_i} \quad i \in \{1, 2, 3\}$$

Assim, a estatística do teste fica

$$\Lambda(K_1, K_2, K_3) = \frac{L(\hat{P}_0, P_0, \hat{P}_0)}{L(P_1, \hat{P}_2, \hat{P}_3)}$$

Assumindo que $-2 \log \Lambda$ têm distribuição chi-quadrado com $r - r_0 = 2$ graus de liberdade ($r_0 = 1$, número de parâmetros sob a hipótese nula; $r = 3$, número de parâmetros sob a hipótese alternativa) podemos encontrar o limiar $\chi^2_{r-r_0}(2)$, como ilustrado no notebook ($\chi^2(0,05) \approx 5,99$).

Exemplo notebook

Teste de Permutação

Esse teste é útil para comparar se duas amostras têm a mesma distribuição. Para apresentá-lo, tome

$$\textcircled{R} \quad X_1, X_2, \dots, X_m \rightsquigarrow F$$

$$\textcircled{R} \quad Y_1, Y_2, \dots, Y_n \rightsquigarrow G$$

Vamos testar as hipóteses:

$$H_0: F = G$$

$$H_1: F \neq G$$

Suponha que a estatística do teste seja

$$T(X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n) = |\bar{X} - \bar{Y}|$$

Sob a hipótese nula, qualquer permutação entre os elementos $(X_1, X_2, \dots, X_m, Y_1, Y_2, \dots, Y_n)$ é igualmente provável. Temos, portanto, $(n+m)!$ possibilidades

para a estatística T

$$\{T_1, T_2, \dots, T_{(n+m)!}\}.$$

A ideia do teste é que esse valores não devem diferir da estatística obtida para os dados reais (t_0). Além disso, como cada T_j tem probabilidade $1/(n+m)!$, de modo que a probabilidade de $T_j > t_0$, fica

$$P(T > t_0) = \frac{1}{(n+m)!} \sum_{j=1}^{(n+m)!} I(T_j > t_0)$$

Note que isso significa o número de vezes que $T_j > t_0$ em todas as possíveis permutações dos elementos ~~de~~ X_i combinados com os Y_i . Se essa quantidade for pequena, rejeitamos H_0 , caso contrário, não rejeitamos. Note ainda que $P(T > t_0)$ é o p-valor do teste. Caso $(n+m)!$ seja muito grande, podemos simplesmente fazer uma amostragem aleatória das permutações, nesse caso costuma-se chamar o teste de random permutation test ou bootstrap hypothesis test. Vale ~~obs~~ observar também que a mesma ideia pode ser usada para comparar outras quantidades estatísticas.

Exemplo notebook

Teste de Wald

Suponha as hipóteses:

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

A estatística desse teste é

$$W = \frac{\hat{\theta}_n - \theta_0}{\text{se}}$$

sendo $\hat{\theta}_n$ o estimador de verossimilhança e se o erro padrão

$$\text{se} = \sqrt{V(\hat{\theta}_n)}$$

De modo geral, $W \xrightarrow{d} N(0,1)$, ou seja, a estatística de Wald é assintoticamente gaussiana, o que permite decidir entre H_0 e H_1 buscando na comparação $|W| > z_{\alpha/2}$ sendo

$$P(|z| > z_{\alpha/2}) = \alpha$$

Exemplo: binomial. Nesse caso, já vimos que

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\text{Var}(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n}$$

Assim, temos

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}}$$

Exemplo notebook

Múltiplos testes de Hipótese

Suponha que tenhamos uma situação envolvendo múltiplas comparações, ou seja,

$$H_0: \theta = \theta_0$$

$$H_1: \theta = \theta_1$$

$$H_2: \theta = \theta_2$$

⋮

$$H_n: \theta = \theta_n$$

Podemos encontrar o p-valor associado a cada comparação entre H_0 e H_k ($k=1, 2, \dots, n$). Suponha que todos as H_k sejam falsas, nesse caso, a probabilidade de rejeitar H_0 em n comparações é

$$FPR = P_{FA} = 1 - (1-\alpha)^n$$

sendo $(1-\alpha)$ a probabilidade de rejeitar H_0 em uma comparação num nível de significância α . O problema de múltiplas comparações é que

$$FPR = P_{FA} \rightarrow 1 \quad \text{se } n \rightarrow \infty$$

Ou seja, existe uma probabilidade muito grande de rejeitar H_0 quando ela é verdadeira apenas ao acaso. Uma das soluções mais difundidas para esse problema é a chamada Bonferroni correction, que consiste em substituir α por α/n . Desse modo,

$$FPR = P_{FA} = 1 - (1 - \frac{\alpha}{n})^n$$

e no limite de $n \rightarrow \infty$

$$P_{FA} = 1 - e^{-\alpha} \approx 1 - (1 - \alpha)$$

$$P_{FA} \approx \alpha$$

Por exemplo, se $\alpha = 0,05$ e $n = 20$, temos

$$FPR = 1 - (1 - 0,05)^{20}$$
$$\approx 0,64$$

Usando a correção de Bonferroni

$$FPR = 1 - \left(1 - \frac{0,05}{20}\right)^{20}$$
$$\approx 0,0488$$

Uma outra possibilidade para contornar o problema de $FPR \rightarrow 1$ se $n \rightarrow \infty$ é o chamado procedimento de Benjamini-Hochberg (BH step-up procedure). Essa abordagem consiste em ordenar p-valores de cada comparação em ordem crescente $\{P^{(1)}, P^{(2)}, \dots, P^{(n)}\}$ e encontrar um P_K tal que:

$$P_K \leq \frac{K\alpha}{n} \quad \text{sendo } K \text{ o índice dos p-valores na lista ordenada}$$

Em seguida, rejeitamos todos as hipóteses nulas $H_0^{(i)}$ para $i \leq K$.

Na prática costuma-se fazer um gráfico de P_K versus K e comparar com a reta $f(K) = \frac{K\alpha}{n}$

