Análise de Dados

Bruno Tebaldi

Mestrado Profissional em Economia

October 23, 2022

1 Aula 7 - Intervalo de confiança

- Inferência
  - População e Amostra
  - Tipos de Amostra
- Propriedades em amostra finita de um estimador
  - Viés de um estimador
  - Erro Quadrático Médio
- Propriedades Assintóticas de um estimador
  - Viés Assintótico
  - Consistência
  - Lei dos Grandes Números Fraca (LGN)
  - Lei dos Grandes Números Fraca (LGN)
  - Teorema do Limite Central
- Intervalo de confiança
  - Intervalo de confiança da média
  - Intervalo de confiança da Proporção
  - Intervalo de confiança da Média com variância desconhecida
  - Intervalo de confiança - Diferença de média com variância conhecida
  - Intervalo de confiança da Variância
- Intervalo de confiança para Razões de variâncias

População e Amostra

Definição: População e Amostra

População é o conjunto de todos elementos ou resultados sob investigação. Amostra é qualquer subconjunto da população.

Parâmetro e Estatística

Seja x uma amostra de uma v.a. $X \sim F$

Parâmetro

Um parâmetro θ é qualquer função (mensurável) de F, ou seja $\theta = \theta(F)$

Estatística

Uma estatística t é qualquer função da amostra $t = t(x)$

Estimador e Estimativa

Seja x uma amostra de uma v.a. $X \sim F$

Estimador

Um estimador $\hat{\theta}$ é uma v.a. (função do espaço amostral via X) que é usada para inferir sobre um parâmetro θ. $\hat{\theta} = \hat{\theta}(X)$

Estimativa

Uma estimativa é um valor particular do estimador associado a uma amostra $(X = x)$. Neste caso uma estimativa é uma estatística: $\hat{\theta}(x) = \hat{\theta}(X = x)$

Tipos de Amostra

Seja $X_1, \ldots, X_n$ um conjunto de v.a. (uma amostra) com pdf conjunta $f(x_1, \ldots, x_n)$

- **Identicamente distribuída:** Uma amostra é dita identicamente distribuída se cada X_i tem a mesma distribuição marginal, ou seja: $f_i = f_j \quad \forall i, j$
- **Independente:** Uma amostra é dita independente se todas as v.a. que a compõem sejam mutualmente independentes, o que implica em $f = f_1 f_2 \ldots f_n$

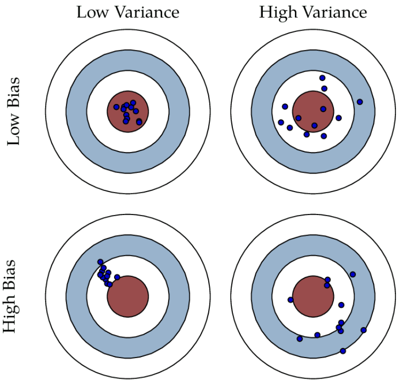
Tipos de Amostra

Podemos, portanto, classificar uma amostra de acordo com sua estrutura de dependência e/ou heterogeneidade como:

- **iid:** Independente e identicamente distribuída (homogênea)
- **niid:** Não independente e identicamente distribuída. (Ex.: Séries de Tempo que apresentam correlação serial)
- **inid:** Independente e não é identicamente distribuída (apresenta heterogeneidade nas marginais).
- **ninid:** Não independente e não é identicamente distribuída.

Propriedades de um estimador

Vamos imaginar um estimador que estima o centro de um alvo. Qual dos cenários abaixo é o mais desejado?



Viés de um estimador

Definição

Formalmente, o viés de um estimador $\hat{\theta}$ para o parâmetro θ é dado por

$$Vies(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

Chamamos um estimador de **não-viesado** se $Vies(\hat{\theta}) = 0$, caso contrário o estimador é dito viesado.

Eficiência de um estimador

- É desejável que, além de não viesado, o estimador seja o mais preciso possível, em outras palavras, tenha a menor variância possível. Este é o conceito de eficiência.
- Dizemos que um estimador é eficiente dentro de uma classe de estimadores se:
  1. for não viesado;
  2. entre os estimadores não viesados da mesma classe, apresentar a menor variância.

Eficiência de um estimador

Example

Para o estimador da média aritmética $\bar{X}$, calcule o viés e a variância, quando temos uma amostra i.i.d., na qual $\mathbb{E}\left[X_i\right] = \mu$ e $\mathbb{V}ar\left[X_i\right] = \sigma^2$.

Erro Quadrático Médio

- Sejam dois estimadores distintos $\hat{\theta}_1$ e $\hat{\theta}_2$ para o parâmetro θ. Qual escolher?
- Utilizaremos o erro quadrático médio para se escolher entre dois estimadores.

Erro Quadrático Médio (MSE)

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Pode-se mostrar que:

$$\text{MSE}(\hat{\theta}) = \mathbb{V}ar\left[\hat{\theta}\right] + \text{Viés}^2(\hat{\theta})$$

A média aritmética é BLUE

Example

Considere os estimadores lineares não viesados da forma:

$$\hat{\theta}_i = \sum_{i=1}^{n} \omega_i X \qquad \text{onde} \sum_{i=1}^{n} \omega_i = 1$$

A variância é dado por

$$\mathbb{V}ar\left[\hat{\theta}_i\right] = \sigma^2 \left(\sum_{i=1}^{n} w_i^2\right)$$

Achar o melhor estimador linear não viesado para a média se reduz a um problema de minimização com restrição.

$$\min \left\{\sum_{i=1}^{n} w_i^2\right\} s.a. \sum_{i=1}^{n} w_i = 1$$

Solução: $w_i = 1/n$, portanto escolhemos $\hat{\theta}_1$ como estimador para θ

Viés Assintótico

Assintóticamente não-viesado

Um estimador é dito assintóticamente não-viesado se

$$\lim_{n \to \infty} \mathbb{E}\left[\hat{\theta}\right] = \theta$$

Viés Assintótico

Example

Considere a variância amostral como estimador de σ^2 dado por $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Note que é um estimador viesado já que

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$$

mas é assintoticamente não-viesado uma vez que:

$$\underset{n\to\infty}{lim}\mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2$$

Por essa razão é comum se dividir por $n-1$ ao invés de n.

Consistência

- Dizemos que um estimador $\hat{\theta}$ é consistente para o parâmetro θ se a medida que a amostra cresce ele converge (em probabilidade) para o parâmetro amostral. Denotaremos por:

$$\hat{\theta} \xrightarrow{p} \theta$$

Definição: Convergência em Probabilidade

Uma sequência de variáveis aleatórias $X_1, X_2, \ldots, X_n$ converge em probabilidade para uma variável aleatória X se, para todo $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n - X| \geq \epsilon\right) = 0$$

Consistência

- Uma maneira de garantirmos a convergência em probabilidade é quando temos um estimador assintoticamente não viesado cuja a variância tende a zero a medida que a amostra cresce.

Consistência (condição suficiente)

Logo, um estimador $\hat{\theta}$ será consistente para θ se

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}\right] = \theta$$

$$\lim_{n\to\infty} \mathbb{V}ar\left[\hat{\theta}\right] = 0$$

- A média amostral é um estimador consistente da média, pois é um estimador não viesado (logo assintóticamente não-viesado) e sua variância converge para zero:

$$\mathbb{V}ar\left[\bar{X}\right] = \frac{1}{n}\sigma^2 \longrightarrow 0$$

Lei dos Grandes Números Fraca (LGN)

Teorema LGN

Seja $X_1; \ldots, X_n$ uma amostra aleatória e g uma função (mensurável) então

$$\frac{1}{n}\sum_{i=1}^{n} g(X) \xrightarrow{p} \mathbb{E}\left[g(X)\right]$$

- Caso $g(X) = X$ temos o resultado padrão da média amostral convergindo para média populacional.

Example

$$\frac{1}{n}\sum_{i=1}^{n} X \xrightarrow{p} \mathbb{E}\left[X\right] = \mu$$

Lei dos Grandes Números Fraca (LGN)

Example

Fazer um exemplo computacional que simule a lei dos grandes numeros.

Teorema do Limite Central

Teorema do Limite Central

Seja $X_1, \ldots, X_n$ uma amostra aleatória, então se $\mathbb{E}\left[X^2\right] < \infty$, temos:

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

Onde $\mu = \mathbb{E}[X]$ e $\sigma^2 = \mathbb{V}ar[X]$

- Note que a v.a. $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ tem média zero e variância 1 por definição.
- A notação "$\xrightarrow{d}$" significa que a cdf de $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ se torna arbitrariamente próxima da cdf de uma N(0; 1) a medida que n cresce.

Resumindo

A grosso modo, estes dois resultados nos mostram que com o aumento do tamanho da amostra temos:

- **Lei dos Grandes Números (LGN):** A média amostral converge para a média populacional.
- **Teorema do Limite Central (TLC):** A distribuição de uma média amostral, devidamente padronizada, se assemelha a uma distribuição normal.

Intervalo de confiança
Introdução

- Aprendemos a encontrar estimadores (pontuais) para um determinado parâmetro de interesse θ
- Imagine que, para uma determinada amostra, tenhamos uma estimativa para um parâmetro de $\hat{\theta} = 7,35$
- Qual a confiabilidade que temos nesta estimativa? Conseguimos dizer que $\theta = 7$? Que garantia temos que $\theta = 0$ ou $\theta = 100$?

Intervalo de confiança da média

com variância conhecida

- Suponha que queiramos estimar a média μ de uma população qualquer, e para tanto usamos a média $\bar{x}$ de uma amostra de tamanho n.
- Assumindo que a população tem distribuição normal (ou recorrendo ao Teorema do Limite Central), sabemos que:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Ou:

$$Z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

- Sabemos da tabela da distribuição Normal Padrão que:

$$\mathbb{P}(|Z| < 1.96) = 0.95 \Leftrightarrow \mathbb{P}\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Intervalo de confiança da média

com variância conhecida

$$\mathbb{P}\left(\underbrace{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}}_{\text{Intervalo de Confiança}}\right) = \underbrace{0.95}_{\text{Confiabilidade}}$$

O que um Intervalo de Confiança de significa:

- Vamos assumir que temos um nível de confiança de 95%. Logo o IC significa que o procedimento que adotamos para construir o intervalo em aproximadamente 95% das vezes contém o parâmetro verdadeiro!

Intervalo de confiança da média

com variância conhecida

Talvez seja mais fácil começar com a pergunta: o que um IC de uma amostra com confiabilidade de $1 - \alpha$ **NÃO** significa:

- Que existe uma probabilidade de $1 - \alpha$ que o parâmetro populacional pertença a este intervalo
- Que existe uma probabilidade de $1 - \alpha$ que o IC contenha o parâmetro de interesse

Lembre-se que o IC de uma amostra é uma realização, ou ele contém ou não o parâmetro de interesse. O nível de confiança é escolhido pelo estatístico e a confiabilidade tem a ver com o procedimento adotado para calcular o IC, e não com o IC em particular.

Intervalo de confiança da média

com variância conhecida

Example (IC para média com variância conhecida)

Seu gerente te apresenta um produto que, segundo ele, tem um retorno excelente. Além disso ele diz que o produto possui um baixo risco. Ele diz que o retorno médio desse produto é de 1,95% ao mês. Numa amostra de 36 meses, você observa uma média de 1,93%. Sabe-se que o desvio-padrão dos retornos desse produto é 0,12%.

- Você pode afirmar que o retorno do tal produto oferecido pelo gerente é igual a 1,95%?
- Vamos construir um intervalo de confiança para o nosso estimador. Isso é, vamos construir um IC (com 90% de confiança) para a média populacional.

$\mathbb{IC}_{0.9} = [1.897; 1.962]$

Intervalo de confiança da média

com variância conhecida

Logo generalizando temos que o intervalo de confiança para uma média com variância conhecida é dado por:

$$\mathbb{IC}_{\gamma} = \bar{X} \pm Z^{c}_{\frac{1-\gamma}{2}} \frac{\sigma}{\sqrt{n}}$$

Aonde γ é o nível de confiança

- Há um *trade-off*: se aumentamos o nível de confiança, a precisão do intervalo cai (a margem de erro aumenta).
- Como fazer para aumentar tanto a precisão do intervalo como a sua confiança, ou, pelo menos, aumentar uma sem diminuir a outra?
- A única maneira é aumentando o número de observações, em outras palavras, aumentar o tamanho da amostra.

Intervalo de confiança da Proporção

$$\mathbb{IC}_\gamma = \hat{p} \pm Z^c_{\frac{1-\gamma}{2}} \sqrt{\frac{p(1-p)}{n}}$$

Note que seria necessário saber $p(1-p)$, porém como não temos, utilizamos o estimador consistente desse valor que é $\hat{p}(1-\hat{p})$

Example

Numa pesquisa de mercado, 400 pessoas foram entrevistadas sobre determinado produto e 60% delas preferiram a marca A.

- Construa um intervalo de confiança para essa proporção com um nível de confiança de 95%

$\mathbb{IC}_{0.95} = 0.6 \pm 0.048$

Intervalo de confiança da média com variância deconhecida

$$\mathbb{IC}_\gamma = \bar{X} \pm t^c_{\frac{1-\gamma}{2};n-1}\frac{S}{\sqrt{n}}$$

Example (Média com variância desconhecida)

Segundo seu gerente o produto que ele te oferece tem um retorno médio de 1,95% ao mês. Numa amostra de 36 meses, você observa uma média de 1,93%. O desvio padrão **dessa amostra** é 0,12%.

- Construa um intervalo de confiança (com 90% de confiança) para a média populacional.

$\mathbb{IC}_{0.9} = 1.93 \pm 0.09 = [1.84; 2.02]$

Intervalo de confiança Diferença de médias

com variância conhecida

$$\mathbb{IC}_\gamma = \left(\bar{X}_1 - \bar{X}_2\right) \pm Z^c_{\frac{1-\gamma}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

onde X_1 são provenientes da 1a população e X_2 são provenientes da 2a população. As populações são independentes e os sorteios i.i.d.

Intervalo de confiança da Variância

Generalizando temos que o intervalo de confiança para a variância é dado por:

$$\mathbb{IC}_\gamma = \left[\frac{(n-1)S^2}{\chi^2_{\text{sup};n-1}}; \frac{(n-1)S^2}{\chi^2_{\text{inf};n-1}} \right]$$

Example (Variância)

O desvio padrão das alturas de 16 estudantes escolhidos aleatoriamente em uma escola com 1000 estudantes é de 8,3 cm.

- Encontre limites de confiança de 95%, 99% do desvio padrão para todos os estudantes do sexo masculino desta escola assumindo que a altura é normalmente distribuída.

$\mathbb{IC}_{0.95} = [6.13; 12.85]$ $\mathbb{IC}_{0.99} = [5.61; 14.99]$

Intervalo de confiança para Razões de variâncias

Generalizando temos que o intervalo de confiança para a variância é dado por:

$$\mathbb{IC}_\gamma = \left[\frac{1}{F_{\frac{\alpha}{2}}}\frac{\hat{S}_1^2}{\hat{S}_2^2}; \frac{1}{F_{1-\frac{\alpha}{2}}}\frac{\hat{S}_1^2}{\hat{S}_2^2}\right]$$

Example

Duas amostras de tamanhos 16 e 10, respectivamente, são extraídas aleatoriamente de duas populações normais. Se duas variâncias estimadas forem 24 e 18 encontre limites de confiança de 98% e 90% para a razão das variâncias.
$\mathbb{IC}_{0.98} = [0.269; 5.188]$ $\mathbb{IC}_{0.90} = [0.444; 3.454]$

Referências

Bussab, Wilton de O.; Morettin, Pedro A. (2014)
Estatística básica 8a ed.
Saraiva
Cap. 11

Meyer, Paul L. (1983)
Probabilidade: Aplicações à Estatística
Livros Técnicos e Científicos Editora
Cap. 14

Casella, George; Berger, Roger L. (2011)
Inferência estatística
Cengage Learning
Cap. 9

2 Aula 8 - Teste de Hipótese

- Teste de Hipótese
  - Hipóteses Nula e Alternativa
  - p-valor
- Teste de Hipótese Exatos
  - Teste de média com variância conhecida
  - Teste de comparação de médias com variância conhecida
  - Teste de variância
  - Teste de média com variância desconhecida
  - Teste de comparação de variância
- Testes Exatos vs. Assintóticos

O que é uma Hipótese

Definition (Hipótese)

Uma afirmação sobre um parâmetro populacional. Matematicamente temos: dado um parâmetro $\theta \in \Theta$ [1]uma hipótese é simplesmente $H : \theta \in \Theta^* \subseteq \Theta$ [2]

[1] Θ é o espaço que contem todos os valores que θ pode assumir

[2] Hipotese: θ pertence a um subconjunto do Θ^* so espaço parametrico Θ

O que é uma Hipótese

Examples

- A média de altura da população é de 1,70m com variância de 0,30;
- A média de salário da população de homens e mulheres é a mesma;
- Retornos futuros de determinado ativo podem ser previstos por retornos passados;
- A distribuição (cdf) que gerou estas duas determinadas amostras é a mesma;

- O objetivo final do teste de hipótese é decidir, baseado em uma amostra da população, se a hipótese deve ser descartada ou não.

Hipóteses Nula e Alternativa

Definition (Hipóteses Nula)

As duas hipóteses **complementares** em um problema de teste de hipóteses são chamadas hipótese nula e hipótese alternativa. Elas são designados H_0 e H_1, respectivamente.

- Em geral, a hipótese que estamos interessados em testar é conhecida como hipótese nula e é denotada por $H_0 : \theta \in \Theta_0 \subseteq \Theta$
- O complementar de H_0 é chamada de hipótese alternativa é denotada por $H_1 : \theta \in \Theta_1 \subseteq \Theta$ (notação H_a também é usada)

Hipóteses Nula e Alternativa

Examples (Percentual de intenção de votos)

Seja θ o percentual de eleitores que deseja votar no candidato A em determinada população, podemos elaborar as seguintes hipóteses:

- $H_0 : \theta = 20\%$ versus $H_1 : \theta \neq 20\%$
- $H_0 : \theta \geq 20\%$ versus $H_1 : \theta < 20\%$

Erro tipo I e tipo II

H_0 pode ser verdadeira ou não. Assim, temos dois tipos de erros:

Definition (Erro tipo I)

Defini-se como **Erro tipo I** ao ato de se rejeitar a hipótese nula, H_0, quando essa é verdadeira. Chamamos de α a probabilidade de cometer esse erro.

$$\alpha = \mathbb{P}\left(\text{erro do tipo I}\right) = \mathbb{P}\left(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}\right)$$

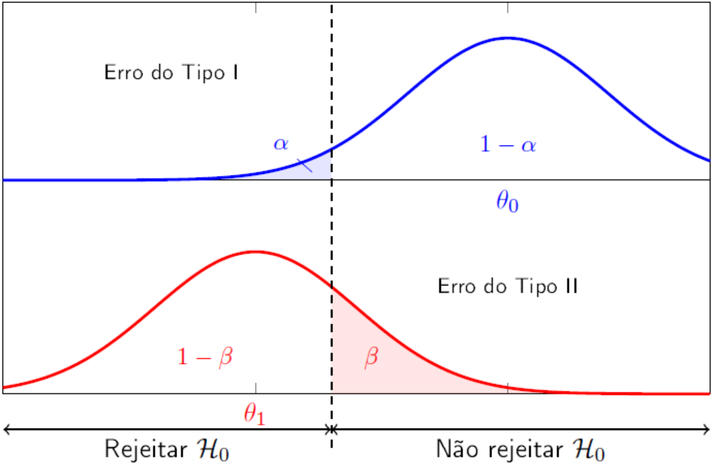
Definition (Erro tipo II)

Defini-se como **Erro tipo II** ao ato de se não rejeitar a hipótese nula, H_0, quando essa é falsa. Chamamos de β a probabilidade de cometer esse erro.

$$\beta = \mathbb{P}\left(\text{erro do tipo II}\right) = \mathbb{P}\left(\text{aceitar } H_0 | H_0 \text{ é falsa}\right)$$

Erro tipo I e tipo II

	Não Rejeita $\mathcal{H}_0$	Rejeita $\mathcal{H}_0$
$\mathcal{H}_0$ verdadeira	Decisão Correta	**Erro tipo I**
$\mathcal{H}_0$ falsa	**Erro tipo II**	Decisão Correta



Teste de hipótese

Definition (Teste de hipótese)

Um procedimento de teste de hipótese ou teste de hipótese é uma regra que especifica:

1. Para quais valores de amostra é tomada a decisão de aceitar H_0 como verdadeiro.
2. Para quais valores de amostra H_0 é rejeitada e H_1 é aceito como verdadeiro.

- A função do teste de hipóteses é dizer, usando uma estatística $\hat{\theta}$, se a hipótese H_0 é ou não aceitável.

Região Crítica

Definition (Região Crítica)

O subconjunto do espaço de amostra para o qual H_0 será rejeitado é chamado de região de rejeição ou região crítica ($\mathcal{RC}$).

- Caso o valor obtido da estatística pertença a região crítica, rejeitamos H_0; caso contrário, não rejeitamos H_0.
- A região crítica é construída de modo que $\mathbb{P}\left(\hat{\theta} \in \mathcal{RC} | H_0\right) = \alpha$.
- α é o nível de significância do teste e é fixado a priori.

Região Crítica

- Dada uma estatística de teste w, definimos um subconjunto $\mathcal{RC} \subseteq \mathbb{R}$ conhecido como região crítica já que:
  - Se $w \in \mathcal{RC}$, então rejeitamos H_0
  - Se $w \notin \mathcal{RC}$, então não rejeitamos H_0
- Caso $\mathcal{RC}$ seja intervalos do tipo (c, ∞) ou $(-\infty, c)$, chamamos $c \in \mathbb{R}$ de valor crítico do teste.

Example (Percentual de intenção de votos)

Seja θ o percentual de eleitores que deseja votar no candidato A em determinada população. Definimos a $\mathcal{RC}$.
$H_0 : \theta = 20\%$ versus $H_1 : \theta \neq 20\%$, se $|\bar{x} - 20\%| \geq c$ rejeitamos H_0 logo $\mathcal{RC} = \{m \in \mathbb{R} \ : |m - 0.2| \geq c\}$

Trade-off entre Erro do Tipo I e II

- Idealmente, gostaríamos de procurar testes que minimizem simultaneamente a probabilidade de ambos os erros.
- Entretanto, é fácil perceber que a medida que diminuímos a probabilidade do erro do Tipo I (diminuindo a região crítica, por exemplo), automaticamente aumentamos a probabilidade do Erro do Tipo II e vice-versa.
- Portanto, na prática, fixamos o nível de significância, por exemplo: $\alpha = 10\%; 5\%; 1\%$ e tentamos minimizar β ou, equivalentemente, maximizar o poder do teste $1 - \beta$.
- A única maneira de reduzirmos ambos simultaneamente é aumentando o tamanho da amostra.

p-valor de um Teste de Hipótese

- Como a rejeição ou não de H_0 é função do nível de significância α, é comum reportar uma estatística conhecida como **p-valor**.

Definition (p-valor)

O p-valor é definido como a probabilidade, sob a hipótese nula, de obter um resultado igual ou mais extremo do que o que foi realmente observado.

- Assim p-valores pequenos são indícios contra H_0 em favor de H_1
- A grande vantagem na utilização de um p-valor e evitar limiares arbitrários. Você deixa o leitor decidir que nível de significância se sente confortável para rejeitar H_0

p-valor

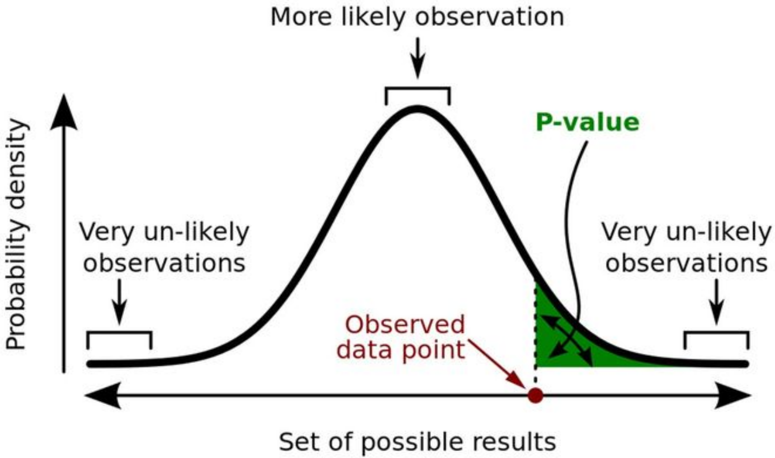


Figure: p-valor de um teste **uni-caudal**

Teste de Hipótese Exatos

- É crucial que conheçamos a distribuição da estatística de teste, para que possamos dizer algo sobre a significância e/ou poder de um teste.
- Assim os chamados Testes Exatos (sem argumento assintótico) são escassos pois mesmo que postulemos uma pdf para nosso amostra aleatória, nada nos garante que saberemos a distribuição da estatística de teste escolhida.
- Portanto, ou assumimos normalidade da amostra e uma estatística de teste simples como uma média ou variância amostral ou teremos que usar resultados assintóticos (Teorema Central do Limite).

Teste de hipótese

Passos para um teste de hipótese

1. Definir a Hipotese nula (e alternativa por complementariedade)
2. Definir a estatistica de teste e a distribuição associada.
3. Escolher o nível de significancia para o teste.
4. Com o nível de significancia e a distribuição determinar o(s) valor(es) crítico(s) e região critica
5. Utilizar os dados da amostra para verificar a hipotese nula e com isso concluir o teste

Teste de média com variância conhecida

Como proceder um teste de média com variância conhecida

- Amostra aleatória de tamanho n de uma população $N(\mu, \sigma^2)$, onde σ^2 é conhecido. Fazemos uma hipotese sobre a média

$$H_0 : \mu = \mu_0$$

- Usaremos como estatística de teste

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- Queremos testar a um nível de significância α, o que implica que os valores críticos c são tais $\mathbb{P}\left(Z \leq c_\alpha\right) = \alpha$
- H_1 será definido como o complementar de H_0, logo teremos que:

$$\mathcal{RC} = \left\{ w \in \mathbb{R} \;\middle|\; |w| > c_{1-\frac{\alpha}{2}} \right\} \qquad H_1 : \mu \neq \mu_0, \quad \text{bicaudal}$$

- Se $Z \in \mathcal{RC}$ rejeito H_0, caso contrario não rejeito.

Teste de média com variância conhecida

- Além do teste de igualdade (bicaudal) podemos também ter testes unicaudais

$$\begin{array}{lcll} H_0 : \mu = \mu_0 & \Rightarrow & H_1 : \mu \neq \mu_0, & \text{bicaudal} \\ H_0 : \mu \geq \mu_0 & \Rightarrow & H_1 : \mu < \mu_0, & \text{cauda inferior} \\ H_0 : \mu \leq \mu_0 & \Rightarrow & H_1 : \mu > \mu_0, & \text{cauda superior} \end{array}$$

- Dependendo de H_1, escolhemos nossa região crítica:

$$\mathcal{RC} = \begin{cases} |w| > c_{1-\alpha/2} & \text{se} \quad H_1 : \mu \neq \mu_0, \quad \text{bicaudal} \\ w > c_{1-\alpha} & \text{se} \quad H_1 : \mu > \mu_0, \quad \text{cauda superior} \\ w < c_{\alpha} & \text{se} \quad H_1 : \mu < \mu_0, \quad \text{cauda inferior} \end{cases}$$

- onde os valores críticos c são tais $\mathbb{P}\left(Z \leq c_{\alpha}\right) = \alpha$

Teste de média com variância conhecida

Example (Teste bicaudal)

Afirma-se que a altura média dos jogadores de basquete que disputam uma determinada liga é 1,95m. Numa amostra de 36 jogadores, foi encontrada uma média de 1,93m.
Sabe-se que o desvio-padrão da altura dos jogadores é 12cm.
Assumindo que a distribuição da sua variável de interesse possui uma distribuição normal, teste, com um nível de significância de 10%, se a afirmação é verdadeira.

- Qual a hipótese nula?
- Qual a hipótese alternativa?
- Qual a estatística de teste você usa nesse caso?

Teste de comparação de médias com variância conhecida

- Duas amostra aleatórias, uma de tamanho n_1 com $X_1 \sim N(\mu_1, \sigma_1^2)$ e outra de tamanho n_2 com $X_2 \sim N(\mu_2, \sigma_2^2)$. Queremos testar a um nível de significância α umas das hipóteses abaixo

$$
\begin{aligned}
H_0 : \mu_1 = \mu_2 \quad &\Rightarrow \quad H_1 : \mu_1 \neq \mu_2 \\
H_0 : \mu_1 \geq \mu_2 \quad &\Rightarrow \quad H_1 : \mu_1 < \mu_2 \\
H_0 : \mu_1 \leq \mu_2 \quad &\Rightarrow \quad H_1 : \mu_1 > \mu_2
\end{aligned}
$$

- Se defnirmos $Y = \bar{X}_1 - \bar{X}_2$, então teremos

$$
\begin{aligned}
\mu_Y &= \mathbb{E}\left[Y\right] = \mu_1 - \mu_2 \\
\sigma_Y^2 &= \mathbb{V}ar\left[Y\right] = \mathbb{V}ar\left[\bar{X}_1 - \bar{X}_2\right] = \mathbb{V}ar\left[\bar{X}_1\right] + \mathbb{V}ar\left[\bar{X}_2\right] \\
\sigma_Y^2 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
\end{aligned}
$$

Teste de comparação de médias com variância conhecida

- Sendo assim podemos redefir o teste de comparacao de médias como:

$$\begin{aligned} H_0 : Y = 0 \quad &\Rightarrow \quad H_1 : Y \neq 0 \\ H_0 : Y \geq 0 \quad &\Rightarrow \quad H_1 : Y < 0 \\ H_0 : Y \leq 0 \quad &\Rightarrow \quad H_1 : Y > 0 \end{aligned}$$

- Usaremos a seguinte estatistica:

$$W = \frac{Y - \mu_Y}{\sigma_Y} \sim N(0,1)$$

- E a regiao critica pode ser definida como:

$$\mathcal{RC} = \left\{ w \in \mathbb{R} \;\middle|\; |w| > c_{1-\frac{\alpha}{2}} \right\} \qquad H_1 : Y \neq 0, \quad \text{bicaudal}$$

- As mesmas considerações devem ser feitas para encontrar valores críticos da Normal, caso a alternativa seja $Y > 0$ ou $Y < 0$.

Teste de comparação de médias com variância conhecida

Example

Fez-se um estudo sobre aluguéis em dois bairros, A e B. No primeiro, em 12 residências, o aluguel médio foi 330. No segundo, em 19 residências, o aluguel médio foi de 280. O desvio padrão dos aluguéis no bairro A é 50 e no bairro B 40. Gostaríamos de saber se a média do aluguel nesses dois bairros é igual a 10% de significância.

Teste de variância

- Amostra aleatória de tamanho n de uma $N(\mu, \sigma^2)$, ambos desconhecidos. Queremos testar a variância a uma significância α, logo teremos uma das hipoteses abaixo:

$$\begin{aligned} H_0 : \sigma^2 = \sigma_0^2 \quad &\Rightarrow \quad H_1 : \sigma^2 \neq \sigma_0^2 \\ H_0 : \sigma^2 \geq \sigma_0^2 \quad &\Rightarrow \quad H_1 : \sigma^2 < \sigma_0^2 \\ H_0 : \sigma^2 \leq \sigma_0^2 \quad &\Rightarrow \quad H_1 : \sigma^2 > \sigma_0^2 \end{aligned}$$

- Usaremos uma estatística de teste baseada em $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ pois não é difícil mostrar que

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Teste de variância

- Lembre-se que a Chi-Quadrado é uma distribuição assimetrica com valores sempre positivos. Dependendo de H_1, escolhemos nossa região crítica:

$$\mathcal{RC} = \begin{cases} w > c_{1-\alpha} & \text{se} \quad H_1 : \sigma^2 > \sigma_0^2, \quad \text{cauda superior} \\ w < c_{\alpha} & \text{se} \quad H_1 : \sigma^2 < \sigma_0^2, \quad \text{cauda inferior} \\ |w| > c_{1-\alpha/2} & \text{se} \quad H_1 : \sigma^2 \neq \sigma_0^2, \quad \text{bicaudal} \end{cases}$$

- Os valores críticos c são tais que $W \sim \chi^2_{n-1}$ e $\mathbb{P}\left(W \leq c_\alpha\right) = \alpha$.

Teste de Variância

Example

Uma caixa de fósforos de uma certa marca vem com a inscrição: "contém, em média, 40 palitos". Segundo o fabricante, o desvio padrão é de, no máximo, 2 palitos. Em uma amostra com 51 caixas foi encontrado um desvio padrão amostral de 3 palitos.

- Supondo que o número de palitos por caixa seja uma variável normal, teste a afirmativa do fabricante utilizando um nível de significância de 1%

Teste de média com variância desconhecida

- Seja $Z \sim N(0,1)$, e $Y \sim \chi^2_k$ e Z independente de Y , definimos:

$$W = \frac{Z}{\sqrt{\frac{Y}{k}}}$$

- Assim W segue distribuição de t-Student com k graus de liberdade e denotamos por $W \sim t_k$, com pdf dada por:

$$f(x|k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right) - \frac{k+1}{2}, \qquad k \in \mathbb{N}$$

Teste de média com variância desconhecida

- Vamos seguir o mesmo setup do teste de média com variaância conhecida, porém agora a variância é desconhecida e estimada por S^2 como no teste de variância.
- Uma amostra aleatória, de tamanho n com $X \sim N(\mu, \sigma^2)$. Queremos testar a um nível de significância α umas das hipoteses abaixo

$$
\begin{aligned}
H_0 : \mu = \mu_0 \quad &\Rightarrow \quad H_1 : \mu \neq \mu_0 \\
H_0 : \mu \geq \mu_0 \quad &\Rightarrow \quad H_1 : \mu < \mu_0 \\
H_0 : \mu \leq \mu_0 \quad &\Rightarrow \quad H_1 : \mu > \mu_0
\end{aligned}
$$

- A estatística de teste segue uma distribuição t-Student com:

$$
W = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{S^2}} \sim t_{n-1}
$$

Teste de média com variância desconhecida

- Dependendo de H_1, escolhemos nossa região crítica:

$$\mathcal{RC} = \begin{cases} w > c_{1-\alpha} & \text{se} \quad H_1 : \mu > \mu_0, \quad \text{cauda superior} \\ w < c_{\alpha} & \text{se} \quad H_1 : \mu < \mu_0, \quad \text{cauda inferior} \\ |w| > c_{1-\alpha/2} & \text{se} \quad H_1 : \mu \neq \mu_0, \quad \text{bicaudal} \end{cases}$$

- Os valores críticos c são tais que $X \sim t_{n-1}$ $\mathbb{P}\left(X \leq c_{\alpha}\right) = \alpha$

Teste de média com variância desconhecida

Example (Teste de Média com Variância Desconhecida)

Gostaríamos de saber se a média de salários de uma determinada empresa não é maior que 1000 reais a um nível de significância de 1% assumindo normalidade dos salários. De uma pesquisa com 5 empregados, temos $\bar{X} = 1030$ e $s = 25$.

Teste de média com variância desconhecida

Example (Teste de Média com Variância Desconhecida)

Gostaríamos de saber se a média de salários de uma determinada empresa não é maior que 1000 reais a um nível de significância de 1% assumindo normalidade dos salários. De uma pesquisa com 5 empregados, temos $\bar{X} = 1030$ e $s = 25$.

$$W = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{1030 - 1000}{25/\sqrt{5}} = 2.68$$

- O que aconteceria caso tivéssemos usado a estimativa da variância como valor verdadeiro?

Teste de média com variância desconhecida

Example (Exemplo no R)

A Tabela abaixo mostra a ingestão média diária de energia em dez dias em 11 mulheres saudáveis com idade entre 22 e 30 anos.

Subject	1	2	3	4	5	
Avg. Energy. intake	5260	5470	5640	6180	6390	

Subject	6	7	8	9	10	11
Avg. Energy. intake	6515	6805	7515	7515	8230	8770

Podemos afirmar que a média de ingestão diária foi inferior a 7725 KJ?

Teste de média com variância desconhecida

- À medida que aumentamos a amostra e, por conseguinte, os graus de liberdade, o valor encontrado na tabela t-Student se aproxima do valor da normal.
- Portanto, se a variância for desconhecida, mas a amostra for grande, fará pouca diferença se usarmos a normal ou a t-Student (e fará menos diferença quanto maior for a amostra).

Teste de comparação de variância

- Seja $Y_1 \sim \chi^2_{k_1}$, $Y_2 \sim \chi^2_{k_2}$ e Y_1 independente de Y_2, definimos:

$$W = \frac{\frac{Y_1}{k_1}}{\frac{Y_2}{k_2}}$$

- Assim W segue distribuição F com k_1 e k_2 graus de liberdade e denotamos por $W \sim F(k_1, k_2)$, com pdf dada por:

$$f(x|k_1,k_2) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1} \left(1+\frac{k_1}{k_2}x\right)^{-\frac{k_1+k_2}{2}}$$

onde $k_1, k_2 > 0$.

Teste de comparação de variância

- Duas amostra aleatórias, uma de tamanho n_1 com $X_1 \sim N(\mu_1, \sigma_1^2)$ e outra de tamanho n_2 com $X_2 \sim N(\mu_2, \sigma_2^2)$. Queremos testar aum nível de significância α umas das hipóteses abaixo.

$$
\begin{aligned}
H_0 : \sigma_1^2 = \sigma_2^2 \quad &\Rightarrow \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \\
H_0 : \sigma_1^2 \geq \sigma_2^2 \quad &\Rightarrow \quad H_1 : \sigma_1^2 < \sigma_2^2 \\
H_0 : \sigma_1^2 \leq \sigma_2^2 \quad &\Rightarrow \quad H_1 : \sigma_1^2 > \sigma_2^2
\end{aligned}
$$

- Se defnirmos $Y = \frac{\sigma_1^2}{\sigma_2^2}$, então teremos:

$$Y = \frac{\sigma_1^2}{\sigma_2^2}$$

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \Rightarrow \quad H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \Rightarrow \quad H_0 : Y = 1$$

Teste de comparação de variância

Se dividirmos suas respectivas variâncias amostrais, devidamente padronizadas, teremos uma estatística de teste com uma distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade:

$$W = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = \frac{s_1^2}{s_2^2} \cdot \frac{1}{Y} \sim \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)} \equiv F_{n_1-1,n_2-1}$$

Lembrando que sob H_0 temos que $Y = 1$!

Teste de comparação de variância

- Sendo assim podemos redefir o teste de comparacao de variância como:

$$
\begin{aligned}
H_0 : Y = 1 \quad &\Rightarrow \quad H_1 : Y \neq 1 \\
H_0 : Y \geq 1 \quad &\Rightarrow \quad H_1 : Y < 1 \\
H_0 : Y \leq 1 \quad &\Rightarrow \quad H_1 : Y > 1
\end{aligned}
$$

- A estatística de teste será $W \sim F_{n_1-1;n_2-1}$ com:

$$W = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

Teste de comparação de variância

- Dependendo de H_1, escolhemos nossa região crítica:

$$\mathcal{RC} = \begin{cases} w > c_{1-\alpha} & \text{se} \quad H_1 : Y > 1, \quad \text{cauda superior} \\ w < c_{\alpha} & \text{se} \quad H_1 : Y < 1, \quad \text{cauda inferior} \\ |w| > c_{1-\alpha/2} & \text{se} \quad H_1 : Y \neq 1, \quad \text{bicaudal} \end{cases}$$

- Os valores críticos c são tais que $W \sim F_{n_1-1;n_2-1}$ e $\mathbb{P}\left(W \leq c_\alpha\right) = \alpha$.

Teste de comparação de variância

Example

Fez-se um estudo sobre a volatilidade (desvio-padrão do retorno) de dois ativos, A e B. No primeiro, em 20 transações, a volatilidade amostral, medida pelo desvio-padrão, foi de 5%. No segundo, em 18 transações, a volatilidade foi de 8%.

- Gostaríamos de saber se os dois ativos têm a mesma volatilidade a 10% de significância, ou se a volatilidade de B é maior que de A?

Testes Exatos vs. Assintóticos

- Na grande maioria dos casos, preferimos nos abster de dizer algo sobre a pdf de uma amostra e usar argumentos assintóticos
- Note que todos os nossos testes partiram da hipótese que $X_1, X_2; \ldots, X_n \sim N(\mu; \sigma^2)$.
- Nesse caso, vimos que $\bar{x} \sim N(\mu; \sigma^2/n)$
- Note que em um teste assintótico, não precisamos impor essa distribuição para fazer inferência. Pois a distribuição da média é normal pelo T.L.C.

Testes Exatos vs. Assintóticos

Teorema

Seja $X_1, \ldots, X_n$ uma amostra aleatória, então

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

se $\mathbb{E}\left[X^2\right] < \infty$. Onde $\mu = \mathbb{E}\left[X\right]$ e $\sigma^2 = \mathbb{V}ar\left[X\right]$

- Note que a v.a. $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ tem média zero e variância 1 por definição.
- A notação "$\xrightarrow{d}$" significa que a cdf de $\sqrt{n}\frac{\bar{X}-\mu}{\sigma}$ se torna arbitrariamente próxima da cdf de uma $N(0,1)$ a medida que n cresce.
- Também existem extensões do TLC para amostras não aleatórias.

Teste sobre proporções

Seja, $X \sim Bernoulli(p)$, pelo teorema central do limite, $\bar{X}$ terá distribuição aproximadamente normal, com média p e variância $\frac{p(1-p)}{n}$, ou seja, podemos utilizar o arcabouço de média com variância conhecida.

Example

Uma pesquisa feita com 300 eleitores revelou que 23% votariam no candidato A. Seu rival, no entanto, o candidato B, afirma que o seu oponente tem, no máximo, 20% dos votos. Teste a afirmação do candidato B, utilizando um nível de significância de 5%.

Referências

Bussab, Wilton de O.; Morettin, Pedro A. (2014)
Estatística básica 8a ed.
Saraiva
Cap. 12, 13

Meyer, Paul L. (1983)
Probabilidade: Aplicações à Estatística
Livros Técnicos e Científicos Editora
Cap. 15

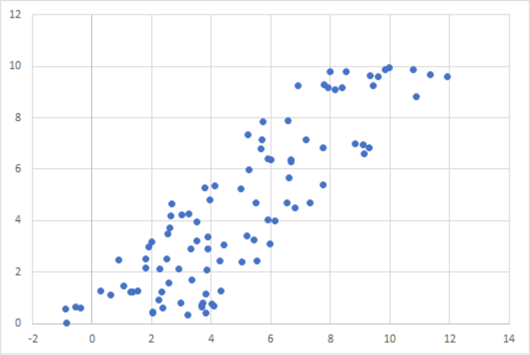
Casella, George; Berger, Roger L. (2011)
Inferência estatística
Cengage Learning
Cap. 8

3 Aula 9 - Mínimos Quadrados Ordináros

- Mínimos Quadrados
  - Motivação
  - Modelo Linear
  - Amostra Aleatória
  - Estimador de Mínimos Quadrados (MQO)
  - Propriedades do Estimador de Mínimos Quadrados

Introdução

- Vamos analisar a dependência de uma v.a. Y em relação a outra v.a. X.
- Em particular vamos assumir que queremos analisar os dados abaixo



Introdução

Motivação

Gostaríamos de:

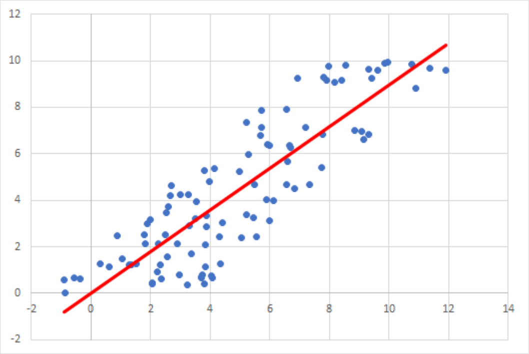
- Analisar o modelo que rege o comportamento dos dados.
- Descobrir os parâmetros que estão associados ao modelo.
- Utilizar o modelo para avaliar o comportamento conjunto das variáveis (econometria e time series)

Modelo

- A nossa **primeira hipótese**, a de que os dados se relacionam por uma relação linear.

Hipóteses

1. Modelo linear



Modelo

- Ao assumir que temos um modelo linear, estamos na realidade assumindo o modelo matemático:

$$Y_i = \alpha + \beta X_i + u_i$$

  onde, para cada elemento i, temos que: Y_i é a variável dependente (resposta), X_i é a variável independente (preditor), e u_i é o erro da nossa relação.
- Note que este modelo descreve a relação entre X_i e Y_i usando dois parâmetros: um intercepto (α) e a inclinação (β).

Modelo

- Podemos pensar nessa relação linear como uma reta que mais se aproxima dos dados.
- Logicamente que para isso a nossa amostra deve ser escolhida aleatoriamente.

Hipóteses

1. Modelo linear
2. A amostra aleatória

Estimador de Mínimos Quadrados

- O nosso objetivo será encontrar estimadores $\hat{\alpha}$ e $\hat{\beta}$ de tal maneira que a soma dos erros ao quadrado seja a menor possível.

$$\min_{\alpha,\beta} \sum_{i=1}^{n} u_i^2$$

$$\min_{\alpha,\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\alpha + \beta X_i))^2 \right\}$$

Estimador de Mínimos Quadrados

- Já sabemos da matemática que para encontrar o mínimo de uma função devemos derivar, igualar a zero e resolver para a variável.
- Para isso é necessário que tenhamos uma condição técnica de estimação, a de que a os dados não podem ser colineares. (em termos práticos, se os dados forem colineares, não haveria erros a serem minimizados). **Essa é a nossa terceira hipótese.**

Hipóteses

1. Modelo linear
2. A amostra aleatória
3. Os dados não são colineares

Estimador de Mínimos Quadrados

A partir das condições de primeira ordem temos:

$$S(\alpha, \beta) = \sum_{i=1}^{n} (Y_i - (\alpha + \beta X_i))^2$$
$$\frac{\partial S(\alpha, \beta)}{\partial \alpha} = 0$$
$$\frac{\partial S(\alpha, \beta)}{\partial \beta} = 0$$

Neste caso nossos estimadores serão dados por:

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$
$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- Para que os estimadores de MQO sejam não viesados é necessário que tenhamos o que conhecemos como exogeneidade ou seja $\mathbb{E}[u_i|X_i] = 0$. Essa condição pode ser relaxada. (Isso será visto com mais detalhes no curso de econometria).
- A hipótese de exogeneidade será nossa **quarta hipótese.**

$$\mathbb{E}\left[\hat{\beta}\right] = \beta$$
$$\mathbb{E}\left[\hat{\alpha}\right] = \alpha$$

Propriedades do Estimador de Mínimos Quadrados

Hipóteses

1. Modelo linear
2. A amostra aleatória
3. Os dados não são colineares
4. Exogeneidade de X_i

Propriedades do Estimador de Mínimos Quadrados

Variância

- Para que os estimadores de MQO sejam eficiente é necessário que tenhamos o que conhecemos como heterocedasticidade. Isso significa que os erros tem variancia constante e não são correlacionados entre si. $\mathbb{V}ar\left[u_i\right] = \sigma$. Essa condição também pode ser relaxada. (Novamente isso será visto com mais detalhes no curso de econometria e econometria de series de tempo).
- A hipótese de homoscedasticidade será nossa **quinta hipótese.**

$$\mathbb{V}ar\left[\hat{\beta}\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathbb{V}ar\left[\hat{\alpha}\right] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \right\}$$

Propriedades do Estimador de Mínimos Quadrados

Hipóteses

1. Modelo linear
2. A amostra aleatória
3. Os dados não são colineares
4. Exogeneidade de X_i
5. Homoscedasticidade

Propriedades do Estimador de Mínimos Quadrados

inferência

- Uma das coisas que é importante é fazermos inferência sobre os valores dos estimadores, $\hat{\alpha}$ e $\hat{\beta}$. Em outras palavras, desejamos testar hipóteses sobre os valores dos estimadores.
- Para isso será necessário que tenhamos uma distribuição e neste ponto assumimos nossa **sexta hipótese** de que os erros são normalmente distribuídos. Essa hipótese também pode ser relaxada pois temos que a distribuição Normal é garantida assintoticamente pelo Teorema do Limite Central.

Propriedades do Estimador de Mínimos Quadrados

Hipóteses

1. Modelo linear
2. A amostra aleatória
3. Os dados não são colineares
4. Exogeneidade de X_i
5. Homoscedasticidade
6. Erros são normalmente distribuídos.

Propriedades do Estimador de Mínimos Quadrados

Example

Utilizando o banco de dados de “Aula OLS”, disponibilizado na página da disciplina, encontre os estimadores dos dados apresentados.

- Determine os coeficientes da regressão.
- Verifique se os coeficientes são significantes;
- Verifique se o coeficiente β poderia ser considerado igual a 0 com 10% de significância;
- Verifique se o coeficiente β poderia ser considerado igual a 0.5 com 10% de significância;
- Verifique se o coeficiente β poderia ser considerado igual a 0.8 com 10% de significância;

Referências

Bussab, Wilton de O.; Morettin, Pedro A. (2014)
Estatística básica 8a ed.
Saraiva
Cap. 11

Meyer, Paul L. (1983)
Probabilidade: Aplicações à Estatística
Livros Técnicos e Científicos Editora
Cap. 14

Casella, George; Berger, Roger L. (2011)
Inferência estatística
Cengage Learning
Cap. 7

4 Aula 10 - Estimação Logit

- Revisão da estimativa linear
- Modelos de resposta binária
  - Modelo Logit
  - Donner Party

Revisão da estimativa linear

Neste ponto, cobrimos:

1. Estimadores pelo método dos momentos
2. Estimadores de máxima verosimilhança
3. Regressão linear simples

A principio sabemos como lidar com modelos de estimativa do tipo:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon \equiv X\beta + \epsilon \tag{1}$$

mas o que fazer quando a variável dependente é binária (1 ou 0)?

Nesta aula, vamos estudar quando a variável dependente é observada como uma variável binária e quando a variável dependente é categórica.

Modelos de resposta binária

Considere uma equação geral do tipo:

$$y_i^* = \beta_0 + \beta_1 x_{1,i} + \epsilon_i \equiv X_i \beta + \epsilon_i \tag{2}$$

O problema é que não observamos y_i. Em vez disso, observamos a variável binária:

$$y_i = \begin{cases} 1 \text{ if } y_i^* = \beta' x_i + \epsilon_i > c \\ 0 \text{ caso contrario} \end{cases} \tag{3}$$

Aqui, a probabilidade de $y_i = 1$ é igual à probabilidade de y_i^* ser maior que uma constante c, onde c é um limite. O limite é facilmente convertido em 0 ajustando o termo constante no modelo. Assim, para simplificar, reescrevemos a condição como

$$y_i = \begin{cases} 1 \text{ if } y_i^* = \beta' x_i + \epsilon_i > 0 \\ 0 \text{ caso contrario} \end{cases} \tag{4}$$

Modelos de resposta binária

Observe que a probabilidade de y_i observado ser um pode ser escrito usando y_i^*

$$\begin{aligned}\mathbb{P}\left(y_i = 1\right) &= \mathbb{P}\left(y_i^* > 0\right) \\ &= \mathbb{P}\left(\epsilon_i > -\beta' x_i\right)\end{aligned} \tag{5}$$

Assumimos que o termo de erro ϵ_i tem uma função de distribuição cumulativa de $F(\epsilon_i)$ e onde $f(\epsilon_i)$ é a função de densidade de probabilidade de $F(\epsilon_i)$. Então nós temos

$$\mathbb{P}\left(\epsilon_i > -\beta' X_i\right) = 1 - F(-\beta' X_i) \tag{6}$$

Se a distribuição for simétrica temos:

$$\mathbb{P}\left(\epsilon_i > -\beta' X_i\right) = F(\beta' X_i) \tag{7}$$

Sendo assim basta escolhermos uma distribuição simétrica para modelarmos a "probabilidade" de ocorrência da variável y_i.
Para isso vamos utilizar a distribuição Logística (contudo seria possível utilizar outra distribuição simétrica). Para a distribuição logística temos:

$$F\left(\beta' x\right) = \frac{e^{\beta' x}}{1 + e^{\beta' x}} \tag{8}$$

Ou seja, para um modelo de uma variável independente, temos o seguinte modelo econométrico:

$$p_i = \frac{e^{(\beta_0 + \beta_1 x_{1,i})}}{1 + e^{(\beta_0 + \beta_1 x_{1,i})}} \tag{9}$$

Exemplo - Donner Party

Em 1846, as famílias Donner e Reed deixaram Springfield, Illinois, para a Califórnia por caravana. Em julho, a *Donner Party*, como ficou conhecida, alcançou Fort Bridger, Wyoming. Lá, seus líderes decidiram tentar uma nova e rota para o Vale do Sacramento. Tendo atingido seu tamanho total de 87 pessoas e 20 vagões, o grupo se atrasou devido a uma difícil travessia do Cordilheira Wasatch e novamente na travessia do deserto a oeste do *Great Salt Lake*. O grupo ficou preso no leste das montanhas de Nevada quando a região foi atingida por fortes nevascas no final de outubro. Quando o último sobrevivente foi resgatado em 21 de abril de 1847, 40 dos 87 membros morreram de fome e exposição ao frio extremo.

Fonte: Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)

Exemplo - Donner Party

Objetivo da analise

Objetivo

- Qual é a relação entre sobrevivência e gênero?
- Qual é a probabilidade de sobrevivência em função da idade?
- Depois de levar em consideração a idade, as mulheres têm mais probabilidade de sobreviver a condições adversas do que os homens?
- A idade afeta a taxa de sobrevivência de homens e mulheres de forma diferente?

Exemplo - Donner Party

Dados

	Age	**Sex**	**Status**
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮			
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Exemplo - Donner Party

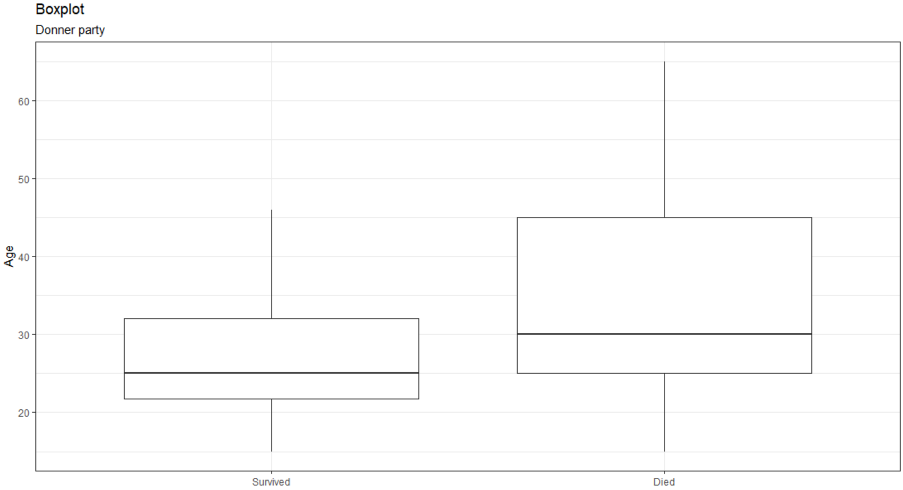
Diagramas de dispersão

Vamos analisar as distribuições marginais e os diagramas de dispersão.

	Female	**Male**
Died	5	20
Survived	10	10

Exemplo - Donner Party

Diagramas de dispersão



Exemplo - Donner Party

Dados

- Parece claro que a idade e o gênero afetam a probabilidade de sobreviver da pessoa, como podemos chegar a um modelo que nos permitirá explorar este relação?
- Uma maneira de pensar sobre o problema - podemos tratar Sobreviveu e Morreu como sucessos e os que não sobreviveram como fracassos.

Modelando a probabilidade de sobrevivência em função da idade

$$\mathbb{P}\left(y = 1|x_1\right) = \beta_0 + \beta_1 x_1 \tag{10}$$

- **Modelo de regressão linear:** Sobrevivência esperada do modelo dada a idade:

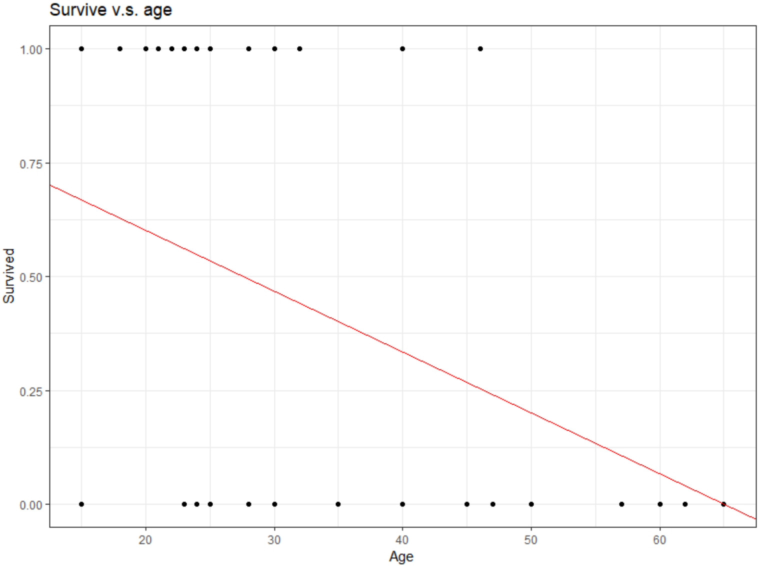
$$\mathbb{E}\left[y = 1|x_1\right] = \beta_0 + \beta_1 x_1 \tag{11}$$

Problemas:

- Não linearidade - um modelo linear pode fornecer valores previstos fora do intervalo (0, 1)
- Heteroscedasticidade - a variância $np(1-p)$ não é constante. Isso se deve ao fato de que y segue uma distribuição de **Bernoulli**.

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade



Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

```
Call:
lm(formula = Survived ~ Age, data = DonnerParty)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66886 -0.49521 -0.06775  0.45136  0.74524

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.869232   0.197106   4.410  6.8e-05 ***
Age         -0.013358   0.005777  -2.312   0.0256 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[...]
```

O modelo estimado é:

$$\mathbb{E}\left[y=1|x_1\right] = 0.8692 - 0.01336 \times x_1 \tag{12}$$

Considere prever a sobrevivência para uma pessoa de 70 anos:

$$\mathbb{E}\left[Survived=1|Age=70\right] = 0.8692 - 0.01336 \times 70 = -0,0658 \tag{13}$$

Este modelo prevê uma probabilidade negativa de sobrevivência.

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

- (Solução adequada): modelo de regressão logística

$$
\begin{aligned}
\pi_i &= \frac{\exp\left\{\beta_0 + \beta_1 X_{1,i}\right\}}{1 + \exp\left\{\beta_0 + \beta_1 X_{1,i}\right\} S} \\
\log \frac{\pi_i}{1 - \pi_i} &= \beta_0 + \beta_1 X_{1,i}
\end{aligned} \tag{14}
$$

onde $\pi_i = \mathbb{P}\left(Survived = 1\right)$ probabilidade de sobrevivência para a pessoa i, e $X_{1,i}$ é a idade de uma pessoa i.

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

```
Call:
glm(formula = Survived ~ Age, family = binomial("logit"), data = DonnerParty)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5401  -1.1594  -0.4651   1.0842   1.7283

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.81852    0.99937   1.820   0.0688 .
Age         -0.06647    0.03222  -2.063   0.0391 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[...]
```

O modelo estimado é

$$\log \frac{\pi_i}{1-\pi_i} = 1.8183 - 0.0665 X_{1i} \tag{15}$$

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

- A probabilidade de sobrevivência esperada de um indivíduo de 70 anos:

$$\hat{\pi} = \Pr(\text{ survival }) = \frac{\exp\{1.8183 - 0.0665 \times 70\}}{1 + \exp\{1.8183 - 0.0665 \times 70\}} = 0.055 \quad (16)$$

- o coeficiente β_1 é negativo, indicando que as chances de sobrevivência diminuem com a idade.

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

- Depois de levar em consideração a idade, as mulheres têm mais probabilidade de sobreviver a condições adversas do que os homens?
- Isso envolverá o ajuste no modelo, mas o modelo final fica:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{17}$$

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

```
Call:
glm(formula = survive ~ age + sex, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7445  -1.0441  -0.3029   0.8877   2.0472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.23041    1.38686   2.329   0.0198 *
age         -0.07820    0.03728  -2.097   0.0359 *
sex         -1.59729    0.75547  -2.114   0.0345 *
---
[...]
```

O modelo estimado é

$$\log \frac{\pi_i}{1-\pi_i} = 3.23041 - 0.07820 X_{1,i} - 1.59729 X_{2,i} \tag{18}$$

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

- As análises anterior assumem que o efeito do gênero na sobrevivência não depende da idade; ou seja, não há interação entre idade e sexo. Para testar a interação, vamos analisar o modelo:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} \qquad (19)$$

Exemplo - Donner Party

Modelando a probabilidade de sobrevivência em função da idade

```
Call:
glm(formula = survive ~ age * sex, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2279  -0.9388  -0.5550   0.7794   1.6998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.24638    3.20517   2.261   0.0238 *
age         -0.19407    0.08742  -2.220   0.0264 *
sex         -6.92805    3.39887  -2.038   0.0415 *
age:sex      0.16160    0.09426   1.714   0.0865 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[...]
```

O modelo estimado é

$$\log \frac{\pi_i}{1-\pi_i} = 7.24638 - 0.19407X_{1,i} - 6.92805X_{2,i} + 0.16160X_{3,i} \quad (20)$$

5 Aula 11 - PCA/Time Series