Notas de aula para o curso de Econometria I

Nota 1: Revisão de estatística

Thiago Fonseca Morello

fonseca.morello@ufabc.edu.br

sala 301, Bloco Delta, SBC

1 Expectativa e variância

1.1 Expectativa e suas propriedades

A expectativa de uma variável aleatória corresponde à média ponderada dos valores possíveis, em que cada valor possível é ponderado por sua probabilidade de ocorrência.

O número de valores que uma variável discreta pode assumir é finito. Sejam estes valores representados por $x_1, x_2, ..., X_N$. Desta maneira, pode-se ter que, por exemplo, x_1 é o menor valor que a variável pode assumir e X_N o maior, enquanto $x_2, ..., X_{N-1}$ são valores intermediários entre estes dois extremos. A expectativa de X, denotada por E[x], é tal que:

$$E[X] = x_1 P(x = x_1) + x_2 P(x = x_2) + \ldots + x_N P(x = x_N) = \sum_{i=1}^{N} P(X = x_i) x_i$$

Caso X seja contínua, como é o caso, por exemplo, de variáveis econômicas como renda anual, despesa anual em consumo, valor investido em novas tecnologias, etc, há um número infinito de valores possíveis. Denotando o domínio de variação (intervalo de valores possíveis) de X por D_X, a expectativa da VA é dada por:

$$E[X] = \int_{D_x} x f_X(x) dx$$

Em que $f_x(x)$ é a FD de X.

Duas propriedades úteis do operador expectativa, sejam X e Y discretas ou contínuas, são:

(Prop.Exp.1, linearidade) Sejam “a” e “b” duas constantes, não se tratando, portanto, de variáveis aleatórias, então tem-se que E[a+bX] = a+bE[X], o que decorre diretamente do princípio de que a expectativa de uma constante é a própria constante.

(Prop.Exp.2, soma de VAs) Sejam X e Y duas VAs, então se aplica a propriedade de que a expectativa da soma de duas VAs é equivalente à soma das expectativas das VAs. E[X + Y] = E[X] + E[Y]. Esta propriedade também se aplica à expectativa da soma de mais de duas VAs.

Outra propriedade diz respeito à expectativa de uma função da VA X.

(Prop.Exp.3, expectativa de uma função de uma VA) Seja g(X) uma função genérica de X. Então o valor esperado (da imagem) desta função E[g(X)] é dado por:

$$E[g(X)]=\int_{D_X}^{\square} g(x)f_X(x)dx\text{, se x for contínua}$$

$$E[g(X)]=\sum_{i=1}^{N} g(x_i)P(X=x_i)\text{, se x for discreta}$$

1.2 Expectativa condicional

1.2.1 Definição

Retomando o tratamento das distribuições conjuntas, sejam tomadas duas variáveis X e Y. A expectativa condicional de Y em relação a X corresponde à média dos valores que X pode assumir ponderada por probabilidades de ocorrência condicionais ao fato de que X assume valor x. Ou seja:

$$E[y \vee x]=y_1P(y=y_1 \vee X=x)+y_2P(y=y_2 \vee X=x)+\ldots+y_NP(y=y_N \vee X=x)=\sum_{i=1}^{N} y_iP(Y=y_i \vee X=x)\text{, com X e Y disc}$$

$$E[Y|X]=\int_{D_Y}^{\square} y f_{Y \vee X}(y)dy\text{, com X e Y contínuas}$$

A função $f_{Y|X}(y)$ é a FD condicional de Y.

A expectativa condicional de Y em relação a X é uma função de X, exclusivamente, ou seja, E[Y|X] = h(x), uma vez que se trata do valor médio de Y correspondente a um dado valor de X. O conceito de expectativa condicional é um dos principais em econometria. Ele será retomado com ênfase na parte II do curso.

1.2.2 Propriedades

Uma propriedade fundamental da expectativa condicional, muito útil em econometria é a lei das expectativas iteradas, a qual pode ser expressa como segue.

(Lei das expectativas iteradas) E[Y] = E[E[Y|X]]

(Expectativa condicional e independência) Sejam X e Y duas VAs independentes, neste caso, E[Y|X] = E[Y] e E[X|Y] = E[X]. A demonstração é fornecida abaixo.

$$E[Y|X]=\int_{D_Y}^{\square} y f_{Y \vee X}(y)dy=\int_{D_Y}^{\square} y\frac{f_{Y,X}(x,y)}{f_X(x)}dy=(\text{assumindo independência})\int_{D_Y}^{\square} y\frac{f_Y(y)f_X(x)}{f_X(x)}dy=\int_{D_Y}^{\square} y f_Y(y)dy=E[Y]$$

O passo crucial da demonstração está no fato de que $f_{Y \vee X}(y)=\frac{f_{Y,X}(x,y)}{f_X(x)}$, o qual, por sua vez, decorre de $f_{Y \vee X}(y)f_X(x)=f_{Y,X}(x,y)$, ou seja, P(A ∩ B)= P(A|B)P(B), conforme visto acima.

A demonstração para X e Y discretas segue o mesmo raciocínio.

Há uma propriedade importante da variância que decorre da covariância. Ela é discutida a seguir.

1.3 Variância

1.3.1 Definição

A variância é uma medida de dispersão, ou seja, ela mede o grau em que os valores possíveis para uma variável aleatória discrepam de um valor de referência, este último dado pelo valor esperado. Trata-se, pois, de uma medida de dispersão em torno da média, quanto maior for a magnitude desta medida, maior é o número de valores consideravelmente distantes do valor médio e/ou mais distantes do valor médio estão alguns dos valores – o que nos diz que o valor médio não é uma boa descrição sucinta dos valores possíveis.

Para compreender porque é desejável tomar em conta uma medida da fidedignidade com que a média resume os dados cabe considerar um exemplo. Seja assumido que, em um dado bairro do ABC, em um período em foi registrada alta taxa de ocorrência de febre em toda a região, a temperatura corporal de metade dos habitantes esteve em torno de 36,5 °C, enquanto que os demais registraram temperatura corporal em torno de 38 °C[1]. Neste caso, a temperatura corporal média dos habitantes, no período, foi de (1/2N36,5 + 1/2N38)/(1/2N+1/2N) = 37,25 °C. Caso a prefeitura se apoie nas temperaturas médias dos bairros para decidir quanto à alocação de seu estoque de medicamentos para tratamento de febre, obviamente será cometido um erro, pois deixarão de ser encaminhados medicamentos para o bairro em questão, mesmo sendo que metade de seus habitantes foi acometida por febre.

Há diversas maneiras de medir a distância em relação ao valor médio, mas a variância se define por tomar por base a média do quadrado da distância, ponderada pela probabilidade de ocorrência do valor associado. Trata-se da seguinte medida:

$$V[X]=(x_1-E[X])^2P(x=x_1)+(x_2-E[X])^2P(x=x_2)+\ldots+(x_N-E[X])^2P(x=x_N)=\sum_{i=1}^{N}(x_i-E[X])^2P(X=x_i)$$

Em termos genéricos, i.e., seja X discreta ou contínua, sua variância é dada por $E[(X-E[X])^2]$.

Caso X seja contínua, o correto é escrever:

$$V[X]=\int_{D_x}^{\square}(x-E[X])^2 f_X(x)\,dx$$

E também se aplica o conceito de variância condicional.

$$V[Y \vee X]=\sum_{i=1}^{N}(x_i-E[X])^2P(X=x_i \vee Y=y)$$

Uma propriedade útil da variância é a obtida da manipulação algébrica a seguir.

$V[X] = E[(X-E[X])^2] = E[X^2+ E[X]^2 - 2XE[X]] = E[X^2] + E[X]^2 - 2E[X]^2 = E[X^2] - E[X]^2 \leftrightarrow V[X] = E[X^2] - E[X]^2$.

Outra propriedade útil é enunciada a seguir.

(Prop.Exp.1, linearidade) Sejam “a” e “b” duas constantes, não se tratando, portanto, de variáveis aleatórias, então tem-se que $V[a+bX] = b^2V[X]$, o que decorre das propriedades da expectativa (a demonstração fica como exercício).

A raiz da variância, $\sqrt{V(X)}=\sqrt{E(X-E[X])^2}$, é denominada por “desvio padrão” e será denotada por DP[X].

[1] A temperatura corporal oscila normalmente entre 36,5°C e 37,5 °C. A febre é identificada com uma temperatura superior a 37,5 °C.

1.3.2 Propriedades

(Variância da soma) A variância da soma de duas variáveis, X e Y, não é equivalente à soma das variâncias, a menos que a covariância entre ambas seja nula. O que pode ser visto com base no raciocínio a seguir. V[X+Y] = E[(X+Y – E[X+Y])^2] = E[(X– E[X] +Y –E[Y])^2] = E[({X– E[X]} +{Y –E[Y]})^2]. Os termos dentro das chaves podem ser tratados como um único termo, basta definir A ≡ X – E[X], B ≡ Y- E[Y]. De modo que V[X+Y] = E[(A+B)^2] = E[A^2+B^2+2AB] = E[A^2]+[B^2]+ 2E[AB]. Agora, reincorporando a definição dos termos A e B à manipulação, tem-se:

V[X+Y] = E[(X– E[X])^2]+[(Y– E[Y])^2]+ 2E[(X– E[X])(Y– E[Y])] = V[X] + V[Y] + 2cov(X,Y). Conclusivamente, pois:

$$V[X+Y] = V[X] + V[Y] + 2cov(X,Y)$$

Se a covariância entre X e Y for nula, portanto, a variância da soma de X e Y é equivalente à soma das variâncias de X e Y.

1.4 Variância condicional

A variância de uma dada variável aleatória Y também pode ser obtida como função de outra variável aleatória X. Ou seja, trata-se de calcular a variância de Y, p.ex., remuneração, para cada nível de anos de estudo, esta última sendo a variável X. Formalmente trata-se de V[Y|X] = f(X). Ou seja:

$$\int_{D(X)}^{\square} (Y - E[Y \vee X])^2 f_{Y \vee X}(y)\, dy \text{, } \textit{para y contínua}$$

$$\sum_{i=1}^{N(x)} (y_i - E[Y|X])^2 P(Y = y_i) \text{, } \textit{para y discreta}$$

2 Estatísticas para a relação entre variáveis

2.1 Introdução

Como a existência de relação entre duas variáveis pode ser verificada a partir de um conjunto de dados? Há diversas técnicas para gerar evidências em tal sentido. Os gráficos e tabelas, por exemplo, permitem um exame visual, exploratório. Abaixo há uma tabela que capta a relação entre renda familiar mensal per capita e a proporção de crianças de zero a cinco anos de idade com altura inferior ao nível "saudável".

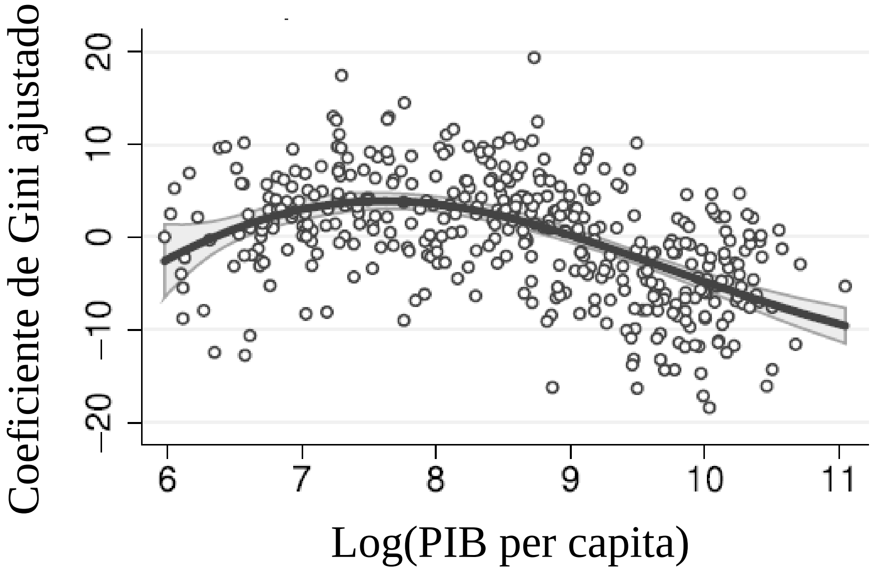
Classes de rendimento total e variação patrimonial mensal familiar per capita (salários mínimos*)	**Prevalência de déficit de altura nas crianças menores de 5 anos de idade (%)**
Até ¼	8,2
Mais de 1/4 a 1/2	6,8
Mais de 1/2 a 1	6,2
Mais de 1 a 2	5,2
Mais de 2 a 5	3,8

Mais de 5	3,1

*O salário mínimo vigente em 15 de Janeiro de 2009, data de referência da pesquisa, era de R$415,00.

Fonte: IBGE, Pesquisa de Orçamentos Familiares (POF). Antropometria, estado nutricional de crianças, adolescentes e adultos no Brasil. IBGE. Disponível em http://www.ibge.gov.br/home/estatistica/populacao/condicaodevida/pof/2008_2009_encaa/pof_20082009_encaa.pdf

Já o gráfico a seguir relaciona, para 113 países do mundo, PIB per capita, uma medida de nível de desenvolvimento econômico, e o coeficiente de Gini, uma medida de desigualdade de renda pessoal. Esta relação foi estudada pioneiramente por Simon Kuznets.



Fonte: Desbordes & Verardi, 2012, Refitting the Kuznets Curve, Economic Letters, 116. Disponível em http://www.sciencedirect.com/science/article/pii/S0165176512000973

2.2 Covariância

2.2.1 Definição

Uma evidência visual é pouco informativa, ou seja, ela representa a extração de baixa proporção do conteúdo informacional dos dados. É possível retirar mais informação dos dados, gerando evidências mais precisas. Uma maneira de fazer isso é obtendo uma medida para o grau em que as duas variáveis estudadas se relacionam. A covariância é um operador matemático que proporciona uma tal medida. Como o próprio nome sugere, trata-se de uma medida para como duas variáveis co-variam, se movimentam em uma mesma direção ou em direções opostas. Mais precisamente, a covariância mede o grau de dependência linear existente entre duas variáveis.

A covariância entre duas variáveis X e Y é dada por cov(X,Y) = E([X-E(X)][Y-E(Y)]). Esta fórmula é uma maneira interessante de quantificar a relação entre duas variáveis por dois motivos:

1. Se X e Y variam na mesma direção, i.e., valores relativamente altos (relativamente baixos) de X correspondem a valores relativamente altos (relativamente baixos) de Y, então a covariância é positiva; caso contrário, i.e., se valores relativamente altos (relativamente baixos) de X correspondem a valores relativamente baixos (relativamente altos) de Y, então a covariância é negativa;
2. Quanto mais relacionados forem X e Y, seja positiva ou negativamente, maior o valor absoluto da covariância.

2.2.2 Propriedades da covariância

Cabe retomar algumas propriedades da covariância.

(Independência implica covariância nula) Se X e Y são independentes ✉ cov(X,Y) = 0. A primeira parte da demonstração corresponde a uma manipulação algébrica similar àquela desenvolvida na decomposição da variância.

Cov(X,Y) = E[X - E[X]][Y – E[Y]] = E[XY – XE[Y] – YE[X] + E[X]E[Y]] = E[XY] – 2E[X]E[Y] + E[X]E[Y] = E[XY] –E[X]E[Y] (a).

Como segundo passo é preciso demonstrar que, se X e Y são independentes, E[XY] = E[X]E[Y]. Para isso, basta considerar as passagens a seguir (Casella e Berger[2], p.144, teorema 4.2.1):

$$E[XY]=\int_{D_Y}^{\square}\int_{D_X}^{\square} XY\, f_{X,Y}(x,y)\, dxdy\, (b)$$

Esta primeira passagem decorre da definição de expectativa condicional para a função de variáveis aleatórias g(X,Y) = XY (Casella e Berger: p.171).

$$\int_{D_Y}^{\square}\int_{D_X}^{\square} XY\, f_{X,Y}(x,y)\, dxdy=\int_{D_Y}^{\square}\int_{D_X}^{\square} XY\, f_X(x) f_Y(y)\, dxdy\, (c)$$

A parte (c) decorre da definição de independência, segunda a qual, se X e Y são independentes, então P(X= x, Y= y) = P(X= x)P(Y=y), o que é equivalente a $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. As demais passagens se resumem à percepção de que $f_X(x)$ e X são funções apenas de X, não variando com Y e um argumento análogo se aplica a $f_Y(y)$ e Y.

$$\int_{D_Y}^{\square}\int_{D_X}^{\square} xy\, f_X(x) f_Y(y)\, dxdy=\int_{D_Y}^{\square} y\, f_Y(y)\left(\int_{D_X}^{\square} x\, f_X(x)\, d\,x\right) dy=\left(\int_{D_X}^{\square} x\, f_X(x)\, dx\right)\left(\int_{D_Y}^{\square} y\, f_Y(y)\, dy\right)=E[X]E[Y](d)$$

Conectando (a) e (c), chega-se a cov(X,Y) = E[XY] – E[X]E[Y] = E[X]E[Y] – E[X]E[Y] = 0, sendo X e Y independentes. A demonstração para variáveis discretas segue o mesmo raciocínio (Casella e Berger, p.144, teorema 4.2.1).

2.2.3 Coeficiente de correlação

Há um inconveniente na fórmula da covariância que é o fato de que ela assume valores dentro de um intervalo da reta real (ou domínio) determinado pelos intervalos da reta real dentro dos quais X e Y variam. Desta maneira, acaba-se com uma medida para o grau de relação entre X e Y que depende dos valores observados para X e Y. Para entender porque se trata de algo inconveniente, considere o seguinte exemplo. Deseja-se determinar o grau em que o balanço comercial do Brasil está correlacionado com o balanço comercial da China. Caso o valor nominal dos dois balanços for expresso em Reais, a magnitude da covariância será diferente da obtida quando os dois balanços são expressos em Yuans. Isto é inconveniente pois a moeda em que os balanços são expressos é uma mera convenção dimensional, a qual não tem qualquer implicação sobre o padrão descrito pelos dois balanços comerciais.

Para eliminar este inconveniente, emprega-se a correlação, esta dada pela razão entre a covariância de X e Y e o produto dos desvios-padrão de cada variável, ou seja, trata-se de:

[2] Casella, G., Berger, R.L., Statistical inference. Pacific Grove, USA: Duxbury: Thomson Learning, 2002.

$$corr(X,Y)=\frac{E[X-E(X)][Y-E(Y)]}{\sqrt{E\left([X-E(X)]^2\right)}\sqrt{E\left([Y-E(Y)]^2\right)}}$$

Esta medida tem seu domínio de variação restrito ao intervalo [-1;1] da reta real. A existência de correlação linear perfeita e positiva corresponde a corr(X,Y) = 1 e de correlação linear perfeita e negativa, corr(X,Y) = - 1.

3 Estimadores e métodos de estimação[3]

Cabe relembrar as definições de “estatística”, “estimador” e “estimativa”. Por “estatística” se entende qualquer função dos dados disponíveis que não seja função de parâmetros desconhecidos, especialmente daqueles que definem a função de distribuição de probabilidades associada à variável aleatória (p.ex., média e variância no caso da distribuição normal). Uma estatística é, pois, geralmente uma operação matemática que resume os dados. Os exemplos mais comuns são a média, a variância, o valor mínimo, o valor máximo e os quartis (entre eles a mediana).

O termo “estimador” designa uma estatística (função dos dados, portanto) que assume valores dentro do domínio de variação de um parâmetro que define a FD conjunta. Os estimadores são utilizados para fazer afirmações acerca dos valores populacionais de medidas relevantes para a análise empírica. Por exemplo, as proporções de eleitores com intenção de voto nos candidatos presidenciais, a elasticidade-preço da demanda por alimentos e a renda média dos brasileiros.

A média amostral, por exemplo, é um estimador para a média populacional. Em primeiro lugar, trata-se de uma estatística, pois sua fórmula contém apenas valores das variáveis observadas. Em segundo lugar, a média amostral tende a assumir valores dentro do domínio de variação da média populacional. É simples entender por que. Se a variável aleatória está definida no intervalo [a,b], então a média populacional está obrigatoriamente definida nesse mesmo intervalo, pois a média é uma operação matemática cujo resultado nunca pode ser menor do que o limite interior dos valores incluídos no cálculo e nem maior do que o limite superior de tais valores. Ou seja, a média populacional está definida no intervalo [a,b] e um estimador para a média populacional, pois, tem obrigatoriamente de estar definida no intervalo [a,b]. Esse é o caso da média amostral, pelo mesmo argumento utilizado para explicar por que a média populacional pertence ao intervalo [a,b].

“Estimativa” se refere ao valor assumido pelo estimador para um conjunto de dados em específico, em uma amostra. O valor da média amostral, por exemplo, varia de acordo com a amostra disponível. Calculando-se o valor médio da renda das famílias brasileiras, este assumirá um valor, com base na amostra da PNAD 2009, possivelmente distinto do valor a que se pode chegar a partir da amostra da PNAD 2010.

Há diversos métodos para encontrar estimadores para um determinado parâmetro. Os três principais são: (i) método de máxima verossimilhança, (ii) método dos momentos e (iii) método de mínimos quadrados ordinários.

4 Propriedades de estimadores

Uma vez que há múltiplos métodos para obter estimadores, e, consequentemente, para estimar um parâmetro populacional, mais de um estimador pode estar disponível. É preciso, pois, ter critérios para estabelecer qual é a melhor opção. Ou seja, é preciso ter indicadores da qualidade, ou melhor, da confiabilidade, dos estimadores, no que tange à geração de estimativas para o parâmetro-alvo.

[3] As definições aqui apresentadas foram retiradas de Bolfarine, Heleno e Sandoval, Monica C., 2001,“Introdução à inferência estatística”. Coleção Matemática Aplicada, Sociedade Brasileira de Matemática.

Há três propriedades cuja verificação é um indicativo de que o estimador é confiável: ausência de viés, eficiência e consistência. Antes de passar às definições é preciso esclarecer que tais propriedades não dizem respeito ao valor assumido pelo estimador na amostra disponível. Mas sim a tendências ou padrões que seriam - atenção para o uso do subjuntivo – observados caso o estimador fosse calculado repetidas vezes em todas as amostras que podem ser retiradas da população dentro do mesmo período de tempo[4]. Por esta razão, as definições das propriedades fazem referência a experimentos mentais. Além disso, por simplicidade, será assumido que, com base em uma amostra para apenas uma característica, $\{X_1,...,X_N\}$, se deseja estimar um parâmetro populacional genérico, θ, recorrendo-se a um estimador igualmente genérico, $\hat{\theta} = f(X_1,...,X_N)$, em que f(.) é uma função qualquer dos dados disponíveis.

A primeira propriedade toma por base a realização de um experimento mental que consiste em obter o valor do estimador, ou estimativa, para cada uma de todas as amostras aleatórias possíveis da população (todas as combinações possíveis de observações) de tamanho N finito (qualquer que seja $1 \leq N < \infty$ e $N < \#\Omega$). Acaba-se, portanto, com a distribuição completa do estimador, ou seja, sua distribuição populacional. Neste caso, pois, o conceito de expectativa se aplica e pode-se tomar $E[\hat{\theta}_N]$. O subscrito "N" vai passar a ser utilizado a partir de agora para indicar o tamanho da amostra em que o estimador é calculado. Esta expectativa tem de ser entendida como a média populacional do estimador. O viés do estimador, $B(\hat{\theta}_N)$, é dado pela diferença entre o valor populacional que se deseja estimar e a expectativa do estimador, i.e., $B(\hat{\theta}_N) = \theta - E[\hat{\theta}_N]$.

O que ocorre se o estimador não atender à propriedade de ausência de viés, i.e., se $B(\hat{\theta}_N) = \theta - E[\hat{\theta}_N] \neq 0$? Basta perceber que, no limite, tomando-se todas as amostras possíveis da população, o conteúdo informacional com base no qual se calcula a média das estimativas geradas para cada amostra, é a população em si, completa. Se, com base nesta, o estimador gera uma estimativa que difere do valor populacional, há algum problema em sua fórmula, i.e., ele é intrinsicamente mal especificado, uma vez que o motivo do erro em que incorre não é a falta de informação. O que dá origem a uma interpretação pertinente do conceito de ausência de viés: um estimador não apresenta viés se ele erra ao estimar o parâmetro populacional única e exclusivamente por tomar por base, para isso, um conteúdo informacional insuficiente.

A propriedade de eficiência requer que a variância das estimativas calculadas a partir de todas as amostras disponíveis – segundo o experimento mental descrito no parágrafo anterior - seja a menor possível entre todos os estimadores alternativos. O estimador que atender a este critério é dito eficiente, ou, de maneira mais precisa, mais eficiente (do que os demais).

A última propriedade, de consistência, é estabelecida com base em um experimento mental em que também são retiradas amostras de um número suficientemente grande da população, mas, porém, o tamanho das amostras varia, tornando-se arbitrariamente grande. Mais precisamente, retiram-se todas as amostras possíveis da população para um determinado tamanho amostral N = 1, por exemplo, então, em um segundo passo, retira-se todas as amostras de tamanho N = 2, e assim sucessivamente com N tendendo ao infinito. As propriedades de estimadores estabelecidas com base neste experimento são denominadas assintóticas e também são referidas como "propriedades em amostras grandes" (intuição retirada de Wooldrigde[5]). A propriedade de consistência estabelece que o valor esperado do estimador se torna progressivamente mais próximo do valor efetivo do parâmetro com o aumento da amostra. De fato, para uma amostra arbitrariamente grande, o quadrado da diferença torna-se nulo. Formalmente, a propriedade de consistência pode ser exprimida de três maneiras.

[4] Não se trata, portanto, de calcular a renda média das famílias com a PNAD 2008, depois com a PNAD 2009, 2010, e assim por diante, mas sim calcular a renda para todas as combinações de brasileiros que compuseram a população em 2008.
[5] 3°edição em inglês do livro, p.178, cap.5.

$$\lim_{N \to \infty} P\left(\hat{\theta}_N - \theta \vee \varepsilon\right) = 0 \text{, } para\ \varepsilon\ muito\ pequeno\,(i)$$

$$\lim_{N \to \infty} E\left[\left(\hat{\theta}_N - \theta\right)^2\right] = 0\,(ii)$$

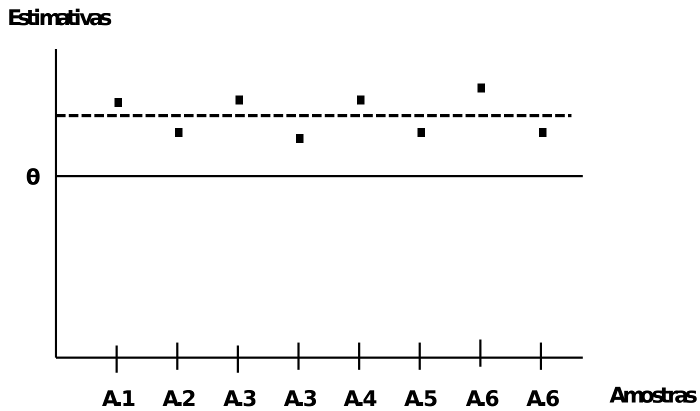
$$\lim_{N \to \infty} E\left[\hat{\theta}_N\right] = \theta\ e \lim_{N \to \infty} V\left[\hat{\theta}_N\right] = 0\,(iii)$$

A primeira notação faz referência ao conceito de convergência em probabilidade. Ela afirma que a probabilidade da diferença entre a estimativa pontual e o valor populacional do parâmetro ser não desprezível tende a zero com o aumento ilimitado da amostra. As duas últimas notações são equivalentes. Elas estabelecem a convergência em média quadrática, a qual pode ser exprimida da seguinte maneira: com o aumento da amostra, o valor esperado do estimador se torna progressivamente mais próximo do valor populacional do parâmetro e a variância do estimador se torna nula. Para entender porque esta última condição é necessária basta ter em conta que o valor populacional do parâmetro é fixo, não varia jamais, de modo que afirmar que as estimativas pontuais se tornam cada vez mais próximas de tal valor é afirmar que elas se tornam progressivamente fixas, ou seja, variam cada vez menos. Colocando de outra maneira, produzir valores cada vez mais próximos de um valor-alvo fixo é produzir valores que discrepam cada vez menos deste valor-alvo e, portanto, discrepam cada vez menos entre si.

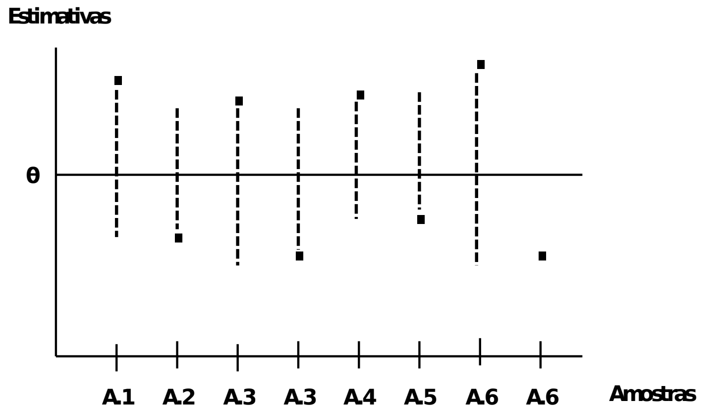
As duas últimas propriedades garantem que os valores gerados pelo estimador para diferentes amostras da mesma população tenham uma volatilidade, medida pela variância, não muito grande, aceitável. Um estimador que gera valores muito discrepantes comparando amostras diferentes não é confiável, porque leva a conclusões acerca do valor populacional que mudam em magnitude considerável com a amostra. Neste caso, as conclusões da análise de dados, portanto, dependem demais da amostra disponível, valendo apenas para "dentro" dela, o que é equivalente a dizer que não se poder retirar nenhuma conclusão em relação à população. Além disso, uma vez que o valor populacional do parâmetro-alvo é fixo, uma alta volatilidade das estimativas significa que a probabilidade destas distanciarem-se do valor populacional que se procura inferir é relevante.

Os gráficos abaixo mostram como os efeitos do viés e da alta volatilidade sobre a qualidade das estimativas se diferenciam.

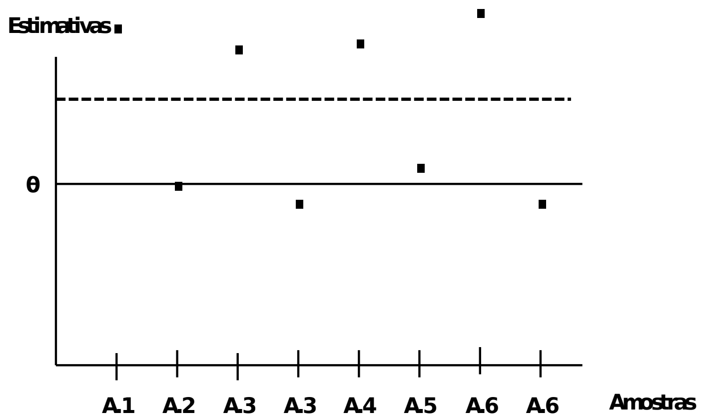
Estimação viesada com baixa volatilidade



Estimação não viesada com alta volatilidade



Estimação viesada com alta volatilidade



5 Teste de hipóteses

Um teste de hipóteses se refere geralmente a duas hipóteses mutuamente excludentes, genericamente:

Hipótese nula (H0): $\theta = \theta_0$

Hipótese alternativa (H0): $\theta \neq \theta_0$

O teste consiste em verificar, com base na evidência disponível, geralmente uma amostra de dados, qual das duas hipóteses é verdadeira. Há dois tipos de erros que podem ser cometidos ao realizar o teste.

- Erro tipo I: rejeitar a hipótese nula quando tal hipótese é verdadeira;
- Erro tipo II: não rejeitar a hipótese nula quanto tal hipótese é falsa.

Tabela 1 Dois tipos de erros em um teste de hipóteses

	Decisão	
Hipótese verdadeira	**Rejeitar H0**	**Não rejeitar H0**
H0	Erro tipo 1	Decisão correta
H1 (H0 é falsa)	Decisão correta	Erro tipo 2

Fonte: Casella & Berger, p.359.

Um teste de hipóteses deve ser estruturado de modo a minimizar as probabilidades dos dois tipos de erro. Tal minimização é claramente observável a partir do procedimento de teste para o erro tipo I, por isso será resumida a seguir. Já a minimização do erro tipo II requer detalhes acerca das propriedades de testes de hipóteses que fogem ao escopo dessa recordação.

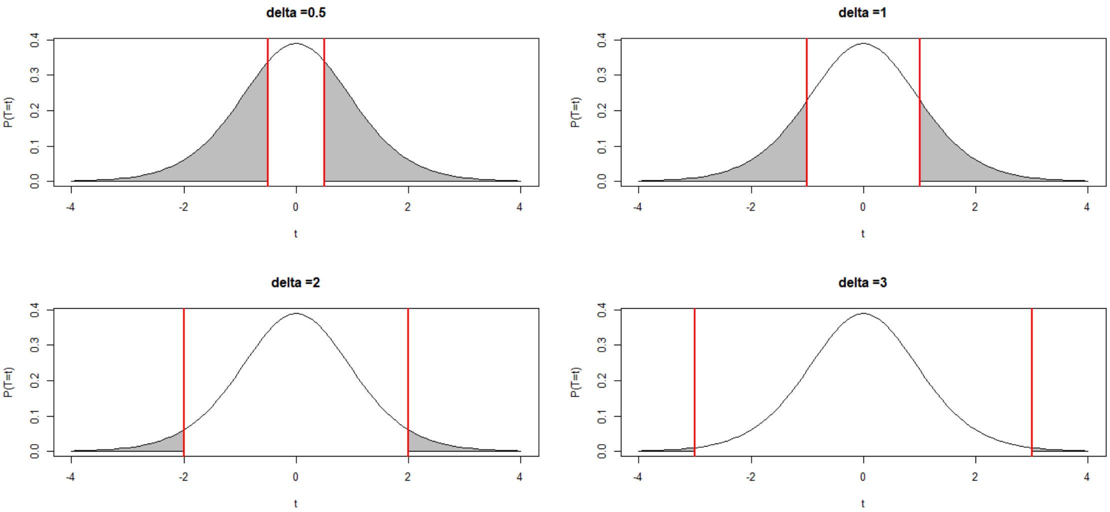
A probabilidade do erro tipo I pode ser escrita como P(rejeitar H0|H0 é verdadeira). O que nos levaria a rejeitar H0? Cabe tomar como exemplo um teste específico, no caso para o coeficiente de correlação populacional. As hipóteses do teste para a nulidade da correlação são:

$$H0: \rho = 0 \text{ vs } H1: \rho \neq 0$$

Recordando que $-1 < \rho < 1$. Se a hipótese em questão estabelece a nulidade da correlação populacional, qual teria de ser o valor da correlação amostral para que ela fosse uma evidência desfavorável à hipótese? [pense antes de continuar a leitura]

A resposta é notória: grandemente positiva ou grandemente negativa. Seja considerado que ε é um limiar de referência, de maneira que números que o ultrapassam em valor absoluto são considerados grandes. Com isso, a probabilidade do erro tipo I para o teste para a nulidade da correlação é P($\hat{\rho}$ ∨ > ε | ρ = 0) = P($\hat{\rho}$ < - ε ou $f\left(\hat{\rho}\right)$ > ε | ρ = 0). O objetivo é minimizar tal probabilidade.

A rigor, como a função de distribuição de probabilidades do coeficiente de correlação amostral é desconhecida. É, pois, preciso aplicar uma transformação matemática a tal estimador, $f\left(\hat{\rho}\right)$. Com isso, a probabilidade do erro de tipo I passa a P($f\left(\hat{\rho}\right)$ ∨ > δ | ρ = 0) = P(f($\hat{\rho}$ < -δ ou $f\left(\hat{\rho}\right)$ > δ | ρ = 0), em que δ = f(ε). Para simplificar a notação, $f\left(\hat{\rho}\right)$ será denotada como “t”. Com isso, a probabilidade a ser minimizada é P(t < -δ ou t > δ | ρ = 0). Tomando por base uma função de distribuição de probabilidade simétrica em torno do valor do parâmetro estabelecido pela hipótese nula – o que de fato se aplica ao teste em questão - é possível perceber que, quanto maior for o valor de δ, menor será tal probabilidade. Isso está indicado na figura abaixo.



Uma vez que uma probabilidade de erro de 5% é considerada tolerável, basta tomar δ tal que P(t < -δ ou t > δ | ρ = 0) = 0.05. Tal valor de δ, ou melhor, os limiares negativo e positivo, -δ e δ a partir dele obtidos, definem a região de rejeição da hipótese nula, ou região crítica do teste. Esta é dada por RC[ρ;0.95] = {-∞, -δ} U {δ, ∞}. Com isso, pois, a probabilidade do erro de tipo I é minimizada ao selecionar-se uma região crítica que ocorre com probabilidade desprezível sendo verdadeira a hipótese nula. Observar um valor para a estatística do teste pertencente à região crítica, significa observar algo que, segundo a hipótese nula, é improvável. Porém, como tudo que ocorre é logicamente provável, tem-se, pois, evidência de que a hipótese nula está equivocada.

6 Procedimentos de para realização de testes de hipóteses 1: teste para o coeficiente de correlação populacional

<u>Passo 1: definição das hipóteses:</u>

O parâmetro-alvo do teste é o coeficiente de correlação populacional linear de Pearson, e representado por ρ, cuja fórmula é:

$$\rho = \frac{cov(X,Y)}{\sqrt{V[X]V[Y]}} = \frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{V[X]V[Y]}}$$

<u>Passo 2: Estatística do teste, distribuição e valor observado</u>

Assume-se (i) AAS $\{(X_1,Y_1),(X_2,Y_2)\ldots,(X_N,Y_N)\}$ e (ii) que a distribuição conjunta é a normal bidimensional, i.e., $(X,Y) \sim N\left(\begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix};\begin{bmatrix}\sigma_X^2 & \rho \\ \rho & \sigma_Y^2\end{bmatrix}\right)$, em que $\mu_X = E[X]$, $\mu_Y = E[Y]$, $\sigma^2_X = V[X]$, $\sigma^2_Y = V[Y]$ (ver Morettin e Bussab, 2010, 8.8, para a fórmula matemática da distribuição). A matriz $\begin{bmatrix}\sigma_X^2 & \rho \\ \rho & \sigma_Y^2\end{bmatrix}$ é denominada por matriz de variância e covariância, e será mencionada frequentemente na disciplina.

2.a) Estatística e função de distribuição (FD)

O coeficiente de correlação amostral, $\hat{\rho}$, é o estimador com propriedades desejáveis para a correlação populacional e sua fórmula é a seguinte:

$$\hat{\rho}=\frac{\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2\right]\left[\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2\right]}}$$

A estatística de um teste é geralmente uma transformação do estimador adequado, e, no caso, tal transformação é a que segue:

$$\hat{T}=\hat{\rho}\sqrt{\frac{N-2}{1-\left(\hat{\rho}\right)^2}}{}^a t\left(N-2\right) sob\, H_0$$

A transformação, pois, tem distribuição assintoticamente equivalente a uma t de Student com N-2 graus de liberdade desde que a hipótese nula seja verdadeira.

2.b) Com base na amostra obtida, deve-se calcular o valor observado da estatística, $\hat{T}$.

Como exemplo, seja considerada uma base de dados que será objeto do primeiro laboratório, referente ao estado da pandemia de COVID-19 no mundo em Julho de 2020. Tomando as variáveis (i) casos (detectados) de infecções por um milhão de habitantes (X) e (ii) óbitos devidos ao coronavírus por um milhão de habitantes (Y), o coeficiente de correção foi de aproximadamente 42%. O valor observado da estatística é:

$$\hat{T}=0.42\sqrt{\frac{N-2}{1-\left(0.42\right)^2}}6.14$$

<u>Passo 3: região crítica e p-valor</u>

<u>3.a) Primeira abordagem para realização do teste de hipóteses: região crítica</u>

A região crítica é dada por RC[ρ;0.95] = {-∞;-t_α} U {t_α; ∞}, tal que P($T \leq -t_\alpha$ ou $T \geq t_\alpha$) = α, sendo T uma variável aleatória com distribuição t de Student com N-2 graus de liberdade, i.e., T ~ t(N-2). Com isso, α é a probabilidade do erro tipo 1.

No exemplo dos dados de COVID-19, o tamanho amostral é N = 178, havendo 176 graus de liberdade. Para determinar qual número possui 95% de probabilidade entre seu valor positivo e negativo, na distribuição t de Student com 176 graus de liberdade, basta utilizar uma tabela de distribuição, como as encontradas no livro de Morettin e Bussab, ou um software estatístico. Seguindo o segundo caminho, opta-se, nessa disciplina, pelo software gratuito R. Neste, o comando que informa o número procurado é qt(α/2,N-2) = qt(0.025,176) = -1.973534. Com isso a região crítica é {-∞;-1.97} U {1.97; ∞}.

<u>3.b) Segunda abordagem para a realização do teste de hipóteses: p-valor</u>

(a) Calcular o valor observado da estatística $\hat{T}=\hat{\rho}\sqrt{\frac{N-2}{1-\left(\hat{\rho}\right)^2}}$;

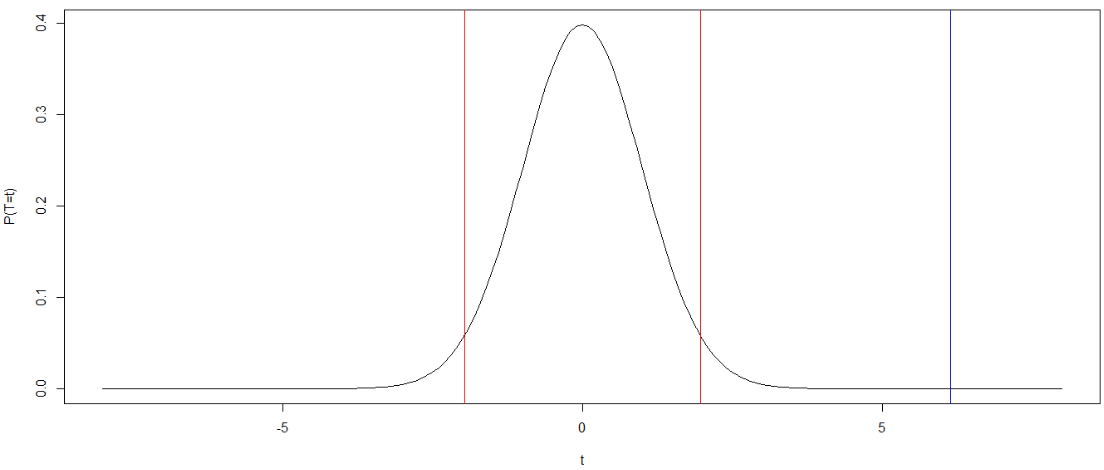
(b) Calcular a probabilidade de obter um valor mais extremo ao observado, ou seja, 2.P(T< $-\widehat{T \vee}$), sendo T ~ t(N-2). Tal probabilidade é denominada por p-valor. Para calculá-la é preciso consultar a FD da estatística do teste. No software R, basta utilizar o comando 2.pt(-|$\hat{T} \vee$, N-2) = pt(-|6.14|, N-2) = 5.425474e-09 = 5.425474 x 10^{-9}

Passo 4: tomada de decisão

4.a) Região crítica:

O critério de decisão é: se $\hat{T} \in RC$, rejeitar H_0, caso contrário, não rejeitar H_0;

Tem-se, no gráfico abaixo, a função de distribuição t de Student para o exemplo dos dados de COVID-19. Os valores críticos estão representados pelas linhas vermelhas e o valor observado pela linha azul. Tem-se, pois, evidência favorável à rejeição de H0.



4.b) P-valor:

O critério de decisão é: se p-valor < α, rejeitar H_0, caso contrário, não rejeitar H_0; em que α é o nível de significância.

Para o exemplo de COVID-19, p-valor = 5.425474 x 10^{-9} < 0.05. Confirma-se, pois, a decisão tomada com base na abordagem de região crítica.