

Probabilidades e Estatística

Conteúdos de Estatística

2019

Nota introdutória

Este é um documento de apoio ao estudo da componente de Estatística para as diversas U.C.'s de Probabilidades e Estatística, servidas pelo Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (DM/FCT/UNL).

O documento tem por base apontamentos elaborados pela Professora Doutora Fátima Miguens nos anos lectivos 2007/2008. Foi revisto pelo Professor Doutor Rui Cardoso nos anos de 2008 a 2010 e objecto das necessárias correções. Entretanto tem vindo a ser actualizado e até a ser revisto pelos Professores Doutores Ayana Furtado Mateus e Gracinda Rita Guerreiro nos anos de 2011 a 2015. A versão que agora se divulga (2019) apresenta reformulações e novidades nos conteúdos agora expostos. A todos os outros Professores Doutores que venham a contribuir para uma edição final, agradeço a sua colaboração e interesse.

Todos os Professores Doutores mencionados são docentes do DM/FCT/UNL.

A leitura deste texto não dispensa a leitura atenta das obras indicadas como referências bibliográficas nas U.C.'s a que se destina.

Conteúdo

1	Inferência Estatística	6
1.1	Introdução	6
1.2	População e amostra aleatória	6
1.2.1	População	6
1.2.2	Amostra	7
2	Estimação Pontual	10
2.1	Estatísticas	10
2.2	Exemplo de estatísticas	11
2.3	Estimação pontual do valor médio $\mu = E(X)$ da população X	12
2.4	Métodos para determinação de estimadores	14
2.4.1	Método dos momentos	14
2.4.2	Método da máxima verossimilhança	15
2.5	Propriedades dos estimadores	20
2.5.1	Distribuição de amostragem de um estimador	20
2.5.2	Enviesamento	21
2.5.3	Eficiência e erro quadrático médio	22
2.5.4	Consistência	23
2.5.5	Propriedades de \bar{X} , S^2 e \hat{P}	25
3	Estimação por Intervalo de Confiança	26
3.1	Introdução	26
3.1.1	Intervalo de confiança $(1 - \alpha)$	26
3.1.2	Método Pivotal	27
3.2	Estimação por intervalo de confiança do valor médio $\mu = E(X)$ da população X	30
3.3	Estimação por intervalo de confiança da variância $\sigma^2 = V(X)$ e do desvio padrão $\sigma = \sigma(X)$, da população X	36
3.4	Estimação por intervalo de confiança da proporção p de observação do acontecimento A	39
3.5	Distribuições de amostragem	42
3.5.1	Média amostral, \bar{X}	42
3.5.2	Variância amostral, S^2	44
3.5.3	Proporção amostral, \hat{P}	44
3.5.4	Diferença de médias de amostras de duas populações, $\bar{X} - \bar{Y}$	45
3.5.5	Quociente de variâncias de amostras de duas populações, S_1^2/S_2^2	45
3.5.6	Diferença de proporções amostrais de duas populações, $\hat{P}_X - \hat{P}_Y$	46

4	Teste de Hipóteses	47
4.1	Introdução	47
4.2	Decisão, regra de decisão e estatística de teste	48
4.3	Erros de decisão e sua probabilidade	49
4.4	Metodologia para realização de um teste de hipóteses paramétricas	51
4.5	p -value ou valor- p	51
4.6	Erros de decisão e sua probabilidade	52
4.7	Teste de hipóteses para o valor médio	52
4.7.1	Teste de hipóteses bilateral para o valor médio	52
4.7.2	Teste de hipóteses unilateral direito para o valor médio	59
4.7.3	Teste de hipóteses unilateral esquerdo para o valor médio	62
4.8	Teste de hipóteses para a variância	65
4.8.1	Teste de hipóteses bilateral para a variância	66
4.8.2	Teste de hipóteses unilateral direito para a variância	67
4.8.3	Teste de hipóteses unilateral esquerdo para a variância	68
4.9	Outros testes de hipóteses	71
4.9.1	Teste de hipóteses para a proporção	71
4.9.2	Teste de hipóteses para a diferença dos valores médios de duas populações	72
5	Teste de ajustamento do Qui-Quadrado	76
5.1	Teste ao pressuposto da normalidade de uma população	79
6	Teste ao pressuposto de aleatoriedade das observações amostrais	84
7	Regressão Linear Simples	88
7.1	Relação entre variáveis	88
7.2	Modelo de regressão linear simples	89
7.3	Método dos mínimos quadrados para estimar β_0 e β_1	90
7.4	Estimação da variância do erro σ^2 e qualidade do ajustamento	92
7.4.1	Estimador para σ^2	93
7.4.2	Qualidade do ajustamento	93
7.5	Distribuição de amostragem dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$	94
7.5.1	Distribuição de amostragem de $\hat{\beta}_1$	94
7.5.2	Distribuição de amostragem de $\hat{\beta}_0$	96
7.6	Inferência sobre os parâmetros do modelo	97
7.6.1	Inferência sobre β_1	97
7.6.2	Inferência sobre β_0	98
7.6.3	Inferência sobre σ^2	99
7.7	Estimação do valor esperado de Y para uma observação x_0 da variável controlada	100
7.8	Previsão do valor da variável resposta Y para um novo valor de x_0 da variável controlada	101
7.9	Alguns modelos linearizáveis	104
7.9.1	O modelo Power e a relação log-log	104
7.9.2	O modelo "Exponencial" e a relação log-lin	104

Lista de Tabelas

2.1	Tabela de estimadores para o valor médio, variância, desvio padrão e proporção	25
3.1	Distribuição de amostragem da média amostral, \bar{X}	43
3.2	Distribuição de amostragem da variância amostral, S^2	44
3.3	Distribuição de amostragem da proporção amostral, \hat{P}	44
3.4	Distribuição de amostragem para a diferença de médias amostrais de duas populações	45
3.5	Distribuição de amostragem para o quociente de variâncias amostrais de duas populações	45
3.6	Distribuição de amostragem para a diferença de proporções amostrais de duas populações	46
4.1	Decisões e erros num teste de hipóteses	49
4.2	Testes de hipóteses bilateral para o valor médio	57
4.3	Testes de hipóteses unilateral direito para o valor médio	62
4.4	Testes de hipóteses unilateral esquerdo para o valor médio	64
4.5	Testes de hipóteses para a variância	70
4.6	Testes para a proporção, p	71
4.7	Testes de hipóteses para a diferença de dois valores médios	72

Lista de Figuras

1.1	Função de probabilidade da população e da amostra	8
3.1	Intervalos de confiança para o valor médio: Situações A, C e D	31
3.2	Intervalos de confiança para o valor médio: Situações A, C e D	32
3.3	Intervalos de confiança para o valor médio: Situação C	34
3.4	Intervalos de confiança para o valor médio: Situações A, B e D	35
3.5	Intervalo de confiança para a variância	38
4.1	Teste bilateral para o valor médio: Situações A, C e D	57
4.2	Teste bilateral para o valor médio: Situação B	58
4.3	Teste unilateral direito para o valor médio: Situações A, C e D	61
4.4	Teste unilateral direito para o valor médio: Situação B	61
4.5	Teste unilateral esquerdo para o valor médio: Situações A, C e D	65
4.6	Teste unilateral esquerdo para o valor médio: Situação B	65
4.7	Teste bilateral para a variância	67
4.8	Teste unilateral direito para a variância	68
4.9	Teste unilateral esquerdo para a variância	69
6.1	Amostras aleatórias e não aleatórias	84

Capítulo 1

Inferência Estatística

1.1 Introdução

A área de estudo sobre inferência estatística consiste no conjunto de métodos utilizados para que seja possível tomarmos decisões ou retirarmos conclusões acerca de uma *população*. Este métodos utilizam a informação contida numa *amostra* seleccionada da população.

A inferência estatística pode ser dividida em duas grandes áreas: *estimação de parâmetros* e *testes de hipóteses*. Como exemplo de um problema sobre estimação de parâmetros, suponhamos que um engenheiro pretende analisar a resistência de uma componente usada no chassis de um automóvel. Uma vez que é natural que a resistência seja variável de componente para componente, isto devido a diferenças que podem ocorrer nos materiais e no processo de fabrico de cada componente assim, como nos métodos de leitura da respectiva resistência, o engenheiro está apenas interessado em estimar a resistência média deste tipo de componentes. Na prática, o engenheiro irá utilizar os dados de uma amostra de resistências para calcular um número que de algum modo será um valor razoável (ou uma predição) da verdadeira resistência média. Este número é denominado *estimativa pontual*. Veremos mais tarde que é possível estabelecer a precisão desta estimativa.

Consideremos a situação em que duas diferentes temperaturas de reacção, digamos t_1 e t_2 , podem ser utilizadas num processo químico. Um engenheiro conjectura que com t_1 produzirá maiores resultados que com t_2 . O teste de hipóteses estatísticas é uma ferramenta que permite resolver questões deste tipo. Neste caso, a hipótese será que o resultado médio quando usada a temperatura t_1 é maior que o resultado médio quando usada a temperatura t_2 . Repare que não é dado ênfase à estimação dos resultados; em vez disso, a atenção é dirigida para a conclusão que se pode retirar acerca da hipótese formulada sobre os resultados médios.

Comecemos por definir *amostra aleatória*, conceito fundamental na inferência estatística. Mais tarde veremos o conceito de *estimador* e *estimativa* de um parâmetro, e finalmente iremos calcular a precisão da estimativa de um parâmetro usando um *intervalo de confiança*.

1.2 População e amostra aleatória

1.2.1 População

Exemplo 1.1. Consideremos o conjunto de alunos da FCT/UNL e a informação acerca do ano de licenciatura em que encontram. Admitamos que 40% dos alunos estão no 1º ano, 30% estão no 2º ano, 20% no 3º ano e 10% no 4º ano. Se formos escolher um aluno ao caso e registarmos o seu

ano de licenciatura, então poderá ocorrer um valor $X = \text{ano de licenciatura}$, com a seguinte função de probabilidade

$$X \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

Se o objectivo for estudar o ano de licenciatura em que se encontram os alunos da FCT/UNL, esse objectivo consiste em estudar a v.a. X .

Esse estudo poderá consistir na estimação da função de probabilidade de X ou, por exemplo, na estimação do ano esperado de licenciatura de um aluno, ou na estimação do desvio padrão de X , ou na estimação do ano de licenciatura em que se encontram pelo menos 80% dos alunos, etc.

No fundo o estudo incide sobre a v.a. X ou seja sobre a população X de ano de licenciatura dos alunos da FCT/UNL.

Definição 1.1. Uma **população** consiste na totalidade das observações do fenómeno em estudo.

Em cada problema, a população pode ser pequena, grande ou infinita. O número de observações na população é designado por *dimensão da população*. Por exemplo, o número de garrafas não completamente cheias produzidas por dia numa companhia de refrigerantes é uma população finita. As observações obtidas por medição do nível diário de monóxido de carbono é uma população infinita.

A estatística dedica-se ao estudo da população, ou seja ao estudo da repartição de probabilidades dos seus valores. Se representarmos por X o conjunto dos valores da população, estudar X será estudar a sua repartição de probabilidades, portanto será estudar as características probabilísticas de uma v.a. X , ou melhor dizendo será estudar a distribuição de uma v.a. X .

Esse estudo poderá passar pela estimação da própria função de distribuição de X , ou pelo estudo do valor do parâmetro de uma certa função de distribuição que se admite ser a mais correcta para X .

Por exemplo, um engenheiro pode considerar que a população das resistências de um elemento do chassis tem distribuição normal com valor médio μ e variância σ^2 . (Quando consideramos este pressuposto, dizemos que temos uma *população normal* ou uma população normalmente distribuída.) O seu objectivo é agora estimar a resistência média μ desse elemento do chassis.

1.2.2 Amostra

Na maioria das situações, é impossível ou impraticável observar a totalidade da população. Por exemplo, não poderíamos estudar a resistência do elemento do chassis, considerando toda a população, porque isso seria demasiado demorado e dispendioso. Além do mais, alguns (por ventura todos) desses elementos não existiriam ainda, no momento em que queremos tirar uma conclusão acerca da sua resistência média.

Portanto, vamos apenas seleccionar alguns elementos da população, e com o estudo das características destas observações, tirar ilacções sobre as características de toda a população.

Ficamos dependentes de um conjunto de observações da população, para podermos tomar decisões acerca dessa mesma população.

Definição 1.2. Uma **amostra** é um conjunto de observações seleccionadas ao acaso de uma população.

Exemplo 1.2. Admitamos que para estudar $X = \text{ano de licenciatura dos alunos da FCT/UNL}$, se recolheu uma amostra de anos de licenciatura de 50 alunos. Destes, 22 estão no 1º ano, 14 no 2º

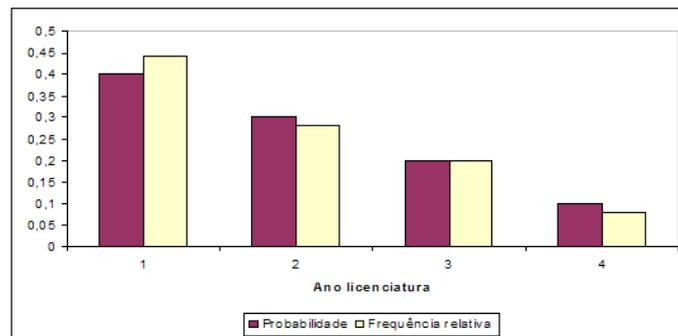
ano, 10 no 3º ano e 4 no 4º ano. O conjunto de frequências relativas desta amostra constitui uma estimativa da função de probabilidade de X .

Valores amostrais	1	2	3	4	Total
Frequência absoluta	22	14	10	4	50
Frequência relativa	0.44	0.28	0.20	0.08	1.00

Se fôssemos analisar a totalidade dos alunos da FCT/UNL, obteríamos

$$X \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

Figura 1.1: Função de probabilidade da população e da amostra



Para que as nossas inferências sejam válidas, a amostra deve ser representativa da população. É por vezes tentador, seleccionar as observações de um modo que seja cómodo ou aplicar critérios na sua escolha. Estas atitudes podem introduzir uma tendência na amostra provocando estimativas sub-avaliadas ou sobre-avaliadas. Para evitar estes problemas, é desejável seleccionar uma *amostra aleatória* de acordo com um mecanismo de escolha ao acaso. Assim sendo, a selecção de uma amostra deve ser resultado de uma experiência aleatória e cada observação dessa amostra é então um valor observado de uma variável aleatória. O modo como se distribuem as observações na população determina a função de distribuição desta variável aleatória.

Exemplo 1.3. Se de facto a função de probabilidade de X é

$$X \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

então será escolhido um aluno de 1º ano com probabilidade 0.4, um aluno do 2º ano com probabilidade 0.3, etc.

Importa agora falarmos do conceito de *amostra aleatória*. Para definirmos uma amostra aleatória, seja X uma variável aleatória que representa o resultado da selecção de uma observação da população. Seja F a função de distribuição de X . Suponhamos que cada observação da amostra é obtida de modo independente, e nas mesmas condições. Isto é, as observações da amostra são obtidas como se observássemos X , independentemente e sob as mesmas condições, por n vezes. Seja X_i a variável aleatória que representa a i -ésima réplica. Então X_1, X_2, \dots, X_n constituem uma amostra aleatória e os valores que se obtêm por concretização desta amostra aleatória são representados por x_1, x_2, \dots, x_n .

As variáveis aleatórias que constituem a amostra aleatória são independentes e têm todas a mesma função de distribuição F , uma vez que se admite que cada observação da amostra é obtida nas mesmas condições.

Exemplo 1.4. *Se no estudo de $X =$ ano de licenciatura dos alunos da FCT/UNL, optássemos por seleccionar uma amostra de 3 alunos, então X_1 representa o ano de licenciatura do 1º aluno que viermos a seleccionar. Claro que, se a função de probabilidade de X for*

$$X \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

então este o ano de licenciatura deste 1º aluno terá função de probabilidade

$$X_1 \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

O ano de licenciatura do 2º aluno que viermos a seleccionar terá função de probabilidade

$$X_2 \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

e o ano de licenciatura do 3º aluno que viermos a escolher terá função de probabilidade

$$X_3 \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

Se a escolha destes 3 alunos for perfeitamente casual, então X_1, X_2 e X_3 são v.a.'s independentes e todas igualmente distribuídas.

Admitamos que, após a escolha dos alunos, se observaram os valores $x_1 = 2$, $x_2 = 1$ e $x_3 = 1$. Isto significa que a amostra aleatória (X_1, X_2, X_3) foi concretizada na amostra observada $(x_1, x_2, x_3) = (2, 1, 1)$.

Definição 1.3. *As variáveis aleatórias (X_1, X_2, \dots, X_n) constituem uma **amostra aleatória** de dimensão n , se*

- (a) X_1, X_2, \dots, X_n são variáveis aleatórias independentes;
- (b) X_1, X_2, \dots, X_n são variáveis aleatórias com a mesma função de distribuição.

Capítulo 2

Estimação Pontual

2.1 Estatísticas

Muitas vezes o propósito da recolha da amostra consiste em obtermos informação acerca do valor dos parâmetros da distribuição da população, caso tenham valor desconhecido. Essa informação é obtida por estimação dos parâmetros, ou seja pela utilização de estatísticas adequadas ao tipo de parâmetros em causa.

Por exemplo, o engenheiro ao considerar que a população das resistências de um elemento do chassis tem distribuição normal, só pretende saber algo acerca da resistência média do elemento do chassis, por isso só pretende estimar o valor médio μ desta distribuição normal. Precisa neste caso de uma estatística para estimar μ .

Suponhamos, por exemplo, que pretendemos chegar a uma conclusão acerca da proporção de pessoas em Portugal que preferem, uma marca de refrigerante, em particular. Representemos por p o valor desconhecido desta proporção. Sendo impraticável interrogar todos os portugueses para determinarmos o verdadeiro valor de p , vamos inferir o seu valor à custa de uma amostra (de tamanho conveniente) e usando a proporção observada \hat{p} , de pessoas que nesta amostra preferem aquela marca de refrigerante.

A proporção amostral, \hat{p} , é calculada dividindo o número total de indivíduos da amostra que preferem a marca de refrigerante, pelo total de indivíduos na amostra (*dimensão da amostra*). Assim sendo, \hat{p} é uma função dos valores observados na amostra. Mas como é possível seleccionar muitas e variadas amostras de uma população, o valor de \hat{p} poderá variar de amostra para amostra. Isto é, \hat{p} é uma variável aleatória a que se dá o nome de estatística.

Definição 2.1. *Uma estatística é uma função das variáveis de uma amostra aleatória.*

Veremos mais tarde, outros exemplos importantes de estatísticas. Uma vez que uma estatística é uma variável aleatória, necessariamente terá uma função de distribuição. A essa função de distribuição é dado o nome de *distribuição de amostragem da estatística*. A noção de distribuição de amostragem é muito importante e será abordada em todos os capítulos e secções futuras.

Uma aplicação muito importante da estatística consiste na *estimação pontual* de parâmetros tais como o valor médio de uma população ou como a variância de uma população. Quando se discutem problemas de inferência estatística sobre parâmetros de uma população é conveniente representar de um modo especial esses mesmos parâmetros. Como tal é usual representá-los por uma letra grega. Por exemplo, μ para o valor médio de uma população, σ para o desvio padrão de uma população. O objectivo da estimação pontual de um parâmetro θ , consiste na atribuição de um valor numérico,

baseado na informao da amostra, que seja um valor plausvel para o parmetro θ . Esse valor numrico ser usado como estimativa pontual do parmetro.

Em geral, se X  uma populao com funo de distribuo F , caracterizada por um parmetro θ de valor desconhecido, e se X_1, X_2, \dots, X_n  uma amostra aleatria de dimenso n da populao X , ento a estatstica $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$  denominada *estimador pontual* de θ . Repare que $\hat{\Theta}$  uma varivel aleatria, porque  funo de variveis aleatrias. Aps uma amostra ter sido seleccionada, $\hat{\Theta}$ toma um valor numrico particular $\hat{\theta}$ chamado *estimativa pontual* de θ .

Definio 2.2. Uma *estimativa pontual* do parmetro θ de uma populao  um nico valor numrico $\hat{\theta}$ de uma estatstica $\hat{\Theta}$.

Exemplo 2.1. Regressemos ao exemplo do ano de licenciatura dos alunos da FCT/UNL.

Suponhamos que quermos saber qual o ano mdio de licenciatura em que se encontram estes alunos.

Se analisssemos toda a populao, sabermos que X tem funo de probabilidade

$$X \begin{cases} 1 & 2 & 3 & 4 \\ 0.4 & 0.3 & 0.2 & 0.1 \end{cases}$$

e portanto sabermos que

$$\mu = E(X) = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.2 + 4 \times 0.1 = 2 \text{  ano}$$

Mas de facto, o que conhecemos  a amostra

Valores amostrais	1	2	3	4	Total
Frequncia absoluta	22	14	10	4	50

Como tal podemos, quando muito apresentar uma estimativa pontual de μ , usando a estatstica

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

$$\bar{x} = \frac{1}{50} \sum_{i=1}^{50} x_i = \frac{1}{50} (1 \times 22 + 2 \times 14 + 3 \times 10 + 4 \times 4) = \frac{96}{50} = 1.92 \text{ ano}$$

2.2 Exemplo de estatsticas

De entre os muitos parmetros que caracterizam uma populao, ou dito de outra forma, que caracterizam a funo de distribuo da varivel aleatria X que representa uma populao, os mais comuns so os parmetros que correspondem ao valor mdio e  varincia da populao (ou de X). Por esta razo, apresentamos os estimadores mais comuns (e melhores de acordo com certos critrios estatsticos fora do nosso mbito de estudo) para o valor mdio e para a varincia de uma populao. Consideremos (X_1, X_2, \dots, X_n) uma amostra aleatria de X .

Estimador do valor mdio μ de uma populao X

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Estimador da varincia σ^2 de uma populao X

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}$$

que tambm pode ser escrito e calculado por

$$S^2 = \frac{1}{n-1} \left(\left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right) = \frac{1}{n-1} \left((X_1^2 + X_2^2 + \dots + X_n^2) - n\bar{X}^2 \right)$$

Estimador do desvio padro σ de uma populao X

$$S = +\sqrt{S^2}$$

As estimativas pontuais destes parmetros, representar-se-o por \bar{x} , s^2 e s , respectivamente.

Estimador da proporo (ou probabilidade) p de realizao de um acontecimento A

Se numa amostra de dimenso n , se observa K vezes o acontecimento A , o estimador de p 

$$\hat{P} = \frac{K}{n}$$

2.3 Estimaco pontual do valor mdio $\mu = E(X)$ da populao X

Admitamos que a populao X tem um valor mdio $\mu = E(X)$ (desconhecido) e uma varincia $\sigma^2 = V(X)$. Face a uma amostra aleatria (X_1, X_2, \dots, X_n) da populao X , podemos estimar o valor mdio μ atravs do estimador

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Contudo os valores de \bar{X} dependem directamente da amostra aleatria (X_1, X_2, \dots, X_n) . Assim \bar{X}  uma varivel aleatria e portanto sofrer de uma certa variabilidade.  desejvel que essa variabilidade seja pequena de modo a que tenhamos algumas garantias de que os valores das estimativas de μ , \bar{X} , no se afastem muito do verdadeiro valor de μ .

Para medirmos a variabilidade de uma qualquer varivel aleatria j sabemos que podemos usar a varincia dessa mesma varivel aleatria. Assim para medirmos a variabilidade de \bar{X} , podemos considerar a sua varincia, $V(\bar{X})$, que passamos a determinar.

Se (X_1, X_2, \dots, X_n)  a amostra aleatria que nos dar a informao para o clculo de \bar{X} , ento sabemos que X_1, X_2, \dots, X_n so variveis aleatrias independentes e todas com a mesma distribuio, distribuio essa que ser a distribuio da populao X em estudo.

Assim sendo, sabemos que X_1, X_2, \dots, X_n so variveis aleatrias independentes e todas com o mesmo valor mdio $E(X_i) = E(X) = \mu$ e a mesma varincia $V(X_i) = V(X) = \sigma^2$. Como tal

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Consequentemente, a variabilidade de \bar{X}  dada por σ^2/n .

Quando pretendemos dar a conhecer a preciso associada a um determinado estimador, devemos indicar a variabilidade desse estimador, mas essa variabilidade deve ser expressa na mesma escala

de medio que a associada ao estimador. Por exemplo, se estamos a estimar a altura mdia dos habitantes de um pas e a escala de medio escolhida so centmetros, ento a indicao sobre a variabilidade das estimativas da altura mdia tambm deve ser fornecida em centmetros. Em resumo, a variabilidade de um estimador deve ser expressa atravs do desvio padro desse estimador, e a que se d o nome de *erro padro* do estimador e se representa por *SE*.

No caso do estimador \bar{X} do valor mdio μ da populao X , a preciso deste estimador  dada pelo respectivo erro padro, ou seja por

$$SE(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Se a varincia da populao $\sigma^2 = V(X)$ tiver um valor desconhecido, no podemos conhecer o valor do erro padro. Neste caso, podemos aceder a um valor estimado do erro padro, substituindo σ^2 pelo seu estimador S^2 , e obter desta maneira o *erro padro estimado*, *SE**

$$SE^*(\bar{X}) = \frac{S}{\sqrt{n}}.$$

Sobre o estimador \bar{X} do valor mdio μ , ter ainda interesse saber algo sobre o seu valor mdio. Determinemos ento o valor mdio de \bar{X} .

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu.$$

Este resultado permite-nos dizer que se considerssemos todas as amostras de dimenso n que  possvel seleccionar de uma populao, e para cada uma delas determinssemos a respectiva mdia, o conjunto formado pelos valores de todas essas mdias tem um ponto de equilbrio que coincide com o ponto de equilbrio da populao (e que sabemos ser $\mu = E(X)$).

Exemplo 2.2. *O nmero de defeitos num painel metlico usado na construo de automveis tem distribuo de Poisson. Seleccionada uma amostra do n de defeitos em 10 paineis, obteve-se os seguintes valores: (2, 7, 15, 8, 7, 6, 3, 7, 3, 4).*

Se pretendemos estimar o parmetro da distribuo da populao, como sabemos que esta  Poisson e o parmetro da distribuo de Poisson coincide com o valor mdio desta distribuo, ento o problema resume-se  estimaco do valor mdio da populao.

Assim, para a amostra obtida, a estimativa do parmetro ser:

$$\bar{x} = \frac{2 + 7 + 15 + 8 + 7 + 6 + 3 + 7 + 3 + 4}{10} = \frac{62}{10} = 6.2$$

ou seja, estimamos que seja de 6.2 o n mdio de defeitos por painel.

Quanto ao erro padro da nossa estimativa, como a varincia da populao  desconhecida, quando muito podemos adiantar um erro padro estimado.

A varincia amostral,  calculvel por

$$\sum_{i=1}^{10} x_i^2 = 2^2 + 7^2 + 15^2 + 8^2 + 7^2 + 6^2 + 3^2 + 7^2 + 3^2 + 4^2 = 510$$

$$s^2 = \frac{1}{10 - 1} (510 - 10 \times 6.2^2) = \frac{125.6}{9} = 13.95555(5)$$

pelo que o erro padro estimado   de

$$SE^*(\bar{x}) = +\sqrt{\frac{s^2}{10}} = 1.181336343.$$

Podemos interpretar estes resultados, dizendo que o n o m edio de defeitos por painel   estimado em 6.2 com uma margem de erro de mais ou menos 1 defeito por painel.

2.4 M etodos para determina o de estimadores

2.4.1 M etodo dos momentos

Momentos centrados

Defini o 2.3. Seja (X_1, X_2, \dots, X_n) uma amostra aleat ria de uma popula o X . Considere $\mu \equiv E(X)$ e $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Defina-se:

- *momento centrado de ordem r por*

$$\mu_r = E[(X - \mu)]^r, \quad r \in \mathbb{N}.$$

- *momento amostral centrado de ordem r por*

$$M_r = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^r, \quad r \in \mathbb{N}$$

Observa es:

- $\mu_1 = 0$ porque $\mu_1 = E[(X - \mu)]^1 = \mu - \mu = 0$;
- $\mu_2 = V(X)$ porque $\mu_2 = E[(X - \mu)^2]$

M etodo dos momentos

Se X   uma popula o cuja distribu o depende de k par metros, $\theta_1, \theta_2, \dots, \theta_k$, o seu valor m edio $E(X)$ e os seus momentos centrados so fun o destes k par metros:

$$E(X) = \psi_1(\theta_1, \theta_2, \dots, \theta_k), \quad \mu_r = \psi_r(\theta_1, \theta_2, \dots, \theta_k), \quad r = 2, 3, \dots$$

O m etodo dos momentos consiste em:

A Constr mos um sistema de k equa es onde igualamos aqueles momentos da popula o aos respectivos estimadores,

$$\begin{cases} E(X) = \bar{X} \\ \mu_2 = M_2 \\ \mu_3 = M_3 \\ \vdots \\ \mu_k = M_k \end{cases} \Leftrightarrow \begin{cases} \psi_1(\theta_1, \theta_2, \dots, \theta_k) = \bar{X} \\ \psi_2(\theta_1, \theta_2, \dots, \theta_k) = M_2 \\ \psi_3(\theta_1, \theta_2, \dots, \theta_k) = M_3 \\ \vdots \\ \psi_k(\theta_1, \theta_2, \dots, \theta_k) = M_k \end{cases}$$

B Resolvermos o sistema de equaces em ordem s incgnitas $\theta_1, \theta_2, \dots, \theta_k$.

Admitindo que o sistema tem uma nica soluo, ento $\theta_1^*, \theta_2^*, \dots, \theta_k^*$ so os *estimadores dos momentos* dos parmetros $\theta_1, \theta_2, \dots, \theta_k$, respectivamente.

Propriedade de invarincia dos estimadores dos momentos: Se θ^*  um estimador dos momentos de θ e $h(\theta)$  uma funo biunvoca de θ , ento $h(\theta^*)$  um estimador dos momentos de $h(\theta)$.

Exemplo 2.3. Seja (X_1, X_2, \dots, X_n) uma amostra aleatria de uma populao X com distribuo $U(a, b)$. Determinemos os estimadores de momentos a^* e b^* , dos parmetros a e b .

Como sabemos, se $X \sim U(a, b)$, ento $E(X) = \frac{a+b}{2}$ e $\mu_2 = V(X) = \frac{(b-a)^2}{12}$. Assim

$$\begin{cases} E(X) = \bar{X} \\ V(X) = M_2 \end{cases} \Leftrightarrow \begin{cases} \frac{a+b}{2} = \bar{X} \\ \frac{(b-a)^2}{12} = M_2 \end{cases} \Leftrightarrow \begin{cases} a = \bar{X} - \sqrt{3M_2} \\ b = \bar{X} + \sqrt{3M_2} \end{cases}$$

Repare que $M_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$.

Os estimadores de momentos para a e b so, respectivamente
$$\begin{cases} a^* = \bar{X} - \sqrt{\frac{3(n-1)}{n}} S \\ b^* = \bar{X} + \sqrt{\frac{3(n-1)}{n}} S \end{cases}.$$

2.4.2 Mtodo da mxima verosimilhana

O mtodo da mxima verosimilhana d origem a estimadores com mais qualidade do que os estimadores dos momentos. Contudo  frequente os dois mtodos possibilitarem os mesmos estimadores.

Para a explicao deste mtodo  obrigatrio conhecer a funo de verosimilhana e entender o seu significado.  tambm necessrio entender o princpio associado a este mtodo. Neste sentido, comeamos por explorar um exemplo.

Exemplo 2.4. Considere X  uma populao com distribuo Binomial de parmetros $(2, 1/4)$. A sua funo de probabilidade :

$$X \begin{cases} 0 & 1 & 2 \\ \frac{9}{16} & \frac{6}{16} & \frac{1}{16} \end{cases}$$

Assim podemos dizer que ao seleccionarmos uma observao de X , o valor mais provvel de ser obtido  $x = 0$.

Suponhamos agora que a probabilidade de sucesso, θ , pode assumir valores $\theta = 1/4$ ou $\theta = 1/2$. No quadro seguinte apresentamos a funo de probabilidade de X para os dois valores de θ :

x	0	1	2
$P(X = x \theta = 1/4)$	$\frac{9}{16}$	$\frac{6}{16}$	$\frac{1}{16}$
$P(X = x \theta = 1/2)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Quando se selecciona uma observao de X e o resultado  $x = 1$, partindo do princpio de que se obtm este valor porque  o mais provvel de ser observado, ento θ dever ser igual a $1/2$ porque

$$P(X = 1 | \theta = 1/2) > P(X = 1 | \theta = 1/4).$$

Admitamos por fim que a probabilidade de sucesso é θ e que não se conhece o seu valor. X terá distribuição Binomial de parâmetros $(2, \theta)$ e a sua função de probabilidade é

$$f(x|\theta) = P(X = x|\theta) = \binom{2}{x} \theta^x (1-\theta)^{2-x}, \quad x = 0, 1, 2, \quad 0 < \theta < 1.$$

Se novamente obtivermos $x = 1$ na selecção de uma observação de X , e repetindo o princípio de que se obtém este valor porque é o mais provável de ser observado, então θ deverá ser o valor que torna máxima a

$$P(X = 1|\theta) = 2\theta(1-\theta), \quad 0 < \theta < 1.$$

Como $\frac{d}{d\theta} 2\theta(1-\theta) = 0 \Leftrightarrow 2(1-2\theta) = 0 \Leftrightarrow \theta = 1/2$ e $\frac{d^2}{d\theta^2} 2\theta(1-\theta) = -4 < 0$, então $P(X = 1|\theta)$ tem valor máximo para $\theta = 1/2$.

Continuemos a considerar o caso da probabilidade de sucesso θ ter um valor desconhecido e admitamos que se selecciona uma amostra de duas observações de X . Sendo (X_1, X_2) a amostra aleatória associada e se a amostra observada foi $(2, 1)$, então a sua probabilidade de realização é

$$\begin{aligned} L(\theta; (2, 1)) &= P(X_1 = 2; X_2 = 1|\theta) = \underbrace{P(X_1 = 2|\theta) P(X_2 = 1|\theta)}_{\text{porque } X_1 \text{ e } X_2 \text{ são independentes}} = \\ &= \underbrace{P(X = 2|\theta) P(X = 1|\theta)}_{\text{porque } X_1 \text{ e } X_2 \text{ têm a mesma distribuição que } X} = f(2|\theta) f(1|\theta) = \\ &= \underbrace{\theta^2}_{P(X=2|\theta)} \underbrace{2\theta(1-\theta)}_{P(X=1|\theta)} = 2\theta^3(1-\theta) \end{aligned}$$

A função $L(\theta; (2, 1))$ é designada por **função de verosimilhança da amostra (2, 1)**.

Mantendo o princípio de que a amostra $(2, 1)$ foi seleccionada por ser a mais provável, então o valor de θ será o valor que corresponde ao máximo da função de verosimilhança da amostra $(2, 1)$. Teremos então de determinar o valor de θ correspondente ao:

$$\max_{0 < \theta < 1} L(\theta; (2, 1)) = \max_{0 < \theta < 1} 2\theta^3(1-\theta).$$

Uma vez que $\frac{d}{d\theta} L(\theta; (2, 1)) = 0 \Leftrightarrow \frac{d}{d\theta} 2\theta^3(1-\theta) = 0 \Leftrightarrow 2\theta^2(3-4\theta) = 0 \Leftrightarrow \theta = 3/4$ e que $\frac{d^2}{d\theta^2} L(\theta; (2, 1))|_{\theta=3/4} < 0$, então $\theta = 3/4$.

Generalizemos agora este exemplo e consideremos que se seleccionam ao acaso e com reposição n observações da população $X \sim B(2, \theta)$. Sendo (X_1, X_2, \dots, X_n) a amostra aleatória associada a essas

selecoes, a probabilidade de ser observada a amostra (x_1, x_2, \dots, x_n) :

$$\begin{aligned}
 L(\theta; (x_1, x_2, \dots, x_n)) &= P(X_1 = x_1; X_2 = x_2; \dots; X_n = x_n | \theta) = \\
 &= \underbrace{P(X_1 = x_1 | \theta) \times P(X_2 = x_2 | \theta) \times \dots \times P(X_n = x_n | \theta)}_{\text{porque } X_1, \dots, X_n \text{ so independentes}} = \\
 &= \underbrace{P(X = x_1 | \theta) \times P(X = x_2 | \theta) \times \dots \times P(X = x_n | \theta)}_{\text{porque } X_1, \dots, X_n \text{ tm a mesma diribuio que } X} = \\
 &= \prod_{i=1}^n f(x_i | \theta) = \prod_{i=1}^n \binom{2}{x_i} \theta^{x_i} (1 - \theta)^{2-x_i} = \\
 &= \prod_{i=1}^n \binom{2}{x_i} \theta^{x_i} \prod_{i=1}^n (1 - \theta)^{2-x_i} = \prod_{i=1}^n \binom{2}{x_i} \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{\sum_{i=1}^n 2-x_i} = \\
 &\quad \text{considerando } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\
 &= \prod_{i=1}^n \binom{2}{x_i} \theta^{n\bar{x}} (1 - \theta)^{n(2-\bar{x})}
 \end{aligned}$$

Assim, a funo de verosimilhana de uma amostra (x_1, x_2, \dots, x_n) da populao $X \sim B(2, \theta)$ :

$$\begin{aligned}
 L(\theta; (x_1, \dots, x_n)) &= \underbrace{\prod_{i=1}^n \binom{2}{x_i}}_{A \equiv A((x_1, \dots, x_n))} \theta^{n\bar{x}} (1 - \theta)^{n(2-\bar{x})}, \quad 0 < \theta < 1, \quad x_i \in \{0, 1, 2\}, \quad i = 1, \dots, n \\
 &= A \theta^{n\bar{x}} (1 - \theta)^{n(2-\bar{x})}, \quad 0 < \theta < 1, \quad x_i \in \{0, 1, 2\}, \quad i = 1, \dots, n
 \end{aligned}$$

Pelo prncipio da mxima verosimilhana, a amostra (x_1, \dots, x_n) foi observada por ser a mais provvel. O valor de θ que permitiu que ela fosse a mais provvel , o valor de θ que maximiza a funo de verosimilhana de uma amostra (x_1, x_2, \dots, x_n) . Assim, θ ter um valor θ_0 tal que:

$$L(\theta_0; (x_1, \dots, x_n)) = \max_{0 < \theta < 1} L(\theta; (x_1, \dots, x_n))$$

Recorrendo aos conhecimentos de Anlise Matemtica, θ_0 ser o valor de θ que satisfaz:

$$\frac{d}{d\theta} L(\theta; (x_1, \dots, x_n)) |_{\theta=\theta_0} = 0, \quad 0 < \theta < 1 \quad \text{e} \quad \frac{d^2}{d\theta^2} L(\theta_0; (x_1, \dots, x_n)) < 0$$

Contudo, a funo de verosimilhana da amostra (x_1, x_2, \dots, x_n) no  mais do que um produto de termos no negativos, pelo que

$$L(\theta; (x_1, \dots, x_n)) = \prod_{i=1}^n P(X = x_i | \theta) = \prod_{i=1}^n f(x_i | \theta) > 0, \quad 0 < \theta < 1, \quad x_i \in \{0, 1, 2\}, \quad i = 1, \dots, n$$

Pela monotonia da funo logaritmo neperiano e por o logaritmo do produto ser igual  soma dos logaritmos, encontrar o valor de θ que maximiza $L(\theta; (x_1, \dots, x_n))$  equivalente a determinar o valor de θ que maximiza a funo log-verosimilhana da amostra (x_1, \dots, x_n) ,

$$\begin{aligned}
 l(\theta; (x_1, \dots, x_n)) &= \ln L(\theta; (x_1, \dots, x_n)) = \ln \left(\prod_{i=1}^n P(X = x_i | \theta) \right) = \ln \left(\prod_{i=1}^n f(x_i | \theta) \right) = \\
 &= \sum_{i=1}^n \ln f(x_i | \theta),
 \end{aligned}$$

isto  , o valor θ_0 tal que:

$$\frac{d}{d\theta}l(\theta; (x_1, \dots, x_n))|_{\theta=\theta_0} = 0, \quad 0 < \theta < 1 \quad e \quad \frac{d^2}{d\theta^2}l(\theta_0; (x_1, \dots, x_n)) < 0$$

Sequem-se os cculos para este exemplo:

$$\begin{aligned} L(\theta; (x_1, \dots, x_n)) &= \underbrace{\prod_{i=1}^n \binom{2}{x_i}}_{A \equiv A((x_1, \dots, x_n))} \theta^{n\bar{x}} (1 - \theta)^{n(2-\bar{x})}, \quad 0 < \theta < 1, \quad x_i \in \{0, 1, 2\}, \quad i = 1, \dots, n \\ &= A\theta^{n\bar{x}} (1 - \theta)^{n(2-\bar{x})}, \quad 0 < \theta < 1, \quad x_i \in \{0, 1, 2\}, \quad i = 1, \dots, n \\ l(\theta; (x_1, \dots, x_n)) &= \ln L(\theta; (x_1, \dots, x_n)) = \ln A + n\bar{x} \ln \theta + n(2 - \bar{x}) \ln(1 - \theta) \\ \frac{d}{d\theta}l(\theta; (x_1, \dots, x_n)) = 0 &\Leftrightarrow \frac{n\bar{x}}{\theta} - \frac{n(2 - \bar{x})}{1 - \theta} = 0 \Leftrightarrow \theta = \frac{\bar{x}}{2} \\ \frac{d^2}{d\theta^2}l(\theta_0; (x_1, \dots, x_n)) &= -n \left(\frac{\bar{x}}{\theta^2} + \frac{2 - \bar{x}}{(1 - \theta)^2} \right) \Big|_{\theta=\frac{\bar{x}}{2}} < 0 \end{aligned}$$

Para uma amostra (x_1, \dots, x_n) , o valor θ que maximiza a funo de verosimilhana   $\theta_0 = \frac{\bar{x}}{2}$.

Para uma amostra aleat ria (X_1, \dots, X_n) da populao $X \sim B(2, \theta)$ o *estimador de mxima verosimilhana de θ*   $\hat{\theta} = \frac{\bar{X}}{2}$, com $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Funo de verosimilhana

Passamos agora a apresentar a definio de funo de verosimilhana.

Definio 2.4 (Funo de verosimilhana).

Seja X uma populao cuja distribuio depende do conhecimento do valor de um parmetro θ . Para uma amostra aleat ria (X_1, X_2, \dots, X_n) da populao X , a funo de verosimilhana de uma amostra observada (x_1, x_2, \dots, x_n)  :

$$L(\theta; (x_1, \dots, x_n)) = \prod_{i=1}^n f(x_i | \theta)$$

sendo:

1. $f(x | \theta) = P(X = x | \theta)$, a funo de probabilidade de X quando X   uma populao discreta;
2. $f(x | \theta)$ a funo densidade de probabilidade de X , quando X   uma populao absolutamente cont ua.

Exemplo 2.5. Considere X uma populao com distribuio $N(\mu, 4)$. Relembre que a funo densidade associada   distribuio de X  :

$$f(x | \mu) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(x-\mu)^2}, \quad \mu \in \mathbb{R}, \quad x \in \mathbb{R}.$$

Para uma amostra (x_1, x_2, \dots, x_n) de X ,

$$\begin{aligned} L(\mu) &\equiv L(\mu; (x_1, \dots, x_n)) = \prod_{i=1}^n f(x_i | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \mu)^2} = \\ &= \left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2}, \quad \mu \in \mathbb{R}, \quad x_i \in \mathbb{R}, \quad i = 1, \dots, n. \end{aligned}$$

Vimos tambm no exemplo 2.4, a importncia da funo log-verosimilhana. Assim,

Definio 2.5 (Funo de log-verosimilhana).

Seja X uma populao cuja distribuico depende do conhecimento do valor de um parmetro θ . Para uma amostra aleatria (X_1, X_2, \dots, X_n) da populao X , a funo de log-verosimilhana de uma amostra observada (x_1, x_2, \dots, x_n) :

$$l(\theta; (x_1, \dots, x_n)) = \ln L(\theta; (x_1, \dots, x_n)),$$

definida para os valores $L(\theta; (x_1, \dots, x_n)) > 0$.

Exemplo 2.6. Na continuao do exemplo 2.5,

$$\begin{aligned} l(\mu) &\equiv l(\mu; (x_1, \dots, x_n)) = \ln \left(\frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2}, \quad \mu \in \mathbb{R}, \quad x_i \in \mathbb{R}, \quad i = 1, \dots, n \\ &= -n \ln \left(\sqrt{2\pi} \right) - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2, \quad \mu \in \mathbb{R}, \quad x_i \in \mathbb{R}, \quad i = 1, \dots, n. \end{aligned}$$

Estimador de mxima verosimilhana

Seja X uma populao cuja distribuico depende de um parmetro θ com valor desconhecido.

Ao seleccionarmos n observaces (ao acaso e com reposio) de X , com resultados (x_1, \dots, x_n) , a funo de verosimilhana desta amostra :

$$L(\theta; (x_1, \dots, x_n)) = \prod_{i=1}^n f(x_i | \theta)$$

e corresponde :

1. probabilidade da amostra aleatria (X_1, \dots, X_n) da populao X se concretizar na amostra (x_1, \dots, x_n) ;
2. funo densidade conjunta da amostra aleatria (X_1, \dots, X_n) da populao X para a amostra observada (x_1, \dots, x_n) .

O **princpio da mxima verosimilhana** estabelece que a amostra (x_1, \dots, x_n)  a mais provvel de ser observada pelo que θ dever ter um valor que corresponda ao mximo de $L(\theta; (x_1, \dots, x_n))$.

Assim, encontrar o estimador de mxima verosimilhana de θ ser determinar o valor $\hat{\theta}$ de θ que torna mxima a funo de verosimilhana, para qualquer amostra (X_1, \dots, X_n) .

A funo de verosimilhana, $L(\theta; (x_1, \dots, x_n))$, dever ser encarada como funo de θ e, caso admita derivada para todos os valores de θ , encontrar o seu mximo corresponde a determinar o valor θ_0 que satisfaz:

$$\frac{d}{d\theta} L(\theta; (x_1, \dots, x_n)) |_{\theta=\theta_0} = 0 \quad \text{e} \quad \frac{d^2}{d\theta^2} L(\theta; (x_1, \dots, x_n)) |_{\theta=\theta_0} < 0 \quad (2.4.1)$$

No final deste processo, θ_0 ser funo da amostra (x_1, \dots, x_n) , isto  $\theta_0 = T(x_1, \dots, x_n)$.

O **estimador de mxima verosimilhana** de θ  ento $\hat{\theta} = T(X_1, \dots, X_n)$.

Para efeitos prticos, e uma vez que a funo logaritmo  uma aplicao montona crescente, o processo 2.4.1 de determinao do mximo da funo de verosimilhana  equivalente ao processo de determinao do mximo da funo log-verosimilhana. Assim, poderemos calcular $\theta_0 = T(x_1, \dots, x_n)$, resolvendo:

$$\frac{d}{d\theta} l(\theta; (x_1, \dots, x_n)) |_{\theta=\theta_0} = 0 \quad \text{e} \quad \frac{d^2}{d\theta^2} l(\theta; (x_1, \dots, x_n)) |_{\theta=\theta_0} < 0 \quad (2.4.2)$$

aps o que, $\hat{\theta} = T(X_1, \dots, X_n)$  **estimador de mxima verosimilhana** de θ .

Exemplo 2.7. Na continuao dos exemplos 2.5 e 2.6, se μ tiver valor desconhecido e quisermos encontrar o seu estimador de mxima verosimilhana, seguem-se os clculos necessrios:

$$\begin{aligned} l(\mu) &= -n \ln(2\sqrt{2\pi}) - \frac{1}{8} \sum_{i=1}^n (x_i - \mu)^2, \quad \mu \in \mathbb{R}, \quad x_i \in \mathbb{R}, \quad i = 1, \dots, n. \\ \frac{d}{d\mu} l(\mu) &= \frac{1}{4} \sum_{i=1}^n (x_i - \mu) = \frac{n}{4} (\bar{x} - \mu), \quad \text{com } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \frac{d}{d\mu} l(\mu) = 0 &\Leftrightarrow \frac{n}{4} (\bar{x} - \mu) = 0 \Leftrightarrow \mu = \bar{x} \\ \frac{d^2}{d\mu^2} l(\mu) &= -\frac{n}{4} < 0 \end{aligned}$$

O estimador de mxima verosimilhana de μ  $\hat{\mu} = \bar{X}$, sendo $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Observaes:

1. Caso a funo de verosimilhana, $L(\theta; (x_1, \dots, x_n))$ (ou a funo de log-verosimilhana, $l(\theta; (x_1, \dots, x_n))$) no tenha derivada para todos os valores de θ , o seu mximo ter de ser determinado por uma anlise que no envolva a sua derivada.
2. **Propriedade de invarincia dos estimadores de mxima verosimilhana:** Se $\hat{\theta}$  o estimador de mxima verosimilhana de θ e $h(\theta)$  uma funo biunvoca de θ , ento $h(\hat{\theta})$  o estimador de mxima verosimilhana de $h(\theta)$.

2.5 Propriedades dos estimadores

J atrs foi dito que, se X  uma populao com funo de distribuio F , caracterizada por um parmetro θ de valor desconhecido, e se X_1, X_2, \dots, X_n  uma amostra aleatria de dimenso n desta populao X , ento a estatstica $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$  denominada **estimador pontual** de θ . Repare que $\hat{\Theta}$  uma varivel aleatria, porque  funo de variveis aleatrias. Aps uma amostra ter sido seleccionada, $\hat{\Theta}$ toma um valor particular $\hat{\theta}$ chamado **estimativa pontual** de θ .

Sendo $\hat{\Theta}$ uma varivel aleatria, ter uma distribuio.

2.5.1 Distribuio de amostragem de um estimador

Definio 2.6. A distribuio de um estimador pontual $\hat{\Theta}$  designada por **distribuio de amostragem**.

Exemplo 2.8. Admita que a população X descreve o número diário de contas a prazo contratualizadas num determinado Banco. Se admitirmos que $X \sim P(\lambda)$ e se quisermos estimar o número médio $E(X) = \lambda$ de contas contratualizadas diariamente, ao propormos $\hat{\Lambda} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ como estatística de avaliação do valor de $E(X) = \lambda$, a distribuição de amostragem de $\hat{\Lambda}$ é:

$$\begin{aligned} P(\hat{\Lambda} = y) &= P(\bar{X} = y) = P\left(\frac{1}{n} \sum_{i=1}^n X_i = y\right) = P\left(\sum_{i=1}^n X_i = ny\right) = \\ &= e^{-n\lambda} \frac{(n\lambda)^{ny}}{ny!}, \quad ny \in \mathbb{N}_0. \end{aligned}$$

porque $\sum_{i=1}^n X_i \sim P(n\lambda)$.

Exemplo 2.9. Admita que a população X descreve a largura (em mm) de uma tablete de chocolate produzida na fábrica ‘‘Cho’’. Se admitirmos que $X \sim N(\mu, 4)$ e se quisermos estimar a largura média $E(X) = \mu$ das tabletas de chocolates produzidas na fábrica ‘‘Cho’’, ao propormos $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ como estimador de $E(X) = \mu$, a distribuição de amostragem de $\hat{\mu}$ é expressa pela sua distribuição. Sendo (X_1, \dots, X_n) uma amostra aleatória de $X \sim N(\mu, 4)$, então $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{4}{n}\right)$.

Resumindo, a distribuição de amostragem de $\hat{\mu}$ é: $\hat{\mu} \sim N\left(\mu, \frac{4}{n}\right)$.

Como $\hat{\Theta}$ tem uma distribuição, podemos falar do $E(\hat{\Theta})$, medida que indica a localização do ponto de equilíbrio da distribuição, da $V(\hat{\Theta})$ e do desvio padrão $\sigma(\hat{\Theta})$, quantidades estas que expressam a dispersão de $\hat{\Theta}$.

Estas medidas permitem estabelecer propriedades importantes para o estimador $\hat{\Theta}$.

2.5.2 Enviesamento

O estimador $\hat{\Theta}$ para o parâmetro θ é considerado um ‘‘bom’’ estimador se o valor que se espera que ele produza coincidir com o verdadeiro valor de θ . Dito de outro modo, $\hat{\Theta}$ é ‘‘bom’’ estimador de θ se $E(\hat{\Theta}) = \theta$.

Definição 2.7. Um estimador $\hat{\Theta}$ para o parâmetro θ diz-se *centrado* (ou *não enviesado*) se $E(\hat{\Theta}) = \theta$.

Definição 2.8. O *enviesamento* de um estimador $\hat{\Theta}$ para o parâmetro θ é $bias(\hat{\Theta}) = E(\hat{\Theta}) - \theta$.

Exemplo 2.10. Suponhamos que X é uma população com distribuição exponencial de parâmetros $(\lambda, 1)$ em que λ tem valor desconhecido.

Dada uma amostra aleatória (X_1, X_2, \dots, X_n) desta população, considerem-se os dois estimadores para λ :

$$\hat{\lambda} = \min(X_1, X_2, \dots, X_n) \quad e \quad \lambda^* = \bar{X} - 1$$

e analisemos o respectivo enviesamento. Como $\hat{\lambda} \sim E\left(\lambda, \frac{1}{n}\right)$:

$$E(\hat{\lambda}) = E(\min(X_1, X_2, \dots, X_n)) = \lambda + \frac{1}{n}$$

$$E(\lambda^*) = E(\bar{X} - 1) = E(X) - 1 = \lambda + 1 - 1 = \lambda$$

Relativamente ao enviesamento, λ^* é um ‘bom’ estimador porque é centrado.

$$\text{O enviesamento de } \hat{\lambda} \text{ é bias}(\hat{\lambda}) = \lambda + \frac{1}{n} - \lambda = \frac{1}{n}.$$

2.5.3 Eficiência e erro quadrático médio

O estimador $\hat{\theta}$ para o parâmetro θ , será tanto ‘melhor’ quanto menor for a sua dispersão em torno do verdadeiro valor de θ , isto é quanto menor for $E\left[(\hat{\theta} - \theta)^2\right]$.

Defina-se então,

Definição 2.9. O erro quadrático médio do estimador pontual $\hat{\theta}$ do parâmetro θ , é

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

Teorema 2.1. Se $\hat{\theta}$ é um estimador do parâmetro θ ,

$$EQM(\hat{\theta}) = V(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 = V(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2.$$

Demonstração:

$$\begin{aligned} EQM(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] = E\left[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right] = \\ &= E\left[\underbrace{(\hat{\theta} - E(\hat{\theta}))^2}_{V(\hat{\theta})}\right] + \underbrace{\left[E(\hat{\theta}) - \theta\right]^2}_{\text{bias}(\hat{\theta})^2} + 2E\left[(\hat{\theta} - E(\hat{\theta}))\underbrace{(E(\hat{\theta}) - \theta)}_{\text{bias}(\hat{\theta})}\right] = \\ &= V(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2 + 2\underbrace{\left[E(\hat{\theta}) - E(\hat{\theta})\right]}_{=0} \text{bias}(\hat{\theta}) = \\ &= V(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2 = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \end{aligned}$$

Se o estimador $\hat{\theta}$ do parâmetro θ for um estimador centrado, isto é se $E(\hat{\theta}) = \theta$, então

$$EQM(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] = V(\hat{\theta}).$$

pele que, ele será tanto ‘melhor’ quanto menor for a sua variância.

Se o estimador $\hat{\Theta}$ é estimador **enviesado** do parâmetro θ , isto é se $E(\hat{\Theta}) \neq \theta$, então

$$EQM(\hat{\Theta}) = E\left[\left(\hat{\Theta} - \theta\right)^2\right] = V(\hat{\Theta}) + \left[E(\hat{\Theta}) - \theta\right]^2.$$

pelo que, ele será tanto ‘melhor’ quanto menor for o seu erro quadrático médio.

Face a dois estimadores, não necessariamente centrados, devemos optar pelo que apresenta menor erro quadrático médio.

Definição 2.10. *Dados dois estimadores $\hat{\Theta}$ e Θ^* para o mesmo parâmetro θ , dizemos que $\hat{\Theta}$ é **melhor** que Θ^* se,*

$$EQM(\hat{\Theta}) < EQM(\Theta^*).$$

Quando temos dois estimadores são centrados, o seu enviesamento é nulo, e então a comparação do seu erro quadrático médio acaba por ser a comparação da sua variância.

Definição 2.11. *Dados dois estimadores **centrados** $\hat{\Theta}$ e Θ^* para o mesmo parâmetro θ , dizemos que $\hat{\Theta}$ é **mais eficiente** que Θ^* se,*

$$V(\hat{\Theta}) < V(\Theta^*).$$

Definição 2.12. *Se $\hat{\Theta}$ é estimador de θ , o seu **erro padrão** é*

$$SE(\hat{\Theta}) = \sqrt{V(\hat{\Theta})}.$$

Exemplo 2.11. *Na continuação do exemplo 2.10,*

$$\begin{aligned} EQM(\lambda^*) &= V(\lambda^*) = V(\bar{X} - 1) = V(\bar{X}) = \frac{V(X)}{n} = \frac{1}{n} \\ EQM(\hat{\lambda}) &= V(\hat{\lambda}) + bias^2(\hat{\lambda}) = \frac{1}{n^2} + \frac{1}{n^2} = \frac{2}{n^2} \end{aligned}$$

Para $n \geq 3$, o estimador $\hat{\lambda}$ é ‘melhor’ do que o estimador λ^* porque $EQM(\hat{\lambda}) < EQM(\lambda^*)$.

2.5.4 Consistência

Quando se procura um estimador $\hat{\Theta}$ consistente para um parâmetro θ , a condição ‘mínima’ que se deseja é a de que, ao aumentarmos a dimensão n da amostra o erro (ou seja, a precisão) do estimador diminua (isto é a sua precisão aumente). Definindo o erro (precisão) do estimador $\hat{\Theta}$ por $|\hat{\Theta} - \theta|$, ele será consistente se, para qualquer um erro $\delta \in \mathbb{R}^+$ fixo à priori:

$$\exists n_o \in \mathbb{N} \text{ tal que para } n \geq n_o : P\left(|\hat{\Theta} - \theta| < \delta\right) = 1.$$

Definição 2.13. *Um estimador $\hat{\Theta}$ de um parâmetro θ é um estimador **consistente** de θ se e só se, qualquer que seja o valor real $\delta > 0$,*

$$\lim_{n \rightarrow +\infty} P\left(|\hat{\Theta} - \theta| < \delta\right) = 1.$$

Vejam os um exemplo

Exemplo 2.12. Admitamos que (X_1, X_2, \dots, X_n) é uma amostra aleatria de uma populao X com distribuo $N(\mu, 4)$. Sendo desconhecido o valor de μ e adotando para o estimador de μ , o estimador $\bar{X} = \frac{1}{n}X_i$, ento:

$$\bar{X} \text{ tem distribuo } N\left(\mu, \frac{4}{n}\right), \text{ ou seja } Z = \frac{\bar{X} - \mu}{2/\sqrt{n}} \sim N(0, 1).$$

Para um valor real $\delta > 0$,

$$\begin{aligned} P(|\bar{X} - \mu| < \delta) &= P\left(\left|\frac{\bar{X} - \mu}{2/\sqrt{n}}\right| < \frac{\delta}{2/\sqrt{n}}\right) = P\left(|Z| < \frac{\delta}{2/\sqrt{n}}\right) = P\left(-\frac{\delta}{2/\sqrt{n}} < Z < \frac{\delta}{2/\sqrt{n}}\right) = \\ &= P\left(Z \leq \frac{\delta}{2/\sqrt{n}}\right) - P\left(Z \leq -\frac{\delta}{2/\sqrt{n}}\right) = \Phi\left(\frac{\delta}{2/\sqrt{n}}\right) - \Phi\left(-\frac{\delta}{2/\sqrt{n}}\right) = \\ &= \Phi\left(\frac{\delta}{2/\sqrt{n}}\right) - \left[1 - \Phi\left(\frac{\delta}{2/\sqrt{n}}\right)\right] = 1 - 2\Phi\left(-\frac{\delta}{2/\sqrt{n}}\right) \end{aligned}$$

Quando $n \rightarrow +\infty$, $\frac{\delta}{2/\sqrt{n}} \rightarrow 0$ e como Φ é uma funo montona crescente, ento

$$\lim_{n \rightarrow +\infty} \Phi\left(\frac{\delta}{2/\sqrt{n}}\right) = \Phi\left(\lim_{n \rightarrow +\infty} \frac{\delta}{2/\sqrt{n}}\right) = \Phi(0).$$

Concluindo

$$\lim_{n \rightarrow +\infty} P(|\bar{X} - \mu| < \delta) = 1, \quad \forall \delta \in \mathbb{R}^+,$$

e portanto $\bar{X} = \frac{1}{n}X_i$ é estimador consistente de μ .

Vamos agora ilustrar a aplicao prtica desta propriedade. Admitamos que queremos estimar o valor mdio μ da populao com um erro que no exceda de $\delta = 0.05$, com probabilidade 0.95. A questo resume-se a determinar a dimenso mnima n da amostra que satisfaz

$$\begin{aligned} P(|\bar{X} - \mu| < 0.05) = 0.95 &\Leftrightarrow 1 - 2\Phi\left(-\frac{0.05}{2/\sqrt{n}}\right) = 0.95 \Leftrightarrow \Phi\left(-\frac{0.05}{2/\sqrt{n}}\right) = 0.025 \Leftrightarrow \\ &\Leftrightarrow \frac{0.05}{2/\sqrt{n}} = \Phi^{-1}(0.975) \Leftrightarrow \frac{0.05}{2/\sqrt{n}} = 1.96 \Leftrightarrow n \geq 6146.56 \end{aligned}$$

Quando pretendemos estimar um parmetro θ , normalmente temos à nossa disposio muitos estimadores que so consistentes. Portanto, aquilo que devemos evitar é a utilizao de estimadores que no so consistentes.

Segue-se uma regra prtica de verificao da consistncia de um estimador.

Teorema 2.2. Seja $\hat{\Theta}$ um estimador de um parmetro θ . Se:

1. $\lim_{n \rightarrow +\infty} E(\hat{\Theta}) = \theta$;
2. $\lim_{n \rightarrow +\infty} V(\hat{\Theta}) = 0$

(isto  , se $\lim_{n \rightarrow +\infty} EQM(\hat{\Theta}) = 0$), ento $\hat{\Theta}$   um estimador *consistente* de θ .

Exemplo 2.13. Retomemos os exemplos 2.10 e 2.11. Como

$$\begin{aligned} E(\lambda^*) &= \lambda \\ V(\lambda^*) &= \frac{1}{n} \xrightarrow{n \rightarrow +\infty} 0 \end{aligned}$$

$\lambda^* = \bar{X} - 1$   um estimador consistente de λ .

Como

$$\begin{aligned} E(\hat{\lambda}) &= \lambda + \frac{1}{n} \xrightarrow{n \rightarrow +\infty} \lambda \\ V(\hat{\lambda}) &= \frac{1}{n^2} \xrightarrow{n \rightarrow +\infty} 0 \end{aligned}$$

$\hat{\lambda} = \min(X_1, X_2, \dots, X_n)$   um estimador consistente de λ .

2.5.5 Propriedades de \bar{X} , S^2 e \hat{P}

Apresentamos na tabela abaixo, os estimadores mais usados para o valor m dio, vari ncia e proporo, indicando tamb m os respectivos valores m dios e vari ncias.

Tabela 2.1: Tabela de estimadores para o valor m dio, vari ncia, desvio padro e proporo

Par�metro	Estimador	Valor m�dio do estimador	Vari�ncia do estimador
θ	$\hat{\Theta}$	$E(\hat{\Theta})$	$V(\hat{\Theta})$
$\mu = E(X)$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	μ	$\frac{\sigma^2}{n}$
$\sigma^2 = V(X)$	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$	σ^2	$\frac{\sigma^4}{n} \left(\alpha_4 - \frac{n-3}{n-1} \right)$
$p = P(A)$	$\hat{P} = \frac{K}{n}$	p	$\frac{p(1-p)}{n}$

α_4   o coeficiente de curtose que tem o valor 3 para a distribuio normal.

Capítulo 3

Estimação por Intervalo de Confiança

3.1 Introdução

Em muitas situações, uma estimação pontual de um parâmetro não fornece informação suficiente sobre esse parâmetro. Vejamos o caso do exemplo 2.2. A estimativa pontual de μ , n.º médio de defeitos por painel, foi $\bar{x} = 6.2$. Mas, é pouco provável que o verdadeiro n.º médio de defeitos seja exactamente 6.2. Portanto é lógico que nos interroguemos acerca da proximidade desta estimativa relativamente ao verdadeiro n.º médio, μ . Como se frisou na secção anterior, o erro padrão (ou o erro quadrático médio, quando o estimador não é centrado) já nos dará uma ideia da precisão da nossa estimativa. Outro tipo de abordagem passa por pretendermos garantir que, para uma grande "percentagem" de todas as amostras que possamos recolher, a diferença em valor absoluto entre a média amostral \bar{X} e o valor médio μ , não ultrapassa um certo valor a (que corresponde ao erro máximo que desejamos para a estimação de μ). Se interpretarmos essa percentagem como a probabilidade de se recolher uma amostra que cumpra o anterior requisito e a representarmos por $1 - \alpha$, então podemos equacionar o problema escrevendo:

$$P(|\bar{X} - \mu| \leq a) = 1 - \alpha.$$

Como $|\bar{X} - \mu| \leq a \Leftrightarrow \bar{X} - a \leq \mu \leq \bar{X} + a$, então o que queremos encontrar é um intervalo $[\bar{X} - a, \bar{X} + a]$ que, com probabilidade $1 - \alpha$ elevada, contenha o valor médio μ .

Designamos esse intervalo por intervalo de confiança $1 - \alpha$ para μ e realizamos assim uma estimação de μ por intervalo de confiança (ou estimação intervalar de μ).

3.1.1 Intervalo de confiança $(1 - \alpha)$

Definição 3.1. Um **intervalo de confiança $1 - \alpha$** para um parâmetro θ (de valor desconhecido), é um intervalo da forma

$$[L(X_1, X_2, \dots, X_n), U(X_1, X_2, \dots, X_n)]$$

onde $L(X_1, X_2, \dots, X_n)$ e $U(X_1, X_2, \dots, X_n)$ são estatísticas (e variáveis aleatórias) que não dependem do valor de θ , e que satisfazem

$$P(L(X_1, X_2, \dots, X_n) \leq \theta \leq U(X_1, X_2, \dots, X_n)) = 1 - \alpha.$$

O intervalo (aleatório) $[L(X_1, X_2, \dots, X_n), U(X_1, X_2, \dots, X_n)]$ é chamado *estimador intervalar* para o parâmetro θ .

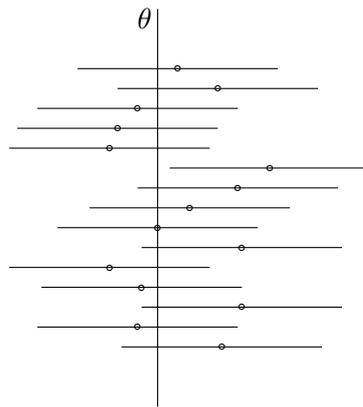
$L(X_1, X_2, \dots, X_n)$ e $U(X_1, X_2, \dots, X_n)$ so denominados **limites de confiana inferior e superior**, respectivamente, e $(1 - \alpha)$  chamado **coeficiente (ou nvel) de confiana do intervalo**.

Representamos o **intervalo de confiana $1 - \alpha$** para um parmetro θ por

$$IC_{(1-\alpha) \times 100\%}(\theta) \equiv [L(X_1, X_2, \dots, X_n), U(X_1, X_2, \dots, X_n)].$$

Podemos interpretar um intervalo de confiana pensando que, se infinitas amostras forem seleccionadas e um intervalo de confiana $(1 - \alpha)$ for calculado para cada uma delas, ento $100(1 - \alpha)$ desses intervalos contm o verdadeiro valor de θ .

Esta situao  ilustrada na figura que se segue, que mostra diversos intervalos de confiana $(1 - \alpha)$ para o parmetro θ de uma populao. Os pontos no centro dos intervalos indicam a estimativa pontual de θ (isto , $\hat{\theta}$). Repare que um dos 15 intervalos falha em conter o verdadeiro valor de θ . Se estes fossem intervalos de 95% de confiana, de entre infinitos intervalos que calculssemos (com base em infinitas amostras) apenas 5% deles no iriam conter o verdadeiro valor de θ .



Na prtica, ns so temos uma amostra (x_1, x_2, \dots, x_n) para a qual determinamos um intervalo de confiana $[l(x_1, x_2, \dots, x_n), u(x_1, x_2, \dots, x_n)]$. Como este intervalo vai conter ou no o verdadeiro valor do parmetro θ , no  razovel associar uma probabilidade a este acontecimento especfico. O que devemos afirmar  que o intervalo observado $[l(x_1, x_2, \dots, x_n), u(x_1, x_2, \dots, x_n)]$ abrange o verdadeiro valor de θ com uma confiana de $(1 - \alpha)$. Esta afirmao tem uma interpretao frequencista; isto , ns no sabemos se, para uma amostra especfica, a afirmao  verdadeira, mas o mtodo usado para obter o intervalo $[l(x_1, x_2, \dots, x_n), u(x_1, x_2, \dots, x_n)]$ permite afirmaoes correctas $100(1 - \alpha)$ das vezes.

A amplitude observada, $u(x_1, x_2, \dots, x_n) - l(x_1, x_2, \dots, x_n)$, de um intervalo de confiana observado  uma importante medida da qualidade da estimao do parmetro. Em particular, a metade da amplitude do intervalo, designada por **preciso** da estimao por intervalo de confiana,  um indicador da disperso da estimativa do parmetro θ . Quanto maior for um intervalo de confiana, mais confiana temos de que esse intervalo contem de facto o verdadeiro valor de θ . Por outro lado, quanto maior for o intervalo de confiana, (menor preciso da estimao) menos informao temos acerca do verdadeiro valor de θ , uma vez que temos uma maior gama de valores possveis para θ . A situao ideal reside num intervalo de pequena amplitude e com elevado coeficiente de confiana.

3.1.2 Mtodo Pivotal

De seguida, apresentamos um mtodo de construo de intervalos de confiana, designado por *mtodo pivotal*. Para o pormos em prtica  necessrio encontrarmos ou conhecermos uma *varivel pivot*.

Varivel Pivot

Definio 3.2. *Seja (X_1, X_2, \dots, X_n) uma amostra aleatria de uma populao cuja distribuo depende de um parmetro θ . Consideremos a varivel aleatria $W \equiv W(X_1, X_2, \dots, X_n, \theta)$, funo da amostra aleatria e de θ (e eventualmente de outros parmetros de valor conhecido). Se a distribuo de W no depende de θ , ento W  uma *varivel pivot* para θ .*

Exemplo 3.1. *Se (X_1, X_2, \dots, X_n) for uma amostra aleatria de uma populao $X \sim N(\mu, 5^2)$, ento*

$$Z = \frac{\bar{X} - \mu}{\sqrt{5^2/n}} \sim N(0, 1)$$

Deste modo podemos afirmar que Z  uma varivel pivot para μ , porque a distribuo de Z  sempre $N(0, 1)$, qualquer que seja o valor de μ .

Exemplo 3.2. *Retomemos o exemplo 2.10. O estimador $\hat{\lambda} = \min(X_1, X_2, \dots, X_n)$ de λ da populao $X \sim E(\lambda, 1)$, tem distribuo $E\left(\lambda, \frac{1}{n}\right)$.*

Como $W = \hat{\lambda} - \lambda$ tem distribuo $E\left(0, \frac{1}{n}\right)$, ento W  varivel pivot porque, qualquer que seja o valor de λ , W tem sempre distribuo $E\left(0, \frac{1}{n}\right)$.

Exemplo 3.3. *Admita que X  uma populao com distribuo de Poisson com parmetro λ . Se λ tiver um valor desconhecido e o quisermos estimar, podemos considerar o seu estimador dos momentos $\lambda^* = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.*

Apelando o Teorema Limite Central, para $n \geq 30$,

$$W = \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \stackrel{a}{\sim} N(0, 1).$$

Assim, W  uma varivel pivot para λ porque qualquer que seja o valor de λ , W tem distribuo assinttica $N(0, 1)$.

Mtodo Pivotal

O *mtodo pivotal* para determinao de um intervalo de confiana $1 - \alpha$ para θ , consiste em:

- I. Conhecer (ou encontrar) uma *varivel pivot* $W \equiv W(X_1, X_2, \dots, X_n, \theta)$ para θ ;
- II. A partir da distribuo de W , *determinar* valores a_1 e a_2 , que satisfaam $a_1 < a_2$ e

$$P(a_1 \leq W \leq a_2) = 1 - \alpha;$$

III. Resolver as desigualdades

$$a_1 \leq W(X_1, X_2, \dots, X_n, \theta) \leq a_2$$

em ordem a θ , de modo a que

$$a_1 \leq W(X_1, X_2, \dots, X_n, \theta) \leq a_2 \Leftrightarrow L(X_1, X_2, \dots, X_n) \leq \theta \leq U(X_1, X_2, \dots, X_n),$$

sendo $L(X_1, X_2, \dots, X_n)$ e $U(X_1, X_2, \dots, X_n)$ estatísticas não dependentes de θ ;

IV.

$$IC_{(1-\alpha) \times 100\%}(\theta) \equiv [L(X_1, X_2, \dots, X_n), U(X_1, X_2, \dots, X_n)]$$

é um *intervalo de confiança* $1 - \alpha$ para θ .

Observação: Normalmente são usados coeficientes de confiança de 90%, 95% e 99%.

NOTA: Para um coeficiente de confiança $1 - \alpha$ fixo, existem diferentes escolhas possíveis para as constantes a_1 e a_2 . Sempre que possível devemos optar por usar aquelas que permitam que o intervalo de confiança tenha amplitude mínima.

- Quando a variável pivot $W \equiv W(X_1, X_2, \dots, X_n, \theta)$ tem uma distribuição simétrica em torno de zero, obtemos um intervalo de amplitude mínima ao considerarmos: $a_1 < a_2$ e:

$$a_1 = -a_2 \quad \text{com } a_2 \quad \text{satisfazendo} \quad P(W > a_2) = \alpha/2. \quad (3.1.1)$$

- Se a variável pivot $W \equiv W(X_1, X_2, \dots, X_n, \theta)$ não tem uma distribuição simétrica, não é simples determinar as constantes a_1 e a_2 que correspondam ao intervalo de confiança com amplitude mínima. Na prática, abdica-se deste objectivo, e opta-se por uma solução aproximada escolhendo as constantes a_1 e a_2 que satisfazem: $a_1 < a_2$ e

$$P(W \leq a_1) = \alpha/2 \quad \text{e} \quad P(W > a_2) = \alpha/2. \quad (3.1.2)$$

Nas secções que se seguem, vamos aplicar o método pivotal para a construção de intervalos de confiança para os casos mais comuns de aplicação, ou seja para a estimação intervalar de:

- Valor médio da população, $E(X) = \mu$;
- Variância da população, $V(X) = \sigma^2$;
- Proporção populacional de ocorrência de um acontecimento A , $p = P(A)$.

3.2 Estimação por intervalo de confiança do valor médio $\mu = E(X)$ da população X

Apliquemos os conceitos sobre intervalo de confiança expostos na secção anterior.

Agora o parâmetro θ será o valor médio $\mu = E(X)$, e adoptemos para estimador deste parâmetro, a média amostral,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Determinar um intervalo de confiança $(1 - \alpha)$ para μ , consiste em determinar os limites de confiança $L \equiv L(X_1, X_2, \dots, X_n)$ e $U \equiv U(X_1, X_2, \dots, X_n)$ que verificam a igualdade

$$P(L \leq \mu \leq U) = 1 - \alpha, \quad 0 < \alpha < 1.$$

Também de acordo com o que foi dito, será a partir de \bar{X} e da sua distribuição de amostragem, que poderemos deduzir os valores de L e de U .

Nas diversas situações que se seguem, vamos deduzir diversos intervalos de confiança para μ , pondo em prática o **método pivot**.

Situação A Admitamos que se sabe que a população X tem distribuição **Normal** de valor médio μ (que se pretende estimar) e **variância σ^2 conhecida**.

I. Variável pivot

Se (X_1, X_2, \dots, X_n) é uma amostra aleatória da população X , então X_1, X_2, \dots, X_n são v.a.'s independentes e identicamente distribuídas (i.i.d.) com distribuição igual à da população X . Assim X_1, X_2, \dots, X_n são v.a.'s independentes e $X_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$.

Podemos então concluir que $\bar{X} \sim N(E(\bar{X}), V(\bar{X})) \equiv N\left(\mu, \frac{\sigma^2}{n}\right)$, e portanto que

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

Como σ^2 tem um valor conhecido,

$$Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \quad \text{é uma variável pivot}$$

e tem distribuição Normal reduzida.

II. Determinação das constantes a_1 e a_2

Definição 3.3. Seja $Z \sim N(0, 1)$. O quantil de probabilidade $1 - q$ da v.a. Z é o valor real z_q que satisfaz:

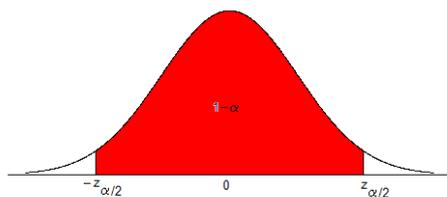
$$P(Z \leq z_q) = 1 - q, \quad 0 < q < 1.$$

Devido à distribuição da variável pivot Z ser simétrica em torno de zero e tendo em conta a nota 3.1.1, consideremos as constantes a_1 e a_2 que satisfazem $a_1 < a_2$ e

$$a_1 = -a_2 \quad \text{com } a_2 \quad \text{tal que} \quad P(Z > a_2) = \alpha/2.$$

De acordo com a anterior definição, $a_2 = z_{\alpha/2}$ e $a_1 = -z_{\alpha/2}$ (ver a figura 3.1).

Figura 3.1: Intervalos de confiança para o valor médio: Situações A, C e D



III. Resolução das desigualdades $a_1 \leq Z \leq a_2$ em ordem a μ

$$\begin{aligned} a_1 \leq Z \leq a_2 &\Leftrightarrow -z_{\alpha/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \leq z_{\alpha/2} \Leftrightarrow \\ &\Leftrightarrow \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{aligned}$$

IV. Intervalo de confiança $(1 - \alpha)$ para μ

$$IC_{(1-\alpha) \times 100\%}(\mu) \equiv \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

**Intervalo de confiança $(1 - \alpha)$ para o valor médio μ
População normal com variância σ^2 conhecida**

$$IC_{(1-\alpha) \times 100\%}(\mu) \equiv \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right]$$

Exemplo 3.4. O tempo que uma máquina leva a executar a sua tarefa em cada peça produzida segue uma distribuição normal de desvio padrão igual a 3 segundos.

Pretendendo-se estimar por intervalo de 95% de confiança, o tempo médio de execução das peças, recolheu-se uma amostra de tempos de execução de 25 peças, cuja média foi de 12 segundos.

Assim,

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow 1 - \alpha/2 = 0.975$$

$$z_{\alpha/2} = z_{0.025} = \Phi^{-1}(0.975) = 1.96$$

$$\bar{x} = 12, \quad \sigma = 3, \quad n = 25$$

Intervalo de confiança 0.95 para μ

$$IC_{95\%}(\mu) = \left[12 - \frac{3}{\sqrt{25}} 1.96, 12 + \frac{3}{\sqrt{25}} 1.96 \right] = [10.824, 13.176]$$

Podemos dizer com 95% de confiança, que o intervalo anterior inclui o verdadeiro tempo médio de execução das peças produzidas pela máquina.

Situação C Admitamos que (X_1, X_2, \dots, X_n) é uma amostra aleatória de dimensão $n \geq 30$, de uma população X cuja distribuição é desconhecida ou, sendo conhecida, não é normal, mas com variância σ^2 conhecida. Seja μ o valor médio da população X , que queremos estimar.

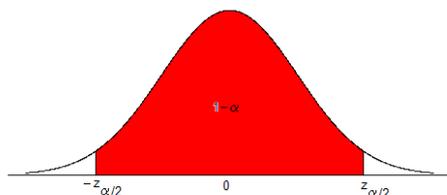
Apesar de se conhecer o valor da variância σ^2 isso por si só não permite o conhecimento da distribuição de \bar{X} . Contudo se a amostra for grande, isto é se tiver uma dimensão $n \geq 30$, por aplicação do Teorema Limite Central, podemos afirmar que

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

(tem uma distribuição aproximadamente normal reduzida).

Portanto, repetindo os raciocínios efectuados na dedução do intervalo da Situação A, obtemos um intervalo (assintótico) de confiança $(1 - \alpha)$ para μ .

Figura 3.2: Intervalos de confiança para o valor médio: Situações A, C e D



**Intervalo de confiança $(1 - \alpha)$ para o valor médio μ
População com distribuição desconhecida ou conhecida mas não normal,
com variância σ^2 conhecida e $n \geq 30$**

$$IC_{(1-\alpha) \times 100\%}(\mu) \equiv \left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right]$$

Situação B Consideremos (X_1, X_2, \dots, X_n) uma amostra aleatória de uma população X com distribuição normal de valor médio μ (que se pretende estimar) e variância σ^2 desconhecida.

Relativamente á Situação A, o que agora se altera é o facto da variância σ^2 ser desconhecida.

Se a variância σ^2 é desconhecida, podemos de imediato pensar em a substituir pela variância amostral, ou seja, por

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right).$$

Como resultado desta substituição o intervalo que se obteve na Situação A passará a ter por expressão

$$\left[\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2} \right].$$

Contudo, como S^2 é um estimador de σ^2 , o seu valor pode não ser igual ao verdadeiro valor de σ^2 . Dito de outro modo, a substituição de σ por S , no intervalo da Situação A, pode introduzir erro no intervalo, e como consequência, não temos garantias de que este novo intervalo permita a estimação de μ com a mesma confiança $(1 - \alpha)$.

De facto o que está em causa é que $P\left(\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2}\right)$ é inferior a $(1 - \alpha)$.

Para que esta probabilidade (confiança) continue a ser $(1 - \alpha)$, devemos aumentar a amplitude do intervalo. O aumento do intervalo que garante que ele tenha uma confiança $(1 - \alpha)$ é conseguido substituindo, no intervalo acima apresentado, $z_{\alpha/2}$ pelo quantil de probabilidade $(1 - \alpha/2)$ da distribuição **t (Student) com $(n - 1)$ graus de liberdade**, que representamos por $t_{n-1;\alpha/2}$.

Recorrendo ao método pivotal e tendo em conta a distribuição de amostragem de \bar{X} apresentada na Situação B da secção 3.5.1 (resultado 3.5.5),

I. Variável pivot

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}.$$

II. Determinação das constantes a_1 e a_2

Definição 3.4. *Seja W uma v.a. com distribuição t (t -Student) com m graus de liberdade, $W \sim t_m$. O quantil de probabilidade $1 - q$ da v.a. W é o valor real $t_{m;q}$ que satisfaz:*

$$P(W \leq t_{m;q}) = 1 - q, \quad 0 < q < 1.$$

Devido à distribuição da estatística pivot T ser simétrica em torno de zero e tendo em conta a nota 3.1.1, consideremos as constantes a_1 e a_2 que satisfazem $a_1 < a_2$ e

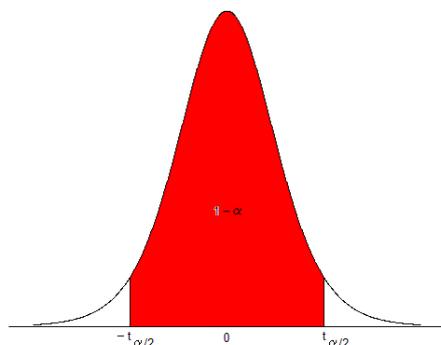
$$a_1 = -a_2 \quad \text{com } a_2 \quad \text{tal que} \quad P(T > a_2) = \alpha/2.$$

De acordo com a anterior definição, $a_2 = t_{n-1;\alpha/2}$ e $a_1 = -t_{n-1;\alpha/2}$ (ver a figura 3.3).

III. Resolução das desigualdades $a_1 \leq T \leq a_2$ em ordem a μ

$$\begin{aligned} a_1 \leq T \leq a_2 &\Leftrightarrow -t_{n-1;\alpha/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{S} \leq t_{n-1;\alpha/2} \Leftrightarrow \\ &\Leftrightarrow \bar{X} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \end{aligned}$$

Figura 3.3: Intervalos de confiança para o valor médio: Situação C



IV. Intervalo de confiança $(1 - \alpha)$ para μ

$$IC_{(1-\alpha)\times 100\%}(\mu) \equiv \left[\bar{X} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right]$$

**Intervalo de confiança $(1 - \alpha)$ para o valor médio μ
População normal com variância σ^2 desconhecida**

$$IC_{(1-\alpha)\times 100\%}(\mu) \equiv \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1;\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1;\alpha/2} \right]$$

Exemplo 3.5. Uma amostra do peso de 8 animais alimentados com um determinado tipo de ração, forneceu os seguintes valores (em kg):

4 6 4.5 4 5.6 6.2 5.8 6

Admitindo que o peso dos animais se comporta de acordo com uma distribuição normal, apresente uma estimativa por intervalo de 90% de confiança para o peso médio dos animais alimentados com este tipo de ração.

$$n = 8 \quad \sum_{i=1}^8 x_i = 42.1 \quad \sum_{i=1}^8 x_i^2 = 227.69$$

$$\bar{x} = \frac{42.1}{8} = 5.2625 \quad s^2 = \frac{1}{7} (227.69 - 8 \times 5.2625^2) = 0.8769657$$

$$s = +\sqrt{s^2} = 0.9364644$$

$$1 - \alpha = 0.9 \Rightarrow \alpha = 0.1 \Rightarrow \alpha/2 = 0.05 \quad t_{7;0.05} = 1.895$$

Intervalo de confiança 0.9 para o peso médio dos animais

$$IC_{90\%}(\mu) = \left[5.2625 - \frac{0.9364644}{\sqrt{8}} \times 1.895, 5.2625 + \frac{0.9364644}{\sqrt{8}} \times 1.895 \right] = [4.63508414, 5.88991586]$$

Situação D Consideremos (X_1, X_2, \dots, X_n) uma amostra aleatória de dimensão $n \geq 30$, de uma população X cuja distribuição é desconhecida ou, sendo conhecida, não é normal e cuja variância σ^2 não é conhecida.

Relativamente á Situação C, o que agora se altera é o facto da variância σ^2 ser desconhecida. Se a variância σ^2 é desconhecida, podemos de imediato pensar em a substituir pela variância amostral, ou seja, por

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\left(\sum_{i=1}^n X_i^2 \right) - n\bar{X}^2 \right).$$

Como resultado desta substituição o intervalo que se obteve na Situação B passará a ter por expressão

$$\left[\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2} \right].$$

Naturalmente que é introduzido um erro no intervalo, devido à substituição realizada. Contudo, como a estimação de σ^2 é feita a partir de uma amostra grande, a sua precisão é razoável, ou seja, não se introduz um erro apreciável neste intervalo ao substituir σ^2 pelo seu estimador S^2 .

Por aplicação do método pivotal e tendo em conta a distribuição de amostragem de \bar{X} apresentada na Situação D da secção 3.5.1 (resultado 3.5.6) vejamos quais as deduções envolvidas: Recorrendo ao método pivotal e tendo em conta a distribuição de amostragem de \bar{X} apresentada na Situação C da secção 3.5.1 (resultado 3.5.7),

I. Variável pivot

$$Z = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \sqrt{n} \frac{\bar{X} - \mu}{S} \stackrel{a}{\sim} N(0, 1).$$

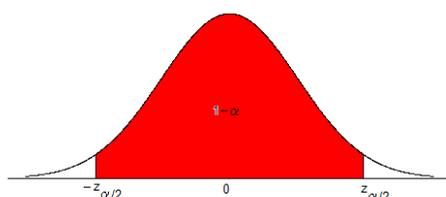
II. Determinação das constantes a_1 e a_2

Devido à distribuição da estatística pivot Z ser simétrica em torno de zero e tendo em conta a nota 3.1.1, consideremos as constantes a_1 e a_2 que satisfazem $a_1 < a_2$ e

$$a_1 = -a_2 \quad \text{com } a_2 \quad \text{tal que} \quad P(Z > a_2) = \alpha/2.$$

De acordo com a definição 3.3, $a_2 = z_{\alpha/2}$ e $a_1 = -z_{\alpha/2}$.

Figura 3.4: Intervalos de confiança para o valor médio: Situações A, B e D



III. Resolução das desigualdades $a_1 \leq Z \leq a_2$ em ordem a μ

$$\begin{aligned} a_1 \leq Z \leq a_2 &\Leftrightarrow -z_{\alpha/2} \leq \sqrt{n} \frac{\bar{X} - \mu}{S} \leq z_{\alpha/2} \Leftrightarrow \\ &\Leftrightarrow \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \end{aligned}$$

IV. Intervalo de confiança $(1 - \alpha)$ para μ

$$IC_{(1-\alpha) \times 100\%}(\mu) \equiv \left[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

**Intervalo de confiança $(1 - \alpha)$ para o valor médio μ
População com distribuição desconhecida ou conhecida mas não normal,
com variância σ^2 desconhecida e $n \geq 30$**

$$IC_{(1-\alpha) \times 100\%}(\mu) \equiv \left[\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} z_{\alpha/2} \right]$$

3.3 Estimação por intervalo de confiança da variância $\sigma^2 = V(X)$ e do desvio padrão $\sigma = \sigma(X)$, da população X

Agora o parâmetro a estimar é a variância da população X , $\sigma^2 = V(X)$, e consideramos o seu estimador

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Determinar um intervalo de confiança $(1 - \alpha)$ para σ^2 , consiste em determinar os extremos L e U que verificam a igualdade

$$P(L \leq \sigma^2 \leq U) = 1 - \alpha, \quad 0 < \alpha < 1.$$

Será a partir de S^2 e da sua distribuição de amostragem, que poderemos deduzir os valores de L e de U .

I. Variável pivot

Consideremos uma amostra aleatória (X_1, X_2, \dots, X_n) da população X . Quando a população X tem distribuição $N(\mu, \sigma^2)$, a estatística

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

tem distribuição do **qui-quadrado com $(n-1)$ graus de liberdade** (e escrevemos de modo abreviado, $X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$).

X^2 é uma variável pivot para σ^2 .

II. Determinação das constantes a_1 e a_2

Observação: A distribuição do qui-quadrado não é simétrica.

Consideremos um coeficiente de confiança $(1 - \alpha)$ e determinemos as constantes a_1 e a_2 que verificam

$$P(a_1 \leq X^2 \leq a_2) = 1 - \alpha.$$

Devido à não simetria da distribuição do qui-quadrado e de acordo com a nota 3.1.2, podemos considerar as constantes que satisfazem as condições:

$$P(X^2 \leq a_1) = \frac{\alpha}{2} \quad \text{e} \quad P(X^2 \geq a_2) = \frac{\alpha}{2}.$$

Definição 3.5. Seja W uma v.a. com distribuição do qui-quadrado com m graus de liberdade, $W \sim \chi_m^2$. O quantil de probabilidade $1 - q$ da v.a. W é o valor real $\chi_{m;q}^2$ que satisfaz:

$$P(W \leq \chi_{m;q}^2) = 1 - q, \quad 0 < q < 1.$$

Os valores de a_1 e de a_2 podem ser lidos numa tabela da distribuição do qui-quadrado e, de acordo com a definição 3.5, serão:

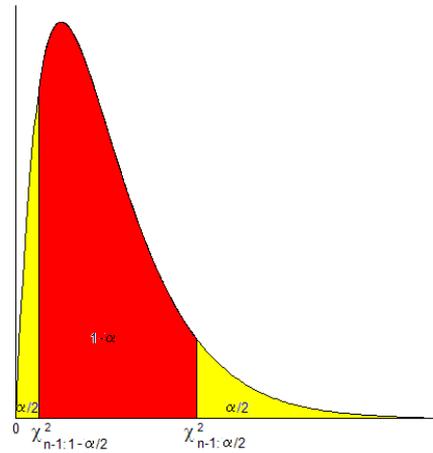
$$a_1 = \chi_{n-1;1-\alpha/2}^2 \quad \text{e} \quad a_2 = \chi_{n-1;\alpha/2}^2.$$

(ver a figura 3.5)

III. Resolução das desigualdades $a_1 \leq X^2 \leq a_2$ em ordem a σ^2

$$a_1 \leq X^2 \leq a_2 \Leftrightarrow \chi_{n-1;1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{n-1;\alpha/2}^2 \Leftrightarrow \frac{(n-1)S^2}{\chi_{n-1;\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{n-1;1-\alpha/2}^2}$$

Figura 3.5: Intervalo de confiança para a variância



IV. Intervalo de confiança $(1 - \alpha)$ para σ^2

$$IC_{(1-\alpha) \times 100\%}(\sigma^2) \equiv \left[\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2} \right]$$

IV. Intervalo de confiança $(1 - \alpha)$ para σ

$$IC_{(1-\alpha) \times 100\%}(\sigma) \equiv \left[\sqrt{\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2}} \right]$$

**Intervalo de confiança $(1 - \alpha)$ para a variância σ^2
População normal com valor médio μ , desconhecido**

$$IC_{(1-\alpha) \times 100\%}(\sigma^2) \equiv \left[\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2} \right]$$

**Intervalo de confiança $(1 - \alpha)$ para o desvio padrão σ
População normal com valor médio μ , desconhecido**

$$IC_{(1-\alpha) \times 100\%}(\sigma) \equiv \left[\sqrt{\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}}, \sqrt{\frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2}} \right]$$

Exemplo 3.6. Considere uma amostra de alturas de 25 pessoas, que apresenta uma média e um desvio padrão de, respectivamente, 172 e 5 centímetros. Admitindo que a altura de qualquer pessoa X , é uma variável com distribuição normal, estimemos por intervalo de 90% de confiança, a variância e o desvio padrão da altura de todas as pessoas.

Sabemos que $s = 5$ e portanto que $s^2 = 25$. Para $n = 25$ e $\alpha = 10\%$,

$$\chi_{24:0.95}^2 = 13.85 \quad e \quad \chi_{24:0.05}^2 = 36.42.$$

A estimativa por intervalo de 90% de confiança para a variância populacional é

$$IC_{90\%}(\sigma^2) \equiv \left[\frac{24 \times 25}{36.42}, \frac{24 \times 25}{13.85} \right] = [16.47446458, 43.32129964]$$

A estimativa por intervalo de 90% de confiança para o desvio padrão da população é

$$IC_{90\%}(\sigma) \equiv \left[\sqrt{16.47446458}, \sqrt{43.32129964} \right] = [4.058874792, 6.581891798]$$

3.4 Estimação por intervalo de confiança da proporção p de observação do acontecimento A

Suponhamos que, como resultado de uma experiência aleatória, queremos observar se ocorre ou não um acontecimento A . Para n realizações independentes da experiência, associemos n variáveis aleatórias X_i , $i = 1, \dots, n$ tais que

$$X_i = \begin{cases} 0 & \text{se não ocorre } A \\ 1 & \text{se ocorre } A \end{cases}$$

A v.a. $X = \sum_{i=1}^n X_i$ regista o total de ocorrências de A nas n experiências. X tem distribuição

Binomial de parâmetros (n, p) .

Consideremos o estimador de p :

$$\hat{P} = \frac{X}{n}.$$

Para deduzirmos o intervalo de confiança para p , precisamos da distribuição de amostragem de \hat{P} . Esta distribuição de amostragem só possibilitará intervalos de confiança de aplicação "cómoda" quando as amostras são "grandes".

I. Variável pivot

Ora para "grandes" amostras, isto é aquelas em $n \geq 30$, podemos aplicar o Teorema Limite Central e ter acesso à distribuição assintótica de \hat{P} . Ou seja, para $n \geq 30$,

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} = \sqrt{n} \frac{\hat{P} - p}{\sqrt{p(1-p)}} \stackrel{a}{\sim} N(0, 1)$$

Saliente-se que Z é uma estatística pivot.

II. Determinação das constantes a_1 e a_2

Devido à distribuição da estatística pivot Z ser simétrica em torno de zero e tendo em conta a nota 3.1.1, consideremos as constantes a_1 e a_2 que satisfazem $a_1 < a_2$ e

$$a_1 = -a_2 \quad \text{com } a_2 \quad \text{tal que} \quad P(Z > a_2) = \alpha/2.$$

De acordo com a definição 3.3, $a_2 = z_{\alpha/2}$ e $a_1 = -z_{\alpha/2}$.

III. Resolução das desigualdades $a_1 \leq Z \leq a_2$ em ordem a p

$$\begin{aligned} a_1 \leq Z \leq a_2 &\Leftrightarrow -z_{\alpha/2} \leq \sqrt{n} \frac{\hat{P} - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2} \Leftrightarrow \\ &\Leftrightarrow \hat{P} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \end{aligned} \quad (3.4.3)$$

Na resolução acima, verificamos que os limites do intervalo dependem do valor de p que não conhecemos. Para resolver este problema sugerimos duas alternativas:

a) A solução exacta

$$\begin{aligned} a_1 \leq Z \leq a_2 &\Leftrightarrow -z_{\alpha/2} \leq \sqrt{n} \frac{\hat{P} - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2} \Leftrightarrow n \frac{(\hat{P} - p)^2}{p(1-p)} \leq z_{\alpha/2}^2 \Leftrightarrow \\ &\Leftrightarrow \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \leq p \leq \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} + \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \end{aligned}$$

b) A solução aproximada

Para uma solução aproximada, substituímos na expressão 3.4.3, as ocorrências de p no desvio padrão de \hat{P} , isto é, em $\sqrt{\frac{p(1-p)}{n}}$, pelo seu estimador \hat{P} . Assim

$$\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq p \leq \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

IV. Intervalo de confiança $(1 - \alpha)$ para p **a) Para uma estimação mais precisa**

$$IC_{(1-\alpha) \times 100\%}(p) \equiv \left[\frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)}, \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} + \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \right]$$

b) Para uma estimação menos precisa

$$IC_{(1-\alpha) \times 100\%}(p) \equiv \left[\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right]$$

Intervalo de confiança $(1 - \alpha)$ para a proporção p **Amostras grandes, $n \geq 30$**

$$IC_{(1-\alpha) \times 100\%}(p) \equiv \left[\frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} - \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)}, \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} + \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{P}(1-\hat{P}) + \frac{z_{\alpha/2}^2}{4n}}}{\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \right]$$

Intervalo de confiança $(1 - \alpha)$ para a proporção p **Amostras grandes, $n \geq 30$**

$$IC_{(1-\alpha) \times 100\%}(p) \equiv \left[\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right]$$

Exemplo 3.7. Num inquérito telefónico destinado a estimar a proporção da população que tem acesso à internet em casa, foram inquiridas 50 pessoas, das quais 32 afirmaram ter este serviço.

A estimativa por intervalo de 95% de confiança para a proporção da população é

$$IC_{95\%}(p) = \left[0.64 - 1.96 \sqrt{\frac{0.64(1-0.64)}{50}}, 0.64 + 1.96 \sqrt{\frac{0.64(1-0.64)}{50}} \right] = [0.507, 0.773]$$

pois $\hat{p} = 32/50 = 0.64$, $z_{0.05/2} = z_{0.025} = 1.96$ e $n = 50$.

Como o extremo inferior deste intervalo está muito perto de 50% (o que implica estar em causa ou não uma maioria de utilizadores da internet em casa), poderemos determinar o intervalo de confiança de forma mais rigorosa, pondo em prática o resultado a).

$$\begin{aligned}
 IC_{95\%}(p) &= \left[\frac{0.64 + \frac{1.96^2}{2 \times 50} - \frac{1.96}{\sqrt{50}} \sqrt{0.64(1-0.64) + \frac{1.96^2}{4 \times 50}}}{1 + \frac{1.96^2}{50}}, \frac{0.64 + \frac{1.96^2}{2 \times 50} + \frac{1.96}{\sqrt{50}} \sqrt{0.64(1-0.64) + \frac{1.96^2}{4 \times 50}}}{1 + \frac{1.96^2}{50}} \right] = \\
 &= [0.50140762, 0.75861437]
 \end{aligned}$$

Como se pode ver, não foi muito compensador usar a solução mais precisa.

Outros intervalos de confiança podem ser deduzidos por aplicação do método pivotal e das distribuições de amostragem constantes na lista que se apresenta da secção seguinte. De acordo com essa lista, poderemos estimar por intervalo de confiança:

- A diferença de valores médios de duas populações X e Y , $\mu_X - \mu_Y$;
- O quociente de variância de duas populações X e Y , σ_X^2 / σ_Y^2 ;
- A diferença de proporções em duas populações X e Y , $p_X - p_Y$.

3.5 Distribuições de amostragem

3.5.1 Média amostral, \bar{X}

Os diferentes intervalos de confiança para o valor médio $E(X) = \mu$ de uma população X , são consequência da distribuição que podemos associar ao estimador de μ , ou seja a \bar{X} . Por sua vez, a distribuição de \bar{X} depende do conhecimento que temos acerca da população e da amostra. De forma resumida, o que foi dito na secção anterior acerca da distribuição de amostragem de \bar{X} , é que

Situação A Se a população X tem distribuição Normal com variância $\sigma^2 \equiv V(X)$ conhecida, então

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ ou de modo equivalente } Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1) \quad (3.5.4)$$

Situação B Se a população X tem distribuição Normal com variância $\sigma^2 \equiv V(X)$ desconhecida, então a estatística (v.a.)

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1} \quad (3.5.5)$$

dito de outro modo, a estatística (v.a.) T tem distribuição t com $n - 1$ graus de liberdade.

Situao C Se a populao X tem distribuico *desconhecida ou conhecida mas no normal*, com $\sigma^2 \equiv V(X)$ conhecida, e para uma amostra de dimenso $n \geq 30$, o T.L.C garante que,

$$W = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \stackrel{a}{\sim} N(0, 1) \quad (3.5.6)$$

Situao D Se a populao X tem distribuico *desconhecida ou conhecida mas no normal*, com $\sigma^2 \equiv V(X)$ desconhecida, e para uma amostra de dimenso $n \geq 30$, o T.L.C garante que,

$$W = \sqrt{n} \frac{\bar{X} - \mu}{S} \stackrel{a}{\sim} N(0, 1) \quad (3.5.7)$$

Tabela 3.1: Distribuico de amostragem da mdia amostral, \bar{X}

Situao	Conhecimento de X e da amostra	Distribuico de \bar{X}
A	$X \sim N(\mu, \sigma^2)$ com σ^2 conhecida	$Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$
B	$X \sim N(\mu, \sigma^2)$ com σ^2 desconhecida	$T = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t_{n-1}$
C	$X \sim ?$ com σ^2 conhecida e $n \geq 30$	$W = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \stackrel{a}{\sim} N(0, 1)$
D	$X \sim ?$ com σ^2 desconhecida e $n \geq 30$	$W = \sqrt{n} \frac{\bar{X} - \mu}{S} \stackrel{a}{\sim} N(0, 1)$

3.5.2 Variância amostral, S^2

Para o estimador S^2 da variância da população $\sigma^2 = V(X)$, tem-se

Tabela 3.2: Distribuição de amostragem da variância amostral, S^2

Conhecimento de X e da amostra	Distribuição de S^2
$X \sim N(\mu, \sigma^2)$ com μ desconhecido	$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

3.5.3 Proporção amostral, \hat{P}

Tabela 3.3: Distribuição de amostragem da proporção amostral, \hat{P}

Conhecimento amostra	Distribuição de \hat{P}
$n \geq 30$	$W = \sqrt{n} \frac{\hat{P} - p}{\sqrt{p(1-p)}} \stackrel{a}{\sim} N(0, 1)$
$n \geq 30$	$W = \sqrt{n} \frac{\hat{P} - p}{\sqrt{\hat{P}(1-\hat{P})}} \stackrel{a}{\sim} N(0, 1)$

3.5.4 Diferença de médias de amostras de duas populações, $\bar{X} - \bar{Y}$

Tabela 3.4: Distribuição de amostragem para a diferença de médias amostrais de duas populações

Situação	Condições de aplicação	Distribuição de $\bar{X} - \bar{Y}$
A	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 conhecidas	$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1)$
B	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$ e desconhecidas	$T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2}$
C	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 conhecidas, n_X e $n_Y \geq 30$	$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \stackrel{a}{\sim} N(0, 1)$
D	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 desconhecidas, n_X e $n_Y \geq 30$	$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \stackrel{a}{\sim} N(0, 1)$
$S_p^2 = \frac{(n_X - 1) S_X^2 + (n_Y - 1) S_Y^2}{n_X + n_Y - 2}$		

3.5.5 Quociente de variâncias de amostras de duas populações, S_1^2/S_2^2

Tabela 3.5: Distribuição de amostragem para o quociente de variâncias amostrais de duas populações

Condições de aplicação	Distribuição de S_X^2/S_Y^2
$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ μ_X, μ_Y desconhecidos	$F = \frac{\sigma_Y^2}{\sigma_X^2} \frac{S_X^2}{S_Y^2} \sim F_{(n_X - 1, n_Y - 1)}$

3.5.6 Diferença de proporções amostrais de duas populações, $\hat{P}_X - \hat{P}_Y$

Tabela 3.6: Distribuição de amostragem para a diferença de proporções amostrais de duas populações

Condições de aplicação	Distribuição de $\hat{P}_X - \hat{P}_Y$
$n_X \geq 30$ e $n_Y \geq 30$	$W = \frac{(\hat{P}_X - \hat{P}_Y) - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{n_X} + \frac{p_Y(1-p_Y)}{n_Y}}} \underset{\sim}{\sim} N(0, 1)$

Capítulo 4

Teste de Hipóteses

4.1 Introdução

No capítulo 2 vimos como estimar pontualmente alguns parâmetros de uma população e apresentámos algumas propriedades que permitem analisar a sua precisão (enviesamento, eficiência e consistência). No capítulo 3 ilustrámos como podemos estimar por intervalo de confiança, o parâmetro de uma população, dando especial atenção aos parâmetros populacionais: valor médio, variância, desvio padrão e proporção.

Outro procedimento muito comum em estatística consiste na realização de um teste sobre uma determinada conjectura que se faça sobre a população.

Exemplo 4.1.

- Num determinado departamento pretende-se estudar o n.º X de faltas ao trabalho (de cada funcionário) durante os 5 dias úteis de uma semana. Trata-se de estudar uma população X com distribuição binomial de parâmetros $(5, p)$ e portanto o que falta conhecer acerca desta população será o valor de p . Podemos então tecer as conjecturas: Será que $p = P(\text{falta num dia}) \leq 0.3$ ou será que $p = P(\text{falta num dia}) > 0.3$?
- Num processo de engarrafamento de refrigerante em latas de 33cl, que queremos controlar, podemos conjecturar: Será que o volume médio de refrigerante por garrafa é igual a 33cl, $\mu = 33$ (boas condições de engarrafamento) ou será que o volume médio de refrigerante por garrafa é diferente de 33cl, $\mu \neq 33$ (más condições de engarrafamento).
- Será que a duração de um pneu de uma determinada marca e tipo, tem distribuição exponencial?

Nos dois primeiros exemplos as conjecturas são feitas sobre o valor dos parâmetros da população, ou melhor dizendo sobre o valor dos parâmetros da distribuição da população X . No terceiro exemplo a conjectura é feita sobre a própria distribuição da população X .

As conjecturas que se fazem sobre a população (quer seja sobre os seus parâmetros, quer seja sobre a própria distribuição) designam-se por hipóteses.

Implicitamente e em cada situação, temos sempre duas hipóteses (conjecturas): A *hipótese nula* representada por H_0 e a *hipótese alternativa* representada por H_1 .

Exemplo 4.2. Para os exemplos atrás apresentados as hipóteses (conjecturas) são:

- $H_0 : p \leq 0.3$ vs $H_1 : p > 0.3$

- $H_0 : \mu = 33 \quad vs \quad H_1 : \mu \neq 33$
- $H_0 : X \sim E(0, \delta) \quad vs \quad H_1 : X \approx E(0, \delta)$

Nota importante: A hipótese nula H_0 deverá corresponder a uma situação de *status quo* na população, ou seja uma situação que não corresponda a alterações que sejam necessárias realizar nessa população.

Vejam os dois primeiros exemplos atrás apresentados:

- Se $p = P(\text{falta num dia}) \leq 0.3$, a proporção diária de falta ao trabalho não é gravosa, só o sendo se $p > 0.3$;
- Se o volume médio μ de refrigerante for diferente de 33cl, deverão ser implementadas alterações no processo de engarrafamento, pois a situação *status quo* de engarrafamento normal exige que $\mu = 33cl$.

4.2 Decisão, regra de decisão e estatística de teste

Um teste das hipóteses H_0 vs H_1 , consiste:

- Na avaliação da ‘‘consistência’’ da informação amostral com a conjectura estabelecida na hipótese H_0 ;
- Decidirmo-nos pela **rejeição** ou pela **não rejeição** da hipótese H_0 de acordo com a ‘‘consistência’’ observada;
- Controlarmos a **probabilidade de tomarmos uma decisão errada**, já que esta é sustentada pelos valores amostrais observados.

Como existem sempre duas hipóteses num teste, quando se rejeita H_0 , sabemos imediatamente que se aceita H_1 e quando se não rejeita H_0 sabemos que se rejeita H_1 .

Considere-se A o conjunto de todas as amostras de dimensão n que é possível seleccionar numa população X .

Um teste das hipóteses H_0 vs H_1 , consiste numa **regra** (ou critério) que permita determinar um subconjunto de $R \subset A$ tal que:

- se $(x_1, x_2, \dots, x_n) \in R$, rejeitamos H_0 ;
- se $(x_1, x_2, \dots, x_n) \notin R$, não rejeitamos H_0 ;

Contudo, na maioria das vezes, realizamos um teste de hipóteses recorrendo a uma **estatística de teste** $W \equiv W(X_1, X_2, \dots, X_n)$ que avalia a ‘‘discrepância’’ (ou a ‘‘consistência’’) da amostra aleatória (X_1, X_2, \dots, X_n) com as conjecturas estabelecidas nas hipóteses H_0 e H_1 .

Se considerarmos E o conjunto de todos os valores da estatística T quando aplicada a todas as possíveis amostras de dimensão n (isto é sobre o conjunto A), então

Definição 4.1. Um teste das hipóteses H_0 vs H_1 , consiste numa *regra* (ou *critério*) que permite determinar um subconjunto $R \subset E$ tal que, sendo $t(x_1, x_2, \dots, x_n)$ o valor da estatística T para a amostra observada (x_1, x_2, \dots, x_n) :

- se $w(x_1, x_2, \dots, x_n) \in R$, rejeitamos H_0 ;
- se $w(x_1, x_2, \dots, x_n) \notin R$, não rejeitamos H_0 ;

O subconjunto R é designado por *região de rejeição* ou *região crítica*.

Pelo que até agora foi dito, para a realização de um teste de hipóteses precisamos de:

- uma hipótese nula, H_0 , que se deverá manter válida até haver evidência estatística que o contradiga;
- uma hipótese alternativa, H_1 , que deverá ser adoptada caso se rejeite H_0 ;
- uma estatística de teste, $W \equiv W(X_1, X_2, \dots, X_n)$ que avalie a discrepância entre a informação amostral e a conjectura expressa na hipótese H_0 ;
- uma região de rejeição R .

4.3 Erros de decisão e sua probabilidade

A decisão acerca da rejeição ou não rejeição da hipótese nula, H_0 assenta na informação contida numa amostra (x_1, x_2, \dots, x_n) . Contudo esta amostra é uma das muitas realizações possíveis da amostra aleatória (X_1, X_2, \dots, X_n) da população X . Assim a nossa decisão está condicionada pelo acaso da amostragem e como tal pode conduzir a decisões erradas. Num teste de hipóteses as decisões erradas são:

Definição 4.2. Num teste das hipóteses H_0 vs H_1 podemos cometer os seguintes erros de decisão:

- **erro de tipo I** (ou erro de 1ª espécie): Rejeitar H_0 quando H_0 é verdadeira;
- **erro de tipo II** (ou erro de 2ª espécie): Não rejeitar H_0 quando H_0 é falsa.

No quadro que se segue, esquematizam-se as decisões (correctas e erradas) de um teste de hipóteses.

Tabela 4.1: Decisões e erros num teste de hipóteses

Decisão	H_0 verdadeira	H_0 falsa
Rejeitar H_0	Erro de tipo I	Decisão correcta
Não rejeitar H_0	Decisão correcta	Erro de tipo II

Existindo sempre uma possibilidade de cometermos estes erros de decisão, podemos associar-lhes a probabilidade de ocorrerem.

Essas probabilidades são:

$$\gamma = P(\text{erro de tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

Ao valor máximo desta probabilidade dá-se o nome de *nível de significância*, habitualmente é representado por α ,

$$\alpha = \max P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

e

$$\beta = P(\text{erro de tipo II}) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa})$$

A

$Q = 1 - \beta = 1 - P(\text{erro de tipo II}) = 1 - P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa})$
dá-se o nome de *potência* (função potência) do teste.

O teste óptimo será aquele em as probabilidades associadas aos dois tipos de erro têm um valor mínimo. Contudo, é matematicamente impossível minimizá-las simultaneamente. De facto, quando a P (erro de tipo I) diminui, a P (erro de tipo II) aumenta e vice-versa.

No que se segue, os testes que realizamos incluem-se nos denominados *testes de significância*, ou seja os testes em que o nível de significância α ,

$$\alpha = \max P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

é estabelecido por nós (e portanto **tem um valor fixo e conhecido**) e para os quais a função potência $Q = 1 - \beta$ tem valor máximo (ou equivalentemente, **β tem valor mínimo**).

Para melhor exposição dos conceitos, vamos começar por abordar os intitulados **testes paramétricos** isto é, aqueles em que as hipóteses incidem sobre o valor θ de um parâmetro que caracteriza a distribuição da população X e tem valor desconhecido.

Formalizando o problema, consideremos que θ pode assumir valores num conjunto Θ (chamado **espaço parâmetro**). Sejam Θ_0 e Θ_1 dois subconjuntos de Θ tais que:

$$\Theta_0 \cup \Theta_1 = \Theta \quad \text{e} \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

As hipóteses sobre os valores do parâmetro θ podem ser apresentadas por:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1,$$

ou

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \notin \Theta_0.$$

Exemplo 4.3. *Continuando com o exemplo 4.2,*

- O parâmetro p tem por espaço parâmetro $\Theta =]0, 1[$. Na hipótese H_0 , $\Theta_0 =]0, 0.3]$ e na hipótese H_1 , $\Theta_1 =]0.3, 1[$.
- Um valor médio μ tem por espaço parâmetro a recta real \mathbb{R} . Assim as hipóteses podem ser escritas $H_0 : \mu \in \{33\}$ vs $H_1 : \mu \in \mathbb{R} \setminus \{33\}$. Neste caso $\Theta = \mathbb{R}$, $\Theta_0 = \{33\}$ e $\Theta_1 = \mathbb{R} \setminus \{33\}$

4.4 Metodologia para realização de um teste de hipóteses paramétricas

Para testarmos as hipóteses:

$$H_0 : \theta \in \Theta_0 \text{ vs } H_1 : \theta \notin \Theta_0.$$

com um nível de significância α , aconselhamos a seguinte metodologia:

- Escolher um estimador $\hat{\Theta}$ para θ .
- Conhecer ou propor uma estatística de teste $W \equiv W(X_1, \dots, X_n)$ que meça a “discrepância” entre o valor de $\hat{\Theta}$ e o valor de θ . Essa estatística W é designada por estatística de teste.

Deduzir ou conhecer a distribuição da estatística de teste.

- Para o nível α de significância escolhido, determinar uma região de rejeição da hipótese H_0 , R_α .
- Face a uma amostra observada (x_1, x_2, \dots, x_n) , calcular o valor observado da estatística de teste $w_{obs} = W(x_1, x_2, \dots, x_n)$ e decidir:
 - Rejeitar H_0 se $w_{obs} = W(x_1, x_2, \dots, x_n) \in R_\alpha$;
 - Não rejeitar H_0 se $w_{obs} = W(x_1, x_2, \dots, x_n) \notin R_\alpha$.

Para cada teste paramétrico que a seguir expomos, iremos propor um estimador para o parâmetro, escolher a estatística de teste $W \equiv W(X_1, X_2, \dots, X_n)$ e indicar a respectiva distribuição de amostragem, determinar a região de rejeição R_α , para um nível de significância α fixo, após o que será possível tomar uma decisão face a uma amostra observada (x_1, x_2, \dots, x_n) .

4.5 p -value ou valor- p

Com a evolução das ferramentas de cálculo, é hoje possível determinar probabilidades de modo expedito e cómodo. Por isso, é agora usual associar e tomar decisões sobre um teste de hipóteses através do conceito de p -value.

Definição 4.3. *Seja (x_1, x_2, \dots, x_n) a concretização de uma amostra aleatória (X_1, X_2, \dots, X_n) e*

$$w_{obs} = W(x_1, x_2, \dots, x_n)$$

o valor observado da estatística de teste. Designa-se por p -value (ou valor- p), a probabilidade de se observarem valores da estatística de teste tão ou mais favoráveis à rejeição de H_0 do que o observado w_{obs} , admitindo que H_0 é verdadeira.

DECISÃO EM FUNÇÃO DO p -value: O p -value é uma medida da concordância entre a hipótese H_0 e as amostras que possamos recolher e que sejam tão ou mais favoráveis à rejeição de H_0 . Quanto menor for o p -value, menor é a consistência da validade de H_0 .

Assim, se p -value $< \alpha$, devemos rejeitar H_0 ao nível de significância α .

4.6 Erros de decisão e sua probabilidade

Em particular para um teste paramétrico das hipóteses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$,

- $\gamma(\theta) = P(\text{erro de tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(W \in R | \theta \in \Theta_0)$
- $\alpha = \max_{\theta \in \Theta_0} \gamma(\theta)$
- $\beta(\theta) = P(\text{erro de tipo II}) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa}) = P(W \notin R | \theta \notin \Theta_0)$
- Potência do teste: $Q(\theta) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa}) = 1 - \beta(\theta) = P(W \in R | \theta \notin \Theta_0)$

4.7 Teste de hipóteses para o valor médio

Nesta secção vamos dar atenção a hipóteses que estabelecem conjecturas sobre o valor médio $E(X) = \mu$ de uma população X .

4.7.1 Teste de hipóteses bilateral para o valor médio

Exemplo 4.4. Estudos sobre o custo de vida, realizados no mês de Janeiro de 2003, permitiram concluir que o gasto semanal em alimentação de famílias com dois filhos, apresentava um valor médio de 100 euros com um desvio padrão de 15 euros. No mês de Agosto do mesmo ano, pretendíamos saber se tinham ocorrido alterações no gasto semanal médio em alimentação das mesmas famílias. Para tal seleccionou-se uma amostra de gastos semanais em alimentação de 25 famílias (com 2 filhos), que revelou uma média $\bar{x} = 108$ euros.

Que conclusões podemos retirar acerca da alteração do gasto médio semanal em alimentação deste tipo de famílias?

A população em estudo é X -gasto semanal em alimentação das famílias com 2 filhos, mas o interesse primordial diz respeito a $\mu = E(X)$ -gasto médio semanal em alimentação das famílias com 2 filhos. A nossa questão reside em saber se μ permanece igual a 100 euros, $\mu = 100$, ou, se em Agosto, μ é diferente de 100 euros, $\mu \neq 100$.

Queremos então testar a validade das hipóteses

$$H_0 : \mu = 100 \quad \text{vs} \quad H_1 : \mu \neq 100$$

Ao teste de hipóteses do tipo

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

dá-se o nome de teste de hipóteses bilateral para o valor médio, μ .

Neste exemplo, $\mu_0 = 100$.

A decisão acerca da validade de alguma destas hipóteses deverá ser feita à custa da informação que a amostra fornecer. Uma vez que as hipóteses dizem respeito ao valor médio da população, devemos considerar a informação que a amostra fornecer sobre μ . Mas já sabemos que a informação amostral sobre μ , pode ser obtida através de um estimador de μ . Se adoptarmos para estimador de μ , a média amostral,

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

O valor de \bar{X} vai-nos permitir decidir se $\mu \neq 100$ ou se $\mu = 100$, isto é, vai-nos permitir decidir se rejeitamos H_0 ou se não rejeitamos H_0 . Como tal, só nos resta saber, quais os valores de \bar{X} que nos levam a rejeitar H_0 ou a não rejeitar H_0 . Em resumo, precisamos de uma regra de decisão.

Regra de decisão

Se \bar{X} tiver um valor muito diferente (ou distante) de 100, é natural que se decida que $\mu \neq 100$. Podemos dizer que \bar{X} é muito diferente de 100, se $|\bar{X} - 100|$ for muito grande, ou seja se o valor de $|\bar{X} - 100|$ ultrapassar uma certa quantidade a ($a > 0$). Então

$$\text{Rejeitamos } \mu = 100 \quad \text{se} \quad |\bar{X} - 100| > a \quad (a > 0)$$

ou de modo equivalente

$$\text{Rejeitamos } H_0 \quad \text{se} \quad |\bar{X} - 100| > a \quad (a > 0)$$

No caso geral do teste de hipóteses bilateral

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

$$\text{Rejeitamos } H_0 \quad \text{se} \quad |\bar{X} - \mu_0| > a \quad (a > 0)$$

Neste exemplo, o que são os erros de decisão?

Admitamos as seguintes situações:

1. Em Agosto, o gasto médio semanal em alimentação permanece igual a 100 euros, $\mu = 100$. Isto é o que acontece na população, mas nós não o sabemos porque não analisamos a população na totalidade.

Suponhamos que o acaso da amostragem, levava a que se obtivessem valores amostrais sobre o gasto semanal em alimentação, muito elevados (muito pequenos). Então \bar{X} teria um valor elevado (pequeno), e de tal modo elevado (pequeno) que $|\bar{X} - 100| > a$. Como consequência, iríamos decidir rejeitar H_0 , ou seja, decidir que $\mu \neq 100$.

A nossa decisão seria errada, porque (baseados na amostra) decidíamos que $\mu \neq 100$ e de facto $\mu = 100$. Estaríamos a cometer um erro de tipo I, nomeadamente a rejeitar $H_0 : \mu = 100$, quando H_0 é verdadeira.

2. Em Agosto, o gasto médio semanal em alimentação sofreu uma alteração e passou a ter um valor $\mu \neq 100$. Isto é o que acontece na população, mas nós não o sabemos porque não analisamos a população na totalidade.

Suponhamos que a média amostral \bar{X} exibia um valor não muito diferente de 100, de tal modo que $|\bar{X} - 100| \leq a$. Como consequência, iríamos decidir não rejeitar $H_0 : \mu = 100$, ou seja, decidir que o gasto médio semanal continuava igual a 100.

Esta decisão seria errada, porque (baseados na amostra) decidíamos que $\mu = 100$ e de facto $\mu \neq 100$. O erro cometido era um erro de tipo II, nomeadamente não rejeitar H_0 , quando H_0 é falsa.

Probabilidade dos erros de decisão

As probabilidades dos erros de decisão são, neste caso

$$\alpha = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = P(|\bar{X} - 100| > a | \mu = 100) \quad (\text{nível de significância})$$

$$\beta(\mu) = P(\text{Não rejeitar } H_0 | H_0 \text{ falsa}) = P(|\bar{X} - 100| \leq a | \mu \neq 100)$$

NOTA: O teste que agora expomos, é um teste que minimiza $\beta(\mu)$, para cada α (nível de significância) que escolhermos.

Os níveis de significância mais usados são $\alpha = 0.1 = 10\%$ para uma decisão *pouco significativa*, $\alpha = 0.05 = 5\%$ para uma decisão *significativa* e $\alpha = 0.01 = 1\%$ para uma decisão *altamente significativa*.

Região de rejeição ou região crítica

Consideremos as hipóteses genéricas para um teste bilateral sobre o valor médio,

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

Já vimos que podemos

$$\text{Rejeitar } H_0 \text{ se } |\bar{X} - \mu_0| > a \quad (a > 0)$$

Mas qual o valor de a ?

Admitamos que escolhíamos um nível de significância α para o nosso teste. Então

$$\alpha = P(|\bar{X} - \mu_0| > a | \mu = \mu_0)$$

Trata-se de uma probabilidade cujo valor conhecemos. O que desconhecemos é o valor de a . Mas se soubermos qual a distribuição da v.a. \bar{X} , podemos trabalhar esta igualdade sobre probabilidades e portanto deduzir o valor de a .

Suponhamos que a população goza das seguintes características:

X tem distribuição normal de valor médio μ e variância conhecida, $\sigma^2 = V(X)$, $X \sim N(\mu, \sigma^2)$

Então a nossa amostra aleatória (X_1, \dots, X_n) é constituída por v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$ e portanto \bar{X} tem distribuição normal de valor médio μ e variância σ^2/n , $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, isto é $Z = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$.

Quando $H_0 : \mu = \mu_0$ é verdadeira, $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$

Agora já podemos determinar o valor de a .

$$\begin{aligned} \alpha &= P(|\bar{X} - \mu_0| > a | \mu = \mu_0) = P\left(\left|\sqrt{n} \frac{\bar{X} - \mu_0}{\sigma}\right| > \sqrt{n} \frac{a}{\sigma}\right) = P\left(|Z| > \sqrt{n} \frac{a}{\sigma}\right) = \\ &= P\left(Z < -\sqrt{n} \frac{a}{\sigma}\right) + P\left(Z > \sqrt{n} \frac{a}{\sigma}\right) = \Phi\left(-\sqrt{n} \frac{a}{\sigma}\right) + 1 - \Phi\left(\sqrt{n} \frac{a}{\sigma}\right) = \\ &= 1 - \Phi\left(\sqrt{n} \frac{a}{\sigma}\right) + 1 - \Phi\left(\sqrt{n} \frac{a}{\sigma}\right) = 2 - 2\Phi\left(\sqrt{n} \frac{a}{\sigma}\right) = 2\left(1 - \Phi\left(\sqrt{n} \frac{a}{\sigma}\right)\right) \end{aligned}$$

ou seja

$$\Phi\left(\sqrt{n} \frac{a}{\sigma}\right) = 1 - \frac{\alpha}{2} \Leftrightarrow \sqrt{n} \frac{a}{\sigma} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = z_{\alpha/2} \Leftrightarrow a = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Regra de decisão para um nível de significância α

$$\text{Rejeitar } H_0 \text{ se } |\bar{X} - \mu_0| > \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

ou de modo equivalente

$$\text{Rejeitar } H_0 \text{ se } \left| \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \right| > z_{\alpha/2}$$

NOTAS:

- Repare que conseguimos deduzir o valor de \underline{a} porque sabemos as características da população e portanto conseguimos saber qual a distribuição de \bar{X} . Repare também que este conhecimento das características da população X corresponde à **situação A** descrita na secção 3.5.1.

- $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma}$ é a estatística de teste.

- A região de rejeição, para um nível de significância α , é $R_\alpha \equiv]-\infty, -z_{\alpha/2} [\cup] z_{\alpha/2}, +\infty [$.

- A regra de decisão, para um nível de significância α será:

– Rejeitar H_0 caso $z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$;

– Não rejeitar H_0 caso $z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \notin R_\alpha$

- Quanto ao p -value, ter-se-á

$$p\text{-value} = P(|Z| > |z_{obs}| | \mu = \mu_0) = 2P(Z > |z_{obs}|),$$

com $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$.

Regra de decisão para um nível de significância α

$$\text{Estatística de teste: } Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$$

$$\text{Região de rejeição: } R_\alpha =]-\infty, -z_{\alpha/2} [\cup] z_{\alpha/2}, +\infty [$$

$$\text{Rejeitar } H_0 \text{ se } z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$$

$$p\text{-value} = 2(P(Z > |z_{obs}|))$$

Exemplo 4.5. Continuação do exemplo 4.4

As hipóteses são

$$H_0 : \mu = 100 \quad \text{vs} \quad H_1 : \mu \neq 100$$

Sabemos que $\sigma = 15$ e que a amostra de $n = 25$ observações forneceu $\bar{x} = 108$.

Admitindo que o gasto semanal em alimentação das famílias com 2 filhos tem distribuição normal,

$$\text{Rejeitamos } H_0 \text{ se } \left| \sqrt{25} \frac{108 - 100}{15} \right| = 2.667 > z_{\alpha/2}$$

Se escolhermos um nível de significância $\alpha = 5\%$ (para uma decisão *significante*),

$$z_{0.05/2} = z_{0.025} = \Phi^{-1}(0.975) = 1.96$$

e como

$$\left| \sqrt{25} \frac{108 - 100}{15} \right| = 2.667 > 1.96 = z_{0.025}$$

decidimos rejeitar $H_0 : \mu = 100$, ao nível de 5% de significância, ou seja, com 5% de significância, concluímos que ocorreram alterações no gasto médio semanal em alimentação das famílias com 2 filhos.

Cálculo e decisão pelo p -value

$$\text{Sendo } z_{obs} = \sqrt{25} \frac{108 - 100}{15} = 2.667,$$

$$p\text{-value} = P(|Z| > |2.667|) = 2P(Z > 2.667) = 2(1 - P(Z \leq 2.667)) = 2(1 - 0.9962) = 0.0076$$

Dado que $p\text{-value} < 0.05$, decidimos rejeitar $H_0 : \mu = 100$, ao nível de 5% de significância.

Outros testes de hipóteses bilateral para o valor médio

Como foi dito na nota importante, a regra de decisão atrás deduzida dependeu do conhecimento das características da população X , nesse caso de $X \sim N(\mu, \sigma^2)$.

Quando esse conhecimento é diferente, a regra de decisão altera-se, mas a alteração depende essencialmente da distribuição que resulta para \bar{X} . Reçamos então nas situações expostas na secção 3.5.1 e portanto as regras de decisão para um teste bilateral para o valor médio μ , com um nível de significância α , vão ser:

Tabela 4.2: Testes de hipóteses bilateral para o valor médio

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

Situação	Conhecimento de X e da amostra	
A	$X \sim N(\mu, \sigma^2)$ σ^2 conhecida	Estatística de teste: $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$ $p\text{-value} = 2 \times P(Z > z_{obs})$
B	$X \sim N(\mu, \sigma^2)$ σ^2 desconhecida	Estatística de teste: $T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \sim t_{n-1}$ Região de rejeição: $R_\alpha =]-\infty, -t_{n-1: \alpha/2}[\cup]t_{n-1: \alpha/2}, +\infty[$ Rejeitar H_0 se $t_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} = 2 \times P(T > t_{obs})$
C	$X \sim ?$ σ^2 conhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$ $p\text{-value} \approx 2 \times P(Z > z_{obs})$
D	$X \sim ?$ com σ^2 desconhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} \approx 2 \times P(Z > z_{obs})$

Figura 4.1: Teste bilateral para o valor médio: Situações A, C e D

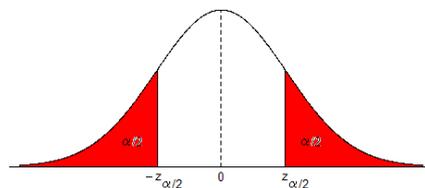
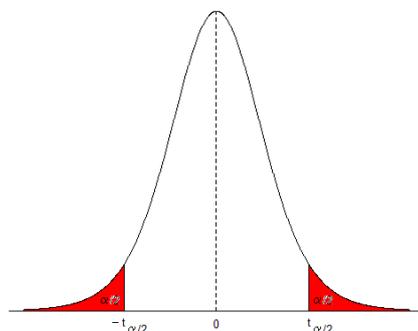


Figura 4.2: Teste bilateral para o valor médio: Situação B



Exemplo 4.6. *Medições de acidez (pH) de amostras de chuva foram registadas em 12 locais de uma região industrial:*

5.1 5.0 3.8 4.8 3.6 4.7
4.3 4.4 4.5 4.9 4.7 4.8

Por estudos anteriores sabe-se que os registos de acidez da chuva nesta região têm distribuição normal.

Poderemos concluir, com 5% de significância, que os níveis actuais de acidez média da chuva saem fora do valor de controlo de 4.5 de acidez média na região?

Pretendemos testar, com $\alpha = 5\%$, as hipóteses

$$H_0 : \mu = 4.5 \quad \text{vs} \quad H_1 : \mu \neq 4.5$$

sendo μ o nível de acidez média da chuva na região.

A amostra possibilita a seguinte informação:

$$n = 12 \quad \bar{x} = \frac{54.6}{12} = 4.55 \quad s^2 = \frac{2.35}{11} = 0.213637$$

O conhecimento da população corresponde à situação C descrita na secção 3.5.1, pelo que a estatística de teste

$$T = \sqrt{n} \frac{\bar{X} - 4.5}{S} \underset{\mu=4.5}{\sim} t_{12-1}$$

Assim a regra de decisão será:

$$\text{Rejeitar } H_0 \text{ se } \left| \sqrt{n} \frac{\bar{X} - \mu_0}{S} \right| > t_{n-1;\alpha/2}$$

Ora $\mu_0 = 4.5$ e para $\alpha = 5\%$, $t_{n-1;\alpha/2} = t_{11;0.025} = 2.201$.

Como $\left| \sqrt{12} \frac{4.55 - 4.5}{\sqrt{0.213637}} \right| = 0.3747 < 2.201$, concluímos que a acidez média da chuva nesta região permanece igual ao valor de controlo de 4.5, com uma significância de 5% nesta decisão.

Cálculo e decisão pelo p -value

$$\text{Sendo } T = \sqrt{n} \frac{\bar{X} - 4.5}{S} \underset{\mu=4.5}{\sim} t_{11} \text{ a estatística de teste e } t_{\text{obs}} = \sqrt{12} \frac{4.55 - 4.5}{\sqrt{0.213637}} = 0.3747,$$

$$p\text{-value} = P(|T| > |0.3747|) = 2P(T > 0.3747) = 2 \times 0.3575 = 0.715$$

Dado que $p\text{-value} > 0.05$, decidimos não rejeitar $H_0 : \mu = 4.5$, ao nível de 5% de significância.

Regra de decisão para um nível de significância α

$$\text{Hipóteses: } H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

$$\text{Estatística de teste: } T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu = \mu_0}{\sim} t_{n-1}$$

$$\text{Região de rejeição: } R_\alpha = [-\infty, -t_{n-1:\alpha/2} [\cup] t_{n-1:\alpha/2}, +\infty [$$

$$\text{Rejeitar } H_0 \text{ se } t_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$$

$$p\text{-value} = 2 \min(P(T > |t_{obs}|))$$

4.7.2 Teste de hipóteses unilateral direito para o valor médio

As hipóteses dum teste bilateral sobre o valor médio μ conjecturam se o valor médio de uma população X tem um valor μ_0 ou se ocorreram alterações e o seu valor actual é diferente de μ_0 . Mas, por vezes tem mais interesse saber se essas alterações ocorreram no sentido do valor de μ ser agora maior que μ_0 .

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

Veja-se o seguinte exemplo:

Exemplo 4.7. *Anuncia-se que um novo tratamento é mais eficaz que o tratamento tradicional para prolongar a vida de doentes em estado terminal sofrendo de cancro. O tratamento tradicional já é usado à algum tempo e sabe-se que a sua aplicação provoca um tempo médio de 4.2 anos de sobrevivência com um desvio padrão de 1.1 anos.*

O novo tratamento foi administrado a 80 pacientes e os tempos registados de sobrevivência à doença desde o começo do tratamento exibiram uma média amostral de 4.5 anos.

Será que esta informação corrobora o anúncio feito ao novo tratamento?

As conjecturas em causa são

$$H_0 : \mu \leq 4.2 \quad \text{vs} \quad H_1 : \mu > 4.2$$

sendo μ o tempo médio de sobrevivência desde o início de um tratamento.

Naturalmente a regra de decisão passa por rejeitarmos a hipótese $H_0 : \mu \leq 4.2$ se $(\bar{X} - 4.2)$ for muito grande, isto é se $(\bar{X} - 4.2) > b$ com $b > 0$.

Mas qual a distribuição de \bar{X} ? Se considerarmos que o desvio padrão se mantém com o valor de $\sigma = 1.1$ anos, estamos no caso da situação C descrita na secção 3.5.1, porque não se conhece a distribuição da população X -tempo de sobrevivência desde o início de um tratamento, mas se conhece a sua variância e se tem uma amostra de dimensão $n = 80 \geq 30$. Portanto, podemos dizer que a estatística de teste,

$$W = \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \text{ tem distribuição aproximada } N(0, 1)$$

ou seja, que

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

Então

$$\begin{aligned}
 \gamma(\mu) &= P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = P(\bar{X} - 4.2 > b | \mu \leq 4.2) = \\
 &= P\left(\sqrt{n} \frac{\bar{X} - 4.2}{\sigma} > \frac{\sqrt{n}}{\sigma} b | \mu \leq 4.2\right) = \\
 &= P\left(\sqrt{n} \frac{\bar{X} - \mu + \mu - 4.2}{\sigma} > \frac{\sqrt{n}}{\sigma} b | \mu \leq 4.2\right) = \\
 &= P\left(W + \sqrt{n} \frac{\mu - 4.2}{\sigma} > \frac{\sqrt{n}}{\sigma} b | \mu \leq 4.2\right) = \\
 &= P\left(W > \frac{\sqrt{n}}{\sigma} (b - \mu + 4.2) | \mu \leq 4.2\right) \approx \\
 &\approx P\left(Z > \frac{\sqrt{n}}{\sigma} (b - \mu + 4.2) | \mu \leq 4.2\right)
 \end{aligned}$$

$\gamma(\mu)$ é uma função crescente para $\mu \in]-\infty, 4.2]$

O nível de significância α , corresponderá a

$$\begin{aligned}
 \alpha &= \max_{\mu \leq 4.2} P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = \max_{\mu \leq 4.2} \gamma(\mu) = \gamma(4.2) \approx \\
 &\approx P\left(Z > \frac{\sqrt{n}}{\sigma} b\right) \approx 1 - \Phi\left(\frac{\sqrt{n}}{\sigma} b\right)
 \end{aligned}$$

Se estabelecermos um valor para α , então

$$\alpha \approx 1 - \Phi\left(\frac{\sqrt{n}}{\sigma} b\right) \Rightarrow \frac{\sqrt{n}}{\sigma} b \approx \Phi^{-1}(1 - \alpha) \Rightarrow \frac{\sqrt{n}}{\sigma} b \approx z_\alpha$$

Regra de decisão para um nível de significância α

$$\text{Rejeitar } H_0 \text{ se } \bar{X} - 4.2 > \frac{\sigma}{\sqrt{n}} z_\alpha$$

ou de modo equivalente

$$\text{Rejeitar } H_0 \text{ se } \sqrt{n} \frac{\bar{X} - 4.2}{\sigma} > z_\alpha$$

Para a amostra observada e para os valores populacionais conhecidos: $n = 80$, $\bar{x} = 4.5$, $\sigma = 1.1$.

Se considerarmos um nível de significância $\alpha = 10\%$, $z_{0.1} = \Phi^{-1}(0.9) = 1.28$,

$$z_{obs} = \sqrt{80} \frac{\bar{X} - 4.2}{1.1} = 2.4393 > 1.28 = z_{0.1}$$

peço que, decidimos que a amostra corrobora o anúncio de que o novo tratamento prolonga a vida dos doentes, com uma significância de 10% na decisão.

- $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu = \mu_0}{\underset{\sim}{\sim}} N(0, 1)$ é a estatística de teste.

- A região de rejeição, para um nível de significância α , é $R_\alpha \equiv]z_\alpha, +\infty[$.
- A regra de decisão, para um nível de significância $\alpha = 10\%$ será a de rejeitar H_0 caso $z_{obs} = \sqrt{n} \frac{\bar{x} - 4.2}{\sigma} \in R_{0.1} \equiv]1.28, +\infty[$.
- Como $z_{obs} = 2.4393 \in R_{0.1}$, decidimos rejeitar H_0 .
- Quanto ao p -value, ter-se-á

$$p\text{-value} = P(Z > z_{obs}) = P(Z > 2.44) = 1 - 0.9927 = 0.0073,$$

com $Z \stackrel{a}{\underset{\mu=4.2}{\sim}} N(0, 1)$. Assim $p\text{-value} < 0.1$ (α) permite-nos decidir pela rejeição de H_0 ao nível de 10% de significância.

NOTA IMPORTANTE: Repare que conseguimos deduzir o valor de \hat{b} porque sabemos as características da população e portanto conseguimos saber qual a distribuição de \bar{X} . Repare também que este conhecimento das características da população X corresponde à **situação C** descrita na secção 3.5.1.

Para esta e outras situações referentes ao conhecimento da população e da amostra tem-se a título de resumo:

Figura 4.3: Teste unilateral direito para o valor médio: Situações A, C e D

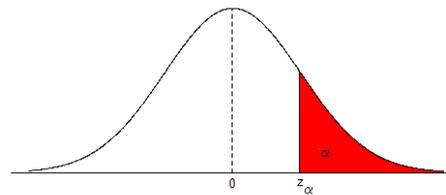


Figura 4.4: Teste unilateral direito para o valor médio: Situação B

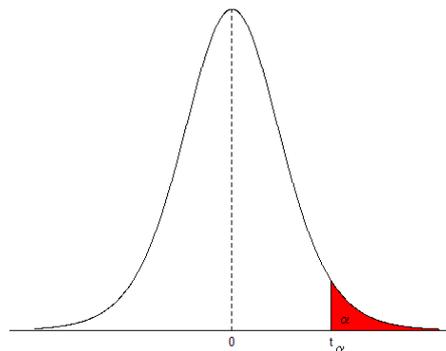


Tabela 4.3: Testes de hipóteses unilateral direito para o valor médio

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

Situação	Conhecimento de X e da amostra	
A	$X \sim N(\mu, \sigma^2)$ σ^2 conhecida	Estatística de teste: $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha =]z_\alpha, +\infty[$ Rejeitar H_0 se $z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$ $p\text{-value} = P(Z > z_{obs})$
B	$X \sim N(\mu, \sigma^2)$ σ^2 desconhecida	Estatística de teste: $T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\sim} t_{n-1}$ Região de rejeição: $R_\alpha =]t_{n-1; \alpha}, +\infty[$ Rejeitar H_0 se $t_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} = P(T > t_{obs})$
C	$X \sim ?$ σ^2 conhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]z_\alpha, +\infty[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} \approx P(Z > w_{obs})$
D	$X \sim ?$ σ^2 desconhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]z_\alpha, +\infty[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} \approx P(Z > w_{obs})$

4.7.3 Teste de hipóteses unilateral esquerdo para o valor médio

Mas também pode ter interesse saber se as alterações de μ ocorrem no sentido do seu valor ser menor que μ_0 , quando antes era $\geq \mu_0$.

$$H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

Vejamos o seguinte exemplo:

Exemplo 4.8. Num processo de fabrico de placas de vidro, produzem-se bolhas que se distribuem aleatoriamente pelas placas. Com base na abundante informação recolhida pelo departamento de qualidade, a densidade média das bolhas estimava-se, até há pouco tempo, em 0.4 bolhas/ m^2 .

Recentemente fez-se uma tentativa de melhorar o processo produtivo, em particular no tocante ao aparecimento deste tipo de defeito. Depois de serem introduzidas alguma alterações no processo de fabrico, recolheu-se uma amostra constituída por 15 placas de $4.5 m^2$, e registou-se o número de bolhas em cada uma delas. A média da amostra foi de $\bar{x} = 0.317$ bolhas/ m^2 e o desvio padrão amostral foi de $s = 0.2254$ bolhas/ m^2 .

Verifiquemos, ao nível de significância de 5%, se a densidade esperada de bolhas por m^2 diminuiu.

Se μ representar a densidade média de bolhas/ m^2 , as hipóteses que estão em causa são:

$$H_0 : \mu \geq 0.4 \quad \text{vs} \quad H_1 : \mu < 0.4$$

Face à presente hipótese nula $H_0 : \mu \geq 0.4$, a regra de decisão mais natural passa por rejeitarmos a hipótese $H_0 : \mu \geq 0.4$ se $(\bar{X} - 0.4)$ for muito menor que 0, isto é se $(\bar{X} - 0.4) < -c$ com $c > 0$.

Mas qual a distribuição de \bar{X} ? Se considerarmos que a distribuição da população X -nº de bolhas/m² tem distribuição normal, desconhecemos a sua variância e portanto estamos no caso da situação C descrita na secção 3.5.1. Assim, a nossa estatística de teste e a sua distribuição (quando $\mu = 0.4$) são:

$$T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\sim} t_{15-1}$$

Para um nível de significância α ,

$$\begin{aligned} \alpha &= \max P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = \max_{\mu \geq 0.4} P(\bar{X} - 0.4 < -c | \mu \geq 0.4) = \\ &= P(\bar{X} - 0.4 < -c | \mu = 0.4) = P\left(\sqrt{n} \frac{\bar{X} - 0.4}{S} < -\sqrt{n} \frac{c}{S}\right) = P\left(T < -\sqrt{n} \frac{c}{S}\right) \end{aligned}$$

$$\text{Logo } \alpha = P\left(T < -\sqrt{n} \frac{c}{S}\right) \Leftrightarrow \alpha = P\left(T \geq \sqrt{n} \frac{c}{S}\right) \Rightarrow \sqrt{n} \frac{c}{S} = t_{n-1;\alpha}$$

Regra de decisão para um nível de significância α

$$\text{Rejeitar } H_0 \text{ se } \bar{X} - 0.4 < -\frac{S}{\sqrt{n}} t_{n-1;\alpha}$$

ou de modo equivalente

$$\text{Rejeitar } H_0 \text{ se } \sqrt{n} \frac{\bar{X} - 0.4}{S} < -t_{n-1;\alpha}$$

Como $n = 15$, $\bar{x} = 0.317$, $s = 0.2254$ e, para $\alpha = 5\%$, $t_{n-1;\alpha} = t_{14;0.05} = 1.76$

$$t_{obs} = \sqrt{n} \frac{\bar{x} - 0.4}{s} = \sqrt{15} \frac{0.317 - 0.4}{0.2254} = -1.42617 > -1.76 = -t_{14;0.05}$$

decidimos não rejeitar H_0 ao nível de significância de 5%, ou melhor dizendo, decidimos que a densidade esperada de bolhas/m² não parece diminuir, sendo de 5% a significância desta conclusão.

Em resumo, e aplicando a metodologia proposta para a realização de um teste de hipóteses paramétrico, tem-se

- Estimador da média μ : a média amostral, \bar{X} .
- $T = \sqrt{n} \frac{\bar{X} - 0.4}{S} \underset{\mu=0.4}{\sim} t_{14}$ é a estatística de teste e a sua distribuição quando $\mu = 0.4$.
- A região de rejeição, para um nível de significância $\alpha = 5\%$, é $R_\alpha \equiv]-\infty, -t_{14;\alpha}[=]-\infty, -1.76[$.
- A regra de decisão, para um nível de significância $\alpha = 5\%$ será a de rejeitar H_0 caso $t_{obs} = \sqrt{n} \frac{\bar{x} - 0.4}{s} \in R_{0.05} \equiv]-\infty, -1.76[$.
- Como $t_{obs} = -1.42617 \notin R_{0.05}$, decidimos não rejeitar H_0 .

- Quanto ao p – value, ter-se-á

$$p\text{-value} = P(T < t_{obs}) = P(T < -1.42617) = P(T > 1.42617) = 0.0879,$$

Assim $p\text{-value} > 0.05$ (α) permite-nos decidir pela não rejeição de H_0 ao nível de 5% de significância.

NOTA IMPORTANTE: Repare que conseguimos deduzir o valor de \underline{c} porque soubemos as características da população e portanto conseguimos saber qual a distribuição de \bar{X} . Repare também que este conhecimento das características da população X corresponde à **situação B** descrita na secção 3.5.1.

Para esta e outras situações referentes ao conhecimento da população e da amostra tem-se a título de resumo:

Tabela 4.4: Testes de hipóteses unilateral esquerdo para o valor médio

$$H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

Situação	Conhecimento de X e da amostra	
A	$X \sim N(\mu, \sigma^2)$ σ^2 conhecida	Estatística de teste: $Z = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_\alpha[$ Rejeitar H_0 se $z_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{\sigma} \in R_\alpha$ $p\text{-value} = P(Z < z_{obs})$
B	$X \sim N(\mu, \sigma^2)$ σ^2 desconhecida	Estatística de teste: $T = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\sim} t_{n-1}$ Região de rejeição: $R_\alpha =]-\infty, -t_{n-1; \alpha}[$ Rejeitar H_0 se $t_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} = P(T < t_{obs})$
C	$X \sim ?$ σ^2 conhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_\alpha[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} \approx P(Z < w_{obs})$
D	$X \sim ?$ σ^2 desconhecida e $n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \underset{\mu=\mu_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_\alpha[$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\bar{x} - \mu_0}{s} \in R_\alpha$ $p\text{-value} \approx P(Z < w_{obs})$

Figura 4.5: Teste unilateral esquerdo para o valor médio: Situações A, C e D

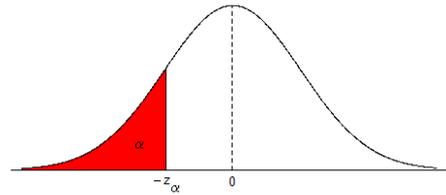
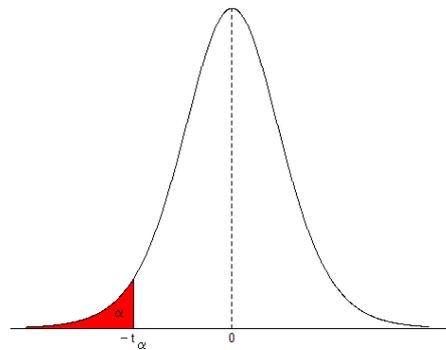


Figura 4.6: Teste unilateral esquerdo para o valor médio: Situação B



4.8 Teste de hipóteses para a variância

Nesta secção vamos dedicar a atenção exclusivamente a hipóteses que estabelecem conjecturas sobre a variância $\sigma^2 = V(X)$ de uma população X .

Os procedimentos e os conceitos são similares aos utilizados nas deduções dos testes para o valor médio.

Os pressupostos a estabelecer sobre a amostra aleatória são:

1. Considerar uma amostra aleatória (X_1, \dots, X_n) de dimensão n da população X ;
2. A população X ter uma distribuição **Normal** com valor médio μ (de valor desconhecido e estimado por \bar{X}) e variância σ^2 que se pretende estimar.

Vamos adoptar o estimador centrado da variância σ^2 ,

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

que, a ser satisfeita a condição 2. sobre a normalidade da população, a estatística

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

X^2 tem distribuição do Qui-Quadrado com $n-1$ graus de liberdade.

4.8.1 Teste de hipóteses bilateral para a variância

Consideremos as hipóteses

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

Naturalmente que devemos rejeitar a hipótese $\sigma^2 = \sigma_0^2$ se a amostra nos fornecer uma estimativa S^2 muito "diferente" de σ_0^2 . Dito de outro modo, se o quociente $\frac{S^2}{\sigma_0^2}$ for muito pequeno ou se for muito grande. Mas se isto acontecer, também o quociente $\frac{(n-1)S^2}{\sigma_0^2}$ deverá ser "demasiado" pequeno ou "demasiado" grande. Numa formulação matemática, deveremos rejeitar a hipótese de $\sigma^2 = \sigma_0^2$ se,

$$\frac{(n-1)S^2}{\sigma_0^2} < a \quad \text{ou} \quad \frac{(n-1)S^2}{\sigma_0^2} > b.$$

Mas qual o valor de a e de b ? Ora, quando $\sigma^2 = \sigma_0^2$, a estatística de teste e respectiva distribuição de amostragem são:

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2 = \sigma_0^2}{\sim} \chi_{n-1}^2$$

(tem distribuição do qui-quadrado com $(n-1)$ graus de liberdade). Então, para um nível de significância α ,

$$\alpha = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = P\left(\frac{(n-1)S^2}{\sigma_0^2} < a\right) + P\left(\frac{(n-1)S^2}{\sigma_0^2} > b\right)$$

Repartindo a probabilidade α em partes iguais pela cauda esquerda e direita da distribuição χ_{n-1}^2 , tem-se

$$a = \chi_{n-1:1-\alpha/2}^2 \quad \text{e} \quad b = \chi_{n-1:\alpha/2}^2.$$

Então, para um nível de significância α , a região de rejeição fica definida por:

$$R_\alpha \equiv \left[0, \chi_{n-1:1-\alpha/2}^2 \right] \cup \left[\chi_{n-1:\alpha/2}^2, +\infty \right[$$

e rejeitaremos H_0 se

$$x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha.$$

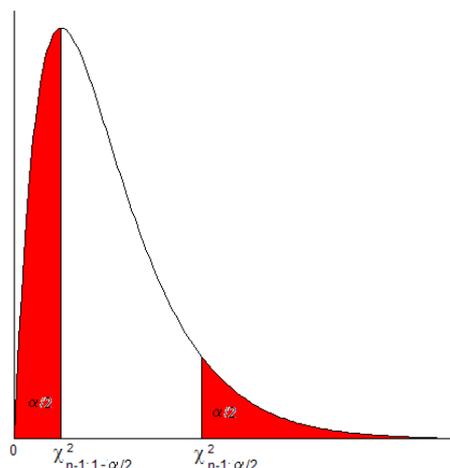
Regra de decisão para um nível de significância α

$$\text{Região de rejeição: } R_\alpha = \left[0, \chi_{n-1:1-\alpha/2}^2 \right] \cup \left[\chi_{n-1:\alpha/2}^2, +\infty \right[$$

$$\text{Rejeitar } H_0 \text{ se } x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha$$

$$\text{p-value} = 2 \min \left(P(X^2 < x_{obs}^2), P(X^2 > x_{obs}^2) \right)$$

Figura 4.7: Teste bilateral para a variância



4.8.2 Teste de hipóteses unilateral direito para a variância

Consideremos as hipóteses

$$H_0 : \sigma^2 \leq \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 > \sigma_0^2$$

Naturalmente que devemos rejeitar a hipótese $\sigma^2 \leq \sigma_0^2$ se a amostra nos fornecer uma estimativa S^2 para a qual o quociente $\frac{S^2}{\sigma_0^2}$ é muito grande. Mas se isto acontecer, também o quociente $\frac{(n-1)S^2}{\sigma_0^2}$ deverá ser "demasiado" grande. Resumindo, deveremos rejeitar a hipótese de $\sigma^2 \leq \sigma_0^2$ se,

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} > a.$$

Mas qual o valor de a ? Ora, quando $\sigma^2 \leq \sigma_0^2$, a estatística de teste $X^2 = \frac{(n-1)S^2}{\sigma_0^2}$ tem distribuição

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2 = \sigma_0^2}{\sim} \chi_{n-1}^2$$

(X^2 tem distribuição do qui-quadrado com $(n-1)$ graus de liberdade)

Para um nível de significância α ,

$$\alpha = \max_{\sigma \leq \sigma_0} P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = P\left(\frac{(n-1)S^2}{\sigma_0^2} > a\right) = P(X^2 > a)$$

o que implica,

$$a = \chi_{n-1; \alpha}^2.$$

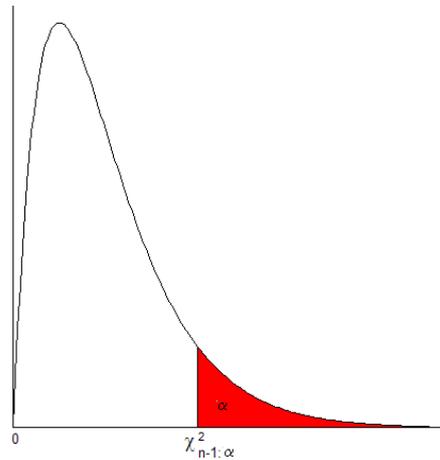
Regra de decisão para um nível de significância α

$$\text{Região de rejeição: } R_\alpha =]\chi_{n-1; \alpha}^2, +\infty[$$

$$\text{Rejeitar } H_0 \text{ se } x_{obs}^2 \in R_\alpha$$

$$\text{p-value} = P(X^2 > x_{obs}^2)$$

Figura 4.8: Teste unilateral direito para a variância



4.8.3 Teste de hipóteses unilateral esquerdo para a variância

Consideremos as hipóteses

$$H_0 : \sigma^2 \geq \sigma_0^2 \quad \text{vs} \quad H_1 : \sigma^2 < \sigma_0^2$$

Neste caso devemos rejeitar a hipótese $\sigma^2 \leq \sigma_0^2$ se a amostra nos fornecer uma estimativa S^2 para a qual o quociente $\frac{S^2}{\sigma_0^2}$ é muito pequeno. Mas se isto acontecer, também o quociente $\frac{(n-1)S^2}{\sigma_0^2}$ deverá ser "demasiado" pequeno. Deveremos então rejeitar a hipótese de $\sigma^2 \geq \sigma_0^2$ se,

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} < a.$$

Mas qual o valor de a ? Ora, quando $\sigma^2 \geq \sigma_0^2$, a estatística de teste passará a ser

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2 = \sigma_0^2}{\sim} \chi_{n-1}^2$$

(X^2 tem distribuição χ_{n-1}^2). Então, para um nível de significância α ,

$$\alpha = \max_{\sigma^2 \geq \sigma_0^2} P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = P\left(\frac{(n-1)S^2}{\sigma_0^2} < a\right) = P(X^2 < a)$$

Isto implica que

$$a = \chi_{n-1; 1-\alpha}^2.$$

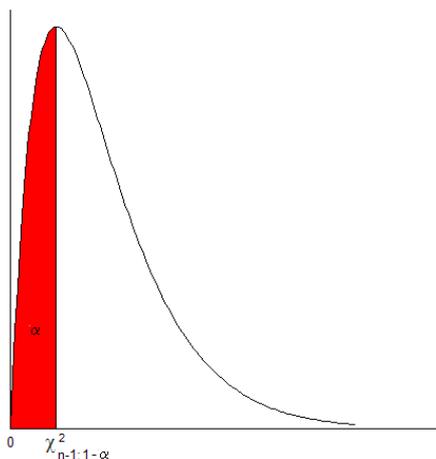
Regra de decisão para um nível de significância α

$$\text{Região de rejeição: } R_\alpha = [0, \chi_{n-1; 1-\alpha}^2[$$

$$\text{Rejeitar } H_0 \text{ se } x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha$$

$$\text{p-value} = P(X^2 < x_{obs}^2)$$

Figura 4.9: Teste unilateral esquerdo para a variância



Exemplo 4.9. A administração de uma SAD reclama que o investimento nas suas ações é seguro e que o desvio padrão do preço das ações é inferior a 2 euros. Suponha que está interessado numa eventual compra de ações desta SAD mas, antes de fazer a compra decide testar a veracidade das afirmações da administração. Para tal escolheu aleatoriamente 30 dias dos últimos 3 anos e registou o preço das ações. A amostra facultou um desvio padrão amostral de $s = 1.70$ euros.

Será que esta estimativa indica, ao nível de 5% de significância, que a administração da SAD está a dar informação verdadeiras?

Queremos testar

$$H_0 : \sigma \geq 2 \quad \text{vs} \quad H_1 : \sigma < 2$$

que é equivalente a testar

$$H_0 : \sigma^2 \geq 4 \quad \text{vs} \quad H_1 : \sigma^2 < 4.$$

A informação amostral disponível é:

$$n = 30 \quad \alpha = 0.05 \quad s^2 = 1.70^2 = 2.89$$

De acordo com a metodologia proposta para a realização do teste, tem-se

Hipóteses: $H_0 : \sigma^2 \geq 4 \quad \text{vs} \quad H_1 : \sigma^2 < 4$

Estimador de σ^2 : $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Estatística de teste: $X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2=\sigma_0^2}{\sim} \chi_{n-1}^2 \Leftrightarrow X^2 = \frac{29S^2}{4} \underset{\sigma^2=4}{\sim} \chi_{29}^2$

Região de rejeição para um nível de significância $\alpha = 5\%$:

$R_{0.05} = [0, c]$, sendo c o valor que satisfaz:

$$\begin{aligned} \alpha = 5\% &= P(X^2 \in R_{0.05}) \Leftrightarrow 0.05 = P(X^2 < c) \Leftrightarrow P(X^2 \geq c) = 0.95 \\ &\Leftrightarrow c = \chi_{29; 0.95}^2 = 17.708 \end{aligned}$$

$$R_{0,05} = [0, 17.708[$$

Decisão: Como $x_{obs}^2 = \frac{29 \times 2.89}{4} = 20.953$ e $x_{obs}^2 \notin R_{0,05}$, não devemos duvidar das afirmações da administração da SAD, com uma significância de 5%.

Para uma decisão fundamentada no p -value, tem-se

$$p\text{-value} = P(X^2 < x_{obs}^2) = P(X^2 < 20.953) = 0.1391 > \alpha \ (\alpha = 5\%)$$

peço que, não devemos duvidar das afirmações da administração da SAD, com uma significância de 5%.

Tabela 4.5: Testes de hipóteses para a variância

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 \neq \sigma_0^2$$

Condições de aplicação	
$X \sim N(\mu, \sigma^2)$, μ desconhecido	Estatística de teste: $X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2=\sigma_0^2}{\sim} \chi_{n-1}^2$ Região de rejeição $R_\alpha = [0, \chi_{n-1:1-\alpha/2}^2 \cup \chi_{n-1:\alpha/2}^2, +\infty[$ Rejeitar H_0 se $x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha$ $p\text{-value} = 2 \times \min(P(X^2 < x_{obs}^2), P(X^2 > x_{obs}^2))$

$$H_0 : \sigma^2 \geq \sigma_0^2 \text{ vs } H_1 : \sigma^2 < \sigma_0^2$$

Condições de aplicação	
$X \sim N(\mu, \sigma^2)$, μ desconhecido	Estatística de teste: $X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2=\sigma_0^2}{\sim} \chi_{n-1}^2$ Região de rejeição $R_\alpha = [0, \chi_{n-1:1-\alpha}^2[$ Rejeitar H_0 se $x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha$ $p\text{-value} = P(X^2 < x_{obs}^2)$

$$H_0 : \sigma^2 \leq \sigma_0^2 \text{ vs } H_1 : \sigma^2 > \sigma_0^2$$

Condições de aplicação	
$X \sim N(\mu, \sigma^2)$, μ desconhecido	Estatística de teste: $X^2 = \frac{(n-1)S^2}{\sigma_0^2} \underset{\sigma^2=\sigma_0^2}{\sim} \chi_{n-1}^2$ Região de rejeição $R_\alpha = [\chi_{n-1:\alpha}^2, +\infty[$ Rejeitar H_0 se $x_{obs}^2 = \frac{(n-1)s^2}{\sigma_0^2} \in R_\alpha$ $p\text{-value} = P(X^2 > x_{obs}^2)$

4.9 Outros testes de hipóteses

Para outros testes de hipóteses usuais, limitamo-nos a apresentar os quadros resumos das estatísticas de teste a utilizar e respectivas regras de decisão.

4.9.1 Teste de hipóteses para a proporção

Tabela 4.6: Testes para a proporção, p

$H_0 : p = p_0$ vs $H_1 : p \neq p_0$	
Condições de aplicação	
$X \sim B(1, p), n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)}} \underset{p=p_0}{\overset{a}{\rightsquigarrow}} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_{\alpha/2} [\cup]z_{\alpha/2}, +\infty [$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \in R_\alpha$ p -value $\approx 2 \times P(Z > w_{obs})$
$H_0 : p \geq p_0$ vs $H_1 : p < p_0$	
Condições de aplicação	
$X \sim B(1, p), n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)}} \underset{p=p_0}{\overset{a}{\rightsquigarrow}} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_\alpha [$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \in R_\alpha$ p -value $\approx P(Z < w_{obs})$
$H_0 : p \leq p_0$ vs $H_1 : p > p_0$	
Condições de aplicação	
$X \sim B(1, p), n \geq 30$	Estatística de teste: $W = \sqrt{n} \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)}} \underset{p=p_0}{\overset{a}{\rightsquigarrow}} N(0, 1)$ Região de rejeição: $R_\alpha \approx]z_\alpha, +\infty [$ Rejeitar H_0 se $w_{obs} = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \in R_\alpha$ p -value $\approx P(Z > w_{obs})$

Exemplo 4.10. Um comerciante admite que a possibilidade de um cliente adquirir pelo menos um produto na sua loja é constante e de valor superior a 0.4. Durante um mês, contou o número de clientes que entraram na loja assim como os que fizeram alguma compra, tendo registado os valores 878 e 495, respectivamente. A informação recolhida permite corroborar as suas suspeitas?

As hipóteses a teste deverão ser

$$H_0 : p \leq 0.4 \text{ vs } H_1 : p > 0.4$$

que vamos testar com um nível de significância $\alpha = 10\%$.

A informação disponível é:

$$p_0 = 0.4 \quad \hat{p} = 495/878 = 0.56 \quad n = 878 \quad z_\alpha = z_{0.10} = 1.28$$

A regra de rejeição é: Rejeitar H_0 se $Z = \sqrt{n} \frac{\hat{P} - 0.4}{\sqrt{0.4(1-0.4)}} > z_{0.1}$

Ora

$$z_{obs} = \sqrt{878} \frac{0.56 - 0.4}{\sqrt{0.4(1-0.4)}} = 9.68 > 1.28$$

Decisão: O comerciante não deve duvidar das suas suspeitas, com uma significância de 10%.

$$p\text{-value} = P(Z > z_{obs}) = P(Z > 9.68) \approx 0 < 0.10 (\alpha)$$

4.9.2 Teste de hipóteses para a diferença dos valores médios de duas populações

Tabela 4.7: Testes de hipóteses para a diferença de dois valores médios

$$H_0 : \mu_X - \mu_Y = d_0 \text{ vs } H_X : \mu_X - \mu_Y \neq d_0$$

Situação	Condições de aplicação	
A	$X \sim N(\mu_X, \sigma_X^2)$ $Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 conhecidas	Estatística de teste: $Z = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \sim N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $z_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} = 2 \times P(Z > z_{obs})$
B	$X \sim N(\mu_X, \sigma_X^2)$ $Y \sim N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$ e desconhecidas	Estatística de teste: $T = \frac{\bar{X} - \bar{Y} - d_0}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \mu_X - \mu_Y = d_0 \sim t_{n_X + n_Y - 2}$ RR: $R_\alpha =]-\infty, -t_{n_X + n_Y - 2; \alpha/2}[\cup]t_{n_X + n_Y - 2; \alpha/2}, +\infty[$ Rejeitar H_0 se $t_{obs} = \frac{\bar{x} - \bar{y} - d_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \in R_\alpha$ $p\text{-value} = 2 \times P(T > t_{obs})$
C	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 conhecidas n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \stackrel{a}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} \approx 2 \times P(Z > w_{obs})$
D	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 desconhecidas n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \stackrel{a}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} \approx 2 \times P(Z > w_{obs})$

$$H_0 : \mu_X - \mu_Y \geq d_0 \text{ vs } H_1 : \mu_X - \mu_Y < d_0$$

Situação	Condições de aplicação	
A	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 conhecidas	Estatística de teste: $Z = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \sim N(0, 1)$ Região de rejeição: $R_\alpha =]-\infty, -z_\alpha[$ Rejeitar H_0 se $z_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p - value = P(Z < z_{obs})$
B	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$ e desconhecidas	Estatística de teste: $T = \frac{\bar{X} - \bar{Y} - d_0}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \mu_X - \mu_Y = d_0 \sim t_{n_X + n_Y - 2}$ Região de rejeição: $R_\alpha =]-\infty, -t_{n_X + n_Y - 2; \alpha}[$ Rejeitar H_0 se $t_{obs} = \frac{\bar{x} - \bar{y} - d_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \in R_\alpha$ $p - value = P(T < t_{obs})$
C	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 conhecidas, n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \stackrel{a}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_\alpha[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p - value \approx P(Z < w_{obs})$
D	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 desconhecidas, n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \mu_X - \mu_Y = d_0 \stackrel{a}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha \approx]-\infty, -z_\alpha[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \in R_\alpha$ $p - value \approx P(Z < w_{obs})$

$$H_0 : \mu_X - \mu_Y \leq d_0 \text{ vs } H_1 : \mu_X - \mu_Y > d_0$$

Situação	Condições de aplicação	Regra de rejeição
A	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 conhecidas	Estatística de teste: $Z = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \underset{\mu_X - \mu_Y = d_0}{\sim} N(0, 1)$ Região de rejeição: $R_\alpha =]z_\alpha, +\infty[$ Rejeitar H_0 se $z_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} = P(Z > z_{obs})$
B	$X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2)$ $\sigma_X^2 = \sigma_Y^2$ desconhecida	Estatística de teste: $T = \frac{\bar{X} - \bar{Y} - d_0}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \underset{\mu_X = \mu_Y}{\sim} t_{n_X + n_Y - 2}$ Região de rejeição: $R_\alpha =]t_{n_X + n_Y - 2; \alpha}, +\infty[$ Rejeitar H_0 se $t_{obs} = \frac{\bar{x} - \bar{y} - d_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \in R_\alpha$ $p\text{-value} = P(T > t_{obs})$
C	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 conhecidas, n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \underset{\mu_X - \mu_Y = d_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha \approx]z_\alpha, +\infty[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} \approx P(Z > w_{obs})$
D	$X \sim ?, Y \sim ?$ σ_X^2, σ_Y^2 desconhecidas, n_X e $n_Y \geq 30$	Estatística de teste: $W = \frac{\bar{X} - \bar{Y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \underset{\mu_X - \mu_Y = d_0}{\overset{a}{\sim}} N(0, 1)$ Região de rejeição: $R_\alpha \approx]z_\alpha, +\infty[$ Rejeitar H_0 se $w_{obs} = \frac{\bar{x} - \bar{y} - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \in R_\alpha$ $p\text{-value} \approx P(Z > w_{obs})$

Exemplo 4.11. A FNN decidiu comprar fatos novos para os atletas. Adquiriu 6 fatos da marca mais cara (Tipo A) e 7 da marca mais barata (TIPO B) e enviou-os para um laboratório, onde se registaram os tempos de duração até romperem. Os registos, em horas, aparecem na tabela que se segue:

Tipo A: 1400 1725 1610 1605 1950 1575

Tipo B: 1615 1665 1730 1755 1632 1606 1790

Admitindo que o tempo de duração dos fatos para cada marca têm uma lei normal com a mesma variância, poderá dizer, com uma significância de 5%, que as durações médias dos fatos das duas marcas são idênticas?

Estime por intervalo de 95% de confiança a diferença entre as durações médias dos fatos de cada marca.

As hipóteses a testar deverão ser

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

que vamos testar com um nível de significância $\alpha = 5\%$.

Supondo que o tempo de duração dos fatos têm distribuição normal e que são iguais, temos:

$$n_1 = 6 \quad \bar{x} = 1644.17 \quad s_1^2 = 182.85^2 \quad n_2 = 7 \quad \bar{y} = 1684.71 \quad s_2^2 = 73.37^2 \quad s_p^2 = 199469.5539$$

$$t_{n_1+n_2-2;\alpha/2} = t_{11;0.025} = 2.201$$

A regra de rejeição é: Rejeitar H_0 se $\left| \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{n_1+n_2-2;\alpha/2}$

Ora

$$\left| \frac{1644.17 - 1684.71}{446.62 \sqrt{\frac{1}{6} + \frac{1}{7}}} \right| = 0.163 < 2.201 = t_{11;0.025}$$

Decisão: Com uma significância de 5%, não existe evidência para dizer que os tempos médios de duração são distintos.

O intervalo de confiança $(1 - \alpha) = 1 - 0.05 = 0.95$ para $\mu_1 - \mu_2$ é a região de não rejeição do teste das hipóteses

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

O intervalo de $(1 - \alpha)$ de confiança para $\mu_1 - \mu_2$ é

$$\left[\bar{X} - \bar{Y} - t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

e a estimativa para a diferença médias das durações dos fatos é

$$\left[1644.17 - 1684.71 - 2.201 \times 446.62 \sqrt{\frac{1}{6} + \frac{1}{7}}, 1644.17 - 1684.71 + 2.201 \times 446.62 \sqrt{\frac{1}{6} + \frac{1}{7}} \right] =$$

$$= [-587.437, 506.357]$$

Capítulo 5

Teste de ajustamento do Qui-Quadrado

Por vezes as conjecturas (hipóteses) sobre uma certa população em estudo, não incidem sobre os parâmetros da distribuição dessa população, mas sim sobre a própria distribuição da população. Os testes destas hipóteses dizem-se **testes não paramétricos**.

Neste capítulo vamos apresentar um teste não paramétrico específico para averiguar se uma dada população (ou v.a.) tem uma distribuição F preconizada, isto é um teste para as hipóteses:

$$H_0 : X \sim F(\cdot) \text{ vs } H_1 : X \not\sim F(\cdot)$$

O teste que aqui apresentamos intitula-se **teste de ajustamento do qui-quadrado** e apresenta a vantagem de se poder aplicar a qualquer distribuição preconizada para uma população, mas a desvantagem de exigir que as amostras sejam grandes.

Consideremos uma amostra (X_1, X_2, \dots, X_n) de uma população X , com dimensão $n \geq 30$.

- Agrupada em m classes A_1, \dots, A_m disjuntas e verificando $\bigcup_{i=1}^m A_i = S_X$, com S_X o suporte de $X \sim F(\cdot)$;
- Sejam $p_i = P(X \in A_i)$, $i = 1, \dots, m$

$$0 < p_i < 1, \quad i = 1, \dots, m \quad \text{e} \quad \sum_{i=1}^m p_i = 1$$

- Seja O_i o n.º observado de valores amostrais que pertencem à classe A_i , $i = 1, \dots, m$.

$$0 < O_i < n, \quad i = 1, \dots, m \quad \text{e} \quad \sum_{i=1}^m O_i = n$$

Os valores O_i , $i = 1, \dots, m$ são designados por **frequências observadas**.

- Admitindo **verdadeira a hipótese H_0** , sejam

$$p_{0i} = P(X \in A_i | H_0 \text{ verdadeira}), \quad i = 1, \dots, m$$

e

$$E_i = n \times P(X \in A_i | H_0 \text{ verdadeira}) = n \times p_{0i}, \quad i = 1, \dots, m$$

$$0 < E_i < n, \quad i = 1, \dots, m \quad \text{e} \quad \sum_{i=1}^m E_i = n$$

E_i é designado por **frequência esperada da classe A_i** , caso H_0 seja verdadeira.

- A estatística do teste do qui-quadrado é

$$X^2 = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i}$$

Nota: Essencialmente, a estatística X^2 é uma medida ponderada das discrepâncias entre as frequências observadas O_i e a frequências que se esperam observar E_i , quando H_0 é verdadeira.

- Quando H_0 é verdadeira,

$$X^2 = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i} \underset{\text{sob } H_0}{\overset{a}{\sim}} \chi_{m-p-1}^2$$

sendo m o número de classes e p o número de parâmetros estimados da distribuição considerada na hipótese nula.

- Sendo necessário estimar parâmetros da distribuição preconizada na hipótese nula, devem ser utilizados os respectivos estimadores de máxima verosimilhança.
- A hipótese H_0 deverá ser rejeitada para grandes valores da estatística de teste. Como tal, a região de rejeição é um intervalo $]a, +\infty[$, $a > 0$.

Para um nível de significância α ,

$$\alpha = P(X^2 > a | H_0 \text{ verdadeira}),$$

pelo que $a = \chi_{m-p-1:\alpha}^2$ e a região de rejeição é:

$$R_\alpha =]\chi_{m-p-1:\alpha}^2, +\infty[$$

- Dado o valor observado da estatística de teste, x_{obs}^2 , e para um nível α de significância, decidimos rejeitar H_0 se:

$$x_{obs}^2 \in R_\alpha, \quad \text{ou seja, se } x_{obs}^2 > \chi_{m-p-1:\alpha}^2.$$

- **Nota:** Quando uma classe tem frequência esperada $E_i < 5$, devemos fundi-la com a classe (ou as classes adjacentes) até que a nova frequência esperada tenha um valor que não inferior a 5. O valor da frequência observada O_i deve ser reajustado, assim como o número m final de classes.

Vejam alguns exemplos.

Exemplo 5.1. Uma máquina de café de utilização doméstica é vendida em 4 cores: preta (A_1), branca (A_2), vermelha (A_3) e castanha (A_4). Pretende-se saber se os consumidores manifestam a mesma preferência por qualquer cor, isto é, se a v.a. X -cor escolhida por um cliente, tem função de probabilidade:

$$X \begin{cases} A_1 & A_2 & A_3 & A_4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{cases}$$

As hipóteses a testar são:

$$H_0 : p_i = P(A_i) = 1/4, \quad i = 1, \dots, 4 \quad \text{vs} \quad H_1 : p_i = P(A_i) \neq 1/4, \text{ para algum } i$$

Para tal, numa sondagem a 300 compradores deste equipamento, foi reunida a seguinte informação acerca da cor da máquina que adquiriram:

X	A_1	A_2	A_3	A_4
O_i	90	88	62	60

Para $i = 1, \dots, 4$, $E_i = n \times p_{0i} = 300 \times 1/4 = 75$ e

$$x_{obs}^2 = \frac{(90 - 75)^2}{75} + \frac{(88 - 75)^2}{75} + \frac{(62 - 75)^2}{75} + \frac{(60 - 75)^2}{75} = 10.5067$$

Para um nível de 5% de significância, $\chi_{4-1;0.05}^2 = 7.815$, pelo que $R_{0.05} =]7.815, +\infty[$.

Como $x_{obs}^2 = 10.5067 \in R_{0.05}$, rejeitamos a hipótese H_0 , ou seja, com 5% de significância concluímos que existe evidência estatística para afirmar que os consumidores manifestam diferentes preferências pela cor da máquina.

Valor-p associado à nossa decisão:

$$p\text{-value} = P(X^2 > x_{obs}^2 | H_0 \text{ verdadeira}) = P(\chi_3^2 > 10.5067) = 0.0147,$$

que sendo inferior a $\alpha = 0.05$, confirma a decisão de rejeitar H_0 .

Exemplo 5.2 (Teste de ajustamento para uma distribuição de Poisson).

O número de avarias registadas diariamente numa máquina, suspeita-se ter distribuição de Poisson. Recolhida informação diária sobre o número de avarias corridas durante 260 dias de laboração, obtiveram-se os seguintes valores:

N.º de avarias	0	1	2	3
N.º de dias, O_i	102	96	48	14

Sendo X -n.º de avarias diárias, queremos testar as hipóteses

$$H_0 : X \sim P(\lambda) \quad \text{vs} \quad H_1 : X \not\sim P(\lambda)$$

Como λ tem valor desconhecido, teremos de o estimar usando o estimador de máxima verosimilhança $\hat{\lambda} = \bar{X}$. Para a presente amostra, obtemos uma estimativa $\hat{\lambda} = 234/260 = 0.9$.

Consideremos as classes $A_1 = \{0\}$, $A_2 = \{1\}$, $A_3 = \{2\}$, $A_4 = \{3, 4, \dots\}$. Na tabela seguinte, apresentamos os valores de: $\hat{p}_{0i} = P(X \in A_i | X \sim P(\hat{\lambda}))$, $i = 1, \dots, 4$ e as frequências esperadas $\hat{E}_i = 300 \times p_{0i}$, $i = 1, \dots, 4$.

Classe	0	1	2	3 ou mais
\hat{p}_{0i}	0.4066	0.3659	0.1647	0.0628
\hat{E}_i	105.716	95.134	42.822	16.328

O valor observado da estatística do qui-quadrado é:

$$x_{obs}^2 = \frac{(102 - 105.716)^2}{105.716} + \frac{(96 - 95.134)^2}{95.134} + \frac{(48 - 42.822)^2}{42.822} + \frac{(14 - 16.328)^2}{16.328} = 1.096542483$$

A estatística de teste é

$$X^2 = \sum_{i=1}^m \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \underset{\text{sob } H_0}{\overset{a}{\sim}} \chi_{4-1-1}^2$$

Para um nível de 10% de significância, a região de rejeição é $R_{0.1} =]\chi_{4-1-1:0.1}^2, +\infty[=]4.605, +\infty[$.

Uma vez que $x_{obs}^2 = 1.096542483 \notin R_{0.1}$, não rejeitamos a hipótese de o n.º de avarias diárias ter distribuição de Poisson.

Valor-p associado à nossa decisão:

$$p\text{-value} = P(X^2 > x_{obs}^2 | H_0 \text{ verdadeira}) = P(\chi_2^2 > 1.096542483) = 0.5779,$$

que sendo superior a $\alpha = 0.10$, confirma a decisão de não rejeitar H_0 .

5.1 Teste ao pressuposto da normalidade de uma população

O teste á normalidade da distribuição de uma população tem particular interesse porque nos permitirá saber qual a distribuição de amostragem a usar para a média amostral \bar{X} (discussão sobre qual das situações, nomeadamente as que temos vindo a identificar como A, B, C ou D, aplicar) ou para a variância amostral S^2 .

Passamos a exemplificar como se utiliza o teste de ajustamento do qui-quadrado para testar as hipóteses

$$H_0: X \sim N(\mu, \sigma^2) \quad \text{vs} \quad H_1: X \not\sim N(\mu, \sigma^2)$$

Exemplo 5.3. Consideremos a amostra de medições da acidez (pH) da água da chuva apresentada no exemplo 4.6 acrescida de mais 18 observações. Nesse exemplo, admitimos que a população X - "Acidez (pH) da água da chuva" tinha distribuição normal.

Vamos agora testar se este pressuposto se verifica ou não, ou seja vamos testar as hipóteses:

$$H_0: X \sim N(\mu, \sigma^2) \quad \text{vs} \quad H_1: X \not\sim N(\mu, \sigma^2)$$

A nossa amostra era

5.1 5.0 4.6 3.8 4.8 3.6 4.7 3.8 4.1 4.2 4.4 4.5 3.6 3.9 4.2
4.3 4.4 4.5 4.9 4.7 4.8 4.1 4.2 4.5 4.5 4.6 4.4 4.1 4.6 4.5

correspondendo-lhe uma média $\bar{x} = 4.55$ e um desvio padrão amostral $s = 0.462208139$.

Começemos por agrupar os dados da amostra. Para tal consideremos os seguintes intervalos (denominados classes) para agrupamento dos dados: $]-\infty, 3.7]$, $]3.7, 4.0]$, $]4.0, 4.3]$, $]4.3, 4.6]$, $]4.6, 4.9]$ e $]4.9, +\infty[$.

As frequências absolutas observados de observações em cada classe são:

Tabela de frequências		
i	Classe A_i	Frequência observada O_i
1	$]-\infty, 3.7]$	2
2	$]3.7, 4.0]$	3
3	$]4.0, 4.3]$	7
4	$]4.3, 4.6]$	11
5	$]4.6, 4.9]$	5
6	$]4.9, +\infty[$	2
	Totais	30

Ora, se a hipótese $H_0: X$ tem distribuição normal, for verdadeira, isto é, se $X \sim N(\mu, \sigma^2)$,

$$\begin{aligned} p_2 = P(A_2) &= P(3.7 < X \leq 4.0) = P(X \leq 4.0) - P(X \leq 3.7) = \\ &= P\left(Z \leq \frac{4.0 - \mu}{\sigma}\right) - P\left(Z \leq \frac{3.7 - \mu}{\sigma}\right) = \\ &= \Phi\left(\frac{4.0 - \mu}{\sigma}\right) - \Phi\left(\frac{3.7 - \mu}{\sigma}\right) \end{aligned}$$

que não podemos calcular porque desconhecemos o valor de μ e de σ .

Contudo, sabemos que os estimadores \bar{X} e S^2 (estimadores de máxima verosimilhança de μ e σ^2 , respectivamente) fornecem boas estimativas para μ e para σ^2 . Portanto, podemos estimar a anterior probabilidade, substituindo μ por $\bar{x} = 4.38$ e σ^2 por $s^2 = 0.152$.

Assim

$$\begin{aligned} \hat{p}_2 = P(A_2) &\approx \Phi\left(\frac{4.0 - \bar{x}}{s}\right) - \Phi\left(\frac{3.7 - \bar{x}}{s}\right) = \\ &= \Phi(-0.97) - \Phi(-1.74) = 0.1251 \end{aligned}$$

Nota: Repare que, para o cálculo da probabilidade, fomos obrigados a usar as estimativas de 2 parâmetros.

Se repetirmos este raciocínio para as restantes classes, obtemos as estimativas da probabilidade de cada classe (se H_0 é verdadeira):

$$\begin{aligned} \hat{p}_1 = P(A_1) &= P(X \leq 3.7) \approx \Phi\left(\frac{3.7 - \bar{x}}{s}\right) = \Phi(-1.74) = 0.0409 \\ \hat{p}_3 = P(A_3) &= P(4.0 < X \leq 4.3) \approx \Phi\left(\frac{4.3 - \bar{x}}{s}\right) - \Phi\left(\frac{4.0 - \bar{x}}{s}\right) = \\ &= \Phi(-0.21) - \Phi(-0.97) = 0.2508 \\ \hat{p}_4 = P(A_4) &= P(4.3 < X \leq 4.6) \approx \Phi\left(\frac{4.6 - \bar{x}}{s}\right) - \Phi\left(\frac{4.3 - \bar{x}}{s}\right) = \\ &= \Phi(0.56) - \Phi(-0.21) = 0.2955 \\ \hat{p}_5 = P(A_5) &= P(4.6 < X \leq 4.9) \approx \Phi\left(\frac{4.9 - \bar{x}}{s}\right) - \Phi\left(\frac{4.6 - \bar{x}}{s}\right) = \\ &= \Phi(1.33) - \Phi(0.56) = 0.1959 \\ \hat{p}_6 = P(A_6) &= P(X > 4.9) \approx 1 - \Phi\left(\frac{4.9 - \bar{x}}{s}\right) = 1 - \Phi(1.33) = 0.0918 \end{aligned}$$

Vamos agora concluir o nosso exemplo. Começamos por construir uma tabela onde apresentamos as frequências observadas, as frequências esperadas e as parcelas $\frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$ de cada classe, assim como o valor observado de X^2 .

i	Classe A_i	Frequência absoluta O_i	Frequência esperada \hat{E}_i	$(O_i - \hat{E}_i)^2 / \hat{E}_i$
1	$]-\infty, 3.7]$	2	1.227	0.4870
2	$]3.7, 4.0]$	3	3.753	0.1511
3	$]4.0, 4.3]$	7	7.524	0.0365
4	$]4.3, 4.6]$	11	8.865	0.5142
5	$]4.6, 4.9]$	5	5.877	0.1309
6	$]4.9, +\infty[$	2	2.754	0.2064
	Totais	30	30	$1.5261 = x_{obs}^2$

Repare que as classes A_1 , A_2 e A_6 têm uma frequência esperada inferior a 5. Devemos então fundir as classes A_1 e A_2 assim como as classes A_5 e A_6 , dando origem à nova informação:

i	Classe A_i	Frequência absoluta O_i	Frequência esperada \hat{E}_i	$(O_i - \hat{E}_i)^2 / \hat{E}_i$
1	$]-\infty, 4.0]$	5	4.980	0.0001
2	$]4.0, 4.3]$	7	7.524	0.0365
3	$]4.3, 4.6]$	11	8.865	0.5142
4	$]4.6, +\infty[$	7	8.631	0.3082
	Totais	30	30	$0.8590 = x_{obs}^2$

A estatística de teste é

$$X^2 = \sum_{i=1}^m \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \underset{\text{sob } H_0}{\sim} \chi_{4-2-1}^2 \equiv \chi_1^2$$

Se considerarmos um nível de significância $\alpha = 0.05$, temos $\chi_{1:0.05}^2 = 3.841$ e como

$$x_{obs}^2 = 0.8590 < 3.841 = \chi_{1:0.05}^2$$

não existem razões para duvidar de que a população X -"acidez (pH) da água da chuva", tem distribuição normal.

Valor-p associado à nossa decisão:

$$p\text{-value} = P(X^2 > x_{obs}^2 | H_0 \text{ verdadeira}) = P(\chi_1^2 > 0.8590) = 0.354,$$

que sendo superior a $\alpha = 0.05$, confirma a decisão de não rejeitar H_0 .

Vejam os outros exemplos, em que o número de parâmetros a estimar para o cálculo das frequências esperadas, é diferente.

Exemplo 5.4. Para o exemplo 4.4, precisaríamos de verificar previamente se a população X —"gasto semanal em alimentação (para famílias com dois filhos) em Agosto de 2003" tem distribuição normal com desvio padrão conhecido e de valor $\sigma = 15$ euros.

As nossas hipóteses são

$$H_0 : X \sim N(\mu, 15^2) \quad \text{vs} \quad H_1 : X \sim N(\mu, 15^2)$$

Consideremos o agrupamento da amostra de gastos em alimentação das $n = 25$ famílias, nas classes $A_1 =]-\infty, 90]$, $A_2 =]90, 95]$, $A_3 =]95, 100]$, $A_4 =]100, 105]$, $A_5 =]105, 110]$, $A_6 =]110, 115]$ e $A_7 =]115, +\infty[$.

Começemos por exemplificar o cálculo da frequência esperada da classe A_3 .

$$\begin{aligned} E_3 &= n \times p_3 = 25 \times P(95 < X \leq 100) = \\ &= 25 \times (P(X \leq 100) - P(X \leq 95)) = \\ &= 25 \times \left(P\left(Z \leq \frac{100 - \mu}{15}\right) - P\left(Z \leq \frac{95 - \mu}{15}\right) \right) = \\ &= 25 \times \left(\Phi\left(\frac{100 - \mu}{15}\right) - \Phi\left(\frac{95 - \mu}{15}\right) \right) \end{aligned}$$

Mas, como desconhecemos o valor de μ , teremos de o substituir pela respectiva estimativa de máxima verosimilhança, $\bar{x} = 108$.

Nota: Repare que para o cálculo das frequências esperadas, somos obrigados a usar a estimativa de 1 parâmetro.

Então

$$\begin{aligned} \hat{E}_3 &= n \times \hat{p}_3 \approx 25 \times \left(\Phi\left(\frac{100 - \bar{x}}{15}\right) - \Phi\left(\frac{95 - \bar{x}}{15}\right) \right) = \\ &= 25 \times (\Phi(-0.53) - \Phi(-0.87)) = 25 \times 0.1059 = 2.6475 \end{aligned}$$

Completando o cálculo das restantes frequências esperadas, obtemos o quadro resumo da informação amostral:

i	Classe A_i	Frequência observada O_i	Frequência esperada \hat{E}_i	$(O_i - \hat{E}_i)^2 / \hat{E}_i$
1	$]-\infty, 90]$	2	2.8775	0.2676
2	$]90, 95]$	2	1.9275	0.0027
3	$]95, 100]$	2	2.6475	0.1584
4	$]100, 105]$	5	3.0650	1.2216
5	$]105, 110]$	7	3.2750	4.2368
6	$]110, 115]$	4	3.2275	0.1849
7	$]115, +\infty[$	3	7.9800	3.1078
	Totais	25	25	9.1798

Desde logo verificamos que é necessário aglutinar classes de modo a que todas tenham valor esperada não inferior a 5. Este processo pode conduzir ao seguinte agrupamento final:

i	Classe A_i	Frequência observada O_i	Frequência esperada \hat{E}_i	$(O_i - \hat{E}_i)^2 / \hat{E}_i$
1	$]-\infty, 95]$	4	4,8050	0.1349
2	$]95, 105]$	7	5.7125	0.2902
3	$]105, 115]$	11	6.5025	3.1107
4	$]115, +\infty[$	3	7.9800	3.0992
	Totais	25	25	6.6350

Estipulando um nível de significância $\alpha = 0.10$, temos $\chi_{4-1-1;0.10}^2 = 4.605$ e como

$$x_{obs}^2 = 6.6350 > 4.605 = \chi_{2;0.10}^2$$

existem razões para duvidar de que a população X - "gasto semanal em alimentação (para famílias com dois filhos) em Agosto de 2003", tenha distribuição normal com desvio padrão conhecido e de valor $\sigma = 15$ euros.

Valor-p associado à nossa decisão:

$$p - \text{value} = P(X^2 > x_{obs}^2 | H_0 \text{ verdadeira}) = P(\chi_2^2 > 6.6350) = 0.0362,$$

que sendo inferior a $\alpha = 0.1$, confirma a decisão de rejeitar H_0 .

Observações: Neste exemplo, é discutível a utilização do teste de ajustamento do qui-quadrado porque a amostra pode não ser suficientemente grande.

Também se optou por não fundir as três primeiras classes de modo a atingir uma frequência esperada superior ou igual a 5, com o objectivo de não diminuir excessivamente o número final de classes. Esta prática é bastante comum, porque pode ser mais danoso usar classes em número muito reduzido do que cumprir estritamente a condição sobre o valor das frequências esperadas.

Nota: 5.1. Existem outros testes para testar a distribuição assumida para uma população (ou v.a.). Só a título de informação, não podemos deixar de referir o teste de Kolmogorov-Smirnov, particularmente conveniente para testar a distribuição de uma população contínua e o teste de Shapiro-Wilk exclusivamente para testar a normalidade de uma população.

Capítulo 6

Teste ao pressuposto de aleatoriedade das observações amostrais

Sistematicamente, em todos os tratamentos estatísticos referidos, viemos a assumir que a amostra (X_1, X_2, \dots, X_n) é uma amostra de v.a.'s. Face a uma amostra observada (x_1, x_2, \dots, x_n) podemos contestar se efectivamente ela resulta de observações obtidas de modo aleatório.

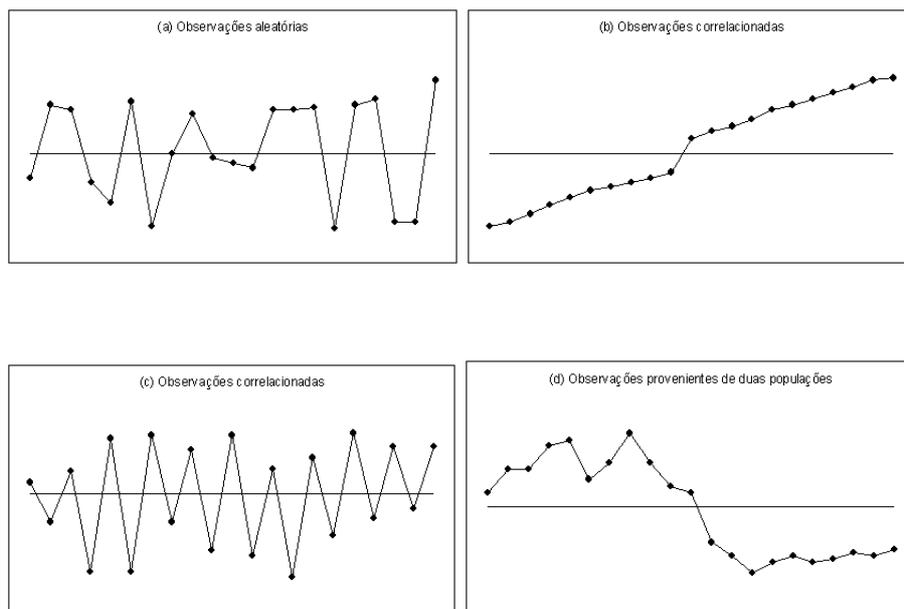
No fundo estamos a interrogar-nos sobre a validade de uma de duas hipóteses:

H_0 : X_1, X_2, \dots, X_n são v.a.'s *vs* H_1 : X_1, X_2, \dots, X_n não são v.a.'s

e que muitas vezes são apresentadas de modo mais simples (ainda que não totalmente correcto)

H_0 : A amostra é aleatória *vs* H_1 : A amostra não é aleatória

Figura 6.1: Amostras aleatórias e não aleatórias



A não aleatoriedade pode verificar-se de muitas maneiras: Nas figuras acima ilustram-se algumas destas situações. Quando a amostra é aleatória, a linha que une as sucessivas observações cruza

frequentemente a mediana amostral (Figura (a)). Contudo, se o número de cruzamentos é demasiado elevado (Figura (c)), isto significa que uma observação acima da linha mediana se segue uma abaixo dessa linha, a que se seguirá outra de novo acima da linha mediana, e assim sucessivamente para a maioria das observações amostrais. Ora isto não é consentâneo com uma situação de aleatoriedade. Também as situações expostas nas figuras (b) e (d), correspondem a um número diminuído de cruzamentos e ilustram outras formas de não aleatoriedade da amostra.

Existem inúmeros testes estatísticos que permitem testar estas hipóteses. Vamos aqui referir apenas um, habitualmente denominado **teste das sequências ascendentes e descendentes**.

Define-se por **sequência** um conjunto de observações consecutivas e idênticas. Consideremos a amostra resultante da classificação de 10 objectos em G -grande e P -pequeno,

G G P P G P G G P G

Nesta amostra aparecem 7 sequências, nomeadamente, GG , PP , G , P , GG , P e G .

De acordo com as observações feitas sobre as características de aleatoriedade de uma amostra, concluímos que uma amostra com um número de sequências demasiado elevado ou demasiado pequeno, não deve ser aleatória.

O teste das sequências ascendentes e descendentes aplica-se a amostras em que os dados são expressos, pelo menos, numa escala ordinal (ou seja, podem ser ordenados).

Uma **sequência ascendente** é uma sequência de observações crescentes e uma **sequência descendente** é uma sequência de observações decrescentes. Sempre que a ordenação altera o seu sentido, começa uma nova sequência.

Por exemplo, se na amostra

(9.4, 3.2, 3.5, 4.4, 4.3, 5.2, 5.4, 6.8, 3.3)

substituímos pelo símbolo + cada observação que é precedida por uma de valor inferior, e pelo símbolo – cada observação que é precedida por outra de valor superior, obtemos

(·, –, +, +, –, +, +, +, –)

onde se registam 5 sequências.

O teste baseia-se no número V de sequências observadas numa amostra de N símbolos.

Neste exemplo, $N = 8$, e V tem um valor observado $v = 5$.

Passemos a discutir a regra de rejeição da aleatoriedade. Quando a amostra não é aleatória, o número V de sequências é muito elevado ou muito pequeno. Assim, devemos

Rejeitar " H_0 : A amostra é aleatória", se $V \leq a$ ou se $V \geq b$.

Mas qual o valor de a e de b ?

A questão da determinação destas constantes vai ser contornada. Na aplicação prática deste teste, o procedimento será o seguinte:

- Determinamos o valor da **estatística de teste** V , que representamos por v_o ;
- Determinamos o valor $t_o = \min(v_o, N + 1 - v_o)$
- Na tabela estatística do teste das sequências ascendentes e descendentes, fazemos a leitura da probabilidade

$$P_{value} = P(V \leq t_o) + P(V \geq t_o)$$

Regra de decisão para um nível de significância α

$$\boxed{\text{Rejeitar } H_0 \text{ se } P_{value} < \alpha}$$

Na tabela estatística do teste das sequências ascendentes e descendentes, só encontramos valores de P_{value} para $N \leq 24$. Quando $N \geq 25$, podemos realizar um teste assintótico, usando a estatística de teste

$$Z = \frac{V - \frac{2N-1}{3}}{\sqrt{\frac{16N-29}{90}}} \xrightarrow{L} Z \sim N(0, 1)$$

e

Regra de decisão para um nível de significância α

$$\boxed{\text{Rejeitar } H_0 \text{ se } |z_{obs}| \geq z_{\alpha/2}}$$

Exemplo 6.1. Consideremos a amostra apresentada no exemplo 4.6 de medições de acidez (pH) de amostras de chuva registadas em 12 locais de uma região industrial:

$$\begin{array}{cccccc} 5.1 & 5.0 & 3.8 & 4.8 & 3.6 & 4.7 \\ 4.3 & 4.4 & 4.5 & 4.9 & 4.7 & 4.8 \end{array}$$

e testemos se estas observações constituem uma amostra aleatória.

A amostra de símbolos é

$$\cdot \quad - \quad - \quad + \quad - \quad + \quad - \quad + \quad + \quad + \quad - \quad + \quad \text{com } N = 11 \text{ símbolos e } v_o = 8 \text{ sequências.}$$

Então $t_o = \min(8, 11 + 1 - 8) = 4$ e $P_{value} = 0.5629$.

Decisão para um nível de significância $\alpha = 5\%$: Como $P_{value} = 0.5629 > 0.05 = \alpha$, não rejeitamos a hipótese de aleatoriedade das observações amostrais.

Exemplo 6.2. Consideremos a amostra apresentada no exemplo 5.3 (teste de ajustamento do qui-quadrado) de medições de acidez (pH) de amostras de chuva registadas em 30 locais de uma região industrial:

$$\begin{array}{cccccccccccccccc} 5.1 & 5.0 & 4.6 & 3.8 & 4.8 & 3.6 & 4.7 & 3.8 & 4.1 & 4.2 & 4.4 & 4.5 & 3.6 & 3.9 & 4.2 \\ 4.3 & 4.4 & 4.5 & 4.9 & 4.7 & 4.8 & 4.1 & 4.2 & 4.5 & 4.5 & 4.6 & 4.4 & 4.1 & 4.6 & 4.5 \end{array}$$

e testemos se estas observações constituem uma amostra aleatória.

A amostra de símbolos é

$$\begin{array}{cccccccccccccccc} \cdot & - & - & - & + & - & + & - & + & + & + & + & - & + & + \\ + & + & + & + & - & + & - & + & + & \cdot & + & - & - & + & - \end{array}$$

com $N = 28$ símbolos e $v_o = 15$ sequências.

Como $N \geq 25$, podemos usar o teste assintótico, para o qual

$$z_{obs} = \frac{15 - \frac{2 \times 28 - 1}{3}}{\sqrt{\frac{16 \times 28 - 29}{90}}} = -1.82$$

Decisão para um nível de significância $\alpha = 5\%$: Como $|z_{obs}| = 1.82 < z_{0.025} = 1.96$, não rejeitamos a hipótese de aleatoriedade das observações amostrais.

O p-value associado à nossa decisão é:

$p_{value} = P(|Z| > |z_{obs}|) = P(|Z| > 1.82) = 2P(Z > 1.82) = 2(1 - 0.9656) = 0.0688 > 0.05 = \alpha$, confirmando-se a decisão de não rejeição da hipótese de aleatoriedade das observações amostrais.

Capítulo 7

Regressão Linear Simples

7.1 Relação entre variáveis

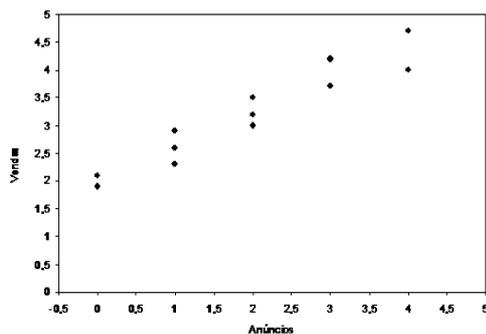
A regressão linear é uma técnica estatística que permite estudar a relação funcional entre uma variável Y (chamada *variável dependente*) e uma ou mais variáveis x, w, \dots (chamadas *variáveis independentes*). Pretendemos estabelecer uma relação funcional que possibilite explicar o valor da variável Y , uma vez conhecidos os valores das variáveis independentes x, w, \dots .

Evidentemente que, tratando-se de uma técnica estatística, a relação a estudar entre a variável dependente e as variáveis independentes é uma relação casuística (ou imprecisa). Ou seja, uma relação em que, para os mesmos valores das variáveis independentes, não é possível dizer exactamente qual o valor de Y .

Exemplo 7.1. Consideremos o seguinte conjunto de dados relativos ao volume mensal de vendas, Y (em milhares de unidades), de uma marca de computadores, e ao número de anúncios, x , que passaram diariamente na televisão em cada mês.

Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
x_i	4	1	3	0	2	4	2	3	1	2	0	1
y_i	4,0	2,3	3,7	2,1	3,0	4,7	3,5	4,2	2,9	3,2	1,9	2,6

O diagrama de dispersão destes dados revela a existência de uma relação linear de natureza probabilística.



7.2 Modelo de regressão linear simples

Quando existe apenas uma variável independente x e a sua relação funcional de natureza probabilística (ou aleatória) com a variável dependente Y é uma relação linear, o modelo matemático implícito é expresso por:

$$Y(x) = \beta_0 + \beta_1 x + E, \quad (7.2.1)$$

(onde E é uma v.a. que expressa a característica aleatória da relação) e dizemos que temos um *modelo de regressão linear simples*.

Dizemos que um modelo de regressão é linear, quando este for linear nos parâmetros que estabelecem a relação entre Y e a/s variáveis independentes x, w, \dots . Por exemplo, o modelo estocástico

$$Y(x) = \beta_0 + \beta_1 x^2 + E$$

é um modelo linear. Mas o modelo matemático

$$Y(x) = \beta_0 + x^{\beta_1} + E$$

já não é um modelo linear porque, apesar de ser linear relativamente a β_0 , já não o é relativamente a β_1 .

Por outro lado, o modelo 7.2.1 é um modelo de regressão *simples* porque nele consta apenas uma variável independente. Por exemplo o modelo de regressão linear

$$Y() = \beta_0 + \beta_1 x + \beta_2 w + E,$$

é dito um modelo de *regressão linear múltipla*.

Analisemos com mais detalhe o modelo de regressão linear simples

$$Y(x) = \beta_0 + \beta_1 x + E$$

A componente $\beta_0 + \beta_1 x$ é a componente determinística do modelo. A componente E expressa a natureza aleatória do modelo da variável dependente Y .

Assim, um modelo estatístico de regressão linear simples fica completo se considerarmos que:

- β_0 e β_1 são os parâmetros do modelo (chamados *coeficientes da regressão*);
- x é a *variável independente* (ou *variável controlada*);
- $Y(x)$ é a *variável dependente* (ou *variável resposta*) e trata-se de uma variável aleatória;
- E é o *erro* e trata-se de uma variável aleatória que se

– pressupõe ter distribuição normal de valor médio nulo e variância σ^2

$$E \sim N(0, \sigma^2).$$

β_0 é a ordenada na origem e β_1 é o declive da recta.

Nota: 7.1. $Y(x)$ acaba por ser uma variável aleatória porque, sendo o erro E a componente aleatória, então $Y(x) = \beta_0 + \beta_1 x + E$ é também variável aleatória.

Evidentemente que, se $E \sim N(0, \sigma^2)$ e, como $Y(x) = \beta_0 + \beta_1 x + E$, também Y tem distribuição normal com parâmetros:

$$\begin{aligned} E[Y(x)] &= E(\beta_0 + \beta_1 x + E) = \beta_0 + \beta_1 x + E(E) = \beta_0 + \beta_1 x \\ V[Y(x)] &= V(\beta_0 + \beta_1 x + E) = V(E) = \sigma^2 \end{aligned}$$

ou seja,

$$Y(x) \sim N(\beta_0 + \beta_1 x, \sigma^2).$$

Devemos também salientar que σ^2 é um parâmetro adicional do modelo que necessita ser estimado, caso não se conheça o seu valor.

7.3 Método dos mínimos quadrados para estimar β_0 e β_1

Aceitando um modelo de regressão linear simples

$$Y(x) = \beta_0 + \beta_1 x + E,$$

importa agora estimar a recta de regressão, ou seja encontrar estimadores para os parâmetros β_0 e β_1 .

Consideremos $((x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n))$ uma amostra aleatória de dimensão n , entendendo-se que

$$Y_i \equiv Y(x_i) = \beta_0 + \beta_1 x_i + E_i$$

Evidentemente que procuramos encontrar a recta que "melhor" se ajuste à amostra de n pares de observações $((x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n))$.

Assim deveremos encontrar estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, para os coeficientes da recta de regressão, para obtermos a recta estimada (ou ajustada)

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

De entre diversos métodos que existem para a dedução dos estimadores dos coeficientes da recta de regressão, vamos aqui abordar o intitulado *método dos mínimos quadrados*. Consiste este método, em determinar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, dos coeficientes de regressão, β_0 e β_1 , que permitam a minimização de

$$SQ(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

conduzam a uma recta que se ajusta ao conjunto de observações *minimizand*o a soma do quadrado dos desvios entre cada observação de (x_i, Y_i) e a recta *ajustada* $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \dots, n$.

Neste método, os desvios

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

são a diferença na vertical entre o valor da observação Y_i e a sua estimativa pela recta de regressão $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$,

A soma dos quadrados de todos os desvios representar-se-á por SQE e encontrar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, ditos *estimadores de mínimos quadrados* de β_0 e β_1 , respectivamente, consiste em resolver o problema

$$\text{minimizar } SQE = \sum_{i=1}^n \hat{E}_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

em ordem a $\hat{\beta}_0$ e $\hat{\beta}_1$.

Demonstra-se que esta minimização é conseguida resolvendo, em ordem a $\hat{\beta}_0$ e $\hat{\beta}_1$, o sistema de equações

$$\begin{cases} \frac{\partial}{\partial \hat{\beta}_0} SQE = 0 \\ \frac{\partial}{\partial \hat{\beta}_1} SQE = 0 \end{cases} \Leftrightarrow \begin{cases} -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2 \sum_{i=1}^n x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

As soluções deste sistema são:

$$\begin{cases} \hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i Y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n Y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \\ \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i \end{cases}$$

que podem ainda ser expressas por

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} \quad \text{e} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

considerando

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ média das observações de X
- $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ média da amostra aleatória de Y
- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$ soma de quadrados para X

- $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2$ soma de quadrados para Y
- $S_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}$ soma de produtos cruzados para (X, Y)

A soma dos quadrados dos desvios pode ainda ser escrita

$$SQE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = S_{YY} - \frac{S_{xY}^2}{S_{xx}} = S_{YY} - \hat{\beta}_1^2 S_{xx}.$$

Dá-se o nome de *recta de regressão de mínimos quadrados* ao estimador da recta de regressão

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{Y} + \hat{\beta}_1 (x - \bar{x}).$$

As estimativas desta recta para as observações x_1, x_2, \dots, x_n da variável independente x serão

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n,$$

em que, sem perda de generalidade, $\hat{\beta}_0$ e $\hat{\beta}_1$ são as estimativas de mínimos quadrados de β_0 e β_1 , respectivamente.

NOTA IMPORTANTE: Só devemos usar esta recta para fazer previsão dos valores da variável resposta para valores de x que estejam dentro do intervalo das observações obtidas para x .

Aos desvios

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

dá-se o nome de *resíduos* e

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

serão os *resíduos observados*.

7.4 Estimação da variância do erro σ^2 e qualidade do ajustamento

Uma vez obtida a recta de regressão de mínimos quadrados e com os valores que ela fornece para cada observação x_i da variável controlada, podemos utilizar os resíduos

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \dots, n$$

para analisar a qualidade do ajustamento da recta.

O que a recta não consegue explicar sobre os valores de $Y(x)$, é considerado observação do erro E e, pode ser usado para estimarmos a variância desse erro. Assim sendo, os resíduos também permitem estimar a variância σ^2 .

7.4.1 Estimador para σ^2

Um estimador de σ^2 é

$$\hat{\sigma}^2 = \frac{SQE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{S_{YY} - \hat{\beta}_1^2 S_{xx}}{n-2}$$

Quando o erro E tem distribuição $N(0, \sigma^2)$,

- $\hat{\sigma}^2 = \frac{SQE}{n-2}$ é um estimador centrado para σ^2
- $(n-2) \frac{\hat{\sigma}^2}{\sigma^2}$ tem distribuição do qui-quadrado com $(n-2)$ graus de liberdade, $(n-2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$

7.4.2 Qualidade do ajustamento

Quanto menores forem os valores dos resíduos

$$\hat{E}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \dots, n$$

melhor é o ajustamento da recta de mínimos quadrados. Por esta razão podemos dizer que, quanto menor for o valor da soma do quadrado dos resíduos, SQE , melhor é o ajustamento.

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 \\ &= \sum_{i=1}^n [(Y_i - \hat{Y}_i)]^2 + \sum_{i=1}^n [(\hat{Y}_i - \bar{Y})]^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{Y}_i - \bar{Y}) \\ &= SQE + \sum_{i=1}^n [\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y}]^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i) (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y}) \\ &= SQE + \sum_{i=1}^n [\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{Y}]^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i) (\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{Y}) \\ &= SQE + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Definição 7.1. Dá-se o nome de *coeficiente de determinação* a

$$R^2 = 1 - \frac{SQE}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{S_{YY} - \hat{\beta}_1^2 S_{xx}}{S_{YY}} = \hat{\beta}_1^2 \frac{S_{xx}}{S_{YY}} = \frac{S_{xY}^2}{S_{xx} S_{YY}}$$

que assume valores $0 \leq R^2 \leq 1$.

Nota: 7.2. A soma de quadrados $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ reflecte a variabilidade de $Y(x)$ quando não se entra em linha de conta com a sua eventual relação com a variável x . Por outro lado, SQE reflecte a variabilidade de $Y(x)$ quando é usado o modelo de regressão para explicar os valores de $Y(x)$ como resposta a x . Por fim, $S_{YY} - SQE$ mede a redução na variabilidade total de $Y(x)$ ao usar x para explicar a resposta $Y(x)$. Então, ao dividirmos $S_{YY} - SQE$ por S_{YY} , obtemos um estimador da redução relativa da variabilidade ao usarmos o modelo para explicarmos $Y(x)$ como função linear de x .

Nota: 7.3. O coeficiente de determinação R^2 , assume valores compreendidos entre zero e um. Vejamos a interpretação que pode ser dada ao seu valor.

•

$$\begin{aligned} \text{Se } R^2 = 1 &\Leftrightarrow SQE = 0 \Leftrightarrow \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \\ &\Leftrightarrow Y_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n \\ &\Leftrightarrow \text{ajuste perfeito} \end{aligned}$$

Conclusão: $R^2 = 1$ quando todas as observações estão sobre a recta de mínimos quadrados (ajustamento perfeito).

•

$$\begin{aligned} \text{Se } R^2 = 0 &\Leftrightarrow SQE = \sum_{i=1}^n (Y_i - \bar{Y})^2 \Leftrightarrow SQE = S_{YY} \Leftrightarrow S_{YY} - \hat{\beta}_1^2 S_{xx} = S_{YY} \\ &\Leftrightarrow \hat{\beta}_1 = 0 \\ &\Leftrightarrow \text{a variável } x \text{ não serve para explicar } Y(x) \end{aligned}$$

Conclusão: $R^2 = 0$ quando o modelo de regressão linear em x não tem utilidade ou seja, a variável x não consegue explicar os valores de $Y(x)$.

Em resumo: Quanto mais próximo R^2 estiver de 1, maior o grau de importância de x na determinação da variável resposta $Y(x)$.

Na prática, consideramos que o ajustamento é razoável se $R^2 \geq 0.8$

7.5 Distribuição de amostragem dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$

7.5.1 Distribuição de amostragem de $\hat{\beta}_1$

Para a dedução da distribuição do estimador $\hat{\beta}_1$ para o coeficiente de regressão β_1 , começamos por o expressar de outro modo. Ora

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$$

mas como

$$S_{xY} = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n x_i Y_i - \bar{x} \sum_{i=1}^n Y_i = \sum_{i=1}^n (x_i - \bar{x}) Y_i,$$

então

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}.$$

Como as observações x_i , $i = 1, \dots, n$ são constantes, também S_{xx} o é, e portanto $\hat{\beta}_1$ não é mais do que uma combinação linear de v.a.'s (Y_i , $i = 1, \dots, n$) independentes e com distribuição normal. Consequentemente $\hat{\beta}_1$ tem distribuição normal, restando saber qual o correspondente valor médio e variância.

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x}) E(Y_i)}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i)}{S_{xx}} = \\ &= \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i}{S_{xx}} = \\ &= \frac{\beta_0 (n\bar{x} - n\bar{x}) + \beta_1 (\sum_{i=1}^n x_i^2 - \bar{x}n\bar{x})}{S_{xx}} = \beta_1 \frac{S_{xx}}{S_{xx}} = \beta_1 \end{aligned}$$

Logo $\hat{\beta}_1$ é estimador centrado para β_1 .

$$\begin{aligned} V(\hat{\beta}_1) &= V\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{S_{xx}}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 V(Y_i)}{S_{xx}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{S_{xx}^2} = \\ &= \sigma^2 \frac{S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Em resumo:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Contudo, na maioria das aplicações, a variância σ^2 dos erros não é conhecida. Nestes casos, podemos estimá-la por $\hat{\sigma}^2 = \frac{SQE}{n-2}$. A substituição de σ^2 pelo seu estimador $\hat{\sigma}^2$ obriga-nos a considerar a seguinte distribuição para $\hat{\beta}_1$:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \sqrt{S_{xx}} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim t_{n-2}$$

(T tem distribuição t com $(n-2)$ graus de liberdade).

7.5.2 Distribuição de amostragem de $\hat{\beta}_0$

Para a dedução da distribuição do estimador $\hat{\beta}_0$ para o coeficiente de regressão β_0 , recordemos que

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Como $\hat{\beta}_1$ tem distribuição normal e \bar{Y} também tem distribuição normal (é uma média aritmética de v.a.'s com distribuição normal), então $\hat{\beta}_0$ tem distribuição normal. Resta saber qual o correspondente valor médio e variância.

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1 \bar{x}) = E(\bar{Y}) - \bar{x}E(\hat{\beta}_1) = \frac{1}{n} \sum_{i=1}^n E(Y_i) - \beta_1 \bar{x} = \\ &= \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

Logo $\hat{\beta}_0$ é estimador centrado para β_0 .

$$\begin{aligned} V(\hat{\beta}_0) &= V(\bar{Y} - \hat{\beta}_1 \bar{x}) = V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} \text{cov}(\bar{Y}, \hat{\beta}_1) = \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} - 2\bar{x} \text{cov}(\bar{Y}, \hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{n} \left(1 + \frac{n\bar{x}^2}{S_{xx}} \right) = \\ &= \frac{\sigma^2}{nS_{xx}} (S_{xx} + n\bar{x}^2) = \frac{\sigma^2}{nS_{xx}} \left(\left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2 + n\bar{x}^2 \right) = \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n x_i^2 \end{aligned}$$

porque

$$\begin{aligned} \text{cov}(\bar{Y}, \hat{\beta}_1) &= \text{cov}\left(\bar{Y}, \frac{S_{xY}}{S_{xx}}\right) = \frac{1}{S_{xx}} \text{cov}(\bar{Y}, S_{xY}) = \\ &= \frac{1}{S_{xx}} \text{cov}\left(\frac{1}{n} \sum_{i=1}^n Y_i, \sum_{i=1}^n (x_i - \bar{x}) Y_i\right) = \text{pela independência de } Y_i \\ &= \frac{1}{nS_{xx}} \sum_{i=1}^n \text{cov}(Y_i, (x_i - \bar{x}) Y_i) = \frac{1}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \text{cov}(Y_i, Y_i) \\ &= \frac{1}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) V(Y_i) = \frac{1}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) \sigma^2 \\ &= \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n (x_i - \bar{x}) = \frac{\sigma^2}{nS_{xx}} (n\bar{x} - n\bar{x}) = 0 \end{aligned}$$

Em resumo:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^n x_i^2\right)$$

Sendo desconhecida a variância σ^2 dos erros, podemos estimá-la por $\hat{\sigma}^2 = \frac{SQE}{n-2}$. A substituição de σ^2 pelo seu estimador $\hat{\sigma}^2$ obriga-nos a considerar a seguinte distribuição para $\hat{\beta}_0$:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{nS_{xx}}}} = \sqrt{\frac{nS_{xx}}{\sum_{i=1}^n x_i^2}} \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \sim t_{n-2}$$

(T tem distribuição t com $(n-2)$ graus de liberdade)

7.6 Inferência sobre os parâmetros do modelo

7.6.1 Inferência sobre β_1

Estimação de β_1 por intervalo de confiança

β_1 é o declive da recta de regressão e, como tal mede o grau de crescimento de $Y(x)$ relativamente aos valores de x . Assim pode ser importante fazer a sua estimação pelos dois processos que conhecemos.

- Estimação Pontual:

Considere-se o estimador de mínimos quadrados

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$$

- Estimação por intervalo de confiança $(1 - \alpha)$.

Para o estimador $\hat{\beta}_1$ de β_1 :

A estatística pivot que devemos usar é: $T = \sqrt{S_{xx}} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim t_{n-2}$.

Dada a simetria em zero da distribuição t , podemos desde logo afirmar que:

$$\begin{aligned} 1 - \alpha &= P(-t_{n-2;\alpha/2} \leq T \leq t_{n-2;\alpha/2}) \Leftrightarrow 1 - \alpha = P\left(-t_{n-2;\alpha/2} \leq \sqrt{S_{xx}} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \leq t_{n-2;\alpha/2}\right) \\ &\Leftrightarrow 1 - \alpha = P\left(\hat{\beta}_1 - t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) \end{aligned}$$

Assim

$$\text{Intervalo de confiança } (1 - \alpha) \text{ para o declive } \beta_1 \\ IC_{1-\alpha}(\beta_1) \equiv \left[\hat{\beta}_1 - t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \hat{\beta}_1 + t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \right]$$

Teste de hipóteses sobre β_1

Já atrás dissemos que β_1 é o declive da recta de regressão e, como tal mede o grau de crescimento de Y relativamente aos valores de x . De particular importância é o caso em que $\beta_1 = 0$. Quando tal acontece, a variável x não é responsável na justificação dos valores de Y . Assim o teste das hipóteses

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

permite testar esta situação.

Mas o teste destas hipóteses inclui-se no teste mais genérico das hipóteses

$$H_0 : \beta_1 = a \text{ vs } H_1 : \beta_1 \neq a$$

que passamos a deduzir.

A estatística de teste é: $T = \sqrt{S_{xx}} \frac{\hat{\beta}_1 - a}{\hat{\sigma}} \underset{\beta_1=a}{\sim} t_{n-2}$

A regra de rejeição, para um nível de significância α é: Rejeitar H_0 se $|T| > c$, $c > 0$

Determinemos o valor de c :

$$\begin{aligned} \alpha &= P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(|T| > c) = P(T < -c) + P(T > c) = \\ &= P(T > c) + P(T > c) = 2P(T > c) \Leftrightarrow P(T > c) = \alpha/2 \Leftrightarrow c = t_{n-2;\alpha/2} \end{aligned}$$

Regra de decisão para um nível de significância α

$$\text{Rejeitar } H_0 \text{ se } |T| = \left| \sqrt{S_{xx}} \frac{\hat{\beta}_1 - a}{\hat{\sigma}} \right| > t_{n-2;\alpha/2}$$

Nota: 7.4. Também se podem deduzir testes de hipóteses unilaterais sobre β_1 , se bem que não têm tanto interesse nas aplicações aos modelos de regressão linear simples.

7.6.2 Inferência sobre β_0

Estimação de β_0 por intervalo de confiança

β_0 é o ponto de intersecção da recta com o eixo das abcissas. A inferência sobre este parâmetro não tem a mesma importância que tem a inferência sobre o declive β_1 da recta de regressão. Mas ainda assim, pode ser necessário estimar β_0 por intervalo de confiança e realizar testes de hipóteses sobre valores que forneçam respostas a questões de utilidade prática.

A estimação pontual feita pelo estimador de mínimos quadrados, é:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Passemos à estimação por intervalo de confiança $(1 - \alpha)$.

A estatística pivot que devemos usar é: $T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{nS_{xx}}}} = \sqrt{\frac{nS_{xx}}{\sum_{i=1}^n x_i^2}} \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \sim t_{n-2}$.

Dada a simetria em zero da distribuição t , podemos desde logo afirmar que:

$$\begin{aligned} 1 - \alpha &= P(-t_{n-2;\alpha/2} \leq T \leq t_{n-2;\alpha/2}) \Leftrightarrow 1 - \alpha = P\left(-t_{n-2;\alpha/2} \leq \sqrt{\frac{nS_{xx}}{\sum_{i=1}^n x_i^2}} \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \leq t_{n-2;\alpha/2}\right) \\ &\Leftrightarrow 1 - \alpha = P\left(\hat{\beta}_0 - t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{nS_{xx}}} \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2;\alpha/2} \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{nS_{xx}}}\right) \end{aligned}$$

Assim

$$\text{Intervalo de confiança } (1 - \alpha) \text{ para o declive } \beta_0$$

$$IC_{1-\alpha}(\beta_0) \equiv \left[\hat{\beta}_0 - t_{n-2;\alpha/2} \sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}, \hat{\beta}_0 + t_{n-2;\alpha/2} \sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}} \right]$$

Teste de hipóteses sobre β_0

Consideremos as hipóteses

$$H_0 : \beta_0 = a \text{ vs } H_1 : \beta_0 \neq a$$

que passamos a deduzir.

$$\text{A estatística de teste é: } T = \sqrt{\frac{nS_{xx}}{\sum_{i=1}^n x_i^2}} \frac{\hat{\beta}_0 - a}{\hat{\sigma}} \underset{\beta_0=a}{\sim} t_{n-2}$$

A regra de rejeição, para um nível de significância α é: Rejeitar H_0 se $|T| > c$, $c > 0$
Determinemos o valor de c :

$$\begin{aligned} \alpha &= P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(|T| > c) = P(T < -c) + P(T > c) = \\ &= P(T > c) + P(T > c) = 2P(T > c) \Leftrightarrow P(T > c) = \alpha/2 \Leftrightarrow c = t_{n-2;\alpha/2} \end{aligned}$$

Regra de decisão para um nível de significância α

$$\text{Rejeitar } H_0 \text{ se } |T| = \left| \sqrt{\frac{nS_{xx}}{\sum_{i=1}^n x_i^2}} \frac{\hat{\beta}_0 - a}{\hat{\sigma}} \right| > t_{n-2;\alpha/2}$$

Nota: 7.5. Também se podem deduzir testes de hipóteses unilaterais sobre β_0 , aplicando o mesmo tipo de conceitos e raciocínios que surgiram nas secções 4.7.2 e 4.7.3.

7.6.3 Inferência sobre σ^2

Estimação de σ^2 por intervalo de confiança

Num modelo de regressão linear simples

$$Y(x) = \beta_0 + \beta_1 x + E,$$

o erro E é a componente aleatória que a parte $\beta_0 + \beta_1 x$ não consegue explicar.

Os pressupostos estocásticos do modelo de regressão linear estabelecem que

$$E \sim N(0, \sigma^2).$$

O que a recta não consegue explicar sobre os valores de $Y(x)$, é considerado observação do erro E e, pode ser usado para estimarmos a variância desse erro. O estimador para σ^2 já foi apresentado na secção 7.4.1. Aí foi dito que, um estimador centrado de σ^2 é

$$\hat{\sigma}^2 = \frac{SQE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{S_{YY} - \hat{\beta}_1^2 S_{xx}}{n-2}$$

e, dado que o erro E tem distribuição $N(0, \sigma^2)$,

- $(n-2) \frac{\hat{\sigma}^2}{\sigma^2}$ tem distribuição do qui-quadrado com $(n-2)$ graus de liberdade, $(n-2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$

Com esta distribuição de amostragem, podemos deduzir um intervalo de confiança $(1-\alpha)$ para a variância σ^2 e para o desvio padrão σ . Usando argumentos idênticos aos apresentados na secção 3.3,

$$\begin{aligned} 1-\alpha &= P\left(\chi_{n-2;1-\alpha/2}^2 \leq (n-2) \frac{\hat{\sigma}^2}{\sigma^2} \leq \chi_{n-2;\alpha/2}^2\right) \Leftrightarrow \\ &\Leftrightarrow 1-\alpha = P\left(\frac{(n-2) \hat{\sigma}^2}{\chi_{n-2;\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-2) \hat{\sigma}^2}{\chi_{n-2;1-\alpha/2}^2}\right) \end{aligned}$$

Assim

Intervalo de confiança $(1-\alpha)$ para σ^2

$$IC_{1-\alpha}(\sigma^2) \equiv \left[\frac{(n-2) \hat{\sigma}^2}{\chi_{n-2;\alpha/2}^2}, \frac{(n-2) \hat{\sigma}^2}{\chi_{n-2;1-\alpha/2}^2} \right]$$

7.7 Estimação do valor esperado de Y para uma observação x_0 da variável controlada

O valor esperado de Y para uma observação x_0 da variável controlada é

$$E[Y(x_0)] = \beta_0 + \beta_1 x_0.$$

que podemos querer estimar.

O estimador pontual para $E(Y|x_0)$ é naturalmente

$$\hat{E}[Y(x_0)] = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

e trata-se de um estimador centrado com distribuição

$$\hat{E}[Y(x_0)] \sim N\left(E[Y(x_0)], \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

Caso a variância σ^2 do erro E não seja conhecida (como é usual nas aplicações), a distribuição de amostragem de $\hat{E}[Y(x_0)]$ é

$$T = \frac{\hat{E}[Y(x_0)] - E[Y(x_0)]}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

que permite deduzir um intervalo de confiança $(1 - \alpha)$ para $E[Y(x_0)]$.

$$IC_{1-\alpha}[E(Y(x_0))] \equiv \left[\hat{E}[Y(x_0)] - t_{n-2;\alpha/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}, \hat{E}[Y(x_0)] + t_{n-2;\alpha/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

Intervalo de confiança $(1 - \alpha)$ para $E(Y|x_0)$

MUITO IMPORTANTE: Só devemos fazer estimação de $E[Y(x_0)]$ para valores x_0 que estejam dentro do intervalo das observações obtidas para x .

7.8 Previsão do valor da variável resposta Y para um novo valor de x_0 da variável controlada

Dada um valor x_0 da variável controlada x , a variável resposta será

$$Y(x_0) = \beta_0 + \beta_1 x_0 + E,$$

onde $E \sim N(0, \sigma^2)$.

O estimador de $Y(x)$, para um valor x_0 , será

$$\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Trata-se de um estimador centrado, com variância $V[\hat{Y}(x_0)] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$. Quando não se conhece o valor de σ^2 , podemos estimar a variância de \hat{Y} por

$$\hat{V}[\hat{Y}(x_0)] = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Nestas circunstâncias, a distribuição de amostragem de \hat{Y} é

$$T = \frac{\hat{Y}(x_0) - Y(x_0)}{\sqrt{\hat{V}(\hat{Y}(x_0))}} = \frac{\hat{Y}(x_0) - Y(x_0)}{\sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

que permite deduzir um intervalo de confiança $(1 - \alpha)$ para Y .

$$IC_{1-\alpha}(Y(x_0)) \equiv \left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2;\alpha/2} \sqrt{\hat{V}(\hat{Y}(x_0))}, \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2;\alpha/2} \sqrt{\hat{V}(\hat{Y}(x_0))} \right]$$

Intervalo de confiança $(1 - \alpha)$ para $Y(x_0)$

MUITO IMPORTANTE: Só devemos fazer estimação de $Y(x_0)$ para valores x_0 que estejam dentro do intervalo das observações obtidas para x .

Exemplo 7.2. Retomemos o exemplo 7.1 e o conjunto de dados relativos ao volume de vendas mensal (em milhares de unidades) de uma marca de computadores, Y e ao número de anúncios que passaram diariamente na televisão em cada mês, x .

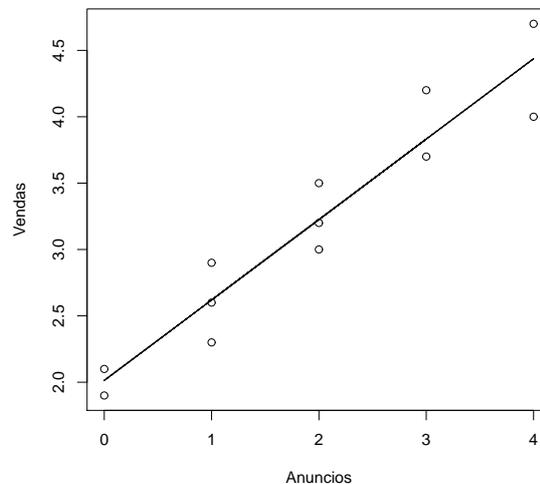
Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
x_i	4	1	3	0	2	4	2	3	1	2	0	1
y_i	4,0	2,3	3,7	2,1	3,0	4,7	3,5	4,2	2,9	3,2	1,9	2,6

Comecemos por estimar a recta de mínimos quadrados. Para tal, vamos usar um método de cálculo bastante rudimentar (que seria o que usaríamos caso a nossa ferramenta de cálculo fosse pouco evoluída).

$$\begin{array}{lllll} \sum_{i=1}^{12} x_i = 23 & \sum_{i=1}^{12} y_i = 38.1 & \sum_{i=1}^{12} x_i^2 = 65 & \sum_{i=1}^{12} y_i^2 = 129.39 & \sum_{i=1}^{12} x_i y_i = 85.7 \\ \bar{x} = 1.9167 & \bar{y} = 3.175 & S_{xx} = 20.9167 & S_{YY} = 8.4225 & S_{xY} = 12.675 \\ \hat{\beta}_0 = 2.013545817 & \hat{\beta}_1 = 0.6059760956 & SQE = 0.7417529885 & \hat{\sigma}^2 = 0.07417529885 & \end{array}$$

Assim a recta estimada é

$$\hat{y}(x) = 2.013545817 + 0.6059760956 x$$



Se estivermos interessados em estimar o volume mensal de vendas num mês em que fossem exibidos 2.5 anúncios teríamos uma estimativa pontual

$$\hat{y}(2.5) = 3.528486056 \text{ milhares de unidades}$$

Verifiquemos agora a qualidade do ajuste, calculando o coeficiente de determinação, R^2 :

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{YY}} = 0.9119319693$$

revela um bom ajustamento do modelo de regressão linear ao conjunto de dados.

Podemos ainda testar se o número de anúncios que passam por mês, x , explicam significativamente o volume de vendas. Trata-se de testar, ao nível de 5% de significância, as hipóteses

•

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

- Estatística de teste: $T = \sqrt{S_{xx}} \frac{\hat{\beta}_1}{\hat{\sigma}} \underset{\beta_1=0}{\sim} t_{12-2} \equiv t_{10}$
- Para $\alpha = 5\%$, $t_{10;0.025} = 2.2281$ e $R_{0.05} =]-\infty, -2.2281[\cup]2.2281, +\infty[$.
- Valor observado da estatística de teste: $t_{obs} = \sqrt{S_{xx}} \frac{\hat{\beta}_1}{\hat{\sigma}} = 10.17588237$.
- Decisão: Como $t_{obs} = 10.17588237 \in R_{0.05}$ decidimos rejeitar H_0 , com 5% de significância. Dito de outro modo, com 5% de significância, não existe evidência para afirmar que $\beta_1 = 0$ e portanto podemos inferir que o número de anúncios que passam mensalmente é uma variável que explica o volume de vendas para esse mês.

Estimemos por intervalo de 90% de confiança:

1. o volume esperado de vendas num mês em que fossem exibidos diariamente 1.5 anúncios, $E[Y(1.5)]$;
2. o volume de vendas num mês em que fossem exibidos diariamente 1.5 anúncios, $Y(1.5)$.

Nas duas situações devemos considerar $x_0 = 1.5$ e $t_{10;0.05} = 1.812$.

1. Com $\hat{E}[Y(1.5)] = 2.013545817 + 0.6059760956 \times 1.5 = 2.92250996$, obteríamos uma banda de valores compreendidos entre o limite inferior

$$2.92250996 - 1.812 \sqrt{0.07417529885 \left(\frac{1}{12} + \frac{(1.5 - 1.9167)^2}{20.9167} \right)} = 2.773121167$$

e o limite superior

$$2.92250996 + 1.812 \sqrt{0.07417529885 \left(\frac{1}{12} + \frac{(1.5 - 1.9167)^2}{20.9167} \right)} = 3.071898754$$

ou seja o intervalo $IC_{90\%}[E(Y(1.5))] \equiv [2.773121167, 3.071898754]$ milhares de unidades de vendas esperadas.

2. Com $\hat{Y}(1.5) = 2.013545817 + 0.6059760956 \times 1.5 = 2.92250996$, obteríamos uma banda de valores compreendidos entre o limite inferior

$$2.92250996 - 1.812 \sqrt{0.07417529885 \left(1 + \frac{1}{12} + \frac{(1.5 - 1.9167)^2}{20.9167} \right)} = 2.406893791$$

e o limite superior

$$2.92250996 + 1.812 \sqrt{0.07417529885 \left(1 + \frac{1}{12} + \frac{(1.5 - 1.9167)^2}{20.9167} \right)} = 3.43812613$$

ou seja o intervalo $IC_{90\%}(Y(1.5)) \equiv [2.406893791, 3.43812613]$ milhares de unidades de vendas.

Por fim podemos ainda calcular os resíduos observados

Tabela de resíduos												
Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
x_i	4	1	3	0	2	4	2	3	1	2	0	1
\hat{e}_i	-0.44	-0.32	-0.13	0.09	-0.23	0.26	0.27	0.37	0.28	-0.03	-0.11	-0.02

A esta amostra de resíduos podíamos aplicar um teste de ajustamento do qui-quadrado para uma distribuição normal, de modo a testar a validade do pressuposto estocástico do modelo, segundo o qual, estes resíduos deverão ser observações do erro, ou seja, observações de uma v.a. $E \sim N(0, \sigma^2)$.

7.9 Alguns modelos linearizáveis

7.9.1 O modelo Power e a relação log-log

$$z(w) = \gamma w^\theta \Leftrightarrow \ln z(w) = \ln \gamma + \theta \ln w, \quad \text{se } z(w) > 0$$

Considerando $x = \ln w$, $y(x) = \ln z(w)$, $\beta_0 = \ln \gamma$ e $\beta_1 = \theta$, podemos estabelecer o modelo de regressão linear simples

$$Y(x) = \beta_0 + \beta_1 x + E$$

sendo E uma v.a. com distribuição $N(0, 1)$.

7.9.2 O modelo "Exponencial" e a relação log-lin

$$z(w) = \gamma \theta^w \Leftrightarrow \ln z(w) = \ln \gamma + \ln \theta w, \quad \text{se } \gamma > 0 \text{ e } \theta > 0$$

Considerando $x = w$, $y(x) = \ln z(w)$, $\beta_0 = \ln \gamma$ e $\beta_1 = \ln \theta$, podemos estabelecer o modelo de regressão linear simples

$$Y(x) = \beta_0 + \beta_1 x + E$$

sendo E uma v.a. com distribuição $N(0, 1)$.