ANÁLISE ESTATÍSTICA DE MEDIDAS EM CIÊNCIAS EXATAS

Vito R. Vanin, Philippe Gouffon, Otaviano Helene

Março 2023

Capítulo 5

Inferência estatística e teste de hipótese

Teste de hipótese, juntamente com a estimativa de parâmetros, representa a maior parte do objeto da *inferência estatística*. A definição formal de *hipótese estatística* — afirmação acerca da f.d.p. de uma ou mais variáveis aleatórias — não ajuda muito a compreender este assunto, de maneira que iniciamos o capítulo diretamente com um exemplo de uma hipótese e o teste (estatístico) relacionado.

Ao longo deste capítulo, vamos supor **sempre** que os dados obtidos sejam estatisticamente independentes e tenham f.d.p. *normal*. Completaremos aqui o estudo das estatísticas calculadas com dados gaussianos, cujas f.d.p.s podem ser obtidas em forma analítica fechada, deduzindo a f.d.p. de F de Fisher, também chamada de F de Fisher-Snedecor, que se aplica à razão entre duas estimativas da variância de uma grandeza. As outras estatísticas que têm f.d.p.s em forma fechada são: a média, $\bar{x}$; a variância, σ^2, e a razão $\frac{\bar{x}-x_0}{\sigma_m}$ (diferença da média e o valor verdadeiro em relação ao desvio padrão da média), cujas f.d.p.s são, respectivamente, a normal (seção 2.6), a de χ^2 (seção 2.9) e a de t de Student (seção 3.5). O conhecimento tão detalhado e em forma fechada das f.d.p.s das estatísticas mais comuns só é possível com dados normais e certamente contribui muito para nossa tendência em adotar essa hipótese.

Descreveremos aqui um pouco da teoria geral do teste de hipótese, mas, principalmente, apresentaremos os testes relacionados às variáveis t, F e χ^2. Repetindo um pouco o que já foi dito no parágrafo precedente, limitamo-nos ao modelo normal, porque conhecemos bem as f.d.p.s das estatísticas, o que facilita a aplicação de um teste de hipótese.

5.1 O teste z

Em muitas situações, pode-se tomar decisões com a aplicação quase direta da tabela de probabilidades da distribuição normal (tabela **??** do Apêndice). Os exemplos abaixo ilustram como se pode fazer isso e procuram destacar a importância de refletir sobre as hipóteses subjacentes, que algumas vezes fazemos de maneira implícita. A simplicidade aparente do problema é enganosa: como é muito mais fácil raciocinar de maneira determinística que probabilística, sempre gostaríamos de saber se alguma coisa é ou não é, mas nas situações abaixo isso é impossível, e teremos que nos contentar em determinar a *probabilidade de acertar* as conclusões.

Um exemplo (fictício) é o do fabricante de rebites com 3 mm de diâmetro. Para verificar o processo, ele mede o diâmetro de 30 peças, cujos valores estão na tabela (5.1). Ao inspecionar os dados, notamos que há mais valores maiores que 3 mm do que menores, o que poderia sugerir que os rebites estejam um pouco grandes. A fim de tornar essa discussão quantitativa, reduzimos os dados do experimento às grandezas estatísticas que contêm toda a informação do conjunto de dados: a média ($\bar{d} = 3,0023$ mm), o desvio-padrão, $\sigma = 0,0106$ mm e o desvio-padrão da média, $\sigma_m = \sigma/\sqrt{N} = 0,0019$ mm. Se os dados obedecem à distribuição normal, a média também, conforme (2.27); no entanto, mesmo que os dados não tenham f.d.p. normal, a média de muitos dados tem distribuição normal, por conta do Teorema Central do Limite, que demonstraremos mais adiante (seção 6.16), de modo que a normalidade dos dados não é muito importante para a validade das conclusões que alcançaremos.

Em seguida, enunciamos a hipótese H_0: diâmetro = $d_0 = 3,0000$ mm, o

Tabela 5.1: Resultados das medições dos diâmetros de 30 rebites, em mm.

2,989	3,012	3,000	3,011	3,003
3,009	3,000	3,014	3,009	2,989
2,979	3,011	3,014	3,006	3,024
2,995	3,009	2,980	3,015	2,990
3,006	3,011	2,998	2,996	2,995
2,993	3,000	2,998	3,007	3,006

que permite calcular a estatística z,

$$z = \frac{\bar{d} - d_0}{\sigma_m} \quad . \tag{5.1}$$

Como foram tomados muitos dados para estimar σ_m, pode-se ignorar sua flutuação estatística e avaliar que z tem distribuição aproximadamente normal com desvio-padrão 1, ou seja, $N(z; 0, 1)$, de modo que a integral dessa função (tabela **??**) permite calcular as probabilidades associadas a intervalos de valores de z. Na próxima seção, veremos o que fazer quando σ_m for estimado a partir de poucos dados ou se a situação for tão delicada que não seja adequado ignorar a flutuação estatística de σ_m.

Após definir a estatística e sua distribuição de probabilidade, a próxima etapa do teste consiste em escolher seu *nível de significância*, simbolizado abaixo por α, que corresponderá *à probabilidade de rejeitar a hipótese se ela for verdadeira*; elaboraremos ao longo do capítulo as razões que orientam a escolha de um valor para α, porque elas dependem de compreender o significado de um teste de hipótese.

Associado à distribuição de z e ao nível de significância α, define-se o valor crítico $z_{\rm critico}$, uma grandeza definida positiva, pela relação

$$\alpha = P(z < -z_{\rm critico} \text{ ou } z > z_{\rm critico}) \quad ,$$

de modo que

$$\begin{aligned} \alpha &= P(\mid z \mid > z_{\rm critico}) = \int_{-\infty}^{-z_{\rm critico}} N(z; 0, 1)\, dz + \int_{z_{\rm critico}}^{\infty} N(z; 0, 1)\, dz \\ \alpha &= 2 \int_{z_{\rm critico}}^{\infty} N(z; 0, 1)\, dz \quad . \end{aligned} \tag{5.2}$$

Quando $\alpha = 0,05$, que é um valor típico, da tabela (**??**) obtém-se $z_{\rm critico} = 1,96$.

Concluímos o teste com o cálculo numérico do valor de z,

$$z = (3,0023 - 3,0000)/0,0019 = 1,2$$

que, como é *menor* que $z_{\rm critico}$ e *maior* que $-z_{\rm critico}$, leva a *não rejeitar* a hipótese.

Note que a definição de $z_{\rm critico}$ significa que 5% das vezes em que a hipótese for verdadeira, teremos $\mid z \mid > z_{\rm critico}$ e *rejeitaremos* a hipótese, que é verdadeira. Assim, este procedimento destina-se a quantificar com que frequência

erraremos a decisão apesar da hipótese ser verdadeira, mas não pode quantificar quantas vezes acertaremos[1], porque isso depende da máquina que fabrica rebites: se ela produz rebites com 3,1 mm, provavelmente rejeitaremos corretamente a hipótese e seguiremos rejeitando os rebites enquanto ela estiver produzindo rebites de 3,1 mm.

Note que foi necessário definir o nível de significância como a probabilidade de encontrar z em todo um intervalo de valores. Isso acontece porque z é uma variável contínua, de modo que a probabilidade de encontrarmos um determinado valor de z é estritamente nula. Na prática, então, toma-se um ou mais intervalos, cujos extremos correspondem aos valores críticos.

Outra situação em que podemos aplicar o teste z é na avaliação de parâmetros ajustados pelo Método dos Mínimos Quadrados (MMQ). Por exemplo, o responsável pelo setor de radioproteção de uma instalação verifica que um objeto está contaminado por um material radioativo, que *pode* ser o isótopo 130 do Iodo. Para verificar se se trata mesmo dele, o supervisor acompanha o decaimento radioativo por um dia inteiro e determina, pelo MMQ, que a meia-vida é $\hat{T} = 11,74(22)$ h. Para quantificar o teste, ele formula a *hipótese estatística*: a meia-vida do radioisótopo é igual à do ^{130}I, T_I=12,36 h [2].

A estatística z é calculada com a fórmula (5.1),

$$z = \frac{11,74 - 12,36}{0,22} = -2,8$$

e, se for adotado o nível de significância igual a 5%, $z_{\text{critico}} = 1,96$ e a hipótese será *rejeitada*, ou seja, concluir-se-á que o contaminante não é o ^{130}I, nesse nível de significância.

As condições necessárias para que z tenha distribuição $N(z; 0, 1)$ com parâmetros ajustados pelo MMQ serão discutidas no capítulo 8.

5.2 O teste t

A seção anterior apresentou um teste de hipótese com base em grandezas calculadas a partir de muitos dados e, portanto, sem considerar a flutuação de σ_m. Nesta, vamos lidar com um número qualquer de dados, o que requer o conhecimento do seu comportamento estatístico, ou seja, da forma da sua função

[1] É por isso que nunca se fala em "aceitar a hipótese".

[2] Os dados nucleares encontram-se em tabelas e, no caso do ^{130}I, o desvio-padrão, 0,01 h, é tão menor que a incerteza da medição do supervisor, 0,22 h, que pode ser ignorado.

densidade de probabilidade. Vamos adotar a hipótese de que eles são gaussianos, de modo que os resultados desta seção somente podem ser aplicados a medições com poucos dados quando eles se distribuem conforme a normal.

Retome o exercício 3.4, acerca de uma medida com 5 dados da intensidade de muons por segundo, cuja média e desvio padrão da média são, respectivamente

$$\bar{x} = 9,30 \qquad \text{e} \qquad \sigma_m = 0,54 \quad .$$

Considere que uma teoria bastante completa sobre a radiação cósmica e sobre o detetor usado prevê uma intensidade detetada de muons igual a 10,5 muons por segundo. Será que os dados suportam esta previsão teórica? Ou será que os dados a contradizem? A pergunta pode ser feita para saber se é correto o modelo tanto da radiação quanto do detetor (normalmente, sua *eficiência* de deteção). O teste de hipótese não pode distinguir entre os dois, mas apenas comparar o resultado medido com o esperado, que é uma composição dos dois modelos.

A proposta da Estatística para realizar essa inferência a partir dos dados do exercício 3.4 é formular uma hipótese estatística — *a intensidade* **média** *do fundo no detetor em questão é igual a 10,5 muons por segundo* — e testá-la.

Nas circunstâncias do experimento descrito no exercício 3.4, os 5 dados obtidos podem ser supostos gaussianos, de maneira que sabemos que a variável aleatória

$$\frac{\bar{x} - x_0}{\sigma_m} = t \quad , \tag{5.3}$$

em que x_0 é o valor verdadeiro por hipótese, tem f.d.p. de t de Student, $\psi_\nu(t)$, com $\nu = 4$ graus de liberdade (veja seção 3.5). A partir de $\psi_\nu(t)$, podemos montar um *teste de hipótese*, como proposto.

Definiremos formalmente *teste de hipótese estatística* como uma regra que, aplicada aos dados experimentais, leva ou à decisão de rejeitar a hipótese em consideração ou à decisão de *não* rejeitá-la. Como há possibilidade da flutuação estatística levar à conclusão errada, e a decisão que se segue é extrema — ou analisamos os dados com esse modelo (hipótese não rejeitada) ou partimos para construir outro modelo —, as consequências práticas são importantes e as discutiremos adiante neste capítulo.

A hipótese sob consideração neste exemplo pode ser expressa matematicamente por

$$\text{hipótese} : \; x_0 = 10,50 \quad .$$

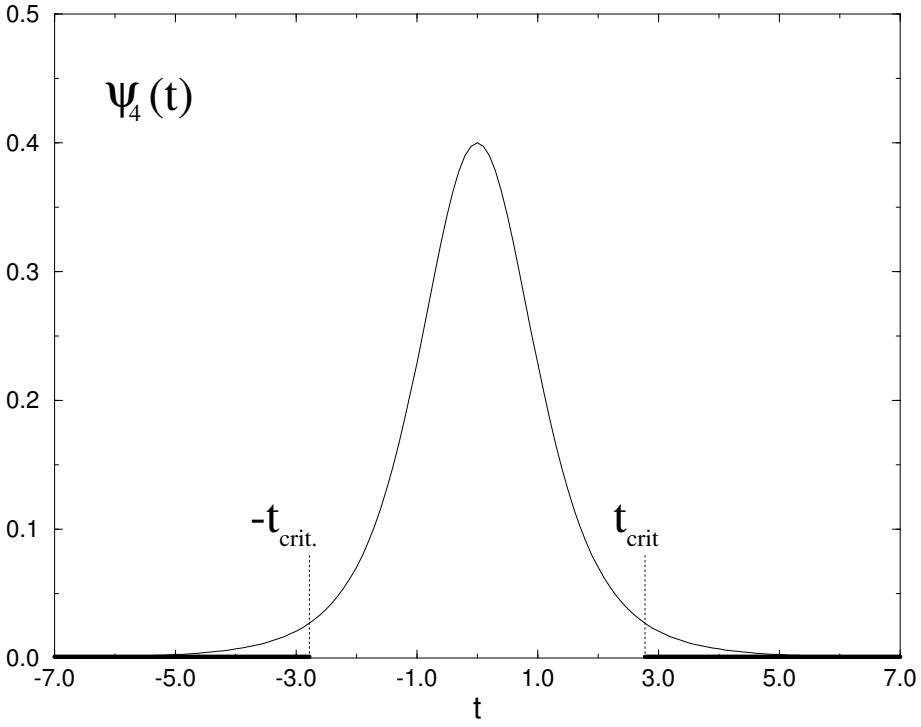


Figura 5.1: Função densidade de probabilidade de t de Student, para $\nu = 4$ graus de liberdade. O valor crítico de t está assinalado (aliás, os *valores críticos*, representados por $t_{critico}$ e seu oposto) e a região crítica está representada como a parte do eixo das abscissas em linha grossa. A probabilidade de rejeição da hipótese, sendo ela verdadeira, é dada pela integral da f.d.p. ao longo da região crítica.

Sabemos que, a cada medida efetuada, um valor diferente de t é obtido pela relação (5.3), em razão da flutuação estatística dos dados. Caso a hipótese seja verdadeira, os valores próximos de zero são os mais prováveis, sendo raro ocorrerem valores com módulos bastante maiores que dois[3]. Para ser preciso, a f.d.p. de t é *exatamente* dada pela expressão (3.25) e está representada na figura 5.1 no caso em que o número de graus de liberdade é $\nu = 4$.

A regra a ser aplicada e que constitui, portanto, o *teste estatístico*, consiste em não rejeitar a hipótese se $\mid t \mid$ é pequeno e rejeitá-la, se $\mid t \mid$ for grande. Porém, a pergunta: quão grande $\mid \ t \ \mid$ precisa ser para rejeitar a hipótese, não tem resposta única. Habitualmente, escolhe-se a *priori* um valor crítico

[3]Quando o número de graus de liberdade é pequeno, podem ocorrer valores significativamente maiores que 2 com boa probabilidade, veja discussão no capítulo 3.

de $\mid t \mid$, t_{critico}, a partir do qual consideramos falsa a hipótese. Este valor t_{critico} define uma região crítica, tal como ilustrada na figura 5.1, que compreende os intervalos $]-\infty, -t_{\text{critico}}]$ e $[t_{\text{critico}}, \infty[$, correspondentes às duas caudas de $\psi_\nu(t)$. O nível de significância, α, é calculado como a integral nessas regiões; quando α é determinado a partir das duas caudas, diz-se que o teste é bi-caudal, e mono-caudal quando se toma a integral ao longo de uma única cauda. O nível de significância

$$\alpha = P(\mid t \mid > t_{\text{critico}}) = \int_{-\infty}^{-t_{\text{critico}}} \psi_\nu(t)dt + \int_{t_{\text{critico}}}^{\infty} \psi_\nu(t)dt = 2\int_{t_{\text{critico}}}^{\infty} \psi_\nu(t)dt, \tag{5.4}$$

onde a última igualdade decorre da simetria de $\psi_\nu(t)$, pode ser interpretado como uma probabilidade, que é identificada como *probabilidade de erro tipo I* e chamada também de *tamanho do teste*. Ao adotar a postura de rejeitar a hipótese sempre que $\mid t \mid$ exceder t_{critico}, incorre-se em *erro* numa fração α das vezes em que o teste for realizado quando a hipótese é verdadeira, porque $\mid t \mid$ supera t_{critico} com probabilidade α. Este tipo de engano—rejeitar uma hipótese quando ela é verdadeira—é chamado de *erro tipo I.* Reflita e perceba que não pode existir um teste que nunca erre! O inconveniente de aumentar t_{critico}, de maneira que α seja praticamente nulo, está relacionado a outro tipo de engano (o *erro tipo II*), que será discutido na seção seguinte. Agora, finalizaremos o exemplo.

A escolha de $t_{\text{critico}} = 2,78$ representada na figura 5.1 corresponde, com 4 graus de liberdade, a um nível de significância

$$\alpha = 0,05 \quad .$$

É comum também escolher valores críticos que dêem α igual a 0,01 ou a 0,001. Os valores de t correspondentes a estes valores mais habituais de α, para diversos valores do número de graus de liberdade, ν, estão apresentados na tabela 5.2. Tabelas para outros valores estão no apêndice (Tabela **??**).

Efetuando finalmente os cálculos do exemplo, levando em conta a hipótese: $x_0 = 10,50$, temos

$$t = \frac{\bar{x} - x_0}{\sigma_m} = \frac{9,30 - 10,50}{0,54} = -2,2 \quad ,$$

o que, por ter módulo menor que 2,78, leva a *não rejeitar* a hipótese. Dizemos que a hipótese $x_0 = 10,50$ *não é rejeitada com nível de significância de* 5%.

Tabela 5.2: Valores de t_α para os quais o nível de significância, $\alpha = 2\int_{t_\alpha}^{\infty} \psi_\nu(t)dt$ onde $\psi_\nu(t)$ é a f.d.p. de t de Student para ν graus de liberdade, tem os valores apresentados na primeira linha da tabela, os mais comuns no teste t. A primeira coluna identifica o número de graus de liberdade ν, que corresponde ao número de dados, N, subtraindo-se 1, devido ao vínculo representado pela média: $\nu = N - 1$ (Eq. (3.24)).

$\alpha =$	0,05	0,01	0,001
ν			
1	12,7	63,7	636
2	4,30	6,97	31,6
3	3,18	5,84	12,9
4	2,78	4,60	8,61
5	2,57	4,03	6,87
10	2,23	3,17	4,59
20	2,09	2,85	3,85
30	2,04	2,75	3,65
∞	1,96	2,58	3,29

Nos cálculos acima, usou-se t com um algarismo significativo além dos que foram usados para representar o resultado final (média e seu desvio padrão), uma vez que a f.d.p. de t já considera *toda* a flutuação estatística das grandezas estimadas e a probabilidade de um certo $t_{\rm critico}$ ser excedido varia bastante com $t_{\rm critico}$. Assim, é necessário um pouco mais de precisão no cálculo de t e, por isso, é prática habitual usar um algarismo significativo a mais, em relação aos do resultado final, para a intensidade média experimental.

Q5.1 *Um experimentador comprou uma pequena folha de ouro cuja pureza é garantida, sendo especificado que mais do que* $99,99\%$ *do material é Au. Para verificar o grau de pureza do material, ele realiza uma medida para determinar a densidade, obtendo os seguintes dados:* $\{19,06; 18,94; 19,08; 19,28; 19,02; 18,90\}$, *todos em* g/cm^3 *e obtidos a* $20\,^{\rm o}$C *de temperatura. Sabendo que a densidade do Au é* $19,32(1)$ g/cm^3 *a* $20\,^{\rm o}$C *e que verificou-se a inexistência de bolhas internas no material por meio de uma chapa de Raio-X, você concordaria que o material adquirido tem o grau de pureza especificado? O teste aplicado é sensível a pequenas misturas de outros materiais? Para fixar idéias, caso o material misturado fosse Ag (*$d_{Ag} = 10,50$ g/cm^3

a 20 °C), a partir de que proporção o teste realizado seria capaz de detetar sua mistura com Au?

Um aspecto difícil do teste de hipótese é a escolha do nível de significância. A tabela (5.2) acima registra os valores mais comuns. Quando se realiza um único teste, é costume adotar $\alpha = 5\%$; valores ainda maiores são usados quando as consequências da decisão são graves, por exemplo na verificação da pureza biológica de sangue para transfusão, quando valores como 20% são comuns. Já quando se repete o teste muitas vezes e o que está em jogo tem consequências mais amenas, como repetir a produção de uma peça de baixo custo, $\alpha = 1\%$ ou mesmo $\alpha = 0,1\%$ são frequentes. Também é comum usar α pequeno quando se testa uma teoria ou hipótese muito sólida. Note, porém, que os valores t_{critico} associados a níveis de significância tão baixos dependem da hipótese sobre a distribuição estatística dos dados; em particular, o uso das tabelas (5.2) e (**??**) requer que a f.d.p. dos dados seja Normal. Mesmo pequenas diferenças no comportamento estatístico podem resultar em mudanças de ordem de grandeza, como transformar um $\alpha = 0,1\%$ em 1% ou 0,01%. Isso é diferente quando se adota $\alpha = 5\%$ e calcula-se t_{critico} com base na gaussiana, uma vez que, se a distribuição dos dados for apenas parecida com a normal, mas diferente, o valor correto de α será diferente, mas, com boa possibilidade, um valor tal como 6%, 7%, 3%, todos com significados qualitativamente semelhantes.

5.3 Erro tipo I e erro tipo II

Considere a seguinte situação. Um laboratório pode sofrer, acidentalmente, contaminação pelo ^{130}I, de meia vida 12,36 h, ou ser afetado por outro radioisótopo, de meia vida 13,02 h, que não causa nenhum problema. Caso seja contaminado pelo isótopo do Iodo e nenhuma providência for tomada, o laboratório terá seu funcionamento prejudicado e todos os resultados obtidos nos próximos 3 dias deverão ser descartados, o que só se saberá a posteriori. Assim, uma vez caracterizada a presença do ^{130}I, o laboratório precisa ser descontaminado, o que demora 1 dia. Caso tenha aparecido o outro radioisótopo, não há problema algum e o laboratório pode continuar operando normalmente, sem necessidade de descontaminação.

Há um detector de radiação que permite saber se houve ou não contaminação, mas para saber se é Iodo ou outro elemento, é necessário medir a meia vida associada à radiação. O protocolo de procedimento do laboratório,

sempre que se deteta uma contaminação, é o seguinte: faz-se uma medida da meia vida e, a partir do resultado, decide-se por continuar operando normalmente ou fazer a descontaminação.

A medida da meia vida, seja qual for a contaminação, leva a um resultado com precisão de 0,22 h. É a partir do resultado obtido que se decide por continuar operando normalmente ou fazer a descontaminação. Como a diferença entre as duas meias vidas, de 0,66 h, não é muito maior do que a precisão experimental, de 0,22 h, é sempre possível haver contaminação por um elemento e o resultado da medida estar mais próximo da meia vida do outro elemento. Assim, ainda depois da medida da meia vida, a decisão se houve ou não contaminação pode ser errada. A questão, portanto, é definir um valor crítico para a meia vida medida, T_{cr}: se o valor experimental for superior a esse valor crítico, admite-se que a contaminação não foi pelo Iodo e opta-se por não descontaminar o laboratório e, caso o resultado seja inferior àquele valor crítico, é feita a descontaminação. Fica claro que qualquer uma das duas decisões pode estar errada.

A situação descrita acima está relacionada ao tipo de erro que se pode cometer ao tomar uma decisão. Vamos chamar de H_0 a hipótese

$$H_0 \equiv \text{ hipótese } : T_0 = 12,36 \text{ h } \text{ (Iodo)}.$$

e de H_1 a hipótese alternativa, de contaminação pelo outro radioisótopo,

$$H_1 \equiv \text{ hipótese } : T_0 = 13,02 \text{ h } \text{ (outro radioisótopo)}.$$

Os dois erros possíveis são rejeitar a hipótese H_0 quando ela é verdadeira, usualmente chamada de *erro do tipo I*, e não rejeitá-la quando é falsa, *erro tipo II*. A hipótese de referência, H_0 neste caso, é chamada de *hipótese nula* em Estatística.

A figura 5.2 apresenta as duas f.d.p.s possíveis dos resultados experimentais, uma delas centrada na meia vida do ^{130}I, 12,36 h, e a outra na meia vida do outro possível contaminante, 13,02 h, ambas com $\sigma = 0,22$ h. As probabilidades dos dois tipos de erros, α e β, quando escolhemos um valor crítico para a meia vida, T_{cr}, também são indicadas. Nesse exemplo, α é a probabilidade de erro tipo I,

$$P(T_{\text{experimental}} > T_{cr} \text{ quando } H_0 \text{ é verdadeira}) = \alpha$$

e β é a probabilidade de erro tipo II,

$$P(T_{\text{experimental}} < T_{cr} \text{ quando } H_1 \text{ é verdadeira }) = \beta \quad .$$

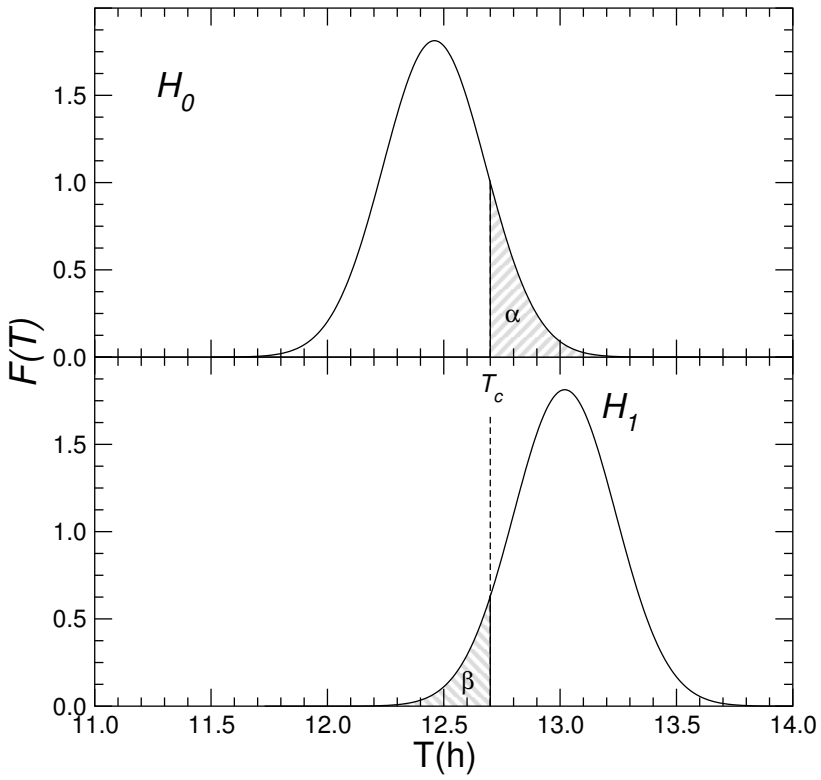


Figura 5.2: Curvas mostrando as duas hipóteses em teste: H_0 vs H_1. T_c é o valor crítico, α a probabilidade de cometer um erro tipo I e β, um erro tipo II.

Fazendo referência à figura 5.2, aumentar o valor de T_{cr} em relação àquele indicado na figura diminui a chance de erro tipo I, α, que é a probabilidade de decidir que não houve contaminação por Iodo quando houve sim; por outro lado, aumentar T_{cr} aumenta β, a chance de não rejeitar a hipótese de que houve contaminação por Iodo quando o outro radioisótopo contaminar o laboratório.

A tarefa delicada, portanto, é a escolha de T_{cr}. *Caso se saiba*, por experiência anterior, que a contaminação pode ocorrer tanto pelo Iodo como pelo outro radioisótopo com igual chance, sem nenhuma outra informação além daquelas já apresentadas nesta seção, um critério possível é escolher T_{cr} de modo que o tempo perdido do laboratório seja mínimo. Esse tempo perdido é proporcional à média entre os 3 dias perdidos caso haja contaminação e mantenha-se

Tabela 5.3: A última coluna dá o tempo perdido no laboratório por uma decisão errada baseada no valor crítico da primeira coluna, calculado a partir das probabilidades de erros tipo I e II nas colunas 2 e 3, respectivamente, que foram deduzidos das f.d.p.s correspondentes às hipóteses H_0 e H_1.

T_{cr} (h)	α	β	$\langle T \rangle$ (d)
12,360	0,500	0,001	1,501
12,426	0,382	0,003	1,150
12,492	0,274	0,008	0,831
12,558	0,184	0,018	0,570
12,624	0,115	0,036	0,381
12,690	0,067	0,067	0,267
12,756	0,036	0,115	0,223
12,822	0,018	0,184	0,238
12,888	0,008	0,274	0,299
12,954	0,003	0,382	0,392
13,020	0,001	0,500	0,504

o laboratório operando com peso α, e o dia perdido, com peso β, caso se opte por descontaminar o laboratório quando isso é desnecessário. Essa escolha é ilustrada pela tabela 5.3, da qual se deduz que o valor crítico de 12,76 h é o melhor de acordo com esse critério, que resulta em que não se descontamina o laboratório quando a meia vida medida estiver acima deste valor e, se estiver abaixo, procede-se à descontaminação. Note que o procedimento de medir a meia-vida e decidir com base no resultado é muito mais econômico em tempo do que simplesmente descontaminar cada vez que se suspeita a contaminação, que redundaria na perda de um dia a cada repetição desse incidente.

Em ciências experimentais encontramos, frequentemente, situações como a desse exemplo, nas quais precisamos tomar uma decisão a partir de resultados experimentais sujeitos a erros estatísticos. Outros exemplos são o da procura de um sinal relativamente fraco na presença de um ruído de fundo, quando se deseja inferir se o sinal foi observado ou o resultado é apenas fruto da flutuação estatística do ruído, e o da busca de uma nova partícula, em que se pretende decidir se os sinais observados revelam sua existência ou são manifestações de outro efeito.

Em todas essas situações, a tomada de decisão depende tanto do conheci-

mento das f.d.p.s envolvidas, as quais permitem definir as probabilidades de erros tipo I e II, bem como de um fator que nos permite quantificar as consequências dos erros cometidos. No exemplo desenvolvido nesta seção, a quantificação das consequências dos erros foi feita com base no tempo perdido do laboratório. Em algumas situações, a maneira de quantificar as consequências pode ser o custo financeiro das diferentes decisões. Em um problema de saúde, podem-se adotar, para quantificar as consequências dos diferentes tratamentos, as f.d.p.s do número de dias que as pessoas demorarão para se recuperar de uma doença nos casos de um ou outro tratamento.

Como um exemplo de outra natureza, suponha que um pesquisador mediu a diferença da velocidade da luz entre as direções norte-sul e leste-oeste e obteve $5,9 \pm 1,9$ m/s, que obedece a uma f.d.p. gaussiana, sem erros sistemáticos no experimento nem mau funcionamento do equipamento. Como a velocidade da luz não depende da direção, essa diferença deveria ser compatível com zero e isso é um fato que se supõe verdadeiro há muito tempo e com inúmeras comprovações experimentais diretas e indiretas. Entretanto, o resultado acima está três desvios padrões afastado de zero, o que corresponde a uma probabilidade de $0,27\%$ em um teste estatístico bicaudal. O pesquisador precisa tomar uma decisão quanto a ter obtido uma evidência de que a velocidade da luz depende da direção de propagação ou não. A quantificação objetiva da decisão envolve aspectos difíceis de avaliar. Se ele decidir que realmente obteve um resultado indicando que a velocidade da luz varia segundo a direção de propagação e divulgar amplamente seu resultado, poderá cair no ridículo caso seu resultado seja fruto apenas de um acaso. Se o resultado for correto e o pesquisador não divulgá-lo, estará privando a humanidade de um importantíssimo resultado. Apesar da dificuldade de avaliação, o processo de decisão requer dar pesos para as atuações possíveis.

Embora este último exemplo seja artificioso, uma vez que, pelo menos dentro da precisão dos resultados apresentados, sabe-se que a velocidade da luz não varia e pareceria mais razoável apostar em um acaso devido à flutuação estatística, problemas com essa característica de envolver questões subjetivas, difiíceis de quantificar, não são raros. Enfim, muitas vezes as escolhas envolvidas em testes de hipóteses são bastante subjetivas e, em particular, a escolha de um nível de significância pequeno corresponde a tomar partido a respeito da hipótese que se testa. Como, porém, não há outra maneira de proceder, tudo o que se pode fazer é enunciar claramente a hipótese sendo testada e o nível de significância do teste.

Observando a figura 5.2, é possível perceber que $1 - \beta$ é a probabilidade de

rejeitar a hipótese nula, H_0, quando ela é, de fato, falsa. Quanto maior essa probabilidade, melhor o teste. Assim, caso esse mesmo procedimento fosse feito com uma medida mais precisa, ou seja, com um desvio padrão inferior a 0,22 h, a discriminação entre as duas hipóteses seria mais fácil e, para um mesmo valor de α, o correspondente valor de $1-\beta$ seria maior. A essa grandeza $1-\beta$ se dá o nome *poder estatístico do teste*.

O poder do teste poderia variar, como dito acima, fazendo experimentos com desvios padrões diferentes. Entretanto, o poder de um teste pode depender, também, da própria escolha do teste. No exemplo desenvolvido nesta seção, não havia outra possibilidade de estruturar a decisão além daquela fornecida pela f.d.p. dos resultados nos casos de contaminação pelo Iodo ou manifestação do outro material, mas há casos em que existem várias possibilidades de escolha da f.d.p. em que se pode basear o teste, quando devemos optar por aquele de maior poder.

Q5.2 *É a gravidade da consequência de um erro tipo I comparada à gravidade de um erro tipo II que define se o nível de significância do teste deve ser grande ou pequeno. Nas situações abaixo, identifique quais delas correspondem a situações onde deve-se procurar reduzir o nível de significância (ou seja, evitar erros tipo I) e quais as situações onde deve-se adotar um nível de significância grande (ou seja, evitar erros tipo II). Em cada caso, deixe claro quais as hipóteses (H_0 e H_1) testadas; pares de hipóteses diferentes levam a respostas diferentes.*

(a). Testar a resistência da estrutura de um prédio de 25 andares.

(b). Testar a contaminação da vacina contra a poliomielite.

(c). Testar a possibilidade de falha do rádio de um automóvel.

(d). Testar que o sistema de freios de um automóvel funciona.

5.4 O teste t na comparação de duas médias

O teste t, aplicado na comparação de um resultado experimental com um possível valor da grandeza medida na seção 5.2, será usado aqui na comparação de duas medidas.

Em uma medida da capacitância de um capacitor, com um número n de dados supostamente normais, o experimentador determinou o valor médio, $\hat{x}$, e o desvio padrão do conjunto de dados, σ (o desvio padrão da média

é, portanto, $\sigma/\sqrt{n}$). Um segundo experimentador mediu a capacitância desse mesmo capacitor, mas tomou um número m de dados supostamente gaussianos, cuja média e desvio padrão foram $\hat{x}'$ e σ', respectivamente.

Como os dois experimentadores mediram o mesmo capacitor, os resultados obtidos devem ser os mesmos, embora os números $\hat{x}$ e $\hat{x}'$ muito provavelmente sejam diferentes devido à flutuação estatística. Espera-se, porém, que eles não difiram em mais do que uma ou duas "barras de incerteza", ou ainda que eles sejam compatíveis "dentro da flutuação estatística"ou "dentro da incerteza experimental". O objetivo do teste de hipótese é transformar essas ideias qualitativas em uma noção objetiva que permita uma avaliação quantitativa. Podemos entender a motivação dessa questão, no caso específico, como o interesse dos dois experimentadores em verificarem a existência de algum erro sistemático, o que resultaria em medidas diferentes.

A fim de formalizar o problema, chamaremos de μ e μ' os valores verdadeiros nas medidas efetuadas pelo primeiro e segundo experimentador, respectivamente. Além disso, as precisões dos procedimentos adotados pelos dois experimentadores serão supostas iguais, de maneira que se adota

$$\sigma_0 = \sigma_0', \text{ e a f.d.p. dos dados é a normal.}$$

Essas duas condições são essenciais para a aplicação do teste t da maneira como segue; o caso em que $\sigma_0 \neq \sigma_0'$ está na próxima seção. Assim, testa-se a

$$\text{Hipótese } : \mu = \mu' \quad ,$$

onde μ e σ_0 *não são especificados.*

Podemos transformar esta hipótese em outra, equivalente, que inclui o fato de μ não ser determinado,

$$\text{Hipótese } : \mu - \mu' = 0 \quad .$$

A estimativa de $\mu - \mu'$ é $\hat{x} - \hat{x}'$, com variância *verdadeira* igual a

$$\text{var}(\hat{x} - \hat{x}') = \text{var}(\hat{x}) + \text{var}(\hat{x}') = \frac{\sigma_0^2}{n} + \frac{\sigma_0^2}{m} = \sigma_0^2 \left(\frac{1}{n} + \frac{1}{m}\right) \quad . \tag{5.5}$$

Para estimar a variância verdadeira, σ_0^2, aplicamos a hipótese das medidas terem o mesmo desvio padrão. Chamando de θ^2 a estimativa global da variância, tem-se

$$\theta^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{x})^2 + \sum_{j=1}^{m}(x_j' - \hat{x}')^2}{n - 1 + m - 1}$$

$$\theta^2 = \frac{(n-1)\sigma^2 + (m-1)\sigma'^2}{n-1+m-1} \quad . \tag{5.6}$$

A variável aleatória t fica, então,

$$t = \frac{(\hat{x} - \hat{x}') - 0}{\theta\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\hat{x} - \hat{x}'}{\theta}\sqrt{\frac{nm}{n+m}}, \tag{5.7}$$

com f.d.p. de t de Student com $n + m - 2$ graus de liberdade.

Assim, o procedimento para comparar as duas medidas consiste em calcular θ pela fórmula (5.6) e, lembrando que o número de graus de liberdade a considerar é $n + m - 2$, comparar t da fórmula (5.7) com o valor crítico escolhido em um teste bi-caudal, conforme discussão das seções anteriores. A tabela (5.2), na seção 5.2, lista os valores críticos para alguns níveis de significância e números de graus de liberdade; uma tabela mais completa está no apêndice (**??**). O exercício 5.3 apresenta um exemplo realista deste procedimento, mas sua solução requer o teste prévio da hipótese $\sigma_0 = \sigma'_0$ com o teste F, que será discutido na seção 5.7 adiante.

Na situação em que as variâncias *são diferentes*, mas conhecidas, ajusta-se a média das medidas pelo método dos mínimos quadrados e realiza-se o teste com base na distribuição da variável χ^2, como será discutido neste capítulo a partir da seção 5.9. A próxima seção lida com a situação em que as variâncias são distintas e suas estimativas têm pouca precisão, em que a estimativa do χ^2 costuma ser uma base muito imprecisa para o teste.

5.5 O teste t na comparação de duas médias com variâncias diferentes

Para avaliar se duas médias podem ou não corresponder a medidas da mesma grandeza, quando elas são obtidas de dois conjuntos de dados que obedecem a funções densidade de probabilidade gaussianas de *mesma* variância, usa-se o teste t, conforme procedimento da seção anterior. A hipótese da igualdade das variâncias é usada na equação (5.5), que é uma estimativa da variância que combina as duas estimativas independentes. Um exemplo de situação na qual essa hipótese é aceitável é a medida das densidades de duas amostras pelo mesmo método, com a mesma balança e os mesmos instrumentos para medição dos volumes, quando é possível testar se as densidades são iguais usando o teste t conforme descrito na seção anterior.

Muitas vezes, porém, a hipótese de igualdade das variâncias pode não ser razoável, ou seja, os desvios padrões dos dois conjuntos de dados são diferentes. No exemplo da medida de densidade do parágrafo anterior, isso pode ocorrer ao usar balanças, instrumentos e/ou métodos diferentes nas duas séries de medições. Esse problema tem uma solução aproximada [Magalhães], que é o assunto desta seção.

Suponha que $\bar{x}_1$, $\sigma_{\bar{x}_1}$, $\bar{x}_2$ e $\sigma_{\bar{x}_2}$ sejam resultados (médias e correspondentes estimativas dos desvios padrões) de duas medidas, com n_1 e n_2 dados, respectivamente. A estatística

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \tag{5.8}$$

com

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} \tag{5.9}$$

obedece, aproximadamente, a uma distribuição t com um número de graus de liberdade dado por [Magalhães]

$$\nu = \frac{\left(\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2\right)^2}{\frac{\sigma_{\bar{x}_1}^4}{n_1 - 1} + \frac{\sigma_{\bar{x}_2}^4}{n_2 - 1}} \quad . \tag{5.10}$$

Assim, usa-se a estatística t da equação (5.8), mas com um número de graus de liberdade dado pela equação (5.10), que não é necessariamente inteiro.

Como exemplo, considere que a energia de uma radiação gama foi medida com um certo detetor, obtendo-se o seguinte conjunto de resultados (em keV): $\{509,1; 512,1; 511,1; 511,6\}$. Em outro experimento e com outro detetor, o conjunto de resultados foi $\{512,0; 514,1; 511,6\}$, todos também em keV. O objetivo do teste é avaliar se as duas energias podem ser iguais.

Os resultados dos experimentos (médias e desvios padrões das médias) são $510,98 \pm 0,66$ e $512,57 \pm 0,78$ keV, respectivamente para o primeiro e segundo experimento. A fórmula (5.8) dá $t = -1,56$. O número de graus de liberdade da fórmula (5.10) é $\nu = 4,4$. As linguagens de programação de alto nível têm essa distribuição e nesse caso você pode calcular a integral diretamente, mas as tabelas de t, como as do Apêndice, só têm entradas para números de graus de liberdade inteiros, quando é necessário interpolar: a probabilidade que t seja menor do que $-1,56$ é $9,7\%$ com 4 graus de liberdade e $9,0\%$ com 5 graus de liberdade, de modo que, para $\nu = 4,4$, o valor interpolado é $P = 9,4\%$. Como

pretende-se testar a igualdade entre as duas grandezas medidas e não se uma delas é maior do que a outra, deve-se usar um teste de duas caudas. Assim, a probabilidade de $\mid t \mid$ ser maior do que 1,56 com 4,4 graus de liberdade é 18,8%. Portanto, com nível de significância de 5%, não é possível descartar a hipótese que as energias das duas transições medidas sejam iguais, pois essa probabilidade, cerca de 19%, é a chance que t seja, em valor absoluto, tão grande quanto se observou ou ainda maior apenas por flutuação estatística quando as duas grandezas são iguais.

5.6 Distribuição da razão de variâncias e F de Fisher

Esta seção apresenta a dedução da f.d.p. da razão de duas estimativas da variância de dados com distribuições normais, obtidas em medidas independentes, que é conhecida como F de Fisher.

Considere uma medida de uma grandeza x de valor verdadeiro x_0, com n dados distribuídos de acordo com uma gaussiana de variância desconhecida, σ^2. Dos dados dessa medida, estima-se a variância da maneira habitual,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \hat{x})^2, \text{ onde } \hat{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad . \tag{5.11}$$

Efetua-se uma segunda medida da mesma grandeza, com n' dados e variância também desconhecida, σ'^2, que pode ser diferente de σ^2; estima-se a variância da maneira habitual, que denominaremos por $\hat{\sigma}'^2$. Note que não estamos usando o sub-índice zero para identificar as variâncias verdadeiras e denotamos as estimativas por um acento circunflexo sobre os símbolos, o que objetiva facilitar a notação no que segue.[4]

A f.d.p. da razão $\hat{\sigma}^2/\hat{\sigma}'^2$ será deduzida a partir das f.d.p.s de $\hat{\sigma}^2$ e de $\hat{\sigma}'^2$ por meio de uma transformação muito parecida com aquela da dedução da f.d.p. de t de Student (seção 3.5), inclusive pelas definições

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{x})^2 \qquad \text{e} \qquad s'^2 = \frac{1}{n'}\sum_{j=1}^{n'}(x'_j - \hat{x}')^2 \quad , \tag{5.12}$$

[4]Consistência notacional completa é difícil de obter e às vezes não ajuda na clareza; os sub-índices $_0$ obscureceriam demais as fórmulas desta seção.

estatísticas muito parecidas com as estimativas das variâncias e mais fáceis de trabalhar. De acordo com a discussão da seção 2.9 e a fórmula (2.45), as f.d.p.s de s^2 e s'^2 são

$$f(s^2) = C \exp\left(-\frac{ns^2}{2\sigma^2}\right) s^{n-3} \quad , \tag{5.13}$$

com

$$C = \left(\frac{1}{2\sigma^2}\right)^{\frac{n-3}{2}} \frac{1}{2 \cdot \Gamma\left(\frac{n-1}{2}\right)} \tag{5.14}$$

e

$$g(s'^2) = C' \exp\left(-\frac{n's'^2}{2\sigma'^2}\right) s'^{\,n'-3} \quad , \tag{5.15}$$

com

$$C' = \left(\frac{1}{2\sigma'^2}\right)^{\frac{n'-3}{2}} \frac{1}{2\Gamma\left(\frac{n'-1}{2}\right)} \quad , \tag{5.16}$$

onde C e C' são constantes porque σ^2 e σ'^2 são os valores verdadeiros, embora desconhecidos, das variâncias. Com a transformação de variáveis

$$\left.\begin{array}{c} s^2 \\ s'^2 \end{array}\right\} \longrightarrow \left\{\begin{array}{l} \phi = \frac{s^2}{s'^2} \\ \omega = s'^2 \end{array}\right. \tag{5.17}$$

obtém-se a f.d.p. de ϕ, ao transformar a f.d.p. conjunta de s^2 e s'^2 na f.d.p. conjunta de ϕ e ω e integrar o resultado em todo o domínio de ω.

A f.d.p. conjunta de s^2 e s'^2 é

$$h(s^2, s'^2) = CC' \exp\left(-\frac{ns^2}{2\sigma^2}\right) s^{n-3} \exp\left(-\frac{n's'^2}{2\sigma'^2}\right) s'^{\,n'-3} \quad .$$

O Jacobiano da transformação é

$$\frac{\partial(\phi, \omega)}{\partial(s^2, s'^2)} = \frac{1}{s'^2} \quad .$$

Calcula-se então a f.d.p. $\eta(\phi, \omega)$ como

$$\eta(\phi, \omega) = CC' \exp\left(-\frac{ns^2}{2\sigma^2}\right) s^{n-3} \exp\left(-\frac{n's'^2}{2\sigma'^2}\right) s'^{n'-1},$$

onde é preciso ainda trocar $s^2 = \phi \cdot \omega$ e $s'^2 = \omega$. Finalmente,

$$\eta(\phi,\omega) = CC' \exp\left(-\frac{n\phi\omega}{2\sigma^2}\right) (\,\phi\cdot\omega\,)^{\frac{n-3}{2}} \exp\left(-\frac{n'\omega}{2\sigma'^2}\right) (\,\omega\,)^{\frac{n'-1}{2}}$$

$$\eta(\phi,\omega) = CC' \exp\left\{-\frac{\omega}{2}\left(\frac{n\phi}{\sigma^2}+\frac{n'}{\sigma'^2}\right)\right\} (\,\omega\,)^{\frac{n+n'-4}{2}} (\,\phi\,)^{\frac{n-3}{2}} .$$

A partir da expressão acima, determina-se a f.d.p. de ϕ integrando para todo ω,

$$g(\phi) = CC' (\,\phi\,)^{\frac{n-3}{2}} \int_0^\infty \exp\left\{-\frac{\omega}{2}\left(\frac{n\phi}{\sigma^2}+\frac{n'}{\sigma'^2}\right)\right\} (\,\omega\,)^{\frac{n+n'-4}{2}} d\omega \quad .$$

A integral fica mais compacta ao substituir o fator que multiplica ω na exponencial por a, ou seja,

$$a = \frac{1}{2}\left(\frac{n\phi}{\sigma^2}+\frac{n'}{\sigma'^2}\right) \quad ,$$

que independe tanto de ω quanto de m, o expoente de ω

$$m = (n+n'-4)/2 \qquad ,$$

o que permite a transformação de variáveis

$$\Omega = a\omega$$

e obter

$$g(\phi) = CC' (\,\phi\,)^{\frac{n-3}{2}} a^{-m-1} \int_0^\infty \exp(-\Omega)\Omega^m d\Omega \quad .$$

A integral da expressão acima é, simplesmente, $\Gamma(m+1)$. Então,

$$g(\phi) = CC' \Gamma(m+1) \frac{(\phi)^{\frac{n-3}{2}}}{\left(\frac{n\phi}{\sigma^2}+\frac{n'}{\sigma'^2}\right)^{\frac{n+n'-2}{2}}} \quad .$$

Essa é uma função de s^2 e s'^2, fórmula (5.12), porque $\phi = \frac{s^2}{s'^2}$, de modo que é preciso reescrever o argumento ϕ em função das estimativas não tendenciosas das variâncias. O procedimento clássico é definir

$$\nu = n-1 \qquad \text{e} \qquad \nu' = n'-1 \quad , \tag{5.18}$$

o que permite escrever $\hat{\sigma}^2$ e $\hat{\sigma}'^2$ em função de s^2 e s'^2 como

$$\hat{\sigma}^2 = \frac{\nu+1}{\nu} s^2 \qquad \text{e} \qquad \hat{\sigma}'^2 = \frac{\nu'+1}{\nu'} s'^2 \quad .$$

e definir a variável F de Fisher como

$$F = \frac{\hat{\sigma}^2}{\sigma^2} \frac{\sigma'^2}{\hat{\sigma}'^2} \quad , \tag{5.19}$$

que se relaciona com ϕ por

$$F = \frac{\hat{\sigma}^2}{\sigma^2} \frac{\sigma'^2}{\hat{\sigma}'^2} = \frac{(\nu+1)\nu'}{(\nu'+1)\nu} \frac{s^2}{s'^2} \frac{\sigma'^2}{\sigma^2} = \frac{n\nu'}{n'\nu} \frac{\sigma'^2}{\sigma^2} \phi \quad .$$

Transforma-se $g(\phi)$ em $f(F)$ pela regra usual,

$$f(F) = \frac{g(\phi)}{\left| \frac{dF}{d\phi} \right|}$$

que, após alguma álgebra para reunir todas as constantes, permite obter, finalmente

$$f(F) = \frac{\nu^{\nu/2} \nu'^{\nu'/2} \Gamma\left(\frac{\nu+\nu'}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \Gamma\left(\frac{\nu'}{2}\right)} \frac{F^{(\nu-2)/2}}{(\nu' + \nu F)^{(\nu+\nu')/2}} \quad . \tag{5.20}$$

O valor médio de F e a variância de F são, respectivamente,

$$< F >= \frac{\nu'}{\nu'-2} \qquad \text{para } \nu' > 2 \qquad \text{e} \tag{5.21}$$

$$\text{var}(F) = \frac{2\nu'^2(\nu+\nu'-2)}{\nu(\nu'-2)^2(\nu'-4)} \qquad \text{para } \nu' > 4 \quad . \tag{5.22}$$

As últimas expressões são calculadas usando a definição da função Beta (veja, por exemplo, [Arfken, cap. 10.4]):

$$B(m+1, n+1) = \int_0^\infty \frac{u^m}{(1+u)^{n+m+2}} \, du \quad , \tag{5.23}$$

$$\text{com} \qquad B(p,q) = \frac{\Gamma(p) \cdot \Gamma(q)}{\Gamma(p+q)} \qquad . \tag{5.24}$$

Exemplo 5.1

Uma determinada grandeza adimensional foi medida com um certo equipamento e obteve-se o conjunto de dados $\{109, 118, 99\}$. A variância desses dados, estimada pela fórmula usual, é $\hat{\sigma}^2 = 90$. Algum tempo depois, a mesma grandeza foi medida com o mesmo equipamento, com o resultado $\{158, 102, 131\}$, cuja variância estimada é $\hat{\sigma}'^2 = 784$. Tendo em vista o grande aumento de variância, pode-se suspeitar que o equipamento se deteriorou.

A fim de aplicar o teste F, supõe-se que não tenha havido deterioração do equipamento, de modo que o valor verdadeiro da variância nas duas medidas seria o mesmo e, portanto, o valor de F observado, equação (5.19), é $8,7$. A tabela ?? mostra, entretanto, que a probabilidade de observar $F > 9,0$ apenas por flutuação estatística quando $\nu = \nu' = 2$ graus de liberdade é 10%. Portanto, apesar do grande aumento da dispersão dos dados entre as duas medidas, não se deve descartar a hipótese do equipamento não ter deteriorado.

5.7 O teste F na comparação de duas estimativas da variância

Vamos ilustrar a aplicação prática do teste F com uma situação real ocorrida por ocasião da compra simultânea de um conjunto de multímetros por uma oficina eletrônica e outro, por um laboratório didático, ambos fabricados pela mesma indústria.

O fabricante dos aparelhos comprados especifica que a precisão dos seus multímetros digitais na medição de resistências na faixa de 32 kΩ é igual a 1,0% do valor medido.

A oficina eletrônica comprou 5 desses multímetros e seus técnicos decidiram verificar se a precisão dos aparelhos estava de acordo com a precisão especificada. Para isso, dispunha de um resistor padrão de $25,000(5)$ kΩ, conhecido, portanto, com precisão muito superior à dos multímetros adquiridos, que serviu para o teste proposto. Os dados obtidos, com os 5 multímetros, foram

$$\{25,06; 25,04; 24,97; 25,26; 25,18\}, \text{ em } k\Omega$$

(um dado para cada um dos multímetros). Desses dados, deduz-se

$$\bar{r} = 25,10 \qquad \text{e} \qquad \hat{\sigma} = 0,116 \quad , \tag{5.25}$$

na notação da seção anterior. A primeira razão de contentamento dos técnicos do laboratório foi verificar que o intervalo $[\bar{r}-t_{II}\sigma_m; \bar{r}+t_{II}\sigma_m] = [24,95; 25,25]$, calculado com t_{II} da tabela 3.3 para 4 graus de liberdade, contém o valor da resistência padrão, e concordamos com seu contentamento. Já a segunda razão de alegria com a compra foi verificar que seu lote de multímetros apresentou uma precisão de 0,4%, melhor que a especificada pelo fabricante, *com o que não concordamos*, em razão do resultado do teste que faremos abaixo. O procedimento correto para verificar se a precisão é melhor que a especificada consiste em formular a hipótese

$$\mathrm{H} : \sigma = 0,25$$

e testá-la. Para isso, escolhe-se um nível de significância e, usando a f.d.p. de $\hat{\sigma}$, o respectivo valor crítico $\sigma_{critico}$, que permitirá decidir pela rejeição ou não de H. A f.d.p. da variância é a mesma de χ^2_{N-1}, e dedicaremos as demais seções a estudar testes baseados na distribuição dessa variável aleatória, de maneira que completaremos esta parte do exemplo depois, mas adiantamos que a hipótese H não pode ser rejeitada com um nível de significância grande.

A polêmica ocorrida e que pretendemos resolver nesta seção precisa ainda ser descrita. O laboratório didático adquiriu 10 desses aparelhos e, a fim de verificar se também seu lote tinha a *mesma* precisão que o lote adquirido pela oficina eletrônica, mediu o mesmo resistor padrão e obteve o seguinte conjunto de dados,

$$\{25,22; 24,81; 24,56; 24,98; 24,73; 24,93; 24,79; 25,39; 25,34; 25,07\}(\mathrm{k}\Omega).$$

De novo, a partir de $\bar{r}' = 24,98$ e $\hat{\sigma}' = 0,273$, determinados a partir dos 10 dados obtidos, nota-se que o valor do resistor padrão está contido no intervalo em torno da média com largura igual a um desvio padrão da média, $\sigma_m = 0,086$. A precisão obtida também é compatível com a especificada pelo fabricante. Porém, comparando seus resultados com os da oficina eletrônica, os técnicos do laboratório didático ficaram descontentes com seu lote de aparelhos, afirmando que *tinham pior precisão que os da oficina.*

A fim de avaliar objetivamente a possibilidade de uma diferença de qualidade entre os dois conjuntos de aparelhos, usamos os métodos da inferência estatística e estabelecemos a hipótese:

Tabela 5.4: Valores de F_α para os quais $\int_{F_\alpha}^{\infty} f(F)dF = 0,05$, para alguns valores de ν e ν', números de graus de liberdade para o numerador e o denominador, respectivamente, da razão F de duas variâncias estimadas a partir de dados gaussianos, com numerador maior que o denominador.

$\nu =$	1	2	3	4	9	19
ν'						
1	161	200	225	230	240	248
2	18,5	19,0	19,2	19,3	19,4	19,4
3	10,1	9,55	9,28	9,12	8,8	8,67
4	7,71	6,94	6,59	6,39	6,00	5,81
9	5,12	4,26	3,86	3,63	3,18	2,95
19	4,38	3,52	3,13	2,90	2,42	2,17

$\mathcal{H} : \sigma = \sigma'$, com σ *desconhecido*

Como a disputa envolve apenas a diferença de precisão obtida, independente da especificação do fabricante (o que foi testado por cada grupo), incluímos na hipótese que σ é desconhecido.

A variável aleatória

$$F = \frac{\hat{\sigma}'^2}{\sigma'^2}\frac{\sigma^2}{\hat{\sigma}^2} = \frac{\hat{\sigma}'^2}{\hat{\sigma}^2} \quad , \tag{5.26}$$

é distribuída como F de Fisher, em que o número de graus de liberdade do numerador é $\nu' = 9$ e do denominador, $\nu = 4$. Conhecendo a f.d.p. de F, fórmula (5.20), baseamos o teste nela, com o nível de significância dado por α,

$$\alpha = \int_{F_{critico}}^{\infty} f(F)dF \qquad .$$

A tabela 5.4 apresenta os valores críticos de F, para alguns valores de n e m (ν e ν'), quando o tamanho do teste é 5%. A tabela foi construída de maneira a colocar a maior estimativa no numerador. Para outros níveis de significância, as tabelas do apêndice (**??** e seguintes) podem ser usadas. Da tabela 5.4, obtemos $F_{critico} = 6,00$ e com os resultados experimentais na fórmula (5.26), obtemos

$$F = 5,54 \qquad ,$$

o que significa que *não podemos rejeitar* $\mathcal{H}$ com nível de significância de 5%. Concluímos, então, que a diferença nas estimativas é possivelmente resul-

tado apenas de flutuação estatística e não de diferenças na fabricação dos multímetros.

Vamos completar o exemplo, avaliando se há informação suficiente para afirmar que os cinco multímetros da oficina eletrônica são mais precisos do que o informado pelo fabricante; embora esse não seja o tema central desta seção, no começo do exemplo ficamos devendo esclarecer essa questão, que requer considerar a f.d.p. de χ^2. A hipótese é que a f.d.p. dos dados seja gaussiana com desvio-padrão $\sigma = 0,25$, que se escreve formalmente[5]

$$\mathrm{H} : f(x_i) = N(x_i; x_0, \sigma) \text{ com } \sigma = 0,25 \quad ,$$

e adotamos um nível de confiança α intermediário, $\alpha = 5\%$. Da expressão de cálculo da variância, relação (1.10), isola-se a variável χ^2,

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} = \sum_{j=1}^{n} \frac{(x_j - \hat{x})^2}{\sigma^2} = \chi^2_{n-1} \quad , \tag{5.27}$$

que tem a f.d.p. de qui-quadrado com $n - 1 = \nu$ graus de liberdade. Como o objetivo é testar a hipótese que a precisão daqueles cinco multímetros é *melhor* do que a especificada pelo fabricante, determina-se um valor crítico de χ^2 tal que a chance de obter valores menores que ele seja pequena. Assim, estabelecemos o valor crítico pela relação

$$\alpha = P(\chi^2 < \chi^2_{\text{critico}}) = \int_0^{\chi^2_{\text{critico}}} F_\nu(\chi^2) d\chi^2 \quad .$$

com $F_\nu\left(\chi^2\right)$ deduzida na seção 2.8. Substituindo o valor de α, escolhido acima, e $\nu = 5 - 1 = 4$ deste exemplo, encontra-se

$$\chi^2_{\text{critico}} = 0,71 \quad .$$

Substituindo na relação (5.27) os valores calculados para o desvio-padrão, $\hat{\sigma} = 0.116$ (veja relação 5.25), $n = 5$ e o valor hipotético $\sigma = 0,25$, obtém-se

$$(n-1)\frac{\hat{\sigma}^2}{\sigma^2} = 0,86 \quad ,$$

que não é menor que o valor crítico, o que *não* indica rejeitar a hipótese de que a precisão dos multímetros seja a especificada pelo fabricante, com nível de significância de 5%.

[5]Lembre-se que neste capítulo estamos simbolizando o desvio padrão verdadeiro por σ sem nenhum adorno.

Em resumo, a f.d.p. da variância é a mesma de χ^2, com as devidas transformações, que deve ser usada para determinar intervalos de confiança para o desvio-padrão. Devemos ter cautela ao fazer avaliações rápidas quando o número de dados é pequeno, como neste exemplo, quando $F_\nu\left(\chi^2\right)$ é uma função bastante assimétrica. Adiante neste capítulo, veremos a aplicação mais conhecida dessa f.d.p. em testes, que corresponde à avaliação da qualidade do ajuste de parâmetros de funções pelo MMQ.

5.8 Teste qualitativo do ajuste de parâmetros

No capítulo precedente, ajustamos os parâmetros $\mathbf{p}$ de uma função $h(\mathbf{x};\mathbf{p})$ a dados experimentais $\{(x_i, y_i, \sigma_i), i = 1, \ldots, N\}$, em que o vetor $\mathbf{x}$ representa o conjunto de variáveis independentes e σ_i, o desvio-padrão de y_i. Os valores medidos y_i são tais que

$$y_i = h(\mathbf{x}, \mathbf{p_0}) + \epsilon_i \quad ,$$

onde $\mathbf{p}_0$ representa os valores verdadeiros dos parâmetros e ϵ_i os erros experimentais, com a propriedade

$$< \epsilon_i^2 >= \sigma_i^2 \quad .$$

Veja as seções 4.5 a 4.8 para os detalhes do modelo.

Pretendemos avaliar se a função calculada com o valores estimados dos parâmetros, $\hat{\mathbf{p}}$, ajusta-se bem aos valores experimentais. Uma idéia mais precisa e objetiva exige os conceitos discutidos nas seções precedentes deste capítulo. Entretanto, pode-se verificar o ajuste por meio de um critério qualitativo baseado na distância do gráfico da função ajustada aos pontos experimentais, considerando-se os desvios padrão σ_i.

Se a f.d.p. dos erros ϵ_i é a gaussiana e os desvios-padrões conhecidos, pode-se calcular a probabilidade de qualquer intervalo conter o valor verdadeiro, em particular, calculamos as probabilidades :

$$\left.\begin{array}{rcl} P(y_i - \sigma_i \leq h(\mathbf{x}_i, \mathbf{p}_0) \leq y_i + \sigma_i) & = & 68,3\% \\ P(y_i - 2\sigma_i \leq h(\mathbf{x}_i, \mathbf{p}_0) \leq y_i + 2\sigma_i) & = & 95,5\% \\ P(y_i - 3\sigma_i \leq h(\mathbf{x}_i, \mathbf{p}_0) \leq y_i + 3\sigma_i) & = & 99,7\% \end{array}\right\} \quad . \tag{5.28}$$

Isto significa que, **em média**, o gráfico da **função verdadeira** deve: cortar cerca de 68,3% das barras de incerteza; passar a menos de 2 barras de incerteza em cerca de 95,5% dos pontos; passar a menos de 3 barras de incerteza na

quase totalidade dos pontos. Lembre-se que as incertezas são representadas convencionalmente por uma barra que se estende um desvio padrão acima e um desvio padrão abaixo do valor experimental. É importante notar que esses números (68% dos pontos, 95% dos pontos, etc) são números médios, o valor real observado em um caso particular é afetado pela flutuação estatística, de maneira análoga ao que discutimos na seção 2.1.

Note, porém, que não temos o gráfico da função verdadeira, apenas o gráfico da função ajustada,

$$h(\mathbf{x}, \hat{\mathbf{p}}) \sim h(\mathbf{x}, \mathbf{p}_0) \quad .$$

Quando substituímos $h(\mathbf{x}_i, \mathbf{p}_0)$ por $h(\mathbf{x}_i, \hat{\mathbf{p}})$ em (5.28), as frações médias reais dos números de pontos cujas barras de incerteza serão cortadas, para os quais a curva ajustada passará a menos de 2 barras de incerteza, etc., são maiores. Isso porque a curva ajustada segue os pontos experimentais melhor que a verdadeira, porque seus parâmetros foram ajustados a eles. No capítulo 8, determinaremos os desvios-padrões desses resíduos, mas, pelo momento, vamos admitir que as relações (5.28) com a substituição dos parâmetros verdadeiros pelos ajustados constituam uma aproximação (frequentemente muito boa, desde que o número de graus de liberdade seja grande) e terminar a discussão.

É principalmente a flutuação estatística das frações observadas que atrapalha a aplicação deste método simples de avaliar a qualidade do ajuste, pela contagem de pontos a uma, duas, e três barras de incerteza da curva ajustada, de forma conclusiva. Por exemplo, quando se ajustam 2 parâmetros a um conjunto de 20 dados experimentais, espera-se que algo como 14 pontos tenham suas barras de incerteza cortadas pelo gráfico da função ajustada, com desvio-padrão aproximadamente igual a $\sqrt{14 \times 0.68 \times 0.32} \approx 2$, de maneira que é possível que tanto 10 barras apenas quanto até mesmo 18 delas tenham sido cortadas pelo gráfico. No entanto, se todas as barras de erro estiverem cortadas, ou apenas meia dúzia, é bem possiível que haja algum problema com o resultado. Assim, o teste é muito útil do ponto de vista qualitativo, especialmente se o número de graus de liberdade é muito maior que o de parâmetros ajustados, quando o impacto da substituição de $h(\mathbf{x}_i, \mathbf{p}_0)$ por $h(\mathbf{x}_i, \hat{\mathbf{p}})$ em (5.28) é pequeno. Enfim, esse método pode fornecer respostas claras, no sentido de revelar se o ajuste é adequado ou não, e se os desvios padrão são razoáveis ou estão subestimados ou superestimados.

5.9 Teste quantitativo do ajuste de parâmetros: a estatística χ^2

O teste que avalia quantitativamente o critério da seção anterior é o de χ^2. Nesta seção, usaremos a mesma notação e consideraremos que o vetor $\mathbf{p}$ representa um número μ de parâmetros,

$$\mathbf{p} = (p_1, p_2, \ldots, p_\mu) \quad .$$

Uma vez que se determinou o vetor $\hat{\mathbf{p}}$ dos parâmetros ajustados, pode-se calcular a variável aleatória

$$\chi^2_{\text{obs}} = \sum_{i=1}^{N} \frac{[y_i - h(\mathbf{x}_i; \hat{\mathbf{p}})]^2}{\sigma_i^2} \quad . \tag{5.29}$$

que obedece à f.d.p. de χ^2_ℓ com

$$\ell = N - \mu$$

graus de liberdade, $F_{N-\mu}(\chi^2)$, da fórmula (2.45), desde que a f.d.p. do erro

$$\epsilon_i = y_i - h(\mathbf{x}_i; \mathbf{p}_0) \tag{5.30}$$

seja gaussiana de desvio padrão σ_i e as covariâncias entre os dados y_i sejam nulas.[6] Descontar o número de parâmetros, μ, do número de pontos para obter o número de graus de liberdade da f.d.p. de χ^2_ℓ é a correção necessária por usar a função ajustada, $h(\mathbf{x}; \hat{\mathbf{p}})$, no lugar da função verdadeira, $h(\mathbf{x}, \mathbf{p}_0)$, como veremos no capítulo 8.

Uma vez que a f.d.p. de χ^2 concentra-se em torno do valor médio, ao menos quando há vários graus de liberdade, esperam-se obter $\chi^2_{\text{obs}} \sim \ell$, sendo pouco prováveis valores próximos de 0 ou valores muito maiores que ℓ. O procedimento para conduzir um teste de hipótese é:

(i). Formular uma hipótese estatística.

(ii). Escolher uma estatística[7] com f.d.p. conhecida e adotar a hipótese como verdadeira.

[6] No capítulo 8, veremos como tratar a questão quando as covariâncias não são nulas.

[7] Estatística, aqui, tem o significado de função cujas únicas variáveis aleatórias são os dados.

(iii). Escolher uma região crítica para a rejeição da hipótese.

A hipótese é

$$\text{H : os erros } \epsilon_i \text{ tem f.d.p. gaussiana de média 0 e desvio padrão } \sigma_i \quad . \tag{5.31}$$

Deve-se observar que essa hipótese já engloba a hipótese da função verdadeira representar corretamente os dados, porque, se não fosse assim, as médias do erros seriam, provavelmente, não nulas (ou a dispersão dos resíduos seria maior que os desvios padrão σ_i).

Uma vez adotada a hipótese, o valor de χ^2_{obs} tem f.d.p. de χ^2 com $\ell = N - \mu$ graus de liberdade, conforme a discussão do início desta seção. É habitual usar um teste monocaudal, de modo que há duas possibilidades para a escolha da região crítica, conforme a cauda escolhida, que correspondem a possibilidades diferentes de violação da hipótese. Embora ambas precisem ser testadas, é costume testá-las isoladamente.

Note que χ^2_{obs} é um somatório de razões, de modo que um valor elevado pode decorrer tanto de numeradores excessivamente grandes quanto de denominadores pequenos. Já um valor muito pequeno de χ^2_{obs} deve decorrer de denominadores excessivamente grandes, não havendo razão para os numeradores serem responsáveis por sua pequenez, uma vez que os numeradores dependem basicamente da adequação da função ajustada.

Ajustar uma forma funcional da relação entre os dados e os parâmetros $h(\mathbf{x}; \mathbf{p})$ que não descreve a função verdadeira aumenta o numerador sistematicamente — ele é a superposição de um erro estatístico à diferença entre o valor de duas funções — o que resulta em aumento de χ^2_{obs}. Quando se escolhe uma região crítica do tipo

$$[\chi^2_{critico}, +\infty[\quad ,$$

verifica-se, principalmente, se a *função modelo* usada, h, não é inadequada.

Da discussão acima percebe-se que há uma outra maneira da hipótese ser falsa, que conduz também a χ^2_{obs} elevado, que corresponde à *subestimação das variâncias.*

Já ao testar uma região crítica do tipo

$$[0, \chi^2_{critico}] \quad ,$$

verifica-se, principalmente, se há *superestimação das variâncias.*

Como exemplo, a figura 5.3 mostra as duas regiões de exclusão para uma curva de χ^2 com 10 graus de liberdade e um nível de significância de 5%. A aplicação do teste a cada uma das regiões críticas está discutida separadamente nas duas próximas sub-seções.

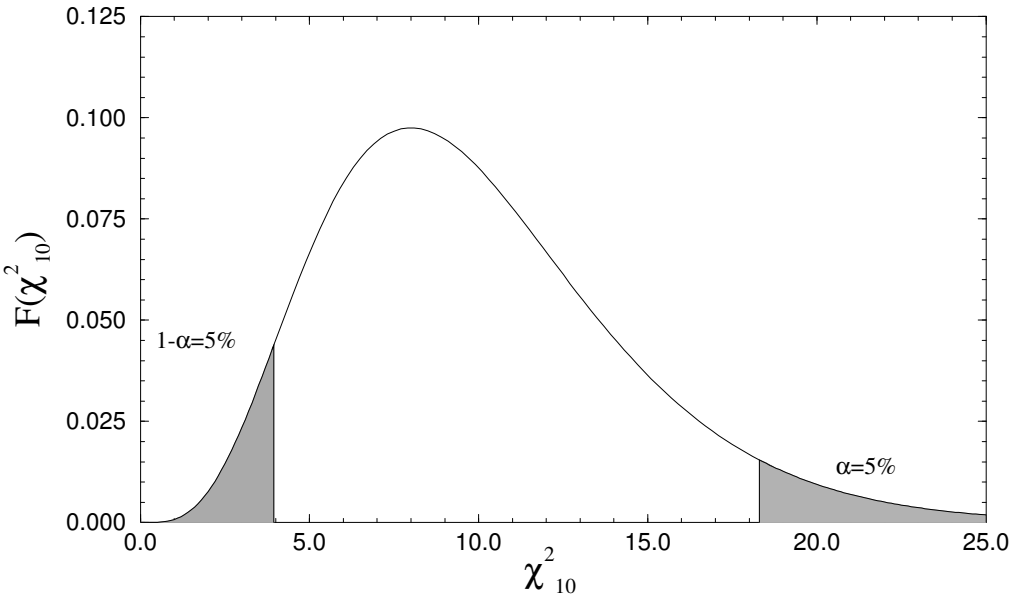


Figura 5.3: Curva de χ^2_{10} com regiões de exclusão de 5%. A região à esquerda exclui um ajuste por ter um $\chi^2_{\rm obs}$ baixo (variâncias superestimadas), caso discutido na seção 5.9.2. A região à direita exclui um ajuste por ter um $\chi^2_{\rm obs}$ alto (variâncias subestimadas ou função que modela inadequadamente a relação entre y e $\mathbf{p}$), caso discutido na seção 5.9.1.

5.9.1 χ^2 "alto"

Ao testar a hipótese (5.31), a região crítica de χ^2 "alto" é o intervalo

$$[\chi^2_{critico}, +\infty[\quad ,$$

ao qual se associa o nível de significância

$$\alpha = P(\chi^2 \geq \chi^2_{critico} \text{ com } \ell \text{ g.l.}) = \int_{\chi^2_{critico}}^{\infty} F_\ell(\chi^2) d\chi^2 \quad . \tag{5.32}$$

A tabela 5.5 apresenta alguns valores de $\chi^2_{critico}$ para os níveis de significância mais comuns e alguns números de graus de liberdade; tabelas mais completas estão no apêndice (**??**). Quando o número de graus de liberdade é grande, pode-se usar aproximações, conforme mencionado na subseção 2.10.3, inclusive a fórmula (2.55). Adiante, discutiremos a escolha do nível de significância.

Quando ajustamos os parâmetros de uma função uma única vez em uma tarefa, é muito comum aplicar o teste de χ^2 sem escolher explicitamente um nível de significância. Podemos calcular qual seria o maior nível de significância

Tabela 5.5: Valores críticos de χ^2 para alguns níveis de significância, α, e alguns valores do número de graus de liberdade, ℓ.

α	0,05	0,01	0,001	α	0,05	0,01	0,001
ℓ				ℓ			
1	3,84	6,63	10,8	8	15,5	20,1	26,1
2	5,99	9,21	13,8	9	16,9	21,7	27,9
3	7,81	11,3	16,3	10	18,3	23,2	29,6
4	9,49	13,3	18,5	15	25,0	30,6	37,7
5	11,1	15,1	20,5	20	31,4	37,6	45,3
6	12,6	16,8	22,5	25	37,7	44,3	52,6
7	14,1	18,5	24,3	30	43,8	50,9	59,7

com o qual H poderia ser rejeitada, que corresponde ao valor acumulado da probabilidade de χ^2 numa região crítica definida pelo valor $\chi^2_{\rm obs}$,

$$\alpha = P(\chi^2 \geq \chi^2_{\rm obs} \text{ com } \ell \text{ g.l.}) = \int_{\chi^2_{\rm obs}}^{\infty} F_\ell(\chi^2) d\chi^2 \quad . \tag{5.33}$$

Esse valor é interpretado como a probabilidade da flutuação estatística originar um conjunto de dados experimentais *ao menos* tão distante da função $h(\mathbf{x}, \hat{\mathbf{p}})$ quanto a medida particular para a qual o ajuste foi efetuado. Não se rejeita, então, o ajuste para o qual α é grande, digamos 5% ou mesmo 1%.

Imagine que $\chi^2_{\rm obs}$ ultrapasse o valor crítico, de modo que o teste efetuado conduza à rejeição da hipótese. Esta rejeição pode corresponder a um erro tipo I (rejeição da hipótese quando ela é verdadeira, por conta da flutuação estatística), mas aconteceu e procura-se uma causa *não aleatória* para a rejeição. Habitualmente, sua origem é atribuída a uma de duas possibilidades:

(i). As variâncias estão subestimadas[8].

(ii). A função h é inadequada.

Em princípio, o engano correspondente ao caso (i) deve ser investigado, voltando-se à etapa de obtenção dos dados experimentais e conferindo os

[8]Covariâncias negativas entre os dados, quando esquecidas, podem dar origem a valores altos de $\chi^2_{\rm obs}$. Veja na seção 8.13 como se incorporam as covariâncias no teste de qui-quadrado.

desvios-padrão, seja efetuando várias observações e estimando os desvios-padrão, seja verificando que todas as fontes de erros que contribuem para σ_i foram incluídas no balanço de incertezas. Além disso, o caso (i) pode ter solução, dentro da hipótese da função ajustada ser adequada e os desvios-padrão dos valores experimentais serem bem conhecidos a menos de um fator comum, estando errados apenas os valores absolutos das variâncias. Trataremos deste caso no capítulo 8, mas adiantamos que ele resulta em *aumentar* os desvios padrão dos parâmetros estimados. Já o caso (ii) leva a procurar outra forma para a função $h(\mathbf{x}, \mathbf{p})$ que representa o modelo do sistema em estudo.

De qualquer maneira, este teste sozinho não permite distinguir se o que está errado é a função ou o conjunto das variâncias. Algumas vezes, o comportamento dos resíduos em função da variável independente x apresenta diferenças sistemáticas, tal como muitos resíduos seguidos com o mesmo sinal. Convém lembrar que o resíduo tem 50% de probabilidade de ser positivo e 50% de ser negativo, de modo que é muito provável ter 2 ou 3 pontos seguidos com resíduos de mesmo sinal, mas 7 ou 8 devem acontecer mais raramente; a probabilidade exata pode ser calculada com a distribuição binomial, mas isto é um outro teste de hipótese, não é o teste com base no χ^2, que tem exatamente a vantagem de dispensar uma análise detalhada.

5.9.2 χ^2 "baixo"

Testar a hipótese (5.31) quando se pretende avaliar se houve superestimação das variâncias corresponde a escolher como região crítica o intervalo

$$[0, \chi^2_{critico}] \quad ,$$

ao qual está associado o nível de significância

$$\alpha = P(\chi^2 \leq \chi^2_{critico} \text{ com } \ell \text{ g.l.}) = \int_0^{\chi^2_{critico}} F_\ell(\chi^2) d\chi^2 \quad . \tag{5.34}$$

A tabela 5.6 apresenta alguns valores de $\chi^2_{critico}$ para níveis de significância de 0,1%, 1% e 5%. Caso se chegue à conclusão que a hipótese deva ser rejeitada, excluindo-se a possibilidade de um erro tipo I, devemos buscar a origem da superestimação das variâncias. Um erro que observamos algumas vezes em dados que correspondem à média de várias observações é usar σ_i igual ao desvio-padrão da série e não ao desvio-padrão da média.[9]

[9]Covariâncias positivas entre os dados, quando esquecidas, podem dar origem a valores

Tabela 5.6: Valores críticos de χ^2 para níveis de significância de 0,1%, 1% e 5%, em um teste baseado na cauda esquerda da f.d.p. de qui-quadrado, para alguns valores do número de graus de liberdade, ℓ.

	$\chi^2_{critico}$				$\chi^2_{critico}$		
ℓ	0,1%	1%	5%	ℓ	0,1%	1%	5%
1	$1{,}57\times10^{-6}$	0,00016	0,0039	8	0,856	1,65	2,74
2	0,002	0,0201	0,104	9	1,152	2,09	3,32
3	0,024	0,115	0,351	10	1,48	2,56	3,94
4	0,920	0,297	0,712	15	3,48	5,23	7,26
5	0,210	0,554	1,145	20	5,92	8,26	10,9
6	0,384	0,872	1,63	25	8,65	11,5	14,6
7	0,595	1,24	2,17	30	11,6	15,0	18,5

Um outro procedimento que conduz a χ^2 baixos é a remoção de pontos "suspeitos" simplesmente por estarem longe da curva ajustada. Isso invalida a hipótese dos erros serem gaussianos e, portanto, invalida o teste também.

Exemplo 5.2

Escolha do nível de significância no teste de χ^2

Quando se faz apenas um ajuste, é muito comum considerá-lo adequado

$$0,05 \leq P(\chi^2_\ell \geq \chi^2_{\text{obs}}) \leq 0,95 \quad .$$

Entretanto, ao efetuar centenas de ajustes, mesmo escolhendo um nível de significância relativamente pequeno como 0,01, provavelmente haverá um ou poucos ajustes com χ^2 dentro da região crítica sem que a hipótese H (5.31) seja falsa, sem que haja algo errado. Simplesmente, o número médio de erros tipo I cresce proporcionalmente com o número de testes efetuados. Ao escolher o nível de significância do teste de χ^2, portanto, é preciso levar em conta o número de vezes que o teste será aplicado. Em particular, quando se efetuam muitos ajustes, normalmente é preciso reduzir esse nível. Vamos discutir este assunto por meio de um exemplo.

baixos de χ^2_{obs}. Veja na seção 8.13 como se incorporam as covariâncias no teste de qui-quadrado.

A referência [Vanin 1989] descreve uma situação onde foram ajustados parâmetros de uma função a muitos conjuntos de dados, cada conjunto gerando um vetor de parâmetros, em um total de $M = 150$ ajustes. Em cada ajuste, calculou-se $\chi^2 = \chi^2_{\rm obs}$ e rejeitou-se ou não a hipótese (5.31) — portanto o ajuste — comparando o valor obtido com o crítico, $\chi^2_{critico}$. Escolhendo o valor crítico de maneira que a probabilidade de se obter um χ^2 maior que ele corresponda ao nível de probabilidade $p = 5\%$,

$$P(\chi^2_\ell \geq \chi^2_{critico}) = \int_{\chi^2_{critico}}^{\infty} f_\ell(\chi^2)\, d\chi^2 = p = 0,05 \qquad ,$$

a função de probabilidade do número m de casos que cairão na faixa crítica de χ^2, $[\chi^2_{critico}, \infty[$, é a binomial

$$P_{M,p}(m) = \binom{M}{m} 0,05^m\ 0,95^{M-m} \qquad \textbf{com } M = 150 \quad ,$$

que indica que em torno de 7 a 8 casos darão χ^2 nessa faixa sem que nem σ_i^2 nem a função sejam inadequadas. Os resultados correspondentes seriam, portanto, rejeitados equivocadamente (erros tipo I).

Aqui, como na seção 5.3, um aumento do valor crítico de χ^2 tem dois efeitos: reduzir a probabilidade de erro tipo I e aumentar a probabilidade de erro tipo II.[10] Não podemos, então, aumentar impunenente o valor crítico de χ^2. Uma possibilidade consiste em escolher um valor crítico tal que a probabilidade *global* de ocorrer um certo número de erros tipo I seja prefixada. Um critério possível corresponde a fixar em 50% a probabilidade de não haver nenhum erro tipo I, o que equivale a fixar também em 50% a probabilidade de haver ao menos um erro tipo I. Neste exemplo, isso corresponde a

$$P(n = 0) = 0,5 \Leftrightarrow$$

$$\binom{M}{0} (1-p)^M = 0,5 \rightarrow$$

$$p \cong 0,5\% \quad .$$

Portanto, ao fixar em 0,5% a probabilidade de erro tipo I em um único ajuste, fixa-se em 50% a probabilidade de não cometer nenhum erro tipo I

[10] A probabilidade de cometer um erro tipo II depende da hipótese alternativa, que poderia ser calculada quando ela fosse definida. No entanto, não é comum conhecer-se a função $h'(\mathbf{x}, \mathbf{p})$ alternativa, de modo que, quase sempre, essa probabilidade fica indeterminada. No entanto, a discussão desta seção não depende desse conhecimento detalhado.

na avaliação do conjunto dos 150 ajustes. Em outras palavras, espera-se não obter *nenhum* $\chi^2_{\rm obs}$ na região crítica ou obter *um* ou *dois* ajustes que resultem em $\chi^2_{\rm obs}$ acima de $\chi^2_{critico}$.

5.10 Qui-quadrado reduzido

O valor esperado de χ^2 é dado pela fórmula (2.46), $< \chi^2 >= \ell$, onde ℓ é o número de graus de liberdade, o que origina a regra habitual segundo a qual o *ajuste é bom quando* o *qui-quadrado reduzido*, χ^2_{red}

$$\chi^2_{red} = \frac{\chi^2}{\ell} \tag{5.35}$$

é *próximo de 1*, uma vez que esse é seu valor médio. Em muitos casos, entretanto, obter o valor de qui-quadrado reduzido igual a 1, ou melhor, numa faixa estreita em torno de 1, digamos no intervalo $[0,9; 1,1]$, é pouco provável. É necessário avaliar o quanto o valor de qui-quadrado reduzido pode diferir de 1 por flutuação estatística.

Uma simplificação é usar o desvio padrão como medida da flutuação estatística. A fórmula (2.47) da seção 2.9 mostra que o desvio padrão da f.d.p. de χ^2 é $\sqrt{2\ell}$. Usando como orientação genérica que os valores mais prováveis são aqueles numa faixa em torno do valor médio de mais ou menos dois desvios-padrão, o intervalo com os valores mais prováveis de qui-quadrado é

$$[\ell - 2\sqrt{2\ell}; \ell + 2\sqrt{2\ell}] \tag{5.36}$$

o que corresponde à faixa de qui-quadrado reduzido

$$\left[1 - 2\sqrt{\frac{2}{\ell}}; 1 + 2\sqrt{\frac{2}{\ell}}\right] \quad . \tag{5.37}$$

Assim, para um número *pequeno* de graus de liberdade, a faixa de qui-quadrado reduzido provável de obter-se é *larga* e, para um número **grande**, é **estreita**. Por exemplo, para 10 graus de liberdade, a faixa é $[0,1; 1,9]$ e para 200 graus de liberdade, é $[0,8; 1,2]$.

Essa regra deve ser usada com cautela para um número pequeno de graus de liberdade (menor que 30) devido à assimetria da f.d.p. de χ^2, mas ela é preferível ao hábito de utilizar-se apenas o número 1 como guia.

Embora aproximada, a regra (5.37) acima fornece intervalos razoavelmente adequados. O que não se pode fazer é tentar estabelecer um intervalo de qui-quadrado reduzido que valha para **qualquer** número de graus de liberdade. Há uma regra prática que muitos experimentadores adotam quando o número de graus de liberdade, no ajuste de parâmetros aos seus dados, é mais ou menos constante, decorrente da pequena variação do intervalo de qui-quadrado reduzido com o número de graus de liberdade **para valores próximos e razoavelmente grandes**, e que é estabelecer uma faixa de qui-quadrado reduzido aceitável para **suas condições experimentais particulares**. Embora regras assim possam ser úteis para avaliações rápidas do andamento de um experimento, elas devem ser usadas com cautela.

EXERCÍCIOS

5.1. Uma maneira possível de identificar que ^{238}U está em uma amostra é pela radiação gama de 1001,03 keV (considere esse valor como exato) emitida após a desintegração do ^{234}Pa, um dos elementos de sua cadeia de decaimento. Para investigar a presença de ^{238}U, usou-se um detetor de raios gama capaz de indicar a energia da radiação com um desvio padrão de 1,8 keV, com uma distribuição gaussiana.

(a) A observação de um único gama resultou em uma energia de 1002,3 keV. Faça um teste z para decidir quanto a rejeitar ou não a hipótese do gama detectado ter 1001,03 keV de energia.

(b) Posteriormente, duas outras medidas resultaram em 1001,7 keV e 1003,5 keV. Considerando os três dados obtidos, faça um novo teste z quanto à presença de ^{238}U na amostra.

(c) Estime quantos dados precisariam ser obtidos com esse detetor para que o resultado pudesse ser usado para testar a hipótese de uma diferença de energia entre a grandeza medida e um valor hipotético de 0,5 keV com um nível de significância de 1%.

5.2. Quatro medições da densidade de um líquido, feitas com um equipamento, obedecem a uma distribuição gaussiana. São eles: 1,21; 1,18; 1,26 e 1,03 g/cm^3.

(a) Use o teste t para verificar a hipótese de a amostra ser de água, cuja densidade é 1,00 g/cm^3.

(b) Outra amostra líquida foi medida com a mesma balança (e, portanto, com dados que devem obedecer à mesma f.d.p.), resultando em: 1,34; 1,52 e 1,06 g/cm^3. Teste a hipótese de que as densidades das duas amostras sejam iguais.

5.3. Em um laboratório didático, duas turmas de estudantes, cada uma com 6 grupos, mediram a aceleração da gravidade local. Os resultados dos grupos das turmas I e II são, respectivamente, $\{8,5; 8,0; 9,2; 8,7; 8,8; 10,3\}$ e $\{9,2; 9,9; 9,9; 9,5; 10,5; 10,0\}$, todos em m/s^2.

(a) Supondo que os resultados da turma I obedeçam a uma mesma f.d.p. gaussiana, use o teste t para verificar se esses dados são compatíveis com o valor $g = 9,796$ m/s^2 dentro de níveis de significância de 5%, 1% e 0,1%.

(b) Faça o mesmo teste com os resultados da turma II.

(c) Use o teste F para verificar se o conjunto de medições da turma I pode ter o mesmo desvio padrão que o conjunto da turma II, com nível de significância de 5%.

(d) Supondo que os desvios padrão das turmas I e II sejam iguais, use o teste t para verificar se ambas as medidas podem corresponder a uma mesma grandeza física.

5.4. Medidas da energia da radiação gama emitida por uma amostra pouco intensa resultaram nos seguintes dados (em keV): 173,5; 172,9; 172,5; 171,6. A resolução do detetor é estimada pelo desvio padrão desse conjunto de valores, no caso igual a 0,80 keV. O mesmo detetor foi usado para medir os raios gama emitidos por outra fonte radioativa, obtendo-se os dados 175,0; 173,1; 176,3, todos em keV. Neste caso, o desvio padrão dos dados é 1,61 keV, correspondendo a outra estimativa da resolução do detector.

(a) Como os desvios padrões dos dados são diferentes nos dois casos, use o teste F para verificar a hipótese de mau funcionamento do detetor, com piora de sua resolução na segunda medida.

(b) Supondo que os valores verdadeiros dos desvios padrões dos dois conjuntos de dados sejam iguais, faça um teste t para a hipótese de que as energias dos raios gama com que foram tomados os dois conjuntos de dados sejam iguais.

5.5. Duas teorias diferentes prevêem valores exatos para quatro grandezas físicas, as quais foram medidas experimentalmente. A tabela abaixo apresenta os dois conjuntos de valores teóricos, comparados com os experimentais.

Tabela do exercicío 5.5. A segunda e a terceira colunas apresentam os valores teóricos de quatro grandezas físicas, cujos valores experimentais estão relacionados na última coluna.

grandeza	Teoria I	Teoria II	Experimento
A	240,0	230,0	217(12)
B	-400,0	-420,0	-415(7)
C	55	18,0	32(9)
D	124	90,0	100(10)

Considerando os valores da tabela, use o teste de χ^2 para avaliar as teorias com níveis de confiança de 5%, 2,5% e 0,5%.

5.6. Uma espécie radioativa emite fótons, cujas energias e intensidades de emissão são caraterísticas particulares da sua espécie, que servem de identificação. Além disso, cada espécie radioativa apresenta um tempo característico de decaimento, normalmente descrito pela *meia vida*.

Numa medida em que diversos nuclídeos estão presentes, as espécies podem ser identificadas, então, pelas energias das radiações emitidas, sendo que as intensidades relativas dessas radiações não dependem do tempo. Isto permite que um conjunto de medidas consecutivas, separadas por intervalos de tempo comparáveis à meia-vida, defina se um certo conjunto de radiações provém do mesmo nuclídeo. Na hipótese delas provirem de uma mesma espécie radioativa, as intensidades relativas dessas radiações devem permanecer constantes ao longo da sequência de medidas.

Uma fonte radioativa, contendo uma ou talvez duas espécies radioativas, foi observada emitindo radiações das seguintes energias: 127, 196, 238, 307 e 507 keV. Mediu-se a intensidade relativa à transição de 307 keV em cinco espectros; radiação dessa energia é caraterística do nuclídeo ^{164m}Ho. A tabela abaixo resume os resultados obtidos em medidas consecutivas, identificadas pelas letras A até E.

Para cada energia, determine a intensidade média, seu desvio padrão e faça um teste de χ^2. Tendo em vista o resultado desse teste, a hipótese destas radiações originarem-se de um único nuclídeo é razoável? (Esses dados são reais, não foram simulados.)

Tabela de intensidade no espectro (exercicío 5.6)

E_γ(keV)	A	B	C	D	E
127	1,22(11)	1,33(11)	1,46(11)	1,25(11)	1,42(11)
196	0,53(6)	0,57(6)	0,67(6)	0,78(6)	0,80(6)
238	0,73(9)	0,53(9)	0,73(9)	0,96(9)	0,81(9)
507	71(8)	71(7)	58(7)	53(6)	69(6)

5.7. A relação entre y e x é da forma $y = ax$. Para estimar o valor de a, foram medidos valores de y para $x = 1$, $x = 2$ e $x = 3$, obtendo-se 2,3; 2,5 e 4,5, respectivamente, todos com mesmo desvio padrão 1,0 e covariâncias nulas.

(a) Determine o valor de a pelo método dos mínimos quadrados. Não deixe de determinar o desvio padrão de a, também.

(b) Realize um teste de qui-quadrado a partir do resultado obtido no ajuste efetuado no item acima.

(c) Supondo que a relação entre y e x seja $y = 3x - 4$, realize um novo teste de qui-quadrado. Compare as duas funções do ponto de vista de qualidade do ajuste.

5.8. Um experimentador ajustou os parâmetros a e b da reta $y = a + b \cdot x$ aos pontos experimentais $(x_i; y_i; \sigma_i)$:

$$\{(1; 4,12; 0,15), (2; 5,32; 0,15), (3; 6,31; 0,15)\}.$$

(a) Calcule as estimativas de a e de b, bem como seus respectivos desvios padrão.

(b) Calcule χ^2 e a probabilidade desse χ^2 ser excedido. Este teste suporta a hipótese da dependência linear entre y e x?

5.9. As figuras I, II e III mostram os dados experimentais, com suas respectivas barras de incerteza supostamente iguais aos desvios padrão, obtidos pela observação da posição de um corpo em Movimento Retilíneo e Uniforme. A equação horária é

$$y = y_0 + v_0 t.$$

Em cada gráfico, está representada a reta

$$y = \hat{y} + \hat{v}t,$$

com os parâmetros ajustados pelo método dos mínimos quadrados. Pergunta-se, em cada caso, se o ajuste obtido é adequado ou não. Justifique suas respostas e aponte eventuais problemas com os dados experimentais (x_i, y_i, σ_i).

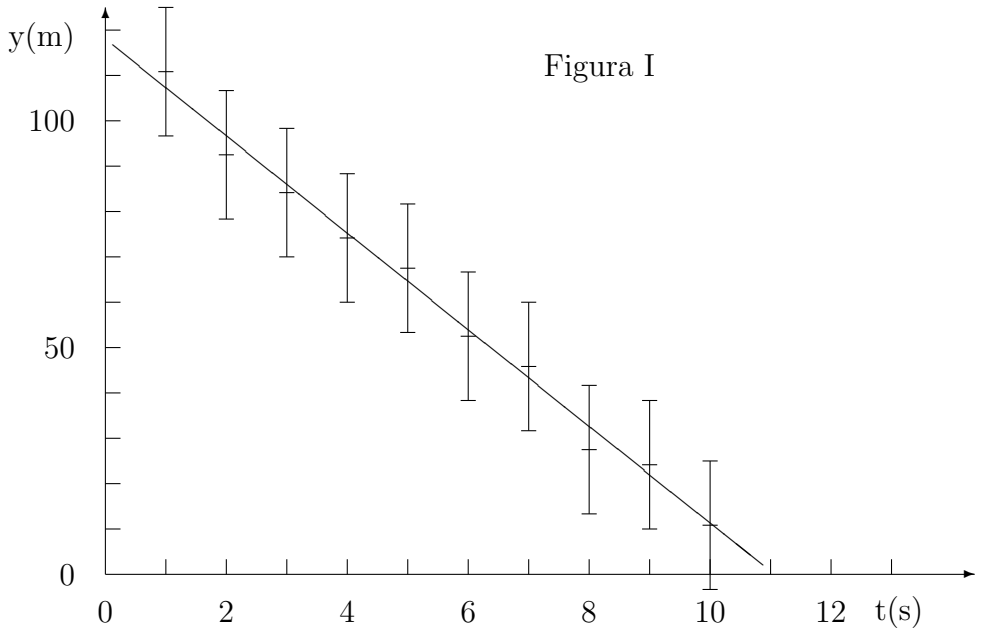
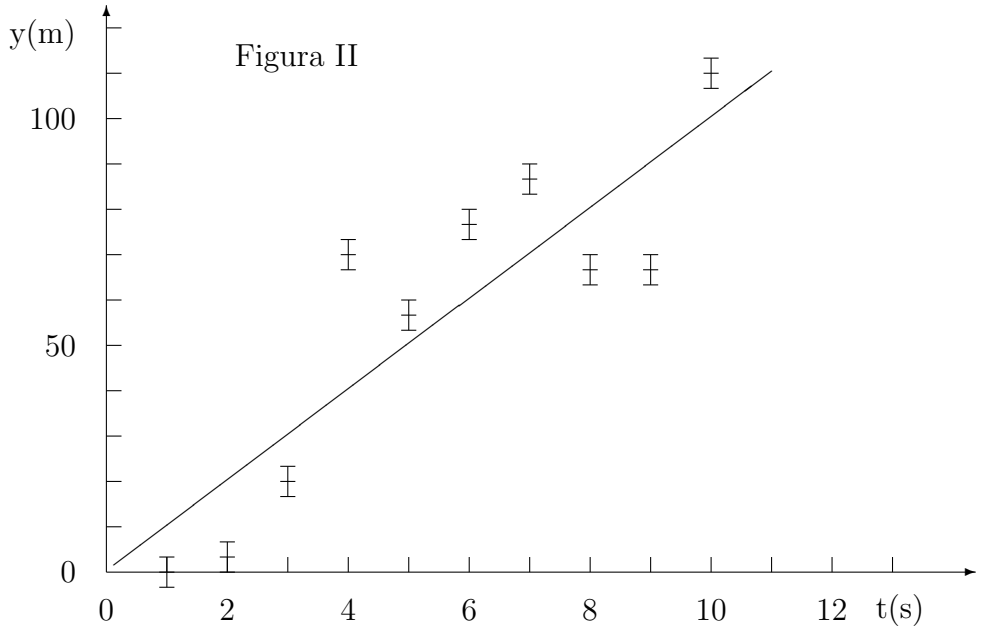
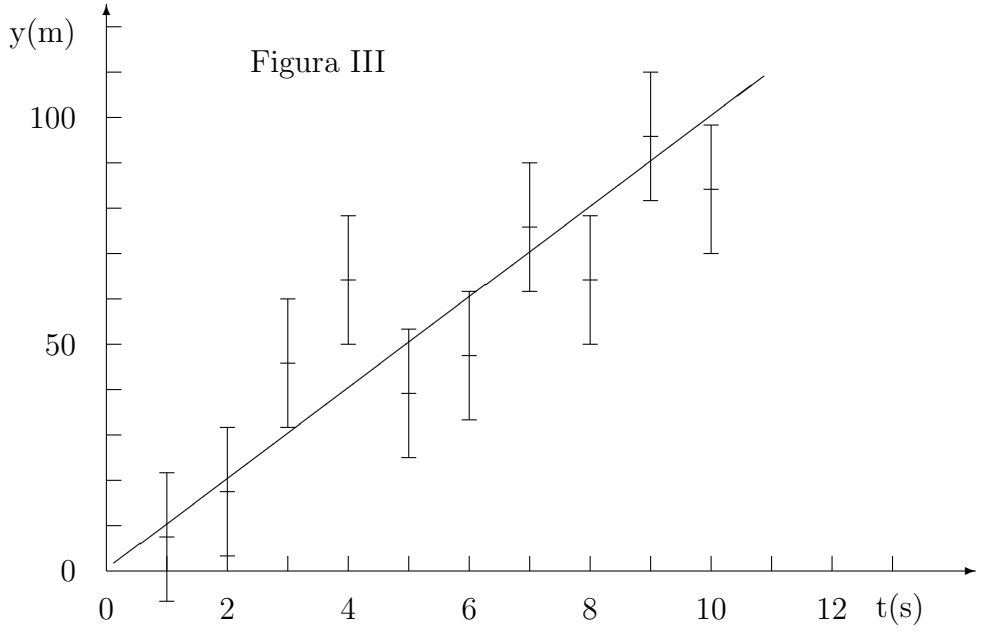


Figura I


y(m)
Figura II
100
50
0
0
2
4
6
8
10
12
t(s)


y(m)
Figura III
100
50
0
0
2
4
6
8
10
12
t(s)

Bibliografia

[Arfken] Mathematical Methods for Physicists, G.Arfken & H.Weber, Academic Press, 4^a edição (1995)

[Bard] Nonlinear Parameter Estimation, Yonathan Bard, Academic Press (1974)

[Benzécri] Histoire et Préhistoire de l'Analyse des Données, J.P.Benzécri, Ed. Bordas, Paris 1982

[Bevington] Data Reduction and Error Analysis for the Physical Sciences, P.Bevington, McGraw-Hill, 1969

[Birge] The calculation of errors by the method of least squares, Raymond T. Birge, Phys Rev *40 (1932) 207-227*

[Conover] Practical Nonparametric Statistics, W.J.Conover, John Wiley & Sons Inc. 1971

[CRC] Handbook of Tables for Probability and Statistics, CRC

[Eadie] Statistical Methods for Physicists, W.T.Eadie et al., North Holland Pub.Co. 1971

[Escoubes] Experimental Signs Pointing to a Bayesian Instead of a Classical Approach for Experiments with Small Number of Events, B.Escoubes, S.De Unamuno e O. Helene, Nuclear Instruments and Methods A257(1987)346

[Feller] Feller, An Introduction to Probability Theory and its Applications, John Wiley, 2^a Ed. (1957)

[Feynman] Lectures on Physics Vol.I, Chap.6, Feynman Leighton & Sands

[Firestone] Analysis of α, β, and γ ray emission probabilities, R.B. Firestone, Nuclear Instruments and Methods *A286*(1990)584

[Forbes] Forbes, Eric G., *Gauss and the Discovery of Ceres*. Journal for the History of Astronomy. 2 (1971) 195?199.

[Frieden] Fisher's Information as the basis for the Schrödinger wave equation, B. Roy Frieden, Am. J.Phys. 57(1989)11

[Geraldo] L.P. Geraldo e D.L Smith, Nuclear Instruments and Methods A290(1990)499

[Grosser] Morton Grosser, The Discovery of Neptune, Harvard University Press, Cambridge, Massachusetts (1962)

[Gray] C.G.Gray, Am. J.Phys. 59(1991)282

[Guimarães-Filho] Z.O. Guimarães-Filho e O. Helene, One Step Self-Calibration Procedure in Gamma-Ray Energy Measurements. Brazilian Journal of Physics, v. 33, n.2, (2003) 280-281.

[Lyons] How to combine correlated estimates of a single physical quantity, L.Lyons, D.Gibaut e P. Clifford, Nuclear Instruments and Methods *A270*(1988)110

[Helene] Tratamento Estatístico de dados em Física Experimental, O.Helene, V. R. Vanin, Ed. Edgard Blücher, 2^a Ed., 1991

[Helene 83] Upper Limit of Peak Area, O.Helene, Nuclear Instruments and Methods 212(1983)319

[Helene 84] Errors in Experiments with Small Number of Events, O.Helene, Nuclear Instruments and Methods 228(1984)120

[Helene 91b] Determination of the Upper Limit of Peak Area, O.Helene, Nuclear Instruments and Methods A300(1991)132

[Helene 91] O que é uma medida?, O. Helene, Shan.P.Tsai, R.P.Teixeira, preprint IFUSP/P-854 (1990) e Revista de Ensino de Física, Vol.13 p.12, SBF (1991).

[Helene 93] O.Helene and V.R.Vanin, Nuclear Instruments and Methods A335(1993)227

[Helene 2013] O. Helene, Método dos Mínimos Quadrados com formalismo matricial, Editora Livraria da Física, São Paulo, 2ª edição (2013).

[James] A review of pseudorandom number generators, F.James, Computer Physics Communications 60(1990)329-344

[Kendall] The Advanced Theory of Statistics, M.Kendall, A.Stuart & J.K.Ord, Charles Griffin & Company Limited, London

[Magalhães] Noções de Probabilidade e Estatística, Marcos N. Magalhães e Antonio Carlos P. Lima, Editora da Universidade de São Paulo - EDUSP, 2011

[Mannhart] A Small Guide to Generating Covariances of Experimental Data, Report PTB–FMRD 84, Berlin, 1981. ISSN 0341–6666

[Marquardt] An Algorithm for Least-Squares Estimation of Nonlinear Parameters, D. Marquardt, SIAM J. Appl. Math. 11, 431-441, 1963

[Merzbacher] Quantum Mechanics, E.Merzbacher, John Wiley & sons, New York 1961

[Mises] Probability, Statistics and Truth, R.von Mises, Dover, 1955

[Moralles] M.Moralles, P.R.Pascholati, V.R.Vanin and O.Helene, Applied Radiation and Isotopes 46-2(1995)133

[Mucciolo] E.R.Mucciolo and O.Helene, Nuclear Instruments and Methods *A256*(1987)153

[Noether] Introdução à Estatística – Uma abordagem não paramétrica, G.E.Noether, Guanabara Dois, 1983

[Smith] D.L. Smith, Nuclear Instruments and Methods *A257*(1987)361

[Stigler] *Gauss and the Invention of Least Squares.* Stephen M. Stigler, Annals of Statistics, *9* (1981) 465–474 - doi:10.1214/aos/1176345451

[Vanin 1989] V.R.Vanin e M.Aiche, Nuclear Instruments and Methods *A284*(1989)452

[Vanin 1997] V.R.Vanin, G.Kenchian, M.Moralles, O.Helene e P.R. Pascholati, Nuclear Instruments and Methods *A391*(1997)338

[Vuolo] Fundamentos da Teoria de Erros, J.H.Vuolo, Ed. Edgard Blücher, 1992

[Youden] Statistical Methods for Chemists, W.J.Youden, John Wiley 1951

[Zar] J.H. Zar, Appl. Statist. 27(1978)n.3, 280-290