4.3.1. Exemplo de um *teste de significância*, usando tachinhas

Para este experimento, usaremos tachinhas comuns (Fig. **2**), porque são mais fáceis de obter do que os outros dados assimétricos. Se lançarmos uma tachinha, ela pode cair em duas posições: com a ponta para cima, ou com a ponta para baixo, tocando o solo. Qual delas é mais provável? Não podemos saber, em princípio.

Consideraremos inicialmente que a tachinha é equilibrada; isto é, que a probabilidade de cair com a ponta para baixo seja igual à probabilidade de cair com a ponta para cima:

P(ponta para cima) = P(ponta para baixo)

Já que as duas probabilidades são iguais, podemos enunciar esta hipótese como:

Hipótese: P(ponta para cima) = 0,5

Note que este tipo de teste é normalmente chamado de teste de *proporções*, porque é frequentemente usado para testar hipóteses sobre a proporção de sucessos encontrados em uma população; por exemplo, sobre a proporção de peças defeituosas produzidas por uma fábrica. No entanto, este teste pode ser considerado equivalente ao teste de uma *probabilidade*: se retiro aleatoriamente uma peça destas, a probabilidade de ela ter defeito é igual à proporção de peças defeituosas, de acordo com a definição frequencista de probabilidade.

Como testar se a hipótese acima é verdadeira? A pergunta que é a base de qualquer teste estatístico é : *se a hipótese for verdadeira, que resultado eu deveria encontrar na amostra, quando fizer um experimento*? Se encontramos este resultado, ótimo, a hipótese parece ter sido corroborada; se encontramos um resultado muito diferente do esperado, desconfiamos que haja algo errado com a hipótese.

Faremos 20 lançamentos da tachinha e contamos o número de vezes em que ela cai com a ponta para cima. Estes 20 lançamentos serão uma *amostra* de lançamentos; a *população* será uma população imaginária, composta dos infinitos lançamentos possíveis da tachinha. O número de "pontas para cima" obtidas (X) será a *variável de teste*; o resultado do teste depende do valor observado nesta variável.

Se a tachinha for equilibrada, podemos calcular as probabilidades de obtermos cada um dos valores possíveis de X (de 0 a 20) por meio de um modelo binomial de distribuição com parâmetros n = 20 tentativas e probabilidade de sucesso p = 0,5. O resultado deste cálculo é a distribuição de probabilidades na Tabela **1**.

Esta distribuição é unimodal e simétrica; o valor de X com maior probabilidade é o valor central, x=10, e os outros valores se distribuem simetricamente em torno deste centro. Se lançarmos a tachinha 20 vezes e o número de pontas para cima estiver no centro da distribuição, em torno do 10, este resultado será considerado coerente com a hipótese; poderemos então concluir que não há evidência de que a hipótese seja falsa. O que acontece, porém, se o número de pontas para cima não estiver no centro da distribuição, mas sim nos extremos – isto, é se for muito menor, ou muito maior que 10? Um resultado destes dificilmente ocorre se a hipótese for verdadeira. Esses valores de X são *muito improváveis*; se eles ocorrem, isto pode ter acontecido porque :

- houve algum erro no experimento (no lançamento da tachinha, na contagem dos lançamentos, no registro dos resultado, etc.)
- quem lançou o dado é muito azarado, e obteve um resultado muito improvável;
- a hipótese inicial (de que a tachinha é equilibrada) não é verdadeira.

Tabela 1. Distribuição de probabilidades da variável *X = número de "pontas para cima"*, em 20 lançamentos de uma tachinha equilibrada

X	p(x)	X	p(x)
00	.0000		
01	.0000	11	.1602
02	.0002	12	.1201
03	.0011	13	.0739
04	.0046	14	.0360
05	.0148	15	.0148
06	.0360	16	.0046
07	.0739	17	.0011
08	.1201	18	.0002
09	.1602	19	.0000
10	.1762	20	.0000

Se temos certeza de que não houve nenhum erro no experimento (e não acreditamos que ninguém possa ser tão azarado), iremos concluir que o problema deve estar na hipótese – este resultado é uma *evidência* contra a hipótese, sugerindo que ela talvez não seja verdadeira e que a tachinha seja desequilibrada. Quanto mais longe de 10 estiver o valor de X encontrado, maior será a evidência de que a hipótese é falsa.

Quando os resultados de um experimento destes são publicados, a posição na distribuição em que situou o valor encontrado de X (o número de vezes em que a tachinha caiu com a ponta para cima) é indicado por uma probabilidade chamada de *p-value*, representado pela letra *p*. (Em português, *p-value* costuma ser traduzido como *valor-p*, *p-valor* ou *valor de p*; usaremos a expressão *valor-p*, que nos parece a tradução mais correta). O valor-*p*, neste problema, será calculado pela soma das probabilidade de todos os pontos, desde o valor observado de X até o extremo da distribuição, multiplicado por dois.

No teste que fizemos com alunos de uma turma, a tachinha caiu 12 vezes com a ponta para cima (X=12). A probabilidade total de todos os valores de 12 até o extremo superior da distribuição é, usando os valores na Tab. **1**:

$$P(X \geq 12) = P(X=12)+P(X=13)+P(X=14)+...+P(X=20) =$$
$$= 0{,}1201+0{,}0739+0{,}0360+...+0{,}0000 = 0{,}2507$$

O valor-*p* é então:

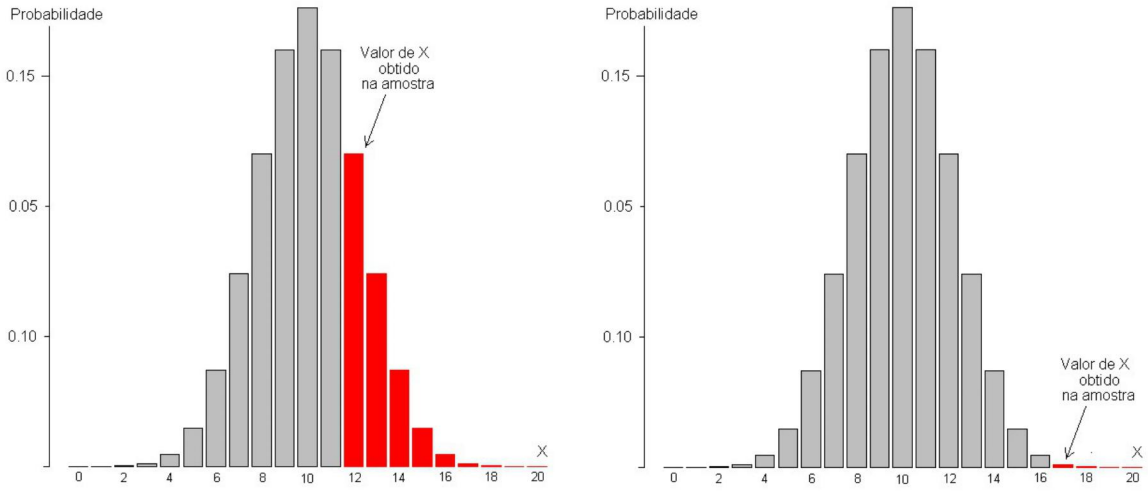
$$p = 2 \times 0{,}2507 = 0{,}5014$$

Por que multiplicar por 2 a soma das probabilidades? Porque o teste, neste caso, é *bilateral*; se a hipótese for verdadeira, o valor de X não deve ser nem muito maior, nem muito menor do que 10. (Isto será explicado melhor na Seção 4.4.3).

O valor-*p* obtido indica que o valor de X encontrado na amostra está localizado perto do centro da distribuição (Fig. **3A**), e que portanto tem grande probabilidade de ocorrer se a hipótese for verdadeira. Ele pode ser considerado como uma evidência *a favor* da hipótese de que a tachinha é equilibrada (mas uma evidência muito fraca, como veremos depois). Um resultado destes é chamado de "não-significativo", já que ele não traz nenhuma evidência *contra* a hipótese inicial.

Suponha agora que a tachinha tenha caído 17 vezes com a ponta para cima. O somatório das probabilidade de todos os valores de 17 até o extremo superior da distribuição (Fig. **3B**) seria, usando os valores na Tab. **1**:

$$P(X \geq 17) = P(X=17) + P(X=18) + P(X=19) + P(X=20) =$$
$$= 0{,}0011+0{,}0002+0{,}0000+0{,}0000= 0{,}0013$$



A. Para X = 12 B. Para X = 17

Figura 3. Exemplos do cálculo do valor-p

O valor-*p* calculado seria então:

$$p = 2\times0{,}0013 = 0{,}0026$$

Este valor-p indica que o número de vezes em que a tachinha caiu com a ponta para cima está n um dos extremos da distribuição (não importa qual deles). Quando isto acontece, chamamos o resultado de "significativo" (*significant*), e dizemos que ele é uma evidência de que a hipótese inicial é falsa, e de que a tachinha não é equilibrada.

Se em vez de 17, tivéssemos encontrado X=16 pontas para cima, o valor-p seria

$$P(X \geq 16) = P(X=16)+P(X=17)+P(X=18)+P(X=19)+P(X=20) = 0{,}0118$$
$$p = 2\times0{,}0118 = 0{,}0236$$

Este valor ainda é significativo, mas menos do que o resultado X=17. Quanto *menor* o valor-*p* encontrado em um experimento, *mais significativo* é o resu-tado, e maior a evidência contra a hipótese nula (isto é, evidência de que ela é falsa). Na prática, a maioria dos pesquisadores usam como referência o valor 0,05 ou o valor 0,01. Um resultado de X=17 seria considerado significativo, se for usada qualquer destas duas referências; um resultado X=16 seria considerado significativo se usarmos como referência o valor 0,01, mas não se usarmos 0,05; o resultado X=12 seria sempre considerado não-significativo.

Este tipo de teste foi proposto por Ronaldo Fisher, nos anos 1920. (Fisher foi provavelmente o personagem mais importante no desenvolvimento da Estatística no século XX). O objetivo de um teste destes não é dizer se a hipótese é *verdadeira* ou *falsa*, mas sim avaliar *quão forte* é a evidência contra a hipótese; o resultado de um experimento é considerado *significativo* quando indica que há um forte evidência contra a teoria; estes testes são por isso chamados de *testes de significância*. Fisher sugeriu que a força da evidência seja classificada de acordo com a escala na Tab. **2** (extraída de [[1]]):

Tabela 2. Escala de significância proposta por Fischer

valor-p	0,10	0,05	0,025	0,01	0,005	0,001
evidência	marginal	moderada	substancial	forte	muito forte	fortíssima

Fisher considerava que a forma de teste que propôs é coerente com o que ocorre no desenvolvimento da ciência. A ciência avança por meio de tentativas, erros e correções de erros. Uma teoria nunca é considerada *verdadeira* ou *falsa* depois de apenas um experimento; uma teoria nova passa a ser aceita de forma progressiva, depois que começam a se acumular evidências a seu favor e contra a teoria velha. (É o que acontece, por exemplo, em Medicina: com frequência vemos remédios já bem conhecidos que desaparecem das farmácias, ou formas de tratamento que são abandonadas, depois que o acúmulo de evidências indica que são menos eficazes do que outras alternativas existentes).

Alguns estatísticos argumentaram, porém, que isto nem sempre é suficiente; é preciso às vezes tomar uma decisão sobre a hipótese: afinal, ela é *verdadeira* ou *falsa*? Um exemplo disto ocorre nos testes de controle de qualidade de produtos, como o que foi mencionado na Seção 4.2.4: Um fabricante produz e vende um tipo de peças, e afirma que apenas 3% delas são defeituosas; o comprador, ao receber lotes destas peças, faz testes para verificar se a afirmação é verdadeira ou não, e decidir se vai aceitar ou não os lotes. O comprador não está interessado em dizer simplesmente que "existe uma forte evidência contra (ou a favor) desta afirmação"; terá que decidir se vai aceitar as peças, ou se vai devolvê-las para o fabricante.

A dificuldade é que, ao tomar esta decisão, o comprador poderá estar cometendo erros, tanto se aceitar quanto se rejeitar a afirmação do fabricante. Se numa amostra ele encontrar um número de peças com defeito muito maior do que seria provável, decide rejeitar as peças. Esta decisão porém será baseada em probabilidades, e nunca será *certa*; o comprador nunca terá *certeza* de que a decisão que tomou é correta (a não ser que ele examine todas as peças, o que geralmente não é possível fazer).

Jerzy Nyman e Egon S. Pearson criaram, a partir dos anos 1930, a teoria que fundamenta os *testes de hipótese*, nos quais as decisões são tomadas com base em probabilidades dos erros: o pesquisador toma a decisão que tem maior probabilidade de estar correta. Para entender o raciocínio destes testes, é necessário apresentarmos novos conceitos, que serão vistos a seguir.

Referência

[1] Bussab, Wilton de O. e Morettin, Pedro A. (2013). *Estatística Básica*. São Paulo: Saraiva.

Resumo

- Os testes de *significância* procuram avaliar quão *significativo* é o resultado encontrado numa amostra.
- Um resultado é *significativo* quando traz evidência de que hipótese que estamos testando é falsa.
- O grau de significância de um resultado é medido pelo *valor-p*.
- Um valor-p muito pequeno indica que o resultado encontrado é improvável, e dificilmente teria ocorrido se a hipótese fosse verdadeira; há evidência, portanto, de que a hipótese é falsa.
- Quanto *menor* o valor-p, *mais* significativo é o resultado, e maior a evidência contra a hipótese.