## 4.4. Testes de proporções



---

Vimos na Seção 3.3.5 que o número $X$ de sucessos (ou a proporção $P$ de sucessos) obtidos em $n$ repetições de uma experimento cujo resultado é binário (cada repetição resulta ou em *sucesso* ou em *fracasso*) tem uma distribuição binomial B($n$,$p$), cujos parâmetros são $p$, a probabilidade de sucesso em cada tentativa, e $n$, o número de tentativas. Na prática, como na maior parte das vezes as amostras usadas para fazer inferências sobre variáveis binárias são grandes (amostras pequenas levam a testes de muito baixo poder, e estimações com pouca precisão), e o modelo binomial não pode ser usado.

Algum leitor porém pode ser perguntar neste ponto: para que serve tudo isto? Que aplicação prática ou científica estes testes podem ter? Afinal, a maioria das pessoas não pretende dedicar parte de suas vidas testando se uma tachinha é equilibrada ou não... Na verdade, testes de hipóteses sobre proporções (ou sobre probabilidades, o que dá no mesmo) tem aplicações em várias áreas. Um exemplo já mencionado, na indústria: um fabricante afirma que a probabilidade das peças que produz terem algum tipo de defeito é de 0,01. Como o comprador pode testar se isto é verdade? Um exemplo nas ciências: Gregor Mendel, o fundador da Genética moderna, fez milhares de experimentos com cruzamentos de ervilhas, estudando como as características da planta se transmitem de geração a geração. Uma destas características foi a superfície da ervilha, que pode ser *lisa* ou *rugosa*. Segundo a teoria que criou, no cruzamentos de ervilhas lisas, a probabilidade de os descendentes serem ervilhas rugosas seria de 0,25. Como testar se esta teoria é verdade?

Como visto na Seção 4.2.1, para testar uma hipótese precisamos de *deduzir* alguma de suas conseqüências (o que aconteceria se a hipótese fosse verdadeira) e depois verificar empiricamente, por experimentos ou por observação, se esta conseqüência prevista realmente ocorreu.

A análise dos resultados de Mendel é muito semelhante, em essência, à que vimos no teste da tachinha, na seção 4.3.1. Se supomos que a tachinha é equilibrada, esperamos que em 20 lançamentos ela caia com a ponta para cima cerca de 10 vezes; se cruzamos 100 pares de ervilhas, esperamos que as descendentes sejam rugosas cerca de 25 vezes. Em ambos estudos, os resultados são aleatórios, e têm que ser interpretados em termos de probabilidades. Assim como a tachinha não vai necessariamente cair 10 vezes com a ponta para cima, as ervilhas não terão necessariamente 25 descendentes rugosas; o que esperamos é que seja *mais provável* que os número de ervilhas rugosas esteja num intervalo em torno de 25. O problema agora é encontrar os limites deste intervalo; como se definem os limites "em torno"? Para isto, precisamos de um modelo de distribuição de probabilidades (quer dizer, de uma fórmula que nos permita calcular as probabilidades da variável que nos interessa).

Esta distribuição das probabilidades dos resultados que podem ocorrer na amostra é chamada de *distribuição amostral*. É ela que nos permite calcular o que seria mais provável encontrar na amostra se a hipótese for verdadeira, e a partir daí fazer testes estatísticos. Se o valor que encontrarmos na amostra for um valor que tinha sido considerado

"provável", concluímos que há evidência a favor da hipótese; se for um valor que tinha sido considerado "improvável", concluímos que a evidência é contra a hipótese.

O experimento de Mendel e o das tachinhas têm o mesmo padrão; em ambos os problemas, o experimento é formado essencialmente por várias repetições de tentativas binárias, e a variável que nos interessa é o número ou a proporção de sucessos obtidos.

No exemplo da tachinha, a distribuição amostral tinha um modelo binomial B(20; 0,5). No exemplo das ervilhas, não poderemos usar a binomial porque o número de repetições é grande demais (Mendel fez milhares de repetições dos cruzamentos, durante um período de oito anos); usaremos em vez disso a distribuição normal como aproximação da binomial, como já foi feito na seção 3.4.4.5. Os resultados de Mendel serão vistos na seção 4.4.5.

### 4.4.1. Distribuição amostral de proporções

Como na maior parte das aplicações que realmente nos interessam as amostras usadas para fazer inferências sobre proporções são grandes, é a aproximação da *binomial* feita pelo modelo *normal* o que usaremos com mais freqüência. Usando esta aproximação, podemos enunciar o seguinte teorema sobre a *distribuição amostral* do número X de sucessos em amostras grandes:

**Teorema 1** *: Distribuição amostral do número X de sucessos*
Se retiramos amostras aleatórias de tamanho *n* (*n* suficientemente grande) de uma população infinita com proporção de sucessos π, o número X de sucessos na amostra terá distribuição que tende para a normal:

$$X \sim N(\mu_X, \sigma_X^2) \text{ quando n} \to \infty$$

cujos parâmetros são $\mu_X = n\pi$ e $\sigma_X^2 = n\pi(1-\pi)$

A variável de teste será a variável padronizada *Z*, dada por:

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \qquad (1)$$

cuja distribuição tende para a normal padrão, Z ~ N(0,1), quando n $\to \infty$

Também é comum trabalharmos com as *proporções P* de sucessos nas amostras ($P = X/n$), ao invés de com os *números X* de sucessos nas amostras. O teorema então pode ser reescrito como:

**Teorema 2:** *Distribuição amostral da proporção de sucessos P*
Se retiramos amostras aleatórias grandes de uma população infinita com proporção de sucessos π, a proporção de sucessos P=X/n nas amostras terá distribuição que tende para a normal:

$$P \to N(\mu_P, \sigma_P^2) \text{ quando n} \to \infty$$

onde: $\mu_P = \pi$ e $\sigma_P^2 = \frac{\pi(1-\pi)}{n}$

A variável de teste será a variável padronizada Z, dada por:

$$Z = \frac{P - \mu_P}{\sigma_P} = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}} \qquad (2)$$

e terá distribuição que tende para a normal padrão, Z ~ N(0,1), quando n $\to \infty$

Assim, para amostras grandes, podemos fazer testes sobre os valores mais prováveis das variáveis X ou P a partir de uma distribuição normal, e não mais de uma binomial.

### 4.4.2. Regiões de rejeição e aceitação

Na curva normal padronizada, os valores de Z que limitam as regiões de rejeição e aceitação podem ser encontrados nas tabelas de curvas normal (seção 3.6.3) ou usando um programa estatístico qualquer. Para os valores mais usuais de α, estes valores críticos são dados na Tab. 1.

**Tabela 1 - Valores críticos de Z**

| valor de α | teste unilateral | teste bilateral |
|---|---|---|
| 0,01 | + 2,33 (teste à direita)<br>– 2,33 (teste à esquerda) | ± 2,58 |
| 0,05 | + 1,64 (teste à direita)<br>– 1,64 (teste à esquerda) | ± 1,96 |

Se o teste é unilateral, a região de rejeição será colocada apenas em um extremo da curva. Para um teste unilateral à direita, por exemplo, com α = 0,05, o valor crítico de Z será de $z_c$ = +1,64, como na Fig. 1.

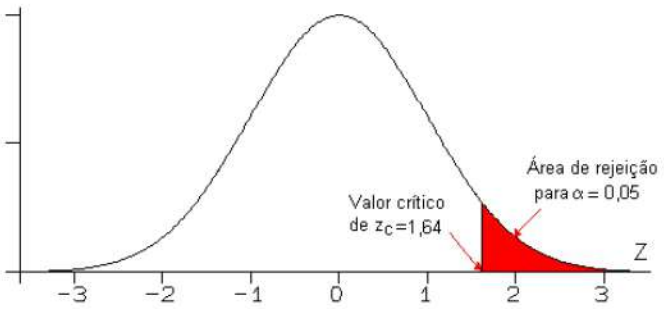


**Figura 1. Região de rejeição em teste unilateral, α=0,05**


**Figura 2. Região de rejeição em teste bilateral, α=0,05**

Se o teste for unilateral à esquerda, com o mesmo α = 0,05, o valor crítico de Z será de $z_c$ = - 1,64, com a área de rejeição do outro lado da curva. Se o teste for bilateral, a região de rejeição é repartida entre os dois extremos da curva normal. Para teste com α=0,05, por exemplo, o valor crítico de Z será de $z_c$ = ± 1,96, como na Fig. 2.

### 4.4.3. Uso do *valor-p* para indicar a significância de um resultado

O valor-p foi criado para indicar quão forte é a evidência encontrada numa amostra *contra* uma hipótese que está sendo considerada, num *teste de significância*. Este valor mede a distância que existe entre o que encontramos numa amostra e o que teria sido mais provável encontrarmos se a hipótese fosse verdadeira; se esta distância for grande, isto quer dizer que o que encontramos foi muito diferente do que era esperado, e que a hipótese provavelmente não é verdadeira.

Suponha um *teste de hipótese* sobre a proporção numa população, baseado na proporção de sucessos $P$ encontrada na amostra. Se a hipótese nula for verdadeira, este $P$ deve estar próximo do $\pi$ dado pela hipótese. A distância entre $P$ e $\pi$ pode ser padronizada por meio da eq. (2); obtemos então a variável $Z$, que é a chamada *distância estatística* entre es-tes dois valores. Se a hipótese nula for verdadeira, esperamos que $Z$ seja próximo de zero; se $Z$ for muito grande, teremos que compará-lo com os valores críticos (que dependem do α escolhido), para decidirmos o que fazer. Se o teste é unilateral à direita, por exemplo, e calculamos $Z$=3,00, isto quer dizer que o que encontramos na amostra difere muito do que era esperado, e que podemos com confiança rejeitar a hipótese nula. Se calculamos $Z$=1,12 a distância entre o que encontramos e o que era esperado não é muito grande, e a hipótese nula se mantém. Se calculamos $Z$=1,90, porém, a decisão é mais complicada: este valor está além do valor crítico se considerarmos α=0,05 ($Z_c$=1,64), mas não se considerarmos α=0,01 ($Z_c$=2,33).

Num artigo descrevendo os resultados deste teste, podemos simplesmente dizer que "os valores encontrados na amostra levaram à rejeição da hipótese nula". Isto porém não indica qual foi a significância do resultado ou qual o grau de confiança que depositamos na conclusão. Se encontramos Z=3,00, teremos praticamente certeza de que a hipótese é falsa; se encontramos Z=1,90, também rejeitamos a hipótese, mas não com tanta certeza.

Uma maneira mais informativa de relatar o resultado seria dizer qual foi o valor de Z encontrado; mas então seria preciso dizer também qual foi a hipótese nula, se o teste foi bilateral ou unilateral, qual foi o nível de significância adotado e (para testes que não usam a distribuição normal) outras informações, como o tamanho da amostra, etc. Com estes dados os leitores – se tiverem conhecimento estatístico e paciência suficientes – poderão deduzir quão significativo foi o resultado.

A método mais usado atualmente para reportar o resultado de um teste é dar o *valor-p* encontrado, o que é apenas uma outra maneira de transmitir a informação dada por Z. Num teste unilateral à direita como na Fig. 1, ao invés de darmos o valor de Z, damos o valor da área sob a curva normal à direita do ponto Z. Se o valor-p é maior do que o tamanho da área de rejeição ($p > \alpha$), isto indica que o Z encontrado ficou fora da área de rejeição, e a hipótese nula não é rejeitada (Fig. 3A). Se porém o valor-p for menor do que o tamanho da região de rejeição ($p < \alpha$), isto indica que o valor Z ficou dentro da região de rejeição, e a hipótese nula deve ser rejeitada (Fig. 3B). Quanto mais afastado da média da distribuição estiver o valor Z, menor será o valor-p, e mais significativo o resultado. Para os três valores de Z acima, os valores-p são (calculados no R):

$Z = 1{,}12 \quad \rightarrow p = 0{,}131$
$Z = 1{,}90 \quad \rightarrow p = 0{,}029$
$Z = 3{,}00 \quad \rightarrow p = 0{,}001$

Num teste unilateral à esquerda, os valores de Z seriam evidentemente negativos, e o valor-p seria igual à área sob a curva normal à esquerda do valor Z.

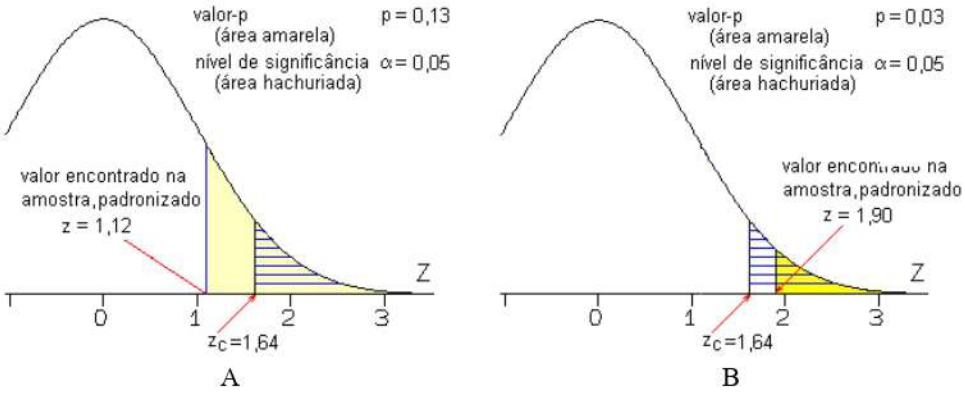


**Figura 3. Comparação entre valor-p e nível de significância, num teste unilateral**

Se o teste for bilateral, o cálculo do valor-p é um pouco diferente. Se por exemplo encontramos o valor Z=1,70 num teste bilateral com α=0,05, temos a situação mostrada na Fig. 4. A região à direita de Z é igual a 0,045; num teste unilateral, este resultado seria significativo, pois esta área é menor do que a região de rejeição (0,045 < 0,05).

No teste bilateral, porém, não podemos fazer esta comparação, pois a região de rejeição foi dividida em duas partes, uma metade em cada extremo da curva; no lado direito a área da região é portanto de 0,025.


**Figura 4. Comparação entre a área limitada por Z e o nível de significância, num teste bilateral**

Para decidirmos se o resultado é significativo, comparando a área à direita de Z com o nível de significância, há duas possibilidades:

(i) comparamos a área à direita de Z com *metade* do valor de α:
0,045 > 0,025, portanto o resultado *não* é significativo;

(ii) comparamos o *dobro* da área à direita de Z com o valor de α:
2×0,045 = 0,09 > 0,05, portanto o resultado *não* é significativo;

O procedimento usualmente adotado é o (ii); portanto, num teste bilateral iremos chamar de valor-p o *dobro* da área à direita de Z. (Se Z for negativo, o dobro da área à esquerda de Z).

Note que neste exemplo o resultado seria significativo se o teste fosse unilateral, mas não seria se fosse bilateral. Vale lembrar aqui o que foi dito na seção 4.3.3: as decisões sobre o nível de significância α e o tipo de teste (unilateral ou bilateral) devem ser to-

madas *antes* de que a amostra tenha sido retirada e que seu resultado tenha sido calculado! Mudar depois as decisões para que um resultado se torne significativo é tão honesto quanto mudar as regras do jogo no meio da partida, para garantir a vitória do time favorito...

### 4.4.4. Exemplos

*(i) Razão de sexo*

Nos exemplos iniciais de probabilidades (seção 3.1.6.2), fizemos a suposição de que a probabilidade de nascerem *meninos* é igual à de nascerem *meninas*, para simplificar. Na realidade, estas probabilidades não são exatamente iguais; a probabilidade de nascimento de *meninos* é quase sempre maior, mas varia de um país para outro. A razão entre o número de meninos nascidos para cada 100 meninas é a chamada de *razão de sexo.* (Veja o verbete "*sex ratio*" na Wikipedia).

Suponha que desejamos testar a hipótese de que a probabilidade de nascerem meninos é maior do que a de nascerem meninas. Fazendo π = P(menino), as hipóteses são:

$H_0$ : $\pi \leq 0{,}5$
$H_1$ : $\pi > 0{,}5$

Para fazer o teste, usamos uma amostra de 495 nascimentos ocorridos em um hospital de São Paulo, num período de duas semanas. Foram encontrados 260 meninos e 235 meninas. Considere que estas crianças sejam uma amostra aleatória simples dos nascimentos ocorridos no país. Se a probabilidade de nascimento de um *menino* for igual à de uma *menina*, qual é a probabilidade de numa amostra aleatória sejam encontrados 260 ou mais meninos? (isto é, qual o valor-p do resultado desta amostra?). O resultado encontrado nesta amostra pode ser considerado uma evidência de que a probabilidade de nascimento de meninos é maior do que a de meninas, se usarmos α=0,05?

Usaremos a aproximação feita pelo modelo normal para a distribuição amostral da *proporção* de meninos na amostra (Teorema **2**). Os valores dados são:

n = 495 tamanho da amostra
X = 260 número de meninos na amostra

Usando um nível de significância α=0,05, podemos calcular:
- Proporção de meninos na amostra

$$P = 260 / 495 = 0{,}5252$$

- Parâmetros da distribuição amostral gaussiana:

$$\mu_P = \pi = 0{,}5$$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}5(1-0{,}5)}{495}} = 0{,}0225$$

Note que nas duas fórmulas acima o valor real de π é (obviamente!) desconhecido, e usamos o valor definido pela hipótese nula.

- Variável padronizada Z, calculada a partir do valor amostral P:

$$z = \frac{P - \mu_P}{\sigma_P} = \frac{P - \mu}{\sqrt{\pi(1-\pi)/n}} = \frac{0{,}5252 - 0{,}5000}{0{,}0225} = 1{,}12$$

Como o valor de P encontrado na amostra, padronizado, está abaixo do valor crítico de $z_c = 1{,}64$, ele caiu fora da região de rejeição. O gráfico na Fig. **3A** mostra a posição deste valor padronizado em relação ao valor crítico, na curva normal. O valor-p correspondente a este valor de Z num teste unilateral será (calculado usando R):

valor-p = P(z ≥ 1,12) = 0,13

Note que a probabilidade de numa amostra aleatória de 495 nascimentos serem encontrados 260 ou mais meninos é razoavelmente alta (p=0,13), se as probabilidades de me-ninos e meninas forem iguais. Portanto, apesar de na amostra haver mais meninos do que meninas, este resultado é *não-significativo*; isto é, este resultado *não* pode ser considerado como evidência de que a probabilidade de nascimento de meninos seja maior do que a de meninas.

*(2) Razão de sexo*

Repetimos agora o teste, usando uma amostra (muito) maior. De 4.065.014 nascimentos registrados nos EUA em 1992, houve 2.081.287 meninos (P=0,5120) e 1.983.727 meninas (P=0,4880) [**1**]. Faremos os cálculos usando agora o *número* de sucessos encontrado na amostra, em vez de a *proporção* de sucessos (Teorema **1**). O teste será unilateral, com as mesmas hipóteses do teste anterior e mesmo α=0,05:

$H_0$ : $\pi \leq 0{,}5$
$H_1$ : $\pi > 0{,}5$

$$X = 2.081.287$$
$$\mu_X = n\pi = 0{,}5 \times 4065014 = 2032507$$
$$\sigma_X = \sqrt{4065014 \times 0{,}5(1-0{,}5)} = 1008{,}1$$
$$z = \frac{X - \mu_X}{\sigma_X} = \frac{2081287 - 2032507}{1008{,}1} = 48{,}4$$

valor-p ≅ 0.0000

O valor-p obtido mostra que o resultado obtido nesta amostra é extremamente *significativo*, e pode ser considerado uma evidência muito forte a favor da hipótese de que a probabilidade de nascimento de meninos é *maior* do que a de meninas (nos EUA, naquela época; esta conclusão não pode porém ser automaticamente estendida para o Brasil atual).

Note que nesta amostra (nascimentos nos EUA) a proporção de meninos foi menor do que no exemplo anterior (nascimentos em SP) – as proporções foram 0,5120 e 0,5252, respectivamente. Mesmo assim o teste com a amostra americana deu resultado altamente significativo, o teste com a amostra de SP não. Por que um teste deu resultado significativo e o outro não? Por causa do tamanho das amostras; a primeira é pequena demais e o teste não teve poder suficiente para detectar a diferença entre as probabilidades de nascimento de meninos e meninas (diferença que geralmente é pequena, apenas 1 ou 2%).

O raciocínio usado aqui é o mesmo do teste feito para verificar se uma tachinha era equilibrada, na seção 4.3.2; a diferença é que lá foi usada uma amostra minúscula (n=20) e a distribuição *binomial* para fazer o teste, e aqui foram usadas duas amostras grandes, e usamos a distribuição *normal* como aproximação da binomial.

*(iii) Lançamentos de tachinhas (de novo!)*

Voltamos agora ao problema do lançamento de tachinhas. No experimento visto na Seção 4.3.1, a tachinha caiu 12 vezes com a ponta para cima, e este resultado foi considerado *não-significativo*; concluímos então que não havia evidência de que a tachinha fosse desequilibrada. Um teste destes, porém, feito com uma amostra de apenas n=20 lançamentos é pouco útil, pois tem um poder muito pequeno.

Fizemos agora 50 amostras de 20 lançamentos, e o gráfico de ramo-e-folhas da Fig. 5 mostra o número de sucessos (pontas para cima) em cada amostra.

```
 7 | 000
 8 | 000
 9 | 0000000
10 | 00000000
11 | 0000000000
12 | 000000
13 | 00000000
14 | 0000
15 |
16 |
17 |
18 | 0
```

**Figura 5. Número de sucessos obtidos em 50 amostras de n=20**

Qual será a conclusão agora, se considerar que todos estes 50×20 lançamentos como uma única amostra de n=1000? Um teste com uma amostra deste tamanho tem muito mais poder, e talvez seja capaz de detectar um pequeno desequilíbrio na tachinha que os testes com n=20 não perceberam.

O número total de sucessos obtidos nos 1000 lançamentos foi X=548. Usaremos este resultado para fazer um teste bilateral, considerando α=0,05 e as hipóteses:

$H_0: \quad \pi = 0{,}5$

$H_1: \quad \pi \neq 0{,}5$

$$P = X/n = 548/1000 = 0{,}548$$

$$\mu_P = \pi = 0{,}5$$

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0{,}5(1-0{,}5)}{1000}} = 0{,}0158$$

$$z = \frac{P-\mu_P}{\sigma_P} = \frac{0{,}548-0{,}500}{0{,}0158} = 3{,}036$$

valor-p = 0,0024

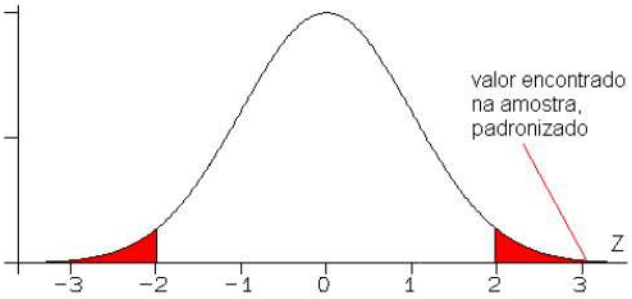


**Figura 6. Posição do valor encontrado na amostra, em relação à distribuição amostral**

O gráfico da Fig. **6** mostra que o valor de X encontrado na amostra de n=1000 é muito pouco provável se a hipótese for verdadeira (se a tachinha for equilibrada). Portanto, ele é uma forte evidência de que a tachinha é *desequilibrada*. Veremos depois (seção 4.8) como fazer uma estimativa da verdadeira probabilidade de a tachinha cair com a ponta para cima, como exemplo das técnicas para *Estimação de Parâmetros*.

---

**referências**

[1] Pagano, M. e Gavreau, K. (2004). *Princípios de Bioestatística*. São Paulo: Thomson.

**Resumo**

- A distribuição amostral é um modelo probabilístico que mostra o que é mais provável encontrarmos na amostra *se a hipótese nula for verdadeira*. As distribuições amostrais são a base da Inferência Estatística.
- Num teste de hipótese sobre a proporção de sucessos numa população, com amostras grandes, a distribuição amostral segue o modelo *normal*.
- O *valor-p* mostra onde o valor encontrado na amostra se localiza, dentro da distribuição amostral.
- Um valor-p pequeno indica que o que foi encontrado na amostra está nos extremos da distribuição amostral, muito longe do que seria mais provável encontrar se a hipótese nula fosse verdadeira.
- Se o que encontramos na amostra é muito improvável, concluímos que a hipótese nula deve ser falsa.