4.3.2. Exemplo de *teste de hipótese*

4.3.2.1. Hipótese *nula* x hipótese *alternativa*
4.3.2.2. Tipos de erros
4.3.2.3. Teste de hipótese unilateral
4.3.2.4. Resumo

Nas próximas seções, analisaremos um exemplo simples de um *teste de hipótese*, que é uma continuação do exemplo de *teste de significância* mostrado na seção 4.3.1. Ambos usam amostras muito pequenas (n = 20) e se baseiam na distribuição binomial; veremos depois que não têm por isso grande aplicação prática, porque o tamanho das amostras é insuficiente. No entanto, servem para apresentar de forma intuitiva, sem grandes complicações matemáticas, todos os conceitos importantes relativos aos testes estatísticos. Além disso, podem ser facilmente reproduzidos por estudantes numa turma, e são uma boa opção para aulas introdutórias sobre Inferência Estatística.

4.3.2.1. Hipótese *nula* x hipótese *alternativa*

Iremos agora aumentar um pouco a complexidade, em termos conceituais, do teste feito na seção anterior. Suponha agora que você acredita que a tachinha seja *desequilibrada*, isto é, que probabilidade de a tachinha cair com a ponta para cima seja diferente da probabilidade de cair com a ponta para baixo:

$$P(\text{ponta para cima}) \neq P(\text{ponta para baixo})$$
$$P(\text{ponta para cima}) \neq 0{,}5$$

Representaremos a probabilidade P(ponta para cima) pela letra grega π. A hipótese que nos interessa testar, portanto, é

$$\pi \neq 0{,}5$$

Como testar esta hipótese? Para fazer o teste, usando o método da seção anterior e o mesmo número n=20 de lançamentos, precisamos de um valor definido para π, a fim de calcularmos a distribuição de probabilidades usando o modelo Binomial com n=20 e p= π, O problema é que, agora, o valor de π não é um número, e sim um intervalo contínuo ($\pi \neq 0.5$) que contém infinitos valores. A solução é usarmos duas hipóteses, ao invés de uma só. A primeira, chamada de *hipótese nula* (representada por H_0), iguala π a um valor definido. Neste exemplo, seria:

$$H_0 \rightarrow \ \pi = P(\text{ponta para cima}) \ = 0{,}5$$

(Esta é a mesma hipótese que usamos para o teste de significância na seção anterior). A segunda hipótese, chamada de *hipótese alternativa* e representada por H_1, será a hipótese que aceitaremos caso a primeira seja rejeitada. No exemplo,

$$H_1 \rightarrow \ \pi \ \neq 0{,}5$$

Os cálculos da distribuição de probabilidades serão feitos com base na hipótese nula, exatamente como fizemos na seção anterior, e o resultado será a distribuição de probabilidades mostrada na Tab. **1**.

Evidentemente, os valores mais prováveis *se a hipótese nula for verdadeira* (se

$\pi = 0,5$) serão aqueles em torno de X=10. Podemos delimitar um intervalo no centro desta distribuição que tenha uma grande probabilidade de conter o valor de X; na tabela, este intervalo está marcado em azul. No intervalo entre 06 e 14, inclusive, estão os valores de X mais prováveis; no total, este intervalo tem uma probabilidade de aproximadamente 0,95:

P(X=6) + P(X=7) + ... + P(X=14) = 0,0360 + 0,0739 + ... + 0,0360 = 0,9586

Os valores de X fora deste intervalo estão nas duas áreas marcadas em vermelho; como a distribuição é simétrica, as probabilidades destas duas áreas serão iguais e, somadas, atingem um pouco menos de 0,05:

P(X<6) = P(X=0) + P(X=1) +... + P(X=5) = 0,0000 + 0,0000 + ... + 0,0148 = 0,0207
P(X>14) = P(X<6) = 0,0207
P(X<6) + P(X>14) = 0,0207 + 0,0207 = 0,0414

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

```
+-----------+
|  X   p(x) |
+-----------+
| 00  .0000 | <----┐   Se a tachinha for equilibrada, é pouco provável que
| 01  .0000 |      +-- o número de "pontas para cima" esteja neste intervalo
| 02  .0002 |      |   (as probabilidades de todos estes valores, somadas,
| 03  .0011 |      |   dão um total de apenas 0,0207).
| 04  .0046 |      |   Região de rejeição
| 05  .0148 | <----┘
| 06  .0360 | <----┐
| 07  .0739 |      |
| 08  .1201 |      |
| 09  .1602 |      |   Se a tachinha for equilibrada, é muito provável que
| 10  .1762 |      +-- o número de "pontas para cima" esteja dentro deste
| 11  .1602 |      |   intervalo(as probabilidades de todos estes valores,
| 12  .1201 |      |   somadas, dão um total de 0,9586)
| 13  .0739 |      |   Região de aceitação
| 14  .0360 | <----┘
| 15  .0148 | <----┐
| 16  .0046 |      |
| 17  .0011 |      |   Como a distribuição de probabilidades é simétrica,
| 18  .0002 |      |   este extremo da distribuição tem a mesma probabilidade
| 19  .0000 |      +-- do outro extremo : 0,0207
| 20  .0000 | <----┘   Região de rejeição
+-----------+
```

A região que contém estes valores de X que seriam mais prováveis caso H_0 seja verdadeira será chamada de *região de aceitação* da hipótese nula; a região que contém os valores que seriam menos prováveis será chamada de *região de rejeição*.

Podemos a partir daí criar uma *regra de decisão*: se ao lançarmos a tachinha 20 vezes encontrarmos um X dentro da região de aceitação (entre 6 ou 14 sucessos), iremos *aceitar* a hipótese H_0 (isto é, considerar que a tachinha é equilibrada). Esta decisão está baseada nas probabilidades: um resultado nesta região seria o que esperaríamos que acontecesse, se a tachinha for equilibrada. Por outro lado, se o número de vezes que a tachinha cair com a ponta para cima estiver dentro da *região de rejeição*, iremos rejeitar a hipótese H_0, porque que este resultado seria muito improvável se a tachinha fosse equilibrada; aceitaremos então a H_1, que afirma que a tachinha é desequilibrada.

No teste feito com 20 lançamentos de uma tachinha (Seção 4.3.1), obtivemos como resultado X=12 sucessos; este valor está dentro da região de aceitação da hipótese nula, portanto aceitamos a hipótese de que a tachinha seja equilibrada.

4.3.2.2. Tipos de erros

O problema que enfrentamos, ao fazer o teste acima, é que toda decisão tem que ser tomada com base em probabilidades, e nunca poderemos ter certeza de que ela será a decisão correta. O que podemos fazer é escolher a decisão que tem a maior probabilidade de estar *certa*; mesmo assim, ainda pode ser que ela esteja *errada*. É um risco que temos que correr, e não há como eliminá-lo. O que caracteriza a Inferência Estatística é o fato de que as probabilidades de erros serem cometidos são conhecidas *a priori*; isto é, sabemos qual o risco que estamos correndo, antes de tomarmos uma decisão.

Num teste de hipótese, existem quatro situações possíveis na tomada de decisão: em duas delas as decisões são corretas, nas outras duas elas são erradas. O Quadro **1** mostra essas situações.

Quadro 1. Resultados possíveis de um teste

		Hipótese nula	
		verdadeira	falsa
Decisão	rejeitar Ho	**Erro tipo I**	ok
	aceitar Ho	ok	**Erro tipo II**

Há portanto dois tipos de erros:

Erro tipo I: rejeitar uma hipótese quando ela é verdadeira;
Erro tipo II: aceitar uma hipótese quando ela é falsa.

É preciso calcularmos as probabilidades de cometer estes erros, antes de tomar qualquer decisão (se estas probabilidades forem altas, será melhor não tomar decisão nenhuma!). Estas probabilidades são designadas convencionalmente pelas letras gregas α (alfa) e β (beta):

$P(\text{erro I}) = \alpha$
$P(\text{erro II}) = \beta$

(a) Erro Tipo I

Uma tachinha *equilibrada* (isto é, uma tachinha para a qual a hipótese nula $\pi=0,5$ é verdadeira) tem uma probabilidade, ainda que pequena, de produzir menos de 6 ou mais de 14 caras. Se isto acontecer, ela será rejeitada no teste, e considerada *desequilibrada*. O erro cometido quando rejeitamos uma tachinha equilibrada é o erro *Tipo I*, cuja probabilidade é representada pela letra α.

No exemplo da seção anterior, qual a probabilidade de isto ocorrer? A Tab. **1** mostra as probabilidades associadas a 20 lançamentos de uma tachinha equilibrada; a probabilidade de $X<6$ ou $X>14$ é a soma de todas as probabilidades marcadas em vermelho; esta é a probabilidade da região de rejeição, igual a 0,0414.

Portanto, neste teste,

$P(\text{erro I}) = \alpha = 0,0414$

Esta probabilidade é chamada de *nível de significância* do teste, o que mostra que os dois tipos de teste – o de *significância* e o de *hipótese* – são relacionados entre si. Lembre-se de que, no teste de significância, um resultado é considerado *significativo* se o seu valor-*p* fosse menor que um valor crítico escolhido, geralmente de α =0,05; dizemos então que há uma forte evidência contra a hipótese nula. Num teste de hipótese, um valor-*p* menor que o α escolhido indica que a variável de teste caiu na região de rejeição, e que portanto a hipótese nula deve ser rejeitada. Um resultado *significativo* num teste, portanto, indica que a hipótese nula foi *rejeitada*.

O nível de significância de um teste é sempre *escolhido* por quem planejou o teste. No exemplo, usamos o valor pouco comum de α = 0,0414, porque é o valor mais próximo de 0,05 que pode ser obtido a partir das probabilidades na Tab. **2**. Poderíamos ter escolhido um tamanho diferente para a região de rejeição; poderíamos, por exemplo, ter decidido que a tachinha só seria rejeitada se caísse com a ponta para baixo menos de 5 ou mais que 15 vezes. A Tab. **2** mostra como seriam então as regiões de aceitação e rejeição.

Tabela 2. Distribuição de probabilidades da variável
***X = número de "pontas para cima"*,**
em 20 lançamentos de uma tachinha equilibrada
(região de rejeição reduzida)

X	p(x)	
00	.0000	<----⎤
01	.0000	+-- Região de rejeição (P = 0,0059)
02	.0002	⎥
03	.0011	⎥
04	.0046	<----⎦
05	.0148	<----+
06	.0360	⎥
07	.0739	⎥
08	.1201	⎥
09	.1602	⎥
10	.1762	+-- Região de aceitação (P = 0,9882)
11	.1602	⎥
12	.1201	⎥
13	.0739	⎥
14	.0360	⎥
15	.0148	<----+
16	.0046	<----⎤
17	.0011	⎥
18	.0002	⎥
19	.0000	+-- Região de rejeição (P = 0,0059)
20	.0000	<----⎦

O valor de α seria então:

$$P(X< 5) = 0,0000 + 0,0000 + 0,0002 + 0,0011 + 0,0046 = 0,0059$$
$$P(X>15) = 0,0000 + 0,0000 + 0,0002 + 0,0011 + 0,0046 = 0,0059$$
$$\alpha = P(X<5) + P(X>15) = 0,0059 + 0,0059 = 0,0118$$

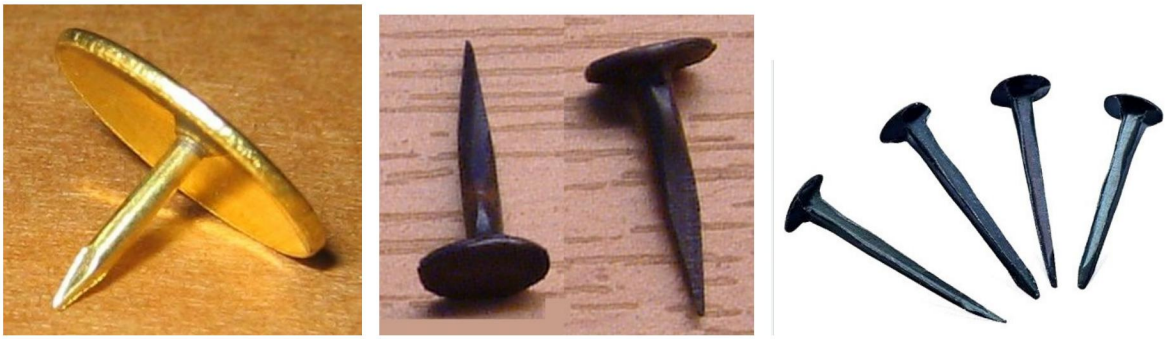
O risco de cometermos o erro do Tipo I, portanto, seria bem menor - apenas cerca de 1%, contra os 4% do teste feito anteriomente. O problema, porém, é que quando alteramos o tamanho da região de *rejeição*, alteramos também o da região de *aceitação*, e em

consequência, também a probabilidade β; esta probabilidade é mais difícil de calcular, porque é variável.

(b) Erro Tipo II

No exemplo, β é a probabilidade de aceitarmos H_0 quando ela é falsa; ou seja, de considerarmos *equilibrada* uma tachinha que na verdade é *desequilibrada*. Qual a probabilidade de isto acontecer? Isto vai depender do tamanho da região de rejeição, e de quão desequilibrada a tachinha é na verdade (ou seja, de qual é o valor verdadeiro de π). Se a tachinha for *quase* equilibrada (por exemplo, π=0,55), é altamente provável que ela caia em torno de 10 vezes com a ponta para cima e que H_0 seja aceita (o valor mais provável seria X=11). Contudo, se a tachinha for *muito* desequilibrada (por exemplo, π=0,05), é altamente provável que ela caia com a ponta para cima apenas uma ou duas vezes e que H_0 seja rejeitada (o valor mais provável seria X=1).

A Fig. **1** mostra três tipos de tachas que, em princípio, podemos considerar que devem ter diferentes probabilidades de cair com a ponta para cima. A da Fig. **1A** é bem semelhante à tacha usada nos exemplos feitos acima, e provavelmente passaria no teste, e seria considerada equilibrada. As da Fig. **1B** talvez consigam cair com a ponta para cima entre 6 e 14 vezes, e passem no teste; as da Fig. **1C**, porém, parecem ser muito desequilibradas, e deve ser altamente improvável que uma tacha destas seja considerada equilibrada. O problema é que o valor real de π para cada uma destas tachas é obviamente desconhecido (se fosse conhecido, não seria preciso fazer testes estatísticos!).



A B C

Figura 1. Diferentes tipos de tachas

O que podemos fazer é calcular a probabilidade β como função do valor verdadeiro de π e do tamanho da amostra. Na verdade, o que normalmente se calcula é o valor (1-β), que é a probabilidade de o teste *não errar*; isto é, de não aceitar a hipótese H_0 quando ela for falsa. Esta probabilidade é chamada de *poder* do teste.

Se você está achando isto tudo muito complicado, está em boa companhia; toda a teoria de Inferência Estatística e de decisão baseada em probabilidades é muito complexa, e gera muita discussão. Na verdade, a maioria dos pesquisadores não se interessa muito por isso, porque preferem fazer apenas testes de significância; o que procuram é o valor-*p*, que lhes indica quão forte é a evidência a favor ou contra a hipótese que querem testar. Como indicado no exemplo sobre testes da proporção de peças defeituosas mencionado antes, o cálculo do poder de um teste geralmente interessa aos engenheiros envolvidos em controle de qualidade; esta teoria aliás foi criada originalmente por engenheiros dos laboratórios Bell, em 1929, e os dois tipos de erros eram chamados de *risco do produtor* (risco do pro-

dutor ter seu produto rejeitado erroneamente) e *risco do consumidor* (risco do consumidor aceitar um produto que não atende às especificações).

Para os pesquisadores de outras áreas, o conceito de poder de um teste é frequentemente útil quando é preciso calcular o tamanho mínimo necessário para uma amostra; quanto maior o tamanho da amostra, maior será o poder do teste. Voltando ao teste das tachinhas: se por exemplo P(ponta para cima)=0,51 e P(ponta para baixo)=0,49, precisaríamos de fazer mais de 10 mil lançamentos para conseguir detectar, a partir da amostra, qual dos dois resultados é o de maior probabilidade. Se por outro lado P(ponta para cima) =0,95, enquanto P(ponta para baixo)=0,05, a diferença entre as duas é tão grande que mesmo uma minúscula amostra de n=20 é capaz de detectá-la. O *poder*, portanto, indica qual é diferença mínima entre as duas probabilidades que o teste seria capaz de detectar, e depende principalmente do tamanho da amostra.

4.3.2.3. Teste de hipótese unilateral

O teste que fizemos acima é um teste *bilateral*: a região de rejeição é composta por duas partes, cada uma situada num dos extremos da distribuição de probabilidades, e aceitamos H_1 se o valor de X encontrado for muito mais alto ou muito mais baixo do que seria esperado se H_0 fosse verdadeira. Há problemas, porém, em que o que nos interessa é testar valores de X que caem em apenas um dos extremos da distribuição, o que gera testes *unilaterais*.

Isto pode ocorrer, por exemplo, nos testes já mencionados de controle de qualidade. Suponha que um fabricante produz e vende um tipo de peça, 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. Este teste geralmente será *unilateral*: se mostrar que a proporção de peças defeituosas é *maior* que 3%, o comprador recusará o lote; se, por outro lado, mostrar que a proporção é *igual* ou *menor* que 3%, o comprador irá aceitá-lo. Para o comprador, não haverá nenhum prejuízo em aceitar lotes nos quais a proporção de peças defeituosas é *menor* do que a afirmada pelo fabricante, ou seja, lotes cuja qualidade é *melhor* do que ele esperava.

Para dar um exemplo simples, voltemos ao problema de 20 lançamentos de dados assimétricos. Suponhamos que ao invés de testar uma tachinha comum, queiramos testar uma tacha de sapateiro, como a da Fig. **1B**. A julgar pela forma desta tacha, julgamos que ela seja muito desequilibrada, e que a probabilidade de ela cair com a ponta para cima seja *menor* do que com a ponta para baixo; faremos por isso um teste unilateral. As hipóteses nula e alternativa podem ser escritas como (π é a probabilidade de a tacha cair com a ponta para cima):

$$H_0: \quad \pi = 0{,}5$$
$$H_1: \quad \pi < 0{,}5$$

Alguns livros preferem escrever a hipótese nula de forma um pouco diferente:

$$H_0: \quad \pi \geq 0{,}5$$
$$H_1: \quad \pi < 0{,}5$$

Esta forma é mais lógica, porque indica que a H_0 abrange não apenas o valor π=0,5 mas também os valores de π>0,5. Isto quer dizer que, se a tachinha for equilibrada, ou desequilibrada para o outro lado (maior probabilidade de cair com a ponta para cima), a H_0 será aceita. Este tipo de H_0 é chamado de *hipótese composta*. De qualquer forma, a distri-

buição será calculada no exemplo por meio de um modelo binomial que usa π=0,5 como a probabilidade de sucessos *p*; a desigualdade π > 0,5 na hipótese nula serve apenas para fazer que as duas hipóteses H_0 e H_1 sejam logicamente complementares. O procedimento de teste é o mesmo, se H_0 : π ≥ 0,5 ou se H_0 : π = 0,5.

A distribuição de probabilidades é a da Tab. **3**. Se o número X de vezes em que a tacha cair com a ponta para cima for menor que 10, iremos verificar se X se encontra ou não na região de rejeição de H_0. Se X for maior ou igual que 10, porém, H_0 será imediatamente aceita.

Tabela 3. Distribuição de probabilidades da variável
X = número de "pontas para cima"
em 20 lançamentos de uma tachinha equilibrada - Teste unilateral à esquerda

X	p(x)		
00	.0000	<---┐	Se a probabilidade de a tachinha cair com a ponta
01	.0000	+--	para cima for π ≥ 0,5 é pouco provável encontrarmos
02	.0002	|	um destes valores de X na amostra. A probabilidade
03	.0011	|	acumulada destes valores é de P= 0,0207.
04	.0046	|	Região de rejeição da hipótese nula **(α = 0.0207)**
05	.0148	<---┘	
06	.0360	<---┐	
07	.0739	|	
08	.1201	|	
09	.1602	|	Se a prob. de a tachinha cair com a ponta para cima
10	.1762	+--	for π ≥ 0,5 é mais provável que eu encontre um destes
11	.1602	|	valores na amostra. A probabilidade acumulada destes
12	.1201	|	valores é de P=0,9793.
13	.0739	|	Região de aceitação da hipótese nula **(1-α = 0,9793)**
14	.0360	|	
15	.0148	|	
16	.0046	|	
17	.0011	|	
18	.0002	|	
19	.0000	|	
20	.0000	<---┘	

O teste também poderia ser feito no outro extremo da curva. Poderíamos por exemplo supor que a P(ponta para cima) seja *maior* que a P(ponta para baixo); neste caso, as hipóteses seriam escritas como:

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

A região de rejeição seria então colocado no extremo superior da distribuição; se mantivermos o mesmo valor α = 0,0207, a região de rejeição conteria então os valores de X≥15.

4.3.2.4. Resumo

Esta é, em síntese, a idéia básica de qualquer teste de hipóteses: escolher uma variável de teste, estudar quais são os valores mais prováveis de seu domínio se a hipótese nula for verdadeira (estes valores formarão a *região de aceitação*) e quais os menos prováveis (estes formarão a *região de rejeição*), retirar uma amostra, calcular o valor assumido pela variável de teste, ver em qual região este valor caiu, e tomar a decisão.

Explicando passo a passo:

a. Escolher uma *variável de teste* X
No exemplo, usamos X = número de vezes em que a tachinha cai com a ponta para cima em 20 lançamentos. (Poderíamos também ter usado a proporção P=X/20 de vezes em que a tachinha cai com a ponta para cima, ou a porcentagem P% = 100×X/n).

b. Decidir se o teste será unilateral ou bilateral, e escolher o nível de significância α.
No exemplo, α=0,05, que é o valor mais usado na prática.

c. Estabelecer a hipótese nula (H_0) e a hipótese alternativa (H_1).
No exemplo: para um teste *bilateral*, a hipótese H_0 é a de que a tachinha é equilibrada, e que
P(ponta para cima) = P(ponta para baixo);
a hipótese H_1 é que estas duas probabilidades são diferentes.
Representando P(ponta para cima) por π, podemos escrever as hipóteses como:

H_0 : $\pi = 0{,}5$
H_1 : $\pi \neq 0{,}5$

Para um teste *unilateral*, se a hipótese que queremos testar é a de
P(ponta para cima) < P(ponta para baixo),
a hipótese H_0 deverá ser:
P(ponta para cima) ≥ P(ponta para baixo);
e estas hipóteses podem ser escritas como:

H_0 : $\pi \geq 0{,}5$
H_1 : $\pi < 0{,}5$

Se porém a hipótese que nos interessa é a de que
P(ponta para cima) > P(ponta para baixo),
a H_0 deverá ser:
P(ponta para cima) ≤ P(ponta para baixo),
e estas hipóteses serão escritas como:

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

d. Delimitar as regiões de aceitação e de rejeição; isto é, verificar quais serão os valores mais prováveis de X na amostra, se H_0 for verdadeira.
No exemplo. se a tachinha for equilibrada, X terá uma distribuição binomial de n=20 e p=0,5; a partir desta distribuição podemos delimitar as regiões de aceitação e rejeição)

e. Criar uma regra de decisão.
No exemplo: se o número estiver dentro da *região de rejeição*, decido *rejeitar a hipótese* de que a tachinha seja equilibrada, e considerá-la *desequilibrada*.

f. Retirar a amostra e calcular o valor de X
No exemplo: lançar a tachinha 20 vezes, e contar o número de vezes que ela cai com a ponta para cima)

g. Verificar se o valor de X está na região de aceitação ou na de rejeição;

h. Aceitar ou rejeitar H_0.

A importante notar que escolhas como a do tamanho da amostra, do tipo de teste (unilateral × bilateral) e do valor de α, têm que ser feitas *antes* de que a amostra seja retirada! Modificar estas escolhas depois que a amostra já foi analisada, para forçar a aceitação (ou rejeição) de uma hipótese que interessa (ou não interessa) é uma prática considerada uma forma de fraude (que às vezes acontece nas publicações, mas com certeza é mal vista).

O teste que fizemos, usando tachinhas, pode parecer um divertimento acadêmico, sem muito interesse no mundo real. Isto em parte verdade; testes de proporções feitos com amostras de n=20 não têm muita utilidade, porque seu poder é muito pequeno; foram usados aqui como exemplo porque a matemática envolvida é muito simples. Os testes de proporções usados na prática exigem amostras muito maiores, geralmente de milhares de elementos. Para analisar dados destas amostras não usaremos a distribuição binomial, mas sim a sua aproximação pela gaussiana, como veremos na Seção 4.4.