4.3.3. Poder de um teste de hipótese

Como visto na Seção 4.3.2.2, quando tomamos a decisão num teste de hipótese, corremos sempre o risco de estar cometendo um erro. Há dois tipos de erros. Em resumo:

Erro Tipo I

O erro do tipo I ocorre quando se *rejeita* uma H_0 verdadeira. A probabilidade de ocorrência deste tipo de erro, representada por α (alfa), é igual à probabilidade de a variável de teste assumir valores dentro da região de rejeição, quando H_0 for verdadeira. Esta probabilidade pode ser reduzida pela diminuição da região de rejeição; contudo, isto aumenta a probabilidade de que seja cometido um erro do Tipo II.

Erro Tipo II

O erro do Tipo II ocorre quando se *aceita* uma H_0 falsa. A probabilidade de ocorrer este tipo de erro é representada por ß (beta); a probabilidade (1-β) de *não* ocorrer este tipo de erro é chamada de *poder* do teste. Estas probabilidades dependem do valor real (obviamente desconhecido) do parâmetro que se está testando (π, no exemplo), e do tamanho da amostra.

Antes de se tomar qualquer decisão, é necessário portanto estudar as probabilidades destes erros; não faz sentido tomarmos decisões que têm uma grande probabilidade de estarem erradas.

No exemplo do teste da tachinha, se usarmos as regiões de rejeição e aceitação como definidas na Tabela **2** da seção 4.3.2.1, a probabilidade do erro Tipo I será o somatório de todas as probabilidades para $X < 6$ ou $X > 14$, no caso de a tachinha ser equilibrada ($\pi = 0{,}5$); valerá α=0,0414. A probabilidade do erro Tipo II será a probabilidade de uma tachinha desequilibrada produzir um valor de X entre $6 \leq X \leq 14$; esta probabilidade dependerá evidentemente do valor real de π. A variável X segue um modelo binomial, cujos parâmetros são π e o tamanho da amostra n=20, $X \sim B(20, \pi)$. As probabilidades de X para cada valor de π podem portanto ser calculadas por:

$$p(x) = C_n^x \pi^x (1-\pi)^{(n-x)}$$

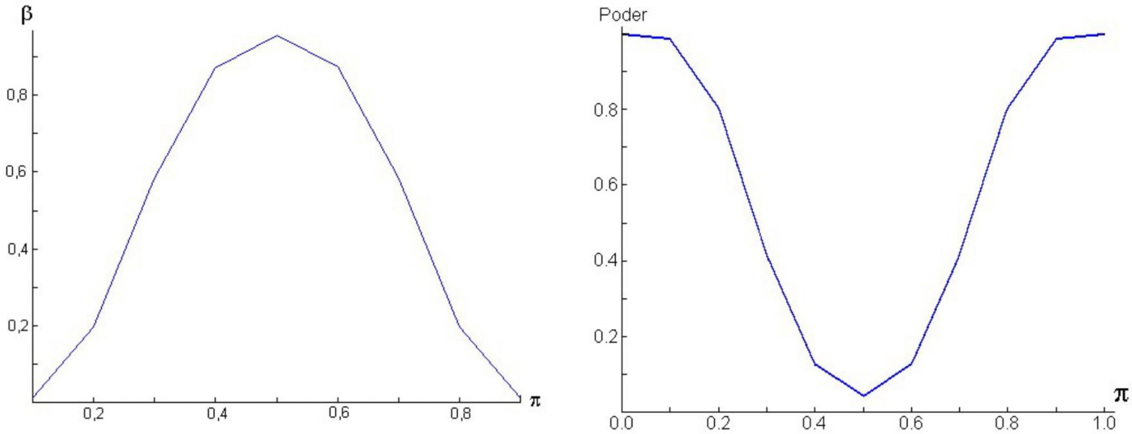
A Tab. **1** mostra as probabilidades de X para diferentes valores de π. Como foi definido antes, a região de aceitação abrange os valores de $6 \leq X \leq 14$ (marcada em azul na tabela); se X cair neste intervalo, a H_0 será aceita e a tachinha será considerada equilibrada. Se uma tachinha tiver π=0,01, as probabilidades associadas aos diferentes valores de X são dadas na primeira coluna. Neste caso, a probabilidade de X cair na região de aceitação será muito pequena (igual ao somatório das probabilidades em azul na tabela); a probabilidade de cometermos o erro do Tipo I (aceitarmos H_0 e concluirmos que a tachinha é equilibrada) será de apenas β=0,0114. O teste portanto tem grande *poder* neste caso; é altamente improvável que tomemos uma decisão errada.

Por outro lado, se uma tachinha tiver π=0,40, a probabilidade de X cair na região de aceitação será muito grande. A probabilidade de cometermos um erro do Tipo I será de β=0,8724; teremos portanto uma grande probabilidade de tomarmos uma decisão errada, concluindo que a tachinha é equilibrada. O teste, neste caso, tem muito pouco poder.

Tabela 1 – Distribuições binomiais para diferentes n=20 e valores de π

X	X/n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
		valores de π									
00	.00	.1216	.0115	.0008	.0000						
01	.05	.2702	.0576	.0068	.0005	.0000					
02	.10	.2852	.1369	.0278	.0031	.0002					região de
03	.15	.1901	.2054	.0716	.0123	.0011	.0000				rejeição
04	.20	.0898	.2182	.1304	.0350	.0046	.0003				
05	.25	.0319	.1746	.1789	.0746	.0148	.0013	.0000			
06	.30	.0089	.1091	.1916	.1244	.0360	.0049	.0002			
07	.35	.0020	.0545	.1643	.1659	.0739	.0146	.0010	.0000		
08	.40	.0004	.0222	.1144	.1797	.1201	.0355	.0039	.0001		
09	.45	.0001	.0074	.0654	.1597	.1602	.0710	.0120	.0005		
10	.50	.0000	.0020	.0308	.1171	.1762	.1171	.0308	.0020	.0000	região de
11	.55		.0005	.0120	.0710	.1602	.1597	.0654	.0074	.0001	aceitação
12	.60		.0001	.0039	.0355	.1201	.1797	.1144	.0222	.0004	
13	.65		.0000	.0010	.0149	.0739	.1659	.1643	.0545	.0020	
14	.70			.0002	.0049	.0360	.1244	.1916	.1091	.0089	
15	.75			.0000	.0013	.0148	.0746	.1789	.1746	.0319	
16	.80				.0003	.0046	.0350	.1304	.2182	.0898	
17	.85				.0000	.0011	.0123	.0716	.2054	.1901	região de
18	.90					.0002	.0031	.0278	.1369	.2852	rejeição
19	.95					.0000	.0005	.0068	.0576	.2702	
20	1.00						.0000	.0008	.0115	.1216	
	β	.0114	.1958	.5836	.8723		.8723	.5836	.1958	.0114	
	1-β	.9886	.8042	.4164	.1274		.1274	.4164	.8042	.9886	

O gráfico que mostra a variação de ß em relação ao valor real do parâmetro π é denominado "curva característica de operação do teste" (Fig. **1A**). Um gráfico usado com mais frequência porém é o gráfico de (1-β), na Fig. **1B**, que mostra a variação do poder do teste, para este tamanho de amostra (n=20).



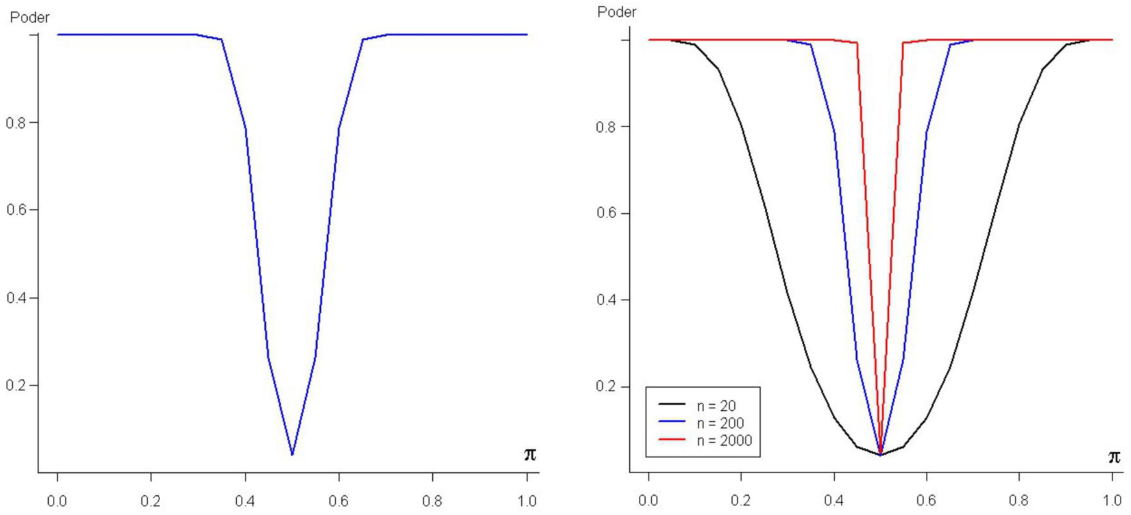
A. Curva característica de operação B. Poder

Figura 1. Curva característica de operação e poder de um teste binomial, para n=20

É possível ver que este teste tem pouco poder, a não ser que a tachinha seja muito desequilibrada. Suponha por exemplo uma eleição onde haja dois candidatos, e queremos testar qual dois tem a maioria dos votos. Mesmo que a diferença entre eles for razoavelmente grande, um teste com uma amostra de n=20 seria praticamente inútil. Se por exemplo um tem 30% e o outro 70% (π=0.3), o poder do teste seria de apenas 1-β = 0.4164, e a probabilidade de concluirmos que eles estão empatados seria de β = 0.5836.

Se usarmos uma amostra de n=200, o curva do poder do teste seria a da Fig. 2A. Se um candidato tem 30% dos votos, a probabilidade de concluirmos que eles estão empatados seria de apenas β = 0,0001; ou seja, este teste teria um poder de 1 - β = 0,9999

No entanto, este teste ainda não é bom o suficiente. Se um dos candidatos tem 40% (π=0.4), a probabilidade de concluirmos que eles estão empatados seria de β=0,2131; o poder do teste seria de 1 - β=0,7868, o que ainda é pouco.



A. Teste com n=200 B. Comparando n=20, n=200, n=2000

Figura 2. Poder de um teste binomial

Por fim, a Fig. 2B compara os poderes de testes feitos com três tamanhos de amostra diferentes, n=20, 200 e 2000. No gráfico, podemos ver que quanto maior o tamanho da amostra, maior o poder do teste. Se um dos candidatos tiver 40% dos votos, uma amostra de n=2000, a probabilidade de o teste errar, e considerar que eles estão empatados, será praticamente nula. Contudo, este teste também pode errar, se a diferença entre os candidatos for muito pequena. Por exemplo, se um dos candidatos tiver 48% dos votos, e o outro 52%, a probabilidade do erro seria de β = 0,5796.

Idealmente, portanto, deveríamos tentar usar sempre o teste de maior poder, e a maior amostra possível. O tamanho da amostra contudo é sempre limitado pela quantidade de recursos disponíveis (material, colaboradores, dinheiro) e do tempo disponível para fazer o trabalho. Na prática, o tamanho escolhido é aquele que permite que o teste tenha um poder aceitável, e possa ser feito dentro dos recursos de que os pesquisadores dispõem.

Resumo

1. O *poder* de um teste é a probabilidade de ele não cometer o erro do Tipo II.
2. Para um dado teste, quanto maior a amostra, maior o poder;
3. O poder não é uma função linear do tamanho da amostra; dobrar o tamanho da amostra não implica em dobrar o poder do teste.
4. Idealmente, deveríamos tentar usar a maior amostra possível, e o teste de maior poder; isto nem sempre é possível, pois o tamanho da amostra é limitado pela quantidade de recursos disponíveis.