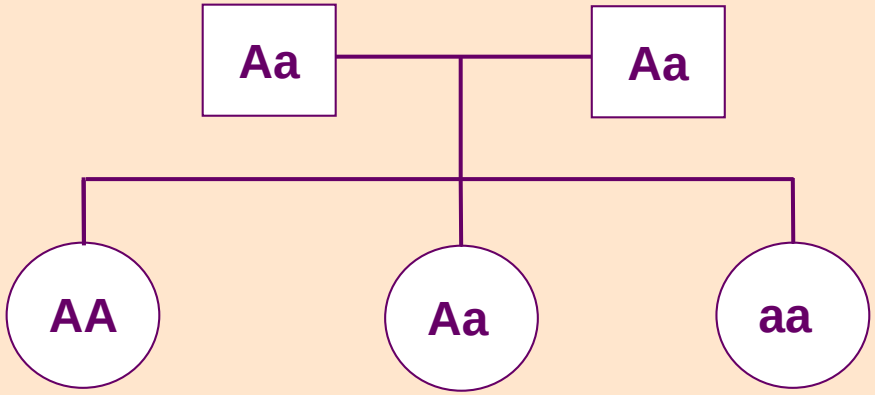
Testes Qui-quadrado

- **Testes de Aderência**
- **Testes de Independência**
- **Testes de Homogeneidade**

1. Testes de Aderência

Objetivo: Testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados

Exemplo 1: Genética – Equilíbrio Hardy-Weinberg



Probabilidades: ¼ ½ ¼
(Modelo teórico)

⇒ **3 categorias: AA, Aa, aa**

Em uma certa população, 100 descendentes foram estudados, fornecendo a tabela a seguir:

Genótipo	Freqüência Observada
AA	26
Aa	45
aa	29
Total	100

Objetivo: Verificar se o modelo genético proposto é adequado para essa população

Se o modelo Hardy-Weinberg for adequado, a freqüência *esperada* de descendentes para o genótipo AA, dentre os 100 indivíduos, pode ser calculada por:

$$100 \times P(\mathrm{AA}) = 100 \times \tfrac{1}{4} = 25$$

Da mesma forma, temos para o genótipo Aa,

$$100 \times P(\mathrm{Aa}) = 100 \times \tfrac{1}{2} = 50$$

E para o genótipo aa,

$$100 \times P(\mathrm{aa}) = 100 \times \tfrac{1}{4} = 25$$

Podemos expandir a tabela de freqüências dada anteriormente:

Genótipo	Freqüência Observada	Freqüência Esperada
AA	26	25
Aa	45	50
aa	29	25
Total	100	100

Podemos afirmar que os valores observados estão suficientemente próximos dos valores esperados, de tal forma que o modelo Hardy-Weinberg é adequado a esta população?

1. Testes de Aderência – Metodologia

Considere uma tabela de freqüências com $k \geq 2$ categorias de resultados:

Categorias	Freqüência Observada
1	O_1
2	O_2
3	O_3
⋮	⋮
k	O_k
Total	n

em que O_i é o total de indivíduos *observados* na categoria *i*, *i* = 1,...,*k*.

Seja p_i a probabilidade associada à categoria *i* , *i=1,...,k*.

O objetivo do teste de aderência é testar as hipóteses

$$H : p_1 = p_{o1} , \quad .... , p_k = p_{ok}$$

A : existe pelo menos uma diferença

sendo p_{oi} a probabilidade associada à categoria *i* , *i* = 1,...,*k*, calculada através do modelo probabilístico de interesse.

Se E_i é o total de indivíduos *esperados* na categoria *i* quando a hipótese H é verdadeira, então:

$$E_i = n \times p_{oi}, \qquad i = 1,\ldots,k$$

Expandindo a tabela de freqüências original, temos

Categorias	Freqüência Observada	Freqüência Esperada
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
⋮	⋮	⋮
k	O_k	E_k
Total	n	n

Quantificação da distância entre as colunas de freqüências:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$ ← **Estatística do teste de aderência.**

Supondo H verdadeira,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_q,$$ **aproximadamente,**

sendo que *q* = *k* – 1 representa o número de graus de liberdade.

Em outras palavras, se *H* é verdadeira, a v.a. χ^2 tem distribuição aproximada qui-quadrado com *q* graus de liberdade.

Obs.: Este resultado é válido para *n grande* e para

$$E_i \geq 5,\ i = 1,\ldots,k\ .$$

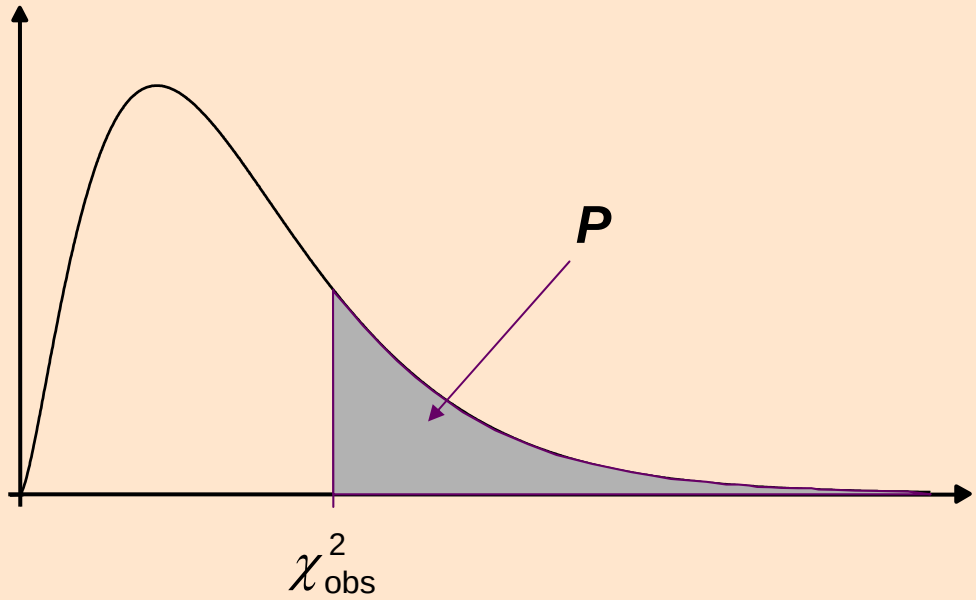
Regra de decisão:

Pode ser baseada no nível descritivo *P*, neste caso

$$P = P(\chi^2_q \geq \chi^2_{\text{obs}}),$$

em que χ^2_{obs} é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2.

Graficamente:



Se, para α fixado, obtemos $P \leq \alpha$, rejeitamos a hipótese H.

Exemplo (continuação): Genética – Equilíbrio Hardy-Weinberg:

Hipóteses:

H: O modelo proposto é adequado a esta situação
A: O modelo não é adequado a esta situação

De forma equivalente, podemos escrever:

H: $P(AA) = \frac{1}{4}$ e $P(Aa) = \frac{1}{2}$ e $P(aa) = \frac{1}{4}$

A: ao menos uma das igualdades não se verifica

A tabela seguinte apresenta os valores observados e esperados (calculados anteriormente).

Genótipo	O_i	E_i
AA	26	25
Aa	45	50
aa	29	25
Total	100	100

Cálculo do valor da estatística do teste ($k = 3$):

$$\chi^2_{obs} = \sum_{1}^{3} \frac{(Oi - Ei)^2}{Ei} = \frac{(26-25)^2}{25} + \frac{(45-50)^2}{50} + \frac{(29-25)^2}{25} =$$

$$= 0{,}04 + 0{,}50 + 0{,}64 = 1{,}18 \ .$$

Usando a distribuição de qui-quadrado com q = k-1 = 2 graus de liberdade,

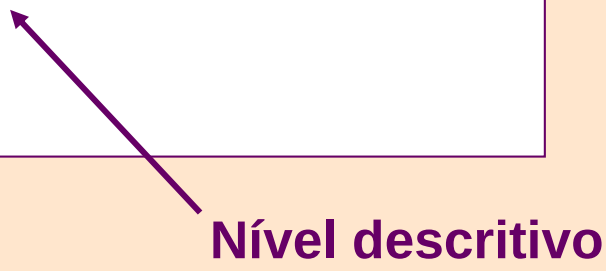
$$P = P(\chi^2_2 \geq 1{,}18) = 0{,}5543.$$

Conclusão: Seja α = 0,05. Como P = 0,5543 > 0,05, não rejeitamos a hipótese H, isto é, essa população segue o equilíbrio Hardy-Weinberg.

O cálculo do *nível descritivo P* pode ser feito no MINITAB, através dos comandos:

```
MTB > cdf 1.18 k1;
SUBC> chisquare 2.
MTB > let k2 = 1 - k1
MTB > print k2

Data Display
K2    0.554327
MTB >
```



Exemplo 2 : **Deseja-se verificar se o número de acidentes em uma estrada muda conforme o dia da semana. O número de acidentes observado para cada dia de uma semana escolhida aleatoriamente foram:**

Dia da semana	No. de acidentes
Seg	20
Ter	10
Qua	10
Qui	15
Sex	30
Sab	20
Dom	35

O que pode ser dito?

Hipóteses a serem testadas:

H: O número de acidentes não muda conforme o dia da semana;
A: Pelo menos um dos dias tem número diferente dos demais.

Se p_i representa a probabilidade de ocorrência de acidentes no *i*-ésimo dia da semana,

H: $p_i = 1/7$ para todo $I = 1,\ldots, 7$
A: $p_i \neq 1/7$ para pelo menos um valor de *i*.

Total de acidentes na semana: $n = 140$.

Logo, se H for verdadeira,

$$\Rightarrow E_i = 140 \times 1/7 = 20,\ i = 1,\ldots,7.$$

Dia da semana	No. de acidentes observados (O_i)	No. esperado de acidentes (E_i)
Seg	20	20
Ter	10	20
Qua	10	20
Qui	15	20
Sex	30	20
Sab	20	20
Dom	35	20

Cálculo da estatística de qui-quadrado:

$$\chi^2_{obs} = \sum_{1}^{7} \frac{(Oi - Ei)^2}{Ei} = \frac{(20-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(15-20)^2}{20} +$$

$$\frac{(30-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(35-20)^2}{20} = 27{,}50 \cdot$$

Neste caso, temos $\chi^2 \sim \chi^2_6$**, aproximadamente.**

O nível descritivo é dado por

$$P = P(\chi^2_6 \geq 27{,}50)$$

e pode ser obtido no MINITAB conforme indicado a seguir:

```
MTB > cdf 27.50 k1;
SUBC> chisquare 6.
MTB > let k2 = 1 - k1
MTB > print k2

Data Display
K2    0.000116680
```

Logo, para α **= 0,05, segue que** ***P*** **= 0,001 <** α **e assim rejeitamos H, e concluímos que o número de acidentes não é o mesmo em todos os dias da semana.**

2. Testes de Independência

Objetivo: Verificar se existe independência entre duas variáveis medidas nas mesmas unidades experimentais.

Exemplo 3: Deseja-se verificar se existe dependência entre a renda e o número de filhos em famílias de uma cidade.

- 250 famílias escolhidas ao acaso forneceram a tabela a seguir:

Renda (R$)	Número de filhos				
	0	1	2	+ de 2	Total
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
5000 ou mais	8	13	9	10	40
Total	48	70	71	61	250

Em geral, os dados referem-se a mensurações de duas características (A e B) feitas em *n* unidades experimentais, que são apresentadas conforme a seguinte tabela:

A \ B	B_1	B_2	…	B_s	Total
A_1	n_{11}	n_{12}	…	n_{1s}	$n_{1\bullet}$
A_2	n_{21}	n_{2s}	…	n_{2s}	$n_{2\bullet}$
…	…	…	…	…	…
A_r	n_{r1}	n_{r2}	…	n_{rs}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet s}$	$n_{\bullet\bullet}$

Hipóteses a serem testadas – Teste de independência:

H: *A* e *B* são variáveis independentes

A: As variáveis *A* e *B* não são independentes

Quantas observações devemos ter em cada casela se *A* e *B* forem independentes?

Se *A* e *B* forem independentes, temos que, para todos os possíveis (A_i e B_j):

$$P(A_i \cap B_j) = P(A_i) \times P(B_j) \qquad i = 1, 2,\ldots, r \quad \text{e} \quad j = 1, 2,\ldots,s.$$

Logo, o *número esperado* de observações com as características (A_i e B_j) entre as $n_{..}$ observações sob a hipótese de independência, é dado por

$$E_{ij} = n_{..} \times p_{ij} = n_{..} \times p_{i.} \times p_{.j} = n_{..} \times \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}}$$

sendo p_{ij} a proporção de observações com as características (A_i e B_j).

Assim,

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

O processo deve ser repetido para todas as caselas (*ij*).

Distância entre os valores observados e os valores esperados sob a suposição de independância:

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Estatística do teste de independência

em que $O_{ij} = n_{ij}$ representa o total de observações na casela (*ij*).

Supondo H verdadeira,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_q$$

sendo $q = (r - 1) \times (s - 1)$ graus de liberdade.

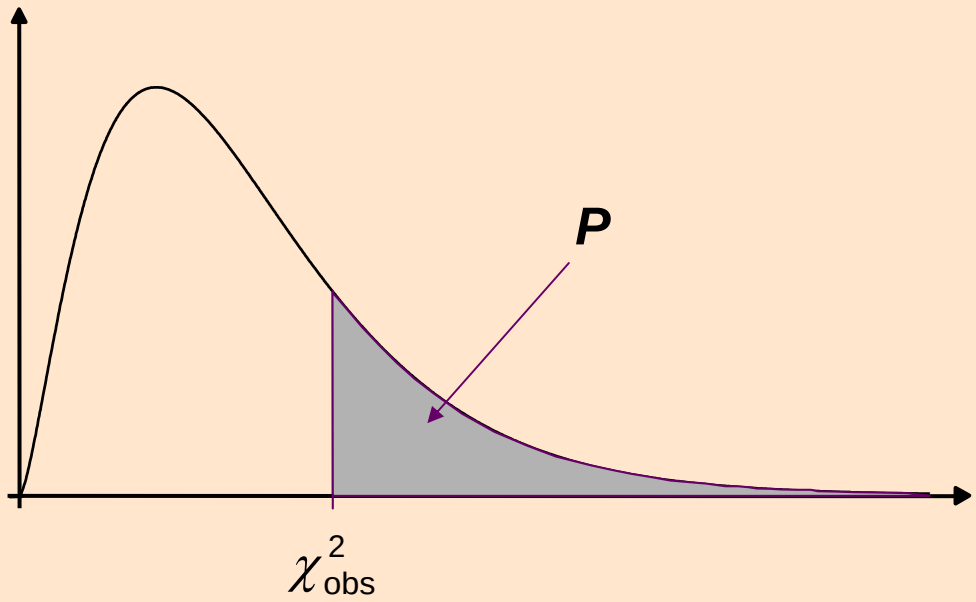
Regra de decisão:

Pode ser baseada no nível descritivo *P*, neste caso

$$P = P(\chi^2_q \geq \chi^2_{\text{obs}})$$

em que χ^2_{obs} é o valor calculado, a partir dos dados, para a estatística do teste.

Graficamente:



Se, para α fixado obtemos $P \leq \alpha$, rejeitamos a hipótese de independência.

Exemplo (continuação):
Estudo da dependência entre renda e o número de filhos
• 250 famílias foram escolhidas ao acaso

Hipóteses **H: O número de filhos e a renda são independentes**
A: Existe dependência entre o número de filhos e a renda

Renda (R$)	Número de filhos				
	0	1	2	+ de 2	Total
menos de 2000	15	27	50	43	135
2000 a 5000	25	30	12	8	75
5000 ou mais	8	13	9	10	40
Total	48	70	71	61	250

Exemplo do cálculo dos valores esperados sob H (independência):

• Número esperado de famílias sem filhos e renda menor que R$ 2000:

$$E_{11} = \frac{48 \times 135}{250} = 25{,}92 \ .$$

Tabela de valores observados e esperados (entre parênteses)

Renda (R$)	Número de filhos				
	0	1	2	+ de 2	Total
menos de 2000	15(25,92)	27(37,80)	50(38,34)	43(32,94)	135
2000 a 5000	25(14,40)	30(21,00)	12(21,30)	8(18,30)	75
5000 ou mais	8(7,68)	13(11,20)	9(11,36)	10(9,76)	40
Total	48	70	71	61	250

1 filho e renda de R$ 2000 a R$ 5000:

$$E_{22} = \frac{70 \times 75}{250} = 21{,}00$$

2 ou + filhos e renda de R$ 5000 ou mais:

$$E_{34} = \frac{61 \times 40}{250} = 9{,}76$$

Lembre-se:

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}}$$

Cálculo da estatística de qui-quadrado:

Renda (R$)	Número de filhos				
	0	1	2	+ de 2	Total
menos de 2000	15(25,92)	27(37,80)	50(38,34)	43(32,94)	135
2000 a 5000	25(14,40)	30(21,00)	12(21,30)	8(18,30)	75
5000 ou mais	8(7,68)	13(11,20)	9(11,36)	10(9,76)	40
Total	48	70	71	61	250

$$\chi^2_{obs} = \frac{(15-25{,}92)^2}{25{,}92} + \frac{(25-14{,}40)^2}{14{,}40} + \frac{(8-7{,}68)^2}{7{,}68} + \frac{(27-37{,}80)^2}{37{,}80} +$$
$$+ \frac{(30-21{,}00)^2}{21{,}00} + \frac{(13-11{,}20)^2}{11{,}20} + \frac{(50-38{,}34)^2}{38{,}34} + \frac{(12-21{,}30)^2}{21{,}30} +$$
$$+ \frac{(12-21{,}30)^2}{21{,}30} + \frac{(9-11{,}36)^2}{11{,}36} + \frac{(43-32{,}94)^2}{32{,}94} + \frac{(8-18{,}30)^2}{18{,}30} +$$
$$+ \frac{(10-9{,}76)^2}{9{,}76} = 36{,}62\ .$$

Determinação do número de graus de liberdade:

- **Categorias de renda: $r = 3$**
- **Categorias de nº de filhos: $s = 4$**

$\Rightarrow$ $q = (r-1)\times(s-1) = 2\times 3 = 6$

Logo, $\chi^2 \sim \chi^2_6$ **e, supondo** $\alpha = 0{,}05$, $P = P(\chi^2_6 \geq 36{,}62) = 0{,}000$

$\therefore$ **Como P = 0,000 <** $\alpha = 0{,}05$**, rejeitamos a independência entre número de filhos e renda familiar.**

Os cálculos podem ser feitos diretamente no MINITAB:

```
Stat →Tables →Chi-Square
test
```

Saída do MINITAB:

Chi-Square Test

Expected counts are printed below observed counts

	C1	C2	C3	C4	Total
1	15	27	50	43	135
	25,92	37,80	38,34	32,94	
2	25	30	12	8	75
	14,40	21,00	21,30	18,30	
3	8	13	9	10	40
	7,68	11,20	11,36	9,76	
Total	48	70	71	61	250

Chi-Sq = 4,601 + 3,086 + 3,546 + 3,072 +
7,803 + 3,857 + 4,061 + 5,797 +
0,013 + 0,289 + 0,490 + 0,006 = 36,621

DF = 6, P-Value = 0,000

Exemplo 4: 1237 indivíduos adultos classificados segundo a pressão sangüínea (mm Hg) e o nível de colesterol (mg/100cm^3).

Verificar se existe independência entre essas variáveis.

Colesterol	Pressão			Total
	< 127	127a 166	>166	
<200	117	168	22	307
200 a 260	204	418	63	685
>260	67	145	33	245
Total	388	731	118	1237

H: Pressão sangüínea e nível de colesterol são independentes;

A: Nível de colesterol e pressão sangüínea são variáveis dependentes.

Saída do MINITAB:

```
    Chi-Square Test
Expected counts are printed below observed counts
            C1       C2       C3    Total
    1      117      168       22      307
         96,29   181,42    29,29

    2      204      418       63      685
        214,86   404,80    65,34

    3       67      145       33      245
         76,85   144,78    23,37
Total      388      731      118     1237

Chi-Sq =  4,452 +  0,993 +  1,812 +
          0,549 +  0,431 +  0,084 +
          1,262 +  0,000 +  3,967 = 13,550
DF = 4, P-Value = 0,009
```

$\therefore$ **Rejeitamos a independência entre pressão sangüínea e nível de colesterol (α = 0,05).**

3. Teste de Homogeneidade

Objetivo: Verificar se uma variável aleatória se comporta de modo similar, ou homogêneo, em várias subpopulações.

Exemplo 5: A reação ao tratamento por quimioterapia está sendo estudada em quatro grupos de pacientes com câncer.
Deseja-se investigar se todos os tipos reagem da mesma maneira.

- Uma amostra de pacientes de cada grupo foi escolhida ao acaso e classificou-se a reação em três categorias:

Câncer	Reação			
	Pouca	Média	Alta	Total
Tipo I	51	33	16	100
Tipo II	58	29	13	100
Tipo IIII	48	42	30	120
Tipo IV	26	38	16	80

Apesar da realização do teste ser semelhante a do Teste de Independência, uma distinção importante se refere à forma como as amostras são coletadas. No teste de homogeneidade fixamos o tamanho da amostra em cada uma das subpopulações e selecionamos uma amostra dentro de cada uma.

Subpopulação	valores	da	variável	Total da linha
1	O_{11}	O_{12}	...	n_1
2	O_{21}	O_{22}	...	n_2
...	...	...	...	...
Total da coluna				Total geral

Hipóteses a serem testadas – Teste de homogeneidade:

H: o comportamento da variável é homogêneo nas subpopulações

A: o comportamento da variável não é homogêneo nas subpopulações

Valores esperados (supondo homogeneidade entre as populações)

$$e_{i,j} = n_i \times \frac{total\ da\ coluna\ \ j}{total\ \ geral}$$

O total da linha ni indica o tamanho da amostra da subpopulação i e o quociente, total da coluna j dividido pelo total geral, representa a proporção de ocorrências do valor da variável correspondente à coluna j.

Caso haja homogeneidade de comportamento da variável, esperamos que essa proporção seja a mesma em todas as subpopulações.

Distância entre os valores observados e os valores esperados sob a suposição de independência:

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{r}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Estatística do teste de homogeneidade

Supondo H verdadeira,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_q$$

sendo $q = (r - 1) \times (s - 1)$ graus de liberdade.

Saída do MINITAB

Expected counts are printed below observed counts

	Pouca	Média	Alta	Total
1	51	33	16	100
	45,75	35,50	18,75	
2	58	29	13	100
	45,75	35,50	18,75	
3	48	42	30	120
	54,90	42,60	22,50	
4	26	38	16	80
	36,60	28,40	15,00	
Total	183	142	75	400

Chi-Sq = 0,602 + 0,176 + 0,403 +
3,280 + 1,190 + 1,763 +
0,867 + 0,008 + 2,500 +
3,070 + 3,245 + 0,067 = 17,173
DF = 6, P-Value = 0,009