MÉTODOS NÃO PARAMÉTRICOS

$$X \sim f(x|\theta), \qquad \theta \in \Theta, \qquad (X_1, X_2, \dots, X_n)$$

- **Inferência não paramétrica:** Ao contrário da inferência paramétrica onde o único aspeto desconhecido erro o valor do(s) parâmetro(s) θ, vamos agora assumir que $f(.)$ é desconhecida.
- A inferência não paramétrica desenvolve-se assim num quadro bastante mais lato do que a inferência paramétrica e cobre múltiplas vertentes como a estimação de probabilidades, a estimação de funções densidade (probabilidade) caso se saiba que a variável aleatória é contínua (discreta), etc …
- No quadro da inferência paramétrica apenas se irão abordar (e de forma muito sintética) 2 tópicos:
    - **Os testes de ajustamento** ou melhor dizendo um dos testes de ajustamento, o **teste do qui-quadrado à bondade do ajustamento**. A ideia é apresentar uma ferramenta que permita testar a hipótese chave que se admitiu nos 2 capítulos anteriores: a função $f(.)$ é conhecida a menos de um conjunto de parâmetros.
    - **O teste de independência do qui-quadrado –** Este segundo tópico aborda, como o nome indica, um dos testes existentes com vista a apurar se é aceitável considerar que dois fatores são independentes.

TESTE DE AJUSTAMENTO

- **Problema**: Recolhida uma amostra casual de determinada população para a qual é proposta determinada distribuição, como testar essa proposta.
- A proposta pode corresponder a uma **hipótese simples** ou a uma **hipótese composta**:
  - Hipótese simples: $f_0(x)$ é completamente especificada

    Exemplos: Poisson com média igual a 10, binomial 5 e 0.3
  - Hipótese composta: $f_0(x)$ não é completamente especificada, isto é, depende de parâmetros desconhecidos e escreve-se $f_0(x \mid \theta_1, \theta_2, ..., \theta_k)$

    Exemplos: Poisson com média desconhecida, binomial 5 e θ

- Assim, o nosso problema pode será:

  Como testar $H_0: X \sim f_0(x)$? ou como testar $H_0: X \sim f_o(x|\theta_1, \theta_2, \dots, \theta_k)$?

  Uma possível solução → Teste do qui-quadrado à bondade do ajustamento.

- O problema vai ser analisado em 3 passos.

1ª situação: X corresponde a um atributo qualitativo com m categorias

Exemplo: Um aspirador vendido em cinco cores: A_1, A_2, A_3, A_4 e A_5.

- Notação:
  - $A_1, A_2, \ldots, A_m \rightarrow$ modalidades (categorias) que o atributo pode assumir.
  - $p_j = P(A_j) \rightarrow$ probabilidade (desconhecida) de um elemento da população, escolhido ao acaso, apresentar a modalidade A_j ($j = 1,2,\ldots,m$).
    - A distribuição é caracterizada pelos m valores desconhecidos $p_1, p_2, \ldots, p_m$
    - Claro que $\sum_{j=1}^{m} p_j = 1$ e $p_j > 0$ para $j = 1,2,\ldots,m$.
- Hipótese nula em teste: $H_0: p_j = p_{0j}$ $(j = 1,2,\ldots,m)$ contra $H_1: p_j \neq p_{0j}$, para algum j, sendo $p_{01}, \ldots, p_{0m}$ conhecidos ($p_{0j} > 0$ com $j = 1,\ldots,m$ e $\sum_{j=1}^{m} p_{0j} = 1$).

- A estatística de teste:

  - Recolhe-se uma amostra casual $(X_1, X_2, \dots, X_n)$ – cada valor de X é um inteiro entre 1 e m já que o fator X só assume m modalidades – e conta-se quantas vezes cada modalidade é observada, seja N_j, $j = 1,2, \dots, m$.

  - $N_j \rightarrow$ v.a. que representa o número de observações (na amostra de dimensão n) que assumem a modalidade A_j, $\sum_{j=1}^{m} N_j = n$
  - Estatística de teste:

$$Q = \sum_{j=1}^{m} \frac{(N_j - np_{0j})^2}{np_{0j}}$$

A estatística de teste mede o **afastamento** (ponderado) **entre os dados observados** (N_j) **e a hipótese em análise** (p_{0j}) - se a modalidade j tem probabilidade p_{0j} de ser observada na população, então o número esperado de observações numa amostra de dimensão n será dado por $n\, p_{0j}$.

Quanto maior for o valor observado Q_{obs}, menos plausível é a hipótese em teste.

- O teste:
  - (Teorema 9.1 do livro) Quando Ho é verdadeira a distribuição assintótica de Q é dada por

$$Q = \sum_{j=1}^{m} \frac{(N_j - np_{0j})^2}{np_{0j}} \overset{a}{\sim} \chi^2(m-1)$$

  - A região de rejeição de dimensão α é $W_\alpha = \{q : q > q_\alpha\}$ onde $q_\alpha : P(Q > q_\alpha) = \alpha$
  - Verificar se $Q_{obs} = \sum_{j=1}^{m} \frac{(n_j - np_{0j})^2}{np_{0j}}$ cai na região critica
  - Alternativamente pode recorrer-se ao cálculo do *valor-p* $=P(Q \geq Q_{obs})$.

- Observação importante: A distribuição de Q é válida quando $n \to +\infty$. Para que a aproximação no caso finito seja válida, deve-se garantir um valor mínimo para $n\, p_{0j}$. Das várias regras existentes na literatura para fixa este mínimo, iremos utilizar $n\, p_{0j} \geq 5$ (**número esperado** de elementos em cada classe é pelo menos 5). Se necessário, agregam-se classes. Cuidado que o número mínimo diz respeito aos p_{0j} e à dimensão da amostra e não aos valores observados.

- **Exemplo** (9.1 do livro) - Um aspirador é vendido em cinco cores: verde (A_1), castanho (A_2), encarnado (A_3), azul (A_4) e branco (A_5). Num estudo de mercado para apreciar a popularidade das várias cores analisou-se uma amostra casual de 300 vendas recentes com o seguinte resultado

A_1	A_2	A_3	A_4	A_5	Total
88	65	52	40	55	300

Pretende testar-se $(\alpha = 0.05)$ a hipótese de que os consumidores não manifestam preferência por qualquer das cores.

Solução:

1. Formalizar a hipótese nula $H_0 : p_{01} = p_{02} = p_{03} = p_{04} = p_{05} = 0.2.$ e a alternativa $H_1 : H_0\ falsa$
2. **Obter as frequências esperadas** e compará-las com as frequências observadas

Modalidades	**Freq. Obs.** (n_j)	**Freq. esp.** (np_{0j})	$\frac{(n_j - np_{0j})^2}{np_{0j}}$
A_1	88	60	13.07
A_2	65	60	0.42
A_3	52	60	1.07
A_4	40	60	6.67
A_5	55	60	0.42
Total	300	300	21.65

3. Efetuar o teste

$\alpha = 0.05$; $m - 1 = 4$ logo $Q_{0.05} = 9.49$. Como $Q_{\text{obs}} = 21.65 > 9.49$ rejeita-se H_0

Alternativamente: valor-p = 0.00023 (computador) logo rejeita-se H_0

Em qualquer dos casos conclui-se que existe preferência por algumas cores em detrimento de outras.

Exemplo – O grau de satisfação dos clientes de determinada operadora de telecomunicações é avaliado na seguinte escala qualitativa: A_1 (muito insatisfeito), A_2 (insatisfeito), A_3 (neutro, nem satisfeito nem insatisfeito), A_4 (satisfeito) e A_5 (muito satisfeito). A administração defende que a estrutura de probabilidades que traduz o grau de satisfação dos clientes é dada por
$p_{01} = 0.025,\ p_{02} = 0.175,\ p_{03} = 0.3,\ p_{04} = 0.35\ e\ p_{05} = 0.15.$

Recolhida uma amostra casual de 100 clientes, observou-se

A_1	A_2	A_3	A_4	A_5	Total
6	15	35	40	4	100

Teste $(\alpha = 0.05)$ a distribuição apresentada pela administração.

Solução:

1. Formalizar as hipóteses

   $H_0 : p_{01} = 0.025,\ p_{02} = 0.175,\ p_{03} = 0.3,\ \ p_{04} = 0.35\ e\ p_{05} = 0.15$ contra $H_1 :\ H_0$ falsa

2. Obter as frequências esperadas e compará-las com as frequências observadas

Modalidades	A_1	A_2	A_3	A_4	A_5	**Total**
Freq. Obs (n_j)	6	15	35	40	4	100
Freq. Esp (np_{0j})	2.5	17.5	30	35	15	100
$\frac{(n_j - np_{0j})^2}{np_{0j}}$						

De acordo com a regra referente ao número mínimo de **elementos esperados** em cada grupo teremos de agregar os 2 primeiros grupos (o mais parecido com "muito insatisfeito" será "insatisfeito") e testar não a distribuição proposta mas uma distribuição "aparentada".

Duas notas antes de reformular o teste:

- **O número mínimo é para a frequência esperada e não para a frequência observada**. Assim o grupo 5 não levanta qualquer problema ao contrário do grupo 1.

- Neste caso concreto (frequência esperada de 2.5) outras regras poderiam ter sido mais flexíveis e autorizar a continuação do teste.

3. Formalizemos então a hipótese "aparentada":

   $H'_0: p_{012} = 0.2,\ p_{03} = 0.3,\ p_{04} = 0.35\ e\ p_{05} = 0.15\ \ contra\ H'_1:\ H'_0$ falsa representando-se por p_{012} a probabilidade do grupo "muito insatisfeito ou insatisfeito" que se representará por A_{12}

4. Obter as frequências esperadas e compará-las com as frequências observadas

Modalidades	**Freq. Obs.** (n_j)	**Freq. esp.** (np_{0j})	$\frac{(n_j - np_{0j})^2}{np_{0j}}$
A_{12}	21	20	0.050
A_3	35	30	0.833
A_4	40	35	0.714
A_5	4	15	8.067
Total	100	100	**9.664**

5. Efetuar o teste

$\alpha = 0.05$; $m - 1 = 3$ logo $Q_{0.05} = 7.815$ e como $Q_{obs} = 9.664 > 7.815$ rejeita-se H'_0.
Alternativamente: valor-p = 0.022 (computador) logo rejeita-se H'_0.
Rejeita-se assim que a distribuição proposta pela administração seja adequada.

Quando se rejeita H'_0, não é problemático rejeitar H_0 mas a inversa pode ser mais preocupante!

2ª situação: H_0 é hipótese simples $X \sim f_0(x)$ (não envolvendo qualquer parâmetro desconhecido)

Ideia base $\rightarrow$ Adaptar a situação ao caso anterior para aplicar a metodologia que se acabou de ver.

- Construir uma partição do domínio de X em m classes, $A_1, A_2, \ldots, A_m$.
- Calcular os valores $p_{0j} = P(A_j)$ ($j = 1, \ldots, m$). Para tal recorre-se a $f_0(x)$, isto é, assume-se H_0 verdadeira.
    - Quando a partição é dada parte-se dela;
    - Quando a partição fica ao nosso cuidado:
        - Variável contínua: constroem-se, tanto quanto possível, classes equiprováveis
        - Variável discreta: Seguem-se, tanto quanto possível os valores com probabilidade positiva (geralmente agregam-se na cauda direita).
- Substituir $H_0 : X \sim f_0(x)$ por $H_0' : p_j = p_{0j}$ ($j = 1, 2, \ldots, m$) utilizar o procedimento anterior para testar H_0'.
- **Cuidado**: Testar H_0' em vez de H_0 pode, por vezes, ser bastante delicado, sobretudo quando o número de classes, m, é pequeno. Rejeitar H_0' implica rejeitar H_0 mas a inversa não é verdadeira.

Exemplo (9. 2 do livro) – Um estudo sobre o tempo de vida em dias de uma amostra de 1000 tubos electrónicos deu os seguintes resultados

Tempo de vida	**Freq. obs.**
$X < 150$	543
$150 \le X < 300$	258
$300 \le X < 450$	120
$450 \le X < 600$	48
$600 \le X < 750$	20
$X \ge 750$	11
Total	1000

O fabricante afirma que o tempo de vida dos tubos, *X*, tem distribuição exponencial com média $\mu = 200$. Suportam os dados esta hipótese?

Uma distribuicao exponencial com

media mu=200

tem parametro

Lambda=1/mu=1/200

Solução:

1. Formalizar a hipótese nula inicial:

$$H_0 : X \sim f_0(x) = \frac{1}{200}\exp\left\{-\frac{x}{200}\right\} = 0.005\,e^{-0.005x} \qquad (x > 0)$$

2. Definir as classes: Já estão definidas

lambda=1/200=0.005

3. Calcular as respectivas probabilidades:
   Recordar que $F(x|\lambda) = 1 - e^{-\lambda x}, x > 0$

$$p_{01} = P(X < 150) = F(150) = 1 - e^{-0.005\times 150} = 1 - e^{-0.75} = 0.52763$$

$$p_{02} = P(150 \le X < 300) = F(300) - F(150) = e^{-0.75} - e^{-1.5} = 0.24924$$

$$p_{03} = P(300 \le X < 450) = F(450) - F(300) = e^{-1.5} - e^{-2.25} = 0.11773$$

$$p_{04} = P(450 \le X < 600) = F(600) - F(450) = 0.05561$$

$$p_{05} = P(600 \le X < 750) = F(750) - F(600) = 0.02627$$

$$p_{06} = P(X \ge 750) = 1 - F(750) = 1 - e^{-0.005\times 750} = 0.02352$$

4. Construir a hipótese a testar: $H_0' : p_{01} = 0.52763;\ p_{02} = 0.24924; \cdots ; p_{06} = 0.02352$

5. Obter as frequências esperadas para as seis classes $1000\times p_{0j}$ (freq. esperadas ≥ 5) e compará-las com as frequências observadas

DADOS Modelo (H0)

Tempo de vida	**Freq. obs.**	**Freq. esp.**
$X < 150$	543	527.63
$150 \leq X < 300$	258	249.20
$300 \leq X < 450$	120	117.73
$450 \leq X < 600$	48	55.61
$600 \leq X < 750$	20	26.27
$X \geq 750$	11	23.52
Total	1000	1000.00

6. Efetuar o teste:

$$Q_{\text{obs}} = \frac{(543-527.63)^2}{527.63} + \frac{(258-249.2)^2}{249.2} + \cdots + \frac{(11-23.52)^2}{23.52} = 10.0004$$

$\alpha = 0.05$, $m-1=5$ logo $Q_{0.05} = 11.1$ (ver tabela qui-quadrado com 5 g.l.) ou valor-p=0.075

Como $Q_{\text{obs}} < 11.1$ não se rejeita H_0' e, então, pode-se admitir que não será de pôr em causa H_0.

O exemplo que acaba de apresentar-se permite sublinhar um aspeto importante.

- De facto, a hipótese testada não foi, $H_0 : X \sim f_0(x) = 0.005e^{-0.005x} \quad (x > 0)$, mas sim a hipótese "aparentada", H_0', dada pelas condições seguintes:

$$p_{01} = P(X < 150) = \int_0^{150} 0.005e^{-0.005x}dx;$$

$$p_{02} = P(150 \leq X < 300) = \int_{150}^{300} 0.005e^{-0.005x}dx;$$

$$\ldots$$

$$p_{06} = P(X \geq 750) = \int_{750}^{+\infty} 0.005e^{-0.005x}dx.$$

- Se se rejeita H_0' então não há dúvida de que também se deve rejeitar H_0. Já o mesmo não pode dizer-se quando não se rejeita H_0', sobretudo se forem consideradas poucas classes.

Exemplo (9.4 do livro) – Determinada empresa seguradora baseia o seu sistema de prémios para determinado risco na premissa de que o número de sinistros por apólice tem distribuição de Poisson de parâmetro $\lambda = 0.2$. Recolhida uma amostra de 1000 apólices referentes ao ano anterior observou-se:

Nº sinistros por apólice	0	1	2	3
Nº apólices	800	175	21	4

A amostra põe em causa a premissa da seguradora?

Solução

1. Formalizar a hipótese nula inicial: $H_0: X \sim Po(0.2)$ contra H_1: H_0 falsa
2. Calcular as probabilidades e ver quantas classes se podem definir:

   Como a amostra tem dimensão 1000, a probabilidade mínima associada a cada classe é dada por $1000 \times p \geq 5$ logo $p \geq \frac{5}{1000} = 0.005$.

   Propriedades da distribuicao Poisson com parametro 0.2

   $P(X=0) = 0.8187;$ $P(X=1) = 0.1637;$ $P(X=2) = 0.0164;$

   Como $P(X > 2) = 1 - P(X \leq 2) = 0.0011 < 0.005$ apenas se poderão considerar 3 classes, $\{x=0\}, \{x=1\}, \{x \geq 2\}$

   P(X>2)=1-P(X<=2)

   =1-[P(X=0)+P(X=1)+P(X=2)]
3. Construir a hipótese a testar: $H'_0: p_{01} = 0.8187;\ p_{02} = 0.1637;\ p_{03} = 0.0175;$

4. Obter as frequências esperadas e compará-las com as frequências observadas

Modalidades	**Freq. Obs.** (n_j)	**Freq. esp.** (np_{0j})	$\frac{(n_j - np_{0j})^2}{np_{0j}}$
$\{x = 0\}$	800	818.7	0.427
$\{x = 1\}$	175	163.7	0.780
$\{x \geq 2\}$	25	17.5	3.214
Total	1000	999.9	**4.421**

5. Efetuar o teste

$\alpha = 0.05$; $m - 1 = 2$ logo $Q_{0.05} = 5.991$ e como $Q_{obs} = 4.421 < 5.991$ não se rejeita H_0

Alternativamente: valor-p = 0.1096 (computador) logo não se rejeita H_0

A conclusão deve ser tirada com precaução já que se testou H'_0 que considera apenas as probabilidades de 3 classes e não H_0

3ª situação: H_0 é uma hipótese composta, $f_0(x)$ envolve parâmetros desconhecidos,

$$H_0 : X \sim f_0(x \mid \theta_1, \theta_2, ..., \theta_k).$$

- Admitir que H_0 é verdadeira e estimar por **máxima verosimilhança** os parâmetros desconhecidos.

- Recorrer ao procedimento anterior, isto é definir as m classes e utilizar agora as probabilidades estimadas $p_j(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$ $(j = 1,2,...,m)$

- Definir a hipótese H_0' e proceder como anteriormente (nomeadamente garantindo que o número esperado de elementos por classe é superior a 5).

- Ao fazer o teste, a qui-quadrado tem agora $m-1-k$ graus de liberdade, isto é desconta-se um grau por cada parâmetro estimado. Tem-se (teorema 9.2)

$$Q = \sum_{j=1}^{m} \frac{\left[N_j - np_j(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)\right]^2}{np_j(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)} \overset{a}{\sim} \chi^2(m-k-1)$$

- Manter a restrição de um mínimo de 5 elementos esperados por cada classe.

- Ter presente, uma vez mais, que não se está a testar H_0 mas sim H_0'.

Exemplo (9.6 do livro) – Numa amostra de 100 peças de fazenda observou-se o número de defeitos por peça tendo-se obtido os resultados seguintes:

Defeitos por peça	0	1	2	3	4	5	**Total**
Freq. observada	20	30	25	10	10	5	100

Será de aceitar ($\alpha = 0.05$) uma distribuição de Poisson?

Solução:

Formalizar o teste → $H_0: X \sim Po(\lambda)$ contra H_1: H_0 falsa

Estimar o parâmetro por MV → $\hat{\lambda} = 1.75$

Recorda-se que o estimador de MV do parâmetro da Poisson é a média da amostra. Logo a estimativa será dada pela média da amostra observada, isto é,

$$\hat{\lambda} = \frac{20 \times 0 + 30 \times 1 + 25 \times 2 + 10 \times 3 + 10 \times 4 + 5 \times 5}{100} = \frac{175}{100} = 1.75$$

Calcular as probabilidades e ver quantas classes se podem definir

$\boldsymbol{x}$	0	1	2	3	4+
$\boldsymbol{P(X = x)}$	0.1738	0.3041	0.2661	0.1552	0.1008

Já que $\boldsymbol{P(X = 4) = 0.0679}$ e $\boldsymbol{P(X > 4) = 0.0329}$

Definir a hipótese em teste:

$H'_0: p_{01} = 0.1738;\ p_{02} = 0.3041;\ p_{03} = 0.2661;\ p_{04} = 0.1552;\ p_{05} = 0.1008;$

Calcular o valor da estatística de teste

Classe	{0}	{1}	{2}	{3}	{4,5,…}	**Total**
Freq. observada	20	30	25	10	15	100
Freq. esperada	17.38	30.41	26.61	15.52	10.08	100
$\boldsymbol{\frac{(Obs-Esp)^2}{Esp}}$	0.395	0.006	0.097	1.963	2.401	**4.862**

Efetuar o teste:

$\alpha = 0.05$; $m = 5$; $k = 1$(estimou-se 1 parâmetro) logo $m - k - 1 = 5 - 1 - 1 = 3$

Assim $Q_{0.05} = 7.815$ e como $Q_{obs} = 4.862 < 7.815$ não se rejeita H_0

alternativamente: valor-p = 0.182 (computador) logo não se rejeita H_0

Tal como no exemplo anterior, mas de forma menos gravosa já que temos um maior número de classes, a conclusão é para ser tomada com precaução. Ter também presente que $\boldsymbol{n = 100}$ não pode ser considerada uma "grande" amostra.

TESTE DE INDEPENDÊNCIA

Problema: Testar se 2 variáveis aleatórias qualitativas são (ou não) independentes numa população. A ideia pode, por vezes, ser generalizada, embora com cautela, para variáveis quantitativas.

TABELA DE CONTINGÊNCIA

Observa-se uma amostra à luz de 2 atributos: O primeiro reveste r modalidades - $A_1, A_2, \ldots, A_r$ - e o segundo s modalidades - $B_1, B_2, \ldots, B_s$. Na célula (i, j) da tabela de contingência regista-se o número de elementos da amostra que verificam o nível i do atributo A e o nível j do atributo B.

Tabela de contingência $r \times s$ observada

	B_1	B_2	$\ldots$	B_s	**Totais**
A_1	n_{11}	n_{12}	$\ldots$	n_{1s}	$n_{1\circ}$
A_2	n_{21}	n_{22}	$\ldots$	n_{2s}	$n_{2\circ}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	$n_{r\circ}$
Totais	$n_{\circ 1}$	$n_{\circ 2}$	$\ldots$	$n_{\circ s}$	$\boldsymbol{n}$

n_{ij} $(i = 1,2,\ldots,r,\ j = 1,2,\ldots,s)$ representa a frequência observada na célula definida por (A_i, B_j).

$n_{i\circ} = \sum_{j=1}^{s} n_{ij}$ $(i = 1,2,\ldots,r)$, $n_{\circ j} = \sum_{i=1}^{r} n_{ij}$ $(j = 1,2,\ldots,s)$ e, claro, $n = \sum_{i=1}^{r} n_{i\circ} = \sum_{j=1}^{s} n_{\circ j}$.

Antes de observar a amostra tem-se

	B_1	B_2	$\ldots$	B_s	**Totais**
A_1	N_{11}	N_{12}	$\ldots$	N_{1s}	$N_{1\circ}$
A_2	N_{21}	N_{22}	$\ldots$	N_{2s}	$N_{2\circ}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_r	N_{r1}	N_{r2}	$\ldots$	N_{rs}	$N_{r\circ}$
Totais	$N_{\circ 1}$	$N_{\circ 2}$	$\ldots$	$N_{\circ s}$	n

Note-se que:

- n é não aleatório já que a dimensão da amostra é fixada.
- as frequências em cada classe são aleatórias: variáveis discretas que assumem os valores $0,1,\cdots,n$.
- $N_{i\circ}$ e $N_{\circ j}$ continuam a ser os totais (aleatórios) em linha e coluna respetivamente.
- Existe uma restrição já que $n=\sum_{j=1}^{s} N_{\circ j} = \sum_{i=1}^{r} N_{i\circ} = \sum_{i=1}^{r}\sum_{j=1}^{s} N_{ij}$.

Teste de independência do qui-quadrado

- Em termos do universo, as probabilidades (desconhecidas) das células (A_i, B_j) representam-se por,

$$p_{ij} = P(A_i, B_j) \ \ (i = 1,2,...,r\,;\, j = 1,2,...,s), \ \ \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1.$$

As respectivas probabilidades marginais são dadas por,

$$p_{i\circ} = \sum_{j=1}^{s} p_{ij} \ \ (i = 1,2,...,r), \ \ \sum_{i=1}^{r} p_{i\circ} = 1;$$

$$p_{\circ j} = \sum_{i=1}^{r} p_{ij} \ \ (j = 1,2,...,s), \ \sum_{j=1}^{s} p_{\circ j} = 1.$$

- Assumir a **independência entre os 2 atributos** equivale a assumir $P(A_i, B_j) = P(A_i)P(B_j)$, logo a hipótese em teste vai ser

$$H_0 : \forall (i, j) : p_{ij} = p_{i\circ}\, p_{\circ j} \quad \text{contra} \quad H_1 : \exists (i, j) : p_{ij} \neq p_{i\circ}\, p_{\circ j}.$$

- Assumindo H_0, pode-se estimar p_{ij} a partir de $p_{i\circ}$ e de $p_{\circ j}$. Os estimadores de Máxima Verosimilhança de $p_{i\circ}$ e de $p_{\circ j}$ são dados por $\hat{p}_{i\circ} = \dfrac{N_{i\circ}}{n}$ $(i = 1,2,...,r)$; $\hat{p}_{\circ j} = \dfrac{N_{\circ j}}{n}$ $(j = 1,2,...,s)$.

- Retomando o teorema 9.2, a estatística de teste vai avaliar a diferença entre a frequência observada e a frequência esperada, assumindo H_0 verdadeiro,

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(N_{ij}-n\,\hat{p}_{i\circ}\hat{p}_{\circ j})^{2}}{n\,\hat{p}_{i\circ}\hat{p}_{\circ j}}\overset{a}{\sim}\chi^{2}\left[(r-1)(s-1)\right].$$

- Os graus de liberdade obtêm-se verificando que existem rs células e que se estimaram $(r-1)$ parâmetros referentes ao fator *A* (o último valor está pré fixado) e $(s-1)$ referentes ao fator *B*. Tem-se assim

$$rs-1-(r-1)-(s-1)=(r-1)(s-1)$$

A região de rejeição vai situar-se, pelas mesmas razões do que no teste do qui-quadrado à bondade do ajustamento na aba direita da distribuição, mantendo-se a restrição referente ao número mínimo esperado de elementos em cada célula (A_i, B_j), isto é, $n\,\hat{p}_{i\circ}\hat{p}_{\circ j}\geq 5$.

Exemplo (9.9 do livro) **–** Barroso, Martins e Macedo (1987), em estudo comparativo da eficiência de empresas agrícolas, consideraram uma amostra de 69 explorações (Ribatejo e Oeste) que classificaram segundo dois atributos:

A – explorações de topo, explorações intermédias, explorações de cauda;

B – explorações vitícolas, explorações frutícolas

As modalidades do atributo A resultaram de uma classificação estabelecida em função de vários indicadores de produtividade e rendibilidade. A tabela de contingência consta do quadro que se segue:

	Vitícolas	Fruticolas	Totais
Topo	6	8	14
Intermédias	10	9	19
Cauda	14	22	36
Totais	30	39	**69**

O objetivo é testar se existe independência **na população** entre o sistema produtivo (atributo B) e a rendibilidade (atributo A).

Solução:

$$H_0 : p_{ij} = p_{i\circ}\, p_{\circ j} \;\; (i=1,2,3;\, j=1,2) \;\; vs \;\; H_1 : \exists_{(i,j)} : p_{ij} \neq p_{i\circ}\, p_{\circ j} \;\; (i=1,2,3;\, j=1,2).$$

Entre parêntesis apresentam-se as frequências esperadas, na hipótese de os atributos serem independentes. Por exemplo $\frac{14\times 30}{69} = 6.09$ ou $\frac{36\times 39}{69} = 20.35$.

	Vitícolas	Fruticolas	Totais
Topo	6 (6.09)	8 (7.01)	14
Intermédias	10 (8.26)	9 (10.74)	19
Cauda	14 (15.65)	22 (20.35)	36
Totais	30	39	**69**

Calculado o valor particular da estatística-teste,

$$Q_{obs} = \frac{(6-6.09)^2}{6.09} + \frac{(8-7.01)^2}{7.01} + \cdots + \frac{(22-20.35)^2}{20.35} = 0.9585,$$

2 graus de liberdade valor-p= $P(Q > 0.9585) = 0.62$ Não se rejeita H_0 logo ….

Observação: Ter presente que a dimensão da amostra é “modesta”

Assumindo independencia (H0)

Prob(Topo & Viticolas)
=Prob(Topo) x Prob(Viticolas)
=14/69 x 30/69

--> freq esperada= Probabilidade * n

Medidas de associação

- Quando se rejeita a independência pode haver interesse em avaliar a intensidade da associação entre os atributos.

- As medidas de associação mais conhecidas, baseadas na estatística Q, são:

a) **Coeficiente de contingência de Pearson**, $C=\sqrt{\dfrac{Q}{Q+n}}$,

que verifica a dupla desigualdade, $0 \le C \le \sqrt{(q-1)/q} < 1,\ q = \min\{r, s\}$.

b) **Coeficiente de Tschuprow**, $T=\sqrt{\dfrac{Q}{n\sqrt{(r-1)(s-1)}}}$,

em que o máximo é 1 apenas no caso em que $r = s$.

c) **Coeficiente de Cramér**, $V=\sqrt{\dfrac{Q}{n(q-1)}}$,

que verifica $0 \le V \le 1$. Note-se que $V \ge T$ ou mais precisamente $V = T$ para $r = s$ e $V > T$ para $r \ne s$.

Exemplo 9.10 – Tendo-se concluído que não se rejeitava a independência entre os atributos na população, não tem grande interesse calcular as medidas de associação. O exemplo serve apenas para ilustrar os cálculos a fazer.

$n = 69; r = 3; s = 2; q = min\{r, s\} = 2; Q_{obs} = 0.9585$

Logo,

$$C = \sqrt{\frac{Q}{Q+n}} = \sqrt{\frac{0.9585}{0.9585+69}} = 0.117$$

$$T = \sqrt{\frac{Q}{n\sqrt{(r-1)(s-1)}}} = \sqrt{\frac{0.9585}{69 \times \sqrt{2 \times 1}}} = 0.099$$

$$V = \sqrt{\frac{Q}{n(q-1)}} = \sqrt{\frac{0.9585}{69 \times (2-1)}} = 0.118$$