O Teste de ajustamento do Qui-quadrado

Um teste muito usado baseado em contagens (frequências) - é o teste do **Qui-quadrado** (K. Pearson).

Considere-se os valores possíveis da característica X, repartidos em k classes, $A_1, A_2, ..., A_k$, mutuamente exclusivas. Seja

- $\boldsymbol{n_i}$ – frequência absoluta observada da classe A_i; $\quad \sum_{i=1}^{k} n_i = n$
- $\boldsymbol{p_i}$ – a probabilidade desconhecida de obter uma observação na classe $A_i, \forall i$;
- $\boldsymbol{p_{0i}}$ – a probabilidade de obter uma observação na classe A_i supondo que a observação foi extraída de uma população com a distribuição especificada em H_0, i.e. $p_{0i} = P(A_i|H_0)$.

O Testes de ajustamento do Qui-quadrado

H_0 : $p_i = p_{i0} \quad i = 1, 2, ...k$ v.s.
H_1 : pelo menos um dos $p_i \neq p_{i0}$

A Estatística do teste é

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

Se H_0 verdadeira $X^2 \sim \chi^2_{(k-1)}$ i.e., a distribuição é assintótica, i.e., válida para dimensões de amostra elevada.

Região Crítica–RC: é unilateral direita

Rejeita-se H_0 se $X^2_{calc} > \chi^2_{\alpha,(k-1)}$

Se houver necessidade de estimar parâmetros a estatística passa a ter assintoticamente distribuição $\chi^2_{(k-\nu-1)}$, onde ν é o número de parâmetros estimados.

O Testes de ajustamento do Qui-quadrado no R

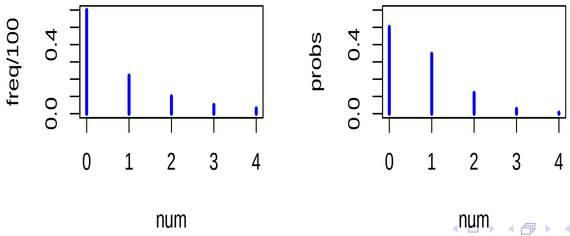
Na tabela seguinte estão representados os resultados de um estudo experimental sobre o efeito do gorgulho Azuki do feijão. Introduziram-se larvas desse orgulho nos feijões que as alimentaram. As crisálidas sairam através de um buraco feito no feijão e, como tal, o n. de buracos por feijão indica-nos o n. de adultos que saíram. Observados 100 feijões obtiveram-se os seguintes resultados:

n. de gorgulhos saídos de 1 feijão	0	1	2	3	4
frequência observada	60	22	10	5	3

Poderá considerar-se o n^o. de gorgulhos por feijão uma v.a. com distribuição de Poisson?

```
>num<-c(0,1,2,3,4); freq<-c(60,22,10,5,3)
>lambda_est<-sum(num*freq)/100;lambda_est
>probs<-c(c(dpois(num[-5],lambda_est)),
+  ppois(3,lambda_est,lower.tail=F))
>chisq.test(freq, p = probs)
```

```
        Chi-squared test for given probabilities
data:  freq
X-squared = 19.678, df = 4, p-value = 0.000578
Warning message:
In chisq.test(freq, p = probs): Chi-squared approximation
                                may be incorrect
>par(mfrow=c(1,2)) #vamos só visualizar
>plot(num,freq/100,type="h", ylim=c(0,.6),lwd=3)
>plot(num,probs,type="h",ylim=c(0,0.6),lwd=3)
```



Outros Testes de ajustamento

- `library(nortest)` é necessário carregar este package
  - `ad.test( )` teste de normalidade de Anderson-Darling
  - `cvm.test()` teste de normalidade de Cramer-Von Mises
  - `lillie.test()` teste de Lilliefors
  - `pearson.test(x.norm)` teste de normalidade do qui-quadrado de Pearson
- `library(vcd)` carregar o package para o próximo teste
  - `goodfit()` ajusta uma distribuição discreta

Tabelas de contingência

Suponhamos que os indivíduos de uma amostra são classificados de acordo com dois critérios (factores) A e B (qualitativos ou quantitativos).

Consideremos ***r*** níveis do critério A e ***c*** níveis do critério B. Portanto os **n** valores observados são classificados de acordo com 2 diferentes factores (critérios).

É costume apresentar as frequências observadas O_{ij} na célula (i, j) de uma tabela a que se chama **tabela de contingência**

	B_1	$\cdots$	B_j	$\cdots$	B_c	
A_1	o_{11}	$\cdots$	o_{1j}	$\cdots$	o_{1c}	$o_{1.}$
A_2	o_{21}	$\cdots$	o_{2j}	$\cdots$	o_{2c}	$o_{2.}$
.	.	.	.	.	.	.
A_r	o_{r1}	$\cdots$	o_{rj}	$\cdots$	o_{rc}	$o_{r.}$
	$o_{.1}$	$\cdots$	$o_{.j}$	$\cdots$	$o_{.c}$	

$\sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} = n$ e o_{ij} representa o número de elementos da amostra classificados nas categorias A_i e B_j.

Tabelas de contingência–testes de independência

Se a tabela de contingência resultou da classificação dos indivíduos da amostra segundo os níveis de cada um dos critérios, regra geral pretende-se com este estudo inferior da eventual existência de alguma relação ou associação entre os dois critérios de classificação. As hipóteses a testar são:

H$_{0:}$ A e B são independentes

H$_{1:}$ A e B não são independentes

A estatística do teste é

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - e_{ij})^2}{e_{ij}},$$

onde e_{ij} representa a estimativa da frequência esperada, se a hipótese H_0 fosse verdadeira, i.e. $e_{ij} = \frac{o_{i.} o_{.j}}{n}$

Se H_0 verdadeira, $\boldsymbol{X^2} \sim \chi^2_{(\boldsymbol{r}-\mathbf{1})(\boldsymbol{s}-\mathbf{1})}$.

Rejeita-se a hipótese H_0 se $X^2_{cal} > \chi^2_{\alpha,(r-1)(s-1)}$

Tabelas de contingência

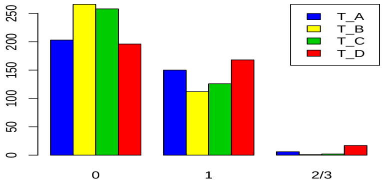
Submeteram-se ramos florais de macieiras “Golden Delicious”, a quatro tratamentos. Contou-se o número de frutos produzidos a fim de verificar se existe ou não uma relação entre os diferentes tratamentos e a frutificação. Vejamos os resultados no seguinte quadro:

Tratamentos	N. de frutos			Totais
	0	1	2ou 3	
A	203	150	6	359
B	266	112	1	379
C	258	126	2	386
D	196	168	17	381
Totais	923	556	26	1505

Pretendemos testar a hipótese nula , de que não há relação entre os tratamentos e a frutificação, ou seja, que são independentes.

Resolução do Exercício no R

```
>frutos<-matrix(c(203,150,6,266,112,1,258,126,2,196,168,17),
+ nc=3,byrow=T,
+dimnames=list(c("T_A", "T_B", "T_C","T_D"),c("0", "1","2/3")))
>frutos
>chisq.test(frutos)
>chisq.test(frutos)$expected
>(chisq.test(frutos)$residuals)^2
>barplot(frutos,names=c("0","1","2/3"),col=c(4,7,3,2),
+ cex.names=1,beside=T)
>legend("topright",c("T_A","T_B","T_C","T_D"),fill=c(4,7,3,2))
```



Notas

Pressupostos a verificar:

- as frequências esperadas em cada classe não devem ser inferiores a 5, quando o número total de observações é ≤ 20;
- se $n > 20$ não deverá existir mais do que 20% das células com frequências esperadas inferiores a 5, nem deverá existir nenhuma com frequência esperada inferior a 1.
- se nos casos anteriores as condições não se verificarem deve-se juntar linhas ou colunas (desde que tal junção tenha significado).
- a realização de um teste de independência não deve terminar com a rejeição da hipótese nula. Deve analisar-se a contribuição de cada célula para o valor de X^2.

Testes Não Paramétricos (só um cheirinho!)

Consideremos agora que pretendemos realizar testes a parâmetros mas as hipóteses da normalidade ou da aproximação à normal não são verificadas. Como já dissemos tais testes têm, habitualmente, a designação "*distribution-free*" mas muitos autores designam-nos por testes não paramétricos.

Vamos apenas deixar aqui uma breve nota sobre dois testes muito usuais que se podem aplicar quando o teste t_S*tudent* não é válido. Estes testes são baseados nas ordens das observações, ou seja, na posição de cada observação na amostra ordenada.

Enquanto os testes paramétricos exigem que as variáveis em causa sejam quantitativas, os testes não paramétricos que vamos usar podem aplicar-se também a variáveis qualitativas, desde que as elas sejam ordinais.

Testes de Wilcoxon

Teste de Wilcoxon - teste não paramétrico para o estudo da mediana de uma população ou para comparar as medianas em duas amostras emparelhadas

Teste de Wilcoxon-Mann-Whitney - teste não paramétrico adequado à comparação de duas amostras independentes

`wilcox.test(A,B)` realiza o teste que atrás chamámos de Wilcoxon-Mann-Whitney para as duas amostras independentes A e B.

`wilcox.test(A,B,paired=T)` realiza o teste que atrás chamámos de Wilcoxon para as duas amostras A e B, mas agora consideradas emparelhadas.

Os testes de Wilcoxon no R

A tabela seguinte dá a percentagem de concentração de zinco, determinada por dois métodos diferentes, em 9 amostras de comida:

Amostra	EDTA tritation	Espectrometria atómica
1	7.2	7.6
2	6.1	6.8
3	4.9	4.8
4	5.9	5.7
5	9.0	9.7
6	8.5	9.1
7	6.6	7.0
8	4	4.7
9	5.2	4.9

Poder-se-á afirmar que existe uma diferença significativa entre os resultados dos dois métodos?