SLIDES DE APOIO

(aulas teóricas)

Estatística Descritiva

É útil dividir a metodologia estatística em duas grandes classes: descritiva e inferencial.

Estatística Descritiva: Métodos visando organizar, apresentar e extrair informação dum conjunto de dados.

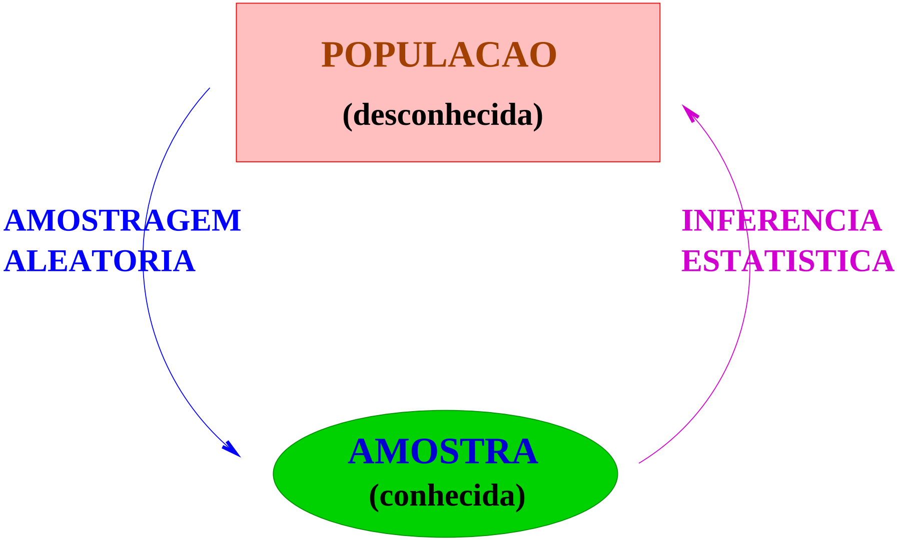
- Os dados podem ser relativos a uma população inteira (censo), a uma amostra (aleatória ou não).
- As conclusões apenas dizem respeito às entidades observadas.
- Exemplos de ferramentas descritivas:
  - Para dados de uma só variável
    - Cálculo de indicadores (média, variância, quantis, etc.).
    - Tabelas de frequências.
    - Histogramas, *boxplots* ou outras ferramentas gráficas.
  - Para dados relativos a duas variáveis
    - Indicadores (Coeficientes de correlação, covariâncias, etc..)
    - Nuvens de pontos (e, se for adequado, rectas de regressão)

Inferência Estatística

O problema conceptualmente mais complexo, de procurar conclusões relativas a um conjunto vasto de elementos (a população), a partir da observação apenas dum subconjunto dessa população (a amostra) designa-se inferência.

- Para que se possa falar em inferência estatística, é necessário que a amostra tenha sido escolhida de forma aleatória.
- A inferência estatística baseia-se na Teoria de Probabilidades, que estuda os fenómenos aleatórios.
- Exemplos de ferramentas inferenciais:
  - Estimadores e estudo das suas propriedades.
  - Intervalos de confiança para parâmetros populacionais.
  - Testes de Hipóteses.

A Inferência Estatística (cont.)



Breve revisão sobre Testes de Hipóteses

Na UC Estatística, dos primeiros ciclos do ISA, estudam-se técnicas de Estatística Descritiva e de Inferência Estatística.

Em particular, estudam-se Testes de Hipóteses para indicadores quantitativos de populações:

- média μ duma população;
- variância σ^2 duma população;
- comparação de médias de duas populações ($\mu_1 - \mu_2$);
- comparação de variâncias de duas populações ($\frac{\sigma_1^2}{\sigma_2^2}$).

As hipóteses dizem respeito à população. Opta-se entre hipóteses alternativas com base numa amostra aleatória dessa população (Estatística 2014, Cap.III - Introdução à Inferência Estatística: Acetatos p.161-185, Apontamentos p.142-147).

Revisão de Testes de Hipóteses

Num teste de Hipóteses há cinco passos a seguir.

Exemplificamos estes cinco passos no contexto dum teste a uma média populacional μ (que se presume ter sido já estudado) .

O objectivo é testar alguma afirmação sobre o valor médio μ duma variável numérica X numa dada população, por exemplo, saber se é admissível que $\mu = 2$.

No primeiro passo, formulam-se hipóteses alternativas em confronto:

Passo 1. As hipóteses em confronto

- Hipótese Nula H_0 (e.g., $\mu = 2$)
- Hipótese Alternativa H_1 (e.g., $\mu \neq 2$)

Testes de Hipóteses (cont.)

Como optar entre as hipóteses?

Precisamos de:

- Informação – que será proveniente duma amostra;
- uma regra de decisão – para optar entre as hipóteses alternativas.

A informação de que se disporá diz respeito a uma amostra concreta:

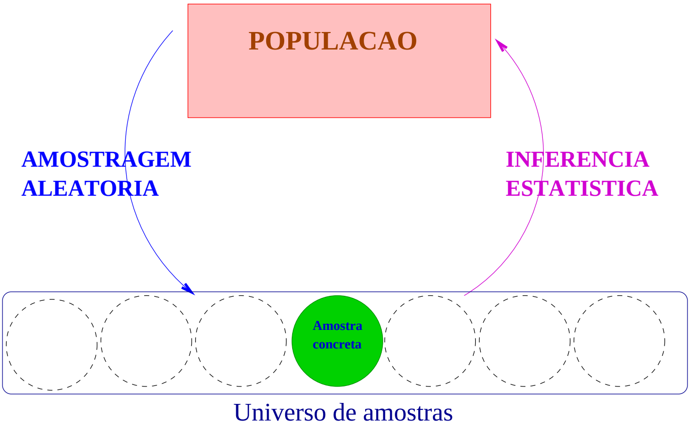
$$(x_1, x_2, ..., x_n) \; ,$$

que se admite ser a concretização dum amostra aleatória:

$$(X_1, X_2, ..., X_n) \; .$$

Uma amostra aleatória é uma colecção de variáveis aleatórias X_i (em geral, independentes e identicamente distribuídas).

Testes de Hipóteses (cont.)



Uma amostra aleatória é a ferramenta para descrever a possibilidade de sair qualquer uma das potenciais amostras concretas.

Testes de Hipóteses (cont.)

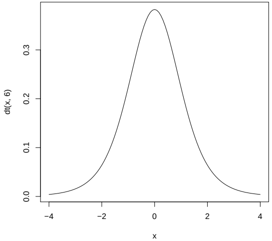
- uma amostra concreta $(x_1, x_2, ..., x_n)$ tem informação:
  - a dimensão da amostra, n;
  - o valor médio na amostra, $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$; e
  - o desvio padrão amostral, $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}$.
- Uma média $\overline{x}$ duma amostra concreta é uma estimativa da média populacional, ou seja, é o valor que o estimador $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ toma, para a amostra concreta.
- a Inferência Estatística exige que se conheça a distribuição de probabilidades do estimador de μ, ou seja, a distribuição dos valores das médias amostrais $\overline{x}$ de cada amostra concreta, ao longo do universo de possíveis amostras. É a Teoria de Probabilidades que permite esse tipo de conhecimento.

Testes de Hipóteses (cont.)

Com uma população Normal, sabe-se que:

$$T \;=\; \frac{\overline{X}-\mu}{S/\sqrt{n}} \quad \cap \quad t_{n-1} \;,$$

sendo $\overline{X}$ e S, respectivamente, a média e desvio padrão amostrais, n o tamanho da amostra e μ o verdadeiro valor da média populacional. O símbolo "$\cap$" indica "com a distribuição", neste caso uma distribuição t-Student com $n-1$ graus de liberdade.



Numa amostra com $\overline{x}$ próximo de μ, $T \approx 0$.

Assim, um hipotético valor $\mu = \mu_{|H_0}$ é plausível se T_{calc} (calculado usando $\mu = \mu_{|H_0}$) fôr próximo de zero. Quanto maior seja $|T_{calc}|$, menos plausível será $\mu_{|H_0}$.

Mas que valores de T_{calc} aconselham a rejeição de $H_0 : \mu = \mu_{|H_0}$?

Testes de Hipóteses: Passo 1 (cont.)

O papel das duas hipóteses em confronto não é simétrico.

- Hipótese Nula H_0 tem o benefício da dúvida
- Hipótese Alternativa H_1 tem o ónus da prova

Porque? Há que distinguir entre:

- a realidade (H_0 ou H_1) que não conhecemos; e
- a decisão (H_0 ou H_1), que podemos controlar.

Existem quatro possíveis combinações:

Realidade	Decisão: Admitir H_0	Decisão: Rejeitar H_0 (optar por H_1)
H_0 verdade	Certo	Erro (Tipo I)
H_0 falso (H_1 verdade)	Erro (Tipo II)	Certo

Testes de Hipóteses: Passo 2

Mas como optar entre H_0 e H_1?

Passo 2. Escolhe-se uma estatística de teste

- É uma quantidade numérica, cujo valor depende apenas da amostra e de H_0.
- É preciso conhecer a distribuição de probabilidades da estatística de teste, caso seja verdade H_0.

Para alguns valores da estatística de teste rejeitar-se-á H_0, para outros não.

Num teste a uma média μ duma população Normal, a estatística de teste usual é (sendo $\mu_{|H_0}$ o valor de μ ao abrigo de H_0):

$$T = \frac{\overline{X} - \mu_{|H_0}}{S/\sqrt{n}} \quad \cap \quad t_{n-1} \, ,$$

Testes de Hipóteses: Passo 3

Passo 3. Define-se o nível de significância do teste (α)

$$\begin{aligned}\alpha &= \text{P[Erro de Tipo I]}\\ \alpha &= \text{P[Rejeitar } H_0 \mid H_0 \text{ verdade]}\end{aligned}$$

O ideal seria $\alpha = 0$, mas isso só é possível nunca rejeitando H_0.
E nesse caso P[Erro Tipo II] = P[Não rejeitar $H_0 \mid H_0$ falso] = 1.

Não é possível reduzir simultaneamente a probabilidade dos dois tipos de erro: diminuir P[Erro Tipo I] significa reduzir a gama de valores que levam à rejeição de H_0, o que leva a aumentar P[Erro Tipo II].

Procedimento: Controlar α.
Corresponde a admitir que o Erro de Tipo I é o mais grave.

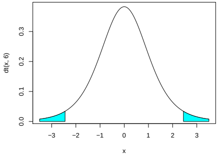
Sendo a probabilidade de um erro, queremos α pequeno:
Valores usuais: $\alpha = 0.05$, $\alpha = 0.01$.

Testes de Hipóteses: Passo 4

Passo 4. Definir a Região Crítica (Região de Rejeição)

- É o conjunto de valores da estatística ao qual associamos a rejeição de H_0;
- É constituída pelos valores da estatística de teste "menos plausíveis", caso fosse verdade H_0 (pode ser bilateral ou unilateral, dependendo da hipótese H_1);
- É uma região de probabilidade α, se fôr verdade H_0.

Exemplo de Região Crítica bilateral (adequada ao exemplo):



Testes de Hipóteses: Passo 5

Passo 5. Conclusão

- Escolhe-se uma amostra concreta;
- Calcula-se o valor da estatística para essa amostra;
- Toma-se a decisão de Rejeitar H_0 ou de Não rejeitar H_0, consoante o valor da estatística calculado para a amostra escolhida recaia, ou não, na Região Crítica.

É o único passo onde é preciso que existam dados.

Os passos 3 a 5 podem ser substituídos pela indicação duma medida de plausibilidade de H_0, designada valor de prova ou *p-value*, definido como a probabilidade de obter um valor tão ou mais extremo quanto o observado na estatística do teste, caso seja verdade H_0.

Quando um *p-value* é muito pequeno, considera-se H_0 irrealista, optando-se pela sua rejeição.

Testes χ^2 de Pearson

Neste primeiro Capítulo estudamos uma classe específica de testes de Hipóteses, que partilham uma mesma estatística de teste: a estatística de Pearson.

Estes testes são também chamados testes χ^2, uma vez que a estatística de teste segue, assintoticamente, uma distribuição qui-quadrado.

Estes "testes χ^2" surgem associados a dados de contagem, que contam as frequências observadas de várias categorias ou classes.

Começamos por considerar o caso de contagens em classes definidas por um só factor/variável.

Dados de contagem unidimensionais

Exemplo 1: Controlo de qualidade numa linha de produção de embalagens (*packs*) de 6 latas de cerveja.

Para cada embalagem, conta-se o número de latas que não passam o controlo de qualidade, havendo assim $k = 7$ possíveis resultados (0, 1, 2, 3, 4, 5 ou 6 latas rejeitadas).

Em $N = 200$ embalagens inspeccionadas, contou-se o número O_i de embalagens com i latas rejeitadas no controlo ($i \in \{0, 1, 2, 3, 4, 5, 6\}$).

Foram obtidos os seguintes valores:

No. latas impróprias	0	1	2	3	4	5	6
No. embalagens (O_i)	141	48	9	2	0	0	0

Hipótese: É admissível que as contagens sigam uma lei Binomial?

Contagens unidimensionais: Exemplo 1 (cont.)

Recordar: a distribuição Binomial surge associada a variáveis aleatórias X que contam o número de êxitos em m provas de Bernoulli, ou seja, experiências aleatórias que se podem repetir indefinidamente em condições análogas e que:

- são efectuadas m vezes de forma independente:
- cada prova só tem dois possíveis resultados ("êxito" e "fracasso");
- cada prova tem igual probabilidade p de "êxito".

No exemplo, cada uma das 200 contagens corresponde ao resultado de repetir $m=6$ vezes uma experiência que resulta no resultado "lata imprópria" (êxito) ou "lata aceite" (fracasso).

A distribuição Binomial será válida se os controlos de cada lata são independentes e com probabilidade constante de êxito.

Testes χ^2 (cont.)

Nos testes χ^2 comparam-se:

- as contagens observadas (indicadas pela letra O); com
- as contagens esperadas ao abrigo de alguma hipótese (no nosso caso a hipótese de distribuição Binomial), indicadas pela letra E.

A maior ou menor proximidade global entre contagens observadas e esperadas contém informação sobre a plausibilidade da hipótese que gerou os valores esperados.

Notação:

- N observações independentes ($N = 200$ no exemplo),
- que podem recair numa de k categorias ($k = 7$).
- O número de observações na categoria i representa-se por O_i.
- O número esperado de observações na categoria i, ao abrigo da hipótese a testar, representa-se por E_i.

Como calcular os valores esperados E_i?

No exemplo, admitimos a hipótese nula que as contagens tenham uma lei Binomial, isto é, $X \cap B(m{=}6\,,\,p)$.

Se assim fôr, a probabilidade de uma observação recair na categoria i $(i = 0, 1, ..., 6)$ é dada por:

$$\pi_i \;=\; P[X{=}i] \;=\; \binom{m}{i} p^i\,(1-p)^{m-i}\;.$$

Ao abrigo dessa hipótese, e tendo em conta o total de N observações, o número esperado de observações na categoria i seria $E_i = N \times \pi_i$.

Falta determinar o segundo parâmetro da Binomial, p, a fim de se poder calcular as probabilidades π_i.

A hipótese nula

Vamos inicialmente admitir que o outro parâmetro da Binomial tem valor $p = 0.04$, ou seja, que a hipótese nula do teste é que o número de latas impróprias em cada embalagem (representado pela variável aleatória X) segue uma distribuição Binomial, de parâmetros $m = 6$ e $p = 0.04$:

$$H_0 \ : \ X \cap B(6\,,0.04)\ .$$

Nesse caso, a probabilidade de haver i latas impróprias numa embalagem de 6 latas é dada por:

$$\pi_i = \binom{6}{i}\, 0.04^i\, 0.96^{6-i}\ , \quad \forall\, i = 0, ..., 6\ .$$

A hipótese alternativa será:

$$H_1 \ : \ \text{outros } \pi_i \qquad \Leftrightarrow \qquad H_1 : X \not\cap B(6\,,0.04).$$

Valores esperados

No exemplo, ter-se-á $E_i = 200\ \pi_i$ e:

i	0	1	2	3	4	5	6
π_i	0.7828	0.1957	0.0204	0.0011	0.0000	0.0000	0.0000
E_i	156.552	39.138	4.077	0.226	0.007	0.000	0.000

comparando-se com os valores observados:

O_i	141	48	9	2	0	0	0

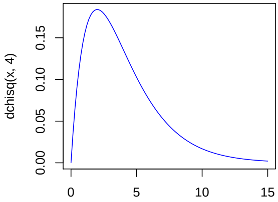
Questão:
a distribuição observada é compatível com a distribuição esperada?

A estatística de Pearson

No contexto agora descrito, Pearson mostrou que a estatística

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

segue assintoticamente (i.e., aproximadamente, para grandes amostras) uma distribuição χ^2_{k-1}, caso H_0 seja verdade.



NOTA: o número de graus de liberdade na distribuição é o número total de categorias (k) menos o número de restrições às contagens nessas categorias, que neste contexto é apenas 1: a soma de todas as contagens tem de ser N (há apenas $k-1$ "contagens livres").

Hipóteses do teste

Em geral, defina-se a hipótese nula como

H_0: a hipótese que gera os valores esperados E_i

e a hipótese alternativa como

H_1: outra distribuição de probabilidades para os π_i.

Quanto mais os O_is e E_is diferirem, mais duvidosa a hipótese nula , i.e.,
quanto maior o valor calculado da estatística X^2_{calc}, mais duvidosa H_0.

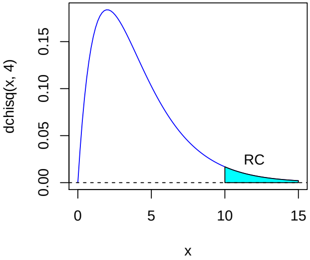
Logo, é natural definir uma Região Crítica unilateral direita.

Região Crítica

Região Crítica unilateral direita:

Rejeitar H_0 (hipótese subjacente aos E_i) se $X^2_{calc} > \chi^2_{\alpha;(k-1)}$,

sendo $\chi^2_{\alpha;(k-1)}$ o valor que, numa distribuição χ^2 com $k-1$ graus de liberdade, deixa à sua direita uma região de probabilidade α.



A probabilidade de recair na Região Crítica, se H_0 é verdade, será α.

Validade da distribuição assintótica

A distribuição da estatística de Pearson é apenas assintótica, ou seja, aproximada para grandes amostras. Há critérios diferentes para quando se considera a aproximação adequada.

Um critério, proposto por Cochran, é:

- nenhum E_i inferior a 1;
- não mais do que 20% dos E_is inferiores a 5.

Caso estas condições não se verifiquem, podem-se agrupar classes de forma a satisfazer o critério.

Exemplo

Seguindo o critério de Cochran (de forma flexível!), no exemplo anterior será necessário agrupar as classes correspondentes a 2 ou mais latas rejeitadas, obtendo-se nova tabela com apenas $k=3$ classes:

i	0	1	≥ 2
π_i	0.7828	0.1957	0.0216
E_i	156.552	39.138	4.311
O_i	141	48	11

A estatística de Pearson calculada tem valor:

$$X^2_{calc} = \frac{(141-156.552)^2}{156.552} + \frac{(48-39.138)^2}{39.138} + \frac{(11-4.311)^2}{4.311} = 13.9327 \,.$$

Numa distribuição χ^2_{3-1} o limiar da região crítica ao nível $\alpha = 0.05$ é $\chi^2_{0.05(2)} = 5.991$, pelo que se rejeita a hipótese nula $X \cap B(6, 0.04)$.

Opta-se pela hipótese H_1 de a distribuição não ser Binomial, ou ser Binomial, mas com outro valor do parâmetro p.

Pearson com a estimação de parâmetros

Por vezes, pode existir uma hipótese incompletamente especificada relativa aos valores esperados E_i. No exemplo anterior, se é natural admitir uma distribuição Binomial cujo primeiro parâmetro seja $m=6$, é discutível qual deva ser o valor do segundo parâmetro p (que representa a probabilidade duma lata individual estar imprópria).

Admitir que o número de latas impróprias por embalagem segue uma distribuição Binomial $B(6, p)$, mas com p desconhecido, não chega para calcular os valores esperados E_i. É necessária a estimação de p.

Quando o cálculo dos valores esperados exige a estimação de um ou mais parâmetros da distribuição (porque a hipótese nula está incompletamente especificada), é necessário retirar um grau de liberdade por cada parâmetro estimado.

A estimação de parâmetros no exemplo

Só é possível calcular os valores esperados E_i conhecendo o valor do parâmetro p. Uma boa forma de estimar p será a partir da expressão para o valor esperado, numa distribuição Binomial. Recorde-se que

$$X \cap B(m,p) \quad \Rightarrow \quad E[X] = mp \,,$$

Pode então usar-se a média amostral $\overline{x}$ para estimar p: $\quad \overline{x} = m\hat{p}$.

Com base nos dados, o número médio de latas impróprias por embalagem, nas 200 embalagens, é $\overline{x} = 0.36$. Como $m = 6$, tem-se

$$\hat{p} \;=\; \frac{\overline{x}}{m} \;=\; \frac{0.36}{6} \;=\; 0.06 \,.$$

Reformulam-se as Hipóteses:

$$H_0 : X \cap B(6\,,0.06) \qquad vs. \qquad H_1 : X \not\cap B(6\,,0.06)$$

Exemplo

Agora, a probabilidade estimada de haver i latas impróprias numa embalagem de 6 latas será dada por:

$$\hat{\pi}_i = \binom{6}{i} 0.06^i 0.94^{6-i} .$$

Para $N = 200$ embalagens, tem-se $\hat{E}_i = 200\ \hat{\pi}_i$.

Reconstruindo a tabela para uma Binomial $B(6\,,0.06)$, tem-se:

i	0	1	2	3	4	5	6
$\hat{\pi}_i$	0.6899	0.2642	0.0422	0.0036	0.0002	0.0000	0.0000
$\hat{E}_i$	137.974	52.841	8.432	0.718	0.034	0.001	0.000

comparando-se com os (mesmos) valores observados:

O_i	141	48	9	2	0	0	0

Exemplo (cont.)

De novo, colapsa-se a tabela para satisfazer o critério de Cochran:

i	0	1	≥ 2
$\hat{\pi}_i$	0.6899	0.2642	0.0459
$\hat{E}_i$	137.974	52.841	9.185
O_i	141	48	11

Sendo necessário estimar r parâmetros, a estatística

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

segue assintoticamente uma distribuição χ^2_{k-1-r}.

A estatística de Pearson calculada tem agora valor: $X^2_{calc} = 0.8686$.
Há aqui $k=3$ classes e $r=1$ parâmetro estimado. Logo $k-1-r=1$.

Pearson com estimação de parâmetros (cont.)

Mantém-se uma Região Crítica unilateral direita:

Rejeita-se H_0 (hipótese subjacente aos $\hat{E}_i$) se $X^2_{calc} > \chi^2_{\alpha;k-1-r}$

Numa distribuição χ^2_1 o limiar duma região crítica ao nível $\alpha = 0.05$ é $\chi^2_{0.05(1)} = 3.841$, pelo que não se rejeita a hipótese de a distribuição subjacente ser Binomial, em particular $B(6, 0.06)$.

Em geral, os graus de liberdade da distribuição são dados por:

número de categorias (k) - número de restrições ($1+r$) .

As restrições neste caso são $r+1 = 2$:

- fixar a dimensão da amostra (1); e
- estimar um parâmetro ($r = 1$).

O teste χ^2 de ajustamento a distribuições discretas

Os exemplos que acabámos de considerar mostram como o teste χ^2, baseado na estatística de Pearson, pode ser usado como um teste de ajustamento duma amostra a uma dada distribuição de probabilidades.

No exemplo considerado, tratava-se duma distribuição discreta (a Binomial). Para outras distribuições discretas (Poisson, Geométrica, Binomial Negativa ou qualquer outra, adequada ao problema sob estudo) pode proceder-se de forma análoga (ver exercícios das práticas).

No caso de distribuições contínuas, o teste pode ainda ser utilizado, definindo classes de possíveis valores da variável.

O teste χ^2 de ajustamento a distribuições contínuas

Definidas k classes de possíveis valores da variável,

- Contam-se as observações que recaem em cada classe. Essas frequências absolutas constituem os valores observados O_i;
- Os valores esperados E_i, calculam-se a partir das probabilidades π_i de recair na classe i, dada a distribuição de H_0: $E_i = N \times \pi_i$.

AVISO: Caso seja necessário estimar algum parâmetro da distribuição, procede-se como antes: calcular valores esperados estimados, ($\hat{E}_i = N \times \hat{\pi}_i$), e retirar um grau de liberdade por cada parâmetro estimado.

AVISO: Para testar a Normalidade, é preferível utilizar o teste de Shapiro-Wilks, já estudado na disciplina de Estatística do 1o. ciclo. Para outras distribuições contínuas, o teste de Kolmogorov-Smirnov constitui uma alternativa.

Teste χ^2 com tabelas de contingência

Admita-se agora um novo contexto para a questão discutida antes: classificam-se observações em várias categorias, mas essas categorias resultam de combinar os níveis de 2 factores.

No contexto dum exemplo de biodiversidade, admita-se que:

- há a locais geográficos, que constituem os níveis de um factor A;
- em cada local se contam as observações de cada uma de b espécies, que definem os níveis dum factor B.

Tabelas de contingência

Assim, as N observações são classificadas de acordo com dois diferentes factores (o que obrigará a usar uma dupla indexação).

Chama-se tabela de contingência a uma tabela com o número O_{ij} de observações em cada célula (i,j) (nível i do factor A e j do factor B):

Níveis do Factor A	Níveis do Factor B					Marginal de A
	1	2	3	$\cdots$	b	
1	O_{11}	O_{12}	O_{13}	$\cdots$	$O_{1,b}$	$N_{1\cdot}$
2	O_{21}	O_{22}	O_{23}	$\cdots$	$O_{2,b}$	$N_{2\cdot}$
3	O_{31}	O_{32}	O_{33}	$\cdots$	$O_{3,b}$	$N_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
a	O_{a1}	O_{a2}	O_{a3}	$\cdots$	$O_{a,b}$	$N_{a\cdot}$
Marginal de B	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	$\cdots$	$N_{\cdot b}$	N

$$N_{i\cdot} = \sum_{j=1}^{b} O_{ij} \qquad\qquad N_{\cdot j} = \sum_{i=1}^{a} O_{ij}$$

Diferentes contextos para tabelas de contingência

Tal como no contexto unidimensional, também no estudo de tabelas de contingência, surgem diferentes situações.

- As probabilidades π_{ij} de recair em cada célula (i,j) podem ser totalmente especificadas por alguma hipótese. Vamos exemplificar esta situação com exemplos ligados à genética.
- As probabilidades π_{ij} de recair em cada célula (i,j) podem ser desconhecidas, mas estimáveis, dada alguma hipótese. Vamos exemplificar esta situação com dois contextos frequentes:
  - testes de homogeneidade.
  - testes de independência.

Situação 1: probabilidades totalmente especificadas

Exemplo: Supõe-se que, em coelhos, existe:

- um gene que controla a cor do pêlo, com:
  - um alelo determinante do cinzento (dominante);
  - um alelo determinante do branco (recessivo).
- outro gene que controla o tipo de pelagem, com:
  - um alelo determinante do pêlo normal (dominante);
  - um alelo determinante da pelagem tipo Rex (recessivo).

Para avaliar esta hipótese, realiza-se uma experiência cruzando coelhos duma população inicial que são heterozigóticos nos dois genes, i.e., têm um alelo de cada cor e um alelo de cada tipo de pelagem.

Exemplo (cont.)

Se a segregação dos genes fôr independente, isto é, se o alelo da cor fôr independente do alelo do tipo de pelagem, seria de esperar que os descendentes desta população surgissem nas seguintes proporções:

- $9/16$ coelhos cinzentos de pelagem normal;
- $3/16$ coelhos cinzentos de pelagem tipo Rex;
- $3/16$ coelhos brancos de pelagem normal;
- $1/16$ coelhos brancos de pelagem tipo Rex;

Nesse caso, o número esperado de observações em cada célula será dado por $E_{ij} = N \times \pi_{ij}$, sendo π_{ij} a probabilidade associada à côr i e pelagem j:

E_{ij}		Pêlo	
		Normal	Rex
Cor	Cinzento	$232 \times \frac{9}{16} = 130.5$	$232 \times \frac{3}{16} = 43.5$
	Branco	$232 \times \frac{3}{16} = 43.5$	$232 \times \frac{1}{16} = 14.5$

Exemplo (cont.)

Neste exemplo existem $a = 2$ cores do pêlo e $b = 2$ tipos de pelagem, num total de $ab = 4$ situações experimentais.

Numa descendência de $N = 232$ coelhos, observaram-se:

		Pêlo	
		Normal	Rex
Cor	Cinzento	134	44
	Branco	42	12

Estes valores observados são compatíveis com os esperados caso sejam verdadeiros os pressupostos de segregação independente e dominância/recessividade dos genes?

Estatística do teste para tabelas bidimensionais

No contexto agora descrito, a estatística de Pearson tem a forma

$$X^2 \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

e segue assintoticamente uma distribuição χ^2_{ab-1}.

Os graus de liberdade são o número de células (ab) menos o número de restrições, que neste caso é apenas um: o número total de observações N que foi utilizado na experiência.

Tem-se uma Região Crítica unilateral direita, ou seja:

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

sendo H_0 a hipótese resultante da teoria genética (a hipótese que gerou os valores esperados E_{ij}).

Exemplo (cont.)

O valor da estatística de Pearson para a amostra referida é

$$X^2_{calc} = \frac{(134-130.5)^2}{130.5} + \frac{(44-43.5)^2}{43.5} + \frac{(42-43.5)^2}{43.5} + \frac{(12-14.5)^2}{14.5} = 0.5823755 \,.$$

A fronteira da região crítica ao nível $\alpha = 0.05$ é

$$\chi^2_{0.05(3)} = 7.814728 \,.$$

Logo, não se rejeita H_0, isto é, não se rejeitam as hipóteses genéticas referidas (dominância/recessividade e segregação independente dos genes).

Segunda situação: Testes de homogeneidade

Consideremos agora situações onde as hipóteses nulas a testar não estão totalmente especificadas, e exigem a estimação de parâmetros. Veremos dois casos particulares frequentes, desta situação.

Inicialmente, admitimos que o número de observações em cada nível de um dos factores é previamente fixado pelo experimentador.

Para fixar ideias, admita-se que os a totais de linha, $N_{i.}$, foram previamente determinados pelo experimentador.

Neste caso, objectivo de interesse pode ser o de ver se as $N_{i.}$ observações de cada linha (nível do factor A) se distribuem de forma análoga (homogénea) pelas b colunas (níveis do factor B).

Um teste com este objectivo chama-se um teste de homogeneidade.

Exemplo - Teste de Homogeneidade

Nos solos duma dada região foi assinalada a presença de larvas de 4 espécies de insectos que afectam as principais culturas da região.

Pretende-se investigar se as frequências relativas das espécies são, ou não, iguais nos vários tipos de solos.

Classificaram-se os solos em três tipos: arenosos, limosos e argilosos (Factor A, com a=3 níveis).

Em cada tipo de solos foram recolhidas 100 larvas, e classificadas de acordo com a respectiva espécie (Factor B, com b=4 níveis).

Exemplo (cont.)

Feita a classificação das larvas, obtiveram-se os seguintes resultados:

		Espécie de larva				
		1	2	3	4	Total
Tipos	Arenosos	27	24	23	26	100
de	Limosos	20	32	18	30	100
solos	Argilosos	13	37	16	34	100
	Total	60	93	57	90	300

O objectivo é o de saber se se pode admitir que, nos três tipos de solos, a distribuição pelas quatro espécies de larvas é idêntica.

Hipóteses num teste de homogeneidade

Seguindo com o exemplo, a hipótese nula é que as probabilidades de cada espécie, condicionais ao solo i, ($\pi_{1|i}$, $\pi_{2|i}$, $\pi_{3|i}$ e $\pi_{4|i}$), não dependem do tipo de solo i dado.

Designando por $\pi_{j|i}$ a probabilidade da espécie j, para o solo i:

$$H_0 \ : \ \begin{cases} \pi_{1|1} = \pi_{1|2} = \pi_{1|3} & [\, = \pi_{.1} \,] \\ \pi_{2|1} = \pi_{2|2} = \pi_{2|3} & [\, = \pi_{.2} \,] \\ \pi_{3|1} = \pi_{3|2} = \pi_{3|3} & [\, = \pi_{.3} \,] \\ \pi_{4|1} = \pi_{4|2} = \pi_{4|3} & [\, = \pi_{.4} \,] \end{cases}$$

A hipótese alternativa H_1 é que pelo menos uma das igualdades acima referidas não é verdadeira.

Mas que valores usar para as probabilidades $\pi_{.j}$ $(j = 1,2,3,4)$?

A estimação das probabilidades

A linha final da tabela, com as frequências absolutas $N_{\cdot j}$ de cada espécie de larva, representa uma base para estimar o que serão as probabilidades de cada espécie de larva, caso haja uma única distribuição das espécies, comum aos três tipos de solo.

		Espécie de larva				
		1	2	3	4	Total
Tipos	Arenosos	27	24	23	26	100
de	Limosos	20	32	18	30	100
solos	Argilosos	13	37	16	34	100
	Total	60	93	57	90	300

A probabilidade estimada da espécie j será $\hat{\pi}_{\cdot j} = \frac{N_{\cdot j}}{N}$, ou seja:

$$\hat{\pi}_{\cdot 1} = \frac{60}{300} = 0.20 \qquad \hat{\pi}_{\cdot 2} = \frac{93}{300} = 0.31 \qquad \hat{\pi}_{\cdot 3} = \frac{57}{300} = 0.19 \qquad \hat{\pi}_{\cdot 4} = \frac{90}{300} = 0.30$$

Os valores esperados (estimados)

Uma vez que em cada tipo de solo há $N_{i.} = 100$ observações, o número esperado de observações na célula (i,j) é dado por

$$\hat{E}_{ij} \quad = \quad N_{i.} \times \hat{\pi}_{.j} \quad = \quad N_{i.} \times \frac{N_{.j}}{N}$$

A tabela com os valores esperados estimados entre parenteses:

		Espécie de larva				
		1	2	3	4	Total
Tipos	Arenosos	27 (20)	24 (31)	23 (19)	26 (30)	100
de	Limosos	20 (20)	32 (31)	18 (19)	30 (30)	100
solos	Argilosos	13 (20)	37 (31)	16 (19)	34 (30)	100
	Total	60	93	57	90	300

Entre as observações O_{ij} e os correspondentes valores esperados estimados ($\hat{E}_{ij}$), existe concordância suficiente para admitir que as espécies se distribuem de forma análoga nos três tipos de solos?

As restrições

Admitindo que se fixou o número de observações em cada linha (níveis do factor A), tal facto impõe a restrições.

A necessidade de estimar as probabilidades dos níveis do outro factor (no nosso caso, as probabilidades de espécie, ou seja as probabilidades marginais de coluna) impõe mais $b-1$ restrições.

(NOTA: Não são b restrições pois a soma dos $\hat{\pi}_i$ tem de ser 1, logo estimar $b-1$ probabilidades determina a última estimativa.)

Assim, ao todo foram impostas $a+b-1$ restrições.

A estatística de Pearson em testes de homogeneidade

No contexto agora descrito, a estatística de Pearson tem a forma

$$X^2 \quad = \quad \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

e segue assintoticamente uma distribuição $\chi^2_{(a-1)(b-1)}$.

NOTA: Os graus de liberdade são o número de células (ab) menos o número de restrições ($a+b-1$), i.e., $ab-(a+b-1) = (a-1)(b-1)$

Tem-se uma Região Crítica unilateral direita, ou seja:

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

sendo H_0 a hipótese de homogeneidade na distribuição das amostras de cada população (a hipótese que gerou os valores esperados $\hat{E}_i$).

Exemplo (cont.)

A estatística de Pearson calculada no exemplo das larvas tem valor

$$X^2_{calc} = 10.10928 \ .$$

Este valor calculado deve ser comparado com o valor que, numa distribuição χ^2_6 (pois $(a-1)(b-1) = 2 \times 3 = 6$), deixa à direita uma região de probabilidade $\alpha = 0.05$:

$$\chi^2_{0.05(6)} = 12.591 \ .$$

Como $X^2_{calc} < \chi^2_{0.05(6)}$ não se rejeita H_0, a hipótese de homogeneidade das distribuições de espécies de larva, nos três tipos de solos.

Tal como nos casos anteriores, pode ser necessário agrupar classes do factor B, se o número esperado de observações nalgumas classes fôr demasiado baixo. Neste exemplo, tal não é necessário.

Terceiro contexto: testes de independência

Numa tabela bidimensional, há independência quando as probabilidades conjuntas são o produto das probabilidades marginais:

$$H_0 \; : \quad \pi_{ij} \quad = \quad \pi_{i.} \times \pi_{.j} \; , \qquad \forall \, i, j$$

onde

- π_{ij} indica a probabilidade duma observação recair na célula (i,j);
- $\pi_{i.}$ indica a probabilidade marginal duma observação recair no nível *i* do factor A (seja qual fôr o nível do outro factor);
- $\pi_{.j}$ indica a probabilidade marginal duma observação recair no nível *j* do factor B (seja qual fôr o nível do outro factor);

Estimação das probabilidades

Pode haver situações onde as probabilidades marginais sejam conhecidas, mas em geral não o são.

É possível estimar as probabilidades marginais a partir das frequências relativas marginais (como foi feito nos testes de homogeneidade, para o factor B):

$$\hat{\pi}_{i.} = \frac{N_{i.}}{N} \quad , \quad \forall\, i = 1,2,...,a$$

$$\hat{\pi}_{.j} = \frac{N_{.j}}{N} \quad , \quad \forall\, j = 1,2,...,b\ ,$$

onde N é o número total de observações (fixo), $N_{i.}$ é o número (livre) de observações no nível i do factor A e $N_{.j}$ é o número (livre) de observações no nível j do factor B.

Valores esperados

Caso se verifique a independência, o número esperado de observações na célula (i,j) é dado por:

$$E_{ij} \quad = \quad N \times \pi_{ij} \quad = \quad N \times \pi_{i.} \times \pi_{.j} \qquad \forall\, i,j.$$

Estimando as probabilidades marginais, caso se verifique a independência, o número esperado estimado de observações na célula (i,j) é:

$$\hat{E}_{ij} = N\,\hat{\pi}_{ij} = N\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} = N\,\frac{N_{i.}}{N}\,\frac{N_{.j}}{N} = \frac{N_{i.}\,N_{.j}}{N} \quad , \quad \forall\, i,j\,.$$

As restrições

Foram estimadas:

- $a-1$ probabilidades marginais do factor A (a última tem de dar a soma 1); e
- $b-1$ probabilidades marginais do factor B (a última tem de dar a soma 1).

Juntamente com

- 1 restrição imposta pelo número total fixo de observações (N),

tem-se um total de $(a-1)+(b-1)+1 = a+b-1$ restrições.

Testes χ^2 de independência (cont.)

Estes valores esperados estimados $\hat{E}_{ij}$ em cada uma das *ab* células serão comparados com os valores observados, O_{ij}, com base na estatística de Pearson.

NOTA: Repare-se que, embora com motivações diferentes,

- as expressões de cálculo dos $\hat{E}_{ij}$ são iguais, nos testes de homogeneidade e nos testes de independência;
- o número de restrições impostas é igual nos dois tipos de teste.

Logo, a estatística X^2 de Pearson terá uma expressão idêntica, e uma distribuição assintótica idêntica, quer nos testes de homogeneidade, quer nos testes de independência.

Mas importa não perder de vista que se trata de contextos diferentes, com hipóteses de referência diferentes e conclusões diferentes.

A estatística do teste

Quer no contexto de testes de homogeneidade, quer no contexto de testes de independência, a estatística de Pearson tem a forma

$$X^2 \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \quad = \quad \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{\left(O_{ij}-\frac{N_{i.}N_{.j}}{N}\right)^2}{\frac{N_{i.}N_{.j}}{N}}$$

e segue assintoticamente uma distribuição $\chi^2_{(a-1)(b-1)}$.

Em ambos os casos tem-se uma Região Crítica unilateral direita, ou seja:

Rejeita-se H_0 se $\quad X^2_{calc} > \chi^2_{\alpha;(a-1)(b-1)}$.

Exemplo - Teste de independência

Um estudo de $N = 6800$ alemães do sexo masculino analisou a cor do cabelo e a cor dos olhos de cada indivíduo. Os resultados foram:

	Cabelo				
Olhos	Louro	Castanho	Preto	Ruivo	Total
Azuis	1768	807	189	47	2811
Cinz./Verde	946	1387	746	53	3132
Castanhos	115	438	288	16	857
Total	2829	2632	1223	116	6800

Pretende-se testar se existe independência entre as características cor do cabelo e cor dos olhos (sendo natural que se rejeite esta hipótese).

Exemplo (cont.)

As frequências marginais de linha dão estimativas das probabilidades marginais de cada cor de olhos ($\hat{\pi}_{i.} = \frac{N_{i.}}{N}$):

$$\hat{\pi}_{1.} = \frac{2811}{6800} = 0.4134 \qquad \hat{\pi}_{2.} = \frac{3132}{6800} = 0.4606 \qquad \hat{\pi}_{3.} = \frac{857}{6800} = 0.1260$$

De forma análoga se obtêm estimativas das probabilidades marginais de cores de cabelo ($\hat{\pi}_{.j} = \frac{N_{.j}}{N}$):

$$\hat{\pi}_{.1} = \frac{2829}{6800} = 0.416\,, \quad \hat{\pi}_{.2} = \frac{2632}{6800} = 0.387\,, \quad \hat{\pi}_{.3} = \frac{1223}{6800} = 0.180\,, \quad \hat{\pi}_{.4} = \frac{116}{6800} = 0.017$$

Os valores esperados estimados em cada célula, caso haja independência, são dados por:

$$\hat{E}_{ij} \; = \; N\,\hat{\pi}_{ij} \; = \; N\,\hat{\pi}_{i.}\,\hat{\pi}_{.j} \; = \; \frac{N_{i.} \times N_{.j}}{N}\;.$$

Por exemplo, $\hat{E}_{11} = \frac{2811 \times 2829}{6800} \; = \; 1169.4587$.

Exemplo (cont.)

A tabela com os valores esperados (estimados) entre parenteses é:

Olhos	Cabelo: Louro	Castanho	Preto	Ruivo	Total
Azuis	1768 (1169.46)	807 (1088.02)	189 (505.57)	47 (47.95)	2811
Cin./Verde	946 (1303.00)	1387 (1212.27)	746 (563.30)	53 (53.43)	3132
Castanhos	115 (356.54)	438 (331.71)	288 (154.13)	16 (14.62)	857
Total	2829	2632	1223	116	6800

A estatística de Pearson será então:

$$X^2_{calc} = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(1768-1169.46)^2}{1169.46} + \ldots + \frac{(16-14.62)^2}{14.62} = 1073.508\,.$$

As dimensões da tabela são iguais às do Exemplo 2, logo a fronteira da região crítica foi dada no acetato 67: $\chi^2_{0.05(6)} = 12.591$.
Como esperado, rejeita-se claramente a hipótese de independência.

Analisando as parcelas da estatística

Em qualquer dos contextos considerados, a região de rejeição é unilateral direita, isto é, são os valores grandes da estatística que rejeitam a hipótese nula, num teste baseado na estatística de Pearson.

No caso de rejeição de H_0, e como a estatística X^2 de Pearson é uma soma de parcelas não-negativas, é possível identificar a(s) categoria(s)/célula(s) que contribuem com parcelas de maior valor e que são, por isso mesmo, maiormente responsáveis pela rejeição de H_0.

Ainda o exemplo de teste de independência

As parcelas individuais da estatística de Pearson, $\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, no caso do teste de independência acima referido, são:

	Cabelo			
Olhos	Louro	Castanho	Preto	Ruivo
Azuis	306.340	72.585	198.222	0.019
Cin./Verde	97.814	25.185	59.257	0.003
Castanhos	163.630	34.059	116.263	0.130

Uma vez que $\chi^2_{0.05(6)} = 12.592$, quase todas as células (excepto as referentes aos ruivos) são, só por si, responsáveis pela rejeição de H_0, com destaque para as associações de olhos azuis com cabelo louro e de olhos azuis com cabelo preto.

Ainda o exemplo da independência (cont.)

No entanto, o sentido destas duas associações é diferente:

- para olhos azuis/cabelo louro, tem-se

$$1768 \; = \; O_{11} \; \gg \; \hat{E}_{11} \; = \; 1169.46 \, .$$

  Trata-se duma associação positiva.

- para olhos azuis/cabelo preto, tem-se

$$189 \; = \; O_{13} \; \ll \; \hat{E}_{13} \; = \; 505.57 \, .$$

  Trata-se duma associação negativa.

A identificação das parcelas que mais contribuem para uma rejeição de H_0 pode ajudar a identificar outras hipóteses, mais realistas, subjacentes às contagens observadas.

Testes usando $p - values$

Em alternativa a fixar previamente o nível de significância α, é possível indicar apenas o valor de prova (ou *p*-value) associado ao valor calculado da estatística dum qualquer teste.

O *p-value* é a

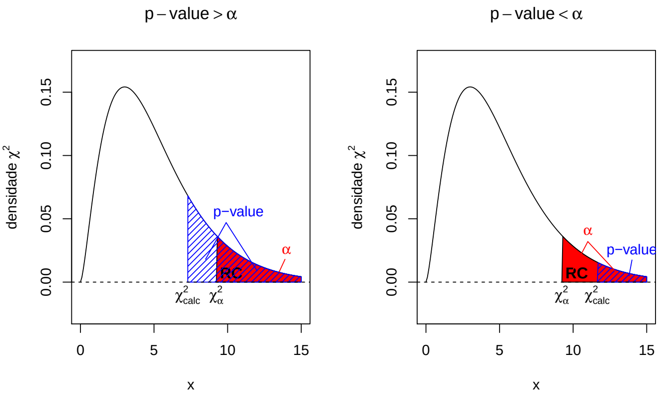
> *probabilidade da estatística de teste tomar valores mais extremos que o valor calculado a partir da amostra, sob H_0*

O cálculo do *p-value* é feito de forma diferente, consoante a natureza das hipóteses nula e alternativa conduza a regiões de rejeição unilaterais ou bilaterais. Mas no contexto dos testes χ^2, baseados na estatística de Pearson, e em que a região crítica é unilateral direita, o *p-value* é sempre calculado como:

$$p \;=\; P[\chi^2 > X^2_{calc}]\,.$$

A relação de *p-values* e níveis de significância

- $p - value \; > \; \alpha \; \Rightarrow$ não rejeição de H_0 ao nível α;
- $p - value \; < \; \alpha \; \Rightarrow$ rejeição de H_0 ao nível α;



Em geral: p-value muito pequeno implica rejeição H_0.