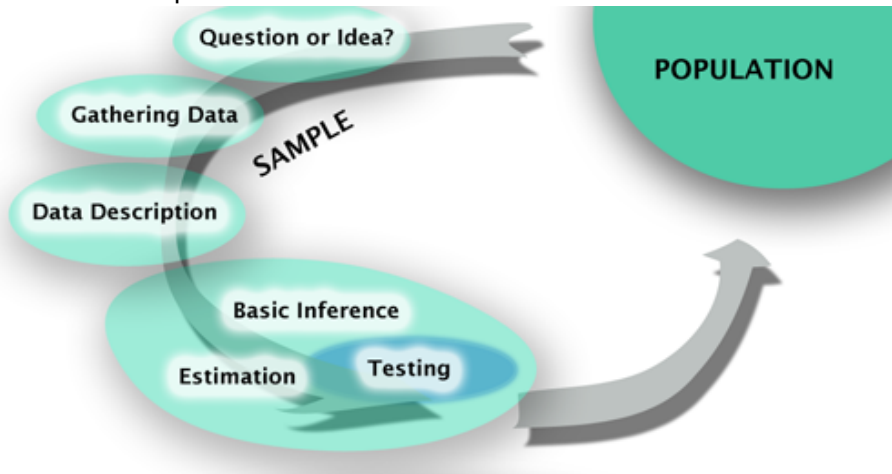


Parte C

Introdução à Inferência Estatística

Inferência Estatística

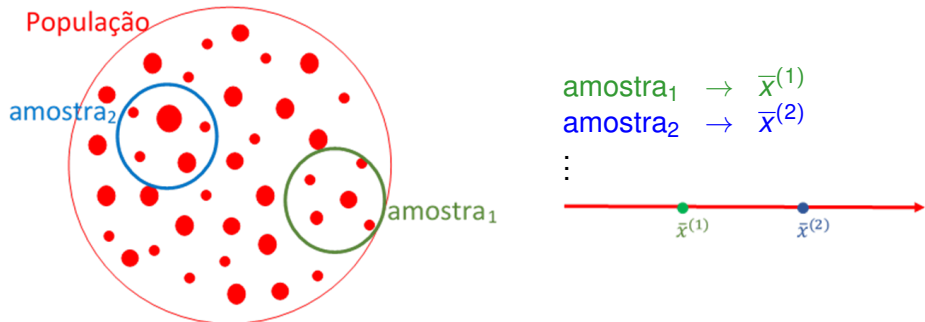
É o processo de tirar conclusões sobre uma **população** com base no conhecimento parcial de **uma amostra**.



(<https://onlinecourses.science.psu.edu/stat200/book/export/html/51>)

Inferência Estatística

Para estimar o valor médio (valor esperado) μ de uma população:



- \bar{x} têm variabilidade resultante da amostragem
- como quantificar esta variabilidade com base numa única amostra?
- onde está o μ ?

Para uma apresentação informal da Inferência Estatística sugere-se a visualização deste vídeo

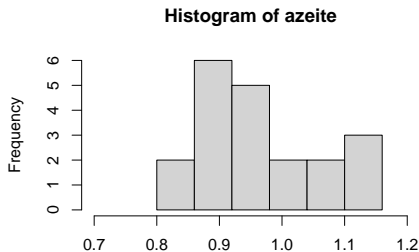
<https://www.youtube.com/watch?v=tFWsu09f74o>

- A inferência estatística é baseada em modelos de probabilidade.
- As amostras devem ser obtidas com intervenção do acaso e ser representativas da população.
- Compreende três tipos de ferramentas:
 - **estimação pontual**
 - **intervalos de confiança**
 - **testes de hipóteses**

Exemplo 10

Numa linha de engarrafamento de azeite, supõe-se que a quantidade despejada em cada garrafa (em litro) é uma variável aleatória X que segue uma distribuição normal, $X \sim \mathcal{N}(\mu, \sigma)$, com μ e σ desconhecidos. Considera-se que o processo está regulado quando $\mu = 1$ e $\sigma < 0.07$. **Como avaliar a concordância dos dados com a suposição $\mu = 1$?**

Escolhe-se ao acaso uma amostra de $n = 20$ garrafas e mede-se a quantidade (litro) de azeite contida em cada uma:



Exemplo 10

Com base nestes 20 valores pode-se:

- **estimar** o valor de μ e afirmar: **uma estimativa para μ é $\bar{x} = 0.96485$** ; (será este valor “muito” diferente de 1?)
- determinar um **intervalo de confiança** para μ e afirmar: **$\mu \in]0.9216, 1.0081[$ com 95% de confiança; como este intervalo contém o valor 1, não há razão para desconfiar que μ seja diferente de 1;**
- **testar as hipóteses** $\mu = 1$ contra $\mu \neq 1$ e afirmar: **com um nível de significância de 5%, rejeita-se a hipótese $\mu = 1$ se $\frac{\bar{x} - 1}{s/\sqrt{20}} > 2.093$. Ou ainda, este teste de hipóteses tem um p -value de 0.1053; ao nível de significância de 0.05 não se rejeita a hipótese de $\mu = 1$.**

Conceitos em Inferência Estatística

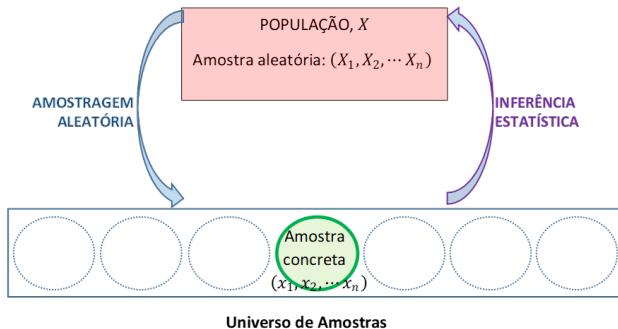
População → conjunto completo de todas os elementos com uma característica comum; a característica que se pretende estudar é considerada uma variável aleatória X . É frequente designar a própria população por X .

Amostra aleatória

Amostra aleatória de dimensão n , é uma coleção (X_1, X_2, \dots, X_n) de n variáveis aleatórias **independentes** e **semelhantes**, i.e., tendo todas a mesma distribuição que é a distribuição da população X em estudo.

Amostra concreta → coleção dos valores efetivamente observados. É considerada uma concretização (de entre as muitas possíveis) da amostra aleatória.

Conceitos em Inferência Estatística



Parâmetro de uma população → **constante desconhecida**, cujo verdadeiro valor se pretende “estimar” ou “validar”. Por exemplo μ , o valor esperado de X .

Conceitos em Inferência Estatística

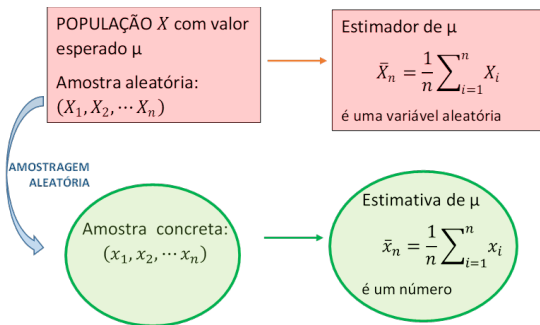
Para obter estimativas dos parâmetros desconhecidos, aceder a informação sobre o erro associado a essas estimativas e validar hipóteses sobre os valores dos parâmetros, utilizam-se **estimadores**.

Estimador e estimativa

Um **estimador de um parâmetro** é uma função da amostra aleatória (X_1, X_2, \dots, X_n) , que serve para obter “valores aproximados” de um parâmetro populacional desconhecido. Quando se aplica essa função (ou procedimento) à amostra concreta, o valor resultante designa-se **uma estimativa**.

Por exemplo, o estimador usual do valor esperado μ de uma população é a média amostral definida por $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Então, para obter uma estimativa de μ , recolhe-se uma amostra de dimensão n e aplica-se a expressão do estimador à amostra concreta:

Conceitos em Inferência Estatística



Cada amostra concreta dá origem a uma nova estimativa \bar{x} , mas em geral apenas se dispõe de **uma única** amostra com a informação:

- n , a dimensão
- \bar{x} , a média
- s , o desvio padrão

Conceitos em Inferência Estatística

Um **estimador** é uma **variável aleatória**, que se caracteriza através da sua função densidade de probabilidade. A Inferência Estatística utiliza a teoria da Probabilidade para determinar a distribuição de probabilidades do estimador, designada **distribuição de amostragem**.

Por exemplo, a distribuição de amostragem de \bar{X}_n modela a distribuição dos valores das médias \bar{x} de cada amostra concreta, ao longo do universo de possíveis amostras.

Sabe-se que se (X_1, X_2, \dots, X_n) é uma amostra aleatória extraída de uma população normal, $X \sim \mathcal{N}(\mu, \sigma)$, então

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma/\sqrt{n}) \Leftrightarrow \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Conceitos em Inferência Estatística

Se a amostragem for feita numa população com distribuição desconhecida com valor médio μ e variância σ^2 , a distribuição de amostragem de \bar{X}_n é ainda aproximadamente normal, desde que a dimensão da amostra seja suficientemente grande (Teorema Limite Central)

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1), \quad n > 30$$

Dada a amostra aleatória, (X_1, X_2, \dots, X_n) extraída de uma população X , os **parâmetros** que vão ser considerados, seus **estimadores** e **estimativas** associadas são:

Parâmetros, estimadores e estimativas

Parâmetro a estimar

Estimador

Estimativa

μ

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

σ^2

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

p

$$\hat{P} = \frac{X^{(a)}}{n}$$

$$\hat{p} = \frac{x^{(b)}}{n}$$

$\mu_1 - \mu_2$

$$\bar{X}_1 - \bar{X}_2$$

$$\bar{x}_1 - \bar{x}_2$$

σ_1^2 / σ_2^2

$$S_1^2 / S_2^2$$

$$s_1^2 / s_2^2$$

$p_1 - p_2$

$$\hat{P}_1 - \hat{P}_2$$

$$\hat{p}_1 - \hat{p}_2$$

^(a) X - v.a. que conta o número de sucessos na amostra de dimensão n

^(b) x - número observado de sucessos na amostra de dimensão n .

Conhecer a **distribuição de um estimador** permite:

- associar a uma estimativa um intervalo de valores acompanhado de uma indicação do grau de confiança em que o verdadeiro valor do parâmetro pertença a esse intervalo → **construir intervalos de confiança para o parâmetro**;
- tomar decisões sobre hipóteses colocadas sobre o valor do parâmetro, controlando o erro associado a essa decisão → **fazer testes de hipóteses ao parâmetro**.

Seguem-se as distribuições de alguns estimadores.

Estimador do valor esperado de uma população, μ

Seja (X_1, X_2, \dots, X_n) uma amostra aleatória (a.a.) extraída de uma população X , com valor esperado μ e variância σ^2 .

Estimador de μ

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Caso A: População normal com σ conhecido (slide 173)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Caso B: População com qualquer distribuição e amostra grande
($n > 30$) (slide 174)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{se } \sigma \text{ conhecido}$$

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{se } \sigma \text{ desconhecido}$$

s é o desvio padrão da amostra.

Estimador da variância de uma população, σ^2

Seja (X_1, X_2, \dots, X_n) uma a.a. extraída de uma população com distribuição normal, $X \sim \mathcal{N}(\mu, \sigma)$.

Estimador de σ^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Distribuição de amostragem:

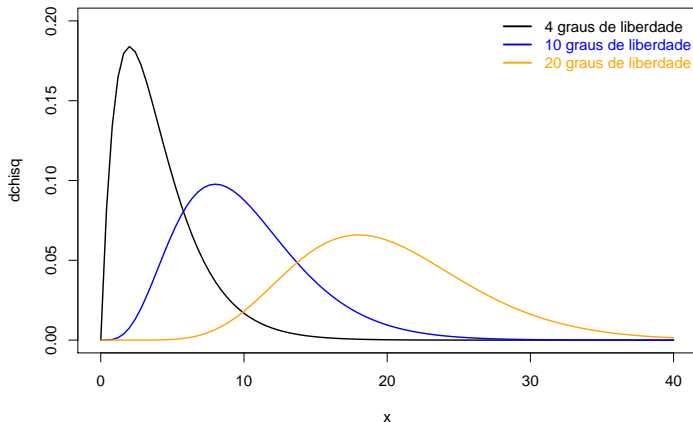
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

a v.a. $\frac{(n-1)S^2}{\sigma^2}$ tem **distribuição qui-quadrado com $n - 1$ graus de liberdade**.

Atenção: Em populações em que não se verifica a normalidade, esta distribuição não é válida.

Estimador da variância de uma população, σ^2

Função densidade da distribuição $\chi^2_{(n)}$:



$$Y \sim \chi^2_{(n)} \Rightarrow E[Y] = n \text{ e } \text{Var}[Y] = 2n$$

Estimador da variância de uma população, σ^2

Note-se que o estimador S^2 verifica:

$$E \left[\frac{(n-1)S^2}{\sigma^2} \right] = n-1 \Leftrightarrow \frac{(n-1)}{\sigma^2} E[S^2] = n-1 \Leftrightarrow E[S^2] = \sigma^2$$

Um estimador cujo valor esperado coincide com o parâmetro a estimar diz-se **estimador centrado**.

Esta é uma propriedade desejável a um bom estimador.

Conclui-se assim que

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ é um estimador centrado de } \sigma^2.$$

É esta propriedade que justifica o denominador de S^2 ser $(n-1)$ e não, por exemplo, n .

Exercício: mostrar que \bar{X} é um estimador centrado de μ .

Estimador do valor esperado de uma população, μ

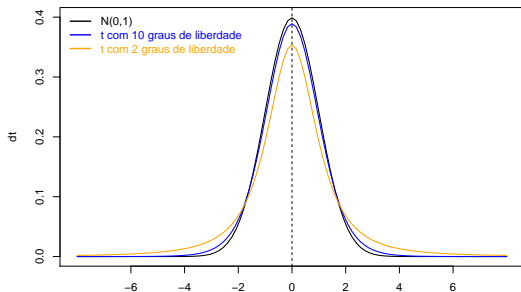
Caso C: População normal com σ desconhecido

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

em que S é a raiz quadrada do estimador de σ^2 , $S = \sqrt{S^2}$.

A v.a. $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ tem **distribuição t-Student com $n - 1$ graus de liberdade**.

Função densidade da distribuição $t_{(n)}$:



A curva é mais achatada do que a $\mathcal{N}(0,1)$ e aproxima-se desta à medida que aumenta o nº de graus de liberdade.

$$Y \sim t_{(n)} \Rightarrow E[Y] = 0 \text{ e } \text{Var}[Y] = \frac{n}{n-2}$$

Distribuições \mathcal{N} , t -Student e χ^2 em Python

- $X \sim \mathcal{N}(\mu, \sigma)$

$\mu \leftrightarrow \text{loc}$ $\sigma \leftrightarrow \text{scale}$

`norm.pdf(x, loc, scale)` # $f(x)$

`norm.cdf(x, loc, scale)` # $F(x)=P(X \leq x)$

`norm.ppf(pr, loc, scale)` # Quantil de probabilidade pr

Exemplo:

```
>>> from scipy.stats import norm
```

```
>>> print(norm.pdf(1,1,0.05)) # note que  $f(x) > 1$ 
```

```
7.978845608028654
```

```
>>> print(norm.cdf(1.96,0,1)) #  $P(X \leq 1.96)$ 
```

```
0.9750021048517795
```

```
>>> print(norm.ppf(0.975,0,1)) #  $F(x)=0.975 \Leftrightarrow x=1.96$ 
```

```
1.959963984540054
```

```
>>> print(norm.ppf(0.025,0,1)) #  $F(x)=0.025 \Leftrightarrow x=-1.96$ 
```

```
-1.9599639845400545
```

Distribuições \mathcal{N} , t -Student e χ^2 em Python

- $X \sim t_n$

$n \leftrightarrow$ df

`t.pdf(x, df)` # $f(x)$

`t.cdf(x, df)` # $F(x)=P(X \leq x)$

`t.ppf(pr, df)` # Quantil de probabilidade pr

Exemplo:

```
>>> from scipy.stats import t
>>> print(t.cdf(2.093, 19))
0.974998810528586
>>> print(t.ppf(0.975, 19))
2.093024054408263
>>> print(1-t.cdf(2.093, 19))
0.025001189471414054
>>> print(t.cdf(-2.093, 19))
0.025001189471413988
```

Distribuições \mathcal{N} , t -Student e χ^2 em Python

- $X \sim \chi_n^2$

$n \leftrightarrow$ df

`chi2.pdf(x, df) # f(x)`

`chi2.cdf(x, df) # F(x)=P(X<=x)`

`chi2.ppf(pr, df) # Quantil de probabilidade pr`

Exemplo:

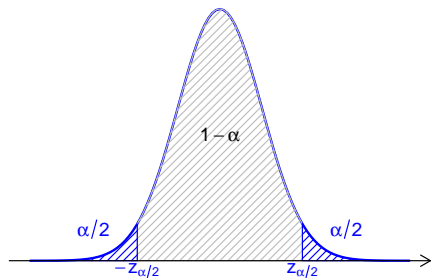
```
>>> from scipy.stats import chi2
>>> print(chi2.cdf(32.852, 19))
0.9749978372652895
>>> print(chi2.ppf(0.975, 19))
32.85232686172969
>>> print(chi2.ppf(0.025, 19))
8.906516481987971
```

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

Sabe-se que

$$\bar{X} \sim \mathcal{N}(\mu, \sigma) \Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

F. densidade de $Z \sim \mathcal{N}(0, 1)$:



$$\text{Confiança} = (1 - \alpha) \times 100\%$$

$$P[-z_{\alpha/2} < Z < z_{\alpha/2}] = 1 - \alpha$$

Valores usuais de confiança:
90%, 95%, 99%

95% de confiança:

$$\Leftrightarrow \alpha = 0.05 \rightarrow z_{\alpha/2} = z_{0.025} = 1.96$$

`scipy.stats.norm.ppf(1-0.025, 0, 1)`

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

Construção do IC:

$$P \left[-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right] = 1 - \alpha$$

$$\Leftrightarrow P \left[-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

$$\Leftrightarrow P \left[-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

$$\Leftrightarrow P \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

Construção do IC (continuação):

$$\mu \in \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \text{ com probabilidade } (1 - \alpha)$$

quando se substitui o estimador \bar{X} pela estimativa \bar{x}

↓

$$\mu \in \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \text{ com } (1 - \alpha) \times 100\% \text{ de confiança}$$

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

Intervalo a $(1 - \alpha) \times 100\%$ de confiança para μ , numa população normal com σ conhecido

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Significa que se se recolhesse muitas amostras de dimensão n e para cada uma delas se calculasse o IC, $(1 - \alpha) \times 100\%$ desses intervalos conteriam o verdadeiro (e desconhecido) valor de μ .

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

$$X \sim \mathcal{N}(2, 1)$$

$(X_1, X_2, \dots, X_{20})$ amostra aleatória de X

Confiança = 95%

$$\Leftrightarrow 1 - \alpha = 0.95 \Leftrightarrow \alpha = 0.05$$

$$Z_{\alpha/2} = Z_{0.025} = 1.96$$

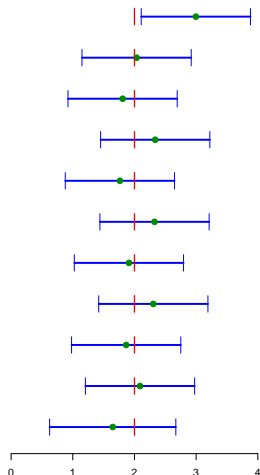
IC a 95% para μ :

$$\left[\bar{x} - 1.96 \frac{1}{\sqrt{20}}, \bar{x} + 1.96 \frac{1}{\sqrt{20}} \right]$$

$(x_1, x_2, \dots, x_{20})$ amostra concreta $\rightarrow \bar{x}$

IC's para 11

amostras concretas:



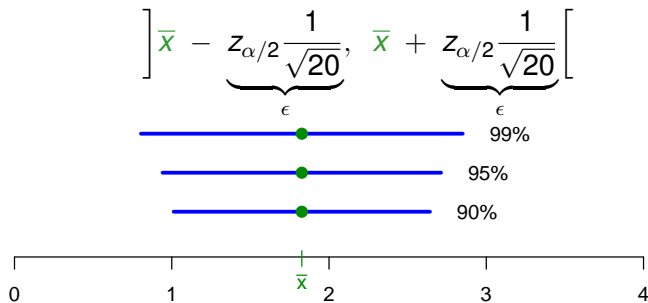
Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

Observações:

- o IC para μ é centrado na estimativa \bar{x}
- a amplitude do IC é $2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- com $(1 - \alpha) \times 100\%$ de confiança, o erro máximo cometido ao estimar μ por \bar{x} é a semi-amplitude do IC, $\epsilon = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- em populações com maior variabilidade (maior σ), os intervalos de confiança têm maior amplitude (são menos precisos)
- amostras de maior dimensão têm mais informação da população e dão origem a intervalos de menor amplitude (mais precisos)

Intervalo de confiança para o valor esperado μ de população normal com σ conhecido

- mantendo a dimensão da amostra, o **aumento da confiança** é acompanhado do **aumento da amplitude** do IC (diminuição da precisão)
- para amostras de dimensão 20, de uma população normal com $\sigma = 1$, os intervalos a $(1 - \alpha) \times 100\%$ de confiança são



No limite, o intervalo com 100% de confiança seria a reta real.

Intervalo de confiança para μ , amostras grandes ($n > 30$)

Para o caso B,

$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$, σ conhecido ou $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0, 1)$, σ desconhecido
em que s é o desvio padrão da amostra,

Intervalo a $(1 - \alpha) \times 100\%$ de confiança para μ , em amostras grandes

$$\left] \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right[\text{ se } \sigma \text{ conhecido}$$

$$\left] \bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right[\text{ se } \sigma \text{ desconhecido}$$

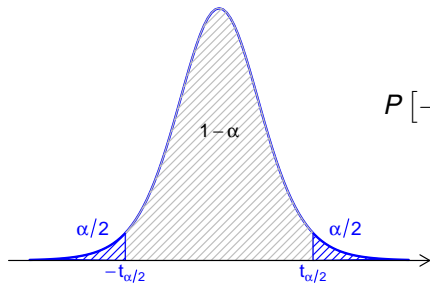
Intervalo de confiança para μ , população normal com σ desconhecido

Para o caso C:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

em que S é o estimador usual de σ .

Função densidade de $T \sim t_{(n-1)}$



$$P[-t_{\alpha/2(n-1)} < T < t_{\alpha/2(n-1)}] = 1 - \alpha$$

Intervalo de confiança para μ , população normal com σ desconhecido

A construção do IC para μ é idêntica à dos casos A e B, substituindo a distribuição $\mathcal{N}(0, 1)$ pela $t_{(n-1)}$.

Intervalo a $(1 - \alpha) \times 100\%$ de confiança para μ , numa população normal com σ desconhecido

$$\left] \bar{x} - t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} \right[$$

em que s é o desvio padrão da amostra.

Note-se que este resultado **não abrange o caso de amostras grandes** se as populações não forem normais. O caso de populações não normais é o B (caso as amostras sejam grandes). Em amostras pequenas de populações não normais é necessário utilizar testes não paramétricos, assunto que não é abordado nesta UC.

Exemplo 10 (continuação)

Para averiguar se $\mu = 1$ pode-se calcular o **intervalo de confiança a 95% para μ** . Como o desvio padrão da população (σ) é desconhecido e a amostra não é grande ($n = 20$), usa-se o intervalo $\left[\bar{x} - t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} \right]$. Note-se que é necessário pressupor que a população tem distribuição normal; o histograma parece estar de acordo com este pressuposto, mas será necessário validar esta hipótese com um teste estatístico (mais à frente).

Informação com base na amostra:

$$n = 20$$

$$\bar{x} = 0.96485 \text{ litro}$$

$$s^2 = 0.0085413 \text{ litro}^2$$

$$\text{Confiança}=95\% \Leftrightarrow 1 - \alpha = 0.95 \Leftrightarrow \alpha/2 = 0.025 \quad \rightarrow \underbrace{t_{0.025(19)} = 2.093}$$

tabela ou $t.ppf(1-0.025, 19)$

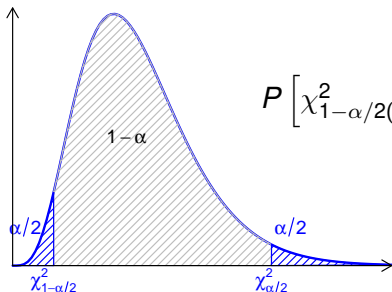
O IC tem semi-amplitude $\epsilon = t_{\alpha/2(n-1)} \frac{s}{\sqrt{n}} = 0.043253$, portanto pode-se afirmar, com 95% de confiança, que $\mu \in]0.9216, 1.0081[$. Como o valor **1** está incluído no intervalo, este é um valor possível para μ .

Intervalo de confiança para σ^2

Sabe-se que para amostras aleatórias de populações normais se tem (slide 179)

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Função densidade de $\chi^2_{(n-1)}$



$$P \left[\chi^2_{1-\alpha/2(n-1)} < \chi^2 < \chi^2_{\alpha/2(n-1)} \right] = 1 - \alpha$$

Intervalo de confiança para σ^2

A construção do IC para σ^2 segue os passos:

$$P \left[\chi_{1-\alpha/2(n-1)}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2(n-1)}^2 \right] = 1 - \alpha$$

$$\Leftrightarrow P \left[\frac{1}{\chi_{1-\alpha/2(n-1)}^2} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi_{\alpha/2(n-1)}^2} \right] = 1 - \alpha$$

$$\Leftrightarrow P \left[\frac{(n-1)S^2}{\chi_{1-\alpha/2(n-1)}^2} > \sigma^2 > \frac{(n-1)S^2}{\chi_{\alpha/2(n-1)}^2} \right] = 1 - \alpha$$

O intervalo aleatório $\left[\frac{(n-1)S^2}{\chi_{\alpha/2(n-1)}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2(n-1)}^2} \right]$ contém σ^2 com probabilidade $1 - \alpha$.

Intervalo de confiança para σ^2

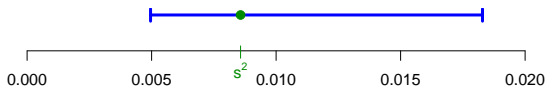
Quando se substitui o **estimador** S^2 pela **estimativa** s^2 calculada a partir de uma amostra concreta, obtém-se o

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para σ^2

$$\left] \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} \right[$$

Note-se que este intervalo não é centrado na estimativa.

Para o Exemplo 10, tem-se $\sigma^2 \in]0.004940, 0.01822[$ com 95% de confiança.



$\chi_{0.025}^2(19) = 32.8523$ e $\chi_{0.975}^2(19) = 8.90655$ (ver tabela e slide 185).

Este IC pressupõe a normalidade da população.

Intervalo de confiança para uma proporção

Pretende-se estimar a **proporção p** de “sucessos” numa população.

Por exemplo, a proporção de pinheiros infetados com uma doença em Portugal ou a proporção de artigos defeituosos numa linha de produção.

Considera-se uma amostra aleatória de dimensão n , (X_1, X_2, \dots, X_n) em que cada X_i , $i = 1, 2, \dots, n$, toma os valores 1 (sucesso) com probabilidade p e 0 (insucesso) com probabilidade $1 - p$.

$X = \sum_{i=1}^n X_i$ representa o número de sucessos na amostra aleatória.

Estimador de p

$$\hat{p} = \frac{X}{n}$$

Uma estimativa para p é $\hat{p} = \frac{x}{n}$, em que x é o número de sucessos observados na amostra.

Intervalo de confiança para uma proporção

Sabe-se que a v.a. X tem distribuição binomial com parâmetros n e p .

Se n grande,

$$X \sim B(n, p) \longrightarrow X \sim \mathcal{N}(np, \sqrt{npq}) \Leftrightarrow \hat{P} = \frac{X}{n} \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$$

Distribuição de amostragem: $Z = \frac{\hat{P} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim \mathcal{N}(0, 1)$

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para p , em amostras grandes

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Testes de hipóteses

Uma **hipótese estatística** é qualquer conjectura sobre aspetos desconhecidos da população (que podem ser parâmetros ou mesmo a forma da distribuição).

Se a hipótese diz respeito a:

- **um parâmetro**, supondo conhecida a forma da distribuição, a hipótese diz-se **paramétrica**.
- **investigar a forma da distribuição**, ou **um parâmetro** sem admitir o conhecimento da forma da distribuição, a hipótese diz-se **não paramétrica**.

Um **teste de hipóteses** é um procedimento que permite decidir se uma dada hipótese formulada sobre a população **é** ou **não** suportada pela informação fornecida pelos dados de uma amostra.

Testes de hipóteses

Em primeiro lugar irão ser estudados testes de hipóteses para parâmetros de populações:

- valor esperado μ de uma população
- variância σ^2 de uma população
- proporção de sucessos p numa população

Num teste de hipóteses há **5 passos** a seguir.

Exemplificam-se estes 5 passos no contexto de um teste a uma média populacional μ .

O objetivo é **testar alguma afirmação sobre o valor esperado μ** de uma variável numérica X numa população; por exemplo, saber se é admissível que $\mu = E[X] = 2$.

Passo 1: Especificar as hipóteses e o nível de significância do teste

As hipóteses em confronto são:

H_0 - Hipótese nula é a hipótese que tem o **benefício da dúvida** (é considerada verdadeira até haver evidência estatística para a sua rejeição, ou seja até os dados testemunharem fortemente contra essa hipótese)

H_1 - Hipótese alternativa é a hipótese que tem o **ónus da prova** onde se especificam o(s) valor(es) a “aceitar” quando se rejeita a hipótese nula.

Se os dados não contradizem a hipótese nula, a conclusão é fraca: os dados não fornecem evidência suficiente contra H_0 , o que pode acontecer com amostras pequenas e/ou em populações com muita variabilidade. Nesse caso não se aceita a hipótese alternativa.

Se a hipótese nula é rejeitada, aceita-se H_1 .

Testes de hipóteses | Passo 1 de 5

A resposta a um teste de hipóteses é dada na forma

- **Rejeitar H_0** - significa que os dados observados testemunham fortemente contra H_0 ; neste caso é **adotada a hipótese H_1** ou
- **Não rejeitar H_0** - significa que não há evidência suficiente para rejeitar H_0 .

Ao tomar decisões sobre a população com base numa amostra corre-se riscos, i.e. cometem-se erros:

Realidade	Decisão	
	Rejeitar H_0	Não rejeitar H_0
H_0 verdadeira	ERRO de tipo I	não há erro
H_0 falsa	não há erro	ERRO de tipo II

Testes de hipóteses | Passo 1 de 5

O **erro do tipo I** é considerado o mais gravoso, por isso é-lhe atribuída uma baixa probabilidade. Define-se **nível de significância** do teste como

$$\alpha = P[\text{erro do tipo I}] = P[\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}].$$

Habitualmente considera-se $\alpha = 0.05$ ou 0.01 .

O **erro do tipo II** está relacionado com a potência do teste:

$$\text{Potência} = 1 - \beta = 1 - P[\text{erro do tipo II}] = P[\text{rejeitar } H_0 \mid H_0 \text{ falsa}].$$

É possível verificar que ao diminuir a probabilidade do erro do tipo I (α), aumenta-se a probabilidade do erro do tipo II, ou seja reduz-se a potência do teste. Uma descrição de fatores que influenciam a potência de um teste pode ser consultada em [https://en.wikipedia.org/wiki/Power_\(statistics\)](https://en.wikipedia.org/wiki/Power_(statistics))

Passo 2: Definir a Estatística do Teste

- Uma **estatística de teste** é uma variável aleatória, função da amostra aleatória e de H_0 , cujo comportamento permite definir as condições que levam à rejeição de H_0 .
- É necessário conhecer a distribuição de probabilidades da estatística de teste quando H_0 é verdadeira.

Por exemplo, num teste ao valor esperado μ de uma população normal com desvio padrão σ desconhecido, a estatística de teste é

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$$

em que μ_0 é o valor de μ ao abrigo de H_0 .

Passo 3: Definir a Região Crítica (ou Região de Rejeição)

- é o conjunto de valores da estatística de teste ao qual se associa a rejeição de H_0 ;
- é constituída pelos valores da estatística de teste “menos plausíveis” caso H_0 seja verdade ;
- a probabilidade de a estatística de teste pertencer à RC é o nível de significância α ;
- a RC pode ser **bilateral** ou **unilateral**, dependendo da hipótese alternativa, H_1 .

Testes de hipóteses | Passo 3 de 5

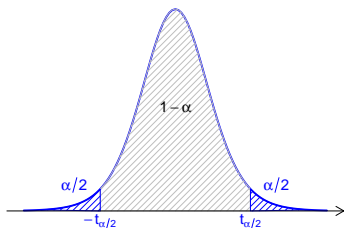
Por exemplo, num teste ao valor esperado μ de uma população normal com desvio padrão σ desconhecido, em que a estatística de teste é $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$, poderá ter-se:

- **Teste bilateral**

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Função densidade de T sob H_0



Região Crítica:

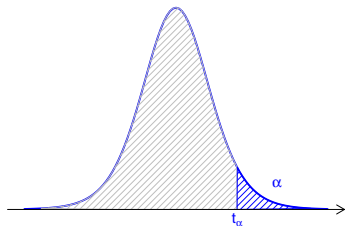
$$T < -t_{\alpha/2} \quad \text{ou} \quad T > t_{\alpha/2}$$

- **Teste unilateral direito**

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

Função densidade de T sob H_0



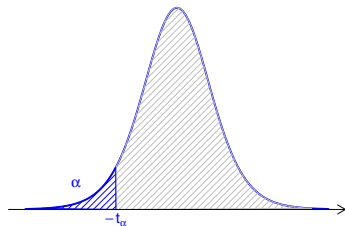
Região Crítica: $T > t_\alpha$

- **Teste unilateral esquerdo**

$$H_0 : \mu \geq \mu_0$$

$$H_1 : \mu < \mu_0$$

Função densidade de T sob H_0



Região Crítica: $T < -t_\alpha$

Passo 4: Calcular o valor da Estatística de Teste

- Escolhe-se uma **amostra concreta** (só neste passo intervêm os dados) e
- calcula-se o valor da estatística de teste para essa amostra

No exemplo,

$$T_{\text{calc}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Passo 5: Conclusão

- Toma-se a decisão de **rejeitar H_0** ou de **não rejeitar H_0** , consoante o valor da estatística de teste, calculado para a amostra observada, recaia ou não na Região Crítica.

No exemplo:

- se $T_{\text{calc}} \in RC \rightarrow$ Rejeita-se H_0 e “aceita-se” H_1
- se $T_{\text{calc}} \notin RC \rightarrow$ Não se Rejeita H_0

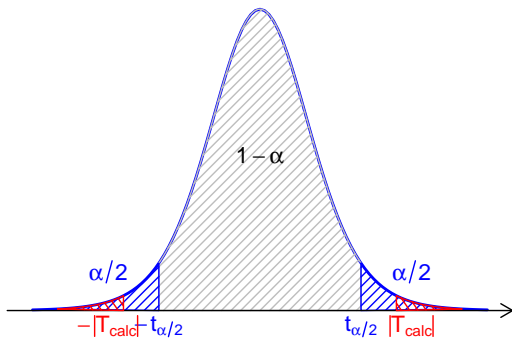
Os passos 3 a 5 podem ser substituídos pela indicação de uma medida de plausibilidade de H_0 , designada **valor de prova** ou ***p-value***, definido como a **probabilidade de obter um valor tão ou mais extremo quanto o observado na estatística do teste, caso H_0 seja verdade**. O ***p-value*** mede a concordância dos dados com H_0 . Quando um ***p-value*** é muito pequeno, considera-se H_0 irrealista, optando-se pela sua rejeição.

p -value e nível de significância

Note-se que:

- $p\text{-value} < \alpha \Leftrightarrow T_{\text{calc}} \in \text{RC} \rightarrow$ Rejeita-se H_0 ao nível de significância α

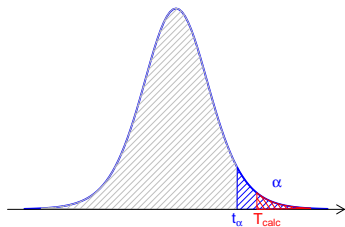
Teste bilateral



$$p\text{-value} = 2P [T > |T_{\text{calc}}|]$$

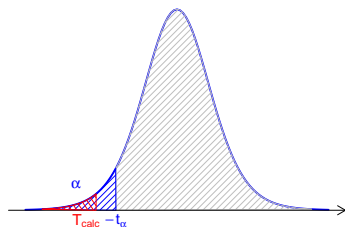
p -value e nível de significância

Teste unilateral direito



$$p\text{-value} = P [T > T_{\text{calc}}]$$

Teste unilateral esquerdo



$$p\text{-value} = P [T < T_{\text{calc}}]$$

Exemplo 10 (continuação)

Para averiguar se $\mu = 1$, uma alternativa ao IC obtido no slide 196 é a realização de um teste de hipóteses bilateral.

1. As hipóteses em confronto são $H_0: \mu = 1$ vs $H_1: \mu \neq 1$.
Nível de significância: $\alpha = P[\text{rejeitar } H_0 \mid H_0 \text{ verdade}] = 0.05$.
2. Estatística de teste: $T = \frac{\bar{X} - 1}{S/\sqrt{n}} \sim t_{(19)}$ sob H_0
3. Para $\alpha = 0.05$, $t_{0.025(19)} = 2.093$. Região crítica: $T < -2.093$
ou $T > 2.093$.
4. O valor da estatística de teste é $T_{\text{calc}} = \frac{\bar{x} - 1}{s/\sqrt{20}} = -1.700899$.
5. Como $T_{\text{calc}} \notin RC$, não se rejeita H_0 ao nível de significância de 5%.

Exemplo 10 (continuação)

O p -value, $2P [T > |T_{\text{calc}}|]$, pode obter-se com o comando Python

```
2*(1-t.cdf(1.700899, 19)) → 0.10527
```

O teste de hipóteses pode ser feito em Python com a função

`ttest_1samp` do módulo `stats` da biblioteca `scipy`:

```
azeite=[1.139, 0.833, 0.943, 0.909, 0.965, 0.928, 1.041, 1.145,  
        0.901, 1.103, 0.897, 0.907, 0.954, 0.855, 0.885, 1.010,  
        1.049, 0.926, 0.880, 1.027]
```

```
from scipy.stats import ttest_1samp  
azeite_res=ttest_1samp(azeite, popmean=1)  
ci = azeite_res.confidence_interval(confidence_level=0.95)  
print(ci)  
print(azeite_res)
```

O resultado é

```
ConfidenceInterval(low=0.9215965245545, high=1.00810347544547)  
TtestResult(statistic=-1.700899055041, pvalue=0.1052715589045,  
df=19)
```

Exemplo 10 (continuação)

Se se pretendesse averiguar se $\mu < 1$, as hipóteses seriam $H_0: \mu \geq 1$ vs $H_1: \mu < 1$. A região crítica seria (para o mesmo nível de significância) $T_{\text{calc}} < -t_{0.05(19)} = -1.729$.

$T_{\text{calc}} = -1.700899 \notin \text{RC}$, \bar{x} não é significativamente inferior a 1.

Em Python:

```
>>> azeite_res_uni=ttest_1samp(azeite, popmean=1,
    alternative='less')

>>> print(azeite_res_uni)
TtestResult(statistic=-1.7008990550410794,
    pvalue=0.05263577945227641, df=19)

>>> print('p-value=',azeite_res_uni[1])
p-value= 0.05263577945227641
```

Exemplo 10 (continuação)

Para averiguar se $\sigma^2 > 0.07^2 = 0.0049$, é preferível realizar um teste de hipóteses unilateral a calcular o IC (slide 199). O IC é equivalente a um TH bilateral, como se viu no exemplo acima.

1. As hipóteses em confronto são $H_0: \sigma^2 \leq 0.0049$ vs $H_1: \sigma^2 > 0.0049$.

Nível de significância: $\alpha = P[\text{rejeitar } H_0 \mid H_0 \text{ verdade}] = 0.05$.

2. Estatística de teste: $\chi^2 = \frac{(n-1)S^2}{0.0049} \sim \chi^2_{(19)}$ sob H_0 .

3. Para $\alpha = 0.05$, $\chi^2_{0.05(19)} = 30.14$. Região crítica: $\chi^2 > 30.14$.

4. O valor da estatística de teste é $\chi^2_{\text{calc}} = \frac{19 \times s^2}{0.0049} = 33.1193$.

5. Como $\chi^2_{\text{calc}} \in \text{RC}$, rejeita-se H_0 ao nível de significância de 5%, e conclui-se que s^2 é significativamente superior a 0.0049.

$$p\text{-value} = P[\chi^2 > \chi^2_{\text{calc}}] = 0.023289 \quad (1\text{-chi2.cdf}(33.1193, 19))$$

Teste de normalidade de Shapiro-Wilk

É um teste de hipóteses não paramétrico, para averiguar se uma população X segue a distribuição normal.

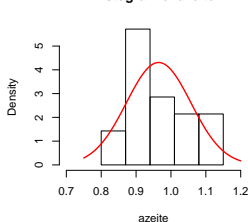
H_0 : X tem distribuição normal

H_1 : X não segue a distribuição normal

Para os dados do Exemplo 10 (slide 167), tem-se

Em Python:

Histogram of azeite



```
>>> from scipy.stats import shapiro
>>> azeite_shapiro=shapiro(azeite)
>>> print(azeite_shapiro)
ShapiroResult(statistic=0.9238094711857526,
               pvalue=0.11732708960912808)
```

Como o p -value é superior a 0.05, a este nível de significância não se rejeita a hipótese da normalidade da quantidade de azeite despejada numa garrafa.

Comparação de parâmetros de duas populações

Sejam X e Y duas populações. Pretende-se, a partir de amostras recolhidas destas populações, comparar os dois valores médios e/ou as duas variâncias populacionais.

As duas **amostras são independentes** quando não há relação entre os elementos de cada uma das amostras

- amostra aleatória de dimensão n extraída da população X :
 (X_1, X_2, \dots, X_n)
- amostra aleatória de dimensão m extraída da população Y :
 (Y_1, Y_2, \dots, Y_m)
- não há relação entre X_i e Y_j
- pode-se alterar a ordem em cada amostra

Amostras independentes e amostras emparelhadas

Um caso particular de amostras **não independentes** é o de **amostras emparelhadas**.

Duas amostras, **necessariamente com a mesma dimensão**, estão emparelhadas quando

- amostra aleatória de dimensão n extraída da população X :
 (X_1, X_2, \dots, X_n)
- amostra aleatória de dimensão n extraída da população Y :
 (Y_1, Y_2, \dots, Y_n)
- X_i e Y_i estão associadas, formam um par (X_i, Y_i)
- não se pode alterar a ordem em cada amostra

Aqui está um vídeo <https://www.youtube.com/watch?v=-6vDjGR41YM> que pode ajudar a visualizar a diferença entre amostras independentes e amostras emparelhadas.

Amostras independentes e amostras emparelhadas

Exemplo: Pretende-se comparar o rendimento médio de duas variedades de milho. A experiência pode ser delineada das duas formas seguintes

Terreno semeado com variedade A

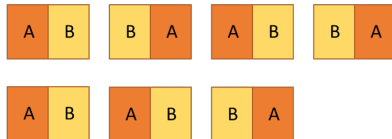


Terreno semeado com variedade B



Amostras independentes

Terreno semeado com as duas variedades



Amostras emparelhadas por talhão

Amostras independentes e amostras emparelhadas

O emparelhamento permite eliminar o efeito de fatores exteriores ao objetivo do estudo.

Neste caso, eventuais diferenças na composição do solo e nas condições de humidade e temperatura, poderiam afetar o rendimento do milho.

O emparelhamento, ao tornar as unidades experimentais mais homogéneas, permite que diferenças que se observem nos rendimentos sejam exclusivamente atribuídas à diferença entre as variedades do milho e não a outros fatores externos.

Comparação dos valores esperados de duas populações quando as amostras são independentes

Duas populações, X e Y tais que

$$E[X] = \mu_X, E[Y] = \mu_Y, \text{Var}[X] = \sigma_X^2, \text{Var}[Y] = \sigma_Y^2.$$

Para comparar μ_X com μ_Y , pode fazer-se

- intervalo de confiança para $\mu_X - \mu_Y$
- teste de hipóteses a $\mu_X - \mu_Y$

Estimador de $\mu_X - \mu_Y$ em amostras independentes

$(X_1, X_2, \dots, X_{n_X})$ amostra aleatória da população X com média \bar{X}
 $(Y_1, Y_2, \dots, Y_{n_Y})$ amostra aleatória da população Y com média \bar{Y}

$$\bar{X} - \bar{Y}$$

Comparação dos valores esperados de duas populações em amostras independentes

Para a distribuição de amostragem do estimador $\bar{X} - \bar{Y}$, será necessário considerar três casos.

Caso A: Populações normais com variâncias conhecidas

$$\bar{X} - \bar{Y} \sim \mathcal{N} \left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \right)$$
$$\iff \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1)$$

Comparação dos valores esperados de duas populações em amostras independentes

Caso B: Amostras grandes

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1) \quad \text{se } \sigma_X \text{ e } \sigma_Y \text{ conhecidos}$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \sim \mathcal{N}(0, 1) \quad \text{se } \sigma_X \text{ e } \sigma_Y \text{ desconhecidos}$$

Caso C: Populações normais com variâncias desconhecidas mas supostas iguais

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{(n_X+n_Y-2)} \quad \text{em que } S_p^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{(n_X+n_Y-2)}$$

Comparação das variâncias de duas populações em amostras independentes

Sejam X e Y duas populações normais:

$X \sim \mathcal{N}(\mu_X, \sigma_X)$ e $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$;

$(X_1, X_2, \dots, X_{n_X})$ amostra aleatória extraída da população X e

$(Y_1, Y_2, \dots, Y_{n_Y})$ amostra aleatória extraída da população Y
duas amostras independentes.

Estimador de σ_X^2/σ_Y^2 em amostras independentes de populações normais

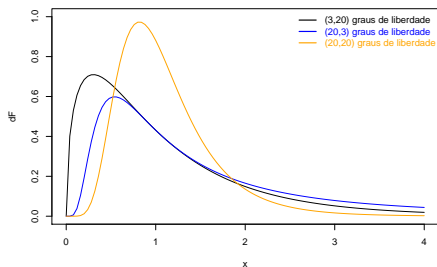
$$\frac{S_X^2}{S_Y^2}$$

A distribuição de amostragem é

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{(n_X-1, n_Y-1)}$$

Comparação das variâncias de duas populações em amostras independentes

Função densidade da distribuição F com (m, n) graus de liberdade, $F_{(m,n)}$



Em Python:

```
>>> from scipy.stats import f
>>> print(f.cdf(2.348,10,20))
0.9500101654895448
>>> print(f.ppf(0.95,10,20))
2.3478775669983114
>>> print(1/f.ppf(1-0.95,20,10))
2.3478775669983114
```

Comparação dos valores esperados de duas populações quando as amostras são emparelhadas

- Duas populações, X e Y tais que $E[X] = \mu_X$, $E[Y] = \mu_Y$
- (X_1, X_2, \dots, X_n) amostra aleatória da população X
- (Y_1, Y_2, \dots, Y_n) amostra aleatória da população Y
- (D_1, D_2, \dots, D_n) amostra das diferenças, $D_i = X_i - Y_i$
($i = 1, \dots, n$)

(D_1, D_2, \dots, D_n) é uma amostra da população das diferenças, $D = X - Y$, com $E[D] = \mu_D = \mu_X - \mu_Y$ e $Var[D] = \sigma_D^2$.

O estimador de μ_D é \bar{D} . Os resultados referentes à estimação de μ numa população, podem ser utilizados para a estimação de μ_D na população das diferenças.

Comparação dos valores esperados de duas populações quando as amostras são emparelhadas

Caso A: População D normal com σ_D conhecido

$$\frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Caso B: População D com qualquer distribuição e amostra grande

$$\frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \sim \mathcal{N}(0, 1) \text{ se } \sigma_D \text{ conhecido}$$

$$\frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \sim \mathcal{N}(0, 1) \text{ se } \sigma_D \text{ desconhecido}$$

s_D é o desvio padrão da amostra das diferenças.

Caso C: População D normal com σ_D desconhecido

$$\frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \sim t_{(n-1)}$$

em que s_D^2 é o estimador de σ_D^2 .

Comparação dos valores esperados de duas populações quando as amostras são emparelhadas

Note-se que, sendo as amostras emparelhadas (não independentes),

- $X_i \sim \mathcal{N}$ e $Y_i \sim \mathcal{N} \not\Rightarrow D_i = X_i - Y_i \sim \mathcal{N}$
- $Var[D_i] = Var[X_i] + Var[Y_i] - 2Cov[X_i, Y_i]$

Comparação dos parâmetros de duas populações em Python

Comparação dos valores esperados em duas amostras independentes de populações normais com variâncias supostas iguais

```
from scipy.stats import ttest_ind
```

- Teste bilateral ou intervalo de confiança:

$H_0: \mu_X - \mu_Y = 0$ vs $H_1: \mu_X - \mu_Y \neq 0$

```
ttest_ind(x, y, alternative='two-sided', equal_var=True)
```

- Teste unilateral esquerdo:

$H_0: \mu_X - \mu_Y \geq 0$ vs $H_1: \mu_X - \mu_Y < 0$

```
ttest_ind(x, y, alternative='less', equal_var=True)
```

- Teste unilateral direito:

$H_0: \mu_X - \mu_Y \leq 0$ vs $H_1: \mu_X - \mu_Y > 0$

```
ttest_ind(x, y, alternative='greater', equal_var=True)
```

Comparação dos parâmetros de duas populações em Python

Comparação dos valores esperados em duas **amostras independentes** de populações normais com variâncias não necessariamente iguais

Caso as variâncias populacionais não se possam considerar iguais, pode-se fazer a comparação dos valores esperados com o **teste de Welch** (ver, por exemplo

https://en.wikipedia.org/wiki/Welch's_t-test).

As instruções em Python são idênticas às do slide anterior, substituindo `equal_var=True` por `equal_var=False`

Comparação dos parâmetros de duas populações em Python

Comparação dos valores esperados em duas amostras emparelhadas, supondo que a população das diferenças tem distribuição normal

```
from scipy.stats import ttest_rel
```

- Teste bilateral ou intervalo de confiança:

$H_0: \mu_X - \mu_Y = 0$ vs $H_1: \mu_X - \mu_Y \neq 0$

```
ttest_rel(x, y, alternative='two-sided')
```

- Teste unilateral esquerdo:

$H_0: \mu_X - \mu_Y \geq 0$ vs $H_1: \mu_X - \mu_Y < 0$

```
ttest_rel(x, y, alternative='less')
```

- Teste unilateral direito:

$H_0: \mu_X - \mu_Y \leq 0$ vs $H_1: \mu_X - \mu_Y > 0$

```
ttest_rel(x, y, alternative='greater')
```


Comparação dos parâmetros de duas populações em Python

Comparação das variâncias em duas amostras independentes de populações normais.

Não existe uma função em Python para comparar variâncias através do teste F do slide 227. No entanto o valor da estatística de teste é simplesmente

$$F_{\text{calc}} = \frac{s_x^2}{s_y^2}$$

O p -value pode obter-se recorrendo à função distribuição cumulativa da distribuição $F_{(n_x-1, n_y-1)}$

Comparação dos parâmetros de duas populações em Python

Comparação das variâncias em duas amostras independentes de populações normais.

Cálculo do *p-value* do teste de hipóteses

```
from scipy.stats import f
```

- Teste bilateral:

$$H_0: \sigma_X^2 / \sigma_Y^2 = 1 \text{ vs } H_1: \sigma_X^2 / \sigma_Y^2 \neq 1$$

$$2 * \min(f.cdf(Fcalc, nX-1, nY-1), 1 - f.cdf(Fcalc, nX-1, nY-1))$$

- Teste unilateral esquerdo:

$$H_0: \sigma_X^2 / \sigma_Y^2 \geq 1 \text{ vs } H_1: \sigma_X^2 / \sigma_Y^2 < 1$$

$$f.cdf(Fcalc, nX-1, nY-1)$$

- Teste unilateral direito:

$$H_0: \sigma_X^2 / \sigma_Y^2 \leq 1 \text{ vs } H_1: \sigma_X^2 / \sigma_Y^2 > 1$$

$$1 - f.cdf(Fcalc, nX-1, nY-1)$$

Comparação de proporções

Pretende-se comparar as proporções de “sucessos” em duas populações com base em duas amostras independentes.

X_1 representa o número de sucessos numa amostra de dimensão n_1 de uma população em que a proporção de sucessos é p_1 ,
 $X_1 \sim \mathcal{B}(n_1, p_1)$

X_2 representa o número de sucessos numa amostra de dimensão n_2 de outra população em que a proporção de sucesso é p_2 ,
 $X_2 \sim \mathcal{B}(n_2, p_2)$

O estimador da diferença entre as duas proporções é

$$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Se n_1 e n_2 grandes,

$$\hat{P}_1 - \hat{P}_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

Testes χ^2 de Pearson

Os slides desta Secção são também baseados em material pedagógico disponibilizado pelo Prof. Jorge Cadima no âmbito da UC Estatística e Delineamento (<https://fenix.isa.ulisboa.pt/downloadFile/563022967868449/aulasQui2.pdf>).

Os testes de qui-quadrado de Pearson são testes de ajustamento para dados nominais ou dados classificados em categorias ou classes.

Trata-se de testes de hipóteses não paramétricos, baseados em contagens, que partilham a mesma estatística de teste, a estatística de Pearson. São também designados testes χ^2 , uma vez que a estatística de teste segue, *assintoticamente*, uma distribuição qui-quadrado.

Vão ser abordados os casos:

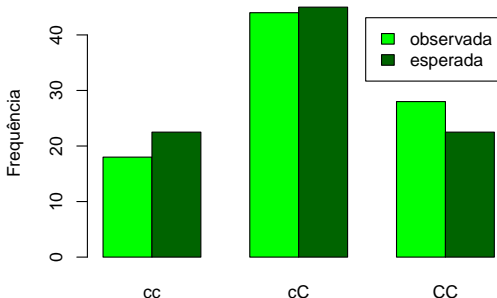
- Testes χ^2 em dados de contagem unidimensionais
- Testes χ^2 em tabelas de contingência (bidimensionais)
 - Probabilidades especificadas por uma teoria (genética, por exemplo)
 - Testes de Independência
 - Testes de Homogeneidade

Exemplo 11 | hipótese genética

A descendência originada pelo cruzamento de dois tipos de plantas pode ser qualquer um dos três genótipos cc , cC e CC . Um modelo teórico de sucessão genética indica que os tipos cc , cC e CC devem aparecer na razão $1 : 2 : 1$. Efetuou-se o cruzamento daqueles dois tipos tendo-se classificado 90 plantas. A sua classificação genética foi registada na tabela:

Genótipos	cc	cC	CC
Num. plantas	18	44	28

Questão: Estes dados estarão de acordo com o modelo genético?



Teste χ^2 em dados de contagem unidimensionais

Suponha-se que uma população é caracterizada por um atributo qualitativo que pode assumir as categorias A_1, A_2, \dots, A_k com probabilidades desconhecidas.

Considerando um conjunto de valores $\pi_i > 0$ ($i = 1, 2, \dots, k$) tais que $\sum_{i=1}^k \pi_i = 1$, pretende-se testar as hipóteses:

$H_0: P(A_i) = \pi_i, \forall i = 1, \dots, k$ versus $H_1: P(A_i) \neq \pi_i$ para algum i

Recolhe-se uma amostra de dimensão N e conta-se o número de elementos da amostra que pertencem a cada categoria. Para a categoria A_i , seja

O_i : a frequência (absoluta) observada na amostra;

$E_i = N\pi_i$: a frequência esperada ao abrigo da hipótese nula.

Teste χ^2 em dados de contagem unidimensionais

Pearson sugeriu, no início do século XX, a

Estatística de teste

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

que é uma medida do afastamento entre os dados e a hipótese nula. Quanto maior for o valor observado de X^2 , menos plausível é a hipótese nula. Por isso a região crítica é unilateral direita. Para definir a região crítica ou de rejeição (ou o p -value) é necessário conhecer a distribuição por amostragem de X^2 , no caso de H_0 ser válida.

Distribuição por amostragem

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)} \quad \text{sob } H_0$$

Teste χ^2 em dados de contagem unidimensionais

A distribuição χ^2 é **assintótica**, isto é, é uma distribuição aproximada, válida apenas para amostras “grandes”. De acordo com Cochran, o critério de validade da distribuição χ^2 é:

Critério de Cochran

- nenhum E_j inferior a 1
- não mais de 20% dos E_j 's inferiores a 5.

Note-se que o critério se aplica às **frequências esperadas** e não às observadas. Quando o critério não se verifica, agrupam-se categorias de forma a atingir as frequências mínimas requeridas.

Teste χ^2 em dados de contagem unidimensionais

O teste χ^2 em dados de contagem unidimensionais pode resumir-se na seguinte tabela:

categoria	A_1	A_2	\dots	A_k	Total
Probabilidades sob H_0	π_1	π_2	\dots	π_k	1
Frequências esperadas	E_1	E_2	\dots	E_k	N
Frequências observadas	O_1	O_2	\dots	O_k	N
Contribuição para a E.T.	$\frac{(O_1 - E_1)^2}{E_1}$	$\frac{(O_2 - E_2)^2}{E_2}$	\dots	$\frac{(O_k - E_k)^2}{E_k}$	χ^2_{calc}

Exemplo 11 | hipótese genética

A descendência originada pelo cruzamento de dois tipos de plantas pode ser qualquer um dos três genótipos cc , cC e CC . Um modelo teórico de sucessão genética indica que os tipos cc , cC e CC devem aparecer na razão $1 : 2 : 1$. Efetuou-se o cruzamento daqueles dois tipos tendo-se classificado 90 plantas. A sua classificação genética foi registada na tabela:

Genótipos	cc	cC	CC
Num. plantas	18	44	28

Questão: Estão estes dados de acordo com o modelo genético?

Ou: $P(cc) = \frac{1}{4}$ e $P(cC) = \frac{2}{4}$ e $P(CC) = \frac{1}{4}$?

Exemplo 11 | hipótese genética

Teste χ^2 de Pearson considerando as $k = 3$ categorias para os genótipos:

- $H_0: \pi_1 = 0.25, \pi_2 = 0.5, \pi_3 = 0.25$ vs H_1 : probab. diferentes
- Estatística de teste:

$$\chi^2 = \sum_{i=1}^3 \frac{(\mathcal{O}_i - E_i)^2}{E_i} \sim \chi_{(2)}^2 \quad \text{sob } H_0$$

em que \mathcal{O}_i é a frequência observada da i -ésima categoria e $E_i = N\pi_i = 90\pi_i$ é a frequência esperada. A distribuição $\chi_{(2)}^2$ só é válida para amostras grandes.

- Região crítica: ao nível de significância $\alpha = P[\text{rejeitar } H_0 \mid H_0 \text{ verdadeira}] = 0.05$, rejeita-se H_0 se $X_{\text{calc}}^2 > \chi_{0.05(2)}^2 = 5.991$.
- Tabela resumo:

genótipo	cc	cC	CC	Total
π_i	0.25	0.5	0.25	1
E_i	22.5	45	22.5	90
\mathcal{O}_i	18	44	28	90
$\frac{(\mathcal{O}_i - E_i)^2}{E_i}$	0.9000	0.0222	1.3444	$X_{\text{calc}}^2 = 2.2667$

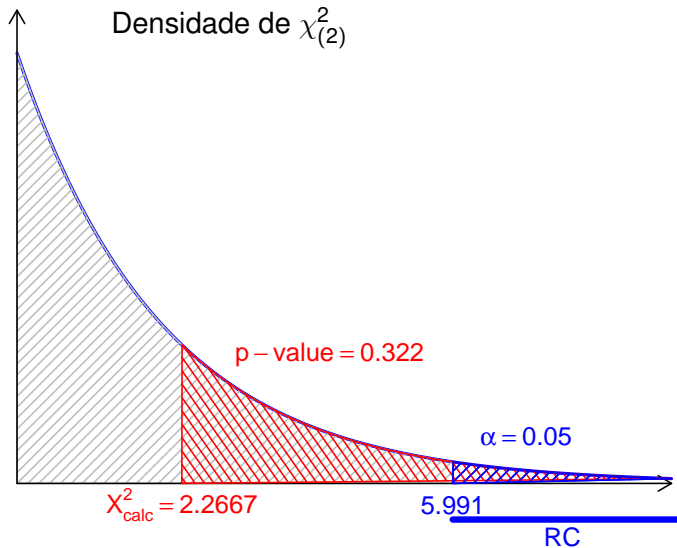
Exemplo 11 | hipótese genética

- Todas as frequências esperadas são superiores a 5, o que garante a validade da distribuição assintótica ao abrigo do critério de Cochran.
- Como $X^2_{\text{calc}} \notin \text{RC}$, não se rejeita H_0 ao nível de significância de 5%, não havendo razões para duvidar da hipótese genética que prevê aquelas proporções para os genótipos.

Em Python, com a biblioteca `scipy`:

```
>>> import numpy as np
>>> from scipy.stats import chisquare
>>> Oi=[18,44,28]
>>> pi=[0.25,0.5,0.25]
>>> Ei=np.array(pi)*sum(Oi)
>>> res=chisquare(f_obs=Oi, f_exp=Ei)
>>> print('X2cal=', res[0])
X2cal= 2.2666666666666666
>>> print('pvalue=', res[1])
pvalue= 0.3219582715376759
```

Exemplo 11 | hipótese genética



Exemplo 12 | hipótese de distribuição uniforme

Na ausência de uma teoria bem definida, pode-se admitir à partida que todas as categorias são igualmente frequentes na população. Por exemplo, pode esperar-se que o número de insectos apanhados numa armadilha a diferentes horas do dia seja constante; ou que a frequência de avistamento de um animal em diferentes habitats seja também contante. Por outras palavras, na ausência de informação, usa-se a **distribuição uniforme**, em que a frequência esperada de cada uma de k categorias é calculada como $E_i = N \times \frac{1}{k} = \frac{N}{k}$.

Exemplo 12 | distribuição uniforme

Uma empresa de comida para cães pretende testar 3 novos sabores de comida para cães para decidir se exclui algum sabor da sua futura produção ou se produz os 3 sabores.

Seleciona uma amostra aleatória de 75 cães, oferece a cada um os 3 sabores e anota a preferência de cada cão. Os resultados foram:

Sabor escolhido	Explosão de Alho	Delícia de Mirtilo	Frescura de Menta
Núm. cães	22	30	23

- 1 Os 75 cães que constituem a amostra têm igual preferência pelos 3 sabores?
- 2 Pode-se concluir que a amostra foi retirada de uma população de cães que escolhem os 3 sabores com igual frequência?
- 3 Baseado nos resultados desta experiência, o presidente da empresa deve eliminar os sabores Explosão de Alho e Frescura de Menta?

Exemplo 13 | ajustamento à Binomial

No controlo de qualidade de uma linha de produção de latas de cerveja, contou-se o número de latas impróprias em cada um de 200 *packs* de 6 latas. Os resultados foram:

Núm. latas impróprias	0	1	2	3	4	5	6
Núm. <i>packs</i>	141	48	9	2	0	0	0

Questão: É admissível que o número de latas impróprias por *pack* siga uma lei Binomial com probabilidade de sucesso $p = 0.04$?

Note-se que cada uma das 200 contagens corresponde ao resultado de repetir 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.

Exemplo 13 | ajustamento à Binomial

Poderá realizar-se um teste de χ^2 de Pearson tomando como “categorias” os valores possíveis para a variável. Neste caso trata-se de um teste de ajustamento a uma distribuição discreta.

Teste χ^2 de Pearson :

- $H_0: X \sim B(6, 0.04)$ vs $H_1: X \not\sim B(6, 0.04)$

- Estatística de teste:

$$\chi^2 = \sum_{i=1}^7 \frac{(\mathcal{O}_i - E_i)^2}{E_i} \sim \chi_{(6)}^2 \quad \text{sob } H_0$$

em que \mathcal{O}_i é a frequência observada do i -ésimo valor da v.a. X e $E_i = N\pi_i = 200\pi_i$ é a frequência esperada. As probabilidades sob H_0 são $P[X = x] = \binom{6}{x} p^x (1-p)^{6-x}$, $x = 0, 1, \dots, 6$. A validade da distribuição $\chi_{(6)}^2$ pode ser apreciada através do critério de Cochran.

Exemplo 13 | ajustamento à Binomial

- Tabela resumo:

x_j	0	1	2	3	4	5	6	Total
π_j	0.7828	0.1957	0.0204	0.0011	0.0000	0	0	1
E_j	156.552	39.138	4.077	0.226	0.007	0	0	200
O_j	141	48	9	2	0	0	0	200

O critério de Cochran não é válido pois há classes com a frequência esperada inferior a 1. **Agrupando as 4 últimas classes**, o critério é aproximadamente válido:

x_j	0	1	≥ 2	Total
π_j	0.7828	0.1957	0.0216	1
E_j	156.5516	39.1379	4.3106	200
O_j	141	48	11	200

Exemplo 13 | ajustamento à Binomial

Será necessário redefinir a estatística de teste:

- Estatística de teste:

$$X^2 = \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} \sim \chi_{(2)}^2 \quad \text{sob } H_0$$

- Região crítica: ao nível de significância $\alpha = 0.05$, rejeita-se H_0 se $X_{\text{calc}}^2 > \chi_{0.05(2)}^2 = 5.991$.
- Tabela resumo:

x_i	0	1	≥ 2	Total
π_i	0.7828	0.1957	0.0216	1
E_i	156.5516	39.1379	4.3106	200
O_i	141	48	11	200
$\frac{(O_i - E_i)^2}{E_i}$	1.5449	2.0067	10.3812	$X_{\text{calc}}^2 = 13.9328$

Exemplo 13 | ajustamento à Binomial

- Como $X_{\text{calc}}^2 = 13.9328 > 5.991$, rejeita-se H_0 , optando-se por H_1 : $X \not\sim B(6, 0.04)$, ou seja pela hipótese de X não ter distribuição binomial ou ter distribuição binomial com outro valor de p . A categoria “ ≥ 2 ” contribui para a estatística de teste com uma parcela com um valor muito elevado que, por si só, levaria à rejeição de H_0 .

Em Python, com este *script*

```
import numpy as np
from scipy import stats
O_i=[141,48,11]
pi=list(stats.binom.pmf([0,1],6,0.04))+
     [1-stats.binom.cdf(1,6,0.04)]
E_i=np.array(pi)*sum(O_i)
print(stats.chisquare(f_obs=O_i, f_exp=E_i))

obtém-se o resultado
Power_divergenceResult(statistic=13.932738649668284,
pvalue=0.0009430706809671183)
```

Exemplo 14 | ajustamento à Poisson

Num ensaio sobre pereiras Rocha, foi testado o sistema de condução Tatura. Foi observado o número de gomos florais em 80 pereiras, tendo-se obtido as contagens abaixo indicadas. Pretende-se saber se é possível considerar que o número de gomos por árvore segue uma lei Poisson, com valor esperado 7.

No. de gomos	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. de árvores	0	1	3	9	6	15	4	8	7	6	4	6	3	4	3	1

Para que as probabilidades, ao abrigo da hipótese de validade da Poisson, somem 1 é necessário acrescentar a categoria “ ≥ 16 ”.

Sendo $X \sim \mathcal{P}(7)$, $P[X = x] = \frac{e^{-7}7^x}{x!}$, $x = 0, 1, 2, \dots$. As frequências esperadas para as duas primeiras categorias são $E_1 = 80 P[X = 0] = 0.073$ e $E_2 = 0.51$, ambas < 1 . Há também demasiadas categorias com $E_i < 5$. É portanto necessário agrupar categorias para que se verifique o critério de Cochran. Sugere-se o seguinte agrupamento em $k = 9$ categorias:

No. de gomos	≤ 3	4	5	6	7	8	9	10	≥ 11
No. de árvores	13	6	15	4	8	7	6	4	17

Exemplo 14 | ajustamento à Poisson

Teste de ajustamento χ^2 de Pearson :

- $H_0: X \sim \mathcal{P}(7)$ vs $H_1: X \not\sim \mathcal{P}(7)$
- Estatística de teste:

$$\chi^2 = \sum_{i=1}^9 \frac{(\mathcal{O}_i - E_i)^2}{E_i} \sim \chi_{(8)}^2 \quad \text{sob } H_0$$

em que \mathcal{O}_i é a frequência observada da i -ésima categoria e $E_i = N\pi_i = 80\pi_i$ é a frequência esperada. As probabilidades sob H_0 são $P[X = x] = \frac{e^{-7}7^x}{x!}$, $x = 0, 1, \dots$. A validade da distribuição $\chi_{(8)}^2$ é justificada pela validade do critério de Cochran (no quadro abaixo).

- Região crítica: ao nível de significância $\alpha = 0.05$, rejeita-se H_0 se $X_{\text{calc}}^2 > \chi_{0.05(8)}^2 = 15.5043$.

Exemplo 14 | ajustamento à Poisson

- Tabela resumo¹:

x_i	≤ 3	4	5	6	7	8	9	10	≥ 11	Total
π_i	0.082	0.091	0.128	0.149	0.149	0.130	0.101	0.071	0.099	1
E_i	6.541	7.30	10.22	11.92	11.92	10.43	8.11	5.68	7.88	80
O_i	13	6	15	4	8	7	6	4	17	80
$\frac{(O_i - E_i)^2}{E_i}$	6.377	0.231	2.239	5.262	1.289	1.128	0.550	0.496	10.549	28.121

- $\chi^2_{\text{calc}} = 28.121$ pertence à região crítica, logo rejeita-se a hipótese de que X tenha distribuição Poisson com valor esperado 7, ao nível de significância de 5 %. A última parcela da estatística de teste tem um valor muito elevado, contribuindo fortemente para a rejeição de H_0 .

¹Os valores de E_i (frequências esperadas) foram obtidos utilizando mais casas decimais em π_i do que as mostradas.

Exemplo 14 | ajustamento à Poisson

Em Python:

```
import numpy as np
from scipy import stats
Oi=[13, 6, 15, 4, 8, 7, 6, 4, 17]
pi=[stats.poisson.cdf(3,7)]+
    list(stats.poisson.pmf(range(4,11),7))+
    [1-stats.poisson.cdf(10,7)]
Ei=np.array(pi)*sum(Oi)
print(stats.chisquare(f_obs=Oi, f_exp=Ei))
```

Resultado:

```
Power_divergenceResult(statistic=28.12210172199837,
pvalue=0.00045158002259350095)
```

Nota:

Em vez de se admitir um valor para o parâmetro p da distribuição binomial ou para o parâmetro λ da distribuição de Poisson, poder-se-ia, em alternativa, estimar-se p e λ a partir dos dados da amostra. Quando se usam estimativas de parâmetros para estimar as frequências esperadas, a distribuição da estatística do teste de Pearson ainda é assintoticamente χ^2 mas o número de graus de liberdade terá que ser adaptado ao número de parâmetros estimados. Este assunto não vai ser desenvolvido.

Testes χ^2 em tabelas de contingência

Uma **tabela de contingência** é uma matriz onde se representam **frequências absolutas de pares** de categorias de dois atributos qualitativos ou discretos. Existem **duas características**, A e B , em que A pode tomar a valores distintos e B pode assumir b valores distintos. Na tabela

	B_1	B_2	\dots	B_b	Soma
A_1	O_{11}	O_{12}	\dots	O_{1b}	$N_{1\cdot}$
A_2	O_{21}	O_{22}	\dots	O_{2b}	$N_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	
A_a	O_{a1}	O_{a2}	\dots	O_{ab}	$N_{a\cdot}$
Soma	$N_{\cdot 1}$	$N_{\cdot 2}$	\vdots	$N_{\cdot b}$	N

O_{ij} representa a frequência observada do par (i -ésima categoria de A , j -ésima categoria de B). $N_{i\cdot}$ e $N_{\cdot j}$ são as frequências marginais observadas, $N_{i\cdot} = \sum_{j=1}^b O_{ij}$ e $N_{\cdot j} = \sum_{i=1}^a O_{ij}$.

Testes χ^2 em tabelas de contingência

Nos testes χ^2 em tabelas de contingência pretende-se **testar hipóteses sobre as probabilidades conjuntas $P(A_i, B_j)$** (desconhecidas) de o atributo A pertencer à categoria A_i e o atributo B pertencer à categoria B_j . Designam-se π_{ij} ($i = 1, \dots, a, j = 1, \dots, b$) essas probabilidades sob a hipótese nula.

- As probabilidades π_{ij} **podem ser totalmente especificadas** por alguma hipótese (genética, por exemplo) - **Situação 1**

Duas situações com interesse em que π_{ij} têm que ser estimadas são:

- **Teste de independência - Situação 2**
- **Teste de homogeneidade - Situação 3**

Situação 1: π_{ij} especificadas por uma hipótese

Considerando um conjunto de valores π_{ij} ($i = 1, \dots, a, j = 1, \dots, b$) tais que $\sum_{i=1}^a \sum_{j=1}^b \pi_{ij} = 1$, pretende-se testar as hipóteses:

$H_0: P(A_i, B_j) = \pi_{ij} \forall (i, j)$ versus $H_1: P(A_i, B_j) \neq \pi_{ij}$ para algum (i, j)

Recolhe-se uma amostra de dimensão N e conta-se o número de elementos da amostra que pertencem a cada par de categorias. Para o par (A_i, B_j) , seja

O_{ij} : a frequência observada na amostra;

$E_{ij} = N\pi_{ij}$: a frequência esperada ao abrigo da hipótese nula.

Situação 1: π_{ij} especificadas por uma hipótese

A estatística do teste e a sua distribuição por amostragem são idênticas ao caso unidimensional:

Estatística de teste

$$\chi^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(\mathcal{O}_{ij} - E_{ij})^2}{E_{ij}}$$

Distribuição por amostragem

$$\chi^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(\mathcal{O}_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(ab-1)} \quad \text{sob } H_0$$

A validade da distribuição assintótica depende da verificação de um critério que define quando é que uma amostra é considerada suficientemente grande. Pode-se aplicar o critério de Cochran às frequências esperadas E_{ij} de cada célula da matriz.

Situação 1: π_{ij} especificadas por uma hipótese

O teste χ^2 em tabelas de contingência pode resumir-se na seguinte tabela:

	B_1	B_2	\dots	B_b	Soma
A_1	$O_{11} (E_{11})$	$O_{12} (E_{12})$	\dots	$O_{1b} (E_{1b})$	$N_{1.}$
A_2	$O_{21} (E_{21})$	$O_{22} (E_{22})$	\dots	$O_{2b} (E_{2b})$	$N_{2.}$
\vdots	\vdots	\vdots	\vdots	\vdots	
A_a	$O_{a1} (E_{a1})$	$O_{a2} (E_{a2})$	\dots	$O_{ab} (E_{ab})$	$N_{a.}$
Soma	$N_{.1}$	$N_{.2}$	\vdots	$N_{.b}$	N

Exemplo 15 | hipótese genética

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

Para avaliar a hipótese de segregação independente e dominância / recessividade dos genes, realiza-se uma experiência cruzando coelhos de uma população inicial, que são heterozigóticos nos dois genes, i.e., têm um alelo de cada cor e um alelo de cada tipo de pelagem.

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

Cor	Tipo	
	Normal	Rex
Cinzento	134	44
Branco	42	12



coelho Branco Rex

Exemplo 15 | hipótese genética

Na hipótese de segregação independente e dominância / recessividade dos 2 genes, espera-se que a proporção de coelhos na descendência com cada par de características seja: 9 : 3 : 3 : 1 de, respetivamente, coelhos cinzentos de pelagem normal : coelhos cinzentos de pelagem Rex : coelhos brancos de pelagem normal : coelhos brancos de pelagem Rex. Portanto as hipóteses do teste são:

- $H_0: \pi_{11} = \frac{9}{16}, \pi_{12} = \pi_{21} = \frac{3}{16}, \pi_{22} = \frac{1}{16}$ versus $H_1: \text{algum } \pi_{ij} \text{ diferente}$

- A estatística do teste é: $X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(4-1)}^2$ sob H_0 .

As frequências esperadas obtêm-se multiplicando as probabilidades pelo número total de observações, $E_{ij} = 232\pi_{ij}$.

- H_0 é rejeitada para valores elevados de discrepância entre frequências observadas e esperadas, a região crítica é unilateral direita. Ao nível de significância $\alpha = 0.05$, rejeita-se H_0 se $X_{\text{calc}}^2 > \chi_{0.05(3)}^2 = 7.8147$.
 $\text{chi2.ppf}(1-0.05, 3)$

Exemplo 15 | hipótese genética

- Tabela resumo:

Cor	Tipo	
	Normal	Rex
Cinzento	134 (130.5)	44 (43.5)
Branco	42 (43.5)	12 (14.5)

- O valor da estatística de teste é então:

$$\chi^2_{\text{calc}} = \frac{(134 - 130.5)^2}{130.5} + \frac{(44 - 43.5)^2}{43.5} + \frac{(42 - 43.5)^2}{43.5} + \frac{(12 - 14.5)^2}{14.5} = 0.5824$$

- Não se rejeita H_0 , i.e., não se rejeitam as hipóteses genéticas referidas (dominância/recessividade e segregação independente dos genes), ao nível de significância de 5%.

O *p-value* é 0.9005 ($1 - \text{chi2.cdf}(0.5824, 3)$)

- Em Python:

```
Oi=[134,44,42,12]
pi=np.array([9,3,3,1])/16
Ei=pi*sum(Oi)
print(chisquare(f_obs=Oi, f_exp=Ei))
```

Situação 2: Teste de independência

Os **testes de independência** têm por objetivo averiguar a existência de uma associação entre dois atributos A e B . A hipótese testada é a hipótese de independência. Se os atributos são independentes, as probabilidades conjuntas são iguais aos produtos das probabilidades marginais:

$$P(A_i, B_j) = P(A_i)P(B_j), \forall (i, j)$$

Portanto as hipóteses em confronto no teste são:

$$H_0: \pi_{ij} = \pi_{i.} \pi_{.j} \forall (i, j) \quad \text{versus} \quad H_1: \text{algum } \pi_{ij} \neq \pi_{i.} \pi_{.j}$$

em que $\pi_{i.} = \sum_{j=1}^b \pi_{ij}$ e $\pi_{.j} = \sum_{i=1}^a \pi_{ij}$ são as probabilidades marginais, em geral desconhecidas.

Este teste difere do anterior no sentido em que a hipótese nula do teste não está completamente especificada.

Situação 2: Teste de independência

As **probabilidades marginais são estimadas** a partir dos dados. Se os dados estão representados numa tabela de contingência como a do slide 261, as probabilidades marginais são estimadas por

$$\hat{\pi}_{i.} = \frac{N_{i.}}{N}, (i = 1, \dots, a) \quad \text{e} \quad \hat{\pi}_{.j} = \frac{N_{.j}}{N}, (j = 1, \dots, b)$$

Assim, as **frequências esperadas** ao abrigo de H_0 , $E_{ij} = N\pi_{ij} = N\pi_{i.}\pi_{.j}$, **são estimadas** como

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

Situação 2: Teste de independência

A estatística do teste tem a expressão idêntica à da Situação 1, a distribuição por amostragem é assintoticamente χ^2 , mas os **graus de liberdade** são diferentes devido à necessidade de estimação de parâmetros:

Estatística do teste e distribuição por amostragem

$$\chi^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_{(a-1)(b-1)} \quad \text{sob } H_0$$

Exemplo 16 | Teste de independência

Análises qualitativas sugerem que a distribuição da vegetação herbácea em torno de árvores isoladas, em climas áridos, depende da orientação geográfica. Pretende-se validar estatisticamente esta suspeita, para um dado ecossistema.



Tribulus terrestris



Zygochloa simplex



Aristida adsencionis

Exemplo 16 | Teste de independência

Contabilizou-se (até totalizar 1000) o número de plantas de três espécies herbáceas que germinaram em torno de árvores isoladas de *Acacia tortilis*, em cada um dos quadrantes com orientação Norte, Este, Sul e Oeste. Obtiveram-se os seguintes resultados:

Espécie	Orientação geográfica				$N_{.j}$
	Norte	Sul	Este	Oeste	
<i>Tribulus terrestris</i>	4	157	12	28	201
<i>Zygophyllum simplex</i>	150	243	26	47	466
<i>Aristida adsencionis</i>	47	73	27	186	333
$N_{.j}$	201	473	65	261	$N = 1000$

Pretende-se testar se existe **independência** entre os 2 atributos: Espécie (com $a = 3$ categorias) e Orientação (com $b = 4$ categorias).

Exemplo 16 | Teste de independência

Teste de hipóteses:

- $H_0: \pi_{ij} = \pi_i \times \pi_j, \forall i = 1, 2, 3, j = 1, 2, 3, 4$ vs. $H_1: \exists(i, j): \pi_{ij} \neq \pi_i \cdot \pi_j$
 π_{ij} é a probabilidade de uma planta ser da i -ésima espécie e ter germinado num quadrante com a j -ésima orientação
 π_i é a probabilidade de uma planta escolhida ao acaso ser da i -ésima espécie
 π_j é a probabilidade de uma planta (de qualquer espécie) ter germinado num quadrante na j -ésima orientação
As probabilidades marginais têm que ser estimadas a partir dos dados.

- Estatística de teste: $X^2 = \sum_{i=1}^3 \sum_{j=1}^4 \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi_{(6)}^2$ sob H_0 ,

O_{ij} é o número de plantas da i -ésima espécie observadas num quadrante com a j -ésima orientação

$\hat{E}_{ij} = N \times \hat{\pi}_i \times \hat{\pi}_j = \frac{N_i \cdot N_j}{1000}$ é a frequência esperada estimada.

Exemplo 16 | Teste de independência

O menor valor de \hat{E}_{ij} é $\hat{E}_{13} = \frac{201 \times 65}{1000} = 13.065$, que é superior a 5, o que justifica a validade da distribuição assintótica, de acordo com o critério de Cochran.

- Região crítica (unilateral direita): ao nível de significância $\alpha = 0.05$, rejeita-se H_0 se $X_{\text{calc}}^2 > \chi_{0.05(6)}^2 = 12.59$.
- $X_{\text{calc}}^2 = 332.79 \gg 12.59$. Trata-se de um valor muito elevado, que leva à rejeição clara da hipótese de independência entre espécie e orientação geográfica.
- $p\text{-value} \simeq 0$ ($1 - \text{chi2.cdf}(332.79, 6)$)
- Dada a rejeição da hipótese de independência, é interessante investigar quais são as associações significativas entre espécie e orientação.

Exemplo 16 | Teste de independência

Por exemplo, para a espécie *Zygophyllum simplex*, tem-se

<i>Zygophyllum simplex</i>	Orientação geográfica				Total
	Norte	Sul	Este	Oeste	
O_{2j}	150	243	26	47	466
\hat{E}_{2j}	93.666	220.418	30.290	121.626	333
$\frac{(O_{2j} - \hat{E}_{2j})^2}{\hat{E}_{2j}}$	33.881	2.314	0.608	45.788	82.591

Só esta espécie contribui com duas parcelas para a estatística de teste que seriam, por si só, suficientes para rejeitar a hipótese nula (cada uma é > 12.59). Note-se que $O_{21} \gg \hat{E}_{21}$, o que indica uma associação positiva entre esta espécie e a orientação Norte. Pelo contrário, $O_{24} \ll \hat{E}_{24}$, o que indica uma associação negativa entre a espécie e a orientação Oeste.

Exemplo 16 | Teste de independência

Em Python:

```
Oij = np.array([[4, 157, 12, 28],
                [150, 243, 26, 47],
                [47, 73, 27, 186]])
# Executar o teste do qui-quadrado
X2calc, pval, gl, Eij = chi2_contingency(Oij)
# Mostrar os resultados
print('Qui-quadrado: ', X2calc)
print('p-value: ', pval)
print('Graus de liberdade:', gl)
print('Frequências esperadas:\n', Eij)
```

Resultado:

```
Qui-quadrado: 332.786020182768
p-value: 7.636494289069907e-69
Graus de liberdade : 6
Frequências esperadas:
[[ 40.401  95.073  13.065  52.461]
 [ 93.666 220.418  30.29  121.626]
 [ 66.933 157.509  21.645  86.913]]
```

Situação 3: Teste de homogeneidade

Suponha-se que uma população é classificada em a subpopulações, de acordo com as categorias do atributo A (A_1, A_2, \dots, A_a). O objetivo do teste de homogeneidade é o de averiguar se a distribuição do atributo B é idêntica (homogénea) para todas as subpopulações, ou seja, pretende-se averiguar se

$$P(B_j|A_1) = P(B_j|A_2) = \dots = P(B_j|A_a), \forall j = 1, 2, \dots, b$$

Designando por $\pi_{j|i}$ a probabilidade de uma observação da subpopulação A_i ser classificada na categoria B_j , as hipóteses em confronto no teste são:

$$H_0 : \begin{cases} \pi_{1|1} = \pi_{1|2} = \dots = \pi_{1|a} [= \pi_{.1}] \\ \pi_{2|1} = \pi_{2|2} = \dots = \pi_{2|a} [= \pi_{.2}] \\ \vdots \\ \pi_{b|1} = \pi_{b|2} = \dots = \pi_{b|a} [= \pi_{.b}] \end{cases} \quad \text{vs} \quad H_1 : \text{alguma igualdade falha}$$

Situação 3: Teste de homogeneidade

Na disposição do slide 261, a hipótese nula é a de as probabilidades condicionais serem iguais em cada coluna, ou seja de homogeneidade de cada coluna (atributo B), através das linhas (subpopulações).

As probabilidades marginais que intervêm em H_0 não são conhecidas e terão que **ser estimadas** a partir dos dados. Neste caso, os dados são constituídos por N_1 elementos da subpopulação A_1 , N_2

elementos da subpopulação A_2, \dots, N_a elementos da subpopulação A_a , ou seja os **totais por linha na tabela de contingência do slide 261 são previamente fixados**. Assim, as **probabilidades marginais de B são estimadas** por

$$\hat{\pi}_{.j} = \frac{N_{.j}}{N}, \quad (j = 1, \dots, b)$$

As frequências esperadas ao abrigo de H_0 são $E_{ij} = N_{i.} \pi_{j|i} = N_{i.} \pi_{.j}$.

As **frequências esperadas estimadas** são $\hat{E}_{ij} = N_{i.} \hat{\pi}_{.j} = N_{i.} \frac{N_{.j}}{N}$.

Situação 3: Teste de homogeneidade

A estatística de teste tem a mesma expressão da Situação 2 e tem a mesma distribuição assintótica (slide 271), isto é,

Estatística do teste e distribuição por amostragem

$$\chi^2 = \sum_{i=1}^a \sum_{j=1}^b \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi_{(a-1)(b-1)}^2 \quad \text{sob } H_0$$

Exemplo 17 | Teste de homogeneidade

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

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

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



Em cada tipo de solo foram recolhidas 100 larvas, que foram classificadas de acordo com a respetiva espécie (atributo B , com $b = 4$ categorias).



Exemplo 17 | Teste de homogeneidade

As frequências observadas das espécies de larvas, para cada tipo de solo, foram:

Tipo de solo	Espécie de larva				$N_{j\cdot}$
	1	2	3	4	
Arenoso	27	24	23	26	100
Limoso	20	32	18	30	100
Argiloso	13	37	16	34	100
$N_{\cdot j}$	60	93	57	90	$N = 300$

O objetivo é averiguar se cada espécie de larva se distribui de forma análoga (homogénea) pelos 3 tipos de solo, ou seja

$$P(j|\text{solo arenoso}) = P(j|\text{solo limoso}) = P(j|\text{solo argiloso}) \quad \forall j = 1, 2, 3, 4.$$

Exemplo 17 | Teste de homogeneidade

Teste de hipóteses (teste de homogeneidade)

- A hipótese nula é: a probabilidade de uma larva encontrada em cada tipo de solo ser da espécie j , ($j = 1, \dots, 4$) é igual para todos os tipos de solo, sendo também igual à probabilidade de a larva ser da espécie da j , independentemente do solo.

$$H_0 : \begin{cases} \pi_{1|\text{solo arenoso}} = \pi_{1|\text{solo limoso}} = \pi_{1|\text{solo argiloso}} [= \pi_{.1}] \\ \pi_{2|\text{solo arenoso}} = \pi_{2|\text{solo limoso}} = \pi_{2|\text{solo argiloso}} [= \pi_{.2}] \\ \pi_{3|\text{solo arenoso}} = \pi_{3|\text{solo limoso}} = \pi_{3|\text{solo argiloso}} [= \pi_{.3}] \\ \pi_{4|\text{solo arenoso}} = \pi_{4|\text{solo limoso}} = \pi_{4|\text{solo argiloso}} [= \pi_{.4}] \end{cases}$$

H_1 : alguma igualdade falha

- Estatística do teste: $X^2 = \sum_{i=1}^3 \sum_{j=1}^4 \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi_{(6)}^2$ sob H_0 ,

O_{ij} é o número de larvas da espécie j observadas no i -ésimo solo

$$\hat{E}_{ij} = N_i \cdot \hat{\pi}_{.j} = 100 \times \frac{N_{.j}}{300}, \quad \hat{E}_{i1} = \frac{100 \times 60}{300}, \quad \hat{E}_{i2} = \frac{100 \times 93}{300}, \quad \hat{E}_{i3} = \frac{100 \times 57}{300}, \\ \hat{E}_{i4} = \frac{100 \times 90}{300}.$$

Exemplo 17 | Teste de homogeneidade

- Região crítica (unilateral direita): ao nível de significância $\alpha = 0.05$, rejeita-se H_0 se $\chi^2_{\text{calc}} > \chi^2_{0.05(6)} = 12.591$.
- Tabela resumo

Tipo de solo	Espécie de larva				$N_{i.}$
	1	2	3	4	
Arenoso	27 (20)	24 (31)	23 (19)	26 (30)	100
Limoso	20 (20)	32 (31)	18 (19)	30 (30)	100
Argiloso	13 (20)	37 (31)	16 (19)	34 (30)	100
$N_{.j}$	60	93	57	90	$N = 300$

Todas as frequências esperadas são superiores a 5, pelo que o critério de Cochran é válido. Não é necessário agrupar células.

- $\chi^2_{\text{calc}} = 10.109$. Este valor não pertence à região crítica, portanto, ao nível de significância de 5 %, não se rejeita a hipótese de homogeneidade das distribuições de espécies de larva, nos três tipos de solos.
- $p\text{-value} = 0.1201$ ($1 - \text{chi2.cdf}(10.109, 6)$)

Exemplo 17 | Teste de homogeneidade

Em Python

```
import numpy as np
from scipy.stats import chi2_contingency
# Criar a matriz de frequências observadas
Oij = np.array([[27, 24, 23, 26],
                [20, 32, 18, 30],
                [13, 37, 16, 34]])
# Executar o teste do qui-quadrado
X2calc, pval, gl, Eij = chi2_contingency(Oij)
# Mostrar os resultados
print('Qui-quadrado: ', X2calc)
print('p-value: ', pval)
print('Graus de liberdade:', gl)
print('Frequências esperadas:\n', Eij)
```

Resultado:

Qui-quadrado: 10.109281267685342
p-value: 0.12012474641823215
Graus de liberdade : 6

Frequências esperadas:

```
[[20. 31. 19. 30.]
 [20. 31. 19. 30.]
 [20. 31. 19. 30.]
```

Nota sobre testes de independência *versus* homogeneidade

Os testes de independência e homogeneidade parecem idênticos na sua forma: partilham a mesma estatística de teste e a mesma distribuição de amostragem. A diferença encontra-se no delineamento das experiências.

No **teste de independência**, os elementos da amostra são extraídos ao acaso de uma população e dois atributos são observados para cada elemento. Só o número total de elementos da amostra é previamente fixado pelo experimentador.

No **teste de homogeneidade** os dados são selecionados ao acaso de cada subpopulação separadamente. Os números de elementos de cada subpopulação são previamente definidos pelo experimentador. A hipótese nula é a de cada subpopulação partilhar a mesma distribuição de um outro atributo.

As diferenças são subtis mas importantes: as hipóteses de referência são diferentes e as conclusões também são diferentes.