

ANADI

Análise de Dados em Informática

Aulas Teóricas - Testes de Hipóteses Paramétricos

Ana Madureira, João Matos

Instituto Superior de Engenharia do Porto

Ano letivo 2023/2024



- Por vezes pretende-se investigar se uma determinada afirmação sobre um determinado parâmetro de uma população é verdadeiro ou falso.
- Os **Testes de Hipóteses** permitem (com uma certa probabilidade) responder a esta questão recorrendo à informação contida em amostras aleatórias.

Exemplo:

- Suponha que pretendemos saber se, nas próximas eleições, um determinado partido político irá exceder os 30% ou irá ficar abaixo dos 30% .
- Então, geralmente, teremos que obter dados (amostra aleatória) que representem a população em causa (conjunto de votantes) e que forneçam informação relativa às proporções dos diferentes partidos políticos.
- Fazer a inferência estatística (tirar conclusões sobre o resultado das eleições a partir da informação obtida pela amostra)

- Dado um parâmetro desconhecido θ de uma população e um valor fixo θ_0 iremos considerar os seguintes três testes de hipótese

Caso	H_0	H_1	
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	teste bilateral
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	teste unilateral (à esquerda)
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	teste unilateral (à direita)

- Onde, H_0 é a hipótese nula e H_1 é hipótese alternativa.
- O θ_0 é um ponto fronteira entre H_0 e H_1 e por vezes escrevemos a hipótese nula nos casos unilaterais (casos (b) e (c)) na forma, $\theta = \theta_0$.

- Num teste de hipóteses podem ocorrer dois tipos de erros:

- ▶ H_0 é verdadeira, mas é rejeitada; **erro tipo I**
- ▶ H_0 não é rejeitada mas é verdadeira; **erro tipo II**

	H_0 é verdadeira	H_0 não é verdadeira
H_0 não é rejeitada	decisão correcta	erro tipo II
H_0 é rejeitada	erro tipo I	decisão correcta

Definição 1

O nível de significância α é a probabilidade de ocorrer um erro do tipo I.

$$\alpha = P(H_1 | H_0)$$



O erro do tipo II (β) depende de parâmetros desconhecidos das populações e do tamanho da amostra.

Como efetuar um teste paramétrico

- 1 Defina as suposições para as distribuições das variáveis aleatórias de interesse
- 2 Formule H_0 e H_1
- 3 Construa uma estatística teste $T(X) = T(X_1, X_2, \dots, X_n)$ onde se conhece a distribuição de T sob H_0
- 4 Calcule o valor da estatística teste $t(x) = t(x_1, x_2, \dots, x_n)$ (usando os valores da amostra $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$)
- 5 **Decidimos se rejeitamos H_0 , ou se não rejeitamos H_0 através de regiões de aceitação e de rejeição ou usando o valor de prova (p-value). Em ANADI iremos privilegiar a decisão usando o p-value.**



Como é usual, usamos letras maiúsculas para variáveis aleatórias e letras minúsculas para realizações de variáveis aleatórias.

Decisões baseadas no p-value


Definição 2 (p-value)


Define-se o **p-value** de um teste estatístico $T(X)$ da seguinte forma

Teste bilateral: $\text{p-value} = P(|T| \geq |t(x)| \mid H_0)$

Teste unilateral: $\text{p-value} = P(T \geq t(x) \mid H_0)$

$$\text{p-value} = P(T \leq t(x) \mid H_0)$$

 *Pode-se interpretar o p-value como sendo o menor nível de significância para o qual a H_0 é rejeitada para o valor observado da estatística teste.*

 *A regra de decisão é: Se $\text{p-value} < \alpha$ rejeita-se H_0 . Caso contrário não se rejeita H_0 .*

Teste para a média quando a variância é conhecida

- Neste teste pretendemos decidir se a média μ de uma variável aleatória X de uma população com distribuição normal $N(\mu, \sigma^2)$, com $\sigma^2 = \sigma_0^2$ conhecida, difere de um valor específico ($\mu = \mu_0$) ou é menor (ou maior) que μ_0 .

- (1) Assumimos que $X \sim N(\mu, \sigma_0^2)$, onde σ_0^2 é conhecida. Assumimos que é extraída uma amostra aleatória de n variáveis X_1, X_2, \dots, X_n i.i.d. (X_i s têm a mesma distribuição de X).
- (2) Definimos uma das três hipóteses H_0 e H_1

$H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ teste bilateral


$H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ teste unilateral à direita


$H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$ teste unilateral à esquerda

Teste para a média quando a variância é conhecida (cont.)

(3) Construir o teste estatístico: Sabe-se que, sob H_0 , tem-se

$$T(X) = \frac{\bar{X} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}} \sim N(0,1) \quad \text{onde,} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

 Se as variáveis aleatórias X_i s não forem normais mas o tamanho da **amostra for grande ($n \geq 30$)** podemos igualmente considerar que estatística teste $T(X) \sim N(0,1)$ (isto é consequência do **Teorema do limite central**)

 Logo este teste pode efectuar-se quando a população é normal ou a amostra tem tamanho igual ou superior a 30.

(4) Cálculo de $t(x)$ (valor observado da estatística teste)

$$t(x) = \frac{\bar{x} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}}, \quad \text{onde} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Teste para a média quando a variância é conhecida (cont.)

(5) Cálculo do p-value:

- ▶ Para testes bilaterais tem-se
$$\text{p-value} = P(|T(X)| \geq t(x) | H_0) = 2P(T(X) \geq |t(x)| | H_0)$$
- ▶ Para testes unilaterais à direita $\text{p-value} = P(T(X) \geq t(x) | H_0)$
- ▶ Para testes unilaterais à esquerda $\text{p-value} = P(T(X) \leq t(x) | H_0)$

Exemplo 1:

A empresa UNLOP garante que se os pneus forem utilizados em condições normais têm uma vida esperada superior a 40000 km. Conhece-se o desvio padrão $\sigma_0 = 8000 \text{ km}$ e a média amostral de uma amostra constituída por 30 pneus é $\bar{x} = 43200 \text{ km}$. Teste, ao nível de significância de 5% se os pneus têm a vida esperada reivindicada pelo fabricante.

Resolução: Tem-se o teste unilateral à direita:

$$H_0 : \mu \leq 40000 \text{ vs } H_1 : \mu > 40000$$

Exemplo 1 (cont)

O valor observado da estatística teste é, $t(x) = \frac{43200-40000}{\frac{8000}{\sqrt{30}}}$ O p-value é,

$$\text{p-value} = P(T(X) > 2.19089)$$

que, usando o Python, é calculado usando as instruções:

```
In [1]: import math
to=(43200-40000)/(8000/math.sqrt(30))
to
Out[1]: 2.1908902300206643
In [1]: import scipy.stats
pv = 1 - scipy.stats.norm(0, 1).cdf(to)
pv
Out[2]: 0.014229868458155326
```

Como o p-value (0.01422987) é inferior a $\alpha = 0.05$ tomamos a decisão de rejeitar a hipótese nula. Ou seja decidimos que os pneus possuem uma vida esperada superior a 40000 km com um nível de significância de 5% .

Teste para a média quando a variância é desconhecida

- Supondo que a **população é normal ou que a amostra é grande (≥ 30)** tem-se que a estatística teste $T(X)$ é dada por,

$$T(X) = \frac{\bar{X} - \mu_0}{\frac{S_X}{\sqrt{n}}} \sim T_{n-1}, \text{ com } S_X^2 = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{n-1}$$

onde T_{n-1} denota a **distribuição t-student com $n-1$ graus de liberdade**.

Exemplo 2.

Um director de qualidade pretende averiguar se o peso dos pacotes de açúcar produzidos corresponde ao valor indicado na embalagem (1 kilo). Supondo que o peso dos pacotes segue uma distribuição normal e que se conhece a amostra aleatória:

1,06 0,98 0,95 1,01 0,97 1,05 0,99 0,95 1,00 1.01

Será que, para $\alpha = 5\%$ se pode dizer que o peso médio corresponde ao peso de 1 kg indicado na embalagem?

Exemplo 2 (cont.)

Resolução: Trata-se do teste bilateral

$$H_0: \mu = 1 \quad \text{versus} \quad H_1: \mu \neq 1$$

Efectuando os cálculos no Python:

```
In [1]: import math
import scipy.stats as stats
import pandas as pd
amostra=pd.DataFrame({'dados':[1.02,0.98,0.97,1.01,0.97,1.02,0.99,0.98,1.00,1.01]})
n=len(amostra) # tamanho da amostra
xm=amostra.mean()['dados'] # media amostral
sx=amostra.std()['dados'] # desvio padrao amostral
mu=1 # media da hipotese nula
to = (xm-mu)/(sx/math.sqrt(n))
print('valor observado da estatistica teste:',to)
valor observado da estatistica teste: -0.8075728530872662
```

```
In [2]: pv=2*(1-stats.t.cdf(abs(to),n-1))
print('valor de prova:',pv)
valor de prova: 0.4401575296513358
```

Ou usando a função `stats.ttest_1samp()`

```
In [2]: amostral=amostra.to_numpy() # pandas.DataFrame to numpy.ndarray
t_statistic, p_value = stats.ttest_1samp(amostral, popmean=mu, alternative='two-sided')
print('valor observado da estatistica teste:',t_statistic)
print('valor de prova:',p_value)
valor observado da estatistica teste: [-0.80757285]
valor de prova: [0.44015753]
```

Exemplo 2 (cont.)

- No output: o `t_statistic` representa o valor observado da estatística teste, e o `p_value` representa o valor de prova.
- Note que os dois processos dão resultados iguais, a menos de arredondamentos.
- A função `scipy.stats.ttest_1samp()` é bastante versátil e dá também para efectuar testes bilaterais à esquerda e à direita dependendo da opção `alternative='greater'` ou `alternative='less'`.
- claro que, neste exemplo, o p-value é bastante alto (0.4402) então decidimos não rejeitar H_0 para todo o nível de significância inferior ao p-value.
- Chamamos a este teste de **t.teste à média de uma amostra** (One-Sample t-Test)

Teste para comparar duas médias (amostras independentes)

- Assumimos que as duas amostras são obtidas de duas variáveis aleatórias X e Y com distribuições $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ com tamanhos n_1 e n_2
- As **variáveis são normais** ($X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$)
- Podemos especificar uma das três hipóteses:
 - $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$ teste bilateral
 - $H_0 : \mu_X \leq \mu_Y$ versus $H_1 : \mu_X > \mu_Y$ teste bilateral à direita
 - $H_0 : \mu_X \geq \mu_Y$ versus $H_1 : \mu_X < \mu_Y$ teste bilateral à esquerda

Distinguimos três casos:

- σ_X^2 e σ_Y^2 são conhecidas
- σ_X^2 e σ_Y^2 são desconhecidas, mas assumem-se iguais ($\sigma_X^2 = \sigma_Y^2$)
- σ_X^2 e σ_Y^2 são desconhecidas, mas assumem-se diferentes ($\sigma_X^2 \neq \sigma_Y^2$)

Caso 1 : Variâncias conhecidas

A estatística teste é,

$$T(X,Y) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}}} \sim N(0,1).$$

Este teste é idêntico ao teste para a média quando a variância é conhecida (ver slide 9).

Caso 2 : Variâncias desconhecidas, mas iguais.

Seja $\sigma^2 = \sigma_X^2 = \sigma_Y^2$, a estatística teste é:

$$T(X, Y) = \frac{\bar{X} - \bar{Y}}{S} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \sim T_{n_1 + n_2 - 2}, \text{ onde, } S^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}$$

- Este teste é idêntico ao teste para a média de uma população quando a variância é conhecida
- Para se fazer este teste em Python usa-se a função (`ttest_ind()`) com a opção `equal_var=TRUE`
- Na prática para decidirmos se estamos no caso 2 ou no caso 3 temos que fazer um teste à igualdade de variâncias das duas amostras (p.e. um [teste de Levene](#)) que veremos mais à frente.

Caso 3 : Variâncias desconhecidas e diferentes (teste de Welsh)

A estatística é

$$T(X, Y) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2}}} \sim T_r$$

onde, os graus de liberdade r são determinados por

$$r = \left(\frac{s_X^2}{n_1} + \frac{s_Y^2}{n_2} \right)^2 / \left(\frac{(s_X^2/n_1)^2}{n_1 - 1} + \frac{(s_Y^2/n_2)^2}{n_2 - 1} \right)$$

Exemplo 3:

Pretende-se averiguar se as médias das notas do exame nacional de Matemática diferem entre os alunos e as alunas. Extraiu-se a seguinte amostra aleatória normal e variância diferentes entre o grupo dos alunos e das alunas.

Aluno	Nota (%)	Género	Aluno	Nota (%)	Género	Aluno	Nota (%)	Género
1	95	M	5	100	F	9	98	M
2	100	F	6	95	M	10	95	F
3	78	M	7	98	M	11	90	F
4	68	M	8	79	M	12	95	F

O teste a efectuar é o teste de Welsh às médias de duas amostras independentes.

Resolução: Seja μ_M a média das notas dos alunos e μ_F a média das notas das alunas. Tem-se:

$$H_0: \mu_M = \mu_F \text{ versus } H_1: \mu_M \neq \mu_F$$

Instruções no Python:

```
import numpy as np
import scipy.stats as stats
M = np.array([95,78,68,95,98,79,98])
F = np.array([100,100,95,90,95])
t_statistic, p_value = stats.ttest_ind(M,F,equal_var = False, alternative = 'two-sided')
print("valor de prova:", p_value)
valor de prova: 0.11605951353428898
```



Com este p -value (0.1160595) não se rejeita H_0 para $\alpha < 11.6\%$

Teste para comparar médias de duas amostras emparelhadas

- Supor que se tem duas variáveis aleatórias X e Y com $E(X) = \mu_X$ e $E(Y) = \mu_Y$.
- As amostras são emparelhadas quando "medimos" a mesma variável duas vezes relativamente ao mesmo assunto.
- Exemplos típicos:
 - ▶ medir o peso de um grupo de pessoas antes e depois de uma dieta
 - ▶ avaliação das despesas de agregados familiares em aparelhos electrónicos em dois anos consecutivos

Neste caso tem-se a estatística do teste

$$T(X, Y) = \frac{\bar{D}}{S_D / \sqrt{n}} \sim T_{n-1}, \text{ onde } D = X - Y$$

e

$$S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}$$

Exemplo 4: Suponha que 10 indivíduos, escolhidos aleatoriamente, foram submetidos a uma dieta. Recolheram-se os dados indicados na tabela (Peso A é o peso antes da dieta e o Peso D é o peso depois da dieta). Pretende-se verificar se houve diferença entre o peso médio antes e depois da dieta.

Pessoa (i)	1	2	3	4	5	6	7	8	9	10
Peso A (x_i)	88	99	102	118	78	94	70	71	84	68
Peso D (y_i)	86	100	99	116	76	93	68	65	85	66

Como os dados antes e depois da dietas são dependentes (peso medido nos mesmos indivíduos) o teste efectuado é um **t-test com amostras emparelhadas**. Usar função **scipy.stats.ttest_rel()**.

Instruções no Python:

```
import numpy as np
import scipy.stats as stats
Peso_A=np.array([88,99,102,118,78,94,70,71,84,68])
Peso_D=np.array([86,100,99,116,76,93,68,65,85,66])
resultado=stats.ttest_rel(Peso_A,Peso_D,alternative='two-sided')
print('valor de prova:',resultado.pvalue)
valor de prova: 0.018719423349875648
```



Rejeita-se que as médias são iguais (H_0) para um nível de significância superior a $\approx 2\%$

Teste Binomial para a proporção p (uma amostra)

- Seja X uma variável aleatória de Bernoulli $B(1; p)$ com $X \in \{0, 1\}$, onde p é a probabilidade de sucesso ($P(X = 1)$)
- Seja $X = (X_1, X_2, \dots, X_n)$ uma amostra de variáveis aleatórias i.i.d. (independentes e identicamente distribuídas com distribuição $B(1; p)$)
- Calculamos frequência relativa

$$\tilde{p} = \frac{1}{n} \sum_{i=1}^n X_i$$

que é um estimador, não enviesado, de p

- Consideramos os três testes,

Caso	H_0	H_1	
(a)	$p = p_0$	$p \neq p_0$	teste bilateral
(b)	$p \geq p_0$	$p < p_0$	teste unilateral (à esquerda)
(c)	$p \leq p_0$	$p > p_0$	teste unilateral (à direita)

- Iremos considerar duas abordagens:

- 1 Solução aproximada: usa-se para grandes amostras ($n \geq 30$ e $np(1-p) \geq 9$). (teste binomial aproximado)
- 2 Solução exata (teste binomial)

Teste binomial aproximado: Consideramos a estatística teste,

$$T(X) = \frac{\tilde{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$$

Exemplo 5:

Numa linha de montagem existe uma máquina que habitualmente sofre avarias em 5% dos turnos. Um técnico de produção decidiu verificar se esta percentagem está a ser cumprida. Para este efeito o técnico recorreu a uma amostra aleatória de 250 turnos em que se detectaram 20 turnos com avaria. Esclareça esta questão usando $\alpha = 0.05$

Exemplo 5 (cont.):

Do ponto de vista do técnico, iremos efetuar o teste unilateral

$$H_0 : p \leq 0.05 \text{ versus } H_1 : p > 0.05$$

As instruções em Python são:

```
from statsmodels.stats.proportion import proportions_ztest
nobs = 250; nsuc = 20; pnull = 0.05
res=proportions_ztest(nsuc,nobs, pnull, prop_var=pnull, alternative='larger')
print('z-value:',round(res[0],4))
print('valor de prova:',round(res[1],4))
z-value: 2.1764
valor de prova: 0.0148
```

Decisão: O p-value (0.0148) é menor que o índice de significância logo rejeitamos H_0 .

Teste binomial exato: A variável aleatória $Y = \sum_{i=1}^n X_i$ (que representa o número de sucessos),

$$T(X) = Y \sim B(n, p_0)$$

Exemplo 6 :

Uma fábrica produz determinado tipo de peças e sabe-se que a percentagem de peças defeituosas é de 20%. O director da linha de montagem procedeu a alguns ajustes no equipamento, com o objectivo de diminuir a percentagem de peças defeituosas. Recolheu-se uma amostra de 20 peças e verificou-se que duas eram defeituosas. Será que há evidências de mudança na percentagem de peças defeituosas?

Tem-se as hipóteses,

$$H_0 : p \geq 0.2 \text{ vs } H_1 : p < 0.2$$

Resolução "longa": O valor da estatística observado é 2 e o p-value:

$$P(Y \leq 2 \mid p_0 = 0.2) = \sum_{i=0}^2 \binom{20}{i} \times (0.2)^i \times (0.8)^{20-i}$$

Para calcular esta probabilidade usando o Python tem-se,

```
pvalue=stats.binom.cdf(2, 20, 0.2)
print('p-value (teste binomial exato)', pvalue)
p-value (teste binomial exato) 0.20608471894847413
```

Resolução "curta":

```
res=stats.binomtest(2,20,0.2, alternative='less ')
print('valor de prova:', res.pvalue)
```

valor de prova: 0.20608471894847413

Decisão: Não rejeitamos a hipótese nula (para todo o $\alpha \leq 0,2$).

Teste Binomial para duas proporções (duas amostras)

- Sejam X e Y duas amostras i.i.d. de duas distribuições de Bernoulli com parâmetros p_1 e p_2 respectivamente

$$X = (X_1, X_2, \dots, X_{n_1}), \quad X_i \sim B(1; p_1)$$

$$Y = (Y_1, Y_2, \dots, Y_{n_2}), \quad Y_i \sim B(1; p_2)$$

- E as somas (número de sucessos)

$$X = \sum_{i=1}^{n_1} X_i \sim B(n_1; p_1), \quad Y = \sum_{i=1}^{n_2} Y_i \sim B(n_2; p_2)$$

- Consideramos os três tipos de testes (semelhantes ao caso anterior)

Caso	H_0	H_1	
(a)	$p_1 = p_2$	$p_1 \neq p_2$	teste bilateral
(b)	$p_1 \geq p_2$	$p_1 < p_2$	teste unilateral (à esquerda)
(c)	$p_1 \leq p_2$	$p_1 > p_2$	teste unilateral (à direita)

- Para amostras de grandes dimensões tem-se que as variáveis aleatórias

$$\frac{X}{n_1} \underset{\sim}{\text{aprox.}} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \frac{Y}{n_2} \underset{\sim}{\text{aprox.}} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

- Agrupando as duas amostras estimamos p por

$$\hat{p} = \frac{X + Y}{n_1 + n_2}$$

e, para amostras grandes e se $np(1-p) > 9$ tem-se

$$D = \frac{X}{n_1} - \frac{Y}{n_2} \underset{\sim}{\text{aprox.}} N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

- Então, admitindo H_0 , tem-se a estatística teste

$$T(X, Y) = \frac{D}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{\sim}{\text{aprox.}} N(0, 1)$$

Exemplo 7:

Duas lotarias concorrentes afirmam que em cada 4 bilhetes tem 1 que é premiado. supor que se pretende testar se as probabilidades de se obter um bilhete premiado são diferentes para as duas lotarias. Obteve-se os seguintes dados:

	<i>n</i>	premiados	não premiados
lotaria A	63	14	49
lotaria B	45	13	32

Resolução: temos um teste bilateral,

$$H_0 : p_A = p_B \quad \text{vs.} \quad H_1 : p_A \neq p_B$$

$$\hat{p}_A = \frac{14}{63}, \quad \hat{p}_B = \frac{13}{45}, \quad \hat{d} = \hat{p}_A - \hat{p}_B = -\frac{1}{15}, \quad \hat{p} = \frac{14+13}{63+45} = 0.25$$

e o valor da estatística teste observada é,

$$t(x,y) = \frac{-\frac{1}{15}}{\sqrt{0.25(1-0.25)\left(\frac{1}{63} + \frac{1}{45}\right)}} = -0.79.$$

Exemplo 7 (cont):

o p-value é $2 * P(T(X, Y) < -0.79)$ que calculado no Python

```
import scipy.stats as stat
pv=2*stat.norm(0,1).cdf(-0.79)
print('p-value: ',pv)
p-value: 0.42952776832727424
```



Usando a função `proportions_ztest()` teríamos

```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest
success_cnts = np.array([14, 13])
total_cnts = np.array([49+14, 32+13])
t_stat, pval=proportions_ztest(success_cnts, total_cnts, alternative='two-sided')
print('Two sided z-test: z = {:.4f}, p-value = {:.4f}'.format(t_stat, pval))
Two sided z-test: z = -0.7888, p-value = 0.4302
```

Teste do χ^2

- O teste do χ^2 baseia-se em comparar as frequências absolutas observadas com as frequências absolutas esperadas, supondo H_0 verdadeira.
- Consideramos uma amostra aleatória $X = (X_1, X_2, \dots, X_n)$ que foi agrupada de k classes, onde cada classe possui n_i , $i = 1, 2, \dots, k$ observações ($\sum_{i=1}^k n_i = n$)
- A estatística do teste define-se da forma

$$T(X) = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \sim \chi_{k-1-r}^2$$

onde:

- ▶ n_i , ($i = 1, 2, \dots, k$) são as frequências absolutas das observações da amostra X na classe i , N_i é uma variável aleatória com realização n_i na amostra observada
- ▶ p_i , ($i = 1, 2, \dots, k$), são calculados da distribuição da hipótese nula
- ▶ np_i são as frequências absolutas esperadas na classe i sob H_0
- ▶ r é o número de parâmetros livres que depende a distribuição de H_0
- ▶ O p -value é calculado por,

$$P(T(X) \geq t(x) | H_0)$$

Exemplo 7 (cont.):

Voltando ao exemplo anterior.

- Tem-se $k = 4$

```
from scipy.stats import chi2
e11=63*27/108; e12=63*81/108
e21=27*45/108; e22=81*45/108 # frequencias absolutas esperadas
To=(14-e11)**2/e11+(49-e12)**2/e12+(13-e21)**2/e21+(32-e22)**2/e22
print('valor da estatistica teste:',To)
pvalue=1-chi2.cdf(To,1)
print('p-value:', pvalue)
valor da estatistica teste: 0.6222222222222222
p-value: 0.43022269113219136
```

Decisão: Dado o valor do p-value não se rejeita H_0 para níveis de significância menores que 0.43.



Pode-se usar a função

statsmodels.stats.proportion.proportions_chisquare()

- Sejam $X = (X_1, X_2, \dots, X_{n_1})$ e $Y = (Y_1, Y_2, \dots, Y_{n_2})$ duas amostras i.i.d. de duas populações normais com variâncias σ_1^2 e σ_2^2 , respectivamente
- O teste à razão entre as duas variâncias, σ_1^2 e σ_2^2 , tem a seguinte formulação

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{versus} \quad H_0 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

ou equivalentemente,

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_0 : \sigma_1^2 \neq \sigma_2^2$$

- A estatística de teste é,

$$P(X, Y) = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

onde S_1^2 e S_2^2 são variâncias amostrais das duas populações e F_{n_1-1, n_2-1} é a distribuição F com $n_1 - 1$ graus de liberdade do numerador e $n_2 - 1$ graus de liberdade do denominador.

- O p-value calcula-se usando a fórmula $2\min\{P(T(X, Y) < t_o), P(T(X, Y) > t_o)\}$

Teste de Levene

- O teste de Levene é usado para verificar se k populações têm variâncias iguais
- O teste de Levene é definido por,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \quad \text{versus} \quad \sigma_i^2 \neq \sigma_j^2 \quad \text{para algum par } (i,j)$$

- Dada uma v.a. X com tamanho n dividida em k subgrupos, com cada subgrupo com tamanho n_i , a estatística teste é

$$T(X) = \frac{n-k}{k-1} \frac{\sum_{i=1}^k n_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Z_{ij} - Z_{i.})^2} \sim F_{k-1, N-k}$$

onde $Z_{i,j}$ pode tomar uma das seguintes formas:

- $Z_{i,j} = |X_{i,j} - \bar{X}_{i.}|$, onde $\bar{X}_{i.}$ é a média do i -ésimo grupo
- $Z_{i,j} = |X_{i,j} - \tilde{X}_{i.}|$, onde $\tilde{X}_{i.}$ é a mediana do i -ésimo grupo
- $Z_{i,j} = |X_{i,j} - X'_{i.}|$, onde $\tilde{X}'_{i.}$ é a média truncada do i -ésimo grupo

e

- ▶ $Z_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{i,j}$
- ▶ $Z_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Z_{i,j}$

- O valor de prova é dado por $P(T(X) > t(x))$



No python usar a função `scipy.stats.levene()`

Análise de variância com um factor (One-Way ANOVA)

- A análise de variância com um factor é usada para comparar médias de duas ou mais amostras independentes e testar se existem diferenças estatisticamente significativas.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k \quad \text{vs} \quad H_1 : \mu_i \neq \mu_j, \text{ para algum par } (i,j)$$

- A one-way ANOVA é uma extensão de um t-teste para amostras independentes.
- Numa ANOVA tem-se:
 - 1 A variável independente é categórica que define os grupos (ou factores) que serão comparados
 - 2 A variável dependente é uma variável numérica, cujas médias serão comparadas

Exemplo:

Usa-se uma one-way ANOVA para verificar se as notas de um exame diferem para diferentes níveis de ansiedade entre os alunos, onde se divide os alunos em três níveis de ansiedade (p.e. baixo, médio e alto).

- O resultado do teste apenas informa se há, ou não, diferenças entre os grupos. Para obter mais informações usa-se um teste **post hoc**.

Pressupostos do teste One-Way ANOVA)

- 1 A variável dependente deve ser contínua
- 2 A variável independente deve ter dois ou mais grupos independentes. Geralmente usa-se apenas um teste ANOVA para três ou mais grupos categóricos independentes uma vez que para dois grupos é mais cómodo fazer t-teste.
- 3 As observações devem ser independentes (Não devem existir relações entre as observações de grupos diferentes).
- 4 As observações não devem ter outliers significativos
- 5 A variável dependente deve ser normalmente distribuída para cada grupo (usar p.e. teste de Shapiro)
- 6 Deve existir homogeneidade de variâncias (usar teste de Levene)

- Geralmente guardam-se os dados numa tabela, com k grupos (ou fatores)

Grupos				
1	2	3	...	k
$x_{1,1}$	$x_{2,1}$	$x_{3,1}$...	$x_{k,1}$
$x_{1,2}$	$x_{2,2}$	$x_{3,2}$...	$x_{k,2}$
\vdots	\vdots	\vdots		\vdots
x_{1,n_1}	x_{2,n_2}	x_{3,n_3}	...	x_{k,n_k}

- n_i : dimensão da amostra retirada do grupo i ($i = 1, 2, 3, \dots, k$)
- n : dimensão global da amostra ($\sum_{i=1}^k n_i$)
- $x_{i,j}$: valor observado, da variável $X_{i,j}$, para a j -ésima observação do i -ésimo grupo ($i = 1, 2, 3, \dots, k$ e $j = 1, 2, 3, \dots, n_i$)
- $SS = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{i,j}^2 - \frac{(\sum_{i=1}^k \sum_{j=1}^{n_i} x_{i,j})^2}{n}$
- $TSS = \sum_{i=1}^k \frac{(\sum_{j=1}^{n_i} x_{i,j})^2}{n_i} - \frac{(\sum_{i=1}^k \sum_{j=1}^{n_i} x_{i,j})^2}{n}$
- $RSS = SS - TSS$

- Tem-se que a estatística do one-way ANOVA é,

$$T(X) = \frac{\frac{TSS}{k-1}}{\frac{RSS}{n-k}} \sim F_{k-1, n-k}$$

- O p-value é dado por,

$$\text{p-value} = P(T(X) \geq t(x) | H_0)$$

Exemplo :

Um analista está interessado em testar se 4 populações de morcegos (S1,S2,S3 e S4) têm pesos médios iguais. Assume-se que o peso das populações são normais e possuem variâncias iguais.

```
import scipy.stats as stats
# pesos dos morcegos em gramas
S1=[9,6,11,14,14]
S2=[12,16,16,12,9]
S3=[8,8,12,7,10]
S4=[17,15,17,16,13]
f_statistic, p_value = stats.f_oneway(S1, S2, S3, S4)
print('p-value:', p_value)
p-value: 0.006523332607387391
```

Decisão: O p-value (0.00652) leva-nos a rejeitar a hipótese nula logo as médias não são todas iguais para todo o nível de significância superior a 0,65%