3. ESTIMAÇÃO E TESTES DE HIPÓTESES

3.1 DISTRIBUIÇÕES DE AMOSTRAGEM

Quando seleccionamos uma amostra aleatória de uma população, às características numéricas que calculamos a partir desta chamamos **estatísticas.**

Definição: A uma função da amostra $(X_1,...,X_n)$ que não dependa de parâmetros desconhecidos chama-se **estatística.**

Exemplos: No capítulo de estatística descritiva tivemos oportunidade de calcular várias estatísticas que nos ajudaram a descrever o conjunto de valores observados. Como exemplo, podemos citar a soma das observações $\sum_{i=1}^{n} X_i$, a média $\overline{X}$, a variância empírica S^2 e o desvio-padrão S, o maior valor $X_{(n)}$ (máximo) e o menor valor observado $X_{(1)}$ (mínimo), os quantis empíricos de ordem p, Q_{p}, etc.

Estas estatísticas cujo valor varia de amostra para amostra, são *variáveis aleatórias* e como tal têm uma distribuição de probabilidade a que se chama **distribuição de amostragem.**

Dada a sua importância, interessa conhecer a distribuição de amostragem da média.

I- Algumas vezes é possível conhecer a distribuição exacta de uma estatística, é por exemplo o caso da distribuição da soma quando

i) as observações têm distribuição Bernoulli, isto é, se $X_1,...,X_n$ são observações independentes e cada $X_i \cap B(p)$, então a estatística

$$\sum_{i=1}^{n} X_i \cap Bi(n,p) \tag{1}$$

ii) as observações têm distribuição *gaussiana*, isto é, se $X_1,...,X_n$ são observações independentes e cada $X_i \cap Gau(\mu,\sigma)$, então

$$\boxed{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \cap Gau(0,1)} \tag{2}$$

Ou equivalentemente,

$$\boxed{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\cap Gau(0,1)} \tag{3}$$

II- Nem sempre é possível conhecer a distribuição exacta da estatística média, mas mesmo quando as observações *não são gaussianas* a distribuição da média pode ser *aproximadamente* gaussiana. É por exemplo o caso de termos observações com distribuição Bernoulli. Pelo teorema de DeMoivre-Laplace sabemos que:

i) se $X_1,...,X_n$ são observações independentes e cada $X_i \cap B(p)$, então para valores "grandes" de n $\frac{\overline{X}-p}{\sqrt{p(1-p)}}\sqrt{n} \sim Gau\,(0,1)$.

ii) Um resultado mais geral, de que este teorema é um caso particular, diz que se a distribuição das observações X_i (i=1,...,n) é tal que existem $E(X)=\mu$ e $Var(X)=\sigma^2$, então a distribuição da soma é *aproximadamente* gaussiana. Isto é, pelo **Teorema Limite Central**

iii)

$$\boxed{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Gau(0,1)}$$

ou equivalentemente

$$\boxed{\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \xrightarrow{d} Gau(0,1)}$$

EXEMPLO: Suponhamos que se lança um dado equilibrado e defina-se a v. a. X que designa o número de pintas observadas na face que fica voltada para cima. Esta v.a. é discreta com f.m.p. $p_k = P(X=k)=\frac{1}{6}, \quad k=1,...,6$ (Dist. Uniforme discreta) e cuja representação gráfica é:

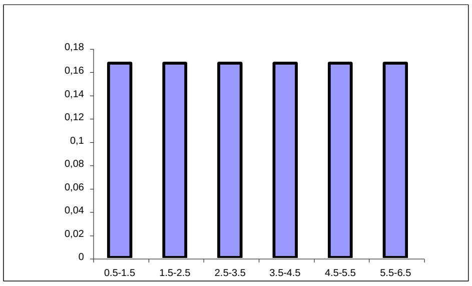

Figura 1- f.m.p. de X

Consideremos agora duas observações independentes desta população X_1, X_2, correspondentes ao número de pintas observadas em cada uma das faces viradas para cima dos dois dados lançados. Calcule-se a soma das pintas obtidas, $\sum_{i=1}^{2} X_i = X_1 + X_2$.

Na tabela abaixo estão representados os valores que a variável soma assume

2º dado	1º dado					
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Tabela 1- valores da v.a. $X_1 + X_2$.

Dada a independência das variáveis X_1, X_2 cada valor da tabela tem probabilidade $P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j) = \frac{1}{36}, \quad i, j = 1,\ldots,6$, e assim é fácil calcular a f.m.p. da variável soma

$X_1 + X_2$	p_k
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Tabela 2- f.m.p. $X_1 + X_2$

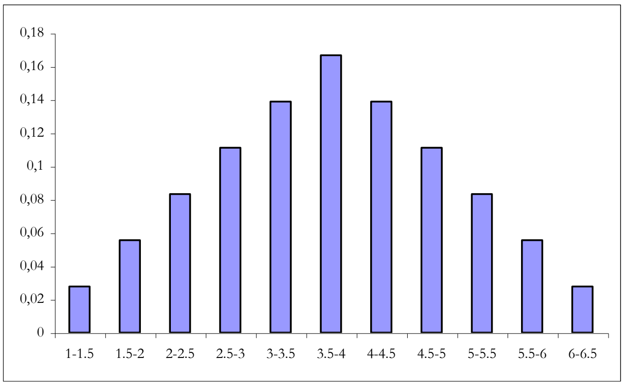

Figura 2-f.m.p. de $\overline{X} = \dfrac{X_1 + X_2}{2}$

A distribuição de amostragem tem agora uma forma diferente, quase campanular, mas continua a ser simétrica em torno do seu valor médio $\mu = E\left(\dfrac{X_1 + X_2}{2}\right) = \dfrac{E(X_1) + E(X_2)}{2} = 3.5$.

Aumentando agora a dimensão da amostra para n=3

$X_1 + X_2 + X_3$	Freq. absoluta
3	1
4	3
5	6
6	10
7	15
8	21
9	25
10	27
11	27
12	25
13	21
14	15
15	10
16	6
17	3
18	1

Tabela 3- f.m.p. $X_1 + X_2 + X_3$

e representando graficamente a distribuição da média $\overline{X} = \dfrac{\sum_{i=1}^{3} X_i}{3}$, obtemos o seguinte diagrama de barras

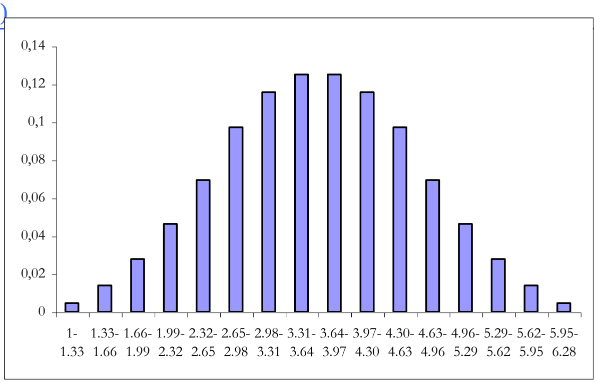

Figura 3- f.m.p. de $\overline{X} = \dfrac{X_1 + X_2 + X_3}{3}$

Tendo-se ainda para n=4

$X_1 + X_2 + X_3 + X_4$	Freq. absoluta
4	1
5	4
6	10
7	20
8	35
9	56
10	80
11	104
12	125
13	140
14	146
15	140
16	125
17	104
18	80
19	56
20	35
21	20
22	10
23	4
24	1

Tabela 4- f.m.p. $X_1 + X_2 + X_3 + X_4$

representando graficamente a distribuição da média $\overline{X} = \dfrac{1}{4}\sum_{i=1}^{4} X_i$,

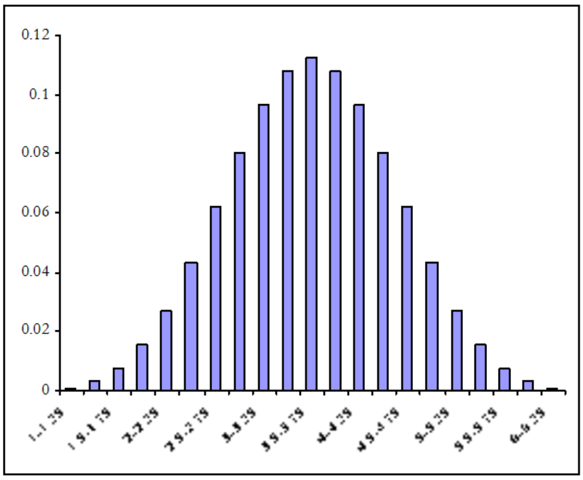

Figura 4- f.m.p. de $\overline{X} = \dfrac{X_1 + X_2 + X_3 + X_4}{4}$

A distribuição da média tem agora uma nítida forma campanular, e a distribuição aproxima-se rapidamente da distribuição gaussiana.

3.2 ESTIMAÇÃO PONTUAL

No capítulo anterior estudámos algumas distribuições discretas e contínuas e pudemos constatar que a f.m.p. ou a f.d.p. destas dependia de constantes, parâmetros, que assumimos conhecidos. Temos, por exemplo, o parâmetro **p** (probabilidade de sucesso) da distribuição binomial, o parâmetro λ (valor médio) da distribuição Poisson, os parâmetros μ e σ (valor médio e desvio-padrão respectivamente) da distribuição gaussiana, etc. Na prática estes parâmetros não são usualmente conhecidos e é necessário estimá-los a partir da amostra observada, podendo-se obter um estimador pontual ou uma região de confiança para o parâmetro genericamente designado por θ, assuntos que iremos estudar nas secções seguintes.

Definição 1: *Estimador* de um parâmetro θ é qualquer estatística, função da amostra $(X_1,...,X_n)$ que não depende de parâmetros desconhecidos, e que toma valores admissíveis para θ.

Há métodos que nos permitem encontrar o estimador de um parâmetro, um deles é muito intuitivo e chama-se *Método dos Momentos*. Tal como o nome indica consiste em igualar os momentos da amostra aos correspondentes momentos teóricos, quando estes existem.

Exemplos: 1- Seja $(X_1,...,X_n)$ uma amostra aleatória de uma população X, de valor médio μ desconhecido que queremos estimar usando o Método dos Momentos:

i) Se $X \cap Gau(\mu,\sigma)$ então bastará fazer $\hat{\mu} = \overline{X}$; *ii)* $X \cap Exp(\lambda)$, $E(X) = \frac{1}{\lambda}$ logo $\hat{\lambda} = \frac{1}{\overline{X}}$; *iii)* no caso de $X \cap P(\lambda)$, $E(X) = \lambda$, logo $\hat{\lambda} = \overline{X}$; *iv)* se $X \cap Bi(n,p)$, $E(X) = np$ donde $\overline{X} = n\hat{p}$ ou $\hat{p} = \dfrac{X}{n}$.

2- Quando queremos estimar a variância $\sigma^2 = E(X-\mu)^2$ o estimador natural e que resulta de aplicar o Método dos Momentos é $S^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.

Quando nos baseamos na amostra para tirar conclusões sobre toda a população estamos a fazer *inferência estatística*. A estimação de um parâmetro é um problema de inferência estatística paramétrica.

Definição 2: À distância entre o estimador $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ e o parâmetro estimado θ chama-se *erro de estimação* e designa-se por $Err(\hat{\theta}) = |\hat{\theta} - \theta|$.

O *erro de estimação* é uma v.a. e se soubermos a sua distribuição podemos determinar para *qualquer* $\varepsilon > 0$ *a* $P(|\hat{\theta} - \theta| < \varepsilon)$, isto é, podemos calcular a probabilidade do erro de estimação não exceder um determinado valor.

Exercício Sejam $X_1, ..., X_n$, v.a.'s independentes com distribuição Bernoulli de parâmetro **p** desconhecido. Supondo que se sabe que o verdadeiro valor de **p** está na vizinhança de 0.2 qual deve ser a dimensão **n** da amostra de modo a ter

$$P(|\overline{X}_n - p| \le 0.1) \ge 0.95 \qquad (1)$$

quando p = 0.2.

RESOLUÇÃO: Tomando p=0.2 (valor de alguma estimativa obtida anteriormente) a expressão (1) é equivalente a

$$P\left(-\frac{0.1\sqrt{n}}{\sqrt{0.2\times 0.8}} \le \frac{\overline{X} - 0.2}{\sqrt{0.2\times 0.8}}\sqrt{n} \le \frac{0.1\sqrt{n}}{\sqrt{0.2\times 0.8}}\right) \ge 0.95 \Leftrightarrow$$

$$\Leftrightarrow P\left(-\frac{\sqrt{n}}{4} \le \frac{\overline{X} - 0.2}{\sqrt{0.2\times 0.8}}\sqrt{n} \le \frac{\sqrt{n}}{4}\right) \ge 0.95$$

Recordando que pelo teorema de DeMoivre-Laplace se tem $\dfrac{\overline{X} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \xrightarrow{d} Gau(0,1)$ e consultando a tabela da f.d. da gaussiana padrão obtemos a seguinte desigualdade $\dfrac{\sqrt{n}}{4} \ge \Phi^{-1}(0.975) = 1.96 \Leftrightarrow n \ge (4 \times 1.96)^2 \approx 61.5$. Assim, para obtermos um erro de estimação não superior a uma décima com uma probabilidade de pelo menos 95%, devemos, se possível, observar uma amostra de dimensão superior ou igual a **62.**

3.3 INTERVALOS DE CONFIANÇA

Por vezes não satisfaz apresentar somente um estimador *pontual* para o parâmetro θ desconhecido, mas tem mais interesse indicar um intervalo aleatório que o contenha com um certo grau de certeza usualmente elevado, ao qual se chama *nível de confiança.*

Definição: A um intervalo aleatório

$$\left(L_1\left(\underset{\sim}{X}\right), L_2\left(\underset{\sim}{X}\right)\right) \subseteq \Re \quad \text{tal que} \quad P\left\{\left(L_1\left(\underset{\sim}{X}\right), L_2\left(\underset{\sim}{X}\right)\right) \ni \theta\right\} = 1-\alpha \qquad (1)$$

onde $\underset{\sim}{X} = (X_1, ..., X_n)$ e $L_i\left(\underset{\sim}{X}\right), i = 1, 2$ são estatísticas, chama-se um *Intervalo de Confiança* (I.C.) de nível de confiança (1-α)×100%.

Geralmente há mais do que um intervalo nestas condições, sempre que possível escolhe-se o melhor, ou seja, o de menor amplitude.
Vamos estudar I.C. para os parâmetros das distribuições binomial e gaussiana, casos que ilustram bem um método geral de construção destes intervalos, o método da variável *fulcral.*

3.3.1 INTERVALOS DE CONFIANÇA PARA O PARÂMETRO p DA DISTRIBUIÇÃO BINOMIAL

Recordemos que a distribuição binomial aparece frequentemente quando pretendemos estimar a *proporção* **p** de indivíduos que têm uma determinada característica, a partir de uma amostra. Consideremos o seguinte

Exemplo - Foi proposta a abertura de uma nova estrada numa certa localidade. As opiniões dividem-se sobre a utilidade desta estrada Assim, os residentes desta localidade resolveram seleccionar aleatoriamente 50 pessoas das listas de voto e perguntaram-lhes as suas opiniões. Verificou-se que 30 foram a favor da nova estrada. Pede-se um intervalo de 95% de confiança para a percentagem **p** de todas as pessoas recenseadas que estão a favor da nova estrada.
Neste caso a distribuição da v.a. X que designa o número de pessoas inquiridas que estão a favor da nova estrada tem distribuição Bi (50, **p**) e observou-se o valor **x = 30.**
Sabemos que para n "grande" (n>25), a distribuição Bi (n, **p**) se pode aproximar pela

distribuição gaussiana, i.e.,

$$X \sim Gau\left(np, \sqrt{np(1-p)}\right) \Leftrightarrow \frac{X-np}{\sqrt{np(1-p)}} \sim Gau(0,1) \qquad (2)$$

e consultando a tabela da f.d. da *Gau* (0, 1) podemos escrever

$$P(-1.96 < \frac{X-np}{\sqrt{np(1-p)}} < 1.96) = 0.95 \qquad (3)$$

resolvendo as desigualdades anteriores em ordem a *p* obtém-se o intervalo aleatório que contém o verdadeiro valor de p com probabilidade aproximadamente igual a 0.95

$$\left(\frac{X}{n} - 1.96\sqrt{\frac{p(1-p)}{n}}, \frac{X}{n} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \qquad (4)$$

No entanto não o podemos utilizar como intervalo de confiança para **p,** dado que os seus limites dependem de p e este é desconhecido. Assim, teremos de substituir **p** pelo seu estimador $\hat{p} = \frac{X}{n}$, vindo finalmente

$$\left(\frac{X}{n} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \frac{X}{n} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \qquad (5)$$

Para o nosso exemplo, substituindo X pelo seu valor observado x = 30 obtém-se o intervalo (determinista) **(0.464, 0.736).**

NOTA: O intervalo obtido tem grande amplitude, incluindo até valores inferiores a 50% o que indica que possivelmente só uma minoria de pessoas estará interessada na nova estrada. Como podemos obter um intervalo de 95% de confiança de menor amplitude e portanto mais informativo? A amplitude do intervalo dado pela expressão (4) é igual a

$$d = 2 \times 1.96\sqrt{\frac{p(1-p)}{n}} \qquad (6)$$

ou seja n = $4x(1.96)^2p(1-p)/d^2$. Assim, substituindo p por uma sua estimativa é possível determinar o valor de n de modo a que a amplitude *d* não ultrapasse um valor por nós estabelecido.

No presente exemplo a amplitude do intervalo é *d*=0.736-0.464=0.272, se quisermos ter *d*=0.15 quantas pessoas devemos seleccionar para responder ao inquérito?

Tomemos para estimativa de p a percentagem de pessoas que na amostra já observada é a favor da construção da nova estrada, i.e., $\frac{x}{n} = \frac{30}{50} = 0.6$, donde $n = \frac{4 \times (1.96)^2 \times 0.6 \times 0.4}{0.0225}$ =163.9, e portanto devemos seleccionar 164 indivíduos. E supondo ainda que a proporção de pessoas a favor da nova estrada se mantém em 0.6, o intervalo de amplitude d=0.15 é (0.6 –0.075, 0.6 +0.075) = (0.525, 0.675) que não inclui o valor 0.5 e é mais informativo do que o anterior.

3.3.2 INTERVALOS DE CONFIANÇA PARA O VALOR MÉDIO DE POPULAÇÕES GAUSSIANAS

Consideremos n observações independentes X_1, X_2,..., X_n de uma população gaussiana de variância σ^2 e valor médio μ desconhecido. Um estimador natural de μ será a média amostral $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$, que aproxima o verdadeiro valor do parâmetro. Pretendemos construir um intervalo de confiança para μ, estudando separadamente os casos: **i)** *variância σ^2 conhecida* e **ii)** *variância σ^2 desconhecida*.

i) POPULAÇÃO GAUSSIANA COM VARIÂNCIA CONHECIDA

Exemplo 1 - Uma instituição de ensino desenvolveu um teste para medir a capacidade de análise e destreza manual de crianças com cinco anos de idade. Para aferir este teste foi recolhida uma amostra aleatória de 10 crianças de 5 anos de idade observando-se os seguintes tempos (em segundos) de execução da dita tarefa:

52.7, 60.8, 40.9, 39.5, 42.6, 47.1, 35.9, 59.2, 66.7, 48.5

Supõe-se que estes valores provêem de uma população normal de variância 100 segundos^2. Pretendemos construir um intervalo de confiança para μ, tempo médio de execução da tarefa para crianças de 5 anos de idade. Recorde-se que

$\bar{X} \cap Gau\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ ou equivalentemente,

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \cap Gau\,(0,\, 1) \tag{1}$$

consultando a tabela da distribuição *Gau* (0, 1) obtém-se

$$P(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95 \qquad (2)$$

e resolvendo as desigualdades em ordem a μ vem

$$P(\bar{X} - 1.96\ \sigma/\sqrt{n} < \mu < \bar{X} + 1.96\ \sigma/\sqrt{n}) = 0.95 \qquad (3)$$

note-se que σ é conhecido. A asserção anterior tem a seguinte leitura: *a probabilidade de que o intervalo aleatório* $(\bar{X} - 1.96\ \sigma/\sqrt{n},\ \bar{X} + 1.96\ \sigma/\sqrt{n})$ *contenha o verdadeiro valor de* μ *é* 0.95.

Substituindo o valor de $\bar{X}$ pela estimativa $\bar{x} = \mathbf{49.4}$, média da amostra observada, e o valor de σ=10, obtém-se o intervalo de 95% de confiança para μ: **(43.2, 55.6)**. A teoria frequencista de probabilidade diz-nos que, se tivermos *100 intervalos construídos desta maneira em média 95* conterão o verdadeiro valor do parâmetro.

Aumentando o nível de confiança do intervalo para 99%, e consultando a tabela da distribuição *Gau* (0, 1) vem

$$P(-2.58 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2.58) = 0.99 \qquad (4)$$

$$P(\bar{X} - 2.58\ \sigma/\sqrt{n} < \mu < \bar{X} + 2.58\ \sigma/\sqrt{n}) = 0.99 \qquad (5)$$

E obtém-se o *intervalo de 99% de confiança* para $\mu : (\bar{X} - 2.58\ \sigma/\sqrt{n},\ \bar{X} + 2.58\ \sigma/\sqrt{n})$. Substituindo o valor de $\bar{X}$ pela estimativa $\bar{x} = \mathbf{49.4}$ e σ=10 obtém-se **(41.3, 57.5)**. Como seria de esperar, a amplitude do intervalo de confiança aumentou, esta é a penalização por termos aumentado o grau de confiança de que o intervalo assim construído conterá o valor μ.

ii) POPULAÇÃO GAUSSIANA COM VARIÂNCIA DESCONHECIDA

Consideremos o caso mais geral, que usualmente ocorre na prática, da variância ser desconhecida. *Como construir o intervalo de confiança para μ?*

As expressões (2) e (4) não podem agora ser utilizadas, dado que dependem de σ desconhecido, a não ser que substituamos σ pelo seu estimador

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

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \cap t_{n-1} \qquad (6)$$

e procedendo analogamente à secção anterior vem

$$P(-t_{0.975} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{0.975}) = 0.95 \qquad (7)$$

onde $t_{0.975}$ (quantil de probabilidade 0.975) é obtido da tabela da distribuição t_{n-1}. Resolvendo as desigualdades anteriores em ordem a μ obtém-se:

$$P(\overline{X} - t_{0.975} S/\sqrt{n} < \mu < \overline{X} + t_{0.975} S/\sqrt{n}) = 0.95 \qquad (8)$$

isto é, $(\overline{X} - t_{0.975} S/\sqrt{n}\ ,\ \overline{X} + t_{0.975} S/\sqrt{n})$ é o *intervalo de 95% de confiança* μ para quando σ^2 é desconhecido. A interpretação probabilística do intervalo de 95% de confiança é igual à dada na secção anterior.

Note-se que a amplitude do intervalo anterior D = $2xt_{0.975})S/\sqrt{n}$ é agora aleatória, o que não acontecia no caso da variância ser conhecida. Isto é devido ao facto de S^2 ser uma variável aleatória.

Exemplo 2: Um industrial de farmácia deseja determinar o tempo médio de absorção de um medicamento pelo organismo. O medicamento é ministrado a 8 pessoas escolhidas aleatoriamente, e são observados os tempos de absorção (em minutos) para que a concentração do medicamento no sangue atinja um determinado nível, obtendo-se os valores:

36, 29, 34, 33, 38, 28, 30, 33

Construa o intervalo de 99% de confiança para o tempo médio de absorção.

RESOLUÇÃO: Os valores das estatísticas $\overline{X}$ e S^2 para esta amostra são $\bar{x}$=32.63 e s^2=11.98 respectivamente e $t_{0.995;\,7}$ =3.499 (valor lido na tabela da distribuição t com parâmetro n-1=7), vindo para I.C de nível 99%

(32.63-4.282, 32.63+4.282)

OBSERVAÇÃO:

Distribuição t-Student:

- É uma distribuição simétrica em torno de 0.
  - A sua f.d.p. tem uma representação gráfica semelhante à da gaussiana.
  - O seu parâmetro é o nº de graus de liberdade

$\rightarrow$ quanto menor este for maior é a dispersão da distribuição.

Nesta aplicação o nº de g.l. é n-1 $\rightarrow$ tínhamos uma amostra de dimensão n para estimar μ, mas perdemos um grau de liberdade ao termos de estimar σ.

À medida que n aumenta mais a distribuição t-Student se assemelha à *Gau*(0, 1)
$\rightarrow$ para n>30 as duas distribuições são mesmo muito idênticas.

O cálculo de quantis e probabilidades associadas a esta distribuição pode ser feito utilizando uma tabela que nos fornece para dados valores de probabilidade α os valores de $t_{1-\alpha;n-1}$.

3.3.3 INTERVALOS DE CONFIANÇA PARA A VARIÂNCIA DE POPULAÇÕES GAUSSIANAS

Consideremos n observações independentes X_1, X_2, ..., X_n de uma população gaussiana de valor médio μ e variância σ^2 desconhecida. Um estimador natural de σ^2 será S^2**.** Pretendemos construir um intervalo de confiança para σ^2, estudando separadamente os casos: **i)** *valor médio* μ *conhecido* e **ii)** *valor médio* μ *desconhecido*.

i) POPULAÇÃO GAUSSIANA COM VALOR MÉDIO CONHECIDO

O estimador da variância é então $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2$ e pode-se demonstrar que:

$$\frac{(n-1)S^2}{\sigma^2} \cap \chi^2_n \qquad (9)$$

onde χ^2_n designa a distribuição qui-quadrado com n graus de liberdade (g.l.). E portanto esta variável é uma *variável fulcral.*

Pretendemos encontrar um par de valores (a, b) tal que:

$$P\left(a<\frac{(n-1)S^2}{\sigma^2}<b\right)=1-\alpha \qquad (10)$$

Mas este problema não tem solução única a não ser que se imponha alguma condição adicional, como por exemplo,

$$P\left(\frac{(n-1)S^2}{\sigma^2}<a\right)=P\left(\frac{(n-1)S^2}{\sigma^2}>b\right)=\frac{\alpha}{2} \qquad (11)$$

e então $a=\chi_{\frac{\alpha}{2};n}$ e $b=\chi_{1-\frac{\alpha}{2};n}$. Resolvendo as desigualdades anteriores em ordem a σ^2 obtém-se o I.C de nível $(1-\alpha)\times 100\%$.:

$$\left(\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2};n}},\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2};n}}\right) \qquad (12)$$

ii) POPULAÇÃO GAUSSIANA COM VALOR MÉDIO DESCONHECIDO

No caso de μ ser desconhecido é necessário estimá-lo e então o estimador da variância é, $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$. A distribuição da variável ainda é qui-quadrado, mas perdemos um g.l.

$$\frac{(n-1)S^2}{\sigma^2}\cap\chi^2_{n-1} \qquad (13)$$

O I.C. é análogo ao obtido anteriormente sendo a única diferença os quantis da distribuição qui-quadrado que são agora iguais $a=\chi_{\frac{\alpha}{2};n-1}$ e $b=\chi_{1-\frac{\alpha}{2};n-1}$.

O I.C de nível $(1-\alpha)\times 100\%$.é:

$$\left(\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2};n-1}},\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2};n-1}}\right) \qquad (14)$$

Exemplo 3: Voltando ao dados do exemplo 2, com $\overline{x}$=32.63 e s^2=11.98 e consultando as tabelas da distribuição qui-quadrado com n-1=7 g.l. (estamos no caso ii)) obtém-se a= $\chi_{0.005;7}$ e $b=\chi_{0.995;7}$ respectivamente iguais a *0.9893* e *20.28*. Logo o I.C. de nível 99% que se obtém é (4.14, 84.77).

OBSERVAÇÃO:

<u>Distribuição qui-quadrado com ν g.l.-</u> χ^2_ν :

- É uma distribuição contínua com f.d.p. igual a:

- $$f_{\chi^2_\nu} = \frac{x^{\frac{\nu}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{\nu}{2}}\, \Gamma\left(\frac{\nu}{2}\right)} \qquad x > 0$$

- O seu parâmetro é o nº de graus de liberdade e

  Se $X \cap \chi^2_\nu$ *então* $E(X) = \nu$ *e* $Var(X) = 2\nu$

- A função de distribuição desta variável encontra-se tabelada.

- Esta distribuição tende para a Gaussiana quando $\nu \to +\infty$, na prática considera-se que esta aproximação é boa para ν >30.

3.4 TESTES DE HIPÓTESES

3.4.1. INTRODUÇÃO

As experiências são usualmente levadas a cabo com o objectivo de testar uma teoria, ou *hipótese*, sobre a natureza do fenómeno em estudo. Se este é um fenómeno aleatório com um modelo probabilístico subjacente, então pode ter interesse formular uma hipótese acerca do valor assumido por um ou mais parâmetros do modelo. Quando os dados estão sujeitos a uma variação de carácter aleatório, não há possibilidade de dizer se uma hipótese é correcta ou não de uma forma *conclusiva*. O que pretendemos então, é arranjar um procedimento ou *teste de hipótese*, que perante os dados observados nos leve a concluir pela evidência ou não da hipótese formulada.

Os erros que podemos cometer são de dois tipos dado que:

		Verdadeira	Verdadeira
		H_0	H_1
Não Rejeitar	H_0	decisão correcta	**decisão incorrecta**
Não Rejeitar	H_1	**decisão incorrecta**	decisão correcta

Consideremos algumas situações em que este problema surge:

Exemplo 1- Um fabricante de pneus está a pensar em mudar o processo de fabrico do seu produto. Por experiência adquirida, ele sabe que em média o tempo de vida X, medido em 1000 km até à ruptura dos pneus, é normal de valor médio 40 e variância 9. Foram fabricados 10 pneus pelo novo processo e observado o seu tempo de vida, com base neste resultado ele pretende decidir se o valor médio de X foi alterado (ele continua a assumir que a variância é 9).

Uma hipótese que pode ter interesse considerar neste caso é a de que o novo processo não produz alteração em $E(X)= \mu$, i. e., o novo processo de fabrico tem o mesmo $\mu = 40$. Uma outra hipótese que pode ter interesse é considerar que o novo processo introduziu uma melhoria que se manifesta por se ter $\mu > 40$. Com base na evidência oferecida pelas observações (tempo de vida dos 10 pneus obtidos pelo novo processo de fabrico), ele deve decidir se vai ou não mudar para o novo processo de fabrico. Note-se que ambas as hipóteses dizem respeito ao valor médio μ do modelo, neste caso *Gau*(μ, σ), do qual nós dispomos somente de 10 observações. Dado que os resultados da experiência estão sujeitos a variabilidade de carácter aleatório não podemos inequivocamente decidir por uma hipótese ou por outra. O melhor que podemos fazer é utilizar um método de escolha que limite, de certo modo, as possibilidades de cometer erros.

Antes de prosseguirmos com estes exemplos estudemos a situação geral. Suponhamos que a nossa amostra aleatória se compõe de observações $X_1, X_2, ..., X_n$ que provêm de uma população *Gau*(μ, σ) onde σ conhecido mas μ desconhecido. Há duas hipóteses relativamente a μ, uma é a *hipótese nula* ou H_0, usualmente da forma $\mu = \mu_0$ onde μ_0 é um valor especificado e que estabelece que não há mudança (no exemplo anterior $\mu_0 = 40$). A outra hipótese, a *hipótese alternativa* ou H_1, diz que houve mudança e especifica a natureza da mudança. Consideraremos quatro tipos de hipóteses alternativas

i) $\mu \neq \mu_0$; *ii)* $\mu > \mu_0$; *iii)* $\mu < \mu_0$,; *iv)* $\mu = \mu_1$, *onde* μ_1 *é um valor especificado diferente de* μ_0.

Considera-se o tipo i) quando qualquer mudança no valor de μ tem interesse; consideram-se os casos ii) e iii) quando só interessa detectar mudanças num determinado sentido e o último tipo interessa quando as observações só podem provir de uma de duas populações *Gau*(μ_0, σ) ou *Gau*(μ_1, σ) As hipóteses i), ii) e iii), nas quais mais do que um valor de μ é consistente com a hipótese alternativa, dizem-se *hipóteses compostas*; as hipóteses do tipo iv) onde a distribuição está univocamente determinada chamam-se *hipóteses simples*.

Na abordagem aos testes de hipóteses que vamos adoptar, supõem-se que a *hipótese nula* está relacionada com a teoria (talvez sugerida pela experiência adquirida) cuja consistência com as observações nós desejamos examinar. Só iremos decidir em desfavor da *hipótese nula* se as observações sugerirem uma forte evidência contra ela. As hipóteses não são pois simétricas.

No exemplo 1 do fabricante de pneus a hipótese nula é que o novo processo de fabrico não altera o valor médio do tempo de vida dos pneus, i.e., H_0: $\mu = 40$, e dado que só interessa aumentar o tempo de vida a hipótese alternativa é H_1: $\mu > 40$.

Veremos agora como podemos decidir se a hipótese nula é ou não aceitável de acordo com as observações. Dado que as hipóteses nula e alternativa dizem ambas respeito ao valor médio do tempo de vida dos pneus, parece sensato utilizar a média amostral $\overline{X}$ para decidir entre as duas. Recorde-se que sendo a população $X \cap Gau(\mu,\sigma)$, então $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$, com X_i observações independentes de X, tem distribuição *Gau*(μ, $\dfrac{\sigma}{\sqrt{n}}$) donde valores de $\overline{X}$ maiores do que 40 dão mais suporte à hipótese H_1 do que à hipótese H_0, mas *quão grande deve ser* $\overline{X}$ *para optarmos por* H_1?

3.4.2 NÍVEL DE SIGNIFICÂNCIA E REGIÃO CRÍTICA

Na abordagem aos testes de hipóteses é dada maior importância ao facto de evitar uma decisão errada relativamente à *Hipótese Nula*, i.e., *rejeitar* H_0 *quando esta é verdadeira*, donde se fixa apriori um valor chamado ***nível de significância***. Este valor é a probabilidade de decidir pela Hipótese Alternativa quando a Hipótese Nula é a verdadeira. Os níveis de significância escolhidos na prática são valores pequenos; 0.05 é o mais usual mas também se usam os valores 0.01 e 0.001 quando é particularmente importante que a Hipótese Nula não seja rejeitada indevidamente.

Voltando ao exemplo do fabricante de pneus, este poderá desejar evitar modificações no processo de fabrico devido a razões que envolvem custos , a não ser que o tempo médio de vida dos pneus sofra um aumento significativo com o novo método. Fixemos em 0.05 ou 5% o nível de significância do teste,

$$P(\text{Rejeitar } H_0 \,/\, H_0 \text{ verdadeira}) = P(\bar{X} > k \,/\, \mu = 40) = 1 - \Phi\left(\frac{k-40}{\sqrt{0.9}}\right) = 0.05$$

donde $\Phi\left(\frac{k-40}{\sqrt{0.9}}\right) = 0.95$ e dado que a função Φ tem inversa vem

$$k = \sqrt{0.9}\ \Phi^{-1}(0.95) + 40,$$

consultando a tabela da distribuição *Gau*(0, 1) obtém-se: $k = 1.64\ \sqrt{0.9} + 40 = 41.48$.

Assim, se $\bar{X} > 41.48$ <u>rejeitaremos</u> a Hipótese Nula H_0: $\mu = 40$ e optaremos pela Hipótese Alternativa H_1: $\mu > 40$, caso contrário <u>não rejeitaremos</u> H_0 : $\mu = 40$. Esta regra para decidir sobre a validade da Hipótese Nula satisfaz as condições que mencionámos anteriormente:

i) somente optaremos por H_1 se $\bar{X}$ é suficientemente grande

ii) se H_0 é verdadeira, a probabilidade de optar erradamente por H_1, i.e.,

$P(\bar{X} > 41.48 \,/\, \mu = 40) = 0.05$ (nível de significância do teste).

À região $\mathbf{R} = \{(X_1, X_2, ..., X_{10}) : \bar{X} > 41.48\}$ chama-se ***região de rejeição*** ou ***região crítica*** do teste.

<u>Exercício</u> 1- *Variemos o nível de significância* do exemplo anterior para 0.01(1%), então $k = \sqrt{0.9}\ \Phi^{-1}(0.99) + 40 = 42.21$, i.e., $\mathbf{R} = \{(X_1, X_2, ..., X_{10}) : \bar{X} > 42.21\}$. O facto de termos diminuído o nível de significância fez com que a evidência contra H_0, em termos de "grandes" valores de $\bar{X}$ tenha agora de ser mais forte.

2- *Variemos a dimensão da amostra* para n = 20 no mesmo exemplo. A variância de $\bar{X}$ é inversamente proporcional a n (= $\frac{\sigma^2}{n}$), portanto o valor k vem agora mais pequeno como seria de esperar dado que a distribuição é menos achatada, de facto k = 41.10. Se n = 100, $\frac{\sigma^2}{n} = 0.09$, donde $k = 1.64 \times 0.3 + 40 = 40.49$. Assim, quanto mais observações temos mais fácil é chegar a uma evidência contra a Hipótese Nula.

Elementos de um teste

1. Hipótese Nula H_0

2. Hipótese Alternativa H_1

3. Estatística de teste

4. Região de Rejeição

3.4.3 TESTES PARA O VALOR MÉDIO DE POPULAÇÕES GAUSSIANAS

I- Com Variância σ^2 conhecida

Ilustremos com exemplos algumas das situações anteriores:

Exemplo 2- Um médico está interessado no efeito de farelo na dieta. Como parte da sua investigação pretende averiguar se uma elevada percentagem de farelo leva a modificações no aumento de peso de ratos em desenvolvimento. Ele sabe que para ratos com uma dieta tipo, o aumento médio de peso ao fim de um período de 30 dias é 10g. Ele vai estudar o efeito do aumento de proporção de farelo na dieta de 16 ratos ao fim de 30 dias. Neste caso H_0: $\mu = 10$, onde μ é o aumento médio de peso dos ratos ao fim de 30 dias com a nova dieta e H_1: $\mu \neq 10$, dado que o médico está interessado em detectar qualquer mudança (aumento ou decréscimo) no aumento médio de peso. A distribuição do aumento médio de peso com ambas as dietas (tipo e experimental) pode supor-se normal com variância 300 g^2. Se a média amostral do aumento médio de peso dos 16 ratos sujeitos à dieta experimental for 19.1 g, poderá rejeitar-se H_0 aos níveis de significância i) 5%; ii) 1% ?

A Hipótese Nula e Alternativa estão especificadas e também está escolhida a estatística de teste $\overline{X} = \dfrac{X_1 + \ldots + X_{16}}{16}$ falta-nos agora construir a região de rejeição.

Note-se que neste caso qualquer modificação no aumento de peso é considerada na hipótese alternativa, e dada a simetria da distribuição de $\overline{X}$ em torno de μ parece sensato rejeitar a hipótese nula quando $|\overline{X} - 10| > k$, dado que 10 é o valor μ sob H_0. Falta-nos agora determinar o valor de k a partir do nível de significância fixado. Consideremos separadamente os casos

i) k : P($|\bar{X} - 10| > k$) = 5% $\Leftrightarrow$ 1-P ($|\bar{X} - 10| \leq k$) = 1- P (-k$\leq \bar{X}$ - 10$\leq$ k) =

$$1- [\Phi(\frac{k}{2.5\sqrt{3}}) - \Phi(-\frac{k}{2.5\sqrt{3}})] = 0.05$$

(note-se que $\bar{X} \cap Gau\left(10, \sqrt{\frac{300}{16}}\right)$. Resolvendo a equação anterior em ordem a k (existe Φ^{-1}) vem k=2.5√3Φ^{-1}(0.975)=2.5√3x1.96=8.5, logo a região de rejeição para o teste de nível 5% é

$$\mathbf{R} = \{(X_1, ..., X_{16}): \bar{X} > 10+2.5\sqrt{3}\ \Phi^{-1}(0.975) \text{ ou } \bar{X} < 10-2.5\sqrt{3}\ \Phi^{-1}(0.975)\}$$

$$\mathbf{R} = \{(X_1, ..., X_{16}): \bar{X} > 18.5 \text{ ou } \bar{X} < 1.5\}$$

ii) k : P($|\bar{X} - 10| > k$) = 1%

Procedendo de igual modo obtemos para região de rejeição:

$$\mathbf{R} = \{(X_1, ..., X_{16}): \bar{X} > 10+2.5\sqrt{3}\ \Phi^{-1}(0.995) \text{ ou } \bar{X} < 10-2.5\sqrt{3}\ \Phi^{-1}(0.995)\} =$$

$$=\{(X_1, ..., X_{16}): \bar{X} > 21.19 \text{ ou } \bar{X} < -1.19\}$$

Exemplo 3- Uma antropóloga descobriu um túmulo pre-histórico com 5 esqueletos pertencentes a adultos machos de um certo tipo de homem primitivo. A teoria em vigor, relacionada com túmulos na área da descoberta, afirma que os esqueletos pertencem ao tipo A, sendo a outra possibilidade estes pertencerem ao tipo B. Os dois tipos diferem no tamanho médio do cérebro, sendo para o tipo A 500 ml e 600 ml para o tipo B. Assim, a Hipótese Nula é H_0: $\mu = 500$, de acordo com a teoria vigente, e a hipótese Alternativa é H_1: $\mu = 600$. A média do tamanho dos cérebros encontrados é $\bar{x}$ = 575 ml. Para ambos os tipos A e B a distribuição do tamanho dos cérebros pode considerar-se normal de variância 15 000 ml^2. Poderá a teoria standard, que os esqueletos são do tipo A, ser rejeitada ao nível de significância de 5%?

A estatística de teste é ainda a média amostral $\bar{X}$, deveremos rejeitar H_0 quando $\bar{X}$ > k, dado que na hipótese alternativa se tem $\mu = 600 > 500$. Como anteriormente, o valor de k é calculado através da igualdade:

$$P(\bar{X} > k/\ \mu = 500) = 5\% \Leftrightarrow 1\text{-}P(\bar{X} \leq k\ /\ \mu = 500) = 1\text{-}P\left(\frac{\bar{X} - 500}{\sqrt{15000/5}} \leq \frac{k - 500}{\sqrt{15000/5}}\right) =$$

$=1-\Phi(\frac{k-500}{54{,}77}) = 0.05 \Leftrightarrow k = 54{,}77 * \Phi^{-1}(0.95) + 500 = 589{,}82$

Logo,

R = $\{(X_1, ..., X_5): \overline{X} > 500+54{,}77 * \Phi^{-1}(0.95)\} = \{(X_1, ..., X_5): \overline{X} > 589{,}82\}$

dado que para a amostra observada se tem $\bar{x} = 575$ ml, conclui-se que não há evidência significativa contra a hipótese H_0, ao nível de significância de 5%.

Mais geralmente, as expressões da região crítica do teste da hipótese H_0: $\mu = \mu_0$ ao nível de significância α, para as várias formas da hipótese alternativa, dado que as observações X_1, X_2, ..., X_n pertencem à distribuição *Gau* (μ, σ) com σ conhecido, estão esquematizadas na tabela seguinte:

Hipótese Alternativa	***Região crítica***
i) H_1: $\mu \neq \mu_0$	$\overline{X} \leq \mu_0 - \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$ ou
	$\overline{X} \geq \mu_0 + \Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$
ii) H_1: $\mu > \mu_0$	$\overline{X} \geq \mu_0 + \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$
iii) H_1: $\mu < \mu_0$	$\overline{X} \leq \mu_0 - \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$
iv) H_1: $\mu = \mu_1$	$\overline{X} \geq \mu_0 + \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$, se $\mu_1 > \mu_0$
	$\overline{X} \leq \mu_0 - \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$, se $\mu_1 < \mu_0$

Tabela I

II- Com Variância σ^2 desconhecida

Na prática o mais usual é que a variância σ^2 seja desconhecida, nestes casos tudo se processa como anteriormente descrito, no entanto a distribuição a utilizar para encontrar a região de rejeição passa a ser a distribuição **t de Student** (situação análoga aos I.C. para o valor médio da distribuição Gaussiana).

Exemplo 4: Retomemos o exemplo da dieta dos ratos, mas supondo agora que a variância é desconhecida e que para a amostra observada se obteve como estimativa desta $s^2 = 295g^2$. Então, a estatística de teste tem $\overline{X} \cap Gau\left(10, \sqrt{\frac{\sigma^2}{16}}\right)$ sendo agora σ^2 desconhecido, assim teremos de o substituir pelo seu estimador S^2 vindo

$$\frac{\overline{X} - \mu}{S}\sqrt{n} \cap t_{n-1} \qquad (1)$$

Deveremos continuar a rejeitar a hipótese nula quando $|\overline{X} - 10| > k$, dado que 10 é o valor μ sob H_0 e a distribuição de $\overline{X}$ é simétrica em torno de μ, onde $k = t_{0,975;15} * \sqrt{\frac{295}{16}}$ ($t_{0,975;15}$ representa o quantil de probabilidade 0,975 da distribuição t com 15 g.l.). Consultando a tabela da distribuição t obtém-se finalmente

SQRT(295/16)	**t(0,975;15)**			
4,294	2,131		**k**	9,152

A região de Rejeição é

$$\mathbf{R}=\{(X_1,\dots,X_{16}) : \overline{X} > 19{,}15 \text{ ou } \overline{X} < 0{,}85\}.$$

Dado que $\overline{x} = 19{,}1 < 19{,}15$ e $\overline{x} = 19{,}1 > 0{,}85$, a amostra observada não pertence à região **R,** logo concluímos pela não rejeição da hipótese nula ao nível de 5%.

Que conclusão tiraríamos se o nível fosse de 1%?

Mais geralmente, as expressões da região crítica do teste da hipótese H_0: $\mu = \mu_0$ ao nível de significância α, para as várias formas da hipótese alternativa, dado que as observações X_1, X_2, ..., X_n têm distribuição *Gau* (μ, σ) com σ <u>desconhecido</u>, estão esquematizadas na tabela seguinte:

Hipótese Alternativa	***Região crítica***
i) H_1: $\mu \neq \mu_0$	$\overline{X} \leq \mu_0 - t_{1-\frac{\alpha}{2};n-1} \dfrac{S}{\sqrt{n}}$ ou
	$\overline{X} \geq \mu_0 + t_{1-\frac{\alpha}{2};n-1} \dfrac{S}{\sqrt{n}}$
ii) H_1: $\mu > \mu_0$	$\overline{X} \geq \mu_0 + t_{1-\alpha;n-1} \dfrac{S}{\sqrt{n}}$
iii) H_1: $\mu < \mu_0$	$\overline{X} \leq \mu_0 - t_{1-\alpha;n-1} \dfrac{S}{\sqrt{n}}$
iv) H_1: $\mu = \mu_1$	$\overline{X} \geq \mu_0 + t_{1-\alpha;n-1} \dfrac{S}{\sqrt{n}}$, se $\mu_1 > \mu_0$
	$\overline{X} \leq \mu_0 - t_{1-\alpha;n-1} \dfrac{S}{\sqrt{n}}$, se $\mu_1 < \mu_0$

Tabela II

3.4.4- TESTES PARA O PARÂMETRO p DA DISTRIBUIÇÃO Bi(n, p)

Até agora temos concentrado a nossa atenção em amostras provenientes de uma população gaussiana, que é uma distribuição contínua. Para as distribuições discretas, a teoria geral dos testes de hipóteses é a mesma, embora sejam necessárias umas certas modificações. Iremos debruçar-nos sobre a distribuição binomial, mas as ideias aplicam-se igualmente a outras distribuições discretas com um só parâmetro desconhecido, como por exemplo a Poisson.

Vejamos o seguinte

Exemplo 1- Num município dos arredores de Lisboa há 3 candidatos à Câmara Municipal, um destes, o João da Silva, reclama que terá 50% dos votos e portanto sairá vencedor. Os seus adversários políticos não acreditam nesta afirmação e tentam encontrar maneira de a refutar, para o que decidem encomendar uma sondagem. Se em **n** pessoas inquiridas, Y manifestarem intenção de votar no João da Silva, que valores para Y devemos observar de modo a podermos concluir pela evidência contra a afirmação do candidato, a um nível de significância de 5%?

A variável aleatória Y que designa o número de pessoas que votam no João segue a distribuição **Bi**(n, p), onde **p** (desconhecido) é a proporção de pessoas que tenciona votar no João, e que é estimado por $\frac{Y}{n}$. Segundo a afirmação do João **p** = 0.5, i.e., interessa-nos testar a hipótese H_0: p = 0.5 contra a alternativa H_1: p < 0.5, e os valores de $\frac{Y}{n}$ mais consistentes com a hipótese alternativa devem ser tais que $\frac{Y}{n} \leq k$ ou equivalentemente $Y \leq k'$ i.e., valores pequenos de Y.

Quão pequeno deve ser Y? A resposta a esta questão é, como anteriormente, dada pela condição:

$$k' : P(Y\leq k'/ p= 0.5) = 0.05 \qquad (1)$$

Para prosseguirmos no cáculo de k' precisamos de concretizar o valor de n. Seja n=25, então sob a validade da hipótese nula $Y \cap Bi(25{,}0.5)$ tendo-se

x	P(X=x)	F(x)
0	0,00000	0,00000
1	0,00000	0,00000
2	0,00001	0,00001
3	0,00007	0,00008
4	0,00038	0,00046
5	0,00158	0,00204
6	0,00528	0,00732
7	0,01433	0,02164
8	0,03223	0,05388

Assim, não existe valor k' que satisfaça a equação (1). Esta é uma situação típica das distribuições discretas. *Como resolver o problema?* Em vez de procurarmos um valor que satisfaça (1), tentaremos antes encontrar um

$$k' : P(Y\leq k'/ p= 0.5) \leq 0.05 \qquad (2)$$

e pela listagem acima, das probabilidades de alguns valores assumidos pela variável B(25, 0.5), obtém-se

$$P(Y\leq 7 / p= 0.5) = 0.02164 < 0.05 \qquad (3)$$

logo rejeita-se a hipótese H_0: p = 0.5 quando o número de pessoas que estão a favor do João da Silva, dentre as 25 pessoas inquiridas, não exceder 7.

(Note-se que a inclusão do valor 8 na região de rejeição iria aumentar o valor da probabilidade em (3) para 0.05388 > 0.05)

Usando a aproximação à normal, obtém-se

$$P\left(\frac{Y-np}{\sqrt{np(1-p)}} \le \frac{k'-np}{\sqrt{np(1-p)}} \,/p=0.5\right) \approx \Phi\left(\frac{k'-12.5}{2.5}\right) = 0.05 \Leftrightarrow k'=2.5\mathrm{x}\Phi^{-1}(0.05)+12.5 =$$

$$=-1.64\mathrm{x}2.5+12.5 = 8.4 \qquad (4)$$

e o problema anterior fica solucionado, dado que estamos a usar uma distribuição contínua.

Esta aproximação pode ainda ser melhorada se tomarmos em (4)

$$P(Y\le k')=P(Y\le k'+0.5)=P\left(\frac{Y-np}{\sqrt{np(1-p)}} \le \frac{k'+0.5-np}{\sqrt{np(1-p)}} \,/p=0.5\right) \approx \Phi\left(\frac{k'-12}{2.5}\right) \approx 0.05 \Leftrightarrow$$

$$\Leftrightarrow k' = 2.5\mathrm{x}\Phi^{-1}(0.05)+12 = 7.9 \qquad (5)$$

3.4.5. RELAÇÃO ENTRE INTERVALOS DE CONFIANÇA E TESTES DE HIPÓTESES

Consultando a tabela I na linha i), verificamos que a região de rejeição do teste das hipóteses H_0: $\mu=\mu_0$ contra H_1: $\mu\neq\mu_0$ (com σ^2 conhecido) é $\overline{X}\le\mu_0-\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$ ou $\overline{X}\ge \mu_0+\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$ ou equivalentemente

Rejeitar H_0 se e só se : $\mu_0 > \overline{X}+\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$ ou $\mu_0<\overline{X}-\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}$

tendo-se

$$P\left(\mu_0 > \overline{X}+\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n} \text{ ou } \mu_0<\overline{X}-\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n} \,/\mu=\mu_0\right)= \alpha \;\Leftrightarrow$$

$$\Leftrightarrow P\left(\overline{X}-\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}\le \mu_0 \le \overline{X}+\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n} \,/\mu=\mu_0\right) = 1-\alpha$$

ou seja

$\left(\overline{X}-\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n},\ \overline{X}+\Phi^{-1}(1-\alpha/2)\sigma/\sqrt{n}\right)$ é o *intervalo de (1- α)% de confiança para* μ

Isto é, a partir do teste da hipótese H_0: $\mu=\mu_0$ contra H_1: $\mu\neq\mu_0$ podemos obter um intervalo de confiança para μ e, vice-versa, se construirmos o intervalo de confiança para μ e desejarmos testar a hipótese H_0: $\mu=\mu_0$ contra H_1: $\mu\neq\mu_0$, podemos concluir pela rejeição da hipótese H_0: $\mu=\mu_0$ se o intervalo de confiança não contém o valor μ_0. Este facto põe em relevo a forte relação existente entre intervalos de confiança e testes de hipóteses.

Exercício: Foram observadas 6 baterias tendo-se registado as suas voltagens:

6.3, 6.2, 6.3, 5.9, 6.1, 5.8

Supondo que as voltagens seguem uma distribuição normal de variância igual 0.04, mostre que um intervalo de 90% de confiança para μ é (5.97, 6.22). Teste ao nível de significância de 10% a hipótese H_0: $\mu=6.0$ contra H_1: $\mu\neq6.0$

3.5 TESTE DE AJUSTAMENTO: TESTE DO QUI-QUADRADO

Falemos agora de um teste não paramétrico, que foi inicialmente desenvolvido como teste de ajustamento de um modelo aos dados classificados em classes.

Exemplo 1:

Num estudo sobre o desenvolvimento de colónias de bactérias numa solução de soro fisiológico procede-se à contagem do número de celas da retícula em que não se observa nenhuma bactéria, e em que há uma, duas, …, com os seguintes resultados:

Número de bactérias na cela	0	1	2	3	4	≥ 5
Número de celas	44	98	107	79	43	29

Pretendemos saber se o modelo $X \cap P(2)$ é adequado.

Se $H_0 : X \cap P(2)$ for verdadeira, as probabilidades de observar k bactérias numa cela são:

$$p_0 = P(X = 0) = e^{-2} \approx 0.135$$

$$p_1 = P(X = 1) = e^{-2} \times 2 \approx 0.271$$

$$p_2 = P(X = 2) = e^{-2} \times 2 \approx 0.271$$

$$p_3 = P(X = 3) = e^{-2} \times \frac{2^3}{3!} \approx 0.180$$

$$p_4 = P(X = 4) = e^{-2} \times \frac{2^4}{4!} \approx 0.090$$

$$p_5 = P(X \geq 5) = 1 - (X \in \{0,1,2,3,4\}) \approx 0.053$$

Seja O_k a variável aleatória que designa o número de celas com k partículas, sob a validade da hipótese nula, $O_k \frown Bi(n, p_k)$ e $e_k = E(O_k | H_o \text{ verdadeira}) = np_k$.

(Note-se que na última classe considerámos $X \geq 5$, de facto devemos ter uma partição para que $\sum p_k = 1$ e consequentemente $\sum e_k = \sum np_k = n = \sum o_k$).

A estatística de teste utilizada é:

$$X_6^2 = \sum_{k=0}^{5} \frac{(O_k - e_k)^2}{e_k} \tag{1}$$

Porquê? Porque esta expressão calcula a discrepância entre a realidade e as nossas suposições, medindo o erro quadrático relativizado global. Se a hipótese nula for verdadeira não esperamos observar grandes discrepâncias entre a realidade e o modelo assumido. Assim, valores excessivamente grandes desta estatística devem levar à rejeição da hipótese nula.

Pode-se mostrar que a distribuição assintótica desta estatística, sob a validade da hipótese nula, é um qui-quadrado com nº de g.l. ν =6-1, χ_5^2. Então para um nível de significância α devemos rejeitar H_0 quando

$$X_6^2 = \sum_{k=0}^{5} \frac{(O_k - e_k)^2}{e_k} > \chi_{1-\alpha;5} \tag{2}$$

Vejamos agora o valor que esta estatística toma para a amostra observada:

k	o_k	$e_k \| H_0$ *verd.*	$\frac{(o_k - e_k)^2}{e_k}$
0	44	54.0	1.897
1	98	108.4	0.974
2	107	108.4	0.015
3	79	72.0	0.645
4	43	36.0	1.323
≥ 5	29	21.2	2.992
Σ	**400**	**400**	$X_6^2(obs.) = \mathbf{7.846}$

$X_6^2(obs.) = \sum_{k=0}^{5} \frac{(o_k - e_k)^2}{e_k}$ =7.846 e dado que o $\chi_{0.95;5} = \mathbf{11.071}$, concluímos que face à amostra observada *não rejeitamos* $H_0 : X \frown P(2)$.

Exemplo 2:

O número Y de acidentes por semana num determinado cruzamento foi registado para n=50 semanas, tendo-se obtido

Y	**frequência**
0	32
1	12
2	6
3 ou mais	0

Pretende-se testar a hipótese $H_0 : X \cap P(\lambda) \quad \lambda > 0$, supondo que as observações são independentes. Vamos tomar α =5%

Uma vez que λ é desconhecido temos de o estimar a partir da amostra observada, usando a estatística $\overline{Y} = \hat{\lambda}$. A média desta amostra é, $\overline{y} = \frac{0 \times 32 + 1 \times 12 + 2 \times 6 + 3 \times 0}{32 + 12 + 6} = \frac{24}{50} = 0,48$ e as estimativas dos p_k são:

$$\hat{p}_0 = P(X = 0) = e^{-0.48} \approx 0,619$$

$$\hat{p}_1 = P(X = 1) = e^{-0.48} \times 0.48 \approx 0,297$$

$$\hat{p}_2 = P(X = 2) = e^{-0.48} \times \frac{0.48^2}{2!} \approx 0,071$$

$$\hat{p}_3 = P(X \geq 3) = 1 - (X \in \{0,1,2\}) \approx 0,013$$

Podemos agora calcular o valor da estatística de teste.

k	o_k	$e_k \mid H_0$ verd.	$\frac{(o_k - e_k)^2}{e_k}$
0	32	30.94	0.036
1	12	14.85	0.547
2	6	3.56	1.665
≥ 3	0	0.65	0.646
Σ	**50**	**50**	$X_4^2(obs.) = \mathbf{2,894}$

Uma vez que se estimou o parâmetro λ, a distribuição assintótica da estatística X_4^2 tem menos um g.l.do que a do exemplo anterior, isto é,

$$X_4^2 = \sum_{k=0}^{3} \frac{(O_k - e_k)^2}{e_k} \overset{a}{\cap} \chi_2^2 \qquad (3)$$

(O número de g.l. ν da distribuição assintótica é igual a ν =4-1-nº parâmetros estimados). Assim o p-value deste teste para a amostra observada é $P(\chi_2^2 \geq 2.894) = 0.235$ ($\chi_{0.95;2=5.991}$), logo não rejeitamos nula $H_0 : X \cap P(\lambda) \quad \lambda > 0$

para os níveis de significância usuais.

Podemos concluir que de uma forma geral a estatística de teste tem distribuição assintótica χ^2_ν, isto é,

$$X^2_N = \sum_{k=1}^{N} \frac{(O_k - e_k)^2}{e_k} \overset{a}{\cap} \chi^2_\nu \qquad (4)$$

Onde o número de graus de liberdade ν é igual a

$$\nu = N - 1 - n^o \text{ de parâmetros estimados} \qquad (5)$$

(N é o número de classes da tabela de frequências). E rejeita-se a hipótese nula ao nível de significância α quando

$$X^2_N = \sum_{k=1}^{N} \frac{(O_k - e_k)^2}{e_k} > \chi_{1-\alpha;\nu} \qquad (6)$$

Exercício: Os dados que se seguem representam as alturas (X) em cm de 40 estudantes da FCUL:

152.2, 167.9, 169.4, 163.5, 163.2, 163.3, 159.2, 167.3, 167.3, 162.3, 162.6, 165.1, 168.5, 165.9, 164.8,163.0, 166.7, 168.2, 166.8, 165.0, 165.4, 166.3, 158.5, 170.8, 172.5, 164.0, 164.2, 164.5, 160.3, 170.8, 175.6, 168.6, 172.4, 161.6, 157.9, 164.6, 158.9, 165.0, 171.9, 159.4

A média, o desvio-padrão, o mínimo e o máximo desta amostra são respectivamente $\bar{x}$=165.13 e s=4.64, $x_{(1)}$=152.2 e $x_{(40)}$ = 175.6.

Teste ao nível de significância de 5%, a hipótese $H_0 : X \cap Gau(\mu,\sigma)$.

Para a resolução siga os seguintes passos:

1- Comece por ordenar a amostra

152.2, 157.9, 158.5, 158.9, 159.2, 159.4, 160.3, 161.6, 162.3, 162.6, 163.0, 163.2, 163.3, 163.5, 164.0, 164.2, 164.5, 164.6, 164.8, 165.0, 165.0, 165.1, 165.4, 165.9, 166.3, 166.7, 166.8, 167.3, 167.3, 167.9, 168.2, 168.5, 168.6, 169.4, 170.8, 170.8, 171.9, 172.4, 172.5, 175.6.

2- Construa uma tabela de frequências absolutas, usando uma amplitude de classe h=4.

3- Estime as probabilidades de X pertencer a cada classe sob a validade da hipótese nula, usando as estimativas obtidas dos parâmetros μ e σ .

4- Calcule as frequências esperadas (e_k) de cada classe.

5- Calcule o valor da estatística de teste para a amostra observada.

6- Que conclui? Justifique a sua resposta.