Índice Geral

1 Variáveis aleatórias e distribuições de probabilidade **5**
1.1 Noção de probabilidade . . . 6
1.2 Probabilidade condicional . . . 9
1.3 Variáveis aleatórias . . . 12
1.4 Função de probabilidade, função densidade de probabilidade e função de distribuição . . . 13
1.5 Probabilidades e variáveis aleatórias - exemplo no caso contínuo . . . 16
1.6 Medidas de localização e escala de uma variável aleatória . . . 20

2 Distribuições de probabilidade especiais **29**
2.1 Distribuições de probabilidade especiais . . . 30
2.1.1 Distribuição Uniforme Discreta: $U(N)$ . . . 31
2.1.2 Distribuição Binomial $B(n,p)$ . . . 32
2.1.3 Distribuição de Poisson: $P(\lambda)$ . . . 48
2.1.4 Distribuição Multinomial $M(n_1,\ldots,n_k;p_1,\ldots,p_k)$ . . . 70
2.1.5 Distribuição Uniforme Contínua: $U(a,b)$ . . . 71
2.1.6 Distribuição Normal: $N(\mu,\sigma^2)$ . . . 73
2.1.7 Distribuição χ^2 de Pearson: $\chi^2(\nu)$ . . . 96
2.1.8 Distribuição t de Student: $t(\nu)$ . . . 106
2.1.9 Distribuição F de Snedecor: $F(\nu_1,\nu_2)$ . . . 112
2.2 Relações entre algumas distribuições - teoremas de aproximação . . . 114

3 Estatística Descritiva **121**
3.1 Amostragem . . . 122

3.2 Medidas de localização de uma a.a. 125
3.3 Medidas de escala de uma a.a. 127
3.4 Erro padrão (amostral) da média 129
3.5 Coeficiente de Variação 130
3.6 Gráficos associados a uma amostra de uma variável contínua 131
3.7 $\hat{F}$, boxplot e quantis: instruções em **SPSS** e exemplos 136
3.8 Estatística Descritiva: Exemplos 142

4 Estatísticas Amostrais **157**
4.1 Estatísticas 158
4.2 Lei dos Grandes Números 159
4.3 Teorema do Limite Central 161
4.4 Distribuição da média amostral num exemplo concreto 165
4.5 Alguma teoria sobre estimadores pontuais* 167
4.6 Algumas estatísticas amostrais mais comuns 170
4.7 Intervalos de confiança 175
4.8 Intervalos de confiança - populações normais 181
4.9 Intervalos de confiança - amostras grandes 183
4.10 Intervalos de confiança - amostras grandes 185
4.11 Intervalos de confiança: Exemplos 187
4.12 Testes de Hipóteses 191

Chapter 1

Variáveis aleatórias e distribuições de probabilidade

1.1. Noção de probabilidade

Experiências aleatórias: os resultados obtidos não são sempre (essencialmente) os mesmos ainda que as condições de realização das experiências se mantenham praticamente as mesmas.

Exemplos:

(a) lançar uma moeda ao ar e registar a face que fica voltada para cima

(b) lançar um dado e registar a face que fica voltada para cima

(c) seleccionar aleatoriamente uma determinada região e várias famílias dessa região com dois filhos, e registar a ordem do sexo dos filhos dessas famílias

(d) seleccionar aleatoriamente um grupo de alunos inscritos em 2009/2010 na UP e registar o peso (em Kg) desses alunos.

Espaço de probabilidade $(\Omega, \mathcal{A}, P)$

- Ω = **espaço amostral**
  = conjunto de todos os resultados de uma experiência aleatória

  (a) $\Omega = \{H, T\}$ H *head*, T *toss*

  (b) $\Omega = \{1, 2, 3, 4, 5, 6\}$

  (c) $\Omega = \{♂♂, ♂♀, ♀♀, ♀♂\}$

  (d) $\Omega = [40, 150]$

- $\mathcal{A}$ = conjunto de partes de Ω
  = conjunto dos **acontecimentos** [a] mensuráveis,

$$\emptyset \in \mathcal{A}, \qquad A \in \mathcal{A} \Longrightarrow \overline{A} = \Omega \setminus A \in \mathcal{A}$$

$$A_n \in \mathcal{A}, \quad n = 1, 2, \ldots \Longrightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$$

  (a) $\mathcal{A} = \{\{H\}, \{T\}, \{H, T\}\}$

  (b) $\mathcal{A} = \{\{1\}, \{2\}, \{3\}, \cdots, \{1, 2\}, \cdots, \{1, 2, 3, 4, 5, 6\}\}$

  (c) $\mathcal{A} = \{\{♂♂\}, \{♂♀\}, \cdots, \{♂♂, ♂♀\}, \{♂♂, ♀♂\}, \cdots$
  $\cdots, \{♂♂, ♂♀, ♀♀\}, \cdots, \{♂♂, ♂♀, ♀♀, ♀♂\}\}$

  Diz-se que:

  – $\emptyset$ é o **acontecimento impossível**

  – Ω é o **acontecimento certo**

  – um acontecimento que consiste de um único ponto é um **acontecimento elementar**

  – $A, B \subset \mathcal{A}$ são acontecimentos **mutuamente exclusivos** [b] se $A \cap B = \emptyset$

- $P : \mathcal{A} \to \mathbb{R}$ **função de probabilidade**

  $P(A)$ = probabilidade de se realizar o acontecimento A

$$P(A) \geq 0$$

$$P(\Omega) = 1$$

$$A_n \cap A_m = \emptyset (n \neq m) \Longrightarrow P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

[a] Um **acontecimento** é então um qualquer subconjunto de Ω, i.e., um qualquer conjunto de resultados possíveis de uma experiência aleatória

[b] Acontecimentos mutuamente exclusivos são também designados como **incompatíveis** ou **disjuntos**

Propriedades: Sejam A e B dois acontecimentos.

(a) $A \subseteq B \Longrightarrow P(A) \leq P(B)$,
$P(B \setminus A) = P(B) - P(A)$

(b) $0 \leq P(A) \leq 1$

(c) $P(\overline{A}) = P(\Omega \setminus A) = 1 - P(A)$

(d) $P(\emptyset) = 0$

(e) $A_1,\ A_2,\ \ldots,\ A_n$ acontecimentos mutuamente exclusivos e $A = A_1 \cup A_2 \cup \cdots \cup A_n \Longrightarrow$
$P(A) = P(A_1) + P(A_2) + \cdots + P(A_n)$

(f) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(g) $B_1, B_2, \ldots, B_n$ acontecimentos mutuamente exclusivos e $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n \Longrightarrow$
$P(A) = P(A \cap \Omega) = \sum_{i=1}^{n} P(A \cap B_i)$.

Questão:
que probabilidade atribuir a cada acontecimento?

Se o espaço amostral consiste apenas dos acontecimentos elementares $A_1,\ A_2,\ \ldots,\ A_n$ então

$$P(A_1) + P(A_2) + \cdots + P(A_n) = 1.$$

Se admitirmos probabilidades iguais para todos os acontecimentos elementares, tem-se

$$P(A_i) = \frac{1}{n}, \qquad i = 1, 2, \ldots, n.$$

Em particular, se A é um acontecimento formado por k acontecimentos elementares então

$$P(A) = k/n.$$

Exemplo 1: No lançamento de um dado não viciado,

(a) a probabilidade de sair a face i, é

$$P(i) = 1/6 \qquad (i = 1, 2, \ldots, 6)$$

dado que qualquer face tem a mesma probabilidade de sair voltada para cima (o dado é não viciado) e existem seis faces possíveis.

(b) a probabilidade de sair 2 ou 5 é

$$P(2 \cup 5) = P(2) + P(5) = 1/3.$$

dado que a saída de 2 e de 5 são acontecimentos mutuamente exclusivos.

(c) a probabilidade de não sair 6 é
$P(\text{não sair } 6) = 1 - P(6) = 1 - 1/6 = 5/6.$

(d) a probabilidade de não sair 1 nem 2 é

$$\begin{aligned} P(\text{não sair 1 nem 2}) &= 1 - P(1 \cup 2) \\ &= 1 - (P(1) + P(2)) \\ &= 1 - 1/3 \\ &= 2/3 \end{aligned}$$

pois o acontecimento que se pretende é o complementar do acontecimento "sair 1 ou sair 2".

Se após n experiências se observarem k ocorrências de um acontecimento A então a **probabilidade empírica** de A é

$$\widehat{P}(A) = \frac{k}{n}.$$

Exemplo 2: se em 100 jogadas de uma moeda, a face "cara" sair 47 vezes, então a probabilidade empírica do acontecimento "sair cara" é

$$\widehat{P}(\text{"sair cara"}) = \frac{47}{100} = 0.47$$

Mesmo que a moeda seja não viciada, esta probabilidade não coincide exactamente com a *probabilidade teórica* de 0.5 de sair cara!

Exemplo 3: Suponhamos que estudos realizados sobre o teste da tuberculina numa determinada região afirmam que a probabilidade deste teste ser positivo é de 0.1. Isto significa que:

- para um número **grande** de pessoas testadas, se espera uma percentagem aproximada de 10% de indivíduos com resultado positivo
- à medida que se aumentar o tamanho da amostra mais o resultado se aproximará dos 10%.
- o resultado será exactamente 10% se se conseguirem avaliar todos os indivíduos dessa região.

Exemplo 4: Um estudo afirma que a probabilidade de aparecimento de cancro da mama num período de 30 anos em mulheres com idade superior a 40 anos e que nunca tiveram esse tipo de cancro é de 1/10. Isto significa que:

- se escolhermos *à sorte* uma mulher com idade superior a 40 anos e que nunca teve cancro da mama, ela tem uma probabilidade de 0.1 de vir a desenvolver esse tipo de cancro durante os próximos 30 anos.
- se considerarmos uma amostra *aleatória* constituída por *muitas* mulheres com idade superior a 40 anos e que nunca tiveram cancro da mama, é de esperar que durante os próximos 30 anos aproximadamente 10% dessas mulheres venham a sofrer de cancro da mama. Quanto maior for a amostra considerada, melhor será a aproximação do valor em causa aos 10%.

1.2. Probabilidade condicional

Sejam A e B dois acontecimentos de um espaço de probabilidade $(\Omega, \mathcal{A}, P)$ e suponhamos que se tem $P(A) > 0$.
A **probabilidade condicional** $P(B|A)$ é a probabilidade de se realizar o acontecimento B na hipótese de A ter ocorrido. Dado que o novo espaço amostral passa a ser A, define-se

$$P(B|A) = \frac{P(B \cap A)}{P(A)}\,^{a}.$$

Exemplo 1:
Determinar a probabilidade de, no lançamento de um dado, sair um número menor do que 5 sabendo que o resultado é um número par.

Sejam A o acontecimento "número par" e B o acontecimento "menor do que 5". Pretende-se determinar $P(B|A)$. Como

$$P(A) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2},$$
$$P(A \cap B) = P(2) + P(4) = \frac{2}{6} = \frac{1}{3}$$

vem

$$P(B|A) = \frac{1/3}{1/2} = \frac{2}{3}.$$

Observe-se que $P(A \cap B)$ é a probabilidade de sair um número menor do que 5 que é par, o que é diferente da probabilidade de sair um número menor do que 5 sabendo que o resultado foi um número par.

[a] Como seria de esperar, $P(B|\Omega) = P(B)$ e $P(B|B) = 1$

Diz-se que A e B são **acontecimentos independentes** se

$$P(B|A) = P(B)$$

ou, equivalentemente,

$$P(A \cap B) = P(A)P(B).$$

Dois acontecimentos independentes são dois eventos em relação aos quais a ocorrência ou não ocorrência de um deles não altera a probabilidade de ocorrência do outro.

Exemplo 2: em dois lançamentos consecutivos de um dado não viciado, os acontecimentos

A = "saída de 1 no primeiro lançamento"
B = "saída de 2 no segundo lançamento"

são independentes. Tem-se $P(A) = P(B) = 6/36 = 1/6$ e portanto a probabilidade de sair (1,2) em dois lançamentos consecutivos é

$$P(A \cap B) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}.$$

Exemplo 3: Sabe-se que, em famílias de uma determinada população, a probabilidade:

- da mãe ser hipertensa é 0.2
- do pai ser hipertenso é 0.25
- de um dos pais ser hipertenso é 0.35.

Os acontecimentos "ser mãe hipertensa" (MH) e "ser pai hipertenso" (PH) são independentes?

Resolução: Os acontecimentos MH e PH serão independentes se

$$P(MH \cap PH) = P(MH)\,P(PH). \tag{1.1}$$

Sabe-se que $P(MH) = 0.2$ e $P(PH) = 0.25$ mas desconhece-se $P(MH \cap PH)$ portanto nada se pode concluir daqui. Contudo, do enunciado resulta também que

$$P(MH \cup PH) = 0.35$$

e viu-se já que, para a reunião de dois acontecimentos, no caso $MH \cup PH$, é válida a fórmula

$$P(MH \cup PH) = P(MH) + P(PH) - P(MH \cap PH)$$

portanto

$$P(MH \cap PH) = 0.2 + 0.25 - 0.35 = 0.1.$$

Voltando à equação (1.1), como

$$0.1 \neq 0.2 \times 0.25$$

conclui-se que os acontecimentos MH e PH não são independentes.

Teorema 1.2.1 *Se A é um acontecimento qualquer e $A_1, \ldots, A_n$ são acontecimentos disjuntos cuja reunião é o espaço amostral ($A_1 \cup \cdots \cup A_n = \Omega$), então*

(a) teorema da probabilidade total

$$P(A) = P(A|A_1)P(A_1) + \cdots + P(A|A_n)P(A_n)$$

(b) teorema de Bayes

$$P(A_k|A) = \frac{P(A|A_k)P(A_k)}{\sum_{j=1}^{n} P(A|A_j)P(A_j)}.$$

Exemplo 4: Todos os indivíduos adultos de uma determinada comunidade são classificados de acordo com o sexo em

feminino (F) ou masculino (M)

e de acordo com o estado civil em

solteiro (S), casado (C) ou outro (O).

Para essa comunidade, sabe-se que:

- 25% dos indivíduos são solteiros e 65% são casados.
- de entre os solteiros, 65% são mulheres
- de entre os casados, 40% são homens
- de entre os indivíduos com estado civil "outro", 45% são homens.

Perguntas:

(a) Escolheu-se aleatoriamente uma mulher da população; qual é a probabilidade de ela ser solteira?

(b) Escolheu-se aleatoriamente um indivíduo da população em causa; qual é a probabilidade de ele ser do sexo masculino?

Resolução: Os dados do problema mostram que:

$$P(S) = 0.25, \quad P(C) = 0.65,$$
$$P(F|S) = 0.65, \quad P(M|C) = 0.4, \quad P(M|O) = 0.45.$$

(a) Pretende-se determinar $P(S|F)$. Pelo teorema de Bayes, tem-se

$$P(S|F) = \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|C)P(C) + P(F|O)P(O)}.$$

Agora

- $P(O) = 1 - P(S) - P(C) = 0.1$
- $P(F|C) = 1 - P(M|C) = 0.6$
- $P(F|O) = 1 - P(M|O) = 0.55$

portanto

$$\begin{aligned} P(S|F) &= \frac{0.65 \times 0.25}{0.65 \times 0.25 + 0.6 \times 0.65 + 0.55 \times 0.1} \\ &\approx 0.27\,. \end{aligned}$$

(b) Pretende-se determinar $P(M)$. Tem-se

$$\begin{aligned} &P(M) \\ &= 1 - P(F) \\ &= 1 - P(F|S)P(S) - P(F|C)P(C) - P(F|O)P(O) \\ &\approx 0.39\,. \end{aligned}$$

1.3. Variáveis aleatórias

Uma variável aleatória (v.a.) é uma função X que atribui um número a cada um dos resultados possíveis de uma experiência aleatória,

$$X : \Omega = \{\text{espaço amostral}\} \longrightarrow \mathbb{R}.$$

Usualmente as variáveis aleatórias são designadas por letras maiúsculas enquanto que os valores que elas podem tomar são designados por letras minúsculas. Por exemplo, $P(X = x)$ representa a probabilidade da v.a. X tomar o valor (teórico) x.

Uma v.a. pode ser:

- **discreta** (ou **categórica** ou **qualitativa**): se toma um número finito de valores ou um número infinito *numerável* de valores.
  Uma v.a. discreta pode ser:
  - **ordinal**: se existe uma ordem natural entre os diferentes valores da v.a.
  - **nominal**: caso contrário.
- **contínua** (ou **quantitativa**): se toma um número infinito *não numerável* de valores.

Exemplos de v.a. discretas:

(a) X v.a. (ordinal) que representa o resultado do lançamento de um dado;

$X : \{1, 2, 3, 4, 5, 6\} \to \mathbb{R}$, $i \mapsto i$.

(b) X v.a. (nominal) que representa o resultado do lançamento de uma moeda;

$X : \{H, T\} \to \mathbb{R}$, $X(H) = 1$ e $X(T) = 0$.

(c) X v.a. (ordinal) que representa a resposta a um determinado tratamento (excelente, boa, razoável, má);

$X : \{\text{exc}, \text{boa}, \text{raz}, \text{má}\} \to \mathbb{R}$,
$X(\text{exc}) = 3$, $X(\text{boa}) = 2$, $X(\text{raz}) = 1$, $X(\text{má}) = 0$.

(d) X v.a. (nominal) que representa o sexo dos filhos em famílias com 2 filhos;

$X : \{♂♂, ♂♀, ♀♀, ♀♂\} \to \mathbb{R}$,
$X(♂♂) = 1$, $X(♂♀) = 2$, $X(♀♀) = 3$, $X(♀♂) = 4$

(e) X v.a. (ordinal) que representa o número de episódios de otites verificados até aos 2 anos de idade.

$X : \{0, 1, 2, 3, \cdots\} \to \mathbb{R}, \quad X(x) = x$

(f) O teste de um novo fármaco passa pela prescrição do mesmo aos primeiros 4 pacientes que necessitem de medicação análoga.

X v.a. (ordinal) que representa o número de pacientes, em 4, que melhoram com o novo fármaco

$X : \{0, 1, 2, 3, 4\} \to \mathbb{R}, \quad X(x) = x.$

Exemplos de v.a. contínuas

(a) X v.a. que representa o peso (em g) dos alunos inscritos na FMUP no ano lectivo 2008/2009;

$X : [45\,000, 130\,000] \to \mathbb{R}, \quad X(x) = x.$

(b) X v.a. que representa o pH de uma solução;

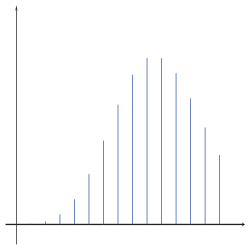
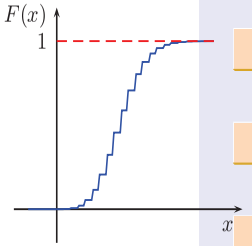
$X : [0, 14] \to \mathbb{R}, \quad X(x) = x.$

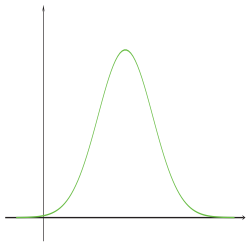
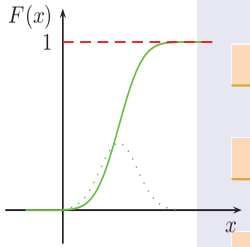
(c) X v.a. que representa a quantidade (em ml) de água ingerida por dia, por indivíduo da população do grande Porto, durante o mês de Outubro;

$X : [0, 10\,000] \to \mathbb{R}, \quad X(x) = x.$

1.4. Função de probabilidade, função densidade de probabilidade e função de distribuição

$X : \Omega = \{\text{espaço amostral}\} \longrightarrow \mathbb{R}$ **variável aleatória**

Variável	Função de prob.	Função de distrib.	Gráfico de f	Gráfico de F
Discreta $X \in \{a_1, a_2, \ldots, a_N\}$ ou $\{a_1, \ldots, a_n, \ldots\}$	$f : \{\text{valores de } X\} \to \mathbb{R}$ $f(x) \geq 0$ $\sum_x f(x) = 1$ $\mathbf{f(x) = P(X = x)}$	$F : \mathbb{R} \longrightarrow [0,1]$ $F(x) = P(X \leq x)$ $= \sum_{u \leq x} f(u)$		

Variável	Função densid. de prob.	Função de distrib.	Gráfico de f	Gráfico de F
Contínua $X \in \mathbb{R}$	$f : \mathbb{R} \longrightarrow \mathbb{R}$ $f(x) \geq 0$ $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ $\mathbf{P(X = x) = 0}$!!	$F : \mathbb{R} \longrightarrow [0,1]$ $F(x) = P(X \leq x)$ $= \int_{-\infty}^{x} f(u)\,du$		

Exemplo - Caso discreto:

Considere-se novamente a v.a. X que conta o número de pacientes nos quais um determinado fármaco produz o efeito desejado, de entre os primeiros 4 a necessitarem de medicação análoga.

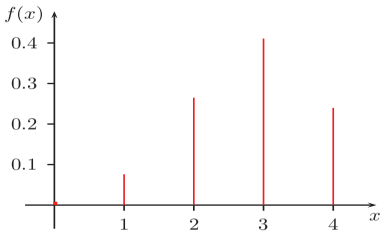
$$X : \{0, 1, 2, 3, 4\} \to \mathbb{R}, \quad X(x) = x.$$

De experiências anteriores (efectuadas em *muitos* grupos de 4 pacientes), sabe-se que:

- a probabilidade de nenhum indivíduo, em 4, melhorar com o novo fármaco é 0.008
- a probabilidade de 1 indivíduo, em 4, melhorar com o novo fármaco é 0.076
- a probabilidade de 2 indivíduos, em 4, melhorarem com o novo fármaco é 0.265
- a probabilidade de 3 indivíduos, em 4, melhorarem com o novo fármaco é 0.411
- a probabilidade de 4 indivíduos, em 4, melhorarem com o novo fármaco é 0.240

A função de probabilidade[a] de X é $f : \{0, 1, 2, 3, 4\} \to \mathbb{R}$

$$f(0) = 0.008, \quad f(1) = 0.076, \quad f(2) = 0.265,$$
$$f(3) = 0.411, \quad f(4) = 0.240.$$



[a] $f(x) = P(X = x)$; repare-se que $f(x) \geq 0$ e $\sum_x f(x) = 1$.

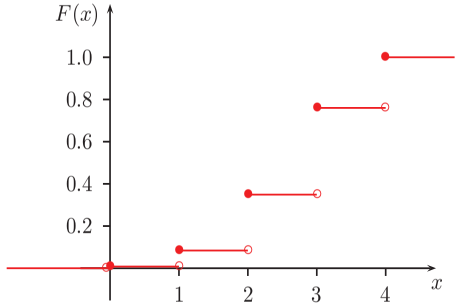
Para o cálculo da função de distribuição recorde-se que, por X ser discreta,

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u).$$

Assim $F : \mathbb{R} \to \mathbb{R}$ é a função definida por

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.008, & 0 \leq x < 1 \\ 0.008 + 0.076 = 0.084, & 1 \leq x < 2 \\ 0.084 + 0.265 = 0.349, & 2 \leq x < 3 \\ 0.349 + 0.411 = 0.760, & 3 \leq x < 4 \\ 0.760 + 0.240 = 1, & 4 \leq x \end{cases}$$

que, tal como em todos os casos discretos, é uma **função em escada**:



Nota: Os valores de $f(x)$ apresentados aqui são teóricos. Isto significa que as frequências relativas obtidas por um clínico que quisesse conduzir uma experiência análoga numa amostra da população considerada não seriam, à partida, exactamente iguais aos valores de f mas tenderiam para esses valores à medida que o tamanho da amostra fosse aumentando.

Algumas cálculos e observações:

- A probabilidade de o fármaco produzir o efeito desejado em 2 indivíduos ou menos (em 4) é

$$P(X \leq 2) = F(2) = 0.349$$

- A probabilidade de o fármaco produzir o efeito desejado em menos de 2 indivíduos (em 4)

$$P(X < 2) = P(X \leq 1) = F(1) = 0.084$$

- A probabilidade de o fármaco produzir o efeito desejado em 2 ou mais indivíduos (em 4) é

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - P(X \leq 1) \\ &= 1 - F(1) \\ &= 1.0.084 \\ &= 0.916 \end{aligned}$$

ou, alternativamente

$$\begin{aligned} P(X \geq 2) &= P(X = 2 \text{ ou } X = 3 \text{ ou } X = 4) \\ &= f(2) + f(3) + f(4) \\ &= 0.265 + 0.411 + 0.240 \\ &= 0.916. \end{aligned}$$

Função de probabilidade de X
$f : \{0, 1, 2, 3, 4\} \to \mathbb{R}$ definida por

$$\begin{array}{lll} f(0) = 0.008, & f(1) = 0.076, & f(2) = 0.265, \\ f(3) = 0.411, & f(4) = 0.240. & \end{array}$$

Função de distribuição de X
$F : \mathbb{R} \to \mathbb{R}$ definida por

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.008, & 0 \leq x < 1 \\ 0.008 + 0.076 = 0.084, & 1 \leq x < 2 \\ 0.084 + 0.265 = 0.349, & 2 \leq x < 3 \\ 0.349 + 0.411 = 0.760, & 3 \leq x < 4 \\ 0.760 + 0.240 = 1, & 4 \leq x \end{cases}$$

1.5. Probabilidades e variáveis aleatórias - exemplo no caso contínuo

Exemplo 1: Seja X a v.a. que representa os valores (mmHg) da pressão arterial diastólica (PAD) na população[a] feminina adulta de uma determinada comunidade. Supondo que

$$P(90 < X < 95) = 0.1, \quad P(X < 90) = 0.7$$
$$P(75 < X < 100) = 0.7, \quad P(X > 100) = 0.15$$

calcule:

(a) $P(X > 90)$

(b) $P(X < 95)$

(c) $P(X < 75)$

(d) $P(75 < X < 90)$

(e) $P(90 < X < 100)$

(f) Dentro do contexto do problema, dê exemplo de 2 eventos mutuamente exclusivos e 2 não mutuamente exclusivos (para X).

Resolução: As alíneas seguintes usam o facto de $P(X = x) = 0$ para uma v.a. contínua.

(a) $P(X > 90) = 1 - P(X \leq 90) = 1 - 0.7 = 0.3$

(b)
$$\begin{aligned} P(X < 95) &= P(90 < X < 95 \quad \text{ou} \quad X < 90) \\ &= P(90 < X < 95) + P(X < 90)^b \\ &= 0.1 + 0.7 \\ &= 0.8 \end{aligned}$$

[a] não é numa amostra.
[b] $P(90 < X < 95) + P(X < 90) - P(90 < X < 95 \text{ e } X < 90)$

(c)
$$\begin{aligned} &P(X < 75) \\ &= 1 - P(X > 75) \\ &= 1 - (P(75 < X < 100) + P(X > 100)) \\ &= 1 - (0.70 + 0.15) \\ &= 0.15 \end{aligned}$$

(d)
$$\begin{aligned} &P(75 < X < 90) \\ &= P(75 < X < 100 \text{ e } X < 90) \\ &= P(75 < X < 100) + P(X < 90) - P(X < 100)^a \\ &= 0.7 + 0.7 - 0.85 \\ &= 0.55 \end{aligned}$$

(e)
$$\begin{aligned} &P(90 < X < 100) \\ &= 1 - [P(X < 90 \text{ ou } X > 100)] \\ &= 1 - \left[P(X < 90) + P(X > 100)^b\right] \\ &= 1 - (0.7 + 0.15) \\ &= 0.15 \end{aligned}$$

(f) 2 eventos mutuamente exclusivos:

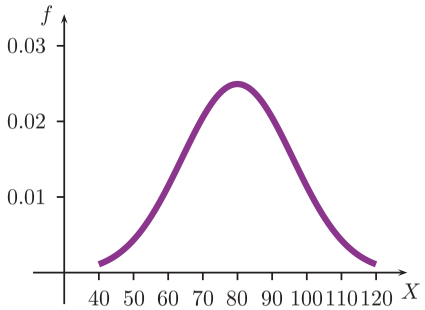
$X < 100$ e $X > 100$ (por exemplo)

2 eventos que não são mutuamente exclusivos:

$75 < X < 100$ e $X < 90$ (por exemplo).

[a] $P(75 < X < 100 \text{ ou } X < 90) = P(X < 100)$
[b] $P(X < 90 \text{ e } X > 100) = 0$

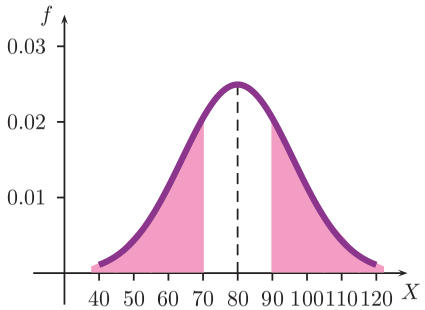
Exemplo 2: Seja X a v.a. que representa os valores (mmHg) da pressão arterial diastólica (PAD) numa população de homens com idades compreendidas entre os 33 e 44 anos e suponha que a f.d.p. de X é a apresentada abaixo.



Várias observações podem ser feitas:

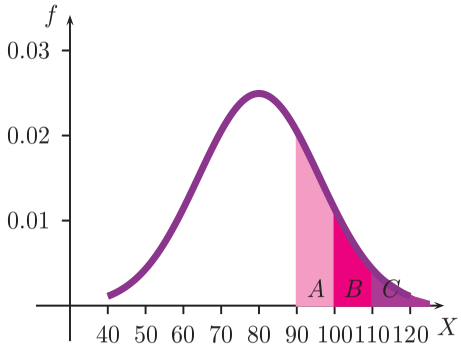
- o gráfico da f.d.p. é em forma de sino
- a f.d.p. é **simétrica** relativamente a $X = 80$:

  $P(X > 80+y) = P(X < 80-y)$ para qualquer $y \geq 0$.



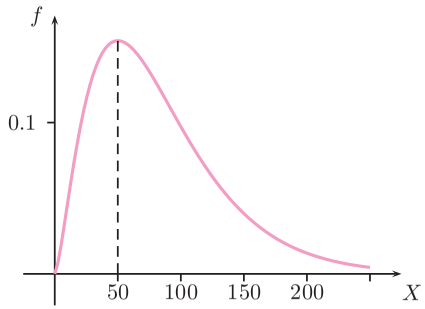
Em particular, $P(X \geq 80) = P(X \leq 80) = 0.5$.

- Os valores mais prováveis de PAD para um indivíduo da população em causa são os que rondam os 80 mmHg sendo que a probabilidade de um indivíduo ter PAD cada vez mais diferente de 80 é cada vez menor.
- A probabilidade de um indivíduo da população em causa ser:
  - ligeiramente hipertenso (90<PAD<100) é $P(90 < X < 100) = \text{área}(A)$
  - moderadamente hipertenso (100<PAD<110) é $P(100 < X < 110) = \text{área}(B)$
  - severamente hipertenso (PAD>110) é $P(X > 110) = \text{área}(C)$



- A f.d.p. apresentada (e as conclusões anteriores) dizem respeito à <u>população</u> de indivíduos do sexo masculino com idades compreendidas entre os 33 e os 44 anos, e não a uma amostra de indivíduos com essas características.

Exemplo 3: Seja X a v.a. que representa os níveis de triglicéridos no sangue (mg/dL) de uma determinada população. Suponha que estudos realizados anteriormente conduziram à seguinte f.d.p. para X:



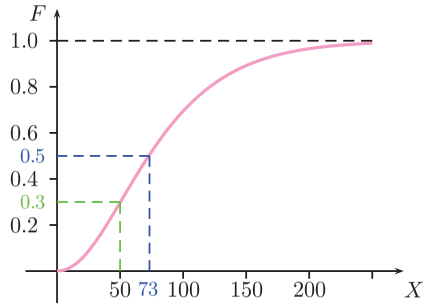
Algumas observações:

- os valores com maior probabilidade de ocorrência rondam os 50 mg/dL
- a distribuição não é simétrica (em relação a nenhum valor). Por exemplo, para $X = 50$, tem-se

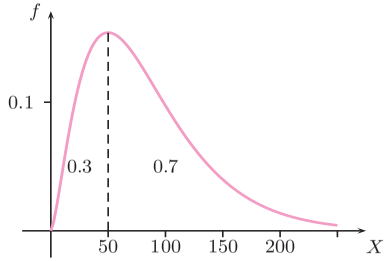
$$P(X \geq 50) > P(X \leq 50)$$

- a distribuição é assimétrica à direita
- a probabilidade de encontrar um indivíduo da população com um nível de triglicéridos superior a 200 mg/mL é muito baixa
- dado um qualquer indivíduo da população, a probabilidade do seu nível de triglicéridos ser inferior a 50 mg/dL é superior à probabilidade desse nível ser superior a 150 mg/dL

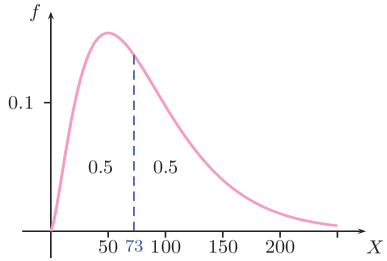
A função de distribuição de X é



- $P(X \leq 50) = F(50) = 0.3 \quad P(X \geq 50) = 0.7$

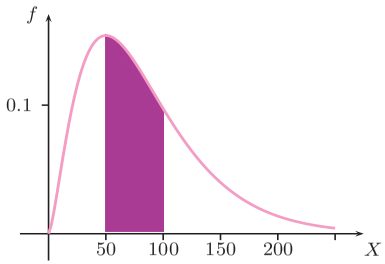


- o ponto que divide a área do gráfico de f em duas partes iguais é $x = 73$



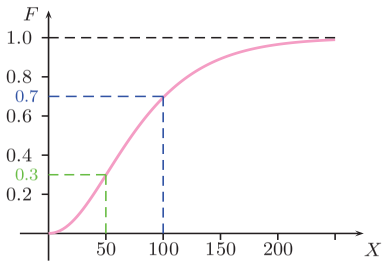
- existe uma probabilidade superior a 0.9 de encontrar na população um indivíduo com um nível de triglicéridos inferior a 200 mg/dL

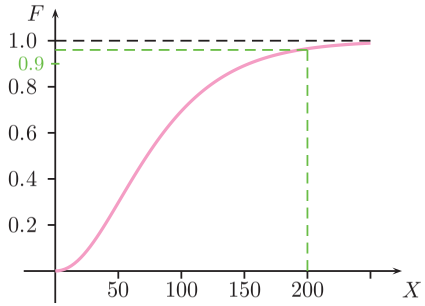
- a probabilidade de encontrar na população um indivíduo com um nível de triglicéridos compreendido entre 50 e 100 mg/dL



pode ser determinada através dos valores da função de distribuição. Tem-se

$$P(50 < X < 100) = F(100) - F(50) = 0.7 - 0.3 = 0.4$$





1.6. Medidas de localização e escala de uma variável aleatória

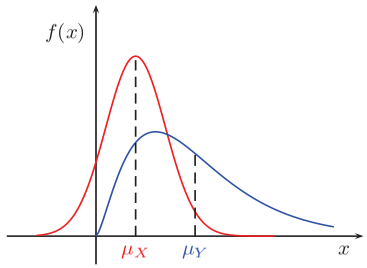
Espaço de Probabilidade $(\Omega = \mathbb{R}, \mathcal{A}, P)$

	X v.a. discreta	X v.a. contínua
Média (ou **Valor Esperado**) $\mu = \mu_X = E(X)$ *Mean* (ou *Expected Value*)	$\mu = \sum_k x_k f(x_k)$	$\mu = \int_{-\infty}^{+\infty} x f(x) dx$
Variância $\sigma^2 = \sigma_X^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$ *Variance*	$\sigma^2 = \sum_k (x_k - \mu)^2 f(x_k)$	$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$
Desvio Padrão *Standard Deviation*	$\sigma = \sqrt{\text{variância}}$	$\sigma = \sqrt{\text{variância}}$
Moda *Mode*	$x \in \mathbb{R} : f(x)$ é máximo (x que ocorre com maior freq.)	$x \in \mathbb{R} : f(x)$ é máximo (x que ocorre com maior freq.)
p-Quantil *p-Quantile*		$q \in \mathbb{R}$ tal que $P(X \leq q) = p$
Mediana $m = \mathrm{med}(X) = q_{0.5}$ *Median*		$m \in \mathbb{R}$ tal que $P(X \leq m) = P(X \geq m) = 0.5$

Nota: O conceito de quantil (e de mediana) pode ser generalizado ao caso discreto mas são os quantis de v.a. contínuas que têm mais interesse. No caso contínuo, a mediana corresponde ao 0.5-quantil.

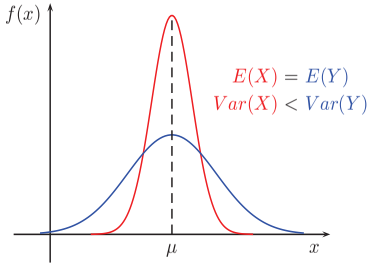
De entre as medidas anteriores:

- a média, a moda e a mediana são medidas de localização (ou tendência central) da distribuição



- a determinação dos p-quantis (e, em particular, da mediana) passa pela determinação da função de distribuição F

- a variância (ou equivalentemente, o desvio-padrão) é uma medida de escala (ou dispersão) dos valores da variável em torno das medidas de localização



Pode-se mostrar que

$$\mathrm{Var}(X) = E\left(X^2\right) - \left(E(X)\right)^2$$

onde, no caso discreto,

$$E\left(X^2\right) = \sum_k x_k^2 \, f(x_k)$$

e, no caso contínuo,

$$E\left(X^2\right) = \int_{-\infty}^{+\infty} x^2 \, f(x) \, dx.$$

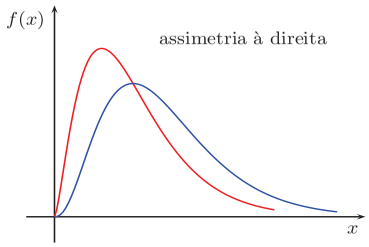
- Existe uma medida - **coeficiente de assimetria** γ_1 - que serve para avaliar a simetria da distribuição (relativamente à sua média). No caso discreto

$$\gamma_1 \quad = \quad \frac{1}{\sigma^3} \sum_k (x_k - \mu)^3 f(x_k)$$

e no caso contínuo

$$\gamma_1 = \frac{1}{\sigma^3} \int_{-\infty}^{+\infty} (x - \mu)^3 f(x) dx.$$

Se a cauda mais longa se situa à direita, a distribuição diz-se **assimétrica à direita** ou **assimétrica positiva**; caso contrário diz-se **assimétrica à esquerda** ou **assimétrica negativa**.



Exemplo 1:
Considere-se novamente a v.a. X que conta o número de pacientes nos quais um determinado fármaco produz o efeito desejado, de entre os primeiros 4 a necessitarem de medicação análoga, e a sua função de probabilidade $f : \{0, 1, 2, 3, 4\} \to \mathbb{R}$ dada por

$$f(0) = 0.008, \quad f(1) = 0.076, \quad f(2) = 0.265,$$
$$f(3) = 0.411, \quad f(4) = 0.240.$$

Em média, o número de pacientes (em 4) nos quais se espera que o fármaco produza efeito é

$$\begin{aligned} E(X) &= \sum_k x_k f(x_k) \\ &= 0(0.008) + 1(0.076) + 2(0.265) \\ &\quad +3(0.411) + 4(0.240) \\ &= 2.80 \end{aligned}$$

Quanto à variância de X, tem-se

$$\begin{aligned} &\mathrm{Var}(X) \\ &= \sum_k (x_k - \mu)^2 f(x_k) \\ &= (0 - 2.80)^2 0.008 + (1 - 2.80)^2 0.076 \\ &\quad +(2 - 2.80)^2 0.265 + (3 - 2.80)^2 0.411 \\ &\quad +(4 - 2.80)^2 0.240 \\ &\approx 0.841. \end{aligned}$$

(Também se podia ter usado a fórmula $\mathrm{Var}(X) = E(X^2) - (E(X))^2$.) Em particular o desvio padrão é

$$\sigma_X = \sqrt{0.841} \approx 0.917.$$

Nota: Para uma amostra *muito grande* de grupos de 4 indivíduos, espera-se que o número médio de indivíduos nos quais o fármaco produz efeito esteja *muito próximo* de 2.8 e a variância *muito próxima* de 0.841 (sendo a aproximação tanto melhor quanto maior for o tamanho da amostra, pois isso implica que a amostra se está a aproximar da população). Veremos mais tarde, no capítulo sobre estatísticas amostrais, os fundamentos teóricos que sustentam esta afirmação.

Exemplo 2:
Considere-se um jogo feito com um dado não viciado em que um jogador ganha 20 euros se sair um 2, 40 euros se sair um 4, perde 30 euros se sair um 6 e não ganha nem perde se aparecer qualquer outra face do dado. Qual é o ganho esperado por um jogador numa jogada?

Seja X a variável aleatória que representa a quantia ganha pelo jogador em qualquer jogada, mais precisamente, seja $X(i)$ o ganho do jogador originado pela saída da face i do dado numa jogada. Então X toma os valores -30, 0, 20 e 40 e a função de probabilidade é

$$f(-30) = \frac{1}{6}, \quad f(0) = \frac{3}{6}, \quad f(20) = \frac{1}{6}, \quad f(40) = \frac{1}{6}.$$

Com estes valores,

$$E(X) = (-30)\left(\frac{1}{6}\right) + 0\left(\frac{3}{6}\right) + 20\left(\frac{1}{6}\right) + 40\left(\frac{1}{6}\right) = 5.$$

Isto significa que numa qualquer jogada um jogador pode esperar ganhar 5 euros.

Observe-se ainda que a dispersão dos valores de X em torno da média é relativamente grande

$$\begin{aligned}\mathrm{Var}(X) &= (-30-5)^2\left(\frac{1}{6}\right)+(0-5)^2\left(\frac{3}{6}\right)\\ &\quad+(20-5)^2\left(\frac{1}{6}\right)+(40-5)^2\left(\frac{1}{6}\right)\\ &= 458.3\end{aligned}$$

Outros cálculos e observações:

- A função de distribuição é $F:\mathbb{R}\to\mathbb{R}$ definida por

$$\begin{array}{rl} x<-30 & F(x)=0\\ -30\le x<0 & F(x)=1/6\\ 0\le x<20 & F(x)=4/6\\ 20\le x<40 & F(x)=5/6\\ 40\le x & F(x)=1.\end{array}$$

- A probabilidade de um indivíduo não perder dinheiro numa qualquer jogada é

  $P(X\ge 0)=1-P(X<0)=1-P(X=-30)=5/6$

- A probabilidade de um indivíduo numa qualquer jogada ter um ganho superior a 10 euros é

  $P(X>10)=1-P(X\le 10)=1-F(10)=1-4/6=2/6$

- A moda de X é 0 (pois f atinge o seu valor máximo em $f(0)$).

Métodos Estatísticos em Saúde - FMUP

Folha de exercícios no. 1:

1. Na jogada de um par de dados não viciados, a soma das faces que ficam voltadas para cima pode ser um dos 11 valores 2, 3, ..., 12.

   (a) Identifique a v.a. X em causa.

   (b) Podemos atribuir a probabilidade 1/11 a cada um dos resultados 2, 3, ..., 12? Em caso negativo, qual deve ser a probabilidade a atribuir a cada resultado?

   (c) Determine a função de probabilidade de X.

   (d) Determine a função de distribuição de X.

   (e) Determine a probabilidade de obter a soma das faces que ficaram voltadas para cima :

   (e.1) inferior a 8

   (e.2) não inferior a 4 e não superior a 11

   (e.3) superior a 6 e inferior a 10.

2. Foi escolhido um grupo de consumidores para ordenar por ordem descrescente três produtos, X, Y e Z, de acordo com a sua capacidade hidratante. Definem-se os acontecimentos:

$$\begin{aligned} A &= \{\text{Produto } X \text{ é preferível a } Y\} \\ B &= \{\text{Produto } X \text{ é ordenado em primeiro}\} \\ C &= \{\text{Produto } X \text{ é ordenado em segundo}\} \\ D &= \{\text{Produto } X \text{ é ordenado em terceiro}\} \end{aligned}$$

   em que uma melhor posição corresponde a uma maior hidratação. Se o grupo de consumidores não for constituído por especialistas e, em consequência, pudermos assumir que atribui ordens aos produtos de uma forma aleatória, será o acontecimento A independente dos acontecimentos B, C e D?

3. Considere a v.a. X que representa o número de filhos do sexo masculino em famílias com 4 filhos. Supondo iguais probabilidades de nascimento para bebés do sexo masculino e feminino:

   (a) mostre que a função de probabilidade de X é $f : \{0, 1, 2, 3, 4\} \to \mathbb{R}$ definida por

$$f(0) = 1/16, \quad f(1) = 1/4, \quad f(2) = 3/8$$
$$f(3) = 1/4, \quad f(4) = 1/16.$$

   (b) determine a função distribuição de X e esboce o seu gráfico.

   (c) calcule a média, a moda e a variância de X.

   (d) determine a probabilidade de uma família com 4 filhos escolhida aleatoriamente ter:

   (d.1) mais de 1 rapaz

   (d.2) menos de 2 rapazes

   (d.3) mais de 3 raparigas.

4. Considere novamente a v.a. X que representa o número de filhos do sexo masculino em famílias com 4 filhos. Supondo agora que a probabilidade de nascer um rapaz é de 0.6 e a probabilidade de nascer uma rapariga é de 0.4, determine a função de probabilidade de X.

5. Suponha que 25% da população portuguesa sofre de hipertensão arterial e que, de entre estes, 75% ingerem bebidas alcoólicas. De entre os que não são hipertensos, 50% ingerem bebidas alcoólicas.

   (a) Qual é a percentagem de pessoas que bebem álcool?

   (b) Qual é a percentagem de pessoas que bebendo álcool sofrem de hipertensão?

   (c) Qual é a probabilidade de um indivíduo da população que bebe álcool ser não hipertenso?

6. Numa determinada população de ratinhos existem 60% de fêmeas das quais 6% têm mais de 200 gr de peso e existem 40% de machos dos quais 2% têm mais de 200 gr de peso. Seleccionou-se um ratinho dessa população e verificou-se que pesava mais de 200 gr. Qual é a probabilidade desse ratinho ser fêmea?

7. Determine a média e a variância do número de pontos que podem aparecer numa jogada de um dado não viciado. Descreva a função distribuição de probabilidade da variável aleatória que considerou.

8. Determine a média e o desvio padrão da soma de pontos obtidos na jogada de um par de dados não viciados.

9. Uma variável aleatória toma os valores 2, 4, 7 e 8 com as probabilidades, respectivamente,

$$\frac{1+3x}{4}, \quad \frac{1-x}{4}, \quad \frac{1+2x}{4}, \quad \frac{1-4x}{4}.$$

   (a) Para que valores de x é que se tem um função densidade de probabilidade?

   (b) Tome $x = 1/8$ e represente graficamente as funções densidade e distribuição.

10. Considere a variável aleatória X com a seguinte função distribuição de probabilidade

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/6 & 0 \leq x < 2 \\ 1/4 & 2 \leq x < 4 \\ 1/2 & 4 \leq x < 6 \\ 1 & x \geq 6 \end{cases}$$

   (a) Classifique a variável aleatória em causa.

   (b) Determine a função de probabilidade de X e esboce o seu gráfico.

   (c) Determine a média, a moda e a variância de X.

   (d*) Calcule $E(-X)$, $E(X^2)$, $E(X-5)$, $\mathrm{Var}(-X)$ e $\mathrm{Var}(X-5)$.
   (sugestão: para determinar $E(Y)$ precisa de conhecer $f_Y(y_i) = P(Y = y_i)$; tente escrever f_Y em função de f_X)

   (e) Calcule $P(X \leq 1)$, $P(X > 5)$, $P(X \geq 1)$, $P(2 \leq X < 6)$, $P(2 < X < 6)$ e $P(0 < X \leq 2)$.

11. Uma variável aleatória X é definida por

$$X = \begin{cases} -2 & \text{com probabilidade } 1/3 \\ 1 & \text{com probabilidade } 1/2 \\ 3 & \text{com probabilidade } 1/6 \end{cases}$$

(a) Determine e esboce as funções densidade e distribuição de probabilidade de X.

(b) Calcule a média, a moda e a variância de X.

(c) Determine $P(X < 4)$, $P(X > -1)$ e $P(X = 0)$.

12. Suponha que a procura diária de aparelhos de medição da tensão arterial numa determinada farmácia do Porto é uma variável aleatória X com a seguinte função de probabilidade:

$$f(x) = \begin{cases} \frac{k2^x}{x!}, & x = 1, 2, 3, 4 \\ 0, & \text{caso contrário} \end{cases}.$$

(a) Classifique a variável aleatória X.

(b) Determine o valor de k.

(c) Determine a média, a variância, o desvio-padrão e a moda de X.

(d) Determine a função distribuição de probabilidade e esboce o seu gráfico.

(e) Qual deverá ser o "stock" mínimo de aparelhos de medição da tensão arterial no início de cada dia para que a procura diária seja satisfeita com uma probabilidade de pelo menos 0.8?

(f) Qual é a probabilidade de que, quando considerados dois dias, a procura de aparelhos seja superior, em cada um deles, a dois aparelhos?

13. O número de comprimidos defeituosos produzidos por hora por uma máquina de um laboratório é uma variável aleatória X com função de probabilidade

$$f(x) = \begin{cases} 4/7, & x = 0 \\ 2/7, & x = 1 \\ 1/7, & x = 2. \end{cases}$$

(a) Classifique a variável aleatória X.

(b) Indique, justificando, o número mais provável de comprimidos defeituosos produzidos pela máquina por hora. Esse número coincide com a média de X? (caso contrário determine também essa média)

(c) Determine a função distribuição de X e esboce o seu gráfico.

(d) Determine a variância e o desvio-padrão de X.

(e) Qual é a probabilidade da máquina fabricar pelo menos um comprimido defeituoso numa hora? E quando muito um comprimido defeituoso?

(f) Suponha que a máquina trabalha 10 horas por dia. Qual é a média e variância da variável aleatória que representa o número de comprimidos defeituosos produzidos pela máquina no final de um dia de trabalho?

14. Suponha que 2 clínicos, A e B, testam todos os pacientes que chegam a uma determinada clínica para a sífilis. Considere os acontecimentos:

A^+: o médico A faz um diagnóstico positivo

B^+: o médico B faz um diagnóstico positivo

Suponha que o médico A e o médico B dignosticam, respectivamente, 10% e 17%, de todos os pacientes como positivos, e ambos os médicos diagnosticam 8% de todos os pacientes como positivos.

(a) Averigúe a independência dos eventos A^+ e B^+.

(b) Suponha que um paciente é aconselhado a realizar outros testes de laboratório sempre que um dos médicos (A ou B) lhe faz um diagnóstico positivo. Qual é a probabilidade de um paciente ser aconselhado a realizar outros testes de laboratório?

15. Considere um estudo sobre a pressão arterial diastólica (PAD) em que se mediram os valores da PAD para os pais (pai e mãe) de todas as famílias de uma determinada região (desde que estes não fossem geneticamente relacionados). Considere os acontecimentos

A: PAD da mãe é ≥ 95

B: PAD do pai é ≥ 95

refectindo casos de hipertensão para a mãe e para o pai, respectivamente. Do estudo efectuado, sabe-se que $P(A) = 0.1$ e $P(B) = 0.2$.

(a) Supondo que A e B são acontecimentos independentes, determine a probabilidade de encontrar, numa das famílias em causa, pelo menos um dos pais hipertenso.

(b) Suponha que a probabilidade de encontrar uma mãe hipertensa numa família em que o pai é hipertenso é 0.25.

(b.1) O que pode afirmar quanto à independência dos acontecimentos A e B?

(b.2) Determine a probabilidade de encontrar uma mãe que não seja hipertensa numa família em que o pai é hipertenso.

(b.3) Determine a probabilidade de encontrar um pai hipertenso numa família em que a mãe não é hipertensa.

16. Suponha que a taxa de diabetes mellitus (DM) tipo II entre indivíduos com idades compreendidas entre os 40 e os 59 anos é de 70% para os Caucasianos, 10% para os Afro-Americanos, 12% para os Hispânicos, e 5% para os Asiático-Americanos. Suponha ainda que a distribuição étnica em Houston, Texas, para os indivíduos dessa faixa etária é 30% de Caucasianos, 25% de Afro-Americanos, 40% de Hispânicos, e 5% de Asiático-Americanos. Qual é a probabilidade de encontrar um habitante de Houston com idade compreendida entre os 40 e os 59 anos que sofra de DM tipo II?

Chapter 2

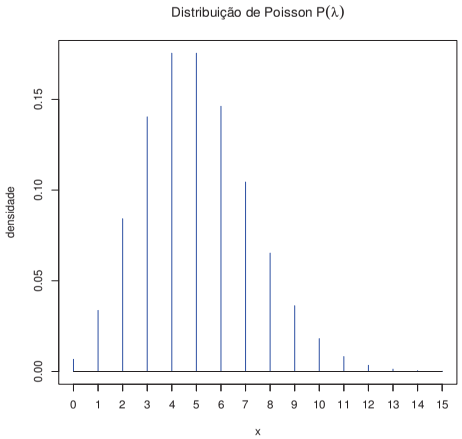
Distribuições de probabilidade especiais

2.1. Distribuições de probabilidade especiais

de uma v.a. **discreta**:

- distribuição uniforme discreta
- distribuição binomial
- distribuição de Poisson
- distribuição multinomial
- distribuição geométrica
- distribuição binomial negativa

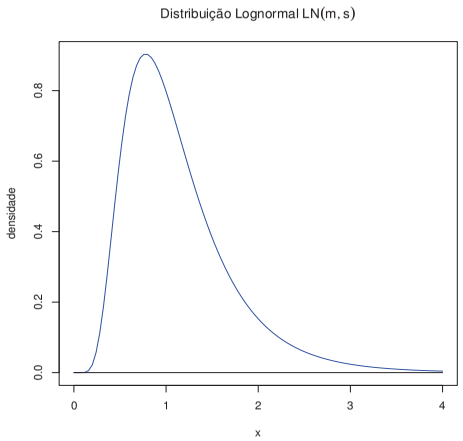
...



de uma v.a. **contínua**:

- distribuição uniforme
- distribuição normal
- distribuição lognormal
- distribuição exponencial
- distribuição χ^2
- distribuição t de Student
- distribuição F de Snedecor

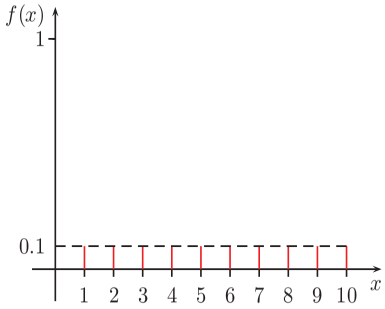
...

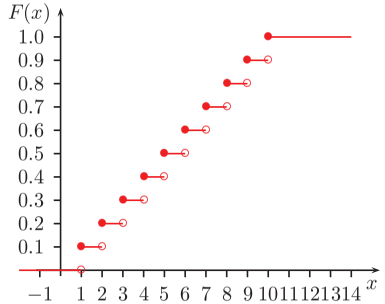


2.1.1. Distribuição Uniforme Discreta: $U(N)$

N = tamanho da população
$f(x_i) = f(x_j), \quad i, j = 1, 2, \ldots, N$, *i.e.*, todos os resultados possíveis da experiência aleatória são equiprováveis.

Função de prob.	$f(x) = \frac{1}{N}$
Função de distrib.	$F(x) = \begin{cases} 0, & x < 1 \\ \frac{[x]}{N}\,^{a}, & 1 \leq x \leq N \\ 1, & x > N \end{cases}$
Domínio	$1, 2, 3, \ldots, N$
Média	$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
Variância	$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} x_i \right)^2$





Exemplos:

1. Considere-se uma população com 10 indivíduos distintos numerados de 1 a 10 e a experiência que consiste da extracção aleatória de um indivíduo dessa população. A variável aleatória X que representa o indivíduo escolhido segue uma distribuição uniforme discreta em 10 pontos. Para $i = 1, 2, \ldots, 10$ tem-se

$$f(i) = P(X = i) = \frac{1}{10}.$$

2. A variável aleatória Y que representa a face voltada para cima resultante de um lançamento de uma moeda não viciada segue uma distribuição discreta uniforme em 2 pontos. Tem-se

$$f(\text{Cara}) = f(\text{Coroa}) = P(Y = \text{Coroa}) = \frac{1}{2}.$$

[a] $[x]$ é a *característica* de x, ie, o menor inteiro que não excede x

2.1.2. Distribuição Binomial $B(n,p)$

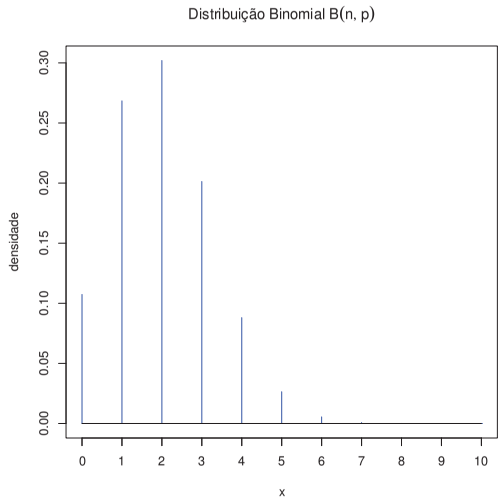
- $n = \#$ ensaios de Bernoulli (experiências independentes e identicamente distribuídas [a], i.i.d., para as quais existem precisamente 2 realizações possíveis - sucesso e insucesso)
- $p =$ probabilidade de sucesso [b]
- $X \sim B(n,p)$ se X representa o número de sucessos em n ensaios

Função de prob.	$b(x) = P(X = x)$ =probabilidade do sucesso acontecer x vezes em n ensaios de Bernoulli $= \binom{n}{x} p^x (1-p)^{n-x}$ $\quad 0 < p < 1$, $x = 0, 1, \ldots, n$
Função de distrib.	$B(x) = P(X \leq x) = \sum_{k=0}^{x} b(k)$
Domínio	$0, 1, 2, 3, \ldots, n$
Média	$\mu = np$
Moda	$p(n+1) - 1 \leq M \leq p(n+1)$
Desvio padrão	$\sigma = \sqrt{np(1-p)}$
Coef. Assimetria	$\lambda = \frac{1-2p}{\sqrt{np(1-p)}}$

[a] **independentes**: os resultados são v.a. independentes; **ident. distribuídas**: a probabilidade de sucesso é a mesma para cada experiência

[b] portanto $1-p$ é a probabilidade de insucesso

A distribuição binomial depende de 2 parâmetros, n e p, designados por **parâmetros da distribuição binomial**. Diferentes valores destes parâmetros correspondem a diferentes distribuições desta família.



Observação: Na expressão de $b(x)$ aparece o termo $\binom{n}{x}$ que representa o número de combinações de x elementos, de entre n, que se podem obter, nomeadamente

$$\binom{n}{x} = C_x^n = \frac{n!}{x!(n-x)!}$$

onde $n! = n.(n-1).(n-2)\ldots 3.2.1.$

Exemplo 1: De um lote de 300 comprimidos em que 9 são defeituosos, são seleccionados aleatoriamente 5 comprimidos. Qual é a probabilidade de obter:

(a) pelo menos um comprimido defeituoso?

(b) no máximo 2 comprimidos defeituosos?

(c) nenhum comprimido defeituoso?

(d) exactamente 3 comprimidos defeituosos?

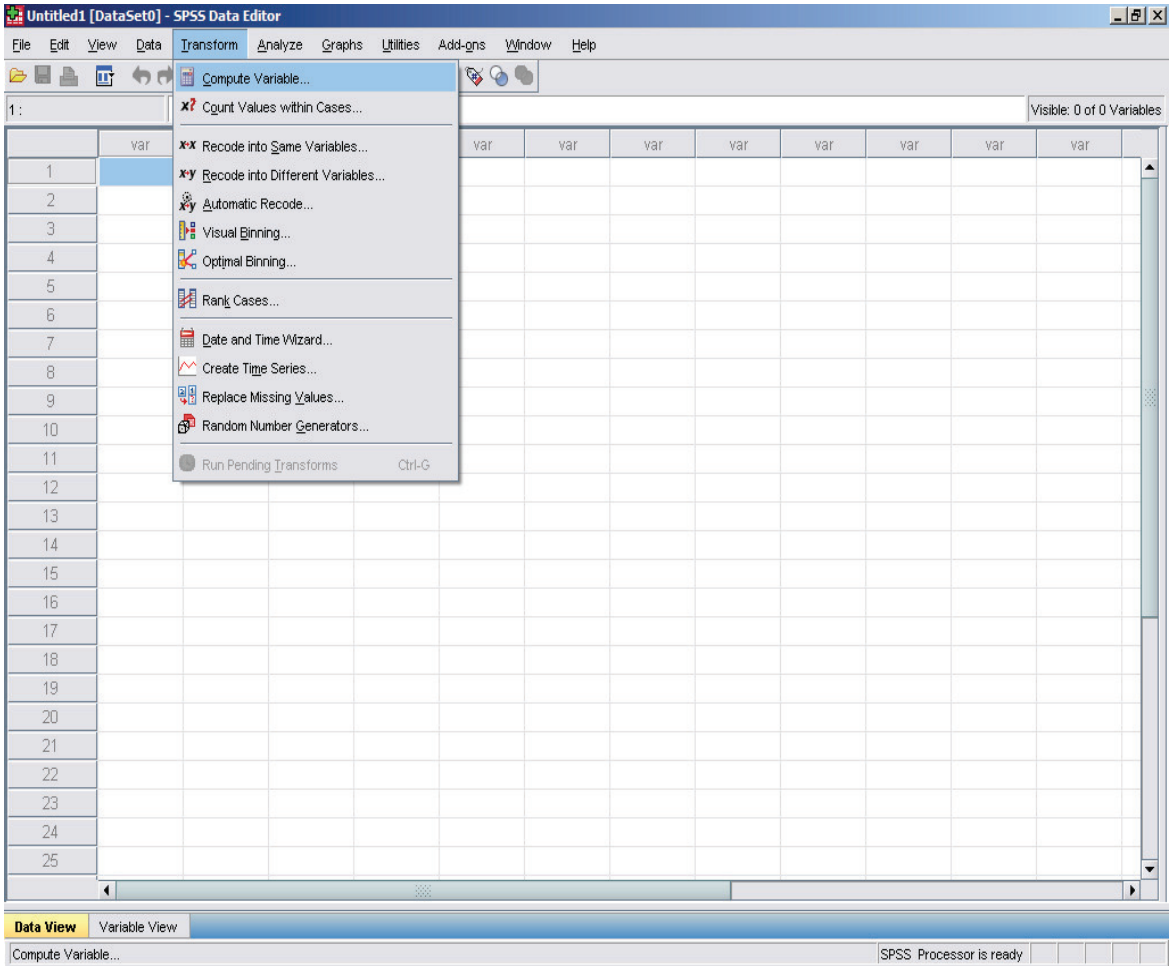
Resolução: Identifiquemos cada ensaio de Bernoulli como sendo a avaliação de cada comprimido do lote de 5 em defeituoso ou normal, e o sucesso desse ensaio como o facto de o comprimido extraído ser defeituoso. Seja X a variável aleatória que representa o número de comprimidos defeituosos, em 5 seleccionados. Como a probabilidade de extrair um comprimido defeituoso do lote é $9/300 = 0.03$ tem-se que

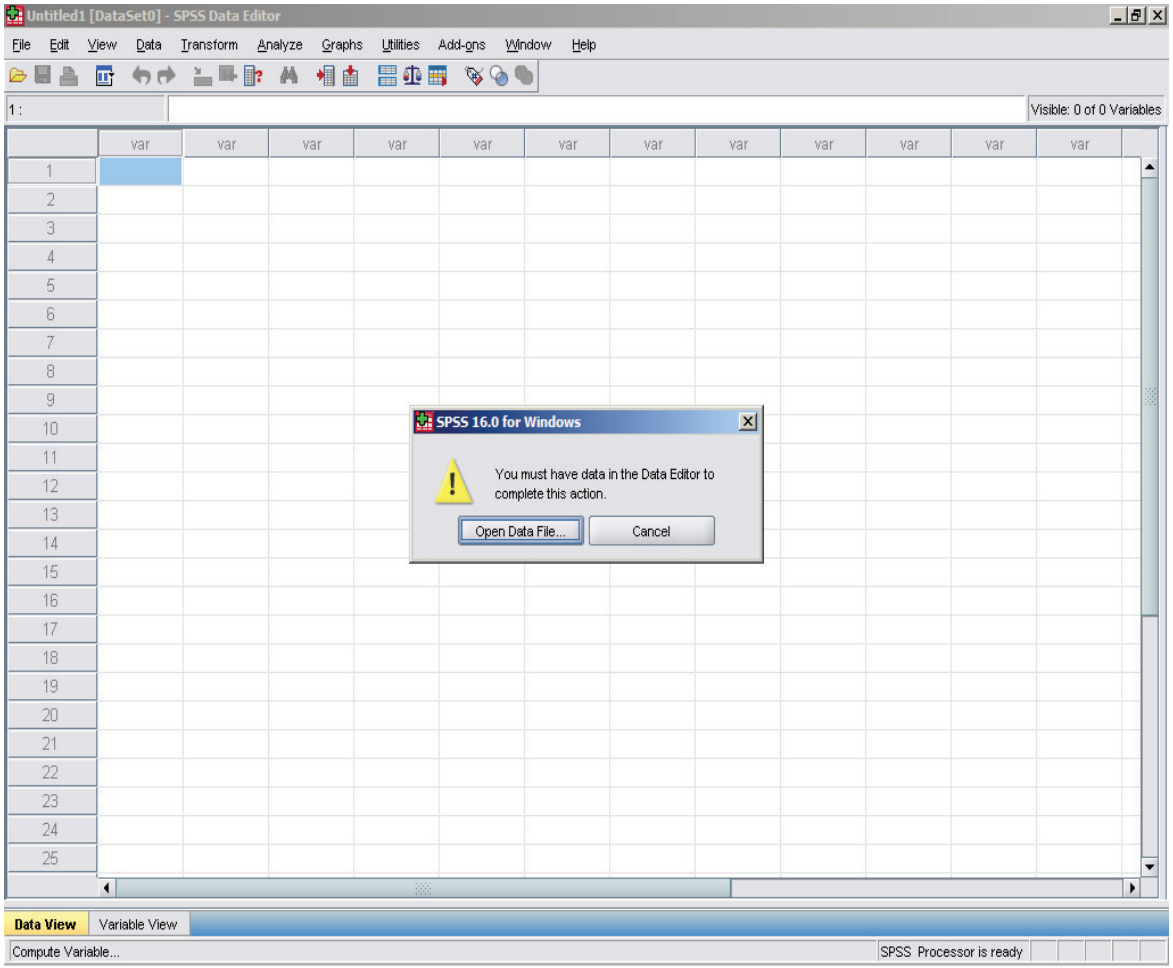
$$X \sim B(5; 0.03)$$

(a) Pretende-se calcular $P(X \geq 1)$. Ora

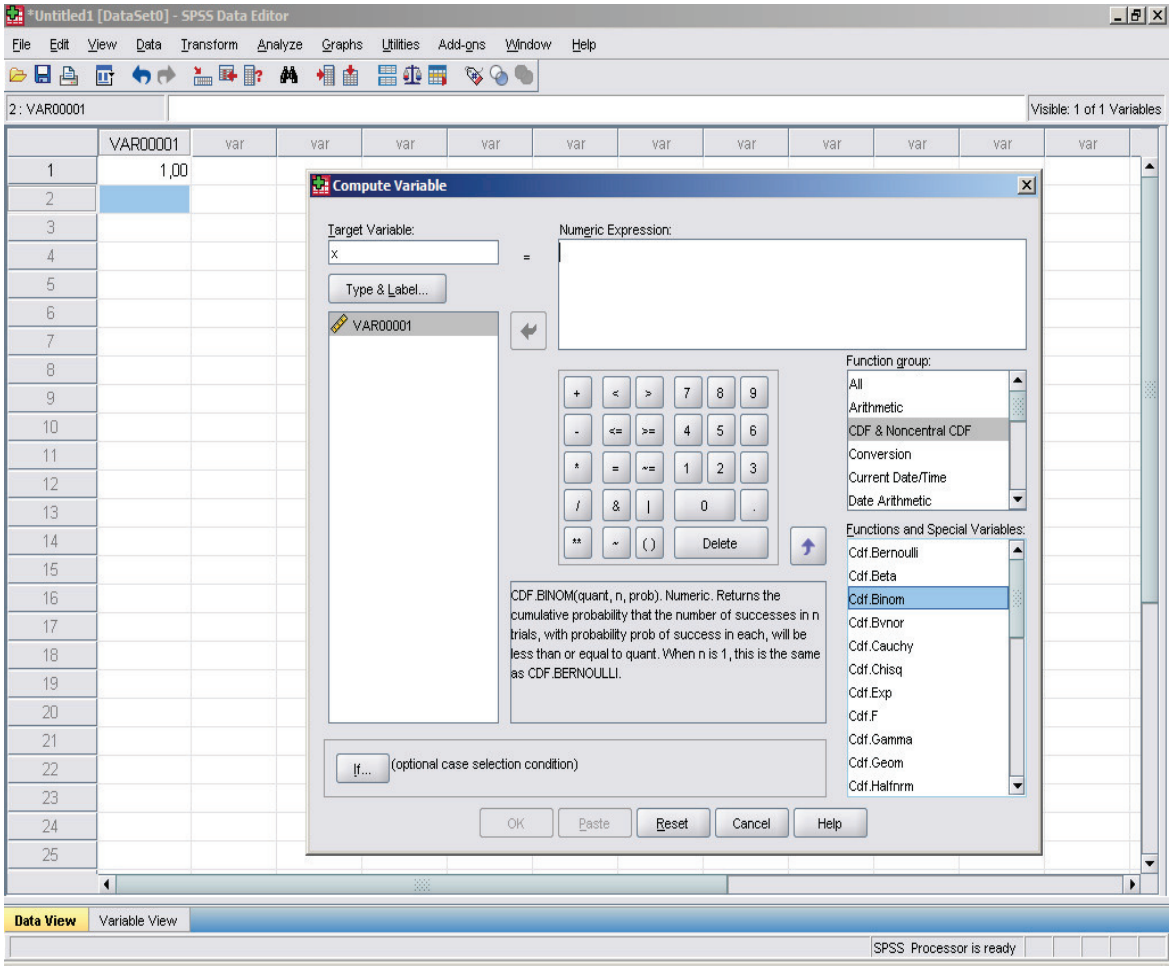
$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) \\ &= 1 - P(X = 0) \\ &= 1 - \binom{5}{0}(0.03)^0(0.97)^5 \\ &\approx 1 - 0.859 \\ &= 0.141. \end{aligned}$$

Para o cálculo de $P(X = 0)$ podemos também usar as instruções em SPSS que se seguem, notando que, nesta situação, $P(X = 0) = P(X \leq 0)$, pelo facto de a variável não tomar valores negativos.

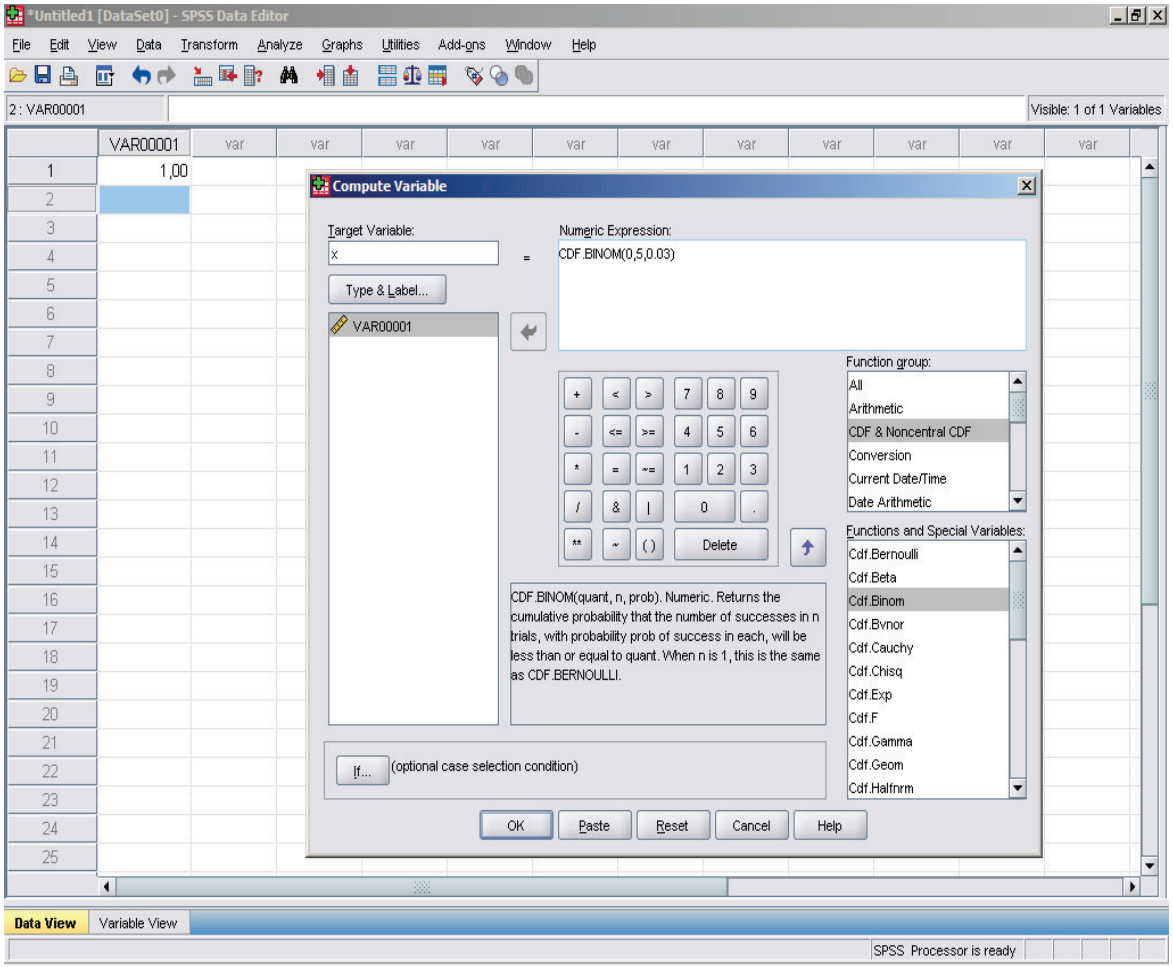

Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Compute Variable...
Count Values within Cases...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Optimal Binning...
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Run Pending Transforms Ctrl-G
Visible: 0 of 0 Variables
Data View
Variable View
Compute Variable...
SPSS Processor is ready



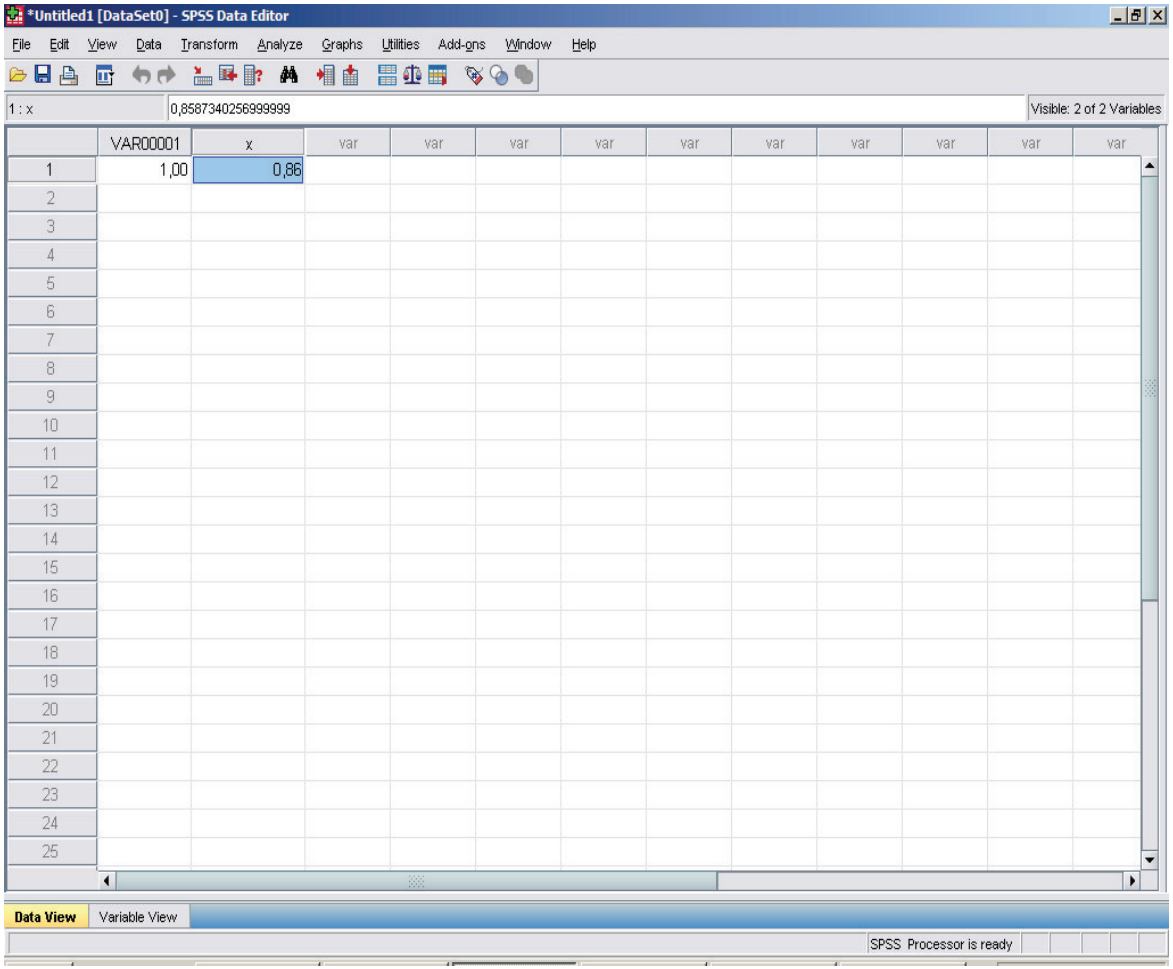
Clicar em "Cancel". É necessário introduzir dados, que para este efeito podem ser arbitrários.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 1 of 1 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Bernoulli
Cdf.Beta
Cdf.Binom
Cdf.Bvnor
Cdf.Cauchy
Cdf.Chisq
Cdf.Exp
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Delete
CDF.BINOM(quant, n, prob). Numeric. Returns the cumulative probability that the number of successes in n trials, with probability prob of success in each, will be less than or equal to quant. When n is 1, this is the same as CDF.BERNOULLI.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

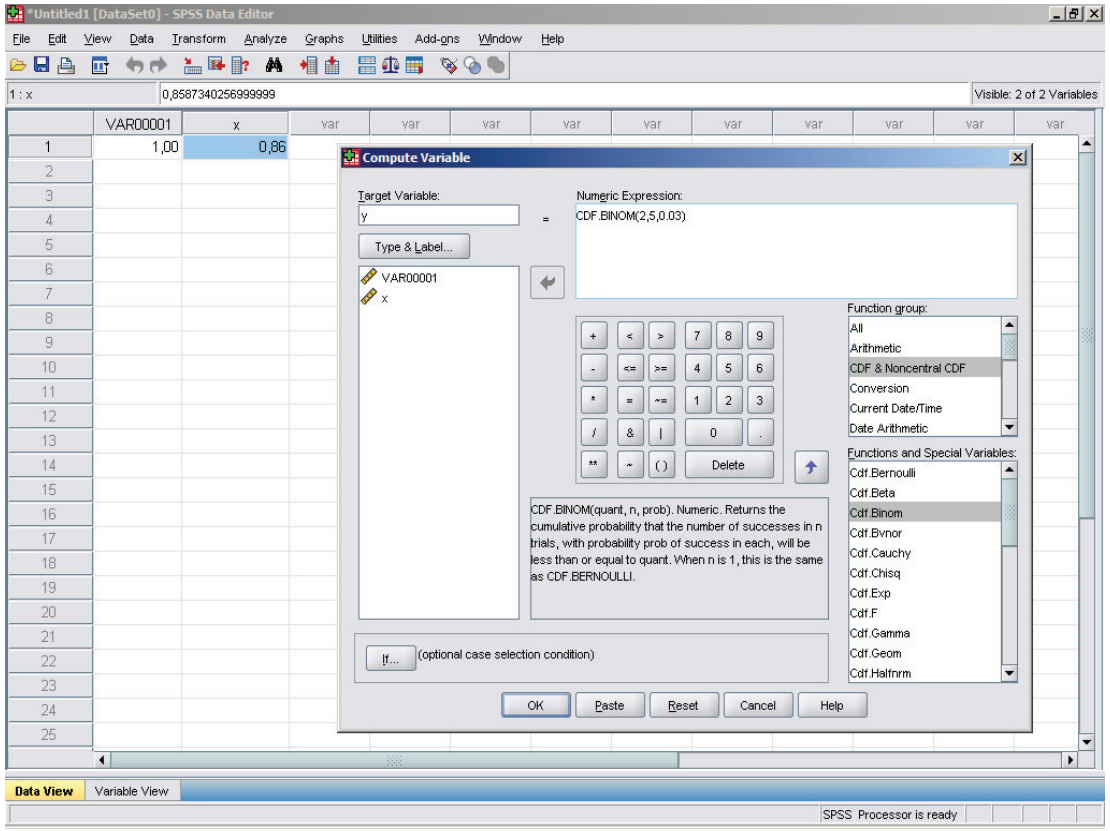
(a)



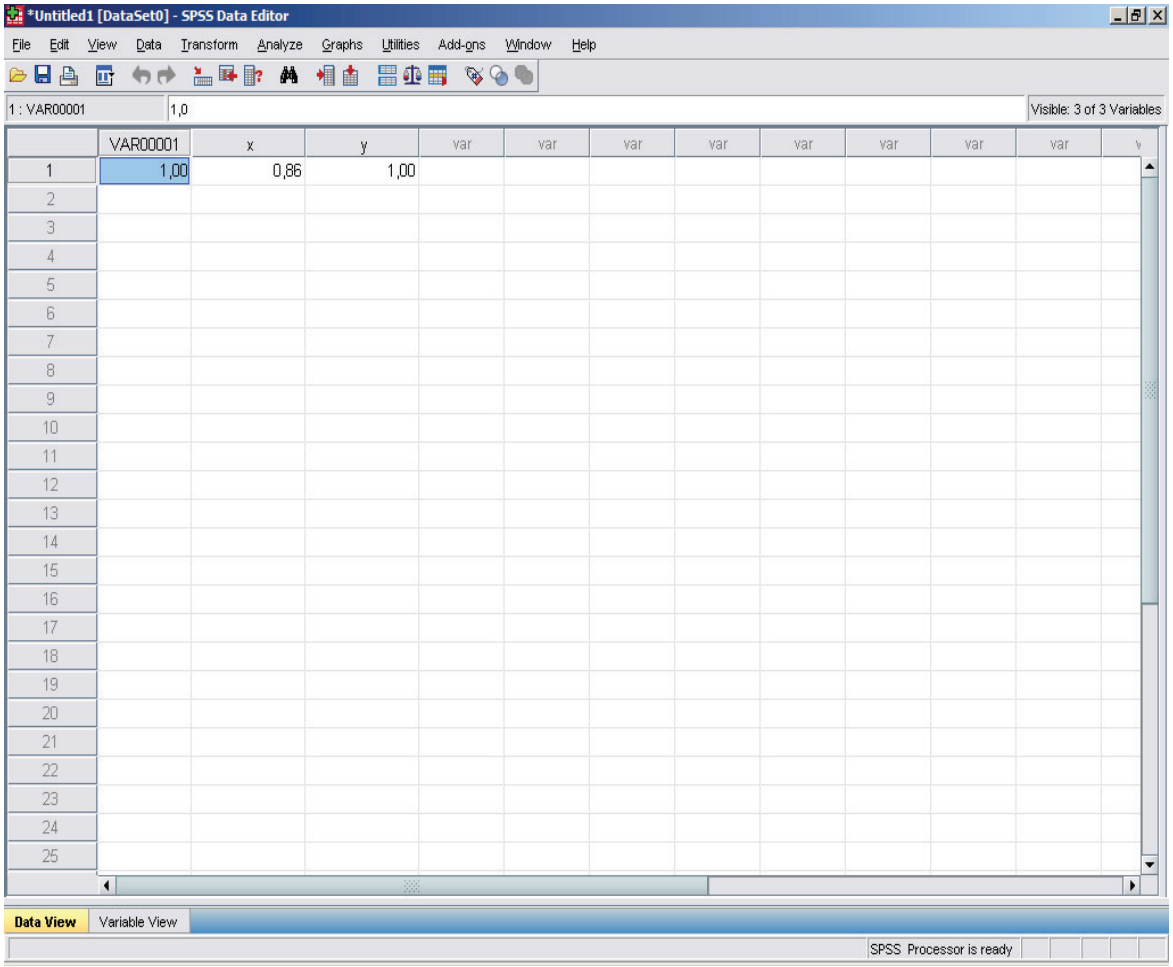
Clicar em "OK". O resultado aparece escrito na coluna correspondente à recém criada variável x.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : x
0,8587340256999999
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,86
Data View
Variable View
SPSS Processor is ready

(b) A probabilidade pretendida é $P(X \leq 2)$ que se obtem com a instrução para a função distribuição de probabilidade da distribuição binomial



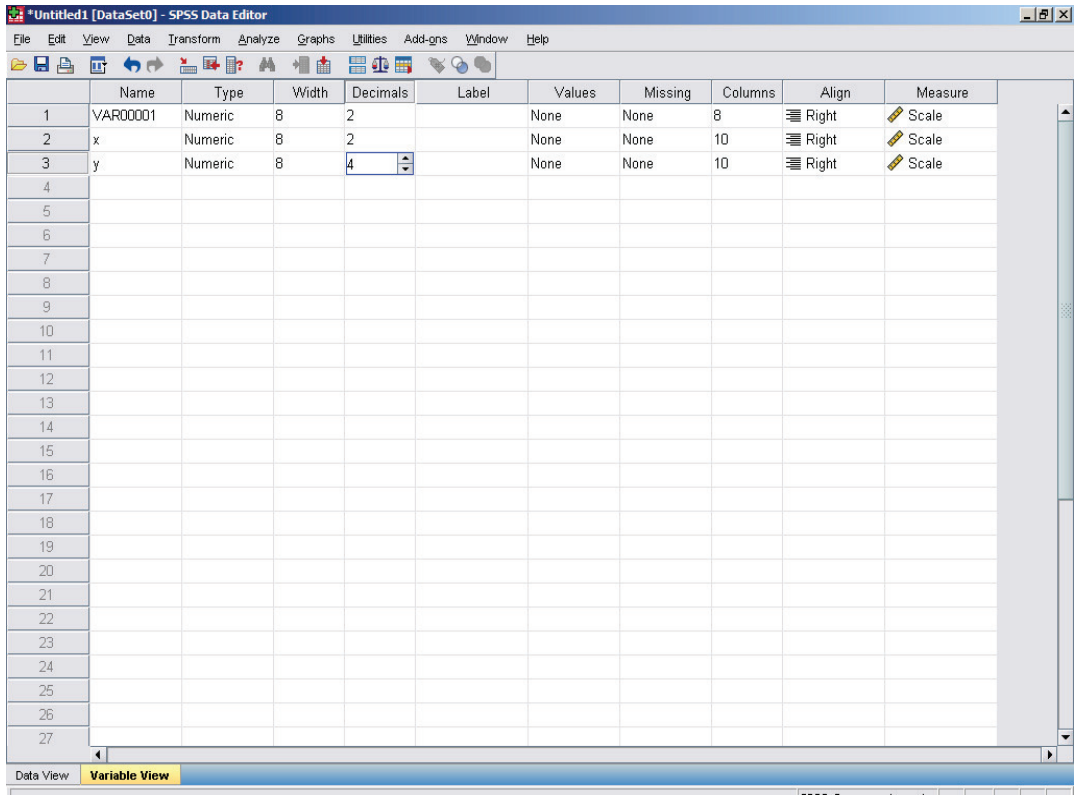
Clicar em "OK". O resultado aparece na coluna correspondente à recém criada variável y.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : VAR00001
1,0
Visible: 3 of 3 Variables
VAR00001 x y var var var var var var var var v
1 1,00 0,86 1,00
Data View
Variable View
SPSS Processor is ready

No caso de se pretenderem mais casas decimais para o resultado anterior:

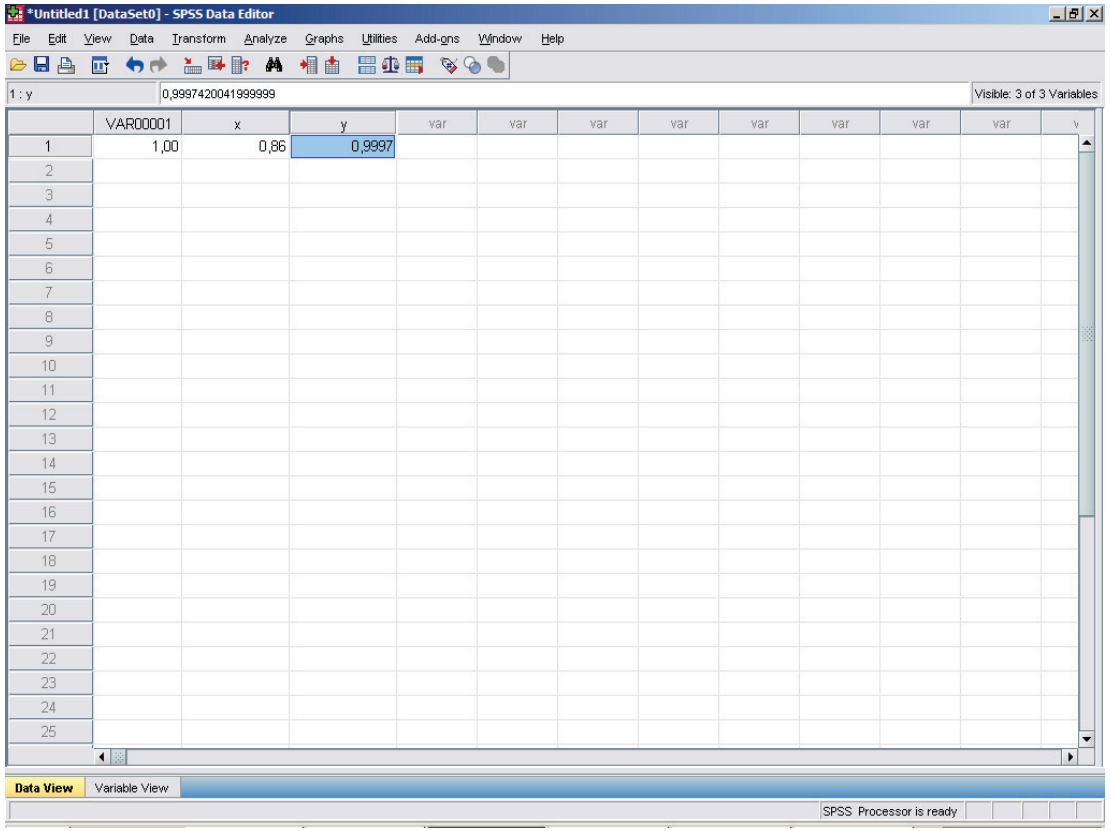
(a) pode-se clicar sobre o valor em causa, na visualização do "Data View", e o valor completo aparece logo abaixo do menu dos comandos principais

(b) pode-se escolher a visualização "Variable View", no canto inferior esquerdo do ecrã, e alterar o número de casas decimais para o valor pretendido (ver abaixo).



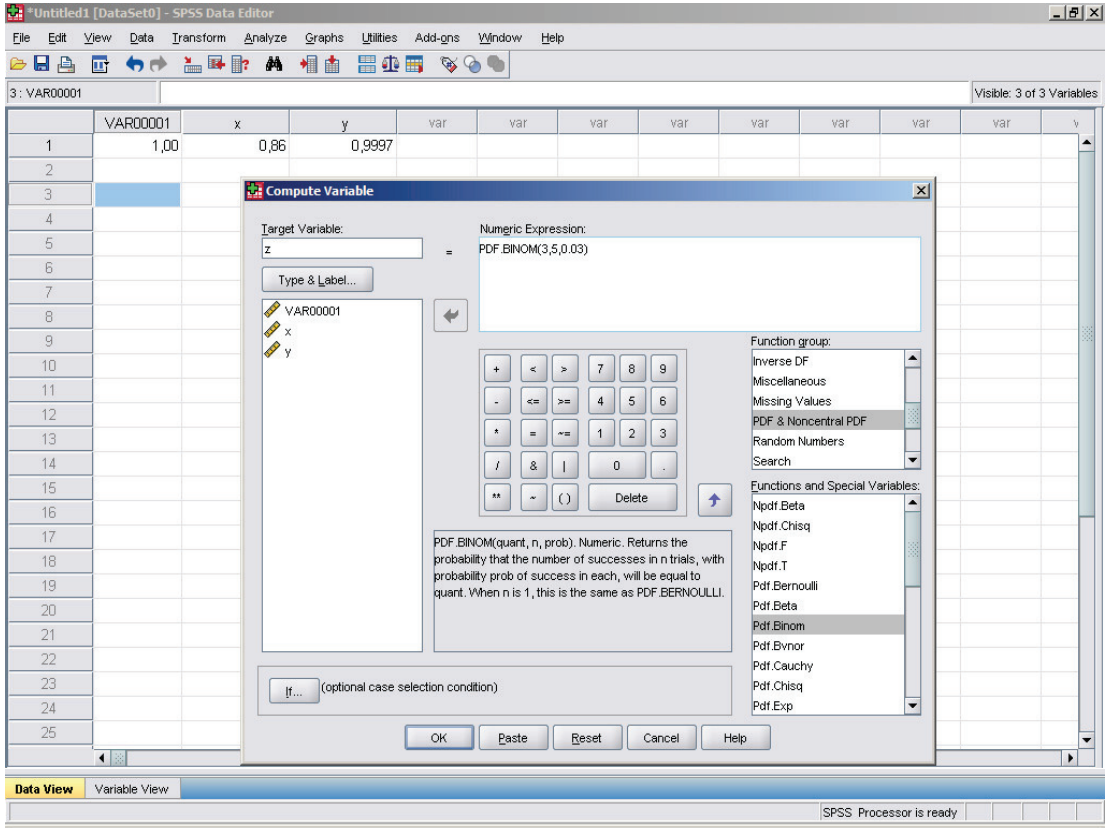
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VAR00001	Numeric	8	2		None	None	8	Right	Scale
2	x	Numeric	8	2		None	None	10	Right	Scale
3	y	Numeric	8	4		None	None	10	Right	Scale

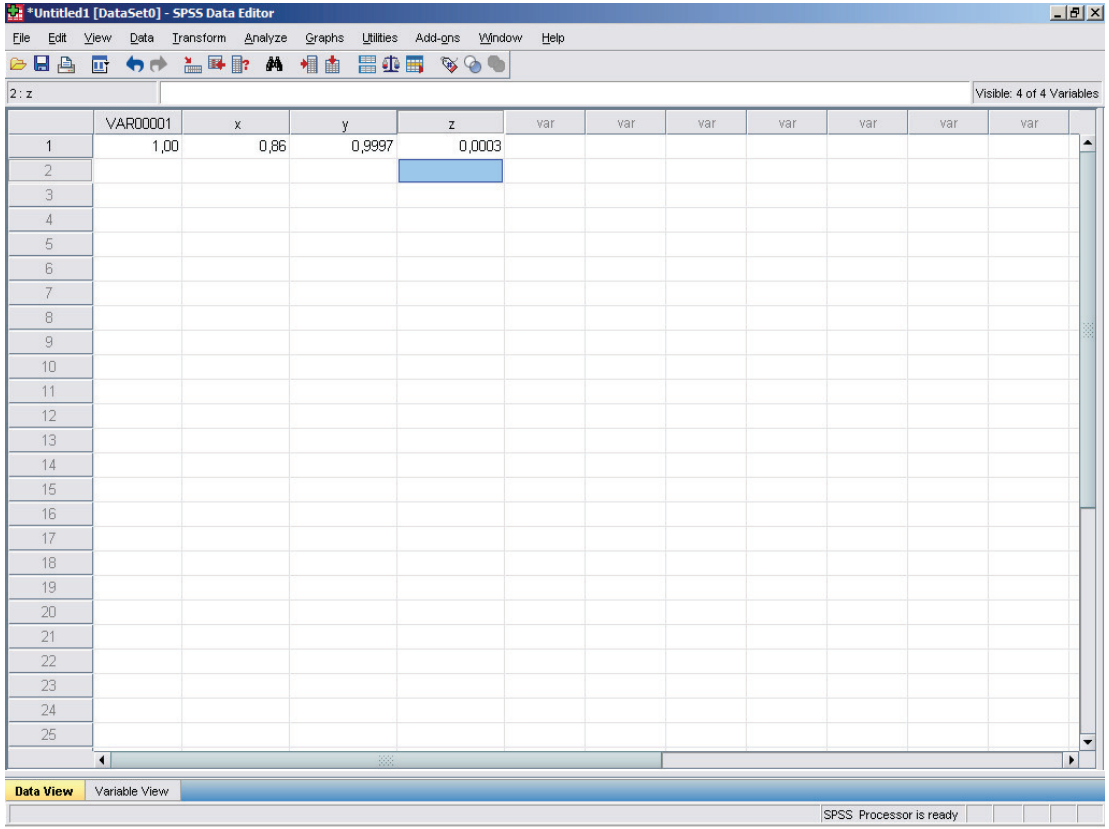
Na situação (b), os resultados são vistos voltando ao esquema inicial de apresentação dos dados, clicando sobre "Data View", que aparece no canto inferior esquerdo do ecrã.



(c) Trata-se de determinar $P(X = 0)$ que já foi calculada na alínea (a) e dá aproximadamente 0.86.

(d) Pretende-se agora $P(X = 3)$, isto é, o valor da função de probabilidade em $X = 3$. Como se observa, a resposta é 0.0003.





Nota:

- CDF = **C***umulative* **D***istribution* **F***unction* = função de distribuição acumulada/função de distribuição
- PDF = **P***robability* **D***istribution* **F***unction* = função densidade de probabilidade/função de probabilidade.

Resolução alternativa: Cada ensaio de Bernoulli continua a ser identificado com a avaliação de cada comprimido do lote de 5 como defeituoso ou normal mas consideramos agora como sucesso desse ensaio o facto de o comprimido extraído ser não defeituoso. Definindo Y como a v.a. que representa o número de comprimidos não defeituosos, em 5 seleccionados, resulta que

$$Y \sim B(5, 0.97).$$

A resposta às questões anteriores seria agora

(a) $P(Y \leq 4)$

(b) $P(Y \geq 3) = 1 - P(Y < 3) = 1 - P(Y \leq 2)$

(c) $P(Y = 5)$

(d) $P(Y = 2)$

que obviamente coincide com os resultados anteriores.

Exemplo 2: Suponha que um grupo de 100 homens com idades compreendidas entre os 60 e os 64 anos recebeu uma nova vacina contra a gripe.

(a) Sabendo que a taxa de mortalidade para homens dessa idade é aproximadamente 0.020, calcule a probabilidade de 10 ou mais homens, desses 100, virem a falecer durante o próximo ano.

(b) Suponha que de entre os 100 indivíduos vacinados se observou que 10 morreram no ano seguinte. Comente os efeitos produzidos pela vacina.

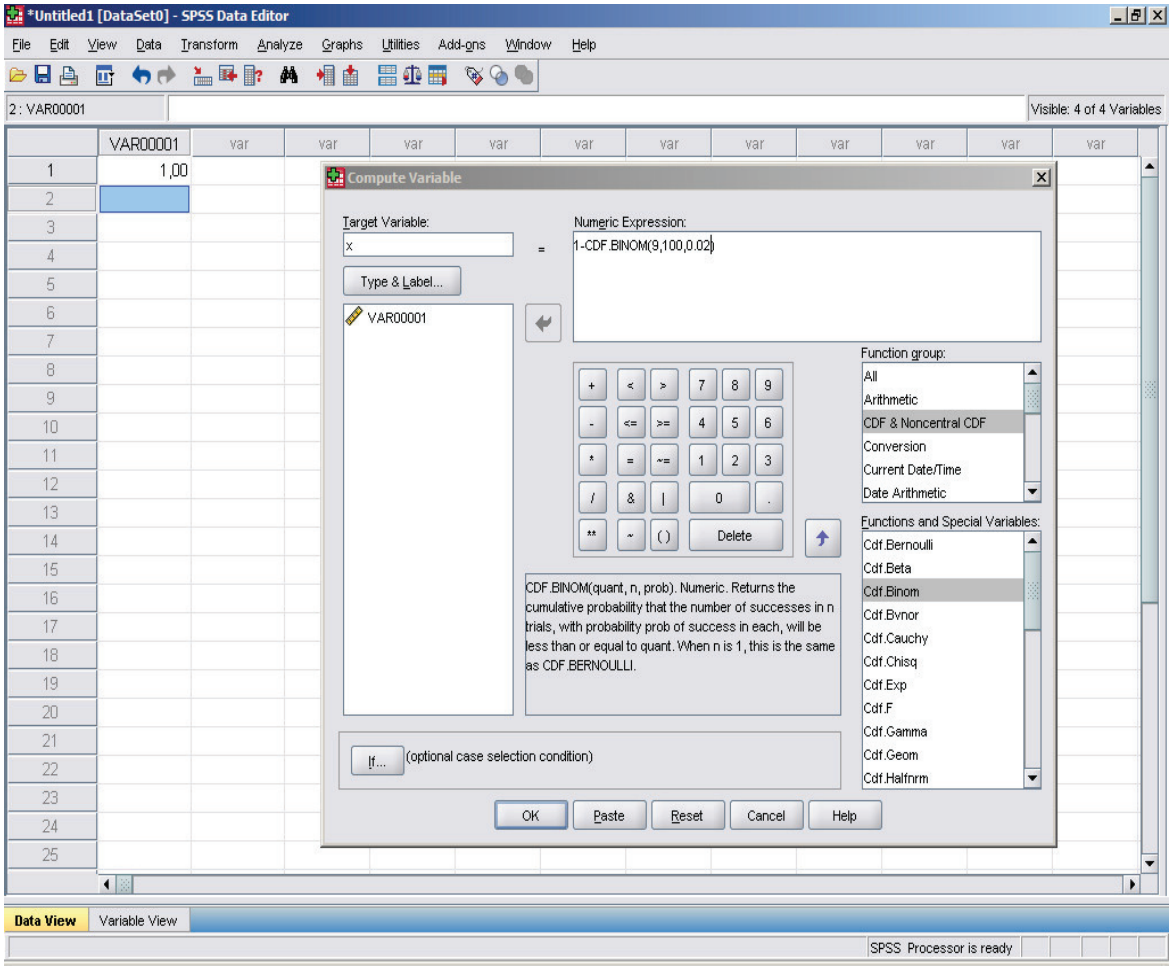
Resolução: Se representarmos por X a v.a. que conta o número de homens com idades compreendidas entre os 60 e os 64 anos, em 100, que morrem no ano seguinte (por qualquer causa), então

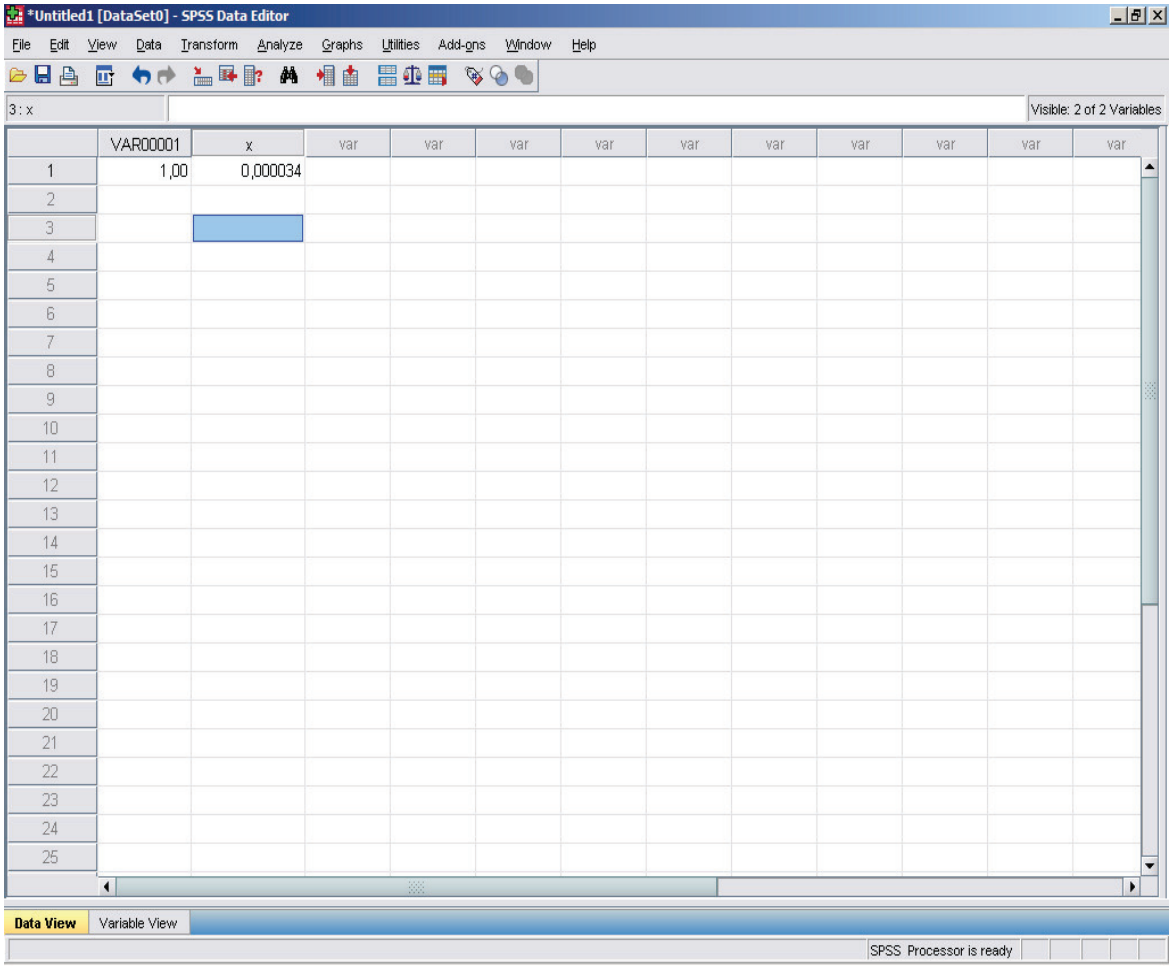
$$X \sim B(100, 0.020).$$

A probabilidade de pelo menos 10 desses homens (em 100) morrerem no ano seguinte é portanto

$$P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9).$$

Pelo que se segue, observamos que esta probabilidade é muito pequena, o que nos leva a sugerir que a administração da vacina deveria ser reavaliada por poder estar a induzir elevadas taxas de mortalidade.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 4 of 4 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
1-CDF.BINOM(9,100,0.02)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Bernoulli
Cdf.Beta
Cdf.Binom
Cdf.Bvnor
Cdf.Cauchy
Cdf.Chisq
Cdf.Exp
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Delete
CDF.BINOM(quant, n, prob). Numeric. Returns the cumulative probability that the number of successes in n trials, with probability prob of success in each, will be less than or equal to quant. When n is 1, this is the same as CDF.BERNOULLI.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

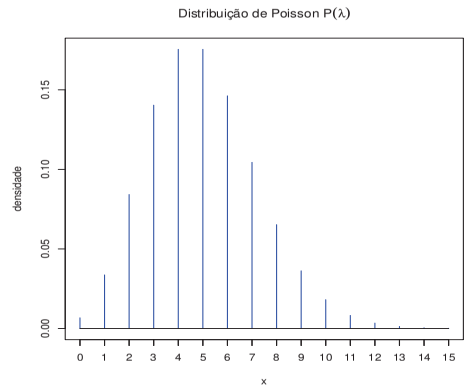

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
3 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,000034
Data View
Variable View
SPSS Processor is ready

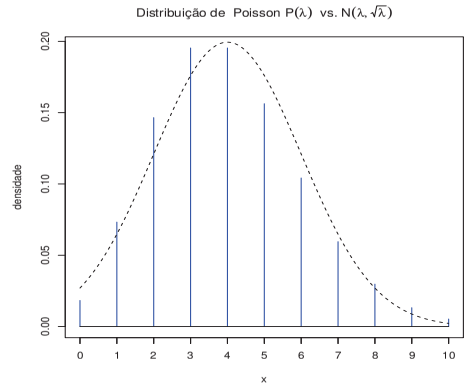
2.1.3. Distribuição de Poisson: $P(\lambda)$

A distribuição de Poisson é normalmente usada para modelar dados de contagens[1]. Geralmente, trata-se de contar o número de ocorrências de um determinado acontecimento, num período de tempo (ou num espaço) especificado, sempre que:

- os eventos ocorram independentemente
- a probabilidade de um evento ocorrer num período de tempo (ou num espaço) muito pequeno seja baixa (*i.e.*, os acontecimentos sejam *raros*)

Função de prob.	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \lambda > 0$
Função de distrib.	$P(x) = \sum_{k=0}^{x} p(k)$
Domínio	$0, 1, 2, 3, \ldots$
Média	$\mu = \lambda$
Moda	maior inteiro $\leq \lambda \quad (= [\lambda])$
Desvio padrão	$\sigma = \sqrt{\lambda}$
Coef. Assimetria	$\lambda = \frac{1}{\sqrt{\lambda}}$
Coef. Achatamento	$\kappa = 3 + \frac{1}{\sqrt{\lambda}}$





[1] *count data*

Observações:

1. Uma v.a. que segue uma distribuição de Poisson tem variância igual à média.

2. É preciso ter algum cuidado com a interpretação de λ em situações reais. Muitas das vezes, λ é descrito como uma taxa, especificada em unidades de "exposição"; por exemplo, λ pode representar o número médio de pessoas que entram numa farmácia para comprar um produto específico, em cada 100 pessoas que entram na farmácia num trimestre. Outras vezes, λ pode corresponder ao número médio dos acontecimentos em causa num determinado período de tempo; por exemplo, λ pode representar o número médio de pessoas que entram numa farmácia para comprar um produto específico, por dia (numa situação em que o período de estudo é, por exemplo, o período natalício).

3. Definindo λ como uma taxa, a escala temporal deve ser sempre mencionada.

4. Assumir uma distribuição de Poisson para uma variável de contagem pode constituir uma hipótese demasiado simplista por causa dos factores que causam muita dispersão. Nesse caso, a **distribuição binomial negativa** será mais adequada pois permite efectuar contagens em situações em que a variância excede a média.

5. Existe um resultado importante sobre a soma de v.a. independentes seguindo, cada uma, uma distribuição de Poisson:

Proposição 2.1.1 *Sejam X_1 e X_2 duas v.a. independentes seguindo distribuições de Poisson, digamos $X_1 \sim P(\lambda_1)$ e $X_2 \sim P(\lambda_2)$. Então*

$$X_1 + X_2 \sim P(\lambda_1 + \lambda_2).$$

Prova: Seja $X = X_1 + X_2$. Se x é um valor que X pode tomar, então podem existir vários $x_1 \in \mathrm{Im}(X_1)$ e $x_2 \in \mathrm{Im}(X_2)$ tais que

$$x = x_1 + x_2.$$

Para $x \in \mathrm{Im}(X)$ seja

$$\mathcal{S}_x = \{(x_1, x_2) \in \mathrm{Im}(X_1) \times \mathrm{Im}(X_2) \mid x = x_1 + x_2\}.$$

Estabelecida esta notação tem-se

$$\begin{aligned}
&P(X = x) \\
&= P(X_1 + X_2 = x) \\
&= \sum_{(x_1,x_2)\in\mathcal{S}_x} P(X_1 + X_2 = x_1 + x_2) \\
&= \sum_{(x_1,x_2)\in\mathcal{S}_x} P(X_1 = x_1)P(X_2 = x_2) \\
&\text{(porque } X_1,\ X_2 \text{ são independentes)} \\
&= \sum_{(x_1,x_2)\in\mathcal{S}_x} e^{-\lambda_1}\frac{\lambda_1^{x_1}}{x_1!}e^{-\lambda_2}\frac{\lambda_2^{x_2}}{x_2!}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{(x_1,x_2)\in\mathcal{S}_x} e^{-(\lambda_1+\lambda_2)} \frac{\lambda_1^{x_1}}{x_1!}\frac{\lambda_2^{x_2}}{x_2!} \\
&= e^{-\lambda} \sum_{(x_1,x_2)\in\mathcal{S}_x} \frac{\lambda_1^{x_1}\lambda_2^{x_2}}{(x_1+x_2)!}\frac{(x_1+x_2)!}{x_1!x_2!} \\
&= \frac{e^{-\lambda}}{x!} \sum_{(x_1,x_2)\in\mathcal{S}_x} \frac{x!}{x_1!x_2!}\lambda_1^{x_1}\lambda_2^{x_2} \\
&= \frac{e^{-\lambda}}{x!} \sum_{x_1} \frac{x!}{x_1!(x-x_1)!}\lambda_1^{x_1}\lambda_2^{x-x_1} \\
&= \frac{e^{-\lambda}}{x!} \sum_{x_1} \binom{x}{x_1}\lambda_1^{x_1}\lambda_2^{x-x_1} \\
&= \frac{e^{-\lambda}}{x!}(\lambda_1+\lambda_2)^{x_1+x_2} \\
&= \frac{e^{-\lambda}}{x!}(\lambda_1+\lambda_2)^{x}
\end{aligned}
$$

portanto

$$X \sim P(\lambda_1+\lambda_2).$$

Exemplo 1: Suponha que o número médio de mortes por febre tifóide numa determinada comunidade num período de um ano foi de 4.6, que a probabilidade de um indivíduo morrer por febre tifóide por dia é muito baixa, e que o número de mortes ocorridas em períodos de tempo disjuntos são v.a. independentes. Qual é a probabilidade de:

(a) mais de 2 indivíduos morrerem por febre tifóide, num ano?

(b) existirem menos de 5 mortes por febre tifóide, num período de 6 meses?

(c) existir mais de uma morte por febre tifóide, num período de 3 meses?

(d) não haver mortes por febre tifóide num período de 1 mês?

Resolução:

(a) Seja X a v.a. que representa o número de mortes por febre tifóide durante um período de 1 ano. Então

$$X \sim P(4.6)$$

e a probabilidade de ocorrerem mais de 2 mortes por febre tifóide num ano é

$$P(X > 2) = 1 - P(X \leq 2).$$

Ora $P(X \leq 2)$ corresponde ao valor da função de distribuição $F_X(2)$ portanto, pelas instruções seguintes em SPSS, tem-se $P(X > 2) \approx 0.837$.

(b) Seja Y a v.a. que representa o número de mortes por febre tifóide num período de 6 meses. Tem-se

$$Y \sim P(2.3)$$

pois 6 meses é metade de 1 ano. O que se pretende agora é

$$P(Y < 5) = P(Y \leq 4)$$

que, como se vê em seguida, é aproximadamente 0.916.

(c) Seja Z a v.a. que representa o número de mortes por febre tifóide num período de 3 meses. Tem-se

$$Z \sim P\left(\frac{4.6}{4}\right) = P(1.15)$$

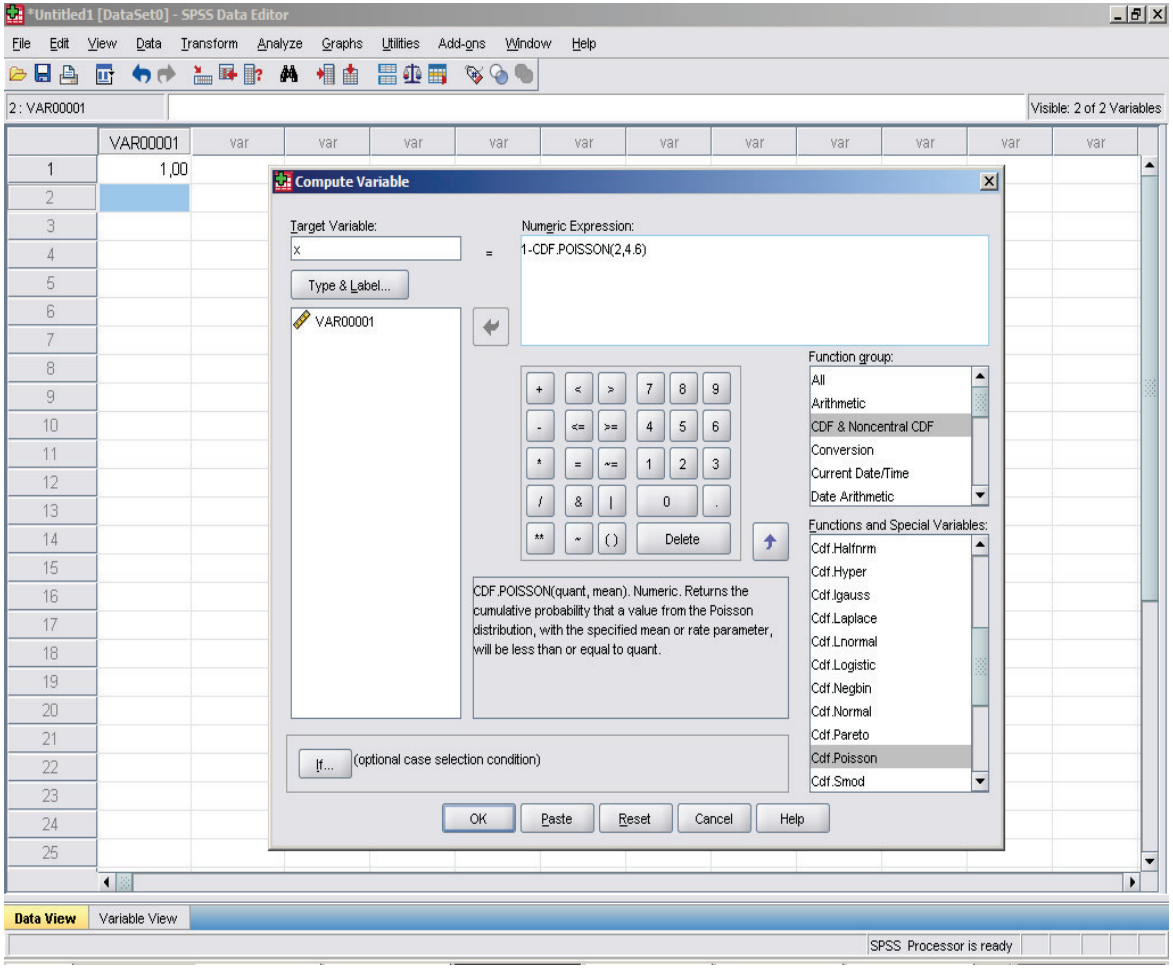
pois 3 meses é um quarto de 1 ano. Pretende-se determinar $P(Z > 1)$, ou seja, $1 - P(Z \leq 1)$, o que é aproximadamente 0.319.

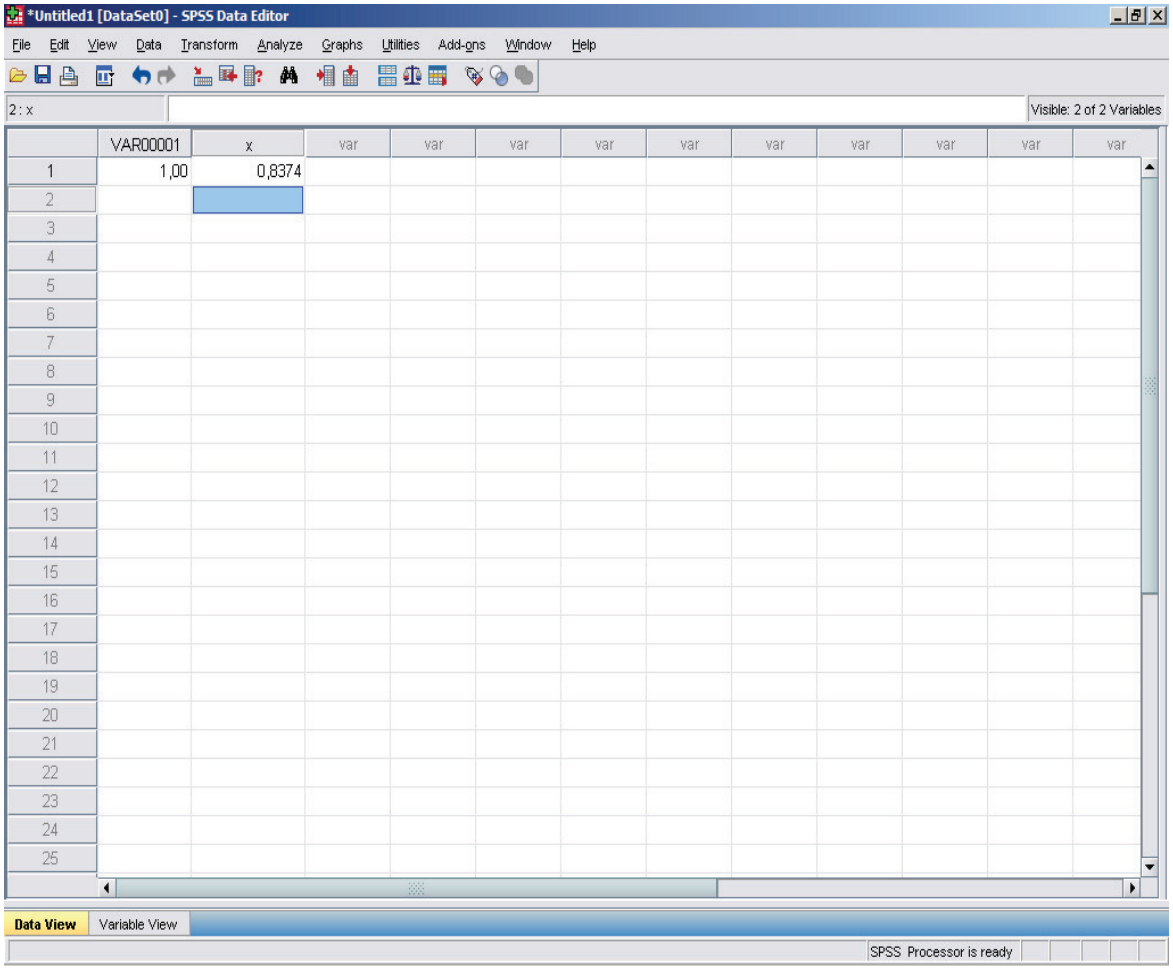
(d) Seja W a v.a. que representa o número de mortes por febre tifóide num período de 1 mês. O número médio de mortes por mês é $\frac{4.6}{12} \approx 0.383$ pelo que

$$W \sim P(0.383).$$

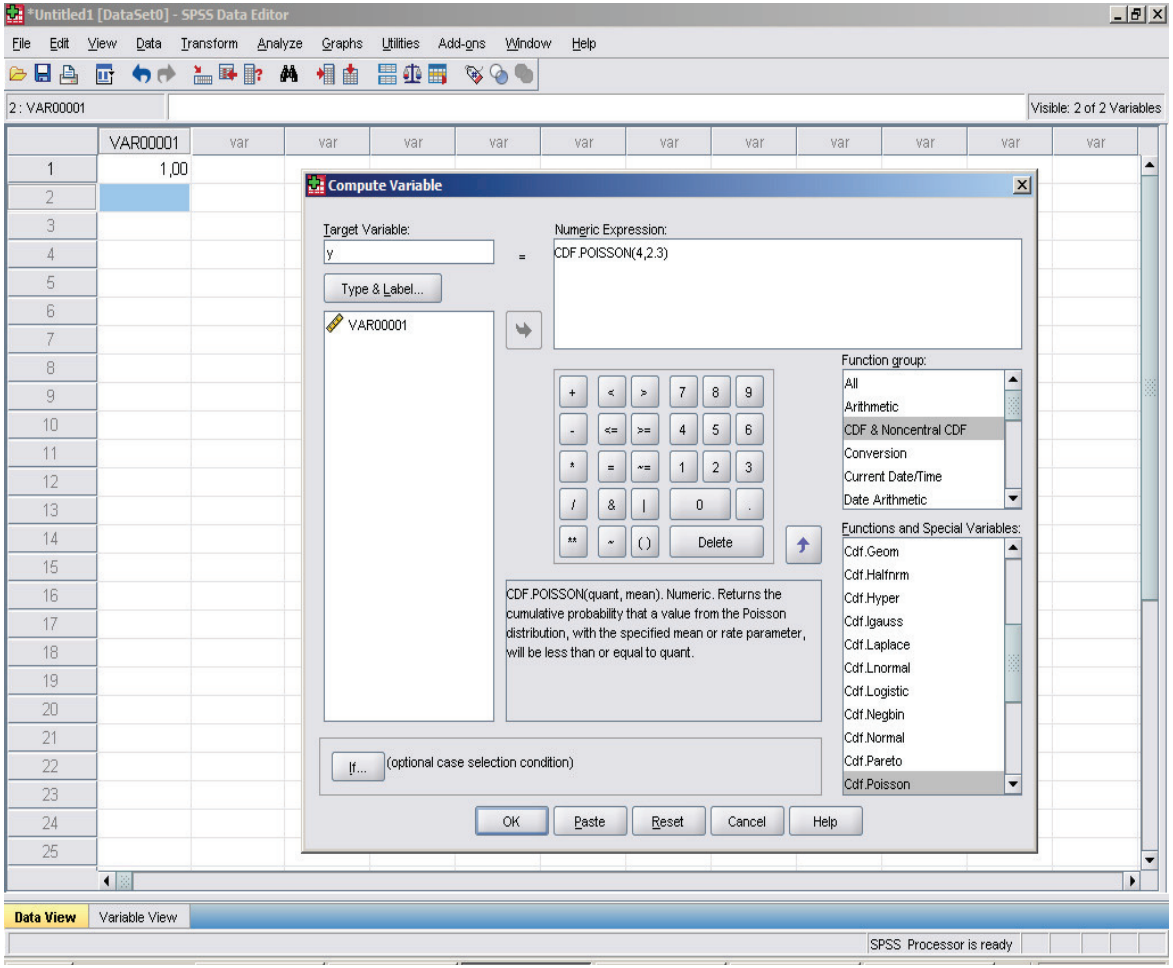
Usando o SPSS, $P(W = 0) \approx 0.682$.

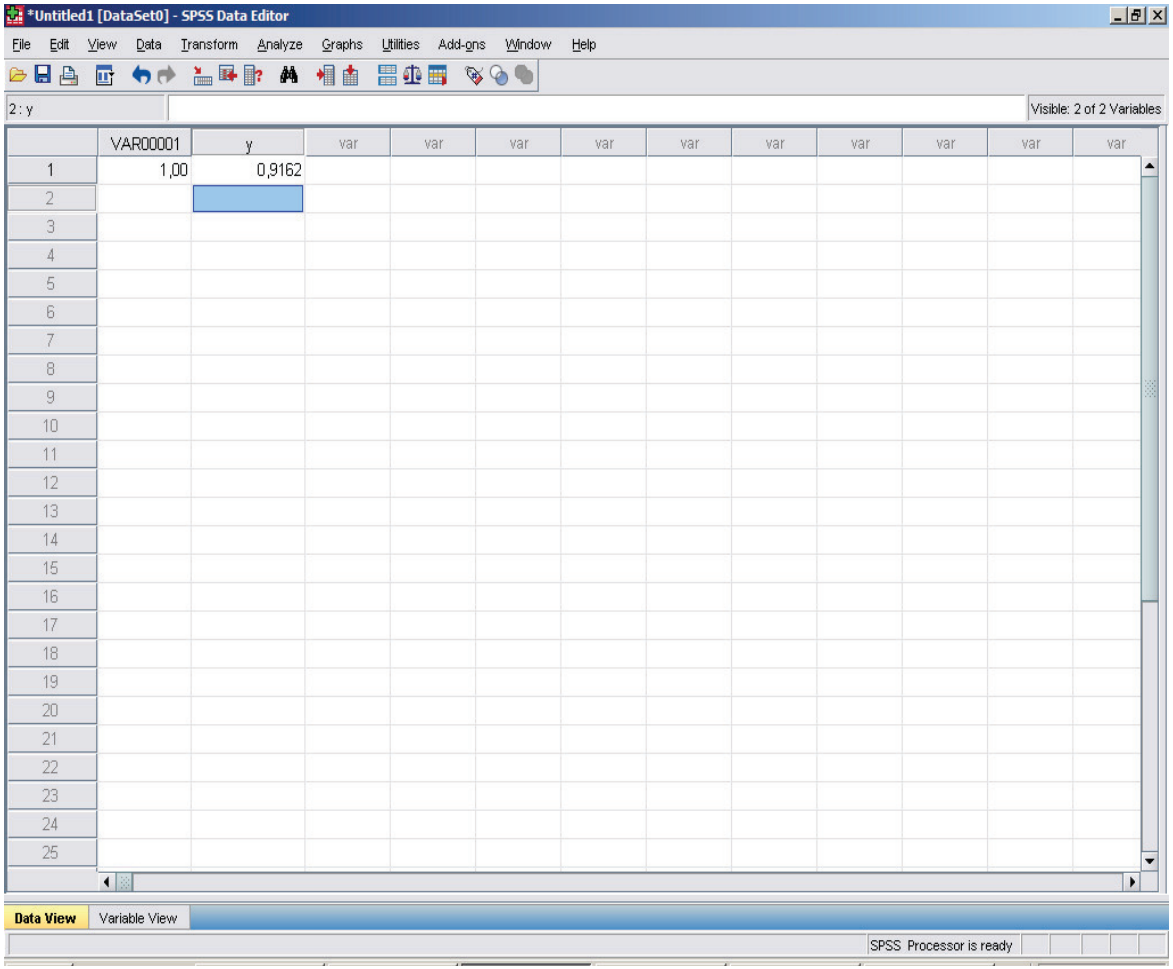
(a)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
1-CDF.POISSON(2,4.6)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Delete
CDF.POISSON(quant, mean). Numeric. Returns the cumulative probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be less than or equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

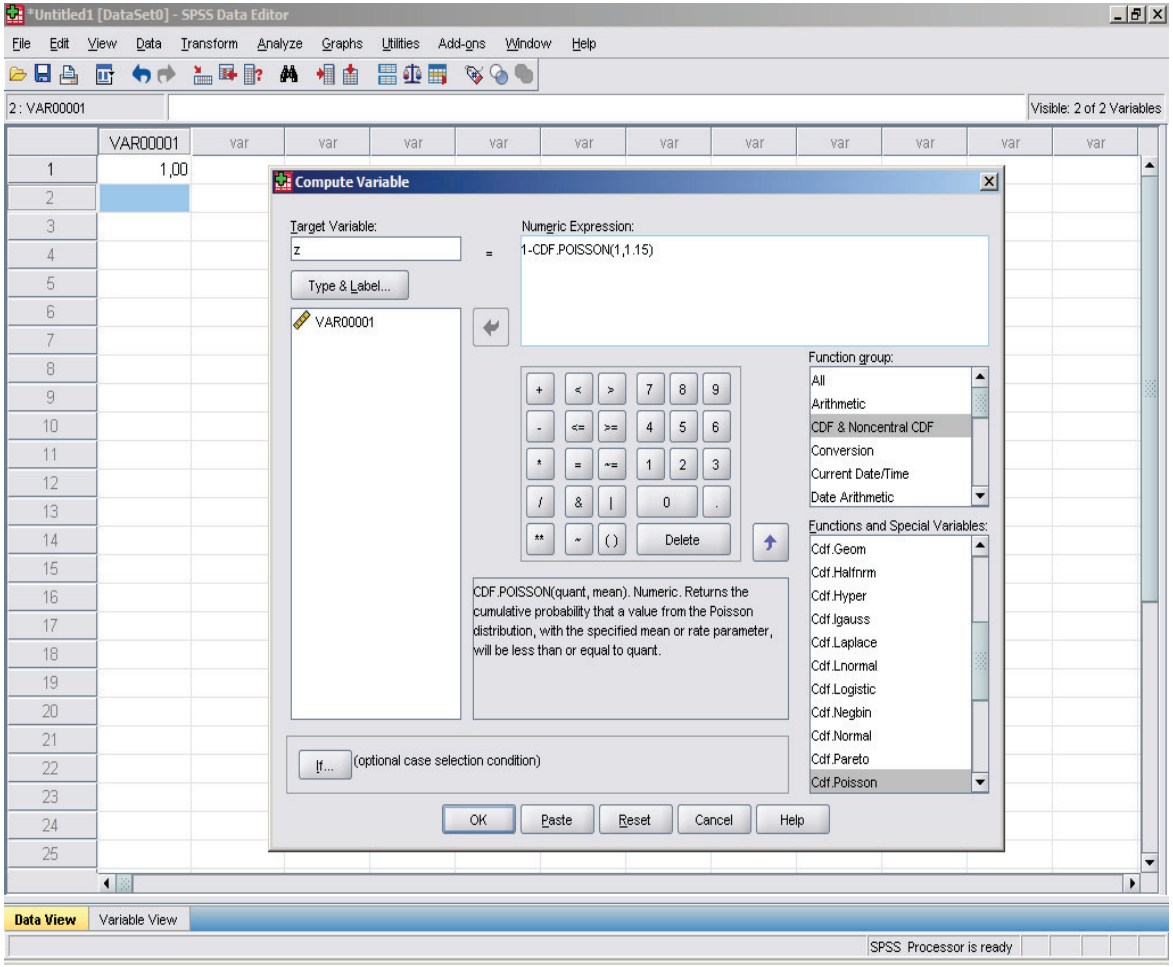

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,8374
Data View
Variable View
SPSS Processor is ready

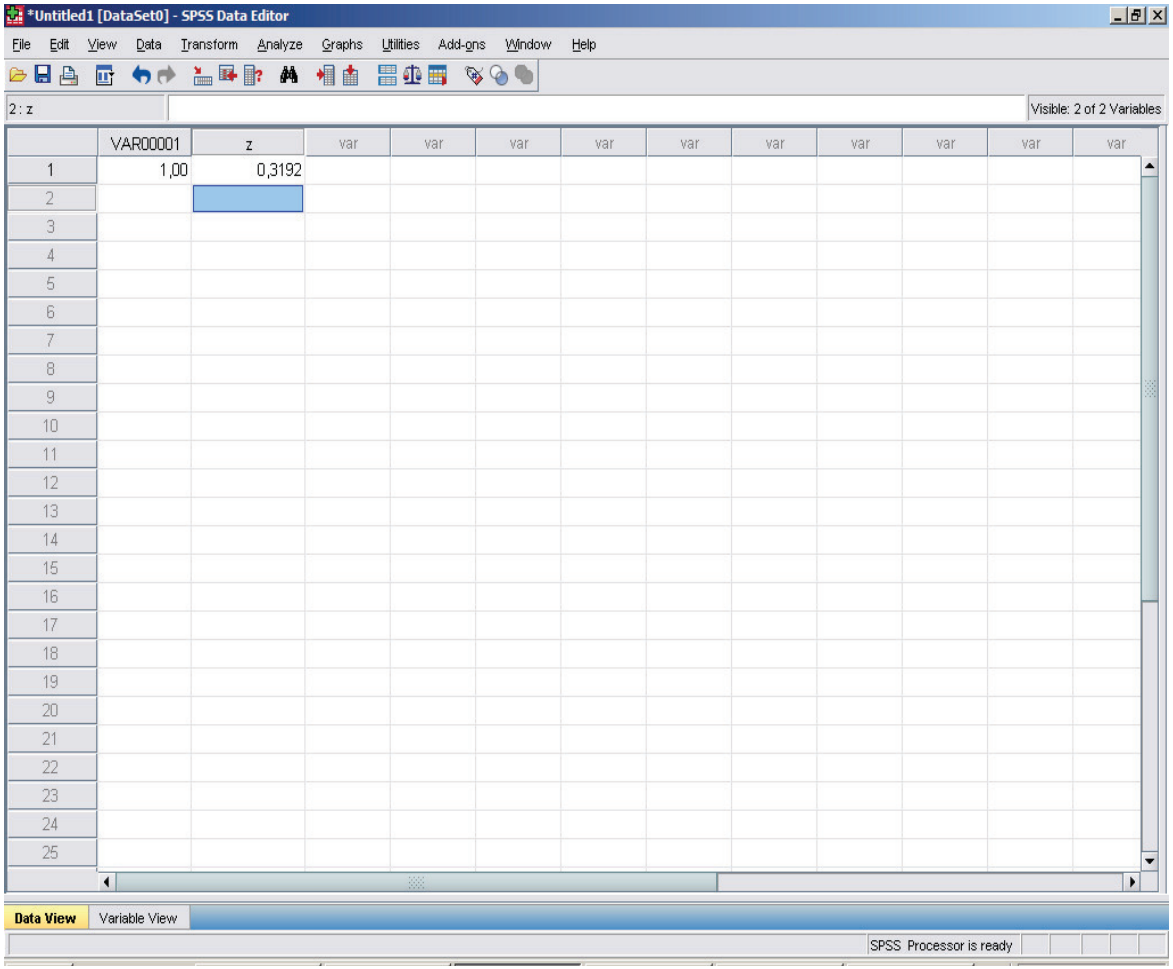
(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
y
Type & Label...
VAR00001
Numeric Expression:
CDF.POISSON(4,2.3)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Delete
CDF.POISSON(quant, mean). Numeric. Returns the cumulative probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be less than or equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

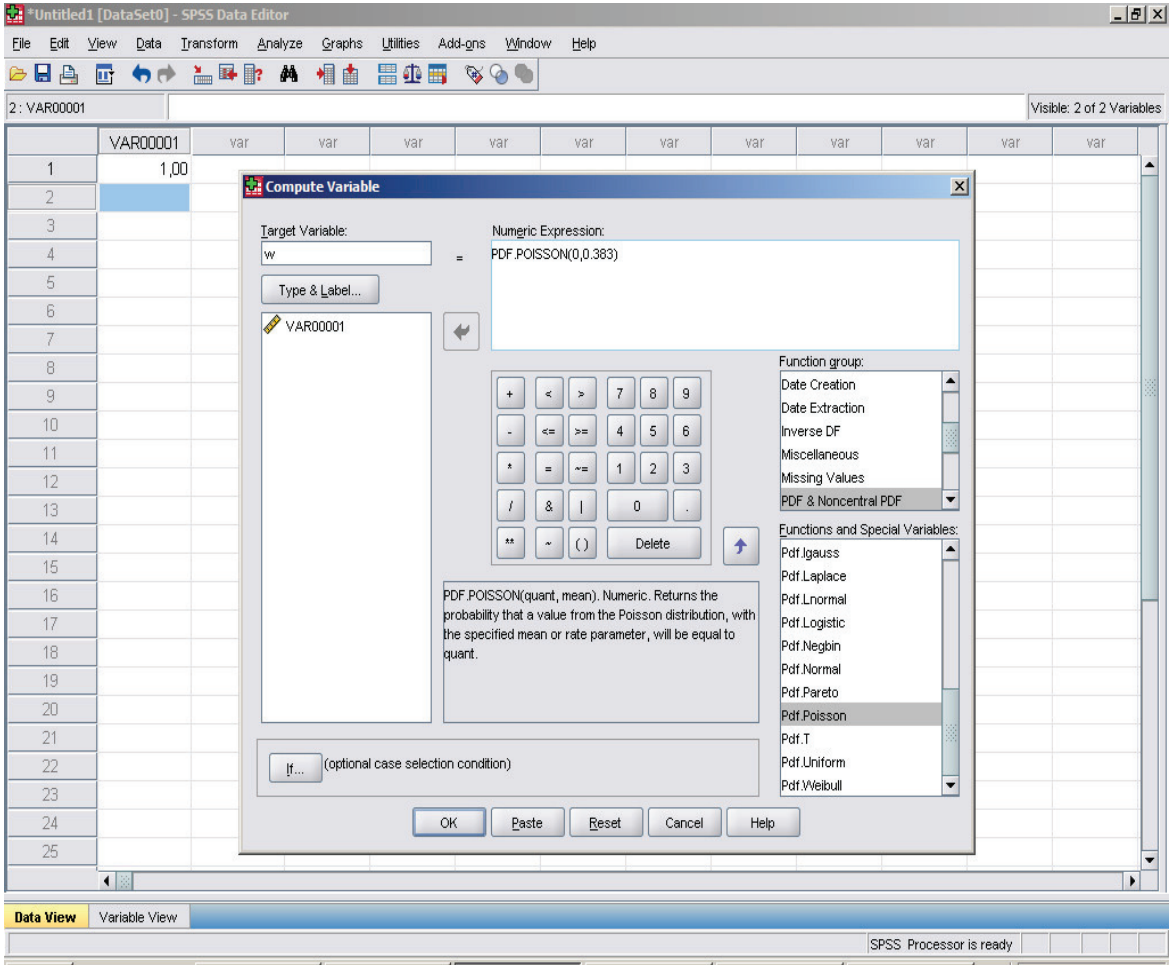

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : y
Visible: 2 of 2 Variables
VAR00001
y
1,00
0,9162
Data View
Variable View
SPSS Processor is ready

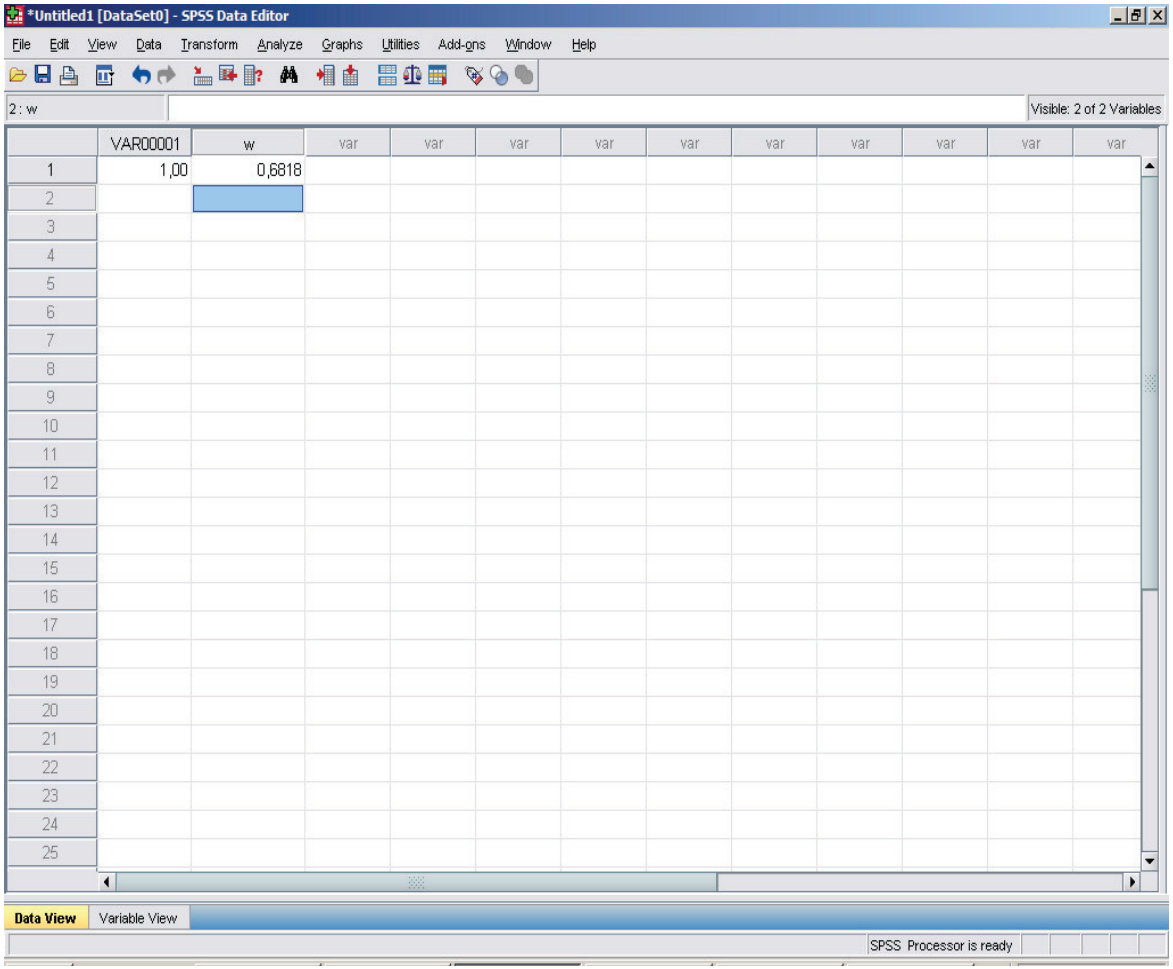
(c)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
z
Type & Label...
VAR00001
Numeric Expression:
1-CDF.POISSON(1,1.15)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Delete
CDF.POISSON(quant, mean). Numeric. Returns the cumulative probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be less than or equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : z
Visible: 2 of 2 Variables
VAR00001
z
1,00
0,3192
Data View
Variable View
SPSS Processor is ready

(d)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
w
Type & Label...
VAR00001
Numeric Expression:
PDF.POISSON(0,0.383)
Function group:
Date Creation
Date Extraction
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Functions and Special Variables:
Pdf.Igauss
Pdf.Laplace
Pdf.Lnormal
Pdf.Logistic
Pdf.Negbin
Pdf.Normal
Pdf.Pareto
Pdf.Poisson
Pdf.T
Pdf.Uniform
Pdf.Weibull
Delete
PDF.POISSON(quant, mean). Numeric. Returns the probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : w
Visible: 2 of 2 Variables
VAR00001
w
1,00
0,6818
Data View
Variable View
SPSS Processor is ready

Exemplo 2: Considere uma placa de agar de 100 cm^2 onde se estão a desenvolver algumas colónias de bactérias. Assuma que a probabilidade de encontrar uma colónia de bactérias num qualquer ponto da placa é muito baixa, e que encontrar colónias de bactérias em pontos diferentes do placa são acontecimentos independentes.
Supondo que se tem uma média de 0.02 colónias por cm^2, determine, para a placa de agar:

(a) a probabilidade de não encontrar nenhuma colónia

(b) a probabilidade de encontrar mais de 3 colónias

(c) a probabilidade de encontrar no máximo 3 colónias

(d) o número de colónias com maior probabilidade de ocorrência.

Resolução: Seja X a v.a. que representa o número de colónias de bactérias na placa de agar. X segue uma distribuição de Poisson $P(\lambda)$ com parâmetro λ (igual à média) dado por $\lambda = 100 \times 0.02 = 2$.

(a) A probabilidade $P(X = 0)$ corresponde ao valor da função de probabilidade em $X = 0$. Pelo SPSS, esse valor é de, aproximadamente, 0.135.

(b) Pretende-se determinar $P(X > 3) = 1 - P(X \leq 3)$. Obtem-se $1 - 0.857 = 0.143$.

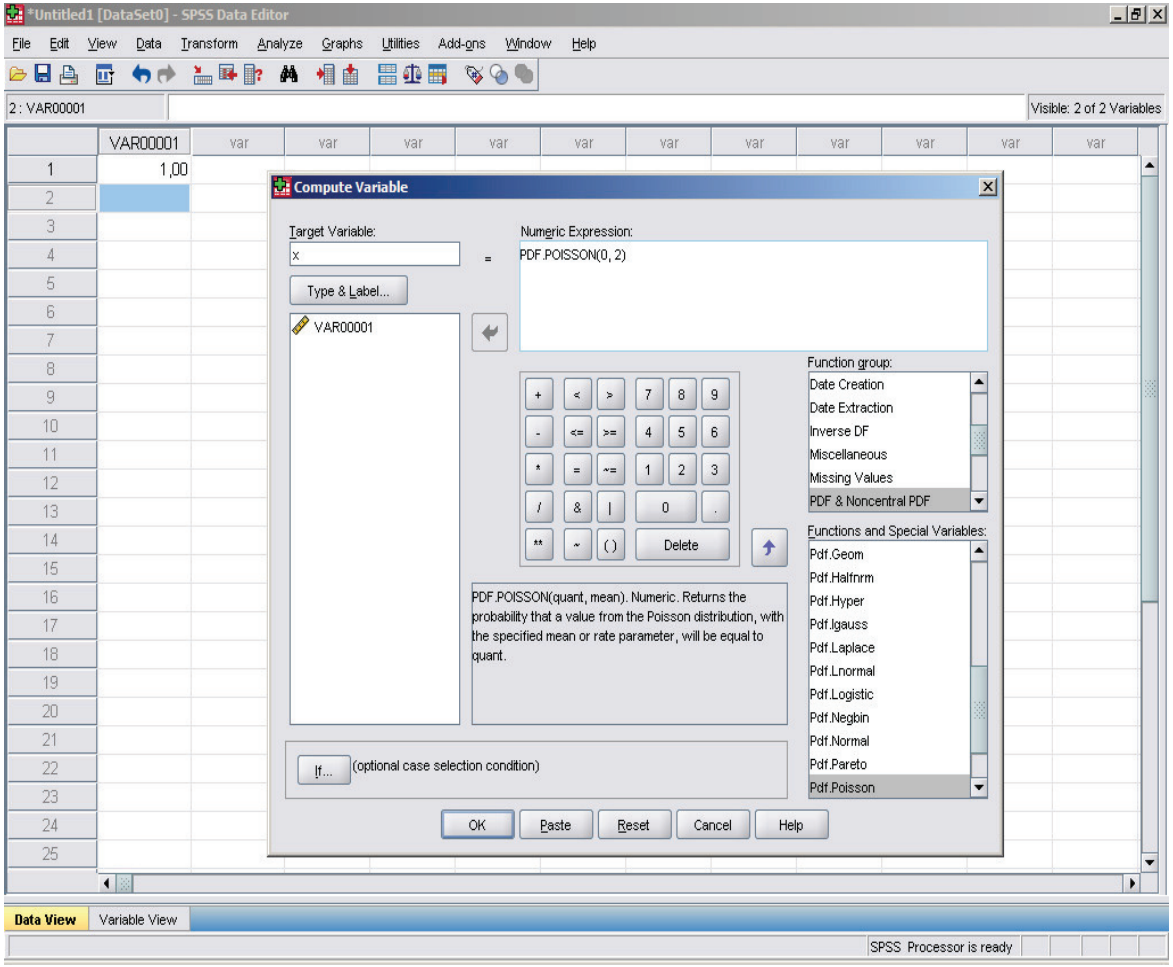
(c) A probabilidade $P(X \leq 3)$ já foi calculada na alínea anterior e corresponde a 0.857.

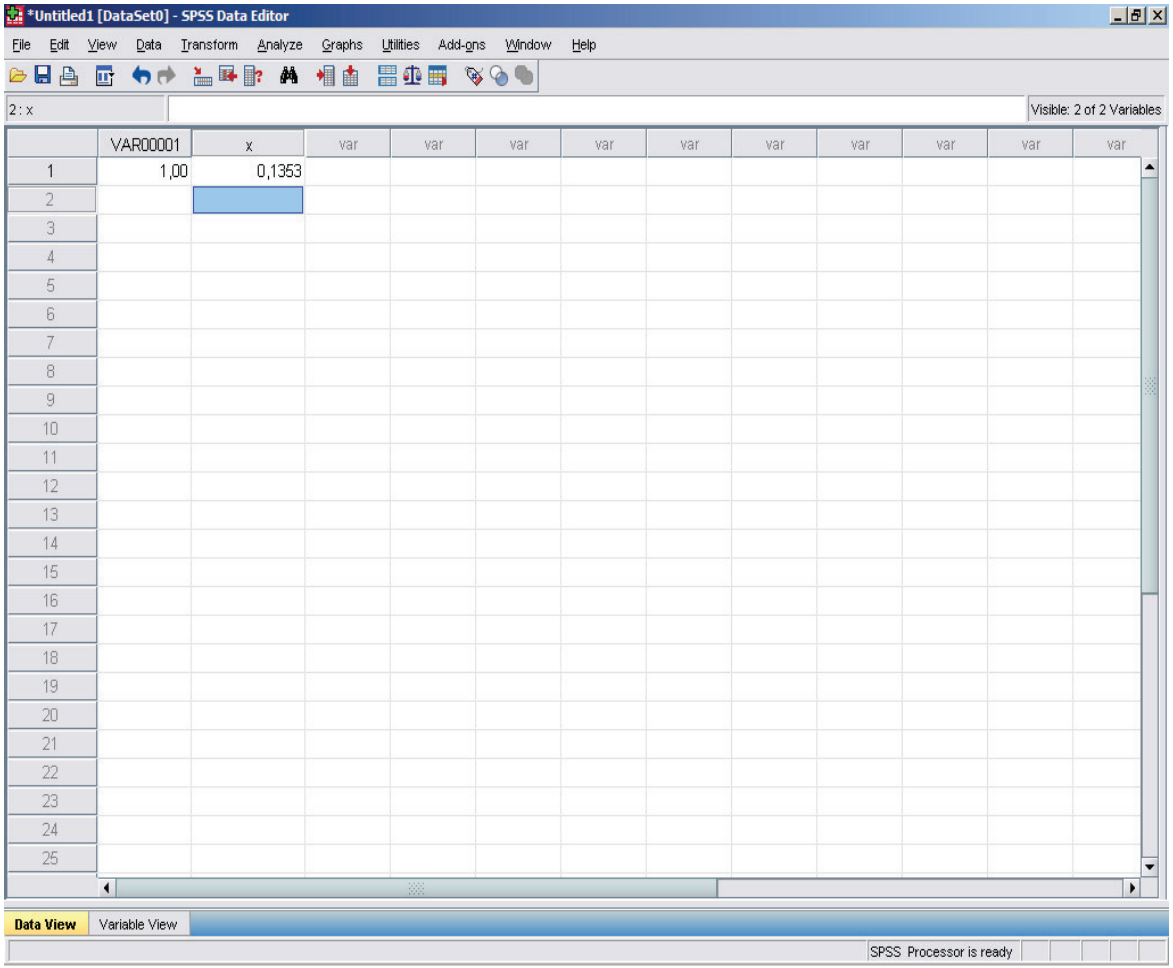
(d) Para determinarmos o número de colónias que ocorre com maior frequência, listamos os primeiros valores da função de probabilidade. Tem-se (apresentamos os cálculos em SPSS apenas para $X = 0$; os restantes valores são obtidos de forma análoga):

$$\begin{aligned} f_X(0) &= 0.135 \\ f_X(1) &\approx 0.271 \\ f_X(2) &\approx 0.271 \\ f_X(3) &= 0.180 \\ f_X(4) &= 0.090 \end{aligned}$$

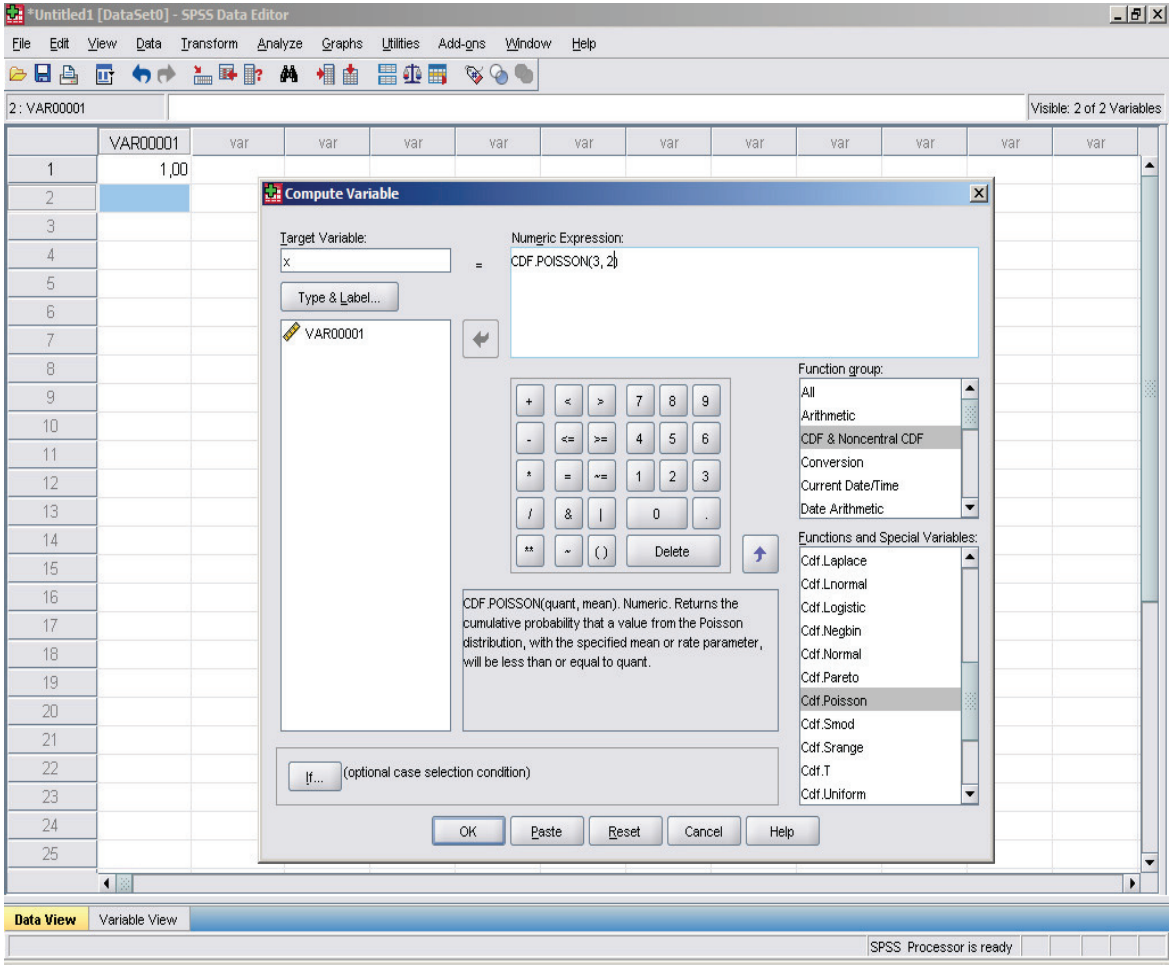
portanto a resposta será 1 ou 2 colónias de bactérias (considerando 3 casas decimais; se aumentarmos o número de casas decimais, verificamos que a moda de X é 2).

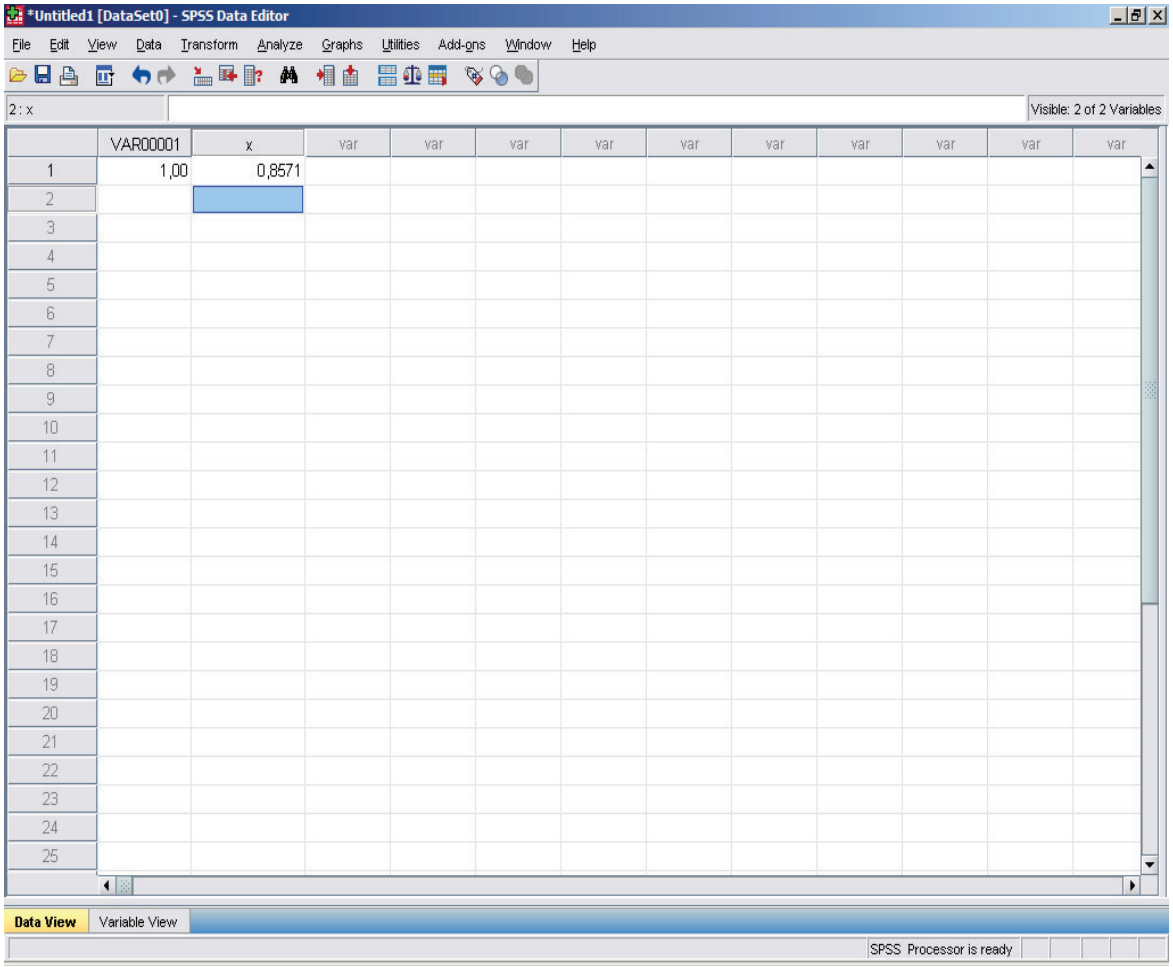
(a)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
PDF.POISSON(0, 2)
Function group:
Date Creation
Date Extraction
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Functions and Special Variables:
Pdf.Geom
Pdf.Halfnrm
Pdf.Hyper
Pdf.Igauss
Pdf.Laplace
Pdf.Lnormal
Pdf.Logistic
Pdf.Negbin
Pdf.Normal
Pdf.Pareto
Pdf.Poisson
Delete
PDF.POISSON(quant, mean). Numeric. Returns the probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

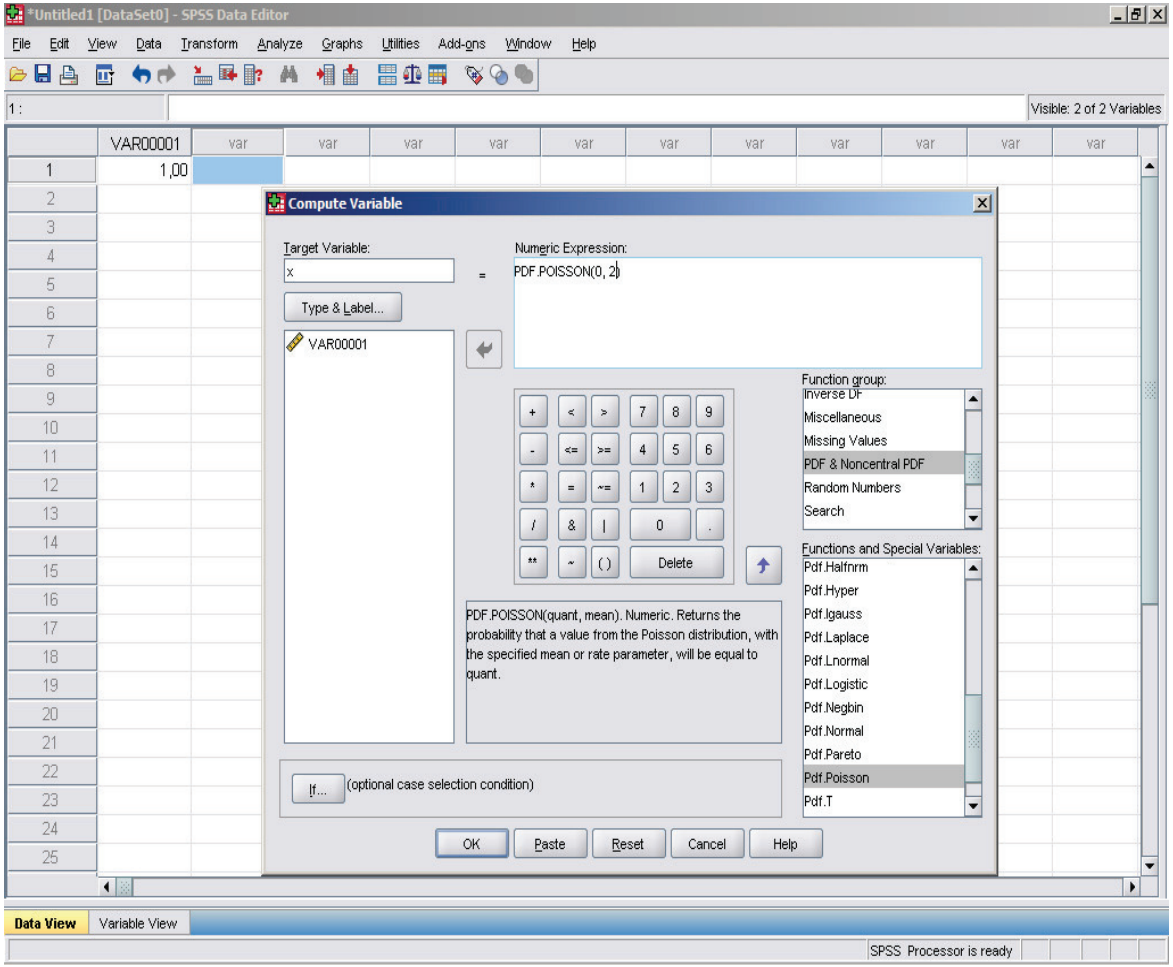

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,1353
Data View
Variable View
SPSS Processor is ready

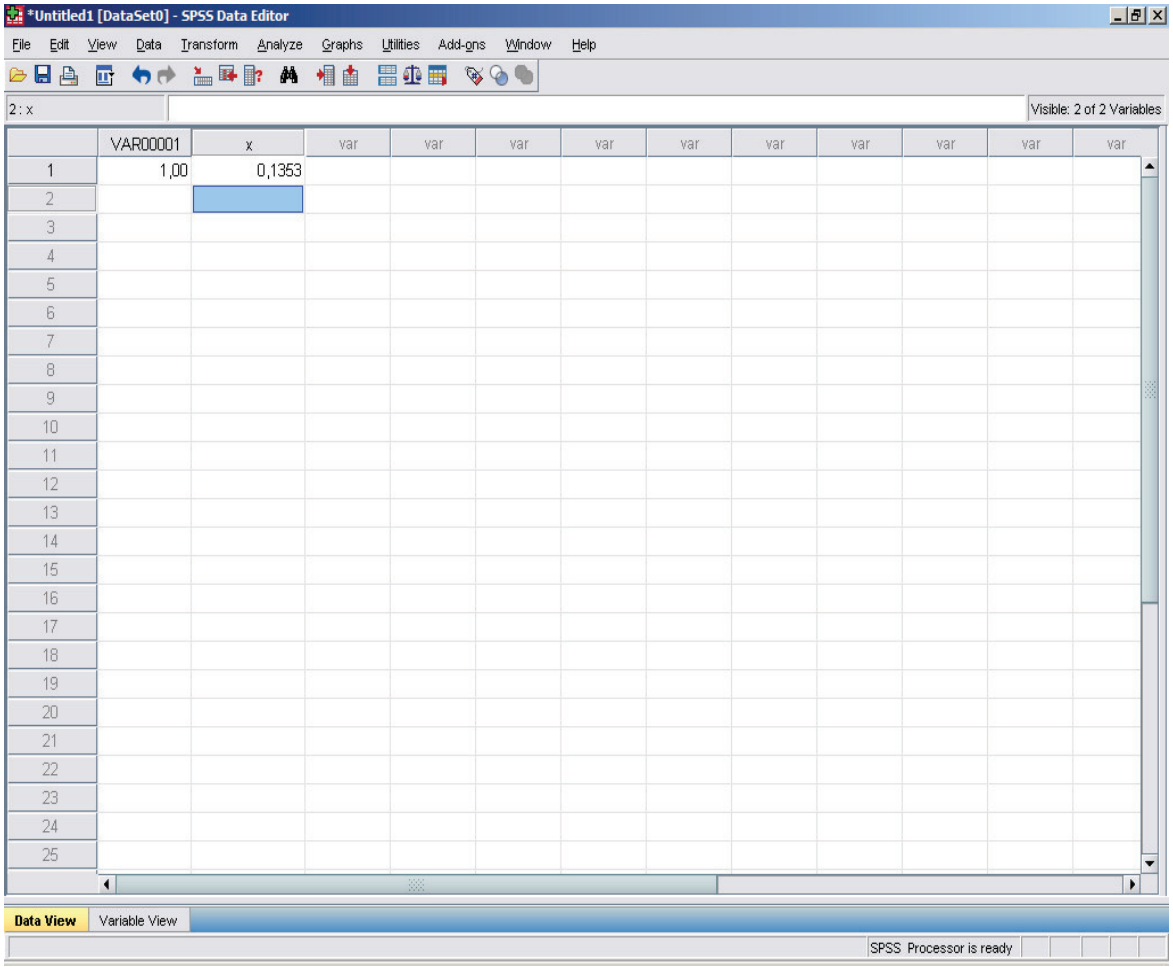
(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
CDF.POISSON(3, 2)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Cdf.Srange
Cdf.T
Cdf.Uniform
Delete
CDF.POISSON(quant, mean). Numeric. Returns the cumulative probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be less than or equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,8571
Data View
Variable View
SPSS Processor is ready

(d)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
PDF.POISSON(0, 2)
Delete
Function group:
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Random Numbers
Search
Functions and Special Variables:
Pdf.Halfnrm
Pdf.Hyper
Pdf.Igauss
Pdf.Laplace
Pdf.Lnormal
Pdf.Logistic
Pdf.Negbin
Pdf.Normal
Pdf.Pareto
Pdf.Poisson
Pdf.T
PDF.POISSON(quant, mean). Numeric. Returns the probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,1353
Data View
Variable View
SPSS Processor is ready

Exemplo 3: Um estudo realizado por Ott *et al* analisou a relação entre a exposição a etileno-dibromido (*ethylene dibromide*, EDB) e morte por cancro. Seguiram-se 161 homens de raça branca, trabalhadores de duas fábricas no Texas e Michigan e expostos ao EDB, de 1940 a 1975. Foram observadas 7 mortes por cancro nessa amostra. Para o período em questão, o número esperado de mortes por cancro a partir de estudos sobre a taxa de mortalidade para indivíduos de raça branca nos EUA era de 5.8.
O número observado de 7 mortes no estudo de Ott *et al* foi excessivo?

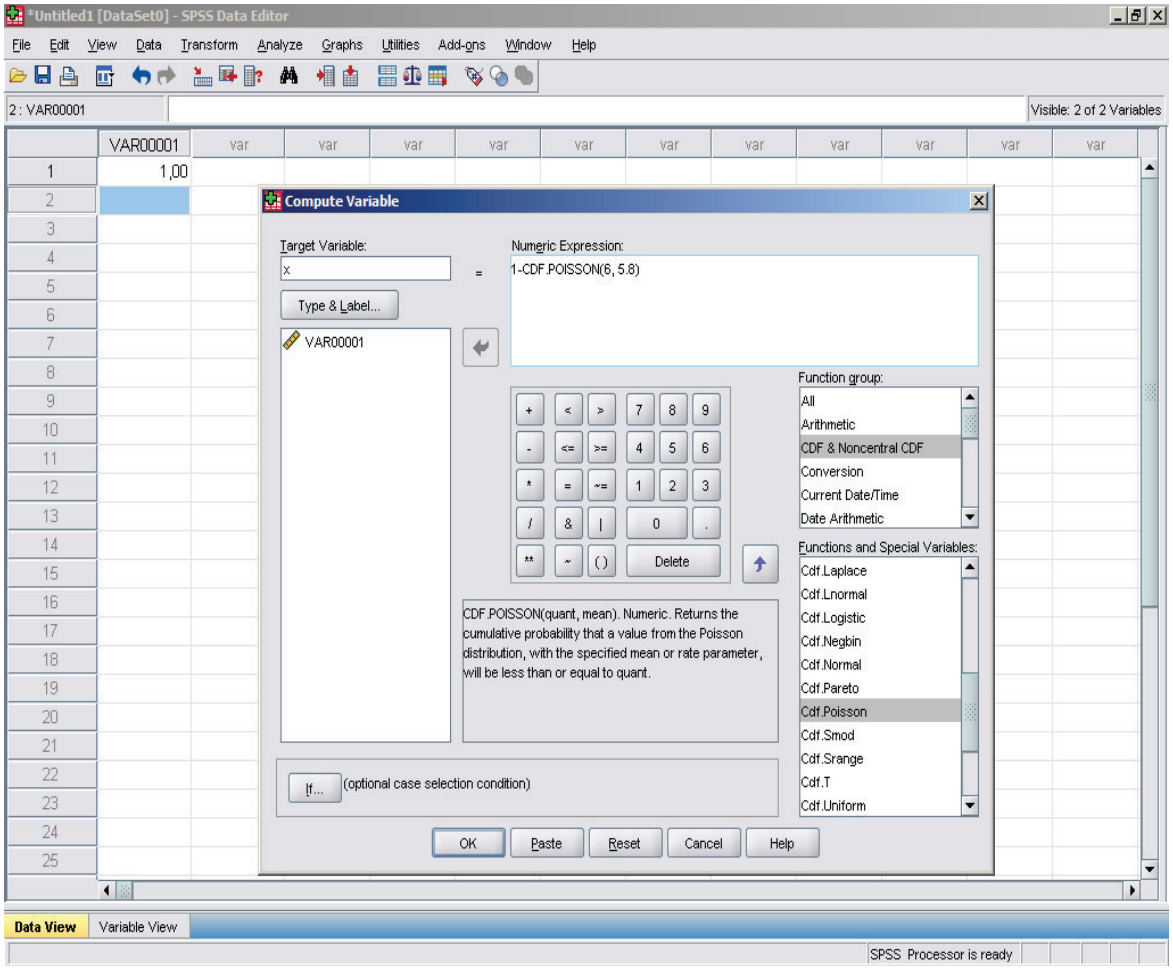
Resolução: Seja X a v.a. que representa o número de mortes por cancro no perído 1940-1975. De estudos anteriores, tem-se então

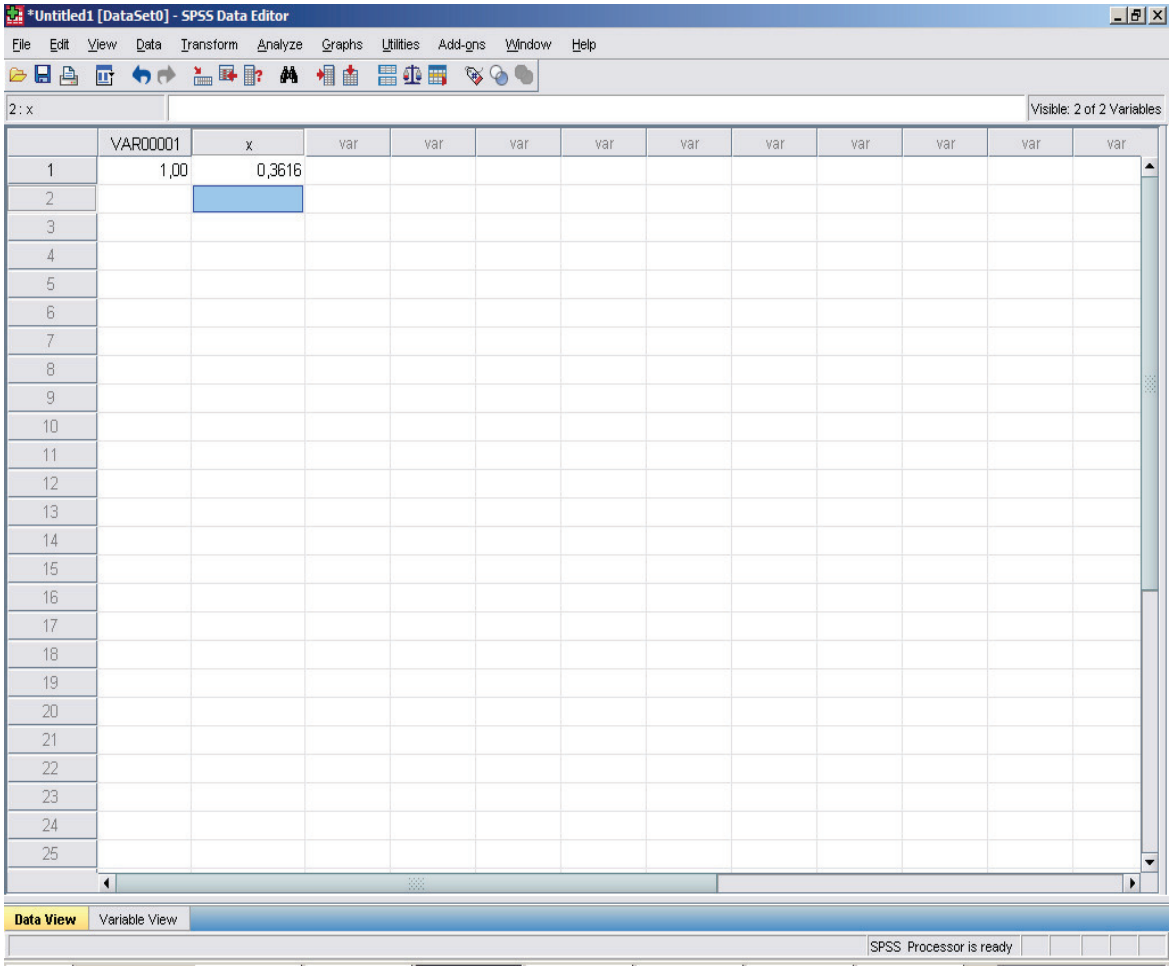
$$X \sim P(5.8).$$

Para avaliar se o número de mortes observadas no estudo de Ott *et al* foi exageramente elevado, vamos calcular a probabilidade de encontrar nesse estudo pelo menos 7 mortes. Ora

$$P(X \geq 7) = 1 - P(X \leq 6) = 0.362$$

portanto, à primeira vista, não nos parece que o número de mortes tenha sido excessivo.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : VAR00001
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
1-CDF.POISSON(6, 5.8)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Cdf.Srange
Cdf.T
Cdf.Uniform
Delete
CDF.POISSON(quant, mean). Numeric. Returns the cumulative probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be less than or equal to quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,3616
Data View
Variable View
SPSS Processor is ready

2.1.4. Distribuição Multinomial $M(n_1, \ldots, n_k; p_1, \ldots, p_k)$

- $A_1, A_2, \ldots, A_k$ acontecimentos mutuamente exclusivos, cuja reunião é todo o espaço amostral
- p_i = probabilidade do acontecimento A_i ocorrer
  $0 < p_i < 1, \qquad p_1 + p_2 + \cdots + p_k = 1$
- n = # ensaios (experiências mutuamente independentes para as quais existem precisamente k realizações possíveis: A_1 ou A_2 ou ... ou A_k)
- n_i = número de ocorrências do acontecimento A_i
  $n_1 + n_2 + \ldots n_k = n$

$$m(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k)$$

é a probabilidade de, em n ensaios, A_1 ocorrer n_1 vezes, A_2 ocorrer n_2 vezes, ..., A_k ocorrer n_k vezes.

Frequência	$m(n_1, \ldots, n_k)$ = prob. A_i acontecer n_i vezes em n ensaios $= \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k} \qquad (1 \leq i \leq k)$
Distribuição	$M(n_1, \ldots, n_k)$ $= \sum_{i_1=0}^{n_1} \cdots \sum_{i_k=0}^{n_k} m(i_1, \ldots, i_k)$
Domínio	$\{(n_1, \ldots, n_k) \in \mathbb{N}_0^k : n_1 + \ldots + n_k = n\}$
Média	$E(n_i) = np_i \qquad (1 \leq i \leq k)$
Variância	$\mathrm{Var}(n_i) = np_i(1 - p_i) \qquad (1 \leq i \leq k)$

A distribuição multinomial é uma generalização da distribuição binomial ($k = 2$).

Exemplo: Suponhamos que, de estudos anteriores, se sabe que de entre a população feminina residente no Grande Porto com idade superior a 30 anos: 8% ingere uma quantidade elevada de carne vermelha, 52% ingere uma quantidade moderada de carne vermelha, e 40% ingere uma quantidade reduzida de carne vermelha. Foi considerada uma amostra de 100 mulheres residentes no Grande Porto, com mais de 30 anos. Determine a probabilidade de, nessa amostra:

(a) encontrar exactamente 8 mulheres com consumo elevado de carne vermelha, 52 com consumo moderado e 40 com consumo reduzido.

(b) encontrar apenas mulheres com consumo moderado.

Resolução: Na experiência em causa, cada ensaio de Bernoulli corresponde à avaliação de cada uma das 100 mulheres quanto à quantidade de carne ingerida. Tem-se $p_1 = 0.08, \quad p_2 = 0.52, \quad p_3 = 0.40$.

(a) Pretende-se determinar $m(8, 52, 40)$. Como o vector das probabilidades correspondentes é $(0.8, 0.52, 0.40)$ a instrução correspondente é

```
> dmultinom(c(8,52,40), prob=c(.8,.52,.4))
[1] 3.409089e-18
```

A probabilidade é aproximadamente zero.

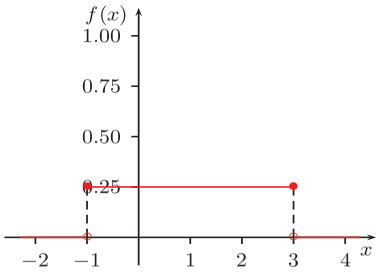
(b)
```
> dmultinom(c(0,100,0), prob=c(.8,.52,.4))
[1] 1.115552e-52
```

A probabilidade é aproximadamente zero.

2.1.5. Distribuição Uniforme Contínua: $U(a,b)$

A variável aletória toma qualquer valor de um intervalo real com a mesma probabilidade. Isto significa que a função densidade de probabilidade é constante e igual ao inverso da amplitude do intervalo.

Função dens. de prob.	$u(x) = \frac{1}{b-a}$
Função de distribuição	$U(x) = \frac{x-a}{b-a}$
Domínio	$a \leq x \leq b$
Média	$\frac{a+b}{2}$
Mediana	$\frac{a+b}{2}$
Desvio padrão	$\sigma = \sqrt{\frac{(b-a)^2}{12}}$
Amplitude	$R = b - a$
Coef. Assimetria	$\lambda = 0$



Exemplo: Suponha que a chegada de clientes ao balcão de uma determinada farmácia do Porto segue uma distribuição uniforme. Sabe-se que durante um período de 30 minutos um certo cliente chegou a essa farmácia.

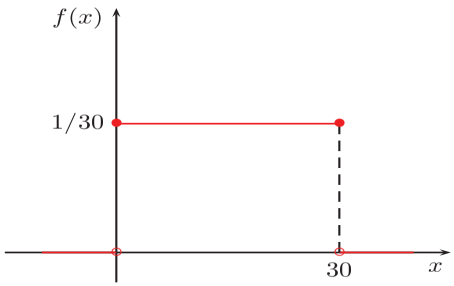
(a) Qual é a probabilidade de que esse cliente tenha chegado durante os últimos 5 minutos daquele período?

(b) Como é que relaciona a probabilidade de que esse cliente tenha chegado nos primeiros cinco minutos com a probabilidade anterior?

(c) Como é que relaciona a probabilidade de que esse cliente tenha chegado entre os primeiros cinco e os dez minutos com a probabilidade calculada em (a)?

Resolução: Por hipótese o tempo real de chegada à farmácia, X, segue uma distribuição uniforme no intervalo $[0, 30]$. A f.d.p. de X é então

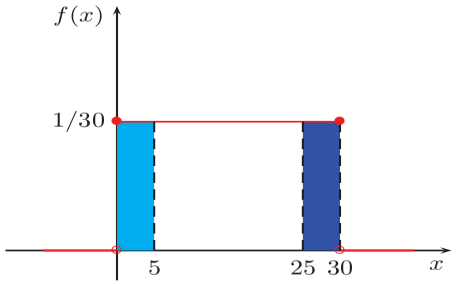
$$f(x) = \begin{cases} 1/30, & x \in [0,30] \\ 0, & \text{caso contrário.} \end{cases}$$



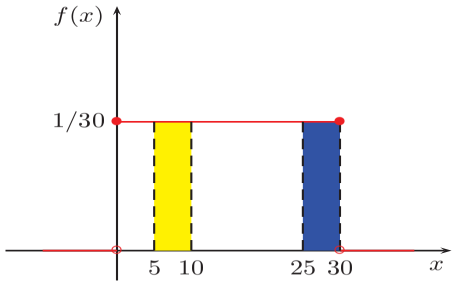
(a) O que se pretende é $P(25 < X < 30)$. Ora

$$\begin{aligned} P(25 < X < 30) &= F(30) - F(25) \\ &= \frac{(30-0)}{30} - \frac{(25-0)}{30} \\ &= \frac{5}{30} = \frac{1}{6}. \end{aligned}$$

(b) $P(0 < X < 5) = F(5) - F(0) = \frac{5-0}{30} - 0 = \frac{1}{6}$.
Portanto $P(0 < X < 5) = P(25 < X < 30)$.



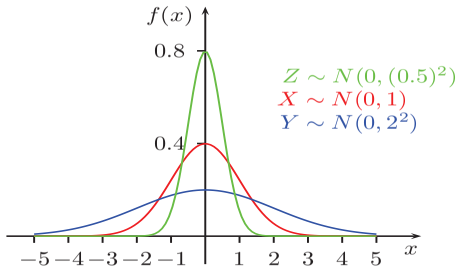
(c) $P(5 < X < 10)$
$= F(10) - F(5) = \frac{10-0}{30} - \frac{1}{6} = \frac{1}{6}$.
Portanto $P(5 < X < 10) = P(25 < X < 30)$.



2.1.6. Distribuição Normal: $N(\mu, \sigma^2)$

É uma das distribuições mais importantes no estudo das probabilidades e da estatística.

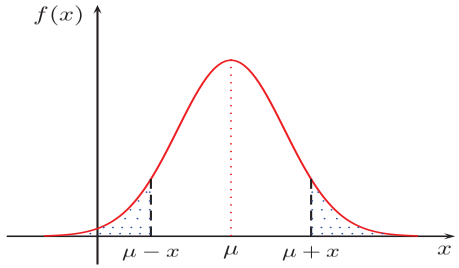
Densidade	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ $(\mu \in \mathbb{R}, \quad \sigma > 0)$
Domínio	$-\infty < x < +\infty$
Média	μ
Mediana	μ
Desvio padrão	σ
Amplitude	∞
Coef. Assimetria	$\lambda = 0$



Propriedades da distribuição normal:

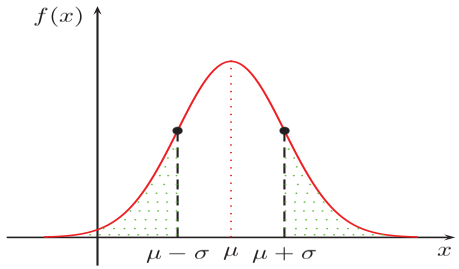
(a) fica completamente determinada pelos parâmetros μ e σ.

(b) é simétrica em relação à média:
$P(X > \mu + x) = P(X < \mu - x), \quad \forall x$



(c) a média, a moda e a mediana são iguais

(d) os pontos de abcissa $\mu \pm \sigma$ são os pontos de inflexão do gráfico da f.d.p.

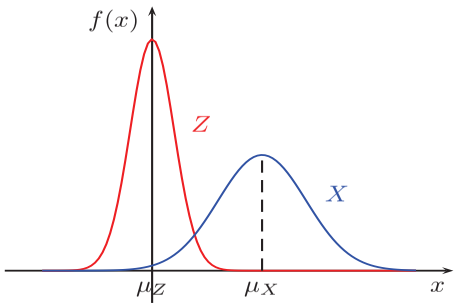


Esta propriedade permite estimar o valor de σ a partir do gráfico da f.d.p.

Redução de uma variável normal:

Se $X \sim N(\mu, \sigma^2)$ e $Z = \frac{X-\mu}{\sigma}$ então $Z \sim N(0,1)$.

A v.a. Z diz-se a variável **reduzida**, **estandardizada** ou **padronizada**.

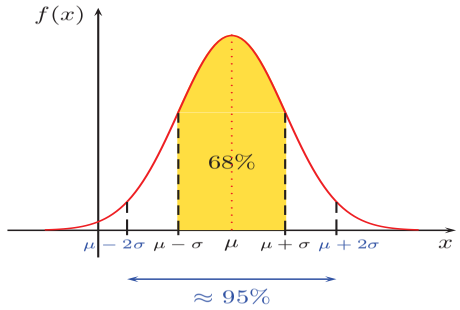


Os valores da função de distribuição de uma v.a. que segue uma distribuição normal podem ser determinados através de uma tabela correspondente ou de um software apropriado. No caso de uma tabela, os valores são os correspondentes à variável reduzida.

A redução de uma v.a. normal permite mostrar que:

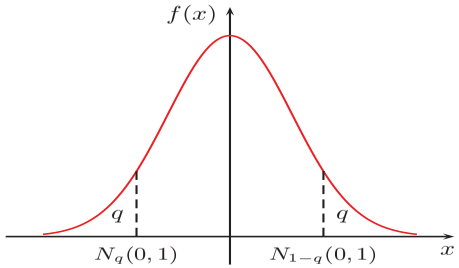
- 68% da área da f.d.p. pertence ao intervalo $[\mu - \sigma, \mu + \sigma]$
- 95% da área da f.d.p. pertence ao intervalo $[\mu - (1.96)\sigma, \mu + (1.96)\sigma]$.

  (aproximadamente 95% da massa probabilística pertence ao intervalo $[\mu - 2\sigma, \mu + 2\sigma]$)



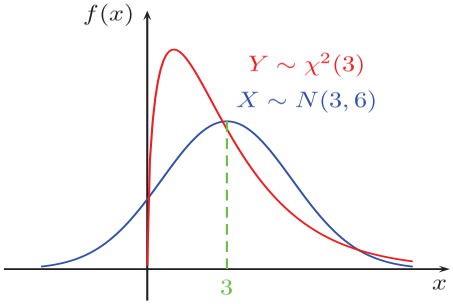
Em particular:

- $N_{0.5}(\mu, \sigma^2)$ [a] $= \mu$
- $N_{0.025}(\mu, \sigma^2) = \mu - (1.96)\sigma$
- $N_{0.975}(\mu, \sigma^2) = \mu + (1.96)\sigma$
- $N_{0.5}(0,1) = 0$
- $N_{0.025}(0,1) = (1.96)\sigma$
- $N_{0.975}(0,1) = (1.96)\sigma$
- $N_q(0,1) = -N_{1-q}(0,1)$



[a] $N_q(\mu, \sigma^2)$ representa o q-quantil de uma v.a. que segue a distribuição normal $N(\mu, \sigma^2)$

Duas distribuições com a mesma média e variância podem ser muito diferentes...



Mas duas distribuições normais com a mesma média e variância têm necessariamente de coincidir!

Proposição 2.1.2 *Sejam X_1, X_2, ..., X_n v.a. independentes e normalmente distribuídas, digamos*

$$X_i \sim N(\mu_i, \sigma_i^2), \qquad i = 1, 2, \ldots, n$$

e seja W a combinação linear

$$W = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

com a_1, a_2, ..., a_n constantes. Então W é uma v.a. que é também normalmente distribuída. Mais precisamente,

$$W \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

A demonstração desta Proposição não é apresentada mas encontra-se com facilidade num manual de estatística que contenha algum detalhe matemático.

Exemplo 1: Para a distribuição normal reduzida, $X \sim N(0,1)$, tem-se

(a) $P(X < 2) = P(X \leq 2) \approx 0.9772$

Este valor pode ser obtido através do software de estatística SPSS, como se mostra nas páginas seguintes, ou por consulta de uma tabela da distribuição normal (método em desuso).

Interpretação: Aproximadamente 97.72% dos valores de X estão entre $-\infty$ e 2; equivalentemente, um valor x escolhido aleatoriamente da população tem uma probabilidade de 0.9772 de estar entre $-\infty$ e 2.

(b) $P(X \geq 2.71) = 1 - P(X < 2.71)$
$= 1 - P(X \leq 2.71)$
$= .0034.$

(c) $P(-2.55 < X < 2.55)$
$= P(X \leq 2.55) - P(X \leq -2.55)$
$= 0.9892$

ou
$= 1 - 2 \times P(X \leq -2.55)$

(d) $P(0.84 \leq X \leq 2.45)$
$= P(X \leq 2.45) - P(X \leq 0.84)$
$= 0.1934$ (instruções em SPSS análogas às de c))

(e) $N_{0.90}(0,1) = 1.282$

Trata-se do quantil 0.90 da distribuição $N(0,1)$, isto é, do ponto x para o qual

$$P(X \leq x) = 0.9.$$

A instrução em SPSS usa a função inversa da função de distribuição (IDF - **I***nverse of the (cumulative)* **D***istribution* **F***unction*)

(f) $N_{0.5}(0,1) = 0$

O quantil 0.5 da distribuição $N(0,1)$ corresponde ao número x para o qual

$$P(X \leq x) = 0.5.$$

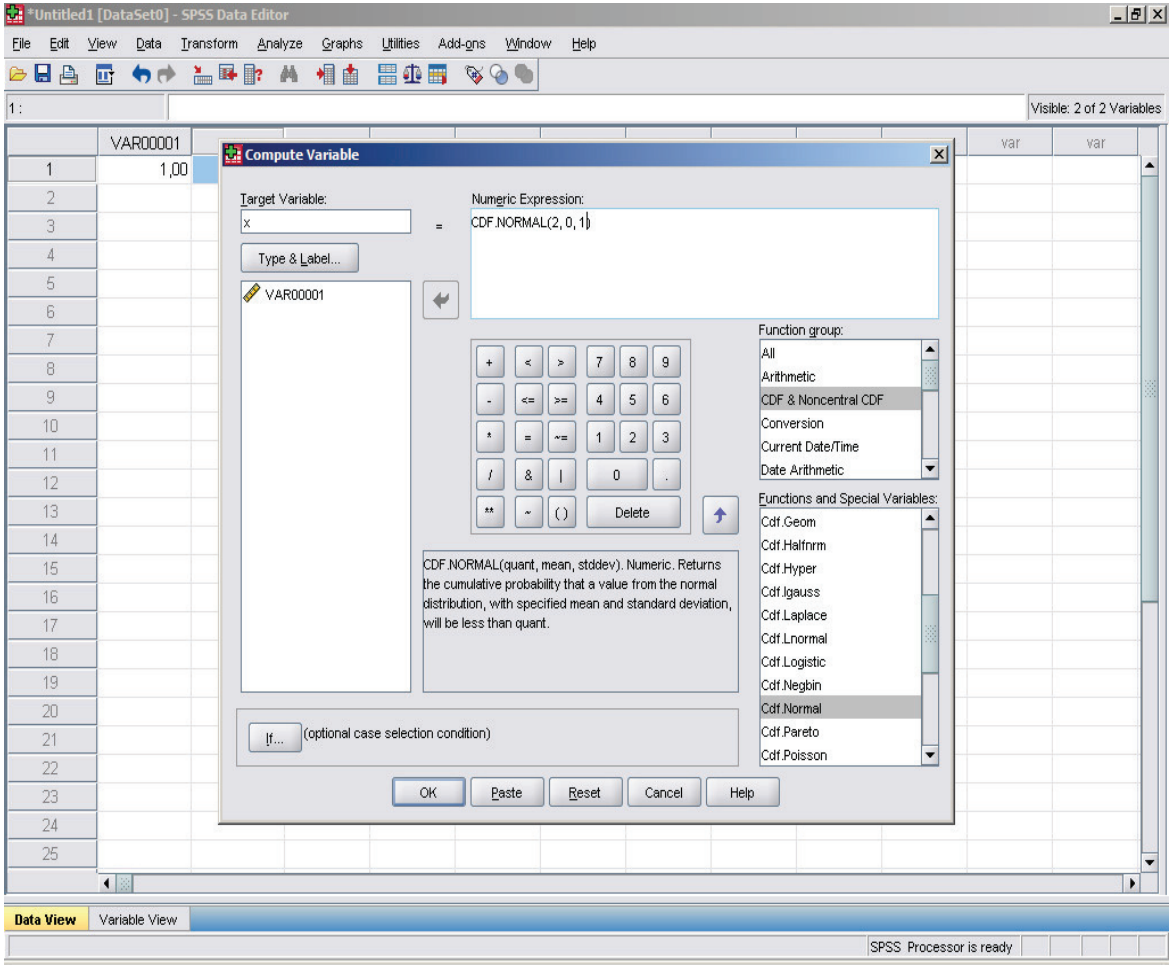
Ora, por definição, tal ponto x é a mediana da distribuição. Como a distribuição normal é simétrica, a mediana coincide com a média e portanto $N_{0.5}(0,1) = 0$.

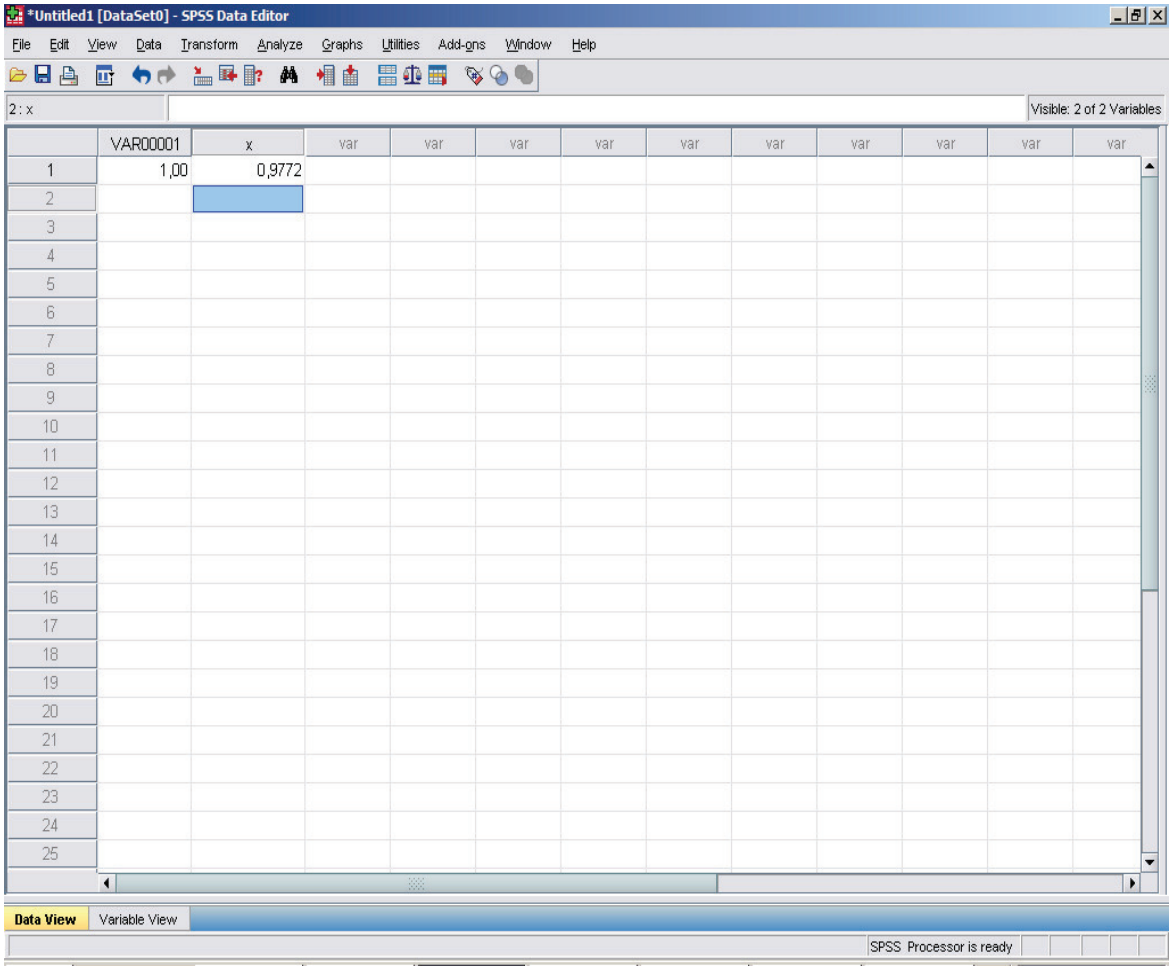
Nota: No **SPSS**, a determinação dos quantis de uma v.a. contínua obtem-se da instrução

IDF - **I***nverse of the (cumulative)* **D***istribution* **F***unction.*

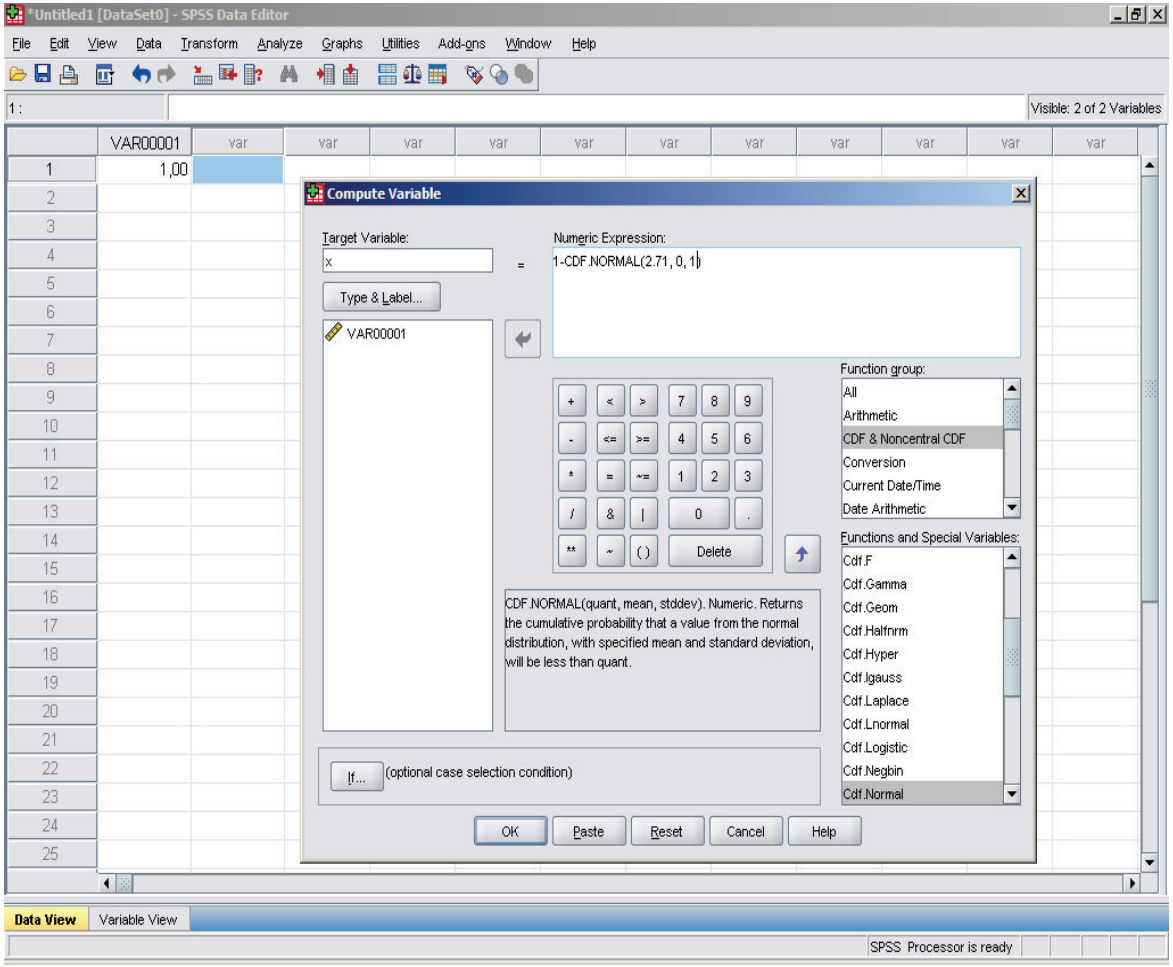
Quantis de v.a. discretas não podem ser obtidos visto que nesse caso as funções de distribuição são funções em escada, portanto não injectivas e consequentemente não invertíveis.

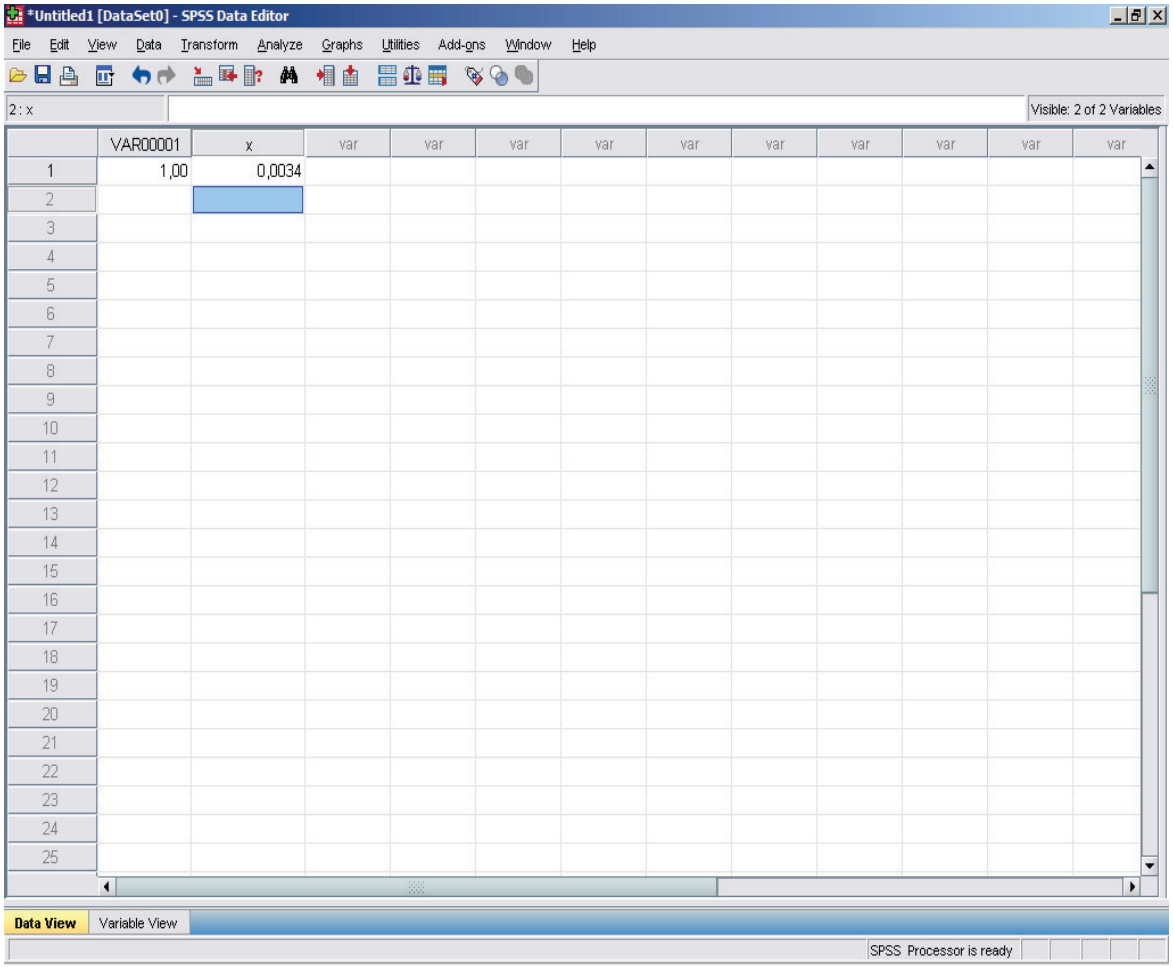
(a)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
CDF.NORMAL(2, 0, 1)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

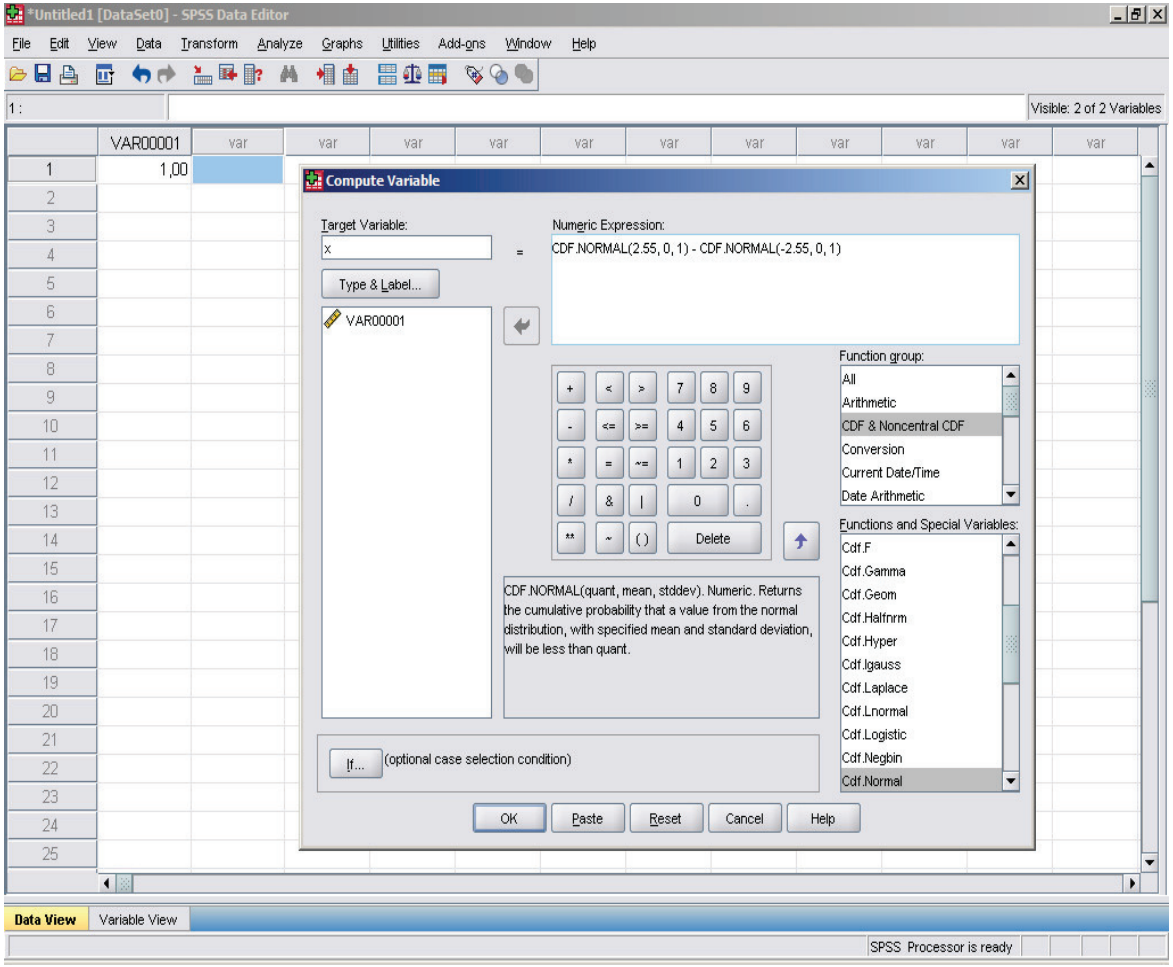

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,9772
Data View
Variable View
SPSS Processor is ready

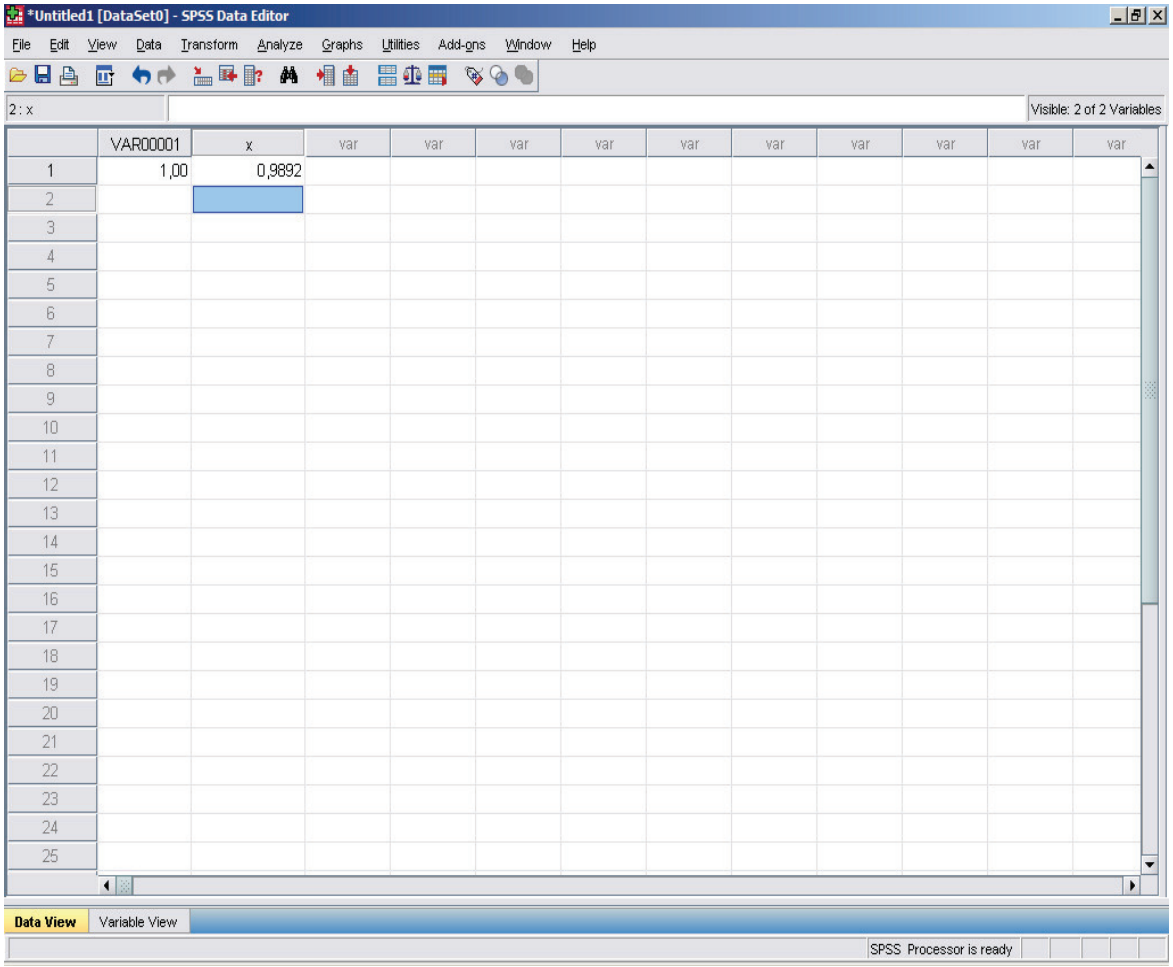
(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
1-CDF.NORMAL(2.71, 0, 1)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

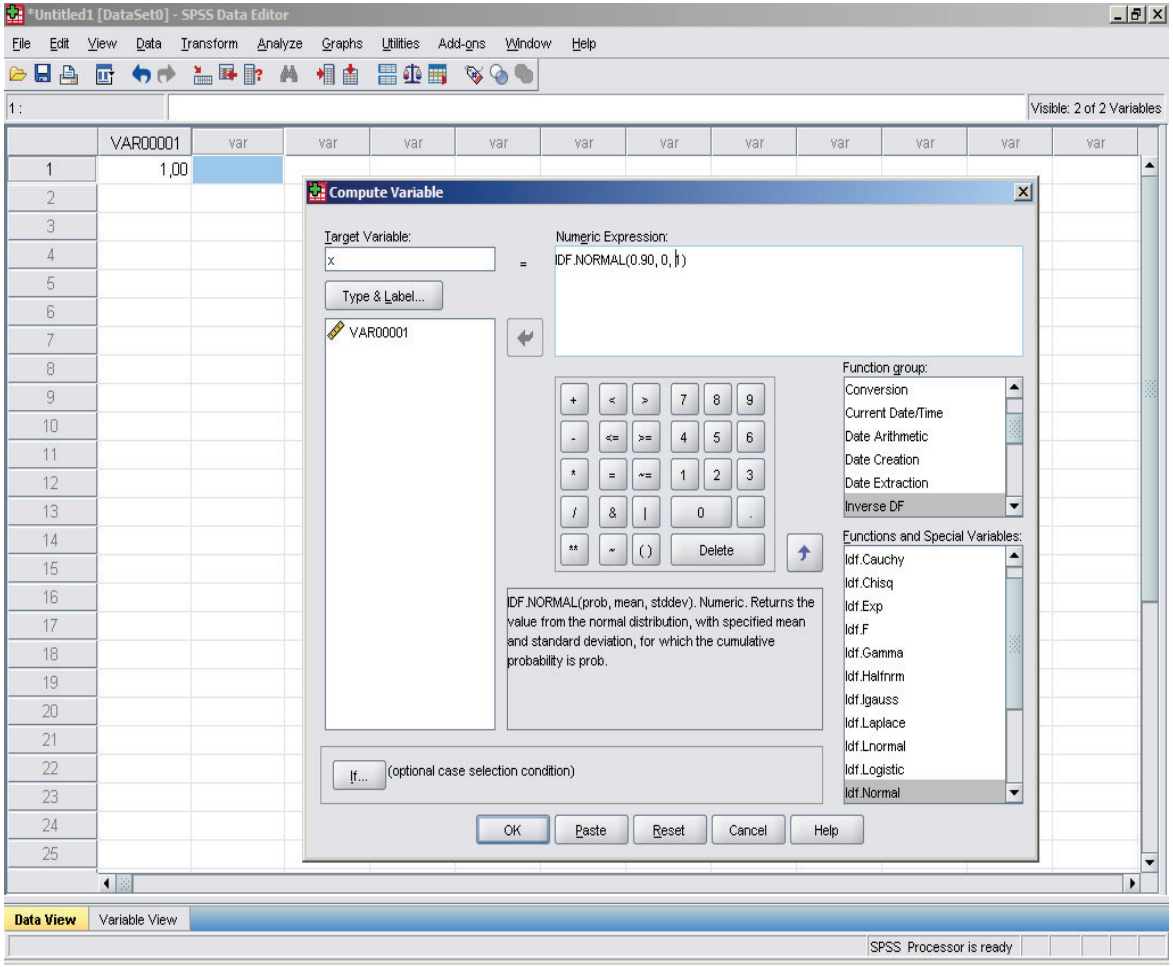

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,0034
Data View
Variable View
SPSS Processor is ready

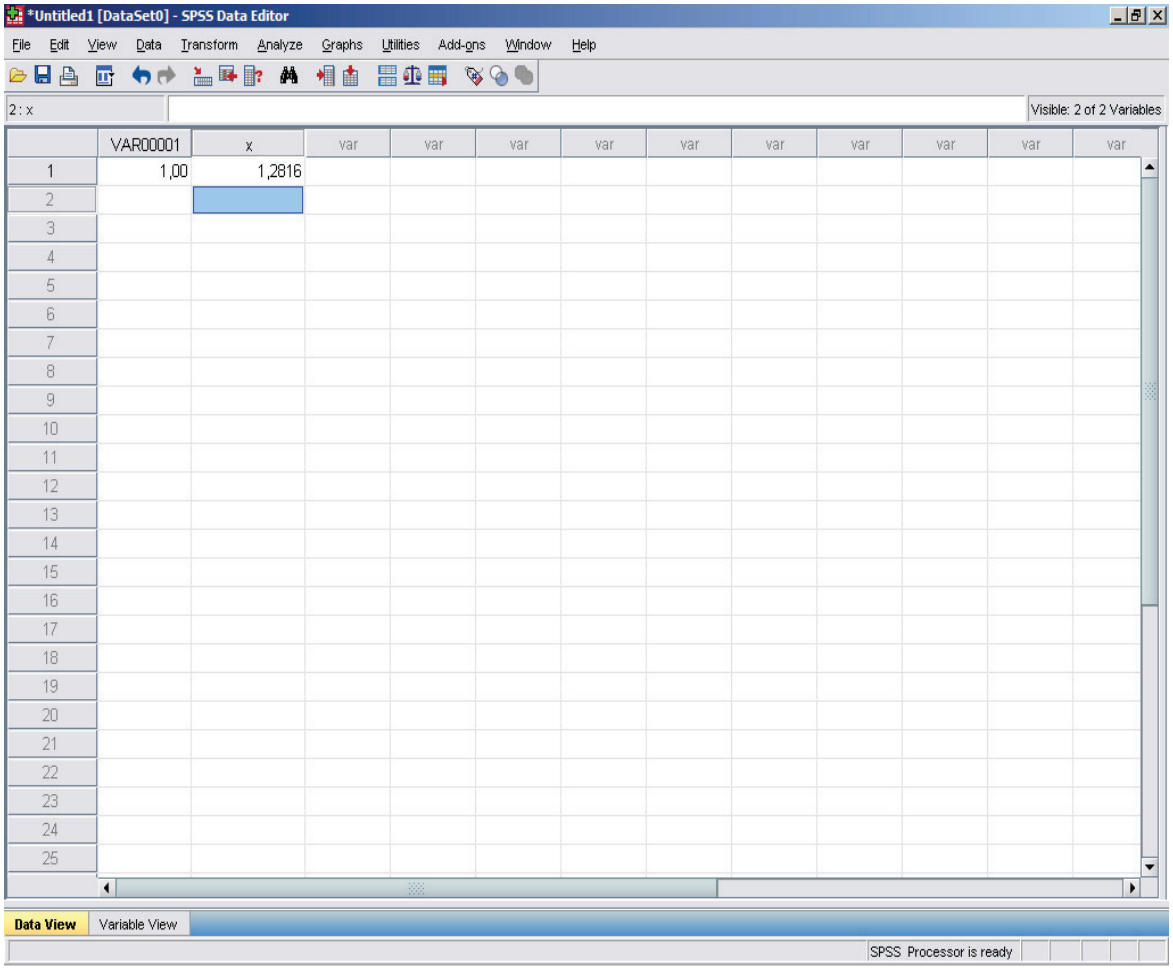
(c)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
CDF.NORMAL(2.55, 0, 1) - CDF.NORMAL(-2.55, 0, 1)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,9892
Data View
Variable View
SPSS Processor is ready

(e)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
IDF.NORMAL(0.90, 0, 1)
Function group:
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Date Extraction
Inverse DF
Functions and Special Variables:
Idf.Cauchy
Idf.Chisq
Idf.Exp
Idf.F
Idf.Gamma
Idf.Halfnrm
Idf.Igauss
Idf.Laplace
Idf.Lnormal
Idf.Logistic
Idf.Normal
Delete
IDF.NORMAL(prob, mean, stddev). Numeric. Returns the value from the normal distribution, with specified mean and standard deviation, for which the cumulative probability is prob.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
1,2816
Data View
Variable View
SPSS Processor is ready

Exemplo 2: Supondo que os valores de colesterol total de uma determinada população seguem aproximadamente uma distribuição normal de média 200 mg/100 ml e desvio padrão 20 mg/100 ml, determine:

(a) a probabilidade de um indivíduo escolhido aleatoriamente desta população ter um valor de colesterol:

(a.1) inferior a 150 mg/100 ml

(a.2) superior a 225 mg/100 ml

(a.3) entre 180 e 200 mg/100 ml

(b) o valor de colesterol a partir do qual se observam apenas 10% dos valores da população.

(c) em que percentil se encontra um indivíduo com um valor de colesterol total de 213 mg/100 ml? Interprete o resultado.

Resolução: Seja X a v.a. que representa os valores de colesterol da população em questão. Por hipótese, $X \sim N(200, 20^2)$.

(a.1) $P(X < 150) \approx 0.006$ [a]

(a.2) $P(X > 225) = 1 - P(X \leq 225) \approx 0.106$

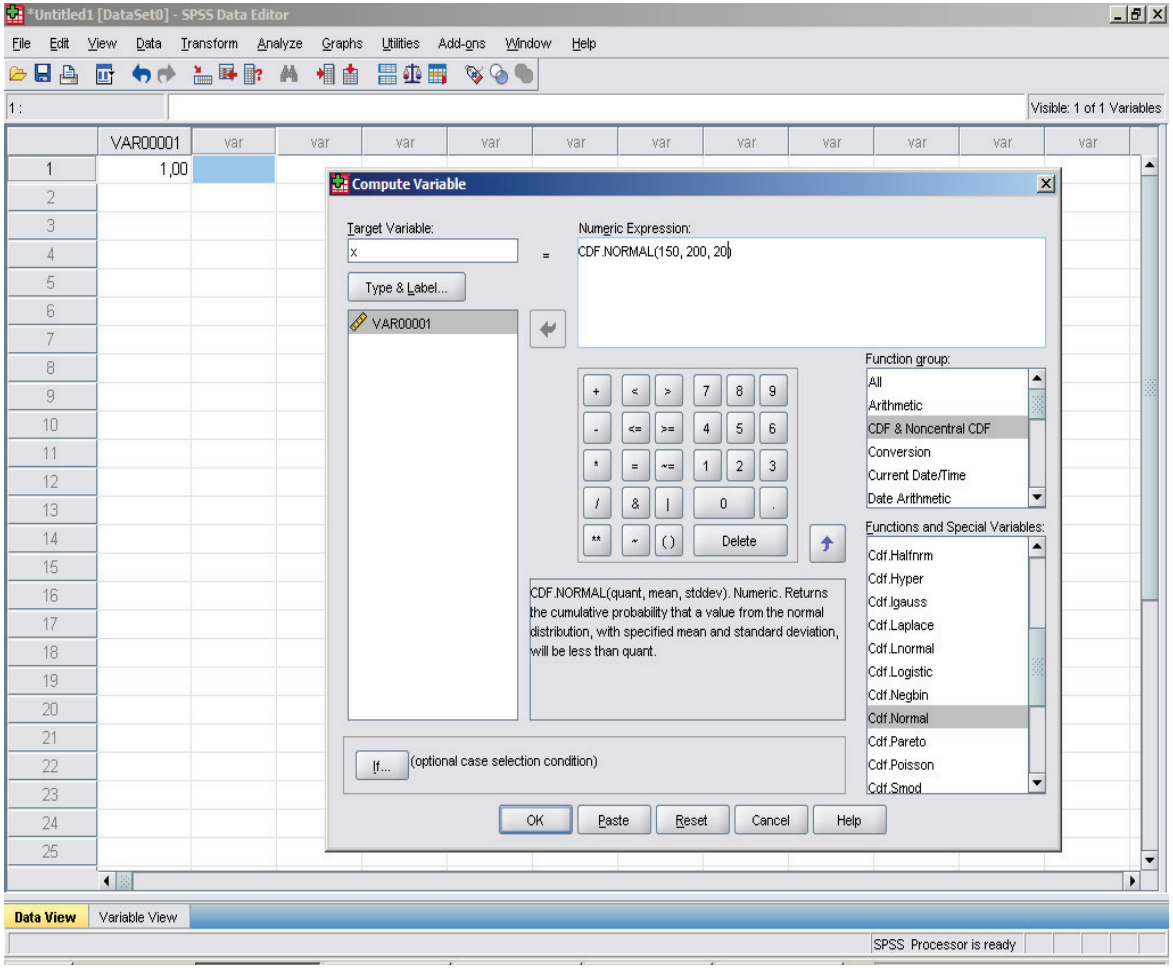
(a.3) $P(180 < X < 200) = P(X \leq 200) - P(X \leq 180) \approx 0.341$

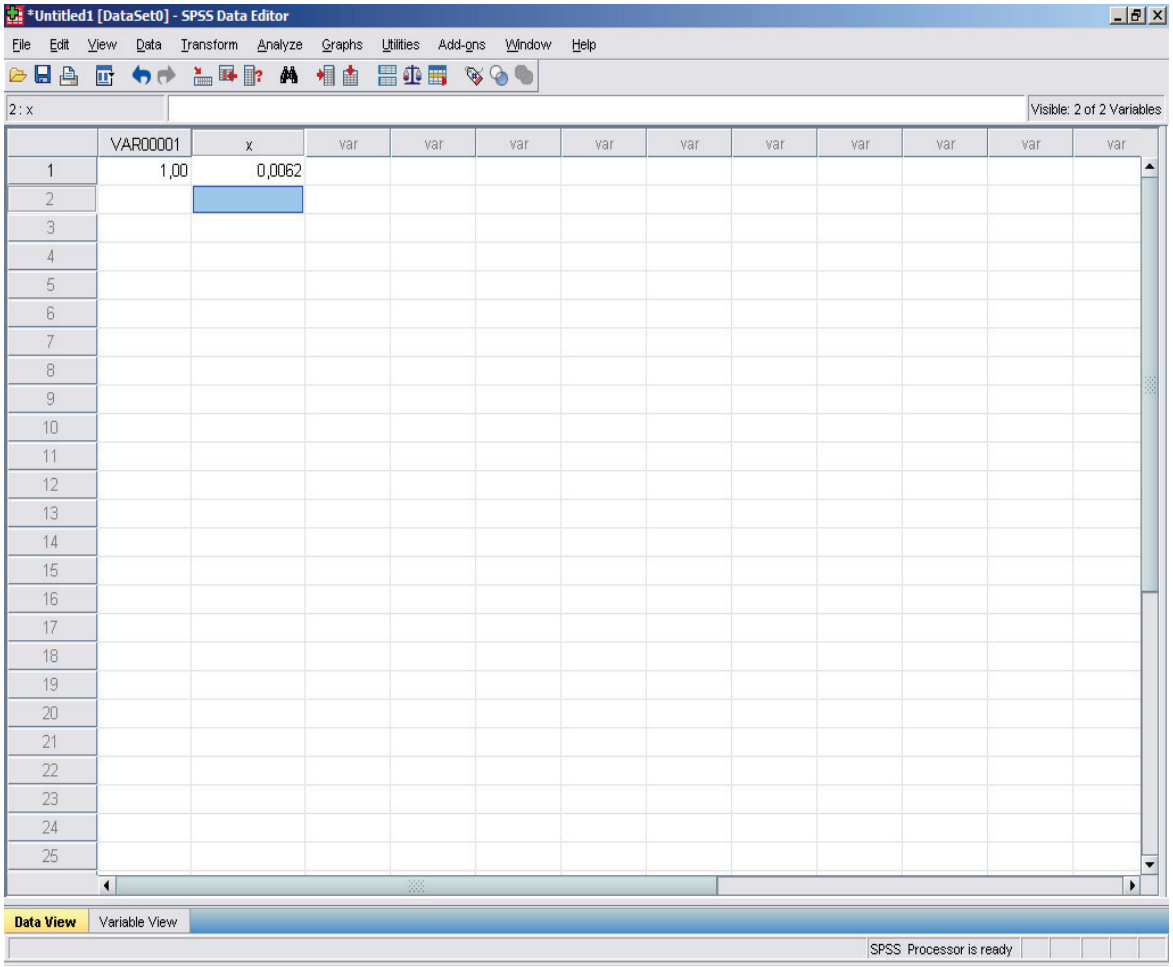
[a] Recorrendo a uma tabela da distribuição normal teríamos de reduzir a variável, considerando $Z = \frac{X-200}{20}$, e fazer $P(X < 150) = P\left(Z < \frac{150-200}{20}\right) = P(Z < -2.5)$

(b) Pretende-se determinar o 0.9-quantil, o que dá aproximadamente 225.6. A instrução em SPSS (à frente) usa a função inversa da função de distribuição.

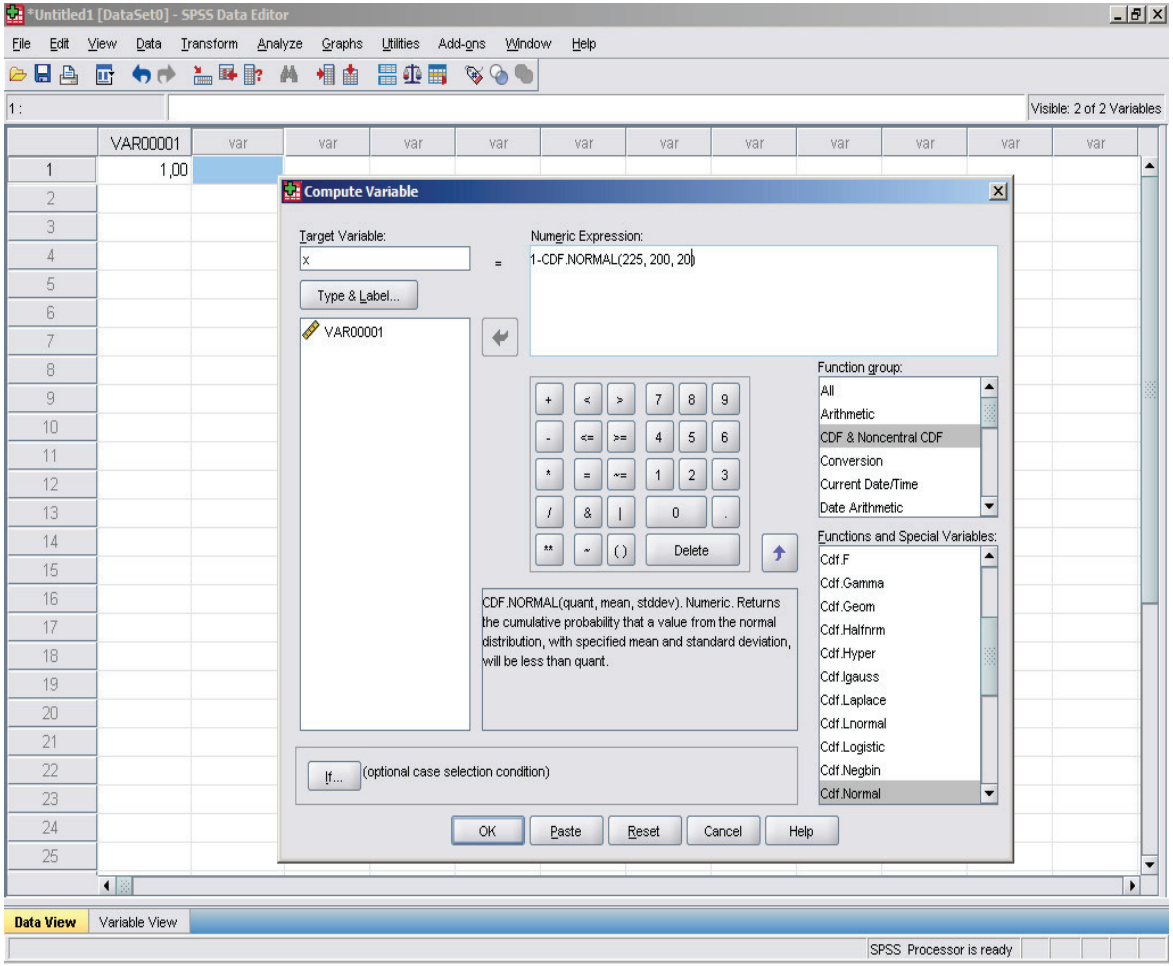
(c) $P(X \leq 213) \approx 0.74$ portanto o indivíduo em causa pertence ao percentil 74. Isto significa que, em 100 indivíduos escolhidos aleatoriamente da população em causa, esperamos encontrar 74 com valores de colesterol total inferiores a 213 mg/100 ml. Em toda a população, exactamente 74% dos indivíduos têm um valor de colesterol inferior a 213 mg/100 ml.

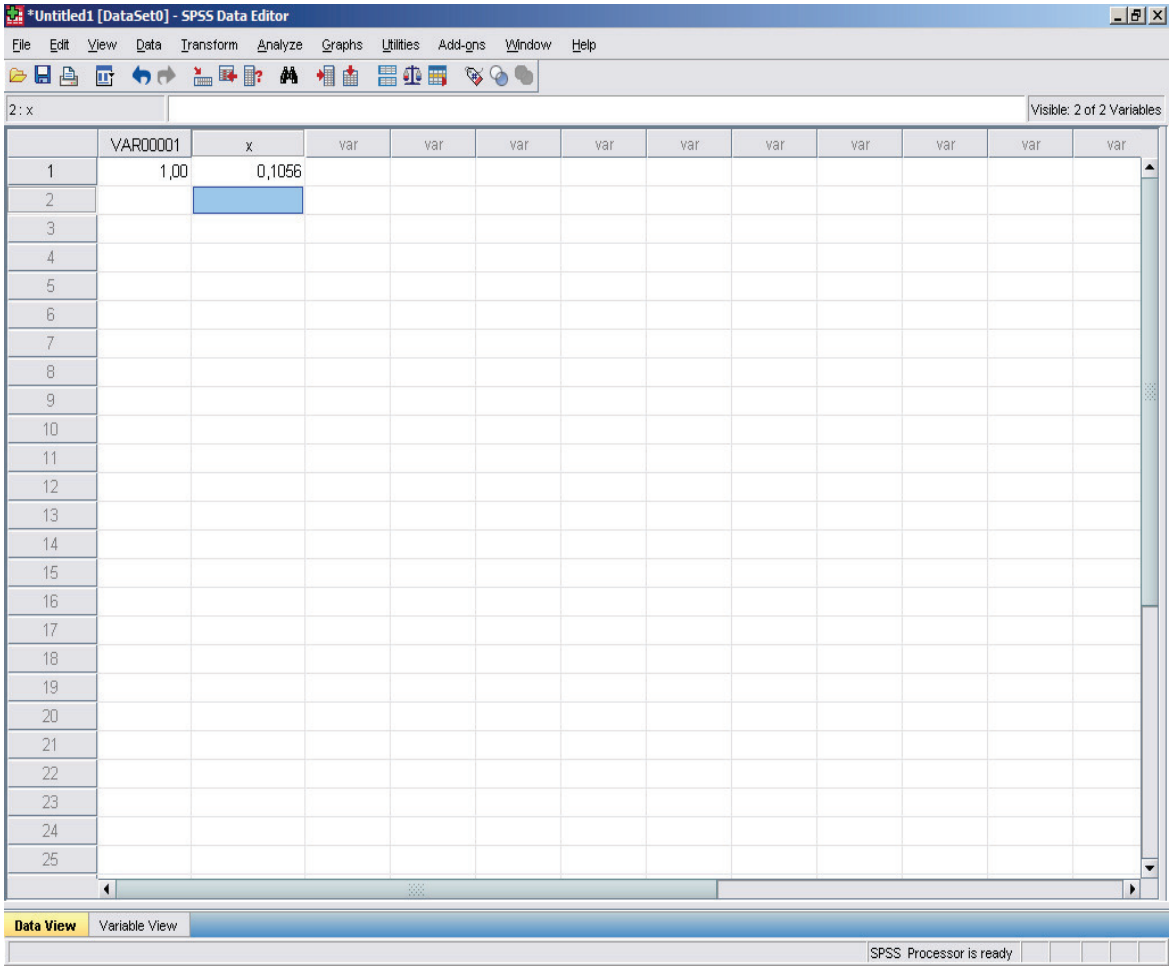
(a.1)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 1 of 1 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
CDF.NORMAL(150, 200, 20)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

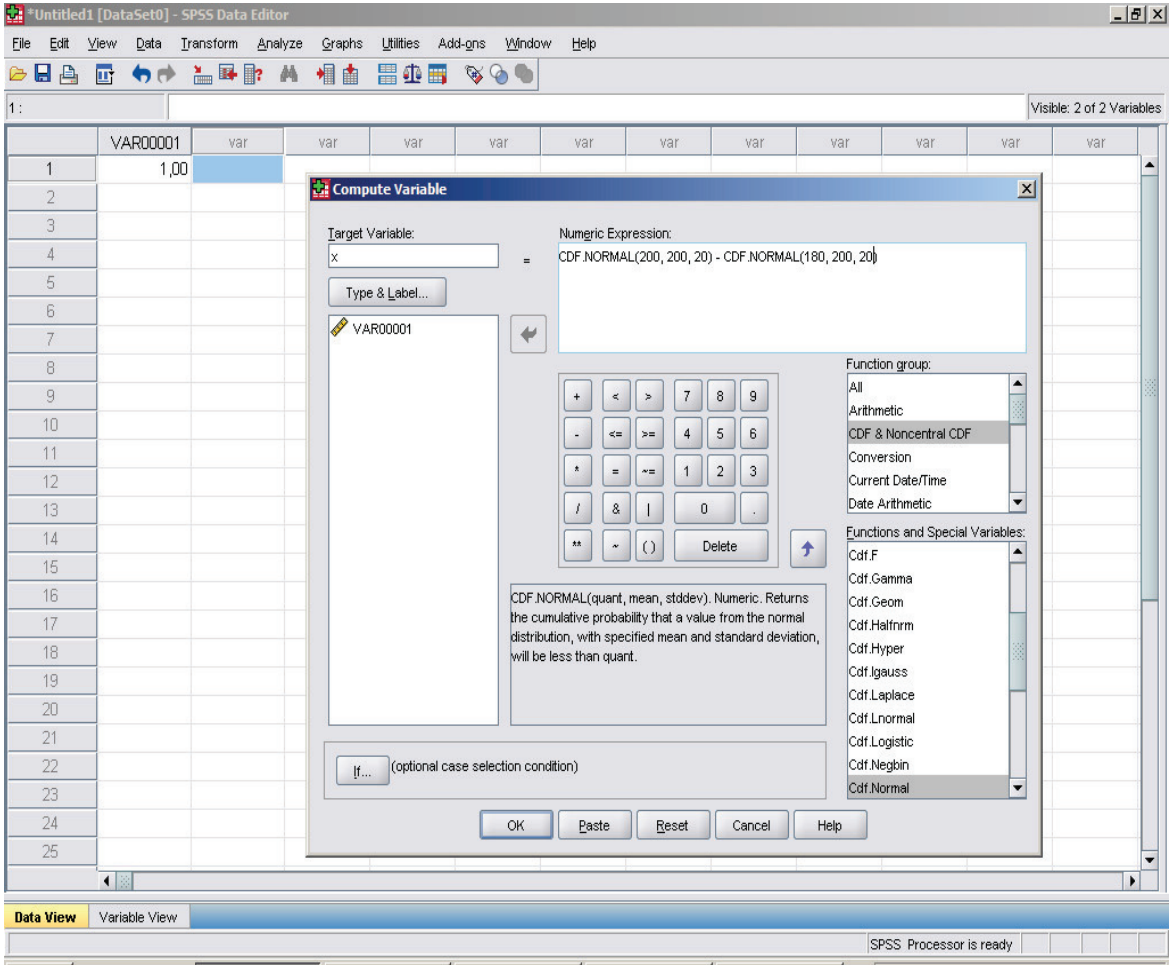

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,0062
Data View
Variable View
SPSS Processor is ready

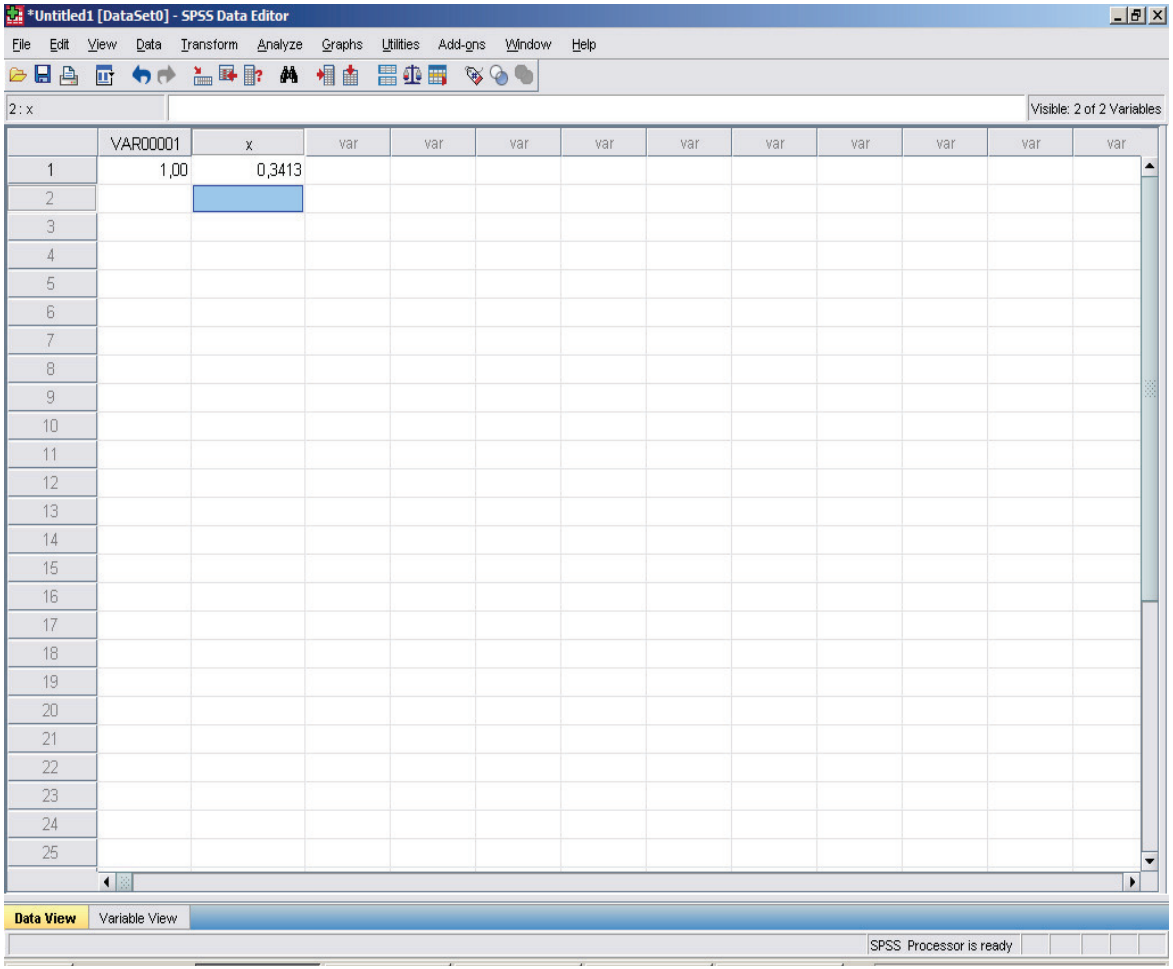
(a.2)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
1-CDF.NORMAL(225, 200, 20)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

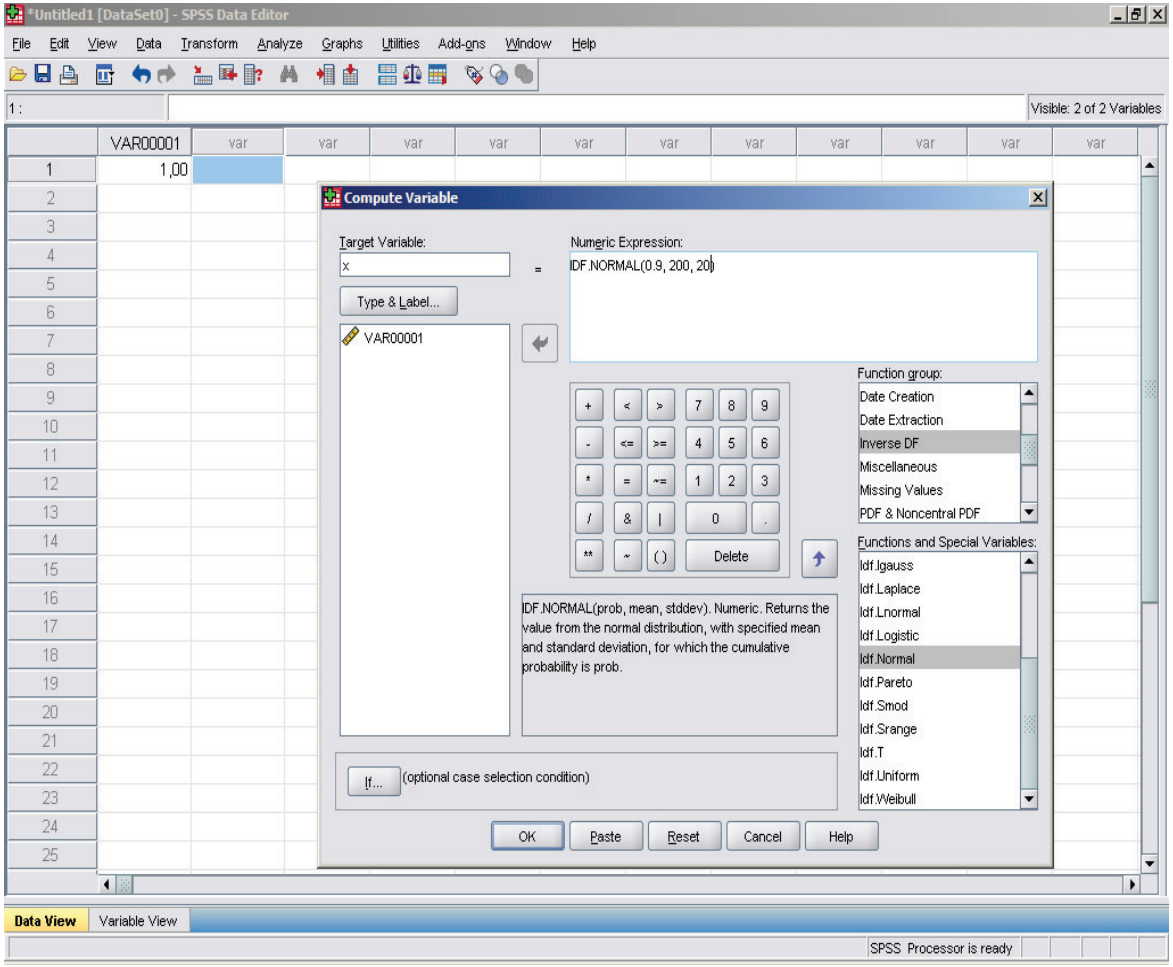

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,1056
Data View
Variable View
SPSS Processor is ready

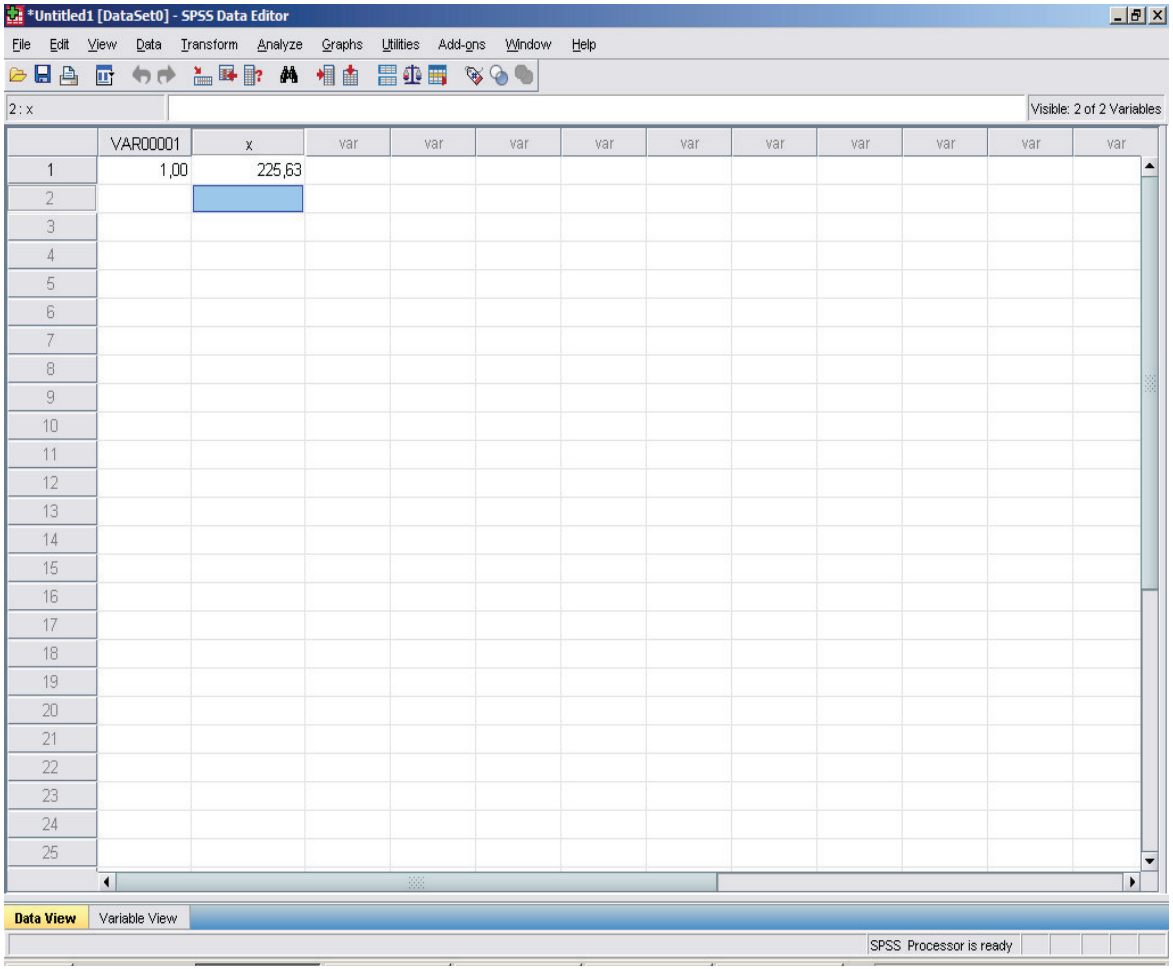
(a.3)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
CDF.NORMAL(200, 200, 20) - CDF.NORMAL(180, 200, 20)
Delete
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

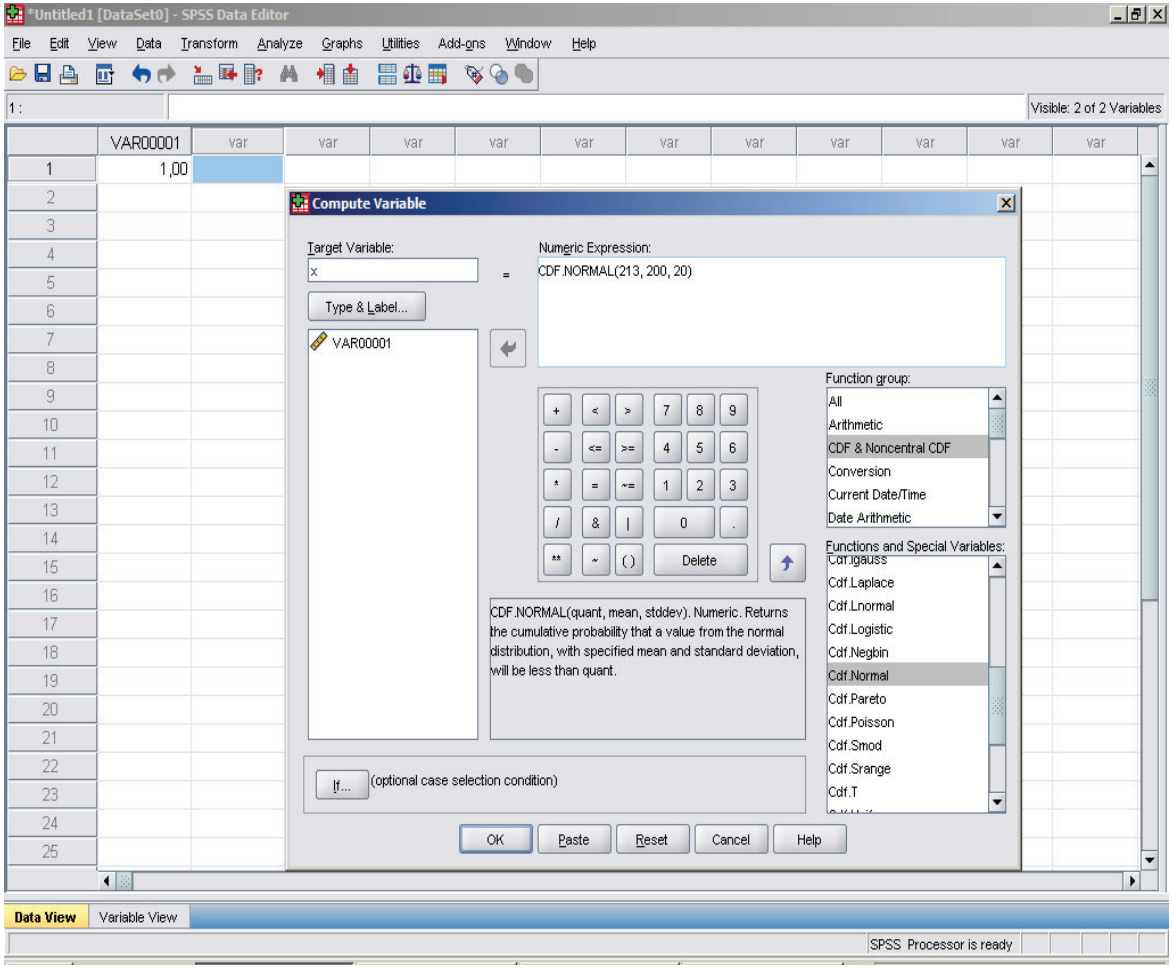

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,3413
Data View
Variable View
SPSS Processor is ready

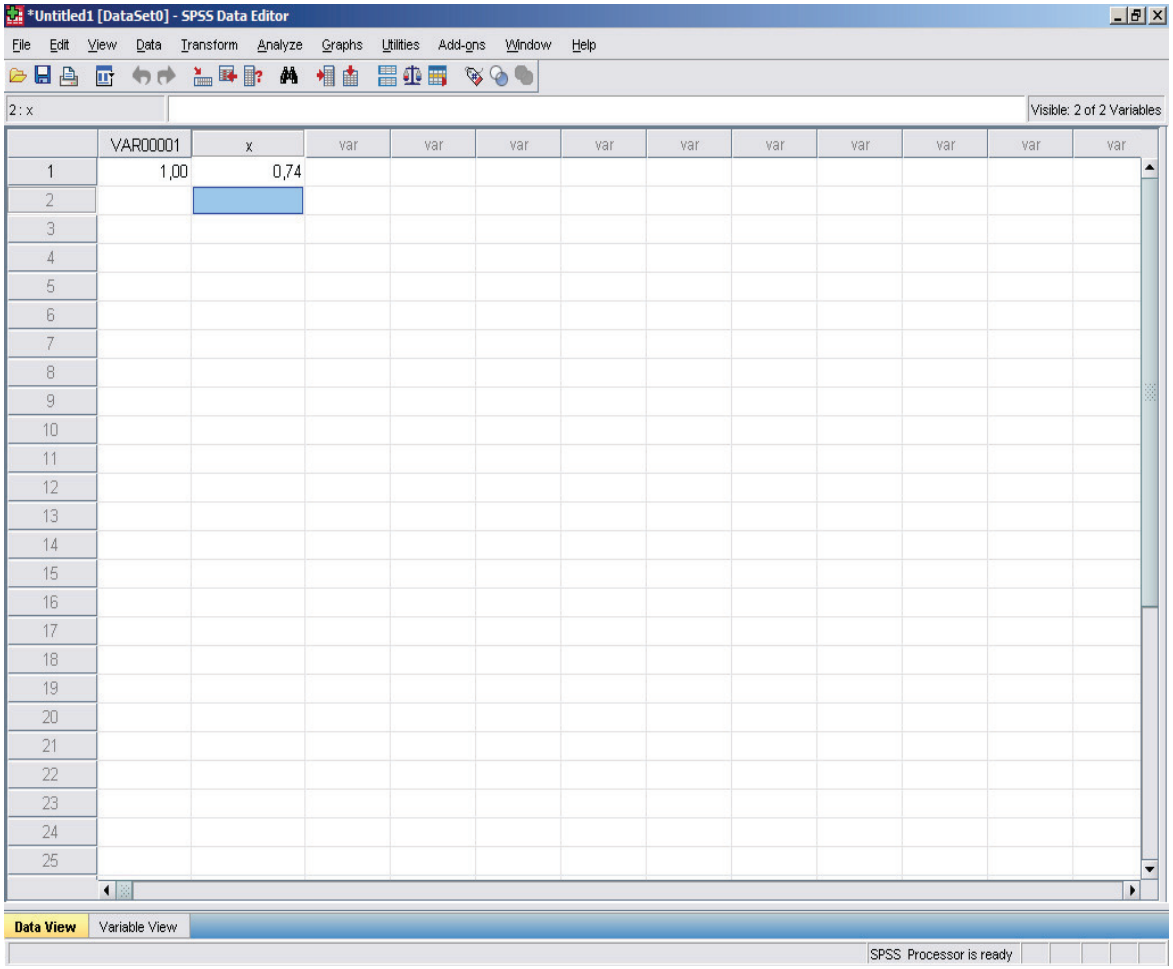
(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
IDF.NORMAL(0.9, 200, 20)
Type & Label...
VAR00001
Function group:
Date Creation
Date Extraction
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Functions and Special Variables:
Idf.Igauss
Idf.Laplace
Idf.Lnormal
Idf.Logistic
Idf.Normal
Idf.Pareto
Idf.Smod
Idf.Srange
Idf.T
Idf.Uniform
Idf.Weibull
Delete
IDF.NORMAL(prob, mean, stddev). Numeric. Returns the value from the normal distribution, with specified mean and standard deviation, for which the cumulative probability is prob.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
225,63
Data View
Variable View
SPSS Processor is ready

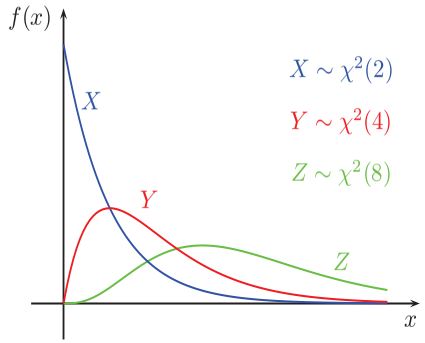
(c)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Type & Label...
VAR00001
Numeric Expression:
CDF.NORMAL(213, 200, 20)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Cdf.Srange
Cdf.T
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,74
Data View
Variable View
SPSS Processor is ready

2.1.7. Distribuição χ^2 de Pearson: $\chi^2(\nu)$

Função dens. prob.	$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} e^{-x/2} x^{(\nu/2)-1}$ $(\nu > 0)$
Domínio	$0 < x < +\infty$
Média	ν
Desvio padrão	$\sqrt{2\nu}$



onde

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

é designada por *função gama*. O parâmetro ν diz-se o **número de graus de liberdade**.

Teorema 2.1.3

(a) $Z_1, \ldots, Z_\nu$ *v.a. independentes,*
$Z_j \sim N(0,1), \quad j = 1, \ldots, \nu$
$\Longrightarrow Z_1^2 + \cdots + Z_\nu^2 \sim \chi^2(\nu)$ [a]

(b) $U_1, \ldots, U_k$ *v.a. independentes,*
$U_j \sim \chi^2(\nu_j), \quad j = 1, \ldots, \nu$
$\Longrightarrow U_1 + \cdots + U_k \sim \chi^2(\nu_1 + \cdots + \nu_k)$

(c) U_1, U_2 *v.a. independentes*
$U_1 + U_2 \sim \chi^2(\nu), \quad U_1 \sim \chi^2(\nu_1), \quad \nu > \nu_1$
$\Longrightarrow U_2 \sim \chi^2(\nu - \nu_1)$.

[a] Este resultado pode ser usado como definição da distribuição $\chi^2(\nu)$.

Uso mais comum: A distribuição do χ^2 é usada para aproximar as distribuições de estatísticas amostrais necessárias à realização de vários testes de hipóteses.

Aproximação: para ν suficientemente grande ($\nu \geq 60$), a distribuição $\chi^2(\nu)$ é aproximada pela distribuição normal $N(\nu, 2\nu)$:

$$\nu \geq 60 \implies \chi^2(\nu) \approx N(\nu, 2\nu).$$

Exemplos: Sejam X e Y duas variáveis aleatórias independentes seguindo a distribuição normal reduzida, $N(0,1)$, e Z a variável aleatória definida por

$$Z = X^2 + Y^2.$$

Determine:

(a) $P(X < -2) + P(X > 2)$

(b) $P(X^2 > 4)$ e justifique que este valor é igual ao obtido em (a)

(c) $P(Z \geq 0.5)$

(d) o 0.10-quantil de Z.

Resolução:

(a) Usamos as instruções em SPSS correspondentes à distribuição normal $N(0,1)$. Obtem-se (ver mais à frente) o valor 0.046.

Uma resolução alternativa consistiria na observação de que $P(X < -2) = P(X > 2)$, pela simetria da distribuição, e que portanto

$$P(X < -2) + P(X > 2) = 2\,P(X < -2).$$

(b) Por (a) do teorema anterior,

$$X^2 \sim \chi^2(1).$$

A instrução correspondente para a função distribuição em SPSS dá (mais à frente) 0.046.

O valor é igual ao obtido na alínea anterior porque

$$X^2 > 4 \iff X < -2 \ \vee \ X > 2$$

Basta agora considerar as distribuições adequadas a cada uma das v.a. X^2 e X.

(c) Por (a) do teorema anterior,

$$Z = X^2 + Y^2 \sim \chi^2(2).$$

Obtem-se (porque a v.a. é contínua)

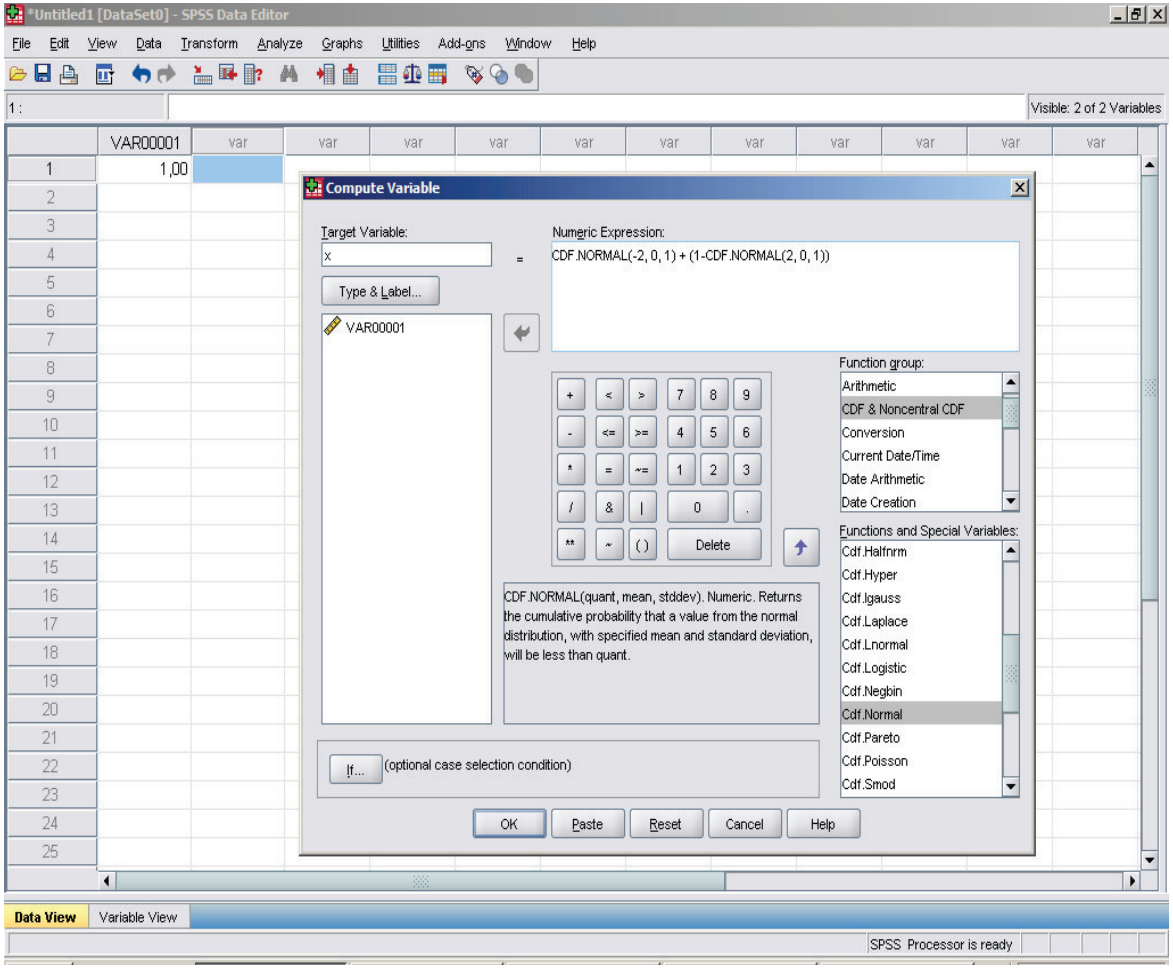
$$P(Z \geq 0.5) = 1 - P(Z \leq 0.5) = 0.779.$$

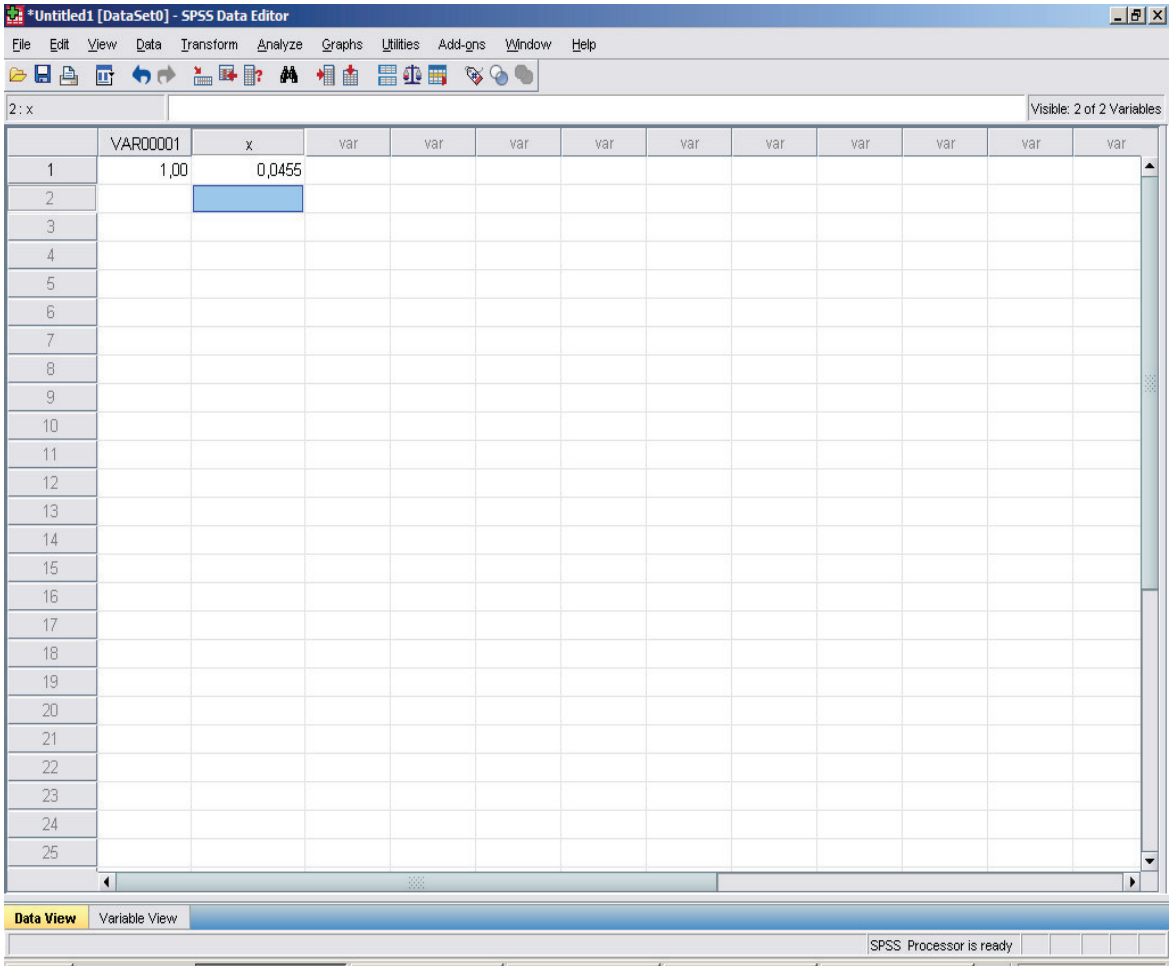
(d) O 0.10-quantil corresponde ao valor z para o qual

$$P(Z \leq z) = 0.10.$$

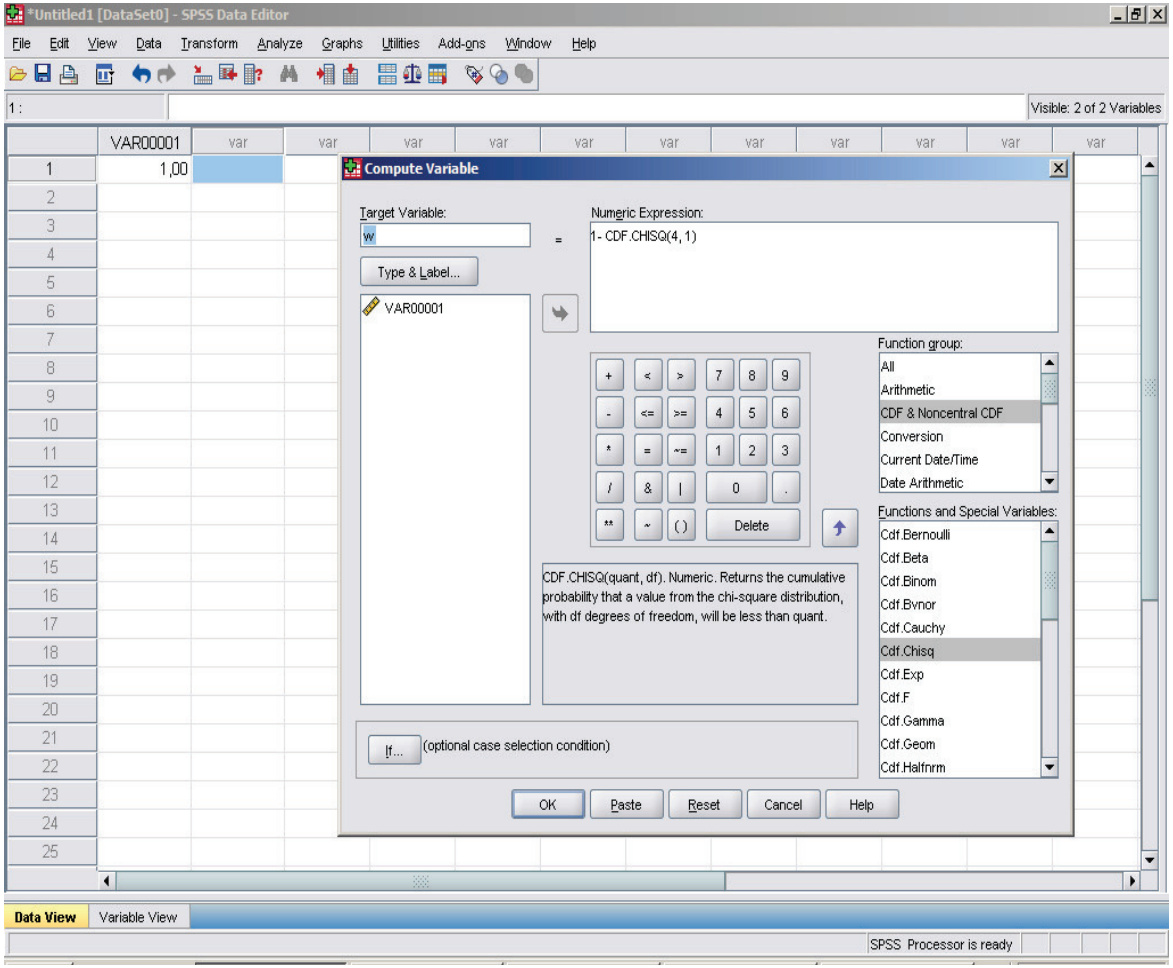
Pela instrução em SPSS correspondente, usando a função inversa da função de distribuição, obtem-se o valor aproximado de 0.211.

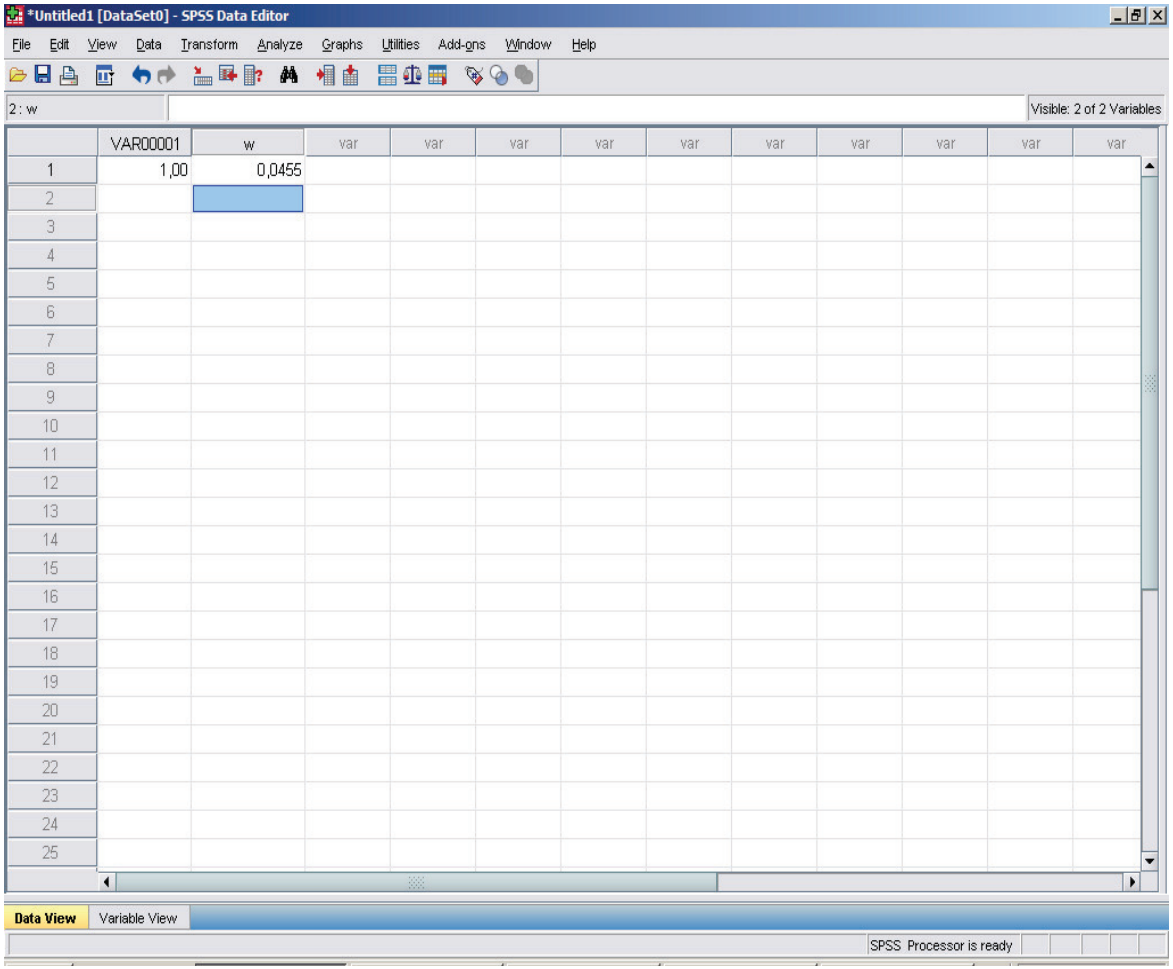
(a)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
x
Numeric Expression:
CDF.NORMAL(-2, 0, 1) + (1-CDF.NORMAL(2, 0, 1))
Type & Label...
VAR00001
Function group:
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Date Creation
Functions and Special Variables:
Cdf.Halfnrm
Cdf.Hyper
Cdf.Igauss
Cdf.Laplace
Cdf.Lnormal
Cdf.Logistic
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Delete
CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

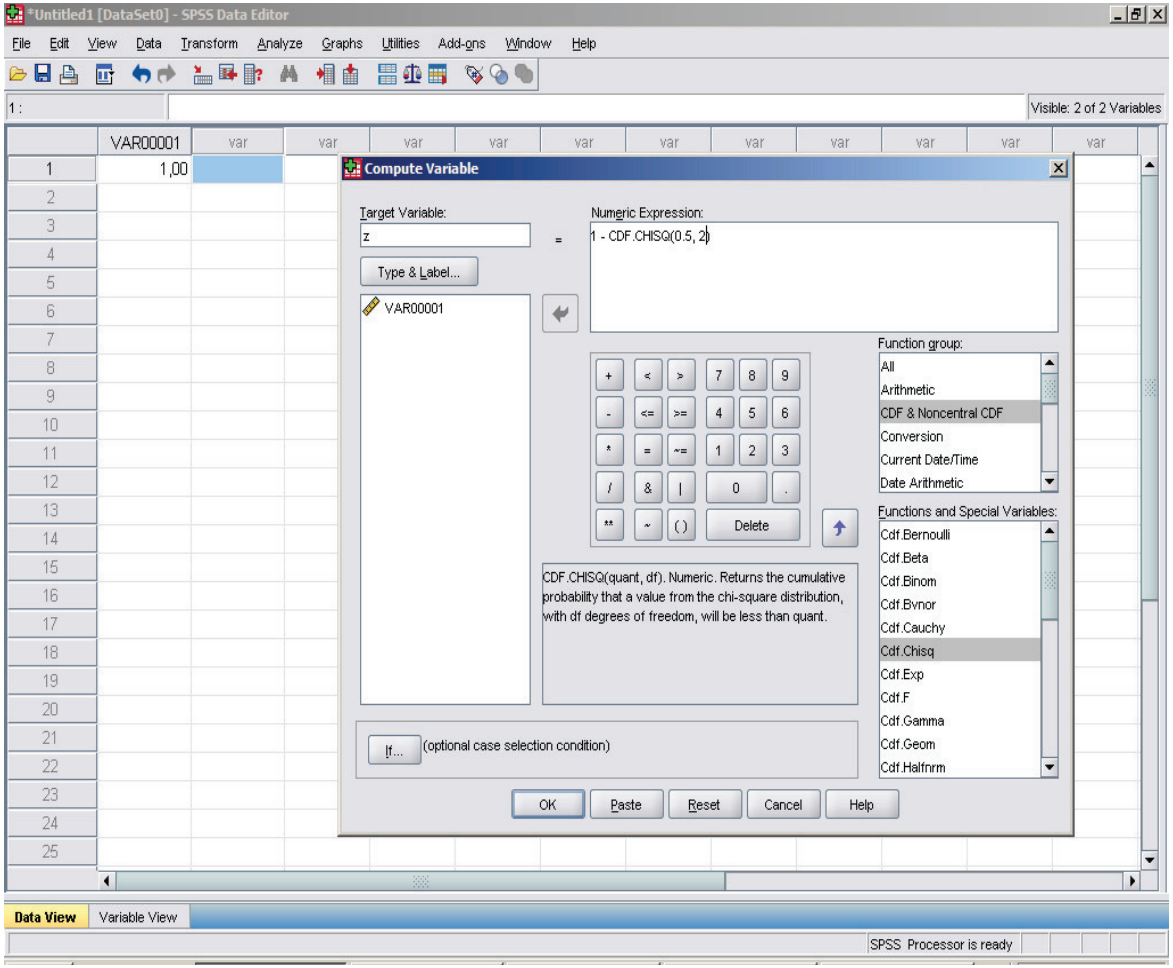

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : x
Visible: 2 of 2 Variables
VAR00001
x
1,00
0,0455
Data View
Variable View
SPSS Processor is ready

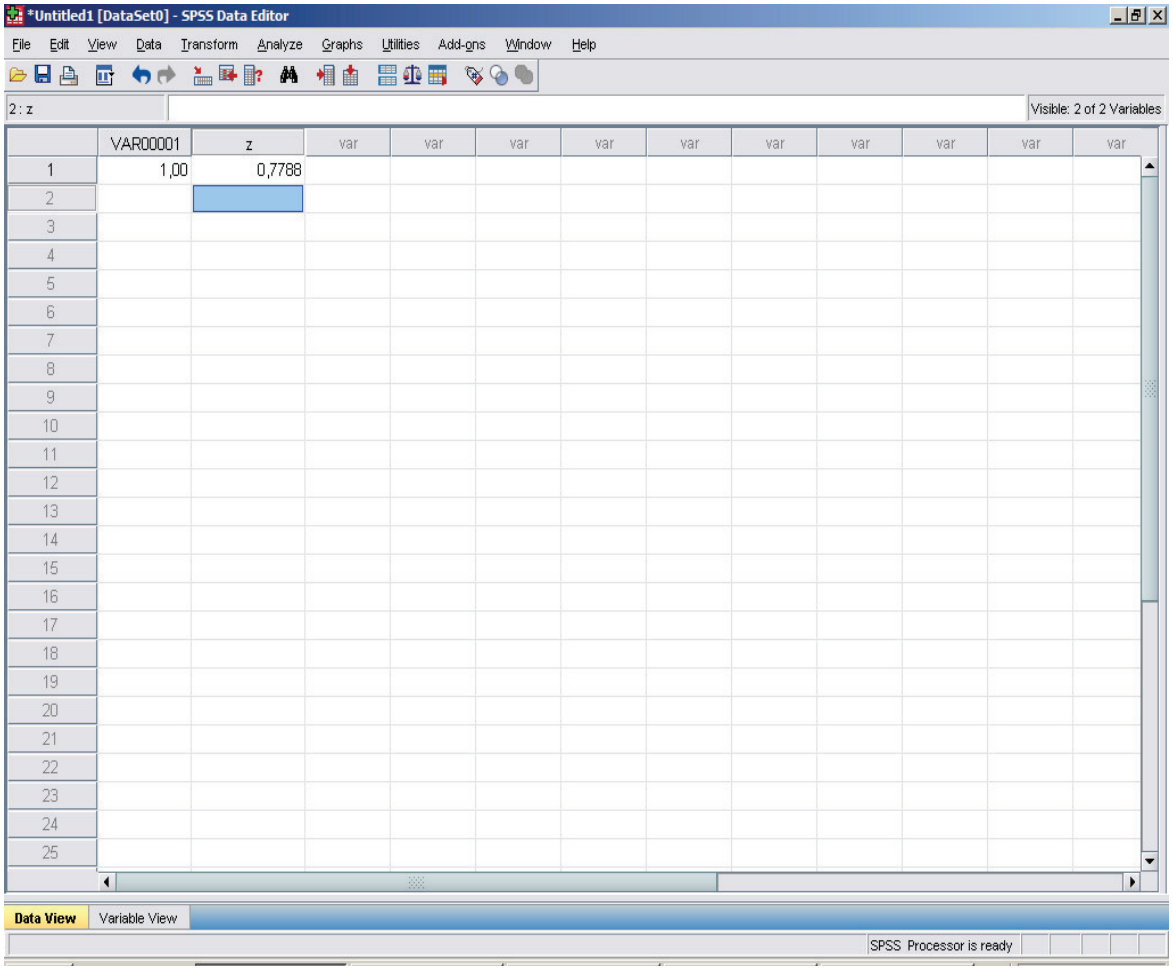
(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
w
Numeric Expression:
1- CDF.CHISQ(4, 1)
Type & Label...
VAR00001
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Bernoulli
Cdf.Beta
Cdf.Binom
Cdf.Bvnor
Cdf.Cauchy
Cdf.Chisq
Cdf.Exp
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Delete
CDF.CHISQ(quant, df). Numeric. Returns the cumulative probability that a value from the chi-square distribution, with df degrees of freedom, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

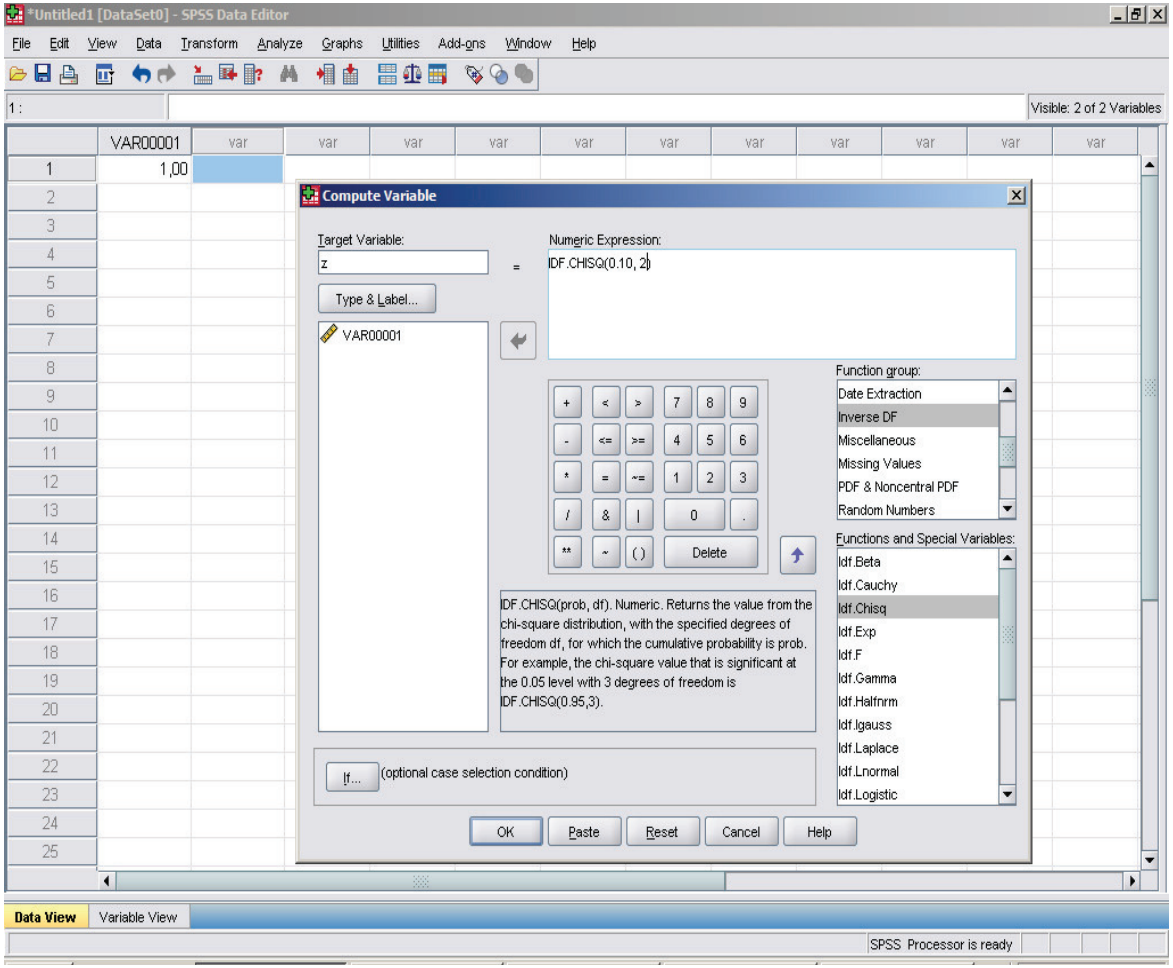

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : w
Visible: 2 of 2 Variables
VAR00001
w
1,00
0,0455
Data View
Variable View
SPSS Processor is ready

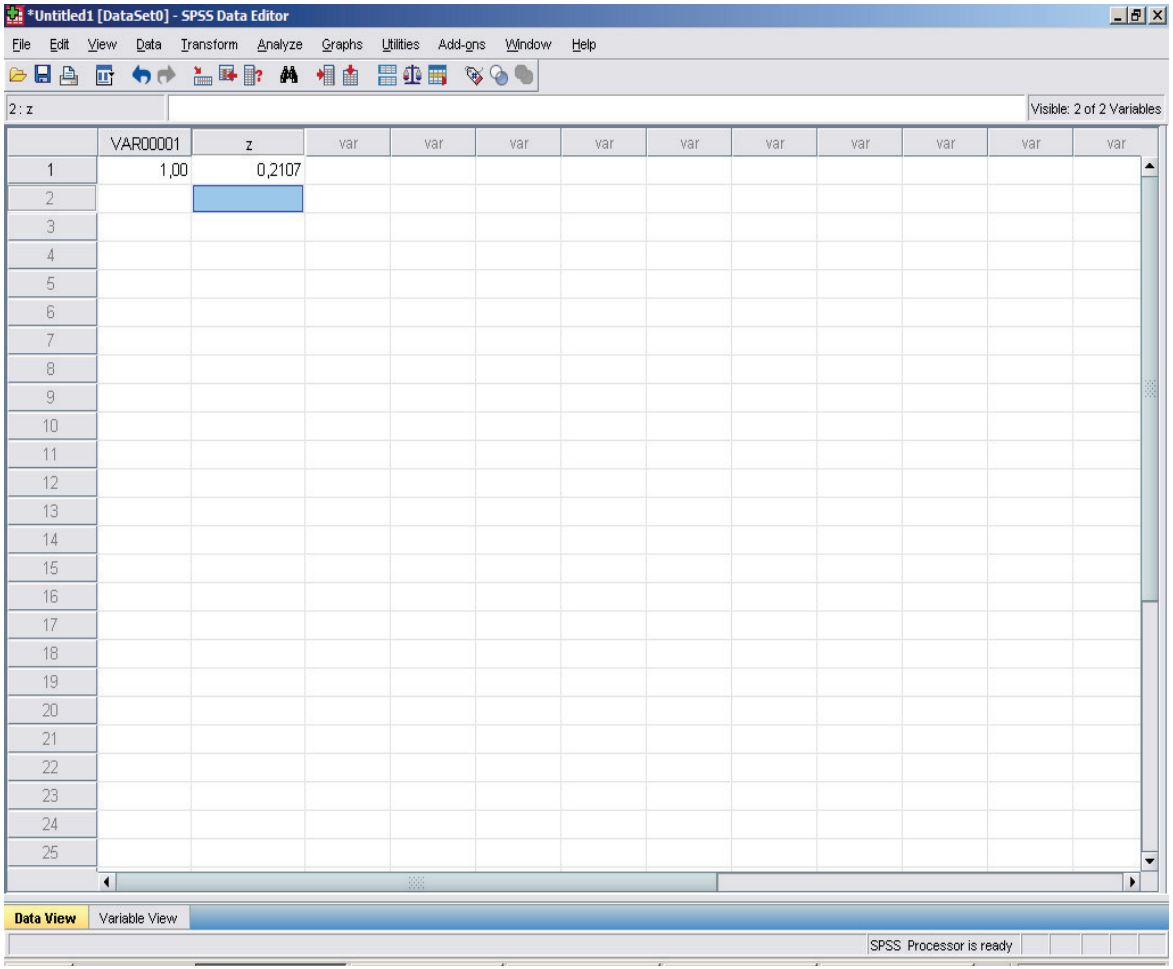
(c)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
z
Type & Label...
VAR00001
Numeric Expression:
1 - CDF.CHISQ(0.5, 2)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Bernoulli
Cdf.Beta
Cdf.Binom
Cdf.Bvnor
Cdf.Cauchy
Cdf.Chisq
Cdf.Exp
Cdf.F
Cdf.Gamma
Cdf.Geom
Cdf.Halfnrm
Delete
CDF.CHISQ(quant, df). Numeric. Returns the cumulative probability that a value from the chi-square distribution, with df degrees of freedom, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : z
Visible: 2 of 2 Variables
VAR00001
z
1,00
0,7788
Data View
Variable View
SPSS Processor is ready

(d)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
z
Numeric Expression:
IDF.CHISQ(0.10, 2)
Type & Label...
VAR00001
Function group:
Date Extraction
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Random Numbers
Functions and Special Variables:
Idf.Beta
Idf.Cauchy
Idf.Chisq
Idf.Exp
Idf.F
Idf.Gamma
Idf.Halfnrm
Idf.Igauss
Idf.Laplace
Idf.Lnormal
Idf.Logistic
Delete
IDF.CHISQ(prob, df). Numeric. Returns the value from the chi-square distribution, with the specified degrees of freedom df, for which the cumulative probability is prob. For example, the chi-square value that is significant at the 0.05 level with 3 degrees of freedom is IDF.CHISQ(0.95,3).
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : z
Visible: 2 of 2 Variables
VAR00001
z
1,00
0,2107
Data View
Variable View
SPSS Processor is ready

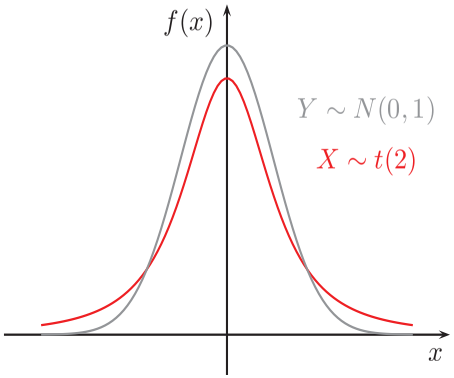
2.1.8. Distribuição t de Student: $t(\nu)$

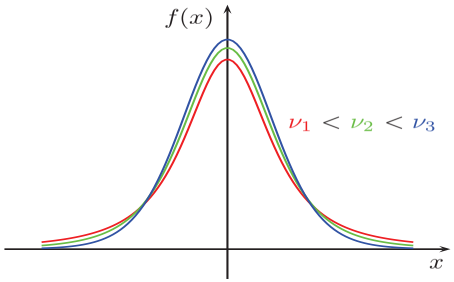
Função dens. prob.	$t(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$
Média	$\mu = 0 \qquad (\nu > 1)$
Desvio Padrão	$\sigma = \sqrt{\frac{\nu}{\nu-2}} \qquad (\nu > 2)$

onde

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$$

é a *função gama* e ν é chamado o **número de graus de liberdade**.
A distribuição t é simétrica em relação à média (zero). Graficamente é bastante parecida com a distribuição normal: apresenta contudo caudas mais altas e um valor máximo inferior e menos arredondado (mais "afiado") que o da distribuição normal.

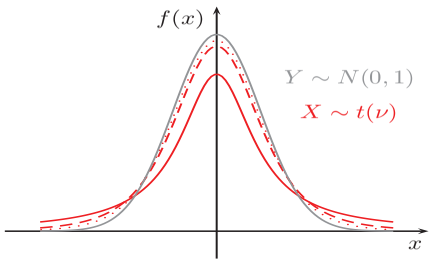




Teorema 2.1.4

(a) *Z, U v.a. independentes,*
$Z \sim N(0,1), \quad U \sim \chi^2(\nu)$
$\Longrightarrow T = \frac{Z}{\sqrt{U/\nu}} \sim t(\nu).$ [a]

(b) $\nu > 60 \Longrightarrow t(\nu) \approx N(0, \frac{\nu}{\nu-2}) \quad (\approx N(0,1))$



Uso mais comum: A distribuição t é usada para aproximar as distribuições de estatísticas amostrais necessárias à realização de vários testes de hipóteses.

[a] Este resultado pode ser usado como definição da distribuição $t(\nu)$.

Exemplo 1: Seja X uma variável aleatória que segue uma distribuição t de Student com 15 graus de liberdade. Determine:

(a) $P(X \leq 0.975)$

(b) o 0.975-quantil.

(Observe que as alíneas (a) e (b) referem-se a valores diferentes).

Resolução: Por hipótese, $X \sim t(15)$.

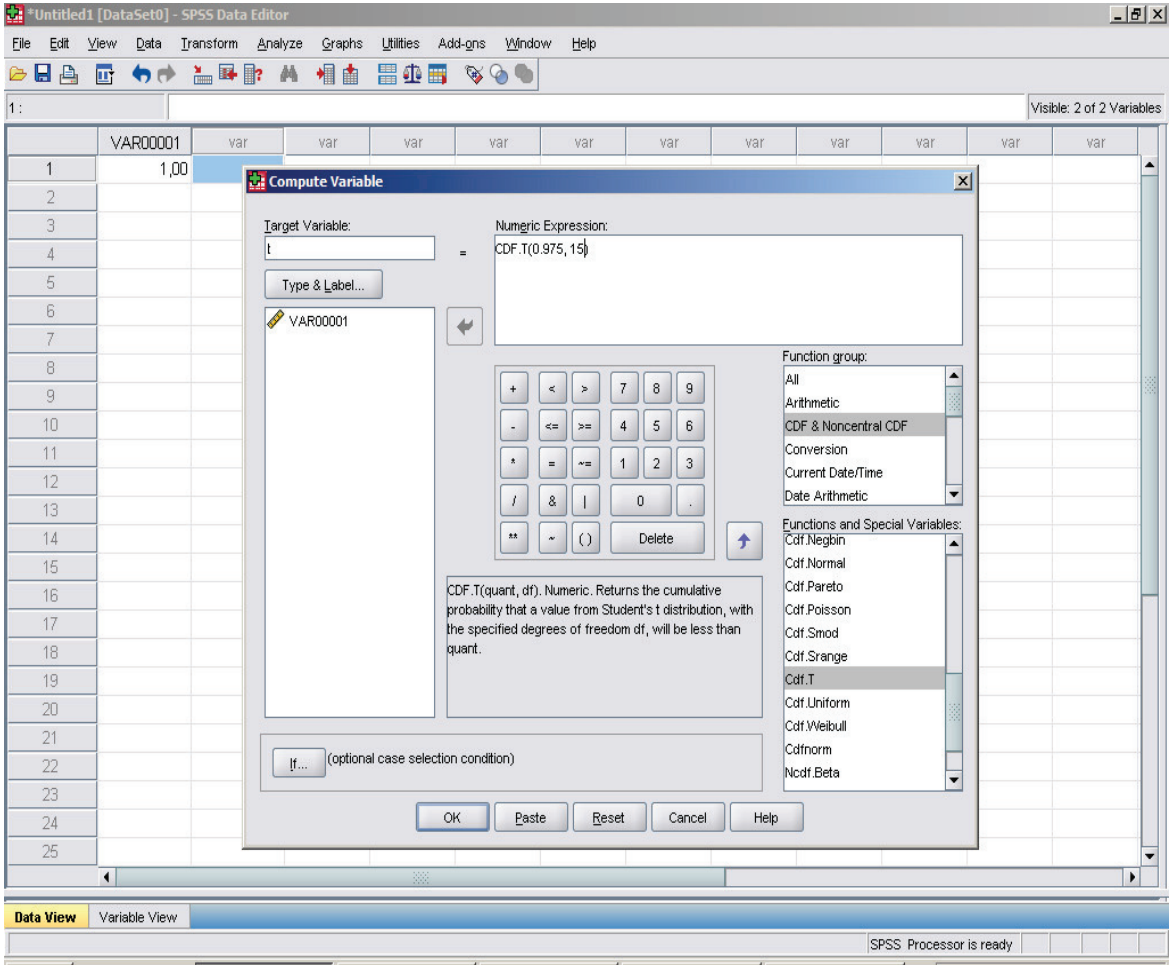
(a) $P(X \leq 0.975)$ corresponde ao valor da função de distribuição em $X = 0.975$. A instruçao correspondente em SPSS (mais à frente) fornece o valor 0.828.

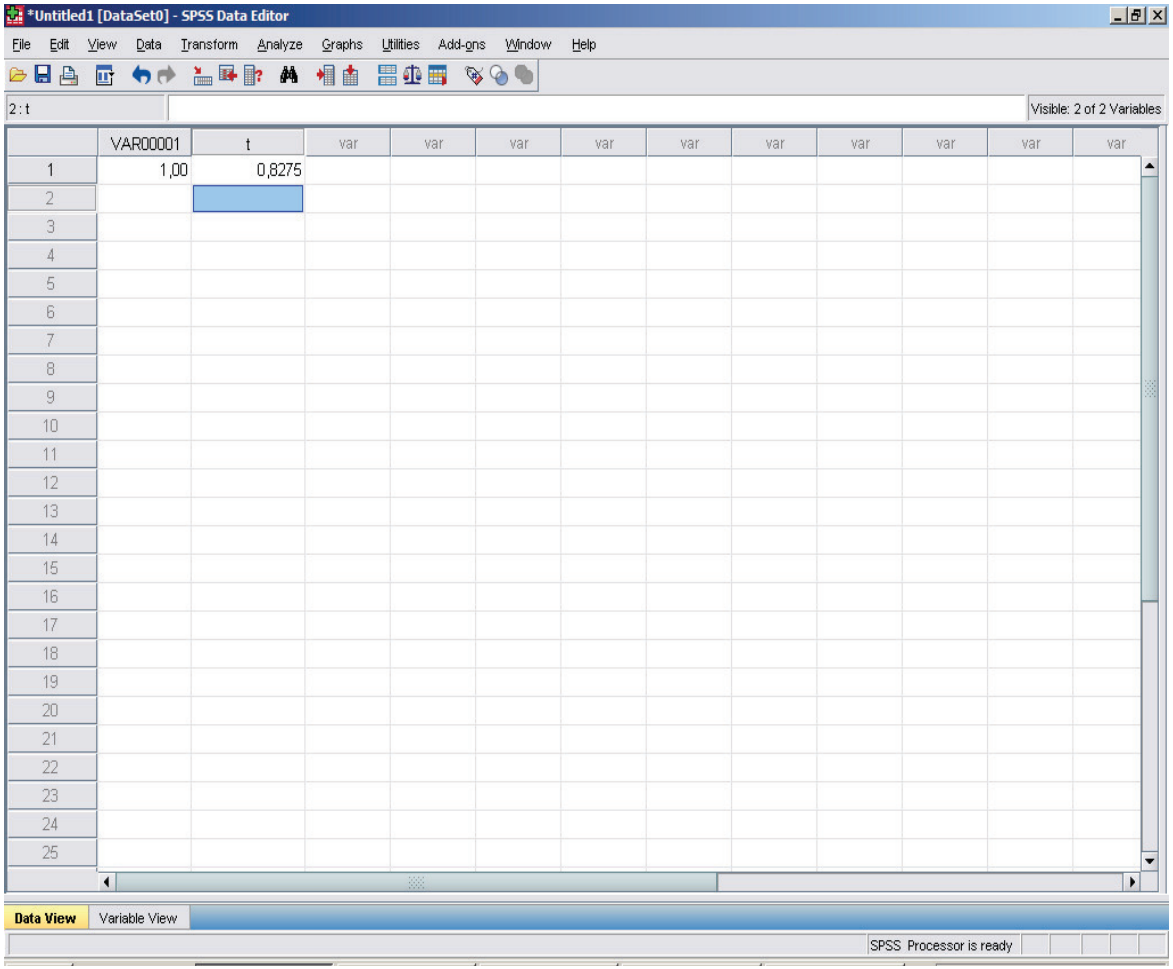
(b) Pretende-se determinar x tal que

$$P(X \leq x) = 0.975.$$

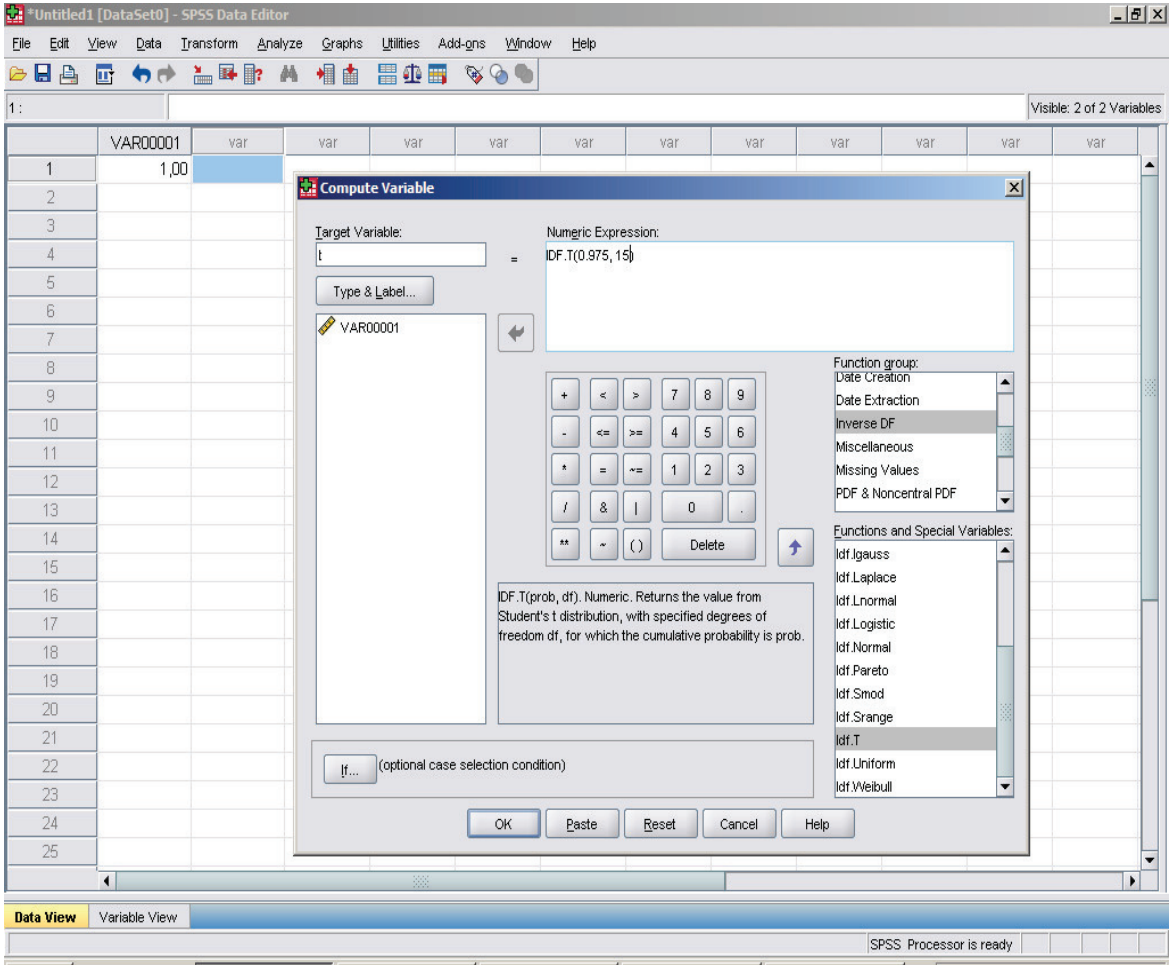
A instrução em SPSS (mais à frente) usa o inverso da função de distribuição. Obtém-se o valor $x \approx 2.131$.

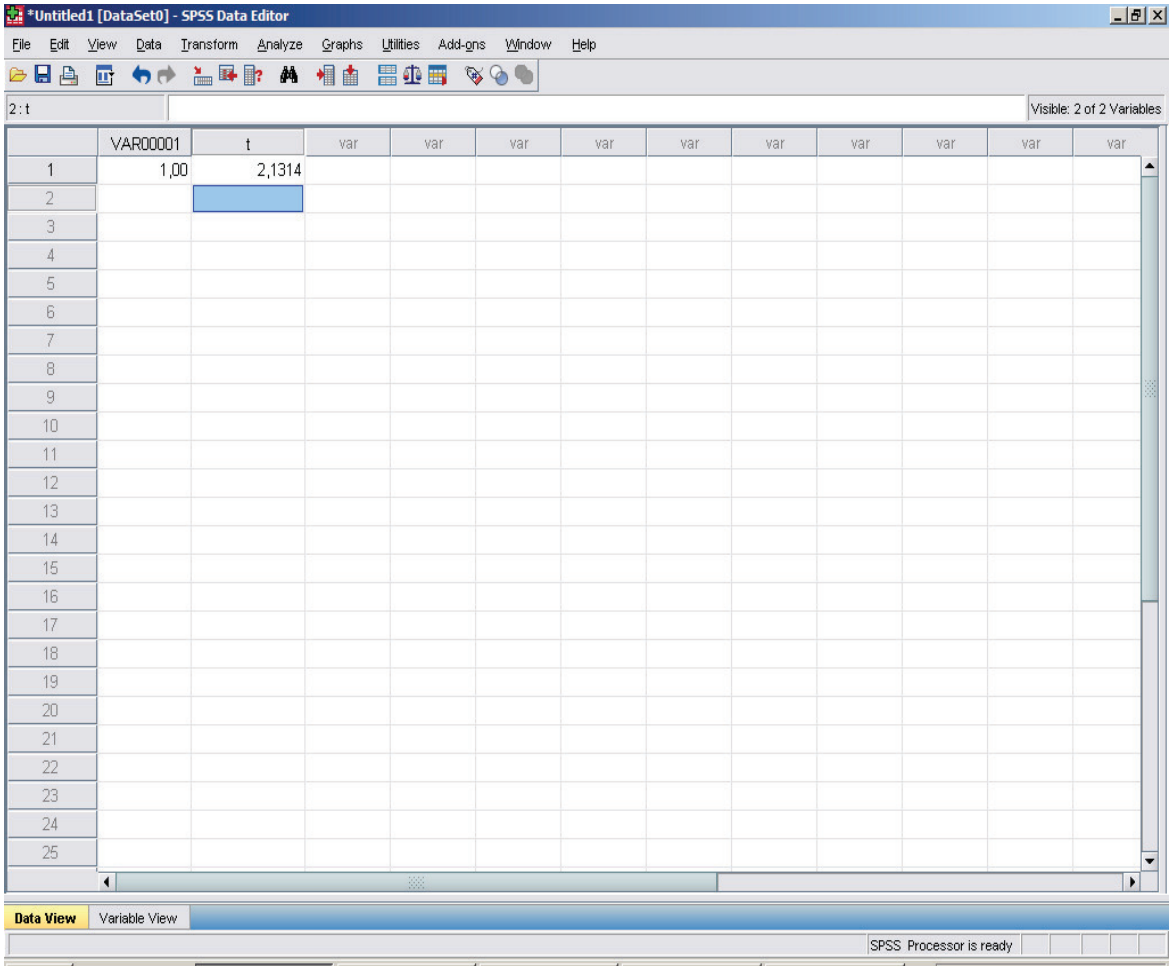
(a)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
t
Type & Label...
VAR00001
Numeric Expression:
CDF.T(0.975, 15)
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Functions and Special Variables:
Cdf.Negbin
Cdf.Normal
Cdf.Pareto
Cdf.Poisson
Cdf.Smod
Cdf.Srange
Cdf.T
Cdf.Uniform
Cdf.Weibull
Cdfnorm
Ncdf.Beta
Delete
CDF.T(quant, df). Numeric. Returns the cumulative probability that a value from Student's t distribution, with the specified degrees of freedom df, will be less than quant.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : t
Visible: 2 of 2 Variables
VAR00001
t
1,00
0,8275
Data View
Variable View
SPSS Processor is ready

(b)


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
VAR00001
1,00
Compute Variable
Target Variable:
t
Type & Label...
VAR00001
Numeric Expression:
IDF.T(0.975, 15)
Function group:
Date Creation
Date Extraction
Inverse DF
Miscellaneous
Missing Values
PDF & Noncentral PDF
Functions and Special Variables:
Idf.Igauss
Idf.Laplace
Idf.Lnormal
Idf.Logistic
Idf.Normal
Idf.Pareto
Idf.Smod
Idf.Srange
Idf.T
Idf.Uniform
Idf.Weibull
IDF.T(prob, df). Numeric. Returns the value from Student's t distribution, with specified degrees of freedom df, for which the cumulative probability is prob.
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
2 : t
Visible: 2 of 2 Variables
VAR00001
t
1,00
2,1314
Data View
Variable View
SPSS Processor is ready

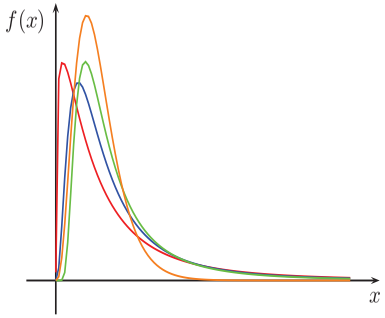
2.1.9. Distribuição F de Snedecor: $F(\nu_1, \nu_2)$

Função	$f(x)$
dens. prob.	$= \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{(\nu_1/2)-1} (\nu_2 + \nu_1 x)^{-(\nu_1+\nu_2)/2}$
Domínio	$0 < x < +\infty$
Média	$\mu = \frac{\nu_2}{\nu_2 - 2} \quad (\nu_2 > 2)$
Variância	$\sigma^2 = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2} \quad (\nu_2 > 4)$
Moda	$\left(\frac{\nu_1-2}{\nu_1}\right)\left(\frac{\nu_2}{\nu_2+2}\right) \quad (\nu_1 > 2)$

onde

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

é a *função gama*. Os parâmetros (ν_1, ν_2) dizem-se os **graus de liberdade** da distribuição F. Independentemente dos valores de (ν_1, ν_2), a distribuição é assimétrica à direita.



Teorema 2.1.5

(a) *U_1, U_2 v.a. independentes*
$U_1 \sim \chi^2(\nu_1)$, $U_2 \sim \chi^2(\nu_2)$

$$\Longrightarrow U = \frac{U_1/\nu_1}{U_2/\nu_2} \sim F(\nu_1, \nu_2)^a$$

(b) $U \sim F(\nu_1, \nu_2) \Longrightarrow \frac{1}{U} \sim F(\nu_2, \nu_1)$

(c) *Relação entre a distribuição-t e a distribuição-F:*

$$Y \sim t(\nu) \Longrightarrow Y^2 \sim F(1, \nu).$$

Uso mais comum: A distribuição F é usada nos testes de hipóteses sobre variâncias de populações normais e, de um modo geral, nos métodos de análise da variância (ANOVA) e regressão.

As instruções em **SPSS** para obtenção dos valores da função densidade de probabilidade, função de distribuição e quantis são análogas às de uma qualquer distribuição contínua já estudada.

[a] Este resultado pode ser usado como definição da distribuição $\chi^2(\nu)$.

Demonstração:

(b) Escreva-se

$$U = \frac{U_1/\nu_1}{U_2/\nu_2}$$

para variáveis aleatórias independentes $U_1 \sim \chi^2(\nu_1)$ e $U_2 \sim \chi^2(\nu_2)$. Tem-se então

$$U^{-1} = \frac{U_2/\nu_2}{U_1/\nu_1}$$

pelo que, pela definição de distribuição F, $U^{-1} \sim F(\nu_2, \nu_1)$.

(c) Por definição da distribuição $t(\nu)$, Escreva-se a v.a. Y na forma

$$Y = \frac{Z}{\sqrt{U/\nu}}$$

para v.a. independentes $Z \sim N(0,1)$ e $U \sim \chi^2(\nu)$. Então

$$Y^2 = \frac{Z^2}{U/\nu} = \frac{Z^2/1}{U/\nu}.$$

Como $Z^2 \sim \chi^2(1)$, $U \sim \chi^2(\nu)$, e Z^2 e U são independentes, então

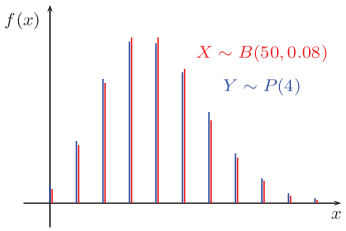
$$Y^2 \sim F(1, \nu).$$

2.2. Relações entre algumas distribuições - teoremas de aproximação

- **distribuição binomial e de Poisson**:

  $p \approx 0$ (sucesso raro) e n grande
  $n \geq 30$ e $np(1-p) \leq 5$

$$B(n,p) \approx P(np)$$

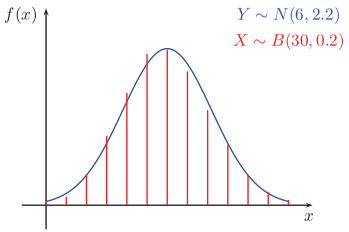


- **distribuição binomial e normal**:

  p e q não muito próximos de zero e n grande
  $n \geq 30$ e $np(1-p) > 5$

$$B(n,p) \approx N(np, np(1-p))$$

i.e., a distribuição binomial tende **assimptoticamente** para a normal quando $n \to +\infty$.

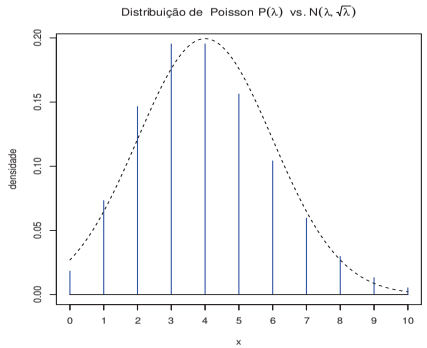


- **distribuição de Poisson e normal**:

  λ grande $\quad \lambda > 20$

$$P(\lambda) \approx N(\lambda, \lambda)$$

i.e., a distribuição de Poisson tende **assimptoticamente** para a normal quando $\lambda \to +\infty$.



- **distribuição t e normal**:

$$\nu \text{ grande} \Longrightarrow t(\nu) \approx N\left(0, \frac{\nu}{\nu - 2}\right)$$

i.e., a distribuição t tende **assimptoticamente** para a normal quando $\nu \to +\infty$.

- **distribuição χ^2 e normal**:

$$\nu \text{ grande} \Longrightarrow \chi^2(\nu) \approx N(\nu, 2\nu)$$

i.e., a distribuição χ^2 tende **assimptoticamente** para a normal quando $\nu \to +\infty$.

Observações: Nas aproximações de uma distribuição discreta por uma distribuição contínua (binomial por normal, ou poisson por normal) deve aplicar-se uma **correcção de continuidade**. Supostas satisfeitas as condições da aproximação:

- $P(X = x)$ quando X é uma v.a. discreta, pode ser calculado considerando

$$P\left(x - \frac{1}{2} \leq \tilde{X} \leq x + \frac{1}{2}\right)$$

onde $\tilde{X}$ é a v.a. contínua correspondente. A justificação de base para este raciocínio resulta de se ter

$$P(X = x) = P(x - 1 < X < x + 1).$$

- $P(a \leq X \leq b)$ quando X é uma v.a. discreta, deve ser calculado considerando

$$P\left(a - \frac{1}{2} \leq \tilde{X} \leq b + \frac{1}{2}\right)$$

onde $\tilde{X}$ é a v.a. contínua correspondente. Mais uma vez, isto está relacionado com o facto de se ter

$$P(a \leq X \leq b) = P(a - 1 < X < b + 1)$$

para a v.a. discreta X.

Há alguns anos atrás, quando se usavam tabelas com os valores das distribuições, as aproximações de distribuições referidas atrás tinham de ser consideradas, ou porque os valores da distribuição exacta não apareciam tabelados, ou porque os cálculos para a distribuição exacta eram demasiado morosos. Actualmente, contudo, já dispomos de computadores com uma larga capacidade de cálculo, o que faz com que as aproximações anteriores ou se tornem desnecessárias ou estejam já incluídas nas rotinas dos softwares que calculam as distribuições de probabilidade.

Folha de exercícios no. 2:

1. Suponha que se sabe que 30% de um determinada população é imune a uma doença. Qual é a probabilidade de se obterem exactamente 4 indivíduos imunes se se extrairem aleatoriamente 10 indivíduos da população?

2. Suponha que se sabe que 10% dos indivíduos de uma determinada população têm problemas renais e se extraem aleatoriamente 25 indivíduos dessa população. Determine a probabilidade de, entre os indivíduos escolhidos,

   (a) cinco ou menos terem problemas renais?

   (b) seis ou mais terem problemas renais?

   (c) entre seis e nove (inclusivé) terem problemas renais?

   (d) dois, três ou quatro terem problemas renais?

3. A probabilidade de um indivíduo com enxaqueca melhorar com um certo fármaco é de 0.9. Suponha que o fármaco foi administrado a treze pessoas com enxaqueca (escolhidas aleatoriamente).

   (a) Qual é a probabilidade do número de pessoas que melhoram da enxaqueca ser:

      (a1) exactamente um?

      (a2) superior a um?

      (a3) não superior a cinco?

      (a4) entre seis e onze, inclusivé?

      (a5) não inferior a dez?

      (a6) superior ao número de pessoas que não melhoram?

   (b) Suponha que este primeiro teste correu bem e que se decidiu administrar o medicamento a 60 pessoas com enxaqueca. Qual é a probabilidade de, nesse grupo de 60 pessoas, encontrar exactamente 6 indivíduos em que o fármaco não tenha surtido efeito? E qual é essa mesma probabilidade para mais de 6 indivíduos?

4. Um estudo realizado numa certa população mostrou que, em média, 13 novos casos de cancro do esófago são diagnosticados por ano. Se a incidência (número de novos casos) anual do cancro do esófago seguir uma distribuição de Poisson, determine a probabilidade de, num dado ano, o número de novos casos deste tipo de cancro ser:

   (a) exactamente 10

   (b) pelo menos 8

   (c) não mais de 12

   (d) inferior a 7

   (e) entre 9 e 15, inclusivé.

5. Suponha que durante um período de vários anos o número médio de mortes, por ano, devido a uma doença não contagiosa foi de 10, e que o número de mortes pela doença num qualquer mês do ano é independente do número de mortes num outro mês. Se o número de mortes por esta doença seguir uma distribuição de Poisson, qual é a probabilidade de:

   (a) durante o ano corrente,

      (a.1) não mais de sete pessoas morrerem dessa doença?

      (a.2) não existirem mortes devidas à doença?

   (b) se verificarem entre 10 e 20 mortes (exclusivé) durante os primeiros 3 meses do ano?

   (c) morrer exactamente uma pessoa no mês de Setembro?

6. Sabendo que $X \sim N(0,1)$ determine:

   (a) a probabilidade de um valor x escolhido aleatoriamente de entre todos os valores possíveis de X estar entre -2.87 e 2.64.

   (b) a área da região limitada pelo eixo dos xx's e o gráfico da função densidade de X entre 0 e 1.43

   (c) $P(X^2 + 1 > 2)$

   (d) $P(X = 0.74)$

   (e) x_1 tal que $P(X \geq x_1) = 0.95$

   (f) $x_2 > 0$ tal que $P(-x_2 < X \leq x_2) = 0.99$.

7. Num estudo realizado por Dusheiko (1975) a propósito da doença de Alzheimer, os dados obtidos fazem crer que o peso do cérebro de vítimas da doença segue uma distribuição normal com média 1076.8 gramas e desvio-padrão 105.76 gramas. Supondo que de facto o peso do cérebro de todas as vítimas de Alzheimer é uma v.a. que segue uma distribuição normal com essa média e desvio-padrão, determine:

   (a) a probabilidade de uma vítima escolhida ao acaso ter um cérebro que pesa:

      (a.1) mais de 800 g e menos de 900 g

      (a.2) 1000 g.

   (b) a percentagem de indivíduos na população em causa com peso do cérebro superior a 1000 g.

8. Suponha que o tempo médio de internamento no hospital devido a uma determinada doença crónica (para pacientes que não apresentem outros problemas físicos) é de 60 dias com um desvio padrão de 15 dias. Seleccionou-se aleatoriamente um indivíduo doente que não apresenta mais nenhuns sintomas para além dos da doença . Na hipótese da variável aleatória "tempo de internamento devido à doença" seguir uma distribuição normal,

   (a) determine a probabilidade do tempo de internamento desse indivíduo

      (a.1) ser superior a 50 dias

      (a.2) estar compreendido entre 30 e 60 dias

      (a.3) ser de 45 ou 46 dias.

(b) determine a proporção de pacientes com tempo de internamento superior a 90 dias.

9. Os pesos X de uma população de mulheres de uma comunidade índia com idades compreendidas entre os 18 e os 40 anos seguem, aproximadamente, uma distribuição normal de média 60 Kg e desvio-padrão 12 Kg.

   (a) Determine:

   (a.1) $X_{0.75}$

   (a.2) a mediana de X

   (a.3) x_1 tal que $P(X \geq x_1) = 0.25$

   (a.4) $x_2 > 0$ tal que $P(60 - x_2 < X < 60 + x_2) = 0.48$.

   (b) Determine a probabilidade de uma mulher seleccionada ao acaso de entre esta população pesar:

   (b.1) exactamente 60 Kg

   (b.2) mais de 70 Kg

   (b.3) 54 Kg ou menos

   (b.4) entre 55 e 65 Kg

   (b.5) no máximo 60 Kg.

   (c) Suponha que as alturas das mulheres dessa comunidade, Y, seguem também uma distribuição normal mas com média 158 cm e desvio-padrão 10 cm, e que X e Y são independentes. Sejam X_r e Y_r as variáveis aleatórias reduzidas de X e Y, respectivamente, e considere $Z = X_r^2 + Y_r^2$.

   (c.1) Qual é a probabilidade de Z tomar valores superiores a 4.6?

   (c.2) Determine $Z_{0.75}$.

   (c.3) Determine z para o qual a probabilidade da variável aleatória Z tomar valores maiores ou iguais a z é 0.975.

   (c.4) Os valores de $P(Z \leq 0.1)$ e de $Z_{0.1}$ coincidem?

10. Suponha que a probabilidade de um indivíduo de uma dada comunidade contrair uma determinada doença é 0.1 e que se extrai uma amostra de 10 indivíduos dessa comunidade.

    (a) Qual é a probabilidade de a amostra possuir exactamente 1 indivíduo com a doença?

    (b) Porque é que a probabilidade anterior é diferente de 0.1?

11. Sabe-se que, com determinado tratamento administrado a pacientes em condições bem definidas, se alcança 70% de cura para certa doença. Se o tratamento for aplicado a 50 pacientes em tais condições, qual é a probabilidade de:

    (a) obter no máximo 35 curas

    (b) obter 30 ou mais curas

    (c) obter um número de curas não inferior a 25 nem superior a 30.

12. Um estudo para avaliar o risco de cancro em pacientes com fibrose cística nos EUA e Canadá consistiu em analisar a incidência de cancro entre estes pacientes entre 1 Janeiro 1985 e 31 Dezembro 1992 e compará-la posteriormente com os valores esperados obtidos do programa SEER, do National Cancer Institute.

    (a) De entre os pacientes com fibrose cística considerados, eram esperados 45.6 casos de cancro e observaram-se 37. Que distribuição deve usar para modelar o número de casos de cancro ocorridos entre os pacientes com fibrose cística?

    (b) O número de cancros observado foi anormalmente baixo?

    (c) No mesmo estudo, foram observados 13 cancros do tracto digestivo enquanto que eram esperados apenas 2. O número de casos de cancro deste tipo entre os pacientes com fibrose cística foi excessivamente elevado?

13. O número de mortes por insuficiência cardíaca (IC) por ano num hospital pode ser modelado através de uma distribuição de Poisson com média igual a 22 mortes/ano. Suponha que em 2002 se observaram 35 mortes por IC num hospital. Esta ocorrência é pouco usual?

Chapter 3

Estatística Descritiva

3.1. Amostragem

Frequentemente precisamos de obter conclusões acerca de um grande grupo de indivíduos ou objectos (chamado **população**). Por escassez de tempo, custos elevados, ou outras razões, não é possível analisar todo o grupo e decidimos escolher apenas uma pequena parte (**amostra**) dessa população. Se for razoável supor que essa amostra representa convenientemente a população, o problema reside em extrair conclusões acerca das características desconhecidas da população com base nas informações observadas na amostra. O processo através do qual se obtêm conclusões acerca de uma população com base em resultados observados numa amostra dessa população é designado por **inferência estatística**. O processo de obtenção ou extração de amostras chama-se **amostragem**. Como a inferência de uma amostra em relação à população correspondente não pode ser considerada absolutamente certa, devemos utilizar a linguagem das probabilidades na formulação de qualquer conclusão.

Exemplos:

1. Podemos querer tirar conclusões acerca da percentagem de comprimidos defeituosos de um determinado medicamento produzido por um laboratório durante uma semana (cinco dias úteis) examinando apenas 50 comprimidos por dia, fabricados em diferentes períodos do dia. Neste caso a população consiste de todos os comprimidos fabricados pelo laboratório durante uma semana e a amostra consiste dos 250 comprimidos que foram sendo recolhidos durante essa semana.

2. Podemos querer tirar conclusões sobre os pesos de todos os estudantes inscritos na FCUP no ano lectivo de 2004-2005 (população) apenas baseados em dados recolhidos de 150 estudantes (amostra) inscritos nesse ano lectivo.

3. Podemos querer tirar conclusões acerca da qualidade de uma moeda (percentagem de caras e coroas) jogando-a repetidamente. A população (infinita) consiste de todas os lançamentos possíveis da moeda. Uma amostra consistiria, por exemplo, dos primeiros 50 lançamentos.

A população pode ser finita ou infinita. O **tamanho da população**, usualmente designado por N, é o número de elementos da população. Analogamente, o **tamanho da amostra**, usualmente designado por n, é o número de elementos da amostra.

A amostragem pode ser feita **com reposição** ou **sem reposição**. No primeiro caso, um elemento da população pode aparecer várias vezes numa amostra enquanto que na amostragem sem reposição essa situação é impossível.
Uma população finita submetida a um processo de amostragem com reposição pode, teoricamente, ser considerada como infinita pois podemos extrair amostras de qualquer tamanho sem nunca esgotarmos a população. Em termos práticos, a amostragem de uma população finita muito grande pode ser considerada como amostragem de uma população infinita.

Um dos problemas da teoria de amostragem consiste obviamente na escolha dos indivíduos que vão fazer parte da amostra. Nem todas as amostras são representativas da população e portanto nem todas as amostras podem ser usadas para inferir acerca da população. É preciso certificarmo-nos de que cada elemento da população (suposta finita) tem a mesma probabilidade de figurar na amostra. Quando esta condição é satisfeita diz-se que a amostra é uma **amostra aleatória**. Equivalentemente, uma amostra de tamanho n de uma população de tamanho N diz-se uma amostra aleatória se qualquer outra amostra possível de tamanho n tem a mesma probabilidade de ser seleccionada.

Como já foi referido atrás, as conclusões extraídas de uma amostra por inferência estatística não podem ser consideradas absolutamente certas pelo que devemos utilizar a linguagem das probabilidades na formulação de qualquer conclusão.
Diz-se então que uma população é considerada **conhecida** quando se conhece a função densidade de probabilidade f da variável aleatória X associada. No exemplo 2. anterior, se X é a v.a. que representa os pesos dos alunos da FCUP então dizemos que conhecemos a população de alunos se conhecermos a função densidade de probabilidade de X. Por comodidade, usaremos a terminologia das distribuições quer para as variáveis aleatórias quer para as populações correspondentes. Assim, no exemplo anterior, tanto diremos que a população dos alunos segue uma distribuição F como que a v.a. X representando o peso dos alunos segue uma distribuição F (sendo F uma distribuição qualquer).

Há certas quantidades que aparecem em $f(x)$, designadas por **parâmetros da distribuição**, que caracterizam completamente a distribuição. Exemplos de parâmetros são μ e σ^2 no caso de uma distribuição normal, n e p no caso de uma distribuição binomial ou λ no caso de uma distribuição de Poisson. De facto qualquer outra quantidade ou informação acerca dessas distribuições (por exemplo, a mediana, a moda, quantis, ...) pode ser obtida usando os parâmetros indicados.
Contudo, na maior parte das vezes, a função densidade não é conhecida com precisão mas, felizmente, tem-se uma ideia qualitativa do seu comportamento. Nesse caso pretende-se estimar os parâmetros da distribuição. Suponhamos, por exemplo, que a amostra dos alunos da FCUP nos faz pensar (embora não tenhamos toda a certeza) que a variável aleatória X que representa o peso dos alunos da FCUP segue uma distribuição normal. O passo seguinte deverá então consistir em **estimar a média e a variância de X (de toda a população) a partir da amostra** dos 150 alunos.

Para obter uma amostra de 150 alunos começamos por escolher aleatoriamente um aluno de entre toda a população (neste caso, todos os alunos inscritos na FCUP no ano lectivo de 2004-2005) e registamos o seu peso, digamos x_1. O valor de x_1 é aleatório pois é um de entre muitos possíveis. Consideremos então a v.a.

$$X_1$$

que representa o peso do primeiro aluno seleccionado da população de todos os alunos da FCUP inscritos em 2004-2005.

O valor x_1 é apenas uma **realização** de X_1. Se tivessemos escolhido um outro indivíduo teríamos obtido possivelmente um valor diferente de x_1.
Em seguida escolhemos, aleatoriamente e de entre a população, o segundo indivíduo e voltamos a registar o seu peso, digamos x_2. Mais uma vez, o valor de x_2 não é previsível e poderia ter sido um outro qualquer de entre muitos possíveis. Seja então

$$X_2$$

a v.a. que representa o peso do segundo aluno seleccionado da população . Tal como acima, x_2 é uma realização de X_2.
Iteramos este processo até chegarmos à escolha do indivíduo número 150. O peso desse indivíduo é x_{150} e corresponde à realização da v.a.

$$X_{150}$$

que representa o peso do aluno número 150 a ser seleccionado de toda a população.
Sem perda de generalidade, suponhamos que a amostragem foi feita com reposição (como o tamanho da amostra é muito menor do que o tamanho da população, a amostragem sem reposição dará sensivelmente os mesmos resultados que a amostragem com reposição).

Mais geralmente, suponhamos que amostramos n elementos de uma população de tamanho N. Para $i = 1, \ldots, n$, seja

$$X_i$$

a v.a. que representa o valor da característica que está a ser avaliada no elemento i. As v.a.

$$X_1, X_2, \ldots, X_n$$

são **independentes** e **identicamente distribuídas** (*i.e.*, qualquer X_i segue a mesma distribuição, digamos F). A sua função densidade de probabilidade conjunta é então

$$\begin{aligned} f_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) &= f_{X_1}(x_1)\ldots f_{X_n}(x_n) \\ &= f(x_1)\ldots f(x_n). \end{aligned}$$

Ao conjunto $\{X_1, X_2, \ldots, X_n\}$ chama-se **amostra aleatória** de tamanho n, pertencente à população com função distribuição F; ao conjunto dos valores observados na amostra

$$\{x_1, x_2, \ldots, x_n\}$$

chama-se uma **realização da amostra aleatória** (é o valor observado de $(X_1, \ldots, X_n)$ para aquela amostra em particular). Assim, dependendo da amostra aleatória escolhida, podemos observar diferentes realizações de $X_1, X_2, \ldots, X_n$:

$$x_1, x_2, \ldots, x_n, \quad x_1', x_2', \ldots, x_n', \quad x_1'', x_2'', \ldots, x_n'', \quad \ldots$$

3.2. Medidas de localização de uma a.a.

Dados: a.a. $x_1, x_2, \ldots, x_n$ de uma v.a. X.

Média *Mean*	$\overline{x} = \frac{1}{n}\sum_i x_i$ [a] $= \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$
Mediana *Median*	n ímpar: observação de ordem $\frac{n+1}{2}$ n par: média das observ. de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$
Moda *Mode*	observação que ocorre mais vezes

Observações:

(a) Nas conclusões de um estudo, deve sempre mencionar-se o tamanho da amostra, n.

(b) A média, a mediana e a moda exprimem-se nas unidades da v.a. correspondente. É comum apresentar estas estatísticas com uma casa decimal a mais do que os dados.

(c) Para o cálculo da mediana, as observações devem ser ordenadas por ordem crescente. A mediana corresponde ao valor para o qual metade das observações têm valores inferiores (e metade têm valores superiores).

[a] Equivalentemente, $\overline{\mu} = \sum_i x_i \, \widehat{f}(x_i)$ onde $\widehat{f}$ representa a frequência empírica relativa.

(d) A mediana é insensível a valores extremos da amostra muito altos ou muito baixos pois para o seu cálculo conta apenas a posição relativa dessas observações extremas e não o seu valor específico.

Trata-se de uma estatística mais influenciada pelos valores centrais da amostra do que pelos seus valores extremos.

(e) A média é muito sensível a valores extremos: valores muito baixos (resp. muito altos) fazem descer (resp. subir) a média, afastando-a do centro da amostra. Existindo muitos valores extremos é preferível apresentar a mediana à média.

(f) A moda é uma estatística usada essencialmente em dados discretos com muitas observações.

Nas situações em que a v.a. é contínua e pode tomar um grande número de valores (peso em g, altura em cm, ...), podem existir várias modas ou a moda pode mesmo não existir, tornando-se uma estatística irrelevante.

(g) Se adicionarmos uma constante a cada uma das observações, ou multiplicarmos cada observação por uma constante, quer a média quer a mediana são alteradas dessa constante.

(h) Quanto mais simétrica for a função de frequência relativa das observações, mais próximos serão os valores da média e da mediana. Neste caso é indiferente apresentar a média ou a mediana.

Nos casos em que a distribuição é assimétrica, é comum apresentar-se a mediana e não a média, porque a média é influenciada pelas caudas da distribuição e portanto fornece menos eficazmente informação sobre a localização das observações.

(i) A média de variáveis ordinais ou nominais tem de ser usada com muito cuidado. Podem ser feitas várias observações a este respeito:

- a média não será necessariamente um dos valores da variável (numa variável contínua, a média toma sempre um dos valores da variável)
- considere-se uma variável ordinal cujas k categorias foram ordenadas por ordem crescente segundo 1, 2, 3, ..., k. Nesta situação a média só pode ser considerada se a passagem de uma categoria para a seguinte tiver sempre o mesmo significado.
- no caso de uma variável binária, ordinal ou nominal, (sem perda de generalidade, tomando 0's e 1's), a média corresponde à proporção de 1's na amostra.

Medidas sumárias descritivas de v.a. categóricas

As medidas sumárias descritivas mais comuns para v.a. categóricas consistem apenas:

- do tamanho da amostra
- da designação das categorias (que, já se sabem, deverão ser numéricas para permitir efectuar análises estatísticas sem problemas no **R** e no **SPSS**)
- do número de observações em cada uma das categorias, e correspondente percentagem relativamente ao total de elementos na amostra (frequências absoluta e relativa de cada categoria).

O gráfico correspondente será simplesmente um gráfico de barras (em contagens ou percentagens, conforme se achar mais adequado).
Na situação em que existem missings:

- se se está a considerar apenas uma v.a. categórica, a indicação do número de observações omissas e da respectiva percentagem pode aparecer como a frequência de mais uma classe da v.a.
- se se estão a considerar várias v.a. categóricas, deve-se incluir uma primeira tabela que contenha, para cada v.a., o número de observações omissas na amostra. Só depois se deve apresentar a tabela com as frequências absoluta e relativa de cada categoria, para todas as v.a. categóricas em causa.

3.3. Medidas de escala de uma a.a.

Dados: a.a. $x_1, x_2, \ldots, x_n$ de uma v.a. X.

Variância *Variance*	$s^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ $= \frac{1}{n-1}\left((x_1-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\right)$
Desvio Padrão *Standard Deviation*	$s = \sqrt{\text{variância}}$
p-Quantil/Percentil $100p$ *p-Quantile/Percentile* $100p$	observação de ordem $p(n+1)$ ou média das observações vizinhas [a]
Amplitude *Range*	(maior observação)-(menor observação)
Amplitude Interquartil *Interquartile Range*	AIQ=(percentil 75)-(percentil 25)

Observações:

(a) É adequado apresentar o máximo e o mínimo da a.a.

(b) É adequado apresentar o desvio padrão com duas casas decimais a mais do que os dados.

(c) A mediana corresponde ao percentil 50.

(d) As unidades da variância são (unidades da v.a.)^2; as unidades do desvio padrão são as unidades da v.a.

(e) Suponhamos que a amostra consta de duas observações. Se somarmos simplesmente a dispersão dos dados em torno da média

$$(x_1 - \overline{x}) + (x_2 - \overline{x})$$

obtemos o valor 0. Mais geralmente, para uma amostra de tamanho n resulta da definição de média que

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Daí a necessidade de considerar os quadrados na fórmula de s^2.

(f) Há estudos em que a dispersão é medida usando $\sum |x_i - \overline{x}|$ mas esta fórmula traz complicações computacionais adicionais.

(g) Se adicionarmos uma constante a cada uma das observações, o valor do desvio padrão permanece inalterado. Se multiplicarmos cada observação por uma constante, o desvio padrão é multiplicado também por essa constante.

(h) Se for razoável assumir que a amostra provém de uma população que segue uma distribuição simétrica, devemos apresentar a média e o desvio-padrão [a]. Caso contrário, devemos apresentar a mediana e alguns percentis superiores e inferiores (por exemplo, os percentis 5, 25, 75 e 95) de forma a fornecer informações sobre a distribuição.

[a] O p-quantil pode também ser calculado a partir da função de distribuição empírica: corresponde ao número q tal que $\widehat{F}(q) = p$, obtido, se necessário, por interpolação linear. O **R** permite escolher a definição de quantil que se pretende, de entre nove possíveis.

[a] Apresentamos o desvio-padrão e não a variância, apenas por ser mais fácil de visualizar.

(i) O facto de dividirmos por $n-1$ e não por n na fórmula da variância prende-se com motivos de ordem técnica (veremos mais tarde que assim definida, a variância amostral é um estimador não tendencioso da variância da população). Pode-se ainda raciocinar em termos de graus de liberdade (as n observações da amostra não estão todas livres porque já se calculou a média, tendo imposto portanto uma condição).

Observação acerca dos quantis:

A definição de p-quantil apresentada na tabela (*type*=2 no **R**) é a mais comum em livros de Bioestatística. É simples de calcular "à mão" e generaliza o método de determinação da mediana mas é mais adequada para variáveis discretas.
Existem outros processos para o cálculo do p-quantil associado a uma amostra de tamanho n: é o número q (não necessariamente na amostra) tal que:

- $\widehat{F}(q) = p$.

  Esta definição é a análoga para o caso amostral da definição de p-quantil de uma v.a. contínua. É a minha preferida; no **R**, *type*=4 para v.a. contínuas, e *type=1* para v.a. discretas.

- $\frac{q-1}{n-1} = p$.

  Definição usada, por defeito, no **S** e no **R** (*type*=7)

- $\frac{q}{n+1} = p$.

  Definição usada no **SPSS**; *type*=6, no **R**.

A diferença entre os vários métodos é pequena e irrelevante para os nossos propósitos.

Simetria da distribuição amostral

A simetria da distribuição amostral (das frequências relativas) pode ser aferida pelos percentis 25 e 75:

- se estes percentis distarem aproximadamente o mesmo valor da mediana, então a distribuição é aproximadamente simétrica.

- se o percentil 75 estiver mais próximo da mediana do que o percentil 25, então a distribuição é assimétrica à esquerda; nestes casos, a média tende a ser menor do que a mediana (porque a média é influenciada pelos valores *baixos* da cauda esquerda).

- se o percentil 25 estiver mais próximo da mediana do que o percentil 75, então a distribuição é assimétrica à direita; nestes casos, a média tende a ser maior do que a mediana.

A a.a. provém de uma distribuição normal?

Mais tarde veremos como testar esta hipótese usando um *teste de hipóteses* adequado.
Para ter uma ideia rápida da resposta à questão podem analisar-se alguns percentis da amostra. Valores muito diferentes dos correspondentes aos da distribuição normal sugerem que apenas a indicação da média e do desvio padrão não é uma boa descrição dos dados. Para a distribuição normal tem-se aproximadamente a seguinte correspondência

$$\begin{array}{rcl} \overline{x} - 2\,s & \approx & \text{percentil } 2.5 \\ \overline{x} - 0.67\,s & \approx & \text{percentil } 25 \\ \overline{x} & \approx & \text{mediana} \\ \overline{x} + 0.67\,s & \approx & \text{percentil } 75 \\ \overline{x} + 2\,s & \approx & \text{percentil } 97.5 \end{array}$$

3.4. Erro padrão (amostral) da média

Dada uma a.a. $x_1, x_2, \ldots, x_n$ de tamanho n, o erro padrão amostral da média (sample standard error of the mean) é representado por $s.e.$ [a] e definido por

$$s.e. = \frac{s}{\sqrt{n}}.$$

Trata-se de uma estatística que estima a precisão com que a média da a.a. estima a média da população.
É sempre inferior ao desvio-padrão amostral.
Por vezes é mencionado nalguns estudos porque fá-los parecer melhores!...em geral, os dados devem ser sumariados usando o desvio-padrão, que quantifica a variabilidade na população, e não o desvio padrão da média.

[a] *S.E.*, *s.e.m.*, *sem* ou *SEM*

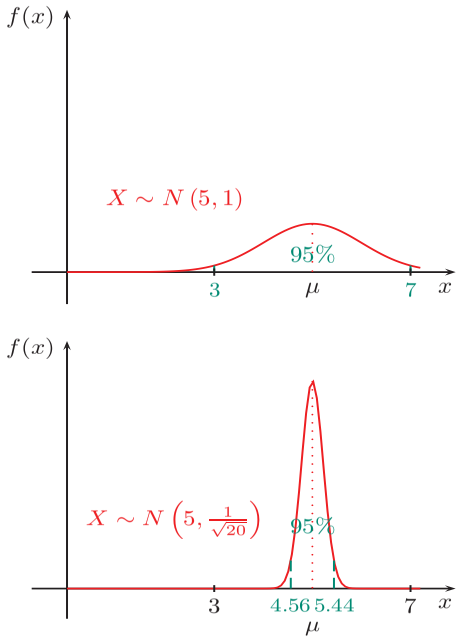
Considere-se uma a.a. de tamanho $n = 20$ com

$$\overline{x} = 5 \qquad \text{e} \qquad s = 1.$$

Tem-se

$$s.e. = \frac{1}{\sqrt{20}} \approx 0.22.$$

Confundir o desvio padrão da média com o desvio padrão, numa população que segue uma distribuição normal com os parâmetros amostrais, confere uma ideia errada acerca da dispersão da distribuição:



3.5. Coeficiente de Variação

Dada uma a.a. $x_1, x_2, \ldots, x_n$ de tamanho n em que todas as observações têm o mesmo sinal, o coeficiente de variação (coefficient of variation), representado por CV é definido por

$$CV = 100\% \times \frac{s}{\overline{x}}.$$

Trata-se de uma quantidade adimensional particularmente útil quando se pretende comparar a dispersão relativa de várias variáveis aleatórias com [a]:

- médias bastante diferentes, ou
- unidades de medida diferentes.

De facto, o desvio padrão dos dados só pode ser interpretado tendo em conta a média amostral: um desvio padrão de 10 numa a.a. com média amostral 20 tem consequências diferentes de um desvio padrão de 10 numa a.a. com média amostral 100. Na primeira situação tem-se $CV = 50\%$ enquando que na segunda $CV = 10\%$ portanto a dispersão relativa é menor no segundo caso.

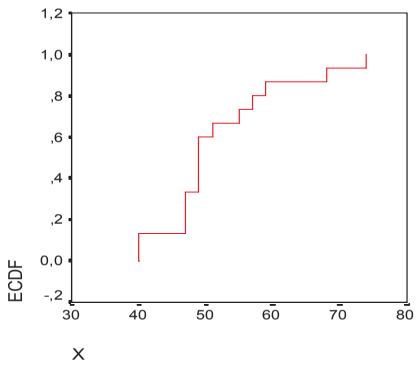
[a] Nesta situação não se devem comparar os desvios padrão.

3.6. Gráficos associados a uma amostra de uma variável contínua

Dados: Uma amostra aleatória de tamanho n de v.a. i.i.d. contínuas, $Y_1, \ldots, Y_n$, seguindo uma distribuição de probabilidade F (não necessariamente conhecida).

A **função distribuição empírica** $\widehat{F}$ da amostra é

$$\widehat{F} : \mathbb{R} \longrightarrow \mathbb{R},\ x \mapsto \frac{1}{n}\,(\#\text{ dos } x_i \leq x)$$



Trata-se sempre de uma função crescente, em (forma de) escada, cujo primeiro valor é 0 e o último é 1.
A função distribuição empírica é a função distribuição de probabilidade associada à função densidade de probabilidade dada pela frequência com que cada observação aparece na amostra. Por este motivo, para o cálculo de $\widehat{F}$ pode construir-se uma tabela como a que se apresenta em seguida: ordenam-se as observações por ordem crescente e para cada observação x_i regista-se o seu número de ocorrências (frequência absoluta) n_i.

valores x_i	frequência absoluta n_i	frequência relativa n_i/n	frequência relativa acumulada $\widehat{F}(x_i) = \sum_{x_j \leq x_i} n_j/n$
...	...	...	...
...	...	...	...
...	...	...	1

A frequência relativa de cada observação obtém-se divindo a sua frequência absoluta pelo tamanho n da amostra, e a frequência relativa acumulada ($\widehat{F}(x_i)$) obtém-se somando as frequências relativas de todas as observações menores ou iguais a x_i.
Observe-se que a soma de todas as frequências absolutas tem de ser igual ao tamanho da amostra (n) e o último valor da frequência relativa acumulada tem de ser 1.
Finalmente, o gráfico da função distribuição empírica é construído a partir dos valores das primeira e última colunas, correspondentes, respectivamente, aos vários valores de x_i e $\widehat{F}(x_i)$.

Boxplot (ou **caixa de bigodes**)
A designação portuguesa para este tipo de gráfico resulta do nome *box-and-whisker plot* (*box*=caixa, *whiskers*=bigodes) actualmente em deduso e substituído pela versão simplificada *boxplot*.
A construção de um boxplot é feita essencialmente a partir dos quartis (0.25, 0.5 e 0.75 quantil) do conjunto de dados e procede-se da seguinte forma.

Representem-se os valores da variável de interesse no eixo vertical de um sistema de coordenadas.
Desenhe-se uma caixa com base um segmento de recta correspondente ao valor do 0.25 quantil e topo um segmento de recta correspondente ao valor do 0.75 quantil (ambos com a mesma largura, arbitrária). No interior dessa caixa construa-se, com uma espessura maior do que a usada anteriormente, um segmento de recta, paralelo ao eixo dos xx´s e correspondente ao valor da mediana (0.5-quantil).
Pode-se agora proceder de 2 formas:

- a mais simples consiste em marcar um traço horizontal correspondente aos valores extremos (menor e maior valor) da amostra e ligar esses traços horizontais à caixa inicial através de segmentos de recta verticais (os bigodes).

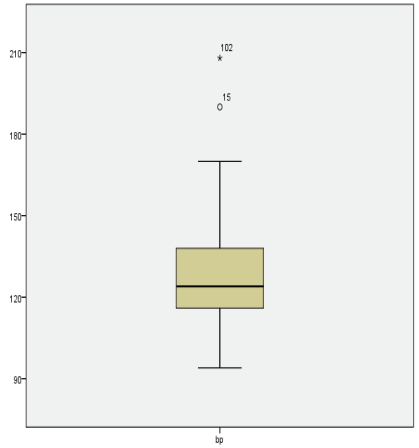
- a segunda forma tem em consideração a existência de observações no conjunto de dados que estão *muito distantes* da maior parte das outras (***outliers***). Determine-se a amplitude interquartil

$$AIQ = q_{0.75} - q_{0.25}$$

e as barreiras internas superior (BIS) e inferior (BII), e as barreiras externas superior (BES) e inferior (BEI):

$$\begin{array}{ll} BII = q_{0.25} - 1.5AIQ, & BIS = q_{0.75} + 1.5AIQ \\ BEI = q_{0.75} - 3AIQ, & BES = q_{0.75} + 3AIQ \end{array}$$

Marquem-se agora dois traços horizontais: um correspondente ao maior valor da amostra que não é superior à BIS, e um outro correspondente ao menor valor da amostra que não é inferior à BII. Liguem-se esses traços horizontais à caixa inicial através de segmentos de recta verticais (os bigodes).



Calculem-se agora as barreiras externas superior e inferior. Todas as observações que se localizem entre as barreiras BIS e BES, ou BII e BEI, são designadas por **outliers moderados** e representadas no boxplot por pequenos círculos abertos, ∘ ; todas as observações que tomem valores para além das barreiras externas (maiores do que a BES ou menores do que a BEI) são designadas por **outliers severos** e representadas no boxplot por asteriscos,∗.
Alguns programas de computador não fazem distinção entre os dois tipos de outliers. Outros apresentam um número associado a cada outlier; esse número representa o número da observação (o número que lhe corresponde no ficheiro).

Por construção, o boxplot é um gráfico que traduz informação acerca da zona de concentração, dispersão e simetria do conjunto de dados observados.

- 50% das observações estão contidas na caixa $(q_{0.75} - q_{0.25})$, e essas são precisamente as observações centrais.
- Para avaliar a dispersão podem-se usar:
  - a amplitude do intervalo de variação: $H = \max - \min$
  - a amplitude interquartil: $AIQ = q_{0.75} - q_{0.25}$.

  A amplitude do intervalo de variação (e o desvio padrão) são sensíveis à presença de outliers. A AIQ é uma medida de dispersão resistente pois no seu cálculo não intervêm as observações maiores nem as menores.
- Quanto ao estudo da simetria da distribuição, convém ter em consideração o seguinte:
  - se as distâncias $q_{0.75} - q_{0.5}$ e $q_{0.5} - q_{0.25}$ forem aproximadamente iguais, então a distribuição é aproximadamente simétrica
  - se se verificar $q_{0.75} - q_{0.5} > q_{0.5} - q_{0.25}$ então a distribuição é assimétrica à direita
  - se se verificar $q_{0.75} - q_{0.5} < q_{0.5} - q_{0.25}$ então a distribuição é assimétrica à esquerda.

Os boxplots são gráficos que se prestam a comparações entre variáveis, desde que estas se exprimam nas mesmas unidades.

Histograma

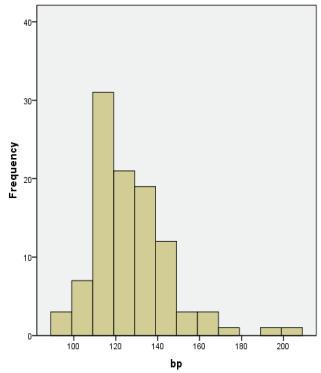
Divide-se a recta real em k intervalos disjuntos consecutivos (classes)

$$I_1 =]-\infty, a_1], \qquad I_2 =]a_1, a_2], \qquad \dots$$
$$\dots, \qquad I_{k-1} =]a_{k-2}, a_{k-1}], \qquad I_k =]a_{k-1}, +\infty[$$

com $-\infty < a_1 < a_2 < \cdots < a_k < +\infty$.

O histograma das observações para estas classes é o gráfico que atribui a cada classe a frequência (número de observações), ou a frequência relativa (proporção das observações), dessa classe.

A determinação das classes deve ser feita usando um compromisso entre precisão relativamente à amostra e variabilidade nas frequências estatisticamente pouco significativas. Existem vários algoritmos que permitem determinar o número de classes a usar. A forma do histograma obtido pode variar consideravelmente conforme o número de classes considerado.

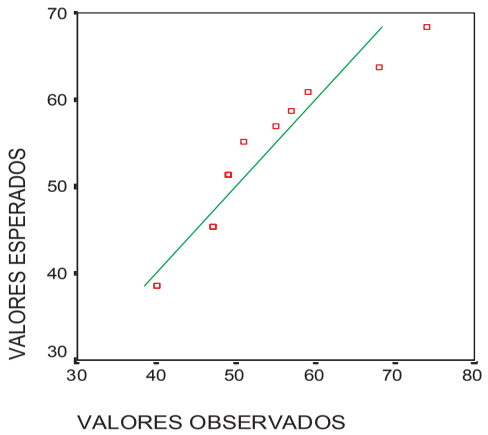


O histograma é uma representação discreta da função densidade de probabilidade. Em particular, permite avaliar a forma da distribuição da v.a. em causa.

Q-Q plot: gráfico dos quantis

Trata-se do gráfico dos quantis de uma distribuição de probabilidade teórica previamente especificada de uma variável aleatória Z (usualmente a distribuição normal) contra os quantis amostrais da distribuição empírica $\widehat{F}$ associada a uma amostra $y_1, \ldots, y_n$. Permite detectar de forma "visual" se a distribuição empírica está ou não próxima da distribuição teórica especificada. Constrói-se do modo a seguir descrito.

Seja y_i uma observação da amostra. Tem-se $\widehat{F}(y_i) = p_i$ para algum $p_i \in [0, 1]$ e, por definição, y_i é o p_i-quantil amostral.
O p_i-quantil (teórico) de Z é o número z_i tal que $P(Z \leq z_i) = p_i$.
O Q-Q plot é o gráfico que contém os pares ordenados (z_i, y_i), para todos os valores $y_1, \ldots, y_n$ da amostra.

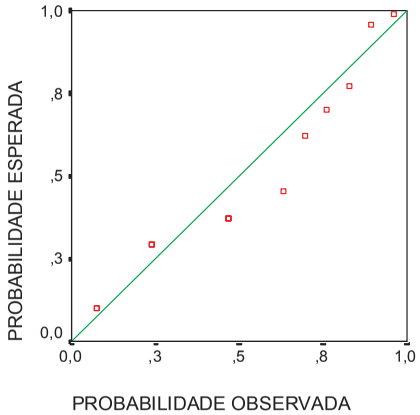


A linha recta no gráfico representa a correspondência perfeita entre os quantis amostrais e os quantis teóricos.

P-P plot: gráfico de probabilidades

Trata-se do gráfico de uma distribuição de probabilidade G de uma variável aleatória Z contra a distribuição empírica $\widehat{F}$ associada a uma amostra $y_1, \ldots, y_n$. Tal como o Q-Q plot, permite também detectar de forma "visual" se $F = G$ ou $F \neq G$. Constrói-se do modo a seguir descrito.

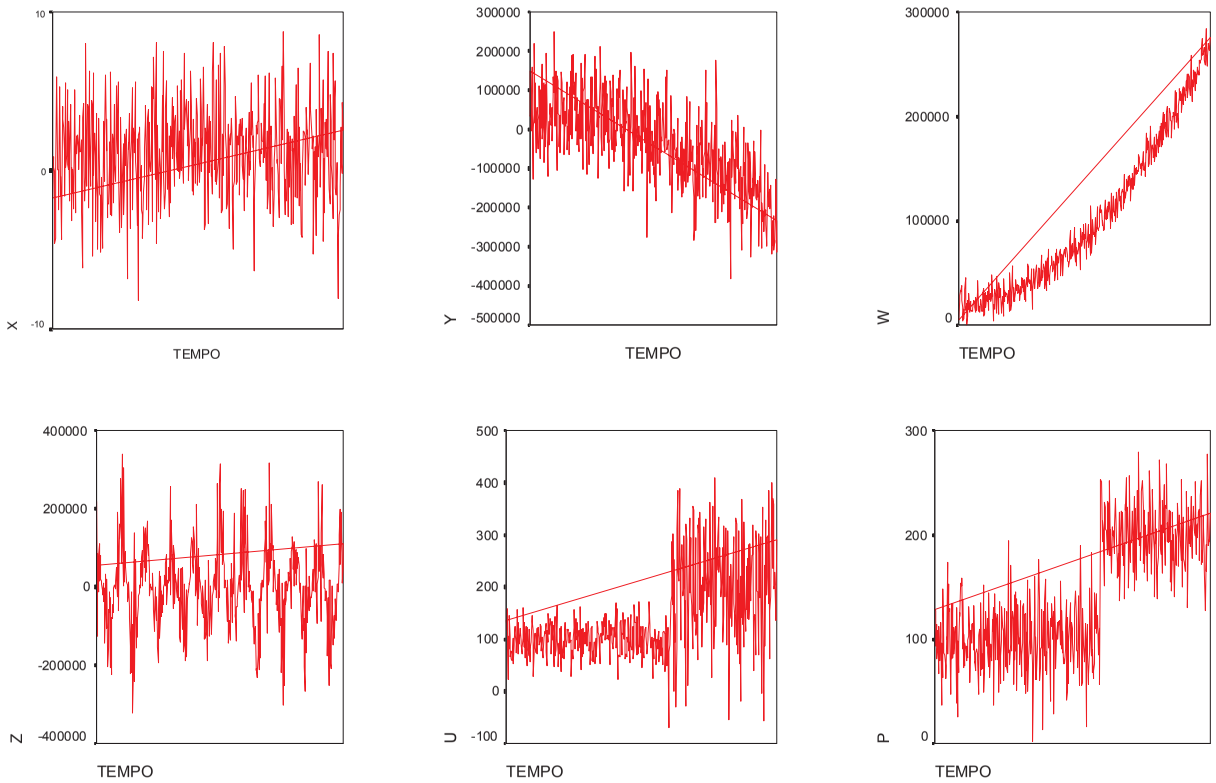
Seja $y \in \mathbb{R}$. A probabilidade empírica de termos $Y \leq y$ é $\widehat{F}(y)$. A probabilidade (teórica) de termos $Z \leq y$ é $G(y)$.
O P-P plot é o gráfico que contém os pares ordenados $(G(y_i), \widehat{F}(y_i))$, para todos os valores $y_1, \ldots, y_n$ da amostra.



A linha recta no gráfico representa a correspondência perfeita entre a distribuição amostral e a distribuição teórica.

Tal como no caso do Q-Q plot, a distribuição teórica usada mais frequentemente é a distribuição normal.

Gráfico sequencial (série temporal):
É o gráfico que atribui ao número de ordem da observação o valor da observação.



Permite a detecção "visual" de não aleatoriedade no processo sequencial que originou as observações (tendência crescente ou decrescente, periodicidade, mudança qualitativa no processo, ...)

3.7. $\hat{F}$, boxplot e quantis: instruções em SPSS e exemplos

Uma equipa de médicos dentistas realizou um estudo em que se pretendeu determinar se ensinar as pessoas a lavar correctamente os seus dentes é ou não uma prática benéfica. Foi escolhida uma clínica de medicina dentária e 12 pares de utentes dessa clínica, cada par com características semelhantes quanto à idade, sexo, inteligência e hábitos de higiene oral. A um dos elementos de cada par escolhido, foi dada instrução sobre a melhor forma de lavar os dentes, entre outras recomendações de boa higiene oral. Seis meses depois, todos os 24 indivíduos foram analisados, e classificados de acordo com a higiene oral apresentada, por um dentista que desconhecia quais dos pacientes tinham recebido instrução prévia. Nessa classificação, valores baixos significam bons níveis de higiene oral. Os resultados obtidos foram os seguintes:

Número do par	**Classificação**		Número do par	**Classificação**	
	C/ instrução	**S/ instrução**		**C/ instrução**	**S/ instrução**
1	1.5	2.0	7	2.0	3.5
2	2.0	2.0	8	1.5	3.0
3	3.5	4.0	9	1.5	2.5
4	3.0	2.5	10	2.0	2.5
5	3.5	4.0	11	3.0	2.5
6	2.5	3.0	12	2.0	2.5

Atente-se apenas na variável que se refere à classificação obtida sem instrução. Os percentis amostrais obtidos do SPSS são

0%	10%	20%	30%	40%	
2.0	2.0	2.3	2.5	2.5	
50%	60%	70%	80%	90%	100%
2.5	2.9	3.05	3.7	4.0	4.0

Para comparação, e por curiosidade, os percentis obtidos do **R** são

0%	10%	20%	30%	40%	
2.00	2.05	2.50	2.50	2.50	
50%	60&	70%	80%	90%	100%
2.50	2.80	3.00	3.40	3.95	4.00

Repare-se, em particular, que os percentis 30, 40 e 50 são coincidentes!
O que se passa (ver valores da função de distribuição empírica a seguir) é que o valor 2.5 aparece várias vezes na amostra e tem uma frequência relativa acumulada de $7/12 \approx 0.58$, enquanto que o valor anterior (2.0) tem uma frequência relativa acumulada apenas de $2/12 \approx 0.17$. [a]

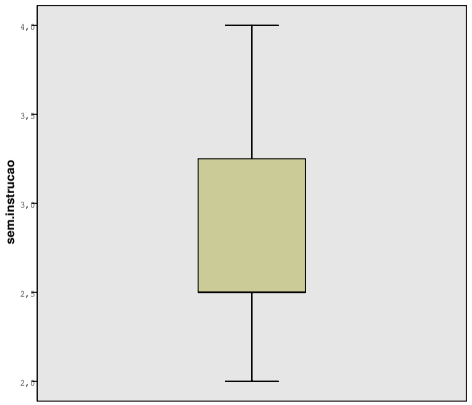
[a] A função no software **R** que, por defeito, calcula os quantis não usa a função distribuição empírica... mas as diferenças são pequenas e irrelevantes para o presente propósito. Em particular, as diferenças serão tão menores quanto maior for o tamanho da amostra.

Nota: A instrução em SPSS para a obtenção de qualquer percentil é

Analyze → Descriptive Statistics
→ Frequencies
→ Statistics
→ Percentile(s)
→ Add ...

O boxplot (o 0.25-quantil e o 0.5-quantil são coincidentes) é obtido de, por exemplo,

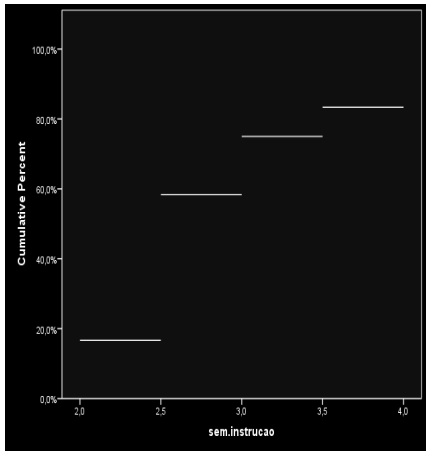
Graphs → Chart Builder
→ Gallery; Choose from: Boxplot
→ X-axis?: sem.instrucao

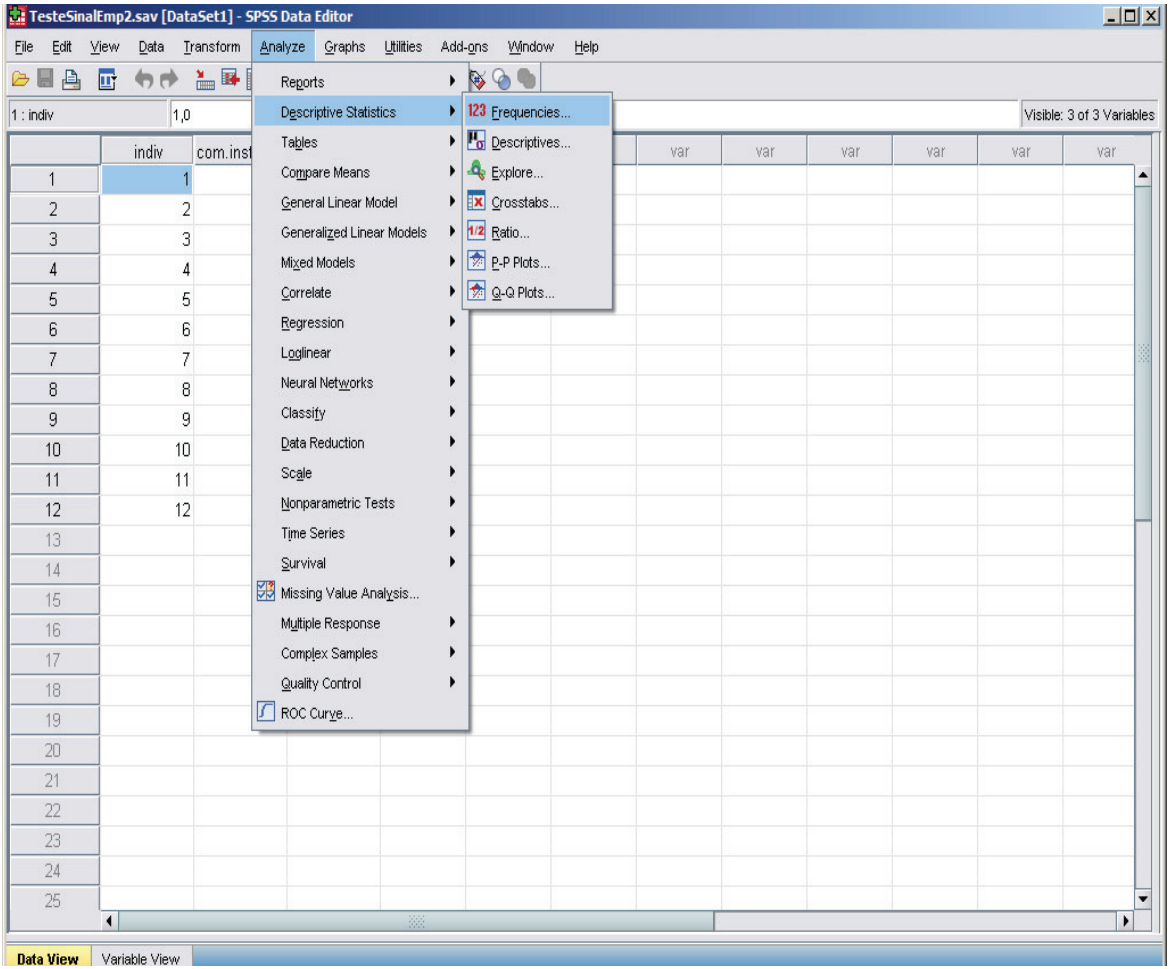


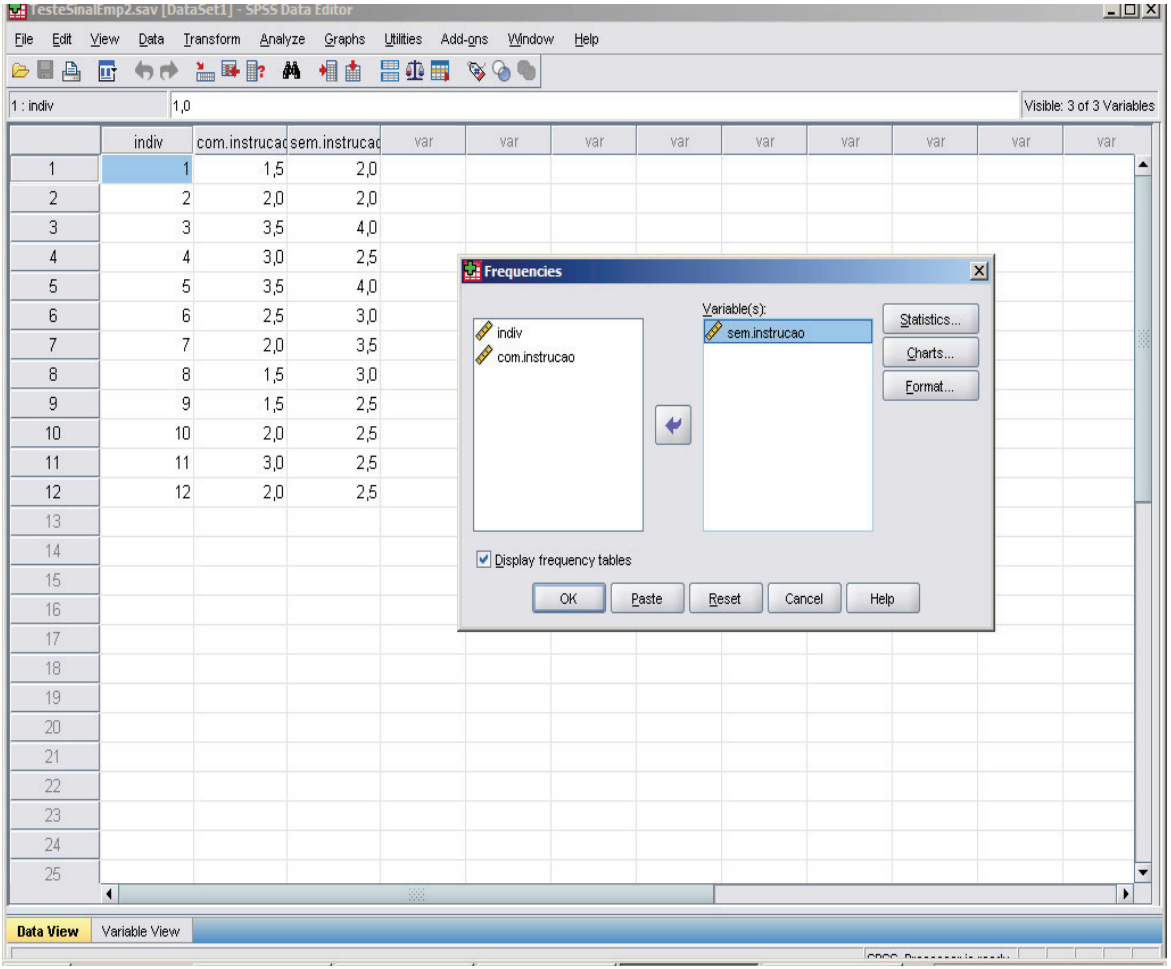
Percentis	($n = 12$)	
25%	50% (mediana)	75%
2.50	2.50	3.375

Quanto aos valores da função de distribuição empírica, $\widehat{F}$, tem-se (instruções em SPSS nas páginas seguintes)

x_i	n_i	$\widehat{F}(x_i)$
2.0	2	$2/12 \approx 0.167$
2.5	5	$7/12 \approx 0.583$
3.0	2	$9/12 \approx 0.750$
3.5	1	$10/12 \approx 0.833$
4.0	2	1.000




TesteSinalEmp2.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : indiv 1,0
Visible: 3 of 3 Variables
indiv com.inst
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...
Data View
Variable View


TesteSinalEmp2.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : indiv 1,0
Visible: 3 of 3 Variables
indiv com.instrucao sem.instrucao var
1 1 1,5 2,0
2 2 2,0 2,0
3 3 3,5 4,0
4 4 3,0 2,5
5 5 3,5 4,0
6 6 2,5 3,0
7 7 2,0 3,5
8 8 1,5 3,0
9 9 1,5 2,5
10 10 2,0 2,5
11 11 3,0 2,5
12 12 2,0 2,5
Frequencies
indiv
com.instrucao
Variable(s):
sem.instrucao
Statistics...
Charts...
Format...
Display frequency tables
OK Paste Reset Cancel Help
Data View
Variable View

Statistics

sem.instrucao

N	Valid	12
	Missing	0

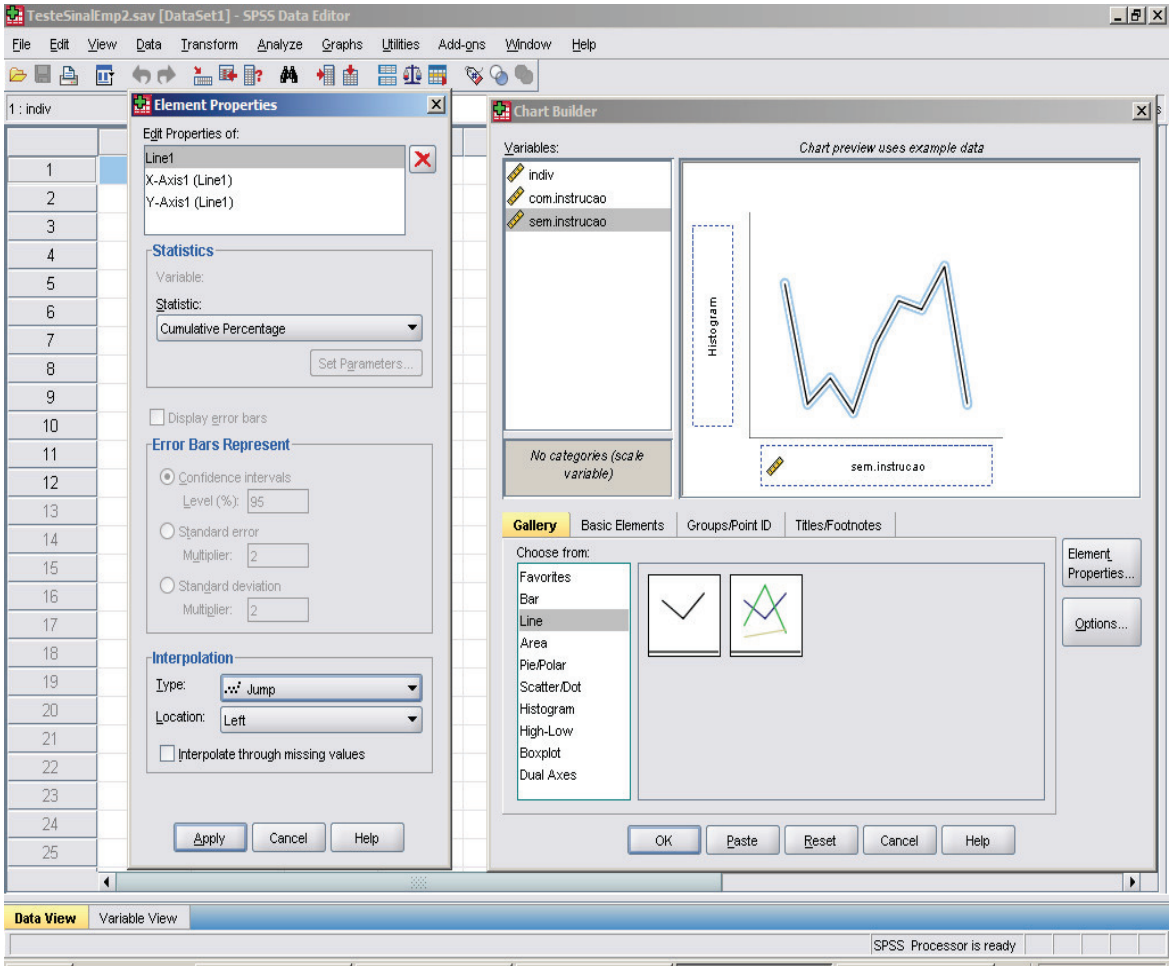
sem.instrucao

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2,0	2	16,7	16,7	16,7
	2,5	5	41,7	41,7	58,3
	3,0	2	16,7	16,7	75,0
	3,5	1	8,3	8,3	83,3
	4,0	2	16,7	16,7	100,0
	Total	12	100,0	100,0	

O gráfico da função distribuição empírica é obtido de

Graphs → Chart Builder

escolhendo "Line" da galeria de gráficos e arrastando a variável de interesse (aqui **sem.instrucao**) para o eixo dos xx's. Em seguida, vai-se a "Element Properties" e escolhe-se "cumulative percentage" do menu Statistics, e "Type: Jump" e "Location: Left" do menu Interpolation. Clica-se em Apply ainda nessa janela, e depois em OK na janela principal.


TesteSinalEmp2.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : indiv
Element Properties
Edit Properties of:
Line1
X-Axis1 (Line1)
Y-Axis1 (Line1)
Statistics
Variable:
Statistic:
Cumulative Percentage
Set Parameters...
Display error bars
Error Bars Represent
Confidence intervals
Level (%): 95
Standard error
Multiplier: 2
Standard deviation
Multiplier: 2
Interpolation
Type: Jump
Location: Left
Interpolate through missing values
Apply
Cancel
Help
Chart Builder
Variables:
Chart preview uses example data
indiv
com.instrucao
sem.instrucao
Histogram
sem.instrucao
No categories (scale variable)
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
Element Properties...
Options...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
SPSS Processor is ready

3.8. Estatística Descritiva: Exemplos

Nesta secção usamos o ficheiro bp.obese.sav [a] Este ficheiro, com 102 linhas e 3 colunas, contém dados recolhidos em adultos americanos de origem mexicana residentes numa cidade da Califórnia, EUA. As variáveis presentes são as seguintes:

- **sex**: factor com 2 níveis que indica o género de cada indivíduo: 0-masculino; 1-feminino.
- **obese**: vector numérico que representa a razão entre o peso actual e o peso ideal (obtido da tabela "New York Metropolitan Life Tables")
- **bp**: vector numérico que representa a pressão arterial sistólica (mm Hg).

Em primeiro lugar, e antes de qualquer análise, deve ir a "Variable View", no canto inferior esquerdo da janela do SPSS, e confirmar que a classificação de cada uma das variáveis (scale, ordinal, nominal) está bem feita. Para não ter problemas com os gráficos, é também conveniente que todos os factores (variáveis categóricas com um número finito de categorias) tenham designações ("labels") - para atribuir esses valores, basta clicar na célula da tabela correspondente à variável em causa e à coluna "Values".
No canto inferior direito da janela aberta pelo SPSS tem ainda a possibilidade "Data View", que lhe permite ver os valores de cada uma das variáveis.

[a] P. Dalgaard, *Introductory Statistics with* **R**.

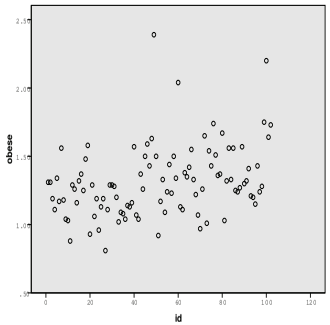
Uma forma de verificar que não se efectuaram erros ao introduzir os dados consiste na consideração de gráficos do tipo *scatter*:

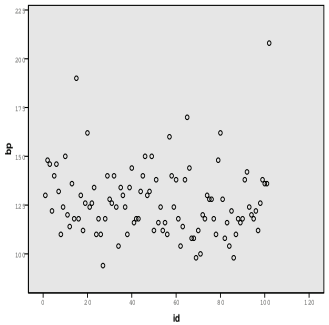
Graphs → Chart Builder

Aparecem duas janelas: fechar a janela "Element Properties" e ficar com "Chart Builder". Nesta janela, escolher

Gallery → Scatter/Dot

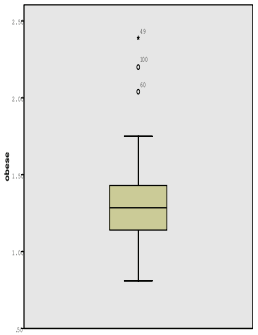
e arrastar o tipo de gráfico que se pretende para cima. Em relação aos eixos, a variável **id**, de identificação dos indivíduos, deve ficar sempre no eixo dos xx's, e para o eixo dos yy's deve arrastar-se a variável de interesse (**obese** e **bp**, nos gráficos apresentados abaixo).

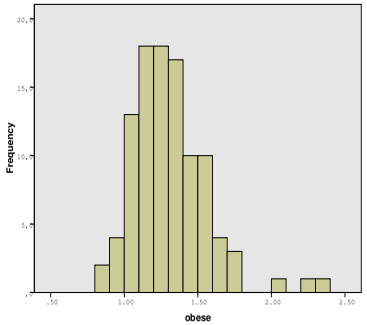




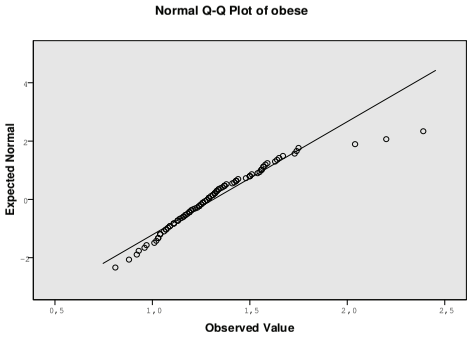
Pontos muito distantes da *nuvem* formada pelos restantes são duvidosos e merecem uma segunda confirmação do seu valor. No presente caso, parece estar tudo bem.

Fazemos agora a análise exploratória da v.a. **obese**, começando por uma análise descritiva gráfica:
A existência de 3 outliers (ver boxplot) altera significativamente a média e essencialmente impede a totalidade dos valores de provirem de uma distribuição simétrica:





Em particular a distribuição de onde provêm os dados não parece ser normal:



Seria portanto mais indicado usar a mediana e alguns quantis para descrever quantitativamente a v.a.:

Percentis ($n = 102$)				
5	25	50 (mediana)	75	95
0.962	1.138	1.285	1.433	1.739

Estes valores obtiveram-se através das seguintes intruções

Analyze → Descriptive Statistics → Explore

e na janela que se abriu, arrastou-se a variável de interesse (**obese**) para o lugar de "Dependent List"; no rectângulo da direita, "Statistics", deu-se a indicação "Percentiles".
(Pelo facto de não se poder admitir que os valores provêm de uma distribuição normal, ver-se-á mais tarde que não será adequado aplicar testes paramétricos, como por exemplo o teste t da média, a esta variável.)
O histograma e o boxplot anteriores obtêm-se de

Graphs → Chart Builder

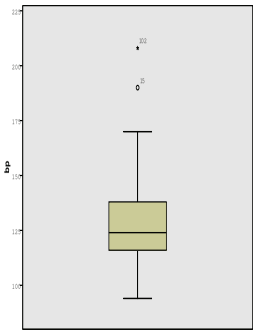
clicando nos gráficos adequados que aparecem na parte inferior da janela e escolhendo a v.a. **obese** para o eixo dos xx's e yy's, respectivamente.
O qq-plot obtem-se de

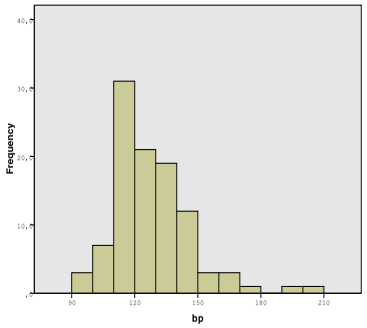
Analyze → Descriptive Statistics → Explore

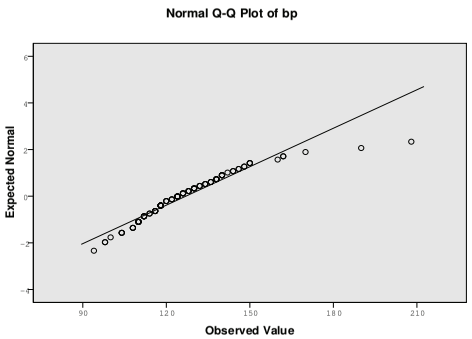
escolhendo "Normality plots with tests" na caixa "Plots". Alternativamente, pode-se fazer

Analyze → Descriptive Statistics → QQ-Plots.

Outra variável contínua no ficheiro é **bp**, que representa os valores da pressão arterial sistólica (mm Hg). Uma análise gráfica semelhante à anterior mostra outra vez que não se deve assumir assimetria (existem 2 outliers à direita) nem normalidade.





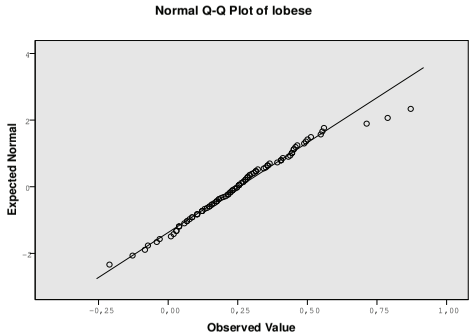


Percentis ($n = 102$)				
5	25	50 (mediana)	75	95
104.0	116.0	124.0	138.0	161.7

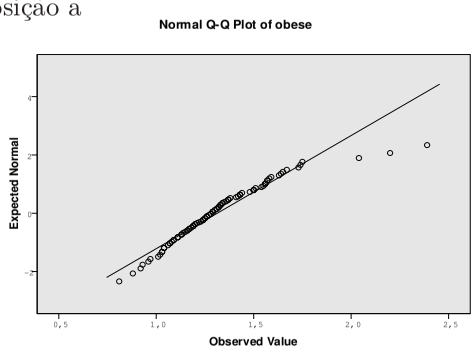
Suponhamos que a normalidade das v.a. anteriores era necessária para a aplicação de determinados métodos estatísticos. Um procedimento usual consiste na transformação das variáveis existentes e na análise de normalidade das variáveis transformadas. É um método que nem sempre funciona, mas funcionando torna-se muito prático. A transformações mais usuais são

$$\ln(y), \qquad 1/y, \qquad \sqrt{y}$$

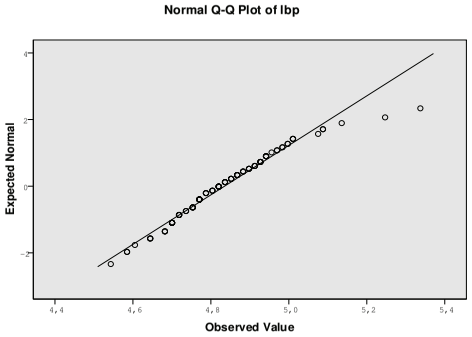
onde se designou por y a v.a. em causa.
Na situação anterior, a função logaritmo parece produzir alguns resultados. Para a normalidade tem-se



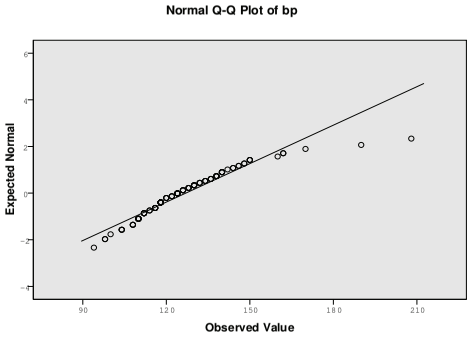
por oposição a



e



por oposição a



(A melhoria não é extraordinária; os exemplos foram apresentados a título elucidativo.) A transformação das variáveis foi conseguida fazendo

Tranform → Compute Variable.

Supondo razoável aceitar que, na população, a distribuição de

lobese=ln(**obese**)

é normal (e portanto simétrica), a descrição quantitativa dessa variável é feita usando o tamanho da amostra, a média e o desvio-padrão:

$$n = 102, \qquad \overline{x} = 0.255, \qquad s = 0.1854.$$

Estes valores foram obtidos de

Analyze → Descriptive Statistics → Explore.

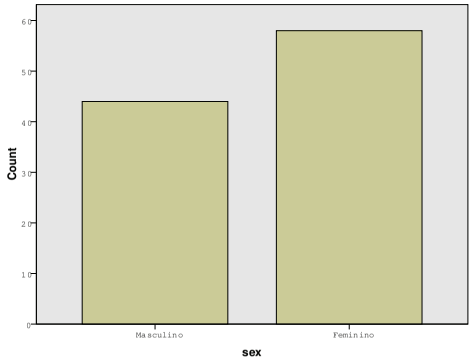
A v.a. **sex** é a uma única variável discreta (nominal) no ficheiro. A descrição deste tipo de variáveis é feita através da instrução

Analyze → Frequencies

que indica o número de elementos em cada categoria:

Sex		
	n	%
Masculino	44	43.1
Feminino	58	56.9
Total	102	100

O gráfico adequado é agora o gráfico de barras



Um gráfico de barras é diferente de um histograma:

- ambos representam contagens mas de tipos diferentes: o gráfico de barras conta o número de indivíduos observados exactamente em cada uma das categorias; o histograma conta o número de indivíduos em cada um dos intervalos reais.
- o gráfico de barras tem as barras separadas umas das outras, para evidenciar o carácter discreto da v.a. correspondente; o histograma tem as barras juntas, para evidenciar o carácter contínuo da v.a. correspondente.

A instrução para construir um gráfico de barras é

Graphs → Chart Buider

escolhendo depois "Bar" na caixa "Gallery" que aparece no canto inferior esquerdo da janela "Chart Builder".

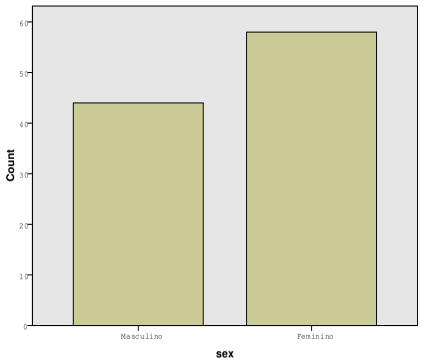
Suponhamos agora que o interesse consiste do estudo da variável **bp** por sexo. Pretende-se caracterizar a v.a. em cada umas das subpopulações e averiguar (visualmente; mais tarde veremos como abordar esta questão de uma forma mais sistemática) se existem diferenças significativas entre as duas subpopulações. Tendo em conta os melhores resultados da variável **lbp**, é com esta última variável que se vai trabalhar. Seguindo a instrução

Analyze → Descriptive Statistics → Explore

e escolhendo, por arrastamento, **lbp** na caixa "Dependent List" e **sex** na caixa "Factor List", obtêm-se as medidas de localização e escala pretendidas.

lbp			
sex	n	$\overline{x}$	s
Masculino	44	4.844	0.1241
Feminino	58	4.828	0.1428

Repare-se que esta instrução fornece também qq-plots para cada um dos géneros e um boxplot *comparativo* entre os géneros. No caso concreto, assumiu-se normalidade (ou pelo menos simetria) para as distribuições em cada subpopulação.



O boxplot anterior também pode ser obtido de

Graphs → Chart Buider

escolhendo o boxplot adequado de entre as várias imagens que aparecem quando se clica em "Boxplot", dentro da caixa "Gallery". Para o eixo dos xx's deve-se arrastar o factor **sex** e para o eixo dos yy's a variável contínua **lbp**.

Suponhamos agora que se pretende criar uma variável categórica associada a **obese**, digamos **obese.c**, definida por

$$\begin{aligned} \textbf{obese.c} &= 1, && \text{se } \textbf{obese} < 1.00 \\ \textbf{obese.c} &= 2, && \text{se } 1.00 \leq \textbf{obese} \leq 1.20 \\ \textbf{obese.c} &= 3, && \text{se } \textbf{obese} > 1.20 \end{aligned}$$

Uma instrução em **SPSS** para isso é

Tranform → Compute Variable

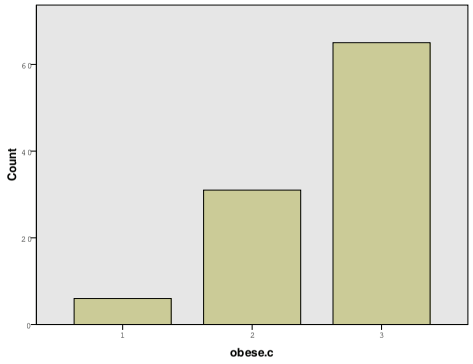
escrever **obese.c** em "Target Variable" e usar a caixa "If" para preencher os valores de **obese.c** em "Numeric Expression". Por exemplo, **obese.c**=1 IF (include if case satisfies condition) **obese**> 1.

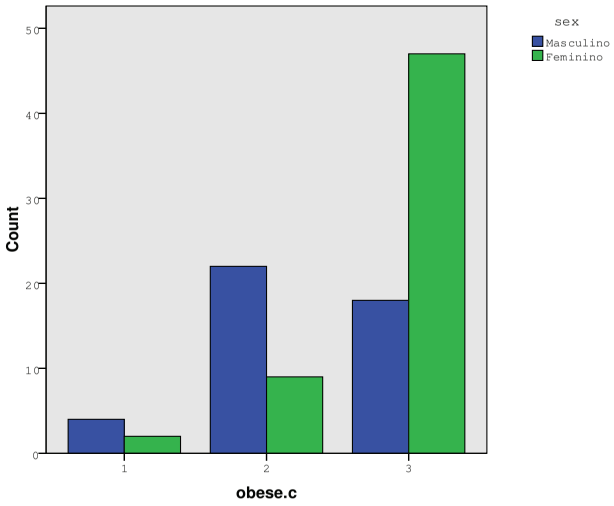
obese.c	1	2	3
n	6	31	65

A distribuição de **obese.c** e de **obese.c** por género são apresentadas nos gráficos de barras seguintes. Ambos foram construídos através de

Graphs → Chart Builder

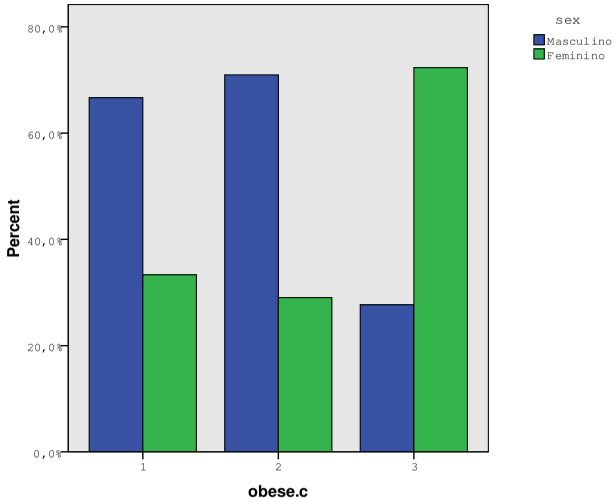
fazendo depois as escolhas adequadas.





No caso em que os vários grupos do factor não têm todos a mesma frequência, um gráfico de barras de percentagens pode ser mais útil do que um gráfico de barras de contagens.

Se no caso da distribuição de **obese.c** por género estivermos interessados na proporção de homens e mulheres dentro de cada categoria, um gráfico mais adequado do que o anterior será



Estes gráficos são obtidos do "Chart Builder" indicando na janela "Element Properties" que a estatística a usar no gráfico é a percentagem (e qual delas, de entre as várias possibilidades apresentadas pelo software).
As tabelas correspondentes a estes gráficos obtêm-se de

Analyze → Descriptive Statistics → Crosstabs

escolhendo, por exemplo, **obese.c** para a caixa das linhas "Row(s)" e **sex** para a caixa das colunas "Column(s)".

	sex		
obese.c	Masculino	Feminino	Total
1	4	2	6
2	22	9	31
3	18	47	65
Total	44	58	102

Outra instrução que permite categorizar a v.a. **obese** de acordo com a lei anterior é

Transform → Recode into different variables.

A variável **obese** tem de ser passada para a caixa central da janela que entretanto se abriu, e à direita, em *output variable*, tem de se incluir o nome da variável que se pretende definir, **obese.c**, e clicar na caixa *Change*. Os valores de **obese.c** são preenchidos na janela que se abre quando se clica em *Old and New Values*. Os *Old Values* referem-se aos valores da variável **obese** (neste caso, teríamos de escolher e preencher o item *Range*) e os *New Values* dizem respeito aos valores da nova variável, **obese.c**. Depois destes valores preenchidos, resta clicar na caixa *Add*, e ir repetindo o processo o número de vezes que for necessário.
A este propósito, convém ainda referir que todas as opções que são disponibilizadas dentro do item *Range* incluem os valores especificados. Mais concretamente,

- **a through b:** altera todos os x tais que $a \leq x \leq b$
- **LOWEST through value** c**:** altera todos os x tais que $x \leq c$
- **value** d **through HIGHEST:** altera todos os x tais que $x \geq d$.

Existe ainda um outro processo de conseguir a categorização da variável **obese** de acordo com as especificações anteriores. A instrução é

$$\text{Transform} \rightarrow \text{Visual Binning}.$$

Na janela que se abre, deve-se escolher **obese** em *Variables to Bin* e clicar em *Continue*. Em seguida clica-se sobre a variável **obese** e esta passa imediamente para o lado direito da janela, para *Current Variable*. O rectângulo correspondente a *Current Variable* deve ser preenchido com o nome da nova variável, **obese.c**. Agora resta preencher a tabela de valores de corte. Pretendendo *labels*, estes são criados pela instrução *Make Labels*.
O comando *Visual Binning* serve também para categorizar variáveis usando intervalos com a mesma amplitude ou com a mesma percentagem de distribuição (percentis). Para isso, deve-se usar *Make Cutpoints* e preencher a janela que se abre.

Métodos Estatísticos em Saúde - FMUP

Folha de exercícios no. 3:

1. Foi registado o peso de 24 alunas da UP seleccionadas aleatoriamente, tendo-se obtido os seguintes resultados:

62	72	51	53	62	49	55	56
58	52	48	72	89	72	66	97
55	52	50	61	50	74	66	62

   (a) Construa um ficheiro no **SPSS** com estes dados.

   (b) Calcule a média, a mediana, a variância e o desvio-padrão da amostra (apresente o resultado com as casas decimais adequadas).

   (c) Multiplique todos os valores da amostra por 2. O que acontece a cada uma das estatísticas amostrais calculadas em (a)?

   (d) Some 5 Kg a cada um dos pesos obtidos. O que acontece a cada uma das estatísticas amostrais calculadas em (a)?

   (e) Construa o boxplot para a variável peso, confirme que está bem construído e que existe um outlier nos dados.

   (f) Substitua a observação correspondente ao outlier por um qualquer valor que não seja um outlier e volte a avaliar a média e a mediana. Comente.

2. Abra o ficheiro bp.obese.sav relativo a um estudo sobre obesidade e pressão arterial conduzido numa cidade da Califórnia, EUA. As variáveis presentes são as seguintes:

   – **sex**: factor com 2 níveis que indica o género de cada indivíduo: 0-masculino; 1-feminino.

   – **obese**: vector numérico que representa a razão entre o peso actual e o peso ideal (obtido da tabela "New York Metropolitan Life Tables")

   – **bp**: vector numérico que representa a pressão arterial sistólica (mm Hg).

   (a) De quantas observações consta o estudo?

   (b) Confirme que cada uma das 3 variáveis está correctamente identificada pelo software (caso contrário proceda às alterações necessárias).

   (c) Apresente, numérica e graficamente, uma descrição dos dados correspondentes à variável sexo.

   (d) Para a variável contínua que representa os valores da pressão arterial sistólica, determine:

      (d.1) o tamanho da amostra

      (d.2) as estatísticas de localização da amostra

      (d.3) o máximo e o mínimo

      (d.4) o desvio-padrão

      (d.5) o erro padrão da média

      (d.6) todos os quartis e os percentis 5, 25, 75 e 95.

      (d.7) o número de observações entre $\overline{\mu} - 2\overline{\sigma}$ e $\overline{\mu} + 2\overline{\sigma}$, e a média dessas observações

      (d.8) um histograma, o boxplot e o qqplot.

(e) Apenas com base nos resultados da alínea anterior, comente a normalidade da variável **bp** na população.

(f) A partir de análises gráficas adequadas, apresente medidas descritivas para a v.a. bp nas mulheres.

(g) O que significa dizer que o percentil 25 (amostral) de **bp** é 116 mmHg? Confirme esse valor.

Discretize a variável obese em duas categorias, correspondentes aos indivíduos com peso superior ao ideal e aos indivíduos com peso igual ou inferior ao ideal. Construa uma *tabela de contingência* para essa nova variável e para o sexo, *i.e.*, uma tabela 2×2 que contenha as frequências absolutas de todas as combinações das duas variáveis.

(h) Apresente gráficos e medidas numéricas descritivos da variável obese nos homens.

(i) Obtenha o gráfico de dispersão da pressão arterial em função da obesidade que representa com cores distintas as observações masculinas e femininas. Confirme que em ambos os sexos se tem uma tendência crescente.

(j) Indique a média e o desvio padrão da pressão arterial de acordo com o sexo dos indivíduos. Qual das duas amostras tem menor erro padrão da média? E maior coeficiente de variação?

(k) Construa histogramas e boxplots correspondentes aos valores da pressão arterial em homens e em mulheres.

(l) Efectue cálculos que confirmem o boxplot obtido para os homens na alínea anterior (pode obter os quartis do SPSS).

3. A tabela seguinte diz respeito aos valores de colesterol (mg/dL) de uma amostra de 24 trabalhadores de um hospital americano. Os trabalhadores concordaram em seguir uma dieta vegetariana durante um mês, e os seus níveis de colesterol foram medidos antes da dieta ter tido início e depois dela ter terminado.

Ind.	Antes	Depois
1	195	146
2	145	155
3	205	178
4	159	146
5	244	208
6	166	147
7	250	202
8	236	215
9	192	184
10	224	208
11	238	206
12	197	169
13	169	182
14	158	127
15	151	149
16	197	178
17	180	161
18	222	187
19	168	176
20	168	145
21	167	154
22	161	153
23	178	137
24	137	125

(a) Determine a variação média dos níveis de colesterol.

(b) Determine o desvio padrão da variação dos níveis de colesterol.

(c) Construa o histograma e o boxplot correspondentes à variação dos níveis de colesterol.

(d) Determine a mediana da variação dos níveis de colesterol.

(e) Com base nas respostas anteriores, comente a simetria da distribuição dos valores de variação do colesterol.

(f) Apresente os boxplots correspondentes aos valores de colesterol antes e depois da dieta.

(g) Alguns investigadores afirmam que as diferenças entre os níveis de colesterol numa experiência deste tipo são mais sentidas nas pessoas com níveis de colesterol *mais altos*.

Separe as observações em dois grupos, de acordo com a relação de ordem entre os valores de colesterol antes da dieta e a sua mediana, e comente a afirmação acima (com recurso apenas a elementos de estatística descritiva).

4. A idade (em anos) de uma amostra de 20 indivíduos que morreram em acidentes de viação numa determinada auto-estrada é apresentada em seguida:

18	41	24	28	71
52	15	20	21	31
16	24	33	44	20
24	16	64	24	32

(a) Construa um histograma e um boxplot para os dados. A distribuição (empírica) é simétrica ou assimétrica (neste caso, para que lado)?

(b) Determine a amplitude, a amplitude interquartil e o desvio padrão dos dados.

(c) Apresente medidas estatísticas descritivas adequadas a estas observações.

5. A tabela abaixo mostra a altura (em cm) de 12 pais e dos seus filhos.

Altura do pai (cm)	Altura do filho (cm)
190	189
184	186
183	180
182	179
179	187
178	184
175	183
174	171
170	170
168	178
165	174
164	165

(a) Compare a altura dos pais e dos filhos, usando medidas descritivas numéricas e gráficas apropriadas.

(b) Construa o gráfico de dispersão de altura dos pais *versus* altura dos filhos e analise o tipo de relação existente entre as duas variáveis.

6. Pretendeu-se estudar a estabilidade física de uma formulação ao longo do tempo e a duas temperaturas distintas (20 ºC e 40ºC). O parâmetro firmeza da formulação foi avaliado após a preparação da formulação e mensalmente durante um período de 3 meses. Cada determinação foi efectuada em quadruplicado.

	firmeza			
Temp. (ºC)	20			
Tempo (meses)	0	1	2	3
	.599	.473	.425	.500
	.577	.479	.441	.483
	.557	.429	.477	.664
	.593	.381	.431	.454
Temp. (º C)	40			
Tempo (meses)	0	1	2	3
	.599	.321	.330	.398
	.571	.308	.328	.373
	.557	.393	.325	.369
	.593	.349	.339	.382

(a) Determine medidas descritivas numéricas adequadas para os resultados obtidos:

(a.1) à temperatura de 20 ºC

(a.2) à temperatura de 40 ºC

(a.3) após a preparação da formulação

(a.4) após a preparação da formulação e à temperatura de 20 ºC

(a.5) no 3º mês e à temperatura de 20 ºC

(b) Compare os resultados obtidos para as diferentes temperaturas (recorrendo também a gráficos adequados).

(c) Apenas recorrendo a gráficos adequados, o que pode conjecturar acerca de

(c.1) simetria e normalidade da distribuição (empírica) dos valores de firmeza

(c.2) simetria e normalidade da distribuição (empírica) dos valores de firmeza à temperatura de 20ºC.

7. Num estudo sobre a actividade proliferativa de um certo tipo de cancro, uma equipa de investigadores usou métodos imunohistoquímicos e o método do anticorpo monoclonal Ki-67.

(a) Num grupo de 30 pessoas com a doença os valores obtidos para o Ki-67 (expressos em percentagens) foram os seguintes:

10.1	10.8	10.5	14.7	10.5
10.1	5.4	25.1	32.6	41.1
17.9	16.6	4.4	25.1	16.6
32.6	10.8	26.3	32.6	41.1
9.3	9.3	7.4	9.3	14.7
21.4	25.1	10.5	17.9	41.1

(a.1) Determine a função de distribuição empírica e construa o seu gráfico.

(a.2) Quantos indivíduos apresentam medições superiores a 20?

(a.3) Determine o 0.90-quantil e o 0.25-quantil.

(a.4) Construa o box-plot associado às medições apresentadas e caracterize a simetria da distribuição.

(a.5) Apresente um histograma representativo dos dados.

(a.6) Determine medidas numéricas descritivas adequadas aos dados.

(a.7) Indique os valores do desvio padrão e do erro padrão da média.

(a.8) Construa o gráfico dos quantis (Q-Q plot) relativamente à distribuição normal. Que conclusão tira?

(a.9) Considere a variável discreta com três categorias definida por

$$\text{Ki67} < 10.0, \quad 10.0 \leq \text{Ki67} < 30.0, \quad \text{Ki67} \geq 30.0.$$

Descreva grafica e numericamente esta nova variável.

(b) A experiência foi repetida num outro grupo de 30 pessoas com a doença. Os valores obtidos para o Ki-67 neste novo grupo foram os seguintes:

28.3	19.5	15.9	15.9	7.1
4.6	63.0	17.8	10.9	2.8
28.3	11.9	28.3	4.5	17.8
2.8	28.3	11.2	52.3	6.4
11.2	28.3	9.8	4.6	4.6
7.1	11.2	35.4	52.3	43.5

(b.1) Determine medidas descritivas numéricas e gráficas que descrevam esta nova amostra.

(c) Qual das duas amostras apresenta maior coeficiente de variação? Interprete.

8. Uma equipa de investigadores mediu a quantidade total de proteínas (μg/ml) em amostras de lavagens broncoalveolares de 80 indivíduos (incluindo indivíduos com rinites alérgicas, indivíduos com asma e indivíduos normais). Os resultados obtidos foram os seguintes:

76.33	57.73	74.78	100.36	73.50
77.63	88.78	77.40	52.25	62.20
101.20	86.24	57.90	72.10	67.20
54.39	54.05	91.47	62.32	44.67
55.47	95.06	71.50	73.53	57.68
52.60	114.30	61.70	47.23	95.33
78.13	53.07	106.00	35.90	55.05
85.40	72.30	61.10	72.30	79.55

(a) Qual é a proporção de observações superiores ou iguais a 75?

(b) Existem outliers na amostra? Em caso afirmativo identifique-os (mencionando também o critério de identificação desses indivíduos).

(c) Descreva a amostra sob o ponto de vista estatístico (usando medidas numéricas e gráficas).

9. Foi recolhida uma amostra aleatória de 5 professores e 5 alunos da Faculdade de Psicologia da Universidade de Sussex e recolhida a informação que consta na tabela seguinte: profissão (aluno ou professor), número de amigos, consumo de álcool, rendimento anual (em libras) e nível de neuroticismo medido numa escala adequada na qual valores mais alto indicam indivíduos mais neuróticos.

Id	Prof	Amigos	Álcool	Rend	Neuro
1	P	5	10	20000	10
2	P	2	15	40000	17
3	P	0	20	35000	14
4	P	4	5	22000	13
5	P	1	30	50000	21
6	A	10	25	5000	7
7	A	12	20	100	13
8	A	15	16	3000	9
9	A	12	17	10000	14
10	A	17	18	10	13

(a) Construa no **SPSS** um ficheiro adequado a estes dados.

(b) Classifique cada uma das v.a. que constam da tabela e inclua essa informação no ficheiro que construiu em (a). Atribua ainda designações adequadas aos níveis dos factores - isto será importante para as análises gráficas posteriores.

(c) Averigue, graficamente, a normalidade e a simetria das variáveis Álcool, Rend e Neuro. De acordo com os resultados, descreva quantitativamente essas variáveis na amostra.

(d) Apresente um gráfico que indique que o rendimento dos professores é superior ao rendimento dos alunos.

(e) Suponha que se pretende comparar o número médio de amigos, consumo médio de álcool e nível médio de neuroticismo entre alunos e professores. O que é que consegue afirmar a esse respeito, baseado apenas em medidas descritivas e gráficos adequados.

(f) Defina uma nova variável **Neuro.c** que vale 1 sempre que Neuro$\leq$ 10, vale 2 sempre que Neuro$>$ 10 e Neuro$<$ 15, e vale 3 sempre que Neuro$\geq$ 15. Estude a relação entre a profissão dos indivíduos e a sua classificação em Neuro.c através de tabelas e gráficos adequados.

Chapter 4

Estatísticas Amostrais

4.1. Estatísticas

Suponhamos que temos uma amostra aleatória

$$x_1, x_2, \ldots, x_n$$

de tamanho n de uma v.a. X com função densidade de probabilidade f desconhecida. Em seguida descrevemos alguns problemas que podem ser colocados, uma vez descrita a amostra.

Podemos estar interessados em usar os valores observados para:

- Estimar a média μ de X.
- Estimar a variância σ^2 de X.
- Estimar uma região que, com probabilidade $\geq 1-\alpha$, contenha a média de X.
- ...
- Estimar a distribuição ou a função densidade de probabilidade de X.
- Calcular uma banda que, com probabilidade $\geq 1-\alpha$, contenha o "gráfico" da distribuição de probabilidade de X.

Se soubermos que a função densidade de probabilidade f pertence a uma família $f(x,\theta)$ de funções densidade de probabilidade que dependem, para completa especificação, de um parâmetro $\theta \in \Theta$ podemos usar os valores observados para:

- Estimar o valor mais "provável" do parâmetro θ da distribuição que gerou as observações.
- Estimar uma região que, com probabilidade $\geq 1-\alpha$, contenha o parâmetro θ da distribuição de X.
- Estimar a dependência de θ de variáveis exógenas (explicativas) $X_1, \ldots, X_p$.

ou para decidir se

- A distribuição de probabilidade tem parâmetro $\theta \in \Theta_0$.

Para resolver os problemas anteriores usamos funções

$$T(x_1, x_2, \ldots, x_n)$$

dos valores observados e do tamanho n da amostra. Como a mesma distribuição da variável X poderia ter gerado outra amostra de tamanho n,

$$x'_1, x'_2, \ldots, x'_n,$$

a função anterior calculada nesta amostra teria o valor

$$T(x'_1, x'_2, \ldots, x'_n),$$

não necessariamente igual ao anterior. Fazendo variar $x = (x_1, \ldots, x_n)$ entre todas as possíveis amostras de tamanho n que podem ser extraídas a partir da distribuição de X obtem-se uma nova variável aleatória

$$T(X_1, X_2, \ldots, X_n).$$

a que chamamos **estatística amostral**.

4.2. Lei dos Grandes Números

Teorema 4.2.1 (Lei dos Grandes Números) *Sejam $X_1, X_2, \ldots$ variáveis aleatórias (discretas ou contínuas) mutuamente independentes, cada uma com média μ e variância σ^2 finitas. Então se*

$$S_n = X_1 + X_2 + \cdots + X_n \qquad (n = 1, 2, \ldots)$$

tem-se, para qualquer $\varepsilon > 0$,

$$\lim_{n\to+\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0.^{a}$$

Algumas observações:

(a) A variável aleatória

$$\frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n}$$

é designada por **média aritmética** das variáveis aleatórias $X_1, \ldots, X_n$.

Em particular, se $X_1, \ldots, X_n$ é uma a.a., $\frac{S_n}{n}$ representa a (v.a.) **média amostral**.

(b) A lei dos grandes números não exige que as v.a. $X_1, X_2, \ldots$ sejam identicamente distribuídas.

[a] Diz-se que $\frac{S_n}{n}$ **converge para μ em probabilidade**.

(c)
$$\begin{aligned} E\left(\frac{S_n}{n}\right) &= \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}\left(E(X_1) + E(X_2) + \cdots + E(X_n)\right) \\ &= \frac{1}{n}(\mu + \mu + \ldots \mu) \\ &= \frac{1}{n}(n\mu) \\ &= \mu \end{aligned}$$

portanto o valor esperado da v.a. média aritmética das v.a. $X_1, \ldots, X_n$ é μ.

(d) O teorema afirma que, para qualquer $\varepsilon > 0$, à medida que n tende para $+\infty$, a probabilidade de a média aritmética S_n/n diferir do seu valor esperado μ mais de ε tende para zero.

(e) O teorema <u>não afirma</u> que

$$\lim_{n\to+\infty} \frac{S_n}{n} = \mu.$$

Esse resultado é mais forte do que a lei dos grandes números e é, na realidade, falso.

(f) Seja $X_1, \ldots, X_n$ uma a.a. de tamanho n de uma qualquer população com média μ e variância finita. Pela lei dos grandes números, a média amostral vai-se tornando cada vez mais próxima da média populacional à medida que o tamanho da amostra aumenta.

Diz-se que a média amostral é um **estimador consistente** da média populacional.

Exemplo:
Suponhamos que a probabilidade de obter "cara" num lançamento de uma moeda é p $(0 < p < 1)$ e que essa moeda é lançada independentemente n vezes. Como é que a partir das observações se consegue estimar p?

Intuitivamente, diríamos que uma boa estimativa para p será dada pelo quociente

$$\frac{\text{número de caras obtidas em } n \text{ lançamentos}}{n}. \quad (4.1)$$

O que garante a veracidade da nossa intuição é precisamente a lei dos grandes números.
Mais concretamente, considerem-se as v.a. X_i $(i = 1, \ldots, n)$ definidas por

$$X_i = \begin{cases} 1, & \text{se sai cara no lançamento } i \\ 0, & \text{se não sai cara no lançamento } i \end{cases}$$

Tem-se

$$E(X_i) = 1(p) + 0(1-p) = p$$

e

$$\mathrm{Var}(X_i) = (1-p)^2 p + (0-p)^2(1-p) = p(1-p)$$

(estes resultados também se obtêm notando que $X_i \sim B(1, p)$).
Ora o quociente em (4.1) pode agora escrever-se na forma

$$\frac{X_1 + \cdots + X_n}{n}$$

e portanto, pela lei dos grandes números, podemos afirmar que (4.1) converge em probabilidade para p.

A lei dos grandes números enunciada atrás é usualmente designada por "lei fraca dos grandes números". Tem funcionado como pilar da estatística frequencista, por mostrar que a frequência relativa com que ocorre um acontecimento converge para a probabilidade de ocorrência desse acontecimento.

4.3. Teorema do Limite Central

As relações entre a distribuição binomial e a normal, e a de Poisson e a normal, levantam naturalmente a questão da existência ou não de mais distribuições que tenham a distribuição normal como limite. O teorema seguinte afirma que existem de facto muitas distribuições com essa propriedade.

Teorema 4.3.1 (Teorema do Limite Central) *Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes e identicamente distribuídas* [a]*, com média μ e variância σ^2 finitas. Definindo*

$$S_n = X_1 + X_2 + \cdots + X_n \qquad (n = 1, 2, \ldots)$$

tem-se, para valores fixos $a < b$,

$$\lim_{n\to+\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du,$$

i.e.,

(a) $S_n \sim N(n\mu, n\sigma^2)$ *assimptoticamente* [b]

(b) $\overline{X} = \frac{S_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ *assimptoticamente*

(c) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ *assimptoticamente.*

[a] *i.e.*, têm todas a mesma f.d.p.

[b] Trata-se de uma **convergência em distribuição**, que significa que, à medida que o n aumenta, a distribuição de S_n vai-se tornando cada vez mais próxima da distribuição normal $N(n\mu, n\sigma^2)$.

Observações:

(a) Como $X_1, X_2, \ldots X_n$ têm a mesma média,

$$\begin{aligned} E(S_n) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= \mu + \mu + \cdots + \mu \\ &= n\mu \end{aligned}$$

e como são independentes e têm variâncias iguais,

$$\begin{aligned} \mathrm{Var}(S_n) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= \sigma^2 + \sigma^2 + \cdots + \sigma^2 \\ &= n\sigma^2 \end{aligned}$$

Mostrar que a a distribuição de S_n é assimptoticamente normal é a parte difícil do teorema.

(b) Pelo teorema, a v.a. média aritmética dos X_i´s

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

segue assimptoticamente uma distribuição normal $N(\mu, \sigma^2/n)$, *i.e.*,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{assimptoticamente.}$$

De facto:

(i) $E\left(\overline{X}\right) = E\left(\frac{S_n}{n}\right) = \frac{1}{n}E(S_n) = \frac{1}{n}(n\mu) = \mu$

(ii) $\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{\sigma^2}{n}$.

(iii) a distribuição assimptoticamente normal é garantida pelo teorema do limite central.

(c) Pode-se provar (resultado muito mais simples do que o teorema do limite central) que qualquer combinação linear de v.a. independentes seguindo distribuições normais é uma v.a. que segue ainda uma distribuição normal.

Em particular, se $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$, então a v.a. $\overline{X}$ segue exactamente uma distribuição normal:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

(d) Existem generalizações do teorema enunciado anteriormente que relaxam as suas hipóteses - dependência fraca em vez de independência, variáveis aleatórias de diferentes tipos e não necessariamente i.i.d. - continuando a observar convergência para uma distribuição normal.

(e) O teorema do limite central não refere a velocidade de convergência da soma para a distribuição normal, em função do n. Assim, é usual tomar-se $n \geq 30$ como regra prática para aceitar a aproximação pela distribuição normal mas várias situações muito diferentes podem surgir. Por exemplo: [a]

(i) a soma de 3 distribuições uniformes é praticamente gaussiana

(ii) a soma de um milhão de Poissons i.i.d. com parâmetro $\lambda = 10^{-27}$ é também uma Poisson, com parâmetro $10^6 10^{-27} = 10^{-21}$ e está muito longe da gaussiana.

Os mesmos autores referem que se as v.a. X_i não forem pronunciadamente assimétricas, a convergência da soma S_n para a distribuição normal é rápida.

(f) Veremos que o teorema do limite central permitirá obter distribuições para as estatísticas amostrais mais comuns no caso de amostras grandes, e que isso será de grande utilidade na determinação dos intervalos de confiança e das estatísticas de teste a usar nos testes de hipóteses. Por um lado, o teorema implica que muitos procedimentos que sejam deduzidos para modelos normais não sejam muito sensíveis à condição de normalidade; por outro, implica que muitas v.a. contínuas (como o peso, o coeficiente de inteligência CI, ...), que podem ser vistas como médias de vários pequenos efeitos, possam ser aproximadamente normalmente distribuídas. Na verdade, o teorema do limite central pode funcionar como uma primeira justificação para o uso tão frequente da distribuição normal em estatística.

(g) Para uma visualização gráfica do TLC ver o Applet CLT em

http://sciris.shu.edu/thinklets/Math/Statistics/CLT/clt.html

[a] Pestana, Velosa, *Introdução à Probabilidade e à Estatística*, 2008.

Exemplo 1:
Uma máquina calculadora soma 1200 números reais após arredondamento para o inteiro mais próximo. Sabendo que o erro cometido em cada arredondamento é uma variável aleatória com distribuição $U\left(-\frac{1}{2}, \frac{1}{2}\right)$, determine a probabilidade aproximada de que o erro na soma exceda 10 unidades.

Resolução: Para $i = 1, 2, \ldots, 1200$, seja E_i a v.a. que representa o erro cometido no i-ésimo arredondamento. Então, para $i = 1, 2, \ldots, 1200$, $E_i \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$ e

$$E(E_i) = 0, \qquad \mathrm{Var}(E_i) = \frac{1}{12}.$$

Se

$$S = \sum_{1=1}^{1200} E_i$$

então, pelo teorema do limite central,

$$E(S) = 1200 \times 0 = 0, \ \mathrm{Var}(S) = 1200 \times \frac{1}{12} = 100$$

e

$$S \sim N(0, 100) \quad \text{assimptoticamente.}$$

Assim

$$\begin{aligned} P(S > 10) &\approx 1 - P(S \leq 10) \\ &= 1 - P\left(Z \leq \frac{10 - 0}{\sqrt{100}}\right) \\ &= 1 - P(Z \leq 1) \\ &= 0.1587. \end{aligned}$$

Exemplo 2:
Os tempos (em minutos) gastos por um caixa no processamento de cada ordem pessoal são variáveis aleatórias independentes com média 1.5 minutos e desvio-padrão 1 minuto. Qual é a probabilidade aproximada de que as ordens de 100 clientes possam ser processadas em menos de 2 horas?

Resolução: Para $i = 1, 2, \ldots, 100$, seja T_i a v.a. que representa o tempo (em minutos) gasto pelo caixa com a ordem da pessoa i. Por hipótese,

$$E(T_i) = 1.5, \qquad \mathrm{Var}(T_i) = 1.$$

Seja

$$T = \sum_{i=1}^{100} T_i.$$

Pelo teorema do limite central,

$$E(T) = 100 \times 1.5 = 150, \quad \mathrm{Var}(T) = 100 \times 1 = 100$$

e

$$T \sim N(150, 100) \quad \text{assimptoticamente.}$$

Portanto a probabilidade aproximada de que as ordens de 100 clientes possam ser processadas em menos de 2 horas (120 minutos) é

$$\begin{aligned} P(T \leq 120) &= P\left(Z \leq \frac{120 - 150}{10}\right) \\ &= P(Z \leq -3) \\ &= P(Z \geq 3) \\ &= 1 - P(Z \leq 3) \\ &= 1 - 0.9987 = 0.0013. \end{aligned}$$

Exemplo 3:
Sejam $X_1, \ldots, X_n$ v.a. independentes, cada uma delas seguindo a distribuição binomial $B(1,p)$. Em particular, tem-se

$$E(X_i) = p \quad \text{e} \quad \mathrm{Var}(X_i) = p(1-p)$$

e pode-se provar que

$$X = X_1 + \cdots + X_n \sim B(n,p).$$

Pelo teorema do limite central,

$$X \sim N\left(np, np(1-p)\right) \quad \text{assimptoticamente}$$

portanto *para n grande*, a distribuição binomial $B(n,p)$ pode ser aproximada pela distribuição normal $N(np, np(1-p))$.
Este resultado é conhecido como **teorema de DeMoivre-Laplace**.

Exemplo 4:
Sejam $X_1, \ldots, X_n$ v.a. independentes, cada uma delas seguindo a distribuição de Poisson $P(\eta)$. Em particular, tem-se

$$E(X_i) = \eta \quad \text{e} \quad \mathrm{Var}(X_i) = \eta$$

e vimos já que

$$X = X_1 + \cdots + X_n \sim P(n\eta).$$

Pelo teorema do limite central,

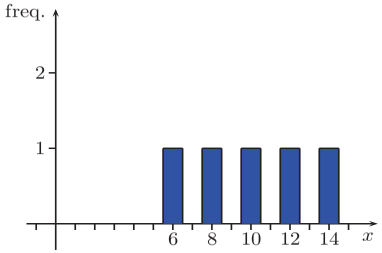
$$X \sim N(n\eta, n\eta) \quad \text{assimptoticamente}$$

portanto, e com a mudança de parâmetro $\lambda = n\eta$ para simplicação, *para λ grande*, a distribuição de Poisson $P(\lambda)$ pode ser aproximada pela distribuição normal $N(\lambda, \lambda)$.

4.4. Distribuição da média amostral num exemplo concreto

Considere-se a população de tamanho $N = 5$ que consiste das idades de cinco crianças que durante o ano de 2004 deram entrada no serviço de pediatria de um determinado hospital com problemas de envenenamento. Seja X a variável aleatória que representa as idades dessas crianças e suponhamos que esses valores são 6, 8, 10, 12 e 14 anos. A função densidade de probabilidade de X é $f : \{6, 8, 10, 12, 14\} \to \mathbb{R}$

$$f(6) = f(8) = f(10) = f(12) = f(14) = \frac{1}{5}.$$



A média μ e a variância σ^2 de X são, respectivamente,

$$\mu = E(X) = \sum_x x f(x) = \frac{6+8+10+12+14}{5} = 10,$$

$$\sigma^2 = \mathrm{Var}(X) = \sum_x (x-\mu)^2 f(x) = \frac{\sum_x (x-10)^2}{5} = 8.$$

No que se segue pretende-se determinar a distribuição da v.a. média amostral,

$$\overline{X} = \frac{X_1 + X_2}{2},$$

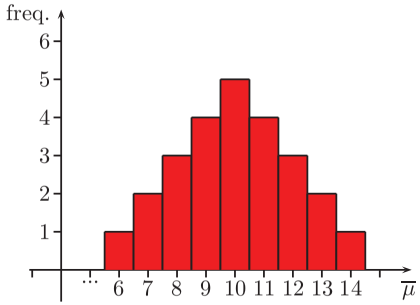
conseguida a partir de amostras aleatórias X_1, X_2 de tamanho $n = 2$ da população das cinco crianças.
Consideremos todas as amostras possíveis (realizadas com reposição) de tamanho $n = 2$ da população das cinco crianças. Essas amostras (juntamente com a média de cada uma, apresentada entre parênteses) estão representadas na tabela seguinte.

X_1: idade da 1a criança	X_2: idade da 2a criança				
	6	8	10	12	14
6	6,6 (6)	6,8 (7)	6,10 (8)	6,12 (9)	6,14 (10)
8	8,6 (7)	8,8 (8)	8,10 (9)	8,12 (10)	8,14 (11)
10	10,6 (8)	10,8 (9)	10,10 (10)	10,12 (11)	10,14 (12)
12	12,6 (9)	12,8 (10)	12,10 (11)	12,12 (12)	12,14 (13)
14	14,6 (10)	14,8 (11)	14,10 (12)	14,12 (13)	14,14 (14)

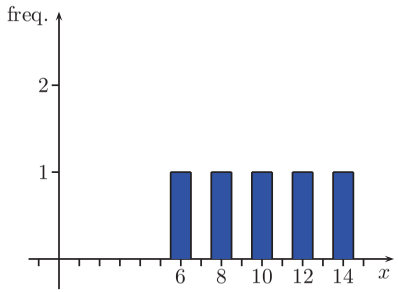
Repare-se que existem 25 amostras possíveis com 2 elementos. Em geral, para uma população com N elementos, se a amostragem for feita com reposição, o número de todas as amostras possíveis de tamanho n é N^n.
Os valores obtidos para a variável aleatória $\overline{X}$ e as correspondentes frequências são:

$\overline{X}$	6	7	8	9	10	11	12	13	14
freq. abs.	1	2	3	4	5	4	3	2	1
freq. rel.	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

O histograma para a distribuição da média amostral $\overline{X}$ é portanto



Observe-se que este histograma é muito diferente do histograma obtido para a v.a. X das idades das crianças (correspondente a uma distribuição uniforme):



Considere-se a variável aleatória média amostral, $\overline{X}$. A sua média é

$$\begin{aligned} E(\overline{X}) &= 6\frac{1}{25} + 7\frac{2}{25} + 8\frac{3}{25} + 9\frac{4}{25} + 10\frac{5}{25} \\ &\quad + 11\frac{4}{25} + 12\frac{3}{25} + 13\frac{2}{25} + 14\frac{1}{25} \\ &= 10, \end{aligned}$$

isto é,

$$E(\overline{X}) = E(X) \quad (= 10) \tag{4.2}$$

Relativamente à variância de $\overline{X}$ tem-se

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= (6-10)^2\frac{1}{25} + (7-10)^2\frac{2}{25} + (8-10)^2\frac{3}{25} \\ &\quad + (9-10)^2\frac{4}{25} + (10-10)^2\frac{5}{25} + (11-10)^2\frac{4}{25} \\ &\quad + (12-10)^2\frac{3}{25} + (13-10)^2\frac{2}{25} + (14-10)^2\frac{1}{25} \\ &= 4. \end{aligned}$$

Em particular tem-se a relação

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} \quad \left(= \frac{8}{2}\right) \tag{4.3}$$

isto é, a variância da média amostral é igual à variância da população a dividir pelo tamanho da amostra. Repare-se ainda, a partir do histograma de $\overline{X}$, que a média amostral parece seguir uma distribuição normal.

As igualdades (4.2) e (4.3) e o facto da distribuição de $\overline{X}$ ser normal não são meras coincidências. Pelo contrário, verificam-se sempre que a população a ser amostrada segue uma distribuição normal ou, caso isso não se verifique, sempre que o tamanho da amostra for suficientemente grande (teorema do limite central). Na situação analisada nenhuma das condições anteriores é verificada mas ainda assim o caso foi escolhido pela simplicidade dos cálculos envolvidos e pela boa elucidação de conceitos a que conduz.

4.5. Alguma teoria sobre estimadores pontuais*

As definições e resultados apresentados nesta secção foram obtidos de

- Pestana, Velosa, *Introdução à Probabilidade e à Estatística*, 2008.
- Arnold, *Mathematical Statistics*, 1990.

Começamos com algumas definições.
Uma **estatística** é uma função das amostras aleatórias que não envolve parâmetros desconhecidos.
Um **estimador** é uma estatística cujo contradomínio é o espaço dos parâmetros. Por outras palavras, a cada amostra observada faz corresponder uma estimativa (uma avaliação) dos parâmetros desconhecidos.
Um **estimador pontual** estima um parâmetro desconhecido declarando um único valor. Exemplos de estimadores não pontuais incluem os estimadores intervalares, que estimam intervalos para o parâmetro em causa.

Seja $\hat{\theta} = \hat{\theta}(X)$ um estimador pontual do parâmetro θ. O estimador $\hat{\theta}$ é **centrado** se e só se

$$E(\hat{\theta}) = \theta.$$

Um estimador não centrado diz-se **enviesado**, e o seu **viés** (*bias*) é definido por

$$B = E(\hat{\theta}) - \theta.$$

Seja $\hat{\theta}_n$ um estimador de θ obtido a partir de uma amostra de dimensão n.
O estimador $\hat{\theta}_n$ diz-se **consistente** se

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{quando} \quad n \to +\infty^{a}$$

Em termos práticos, significa que observando mais unidades amostrais estamos a obter estimativas que provavelmente estarão mais próximas do verdadeiro valor do parâmetro.

Um estimador $\widehat{\theta}$ de θ diz-se **eficiente** se possui a menor variância de entre todos os estimadores não-enviesados de θ. Isto é,

(i) $\widehat{\theta}$ é um estimador não-enviesado, $E(\widehat{\theta}) = \theta$

(ii) $\text{var}(\widehat{\theta}) \leq \text{var}(\widetilde{\theta})$, para qualquer estimador não-enviesado $\widetilde{\theta}$ de θ.

---

[a] **Convergência em probabilidade:**
Seja $X_1, X_2, \ldots$ uma sucessão de v.a., cada uma com função distribuição $F_n = F_n(x)$, $n = 1, 2, \ldots$.
Diz-se que X_n *converge em probabilidade* para b ($b \in \mathbb{R}$), e escreve-se $X_n \xrightarrow{P} b$ se

$$\lim_{n \to +\infty} F_n(x) = \begin{cases} 0, & x < b \\ 1, & x > b \end{cases}$$

Pode-se provar que $X_n \xrightarrow{P} b$ se e só se, para qualquer $\varepsilon > 0$,

$$P(|X_n - b| > \varepsilon) \to 0, \quad \text{quando } n \to +\infty.$$

Teorema 4.5.1 *Seja $X_1, X_2, \ldots$ uma sucessão de v.a. com $E(X_n) = \mu_n$ e $\mathrm{Var}(X_n) = \sigma_n^2$. Se $\mu_n \to a$ e $\sigma_n^2 \to 0$, quando $n \to +\infty$, então*

$$X_n \xrightarrow{P} a \qquad \text{quando } n \to +\infty.$$

Demonstração: Tem-se

$$\begin{aligned}
&E\left((X_n - a)^2\right) \\
&= E\left((X_n - \mu_n + \mu_n - a)^2\right) \\
&= E\left((X_n - \mu_n)^2 + (\mu_n - a)^2 \right. \\
&\qquad \left. + 2(X_n - \mu_n)(\mu_n - a)\right) \\
&= \sigma_n^2 + (\mu_n - a)^2
\end{aligned}$$

dado que $E(X_n - \mu_n) = E(X_n) - \mu_n = 0$. Portanto, pela desigualdade de Markov, [a]

$$\begin{aligned}
0 &\leq P\left(|X_n - a| > \varepsilon\right) \\
&= P\left((X_n - a)^2 > \varepsilon^2\right) \\
&\leq \frac{E\left((X_n - a)^2\right)}{\varepsilon^2} \\
&= \frac{\sigma_n^2 + (\mu_n - a)^2}{\varepsilon^2} \\
&\to 0.
\end{aligned}$$

[a] **Desigualdade de Markov**: Seja X uma v.a. que toma apenas valores não negativos. Para qualquer $t > 0$ tem-se

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Resulta do teorema 4.5.1 que a convergência usual numérica implica convergência em probabilidade. Mais precisamente,

Corolário 4.5.2 *Seja $(a_n)_{n\in\mathbb{N}}$ uma sucessão numérica de números reais convergente para $b \in \mathbb{R}$, i.e., $\lim_{n\to\infty} a_n = b$. Então a_n converge também em probabilidade para b,*

$$a_n \xrightarrow{P} b.$$

Demonstração: Considere-se a sucessão de v.a. X_n definidas por $X_n = a_n$. Tem-se que, para qualquer $n \in \mathbb{N}$, $E(X_n) = a_n \to b$ e $\mathrm{Var}(X_n) = 0$ pelo que, pelo teorema anterior, $X_n = a_n \xrightarrow{P} b$.

Definição 4.5.3 *Seja T_n uma sucessão de estimadores de $\tau(\theta)$, onde τ representa uma função real de variável real e θ uma parâmetro populacional. ($\tau(\theta)$ é muitas vezes referido como um parâmetro de dimensão 1). Diz-se que a sucessão T_n é* **consistente** *se T_n converge em probabilidade para $\tau(\theta)$,*

$$T_n \xrightarrow{P} \tau(\theta).$$

O teorema que apresentamos a seguir dá-nos uma forma rápida e fácil de estabelecer a consistência de uma sucessão de estimadores para muitos modelos. A sua demonstração resulta directamente do teorema 4.5.1.

Teorema 4.5.4 *Seja T_n uma sucessão de estimadores de um parâmetro $\tau(\theta)$ de dimensão 1.*

(a) Se vies$(T_n) \to 0$ e $\mathrm{Var}(T_n) \to 0$, então T_n é uma sucessão de estimadores consistentes.

(b) se T_n é uma sucessão consistente de $\tau(\theta)$ e $h(t)$ é uma função contínua, então $h(T_n)$ é uma sucessão consistente de $h(\tau(\theta))$.

Definição 4.5.5 *Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$, um estimador $\widehat{\theta}$ de θ diz-se* **suficiente** *se a distribuição condicional (conjunta) de $X_1, X_2, \ldots, X_n$ dado $\widehat{\theta}$ não depende de θ. Formalmente,*

$$f(x_1, x_2, \ldots, x_n \,|\, \widehat{\theta}) = \frac{f(x_1, x_2, \ldots, x_n)}{g(\widehat{\theta})}.$$

Segundo Fisher, "um estimador suficiente é aquele que contém toda a informação contida na amostra, sendo desnecessário considerar qualquer outro estimador".

Por exemplo, o EMV é um estimador suficiente. A função de verosimilhança

$$\prod_{i=1}^{n} f(\theta; x_i)$$

pode ser escrita na forma

$$\prod_{i=1}^{n} f(\theta; x_i) = g(x_1, x_2, \ldots, x_n \,|\, \widehat{\theta})\, h(\widehat{\theta}, \theta)$$

onde a distribuição de $(X_1, X_2, \ldots, X_n)$ dado $\widehat{\theta}$ é independente do parâmetro θ e $h(\widehat{\theta}, \theta)$ é a distribuição conjunta de $\widehat{\theta}$ e θ, sendo $\widehat{\theta}$ o EMV de θ.

4.6. Algumas estatísticas amostrais mais comuns

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de tamanho n (> 1) de uma variável aleatória X associada a uma população infinita.

- **Média amostral:**

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Sendo $\overline{X}$ uma função da amostra aleatória $X_1, \ldots, X_n$, $\overline{X}$ é uma estatística. Por razões óbvias é designada por **média amostral**.

(1) Se $X \sim N(\mu, \sigma^2)$ então a distribuição de $\overline{\mu}$ é também normal com média e variância

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

respectivamente, *i.e.*,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{4.4}$$

(2) Sob condições muito gerais sobre a distribuição de X (não necessariamente normal) e quando o tamanho n da amostra é *grande* (é usual considerar-se para este efeito $n \geq 30$),

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1) \quad \text{(assimptoticamente)}.$$

Este resultado é uma consequência directa do teorema do limite central.

- **Variância amostral bruta:**

$$\widehat{\sigma}^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Esta variável aleatória tem média e variância respectivamente iguais a

$$E(S_n^2) = \frac{n-1}{n}\sigma^2, \quad \mathrm{Var}(S_n^2) = \frac{\mu_4 - 3\sigma^2}{n^3}.$$

Se $X \sim N(\mu, \sigma^2)$ então

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

- **Variância amostral:**

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

Como

$$S^2 = \frac{n}{n-1}S_n^2$$

a variável aleatória S^2 tem então média e variância respectivamente iguais a

$$E(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{n^2}{(n-1)^2}\mathrm{Var}(S_n^2).$$

Se $X \sim N(\mu, \sigma^2)$ então

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

- **Ratio de Student:**

  Relembre-se que, se

$$U \sim N(0,1) \quad \text{e} \quad V \sim \chi^2(n-1)$$

  e U e V são independentes, então

$$T = \frac{U}{\sqrt{V/(n-1)}} \sim t(n-1).$$

  Ora, se $X \sim N(\mu, \sigma^2)$ então

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

  e

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

  portanto

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \qquad (4.5)$$

  Repare-se que esta condição é diferente de (4.4). Em (4.4) a variância σ^2 da população é conhecida enquanto que em (4.5) a variância da população não é conhecida (e portanto usa-se a variância amostral S^2 para a estimar).

  Se X não segue uma distribuição normal mas o tamanho n da amostra é *grande*, a convergência de $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ para a distribuição $N(0,1)$ é assimptótica:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  (a convergência é garantida pelo facto de S ser um estimador consistente de σ)

- **Razão de variâncias amostrais**

  Relembre-se que, se

$$U \sim \chi^2(m-1) \quad \text{e} \quad V \sim \chi^2(n-1)$$

  então

$$F = \frac{U/(m-1)}{V/(n-1)} \sim F(m-1, n-1).$$

  Se $X_1, \ldots, X_m$ é uma amostra aleatória de tamanho m da variável aleatória X e $Y_1, \ldots, Y_n$ é uma amostra aleatória de tamanho n da variável aleatória Y, a razão de variâncias amostrais é

$$F = \frac{{S_X}^2/\sigma_X^2}{{S_Y}^2/\sigma_Y^2}$$

  Se $X \sim N(\mu_X, \sigma_X^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$ então

$$U = \frac{(m-1){S_X}^2}{\sigma_X^2} \sim \chi^2(m-1)$$

$$V = \frac{(n-1){S_Y}^2}{\sigma_Y^2} \sim \chi^2(n-1)$$

  e portanto

$$F = \frac{U/(m-1)}{V/(n-1)} = \frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F(m-1, n-1).$$

- **Proporção amostral**

  Seja $X_1, X_2, \ldots, X_n$ uma a.a. de tamanho n de uma população binomial: $X_i \sim B(1,p)$, para $i = 1, \ldots n$. A variável aleatória

$$Y = \sum_{i=1}^{n} X_i$$

  representa o **número de sucessos na amostra**, e tem-se (resultado teórico)

$$Y \sim B(n,p).$$

  Relembre-se que se p e q não forem muito próximos de zero e n for grande [a] então

$$B(n,p) \approx N(np, np(1-p))$$

  e portanto

$$\frac{Y - np}{\sqrt{np(1-p)}} \overset{a}{\sim} N(0,1).$$

  Para a variável aleatória

$$\overline{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{Y}{n},$$

  que representa a **proporção de sucessos na amostra**, ou a **proporção amostral**, o resultado anterior escreve-se

$$\frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0,1).$$

[a] É usual considerar $n \geq 30$ e $np(1-p) > 5$.

- **Diferença de proporções amostrais**

  Dadas duas a.a. de duas populações binomiais (independentes) com parâmetros (n_1, p_1) e (n_2, p_2), respectivamente,

$$X_{11}, X_{12}, \ldots, X_{1,n_1}, \qquad X_{21}, X_{22}, \ldots, X_{2,n_2}$$

  considerem-se as (v.a.) proporções amostrais correspondentes:

$$\overline{p}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}, \qquad \overline{p}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}.$$

  Se $n_1, n_2 \geq 30$ e $np_1(1-p_1), np_2(1-p_2) > 5$,

$$\begin{aligned} \overline{p}_1 &\overset{a}{\sim} N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \\ \overline{p}_2 &\overset{a}{\sim} N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right) \end{aligned}$$

  e que

$$\begin{aligned} E(\overline{p}_1 - \overline{p}_2) &= E(\overline{p}_1) - E(\overline{p}_2) = p_1 - p_2 \\ \mathrm{Var}(\overline{p}_1 - \overline{p}_2) &= \mathrm{Var}(\overline{p}_1) + \mathrm{Var}(\overline{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \end{aligned}$$

  Resulta então que (notando que a diferença de duas v.a. independentes e aproximadamente normais é uma v.a. aproximadamente normal)

$$\frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{a}{\sim} N(0,1).$$

Observação: O desvio padrão de $\overline{p}$ (resp. $\overline{p}_1$ e $\overline{p}_2$) contém o parâmetro p (em geral, desconhecido). Para efectuar cálculos (por exemplo, para deduzir intervalos de confiança e para realizar testes de hipóteses), é comum substituir p (resp. p_1, p_2) por $\overline{p}$ (resp. $\overline{p}_1$, $\overline{p}_2$), um seu estimador consistente. Despreza-se o erro cometido porque a substituição é feita para uma grande amostra e numa expressão que está dividida por um n grande. Assim

$$\frac{\overline{p}-p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \overset{a}{\sim} N(0,1),$$

$$\frac{(\overline{p}_1-\overline{p}_2)-(p_1-p_2)}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1}+\frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}} \overset{a}{\sim} N(0,1).$$

- **Diferença de médias amostrais (variâncias conhecidas)**

  Considerem-se duas a.a. $X_{11}, X_{12}, \ldots, X_{1,n_1}$ e $X_{21}, X_{22}, \ldots, X_{2,n_2}$, independentes, de duas variáveis aleatórias X_1 e X_2, respectivamente.

  (1) Se $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$, então sabe-se já que

  $$\overline{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \overline{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

  Tem-se também

  $$E(\overline{X}_1-\overline{X}_2) = E(\overline{X}_1)-E(\overline{X}_2) = \mu_1-\mu_2$$

  e, pela independência das amostras,

  $$\begin{aligned} \mathrm{Var}(\overline{X}_1-\overline{X}_2) &= \mathrm{Var}(\overline{X}_1)+\mathrm{Var}(\overline{X}_2) \\ &= \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}. \end{aligned}$$

  Dado que a soma de v.a. normais é ainda uma v.a. normal, resulta que

  $$\overline{X}_1-\overline{X}_2 \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right),$$

  isto é,

  $$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

(2) Quando o tamanho n_1 e n_2 das amostras é *grande* e as amostras não apresentam assimetrias *exageradas*, o teorema do limite central permite substituir as distribuições normais (exactas) anteriores por distribuições normais assimptóticas, obtendo-se

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{a}{\sim} N(0,1).$$

- **Diferença de médias amostrais (variâncias desconhecidas)**

  Considerem-se duas a.a. independentes X_{11}, X_{12}, ..., X_{1,n_1} e $X_{21}, X_{22}, \ldots, X_{2,n_2}$ de duas variáveis aleatórias X_1 e X_2, respectivamente.

  (1) Se $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$,

  $$\frac{\overline{X}_1 - \mu_1}{S_1/\sqrt{n}} \sim t(n_1 - 1), \quad \frac{\overline{X}_2 - \mu_2}{S_2/\sqrt{n}} \sim t(n_2 - 1)$$

  e

  $$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2.$$

  Contudo, a determinação da distribuição exacta da estatística

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  é um problema complexo - **problema de Behrens-Fisher**.

  Uma solução possível consiste em recorrer à aproximação de Welch-Satterwaite

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \simeq t(\nu)$$

  (distribuição **aproximada**) onde ν é o maior inteiro que não excede

  $$\nu' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}.$$

  (2) Quando o tamanho n_1 e n_2 das amostras é *grande* e as amostras não apresentam assimetrias *exageradas*, de forma análoga ao caso em que as variâncias eram conhecidas, pode-se concluir que

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} N(0,1).$$

  (substituíram-se as variâncias populacionais pelas amostrais e o erro é desprezável porque os estimadores são consistentes e as amostras são grandes.)

4.7. Intervalos de confiança

Seja X uma variável aleatória com média e variâncias designadas, respectivamente, por

$$E(X) = \mu \quad \text{e} \quad \text{Var}(X) = \sigma^2.$$

Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de tamanho n da variável aleatória X e $x_1, \ldots, x_n$ é uma realização dessa amostra aleatória, definimos a média amostral e a variância amostral associada a esta realização por

$$\begin{aligned} \overline{X}(x_1, \ldots, x_n) &= \frac{1}{n}(x_1 + \cdots + x_n) \\ \overline{\sigma}^2(x_1, \ldots, x_n) &= \frac{1}{n-1}[(x_1 - \overline{X})^2 + \cdots + (x_n - \overline{X})^2] \end{aligned}$$

respectivamente. Sempre que for implícito que estamos a trabalhar com a realização da amostra, usaremos a notação

$$\overline{X} \quad \text{e} \quad \overline{\sigma}^2$$

para designar $\overline{X}(x_1, \ldots, x_n)$ e $\overline{\sigma}^2(x_1, \ldots, x_n)$, respectivamente. Note-se que estas quantidades são *números* e não devem ser confundidas com as *variáveis aleatórias* $\overline{X}$ e $\overline{\sigma}^2$.

Suponhamos que se sabe que a função densidade de probabilidade f de X pertence a uma família $f(x;\theta)$ de densidades de probabilidade que dependem, para completa especificação, de um parâmetro $\theta \in \Theta$ (θ pode ser um único parâmetro ou um vector de parâmetros $\theta = (\theta_1, \ldots, \theta_k)$).
Podemos colocar o problema de usar os valores observados de uma amostra aleatória da população para:

- estimar o valor "mais provável" do parâmetro θ da distribuição que gerou as observações (teoria sobre **estimação pontual** ou sobre **testes de hipóteses**)
- estimar θ através de uma região que, de alguma forma, transmita informação sobre a provável magnitude de θ (teoria sobre **intervalos de confiança**).

No que se segue, consideramos que θ é a média da população e deduzimos pormenorizadamente um intervalo *aleatório* que contenha essa média com um determinado nível de confiança pré-especificado.
Os intervalos de confiança (centrados) para os restantes parâmetros populacionais mais usados são apresentados em seguida. Não se inclui a dedução dessas fórmulas, por resultarem das estatísticas amostrais correspondentes de forma absolutamente análoga ao que foi feito para a média.
A apresentação da teoria sobre testes de hipóteses será feita no final deste capítulo.

- **Intervalo de confiança para a média - σ^2 conhecida**

Como vimos, se $X \sim N(\mu, \sigma^2)$ a média amostral segue também uma distribuição normal, independentemente do tamanho da amostra; se X não segue uma distribuição normal mas o tamanho n das amostras é suficientemente grande ($n \geq 30$) então a distribuição da média amostral é assimptoticamente normal. Para cada uma das situações tem-se

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{(resp. assimpt.)}$$

ou equivalentemente

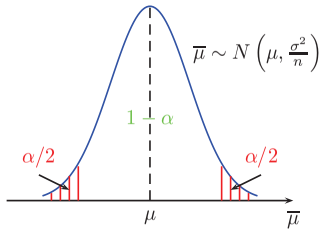
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{(resp. assimpt.)}.$$

Fixemos agora um número α com $0 < \alpha < 1$. O problema que se coloca é o da determinação de um intervalo que contenha a média da população com $(1-\alpha)100\%$ de confiança.

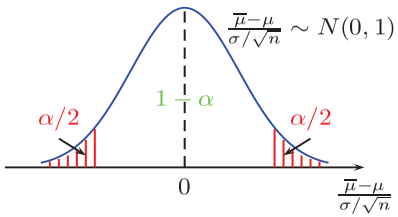
Como a distribuição normal é simétrica, é usual (mas não obrigatório) pensar-se em intervalos centrados em μ. É dessa situação que vamos tratar em seguida.

Começamos por dividir α de igual forma por entre as duas caudas da distribuição de $\overline{X}$, definindo um intervalo $[\mu_L, \mu_U]$ que contém $(1-\alpha)100\%$ de todos os possíveis valores de $\overline{X}$:

$$P(\mu_L < \overline{X} < \mu_U) = 1 - \alpha.$$



Equivalentemente,



Resolvendo a inequação

$$N_{\alpha/2}(0,1) \leq \frac{\overline{X} - \mu}{\sigma\sqrt{n}} \leq N_{1-\alpha/2}(0,1)$$

em ordem a μ, obtem-se

$$\frac{\sigma}{\sqrt{n}} N_{\alpha/2}(0,1) \leq \overline{X} - \mu \leq \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1)$$

donde

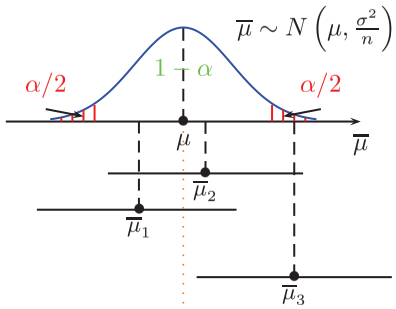
$$\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1) \leq \mu \leq \overline{X} - \frac{\sigma}{\sqrt{n}} N_{\alpha/2}(0,1)$$

ou, dado que $N_{\alpha/2}(0,1) = -N_{1-\alpha/2}(0,1)$,

$$\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}.^{a} \tag{4.6}$$

As desigualdades (4.6) representam um **intervalo aleatório** pois $\overline{X}$ é uma variável aleatória e como tal pode tomar uma infinidade de valores! A questão que se coloca é a de passar do intervalo aleatório para um intervalo concreto.

Suponhamos que temos uma amostra de tamanho n de X. Para essa amostra, calculamos (o escalar) $\overline{X}$ e o intervalo correspondente a (4.6). Em seguida, extraímos outra amostra e voltamos a calcular o intervalo correspondente a (4.6). E fazemos o mesmo noutra amostra. E noutra amostra...



O número de amostras que podemos extrair da população (com reposição) é infinito, podendo portanto encontrar um número infinito de concretizações de intervalos do tipo (4.6), todos com a mesma amplitude.

[a] Para simplificação de notação, representamos $N_q(0,1)$ apenas por N_q, para qualquer $q \in [0,1]$.

Dependendo do valor de $\overline{X}$ na amostra, alguns desses intervalos conterão a média μ da população, outros não.

Dada uma amostra, o intervalo (concreto)

$$IC_{(1-\alpha)100\%} = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1), \overline{X} + \frac{\sigma}{\sqrt{n}} N_{1-\alpha/2}(0,1)\right]$$

diz-se o **intervalo de confiança a $(1-\alpha)100\%$ para a média**, e é representado por $IC_{(1-\alpha)100\%}$.

Tecem-se agora as seguintes considerações:

- a probabilidade do intervalo aleatório (4.6) conter a média da população é igual a $1-\alpha$
- a média da população pode ou não estar contida num intervalo de confiança (concreto)
- não se pode afimar que um intervalo de confiança a $(1-\alpha)100\%$ contém a média com probabilidade $1-\alpha$
- de entre todos os intervalos de confiança possíveis de construir a partir de amostras aleatórias (com reposição) de tamanho n, exactamente $(1-\alpha)100\%$ desses intervalos contêm a média μ da população.
- utilizando um argumento frequencista, de entre N intervalos de confiança a 95% para a média, *aproximadamente* $N \times 0.95$ desses intervalos conterão a média da população

– qualquer intervalo a 99% de confiança merece mais **confiança**, no que diz respeito a cobrir μ, do que qualquer outro intervalo a 95% de confiança

– não se conhece a **probabilidade** com que um determinado intervalo de confiança para a média contém μ (apesar dessa probabilidade só poder ser 0 ou 1).

– em termos práticos, para facilitar a interpretação mas usando linguagem um pouco imprecisa, diz-se que dado um intervalo de confiança para a média a $(1-\alpha)100\%$, temos $(1-\alpha)100\%$ de ***confiança*** que a média da população esteja contida algures nesse intervalo.

- **Intervalo de confiança para a média - σ^2 desconhecida**

No caso em que a variância da população não é conhecida, tem de ser trabalhar com a variância amostral

$$\overline{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Sabe-se já que, se $X \sim N(\mu, \sigma^2)$, então

$$\frac{\overline{X}-\mu}{\overline{\sigma}/\sqrt{n}} \sim t(n-1). \tag{4.7}$$

Repetindo o argumento do caso anterior, resolvemos a inequação

$$t_{\alpha/2}(n-1) \leq \frac{\overline{X}-\mu}{\overline{\sigma}/\sqrt{n}} \leq t_{1-\alpha/2}(n-1)$$

em ordem a μ para encontrar o intervalo aleatório que contém a média μ de X com uma probabilidade de $(1-\alpha)100\%$:

$$\overline{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1).$$

Daqui resulta que o intervalo de confiança para a média na situação em que a variância é desconhecida é, para uma qualquer amostra,

$$\begin{aligned} &IC_{(1-\alpha)100\%} \\ &= \left[\overline{X} + \frac{\overline{\sigma}}{\sqrt{n}} t_{\alpha/2}(n-1), \overline{X} + \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right] \\ &= \left[\overline{X} - \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\alpha/2}(n-1), \overline{X} + \frac{\overline{\sigma}}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right] \end{aligned}$$

A igualdade entre os dois intervalos anteriores resulta da simetria da distribuição t.

Se a distribuição não for normal mas não apresentar assimetrias *exageradas* e n for suficientemente grande ($n \geq 30$), então tem-se uma convergência assimptótica para a distribuição $N(0,1)$. A aproximação assimptótica da distribuição t pela normal reduzida para amostras *grandes* [a], permite substituir o quantil $t_{1-\alpha/2}(n-1)$ na fórmula anterior por $N_{1-\alpha/2}(0,1)$.

[a] Há autores que usam $n > 200$; por ex., Rosner, Fundamentals of Biostatistics. Essencialmente dependerá das características da distribuição empírica.

Mais uma vez, de entre todos os intervalos de confiança possíveis de contruir a partir de amostras da população, com reposição e com o mesmo tamanho, $(1-\alpha)100\%$ deles conterão a média μ da população.

Observação 1: A amplitude do intervalo de confiança para a média reflecte a precisão com que $\overline{X}$ estima (pontualmente) μ e depende dos valores de n, $\overline{\sigma}$ e α. Têm-se as seguintes considerações:

- quanto maior for o tamanho da amostra menor será a amplitude do intervalo (uma vez fixados $\overline{\sigma}$ e α). De facto, quantos mais indivíduos conseguirmos amostrar, maior será a probabilidade de a amostra ser semelhante à população e consequentemente maior será a semelhança entre os parâmetros amostrais e populacionais.
- quanto maior for $\overline{\sigma}$ maior será a amplitude do intervalo de confiança. Isto relaciona-se com o facto de o desvio-padrão (amostral) reflectir a variabilidade da distribuição empírica.
- quanto maior for α (menor será $1-\alpha$), menor será a amplitude do intervalo de confiança.

Observação 2: As fórmulas para os intervalos de confiança unilaterias a $(1-\alpha)100\%$ de confiança são

$$\left(\overline{X} - \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1), +\infty\right)$$
$$\left(-\infty, \overline{X} + \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$$

- **Intervalo de confiança para a proporção**

  Sendo $X_i \sim B(1,p)$ tem-se $E(X_i) = p$ e $\mathrm{Var}(X_i) = p(1-p)$. Considere-se a v.a.

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

Pelo teorema de De Moivre-Laplace,

$$X \sim B(n,p) \overset{a}{\sim} N\left(np, np(1-p)\right)$$

portanto a proporção amostral $\overline{p} = X/n$ para amostras aleatórias de tamanho n suficientemente grande ($n \geq 30$ e $np(1-p) > 5$) satisfaz

$$\frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0,1).$$

Em particular, note-se também que $E(\overline{p}) = p$. Pelo raciocínio seguido nos casos anteriores, o intervalo aleatório centrado no qual, com probabilidade $1-\alpha$, está a verdadeira proporção p da população é conseguido resolvendo em ordem a p a inequação

$$N_{\alpha/2}(0,1) \leq \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq N_{1-\alpha/2}(0,1)$$

e estimando p por $\overline{p}$ em

$$\sqrt{\frac{p(1-p)}{n}}.$$

Para uma amostra concreta, obtem-se

$$
\begin{aligned}
&IC_{(1-\alpha)100\%} \\
&= \left[\overline{p} + \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} N_{\alpha/2}, \overline{p} + \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} N_{1-\alpha/2}\right] \\
&= \left[\overline{p} - \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} N_{1-\alpha/2}, \overline{p} + \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} N_{1-\alpha/2}\right]
\end{aligned}
$$

onde $N_{\alpha/2}$ e $N_{1-\alpha/2}$ representam, respectivamente, os $\alpha/2$ e $1-\alpha/2$ quantis da distribuição normal $N(0,1)$.

De entre todos os intervalos de confiança a $(1-\alpha)100\%$ para a proporção assim construídos, apenas $\alpha 100\%$ desses intervalos não conterão o parâmetro p.

- **Intervalo de confiança para a variância**

Sabemos já que, se $X \sim N(\mu, \sigma^2)$, então

$$n\frac{\widehat{\sigma}}{\sigma^2} = (n-1)\frac{\overline{\sigma}^2}{\sigma^2} \sim \chi^2(n-1).$$

Para qualquer $0 < \alpha < 1$, resulta então que o intervalo aleatório (centrado) que com probabilidade $1-\alpha$, contém a variância σ^2 da população é o conjunto dos σ^2 que satisfazem

$$\chi^2_{\alpha/2}(n-1) \leq (n-1)\frac{\overline{\sigma}^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}(n-1)$$

isto é

$$\frac{1}{(n-1)\overline{\sigma}^2}\chi^2_{\alpha/2} \leq \frac{1}{\sigma^2} \leq \frac{1}{(n-1)\overline{\sigma}^2}\chi^2_{1-\alpha/2}$$

e portanto o $IC_{(1-\alpha)100\%}$ para uma amostra é

$$
\begin{aligned}
&IC_{(1-\alpha)100\%} \\
&= \left[\frac{(n-1)\overline{\sigma}^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)\overline{\sigma}^2}{\chi^2_{\alpha/2}(n-1)}\right].
\end{aligned}
$$

Repare-se que a distribuição χ^2 não é simétrica e portanto não existe nenhuma relação privilegiada entre $\chi^2_{\alpha/2}(n-1)$ e $\chi^2_{1-\alpha/2}(n-1)$.

Para n grande (para distribuições *razoáveis* pode pensar-se em $n-1 > 60$), pode-se usar o facto de uma distribuição $\chi^2(\nu)$ (com média ν e variância 2ν) poder ser aproximada por uma distribuição normal, isto é,

$$\chi^2(n-1) \approx N(n-1, 2(n-1)).$$

4.8. Intervalos de confiança - populações normais

- **Parâmetro**: Média μ (σ^2 conhecida)

  Condições de aplicação:
  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

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

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1), \overline{X} + \frac{\sigma}{\sqrt{n}} N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Média μ (σ^2 desconhecida)

  Condições de aplicação:
  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

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

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{X} - \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right)$$ [a]

[a] Há autores que escrevem $t_{n-1,1-\frac{\alpha}{2}}$ para designar o $(1-\frac{\alpha}{2})$-quantil da distribuição $t(n-1)$, valor acima designado por $t_{1-\frac{\alpha}{2}}(n-1)$.

- **Parâmetro**: Variância σ^2

  Condições de aplicação:
  $X \sim N(\mu, \sigma^2)$

  Variável fulcral:

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

  IC a $(1-\alpha)100\%$:

  $$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)$$

- **Parâmetro**: Diferença de médias $\mu_1 - \mu_2$ (pop. independentes; σ_1^2, σ_2^2 conhecidas)

  Condições de aplicação:
  $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

  $$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

  IC a $(1-\alpha)100\%$:

  $$\left(\overline{X}_1 - \overline{X}_2 - \sigma^* N_{1-\frac{\alpha}{2}}(0,1), \overline{X}_1 - \overline{X}_2 + \sigma^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $\sigma^* = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- **Parâmetro**: Diferença de médias $\mu_1 - \mu_2$ (pop. independentes; σ_1^2, σ_2^2 desconhecidas)

  Condições de aplicação:

  $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{a}{\sim} t(\nu)$$

  onde ν é o maior inteiro que não excede

$$\nu' = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{X}_1 - \overline{X}_2 - S^* t_{1-\frac{\alpha}{2}}(\nu), \overline{X}_1 - \overline{X}_2 + S^* t_{1-\frac{\alpha}{2}}(\nu)\right)$$

  onde $S^* = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$.

- **Parâmetro**: Razão de variâncias $\frac{\sigma_2^2}{\sigma_1^2}$

  Condições de aplicação:

  $X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2)$

  Variável fulcral:

$$\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1 - 1, n_2 - 1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\frac{S_2^2}{S_1^2} F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1), \frac{S_2^2}{S_1^2} F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)\right)$$

- **Parâmetro**: Diferenças de médias $\mu_1 - \mu_2$ (amostras emparelhadas)

  Condições de aplicação:

  $D = X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$

  Variável fulcral:

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{X}_1 - \overline{X}_2 - \frac{S_D}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \overline{X}_1 - \overline{X}_2 + \frac{S_D}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right)$$

4.9. Intervalos de confiança - amostras grandes

- **Parâmetro**: Média μ (σ^2 finita)

  Variável fulcral:

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{X}-\frac{S}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{X}+\frac{S}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Proporção p

  Condições de aplicação:

  $n \geq 30$ e $n\overline{p}(1-\overline{p}) > 5$ [a]

  Variável fulcral:

$$\frac{\overline{p}-p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{p}-\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}N_{1-\frac{\alpha}{2}}, \overline{p}+\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}N_{1-\frac{\alpha}{2}}\right)$$

  onde $N_{1-\frac{\alpha}{2}}$ representa o $\left(1-\frac{\alpha}{2}\right)$-quantil da distribuição $N(0,1)$.

[a] Caso contrário, usa-se a distribuição binomial. Há autores que consideram $n\overline{p}(1-\overline{p}) \geq 10$, ou $n\overline{p} > 5$ e $n\overline{p}(1-\overline{p}) > 5$.

- **Parâmetro**: Diferença de médias $\mu_1 - \mu_2$ (pop. independentes; σ_1^2, σ_2^2 finitas)

  Variável fulcral:

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{X}_1-\overline{X}_2-S^*N_{1-\frac{\alpha}{2}}(0,1), \overline{X}_1-\overline{X}_2+S^*N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $S^* = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$.

- **Parâmetro**: Diferenças de médias $\mu_1 - \mu_2$ (amostras emparelhadas)

  Variável fulcral:

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S_D/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{X}_1-\overline{X}_2-\frac{S_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{X}_1-\overline{X}_2+\frac{S_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Diferença de proporções $p_1 - p_2$

  Condições de aplicação:

  $n_1, n_2 \geq 30$, $n_1\overline{p}_1(1-\overline{p}_1) > 5$ e $n_2\overline{p}_2(1-\overline{p}_2) > 5$ [a]

  Variável fulcral:

$$\frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}} \overset{a}{\sim} N(0,1)^{b}$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{p}_1 - \overline{p}_2 - \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1), \overline{p}_1 - \overline{p}_2 + \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $\overline{p}^* = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$

[a] Caso contrário, usa-se o teste exacto de Fisher que usa a distribuição hipergeométrica.

[b] Com correcção de continuidade, a variável fulcral é

$$\frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2) - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}} \overset{a}{\sim} N(0,1)$$

Para amostras *bastante grandes*, a correcção de continuidade é usualmente ignorada (reparar que esta conduz a uma alteração da estatística quase insignificante).

Observação: Um intervalo de confiança **exacto** a $(1-\alpha)100\%$ para a proporção p é (p_1, p_2) onde p_1 e p_2 são tais que:

$$P(X \geq x | p = p_1) = \frac{\alpha}{2} = \sum_{i=x}^{n} \binom{n}{i} p_1^i (1-p_1)^{n-i}$$

$$P(X \leq x | p = p_2) = \frac{\alpha}{2} = \sum_{i=0}^{x} \binom{n}{i} p_2^i (1-p_2)^{n-i}$$

(aqui $X \sim B(n,p)$ é a v.a. que conta o número de sucessos na amostra de tamanho n.)

4.10. Intervalos de confiança - amostras grandes

- **Parâmetro**: Média μ (σ^2 finita)

  Variável fulcral:

$$\frac{\overline{\mu}-\mu}{\overline{\sigma}/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{\mu}-\frac{\overline{\sigma}}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{\mu}+\frac{\overline{\sigma}}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

- **Parâmetro**: Proporção p

  Condições de aplicação:

  $n \geq 30$ e $n\overline{p}(1-\overline{p}) > 5$

  Variável fulcral:

$$\frac{\overline{p}-p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{p}-\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}N_{1-\frac{\alpha}{2}}, \overline{p}+\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}N_{1-\frac{\alpha}{2}}\right)$$

  onde $N_{1-\frac{\alpha}{2}}$ representa o $\left(1-\frac{\alpha}{2}\right)$-quantil da distribuição $N(0,1)$.

- **Parâmetro**: Diferença de médias $\mu_1-\mu_2$ (pop. independentes; σ_1^2, σ_2^2 finitas)

  Variável fulcral:

$$\frac{(\overline{\mu}_1-\overline{\mu}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\overline{\sigma}_1^2}{n_1}+\frac{\overline{\sigma}_2^2}{n_2}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{\mu}_1-\overline{\mu}_2-\overline{\sigma}^* N_{1-\frac{\alpha}{2}}(0,1), \overline{\mu}_1-\overline{\mu}_2+\overline{\sigma}^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

  onde $\overline{\sigma}^* = \sqrt{\frac{\overline{\sigma}_1^2}{n_1}+\frac{\overline{\sigma}_2^2}{n_2}}$.

- **Parâmetro**: Diferenças de médias $\mu_1-\mu_2$ (amostras emparelhadas)

  Variável fulcral:

$$\frac{(\overline{\mu}_1-\overline{\mu}_2)-(\mu_1-\mu_2)}{\overline{\sigma}_D/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

$$\left(\overline{\mu}_1-\overline{\mu}_2-\frac{\overline{\sigma}_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1), \overline{\mu}_1-\overline{\mu}_2+\frac{\overline{\sigma}_D}{\sqrt{n}}N_{1-\frac{\alpha}{2}}(0,1)\right)$$

Nota: Os resultados apresentados nesta secção - excepto para a proporção e diferença de proporções - são válidos para amostras grandes que não sejam *exageradamente* assimétricas.

- **Parâmetro**: Diferença de proporções $p_1 - p_2$

  Condições de aplicação:

  $n_1, n_2 \geq 30$, $n_1\overline{p}_1(1-\overline{p}_1) > 5$ e $n_2\overline{p}_2(1-\overline{p}_2) > 5$

  Variável fulcral:

$$\frac{(\overline{p}_1 - \overline{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}} \overset{a}{\sim} N(0,1)$$

  IC a $(1-\alpha)100\%$:

  $\left(\overline{p}_1 - \overline{p}_2 - \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1), \overline{p}_1 - \overline{p}_2 + \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1)\right)$

  onde $\overline{p}^* = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$.

4.11. Intervalos de confiança: Exemplos

Exemplo 1: Suponhamos que um investigador, interessado em estimar o nível médio de uma determinada enzima numa população de indivíduos, considera uma realização $x_1, \ldots, x_{10}$ de uma amostra aleatória de tamanho 10, determina o nível da enzima em cada uma dessas pessoas e calcula a média desses valores obtendo $\overline{x} = 22$. Suponhamos ainda que é razoável assumir que a variável "nível da enzima" segue uma distruibuição normal com uma variância igual a 45. Determinar um intervalo de confiança para a média da população:

(a) a 95% de confiança

(b) a 99% de confiança.

Resolução: Temos $\overline{x} = 22$, $\sigma^2 = 45$ e $n = 10$.
O intervalo de confiança a 95% para a média da população é

$$\begin{aligned}&\left(22 - \frac{\sqrt{45}}{\sqrt{10}} N_{0.975}, 22 + \frac{\sqrt{45}}{\sqrt{10}} N_{0.975}\right)\\ &= (22 - 2.12(1.96), 22 + 2.12(1.96))\\ &= (17.85, 26.15).\end{aligned}$$

```
> n <- 10
> mean.enz <- 22
> var.enz <- 45
> IC.low.95 <- mean.enz -
+ sqrt(var.enz)/sqrt(n)*qnorm(0.975)
> IC.high.95 <- mean.enz +
+ sqrt(var.enz)/sqrt(n)*qnorm(0.975)
> cat("[", IC.low.95, ",", IC.high.95, "]", "\n")
[ 17.84229 , 26.15771 ]
```

De forma análoga, o intervalo de confiança a 99% para a média da população é

$$\begin{aligned}&\left(22 - \frac{\sqrt{45}}{\sqrt{10}} N_{0.995}, 22 + \frac{\sqrt{45}}{\sqrt{10}} N_{0.995}\right)\\ &= (22 - 2.12(2.57), 22 + 2.12(2.57))\\ &= (16.54, 27.46).\end{aligned}$$

```
> IC.low.99 <- mean.enz -
+ sqrt(var.enz)/sqrt(n)*qnorm(0.995)
> IC.high.99 <- mean.enz +
+ sqrt(var.enz)/sqrt(n)*qnorm(0.995)
> cat("[", IC.low.99, ",", IC.high.99, "]", "\n")
[ 16.53584 , 27.46416 ]
```

Repare-se que qualquer intervalo de confiança a 99% é maior do que o correspondente intervalo a 95% de confiança. Ganha-se confiança mas perde-se precisão.

Exemplo 2: Um grupo de investigadores recolheu os seguintes dados para o peso da glândula pituitária numa amostra constituída por quatro ratinhos:

média: 9.0 mg, desvio padrão: 0.6.

Pretende-se determinar um intervalo de confiança a 95% para o peso médio da glândula pituitária da população de ratinhos. Sob que hipóteses é que este intervalo é válido?

Resolução: Temos $n = 4$, $\overline{x} = 9.0$, $s^2 = 0.36$, $\alpha = 0.05$ e não se conhece a variância da população. Além disso não se sabe se o peso da glândula pituitária nos ratinhos segue uma distribuição normal e o tamanho da amostra não é grande. Nestas condições não poderíamos deduzir um intervalo de confiança para a média da população.
Na hipótese de ser razoável assumir que a população segue uma distribuição normal, o intervalo de confiança a 95% será

$$\left(9.0 - \frac{0.6}{2} t_{0.975}(3), 9.0 + \frac{0.6}{2} t_{0.975}(3)\right)$$
$$= (9.0 - 0.3(3.182), 9.0 + 0.3(3.182))$$
$$= (8.05, 9.95)$$

```
> alpha <- 0.05
> n <- 4
> mean.peso <- 9
> sd.peso <- 0.6
> IC.low.alpha <- mean.peso -
+ sd.peso/sqrt(n)*qt(1-alpha/2, df=n-1)
> IC.high.alpha <- mean.peso +
+ sd.peso/sqrt(n)*qt(1-alpha/2, df=n-1)
> cat("[", IC.low.alpha, ",", IC.high.alpha, "]", "\n")
[ 8.045266 , 9.954734 ]
```

Exemplo 3: Numa amostra aleatória constituída por 591 pacientes de um hospital psiquiátrico verificou-se que 204 desses pacientes admitiram ter consumido cannabis pelo menos uma vez na vida. Determinar o intervalo de confiança a 90% para a proporção de indivíduos que alguma vez consumiu cannabis considerando como população todos os indivíduos inscritos nesse hospital psiquiátrico.

Resolução: Temos $n = 591$, $\overline{p} = 204/591 \approx 0.34$ e $\alpha = 0.1$. Sendo a amostra grande e p não muito próximo de 0 nem 1, o intervalo de confiança a 90% para a proporção da população é

$$\left(0.34 - \sqrt{\frac{0.34(0.66)}{591}} N_{0.95}, 0.34 + \sqrt{\frac{0.34(0.66)}{591}} N_{0.95}\right)$$
$$= (0.34 - 0.02(1.645), 0.34 + 0.02(1.645)).$$
$$= (0.31, 0.38).$$

```
> alpha <- 0.1
> n <- 591
> p <- 204/591
> IC.low.alpha <- p - sqrt(p*(1-p)/n)*qnorm(1-alpha/2)
> IC.high.alpha <- p + sqrt(p*(1-p)/n)*qnorm(1-alpha/2)
> cat("[", IC.low.alpha, ",", IC.high.alpha, "]", "\n")
[ 0.3130102 , 0.3773451 ]
```

Exemplo 4: Uma equipa de investigação está interessada na diferença entre os níveis de ácido úrico em pacientes com e sem sindrome de Down. Uma amostra de 12 indivíduos com a doença foi seleccionada de um hospital psiquiátrico fornecendo uma média de 4.5 mg/100 ml e um desvio padrão de 1 mg/100 ml. Num outro hospital foi recolhida uma amostra de 15 indivíduos sem a doença, com as mesmas idades e o mesmo sexo, e a média e o devio padrão obtidos foram de 3.4 mg/100 ml e 1.5 mg/100 ml, respectivamente. Determine um intervalo de confiança a 99% para diferença entre os valores médios de ácido úrico entre doentes e não doentes.

Resolução: As amostras não são grandes portanto um intervalo de confiança para a diferença de médias só será conseguido se se assumir que as duas populações em causa (independentes) seguem distribuições normais. Dos dados, resulta que

$$\begin{array}{lll} n_1 = 12 & \overline{x}_1 = 4.5 & s_1 = 1 \\ n_2 = 15 & \overline{x}_2 = 3.4 & s_2 = 1.5. \end{array}$$

O intervalo de confiança a 99% para a diferença de médias é

$$\left(\overline{x}_1 - \overline{x}_2 - s^* t_{1-\frac{\alpha}{2}}(\nu), \overline{x}_1 - \overline{x}_2 + s^* t_{1-\frac{\alpha}{2}}(\nu)\right)$$

onde ν é o maior inteiro que não excede

$$\nu' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

Ora

$$\nu' = \frac{\left(\frac{1^2}{12} + \frac{(1.5)^2}{15}\right)^2}{\frac{1}{11}\left(\frac{1^2}{12}\right)^2 + \frac{1}{14}\left(\frac{(1.5)^2}{15}\right)^2} \approx 24.322$$

portanto $\nu = 24$, e

$$S^* = \sqrt{\frac{1^2}{12} + \frac{(1.5)^2}{15}} \approx 0.483.$$

Vem então que o intervalo pedido é

$$\begin{aligned} &(4.5 - 3.4 - 0.483\, t_{0.995}(24), 4.5 - 3.4 + 0.483\, t_{0.995}(24)) \\ &= (-0.28, 2.48). \end{aligned}$$

```
> IC.L <- 4.5 -3.4-0.493*qt(0.995, df=24)
> IC.L
[1] -0.2788912

> IC.U <- 4.5 -3.4+0.493*qt(0.995, df=24)
> IC.U
[1] 2.478891
```

Podemos dizer, em termos práticos, que temos 99% de *confiança* que o intervalo $(-0.28, 2.48)$ contenha a diferença de médias entre as duas populações. Ora isto implica que, para esse nível de confiança, essa diferença pode ser 0! Resulta então que, neste caso, não se consegue afirmar que existem diferenças entre os valores médios de ácido úrico nos doentes e nos não doentes.

Exemplo 5: Connor *et al* (2003) estudaram as diferenças de género entre agressões proactivas e reactivas numa amostra de 323 crianças e adolescentes (68 raparigas e 255 rapazes). Dos relatórios sobre os indivíduos amostrados, constatou-se que 31 das raparigas e 53 dos rapazes tinham sofrido abusos sexuais.
Construa um intervalo de confiança a 95% para a diferença entre as proporções de abuso sexuais entre as duas populações consideradas. Interprete o resultado.

Resolução: Tem-se

$$\begin{aligned} n_1 &= 68, \quad \overline{p}_1 = 31/68 \approx 0.456 \\ n_2 &= 255, \quad \overline{p}_2 = 53/255 \approx 0.208. \end{aligned}$$

Como as amostras são *grandes* e as proporções amostrais não são muito próximas dos extremos (0 ou 1), o intervalo de confiança a 95% para $p_1 - p_2$ é

$$\left(\overline{p}_1 - \overline{p}_2 - \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1), \overline{p}_1 - \overline{p}_2 + \overline{p}^* N_{1-\frac{\alpha}{2}}(0,1)\right)$$

onde $\overline{p}^* = \sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$. No caso em questão, vem então

$$\overline{p}^* = \sqrt{\frac{\frac{31}{68}\left(1-\frac{31}{68}\right)}{68} + \frac{\frac{53}{255}\left(1-\frac{53}{255}\right)}{255}} \approx 0.065$$

e portanto o intervalo de confiança a 95% é

$$\begin{aligned} &\left(\frac{31}{68} - \frac{53}{255} - 0.065\, N_{0.975}, \frac{31}{68} - \frac{53}{255} + 0.065\, N_{0.975}\right) \\ &= \left(\frac{31}{68} - \frac{53}{255} - 0.065 \times 1.96, \frac{31}{68} - \frac{53}{255} + 0.065 \times 1.96\right) \\ &= (0.0791, 0.4171). \end{aligned}$$

Conclui-se que, com 95% de *confiança*, a diferença de proporções populacionais $p_1 - p_2$ situa-se alguns entre 0.0791 e 0.4171. Dado que este intervalo não contém o 0, conseguimos afirmar, com esse nível de confiança, que a proporção de raparigas que sofrem abusos sexuais é diferente da proporção de rapazes que sofrem abusos sexuais.

4.12. Testes de Hipóteses

Suponhamos que X é uma variável aleatória com densidade de probabilidade $f(X, \theta)$ e que Θ é o espaço dos possíveis valores para θ. Suponhamos ainda que o espaço Θ está naturalmente (pelo problema em estudo) dividido em dois subespaços disjuntos Θ_0 e Θ_1,

$$\Theta = \Theta_0 \cup \Theta_1, \quad \Theta_0 \cap \Theta_1 = \emptyset.$$

Problema: Com base numa amostra aleatória de tamanho n da variável aleatória X,

$$x_1, \ldots, x_n,$$

verificar se existe evidência estatística para rejeitar a chamada **hipótese nula** [a]

$$H_0 : \theta \in \Theta_0$$

e aceitar a chamada **hipótese alternativa**

$$H_1 : \theta \in \Theta_1$$

ou se, pelo contrário, não podemos rejeitar H_0.
Pela forma como o problema está colocado, a resposta só pode ser dada de duas formas:

- **rejeitar H_0 e aceitar H_1**

ou

- **não rejeitar H_0 (e nada afirmar sobre H_1).**

[a] Para simplificação da apresentação da teoria, consideramos para já testes paramétricos

Para resolver o problema vamos assumir H_0 e construir a partir daí uma estatística (que é uma variável aleatória), designada por **estatística de teste**,

$$T(x_1, \ldots, x_n)^{a}$$

que vai servir para diagnosticar a situação.
A ideia básica consiste em construir uma subregião W no espaço dos possíveis valores que T pode tomar quando aplicado a uma amostra $x_1, \ldots, x_n$ de tamanho n e quando o parâmetro θ está em Θ, e estabelecer uma regra de decisão do seguinte tipo:

- Se $T(x_1, \ldots, x_n) \in W$ rejeita-se H_0 e aceita-se H_1, *i.e.*, aceita-se que $\theta \in \Theta_1$
- Se $T(x_1, \ldots, x_n) \notin W$ não se pode rejeitar H_0.

Por esta razão se diz que W é a **região de rejeição** de H_0.

	$\theta \in \Theta_0$ H_0 é verdadeira	$\theta \in \Theta_1$ H_1 é verdadeira
Rejeitar H_0 $(T \in W)$	Erro de tipo I	Decisão correcta
Não rejeitar H_0 $(T \in W)$	Decisão correcta	Erro de tipo II

[a] Uma grande parte das estatísticas de teste consiste do quociente

$$T = \frac{\text{variância explicada pelo modelo}}{\text{variância não explicada pelo modelo}}.$$

Seja

$$\begin{aligned}\alpha &= P\,(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira})\\ &= P(T(x) \in W \mid \theta \in \Theta_0)\end{aligned}$$

a probabilidade de se cometer um erro de tipo I, *i.e.*, a probabilidade de se cometer um erro ao rejeitar a hipótese nula. O valor α diz-se o **nível de significância** ou a **dimensão do teste**.[a] Seja

$$\begin{aligned}\beta &= P\,(\text{não rejeitar } H_0 \mid \ H_1 \text{ é verdadeira})\\ &= P(T(x) \notin W \mid \theta \in \Theta_1)\end{aligned}$$

a probabilidade de se cometer um erro de tipo II, *i.e.*, a probabilidade de cometer um erro ao não rejeitar a hipótese nula. À probabilidade

$$\begin{aligned}1-\beta &= P\,(\text{rejeitar } H_0 \mid H_0 \text{ é falsa})\\ &= P(T(x) \in W \mid \theta \in \Theta_1),\end{aligned}$$

que representa a probabilidade de acertar ao rejeitar a hipótese nula, chama-se a **potência do teste**.

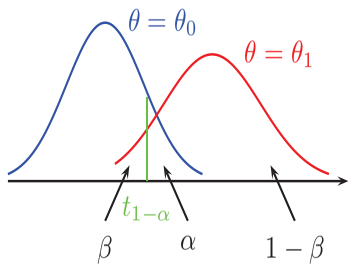
	$\theta \in \Theta_0$ H_0 é verdadeira	$\theta \in \Theta_1$ H_1 é verdadeira
Rejeitar H_0 $(T \in W)$	Erro de tipo I Prob $= \alpha$	Decisão correcta Prob $= 1-\beta$
Não rejeitar H_0 $(T \in W)$	Decisão correcta Prob $= 1-\alpha$	Erro de tipo II Prob $= \beta$

[a] ou o **tamanho de um erro de tipo I**.

Podem agora levantar-se vários problemas:

- como construir a região de rejeição W de modo a que a probabilidade de se cometer um erro de tipo I seja inferior a um certo valor α?
- como construir a região de rejeição W de modo a que a potência do teste seja superior a um certo valor β?
- se para um nível de significância α dado, um teste é mais potente quando aumenta o tamanho da amostra, qual deverá ser o tamanho da amostra para que o teste tenha potência β especificada?

Na maior parte dos casos existe um procedimento que permite construir um teste (*i.e.*, uma região de rejeição R_α) para um qualquer nível de significância $\alpha \in [0,1]$ especificado. Tipicamente, essa região vai-se tornando mais pequena à medida que α descresce. Por outro lado, se reduzirmos α fazendo com que a região de rejeição seja menor, em geral aumentamos β. O único meio de reduzir ambos os erros é aumentar o tamanho da amostra, o que não é sempre possível.



Na prática, costuma adoptar-se um nível de significância de 0.05 ou 0.01, embora quaisquer outros valores entre 0 e 1 possam ser possíveis. Se ao delinearmos um teste escolhermos, por exemplo, um nível de significância de 0.05, ou 5%, isso significa que em cerca de 5 vezes em 100 rejeitaríamos a hipótese nula quando ela deveria ter sido ser aceite. Equivalentemente, poderíamos ter 95% de confiança em termos tomado a decisão correcta. Na situação anterior, dizemos que H_0 foi rejeitada com um nível de significância de 5%, o que significa que podemos ter errado com uma probabilidade de 5%.

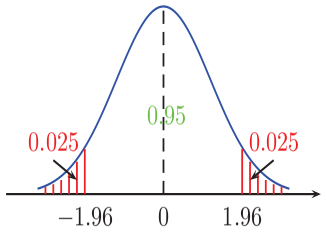
Para um dado teste de hipóteses, o seu **valor-p** (ou **valor de prova**) é o menor α para o qual $X \in R_\alpha$, *i.e.*, é o menor nível de significância com que H_0 é rejeitada [a] , dado X. Conhecido o valor-p de um teste para uma variável aletória X procedemos da seguinte forma:

- se $\mathbf{p} \leq \alpha$ então **rejeitamos H_0 com um nível de significância α**
- se $\mathbf{p} > \alpha$ então **não rejeitamos H_0 com um nível de significância α**.

Note-se que o valor-p é uma estatística (o seu valor varia com a amostra considerada).

Exemplo: Suponhamos que, assumindo H_0 a estatística de teste Z segue a distribuição normal standard $N(0,1)$.

[a] na verdade, corresponde ao nível de significância para o qual seria indiferente rejeitar ou não H_0, dado que a estatística de teste está na fronteira da região de rejeição.



A figura considera um nível de significância de 0.05. Dada uma amostra aleatória, podemos ter 95% de confiança em que o valor de z esteja entre -1.96 e 1.96 (a região de não rejeição). Se na escolha de uma amostra aleatória verificarmos que o valor de z está na região de rejeição então, supondo que H_0 é verdadeira, esse valor amostral ocorreria apenas em 5% de todos os casos possíveis. Estando dispostos a aceitar um erro de 0.05, rejeitaríamos H_0.

Os testes de hipóteses podem ser:

- **bilaterais:** H_1 é uma reunião de dois conjuntos disjuntos

  Podemos ter uma região de rejeição (caso do teste de ajustamento do χ^2) ou duas regiões de rejeição (caso do teste t para a média de uma distribuição normal), dependendo da estatística de teste que está a ser usada.

- **unilaterais:** H_1 corresponde a um único intervalo real.

  Temos apenas uma região de rejeição, situada sobre a cauda da distribuição da estatística do teste que sob o que se conhece a priori tem menor probabilidade de ocorrer.

A decisão de efectuar um teste bilateral ou unilateral deverá ser tomada antes dos dados serem recolhidos, face à investigação que está a ser conduzida. Essa decisão nunca deverá depender dos dados obtidos após a experiência e muito menos ser influenciada pelo valor observado da estatística de teste proveniente da amostra. Por exemplo, prescindir de um teste bilateral em favor de um teste unilateral apenas porque o segundo fornece um menor valor-p e permite rejeitar a hipótese nula, é uma prática a evitar! (facilmente será interpretada como desonesta...). Na situação de uma amostra sugerir novas pistas de investigação após a realização de um dado teste de hipóteses, essas devem ser testadas recolhendo para o efeito outra amostra, com o planeamento experimental adequado.

Os testes unilaterais devem ser usados com muito cuidado e idealmente só nas situações em que existe uma evidência clara para condicionar os parâmetros que se querem testar numa determinada direcção. Por exemplo, um psiquiatra quer testar se o tempo médio de efeito de dois ansiolíticos é o mesmo. Porém, ele desconfia (baseado nalguns argumentos) que o ansiolítico A produz efeito durante menos tempo, em média, do que o ansiolítico B. Ao fazer o planeamente experimental, ele pode escolher a realização de um teste unilateral para comparação de médias.

Há testes de hipóteses que só são válidos sob determinadas condições, aquilo a que chamamos as suas **condições de aplicação**. Se estas condições não forem satisfeitas e ainda assim aplicarmos o teste, corremos o risco de tirar uma conclusão completamente errada para o problema em causa.

Um teste de hipóteses diz-se:

- **robusto** se estimar correctamente o seu valor p de uma forma relativamente independente do facto de as suas condições teóricas de aplicação se verificarem ou não.

- **conservativo** se sobreestimar o seu valor p na situação em que algumas das condições de aplicação não forem satisfeitas. Isto significa que o nível de significância do teste (realizado sob condições diferentes das condições teóricas de aplicabilidade) é, quando muito, igual ao nível α fixado para as condições teóricas ideais. Por outras palavras, a medida da região crítica, sob H_0, é quando muito igual a α.

  Um teste conservativo pode incorrectamente falhar na rejeição da hipótese nula e portanto ser menos potente do que o esperado (recorde-se que a potência de um teste é a probabilidade de, correctamente, rejeitar a hipótese nula quando ela é de facto falsa).

- **liberal** ou **não conservativo** se subestimar o seu valor p se algumas das hipóteses de aplicação não forem satisfeitas.

O ideal seria, obviamente, que todos os testes fossem robustos.

Os testes de hipóteses são muitas vezes classificados numa das seguintes categorias:

- **paramétricos:** assume-se que os dados provêm de uma determinada distribuição e a hipótese nula incide sobre um parâmetro dessa distribuição.
- **não paramétricos:** não se assumem hipóteses (ou assumem-se hipóteses muito ligeiras) sobre a distribuição de onde provêm os dados.

A maior parte dos testes não paramétricos usa estatísticas de teste que resultam de uma ordenação das observações por ordem crescente (*rank statistics*). Outros usam frequências ou percentis. Este tipo de testes são usados, em geral, quando o tamanho da amostra é pequeno, quando não pode ser assumido que os dados provêm da distribuição exigida (normalmente a normal) pelos testes paramétricos *correspondentes*, ou quando os dados recolhidos são ordinais ou nominais.

Não havendo (quase) condições de aplicação para os teste não paramétricos, pode-se pensar numa aplicação sistemática destes testes em detrimento das versões *análogas* de testes paramétricos. Este não é um procedimento correcto; pode-se mostrar, por exemplo, que a aplicação de um teste não paramétrico para a mediana a dados que provêm de uma população que segue uma distribuição normal fornece um valor-p que é superior ao valor-p do teste paramétrico *correspondente* (o teste-t). Em particular, poder-se-ía dar o caso de rejeitar H_0 no caso do teste t e não poder fazer o mesmo no caso do teste não paramétrico. Em regra, os métodos não-paramétricos têm uma potência inferior (i.e., menor probabilidade de detectarem um efeito verdadeiro sempre que este existe) aos seus equivalentes métodos paramétricos.

Em conclusão, sempre que as hipóteses de aplicação de um teste paramétrico se verificarem, esse deve ser o teste a usar.

Uma inspecção das tabelas das estatísticas de teste de procedimentos não paramétricos indica que muito dificilmente se conseguem obter testes significativos, para $\alpha = 0.05$, em amostras demasiado pequenas. [a] A ser necessário efectuar algum tipo de análise estatística com tamanhos amostrais pequenos (digamos inferiores a 8), terá de se assumir normalidade ou obtê-la de estudos anteriores. Caso isto não se consiga, o melhor é optar por análises meramente descritivas.

O resultado de um teste de hipóteses diz-se **estatisticamente significativo** quando se pode rejeitar a hipótese nula para o nível de significância especificado no início. Quando tal não acontece, o que se tem é que as diferenças entre a amostra e aquilo que se quer testar em H_0 não fornecem evidência estatística *suficiente* para H_0 poder ser rejeitada (para o α especificado). Diz-se então que essas diferenças se devem ao acaso, relacionado com a selecção aleatória da(s) amostra(s).

O facto do resultado de um teste de hipóteses ser estatisticamente significativo não implica necessariamente que seja **clinicamente significativo**. De facto, ser estatisticamente significativo apenas garante a existência de diferenças mas não indica a magnitude nem a precisão dessas diferenças. Por exemplo, se um teste de hipóteses mostra que um novo método clínico permite baixar o valor da pressão arterial, a aplicação clínica desse novo método dependerá do valor concreto da redução que produz.

[a] Bland, Altman, *An Introduction to Medical Statistics.*

Para poder ter uma ideia do significado clínico do resultado do teste, tem sido ultimamente indicada a divulgação, juntamente com o resultado do teste de hipóteses, do intervalo de confiança correspondente. Na verdade, apesar do valor-p funcionar como uma medida da força de uma associação ou efeito, o intervalo de confiança permite perceber entre que valores é que varia esse efeito.
A este propósito, notar que um nível de significância de α para um teste de hipóteses retribui um intervalo de confiança correspondente a $(1-\alpha)\%$ de confiança. O resultado do teste de hipóteses é significativo sempre que o intervalo de confiança não contém o valor especificado na hipótese nula (muitas vezes o zero, mas não sempre!). (A relação entre testes de hipóteses e intervalos de confiança vai ser mostrada com detalhe no caso do teste de hipóteses para a média).
Há contudo situações em que não é possível determinar os desvios padrão das estatísticas de teste e os intervalos de confiança são contruídos a partir das variâncias assimptóticas. Nesta situação, as propriedades estatísticas dos intervalos de confiança assim contruídos são menos interessantes do que o resultado ("isolado") do teste de hipóteses.
Ainda outra questão relacionada com os intervalos de confiança tem a ver com a sua simetria em torno da estatística amostral...
A propósito de testes de hipóteses e intervalos de confiança, podemos sempre seguir a recomendação de Knapp (1999):

use a hypothesis test if you have a hypothesis to test; otherwise, use interval estimation.

Finalmente, refira-se que, nas conclusões dos estudos, o autor deve referir explicitamente os valores de prova dos testes realizados, de forma a que o leitor possa aferir, em pormenor, a significância estatística dos eventuais efeitos encontrados.

Para a resolução de exercícios que envolvam a consideração de um teste de hipóteses, os alunos devem seguir o esquema seguinte:

1. escrever o nome do teste
2. identificar as v.a. envolvidas no problema
3. enunciar as condições de aplicação do teste (caso existam) e verificar que todas são satisfeitas
4. formular H_0 e H_1
5. determinar à mão ou identificar nos resultados do software (dependendo da situação) o valor da estatística de teste e indicar a sua distribuição
6. desenhar uma figura que ilustre a distribuição da estatística e a zona de rejeição

   (este item é facultativo mas é muito útil para ajudar a obter a conclusão)
7. escrever a conclusão do teste de hipóteses e referir também (se possível) o intervalo de confiança correspondente
8. se pedido no enunciado, determinar (à mão) o valor-p do teste (pelo menos nos casos mais simples; por exemplo, nos testes sobre a(s) média(s)).

Observações finais sobre o valor-p:

- é uma medida da força de uma associação ou efeito; não é uma medida do tamanho desse efeito.
- não deve ser usado isoladamente de uma interpretação dos resultados experimentais (significância estatística versus significância clínica...)
- o valor-p de um teste é afectado pela magnitude do efeito e pelo tamanho amostral do estudo a que diz respeito, pelo que deve ser interpretado com cuidado.

Métodos Estatísticos em Saúde - FMUP

Folha de exercícios no. 4:

1. De entre os 1546 estudantes do sexo feminino de uma determinada faculdade escolheu-se uma amostra aleatória constituída por 100 estudantes. A altura média e o desvio padrão obtidos para os indivíduos da amostra foram, respectivamente, 1.63 cm e 8 cm. Determine dois intervalos de confiança, um a 95% e outro a 99%, para a altura média das estudantes dessa faculdade, especificando as hipóteses de aplicação desses intervalos. Interprete ainda os resultados obtidos.

2. Os valores dos diâmetros de uma amostra aleatória de 200 comprimidos esféricos produzidos por um determinado laboratório durante uma semana acusa média de 2.8 mm e desvio padrão 0.3 mm.

   (a) Determine um intervalo de confiança a 99% para o diâmetro médio de todos os comprimidos produzidos pelo laboratório nessa semana.

   (b) Suponha que pretende obter um intervalo de confiança com uma amplitude ainda mais pequena do que a do anterior. Como pode proceder?

3. Ao medir um tempo de reacção, um psicólogo estima o desvio padrão em 0.06 segundos. Qual o tamanho de uma amostra a considerar a fim de que possa ter 95% de confiança em que o *erro* da sua estimativa para o tempo médio de reacção não seja superior a 0.01 segundos?

4. De entre um total de 200 classificações numa determinada disciplina, uma amostra aleatória de 50 alunos acusa média de 17 valores e desvio padrão de 1 valor.

   (a) Quais os valores limite que constituem um intervalo confiança de 95% para a média das classificações obtidas dos 200 alunos?

   (b) Com que grau de confiança podemos afirmar que a média das 200 notas está contida algures no intervalo 17± 0.3?

5. Uma amostra de 10 comprimidos esféricos de entre todos os produzidos por um laboratório num dia apresenta diâmetros com média 3.38 mm e desvio padrão 0.06 mm. Determine um intervalo de confiança a 99% para o valor médio dos diâmetros dos comprimidos produzidos pelo laboratório nesse dia. Sob que condições é que esses intervalos são válidos? (compare a estrutura do problema com a do item 2.)

6. Uma amostra aleatória de 100 indivíduos com uma determinada patologia experimentou um novo medicamento e em 55% dos casos observaram-se melhorias significativas. Determine limites de confiança, a 95% e a 99%, para a percentagem de indivíduos da população com a patologia que vão acusar melhoras ao tomarem o medicamento.

7. O desvio padrão da validade de uma amostra de 200 comprimidos de um determinado medicamento, analisados em laboratório, é de 0.08/ano. Determine um intervalo de confiança, a 95% para o desvio padrão da validade desse medicamento.

8. Em 40 lançamentos de uma moeda apareceram 24 caras. Determine intervalos de confiança a 90% e para a proporção de caras numa sequência ilimitada de lançamentos da moeda.

9. Um investigador, interessado em estimar o nível médio de uma determinada enzima numa população de indivíduos, considera uma amostra constituída por 10 pessoas, determina o nível da enzima em cada uma dessas pessoas e obtem o valor 22 como média do nível da enzima. Supondo que a variável "nível da enzima" é normalmente distribuída com uma variância igual a 45, determine um intervalo que, com 95% de confiança, contenha a média do nível de enzima da população.

10. Um grupo de investigadores recolheu os seguintes dados para o peso da glândula pituitária numa amostra constituída por quatro ratinhos:

    média: 9.0 mg, erro padrão da média: 0.3.

    Pretende-se determinar:

    (a) o desvio padrão amostral.

    (b) um intervalo a 95% de confiança para a média dos pesos da glândula pituitária da população de ratinhos. Sob que hipóteses é que este intervalo é válido? Comente.

11. Numa amostra aleatória constituída por 591 pacientes de um hospital psiquiátrico verificou-se que 204 desses pacientes admitiram ter consumido cannabis pelo menos uma vez na vida. Determine um intervalo a 95% de confiança para a proporção de indivíduos desse hospital que alguma vez consumiu a droga.

12. Foi realizado um estudo sobre tratamento em indivíduos com doença coronária. Depois de (aleatoriamente) escolhida uma amostra, os indivíduos foram distribuídos por 2 grupos: um grupo de 49 indíviduos seguiu um tratamento convencional; um outro grupo de 51 indivíduos mudou completamente de medicação e foi-lhes iniciado um novo tratamento com Diltiazem a 180 mg/dia. Uma das característica de interesse neste estudo foi a diferença de percentagens de pacientes com história de insuficiência cardíaca congestiva (ICC) entre os dois grupos, em *baseline* (no início do estudo). No primeiro grupo, 16 indivíduos apresentaram história de ICC enquanto que no grupo que iría ser exposto ao Diltiazem esse número foi de 12.

    Construa um intervalo de confiança a 95% para a diferença entre as proporções de indivíduos com história de ICC em cada uma das populações consideradas. Indique sob que condições é que esse intervalo de confiança é válido e interprete o resultado obtido.

13. Um nutricionista pretende conduzir um estudo numa população de adolescentes para determinar o consumo médio diário de proteínas (medido em gramas). Qual o tamanho da amostra que o nutricionista deve considerar se:

    - pretende um intervalo de confiança para o consumo médio com uma amplitude inferior ou igual a 10 g
    - o coeficiente de confiança é de 95%
    - de experiências passadas, se pode considerar que o desvio padrão populacional para o consumo de proteínas é de 20 g.

14. Está a ser planeado um inquérito para averiguar a proporção de famílias de uma determinada área que tem necessidades médicas urgentes. Acredita-se que essa proporção não deve ser superior a 0.35. Qual deve ser o tamanho da amostra a considerar caso se pretenda obter um intervalo de confiança a 95% com uma amplitude menor ou igual a 0.1?

15. Foi recolhida uma amostra de 50 indivíduos com problemas mentais severos e submeteu-se (depois de aprovação) essa amostra a um novo programa de tratamento. Uma das variáveis estudadas foi o número de dias de internamento em hospital observados no ano seguinte à conclusão do novo tratamento. De entre 18 indivíduos com esquizofrenia, o número médio de dias de internamento foi 4.7, com um desvio padrão de 9.3; de entre 10 indivíduos com doença bipolar, o correspondente número médio de dias foi 8.8 e o desvio padrão 11.5. Construa um intervalo de confiança a 95% para diferença entre o número médio de dias de internamente nas populações correspondentes às consideradas na amostra. Enuncie as hipóteses de aplicação desse intervalo de confiança e interprete o resultado obtido.

16. Foram realizados testes de acuidade visual em dois grupos de crianças. O primeiro grupo consistiu de 60 crianças que recorriam, essencialmente, a serviços privados de saúde; o valor médio do teste visual para este grupo foi de 26, com um desvio padrão de 5.

    O segundo grupo consistiu de 110 crianças que recorriam, essencialmente, ao serviço nacional de saúde; o valor médio do teste visual para este grupo foi de 21, com um desvio padrão de 7.

    Construa um intervalo de confiança a 99% para a diferença de acuidades visuais médias entre as crianças que recorrem a serviços privados de saúde e as que recorrem ao serviço nacional de saúde. Comente.

17. Num estudo sobre os factores que se pensam estar relacionados com os efeitos adversos do acto de fumar na reprodução humana, foram determinados os níveis de cádmio (nanogramas por grama) no tecido da placenta de uma amostra de 14 mães fumadoras e de uma amostra independente de 18 mães não fumadores. Os resultados foram os seguintes:

Não fumadoras	10.0	8.4	12.8	25.0	11.8
	12.5	15.4	23.5	9.4	25.1
	25.5	9.8	7.5	11.8	12.2
	9.8	19.5	15.0		
Fumadoras	30.0	30.1	15.0	24.1	30.5
	16.8	14.8	13.4	28.5	17.5
	12.5	20.4	17.8	14.4	

Apenas a partir do que sabe sobre intervalos de confiança, parece-lhe razoável concluir que o nível médio de cádmio é maior em 95% das vezes nas mães fumadoras do que nas mães não fumadoras? Justifique convenientemente.

18. Foi realizado um estudo em que se pretendeu avaliar a diferença em relação à adesão ao tratamento em indivíduos com depressão recebendo tratamento convencional e em indivíduos recebendo tratamento com acompanhamento constante de um médico. De entre os 50 indivíduos considerados sob o regime geral, 24 aderiram ao tratamento prescrito; de entre os 75 indivíduos considerados sob o tratamento acompanhado, 50 aderiram ao tratamento prescrito.

    Construa um intervalo de confiança a 90% para a diferença entre as proporções de adesão ao tratamento para as duas populações em causa, e interprete os valores obtidos.

19. Foi realizado um inquérito a 500 adultos em que se perguntava o motivo da sua última visita ao dentista. De entre os 220 indivíduos que não tinham formação superior, 44 mencionaram razões de prevenção.

    De entre os restantes 280 indivíduos, com formação superior, 150 afirmaram ser por razões de prevenção.

    Construa um intervalo de confiança a 90% para a diferença entre as proporções de indivíduos que recorrem ao dentista por motivos de prevenção, entre as populações com e sem formação superior. Interprete o resultado obtido.