CE-003: Estatística II - Turma O2 - Avaliações Semanais (1^o semestre 2013)

1. Considere que será feita uma pesquisa aplicando-se um questionário aos alunos do BCC/UFPR sobre o curso e as características e opiniões dos alunos.

   (a) Liste possíveis questões deste questionário certificando-se que sejam incluídas ao menos duas de cada tipo de variáveis conforme discutido em aula (qualitativas nominal/ordinal e quantitativas discreta/contínua).

   (b) Imagine agora que o questionário foi aplicado e as respostas tabuladas para análises. Indique/esboce como seria analisada (separadamente) cada uma das variáveis do questionário.

   (c) Indique ao menos três questões de interesse envolvendo duas ou mais variáveis a serem investigadas no questionário e qual análise dos dados permitiria investigar estas questões.

---

2. Foram coletados dados[1] sobre indicadores sociais em 97 países. Os atributos[2] são: *Nat: taxa de natalidade (1.000 hab.)*, *Mort: taxa de mortalidade (1.000 hab.)*, *MI: mortalidade infantil (1.000 hab)*, *ExpM: expectativa de vida para homens*, *ExpF: expectativa de vida para mulheres*, *Renda: renda per capta anula* e *Região: região geográfica* sendo consideradas: "EUOr"(Europa Oriental),"SA"(América Latina e México),"PM"("Primeiro Mundo"),"OrMd"(Oriente Médio), "Asia"e "Africa". A renda *per capta* foi também dividida em classes: $[0, 500)$, $[500, 2.000)$, $[2.000, 10.000)$ e $[10.000, 35.000)$. Um cabeçalho do arquivo de dados e um resumo das variáveis são mostrados a seguir.

```
                   Nat Mort   MI ExpM ExpF Renda Regiao    GrupoRenda
Albania           24.7  5.7 30.8 69.6 75.5   600   EUOr   (500,2e+03]
Bulgaria          12.5 11.9 14.4 68.3 74.7  2250   EUOr (2e+03,1e+04]
Czechoslovakia    13.4 11.7 11.3 71.8 77.7  2980   EUOr (2e+03,1e+04]
Former_E._Germany 12.0 12.4  7.6 69.8 75.9    NA   EUOr          <NA>
Hungary           11.6 13.4 14.8 65.4 73.8  2780   EUOr (2e+03,1e+04]
Poland            14.3 10.2 16.0 67.2 75.7  1690   EUOr   (500,2e+03]
```

```
      Nat            Mort            MI             ExpM          ExpF
 Min.   : 9.7   Min.   : 2.2   Min.   :  4.5   Min.   :38.1   Min.   :41.2
 1st Qu.:14.5   1st Qu.: 7.8   1st Qu.: 13.1   1st Qu.:55.8   1st Qu.:57.5
 Median :29.0   Median : 9.5   Median : 43.0   Median :63.7   Median :67.8
 Mean   :29.2   Mean   :10.8   Mean   : 54.9   Mean   :61.5   Mean   :66.2
 3rd Qu.:42.2   3rd Qu.:12.5   3rd Qu.: 83.0   3rd Qu.:68.6   3rd Qu.:75.4
 Max.   :52.2   Max.   :25.0   Max.   :181.6   Max.   :75.9   Max.   :81.8
```

```
     Renda          Regiao             GrupoRenda
 Min.   :   80   EUOr  :11   (0,500]        :24
 1st Qu.:  475   SA    :12   (500,2e+03]    :24
 Median : 1690   PM    :19   (2e+03,1e+04]  :22
 Mean   : 5741   OrMd  :11   (1e+04,3.5e+04]:21
 3rd Qu.: 7325   Asia  :17   NA's           : 6
 Max.   :34064   Africa:27
 NA's   :6
```

A seguir são mostrados alguns gráficos e resumos dos dados. Inicialmente são fornecidos resumos das taxas de natalidade (`NAT`) para cada faixa de renda. A seguir uma tabela relaciona o grupo de renda com a região geográfica. Os gráficos ilustram relacionamentos entre algumas das variáveis. As últimas matrizes são de correlação de Pearson e Spearman respectivamente.

   (a) Faça interpratações estatísticas, no contexto do problema, de cada um dos resultados mostrados.

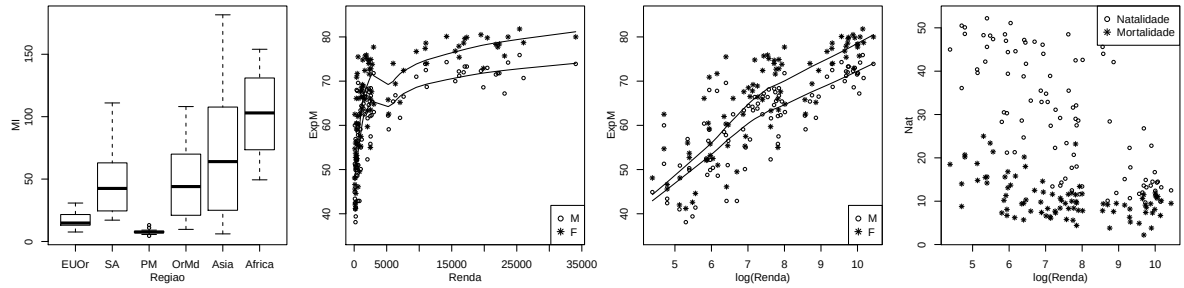
   (b) Comente ainda ao menos mais duas (2) questões de interesse que poderiam ser investigadas e que não foram abordadas nos resultados já mostrados. Indique como seriam utilizados os dados (tipo de análise) para abordar estas questões.

[1] http://www.amstat.org/publications/jse/datasets/poverty.dat.txt
[2] http://www.amstat.org/publications/jse/datasets/poverty.txt

```
$`(0,500]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21.2    38.6    44.8    41.7    48.3    52.2

$`(500,2e+03]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.4    24.4    32.9    31.8    39.6    47.2

$`(2e+03,1e+04]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   10.1    15.8    28.5    27.7    40.4    48.5

$`(1e+04,3.5e+04]`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    9.7    12.0    13.6    14.7    14.9    26.8
```

GrupoRenda	Regiao EUOr	SA	PM	OrMd	Asia	Africa
(0,500]	0	1	0	0	8	15
(500,2e+03]	5	6	0	2	3	8
(2e+03,1e+04]	4	5	3	5	1	4
(1e+04,3.5e+04]	0	0	16	3	2	0

```
X-squared
    87.64
```



	Nat	Mort	MI	ExpM	ExpF	Renda
Nat	1.0000	0.4862	0.8584	-0.8665	-0.8944	-0.6291
Mort	0.4862	1.0000	0.6546	-0.7335	-0.6930	-0.3028
MI	0.8584	0.6546	1.0000	-0.9368	-0.9554	-0.6016
ExpM	-0.8665	-0.7335	-0.9368	1.0000	0.9826	0.6430
ExpF	-0.8944	-0.6930	-0.9554	0.9826	1.0000	0.6500
Renda	-0.6291	-0.3028	-0.6016	0.6430	0.6500	1.0000

	Nat	Mort	MI	ExpM	ExpF	Renda
Nat	1.0000	0.4045	0.8861	-0.8823	-0.9018	-0.7342
Mort	0.4045	1.0000	0.4930	-0.5942	-0.5346	-0.4473
MI	0.8861	0.4930	1.0000	-0.9481	-0.9622	-0.8363
ExpM	-0.8823	-0.5942	-0.9481	1.0000	0.9784	0.8240
ExpF	-0.9018	-0.5346	-0.9622	0.9784	1.0000	0.8391
Renda	-0.7342	-0.4473	-0.8363	0.8240	0.8391	1.0000

---

3. (a) Em um levantamento do volume de vendas diário (em unidades de milhares de reais) de um site foram coletados os seguintes valores durante um certo período:

   2.8 6.8 17.6 0.4 18.1 20.5 67.8 1.0 11.4 2.7 32.3 49.4 24.5 12.2 14.0

   3.4 13.4 18.6 2.9 8.6 2.9 46.4 14.1 37.9 9.2 4.2 15.1 4.1 3.8 16.9

   i. obtenha o teor médio e o desvio padrão,

   ii. obtenha os quantis e a amplitude,

   iii. obtenha o coeficiente de variação,

   iv. obtenha um histograma,

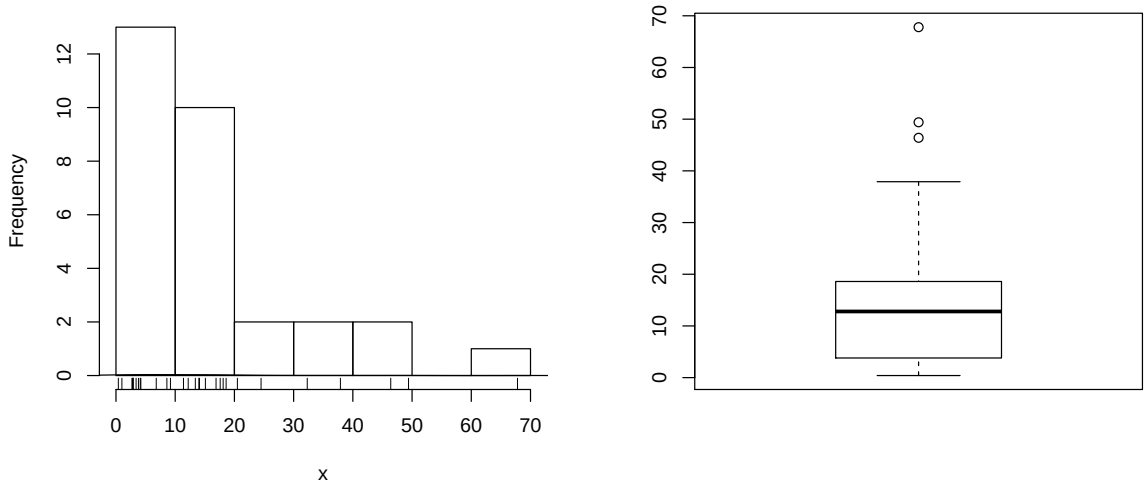


Figura 1: (d) histograma (esquerda) e (e) *box-plot* (direita) dos dados

v. obtenha um box-plot,
vi. obtenha um diagrama de ramo-e-folhas,
vii. comente sobre o padrão da distribuição dos dados e se voce consideraria alguma outra forma de analisá-los.

Solução:

i. $\overline{x} = 16.1$ e $S_x = 16.15$
ii.
```
     Q1      md      Q3 Amplitude
    3.8    12.8    18.6      67.4
```
iii. $C.V. = 100\%$
iv.
v.
vi.
```
> stem(x)
  The decimal point is 1 digit(s) to the right of the |

  0 | 0133333444799
  1 | 1234457889
  2 | 15
  3 | 28
  4 | 69
  5 |
  6 | 8
```
vii. comentários

(b) Uma cidade recebeu críticas à sua excessiva descarga de esgoto não tratado em um rio. Um microbiologista tomou 45 amostras na água depois da passagem pela planta de tratamento de esgoto e mediu a quantidade de coliformes (bactéria) presente nas amostras.

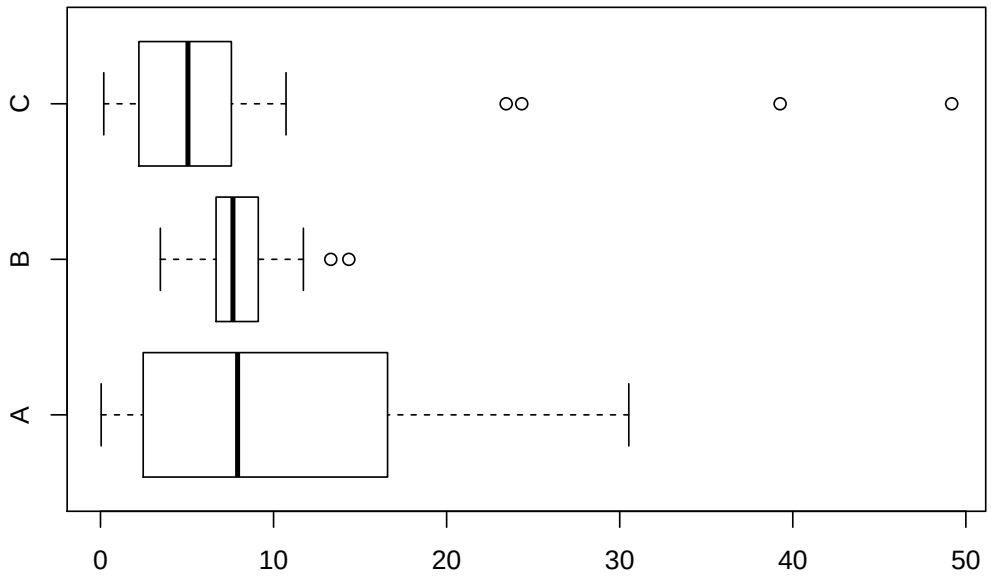
Número de Bactérias	Número de amostras
20-30	5
30-40	20
40-50	15
50-60	5

i. Obtenha a média
ii. Obtenha a mediana

Solução:

i. $\overline{x} = 39.44$
ii. $\text{md}(x) = \frac{10*(22,5-5)}{20} = 38.75$

(c) Os tempos de atendimento e solução de problemas foram medidos em três *call-centers* distintos de uma mesma empresa e os dados foram representados no gráfico a seguir. Baseando-se no gráfico, avalie cada uma das afirmações a seguir, dizendo se está certa ou errada, justificando sua resposta e corrigindo as afirmações erradas.



( ) Os valores no local C possuem uma distribuição simétrica.
( ) Os dados discrepantes do local A afetam (aumentam) a mediana do local.
( ) Os locais B e C possuem médias e desvios padrão semelhantes.
( ) O local B possui o menor coeficiente de variação.
( ) As médias dos três locais devem ser semelhantes.

---

4. (a) Três indivíduos tentam, de forma independente, resolver um problema. O primeiro tem 50% de chance de resolver, o segundo tem 65% e o terceiro tem 30%. Qual a probabilidade do problema ser resolvido?

Solução:

A : o primeiro resolve o problema $\quad B$: o segundo resolve o problema $\quad C$: o terceiro resolve o problema

$P(A) = 0,50 \quad P(B) = 0,65 \quad P(C) = 0,30$

$$P(A \cup B \cup C) = 1 - P(\overline{A} \cap \overline{B} \cap \overline{C}) \overset{ind}{=} 1 - P(\overline{A}) \cdot P(\overline{B}) \cdot P(\overline{C}) = 1 - (1 - 0,50)(1 - 0,65)(1 - 0,70) = 0.878$$

(b) Em um teste múltipla escolha, marca-se uma alternativa em cada uma das 5 questões, cada uma com quatro alternativas da qual apenas uma é correta. Qual a probabilidade de um indivíduo acertar por mero acaso alguma questão?

Solução:

$$A_i : \text{acerta a } i\text{-ésima questão } \ i = 1, \ldots, 5$$
$$P(A_i) = 0,25 \ \ P(\overline{A}_i) = 0,75 \ \ \forall i$$
$$P(\text{acertar alguma}) = 1 - P(\text{errar todas}) = 1 - P(\overline{A_1} \cap \overline{A_2} \cap \overline{A_3} \cap \overline{A_4} \cap \overline{A_5}) \overset{ind}{=} 1 - P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot P(\overline{A_3}) \cdot P(\overline{A_4}) \cdot P(\overline{A_5}) =$$
$$= 1 - (0,75)^5 = 0.763$$

(c) Dentre seis números inteiros pares e oito ímpares, dois números são escolhidos ao acaso e multiplicados. Qual a probabilidade de que o produto seja par?

Solução:

Evento	$Par \cap Par$	$Par \cap Impar$	$Impar \cap Impar$	$Impar \cap Impar$
Produto	Par	Par	Par	$Impar$
Probabilidade	$\frac{6}{14}\frac{5}{13}$	$\frac{6}{14}\frac{8}{13}$	$\frac{8}{14}\frac{6}{13}$	$\frac{8}{14}\frac{7}{13}$

$$P[ProdutoPar] = 1 - P[ProdutoImpar] = 1 - \frac{8}{14}\frac{7}{13} = 0.692$$

(d) Forneça exemplos que ilustrem situações nas quais probabilidades são avaliadas pelas definições a) clássica, b) frequentista, c) subjetiva.

---

5. (a) Considere o problema a seguir de uma avaliação semanal anterior.
Em um teste múltipla escolha, marca-se uma alternativa em cada uma das cinco questões, cada uma com quatro alternativas, entre as quais apenas uma é correta. Qual a probabilidade de um indivíduo acertar por mero acaso alguma questão?

i. Indique como fica o espaço amostral do experimento (sem necessariamente listar todos os elementos).
ii. Defina a variável aleatória (v.a) adequada ao interesse do problema.
iii. Monte uma tabela com a distribuição de probabilidades desta variável
iv. Caso possível identifique a distribuição de probabilidades desta variável e fornecendo a equação da distribuição.
v. Mostre como obter a probabilidade solicitada a partir do resultado de alguns dos itens anteriores.

Solução:

i. $\Omega = (\overline{AAAAA}), (\overline{AAAA}A), (\overline{AAA}A\overline{A}), \ldots (AAAA\overline{A}), (AAAAA) \quad n(\Omega) = 2^5 = 32$
ii. X : número de acertos
iii.

x	0	1	2	3	4	5
$P[X = x]$	$(0,75)^5$	$\binom{5}{1}(0,25)^1(0,75)^4$	$\binom{5}{2}(0,25)^2(0,75)^3$	$\binom{5}{3}(0,25)^3(0,75)^2$	$\binom{5}{4}(0,25)^4(0,75)^1$	$(0,25)^5$

iv. $X \sim B(n = 5, p = 0,25) \quad P[X = x] = \binom{5}{x}(0,25)^x(1 - 0,25)^{5-x}$
v. $P[X \geq 1] = 1 - P[X = 0] = 1 - \binom{5}{0}(0,25)^0(1 - 0,25)^{5-0} = 0.763$

(b) Identifique a v.a., liste seus possíveis valores e forneça a expressão da função de probabilidades nas situações a seguir.

i. Sabe-se que a proporção de respondentes a um anúncio é de 5%. Vou verificar quantos acessos serão feitos sem obter resposta até que seja obtida a marca de 10 respondentes.
ii. Vou escolher ao acaso 500 habitantes de Curitiba e verificar quantos sabem o nome do vice-prefeito(a) para estimar a proporção dos que conhecem.
iii. Supondo que a proporção da população que possua um determinado tipo de sangue seja de 12%, vou verificar quantos doadores vou receber até conseguir um que tenha o tipo desejado.

Solução:

i.

$$X : \text{número de acessos não respondentes até obter } 10^o \text{ respondente}$$
$$X \sim BN(k = 10, p = 0,05)$$
$$P[X = x] = \binom{x + k - 1}{x}(0,05)^{10}(0,95)^x$$

ii.

$$X : \text{número que conhecem entre os 500 entrevistados}$$
$$X \sim B(n = 500, p)$$
$$P[X = x] = \binom{500}{x}(p)^{10}(1 - p)^x$$

iii.

$$X : \text{número de doadores que não possuem o sangue desejado, até obter o que possui}$$
$$X \sim G(p = 0,12)$$
$$P[X = x] = (0,12)(1 - 0,12)^x$$

---

6. Seja uma v.a. contínua com função de distribuição de probabilidades (f.d.p) $f(x) = k(1 - x^2)\mathrm{I}_{(0,1]}(x)$, obtenha:
   (a) valor de k para que $f(x)$ seja uma f.d.p. válida,
   (b) a média de X,
   (c) a mediana de X,
   (d) a função de distribuição (acumulada) $F(x)$,
   (e) $P[X > 1/2]$,
   (f) $P[X < 0,75]$,
   (g) o primeiro quartil,
   (h) o terceiro quartil,
   (i) $P[0,25 < X < 0,75]$,
   (j) $P[X < 0,75|X > 0,5]$,

Solução:

(a)

$$\int_0^1 f(x)\mathrm{d}x = 1$$
$$k[(1 - 0)\frac{1}{3}(1^3 - 0^3)] = 1$$
$$k = \frac{3}{2}$$

(b)

$$E[X] = \int_0^1 x \cdot f(x)\mathrm{d}x = \frac{3}{2}[\frac{1}{2}(1^2 - 0^2) - \frac{1}{4}(1^4 - 0^4)] = \frac{3}{8} = 0,375$$

(c)

$$\int_0^{Md} f(x)\mathrm{d}x = 0,5$$
$$\frac{3}{2}[(Md - 0) - \frac{1}{3}(Md^3 - 0^3)] = 0,5$$
$$Md = 0.347$$

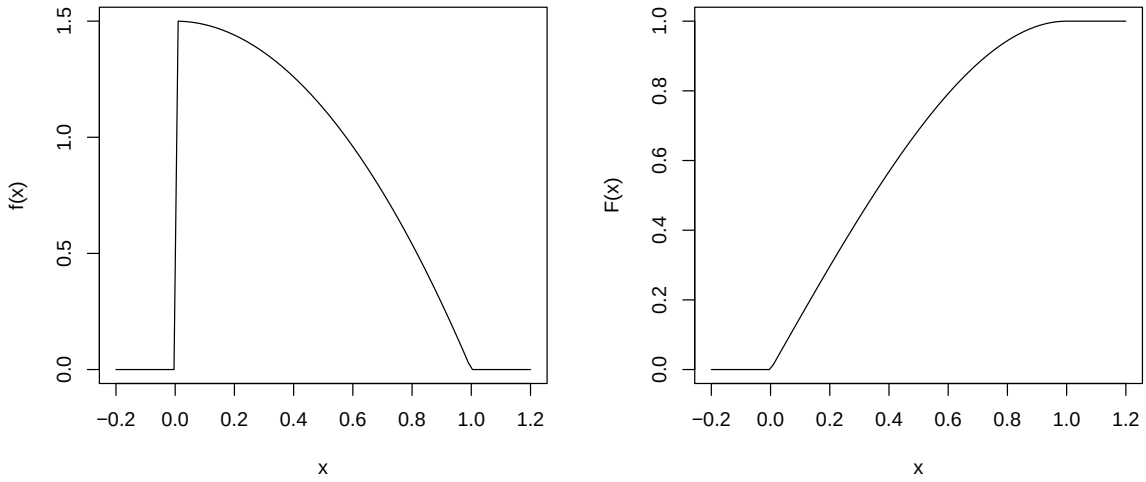


Figura 2: Função de densidade de probabilidade (esquerda) e função de distribuição.

(d) a função de distribuição (acumulada) $F(x)$,

$$F(x) = \int_0^x f(x)\mathrm{d}x = \frac{3}{2}[(x-0) - \frac{1}{3}(x^3 - 0^3)] = \frac{1}{2}(3x - x^3)$$

(e)

$$P[X > 1/2] = \int_{1/2}^1 f(x)\mathrm{d}x = 1 - F(1/2) = 0.312$$

(f)

$$P[X < 0,75] = \int_0^{0,75} f(x)\mathrm{d}x = F(0,75) = 0.914$$

(g)

$$\int_0^{Q_1} f(x)\mathrm{d}x = 0,25$$
$$\frac{3}{2}[(Q_1 - 0) - \frac{1}{3}(Q_1^3 - 0^3)] = 0,25$$
$$Q_1 = 0.168$$

(h)

$$\int_0^{Q_3} f(x)\mathrm{d}x = 0,5$$
$$\frac{3}{2}[(Q_3 - 0) - \frac{1}{3}(Q_3^3 - 0^3)] = 0,5$$
$$Q_3 = 0.558$$

(i)

$$P[0,25 < X < 0,75] = \int_{0,25}^{0,75} f(x)\mathrm{d}x = F(0,75) - F(0,25) = 0.547$$

(j)

$$P[X < 0,75 | X > 0,5] = \frac{P[0,50 < X < 0,75]}{P[X > 0,50]} = \frac{\int_{0,50}^{0,75} f(x)\mathrm{d}x}{\int_{0,50}^{1} f(x)\mathrm{d}x} = \frac{F(0,75) - F(0,50)}{1 - F(0,50)} = 0.725$$

Resoluções computacionais:

```
> require(MASS)
> ## a)
> kfx <- function(x) ifelse(x > 0 & x <= 1, (1-x^2), 0)
> fractions(1/integrate(kfx, 0, 1)$value)
```

```
[1] 3/2

> fx <- function(x) ifelse(x > 0 & x <= 1, (3/2)*(1-x^2), 0)
> integrate(fx, 0, 1)$value

[1] 1

> ## b)
> Ex <- function(x) ifelse(x > 0 & x <= 1, x*fx(x), 0)
> integrate(Ex, 0, 1)$value

[1] 0.375

> ## c)
> Qx <- function(x, quantil) (integrate(fx, 0, x)$value - quantil)^2
> (md <- optimize(Qx, c(0,1), quantil=0.5)$min)

[1] 0.3473

> ## d)
> Fx <- function(x) ifelse(x>0, ifelse(x<=1, (3*x - x^3)/2,1), 0)
> Fx(1)

[1] 1

> ## e)
> 1-Fx(1/2)

[1] 0.3125

> ## f)
> Fx(0.75)

[1] 0.9141

> ## g)
> (q1 <- optimize(Qx, c(0,1), quantil=0.25)$min)

[1] 0.1683

> ## h)
> (q3 <- optimize(Qx, c(0,1), quantil=0.75)$min)

[1] 0.5579

> ## i)
> Fx(0.75) - Fx(0.25)

[1] 0.5469

> ## j)
> (Fx(0.75) - Fx(0.5))/(1-Fx(0.5))

[1] 0.725
```

Outra forma para quantis:

```
> require(rootSolve)
> quantil <- function(p){q <- Re(polyroot(c(2*p,-3,0,1)));q[q>0&q<=1]}
> quantil(0.25)

[1] 0.1683

> quantil(0.5)

[1] 0.3473

> quantil(0.75)
```

```
[1] 0.5579
```

---

7. (a) Um sistema de climatização e refrigeração funciona continuamente e pode ocorrer uma interrupção a qualquer instante do dia com igual probabilidade.
   i. Qual a probabilidade de ocorrer falhas no período da noite, entre 20 : 00 e 6 : 00?
   ii. Qual a probabilidade de ocorrer falhas nos horários de pico de uso entre $9:00-12:00$ ou $14:00-17:30$?
   iii. Se houve falha na primeira metade do dia, qual a probabilidade de que tenha sido no horário comercial entre $8:30-12:00$?
   iv. Se houve uma falha fora do horário comercial de $9:00-18:00$, qual a probabilidade de que tenha sido de madrugada entre $0:00-5:00$
   v. Os custos de reparo variam em função do horário do dia. É de R$ 200,00 se a falha é notificada entre $9:00-17:30$, R$ 250,00 se a falha é notificada entre $6:00-9:00$ ou $17:30-20:00$ e R$350,00 para outros horários do dia. Qual o valor esperado para o pagamento de 100 reparos?

Solução:

$$X: \text{horário da falha/interrupção}$$
$$X \sim U_c[0:00, 24:00]$$
$$f(x) = \frac{1}{24-0} = \frac{1}{24}\mathrm{I}_{[0,24]}(x) \qquad F(x) = \frac{x-0}{24-0} = \frac{x}{24}$$

i. $P[20:00 < X < 24:00] + P[0:00 < X < 6:00] = \frac{4}{24} + \frac{6}{24} = \frac{10}{24} = 0.42$
ii. $P[9:00 < X < 12:00] + P[14:00 < X < 17:30] = \frac{3}{24} + \frac{3{,}5}{24} = \frac{6{,}5}{24} = 0.27$
iii. $P[8:30 < X < 12:00 | X < 12:00] = \frac{3{,}5/24}{12/24} = \frac{3{,}5}{12} = 0.29$
iv. $P[(0:00 < X < 5:00) | (0:00 < X < 9:00) \cup (18:00 < X < 24:00)] = \frac{5/24}{(9/24)+(6/24)} = \frac{5}{15} = 0.33$
v.

$$Y \sim \text{custo do reparo} \quad y \in \{200, 250, 350\}$$

y	200,00	250,00	350,00
$P[Y=y]$	$P[9:00 < X < 17:30]$	$P[(6:00 < X < 9:00) \cup (17:30 < X < 20:00)]$	$P[(0:00 < X < 6:00) \cup (20:00 < X <$
$P[Y=y]$	$\frac{8}{24}$	$\frac{6}{24}$	$\frac{10}{24}$

$$100 \cdot \mathrm{E}[Y] = 100(200 \cdot \frac{8,5}{24} + 250 \cdot \frac{5,5}{24} + 350 \cdot \frac{10}{24}) = R\$27.395,83$$

(b) Assume-se que o tempo entre conexões a um servidor tem distribuição com média de 2,5 segundos.
   i. Qual a probabilidade de se passarem 10 segundos sem conexão alguma?
   ii. Tendo havido uma conexão, qual a probabilidade de a próxima conexão não ocorrer antes de 1,5 segundos?
   iii. Qual a probabilidade do intervalo entre duas conexões ultrapassar 4 segundos?
   iv. Se já se passaram 2 segundos sem conexão, qual a probabilidade de se passaram mais 4 segundos adicionais sem conexão?
   v. Qual a probabilidade do intervalo entre conexões superar 4 segundos se já se passaram 2,5 segundos sem conexão?

Solução:
Não se especificou a distribuição e vamos assumir a distribuição exponencial considerando: (i) que devem ser valores positivos, (ii) pela possibilidade de cálculos com as informações fornecidas.

$$X: \text{intervalo de tempo entre conexões (segundos)}$$
$$X \sim \mathrm{Exp}(\lambda = 1/2,5 = 2/5)$$
$$f(x) = \frac{2}{5}\mathrm{e}^{-2x/5}\ \mathrm{I}_{(0,\infty)}(x) \qquad F(x) = 1 - \mathrm{e}^{-2x/5}$$

i. $P[X > 10] = \int_{10}^{\infty} f(x)\mathrm{d}x = 1 - F(10) = 0.018$
ii. $P[X > 1,5] = \int_{1,4}^{\infty} f(x)\mathrm{d}x = 1 - F(1,5) = 0.55$
iii. $P[X > 4] = \int_{4}^{\infty} f(x)\mathrm{d}x = 1 - F(4) = 0.2$

iv. $P[X > 6|X > 2] = \frac{\int_6^\infty f(x)\mathrm{d}x}{\int_2^\infty f(x)\mathrm{d}x} = {}^3P[X > 4] = 1 - F(4) = 0.2$

v. $P[X > 4|X > 2,5] = \frac{\int_4^\infty f(x)\mathrm{d}x}{\int_{2,5}^\infty f(x)\mathrm{d}x} = {}^3P[X > 1,5] = 1 - F(1,5) = 0.55$

---

[3] propriedade de falta de memória da exponencial