CE-009: Introdução a Estatística - Turma A
Avaliações Periódicas (1^o semestre 2017)

Avaliação 01

1. Um determinado exame tem probabilidade de 0,85 de detectar uma doença quando esta está presente enquanto que um segundo tipo de exame tem probabilidade de 0,70. A doença é considerada detectada se algum dos exames é positivo. Considere que um material com a doença vai ser testado por ambos exames.
   (a) Descreva os eventos relevantes com uma notação apropriada.
   (b) Forneça o espaço amostral na forma de um conjunto e aponte suas características (finito ou infinito, enumerável ou não enumerável, equiprovável ou não).
   (c) Defina em notação o evento "a doença é detectada" e forneça o conjunto que define este evento.
   (d) Qual a probabilidade da doença ser detectada?
   (e) Qual a suposição feita no cálculo da probabilidade do anterior?
   (f) Se três materiais com a doença forem testados com ambos exames, qual a probabilidade de que todos deem (falso) "negativo".

   Solução:

   (a) **Notação**:

   $$\begin{aligned} A &: \text{doença detectada no primeiro exame} \\ B &: \text{doença detectada no segundo exame} \\ D &: \text{a doença é detectada} \\ \overline{D} &: \text{a doença não é detectada} \\ P[A] &= 0,85 \\ P[B] &= 0,70 \end{aligned}$$

   (b) $\Omega = \{(AB), (A\overline{B}), (\overline{A}B), (\overline{AB})\}$.
   O espaço amostral é *finito*, *enumerável* e *não equiprovável*.
   (c) O evento "a doença é detectada" é definido pelo conjunto $D = \{(AB), (A\overline{B}), (\overline{A}B)\}$.
   (d) $P[D] = P[A \cup B] = P[A] + P[B] - P[A \cap B] = P[A] + P[B] - P[A] \cdot P[B] = 0,85 + 0,70 - 0,85 \cdot 0,70 = 0.955$
   (e) Independência entre os resultados dos dois exames.
   (f) Supondo independência entre materiais e exames:
   $P[\overline{D}]^3 = (1 - P[D])^3 = (1 - P[A \cup B])^3 = 0.00009113$

---

2. Um material genético de feijão em desenvolvimento foi testado quanto à resistência a duas doenças que afetam comumente a cultura. Os resultados de 100 exames são resumidos na tabela a seguir.

resistência à doença B	resistência à doença A: alta	resistência à doença A: baixa
alta	80	9
baixa	6	5

   Denote por A o evento *o material tem alta resistência à doença A* e por B o evento *o material alta resistência à doença B*.
   (a) Obtenha: $P[A]$, $P[A \cap B]$, $P[A^c]$, $P[A^c \cap B^c]$, $P[A^c \cup B]$.
   (b) Obtenha: $P[A|B]$, $P[B|A]$, $P[A|B^c]$, $P[B^c|A]$, $P[B|A^c]$.
   (c) Se um material é selecionado ao acaso qual a probabilidade de ter:
   - alta resistência a A e baixa a B?
   - alta resistência a B e baixa a A?

   (d) os eventos alta resistência a ambas doenças são mutuamente exclusivos? (justifique)
   (e) os eventos alta resistência a ambas doenças são independentes? (justifique)

Solução:

```
> m <- matrix(c(80,6,9,5),ncol=2,dimnames=list(c("A","A^c"),c("B","B^c")))
> m
     B B^c
A   80   9
A^c  6   5
> mp <- prop.table(m)
> mp
       B  B^c
A   0.80 0.09
A^c 0.06 0.05
```

(a)
- $P[A] = 0.86$
- $P[A \cap B] = 0.8$
- $P[A^c] = 0.14$
- $P[A^c \cap B^c] = 0.05$
- $P[A^c \cup B] = P[A^c] + P[B] - P[A^c \cap B] = 0.94$

(b)
- $P[A|B] = \frac{P[A \cap B]}{P[B]} = 0.9$
- $P[B|A] = \frac{P[A \cap B]}{P[A]} = 0.93$
- $P[A|B^c] = \frac{P[A \cap B^c]}{P[B^c]} = 0.55$
- $P[B^c|A] = \frac{P[B^c \cap A]}{P[A]} = 0.43$
- $P[B|A^c] = \frac{P[A^c \cap B]}{P[A^c]} = 0.64$

(c)
- $P[B \cap A^c]$ =0.06
- $P[A \cap B^c]$ =0.09

(d) Não, pois $P[A \cap B] \neq 0$, isto é, os eventos ter alta resistência a ambas doenças possuem intersecção, por isso não são mutuamente exclusivos. No contexto do exemplo, isto significa, por exemplo, que é possível ter resistência a ambas doenças ao mesmo tempo.

(e) $P[A \cap B] \neq P[A] \cdot P[B]$, isto é, o produto das marginais difere dos valores observados, por isso sabemos que os eventos não são independentes. No contexto do exemplo, as chances de ter resistência a uma doença são diferentes quando se tem ou não resistência à outra. Dizendo de outra forma, a probabilidade *marginal* (em todo universo) de se ter uma das doenças é diferente da *condicional* (no subgrupo que tem a outra doença), i.e. $P[A] \neq P[A|B]$.

```
> addmargins(mp)
       B  B^c  Sum
A   0.80 0.09 0.89
A^c 0.06 0.05 0.11
Sum 0.86 0.14 1.00
> outer(rowSums(mp), colSums(mp), "*")
         B    B^c
A   0.7654 0.1246
A^c 0.0946 0.0154
> ## note tb a ordem entre os termos!
> #outer(apply(m/sum(m),2,sum),apply(m/sum(m),1,sum),"*")
```

Avaliação 02

1. Em um lote estão misturadas 20 sementes de uma determinada cultivar (A) de uva com 5 de outra (B). Não é possível distinguir facilmente as sementes, o que só pode ser feito em um exame detalhado. Considere as diferentes possibilidades abaixo e calcule as probabilidades solicitadas.

   (a) Se forem retiradas ao acaso de uma só vez quatro sementes para inspeção, qual a probabilidade de obter ao menos uma de B?

   (b) Se forem retiradas ao acaso e inspecionadas as sementes uma a uma, retornando a semente retirada ao lote, qual a probabilidade de que sejam retiradas três ou mais de A antes de se retirar a primeira de B ?

   (c) Finalmente, considere que as sementes serão retiradas ao acaso e inspecionadas uma a uma, retornando a semente retirada ao lote antes da próxima retirada e serã feitas exatamente quatro retiradas, Qual a probabilidade de obter ao menos uma de B?

Solução:

(a)

$$X: \text{Número sementes do tipo } B \text{ entre quatro retiradas}$$
$$x \in \{0,1,2,3,4\}$$
$$X \sim \text{HG}(\text{N}=25, \text{K}=5, \text{n}=4)$$
$$P[X=x] = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} == \frac{\binom{5}{x}\binom{25-5}{4-x}}{\binom{25}{n}}$$
$$P[X \geq 1] = P[X=1] + P[X=2] + P[X=3] + P[X=4] = 1 - P[X=0] = 1 - \frac{\binom{5}{0}\binom{20}{4}}{\binom{24}{4}} = 0.617$$

(b)

$$X: \text{Número de sementes retiradas de } A \text{ até retirar a primeira de } B$$
$$X \sim \text{G}(\text{p} = 5/25 = 0,2)$$
$$x \in \{0,1,2,3,4,5,\ldots\}$$
$$P[X=x] = (1-p)^x \cdot p = (0,8)^x 0,2$$
$$P[X \geq 3] = P[X=3] + P[X=4] + P[X=5] + \ldots = 1 - P[X<3] = 1 - \{P[X=0] + P[X=1] + P[X=2]\} =$$
$$= 1 - \{(0,8)^0 0,2 + (0,8)^1 0,2 + (0,8)^2 0,2\} = 0.422$$

(c)

$$X: \text{Número sementes do tipo } B \text{ entre quatro retiradas (com reposição)}$$
$$x \in \{0,1,2,3,4\}$$
$$X \sim \text{B}(\text{n}=4, \text{p}=0,20)$$
$$P[X=x] = \binom{n}{x} p^x (1-p)^{n-x} = \binom{4}{x} 0,2^x 0,8^{4-x}$$
$$P[X \geq 1] = P[X=1] + P[X=2] + P[X=3] + P[X=4] = 1 - P[X=0] = 1 - \binom{4}{0} 0,2^0 0,8^4 = 0.59$$

Soluções computacionais com o programa R:

```
> (pa <- phyper(0, m=5, n=20, k=4, lower=F))
[1] 0.617
> (pb <- pgeom(2, p=0.25, lower=F))
[1] 0.4219
> (pc <- pbinom(0, size=4, prob=0.2, lower=F))
[1] 0.5904
```

---

2. Registros mostram que em determinado bairro são registradas, em média, 1,7 violações de residências por dia. Assuma uma distribuição de probabilidades possivelmente adequada e calcule as probabilidades de que hajam
   (a) três ou mais ocorrências em um dia;
   (b) nenhuma ocorrência em um dia;
   (c) entre uma e quatro ocorrências em um dia.

Solução:

$$X: \text{Número diário de violações}$$
$$x \in \{0,1,2,3,4,\ldots\}$$
$$X \sim \text{P}(\lambda = 1,7)$$
$$P[X=x] = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1.7}1.7^x}{x!}$$

(a) $P[X \geq 3] = P[X=3] + P[X=4] + P[X=5] \ldots = 1 - \{P[X=0] + P[X=1] + P[X=2]\} = 1 - \{\frac{e^{-1.7}1.7^0}{0!} + \frac{e^{-1.7}1.7^1}{1!} + \frac{e^{-1.7}1.7^2}{2!}\} = 1 - e^{-1.7}\{1 + 1.7^1 + \frac{1.7^2}{2!}\} = 0.243$

(b) $P[X=0] = \frac{e^{-1.7}1.7^0}{0!} = 0.183$

(c) $P[1 \leq X \leq 4] = P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] \ldots == \{\frac{e^{-1.7}1.7^1}{1!} + \frac{e^{-1.7}1.7^2}{2} + \frac{e^{-1.7}1.7^3}{3!} + \frac{e^{-1.7}1.7^4}{4!}\} = 1 - e^{-1.7}\{1.7^1 + \frac{1.7^2}{2} + \frac{1.7^3}{6} + \frac{1.7^4}{24}\} = 0.788$

Soluções computacionais com o programa R:

```
> (pa <- ppois(2, lam=1.7, lower=F))
[1] 0.2428
> (pb <- dpois(0, lam=1.7))
[1] 0.1827
> (pc <- diff(ppois(c(0,4), lam=1.7)))
[1] 0.7877
```

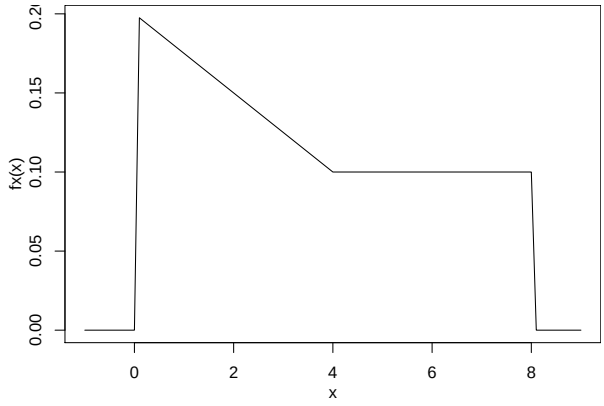
Avaliação 03

1. Uma função de densidade de probabilidade (f.d.p.) tem a expressão:

$$f(x) = \begin{cases} -0,025x + 0,2 & \text{se } 0 < x \leq 4 \\ 0,1 & \text{se } 4 < x \leq 8 \\ 0 & \text{para outros valores de x} \end{cases}$$

(a) Mostre que $f(x)$ é uma f.d.p. válida.
(b) Calcule $P[X < 3]$
(c) Calcule $P[X > 7]$
(d) Calcule $P[X < 6,5]$
(e) Calcule $P[X > 2]$
(f) Calcule $P[3 < X < 7]$
(g) Calcule $P[X < 7|X > 4]$
(h) Obtenha a tal que $P[X < a] = 0,80$
(i) Obtenha a mediana de X
(j) Estime uma valor (aproximado) para a média ($E[X]$). Este valor deve ser menor, maior ou igual à mediana? Justifique.
(k) Obtenha os quartis da distribuição de X

Solução:



(a) i. Pelo gráfico da função é possível verificar que $f(x) \geq 0 \, \forall x$,
ii. A área sob a função deve ser igual a 1. Isto pode ser verificado geometricamente:

$$A = \frac{(0,2 + 0,1) \cdot 4}{2} + 0,1 \cdot 4 = 0,6 + 0,4 = 1.$$

Alternativamente pode-se integrar a função no intervalo e verificar que

$$\int_0^8 f(x)\,\mathrm{d}x = \ldots = 1$$

(b) $P[X < 3] = 0.4875$
(c) $P[X > 7] = 0.1$

(d) $P[X < 6,5] = 0.85$
(e) $P[X > 2] = 0.65$
(f) $P[3 < X < 7] = 0.4125$
(g) $P[X < 7|X > 4] = 0.75$
(h) $P[X < a] = 0,80 \longrightarrow a = 6$
(i) $md(X) = 3.1$
(j) $E[X] = 3.47$. Maior que a mediana pois a distribuição é assimétrica (com maior cauda à direita)
(k) $Q_1(X) = 1.37$ e $Q_3(X) = 5.5$

Soluções computacionais com o programa R:

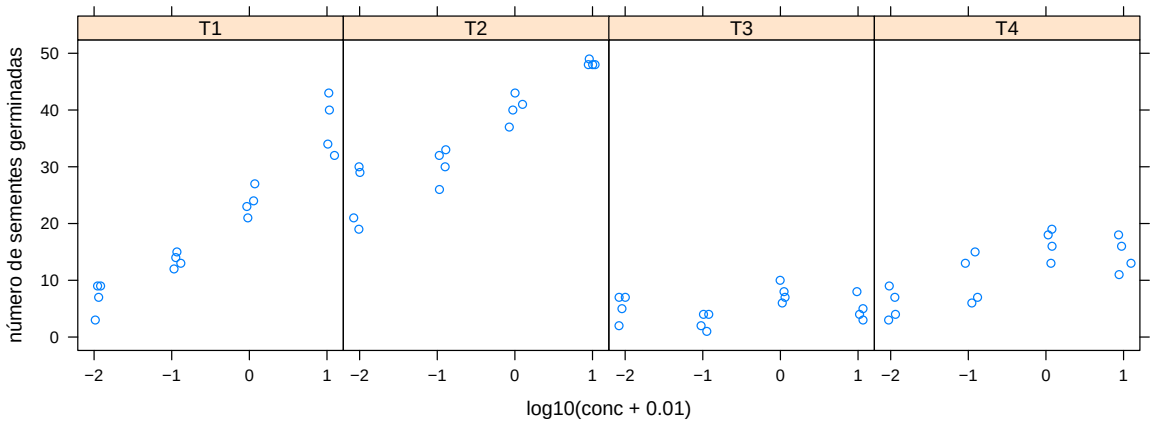
```
> fx <- function(x){
+     y <- numeric(length(x))
+     y[x > 0 & x <= 4] <- -0.025*x[x > 0 & x <= 4] + 0.2
+     y[x > 4 & x <= 8] <- 0.1
+     return(y)
+ }
> (ita <- integrate(fx, 0, 8)$value)

[1] 1

> (itb <- integrate(fx, 0, 3)$value)

[1] 0.4875

> (itc <- integrate(fx, 7, 8)$value)

[1] 0.1

> (itd <- integrate(fx, 0, 6.5)$value)

[1] 0.85

> (ite <- integrate(fx, 2, 8)$value)

[1] 0.65

> (itf <- integrate(fx, 3, 7)$value)

[1] 0.4125

> (itg <- integrate(fx, 4, 7)$value/integrate(fx, 4, 8)$value)

[1] 0.75

> Qx <- function(x, quantil) (integrate(fx, 0, x)$value - quantil)^2
> (q80 <- optimize(Qx, c(0,8), quantil=0.80)$min)

[1] 6

> (md <- optimize(Qx, c(0,8), quantil=0.5)$min)

[1] 3.101

> Ex.f <- function(x) ifelse(x > 0 & x <= 8, x*fx(x), 0)
> (Ex <- integrate(Ex.f, 0, 8)$value)

[1] 3.467

> (q1 <- optimize(Qx, c(0,8), quantil=0.25)$min)

[1] 1.367

> (q3 <- optimize(Qx, c(0,8), quantil=0.75)$min)

[1] 5.5
```

Avaliação 04

1. O conjunto da dados `agridat::mead.germination` do programa **R** contém os resultados de um experimento agronômico no qual foi verificado o efeito da concentração de um elemento químico (`conc`) e do regime de temperatura (`temp`) na germinação de sementes contando-se o número de sementes germinadas (`germ`) dentre 50 (`seeds`) inspecionadas em cada lote. Os lotes eram definidos pelas diferentes combinações das condições de temperatura e concentração, havendo ainda quatro replicações (`rep`) das diferentes condições. A seguir vemos um extrato dos dados.

```
   temp rep conc germ seeds
1    T1  R1  0.0    9    50
2    T1  R1  0.1   13    50
3    T1  R1  1.0   21    50
4    T1  R1 10.0   40    50
5    T2  R1  0.0   19    50
...
22   T2  R2  0.1   32    50
23   T2  R2  1.0   40    50
24   T2  R2 10.0   48    50
...
62   T4  R4  0.1    7    50
63   T4  R4  1.0   19    50
64   T4  R4 10.0   16    50
```

O gráfico a seguir foi feito para examinar os dados.[1]



(a) Quais as variáveis representadas do gráfico e quais os seus "tipos"?

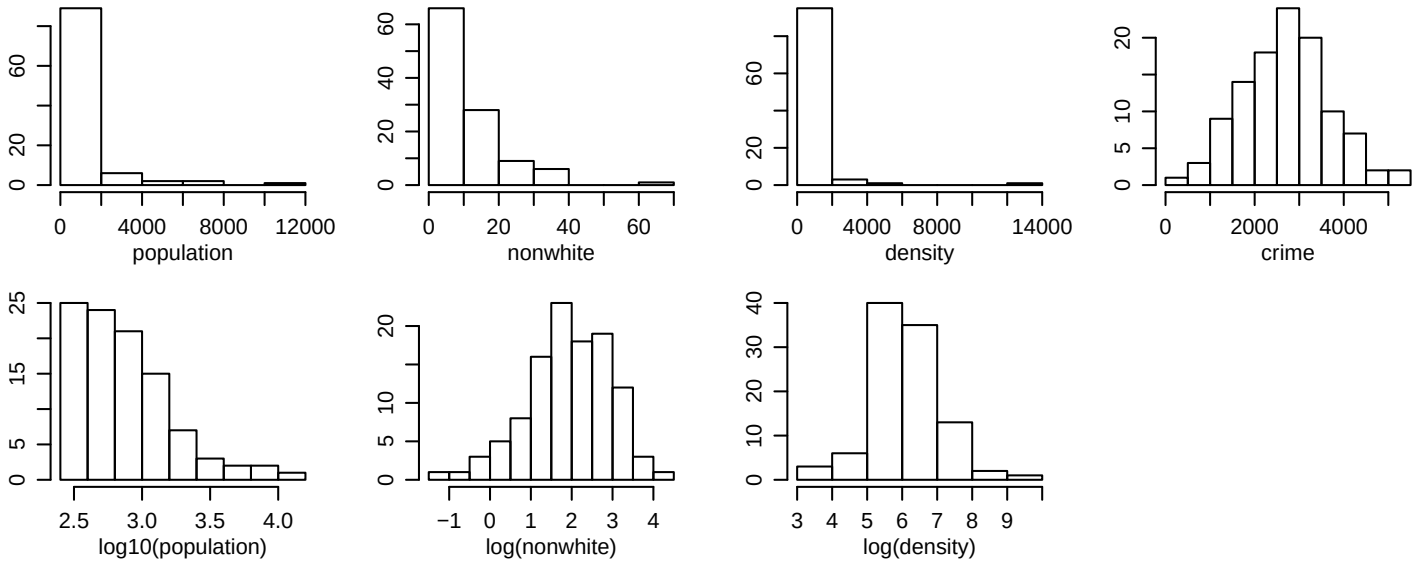
(b) Interprete o gráfico dizendo o que ele sugere em relação ao objetivo do experimento.

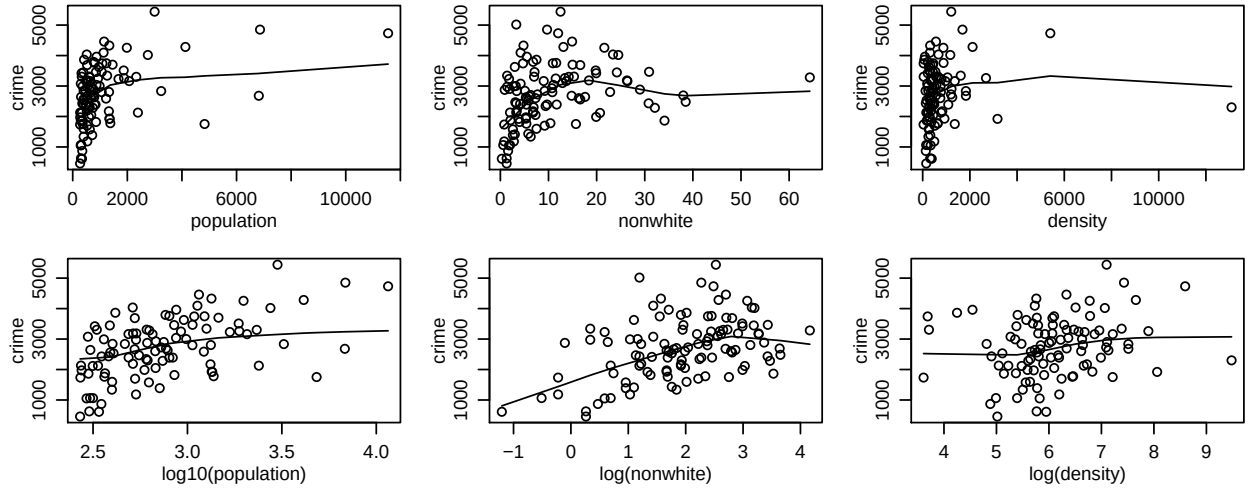
(c) Discuta porque optou-se por utilizar a concentração como `log10(conc+0.01)`.

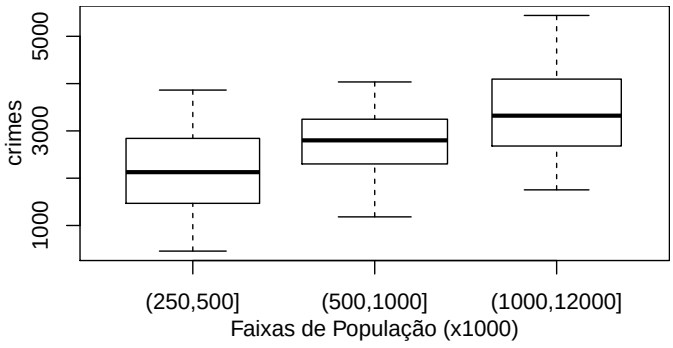
[1] os pontos foram levemente deslocados no eixo-x (*jittered*) para evitar sobreposição.

2. O conjunto de dados `car::Freedman` do programa **R** possui registros da população (`population`) em milhares de habitantes, porcentagem de não brancos (`nonwhite`), densidade populacional (`density`) e número de crimes (`crimes`) em 110 áreas metropolitanas com população acima de 250 mil habitantes dos Estados Unidos no ano de 1968. A tabela de medidas estatísticas e gráficos abaixo apresentam resumos dos dados a serem interpretados. Comece esboçando como seria o formato da tabela dos dados. Identifique os tipos de variáveis e discuta todos os resultados. Inclua ainda nos comentários o que voce espera dos valores de correlação entre número de crimes e demais variáveis.[2]

	n	media	desvioP	min	max	amplitude	Q0.25	Q0.5	Q0.75	CV
population	100	1136.0	1560.14	270.0	11551.0	11281	398.8	664.0	1167.75	137.34
nonwhite	110	10.8	10.26	0.3	64.3	64	3.4	7.2	14.88	94.97
density	100	765.7	1441.95	37.0	13087.0	13050	266.5	412.0	773.25	188.33
crime	110	2714.1	991.40	458.0	5441.0	4983	2066.8	2698.0	3305.00	36.53







[2] `log10()` é o logarítmo na base 10 enquanto que `log()` é o logarítmo neperiano

Avaliação 05

1. Foram escolhidos ao acaso 500 animais (bovinos) de uma região para estimar a proporção de com propenção à uma certa doença. Destes, 120 testaram positivo.
   (a) Obtenha a estimativa pontual do percentual de susceptíveis na população.
   (b) Obtenha a estimativa intervalar (com confiança de 95%) do percentual de susceptíveis na população.
   (c) Idem anterior porém com confiança de 80%.
   (d) Deseja-se extender o levantamento para obter uma margem de erro de 1,5% para 95% de confiança. Quantos animais adicionais devem ser selecionados e testados?

   População:

   X : propenção à doença, 0 - não propenso, 1 - propenso
   $X \sim B(p)$
   p : parâmetro desconhecido

   Amostra:

   Dados:

$$n = 500 \ ; \ \sum_{i=1}^{500} x_i = 120$$

   Estimador:

$$\hat{p} = \frac{\sum_{i=1}^{500} x_i}{n}$$

   Estimativa:

$$\hat{p} = \frac{120}{500} = 0,24$$

   Distribuição amostral:

$$\hat{p} \sim \mathrm{N}(\mu_{\hat{p}} = p, \sigma^2_{\hat{p}} = \frac{p(1-p)}{n})$$

   (a) $\hat{p} = 0,24$
   (b) Intervalo assintótico (tomando $p = \hat{p}$)

$$\hat{p} \pm z_{0.95}\sqrt{\frac{p(1-p)}{n}}$$
$$0.24 \pm 1.96 \cdot 0.019$$
$$0.24 \pm 0.0374$$
$$(0.203 \ ; \ 0.277)$$

   (c)

$$\hat{p} \pm z_{0,80}\sqrt{\frac{p(1-p)}{n}}$$
$$0.24 \pm 1.28 \cdot 0.019$$
$$0.24 \pm 0.0245$$
$$(0.216 \ ; \ 0.264)$$

   (d)

$$\mathrm{ME} = z_{0.95}\sqrt{\frac{p(1-p)}{n}}$$
$$0,015 = 1.96^2\sqrt{\frac{0.24(1-0.24)}{n}}$$
$$n = \lceil 1.96^2 \frac{0.24(1-0.24)}{0,015^2}$$
$$n = 3115$$

   Portanto 2615 novas amostras.

Soluções computacionais com o programa R:

```
> prop.test(120, 500, conf=0.95)$conf
[1] 0.2037 0.2804
attr(,"conf.level")
[1] 0.95
> prop.test(120, 500, conf=0.80)$conf
[1] 0.2154 0.2663
attr(,"conf.level")
[1] 0.8
```

---

2. Um pecuarista possui um rebanho e monitora o peso dos animais periodicamente. Como o rebanho é muito grande ele escolhe alguns animais ao acaso de cada vez. Um uma determinada pesagem ele seleciona 20 animais que mostram um peso médio de 430 kg com desvio padrão de 20 kg.
   (a) Obtenha uma estimativa intervalar (90% de confiança) para o peso do rebanho.
   (b) Se ele pesasse 10 animais, qual seria o intervalo de confiança?
   (c) Obtenha um intervalo de confiança (90% de confiança) para a variância do peso.

   População:

$$\begin{aligned} &\text{X : peso de cada animal} \\ &\text{X} \sim \text{N}(\mu, \sigma^2) \\ &\mu \,,\, \sigma^2 \text{ : parâmetros desconhecidos} \end{aligned}$$

   Amostra:

$$\begin{aligned} &\text{Dados:} \\ &n = 20 \;\; ; \;\; \overline{x} = 430 \\ &\text{Estimadores:} \\ &\hat{\mu} = \overline{x} = \frac{\sum_{i=1}^{500} x_i}{n} \;\; ; \;\; \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{500} (x_i - \overline{x})^2}{n-1} \\ &\text{Estimativas:} \\ &\overline{x} = 430 \;\; ; \;\; S^2 = 20^2 \\ &\text{Distribuições amostrais:} \\ &\overline{x} \sim \text{N}(\mu, \sigma^2/n) \;\; ; \;\; S^2 \sim \chi^2(\nu = n-1) \end{aligned}$$

   (a)

$$\begin{aligned} &\overline{x} \pm t_{0.90,19} \frac{S}{\sqrt{n}} \\ &430 \pm 1.73 \cdot 4.47 \\ &430 \pm 7.73 \\ &(422.3 \;\; ; \;\; 437.7) \end{aligned}$$

   (b)

$$\begin{aligned} &\overline{x} \pm t_{0.90,9} \frac{S}{\sqrt{n}} \\ &430 \pm 1.83 \cdot 6.32 \\ &430 \pm 11.6 \\ &(418.4 \;\; ; \;\; 441.6) \end{aligned}$$

   (c)

$$\begin{aligned} &\left(\frac{(n-1)S^2}{\chi^2_{sup}} \;\; ; \;\; \frac{(n-1)S^2}{\chi^2_{inf}} \right) \\ &\left(\frac{(20-1)20^2}{30.1} \;\; ; \;\; \frac{(20-1)20^2}{10.1} \right) \\ &(252 \;\; ; \;\; 751) \end{aligned}$$

   E para o desvio padrão teríamos:

$$\left(\sqrt{252} \;\; ; \;\; \sqrt{751} \right) \longrightarrow (15.9 \;\; ; \;\; 27.4)$$
