MI404 - Métodos Estatísticos
Mariana Rodrigues Motta
Departamento de Estatística, Universidade de Campinas (UNICAMP)

PROVA 1
-
EXERCÍCIOS

Henrique Aparecido Laureano

Abril de 2017

Sumário

Exercício 1

Considere uma variável aleatória X = consumo (em gramas) de vegetais de $n = 188$ indivíduos. A seguir considere o gráfico abaixo e responda.

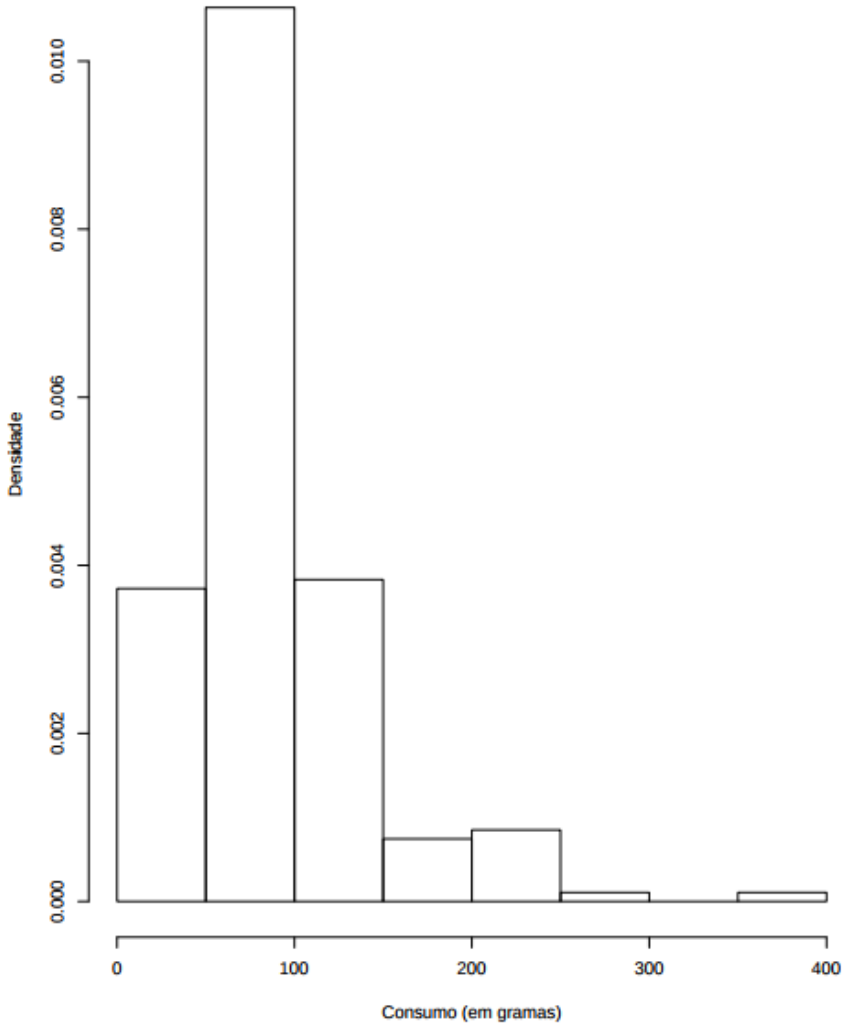

Figura 1: Histograma do consumo de vegetais (em gramas).

(a)

O que você observa sobre o histograma acima?

Observa-se inicialmente que o consumo de vegetais dos indivíduos sob estudo varia de O à 400 gramas, com maior concentração entre 50 e 100 gramas. Este intervalo de maior concentração apresenta mais que o dobro de indivíduos que o segundo intervalo com maior frequência de consumo (intervalo 100 à 150 gramas). Feita esta observações, podemos afirmar que o consumo de vegetais destes indivíduos segue uma distribuição unimodal (com uma única moda).

(b)

O que significa a escala do eixo y ser chamada de densidade?

Significa que a escala do histograma apresentado é dada em densidade de probabilidade, desta modo, o histograma tem uma área total de um.

(c)

Neste exemplo, quem será maior: a média ou a mediana? Justifique.
Considere $\bar{X} = 87.2$ e $S = 49.14$, em que $\bar{X}$ é a média amostral e S o desvio padrão amostral.

Este histograma apresenta assimetria a direita (cauda direita mais longo e fina), e em histogramas com tal característica a média é maior que a mediana. Isso é dado pelo fato do conjunto de dados possuir poucas observações com valores altos que aumentam a média mas que não afetam a localização do meio exato do conjunto de dados (i.e., a mediana).

Para ter uma noção da mediana, $\tilde{x}$, podemos usar a seguinte inequação:

$$\max(x_{(1)}, \bar{x} - s_n) \leq \tilde{x} < \bar{x},$$

em que

$$s_n = \sqrt{\frac{n-1}{n}} s_{n-1}.$$

O menor valor da amostra é x_1, o desvio padrão viesado da amostra é s_n e o desvio padrão não viesado é s_{n-1}.

Assim, $\max(0, 87.2 - \sqrt{187/188} \cdot 49.14) \leq \tilde{x} < 87.2 \rightarrow 38.19 \leq \tilde{x} < 87.2$.

(d)

Sabendo que '*skewness* é uma medida de assimetria de uma determinada distribuição de frequência' e que na distribuição normal a *skewness* é zero, a estimativa da *skewness* da distribuição dos dados apresentados no histograma acima deve ser positiva, negativa ou zero? Justifique.
Note que *skewness* $= \mu_3/\sigma^3$, em que $\mu_3 = E[(X - E(X))^3]$ e σ é o desvio padrão populacional. Aproximadamente 60% das observações estão abaixo da média.

Estamos numa situação em que a mediana é menor que a média, desta forma, a *skewness* da distribuição dos dados é positiva.
A *skewness* pode ser mensurada por

$$\text{Skew} = 3 \cdot \frac{\bar{x} - \tilde{x}}{s_{n-1}}.$$

Como sabemos que a média é maior que a mediana, $\bar{x} > \tilde{x}$, a *skewness* será maior que zero, Skew > 0.

(e)

As distribuições lognormal e gama seriam duas candidatas para ajustar esses dados. Usando métodos computacionais e gráficos, como você escolheria a mais apropriada? Justifique.

Pensando em ajuste de modelos de regressão, poderíamos ajustar dois modelos lineares generalizados, um com resposta lognormal e outro com resposta gama. O modelo com melhor ajuste, baseado numa análise de resíduos, responderia a pergunta de qual distribuição é mais adequada aos dados.

Não pensando em ajuste de modelos de regressão, dois gráficos poderiam ser feitos pra responder essa pergunta. Um `q-q plot` que compararia as duas funções de distribuição através da comparação gráfica dos quantis das distribuições. Esse gráfico ajudaria na visualização de diferenças nas caudas das distribuições. O outro gráfico útil seria um gráfico normal de probabilidade que representaria a comparação dos quantis de cada distribuição com uma normal padrão. Tal gráfico também seria válido na comparação das caudas das distribuições.

Na Figura 2 temos a aplicação desses dois gráficos em duas amostras simuladas de tamanho 188 das distribuições lognormal e gama com igual média, 90, e desvio padrão, 50. Na comparação das distribuições observamos uma grande fuga na calda direita da distribuição lognormal.

Pensando no conjunto de dados do exercício e olhando para os resultados dos gráficos, escolheríamos a distribuição lognormal como mais apropriada, já que seus dados simulados apresentam a mesma característica dos dados originais.

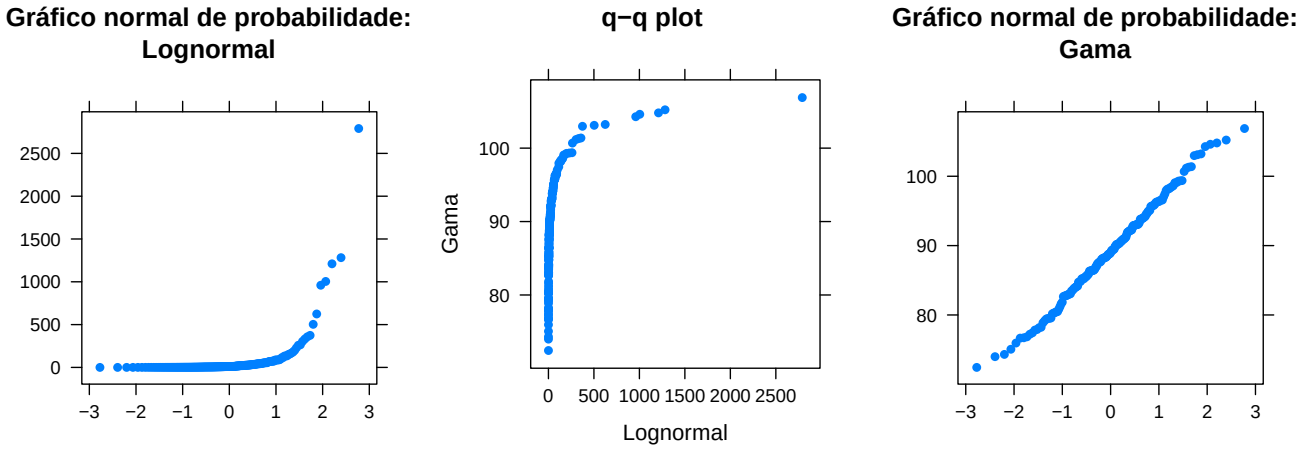

Figura 2: q-q plot e gráficos normais de probabilidade para duas amostras simuladas de tamanho 188 de distribuições lognormal e gama de igual média, 90, e desvio padrão, 50.

Exercício 2

Seja X uma variável aleatória de distribuição gama com parâmetros a e b dada por

$$f(x;a,b) = \frac{b}{\Gamma(a)}(bx)^{a-1}e^{-bx}, \tag{1}$$

tal que $E(X) = a/b$ e $Var(X) = a/b^2$. Reparametrize a distribuição em (1) de tal forma que $E(X) = \mu$ e $Var(X) = \mu/\phi$. Escreva a distribuição em (1) a partir desta nova parametrização.

$$\mu = \frac{a}{b} \qquad \text{e} \qquad \frac{\mu}{\phi} = \frac{a}{b^2},$$

logo,

$$a = \mu b \qquad \text{e} \qquad \frac{\mu}{\phi} = \frac{\mu b}{b^2} = \frac{\mu}{b} \Rightarrow \boxed{\phi = b} \qquad \text{e} \qquad \boxed{a = \mu\phi}.$$

$$\boxed{f(x;\mu,\phi) = \frac{\phi}{\Gamma(\mu\phi)}(\phi x)^{\mu\phi-1}e^{-\phi x}}.$$

Exercício 3

Suponha que indivíduos entrevistados sobre o consumo de vegetais informem consumos positivos ou zero (sem consumo) e que você tenha decidido pela distribuição gama para modelar os dados de consumo de vegetais positivos. Escreva uma distribuição de probabilidade para estes dados, considerando que $f(x)$ é a distribuição gama em (1) com parâmetros μ e ϕ, onde μ é a média e ϕ parâmetro de escala. Para construir a distribuição de probabilidade dos dados positivos e zeros use a idéia do modelo ZIP visto em sala de aula.

$$P(X = x) = \begin{cases} p & ; x = 0 \\ (1-p)f(x;\mu,\phi) & ; x \in (0,\infty) \end{cases},$$

$$\text{ZIG}(x;p,\mu,\phi) = p^{\,\text{I}(x=0)}\left[(1-p)f(x;\mu,\phi)\right]^{\,\text{I}(x>0)}.$$

ZIG: *Zero-Inflated Gamma.*

Em que p é a probabilidade de ter consumo de vegetais.

Exercício 4

Considere uma normal tetravariada originada da seguinte forma:

$$y_1 = a_1 + e_1 \qquad (2)$$
$$y_2 = a_2 + e_2$$

em que $E(a_i) = E(e_i) = 0$, $Cov(a_i, e_j) = 0$, para $i, j = 1, 2$. Além disso, assuma que $Cov(e_i, e_j) = 0$, $i \neq j$, $Var(a_i) = \sigma_a^2$ e $Var(e_i) = \sigma_e^2$, $i, j = 1, 2$. Assuma que a distribuição de a_i e e_i seja normal.

(a)

Qual a distribuição do vetor $\mathbf{d} = (y_1, y_2, a_1, a_2)$?

$$D = \begin{bmatrix} y_1 \\ y_2 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 + e_1 \\ a_2 + e_2 \\ a_1 \\ a_2 \end{bmatrix} \sim N_4(\mu, \Sigma),$$

em que

$$\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_a^2 + \sigma_e^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} \\ \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 + \sigma_e^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 \\ \sigma_a^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} \\ \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 & \rho_{a_1a_2}\sigma_{a_1}\sigma_{a_2} & \sigma_a^2 \end{bmatrix}.$$

Outra maneira de representar:

$$D = \begin{bmatrix} y_1 \\ y_2 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 + e_1 \\ a_2 + e_2 \\ a_1 \\ a_2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ e_1 \\ e_2 \end{bmatrix}}_{X} = AX,$$

$$AX \sim N_4(A\mu, A\Sigma A').$$

(b)

Usando a ideia de probabilidade condicional, mostre como encontrar a distribuição $p(y_1, y_2|a_1, a_2)$. Pode deixar indicado quando tiver alcançado uma resposta que depende das distribuições conhecidas no problema.

$$p(y_1, y_2|a_1, a_2) = \frac{p(y_1, y_2, a_1, a_2)}{p(a_1, a_2)} = \frac{p(y_1, y_2, a_1, a_2)}{\int\int p(y_1, y_2, a_1, a_2)\mathrm{d}y_1\mathrm{d}y_2}.$$

(c)

Encontre $E(y_i)$ e $Var(y_i)$ usando o método da iteração a partir da $E(y_i|a_i) = a_i$ e $Var(y_i|a_i) = \sigma_e^2$. Mostre que $E(y_i) = 0$ e $Var(y_i) = \sigma_a^2 + \sigma_e^2$.

$$E(y_i) = E_{a_i}[E(y_i|a_i)] = E_{a_i}[a_i] = 0.$$

$$
\begin{aligned}
Var(y_i) &= E_{a_i}[Var(y_i|a_i)] + Var_{a_i}[E(y_i|a_i)] \\
&= E_{a_i}[\sigma_e^2] + Var_{a_i}[a_i] \\
&= \sigma_e^2 + \sigma_a^2.
\end{aligned}
$$

Exercício 5

Considere o seguinte mecanismo de amostragem: (1) Jogue uma moeda em que a probabilidade de ser 'cara' é 0.3; (2) Se você observa cara, considere uma realização $X \sim N(2, 1)$; se for 'coroa' considere uma realização $X \sim N(8, 1)$. O código abaixo ilustra este experimento a partir de simulação:

```
# <code r> ============================================================ #
coin <-
  rbinom(1000
         , size = 1, prob = 0.3) # gera 1000 amostras de uma v.a. Ber(0.3)
n1 <- rnorm(sum(coin), 2, 1) # gera um total de soma(coin) de N(2, 1)
n2 <- rnorm(1000 - sum(coin)
            , 8, 1) # gera um total de [1000 - soma(coin)] de N(8, 1)
# </code r> ============================================================ #
```

O histograma destas 1000 amostras é dado por

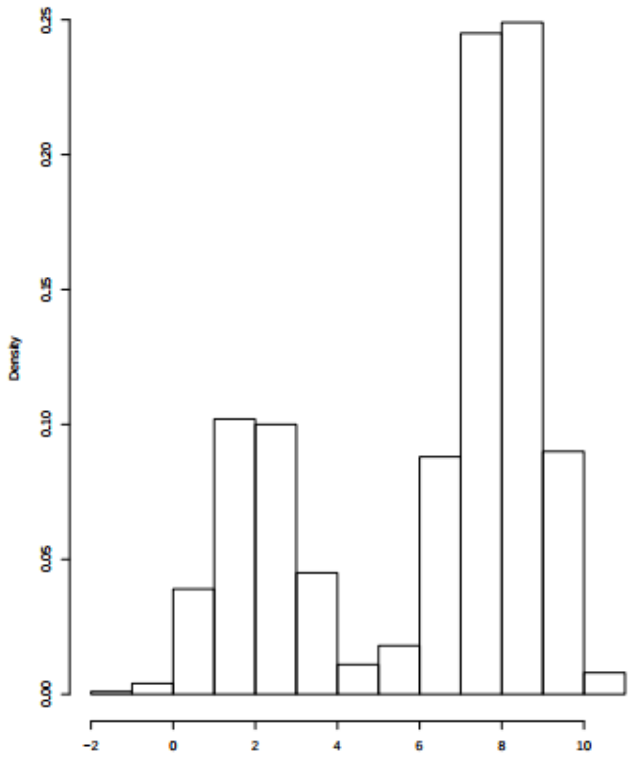

Figura 3: Histograma das 1000 v.a.'s geradas pelo processo acima.

Considere que uma v.a. X gerada através do processo acima. Escreva a distribuição de probabilidade de X.

$$f_{X|\theta}(x;\theta) = [0.3 \cdot N(2,1)]^{\mathrm{I}(\theta=0)} \cdot [0.7 \cdot N(8,1)]^{\mathrm{I}(\theta=1)},$$

Em que θ é o resultado de jogar a moeda (cara = 0, coroa = 1).
