ACH0021 – Tratamento e Análise de Dados/Informações (2014.1)

Primeira Prova – Maio/2014

Nome: ______________________________ Nº USP: ____________

Turma/Horário: ____________________ Curso: ____________________

Observação 1: Duração da prova: **75** (**setenta e cinco**) minutos.
Observação 2: O uso de calculadora é opcional, e seu empréstimo durante a prova é **proibido**.

Formulário (conjunto de n elementos $\{w_i\}$)

Média: $\overline{w} = \frac{1}{n}\sum_{k=1}^{n} w_k$ Variância: $\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(w_k - \overline{w})^2$ Desvio padrão: σ

1) Uma empresa de transportes mediu o tempo que um ônibus necessita para ir da cidade X à cidade Y, registrando o tempo de percurso 100 vezes, constatando que este tempo varia bastante conforme as condições do tempo e estrada. Os dados de sua pesquisa estão organizados na tabela abaixo.

Tempo de percurso (minutos)	Frequência absoluta
$04q \vdash 06q$	10
$06q \vdash 11q$	40
$11q \vdash 14q$	30
$14q \vdash 18q$	20
TOTAL	100

a) [2,5 pontos] Estimar a média $\overline{t}$ do tempo de percurso e o desvio padrão σ.

b) [3,0 pontos] Estimar o **número de viagens** cujo tempo de percurso foi inferior a $7,5q$ minutos.

Nota: Explicitar/justificar o raciocínio na resolução.

Prova A: $q = 6$ Prova B: $q = 8$ Prova C: $q = 10$ Prova D: $q = 12$

1a) Assumindo uma distribuição uniforme dos dados em cada intervalo da variável e tomando o ponto médio t_i de cada um destes como sendo o respectivo representante, tem-se

Tempo de percurso (min)	t_i (min)	Frequência absoluta (n_i)	Frequência relativa (f_i)	Amplitude (Δ_i)	Densidade (d_i)
$04q \vdash 06q$	$05{,}00q$	10	$10/100 = 0{,}10$	$2q$	$1/(20q)$
$06q \vdash 11q$	$08{,}50q$	40	$40/100 = 0{,}40$	$5q$	$2/(25q)$
$11q \vdash 14q$	$12{,}50q$	30	$30/100 = 0{,}30$	$3q$	$1/(10q)$
$14q \vdash 18q$	$16{,}00q$	20	$20/100 = 0{,}20$	$4q$	$1/(20q)$
TOTAL	-	100	1,00	-	-

Estimativa da média dos $n = 100$ dados:

$$\overline{t} = \frac{1}{n}\sum_i n_i t_i = \frac{1}{100}\left[10 \cdot 5,00q + 40 \cdot 8,50q + 30 \cdot 12,50q + 20 \cdot 16,00q\right] = \frac{217q}{20} = 10,85q \text{ (min)}.$$

Estimativa da variância:

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_i n_i\left(t_i - \overline{t}\right)^2 = \frac{1}{100}\Big[10\,(5,00q - 10,85q)^2 + 40\,(8,50q - 10,85q)^2 + \\ &\quad +30\,(12,50q - 10,85q)^2 + 20\,(16,00q - 10,85q)^2\Big] = \frac{4701q^2}{400},\end{aligned}$$

que implica um desvio padrão de $\sigma = \sqrt{\frac{4701}{400}q}$ (min). Em suma, tem-se

$$\begin{cases} \text{Prova A } (q=6): & \text{Estimativas: } \bar{t} = 65,1 \text{ (minutos) e } \sigma \approx 20,6 \text{ (minutos)} \\ \text{Prova B } (q=8): & \text{Estimativas: } \bar{t} = 86,8 \text{ (minutos) e } \sigma \approx 27,4 \text{ (minutos)} \\ \text{Prova C } (q=10): & \text{Estimativas: } \bar{t} = 108,5 \text{ (minutos) e } \sigma \approx 34,3 \text{ (minutos)} \\ \text{Prova D } (q=12): & \text{Estimativas: } \bar{t} = 130,2 \text{ (minutos) e } \sigma \approx 41,1 \text{ (minutos)} \end{cases}.$$

1b) O ponto $7,5q$ situa-se na faixa $06q \vdash 11q$, e deseja-se saber $f_{06q \vdash 7,5q}$ que, juntamente com $f_{04q\vdash 06q} = 0,10$ (vide tabela do exercício (1a)), fornece $f_{04q\vdash 7,5q} = f_{04q\vdash 06q} + f_{06q\vdash 7,5q}$, que é a fração das viagens com duração inferior a $7,5q$ minutos. Assumindo uma distribuição uniforme dos dados em cada intervalo da variável, e da tabela do exercício (1a), tem-se:

$$d_{06q\vdash 7,5q} = d_{06q\vdash 11q} \Leftrightarrow \frac{f_{06q\vdash 7,5q}}{\underbrace{\Delta_{06q\vdash 7,5q}}_{=1,5q}} = \frac{2}{25q},$$

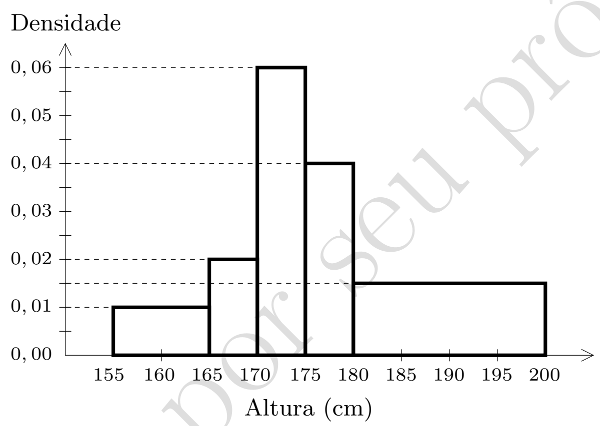
donde é imediato que $f_{06q\vdash 7,5q} = 0,12$. Logo, estima-se que $f_{04q\vdash 7,5q} = f_{04q\vdash 06q} + f_{06q\vdash 7,5q} = 0,10 + 0,12 = 0,22$ do total de dados foi inferior a $7,5q$ minutos, e isto corresponde a $100 \times 0,22 = 22$ viagens.

Nota: Notar que o resultado do exercício (1b) independe do valor de q, e a resposta é a mesma para as quatro provas.

---

2) Um diretor de cinema deseja selecionar algumas atrizes que tenham altura entre a cm e h cm (com $h > a$), sendo que as alturas das candidatas que tentam o papel é dada no histograma abaixo.

Histograma da altura das atrizes (cm)



a) [3,5 pontos] Determianar h para que sejam selecionadas somente p do total de candidatas. Dar a resposta com ao menos **uma** casa decimal.

b) [1,0 ponto] Considere um conjunto $\{x_i\}$ que contém $n-1>0$ elementos. Acrescenta-se y a este conjunto, que passa a ter n elementos. Mostrar, **através de cálculos**, qual deve ser o valor de y para que a dispersão dos n dados seja mínima.

Hint: $az^2 + bz + c = a\left(z + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$

Nota: A questão (2b) é **independente** de (2a).

Prova A: $a = 156$ e $p = 74\%$
Prova B: $a = 158$ e $p = 78\%$
Prova C: $a = 162$ e $p = 82\%$
Prova D: $a = 164$ e $p = 87\%$

2a) O histograma assume uma distribuição uniforme dos dados em cada barra, e como $a \in [155, 165)$ (em cm), tem-se

$$d_{a\vdash 165} = d_{155\vdash 165} \Rightarrow \frac{f_{a\vdash 165}}{\Delta_{a\vdash 165}} = 0,01 \Rightarrow f_{a\vdash 165} = \frac{165-a}{100}.$$

Ademais, sabe-se que $f_{165\vdash 170} = d_{165\vdash 170}\Delta_{165\vdash 170} = 0,02\cdot 5 = 0,10$, $f_{170\vdash 175} = d_{170\vdash 175}\Delta_{170\vdash 175} = 0,06\cdot 5 = 0,30$ e $f_{175\vdash 180} = d_{175\vdash 180}\Delta_{175\vdash 180} = 0,04\cdot 5 = 0,20$. Logo, $f_{165\vdash 180} = f_{165\vdash 170} + f_{170\vdash 175} + f_{175\vdash 180} =$

$0,10+0,30+0,20=0,60$. Logo, a altura h tal que $f_{a\vdash h}=p$ situa-se na faixa $180\vdash 200$. Por conseguinte,

$$f_{180\vdash h}=p-f_{a\vdash 180}=p-f_{a\vdash 165}-f_{165\vdash 180}=p-\frac{165-a}{100}-0,60\,.$$

Como se tem uma distribuição uniforme dos dados nas barras do histograma,

$$d_{180\vdash h}=d_{180\vdash 200}\Rightarrow\frac{f_{180\vdash h}}{\Delta_{180\vdash h}}=0,015\Rightarrow\Delta_{180\vdash h}=\frac{1}{0,015}f_{180\vdash h}=\frac{1}{0,015}\left(p-\frac{165-a}{100}-0,60\right)\,,$$

donde

$$\begin{aligned}h&=180+\frac{1}{0,015}\left(p-\frac{165-a}{100}-0,60\right)\\&=30+\frac{2a}{3}+\frac{200p}{3}\quad(\text{cm})\,.\end{aligned}$$

Desta forma,

$$\begin{cases}\text{Prova A } (a=156 \text{ e } p=0,74): & h=\dfrac{550}{3}\approx 183,3\text{ cm}\\[2ex]\text{Prova B } (a=158 \text{ e } p=0,78): & h=\dfrac{562}{3}\approx 187,3\text{ cm}\\[2ex]\text{Prova C } (a=162 \text{ e } p=0,82): & h=\dfrac{578}{3}\approx 192,7\text{ cm}\\[2ex]\text{Prova D } (a=164 \text{ e } p=0,87): & h=\dfrac{592}{3}\approx 197,3\text{ cm}\end{cases}\,.$$

2b) Sendo a média $\overline{x}$ dos n dados igual a

$$\overline{x}=\frac{1}{n}\left(S+y\right)\,,\qquad\text{onde}\qquad S:=\sum_{i=1}^{n-1}x_i\,,$$

a variância σ^2 é

$$\begin{aligned}\sigma^2&=\frac{1}{n}\left\{\left[x_1-\frac{1}{n}\left(S+y\right)\right]^2+\cdots+\left[x_{n-1}-\frac{1}{n}\left(S+y\right)\right]^2+\left[y-\frac{1}{n}\left(S+y\right)\right]^2\right\}\\&=\frac{1}{n}\left[\sum_{i=1}^{n-1}x_i^2+y^2-2\sum_{i=1}^{n-1}x_i\left(\frac{S+y}{n}\right)-2y\left(\frac{S+y}{n}\right)+n\left(\frac{S+y}{n}\right)^2\right]\\&=\frac{1}{n^2}\left[\left(n-1\right)y^2-2Sy+\left(nT-S^2\right)\right]\,,\end{aligned}$$

onde $T:=\sum_{i=1}^{n-1}x_i^2$. Logo,

$$\sigma^2=\frac{n-1}{n^2}\left[\left(y-\frac{S}{n-1}\right)^2+\frac{nT-S^2}{n-1}-\left(\frac{S}{n-1}\right)^2\right]\,,$$

e esta função de y assume o valor mínimo em $y_{\min}$ quando o termo não-negativo $\left(y-\frac{S}{n-1}\right)^2$ for nulo, conduzindo a

$$y_{\min}=\frac{1}{n-1}\sum_{i=1}^{n-1}x_i\,,$$

que é a média dos $n-1$ dados iniciais. Geometricamente, o gráfico de σ^2 é uma parábola convexa em y, e o seu ponto mínimo ocorre em $y=y_{\min}$.