1a Prova de Estatística Computacional CE223. 16/04/2008

A solução fornecida não é única.

1. A tabela abaixo mostra a distribuição de frequências de 499 trabalhadores desempregados em função da idade, sexo e duração do período de desemprego.

duração	Abaixo de 35 anos		Acima de 35 anos	
(dias)	Feminino	Masculino	Feminino	Masculino
1-7	36	48	44	43
8-30	48	42	38	49
Mais de 30	30	43	42	36

Mostre os comandos do `R` para

(a) entrar com estes dados em uma estrutura adequada,

(b) encontrar o número total de pessoas em cada faixa etária,

(c) encontrar a proporçao de homens e mulheres que ficou desempregada por mais de 30 dias,

(d) encontrar a probabilidade de uma pessoa sorteada ao acaso deste grupo ter menos de 35 anos e ser do sexo masculino,

Trata-se de uma tabela de tripla entrada, ou seja a distribuição conjunta das frequencias de 3 variaveis categoricas (duração com 3 valores, faixa etaria com 2 valores e sexo com 2 valores). Então uma estrutura adequada para estes dados é um `array` com 3 dimensões.

(a)

```
> dados = c(36,48,30,48,42,43,44,38,42,43,49,36)
> tab= array(dados,dim=c(3,2,2), dimnames=list(
+          c('1-7','8-30','mais de 30'),
+          c('Feminino','Masculino'),
+          c('abaixo de 35','acima de 35')))
```

(b) A variável faixa etária está na terceira dimensão, então o total é

```
> sum(tab[,,1])
[1] 247
> sum(tab[,,2])
[1] 252
```

ou

```
> apply(tab,3,sum)
abaixo de 35  acima de 35
         247          252
```

(c) A variavel "duração" está na primeira dimensão, então a proporção é

```
> prop.table(tab[3,,])
          abaixo de 35 acima de 35
Feminino     0.1986755   0.2781457
Masculino    0.2847682   0.2384106
```

ou por faixa etaria

```
> prop.table(tab[3,,1])
 Feminino Masculino
0.4109589 0.5890411
> prop.table(tab[3,,2])
 Feminino Masculino
0.5384615 0.4615385
```

(d) $P(A, B) = P(B|A)P(A)$, A="abaixo de 35", B="sexo masculino". Então a frequencia do sexo masculino dentre os que estao abaixo de 35 anos é

```
> sum(tab[,2,1])
[1] 133
```

e $P(B|A)$ é

```
> AB = sum(tab[,2,1])/sum(tab[,,1])
> AB
[1] 0.5384615
```

A frequencia dos que estao abaixo de 35 anos é

```
> sum(tab[,,1])
[1] 247
```

e $P(A)$ é

```
> A = sum(tab[,,1])/sum(tab)
> A
[1] 0.49499
```

Então a probabilidade é

```
> AB * A
[1] 0.2665331
```

2. Suponha que foi criada a variável `m` com o seguinte comando do `R`,

```
> m = data.frame(peso=c(62,73,74,65,62,68,73,61,66,64),
                 altura=c(154,173,145,145,155,185,156,151,156,180),
                 sexo=c(2,1,2,2,2,1,2,2,2,1))
```

sendo que `sexo=1` significa sexo masculino e `sexo=2` sexo feminino. Mostre os comandos do `R` para,

(a) transformar a variável `sexo` em um fator com 2 níveis,

(b) obter um resumo da variável peso segundo o sexo,

(c) fazer um boxplot da variável altura segundo o sexo,

(d) fazer o histograma com a densidade estimada superimposta para a variável peso.

(a) Criando o fator,

```
> m$sexo = factor(m$sexo,levels=1:2,labels=c('Masculino','Feminino'))
> head(m)
  peso altura      sexo
1   62    154  Feminino
2   73    173 Masculino
3   74    145  Feminino
4   65    145  Feminino
5   62    155  Feminino
6   68    185 Masculino
```

(b) Resumo de peso por sexo,

```
> by(m$peso,m$sexo,summary)
m$sexo: Masculino
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.00   66.00   68.00   68.33   70.50   73.00
------------------------------------------------
```

```
m$sexo: Feminino
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  61.00   62.00   65.00   66.14   69.50   74.00
```

(c) 2 boxplots, um para cada sexo

```
> boxplot(m$altura ~ m$sexo)
```

(d) histograma usando frequencias relativas com a densidade estimada,

```
> hist(m$peso, prob=TRUE)
> lines(density(m$peso))
```

3. Escreva uma função que receba dois vetores: o primeiro é um vetor de dados e o segundo um vetor que indica o grupo a qual pertence cada dado correspondente do primeiro vetor. O vetor de grupos deve admitir apenas 2 grupos. A função deve retornar a média, desvio padrão, coeficiente de variação, mediana, primeiro e terceiro quartis para cada um dos grupos. A função deve retornar estas estatísticas mesmo que haja valores perdidos (NA) nos dados. Além disto a função deve ter um argumento opcional para definir se as mesmas estatística serão também retornadas para todos os dados, sem separação por grupos, com *default* para não retornar.