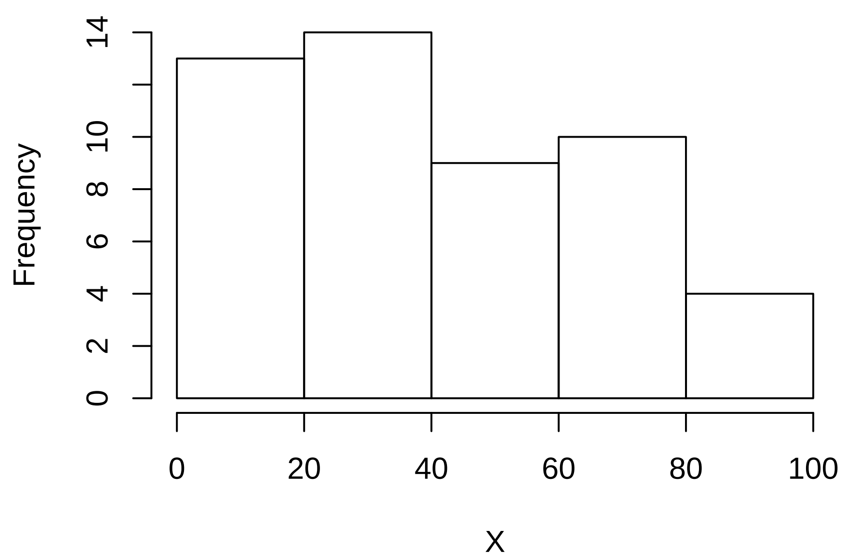
# Introdução à Probabilidade e à Estatística I
# Resolução Lista 1

Professor: Pedro Morettin & Chang Chiann

1.

(a) Podemos inserir dados no software R e construir um histograma com 5 intervalos:




(b) Uma medida de posição é a média $\bar{X}$ e uma medida de dispersão é a variância $Var(X) = s_X^2$. No software R, o comando é o seguinte (também pode ser feito no Excel ou outro pacote estatístico):

```
> mean(X)

[1] 40.42
```

```
> var(X)

[1] 680.0037
```

2.

(a) Quando cada observação é multiplicada por 2, temos o seguinte para as medidas (assume-se que X é a matriz de dados original e Y a matriz multiplicada por 2, ou seja, $Y_i = 2X_i$):

*Média*:
$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
$\bar{Y} = \frac{\sum_{i=1}^{n} 2 \cdot X_i}{n} = 2 \cdot \bar{X}$

*Variância*:
$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2}{n-1}$

$s_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} (2 \cdot X_i)^2 - n\bar{Y}^2}{n-1}$

Segue que:
$s_Y^2 = \frac{\sum_{i=1}^{n} 4 \cdot {X_i}^2 - 4n\bar{X}^2}{n-1} = 4 \cdot s_X^2$

*Mediana*:

$med(Y) = 2 \cdot med(X)$

Note que usando os dados do problema 1, a mediana é a média aritmética entre a observação de número 25 e a observação de número 26, após a ordenação dos dados em ordem crescente. Logo, qualquer transformação que não altere a ordenação e que seja realizada em todas as observações (neste caso, a multiplicação por 2) impactará da mesma forma a mediana.

(b) Seguindo os passos do item anterior poderíamos provar que (assume-se que X é a matriz de dados original e Y a matriz somada do valor 10, ou seja, $Y_i = X_i + 10$):

*Média*: $\bar{Y} = \bar{X} + 10$

*Variância*: $s_Y^2 = s_X^2$ (variância de uma constante é zero)

*Mediana*: $med(Y) = med(X) + 10$

(c) Agora suponha que $Y_i = (X_i - \bar{X})$. Assim, temos:

*Média*: $\bar{Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - \frac{n\bar{X}}{n} = 0$

*Variância*: $s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = s_X^2$

(novamente, dado que $\bar{X}$ é uma constante, não afeta a variância)

*Mediana*: $med(Y) = med(X) + \bar{X}$

(d) Agora suponha que $Y_i = \frac{(X_i - \bar{X})}{s_X}$

*Média*: $\bar{Y} = \frac{\sum_{i=1}^{n} \frac{(X_i - \bar{X})}{s_X}}{n} = \frac{\sum_{i=1}^{n} X_i}{n s_X} - \frac{n\bar{X}}{n s_X} = 0$

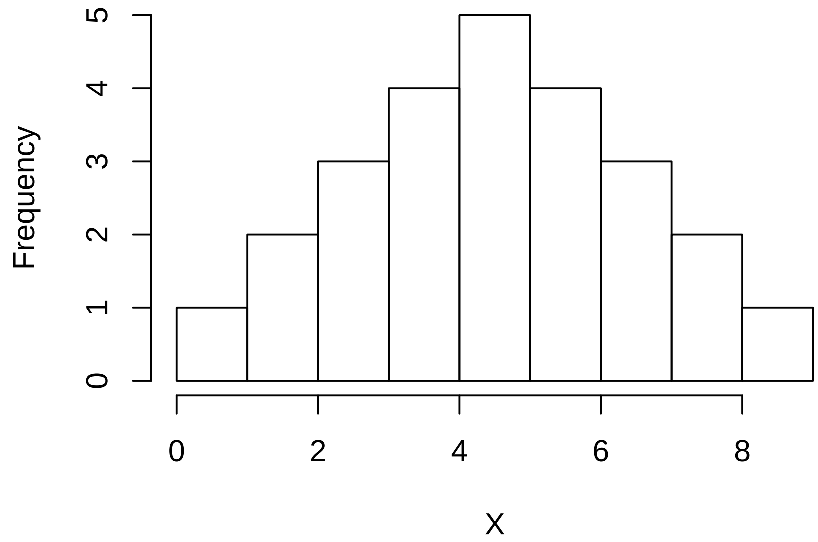
*Variância*: $s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)s_X^2} = \frac{s_X^2}{s_X^2} = 1$

3.

(a) A mediana é uma medida resistente, pois é pouco afetada por valores extremos, ao passo que a média não. Por exemplo, se temos os seguintes dados observados: 5,7,8,10,12,15. A mediana $med(X)$ é 9 e a média $\bar{X}$ é 9,5. Porém, se último valor ao invés de 15 torna-se 150, teríamos $med(X) = 9$ e $\bar{X} = 32$. Ou seja, a mediana é mais adequada quando a distribuição é assimétrica à direita ou à esquerda.

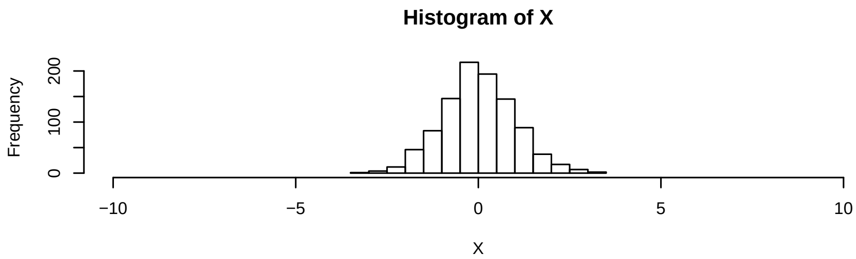
(b) Sempre que os dados forem simétricos, teremos $med(X) = \bar{X}$. Por exemplo, usando o software R:

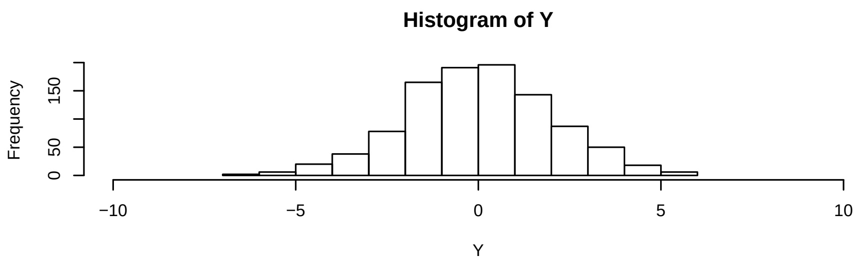
```
> X <- c(1,2,2,3,3,3,4,4,4,4,5,5,5,5,5,6,6,6,6,7,7,7,8,8,9)
> median(X)

[1] 5

> mean(X)

[1] 5

> hist(X,breaks=c(0,1,2,3,4,5,6,7,8,9),main="")
```

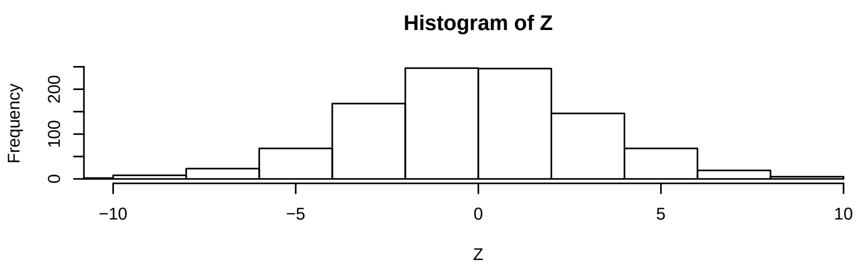



(c) Podemos usar variáveis normais com médias iguais e variâncias crescentes. No software R:

```
> X<-rnorm(1000,mean=0,sd=1)
> Y<-rnorm(1000,mean=0,sd=2)
> Z<-rnorm(1000,mean=0,sd=3)
> par(mfrow=c(3,1))
> hist(X,xlim=c(-10,10))
> hist(Y,xlim=c(-10,10))
> hist(Z,xlim=c(-10,10))
```




Ou seja, o histograma tem aparência mais dispersa, quanto maior a variância.

4.

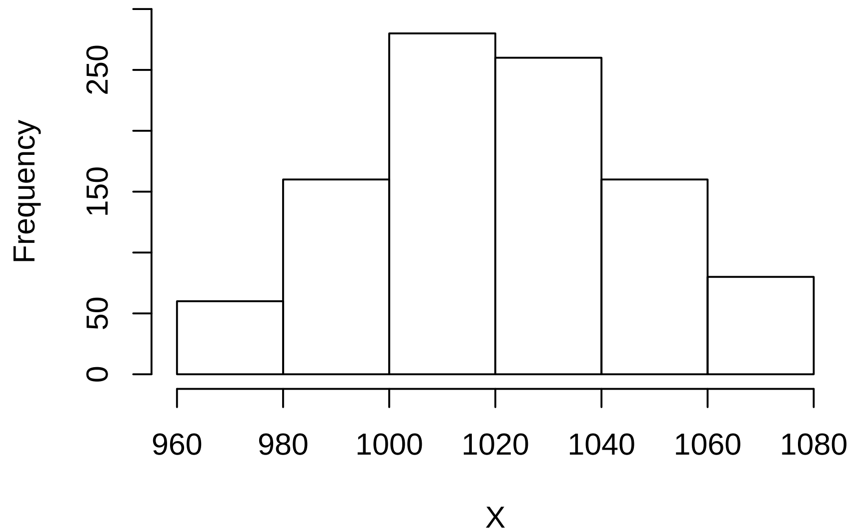
(a) Denominando por $X$ a variável aleatória do peso dos frangos, temos que a média aproximada $Me(X)$ neste caso é obtida usando os pontos médios $(x_1, x_2, ..., x_6)$ de cada classe e as respectivas frequências $(n_1, n_2, ..., n_6)$ com que cada classe de peso ocorre:

$$Me(X) = \frac{970 \cdot 60 + 990 \cdot 160 + 1010 \cdot 280 + 1030 \cdot 260 + 1050 \cdot 160 + 1070 \cdot 80}{1000} = 1020,8$$

(b) A variância aproximada é obtida de forma semelhante, usando como se os pontos médios $(x_1, x_2, ..., x_6)$ fossem as observações e considerando a média obtida no item (a).

$Var(X) = \frac{\sum_{i=1}^{n} n_i (x_i - Me(X))^2}{n} = 691,36$

(c) O histograma fica o seguinte:




(d) Podemos calcular os quantis/ percentis e achar os limites das quatros categorias da seguinte forma (note que para isso tivemos que acumular os percentuais). No software R, usamos o código:

```
> quantile(X,probs=c(0.2,0.5,0.8,1))

 20%  50%  80% 100%
 990 1020 1050 1070
```

Logo, o limites de peso ($X$) das categorias são os seguintes:
A: $X <= 990$ gramas
B: $990 < X <= 1020$ gramas
C: $1020 < X <= 1050$ gramas
D: $X > 1050$ gramas

(e) Sabemos que a média é de 1020,8 gramas e a variância é de 691,36, logo o desvio padrão é de 26,29. Assim, o granjeiro decide separar os animais com peso inferior à (1020,8-2*26,29)=968,21 para receberem ração reforçada e os com peso superior à 1060,24 para serem reprodutores. Podemos então obter os percentuais de frangos sujeitos a esses tratamentos específicos calculando a área do histograma abaixo de 968,21 e a área acima de 1060,24 e dividindo pela área total. Assim, temos que 2,46% dos frangos receberão ração reforçada e 7,9% serão separados para reprodução.

5.

(a) Temos que calcular a média aproximada da distribuição, considerando os pontos médios de cada intervalo, conforme o exercício 4.

$Me(X) = \dfrac{19 \cdot 18 + 21 \cdot 12 + 24 \cdot 10 + 28 \cdot 8 + 33 \cdot 2}{50} = 22,48$

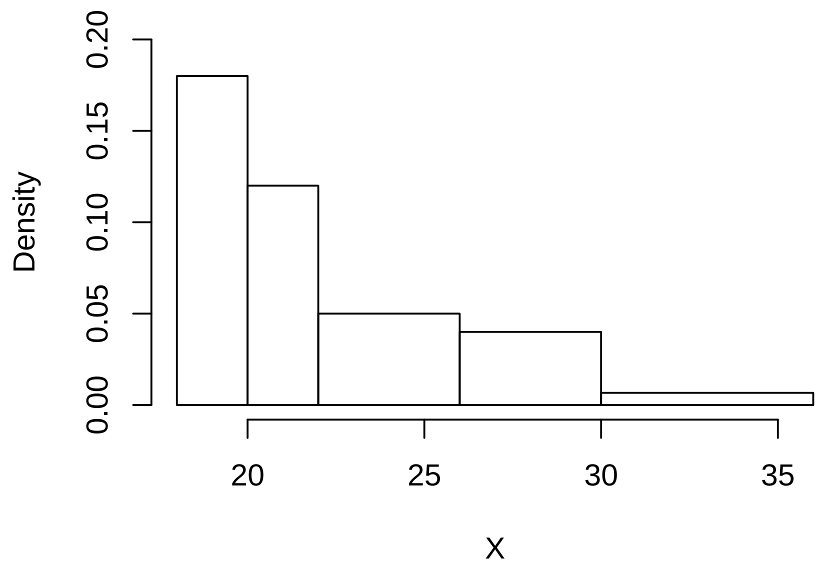
Sendo assim, como a média aproximada é superior a 22, temos que a campanha surtiu algum efeito.

(b) Vamos calcular a variância aproximada considerando os pontos médios:

$Var(X) = \frac{\sum_{i=1}^{n} n_i (x_i - Me(X))^2}{n} = 14,64$

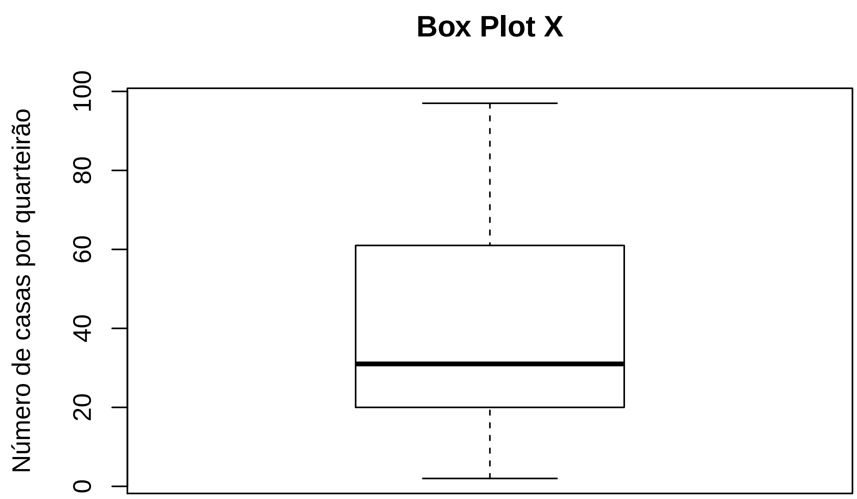
Logo, $dp(X) = 3.82$ e $\dfrac{2dp(X)}{sqrt(n)} = 1,083$. Como $Med(X) - 22 = 0,48$, temos que a campanha não surtiu efeito, segundo esta regra.

(c) O histograma, considerando os pontos médios é o seguinte. Nota: o software R somente produz gráficos com valores equidistantes no eixo $x$, porém os intervalos das classes não são equidistantes. Contudo, é possível notar que a primeira classe vai de 18 a 20, a segunda vai de 20 a 22, a terceira de 22 a 26 e assim por diante, terminando em X=36.



Density
0.00
0.05
0.10
0.15
0.20
20
25
30
35
X


6.



Box Plot X
Número de casas por quarteirão
0
20
40
60
80
100

7.

(a) Vamos calcular os quantis da distribuição do Ex.1 usando o software R:

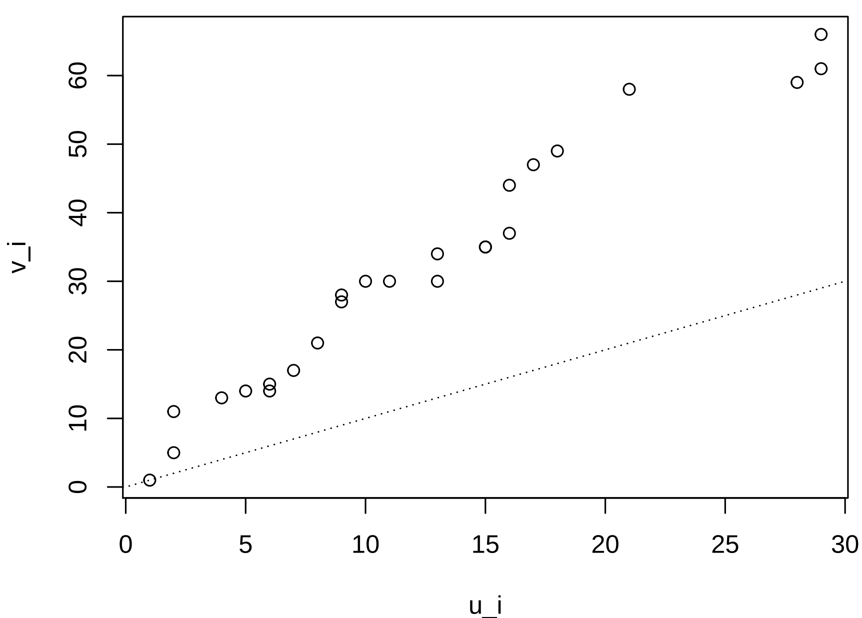
```
> quantile(X,probs=c(0.1,0.4,0.8,0.9))

 10%  40%  80%  90%
13.9 25.6 65.2 78.2
```

(b) Usamos o mesmo comando, porém agora com as probabilidades dos quartis:

```
> quantile(X,probs=c(0.25,0.5,0.75))

  25%   50%   75%
20.25 31.00 61.00
```

(c) O desvio absoluto mediano é igual à mediana dos desvios absolutos dos dados em relação à mediana. Usando os dados, temos que $med(X) = 31$. Logo, a medida é igual à 17.

8.



Ou seja, os dados claramente não são simétricos.

9.

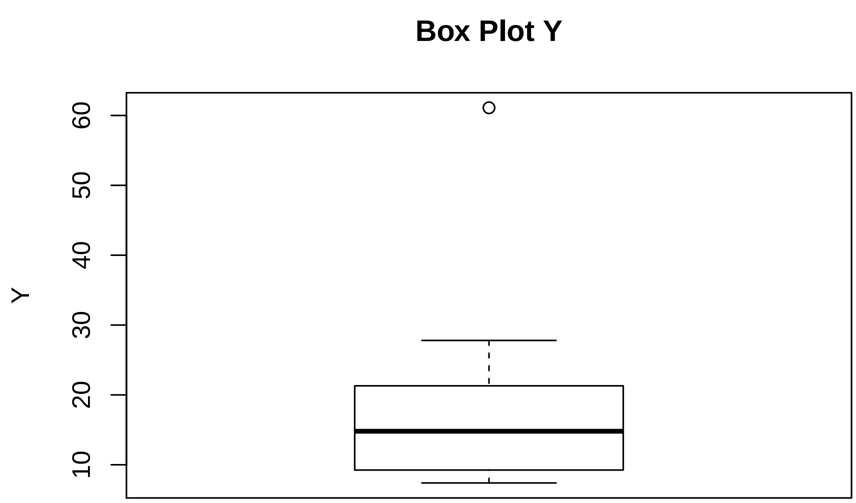
(a) Chamando de Y a variável que representa os dados, temos que a média, a mediana, o desvio padrão e o desvio médio são, respectivamente:

$Me(Y) = 18,4$
$Med(Y) = 14,8$
$dp(Y) = 13,066$
$dm(Y) = 8,76$

(b) A média aparada a 10% é obtida eliminando 10% das menores e das maiores observações e calculando a média aritmética dessa distribuição aparada. Como se tratam de 15 observações, vamos eliminar as duas menores e as duas maiores observações, obtendo uma média aparada de 15,66.
Já o desvio absoluto mediano é calculado da seguinte forma: calculam-se os desvios absolutos de cada observação com relação à mediana e daí a mediana desses desvios. Como Med(Y)=14,8, temos que o desvio absoluto mediano é 7,4.

(c) Pelo gráfico podemos ver que há um valor atípico:

```
> boxplot(Y,main = "Box Plot Y",ylab = "Y")
```



10. Podemos calcular algumas medidas usando o software R:

```
> #medidas de posição
>
> media<-mean(ibv)

[1] 18619.97

> mediana<-median(ibv)

[1] 13245

> #medidas de dispersão
>
> desvio_padrao<-sd(ibv)

[1] 14823.54

> desvio_absoluto_mediano<-mad(ibv, center = median(ibv),constant=1)

[1] 6445
```

(observação: o desvio padrão no R é calculado dividindo por n-1, ou seja, corresponde ao desvio padrão amostral $s_X^2$)