UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos

ZAB1111 – Estatística Básica

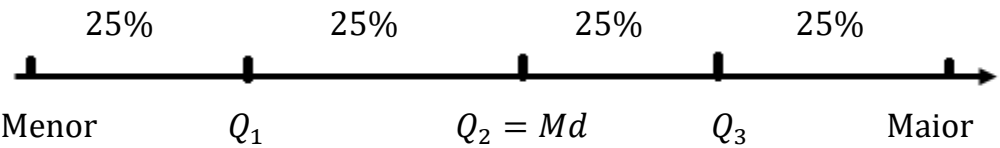
Separatrizes e *box-plot*

SEPARATRIZES são medidas de ordem que têm a propriedade de deixar à sua esquerda certa porcentagem das observações da série ordenada. As principais separatrizes são os **quartis** e os **percentis**.

QUARTIS (Q_j, $j = 1$, 2 e 3) são os valores que dividem a série ordenada em 4 partes iguais (Galton,1882).

- Q_1: é o valor da série que ocupa a posição $(n + 1)/4$ e deixa 25% dos valores à sua esquerda e 75% dos valores à sua direita.
- Q_2: é o valor da série que ocupa a posição $2(n + 1)/4$ e deixa 50% dos valores tanto à sua esquerda quanto à sua direita, ou seja, Q_2 é a mediana da série.
- Q_3: é o valor da série que ocupa a posição $3(n + 1)/4$ e deixa 75% dos valores à sua esquerda e 25% dos valores à sua direita.

Ilustrando:



PERCENTIS (P_j, $j = 1, 2, \ldots, 99$) são os valores que dividem a série ordenada em 100 partes iguais (Galton, 1885).

- P_j é o valor que ocupa a posição $j(n + 1)/100$ e deixa j% dos valores da série ordenada à sua esquerda e $(100-j)$% dos valores à sua direita.

Note que: $Q_1 = P_{25}$, $Q_2 = Md(X) = P_{50}$ e $Q_3 = P_{75}$

Exemplo 1. Altura de 45 alunos de Engenharia de Biossistemas

1,45	1,48	1,53	1,53	1,59	1,60	1,61	1,61	1,62	1,62
1,63	1,63	1,64	1,64	1,64	1,65	1,65	1,66	1,66	1,66
1,66	1,67	1,67	1,67	1,68	1,68	1,69	1,71	1,71	1,71
1,71	1,71	1,72	1,72	1,73	1,77	1,78	1,79	1,80	1,81
1,82	1,83	1,86	1,87	1,88					

1) Qual é a altura limite que deixa à sua esquerda 10% dos alunos de menor estatura?

   O P_{10} é o elemento da série ordenada que ocupa a posição:

$$10(45 + 1)/100 = 4{,}6º$$

Usando interpolação simples para obtenção do P_{10}:

$$P_{10} = x_4 + 0{,}6(x_5 - x_4) = 1{,}53 + 0{,}6(1{,}59 - 1{,}53) = 1{,}57\text{m}$$

⇒ Podemos concluir que "1,57m é a altura abaixo da qual estão 10% dos alunos de EB" ou que "90% dos alunos de EB têm altura acima de 1,57m".

2) Calcular e interpretar P_{85}

P_{85} é a altura que ocupa a posição $85(45+1)/100 = 39{,}1º$

$$P_{85} = x_{39} + 0{,}1(x_{40} - x_{39})$$
$$= 1{,}80 + 0{,}1(1{,}81 - 1{,}80) = 1{,}80\text{m}$$

⇒ "1,80m é a altura abaixo da qual estão 85% dos alunos de EB" ou "acima de 1,80m estão 15% dos alunos de EB".

Exemplo 2. Calcular Q_1, $Q_2 = Md$, Q_3, P_{10} e P_{90} dos pesos de coelhos ao desmame apresentados a seguir:

492	552	560	583	657	657	666	697	699	716
727	731	737	750	770	798	808	817	823	823
830	842	842	860	873	878	880	883	900	910
940	960	960	963	992	1000	1000	1000	1020	1040

Separatriz	Posição	Peso (g)
$Q_1 = P_{25}$	10,25º	$716 + 0{,}25(727 - 716) = 718{,}8g$
$Q_2 = Md = P_{50}$	20,50º	$823 + 0{,}50(830 - 823) = 826{,}5g$
$Q_3 = P_{75}$	30,75º	$910 + 0{,}75(940 - 910) = 932{,}5g$

Continuando:

Separatriz	Posição	Peso (g)
P_{10}	4,10º	$583 + 0{,}10(657-583) = 590{,}4g$
P_{90}	36,90º	$1000 + 0{,}9(1000-1000) = 1000g$

Comentários:

- Se quisermos descartar 10% dos coelhos mais leves, separamos os animais com pesos iguais ou inferiores a $590{,}4g$.
- Se quisermos manter no plantel 10% dos coelhos mais pesados, separamos os animais com pesos iguais ou superiores a $1000g$.

Se os dados de uma variável contínua estiverem classificados em uma distribuição de frequências usamos a fórmula:

- $P_j = L_{P_j} + \dfrac{\frac{jn}{100} - F_a}{f_{P_j}} h$

  Em que L_{P_j} é o limite inferior da classe que contem P_j; j é a ordem do percentil; n é o número de elementos da série; F_a é a frequência acumulada da classe anterior à classe que contem P_j; f_{P_j} é a frequência absoluta da classe que contem P_j e h é a amplitude desta classe.

  O P_j (percentil de ordem j) estará na classe que contém o elemento de ordem $\dfrac{jn}{100}$.

Podemos calcular a porcentagem de valores abaixo de um certo valor P_j da série usando a fórmula:

$$j = \left[\frac{\left(P_j - L_{P_j}\right) f_{P_j}}{h} + F_a\right] \frac{100}{n}$$

Exemplo 3. Calcular P_{25}, P_{75} e P_{90} dos pesos de coelhos ao desmame já classificados na distribuição de frequências. Qual a porcentagem de coelhos com peso superior a 1000 gramas?

Peso (g)	f_i	F_i	
$490 \vdash 590$	4	4	1º até 4º
$590 \vdash 690$	3	7	5º até 7º
$690 \vdash 790$	8	15	8º até 15º
$790 \vdash 890$	13	28	16 até 28º
$890 \vdash 990$	6	34	29º até 34º
$990 \vdash 1090$	6	40	35º até 40º
Total	40		

P_{25} é o peso que ocupa a $25(40)/100 = 10ª$ posição

$$P_{25} = 690 + \frac{(10-7)}{8}100 = 727{,}5 \text{ gramas} \Rightarrow 25\% \text{ dos coelhos têm peso inferior a } 727{,}5 \text{ gramas, ao desmame.}$$

P_{75} é o peso que ocupa a $75(40)/100 = 30ª$ posição

$$P_{75} = 890 + \frac{(30-28)}{6}100 = 923{,}3 \text{ gramas} \Rightarrow 75\% \text{ dos coelhos têm peso inferior a } 923{,}3 \text{ gramas, ao desmame.}$$

P_{90} é o peso que ocupa a $90(40)/100 = 36^{a}$ posição

$$P_{90} = 990 + \frac{(36-34)}{6}\,100 = 1023{,}3 \text{ gramas} \Rightarrow 90\% \text{ dos coelhos}$$

têm peso inferior a 1023,3 gramas, ao desmame.

Qual é a porcentagem de coelhos com peso <u>inferior</u> a 1000 gramas?

$$j = \left[\frac{(1000-990)6}{100} + 34\right]\frac{100}{40} = 86{,}5\%$$

⇒ 86,5% é a porcentagem de coelhos com peso <u>inferior</u> a 1kg.

⇒ A porcentagem de coelhos com peso <u>superior</u> a 1kg é igual a

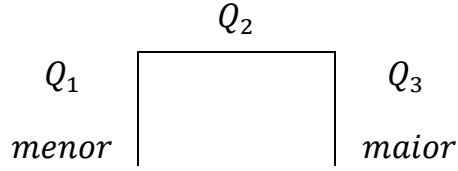
$$100 - 86{,}5 = 13{,}5\%$$

DESENHO ESQUEMÁTICO OU *Box-plot*

É um dispositivo gráfico utilizado para mostrar a (as)simetria de um conjunto de dados brutos, a dispersão dos seus valores e a presença ou não de pontos atípicos ou discrepantes (*outliers*).

Para construir o *box-plot* nós precisamos do menor valor, do maior valor da série e dos três quartis (Q_1, Q_2 e Q_3).

Com essas informações montamos um diagrama de 5 números:



A seguir calculamos:

- Intervalo interquartílico: $df = Q_3 - Q_1$
- Limite crítico inferior: L.C.I. $= Q_1 - 1{,}5df$
- Limite crítico superior: L.C.S. $= Q_3 + 1{,}5df$

Para construir o $box\text{-}plot$ nós devemos:

1. Desenhar uma caixa (box) com as faces laterais localizadas em Q_1 e Q_3 sobre um eixo com escala compatível com a variável estudada. Indicar a mediana (Q_2) por uma linha vertical no interior da caixa.
2. Traçar uma linha horizontal à esquerda da caixa ligando a face relativa ao Q_1 com o menor valor da série ou com o L.C.I. (o que aparecer primeiro!).

3. Traçar uma linha horizontal à direita da caixa ligando a face relativa ao Q_3 com o maior valor da série ou com o L.C.S. (o que aparecer primeiro!).

4. Verificar se na série existe algum valor inferior ao L.C.I. ou superior ao L.C.S.

Cada valor na série que seja inferior ao L.C.I. ou superior ao L.C.S. será considerado *outlier* e a sua posição deverá ser indicada no gráfico pelo símbolo $\bullet$ ou $*$, como na Figura 1.

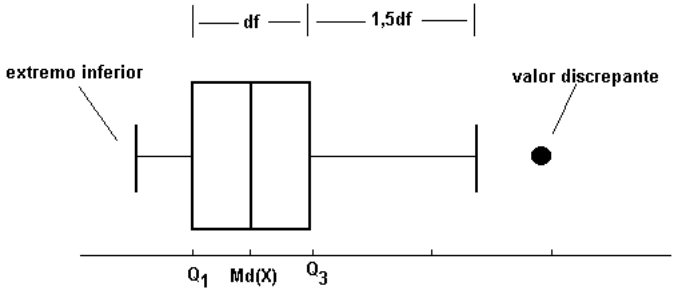


Figura 1. Box-plot

A localização da mediana, dos quartis inferior (Q_1) e superior (Q_3) e o comprimento das linhas horizontais traçadas à esquerda e à direita da caixa (box) podem sugerir uma assimetria da distribuição e indicar uma grande ou pequena dispersão dos dados.

EXEMPLO: Para os dados brutos de peso ao desmame de coelhos:

492	552	560	583	657	657	666	697	699	716
727	731	737	750	770	798	808	817	823	823
830	842	842	860	873	878	880	883	900	910
940	960	960	963	992	1000	1000	1000	1020	1040

Já sabemos que: $Q_1 = 718{,}8$, $Q_2 = Md(X) = 826{,}5$ e $Q_3 = 932{,}5$

Então:

- $df = Q_3 - Q_1 = 932{,}5 - 718{,}8 = 213{,}7$
- L.C.I. $= Q_1 - 1{,}5df = 398{,}2$ e L.C.S. $= Q_3 + 1{,}5df = 1253{,}0$
- Da série de dados brutos: $Min = 492{,}0$ e $Max = 1040{,}0$

Como não existe nenhum peso inferior ao L.C.I. (398,2), nem superior ao L.C.S. (1253,0) concluímos que não existem *outliers* na série e que a linha horizontal deve parar no menor e no maior dos pesos.

Percebe-se no *box-plot* dos pesos uma leve assimetria à esquerda, porque a cauda à esquerda (pesos menores) é mais longa que a da direita (pesos maiores), confirmando o comentário feito sobre a forma do histograma dos pesos.

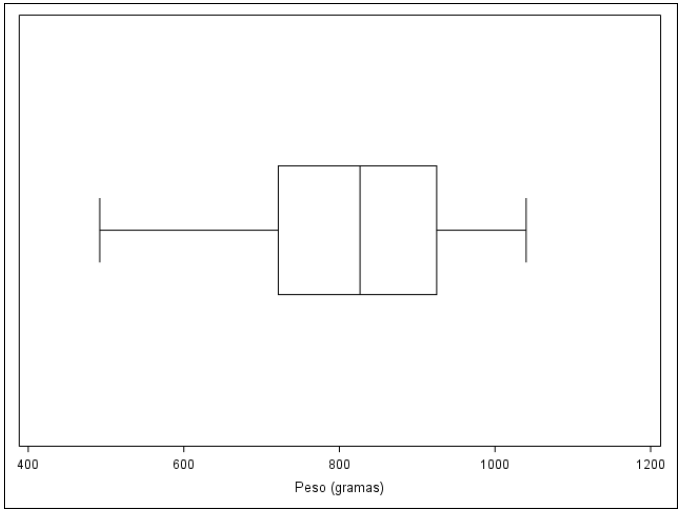


Figura 2. *Box-plot* dos pesos médios das ninhadas de coelhos desmamados no primeiro trimestre de 1989.