



Quick  
index  
[main](#)  
[eev](#)  
[maths](#)  
[blogme](#)  
[dednat4](#)  
[littlelangs](#)  
[PURO](#)  
([MD](#), [GA](#),  
[BE](#), etc)  
([Chapa 1](#))  
[emacs](#)  
[lua](#)  
([la](#))[tex](#)  
[fvwm](#)  
[tcl](#)  
[forth](#)  
[icon](#)  
[debian](#)  
[irc](#)  
[contact](#)  
[☞](#)

## Bioestatística - 2012.1

Horários do curso: 5<sup>as</sup>, 12-14, Sala 2.

Plano de aulas / resumo do que já aconteceu:

[1ª aula](#) (29/mar): [ainda não transcrevi as notas]

[2ª aula](#) (05/abr): [ainda não transcrevi as notas]

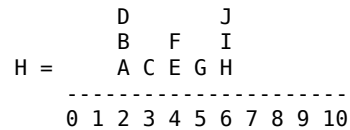
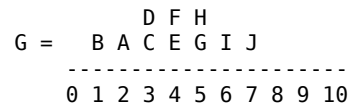
?ª aula (12/abr):

Não teve (tive uma audiência de um processo trabalhista no Rio)

[3ª aula](#) (19/abr):

Exercícios do fim da aula passada:

calculam a média e a variância das distribuições G e H, que têm estes histogramas:



Fórmulas:

$$\bar{A} = \frac{\sum_{i=1}^N A_i}{N} = \left( \frac{\sum_{i=1}^N A_i}{N} \right) / N$$

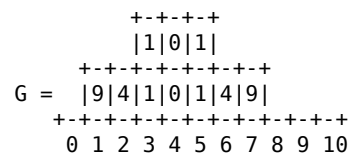
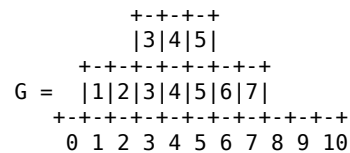
$$\text{Var}(A) = \frac{\sum_{i=1}^N (A_i - \bar{A})^2}{N - 1} = \left( \frac{\sum_{i=1}^N A_i}{N} \right) / N$$

Obs: não tentem calcular  $(a+b)^2$  fazendo  $a^2+b^2$ ,  
nem  $(a-b)^2$  fazendo  $a^2-b^2$ !...

Em geral  $(a+b)^2 \neq a^2+b^2$   
e  $(a-b)^2 \neq a^2-b^2$  ...

Truque: como cada termo de um somatório como  $\sum A_i$  ou  $\sum (A_i - \bar{A})^2$  corresponde a um quadradinho podemos pôr o resultado do termo dentro do quadradinho...

Calculamos a média e a variância das distribuições G e H via diagramas:



D            J

```

      B   F   I
H =   A C E G H
-----
      0 1 2 3 4 5 6 7 8 9 10

```

Exercícios: calcule a média e a variância de:

```

5 5 5 6 7 7 7,
4 4 4 6 8 8 8,
3 4 5 6 7 8 9.

```

Probabilidade (trailer da próxima aula):

Até agora vimos eventos "reais" - fingi que alunos tinham feito uma prova, tirado certas notas, etc...

Agora vamos ver eventos "ideais" (imaginários).

Vamos começar pensando em termos de jogadas de dados.

Vamos fazer uma tabela com todos os resultados possíveis - e cada linha da tabela vai ter a mesma "probabilidade de acontecer".

Exemplo:

Dado de verdade:			Dado ideal:	
i	A <sub>i</sub>	(resultado)	i	B <sub>i</sub>
-----				
1	5	<-- o 5 apareceu	1	1
2	4	/ duas vezes!	2	2
3	1		3	3
4	5	<-/	4	4
5	6		5	5
6	2		6	6

Avisei que vamos começar trabalhando com uma tabela de 36 linhas, com os resultados possíveis quando jogamos dois dados - e mostrei uma representação gráfica (bidimensional) dos quadradinhos dela.

Aula que vem: probabilidade, eventos, probabilidade condicional!

[4ª aula](#) (26/abr): Hoje: probabilidade condicional.

Estamos usando este livro:

Arminda Siqueira, Jacqueline Tibúrcio,  
"Estatística na Área da Saúde"

Probabilidade condicional aparece na p.145.

Probabilidade

=====

A notação:  $Pr(A|B)$

quer dizer: "a probabilidade de A acontecer dado que B já ocorreu"

Vamos voltar ao exemplo dos dois dados, do fim da aula passada.

Vamos usar "X" pro resultado do primeiro dado,

"Y" pro resultado do segundo,

"S" pra soma dos dois ( $S_i = X_i + Y_i$ ).

Tabela:

i	X <sub>i</sub>	Y <sub>i</sub>	S <sub>i</sub>
-----			
1	1	1	2
2	2	1	3
3	3	1	4
4	4	1	5
5	5	1	6
6	6	1	7
7	1	2	3
8	2	2	4
9	3	2	5
10	4	2	6
11	5	2	7
12	6	2	8
13	1	3	4
(...)			

36 6 6 12

Um diagrama pro valor do S:

Y=6	7	8	9	10	11	12
Y=5	6	7	8	9	10	11
Y=4	5	6	7	8	9	10
Y=3	4	5	6	7	8	9
Y=2	3	4	5	6	7	8
Y=1	2	3	4	5	6	7
	X=1	X=2	X=3	X=4	X=5	X=6

Eventos

=====

Exemplo:  $S \geq 8$  (ou:  $S_i \geq 8$ )

Qual é a probabilidade de  $S \geq 8$  acontecer?

Podemos marcar as linhas da tabela nas quais  $S \geq 8$  é verdade e dividir pelo total de linhas.

Melhor: vamos marcar os quadrados nos quais  $S \geq 8$ , já que cada quadrado corresponde a uma linha...

Truque: 1 vai querer dizer "verdadeiro",  
0 vai querer dizer "falso".

Um diagrama pro "valor" do  $S \geq 8$

(lembre que  $(S \geq 8) = 1$  quando  $S \geq 8$  é verdadeiro,  
 $(S \geq 8) = 0$  quando  $S \geq 8$  é falso):

Y=6	0	1	1	1	1	1
Y=5	0	0	1	1	1	1
Y=4	0	0	0	1	1	1
Y=3	0	0	0	0	1	1
Y=2	0	0	0	0	0	1
Y=1	0	0	0	0	0	0
	X=1	X=2	X=3	X=4	X=5	X=6

Daí:  $S_i \geq 8$  é verdadeiro em 15 quadradinhos (15 linhas da tabela)  
e portanto:

$$\begin{aligned} P(S \geq 8) &= 15/36 \\ &= 0.416666... \\ &= 41.6666... / 100 \\ &= 41.6666 \% \end{aligned}$$

Outro evento:

$X_i \leq 3$

Exercício: represente-o num diagrama similar ao anterior usando "0"s e "1"s, e calcule  $P(X \leq 3)$ .

Y=6	1	1	1	0	0	0
Y=5	1	1	1	0	0	0
Y=4	1	1	1	0	0	0
Y=3	1	1	1	0	0	0

Y=2	1	1	1	0	0	0
Y=1	1	1	1	0	0	0
X=1	X=2	X=3	X=4	X=5	X=6	

Probabilidade condicional

Fórmula do livro (p.145):

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad (\text{A e B não são variáveis - são "eventos"!!})$$

Vamos tentar interpretar:

$$\Pr(X \leq 3 \mid S \geq 8) = \frac{\Pr(X \leq 3 \cap S \geq 8)}{\Pr(S \geq 8)}$$

Porque o livro usa o sinal de interseção?

Ele interpreta eventos como conjuntos.

Lembrando o que são interseção e união:

$$\begin{aligned} \{1,2,3\} \cap \{2,3,4\} &= \{2,3\} \\ \{1,2,3\} \cup \{2,3,4\} &= \{1,2,3,4\} \end{aligned}$$

Um evento é um subconjunto do nosso espaço amostral...

O modo mais fácil de entender isto é pensar que o nosso espaço amostral é um conjunto de linhas de uma tabela, e eventos são subconjuntos disto - isto é, só algumas linhas.

Truque: quando numeramos as linhas da tabela (e portanto os alunos, os resultados dos dois dados...) podemos pensar no espaço amostral como um conjunto de números.

Vamos representar dentro de cada quadradinho o "i" correspondente.

Y=6	31	32	33	34	35	36
Y=5	25	26	27	28	29	30
Y=4	19	20	21	22	23	24
Y=3	13	14	15	16	17	18
Y=2	7	8	9	10	11	12
Y=1	1	2	3	4	5	6
X=1	X=2	X=3	X=4	X=5	X=6	

Espaço amostral todo:

$$E = \{1,2,3,\dots,36\}$$

Eventos (como conjuntos):

$$\begin{aligned} (S \geq 8) &= \{12,17,18,22,23,24,27,28,29,30,32,33,34,35,36\} \\ (X \leq 3) &= \{1,2,3,7,8,9,13,14,15,19,20,21,25,26,27,31,32,33\} \\ (X \leq 3) \cap (S \geq 8) &= \{27,32,33\} \end{aligned}$$

Daí:

$$\Pr(X \leq 3 \cap S \geq 8) = \frac{3}{36} = 0.083333\dots$$

e:

$$\Pr(X \leq 3 \mid S \geq 8) = \frac{\Pr(X \leq 3 \cap S \geq 8)}{\Pr(S \geq 8)} = \frac{0.083333\dots}{0.416666\dots} = 0.2 = 20\%$$

### 5ª aula (02/mai):

Na aula passada vimos probabilidade condicional...

$\Pr(A|B)$  é a probabilidade do evento A acontecer dado que o evento B já aconteceu. Vimos que pra entender isto precisávamos fazer outras distribuições (tabelas), restringindo as linhas da tabela original, ou, equivalentemente, o espaço amostral... Agora

vamos usar uma idéia parecida (com uma notação diferente da do livro).

Lembre que estamos trabalhando com uma tabela de 36 linhas (a dos resultados possível em dois dados), e variáveis X, Y e S=X+Y.

A partir de agora toda vez que escrevermos uma variável - como X e Y - vamos prestar atenção ao espaço amostral no qual ela está definida.

Podemos construir novas variáveis restringindo o espaço amostral - e para cada variável nova destas podemos calcular a sua média, variância, etc, e comparar estes valores com os das variáveis originais.

Exemplos: Y|X=2 "Y quando X=2",  
 Y|X>=3 "Y quando X>=3",  
 X|S>=8 "X quando X>=8"

Aí vimos como representar estas variáveis em diagramas parecidos com os anteriores; às vezes usávamos "." pra indicar onde elas não estavam definidas, às vezes não usávamos nada.

X=	1 2 3 4 5 6	Y=	6 6 6 6 6 6	S=	7 8 9 10 11 12
	1 2 3 4 5 6		5 5 5 5 5 5		6 7 8 9 10 11
	1 2 3 4 5 6		4 4 4 4 4 4		5 6 7 8 9 10
	1 2 3 4 5 6		3 3 3 3 3 3		4 5 6 7 8 9
	1 2 3 4 5 6		2 2 2 2 2 2		3 4 5 6 7 8
	1 2 3 4 5 6		1 1 1 1 1 1		2 3 4 5 6 7

(X=2)=	0 1 0 0 0 0	(Y X=2)=	. 6 . . . .
	0 1 0 0 0 0		. 5 . . . .
	0 1 0 0 0 0		. 4 . . . .
	0 1 0 0 0 0		. 3 . . . .
	0 1 0 0 0 0		. 2 . . . .
	0 1 0 0 0 0		. 1 . . . .

(X>=3)=	0 1 0 0 0 0	(Y X>=3)=	6 6 6 6
	0 1 0 0 0 0		5 5 5 5
	0 1 0 0 0 0		4 4 4 4
	0 1 0 0 0 0		3 3 3 3
	0 1 0 0 0 0		2 2 2 2
	0 1 0 0 0 0		1 1 1 1

(S>=8)=	0 1 1 1 1 1	(X S>=3)=	2 3 4 5 6
	0 0 1 1 1 1		3 4 5 6
	0 0 0 1 1 1		4 5 6
	0 0 0 0 1 1		5 6
	0 0 0 0 0 1		6
	0 0 0 0 0 0		

Pra calcular a média do (Y|X=2), fazemos:

$$\overline{(Y|X=2)} = \frac{\sum_{j=1}^N Y_j}{N} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$



ordenados por nota, dividindo esta lista em quatro partes iguais (cada uma com 5.25 alunos!),

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| A| B| C| D| E| F| G| H| I| J| K| L| M| N| O| P| Q| R| S| T| U|
| 1| 2| 2| 3| 4| 4| 5| 5| 5| 5| 6| 7| 7| 8| 8| 8| 8| 9| 9| 9|10|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
                        ^           ^           ^
                        Q1          Q2          Q3

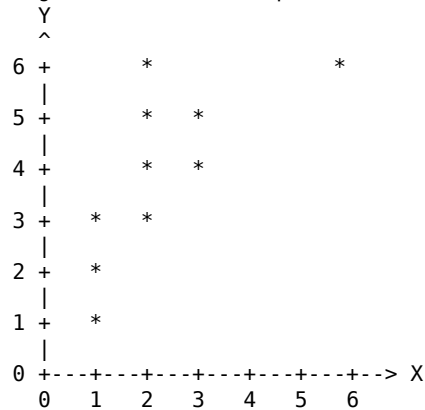
```

e aí vendo quais são as notas dos alunos nestas divisões - neste caso Q1=4, Q2=6 (obs: Q2 é a mediana), e Q3=8... Quando as divisões caem entre dois alunos (pense no que aconteceria se tivéssemos 10 alunos, ou 8), pegamos a média dos alunos dos dois lados da divisão.

Exercício: desenhe o boxplot (p.97) para a distribuição acima (desenhe-o sob o histograma).

9ª aula (?) (04/out):

Num gráfico com este aqui temos 10 "observações" (10 pessoas),



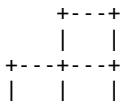
e podemos pensar que este gráfico representa uma tabela...

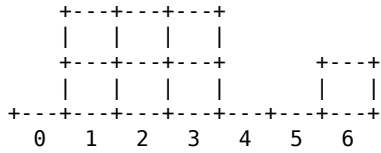
Só que uma tabela tem informações a mais! Em sala nós anotamos do lado de cada um desses pontos um nome de uma pessoa - pra ficar mais fácil entender cada ponto como uma pessoa - e um índice, i, que dizia em que linha essa pessoa aparecia na tabela... e fizemos uma tabela com todas estas informações; cada ponto tinha quatro informações - nome, i, X e Y - e no gráfico só aparecem o X e Y. Lembre que um muitas fórmulas que já vimos, por exemplo, a da média,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

o índice - no caso, "i" - aparece explicitamente.

A distribuição acima - vamos chamá-la de distribuição A - é uma distribuição em várias variáveis. Quando escrevemos X(A) estamos falando de uma distribuição em uma variável, com o mesmo número de observações que a distribuição A (isto é, 10 pessoas); a gente obtém a distribuição X(A) a partir da distribuição A ficando só com a informação do X de cada pessoa, e descartando o resto. Dá pra fazer um histograma pro X(A), e obtemos isto:





Pra calcularmos a mediana e os quartis de  $X(A)$ , isto é,  $Q2(X(A))$ ,  $Q1(X(A))$  e  $Q3(X(A))$ , é melhor fazermos uma outra tabela, na qual as pessoas aparecem ordenadas pelo seu "X" (do mesmo modo que ordenamos os alunos por nota na aula anterior) - e pra isto é melhor inventarmos um outro índice... podemos chamá-lo de "j".

...Aí a gente acrescentou mais uma informação pra cada um dos pontos do gráfico do início da aula: um índice "j" pra cada pessoa - e numa tabela indexada pelo j temos:

j	X(A)_j
1	1
2	1
3	1
4	2
5	2
6	2
7	2
8	3
9	3
10	6

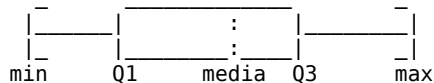
Então: agora a gente já sabe inventar índices novos, e também dá pra inventar um outro índice, k, pros "Y"s ficarem em ordem e a gente poder calcular os quartis da distribuição em uma variável  $Y(A)$  -  $Q1(Y(A))$ ,  $Q2(Y(A))$ ,  $Q3(Y(A))$ .

Agora a idéia mais importante. Dá pra gente criar novas distribuições a partir da distribuição A fazendo \_restrições\_. A notação é com uma barra vertical, que vamos pronunciar como "tal que". A distribuição  $A|X \leq 2$  só tem 7 observações - só os pontos cujo valor de "X" é menor ou igual a 2 - e a distribuição  $A|X \geq 3$  só tem 3 observações, os pontos cujo valor de "X" é maior ou igual a 3. Dá pra calcular a média e os quartis destas distribuições:

$$\overline{Y(A|X \leq 2)}, \quad Q1(Y(A|X \leq 2)), \quad Q2(Y(A|X \leq 2)), \quad Q3(Y(A|X \leq 2)),$$

$$\overline{Y(A|X \geq 3)}, \quad Q1(Y(A|X \geq 3)), \quad Q2(Y(A|X \geq 3)), \quad Q3(Y(A|X \geq 3)),$$

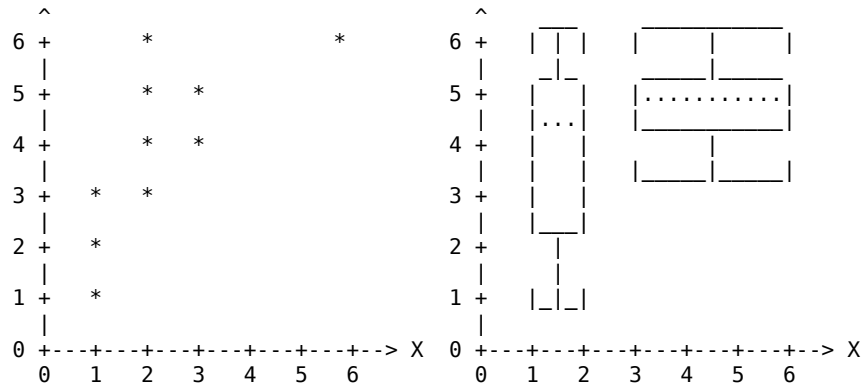
Existem várias convenções para boxplots, e vamos usar - temporariamente - uma diferente da do livro. Na nossa vamos representar o valor mínimo, o Q1, a média (não a mediana!), o Q3 e o valor máximo, e o desenho vai ser assim, com "[---", no mínimo, depois um retângulo com parede esquerda no Q1, uma linha pontilhada na média e parede direita no Q2, depois um "---]" até o valor máximo:



Podemos desenhar estes boxplots sobre o gráfico da distribuição A, e aí vamos ter algo como a figura da direita abaixo (obs: as alturas estão erradas, isso é só pra dar uma noção de como podemos ter vários boxplots verticais num gráfico só, um pra cada região do X):







Exemplos da Wikipedia:

<http://upload.wikimedia.org/wikipedia/commons/f/fa/Sstcurve.jpg>

<http://upload.wikimedia.org/wikipedia/commons/7/71/Sstpoint.jpg>

[10ª aula](#) (?) (11/out): P1.

Uma questão MUITO importante da prova vai ser sobre traçar boxplots para distribuições em duas variáveis. Outra vai ser sobre notação - você vai ter que interpretar uma fórmula envolvendo um " $\Sigma$ ". Não posso dar mais dicas além destas. =)

[11ª aula](#) (?) (18/out):

Revimos probabilidade condicional - a matéria da P2 vai ser a da P1 e mais probabilidade condicional. Fotos do quadro:

[http://angg.twu.net/BE/2012-10-18\\_BE1.jpg](http://angg.twu.net/BE/2012-10-18_BE1.jpg)

[http://angg.twu.net/BE/2012-10-18\\_BE2.jpg](http://angg.twu.net/BE/2012-10-18_BE2.jpg)

[http://angg.twu.net/BE/2012-10-18\\_BE3.jpg](http://angg.twu.net/BE/2012-10-18_BE3.jpg)

Depois eu limpo as anotações e passo pra cá numa forma em que dê pra imprimí-las!

[12ª aula](#) (?) (25/out): P2.

[13ª aula](#) (?) (01/nov): VR e VS.