Estatística Aplicada à Medicina - Listas de exercícios

Profs.: Ronado Rocha Bastos e Henrique S. Hippert

Lista 1 : Probabilidades

1) Numa pesquisa feita entre assinantes de revistas, verificou-se que 205 pessoas assinam a revista "A", 283 assinam a "B". Se um total de 401 pessoas foram entrevistadas, qual é o número de pessoas que assina ambas as revistas?

2) Um grupo de pessoas é constituído de 5 homens brasileiros e 4 estrangeiros, 6 mulheres brasileiras e 3 estrangeiras. Uma destas pessoas será sorteada. Definimos estes eventos: (a notação A' representa o *complementar* de A)

B= {a pessoa é brasileira} *E= {a pessoa é estrangeira}*

H= (a pessoa é homem} *M= {a pessoa é mulher}*

a) Explique os que representam os eventos representados pelas notações abaixo. Verifique se algum destes conjuntos é vazio.

(i) $B \cup H$ (ii) B' (iii) $B \cap H$ (iv) $H \cap M$ (v) M' (vi) $M' \cup E'$ (vii) $E-M$

b) Verifique se existe alguma maneira diferente de notar estes conjuntos.

c) Calcule a probabilidade de cada um destes eventos.

d) Calcule P(B/H), P(M/E').

3) Suponha que a probabilidade de uma pessoa ser do tipo sangüíneo "O" seja de 0.4, do tipo "A" seja 0.3, e do tipo "B"seja de 0.2. Suponha ainda que a probabilidade do fator Rh ser positivo seja de 0.9, independentemente do tipo sangüíneo. Qual será então a probabilidade de uma pessoa tomada ao acaso ser do tipo: a) O+ b) AB-

4) Dentre 6 números positivos e 8 negativos, 2 são sorteados, sem reposição, e multiplicados entre si. Qual a probabilidade de que o resultado da multiplicação seja positivo ?

5) Suponha que num hemograma sejam feitos 12 testes. Em cada um deles, é medida uma variável do sangue (número de hemácias, nível de colesterol, etc.). Para cada uma destas variáveis, existem valores v_1 e v_2 de referência; a probabilidade de a variável assumir valores abaixo de v_1 é de 0,05, a probabilidade de a variável ficar entre v_1 e v_2 é de 0,90 (se isto acontece, o resultado do teste é considerado "normal"), e a probabilidade de ficar acima de v_2 é de 0,05. Qual será a porcentagem da população que obteria valores "normais" em todos os 12 testes do exame?

6) Um dado é lançado 6 vezes. Quais são as probabilidades de que:

a) somente apareçam números pares?

b) somente apareçam números menores que 5?

c) não apareça nenhuma vez o numero "6" ?

d) o numero "6" apareça em todos os 3 primeiros lançamentos, e somente neles ?

e) apareça a seqüência "1,2,3,4,5,6" ?

7) Um animal foi treinado para executar uma tarefa. A probabilidade de que ele falhe na sua primeira tentativa é de 0,40. Se o animal falhar, ele faz nova tentativa; o animal aprende com cada erro cometido, de modo que a probabilidade de um erro em cada tentativa é apenas metade da probabilidade de erro na tentativa fracassada anterior. Se é permitido ao animal fazer apenas 3 tentativas, qual é a probabilidade de que ele consiga realizar a tarefa?

8) Três prêmios são sorteados, com reposição, entre as 10 pessoas de um grupo.
a) Qual é a probabilidade de que a pessoa mais velha do grupo ganhe os 3 prêmios ?
b) Qual é a probabilidade de que uma mesma pessoa ganhe os 3 prêmios ?

9) Uma pessoa tem 10 pares de meias, que guarda misturadas numa caixa. De manhã, ao acordar, pega duas meias aleatoriamente e as calça. Qual é a probabilidade de que estas duas meias formem um par?

10) Depois de uma festa do fundo de formatura, sobraram 4 caixas de cerveja. Foi decidido sorteá-las, uma de cada vez, entre os alunos da turma (16 homens e 4 mulheres). O sorteio foi sem reposição. Calcule as probabilidades de:
a) nenhuma das mulheres ser premiada; b) todas as mulheres serem premiadas.

11) São retiradas 4 cartas de um baralho completo. Calcule as probabilidades de que sejam todas do naipe de ouros, se a retirada for
a) sem reposição
b) com reposição.

12) Uma pessoa lanca um dado octoédrico 3 vezes consecutivas. Qual é a probabilidade de que este dado mostre a mesma face nos 3 lancamentos?

13) Dois atiradores disparam um tiro cada um, contra o mesmo alvo. Se cada um deles tem 0,5 de probabilidade de acertar, qual é a probabilidade de o alvo ser atingido ? E se forem três atiradores?

14) Para bebês americanos descendentes de mexicanos nascidos no Arizona em 1986 e 1987, a probabilidade de a idade gestacional ser menor que 37 semanas é 0,142; a probabilidade de o peso ao nascer ser menor que 2500 gramas é 0,051; a probabilidade de que esses dois eventos ocorram simultaneamente é 0,031.
a) Seja A o evento "a idade gestacional do bebê é menor do que 37 semanas" e B o evento "o peso ao nascer do bebê é menor do que 2500 gramas". Construa um diagrama de Venn para ilustrar a relação entre os eventos A e B.
b) Se um bebê é selecionado aleatoriamente nesta população, qual é a probabilidade de que ele tenha nascido com menos de 37 semanas de idade gestacional e menos de 2500g de peso?
c) Se um bebê é selecionado aleatoriamente nesta população, qual é a probabilidade de que ele tenha nascido **ou** com menos de 37 semanas de idade gestacional, **ou** com menos de 2500g de peso, **ou** com ambas as características?
d) Se um bebê nasce, nesta população, com idade gestacional de 37 semanas ou mais, qual é a probabilidade de que ele tenha baixo peso?
e) Se um bebê nasce, nesta população, com idade gestacional de menos de 37 semanas, qual é a probabilidade de que ele tenha baixo peso?
f) Comparando suas respostas aos itens (d) e (e), você pode dizer que os eventos A e B são *dependentes* ou *independentes*, probabilisticamente?

(Pagano e Gauvreau, p.141, modificado)

Lista 1 - Respostas

1) $n(A \cap B) = 87$
2) (a) ------
   (b) ------
   (c.i) P = 5/6
   (c.ii) P = 7/18
   (c.iii) P = 5/18
   (c.iv) P = 0
   (c.v) P = 1/2
   (c.vi) P = 5/6
   (c.vii) P =2/9
   (d) P = 5/9 , 6/11
3) (a) P(O+) = 0,36
   (b) P(AB-) = 0,01
4) P(resultado positivo)= 0,472
5) $P= (0,9)^{12} = 0,282$ (desde que os testes sejam independentes)
6) a) P(somente pares) = 0,0156
   b) P(só números menores que 5) = 0,0877
   c) P(nenhum 6) = 0,335
   d) P(666xxx) = 0,00267
   e) P(123456) = 0,0000214
7) P(conseguir) = 0.9920
8) a) P(pessoa mais velha) = 0,001
   b) P(mesma pessoa) = 0,010
9) P(formar par) = 1/19
10) (a) P(nenhuma mulher) = 0,376
    (b) P(todas as mulheres) = 0,000206
11) (a) P(sem reposição) = 0,00264
    (b) P(com reposição) = 0,00390
12) P(mesma face) = 0,0156
13) (a) com dois atiradores, P(atingir) = 0,750
    (b) com três atiradores, P(atingir) = 0,875
14) (a) --
    (b) 0,031
    (c) 0,162
    (d) 0,0233
    (e) 0,218
    (f) Dependentes.

Lista 2 : Razão de Chances e Risco Relativo

1) página 246, exemplo 7.5 (Soares e Siqueira)
Este exemplo mostra o cálculo de RR para dados fornecidos. Monte uma tabela 2x2 com os dados fornecidos (isto é, uma tabela *uso de aspirina* x *infartos*). Calcule a razão de chances, e compare com o risco relativo calculado.

2) página 259, exercício 7 (Soares e Siqueira)
Calcule a *odds-ratio*, como pedido; não se preocupe em calcular o intervalo de confiança (este tópico ainda não foi visto). Que conclusão você pode tirar deste resultado?

3) Exercício no. 18 (Pagano & Gauvreau, p.144)
Por diversos métodos de contracepção, as probabilidades de uma mulher casada tenha uma gravidez não planejada durante o primeiro ano de uso estão abaixo:

Método	Probabilidade de gravidez
nenhum	0,431
diafragma	0,149
preservativo	0,106
DIU	0,071
pílula	0,037

Para cada método, calcule o risco relativo de gravidez para mulheres que usam o método, em relação às que não usam nenhum método. Como o risco se modifica, em relação ao método de contracepção usado?

4) Exercício no. 19 (Pagano & Gauvreau, p.144)
Um estudo sobre doenças respiratórias durante o primeiro ano de vida foi conduzido em uma comunidade na Carolina do Norte. Como parte deste estudo, um grupo de crianças foi classificado de acordo com o status socioeconômico da família. Os números de crianças em cada grupo que experimentaram sintomas respiratórios persistentes estão mostrados abaixo:

status	número de crianças	crianças com sintomas
baixo	79	31
médio	122	29
alto	192	27

a) Use esses dados para calcular a probabilidade de cada grupo socio-econômico sofrer de sintomas respiratórios persistentes. Suponha que os números são suficientemente grandes para satisfazer a definição frequentista de probabilidade.
b) Calcule os riscos relativos de os grupos socioeconômicos baixo e médio experimentarem sintomas respiratórios persistentes em relação ao grupo socioeconômico alto.
c) Calcule as razões de chances de os grupos socioeconômicos baixo e médio experimentarem sintomas respiratórios persistentes em relação ao grupo socioeconômico alto.
d) Entre o status socioeconômico e os sintomas respiratórios, parece haver alguma associação?

Lista 2 - Respostas

1) Soares e Siqueira p.246, ex. 7.5

tabela:

uso da aspirina	infarto (D)		total
	D+	D-	
F+	139	10898	11037
F-	239	10795	11034
total	378	21693	22071

Cálculo da razão de chances:

OR = (139/10898)/(239/10795) = 0,5761

Portanto, a chance de uma pessoa que toma aspirina vir a ter um infarto é 0,58 da chance de uma pessoa que não toma aspirina.
Considerando o inverso da OR:

1/OR = 1/0,5761 = 1,74

Portanto, uma pessoa que não toma aspirina tem uma chance de infarto que é 74% maior que a da pessoa que toma. O RR calculado, no exemplo, é de 1,69, bastante semelhante ao OR. Isto ocorre quando a prevalência da doença (infarto) é baixa.

2) Soares e Siqueira, p.259, ex. 7

Razão de chances:

OR = (24/15)/(32/49) = 2,45

A chance de um recém-nascido com baixo peso (<1500g) sofrer de HPIV é 2,45 vezes a chance de um recém-nascido de peso entre 1500-2000g.

3) Pagano e Gauvreau, p.144, Exercício no. 18

Resposta:

RR(diafragma) = 0,149/0,431= 0,346

ou, calculando pelo recíproco:

1/RR = 1/0,346 = 2,893

Portanto, a mulher que usa diafragma tem um risco de gravidez que é aproximadamente 35% do risco que corre a mulher que não usa nenhum método anticoncepcional. Ou, usando o recíproco, a mulher que não usa nenhum método tem um risco de gravidez que é 2,9 vezes o da mulher que usa diafragma.

RR(preservativo) = 0,106/0,431 = 0,246
RR(DIU) = 0,071/0,431 = 0,165
RR(pilula) = 0,037/0,431 = 0,086

4) Pagano e Gauvreau, p.144, Exercício no. 19

a) Probabilidade de uma criança do grupo "baixo *status*" ter os sintomas:
P = 31/79 = 0,392
Probabilidade de uma criança do grupo "médio *status*" ter os sintomas:
P = 29/122 = 0,238
Probabilidade de uma criança do grupo "alto *status*" ter os sintomas:
P = 27/192 = 0,141

b) Riscos relativos : os riscos relativos (RR) podem ser calculados desta forma:

	sintomas		
status	D+	D-	total
Baixo	31	48	79
Alto	27	165	192

a) RR(*status* baixo / alto) = (31/79)/(27/192) = 2,79

	sintomas		
status	D+	D-	total
Médio	29	93	122
Alto	27	165	192

RR(*status* médio / alto) = (29/122)/(27/192) = 1,69

Portanto (se for verificado que estes resultados são *estatisticamente significativos*), podemos dizer que as crianças de *status* baixo tem um risco de terem sintomas 2,79 vezes o risco das crianças de *status* alto; as crianças de *status* médio tem um risco 1,69 vezes o risco das de *status* alto. Portanto, quanto mais baixo o *status* social, maior o risco de sintomas.

c) Razão de chances (*odds ratio*): as razões de chances (OR) podem ser calculadas a partir das mesmas tabelas do item anterior:

OR(*status* baixo/alto) = (31/48) / (27/165) = 3,95
OR(*status* médio/alto) = (29/93) / (27/165) = 1,91

d) Sim; os *status* mais baixos estão associados a maiores riscos de sintomas. Tanto o RR quanto a OR podem ser usados para quantificar esta relação (embora, é claro, seja preciso calcular a significância estatística destes valores).

Pagano M, Gauvreau K. (2004). *Princípios de Bioestatística*. São Paulo: Pioneira Thomson.
Soares J.F., Siqueira A.L. (2002). *Introdução à Estatística Médica*. 2ª. ed. Belo Horizonte: COOPMED.

Lista 3 : Qualidade de testes diagnósticos

1) Pesquisadores que tratam de doenças hepáticas em uma clínica especializada sugeriram um novo teste para detectar câncer no fígado. Os resultados do experimento, numa amostra de 2.225 pacientes atendidos nesta clínica, foram os seguintes:

Câncer hepático	Teste		Total
	Positivo	Negativo	
Presente	90	17	107
Ausente	39	2.079	2.118
Total	129	2.096	2.225

c) Pelo modo como a amostra foi selecionada, a prevalência da doença pode ser calculada a partir da tabela. Calcule-a.
d) Calcule a probabilidade de que um paciente atendido nesta clínica, que não têm câncer de fígado, obtenha um resultado positivo no teste.
e) Calcule a VPP e VPN.

Os pesquisadores que desenvolveram o teste sugeriram seu uso como um meio simples para os clínicos em geral decidirem se devem ou não encaminhar o paciente para uma clínica especializada. Comente sobre esta recomendação. (Soares e Siqueira, p.123)

2) Um estudo registrou que a sensibilidade da mamografia como teste de triagem para detecção de câncer de mama é de 0,85, enquanto sua especificidade é de 0,80.
c) Qual é a probabilidade de um resultado de teste falso negativo?
d) Qual é a probabilidade de um resultado falso positivo?
e) Se na população a probabilidade de que uma mulher tenha câncer de mama é de 0,0025, qual a probabilidade de que uma mulher tenha este tipo de câncer, dado que sua mamografia for positiva? Como se chama esta probabilidade?
f) Nesta população, qual é a probabilidade de a paciente não ter câncer, dado que sua mamografia foi negativa? Como se chama esta probabilidade?
g) Comparando as respostas dos itens (c) e (d), que conclusões você pode tirar sobre a utilidade deste teste? (Pagano e Gauvreau, p.142, modificado)

3) Os dados seguintes são tomados de um estudo que investiga o uso de uma técnica chamada "ventriculografia radionuclídica" como teste de diagnóstico para detecção de doença da artéria coronária.
g) Qual foi a sensibilidade da ventriculografia radionuclídica estimada por esse estudo?
h) Qual foi a especificidade?
i) Numa população, a prevalência da doença da artéria coronária é de 0,10. Calcule a probabilidade de que um indivíduo tenha a doença, dado que ele apresentou resultado positivo na ventriculografia radionuclídica.
j) Neste estudo, qual será o valor preditivo de um teste negativo? (Pagano e Gauvreau, p.143)

Teste	Doença presente	Doença ausente	Total
Positivo	302	80	382
Negativo	179	372	551
Total	481	452	933

Referências:

Soares, JF; Siqueira, AL. *Introdução à Estatística Médica*. Belo Horizonte: Coopmed, 2002.
Pagano, M; Gauvreau, K. *Princípios de Bioestatística*.São Paulo: Pioneira Thomson, 2004.

Lista 3 - Respostas

1.
a) p = 0,048
b) P(T+|D-) = 0,018
c) VPP = 0,698
   VPN = 0,992
d) Como o VPP e VPN foram calculados na clinica especializada, onde provavelmente a prevalência é bem maior do que na clinica geral, devemos nos basear na sensibilidade (0,8411) e especificidade (0,9815) do teste. Já que a especificidade, que é o principal determinante do VPP em condições de baixa prevalência, é especialmente boa, a recomendação parece ter fundamento. Porem, o paciente com teste positivo deve ser avisado de que ainda há uma probabilidade razoável de que ele não tenha a doença (aproximadamente 1 para 2 se a prevalência for igual a da clinica especializada, 1 para 4 se a prevalência for 10% daquela da clinica especializada, etc.), que precisa ser investigada pela clinica especializada.

2.
a) 0,15
b) 0,20
c) A probabilidade se chama “valor preditivo positivo”, e é de 0,0105
d) A probabilidade se chama “valor preditivo negativo”, e é de 0,9995
e) O teste tem utilidade limitada, pois, como se trata de uma situação de baixa prevalência, o principal objetivo seria obter um bom VPP, algo que não ocorre.

3.
a) 0,6278
b) 0,8230
c) 0,2827
d) 0,9521

Lista 4 : Estatística Descritiva

1) O diagrama de ramo-e-folhas abaixo mostra a população atendida em cada uma das 40 Unidades Básicas de Saúde de Juiz de Fora, em 2004. Neste gráfico, cada "folha" contém os algarismos das centenas, dezenas e unidades, enquanto o algarismo dos milhares fica no eixo vertical.

a) Qual foi o menor número de atendimentos registrado neste gráfico? Qual foi o maior?
b) Refaça este gráfico, de uma forma mais simples, *arredondando* os valores, de modo a colocar apenas um algarismo por "folha".
c) Refaça-o novamente, desta vez *truncando* os valores.
d) Existe muita diferença entre os gráficos dos itens (b) e (c) ? Quais as vantagens e desvantagens que existem no *arredondamento*, em relação ao *truncamento*?
e) Que relação deve existir entre a média dos dados originais, a média dos dados arredondados e a média dos dados truncados?
f) Calcule, a partir do gráfico abaixo, o número mediano de atendimentos nas UBS de Juiz de Fora.
g) Calcule os quartis desta distribuição e, a partir deles, verifique se o valor extremo 16.880 pode ser considerado discrepante, de acordo com a regra de Tukey.

```
 2*** | 981
 3*** | 284,993
 4*** | 022,158,449,968
 5*** | 336,450,494,606,610,765
 6*** | 040,100,102,167,259,385,405,625,677,813
 7*** | 051,200,338
 8*** | 015,368,431,726
 9*** | 708,964
10*** | 098,596,913
11*** | 148,658
12*** | 102,172
13*** |
14*** |
15*** |
16*** | 880
```

Arquivo: ubs.dta

2) O gráfico de ramo-e-folhas abaixo mostra os níveis de colesterol (mg/dl) medidos numa amostra de 100 pessoas do Havaí, em 1969. Este gráfico foi feito pelo programa *Stataquest*, e nele cada dezena é representada em uma linha.

a) Calcule, a partir do gráfico, a mediana do nível de colesterol destas pessoas.
b) A julgar pelo gráfico, qual deve ser aproximadamente o nível *médio* de colesterol das pessoas desta amostra?
c) Considerando que o nível máximo recomendado de colesterol no sangue é de 200 mg/dl, qual é a proporção de pessoas nesta amostra que tem nível acima do recomendado?
d) Refaça o gráfico, usando uma escala mais condensada, em que cada centena seja representada em cinco linhas. Para isto, trunque os dados, eliminando o algarismo das unidades. As três primeiras e a última linhas deste no gráfico estão mostradas abaixo, assim como sua legenda.

```
1* |
1t | 3
1f | 45
   .
   .
3- | 8
```

legenda: 2s|6 representa 260-269 mg/dl

```
13* | 4
14* | 7
15* | 7
16* | 12456
17* | 1366899
18* | 011345666799
19* | 0024568999
20* | 134569
21* | 011235667788999
22* | 1133458
23* | 001112448899
24* | 000368
25* | 15569
26* | 178
27* | 29
28* | 679
29* | 068
30* |
31* |
32* |
33* |
34* |
35* |
36* |
37* |
38* | 2
```

legenda: 26*|1 representa 261 mg/dl
(dados de Soares e Siqueira, 2002)

e) Refaça novamente o gráfico, de modo que cada centena ocupe apenas duas linhas. As duas primeiras deste no gráfico estão mostradas abaixo, assim como sua legenda.

```
1 | 34
- | 566666777777788888888888889999999999
```

legenda: 2|6 representa 260-269 mg/dl

f) Qual destes gráficos lhe parece o mais útil para representar estes dados? Qual lhe parece o menos útil? Por que?

3) Em uma pesquisa sobre a concentração de minerais no leite materno, foram coletados dados de 55 mães do Hospital Maternidade Odete Valadares em Belo Horizonte, no período de 1984 a 1985. As mães foram divididas em dois grupos, segundo o período de lactação: colostro e leite maduro. Os minerais considerados foram, cobre, magnésio e zinco. Os dados a seguir referem-se às concentrações de cálcio e zinco.

Cálcio (μg/ml de leite) – Grupo colostro

113	181	254	311	334	145	221	256	312	344
163	225	275	313	372	163	231	296	323	375
167	241	303	325	375	437				

Cálcio (μg/ml de leite) – Grupo maduro

159	175	181	188	200	206	213	214	217	231
238	238	242	244	256	259	260	263	264	275
277	279	281	293	302	303	314	344	394	

Zinco (μg/ml de leite) – Grupo colostro

1,07	1,20	1,30	3,13	3,20	3,70	4,40	4,57	5,20	5,82
6,07	6,13	6,50	6,82	6,82	6,90	7,23	7,42	7,43	7,43
7,45	8,25	8,40	8,77	9,54					

Zinco (μg/ml de leite) – Grupo maduro									
0,52	0,60	0,78	0,86	1,08	1,23	1,28	1,31	1,40	1,55
1,55	1,60	1,92	2,15	2,27	2,40	2,41	2,52	2,57	2,60
2,63	2,92	3,02	3,05	3,57	3,88	5,50	7,38		

a) Compare, por meio de gráficos de ramo-e-folhas, os grupos "maduro" e "colostro", quanto às concentrações de zinco e de cálcio.
b) Compare novamente, usando diagramas de Tukey. Lembre-se de destacar os pontos discrepantes existentes.

A partir destes gráficos, descreva a diferença existente entre leite maduro e colostro, quanto à concentração de zinco e cálcio.

(dados de Soares e Siqueira, 2002, p. 82)

4) Os dados abaixo mostram os teores de gordura fecal (em g/24h) medidos em 43 crianças sadias.
   a) Organize estes dados numa tabela de distribuição de freqüências, com intervalo de classe = 0.5, e classes fechadas à esquerda.
   b) Faça o histograma desta distribuição de freqüências.
   c) Represente os dados por um gráfico de ramo-e-folhas que tenha duas linhas por dezena.
   d) Observe que gráfico do item (c) terá exatamente a mesma forma que o histograma do item (b). Isto mostra que o ramo-e-folhas e o histograma estão relacionados; ambos são gráficos que representam os dados de forma agrupada. Quais as vantagens e desvantagens de cada um destes gráficos em relação ao outro?

3,7	1,6	2,5	3,0	3,9	1,9	3,8	1,5	1,1	1,8
1,4	2,7	2,1	3,3	3,2	2,3	2,3	2,4	0,8	3,1
1,8	1,0	2,0	2,0	2,9	3,2	1,9	1,6	2,9	2,0
1,0	2,7	3,0	1,3	1,5	4,6	2,4	2,1	1,3	2,7
2,1	2,8	1,9							

(dados de Soares e Siqueira, 2002, p.36)

5) A tabela a seguir mostra a distribuição do peso ao nascer dos 7.126 crianças nascidas vivas no ano de 1992, na região urbana do distrito sede de Juiz de Fora. Não foram incluídas na tabela 99 crianças com peso ao nascer ignorado. Represente esta distribuição em um polígono de freqüências. Como você a descreveria, em termos de moda e de simetria?

Peso ao nascer (g)			f
500	⊢	1000	33
1000	⊢	1500	60
1500	⊢	2000	147
2000	⊢	2500	614
2500	⊢	3000	2077
3000	⊢	3500	2875
3500	⊢	4000	1113
4000	⊢	4500	184
4500	⊢	5000	20
5000	⊢	5500	1
5500	⊢	6000	2
Total			7126

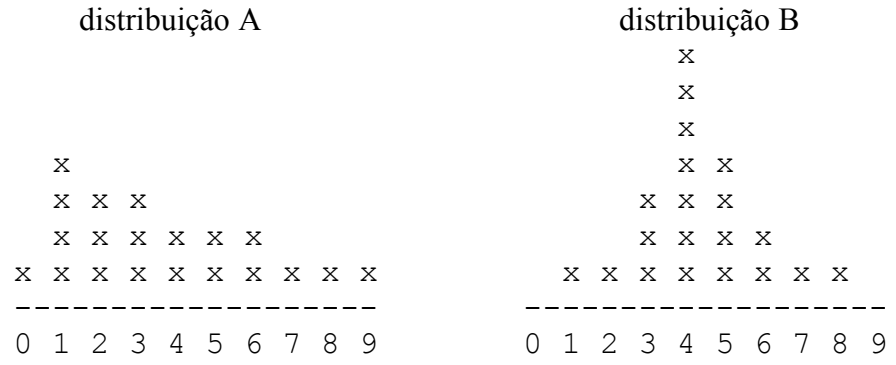
(Dados em Soares e Siqueira, 2002, p85)

6) A tabela abaixo mostra a distribuição das idades de 302 indivíduos que tentaram suicídio, obtida num levantamento feito em hospital de São Paulo em 1992.
   a) Complete as colunas em branco com as freqüências pedidas.
   b) Numa tabela de distribuição de freqüências como esta, que medidas de tendência central podem ser usadas? Calcule seus valores aproximados. Que medidas não podem ser usadas? Por que?
   c) Que medidas de dispersão que podem ser usadas? Calcule seus valores aproximados. Que medidas não podem ser usadas? Por que?
   d) Considerando a assimetria da distribuição, você espera que sua média seja maior ou menor do que a mediana ?
   e) Por que não é possível fazer o histograma desta tabela?

idade (anos)	f	fr	fr%	F
10 ⊢ 20	57			
20 ⊢ 30	113			
30 ⊢ 40	59			
40 ⊢ 50	32			
50 ⊢ 60	19			
60 ⊢ 70	7			
≥ 70	2			
indeterminada	13			
total				

(Soares e Siqueira, 2002, p.41)

7) A partir dos gráficos de pontos mostrado abaixo, encontre os valores dos quartis das duas distribuições e faça os diagramas de Tukey equivalentes. Você espera que as médias aritméticas destas distribuições sejam maiores ou menores que a mediana? Por que? Compare as formas das duas distribuições



8) A tabela abaixo mostra uma parte os resultados de uma pesquisa feita com uma amostra de 100 indivíduos. São exibidos as estatísticas dos níveis de glicose e de colesterol no sangue, e de pressão sistólica.

a) A julgar por estes resultados, qual destas três variáveis tem maior dispersão ?
b) Qual destas distribuições provavelmente tem pontos discrepantes?
c) Que se pode dizer sobre a assimetria destas distribuições? (Calcule os coeficientes de assimetria)

	média	mediana	desvio-padrão	intervalo quartílico	amplitude total
glicose	152	143,5	54,8	64,0	384
colesterol	215	216	35,3	51,0	164
pressão	129	128	19,8	24,0	98

9) Uma turma de 50 alunos obteve nota média igual a 5,5 numa prova. Outra turma, de 60 alunos, obteve média 6,8 na mesma prova. Se juntarmos as duas turmas, qual terá sido a média geral obtida pelos 110 alunos ?

10) É possível que uma distribuição venha a ter:
   a) a média igual a zero ?
   b) a média negativa ?
   c) a média aritmética com um valor menor do que o da mediana ?
   d) a média aritmética com um valor menor do que o do primeiro quartil ?
   e) o desvio-padrão igual a zero ?
   f) o desvio-padrão negativo?
   g) o coeficiente de variação maior que 100 % ?
   h) a mediana menor do que o primeiro quartil?
   i) a mediana igual a um dos quartis?
   j) a mediana igual ao primeiro quartil?
   k) a mediana igual a um dos valores discrepantes ?
   l) duas medianas? Duas modas?
   No caso afirmativo, que características teriam estas distribuições? Faça esboços dos gráficos destas distribuições.

11) As integrantes da seleção brasileira feminina de vôlei tinham, em 1986 as alturas:

180 179 178 178 178 178 177 176 176 175 174 172

enquanto que as integrantes da seleção de 1989 tinham as alturas:

188 185 184 183 180 180 180 178 178 177 176 176

Todos os dados estão em cm. A altura das jogadoras, em geral, aumentou ou diminuiu neste intervalo de 3 anos? Verifique graficamente (fazendo um gráfico de pontos duplo) e numericamente (calculando uma medida de tendência central). Calcule os desvios-padrões das duas distribuições. Qual das duas seleções você conclui ser mais homogênea, quanto a altura de suas jogadoras ?

(dados publicados pela revista *IstoE´*, 11 out. 1989).

12) Uma variável tem sua distribuição caracterizada por média = 8 e desvio padrão = 0. Suponha que esta variável assuma apenas três valores. Quais serão eles? Explique sua resposta.

13) Uma das medidas que um pediatra pode usar para avaliar o grau de controle do diabetes é a dosagem de glicohemoglobina, medida em % da hemoglobina total. Valores acima de 12 indicam que o paciente não está sob controle. Num levantamento feito num grupo de 15 crianças, um pediatra obteve os seguintes resultados para 13 delas:

10,5	11,2	11,7	12,1	12,3
12,3	12,3	12,4	12,4	12,5
13,1	13,1	13,5	*	*

Para as duas crianças restantes, sabe-se apenas que os valores são maiores de 14% (o aparelho de análise não consegue registrar valores além de 14%).
Para descrever estes dados, o pediatra igualou os dois dados faltantes a 15,5%, e calculou a média dos 15 dados. Você acha que este procedimento é válido? Existe alguma forma alternativa de descrever estes dados?

(Adaptado de Soares e Siqueira, 2002, p.83)

Referências:

Soares J.F., Siqueira A.L. (2002) *Introdução à Estatística Médica*, 2ª. ed. Belo Horizonte: Coopmed. (dados também disponíveis no site *www.est.ufmg.br/~estmed*)

Guedes M.L.S., Guedes J.S. (1988) Bioestatística para profissionais de saúde. Rio de Janeiro: Ao Livro Técnico (co-ed MCT/CNPq).

Lista 4 - Respostas

a) menor: 2981 maior: 16880

b)

```
 3 | 0 3
 4 | 0 0 2 4
 5 | 0 3 4 5 6 6 8
 6 | 0 1 1 2 3 4 4 6 7 8
 7 | 1 2 3
 8 | 0 4 4 7
 9 | 7
10 | 0 1 6 9
11 | 1 7
12 | 1 2
13 |
14 |
15 |
16 | 9
```

c)

```
 2 | 9
 3 | 2 9
 4 | 0 1 4 9
 5 | 3 4 4 6 6 7
 6 | 0 1 1 1 2 3 4 6 6 8
 7 | 0 2 3
 8 | 0 3 4 7
 9 | 7 9
10 | 0 5 9
11 | 1 6
12 | 1 1
13 |
14 |
15 |
16 | 8
```

d) Não há praticamente nenhuma diferença entre os dois gráficos. O arredondamento é um processo mais trabalhoso do que o truncamento; mas introduz um pequeno erro sistemático no resultado.

e) A média dos dados truncados será provavelmente menor do que a dos dados originais. A média dos dados arredondados estará em geral um pouco mais próxima da média real do que a média dos dados truncados.

f) mediana = 6.515

g) Q1= 5.550; Q2= 6515; Q3= 9217. Sim, 16.880 é um valor discrepante

2)

a) mediana = 216

b) média = cerca de 215 (o valor real é 216,96)

c) 63%

d)

1*	
1t	3
1f	4 5
1s	6 6 6 6 6 7 7 7 7 7 7 7
1-	8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9
2*	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2t	2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3
2f	4 4 4 4 4 4 5 5 5 5 5
2s	6 6 6 7 7
2-	8 8 8 9 9 9
3*	
3t	
3f	
3s	
3-	8

e)

1	3 4
-	5 6 6 6 6 6 7 7 7 7 7 7 7 8 8 8 8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9
2	0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4
-	5 5 5 5 5 6 6 6 7 7 8 8 8 9 9 9
3	
-	8

f) Os dois primeiros parecem mais úteis. O terceiro parece menos útil, pois tem uma escala muito reduzida, o que mascara a real forma da distribuição, escondendo alguns detalhes.

3)

a) Cálcio

Cálcio

Colostro		Maduro
45 13	1	
63 63 63	-	59 75 81 88
41 31 25 21	2	0 06 13 14 17 31 38 38 42 44
96 75 56 54	-	56 59 60 63 64 75 77 79 93
44 34 25 23 13 12 11 03	3	02 03 14 44
75 75 72		94
37	4	

Outra representação dos mesmos dados, usando uma escala diferente:

```
                    colostro        maduro
                           1|  1*  |
                            |   t  |
                           4|   f  |5
                         666|   s  |7
                           8|   .  |88
                            |  2*  |00111
                         322|   t  |333
                         554|   f  |4455
                           7|   s  |666777
                           9|   .  |89
                        1110|  3*  |001
                         322|   t  |
                           4|   f  |4
                         777|   s  |
                            |   .  |9
                            |  4*  |
                           3|   t  |
                            |   f  |
                            |   s  |
                            |   .  |
                            |  5*  |

n:                  26          29

                 legenda: 1|2 representa 120, leaf unit: 10
```

Zinco

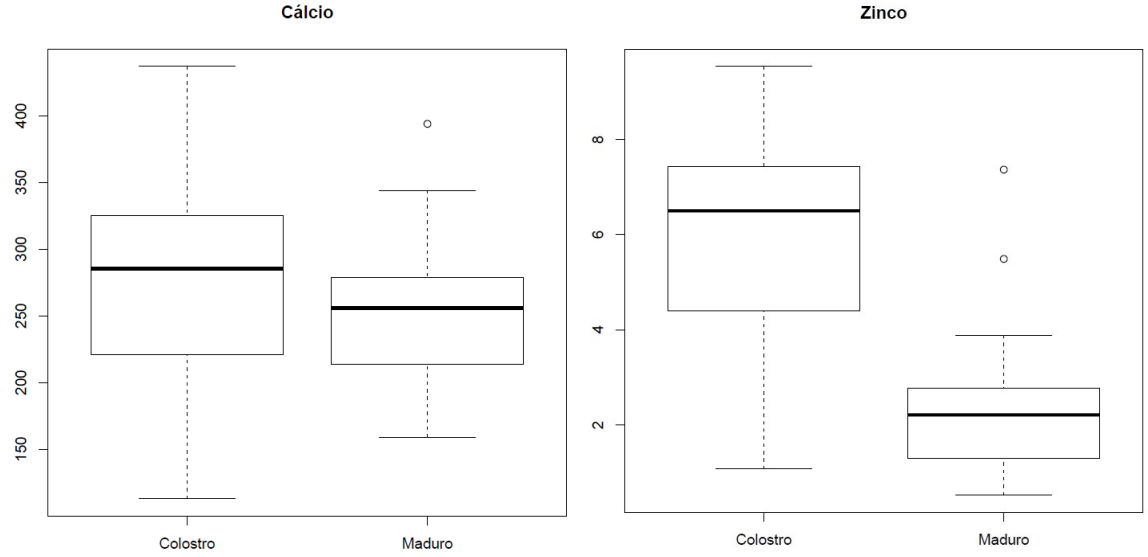
Colostro		Maduro
	0	52 60 78 86
30 20 07	1	08 23 28 31 40 55 55 60 92
	2	15 27 40 41 52 57 60 63 92
70 20 13	3	02 05 57 88
54 40	4	
82 20	5	50
90 82 82 50 13 07	6	
45 43 43 42 23	7	38
77 40 25	8	
54	9	

Zinco: colostro x maduro

```
                colostro          maduro
                       |   0 |5678          4
           3         320|   1 |022345569    13
                       |   2 |124455669    (9)
           6         721|   3 |0058          6
           8          54|   4 |
          10          82|   5 |5             2
         (6)      988510|   6 |
           9       44442|   7 |3             1
           4         742|   8 |
           1           5|   9 |
                       |  10 |

                legend : 1|2 represents 1.2
```

b) Diagramas de Tukey



c) Quanto à concentração de cálcio, não parece haver diferença significativa, embora o grupo maduro tenha menor dispersão. Já a concentração de zinco é claramente maior no grupo colostro.

4) a) Tabela:

classe	f
0,5 |- 1,0	1
1,0 |- 1,5	6
1,5 |- 2,0	9
2,0 |- 2,5	10
2,5 |- 3,0	7
3,0 |- 3,5	6
3,5 |- 4,0	3
4,0 |- 4,5	0
4,5 |- 5,0	1

c) Gráfico de ramo-e-folhas:

```
1|2 represents 1.2
 leaf unit: 0.1     n: 43

    1    0. | 8
    7    1* | 001334
   16    1. | 556688999
  (10)   2* | 0001113344
   17    2. | 5777899
   10    3* | 001223
    4    3. | 789
         4* |
    1    4. | 6
```

d) Pode-se citar, por exemplo, que o histograma permite fácil visualização de classes com intervalos diferentes, e o gráfico de ramo-e-folhas permite reter alguma informação adicional sobre os membros dos ramos (classes).

5) Unimodal, com assimetria negativa.

6)
a) Tabela:

classe	f	fr	F
10|-20	57	0,1887	0,1887
20|-30	113	0,3741	0,5629
30|-40	59	0,1953	0,7582
40|-50	32	0,1059	0,8642
50|-60	19	0,0629	0,9271
60|-70	7	0,0231	0,9503
>=70	2	0,0066	0,9569
indet	13	0,0430	1,0000

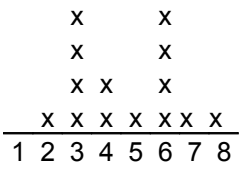
b. Pode-se usar a mediana (entre 20 e 30, mais próximo de 30), mas não é bom usar a média pois seria necessário conjecturar valores, o que é ainda mais problemático na faixa >=70 (sem limite superior).
c. Pode-se usar o intervalo quartílico (entre 0 e 20, provavelmente pouco mais que 10), mas não é bom usar o desvio-padrão, pois seria necessário conjecturar valores para cada uma das faixas.
d. Pela assimetria positiva, conclui-se que a média deve ser maior do que a mediana.
e. Por causa da faixa >=70, sem limite superior.

7)
A: Q1 = 1,5 Q2 = 3 Q3 = 5,5
média = 3,65
B: Q1 = 3,5 Q2 = 4 Q3 = 5
média = 4,35

8)
a. A glicose.
b. Possivelmente a glicose, e talvez também a pressão.
c. Glicose: A = 0,4653
Colesterol: A = -0,08498
Pressão: A = 0,1515

9) 6,21

10)
a) Sim b) Sim c) Sim d) Sim e) Sim f) Não
g) Sim h) Não i) Sim j) Sim k) Não
l) Não. Sim, uma distribuição pode ter duas modas, como por exemplo:



11)

						X										
						X										
				X		X										
X		X	X	X	X	X	X	X								
172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188
				X	X	X		X			X	X	X			X
				X		X		X								
								X								

Média: em 1986: 176,75
em 1989: 180,41

Mediana: em 1986: 177,5
em 1989: 180,0

Moda: em 1986: 178
em 1989: 180

Desvio padrão: em 1986: 2,26
em 1989: 3,82

12) Todos os valores são iguais a 8.

13) Não; conjeturar dados é sempre perigoso, pois pode introduzir viés nos resultados. O melhor seria usar a mediana, pois para calculá-la não precisamos dos valores exatos dos casos que estão acima de 14 %.

Lista 5 : Variáveis aleatórias

1) Dois *hamsters* dourados, que acredita serem heterozigotos para a cor, foram cruzados entre si. (*Hamsters* dourados heterozigotos têm genótipo **Aa**, **A** para a cor "dourada" e **a** para "creme"). Numa ninhada de oito filhotes,

h) Qual é o número mais provável de *hamsters* de cor creme?
i) Qual é a probabilidade de que haja exatamente 3 animais de cor creme?
j) Qual é a probabilidade de que não haja nenhum animal de cor creme?
k) Qual a probabilidade de todos estes filhotes serem fêmeas douradas?

2) Numa família de 5 crianças, calcule as probabilidades de que haja entre elas:
a) exatamente dois meninos b) exatamente quatro meninos c) mais de 3 meninos.
Considere P(menino) = P(menina) = 0,5.

3) Numa população animal, acredita-se que 10% dos animais estejam infectados com uma bactéria "X". Se uma amostra aleatória de 12 animais é retirada desta população,

a) será provável que 5 ou mais destes animais estejam infectados?
b) qual será o número mais provável de animais infectados na amostra?

4) Numa pesquisa feita entre os funcionários da fábrica da Philips, em Guarulhos (SP) em 1965, a altura média dos homens foi de 169.7 cm, com desvio padrão de 7.5 cm, enquanto a altura média das mulheres foi de 157.3 cm, com desvio padrão de 5.8 cm. Supondo que estes dados sejam válidos para a população brasileira, e que a distribuição das alturas seja normal, calcule as probabilidades de que:

f) um homem, escolhido ao acaso, tenha mais de 175 cm de altura;
g) uma mulher, escolhida ao acaso, tenha menos de 150 cm de altura;
h) num grupo de três homens escolhidos ao acaso, todos tenha alturas acima de 175 cm;
i) cite pelo menos três razões pelas quais a amostra acima não pode ser considerada com representativa da população brasileira, hoje.

5) Uma regra usada para identificar pontos discrepantes em uma amostra é a que considera discrepante todos os pontos que estiverem acima de $Q_3 + 1{,}5 \times (Q_3-Q_1)$ ou abaixo de $Q_1 - 1{,}5 \times (Q_3-Q_1)$, onde os "Q" são os quartis Se esta regra for aplicada aos valores de uma distribuição normal, qual será a porcentagem de valores que serão considerados discrepantes?

6) Uma pesquisa mostrou que o nivel de colesterol total dos habitantes de uma região tem distribuição aproximadamente normal, com média 180 mg/dL e desvio padrão de 32 mg/dL.

4) Qual é a proporção de habitantes desta região que tem níveis de colesterol acima de 200 mg/dL?
5) Qual é a proporção que tem níveis entre 200 e 250 mg/dL?
6) Qual é a proporção que tem níveis abaixo de 150 mg/dL?
7) Quais são os quartis desta distribuição? Qual é o intervalo quartílico?
8) Acima de que valor o nível de colesterol pode ser considerado "discrepante" ? (use a regra máx = Q3 + 1.5xIQ).

7) A tabela abaixo apresenta os resultados de exames laboratoriais solicitados a duas pacientes, mãe e filha, com respectivamente 60 e 40 anos de idade. Também são apresentados os parâmetros da população de mulheres adultas.

Média e desvio-padrão para adultos do sexo feminino e resultados de exames laboratoriais de mãe e filha

variável	parâmetros da população		valores observados	
	média	desvio padrão	mãe	filha
Glicemia em jejum	85	12,5	90	79
Ácido úrico	4,2	0,9	3,5	3,1
Triglicérides	105	30	97	66
Colesterol total	200	25	251	185

12) *padronize* os valores obtidos por estas duas pacientes, supondo que as distribuições destas quatro variáveis sejam gaussianas.
13) identifique se algum dos valores obtidos por estas pacientes pode ser considerado anormal (i.e., está afastado da média por mais de dois desvios padrões)

(Soares e Siqueira)

8) Uma das primeiras aplicações da distribuição normal foi na *Teoria dos Erros*. Basicamente, esta teoria diz que toda medida feita por um instrumento contém um erro. Qualquer medida é apenas uma *estimativa* do valor real da variável que se está medindo; este valor real, contudo, nunca poderá ser conhecido. Gauss descobriu, no século XIX, que medidas de uma mesma variável astronômica feitas por pesquisadores diferentes seguiam uma distribuição aproximadamente normal.
O erro de medição de nível de açúcar no sangue por um instrumento é uma variável aproximadamente normal, de média 0,05 e desvio-padrão de 1,5. O *erro* de medição é a diferença entre o *valor medido* pelo instrumento e o *valor real* do nível de açúcar.

k) qual será a probabilidade de o valor medido pelo instrumento, num paciente escolhido ao acaso, seja maior do que o valor real ? (isto é, de que o erro de medida seja positivo).
l) Suponha que o erro seja considerado "grave" quando a diferença (em módulo) entre o valor medido e o valor real seja superior a 2,8. Qual é a porcentagem de medições que conterão erros graves?

(Bhattacharyya & Johnson, p.227)

Lista 5 - Respostas

1) a) número mais provável =2
b) p(3) = 0,2076
c) p(0) = 0,1001
d) p(todas fêmeas douradas) = 0,000391

2) a) p(2) = 0,3125
b) p(4) = 0,1562
c) P(X>3) = p(4) + p(5) = 0,1875

3) a) não, já que o valor esperado é de apenas 1,2 animais. A probabilidade de X=7 seria apenas p = 0,004329
b) valor mais provável = 1 animal

4) a) P = 0,240
b) P = 0,104
c) P = 0,014
d) Esta amostra provavelmente não é representativa da população brasileira em geral, por várias razões:
- os dados foram todos obtidos em uma mesma cidade, em vez de cobrirem todo o país
- as pessoas pertencem todas a um mesmo grupo social (trabalhores de fábrica)
- os dados são antigos (1965), e a população provavelmente é mais alta hoje

5) porcentagem = 0,006976

6) (a) P(X>200) = 0,2660
(b) P(200<X<250) = 0,2516
(c) P(X<150) = 0,1743

8) Q1 = 158,4 mg/dL Q3= 201,6 mg/dL IQ= 43,2 mg/dL
(e) discrepantes: valores acima de 266,4 mg/dL

7) (a) Valores padronizados:

	mae	filha
Glc	0,40	-0,48
ac.ur.	-0,78	-1,22
TAG	-0,27	-1,30
CT	2,04	-0,60

(b) Sim, o colesterol total da mae está um pouco fora da faixa de normalidade.

8) (a) 0,5132
(b) 6,72 %