Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Exercícios Resolvidos em Introdução à Bioestatística

Edna Afonso Reis
Ilka Afonso Reis

Relatório Técnico – RTE-03/2000
Sexta Edição – outubro/2012

Índice Geral

Agradecimento

Gostaríamos de agradecer ao Prof. Aloísio J. F. Ribeiro, do Departamento de Estatística da UFMG, pela revisão cuidadosa da primeira edição deste trabalho e pelas valiosas contribuições, principalmente na complementação da solução dos exercícios da seção 11.

Primeira Parte:

Enunciado dos Exercícios

"*Se ouço, esqueço*; *se vejo, recordo;* ***se faço***, ***aprendo***."
(Provérbio chinês)

Índice da Primeira Parte - Enunciado dos Exercícios

Observação: Os exercícios marcados com asterisco (*) foram adaptados da apostila ***Introdução à Bioestatística***, de Nogueira et alli, Edição de 1997.

Seção 1: Tipos de Estudos e Variáveis

**1.1) Classifique as seguintes variáveis em :
Quantitativas (Discretas ou Contínuas) ou Qualitativas (Nominais ou Ordinais).**

a) A cor da pele de pessoas (ex.: branca, negra, amarela).

Variável do tipo ______________________ e ______________________

b) O número de consultas médicas feitas por ano por um associado de certo plano de saúde.

Variável do tipo ______________________ e ______________________

c) O teor de gordura, medido em gramas por 24 horas, nas fezes de crianças de 1 a 3 anos de idade.

Variável do tipo ______________________ e ______________________

d) O tipo de droga que os participantes de certo estudo tomaram, registrados como: Droga A, Droga B e placebo.

Variável do tipo ______________________ e ______________________

e) A pressão intra-ocular, medida em mmHg, em pessoas.

Variável do tipo ______________________ e ______________________

f) O número de filhos das pacientes participantes de certo estudo.

Variável do tipo ______________________ e ______________________

1.2) Classifique os seguintes estudos como observacionais ou experimentais. Identifique os grupos de comparação

a) Viagra para os diabéticos (Revista Istoé nº 1535 de 03/03/1999)
A famosa pílula azul pode também ser eficaz para diabéticos que têm a função erétil comprometida. Estudos preliminares haviam descartado a eficiência do Viagra nesses casos. Mas uma pesquisa realizada com 268 homens pela Universidade de Creighton, nos Estados Unidos, mostrou que 56% dos pacientes que tomaram Viagra tiveram melhora contra 10% dos que ingeriram placebo (pílula inócua). Mas em hipótese nenhuma se recomenda o uso do medicamento sem orientação médica.

Estudo do tipo ________________________________

Grupos de Comparação ____

____

b) Observar imagens de animais 'fofinhos' pode aumentar produtividade, diz estudo (Veja on-line, 02/10/2012)
Filhotes de cães e gatos são *kawaii*. Esta palavra japonesa pode ser traduzida para o português como aquilo que é fofinho. E tudo aquilo que é fofinho produz sentimentos positivos nas pessoas. Mas qual a influência dessas boas sensações no comportamento humano? (...) Os cientistas pediram a estudantes da universidade que realizassem três atividades que envolviam concentração. A cada exercício, os participantes eram divididos em duas equipes. Uma delas observava, antes do início das tarefas, um conjunto de fotografias de filhotes de cães e gatos. Aos demais eram entregues fotografias neutras. Em todos os casos, os alunos que viram imagens mais *kawaii* tiveram melhor desempenho do que os seus colegas. (...)
Artigo original: PLoS ONE 7(9): e46362. doi:10.1371/journal.pone.0046362 (URL: http://www.plosone.org/)

Estudo do tipo ________________________________

Grupos de Comparação ____
____

c) Alerta da pele (Revista Istoé nº 1537 de 17/03/1999)
Quem já teve câncer de pele deve redobrar os cuidados para não ser vítima de um outro tipo de tumor. Um estudo publicado no Jornal da Associação Médica Americana revelou que aqueles que tiveram câncer dermatológico estão 25% a 30% mais propensos a desenvolver um outro câncer até 12 anos depois de se terem curado. Acredita-se que o tumor de pele aumente a suscetibilidade geral do organismo a novos episódios da doença.

Estudo do tipo ________________________________

Grupos de Comparação ____
____

d) Gene da gordura (Revista Istoé nº 1537 de 17/03/1999)
Cientistas americanos anunciaram na semana passada ter descoberto em ratos o primeiro gene que suprime a obesidade e regula a queima de calorias. Essa pode ser a chave para o desenvolvimento de uma droga para manter as pessoas em forma. Na verdade, esse é o sexto gene relacionado com a obesidade, mas, de acordo com os pesquisadores, é o primeiro que age no metabolismo e consegue gastar energia. Eles submeteram dois grupos de ratos a testes com alimentos gordurosos. Aqueles com uma mutação nesse gene não ganharam peso enquanto que os normais engordaram.

Estudo do tipo ________________________________

Grupos de Comparação ____
____

e) Colesterol na medida (Revista Istoé nº 1536 de 10/3/1999)
Níveis muito baixos de colesterol podem ser prejudiciais, afirma um estudo divulgado na semana passada, no congresso da American Heart Association. A pesquisa comparou 714 vítimas de derrame com 3.743 pessoas saudáveis. Quem tinha colesterol acima de 280 estava duas vezes mais suscetível a sofrer derrame isquêmico (bloqueio de vaso sangüíneo). Aqueles com colesterol abaixo de 180 estavam duas vezes mais propensos a ter derrame hemorrágico. Explica-se: o colesterol ajuda na estrutura das veias e evita que elas se rompam. O ideal é mantê-lo no nível médio (200), como recomendam os órgãos de saúde.

Estudo do tipo ________________________________

Grupos de Comparação ____
____

f) Efeito protetor da vacina BCG em crianças (Boletim OPAS 1986)
Para avaliar o efeito protetor da vacina BCG em crianças com menos de 15 anos de idade, na cidade de Buenos Aires (Argentina), estudaram-se as crianças que receberam algum tratamento antituberculose durante o ano de 1981, tanto internados em hospitais ou tratados na forma ambulatorial. Para cada uma destas crianças, encontrou-se outra criança de mesma idade, sexo, condição sócio-econômica e que tinha tido alguma doença aguda, diferente da tuberculose, no mesmo período e que havia sido tratada no mesmo estabelecimento. Em ambos os grupos, considerou-se como vacinados os que tinham a cicatriz correspondente à vacina BCG em uma ou ambas regiões deltoidianas.

Estudo do tipo ______________________

Grupos de Comparação ____
____

g) Torcida bastante eficaz (Revista Istoé nº 1581 de 19/01/2000)
O sabor de vitória tem um efeito químico muito mais benéfico para a alma do que se acreditava. A conclusão é de uma equipe de pesquisadores americanos. Eles mediram o nível de testosterona (hormônio masculino) em torcedores de futebol e basquete e constataram um aumento de 20% do hormônio quando seus times vencem (quando os times perdem há uma queda de 20%). Como o hormônio regula o humor, a sensação de bem-estar e o interesse sexual, uma dose extra de testosterona vai bem.

Estudo do tipo ______________________

Grupos de Comparação ____
____

h) Animais contra alergia (Revista Istoé nº 1569 de 27/10/1999)
Brincar na fazenda, onde vivem animais como vacas, galinhas e outros bichos, diminui as chances de a criança desenvolver alergias. A constatação é de pesquisadores austríacos, que estudaram 2.283 crianças. Aquelas que tinham contato com animais eram três vezes menos sensíveis a problemas alérgicos e respiratórios, como a asma, do que as que vivem na zona urbana. A hipótese é a de que o contato precoce com os animais aumente a tolerância das células de defesa do organismo a bactérias e ácaros.

Estudo do tipo ______________________

Grupos de Comparação ____
____

Seção 2.1: Análise Descritiva e Exploratória de Dados
(Tabelas e Gráficos)

2.1.1) As Tabelas 2.1 e 2.2 mostram a Distribuição de Freqüências da idade dos alunos do curso de inglês, 150 estudantes da escola A e 200 estudantes da escola B, respectivamente.

Complete as tabelas, faça um histograma e uma ogiva para cada uma e compare os estudantes das duas escolas quanto à variável "idade dos alunos do curso de inglês".

Tabela 2.1: Idade dos alunos do curso de inglês para 150 estudantes da escola A

Idade	Ponto Médio da classe	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Absoluta Acumulada	Freqüência Relativa Acumulada (%)
10 |- 15		5			
15 |- 20		57			
20 |- 25		42			
25 |- 30		28			
30 |- 35		17			
35 |- 40		1			
Total		150			

Tabela 2.2: Idade dos alunos do curso de inglês para 200 estudantes da escola B

Idade	Ponto Médio da classe	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Absoluta Acumulada	Freqüência Relativa Acumulada (%)
10 |- 15		4			
15 |- 20		18			
20 |- 25		43			
25 |- 30		76			
30 |- 35		43			
35 |- 40		16			
Total		200			

Obs: lembre-se de que a ogiva deve ser feita com o limite superior de cada classe no eixo X.

2.1.2) Num estudo sobre a associação entre tromboembolismo e tipo sangüíneo, participaram 200 usuárias de contraceptivo oral. Dessas mulheres, 55 tinham tromboembolismo. Quanto ao grupo sanguíneo, o tipo A foi o mais numeroso, com 83 mulheres, seguido dos grupos O e B, com 79 e 27 mulheres, respectivamente. Das pacientes sadias, 70 eram do grupo O, 51 do grupo A e 19 do grupo B.

a) A partir dessas informações, preencha a Tabela 2.3 abaixo (Dica: comece encontrando os totais de linha e coluna)

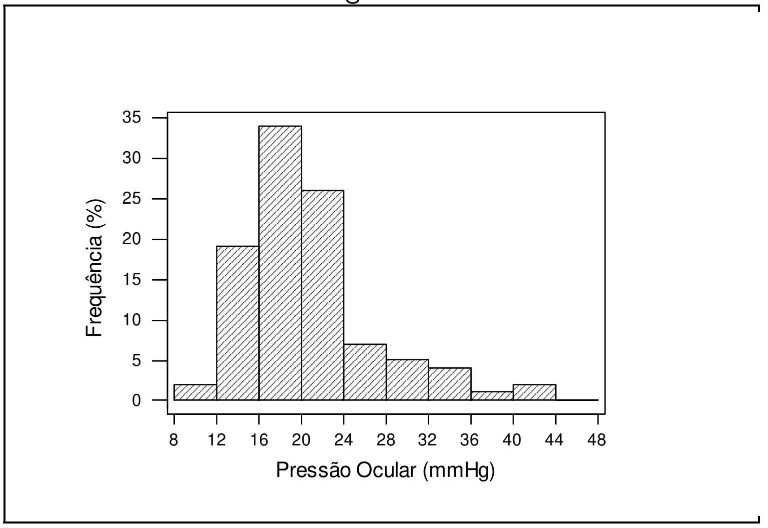
b) Utilizando a tabela preenchida no item a), compare os dois grupos de mulheres (sadias e doentes) de forma gráfica e/ou numérica.

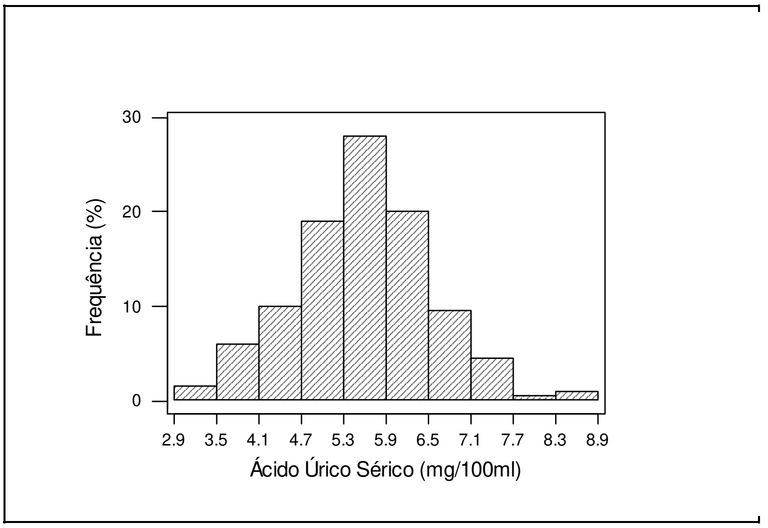
Tabela 2.3

Grupo Sangüíneo	Tromboembolismo		Total
	Doente	Sadia	
A			
B			
AB			
O			
Total			200

2.1.3) *Analisando os histogramas apresentados a seguir, comente sobre a distribuição da pressão ocular e do nível de ácido úrico sérico.

Figura 2.1





2.1.4) *O tempo (em meses) entre a remissão (cura) de uma doença e a recidiva (volta) de 48 pacientes de uma determinada clínica médica foi registrado. Os dados ordenados são apresentados a seguir, para os 24 homens e as 24 mulheres participantes do estudo.

Homens:	2 2 3 4 4 4 4 7 7 7 8 9 9 10 12 15 15 15 16 18 18 22 22 24
Mulheres:	2 2 3 3 4 4 5 5 6 6 7 7 7 7 8 8 8 8 10 10 11 11 12 18

a) Construa o diagrama de pontos para cada sexo e comente-os.

b) Construa um ramo-e-folhas para cada sexo usando 10 meses como escala e outro usando 5 meses.

Exemplo:

Escala de 10 meses	Escala de 5 meses
0 |	0 |
1 |	0 |
2 |	1 |
	1 |
	2 |
	2 |

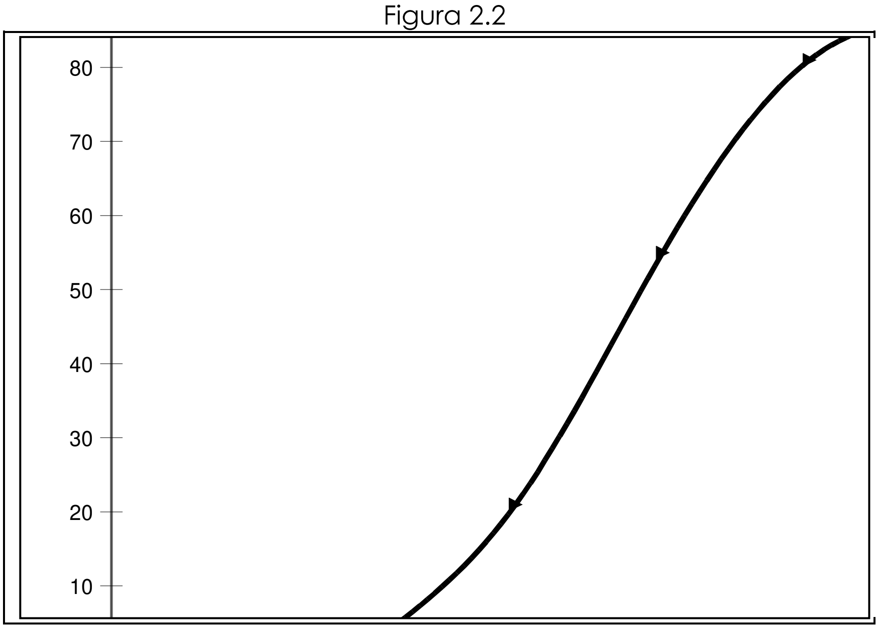
Seção 2.2: Análise Descritiva e Exploratória de Dados
(Síntese Numérica e *Boxplot*)

2.2.1) Com os dados apresentados no exercício 2.1.4 (Parte 1) :

Calcule a média, o desvio padrão, a mediana e o coeficiente de variação do tempo entre a cura e a recidiva da doença para cada sexo. Comente essas estatísticas descritivas, comparando os grupos.

2.2.2) *Utilizando a ogiva apresentada a seguir (Figura 2.2), estime a mediana, o primeiro e o terceiro quartis e o percentil de ordem 95. Interprete estes valores.

Figura 2.2



2.2.3) *Em um quartel do exército, a média do nível de colesterol de um grupo de jovens recrutas é 205 mg/dl e o desvio padrão é 22 mg/dl. Para um grupo de oficiais, entretanto, a média obtida foi 244 mg/dl e o desvio padrão 45 mg/dl.

a) Compare os dois grupos quanto à homogeneidade da variável nível de colesterol.

b) Um militar foi transferido para este quartel e teve o seu nível de colesterol avaliado, resultando em 224,5 mg/dl. Considerando seu nível de colesterol, de qual dos dois grupos (oficiais ou recrutas) esse novo membro mais se aproxima?

2.2.4) Considere as alturas dos cascos de uma amostra de 48 tartarugas pintadas (24 machos e 24 fêmeas).

Machos	35	35	35	37	37	38	38	39	39	40	40	40
	40	41	41	41	42	43	44	45	45	45	46	47
Fêmeas	38	38	42	42	44	46	48	49	50	51	51	51
	51	51	53	55	56	57	60	61	62	63	63	67

a) Complete a tabela de estatísticas descritivas para a variável "altura de casco", segundo o sexo da tartaruga.

Grupo	Média	Desvio padrão	C. V.	Mediana	Primeiro Quartil	Terceiro Quartil	Percentil 10
Macho	40,54	3,54					
Fêmea	52,04	8,05					

b) Construa um box-plot para as medidas de "altura de casco" de cada grupo (Utilize a mesma escala nos dois gráficos para que eles sejam comparáveis).
c) Usando os resultados do item (a) e os box-plots em (b), *compare* os dois grupos de tartaruga pintada quanto a variável "altura do casco".

Exemplo de uso de percentis: Faixa de Referência

Uma *Faixa de Referência* é um intervalo de valores de determinada característica que contém uma certa porcentagem de indivíduos da população sadia. Por exemplo, uma faixa de referência de 90% para a taxa de glicose no sangue de homens adultos vai de 80 a 100 mg/dl. Isto quer dizer que 90% dos homens adultos sadios possuem taxa de glicose entre 80 e 100 mg/dl.

Uma das maneiras de construir uma faixa de referência é usar *percentis* como os limites inferior e superior da faixa. A abordagem mais comum é construir uma *faixa de referência simétrica*, de modo que a proporção de indivíduos abaixo do limite inferior seja igual à proporção de indivíduos acima do limite superior da faixa. Assim, se desejamos uma faixa de referência de 90%, então 10% dos indivíduos sadios ficarão de fora da faixa. Destes, 5% terão valores inferiores ao limite inferior e 5% terão valores acima do limite superior.

Isto nos remete à idéia de percentil. O limite *inferior* da faixa deve ser o valor que deixa 5% *abaixo* dele, ou seja, o percentll de ordem 5. Do mesmo modo, o limite *superior* da faixa deve ser o valor que deixa 5% *acima* dele, ou seja, 95% *abaixo* dele, isto é, o percentil de ordem 95.

De modo geral, se desejarmos uma Faixa de Referência de (100-α)%, onde α é a porcentagem de indivíduos que ficarão do lado de fora da faixa, então os limites inferior e superior da faixa serão dados, respectivamente, pelos percentis de ordem $\alpha/2$ e de ordem (100-$\alpha/2$).

Faixa de Referência de (100-α)% = [$P_{\alpha/2}$; $P_{(100-\alpha/2)}$]

Este é o chamado *Método dos Percentis* para o cálculo de uma Faixa de Referência.
Exemplo: Faixa de Referência de 95% = [$P_{2.5}$; $P_{97.5}$], pois $\alpha = 5$, $\alpha/2 = 2.5$ e $100-\alpha/2=97.5$
Faixa de Referência de 99% = [$P_{0.5}$; $P_{99.5}$], pois $\alpha = 1$, $\alpha/2 = 0.5$ e $100-\alpha/2=99.5$

2.2.5) *Um pesquisador deseja criar um padrão para identificação de infecção bacteriana (*pseudomonas sp*) no trato respiratório, através de cultura de escarro. Para isto, coletou dados de pessoas sabidamente sadias e determinou o número de colônias encontradas em cada cultura de escarro. Os resultados foram os seguintes:

17	22	23	23	23	23	24	24	24	24	24	24	25	25	25	25	25	25	25	26	28	28
29	30	30	31	31	35	35	35	36	40	41	41	41	42	51	54	56	56	58	60	68	79

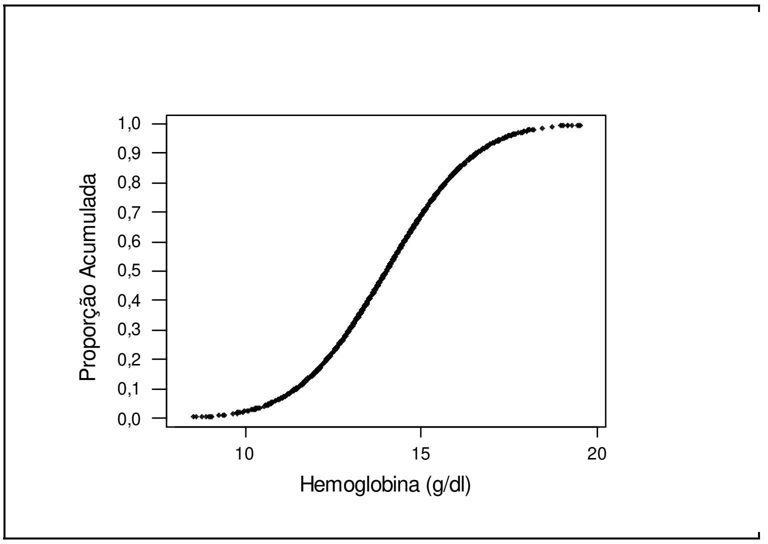
(n = 44 pessoas; $\bar{x}$ = 34,3 colônias por cultura de escarro; s =14,2 colônias por cultura de escarro)

Construa um ramo-e-folhas para representar graficamente estes dados e determine uma faixa de referência de 95% para o número de colônias de bactérias no trato respiratório de pessoas sadias, usando o Método dos Percentis.

2.2.6) Um dos parâmetros hematológicos de uso rotineiro na clínica médica para diagnóstico da anemia é o valor de hemoglobina. A figura abaixo apresenta a ogiva de proporções acumuladas para 1000 mulheres sadias de uma amostra, na qual o valor médio de hemoglobina foi de 14 g/dl e o desvio padrão de 2 g/dl.

Construa uma Faixa de Referência de 80% para o valor de hemoglobina em mulheres sadias usando o Método dos Percentis. Interprete o resultado obtido.

Figura 2.3 - Ogiva de proporções acumuladas para hemoglobina (g/dl) em uma amostra de 1000 mulheres sadias



Seção 2.3: Análise Descritiva e Exploratória de Dados
(Associação entre Variáveis)

2.3.1) *Como resultado de um programa de fortificação isométrica desenvolvido em 10 semanas, alunos da oitava série foram avaliados em duas ocasiões, antes e após o programa, quanto a sua habilidade em executar abdominais em dois minutos. Os dados são apresentados a tabela a seguir. Quanto maior o escore, maior é a habilidade do aluno em executar abdominais de dois minutos.

Analise graficamente a efetividade ou não do programa isométrico no aumento da habilidade em executar abdominais nestes alunos. (Use o gráfico de dispersão Antes *versus* Depois [gráfico XY] e desenhe a reta x=y).

Tabela 2.4

No do aluno	Escores de abdominais	
	Antes	Depois
1	12	15
2	10	9
3	23	25
4	25	25
5	29	31
6	32	30
7	14	16
8	17	20
9	19	22
10	20	22

2.3.2) A tabela a seguir apresenta a resposta de pacientes que tomaram vacina ou placebo para o tratamento de gripe.

Tabela 2.5: Distribuição da Resposta ao Tratamento de Gripe segundo Grupo de Tratamento

Resposta	Placebo	Vacina	Total
Baixa	25	6	31
Moderada	8	18	26
Alta	5	11	16
Total	38	35	73

a) Calcule as percentagens por linha e por coluna.

b) Analisando as percentagens por coluna, você acredita que haja associação entre a intensidade da resposta e o grupo do paciente (vacina ou placebo)?

2.3.3) Quatorze estudantes do segundo ano do curso de Nutrição aferiram a pressão sanguínea do aciente. Os resultados estão listados a seguir.

Sistólica	138	130	135	140	120	125	120	130	130	144	143	140	130	150
Diastólica	82	91	100	100	80	90	80	80	80	98	105	85	70	100

a) Faça o diagrama de dispersão e analise a relação entre as duas medidas.

b) A correlação entre os valores de pressão sistólica e diastólica é nula, positiva ou negativa? Calcule o coeficiente de correlação de Pearson e comente.

Seção 3: Probabilidade

3.1) Considere um baralho com 52 cartas numeradas, 13 para cada um dos naipes (ouros, copas, espada e paus). Seja o experimento de retirar uma carta aleatoriamente, observando seu naipe, número e/ou cor (vermelha ou preta).
Sejam os seguintes eventos:

A = [a carta retirada é um ás];
V = [a carta retirada é vermelha] e
E = [a carta retirada é de espada].

Calcule:

a) P(A), P(V) e P(E).

b) P(A $\cap$ V), P(A $\cap$ E) e P(V $\cap$ E).

c) P(A $\cup$ V), P(A $\cup$ E) e P(V $\cup$ E).

d) P(A | V) . Os eventos A e V são independentes?

e) P(V | E). Os eventos V e E são independentes?

f) Suponha que você retire do baralho, aleatoriamente, duas cartas do seguinte modo: retira uma, observa seu naipe, número e cor, e a coloca de volta. Em seguida, retira a segunda carta, observa seu naipe, número e cor, e a coloca de volta. Sejam os eventos:
A1 = [a primeira carta retirada é um ás] e A2 = [a segunda carta retirada é um ás].

f.1) Sem fazer cálculos, você acha que os eventos A1 e A2 são independentes ? Ou seja, você acha que o fato da primeira carta retirada ter sido um ás altera a probabilidade de que a segunda carta seja um ás ? Então, qual é o valor de P(A2 | A1) ?

f.2) Qual é a probabilidade das duas cartas retiradas serem um ás ? Ou seja, calcule P(A1 $\cap$ A2).

3.2) Considere o seguinte experimento aleatório: "sortear 3 pessoas ao acaso em uma população e verificar se elas são daltônicas ou não". A probabilidade de uma pessoa ser daltônica nesta população é ¼. Considere também que todas três sorteadas têm a mesma probabilidade de ser daltônica, ou seja, ¼, e que as três pessoas são independentes entre si

a) Enumere o espaço amostral deste experimento *(Dica: veja que os elementos deste espaço amostral são eventos compostos. Se chamarmos de* D *o evento "a pessoa é daltônica" e de* $\overline{D}$ *o evento "a pessoa não é daltônica", um dos pontos do espaço amostral é (*$D, \overline{D}, \overline{D}$*), ou seja, a primeira pessoa é daltônica e duas últimas não o são).*

b) Calcule a probabilidade de:
- b.1) Sortear nenhuma pessoa daltônica;
- b.2) Sortear uma pessoa daltônica, não importando a ordem em que ela apareça;
- b.3) Sortear duas pessoas daltônicas, não importando a ordem em que elas apareçam;
- b.4) Sortear todas as três pessoas daltônicas.

(Dica: primeiramente, calcule a probabilidade de cada um dos pontos do espaço amostra identificado na letra a).

c) Você consegue notar algum padrão no cálculo das probabilidades em b)? Se sim, qual é este padrão?

3.3) *É bem conhecido que o daltonismo é hereditário. Devido ao fato do gene responsável ser ligado ao sexo, o daltonismo ocorre mais freqüentemente nos homens do que nas mulheres. As 10.000 pessoas de uma amostra aleatória de uma população foram classificadas de acordo com seu sexo e se sofrem ou não de daltonismo da cor vermelha-verde. Os resultados são mostrados na Tabela 3.1.

Tabela 3.1

Daltonismo	Sexo		Total
	Masculino	Feminino	
Presente	423	65	
Ausente	4848	4664	
Total			10.000

Notação: Sexo Masculino: M   Daltonismo Presente: D

Sexo Feminino: F   Daltonismo Ausente: $\overline{D}$

I) Complete a Tabela 3.1.

II) Uma pessoa é escolhida ao acaso desta população. Estime a probabilidade de esta pessoa ser:

a) Daltônica
b) Não daltônica
c) Do sexo masculino
d) Do sexo feminino
e) Daltônica e do sexo masculino
f) Daltônica e do sexo feminino
g) Não daltônica ou do sexo masculino
h) Não daltônica ou do sexo feminino
i) Daltônica dado que é do sexo masculino
j) Daltônica dado que é do sexo feminino
k) Não daltônica dado que é do sexo masculino
l) Não daltônica dado que é do sexo feminino

III) Os eventos "ser daltônica" e "ser do sexo masculino" são independentes ?

3.4) Dentre os recém-nascidos de determinado grupo étnico, a probabilidade de um bebê nascer com menos de 37 semanas de gestação (prematuro) é 0,142. Dentre os bebês que nasceram com menos de 37 semanas de gestação, a probabilidade de nascer com baixo peso (menos de 2500 gramas) é 0,218 e, dentre os que nasceram com 37 semanas ou mais, a probabilidade de nascer com baixo peso é 0,023.

a) Nesse grupo étnico, qual é a probabilidade de que um bebê nasça com baixo peso
a.1) e seja prematuro?
a.2) mas não seja prematuro?

b) Nesse grupo étnico, qual é a probabilidade de que um bebê nasça com baixo peso?

c) Sabendo que um bebê nasceu com baixo peso, qual é a probabilidade de que sua gestação tenha durado menos do que 37 semanas?

3.5) A detecção precoce do câncer cervical é crucial para o tratamento e cura da paciente. As 600 mulheres de amostra aleatória foram classificadas em um de dois grupos: "com câncer" ou "sem câncer" através de biópsia cervical (Tabela 3.2). Outro teste que pode ser usado no diagnóstico do câncer cervical é o papanicolau, mais barato e mais rápido que a biópsia cervical. Para avaliar a qualidade de diagnóstico do papanicolau, as 600 mulheres mencionadas anteriormente foram submetidas a este teste. Os resultados do teste papanicolau são mostrados na Tabela 3.2 ("Positivo" indica que o teste classifica a paciente como portadora do câncer; "negativo", caso contrário). Assuma que o resultado da biópsia cervical é certo.

Tabela 3.2

Situação da paciente	Resultado do Papanicolau		Total
	Positivo	Negativo	
Com câncer	94	6	100
Sem câncer	250	250	500
Total	344	256	600

a) Estime a proporção de mulheres que têm câncer cervical na população de onde foi retirada esta amostra (ou seja, a prevalência do câncer na população).

b) Para quantas pacientes o teste papanicolau acertou o diagnóstico ?

c) Para quantas pacientes o teste papanicolau errou o diagnóstico ?

d) Qual é a probabilidade do teste papanicolau ter resultado positivo dentre as pacientes que realmente têm câncer ? (Esta probabilidade é chamada sensibilidade do teste)

e) Qual é a probabilidade do teste papanicolau ter resultado negativo dentre as pacientes que não têm câncer ? (Esta probabilidade é chamada especificidade do teste)

f) Qual é a probabilidade de uma paciente realmente ter câncer dentre aquelas com resultado positivo no teste papanicolau? (Esta probabilidade é chamada valor de predição positiva do teste)

g) Qual é a probabilidade de uma paciente realmente não ter câncer dentre aquelas com resultado negativo no teste papanicolau? (Esta probabilidade é chamada valor de predição negativa do teste)

h) Qual é a probabilidade de uma paciente realmente não ter câncer dentre aquelas com resultado positivo no teste papanicolau? (Esta é a proporção de falsos positivos do teste)

i) Qual é a probabilidade de uma paciente realmente ter câncer dentre aquelas com resultado negativo no teste papanicolau? (Esta é a proporção de falsos negativos do teste)

Seção 4: Avaliação da Qualidade de Testes Clínicos

4.1) No contexto da avaliação da qualidade de um teste clínico, associe as definições a seguir com: sensibilidade, especificidade, valor de predição positiva, valor de predição negativa, proporção de falsos positivos ou proporção de falsos negativos.

a) A probabilidade de um paciente com resultado positivo não estar doente.

Definição de: ____

b) A probabilidade do paciente não estar doente dado que seu resultado no teste foi negativo.

Definição de: ____

c) A probabilidade do resultado do teste ser negativo dentre os pacientes realmente não doentes.

Definição de: ____

d) A probabilidade do paciente realmente estar doente dado que seu resultado no teste foi positivo.

Definição de: ____

e) A probabilidade de um paciente com resultado negativo estar doente.

Definição de: ____

f) A probabilidade do resultado do teste ser positivo dentre os pacientes realmente doentes.

Definição de: ____

4.2) Um dos testes utilizados para detectar a doença de Aujeszky em suínos, também conhecida como pseudo-raiva, é o teste ELISA. Na tabela abaixo são apresentados os resultados deste teste para 52 suínos portadores da doença e 238 não portadores da doença.

Tabela 4.1

Teste ELISA	Doença de Aujeszky		Total
	Doente (D)	Doente ($\overline{D}$)	
Positivo (+)	51	6	57
Negativo (-)	1	232	233
Total	52	238	290

a) Calcule a sensibilidade e a especificidade do teste.

b) Se a prevalência dessa doença (na população) é de 17,9%, você pode calcular o VPP e o VPN diretamente da tabela ? Por quê ? Calcule estes índices.

c) Suponha que a prevalência da doença seja bem menor que 17,9%. Sabendo que o tratamento da pseudo-raiva é relativamente caro, qual atitude deveria ser tomada para um suíno que apresentasse resultado positivo no teste ELISA?

4.3) *A creatinina fosfacinase (CFC) é um marcador para o diagnóstico de infarto agudo do miocárdio. Pacientes com infarto agudo do miocárdio apresentam valores elevados de CFC. A tabela a seguir apresenta os valores de CFC para 360 pacientes de um hospital do coração, sendo 230 com infarto agudo do miocárdio. Considere que a prevalência de infarto agudo do miocárdio em hospitais do coração seja igual a desse estudo.

Tabela 4.2

Infarto agudo do miocárdio	Valores de CFC			Total
	CFC menor do que 80	CFC entre 80 e 280	CFC maior do que 280	
Doente (D)	15	118	97	**230**
Não doente ($\overline{D}$)	114	15	1	**130**
Total	**129**	**133**	**98**	**360**

Sejam dois testes de diagnóstico de infarto agudo do miocárdio baseados no valor de CFC do paciente:

Teste 1: O resultado é positivo se os valores de CFC são maiores do que 80 e negativo caso contrário;
Teste 2: O resultado é positivo se os valores de CFC são maiores do que 280 e negativo caso contrário.

a) Calcule a sensibilidade e a especificidade do Teste 1 e do Teste 2. Comente.

b) Calcule o VPP, VPN, PFP e PFN do Teste 1 e do Teste 2.

c) Considerando a prevalência de infarto do miocárdio calculada na tabela 4.2, qual dos dois testes (1 ou 2) será mais útil para
c.1) confirmar a presença do infarto do miocárdio? Por quê?
c.2) descartar a presença do infarto do miocárdio? Por quê?

4.4) Sabe-se que o Valor de Predição Positiva (VPP) e o Valor de Predição Negativa (VPN) de um teste clínico de diagnóstico de uma doença depende da Prevalência (p) da doença na população, da Sensibilidade (s) e da Especificidade (e) do teste.

Estas relações podem ser expressas através das equações:

$$VPP = \frac{sp}{sp + (1-e)(1-p)} \quad \textbf{e} \quad VPN = \frac{e(1-p)}{e(1-p) + (1-s)p}$$

A Figura 4.1 mostra a relação entre VPP e Especificidade e entre VPN e Especificidade para vários valores de Sensibilidade, mantendo-se p=0,05. Analogamente, estas mesmas relações são mostradas Figura 4.2, com p=0,5.

a) Considere as figuras 4.1(a) e 4.2(a). Para um valor fixo de prevalência e sensibilidade, o que acontece com o VPP à medida que a especificidade aumenta ?

b) Considere as figuras 4.1(b) e 4.2(b). Para um valor fixo de prevalência e sensibilidade, o que acontece com o VPN à medida que a especificidade aumenta ?

c) Considere a Figura 4.2(a), onde a prevalência está fixada em p=0,5.

   i. Para um valor fixo de sensibilidade, qual é, aproximadamente, o aumento que se tem no VPP quando se passa de e=0,7 para e=0,95 ?

   ii. Para um valor fixo de especificidade, qual é, aproximadamente, o aumento que se tem no VPP quando se passa de s=0,7 para s=0,95 ?

   iii. O teste será poderoso para confirmar a presença da doença se ele tiver PFP grande ou pequena?

   iv. Suponha que se queria que o teste fosse poderoso para confirmar a presença da doença em uma população com p=0,5. Baseado na sua resposta em (iii) e comparando suas respostas em (i) e (ii), você acha melhor trabalhar para aumentar a sensibilidade ou a especificidade do teste?

d) Considere a Figura 4.1(b), onde a prevalência está fixa em p=0,05

   i. Para um valor fixo de sensibilidade, qual é, aproximadamente, o aumento que se tem no VPN quando se passa de e=0,7 para e=0,95 ?

   ii. Para um valor fixo de especificidade, qual é, aproximadamente, o aumento que se tem no VPN quando se passa de s=0,7 para s=0,95 ?

   iii. O teste será poderoso para descartar a presença da doença se ele tiver PFN grande ou pequena?

   iv. Suponha que se queira que o teste seja poderoso para descartar a presença da doença em uma população com p=0,05. Baseado na sua resposta em (iii) e comparando suas respostas em (i) e (ii), você acha melhor trabalhar para aumentar a sensibilidade ou a especificidade do teste?

Figura 4.1

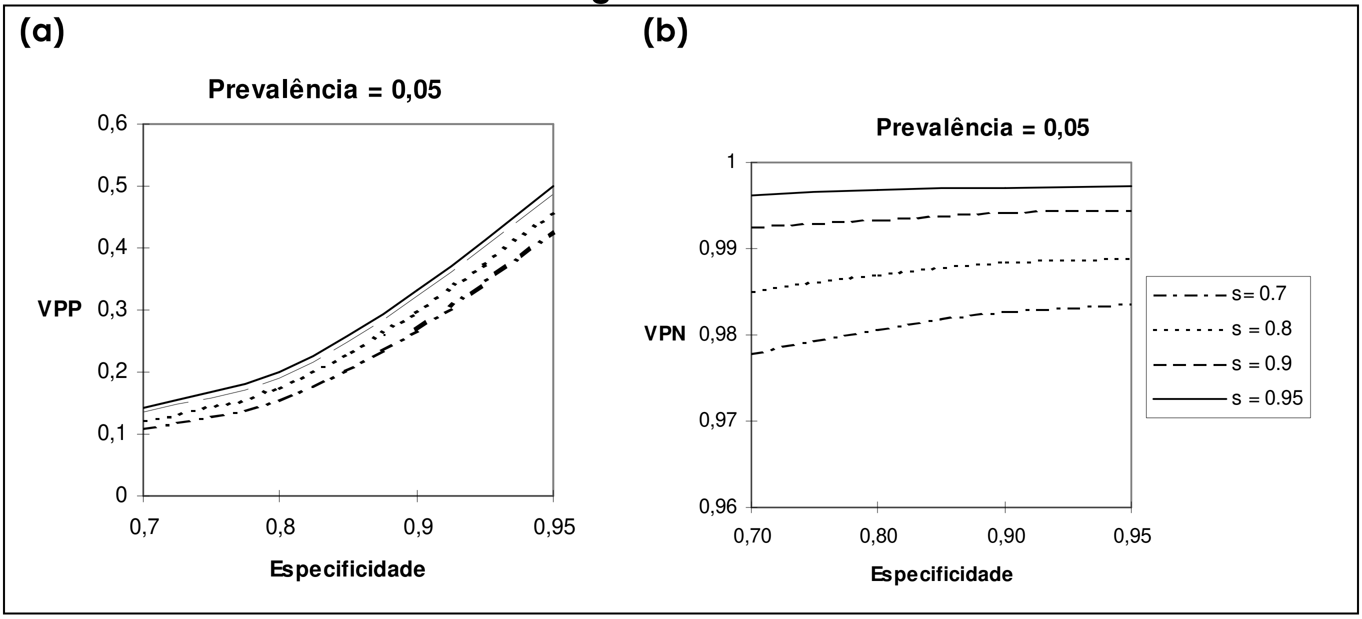
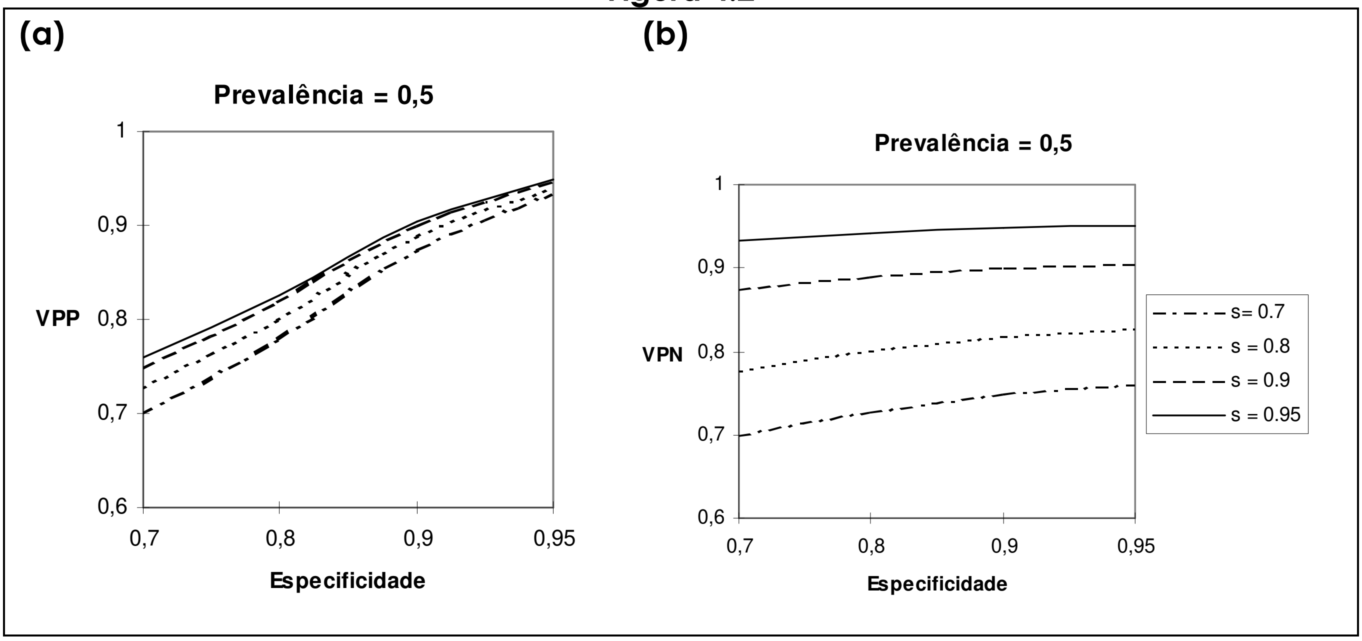


Figura 4.2



4.5) Considere os dados do Exercício 2.2.6 da Seção 2.2 (valores de hemoglobina em mulheres sadias)

Considerando o limite inferior da faixa de referência calculado no exercício, responda:

a) que porcentagem de mulheres sadias estará acima do limite inferior? E abaixo dele?

b) qual seria a especificidade de um método de diagnóstico de anemia baseado nesse limite inferior de referência?

c) qual seria a sensibilidade?

Seção 5: Distribuições de Probabilidade: Binomial e Poisson

5.1) *Em uma determinada população, a probabilidade de um indivíduo ter sangue Rh negativo é de 0,10. Qual é a probabilidade de 4 indivíduos dessa população que se apresentarem para o exame de sangue serem todos Rh negativo ?

5.2) A probabilidade de que um casal com olhos azuis escuros tenha filhos com olhos azuis é de ¼ . Se esse casal tiver 3 filhos, qual é a probabilidade de que pelo menos 2 tenham olhos azuis?

5.3) *Suponha que o número de unidades vendidas de um medicamento por dia numa rede de drogarias tenha uma distribuição de Poisson com média 3 ($\lambda = 3$).

a) Construa o gráfico para a distribuição de probabilidade do número de unidades vendidas desse medicamento por dia (use a tabela de Poisson).
b) Calcule a probabilidade de que sejam vendidas entre 7 e 11 unidades por dia.
c) Qual é distribuição do número de unidades vendidas desse medicamento por semana ?
d) Qual é a probabilidade de que 100 unidades desses medicamento sejam vendidas em uma semana ?

5.4) Em uma certa população, a probabilidade de um menino ser daltônico é 0,08. Num grupo de 4 meninos vindos dessa população, qual é a probabilidade de 3 não serem daltônicos ?

5.5) A probabilidade de um animal sobreviver durante um experimento cirúrgico é 2/3. Seja X o número de animais que sobrevivem quando 5 animais são submetidos à cirurgia.

a) Determine a distribuição de probabilidade de X.

b) Determine a probabilidade de :
   i. Exatamente 3 animais sobreviverem
   ii. No mínimo 1 animal sobreviver
   iii. Mais de 2 animais não sobreviverem

c) Se 60 animais se submeterem a essa cirurgia, espera-se que, em média, quantos não sobrevivam?

5.6) Um produtor de sementes vende pacotes com 10 sementes cada. Os pacotes que apresentam mais de quatro sementes sem germinar serão indenizados. A probabilidade de uma semente germinar é 0,8.

a) Qual é a probabilidade de um pacote ser indenizado ?

b) Se o produtor vende 1000 pacotes, qual é o número esperado de pacotes indenizados ?

5.7) Suponha que o número médio de colônias de bactérias por 10 ml de água de um lago seja 3.

a) Qual a probabilidade de não se achar nenhuma colônia em 10 ml de água desse lago ?

b) Qual a probabilidade de se achar pelo menos duas colônias em 10 ml de água desse lago ?

c) Qual a probabilidade de que no máximo 4 colônias sejam achadas em uma amostra de 30 ml de água desse lago?

5.8) Para estudar a relação entre a saúde da mulher e o tamanho de sua prole, um grupo de pesquisas precisa selecionar mulheres que tenham tido, pelo menos, 3 filhos. Na população-alvo do estudo, o número médio de filhos por mulher é 2,5. Considerando que o número de filhos por mulher nessa população siga a distribuição de Poisson, responda:

a) Qual é a probabilidade de que uma mulher selecionada aleatoriamente da população-alvo atenda ao critério de seleção para participação no estudo?

b) Um grupo de 10 mulheres mostrou interesse em participar do estudo. Sabendo que elas fazem parte da população-alvo, qual é a probabilidade de que mais de 7 delas atendam ao critério de seleção do estudo? (*Deixe os cálculos indicados até o ponto de usar a calculadora, não sendo necessário terminá-los*).

Seção 6: Distribuições de Probabilidade: Normal

6.1) Suponha que a quantidade de ferro sérico de indivíduos sadios de uma população (variável X) tenha distribuição Normal com parâmetros μ= 100 mcg/dl e σ= 25 mcg/ dl.

a) Qual é o valor da quantidade média de ferro sérico em indivíduos desta população ?
E qual é o valor do desvio padrão da quantidade de ferro sérico nesta população ?

b) Faça um esboço da curva normal X definida no enunciado, representando os intervalos simétricos em torno da média correspondentes às probabilidades 0,683, 0,954 e 0,997 ($\mu \pm \sigma$; $\mu \pm 2\sigma$; $\mu \pm 3\sigma$).

c) Como é a equação da variável normal padrão Z neste caso ?

d) Complete o quadro a seguir:

Um valor de X igual a ...	Corresponde a Z igual a ...	$P(X \leq x)$ =
25		
	-2	
		0,1587
59		
	-1,28	
		0,2482
90		

Um valor de X igual a ...	Corresponde a Z igual a ...	$P(X \geq x)$ =
	3	
		0,0228
125		
	1,64	
		0,1001
117		
	0,4	

Obs: Use duas casas decimais para Z, nenhuma para X e quatro para $P(X \leq x)$ e $P(X \geq x)$

i. Qual é a probabilidade de X assumir um valor entre 68 e 110 mcg/dl ?
Se tomarmos uma amostra aleatória de 500 indivíduos sadios desta população, quantos indivíduos podemos esperar que tenham quantidade de ferro sério entre 68 e 110 mcg/dl ?

ii. Qual é o intervalo $[x_1;x_2]$ simétrico em torno da média, que contém 50% dos valores de X ?

iii. Qual é o intervalo $[x_3;x_4]$ simétrico em torno da média, que contém 95% dos valores de X ?

6.2) *Suponha que a concentração sérica de tiroxina T4(D) em cães machos sadios tenha distribuição Normal com média 2,04 mcg/100ml e desvio padrão 0,78 mcg/100ml.

a) Determine a probabilidade de um cão macho sadio apresentar concentração sérica de tiroxina:

i Inferior a 2,81 mcg/100ml
ii Superior a 1,8 mcg/100ml
iii Entre 1,01 e 2,50 mcg/100ml

b) Se considerarmos 200 desses cães, quantos se poderia esperar que tivessem uma concentração sérica entre 2,20 e 3,80 mcg/100ml ?

c) Qual intervalo de valores, simétrico em torno da média, abrange 98% dos cães sadios ?

6.3) [1]Os prazos de duração de gravidez têm distribuição Gaussiana com média de 268 dias e desvio-padrão de 15 dias. Definindo como prematura uma criança que nascer com menos de 247 dias de gestação, responda :

a) Qual é a porcentagem de crianças nascidas prematuramente?

b) Se desejássemos mudar a definição de uma criança prematura como sendo "aquela cujo o período de gestação está entre os 4% menores", qual seria o tempo mínimo de gestação para que uma criança não fosse considerada prematura?

6.4) [1]Os escores de Q.I. têm distribuição Gaussiana com média 100 pontos e desvio-padrão 15 pontos. Uma organização só admite pessoas de Q.I. elevado, para ela, maiores do que 131,5 . Com base nessas informações, responda:

a) Escolhida uma pessoa aleatoriamente, qual é a probabilidade de que ela seja admita por essa organização?

b) Definindo como "gênio" uma pessoa que tenha seu escore de Q.I. situado entre os 1% mais altos, qual seria o valor para o escore de Q.I. que separaria os "gênios" das pessoas comuns?

6.5) Considere o problema e os dados apresentados no exercício 2.2.5 (Seção 2.2)

a) Verifique a suposição de normalidade destes dados.

b) Determine uma faixa de normalidade de 95% para o número de colônias de bactérias no trato respiratório de pessoas sadias, usando:

   i. Método dos Percentis (já construída no exercício 2.2.5)
   ii. Método da Curva de Gauss

c) Qual dos dois métodos é o mais indicado nesse caso? Justifique.

[1] Exercício adaptado de Triola, M. F. (1999) - *Introdução à Estatística* - 7ª Edição - LTC

Seção 7: Intervalos de Confiança

7.1) Num estudo para descrever o perfil dos pacientes adultos atendidos no ambulatório de um posto de saúde, uma amostra de 70 pacientes adultos foi selecionada ao acaso entre o total de pacientes atendidos no posto durante os últimos três anos, coletando-se dos prontuários desses pacientes dados relativos à idade, à escolaridade e a outros fatores de interesse.

Para a variável idade, observou-se uma média amostral de 36,86 anos com um desvio padrão amostral de 17,79 anos. Para a variável escolaridade, observou-se que 19 pacientes da amostra eram analfabetos.

a) Defina a população e a amostra.

b) Forneça uma estimativa pontual, um intervalo de 90% de confiança e um intervalo de 95% de confiança para a idade média dos adultos atendidos neste ambulatório nos últimos três anos. Interprete e compare os intervalos de confiança.

c) Forneça uma estimativa pontual, um intervalo de 90% de confiança e um intervalo de 95% de confiança para proporção de analfabetos dentre os adultos atendidos neste ambulatório nos últimos três anos. Interprete e compare os intervalos de confiança.

7.2) A produção de leite na primeira lactação foi medida em 20 vacas selecionadas aleatoriamente dentre as vacas de uma fazenda. A produção média nesta amostra de 1500 litros e o desvio padrão de 300 litros. Construa e interprete um intervalo de 98% de confiança para a produção média de leite na primeira lactação das vacas dessa fazenda.

7.3) Considere o Intervalo de 100(1-α)% Confiança para a média (μ) de uma variável com distribuição Normal e desvio-padrão (σ) conhecido :

$$IC_{\mu}^{100(1-\alpha)}\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} ; \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

Note que o intervalo é simétrico em torno da média amostral $\bar{x}$ e que, quanto menor a parcela $z_{\alpha/2} \cdot \sigma/\sqrt{n}$, menor será a amplitude do intervalo.

a) Por que um intervalo com grande amplitude não é útil para o pesquisador fazer inferência sobre a média da variável estudada ?

b) Mantendo fixos o tamanho de amostra (n) e o desvio-padrão (σ), o que aconteceria com esta parcela se reduzirmos o nível de confiança (100(1-α)) requerido para o intervalo ?

c) Mantendo fixos o tamanho de amostra (n) e o nível de confiança (100(1-α)) requerido para o intervalo, o que aconteceria com esta parcela se o desvio-padrão (σ) fosse menor ?

d) Mantendo fixos o desvio-padrão (σ) e o nível de confiança (100(1-α)) requerido para o intervalo, o que aconteceria com esta parcela se o tamanho de amostra (n) fosse maior ?

e) Qual(is) dos três elementos que compõem esta parcela (n, σ e α), o pesquisador é capaz de alterar para conseguir um intervalo mais curto para a média da variável estudada ?

f) Caso o pesquisador queira um intervalo curto e com nível de confiança alto (ex. 98%, 99%), o que ele deveria fazer ?

Seção 8: Conceitos Básicos de Testes de Hipóteses

Para cada uma das situações abaixo, identifique:

a) O parâmetro que está sendo testado (média ou proporção);
b) As hipóteses nula e alternativa;
c) Os erros Tipo I e Tipo II.

Situação 1 - Um método padrão para identificação de bactérias em hemoculturas vem sendo utilizado há muitos anos e seu tempo médio de execução (desde a etapa de preparo das amostras até a identificação do gênero e espécie) é de 40,5 horas. Um microbiologista propôs uma nova técnica que ele afirma ter menor tempo de execução que o método padrão. A nova técnica foi aplicada em uma amostra de 18 hemoculturas e para cada uma mediu-se o tempo de execução. A média amostral foi 39,42 horas e o desvio padrão amostral foi 1,96 horas.

Situação 2 - Estudos sobre mortalidade de homens com idade superior a 65 anos de uma cidade mostram que 4% deles morrem dentro de um ano. Num grupo de 1000 indivíduos selecionados dessa população, 60 morreram no período de um ano. Suspeita-se de que houve um aumento da mortalidade anual nessa população.

Situação 3 - Um restaurante compra frangos abatidos inteiros com peso médio de 3 quilos há vários anos de um fornecedor. Outro fornecedor propõe ao gerente do restaurante vender frangos com peso médio maior que 3 quilos ao mesmo preço do fornecedor antigo. Antes de mudar de fornecedor, o gerente do restaurante decidiu comprar 25 frangos do novo fornecedor e pesá-los. Encontrou um peso médio de 3,2 quilos com um desvio padrão de 0,4 quilos.

Situação 4 - Uma indústria farmacêutica especifica que em certo analgésico a quantidade média de ácido acetil salicílico deve ser 5,5 mg por comprimido. A indústria suspeita que houve problemas na produção de um determinado lote e que, nesse lote, a quantidade média dessa substância está diferente da especificada. Para verificar essa suspeita, a indústria selecionou uma amostra aleatória de 40 comprimidos desse lote, observando uma quantidade média de ácido acetil salicílico igual a 5,2 mg e um desvio padrão de 0,7 mg.

Situação 5 - Um vendedor de sementes de milho garantiu a um agricultor que a proporção de sementes de sua marca que realmente chegam a germinar é 95%. O agricultor desconfia que na verdade esta proporção é menor do que a anunciada pelo vendedor. Antes de efetuar uma grande compra, o agricultor comprou um pacote com 1000 sementes e plantou, observando mais tarde que 940 sementes germinaram.

Seção 9: Testes de Hipóteses para Uma População

Para cada uma das situações descritas na Seção 8, responda as questões abaixo, utilizando:

a) Método Tradicional de Teste de Hipóteses ;
b) Método do Valor P;
c) Método do Intervalo de Confiança, quando for adequado.

Utilize um nível de significância α=0,05.

Situação 1 - A nova técnica reduz o tempo para identificação de bactérias?

Situação 2 - Existe evidência de que houve um aumento da mortalidade anual nesta população?

Situação 3 - A afirmação do novo fornecedor é confirmada pelos dados coletados pelo gerente?

Situação 4 - Os dados confirmam a suspeita da indústria?

Situação 5 - O resultado do experimento do agricultor confirma sua desconfiança?

Seção 10: Testes de Hipóteses para Duas Populações

10.1) Como resultado de um programa de fortificação isométrica desenvolvido em 10 semanas, alunos da oitava série foram avaliados em duas ocasiões, antes e após o programa, quanto a sua habilidade em executar abdominais em dois minutos. Os dados são apresentados a tabela abaixo. Quanto maior o escore, maior é a habilidade do aluno em executar abdominais de dois minutos.

Tabela 10.1

No do aluno	Escores de abdominais		
	Antes	Depois	
1	12	15	
2	10	9	
3	23	25	
4	25	25	
5	29	31	
6	32	30	
7	14	16	
8	17	20	
9	19	22	
10	20	22	

a) Faça um teste de hipóteses (ao nível de significância α = 5%) para verificar se o programa de fortificação isométrica aumenta a habilidade em executar abdominais em dois minutos.
(Dica: tome a diferença de "depois-antes" e teste se $\mu_{depois} > \mu_{antes}$*)*

b) Calcule o *valor p* do teste de hipóteses acima. Como você chegaria à conclusão do teste do item (a) usando a informação do *valor p* ?

c) Construa e interprete um Intervalo de 95% de Confiança para a diferença entre a habilidade média depois do programa e a habilidade média antes do programa.

10.2) *Em um experimento, dois grupos de ratos fêmeas foram alimentados com dietas apresentando alto e baixo conteúdo de proteína. O quadro abaixo fornece, para cada rato, o ganho de peso, em gramas, entre o 28º e o 84º dia de vida.

Conteúdo de proteína	Ganho de peso
Alto	123 134 146 104 119 124 161 107 83 113 129 97
Baixo	70 118 101 85 107 132 94

a) Ao nível de significância de 1%, há evidência estatística de que a dieta com alto conteúdo de proteína aumenta o ganho de peso?

b) Calcule o valor P e use-o para responder à pergunta do item (a).

c) Construa e interprete um intervalo de 95% de confiança para a diferença entre os ganhos médios de peso com as dietas de alto e baixo conteúdo de proteína.

10.3) *Em um estudo sobre a influência do uso de cocaína no peso de crianças nascidas de mães dependentes, pesquisadores trabalharam com dois grupos de crianças nascidas a termo: o primeiro grupo era composto de mães que usaram regularmente a droga durante toda a gravidez (Grupo I) e o segundo, de mães que não tinham história ou evidência de uso de cocaína (Grupo II). A hipótese dos pesquisadores era de que o peso médio de crianças de mães dependentes é menor do que o peso médio de crianças de mães não-dependentes. Os resultados são apresentados abaixo:.

Grupo	Tamanho da amostra	Peso médio (g)	Desvio padrão (g)
I	36	2829	708
II	39	3436	628

Fonte: Chanoff, I. J. et al. (1989) "Temporal patterns of cocaine use in pregnancy - perinatal outcome", JAMA, março.

a) Usando um nível de significância igual a 5%, teste a hipótese dos pesquisadores. (Para isso, estabeleças as hipóteses nula e alternativa adequadas, construa a região de rejeição, calcule o valor da estatística de teste e conclua).

b) Calcule o Valor P do teste. Use o Valor P obtido para responder à questão formulada no enunciado do problema, com nível de significância de 1%.

10.4) *Em um estudo publicado no "Canadian Medical Association Journal" em novembro de 1972, procurou-se investigar o efeito do uso da vitamina C na prevenção de resfriados. Para isso, realizou-se o seguinte experimento: por um determinado período de tempo, 407 indivíduos tomaram fortes doses de vitamina C e 411 receberam placebo. No grupo da vitamina, 105 participantes ficaram livres de doenças do trato respiratório, enquanto, no grupo placebo, esse número foi de apenas 76 participantes. O que os pesquisadores puderam concluir? (Use α = 0,05 e calcule o Valor P)

10.5) Em um estudo para determinar se o tratamento com estrogênio e progestina altera o risco de eventos de doença coronariana (DC) em mulheres na pós-menopausa com doença coronariana estabelecida, um total de 2.763 destas mulheres foram divididas aleatoriamente em dois grupos: 1.380 fizeram uso desses hormônios e o restante fez uso de um placebo.

A Figura 11.1 mostra os resultados publicados na edição brasileira do *Journal of American Medical Association* em setembro de 1999. Para cada desfecho considerado nas linhas da tabela, corresponde, na última coluna, o Valor P do teste:

H_0: O desfecho não está associado ao uso dos hormônios;
H_1: O desfecho está associado ao uso dos hormônios.

Analisando o Valor P, quais os desfechos podem ser considerados associados ao uso de reposição hormonal, ao nível de significância de 5% ?

Figura 10.1

Tabela 4. Óbito e Desfechos não Cardiovasculares Secundários por Grupo de Tratamento*

	Grupo de Tratamento			
Desfechos	**Estrogênio-Progestina (N = 1.380)**	**Placebo (N = 1.383)**	**RR (IC de 95%)**	**Valor de *P***
Óbito				
Óbito por DC	71	58	1,24 (0,87-1,75)	0,23
Óbito por câncer	19	24	0,80 (0,44-1,46)	0,47
Óbito não-DC, não-câncer	37	36	1,04 (0,66-1,64)	0,87
Óbito não julgado	4	5	...	...
Total de óbitos	131	123	1,08 (0,84-1,38)	0,56
Evento tromboembólico venoso				
Trombose venosa profunda	25	8	3,18 (1,43-7,04)	0,004
Embolismo pulmonar	11	4	2,79 (0,89-8,75)	0,08
Qualquer evento tromboembólico	34	12	2,89 (1,50-5,58)	0,002
Câncer				
Mama	32	25	1,30 (0,77-2,19)	0,33
Endometrial	2	4	0,49 (0,09-2,68)	0,41
Outro	63	58	1,10 (0,77-1,57)	0,60
Qualquer tipo de câncer	96	87	1,12 (0,84-1,50)	0,44
Fratura				
Quadril	12	11	1,10 (0,49-2,50)	0,82
Outra	119	129	0,93 (0,73-1,20)	0,59
Qualquer tipo de fratura	130	138	0,95 (0,75-1,21)	0,70
Doença da vesícula biliar	84	62	1,38 (1,00-1,92)	0,05

*RR indica risco relativo; IC, intervalo de confiança e DC, doença coronariana. Cada coluna representa o número de mulheres com o evento designado; as mulheres com mais de um tipo de evento podem aparecer em mais de uma coluna.

Fonte: JAMABrasil, set 1999, v.3, n. 8

Seção 11: Teste Qui-Quadrado

11.1) Desejando verificar se duas vacinas contra brucelose* (uma padrão e um nova) são igualmente eficazes, pesquisadores realizaram o seguinte experimento: um grupo de 14 bezerras tomou a vacina padrão e outro grupo de 16 bezerras tomou a vacina nova. Considerando que os dois grupos estavam igualmente expostos ao risco de contrair a doença, após algum tempo, verificou-se quantos animais, em cada grupo, havia contraído a doença. Do grupo que tomou a vacina nova, 5 animais ficaram doentes. No outro grupo, 10 animais ficaram doentes.

Construa uma tabela para organizar os dados descritos acima e responda:

Existe diferença estatisticamente significativa entre as proporções de bezerras que contraíram brucelose usando a nova vacina e a vacina padrão? (Use o teste Qui-Quadrado, com nível de significância de 5%, e calcule o valor P).

*A brucelose é uma doença crônica causada pela bactéria *Brucella*, cuja principal espécie atinge os rebanhos bovino, suíno, ovino, e caprino. Incurável, leva as fêmeas a abortarem seus filhotes e provoca, nos machos, inflamações nos testículos. Também atinge o homem, causando febre, lesões articulares e insônia.

11.2) *Com o objetivo de examinar a existência do efeito de determinado fertilizante na incidência da *Bacterium phithotherum* em plantação de batatas, foi realizado o seguinte experimento: pés de batata tratados com diferentes fertilizantes foram classificados, ao final do estudo, como contaminados ou livres de contaminação (Tabela 11.2).

Tabela 11.2

Fertilizante	Contaminação		Total
	Sim	Não	
Nenhum	16	85	101
Nitrogênio	10	85	95
Esterco	4	109	113
Nitrogênio e esterco	14	127	141
Total	44	406	450

Existe efeito de fertilizante na incidência desse tipo de bactéria nas plantações de batata? (Use o teste Qui-Quadrado, com nível de significância de 5%, e calcule o valor P).

11.3) Pesquisadores de doenças parasitárias em animais de grande porte suspeitam que a incidência de certos parasitas esteja associada à raça do animal, pura ou não-pura. Num estudo para verificar essa suspeita, 1355 animais selecionados aleatoriamente de uma grande fazenda foram classificados segundo sua raça e incidência de parasitose (berne). Dos 805 animais de raça pura, 105 tiveram parasitose. Nos 550 animais restantes (raça não-pura), 50 tiveram parasitose.

Construa uma tabela para organizar os dados descritos acima e responda:

Existem evidências estatísticas suficientes nesses dados para verificar a hipótese de que a raça do animal e a incidência de parasitas estejam associadas? (Use nível de significância igual a 1% e calcule o valor P).

11.4) Em estudo da associação entre a ocorrência de doença coronariana e grupo sanguíneo, voluntários do grupo A (450), do grupo B (450), do grupo AB (300) e do grupo O (600) foram acompanhados por vários anos. Ao final do período do estudo, foram registrados quantos voluntários em cada grupo desenvolveram doença coronariana. (Tabela 11.4).** Existem evidências estatísticas suficientes para a hipótese de que a ocorrência de doença coronariana e o grupo sanguíneo estejam associados? (Use nível de significância igual a 1% e calcule o valor P).

Tabela 11.4

Grupo Sangüíneo	Doença Coronariana		Total
	Sim	Não	
A	9	441	450
B	12	438	450
AB	9	291	300
O	12	588	600
Total	42	1758	1800

** Reprodução baseada em resultados do trabalho de He et al. (2012) *ABO Blood Group and Risk of Coronary Heart Disease in Two Prospective Cohort Studies*, Arteriosclerosis, Thrombosis, and Vascular Biology, v. 32: pp. 2314-2320

11.5) Para cada um dos problemas anteriores, calcule a medida de associação mais adequada (Risco Relativo ou Razão de Chances) e justifique sua escolha.

Intervalos de Confiança

Uma população

$$IC_{p}^{100(1-\alpha)}=\left[\hat{p}\pm z_{\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$IC_{\mu}^{100(1-\alpha)}=\left[\bar{x}\pm z_{\alpha/2}\cdot\frac{s}{\sqrt{n}}\right]$$

$$IC_{m}^{100(1-\alpha)}=\left[\bar{x}\pm t_{(n-1;\alpha/2)}\cdot\frac{s}{\sqrt{n}}\right]$$

Duas populações (amostras emparelhadas ou dependentes)

$$IC_{\mu_1-\mu_2}^{100(1-\alpha)}=\left[\bar{d}\pm z_{\alpha/2}\text{¿}\frac{s_d}{\sqrt{n}}\right]$$

$$IC_{m_1-m_2}^{100(1-\alpha)}=\left[\bar{d}\pm t_{(n-1;\alpha/2)}\text{¿}\frac{s_d}{\sqrt{n}}\right]$$

Duas populações (amostras independentes)

$$IC_{m_1-m_2}^{100(1-\alpha)}=\left[(\bar{x}-\bar{y})\pm z_{\alpha/2}\text{¿}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]$$

$$IC_{m_1-m_2}^{100(1-\alpha)}=\left[(\bar{x}-\bar{y})\pm t_{(n_1+n_2-2;\alpha/2)}\text{¿}\sqrt{s_{comb}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]$$

onde $s_{comb}^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

$$IC_{p_1-p_2}^{100(1-\alpha)}=\left[(\hat{p}_1-\hat{p}_2)\pm z_{\alpha/2}\text{¿}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

Testes de Hipóteses

Teste de Hipóteses para uma Média (amostra grande)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
H_0: $\mu = \mu_0$ H_a: $\mu < \mu_0$		$Z_{obs} < -Z_{\alpha}$	$P(Z < Z_{obs})$
H_0: $\mu = \mu_0$ H_a: $\mu > \mu_0$	$Z_{obs} = \frac{x - \mu_o}{s / \sqrt{n}}$	$Z_{obs} > Z_{\alpha}$	$P(Z > Z_{obs})$
H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$		$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	$2 \times P(Z > \lvert Z_{obs} \rvert)$

Teste de Hipóteses para uma Média (amostra pequena)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
H_0: $\mu = \mu_0$ H_a: $\mu < \mu_0$		$T_{obs} < -t_{(n-1;\alpha)}$	$P(T_{(n-1)} < T_{obs})$
H_0: $\mu = \mu_0$ H_a: $\mu > \mu_0$	$T_{obs} = \frac{x - \mu_o}{s / \sqrt{n}}$	$T_{obs} > t_{(n-1;\alpha)}$	$P(T_{(n-1)} > T_{obs})$
H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$		$T_{obs} < -t_{(n-1;\frac{\alpha}{2})}$ ou $T_{obs} > t_{(n-1;\frac{\alpha}{2})}$	$2 \times P(T_{(n-1)} > \lvert T_{obs} \rvert)$

Teste de Hipóteses para uma Proporção (amostra grande)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
H_0: $p = p_0$ H_a: $p < p_0$		$Z_{obs} < -Z_{\alpha}$	$P(Z < Z_{obs})$
H_0: $p = p_0$ H_a: $p > p_0$	$Z_{obs} = \frac{\hat{p} - p_o}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	$Z_{obs} > Z_{\alpha}$	$P(Z > Z_{obs})$
H_0: $p = p_0$ H_a: $p \neq p_0$		$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	$2 \times P(Z > \lvert Z_{obs} \rvert)$

Teste de Hipóteses para duas Médias (amostra emparelhada e grande)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
H_0: $\mu_D = 0$ H_a: $\mu_D < 0$		$Z_{obs} < -Z_{\alpha}$	$P(Z < Z_{obs})$
H_0: $\mu_D = 0$ H_a: $\mu_D > 0$	$Z_{obs} = \frac{d}{s_d / \sqrt{n}}$	$Z_{obs} > Z_{\alpha}$	$P(Z > Z_{obs})$
H_0: $\mu_D = 0$ H_a: $\mu_D \neq 0$		$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	$2 \times P(Z > \lvert Z_{obs} \rvert)$

Teste de Hipóteses para duas Médias (amostra emparelhada e pequena)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
H_0: $\mu_D = 0$ H_a: $\mu_D < 0$		$T_{obs} < -t_{(n-1;\alpha)}$	$P(T_{(n-1)} < T_{obs})$
H_0: $\mu_D = 0$ H_a: $\mu_D > 0$	$T_{obs} = \frac{d}{s_d / \sqrt{n}}$	$T_{obs} > t_{(n-1;\alpha)}$	$P(T_{(n-1)} > T_{obs})$
H_0: $\mu_D = 0$ H_a: $\mu_D \neq 0$		$T_{obs} < -t_{(n-1;\frac{\alpha}{2})}$ ou $T_{obs} > t_{(n-1;\frac{\alpha}{2})}$	$2 \times P(T_{(n-1)} > \lvert T_{obs} \rvert)$

Teste de Hipóteses para duas Médias (amostras independentes e grandes)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 < \mu_2$	$Z_{obs} = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	$Z_{obs} < -Z_{\alpha}$	$P(Z < Z_{obs})$
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 > \mu_2$		$Z_{obs} > Z_{\alpha}$	$P(Z > Z_{obs})$
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$		$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	$2 \times P(Z > \lvert Z_{obs} \rvert)$

Teste de Hipóteses para duas Médias (amostras independentes e pequenas)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 < \mu_2$	$T_{obs} = \dfrac{\bar{x} - \bar{y}}{\sqrt{s_{comb}^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}}$	$T_{obs} < -t_{(gl;\alpha)}$	$P(T_{gl} < T_{obs})$
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 > \mu_2$		$T_{obs} > t_{(gl;\alpha)}$	$P(T_{gl} > T_{obs})$
$H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$	$s_{comb}^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$T_{obs} < -t_{(gl;\frac{\alpha}{2})}$ ou $T_{obs} > t_{(gl;\frac{\alpha}{2})}$	$2 \times P(T_{gl} > \lvert T_{obs} \rvert)$

$gl = n_1 + n_2 - 2$

Teste de Hipóteses para duas Proporções (amostras independentes e grandes)

Hipóteses	Estatística de Teste	Região de Rejeição de H_0	Valor P
$H_0: p_1 = p_2$ $H_a: p_1 < p_2$	$Z_{obs} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	$Z_{obs} < -Z_{\alpha}$	$P(Z < Z_{obs})$
$H_0: p_1 = p_2$ $H_a: p_1 > p_2$		$Z_{obs} > Z_{\alpha}$	$P(Z > Z_{obs})$
$H_0: p_1 = p_2$ $H_a: p_1 \neq p_2$		$Z_{obs} < -Z_{\frac{\alpha}{2}}$ ou $Z_{obs} > Z_{\frac{\alpha}{2}}$	$2 \times P(Z > \lvert Z_{obs} \rvert)$

Tabela de Poisson (acumulada)

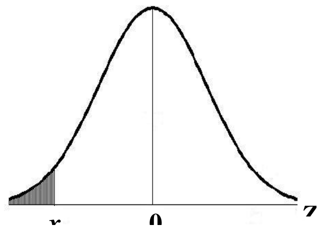
$$P(X \le x) = \sum_{r=0}^{x} \lambda^{r} \frac{e^{-\lambda}}{r!}$$

$\lambda=$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.4	1.6	1.8
$x=$ 0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3012	0.2466	0.2019	0.1653
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6626	0.5918	0.5249	0.4628
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.8795	0.8335	0.7834	0.7306
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9662	0.9463	0.9212	0.8913
4	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9923	0.9857	0.9763	0.9636
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9985	0.9968	0.9940	0.9896
6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994	0.9987	0.9974
7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9994
8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$\lambda=$	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.5	5.0	5.5
$x=$ 0	0.1353	0.1108	0.0907	0.0743	0.0608	0.0498	0.0408	0.0334	0.0273	0.0224	0.0183	0.0111	0.0067	0.0041
1	0.4060	0.3546	0.3084	0.2674	0.2311	0.1991	0.1712	0.1468	0.1257	0.1074	0.0916	0.0611	0.0404	0.0266
2	0.6767	0.6227	0.5697	0.5184	0.4695	0.4232	0.3799	0.3397	0.3027	0.2689	0.2381	0.1736	0.1247	0.0884
3	0.8571	0.8194	0.7787	0.7360	0.6919	0.6472	0.6025	0.5584	0.5152	0.4735	0.4335	0.3423	0.2650	0.2017
4	0.9473	0.9275	0.9041	0.8774	0.8477	0.8153	0.7806	0.7442	0.7064	0.6678	0.6288	0.5321	0.4405	0.3575
5	0.9834	0.9751	0.9643	0.9510	0.9349	0.9161	0.8946	0.8705	0.8441	0.8156	0.7851	0.7029	0.6160	0.5289
6	0.9955	0.9925	0.9884	0.9828	0.9756	0.9665	0.9554	0.9421	0.9267	0.9091	0.8893	0.8311	0.7622	0.6860
7	0.9989	0.9980	0.9967	0.9947	0.9919	0.9881	0.9832	0.9769	0.9692	0.9599	0.9489	0.9134	0.8666	0.8095
8	0.9998	0.9995	0.9991	0.9985	0.9976	0.9962	0.9943	0.9917	0.9883	0.9840	0.9786	0.9597	0.9319	0.8944
9	1.0000	0.9999	0.9998	0.9996	0.9993	0.9989	0.9982	0.9973	0.9960	0.9942	0.9919	0.9829	0.9682	0.9462
10	1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9995	0.9992	0.9987	0.9981	0.9972	0.9933	0.9863	0.9747
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9996	0.9994	0.9991	0.9976	0.9945	0.9890
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9992	0.9980	0.9955
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9993	0.9983
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

$\lambda=$	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	11.0	10.0	12.0	14.0	15.0
$x=$ 0	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008	0.0005	0.0002	0.0005	0.0001	0.0000	0.0000
2	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042	0.0028	0.0012	0.0028	0.0005	0.0001	0.0000
3	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149	0.0103	0.0049	0.0103	0.0023	0.0005	0.0002
4	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403	0.0293	0.0151	0.0293	0.0076	0.0018	0.0009
5	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885	0.0671	0.0375	0.0671	0.0203	0.0055	0.0028
6	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649	0.1301	0.0786	0.1301	0.0458	0.0142	0.0076
7	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687	0.2202	0.1432	0.2202	0.0895	0.0316	0.0180
8	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918	0.3328	0.2320	0.3328	0.1550	0.0621	0.0374
9	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218	0.4579	0.3405	0.4579	0.2424	0.1094	0.0699
10	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453	0.5830	0.4599	0.5830	0.3472	0.1757	0.1185
11	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520	0.6968	0.5793	0.6968	0.4616	0.2600	0.1848
12	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364	0.7916	0.6887	0.7916	0.5760	0.3585	0.2676
13	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981	0.8645	0.7813	0.8645	0.6815	0.4644	0.3632
14	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400	0.9165	0.8540	0.9165	0.7720	0.5704	0.4657
15	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665	0.9513	0.9074	0.9513	0.8444	0.6694	0.5681
16	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823	0.9730	0.9441	0.9730	0.8987	0.7559	0.6641
17	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911	0.9857	0.9678	0.9857	0.9370	0.8272	0.7489
18	1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957	0.9928	0.9823	0.9928	0.9626	0.8826	0.8195
19	1.0000	1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980	0.9965	0.9907	0.9965	0.9787	0.9235	0.8752
20	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	0.9953	0.9984	0.9884	0.9521	0.9170
21	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9996	0.9993	0.9977	0.9993	0.9939	0.9712	0.9469
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9997	0.9990	0.9997	0.9970	0.9833	0.9673
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9995	0.9999	0.9985	0.9907	0.9805
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	1.0000	0.9993	0.9950	0.9888
25	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000	0.9997	0.9974	0.9938
26	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9987	0.9967
27	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9994	0.9983
28	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9991
29	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996
30	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998
31	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
32	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabela Normal Padrão: Normal ($\mu = 0$; $\sigma = 1$)



x	0	1	2	3	4	5	6	7	8	9
	Segunda casa decimal de *x*									
-2,9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
-2,8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
-2,7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
-2,6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
-2,5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
-2,4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
-2,3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
-2,2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
-2,1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
-2,0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
-1,9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
-1,8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
-1,7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
-1,6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
-1,5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
-1,4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
-1,3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
-1,2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
-1,1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
-1,0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
-0,9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
-0,8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
-0,7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
-0,6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
-0,5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
-0,4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
-0,3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
-0,2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
-0,1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0,0		4960	4920	4880	4840	4801	4761	4721	4681	4641

x	0	1	2	3	4	5	6	7	8	9
	Segunda casa decimal de *x*									
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0,5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0,6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0,7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0,8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0,9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015
1,3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319
1,5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986

Obs: os números dentro da tabela referem-se as 4 quatro casas decimais de P[Z < *x*]. Ex: “0179” leia-se 0,0179 .

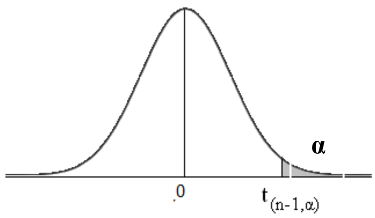


Tabela *t-Student*

g.l.	α										
	0.25	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Distribuição Qui-Quadrado

TABELA - Distribuição Qui-Quadrado

φ = graus de liberdade

φ \ α	0,995	0,99	0,975	0,95	0,90	0,75	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,001
1	0,0004	0,002	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879	10,828
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597	13,816
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838	16,266
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860	18,467
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,071	12,833	15,086	16,750	20,515
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548	22,458
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278	24,322
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955	26,125
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589	27,877
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188	29,588
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,920	24,725	26,757	31,264
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,299	32,909
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819	34,528
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319	36,123
15	4,601	5,229	6,262	7,261	8,547	11,036	14,339	18,245	22,307	24,996	27,488	30,578	32,801	37,697
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267	39,252
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718	40,790
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156	43,312
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582	43,820
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997	45,315
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401	46,797
22	8,643	9,542	10,982	12,338	14,042	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796	48,268
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181	49,728
24	9,886	10,856	12,401	13,848	15,659	19,037	22,337	28,241	33,196	36,415	39,364	42,980	45,559	51,179
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928	52,620
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,434	35,563	38,885	41,923	45,642	48,290	54,052
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,194	46,963	49,645	55,476
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,993	56,892
29	13,121	14,257	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,336	58,302
30	13,787	14,954	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672	59,703
31	14,458	15,655	17,539	19,281	21,434	25,390	30,336	35,887	41,422	44,985	48,232	52,191	55,003	61,098
32	15,134	16,362	18,291	20,072	22,271	26,304	31,336	36,973	42,585	46,194	49,480	53,486	56,328	62,487
33	15,815	17,074	19,047	20,867	23,110	27,219	32,336	38,058	43,745	47,400	50,725	54,776	57,648	63,870
34	16,501	17,789	19,806	21,664	23,952	28,136	33,336	39,141	44,903	48,602	51,966	56,061	58,964	65,247
35	17,192	18,509	20,569	22,465	24,797	29,054	34,336	40,223	46,059	49,802	53,203	57,342	60,275	66,619
36	17,887	19,233	21,336	23,269	25,643	29,973	35,336	41,304	47,212	50,998	54,437	58,619	61,581	67,985
37	18,586	19,960	22,106	24,075	26,492	30,893	36,336	42,383	48,363	52,192	55,668	59,892	62,883	69,346
38	19,289	20,691	22,878	24,884	27,343	31,815	37,335	43,462	49,513	53,384	56,896	61,162	64,181	70,701
39	19,996	21,426	23,654	25,695	28,196	32,737	38,335	44,539	50,660	54,572	58,120	62,428	65,476	72,055
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766	73,402
41	21,421	22,906	25,215	27,326	29,907	34,585	40,335	46,692	52,949	56,942	60,561	64,950	68,053	74,745
42	22,138	23,650	25,999	28,144	30,765	35,510	41,335	47,766	54,090	58,124	61,777	66,206	69,336	76,084
43	22,859	24,398	26,785	28,965	31,625	36,436	42,335	48,840	55,230	59,304	62,990	67,459	70,616	77,419
44	23,584	25,148	27,575	29,787	32,487	37,363	43,335	49,913	56,369	60,481	64,201	68,710	71,893	78,750
45	24,311	25,901	28,366	30,612	33,350	38,291	44,335	50,985	57,505	61,656	65,410	69,957	73,166	80,077
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490	86,661
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952	99,607
70	43,275	45,442	48,758	51,739	55,329	61,698	69,335	77,577	85,527	90,531	95,023	100,425	104,215	112,317
80	51,172	53,540	57,153	60,391	64,278	71,145	79,335	88,130	96,578	101,879	106,629	112,329	116,321	124,839
90	59,196	61,754	65,647	69,126	73,291	80,625	89,335	98,650	107,565	113,145	118,136	124,116	128,299	137,208
100	67,328	70,065	74,222	77,929	82,358	90,133	99,335	109,141	118,498	124,342	129,561	135,807	140,169	149,449

Segunda Parte:

Resolução dos Exercícios

Índice

Seção 1: Tipos de Estudos e Variáveis

Exercício (1.1)

a) Qualitativa e Nominal

b) Quantitativa e Discreta

c) Quantitativa e Contínua

d) Qualitativa e Nominal

e) Quantitativa e Contínua

f) Quantitativa e Discreta

Exercício (1.2)

a) Estudo do tipo experimental.

Variável resposta: Alguma medida da disfunção erétil.

Grupos comparados: - Grupo Tratamento, formado pelos pacientes que tomaram Viagra;
- Grupo Controle, formado pelos pacientes que tomaram placebo.

b) Estudo do tipo experimental.

Variável resposta: Desempenho dos alunos na realização das tarefas propostas.

Grupos comparados:
- Grupo Tratamento, alunos que observavam um conjunto de fotografias de filhotes de cães e gatos antes das tarefas;
- Grupo Controle, alunos que observavam um conjunto de fotografias neutras.

c) Estudo do tipo observacional.

Variável resposta: Ocorrência de algum tumor não dermatológico.

Grupos comparados: - Pessoas que tiveram câncer dermatológico;
- Pessoas que não tiveram câncer dermatológico.

d) Estudo do tipo experimental, se a mutação no gene foi feita no laboratório;
Estudo do tipo observacional, caso contrário.

Variável resposta: Ganho de peso

Grupos comparados: - Ratos com mutação no tal gene;
- Ratos sem mutação no tal gene.

e) Estudo do tipo observacional.

Para a variável resposta "ocorrência de derrame isquêmico", temos a comparação entre os seguintes grupos:

- Pessoas com nível de colesterol acima de 280;
- Pessoas com nível de colesterol abaixo de 280.

Para a variável resposta "ocorrência de derrame hemorrágico" temos a comparação entre os seguintes grupos:

- Pessoas com nível de colesterol abaixo de 180;
- Pessoas com nível de colesterol acima de 180.

f) Estudo do tipo observacional.

Variável resposta: Presença da marca da vacina BCG

Grupos comparados:
- Casos: crianças que receberam tratamento contra tuberculose;
- Controles: crianças que receberam outro tratamento diferente do tratamento contra tuberculose.

g) Estudo do tipo observacional.

Variável resposta: Nível de testosterona

Grupos comparados:
- Torcedores de futebol e basquete quando seus times ganharam;
- Torcedores de futebol e basquete quando seus times perderam;

h) Estudo do tipo observacional.

Variável resposta: Ocorrência de problemas respiratórios e/ou alergias.

Grupos comparados:
- Crianças que tinham contato com animais;
- Crianças que viviam na zona urbana.

Seção 2.1: Análise Descritiva e Exploratória de Dados

Exercício (2.1.1)

Tabela 2.1: Idade dos alunos do curso de inglês para 150 estudantes da escola A

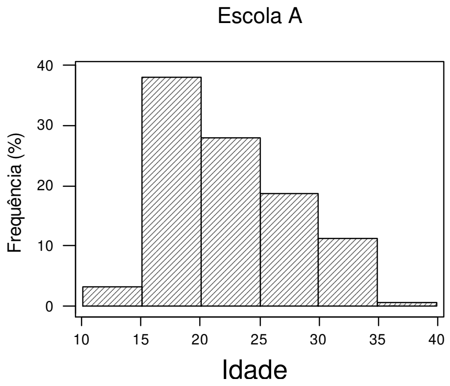
Idade	Ponto Médio	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Absoluta Acumulada	Freqüência Relativa Acumulada (%)
10 |- 15	12,5	5	3,33	5	3,33
15 |- 20	17,5	57	38,00	62	41,33
20 |- 25	22,5	42	28,00	104	69,33
25 |- 30	27,5	28	18,67	132	88,00
30 |- 35	32,5	17	11,33	149	99,33
35 |- 40	37,5	1	0,67	150	100,00
Total		150	100,00	------	------

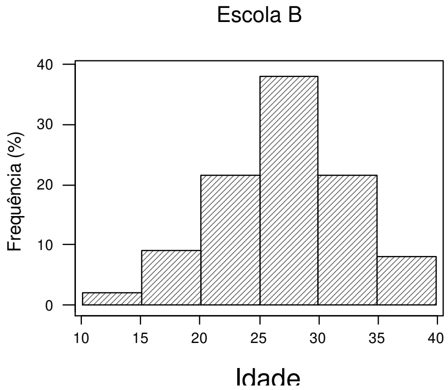
Tabela 2.2: Idade dos alunos do curso de inglês para 200 estudantes da escola B

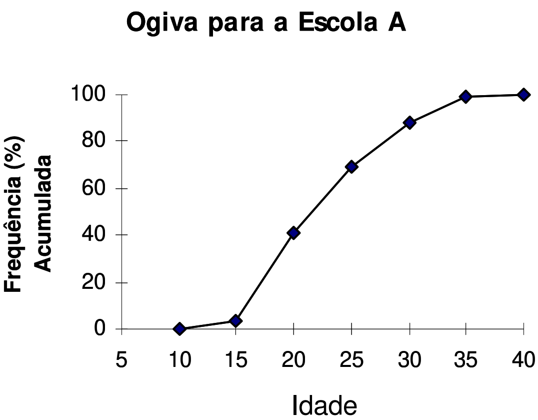
Idade	Ponto Médio	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Absoluta Acumulada	Freqüência Relativa Acumulada (%)
10 |- 15	12,5	4	2,00	4	2,00
15 |- 20	17,5	18	9,00	22	11,00
20 |- 25	22,5	43	21,50	65	32,50
25 |- 30	27,5	76	38,00	141	70,50
30 |- 35	32,5	43	21,50	184	92,00
35 |- 40	37,5	16	8,00	200	100,00
Total		200	100,00	------	-----

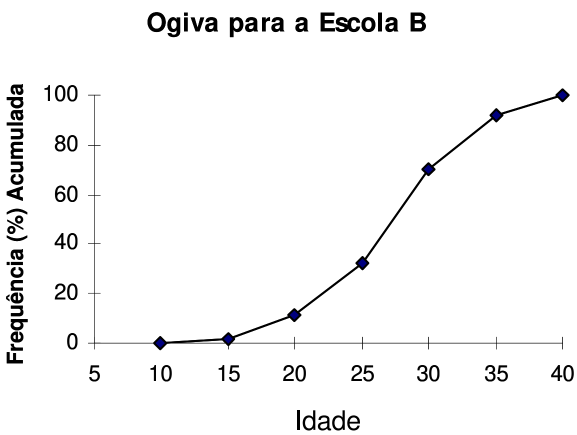
<u>Observação:</u> a soma da coluna "Frequência Relativa (%)" tem que ser 100,00. No entanto, por causa de arredondamentos, algumas vezes os valores dessa coluna somarão 100,01 (ou 99,99). Para corrigir o problema, podemos subtrair (ou somar) 0,01 à frequência relativa da classe com maior frequência. A mesma observação vale para quando estivermos usando somente uma casa decimal após a vírgula. Nesse caso, o total da coluna "Frequência Relativa (%)" tem que ser 100,0 e, se ocorrer o problema, a correção pode ser feita subtraindo-se (ou somando-se) 0,1 à frequência relativa da classe com maior frequência.

Como podemos notar pela análise dos histogramas a seguir, a distribuição de frequência da idade dos alunos do curso de inglês da Escola A é assimétrica com concentração à esquerda, enquanto a distribuição de frequência da idade dos alunos do curso de inglês da Escola B é razoavelmente simétrica em torno da classe 25 a 30 erros. A assimetria com concentração à esquerda dos dados da Escola A também pode ser percebida através da sua ogiva, que "cresce" mais rápido do que a ogiva da Escola B, que tem distribuição simétrica. Através desses gráficos, podemos observar que os alunos da Escola A tendem a ser mais jovens do que os alunos a Escola B.









Observação 1: Como gostaríamos de comparar as duas escolas, os histogramas e as ogivas foram feitos usando a frequência relativa, pois os tamanhos das amostras são diferentes. Se os tamanhos das amostras fossem iguais, a frequência absoluta poderia ser usada, embora o uso da frequência relativa torne os gráficos mais, digamos, úteis, pois eles poderão ser comparados com outros que também usem a frequência relativa, mesmo que as amostras sejam de tamanhos diferentes.

Observação 2: O primeiro ponto na ogiva deve ser o limite inferior da primeira classe, que corresponde à frequência acumulada igual a 0, ou seja, abaixo desse limite não existe nenhum dado.

Exercício (2.1.2)

a)

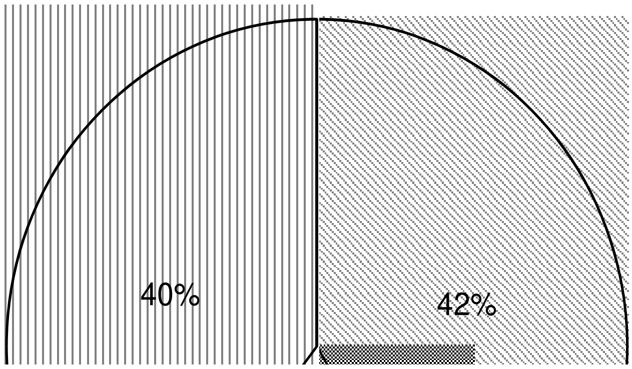
Tabela 2.3

Grupo Sangüíneo	Tromboembolismo		Total
	Doente	Sadia	
A	32	51	83
B	8	19	27
AB	6	5	11
O	9	70	79
Total	55	145	200

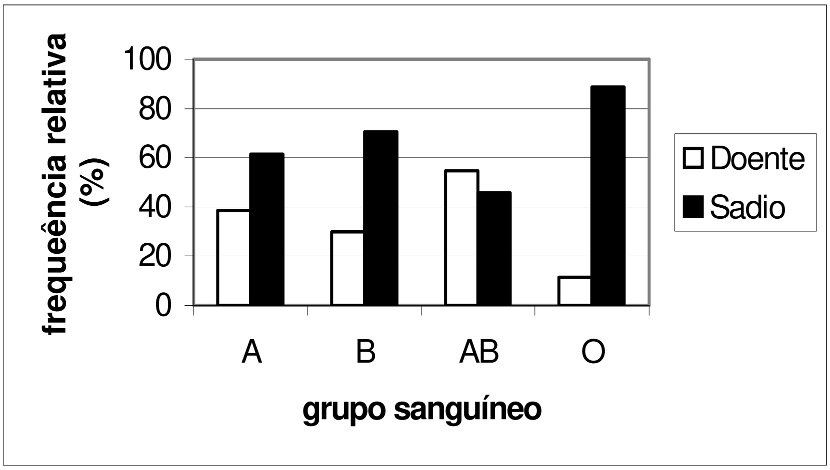
Percebe-se que grande parte das pessoas nesse estudo é sadia (72,5%).

Como as variáveis em questão (presença de tromboembolismo e grupo sanguíneo) são do tipo qualitativas, poderemos usar os gráficos de colunas (barras) ou de setores (torta) para representar esses dados.

O gráfico abaixo revela que a maioria dos pacientes da amostra tem o sangue do tipo A, seguido pelas pessoas de sangue tipo O e tipo B, sendo que a minoria tem o sangue do tipo AB.



O gráfico a seguir representa a frequência relativa da doença dentro de cada grupo sanguíneo. A frequência relativa tem que ser usada porque os grupos de sangue têm tamanhos diferentes. Usamos o gráfico de colunas por ser de melhor visualização, nesse caso em que temos 4 categorias (os tipos sanguíneos), ao invés de usarmos quatro gráficos de setores.



Pela tabela do exercício, temos que os pacientes sadios representam a maioria (72,5%) dos 200 pacientes amostrados. No gráfico acima, podemos observar que, nos grupos sanguíneos A, B e O, essa maioria se mantém, embora em proporções diferentes (no grupo O, por exemplo, em torno de 90% das pessoas são sadias). No entanto, no grupo AB, essas proporções se invertem, sendo as pessoas doentes mais frequentes. Isso pode ser um indício de que a distribuição de frequências do estado do paciente depende do tipo sanguíneo, ou seja, de que as variáveis tipo sanguíneo e presença de tromboembolismo podem estar associadas. Esses conceitos serão mais bem elaborados mais adiante, quando tratarmos de *independência de eventos* e *associação de variáveis*.

Exercício (2.1.3)

Os valores de pressão intraocular têm distribuição de frequências assimétrica com concentração à esquerda. Isso significa que a maioria das pessoas avaliadas possui valores pequenos para a pressão intraocular e algumas poucas pessoas possuem valores mais altos.

Já os valores de ácido úrico sérico têm distribuição de frequências que pode ser considerada simétrica em torno dos valores de 5,3 a 5,9 mg/100ml. Isto é, a maioria das pessoas possui medidas de ácido úrico sérico em torno desses valores, que podem ser considerados típicos. Algumas poucas pessoas possuem valores mais altos (cauda direita do histograma) e outras poucas pessoas possuem valores mais baixos (cauda esquerda). Essa questão dos valores que podem ser considerados típicos será mais bem discutida mais adiante quando tratarmos do conceito e construção das *Faixas de Referência.*

Exercício (2.1.4)

a) Diagrama de Pontos para os Tempos entre a Remissão e a Recidiva de uma Doença (em meses)

```
   :    .           .
: .:    :. :.  .    :.  :     :  .
-----+---------+---------+---------+---------+---------+- Homens

        ::
: :: :: ::  : :.          .
-----+---------+---------+---------+---------+---------+- Mulheres
    4,0       8,0      12,0      16,0      20,0      24,0
```

O grupo de mulheres tem, em geral, tempos menores entre a remissão e a recidiva dessa doença se comparado ao grupo dos homens , sendo esses tempos mais homogêneos no grupo feminino do que no grupo masculino.

b)

Ramo-e-folhas usando escala de 10 meses

```
Homens                    Mulheres

0 | 2234444777899         0 | 223344556677778888
1 | 02555688               1 | 001128
2 | 224
              Legenda: 0|2 : leia-se 2 meses
```

Ramo-e-folhas usando escala de 5 meses

```
Homens               Mulheres

 0 | 2234444         0 | 223344
 0 | 777899          0 | 556677778888
 1 | 02              1 | 00112
 1 | 555688          1 | 8
 2 | 224
              Legenda: 0|2 : leia-se 2 meses
```

Comentário: Através do ramo-e-folhas com escala de 10 meses, a distribuição de frequências dos tempos nos dois grupos poderiam ser consideradas assimétricas com concentração à esquerda. No entanto, o ramo-e-folhas com escala mais refinada (5 meses) permite um melhor detalhamento da distribuição dos tempos nos dois grupos. A distribuição dos tempos no grupo feminino ainda pode ser considerada assimétrica com concentração à esquerda. Porém, a análise do ramo-e-folhas do grupo masculino nos revela que o grupo dos homens parece ser a união de dois grupos: um, que tem tempos "baixos"e, outro grupo que tem tempos "altos". O ponto de corte para a definição de um tempo "baixo" ou "alto" ficaria a critério do pesquisador, mas poderia ser, por exemplo, 12 meses. Essa subdivisão do grupo masculino também pode ser notada no diagrama de pontos, embora de maneira mais sutil.

A escolha da escala a ser usado num gráfico deve ser guiada pelo bom senso: nem muito grande, o que tornaria a representação grosseira; nem muito pequena, o que diminuiria o poder de resumo da representação gráfica.

Seção 2.2: Análise Descritiva e Exploratória de Dados

Exercício (2.2.1)

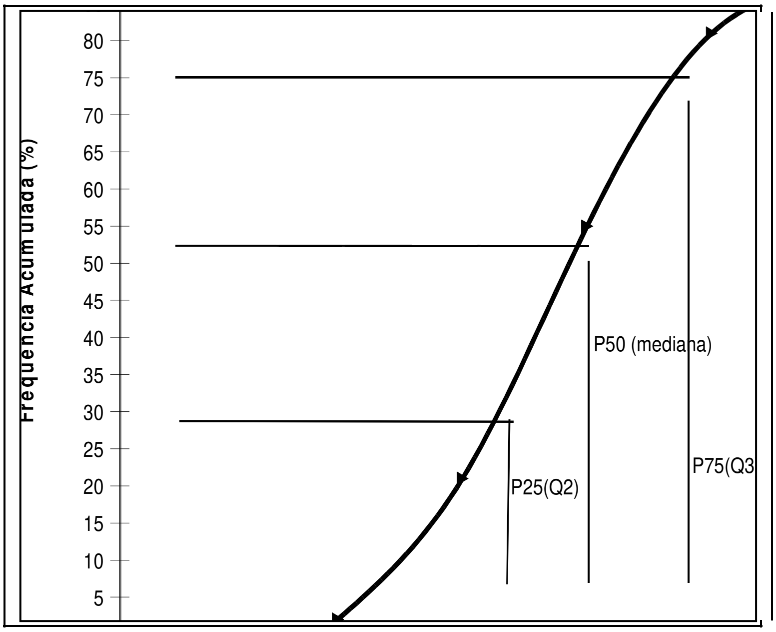
Grupo	Média	Mediana	Desvio-padrão	Coeficiente de Variação
Homens	10,71	9,00	6,82	0,64
Mulheres	7,17	7,00	3,67	0,51

Os homens possuem tempos entre a remissão e a recidiva da doença maiores, em média e em mediana, do que os tempos das mulheres. A análise do coeficiente de variação confirma o que já havíamos observado no diagrama de pontos: o grupo feminino possui tempos mais homogêneos (C.V. menor) do que o grupo masculino.

Exercício (2.2.2)

Queremos encontrar a mediana, que é o valor que deixa 50% dos dados abaixo dela. Essa também é a definição do percentil de ordem 50. O primeiro quartil é o valor que deixa um quarto dos dados (25%) abaixo dele e, portanto, é o percentil de ordem 25. O terceiro quartil é o valor que deixa três quartos dos dados (75%) abaixo dele e, portanto, é o percentil de ordem 75. O percentil de ordem 90 é o valor que deixa 90% dos dados abaixo dele.

O cálculo de percentis através da ogiva é aproximado, pois os dados foram agrupados para que o ogiva pudesse ser construída. Para encontrar o percentil de ordem 50, por exemplo, devemos traçar no gráfico, a partir do valor 50 no eixo das frequências acumuladas, uma reta paralela ao eixo dos valores (eixo X) até que essa reta encontre a curva (veja no gráfico). No ponto em que a primeira reta encontrar a curva, devemos traçar outra reta, agora vertical, paralela ao eixo das frequências (eixo Y), até que essa reta encontre o eixo dos valores (eixo X). O ponto em que essa segunda reta encontrar o eixo X será o percentil de ordem 50. No gráfico, ele é aproximadamente o valor 19,0. Usamos o mesmo raciocínio para encontrar qualquer outro percentil. No gráfico, também estão ilustrados os procedimentos para encontrar os valores dos percentis de ordem 25, 75 e 90 (P25, P75 e P90, respectivamente). Os valores aproximados para esses percentis são P25 igual a 16,5 , P75 igual a 22,5 e P95 igual a 34,5.



Observe que o primeiro e o terceiro quartis (que deixam a mesma porcentagem de dados, 25%, abaixo e acima deles, respectivamente) estão praticamente à mesma distância da mediana. Isso é um índicio de que a distribuição desses dados de pressão intraocular pode ser considerada simétrica. Outro indício de simetria é que o percentil de ordem 5 (calcule!!!) e percentil de ordem 95 (que deixam a mesma porcentagem de dados, 5%, abaixo e acima deles, respectivamente) também estão praticamente à mesma distância da mediana.

Exercício (2.2.3)

a) Para comparar a homogeneidade de dois grupos com médias diferentes, precisamos calcular o coeficiente de variação, que é uma medida de variabilidade que considera a escala em que a variável está sendo medida.

Coeficiente de variação no grupo de recrutas: 22/205 = 0,11
Coeficiente de variação no grupo de oficiais: 45/244 = 0,18

O grupo de oficiais, além de ter um nível médio de colesterol maior do que o nível médio dos recrutas, também é um grupo mais heterogêneo com respeito a essa variável (C.V. maior do que o dos recrutas). Lembrando que o nível de colesterol de um indivíduo está relacionado também ao seu grau de sedentarismo, a maior heterogeneidade do nível de colesterol entre os oficiais pode ser explicada pela diversidade de funções que esses oficiais desempenham, podendo implicar em maior ou menor sedentarismo. Já os jovens recrutas passam por constantes treinamentos e exercícios, o que diminui a heterogeneidade entre eles desse importante fator associado ao nível de colesterol.

b) O nível de colesterol do novo membro está à mesma distância do valor médio dos dois grupos (19,5 mg/dl). Porém, a variabilidade do nível de colesterol é diferente nos dois grupos. Assim, para avaliar essa distância, usaremos o Escore Padronizado (EP), onde $EP = \dfrac{x_i - \overline{x}}{s}$.

Desse modo, temos que $EP_{recrutas} = \dfrac{224,5-205}{22} = 0,89$ e $EP_{oficiais} = \dfrac{224,5-244}{45} = -0,43$.

Assim, vemos que o nível de colesterol do novo membro está 0,89 desvios-padrão acima do nível de colesterol médio dos recrutas, mas está 0,43 desvios-padrão abaixo do nível de colesterol médio dos oficiais. Assim, apesar de estar de o nível de colesterol do novo membro estar à mesma distância do valor médio dos dois grupos, a análise dos escores padronizados nos revela que ele se distancia menos do grupo dos oficiais, pois $|EP_{oficiais}| < |EP_{recrutas}|$.

Exercício (2.2.4)

a)

Grupo	Média	Desvio padrão	Coeficiente	Mediana	Primeiro	Terceiro	P_{10}
Macho	40,54	3,54	0,09	40,00	38,00	43,50	35
Fêmea	52,04	8,05	0,15	51,00	47,00	58,50	42

<u>Exemplos de cálculo dos percentis:</u>

Cálculo do 1º quartil (grupo das tartarugas machos):
0,25 x 24 = 6 (número inteiro). O 1º quartil está o 6º valor e o 7º valor, os números 38 e 38, respectivamente. Tomaremos a média desses valores, que é o próprio 38.

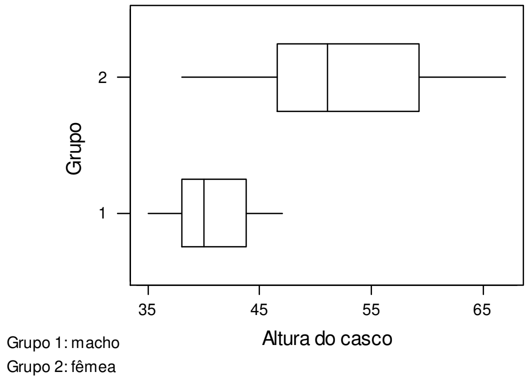
Cálculo do 3º quartil (grupo das tartarugas fêmeas):
075 x 24 = 18 (número inteiro). O 3º quartil está o 18º valor e o 19º valor, os números 57 e 60, respectivamente. Tomaremos a média desses valores, que é 58,50.

Cálculo do Percentil de ordem 10 (grupo das tartarugas fêmeas):
0,10 x 24 = 2,4 (número fracionário⇒arredondar para cima) O percentil de ordem 10 é o 3º valor, o número 42. (isto significa que aproximadamente 10% das tartarugas fêmeas possuem cascos com alturas inferiores ou iguais a 42).

Cálculo do Percentil de ordem 90 (grupo das tartarugas fêmeas):
0,90 x 24 = 21,6 (número fracionário⇒arredondar para cima) O percentil de ordem 90 é o 22º valor, o número 63 (isto significa que aproximadamente 90% das tartarugas fêmeas possuem cascos com alturas inferiores ou iguais a 63.

Observação: existem vários métodos para o cálculo de percentis e todos resultarão em valores aproximados. O método apresentado acima é o mais simples de todos.

b)



c) As fêmeas têm, em geral, cascos mais altos que os machos, pois os valores da média e da mediana são maiores para as fêmeas. Isto também pode ser visto nos box-plots, onde a maior parte das medidas das fêmeas encontram-se acima das medidas dos machos: o terceiro quartil dos machos é menor que o primeiro quartil das fêmeas. Além disso, os box-plots mostram que a altura do casco é bem mais variável para as fêmeas do que para os machos (a largura da "caixa" é maior para as fêmeas), o que é confirmado comparando-se os coeficientes de variação. O grupo de fêmeas é menos homogêneo quanto à altura do casco.

Exercício (2.2.5)

Ramo-e-Folhas para número colônias por cultura de escarro

n = 44 Escala: |1|7 = 17

|1|7
|2|23333444444
|2|55555556889
|3|0011
|3|5556
|4|01112
|4|
|5|14
|5|668
|6|0
|6|8
|7|
|7|9

Podemos ver que a distribuição de freqüência do número de colônias por cultura de escarro é muito assimétrica.

Faixa de Referência de 95%

$(100-\alpha)\% = 95\%$
$\alpha = 5$
$\alpha/2 = 5/2 = 2{,}5$
$100 - \alpha/2 = 100 - 2{,}5 = 97{,}5$

Faixa de Referência de $(100-\alpha)\%$: [Percentil de ordem $\alpha/2$; Percentil de ordem $(100-\alpha/2)$]
Assim, a Faixa de Referência de 95% : [Percentil de ordem 2,5; Percentil de ordem 97,5]

- Percentil de ordem 2,5 (2,5%): 44 x 0,025 = 1,1 (arredonda para 2)
  O 2^{O} valor (em ordem crescente) é X = 22
  Portanto: $P_{2,5} = 22$

- Percentil de ordem 97,5 (97,5%): 44 x 0,975 = 42,9 (arredonda para 43)
  O 43^{O} valor (em ordem crescente) é X = 68
  Portanto: $P_{97,5} = 68$

Portanto, a Faixa de Referência de 95% para o número de colônias por cultura de escarro em pessoas sadias é [22; 68] colônias, pelo método dos percentis.

Observação: Fazendo pelo método dos percentis ensinado na apostila *Introdução à Bioestatística*.

Na tabela a seguir, vemos que as ordem de percentis que mais se aproximam de 0,025 e 0,975 são, respectivamente 0,034091 e 0,965909 correspondendo aos valores de X ($n^{\underline{o}}$ de colônias por cultura de escarro) X = 22 e X = 68.

Portanto, a Faixa de Referência de 95% para o número de colônias por cultura de escarro em pessoas sadias é [22 ; 68], pelo método dos percentis da Apostila.

$N^{\underline{o}}$ de ordem i	X: $n^{\underline{o}}$ de colônias por culura de escarro	Ordem do percentil (i- 0,5)/44	
1	17	0,011364	
2	22	0,034091	→ Mais próximo do $P_{2,5}$
3	23	0,056818	
.....	.....	.....	
42	60	0,943182	
43	68	0,965909	→ Mais próximo do $P_{97,5}$
44	79	0,988636	

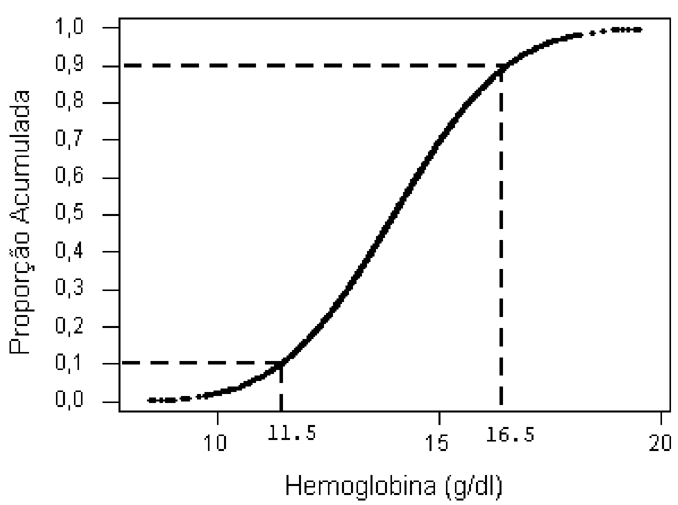
Observação: Note que, na verdade, a faixa de referência encontrada acima é de
(96, 5909 - 3,4091) ≅ 93%

Exercício (2.2.6)

A Faixa de Referência de 80% será dada por FR(80%) = [P_{10}; P_{90}], pois:

$$(100-\alpha)\% = 80\%$$
$$\alpha = 100- 80 = 20$$
$$\alpha /2 = 20/2 = 10$$
$$100 - \alpha /2 = 100- 10 = 90$$

O percentil de ordem 10% (P_{10}) pode ser encontrado procurando-se o valor de hemoglobina ao qual corresponde a proporção acumulada igual a 0,10 na ogiva. Este valor é aproximadamente 11 g/dl.



O percentil de ordem 90% (P_{90}) pode ser encontrado procurando-se o valor de hemoglobina ao qual corresponde a proporção acumulada igual a 0,90 na ogiva. Este valor é aproximadamente 17 g/dl.

A Faixa de Referência de 80% são os valores de hemoglobina de 11 a 17 g/dl. Isto significa que aproximadamente 80% das mulheres sadias desta população têm valor de hemoglobina entre 11 e 17 g/dl.

Seção 2.3: Análise Descritiva e Exploratória de Dados
(Associação entre Variáveis)

Exercício (2.3.1)

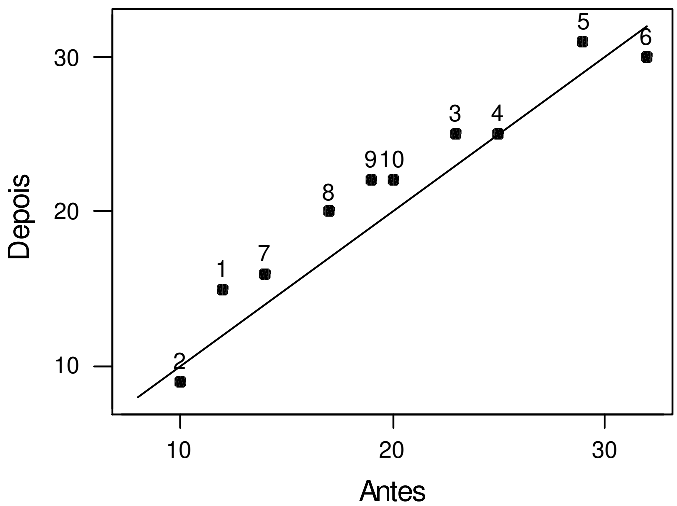
No gráfico a seguir, verificamos que a habilidade aumenta após o programa de fortificação para a maioria dos alunos (pontos acima da linha x=y). Para apenas três deles, a habilidade é igual ou menor após o programa. Portanto, o programa se mostrou eficaz para a maioria dos alunos. Note que os pontos tendem a se posicionar ao longo de uma linha reta (no caso, a linha x=y).

Algumas considerações sobre o relacionamento entre variáveis quantitativas: no problema acima, dizemos que os escores antes e depois do programa estão *linearmente correlacionados*. O Diagrama de Dispersão é um modo de visualizar essa correlação. Para medir o grau de correlação, usa-se o *Coeficiente de Correlação Linear*, que é um número entre -1 e 1. Um coeficiente de correlação positivo (> 0) acontece se a relação entre as duas variáveis é do tipo: quando o valor de uma variável aumenta, o valor da outra aumenta também; ou, quando o valor de uma variável diminui, o valor da outra também diminui. Já um coeficiente de correlação negativo (< 0) acontece se a relação entre as duas variáveis é do tipo: quando o valor de uma variável aumenta, o valor da outra diminui; ou vice-versa. Um coeficiente de correlação linear igual a zero significa que as duas variáveis não possuem nenhuma correlação do tipo linear, mas podem estar relacionadas de uma maneira não linear (quadrática, por exemplo). Isso poderá ser percebido através do exame do Diagrama de Dispersão. Assim, o Coeficiente de Correlação Linear e o Diagrama de Dispersão são ferramentas que devem ser usadas conjuntamente. Para mais detalhes sobre correlação, veja *Soares, J. F. e colegas (Introdução à Estatística)* e também *Triola, M. F. (Introdução à Estatística, Editora LTC)* . O coeficiente de correlação entre os escores antes e depois nesse exercício é de 0,97, indicando uma forte correlação positiva entre as medidas antes e depois do programa de fortificação isométrica.

Gráfico de Dispersão

Escores de abdominais antes e depois do programa

(Linha x=y)



Exercício (2.3.2)

a)

Distribuição de freqüências calculadas por coluna

Resposta	Placebo	Vacina	Total
Baixa	25 (65,79%)	6 (17,14%)	31 (42,46%)
Moderada	8 (21,05%)	18 (51,43%)	26 (35,62%)
Alta	5 (13,16%)	11 (31,43%)	16 (21,92%)
Total	38 (100%)	35 (100%)	73 (100%)

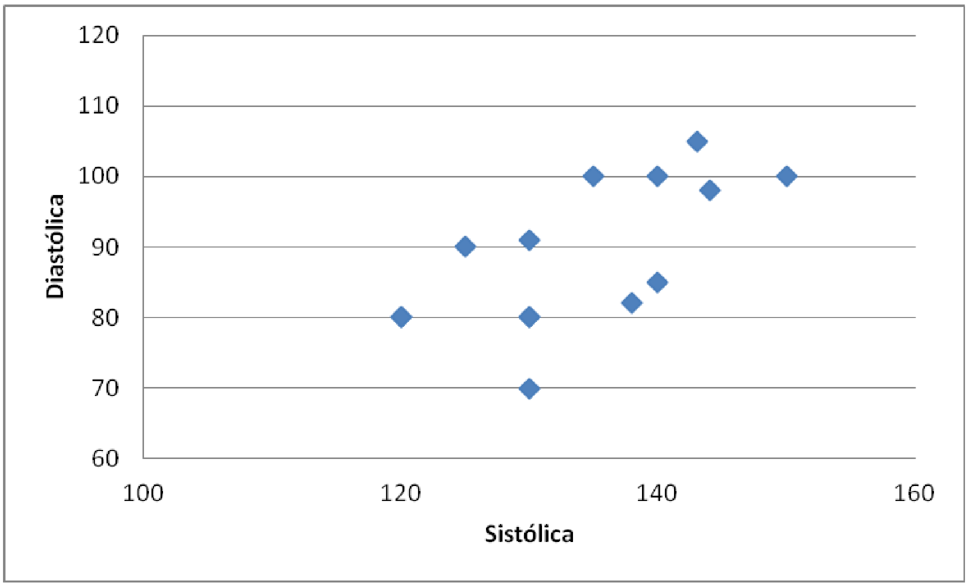
Distribuição de freqüências calculadas por linha

Resposta	Placebo	Vacina	Total
Baixa	25 (80,64%)	6 (19,36%)	31 (100%)
Moderada	8 (30,77%)	18 (69,23%)	26 (100%)
Alta	5 (31,25%)	11 (68,75%)	16 (100%)
Total	38 (52,06%)	35 (47,94%)	73 (100%)

b)

Analisando as porcentagens por coluna, observa-se que existe associação entre intensidade de resposta e o grupo do paciente (vacina ou placebo), pois as respostas variam de acordo com o grupo de pacientes. Ou seja, em relação aos pacientes do grupo placebo a maioria obteve resultado de intensidade baixa, e nos pacientes do grupo vacina a maioria obtiveram resultado de intensidade moderada no tratamento de gripe.

Exercício (2.3.3)



Percebemos que as medidas têm correlação positiva, pois quando a pressão sanguínea sistólica de um paciente aumenta, a pressão diastólica também tende a aumentar.

b)

i	x_i	y_i	$(x_i$-x(barra))(yi-y(barra))	(xi-x(barra))²	(yi-y(barra))²
1	138	82	-3,79	16,65	0,86
2	130	91	-31,63	15,37	65,12
3	135	100	18,44	1,17	291,38
4	140	100	103,79	36,97	291,38
5	120	80	40,79	193,77	8,58
6	125	90	63,06	79,57	49,98
7	120	80	40,79	193,77	8,58
8	130	80	11,49	15,37	8,58
9	130	80	11,49	15,37	8,58
10	144	98	151,91	101,61	227,1
11	143	105	200,4	82,45	487,08
12	140	85	12,59	36,97	4,28
13	130	70	50,69	15,37	167,18
14	150	100	274,49	258,57	291,38
Soma	1875	1241	944,51	1062,98	1910,06

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right)\left(\sum (y_i - \bar{y})^2\right)}}$$

$$r = \frac{944.51}{\sqrt{1062.98 \times 1910.06}} = 0.66$$

O coeficiente de correlação de Pearson é 0.66, o que indica uma correlação positiva e moderada entre a pressão sistólica e diastólica dos pacientes.

Seção 3: Probabilidade

Exercício (3.1)

a) Existem quatro ases em um baralho de 52 cartas. Assim P(A) = 4/52 = 1/13.
Metade as cartas (as de naipe ouros e copas) são vermelhas. Assim, P(V) = 26/52 = 1/2.
Cada naipe tem 13 cartas. Assim, P(E) = 13/52 = 1/4.

b) Há duas cartas que são, ao mesmo tempo, ases e vermelhas (ás de ouros e ás de copas). Assim, $P(A \cap V) = 2/52 = 1/26$.
Há somente uma carta que é, ao mesmo tempo, ás e de espada. Assim, $P(A \cap E) = 1/52$.
Nenhuma carta é, ao mesmo tempo, vermelha e de espada. Assim, $P(V \cap E) = 0$, ou seja, os eventos V e E são mutuamente exclusivos.

c) $P(A \cup V) = P(A) + P(V) - P(A \cap V) = 1/13 + 1/2 - 1/26 = (2 + 13 - 1)/26 = 14/26 = 7/13$.
$P(A \cup E) = P(A) + P(E) - P(A \cap E) = 1/13 + 1/4 - 1/52 = (4 + 13 -1)/52 = 16/52 = 4/13$.
$P(V \cup E) = P(V) + P(E) - P(V \cap E) = 1/2 + 1/4 - 0 = (2+1)/4 = 3/4$.

d) $P(A | V) = P(A \cap V)/P(V) = (1/26)/(1/2) = 1/13$. Ou seja, como, das 26 cartas vermelhas, duas são ases, então $P(A | V) = 2/26 = 1/13$.
Como $P(A | V) = P(A)$, os eventos A e V são independentes.

e) $P(V|E) = P(V \cap E)/P(E) = 0/(1/4) = 0$. Ou seja, como nenhuma carta de espada é vermelha, então $P(V|E) = 0$.
Como $P(V|E) \neq P(V)$, os eventos V e E não são independentes.

f.1) O fato da primeira carta retirada ter sido um ás não altera a probabilidade de que a segunda carta seja um ás, pois a primeira carta foi devolvida ao baralho, tornando, assim, a segunda retirada um experimento idêntico àquele da primeira retirada. Então, os eventos A1 e A2 são independentes: $P(A2|A1) = P(A2) = 4/52 = 1/13$.

f.2) Como os eventos A1 e A2 são independentes, $P(A1 \cap A2) = P(A1) \times P(A2) = 1/13 \times 1/13 = 1/169$.

Exercício (3.2)

a) Espaço amostral do experimento:

$$
\begin{aligned}
E = \{ & (D, D, D); \\
& (D, \overline{D}, \overline{D}); (\overline{D}, D, \overline{D}); (\overline{D}, \overline{D}, D); \\
& (D, D, \overline{D}); (D, \overline{D}, D); (\overline{D}, D, D); \\
& (\overline{D}, \overline{D}, \overline{D}) \},
\end{aligned}
$$

onde a vírgula substitui o símbolo da interseção "$\cap$".

b) A probabilidade de uma pessoa daltônica é $P[D]=1/4$ e $p[\overline{D}]=(1-1/4)= 3/4$.
Seja X o número de pessoas daltônicas. Então, X pode assumir os valores 0, 1, 2 ou 3.

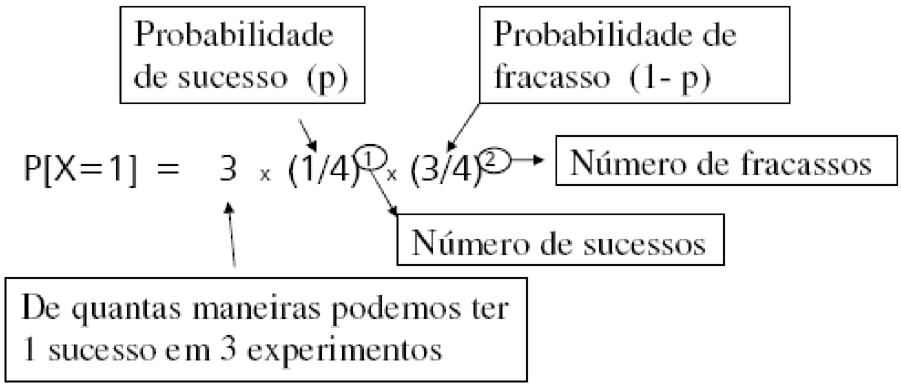
$$
\text{b.1) } P[X=0] = P[(\overline{D}, \overline{D}, \overline{D})] = P[\overline{D}] \times P[\overline{D}] \times P[\overline{D}] = (P[\overline{D}])^3 = (3/4)^3
$$

$$
\begin{aligned}
\text{b.2) } P[X=1] &= P[(D, \overline{D}, \overline{D}) \cup (\overline{D}, D, \overline{D}) \cup (\overline{D}, \overline{D}, D)] \\
&= P[(D, \overline{D}, \overline{D})] + P[(\overline{D}, D, \overline{D})] + P[(\overline{D}, \overline{D}, D)] \\
&= \{ P[D] \times P[\overline{D}] \times P[\overline{D}] \} + \{ P[\overline{D}] \times P[D] \times P[\overline{D}] \} + \{ P[\overline{D}] \times P[\overline{D}] \times P[D] \} \\
&= (1/4) \times (3/4) \times (3/4) + (3/4) \times (1/4) \times (3/4) + (3/4) \times (3/4) \times (1/4) \\
&= (1/4) \times (3/4)^2 + (1/4) \times (3/4)^2 + (1/4) \times (3/4)^2 \\
&= 3 \times (1/4) \times (3/4)^2
\end{aligned}
$$

$$
\begin{aligned}
\text{b.3) } P[X=2] &= P[(D, D, \overline{D}) \cup (D, \overline{D}, D) \cup (\overline{D}, \overline{D}, \overline{D})] \\
&= P[(D, D, \overline{D})] + P[(D, \overline{D}, D)] + P[(\overline{D}, \overline{D}, D)] \\
&= (1/4) \times (1/4) \times (3/4) + (1/4) \times (3/4) \times (1/4) + (3/4) \times (1/4) \times (1/4) \\
&= (1/4)^2 \times (3/4) + (1/4)^2 \times (3/4) + (1/4)^2 \times (3/4) \\
&= 3 \times (1/4)^2 \times (3/4)
\end{aligned}
$$

$$
\begin{aligned}
\text{b.4) } P[X=3] &= P[(D, D, D)] \\
&= P[D] \times P[D] \times P[D] \\
&= (P[D])^3 = (1/4)^3
\end{aligned}
$$

c) Se definirmos como "sucesso" o evento D, teremos como "fracasso" o evento $\overline{D}$.
Assim, podemos identificar o seguinte padrão no cálculo das probabilidades em b) :



O mesmo ocorre com as outras probabilidades

P[X=0] = $(3/4)^3 = 1 \times (1/4)^0 \times (3/4)^3$

P[X=2] = $3 \times (1/4)^2 \times (3/4)^1$

P[X=3] = $(1/4)^3 = 1 \times (1/4)^3 \times (3/4)^0$

Este padrão é o que vamos definir como o *Modelo Probabilístico Binomial*, na seção 5.
Com um modelo para calcular as probabilidades que desejamos, não necessitaremos de passar por todo o processo de enumerar o espaço amostral.
Um experimento aleatório com somente 3 pessoas gerou um espaço amostral de tamanho 8. Se sortearmos 4 pessoas, o espaço amostral terá 16 pontos. Com 5 pessoas, serão 32 pontos e, assim por diante, crescendo sempre em potência de 2. No entanto, usar um modelo probabilístico para calcular probabilidades nos poupará deste trabalho.

Exercício (3.3)

Tabela 3.1

Daltonismo	Sexo		Total
	Masculino	Feminino	
Presente	423	65	488
Ausente	4848	4664	9512
Total	5271	4729	10.000

Sejam os eventos:
D = {a pessoa escolhida é Daltônica}
ND = {a pessoa escolhida é Não Daltônica}
M = {a pessoa escolhida é do sexo Masculino}
F = {a pessoa escolhida é do sexo Feminino}

Item (I)

a) $P(D) = 488/10.000 = 0{,}0488$
b) $P(ND) = 9512/10.000 = 0{,}9512$

Observe que $P(D) + P(ND) = 1$

c) $P(M) = 5271/10.000 = 0{,}5271$
d) $P(F) = 4729/10.000 = 0{,}4729$

Observe que $P(M) + P(F) = 1$

e) $P(D \cap M) = 423/10.000 = 0{,}0423$
f) $P(D \cap F) = 65/10.000 = 0{,}0065$

g) $P(ND \cup M) = P(ND) + P(M) - P(ND \cap M) =$
$9512/10.000 + 5271/10.000 - 4848/10.000 = 0{,}9935$
h) $P(ND \cup F) = P(ND) + P(F) - P(ND \cap F) =$
$9512/10.000 + 4729/10.000 - 4664/10.000 = 0{,}9577$

Observe que:
$P(D \cap M) + P(D \cap F) +$
$+ P(ND \cap M) + P(ND \cap F) = 1$

i) $P(D|M) = P(D \cap M)/P(M) = 0{,}0423/0{,}5271 = 0{,}0803$
j) $P(D|F) = P(D \cap F)/P(F) = 0{,}0065/0{,}4729 = 0{,}0137$

k) $P(ND|M) = P(ND \cap M)/P(M) = 0{,}4848/0{,}5271 = 0{,}9197$
l) $P(ND|F) = P(ND \cap F)/P(F) = 0{,}4664/0{,}4729 = 0{,}9863$

Observe que:
$P(D|M) + P(ND|M) = 1$
$P(D|F) + P(ND|F) = 1$

Item (II)

$P(D) = 0{,}0488$
$P(D|M) = 0{,}0803 \neq P(D)$

Portanto, os eventos “ser daltônica” e “ser do sexo masculino” não são independentes

Exercício (3.4)

Considere os seguintes eventos de interesse:

S = nascer com menos de 37 semanas de gestação (prematuro)

$\overline{S}$ = nascer com mais de 37 semanas de gestação (não-prematuro)

B = nascer com menos de 2500 gramas (baixo peso)

$\overline{B}$ = nascer com 2500 gramas ou mais (peso normal)

De acordo com o enunciado do problema, temos que

$P[S]=0{,}142$. Então, $P\left[\overline{S}\right]=1-P[S]=1 - 0{,}142 = 0{,}858$.

$P[B\,|\,S]=$ 0,218 e $P\left[B\,|\,\overline{S}\right]=$ 0,023.

a.1) $P[B\cap S]=?$

Da regra que define a probabilidade condicional de um evento, temos que

$$P[B\,|\,S]=\frac{P[B\cap S]}{P[S]}\text{. Então, } P[B\cap S]=P[B\,|\,S]\times P[S]$$

$$\text{Assim, } P[B\cap S]=P[B\,|\,S]\times P[S]=\ 0{,}218\times 0{,}142=0{,}031$$

a.2) $P\left[B\cap\overline{S}\right]=?$. Usando a mesma regra usada no item a.1 :

$$P\left[B\cap\overline{S}\right]=P\left[B\,|\,\overline{S}\right]\times P\left[\overline{S}\right]=\ 0{,}023\times 0{,}858=0{,}020$$

b) Probabilidade de que um bebê nasça com baixo peso

O evento B é a união de dois eventos mutuamente exclusivos : "nascer com baixo peso e não prematuro" $\left[B\cap\overline{S}\right]$ e "nascer com baixo peso e prematuro" $[B\cap S]$. Assim, pela regra do cálculo da probabilidade da união de dois eventos, temos que

$$\begin{aligned} P(B) &= P\left([B\cap S]\cup\left[B\cap\overline{S}\right]\right)=P\left([B\cap S]\right)+P\left(\left[B\cap\overline{S}\right]\right) \\ &= 0{,}031+0{,}020=0{,}051 \end{aligned}$$

c) $P[S\,|\,B]=?$

Pela regra da probabilidade condicional, $P[S\,|\,B]=\dfrac{P[B\cap S]}{P[B]}$. Assim, usando as probabilidades calculadas nos itens anteriores, temos que

$$P[S\,|\,B]=\frac{0{,}031}{0{,}051}=0{,}608$$

A probabilidade de que a gestação de um bebê tenha durado menos do que 37 semanas, dado que ele nasceu com baixo peso é de 0,608.

Exercício (3.5)

Tabela 3.2

Situação da paciente	Resultado do Papanicolau		Total
	Positivo	Negativo	
Com câncer	94	6	100
Sem câncer	250	250	500
Total	344	256	600

Sejam os eventos: C = {ter câncer}
S = {não ter câncer}
P = {teste papanicolau positivo}
N = {teste papanicolau negativo}

a) Prevalência do câncer = 100/600 = 0,1617

b) (Pessoas com câncer e teste positivo) + (Pessoas sem câncer e teste negativo) = 94 + 250 = 344

c) (Pessoas com câncer e teste negativo) + (Pessoas sem câncer e teste positivo) = 6 + 250 = 256

d) Sensibilidade: $P(P|C)$ = 94/100 = 0,94

e) Especificidade: $P(N|S)$ = 250/500 = 0,50

f) Valor de Predição Positiva: $P(C|P)$ = 94/344 = 0,27

g) Valor de Predição Negativa: $P(S|N)$ = 250/256 = 0,98

h) Proporção de Falsos Positivos: $P(S|P)$ = 250/344 = 0,73

i) Proporção de Falsos Negativos: $P(C|N)$ = 6/256 = 0,02

Observe que:

Valor de Predição Positiva + Proporção de Falsos Positivos = 1

Valor de Predição Negativa + Proporção de Falsos Negativos = 1

Seção 4: Avaliação da Qualidade de Testes Clínicos

Exercício (4.1)

a) Proporção de falsos positivos: PFP = P($\overline{D}$ | +)

b) Valor de Predição Negativa: VPN = P($\overline{D}$ | -)

c) Especificidade: e = P(- | $\overline{D}$)

d) Valor de Predição Positiva: VPP = P(D | +)

e) Proporção de falsos negativos: PFN = P(D | -)

f) Sensibilidade: s = P(+ | D)

Exercício (4.2)

a) Sensibilidade : $s = p(+ | D) = \dfrac{51}{52} = 0{,}98$

Especificidade: $e = P(- | \overline{D}) = \dfrac{232}{238} = 0{,}97$

b) Prevalência na população:p = 0,179

Prevalência na amostra: $\dfrac{\text{número de doentes na amostra}}{\text{tamanho total da amostra}} = \dfrac{52}{290} = 0{,}1793 \simeq p$

Como a prevalência da doença na amostra é próxima da prevalência na população, podemos calcular VPP e VPN diretamente da tabela.

$VPP = P(D | +) = \dfrac{51}{57} = 0{,}89$ PFP = 1-VPP = 0,11

$VPN = P(\overline{D} | -) = \dfrac{232}{233} = 0{,}996$ PFN = 1-VPN = 0,004

c) Suponha que p <<0,179. Então, VPP << 0,89 e, daí, PFP >>0,11. Portanto, um resultado positivo no teste tem uma probabilidade considerável de ser um falso positivo. Como o tratamento é caro, não podemos correr um risco de estar tratando um "falso positivo" com muita frequência (PFP). Assim, o melhor procedimento diante de um resultado positivo é fazer mais investigações com outros testes de diagnóstico.

Exercício (4.3)

Resultados do Teste 1

Infarto agudo do miocárdio	Diagnóstico do Teste		Total
	Positivo (CFC ≥ 80)	Negativo (CFC < 80)	
Doente (D)	215	15	230
Não doente $(\overline{D})$	16	114	130
Total	231	129	360

Resultados do Teste 2

Infarto agudo do miocárdio	Diagnóstico do Teste		Total
	Positivo (CFC ≥ 280)	Negativo (CFC < 280)	
Doente (D)	97	133	230
Não doente $(\overline{D})$	1	129	130
Total	98	262	360

a)

Teste 1:

$$s_1 = P(+|D) = \frac{215}{230} = 0,93$$

$$e_1 = P(-|\overline{D}) = \frac{114}{130} = 0,88$$

Teste 2:

$$s_2 = P(+|D) = \frac{97}{230} = 0,42$$

$$e_2 = P(-|\overline{D}) = \frac{129}{130} = 0,99$$

O Teste 1 tem sensibilidade e especificidade bastante altas, o que não ocorre com o Teste 2, que tem uma especificidade muito alta e uma sensibilidade muito baixa. Assim, o Teste 1 é o teste mais sensível dentre os dois e o Teste 2 é o teste mais específico. A qualidade do diagnóstico baseado nesses testes, medida pelos Valores de Predição Positiva e Negativa, depende da prevalência da doença na população onde eles serão usados. Conforme for o objetivo do procedimento (descartar ou confirmar a presença da doença), um ou outro teste será mais útil, como veremos no próximo item.

b) Considerando que os testes seriam aplicados numa população cuja a prevalência de infarto agudo no coração é aproximadamente igual à prevalência da amostra (230/360=0,64), poderemos calcular os valores de predição positiva e negativa diretamente da tabela.

Teste 1: $VPP_1 = P(D|+) = \frac{215}{231} = 0,93$ PFP$_1$ = 1 – 0,93 = 0,07

$VPN_1 = P(\overline{D}|-) = \frac{114}{129} = 0,88$ PFN$_1$ = 1 – 0,88 = 0,12

Observação: a coincidência do valor da sensibilidade com o valor do VPP, assim como do valor da especificidade com o valor do VPN, para o teste 1, não é regra. Apenas aconteceu porque os denominadores são muito parecidos. Isso não ocorrerá com o teste 2, como veremos a seguir.

Teste 2: $VPP_2 = P(D|+) = \frac{97}{98} = 0,99$ PFP$_2$ = 1 – 0,99 = 0,01

$VPN_2 = P(\overline{D}|-) = \frac{129}{262} = 0,49$ PFN$_2$ = 1 – 0,49 = 0,51

c)

c.1) O Teste 2 será mais útil para confirmar a doença do que o Teste 1, pois o VPP do Tese 2 é mais alto (99%) do que o VPP do Teste 1 (93%).

c.2) O Teste 1 será mais útil para descartar a doença do que o Teste 2, pois o VPN do Tese 1 é mais alto (88%) do que o VPN do Teste 2 (49%).

Comentários sobre os testes: na escolha de qual teste deve ser usado, devemos estar atento ao objetivo do procedimento e à prevalência da doença na população onde ele for aplicado. Nesta análise, vamos considerar que a população de trabalho tem uma prevalência de infarto no coração aproximadamente igual à prevalência desse estudo (64%). Se o objetivo do procedimento diagnóstico for o de confirmar a doença, o Teste 2 deve ser usado, pois seu VPP é bastante alto (99%), isto é, um paciente que tenha o resultado positivo no Teste 2 tem grande probabilidade de estar doente. Por outro lado, se o objetivo for descartar a doença, o Teste 2 não deve ser mais usado, pois seu VPN é muito baixo, levando a uma Proporção de Falsos Negativos muito alta (51%). Assim, será preferível usar o Teste 1, que apesar de ter uma Proporção de Falsos Negativos considerável (12%), é melhor do que o Teste 2 nesse ponto. Numa população onde a prevalência de infarto for menor do que 64%, os valores de VPN para os dois testes vão aumentar e, então, a utilização do Teste 1 para descartar a presença da doença será mais confiável (a proporção de falsos negativos irá diminuir). Como já sabemos, quanto menor é a prevalência, maior é o valor de VPN e menor é o valor de VPP.

Exercício (4.4)

a) O VPP aumenta à medida que a especificidade aumenta, com sensibilidade e prevalência fixas.

b) O VPN aumenta à medida que a especificidade aumenta, com sensibilidade e prevalência fixas.

c)

i) VPP aumenta aproximadamente em 0,23 quando se passa de um teste com especificidade igual a 0,7 para 0,95, com p = 0,5.

ii) VPP aumenta aproximadamente entre 0,02 e 0,06, dependendo do valor da especificidade, quando se passa de um teste com sensibilidade igual a 0,7 para 0,95, com p = 0,5.

iii) O teste será poderoso para confirmar a presença da doença se ele tiver uma PFP pequena, o que significa que seu VPP é grande. Ou seja, haverá uma alta probabilidade de um paciente ser doente dado que seu resultado no teste foi positivo.

iv) Deseja-se, neste caso, um teste com VPP alto. Desse modo, devemos aumentar a especificidade do teste, pois, como foi mostrado nos itens (i) e (ii), o aumento da especificidade provoca um maior acréscimo no VPP.

d)

i) VPN aumenta aproximadamente em 0,003 quando se passa de um teste com especificidade igual a 0,7 para 0,95, com p = 0,05.

ii) VPN aumenta aproximadamente em 0,019, quando se passa de um teste com sensibilidade igual a 0,7 para 0,95, com p = 0,05.

iii) O teste será poderoso para descartar a presença da doença se ele tiver uma PFN pequena, o que significa que seu VPN é grande. Ou seja, haverá uma alta probabilidade de um paciente não ser doente dado que seu resultado no teste foi negativo.

iv) Deseja-se, neste caso, um teste com VPN alto. Desse modo, devemos aumentar a sensibilidade do teste, pois, como foi mostrado nos itens (i) e (ii), o aumento da sensibilidade provoca um maior acréscimo no VPN.

Exercício (4.5)

a) Se a faixa de referência é de 80%, o limite inferior é o percentil 10. Assim, espera-se que 90% das mulheres sadias esteja acima do limite inferior e 10% esteja abaixo deste limite.

b) Um método para diagnóstico de anemia baseado neste valor inferior de referência daria resultado:

negativo, se a paciente tivesse valor de hemoglobina maior do que 11 g/dl;
positivo, se a paciente tivesse esse valor menor do que 11 g/dl.

Como sabemos, a especificidade de um teste é estimada como sendo a frequência de negativos entre as pessoas sadias. Nesse problema, todas as mulheres são sadias, e 90% têm valores maiores do que 11 g/dl, sendo consideradas negativas. Assim, estimativa da especificidade é justamente 90%. Esse é um valor aproximado, já que a própria faixa é aproximada.

c) Como também sabemos, a sensibilidade de um teste é estimada como sendo a frequência de positivos entre as pessoas doentes. Aqui, a sensibilidade desse método não pode ser calculada somente com esses dados, pois precisamos de uma amostra de pessoas doentes para sabermos quantas delas serão consideradas positivas pelo método.

Seção 5: Distribuições de Probabilidade: Binomial e Poisson

Exercício (5.1)

X: n^{O} de pessoas com Rh negativo em um grupo de n=4

Valores que X pode assumir: x = 0,1,2,3 ou 4
Probabilidade de uma pessoa ser Rh negativo: p = 0,10

Desse modo, X ~ Binomial (n=4,p=0,10)

Assim, $P(X=x)=\binom{4}{x}\cdot 0,10^{x}\cdot(1-0,10)^{4-x}$ para x = 0,1,2, 3 ou 4

Estamos interessados em saber qual é a probabilidade de que todas as 4 pessoas desse grupo sejam Rh negativo, ou seja, estamos interessados em saber qual é a probabilidade da variável X assumir o valor 4 (P[X=4]).

$$P(X=4)=\binom{4}{4}\cdot 0,10^{4}\cdot(1-0,10)^{4-4}=\frac{4!}{4!\,0!}\cdot 0,10^{4}\cdot(0,90)^{0}=1\cdot 0,0001\cdot 1=0,0001$$

Quando a probabilidade de uma pessoa ser Rh negativo é 10%, a probabilidade de que todas as 4 pessoas de um grupo sejam Rh negativo é 0,01%.

Exercício (5.2)

Y: n^{O} de crianças com olhos azuis numa família de 3 filhos (n=3)

Valores que Y pode assumir: y = 0,1,2 ou 3
Probabilidade de uma criança ter olhos azuis: p = 0,25

Desse modo, Y ~ Binomial (n=3,p=0,25)

Assim, $P(Y=y)=\binom{3}{y}\cdot 0,25^{y}\cdot(1-0,25)^{3-y}$ para y = 0,1,2 ou 3

Pelo menos duas crianças devem ter olhos azuis, ou seja, 2 ou todas as 3 crianças. Assim, estamos interessados na probabilidade de Y assumir valores maiores ou iguais a 2, ou seja, os valores 2 ou 3.

P[Y≥2] = {P[Y=2]+P[Y=3]} = $\binom{3}{2}\cdot 0,25^{2}\cdot(1-0,25)^{3-2} + \binom{3}{3}\cdot 0,25^{3}\cdot(1-0,25)^{3-3}$

$= \frac{3!}{2!\,1!}\cdot 0,25^{2}\cdot(0,75)^{1} + \frac{3!}{3!\,0!}\cdot 0,25^{3}\cdot(0,75)^{0}$

= 3 . 0,0625 . 0,75 + 1 . 0,0156 . 1 = 0,1406 + 0,0156 = 0,1562

A probabilidade de que pelo menos 2 das 3 crianças de um casal de olhos azuis também tenham olhos azuis é de 15,62 %.

Exercício (5.3)

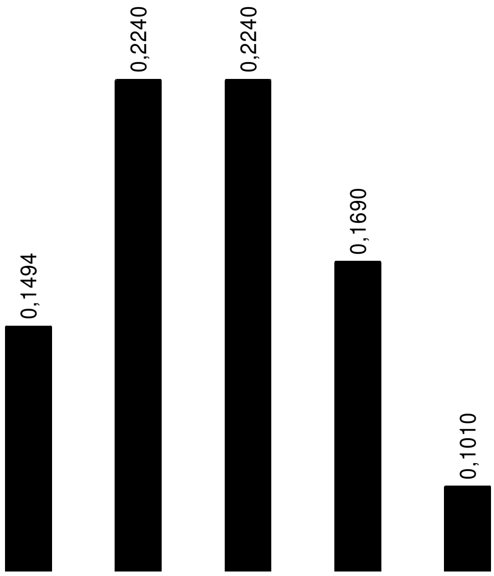
X: n^{o} de unidades do medicamento vendidas em 1 dia

Valores que X pode assumir: x = 0, 1, 2, ... (até um limite máximo desconhecido)
Número médio de unidades do medicamento vendidas por dia: 3 (λ = 3)

Supondo X ~ Poisson (λ = 3)

Assim, $P(X=x)=\dfrac{e^{-3} \cdot 3^{x}}{x!}$ para x = 0, 1, 2, ... (e ≈ 2,72)

a)



Número de unidades vendidas por dia

Note que a distribuição de p camento vendidas por dia é assimétrica com concentração à esquerda, isto é, nos valores mais próximos do valor esperado (3). Para valores maiores ou iguais a 12, a probabilidade é tão pequena que pode ser considerada igual a zero.

b) P[X entre 7 e 11] = P[7 ≤ X ≤ 11] = {P[X=7]+ P[X=8]+ P[X=9]+ P[X=10]+ P[X=11]}

Do gráfico em a), temos que :

P[X entre 7 e 11] = 0,0216 + 0,0081 + 0,0027 + 0,0008 + 0,0002 = 0,0334

A probabilidade de que sejam vendidas entre 7 e 11 unidades do medicamento por dia é de 3,34%.

c) Considerando que a procura pelo medicamento durante os dias da semana se mantém constante e que o que ocorre em um dia é independente do que ocorre em outro dia, podemos concluir que o número médio de unidades do medicamento vendidas em 1 semana é de 7 x 3 = 21 unidades. Assim, o número de unidades do medicamento vendidas em 1 semana tem distribuição de Poisson com média igual a 21.

d) Y: nº de unidades do medicamento vendidas em 1 semana (7 dias)

Valores que Y pode assumir: y = 0, 1, 2, ... (até um limite máximo desconhecido)
Número médio de unidades do medicamento vendidas em 1 semana: 21 (λ = 21)

Supondo Y~ Poisson (λ = 21).

Assim, $P(Y=y)=\dfrac{e^{-21}\cdot 21^{y}}{y!}$, para y = 0, 1, 2, ... (e ≈ 2,72)

$P(Y=100)=\dfrac{e^{-21}\cdot 21^{100}}{100!}$ = 0,00000000000000000000000000000000000135 (1,35 x 10^{-35})

(uma probabilidade baixíssima!!!)

Exercício (5.4)

X: nº de meninos de daltônicos em um grupo de n=4
Valores que X pode assumir: x = 0,1,2,3 ou 4
Probabilidade de um menino ser daltônico: p = 0,08

Portanto, X ~ Binomial (n=4,p=0,08)

Assim, $P(X=x)=\binom{4}{x}\cdot 0{,}08^{x}\cdot(1-0{,}08)^{4-x}$ para x = 0,1,2, 3 ou 4

A probabilidade de 3 meninos não serem daltônicos num grupo de 4 meninos é igual à probabilidade de 1 menino ser daltônico num grupo de 4 meninos, ou seja P(X=1):

$$P(X=1)=\binom{4}{1}\cdot 0{,}08^{1}\cdot(1-0{,}08)^{3}=\frac{4!}{1!3!}\cdot 0{,}08\cdot(0{,}92)=4\cdot 0{,}08\cdot 0{,}78=0{,}25$$

Exercício (5.5)

X: nº de animais que sobrevivem em n animais submetidos ao experimento cirúrgico
Valores que X pode assumir: x = 0,1,2,...,n
Probabilidade de um animal sobreviver é p = 2/3.

a) Distribuição de probabilidade de X: X ~ Binomial (n=5 , p=2/3)

$$P(X=0)=\binom{5}{0}\cdot\left(\frac{2}{3}\right)^{0}\cdot\left(\frac{1}{3}\right)^{5}=\frac{5!}{0!5!}\cdot 1\cdot\frac{1}{243}=1\cdot\frac{1}{243}=0{,}004$$

$$P(X=1)=\binom{5}{1}\cdot\left(\frac{2}{3}\right)^{1}\cdot\left(\frac{1}{3}\right)^{4}=\frac{5!}{1!4!}\cdot\frac{2}{3}\cdot\frac{1}{81}=5\cdot\frac{2}{243}=0{,}041$$

$$P(X=2)=\binom{5}{2}\cdot\left(\frac{2}{3}\right)^{2}\cdot\left(\frac{1}{3}\right)^{3}=\frac{5!}{2!3!}\cdot\frac{4}{9}\cdot\frac{1}{27}=10\cdot\frac{4}{243}=0{,}165$$

$$P(X=3)=\binom{5}{3}\cdot\left(\frac{2}{3}\right)^3\cdot\left(\frac{1}{3}\right)^2=\frac{5!}{3!\,2!}\cdot\frac{8}{27}\cdot\frac{1}{9}=10\cdot\frac{8}{27}\cdot\frac{1}{9}=0,329$$

$$P(X=4)=\binom{5}{4}\cdot\left(\frac{2}{3}\right)^4\cdot\left(\frac{1}{3}\right)^1=\frac{5!}{4!\,1!}\cdot\frac{16}{81}\cdot\frac{1}{3}=5\cdot\frac{16}{243}=0,329$$

$$P(X=5)=\binom{5}{5}\cdot\left(\frac{2}{3}\right)^5\cdot\left(\frac{1}{3}\right)^0=\frac{5!}{5!\,0!}\cdot\frac{32}{243}\cdot 1=1\cdot\frac{32}{243}=0,132$$

Note que P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = 1

b)

i. P(X = 3) = 0,33
ii. P(X ≥ 1) = 1- P(X = 0) = 1- 0,004 = 0,996
iii. "Mais de 2 animais não sobreviverem" é o mesmo que no "no máximo 2 animais sobreviverem": P(X ≤ 2) = P(X = 0) + P(X = 1) +P(X = 2) = 0,004 + 0,041 + 0,165 = 0,210

c) Se 60 animais se submeterem a essa cirurgia, espera-se que, em média, 60 x p = 60 x 1/3 = 20 não sobrevivam.

Exercício (5.6)

X: nº de sementes que <u>não</u> germinam em um pacote com 10 sementes

Valores que X pode assumir: x = 0,1,2,...,10
Probabilidade de uma semente <u>não</u> germinar: p = 0,2

X ~ Binomial (n=10,p=0,2)

Assim, $P(X=x)=\binom{10}{x}\cdot 0{,}2^{x}\cdot 0{,}8^{10-x}$ para x = 0,1,2,...,10

a) $P(X>4)=1-P(X\le 4)=1-\left(P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)\right)$

= 1 - 0,9672 = 0,0328

$$P(X=0)=\binom{10}{0}\cdot 0{,}2^{0}\cdot 0{,}8^{10}=0,1074$$

$$P(X=1)=\binom{10}{1}\cdot 0{,}2^{1}\cdot 0{,}8^{9}=0,2684$$

$$P(X=2)=\binom{10}{2}\cdot 0{,}2^{2}\cdot 0{,}8^{8}=0,3020$$

$$P(X=3)=\binom{10}{3}\cdot 0{,}2^{3}\cdot 0{,}8^{7}=0,2013$$

$$P(X=4)=\binom{10}{4}\cdot 0{,}2^4\cdot 0{,}8^6=0{,}0881$$

b) O número esperado de pacotes indenizados quando 1000 são vendidos é 1000 vezes a probabilidade de um pacote ser indenizado, ou seja, 1000 x 0, 0328 ≈ 33 pacotes.

Exercício (5.7)

X: n^{o} de colônias de bactérias por 10 ml de água de um lago.
Valores que X pode assumir: x = 0, 1, 2, ... (até um limite máximo desconhecido)

Supondo X ~ Poisson (λ = 3)

Assim, $P(X=x)=\dfrac{e^{-3}\cdot 3^{x}}{x!}$ para x = 0, 1, 2, ... (e ≈ 2,72)

a) Qual a probabilidade de não se achar nenhuma colônia em 10 ml de água desse lago ?

$$P(X=0)=\frac{e^{-3}\cdot 3^{0}}{0!}=\frac{2,72^{-3}\cdot 1}{1}=\frac{1}{2,72^{3}}=0,05$$

b) Qual a probabilidade de se achar pelo menos duas colônias em 10 ml de água desse lago ?

$$P(X\geq 2)=1-P(X\leq 1)=1-\left(P(X=0)+P(X=1)\right)=1-(0,05+0,15)=1-0,20=0,80$$

$P(X=0)=0,05$ pelo item (a)

$$P(X=1)=\frac{e^{-3}\cdot 3^{1}}{1!}=\frac{2,72^{-3}\cdot 3}{1}=\frac{3}{2,72^{3}}=0,15$$

c) O volume de água triplicou.
X: n^{o} de colônias de bactérias por <u>30 ml</u> de água de um lago.
Portanto, a média de X passa a ser λ = 3 x 3 = 9 colônias por 30 ml de água.

Assim, $P(X=x)=\dfrac{e^{-9}\cdot 9^{x}}{x!}$ para x = 0, 1, 2, ... (e ≈ 2,72)

$$P(X\leq 4)=\left(P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)\right)$$
$$=\left(\frac{e^{-9}\cdot 9^{0}}{0!}+\frac{e^{-9}\cdot 9^{1}}{1!}+\frac{e^{-9}\cdot 9^{2}}{2!}+\frac{e^{-9}\cdot 9^{3}}{3!}+\frac{e^{-9}\cdot 9^{4}}{4!}\right)=$$
$$(0,000123+0,001111+0,004998+0,014994+0,033737)=0,056074$$

Exercício (5.8)

a) Seja X o número de filhos por mulher na população em estudo. X ~ Poisson(λ= 2.5). Para atender ao critério de seleção, a mulher deve ter 3 filhos ou mais. Assim, a probabilidade que uma mulher selecionada aleatoriamente da população atenda ao critério de seleção é calculada como

$$P(X\geq 3)=1-P(X<3)=1-P(X\leq 2)=1-[P(X=0)+P(X=1)+P(X=2)]$$

$$P(X\geq 3)=1-P(X<3)=1-[0.082+0.205+0.257]=1-0.544=0.456$$

b) Seja Y o número de mulheres, em 10 selecionadas na população em estudo, que atendem ao critério de seleção. A distribuição de probabilidade de Y pode ser o modelo binomial com n=10 e probabilidade de sucesso p=0.456. Assim, a probabilidade que mais de 7 mulheres atendam ao critério de seleção é calculada como

$$Y \sim Bin(n = 10; p = 0.456)$$
$$P[Y > 7] = P[Y = 8] + P[Y = 9] + P[Y = 10]$$

Seção 6: Distribuições de Probabilidade: Normal

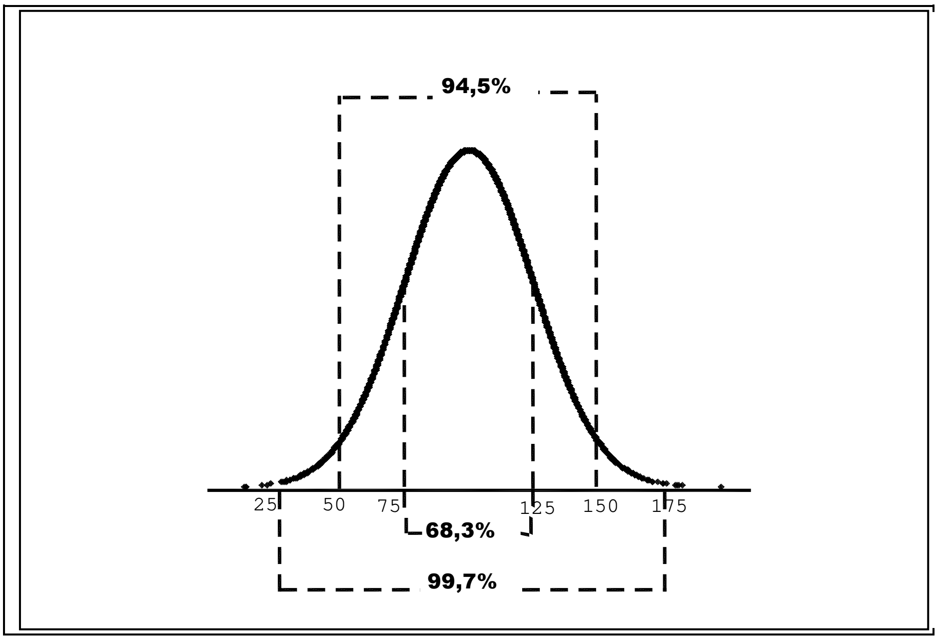
Exercício (6.1)

X: quantidade de ferro sérico de indivíduos sadios
X ~ Normal (μ = 100 ; σ = 25)

a) Quantidade <u>média</u> de ferro sérico em indivíduos desta população: <u>μ = 100 mcg/dl</u>
<u>Desvio padrão</u> da quantidade de ferro sérico nesta população: <u>σ = 25 mcg/dl</u>

b) Representação gráfica da função de densidade de probabilidade da variável X:

68,3% dos valores de X estão dentro do intervalo [μ ± σ] = [100 ± 25] = [75 ; 125] mcg/dl
95,4% dos valores de X estão dentro do intervalo [μ ± 2σ] = [100 ± 50] = [50 ; 150] mcg/dl
99,7% dos valores de X estão dentro do intervalo [μ ± 3σ] = [100 ± 75] = [25 ; 175] mcg/dl



c) Equação da variável normal padrão Z: $\boxed{Z = \frac{X - \mu}{\sigma}}$.

Neste caso, $\boxed{Z = \frac{X - 100}{25}}$

d)

Um valor de X igual a ...	Corresponde a Z igual a ...	P(X ≤ x) =
25	$\frac{25-100}{25}=-3$	P(Z ≤ -3) = 0,0014
100 – 2 x 25 = 50	-2	P(Z ≤ -2) = 0,0228
100 – 1 x 25 = 75	Da tabela Z = -1	0,1587
59	$\frac{59-100}{25}=-1,6$	P(Z ≤ -1,64) = 0,0505
100 – 1,28 x 25 = 68	-1,28	P(Z ≤ -1,28) = 0,1001
100 – 0,68 x 25 = 83	Da tabela Z = -0,68	0,2483
90	$\frac{90-100}{25}=-0,4$	P(Z ≤ -0,4) = 0,3446

Um valor de X igual a ...	Corresponde a Z igual a ...	P(X ≥ x) =
100 + 3 x 25 = 175	3	P(Z≥3) = P(Z≤-3) =0,0014
100 + 2 x 25 = 150	Da tabela Z = 2	0,0228
125	$\frac{125-100}{25}=1$	P(Z≥1) = P(Z≤-1) = 0,1587
100 + 1,64 x 25 = 141	1,64	P(Z≥1,64)=P(Z≤-1,64) = 0,0505
100 + 1,28 x 25 = 132	Da tabela Z = 1,28	0,1001
117	$\frac{117-100}{25}=0,6$	P(Z≥0,68) = P(Z≤-0,68) 0,2483
100 + 0,4 x 25 = 110	0,4	P(Z≥0,4) = P(Z≤-0,4) = 0,3446

(i)

$$P(68 \le X \le 110) = P\left(\frac{68-100}{25} \le Z \le \frac{110-100}{25}\right) = P(-1,28 \le Z \le 0,40)$$

$$= P(Z \le 0,40) - P(Z \le -1,28) = 0,6554 - 0,1005 = 0,5551$$

Em uma amostra aleatória de 500 indivíduos sadios desta população, esperamos que
500 x 0,555149 ≈ 276
tenham quantidade de ferro sério entre 68 e 110 mcg/dl.

(ii)

Devemos ter $P(x_1 < X < x_2)= 0,50$ e $(100 – x_1) = (x_2 – 100$??pela simetria em torno da média.

$$P(x_1 < X < x_2) = P\left(\frac{x_1-100}{25} < Z < \frac{x_2-100}{25}\right) = P(z_1 < Z < z_2) = 0,50$$

Pela simetria de Z em torno da média, $z_1= -z_2$ e então: $P(-z_2 < Z < z_2) = 0,50$

$$P(Z > z_2) = P(Z < -z_2) = \frac{1-0,50}{2} = \frac{0,50}{2} = 0,25$$

Na Tabela Z vemos que P(Z < -0,67) = 0,251429 ≈ 0,25. Portanto, $-z_2 = -0{,}67$ e $z_2 = 0{,}67$.

Assim:

$$\frac{x_1 - 100}{25} = -z_2 \Rightarrow x_1 = 100 - z_2 \cdot 25 = 100 - 0{,}67 \cdot 25 = 100 - 17 = 83$$

$$\frac{x_2 - 100}{25} = z_2 \Rightarrow x_2 = 100 + z_2 \cdot 25 = 100 + 0{,}67 \cdot 25 = 100 + 17 = 117$$

Portanto, o intervalo simétrico em torno da média 100 que abrange 50% das quantidades de ferro em indivíduos sadios é [83;117] mcg/dl.

(iii)

Devemos ter $P(x_1 < X < x_2) = 0{,}95$ e $(100 - x_1) = (x_2 - 100)$ pela simetria em torno da média.

$$P\left(x_1 < X < x_2\right) = P\left(\frac{x_1 - 100}{25} < Z < \frac{x_2 - 100}{25}\right) = P\left(z_1 < Z < z_2\right) = 0{,}95$$

Pela simetria de Z em torno da média, $z_1 = -z_2$ e então: $P\left(-z_2 < Z < z_2\right) = 0{,}95$

$$P(Z > z_2) = P(Z < -z_2) = \frac{1 - 0{,}95}{2} = \frac{0{,}05}{2} = 0{,}025$$

Na Tabela Z vemos que P(Z < -1,96) = 0,024998 ≈ 0,025. Portanto, $-z_2 = -1{,}96$ e $z_2 = 1{,}96$.

Assim:

$$\frac{x_1 - 100}{25} = -z_2 \Rightarrow x_1 = 100 - z_2 \cdot 25 = 100 - 1{,}96 \cdot 25 = 100 - 49 = 51$$

$$\frac{x_2 - 100}{25} = z_2 \Rightarrow x_2 = 100 + z_2 \cdot 25 = 100 + 01{,}96 \cdot 25 = 100 + 49 = 149$$

Portanto, o intervalo simétrico em torno da média 100 que abrange 95% das quantidades de ferro em indivíduos sadios é [51;149] mcg/dl.

Exercício (6.2)

X: concentração sérica de tiroxina T4(D) em cães machos sadios
X ~ Normal (μ = 2,04 ; σ = 0,78)

a)

i) $P(X < 2{,}81) = P\left(Z < \frac{2{,}81 - 2{,}04}{0{,}78}\right) = P(Z < 0{,}99) = 0{,}8390$

ii) $P(X > 1{,}8) = P\left(Z > \frac{1{,}8 - 2{,}04}{0{,}78}\right) = P(Z > -0{,}31) = P(Z < 0{,}31) = 0{,}6217$

iii) $P(1{,}01 \le X \le 2{,}50) = P\left(\frac{1{,}01 - 2{,}04}{0{,}78} \le Z \le \frac{2{,}50 - 2{,}04}{0{,}78}\right) = P(-1{,}32 \le Z \le 0{,}59)$

$= P(Z \le 0{,}59) - P(Z \le -1{,}32) = 0{,}7224 - 0{,}0934 = 0{,}6290$

b)

$$P(2{,}20 \le X \le 3{,}80) = P\left(\frac{2{,}20 - 2{,}04}{0{,}78} \le Z \le \frac{3{,}80 - 2{,}04}{0{,}78}\right) = P(0{,}21 \le Z \le 2{,}26)$$

$$= P(Z \le 2{,}26) - P(Z \le 0{,}21) = 0{,}9881 - 0{,}5832 = 0{,}4049$$

Assim, em 200 cães sadios, esperamos 200 x 0,4049≈ 81 cães com concentração sérica entre 2,20 e 3,80 mcg/100ml.

c)

Devemos ter $P(x_1 < X < x_2) = 0{,}98$ e $(2{,}04 - x_1) = (x_2 - 2{,}04)$ pela simetria em torno da média.

$$P\left(x_1 < X < x_2\right) = P\left(\frac{x_1 - 2{,}04}{0{,}78} < Z < \frac{x_2 - 2{,}04}{20{,}78}\right) = P\left(z_1 < Z < z_2\right) = 0{,}98$$

Pela simetria de Z em torno da média, $z_1 = -z_2$ e então: $P\left(-z_2 < Z < z_2\right) = 0{,}98$.

$$P(Z > z_2) = P(Z < -z_2) = \frac{1 - 0{,}98}{2} = \frac{0{,}02}{2} = 0{,}01$$

Na Tabela Z vemos que $P(Z < -2{,}33) = 0{,}0099 \approx 0{,}01$. Portanto, $-z_2 = -2{,}33$ e $z_2 = 2{,}33$.

Assim:

$$\frac{x_1 - 2{,}04}{0{,}78} = -z_2 \Rightarrow x_1 = 2{,}04 - z_2 \cdot 0{,}78 = 2{,}04 - 2{,}33 \cdot 0{,}78 = 2{,}04 - 1{,}82 = 0{,}22$$

$$\frac{x_2 - 2{,}04}{0{,}78} = z_2 \Rightarrow x_2 = 2{,}04 + z_2 \cdot 0{,}78 = 2{,}04 + 2{,}33 \cdot 0{,}78 = 2{,}04 + 1{,}82 = 3{,}86$$

Portanto, o intervalo simétrico em torno da média 2,04 que abrange 98% das concentrações séricas em cães sadios é [0,22 ; 3,86] mcg/100ml.

Exercício (6.3)

X: tempo de gestação de um bebê, em dias.
X ~ Normal ($\mu = 268$; $\sigma = 15$)

a) Uma criança será considerada prematura se o seu tempo de gestação for inferior a 247 dias, ou seja, se X<247. Então, devemos calcular P[X<247].

$$P(X < 247) \xrightarrow[padronixando]{} = P\left(\frac{X - 268}{15} < \frac{247 - 268}{15}\right) = P(Z < -1{,}40) = 0{,}0808$$

Assim, a porcentagem de crianças nascidas prematuramente será de 8,08% .

b) Para que uma criança não seja considerada prematura, seu tempo de gestação tem que ser maior do que um valor, digamos a . E o valor a é aquele que deixa 4% dos tempos de gestação abaixo dele. Ou seja, o valor a é tal que P(X < a) = 0,04. O valor a é o percentil de ordem 4 (ou 4%).

Encontrando o valor de a

$$P(X < a) \xrightarrow[padronixando]{} = P\left(\frac{X - 268}{15} < \frac{a - 268}{15}\right) = P\left(Z < \frac{\text{a-268}}{15}\right) = 0{,}04$$

Na tabela Z, o valor que deixa, aproximadamente, 4% da área abaixo dele é o -1,75.
Assim, (a - 268)/15 = -1,75. Então, a = (-1,75 *x* 15) + 268 = 241,75.

Desse modo, para que uma criança não seja considerada prematura pelo novo critério, seu tempo de gestação tem que ser, no mínimo, de 242 dias.

Exercício (6.4)

X: escore de Q.I.
X ~ Normal (μ = 100 ; σ = 15)

a) Uma pessoa será admitida nessa organização se o seu escore de Q.I. for superior a 131,5 pontos, ou seja, se X>131,5. Então, devemos calcular P[X>131,5].

$$P(X > 131,5) \xrightarrow{\text{padronizando}} = P\left(\frac{X-100}{15} > \frac{131,5-100}{15}\right) =$$
$$¿\, P(Z > 2,10) \xrightarrow{\text{por simetria}} = P(Z < -2,10) \xrightarrow{\text{na tabela}} = 0,0179$$

Assim, a probabilidade de uma pessoa escolhida aleatoriamente ser aceita nessa organização é de apenas 1,79%.

b) Para que uma pessoa seja considerada um "gênio", ela deve ter um Q.I superior a um valor, digamos b . Esse valor b é aquele que deixa 1% dos escores de Q.I. acima dele. Ou seja, o valor b é tal que P(X > b) = 0,01 e, portanto, P(X < b) = 0,99. O valor b é o percentil de ordem 99 (ou 99%).
Encontrando o valor b

$$P(X < b) \xrightarrow{\textit{padronixando}} = P\left(\frac{X-100}{15} < \frac{b-100}{15}\right) = P\left(Z < \frac{\text{b-100}}{15}\right) = 0,99$$

O valor, na tabela Z, que deixa, aproximadamente, 99% da área abaixo dele é o 2,33.
Assim, (b - 100)/15 = 2,33. Então, b = (2,33 x 15) + 100 = 134,95.

Desse modo, para que uma pessoa seja considerada um "gênio", ela deve ter um Q.I superior a 134,95 pontos.

Exercício (6.5)

X: número de colônias por cultura de escarro em pessoas sadias

a) Ramo-e-Folhas para número colônias por cultura de escarro

n = 44 Escala: |1|7 = 17

|1|7
|2|23333444444
|2|55555556889
|3|0011
|3|5556
|4|01112
|4|
|5|14
|5|668
|6|0
|6|8
|7|
|7|9

Podemos ver que a distribuição de freqüência do número de colônias por cultura de escarro é muito assimétrica. Portanto, não podemos supor que siga uma esta variável siga a distribuição Normal

b)

Faixa de Referência de 95%

$(100- \alpha)\% = 95\%$
$\alpha = 5$
$\alpha/2 = 5/2 = 2{,}5$
$100 - \alpha/2 = 100 - 2{,}5 = 97{,}5$

(i) Método dos Percentis

Calculada no exercício 2.2.5.

Portanto, a Faixa de Referência de 95% para o número de colônias por cultura de escarro em pessoas sadias é [22; 68] colônias, pelo método dos percentis.

(ii) Método da Curva de Gauss

Faixa de Referência de 100(1- α)% : $[\mu - z_{(\alpha/2)} \cdot \sigma \,;\, \mu + z_{(\alpha/2)} \cdot \sigma]$

Como μ e s são desconhecidos, vamos estimá-los por $\bar{x}$ = 34,3 e s =14,2 e, assim, a Faixa de Referência de 95% torna-se: $[\bar{x} - z_{(\alpha/2)} \cdot s \,;\, \bar{x} - z_{(s/2)} \cdot s] = [34{,}3 - z_{(0{,}025)}(14{,}2) \,;\, 34{,}3 + z_{(0{,}025)}(14{,}2)]$

Vamos descobrir na Tabela da Distribuição Normal Padronizada (Tabela Z) quem é $z_{(0{,}025)}$:
Pela definição, $z_{(0{,}025)}$ é o valor de Z tal que $P(Z > z_{(0{,}025)}) = 0{,}025$.
Pela simetria em torno de zero da distribuição de Z, temos: $P(Z < -z_{(0{,}025)}) = 0{,}025$.
Na tabela Z temos que $P(Z < -1{,}96) = 0{,}024998 \cong 0{,}025$
Portanto, $-z_{(0{,}025)} = -1{,}96$ e assim, $z_{(0{,}025)} = 1{,}96$.

A Faixa de Referência de 95% torna-se:
$[34{,}3 - z_{(0{,}025)}(14{,}2) \,;\, 34{,}3 + z_{(0{,}025)}(14{,}2)] = [34{,}3 - 1{,}96(14{,}2) \,;\, 34{,}3 + 1{,}96(14{,}2)] = [6 \,;\, 62]$ colônias

Portanto, a Faixa de Referência de 95% para o número de colônias por cultura de escarro em pessoas sadias é [6; 62], pelo método da curva de Gauss.

Note que as Faixas de Referência de 95% para o número de colônias por cultura de escarro em pessoas sadias fornecida pelo dois métodos diferem bastante no limite inferior:

Método da Curva de Gauss: FR(95%) : [6; 62] colônias
Método dos Percentis: FR(95%): [22 ; 68] colônias

Como vimos no item (a) a Curva Normal (Curva de Gauss) não se ajusta bem a esses dados. Assim, o método mais indicado neste caso é o Método dos Percentis.

Seção 7: Intervalos de Confiança

Exercício (7.1)

a) População: todos os pacientes atendidos no posto durante os últimos três anos.
Amostra: os 70 pacientes adultos selecionados.

b) Variável X: idade dos pacientes desta população (variável contínua)
Parâmetro μ : <u>idade média</u> dos pacientes desta população

- <u>Estimativa Pontual para μ</u>: $x = 36{,}86$ anos (média amostral)

Intervalo de 100(1-α)% de Confiança para μ: $\left[\bar{x} - Z_{(\alpha/2)} \cdot \frac{s}{\sqrt{n}}\,;\, \bar{x} + Z_{(\alpha/2)} \cdot \frac{s}{\sqrt{n}}\right]$ com 0<α<1

s = 17,79 anos (desvio padrão amostral)
n = 70 (tamanho da amostra)

- <u>Intervalo de 90% de Confiança para μ</u>: (a amostra é grande n = 70 > 30)

100(1- α) = 90
1- α = 0,90
α = 0,10
α/2 = 0,05
$Z_{0,05}$ = 1,64 pois P(Z > 1,64) ≈ 0,05

$$IC_{\mu}^{90} = \left[\bar{x} - Z_{(0,05)} \cdot \frac{s}{\sqrt{n}}\,;\, \bar{x} + Z_{(0,05)} \cdot \frac{s}{\sqrt{n}}\right]$$
$$\text{¿}\left[36,86 - 1,64 \cdot \frac{17,79}{\sqrt{70}}\,;\, 36,86 + 1,64 \cdot \frac{17,79}{\sqrt{70}}\right]$$
$$\text{¿}[36,86 - 3,49\,;\, 36,86 + 3,49] = [33,37\,;\, 40,35]$$

Interpretação: A idade de média dos adultos que freqüentam o posto está entre 33,37 e 40,35 anos com 90% de confiança.

- <u>Intervalo de 95% de Confiança para μ</u> :

100(1-α) = 95
1- α = 0,95
α = 0,05
α /2 = 0,025
$Z_{0,025}$ = 1,96 pois P(Z > 1,96) ≈ 0,025

$$IC_{\mu}^{95} = \left[\bar{x} - Z_{(0,025)} \cdot \frac{s}{\sqrt{n}}\,;\, \bar{x} + Z_{(0,025)} \cdot \frac{s}{\sqrt{n}}\right]$$
$$\text{¿}\left[36,86 - 1,96 \cdot \frac{17,79}{\sqrt{70}}\,;\, 36,86 + 1,96 \cdot \frac{17,79}{\sqrt{70}}\right]$$
$$\text{¿}[36,86 - 4,17\,;\, 36,86 + 4,17] = [32,69\,;\, 41,03]$$

Interpretação: A idade de média dos adultos que freqüentam o posto está entre 32,69 e 41,03 anos com 95% de confiança.

Comparando os dois intervalos: o IC de 95% é mais amplo que o IC de 90%.

c) Variável Y: ser ou não analfabeto (variável categórica)
Parâmetro p: proporção de analfabetos nesta população

- <u>Estimativa Pontual para p</u>: $p = 19/70 = 0,27$
  ou seja, 27% dos pacientes deste posto são analfabetos

- <u>Intervalo de 90% de Confiança para p</u>: (a amostra é grande n = 70 > 30)

100(1- α) = 90
1- α = 0,90
α = 0,10
α/2 = 0,05
$Z_{0,05}$ = 1,64 pois P(Z > 1,64) ≈ 0,05

$$IC_p^{90} = \left[\hat{p} \pm Z_{(\alpha/2)}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[0,27 \pm 1,64\sqrt{\frac{0,27(0,73)}{70}}\right]$$
$$¿\left[0,27 \pm 0,09\right] = \left[0,18\,;0,36\right]$$

Interpretação: A proporção de frequentadores adultos do posto que são analfabetos está entre 18% e 36% com 90% de confiança.

- <u>Intervalo de 95% de Confiança para p</u>:

100(1-α) = 95
1- α = 0,95
α = 0,05
α /2 = 0,025
$Z_{0,025}$ = 1,96 pois P(Z > 1,96) ≈ 0,025

$$IC_p^{95} = \left[0,27 \pm 1,96\sqrt{\frac{0,27(0,73)}{70}}\right]$$
$$= \left[0,27 \pm 0,10\right] = \left[0,17\,;0,37\right]$$

Interpretação: A proporção de frequentadores adultos do posto que são analfabetos está entre 17% e 37% com 95% de confiança.

Comparando os dois intervalos: o IC de 95% é mais amplo que o IC de 90%.

Exercício (7.2)

Variável X: produção de leite (em litros) na primeira lactação de vacas desta fazenda
(variável contínua)
Parâmetro μ : <u>produção média</u> de leite na primeira lactação de vacas desta fazenda

- <u>Estimativa Pontual para μ</u> : $x = 1500$ litros (média amostral)

s = 300 litros (desvio padrão amostral)
n = 20 (tamanho da amostra)

- <u>Intervalo de 98% de Confiança para μ :</u> (a amostra é pequena n = 20 < 30)

100(1-α) = 98
1-α = 0,98
α = 0,02
α/2 = 0,01
gl = n-1 = 20-1 =19
$t_{(19;\,0,01)}$ = 2,539

$$IC_\mu^{98} = \left[\bar{x} \pm t_{19;\,0,01}\cdot\frac{s}{\sqrt{n}}\right] = \left[1500 \pm 2,539\cdot\frac{300}{\sqrt{20}}\right]$$
$$= \left[1500 \pm 170\right] = \left[1330\,;1670\right] \text{ litros .}$$

Interpretação: A produção média de leite na primeira lactação de vacas desta fazenda está entre 1330 e 1670 litros, com 98% de confiança.

Exercício (7.3)

a) Porque a informação sobre o parâmetro (neste caso, a média da variável na população) é muito vaga. Por exemplo, se o intervalo de confiança para a altura média das pessoas adultas de uma cidade vai de 1 a 2,50 metros, não saberemos dizer se esta cidade é habitada por adultos que tendem a ser altos ou baixos. Além disso, não seria necessário um estudo para obter esta informação, pois já sabemos que os adultos não medem menos que um metro nem mais que 2,50 metros.

b) Reduzir o nível de confiança $100(1-\alpha)$ significa aumentar o valor de α, o que leva a valor menor de $z_{\alpha/2}$, reduzindo a amplitude do intervalo.
Por exemplo: passar de $100(1-\alpha) = 95$ para $100(1-\alpha) = 90$ significa passar de $\alpha = 0{,}05$ para $\alpha = 0{,}10$.
Assim, passamos de $Z_{\alpha/2} = Z_{0,025} = 1{,}96$ para $Z_{\alpha/2} = Z_{0,05} = 1{,}64$.
Desse modo, quanto menor o nível de confiança requerido, menor será a amplitude do intervalo.

c) Esta parcela seria menor. Desse modo, quanto menor a variabilidade da variável, menor será a amplitude do intervalo.

d) Esta parcela seria menor. Desse modo, quanto maior o tamanho da amostra, menor será a amplitude do intervalo.

e) Apenas o tamanho da amostra (n), que ele pode aumentar, e o nível de confiança do intervalo ($100(1-\alpha)$), que ele pode reduzir. O pesquisador não tem controle sobre a variabilidade (σ) da característica estudada.

f) Aumentar o tamanho da amostra (n).

Seção 8: Conceitos Básicos de Testes de Hipóteses

Situação 1:

a) Parâmetro a ser testado μ : tempo médio (em horas) da execução de uma nova técnica para identificar bactérias em hemoculturas.

b) Hipótese Nula H_0: μ = 40,5 horas
Hipótese Alternativa H_a: μ < 40,5 horas

c) Erro Tipo I: Concluir que o novo método tem tempo médio de execução menor do que 40,5 horas quando, na verdade, seu tempo médio é igual a 40,5 horas.
Erro Tipo II: Concluir que novo método tem tempo médio de execução igual a 40,5 horas Quando, na verdade, seu tempo médio é menor do que 40,5 horas.

Situação 2:

a) Parâmetro a ser testado p: proporção de homens com mais de 65 anos de uma cidade que morrem dentro de um ano.

b) Hipótese Nula H_0: p = 0,04
Hipótese Alternativa H_a: p > 0,04

c) Erro Tipo I: Concluir que a proporção de homens com mais de 65 anos que morrem dentro de um ano nessa cidade é maior do que 0,04 quando, na verdade, essa proporção é igual a 0,04.
Erro Tipo II: Concluir que a proporção de homens com mais de 65 anos que morrem dentro de um ano nessa cidade é igual a 0,04 quando, na verdade, essa proporção aumentou.

Situação 3:

a) Parâmetro a ser testado μ: peso médio (em quilos) de frangos vendidos pelo fornecedor novo

b) Hipótese Nula H_0: μ = 3 quilos
Hipótese Alternativa H_a: μ > 3 quilos

c) Erro Tipo I: Concluir que o peso médio dos frangos do novo fornecedor é maior do que 3 quilos quando, na verdade, o peso médio é igual a 3 quilos.
Erro Tipo II: Concluir que o peso médio dos frangos do novo fornecedor é igual a 3 quilos quando, na verdade, peso médio é maior do que 3 quilos.

Situação 4:

a) Parâmetro a ser testado μ : quantidade média de ácido acetil salicílico (mg por comprimido) de certo analgésico.

b) Hipótese Nula H_0: $\mu = 5{,}5$ mg
Hipótese Alternativa H_a: $\mu \neq 5{,}5$ mg

c) Erro Tipo I: Concluir que a quantidade média de ácido acetil salicílico nos comprimidos do analgésico é diferente da especificada quando, na verdade, ela é igual a 5,5 mg.
d) Erro Tipo II: Concluir que a quantidade média de ácido acetil salicílico nos comprimidos do analgésico é igual à especificada quando, na verdade, ela é diferente de 5,5 mg.

Situação 5:

a) Parâmetro a ser testado: p: proporção de sementes que germinam

b) Hipótese Nula H_0: $p = 0{,}95$
Hipótese Alternativa H_a: $p < 0{,}95$

c) Erro Tipo I: Concluir que a proporção de sementes que germinam é menor de 0,95 quando, na verdade, essa proporção é igual a 0,95.
Erro Tipo II: Concluir que a proporção de sementes que germinam é igual a 0,95 quando, na verdade, essa proporção é menor do que 0,95.

Seção 9: Testes de Hipóteses para Uma População

Situação 1:

H_0: μ = 40,5 horas n = 18 (amostra pequena)
H_a: μ < 40,5 horas média amostral = 39,42;
desvio-padrão amostral = 1,96

$$T_{obs} = \frac{39,42 - 40,50}{1,96/\sqrt{18}} = -2,34$$

RR: $T_{obs} < -t_{(17;0,05)}$ que seja $T_{obs} < -1,74$

Como o valor de T_{obs} está na Região de Rejeição, rejeitamos H_0, ao nível de significância de 5%.

Valor P = P[t_{17} < -2,34] = ?
Deve-se encontrar a P[t_{17} > 2,34], que, por simetria, é igual a P[t_{17} < -2,34]).
Na linha 17 da tabela t-student, não existe o valor 2,34. Ele está entre os valores 2,110 e 2,567, que correspondem às colunas 0,025 e 0,01, respectivamente. Assim, P[t_{17} > 2,34] está entre 0,01 e 0,025. Conseqüentemente, o valor P está entre 1% e 2,5%.

Conclusão em termos do problema: Rejeita-se a hipótese de que o tempo médio de execução do novo método é igual a 40,5 horas, em favor da hipótese de que ele é menor do que 40,5 horas, ao nível de significância de 5% (0,01 < valor P < 0,025).

Situação 2:

H_0: p = 0,04
H_a: p > 0,04 n = 1000 (amostra grande); proporção amostral = 60/1000 = 0,06

$$Z_{obs} = \frac{0,06 - 0,04}{\sqrt{\frac{0,06(1-0,06)}{1000}}} = 2,66$$

RR: $Z_{obs} > z_{0,05}$

$Z_{obs} > 1,64$

Como o valor de Z_{obs} (2,66) está na Região de Rejeição, rejeitamos H_0, ao nível de significância de 5%.

Valor P = P[Z > 2,66] = P[Z < -2,66] = 0,0039

Conclusão em termos do problema: rejeitamos a hipótese de que a proporção de idosos que morrem por ano nessa cidade é igual a 4%, em favor da hipótese de que essa proporção é maior 4%, ao nível de significância de 5% (valor P = 0,0039).

Para pensar: a hipótese nula também seria rejeitada a 1% de significância?

Obs: a resolução deste exercício usa a *Estatística de Teste de Wald*: $$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

O resultado da estatística de teste alternativa ao teste de Wald $\left(Z_{obs} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \right)$ é

$$Z_{obs} = \frac{0,06 - 0,04}{\sqrt{\dfrac{0,04(1-0,04)}{1000}}} = 3,23$$

Valor P = P[Z > 3,23] < 0,0014 (pois o maior valor tabela é 2,99)

<u>Conclusão em termos do problema:</u> Ao nível de 5% de significância, não rejeitamos a hipótese de que a proporção de sementes que germinam é igual a 95% (valor P < 0,0014).

Situação 3:

H_0: μ = 3 quilos
H_a: μ > 3 quilos   n = 25 (amostra pequena); média amostral = 3,2; desvio-padrão amostral = 0,4

$$T_{obs} = \frac{3,2 - 3,0}{0,4/\sqrt{25}} = 2,5$$

RR: $T_{obs} > t_{24;0,05}$

$T_{obs} > 1,711$

Como o valor de T_{obs} está na Região de Rejeição, rejeitamos H_0, ao nível de significância de 5%.

Valor P = P[t_{24} > 2,5] = ?
Na linha 24 da tabela t-student, não existe o valor 2,5. Ele está entre as colunas do 0,01 e 0,005. Assim, 0,005 < P[t_{24} > 2,5] < 0,01. Conseqüentemente, 0,005 < valor P < 0,01.

<u>Conclusão em termos do problema:</u> Ao nível de significância de 5%, rejeitamos a hipótese de que o peso médio de frangos vendidos pelo fornecedor novo é igual a 3 quilos, em favor da hipótese de que o peso médio desses frangos é maior do que 3 quilos (0,005 < valor P < 0,01).

Situação 4:

H_0: μ = 5,5 mg
H_a: $\mu \neq$ 5,5 mg   n = 40 (amostra grande); média amostral = 5,2 ; desvio-padrão amostral = 0,7

$$Z_{obs} = \frac{5,2 - 5,5}{0,7/\sqrt{40}} = -2,71$$

RR: $Z_{obs} < -z_{0,025}$ ou $Z_{obs} > z_{0,025}$

$Z_{obs} < -1,96$ ou $Z_{obs} > 1,96$

Como o valor de Z_{obs} está na Região de Rejeição, rejeitamos H_0, ao nível de significância de 5%.

Valor P = 2 x P[Z > |-2,71|] = 2 x P[Z > 2,71] = (por simetria) 2 x P[Z <-2,71] = 2 x 0,0034 = 0,0068.

<u>Conclusão em termos do problema:</u> rejeitamos a hipótese de que a quantidade média de ácido acetil salicílico (mg por comprimido) de certo analgésico é igual a 5,5 mg ao nível de significância de 5% (valor P = 0,0068).

Nessa situação, podemos usar o intervalo de confiança para realizar o teste de hipóteses, pois a hipótese alternativa é bilateral. Como queremos um teste a 5% de significância, calcularemos um intervalo de 95% de confiança para a quantidade média de ácido acetil salicílico, por comprimido.

Intervalo de 95% de Confiança para μ : (a amostra é grande n = 40 > 30)

100(1-α) = 95
1-α = 0,95
α = 0,05
α/2 = 0,025
$Z_{0,025}$ = 1,96

$$IC_m^{95} = \left[\bar{x} \pm Z_{0,025} \cdot \frac{s}{\sqrt{n}}\right]$$

$$IC_m^{95} = \left[5,2 \pm 1,96 \cdot \frac{0,7}{\sqrt{40}}\right] = [5,2 \pm 0,2]$$

$$IC_m^{95} = [5,0 \;\; ; 5,4]$$

Interpretação: A quantidade média de ácido acetil salicílico, por comprimido, está entre 5,0 e 5,4 mg, com 95% de confiança.

Teste de hipóteses baseado no intervalo de confiança: como o valor 5,5 não pertence ao intervalo de 95% de confiança para a quantidade média de ácido acetil salicílico, por comprimido, rejeitamos a hipótese de que a quantidade média de ácido acetil salicílico de certo analgésico é igual a 5,5 mg ao nível de significância de 5%.

Situação 5:

H_0: p = 0,95
H_a: p < 0,95 n = 1000 (amostra grande); proporção amostral = 940/1000 = 0,94

$$Z_{obs} = \frac{0,94 - 0,95}{\sqrt{\frac{0,94(1-0,94)}{1000}}} = -1,33$$

RR: $Z_{obs} < -z_{0,05}$

$Z_{obs} < -1,64$

Como o valor de Z_{obs} não está na Região de Rejeição, não rejeitamos H_0, ao nível de significância de 5%.

Valor P = P[Z < -1,33] = 0, 0918

Conclusão em termos do problema: Ao nível de 5% de significância, não rejeitamos a hipótese de que a proporção de sementes que germinam é igual a 95% (valor P=0, 0918).

Para pensar: a hipótese nula seria rejeitada a 1% de significância?

Obs: a resolução deste exercício usa a *Estatística de Teste de Wald*: $Z_{obs} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$

O resultado da estatística de teste alternativa ao teste de Wald $\left(Z_{obs} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \right)$ é

$$Z_{obs} = \frac{0,94 - 0,95}{\sqrt{\frac{0,95(1-0,95)}{1000}}} = -1,45$$

Valor P = P[Z < -1,45] = 0,0735

Conclusão em termos do problema: Ao nível de 5% de significância, não rejeitamos a hipótese de que a proporção de sementes que germinam é igual a 95% (valor P=0,0735).

Seção 10: Testes de Hipóteses para Duas Populações

Exercício (10.1)

a)
H0: a média da diferença dos escores depois e antes do programa é igual a zero (o programa não funciona);
Ha: a média da diferença dos escores depois e antes do programa é maior do que zero (o programa aumenta a habilidade);

Nº do aluno	Escores de abdominais Antes	Depois	Diferenças (d) (Depois – Antes)
1	12	15	15 - 12 = **3**
2	10	9	9 - 10 = **-1**
3	23	25	25 - 23 = **2**
4	25	25	25 - 25 = **0**
5	29	31	31 - 29 = **2**
6	32	30	30 - 32 = **-2**
7	14	16	16 - 14 = **2**
8	17	20	20 - 17 = **3**
9	19	22	22 - 19 = **3**
10	20	22	22 - 20 = **2**

$\bar{d} = 1{,}4$

$s_d = 1{,}8$

Média das diferenças: $\bar{d} = \dfrac{3-1+2+0+2-2+2+3+3+2}{10} = \dfrac{14}{10} = 1{,}4$

Desvio padrão das diferenças:
$$s_d = \sqrt{\frac{(3-1{,}4)^2 + (-1-1{,}4)^2 + \ldots + (3-1{,}4)^2 + (2-1{,}4)^2}{10-1}} = 1{,}8$$

$$T_{obs} = \frac{1{,}4-0}{1{,}8/\sqrt{10}} = 2{,}45$$

RR: $T_{obs} > t_{(10-1);0,05} = 1{,}833$

Como T_{obs} está na região de rejeição, rejeitamos H_0 ao nível de 5% de significância.

b) Valor P = P[t_9>2,45] = ? O valor 2,45 não existe na linha 9 da tabela t-Student. Ele está entre os valores 2,262 e 2,821, correspondentes às colunas 0,025 e 0,01. Assim, a P[t_9>2,45] está entre 1% e 2,5% .
Como o valor P é menor do que o nível de significância (5%), rejeitamos H0.

<u>Conclusão em termos do problema:</u> Ao nível de 5% de significância, rejeitamos a hipótese de que a média dos escores depois do programa é igual à média dos escores antes do programa, em favor da hipótese de que a média dos escores depois do programa é maior do que média dos escores antes do programa (0,01< valor p < 0,025).

c) $100(1-\alpha)\% = 95\%$; $1-\alpha = 0{,}95$; $\alpha=0{,}05$; $\alpha/2 = 0{,}025$;
$t_{9;0,025} = 2{,}262$

$$[1{,}4 \pm 2{,}262 \times 1{,}8/\sqrt{10}] = [1{,}4 \pm 2{,}262 \times 0{,}57] = [1{,}4 \pm 1{,}28]$$

[0,12 ; 2,68]

A diferença média entre os escores depois e antes do programa está entre 0,12 e 2,68, com 95% de confiança.

Exercício (10.2)

$\mu 1$: peso médio no grupo I;
$\mu 2$: peso médio no grupo II;

a)
H0: $\mu 1 = \mu 2$
Ha: $\mu 1 > \mu 2$

$\overline{x_1} = 120$ s = 21,39 n1 = 12
$\overline{x_2} = 101$ s = 20,62 n2 = 7

$$T_{obs} = \frac{120 - 101}{21,12\sqrt{\frac{1}{12} + \frac{1}{7}}} = 1,89 \qquad S_{comb} = \sqrt{\frac{(12-1)(21,29)^2 + (7-1)(20,62)^2}{12+7-2}} = 21,12$$

RR: $T_{obs} > t_{(12+7-2);0,01}$
$T_{obs} > 2,57$

Como T_{obs} não está na região de rejeição, não rejeitamos H_0, ao nível de 1% de significância.

<u>Conclusão em termos do problema</u>: Ao nível de 1% de significância, não rejeitamos a hipótese de que o ganho médio de peso no grupo alimentado com dieta de alto conteúdo protéico é igual ao ganho médio de peso no grupo alimentado com dieta de baixo conteúdo protéico.

b) Valor P= $P[t_{17} > 1,89]$ = ? Na linha 17 da tabela t-Student, não existe o valor 1,89. Ele está entre os valores 1,74 e 2,11, correspondentes às colunas 0,05 e 0,025, respectivamente. Assim, $P[t_{17} > 1,89]$ está entre 2,5% e 5%. O valor P está entre 0,025 e 0,05.
Como o valor P é maior do que 1%, não rejeitamos H_0.

<u>Conclusão em termos do problema</u>: Ao nível de 1% de significância, não existem evidências para rejeitarmos a hipótese de que o ganho médio de peso no grupo alimentado com dieta de alto conteúdo protéico é igual ao ganho médio de peso no grupo alimentado com dieta de baixo conteúdo protéico (0,025 < valor P < 0,05).

c) $$\left[(120 - 101) \pm 2,11 \times 21,12\sqrt{\frac{1}{12} + \frac{1}{7}}\right] = [19 \pm 21,18]$$

[-2,18 ; 40,18]

A diferença entre o ganho médio de peso no grupo alimentado com dieta de alto conteúdo protéico e o ganho médio de peso no grupo alimentado com dieta de baixo conteúdo protéico está entre -2,18 gramas e 40,18 gramas, com 95% de confiança.

Observação para todo o exercício: se estivéssemos testando a H_0: $\mu 1 = \mu 2$, a 5% de significância, nós rejeitaríamos H_0 em favor de H_1: $\mu 1 > \mu 2$, pois o valor P < 0,05. Porém, se estivéssemos testando a H_1: $\mu 1 \neq \mu 2$, não teríamos evidências suficientes para rejeitar H_0, a 5% de significância, pois o intervalo de 95% de confiança para a diferença entre as duas médias contém o valor zero. Isso mostra que a não indicação de uma direção para o valor da diferença entre as duas médias (maior ou menor do que zero) reduz o poder do teste para rejeitar H_0. Ou seja, quando o teste é unilateral, estamos fornecendo uma informação a mais para o teste. Sendo assim, não precisaremos de evidências amostrais tão fortes para rejeitar H_0 quanto precisaríamos se o teste fosse bilateral, onde não fornecemos nenhuma informação a mais.

Exercício (10.3)

μ_1 : peso médio dos bebês no grupo de mães que usaram cocaína durante toda a gravidez;
μ_2 : peso médio dos bebês no grupo de mães que não têm história de uso de cocaína;

a)

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 < \mu_2$

$\overline{x_1} = 2829$ $s_1 = 708$ $n_1 = 36$
$\overline{x_2} = 3436$ $s_2 = 628$ $n_2 = 39$

$$T_{obs} = \frac{2829 - 3436}{667.55\sqrt{\frac{1}{36} + \frac{1}{39}}} = \frac{-607}{154.29} = \text{-}3.93 \qquad S_{comb} = \sqrt{\frac{(36-1)(708)^2 + (39-1)(628)^2}{36+39-2}} =$$

RR: $T_{obs} < - t_{(36+39-2);0,05}$
$T_{obs} < - t_{(73);0,05}$

Como não existe g.l. = 73 na tabela t-student, vamos usar a tabela Normal-padrão como aproximação para os percentis da distribuição t-student e também no cálculo do Valor P. Assim,

RR: $T_{obs} < - z_{0,05}$
$T_{obs} < - 1,64$

Como T_{obs} está na região de rejeição, rejeitamos a hipótese de que o peso médio dos bebês no grupo de mães que usaram cocaína durante toda a gravidez é igual ao peso médio dos bebês no grupo de mães que não têm história de uso de cocaína, em favor da hipótese de que o peso médio dos bebês do primeiro grupo de mães é menor do que o peso médio dos bebês do segundo grupo de mães, ao nível de 5 % de significância.

b) Valor P = P[Z < -3,93] < 0,0014 , pois o menor valor da tabela Normal padrão é -2,99, que deixa uma área de 0,0014 abaixo dele.

<u>Conclusão em termos do problema</u>: ao nível de 1% de significância, rejeitamos a hipótese de que o peso médio dos bebês no grupo de mães que usaram cocaína durante toda a gravidez é igual ao peso médio dos bebês no grupo de mães que não têm história de uso de cocaína, em favor da hipótese de que o peso médio dos bebês do primeiro grupo de mães é menor do que o peso médio dos bebês do segundo grupo de mães (valor P < 0,0014).

Exercício (10.4)

p_C : proporção de pessoas que tomaram vitamina C e ficaram livres de doenças do trato respiratório

p_P : proporção de pessoas que tomaram o placebo e ficaram livres de doenças do trato respiratório

H0: $p_C = p_P$ $\qquad \hat{p}_C = \frac{105}{407} = 0,26 \qquad \hat{p}_p = \frac{76}{411} = 0,18$

Ha: $p_C > p_P$

$$Z_{obs} = \frac{0,26 - 0,18}{\sqrt{\frac{0,26(1-0,26)}{407} + \frac{0,18(1-0,18)}{411}}} = \frac{0,07}{0,028} = 2,77$$

RR: $Z_{obs} > z_{0,025}$

RR: $Z_{obs} > 1,64$

Como Z_{obs} está na região de rejeição, rejeitamos H0, ao nível de significância de 5%.

Valor P = P[Z>2,77] = P[Z < -2,77] = 0,0028.

<u>Conclusão em termos do problema</u>: Ao nível de significância de 5%, rejeitamos a hipótese de que a proporção de pessoas que tomaram vitamina C e ficaram livres de doenças do trato respiratório é igual à proporção de pessoas que tomaram placebo e ficaram livres de doenças do trato respiratório, em favor da hipótese de que a proporção de pessoas livres de doenças no trato respiratório entre as que tomaram vitamina C é maior do que essa proporção entre as que tomaram placebo (valor P = 0,0028).

Observação: Também rejeitaríamos a hipótese nula se o nível de significância fosse 1%.

Exercício (10.5)

Para que um desfecho possa ser considerado associado ao uso de reposição hormonal, devemos rejeitar a hipótese nula no seu teste, ou seja, o Valor P de seu teste deve ser menor que o nível de significância, neste caso, 5%.

Assim, "qualquer evento tromboembólico" pode ser considerado, ao nível de significância de 5%, associado ao uso de reposição hormonal (Valor P = 0,002). Em especial, a "trombose venosa profunda" (Valor P = 0,004).

Observações:

1) O evento "doença da vesícula biliar" tem Valor P igual a 5%, e pode ser considerado sob suspeita.
2) Note que 34 das 1.380 (2,46%) mulheres que fizeram uso desses hormônios tiveram algum evento tromboembólico, contra apenas 12 das 1383 (0,87%) mulheres do grupo placebo. Assim, o risco de ter algum evento tromboembólico é 2,46/0,87= 2,84 vezes este risco para as mulheres que usam reposição hormonal. O valor 2,84 é aproximadamente aquele que está na coluna do Risco Relativo (RR). A mesma conclusão pode ser feita para o desfecho "trombose venosa profunda", cujo risco é 218% maior entre as mulheres que fizeram reposição hormonal do que entre as mulheres do grupo placebo.

Seção 11: Teste Qui-Quadrado

Exercício (11.1)

Tabela de contingência para os dados do exercício 11.1

Vacina	Brucelose		Total
	Sim	Não	
Padrão	10	4	14
Nova	5	11	16
Total	15	15	30

Este é um Teste de Homogeneidade, pois deseja comparar a eficácia das vacinas padrão e nova.

H_0: As proporções de animais que contraíram a doença são iguais para as duas vacinas
H_a: As proporções de animais que contraíram a doença são diferentes para as duas vacinas

Tabela de valores esperados sob a hipótese de homogeneidade

Vacina	Brucelose		Total
	Sim	Não	
Padrão	14x15/30 =7	14x15/30=7	14
Nova	16x15/30=8	16x15/30=8	16
Total	15	15	30

Estatística de Teste:

$$X^2 = \frac{(10-7)^2}{7} + \frac{(4-7)^2}{7} + \frac{(5-8)^2}{8} + \frac{(11-8)^2}{8}$$

$$X^2 = 1,286 + 1,286 + 1,125 + 1,125 = 4,821$$

Região de Rejeição: $X^2_{obs} > 3,84$ ($X^2_{1;0,05} = 3,84$)

Verificação: Como 4,821 está na região de rejeição, rejeitamos a hipótese nula ao nível de 5% de significância.

Cálculo do Valor P: Valor P = $P[X^2_1 > 4,821]$ = ?
Na tabela Qui-quadrado, na linha 1, não existe o valor 4,821. Ele está entre os valores 3,84 e 5,024, correspondentes às colunas do 5% e 2,5%, respectivamente. Assim, $P[X^2_1 > 4,821]$ está entre 0,025 e 0,05.

Conclusão em termos do problema: ao nível de significância de 5%, rejeitamos a hipótese de homogeneidade entre as proporções de animais que contraíram a doença nos grupos vacina padrão e vacina nova, em favor da hipótese de que essas proporções são diferentes (0,025 < valor P < 0,05).

Medindo a associação entre variáveis: o Risco Relativo

O risco (probabilidade) de uma bezerra que recebeu a vacina padrão contrair brucelose é estimado em 10/14=0,71. Para as bezerras que receberam a vacina nova, este risco é estimado por 5/16=0,31. Logo, as bezerras vacinadas com vacina padrão apresentam um risco de desenvolver brucelose 2,29 (0,71/0,31) vezes o risco das bezerras vacinadas com a vacina nova. O valor 2,29 é uma estimativa do *risco relativo* (RR). O risco relativo é definido como RR=p_1/p_2, onde p_1 e a probabilidade de ocorrência do evento (brucelose) no grupo 1 (vacina padrão) e p_2 é a probabilidade de ocorrência do evento (brucelose) no grupo 2 (vacina nova). O risco relativo é calculado somente para estudos prospectivos, onde a ocorrência do evento de interesse (brucelose) é precedida pelos tratamentos (vacina nova e vacina padrão). As hipóteses de teste χ^2 podem ser escritas em função de RR como :

$$H_0 : RR = 1 \qquad x \qquad H_a: RR \neq 1,$$

pois, se não existe associação entre as variáveis (H0), os riscos nos dois grupos são iguais e RR=1.

Exercício (11.2)

Teste de Homogeneidade

H_0: As proporções de contaminação são iguais para todos os tipos de fertilizantes
H_a: As proporções de contaminação são diferentes para todos os tipos de fertilizantes

Tabela de valores esperados sob a hipótese de homogeneidade

Fertilizante	Contaminação		Total
	Sim	Não	
Nenhum	101x44/450=9,88	101x406/450=91,12	101
Nitrogênio	95x44/450=9,29	95x406/450=85,71	95
Esterco	113x44/450=11,05	113x406/450=101,95	113
Nitrogênio e Esterco	141x44/450=13,79	141x406/450=127,21	141
Total	44	406	450

Estatística de Teste:

$$X^2 = \frac{(16-9,88)^2}{9,88} + \frac{(85-91,12)^2}{91,12} + \frac{(10-9,29)^2}{9,29} + \frac{(85-85,71)^2}{85,71} + \frac{(4-11,05)^2}{11,05} + \frac{(109-101,95)^2}{101,95} + \frac{(14-13,79)^2}{13,79} + \frac{(127-127,21)^2}{127,21}$$

$$X^2 = 3.798 + 0.412 + 0.054 + 0.006 + 4.497 + 0.487 + 0.003 + 0.000 = 9.258$$

Região de Rejeição: $X^2_{obs} > X^2_{(4-1)(2-1);0,05}$

$X^2_{obs} > X^2_{3;0,05}$

$X^2_{obs} > 7,82$

Verificação: Como 9,258 está na região de rejeição, rejeitamos a hipótese nula ao nível de 5% de significância.

Conclusão em termos do problema: ao nível de significância de 5%, rejeitamos a hipótese de homogeneidade entre as proporções de contaminação dos tipos de fertilizantes estudados, em favor da hipótese de que essas proporções são diferentes (0,025< valor P <0,05).

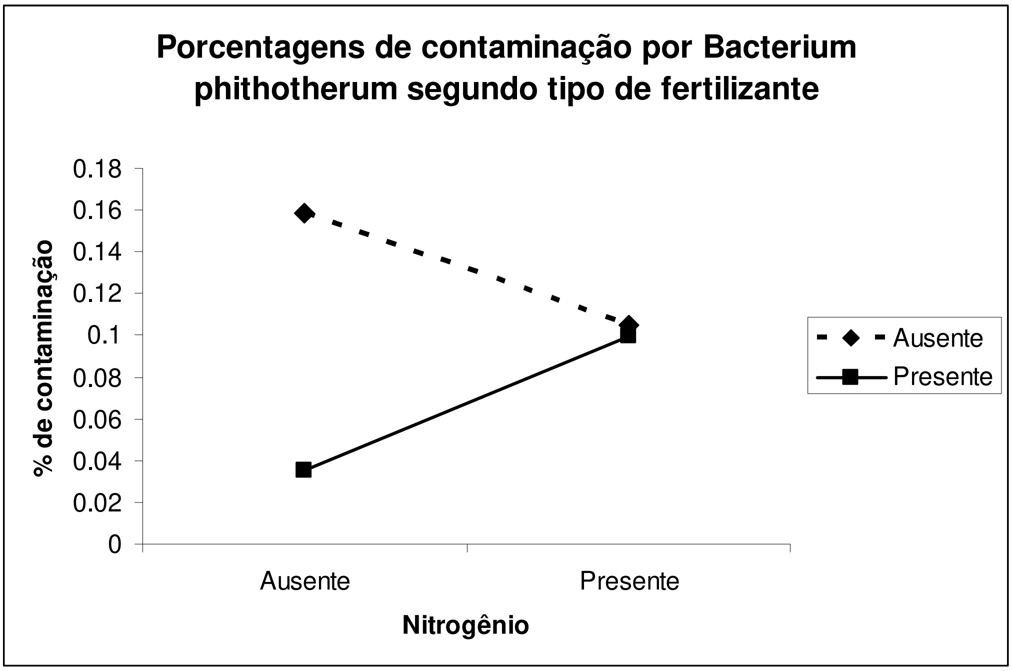
Cálculo do Valor P: Valor P = $P[X^2_3 > 9,258]$ = ?.
Na tabela Qui-quadrado, na linha 3, não existe o valor 9,258.
Ele está entre os valores 7,82 e 9,35, correspondentes às colunas do 5%e 2,5%, respectivamente.
Assim, $P[X^2_3 > 9,258]$ está entre 0,025 e 0,05.

Observação
Os tratamentos "Nenhum", "Nitrogênio", "Esterco" e "Nitrogênio e Esterco" podem ser vistos como a combinação de dois fatores: Nitrogênio (presente e ausente) e Esterco (presente e ausente), pois

				Tratamento
Nitrogênio ausente	+	Esterco ausente	=	Nenhum
Nitrogênio ausente	+	Esterco presente	=	Esterco
Nitrogênio presente	+	Esterco ausente	=	Nitrogênio
Nitrogênio presente	+	Esterco presente	=	Nitrogênio e Esterco

A estrutura desse experimento é chamada *fatorial 2 a 2*, pois existem dois *fatores* (Nitrogênio e Esterco), cada um com dois *níveis* (ausente e presente).
Os experimentos fatoriais são muito comuns nas áreas industrial, agronômica e de experimentação animal.

O gráfico a seguir mostra as porcentagens de contaminação para os 4 tratamentos.



O gráfico indica que, na ausência de esterco (linha pontilhada), a adição de nitrogênio diminui a porcentagem de contaminação, mas, na presença do esterco (linha cheia), a adição do nitrogênio aumenta a porcentagem de contaminação. Em outras palavras, o efeito do nitrogênio é de aumentar a porcentagem de contaminação quanto o esterco está presente e de diminuir essa porcentagem quanto o esterco está ausente. Esse fato caracteriza a *interação* entre os fatores Esterco e Nitrogênio, isto é, o efeito do Nitrogênio na porcentagem de contaminação não é o mesmo nos dois *níveis* de Esterco. Se não houvesse interação entre esse dois fatores, as retas do gráfico seriam paralelas, ou seja, o efeito do nitrogênio seria o mesmo nos dois níveis de esterco.

Exercício (11.3)

Tabela de contingência para os dados do exercício 11.3

Raça	Incidência de parasitas		Total
	Sim	Não	
Pura	105	700	805
Não-pura	50	500	550
Total	155	1200	1355

Esse é um teste de independência entre as variáveis "incidência de parasitose" e "raça".

Hipóteses:

H_0: A incidência de parasitose independe da raça do animal
(ou seja, as duas variáveis não estão associadas)

H_a: A incidência de parasitose depende da raça do animal
(ou seja, as duas variáveis estão associadas)

Tabela de valores esperados sob a hipótese de independência (H_0)

Raça	Incidência de parasitas		Total
	Sim	Não	
Pura	805x155/1355=92.08	805x1200/1355=712.92	805
Não-pura	550x155/1355=62.92	550x1200/1355=487.08	550
Total	155	1200	1355

[2,]

Estatística de Teste:

$$X^2 = \frac{(105-92.08)^2}{92.08} + \frac{(700-712.92)^2}{712.92} + \frac{(50-62.92)^2}{62.92} + \frac{(500-487.08)^2}{487.08}$$

$$X^2 = 1.81 + 0.23 + 2.65 + 0.34 = 5.03$$

Região de Rejeição: Rejeita-se H_0 se $X^2_{obs} > \chi^2_{1;0,01} = 6.64$

Verificação: Como 5.03 não está na região de rejeição (5.03 > 6.64), não rejeitamos a hipótese nula ao nível de significância de 1%.

Cálculo do Valor P: Valor P = $P[X^2_1 > 5.03]$ = ?

Na tabela Qui-Quadrado, na linha 1, não existe o valor 5.03.
Mas 5.03 está entre os valores 5.024 e 6.635, correspondentes às colunas de 2.5% e 1% de probabilidade, respectivamente. Assim, $P[X^2_1 > 5.03]$ está entre 0.01 e 0.025.
Desse modo, 0.01 < Valor P < 0.025.
Como Valor P > 0,01 (α), não rejeitamos a hipótese nula ao nível de significância de 1%, confirmando (como esperado) a conclusão anterior.

Conclusão em termos do problema: Ao nível de significância de 1%, não rejeitamos a hipótese de independência entre a incidência de parasitose e a raça do animal (0.01 < valor P < 0.025).

Medindo a associação entre variáveis: a Razão das Chances

Através de teste Qui-quadrado verificou-se a existência de associação entre a "incidência de parasitose" e "raça". Como quantificar esta associação? Quando as variáveis são quantitativas, pode-se medir a associação linear entre duas variáveis através do coeficiente de correlação. Quando duas variáveis são qualitativas, essa medida seria o Risco Relativo. Porém, o Risco Relativo só pode ser calculado em estudos prospectivos, que não é o caso do estudo desse problema, pois os animais foram classificados simultaneamente quanto à raça e à incidência de parasitose (os grupos de raça não-pura e pura não foram fixados previamente e depois contada a incidência de parasitose). A medida alternativa ao Risco Relativo nesses casos é a Razão das Chances (RC). Chance é definida como razão de probabilidades. A chance de ter parasitose no grupo de raça pura, por exemplo, é definido como a razão entre a probabilidade de ter parasitose e a probabilidade de não ter parasitose, ou seja, (105/805) / (700/805) = 105/700 = 0.15. No grupo de raça não-pura, a chance de ter parasitose é 50/500 = 0.10. Comparando as chances nos dois grupos, temos que 0.15/0.10 = 1.50. Essa a estimativa da Razão das Chances de parasitose entre os grupos de raça pura e não-pura. Interpretando esse número, vemos que a chance de parasitose no grupo de raça pura é 50% maior do que no grupo de raça não-pura. De maneira geral, no caso de tabelas 2x2 como a que apresentamos a seguir

	Resposta de Interesse		
Grupo	Sucesso	Fracasso	Total
A	a	b	a+b
B	c	d	c+d
Total	a+c	b+d	a+b+c+d

a razão das chances de sucesso entre o grupo A e B é estimada como

$$RC = \frac{\text{chance de sucesso no grupo A}}{\text{chance de sucesso no grupo B}} = \frac{\left.\frac{a}{a+b}\middle/\frac{b}{a+b}\right.}{\left.\frac{c}{c+d}\middle/\frac{d}{c+d}\right.} = \frac{a/b}{c/d} = \frac{a\cdot d}{b\cdot c}\ .$$

No caso do problema,

$$RC = \frac{\text{chance de parasitose no grupo raca pura}}{\text{chance de parasitose no grupo raca nao-pura}} = \frac{105/700}{50/500} = \frac{105\cdot 500}{50\cdot 700} = 1.50$$

O termo Razão de Chances vem do inglês *Odds Ratio.* Alguns autores preferem não traduzir "odds" por "chance" e acabam por trabalhar com o termo em inglês, ou com sua abreviatura, OR, ou ainda com uma mistura de termos (*Razão de Odds*).

Exercício (11.4)

Esse é um teste de independência entre as variáveis "ocorrência de doença coronariana" e "grupo sanguíneo".

Hipóteses:

H_0: A ocorrência de doença coronariana independe do grupo sanguíneo (ou seja, as duas variáveis não estão associadas)

H_a: A ocorrência de doença coronariana depende do grupo sanguíneo (ou seja, as duas variáveis estão associadas)

Tabela de valores esperados sob a hipótese de independência (H_0)

Grupo Sanguíneo	Ocorrência de doença coronariana		Total
	Sim	Não	
A	450 x 42/1800 = 10,50	450 x 1758/1800 = 439,50	450
B	450 x 42/1800 = 10,50	450 x 1758/1800 = 439,50	450
AB	300 x 42/1800 = 7,00	300 x 1758/1800 = 293,00	300
O	600 x 42/1800 = 14,00	600 x 1758/1800 = 586,00	600
Total	42	1758	1800

Estatística de Teste:

$$X^2 = \frac{(9-10,50)^2}{10,50} + \frac{(141-439,50)^2}{439,50} + \frac{(12-10,50)^2}{10,50} + \frac{(146-439,50)^2}{439,50} + $$
$$+ \frac{(9-7,00)^2}{7,00} + \frac{(291-293,00)^2}{293,00} + \frac{(12-14,00)^2}{14,00} + \frac{(588-586,00)^2}{586,00}$$

$X^2 =$ 1,316

Região de Rejeição: Rejeita-se H_0 se $X^2_{obs} > \chi^2_{3;0,01} = 11,345$

Verificação: Como 1,316 está fora da região de rejeição (1,316 < 11,345), não rejeitamos a hipótese nula ao nível de significância de 1%.

Cálculo do Valor P: Valor P = $P[X^2_3 > 1,316]$ = ?

Na tabela Qui-Quadrado, na linha 3, não existe o valor 1,316.
Ele está entre os valores 1,213 e 2,366, correspondentes às coluna do 75% e 50%, respectivamente. Assim, $P[X^2_3 > 1,316]$ está entre 50% e 75%. Desse modo, 0,50 < Valor P < 0,75.

Conclusão em termos do problema: Ao nível de significância de 1%, não rejeitamos a hipótese de independência a ocorrência de doença coronariana e o grupo sanguíneo (0,50 < Valor P < 0,75).

Mesmo não tendo encontrado evidências de associação entre o grupo sanguíneo e a ocorrência de doença coronariana, vamos calcular as medidas de associação a título de exercício.

Medindo a associação entre variáveis: o Risco Relativo

A estimativa de risco pode ser feita com os dados deste estudo, pois ele é do tipo prospectivo, tendo partido da exposição (grupo sanguíneo) para a observação do evento (doença coronariana).

Na tabela a seguir, são estimados os riscos de doença coronariana em cada grupo sanguíneo e também o risco relativo, tendo o grupo sanguíneo O como referência.

Grupo	Cálculo do Risco	Risco Relativo
A	9/450 = 0,020	0,020/0,020= 1,00
B	12/450 = 0,027	0,027/0,020= 1,35
AB	9/300 = 0,030	0,030/0,020= 1,50
O	12/600 = 0,020	1

Interpretando a tabela, temos que o risco de doença coronária entre os voluntários do grupo sanguíneo A é igual ao risco de doença coronária dos voluntários de sangue tipo O. Os voluntários do grupo B tem um risco de doença coronária que é 1,35 vezes esse risco no grupo O (35% maior). Já os voluntários do grupo AB têm um risco de doença coronária que é 1,50 vezes o risco dos voluntários do grupo O (50% maior).

Se quiséssemos o grupo AB com referência, a tabela ficaria

Grupo	Risco	Risco Relativo
A	0,020	0,020/0,030= 1,50
B	0,027	0,027/0,030= 0,90
AB	0,030	1
O	0,020	0,020/0,030= 1,50

Interpretando a tabela, temos que o risco de doença coronária entre os voluntários tanto do grupo sanguíneo A quanto do grupo O é 1,50 vezes o risco de doença coronária dos voluntários de sangue tipo AB. Os voluntários do grupo B têm um risco de doença coronária que é 0,90 vezes esse risco no grupo AB (10% menor).

Medindo a associação entre variáveis: a Razão das Chances

Embora o Risco Relativo possa ser estimado com os dados desse estudo, a Razão de Chances também pode ser calculada.

Se escolhermos o grupo O como referência, teremos a seguinte tabela:

Grupo	Cálculo	Razão de Chances
A	(9 x 588)/ (12 x 441)	1,00
B	(12 x 588) / (12 x 438)	1,34
AB	(9 x 588) / (12 x 291)	1,52
O	----------------------	1

Interpretando a tabela, vimos que a chance de doença coronária entre os voluntários do grupo sanguíneo A é igual à chance de doença coronária entre os voluntários de sangue tipo O. Os voluntários do grupo B têm uma chance de doença coronária que é 34% maior do que essa chance no grupo O. Já os voluntários do grupo AB têm uma chance de doença coronária que é 1,52 vezes a chance dos voluntários do grupo O (52% maior).

Nota: Observe que os valores do Risco Relativo e os da Razão de Chances possuem valores muito similares nesse problema. Isto acontece porque o risco do evento (doença coronariana) em todos os grupos é pequeno (entre 2% e 3%), o que faz as duas medidas de associação terem valores parecidos.

Referências Bibliográficas

Nogueira, M. L. G. et alli (1997), ***Introdução à Bioestatística***, apostila do Instituto de Ciências Exatas da UFMG.

Montgomery, D. C. (2001), ***Design and Analysis of Experiments***, 5ª Edição, John Wiley and Sons, New York.

Soares, J. F. et alli ( ), ***Introdução à Estatística***, Editora Guanabara Dois

Triola, M. F. (1998), ***Introdução à Estatística***, 7ª Edição, LTC.