UNIVERSIDADE
FEDERAL DE JUIZ DE FORA

Programa de Pós-Graduação em
Modelagem Computacional

Métodos Estatísticos

Trabalho 2
Análise de Dois Bancos de Dados Reais

Aluno: Iago Gomes de Lima Rosa

Professor: Henrique Steinherz Hippert

Juiz de Fora

Setembro de 2019

1 INTRODUÇÃO

Este trabalho apresenta o estudo de duas bases de dados reais: *cancer.txt* e *lowbw.txt*. Espera-se aplicar diferentes técnicas e conceitos demonstrados ao longo da disciplina.

A primeira base de dados é apresentada e disponibilizada pela principal referência da disciplina: (MAGALHÃES; DE LIMA, 2002) e referem-se a dados de uma pesquisa sobre incidência de câncer.

A segunda base de dados referem-se aos dados de uma pesquisa que procura identificar os fatores de risco associados ao nascimento de um bebê com baixo peso (peso menor que 2500 g).

Os códigos foram desenvolvidos na linguagem de programação *python3* utilizando principlamente as bibliotecas: *scipy*, *statsmodels*, *pandas*, *matplotlib* e *seaborn*. Estes códigos estão disponíveis no *github* através do link: `https://github.com/iagorosa/metodos-estatisticos/tree/master/Trabalho2`. O trabalho foi resolvido través das principais referencias da disciplina (MAGALHÃES; DE LIMA, 2002; HIPPERT, 2019).

2 ARQUIVO *cancer.txt*

A primeira base de dados considerada foi sobre câncer. Os itens a seguir referem-se aos dados contidos no arquivo de nome *cancer.txt*. Esse arquivo é apresentado em 9 colunas representando as seguintes variáveis de interesse:

coluna 1: identificação do paciente.
coluna 2: diagnóstico:
1 = falso-negativo: diagnosticados como não tendo a doença quando na verdade a tinham.
2 = negativo: diagnosticados como não tendo a doença quando de fato não a tinham.
3 = positivo: diagnosticados corretamente como tendo a doença.
4 = falso-positivo: diagnosticados como tendo a doença quando na verdade não a tinham.
coluna 3: idade.
coluna 4: espectro qumico da análise do sangue-alkaliine phosphatose (AKP).
coluna 5: concentração de fosfato no sangue (P).
coluna 6: enzima, lactate dehydrogenase (LDH).
coluna 7: albumina (ALB).
coluna 8: nitrogênio na uréia (N).
coluna 9: glicose (GL).

Exercício 24, Capítulo 1 - Página 45

letra b)

Médicos afirmam que o grupo dos falso-positivos é mais jovem do que o dos falso-negativos. Para verificar essa informação, são examinados os histogramas de cada um dos grupos, representado pela Figura 2.1 e pela comparação desses grupos através do *boxplot* da Figura 2.2.

Pela Figura **??**, pode-se perceber uma concentração de frequências mais altas entre 40 a 70 anos para a idade dos falso-positivos. Além disso, percebe-se que a idade máxima para pacientes do grupo falso-positivos é 90 anos. Por outro lado, para o grupo de falso-negativos as idades entre 60 e 70 anos tem frequência alta e também observa-se alguns valores maiores que 90 anos.

Os *boxplots* da Figura 2.2 mostram que a mediana do grupo dos falso-negativos é maior do que a do grupo de falso-positivos. Também é possível perceber pelo *boxplot* que o limite superior do grupo dos falso-negativos é maior do que o outro grupo. Através destas imagens, há evidências para concordar com a afirmação dos médicos. Utilizando o recurso da tabela de frequências, encontra-se ainda mais evidências para concordar com a afirmação.

Através da Tab. 2.1 e Tab. 2.2 pode-se confirmar as observações descritas acima. Os intervalos de [50, 60) e [60, 70) apresentam frequências elevadas em ambos grupos da Fig. **??**. Sendo assim, pode-se considerar esses intervalos como críticos para avaliar se o grupo de falso-positivos é mais jovem do que o grupo dos falso-negativos. Avaliando a coluna de *Frequência Relativa Acumulada* das tabelas mencionadas, percebe-se que a porcentagem de idade de até 60 anos é de 67,692% para os falso-positivos contra 57,148% para os falso-negativos. Sendo assim, pode-se dizer que o grupo de falso-positivos é mais jovem que o grupo de falso-negativos, já que existem mais pessoas com até 60 anos.

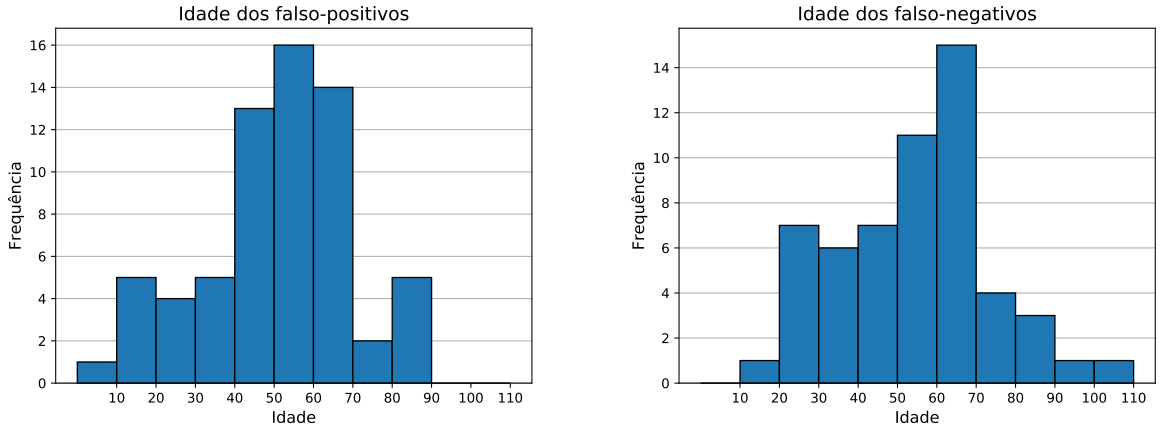

Figura 2.1: Comparação entre os histogramas da idade dos pacientes de cada grupo.

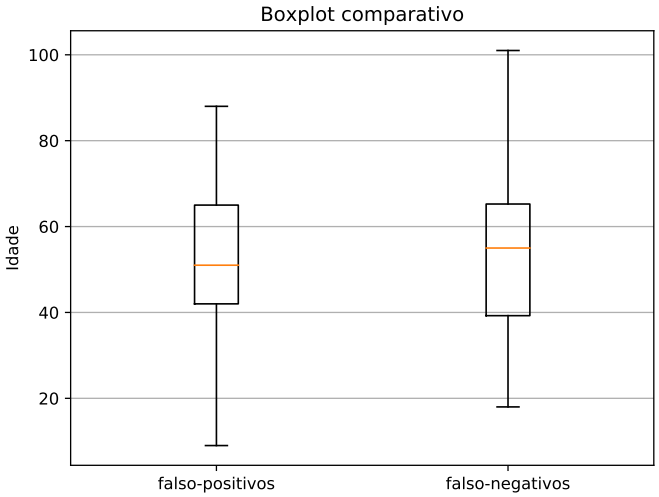

Figura 2.2: Comparação através do *boxplot* da idade dos pacientes de cada grupo.

Tabela 2.1: Tabela de Frequência de idade do grupo falso-positivos

	Frequência	Freq. Acumulada	Freq. Rel. Acum. (%)
[0, 10)	1	1	1.538
[10, 20)	5	6	9.231
[20, 30)	4	10	15.385
[30, 40)	5	15	23.077
[40, 50)	13	28	43.077
[50, 60)	16	44	67.692
[60, 70)	14	58	89.231
[70, 80)	2	60	92.308
[80, 90)	5	65	100.000

Tabela 2.2: Tabela de Frequência de idade do grupo falso-negativos

	Frequência	Freq. Acumulada	Freq. Rel. Acum. (%)
[0, 10)	0	0	0.000
[10, 20)	1	1	1.785
[20, 30)	7	8	14.285
[30, 40)	6	14	25.000
[40, 50)	7	21	37.500
[50, 60)	11	32	57.148
[60, 70)	15	47	83.929
[70, 80)	4	51	91.071
[80, 90)	3	54	96.429
[90, 100)	1	55	98.214
[100, 110)	1	56	100.000

Exercício 36, Capítulo 6 - Página 220

Neste momento, é considerado a variável *enzima, lactate dehydrogenase* (LDH) para as pessoas com idade menor que 40.

letra a)

A Tab. 2.3 demonstra algumas medidas descritivas referentes a variável em questão. Percebe-se uma amplitude dos dados de aproximadamente 100 unidades de LDH com média 18.26 e desvio padrão 11.76. A Fig. 2.3 demonstra de maneira gráfica informações similares as dispostas na Tab. 2.3.

Tabela 2.3: Medidas descritivas para LDH

Medida	**Valores LDH**
Quantidade	362
Média	18.260
Dev. Padrão	11.760
Mínimo	0.000
25%	13.225
50%	15.150
75%	18.400
Máximo	99.900

letra b)

Seja a variável aleatória *X:* níveis de enzima, lactate dehydrogenase (LDH) nos pacientes com menos de 40 anos. É verificado se X pode ser modelado por uma distribuição normal. As hipóteses para esse teste, com nível de significância de $\alpha = 5\%$, são:

$$H_0 : \text{A população pode ser modelada por uma distribuição normal}$$
$$H_1 : \text{A população pode ser modelada por uma não distribuição normal}$$

Para isso, são utilizados os testes de Shapiro-Wilk e Lilliefors considerando um nível de significância de $\alpha = 0.05$. Os seguintes valores foram encontrados:

Tabela 2.4: P-valores dos testes de normalidade de X

Teste	**p-valor**
Shapiro	0.0
Lillefors	0.0

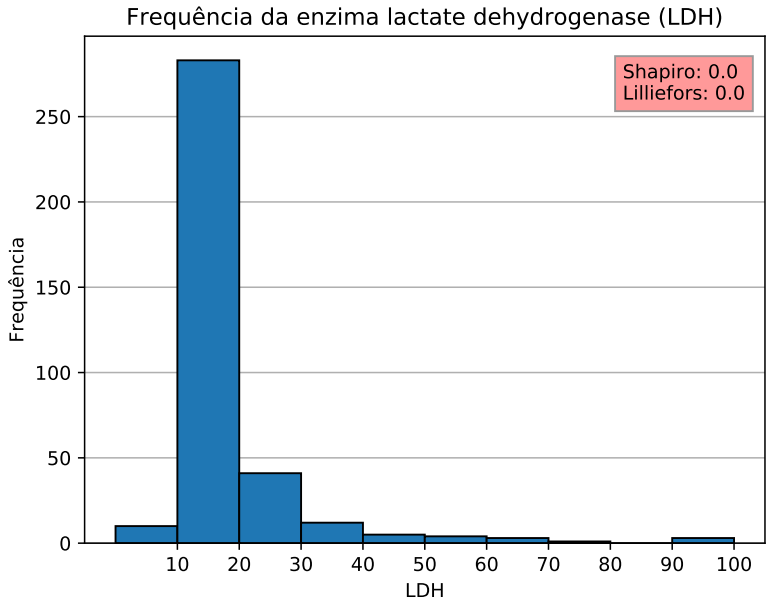

Figura 2.3: Histograma da variável LDH.

Ambos resultados de p-valor encontrados com os testes de Shapiro-Wilk e Lilliefors foram menores do que o valor de α. Portanto, rejeita-se a hipótese nula, ou seja, há evidências nos dados que leva a inferir que X não é ajustada por um modelo normal com coeficiente de confiança de 95%. Pode-se verificar essa mesma informação através da Fig. 2.4, onde há um histograma e uma curva normal, em vermelho, utilizando medidas de média e desvio padrão da variável LDH. Pode-se notar que este modelo de distribuição normal não se ajusta bem aos dados.

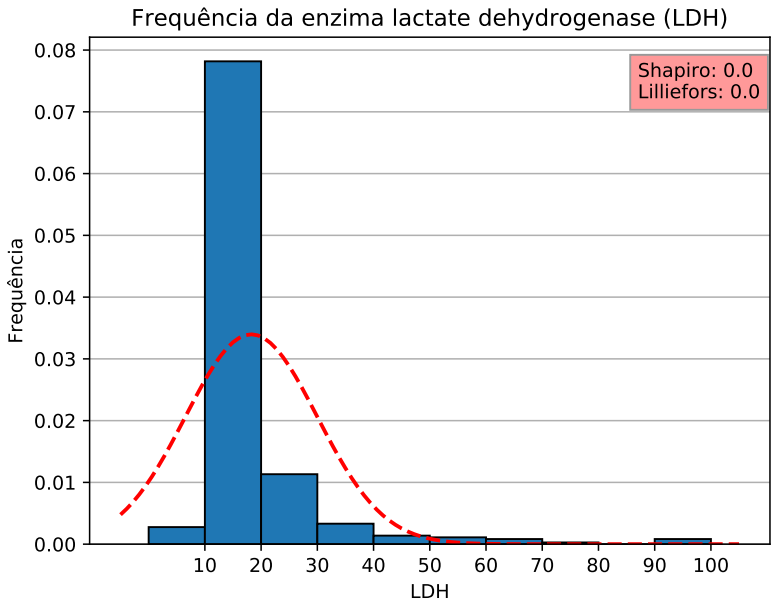

Figura 2.4: Modelo de distribuição normal baseado nas medidas dos dados LDH.

Exercício 41, Capítulo 8 - Página 307

Foram definidos dois grupos para os próximos resultados: o primeiro é de pacientes jovens, com idades inferiores ou iguais a 54 anos, e o segundo grupo de pacientes idosos com idades superiores a 54 anos. Para esses grupos, realiza-se análises sobre a variável *nitrogênio na ureia* (N).

letra a)

Inicialmente, são extraídas algumas medidas descritivas da variável em questão. Os resultados são observados pela Tab. 2.5.

Tabela 2.5: Medidas descritivas para N

	Jovens	**Idosos**
Quantidade	191	171
Média	12.822	17.228
Dev. Padrão	3.900	7.465
Minímo	3.000	4.000
25%	10.000	13.000
50%	13.000	16.000
75%	15.000	19.000
Máximo	33.000	58.000

A Fig. 2.5 apresenta o *boxplot* comparativo entre os grupos de jovens e idosos para a variável de nitrogênio na ureia. Percebe-se que a medida de nitrogênio na ureia dos idosos é mais elevada do que em jovens. Além disso, observa-se várias instâncias com o valor de N superior a 25 para o grupo de idosos, o que não ocorre para o grupo de jovens (exceto por um *outlier* observado). Observa-se pela Tab. 2.5 uma média com desvio padrão de N para os idosos bem maior em relação aos jovens.

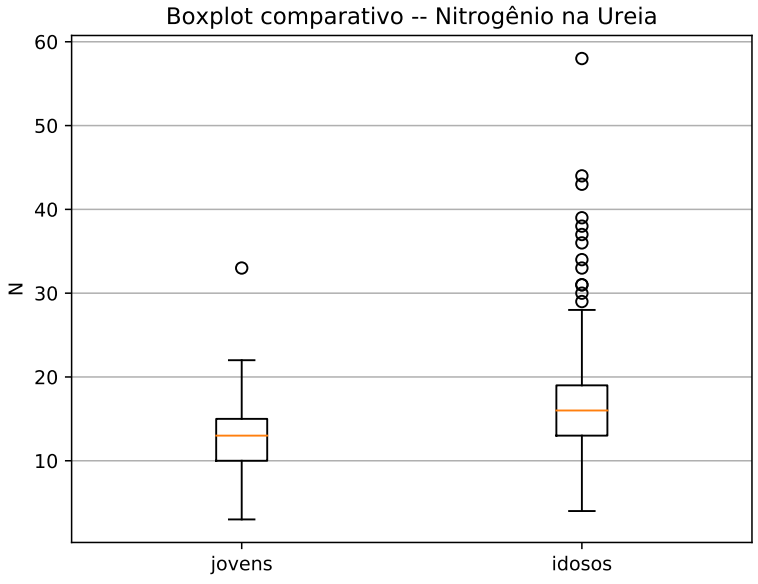

Figura 2.5: Comparação através do *boxplot* da variável nitrogênio na ureia para os grupos de jovens e idosos.

letra b)

Seja a variável aleatória *X:* concentração de nitrogênio na ureia de pacientes idosos. Deseja-se verificar se a média populacional de N da população de pacientes idosos é superior a 15, supondo que a população possa ser

modelada adequadamente por uma distribuição Normal com desvio padrão $\sigma = 7$ com nível de significância $\alpha = 0.001$. Para isso, o seguinte teste é realizado:

$$H_0 : \ \mu_X \leq 15$$
$$H_1 : \ \mu > 15$$

Os resultados obtidos podem ser verificado pela Tab. 2.6. Através desta, pode-se perceber que o valor de Z observado é maior do que o Z crítico. Sendo assim, o valor de Z observado entra na área de rejeição do teste. O mesmo pode ser observado com o p-valor encontrado, cujo valor é menor do que o nível de significância $\alpha = 0.001$. Portanto, rejeita-se a hipótese nula e há evidências pela amostra para concluir que a média concentração na ureia é maior do que 15 para a população de idosos com coeficiente de confiança de 99.9%. A Fig. 2.6 apresenta um modelo normal com desvio padrão $\sigma = 7$, em vermelho.

Tabela 2.6: Resultados obtidos para o teste realizado com idosos

	Valor
Z crítico	3.090
Z observado	4.162
P-valor	0.000

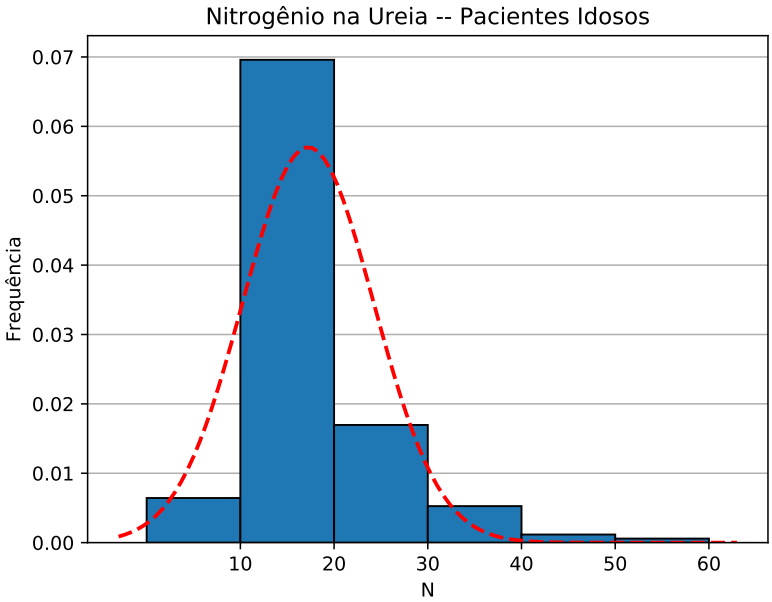

Figura 2.6: Modelo de distribuição normal baseado na média dos dados de N dos idosos e desvio padrão $\sigma = 7$.

letra c)

Seja a variável aleatória *X:* concentração de nitrogênio na ureia de pacientes jovens. Deseja-se verificar se a média populacional de N para o grupo de pacientes jovens é inferior a 15, supondo que a população possa ser modelada adequadamente por uma distribuição Normal com desvio padrão $\sigma = 5$ ao nível de significância $\alpha = 0.001$. Para isso, o seguinte teste é realizado:

$$H_0 : \ \mu_X \geq 15$$
$$H_1 : \ \mu < 15$$

Os resultados obtidos podem ser verificado pela Tab. 2.7. Através desta, pode-se perceber que o valor de Z observado é menor do que o Z crítico. Sendo assim, o valor de Z observado entra na área de rejeição

do teste. O mesmo pode ser observado com o p-valor encontrado, cujo valor é menor do que o nível de significância $\alpha = 0.001$. Portanto, rejeita-se a hipótese nula e há evidências pela amostra para concluir que a média concentração na ureia é menor do que 15 para a população de jovens com coeficiente de confiança de 99.9%. A Fig. 2.7 apresenta um modelo normal com desvio padrão $\sigma = 5$, em vermelho.

Tabela 2.7: Resultados obtidos para o teste realizado com jovens

	Valor
Z crítico	-3.090
Z observado	-6.020
P-valor	0.000

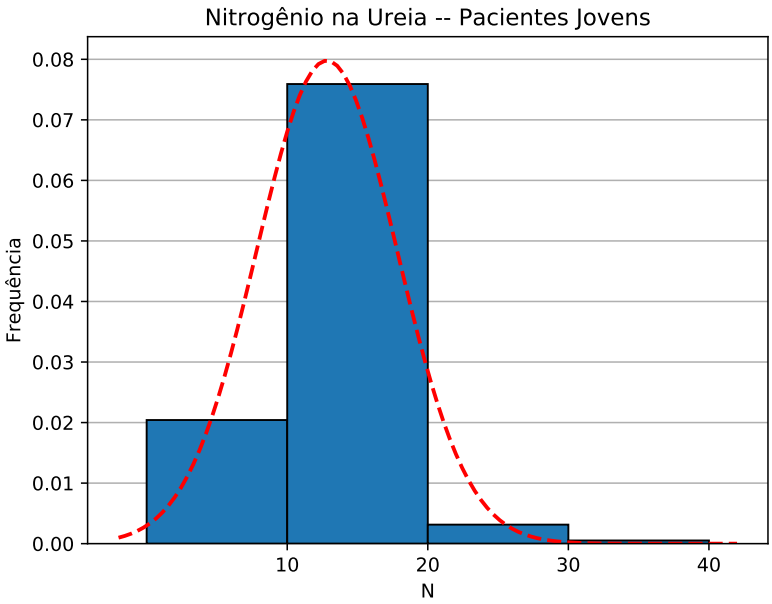

Figura 2.7: Modelo de distribuição normal baseado na média dos dados de N dos jovens e desvio padrão $\sigma = 5$.

letra d)

Pelos resultados obtidos em *(b)* e *(c)*, percebe-se que a população de idosos tem média de concentração de nitrogênio na ureia maior do que 15 e os jovens tem a concentração menor do que 15. Isso foi possível inferir pois os testes rejeitaram a hipótese nula. Porém, observando a tabela 2.5, nota-se que a média do grupo dos jovens é de 12.822, que é menor do que 15, e a média do grupo de idosos é de 17.228, que é maior do que 15. Sendo assim, já era de se esperar a rejeição das hipóteses nulas.

Exercício 30, Capítulo 9 - Página 368

Neste momento, deseja-se realizar testes a fim de verificar se conforme aumenta a idade, muda-se a concentração de nitrogênio na ureia. Aqui, considerou-se um grupo de doentes formado por pacientes com diagnóstico de falso-negativo ou positivo. Seguindo a mesma lógica, considerou-se um grupo de não doentes formados por pacientes com diagnóstico negativo ou falso-positivo.

letra a)

A Fig. 2.8 apresenta o gráfico de dispersão para os pacientes doentes. Conforme a idade avança, não percebe-se um incremento significativo da concentração de nitrogênio, ou seja, mantém-se constante entre 5

e 25 uniades de N. Também é possível perceber essa informação através do valor de correlação de 0.263 para esse caso. Sendo assim, fica explícito que há uma correlação baixa entre a variável idade e concentração de nitrogênio. Há uma observação interessante sobre a concentração de nitrogênio na ureia para pacientes entre 50 a 70 anos. Nessa faixa de valor ocorrem alguns *outliers* com valores superiores a 30 unidades de N. Nesta mesma faixa pode-se perceber uma quantidade maior de observações.

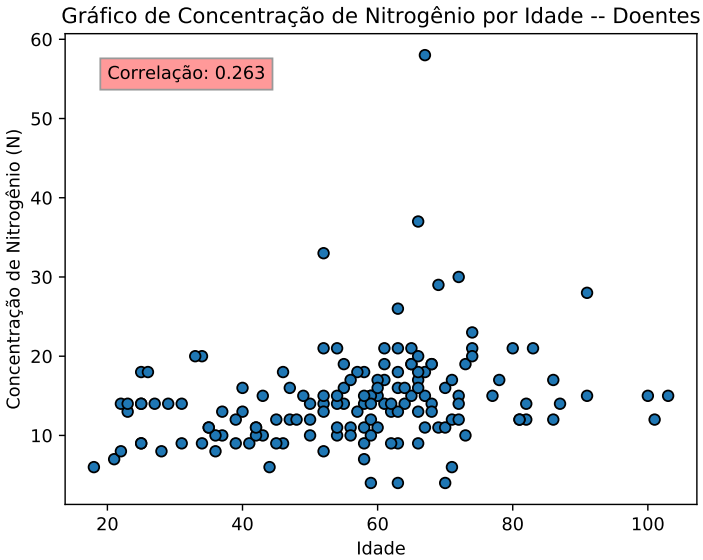

Figura 2.8: Gráfico de dispersão da idade por N para os pacientes doentes.

letra b)

A Fig. 2.9 demonstra a regressão linear realizada para a dispersão dos dados. Nota-se que o coeficiente angular da reta ajustada apresenta um valor baixo, 0.094. Esse valor corrobora com a dedução de constância dos dados discutido acima. Esse coeficiente diz que ao avançar na idade, a concentração de nitrogênio aumenta a uma taxa de 0.094.

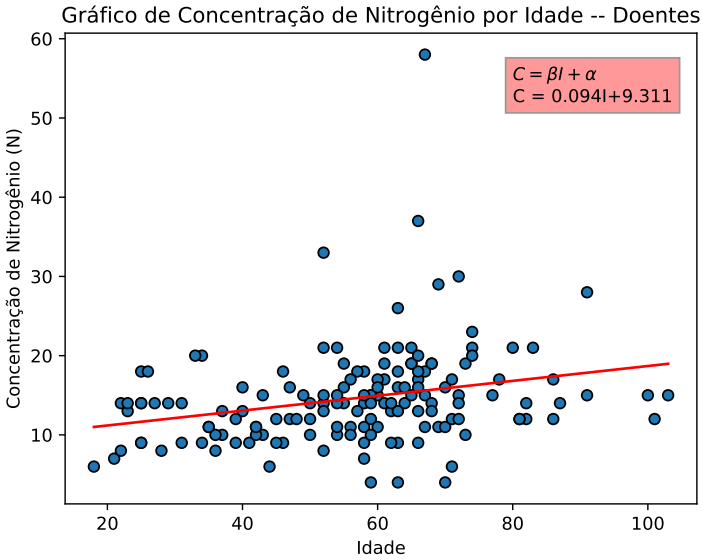

Figura 2.9: Regressão linear da idade por N para os pacientes doentes.

letra c)

A tabela ANOVA é utilizada aqui para verificar o resultado da regressão realizada. É feito um teste de hipótese que consiste em verificar a influência da idade na concentração de nitrogênio na ureia (N) para pacientes doentes. Assume-se que um modelo normal pode ser ajustado aos resíduos da regressão linear. O teste é realizado da seguinte maneira, com nível de significância $\alpha = 5\%$:

$$H_0 : \text{A idade não influencia diretamente na variação de } N$$
$$H_1 : \text{A idade influencia diretamente na variação de } N$$

Os resultados obtidos encontram-se na Tab. 2.8. O p-valor do teste (coluna $PR(>F)$) foi menor do que o nível de significância α, isto é, p-valor $= 0.001 < \alpha = 0.05$. Portanto, rejeita-se a hipótese nula. Sendo assim, existem evidências para afirmar que a idade influencia diretamente na variação de N para pacientes doentes com coeficiente de confiança de 95%. Porém, os resultados obtidos nas letras *(a)* e *(b)* dessa questão demonstram o contrário.

Tabela 2.8: Tabela ANOVA para os pacientes doentes.

Tabela ANOVA				
	SS	**df**	**F**	**PR(>F)**
N	3,254.173	1.000	11.103	0.001
Resíduo	43,668.569	149.000	-	-

letra d)

A Fig. 2.10 apresenta o gráfico de dispersão para os pacientes não doentes. Conforme a idade avança, não percebe-se um incremento significativo da concentração de nitrogênio na ureia, ou seja, mantém-se constante entre 5 e 25 unidades de N. Também é possível perceber essa informação através do valor de correlação de 0.438 para esse caso. Apesar de ser um valor um pouco mais alto do que o encontrado para os pacientes doentes, ainda apresenta uma correlação baixa entre a variável idade e concentração de nitrogênio. Há uma observação interessante sobre a concentração de nitrogênio na ureia para pacientes entre 55 a 90 anos. Nessa faixa de valor ocorrem alguns *outliers* com valores superiores a 30 unidades de N.

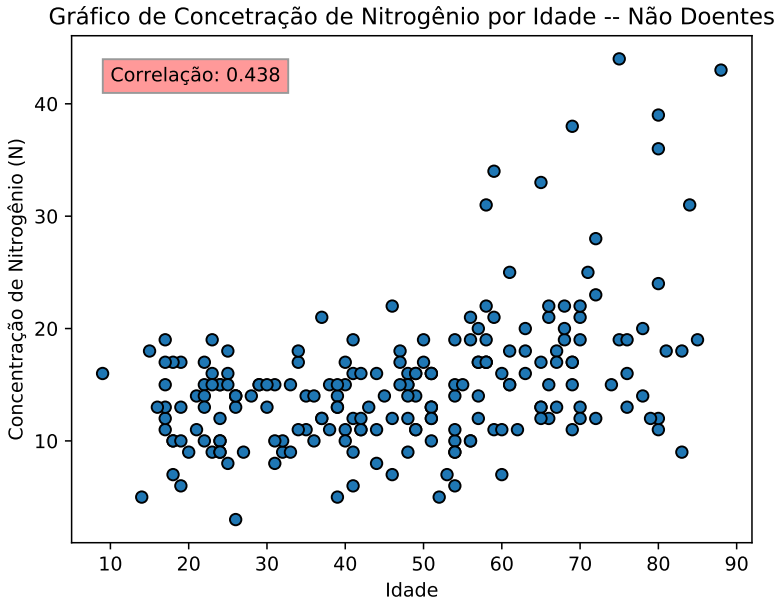

Figura 2.10: Gráfico de dispersão da idade por N para os pacientes não doentes.

letra e)

A Fig. 2.11 demonstra a regressão linear realizada para a dispersão dos dados. Nota-se que o coeficiente angular da reta ajustada apresenta um valor baixo, 0.142. Esse valor corrobora com a dedução de constância dos dados discutido acima. Esse coeficiente diz que ao avançar na idade, a concentração de nitrogênio na ureia aumenta a uma taxa de 0.142.

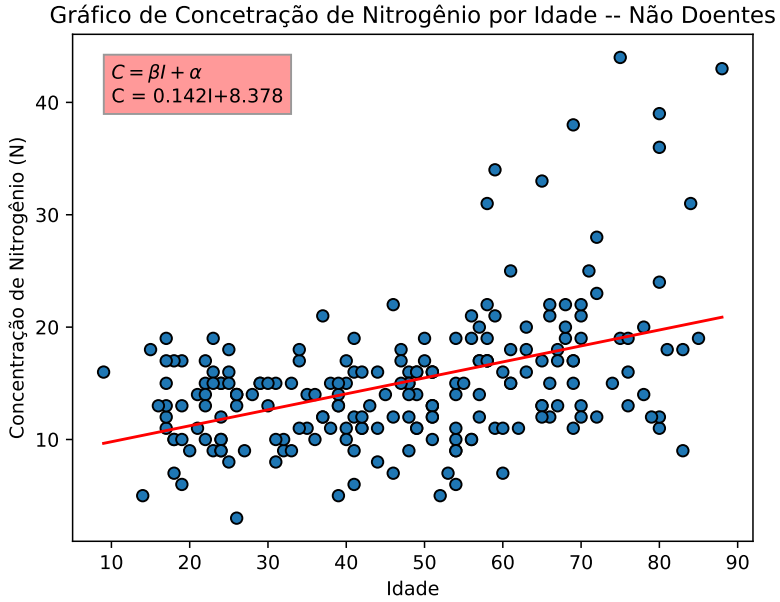

Figura 2.11: Regressão linear da idade por N para os pacientes não doentes.

letra f)

Supondo que a variável dependente é a concentração de nitrogênio e que a covariável é a idade do paciente, serão realizados testes sobre a evidência estatística de que a idade influência na concentração de nitrogênio. Considera-se ainda que um modelo normal pode ser ajustado aos resíduos da regressão linear.

A tabela ANOVA é utilizada aqui para verificar o resultado da regressão realizada. É feito um teste de hipótese que consiste em verificar a influência da idade na concentração de nitrogênio na ureia (N) para pacientes não doentes. O teste é realizado da seguinte maneira, com nível de significância $\alpha = 5\%$:

$$H_0 : \text{A idade não influencia diretamente na variação de } N$$
$$H_1 : \text{A idade influencia diretamente na variação de } N$$

Os resultados obtidos encontram-se na Tab. 2.9. O p-valor do teste (coluna $PR(>F)$) foi menor do que o nível de significância α, isto é, p-valor $= 0.000 < \alpha = 0.05$. Portanto, rejeita-se a hipótese nula. Sendo assim, existem evidências para afirmar que a idade influencia diretamente na variação de N para pacientes não doentes com coeficiente de confiança de 95%. Porém, o resultado obtido nas letras *(d)* e *(e)* dessa questão demonstra o contrário.

Tabela 2.9: Tabela ANOVA para os pacientes não doentes.

Tabela ANOVA				
	SS	**df**	**F**	**PR(>F)**
N	14,792.296	1	49.521	0.000
Residual	62,429.894	209	-	-

letra g)

Comparando as retas das regressões lineares para os pacientes doentes e não doentes pela Fig. 2.12, percebe-se que o efeito da idade, na concentração de nitrogênio na ureia, apresenta uma pequena diferença entre pacientes doentes e não doentes. Nota-se que o efeito é um pouco maior para pacientes não doentes em relação a pacientes doentes.

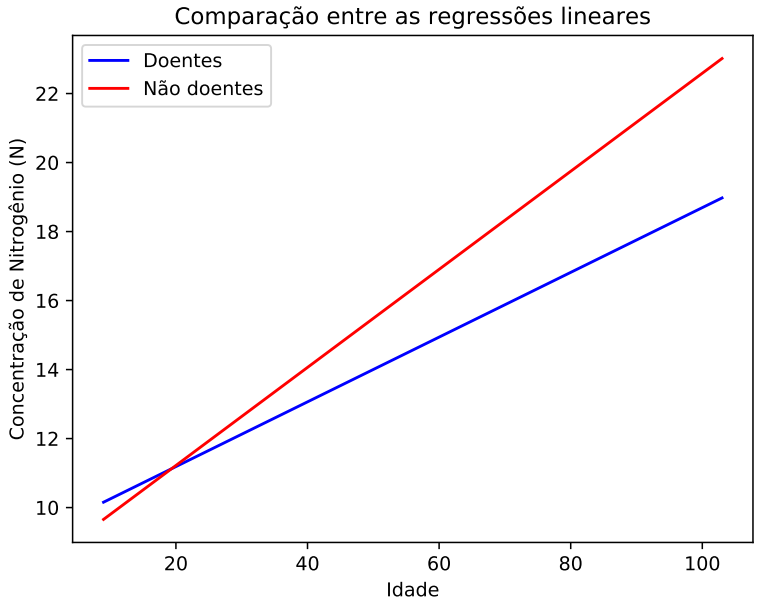

Figura 2.12: Comparação entre as regressões lineares de pacientes doentes e não doentes.

3 ARQUIVO *lowbw.txt* (low birth weight)

O baixo peso em recém-nascidos tem sido uma preocupação para os médicos há muitos anos, uma vez que as taxas de mortalidade infantil e de ocorrência de problemas de má formação são muito elevadas em bebês que nascem com baixo peso. De acordo com a literatura da área médica, o comportamento das mães durante a gravidez (incluindo a dieta, o hábito do fumo, e o atendimento pré-natal) pode alterar as chances de o bebê nascer no prazo correto e com o peso normal. Os dados neste trabalho provém de um estudo que procura identificar os fatores de risco associados ao nascimento de um bebê com baixo peso (peso menor que 2500 g). Estudos realizados anteriormente haviam indicado que os fatores listados no quadro abaixo poderiam ser importante

variáveis	fator	codificação
ID	Identificação da mãe	-
low	Peso baixo ao nascer	0 se peso $\geq$ 2500g; 1 se peso < 2500g
age	Idade da mãe (em anos)	-
lwt	Peso da mãe na época da última menstruação (libras)	-
race	Raça da mãe	1 = branca; 2 = outra; 3 = negra
smoke	hábito do fumo durante a gravidez	1 = sim; 0 = não
ptl	histórico de partos prematuros	0 = nenhum; 1 = um; 2 = dois; etc.
ht	histórico de hipertensão	1 = sim; 0 = não
ui	presença de irritabilidade uterina	1 = sim; 0 = não
ftv	número de visitas ao médico durante o primeiro trimestre da gravidez	0 = nenhuma; 1 = uma; 2 = duas; etc.
bwt	Peso do bebê ao nascer (em gramas)	-

Estas variáveis foram originalmente observadas para um grupo de 189 mulheres (HOSMER; LEMESHOW; STURDIVANT, 1989).

Descrição e Modelagem dos Dados

Questão 1)

São apresentados aqui as distribuições de frequência e gráficos de barras para as variáveis: **(a)** *ptl* (histórico de parto prematuro) e **(b)** *ftv* (número de visitas ao médico durante o primeiro trimestre de gravidez).

A Tab. 3.1 e a Tab. 3.2 demonstram as tabelas de frequências para as variáveis *ptl* e *ftv*. Já a Fig. 3.1 demonstra o gráfico de barras das duas variáveis em questão para mostrar a quantidade de casos que ocorrem em cada um dos diferenres valores presentes nos dados.

Tabela 3.1: Tabela de Frequência do histórico de parto prematuro (*ptl*).

	Frequência	Freq. Acum.	Freq. Rel. Acum. (%)
0	159	159	84.127
1	24	183	96.825
2	5	188	99.471
3	1	189	100.000

Pode-se perceber com os resultados acima que os dados apresentam uma forte assimetria positiva com valores mais frequentes (moda) no valor 0 em ambos casos. Percebe-se que a frequência vai diminuindo consideravelmente com o avanço do valor das variáveis. A queda ocorre de maneira mais brusca para os dados *ptl*. Um resultado que destacou-se foi que cerca de 97% dos dados tem histórico de parto prematuro igual a 1. Já para o número de visitas ao médico, aproximadamente a mesma porcentagem ocorre com 3 visitas.

Tabela 3.2: Tabela de Frequência do número de visitas ao médico durante o primeiro trimestre da gravidez (*ftv*).

	Frequência	**Freq. Acum.**	**Freq. Rel. Acum. (%)**
0	100	100	52.910
1	47	147	77.778
2	30	177	93.651
3	7	184	97.354
4	4	188	99.471
6	1	189	100.000

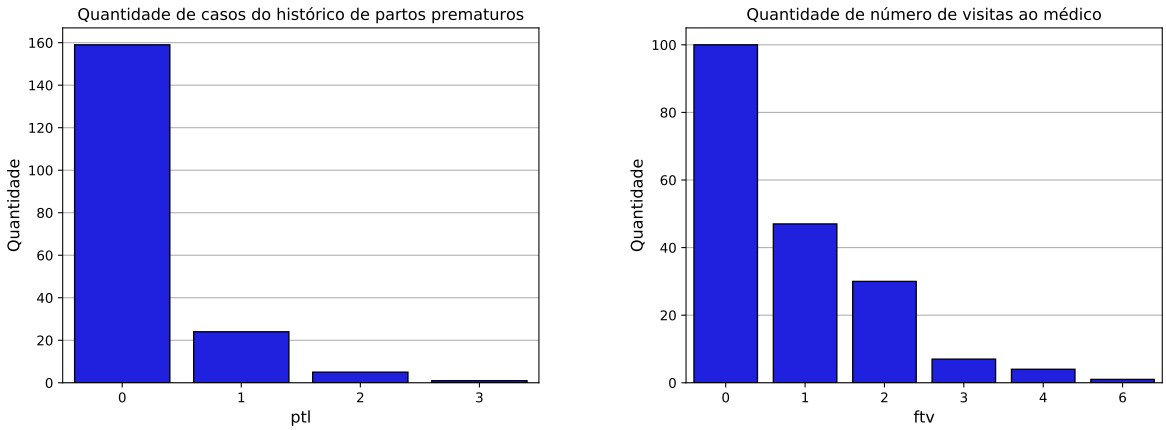

Figura 3.1: Gráfico de barras da quantidade de casos das variáveis *ptl* e *ftv*.

Questão 2)

Sejam as as variáveis aleatórias X: histórico de partos prematuros e Y: número de visitas ao médico durante o primeiro trimestre da gravidez. São avaliados se as variáveis podem ser modeladas por um modelo de Poisson. Os testes de hipóteses realizados, com nível de significância $\alpha = 5\%$, são:

$$H_0 : \text{a variável pode ser modelada por uma distribuição de Poisson}$$
$$H_1 : \text{a variável não pode ser modelada por uma distribuição de Poisson}$$

As classes foram agrupadas conforme necessário (quando a quantidade de um grupo é menor do que 5) reduzindo o grau de liberdade em 1. Outro grau de liberdade é reduzindo ao estimar o parâmetro λ através do valor da média dos dados. O valor qui-quadrado é calculado da seguinte forma:

$$Q^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Onde o representa os dados observados e e os valores esperado. Ambos vetores foram omitidos nesse relatório, o qual avalia apenas os resultados finais obtidos com o teste.

Tabela 3.3: Testes qui-quadrado para as variáveis *ptl* e *ftv*.

	X	**Y**
λ	0.196	0.794
df	1	3
Q^2	3.935	15.860
q crítico	3.841	7.815
p-valor	0.047	0.001

Pela Tab. 3.3 pode-se verificar que o valor observado do Q^2 é maior do que os valores de q crítico. Isso indica que ambos valores, tanto para X quanto para Y, estão na área de rejeição de H_0. O mesmo pode-se

observar com os p-valores, os quais são menores que $\alpha = 0.05$. Portanto, a hipótese nula é rejeitada em ambos casos e pode-se dizer que há evidências para afirma que ambas variáveis, X e Y, não podem ser modeladas por uma distribuição de Poisson com coeficiente de confiança de 95%. Os resultados do modelo ajustado com os valores de λ da Tab. 3.3 são apresentados na Fig. 3.2.

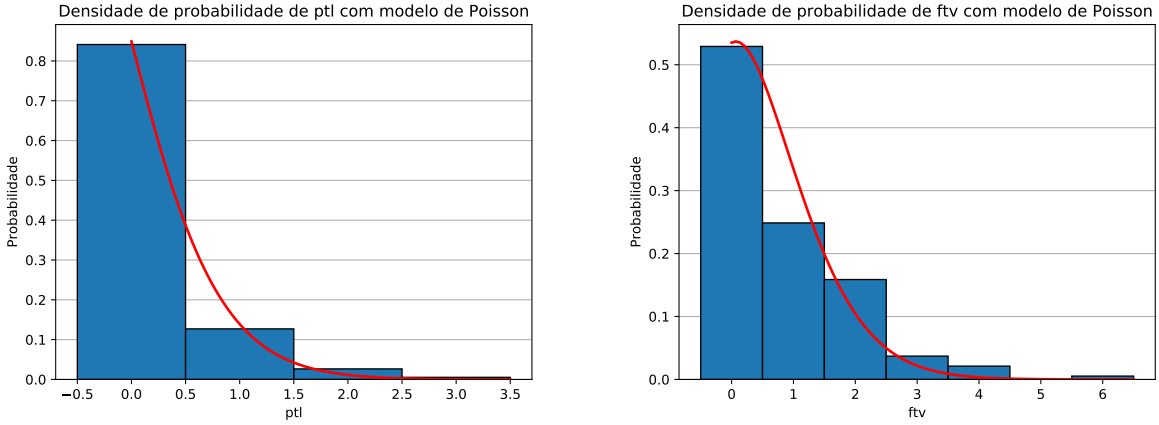

Figura 3.2: Modelo de Poisson utilizando valores de λ estimados para as variáveis *ptl* e *ftv*.

Questão 3)

São apresentados aqui as distribuições de frequência e o histograma para as variáveis: **(a)** *lwt* (peso da mãe na época da última menstruação) e **(b)** *bwt* (peso ao nascer da criança).

A Tab. 3.4 demonstram algumas mediadas descritivas para as variáveis *lwt* e *bwt*. Já a Fig. 3.3 demonstra os histogramas das duas variáveis em questão para mostrar a quantidade de casos que ocorrem em cada um dos diferentes valores presentes nos dados.

Tabela 3.4: Algumas medidas descritivas para as variáveis *lwt* e *bwt*

	lwt	**bwt**
Quantidade	189	189
Média	129.815	2,944.656
Dev. Padrão	30.579	729.022
Mínimo	80	709
25%	110	2,414
50%	121	2,977
75%	140	3,475
Máximo	250	4,990
Skewness	1.391	-0.208
Kurtosis	5.309	2.889

A Tab. 3.4 demonstra algumas informações descritivas importantes. Percebe-se uma amplitude muito maior dos dados quando considerado o peso das crianças em relação ao peso das mães. É interessante notar que os dados de *bwt* estão em gramas enquanto o *lwt* estão em libras.

As medidas de *Skewness* e *Kurtosis* da Tab. 3.3 são referentes às informações sobre assimetria dos dados. Há um indício teórico que diz que os dados seguem uma distribuição normal quando *Skewness* se aproxima de 0 e *Kurtosis* se aproxima de 3. Ao comparar essa informação teórica com os resultados obtidos na Tab. 3.4, percebe-se que a variável *bwt* se aproxima dos valores em questão. Sendo assim, há indícios que os dados *bwt* seguem uma distribuição normal. Essa informação ganha ainda mais força ao observar a Fig. 3.3 e perceber que, visualmente, os dados de *bwt* realmente aparentam seguir uma distribuição normal.

As observações feitas para *bwt* não podem ser estendidas para *lwt* da mesma maneira. Os dados da variável em questão apresenta valores de *Kurtosis* muito elevado (cerca de 5.3). Isso indica que os dados seguem uma distribuição assimétrica. Essa indicação é corroborada com o histograma da variável *lwt* pela

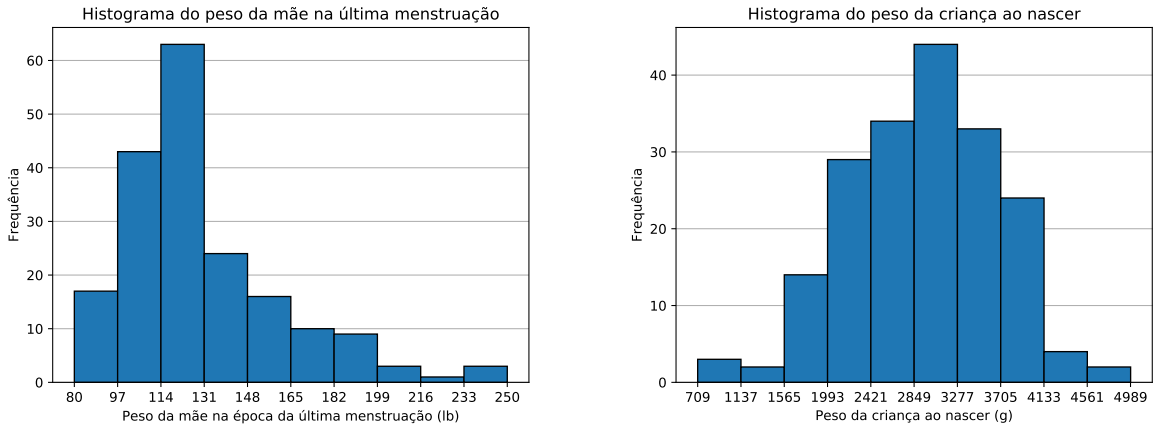

Figura 3.3: Algumas medidas descritivas para as variáveis *lwt* e *bwt*.

Fig. 3.3. Percebe-se visualmente pela figura em questão que os dados apresentam uma forte assimetria positiva.

Questão 4)

Sejam as as variáveis aleatórias X: peso da mãe na época da última menstruação e Y: peso ao nascer da criança. São avaliados se as variáveis podem ser modeladas por um modelo Normal. Os testes de hipóteses realizados, com nível de significância $\alpha = 5\%$, são:

$$H_0 : \text{a variável pode ser modelada por uma distribuição Normal}$$
$$H_1 : \text{a variável não pode ser modelada por uma distribuição Normal}$$

Para isso, são utilizados os testes de Shapiro-Wilk e Lilliefors considerando um nível de significância de $\alpha = 0.05$. Os seguintes valores foram encontrados:

Tabela 3.5: Testes de normalidade para as variáveis *lwt* e *bwt*

P-valores		
	X	**Y**
Shapiro	0.000	0.438
Lillefors	0.000	0.200

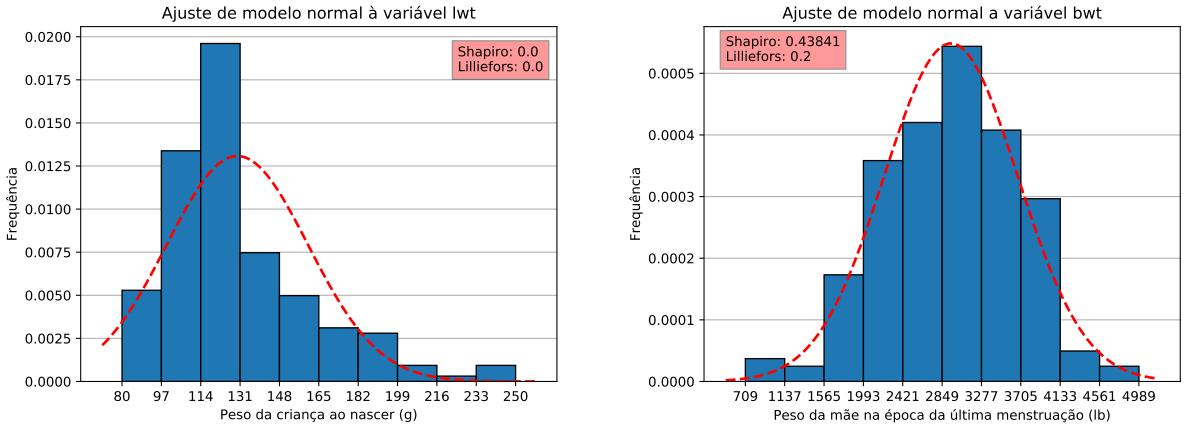

Figura 3.4: Ajuste do modelo Normal para as variáveis *lwt* e *bwt*.

Os resultados dos p-valores encontrados com os testes de Shapiro-Wilk e Lilliefors foram diferentes para X e para Y. Para X, os p-valores encontrado foi menor que o nível de significância α em ambos testes. Portanto, rejeita-se a hipótese nula e há evidências para inferir que X não é ajustado por um modelo normal com coeficiente de confiança de 95%. Já para Y os p-valores encontrados em ambos testes foram maiores que α. Portanto, não se pode rejeitar H_0, ou seja, existem evidências nos dados que leve a rejeitar a hipótese que a população siga o modelo normal com coeficiente de confiança de 95%. Pode-se verificar essas mesmas informações através da Fig. 3.4, onde há um histograma e uma curva normal, em vermelho, utilizando medidas das médias e desvios-padrão das variáveis *lwt* e *bwt*. Na figura, observa-se que o modelo normal se ajusta bem a variável *bwt* enquanto o mesmo não ocorre para *lwt*.

Transformações e recodificações dos dados

Neste momento, foram criados nova variáveis para a base de dados. Estas foram criadas e adicionadas a base de dados original com as seguintes regras e descrições:

- Variável *lwtkg:* transformação do peso *lwt* (peso da mãe na época da última menstruação) de libras para quilogramas (1 libra = 0,453 kg).
- Variável *race2:* reagrupamento das mães, a partir da variável *race*, em apenas duas classes (brancas / não-brancas).
- Variável *ptl2:* reagrupamento das mães, a partir da variável *ptl*, em apenas duas classes (nenhum parto prematuro / um ou mais partos prematuros).
- Variável *ftv2:* reagrupamento das mães, a partir da variável *ftv*, em apenas duas classes (nenhuma visita ao médico / uma ou mais visitas).

Análise usando *bwt* (variável quantitativa)

Nesta seção o objetivo é verificar se o peso ao nascer da criança difere entre as mães que estão expostas a um dado fator de risco, e as que não estão exposta. A Fig. 3.5 demonstram a comparação entre as variáveis binárias do eixo x, respondidas com sim ou não para os *labels* 1 ou 0, respectivamente.

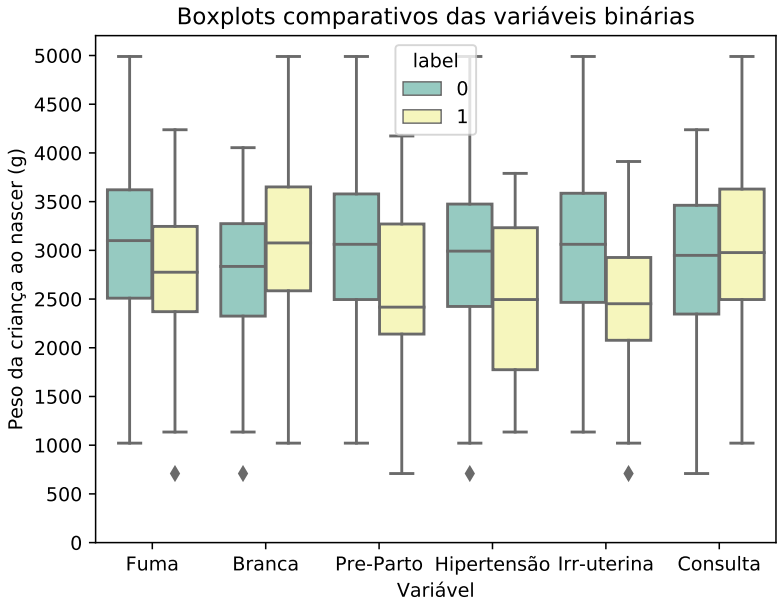

Figura 3.5: Boxplots comparativos entre as variáveis binárias em relação ao peso da criança ao nascer em gramas.

Os resultados, em geral, parecem fazer sentido. Os mais interessantes obtidos são aqueles referentes a mães fumantes e não fumantes, onde aquelas que não fumam têm filhos nascendo com pesos maiores do mães fumantes. O mesmo acontece com mães com hipertensão. Aquelas que não têm hipertensão apresentam filhos mais pesados em comparação a mães que têm hipertensão. O mesmo comportamento também aparece para mães que não têm irritação uterina. Todos esses resultados fazem sentido, visto que apresentar as características destacadas tende a ser algo, de alguma forma, prejudicial à saúde.

A variável que diz se uma mãe é branca mostra que, quando verdadeiro, os filhos nascem mais pesados. Isso talvez possa ser visto como um efeito social, uma vez que pessoas brancas tende a ter uma condição social melhor do que pessoas não brancas. As outras variáveis, mães que já tiveram filhos prematuros (pre-parto) e mães que visitaram o médico durante a gravidez (consulta) também apresenta resultados que fazem sentido, visto que essas variáveis estão relacionadas a histórico médicos.

Seja as varáveis aleatórias X: peso da criança quando a resposta é afirmativa para a variável binária e Y: peso da criança quando a resposta é negativa para a variável binária, deseja-se realizar o teste de média para os pesos das crianças ao nascer verificando, para cada uma das variáveis binárias, se apresentam valores de médias iguais quando a resposta é positiva ou negativa. Sendo assim, realiza-se o seguinte teste de hipótese com nível de significância $\alpha = 5\%$:

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

Os resultados do teste para cada variável encontra-se na Tab. 3.6. Pode-se observar que para todas as variáveis (exceto a variável consulta) os valores de Z observado foram maiores do que o Z crítico. Com isso, os valores entram na área de rejeição de H_0. Isso também fica evidente através dos p-valores encontrados. Para todas as variáveis (exceto a variável consulta) os p-valores são menores que o nível de significância $\alpha = 0.05$. Sendo assim, para essas variáveis, rejeita-se a hipótese nula. Portanto, existem evidências pelas amostras para dizer que a média de pesos das crianças ao nascer são diferentes quando considera-se o valor da variável binária igual a 1 ou igual a 0 com coeficiente de confiança de 95%.

Já para a variável consulta (visitas ao médico durante a gravidez), o valor de Z observado é menor do que o Z crítico. Sendo assim, rejeita-se H_0. Através do p-valor, observa-se que o valor encontrado de 0.111 é maior que o nível de significância $\alpha = 0.05$, reforçando a informação da rejeição da hipótese nula. Portanto, nesta variável, há evidências para acreditar que a média de peso das crianças ao nascer são iguais para casos onde a mulher foi ao médico ou não no primeiro trimestre da gravidez com coeficiente de confiança de 95%.

Tabela 3.6: Testes de médias para as variáveis binárias

Variável	**Z observado**	**Z crítico**	**p-valor**
Fuma	2.634	1.645	0.008
Branca	3.118	1.645	0.002
Pre-Parto	3.058	1.645	0.002
Hipertensão	2.019	1.645	0.043
Irr-uterina	4.042	1.645	0.000
Consulta	1.593	1.645	0.111

Análise usando *low* (variável qualitativa)

A Tab. 3.7 apresenta as tabelas 2x2 associando o fato de a criança ter tido baixo peso ao nascer (*low*) e outras variáveis nominais binárias. As variáveis consideradas são: *smoke, race2, ptl2, ht, ui* e *ftv2*. As tabelas consideram os valores 0 para falso e 1 para verdadeiro.

As subtabelas da Tab. 3.7 parecem fazer sentido. Abaixo uma explicação dessa afirmação, detalhando a relação entre cada variável binária o fato da criança ter tido baixo peso ao nascer:

- Com relação ao fato de a mãe fumar ou não, percebe-se que para as mães que não fumam 86 crianças nasceram com peso normal, enquanto 29 com baixo peso. Além disso, 30 crianças nascem com peso baixo para as mães que fumam, enquanto 29 para as mães que não fumam.

- Com relação à raça, tanto para mães bancas quanto para mães não-brancas são maiores os números de crianças que nascem com peso normal. Entretanto, verifica-se que para mães brancas o número

Tabela 3.7: Tabelas 2x2 associando o fato de a criança ter tido baixo peso ao nascer (*low*) e outras variáveis nominais binárias

low *smoke*	**0**	**1**	**All**
0	86	29	115
1	44	30	74
All	130	59	189

low *race2*	**0**	**1**	**All**
0	57	36	93
1	73	23	96
All	130	59	189

low *ptl2*	**0**	**1**	**All**
0	118	41	159
1	12	18	30
All	130	59	189

low *ht*	**0**	**1**	**All**
0	125	52	177
1	5	7	12
All	130	59	189

low *ui*	**0**	**1**	**All**
0	116	45	161
1	14	14	28
All	130	59	189

low *ftv2*	**0**	**1**	**All**
0	64	36	100
1	66	23	89
All	130	59	189

de crianças que nascem com peso normal é maior do que para mães não brancas. O inverso ocorre quando considera-se crianças que nascem com baixo peso. Portanto, também há certa associação entre as variáveis "peso" e "raça"que pode estar ligado ao fato social.

- Para as mães que não tiveram nenhum parto prematuro é maior o número de crianças que nascem com peso normal, se comparado ao número de crianças de baixo peso. Em contrapartida, para mães que tiveram um ou mais partos prematuras, o número de crianças abaixo do peso é maior, o que nos mostra a relação e associação entre as variáveis "peso da criança"e "histórico de partos prematuros". Talvez isso ocorra por alguma questão médica.

- O fato de a mãe ser hipertensa apresenta associação significativa com a variável peso da criança. Nota-se que o número de crianças com baixo peso nascidas de mães hipertensas é maior se comparado ao número de crianças com peso normal. Por outro lado, para as mães não hipertensas, 125 crianças nasceram com peso normal, enquanto apenas 52 com baixo peso, o que confirma esta associação. Isso pode ocorrer por questões de saúde.

- A irritabilidade uterina da mãe também tem dependência com a variável peso da criança, uma vez que, para mães que não tem irritabilidade o número de crianças que nascem com peso normal é maior se comparado àquelas que nascem fora do peso. Isso também pode ocorrer por questões de saúde.

- O número de visitas ao médico parece não fazer associação com a variável peso da criança, uma vez que tanto para mães que visitaram quanto para as que não realizaram nenhuma visita, o número de crianças com peso normal é significativamente maior se comparado às crianças com baixo peso. Talvez isso também ocorra por alguma questão médica.

Para confirmar as observações de dependências descritas acima, realiza-se testes de dependências entre as variáveis. Seja as varáveis aleatórias X : informação de uma criança nascer com baixo peso e Y : cada uma das variáveis binárias, o teste de dependência é realizado seguindo as hipóteses com nível de significância $\alpha = 5\%$:

$$H_0 : \text{as variáveis } X \text{ e } Y \text{ são independentes}$$
$$H_1 : \text{as variáveis } X \text{ e } Y \text{ são dependentes}$$

O teste é realizado utilizando a tabela qui-quadrado (Q^2). O valor de Q^2 é calculado da seguinte forma:

$$Q^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Onde o representa os dados observados e e os valores esperado. Ambos vetores foram omitidos nesse relatório, o qual avalia apenas os resultados finais obtidos com o teste.

A Tab. 3.8 demonstra os valores obtidos com os testes para cada um das variáveis binárias. Considerando os resultados encontrados com um nível de significância $\alpha = 5\%$, pode-se perceber que apenas as variáveis *ht* (histórico de hipertensão) e *ftv2* (visitas ao médico durante o primeiro trimestre da gravidez) não apresentaram p-valores menores do que os α. Portanto, para essas variáveis, não rejeita-se H_0. Sendo assim, não

existem evidências nas amostras para relacionar o fato das crianças que nascem com baixo peso com o fato das mães serem hipertensas ou terem consultas com médicos com coeficiente de confiança de 95%.

Tabela 3.8: Resultados do teste de dependência para cada variável binária

Variável	**Q2**	**p-valor**	**df**
smoke	4.235	0.0396	1
race2	4.125	0.0423	1
ptl2	12.21	0.0004	1
ht	3.143	0.0763	1
ui	4.423	0.0355	1
ftv2	1.814	0.1780	1

Para as outras variáveis, há evidências para não rejeitar a hipótese nula. Sendo assim, existem evidências nas amostras para acreditar que as variáveis X e Y são independentes com coeficiente de confiança de 95%.

Análise de variância

Para verificar se o peso ao nascer da criança (*bwt*) depende da raça da mãe (codificada na variável *race* em três classe : 1 = *white*, 2 = *black* e 3 = *other*), é realizado o teste de variância pela tabela ANOVA. Os pressupostos para esse testes são:

- A variável de interesse para comparação deve ser contínua. Nesse caso, *bwt* é a variável;
- Os testes comparativos baseiam-se na separação de subconjunto de valores em duas ou mais variáveis categóricas. Nesse caso a variável é a *race*;
- As observações de *bwt* devem ser únicas em *race*, isto é, uma observação não pode estar presente em mais de um grupo ao mesmo tempo;
- A variável de interesse, *bwt*, devem seguir distribuições próximas a Normais para cada uma das categórias de *race*. Os testes são realizados com a hipótese de que os dados podem ser modelados por uma distribuição normal. Dessa forma, usa-se os testes de Shapiro e Lillefors para verificar se os p-valores de ambos são maiores do que o nível de significância α desejado.
- As variâncias de cada grupo devem ser consideradas aproximadamente iguais. Para isso, realiza-se a verificação se há uma relação de pelo menos 4 vezes uma variável da outra ou realiza-se o teste de *Bartlett*.

Neste momento, é assumido que todos os pressupostos são verdadeiros para as amostras. Considerando as classes categórias de *race* como x, y, z, é realizado o seguinte teste de hipóteses para as médias:

$$H_0 : \mu_X = \mu_Y = \mu_Z$$
$$H_1 : \mu_X \neq \mu_Y \neq \mu_Z$$

As conclusões para este teste podem ser retiradas do teste ANOVA, cujo resultados seguem na Tab. 3.9.

Tabela 3.9: Tabela ANOVA para dependência do peso da criança ao nascer com a raça.

Tabela ANOVA				
	SS	**df**	**F**	**PR(>F)**
C(race)	5,070,607.632	2	4.972	0.008
Residual	94,846,445.013	186	-	-

Considerando um nível de significância $\alpha = 5\%$ e observando o p-valor encontrado de $0.008 < \alpha$, rejeita-se H_0. Portanto, existem evidências pelas amostras para acreditar que as médias de *bwt* sejam diferentes entre os diferentes grupos de raça com coeficiente de confiança de 95%.

Correlação e regressão

Neste momento, deseja-se investigar se o peso ao nascer da criança (*bwt*) está relacionado ao peso da mãe (*lwt*) ou à sua idade (*age*). Para isso, são utilizados modelos de regressão linear. Primeiramente apresentam-se os gráficos de dispersão para as relações (*bwt, lwt*) e (*bwt, age*) através da Fig. 3.6.

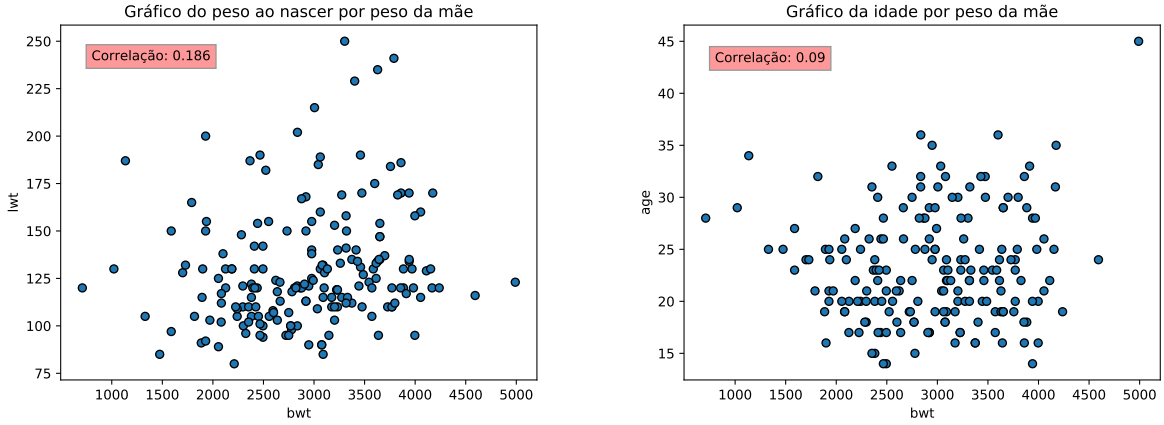

Figura 3.6: Ajuste do modelo Normal para as variáveis *lwt* e *bwt*.

Percebe-se que a correlação entre o peso da criança ao nascer com o peso da mãe tem o valor pouco maior que a correlação entre peso da criança com a idade da mãe. Isso ocorre pois os valores de *lwt* e *age* estão limitados em uma faixa de valor de *bwt*.

Para verificar se esses resultados são estatisticamente significativos, realiza-se testes com ANOVA para verificar dependência entre as variáveis. Considerando as variáveis aleatórias X: peso da criança ao nascer e Y: o peso da mãe ou idade, o teste de dependência é realizado seguindo as hipóteses, com nível de significância de $\alpha = 5\%$:

$$H_0 : \text{as variáveis } X \text{ e } Y \text{ são independentes}$$
$$H_1 : \text{as variáveis } X \text{ e } Y \text{ são dependentes}$$

Realizando o teste ANOVA para ambos pares [(*bwt, lwt*) e (*bwt, age*)], encontram-se os valores conforme a Tab. 3.10 e Tab. 3.11.

Tabela 3.10: Tabela ANOVA para dependência do peso da criança ao nascer com o peso da mãe.

Tabela ANOVA				
	SS	**df**	**F**	**PR(>F)**
lwt	3,448,881.301	1	6.686	0.010
Residual	96,468,171.344	187	-	-

Tabela 3.11: Tabela ANOVA para dependência do peso da criança ao nascer com a idade da mãe.

Tabela ANOVA				
	SS	**df**	**F**	**PR(>F)**
age	806,926.913	1	1.523	0.219
Residual	99,110,125.732	187	-	-

Os resultados de ambas tabelas (Tab. 3.10 e Tab. 3.11) retornam p-valores menores do que o nível de significância $\alpha = 0.05$ considerado. Isso indica que a hipótese nula é rejeitado no teste de dependência entre (*bwt, lwt*). Portanto, existem evidencias para acreditar que uma variável é capaz de explicar bem a outra com coeficiente de confiança de 95%. Já para a correlação entre (*bwt, age*) demonstra que não há evidências para rejeitar H_0 e, portanto, pode-se acreditar que as variáveis são independentes com coeficiente de confiança de 95%.

Os resultados foram um pouco surpreendentes visto que os coeficientes de correlações das variáveis eram baixos. Sendo assim, esperava-se que não fosse possível rejeitar a hipótese nula em ambos os casos. Também foram realizados regressões lineares nos dois pares de correlações através da Fig. 3.7.

Os resultados de coeficiente linear em ambos os casos, expressos pela Fig. 3.7, corroboram com a ideia de independência das variáveis levantada pela Fig. 3.6, onde demonstrou que os dados tinham correlações baixas. Porém os resultados do modelo contrariam os resultados obtidos com a tabela ANOVA na Tab. 3.10 e Tab. 3.11.

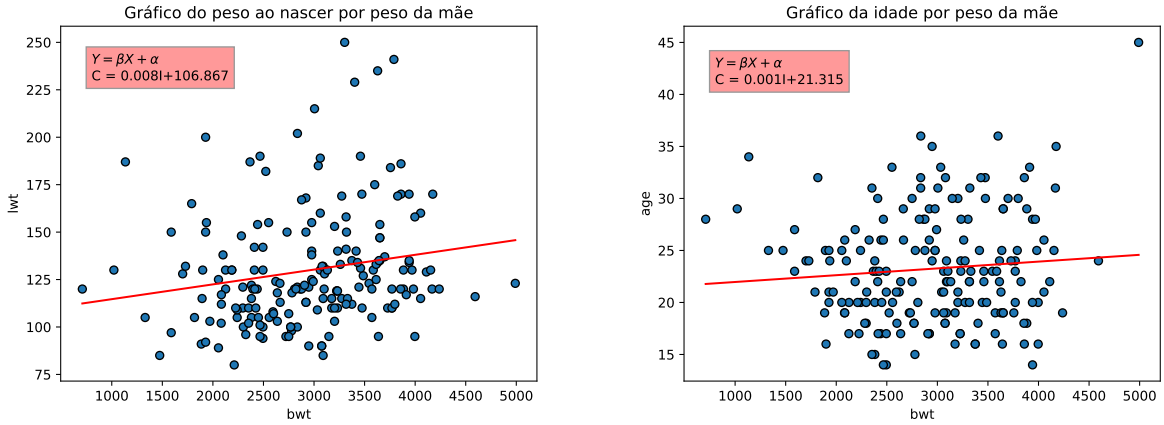

Figura 3.7: Ajuste do modelo Normal para as variáveis *lwt* e *bwt*.

Referências

Hippert, H. S. **Notas de aula:** Métodos Estatísticos, 2019.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. The multiple logistic regression model. **Applied logistic regression**, v. 1, p. 25-37, 1989.

MAGALHÃES, M. N.; DE LIMA, A. C. P. **Noções de probabilidade e estatística**. 7 ed. São Paulo: Editora da Universidade de São Paulo, 2002.