UNIVERSIDADE DE ÉVORA

Trabalho de IPE

Professora Dulce Gomes

1º ano

Engenharia Informática

(2012)

Ricardo Fusco, nº 29263 EI
Marcus Santos, nº 29764 EI

Introdução

Este trabalho, no âmbito da disciplina de IPE, tem como objectivo principal a análise e processamento de dados provenientes de uma base de dados acerca de pedidos de indemnização junto de uma seguradora devido a acidentes de automóvel que nos foi disponibilizada pela docente da disciplina ($n=128$), tendo também como objectivo, obviamente, a familiarização com o funcionamento do programa SPSS com o qual irão ser tratados os dados.

Na primeira fase do trabalho compete-nos estudar e classificar devidamente as variáveis em estudo seguido da selecção das variáveis que necessitam de ser agrupadas em classes visto que têm um leque de valores muito elevado resultando numa tabela de frequências desproporcionalmente grande em relação às outras variáveis, e por fim elaborar as tabelas de dupla entrada necessárias para podermos, eficientemente, estudar os dados e retirar conclusões.

Numa segunda fase do trabalho iremos proceder à analise estatistica descritiva de todas as variáveis, verificando as tabelas, todos os quantis, os outliers, os histogramas, entre outros. Poderá ser também útil converter as frequências absolutas ou relativas, nos casos em que as classes têm amplitudes diferentes, para que efectivamente se verifique a proporcionalidade entre a altura das barras e a sua base e, também, para que se garanta que a área das barras seja igual a n ou a 1 (n se se estiver a usar as frequências absolutas e 1 no caso das frequências relativas)

Após todos estes passos poderemos então retirar conclusões que nos permitirão responder às questões colocadas no enunciado do trabalho.

Desenvolvimento

Neste tipo de estudos é extremamente importante identificar a natureza de todas as variáveis avaliadas para que não surjam erros. Relativamente às variáveis a avaliar podemos afirmar que o Nº de acidentes nos últimos 5 anos se trata de uma variável quantitativa discreta visto que assume valores dentro de um tempo finito ou enumerável, sendo estes valores tipicamente números inteiros, e quanto às restantes, a Idade do polícia, a Idade do veículo, o Nº de coimas, a idade do condutor e o Valor das coimas são variáveis quantitativas contínuas, o tipo de veículo e o género são ambas variáveis qualitativas nominais visto que não obedecem uma ordenação específica.

Estatística descritiva

Antes de efectuarmos as tabelas de frequências há a necessidade de agrupar em classes as variáveis nº de coimas, valor das coimas e a idade do condutor (como já tinhamos referido anteriormente na introdução):

- Nº de coimas ($n=128$):

$$k = \left[\frac{\ln 128}{ln2}\right] + 1 = 8 \qquad \Delta = max - min = 434 - 0 = 434$$

$$A = \frac{\Delta}{k} = \frac{434}{8} = 54.25$$

- Valor das coimas ($n=128$):

$$k = \left[\frac{\ln 128}{ln2}\right] + 1 = 8 \qquad \Delta = max - min = 850 - 0 = 850$$

$$A = \frac{\Delta}{k} = \frac{850}{8} = 106.25$$

- Idade do condutor ($n=128$):

$$k = \left[\frac{\ln 128}{ln2}\right] + 1 = 8 \qquad \Delta = max - min = 49 - 22 = 27$$

$$A = \frac{\Delta}{k} = \frac{27}{8} = 3.375$$

Statistics

		Idade do policia	Idade do veículo	Valor das coimas	Número de coimas	Idade Condutor	Número de acidentes nos últimos 5 anos
N	Valid	128	128	123	128	128	128
	Missing	0	0	5	0	0	0
Mean		4,50	2,50	231,14	69,86	34,89	1,97
Median		4,50	2,50	213,00	35,50	35,00	2,00
Mode		1[a]	1[a]	110[a]	1[a]	38	1[a]
Std. Deviation		2,300	1,122	117,048	91,852	5,780	1,516
Variance		5,291	1,260	13700,218	8436,783	33,405	2,298
Skewness		,000	,000	2,313	2,049	,038	,536
Std. Error of Skewness		,214	,214	,218	,214	,214	,214
Kurtosis		-1,239	-1,366	9,225	4,148	-,240	-,416
Std. Error of Kurtosis		,425	,425	,433	,425	,425	,425
Minimum		1	1	11	0	22	0
Maximum		8	4	850	434	49	6
Percentiles	25	2,25	1,25	157,00	9,00	31,00	1,00
	50	4,50	2,50	213,00	35,50	35,00	2,00
	75	6,75	3,75	274,00	96,75	38,00	3,00

a. Multiple modes exist. The smallest value is shown

Figura 1 – Tabela com os valores dos quantis, da média, da mediana, da moda, do desvio padrão e da variância e outras medidas descritivas (Nota: Os valores dos quantis, quando calculados à mão deram ligeiramente diferentes dos valores fornecidos pelo SPSS).

Statistics

		Número de coimas (em classes)	Valor das coimas (em classes)	Idade Condutor (em classes)
N	Valid	128	123	128
	Missing	0	5	0
Mean		1,91	2,69	4,30
Median		1,00	3,00	4,00
Mode		1	2	5
Std. Deviation		1,619	1,095	1,713
Variance		2,621	1,199	2,935
Skewness		2,190	2,051	,033
Std. Error of Skewness		,214	,218	,214
Kurtosis		4,580	7,879	-,353
Std. Error of Kurtosis		,425	,433	,425
Minimum		1	1	1
Maximum		8	8	8
Percentiles	25	1,00	2,00	3,00
	50	1,00	3,00	4,00
	75	2,00	3,00	5,00

Figura 2 – Tabela com os valores dos quantis, da média, do desvio padrão e da variância e outras medidas descritivas (Nota: Os valores dos quantis, quando calculados à mão deram ligeiramente diferentes dos valores fornecidos pelo SPSS).

Idade do veículo

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0-3	32	25,0	25,0	25,0
	4-7	32	25,0	25,0	50,0
	8-9	32	25,0	25,0	75,0
	10+	32	25,0	25,0	100,0
	Total	128	100,0	100,0	

Figura 3 – Tabela de frequências da variável *IdVeículo*.

Idade do policia

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	17-20	16	12,5	12,5	12,5
	21-24	16	12,5	12,5	25,0
	25-29	16	12,5	12,5	37,5
	30-34	16	12,5	12,5	50,0
	35-39	16	12,5	12,5	62,5
	40-49	16	12,5	12,5	75,0
	50-59	16	12,5	12,5	87,5
	60+	16	12,5	12,5	100,0
	Total	128	100,0	100,0	

Figura 4 – Tabela de frequências da variável *Idpolícia*.

✓ Podemos observar nas duas tabelas anteriores (figuras 3 e 4) que há uma probabilidade igual para cada uma das classes, ou seja, cada classe tem a mesma frequência apesar de não terem a mesma amplitude, facto este que vamos explorar mais à frente para podermos obter um histograma "fidedigno", pois os valores destas duas tabelas são subtilmente falaciosos no que diz respeito à interpretação dos seus respectivos histogramas.

Número de acidentes nos últimos 5 anos

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	24	18,8	18,8	18,8
	1	31	24,2	24,2	43,0
	2	31	24,2	24,2	67,2
	3	19	14,8	14,8	82,0
	4	15	11,7	11,7	93,8
	5	6	4,7	4,7	98,4
	6	2	1,6	1,6	100,0
	Total	128	100,0	100,0	

Figura 5 – Tabela de frequências da variável *Acidentes*.

Tipo de veículo

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	32	25,0	25,0	25,0
	B	32	25,0	25,0	50,0
	C	32	25,0	25,0	75,0
	D	32	25,0	25,0	100,0
	Total	128	100,0	100,0	

Figura 6 – Tabela de frequências da variável *Tipo de veículo*.

Género do condutor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Masculino	58	45,3	45,3	45,3
	Feminino	70	54,7	54,7	100,0
	Total	128	100,0	100,0	

Figura 7 – Tabela de frequências da variável *Género*.

✓ De acordo com as tabelas anteriores podemos concluir que:

- Mais de metade das pessoas seguradas tiveram entre zero e dois acidentes (inclusivé) nos últimos cinco anos sendo menos frequente terem tido entre três e seis acidentes (inclusivé) nos últimos cinco anos (fig. 5).
- Não há um tipo de veiculo que seja mais frequente entre os segurados visto que a frequência de pessoas com diferentes tipos de veiculos é idêntica (figs. 6 e 25).
- Dos 128 segurados 70 são do sexo feminino e 58 são do sexo masculino (figs. 7 e 26).
- Tendo em conta que as variáveis Género do condutor e Tipo de veículo (fig. 6 e 7) são variáveis qualitativas não faria qualquer sentido estar a calcular médias, desvio – padrão, quantis, etc. Foi esta a razão pelo qual excluimos estas duas variáveis das duas primeiras tabelas (fig. 1 e 2).

Número de coimas (em classes)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 54	81	63,3	63,3	63,3
	54 - 108	21	16,4	16,4	79,7
	109 - 162	7	5,5	5,5	85,2
	163 - 216	9	7,0	7,0	92,2
	217 - 270	4	3,1	3,1	95,3
	271 - 325	1	,8	,8	96,1
	326 - 379	2	1,6	1,6	97,7
	380+	3	2,3	2,3	100,0
	Total	128	100,0	100,0	

Figura 8 – Tabela de frequências da variável *Ncoimas (em classes)*.

Valor das coimas (em classes)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 106	7	5,5	5,7	5,7
	106 - 212	54	42,2	43,9	49,6
	213 - 318	42	32,8	34,1	83,7
	319 - 424	17	13,3	13,8	97,6
	531 - 637	1	,8	,8	98,4
	744+	2	1,6	1,6	100,0
	Total	123	96,1	100,0	
Missing	System	5	3,9		
Total		128	100,0		

Figura 9 – Tabela de frequências da variável Valor das coimas (em classes).

Idade Condutor (em classes)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	< 25	8	6,3	6,3	6,3
	25 - 28	11	8,6	8,6	14,8
	29 - 31	24	18,8	18,8	33,6
	32 - 35	23	18,0	18,0	51,6
	36 - 38	32	25,0	25,0	76,6
	39 - 41	20	15,6	15,6	92,2
	42 - 45	4	3,1	3,1	95,3
	46+	6	4,7	4,7	100,0
	Total	128	100,0	100,0	

Figura 10 – Tabela de frequências da variável Idcondutor (em classes).

✓ Com todas as tabelas já preparadas e com as devidas variáveis já agrupadas em classes (fig. 8, 9 e 10), podemos avançar então para o estudo dos outliers de cada variável para podermos responder adequadamente à segunda questão. À partida podemos facilmente observar que existem duas variáveis que nos vão impossibilitar a detecção de outliers visto que são variáveis qualitativas que é o caso das variáveis *Género* e *Tipo de automóvel*. Tendo isto em conta precederemos então ao calculo dos outliers para as restantes variáveis. Para este efeito recorremos às tabelas com os outliers e às caixas de bigodes

Extreme Values

			Case Number	Value
Idade do policia	Highest	1	52	8
		2	54	8
		3	61	8
		4	69	8
		5	79	8[a]
	Lowest	1	65	1
		2	51	1
		3	44	1
		4	35	1
		5	33	1[b]

a. Only a partial list of cases with the value 8 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

Extreme Values

			Case Number	Value
Idade do veículo	Highest	1	4	4
		2	7	4
		3	12	4
		4	16	4
		5	20	4[a]
	Lowest	1	125	1
		2	121	1
		3	120	1
		4	119	1
		5	104	1[b]

a. Only a partial list of cases with the value 4 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

Figuras 11 e 12 – Extremos das variáveis Idpolicia e Idveiculo.

Extreme Values

			Case Number	Value
Número de acidentes nos últimos 5 anos	Highest	1	23	6
		2	76	6
		3	28	5
		4	33	5
		5	95	5[a]
	Lowest	1	127	0
		2	123	0
		3	118	0
		4	117	0
		5	114	0[b]
Idade Condutor (em classes)	Highest	1	123	8
		2	124	8
		3	125	8
		4	126	8
		5	127	8[c]
	Lowest	1	8	1
		2	7	1
		3	6	1
		4	5	1
		5	4	1[d]

a. Only a partial list of cases with the value 5 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

c. Only a partial list of cases with the value 8 are shown in the table of upper extremes.

d. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

Figura 14 – Extremos das variáveis Acidentes e Idcondutor (em classes).

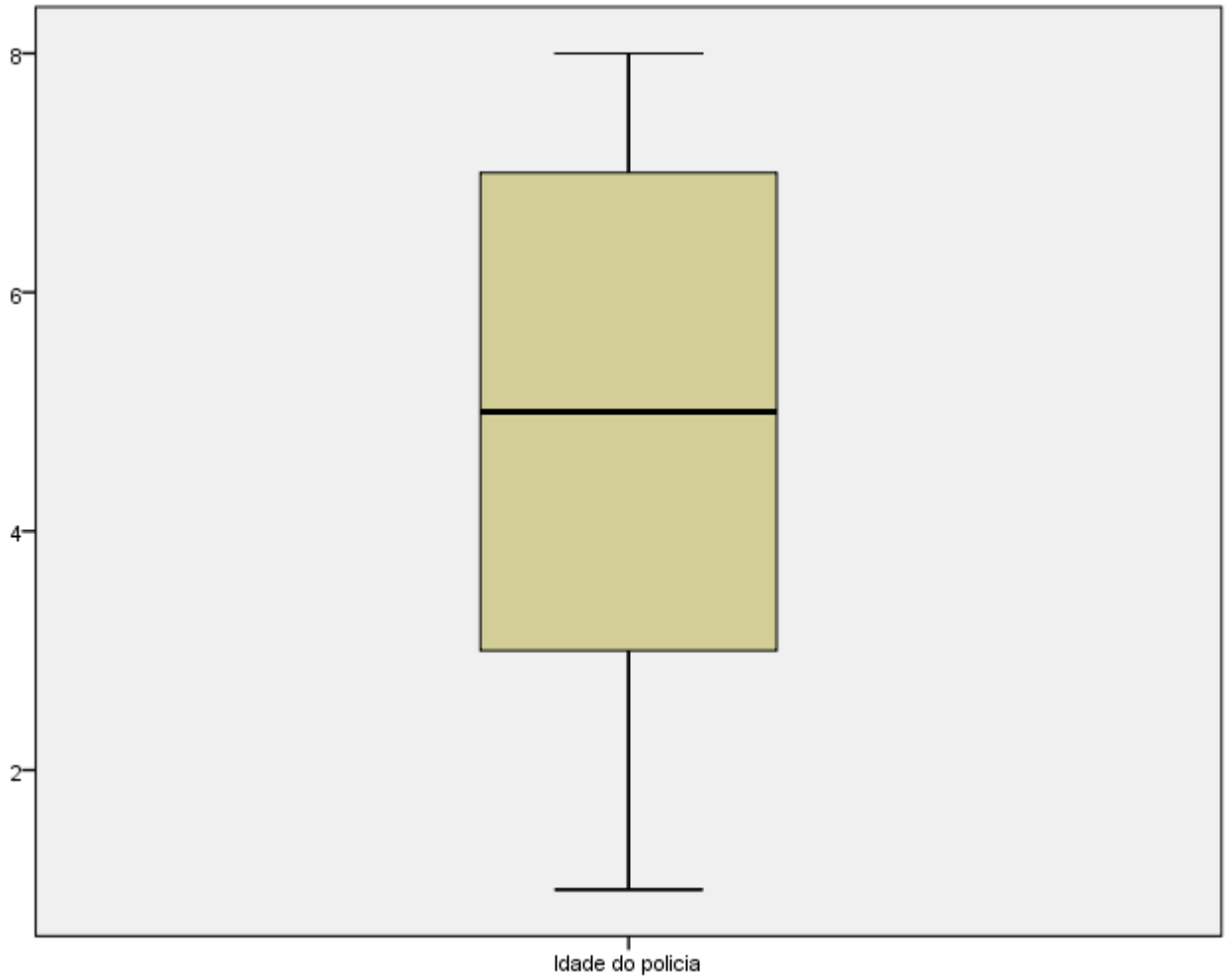


Figura 15 – Caixa de bigodes da variável Idpolícia.

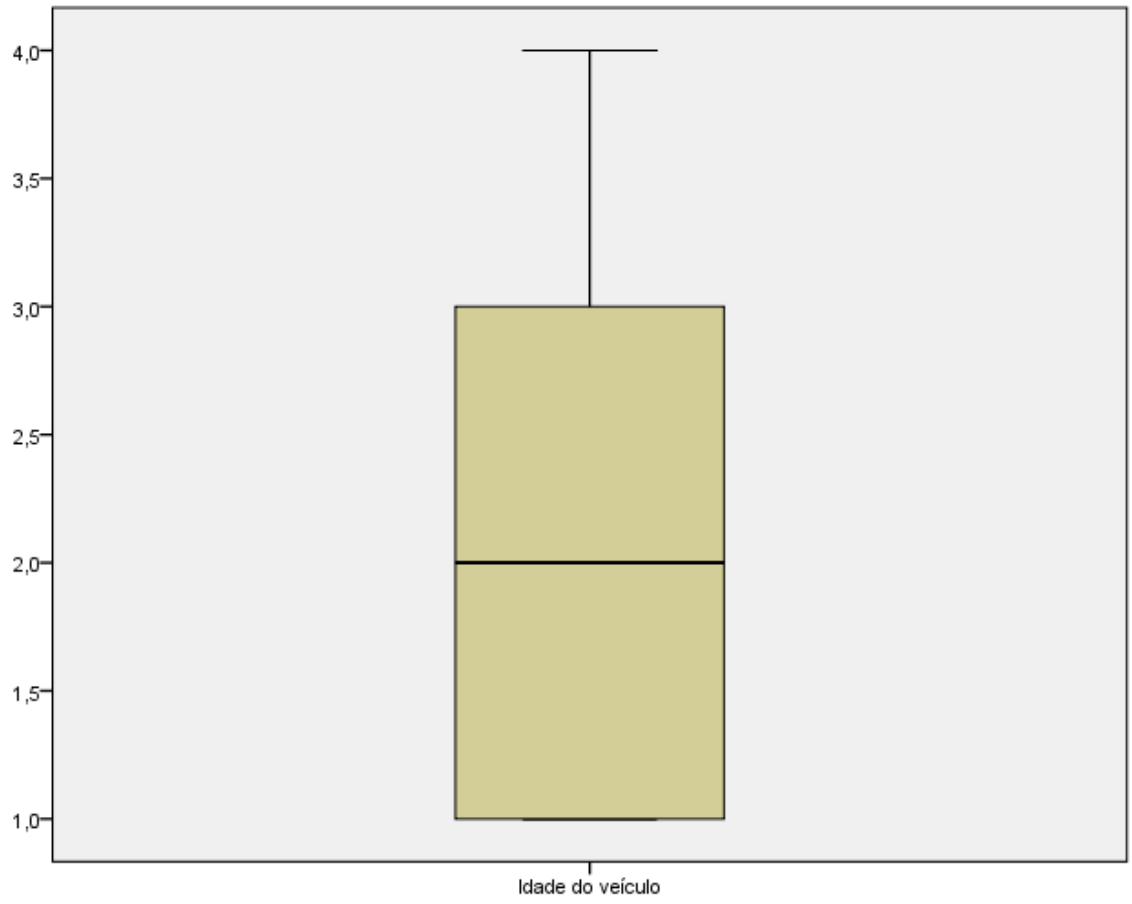


Figura 16 – Caixa de bigodes da variável Idveículo.

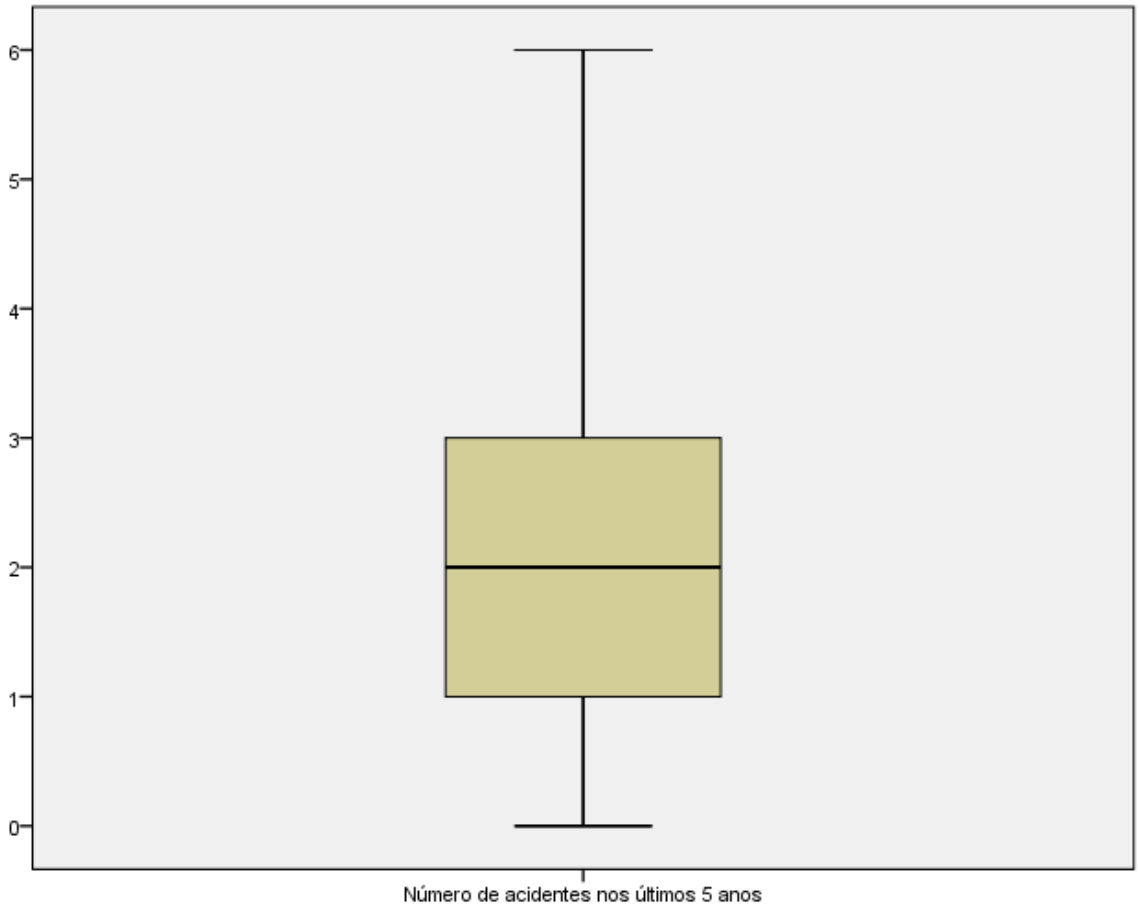


Figura 17 – Caixa de bigodes da variável Acidentes.

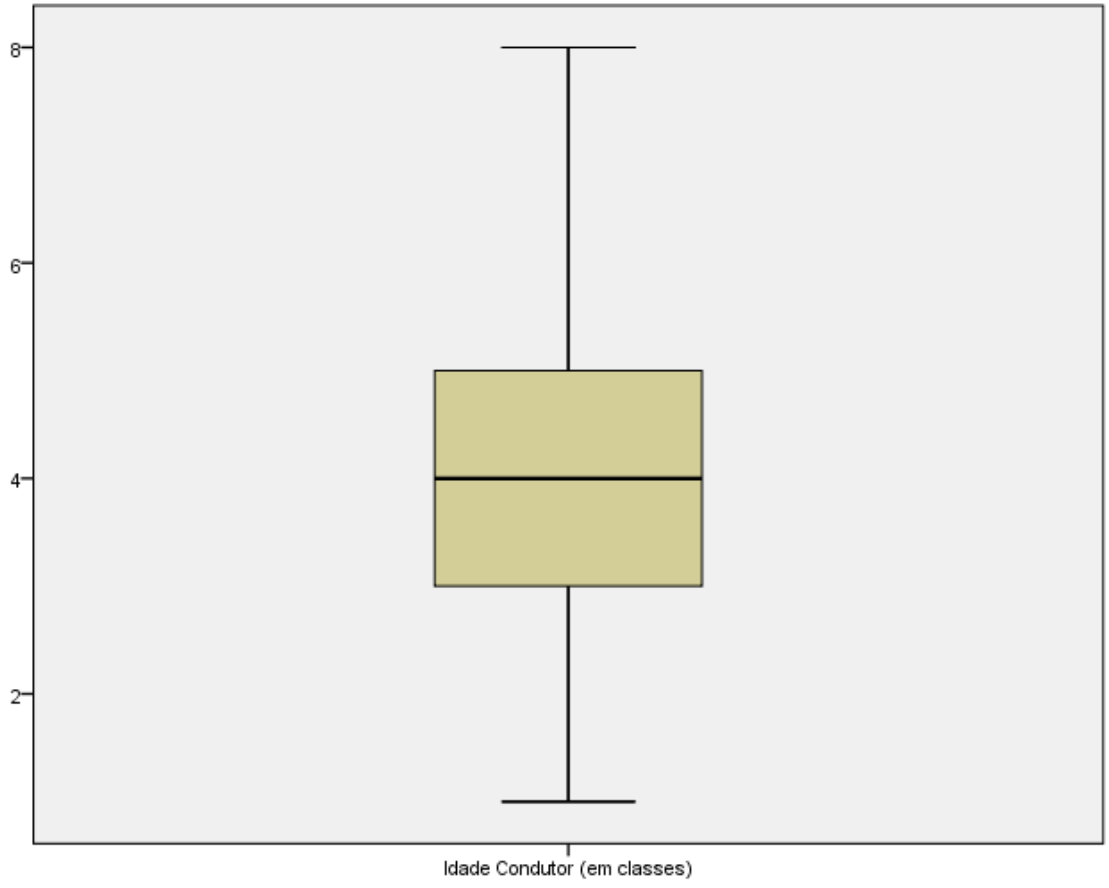


Figura 18 – Caixa de bigodes da variável Idcondutor (em classes).

Extreme Values

			Case Number	Value
Valor das coimas	Highest	1	28	850
		2	15	763
		3	38	636
		4	14	420
		5	37	407
	Lowest	1	29	11
		2	30	65
		3	110	98
		4	75	98
		5	52	101
Número de coimas	Highest	1	124	434
		2	80	401
		3	92	380
		4	70	366
		5	101	353
	Lowest	1	114	1
		2	38	1
		3	29	1
		4	10	1
		5	7	1[a]
Valor das coimas (em classes)	Highest	1	15	8
		2	28	8
		3	38	6
		4	14	4
		5	19	4[c]
	Lowest	1	110	1
		2	76	1
		3	75	1
		4	52	1
		5	45	1[a]

Extreme Values

			Case Number	Value
Idade Condutor	Highest	1	128	49
		2	126	48
		3	127	48
		4	125	47
		5	123	46[b]
	Lowest	1	1	22
		2	3	23
		3	2	23
		4	5	24
		5	4	24
Número de coimas (em classes)	Highest	1	80	8
		2	92	8
		3	124	8
		4	70	7
		5	101	7
	Lowest	1	128	1
		2	127	1
		3	126	1
		4	123	1
		5	122	1[a]

a. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

b. Only a partial list of cases with the value 46 are shown in the table of upper extremes.

c. Only a partial list of cases with the value 4 are shown in the table of upper extremes.

Figura 19 – Extremos das variáveis Coimas, Ncoimas, Coimas (em classes), Idcondutor e Ncoimas (em classes)

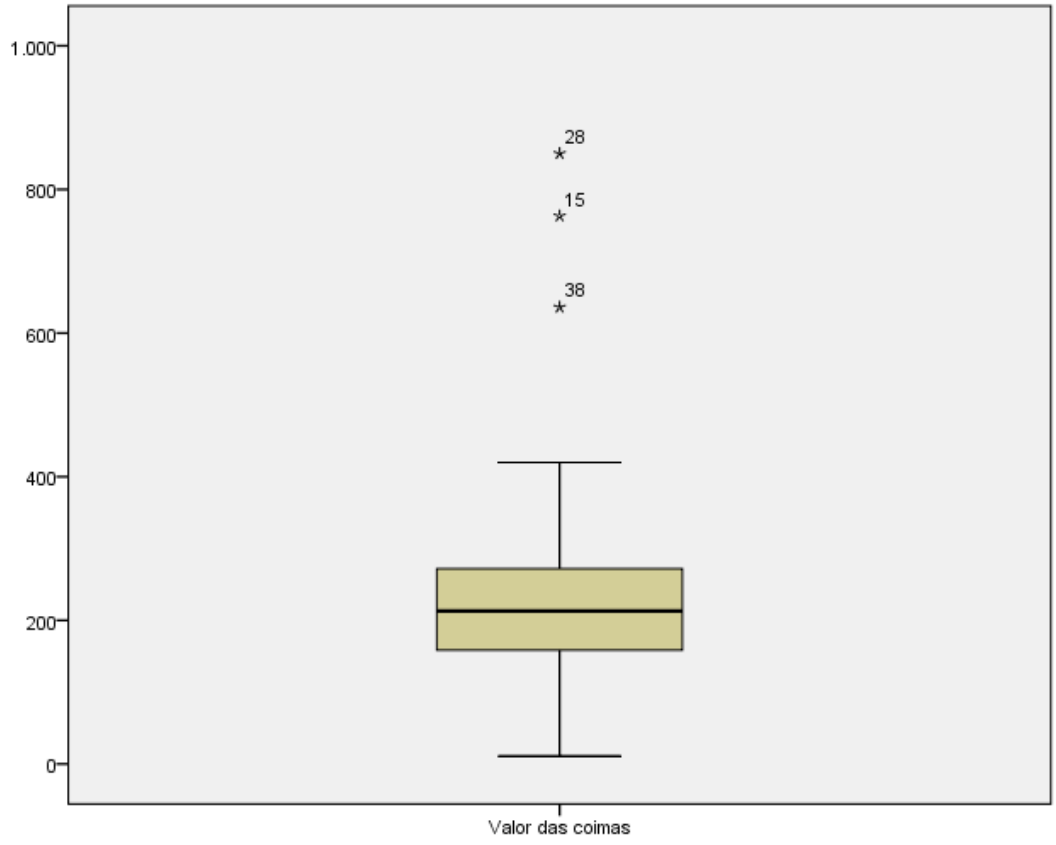


Figura 20 – Caixa de bigodes da variável Coimas.

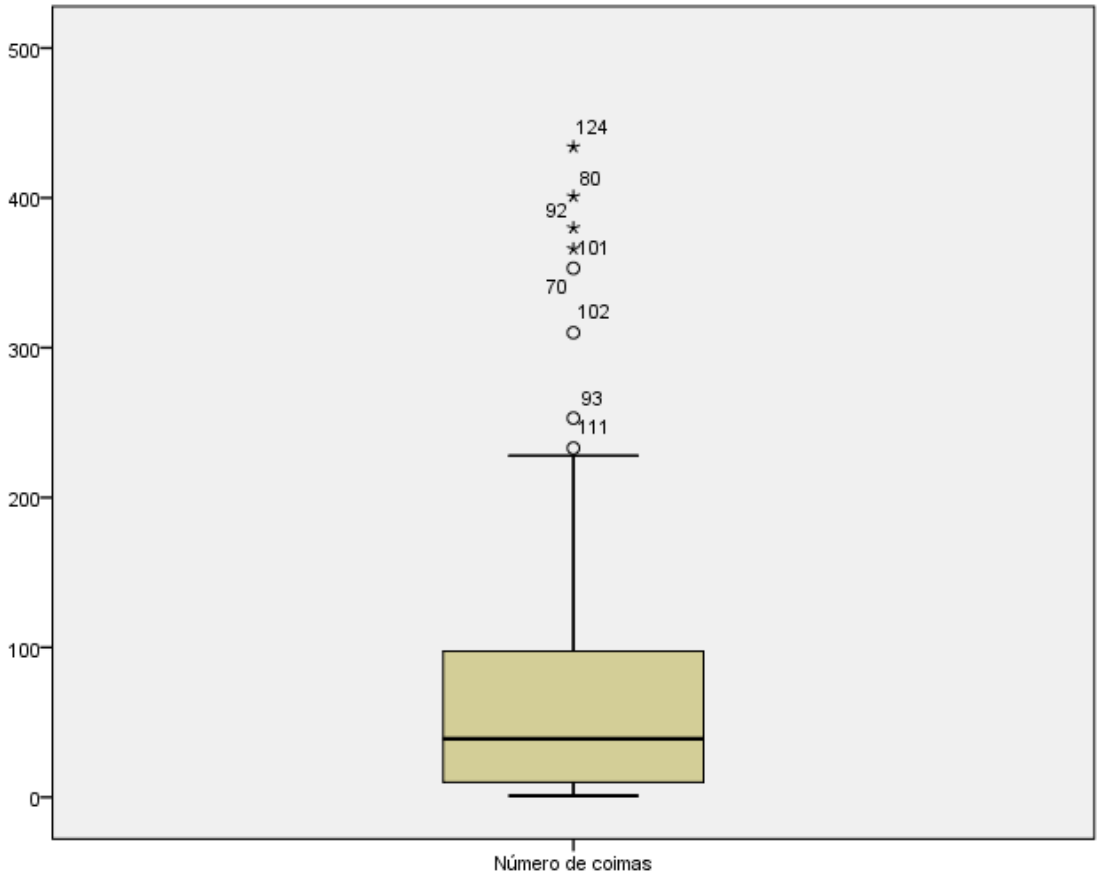


Figura 21 – Caixa de bigodes da variável Ncoimas.

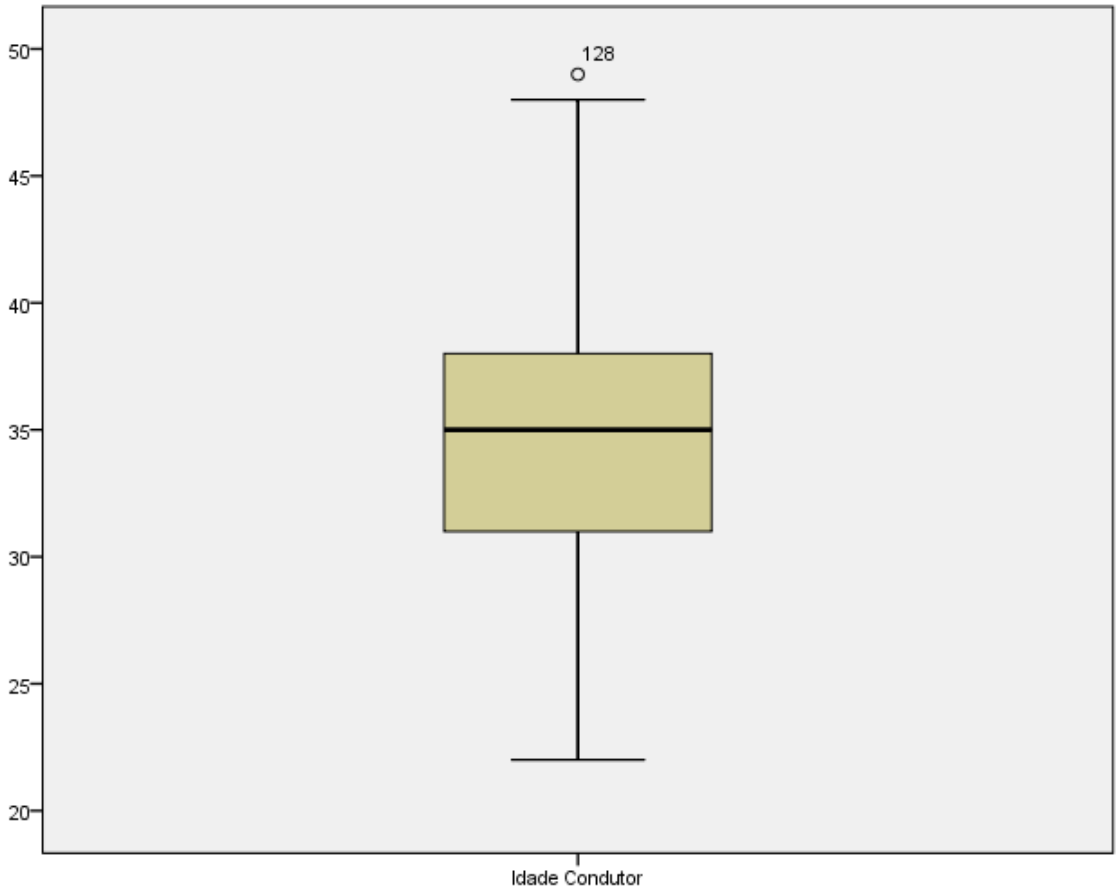


Figura 22 – Caixa de bigodes da variável Idcondutor.

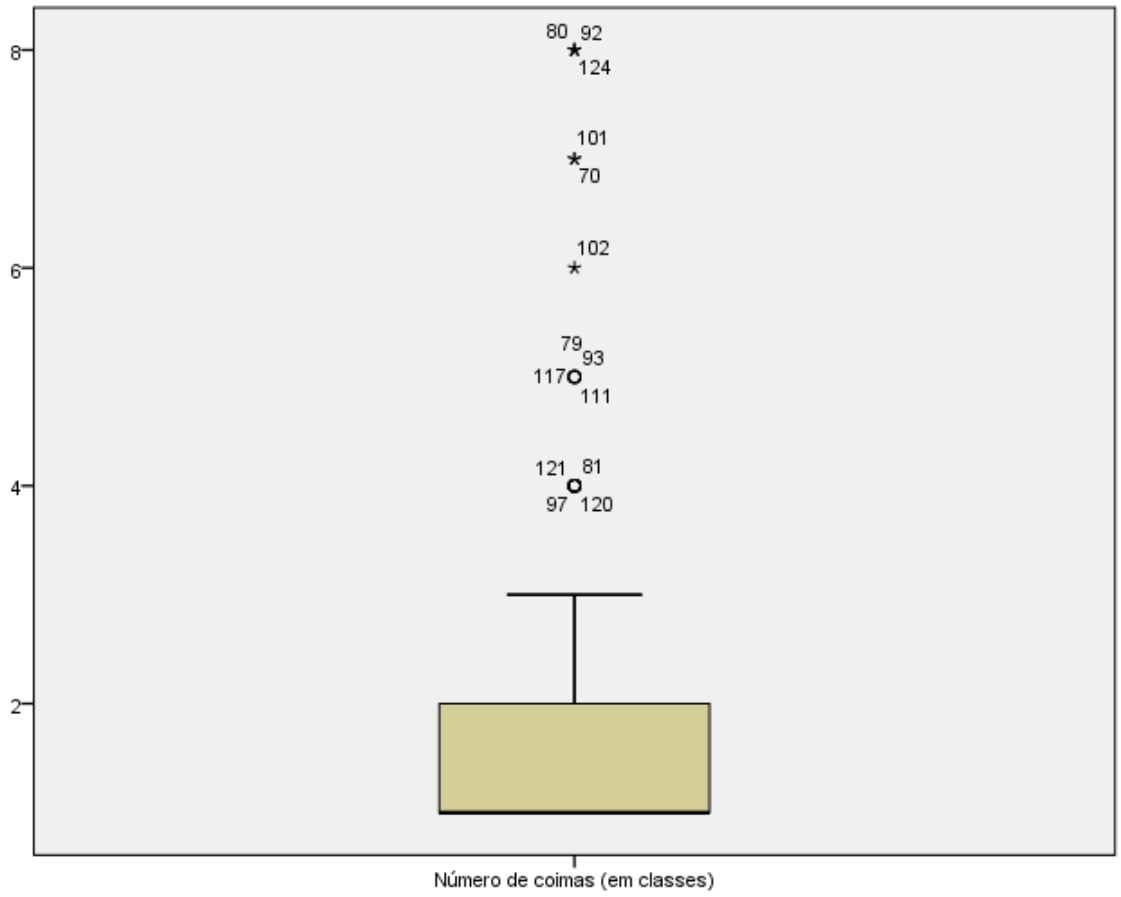


Figura 23 – Caixa de bigodes da variável Ncoimas (em classes).

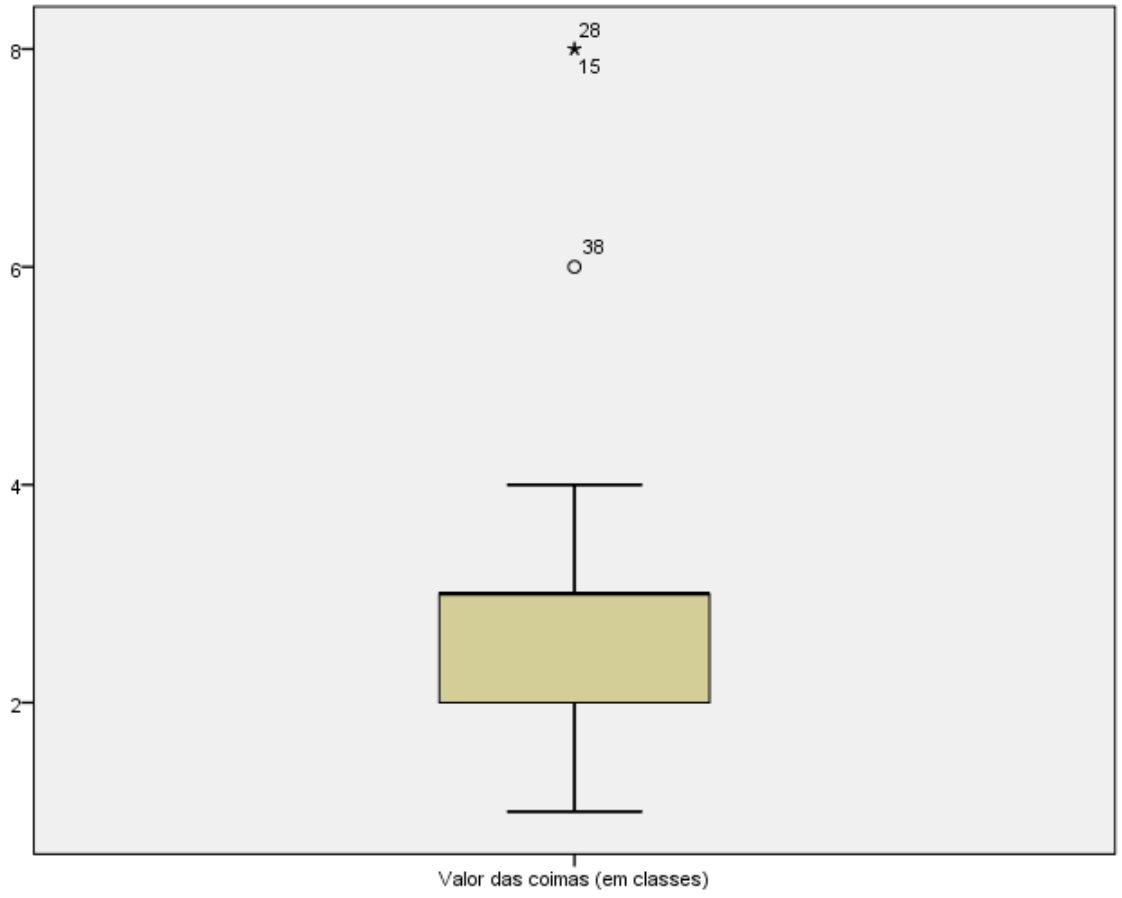


Figura 24 – Caixa de bigodes da variável Coimas (em classes).

✓ Como pudemos observar nas tabelas de extremos e nas caixas de bigodes anteriores (figs. 11-24) pudemos finalmente concluir que:

- As variáveis "Idade do polícia", "Idade do veículo", "Nº de acidentes nos ultimos cinco anos" e "Idade do condutor (em classes)" não têm outliers de acordo com as figuras 11-18.
    - Se observarmos a tabela de frequências da variável "Idade do condutor (em classes)" podemos constatar que grande parte dos condutores (aprox. 77 %) segurados têm idades entre os 29 e 41 anos (fig. 10).
- As variáveis "Valor das coimas", "Nº de coimas", "Idade do condutor", "Nº de coimas (em classes)" e "Valor das coimas (em classes)" têm outliers de acordo com as figuras 19-24.
- Para identificar os outliers nas caixas de bigodes basta olhar para os asteriscos que representam os valores atípicos severos (*) e moderados (°). Relativamente à variável "Valor das coimas" verificámos três outliers todos eles valores atípicos severos sendo eles o 15º, 28º e o 38º caso que correspondem às multas de 763€, 850€ e 636€ respectivamente (figs. 19 e 20).
- Relativamente à variável "Nº de coimas" obtivemos 8 outliers, de acordo com a tabela de extremos e a caixa de bigodes (figs. 19 e 21), dos quais metade são valores atípicos moderados (70º, 102º, 93º e 111º casos) e a outra metade valores atípicos severos (124º, 80º, 92º e 101º casos).
- Relativamente à variável "Idade do condutor" apenas tem um outlier, neste caso um valor atípico moderado, o 128º caso, ou seja, o ultimo caso registado em que a idade do condutor é 49 anos (figs. 19 e 22).
- No caso da variável "Nº de coimas (em classes)" há 14 outliers dos quais 6 são valores atípicos severos (70º, 80º, 92º, 101º, 102º e 124º casos). Os 80º, 92º e 124º casos são os casos em que o número de coimas foi superior a 380, ou seja, corresponde à 8ª classe criada pelo spss como se pode observar na caixa de bigodes. Os 70º, 101º casos são os casos em que o número de coimas esteve entre 326 e 379, ou seja, corresponde à 7ª classe. O 102º caso é o caso em que o número de coimas está entre 271 e 325 que corresponde à 6ª classe. Além destes severos ainda faltam os valores atípicos moderados (79º, 81º, 93º, 97º, 111º, 117º, 120º e 121º casos). Os 79º, 93º, 111º e 117º casos são os casos em que o número de coimas está entre 217 e 270 que corresponde à 5ª classe. Os 81º, 97º, 120º e 121º casos são os casos em que o número de coimas está entre 163 e 216 que corresponde à 4ª classe. De todos estes valores podemos concluir que a maior

parte dos dados está concentrada nas primeiras 3 classes, ou seja, a maior parte das pessoas seguradas já teve até 162 coimas.

- Em relação à variável "Valor das coimas (em classes)" podemos afirmar que esta tem três outliers dos quais dois são valores atípicos severos – o 15º e o 28º caso que são ambos os casos em que as coimas foram entre 744€ e 850€ – e o outro um valor atípico moderado – 38º caso em que a coima teve um valor entre 531€ e 637€ (figs. 19 e 24). Se tivermos em conta os dados obtidos anteriormente para a variável "Valor das coimas" (figs. 19 e 24) podemos efectivamente observar que dois dos outliers se encontram entre 744€ e 850€ e um desses três valores se encontra entre 531€ e 637€. A única diferença é que o outlier que se encontra no último intervalo referido para a variável agrupada em classes é um valor atípico moderado e para a variável antes de ser agrupada em classes é um valor atípico severo. Concluindo, observa-se que a maior parte dos dados se encontra concentrados nas primeiras 4 classes, ou seja, a maior parte dos segurados pagou coimas até ao valor de 424€.

✓ Tendo como base os valores já obtidos nas figuras 1 e 2 podemos afirmar que:

- As variáveis "Idade do polícia", "Idade do veículo", "Idade do condutor" e "Idade do condutor (em classes)" apresentam uma simetria da distribuição de acordo com o coeficiente de assimetria amostral ($\boldsymbol{g_{spss}}$) que consiste na divisão do valor de Skewness pelo erro padrão de Skewness.

$$|\boldsymbol{g_{spss}}| = \left|\frac{\boldsymbol{Skewness}}{\boldsymbol{Std.\ Error\ of\ Skewness}}\right| < \mathbf{1.96}$$

- As variáveis "Valor das coimas", "Nº de coimas", "Nº de acidentes nos últimos 5 anos", "Nº de coimas (em classes)" e "Valor das coimas (em classes)" apresentam uma distribuição assimétrica positiva de acordo com o coeficiente de assimetria amostral ($\boldsymbol{g_{spss}}$).

$$\boldsymbol{g_{spss}} = \frac{\boldsymbol{Skewness}}{\boldsymbol{Std.\ Error\ of\ Skewness}} > \mathbf{1.96}$$

- As variáveis "Valor das coimas", "Nº de coimas", "Valor das coimas (em classes)" e "Nº de coimas (em classes)" podemos observar que a distribuição é leptocúrtica de acordo com o coeficiente de kurtosis amostral ($\boldsymbol{k_{spss}}$).

$$\boldsymbol{k_{spss}} = \frac{\boldsymbol{Kurtosis}}{\boldsymbol{Std.\ Error\ of\ Kurtosis}} > \mathbf{1.96}$$

- No caso das variáveis "Idade do polícia" e "Idade do veículo" pode-se afirrmar que estas apresentam uma distribuição platicúrtica de acordo com o coeficiente de kurtosis amostral ($\boldsymbol{k_{spss}}$).

$$\boldsymbol{k_{spss} = \frac{Kurtosis}{Std.\ Error\ of\ Kurtosis} < -1.96}$$

- As restantes variáveis, "Idade do condutor", "Nº de acidentes nos últimos 5 anos" e "Idade do condutor (em classes)", apresentam uma distribuição mesocúrtica de acordo com o coeficiente de kurtosis amostral ($\boldsymbol{k_{spss}}$).

$$\boldsymbol{|k_{spss}| = \left|\frac{Kurtosis}{Std.\ Error\ of\ Kurtosis}\right| < 1.96}$$

Para uma melhor análise dos dados iremos de seguida interpretar os graficos circulares (variáveis qualitativas) e os histogramas (variáveis quantitativas contínuas).

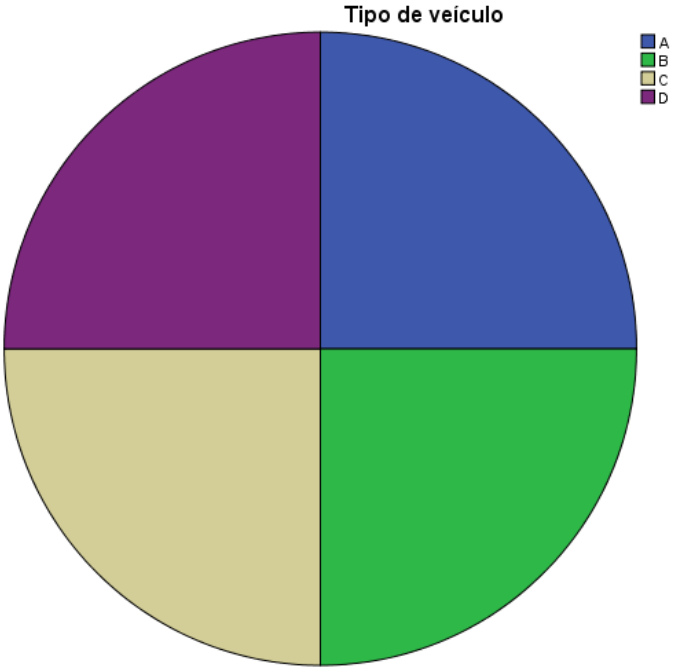


Figura 25 – Gráfico circular da variável "Tipo de Veículo".

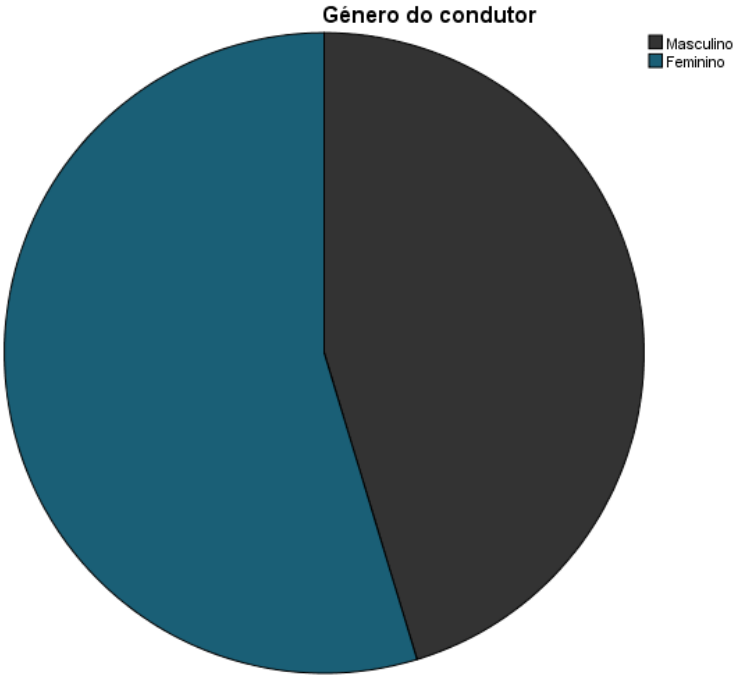


Figura 26 – Gráfico circular da variável "Género do condutor".

Como já foi referido anteriormente, para podermos interpretar os histogramas das variáveis cujas classes possuem amplitudes diferentes, que é o caso das variáveis "Idade do polícia" e "Idade do veículo", teremos que converter as frequências para obter a proporcionalidade entre a altura das barras e a sua base desejada ($\boldsymbol{n_i}^* = \frac{\boldsymbol{n_i}}{\boldsymbol{a_i}}$).

IdPolícia	IdVeículo
$\boldsymbol{n_1}^* = \frac{16}{4} = 4$	$\boldsymbol{n_1}^* = \frac{32}{4} = 8$
$\boldsymbol{n_2}^* = \frac{16}{4} = 4$	$\boldsymbol{n_2}^* = \frac{32}{4} = 8$
$\boldsymbol{n_3}^* = \frac{16}{5} = 3.2$	$\boldsymbol{n_3}^* = \frac{32}{2} = 16$
$\boldsymbol{n_4}^* = \frac{16}{5} = 3.2$	$\boldsymbol{n_4}^* = \frac{32}{6} = 5.3$
$\boldsymbol{n_5}^* = \frac{16}{5} = 3.2$	
$\boldsymbol{n_6}^* = \frac{16}{10} = 1.6$	
$\boldsymbol{n_7}^* = \frac{16}{10} = 1.6$	
$\boldsymbol{n_8}^* = \frac{16}{11} = 1.45$	

Nota: Para a última classe da variável *IdPolícia* foi assumido o intervalo [60 - 70] tendo em conta que a idade da reforma é aos 65 anos, extendemos o limite até aos 70 anos visto que também há polícias acima da idade da reforma que ainda exercem funções. Para a última classe da variável *IdVeículo* assumimos o intervalo [10 - 15] visto que o spss assumiu na ultima classe o intervalo de 10 a infinito tivemos que impôr um valor.

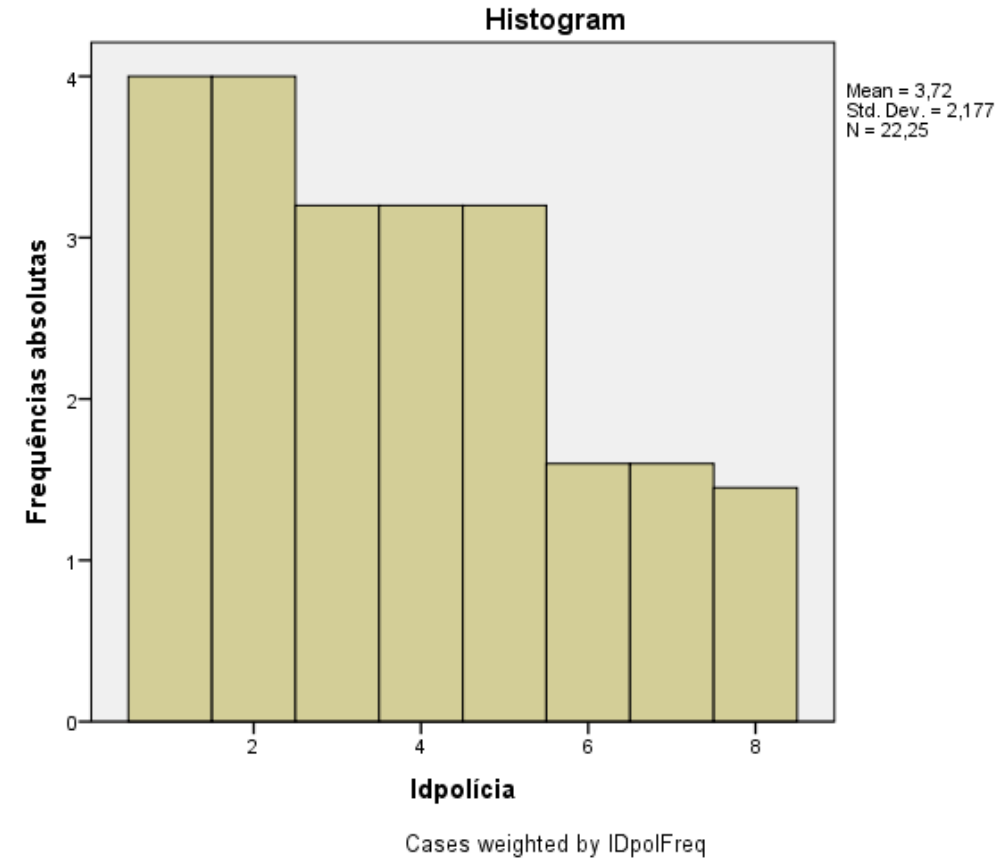


Figura 27 – Histograma da variável Idpolícia.

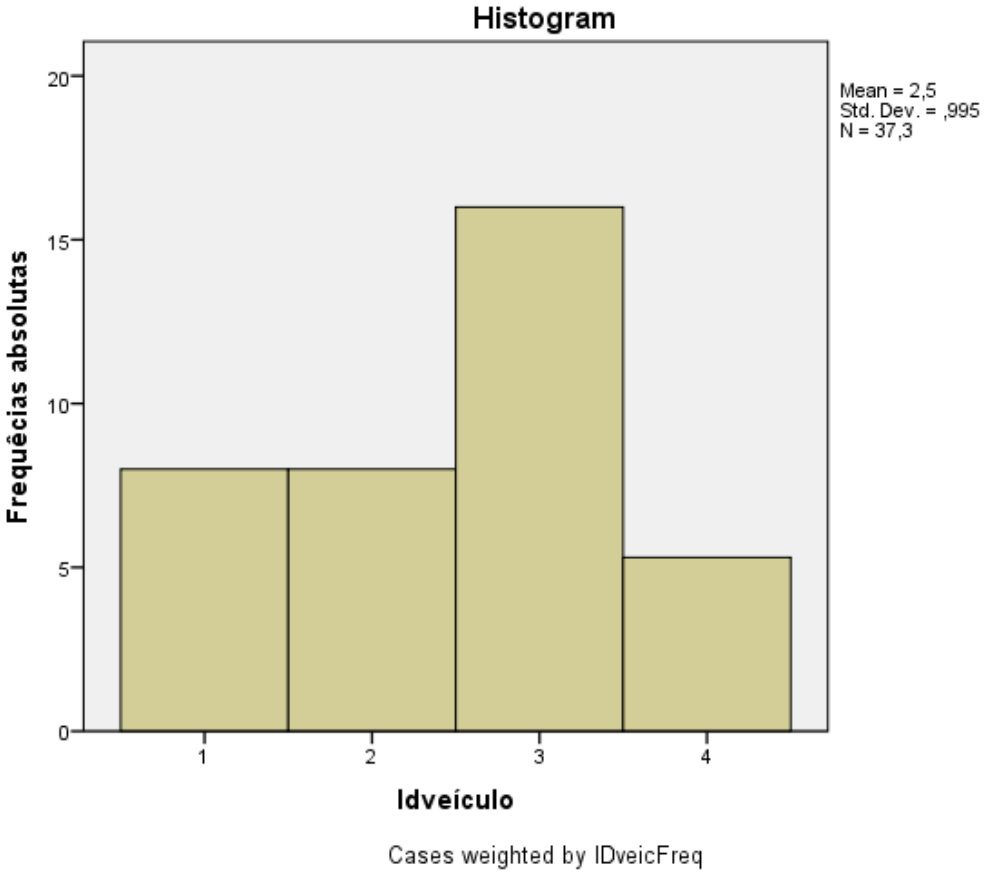


Figura 28 – Histograma da variável Idveículo.

Como podemos observar a distribuição dos dados segundo os histogramas não é tão "linear" assim como iria parecer se fizesse um histograma sem converter as frequencias. Iriamos ter um histograma com as barras todas à mesma altura para classes com diferentes amplitudes, o que poderia originar erros na interpretação dos dados levando-nos a pensar de imediato que se trata de uma distribuição dos dados simétrica. Neste caso verifica-se que é de facto uma distribuição simetrica tendo em conta o valor do coeficiente de assimetria amostral calculado anteriormente ($\boldsymbol{g_{spss}}$).

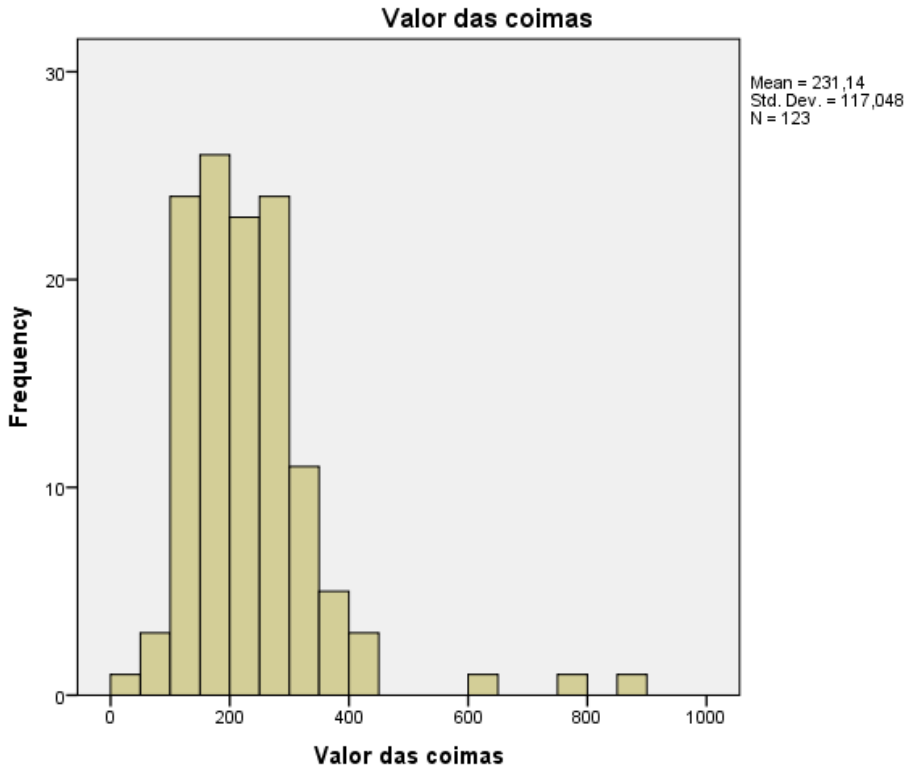


Figura 29 – Histograma da variável Coimas.

✓ Segundo o histograma anterior (fig. 29) a maior parte das coimas ronda o valor de 200€ verificando-se uma distribuição assimétrica positiva ou enviesada à esquerda.

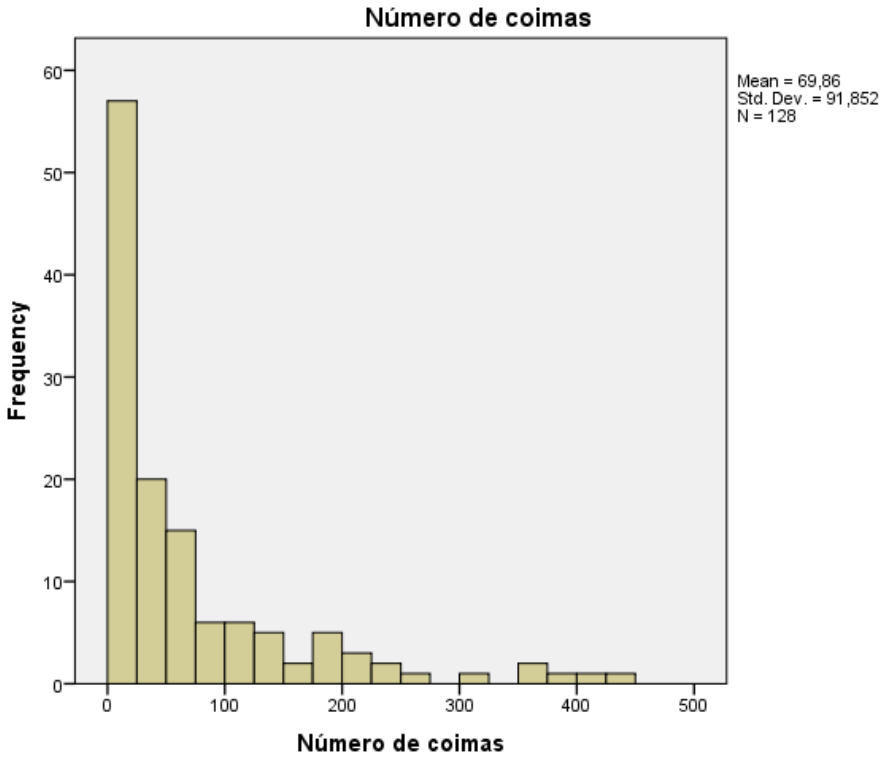


Figura 30 – Histograma da variável NºCoimas.

✓ No caso deste histograma (fig. 30) podemos observar que a maior parte das pessoas seguradas nesta amostra cometeram poucas infracções, ou seja, pagaram poucas coimas verificando-se também uma distribuição assimétrica positiva ou enviesada à esquerda de acordo com os dados de Skewness.

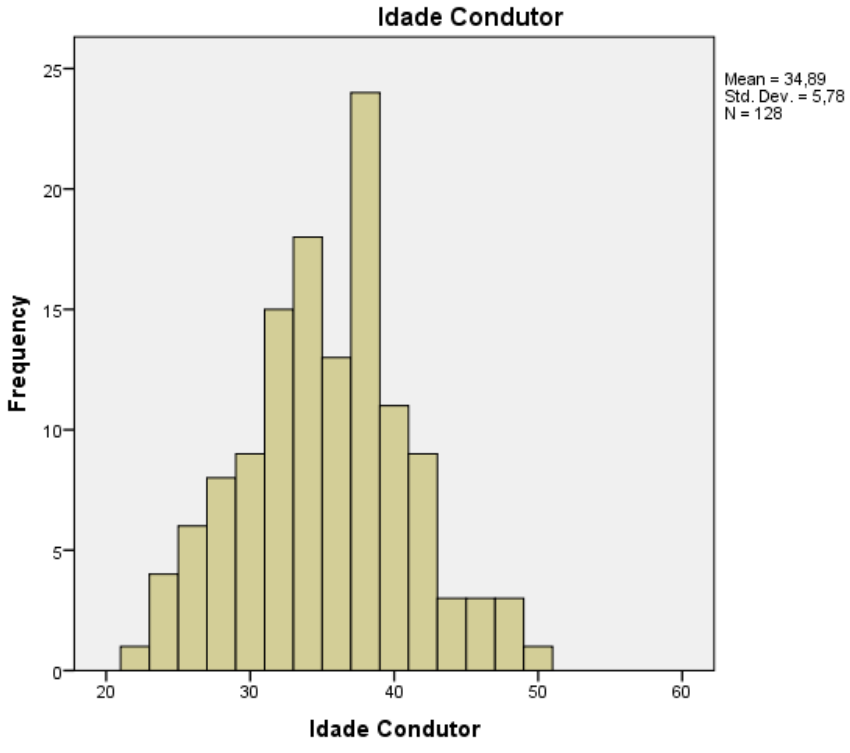


Figura 31 – Histograma da variável Idcondutor.

✓ Já neste caso (fig. 31) podemos facilmente constatar que se trata de uma distribuição simétrica, pois neste histograma esta aproxima-se bastante da distribuição Normal. Segundo este histograma a maior parte dos condutores segurados têm idades entre os 30 e 40 anos aproximadamente complementando as conclusões anteriores acerca da tabela de frequências.

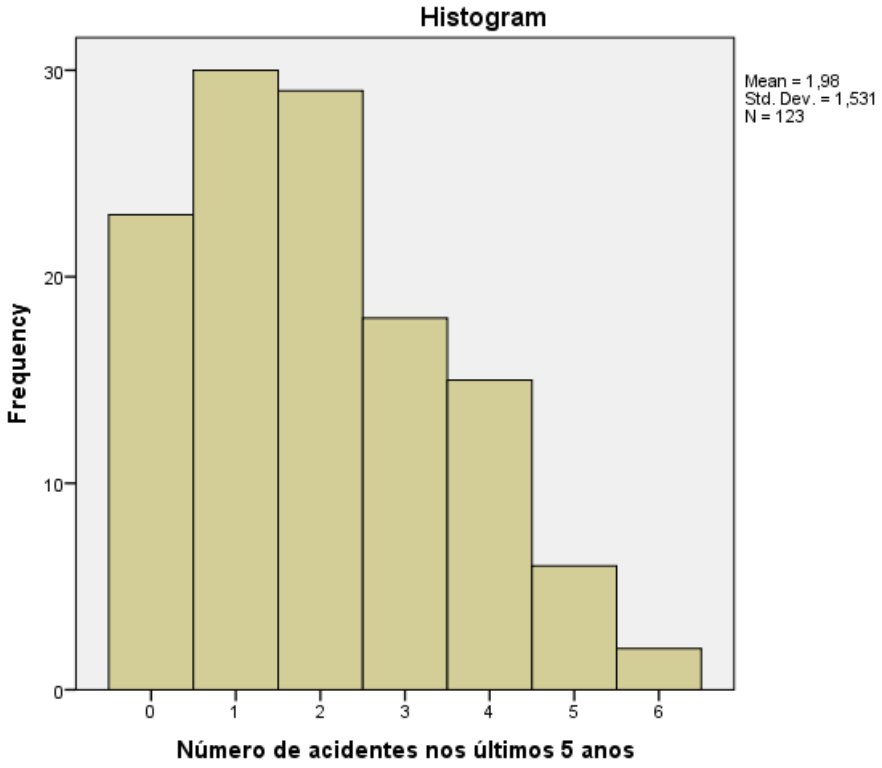


Figura 32 – Histograma da variável Acidentes.

✓ De acordo com este histograma (fig. 32) podemos observar que se trata de uma distribuição assimétrica positiva ou enviesada à esquerda estando de acordo com os dados obtidos através do coeficiente de assimetria amostral apesar de não ser muito adequado fazer um histograma para uma variável contínua como esta. Estes dados demonstram que a maior parte dos condutores segurados tiveram poucos ou nenhuns acidentes nos últimos anos sendo que a maior parte teve desde zero a dois acidentes.

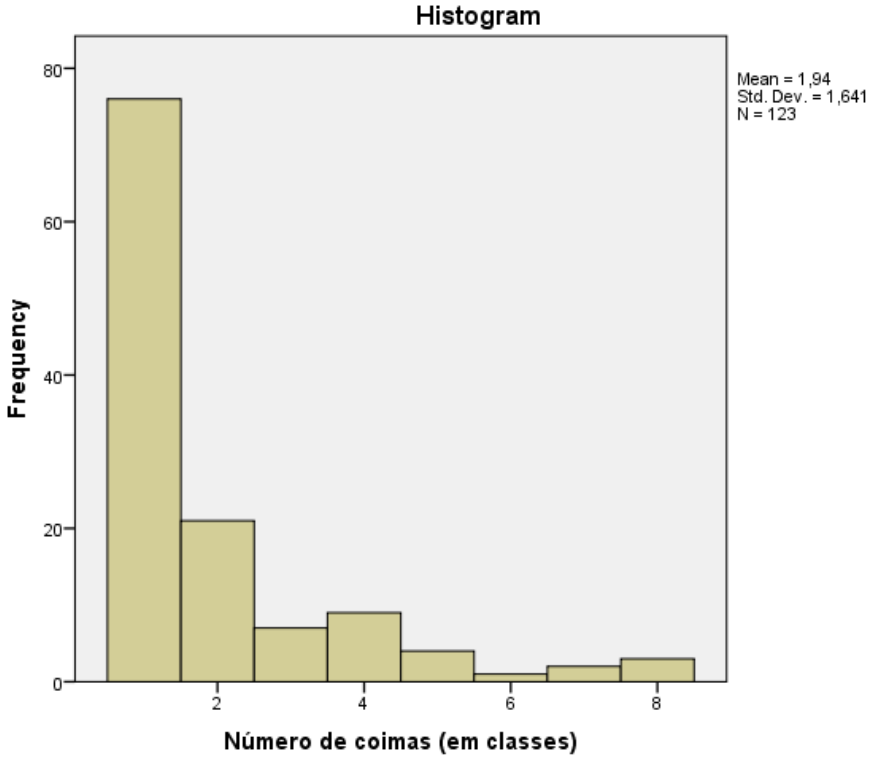


Figura 33 – Histograma da variável Ncoimas(em classes).

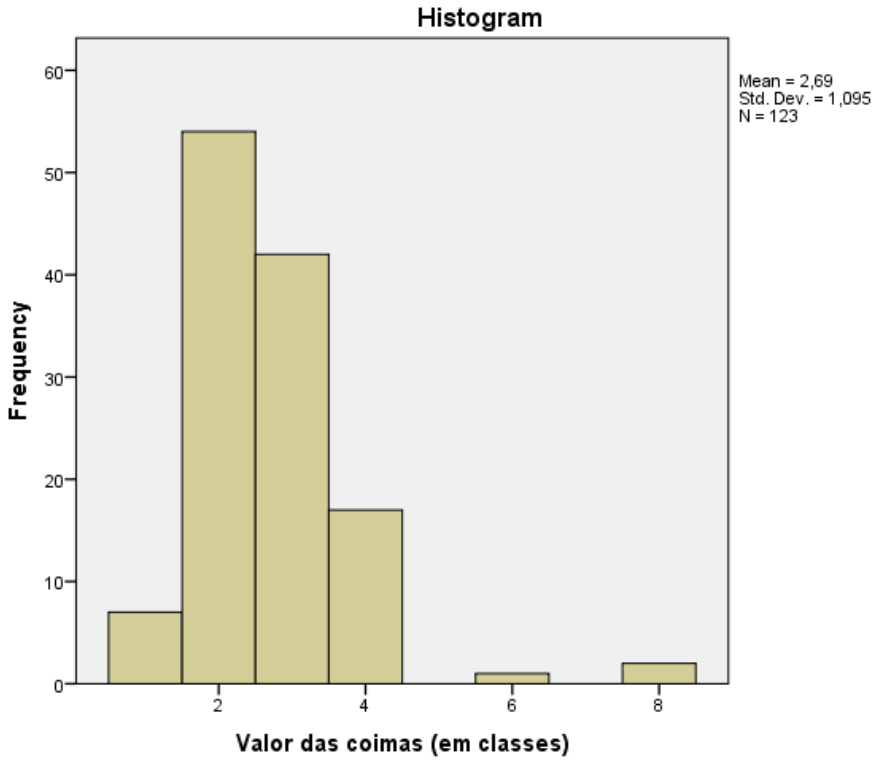


Figura 34 – Histograma da variável Coimas(em classes).

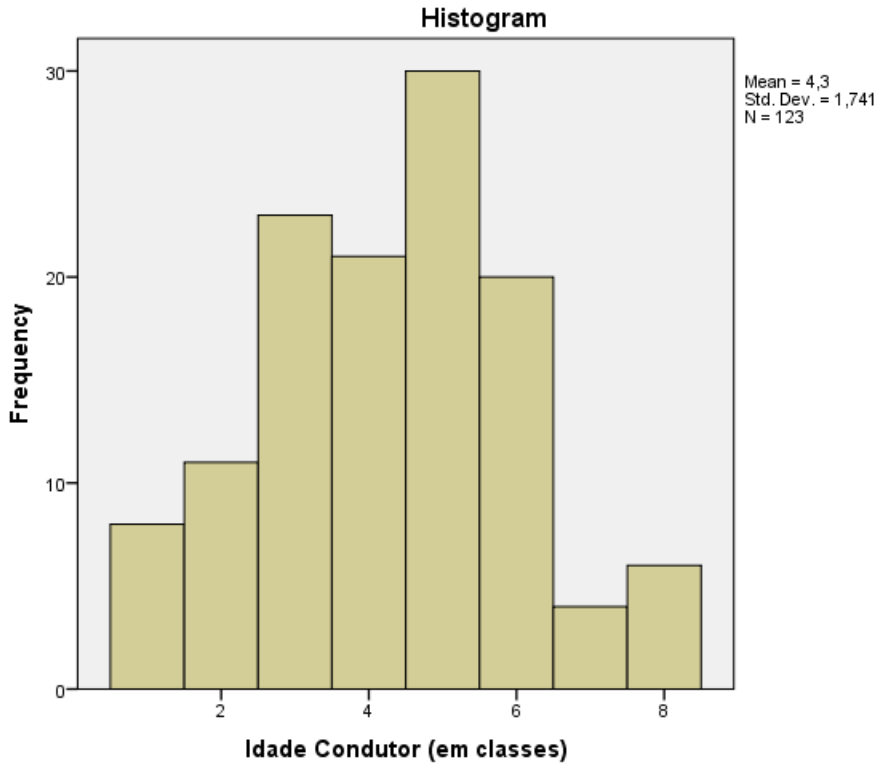


Figura 35 – Histograma da variável Idcondutor(em classes).

✓ As conclusões que se retiram destes histogramas (figs. 33, 34 e 35) são practicamente as mesmas que retirámos dos anteriores não agrupados em classes (figs. 29, 30 e 31), a maior parte dos condutores segurados pagaram entre zero e cinquenta e quatro coimas; a maior parte das coimas assume valores entre 106€ e 318€ aproximadamente; e a maior parte dos condutores têm idades entre 29 e 41 anos aprox. que corresponde às classes 3, 4, 5 e 6. Em termos da distribuição dos dados as conclusões continuam a ser exactamente as mesmas.

One-Sample Statistics

Tipo de veículo		N	Mean	Std. Deviation	Std. Error Mean
A	Valor das coimas	32	190,38	64,131	11,337
B	Valor das coimas	32	201,94	78,281	13,838
C	Valor das coimas	31	221,32	67,826	12,182
D	Valor das coimas	28	321,96	182,729	34,533

One-Sample Test

		Test Value = 0					
						95% Confidence Interval of the Difference	
Tipo de veículo		t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
A	Valor das coimas	16,792	31	,000	190,375	167,25	213,50
B	Valor das coimas	14,593	31	,000	201,938	173,71	230,16
C	Valor das coimas	18,168	30	,000	221,323	196,44	246,20
D	Valor das coimas	9,324	27	,000	321,964	251,11	392,82

Figura 36 – Tabela das medidas de tendencia central e da estimativa pontual e intervalar.

✓ Tendo em conta os valores obtidos para o I.C. do valor das coimas segundo o tipo de veículo podemos afirmar que:

- Com 95 % de confiança, os condutores de veículos do tipo A pagaram, em média, coimas com um valor entre 167.25€ e 213.50€ e um erro padrão da média associado ao I.C. de 11.34.
- Com 95 % de confiança, os condutores de veículos do tipo B pagaram, em média, coimas com um valor entre 173.71€ e 230.16€ e um erro padrão da média associado ao I.C. de 13.84.
- Com 95 % de confiança, os condutores de veículos do tipo C pagaram, em média, coimas com um valor entre 196.44€ e 246.20€ e um erro padrão da média associado ao I.C. de 12.18.
- Com 95 % de confiança, os condutores de veículos de tipo D pagaram, em média, coimas com um valor entre 251.11€ e 392.82€ e um erro padrão da média associado ao I.C. de 34.
- Se tivermos em conta os erros-padrão da média pode-se concluir que o intervalo de confiança mais preciso é o intervalo para o valor das coimas de condutores de veículos de tipo A, pois é aquele em que o erro-padrão da média é menor e o intervalo de confiança menos preciso é aquele para condutores de veículos do tipo D visto ser aquele em que o erro-padrão é maior. Podemos chegar à mesma conclusão se subtrairmos o lower value ao upper value do intervalo de confiança.

Considerando a variável "Valor das coimas" separada por "Género do condutor":

Group Statistics

	Género do condutor	N	Mean	Std. Deviation	Std. Error Mean
Valor das coimas	Masculino	58	199,34	72,993	9,584
	Feminino	65	259,51	140,134	17,381

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
Valor das coimas	Equal variances assumed	7,053	,009
	Equal variances not assumed		

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Valor das coimas	Equal variances assumed	-2,933	121	,004	-60,163	20,513	-100,773	-19,553
	Equal variances not assumed	-3,031	98,601	,003	-60,163	19,849	-99,549	-20,777

Figura 37 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar.

✓ Para averiguar se, com 95 % de confiança, existe evidência de diferença entre as médias do valor das coimas por género, pode recorrer-se ao intervalo de confiança ou a um teste de hipóteses (neste caso trata-se de um teste bilateral), para a comparação das duas médias de amostras independentes. O primeiro passo a efectuar na análise dos dados obtidos anteriormente (fig. 37) é verificar se a igualdade das variâncias pode ou não ser rejeitada:

- Se observarmos o teste de Levene (p-value = 0.009) podemos assumir que as variâncias são iguais. Sendo assim teremos apenas em conta os dados da primeira linha (Equal variances assumed).

✓ Analisando o p-value = 0.004, podemos afirmar que não existe evidência da igualdade entre as médias dos valores das coimas dos dois sexos, ou seja, rejeita-se Ho visto que o p-value é menor que o nível de significância ($\alpha = 0.05$). Podemos também alcançar esta mesma conclusão através do I.C. a 95 % que não contém o valor 0 excluindo logo de imediato a hipótese da igualdade entre as médias estando a margem da possível diferença entre o valor das coimas da população masculina e feminina situada no intervalo $]-100.773;\ -19.553[$.

Group Statistics

	Género do condutor	N	Mean	Std. Deviation	Std. Error Mean
Número de acidentes nos últimos 5 anos	Masculino	58	1,81	1,317	,173
	Feminino	70	2,10	1,661	,198

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
Número de acidentes nos últimos 5 anos	Equal variances assumed	2,306	,131
	Equal variances not assumed		

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	99% Confidence Interval of the Difference	
							Lower	Upper
Número de acidentes nos últimos 5 anos	Equal variances assumed	-1,077	126	,284	-,290	,269	-,993	,414
	Equal variances not assumed	-1,100	125,781	,273	-,290	,263	-,978	,399

Figura 38 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar.

✓ Vamos então desta vez averiguar se, com 99 % de confiança, a diferença entre as médias do nº de acidentes nos últimos cinco anos para homens e para mulheres é ou não maior que zero, ou seja, averiguar se a média do nº de acidentes nos últimos cinco anos é maior em homens do que em mulheres recorrendo também ao I.C ou a um teste de hipóteses (neste caso trata-se de um teste unilateral esquerdo) ou a ambos (fig. 38):

- Observando o teste de Levene (p-value = 0.131) concluir-se-á que as variâncias são iguais. Utilizaremos então apenas os dados da primeira linha (Equal variances assumed).

✓ Pela análise do p-value = 0.284, podemos comprovar que não existe evidência que suporte a hipótese da diferença entre as médias do nº de acidentes nos últimos cinco anos para homens e para mulheres ser menor que zero, ou seja, podemos afirmar que o número de acidentes nos ultimos cinco anos é, em média, maior para os homens do que para as mulheres. No fundo não se rejeita Ho visto que o p-value é maior que o nível de significância ($\alpha = 0.01$). Esta conclusão é também baseada no I.C. a 99 %, que contém o valor 0, onde a margem da possivel diferença entre o número de acidentes nos últimos cinco anos para homens e para mulheres se situa no intervalo $]-0.993;\ 0.414[$.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Número de coimas	58	76,53	100,146	13,150

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Número de coimas	5,820	57	,000	76,534	50,20	102,87

Figura 39 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar para o sexo masculino.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Número de coimas	70	64,33	84,705	10,124

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Número de coimas	6,354	69	,000	64,329	44,13	84,53

Figura 40 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar para o sexo feminino.

✓ De acordo com os dados das duas figuras anteriores anteriores (figs. 39 e 40) podemos afirmar que, com 95 % de confiança, o nº de coimas médio da população masculina se situa entre 50 e 103 coimas, aproximadamente, e o nº de coimas médio da população feminina se encontra entre 44 e 85 coimas, aproximadamente. Tendo em conta estes valores obtidos podemos concluir que as mulheres cometem menos infracções do que os homens visto que estas têm, em média, menos coimas que os homens.

Para avaliar a existência de relação entre a Idade do condutor e o Nº de acidentes nos ultimos 5 anos necessitaremos dos coeficientes de correlação visto que se tratam de duas variáveis quantitativas. Para podermos decidir se será mais correcto utilizar o coeficiente de correlação de Spearman ou de Pearson, deverá analisar-se a "normalidade" das distribuições fazendo um teste de Kolmogorov-Smirnov.

One-Sample Kolmogorov-Smirnov Test

		Número de acidentes nos últimos 5 anos	Idade Condutor (em classes)
N		128	128
Normal Parameters[a,b]	Mean	1,97	4,30
	Std. Deviation	1,516	1,713
Most Extreme Differences	Absolute	,168	,144
	Positive	,168	,111
	Negative	-,097	-,144
Kolmogorov-Smirnov Z		1,904	1,625
Asymp. Sig. (2-tailed)		,001	,010

a. Test distribution is Normal.

b. Calculated from data.

Figura 41 – Tabela de teste de Normalidade - Kolmogorov-Smirnov.

Como o tamanho da amostra é superior a 50 optamos, obviamente, pelo teste de Kolmogorov-Smirnov. A normalidade é rejeitada para ambas as variáveis analisadas (p-values ≤ 0.01), logo deverá utilizar-se o teste de correlação de Spearman.

Correlations

			Número de acidentes nos últimos 5 anos	Idade Condutor (em classes)
Spearman's rho	Número de acidentes nos últimos 5 anos	Correlation Coefficient	1,000	-,029
		Sig. (2-tailed)	.	,748
		N	128	128
	Idade Condutor (em classes)	Correlation Coefficient	-,029	1,000
		Sig. (2-tailed)	,748	.
		N	128	128

Figura 42 – Tabela de teste de correlação de Spearman.

Como o p-value ≤ 0.748, ou seja, é muito maior que o nível de significância o que significa que as correlações são estatisticamente insignificantes. Se observarmos o coeficiente de correlação concluimos que a relação entre estas duas variáveis é negativa e muito fraca practicamente inexistente.

One-Sample Kolmogorov-Smirnov Test

		Idade do policia	Valor das coimas (em classes)
N		128	123
Normal Parameters[a,b]	Mean	4,50	2,69
	Std. Deviation	2,300	1,095
Most Extreme Differences	Absolute	,118	,232
	Positive	,118	,232
	Negative	-,118	-,207
Kolmogorov-Smirnov Z		1,333	2,573
Asymp. Sig. (2-tailed)		,057	,000

a. Test distribution is Normal.

b. Calculated from data.

Figura 43 – Tabela de teste de Normalidade - Kolmogorov-Smirnov.

De acordo com estes dados teremos que recorrer ao teste de correlação de Pearson, pois a Normalidade não é rejeitada.

Correlations

		Idade do policia	Valor das coimas (em classes)
Idade do policia	Pearson Correlation	1	-,195*
	Sig. (2-tailed)		,030
	N	128	123
Valor das coimas (em classes)	Pearson Correlation	-,195*	1
	Sig. (2-tailed)	,030	
	N	123	123

*. Correlation is significant at the 0.05 level (2-tailed).

Figura 44 – Tabela de teste de correlação de Pearson.

Tendo em conta a correlação entre as duas variáveis anteriores pode-se concluir que existe uma relação negativa e moderadamente fraca entre a idade do polícia e o valor das coimas.

Questões Colocadas

1. **Analise a tabela de dupla entrada e o gráfico de barras da variável "Tipo de veículo" dividida por "Género" e retire conclusões.**

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tipo de veículo * Género do condutor	128	100,0%	0	0,0%	128	100,0%

Tipo de veículo * Género do condutor Crosstabulation

			Género do condutor		
			Masculino	Feminino	Total
Tipo de veículo	A	Count	32	0	32
		% within Tipo de veículo	100,0%	0,0%	100,0%
	B	Count	26	6	32
		% within Tipo de veículo	81,2%	18,8%	100,0%
	C	Count	0	32	32
		% within Tipo de veículo	0,0%	100,0%	100,0%
	D	Count	0	32	32
		% within Tipo de veículo	0,0%	100,0%	100,0%
Total		Count	58	70	128
		% within Tipo de veículo	45,3%	54,7%	100,0%

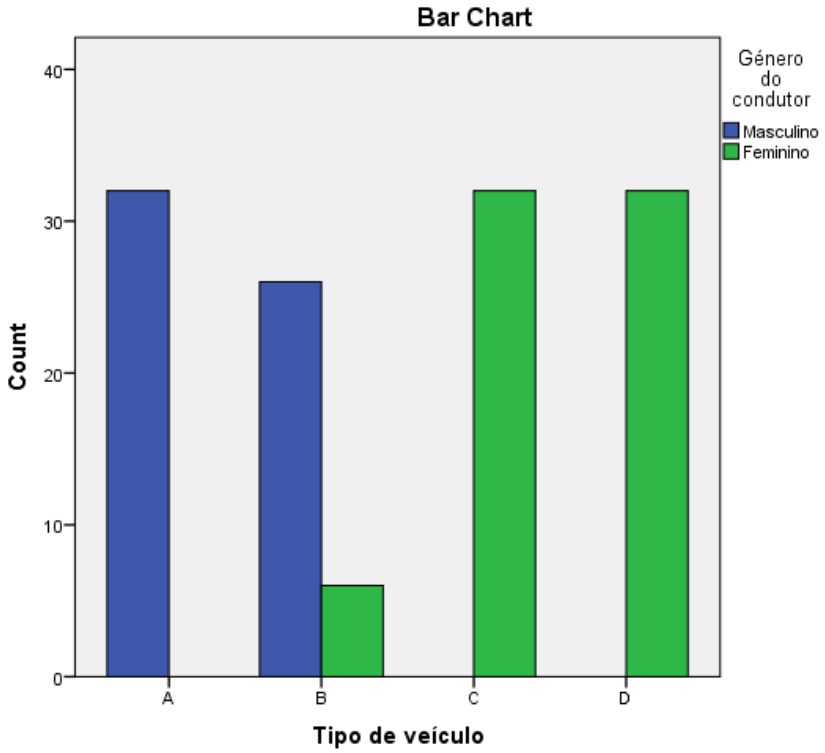


Figura 45 – Tabela de dupla entrada e gráfico de barras da var. TVeíc por género.

- O conjunto de dados apresentado anteriormente (fig. 45) é particularmente interessante, pois se repararmos bem, os veículos que as mulheres mais conduzem nesta amostra são veículos do tipo C e D, os veículos que os homens mais conduzem são do tipo A e B, não há um único homem que conduza um veículo do tipo C e D e não há uma única mulher que conduza um veículo do tipo A.

2. **Construa o intervalo de confiança a 95 % para o número médio de acidentes nos ultimos cinco anos. Retire conclusões.**

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Número de acidentes nos últimos 5 anos	58	1,81	1,317	,173

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Número de acidentes nos últimos 5 anos	10,466	57	,000	1,810	1,46	2,16

Figura 46 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar para o sexo masculino.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Número de acidentes nos últimos 5 anos	70	2,10	1,661	,198

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Número de acidentes nos últimos 5 anos	10,580	69	,000	2,100	1,70	2,50

Figura 47 – Tabela da média, do teste de hipóteses e da estimativa pontual e intervalar para o sexo feminino.

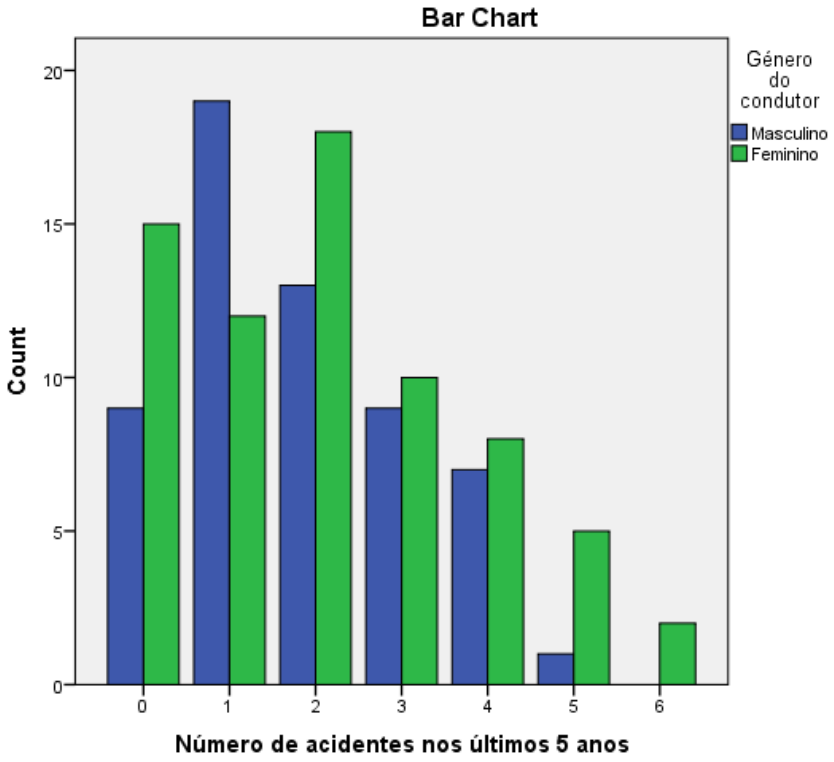


Figura 48 – Gráfico de barras da variável Acidentes por Género.

- Podemos concluir, a partir dos dados anteriores, que com 95 % de confiança o número médio de acidentes nos ultimos cinco anos da população masculina se situa entre um e dois acidentes aproximadamente, e da população feminina encontra-se entre dois e três acidentes aproximadamente, o que significa que nos ultimos cinco anos quem teve mais acidentes, em média, foram as mulheres.

3. **Pode-se afirmar que, em média, são os homens que tiveram mais multas que as mulheres? Considere um nível de confiança de 2 %.**

Group Statistics

	Género do condutor	N	Mean	Std. Deviation	Std. Error Mean
Número de coimas	Masculino	58	76,53	100,146	13,150
	Feminino	70	64,33	84,705	10,124

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
							98% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Número de coimas	Equal variances assumed	,184	,669	,747	126	,456	12,206	16,338	-26,290	50,702
	Equal variances not assumed			,735	112,073	,464	12,206	16,596	-26,961	51,373

Figura 49 – Tabela da média, desvio padrão e erro-padrão, do teste de hipóteses e da estimativa pontual e intervalar.

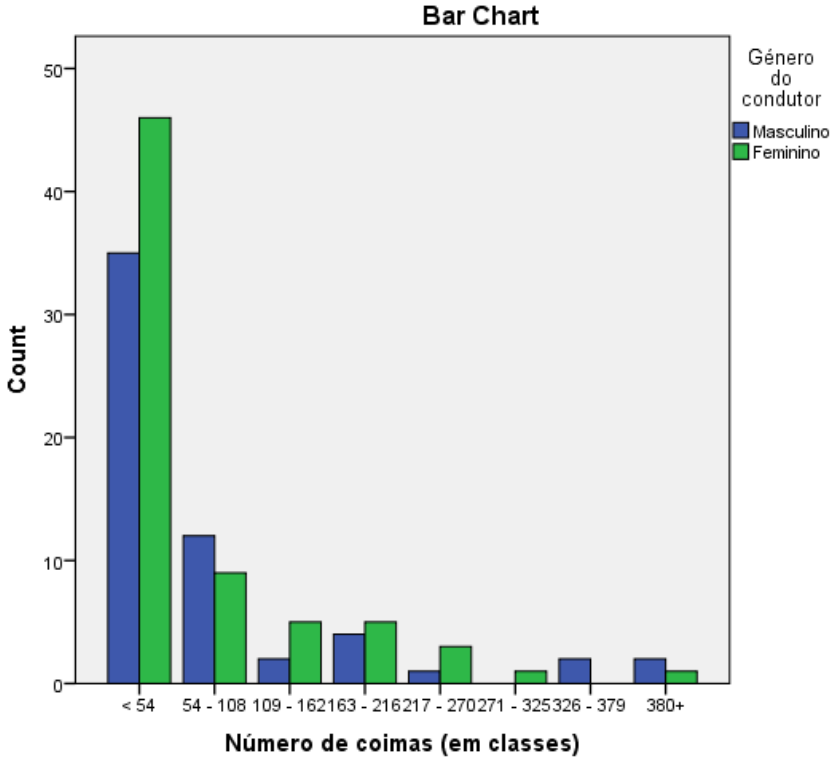


Figura 50 – Gráfico de barras da variável Ncoimas por Género.

✓ Observando o teste de Levene (p-value = 0.669) concluir-se-á que as variâncias são iguais. Utilizaremos então apenas os dados da primeira linha (Equal variances assumed).

✓ Analisando o p-value = 0.456, podemos comprovar que não existe evidência que suporte a hipótese da diferença entre as médias do nº de coimas para homens e para mulheres ser menor que zero, ou seja, podemos afirmar que o número de coimas é, em média, maior para os homens do que para as mulheres. No fundo não se rejeita Ho visto que o p-value é maior que o nível de significância ($\alpha = 0.02$). Esta conclusão é também baseada no I.C. a 98 %, que contém o valor 0, onde a margem da possivel diferença entre o número de acidentes nos últimos cinco anos para homens e para mulheres se situa no intervalo $]-26.290;\ 50.702[$.

4. **De que forma poderá a idade do veículo explicar linearmente o número de coimas?**

Para descrever a relação linear entre estas duas variáveis, iremos utilizar a técnica de regressão linear simples.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,528[a]	,278	,273	78,332

a. Predictors: (Constant), Idade do veículo

b. Dependent Variable: Número de coimas

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	298339,256	1	298339,256	48,621	,000[b]
	Residual	773132,212	126	6135,970		
	Total	1071471,469	127			

a. Dependent Variable: Número de coimas

b. Predictors: (Constant), Idade do veículo

Coefficients[a]

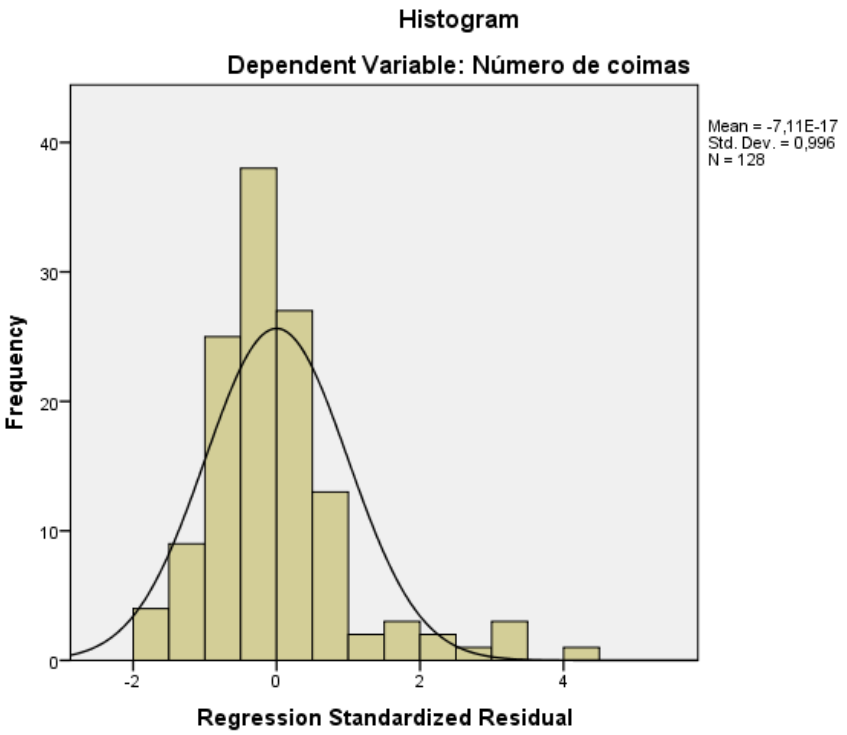
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	177,813	16,959		10,485	,000	144,250	211,375
	Idade do veículo	-43,181	6,193	-,528	-6,973	,000	-55,436	-30,926

a. Dependent Variable: Número de coimas

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5,09	134,63	69,86	48,468	128
Residual	-131,631	342,550	,000	78,023	128
Std. Predicted Value	-1,336	1,336	,000	1,000	128
Std. Residual	-1,680	4,373	,000	,996	128

a. Dependent Variable: Número de coimas

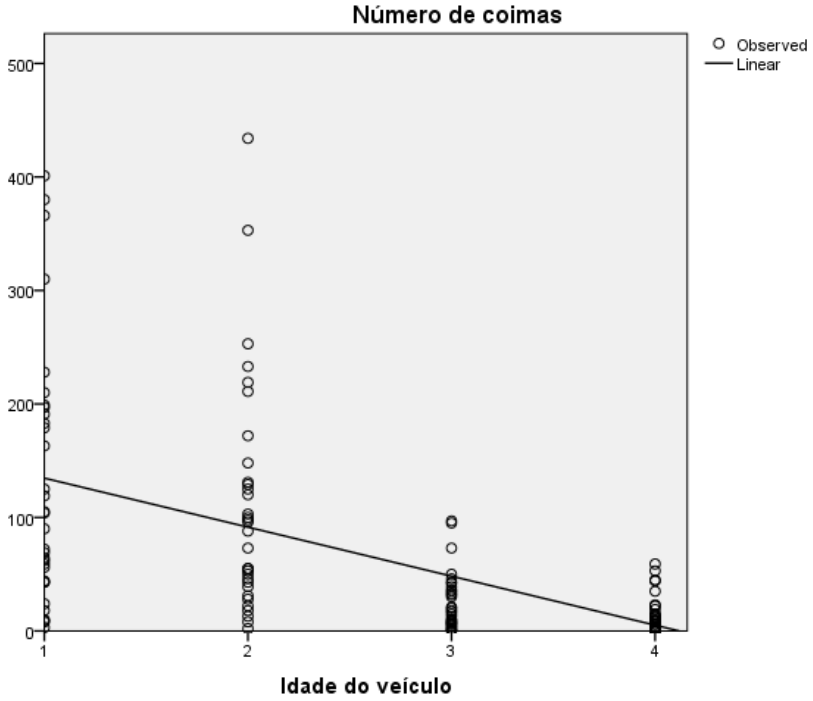


Model Summary and Parameter Estimates

Dependent Variable: Número de coimas

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	,278	48,621	1	126	,000	177,813	-43,181

The independent variable is Idade do veículo.



✓ Pela análise do gráfico de dispersão pode-se constatar que os pontos parecem de certa forma dispor-se em torno de uma recta de declive negativo, verificando-se que à medida que aumenta a idade do veículo o número de coimas diminui e vice versa, existindo assim uma relação linear negativa fraca entre as variáveis. Este modelo, estatisticamente significativo, visto que o p-value é zero na tabela ANOVA, apresenta um poder explicativo

fraco de apenas 27.3 % (R square = 0.273). A linha que se observa no gráfico de dispersão representa o modelo linear ajustado aos dados, cuja equação é:

- $\widehat{Ncoimas} = 177.813 - 43.181 \times Idveic$

✓ Estes coeficientes são significativos (p-values ≈ 0). Os resíduos apresentam média igual a zero e não se desviam muito da distribuição Normal (Histograma dos resíduos).