My Class Notes

Marcelo Bezerra

October 2023

Contents

List of Tables

List of Figures

Chapter 1

Descriptive statistics

1.1 Describing data sets

[?]

Starting yearly salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Table 1.1: Frequency table for starting yearly salaries

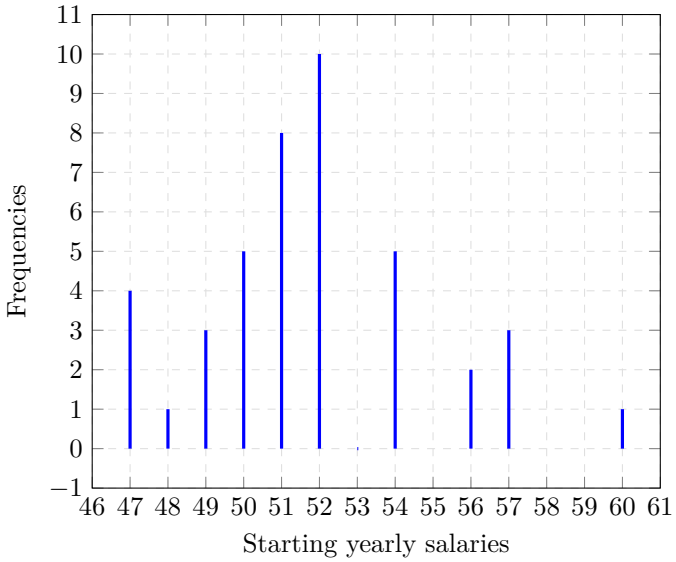

Figure 1.1: Line graph for starting salary data

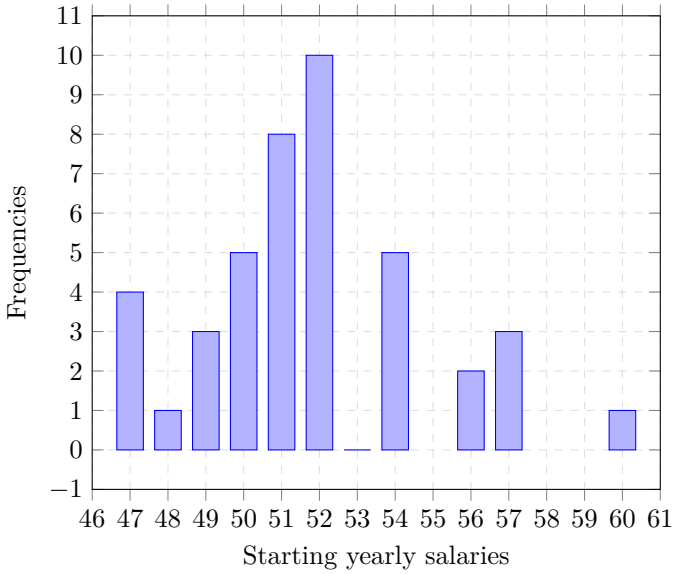

Figure 1.2: Bar graph for starting salary data

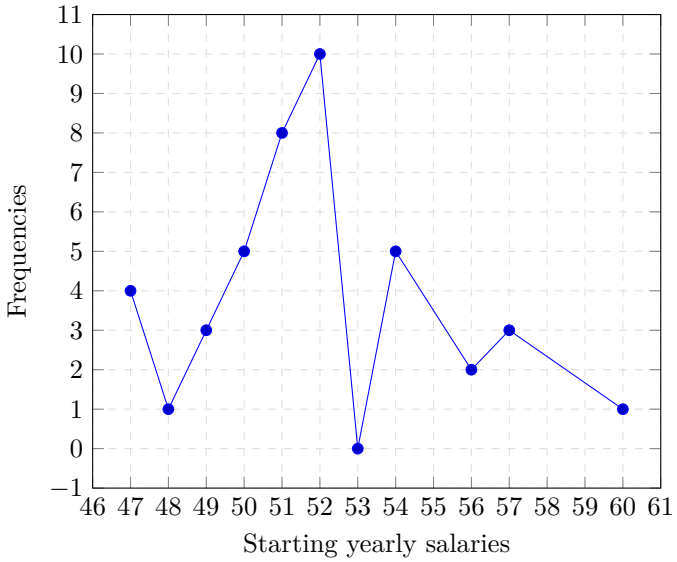

Figure 1.3: Frequency polygon for starting salary data

Example 1. *The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer. These data are represented in the frequency table 1.2 and also in the pie chart 1.4.*

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

Table 1.2: Frequency table for different types of cancers

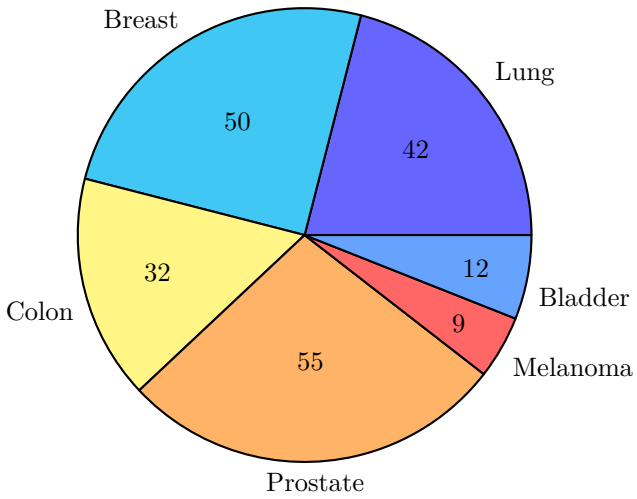

Figure 1.4: Pie chart for different types of cancers

1.1.1 Grouped Data

1.067	919	1.196	785	1.126	936	918	1.156	920	948
855	1.092	1.162	1.170	929	950	905	972	1.035	1.045
1.157	1.195	1.195	1.340	1.122	938	970	1.237	956	1.102
1.022	978	832	1.009	1.157	1.151	1.009	765	958	902
923	1.333	811	1.217	1.085	896	958	1.311	1.037	702
521	933	928	1.153	946	858	1.071	1.069	830	1.063
930	807	954	1.063	1.002	909	1.077	1.021	1.062	1.157
999	932	1.035	944	1.049	940	1.122	1.115	833	1.320
901	1.324	818	1.250	1.203	1.078	890	1.303	1.011	1.102
996	780	900	1.106	704	621	854	1.178	1.138	951
1.187	1.067	1.118	1.037	958	760	1.101	949	992	966
824	653	980	935	878	934	910	1.058	730	980
844	814	1.103	1.000	788	1.143	935	1.069	1.170	1.067
1.037	1.151	863	990	1.035	1.112	931	970	932	904
1.026	1.147	883	867	990	1.258	1.192	922	1.150	1.091
1.039	1.083	1.040	1.289	699	1.083	880	1.029	658	912
1.023	984	856	924	801	1.122	1.292	1.116	880	1.173
1.134	932	938	1.078	1.180	1.106	1.184	954	824	529
998	996	1.133	765	775	1.105	1.081	1.171	705	1.425
610	916	1.001	895	709	860	1.110	1.149	972	1.002

Table 1.3: Life in hours of 200 incandescent lamps

500	–	600	2
600	–	700	5
700	–	800	12
800	–	900	25
900	–	1000	58
1000	–	1100	41
1100	–	1200	43
1200	–	1300	7
1300	–	1400	6
1400	–	1500	1

Table 1.4: A class frequency table

1.2 Chebyshev's Inequality

Theorem 1 (Chebyshev's Inequality)**.** *Let $\bar{x}$ and s_x be the sample mean and the sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$, where $s_x > 0$. Also, for any $k \geqslant 1$, take*

$$S_k = \{i \in \{1, \ldots, n\} : 1 \leqslant i \leqslant n, |x_i - \bar{x}| < ks_x\},$$

and let $N(S_k)$ be the number of elements in S_k. Then, we have that

$$\frac{N(S_k)}{n} \geqslant 1 - \frac{n-1}{nk^2} \geqslant 1 - \frac{1}{k^2}.$$

Proof. Notice that we have

$$\begin{aligned}
(n-1)s_x^2 &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \sum_{i \in S_k} (x_i - \bar{x})^2 + \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\geqslant \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\geqslant \sum_{i \notin S_k} k^2 s_x^2 \\
&= N(S_k)\, k^2 s_x^2
\end{aligned}$$

from what it follows that

$$\frac{N(S_k)}{n} \geqslant 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}.$$

The proof is now complete. □

Ford F Series	44813
Toyota Camry	40016
Chevrolet Silverado	37231
Honda Accoord Hybrid	35075
Toyota Corolla Matrix	32535
Honda Civic Hybrid	31710
Chevrolet Impala	26728
Dodge Ram	24206
Ford Focus	23850
Nissan Altima Hybrid	22630

Table 1.5: Top Selling Vehicles in the United States in 2008

Example 2.

Theorem 2 (One-Sided Chebyshev's Inequality)**.** *Let $\bar{x}$ and s_x be the sample mean and the sample standard deviation of the data set $x_1, \ldots, x_n$. Also, for any $k > 0$, let*

$$N(k) = \{i \in \{1, \ldots, n\} : x_i - \bar{x} \geqslant k s_x\}.$$

Then, we have that

$$\frac{N(k)}{n} \leqslant \frac{1}{1+k^2}.$$

Proof. Let $y_i = x_i - \bar{x}$, $i = 1, \ldots, n$. For any $b > 0$, we have that

$$\sum_{i=1}^{n} (y_i + b)^2 \geqslant \sum_{i : y_i \geqslant k s_x} (y_i + b)^2$$

□

1.3 Correção de um exercício da lista

Exercise 1. *The sample mean and sample variance of five data values are, respectively, $\bar{x} = 104$ and $s_x^2 = 4$. If three of the data values are $102, 100, 105$ what are the other two data values?*

Suponha que $x_1^0, x_2^0, x_3^0, \bar{x}, s_x^2 \in \mathbb{R}$ sejam constantes, onde $s_x^2 \geqslant 0$, e que $x_4, x_5 \in \mathbb{R}$ sejam números reais a serem encontrados de forma que

$$\begin{cases} \bar{x} &= \dfrac{x_1^0 + x_2^0 + x_3^0 + x_4 + x_5}{5} \\ s_x^2 &= \dfrac{(x_1^0 - \bar{x})^2 + (x_2^0 - \bar{x})^2 + (x_3^0 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2}{4} \end{cases}$$

Observe, neste caso, que:

$$x_4 + x_5 = 5\bar{x} - \sum_{i=1}^{3} x_i^0 \tag{1.1}$$

$$(x_4 - \bar{x})^2 + (x_5 - \bar{x})^2 = 4s_x^2 - \sum_{i=1}^{3} (x_i^0 - \bar{x})^2 \tag{1.2}$$

de onde segue que:

$$x_4^2 + x_5^2 - 2\bar{x}(x_4 + x_5) + 2\bar{x}^2 = 4s_x^2 - \sum_{i=1}^{3}(x_i^0)^2 + 2\bar{x}\sum_{i=1}^{3}x_i^0 - 3\bar{x}^2,$$

e, assim, que:

$$x_4^2 + x_5^2 - 2\bar{x}\left(5\bar{x} - \sum_{i=1}^{3}x_i^0\right) + 2\bar{x}^2 = 4s_x^2 - \sum_{i=1}^{3}(x_i^0)^2 + 2\bar{x}\sum_{i=1}^{3}x_i^0 - 3\bar{x}^2.$$

Portanto, tem-se que:

$$5\bar{x}^2 + 4s_x^2 - \sum_{i=1}^{3}(x_i^0)^2 = x_4^2 + x_5^2 \geqslant 0.$$

Isto mostra que os valores $x_1^0, x_2^0, x_3^0, \bar{x}, s_x$ precisam satisfazer à relação:

$$\frac{\bar{x}^2}{a^2} + \frac{s_x^2}{b^2} \geqslant 1,$$

onde

$$a = \frac{\sqrt{\sum_{i=1}^{3}(x_i^0)^2}}{\sqrt{5}} \quad \text{e} \quad b = \frac{\sqrt{\sum_{i=1}^{3}(x_i^0)^2}}{2}.$$

Sabendo disso, você consegue agora escolher valores adequados?