Statistics and Data Analysis
Exercises

Edgar Bueno

Exercise 2.21. (From Newbold, modified) Navigating through an airport can be quite chaotic. You could instead choose to spend your waiting time at one of the airport lounges —hospitality areas where you can relax and enjoy food and beverages before you jet off. 10 airport lounges are selected, their entry fees are listed below (in euros):

15 30 25 32 40 18 22 28 25 35

1. Find the mean entry fee.
2. Find the median entry fee.
3. Find the mode of the entry fees.
4. Calculate the range of the entry fee.
5. Calculate the interquartile range of the entry fee.
6. Calculate the standard deviation $S_{x,U}$

Solution 2.21.

1. The mean is calculated as follows:

$$\bar{x}_U = \frac{1}{N}\sum_U x_i = \frac{1}{10}(15+30+\cdots+35) = \frac{270}{10} = 27.$$

2. In order to find the median, first we need to find the x-ordered population:

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	15	18	22	25	25	28	30	32	35	40

Now, as the population size ($N = 10$) is even, we have

$$\breve{x}_U = \frac{1}{2}\left(x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}\right) = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}(25+28) = 26.5.$$

3. The mode is $\dot{x} = 25$ which is the only observation with more than one occurrence.

4. The range is

$$\text{range}_{x,U} = x_{(N)} - x_{(1)} = 40 - 15 = 25.$$

5. In order to find the interquartile range (IQR) first we need to find the first and the third quartile. As the first quartile is the 25th percentile, we have $c = (N-1)p+1 = (10-1)0.25+1 = 3.25$, thus $a = 3$ and $b = 0.25$ and

$$\breve{x}_{0.25} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.25)x_{(3)} + 0.25x_{(4)} = 0.75 \cdot 22 + 0.25 \cdot 25 = 22.75.$$

As the third quartile is the 75th percentile, we have $c = (N-1)p+1 = (10-1)0.75+1 = 7.75$, thus $a = 7$ and $b = 0.75$ and

$$\breve{x}_{0.75} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.75)x_{(7)} + 0.75x_{(8)} = 0.25 \cdot 30 + 0.75 \cdot 32 = 31.5.$$

The interquartile range is then

$$IQR_{x,U} = \breve{x}_{0.75} - \breve{x}_{0.75} = 31.5 - 22.75 = 8.75.$$

6. The variance is

$$S^2_{x,U} = \frac{1}{N-1}\sum_U (x_i - \bar{x}_U)^2 = \frac{1}{10-1}\left((15-27)^2 + (30-27)^2 + \cdots + (35-27)^2\right) = \frac{1}{9}\left((-12)^2 + 3^2 + \cdots + 8^2\right) = \frac{1}{9}(144 + 9 + \cdots + 64) = \frac{526}{9} = 58.4.$$

Thus, the standard deviation is $S_{x,U} = 58.4^{0.5} = 7.64$.

Exercise 2.1. Labor status: The table below gives the numbers of Maltese persons aged 15 and over during the first quarter of 2019 by labor status.

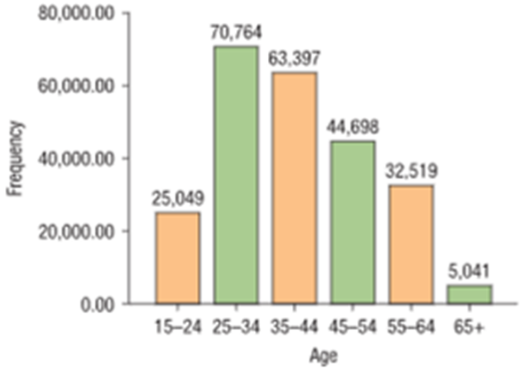
Labor status	Frequency
Employed	241 468
Unemployed	8853
Inactive	165 235

Convert the table to a relative frequency table.

Solution 2.1. We just divide the frequency by the number of observations. For instance the relative frequency of the category *Employed* is 241 468/415 556 = 0.581.

Labor status	Frequency	Relative
Employed	241 468	0.581
Unemployed	8853	0.021
Inactive	165 235	0.398
Total	415 556	1

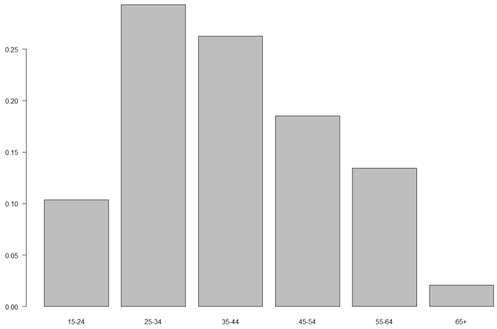
Exercise 2.2. Employment by age: The distribution of employed persons with a main job by age group is shown in the following bar chart. Change this to a bar chart of relative frequencies.



Solution 2.2. First we have to find the relative frequency.

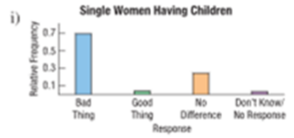
Age	Frequency	Relative
15—24	25049	0.104
25—34	70764	0.293
35—44	63397	0.263
45—54	44698	0.185
55—64	32519	0.135
65+	5041	0.021
Total	241 468	1

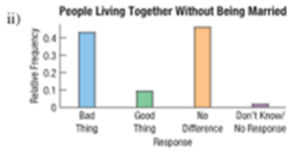
Now we can create the plot:

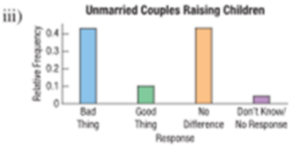


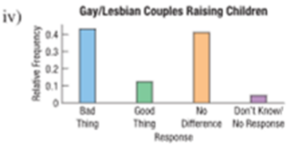
Exercise 2.6. Marriage in decline: Changing attitudes about marriage and families prompted Pew Research to ask how people felt about particular recent trends. For each trend, participants were asked whether the trend "is a good thing", "is a bad thing", or "makes no difference". Some participants said they didn't know or chose not to respond. The following bar charts show the number of each response for each trend. The pie charts show the same data without the trends identified. Match each pie chart with the correct trend and bar chart.

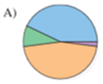
Match each pie chart with the correct trend and bar chart.

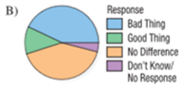


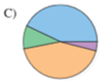


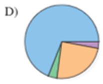






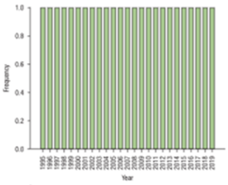


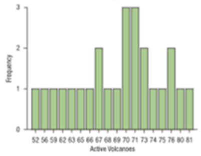




Solution 2.6. The pairs are: i and D, ii and A, iii and C and iv and B.

Exercise 2.7. **Active volcanoes:** Here are two histograms showing the annual number of active volcanoes. One plots the years, and the other plots the number of active volcanoes.



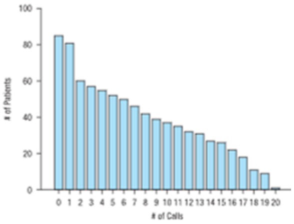


1. For which years is data about active volcanoes available?
2. Why do all the bars in the Year histogram have a height of one unit?
3. Explain what the Active Volcanoes histogram says about active volcanoes.

Solution 2.7. The dataset seems to be a table containing years in the rows. Among the variables we have the year and the number of active volcanoes.

1. From 1995 to 2019;
2. Because each year appears only once in the dataset. This variable is an identifier in this case;
3. It indicates in the vertical axis the number of years having as many active volcanoes as indicated in the horizontal axis. For instance, during one year there were 52 active volcanoes, during three years there were 70 active volcanoes.

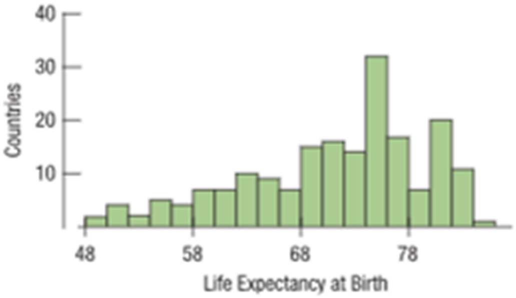
Exercise 2.11. Patient calls: During a particular month, a gastroenterology provider kept record of the number of telephone calls received by his patients. The following histogram shows the distribution of telephone calls received.



What does the above histogram say about the distribution of patient calls received by the provider?

Solution 2.11. The histogram shows a distribution that is skewed to the right or a right-tailed distribution. Many patients receive a few calls and a few patients receive a large number of calls.

Exercise 2.15. Life expectancy: Here are the life expectancies at birth in 190 countries (2014) as collected by the World Health Organization.



1. Describe the shape.
2. How many modes do you see?

Solution 2.15.

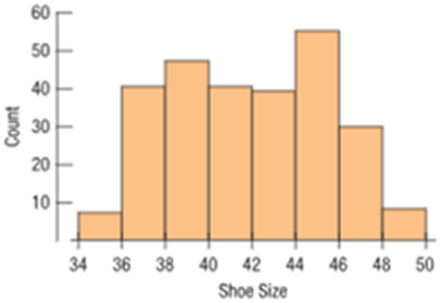
1. Life expectancy is skewed to the left: there are many counties having a large life expectancy and a few having a small one;
2. Not straightforward: I see a unimodal distribution. One of my colleagues sea a multimodal distribution with at least two peaks "Perhaps there are clear groups of countries in this data, where, for example, different regions have different unimodal distributions".

Exercise 2.16. Shoe sizes: A last is a form, traditionally made of wood, in the shape of the human foot. Lasts of various sizes are used by shoemakers to make shoes. In the United States, shoe sizes are defined differently for men and women:

- U.S. men's shoe size = (last size in inches ×3) - 24;
- U.S. women's shoe size = (last size in inches ×3) - 22.5.

But in Europe, they are both: Euro size = last size in cm $\times 3/2$.

Here is a histogram of the European shoe sizes of 269 college students (converted from their reported U.S. shoe sizes):



1. Describe the shape.

2. How many modes do you see?

Solution 2.16.

1. The distribution is fairly symmetric;

2. I see two modes. Probably corresponding to both sexes which define two subpopulations. The first subpopulation (women) seems to be centered around 39; the second one (men) is centered around 45.

Exercise 2.17. Life expectancy II: For the 146 life expectancies in Exercise 15,

1. Which would you expect to be larger: the median or the mean? Explain briefly.

2. Which would you report: the mean or the median? Explain briefly.

Solution 2.17. Weren't they 190 life expectancies?

1. I expect the median to be larger. Those observations that are very small in relation to the rest of the population will pull the mean down;

2. Probably the authors of the book expect me to say "the median because the mean is affected by those very small values" but I would report both values and just make sure to emphasize the fact that the mean is affected by those small observations.

Exercise 2.20. Shoe sizes II: For the shoe sizes in Exercise 16, what might be the problem with either the mean or the median as a measure of center?

Solution 2.20. As it seems like the observations are better described in terms of two subpopulations (men and women), it would be more adequate to report parameters for each of these subpopulations, not necessarily for the whole population.

Exercise 2.21. Life expectancy III For the 190 life expectancies in Exercise 15,

1. Would you report the standard deviation or the IQR?

2. Justify your answer briefly.

Solution 2.21. Probably the authors of the book expect me to say "the IQR because the standard deviation is affected by those very small values" but I would report both values and just make sure to emphasize the fact that the standard deviation is affected by those small observations.

Exercise 2.24. Shoe sizes III: For the shoe sizes in Exercise 16, what might be the problem with either the standard deviation or the IQR as a measure of spread?

Solution 2.24. As it seems like the observations are better described in terms of two subpopulations (men and women), it would be more adequate to report parameters for each of these subpopulations, not necessarily for the whole population.

Exercise 3.1. Regulation amendments The administration of a particular University asked 4000 of its students to express their opinion as to whether they agree or disagree with a set of amendments to the regulations. Students had to choose from "Strongly Agree", "Agree", "Not Sure", "Disagree", and "Strongly Disagree".

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree
Males	499	616	47	395	343
Females	689	571	92	436	312

1. What percent of male students strongly agree with the regulation amendments?
2. What percent of male students disagree or strongly disagree with the regulation amendments? What is the comparable percentage of female students?
3. Compare the distribution of opinions between male and female students.
4. Is it reasonable to conclude that 30% of all University students strongly agree with the regulation amendments? Why or why not?

Solution 3.1. The question concerns two categorical variables *Sex* and *Opinion.* In question 1 we are only interested in the male students, this means that we have to condition on sex. Before doing that, let us add the marginals to the table:

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree	Total Total
Males	499	616	47	395	343	1900
Females	689	571	92	436	312	2100
Total	1188	1187	139	831	655	4000

The following table shows the distribution of the opinion conditioned on sex:

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree	Total Total
Males	0.263	0.324	0.025	0.208	0.181	1
Females	0.328	0.272	0.044	0.208	0.149	1
Total	0.297	0.297	0.035	0.208	0.164	1

Now we are ready to answer the questions:

1. 26.3% of the male students strongly agree with the amendments.
2. 38.8% (0.208+0.181) of the male students disagree or strongly disagree with the amendments, whereas 35.6% (0.208+0.149) of the female students express the same opinion.
3. Although there are some differences, in general terms, the distribution is quite similar: around 60% of students of either sex are in agreement with the amendments and around 36% are in disagreement.
4. Yes. We have observed a *sample* of 4000 students. In this sample, around 30% (0.297) of the students have expressed that they strongly agree with the amendments. Although this true value may be different in the population, we expect it to be quite close to what we have observed in the sample. (If the sample has been selected properly.)

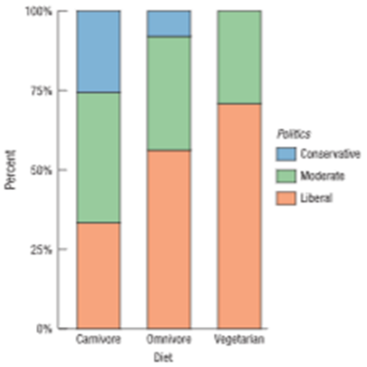
Exercise 3.3. Regulation amendments again Consider the sample data in Exercise 3.1.

1. What is the conditional distribution (in percentages) of the opinions of female students on the regulation amendments?
2. Find the conditional distribution of the opinions of male students when the categories are combined into Positive (Strongly Agree or Agree), Neutral (Not Sure), and Negative (Disagree or Strongly Disagree).

Solution 3.3.

1. See the second row of the conditional distribution table in Exercise 3.1. above.
2. We will have that 58.7% (0.263+0.324) of the male students have a positive opinion, 2.5% have a neutral opinion and 38.8% (0.208+0.181) have a negative opinion.

Exercise 3.5. Diet and politics A survey of 299 undergraduate students asked about respondents' diet preference (Carnivore, Omnivore, Vegetarian) and political alignment (Liberal, Moderate, Conservative). Here is a stacked bar chart:

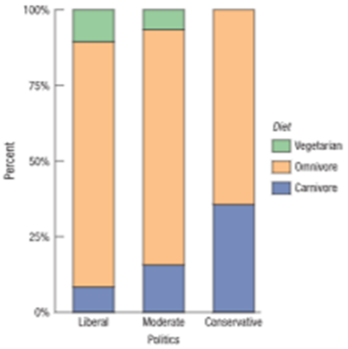


1. Describe what this plot shows using the concept of a conditional distribution.
2. Do you think the differences here are real? Explain.

Solution 3.5.

1. The stacked bar chart shows the distribution of political alignment conditioned on diet preference.
2. The question is quite ambiguous: what do they mean by *real*? If they mean "big" as for concluding that the differences are "significant" and not due to some randomness, we will cover that topic later in the course. If they mean "meaningful" I would answer yes. In our sample a larger proportion of vegetarian students are liberal compared to carnivore students, for instance.

Exercise 3.6. Diet and politics revisited Here are the same data as in Exercise 5 but displayed differently:

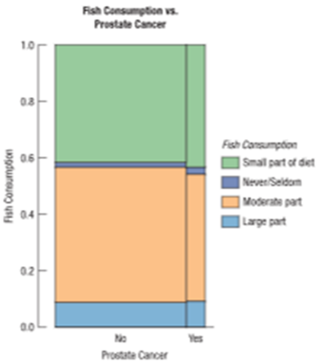


1. Describe what this plot shows using the concept of a conditional distribution.
2. Do you think the differences here are real? Explain.

Solution 3.6.

1. The stacked bar chart shows the distribution of diet preference conditioned on political alignment.
2. Once again the question is quite ambiguous: what do they mean by *real*? If they mean "big" as for concluding that the differences are "significant" and not due to some randomness, we will cover that topic later in the course. If they mean "meaningful" I would answer yes. In our sample a larger proportion of conservative students are carnivore compared to liberal students, for instance.

Exercise 3.7. Fish and prostate cancer revisited Here is a mosaic plot of the data on Fish consumption and Prostate cancer from the Step-by-Step Example on page 97



1. From the mosaic plot, about what percent of all men in this survey were diagnosed with prostate cancer?

2. Are there more men who had cancer and never/seldom ate fish, or more who didn't have cancer and never/seldom ate fish?

3. Which is higher: the percent of men with cancer who never/seldom ate fish, or the percent of men without cancer who never/seldom ate fish?

Solution 3.7.

1. The width of the bar for the category *Yes* is more or less one seventh of the overall width, so around 14%.

2. No.

3. The question asks about the frequency of never/seldom eating fish conditioning on having cancer or not. Thus, in order to answer the question we have to look at the height of the bars, not the area: the proportion of men with cancer who never/seldom ate fish is higher than the proportion of men without cancer who never/seldom ate fish.

Exercise 3.9. Diet and politics III Are the patterns seen in Exercises 5 and 6 relating diet to political opinion the same for men and women? Here are two contingency tables:

Men	Carnivore	Omnivore	Vegetarian
Liberal	9	74	5
Moderate	12	54	1
Conservative	9	14	0

Women	Carnivore	Omnivore	Vegetarian
Liberal	4	53	12
Moderate	4	26	6
Conservative	1	4	0

1. Are women or men more likely to be conservative carnivores?

2. Are liberal vegetarians more likely to be women or men?

Solution 3.9. Now we are analyzing three categorical variables: sex, diet and political opinion.

1. The first question is asking to condition on sex. First let us add the marginals to the table:

Men	Carnivore	Omnivore	Vegetarian	Total
Liberal	9	74	5	88
Moderate	12	54	1	67
Conservative	9	14	0	23
Total	30	142	6	178

Women	Carnivore	Omnivore	Vegetarian	Total
Liberal	4	53	12	69
Moderate	4	26	6	36
Conservative	1	4	0	5
Total	9	83	18	110

Total	Carnivore	Omnivore	Vegetarian	Total
Liberal	13	127	17	157
Moderate	16	80	7	103
Conservative	10	18	0	28
Total	39	225	24	288

Now we can find the distribution of diet and political opinion conditioned on sex:

Men	Carnivore	Omnivore	Vegetarian	Total
Liberal	0.051	0.416	0.028	0.494
Moderate	0.067	0.303	0.006	0.376
Conservative	0.051	0.079	0.000	0.129
Total	0.169	0.798	0.034	1

Women	Carnivore	Omnivore	Vegetarian	Total
Liberal	0.036	0.482	0.109	0.627
Moderate	0.036	0.236	0.055	0.327
Conservative	0.009	0.036	0.000	0.045
Total	0.082	0.755	0.164	1

We observe that 5% of men are conservative carnivore whereas 0.9% of women are conservative carnivore. It is worth mentioning that we are simply describing the data that we have, we cannot generalize these results to any other population in a straightforward manner.

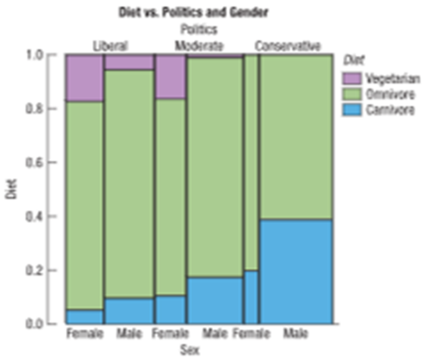
2. The second question is asking to condition on political view and diet:

Men	Carnivore	Omnivore	Vegetarian	Total
Liberal	0.692	0.583	0.294	0.561
Moderate	0.75	0.675	0.143	0.65
Conservative	0.9	0.778	—	0.821
Total	0.769	0.631	0.25	0.618

Women	Carnivore	Omnivore	Vegetarian	Total
Liberal	0.308	0.417	0.706	0.439
Moderate	0.25	0.325	0.857	0.35
Conservative	0.1	0.222	—	0.179
Total	0.231	0.369	0.75	0.382

Now we see that 70.6% of liberal vegetarians are women and 29.4 are men.

Exercise 3.25. Diet and politics IV Here is a mosaic plot of the data on Diet and Politics from Exercise 5 combined with data on Gender.



1. Are there more men or women in the survey? Explain briefly.
2. Does there appear to be an association between Politics and Gender? Explain briefly.
3. Does there appear to be an association between Politics and Diet? Explain briefly.
4. Does the association between Politics and Diet seem to differ between men and women? Explain briefly.

Solution 3.25.

1. There are more men than women: each bar for males is wider than their female counterpart.
2. Yes, the width of the bars for females become proportionally smaller with respect to males as long as we move towards "conservative".
3. We already saw this before but now in the context of a mosaic plot: the proportion of vegetarians become smaller as we move towards the right.
4. It is not easy for me to say.

Before we close this exercise I believe it is important to mention that contingency tables and mosaic plots are not the most adequate methods for multidimensional analysis. We have "only" three variables here and things became really messy. There are other methods that are more adequate e.g. multiple correspondence analysis.

Exercise 3.39. Test answers A student had to undergo a test which consisted of 50 Yes/No questions. The following contingency table compares the student's answers with the actual correct answers:

		Actual answers	
		Yes	No
Student's	Yes	18	9
answers	No	6	17

1. For what percentage of questions was the actual answer a No?

2. For what percentage of questions did the student obtain a correct answer?

3. For what percentage of questions did the student answer an incorrect Yes answer?

4. Do you see evidence of an association between the type of answer and the ability of the student to choose the correct answer? Write a brief explanation, including an appropriate graph.

Solution 3.39. Let us start by adding the marginals:

		Actual answers		
		Yes	No	Total
Student's	Yes	18	9	27
answers	No	6	17	23
	Total	24	26	50

1. Let us find the joint relative frequency table:

		Actual answers		
		Yes	No	Total
Student's	Yes	0.36	0.18	0.54
answers	No	0.12	0.34	0.46
	Total	0.48	0.52	1

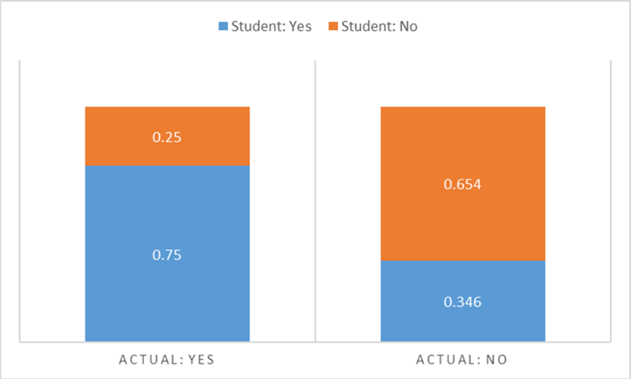
   So for 52% of the questions the actual answer was No.

2. The student obtained a correct answer for 70% (0.36+0.34) of the questions.

3. 18%.

4. Note that we would like to see some association: the student should answer Yes when the actual answer is Yes. Let us find the distribution of the student's answer conditioned on the actual answer:

		Actual answers		
		Yes	No	Total
Student's	Yes	0.75	0.346	0.54
answers	No	0.25	0.654	0.46
	Total	1	1	1



There is a clear association between the type of actual answer and the ability of the student: the student is more likely to choose Yes when the actual answer is, indeed, yes.

Exercise 3.41. Blood pressure A company held a blood pressure screening clinic for its employees. The results are summarized in the table below by Age and Blood pressure:

		Age		
		Under 30	30—49	Over 50
	Low	27	37	31
Blood	Normal	48	91	93
pressure	High	23	51	73

1. Find the marginal distribution of blood pressure level.

2. Find the conditional distribution of blood pressure level within each age group.

3. Compare these distributions with a segmented bar graph.

4. Write a brief description of the association between age and blood pressure among these employees.

5. Does this prove that people?s blood pressure increases as they age? Explain.

Solution 3.41.

1. Let us add both marginals to the table

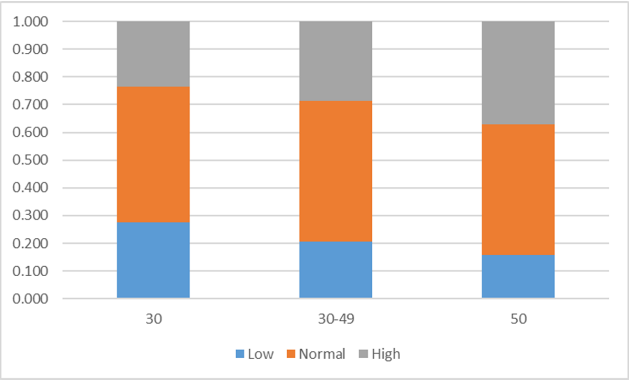
		Age			
		Under 30	30—49	Over 50	Total
	Low	27	37	31	95
Blood	Normal	48	91	93	232
pressure	High	23	51	73	147
	Total	98	179	197	474

Now we can find the distribution of blood pressure level conditioned on age group:

		Age			
		Under 30	30—49	Over 50	Total
	Low	0.276	0.207	0.157	0.200
Blood	Normal	0.490	0.508	0.472	0.489
pressure	High	0.235	0.285	0.371	0.310
	Total	1	1	1	1

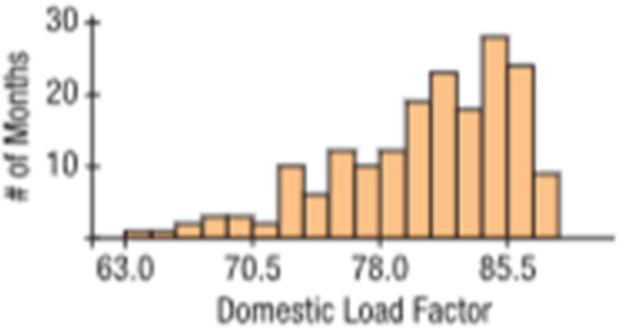
The marginal distribution of blood pressure level is shown in the last column.

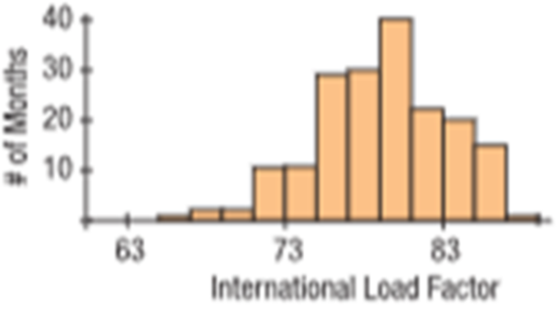
2. It was shown in the previous table.
3. The following plot shows the segmented (or stacked) bar chart. We can see that a higher age is associated with a higher blood pressure



4. We can see association, not causality.

Exercise 4.1. Load factors 2017 The Research and Innovative Technology Administration of the Bureau of Transportation Statistics reports load factors (passenger–miles as a percentage of available seat–miles) for commercial airlines for every month from October 2002 through 2017. Here are histograms and summary statistics for the domestic and international load factors for this time period:





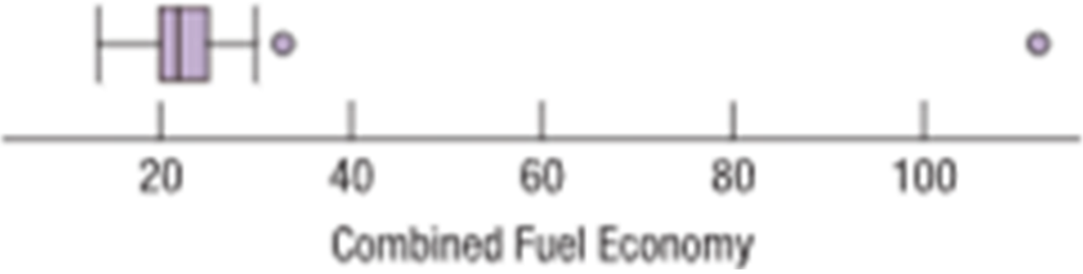
Variable	Count	Mean	Median	StDev	IQR
Domestic load factor	183	80.6441	81.54	5.23262	7.3775
International load factor	183	79.2046	79.27	4.22962	5.7900

Compare and contrast the distributions.

Solution 4.1.

- Both populations are left–skewed and both have a similar range (they take values more or less between 65 and 90, for a range of, more or less, 25), but the domestic flights are more skewed than the international flights.
- Both populations are unimodal.
- The domestic flights are centered at values slightly larger than the international flights.
- The observations vary less in the international flights than in the domestic flights.

Exercise 4.3. Fuel economy The boxplot shows the fuel economy ratings for 67 model year 2012 subcompact cars. Some summary statistics are also provided. The extreme outlier is the Mitsubishi i-MiEV, an electric car whose electricity usage is equivalent to 112 miles per gallon.

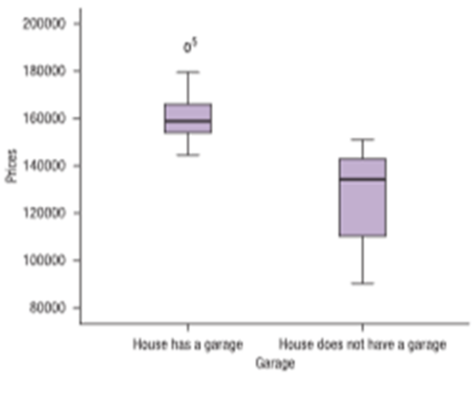


Mean	SD	Min	Q1	Med	Q3	Max	n
23.76	11.87	14	20	22	25	112	67

If that electric car is removed from the dataset, how will the standard deviation be affected? The IQR?

Solution 4.3. By removing that outlier the standard deviation will decrease "significantly", whereas the IQR is expected to remain almost the same.

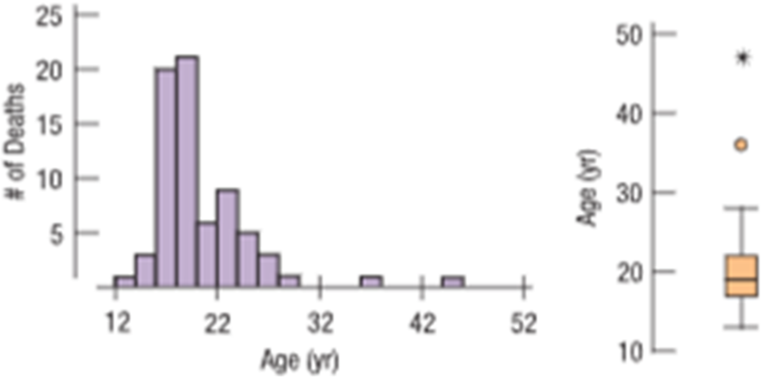
Exercise 4.14. House prices The following boxplot displays the prices (in British pounds) of 35 houses, 19 of which have a garage and 16 of which do not. Compare the two groups and interpret what you find.



Solution 4.14.

- The price of houses with and without a garage largely differ: the boxplots do not "overlap" each other. The price of houses with a garage is evidently larger than the price of houses without a garage.
- The set of houses with a garage show a positive skewness, whereas the set of houses with a garage show a negative skewness.
- The price of the houses with a garage varies less than the price of houses without a garage.

Exercise 4.17. Rock concert deaths Crowd Management Strategies (www.crowdsafe.com) monitors accidents at rock concerts. In their database, they list the names and other variables of victims whose deaths were attributed to "crowd crush" at rock concerts. Here are the histogram and boxplot of the victims' ages for data from a one-year period:

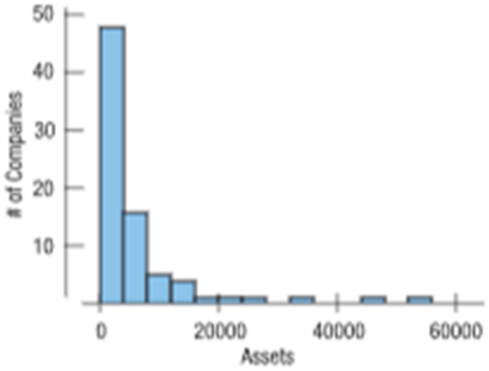


1. What features of the distribution can you see in both the histogram and the boxplot?
2. What features of the distribution can you see in the histogram that you cannot see in the boxplot?
3. What summary statistic would you choose to summarize the center of this distribution? Why?
4. What summary statistic would you choose to summarize the spread of this distribution? Why?

Solution 4.17.

1. Both plots allow to see that **i.** the distribution is skewed to the right, **ii.** although a bit easier in the boxplot than in the histogram, the median is around 18—19 years; **iii.** there are a couple of extreme observations; **iv.** the observations vary between 12 and 45 years, for a range of 33 years; **v.** although a bit easier in the boxplot than in the histogram, the first quartile is around 16 years, the third quartile is around 24 years, so the IQR is around 8 years.
2. The histogram shows a bimodal distribution with a "big" mode and a "minor one". The "big one" is around 18 years and the another one is around 23 years.
3. The median, as the extreme observations will not affect it too much.
4. The IQR, for the same reason as above.

Exercise 4.45. Assets Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the Forbes list of the nation's top corporations: (Data in Companies)



1. What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
2. What would you suggest doing with these data if we want to understand them better?

Solution 4.45.

1. The population is very skewed to the right, therefore most companies fall in the first bar. This makes it hard to visually identify measures of centrality or spread with precision. For instance, all we can see about the first quartile and the median is that both fall somewhere between 0 and 4000 millions of dollars (a quite large interval). The third quartile falls somewhere between 4000 and 8000 millions of dollars.
2. We can either transform the data (using, for instance, a log transformation) or describe the population after removing the very large outliers.

Exercise 5.1. Stats test The mean score on the Stats exam was 75 points with a standard deviation of 5 points, and Gregor's z-score was -2. How many points did he score?

Solution 5.1. Let x_i be the score on the exam of the ith student ($i = 1, 2, \cdots, N$). In particular, let x_1 be Gregor's score. We know that $\bar{x} = 75$, $S_x = 5$ and $z_1 = -2$.

We have that

$$-2 = z_1 = \frac{x_1 - \bar{x}}{S_x} \Longrightarrow x_1 = z_1 S_x + \bar{x} = -2 \cdot 5 + 75 = 65.$$

Exercise 5.2. Mensa People with z-scores above 2.5 on an IQ test are sometimes classified as geniuses. If IQ scores have a mean of 100 and a standard deviation of 15 points, what IQ score do you need to be considered a genius?

Solution 5.2. Let x_i and z_i be, respectively, the IQ score and the z-score of the ith individual ($i = 1, \cdots, N$). We have that a person with a z-score larger than 2.5 is considered a genius, we are looking for the corresponding x-value:

$$z = \frac{x - \bar{x}}{S_x} > 2.5 \Longrightarrow x > 2.5 S_x + \bar{x} = 2.5 \cdot 15 + 100 = 137.5$$

People with IQ scores above 137.5 would be considered as geniuses.

Exercise 5.4. Placement exams Clara took her college's placement exams in Physics and Statistics. In Physics, she scored 85 and in Statistics 88. The overall results on the Physics exam had a mean of 74 and a standard deviation of 6, while the mean Statistics score was 80, with a standard deviation of 7. On which exam did she do better compared to the other students?

Solution 5.4. Let x_i and y_i be, respectively, the score of the ith student in Physics and in Statistics. In particular, let $x_1 = 85$ and $y_1 = 88$ be Clara's results. We know that $\bar{x} = 74$, $S_x = 6$, $\bar{y} = 80$ and $S_y = 7$. Let us find Clara's z-scores on each exam.

$$z_1^p = \frac{x_1 - \bar{x}}{S_x} = \frac{85 - 74}{6} = 1.83 \qquad \text{and} \qquad z_1^s = \frac{y_1 - \bar{y}}{S_y} = \frac{88 - 80}{7} = 1.14.$$

So compared to the other students, Clara performed better in the Physics exam.

Exercise 5.5. Food packages A food company that ships food packages worldwide claims that its food packages have a median weight of 3 kilograms and an IQR of 1.9 kilograms.

1. The company plans to include a new advertisement sheet weighing 100 grams in each package. What will be the new median and IQR?

2. If the company recorded the shipping weights of these new packages in pounds instead of kilograms, what would be the median and IQR? (0.4536 kg = 1 lb)

Solution 5.5. Let x_i be the *original* weight of the ith package. We have that $\breve{x} = 3$ and $IQR_x = 1.9$.

1. Let y_i be the weight of the ith package with the advertisement. As 100 grams are equivalent to 0.1 kg, we have that $y_i = x_i + 0.1$. Thus

$$\breve{y} = \breve{x} + 0.1 = \breve{3} + 0.1 = 3.1 \qquad \text{and} \qquad IQR_y = IQR_x = 1.9.$$

2. Let w_i be the weight of the ith package with the advertisement in pounds. As one pound is equivalent to 0.4536 kg then one kg is equivalent to $1/0.4536 = 2.205$ pounds. We have that $w_i = 2.205(x_i + 0.1)$. Thus

$$\breve{w} = 2.205(\breve{x} + 0.1) = 2.205(3 + 0.1) = 6.834 \qquad \text{and} \qquad IQR_w = 2.205\, IQR_x = 4.189.$$

Exercise 5.8. Women's shoe sizes The shoe size data for women has a mean of 38.46 and a standard deviation of 1.84. To convert to U.S. sizes, use

$$USsize = EuroSize \times 0.7865 - 22.5.$$

1. What is the mean women's shoe size for these respondents in U.S. units?

2. What is the standard deviation in U.S. units?

Solution 5.8. Let x be the EuroSize and y be the USsize. We know that $\bar{x} = 38.46$ and $S_x = 1.84$.

We have $y = 0.7865x - 22.5$. The expression we have used in the course is of the type $y = b(x + a)$, first we shift and then we scale. This expression is of the type $y = cx + d = 0.7865\,x - 22.5$, first we scale and then we shift. We can answer the solution in at least two ways: the "short way" or the long way.

We know that scaling will equally affect the mean and the standard deviation. By shifting next, only the mean will be affected, not the standard deviation, so we have

$$\bar{y} = c\bar{x} + d = 0.7865 \cdot 38.46 - 22.5 = 0.7865 \cdot 38.46 - 22.5 = 7.75$$

and

$$S_y = cS_x = 0.7865 \cdot 1.84 = 1.45.$$

If we don't see it this way we can rewrite the expression given so it fits into the form we are familiar with:

$$y = 0.7865\,x - 22.5 = 0.7865\left(x - \frac{22.5}{0.7865}\right) = 0.7865\,(x - 28.61)\,,$$

which is of the form $y = b(x + a)$. This yields

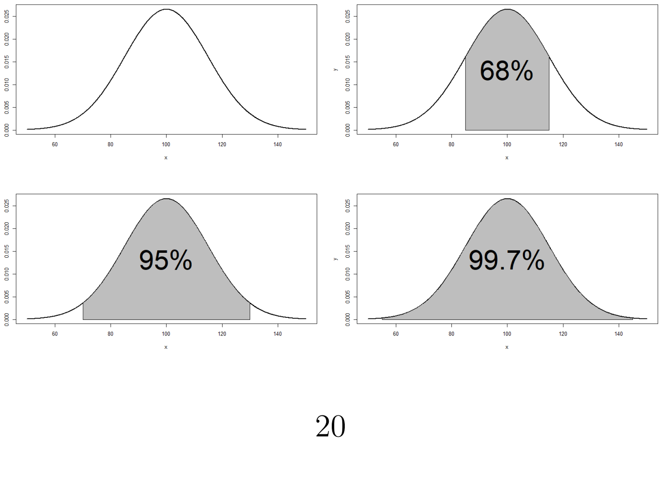
$$\bar{y} = b(\bar{x} + a) = 0.7865(38.46 - 28.61) = 7.75 \qquad \text{and} \qquad S_y = bS_x = 0.7865 \cdot 1.84 = 1.45.$$

Exercise 5.10. IQ Some IQ tests are standardized to a Normal model, with a mean of 100 and a standard deviation of 15.

1. Draw the model for these IQ scores. Clearly label it, showing what the 68–95–99.7 Rule predicts.
2. In what interval would you expect the central 95% of IQ scores to be found?
3. About what percent of people should have IQ scores above 115?
4. About what percent of people should have IQ scores between 70 and 85?
5. About what percent of people should have IQ scores above 130?

Solution 5.10. We have not discussed the Normal distribution, but we can still solve the exercise. Just read the exercise as "Some IQ tests follow a symmetric, unimodal and bell-shaped distribution with a mean of 100 and a standard deviation of 15".

What the 68–95–99.7 Rule says is that for a variable with a symmetric, unimodal and bell-shaped distribution, approximately 68% of the observations will lie within one standard deviation of the mean, approximatley 95% of the observations will lie within two standard deviations of the mean and approximately 99.7% of the observations will lie within three standard deviations of the mean. This is illustrated in the figure below.



1. We expect the central 95% of IQ score to be found in the interval
$$(\bar{x} - 2S_x\,, \bar{x} + 2S_x) = (100 - 2 \times 15\,, 100 + 2 \times 15) = (70\,, 130).$$
2. We expect that 32% of people will be under 85 or over 115. As the distribution is symmetric, we expect 16% of people will be over 115.
3. We expect that 27% of people will be between 70 and 85 or between 115 and 130. As the distribution is symmetric, we expect 13.5% of people will be between 70 and 85.
4. We expect that 5% of people will be under 70 or over 130. As the distribution is symmetric, we expect 2.5% of people will be over 130.

Exercise 5.14. IQs revisited Based on the Normal model $N(100\,, 15)$ describing IQ scores, what percent of people's IQs would you expect to be

1. over 80?
2. under 90?
3. between 112 and 132?

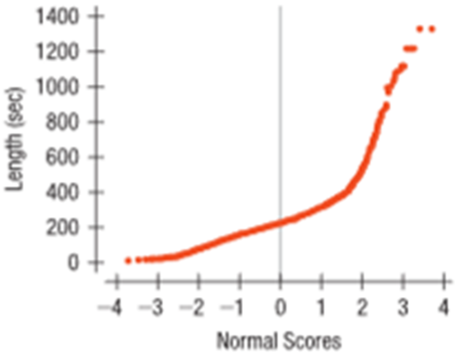
Solution 5.14. We have relegated the discussion of the normal distribution to the second part of the course. So we cannot solve this exercise yet.

Exercise 5.16. Annual incomes The mean annual income of a group of assistant chefs is EURO 20 500, while the median income is EURO 24 595. The standard deviation is EURO 1 800.

1. If a Normal model is used for these incomes, what would be the annual income of the top 1%?
2. How confident are you in the answer of part 1?
3. Do you think that the Normal model is appropriate for these incomes? Explain.

Solution 5.16. We have relegated the discussion of the normal distribution to the second part of the course. So we cannot solve this exercise yet.

Exercise 5.17. Music library Corey has 4929 songs in his computer's music library. The lengths of the songs have a mean of 242.4 seconds and standard deviation of 114.51 seconds. A Normal probability plot of the song lengths looks like this:



1. Do you think the distribution is Normal? Explain.

2. If it isn't Normal, how does it differ from a Normal model?

Solution 5.17. We have relegated the discussion of the normal distribution to the second part of the course. So we cannot solve this exercise yet.

Exercise 5.30. Car speeds 100 John Beale of Stanford, California, recorded the speeds of cars driving past his house, where the speed limit read 20 mph. The mean of 100 readings was 23.84 mph, with a standard deviation of 3.56 mph. (He actually recorded every car for a two-month period. These are 100 representative readings.)

1. How many standard deviations from the mean would a car going under the speed limit be?

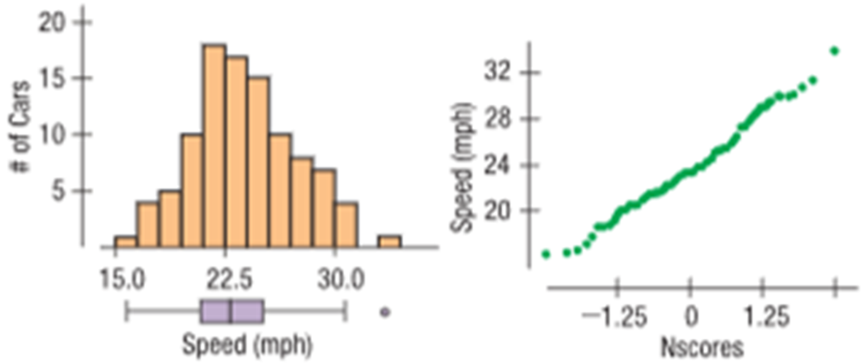
2. Which would be more unusual, a car traveling 34 mph or one going 10 mph?

Solution 5.30. Let x_i be the speed of the ith car, we know that $\bar{x} = 23.84$ and $S_x = 3.56$.

1. We have
$$z = \frac{x - \bar{x}}{S_x} = \frac{20 - 23.84}{3.56} = -1.07.$$
So a car going under the speed limit would be running at least 1.07 standard deviations under the mean.

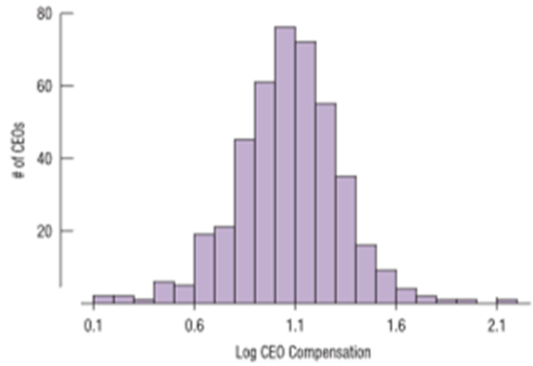
2. Let us say that $x_1 = 34$ and $x_2 = 10$, we have
$$z_1 = \frac{x_1 - \bar{x}}{S_x} = \frac{34 - 23.84}{3.56} = 2.85 \qquad \text{and} \qquad z_2 = \frac{x_2 - \bar{x}}{S_x} = \frac{10 - 23.84}{3.56} = -3.89.$$
So a car going at 10 mph is more remarkable than one going at 34, as it is further away from the mean.

Exercise 5.42. Car speeds 100, the picture For the car speed data in Exercise 30, here are the histogram, boxplot, and Normal probability plot of the 100 readings. Do you think it is appropriate to apply a Normal model here? Explain.



Solution 5.42. We have relegated the discussion of the normal distribution to the second part of the course. So we cannot solve this exercise yet.

Exercise 5.48. CEO compensation logged and sampled Suppose we take logarithms of the CEO compensations in Exercise 47. The histogram of log Compensation looks like this:

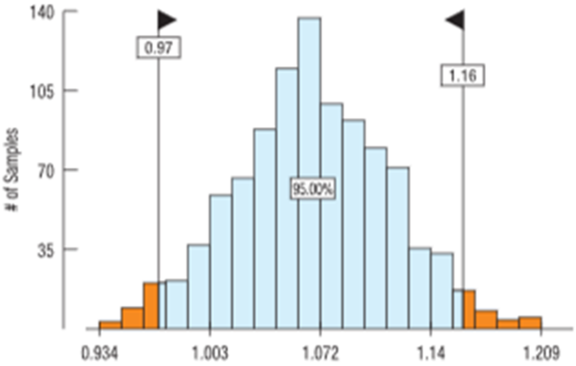


with a mean of 1.07 and a standard deviation of 0.26.

1. According to the Normal model, what percent of CEOs would you expect to earn more than 2 standard deviations above the mean compensation?

2. Is that percentage appropriate for these data?

Now let's draw samples of 30 CEOs from the logged data. We drew 1000 samples and found their means. The distribution of means looks like this:



with a standard deviation of 0.05.

3. Do you think the 68–95–99.7 Rule applies to these means?

Solution 5.48. We have not discussed the Normal distribution, but we can still solve the exercise. Let x_i be the log Compensation of the ith CEO. We have $\bar{x} = 1.07$ and $S_x = 0.26$.

1. Just read the exercise as "According to the 68–95–99.7 Rule, what percent of CEOs would you expect to earn more than 2 standard deviations above the mean compensation?".

   According to the rule we expect around 95% of the observations to lie within two standard deviations of the mean. Equivalently, we expect around 5% of the observations to lie outside two standard deviations of the mean. By symmetry, we expect 2.5% of the observations to lie over two standard deviations of the mean.

2. Yes and No. We can expect around 2.5% of the log Compensations to be over two standard deviations of the mean. But we cannot expect around 2.5% of the *true* Compensations to be over two standard of the mean.

3. Yes. The histogram looks symmetric, unimodal and bell-shaped with a mean of (around) $\bar{x} = 1.07$ and a standard deviation of $S_x = 0.05$. According to the Rule we would have that around 95% of the observations would lie in the interval

   $$(\bar{x} - 2S_x \, , \bar{x} + 2S_x) = (1.07 - 2 \cdot 0.05 \, , 1.07 + 2 \cdot 0.05) = (0.97 \, , 1.17).$$

   Which coincides pretty well with the interval reported in the plot.