Exercise sheet 1
Solutions

Statistics Pre-course
Alfonso Russo

Exercise 1

Consider the following observations:

$$0,\ 43,\ 70,\ 14,\ 22,\ 13,\ 90,\ 50$$

(a) Find the mean.
Answer: the mean is given by

$$(0 + 43 + 70 + 14 + 22 + 13 + 90 + 50)/8 = 37.35$$

(b) FInd the median. **Answer:** First, sort the observations:

$$0,\ 13,\ 14,\ 22,\ 43,\ 50,\ 70,\ 90.$$

Since n is even, the median is given by

$$(22 + 43)/2 = 32.5$$

Exercise 2

The owner of a company in downtown Rome is concerned about the large use of gasoline by her employees due to urban sprawl and traffic congestion. She'd like to promote the use of public transportation. She decides to investigate how many kilometers her employees travel on public transportation during a typical day. The values for her 10 employees (recorded to the closest kilometer) are

$$0,\ 0,\ 4,\ 0,\ 0,\ 0,\ 10,\ 0,\ 6,\ 0$$

(a) Find and interpret the mean, mode and median.
Answer:

- Mean:
$$(0 + 0 + 4 + 0 + 0 + 0 + 10 + 0 + 6 + 0)/10 = 2$$
- Median:
Sort the data
$$0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 4,\ 6,\ 10$$
The median is the average of the two middle values, which is 0

- Mode:

x_i	n_i
0	7
4	1
6	1
10	1

(b) She hires an additional employee who lives in a different city and travels 90km a day on public transport. Recompute the mean and median and comment.
Answer:
The mean in now $(0+0+4+0+0+0+10+0+6+0+90)/11 = 10$.
The median is still 0. Sort the data $(0,0,0,0,0,0,0,4,6,10,90)$ and pick the middle value.

Exercise 3

In a sample of 4838 individuals, subjects were asked how many people who they personally know have completed their PhD in the past six months. The table summarizes response.

People Dismissed	Frequency
0	3944
1	279
2	97
3	40
4 or more	23
Total	**4383**

(a) Find the mean (use 4.5 as score for the class "4 or more").
Answer:

x_i	n_i	$x_i * n_i$
0	3944	0
1	279	279
2	97	194
3	40	120
4.5	23	103.5
Total	**4383**	**696.5**

Thus the mean is 696.6/4838 = 0.16.

(b) Find the median. **Answer:**

x_i	n_i	f_i	F_i
0	3944	0.90	0.90
1	279	0.064	0.96
2	97	0.022	0.986
3	40	0.009	0.995
4 or more	23	0.005	1
Total	**4383**	**696.5**	

Exercise 4

Consider the following two sets of observations:

$$\mathcal{S}_1 = 2,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4$$
$$\mathcal{S}_2 = 2,\ 3,\ 3,\ 3,\ 3,\ 3,\ 4$$

(a) Determine which set has the higher variability.
Answer:

- $\mathcal{S}_1$ has $\bar{x} = 3.29$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-1.2857143	1.65306122
3	-0.2857143	0.08163265
3	-0.2857143	0.08163265
3	-0.2857143	0.08163265
4	0.7142857	0.51020408
4	0.7142857	0.51020408
4	0.7142857	0.51020408
Total	0	3.43

The variance is then $3.43/7 = 0.49$

- $\mathcal{S}_2$ has $\bar{x} = 3$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-1	1
3	0	0
3	0	0
3	0	0
3	0	0
3	0	0
4	1	1
Total	0	2

The variance is then $2/7 = 0.29$

Exercise 5

A Professor recorded the following scores from a Statistics exam:

$$70,\ 84,\ 59,\ 73,\ 86,\ 35,\ 81,\ 75.$$

(a) Find the mean and the median.
Answer:
The mean is given by

$$(70 + 84 + 59 + 73 + 86 + 35 + 81 + 75)/8 = 70.375$$

The median is

$$35, 59, 70, 73, 75, 81, 84, 86 \implies (73 + 75)/2 = 74.$$

(b) Based on your results from the previous point, can you tell something about the distribution of the data?
Answer: Since the median is greater than the mean we suppose that the distribution is asymmetric and left skewed.

(c) Find the standard deviation. **Answer:**

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
35	-35.375	1251.39
59	-11.375	19.39
70	-0.375	0.14
73	2.625	6.89
75	4.625	21.39
81	10.625	112.89
84	13.625	185.64
86	15.625	244.14
Total	0	1951.875

The standard deviation is given by $s = \sqrt{1951.875/8} \approx 15.62$

Exercise 6

The following table reports answers to the question "How many times have you flown overseas in the last 12 months?"

Flights	0	1	2	3	4
Count	25	15	20	5	0

(a) Find mean, variance and standard deviation.

Answer:

x_i	n_i	$n_i * x_i$	$(x_i - \bar{x})^2$	$n_i * (x_i - \bar{x})^2$
0	25	0	-1.08^2	29.16
1	15	15	-0.08^2	0.096
2	20	40	0.92^2	16.928
3	5	15	1.92^2	18.432
4	0	0	2.92^2	0
Total	65	70		64.616

The mean is $\bar{x} = 70/65 = 1.08$. The variance is obtained as $s^2 = 64.616/65 = 0.99$ and therefore the standard deviation is $s = \sqrt{0.99} = 0.995$.

Ecercise 7

We report fertility rates for 18 countries worldwide in the following table:

Country	Fertility	Country	Fertility
Austria	1.4	Netherlands	1.7
Belgium	1.7	Norway	1.8
Denmark	1.8	Spain	1.3
Finland	1.7	Sweden	1.6
France	1.9	Switzerland	1.4
Germany	1.3	United Kingdom	1.7
Greece	1.3	United States	2.0
Ireland	1.9	Canada	1.5
Italy	1.3	Mexico	2.4

(a) Find quartiles (Q_1, Q_2, Q_3) for the fertility rates.
Answer:
Sort the data

| 1.3, 1.3, 1.3, 1.3, 1.4, 1.4, 1.5, 1.6, 1.7, | 1.7, 1.7, 1.7, 1.8, 1.8, 1.9, 1.9, 2.0, 2.4 |

$Q_1 = 1.4$; $Q_2 = 1.7$; $Q_3 = 1.8$

(b) Find the interquartile range.
Answer: $Q_3 - Q_1 = 1.8 - 1.4 = 0.4$.

Exercise 8

An instructor of Statistics collected data from one of her classes in Spring 2016 to investigate the relationship between Study time per week (number of hours) to predict the final grade. For the 8 students in her class the data were as shown in the Table.

Student	Time	Grade
1	14	26
2	25	30
3	15	20
4	5	18
5	10	23
6	12	25
7	5	21
8	21	28

(a) Identify response variable and explanatory variable and their natures.
Answer: Grade is the discrete response variable; Study Time is the continuous explanatory variable.

(b) Investigate correlation and interpret the results.
Answer:

Student	Time	Grade	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
1	14	26	0.62	2.12	1.31	0.38	4.49
2	25	30	11.62	6.12	71.11	135.02	37.45
3	15	20	1.62	-3.88	-6.29	2.62	15.05
4	5	18	-8.38	-5.88	49.27	70.22	34.57
5	10	23	-3.38	-0.88	2.97	11.42	0.77
6	12	25	-1.38	1.12	-1.55	1.90	8.29
7	5	21	-8.38	-2.88	24.13	70.22	8.29
8	21	28	7.62	4.12	31.39	58.06	16.97
					172.34	**349.84**	**118.84**

The correlation coefficient is therefore $r = 172.34/\sqrt{(349.84 * 118.84)} \approx 0.8452$. We detect positive correlation and expect the Grade to grow when study time increases.