Chapter 1

Introduction, descriptive statistics, Python and data visualization

Exercises

Contents

1.1 Infant birth weight

|||| Exercise 1.1 Infant birth weight

In a study of different occupational groups the infant birth weight was recorded for randomly selected babies born by hairdressers, who had their first child. The following table shows the weight in grams (observations specified in sorted order) for 10 female births and 10 male births:

Females (x)	2474	2547	2830	3219	3429	3448	3677	3872	4001	4116
Males (y)	2844	2863	2963	3239	3379	3449	3582	3926	4151	4356

Solve at least the following questions a)–c) first manually and then by Python. It is OK to use Python as an alternative to your pocket calculator for the manual part, but avoid built-in functions that produce the result without forcing you to think about how to compute it during the manual part.

a) What is the sample mean, variance and standard deviation of the female births? Express in your own words the story told by these numbers.

|||| Solution

We have $n = 10$, hence, using the sample mean formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

the sample mean is

$$\begin{aligned}\bar{x} &= \frac{1}{10}(2474 + 2547 + 2830 + 3219 + 3429 + 3448 + 3677 + 3872 + 4001 + 4116) \\ &= 3361.3,\end{aligned}$$

using the sample variance formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

the sample variance is

$$s^2 = \frac{1}{9}\Big[(2474-3361.3)^2 + (2547-3361.3)^2 + (2830-3361.3)^2 + (3219-3361.3)^2 + (3429-3361.3)^2 + (3448-3361.3)^2 + (3677-3361.3)^2 + (3872-3361.3)^2 + (4001-3361.3)^2 + (4116-3361.3)^2\Big] = 344920.5,$$

and using $s = \sqrt{s^2}$, the sample standard deviation is

$$s = \sqrt{344920.5} = 587.30.$$

With Python the answer would look like this:

```
x = np.array([2474, 2547, 2830, 3219, 3429, 3448, 3677,
   3872, 4001, 4116])
print(np.mean(x))
3361.3
print(np.var(x, ddof=1))
344920.5
print(np.std(x, ddof=1))
587.30
```

Interpretation: the sample mean $\hat{\mu} = 3361$ g estimates the population mean weight. Individual births will differ from this value; they are estimated to deviate from the mean by $s = 587$ g on average, so most births are expected to fall within $\bar{x} \pm 2s = [2187, 4535]$ g.

An average absolute difference to the mean (estimated by s) matches that individual observations distribute from roughly $\bar{x} - 2s$ to $\bar{x} + 2s$ (at least for a symmetric, unimodal distribution).

b) Compute the same summary statistics for the male births, and compare the mean and standard deviation to the results for the female births. Based on these ten observations alone, would you say the difference is large?

|||| Solution

Applying the same three formulas to the male data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad s_y = \sqrt{s_y^2},$$

gives

$$\bar{y} = \frac{1}{10}(2844 + 2863 + 2963 + 3239 + 3379 + 3449 + 3582 + 3926 + 4151 + 4356) = 3475.2,$$
$$s_y^2 = 283158,$$
$$s_y = 532.13.$$

With Python the answer would look like this:

```
y = np.array([2844, 2863, 2963, 3239, 3379, 3449, 3582,
    3926, 4151, 4356])
print(np.mean(y))
3475.2
print(np.var(y, ddof=1))
283158
print(np.std(y, ddof=1))
532.13
```

Comparison: male birth weights are on average somewhat higher ($\bar{y} - \bar{x} = 113.9$ g), while the spread (standard deviation) is slightly smaller for males ($s_y = 532$ g vs. $s = 587$ g for females).

An important goal of the course is to provide methods that allow a statistically rigorous comparison of these numbers – accounting for the fact that another sample of 2×10 births could give different results.

c) Find the five quartiles for each sample and draw the two box plots with pen and paper (i.e. not using Python).

|||| Solution

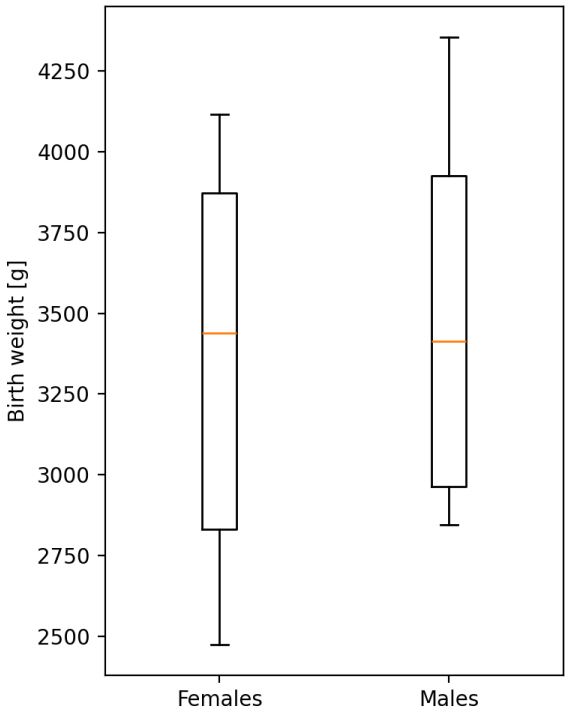
The data are already sorted. With $n = 10$ the relevant values of np are:

	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$
np	0	2.5	5	7.5	10

Since np is non-integer for $p = 0.25, 0.5, 0.75$, we round up and read off the corresponding order statistic:

	$p = 0$	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 1$
Quartile	Min	Q_1	Median	Q_3	Max
Females	2474	2830	$(3429 + 3448)/2 = 3438.5$	3872	4116
Males	2844	2963	$(3379 + 3449)/2 = 3414.0$	3926	4356

These five numbers per group are sufficient to draw the basic box plots by hand:



d) Are there any "extreme" observations in the two samples (use the *modified box plot* definition of extremeness)?

|||| Solution

An observation is extreme if it lies further than $1.5 \cdot \text{IQR}$ from the box. Computing the IQR:

$$\begin{aligned} \text{IQR}_x &= Q_3 - Q_1 = 3872 - 2830 = 1042, \\ \text{IQR}_y &= Q_3 - Q_1 = 3926 - 2963 = 963. \end{aligned}$$

The whisker limits are $[Q_1 - 1.5 \cdot \text{IQR},\ Q_3 + 1.5 \cdot \text{IQR}]$:

$$\begin{aligned} \text{Females:} \quad & [2830 - 1.5(1042),\ 3872 + 1.5(1042)] = [1267,\ 5435], \\ \text{Males:} \quad & [2963 - 1.5(963),\ 3926 + 1.5(963)] = [1518.5,\ 5370.5]. \end{aligned}$$

Comparing to the actual range of each sample (females: 2474 to 4116; males: 2844 to 4356), all observations fall well inside these limits. Hence there are no extreme observations in either sample.

e) What are the coefficients of variation in the two groups?

Solution

The coefficient of variation (CV) expresses the standard deviation relative to the mean:

$$CV_{\text{female}} = \frac{s_x}{\bar{x}} \cdot 100\% = \frac{587.30}{3361.3} \cdot 100\% = 17.5\%,$$
$$CV_{\text{male}} = \frac{s_y}{\bar{y}} \cdot 100\% = \frac{532.13}{3475.2} \cdot 100\% = 15.3\%.$$

With Python the answer would look like this:

```
print(np.std(x, ddof=1) / np.mean(x) * 100)
17.5
print(np.std(y, ddof=1) / np.mean(y) * 100)
15.3
```

Male birth weights are slightly more homogeneous relative to their mean.

1.2 Course grades

Exercise 1.2 Course grades

To compare the difficulty of two different courses at a university the following grade distributions (given as number of pupils who achieved each grade) were registered:

	Course 1	Course 2	Total
Grade 12	20	14	34
Grade 10	14	14	28
Grade 7	16	27	43
Grade 4	20	22	42
Grade 2	12	27	39
Grade 0	16	17	33
Grade -3	10	22	32
Total	108	143	251

a) What is the median of the 251 achieved grades?

Solution

Looking at the Total column, the 251 grades are already ordered (from -3 to 12). The median is the 126th ordered observation from below. Cumulating from below:

Grade	Count	Cumulative (from below)
-3	32	32
0	33	65
2	39	104
4	42	146

The cumulative count reaches 104 after Grade 2, and 146 after Grade 4. Since $104 < 126 \leq 146$, the 126th observation is a grade of **4**. The median is **4**.

This is the *sample median*, which serves as an estimate of the *population median* in exactly the same way that the sample mean estimates the population mean.

b) What are the quartiles and the IQR (Inter Quartile Range)?

|||| Solution

With $n = 251$, the quartile indices are

$$n \cdot 0.25 = 62.75 \quad \Rightarrow \quad Q_1 = x_{(63)},$$
$$n \cdot 0.75 = 188.25 \quad \Rightarrow \quad Q_3 = x_{(189)}.$$

Extending the cumulative table:

Grade	Count	Cumulative (from below)
-3	32	32
0	33	65
2	39	104
4	42	146
7	43	189
10	28	217
12	34	251

The 63rd observation falls in the Grade 0 category (cumulative 65), and the 189th falls at exactly the top of Grade 7 (cumulative 189). Therefore

$$Q_1 = 0,$$
$$Q_3 = 7,$$
$$\text{IQR} = Q_3 - Q_1 = 7 - 0 = 7.$$

Strictly speaking these are the *sample quartiles*, estimating the corresponding *population quartiles*. For simplicity (following convention) we omit the "sample" qualifier.

1.3 Cholesterol

|||| Exercise 1.3 Cholesterol

In a clinical trial of a cholesterol-lowering agent, 15 patients' cholesterol (in mmol L^{-1}) was measured before treatment and 3 weeks after starting treatment. Data are listed below:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Before	9.1	8.0	7.7	10.0	9.6	7.9	9.0	7.1	8.3	9.6	8.2	9.2	7.3	8.5	9.5
After	8.2	6.4	6.6	8.5	8.0	5.8	7.8	7.2	6.7	9.8	7.1	7.7	6.0	6.6	8.4

a) What is the median of the cholesterol measurements for the patients before treatment, and similarly after treatment?

|||| Solution

With Python the answer would look like this:

```
before = np.array([9.1, 8.0, 7.7, 10.0, 9.6, 7.9, 9.0, 7.1,
   8.3, 9.6, 8.2, 9.2, 7.3, 8.5, 9.5])

after = np.array([8.2, 6.4, 6.6, 8.5, 8.0, 5.8, 7.8, 7.2,
   6.7, 9.8, 7.1, 7.7, 6.0, 6.6, 8.4])

print(np.median(before))
8.5

print(np.median(after))
7.2
```

Alternatively, by hand:

With $n = 15$ (odd), the median is the 8th ordered observation $x_{(8)}$. Sorting both samples by hand:

Before (sorted): 7.1, 7.3, 7.7, 7.9, 8.0, 8.2, 8.3, **8.5**, 9.0, 9.1, 9.2, 9.5, 9.6, 9.6, 10.0,

After (sorted): 5.8, 6.0, 6.4, 6.6, 6.6, 6.7, 7.1, **7.2**, 7.7, 7.8, 8.0, 8.2, 8.4, 8.5, 9.8.

Reading off the 8th value in each sorted list:

$$\text{Median before} = 8.5 \text{ mmol L}^{-1},$$
$$\text{Median after} = 7.2 \text{ mmol L}^{-1}.$$

The median cholesterol level decreased by 1.3 mmol L^{-1} after treatment.

b) Find the standard deviations of the cholesterol measurements of the patients before and after treatment.

|||| Solution

With Python the answer would look like this:

```
print(np.std(before, ddof=1))
0.9024

print(np.std(after, ddof=1))
1.0901
```

Alternatively, using the sample mean formula $\bar{x} = \frac{1}{n}\sum x_i$ and sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$:

$$\bar{x}_{\text{before}} = \frac{1}{15}(9.1 + 8.0 + \cdots + 9.5) = 8.600,$$
$$s_{\text{before}} = 0.9024 \text{ mmol L}^{-1},$$
$$\bar{x}_{\text{after}} = \frac{1}{15}(8.2 + 6.4 + \cdots + 8.4) = 7.387,$$
$$s_{\text{after}} = 1.0901 \text{ mmol L}^{-1}.$$

Note that the variability is slightly larger after treatment.

c) Find the sample covariance between the cholesterol measurements before and after treatment.

Solution

With Python the answer would look like this:

```
cov_matrix = np.cov(before, after, ddof=1)

print(cov_matrix[0, 1])
0.7964
```

Alternatively, denoting before-treatment values as x_i and after-treatment values as y_i, the sample covariance is

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{14}\sum_{i=1}^{15}(x_i - 8.600)(y_i - 7.387) = \frac{11.15}{14} = 0.7964. \end{aligned}$$

The positive covariance indicates that patients with higher cholesterol before treatment also tend to have higher cholesterol after treatment.

d) Find the sample correlation between the cholesterol measurements before and after treatment.

Solution

With Python the answer would look like this:

```
print(np.corrcoef(before, after)[0, 1])
0.810
```

Alternatively, the sample correlation coefficient is

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{0.7964}{0.9024 \times 1.0901} = 0.810.$$

The strong positive correlation ($r = 0.81$) confirms that patients with higher initial cholesterol levels also tend to have higher cholesterol after treatment.

e) Compute the 15 differences (Dif = Before − After) and do various summary statistics: sample mean, sample variance, sample standard deviation. Also draw a boxplot.

Solution

With Python the answer would look like this:

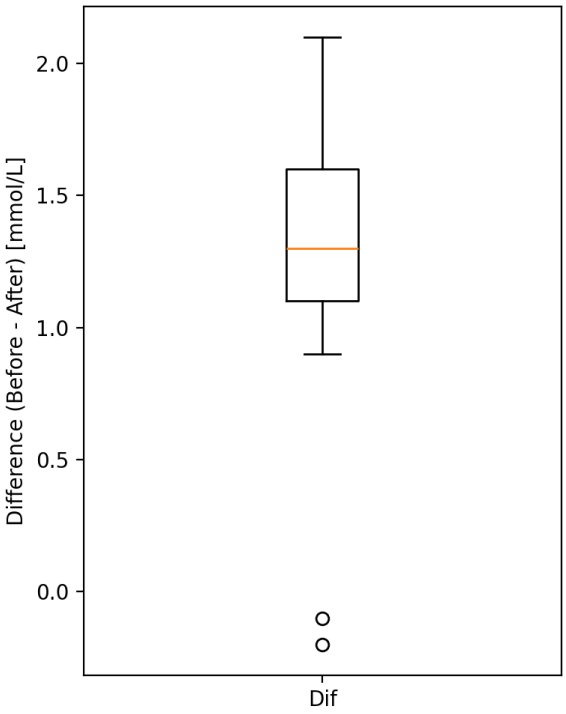
```
dif = before - after

print(np.mean(dif))
1.213

print(np.var(dif, ddof=1))
0.408

print(np.std(dif, ddof=1))
0.639

plt.boxplot(dif)
plt.ylabel("Difference (Before - After) [mmol/L]")
```



Alternatively, the 15 differences are:

Patient	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Dif	0.9	1.6	1.1	1.5	1.6	2.1	1.2	−0.1	1.6	−0.2	1.1	1.5	1.3	1.9	1.1

Summary statistics of the differences:

$$\bar{d} = \frac{1}{15}\sum d_i = \frac{18.2}{15} = 1.213 \text{ mmol L}^{-1},$$
$$s_d^2 = \frac{1}{14}\sum (d_i - \bar{d})^2 = 0.408,$$
$$s_d = \sqrt{0.408} = 0.639 \text{ mmol L}^{-1}.$$

The mean effect (decrease in cholesterol due to treatment) is estimated at 1.21 mmol L^{-1}. Note that two patients (patients 8 and 10) show no beneficial effect (negative differences), making them likely outliers in a boxplot.

f) Considering such data, the key question is whether the average decrease in cholesterol can be "shown statistically". Which summary statistics and/or plots would you examine to get an initial idea of the answer? (The formal answer is given in Chapter 3.)

|||| Solution

The difference-based analysis is the most informative approach. The mean difference $\bar{d} = 1.21$ and standard deviation $s_d = 0.64$ can be compared: the mean is almost twice the standard deviation, suggesting the treatment effect is real – but patients 8 and 10 (with $d = -0.1$ and $d = -0.2$) are exceptions.

We recommend the *paired* (difference) approach over examining the two groups separately. The paired approach removes between-patient variability (e.g. differences in body mass, genetics), leaving only the treatment effect signal. This is reflected in the much smaller standard deviation of the differences ($s_d = 0.64$) compared to the within-group standard deviations ($s_{\text{before}} = 0.90$, $s_{\text{after}} = 1.09$).

1.4 Project start

Exercise 1.4 Project start

a) Go to DTU Learn and read the first project description. Follow the steps to import the data into Python and get started with the exploratory data analysis.

Solution

We provide no solutions for this exercise – you carry it out as a project.

The project descriptions are available at https://02402.compute.dtu.dk/projects.

1.5 Descriptive statistics

|||| Exercise 1.5 Descriptive statistics

The table below shows a sample of 10 measurements. The data is reported in sorted order.

Measurement (x)	256	258	266	296	314	318	326	353	380	391

a) What is the mean of the sample?

|||| Solution

The sample mean is

$$\bar{x} = \frac{1}{10}(256 + 258 + 266 + 296 + 314 + 318 + 326 + 353 + 380 + 391) = \frac{3158}{10} = 315.8.$$

b) Find the minimum, maximum and the quartiles (Q_1, Q_2 and Q_3) of the sample.

|||| Solution

The data are already sorted, so it is easy to read of the minimum (256) and the maximum (391).

Q_2 is the *median*. Since the number of observation is even there is no "middle observation". The median is the average of the two values closest to the middle, i.e., $Q_2 = (314 + 318)/2 = 316$.
Another way to see this is to realize that Q_2 is the 0.50-quantile, $q_{0.50}$, and follow Definition 1.7. We have $p = 0.50$, and $np = 10 \cdot 0.50 = 5$, which is an *integer*. We then compute $Q_2 = q_{0.50} = (x_5 + x_6)/2 = (314 + 318)/2 = 316$.

Q_1 is the 0.25-quantile, $q_{0.25}$. We have $p = 0.25$, and $np = 10 \cdot 0.25 = 2.5$, which is *non-integer*. We then compute $Q_1 = q_{0.25} = x_3 = 266$.

Q_3 is the 0.75-quantile, $q_{0.75}$. We have $p = 0.75$, and $np = 10 \cdot 0.75 = 7.5$, which is *non-integer*. We then compute $Q_3 = q_{0.75} = x_8 = 353$.

Below is an attempt to provide a visualization of the different quantiles:

256	258	266	296	314	318	326	353	380	391
↑		↑		↑			↑		↑
min		Q_1		Q_2			Q_3		*max*

c) What is the sample range and the IQR (Inter Quartile Range)?

|||| Solution

$$\text{Range} = \max - \min = 391 - 256 = 135,$$
$$\text{IQR} = Q_3 - Q_1 = 353 - 266 = 87.$$

d) Compute the 35th percentile of the sample.

|||| Solution

We compute $np = 10 \times 0.35 = 3.5$. Since this is not an integer, we round up to 4, so the 35th percentile is the 4th ordered observation:

$$x_{(4)} = 296.$$

e) Are there any "extreme" values in the sample? (Use the *modified box plot* definition of extremeness.)

Solution

An observation is extreme if it lies more than $1.5 \cdot \text{IQR}$ from the box. The whisker limits are:

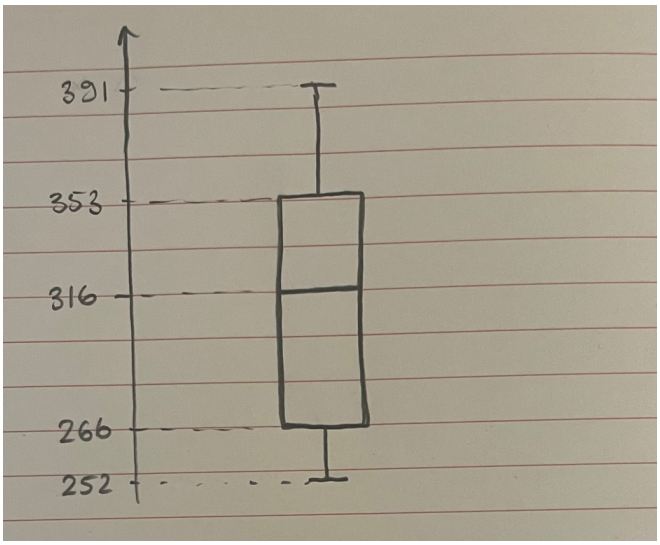
$$Q_1 - 1.5 \cdot \text{IQR} = 266 - 1.5 \times 87 = 266 - 130.5 = 135.5,$$
$$Q_3 + 1.5 \cdot \text{IQR} = 353 + 1.5 \times 87 = 353 + 130.5 = 483.5.$$

All observations lie within $[135.5,\ 483.5]$, so there are no extreme values.

f) Draw a boxplot of the sample (using pen and paper).

Solution

The five-number summary $\{256,\ 266,\ 316,\ 353,\ 391\}$ together with the absence of extremes (whiskers extend to min and max) determines the boxplot fully. No extreme observations are marked.



The table below shows another sample of only 5 measurements (not reported in sorted order).

Measurement (y)	2	5	3	4	3

g) Draw a cumulative distribution plot of the sample (using pen and paper).

Solution

Sorting the sample: 2, 3, 3, 4, 5. The empirical CDF is a step function that jumps by $1/5$ at each observation, so the CDF takes values $0.2, 0.4, 0.6, 0.8, 1.0$ at values $2, 3, 4, 5$ (with a double step of 0.4 at $y = 3$).

1.6 Reading speed

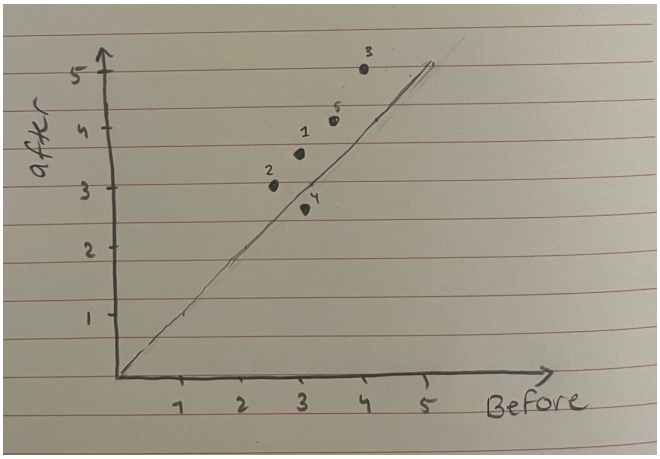
|||| Exercise 1.6 Reading speed

In an elementary school, 5 pupils participated in a short training class to improve reading speed. The number of pages each pupil could read within a certain amount of time was measured before and after the training:

Pupil	1	2	3	4	5
Before training	3	2.5	4	3	3.5
After training	3.5	3	5	2.5	4

a) Draw a scatter plot of the data (using pen and paper). Express in your own words the story told by this plot.

|||| Solution



Plot "After" on the y-axis against "Before" on the x-axis. Most points lie above the diagonal (After $>$ Before), indicating an improvement for four out of five pupils. Pupil 4 shows a slight decrease in reading speed, visible as the one point below the diagonal.

b) What is the mean number of pages read, both before and after the training? By how much did the mean increase/decrease?

Solution

$$\bar{x}_{\text{before}} = \frac{1}{5}(3 + 2.5 + 4 + 3 + 3.5) = \frac{16.0}{5} = 3.2,$$
$$\bar{x}_{\text{after}} = \frac{1}{5}(3.5 + 3 + 5 + 2.5 + 4) = \frac{18.0}{5} = 3.6.$$

The mean improved by 0.4 pages after training.

c) What is the standard deviation of the number of pages read, both before and after the training? How does the spread compare?

Solution

Using $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$ with $n = 5$:

$$s_{\text{before}} = \sqrt{\frac{(-0.2)^2 + (-0.7)^2 + (0.8)^2 + (-0.2)^2 + (0.3)^2}{4}} = \sqrt{\frac{1.30}{4}} = 0.570,$$
$$s_{\text{after}} = \sqrt{\frac{(-0.1)^2 + (-0.6)^2 + (1.4)^2 + (-1.1)^2 + (0.4)^2}{4}} = \sqrt{\frac{3.70}{4}} = 0.962.$$

The spread in reading speed increased after training.

d) What is the CV (coefficient of variation) of the number of pages read, both before and after the training? What does the comparison tell you?

Solution

$$\text{CV}_{\text{before}} = \frac{s_{\text{before}}}{\bar{x}_{\text{before}}} = \frac{0.570}{3.2} = 0.18 \quad (18\%),$$
$$\text{CV}_{\text{after}} = \frac{s_{\text{after}}}{\bar{x}_{\text{after}}} = \frac{0.962}{3.6} = 0.27 \quad (27\%).$$

While the mean improved, variability relative to the mean increased substantially after training.

e) Express in your own words the story told by the numbers calculated above.

Solution

On average, pupils read 0.4 more pages per unit time after training ($\bar{x}$ increased from 3.2 to 3.6). However, the standard deviation nearly doubled (from 0.57 to 0.96), meaning the effect of the training was very uneven across pupils: some improved considerably while at least one pupil (pupil 4) actually performed worse. The CV confirms this: relative variability rose from 18% to 27%.

f) Find the sample covariance between observations before and after training. What does the sign of this number tell you about the relationship between pupils' reading speed before and after training?

Solution

Using Definition 1.18 with $n = 5$, $\bar{x}_{\text{before}} = 3.2$, $\bar{x}_{\text{after}} = 3.6$:

$$\begin{aligned} s_{xy} &= \frac{1}{4}\sum_{i=1}^{5}(x_{\text{before},i} - 3.2)(x_{\text{after},i} - 3.6) \\ &= \frac{1}{4}\big[(-0.2)(-0.1) + (-0.7)(-0.6) + (0.8)(1.4) + (-0.2)(-1.1) + (0.3)(0.4)\big] \\ &= \frac{0.02 + 0.42 + 1.12 + 0.22 + 0.12}{4} = \frac{1.90}{4} = 0.475. \end{aligned}$$

The positive covariance shows that pupils who read more pages before training also tended to read more pages after training.

g) Find the sample correlation between observations before and after training. Express in your own words how you would interpret this result. Does the result fit with the plot you made in question (a)?

Solution

$$r = \frac{s_{xy}}{s_{\text{before}} \cdot s_{\text{after}}} = \frac{0.475}{0.570 \times 0.962} = \frac{0.475}{0.548} = 0.87.$$

The strong positive correlation ($r = 0.87$) conveys the same qualitative information as the covariance: pupils who read more pages before the training also tended to read more pages after the training. However, because the correlation coefficient is always between -1 and 1, we can conclude that the association between the two measurements is quite strong. This result is consistent with the scatter plot: points follow a positive upward trend with little scatter around it.

h) Calculate the 5 differences (Dif = After − Before) to obtain the achieved "effect" for each pupil. Calculate the mean of these differences. Express in your own words how you could use these differences (instead of the raw data) to analyse the effect of the training class on reading speed.

Solution

The 5 differences are:

Pupil	1	2	3	4	5
Before training	3	2.5	4	3	3.5
After training	3.5	3	5	2.5	4
Dif	0.5	0.5	1	−0.5	0.5

The mean difference is:

$$\bar{d} = \frac{0.5 + 0.5 + 1.0 + (-0.5) + 0.5}{5} = \frac{2.0}{5} = 0.4.$$

By reducing the data to differences, we remove between-pupil variability (e.g. natural reading ability), leaving only the training effect. This paired approach focuses directly on the question of interest: did each individual improve? The average improvement is 0.4 pages.

i) Observing such data, the big question is whether an average increase in reading speed can be "shown statistically". How to formally answer this question is presented later in the course, but consider now how you would answer such a question.

Solution

An informal assessment: the average difference is $\bar{d} = 0.4$ and four out of five pupils improved. However, the standard deviation of the differences is substantial relative to the mean, and the sample is very small ($n = 5$). A formal test (paired t-test, Chapter 3) would quantify whether $\bar{d} = 0.4$ is large enough relative to its uncertainty to conclude a real effect.