1 Descriptive Statistics

A *data set* considered in statistics typically consists of a collection of values

$$x_i, \qquad i = 1, \ldots, N$$

for some $N \in \mathbb{N}$. The x_i's need not to be pairwise different values. (E.g. consider the results of rolling a die for $N = 60$ times.) Therefore, such data sets can not simply be described by sets in a mathematically strict sense, but are better described by a vector of length N or a multiset with underlying set $\{x_i | i = 1, \ldots, N\}$.

(1.1) Example. The results of throwing a die for 60 times have been recorded and are given in the following list:

$$\begin{array}{lll} 3,5,1,2,1,1,2,3,3,4, & 3,5,2,6,1,5,3,4,1,2, & 5,6,2,5,6,6,1,1,2,6, \\ 1,3,6,4,5,2,5,6,1,5, & 6,5,2,5,1,3,6,2,2,1, & 1,5,2,2,3,1,4,1,4,5 \end{array}$$

Such a data set can also be easily plotted, see Figure 1. If the order of the values does not

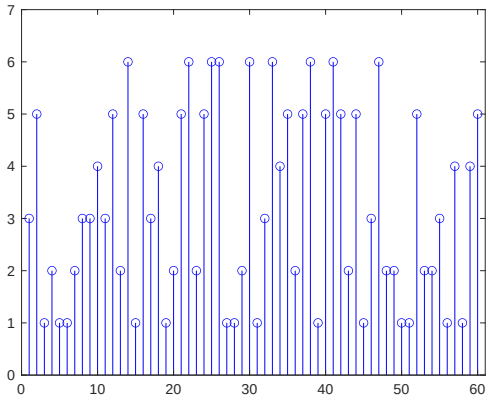

Figure 1: Rolling a die 60 times – Data set

matter, it makes sense to simply count the number of occurrences of all values. Such an aggregation of the data set gives rise to the display of the data set as a *frequency table* (see Table 1) or a *frequency plot* (see Figure 2 and Figure 3).

value	frequency	relative frequency
1	14	0.233
2	12	0.2
3	8	0.133
4	5	0.083
5	12	0.2
6	9	0.15

Table 1: Frequency table

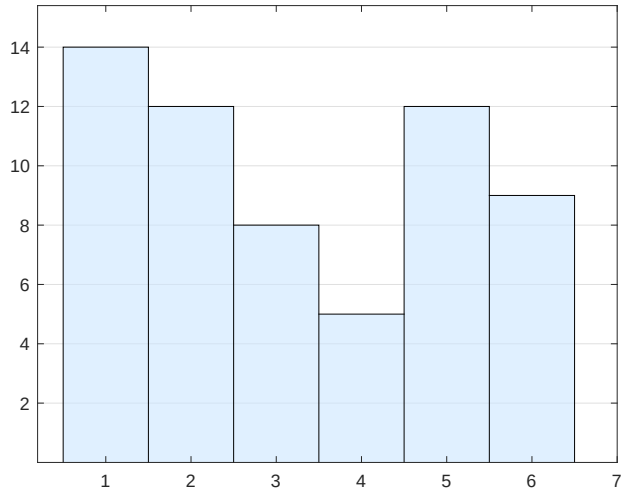

Figure 2: Rolling a die 60 times – Frequency plot

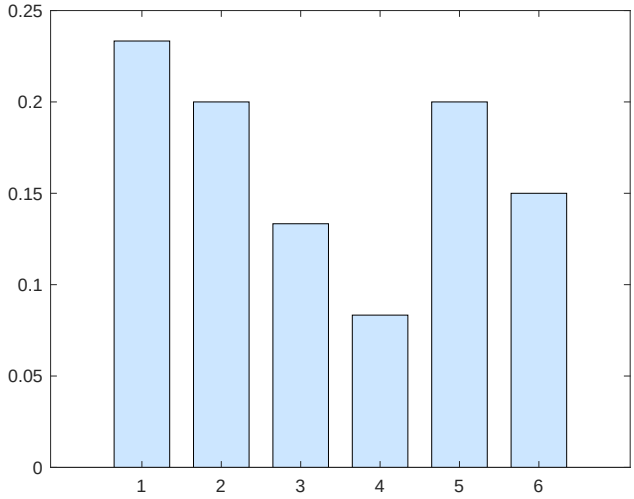

Figure 3: Rolling a die 60 times – Relative frequency plot

(1.2) Definition. Given a data set $x_1, x_2, \ldots, x_N$, the values that occur with the greatest frequency are called *modal values*. If there is only one such modal value, it is also called the *sample mode*.

(1.3) Example. The data set in (1.1) has sample mode 1.

(1.4) Example. The lifetimes of 1000 incandescent lamps have been determined and the data set of this measurement can be seen in Figure 4. Although 1000 different values have been recorded, there is no single pair of equal values. As a consequence a frequency table or frequency plot (see Figure 5) is not particularly revealing. Instead, by dividing the range of measured values into intervals and counting the distribution of the values from the data set with respect to these intervals, a *histogram* can be drawn, which might be quite informative, see Figure 6.

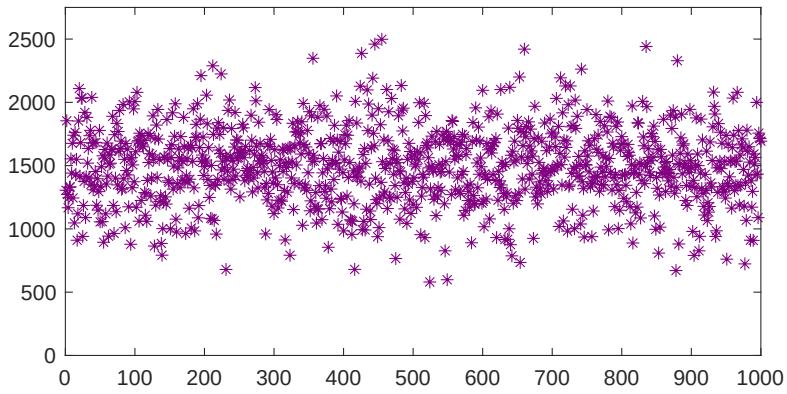

Figure 4: Lifetime of 1000 incandescent lamps – Data set

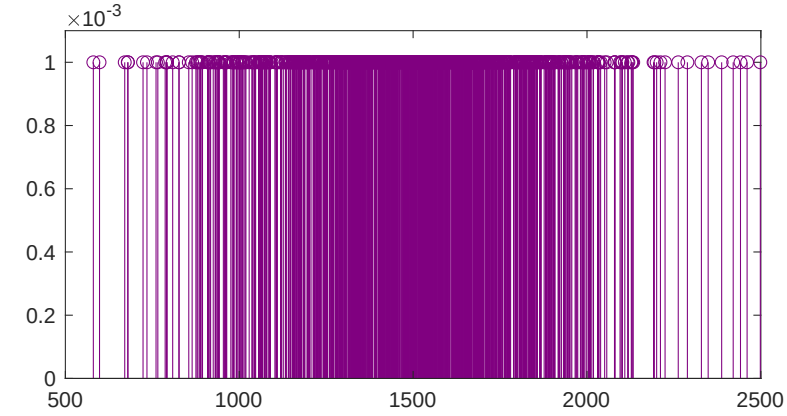

Figure 5: Lifetime of 1000 incandescent lamps – Frequency plot

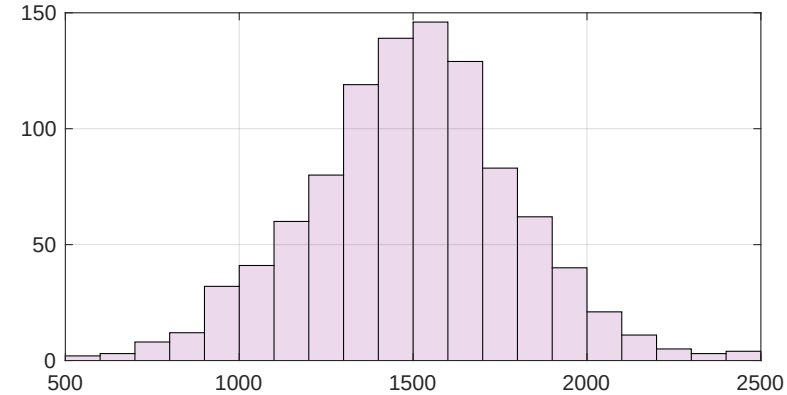

Figure 6: Lifetime of 1000 incandescent lamps – Histogram

(1.5) MatLab. A histogram can simply be plotted with MatLab's `histogram()` function. For example, Figure 6 has been generated by calling:

```
histogram(data)
```

The `histogram()` function determines a decomposition of a suitable part of the real line into subintervals (bins) and counts for every subinterval the number of elements of `data` it contains. You may also explicitly define the subintervals by supplying the edge values to a call of the `histogram()` function. E. g., calling `histogram(data)` in our example is equivalent to:

```
histogram(data, 500:100:2500)
```

MatLab's `histogram()` function follows the convention of including the left end, so in our example the subintervals (bins) look like this:

$$[500, 600),\ [600, 700), \ldots,\ [2300, 2400),\ [2400, 2500]$$

Note: The last interval also includes the right edge.

(1.6) Definition. A *cumulative frequency plot* of a data set $x_1, x_2, \ldots, x_N$ is a plot of the graph of $f : \mathbb{R} \to \mathbb{N}$ defined by:

$$f(x) \quad = \quad |\{i \mid x_i \leq x\}|$$

(1.7) Example. Cumulative frequency plots for the data sets given in Examples (1.1) and (1.4) are displayed in Figure 7 and Figure 8, respectively.

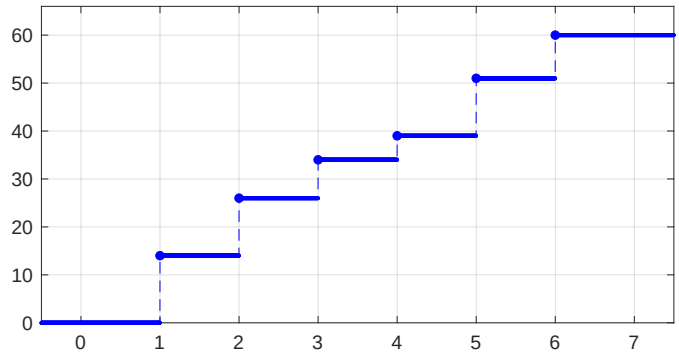

Figure 7: Rolling a die 60 times – Cumulative frequency plot

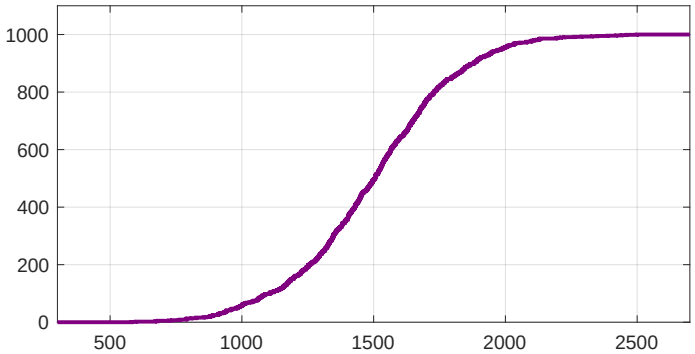

Figure 8: Lifetime of 1000 incandescent lamps – Cumulative frequency plot

(1.8) Definition. Given an ordered data set

$$x_1 \leq x_2 \leq \cdots \leq x_N$$

the *quantile* function

$$x : [0,1] \to \mathbb{R}$$

is defined as follows:

$$x(p) := \begin{cases} x_1 & \text{if } p = 0 \\ x_{\lceil p \cdot N \rceil} & \text{if } p > 0 \text{ and } p \cdot N \notin \{1, 2, \cdots, N-1\} \\ \frac{x_{(p \cdot N)} + x_{(p \cdot N+1)}}{2} & \text{if } p \cdot N \in \{1, 2, \cdots, N-1\} \end{cases}$$

Alternatively $x(p)$ is called the $(100\,p)$*th percentile* of the ordered data set.

(1.9) Remark. If $p \cdot N \notin \{1, 2, \cdots, N-1\}$, then $x(p)$ is the unique value from the data set, such that at least $100p$ percent of the x_i's are less than or equal to $x(p)$ and at least $100(1-p)$ percent of the x_i's are greater than or equal to $x(p)$.

If $p \cdot N \in \{1, 2, \cdots, N-1\}$, there are at most two such values from the data set and $x(p)$ is defined to be the mean of these values.

(1.10) Example. For the data sets given in Examples (1.1) and (1.4), the quantile functions are shown in Figure 9 and Figure 10, respectively. Note, that these graphs can essential be obtained from a reflection along the line $y = x$ of the cumulative relative frequency graphs (cf. with Figure 7 and Figure 8).

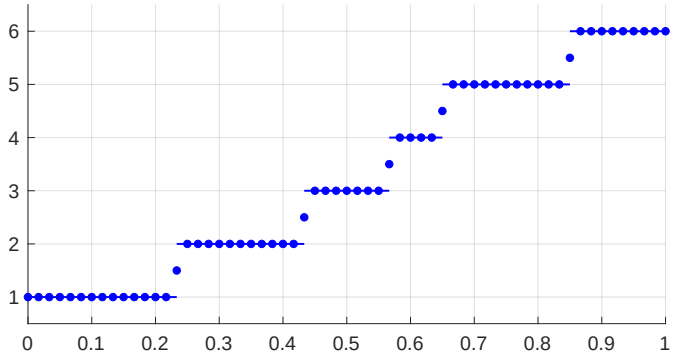

Figure 9: Rolling a die 60 times – Quantiles

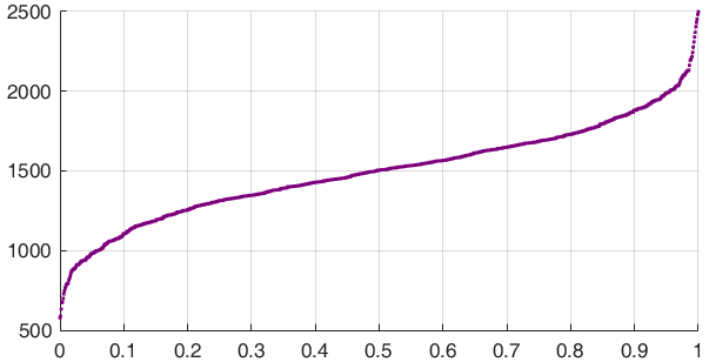

Figure 10: Lifetime of 1000 incandescent lamps – Quantiles

(1.11) Definition. Let $x(p)$ be the quantile function of an ordered data set. Then:

(i) $x(0.25)$ is called the *first quartile*,

(ii) $x(0.5)$ is called the *median* (or *second quartile*), and

(iii) $x(0.75)$ is called the *third quartile* of the data set.

(1.12) Example. For the data set given in Example (1.1) the first quartile, median, and third quartile are respectively: 2, 3 and 5.

For the data set given in Example (1.4) the first quartile, median, and third quartile are:

$$\begin{aligned} x(0.25) &= 1314.28 \\ x(0.5) &= 1505.05 \\ x(0.75) &= 1685.90 \end{aligned}$$

(1.13) Notation (Box plot). The range of values of a data set and the location of its quartiles are often visualized with a so-called *box plot*.

(1.14) Example. A box plot for the data set given in Example (1.4) is shown in Figure 11.

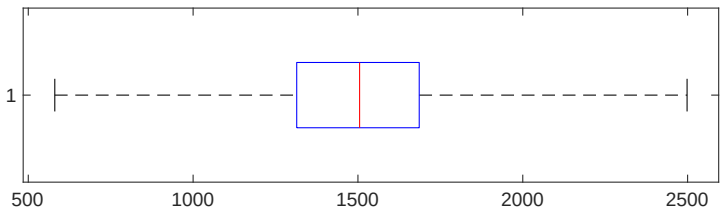

Figure 11: Lifetime of 1000 incandescent lamps – Box plot

(1.15) Remark. The definition of a quantile function is not at all unique. In a paper of Hyndman and Fan (Hyndman, R. J., Fan, Y. (1996) Sample quantiles in statistical packages. American Statistician, 50, 361-365) nine different quantile functions $\hat{Q}_1, \ldots, \hat{Q}_9$ are explicitly discussed. Three functions are discontinuous step functions and six functions are continuous. The definition of $x(p)$ in (1.8) is equal to the definition of $\hat{Q}_2(p)$.

MatLab's `quantile` function corresponds to $\hat{Q}_5$, whereas the CAS Maxima implements a `quantile` function in its package `descriptive` that corresponds to $\hat{Q}_7$. Figure 12 shows the graphs of the above mentioned functions $\hat{Q}_2(p)$, $\hat{Q}_5(p)$ and $\hat{Q}_7(p)$ for a small ordered data set with values:

$$x_1 = 1, \quad x_2 = x_3 = 3, \quad x_4 = 4, \quad x_5 = 5.5, \quad x_6 = 6.5$$

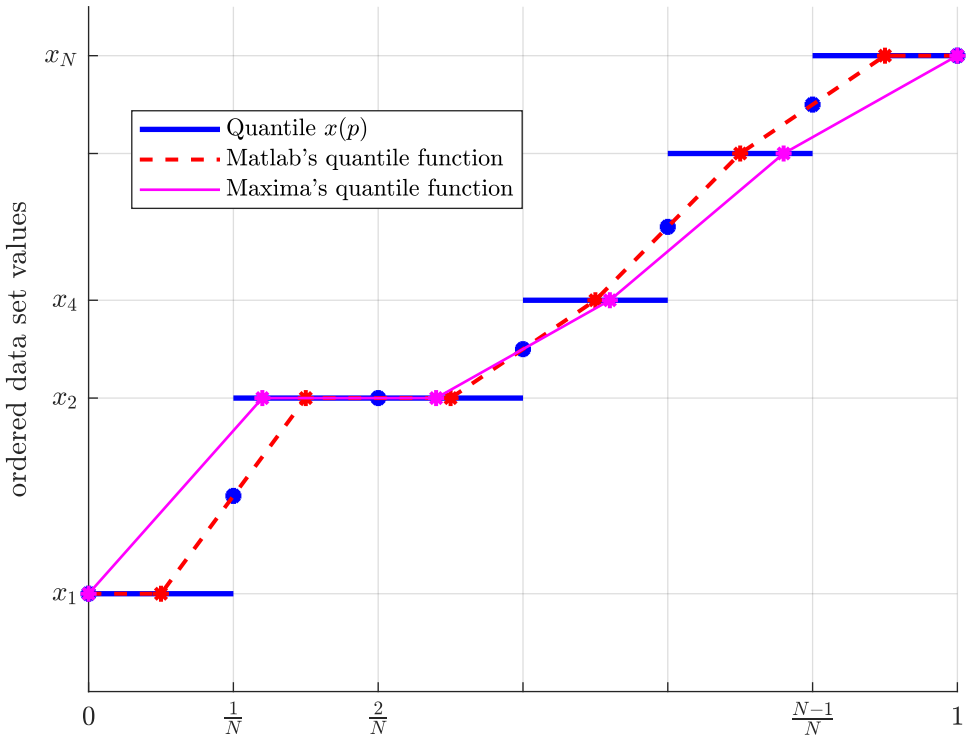

Figure 12: Quantile functions $\hat{Q}_2(p)$, $\hat{Q}_5(p)$ and $\hat{Q}_7(p)$ for a data set with $N = 6$ values

(1.16) Definition. For a given data set $x_1, x_2, \ldots, x_N$ the *mean* $\overline{x}$, the *variance* s^2, and the *standard deviation* s are defined as follows:

(i) $\overline{x} \;:=\; \dfrac{1}{N}\sum_{i=1}^{N} x_i$

(ii) $s^2 \;:=\; \dfrac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2 \;=\; \dfrac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right)$

(iii) $s \;:=\; \sqrt{\dfrac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2}$

(1.17) Example. For the data set given in Example (1.1) the mean, variance, and standard deviation are respectively:

(i) $\overline{x} \;=\; \frac{196}{60} \;\approx\; 3.267,$

(ii) $s^2 \;=\; \frac{(14+12\cdot 2^2+8\cdot 3^2+5\cdot 4^2+12\cdot 5^2+9\cdot 6^2) \,-\, 60\cdot\left(\frac{196}{60}\right)^2}{59} \;\approx\; 3.351$

(iii) $s \;=\; \sqrt{s^2} \;\approx\; 1.831$

(1.18) Example. For the data set given in Example (1.4) the mean, variance, and standard deviation are respectively:

$$\overline{x} \;\approx\; 1496.02, \quad s^2 \;\approx\; 89726.6, \quad s \;\approx\; 299.54$$

(1.19) Lemma. Let $x_1, x_2, \ldots, x_N$ be a data set and

$$y_i \;=\; a \,+\, b\,x_i \qquad \text{for } i = 1, \ldots, N$$

for some constant values a and b. If $\overline{x}$ and $\overline{y}$ denote the mean values and ${s_x}^2$ and ${s_y}^2$ denote the variances of the x_i's and y_i's respectively, than:

(i) $\overline{y} \;=\; a + b\,\overline{x}$

(ii) ${s_y}^2 \;=\; b^2\,{s_x}^2$

(1.20) Theorem (Chebyshev's inequality). Let $\overline{x}$ be the mean and s the standard deviation of the data set $x_1, x_2, \ldots, x_N$. Then the following inequalities hold:

(i)
$$\frac{|\{i \mid |x_i - \overline{x}| < k \cdot s\}|}{N} \;\geq\; 1 - \frac{N-1}{N\,k^2} \;>\; 1 - \frac{1}{k^2} \qquad \text{for all } \; k \geq 1$$

(ii)
$$\frac{|\{i \mid |x_i - \overline{x}| \geq k \cdot s\}|}{N} \;\leq\; \frac{N-1}{N\,k^2} \;<\; \frac{1}{k^2} \qquad \text{for all } \; k \geq 1$$

(iii)
$$\frac{|\{i \mid (x_i - \overline{x}) \geq k \cdot s\}|}{N} \;<\; \frac{1}{1+k^2} \qquad \text{for all } \; k > 0$$

Proof:

(i), (ii): Let

$$I_k \;:=\; \{i \mid |x_i - \overline{x}| < k \cdot s\} \qquad \text{and} \qquad O_k \;:=\; \{i \mid |x_i - \overline{x}| \geq k \cdot s\}\,.$$

Then

$$(N-1) \cdot s^2 \;=\; \sum_{i=1}^{N} (x_i - \overline{x})^2 \;\geq\; \sum_{i \in O_k} (x_i - \overline{x})^2 \;\geq\; \sum_{i \in O_k} (k \cdot s)^2 \;=\; |O_k| \cdot k^2 \cdot s^2$$

and therefore:

$$|O_k| \;\leq\; \frac{N-1}{k^2} \qquad \text{and} \qquad |I_k| \;=\; N - |O_k| \;\geq\; N - \frac{N-1}{k^2}$$

Hence:

$$\frac{|I_k|}{N} \;\geq\; 1 - \frac{N-1}{N\,k^2} \;>\; 1 - \frac{1}{Nk^2} \qquad \text{and} \qquad \frac{|O_k|}{N} \;\leq\; \frac{N-1}{N\,k^2} \;<\; \frac{1}{k^2}$$

(iii): Let

$$d_i \;:=\; x_i - \overline{x} \;\text{ for }\; i = 1, \ldots, N \qquad \text{and} \qquad R_k \;:=\; \{i \mid d_i \geq k \cdot s\}\,.$$

Then, for every $b > 0$:

$$\sum_{i=1}^{N} (d_i + b)^2 \;\geq\; \sum_{i \in R_k} (d_i + b)^2 \;\geq\; \sum_{i \in R_k} (k \cdot s + b)^2 \;=\; |R_k|(k \cdot s + b)^2$$

Together with

$$\sum_{i=1}^{N} (d_i + b)^2 \;=\; \sum_{i=1}^{N} d_i^2 + 2b \sum_{i=1}^{N} d_i + Nb^2 \;=\; (N-1)s^2 + 0 + Nb^2$$

it follows that:

$$\frac{|R_k|}{N} \;\leq\; \frac{1}{N} \cdot \frac{(N-1)s^2 + Nb^2}{(k \cdot s + b)^2} \;<\; \frac{s^2 + b^2}{(k \cdot s + b)^2}$$

In particular, setting $b = \frac{s}{k}$:

$$\frac{|R_k|}{N} \;<\; \frac{s^2 + \frac{s^2}{k^2}}{(k \cdot s + \frac{s}{k})^2} \;=\; \frac{1 + \frac{1}{k^2}}{(k + \frac{1}{k})^2} \;=\; \frac{k^2 + 1}{(k^2 + 1)^2} \;=\; \frac{1}{k^2 + 1}$$

Corrolated data sets

(1.21) Example. Twenty students attended the lectures on probability and statistics. Points for their homework and their exam grades are listed in the following table.

Student No.	1	2	3	4	5	6	7	8	9	10
Homework Pts.	51	120	123	127	103	137	163	115	123	132
Exam Grade	5.0	3.3	3.0	3.3	3.7	2.3	1.7	2.7	2.7	2.7

Student No.	11	12	13	14	15	16	17	18	19	20
Homework Pts.	160	136	0	87	182	173	55	181	186	0
Exam Grade	2.3	3.3	5.0	3.3	1.0	2.3	4.0	1.7	1.7	5.0

The two values corresponding to a single student can be visualized as a point in the two dimensional space. All these points are shown in a so-called scatter diagram (see Figure 13). (Note that only 19 points are visible, as multiple occurrences of the same point can not be distinguished from single occurrences in this simple scatter diagram.) Furthermore, the line that fits the data sets best (see Lemma (1.22)) is shown in Figure 13.

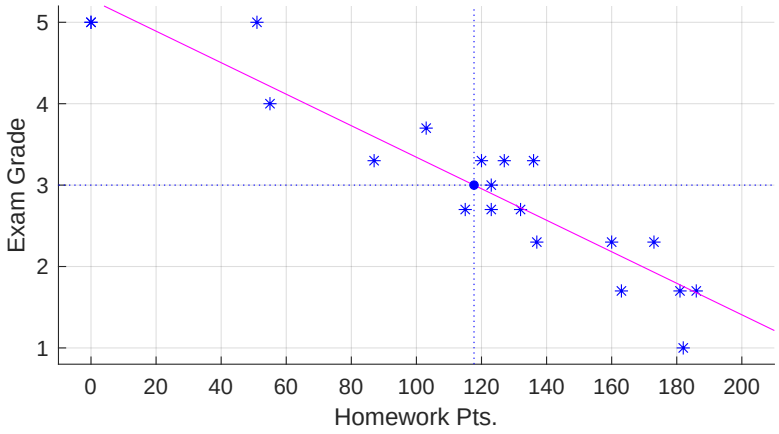

Figure 13: Scatter diagram for the data sets from Example (1.21) and the line of best fit

(1.22) Lemma. For given non-constant data sets $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ the *line of best fit*

$$y \;=\; a \,+\, b \cdot x$$

is defined by the *least squares fitting* property, i.e. by minimizing the sum:

$$\sum_{i=1}^{N} (y_i - (a + b \cdot x_i))^2$$

The following holds:

$$\text{(i)} \quad b \;=\; \frac{\sum_{i=1}^{N} x_i y_i \;-\; N\overline{x}\,\overline{y}}{\sum_{i=1}^{N} x_i^2 \;-\; N\overline{x}^2} \;=\; \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{(N-1)\,{s_x}^2}$$

$$\text{(ii)} \quad a \;=\; \overline{y} \;-\; b\overline{x}$$

Proof: If $y = a + b \cdot x$ is the line of best fit and

$$d \;:=\; \sum_{i=1}^{N}(y_i \;-\; (a + b \cdot x_i))^2$$

then:

$$\frac{\partial\, d}{\partial\, a} \;=\; \frac{\partial\, d}{\partial\, b} \;=\; 0$$

(ii):

$$0 \;=\; \frac{\partial\, d}{\partial\, a} \;=\; -2\sum_{i=1}^{N}(y_i - (a + b \cdot x_i)) \;=\; -2\,(N\overline{y} - Na - bN\overline{x})$$

$$\Longrightarrow \qquad 0 \;=\; \overline{y} - a - b\overline{x} \qquad \Longrightarrow \qquad a \;=\; \overline{y} - b\overline{x}$$

(i):

$$0 \;=\; \frac{\partial\, d}{\partial\, b} \;=\; -2\sum_{i=1}^{N} x_i\,(y_i - (a + b \cdot x_i)) \;=\; -2\left(\sum_{i=1}^{N} x_i y_i - Na\overline{x} - b\sum_{i=1}^{N} x_i^2\right)$$

$$\begin{aligned}\Longrightarrow \quad 0 \;&=\; \sum_{i=1}^{N} x_i y_i - N(\overline{y} - b\overline{x})\overline{x} - b\sum_{i=1}^{N} x_i^2 \\ &=\; \sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y} - b\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right)\end{aligned}$$

$$\Longrightarrow \quad b \;=\; \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} \;=\; \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{(N-1)\,{s_x}^2}$$

The last equality follows from (1.16)(ii) and:

$$\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y}) \;=\; \sum_{i=1}^{N} x_i y_i \;-\; \overline{x}\sum_{i=1}^{N} y_i \;-\; \overline{y}\sum_{i=1}^{N} x_i \;+\; N\overline{x}\,\overline{y} \;=\; \sum_{i=1}^{N} x_i y_i \;-\; N\overline{x}\,\overline{y}$$

(1.23) Definition. The normalized form of a non-constant data set $x = (x_1, x_2, \ldots, x_N)$ is defined by:

$$x^o \;:=\; ({x_1}^o, {x_2}^o, \ldots, {x_N}^o), \qquad {x_i}^o \;:=\; \frac{x_i - \overline{x}}{s_x}$$

Note: $\overline{x^o} = 0$ and $s_{x^o} = 1$.

(1.24) Definition. Given non-constant data sets $x = (x_1, x_2, \ldots, x_N)$ with mean $\overline{x}$ and $y = (y_1, y_2, \ldots, y_N)$ with mean $\overline{y}$, the *correlation coefficient* is defined by:

$$r \;:=\; r_{x,y} \;:=\; \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}} \;=\; \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y} \;=\; \frac{\sum_{i=1}^{N} {x_i}^o \cdot {y_i}^o}{N-1}$$

Note: $r_{x^o,y^o} = r_{x,y}$

(1.25) Lemma. Let x and y be non-constant data sets with correlation coefficient $r = r_{x,y}$. Then the line of best fit for the normalized data sets x^o and y^o is given by:

$$y \;=\; r \cdot x$$

Proof. By Lemma (1.22) the slope of the line of best fit for the normalized data sets and x^o and y^o is given by:

$$\frac{\sum_{i=1}^{N}({x_i}^o - \overline{x^o})({y_i}^o - \overline{y^o})}{(N-1)s_{x^o}^2} \;=\; \frac{\sum_{i=1}^{N} {x_i}^o \, {y_i}^o}{(N-1)} \;=\; r$$

(1.26) Lemma. Let $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ as in (1.24) and r the correlation coefficient of x and y. Then, the slope of the line of best fit for the data sets x and y is given by:

$$b \;=\; r \cdot \frac{s_y}{s_x}$$

(1.27) Lemma. Let $x = (x_1, x_2, \ldots, x_N)$ be a data set and

$$\xi_i \;=\; \alpha \,+\, \beta\, x_i \qquad \text{for } i = 1, \ldots, N$$

for some constant values α and β. Then:

$${\xi_i}^o \;=\; \operatorname{sgn}(\beta) \cdot {x_i}^o \qquad \text{for } i = 1, \ldots, N$$

(1.28) Lemma. Let $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ be non-constant data sets with mean $\overline{x}$ and $\overline{y}$, respectively. Moreover, let $r = r_{x,y}$ be the correlation coefficient of x and y. Then:

(i) $r \in [-1, 1]$

(ii) $|r| = 1$ iff and only if there exist constants a, b, such that

$$y_i = a + bx_i \qquad \text{for all } i \in \{1, \ldots, N\} .$$

(In this case $r = \text{sgn}(b)$.)

(iii) Let a, b, c and d constant values with $b \cdot d > 0$. Set

$$\xi_i = a + b \cdot x_i, \qquad \eta_i = c + d \cdot y_i$$

for $i = 1, 2, \ldots, N$. Then $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ and $\eta = (\eta_1, \eta_2, \ldots, \eta_N)$ have the same correlation coefficient as x and y:

$$r \;=\; r_{x,y} \;=\; r_{\xi,\eta}$$

Proof:

(i): Consider the vectors

$$\mathbf{v} = (x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x})$$

and

$$\mathbf{w} = (y_1 - \overline{y}, y_2 - \overline{y}, \ldots, y_N - \overline{y}) .$$

If α denotes the angle between $\mathbf{v}$ and $\mathbf{w}$, then:

$$r \;=\; \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|} \;=\; \cos(\alpha) \;\in\; [-1, 1]$$

(ii): $|r| = 1$, if and only if $\mathbf{v}$ and $\mathbf{w}$ are collinear, i.e. if and only if there exists some $b \in \mathbb{R}$ such that:

$$\begin{aligned} & & y_i - \overline{y} &= b(x_i - \overline{x}) && \text{for all } i = 1, 2, \ldots, N \\ \iff & & y_i &= (\overline{y} - b\overline{x}) + b\, x_i && \text{for all } i = 1, 2, \ldots, N \\ \iff & \text{for some } a: & y_i &= a + b\, x_i && \text{for all } i = 1, 2, \ldots, N \end{aligned}$$

For the last equivalence note, that $y_i = a + b\, x_i$ for all i implies:

$$a \;=\; y_i - b\, x_i \text{ for all } i \qquad \Longrightarrow \qquad a \;=\; \frac{1}{N} \sum_{i=1}^{N} (y_i - b\, x_i) \;=\; \overline{y} - b\overline{x}$$

(iii): Lemma (1.27) gives

$$\xi_i{}^o \;=\; \text{sgn}(b) \cdot x_i{}^o \quad \text{ and } \quad \eta_i{}^o \;=\; \text{sgn}(d) \cdot y_i{}^o \qquad \text{for } i = 1, \ldots, N$$

and therefore:

$$r_{\xi,\eta} \;=\; \frac{\sum_{i=1}^{N} \xi_i{}^o \, \eta_i{}^o}{(N-1)} \;=\; \frac{\text{sgn}(b) \cdot \text{sgn}(d) \cdot \sum_{i=1}^{N} x_i{}^o \, y_i{}^o}{(N-1)} \;=\; r_{x,y}$$