Probability and Statistics

1 – Descriptive Statistics

Stefan Heiss

Technische Hochschule Ostwestfalen-Lippe
Dep. of Electrical Engineering and Computer Science

October 10, 2023

Chebyshev's inequalities

Theorem (1.20)

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:

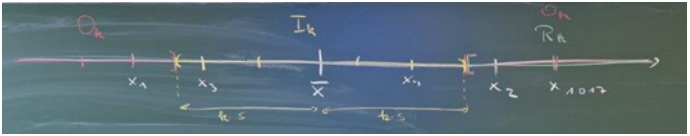
(i)
$$\frac{|I_k|}{N} = \frac{|\{i \mid |x_i - \overline{x}| < k \cdot s\}|}{N} \geq 1 - \frac{N-1}{N\,k^2} > 1 - \frac{1}{k^2} \qquad \textit{for all } k \geq 1$$

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

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

Proof of Chebyshev's inequalities

(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

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

and therefore:

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

Proof of Chebyshev's inequalities

(iii): Let

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

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

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

$$\sum_{i=1}^{N} (x_i - \bar{x}) = \sum_{i=1}^{N} x_i - N\bar{x} = N \cdot \left(\frac{1}{N} \sum_{i=1}^{N} x_i - \bar{x} \right)$$
$$= N(\bar{x} - \bar{x}) = 0$$

Proof of Chebyshev's inequalities

(iii): Let

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

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

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

it follows that:

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

Proof of Chebyshev's inequalities

(iii)

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

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

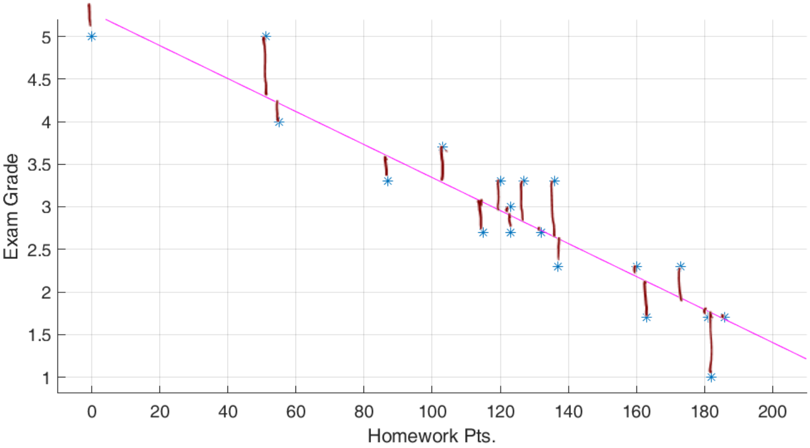
Example (1.21)

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

Scatter Diagram



Line of Best Fit

Definition

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:

(x_i, y_i)

$$\sum_{i=1}^{N} (y_i - \underbrace{(a + b \cdot x_i)}_{\text{value of "line" at } x_i})^2$$

Lemma (1.22)

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

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

are given by:

(i)

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

(ii)

$$a = \overline{y} - b\overline{x}$$

$\Rightarrow$ $(\bar{x}, \bar{y})$ is a point on the line of best fit

Proof of Lemma (1.22)

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):

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

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

Proof of Lemma (1.22)

(i):

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

$$0 = \sum_{i=1}^{N} x_i y_i - N\overline{x}a - b\sum_{i=1}^{N} x_i^2 \overset{(ii)}{=} \sum_{i=1}^{N} x_i y_i - N\overline{x}(\overline{y} - b\overline{x}) - b\sum_{i=1}^{N} x_i^2$$

$$0 = \sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y} - b\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right) \implies b = \frac{\displaystyle\sum_{i=1}^{N} x_i y_i - N\overline{x}\,\overline{y}}{\displaystyle\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}$$

Proof of Lemma (1.22)

(i):

$$\frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{N}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{N}x_iy_i \;-\; \overline{x}\overbrace{\sum_{i=1}^{N}y_i}^{N\cdot\bar y} \;-\; \overline{y}\overbrace{\sum_{i=1}^{N}x_i}^{N\cdot\bar x} \;+\; N\overline{x}\,\overline{y}}{(N-1)\,{s_x}^2} = \frac{\sum_{i=1}^{N}x_iy_i - N\overline{x}\,\overline{y}}{(N-1)\,{s_x}^2} = b$$

$$b = \frac{\sum_{i=1} x_i y_i - N\bar x \bar y}{\sum_{i=1}^{N} x_i^2 - N\bar x^2} = \text{proof as exercise}$$

Normalized Data Set

Definition (1.23)

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 $s_{x^o} = 1$.

Correlation Coefficient

Definition (1.24)

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:

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

Correlation Coefficient

Definition (1.24)

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:

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

$$\Big\updownarrow =$$

$$\frac{\sum ({x_i}^o)^o ({y_i}^o)^o}{N-1}$$

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

Lemma (1.25)

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:

$$y = r \cdot x$$

Lemma 1.22

$y = a + bx$

$a = \bar{y} - b\bar{x}$

$\underbrace{\overline{y^o}}_{0} - b\,\underbrace{\overline{x^o}}_{0}$

$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{(N-1)\,s_x^2}$

$$r = \frac{\sum_{i=1}^{N}(x_i{}^o - \overset{0}{\overset{\shortparallel}{\overline{x^o}}})(y_i{}^o - \overset{0}{\overset{\shortparallel}{\overline{y^o}}})}{(N-1)\underset{1}{\underset{\shortparallel}{s_{x^o}^2}}} = \frac{\sum_{i=1}^{N} x_i{}^o\, y_i{}^o}{(N-1)} = r$$

Lemma (1.26)

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 mean* $\overline{y}$*, respectively. Furthermore let* r *be 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:*

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

$$r \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y} \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x^2} = b$$