Probability and Statistics

1 – Descriptive Statistics

Stefan Heiss

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

October 6, 2023

Data Sets

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

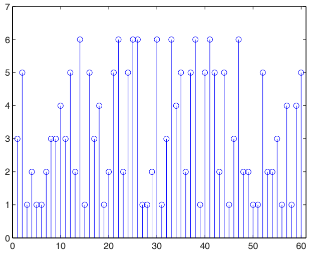
$$x_i, \qquad i = 1, \ldots, N$$

for some $N \in \mathbb{N}$. The x_i's need not to be pairwise different values. (E.g. consider the results of throwing a die for $N = 50$ 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 $\{x_i | i = 1, \ldots, N\}$.

Example (1.1)

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

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

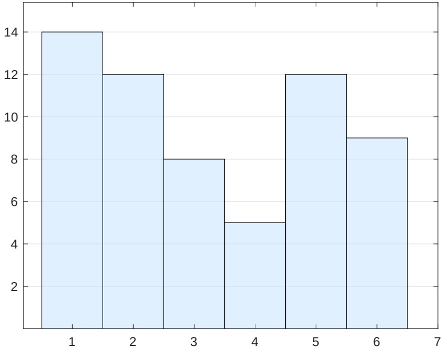


pas01_dice.m

Frequency Table and Frequency Plots

If the order of the values does not 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* or a *frequency plot*.

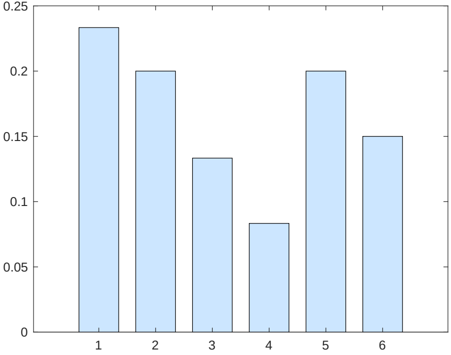
i	n_i	$n_i/60$
value	frequency	relative frequency
1	14	0.233
2	12	0.2
3	8	0.133
4	5	0.083
5	12	0.2
6	9	0.15
	60	$0.999 \approx 1.0$



Frequency Table and Frequency Plots

If the order of the values does not 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* or a *frequency plot*.

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



Modal Values, Sample Mode

Definition (1.2)

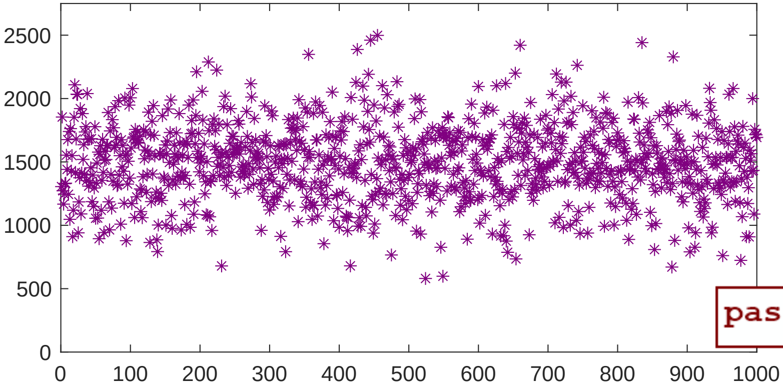
Given a data set $x_1, x_2, \ldots, x_N$, the values that occur with the greatest frequency are called *modal values*. If there is only one such modal value, it is also called the *sample mode*.

Example (1.3)

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

Example (1.4)

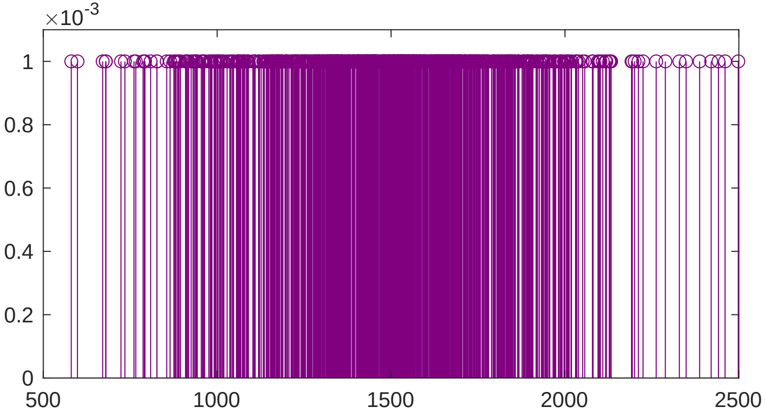
The lifetimes of 1000 incandescent lamps have been determined and the data set of this measurement can be seen in the figure below:



pas01_lamps.m

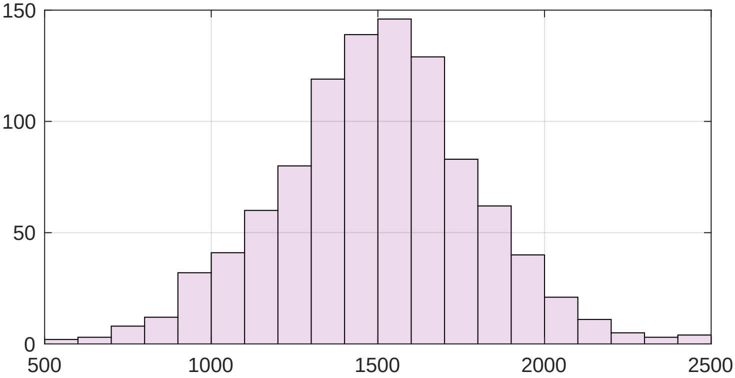
Example (1.4)

Although 1000 different values have been recorded, there is no single pair of equal values. As a consequence a frequency table or frequency plot is not really enlightening:



Histogram

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:



Histogram

MATLAB

A histogram can simply be plotted with MATLAB's `histogram()` function. 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)
```

(500, 600, ... , 2500)

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

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

Note: The last interval also includes the right edge.

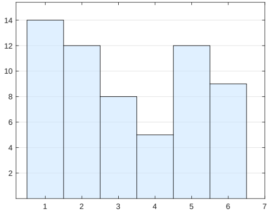
Cumulative Frequency Plot

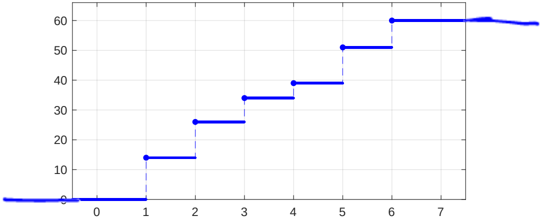
Definition (1.6)

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:

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

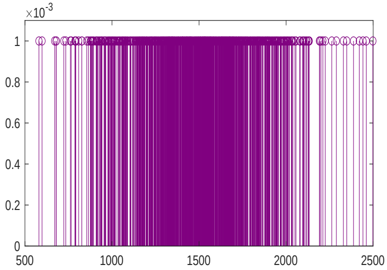
Example

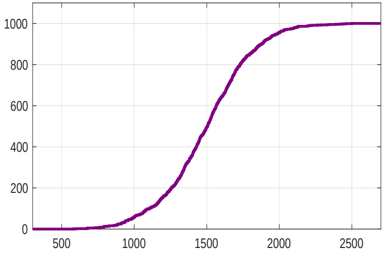




Cumulative Frequency Plot

Example





Quantiles

Definition (1.8)

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:

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

Percentiles

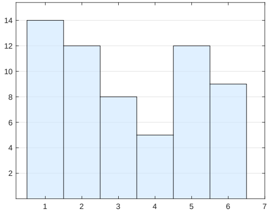
Definition (1.8)

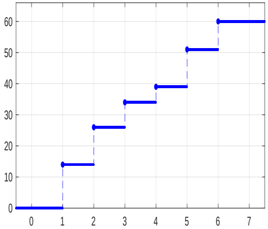
Alternatively $x(p)$ is called the $(100\,p)$*th percentile* of the ordered data set.

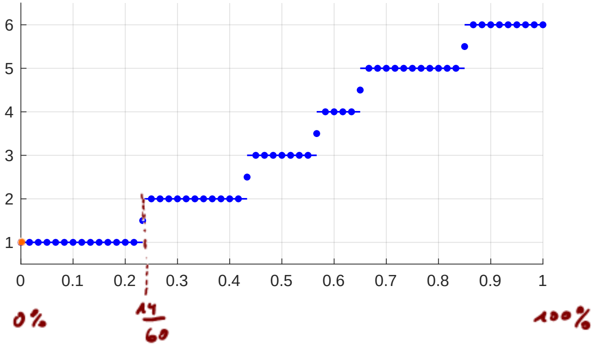
Remark (1.9)

! If $p \cdot N \notin \mathbb{N}$ or $p \in \{0, 1\}$, 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 $\boxed{x(p)}$ and at least $100(1-p)$ percent of the x_i's are greater than or equal to $\boxed{x(p)}$.
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.

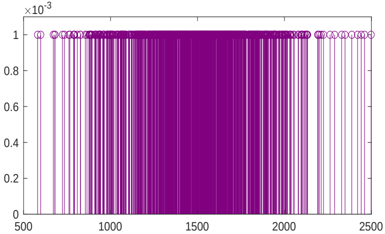
Example (1.10)

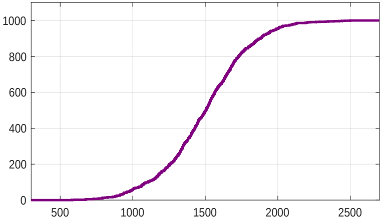

14
12
10
8
6
4
2
1
2
3
4
5
6
7

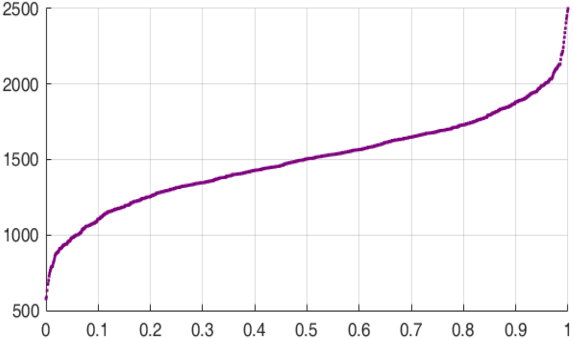

60
50
40
30
20
10
0
0
1
2
3
4
5
6
7


6
5
4
3
2
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0%
14/60
100%

Example (1.10)


×10^-3
1
0.8
0.6
0.4
0.2
0
500
1000
1500
2000
2500


1000
800
600
400
200
0
500
1000
1500
2000
2500


2500
2000
1500
1000
500
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Median and Quartiles

Definition (1.11)

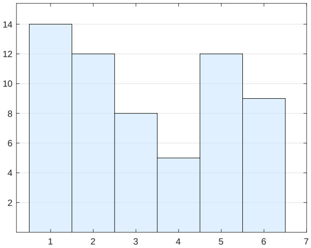
Let $x(p)$ be the quantile function of an ordered data set. Then:

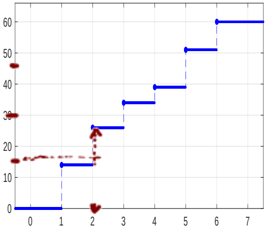
(i) $x(0.25)$ is called the *first quartile*,

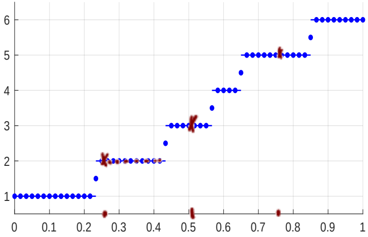
(ii) $x(0.5)$ is called the *median* (or *second quartile*), and

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

Example (1.12)



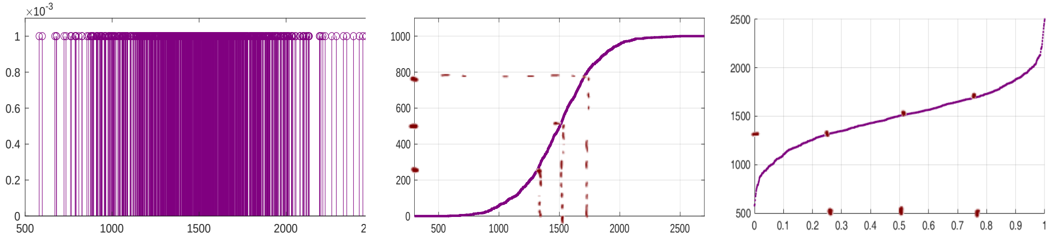




The first quartile, median, and third quartile are:

$$x(0.25) = 2$$
$$x(0.5) = 3$$
$$x(0.75) = 5$$

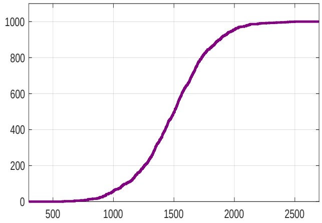
Example (1.12)



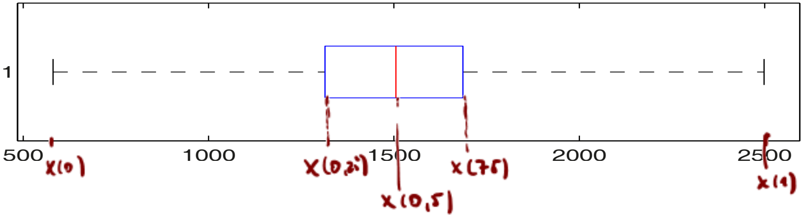
The first quartile, median, and third quartile are:

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

Box Plots



$$
\begin{aligned}
x(0) &= 580.44\\
x(0.25) &= 1314.28\\
x(0.5) &= 1505.05\\
x(0.75) &= 1685.90\\
x(1) &= 2498.40
\end{aligned}
$$



Quantiles

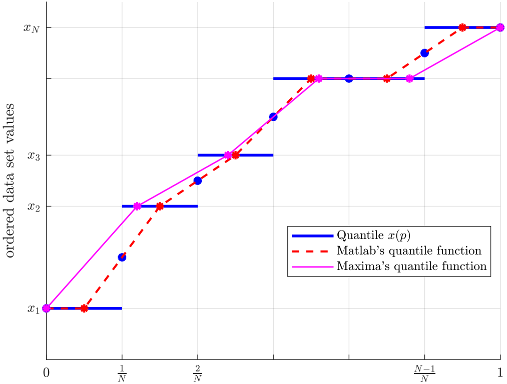
Remark (1.15)

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 $\hat{Q}_1, \dots, \hat{Q}_9$ 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 $\hat{Q}_7$.

Quantiles

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



Mean, Variance and Standard Deviation

Definition (1.16)

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

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

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

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

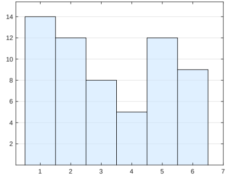
$$\sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum_{i=1}^{N} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + N\cdot\bar{x}^2$$

$$= \sum_{i=1}^{N} x_i^2 - 2\bar{x}\cdot N\bar{x} + N\bar{x}^2$$

$$= \sum x_i^2 - N\bar{x}^2$$

Example (1.17)

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



(i) $$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{60}\left(1\cdot 14 + 2\cdot 12 + \ldots + 6\cdot 9\right) = \frac{196}{60} =$$

(ii) $$s^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right) = \frac{1}{59}\left(1\cdot 14 + 4\cdot 12 + \ldots + 36\cdot 9 - \frac{196^2}{60}\right) = \ldots$$

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

Lemma (1.19)

Let $x_1, x_2, \dots, x_N$ be a data set and

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

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

(iii) $s_y = |b|\cdot s_x$

(i) $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(a+bx_i) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} a}_{N\cdot a} + \underbrace{\frac{1}{N} b \sum_{i=1}^{N} x_i}_{b\bar{x}}$

$= a + b\bar{x}$

(ii) $s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2 \overset{(i)}{=} \frac{1}{N-1}\sum(\cancel{a}+bx_i-(\cancel{a}+b\bar{x}))^2$

$= \frac{1}{N-1}\sum_{i=1}^{N} b^2(x_i-\bar{x})^2 = \frac{1}{N-1}\cdot b^2 \sum_{i=1}^{N}(x_i-\bar{x})^2 = b^2\cdot s_x^2$