Chapter 2

Frequency distributions

The first task at hand in unravelling the intrinsic structure which resides in a given raw data set $\{x_i\}_{i=1,\ldots,n}$ for some statistical variable X corresponds to Cinderella's task of separating the good peas from the bad peas, and collecting them in respective bowls (or bins). This is to say, the first question to be answered requires determination of the frequencies with which a value a_j in the spectrum of possible values of X occurred in the statistical sample S.

2.1 Absolute and relative frequencies

Definition 2.1.1 *Let X be a nominally, ordinally or metrically scaled 1-D statistical variable, with a spectrum of k different values or attributes a_j resp. k different categories (or bins) K_j $(j = 1, \ldots, k)$. If, for X, we have a raw data set comprising n observed values $\{x_i\}_{i=1,\ldots,n}$, we define by*

$$o_j := \begin{cases} o_n(a_j) & = & \text{number of } x_i \text{ with } x_i = a_j \\ o_n(K_j) & = & \text{number of } x_i \text{ with } x_i \in K_j \end{cases} \tag{2.1}$$

$(j = 1, \ldots, k)$ the absolute (observed) frequency of a_j resp. K_j, and, upon division of the o_j by the sample size n, we define by

$$h_j := \begin{cases} \dfrac{o_n(a_j)}{n} \\ \dfrac{o_n(K_j)}{n} \end{cases} \tag{2.2}$$

$(j = 1, \ldots, k)$ the relative frequency of a_j resp. K_j. Note that for all $j = 1, \ldots, k$, we have $0 \leq o_j \leq n$ with $\sum_{j=1}^{k} o_j = n$, and $0 \leq h_j \leq 1$ with $\sum_{j=1}^{k} h_j = 1$. The k value pairs $(a_j, o_j)_{j=1,\ldots,k}$ resp. $(K_j, o_j)_{j=1,\ldots,k}$ represent the distribution of absolute frequencies, the k value pairs $(a_j, h_j)_{j=1,\ldots,k}$ resp. $(K_j, h_j)_{j=1,\ldots,k}$ represent the distribution of relative frequencies of the a_j resp. K_j in S_Ω.

Typical graphical representations of relative frequency distributions, regularly employed in making results of descriptive statistical data analyses public, are the

- histogram for metrically scaled data,
- bar chart for ordinally scaled data,
- pie chart for nominally scaled data.

It is standard practice in Statistics to compile from the relative frequency distribution $(a_j, h_j)_{j=1,\ldots,k}$ resp. $(K_j; h_j)_{j=1,\ldots,k}$ of data for some ordinally or metrically scaled 1-D variable X the associated empirical cumulative distribution function. Hereby it is necessary to distinguish the case of data for a variable with a discrete spectrum of values from the case of data for a variable with a continuous spectrum of values. We will discuss this issue next.

2.2 Empirical cumulative distribution function (discrete data)

Definition 2.2.1 *Let X be an ordinally or metrically scaled 1-D statistical variable, the spectrum of values a_j $(j = 1, \ldots, k)$ of which vary discretely. Suppose given for X a statistical sample S_Ω of size $|S_\Omega| = n$ comprising observed values $\{x_i\}_{i=1,\ldots,n}$, which we assume ordered in increasing fashion according to $a_1 < a_2 < \ldots < a_k$. The corresponding relative frequency distribution is $(a_j, h_j)_{j=1,\ldots,k}$. For all real numbers $x \in \mathbb{R}$, we then define by*

$$F_n(x) := \begin{cases} 0 & \text{for } x < a_1 \\ \sum_{i=1}^{j} h_n(a_i) & \text{for } a_j \leq x < a_{j+1} \quad (j = 1, \ldots, k-1) \\ 1 & \text{for } x \geq a_k \end{cases} \tag{2.3}$$

the empirical cumulative distribution function for X. The value of F_n at $x \in \mathbb{R}$ represents the cumulative relative frequencies of all a_j which are less or equal to x. $F_n(x)$ has the following properties:

- *its domain is $D(F_n) = \mathbb{R}$, and its range is $W(F_n) = [0; 1]$; hence, F_n is bounded from above and from below,*
- *it is continuous from the right and monotonously increasing,*
- *it is constant on all half-open intervals $[a_j; a_{j+1})$, but exhibits jump discontinuities at all a_{j+1}, of size $h_n(a_{j+1})$, and,*
- *asymptotically, it behaves as $\lim_{x \to 1} F_n(x) = 0$ and $\lim_{x \to +\infty} F_n(x) = 1$.*

Exercise 1 Calculate and draw the empirical distribution function of the following sample:

(-15.4,-8.8,8.2,3.4,-7.1,4.5,-12.7,5.2,-10.6,-11.2)

Property 2.2.1 *Computational rules for $F_n(x)$*

1. $h(x \leq d) = F_n(d)$
2. $h(x < d) = F_n(d) - h_n(d)$
3. $h(x \geq c) = 1 - F_n(c) + h_n(c)$
4. $h(x > c) = 1 - F_n(c)$
5. $h(c \leq x \leq d) = F_n(d) - F_n(c) + h_n(c)$
6. $h(c < x \leq d) = F_n(d) - F_n(c)$
7. $h(c \leq x < d) = F_n(d) - F_n(c) - h_n(d) + h_n(c)$
8. $h(c < x < d) = F_n(d) - F_n(c) - h_n(d)$,

wherein c denotes an arbitrary lower bound, and d denotes an arbitrary upper bound, on the argument x of $F_n(x)$.

Exercise 2 Analyse a sample composed of the number of siblings of 20 students of a specific class:

1,1,2,1,0,3,4,2,3,1,0,2,1,1,0,1,1,0,3,2

2.3 Empirical cumulative distribution function (continuous data)

Definition 2.3.1 *Let X be a metrically scaled 1-D statistical variable, the spectrum of values of which vary continuously, and let observed values $\{x_i\}_{i=1,\ldots,n}$ for X from a statistical sample S of size $|S_\Omega| = n$ be binned into k class intervals (or bins) K_j $(j = 1, \ldots, k)$, of width b_j, with lower boundary u_j and upper boundary o_j. The distribution of relative frequencies of the class intervals be $(K_j; h_j)_{j=1,\ldots,k}$. Then, for all real numbers $x \in \mathbb{R}$,*

$$\tilde{F}_n(x) := \begin{cases} 0 & \text{for } x < u_1 \\ \sum_{i=1}^{j-1} h_i + \frac{h_j}{b_j}(x - u_j) & \text{for } x \in K_j \\ 1 & \text{for } x > o_k \end{cases} \tag{2.4}$$

defines the empirical cumulative distribution for X. $\tilde{F}_n(x)$ has the following properties:

Property 2.3.1 *1. its domain is $D(\tilde{F}_n) = \mathbb{R}$, and its range is $W(\tilde{F}_n) = [0; 1]$; hence, $\tilde{F}_n$ is bounded from above and from below,*

2. it is continuous and monotonously increasing, and,

3. asymptotically, it behaves as $\lim_{x\to-\infty} \tilde{F}_n(x) = 0$ and $\lim_{x\to+\infty} \tilde{F}_n(x) = 1$.

Property 2.3.2 *Computational rules for $\tilde{F}_n(x)$*

1. $h(x < d) = h(x \leq d) = \tilde{F}_n(d)$
2. $h(x > c) = h(x \geq c) = 1 - \tilde{F}_n(c)$
3. $h(c < x < d) = h(c \leq x < d) = h(c < x \leq d) = h(c \leq x \leq d) = \tilde{F}_n(d) - \tilde{F}_n(c)$,

wherein c denotes an arbitrary lower bound, and d denotes an arbitrary upper bound, on the argument x of $\tilde{F}_n(x)$.

Exercise 3 25 students of the X-University in X-town were asked in June 2012 about their field of study, number of siblings and income. The outcome was as follows in the table below.

1. What is the statistical population and units in this survey? What characteristics of identification can you define in this population?
2. How is the variable/category "Field of study" scaled?
   Calculate its absolute and relative frequency. Plot the results.
3. How is the variable/category "Number of siblings" scaled?
   Calculate its absolute and relative frequency.
   Calculate empirical cumulative distribution function.
   Plot the results.
4. How many student have at most 2 siblings?
5. What percentage of students has at least two siblings?
6. What percentage of students has 1 or 2 sibling?
7. How is the variable/category "Income" scaled?
   With respect to following grouping:

[600;650);[650;700);[700;900);[900;1200);[1200;1450]

Calculate its absolute and relative frequency.
Calculate empirical c.d.f.
Plot the results.

8. Taking the results of previous task 7., compute:
What percentage of students has income from 750 to 1300€?
What percentage of students has income more than 800€?
What is the highest income of the 50% of the students with the lowest income? What is the smallest income of the 20% of the students with the highest income?

No.	Name	Studies	No. of Siblings	Income
1	Martin A.	Economics	0	924
2	Ute A.	Social S.	1	789
3	Wilhelm A.	Business	0	1365
4	Kurt B.	Business	1	683
5	Sylvia B.	Polit. S.	1	744
6	Elke D.	Polit. S.	2	640
7	Klaus D.	Social S.	2	631
8	Theo E.	Economics	1	814
9	Jean F.	Polit. S.	1	778
10	Elvira G.	Business	0	1062
11	Karl H.	Business	0	1230
12	Andreas K.	Economics	1	700
13	Thomas K.	Business	0	850
14	Chris L.	Social S.	3	641
15	Uwe L.	Polit. S.	2	640
16	Axel M.	Business	0	850
17	Maria M.	Business	1	683
18	Ruth M.	Social S.	0	616
19	Brbel N.	Business	1	683
20	Armin R.	Business	2	683
21	Christa R.	Economics	1	660
22	Bernd S.	Business	1	1440
23	Claudia S.	Social S.	3	794
24	Udo T.	Economics	0	660
25	Clausia W.	Polit. S.	1	640

Our next step is to introduce a set of scale level-dependent standard descriptive measures which characterise specific properties of univariate and bivariate relative frequency distributions of statistical variables X resp. $(X;Y)$.