2. Frequency Measures and Graphical Representation of Data

We highlighted that diffrent variables contains diffrent levels of information. When summarizing or visualizing one or more variables, it is this information which determines the appropriate statistical methods to use.

- Absolute & Relative frequencies:

The number of observations in a particular category is called the absolute frequency.

eg: Ten people in a city. Each of them is either coded as "F" (if the person is female) or "M" (if the person is male).

M, F, M, M, M, F, F, F, F, F.

- There are now two categories in the data male (M) and Female (F). a_1 refer to the male category
a_2 refer to the female category

4 values in category a_1, denoted by n_1
6 values in category a_2, denoted by n_2

$n_1 = 4, \quad n_2 = 6.$

relative frequencies of a_1 & a_2 as $f_1 = f(a_1) = \frac{n_1}{n} = \frac{4}{10} = 0.4$

$f_2 = f(a_2) = \frac{n_2}{n} = \frac{6}{10} = 0.6.$

We now extend these concepts to a general framework for the summary of data on <u>discrete variables</u>.

Suppose there are k categories denoted as $a_1, a_2, \cdots, a_k$ with n_j $(i=1,2,\cdots,k)$ observations in category a_j.

The absolute frequency n_j is defined as the number of units in the j^{th} category a_j. The sum of absolute frequencies equals the total number of units in the data : $\sum_{j=1}^{k} n_j = n$. The relative frequencies of the j^{th} class are defined as $f_j = f(a_j) = \frac{n_j}{n}$, $j=1,2,\cdots,k$

The relative frequencies always lie between 0 and 1 & $\sum_{j=1}^{k} f_j = 1$.

- <u>Grouped continuous Data</u>:

Data on continuous variables usually has a large number (k) of diffrent values. Sometimes k may be the even be the same as n and in such a case the relative frequencies become $f_i = \frac{1}{n}$ for all j. However, it is possible to define intervals in which the observed values are contained.

students 60 Marks. make group Class intervals such as 0-10, 10-20, ··· 90-100

Class intervals	0-10	10-20	···	90-100
absolute frequencies	$n_1 =$	$n_2 =$		
Relative frequencies	$f_1 = \frac{}{60}$			

Frequency distribution for discrete data:

Class intervals (a_j)	a_1	a_2	...	a_k
Absolute frequencies n_j	n_1	n_2	...	n_k
Relative frequencies f_j	f_1	f_2	...	f_k.

Now, suppose the n observations can be classified into k class intervals $a_1, a_2, \dots, a_k$, where a_j ($j = 1, 2, \dots k$) contains n_j observations with $\sum_{j=1}^{k} n_j = n$ & $\sum_{j=1}^{k} f_j = 1$

next page:

ECDF for ordinal variables:

Car service company

~~200~~ cars

overall satisfaction of 200 customers.

1. not satisfied at all
2. unsatisfied
3. satisfied
4. very satisfied
5. percetly satisfied.

- ## Empirical Cumulative Distribution function:

Another approch to summarize and visualize the (frequency) distribution of variable is empirical cumulative distribution function (ECDF).

Before discussing the empirical cumulative distribution function in a more general framework, let us first understand the concept of ordered values.

Consider n observations $x_1, x_2, \cdots, x_n$ of a variable X, which are arranged in ascending order as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ (and are thus on an at least ordinal scale).

The empirical cumulative distribution function $F(x)$ is defined as the cumulative relative frequencies, of all values a_j, which are smaller than or equal to, x:

$$F(x) = \sum_{a_j \leq x} f(a_j) = \underline{f_j}$$

This definition implies that $F(x)$ is a monotonically non-decreasing function, $0 \leq F(x) \leq 1$, $\lim_{x \to -\infty} F(x) = 0$ (the lower limit of F is 0), $\lim_{x \to +\infty} F(x) = 1$ (the upper limit of F is 1) & $F(x)$ is right continuous.

* The empirical cumulative distribution function of ordinal variables is a step function.

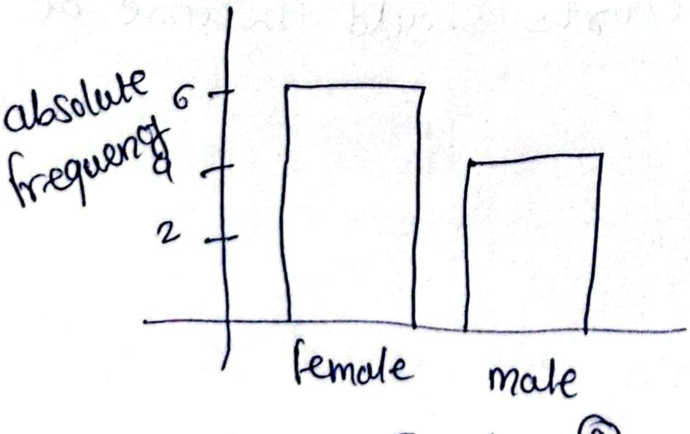
Graphical Representation of a variable:

Frequency tables and empirical cumulative distribution function are useful in providing a numerical summary of a variable. Graphs are an alternative way to summarize a variable's information. In many situation, they have the advantage of conveying the information hidden in the data more compactly.

Bar Chart:

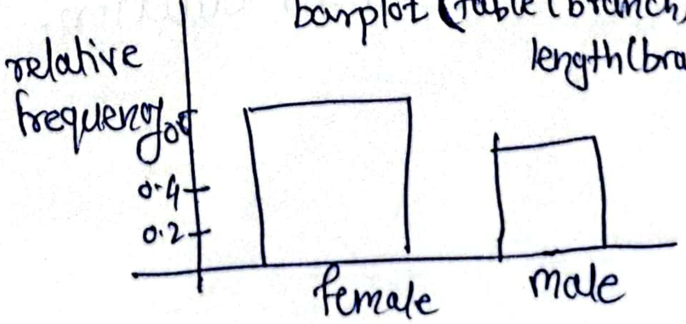
A simple tool to visualize the relative or absolute frequencies of observed values of a variables is a bar chart. A bar chart can be used for nominal and ordinal variables, as long as the number of categories is not very large.

It consist of one bar for each category. The height of each bar is determined by either the absolute frequency or relative frequency of the respective category and is shown on the y-axis.



(a)

```
barplot(table(branch))
barplot(table(branch)/length(branch))
```



(b)

Pie Chart: Pie charts are another option to visuali[ze] the absolute and relative frequencies of nominal & ordinal variables. A pie chart is a circle partitioned, into segments where each of the segments represents a category. The size of each segment depends upon relative frequencies & is determined by the angle $f_j \cdot 360°$.

$0.6 \times 360 = 216$
$0.4 \times 360 = 144$

pie(table(sv)).

Remark: Note that the area of a segment is not proportional to the absolute frequency of the respective category. Instead, the area of the segment is proportional to the angle $f_j \cdot 360°$. (& depends also on ther radius of the whole circle. (This may cause improper interpretations as the human eye may catch the segment's area more easily than the angle of a segment. Pie charts should therefore be used with caution.

Histogram:

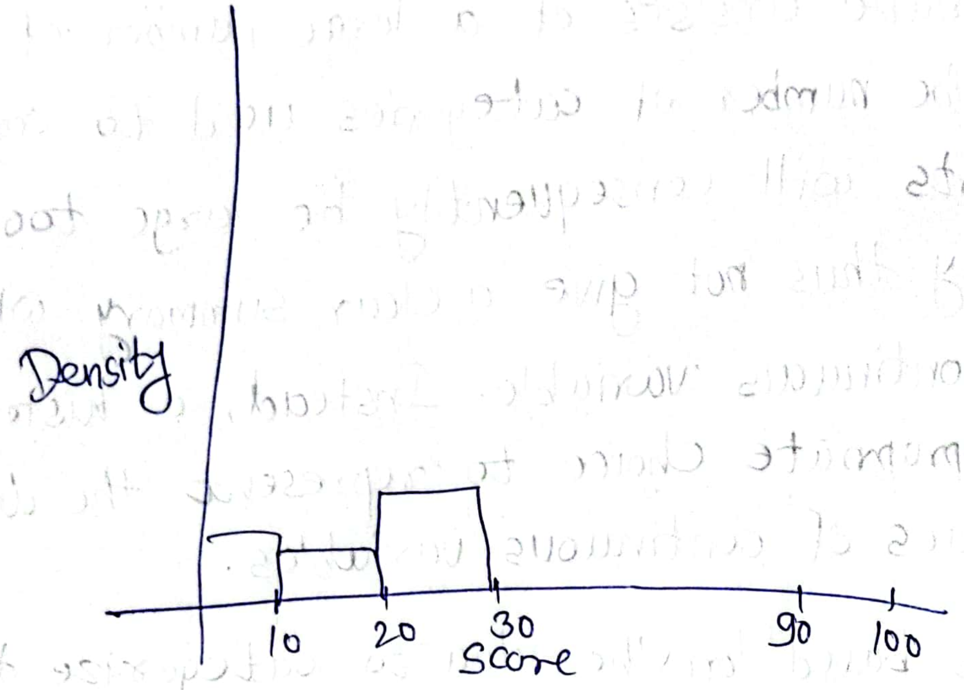
If a variable consists of a large number of diffrent values, the number of categories used to construct bar charts will consequently be large too. A bar Chart may thus not give a clear summary when applied to a continuous variable. Instead, a histogram is the appropriate choice to represent the distribution of values of continuous variables.

It is based on the idea to categorize the data into diffrent groups and plot the bars for each category with height $h_j = f_j/d_j$ where $d_j = e_j - e_{j-1}$ denotes the width of the j^{th} interval or category. An important consideration for this concept is that the area of the bars = (height × width) is proportional to the relative frequency.

width of bars need not to be the same.

Class intervals	0-10	10-20	- - - -	90-100
Absolute frequency	$n_1 =$	$n_2 =$		
Relative frequency	$f_1 =$	$f_2 =$		
Height	$h_1 = \frac{f_1}{10}$ $d_i = 10$			

Histogram for the scores of the people :



exa: Suppose we have following information to construct a histogram for a continuous variable with 2000 observations

j	e_{j-1}	e_j	d_j	h_j	$f_j = d_j \cdot h_j$	absolute freq
1	0	1	1	0.125	0.0125	250
2	1	4	3	0.125	0.375	750
3	4	7	3	0.125	0.375	750
4	7	8	1	0.125	0.385	750 .

(a) Determine the relative frequencies for each interval.

(b) Determine the absolute frequencies.