Frequency Distributions

Frequency distributions

Let X be a variable, either qualitative or quantitative, and let $x_1, \dots, x_k$ be the values it takes in the data.

Assume that n_j is the number of units for which $X = x_j$. n_j is called the **frequency** of x_j

Frequency distribution: a table that lists the values of X along with the corresponding frequencies.

Value of X	Frequency
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k
Total	n

This is the frequency distribution of X

$$n = n_1 + n_2 + \cdots + n_k$$

n is the number of units in the sample

Frequency distributions: example

Cycle of chemotherapy	number of patients (frequency)
1	20
2	19
3	20
4	29
5	24
6	26
Total	138

With the construction of a frequency distribution, we move from raw data to a concise form useful for both the presentation of data and the understanding the main characteristics of phenomenon under study.

Relative frequency distributions

Being n_j is the number of units for which $X = x_j$ and n th number of all units in the sample, the **relative frequency** of x_j is given by

$$f_j = \frac{n_j}{n}$$

The relative frequency of x_j is often calculated as a percentage

$$p_j = 100 \times \frac{n_j}{n}$$

Relative Frequency distribution: a table that lists the values of X along with the corresponding relative frequencies.

Relative frequency and percentage distributions: example

Cycle of chemotherapy	Number of Patients (frequency)	**Relative Frequency**	**Percentages**
1	20	**0.14**	**14.5**
2	19	**0.14**	**13.8**
3	20	**0.14**	**14.5**
4	29	**0.21**	**21.0**
5	24	**0.17**	**17.4**
6	26	**0.19**	**18.8**
Total	138	**1.00**	**100.0**

Cumulative frequency distribution

Assume n_j the number of units for which $X = x_j$.

The cumulative frequency of x_j is the number of units for which X is less than or equal to x_j: $N_j = n_1 + n_2 + \cdots + n_j$ where $j \leq k$

It clearly holds $N_k = n$

Value of X	Cumulative frequencies
x_1	$N_1 = n_1$
x_2	$N_2 = n_1 + n_2$
$\vdots$	$\vdots$
x_k	$N_k = \sum_{i=1}^{k} n_i = n$

Cumulative frequency distribution: a table that lists the values of X along with the corresponding **cumulative** frequencies

Cumulative relative frequency distribution

Let n_j be the number of units for which $X = x_j \quad j = 1, \ldots, k$

Let N_j be the cumulative frequency of x_j

The cumulative relative frequency x_j is given by $P_j = \frac{N_j}{n}$

It clearly holds $P_k = n$

Cumulative relative frequency distribution: a table that cumulates relative frequencies up to a certain value of X.

If expressed in percentages, it is the percentage of units in the sample for which $X \leq x_j$.

Cumulative and Cumulative relative frequency distributions: example

Cycle of chemotherapy	Number of patients (frequency)	Relative Frequency	Percentages	Cumulative Frequency Distribution	Cumulative Relative Distribution	Cumulative Percentages
1	20	0.14	14.5	**20**	**0.14**	**14**
2	19	0.14	13.8	**39**	**0.28**	**28**
3	20	0.14	14.5	**59**	**0.43**	**43**
4	29	0.21	21.0	**88**	**0.64**	**64**
5	24	0.17	17.4	**112**	**0.81**	**81**
6	26	0.19	18.8	**138**	**1.00**	**100**
Total	138	1.00	100.0			

Note that cumulative distributions can be calculated only if the variable is numerical or ordinal.

Grouped variables

The observation of a numerical variable can result in a series of measurements that are significantly varied from one another, so constructing a frequency distribution using the procedure illustrated earlier may produce unsatisfactory results.

Consider, for example, the height of 100 male individuals. With a reasonably accurate instrument we may have 50 or 60 different measurements each of which having a very low frequency. As a result, the frequency table considered previously would not be suitable for summarising and interpreting the variable

In such cases, it is more effective to cut the variable values into classes and use the categorised version of it to construct the frequency distributions for describing the variable in the sample

Group frequency distributions

Let X be a numerical variable.

Let $[c_i - c_{j+1})$ be an interval of values of X where $j = 0, \cdots, k$ i.e. we have k intervals or classes.

Assume that n_j is the number of units of the sample for which $c_i \leq X < c_{j+1}$.

n_j is called the **frequency** of the category $c_i - c_{j+1}$.

Class	Frequency
c_0-c_1	n_1
c_1-c_2	n_2
$\vdots$	$\vdots$
c_{k-1}-c_k	n_k
Total	n

Group Frequency distribution: a table that lists the classes into which the values of X are grouped along with the corresponding frequencies.

In a similar manner we can calculate group relative frequency and percentage distributions as well as cumulative group distribution

Group frequency distributions: example

Age	Frequency
22	2
23	1
24	1
28	1
29	2
30	1
31	1
32	3
33	1
34	4
….	….
71	3
73	2
74	1
Total	138

Frequency distribution of the patient's age.

Frequency distribution is not much useful in this case.

There are too many values of age in the sample.

Grouping data gives a clearer information on the distribution of patient age.

Age Class	Frequency
<35	17
35-50	45
>50	76
Total	138

Group cumulative frequency distributions: example

Age Class	Frequency	Cumulative Frequency
<35	17	17
35-50	45	62
>50	76	138
Total	138	

The categories of the variables should be at least ordered.

Similarly, one can calculate relative frequencies and percentages (cumulative) distributions

Bivariate (and multivariate) frequency distributions

Let X and Y be two discrete or categorical variables.

Assume that X takes s different values $\tilde{x}_1, \dots, \tilde{x}_s$, and Y takes t different values $\tilde{y}_1, \dots, \tilde{y}_t$

Let $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$, be the values that X and Y take on each unit of the data set.

Assume that n_{ij} is the number of units of the sample for which $X = \tilde{x}_i$ and $Y = \tilde{y}_j$

n_{ij} is called the **joint frequency** of the pair $(\tilde{x}_i, \tilde{y}_j)$.

	$\tilde{y}_1$		$\tilde{y}_j$		$\tilde{y}_t$	Marginal X
$\tilde{x}_1$	$\boldsymbol{n_{11}}$	$\cdots$	$\boldsymbol{n_{1j}}$	$\cdots$	$\boldsymbol{n_{1t}}$	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_i$	$\boldsymbol{n_{i1}}$	$\cdots$	$\boldsymbol{n_{ij}}$	$\cdots$	$\boldsymbol{n_{it}}$	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\tilde{x}_s$	$\boldsymbol{n_{s1}}$	$\cdots$	$\boldsymbol{n_{sj}}$	$\cdots$	$\boldsymbol{n_{st}}$	$n_{s.}$
Marginal Y	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.t}$	$\boldsymbol{n}$

The joint frequency distribution is a rectangular $s \times t$ table that lists all the pairs of values/classes of X and Y along with their corresponding frequencies .

In a similar manner we can arrange the data when we have three or more variables

Bivariate frequency distributions

Frequency Distribution

	ECOG PS			
Randomisation	0	1	2	Randomisation Distrib.
treatment A	47	24	3	74
treatment B	48	10	6	64
ECOG PS distrib	95	34	9	138

Relative Frequency Distribution

	ECOG PS			
Randomisation	0	1	2	Distribution of Randomisation
treatment A	0.34	0.17	0.02	0.54
treatment B	0.35	0.07	0.04	0.46
Distribution of ECOG PS	0.69	0.25	0.07	**1**

Variable Description

ECOG PS Performance status score:

ECOG PS=0: Fully active

ECOG PS=1: Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature

ECOG PS=2: Ambulatory and capable of all self care but unable to carry out any work activities.

ECOG PS is an ordered variables

Randomisation is measured on a nominal scale