Descriptive Statistics

Dr. Abbas Adigun (PhD)

Biostatistician

19th May 2017

Descriptive statistics

It is a series of procedures designed to illuminate the data, so that its primary characteristics and main features are revealed. it simply describe some numerical property of a set of data, with no indication on how that data relate to our hypotheses.

This may mean sorting the data by size; perhaps putting it into a table, maybe presenting it in an appropriate chart, or summarizing it numerically; and so on.

An important consideration in this process is the type of variable concerned. The data from some variables are best described with a table, some with a chart, some, perhaps, with both. With other variables, a numeric summary is more appropriate.

Descriptive statistics are broken down into measures of central tendency, or location and measures of variability, or spread.

Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation or variance, the minimum and maximum variables, and skewness.

All descriptive statistics, whether they be the mean, median, mode, standard deviation, kurtosis or skewness, are either measures of central tendency or measures of variability.

Describing data with tables

- **Frequency table for nominal data**

  Suppose we have used 95 questionnaires to collect data, and one of the variable being the child's blood type A, B, AB and O.

  The resulting *frequency table* for the four blood type categories is shown in Table 2.1.

  As you know, the ordering of nominal categories is arbitrary, and in this example they are shown by the number of children in each – largest first.

  Notice that total frequency (*n* = 95), is shown at the top of the frequency column. This is helpful to any reader and is good practice.

Table 2.1 Frequency table showing the distribution of blood type of the 95 children in a study.

Category (Blood group)	Frequency (no of children)
AB	49
A	27
B	15
O	4

Table 2.2 Relative frequency table, showing the *percentage* of children in each blood type category

Category (Blood group)	Frequency (no of children)	Relative frequency (% of children in each category)
AB	49	51.6
A	27	28.4
B	15	15.8
O	4	4.2

Ordinal variables – organizing the data into ordered categories

When the variable in question is ordinal, we can allocate the data into ordered categories.

For instance, table 2.3 shows the frequency distribution for the variable, household socio-economic status of women of child bearing age in Nigeria, obtained from malaria indicator survey data (MIS 2010). The variable has five categories as shown.

Socio-economic status is clearly an ordinal variable. It cannot be properly measured, and has no units. But the categories can be meaningfully ordered, as they have been here.

Socio-Economic status	Frequency N=7754
Poorest	1054
Poorer	1351
Middle	1676
Less poor	1844
Least poor	1816

Continuous metric variables – organizing the data by value

Organising raw metric *continuous* data into a frequency table is usually impractical, because there are such a large number of possible values. Indeed, there may well be no value that occurs more than once.

This means that the corresponding frequency table is likely to have a large, and thus unhelpful, number of rows.

Not of much help in uncovering any pattern in the data.

The most useful approach with metric continuous data is to *group* them first, and then construct a frequency distribution of the grouped data.

Let’s see how this works. The table below show some characteristic associated with 30 infants including birthweight.

Table 2.5

Infant ID	Birthweight (Kg)	Apgar score	Sex	Mother smoked	Mothers parity
1	3710	8	m	no	1
2	3650	7	f	no	1
3	4490	8	m	no	0
4	3421	6	f	yes	1
5	3399	6	f	no	2
6	4094	9	m	no	3
7	4006	8	m	no	0
8	3287	5	f	yes	5
9	3594	7	f	no	2
10	4206	9	m	no	4
11	3508	7	f	no	0
12	4010	8	m	no	2
13	3896	8	m	no	0
14	3800	8	f	no	0

15	2860	4	m	no	6
16	3798	8	f	no	2
17	3666	7	f	no	0
18	4200	9	m	yes	2
19	3615	7	m	no	1
20	3193	4	f	yes	1
21	2994	5	f	yes	1
22	3266	5	m	yes	1
23	3400	6	f	no	0
24	4090	8	m	no	3
25	3303	6	f	yes	0
26	3447	6	m	yes	1
27	3388	6	f	yes	1
28	3613	7	m	no	1
29	3541	7	m	no	1
30	3886	8	m	yes	1

Among the 30 infants there are *none* with the same birthweight, and a frequency table with 30 rows and a frequency of 1 in every row would add very little to what you already know from the raw data (apart from telling you what the minimum and maximum birth weights are).

One solution is to *group* the data into (if possible) groups of equal width, to produce a *grouped frequency distribution*.

The resulting grouped frequency table for birthweight is shown in table below. This gives us a much better idea of the data's main features than did the raw data.

For example, you can now see that most of the infants had a birthweight around the middle of the range of values, about 3600g, with progressively fewer values above and below this

The resulting grouped frequency table for birthweight is shown in table below. This gives us a much better idea of the data’s main features than did the raw data. For example, you can now see that most of the infants had a birthweight around the middle of the range of values, about 3600g, with progressively fewer values above and below this.

Birthweight	No of infants
2700 - 2999	2
3000 - 3299	3
3300 - 3599	9
3600 - 3899	9
3900 - 4199	4
4200 - 4499	3

Exercise

Construct a grouped frequency table of percentage mortality using five groups. What do you observe?

ICU	1	2	3	4	5	6	7	8	9	10	11	12	13
% mort.	15.2	31.3	14.9	16.3	19.3	18.2	20.2	12.8	14.7	29.4	21.1	20.4	13.6
ICU	14	15	16	17	18	19	20	21	22	23	24	25	26
% mort.	22.4	14.0	14.3	22.8	26.7	18.9	13.7	17.7	27.2	19.3	16.1	13.5	11.2

Frequency tables with discrete metric variables

Constructing frequency tables for metric *discrete* data is often less of a problem than with continuous metric data, because the number of possible values which the variable can take is often limited (although, if necessary, the data can be grouped in just the same way).

As an example, Table 2.8 is a frequency table showing the number of times in the past 24 hours that 53 asthmatic children used their inhaler. We can easily see that most used their inhaler once or twice. Notice the open-ended row showing that six children had used their inhaler five or more times.

Table 2.8

Number of times of inhalers used	Frequency (No of children)
0	6
1	16
2	12
3	8
4	5
>=5	6

Exercise

The data below are the *parity* (the number of previous live births) of 40 women chosen at random from the 332 women in the stress and breast cancer study. (a) Construct frequency and relative frequency tables for this parity data.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Parity	4	0	2	3	2	2	3	3	0	3	1	2	8	3	4	2	1	2	2	2
ID	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Parity	2	2	3	2	2	3	0	3	2	4	0	1	3	5	1	1	0	3	2	1

Cross tabulation

Each of the table describe so far only provides description on frequency distribution on single variable. Sometimes, you will want to examine the association between *two* variables, within a *single* group of individuals. This is made easy by putting the data into a table of cross-tabulations.

Example

To illustrate the idea, let's return to the 30 infants whose data is recorded in Table 2.5. Suppose you are particularly interested in a possible association between infants whose Apgar score is less than 7 (since this is an indicator for potential problems in the infant's well-being), and whether during pregnancy the mother smoked or not. Notice that we have only one group here, the 30 infants, but two sub-groups, those with an Apgar score of less than 7, and those with a score of 7 or more.

Mother smoked during pregnancy		Apgar score<7	
		Yes	No
	Yes	8(72.7)	2(10.5)
	no	3(27.3)	17(89.5)

Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\infant.dta", clear

gen apgar_score_gp=apgar_score

recode apgar_score_gp (min/6=1)  (7/max=2)

label define apgar_score_gp 1 "yes" 2  "no"

label value apgar_score_gp apgar_score_gp

tab smoked_during_preg apgar_score, col
```

Exercise

Use breast_cancer.dta to see if there is association between breast cancer diagnosis and parity using categories (‘two or fewer children’, and ‘more than two children’).

CORBIS/Brian Leng (05065)