Contents

2. Descriptive Statistics

Descriptive statistics deals with presenting and summarizing collected data or information with the purpose of understanding the general properties of the data set. It includes the construction of graphs, charts, and tables, and the calculation of descriptive measures such as means and standard deviations.

2.1 Frequency Distribution and Graphical Representation

In this section, we review some techniques of graphical presentation of data: frequency distribution, cumulative frequency distribution, and histogram.

2.1.1 Frequency Distribution and Cumulative Frequency Distribution

Some statistics about first 20 Normal-type Pokemon are shown in the following table:

Name	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
Pidgey	251	40	45	40	35	35	56
Pidgeotto	349	63	60	55	50	50	71
Pidgeot	479	83	80	75	70	70	101
PidgeotMega	579	83	80	80	135	80	121
Rattata	253	30	56	35	25	35	72
Raticate	413	55	81	60	50	70	97
Spearow	262	40	60	30	31	31	70
Fearow	442	65	90	65	61	61	100
Jigglypuff	270	115	45	20	45	25	20
Wigglytuff	435	140	70	45	85	50	45
Meowth	290	40	45	35	40	40	90
Persian	440	65	70	60	65	65	115
Farfetch'd	352	52	65	55	58	62	60
Doduo	310	35	85	45	35	35	75
Dodrio	460	60	110	70	60	60	100
Lickitung	385	90	55	75	60	75	30
Chansey	450	250	5	5	35	105	50
Kangaskhan	490	105	95	80	40	80	90
KangaskhanMega	590	105	125	100	60	100	100
Tauros	490	75	100	95	40	70	110

Table: Statistics of first 20 Normal-type Pokemon

We could hardly have any insight just from the raw data. Hence, data need to be presented in an appropriate way before exploring the abilities (e.g. Attack, Defense, etc.) of them. For example, if we want to analyze the Attack of them, we first extract the relevant data:

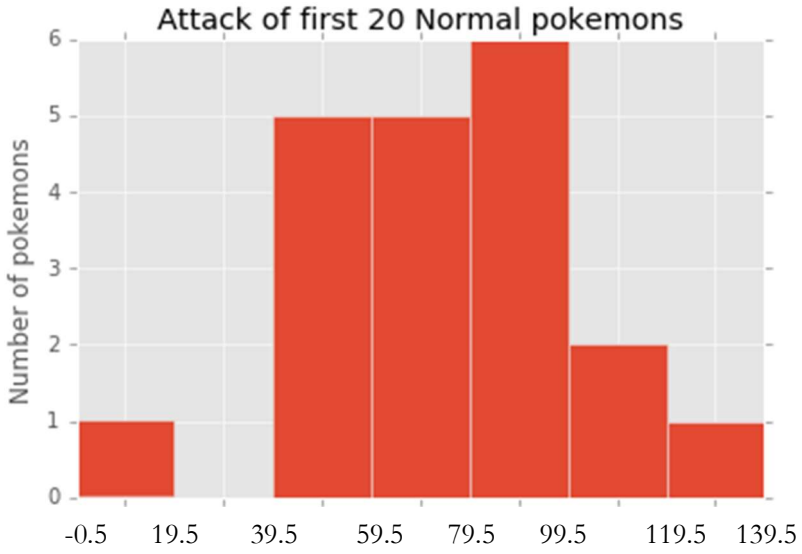
45, 60, 80, 80, 56, 81, 60, 90, 45, 70, 45, 70, 65, 85, 110, 55, 5, 95, 125, 100

Then we set up a number of classes: 0-19, 20-39, …, 120-139, and count the frequencies for the classes.

Attack	Number of Pokemon
0-19	1
20-39	0
40-59	5
60-79	5
80-99	6
100-119	2
120-139	1

The above table is called the frequency distribution of the data. From this table, we have some idea about the "spread" of these data.

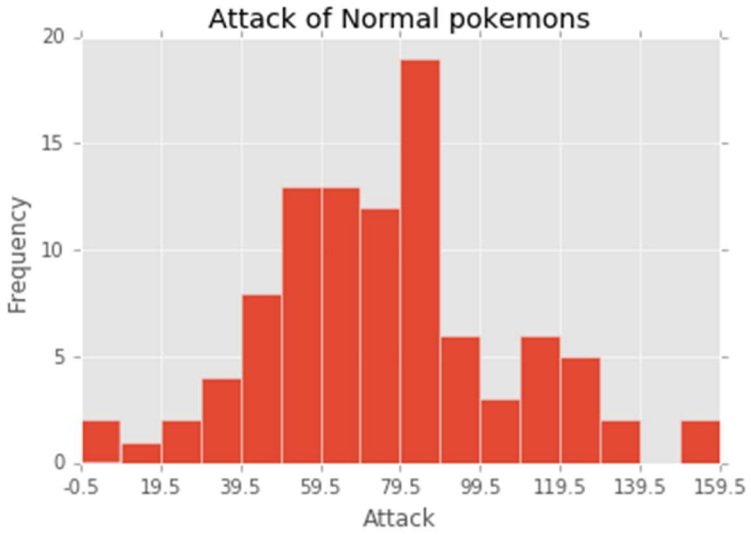
We can plot a histogram showing the "shape" of the data set, as follows:



Sometimes, we want to present the data in a cumulative manner so that we can know how many of the data are more than or less than some particular values. A cumulative frequency distribution can serve this purpose. According to a certain data set, there are 98 Normal-type Pokemon. If we want to analyse a particular Pokemon and its ranking among all Normal-type Pokemon, we can make use of the cumulative frequency distribution.

After grouping the data, a frequency distribution and a histogram are constructed as follows:

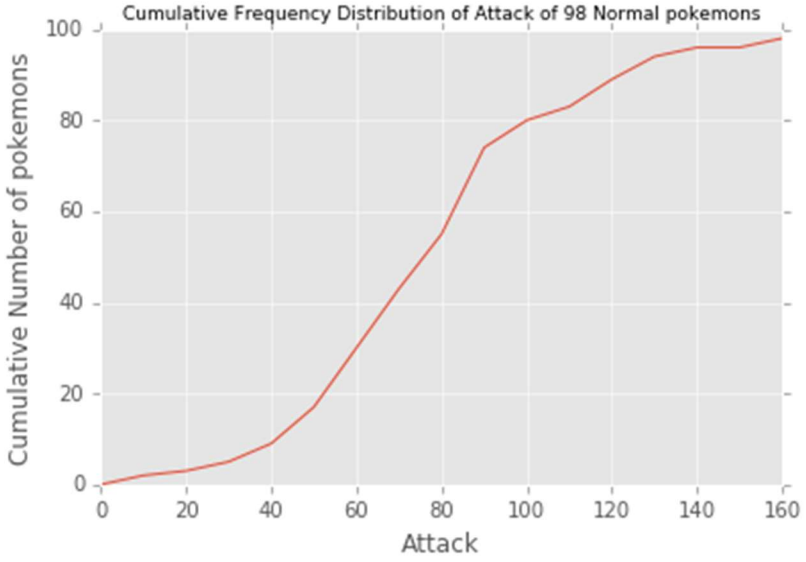
Attack	Number of Pokemon
0-9	2
10-19	1
20-29	2
30-39	4
40-49	8
50-59	13
60-69	13
70-79	12
80-89	19
90-99	6
100-109	3
110-119	6
120-129	5
130-139	2
140-149	0
150-159	2



Then we construct the following cumulative frequency:

Attack (less than or equal to)	Cumulative Frequency
9	2
19	3
29	5
39	9
49	17
59	30
69	43
79	55
89	74
99	80
109	83
119	89
129	94
139	96
149	96
159	98

The cumulative frequency distribution shows the numbers of Pokemon with Attack less than or equal to some particular values. The corresponding cumulative frequency polygon is constructed as follows:

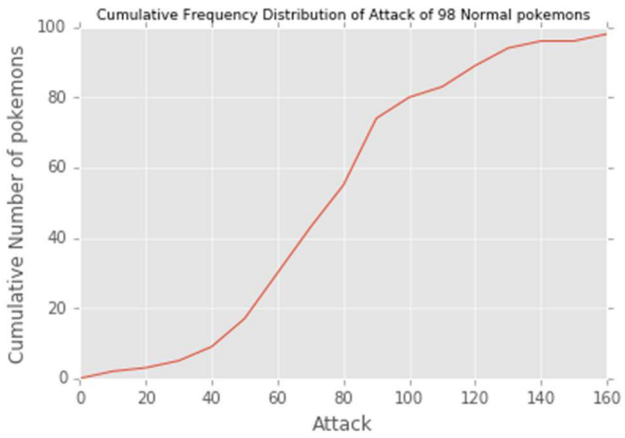


The cumulative frequency polygon can present information that cannot be easily obtained from a histogram. For example, from the graph, we know that there are about 80 pokemon with Attack less than or equal to 100. This information cannot be easily grasped from the histogram.

2.1.2 Quartile

Quartiles are the three values dividing the whole set of ordered data into four groups having the same amount of data. Quartiles can also be found easily from a cumulative frequency polygon.

We use the cumulative frequency polygon of 98 Pokemon of type Normal again as an example.



From the above figure, we find that there are about 25% Pokemon of type Normal with Attack lower than 55. This number is called the first quartile, denoted as Q_1. Also, there are 25% of Pokemon lies between 55 and 74, and 74 is called the second quartile, Q_2. As there are 75% of Pokemon with Attack lower than 89, 89 is called the third Quartile, Q_3.

More formally, in a distribution, 25% of the data are $\leq Q_1$; 50% of the data are $\leq Q_2$, and 75% of the data are $\leq Q_3$.

Another name of Q_1 is the lower quartile, while Q_3 is called the upper quartile. Moreover, both median and Q_2 means that 50% of the data are less than or equal to that value. In other words, they are the same.

Let's explore more properties of quartiles through examples.

Example 1

The grade point distribution in the best five subjects of 2015 HKDSE are listed below.

Total Grade Points	Number of Candidates (Percentage)
0-9	19747 (31.1%)
10-12	4947 (7.8%)
13-15	9443 (14.9%)
16-18	10677 (16.8%)
19-21	8623 (13.6%)
22-24	5075 (8.0%)
25-27	2760 (4.3%)
28-30	1326 (2.1%)
31-33	733 (1.2%)
34-35	196 (0.3%)

Now, from that pool of candidates, we randomly draw a sample of 23 candidates and their total grade points are listed below:

27, 16, 21, 30, 28, 6, 21, 12, 13, 17, 15, 24, 19, 14, 17, 16, 25, 12, 16, 7, 18, 20, 33

In order to find the quartiles, these sample values are ordered as follows:

6, 7, 12, 12, 13, 14, 15, 16, 16, 16, 17, 17, 18, 19, 20, 21, 21, 24, 25, 27, 28, 30, 33

$$\begin{aligned} Q_1 &= \frac{(n+1)}{4}\text{th value} \\ &= \frac{(23+1)}{4}\text{th value} \\ &= 6\text{th value} \end{aligned}$$

where n is the number of values in the sample. Hence, Q_1 is 14.

Similarly,

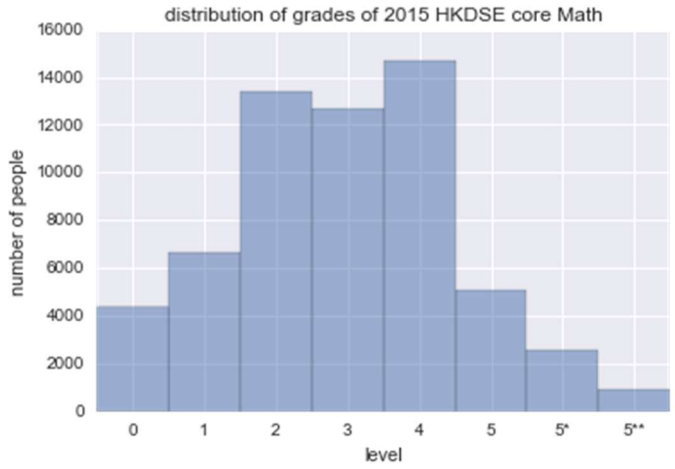
$$Q_2 = \frac{(n+1)}{2}\text{th value} = 12\text{th value}$$

$$Q_3 = \frac{3(n+1)}{4}\text{th value} = 18\text{th value}$$

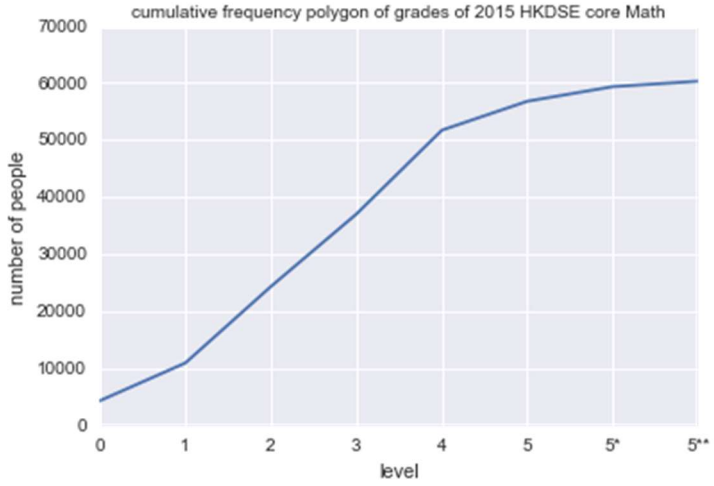
Hence, in this sample, Q_2 is 17 and Q_3 is 24.

Example 2

The frequency distribution of grades of 2015 HKDSE core Mathematics is shown in the following histogram:



From the histogram, the following cumulative frequency polygon is obtained:



As approximately 60000 candidates attended the examination, the quartiles Q_1, Q_2, and Q_3 would be the grades of the 15000th, 30000th and the 45000th candidates, respectively.

Interpretation from the cumulative frequency polygon: Q_1 is higher than level 1 but not higher than level 2, that is, at level 2; similarly, Q_2 is at level 3, and Q_3 is at level 4.

Please attempt Interactive Exercise 2.1.

2.2 Central Tendency and Dispersion

Central tendency means the "middle" or the typical value of the data collected. Mean, mode and median are commonly used measurements of central tendency.

While central tendency is the tendency of quantitative data to cluster around some central value, dispersion refers to how data are scattered. Typical measures of dispersion are standard deviation and variance.

2.2.1 Central Tendency

Suppose a survey of the salaries of fresh graduates has been conducted. Twenty values have been randomly selected from the survey data and the results are listed below:

16500, 13800, 14900, 17400, 16700, 11000, 14900, 12700, 12700, 13800,

13200, 15900, 14500, 13200, 13800, 13600, 15900, 12500, 13600, 11200

Then, the mean of the values in the sample can be found as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $\bar{x}$ is the sample mean and n is the sample size.

Remark:

We usually denote sample mean as $\bar{x}$ and population mean as μ, which is found as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

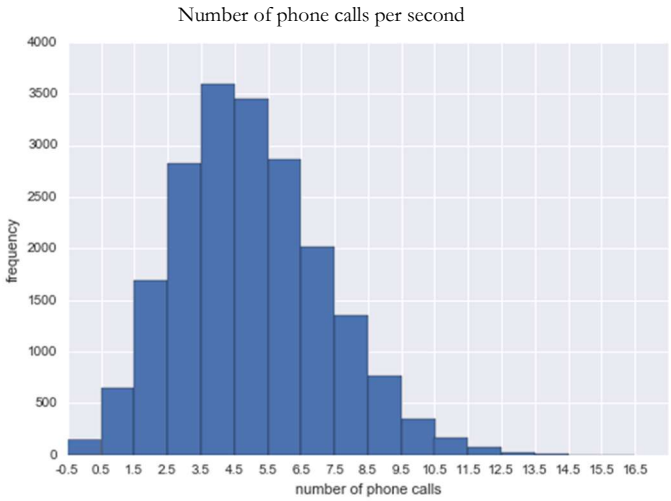
where N is the population size.

In this sample of size 20, the mean is \$14090. Median is the middle data in an ordered sequence of data. Mode is the value that happens most frequently in a set of data. Hence, the median is \$13800 and the mode is also \$13800 in this set of data.

Let us draw a histogram to see the shape of these data.

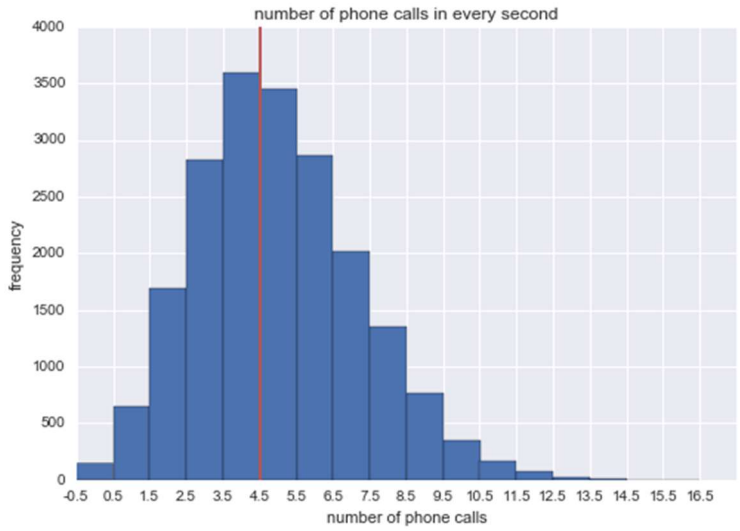
The mode of the data is $13800 which is in the class of $12999.5-$13999.5. The shape of the distribution is roughly symmetric with the class of the highest frequency near the middle of the set of data. The values of mean, mode and median are close to one another.

Let us explore another set of data. A certain company has a hotline for consultation of maintenance. Every day, there are thousands of phone calls asking for consultation. The frequency distribution of the number of phone calls per second for a random sample of 20000 one-second intervals is shown in the following histogram:



In the histogram, "frequency" is the number of one-second intervals, e.g., there are approximately 600 one-second intervals having one phone call each.

The mean number of phone calls per second is indicated by the red line in the following histogram:



In this case, the distribution of the number of phone calls is not symmetric. The mean is not near the middle value of the set of data.

2.2.2 Dispersion

Dispersion or variation is the measurement of how data are scattered. Standard deviation and variance are popular measures of dispersion. In this section, we will review population variance and population standard deviation only. (Sample variance and sample standard deviation will be introduced in the course of AMA1501.)

Population variance, denoted by σ^2, is the average of the squared "distances" of the values from the mean. It can be found by the following formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Population standard deviation, denoted by σ, is the square root of population variance and can be calculated by the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Example 1

In a midterm examination, the marks of 30 students are listed below:

87, 74, 79, 92, 88, 60, 79, 68, 68, 74, 71, 84, 77, 71, 74, 73, 84, 67, 73, 61, 44, 76, 78, 62, 92, 55, 70, 68, 85, 84.

The mean of these marks is given by:

$$\text{mean} = \frac{87 + 74 + \ldots + 85 + 84}{30} = 73.93$$

Therefore, the variance and the standard deviation are:

$$\text{variance} = \frac{(87 - 73.93)^2 + (74 - 73.93)^2 + \ldots + (85 - 73.93)^2 + (84 - 73.93)^2}{30} = 115.86$$

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{115.86} = 10.76$$

Example 2

Assume that after the midterm examination, the students are required to attend the final examination. The marks for the final examination are listed below:

73, 77, 52, 30, 63, 73, 94, 94, 62, 63, 49, 41, 35, 59, 61, 44, 85, 37, 65, 52, 77, 59, 46, 69, 78, 71, 76, 57, 62, 100.

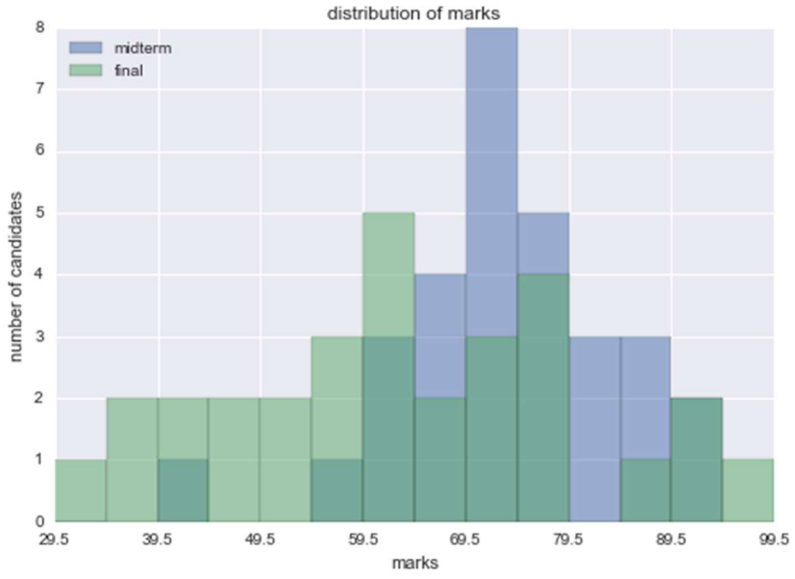
The mean, variance and standard deviation are:

$$\text{mean} = \frac{73+77+\ldots+62+100}{30} = 63.46$$

$$\text{variance} = \frac{(73-63.46)^2+(77-63.46)^2+\ldots+(62-63.46)^2+(100-63.46)^2}{30} = 301.78$$

$$\text{standard deviation} = \sqrt{\text{variance}} = \sqrt{301.78} = 17.37$$

The standard deviation of the marks for the final examination is much larger than that for the midterm examination. The histograms for the distributions of marks for the two examinations are plotted in the same graph.



From the histograms, it is clear that the marks for the final examination is more dispersed than that for the midterm examination. This illustrates that standard deviation is a valid measure of dispersion.

Please attempt Interactive Exercise Question 2.2.