Sample Percentile & Box plots ⇒

The sample 100p percentile is that data value such that 100p percent of the data are less than or equal to it and 100(1-p) percent of the data are greater than or equal to it. If two data values satisfy this condition, then the sample 100p percentile is the arithmetic average of these two values.

Method for finding Sample percentile ⇒

* First arrange the data in increasing order.

** If np is not an integer, then the data value whose position is the smallest integer exceeding np is the sample 100p percentile.

*** On the otherhand, if np is an integer, then the sample 100p percentile is the average of the values at position np & np+1.

Quartile ⇒ The sample 25 percentile is called the first quartile; the sample 50 percentile is called the sample median or the second quartile; the sample 75 percentile is called the third quartile.

The quartiles break up the data into four parts, with roughly 25 percent of the data being less than the first quartile, 25 percent b/w first and median, 25 percent median, & third quartile & 25 percent being greater than the third quartile.

Problem ⇒ The following data give noise levels measured at 36 different times directly outside of Grand Central station in New York.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 85, 114, 85, 69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

$\frac{89+90}{2}$ $\frac{82+83}{2} = \frac{165}{2}$

(3)

Solution: A stem & leaf plot of the data is as follows:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 6, 5

The first quartile (25 percentile) implies $100p = 25$ so $p = \frac{25}{100} = .25$.

Now $np = 36 \times (.25) = 9$

$np+1 = 9+1 = 10$.

so first quartile = $\frac{75+79}{2}$ = average of 9th & 10th smallest values.

$= \frac{75+77}{2} = 76$.

Second quartile (50 percentile) implies $100p = 50 \Rightarrow p = 0.5$

$np = 36(0.5) = 18$

$np+1 = 18+1 = 19$, so second quartile is 84.5.

Third quartile corresponds $100p = 75$ or $p = \frac{75}{100} = 0.75$.

Then $np = 36(0.75) = 27$

$np+1 = 28$.

So third quartile is the average of 27th & 28th smallest value, which is 104.5.

Box Plot ⇒ A box plot is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis. Imposed on the line is a box which starts at the first quartile & ends at the third one. The value of the second quartile should be indicated by a vertical line inside the box.

Starting Salary	frequency
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
65	2
66	3
70	1

first quartile $\Rightarrow$ implies $p = 0.25$

$42(0.25) = 10.5$

which is not an integer

This first quartile = 60

Second quartile implies $p = 0.5$

so $42(0.5) = 21$. Thus

Second quartile = (average of 21th & 22th value)

$= \frac{61+62}{2} = \frac{123}{2} = 61.5$.

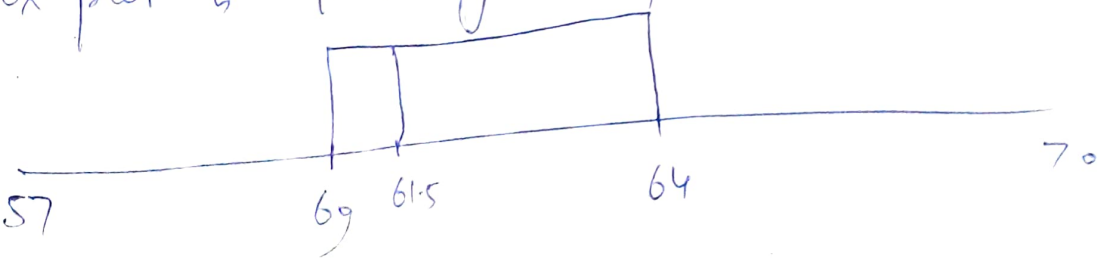
Third quartile implies $p = 0.75$

$np = (42)(0.75) = 31.5$

so 3rd quartile = 64

$n = 42$, so the 42 values go from 57 to 70 on the horizontal line. The value of the first quartile 60, second quartile 61.5 & the third one is 64.

The box plot is then given by

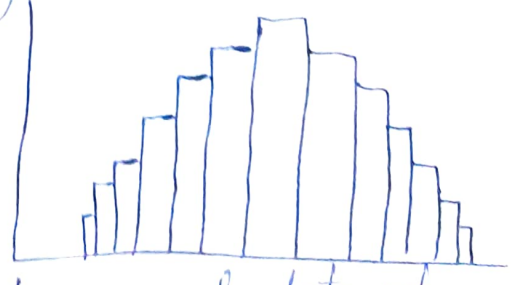


(6)

The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the range of the data. Also, the length of the box (Q3 − Q1), equal to third quartile minus first quartile, is called the interquartile range.

Normal data Sets ⇒ A data set is said to be normal if a histogram describing it has the following properties:

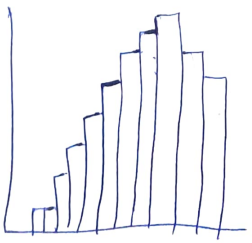
1. It is highest at the middle.
2. Moving from a middle interval in either direction, the height decreases in such a way that the entire histogram is bell-shaped.
3. The histogram is symmetric about its middle interval.



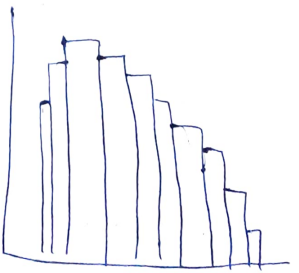
Histogram of a normal data set.

* If the histogram of a data set is close to being a normal histogram, then we say that the data set is approximately normal. For example

Any data set that is not approximately symmetric about its sample median is said to be skewed. It is skewed to the right if it has long tail to the right. Similarly it is skewed to the left if it has long tail to the left.



Skewed to the left



Skewed to the right.

The Empirical Rule ⇒

If a data set is approximately normal with the sample mean $\bar{x}$ & sample standard deviation s, then the following statements are true:

(1) Approximately 68% of the observations lies within $\bar{x} \pm s$.

(2) Approximately 95 percent of the observations lie within $\bar{x} \pm 2s$.

(3) Approximately 99.7% of the observations lie within $\bar{x} \pm 3s$.

Problem: The scores of 25 students on a history examination are listed on the following stem & leaf plot.

9	0, 0, 4
8	3, 4, 4, 6, 6, 9
7	0, 0, 3, 5, 5, 8, 9
6	2, 2, 4, 5, 7
5	0, 3, 5, 8

(9)

Solution:-

$$\bar{x} = \frac{\sum x_i}{n} = [90+90+94+83+84+84+86+86+89 + 70+70+73+75+75+78+79 + 62+62+64+65+67+50+53 + 55+58] / 25 = 73.68$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{n\bar{x}^2}{n-1}} = 12.80$$

$\bar{x} - S = 73.68 - 12.80 = 60.88$

$\bar{x} + S = 73.68 + 12.80 = 86.48.$

Since 17 of the observation lie with in 60.88 & 86.48, the actual percentage $\left(\frac{17}{25}\right)(100) = 68\%$.

Now

$\bar{x} + 2S = 98.28$, $\bar{x} - 2S = 48.08$

with in this interval $[\bar{x} - 2S, \bar{x} + 2S]$

Since 100% of the data falls in this range so the statement of emphirical rule, i.e., 95% of data falls in $[\bar{x} - 2S, \bar{x} + 2S]$ holds. Hence