3.1

Mean:

$$\frac{\text{Sum of observations}}{\text{Number of observations.}}$$

Median:

- Arrange data in increasing order.
- if no of observation is odd, then median is exactly the middle of ordered list.
- if no of observations is even, then median is the mean of two middle observations in ordered list.

$$\frac{n+1}{2}$$

Sample Mean:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$n \longrightarrow$ sample size or ~~number of observation~~

3.2

Range of Data set:

Range = Max − Min

Standard deviation:

Deviation from mean

$x_i - \bar{x}$

Height x	Deviation from mean $x_i - \bar{x}$
72	-3
73	-2
76	1
76	1
78	3

$\bar{x} = \frac{\Sigma x_i}{n} = 75$ inches

sum always equal to zero.

To obtain quantities that do not sum to zero, we square the deviation from mean. The sum of the squared deviations from mean, $\Sigma(x_i - \bar{x})^2$, is called sum of squared deviations.

Height (x)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
72	-3	9
73	-2	4
76	1	1
76	1	1
78	3	9

$\Sigma(x_i - \bar{x})^2 = 24$

Sample standard Deviation

we do by dividing sum of squared deviations by n-1 or 1 less than the sample size. The resulting quantity is called sample variance and is denoted by S_x^2 or when no confusion can arise, s^2

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

⇒ The more variation that there is in a data set, the larger is its standard deviation.

⇒ S.D value can be affected by a few extreme observations.

Computing formula for sample standard deviation

$$s = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/n}{n-1}}$$

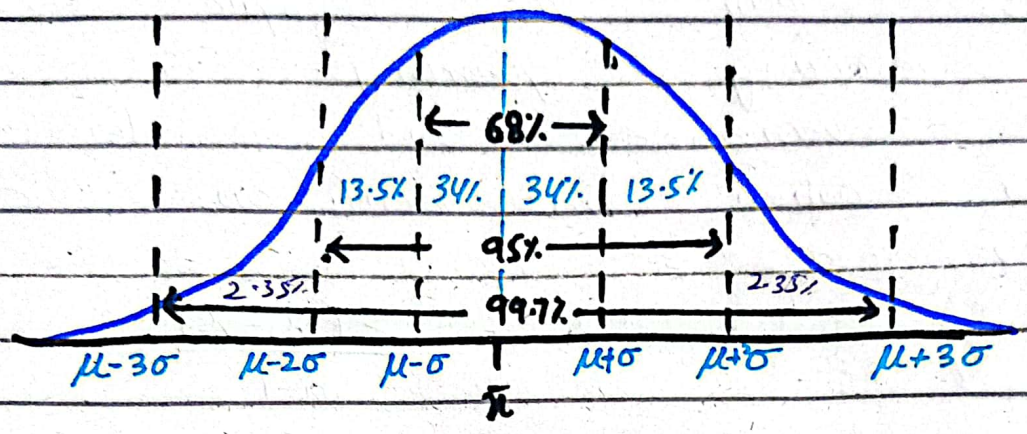
- first compute $(\Sigma x_i)^2/n$
- then $\Sigma x_i^2 - (\Sigma x_i)^2/n$

first square then sum all values
sum observations then square it

both formulas give same answer.

Standard deviation = Measure of variation

Three standard deviations rule



for calculating z-score

$$z = \frac{x - \mu}{\sigma}$$

$\mu \rightarrow$ mean

$\sigma \rightarrow$ standard deviation

$\rightarrow x = \mu + z\sigma$

Ex 1 $X \sim N(50, 10)$

$X \sim N(\mu, \sigma)$

a) $\mu = 50$, $\sigma = 10$

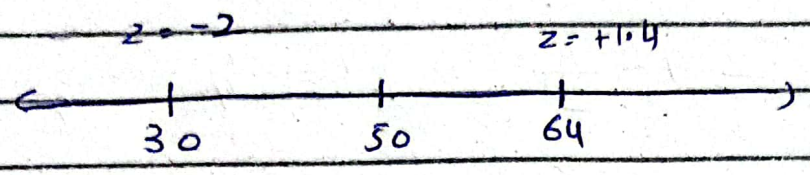
$z = 1.4$, $x = ?$

$$x = \mu + z\sigma$$
$$= 50 + (1.4)(10)$$
$$= 64$$

b) $x = 30$ $z = ?$

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{30 - 50}{10}$$
$$z = -2$$

c) difference between positive and negative z value.



Empirical rule is only useful for distribution and bell shape.

Empirical rule gives us an estimate for how much data falls within 1, 2 or 3 standard deviations of our means. We can only use the empirical rule if our data has a normal distribution, which is bell shapped distribution.

Chebyshev's Theorem

Just as the empirical rule gives us an estimate for how much data fall within 1, 2 or 3 standard deviations of our mean for a bell shaped distribution, chebyshev's Theorem gives us similar estimates for non-normal data.

The proportion of data that lie within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$ for $k > 1$.

$$k = 2 : 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

$$k = 3 : 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} \approx 88.9\%$$

$$k = 4 : 1 - \frac{1}{4^2} = 1 - \frac{1}{16} = \frac{15}{16} = 93.75\%$$

CamScanner

The five number Summary; Boxplots

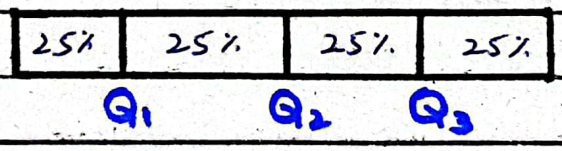
=> Quartiles

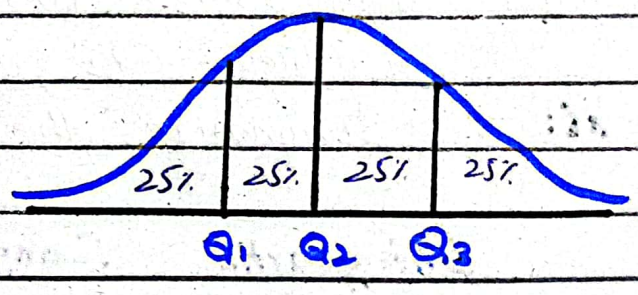
Quartiles are the 25th, 50th, and 75th percentiles and divide data set into quarters (four equal parts)

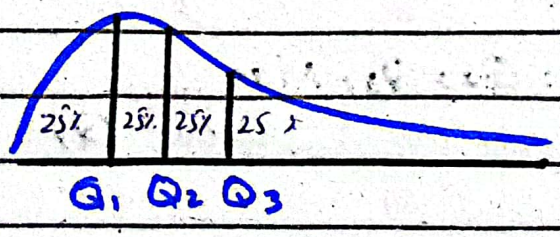
Q_1: Number that divides bottom 25% of data from top 75%

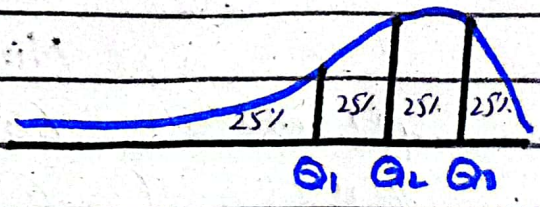
Q_2: is the median, the number that divide bottom 50% of data from top 50%.

Q_3: Number that divides bottom 75% of data from top 25%.









CamScanner

Q. 8 8 9 10 11 | 11 11 11 14 15 (even)

$Q_2 = \frac{10+1}{2} = 5.5 = 11$

$Q_1 = \frac{5+1}{2} = 3 = 9$

$Q_3 = \frac{5+1}{2} = 3 = 11$

Q. 8 8 9 10 11 11 11 11 14 (odd number)

$Q_2 = \frac{n+1}{2} = \frac{9+1}{2} = 5th \rightarrow 11$ (Q_1, Q_2)

$Q_1 = \frac{4+1}{2} = 2.5 = \frac{8+9}{2} = 8.5$

$Q_3 = \frac{4+1}{2} = 2.5 = 11$

Determine the Quartiles

Step 01: Arrange the data in increasing order.

Step 02: After step 01, find the middle number. The middle number is called median or Q_2. Its like midpoint of your data.

Step 03: After finding median, split your list into two halves. If you have an odd number of numbers, include median in both halves. This gives you a bottom half and a top half.

Step 04: Find median of bottom half of data set. This value is first quartile Q_1.

Step 05: Find median of top half of data set. This value is third quartile. Q_3.

Step 06: Summarize the result.

Interquartile Range

$$IQR = Q_3 - Q_1$$

Five number summary

Min, Q_1, Q_2, Q_3, Max

spread of first quarter = $Min - Q_1$

Spread of fourth quarter = $Q_3 - Max$

Outliers

- Outliers are data points that are very different from rest of data.
- They could be due to errors, come from a different group, or just be very extreme.

- For example, Bill gates wealth might be an outlier in a dataset of U.S resident wealth.
- When you see ~~out~~ outliers, we need to figure out why its there.

- if it's mistake or clearly doesnot fit we can remove it, but if not, it's a tough call.

- We can use quartiles and IQR to find potential outliers.

  Lower limit = $Q_1 - 1.5IQR$

  Upper limit = $Q_3 + 1.5IQR$

- Any data points below the lower limit or above the upper limit could be outlier

- To decide if they really are outliers, we need to do more analysis, like making graphs.

CamScanner

Boxplots

Boxplot also called box-and-whisker diagram.

Step 01: Determine quartiles

Step 02: Determine potential outlier and adjacent values

↓

AV are nearest data points to quartiles that are not classified as outliers.

Lower adjacent value is the closest data point to lower quartile (Q1) that is not considered outlier.

Upper adjacent value is closest data point to upper quartile (Q3) that is not considered outlier.

Step 03:-

- Draw a straight horizontal line.
- Mark points on the line to show where quartiles (Q1, Q2, Q3) and adjacent values are located.
- Quartiles are marked with vertical lines, and adjacent value are marked with dot or small lines. This line serves as reference for positioning these values on graph.

step 04: Connect the quantiles to make a box, and then connect the box to adjacent values with lines.

step 05: plot each potential outlier with an asterisk.

Note:

If data set has no potential outlier then adjacent values are just minimum and maximum observation

Finding adjacent values

For lower adjacent value, find nearest data point greater or equal to Q1 but less than or equal to lower limit.

For upper adjacent value, find nearest data point less than or equal to Q3 but greater than or equal to upper limit.

3.5

Descriptive measures for populations:
Use of samples

Population mean

$$\mu = \frac{\Sigma x_i}{N}$$

Sample mean $\longrightarrow \bar{x}$ ✓
population mean $\longrightarrow \mu$ ✓

population standard deviation

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

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

$$\sigma = \sqrt{\frac{\Sigma x_i^2}{N} - \mu^2}$$

$s^2 \longrightarrow$ sample variance
$\sigma^2 \longrightarrow$ population variance

Parameter : descriptive measure for a population

Statistic : descriptive measure for sample

From any variable x, we can form new variable z,

$$z = \frac{x - \mu}{\sigma}$$

It is called standard version of x or standardized variable corresponding to x.

Standarized variable always has mean 0 and standard deviation 1.

Z- score

for observed value of variable x, corresponding value of standardized variable z is called z-score of observation. The term standarized score is often used instead of z-score.

negative z-score indicates that observation is below (less than) the mean.

positive show above the mean