Chapter 3:

Measures of central Tendency and Dispersion.

A data set contain many variables and observations. We are not always interested in each of the measured values but rather in summary which interprets the data. Statistical functions fulfill the purpose of summarizing data in a meaningful yet concise way.

Now, we focus on the most important statistical concepts to summarize data: these are measures of central tendency and variability. The applications of each measure depends on the scale of the variables of interest.

* Measures of Central Tendency:

A natural human tendency is to make comparisons with the average.

e.g: average score, mean temparature in April in Pune, average income, etc.

- Various statistical concepts refer to the average of the data, but the right choice depends upon the nature and scale of the data as well as the objective of the study. We call statistical functions which describe the average or centre of the data location parameters or measures of central tendency.

CamScanner

Arithmetic Mean:

The arithmatic mean is one of the most intuitive me of central tendency. Suppose a variable of size n consists of the value $x_1, x_2, \cdots, x_n$. The arithmetic mean of this data is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \qquad \text{—— (A)}$$

In informal languge, we often speak of the average or just mean when using the formula (A).

Arithmetic mean for grouped data:

Class intervals a_j	$a_1 = e_0 - e_1$	$a_2 = e_1 - e_2$	$\cdots$	$a_k = e_{k-1} - e_k$
Absolute freq. n_j	n_1	n_2	$\cdots$	n_k
Relative freq. f_j	f_1	f_2		f_k.

Note that $a_1, a_2, \cdots, a_k$ are the k class intervals & each interval a_j $(j = 1, 2, \cdots, k)$ contains n_j observations with $\sum_{j=1}^{k} n_j = n$. The relative frequency of the j^{th} class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. The mid-value of the j^{th} class interval is defined as $m_j = \frac{e_{j-1} + e_j}{2}$, which is the mean of the lower and upper limits of the interval. The weighted arithmetic mean for grouped data is defined as

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j.$$

CamScanner

The results of the mean and the weighted mean differ because we use the middle of each class as an approximation of the mean within the class. The implication is that we assume that the value are uniformly distributed within each interval. This assumption is obviously not met. If we had knowlege about the mean in each class, like in this example, we would obtain the correct result.

However, the weighted mean is meant to estimate the arithmetic mean in those situation where only grouped data is available. It is therefore typically used to obtain an approximation of the true mean.

* Properties of the Arithmetic mean:

① The sum of the deviation of each variable around the arithmetic mean is zero.

$$\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

② If the data is linearly transformed as $y_i = a+b x_i$; where a and b are known constant, it holds that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a+b x_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} x_i$$

$$= a+b\bar{x}.$$

temparature, °C we want °F $\boxed{°F = 32 + 1.8 \cdot x\,°C.}$

Median and Quantiles:

The median is the value which divides the observ into two equal parts such that at least 50% of values are greater than or equal to the median and at least 50% of the values are less than or equal to the median. The median is denoted by $\tilde{x}_{0.5}$; then, in terms of the empirical cumulative distribution function, the condition $F(\tilde{x}_{0.5}) = 0.5$ is satisfied.

Consider the n observations $x_1, x_2, \cdots, x_n$ which can be ordered as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The calculations of the median depends on whether the number of observations n is odd or even. When n is odd, then $\tilde{x}_{0.5}$ is the middle ordered value. When n is even, then $\tilde{x}_{0.5}$ is the arithmetic mean of the two middle ordered values.

$$\tilde{x}_{0.5} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even.} \end{cases}$$

e.g: $x_1 \cdots x_{10}$, arrange in order $x_{(1)}, \cdots, x_{(10)}$
Find median.

– If ~~we~~ deal

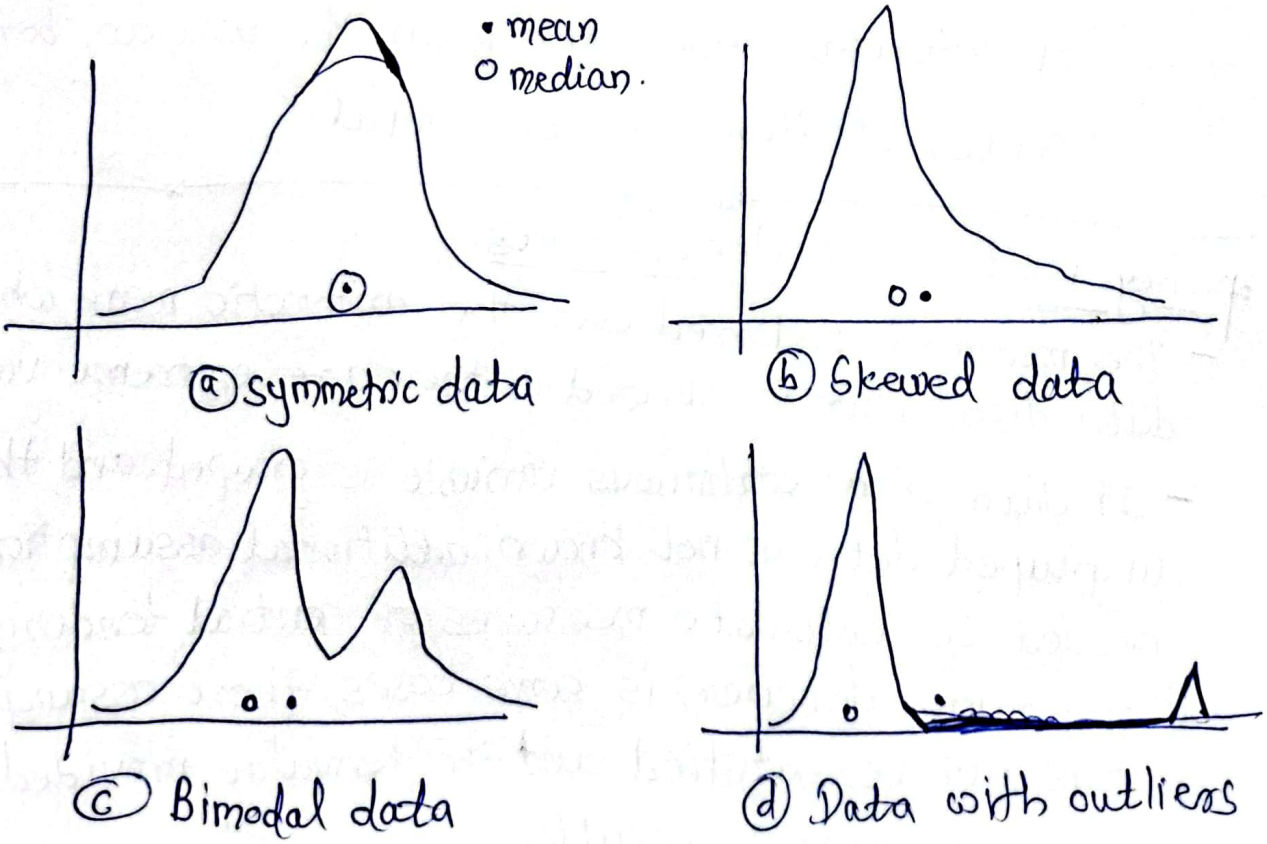
r we deal with grouped data, we can calculate the median under the assumption that the values within each class are equally distributed. Let $K_1, K_2, \ldots, K_k$ be k classes with observations of size $n_1, n_2, \ldots, n_k$, respectively. First, we need to determine which class is the median class, e.e., the class that includes the median. We define the median class as the class K_m for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j \geq 0.5 \quad \text{hold.}$$

Then, we can determine the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m,$$

where e_{m-1} denotes the lower limit of the interval K_m and d_m is the width of the interval K_m.



Comparing mean and median:

In ⓐ, the raw data is summarized by using ticks at the bottom of the graph and by using a kernel density estimator. The mean and the median are similar here because the distribution of the observations is symmetric around the centre. If we have skewed data ⓑ, then the mean and the median may differ. If the data has more than one centre, such as in ⓒ neither the median nor the mean has meaningful interpretations. If we have outliers (@), then it is wise to use the median because the median because the mean is sensitive to outliers.

- These examples show that depending on the situation of interest either the mean, the median, both or neither of them can be useful.

---

* Key points & further Issues:

- The median is preferred over the arithmetic mean when the data distribution is skewed or there are extreme values.
- If data of a continuous variable is grouped, and the original ungrouped data is not known, additional assumptions are needed to calculate measures of central tendency and dispersion. However, is some cases, these assumptions may not be satisfied, and the formulae provided may give imprecise results.

CamScanner

Quantiles:

Quantiles are a generalization of the idea of the median. The median is the value which splits the data into two equal parts. Similarly, a quantiles partitions the data into other proportions.

e.g:, a 25%-quantile splits the data into two parts such that at least 25% of the values are less than or equal to the quantiles and at least 75% of the values are greater than or equal to the quantile.

In general, let α be a number between zero and one. The $(\alpha \times 100)\%$-quantile, denoted as $\tilde{x}_\alpha$, is defined as the value which divides the data in proportions of $(\alpha \times 100)\%$ and $(1-\alpha) \times 100\%$ such that at least $\alpha \times 100\%$ of the values are less than or equal to the quantile and at least $(1-\alpha) \times 100\%$ of the values are greater than or equal to the quantile.

- In terms of the empirical cumulative distribution function, we can write $F(\tilde{x}_\alpha) = \alpha$.

- It follows immediately that for n observations, at least $n\alpha$ values are less than or equal to $\tilde{x}_\alpha$ and at least $n(1-\alpha)$ observations are greater than or equal to $\tilde{x}_\alpha$.

CamScanner

- The median is the 50% quantiles $\tilde{x}_{0.5}$.

- If α takes the values $0.1, 0.2, \ldots, 0.9$, the quantiles are called deciles.

- If $\alpha \cdot 100$ is an integer number (e.g. $\alpha \times 100 = 95$), the quantiles are called percentiles, i.e., the data is divided into 100 equal parts.

- If α takes the value 0.2, 0.4, 0.6 & 0.8, the quantiles are known as quaintiles and they divide the data into five equal parts.

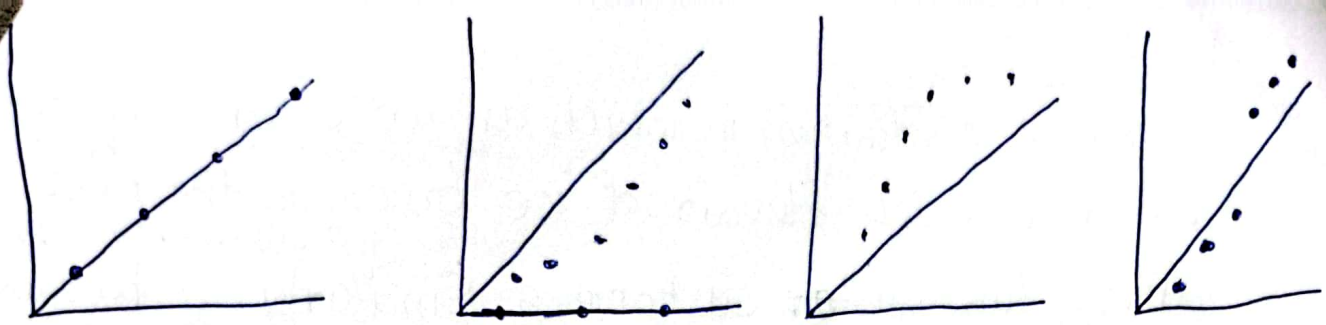
- If α takes the values 0.25, 0.5 and 0.75, the quantiles are called quartiles.

Consider n ordered observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
The $\alpha \cdot 100\%$-quantile $\tilde{x}_\alpha$ is calculated as

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer number, choose } k \text{ as the smallest integer} > n\alpha. \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{if } n\alpha \text{ is an integer} \end{cases}$$

* Quantile - Quantile Plots (QQ Plots).

If we plot the quantiles of two variables against each other, we obtain a Quantile-Quantile plot (QQ-Plot). This provides a simple summary of whether the distributions of the two variables are similar with respect to their location or not.



ⓐ setting 1 ⓑ Setting 2 ⓒ Setting 3 ⓓ Setting 4

As a summary, let us consider four important patterns:

ⓐ If all the pairs of quantiles lie (nearly) on a straight line at an angle of 45° from the x-axis, then the two samples have similar distributions.

ⓑ If the y-quantiles are lower than the x-quantiles, then y-values have a tendency to be lower than x-values.

ⓒ If the x-quantiles are lower than the y-quantiles, then the x-values have a tendency to be lower than the y-values.

ⓓ If QQ plot is like ⓓ, it indicates that there is a break point upto which the y-quantiles are lower than the x-quantiles and after that point, the y-quantiles are higher than the x-quantiles.

CamScanner

Mode:

Consider a situation in which an ice cream shop owner to know which flavour of ice cream is the most popu among his ~~custom~~ customers. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand. To answer this type of question, one can use the mode which is another measure of central tendency.

- The mode $\bar{x}_M$ of n observations $x_1, x_2, x_3, \cdots, x_n$ is the value which occurs the most compared with all other value, i.e., which has maximum absolute frequency. It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined.
- A formal definition of the mode is

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, n_2, \cdots, n_K\}.$$

The mode is typically applied to any type of variable for which the number of different values is not too large. If continuous data is summarized in groups, then the mode can be used as well.

ometric Mean:

Consider n observations $x_1, x_2, \ldots, x_n$ which are all positive and collected on a quantitative variable. The geometric mean $\bar{x}_G$ of this data is defined as

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}.$$

Time t	Inventory B_t	Growth factor x_t	Growth rate r_t
0	B_0		
1	B_1	$x_1 = B_1/B_0$	$((x_1-1)\cdot 100)\%$
2	B_2	$x_2 = B_2/B_1$	$((x_2-1)\cdot 100)\%$
⋮			
T	B_T	$x_T = \frac{B_T}{B_{T-1}}$	$((x_t-1)\,100)\%$.

The ratio of B_t and B_{t-1}, $x_t = \frac{B_t}{B_{t-1}}$ is called the t^{th} growth ~~rate~~ factor.

The growth rate r_t is defined as $r_t = ((x_t-1)\cdot 100)\%$ & gives us an idea about the growth or decline of our value at time t.

Geometric mean $\bar{x}_G = \sqrt[T]{x_1 x_2 \cdots x_T}$

Thus, B_t at time t can be calculated as

$$B_t = B_0 \cdot \bar{x}_G^t$$

Harmonic Mean:

The harmonic mean is typically used whenever di... x_i contributes to the mean with a diffrent weight ... i.e., when we implicitly assume that the weight of each x_i is not one. It can be calculated as

$$\bar{x}_H = \frac{w_1 + w_2 + \cdots + w_k}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_k}{x_k}} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} \frac{w_i}{x_i}}$$

e.g: When calculating the average speed, each weight relates to the relative distance travelled, n_i/n, with speed x_i. Using $w_i = n_i/n$ & $\sum_i w_i = \sum_i \frac{n_i}{n} = 1$, the harmonic mean can be written as

$$\bar{x}_H = \frac{1}{\sum_{i=1}^{k} \frac{w_i}{x_i}}.$$

CamScanner

asures of Dispersion:

Measures of central tendency, as introduced earlier, give us an idea about the location where most of the data is concentrated. However, two diffrent data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have diffrent concentrations around the mean.

- In this case, the location measures may not be adequate enough to describe the distribution of the data.
- The concentration or dispersion of observations around any particular value is another property which characterizes the data and it's distribution.

Now, we introduce statistical methods which describe the variability or dispersion of data.

* Range & Interquartile Range:

Consider a variable X with n observations $x_1, x_2, \cdots, x_n$. Order these n observations as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Range: The range is a measure of dispersion defined as the difference between the maximum and minimum

value of the data as

$$R = x_{(n)} - x_{(1)}.$$

Interquartile Range:

The interquartile range is defined as the difference between the 75^{th} and 25^{th} quartiles as

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

It covers the centre of distribution and contains 50% of the observations.

* Absolute Deviation, Variance and Standard Deviation:

Another measure of dispersion is the variance. The variance is one of the most important measures in statistics & is needed throughout this course.

— We use the idea of absolute deviation to give some more background and motivation for understanding the variance as a measure of dispersion.

Consider the deviation of n observations around a certain value A and combine them together, for instance, via the arithmetic mean of all the deviations:

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - A) \qquad \text{——— (A)}$$

This measure has the drawback that the deviations $(x_i - A)$, $i = 1, 2, \cdots, n$, can be either positive or negative and consequently their sum can potentially

very small or even zero. Thus, using D as a measure of variability is therefore not a good idea since D may be small even for a large variability in the data.

- Using absolute values of the deviations solves this problem and we introduce the following measure of dispersion.

$$D(A) = \frac{1}{n} \sum_{i=1}^{n} |x_i - A|. \quad \text{——— (B)}$$

- It can be shown that the absolute deviation attains it's minimum when A corresponds to the median of the data

$$D(\tilde{x}_{0.5}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}_{0.5}|. \quad \text{——— (C)}$$

* We call $D(\tilde{x}_{0.5})$ the <u>absolute median deviation.</u>

* When $A = \bar{x}$, we called <u>absolute mean deviation</u> given by $D(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$. ——— (D)

= Another solution to avoid the positive and negative signs of deviation in (A) is to consider the squares of deviations $x_i - A$, rather than using the absolute value. This provides another measure of dispersion as

$$s^2(A) = \frac{1}{n} \sum_{i=1}^{n} (x_i - A)^2$$

which is known as the mean squared error (MSE) w.r.t. A.

- The MSE (mean squared error) is another import measure is statistics, see Unit IV.

- It can be shown that $S^2(A)$ attains it's minimum when $A = \bar{x}$. This is the (sample) variance

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

After expanding, we get

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

The positive square root of the variance is called the (sample) standard deviation, defined as

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

- The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations.
e.g: if X is distance, measured in km, then $\bar{x}$ & $\bar{s}$ are also measured in km, whereas $\tilde{S}^2$ is measured in km^2 (which may be difficult to interpret).

* The variance is a measure which we use in next units to obtain measures of association between variables and to draw conclusions from a sample about a population of interest.

CamScanner

standard deviation is typically preferred for a descriptive summary of the dispersion of data.

- The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean.
- A low value of the standard deviation indicates that the value are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may be far away from the mean.
- If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median. In such a case, the absolute median deviation © may be preferred over the standard deviation.

exa: Suppose three student $S1, S_2, S_3$ arrive at different times in the class to attend their lectures. Let us look at their arrival time in the class after or before the starting time of lecture, i.e., let us look how early or late they were (in minutes)

Week	1	2	3	4	5	6	7	8
S1								
S2								
S3								

week	1	2	3	4	5	6	7	8	9	
M1	0	0	0	0	0	0	0	0	0	
M2	-10	+10	-10	+10	-10	+10	-10	+10	-10	+
M3	3	5	6	2	4	6	8	4	5	7

We calculate the variance & absolute median deviation:

$$\tilde{s}_{M1} = \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{1}{10}\left((10-0)^2+\cdots+(0-0)^2\right) = 0$$

$$\tilde{s}_{M2} = \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{1}{10}\left((-10-0)+\cdots+(10-0)^2\right) \approx 111.1$$

$$s_{M3} = \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 \approx 3.3 \quad (\because \bar{x}=5)$$

$$D(\tilde{x}_{0.5}, M1) = \frac{1}{10}\sum_{i=1}^{10}|x_i-\tilde{x}_{0.5}| = \frac{1}{10}(|10-0|+\cdots+|10-0|)$$
$$= 0$$

$$D(\tilde{x}_{0.5}, M2) = \frac{1}{10}\sum_{i=1}^{10}|x_i-\tilde{x}_{0.5}| = \frac{1}{10}(|-10-0|+\cdots+|10-0|)$$
$$= 10$$

$$D(\tilde{x}_{0.5}, M3) = \frac{1}{10}\sum_{i=1}^{10}|x_i-\tilde{x}_{0.5}| = (|3-5|+\cdots+|7-5|)\frac{1}{10}$$
$$= 1.4$$

- Using the arithmetic mean, we concluded that both M_1, M_2, M_3 arrive on time, whereas M_3 is always late. But we saw the variation of arrival times differs substaintially among the three students.
- We observe that the variation/dispersion/variability is the lowest for M_1 & higest for M_2. Both median absolute deviation and variance allow a comparison between the two students.

CamScanner

A hiking enthusiast has a new app for his smartphone which summarizes his hikes by using a GPS device. Let us look at the distance hiked (in km) and maximum altitude (in m) for the last 10 hikes.

Distance	12.5	29.9	14.8	18.7	7.6	16.2	16.5	27.4	12.1	17.5
Altitude	342	1245	502	555	398	670	796	912	238	461

(a) Calculate the arithmetic mean and median for both distance and altitude. $\bar{x}_D = 17.32$ | $\tilde{x}_{0.5,D} = 16.35$; $\bar{x}_A = 612.4$ | $\tilde{x}_{0.5,A} = 528.5$

(b) Determine the arithmetic mean & median for the distance and the altitude variables. (Discuss the shape of the distribution given the results of (a) & (b).
$\tilde{x}_{0.25,D} = 12.5$, $\tilde{x}_{0.25,A} = 398$, $\tilde{x}_{0.75,D} = 18.85$, $\tilde{x}_{0.75,A} = 796$

(c) Calculate the interquartile range, absolute median, deviation & standard deviation for both variables. What is your conclusion about the variability of the data? $d_{\tilde{x},A} = 398$, $d_{\tilde{x},D} = 6$ = absolute median deviation
$D_D(\tilde{x}_{0.5}) = 4.68$
$D_A(\tilde{x}_{0.5}) = 223.2$
$\tilde{s}_D^2 \approx 41.5$, $\tilde{s}_D = \sqrt{41.5}$; $\tilde{s}_A^2 = 82314$, $\tilde{s}_A = \sqrt{82314}$

(d) One metre corresponds to approximately 3.28 ft. What is the average altitude when measured in feet rather than in meters. $\bar{y} = 3.28 \times 528.5 = 1722.48$

(e) Assume distance is measured as only short (5-15 km) moderate (15-20 km) & long (20-30 km). Summarize the ~~data~~ grouped data in a frequency table. Calculate the weighted arithmetic mean under the assumption that the raw data is not known.

CamScanner

Class intervals	(5;15]	(15;20]	(20;30]
n_j	4	4	2
f_j	4/10	4/10	2/10
Σf_j	4/10	8/10	1

$$\bar{x} = \sum_{j} f_j m_j = 16$$

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 16.25$$

(2) Suppose maximum temparature during the day (in degree Celsius) for ~~December~~ July 1-31 as follows.

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

Samarize this data into three categories. Arithmetic mean = 26.48°c

Class intervals	(0-25]	(25,30]	(30,35]
Absolute freq.	$n_1 = 12$	$n_2 = 18$	$n_3 = 1$
Relative freq.	$f_1 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_3 = \frac{1}{31}$

Weighted arithmetic mean : $\bar{x} = \sum_{j=1}^{k} f_j m_j \approx 25.7$

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m \approx 25.97$$

for grouped data (median). / median = 26.

$\tilde{x}_{0.25} = 25$, $\tilde{x}_{0.5} = 26$, $\tilde{x}_{0.75} = 29$

The following frequency table gives the values obtained in 40 rolls of a die

Value	frequency	
1	9	9
2	8	7
3	5	4
4	5	4
5	6	4
6	7	12

Find ⓐ mean 3.05
ⓑ median — 3
ⓒ mode → 1
ⓓ

20th & 21st value average

$$\text{Mean} = \frac{1\times9 + 2\times7 + 3\times4 + 4\times4 + 5\times4 + 6\times12}{40}$$

$$= \frac{9+14+12+16+20+72}{40}$$

$$= \frac{143}{40}$$

$$= 3.575.$$

1-2	3-4	5-6
16	8	16

[1,2]	(2,4]	(4,6]
16	8	16

$$\frac{1.5\times16 + 3.5\times8 + 5.5\times16}{40}$$

$1.5\times16 + 3\times8 + 5\times16$

$= 128$

$\frac{128}{40}$

$= 3.2$

Weighted Mean $= \frac{140}{40} = \frac{14}{4} = 3.5$

Median $= \frac{3+4}{2} = 3.5$, Mode - 6

CamScanner