Measure of Dispersion

Md. Sabbir Hossain

MNS, BRAC UNIVERSITY

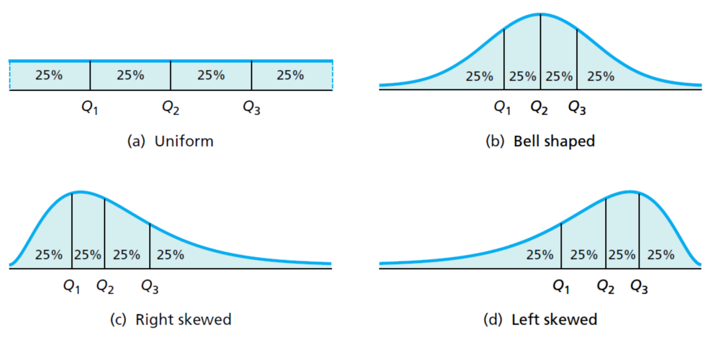
June 28, 2023

Outline

- Quartiles, Deciles, Percentiles
  - Quartiles
  - Percentile & Decile
- Measure of Dispersion
  - Range
  - Inter-quartile Range
  - Mean Deviation from mean
  - Variance
  - Standard Deviation
- Relative measure of dispersion

Quartiles, Deciles, Percentiles

Quartiles



(a) Uniform

(b) Bell shaped

(c) Right skewed

(d) Left skewed

- The first quartile, Q_1, is the number that divides the bottom 25% of the data from the top 75%;
- The second quartile, Q_2, is the median, which, as you know, is the number that divides the bottom 50% of the data from the top 50%; and
- The third quartile, Q_3, is the number that divides the bottom 75% of the data from the top 25%.

Quartiles

- The first quartile (Q_1) is the median of the bottom half of the data set,
- the second quartile (Q_2) is the median of the entire data set and
- the third quartile (Q_3) is the median of the top half of the data set.

To Determine the Quartiles

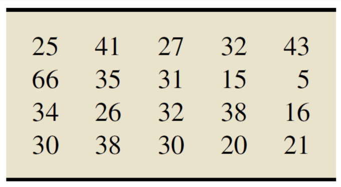
- Step 1 Arrange the data in increasing order.
- Step 2 Find the median of the entire data set. This value is the second quartile, Q_2.
- Step 3 Divide the ordered data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

Quartiles

To Determine the Quartiles

- Step 4 Find the median of the bottom half of the data set. This value is the first quartile, Q_1.
- Step 5 Find the median of the top half of the data set. This value is the third quartile, Q_3.
- Step 6 Summarize the results.

- A sample of 20 people yielded the weekly viewing times, in hours, displayed in the table. Determine and interpret the quartiles for these data.



25	41	27	32	43
66	35	31	15	5
34	26	32	38	16
30	38	30	20	21

Quartiles: Example 1 (n is even)

- Step 1: Arrange the data in increasing order.

  5 15 16 20 21 25 26 27 30 30 31 32 32 34 35 38 38 41 43 66

- Step 2: Find the median of the entire data set. This value is the second quartile, Q_2.

  The number of observations is 20 and, consequently, the median is at position $(20+1)/2 = 10.5$, halfway between the tenth and eleventh observations in the ordered list. Thus, the median of the entire data set is $(30+31)/2 = 30.5$. So, $Q_2 = 30.5$.

- Step 3: Divide the ordered data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

  Referring to the ordered list in Step 1, we see that the bottom half and top half of the data set are as follows:

$$\underbrace{5\ 15\ 16\ 20\ 21\ 25\ 26\ 27\ 30\ 30}_{\text{Bottom half}} \qquad \underbrace{31\ 32\ 32\ 34\ 35\ 38\ 38\ 41\ 43\ 66}_{\text{Top half}}$$

Quartiles: Example 1 (n is even) (Cont...)

- Step 4: Find the median of the bottom half of the data set. This value is the first quartile, Q_1.

  Referring to Step 3, we see that the bottom half of the data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations in the ordered list. Thus, the median of this data setand hence the first quartile is $(21 + 25)/2 = 23$; that is, $Q_1 = 23$.

- Step 5: Find the median of the top half of the data set. This value is the third quartile, Q_3.

  Referring again to Step 3, we see that the top half of the data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations in the ordered list. Thus, the median of this data setand hence the third quartile is $(35 + 38)/2 = 36.5$; that is, $Q_3 = 36.5$.

Quartiles: Example 1 (n is even) (Cont...)

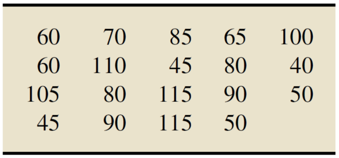
- Step 6: Summarize the results.

  In summary, then, the three quartiles for the TV-viewing times in Table are $Q_1 = 23$ hours, $Q_2 = 30.5$ hours, and $Q_3 = 36.5$ hours.

- Interpretation: We see that 25% of the TV-viewing times are less than 23 hours, 25% are between 23 hours and 30.5 hours, 25% are between 30.5 hours and 36.5 hours, and 25% are greater than 36.5 hours.

Quartiles: Example 2 (n is odd)

- From Tropical Cyclone Reports, published by the National Hurricane Center, we obtained the data shown in the table on maximum wind speeds, in miles per hour (mph), for one years tropical cyclones in the Atlantic Basin. Determine and interpret the quartiles for these data.



60	70	85	65	100
60	110	45	80	40
105	80	115	90	50
45	90	115	50	

- Step 1: Arrange the data in increasing order.
  40 45 45 50 50 60 60 65 70 80 80 85 90 90 100 105 110 115 115

Quartiles: Example 1 (n is odd)

- Step 2: Find the median of the entire data set. This value is the second quartile, Q_2.

  The number of observations is 19 and, consequently, the median is at position $(19+1)/2 = 10$, the tenth observation in the ordered list. Thus, the median of the entire data set is 80. So, $Q_2 = 80$.
- Step 3: Divide the ordered data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

  Because the number of observations is 19, which is an odd number, we include the median in both the bottom and top halves of the data set. So, we see by referring to the ordered list in Step 1 that the bottom half and top half of the data set are as follows:

$$\underbrace{40\ 45\ 45\ 50\ 50\ 60\ 60\ 65\ 70\ 80}_{\text{Bottom half}} \quad \underbrace{80\ 80\ 85\ 90\ 90\ 100\ 105\ 110\ 115\ 115}_{\text{Top half}}$$

Quartiles: Example 1 (n is odd) (Cont...)

- Step 4: Find the median of the bottom half of the data set. This value is the first quartile, Q_1.

  Referring to Step 3, we see that the bottom half of the data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations in the ordered list. Thus, the median of this data setand hence the first quartile is $(50 + 60)/2 = 55$; that is, $Q_1 = 55$.

- Step 5: Find the median of the top half of the data set. This value is the third quartile, Q_3.

  Referring again to Step 3, we see that the top half of the data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations in the ordered list. Thus, the median of this data set and hence the third quartile is $(90 + 100)/2 = 95$; that is, $Q_3 = 95$.

Quartiles: Example 2 (n is odd) (Cont...)

- Step 6: Summarize the results.

  In summary, the three quartiles for the maximum wind speeds in the table are $Q_1 = 55$ mph, $Q_2 = 80$ mph, and $Q_3 = 95$ mph.

- Interpretation: We see that roughly 25% of the maximum wind speeds are less than 55 mph, roughly 25% are between 55 mph and 80 mph, roughly 25% are between 80 mph and 95 mph, and roughly 25% are greater than 95 mph

Determination of Quartile: Another method

Quartile

- For raw data,
  - if $i = 1, 2, 3$
  - $n =$ number of values
- when $\frac{i \times n}{4}$ is an integer

$$i^{th} quartile, Q_i = \frac{1}{2}\left[\left(\frac{i \times n}{4}\right)^{th} \text{ value} + \left(\frac{i \times n}{4} + 1\right)^{th} \text{ value}\right],$$

- when $\frac{i \times n}{4}$ is not an integer

$$i^{th} quartile, Q_i = \text{next integer value of } \frac{i \times n}{4},$$

Percentile & Decile

- *pth* percentile is the number that divides the bottom p% of the data from the top (100 - p)%.
- We use the letter P, subscripted with the percent in question, to denote a percentile.
- For example, the 30th percentile is denoted P_{30} and is the number that divides the bottom 30% of the data from the top 70%; the $2.5th$ percentile is denoted $P_{2.5}$ and is the number that divides the bottom 2.5% of the data from the top 97.5%. Note that the median is also the 50th percentile.
- Certain percentiles are particularly important: the 10th, 20th,..., 90th percentiles are called the deciles and divide a data set into tenths (10 equal parts).
- Note: Median $= Q_2 = D_5 = P_{50}$

Decile

Percentile

- For raw data,
  - if $i = 1, 2, 3, \ldots, 9($ for decile)
  - $n =$ number of values
- when $\frac{i \times n}{10}$ is an integer

$$i^{th} decile, D_i = \frac{1}{2}\left[\left(\frac{i \times n}{10}\right)^{th} \text{ value} + \left(\frac{i \times n}{10} + 1\right)^{th} \text{ value}\right],$$

- when $\frac{i \times n}{10}$ is not an integer

$$i^{th} decile, D_i = \text{next integer value of } \frac{i \times n}{10}.$$

Inspiring Excellence

Percentile

Decile

- For raw data,
  - if $i = 1, 2, 3, \ldots, 99$(for percentile)
  - $n =$ number of values
- when $\frac{i \times n}{100}$ is an integer

$$i^{th} percentile, P_i = \frac{1}{2}\left[(\frac{i \times n}{100})^{th} \text{ value} + (\frac{i \times n}{100} + 1)^{th} \text{ value}\right],$$

- when $\frac{i \times n}{100}$ is not an integer

$$i^{th} percentile, P_i = \text{next integer value of } \frac{i \times n}{100}.$$

Example

- Following is the marks in STA-201 obtained by 20 students in Fall 2022.

99	75	84	33	45	66	97	69	55	61
72	91	74	93	54	76	62	91	77	68

- Find out
  - 1^{st} quartile & 3^{rd} quartile
  - 3^{rd} decile, 6^{th} decile
  - 20^{th} percentile, 37^{th} percentile & 86^{th} percentile

Solution

- First, arrange the data

33	45	54	55	61	62	66	68	69	72
74	75	76	77	84	91	91	93	97	99

- Here, $n = 20$
- $1^{st} quartile$: for $i = 1,\ \frac{i \times n}{4} = \frac{1 \times 20}{4} = 5,$ an integer
- so,

$$\begin{aligned} 1^{st} \text{ quartile} &= \frac{1}{2}\left[5^{th} \text{ value} + (5+1)^{th} \text{ value}\right] \\ &= \frac{1}{2}(61 + 62) = 61.5 \end{aligned}$$

Solution

- 6^{th} Decile : for $i = 6,\ \frac{i \times n}{10} = \frac{6 \times 20}{10} = 12,$ an integer
- so,

$$
\begin{aligned}
6^{th} \text{ Decile}, D_6 &= \frac{1}{2}\left[12^{th} \text{ value} + (12+1)^{th} \text{ value}\right] \\
&= \frac{1}{2}(75 + 76) = 75.5
\end{aligned}
$$

- 37^{th} percentile : for $i = 37,\ \frac{i \times n}{100} = \frac{37 \times 20}{100} = 7.4$
- not an integer. So, next integer $= 8$.

$$37^{th} \text{ percentile}, P_{37} = 8^{th} \text{ value} = 68$$

Measure of Dispersion

Dispersion: Concepts

Consider the height of the basketball players of two teams

Team 1:	72	73	76	76	78
Team 2:	67	72	76	76	84

Summary

- The measures of the central tendency of the two data sets:

	Team 1	**Team 2**
Mean	75	75
Median	76	76
Mode	76	76

Measures of Dispersion

Dispersion

- We can see that both teams have the same mean, median and mode but the two teams are definitely not same.
- The characteristic that make the two data sets different seems to be the spread of the two sets.
- This spread of the data is known as the dispersion of the data.
- The measures of dispersion are of very much importance in statistical studies.

Dierent Measures of Dispersion

- The following measures of dispersion are widely used: 1. Range 2. Inter-quartile range 3. Mean deviation from mean 4. Variance 5. Standard deviation.

Range

- The range of a set of numbers is the difference between the largest and smallest numbers in the set.
- As a primary and quick idea of dispersion, range is of great use.
- Statistically, range, by taking into consideration only the two observations (minimum and maximum) is rated as a crude measure of dispersion.
- The measure can be influenced very badly by only one extreme observation.

- For a set of observations $x_1, x_2, \ldots, x_n$, the following is the formula for the Range,
  Range = Highest observation - lowest observation

The Inter-quartile Range

- The Inter-quartile Range of a set of numbers is the difference between the third and the first quartile of the data set.
- In other words, the Inter-quartile Range is the simple range of the central 50% of the ordered data. Inter quartile range, by avoiding the 25% edge observations from both side of the data becomes free from the influence of extreme observations, but it does not reflect the total available information.

- For a set of observations $x_1, x_2, \ldots, x_n$, the following is the formula for the Inter-quartile Range (IQR),

$$\text{IQR} = Q_3 - Q_1,$$

where $Q_3 = \text{Third Quartile}, Q_1 = \text{First Quartile}$

The Mean Deviation from mean

- Mean deviation from mean (MD) is the arithmetic mean of the absolute deviations of each of values from the mean of the data.
- MD, unlike range and IQR includes all the observations in its computation and as a result is not affected by the extreme observations.
- But the complicacy in its algebraic manipulation mainly kept it out of much wider application.

- For a set of observations $x_1, x_2, \ldots, x_n$, the formula for the Mean Deviation from mean,

$$\text{MD} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|.$$

The Variance

- Variance has almost all the desirable properties of a good measure of dispersion.
- It involves all the observations, It has nice mathematical expression and can easily be used in algebraic manipulation.
- It does not suffer from any influence of extreme values.
- Because of squaring the deviation, the units of measurement of the variance differ from units of measurement original observation.

- For a set of observations $x_1, x_2, \ldots, x_n$, the following is the formula for the Variance,

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

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

The variance: Formula

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)\\
&= \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\bar{x}+n\bar{x}^2)\\
&= \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2)[\because \frac{\sum_{i=1}^{n}x_i}{n}=\bar{x}]\\
&= \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)\\
&= \frac{1}{n-1}(\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n})
\end{aligned}$$

The Standard Deviation

- The sample Standard Deviation (SD), denoted by S, is the positive square root of the sample variance.
- Sample standard deviation includes all the observations and algebraically very easily manipulated, in addition it does not suffer from the problem of units of measurement like variance.
- Although being considered as the best measure of dispersion, standard deviation is strongly affected by extreme observations, and is hardly applicable in comparing different data sets.

- For a set of observations $x_1, x_2, \ldots, x_n$, the following is the formula for the Variance,

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

The Standard Deviation

Properties of Standard Deviation

- It describes the square root of the mean of the squares of all deviations in a data set and is also called the root-mean-square deviation.
- The smallest value of the standard deviation is 0 since it cannot be negative.
- When the data values of a group are similar, then the standard deviation will be very low or close to zero.
- But when the data values vary with each other, then the standard variation is high or far from zero.

Variance and Standard Deviation from Population Data

- For a set of observations $x_1, x_2, \ldots, x_n$,that constitute the whole population with a mean μ, the following formulas for the Variance and Standard Deviation (SD) are used:

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

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

Note: S^2 and S are often called the sample variance and sample standard deviation, but commonly stated by only variance and standard deviation.

Why divide by n-1 in sample variance instead of n

- The reason we use $n-1$ rather than n is so that the sample variance (S^2) will be what is called an unbiased estimator of the population variance σ^2.
- To explain what this means, we first define the term estimator.
- An estimator is a statistic used for the purpose of estimating an unknown parameter. An estimator is a function of the data in a sample.. Example: sample mean $\bar{x}$ is an estimator of population mean μ.
- Note that the concepts of estimate and estimator are related but not the same: a particular value (calculated from a particular sample) of the estimator is an estimate.
- An unbiased estimator is an estimator whose expected value (i.e., the mean of the distribution of the estimator) is the parameter being estimated.

$$E(\bar{x}) = \mu$$

$$E(S^2) = \sigma^2$$

Example: Measures of Dispersion

Example

- Marks obtained by 8 students in a class test: 10, 19, 12, 21, 18, 20, 11, and 19. Find all measures of dispersion.

Range

$$\begin{aligned}\text{Range}, R &= \text{Highest observation - Lowest observation}\\ &= 21 - 10 = 11\end{aligned}$$

IQR

$$\begin{aligned}\text{IQR}, &= Q_3 - Q_1\\ &= ?\end{aligned}$$

Example: Measures of Dispersion

Mean deviation from Mean

First, we have to calculate mean,

$$\bar{x} = \frac{(10 + 19 + 12 + 21 + 18 + 20 + 11 + 19)}{8} = 16.25$$

Then We construct the following Table:

x	$\lvert x_i - \bar{x}\rvert$	**x**	$\lvert x_i - \bar{x}\rvert$
10	6.25	18	1.75
19	2.75	20	3.75
12	4.25	11	5.25
21	4.75	19	2.75
Total			31.5

Mean Deviation from mean, MD $= \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert = \frac{1}{8}\times 31.5 =$ 3.9375.

Example: Measures of Dispersion

Variance and Standard deviation

We construct the following Table:

x	$\left(x_i - \bar{x}\right)^2$	**x**	$\left(x_i - \bar{x}\right)^2$
10	39.0625	18	3.0625
19	7.5625	20	14.0625
12	18.0625	11	27.5625
21	22.5625	19	7.5625
Total			139.5

$$\text{Variance,}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{8-1} \times 139.5 = 19.9286.$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{19.9286} = 4.4641.$$

Example: Measures of Dispersion

Variance and Standard deviation from Computational formula

We construct the following Table:

x	x^2	**x**	x^2
10	100	18	324
19	361	20	400
12	144	11	121
21	441	19	361
Total		130	2252

$$\text{Variance,} S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x\Big)^2\right]$$

$$= \frac{1}{8-1}\left[2252 - \frac{1}{8}(130)^2\right] = 19.9286.$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{19.9286} = 4.4641.$$

Relative measure of dispersion

Relative measure of dispersion

Coefficient of Variation

- The measures of dispersion like standard deviation are not helpful in comparing variability of two different variables.
- Difference in the units of measurement and difference in the magnitude of values are the main reason for this.
- That is why a relative measure of dispersion which is free of units of measurement and also of magnitude of the data values is required.
- Coefficient of variation is one such measure.

RRelative measure of dispersion (Cont.)

Coefficient of Variation

The ratio of the standard deviation to the arithmetic mean, expressed in percent is known as Coefficient of variation (CV). In terms of formula,

$$CV = \frac{S}{\bar{x}} \times 100$$

Coefficient of variation (CV) can be interpreted as the variation per unit mean of a data.

Relative measure of dispersion

Coefficient of Variation

- The coefficient of variation shows the extent of variability of data in a sample in relation to the mean of the population.
- Advantage The advantage of the CV is that it is unitless. This allows CVs to be compared to each other in ways that other measures, like standard deviations or root mean squared residuals, cannot be measured.
- Dis-advantage Unlike the standard deviation, it cannot be used directly to construct confidence intervals for the mean.

Example: Coefficient of Variation

Coefficient of Variation

- The weekly sales of two products A and B were recorded as follow :

Product A	59	75	27	63	27	28	56
Product B	150	200	125	310	330	250	225

Find out which of the two products shows greater fluctuation in sales.

Example: Coefficient of Variation

Product A		Product B	
x	x^2	y	y^2
59	3481	150	22500
75	5625	200	40000
27	729	125	15625
63	3969	310	96100
27	729	330	108900
28	784	250	62500
56	3136	225	50625
$\sum x = 335$	$\sum x^2 = 18453$	$\sum y = 1590$	$\sum y^2 = 396250$

Here, $\bar{x} = 47.86$, $\bar{y} = 227.14$, $S_x = 20.09$, $S_y = 76.48$.
Now,

$$CV_x = \frac{S_x}{\bar{x}} \times 100 = \frac{20.09}{47.86} \times 100 = 41.98\%$$

$$CV_y = \frac{S_y}{\bar{y}} \times 100 = \frac{76.48}{227.14} \times 100 = 33.67\%$$

Product A has more fluctuations than product B.

References and Acknowledgement

- Reference Books:
  1. Introductory STATISTICS by Neil A. Weiss, Ph.D.
  2. INTRODUCTORY STATISTICS by PREM S. MANN
- Acknowledgement:
  1. Syed S. Hossain, Professor, ISRT, University of Dhaka, Bangladesh.

THANK YOU!