Measures of variability

Measures of variability give you a sense of how spread out the response values are. The range, standard deviation and variance each reflect different aspects of spread.

Range

The range gives you an idea of how far apart the most extreme response scores are. To find the range, simply subtract the lowest value from the highest value.

Range of visits to the library in the past year**Ordered data set:** 0, 3, 3, 12, 15, 24

Range: $24 - 0 = \mathbf{24}$

Standard deviation

The standard deviation (*s*) is the average amount of variability in your dataset. It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

There are six steps for finding the standard deviation:

1. List each score and find their mean.
2. Subtract the mean from each score to get the deviation from the mean.
3. Square each of these deviations.
4. Add up all of the squared deviations.
5. Divide the sum of the squared deviations by $N - 1$.
6. Find the square root of the number you found.

Standard deviations of visits to the library in the past yearIn the table below, you complete **Steps 1 through 4**.

Raw data	Deviation from mean	Squared deviation
15	15 – 9.5 = 5.5	30.25
3	3 – 9.5 = -6.5	42.25
12	12 – 9.5 = 2.5	6.25
0	0 – 9.5 = -9.5	90.25
24	24 – 9.5 = 14.5	210.25
3	3 – 9.5 = -6.5	42.25

Raw data	Deviation from mean	Squared deviation
M = 9.5	Sum = 0	Sum of squares = 421.5

Step 5: 421.5/5 = 84.3

Step 6: √84.3 = 9.18

From learning that ***s* = 9.18**, you can say that on average, each score deviates from the mean by 9.18 points.

Variance

The variance is the average of squared deviations from the mean. Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.

To find the variance, simply square the standard deviation. The symbol for variance is s^2.

Variance of visits to the library in the past year**Data set:** 15, 3, 12, 0, 24, 3

s = 9.18

s^2 = **84.3**

Univariate descriptive statistics

Univariate descriptive statistics focus on only one variable at a time. It's important to examine data from each variable separately using multiple measures of distribution, central tendency and spread. Programs like SPSS and Excel can be used to easily calculate these.

	Visits to the library
N	6
Mean	9.5
Median	7.5
Mode	3
Standard deviation	9.18

	Visits to the library
Variance	84.3
Range	24

If you were to only consider the mean as a measure of central tendency, your impression of the “middle” of the data set can be skewed by outliers, unlike the median or mode.

Likewise, while the range is sensitive to extreme values, you should also consider the standard deviation and variance to get easily comparable measures of spread.

Bivariate descriptive statistics

If you’ve collected data on more than one variable, you can use bivariate or multivariate descriptive statistics to explore whether there are relationships between them.

In bivariate analysis, you simultaneously study the frequency and variability of two variables to see if they vary together. You can also compare the central tendency of the two variables before performing further statistical tests.

Multivariate analysis is the same as bivariate analysis but with more than two variables.

Contingency table

In a contingency table, each cell represents the intersection of two variables. Usually, an independent variable (e.g., gender) appears along the vertical axis and a dependent one appears along the horizontal axis (e.g., activities). You read “across” the table to see how the independent and dependent variables relate to each other.

	Number of visits to the library in the past year				
Group	0–4	5–8	9–12	13–16	17+
Children	32	68	37	23	22
Adults	36	48	43	83	25

Interpreting a contingency table is easier when the raw data is converted to percentages. Percentages make each row comparable to the other by making it seem as if each group had only 100 observations or participants. When creating a percentage-based contingency table, you add the *N* for each independent variable on the end.

Visits to the library in the past year (Percentages)

Group	0–4	5–8	9–12	13–16	17+	*N*
Children	18%	37%	20%	13%	12%	182
Adults	15%	20%	18%	35%	11%	235

From this table, it is more clear that similar proportions of children and adults go to the library over 17 times a year. Additionally, children most commonly went to the library between 5 and 8 times, while for adults, this number was between 13 and 16.

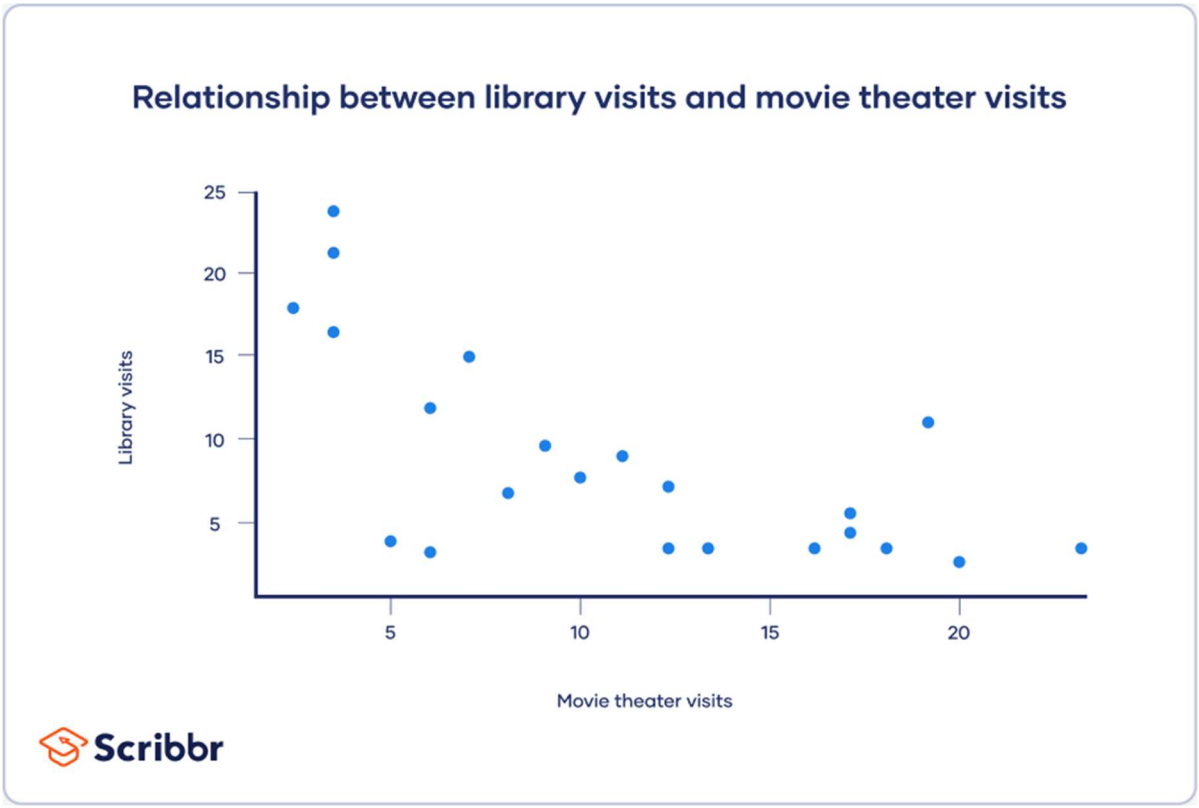
Scatter plots

A scatter plot is a chart that shows you the relationship between two or three variables. It's a visual representation of the strength of a relationship.

In a scatter plot, you plot one variable along the x-axis and another one along the y-axis. Each data point is represented by a point in the chart.

Scatter plot example: Library visits and movie theater visitsYou investigate whether people who visit the library more tend to watch a movie at a theater less. You plot the number of times participants watched movies at a theater along the x-axis and visits to the library along the y-axis.

From your scatter plot, you see that as the number of movies seen at movie theaters increases, the number of visits to the library decreases. Based on your visual assessment of a possible linear relationship, you perform further tests of correlation and regression.



Descriptive Statistics

This page describes graphical and pictorial methods of descriptive statistics and the three most common measures of descriptive statistics (central tendency, dispersion, and association).

Descriptive statistics can be useful for two purposes: 1) to provide basic information about variables in a dataset and 2) to highlight potential relationships between variables. The three most common descriptive statistics can be displayed graphically or pictorially and are measures of:

- Graphical/Pictorial Methods
- Measures of Central Tendency
- Measures of Dispersion
- Measures of Association

There are several graphical and pictorial methods that enhance researchers' understanding of individual variables and the relationships between variables. Graphical and pictorial methods provide a visual representation of the data. Some of these methods include:

- Histograms
- Scatter plots
- Geographical Information Systems (GIS)
- Sociograms

Histograms

- Visually represent the frequencies with which values of variables occur
- Each value of a variable is displayed along the bottom of a histogram, and a bar is drawn for each value
- The height of the bar corresponds to the frequency with which that value occurs

Scatter plots

- Display the relationship between two quantitative or numeric variables by plotting one variable against the value of another variable
- For example, one axis of a scatter plot could represent height and the other could represent weight. Each person in the data would receive one data point on the scatter plot that corresponds to his or her height and weight

Geographic Information Systems (GIS)

- A GIS is a computer system capable of capturing, storing, analyzing, and displaying geographically referenced information; that is, data identified according to location
- Using a GIS program, a researcher can create a map to represent data relationships visually

Sociograms

- Display networks of relationships among variables, enabling researchers to identify the nature of relationships that would otherwise be too complex to conceptualize
-
-
-
-

Measures of central tendency are the most basic and, often, the most informative description of a population's characteristics. They describe the "average" member of the population of interest. There are three measures of central tendency:

- **Mean** -- the sum of a variable's values divided by the total number of values
  Median -- the middle value of a variable
  Mode -- the value that occurs most often
- **Example:**
  The incomes of five randomly selected people in the United States are $10,000, $10,000, $45,000, $60,000, and $1,000,000.
- Mean Income = (10,000 + 10,000 + 45,000 + 60,000 + 1,000,000) / 5 = $225,000
  Median Income = $45,000
  Modal Income = $10,000
- The mean is the most commonly used measure of central tendency. Medians are generally used when a few values are extremely different from the rest of the values (this is called a skewed distribution). For example, the median

income is often the best measure of the average income because, while most individuals earn between $0 and $200,000, a handful of individuals earn millions.

Measures of dispersion provide information about the spread of a variable's values. There are four key measures of dispersion:

- Range
- Variance
- Standard Deviation
- Skew

Range is simply the difference between the smallest and largest values in the data. The interquartile range is the difference between the values at the 75th percentile and the 25th percentile of the data.

Variance is the most commonly used measure of dispersion. It is calculated by taking the average of the squared differences between each value and the mean.

Standard deviation, another commonly used statistic, is the square root of the variance.

Skew is a measure of whether some values of a variable are extremely different from the majority of the values. For example, income is skewed because most people make between $0 and $200,000, but a handful of people earn millions. A variable is positively skewed if the extreme values are higher than the majority of values. A variable is negatively skewed if the extreme values are lower than the majority of values.

Example:
The incomes of five randomly selected people in the United States are $10,000, $10,000, $45,000, $60,000, and $1,000,000:

Range = 1,000,000 - 10,000 = 990,000
Variance = [(10,000 - 225,000)2 + (10,000 - 225,000)2 + (45,000 - 225,000)2 + (60,000 - 225,000)2 + (1,000,000 - 225,000)2] / 5 = 150,540,000,000
Standard Deviation = Square Root (150,540,000,000) = 387,995
Skew = Income is positively skewed

Measures of association indicate whether two variables are related. Two measures are commonly used:

- Chi-square

- Correlation

Chi-Square

- As a measure of association between variables, chi-square tests are used on nominal data (i.e., data that are put into classes: e.g., gender [male, female] and type of job [unskilled, semi-skilled, skilled]) to determine whether they are associated*
- A chi-square is called significant if there is an association between two variables, and nonsignificant if there is not an association

To test for associations, a chi-square is calculated in the following way: Suppose a researcher wants to know whether there is a relationship between gender and two types of jobs, construction worker and administrative assistant. To perform a chi-square test, the researcher counts up the number of female administrative assistants, the number of female construction workers, the number of male administrative assistants, and the number of male construction workers in the data. These counts are compared with the number that would be expected in each category if there were no association between job type and gender (this expected count is based on statistical calculations). If there is a large difference between the observed values and the expected values, the chi-square test is significant, which indicates there is an association between the two variables.

*The chi-square test can also be used as a measure of goodness of fit, to test if data from a sample come from a population with a specific distribution, as an alternative to Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. As such, the chi square test is not restricted to nominal data; with non-binned data, however, the results depend on how the bins or classes are created and the size of the sample

Correlation

- A correlation coefficient is used to measure the strength of the relationship between numeric variables (e.g., weight and height)
- The most common correlation coefficient is **Pearson's r**, which can range from -1 to +1.
- If the coefficient is between 0 and 1, as one variable increases, the other also increases. This is called a positive correlation. For example, height and weight are positively correlated because taller people usually weigh more
- If the correlation coefficient is between -1 and 0, as one variable increases the other decreases. This is called a negative correlation. For example, age and hours slept per night are negatively correlated because older people usually sleep fewer hours per night

This means we have a sample size of 5 and in this case, we use the standard deviation equation for the sample of a population.

Consider the number of gold coins 5 pirates have; 4, 2, 5, 8, 6.

Mean:

$=\overline{x} = \frac{\Sigma x}{n}$

$$\frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots + x_n}{n}$$

= (4 + 2 + 5 + 6 + 8) / 5

= 5

for every value of the sample:

$$x_n - \overline{x}$$

= 20

Standard deviation:

=

= √5

= 2.236

Standard deviation of Grouped Data

In case of grouped data or grouped frequency distribution, the standard deviation can be found by considering the frequency of data values. This can be understood with the help of an example.

Question: Calculate the mean, variance and standard deviation for the following data:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	27	10	7	5	4	2

Solution:

Class Interval	Frequency (f)	Mid Value (x_i)	fx_i	fx_i^2

0 – 10	27	5	135	675
10 – 20	10	15	150	2250
20 – 30	7	25	175	4375
30 – 40	5	35	175	6125
40 – 50	4	45	180	8100
50 – 60	2	55	110	6050
	$\sum f = 55$		$\sum fx_i = 925$	$\sum fx_i^2 = 27575$

$N = \sum f = 55$

Mean = $(\sum fx_i)/N = 925/55 = 16.818$

Variance = $1/(N - 1) [\sum fx_i^2 - 1/N(\sum fx_i)^2]$

$= 1/(55 - 1) [27575 - (1/55) (925)^2]$

$= (1/54) [27575 - 15556.8182]$

$= 222.559$

Standard deviation = √variance = √222.559 = 14.918

Practice Problems on Standard Deviation

1. Calculate the standard deviation of the following values:

5, 10, 25, 30, 50

1. Find the mean and standard deviation for the following data.

x	60	61	62	63	64	65	66	67	68
f	2	1	12	29	25	12	10	4	5

1. The diameters of circles (in mm) drawn in a design are given below:

Diameters	33 – 36	37 – 40	41 – 44	45 – 48	49 – 52
No.of circles	15	17	21	22	25

Calculate the standard deviation and mean diameter of the circles.

[Hint: First make the data continuous by making the classes as 32.5-36.5, 36.5-40.5, 40.5-44.5, 44.5 – 48.5, 48.5 – 52.5 and then proceed.]

Check out more problems on variance and standard deviation of grouped data and Statistics, register with BYJU'S – The Learning App to learn with ease.