STA 405 (Engineering Statistics)

First Material

Descriptive Statistics: Definition, Overview, Types, and Examples

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population.

In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

The next step is **inferential statistics**, which help you decide whether your data confirms or refutes your hypothesis and whether it is generalizable to a larger population.

Types of descriptive statistics

There are 3 main types of descriptive statistics:

The **distribution** concerns the frequency of each value.

The **central tendency** concerns the averages of the values.

The **variability** or dispersion concerns how spread out the values are.

You can apply these to assess only one variable at a time, in univariate analysis, or to compare two or more, in bivariate and multivariate analysis.

Research example, You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:

- ✓ **Go to a library**
- ✓ **Watch a movie at a theater**
- ✓ **Visit a national park**

Your data set is the collection of responses to the survey. Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Frequency distribution

A data set is made up of a distribution of values, or scores. In tables or graphs, you can summarize the frequency of every possible value of a variable in numbers or percentages. This is called a **frequency distribution**.

Simple frequency distribution table

Grouped frequency distribution table

For the variable of gender, you list all possible answers on the left hand column. You count the number or percentage of responses for each answer and display it on the right hand column.

Gender	Number
Male	182
Female	235
Other	27

From this table, you can see that more women than men or people with another gender identity took part in the study.

Measures of central tendency

Measures of central tendency estimate the center, or average, of a data set. The mean, median and mode are 3 ways of finding the average.

Here we will demonstrate how to calculate the mean, median, and mode using the first 6 responses of our survey.

Mean

Median

Mode

The **mean**, or *M*, is the most commonly used method for finding the average.

To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called *N*.

Mean number of library visits

Data set 15, 3, 12, 0, 24, 3

Sum of all values 15 + 3 + 12 + 0 + 24 + 3 = 57

Total number of responses $N = 6$

Mean Divide the sum of values by *N* to find *M*: 57/6 = **9.5**

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 = **24**

Standard deviation

The standard deviation (*s* or *SD*) 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:

List each score and find their mean.

Subtract the mean from each score to get the deviation from the mean.

Square each of these deviations.

Add up all of the squared deviations.

Divide the sum of the squared deviations by $N - 1$.

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
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

$\boldsymbol{s^2}$ = **84.3**

Prevent plagiarism. Run a free check.

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
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 outliers, 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.

Simple Linear Regression

Simple linear regression is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Simple linear regression example, You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

If you have more than one independent variable, use multiple linear regression instead.

Assumptions of simple linear regression

Simple linear regression is a **parametric test**, meaning that it makes certain assumptions about the data. These assumptions are:

1. **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. **Normality**: The data follows a normal distribution.

Linear regression makes one additional assumption:

4. The relationship between the independent and dependent variable is **linear**: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

If your data do not meet the assumptions of homoscedasticity or normality, you may be able to use a nonparametric test instead, such as the Spearman rank test.

Example: Data that doesn't meet the assumptions, You think there is a linear relationship between cured meat consumption and the incidence of colorectal cancer in the U.S. However, you find that much more data has been collected at high rates of meat consumption than at low rates of meat consumption, with the result that there is much more variation in the estimate of cancer rates at the low range than at the high range. Because the data violate the assumption of homoscedasticity, it doesn't work for regression, but you perform a Spearman rank test instead.

If your data violate the assumption of independence of observations (e.g., if observations are repeated over time), you may be able to perform a linear mixed-effects model that accounts for the additional structure in the data.

How to perform a simple linear regression

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B_0** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B_1** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Derivation of the normal equations for fitting a straight line $y = \beta_0 + \beta_1 X + \epsilon$

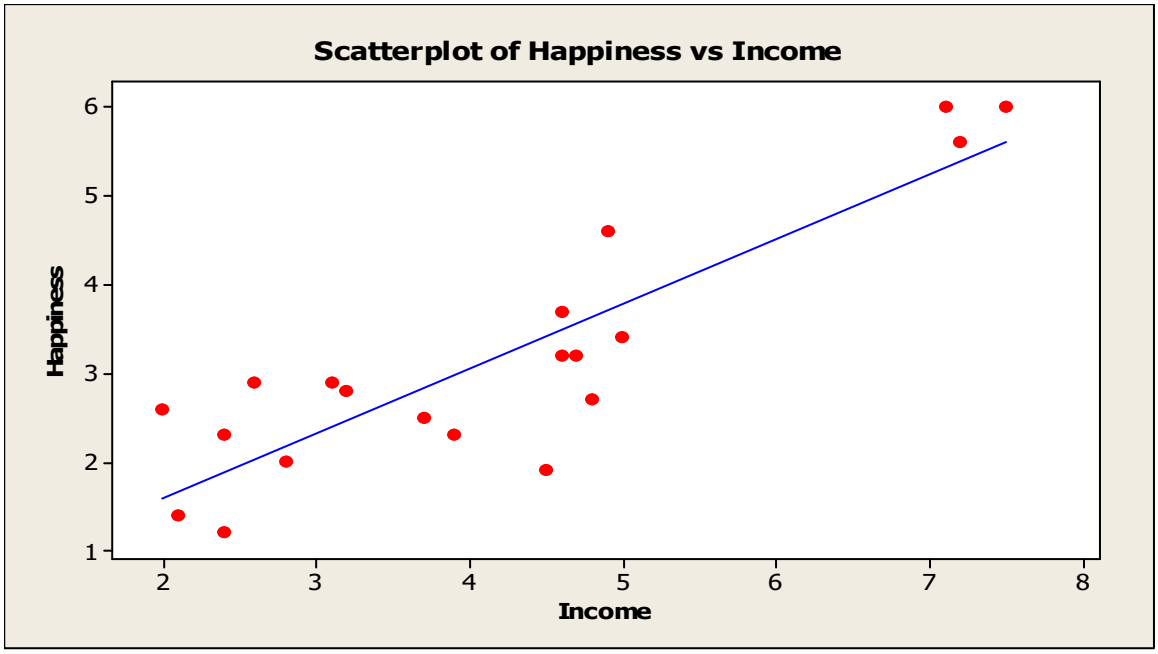
(To be done in the class)

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model.

While you can perform a linear regression by hand, this is a tedious process, so most people use statistical programs to help them quickly analyze the data.

Simple Linear Regression

S/No	Income (10,000)	Happiness
1	3.9	2.3
2	5.0	3.4
3	4.9	4.6
4	3.2	2.8
5	7.2	5.6
6	3.7	2.5
7	4.7	3.2
8	4.5	1.9
9	3.1	2.9
10	4.6	3.7
11	4.6	3.2
12	2.8	2.0
13	7.1	6.0
14	7.5	6.0
15	2.1	1.4
16	2.6	2.9
17	2.4	1.2
18	2.4	2.3
19	4.8	2.7
20	2.0	2.6



Can we study the relationship between Happiness and Income?

Yes, this is because the Scatter Plot above shows that there exists a linear (straight line) relationship between Happiness and Income because the numbers of points above and below the straight line are equal.

Studying the Relationship between Happiness and Income

Regression Analysis: happiness versus income

```
The regression equation is
happiness = 0.150 + 0.725 income

Predictor      Coef  SE Coef      T      P
Constant     0.1497   0.4142   0.36  0.722
income      0.72548  0.09288   7.81  0.000


R-Sq = 77.2%   R-Sq(adj) = 76.0%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  28.076  28.076  61.01  0.000
Residual Error   18   8.283   0.460
Total            19  36.359
```

This output above first presents the Linear Regression Equation (**happiness = 0.150 + 0.725 income**) for the relationship between Happiness and Income.

You can use this regression equation if you want to predict happiness values across the range of income that you have observed:

happiness = 0.150 + **0.725*income**

Interpretation of the Model

The number in the equation (0.725) tells us that for every one unit increase in income (where one unit of income = 10,000) there is a corresponding 0.725-unit increase in reported happiness (where happiness is a scale of 1 to 10).

The R-square is the estimated **effect**, also called the **regression coefficient** or r^2 value. This number shows how much variation there is in our estimate of the relationship between income and happiness.

The P column shows the ***p*** **value**. This number tells us how likely we are to see the estimated effect of income on happiness if the null hypothesis of no effect were true.

Because the *p* value is so low ($p < 0.001$), we can **reject the null hypothesis** and conclude that income has a statistically significant effect on happiness.

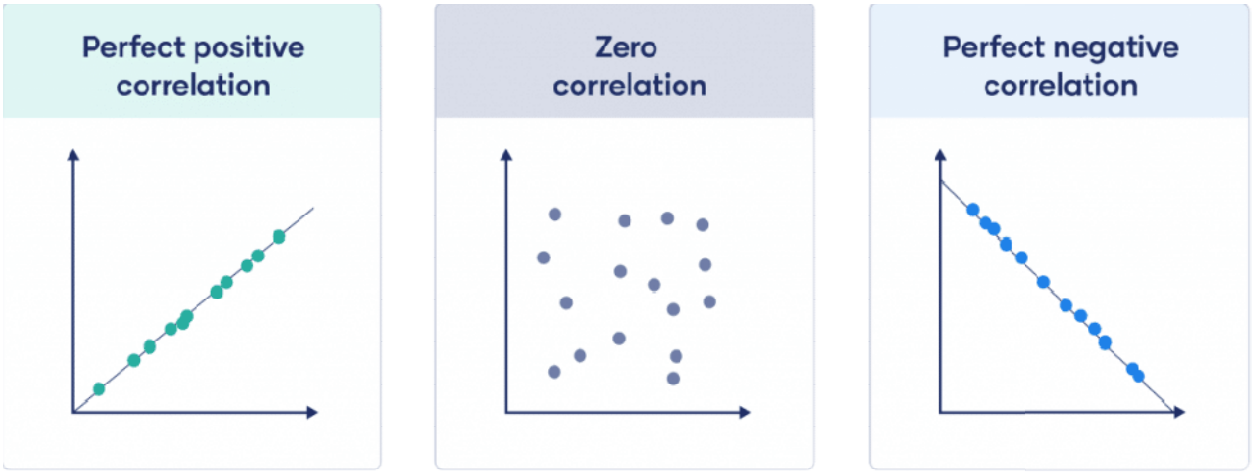
The Analysis of Variance Table provides the test of significance about the model as a whole. The most important thing to notice here is the P column called the *p* value of the model. Here it is significant ($p < 0.001$), which means that this model is a good fit for the observed data. Therefore, the estimated model can be use for forecasting the future Income or Happiness.

Correlation Analysis

A **correlation coefficient** is a number between -1 and 1 that tells you the strength and direction of a relationship between variables.

In other words, it reflects how similar the measurements of two or more variables are across a dataset.

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.



What does a correlation coefficient tell you?

Correlation coefficients summarize data and help you compare results between studies.

Summarizing data

A correlation coefficient is a **descriptive statistic**. That means that it summarizes sample data without letting you infer anything about the population. A correlation coefficient is a bivariate statistic when it summarizes the relationship between two variables, and it's a multivariate statistic when you have more than two variables.

If your correlation coefficient is based on sample data, you'll need an inferential statistic if you want to generalize your results to the population. You can use an F test or a *t* test to calculate a test statistic that tells you the statistical significance of your finding.

Comparing studies

A correlation coefficient is also an **effect size** measure, which tells you the practical significance of a result.

Correlation coefficients are unit-free, which makes it possible to directly compare coefficients between studies.

Using a correlation coefficient

In correlational research, you investigate whether changes in one variable are associated with changes in other variables.

Correlation example

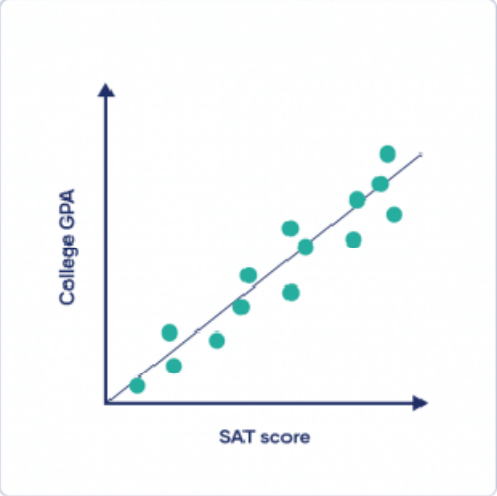
You investigate whether standardized scores from high school are related to academic grades in college. You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

After data collection, you can visualize your data with a scatterplot by plotting one variable on the x-axis and the other on the y-axis. It doesn't matter which variable you place on either axis.

Visually inspect your plot for a pattern and decide whether there is a linear or non-linear pattern between variables. A linear pattern means you can fit a straight line of best fit between the data points, while a non-linear or curvilinear pattern can take all sorts of different shapes, such as a U-shape or a line with a curve.

Visual inspection example

You gather a sample of 5,000 college graduates and survey them on their high school SAT scores and college GPAs. You visualize the data in a scatterplot to check for a linear pattern:



There are many different correlation coefficients that you can calculate. After removing any outliers, select a correlation coefficient that's appropriate based on the general shape of the scatter plot pattern. Then you can perform a correlation analysis to find the correlation coefficient for your data.

You calculate a correlation coefficient to summarize the relationship between variables without drawing any conclusions about causation.

Correlation analysis example

You check whether the data meet all of the assumptions for the Pearson's r correlation test.

Both variables are quantitative and normally distributed with no outliers, so you calculate a Pearson's *r* correlation coefficient.

The correlation coefficient is strong at .58.

Interpreting a correlation coefficient

The value of the correlation coefficient always ranges between 1 and -1, and you treat it as a general indicator of the strength of the relationship between variables.

The **sign** of the coefficient reflects whether the variables change in the same or opposite directions: a positive value means the variables change together in the same direction, while a negative value means they change together in opposite directions.

The **absolute value** of a number is equal to the number without its sign. The absolute value of a correlation coefficient tells you the magnitude of the correlation: the greater the absolute value, the stronger the correlation.

There are many different guidelines for interpreting the correlation coefficient because findings can vary a lot between study fields. You can use the table below as a general guideline for interpreting correlation strength from the value of the correlation coefficient.

While this guideline is helpful in a pinch, it's much more important to take your research context and purpose into account when forming conclusions. For example, if most studies in your field have correlation coefficients nearing .9, a correlation coefficient of .58 may be low in that context.

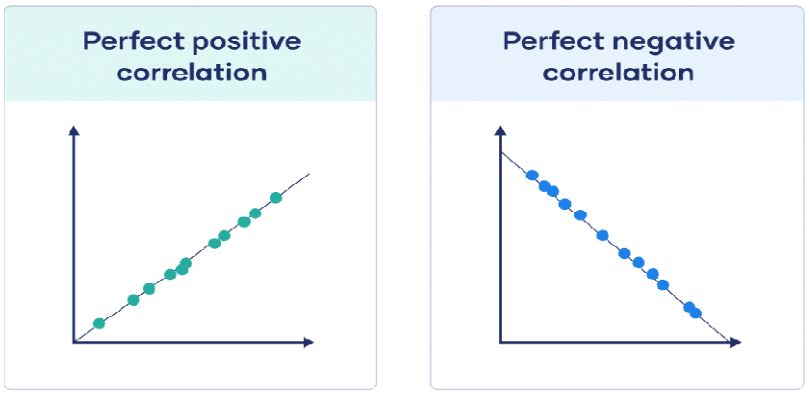
Correlation coefficient	Correlation strength	Correlation type
-0.7 to -1	Very strong	Negative
-0.5 to -0.7	Strong	Negative
-0.3 to -0.5	Moderate	Negative
0 to -0.3	Weak	Negative
0	None	Zero
0 to 0.3	Weak	Positive
0.3 to 0.5	Moderate	Positive
0.5 to 0.7	Strong	Positive
0.7 to 1	Very strong	Positive

Visualizing linear correlations

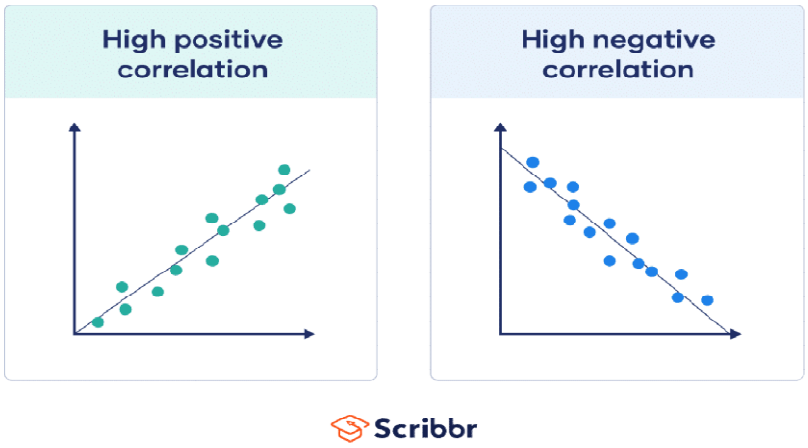
The correlation coefficient tells you how closely your data fit on a line. If you have a linear relationship, you'll draw a straight line of best fit that takes all of your data points into account on a scatter plot.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

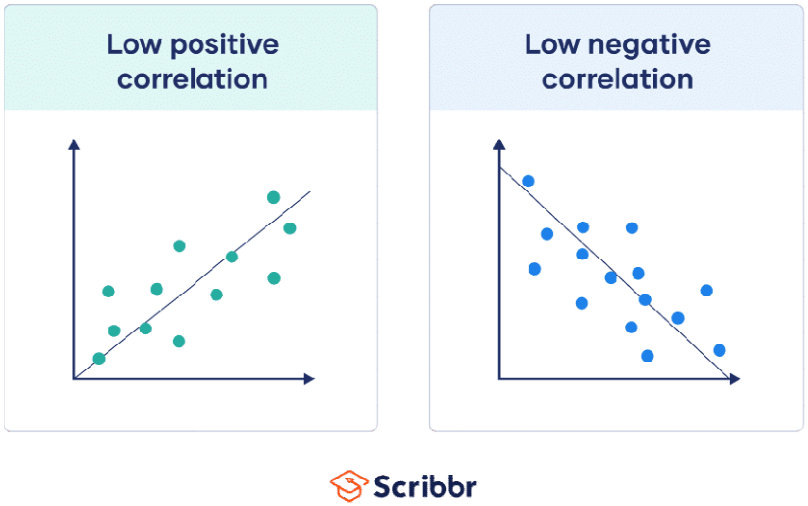
If all points are perfectly on this line, you have a **perfect** correlation.



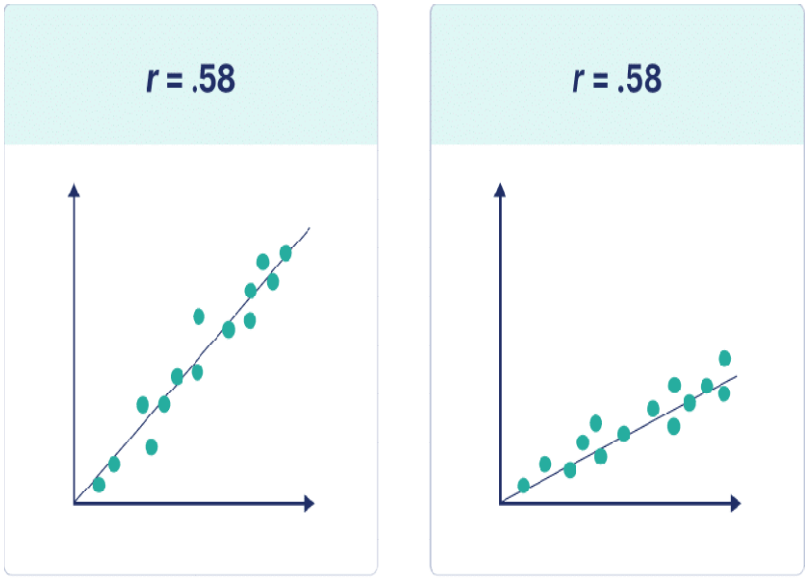
If all points are close to this line, the absolute value of your correlation coefficient is **high**.



If these points are spread far from this line, the absolute value of your correlation coefficient is **low**.



Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines with very different slopes.



Types of correlation coefficients

You can choose from many different correlation coefficients based on the linearity of the relationship, the level of measurement of your variables, and the distribution of your data.

For high statistical power and accuracy, it's best to use the correlation coefficient that's most appropriate for your data.

The most commonly used correlation coefficient is Pearson's r because it allows for strong inferences. It's parametric and measures linear relationships. But if your data do not meet all assumptions for this test, you'll need to use a non-parametric test instead.

Non-parametric tests of rank correlation coefficients summarize non-linear relationships between variables. The Spearman's rho and Kendall's tau have the same conditions for use, but Kendall's tau is generally preferred for smaller samples whereas Spearman's rho is more widely used.

The table below is a selection of commonly used correlation coefficients, and we'll cover the two most widely used coefficients in detail in this article.

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

Pearson's r

The Pearson's product-moment correlation coefficient, also known as Pearson's r, describes the linear relationship between two quantitative variables.

These are the assumptions your data must meet if you want to use Pearson's r:

- Both variables are on an interval or ratio level of measurement
- Data from both variables follow normal distributions
- Your data have no outliers
- Your data is from a random or representative sample
- You expect a linear relationship between the two variables

The Pearson's r is a parametric test, so it has high power. But it's not a good measure of correlation if your variables have a nonlinear relationship, or if your data have outliers, skewed distributions, or come from categorical variables. If any of these assumptions are violated, you should consider a rank correlation measure.

The formula for the Pearson's r is complicated, but most computer programs can quickly churn out the correlation coefficient from your data. In a simpler form, the formula divides the covariance between the variables by the product of their standard deviations.

Formula	Explanation
$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$	• r_{xy}= strength of the correlation between variables x and y • n= sample size • $\sum$= sum of what follows… • X= every x-variable value • Y = every y-variable value • XY = the product of each x-variable score and the corresponding y-variable score

Pearson sample vs population correlation coefficient formula

When using the Pearson correlation coefficient formula, you'll need to consider whether you're dealing with data from a sample or the whole population.

The sample and population formulas differ in their symbols and inputs. A sample correlation coefficient is called *r*, while a population correlation coefficient is called rho, the Greek letter ρ.

The sample correlation coefficient uses the sample covariance between variables and their sample standard deviations.

Sample correlation coefficient formula	Explanation
$r_{xy} = \frac{cov(x,y)}{s_x s_y}$	• r_{xy}= strength of the correlation between variables x and y • $cov(x,y)$ = covariance of x and y • s_x = sample standard deviation of x • s_y = sample standard deviation of y

The population correlation coefficient uses the population covariance between variables and their population standard deviations.

Population correlation coefficient formula	Explanation
$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	• ρ_{XY}= strength of the correlation between variables X and Y • $cov(X,Y)$ = covariance of X and Y • σ_X = population standard deviation of X • σ_Y = population standard deviation of Y

Spearman's rho

Spearman's rho, or Spearman's rank correlation coefficient, is the most common alternative to Pearson's *r*. It's a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

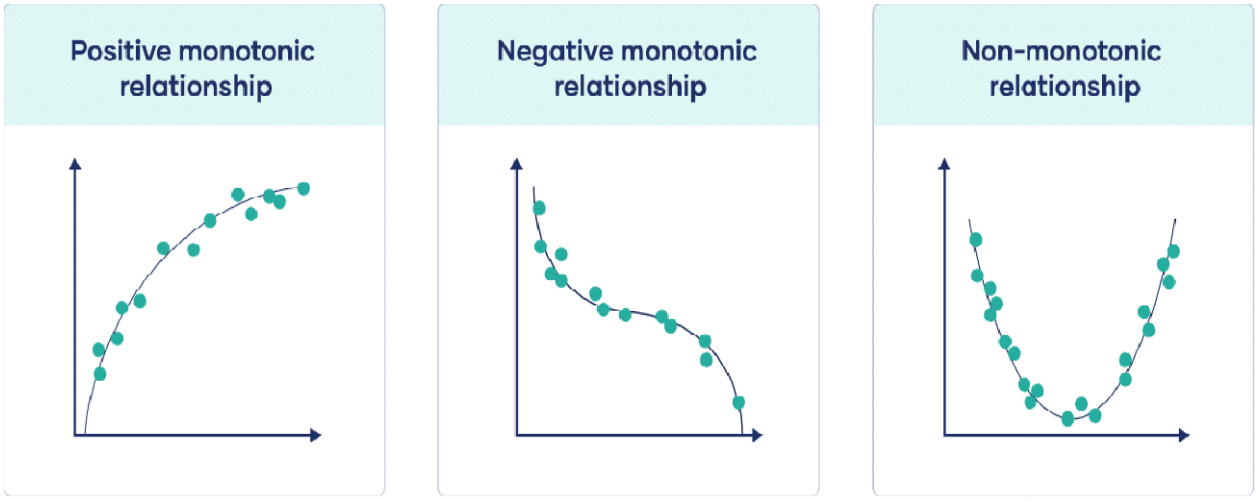
You should use Spearman's rho when your data fail to meet the assumptions of Pearson's *r*. This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

While the Pearson correlation coefficient measures the linearity of relationships, the Spearman correlation coefficient measures the monotonicity of relationships.

In a linear relationship, each variable changes in one direction at the same rate throughout the data range. In a monotonic relationship, each variable also always changes in only one direction but not necessarily at the same rate.

- Positive monotonic: when one variable increases, the other also increases.
- Negative monotonic: when one variable increases, the other decreases.

Monotonic relationships are less restrictive than linear relationships.



Spearman's rank correlation coefficient formula

The symbols for Spearman's rho are ρ for the population coefficient and r_s for the sample coefficient. The formula calculates the Pearson's *r* correlation coefficient between the rankings of the variable data.

To use this formula, you'll first rank the data from each variable separately from low to high: every datapoint gets a rank from first, second, or third, etc.

Then, you'll find the differences (d_i) between the ranks of your variables for each data pair and take that as the main input for the formula.

Spearman's rank correlation coefficient formula	Explanation
$r_s = 1 - \frac{6 \sum d_i^2}{(n^3 - n)}$	• r_s= strength of the rank correlation between variables • d_i = the difference between the x-variable rank and the y-variable rank for each pair of data • $\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks • n = sample size

If you have a correlation coefficient of 1, all of the rankings for each variable match up for every data pair. If you have a correlation coefficient of -1, the rankings for one variable are the exact opposite of the ranking of the other variable. A correlation coefficient near zero means that there's no monotonic relationship between the variable rankings.

Other coefficients

The correlation coefficient is related to two other coefficients, and these give you more information about the relationship between variables.

Coefficient of determination

When you square the correlation coefficient, you end up with the correlation of determination (r^2). This is the proportion of common variance between the variables. The coefficient of determination is always between 0 and 1, and it's often expressed as a percentage.

Coefficient of determination	Explanation
r^2	The correlation coefficient multiplied by itself

The coefficient of determination is used in regression models to measure how much of the variance of one variable is explained by the variance of the other variable.

A regression analysis helps you find the equation for the line of best fit, and you can use it to predict the value of one variable given the value for the other variable.

A high r^2 means that a large amount of variability in one variable is determined by its relationship to the other variable. A low r^2 means that only a small portion of the variability of one variable is explained by its relationship to the other variable; relationships with other variables are more likely to account for the variance in the variable.

The correlation coefficient can often overestimate the relationship between variables, especially in small samples, so the coefficient of determination is often a better indicator of the relationship.

Coefficient of alienation

When you take away the coefficient of determination from unity (one), you'll get the coefficient of alienation. This is the proportion of common variance not shared between the variables, the unexplained variance between the variables.

Coefficient of alienation	Explanation
$1 - r^2$	One minus the coefficient of determination

A high coefficient of alienation indicates that the two variables share very little variance in common. A low coefficient of alienation means that a large amount of variance is accounted for by the relationship between the variables.