Overview (/study/ap 122- cid- 754029/

Teacher view

Index

The big picture

Does early success indicate a good finish?

Credit: Chris Ryan Getty Images

Runners stand at the starting line, tensely waiting for their race to begin. The starting gun fires and the racers lunge forward. As they pass the half-way point, some runners have pushed out ahead of the others, while some choose just to keep up with the rest of the pack. But how likely is it that the runner in the lead now will win? Is there any relationship between the order the runners are in now and the order they will finish?

Student view

Overview (/study/ap 122- cid- 754029/

Two executives are evaluating employees using several criteria to determine who is most deserving of promotions, raises or further training. After examining the scores, it is clear that one executive was much more critical in their evaluation, giving much lower scores. Are the results now useless? If the goal is simply to rank employees from best to worst, can they determine whether their assessments would produce consistent rankings?

In subtopic 4.4 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26239/) you learned how to determine the significance of linear relationships, or correlations, between sets of data. The Spearman's rank correlation coefficient, r_s, also measures the strength of a relationship, but where rank is important rather than actual values, as in the scenarios described above. In this subtopic, you will learn when to use Spearman's rank correlation coefficient and how to calculate and interpret it.

Making connections

Statistics is used to verify results and test for bias in numerous applications. Spearman ' s rank correlation is one tool that can be used to test for consistency in subjectively rated assessments. Why do you think it would be important to determine if ratings are consistent? Can you think of situations besides the employee evaluations described above where ratings are used?

Concept

When researchers notice trends in data, various statistical tests are used to determine the validity of the hypotheses they make. While there may *seem* to be a relationship between two variables, the question of how *confident* we can be in claiming that relationship exists.

4. Probability and statistics / 4.10 Spearman's rank correlation coefficient,

Understanding Spearman's rank correlation coefficient

Student view

Overview (/study/ap 122-cid-754029/

Choosing a test

Types of data

The least specific type of data is nominal data, which includes data that are separated into categories that do not have any particular order.

Ordinal data are data placed first, second, third, ... in order, or rank, without actual values. Therefore, the difference between the quality or value of things ranked 1 and 2 is not necessarily the same as the difference between the things ranked 2 and 3. Ordinal data can also include categories that have an implied order, such as size, or the degree to which someone agrees with something, like a survey question.

Data that indicate value and have a consistent distance between consecutive values are known as interval data.

If the data contain an absolute 0, then we can relate the values proportionally to each other, which is why it is called ratio data.

Types of test

Specific types of data require specific statistical tests. The table below shows a summary of the four general types of data and the tests we use for them.

Type	Examples	Use
Nominal	male and female left-handed, right-handed, ambidextrous cars, trucks, motorcycles, vans	χ^2 (chi-squared) (subtopic 4.11 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-27558/))
Ordinal	first, second, third... small, medium, large disagree, ambivalent, agree	r_s (Spearman 's rank correlation)

Student view

Type	Examples	Use
Interval	year (1096, 1611, etc.) temperature	Pearson's correlation (subtopic 4.4 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26239/))
Ratio	measurements of length, distance, area, time, speed or acceleration	Pearson's correlation (subtopic 4.4 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-26239/))

The four types of data in the table above have been arranged from the most general (nominal) to most precise (ratio). You can convert data to a simpler type, but not to a more precise type. For example, if you have ratio data, you can order it and assign ranks to it to make it ordinal data to perform a Spearman's rank test, or you can divide it into ranges of values to create categories for use in a χ^2 test.

Be aware

If you are collecting data to analyse, it is best to gather it in the most precise form you can and convert it to a more general form as needed. If you initially planned a χ^2 test, but then decided to find Pearson ' s correlation coefficient instead, it would be impossible if all your data were nominal.

Theory of Knowledge

The four types of data — nominal, ordinal, interval and ratio — provide one example of how language is used within an area of knowledge to indicate different meanings. In order for a data set to be useful for a particular analysis, it must contain the proper type of data.

Categorising data like this enables us to communicate requirements for mathematical tests clearly without having to explain the concept each time. How do you think language comes to be used this way in the area of mathematics? Can you see any parallels between the use of language and the use of notation in mathematics? What would the study of mathematics be like if we did not have vocabulary and notation like this?

Overview (/study/ap 122- cid- 754029/

The purpose of Spearman's rank correlation

Spearman's rank correlation is used to find the strength of a correlation between two sets of ordinal data that show a monotonic trend. Why would you want to find a correlation between ranks instead of the values? Consider some of the examples described in section 4.10.0 (/study/app/m/sid-122-cid-754029/book/the-big-picture-id-27583/): does early success in a race correlate to success at the finish? Are two judges in a competition ranking the competitors in approximately the same order?

Other examples might include finding correlations between the level of education reached and annual salary or between someone's position in a company and their years of experience. Variables like level of education and position in a company may not initially seem like ordinal data, but remember that order can be implied by increased position or some other measure of quality.

International Mindedness

Numerous sports in the Olympics and other international competitions are determined by judging, which naturally brings an element of subjectivity and possible bias. Figure skating is one competition that has experienced several instances of questionable judging, as the research cited in this article (https://www.theatlantic.com/entertainment/archive/2014/02/why-people-think-adelina-sotnikovas-figure-skating-gold-medal-was-rigged/358344/) explains.

Can you think of what might cause professional judges to be biased? How could you tell?

Calculating Spearman's rank correlation coefficient

Finding r_s using the formula

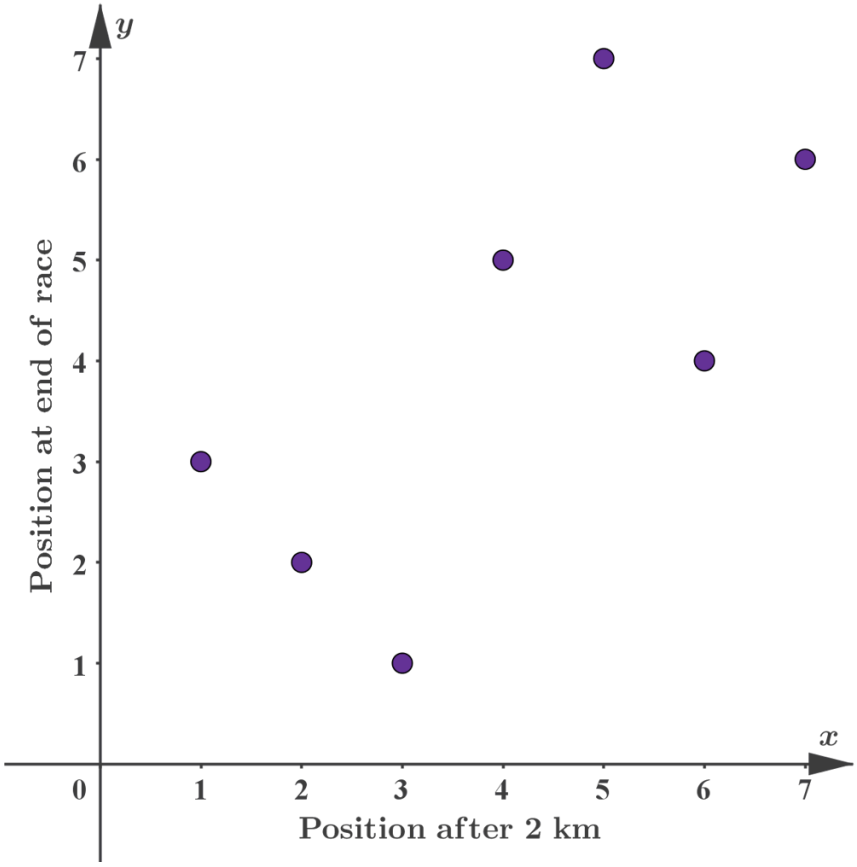
Consider the data in the table below, which shows seven runners' positions in a 5 km race after 2 km and at the end of the race.

Student view

Runner	After 2 km (x)	End of race (y)
A	1	3

Overview (/study/ap 122- cid- 754029/

Runner	After 2 km (x)	End of race (y)
B	2	2
C	3	1
D	4	5
E	5	7
F	6	4
G	7	6



More information

The image is a scatterplot that depicts the positions of racers at two points: 'Position after 2 km' on the x-axis and 'Position at end of race' on the y-axis. The x-axis ranges from 0 to 7 and the y-axis also ranges from 0 to 7. Each point on the scatterplot represents a racer's position after 2 km versus their final position. The data points show a general upward trend, suggesting that runners who were better positioned after 2 km tended to maintain or improve their

Student view

Overview (/study/ap 122- cid- 754029/

positions by the end of the race. Despite the upward trend, the relationship is not perfectly linear as only one racer maintained the same position at both checkpoints, highlighting variability among the runners' performances over time. This scatterplot is relevant for exploring potential correlations using Spearman's rank correlation coefficient.

[Generated by AI]

As you can see, only one of the racers held the same position at both points in the race. It seems possible from the scatterplot that runners who started off better tended to finish better, but you can't be certain. There does seem to be an upward trend, so it is reasonable to think there might be a monotonic relationship. This is a situation suited to testing with Spearman's rank correlation coefficient, r_s.

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$, where d is the difference between the ranks of the independent variable and dependent variable and n is the number of data points.

Example 1

★☆☆

Calculate the Spearman's rank coefficient for the example of the runners above.

A table, such as the one below, can be helpful to organise the data for your calculations.

After 2 km (x)	End of race (y)	d	d^2
1	3	-2	4
2	2	0	0
3	1	2	4
4	5	-1	1
5	7	-2	4
6	4	2	4

Student view

After $2\ \text{km}$ (x)	End of race (y)	d	d^2
7	6	1	1

This table helps you find $\sum d^2 = 4 + 0 + 4 + 1 + 4 + 4 + 1 = 18$.

There are seven runners, so $n = 7$. Now you can use the formula to find r_s.

$$\begin{aligned} r_\text{s} &= 1 - \frac{6 \sum d^2}{n\left(n^2 - 1\right)} \\ &= 1 - \frac{6 \times 18}{7\left(7^2 - 1\right)} \\ &= 1 - \frac{108}{336} \\ &= \frac{19}{28} \\ &= 0.679 \ (3 \text{ significant figures}) \end{aligned}$$

Exam tip

You will not be required to remember or use the formula to find r_s . Section 4.10.2 (/study/app/m/sid-122-cid-754029/book/calculating-spearmans-rank-correlation-coefficient-id-27585/) shows you how to find it using technology instead.

One important thing to note is that this formula is an **alternative form** of Pearson's correlation that only works when these conditions are met:

- The data are ordinal and monotonic.
- Each element of a set occurs only once; in other words, there are no ties.

Important

The formula $r_\text{s} = 1 - \frac{6 \sum d^2}{n\left(n^2 - 1\right)}$ works only when there are no repeated values in the data for either variable.

If there is a tie, you must average the ranks and use least-squares regression, as you will examine in the next section.

Calculating Spearman's rank correlation coefficient using technology

Learning to calculate Spearman's rank correlation coefficient by hand is helpful for understanding the formula and the process of comparing ranks. However, you need to be able to calculate it using technology. This is faster and can help to handle situations in which there is a tie in the rankings.

The table below includes data in which certain values are repeated. When this occurs, find the mean of the rankings those values would hold as shown. Note that after averaging the ranks of equal values, the next value takes the rank following the largest one averaged.

Set A	Rankings of A	Set B	Rankings of B
3	1	4	$\frac{(2+3+4)}{3}=3$
5	2	4	$\frac{(2+3+4)}{3}=3$
6	$\frac{(3+4)}{2}=3.5$	2	1
6	$\frac{(3+4)}{2}=3.5$	4	$\frac{(2+3+4)}{3}=3$
8	5	9	6
11	6	7	5

Remember that if there are equal values, you cannot use the formula given in the previous section. Once you have ranked the values, use the least-squares regression method that you learned in section 4.4.2 (/study/app/m/sid-122-cid-754029/book/pearsons-r-correlation-

Overview (/study/ap 122- cid- 754029/

coefficient-id-26241/) with the Linear Regression function on the calculator.

Example 1

★☆☆

Use the rankings from the table above to find Spearman's rank correlation coefficient for Sets A and B.

Use the Linear Regression function on the calculator with the rankings found in the table above.

Rankings of A	Rankings of B
1.0	3
2.0	3
3.5	1
3.5	3
5.0	6
6.0	5

From the calculator, you find $r_s \approx 0.585$.

Example 2

★★☆

The table below shows the score given by two judges for skating skills for the twenty finalists of the ice dance free dance competition of the 2021 World Figure Skating Championships.

Student view

Overview (/study/ap 122- cid- 754029/

Name	Score, judge 1	Score, judge 2
Sinitsina/Katsalapov	9.50	9.75
Gilles/Poirier	9.50	9.50
Hubbel/Donohue	9.75	9.50
Chock/Bates	9.00	9.00
Stepanova/Bukin	9.25	9.25
Guignard/Fabbri	9.00	9.25
Fear/Gibson	8.50	9.00
Fournier Beaudry/Sorensen	9.00	9.00
Hawayek/Baker	8.25	8.75
Zagorski/Guerreiro	8.25	8.75
Hurtado/Khaliavin	8.25	8.50
Wang/Liu	7.25	8.50
Lajoie/Lagha	8.50	8.50
Kaliszek/Spodyriev	8.00	8.75
Reed/Ambrulevicius	8.00	8.25
Galyavieva/Thauron	7.75	8.25
Lopareva/Brissaud	7.50	7.75
Nazrova/Nikitin	7.00	7.25
Mueller/Dieck	7.75	8.00
Komatsubara/Koleto	7.25	7.25

Student view

Overview (/study/ap 122- cid- 754029/

Find the Spearman's rank correlation coefficient and comment on the result.

To find the Spearman's rank correlation coefficient, the first step is to rank the data. Remember to use averages for equal scores.

Judge 1			Judge 2		
Score	Position	Rank	Score	Position	Rank
9.50	$2-3$	2.5	9.75	1	1
9.50	$2-3$	2.5	9.50	$2-3$	2.5
9.75	1	1	9.50	$2-3$	2.5
9.00	$5-7$	6	9.00	$6-8$	7
9.25	4	4	9.25	$4-5$	4.5
9.00	$5-7$	6	9.25	$4-5$	4.5
8.50	$8-9$	8.5	9.00	$6-8$	7
9.00	$5-7$	6	9.00	$6-8$	7
8.25	$10-12$	11	8.75	$9-11$	10
8.25	$10-12$	11	8.75	$9-11$	10
8.25	$10-12$	11	8.50	$12-14$	13
7.25	$18-19$	18.5	8.50	$12-14$	13
8.50	$8-9$	8.5	8.50	$12-14$	13
8.00	$13-14$	13.5	8.75	$9-11$	10
8.00	$13-14$	13.5	8.25	$15-16$	15.5
7.75	$15-16$	15.5	8.25	$15-16$	15.5

Student view

Overview (/study/ap 122- cid- 754029/

Judge 1			Judge 2		
Score	Position	Rank	Score	Position	Rank
7.50	17	17	7.75	18	18
7.00	20	20	7.25	$19-20$	19.5
7.75	$15-16$	15.5	8.00	17	17
7.25	$18-19$	18.5	7.25	$19-20$	19.5

Use the linear regression option of your calculator to find the correlation coefficient of the ranks.

The Spearman's rank correlation coefficient is $r_S = 0.933$.

This indicates a strong correlation of the ranks, so even though the two judges did not always agree on the scores, there is a strong connection between them in ranking the competitors.

Activity

People often have very different opinions about new movies. Professional critics give their opinions, and then the public assesses the movie in the form of purchasing tickets or posting online reviews. Look up critic reviews and ticket sales for at least five recent movies and use Spearman's rank to determine if critical approval is a significant indicator of public approval.

2 section questions

4. Probability and statistics / 4.10 Spearman's rank correlation coefficient.

Checklist

Student view

Section Student... (0/0) Feedback Print (/study/app/m/sid-122-cid-754029/book/checklist-id-27586/print/) Assign

Overview (/study/ap 122- cid- 754029/

What you should know

By the end of this subtopic you should be able to:

- determine the appropriate context in which to use Spearman's rank coefficient,
- know how to rank the data with or without repeated values,
- use technology to find r_s for any sets of data.

Section Student... (0/0) Feedback Print (/study/app/m/sid-122-cid-754029/book/calculating-spearmans-rank-correlation-coefficient-id-27585/print/) Assign

Section Student... (0/0) Feedback Print (/study/app/m/sid-122-cid-754029/book/understanding-spearmans-rank-correlation-id-27584/print/) Assign

Investigation

Section Student... (0/0) Feedback Print (/study/app/m/sid-122-cid-754029/book/investigation-id-27587/print/) Assign

In this subtopic two ways of calculating the Spearman rank correlation coefficient were mentioned. One way was to use a formula, the other way was to rank the variables and use the calculator to find the Pearson correlation coefficient of the ranked variables.

- These two methods give the same value if there are no repeated values in the data. Pick a data set with no repeating values and check this by finding the Spearman rank correlation coefficients using both methods.
- Investigate what happens if there are repeated values. Pick data sets with repeated values and find the Spearman rank the two methods give. Do these values differ by much? Does the difference depend on the percentage of repeated values in the data?
- You can also check out the different online tools and spreadsheets that are available to find the Spearman rank correlation coefficient. Which method do they use? Do they rank the data using the averaging method used in this subtopic?

Exam tip

This investigation asks you to explore different ways to find the Spearman rank correlation coefficient. On exams the expectation is that you use the averaging method to rank the data and use your calculator to find the Pearson correlation

Student view

Overview (/study/ap 122- cid- 754029/

coefficient of the ranked data.

When you explore different online tools to find the Spearman rank correlation coefficient, you may notice that these tools also report a p-value besides the coefficent. This p-value is used by statisticians to determine the significance of the coefficient. This is beyond the syllabus, but if you are interested, you can investigate how to interpret this p-value.

Student view