2023. V1

Data Analysis Addendum

for UK Fire Rescue Services

NFCC
National Fire
Chiefs Council

Introduction

> **Many decisions within the FRS are based on the interpretation of data**

The robustness of these decisions is reliant on both the quality of the data and how well it was analysed. This section builds on the guidance on data analysis given in the main document.

Go to Evaluation Methods document →

It provides additional guidance on how to complete a range of common statistical tests frequently used in the evaluation process and how to undertake the analysis of qualitative data.

Who is this section for?

The quantitative content aims to support fire and rescue personnel who already have a level of statistical knowledge and who are confident users of Excel but may not be familiar with common statistical tests used in evaluation. The qualitative sub-section provides a simple introduction and is suitable for anybody new to the analysis of this type of data.

Software requirements

This document focuses on the use of Excel and requires that the Excel Analysis ToolPak add-on has been installed. This is a Microsoft package that comes as part of the Excel package from 2011 onwards. Instructions on how to do this can be found on the Microsite website here. As well as Excel, there is a range of other packages that could be considered, these are not discussed in this document, and are subject to pricing and agreements.

These include, but are not limited to:

Quantitative analysis

As well as a range of plug-in packages for Excel consideration could be given to:

IBM's SPSS: https://www.ibm.com/products/spss-statistics
JMP: https://www.jmp.com/en_gb/home.html

Both these software packages allow the user to complete all the tests covered in this document but also include a range of more sophisticated tests.

Qualitative analysis

NVIVO: https://support.qsrinternational.com/s/article/How-do-I-download-NVivo

This package can aid the coding of qualitative data and is widely used in thematic analysis.

Other packages are available, and it is recommended that a review is completed by a FRS before purchasing to identify what package is best suited to their needs.

What questions is the analysis trying to answer?

Before starting the analysis, the evaluator should ensure they are fully conversant with the aims of the evaluation (Further information about the aims and objectives of evaluation is available in the main document).

This will then allow the evaluator to develop a specific list of questions to be answered. This does not mean that these will be the only areas of enquiry, but having a list of questions will help to focus the initial analysis of the data. Typical questions may include:

- How often did something occur?
- Were there any differences found:
  - Before and after the intervention?
  - Between groups (for example, between a group that received the intervention and one that did not)?
- Are there any relationships between different variables (for example, between age and the risk of dying as a result of a fire)?
- How did people react to the intervention?
- How could the intervention be further developed?

How to use this document

After this introductory section, the document includes a navigation flowchart.

Navigation Flowchart

This flow chart provides a quick way of navigating through this document. Click on the heading to go to the section required. Each test provides an outline of why and when a test should be used and is related to an example.

The menu bar below will take you to other sections of the document, including the contents page indicated by the ‘home’ icon, flowchart, quantitative analysis, qualitative analysis and the glossary.

The glossary is important as some of the terms used in statistics require a detailed explanation which is included in that section. If you are not familiar with these terms, it may be worthwhile reading the glossary before proceeding.

Supporting

This document also has supporting information at the end of the document including glossary of terms, useful links, test formulas and references.

The document looks at the analysis of both qualitative and quantitative data.

Quantitative Analysis

The quantitative content aims to support fire and rescue personnel who already have a level of statistical knowledge and who are confident users of Excel but may not be familiar with common statistical tests used in evaluation.

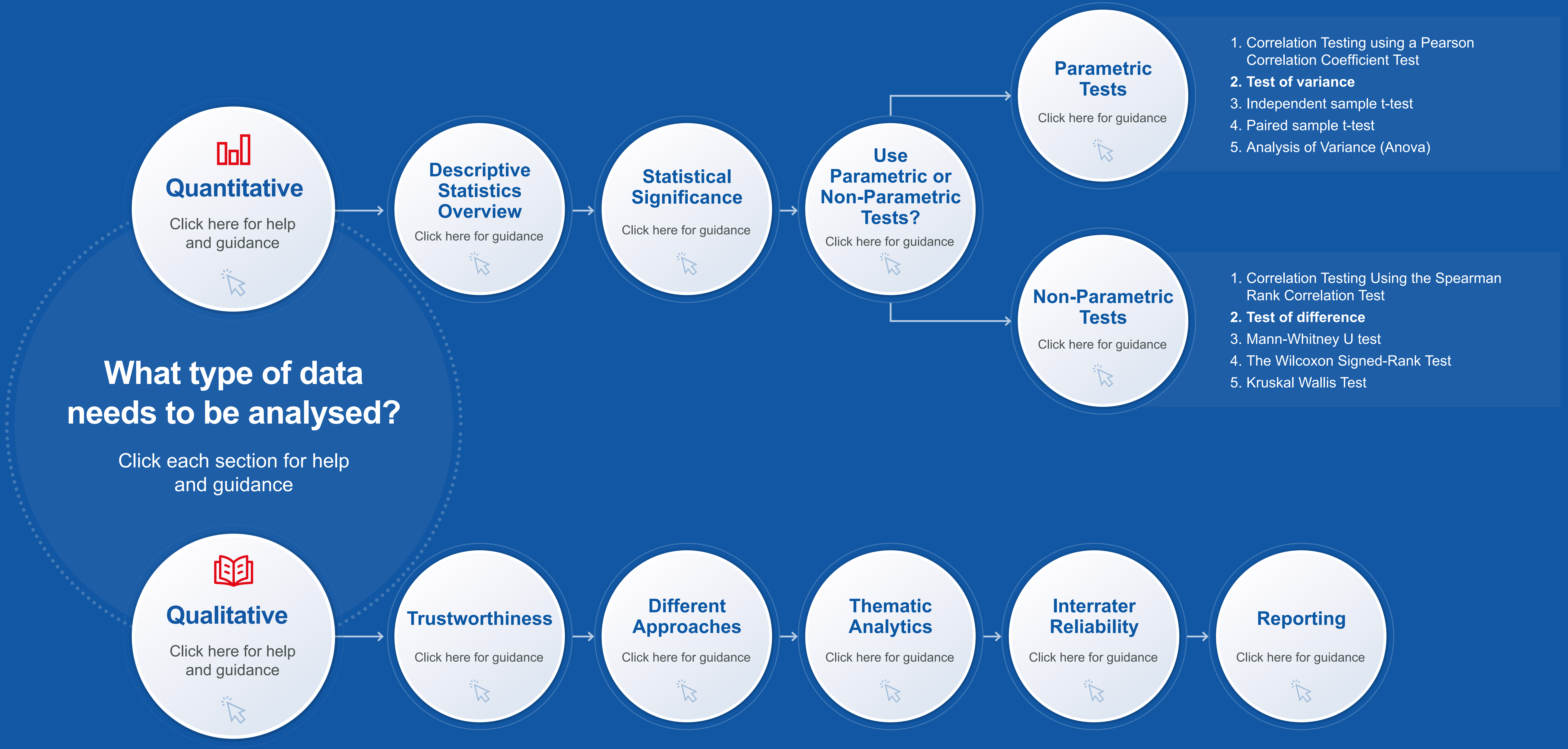
Qualitative Analysis

The qualitative sub-section provides a simple introduction and is suitable for anybody new to the analysis of this type of data.

Go to Contents →

Contents

NFCC
National Fire
Chiefs Council



What type of data needs to be analysed?

Click each section for help and guidance

Quantitative

Click here for help and guidance

Descriptive Statistics Overview

Click here for guidance

Statistical Significance

Click here for guidance

Use Parametric or Non-Parametric Tests?

Click here for guidance

Parametric Tests

Click here for guidance

1. Correlation Testing using a Pearson Correlation Coefficient Test
2. **Test of variance**
3. Independent sample t-test
4. Paired sample t-test
5. Analysis of Variance (Anova)

Non-Parametric Tests

Click here for guidance

1. Correlation Testing Using the Spearman Rank Correlation Test
2. **Test of difference**
3. Mann-Whitney U test
4. The Wilcoxon Signed-Rank Test
5. Kruskal Wallis Test

Qualitative

Click here for help and guidance

Trustworthiness

Click here for guidance

Different Approaches

Click here for guidance

Thematic Analytics

Click here for guidance

Interrater Reliability

Click here for guidance

Reporting

Click here for guidance

NFCC
National Fire
Chiefs Council

Quantitative Analysis

Descriptive statistics

This term refers to a group of statistics that look at how often something occurs, where it occurs, what time it occurs, who is involved and the distribution of the data. For example, how many Road Traffic Collisions (RTC) have occurred in an area? What time of day did they occur? How many people were involved? What were their age, gender and ethnicity, etc.? These statistics should be explored before a more in-depth analysis is completed, as this will help the evaluator to identify areas for more detailed analysis and assist in finding any anomalies or errors in the data that may need to be addressed.

Excel can quickly and simply produce a wide range of descriptive statistics. In particular, the use of pivot tables and pivot graphs can be very useful. Further information about these tools and their use can be found at: https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576

As with all statistical approaches, it is important to recognise the limitations of descriptive statistics. Most notably, they do not explain why something has occurred or if the finding is significant.

Significance

Statistical significance is a term used to describe the probability that a result is real and not a chance occurrence. This is expressed as a p-value, the lower the value of p, the lower the probability that the finding is due to chance. The point at which something is accepted as being statistically significant is called the alpha value. The accepted alpha value for most research is .05. Therefore, to be regarded as being significant, a p-value would need to be at or below .05. This value would indicate there is a 95% probability that any difference found is real and not a chance occurrence. It should be noted that this does not mean it could not have occurred by chance; there is still a 1 in 20 chance of an error with an alpha value of .05. There is a range of tests that can be used to work out the p-value. These tests fall into two groups, parametric and non-parametric.

Should you use Parametric or Non-Parametric tests?

Parametric tests

These tests are based on means and the distribution of the data points. They assume that the data:

1. Is normally distributed
2. The variance is approximately equal between each data set
3. There are no outliers in the data

Parametric tests are widely used as they are very robust, even when the distribution is less than perfect (Langdridge, 2004). These tests include t-tests and Analysis of Variance (ANOVA).

Non-parametric test

Non-parametric tests, sometimes referred to as assumption-free tests, make fewer assumptions about the distribution of the data and therefore are less prone to the impact of outliers in the data and are better able to deal with uneven distributions. In most cases, these tests achieved this by ranking the data to identify any difference between data sets. However, non-parametric tests can be less powerful than parametric tests (Field, 2018). Consideration should be given to the use of non-parametric tests if:

> There are outliers in the data that cannot be removed and/or if the median is a more useful measure than the mean. For example, the median figure for household income is probably a more representative figure than the mean, which would be influenced by high earners.

The sample is too small to use a parametric test. What would be regarded as too small of a sample is complex, but as a rule of thumb, consideration to using non-parametric tests should be given if the sample is less than 20.

The data is ordinal in nature. For example, if a rating system is used and the difference between the grading is unclear, for example: much worse, slightly worse, no change, slightly improved, or much improved (although, this could be used as scale data if the distribution was acceptable).

Parametric Test

Correlation Testing using a Pearson Correlation Coefficient Test

Why complete a correlation test?

It is often useful to know if there is a correlation **(a relationship)** between one variable and another. For example, an evaluator may want to know if there was a relationship between age and the number of fire deaths in the UK. If there is, resources can then be targeted appropriately. However, it is important to recognise that a **correlation does not mean that one variable caused another.** For example, the number of firefighters attending a fire may correlate with the amount of damage caused. This would not mean they cause the damage; it is likely that larger fires require more firefighters to attend, and it is the fire that causes the damage. **As this example demonstrates, it is important that great care is taken by the evaluator when considering if one variable caused another or whether they simply occurred together.**

Conducting a Pearson correlation coefficient calculation in Excel

A Pearson correlation coefficient calculation is a parametric test used to test for a linear relationship between two variables. Click here to see the formula for the test. It provides a coefficient score (how strong the relationship is) ranging from +1 to -1. A score of +1 would indicate a perfect positive relationship with both variables moving in the same direction at the same rate (i.e. as one gets larger, the other gets larger at the same rate). A coefficient of -1 would indicate a perfect negative relationship between the variables, with the variables moving in the opposite direction to each other (i.e., one is increasing as one is decreasing).

In this example, the evaluator wants to know if there is a relationship between age and the number of fire deaths in the UK. A visual check would indicate there may be a linear positive relationship (both going up together), but this visualisation does not indicate how strong any relationship is and if it is statistically significant (Figure 1). To answer these questions, the coefficient (how strong the relationship is) would need to be calculated, as would the level of significance.

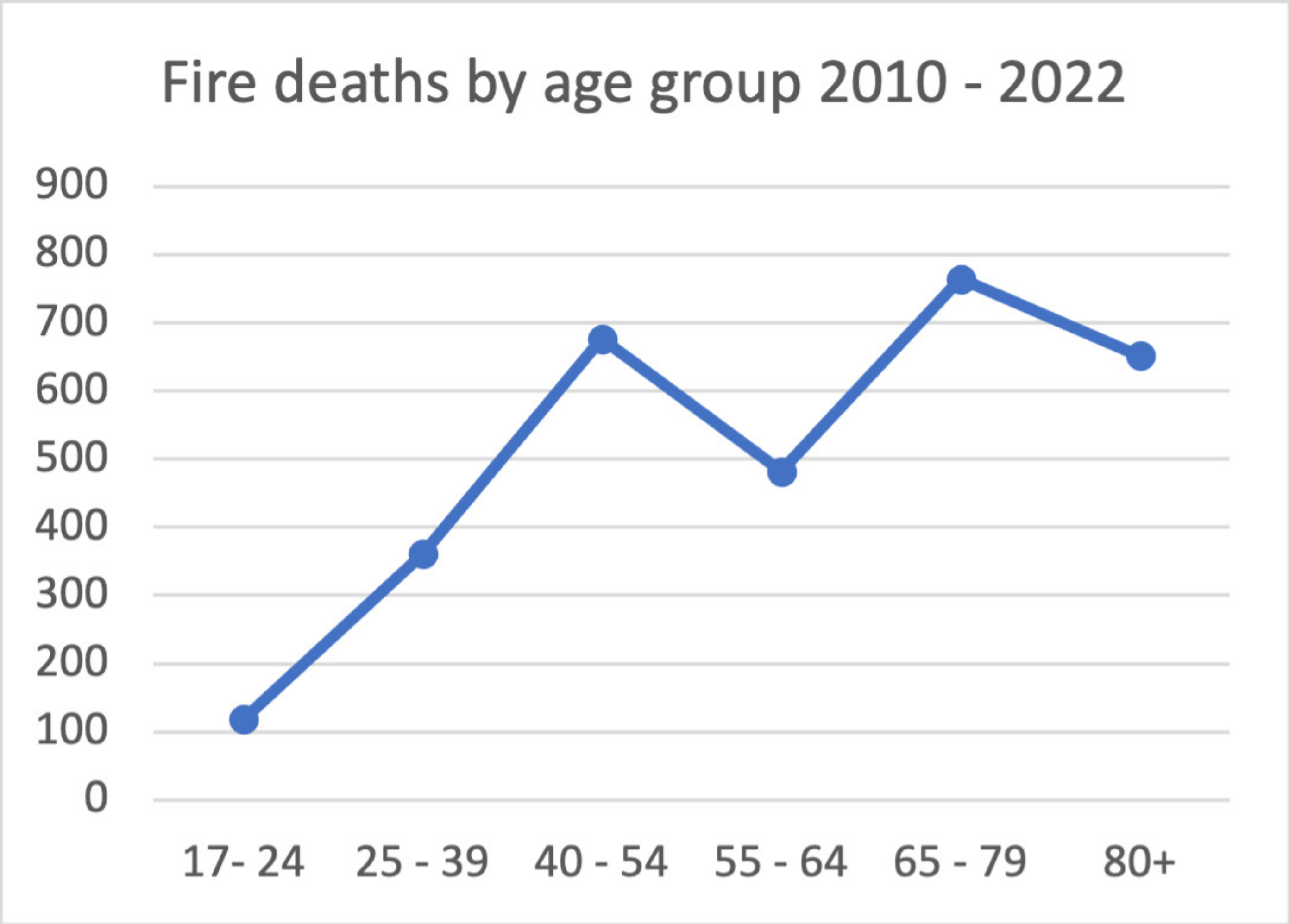
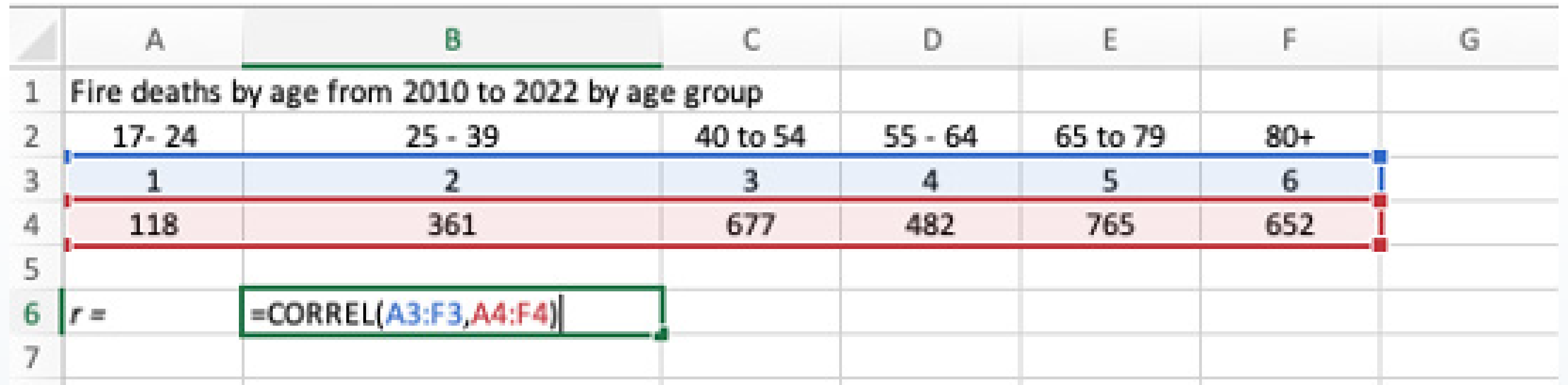


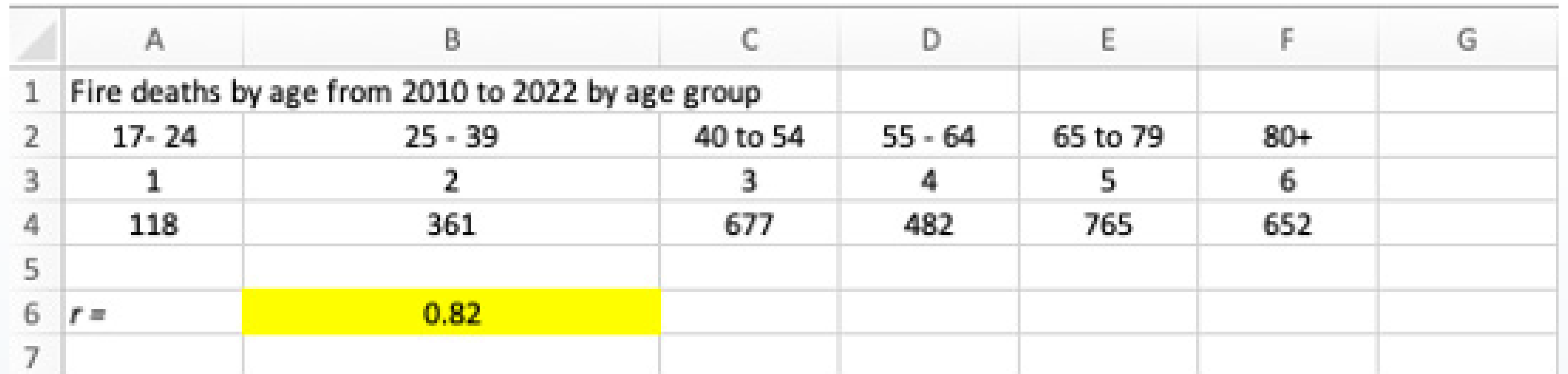
Figure 1. Fire-related deaths by age group from 2010 - 2022

The formula for working out the coefficient, expressed as an r-value, in Excel is =CORREL(array1, array2). Figure 2 shows how to apply this. **Note: the age ranges have been coded 1 to 6, allowing the data to be used as a scale rather than ordinal data.** Figure 3 gives the coefficient as r = .82. Given that the maximum score is 1, this would be regarded as a strong correlation (Table 1). To identify if this is a significant finding, it is necessary to complete a t-test and a p-value calculation. Once again, both are simple to perform on Excel.



	A	B	C	D	E	F	G
1	Fire deaths by age from 2010 to 2022 by age group						
2	17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+	
3	1	2	3	4	5	6	
4	118	361	677	482	765	652	
5							
6	r =	=CORREL(A3:F3,A4:F4)					
7							

Figure 2. The Excel coefficient formula =CORREL(array1, array2)



	A	B	C	D	E	F	G
1	Fire deaths by age from 2010 to 2022 by age group						
2	17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+	
3	1	2	3	4	5	6	
4	118	361	677	482	765	652	
5							
6	r =	0.82					
7							

Figure 3. Coefficient answer

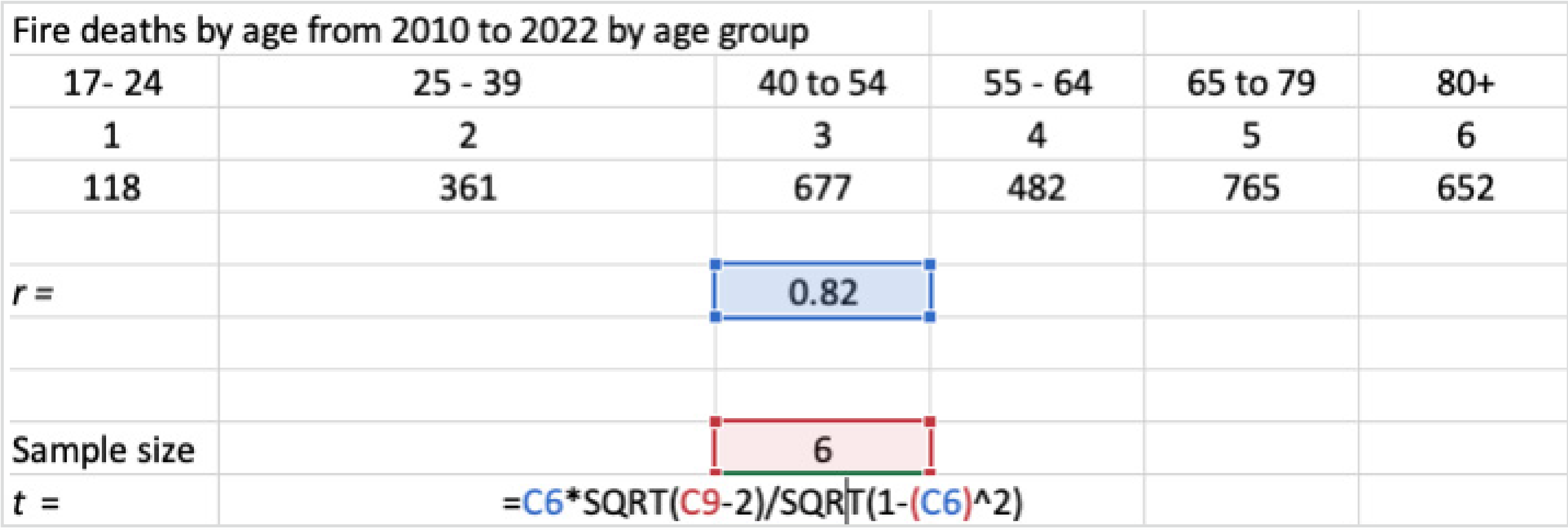
Table 1. Description of correlation strength

Value of r	Strength of Relationship
R < 0.3	None or very weak
0.3 to < 0.5	Weak
0.5 to < 0.7	Moderate
< 0.7	Strong

Note - The same wording would be used for a negative r-value.

A t-test compares the means between the two groups to identify how related they are. The test produces a t-value that can then be used to calculate the p-value. Two bits of data are needed to complete a t-test, the sample size and the r-value.

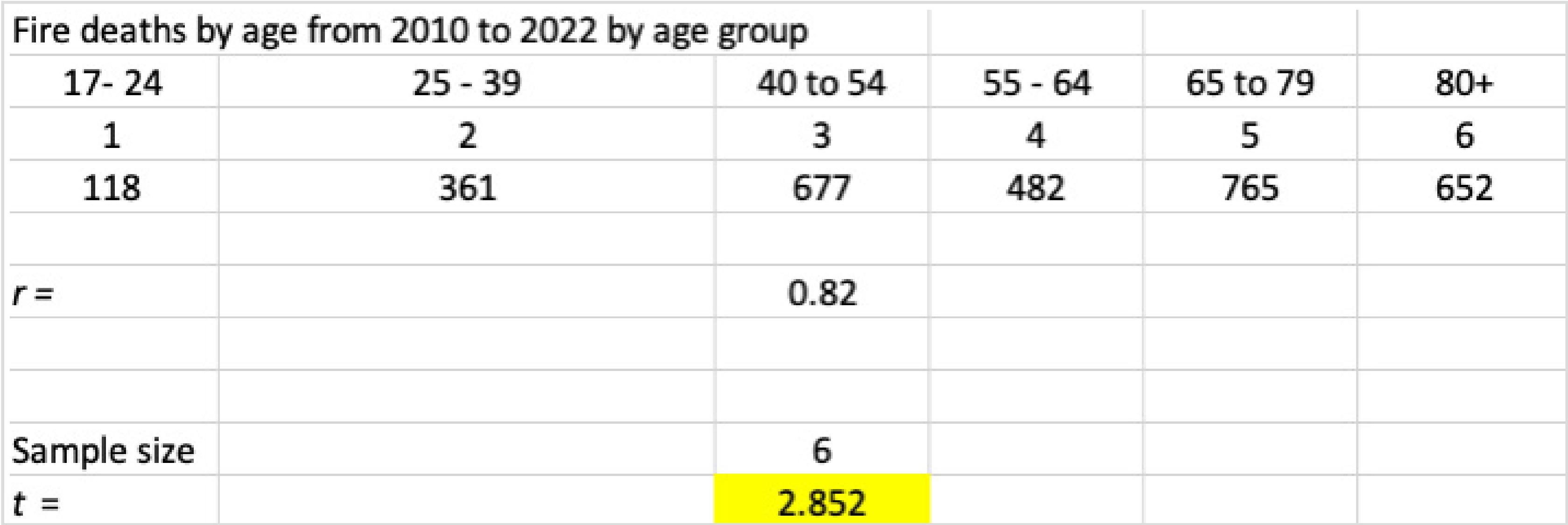
The sample size (N) in this example is 6, as there are 6 age groups, with the r-value being 0.82. The formula used in Excel is: =r*SQRT(number-2)/SQRT(1-(number)^2) (Figures 4 and 5).



Fire deaths by age from 2010 to 2022 by age group

17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+
1	2	3	4	5	6
118	361	677	482	765	652
r =		0.82			
Sample size		6			
t =		=C6*SQRT(C9-2)/SQRT(1-(C6)^2)			

Figure 4. The Excel formula for the t-test =r*SQRT(number-2)/SQRT(1-(number)^2)



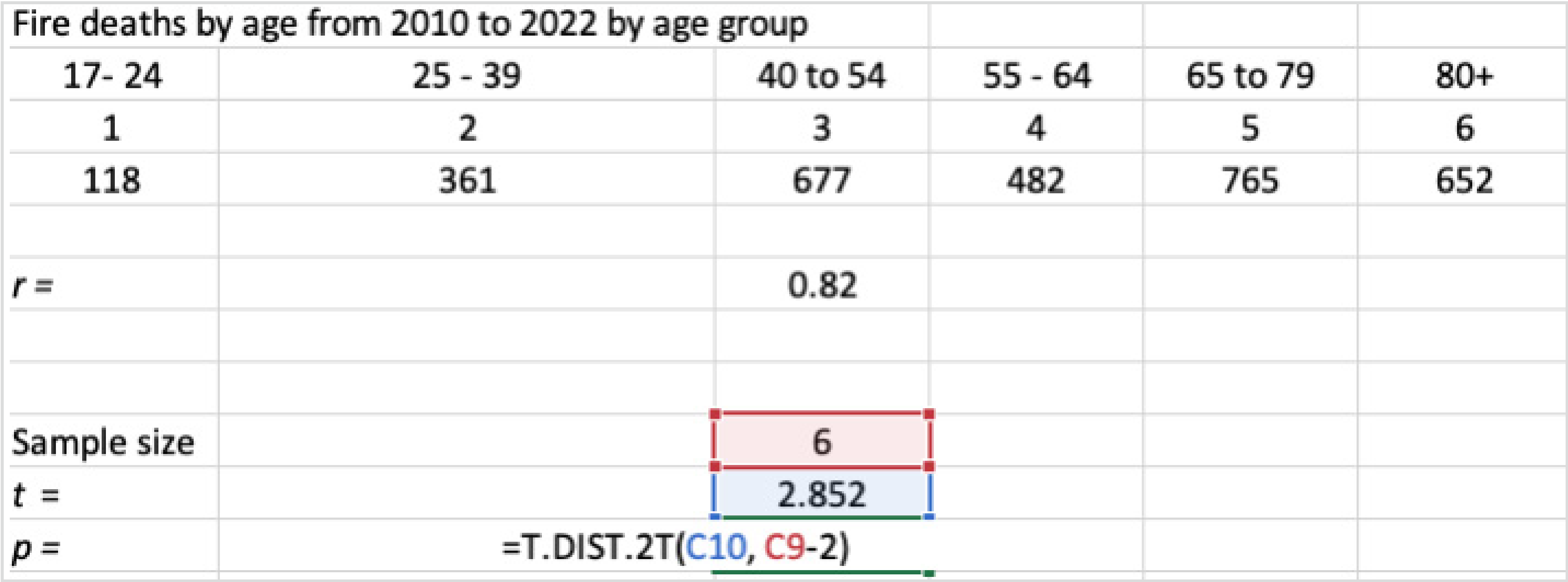
Fire deaths by age from 2010 to 2022 by age group

17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+
1	2	3	4	5	6
118	361	677	482	765	652
r =		0.82			
Sample size		6			
t =		2.852			

Figure 5. The t-value

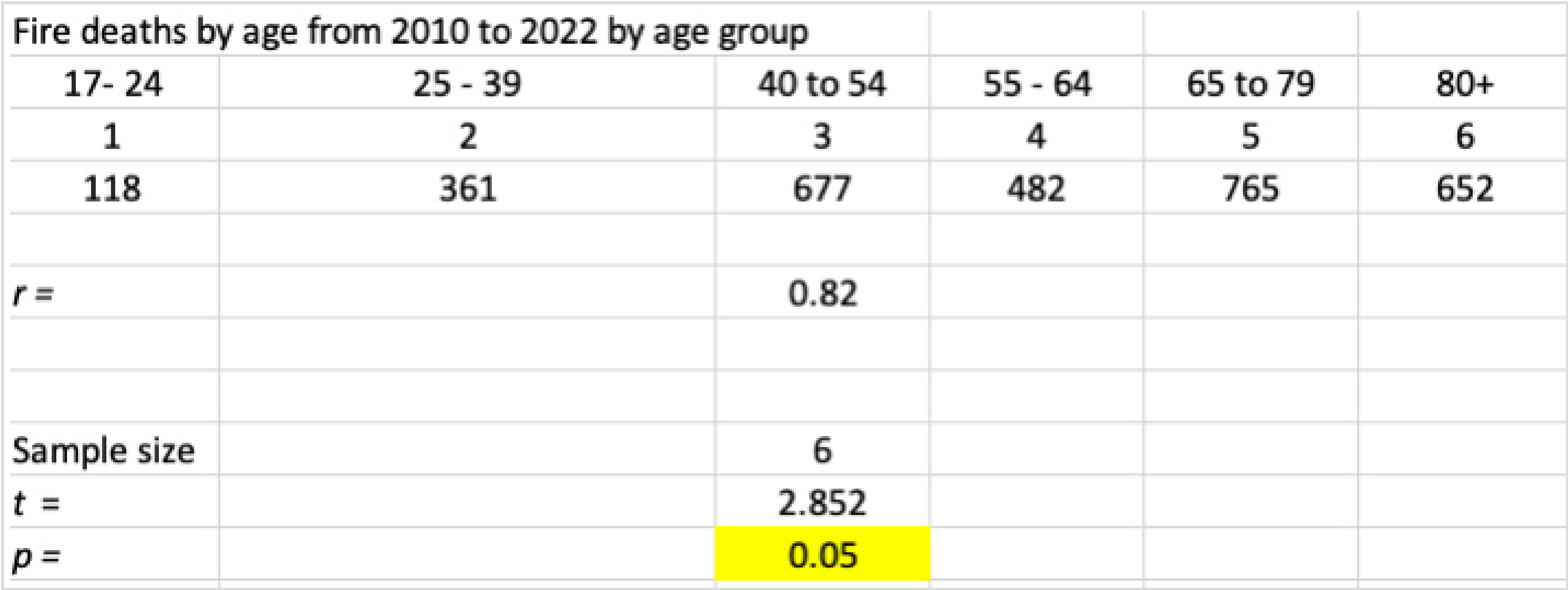
P calculation

Using the t value and the sample size, it is now possible to calculate the value of p. The Excel formula is =T.DIST.2T(x, deg_freedom-2). X is the t value, and the degrees of freedom are worked out automatically on adding the sample size, the 2 indicates this is a 2-way test (Figure 6). Figure 7 shows the p-value is .05. which means the result is statistically significant. To be significant, the value of p needs to be at or below .05.



Fire deaths by age from 2010 to 2022 by age group					
17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+
1	2	3	4	5	6
118	361	677	482	765	652
r =		0.82			
Sample size		6			
t =		2.852			
p =		=T.DIST.2T(C10, C9-2)			

Figure 6. The formulate for the p calculation=T.DIST.2T(x, deg_freedom-2)



Fire deaths by age from 2010 to 2022 by age group					
17- 24	25 - 39	40 to 54	55 - 64	65 to 79	80+
1	2	3	4	5	6
118	361	677	482	765	652
r =		0.82			
Sample size		6			
t =		2.852			
p =		0.05			

Figure 7. Value of p.

The results of the correlation can then be reported in the following way:

> A Pearson correlation test found a strong positive correlation between age and the number of fire-related deaths (r = .82, p = .05).

The relationship is described as positive, as the r-value is a positive number. The word strong is used, which is a statistical term based on accepted levels (Table 1). These terms can be used to describe both positive and negative r-values.

Independent sample t-test

Why use an independent t-test?

An independent sample t-test (sometimes referred to as a two-sample t-test) compares the means between two unrelated groups to identify if any difference is significant. The test produces a t-value that can then be used to calculate the p-value. The independent sample t-test assesses differences between **unrelated** groups, for example, between a control group and an intervention group, to ascertain if any difference is real and not simply a chance occurrence.

Conducting an Independent sample t-test in Excel

Please note that this is an imagined example for demonstration purposes.

There are two types of independent t-tests, you can see the formula for both tests here. The selection of which test to use is dependent on whether the variance in the data is equal. This is explained in more depth in a moment. An evaluation may want to identify if a group have retained the knowledge gained 3 months after receiving an educational intervention on water safety. To assess this, the group that received the intervention, the Intervention Group (IG), were compared with a group who have not received the intervention, the Control Group (CG). The data being analysed came from the results of a knowledge test administered to both groups, with 40 people completing the test in both the IG and CG. The test contains 10 questions, with one point being allocated for each correct answer. Scores were then totalled to provide an overall score. A review of the distribution showed the data to be normally distributed.

As previously mentioned, there are two types of independent samples t-tests, one assumes equal variance, and the second assumes unequal variance. Variance is important in t-tests as the test makes assumptions about the distribution of the data, therefore it is important to select the correct test. To do this is necessary to complete an F-test, which will identify if the variance is or is not equal.

To complete the F-test, select Data from the worksheet toolbar and then Data Analysis. The menu shown in Figure 8 will then appear. From the menu, select the F-Test Two-Sample for Variance. If this menu box does not appear, ensure the Analysis ToolPak has been downloaded, instructions on how to do this can be found at: https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

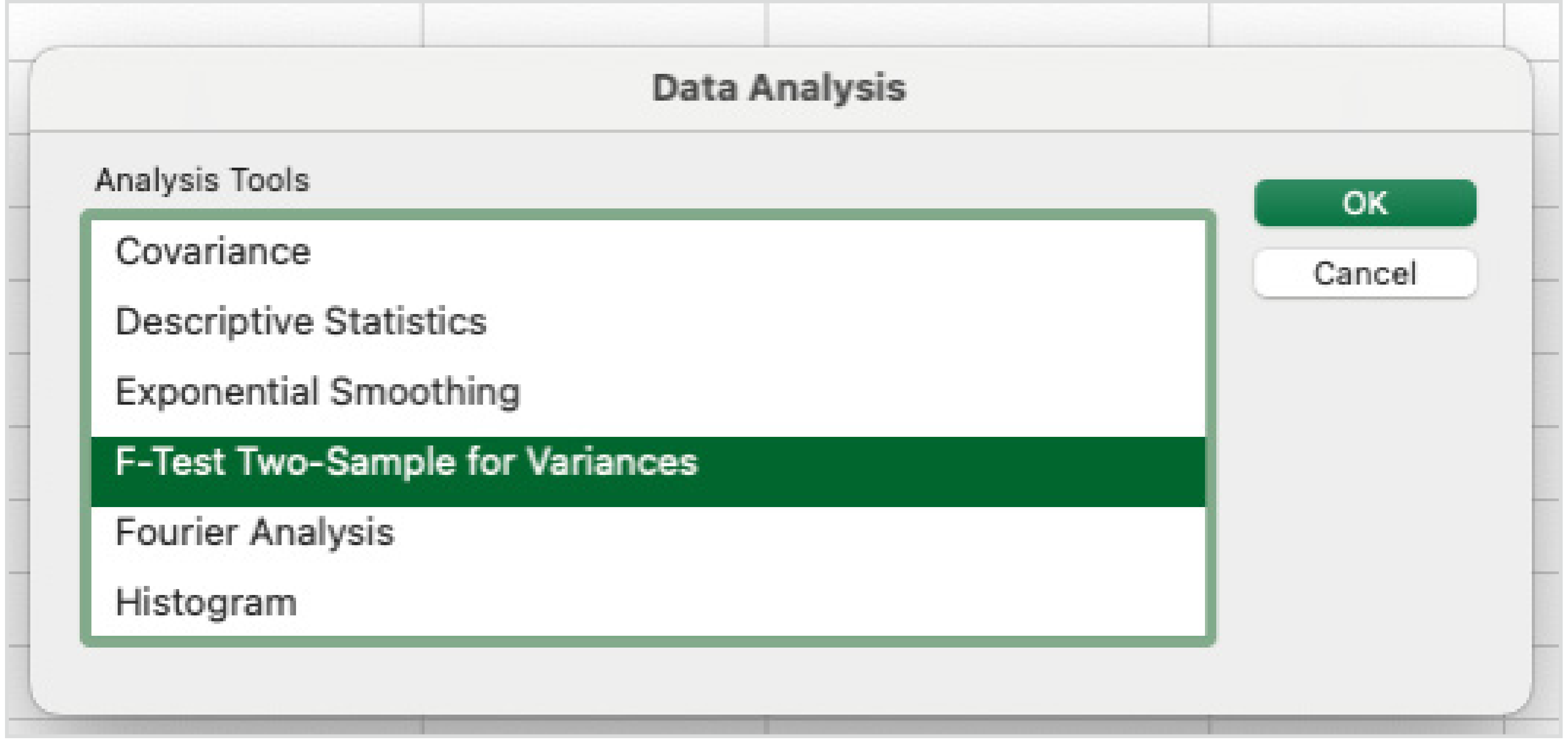
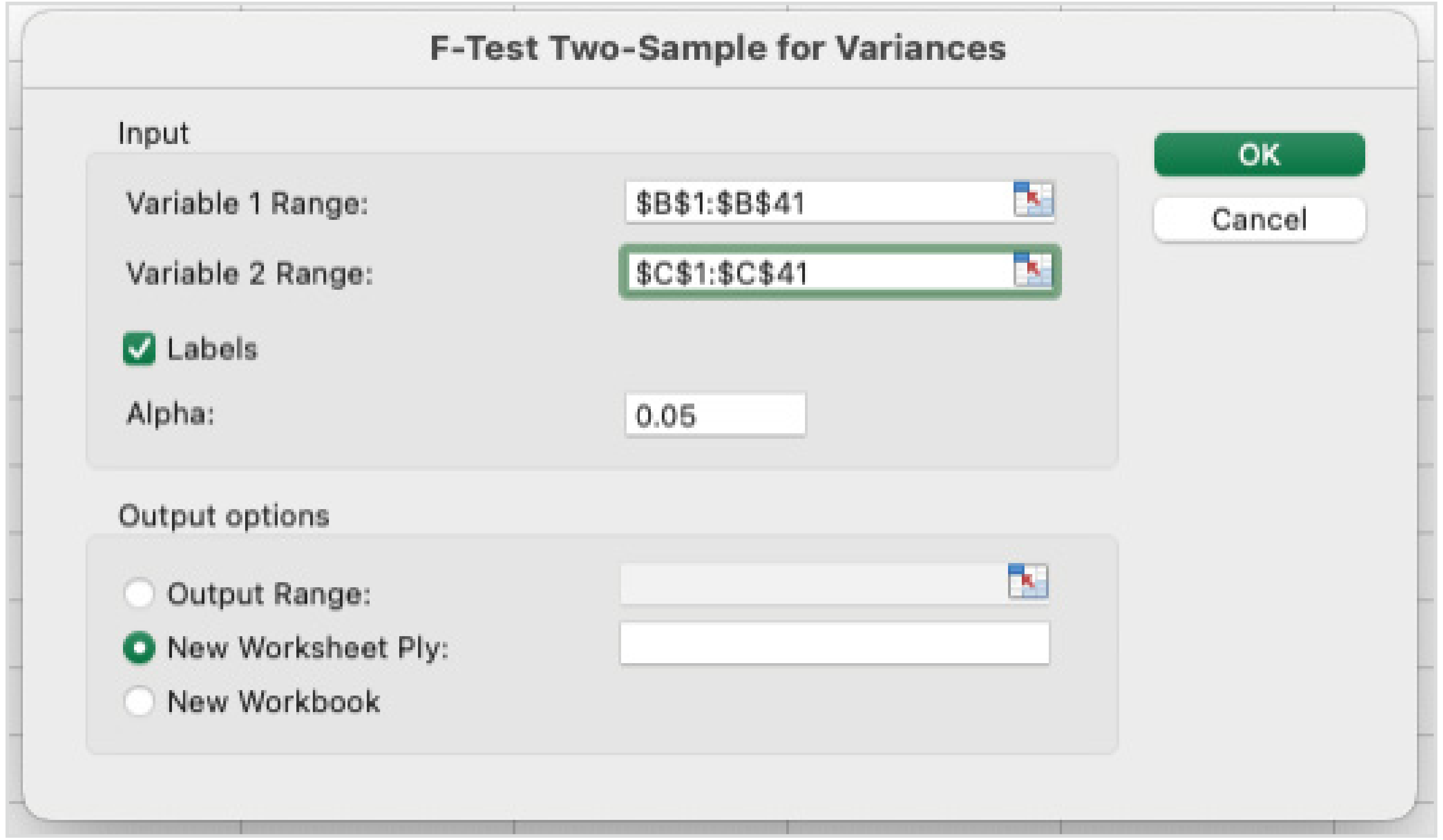


Figure 8. Data Analysis Menu

This opens the test input box (Figure 9). In Variable 1 Range select the data range for one of the sample groups and then do the same for the second sample groups in the Variable 2 Range box. If you have included the names of the variables when selecting the data ranges, tick the Labels. The name of the data range will then be shown in the output. The alpha level by default is set at .05, which is correct. The output options can be left as the default, this means the result will be displayed on a new worksheet. Then click OK.



F-Test Two-Sample for Variances

Input

Variable 1 Range: $B$1:$B$41

Variable 2 Range: $C$1:$C$41

☑ Labels

Alpha: 0.05

Output options

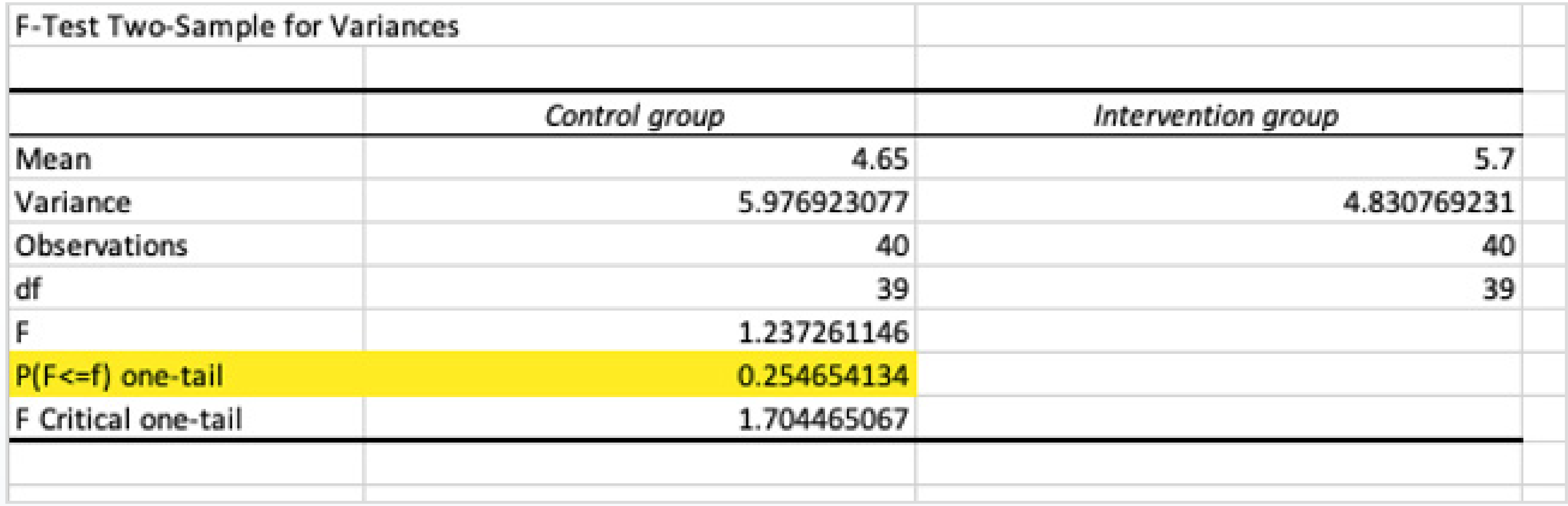
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 9. F-Test - Sample for Variance input box

Figure 10 shows the test output with the p-value highlighted. If the p-value is above .05, then the variance between the two groups is not significant, therefore the **t-Test: 2 Sample Assuming Equal Variances** should be used. If this p-value is less than .05, then the variance between the sample is significant, and the **t-test: 2 Sample Assuming Unequal Variances** should be used. Both tests are conducted in the same way.

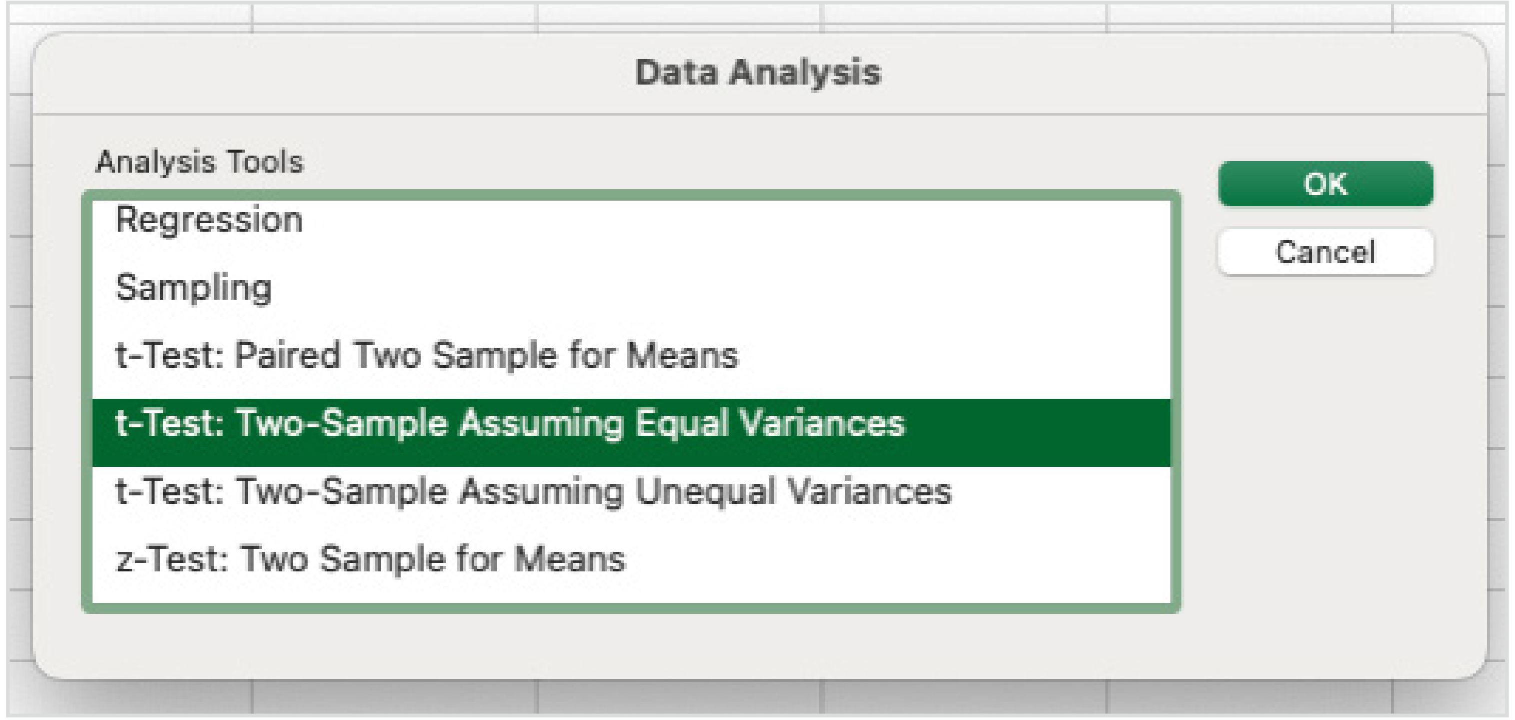


F-Test Two-Sample for Variances

	Control group	*Intervention group*
Mean	4.65	5.7
Variance	5.976923077	4.830769231
Observations	40	40
df	39	39
F	1.237261146	
P(F<=f) one-tail	0.254654134	
F Critical one-tail	1.704465067	

Figure 10. F-test results output

To select the t-test select Data, then Analyse data to re-open the test menu (Figure 11).



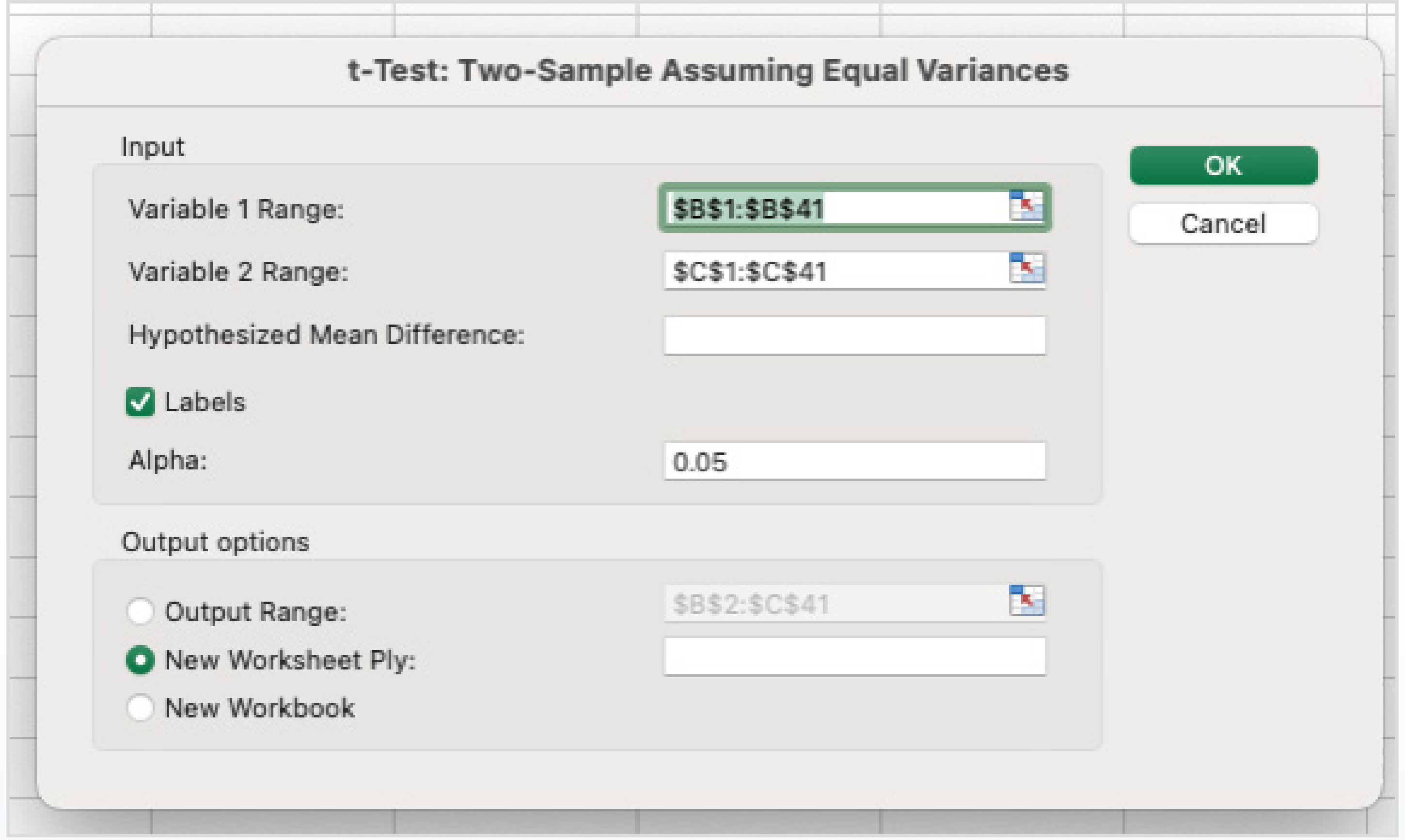
Data Analysis

Analysis Tools

- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means

Figure 11. Data Analysis Menu

Once the required t-test has been selected, the test input box will appear (Figure 12). The data range for groups 1 and 2 can then be entered into the **Variable Range** boxes. Labels should be clicked if the column's names were included in the range selections. The **Alpha** value is .05. No input is required in the **Hypothesized Mean Difference** box. No changes are needed in the **Output Options** boxes.



t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $B$1:$B$41

Variable 2 Range: $C$1:$C$41

Hypothesized Mean Difference:

☑ Labels

Alpha: 0.05

Output options

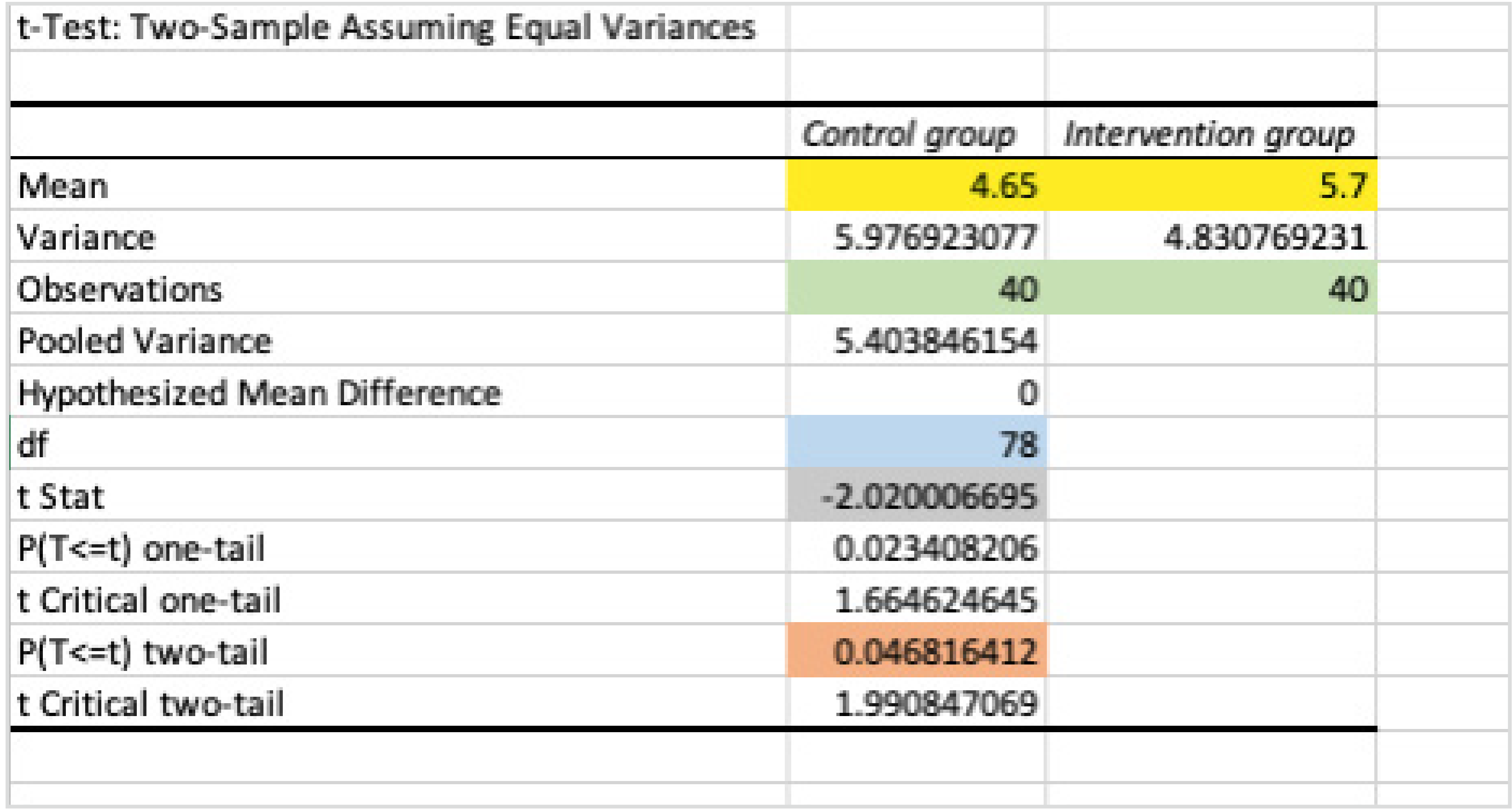
○ Output Range: $B$2:$C$41

● New Worksheet Ply:

○ New Workbook

Figure 12. t-Test: Two-Sample Assuming Equal Variance- input box

The results table (Figure 13) provides a wide range of statistical information that supports the calculations behind the test. The outputs needed for interpreting and reporting the results of the test have been highlighted. The results highlighted in yellow are the mean scores for each group. The results in green are the number of observations which in this case are the number of people in the sample. The blue highlight shows the degrees of freedom (df), the t-statistic is highlighted in grey, with the p-value highlighted in orange. As the value of p is less than .05 the change between the two mean scores is statistically significant. **Note, the Standard Deviation is not given but is needed to correctly report the results** (click here to see how to calculate a Standard Deviation in Excel).



t-Test: Two-Sample Assuming Equal Variances

	Control group	*Intervention group*
Mean	4.65	5.7
Variance	5.976923077	4.830769231
Observations	40	40
Pooled Variance	5.403846154	
Hypothesized Mean Difference	0	
df	78	
t Stat	-2.020006695	
P(T<=t) one-tail	0.023408206	
t Critical one-tail	1.664624645	
P(T<=t) two-tail	0.046816412	
t Critical two-tail	1.990847069	

Figure 13. t-Test: Two-Samples Assuming Variance Results output

The results of the test could be written in the following way:

> Three months post-intervention, the mean score for the control group was 4.65 (SD = 2.44), with the intervention group achieving a higher mean score of 5.5 (SD = 2.20). An independent t-test found this difference to be significant (t = -2.02, df = 78, p = .05).

Paired sample t-test

Why use a paired sample t-test?

A paired sample t-test should be used when the sample is **related** and can be matched. For example, if the same group were asked to complete a questionnaire on two occasions, the two questionnaires could be paired up by a unique identifier, such as an email address.

A paired sample t-test compares the mean scores between the same group at different time points to identify if any difference found in the mean scores is significant and not simply a chance occurrence. The test produces a t-value that can then be used to calculate the p-value. The test formula can be viewed here.

Conducting a Paired Sample t-test in Excel

Please note that this is an imagined example for demonstration purposes.

In this example, the evaluator is looking to identify what impact attending a FRS advanced driving course has had on the attending firefighter's willingness to speed. To do this, they send out a questionnaire just prior to the firefighters attending the course and then another immediately after the course. The questions use a 5-point scale asking them to rate how willing they would be to break the speed limit in several different scenarios, with 1 being very willing and 5 being very unwilling.

Question 1, at both Time 1 (pre-intervention) and Time 2 (post-intervention), asks the firefighters, 'How willing would you be to break the speed limited if you were late for an important appointment'. The mean score pre-course was 2.7 (SD = 0.94), post-course the mean score had improved to 3.15 (SD = 1.05), a difference between means of 0.45. To identify if this change was significant, a paired sample t-test could be performed as the pre- and post-questionnaires could be matched, and a review of the distribution found that the data was normally distributed.

To do this test on Excel, select Data, then Data Analysis. This will open the test menu screen (Figure 14). (If this menu is not available, download and install the Analysis Tool pack at https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4).

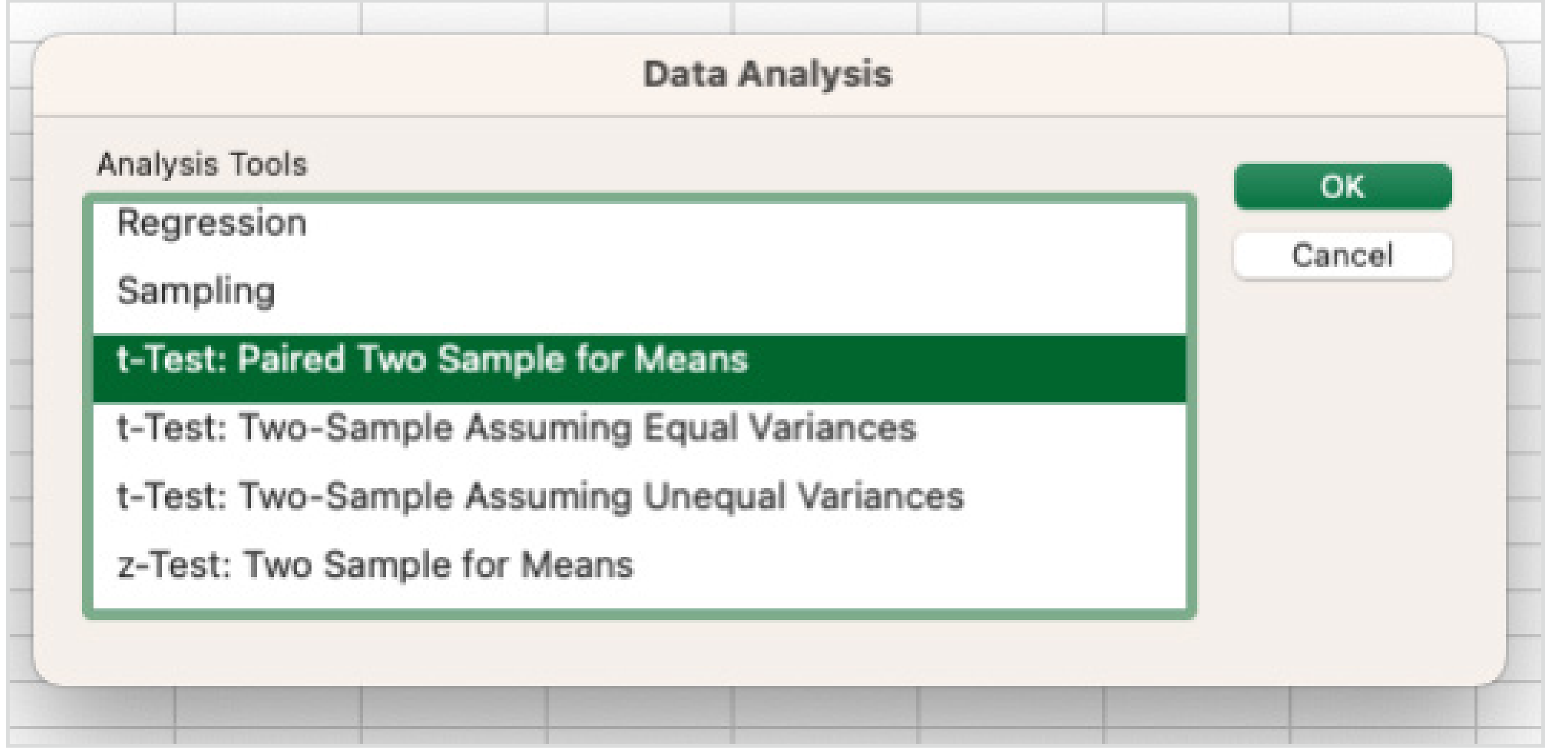
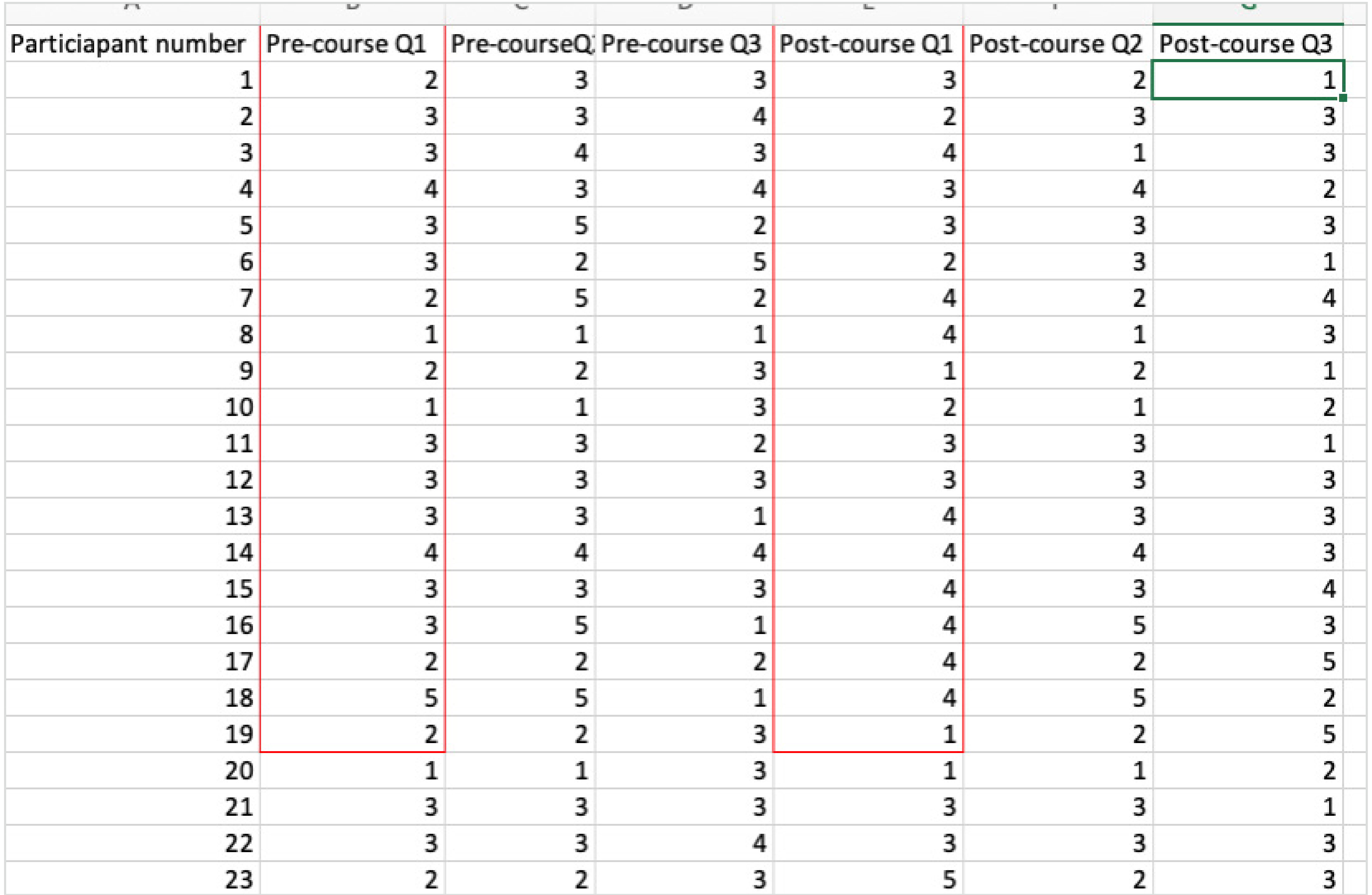
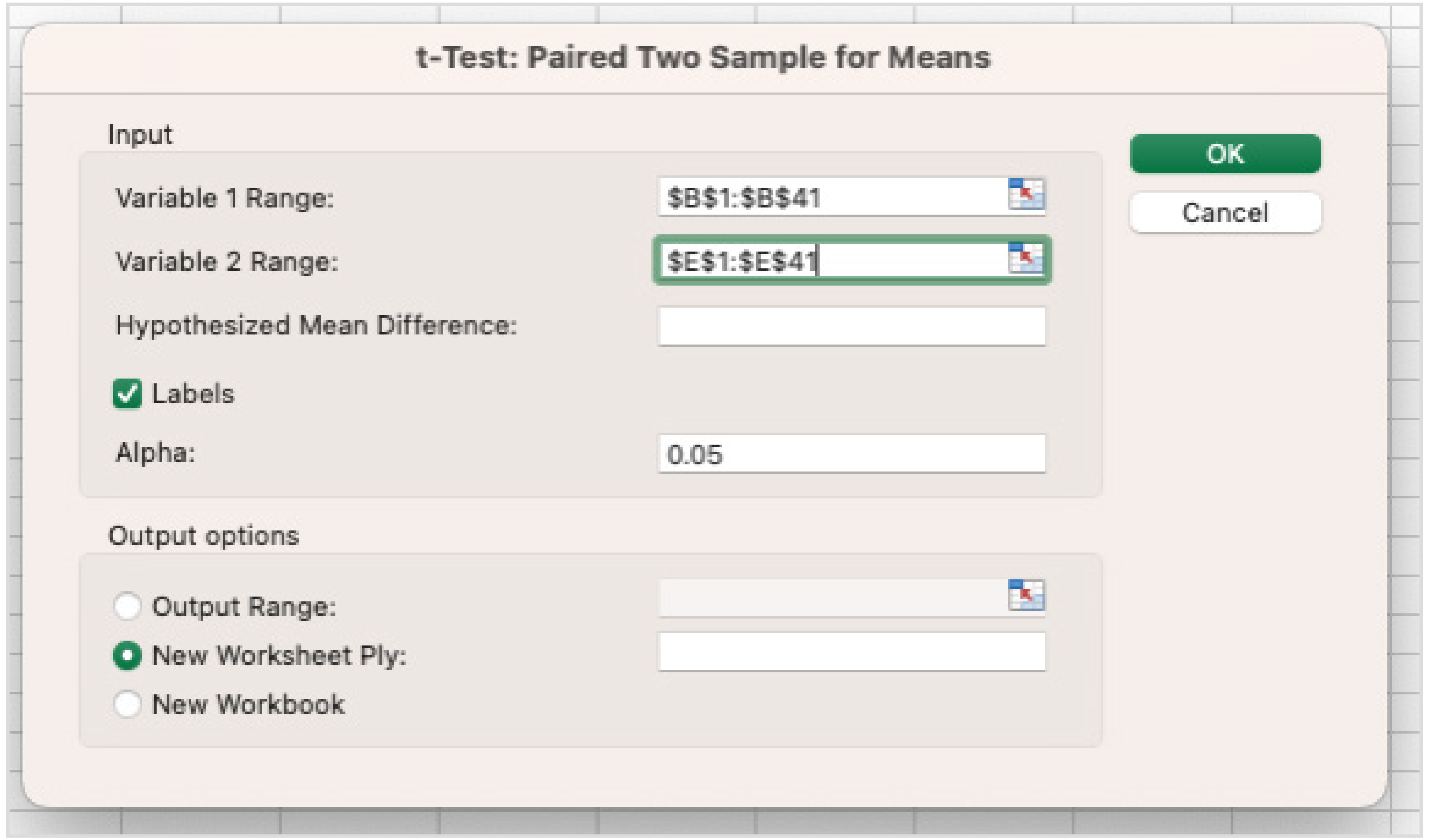

Data Analysis
Analysis Tools
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel

Figure 14. Data Analysis Menu

From the menu select **t-Test: Paired Two Sample for Means,** then click OK. This will open the test input box (Figure 15). Select the **Variable 1 Range** for the pre-intervention data and the **Variable 2 Range** data for the post-intervention results. These are part shown in Figure 15, as it is difficult to show the full 40 entries. If **Labels** is selected, the name of the data range given at the top of the column will be shown if included in the data range selection. The **Hypothesized Mean Difference** box should be left empty. The **alpha** level is already set as .05 by default, which is acceptable for this type of analysis. The Output options can be left as the default, the result will be displayed in new worksheet.



Particiapant number	Pre-course Q1	Pre-courseQ	Pre-course Q3	Post-course Q1	Post-course Q2	Post-course Q3
1	2	3	3	3	2	1
2	3	3	4	2	3	3
3	3	4	3	4	1	3
4	4	3	4	3	4	2
5	3	5	2	3	3	3
6	3	2	5	2	3	1
7	2	5	2	4	2	4
8	1	1	1	4	1	3
9	2	2	3	1	2	1
10	1	1	3	2	1	2
11	3	3	2	3	3	1
12	3	3	3	3	3	3
13	3	3	1	4	3	3
14	4	4	4	4	4	3
15	3	3	3	4	3	4
16	3	5	1	4	5	3
17	2	2	2	4	2	5
18	5	5	1	4	5	2
19	2	2	3	1	2	5
20	1	1	3	1	1	2
21	3	3	3	3	3	1
22	3	3	4	3	3	3
23	2	2	3	5	2	3



t-Test: Paired Two Sample for Means

Input

Variable 1 Range: $B$1:$B$41

Variable 2 Range: $E$1:$E$41

Hypothesized Mean Difference:

☑ Labels

Alpha: 0.05

Output options

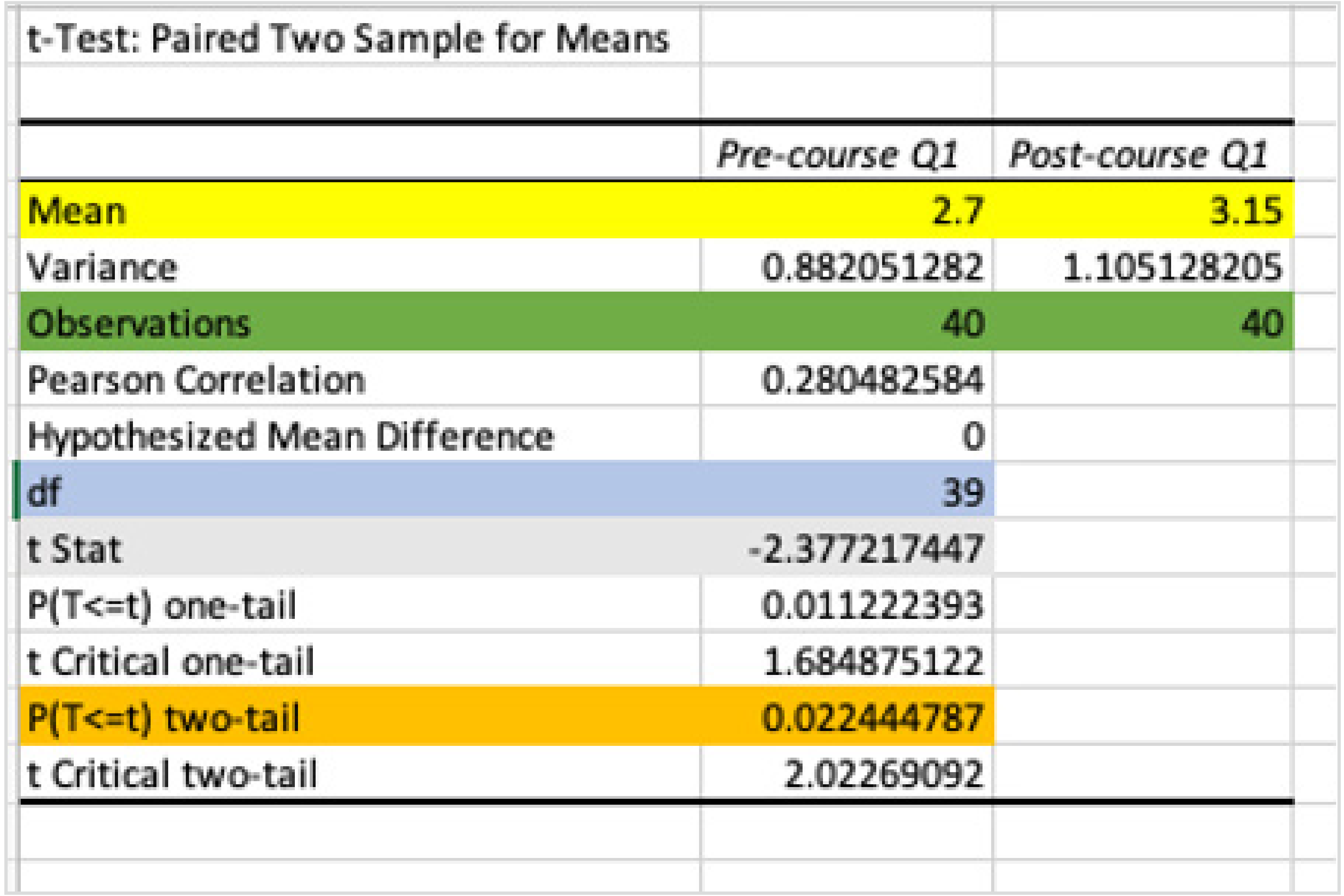
○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 15. t-Test: Paired Two Sample for Means – input box and Worksheet.

The results table (Figure 16) provides a range of statistical information that supports the calculations. The outputs needed for interpreting and reporting the results of the test have been highlighted. The results highlighted in yellow are the mean scores for each time point. The results in green are the number of observations, in this case, the number of people in the sample. Blue highlights the degrees of freedom (df), the t-statistic is in grey, with the two-tail p-value being in orange. As the value of p is less than .05, the change seen between the two mean scores is statistically significant. **Note** the Standard Deviation is not given but is needed to correctly report the results (click here to see how to calculate this).



t-Test: Paired Two Sample for Means		
	Pre-course Q1	*Post-course Q1*
Mean	2.7	3.15
Variance	0.882051282	1.105128205
Observations	40	40
Pearson Correlation	0.280482584	
Hypothesized Mean Difference	0	
df	39	
t Stat	-2.377217447	
P(T<=t) one-tail	0.011222393	
t Critical one-tail	1.684875122	
P(T<=t) two-tail	0.022444787	
t Critical two-tail	2.02269092	

Figure 16. t-Test: Paired Sample for Means - results output

The results of the test can then be written up in the following way.

> Forty participants were tested pre- and immediately post-training. Pre-training, the mean score for question 1 was 2.7 (SD = 0.94). Immediately post-course, this had improved to a mean of 3.15 (SD = 1.05). A paired sample t-test found the improvement to be significant (t = -2.37, df = 39, p = .02). This result indicates that the participants were less willing to speed when late for a meeting after they had completed the training course.

Analysis of Variance (Anova)

Why use an Analysis of Variance test?

An evaluation design often contains the same measure administered at different time points. For example, an evaluation of an open water safety intervention may want to assess if the participant's knowledge initially improved and if any improvement was maintained over time. To do this, the participant could be assessed at three time points: just prior to commencing the intervention, immediately post-intervention and 3-month post-intervention.

To identify if there is a significant difference, several paired sample t-tests comparing the three time points could be completed. However, a p-value of .05 means there is a 1 in 20 chance of an error occurring. Therefore, completing multiple t-tests introduces a large element of chance into the results. To reduce the possibility of error, not only should the t-tests be performed but so should an Analysis of Variance (Anova). This test looks specifically for differences with 3 or more scores. There are a range of Anova tests, and the evaluator will need to consider which test to use, more information about these tests can be found at: https://www.scribbr.com/statistics/one-way-anova/

Conducting a repeat measure ANOVA test in Excel
To perform an ANOVA, go to Data then Data Analysis which will open the menu box shown in Figure 17. From the menu, select Anova: Two-factor Without Replication (Figure 17).

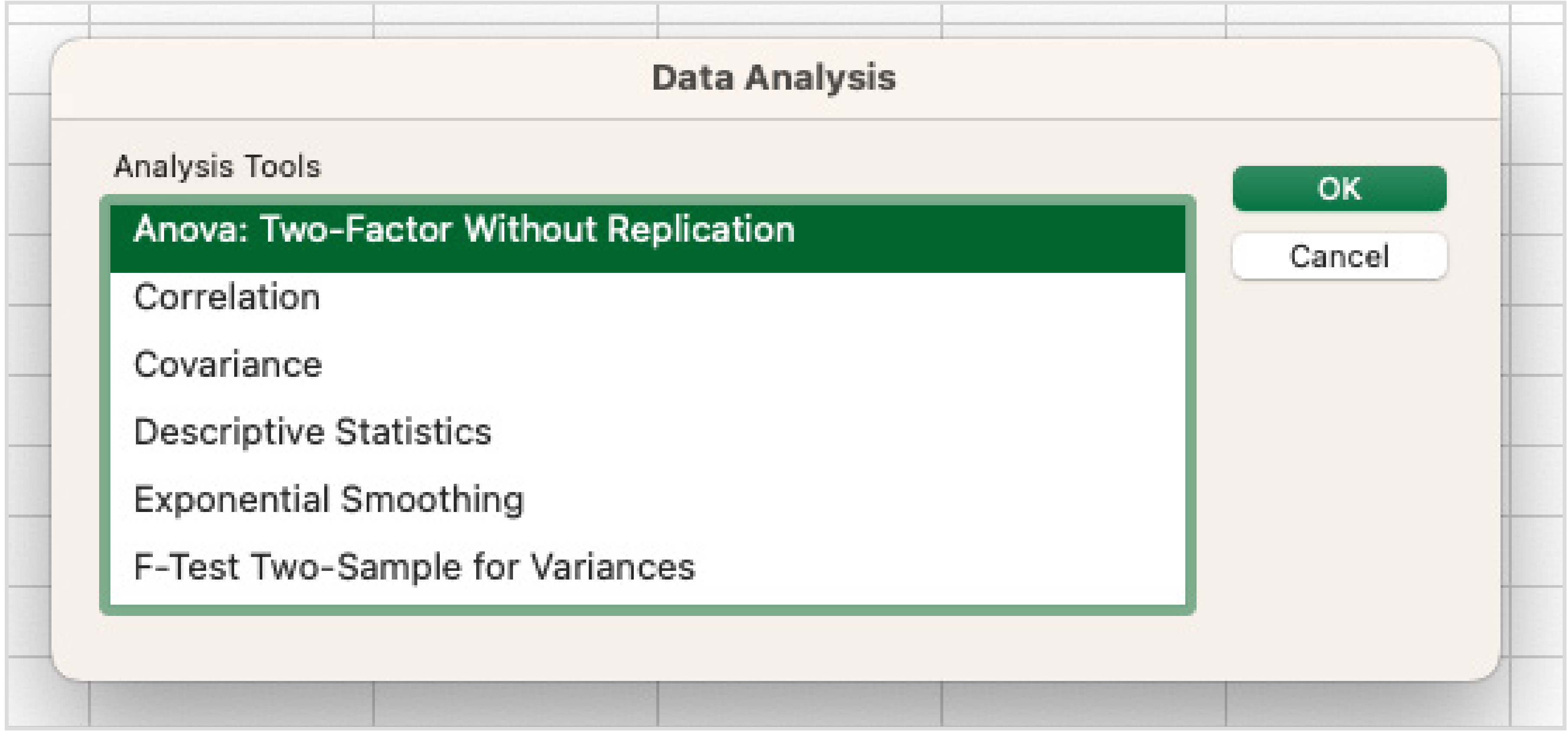
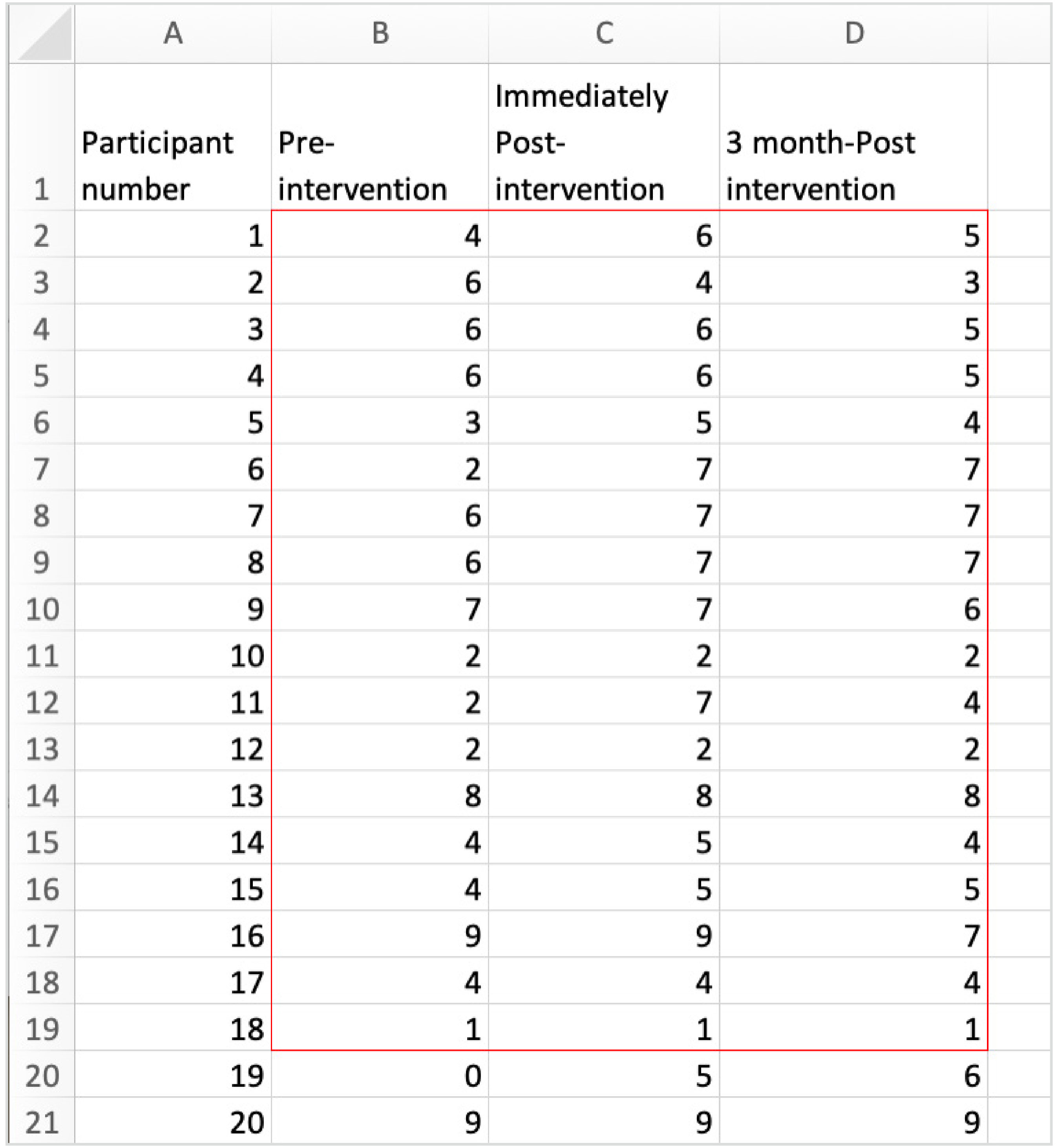
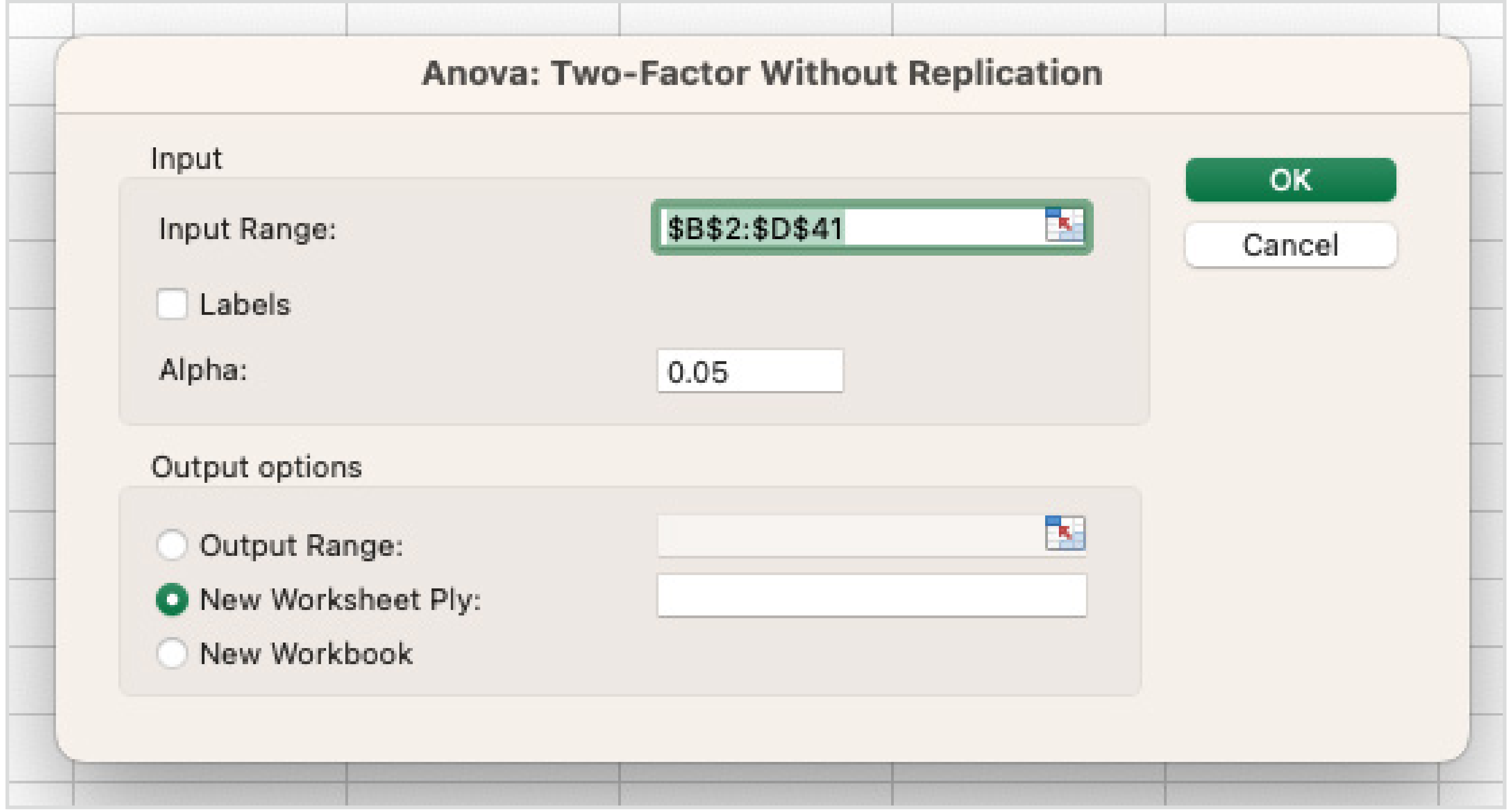


Figure 17. Data Analysis Menu

On selecting OK, this input box shown in Figure 18 will appear. In the **Input Range**, select the data range to be used. **In the example shown, only part of the selection is shown, there are 40 participants in the sample.**



	A	B	C	D
1	Participant number	Pre-intervention	Immediately Post-intervention	3 month-Post intervention
2	1	4	6	5
3	2	6	4	3
4	3	6	6	5
5	4	6	6	5
6	5	3	5	4
7	6	2	7	7
8	7	6	7	7
9	8	6	7	7
10	9	7	7	6
11	10	2	2	2
12	11	2	7	4
13	12	2	2	2
14	13	8	8	8
15	14	4	5	4
16	15	4	5	5
17	16	9	9	7
18	17	4	4	4
19	18	1	1	1
20	19	0	5	6
21	20	9	9	9



Anova: Two-Factor Without Replication

Input

Input Range: $B$2:$D$41

□ Labels

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Figure 18. Anova: Two-factor without replication - input box and worksheet

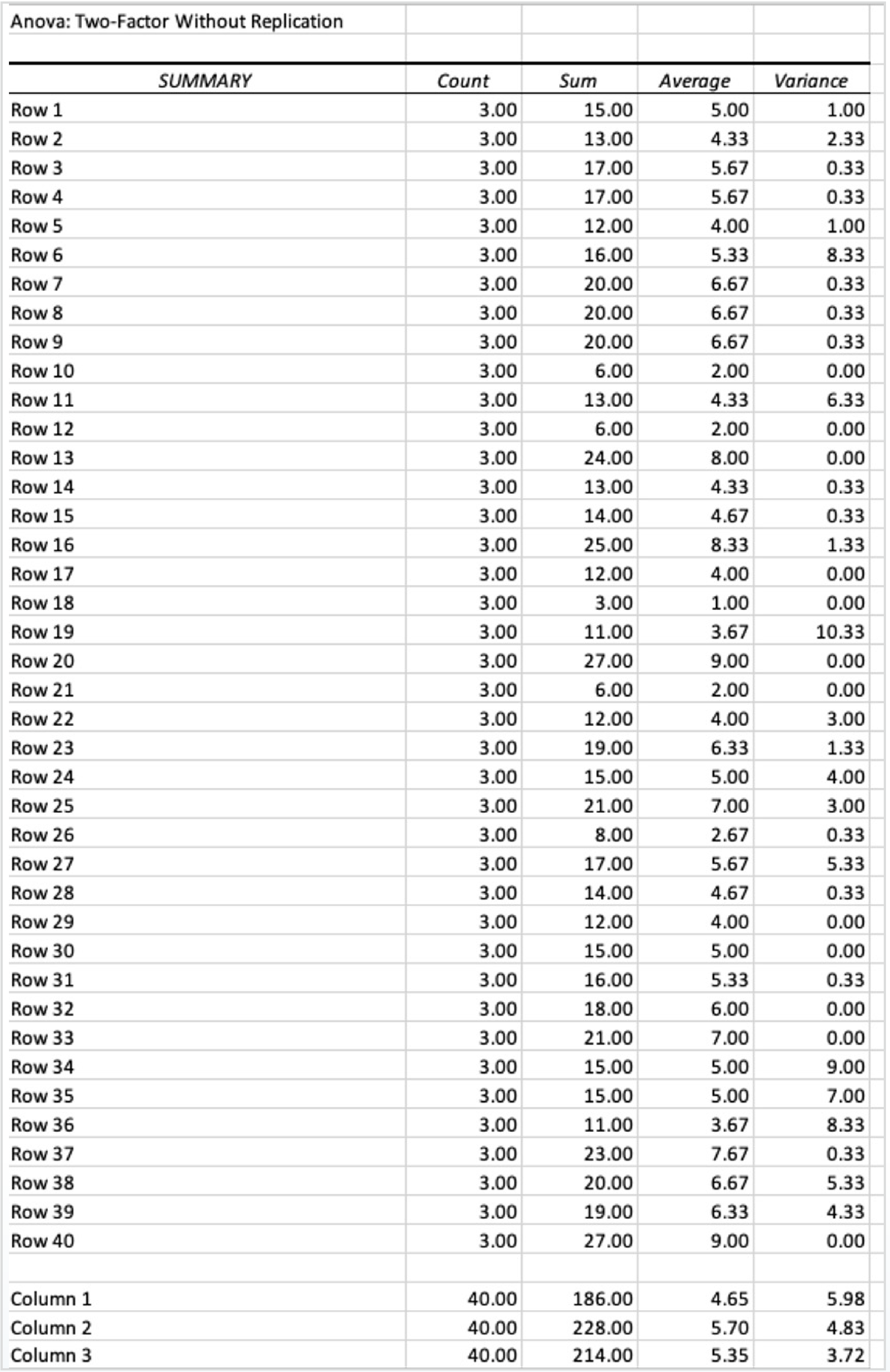
Leave the Labels tick box blank. The Alpha values should be set to 0.05. Select where the Output is to be shown. Then press OK.

Two tables will then appear, a summary table and a results table. The summary table (Figure 19) shows each participant as a row. In this instance, Count refers to how many of the questionnaires the participant had completed. The Sum is the participant’s total score across all three assessments. Average is the mean score across all 3 assessment points, Variance is the amount of difference across the scores. The scores are summarised at the bottom of the summary table. In this instance, there were 40 participants. The mean score for Column 1 (which was the pre-intervention assessment) was 4.65, the mean score one-month post-intervention assessment was 5.7, with the mean score reducing to 5.35 at the 3-month point. The results of the ANOVA test are shown in the ANOVA table (Figure 20).

As the data was arranged in columns (See Figure 18), it is the column score that needs to be interpreted and reported. The F-statistic is 6.06, with 2 degrees of freedom with 78 degrees of error, the p-value is .003. As the p-value is less than .05, it can be concluded that any change found in the scores is significant. A set of paired sample t-tests should now be completed, comparing the:

- Pre-intervention scores with the immediate post-intervention scores
- Pre-intervention scores with the 3-month post-intervention scores
- The immediate post-intervention scores with the 3-month post-intervention scores

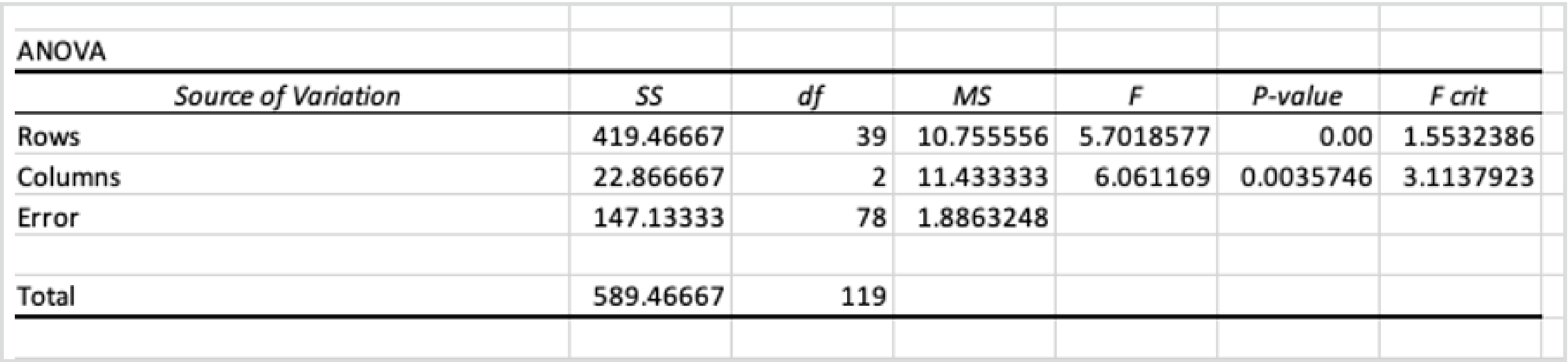
However, a Bonferroni correction needs to be applied to the results of the t-tests. This correction aims to reduce the likelihood of error occurring when multiple t-test are conducted by adjusting the alpha value. The correction is calculated by dividing the alpha value by the number of tests. In this case, the alpha value is .05, and there will be 3 tests conducted. Therefore, to be significant, the p-value for the t-tests would have to be at or below .02. (.05 / 3 = .02 when rounded to 2 decimal points).



Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
Row 1	3.00	15.00	5.00	1.00
Row 2	3.00	13.00	4.33	2.33
Row 3	3.00	17.00	5.67	0.33
Row 4	3.00	17.00	5.67	0.33
Row 5	3.00	12.00	4.00	1.00
Row 6	3.00	16.00	5.33	8.33
Row 7	3.00	20.00	6.67	0.33
Row 8	3.00	20.00	6.67	0.33
Row 9	3.00	20.00	6.67	0.33
Row 10	3.00	6.00	2.00	0.00
Row 11	3.00	13.00	4.33	6.33
Row 12	3.00	6.00	2.00	0.00
Row 13	3.00	24.00	8.00	0.00
Row 14	3.00	13.00	4.33	0.33
Row 15	3.00	14.00	4.67	0.33
Row 16	3.00	25.00	8.33	1.33
Row 17	3.00	12.00	4.00	0.00
Row 18	3.00	3.00	1.00	0.00
Row 19	3.00	11.00	3.67	10.33
Row 20	3.00	27.00	9.00	0.00
Row 21	3.00	6.00	2.00	0.00
Row 22	3.00	12.00	4.00	3.00
Row 23	3.00	19.00	6.33	1.33
Row 24	3.00	15.00	5.00	4.00
Row 25	3.00	21.00	7.00	3.00
Row 26	3.00	8.00	2.67	0.33
Row 27	3.00	17.00	5.67	5.33
Row 28	3.00	14.00	4.67	0.33
Row 29	3.00	12.00	4.00	0.00
Row 30	3.00	15.00	5.00	0.00
Row 31	3.00	16.00	5.33	0.33
Row 32	3.00	18.00	6.00	0.00
Row 33	3.00	21.00	7.00	0.00
Row 34	3.00	15.00	5.00	9.00
Row 35	3.00	15.00	5.00	7.00
Row 36	3.00	11.00	3.67	8.33
Row 37	3.00	23.00	7.67	0.33
Row 38	3.00	20.00	6.67	5.33
Row 39	3.00	19.00	6.33	4.33
Row 40	3.00	27.00	9.00	0.00
Column 1	40.00	186.00	4.65	5.98
Column 2	40.00	228.00	5.70	4.83
Column 3	40.00	214.00	5.35	3.72

Figure 19. Anova: Two-factor Without Replication - Summary



ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	419.46667	39	10.755556	5.7018577	0.00	1.5532386
Columns	22.866667	2	11.433333	6.061169	0.0035746	3.1137923
Error	147.13333	78	1.8863248			
Total	589.46667	119				

Figure 20. Anova results table

The results of the tests could then be reported as (the numbers in brackets next to the letter F refer to the degrees of freedom and the degrees of error):

An ANOVA test was completed comparing the participant's scores across the three time points. The results identified that there was a significant difference in scores (F(3, 78) = 6.06, p = .003) between time points. A set of paired sample t-tests were then performed incorporating the Bonferroni correction, which identified a significant improvement in scores immediately post-intervention compared to the pre-intervention scores. Whilst there was an improvement in the mean score 3 months post-intervention compared to the pre-intervention mean score, this was not found to be significant. No significant difference was detected between the immediate post-intervention score and the 3-month post-intervention score.

Table 2 shows how the results of the paired sample t-test could be presented in tabulated form.

Table 2. Results of a paired sample t-test

	N =	Mean	SD	t	df	p=
Pre intervention	40	4.65	2.44	-.374	39	.003*
Immediately - post intervention	40	5.7	2.20			
Pre-intervention	40	4.65	2.44	-1.983	39	.05
3-month post-intervention	40	5.35	1.92			
Immediately - post intervention	40	4.65	2.44	1.617	39	.11
3-month post-intervention	40	5.35	1.92			

* Indicates significant (Bonferroni correction applied)

Non-parametric Tests

Correlation Testing Using the Spearman Rank Correlation Test

Why use the Spearman Rank Correlation Test

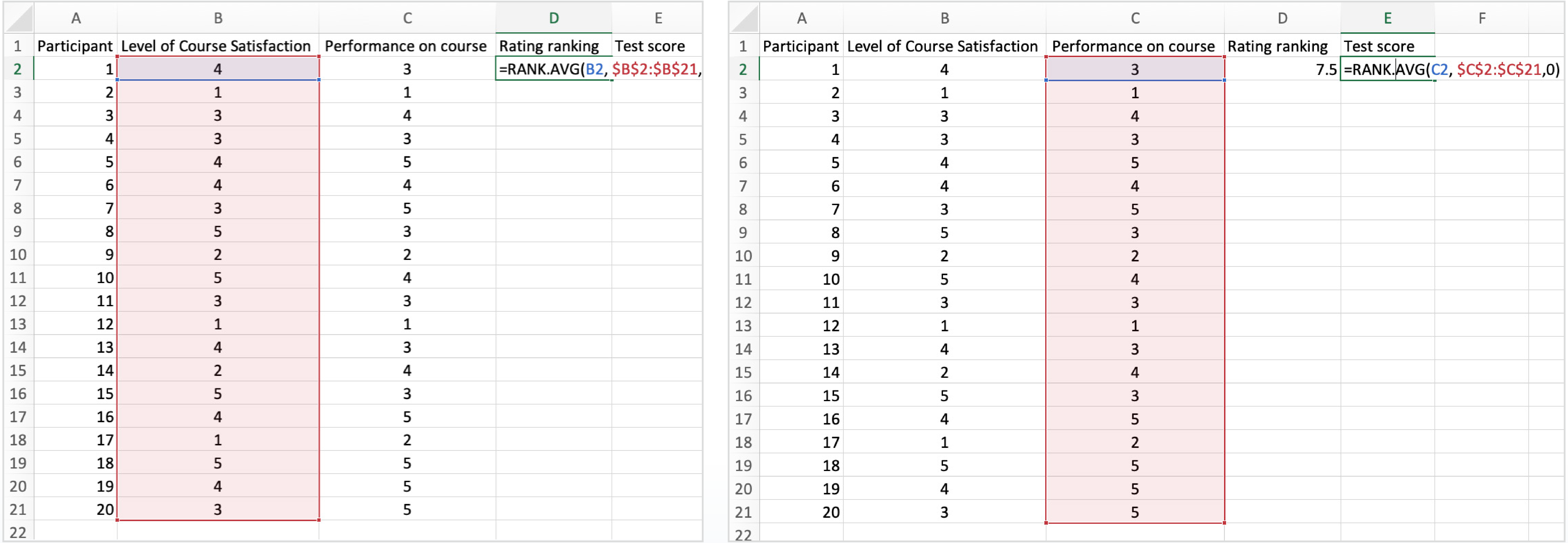
The Spearman Rank Correlation test is a non-parametric alternative to the Pearson Coefficient test. This test should be used when an ordinal scale is being used.

An evaluator may wish to know if there is a correlation **(a relationship)** between one variable and another. For example, an evaluator may want to know if there was a relationship between age and the number of fire deaths in the UK. If there is, then resources can then be targeted appropriately. However, it is important to recognise that a **correlation does not mean that one variable caused another.** For example, the number of firefighters attending a fire may correlate with the amount of damage caused. This would not mean they cause the damage; it is because larger fires are likely to require more firefighters to attend. **As this example demonstrates, it is important that great care is taken by the evaluator when considering if one variable caused another or whether they simply occurred together.**

The test provides a coefficient score (how strong the relationship is) ranging from +1 to -1. A score of +1 would indicate a perfect positive relationship with both variables moving in the same direction at the same rate. A coefficient of -1 would indicate a perfect negative relationship between the variables, with the variables moving in the opposite direction to each other (i.e. one is increasing as one is decreasing). The test achieves this by comparing the ranked variables.

In this **fictitious example,** the evaluator wanted to know if the participant’s overall satisfaction with the training course correlated with how well the participants performed. At the end of the course, the participants were given a grade by the trainers using a 1 to 5 scale, with 1 being a very poor level of performance to 5 being a very high level of performance. The participants were asked, prior to receiving their grading, to give the course an overall satisfaction rating using a 5-point scale, with 1 being very poor and 5 being very good. As the exact margin between the rating on both scales was subjective, and the sample size was small, a Spearman Rank Correlation was used.

The first step is to rank cases, to do this, based on the example shown in Figure 21, the following formula should be entered into Excel =RANK.AVG(B2, $B$2:$B$21,0) for the levels of customer satisfaction and with the formula for Performance rating being =RANK.AVG(C2, $C$2:$C$21,0).



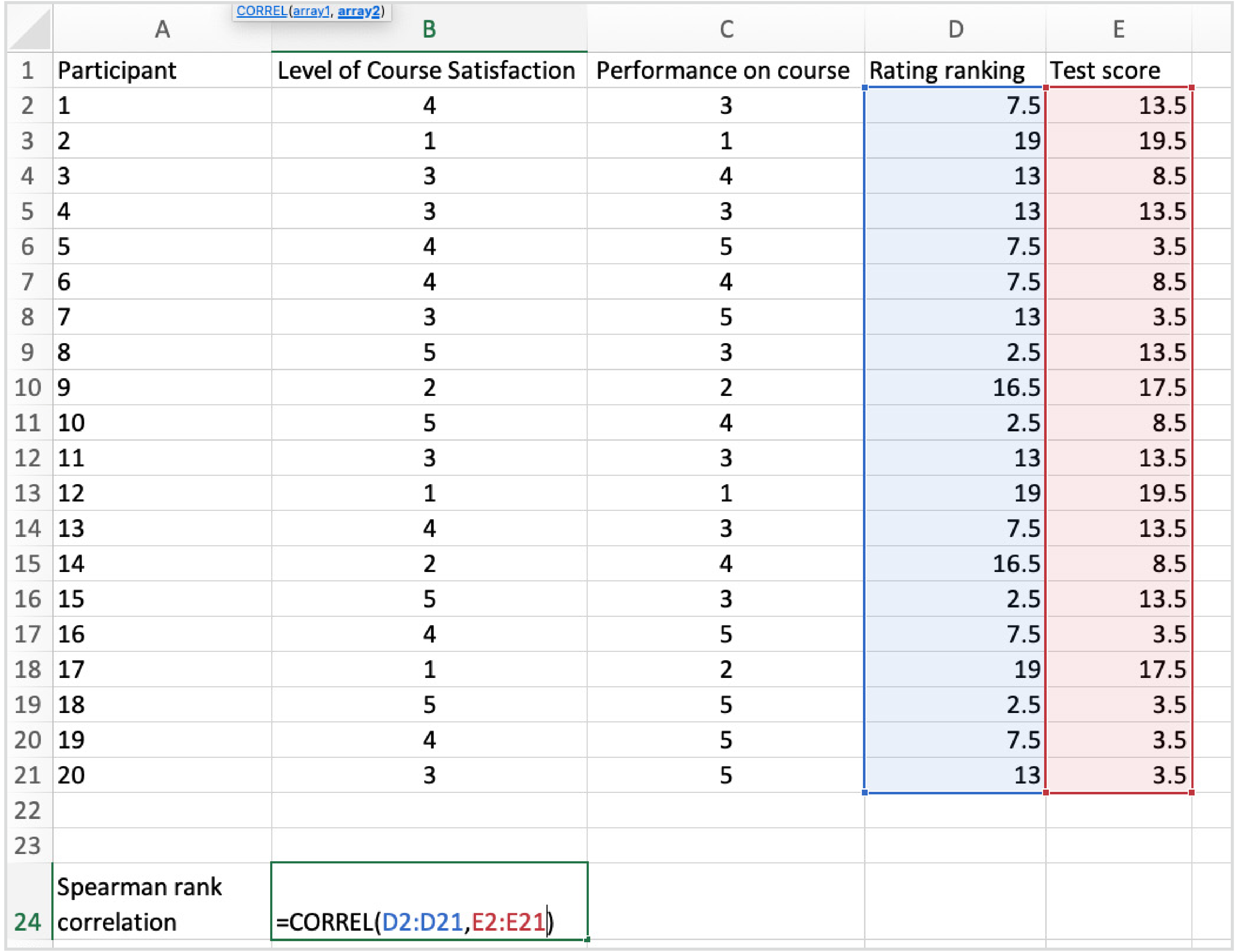
	A	B	C	D	E
1	Participant	Level of Course Satisfaction	Performance on course	Rating ranking	Test score
2	1	4	3	=RANK.AVG(B2, $B$2:$B$21,	
3	2	1	1		
4	3	3	4		
5	4	3	3		
6	5	4	5		
7	6	4	4		
8	7	3	5		
9	8	5	3		
10	9	2	2		
11	10	5	4		
12	11	3	3		
13	12	1	1		
14	13	4	3		
15	14	2	4		
16	15	5	3		
17	16	4	5		
18	17	1	2		
19	18	5	5		
20	19	4	5		
21	20	3	5		
22					

	A	B	C	D	E	F
1	Participant	Level of Course Satisfaction	Performance on course	Rating ranking	Test score	
2	1	4	3	7.5	=RANK.AVG(C2, $C$2:$C$21,0)	
3	2	1	1			
4	3	3	4			
5	4	3	3			
6	5	4	5			
7	6	4	4			
8	7	3	5			
9	8	5	3			
10	9	2	2			
11	10	5	4			
12	11	3	3			
13	12	1	1			
14	13	4	3			
15	14	2	4			
16	15	5	3			
17	16	4	5			
18	17	1	2			
19	18	5	5			
20	19	4	5			
21	20	3	5			
22						

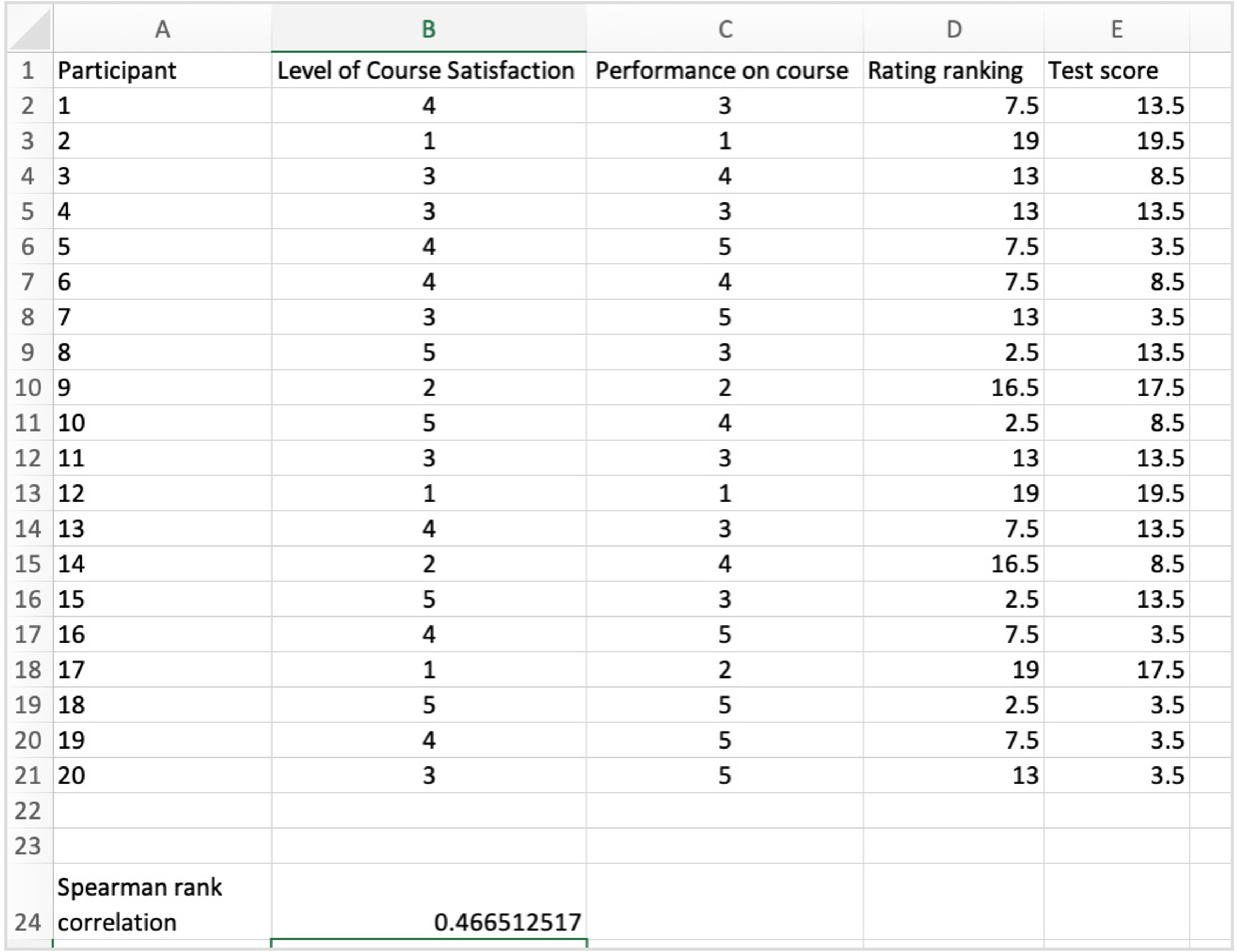
Figure 21. Formulas for ranking

Once the formulas have been entered, the drag-down function can be used to populate the rows below.

The next step is to work out the correlation, the input in this example would be =CORREL(D2:D21,E2:E21) (Figure 22). This provided a coefficient of 0.47 (rounded up to two decimal points).



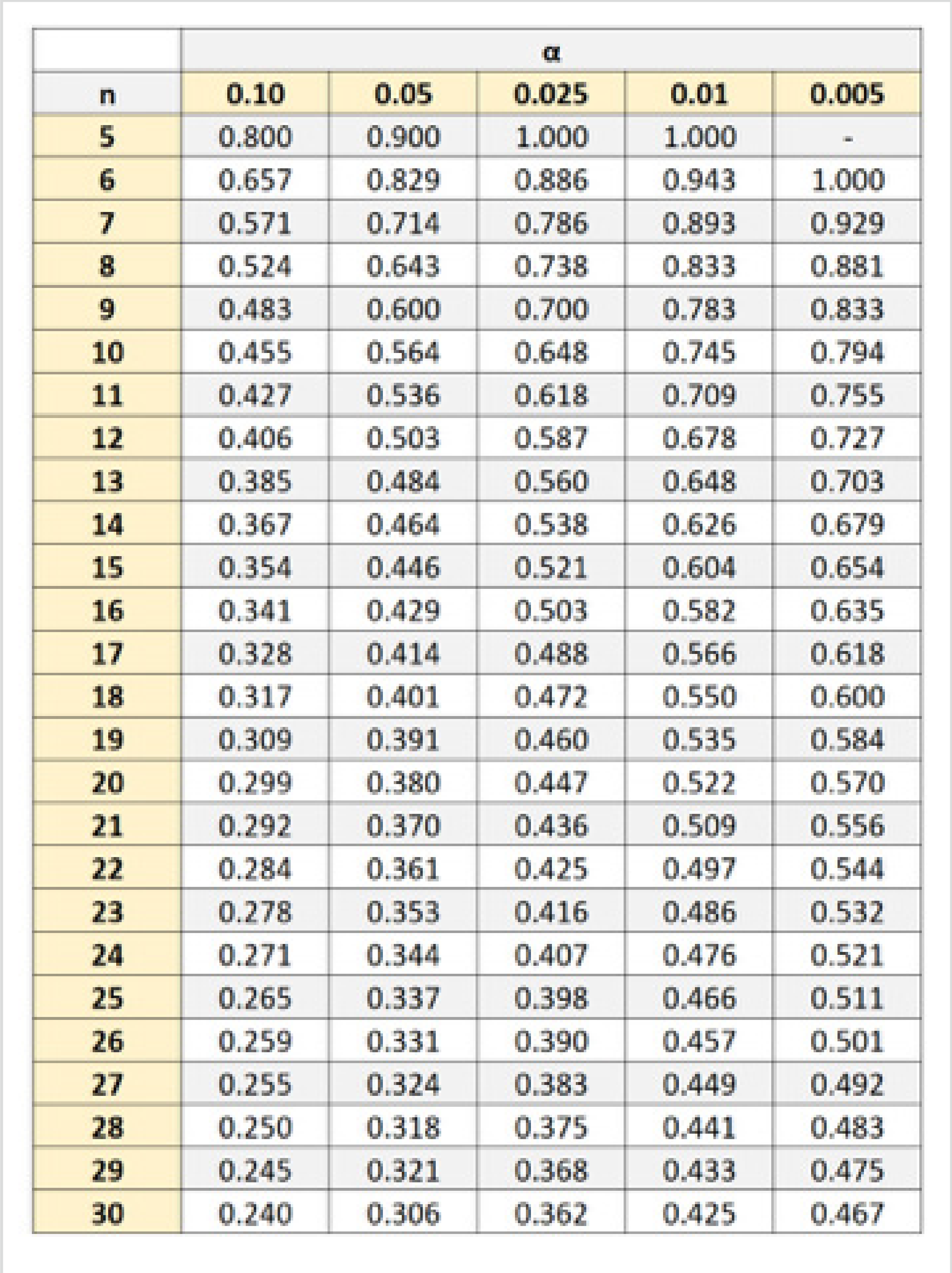
	A	B	C	D	E
1	Participant	Level of Course Satisfaction	Performance on course	Rating ranking	Test score
2	1	4	3	7.5	13.5
3	2	1	1	19	19.5
4	3	3	4	13	8.5
5	4	3	3	13	13.5
6	5	4	5	7.5	3.5
7	6	4	4	7.5	8.5
8	7	3	5	13	3.5
9	8	5	3	2.5	13.5
10	9	2	2	16.5	17.5
11	10	5	4	2.5	8.5
12	11	3	3	13	13.5
13	12	1	1	19	19.5
14	13	4	3	7.5	13.5
15	14	2	4	16.5	8.5
16	15	5	3	2.5	13.5
17	16	4	5	7.5	3.5
18	17	1	2	19	17.5
19	18	5	5	2.5	3.5
20	19	4	5	7.5	3.5
21	20	3	5	13	3.5
22					
23					
24	Spearman rank correlation	=CORREL(D2:D21,E2:E21)			



	A	B	C	D	E
1	Participant	Level of Course Satisfaction	Performance on course	Rating ranking	Test score
2	1	4	3	7.5	13.5
3	2	1	1	19	19.5
4	3	3	4	13	8.5
5	4	3	3	13	13.5
6	5	4	5	7.5	3.5
7	6	4	4	7.5	8.5
8	7	3	5	13	3.5
9	8	5	3	2.5	13.5
10	9	2	2	16.5	17.5
11	10	5	4	2.5	8.5
12	11	3	3	13	13.5
13	12	1	1	19	19.5
14	13	4	3	7.5	13.5
15	14	2	4	16.5	8.5
16	15	5	3	2.5	13.5
17	16	4	5	7.5	3.5
18	17	1	2	19	17.5
19	18	5	5	2.5	3.5
20	19	4	5	7.5	3.5
21	20	3	5	13	3.5
22					
23					
24	Spearman rank correlation	0.466512517			

Figure 22. Spearman rank correlation score

This coefficient figure is then used to identify the p-value using the Spearman Rank correlation table (Figure 23). The chart shows the alpha values across the top with the sample size down the side. The alpha value used for this test would be .05 and the sample size of 20. This means the coefficient would need to be at or above .380 to be regarded as significant. As the coefficient in this test was .47 it is significant.



	α				
n	0.10	0.05	0.025	0.01	0.005
5	0.800	0.900	1.000	1.000	-
6	0.657	0.829	0.886	0.943	1.000
7	0.571	0.714	0.786	0.893	0.929
8	0.524	0.643	0.738	0.833	0.881
9	0.483	0.600	0.700	0.783	0.833
10	0.455	0.564	0.648	0.745	0.794
11	0.427	0.536	0.618	0.709	0.755
12	0.406	0.503	0.587	0.678	0.727
13	0.385	0.484	0.560	0.648	0.703
14	0.367	0.464	0.538	0.626	0.679
15	0.354	0.446	0.521	0.604	0.654
16	0.341	0.429	0.503	0.582	0.635
17	0.328	0.414	0.488	0.566	0.618
18	0.317	0.401	0.472	0.550	0.600
19	0.309	0.391	0.460	0.535	0.584
20	0.299	0.380	0.447	0.522	0.570
21	0.292	0.370	0.436	0.509	0.556
22	0.284	0.361	0.425	0.497	0.544
23	0.278	0.353	0.416	0.486	0.532
24	0.271	0.344	0.407	0.476	0.521
25	0.265	0.337	0.398	0.466	0.511
26	0.259	0.331	0.390	0.457	0.501
27	0.255	0.324	0.383	0.449	0.492
28	0.250	0.318	0.375	0.441	0.483
29	0.245	0.321	0.368	0.433	0.475
30	0.240	0.306	0.362	0.425	0.467

Figure 22. Spearman rank correlation score

The result can then be reported as:

> A Spearman Rank calculation found a positive correlation between the trainer’s assessment of the participant’s performance and the participant’s levels of satisfaction with the training course (r = .47, p <.05).

Mann-Whitney U Test

Why use an Mann-Whitney U Test?

The Mann-Whitney U-test is a non-parametric alternative to an independent sample t-test and can be used where the data is ordinal and/or if the distribution or sample size does not allow a t-test to be conducted.

Using Excel to complete a Mann-Whitney U test

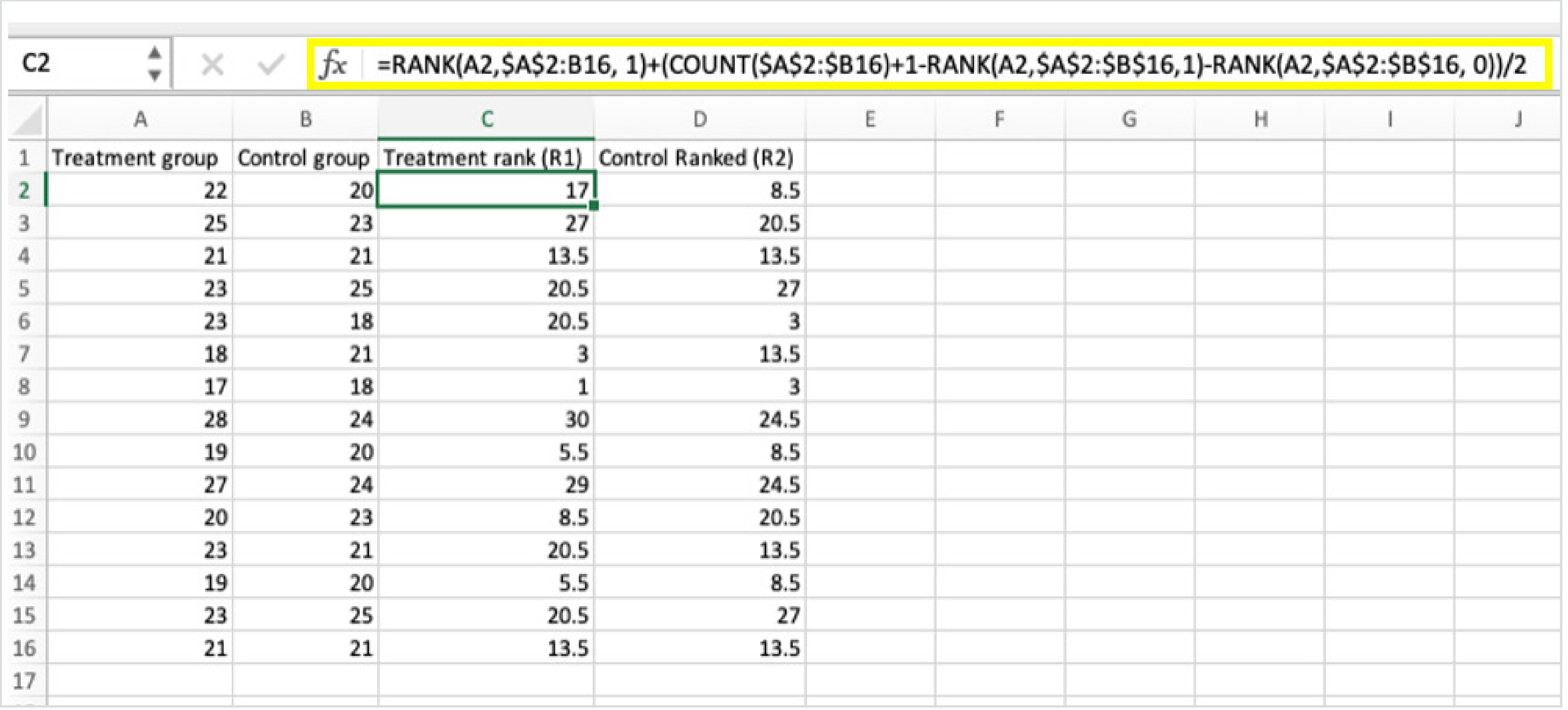
Please note that this is a fictitious example for demonstration purposes.
An evaluation may want to identify if a group have retained the knowledge gained 3 months after receiving an educational intervention on water safety. To assess this, a group that received the intervention, the treatment group (TG) were compared with a group that had not received the intervention, the Control Group (CG). The data being analysed came from the results of a knowledge test administered to both groups, with 20 people completing the test in both the TG and CG. The test contains 10 questions, with one point being allocated for each correct answer. Scores were then totalled to provide an overall score. A review of the distribution showed the data was not normally distributed. Therefore a Mann-Whitney test was used as the data did not appear to be normally distributed.

The test ranks cases, and this is the first step in completing this test using Excel. The following equations should be used to rank the cases based on the example shown in Figure 24 and 25.

For the treatment group it would be =RANK(A2,$A$2:B16, 1)+(COUNT($A$2:$B16)+1-RANK(A2,$A$2:$B$16,1)-RANK(A2,$A$2:$B$16, 0))/2 (Figure 24)

For the Control group is would be =RANK(B2,$A$2:B16, 1)+(COUNT($A$2:$B16)+1-RANK(B2,$A$2:$B$16,1)-RANK(B2,$A$2:$B$16, 0))/2 (Figure 25).

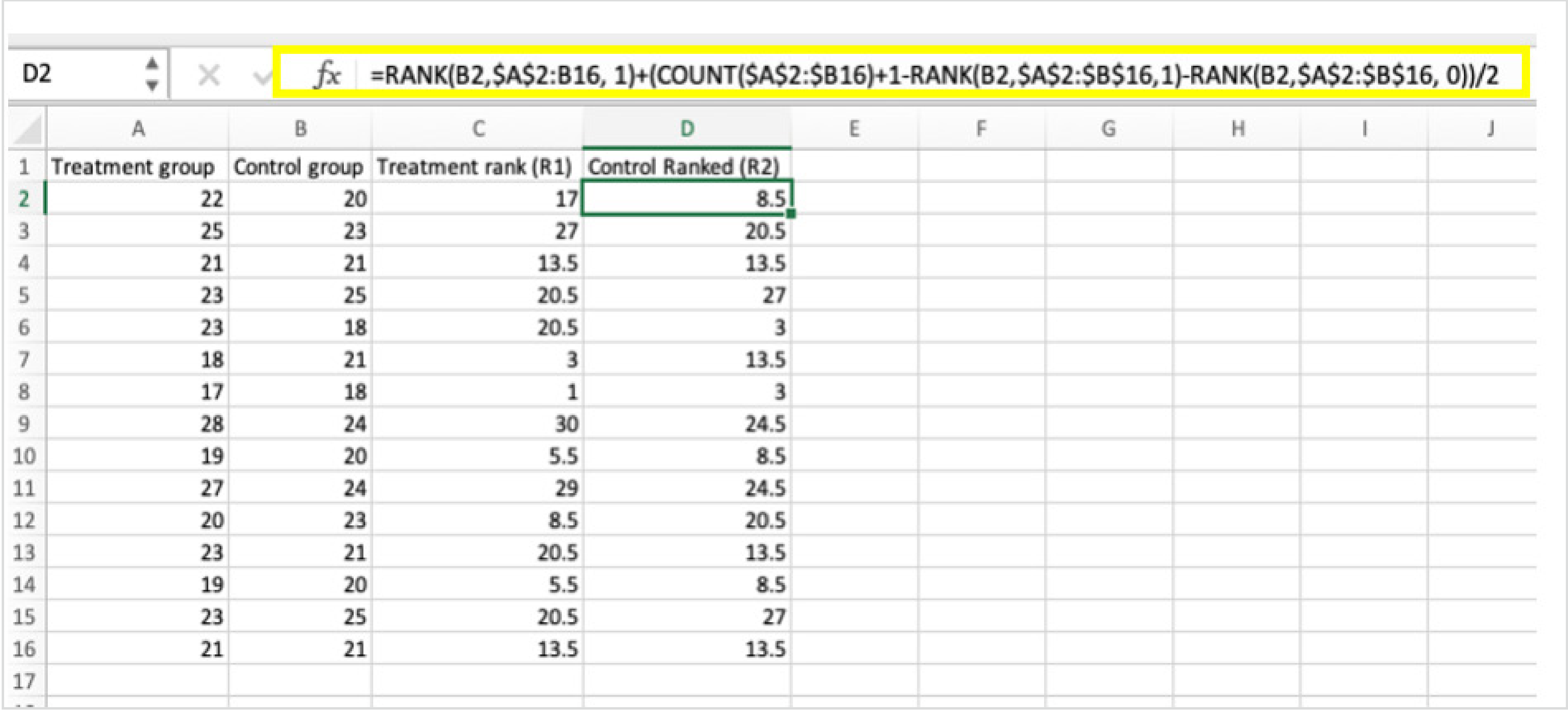
Once these equations have been inputted for the first row of data, these can be dragged down for the remaining rows, meaning these only need inputting once.



C2 =RANK(A2,$A$2:B16, 1)+(COUNT($A$2:$B16)+1-RANK(A2,$A$2:$B$16,1)-RANK(A2,$A$2:$B$16, 0))/2

	A	B	C	D
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)
2	22	20	17	8.5
3	25	23	27	20.5
4	21	21	13.5	13.5
5	23	25	20.5	27
6	23	18	20.5	3
7	18	21	3	13.5
8	17	18	1	3
9	28	24	30	24.5
10	19	20	5.5	8.5
11	27	24	29	24.5
12	20	23	8.5	20.5
13	23	21	20.5	13.5
14	19	20	5.5	8.5
15	23	25	20.5	27
16	21	21	13.5	13.5

Figure 24. Ranking for the treatment group

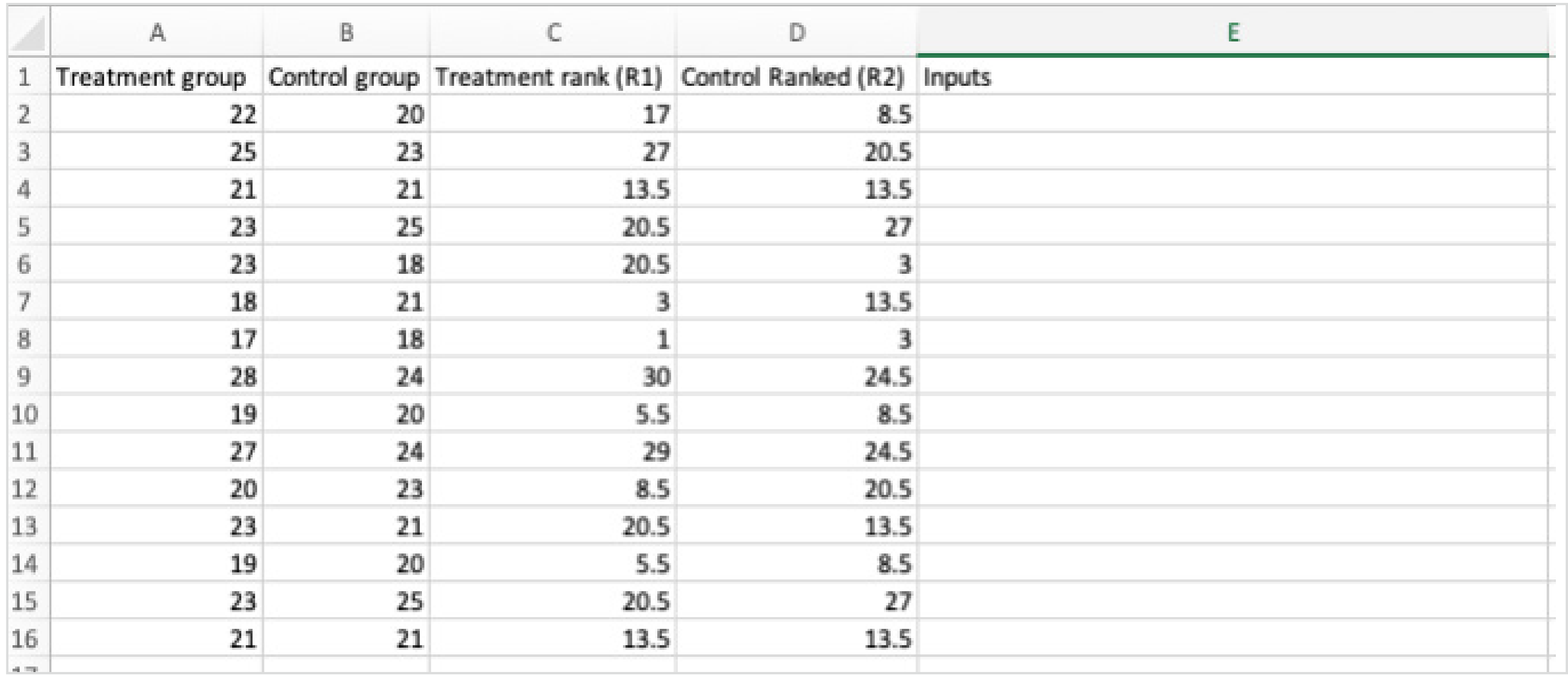


D2 =RANK(B2,$A$2:B16, 1)+(COUNT($A$2:$B16)+1-RANK(B2,$A$2:$B$16,1)-RANK(B2,$A$2:$B$16, 0))/2

	A	B	C	D
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)
2	22	20	17	8.5
3	25	23	27	20.5
4	21	21	13.5	13.5
5	23	25	20.5	27
6	23	18	20.5	3
7	18	21	3	13.5
8	17	18	1	3
9	28	24	30	24.5
10	19	20	5.5	8.5
11	27	24	29	24.5
12	20	23	8.5	20.5
13	23	21	20.5	13.5
14	19	20	5.5	8.5
15	23	25	20.5	27
16	21	21	13.5	13.5

Figure 25. Ranking for the control group

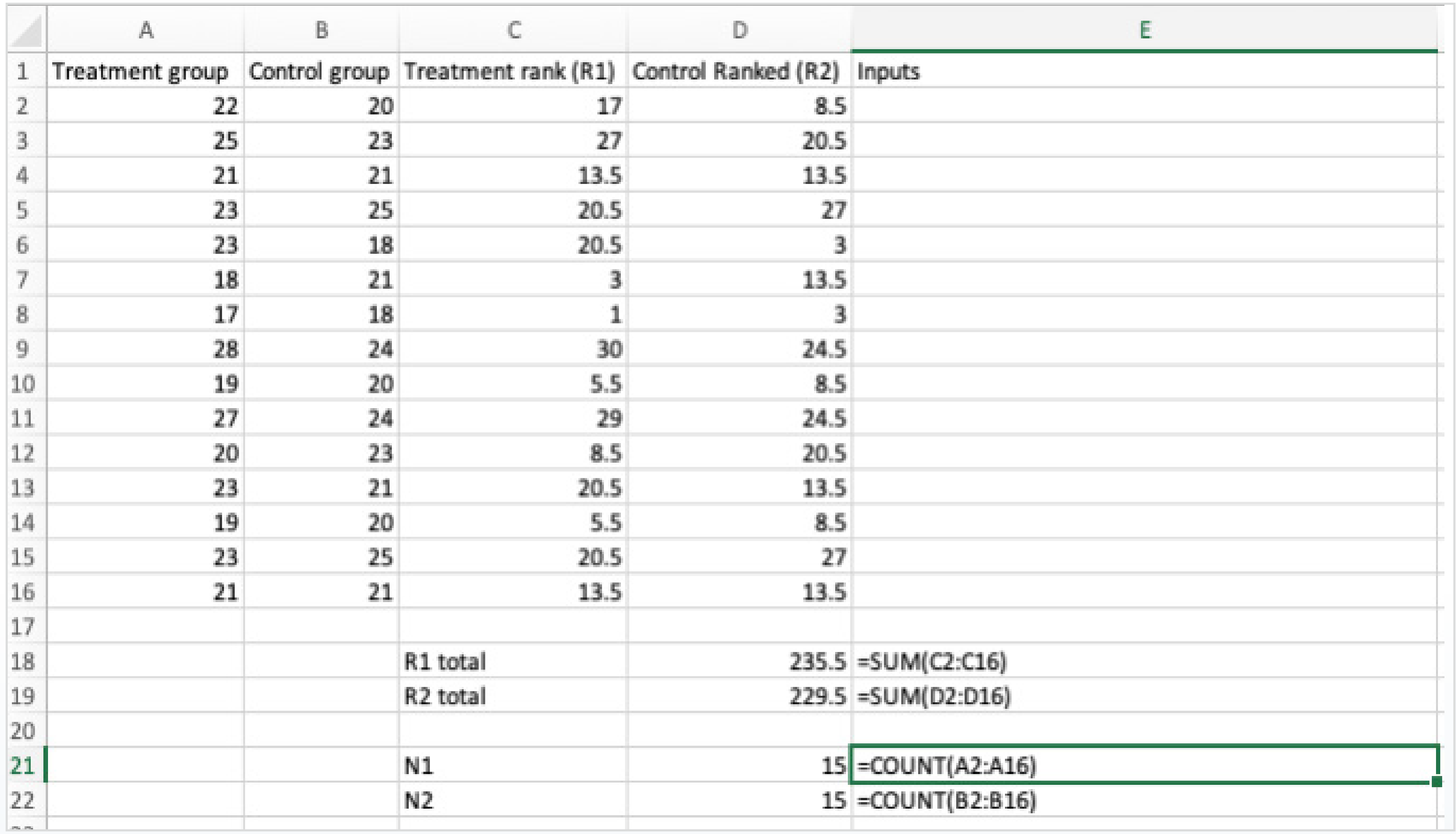
The next stage is to Rank 1 (R1) total and Rank (R2) Total. This can be done using the following formula, for the R1 total =SUM(C2:C16), for the Rank 2 total =SUM(D2:D16) (Figure 26).



	A	B	C	D	E
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)	Inputs
2	22	20	17	8.5	
3	25	23	27	20.5	
4	21	21	13.5	13.5	
5	23	25	20.5	27	
6	23	18	20.5	3	
7	18	21	3	13.5	
8	17	18	1	3	
9	28	24	30	24.5	
10	19	20	5.5	8.5	
11	27	24	29	24.5	
12	20	23	8.5	20.5	
13	23	21	20.5	13.5	
14	19	20	5.5	8.5	
15	23	25	20.5	27	
16	21	21	13.5	13.5	

Figure 26. Total ranking for R1 and R2

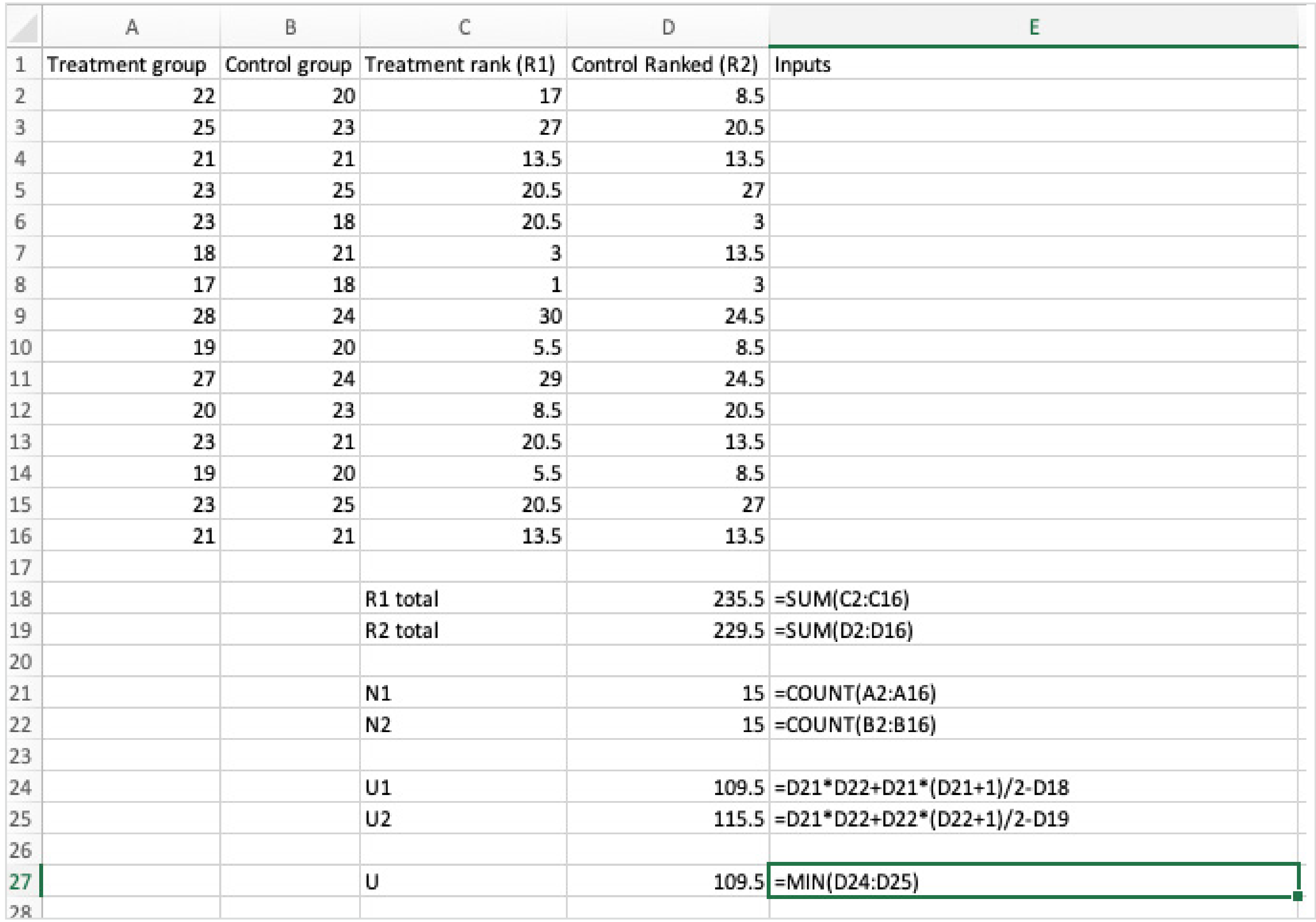
It is then necessary to calculate the number in the sample for each of the groups. In this case, the count is 15 (Figure 27).



	A	B	C	D	E
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)	Inputs
2	22	20	17	8.5	
3	25	23	27	20.5	
4	21	21	13.5	13.5	
5	23	25	20.5	27	
6	23	18	20.5	3	
7	18	21	3	13.5	
8	17	18	1	3	
9	28	24	30	24.5	
10	19	20	5.5	8.5	
11	27	24	29	24.5	
12	20	23	8.5	20.5	
13	23	21	20.5	13.5	
14	19	20	5.5	8.5	
15	23	25	20.5	27	
16	21	21	13.5	13.5	
17					
18			R1 total	235.5	=SUM(C2:C16)
19			R2 total	229.5	=SUM(D2:D16)
20					
21			N1	15	=COUNT(A2:A16)
22			N2	15	=COUNT(B2:B16)

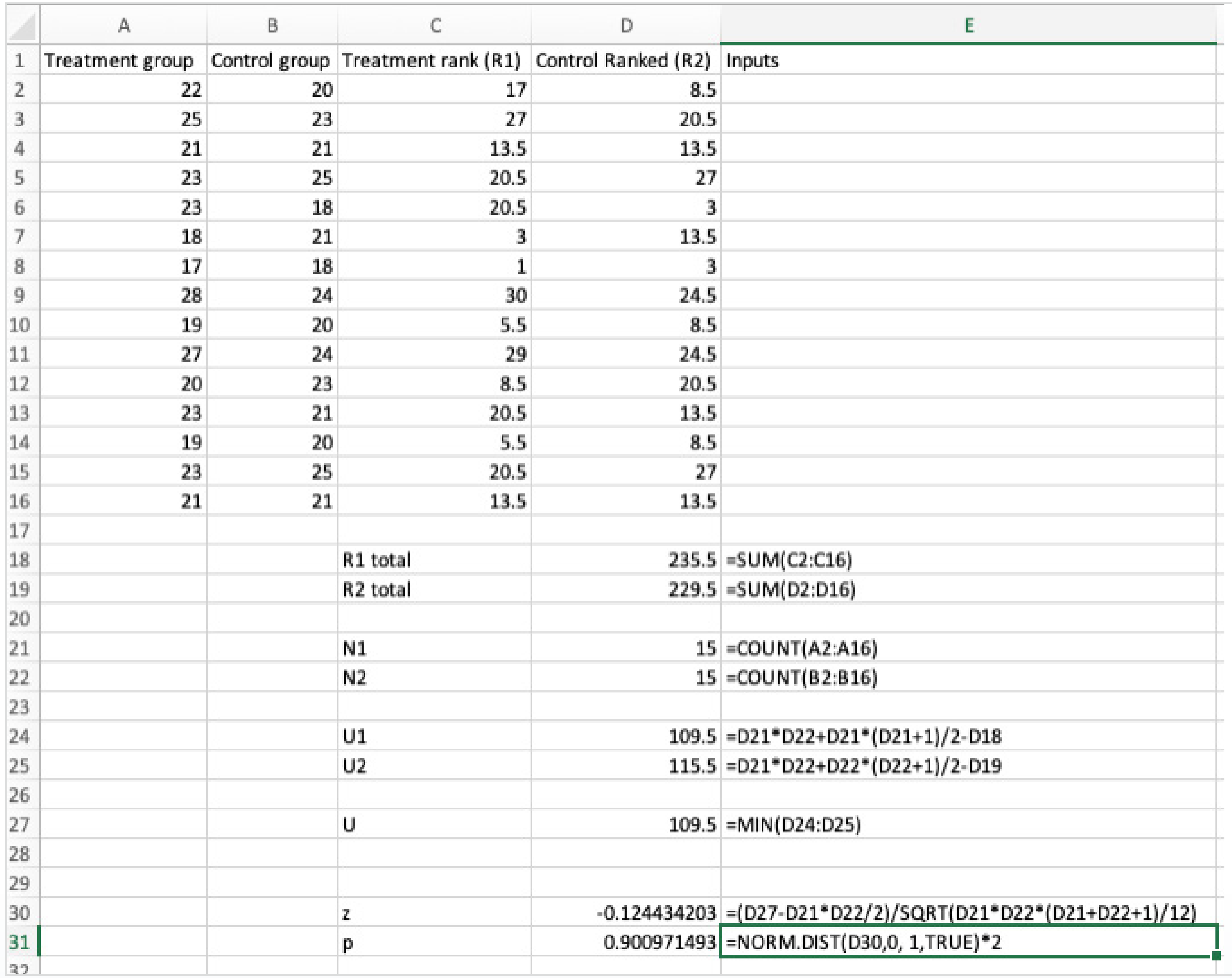
Figure 27. Count of sample (N1 and N2)

It is now possible to work out the test statistics, referred to as the U-value. This must be calculated for both groups initially, these scores are then used to calculate the overall value of U. In this example the formula for U1 (Treatment group) would be D21*D22+D21*(D21+1)/2-D18. For U2 (Control group) it would be =D21*D22+D22*(D22+1)/2-D19. To finalise the U the input would be =MIN(D24:D25). (Figure 28)



	A	B	C	D	E
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)	Inputs
2	22	20	17	8.5	
3	25	23	27	20.5	
4	21	21	13.5	13.5	
5	23	25	20.5	27	
6	23	18	20.5	3	
7	18	21	3	13.5	
8	17	18	1	3	
9	28	24	30	24.5	
10	19	20	5.5	8.5	
11	27	24	29	24.5	
12	20	23	8.5	20.5	
13	23	21	20.5	13.5	
14	19	20	5.5	8.5	
15	23	25	20.5	27	
16	21	21	13.5	13.5	
17					
18			R1 total	235.5	=SUM(C2:C16)
19			R2 total	229.5	=SUM(D2:D16)
20					
21			N1	15	=COUNT(A2:A16)
22			N2	15	=COUNT(B2:B16)
23					
24			U1	109.5	=D21*D22+D21*(D21+1)/2-D18
25			U2	115.5	=D21*D22+D22*(D22+1)/2-D19
26					
27			U	109.5	=MIN(D24:D25)

Figure 28. Calculating the U values



	A	B	C	D	E
1	Treatment group	Control group	Treatment rank (R1)	Control Ranked (R2)	Inputs
2	22	20	17	8.5	
3	25	23	27	20.5	
4	21	21	13.5	13.5	
5	23	25	20.5	27	
6	23	18	20.5	3	
7	18	21	3	13.5	
8	17	18	1	3	
9	28	24	30	24.5	
10	19	20	5.5	8.5	
11	27	24	29	24.5	
12	20	23	8.5	20.5	
13	23	21	20.5	13.5	
14	19	20	5.5	8.5	
15	23	25	20.5	27	
16	21	21	13.5	13.5	
17					
18			R1 total	235.5	=SUM(C2:C16)
19			R2 total	229.5	=SUM(D2:D16)
20					
21			N1	15	=COUNT(A2:A16)
22			N2	15	=COUNT(B2:B16)
23					
24			U1	109.5	=D21*D22+D21*(D21+1)/2-D18
25			U2	115.5	=D21*D22+D22*(D22+1)/2-D19
26					
27			U	109.5	=MIN(D24:D25)
28					
29					
30			z	-0.124434203	=(D27-D21*D22/2)/SQRT(D21*D22*(D21+D22+1)/12)
31			p	0.900971493	=NORM.DIST(D30,0, 1,TRUE)*2

Figure 29. Calculating the Z and p values

It is now possible to determine the Z-score and the p-value. The input for Z-score would be =(D27-D21*D22/2)/SQRT(D21*D22*(D21+D22+1)/12). The input to calculate the p-value would be =NORM.DIST(D30,0, 1,TRUE)*2 (Figure 29).

As the p-value is below .05, it can be concluded that any change in the scores was not significant.

The results of the test can now be reported:

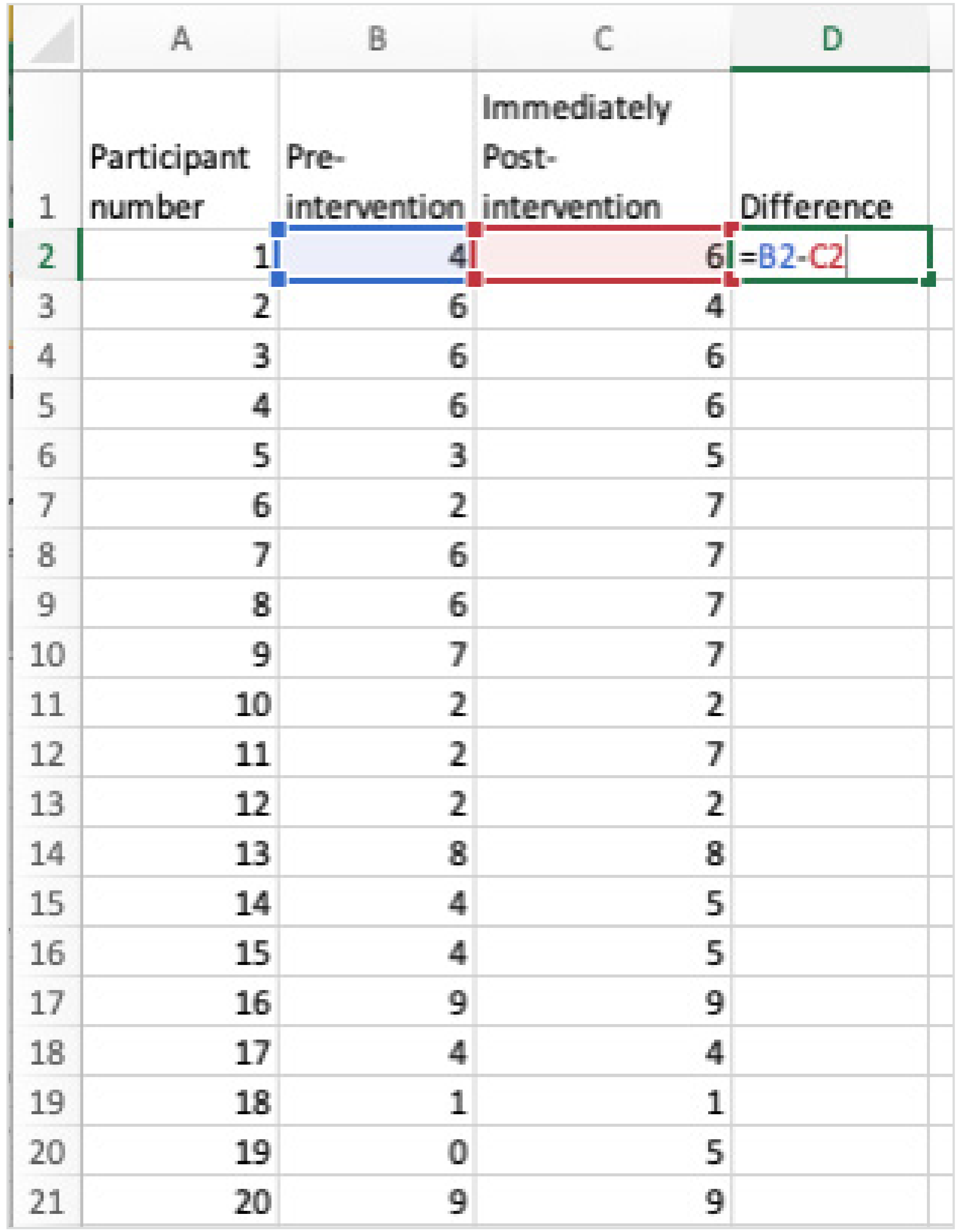
> A Mann-Whitney U test found no significant difference between groups (z = -0.12, p = .90)

The Wilcoxon Signed-Rank Test

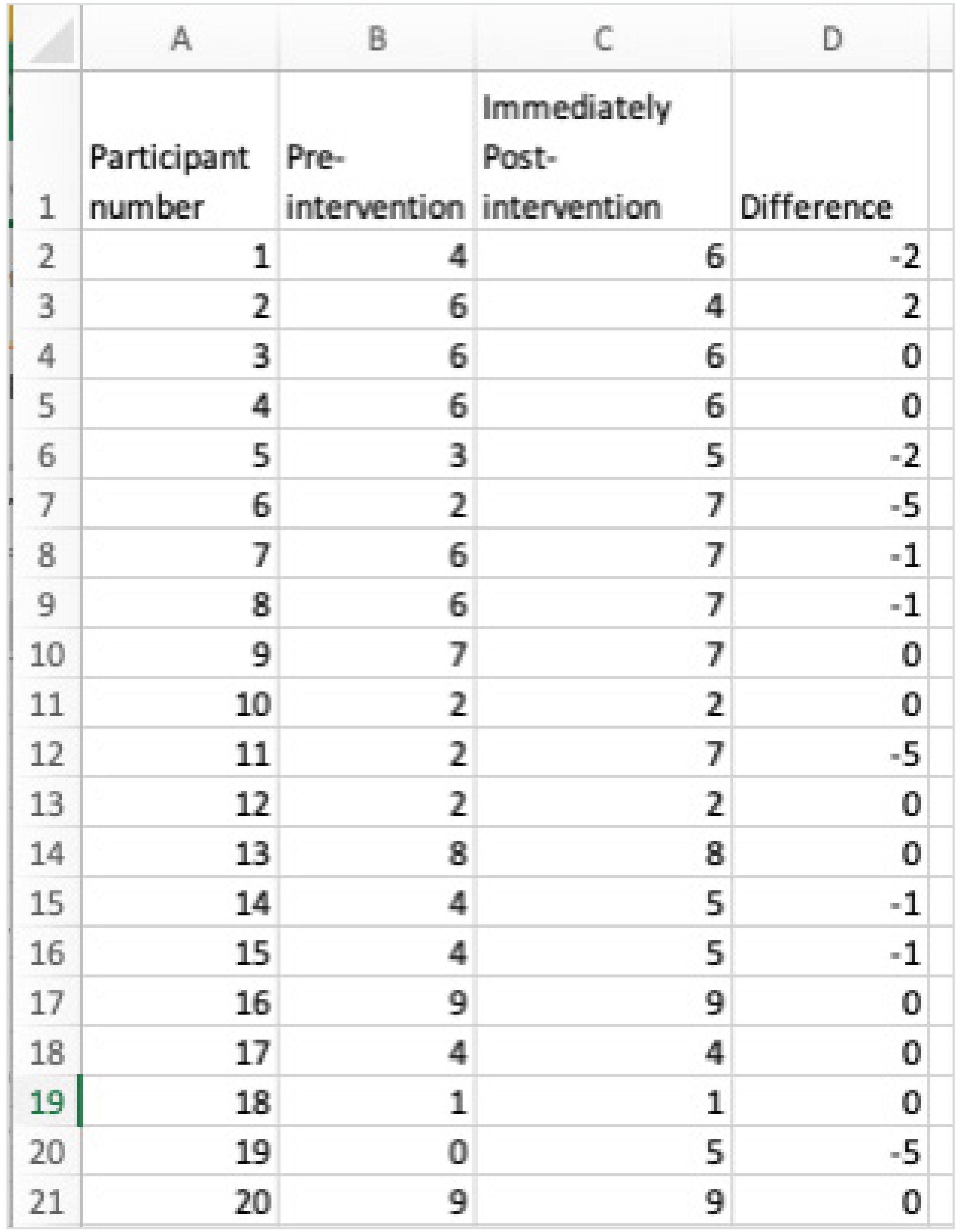
Why Use a Wilcox Signed-Rank Test?

This test should be used when the sample is related and can be matched. It should be used as an alternative to a paired sample t-test when the distribution is not normally distributed or if the sample size is small. In this example, the evaluator is looking to identify if a new FRS breathing apparatus course improved a Firefighter’s knowledge. To do this, they sent out a knowledge test to the participants before the start of the course and then again immediately upon completion the pre-course score and pos-course scores can be matched. Please note that this is a fictitious example for demonstration purposes.

Using Excel to complete a Wilcox Signed-Test
The first step in completing the tests is to calculate the difference between the two sets of scores (Figure 30).



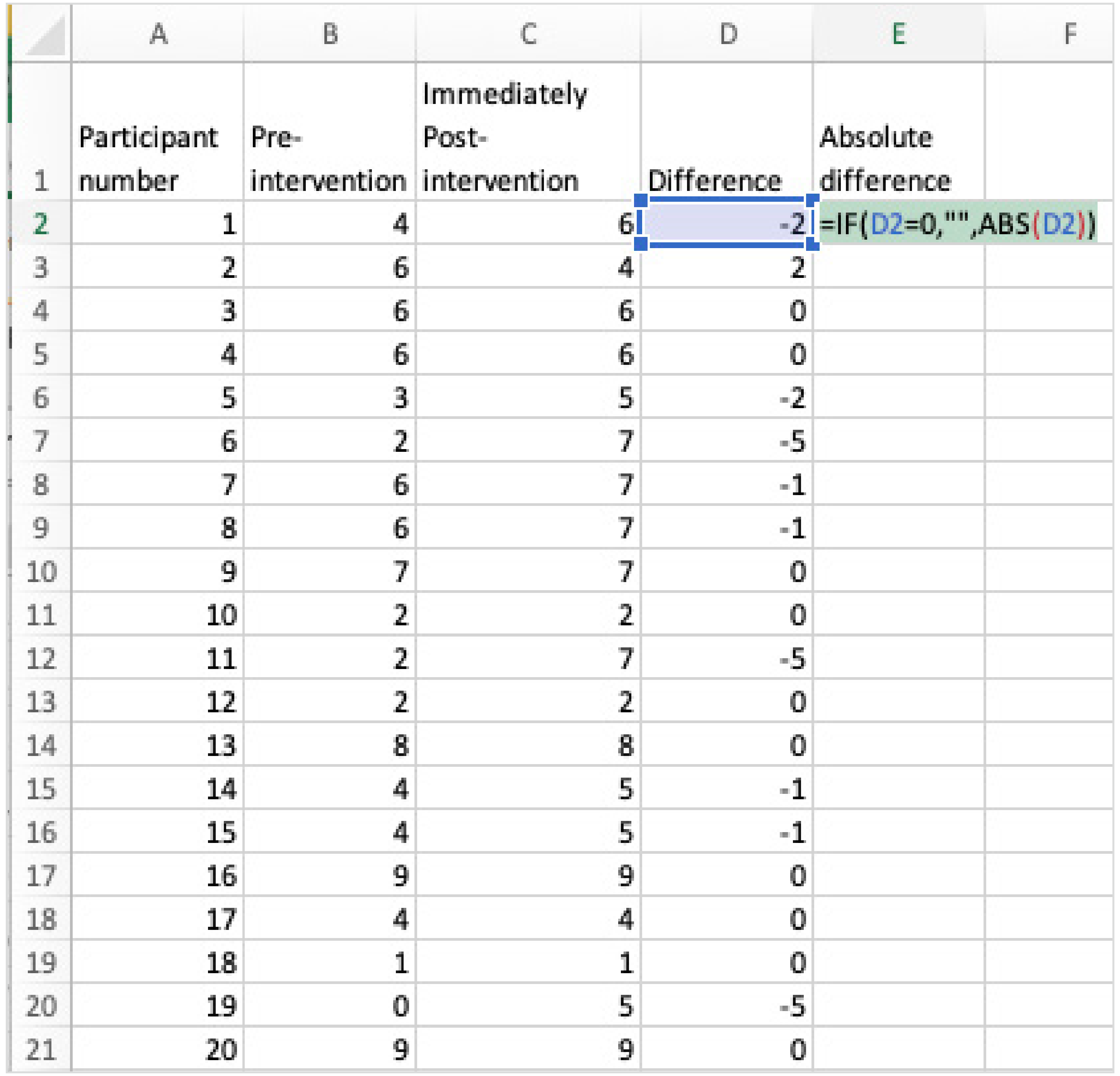
	A	B	C	D
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference
2	1	4	6	=B2-C2
3	2	6	4	
4	3	6	6	
5	4	6	6	
6	5	3	5	
7	6	2	7	
8	7	6	7	
9	8	6	7	
10	9	7	7	
11	10	2	2	
12	11	2	7	
13	12	2	2	
14	13	8	8	
15	14	4	5	
16	15	4	5	
17	16	9	9	
18	17	4	4	
19	18	1	1	
20	19	0	5	
21	20	9	9	



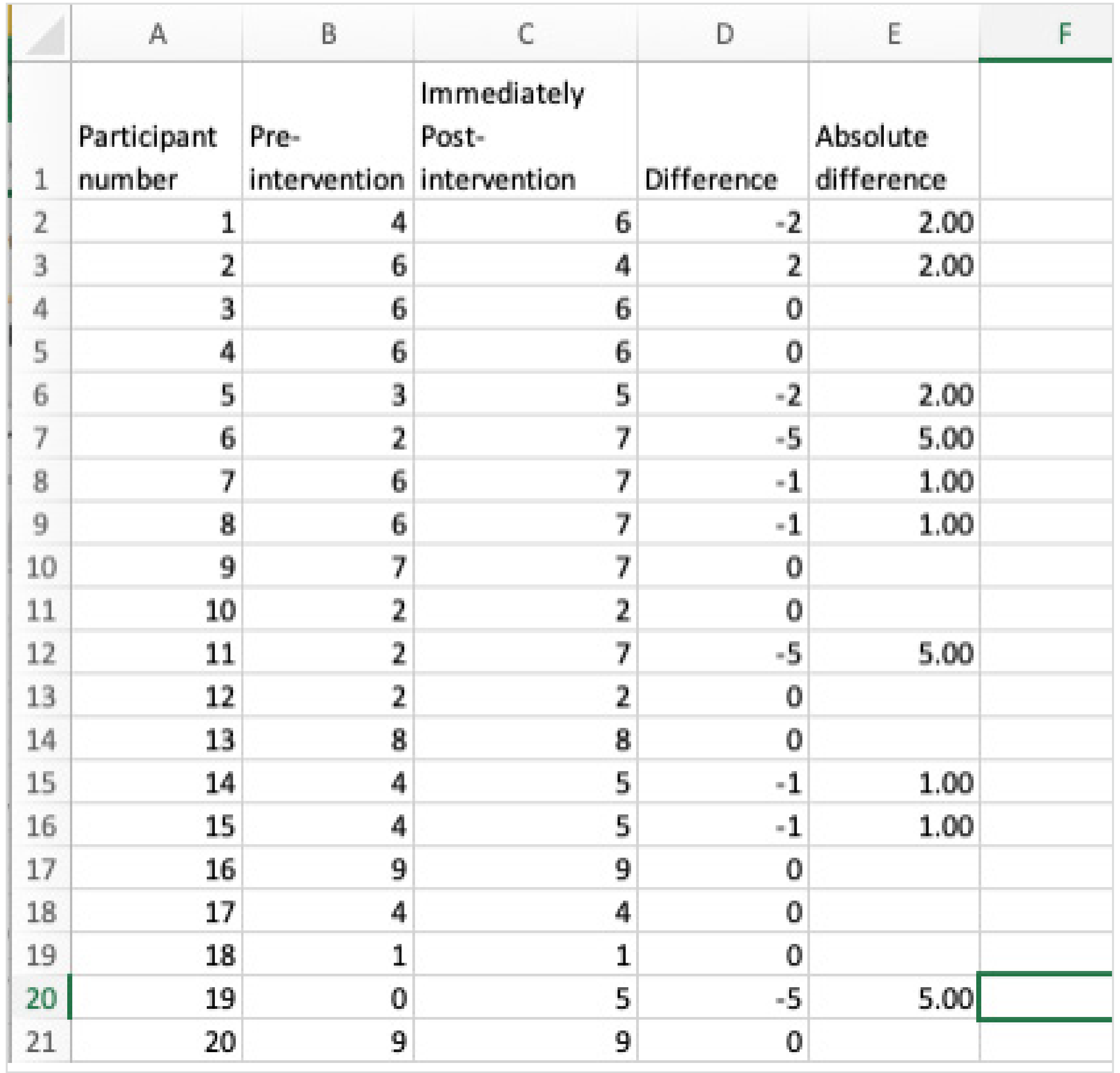
	A	B	C	D
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference
2	1	4	6	-2
3	2	6	4	2
4	3	6	6	0
5	4	6	6	0
6	5	3	5	-2
7	6	2	7	-5
8	7	6	7	-1
9	8	6	7	-1
10	9	7	7	0
11	10	2	2	0
12	11	2	7	-5
13	12	2	2	0
14	13	8	8	0
15	14	4	5	-1
16	15	4	5	-1
17	16	9	9	0
18	17	4	4	0
19	18	1	1	0
20	19	0	5	-5
21	20	9	9	0

Figure 30. The formula for working out the difference between scores

The difference scores need to be altered to the absolute difference; this alters any negative numbers to positive numbers. The formula for this is =IF(D2=0,"",ABS(D2)) (Figure 31).



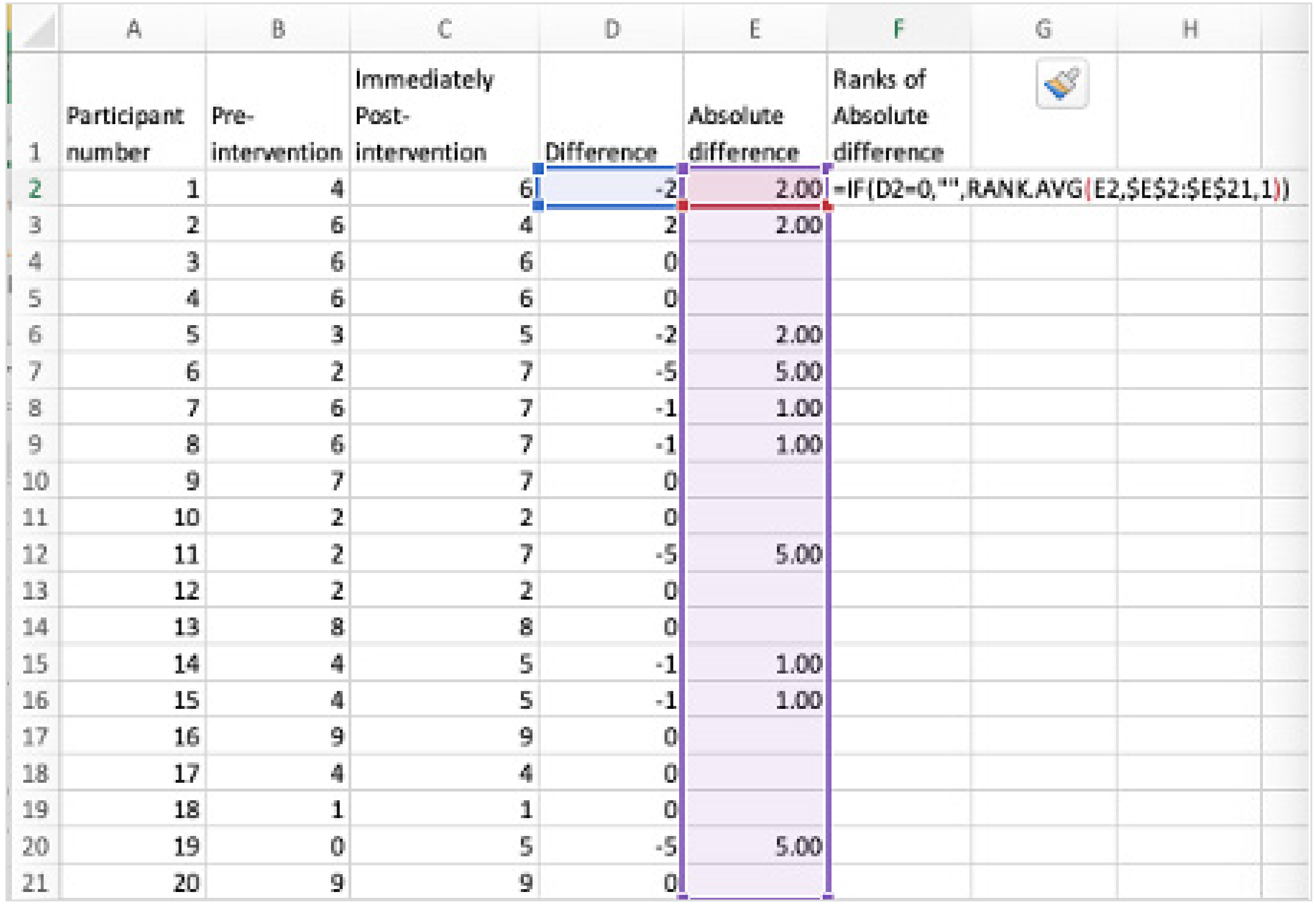
	A	B	C	D	E	F
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	
2	1	4	6	-2	=IF(D2=0,"",ABS(D2))	
3	2	6	4	2		
4	3	6	6	0		
5	4	6	6	0		
6	5	3	5	-2		
7	6	2	7	-5		
8	7	6	7	-1		
9	8	6	7	-1		
10	9	7	7	0		
11	10	2	2	0		
12	11	2	7	-5		
13	12	2	2	0		
14	13	8	8	0		
15	14	4	5	-1		
16	15	4	5	-1		
17	16	9	9	0		
18	17	4	4	0		
19	18	1	1	0		
20	19	0	5	-5		
21	20	9	9	0		



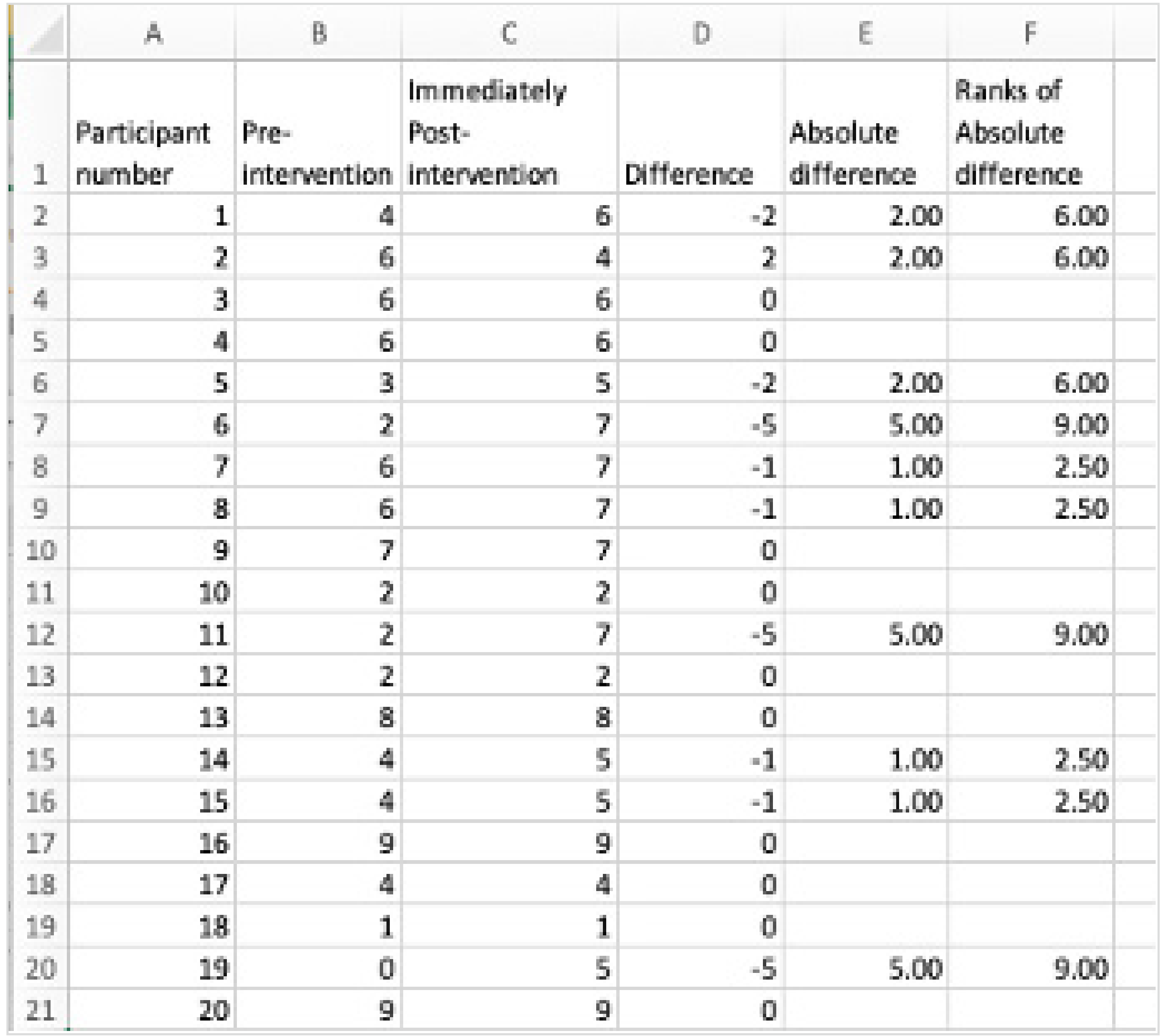
	A	B	C	D	E	F
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	
2	1	4	6	-2	2.00	
3	2	6	4	2	2.00	
4	3	6	6	0		
5	4	6	6	0		
6	5	3	5	-2	2.00	
7	6	2	7	-5	5.00	
8	7	6	7	-1	1.00	
9	8	6	7	-1	1.00	
10	9	7	7	0		
11	10	2	2	0		
12	11	2	7	-5	5.00	
13	12	2	2	0		
14	13	8	8	0		
15	14	4	5	-1	1.00	
16	15	4	5	-1	1.00	
17	16	9	9	0		
18	17	4	4	0		
19	18	1	1	0		
20	19	0	5	-5	5.00	
21	20	9	9	0		

Figure 31. Formula for absolute difference

The scores then need to be ranked. The formula for this is =IF(D2=0,””,RANK.AVG(E2,$E$2:$E$21,1)) (Figure 32)



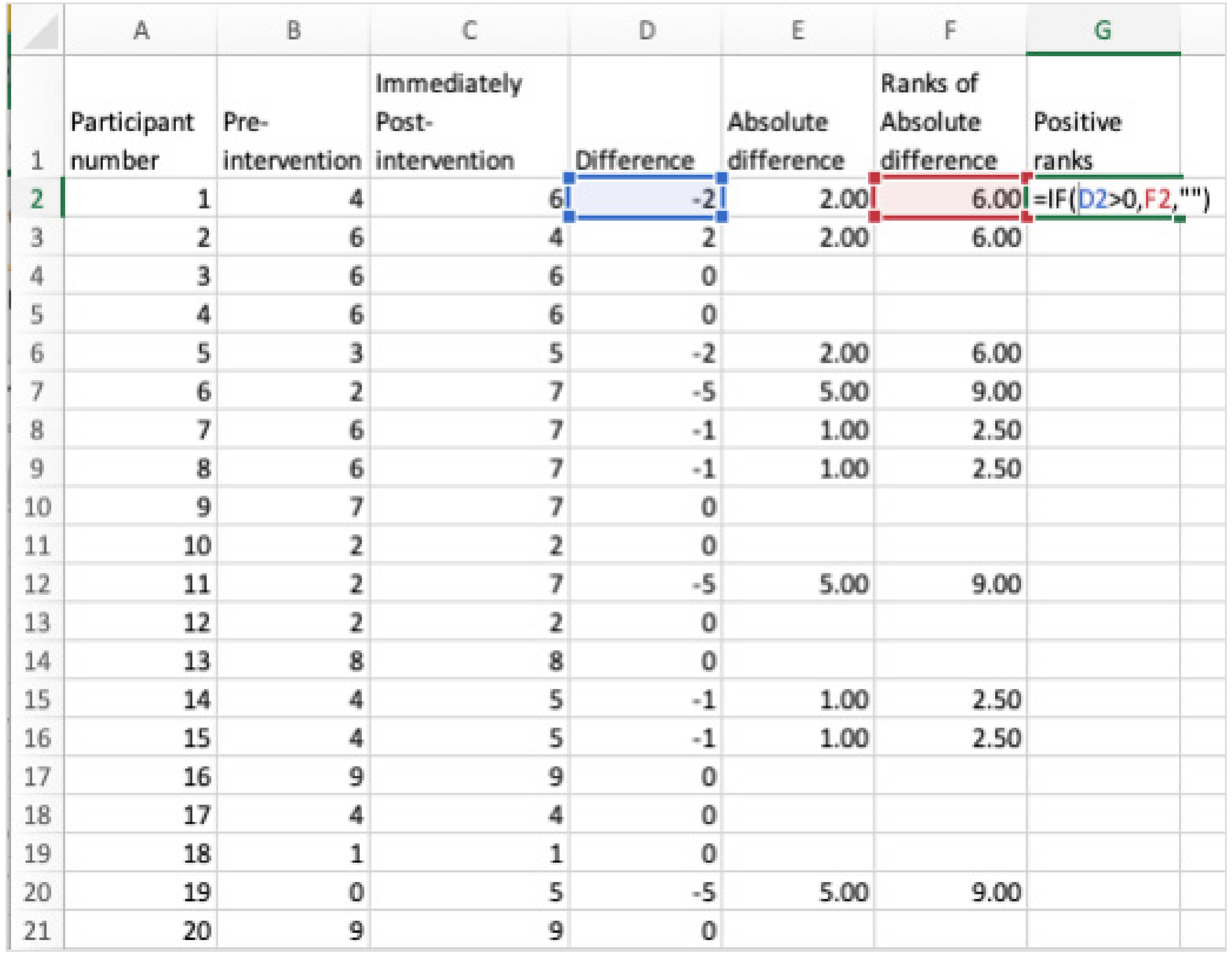
	A	B	C	D	E	F	G	H
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference		
2	1	4	6	-2	2.00	=IF(D2=0,"",RANK.AVG(E2,$E$2:$E$21,1))		
3	2	6	4	2	2.00			
4	3	6	6	0				
5	4	6	6	0				
6	5	3	5	-2	2.00			
7	6	2	7	-5	5.00			
8	7	6	7	-1	1.00			
9	8	6	7	-1	1.00			
10	9	7	7	0				
11	10	2	2	0				
12	11	2	7	-5	5.00			
13	12	2	2	0				
14	13	8	8	0				
15	14	4	5	-1	1.00			
16	15	4	5	-1	1.00			
17	16	9	9	0				
18	17	4	4	0				
19	18	1	1	0				
20	19	0	5	-5	5.00			
21	20	9	9	0				



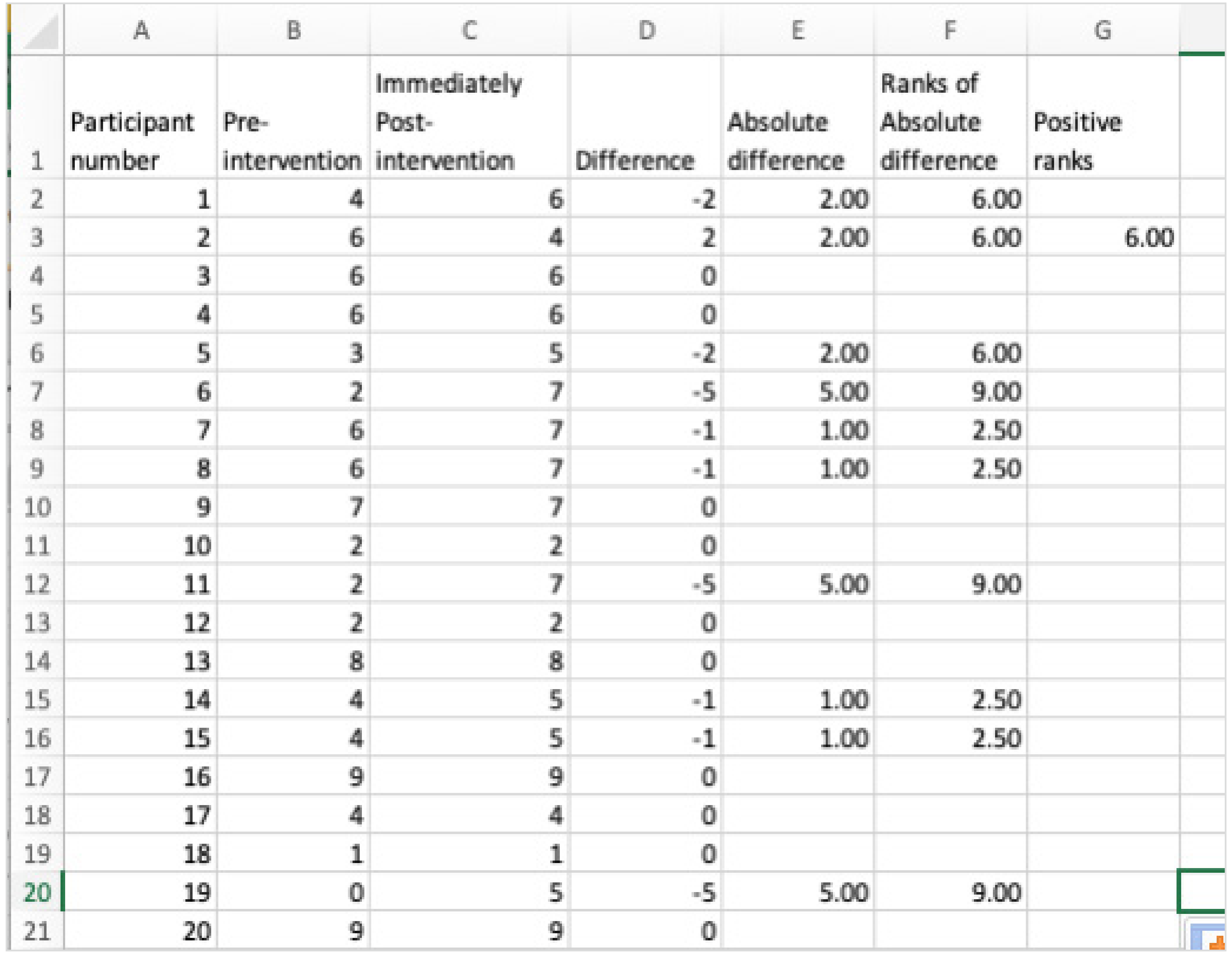
	A	B	C	D	E	F
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference
2	1	4	6	-2	2.00	6.00
3	2	6	4	2	2.00	6.00
4	3	6	6	0		
5	4	6	6	0		
6	5	3	5	-2	2.00	6.00
7	6	2	7	-5	5.00	9.00
8	7	6	7	-1	1.00	2.50
9	8	6	7	-1	1.00	2.50
10	9	7	7	0		
11	10	2	2	0		
12	11	2	7	-5	5.00	9.00
13	12	2	2	0		
14	13	8	8	0		
15	14	4	5	-1	1.00	2.50
16	15	4	5	-1	1.00	2.50
17	16	9	9	0		
18	17	4	4	0		
19	18	1	1	0		
20	19	0	5	-5	5.00	9.00
21	20	9	9	0		

Figure 32. The formula for ranking of absolute difference

Next, the scores need to be ranked positively and negatively. The input for positive ranking is =IF(D2>0,F2,””) (Figure 33) and the input for negative ranking is =IF(D2<0,F2,””) (Figure 34).

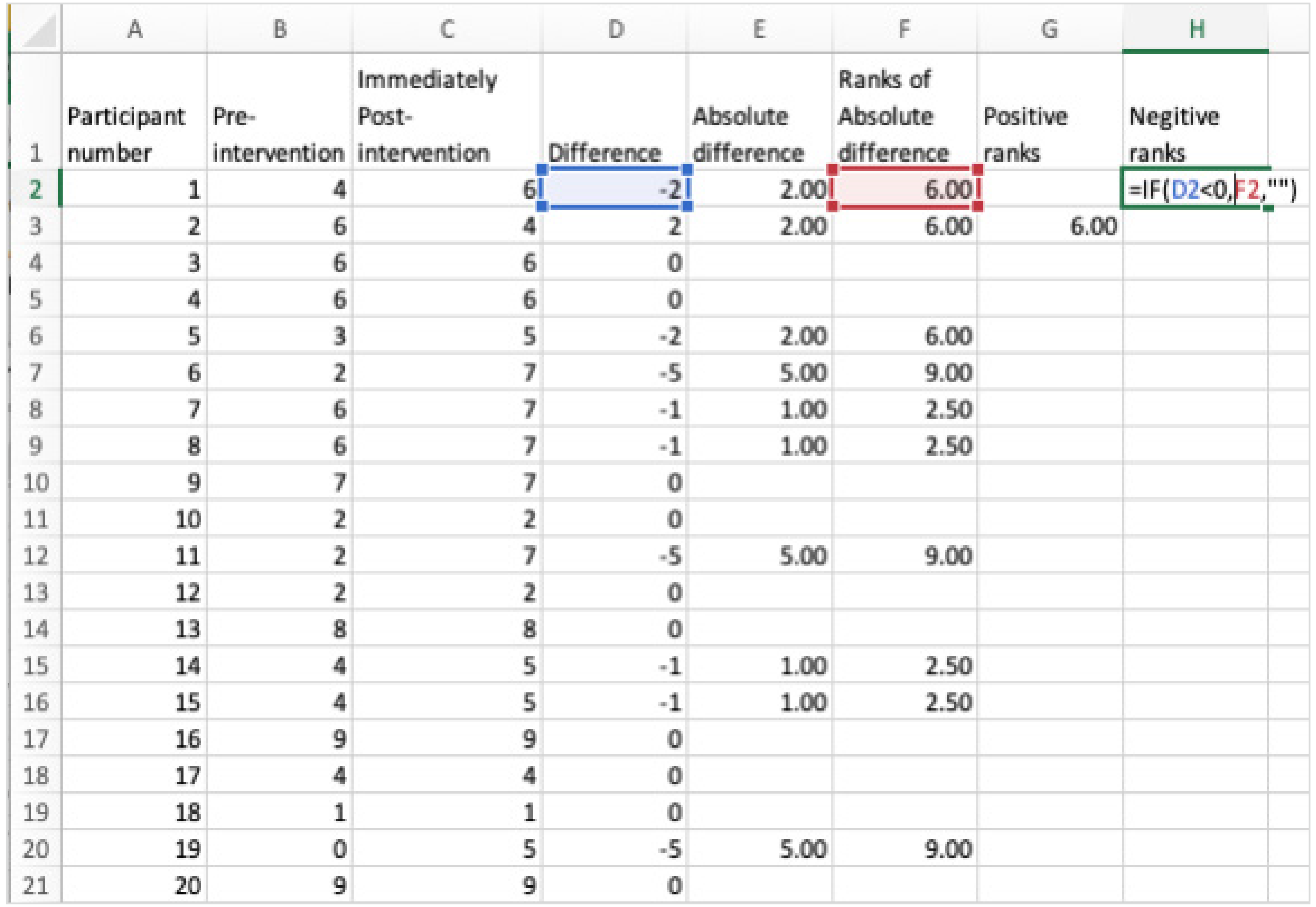


	A	B	C	D	E	F	G
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks
2	1	4	6	-2	2.00	6.00	=IF(D2>0,F2,"")
3	2	6	4	2	2.00	6.00	
4	3	6	6	0			
5	4	6	6	0			
6	5	3	5	-2	2.00	6.00	
7	6	2	7	-5	5.00	9.00	
8	7	6	7	-1	1.00	2.50	
9	8	6	7	-1	1.00	2.50	
10	9	7	7	0			
11	10	2	2	0			
12	11	2	7	-5	5.00	9.00	
13	12	2	2	0			
14	13	8	8	0			
15	14	4	5	-1	1.00	2.50	
16	15	4	5	-1	1.00	2.50	
17	16	9	9	0			
18	17	4	4	0			
19	18	1	1	0			
20	19	0	5	-5	5.00	9.00	
21	20	9	9	0			

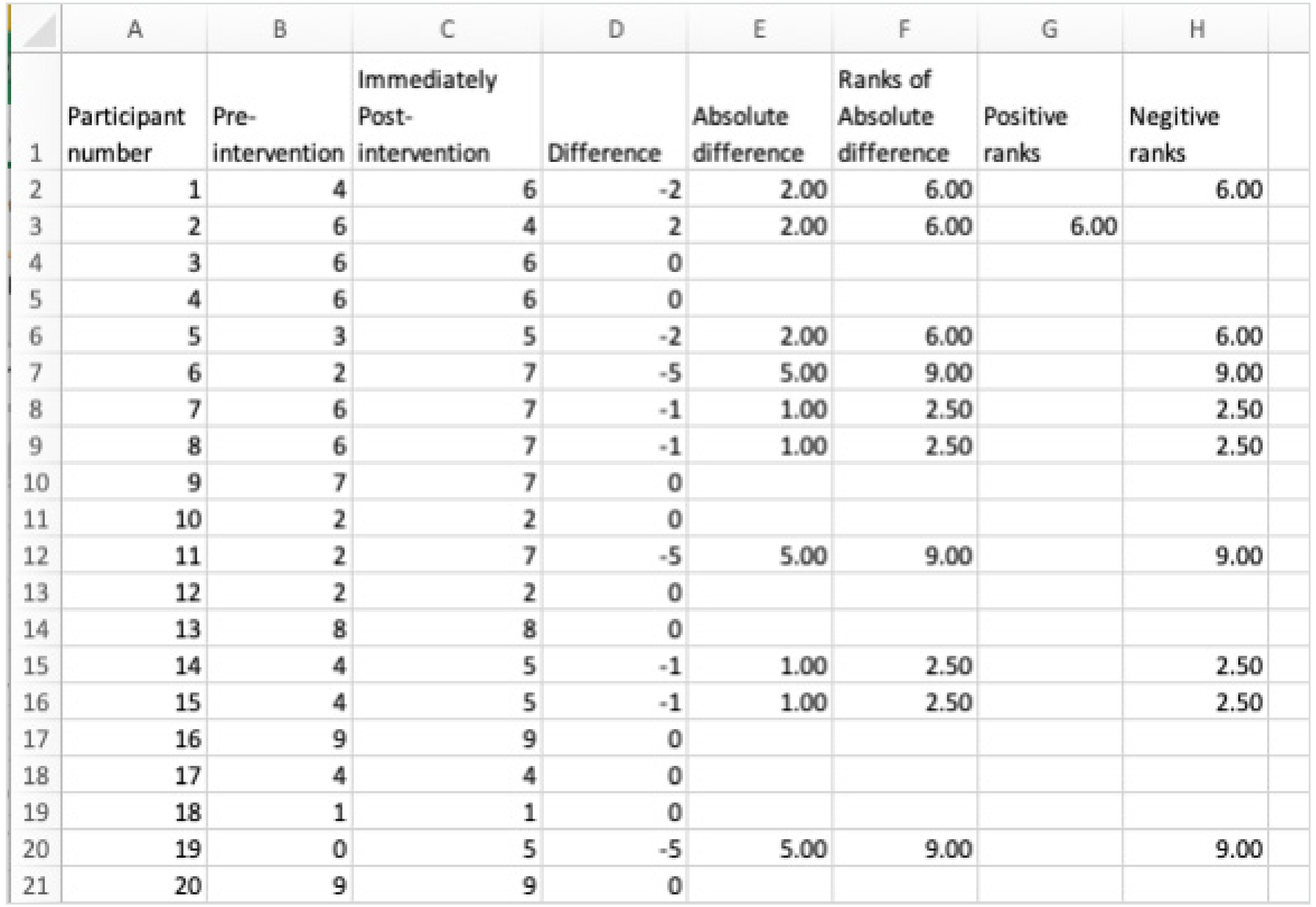


	A	B	C	D	E	F	G
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks
2	1	4	6	-2	2.00	6.00	
3	2	6	4	2	2.00	6.00	6.00
4	3	6	6	0			
5	4	6	6	0			
6	5	3	5	-2	2.00	6.00	
7	6	2	7	-5	5.00	9.00	
8	7	6	7	-1	1.00	2.50	
9	8	6	7	-1	1.00	2.50	
10	9	7	7	0			
11	10	2	2	0			
12	11	2	7	-5	5.00	9.00	
13	12	2	2	0			
14	13	8	8	0			
15	14	4	5	-1	1.00	2.50	
16	15	4	5	-1	1.00	2.50	
17	16	9	9	0			
18	17	4	4	0			
19	18	1	1	0			
20	19	0	5	-5	5.00	9.00	
21	20	9	9	0			

Figure 33. The formula for positive ranking



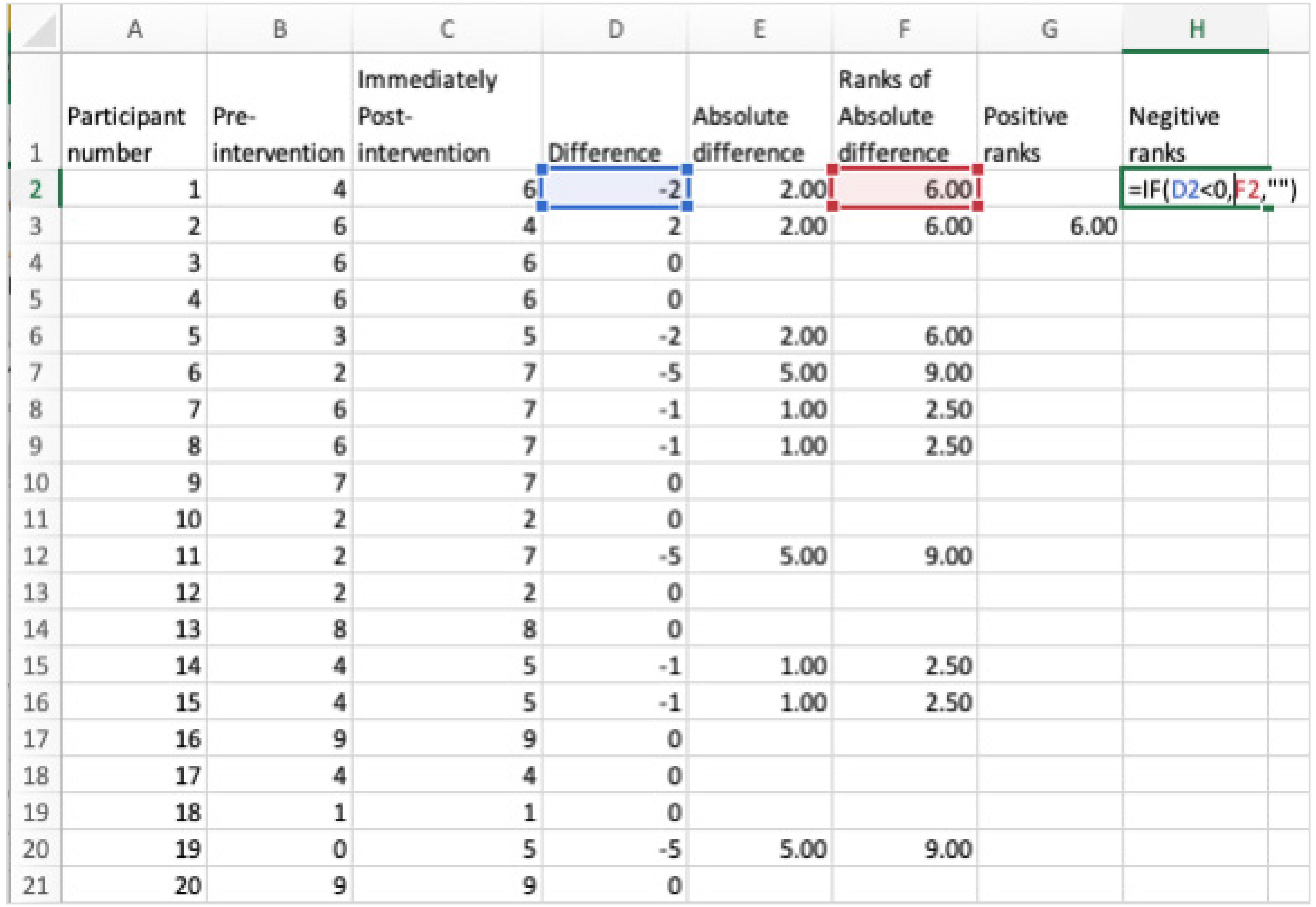
	A	B	C	D	E	F	G	H
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks
2	1	4	6	-2	2.00	6.00		=IF(D2<0,F2,"")
3	2	6	4	2	2.00	6.00	6.00	
4	3	6	6	0				
5	4	6	6	0				
6	5	3	5	-2	2.00	6.00		
7	6	2	7	-5	5.00	9.00		
8	7	6	7	-1	1.00	2.50		
9	8	6	7	-1	1.00	2.50		
10	9	7	7	0				
11	10	2	2	0				
12	11	2	7	-5	5.00	9.00		
13	12	2	2	0				
14	13	8	8	0				
15	14	4	5	-1	1.00	2.50		
16	15	4	5	-1	1.00	2.50		
17	16	9	9	0				
18	17	4	4	0				
19	18	1	1	0				
20	19	0	5	-5	5.00	9.00		
21	20	9	9	0				



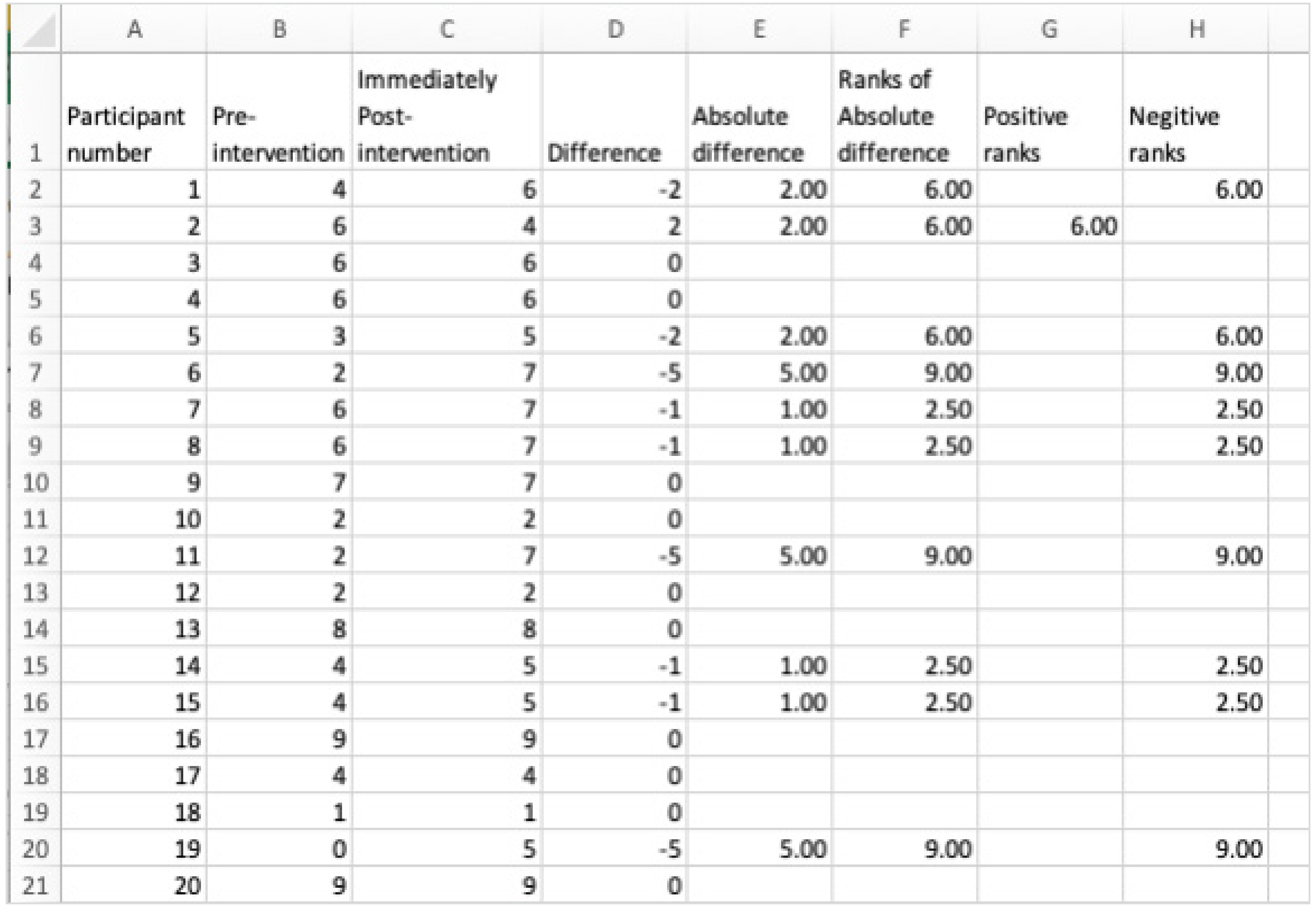
	A	B	C	D	E	F	G	H
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks
2	1	4	6	-2	2.00	6.00		6.00
3	2	6	4	2	2.00	6.00	6.00	
4	3	6	6	0				
5	4	6	6	0				
6	5	3	5	-2	2.00	6.00		6.00
7	6	2	7	-5	5.00	9.00		9.00
8	7	6	7	-1	1.00	2.50		2.50
9	8	6	7	-1	1.00	2.50		2.50
10	9	7	7	0				
11	10	2	2	0				
12	11	2	7	-5	5.00	9.00		9.00
13	12	2	2	0				
14	13	8	8	0				
15	14	4	5	-1	1.00	2.50		2.50
16	15	4	5	-1	1.00	2.50		2.50
17	16	9	9	0				
18	17	4	4	0				
19	18	1	1	0				
20	19	0	5	-5	5.00	9.00		9.00
21	20	9	9	0				

Figure 34. The formula for negative ranking

The final stage is to work out the value of the test statistic/Z statistic and the sample size (Figures 35 and 36). The input for the test statistic is =MIN(SUM(G2:G21),SUM(H2:H21)) and the input for the sample size is =COUNT(G2:H20). These calculations give the following values: Z statistic = 6, with the sample size being 10.

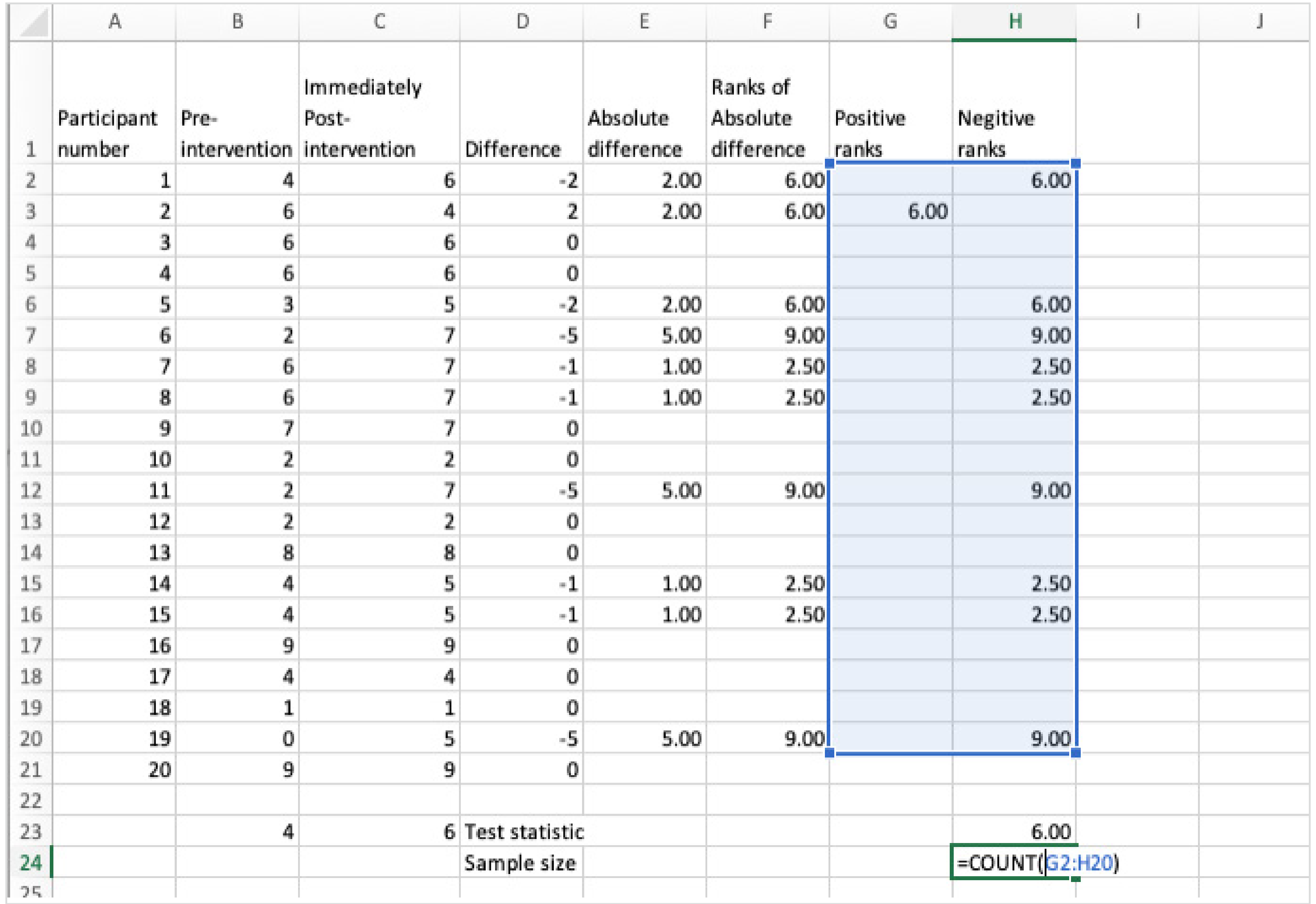


	A	B	C	D	E	F	G	H
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks
2	1	4	6	-2	2.00	6.00		=IF(D2<0,F2,"")
3	2	6	4	2	2.00	6.00	6.00	
4	3	6	6	0				
5	4	6	6	0				
6	5	3	5	-2	2.00	6.00		
7	6	2	7	-5	5.00	9.00		
8	7	6	7	-1	1.00	2.50		
9	8	6	7	-1	1.00	2.50		
10	9	7	7	0				
11	10	2	2	0				
12	11	2	7	-5	5.00	9.00		
13	12	2	2	0				
14	13	8	8	0				
15	14	4	5	-1	1.00	2.50		
16	15	4	5	-1	1.00	2.50		
17	16	9	9	0				
18	17	4	4	0				
19	18	1	1	0				
20	19	0	5	-5	5.00	9.00		
21	20	9	9	0				

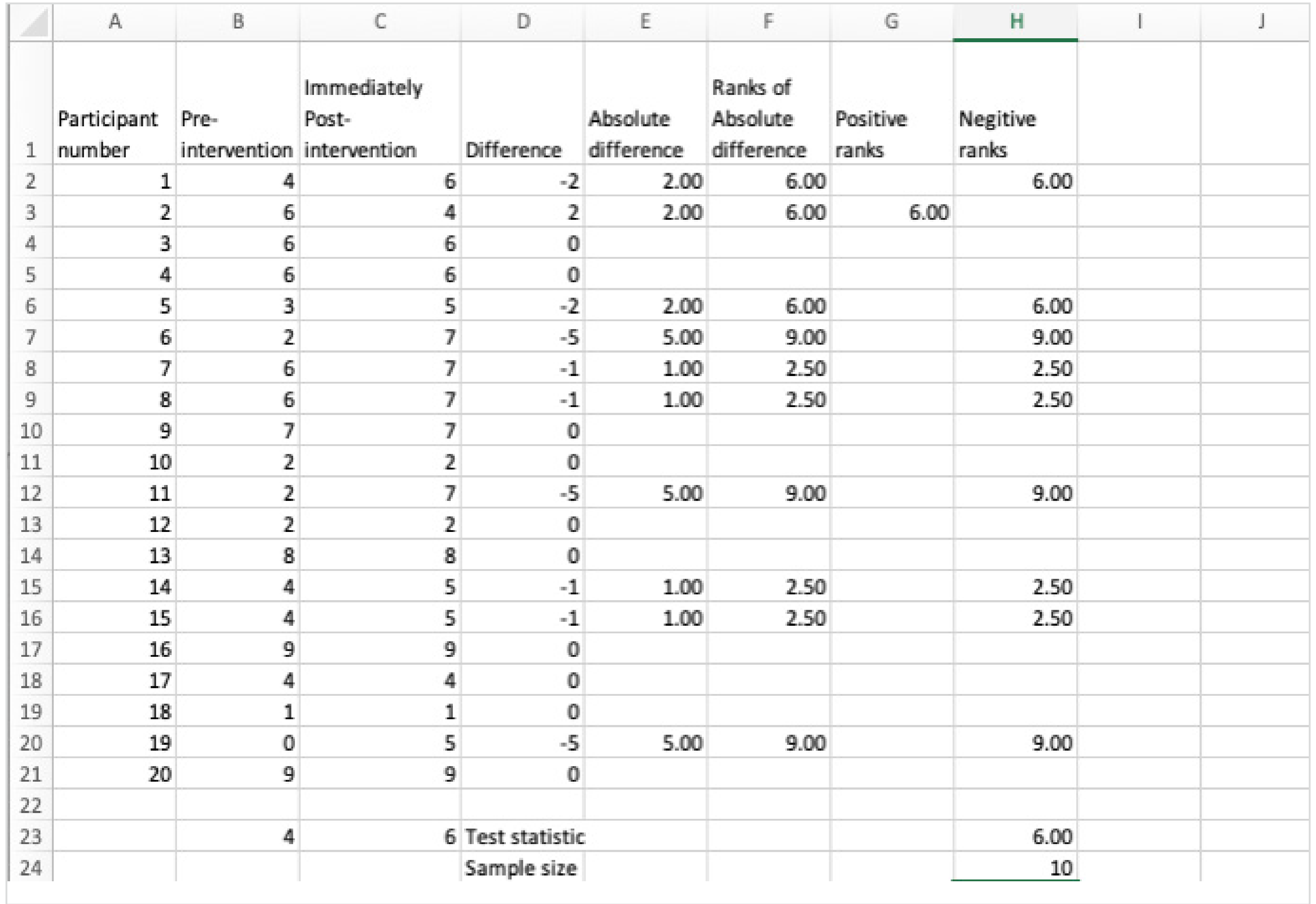


	A	B	C	D	E	F	G	H
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks
2	1	4	6	-2	2.00	6.00		6.00
3	2	6	4	2	2.00	6.00	6.00	
4	3	6	6	0				
5	4	6	6	0				
6	5	3	5	-2	2.00	6.00		6.00
7	6	2	7	-5	5.00	9.00		9.00
8	7	6	7	-1	1.00	2.50		2.50
9	8	6	7	-1	1.00	2.50		2.50
10	9	7	7	0				
11	10	2	2	0				
12	11	2	7	-5	5.00	9.00		9.00
13	12	2	2	0				
14	13	8	8	0				
15	14	4	5	-1	1.00	2.50		2.50
16	15	4	5	-1	1.00	2.50		2.50
17	16	9	9	0				
18	17	4	4	0				
19	18	1	1	0				
20	19	0	5	-5	5.00	9.00		9.00
21	20	9	9	0				

Figure 35. The formula for the test statistic



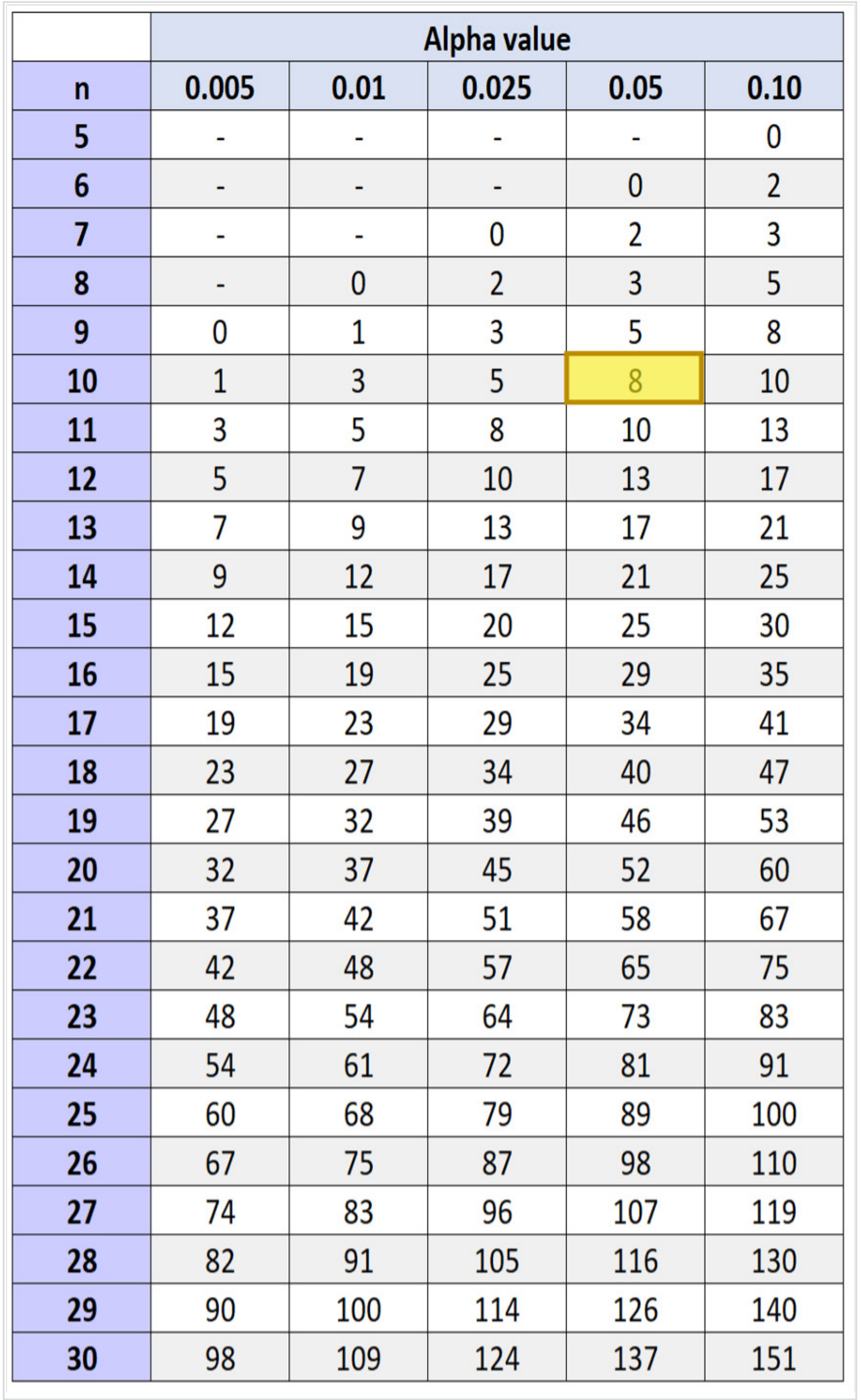
	A	B	C	D	E	F	G	H	I	J
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks		
2	1	4	6	-2	2.00	6.00		6.00		
3	2	6	4	2	2.00	6.00	6.00			
4	3	6	6	0						
5	4	6	6	0						
6	5	3	5	-2	2.00	6.00		6.00		
7	6	2	7	-5	5.00	9.00		9.00		
8	7	6	7	-1	1.00	2.50		2.50		
9	8	6	7	-1	1.00	2.50		2.50		
10	9	7	7	0						
11	10	2	2	0						
12	11	2	7	-5	5.00	9.00		9.00		
13	12	2	2	0						
14	13	8	8	0						
15	14	4	5	-1	1.00	2.50		2.50		
16	15	4	5	-1	1.00	2.50		2.50		
17	16	9	9	0						
18	17	4	4	0						
19	18	1	1	0						
20	19	0	5	-5	5.00	9.00		9.00		
21	20	9	9	0						
22										
23		4	6	Test statistic				6.00		
24				Sample size				=COUNT(G2:H20)		
25										



	A	B	C	D	E	F	G	H	I	J
1	Participant number	Pre-intervention	Immediately Post-intervention	Difference	Absolute difference	Ranks of Absolute difference	Positive ranks	Negitive ranks		
2	1	4	6	-2	2.00	6.00		6.00		
3	2	6	4	2	2.00	6.00	6.00			
4	3	6	6	0						
5	4	6	6	0						
6	5	3	5	-2	2.00	6.00		6.00		
7	6	2	7	-5	5.00	9.00		9.00		
8	7	6	7	-1	1.00	2.50		2.50		
9	8	6	7	-1	1.00	2.50		2.50		
10	9	7	7	0						
11	10	2	2	0						
12	11	2	7	-5	5.00	9.00		9.00		
13	12	2	2	0						
14	13	8	8	0						
15	14	4	5	-1	1.00	2.50		2.50		
16	15	4	5	-1	1.00	2.50		2.50		
17	16	9	9	0						
18	17	4	4	0						
19	18	1	1	0						
20	19	0	5	-5	5.00	9.00		9.00		
21	20	9	9	0						
22										
23		4	6	Test statistic				6.00		
24				Sample size				10		

Figure 36. The formula for the sample size

These numbers can then be checked against the alpha value chart for this test (Figure 37). This is done by identifying the required alpha value and then the number in the sample. For the finding to be statistically significant at the alpha value of .05 the test would need to be equal to or lower than 8 (highlighted in yellow). As the Z score was 6, there is a significant difference between groups.



	Alpha value				
n	0.005	0.01	0.025	0.05	0.10
5	-	-	-	-	0
6	-	-	-	0	2
7	-	-	0	2	3
8	-	0	2	3	5
9	0	1	3	5	8
10	1	3	5	8	10
11	3	5	8	10	13
12	5	7	10	13	17
13	7	9	13	17	21
14	9	12	17	21	25
15	12	15	20	25	30
16	15	19	25	29	35
17	19	23	29	34	41
18	23	27	34	40	47
19	27	32	39	46	53
20	32	37	45	52	60
21	37	42	51	58	67
22	42	48	57	65	75
23	48	54	64	73	83
24	54	61	72	81	91
25	60	68	79	89	100
26	67	75	87	98	110
27	74	83	96	107	119
28	82	91	105	116	130
29	90	100	114	126	140
30	98	109	124	137	151

Figure 37. Alpha value chart for a Wilcox signed rank test

The result can then be reported as

> A Wilcoxon signed rank test showed that there was a significant difference (Z = 6, p < .05) between the scores, with the median score pre-intervention being 4 compared to a post-intervention score of 6.

Kruskal Wallis Test

Why use a Kruskal Wallis Test?

An evaluation design often contains the same measure administered at different time points or to different groups to identify if anything had altered. However, running multiple tests on the data increases the chance of an error occurring, a p-value of .05 means there is a 1 in 20 chance of an error occurring. Therefore, completing multiple t-tests introduces a large element of chance into the results. To reduce the possibility of error, not only should the appropriate test of difference but also, on non-parametric data, a Kruskal Wallis Test, which if the equivalent of the parametric ANOVA test.

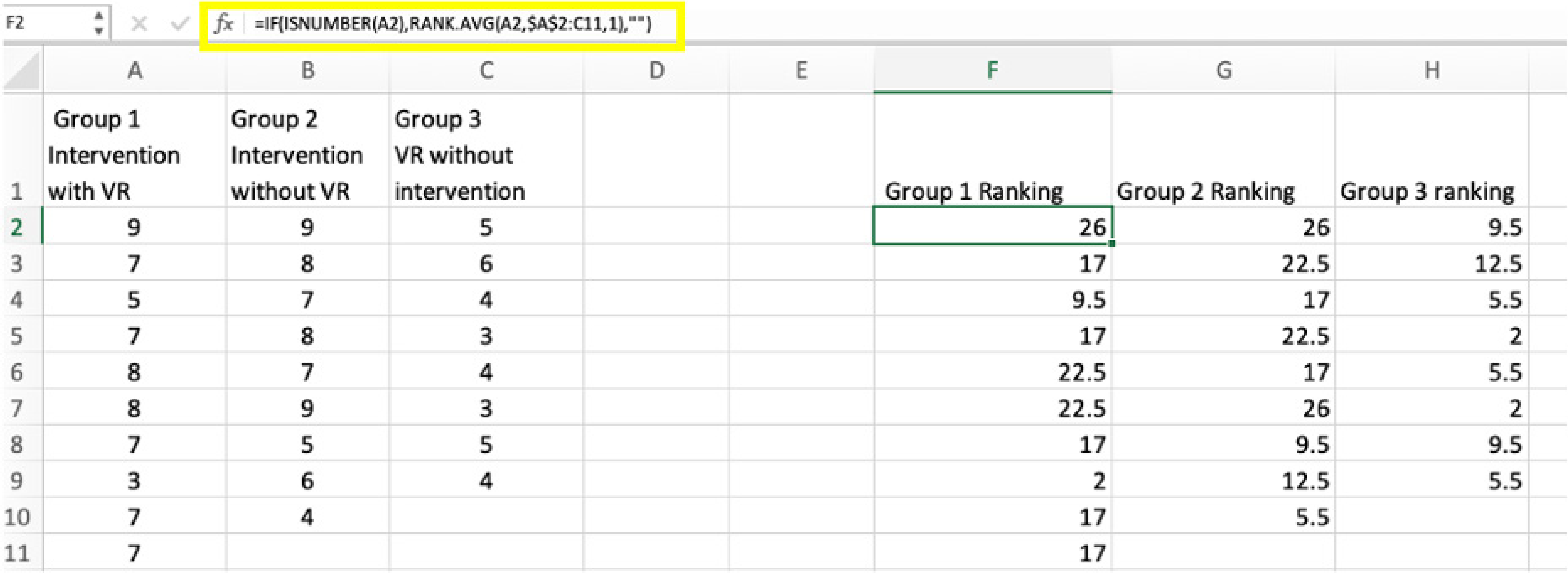
In this example, the evaluator wants to know if the inclusion of a Virtual Reality (VR) film for a firefighter training session enhances the training outcomes. To evaluate this 3 groups are used: Group 1 receives the training with the VR, Group 2 receives the training without the VR, and Group 3 just receives the VR film without the training. **Please note that this is a fictitious example for demonstration purposes.**

Using Excel to complete a Kruskal Wallis test

The first step in completing this test is to rank the scores. The input for this is below, with Figure 38 showing it applied to Group 1:

Group 1 =IF(ISNUMBER(A2),RANK.AVG(A2,$A$2:C11,1),””)
Group 2 =IF(ISNUMBER(B2),RANK.AVG(B2,$A$2:C11,1),””)
Group 3 =IF(ISNUMBER(C2),RANK.AVG(C2,$A$2:C11,1),””)

Once inputted for the first rows of each of the three ranking columns, they can be dragged down to populate the rows beneath (Figure 38).



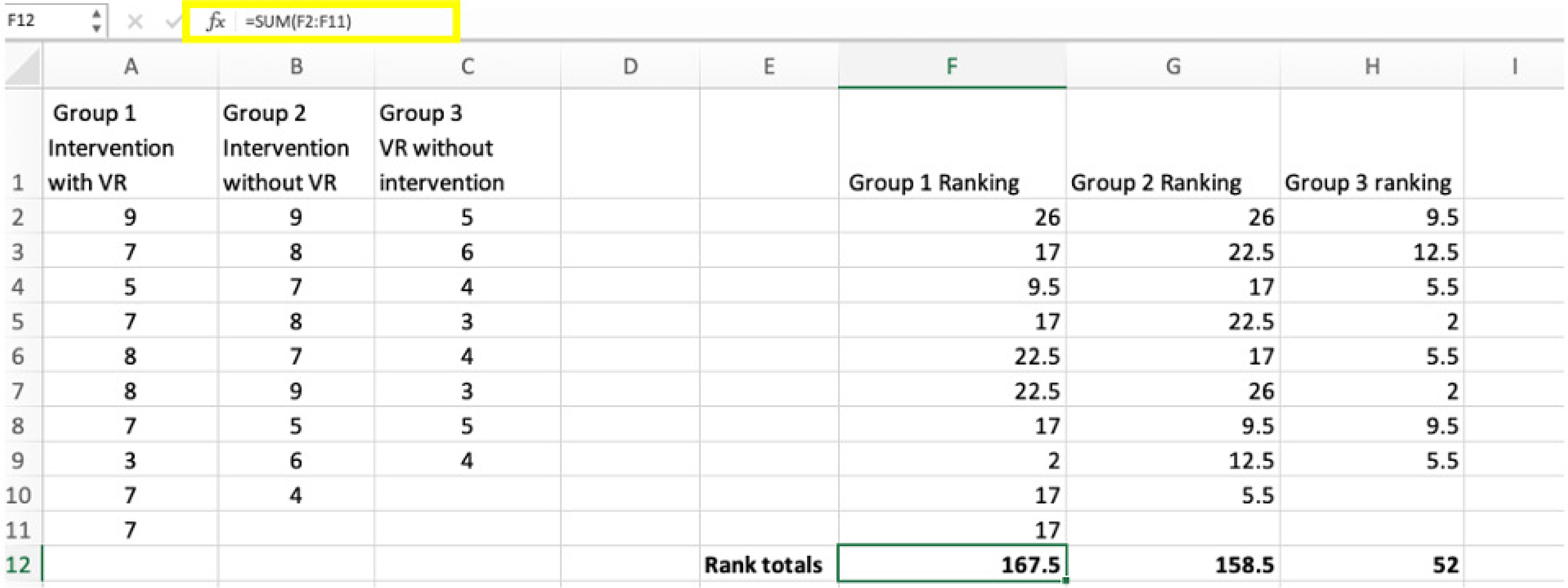
F2

```
=IF(ISNUMBER(A2),RANK.AVG(A2,$A$2:C11,1),"")
```

	A	B	C	D	E	F	G	H
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking
2	9	9	5			26	26	9.5
3	7	8	6			17	22.5	12.5
4	5	7	4			9.5	17	5.5
5	7	8	3			17	22.5	2
6	8	7	4			22.5	17	5.5
7	8	9	3			22.5	26	2
8	7	5	5			17	9.5	9.5
9	3	6	4			2	12.5	5.5
10	7	4				17	5.5	
11	7					17		

Figure 38. Ranking scores

The next step is to total the ranks (Figure 39), the input for this for Group 1 Ranking 1 would be =SUM(F2:F11), this can then be dragged over for the other 2 Ranking groups.

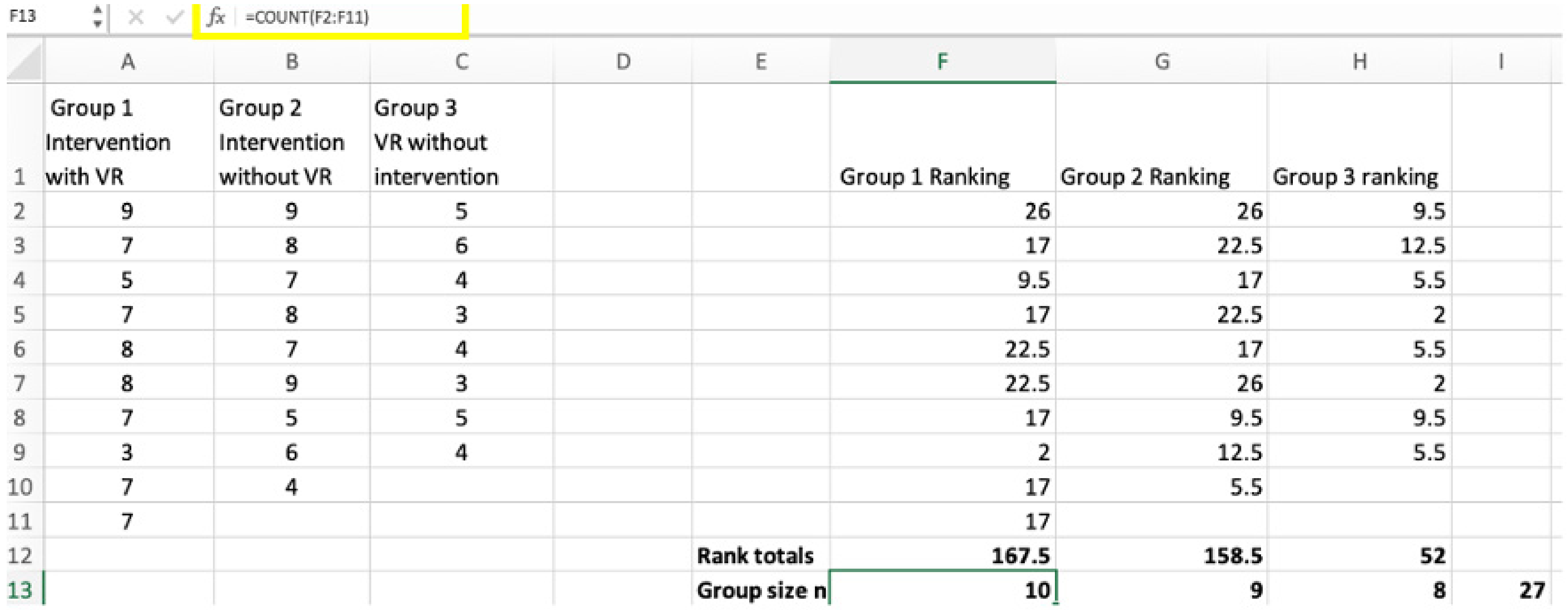


F12 | *fx* =SUM(F2:F11)

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					**Rank totals**	**167.5**	**158.5**	**52**	

Figure 39. Rank totals

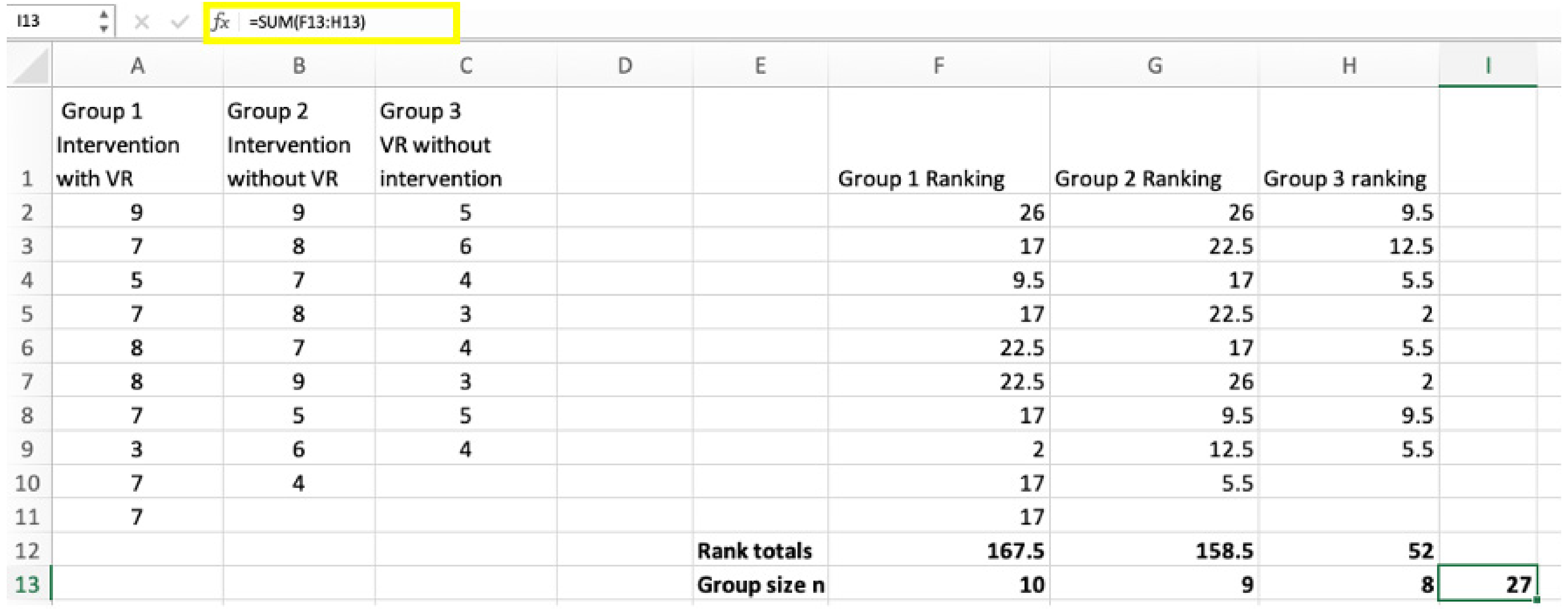
The next stage is to enter the Group size n values, which is the total for the 3 samples. This can be calculated using Excel for the Group 1 Ranking column using =COUNT(F2:F11), this can then be dragged across to the other groups (Figure 40). The total can be calculated using =SUM(F13:H13) (Figure 41)



F13 =COUNT(F2:F11)

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					Rank totals	167.5	158.5	52	
13					Group size n	10	9	8	27

Figure 40. Individual Group size n

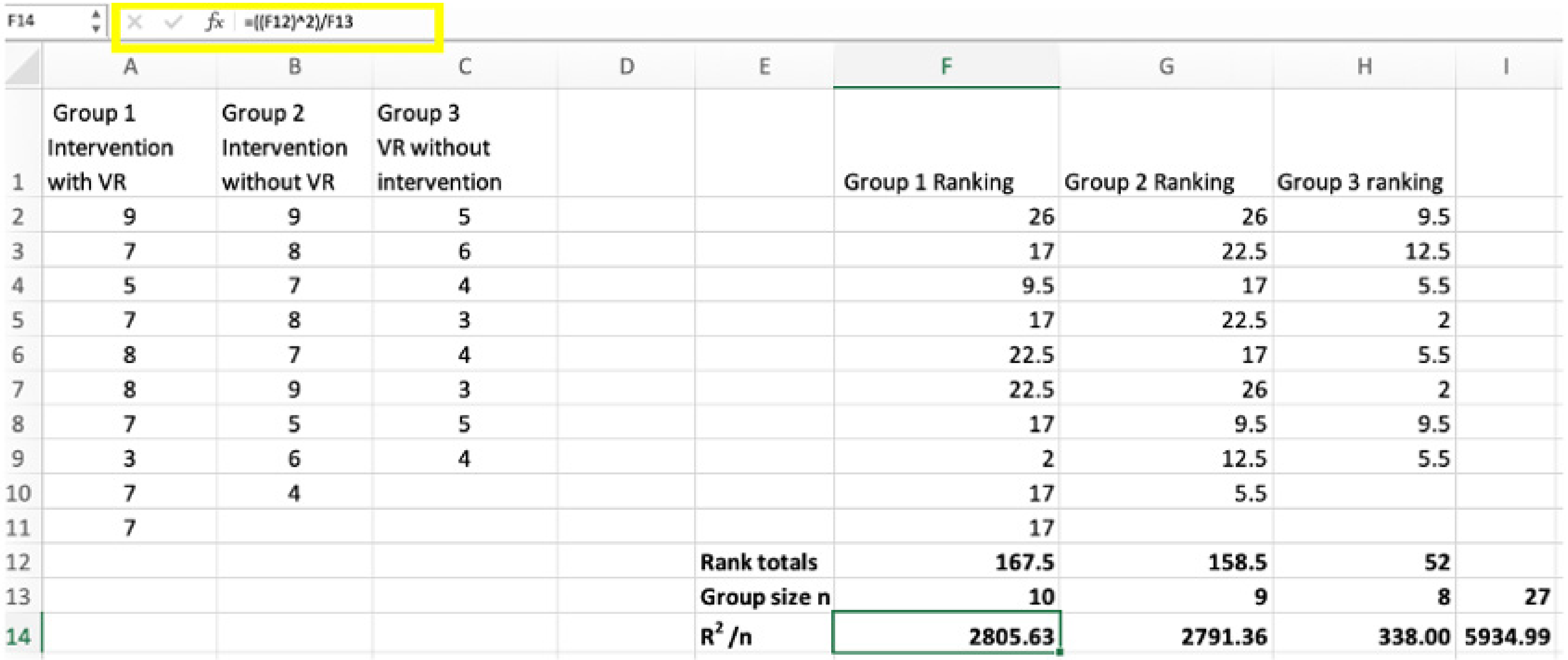


I13 =SUM(F13:H13)

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					Rank totals	167.5	158.5	52	
13					Group size n	10	9	8	27

Figure 41. Total Group size n

The next step is to work out the R2/n Figure. The equation for Group 1 Ranking column is =((F12)^2)/F13. Once inputted, this can be dragged over to the other columns (Figure 42). The total, in this case 5934.99, is the sum of the 3 Ranking columns =SUM(F14:H14).

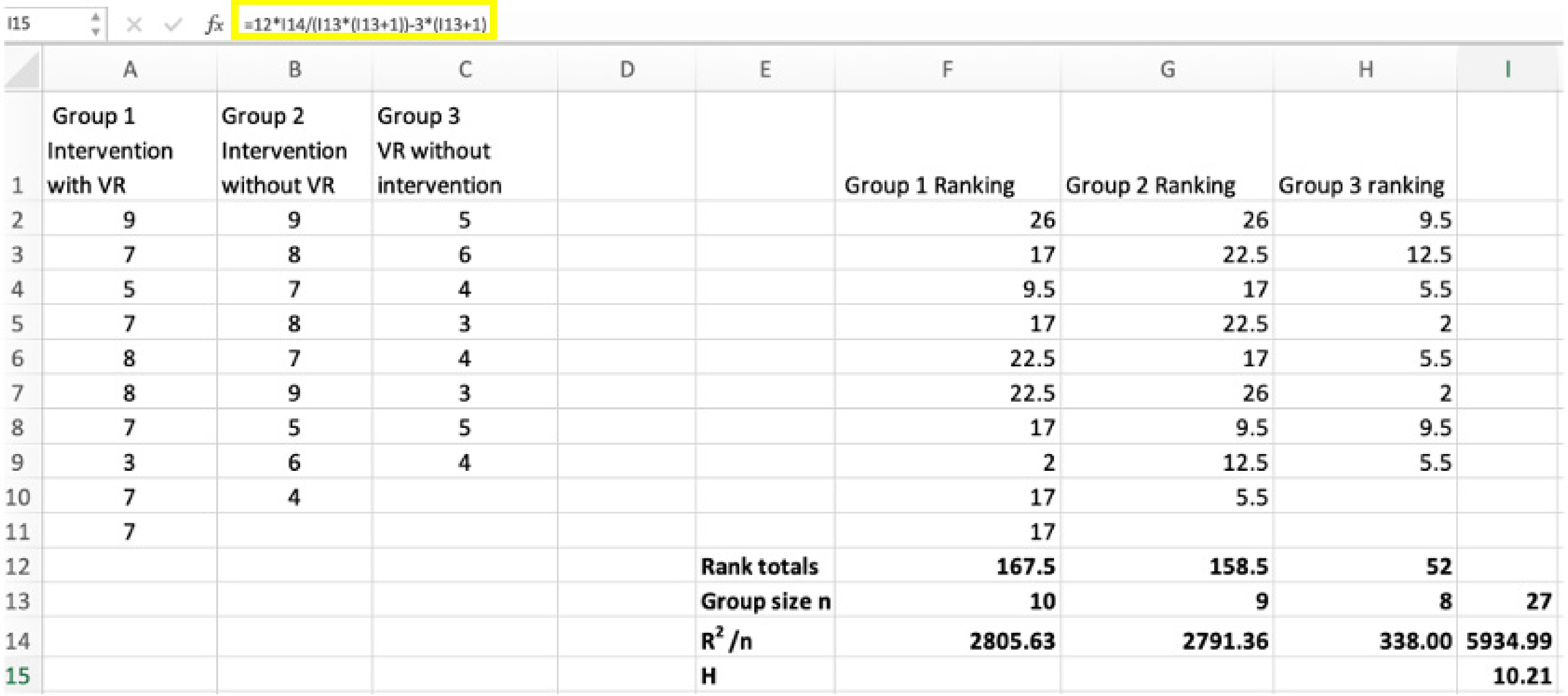


F14 fx =((F12)^2)/F13

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					Rank totals	167.5	158.5	52	
13					Group size n	10	9	8	27
14					R^2 /n	2805.63	2791.36	338.00	5934.99

Figure 42. R2/n calculation

The group size n figure and the R2 /N figures are used to calculate the test statistic (Figure 43), referred to in this test as the H-value. The input is =12*I14/(I13*(I13+1))-3*(I13+1) A couple of issues to note. The 12 and the +1 figures are fixed in the equation, these do not alter. The – 3 figure refers to the number of groups. The H value, in this case, was 10.21.

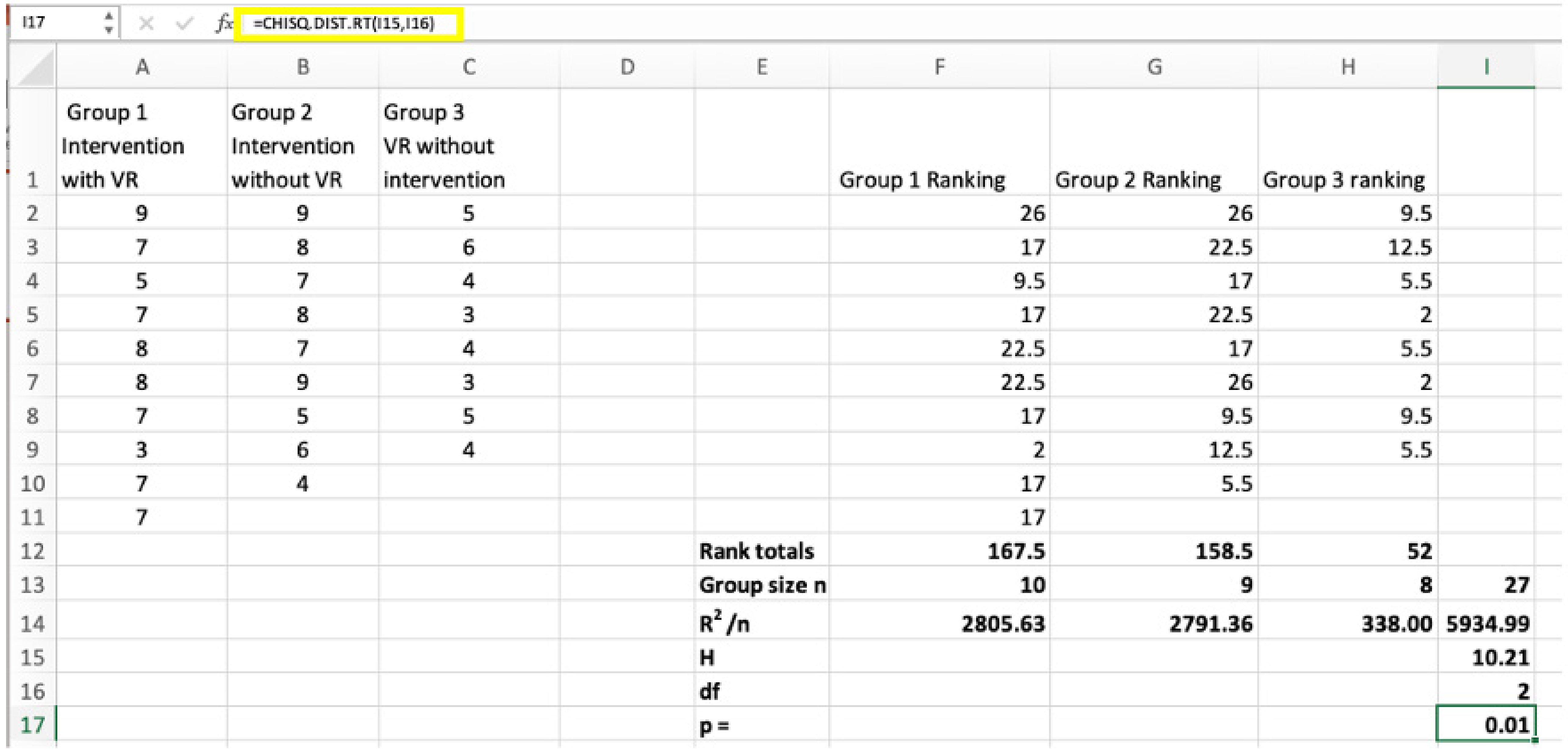


=12*I14/(I13*(I13+1))-3*(I13+1)

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					**Rank totals**	**167.5**	**158.5**	**52**	
13					**Group size n**	**10**	**9**	**8**	**27**
14					**R^2 /n**	**2805.63**	**2791.36**	**338.00**	**5934.99**
15					**H**				**10.21**

Figure 43. H calculation

The final element is to work out the degrees of freedom (df) and the p-value. The df for this test would be 2, as the test is comparing one score against the two other scores. The p-value input is, therefore, =CHISQ.DIST.RT(I15,I16), which uses the H score and the df, giving a p-value of .01, meaning the difference was significant (Figure 44).



I17 =CHISQ.DIST.RT(I15,I16)

	A	B	C	D	E	F	G	H	I
1	Group 1 Intervention with VR	Group 2 Intervention without VR	Group 3 VR without intervention			Group 1 Ranking	Group 2 Ranking	Group 3 ranking	
2	9	9	5			26	26	9.5	
3	7	8	6			17	22.5	12.5	
4	5	7	4			9.5	17	5.5	
5	7	8	3			17	22.5	2	
6	8	7	4			22.5	17	5.5	
7	8	9	3			22.5	26	2	
8	7	5	5			17	9.5	9.5	
9	3	6	4			2	12.5	5.5	
10	7	4				17	5.5		
11	7					17			
12					**Rank totals**	**167.5**	**158.5**	**52**	
13					**Group size n**	**10**	**9**	**8**	**27**
14					**R^2 /n**	**2805.63**	**2791.36**	**338.00**	**5934.99**
15					**H**				**10.21**
16					**df**				**2**
17					**p =**				**0.01**

Figure 44. P value calculation

A set of Mann-Whitney U tests would then need to be performed to identify which groups differ. However, to reduce the change of an error occurring these need to be completed using a Bonferroni correction which adjusts the alpha value. The correction is calculated by dividing the alpha value by the number of tests. In this case, the alpha value is .05, and there are 3 tests being conducted. Therefore, to be significant the p-value for the t-tests would have to be at or below .02. (.05 / 3 = .02 when rounded to 2 decimal points).

The results could then be reported as:

> A Kruskal-Wallis test found a significant difference between groups (H(2)=10.21, p = .01). A set of Mann-Whitney U tests were performed using the Bonferroni correction to identify which Groups differed.
>
> No difference was found between Group 1 and Group 2, with both groups having a medium of 7 (Z = -0.40, p = 0.69).
>
> Group 3 had a median score of 4, and this was found to be significantly lower than both Group 1 (Z = -75, p = .01) and Group 2 (Z = -2.79, p = .01).
>
> It is therefore concluded that the inclusion of the film had no impact on the outcomes of the intervention, and the use of the film alone did not have as strong an effect as the training with or without the inclusion of the VR film.

Qualitative Analysis

Introduction

Qualitative approaches can provide deep insight into why something has or is occurring. However, the quality of the insight is dependent on the quality of the data and the robustness of the analysis. This section provides an overview of how to complete a thematic analysis, one of the most widely used qualitative methods.

Trustworthiness

Whereas quantitative approaches to evaluation consider the reliability and validity of the data, qualitative approaches focus on the trustworthiness of the data and the analytical process used. What would be regarded as trustworthy will be dependent on the nature of the evaluation, but consideration should be given to:

Credibility
How accurately the analysis represents the data it is reporting.

Transferability
Whether the findings could be transferred to a wider context.

Dependability
Was a clear process followed that demonstrates how the conclusions were drawn.

Confirmability
Did the findings come from the data? This is achieved through the application of the previous 3 criteria.

Further information about these concepts can be found at: https://journals.sagepub.com/doi/pdf/10.1177/1609406917733847

Approached to Qualitative Analysis

There is a wide range of approaches that can be undertaken for the analysis of qualitative data. Most approaches have the aim of distilling the information contained in the data into a format that can be clearly understood. For this reason, many of the approaches overlap, although the focus of the analysis may differ. These approaches include:

Content analysis
This approach looks to identify keywords used in a text. For example, an interview transcript relating to how quickly the FRS attended an incident could be reviewed for keywords such as: quickly, time, speed, fast, etc. This type of analysis can be quickly completed on Word using the search function. This approach would be seen as a very basic form of analysis and would need to be used with caution, as it is likely to miss key ideas and concepts that a more in-depth analysis would have identified. More complex forms of content analysis can be undertaken, but these start to move toward a thematic analysis. More information about content analysis is available at: https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis

Thematic analysis
This is the most common form of analysis and the one this document will focus on. The approach aims to condense the data into a set of themes that reflect the ideas and meaning contained in the data. A thematic approach explores the data in-depth, resulting in a more detailed and trustworthy outcome.

Narrative analysis
This term includes a range of approaches (Riessman, 2008) that aim to explore a participant's stories, whether they are written or in oral form. The concept behind this approach is that the stories help people make sense of their lives (Figgou &

Pavlopoulos, 2015), and therefore understanding these stories will improve our understanding of the social world. Further information about this approach can be found at: https://www.tandfonline.com/doi/full/10.1080/21642850.2018.1515017

Discourse analysis

This approach looks at how language is used in a social context to convey meaning, influence others and establish norms. More information about this approach can be found at https://research.ncl.ac.uk/methodshub/methods/discourseanalysis/

Ground theory

Grounded Theory is a complex concept that is constantly developing. The simplest explanation is that it looks to generate theories and concepts from the data in an iterative process as opposed to more traditional approaches, which tend to be more theory-driven. More information on grounded theory can be found at: https://www.personal.psu.edu/wxh139/grounded.htm

Thematic Analysis

A thematic analysis aims to identify, analyse, organise and report themes found in the data (Braun and Clark, 2006). Thematic analysis is a six-phase process, although it is important to recognise that it will be necessary to move back and forth through the steps during the analysis.

Thematic Analysis Steps	Means of Establishing Trustworthiness
1. Data familiarisation	• Prolong engagement with the data • Triangulate different data modes • Document thoughts about potential codes/themes • Store raw data in well-organised archives • Keep notes, transcripts, etc.
2. Initial codes	• Work with others to code and cross-check the data • Keep notes on why a code was given and if the coding changes
3. Generating themes	• Researcher triangulation • Cross-check themes with the raw data
4. Reviewing themes	• Research triangulation • Themes and subthemes checked by others
5. Defining and naming themes	• Researcher triangulation • Define themes and cross check with coded and raw data
6. Producing the report	• Describe processing of coding and analysis in sufficient detail • Explain context in detail

Table 3. Establishing trustworthiness using a Thematic Analysis (Based on Nowell et al., 2027)

Table 3 is taken from Nowell et al. (2017) and shows the phases that needs to be undertaken and steps that should be taken to maximise the trustworthiness of the analysis.

Step 1- Data familiarisation

Prior to commencing analysis, the evaluator needs to be familiar with the data. To do this, the data should be in a written format that allows the evaluator to read and reread the participant's respondence. The recording should also be listened to, as the transcripts are unlikely to convey the importance given to the words used by the participant. It is often useful for the evaluator to transcribe the interview themselves as this will help them with the familiarisation process.

Whilst reading the transcripts, the comment facility in Word can be used to record initial thoughts that may occur to the evaluator. The text, audio and video should be reviewed repeatedly until the evaluator is happy that the data can provide no new insights. All notes and records should be kept ensuring the analytical process is open and trustworthy.

Step 2 - Initial Coding

Once the evaluator is fully familiarised with the data, the process of coding can begin. Phrases or sentences in the text should be highlighted and coded as shown below. This text is based on an interview with a driver who had been asked to attend a driver intervention scheme that aims to reduce reoffending.

> (Interviewer) What were your initial thoughts when you first heard about the scheme?
>
> Participant 1: **I was very sceptical because you get a load of jargon, you know, people who go, oh we can do this and that. I've never had a better service in all honesty, so I started off with, you know, a pessimistic view about it but then slowly realised, no, this is good, this is a good service.**

Concerns about the scheme on enrolment | Clarity of language | Recognition of benefits | Level of service

As this process requires the evaluator to interpret what was meant by the participant to develop a set of initial codes. Do the subjective nature of the process, ideally, it should be done by two people working independently. They can then come together to cross-check their coding and agree on any areas of difference.

Step 3 - Generating themes

A theme should provide an overview of the meaning derived from the coding process and the raw data. This is important, as the analysis must accurately reflect the meaning contained in the data. Whilst a single code can be used as a theme, a theme will often encompass multiple codes. An example of the initial themes derived from the analysis of the complete interview is shown in Table 4. Table 4 includes more codes than the ones shown above, as it is based on more answers than the one given above.

Initial themes	Coding
Benefits	• Anger management • Psychological well-being • Recognition of benefits • Physical wellbeing
Driving behaviour	• Drink driving • Driving offences
Levels of service	• Level of service • Support provided
Uncertainty	• Understanding scheme • Clarity of language • Concerns about scheme enrolment

Table 4. Initial themes generation based on initial coding

Step 4 - Reviewing themes

To ensure the themes reflect the data, the themes should be cross-checked with the codes and the raw data to ensure that no meaning has been lost in the process. During this step, new themes may be added, and the existing themes may be reworded to better reflect the data. At this point, themes should be reviewed by peers to ensure that they are representative of the data. Table 5 shows how the initial themes were further developed.

Themes	Coding
Wellbeing on referral	• Anger management • Psychological wellbeing • Physical wellbeing • Drink driving • Driving offences
Client satisfaction	• Level of service • Support provided • Recognition of benefits
Scheme recognition	• Understanding the scheme • Clarity of language • Concerns about scheme enrolment

Table 5. Review of themes linked to coding

Step 5 - Defining and naming themes

Once a list of themes has been developed, their wording should be reviewed to ensure that what they are conveying is clear and accurate. This process is likely to include developing a set of wording that defines the theme and what it is telling the reader about the data, and how it is relevant to the evaluation (Table 6). Once again, it may be necessary to review the themes during this process. Including others in this process is important to ensure that the themes are reflective of the content and detail contained within the data.

Final theme	Coding	Theme definition
Wellbeing on referral	• Anger management • Psychological well-being • Physical wellbeing • Drink driving • Driving offences • Drug driving	The participants physical and psychological state on referral.
Scheme recognition	• Understanding the scheme • Clarity of language • Concerns about scheme enrolment	The participants' awareness and understanding of the scheme prior to enrolment.
Participant reaction to the scheme	• Level of service • Support provided • Recognition of benefits	How the participants viewed the support given to them by the scheme.

Table 6. Final themes, coding and definitions

Step 6 - Reporting

The report should give a clear outline of why and how the data was collected and the methodology used in the analysis. The method should cover the previous 5 steps so that the reader is able to follow the process from the initial data to the conclusions presented. The report should include quotes from the data to bring themes to life and provide insight into what was said and the context in which it was said.

Whilst, in the context of evaluation, it is likely that some simple statistic of how many times a theme appeared during the analysis will be reported, the reporting of the data should be wider to reflect the richness of the data.

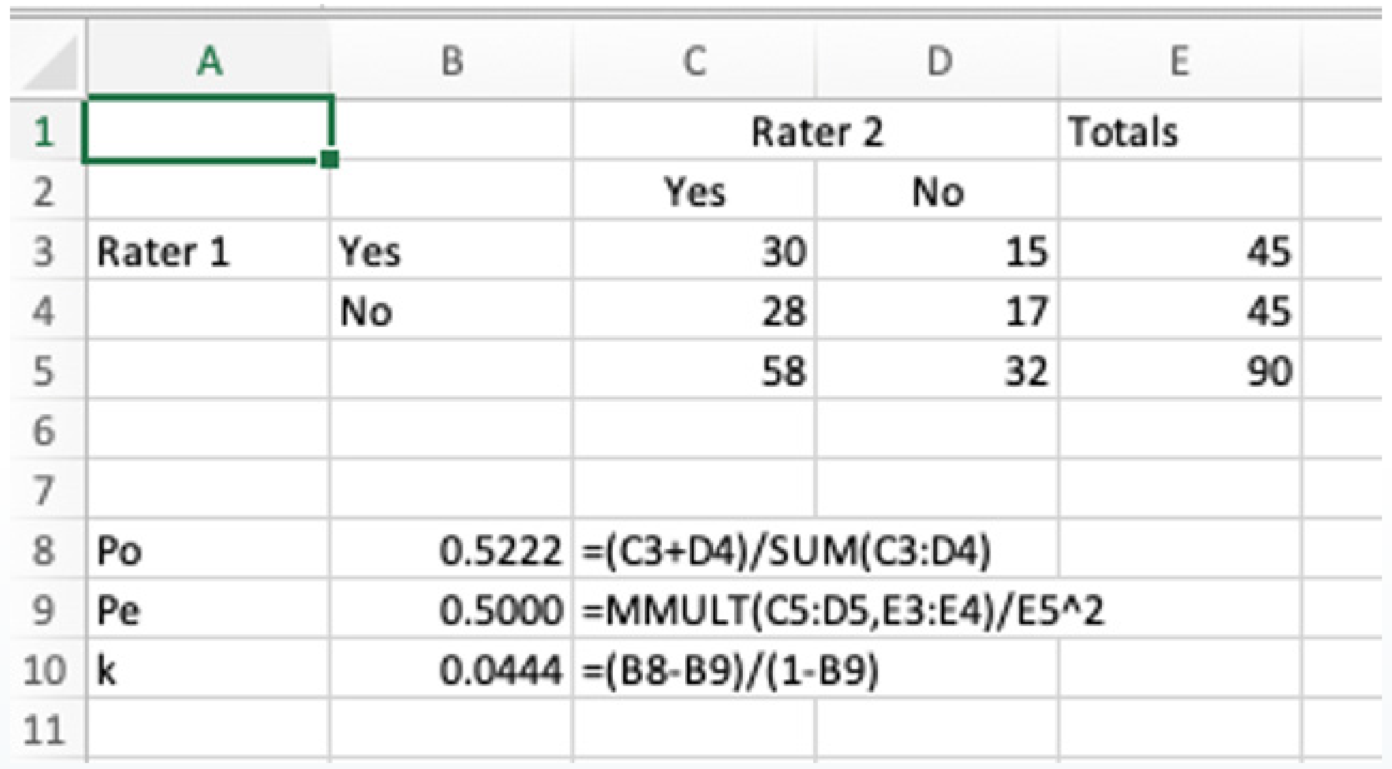
Interrater Reliability

Throughout the process, consideration should be given to ensuring the consistency of the coding where 2 coders are used. This consistency is referred to as interrater reliability. A type of correlation test called a Cohen kappa can be completed to check the level of interrater reliability.

In the example shown in Figure 45, coder 1 coded 45 pieces of data. Coder 2 agreed on the coding 28 times but disagreed on 17 occasions, coding it to another category. The equation for the test is κ = (Po – Pe)/(1-Pe):

- Po = Relative observed agreement
- Pe = Hypothetical probability of chance agreement

The Kappa statistic (κ) in the example is 0.44, Table 7 provides the interpretation.



	A	B	C	D	E
1			Rater 2		Totals
2			Yes	No	
3	Rater 1	Yes	30	15	45
4		No	28	17	45
5			58	32	90
6					
7					
8	Po	0.5222	=(C3+D4)/SUM(C3:D4)		
9	Pe	0.5000	=MMULT(C5:D5,E3:E4)/E5^2		
10	k	0.0444	=(B8-B9)/(1-B9)		
11					

Figure 45. Final themes, coding and definitions

Cohen’s Kappa	Interpretation
0	No agreement
0.10 - 0.20	Slight agreement
0.21 -0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near perfect agreement
1	Perfect agreement

Table 7. Cohen kappa statistic interpretation

The result can then be included in the reporting section in the following way:

> A Cohen Kappa test was performed to check the interrater reliability between researchers was moderate (κ = .44).

Full details for this Cohen Kappa test and the related Fleiss kappa (with is used if there are more than two coders) at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Reporting Qualitative Analysis

The aim of the example below is to show how the results of a qualitative analysis can be reported. This section only covers the methodology and results section of a report. A full report is likely also to include the following sections: Executive summary, literature review and background to the intervention, description of the intervention, a full description of the methodology used, results, discussion and recommendations.

Method

Five 15-minute semi-structured telephone interviews were conducted with scheme participants. The participants were sent an information sheet and consent form, which had to be returned prior to the interview appointment being made. Before commencing the interview, the evaluator confirmed over the phone that they were happy to proceed as outlined in the information sheet and reminded them that the interview would be recorded.

The evaluator used a semi-structured interview template (This should be added to the report as an appendix) to guide the interview but was allowed to use follow-up questions to explore areas of interest. On completion, the recording was transcribed by the evaluator. A thematic analysis was then completed using the approach advocated by Braun and Clark (2006), which advocates a six-stage approach, with the final stage being reporting. The process used is outlined below:

Data familiarisation

The transcripts and interview recordings were read and listened to repeatedly. Initial thoughts and reflections were recorded as notes in the transcriptions.

Initial codes

Based on the initial notes and reflection, a range of codes were developed. These codes were then used to code all 5 transcripts. A second evaluator reviewed the coding and the transcripts, this resulted in further development of the codes being used.

Generating themes

The codes were then placed under broad headings. This resulted in four themes being developed.

Reviewing themes

The initial four themes were then reviewed against the raw data to see if anything was being missed and if the themes accurately reflected the data. This resulted in all the themes being reworded, which in turn reduced the number of themes to 3.

Defining and naming themes

Clear definitions were added to the themes, and the themes were then again reviewed against the data. This review was completed by both the evaluator and a member of the wider team.

Results

The three themes that emerged from the analysis were:

- Wellbeing on referral
- Scheme identity
- Participant reaction

Wellbeing on referral

The well-being on referral theme was defined in the evaluation as 'The participant's physical and psychological state on referral'. All but one of the participants mentioned that they were experiencing some physical and/or psychological issues that they had been struggling to address. These issues included anger management, depression and addiction:

> "Well when I was in a real struggle, I was on a hell of a lot of medication, I'd gone through a lot health wise and I was struggling to deal with things"
> **- Participant 2**

> "Instead of punching a wall, if you get what I mean and headbutting and throwing your phone up against the wall, getting annoyed."
> **- Participant 1**

These issues were often linked by the participants to the driving offence that had led them to participate in the scheme.

> "I just think general ill health, like, sort of addiction, depression and God knows what and also with the nature of how much driving offences I've committed."
> **- Participant 4**

As the scheme aims to reduce reoffending by offering support to people who have a range of complex physical and mental health needs, the emergence of this theme provides credence that the scheme is successfully engaging the target group. It also offers support to the idea that driving offences are often associated with a range of wider issues that, left unaddressed, are likely to result in further offences being committed.

Scheme identity

This theme was defined as 'The participants' awareness and understanding of the scheme prior to enrolment'. None of the participants appeared to have a clear understanding of what the scheme was offering and how it related to them.

> "I didn't have a clue about it, so I started off with, thinking this going to be a waste but I don't have much to lose view then slowly I realised, no, this is good."
> **- Participant 1**

> "I didn't know what I'd gain, but I gained a lot of help"
> **- Participant 4**

The language used to describe the scheme may not have aided this process:

> "I was very sceptical because you get a load of jargon, you know,"
> **- Participant 5**

These findings would indicate that the scheme needs to explore how it can present its content in a concise way. It should be remembered that the people being asked to enrol on the scheme may be vulnerable and concerned about their future when they are first approached, therefore the initial description needs to be clear about how the scheme can assist them.

Participant reaction to the scheme

This theme was defined as 'How the participant viewed the support given to them by the scheme'. Without exception, the scheme was viewed positively:

> "I've never had a better service in all honesty."
> **- Participant 4**

> "I'm like brilliant, you know. It took the weight off my shoulders with everything, my driving and all sorts."
> **- Participant 2**

The client also felt the scheme was delivering positive benefits to them:

> "I learnt so it's just been a learning curve really."
> **- Participant 4**

> "In a bad situation taken a lot of weight of my shoulders and help, advice and guidance that I might not have found without the help"
> **- Participant 3**

The positive feedback was resounding and would indicate that the scheme is well received and appears to be delivering a high level of service to the participant.

Glossary of Terms

Alpha value
The alpha value is the threshold for statistical significance. The generally accepted alpha value is .05, meaning a p-value at or below .05 would indicate that any difference between values is significant.

Bonferroni correction
This correction lowers the alpha value and is used where multiple t-tests are required. The correction is calculated by dividing the alpha value by the number of tests conducted.

Analysis of Variance (ANOVA)
A set of statistical tests to analyse differences among means.

Central tendencies
How closely the data is clustered around the mean, median or mode.

Coefficient
A value used to multiply a variable.

Cohen’s Kappa
A correlation test is used to assess Interrater reliability.

Correlation
A measure that shows how closely two variables are linearly related. It should be remembered that a correlation does not mean one variable cause another.

Degrees of Error
The margin of error.

Degrees of Freedom

Is the number of pieces of information used to calculate a statistic. It is calculated by the sample size less the number of parameters being used in the test.

Distribution

Distribution is a term used to describe how frequently values occur in a data field. Figure 46 shows a normal distribution. In a perfect distribution, the mean, mode and median would all be the same, in this example, their value would all be 4. The black line drawn onto the graph is to demonstrate the classic bell-shaped curve that illustrates a normal distribution.

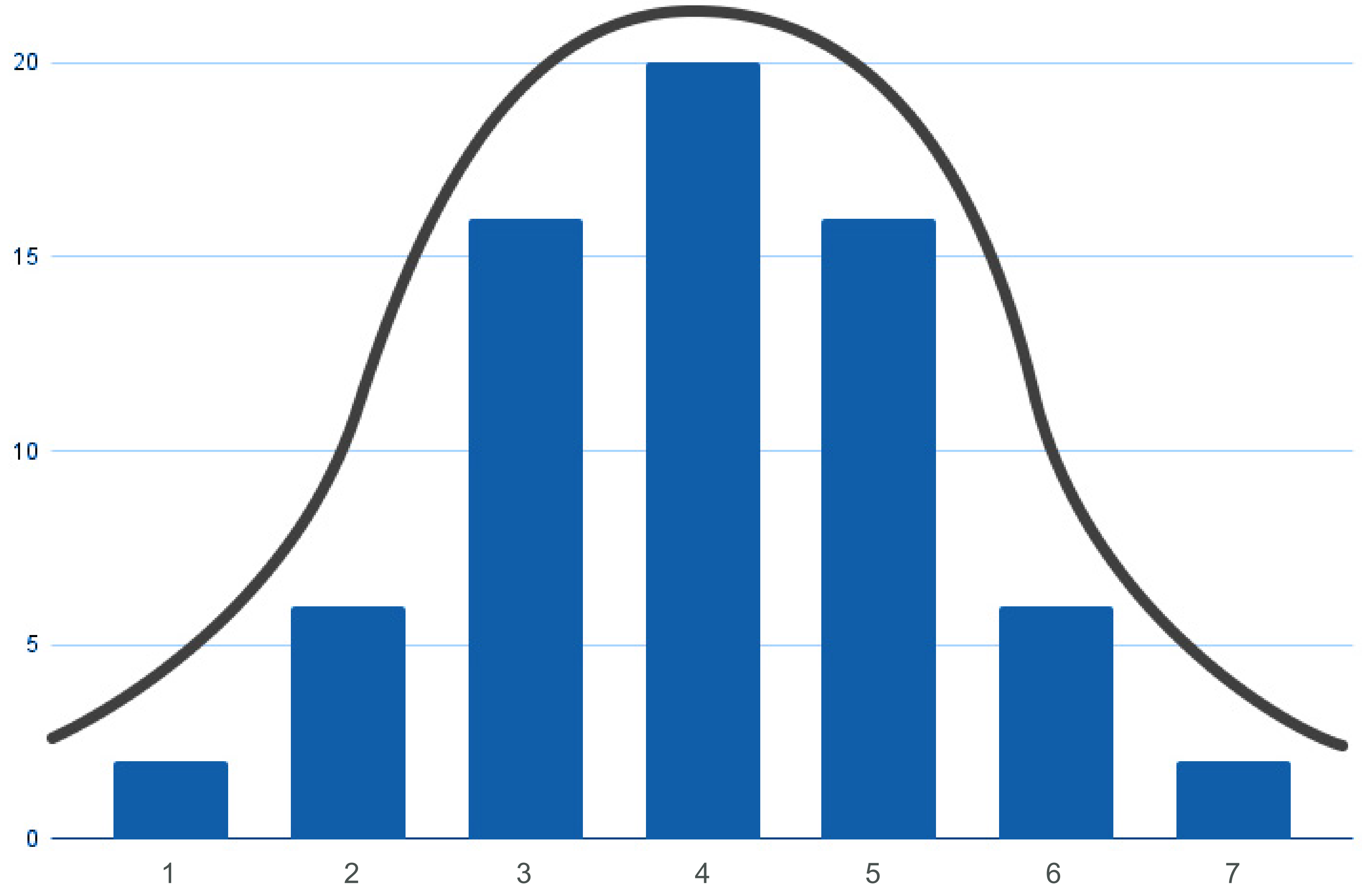


Figure 46. Normal distribution

If the distribution is similar to the histograms shown in Figure 47, it would not be regarded as being normally distributed. Understanding how data is distributed is important, as it will inform the type of test that should be used to identify if there are any statistically significant differences between one set of results and another.

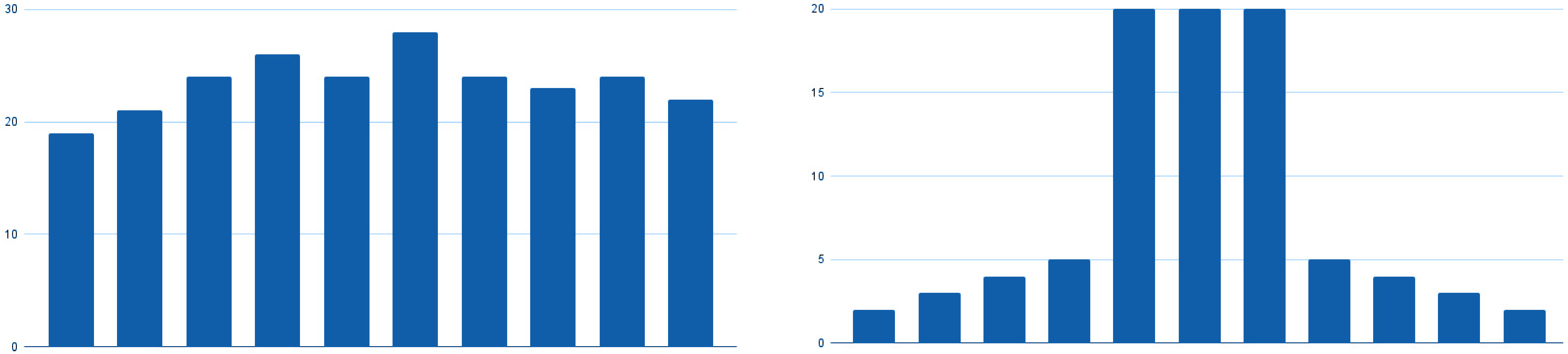


Figure 47. Uneven distributions

F-test

A test to assess the ratio of the variance between samples.

Histogram

A histogram is a type of graph that shows how data is distributed. To generate a histogram on Excel, highlight the data to be explored, then go to Insert and select the statistical graph icon and select histogram (Figure 48).

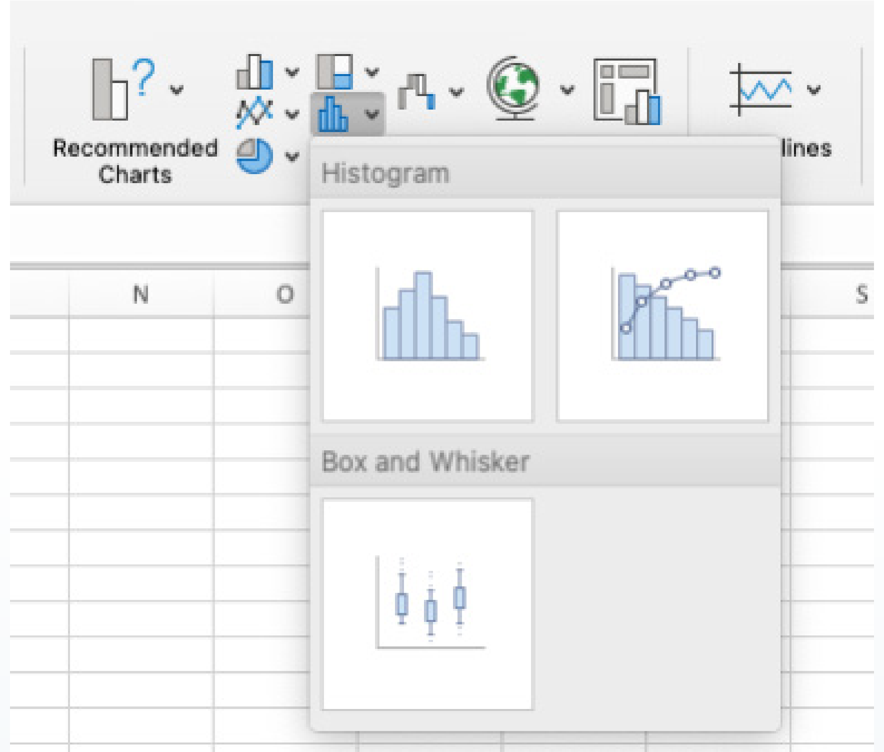


Figure 48. Producing a histogram

Figure 49 shows a normal distribution, Figure 50 shows a left-skewed distribution, with Figure 50 showing a right-skewed distribution.

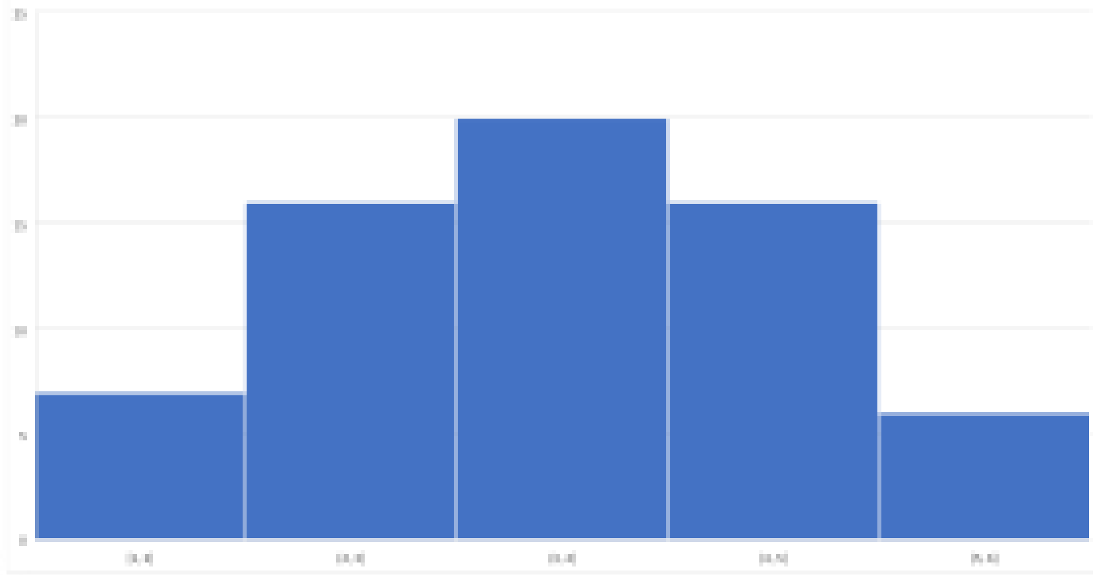


Figure 49. Histogram showing a normal distribution

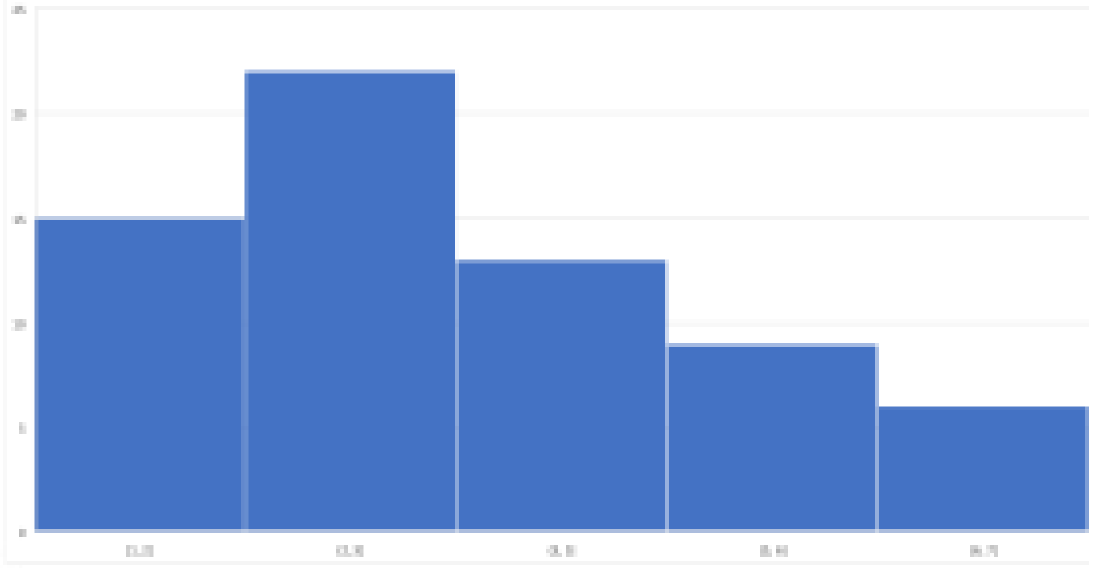


Figure 50. Left skewed distribution.

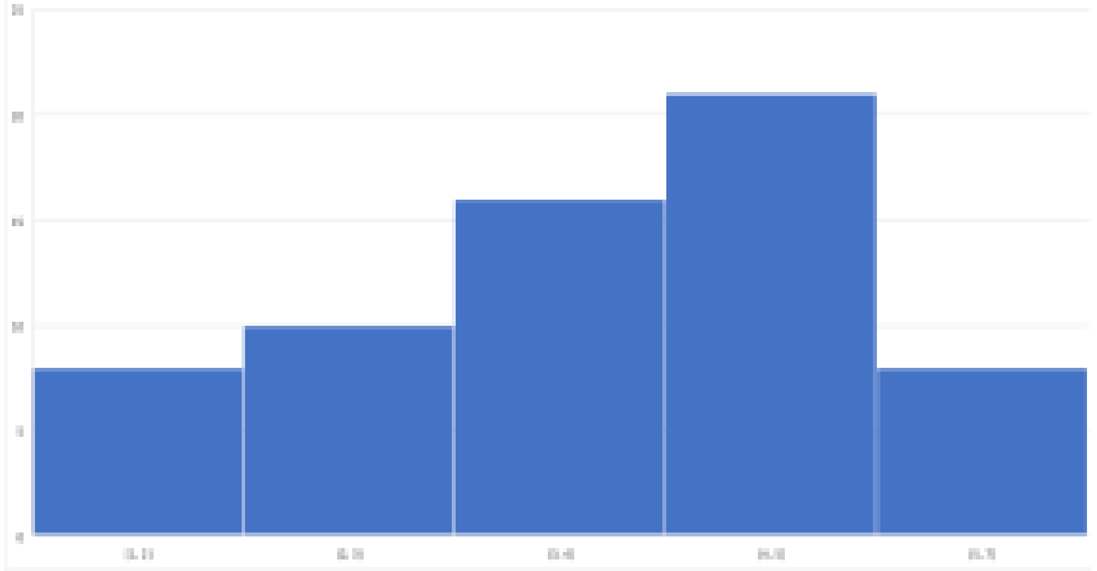
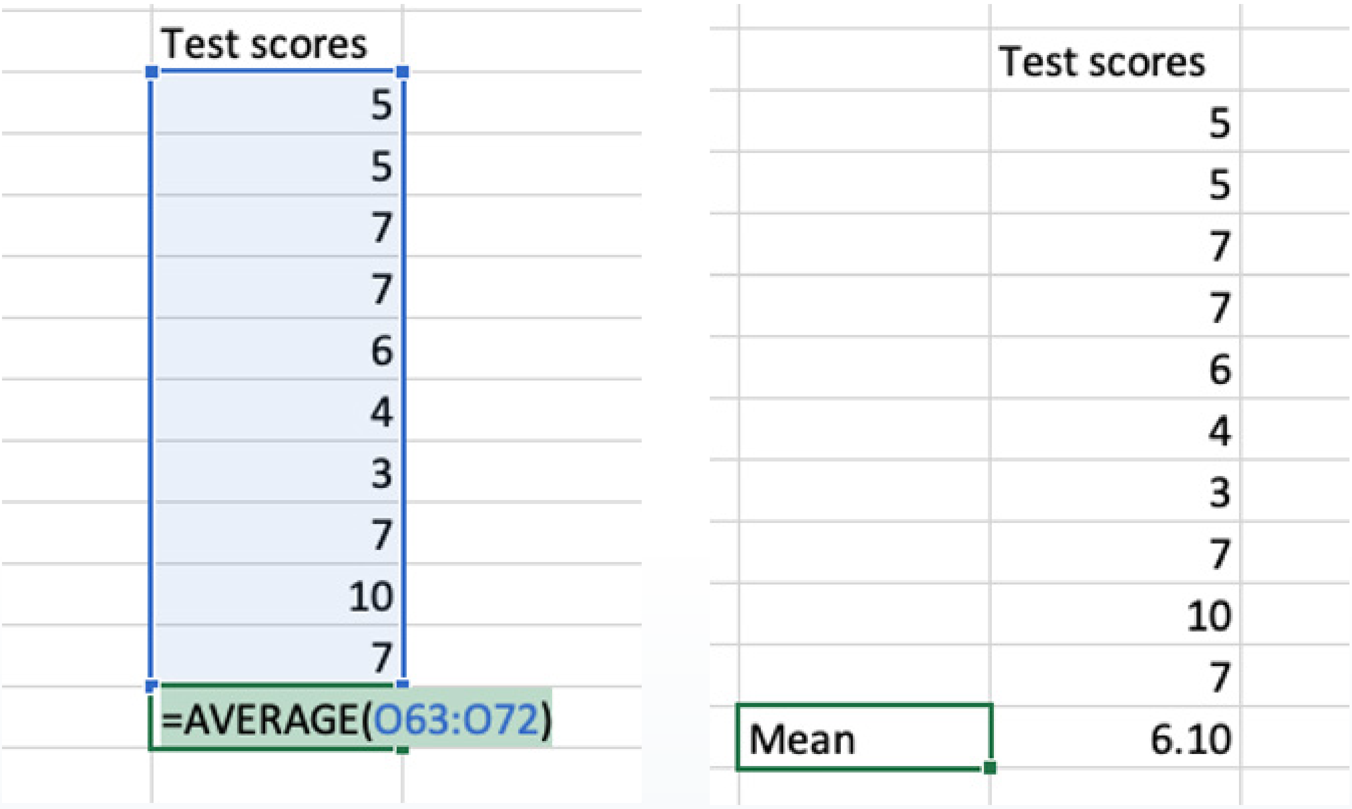


Figure 51. Right skewed distribution

Mean

The mean is what most people are referring to when they talk about the average. A simple example would be calculating the mean score on a test paper. To calculate the mean, all the test scores would be added together and then divided by the number of people who took the test. Excel allows this to be done by typing in =AVERAGE(number1, (number2),…)) and then selecting the data range, in this case, O63:O72. (Figure 52) which is 6.10. The function can also be accessed via the auto-sum function.



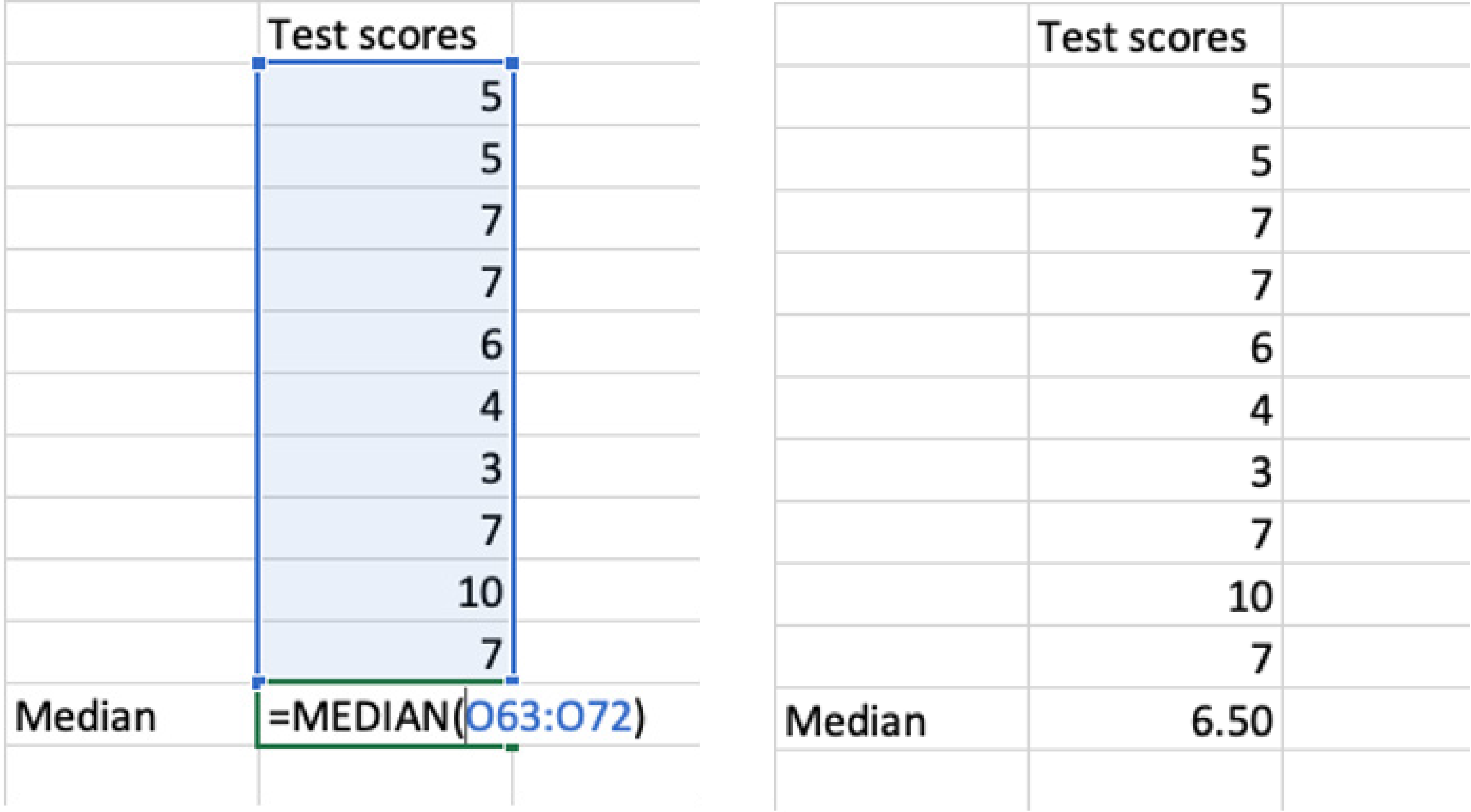
Test scores
5
5
7
7
6
4
3
7
10
7
=AVERAGE(O63:O72)

	Test scores
	5
	5
	7
	7
	6
	4
	3
	7
	10
	7
Mean	6.10

Figure 52. Calculating the mean

Median

This is the number that falls in the centre of a data set, with 50% of the data being lower and 50% of the data being higher than this point. If the data shown in Figure 52 is arranged in number order, the data will look like this: 3, 4, 5, 5, 6, 7, 7, 7, 7, 10. The highlighted area shows the central point, where there is the same number of entries on both sides. Therefore, the median would be midway between these numbers, at 6.5. Once again, excel allows this to be easily calculated using =MEDIAN((number1, (number2),…) (Figure 53).



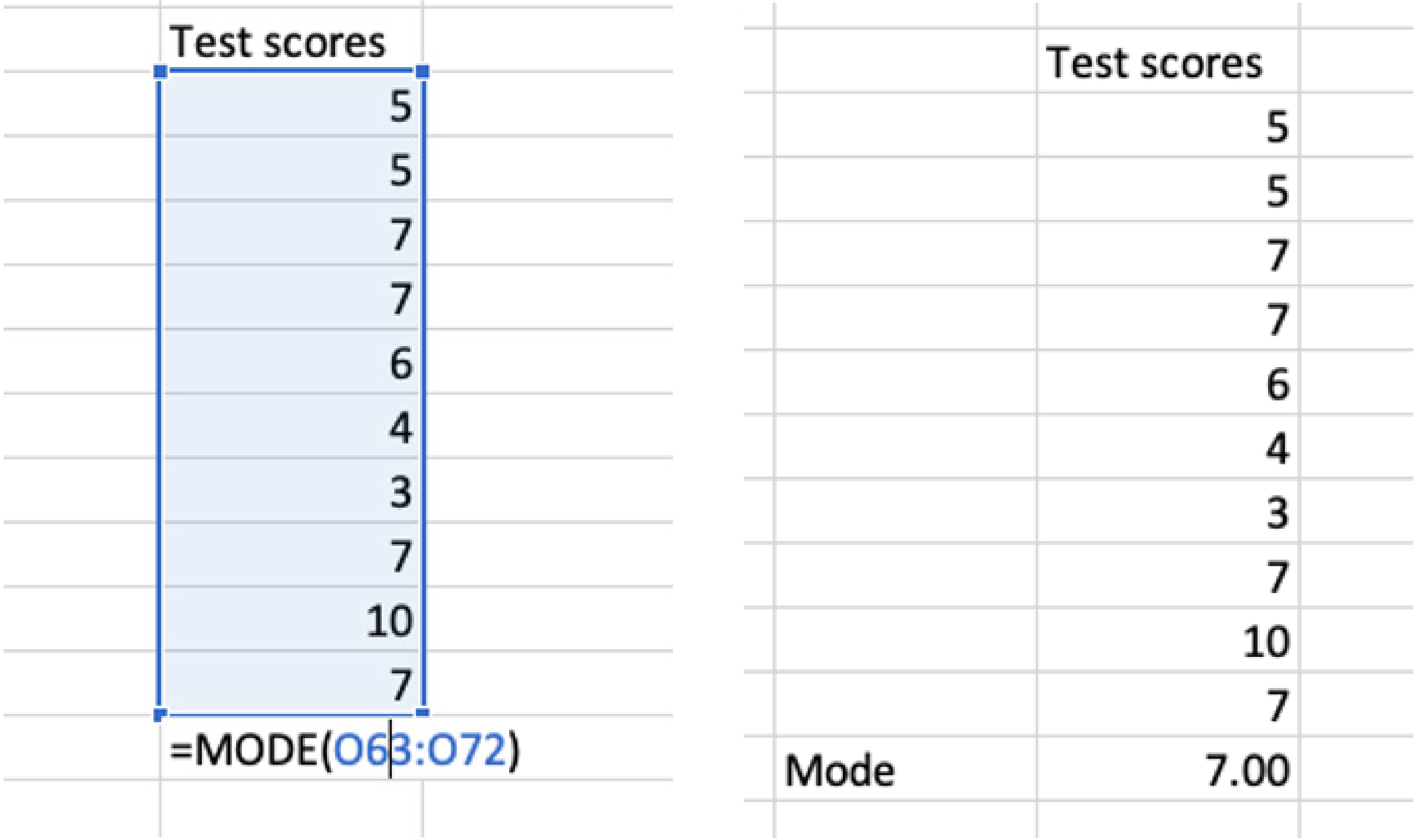
	Test scores
	5
	5
	7
	7
	6
	4
	3
	7
	10
	7
Median	=MEDIAN(O63:O72)

	Test scores
	5
	5
	7
	7
	6
	4
	3
	7
	10
	7
Median	6.50

Figure 53. Calculating the median

Mode

This term refers to the most common value in a data set. Excel calculates this by using =MODE(number1, (number2),…). The mode for the data set shown in Figure 54 is 7, which occurs 4 times.



Test scores
5
5
7
7
6
4
3
7
10
7
=MODE(O63:O72)

	Test scores
	5
	5
	7
	7
	6
	4
	3
	7
	10
	7
Mode	7.00

Figure 54. Calculating the mode

Nominal Data

This is where a number is used to identify something, for example, the number of a football shirt identifies the player. In evaluation, this type of data is often used when coding groups, by aspects such as gender, ethnicity, etc.

One-sided or two-sided tests

Some tests can be performed as one- or two-tailed tests. A one-tail test looks for a difference in a specific direction, whereas two-tailed tests look for differences at both ends of a distribution. For this reason, it is usually better to use the two-tailed version of a test and report these results. More information on one and two-tailed tests can be found at: https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

Ordinal data

Ordinal data is where it is known that something is bigger or smaller, but the difference is not clearly defined. For example, a scale that used: much worse, slightly worse, no change, slightly improved, or much improved could be considered as ordinal data.

Outliners

A data point that is an abnormal distance from the other values.

P-values

This is a term used to describe the probability that a result is real and not simply a chance occurrence. The lower the p-value the lower the probability that the change is due to chance. The accepted level (called the alpha value) for a difference to be seen as being significant is a score at or below .05. This number means there is a 95% probability that any difference found is real and not down to chance. It should be noted that this does not mean it could not have occurred by chance, in fact, there is still a 1 in 20 chance it may have done at p = .05.

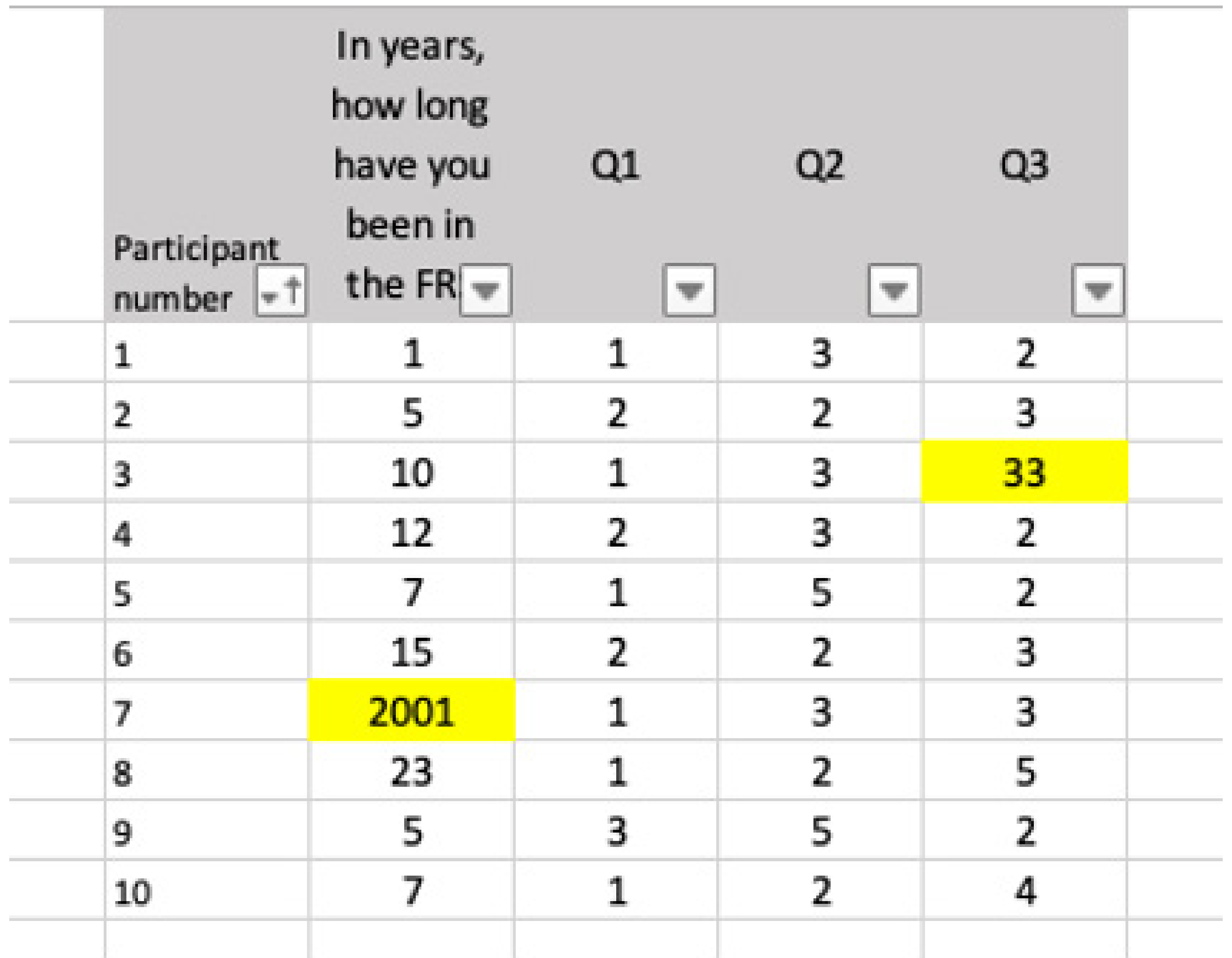
Reliability

This term covers the following concepts:

> **Measure reliability** - Measures should provide a consistent result, for example, two people with the same level of IQ should score the same on an IQ test. The measure should provide consistency, for example, if the same person completes an IQ test and then takes the same test a week later, it should provide a very similar result, unless they have done something to improve their IQ. Measure reliability is complex and requires an understanding of a range of statistical tests that fall outside of the scope of this document. Further information on this type of testing can be accessed at: https://www.simplypsychology.org/reliability.html.
>
> **Data reliability** - Is the data set complete, free from duplication and accurate?

Under the concept of data reliability, before any analysis is started, it is necessary to check that the data is accurate. This is particularly important if the data has been coded manually. This is often done by cross-checking a sample of the data usually between 5% and 10%. As well as this check, a visual check of the data should be made for input errors.

If the data set is small, a check can be carried out visually, but this can be very time-consuming on large data sets. On larger data sets it is worth filtering each column in turn, to see if there are any inputs that fall outside of what would be expected. Figure 55 shows some common anomalies (highlighted in yellow). Someone could not have been in the FRS for 2001 years. The maximum possible score in Q3 is 5, as the question uses a 5-point scale, therefore it is not possible to have a score of 33. Once identified, incorrect data can either be removed or checked at the data source and corrected.



Participant number	In years, how long have you been in the FR	Q1	Q2	Q3
1	1	1	3	2
2	5	2	2	3
3	10	1	3	33
4	12	2	3	2
5	7	1	5	2
6	15	2	2	3
7	2001	1	3	3
8	23	1	2	5
9	5	3	5	2
10	7	1	2	4

Figure 55. Reviewing data for errors and anomalies

As well as the checks outlined above, the evaluator should look at how the data is distributed as this will help not only to identify potential errors but also inform the type of statistical tests that should be undertaken.

Scale data

Scale data has known and equal increments. For example, the difference between 1o and 2o centigrade is the same as the difference between 2o and 3o centigrade, etc.

Standard Deviation

A Standard Deviation (SD) is a measure of how spread the data is in relation to the mean. The smaller the SD the more tightly the data is clustered around the mean. A standard distribution has known properties, with 68.26% of scores falling within one SD (+ or – of the mean) and 95.44% will fall into 2 SD from the mean (Figure 56). The areas that are more than two SD (the unshaded areas if Figure 56) are referred to as the tails of the distribution. The likelihood of a result falling into the tails is low. For example, if this were representative of IQ scores, someone would need to achieve a very high score to fall into the tail on the right or a very low score to fall into the tail on the left.

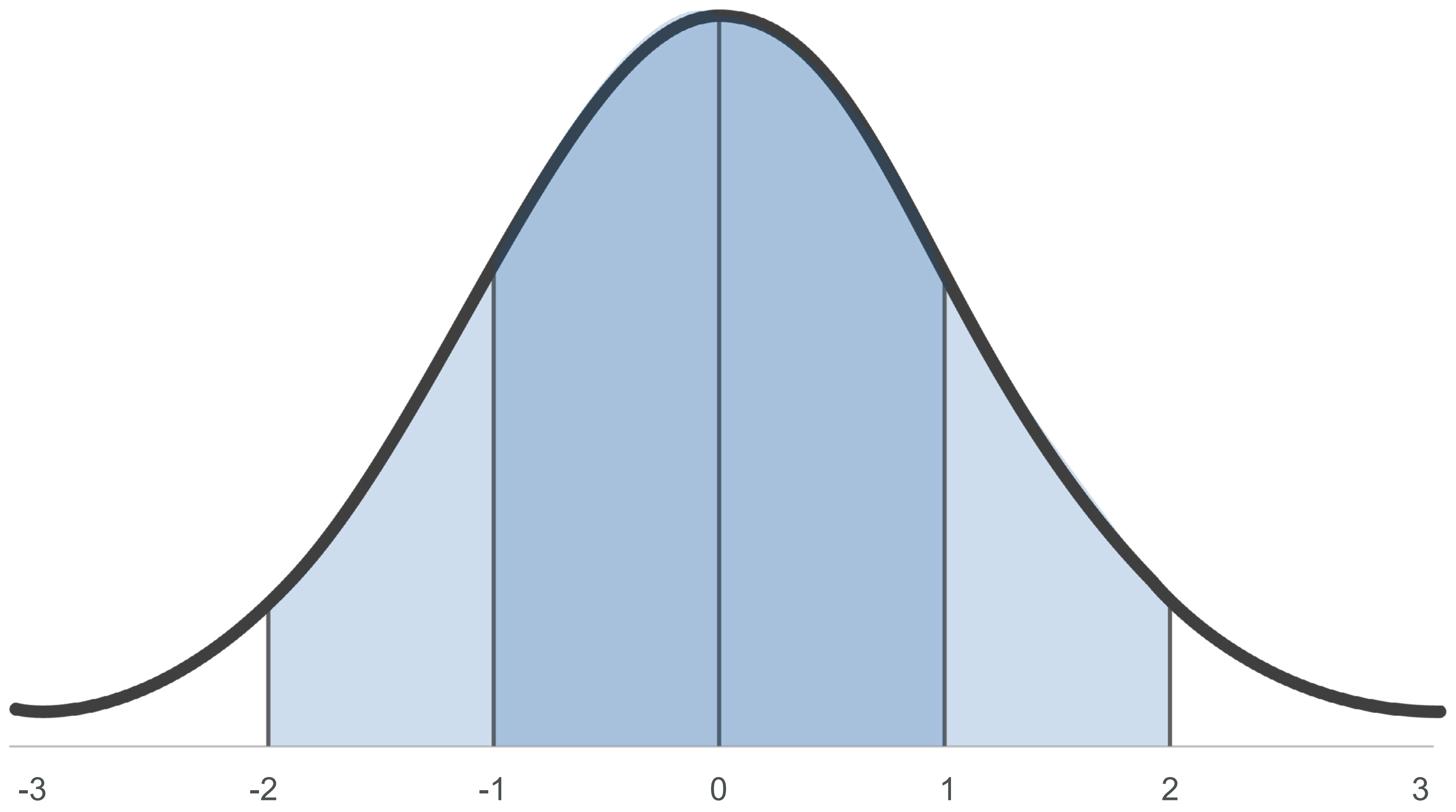
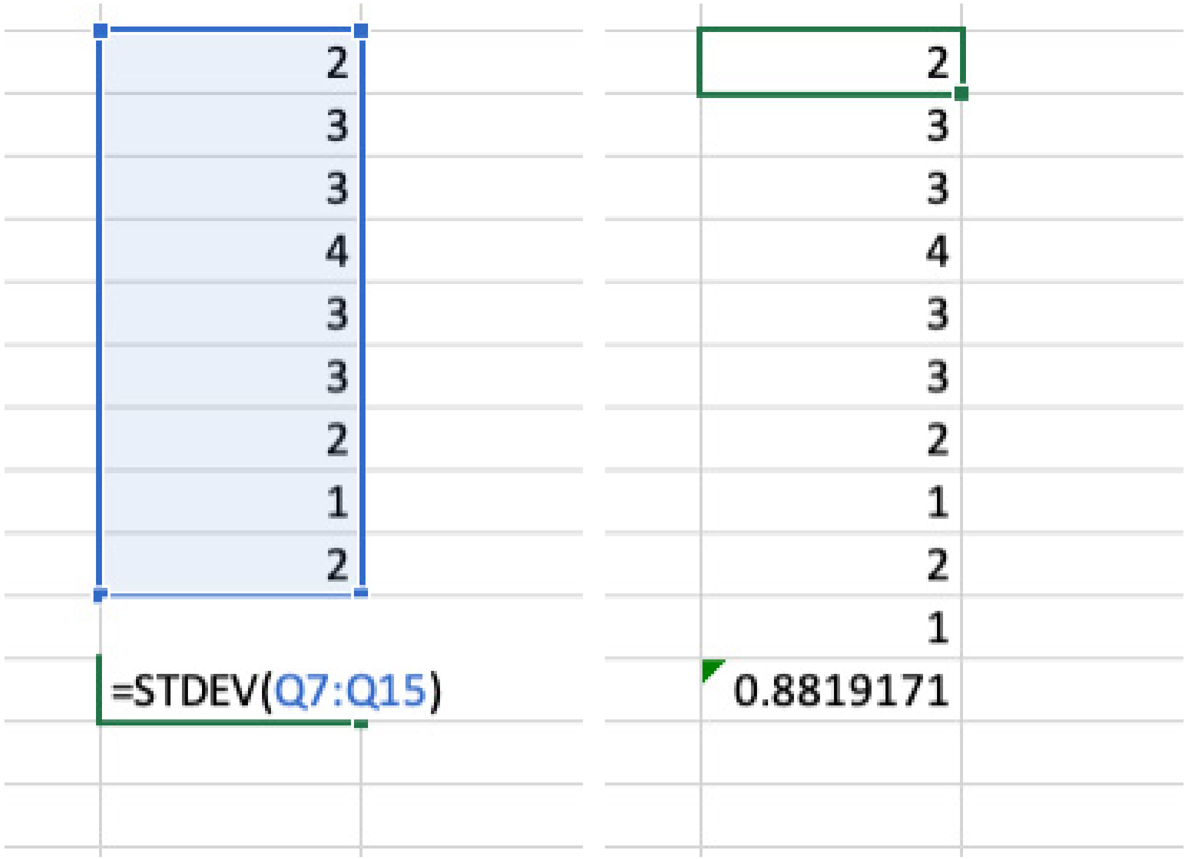


Figure 56. Standard deviation and the tails of the distribution



2	2
3	3
3	3
4	4
3	3
3	3
2	2
1	1
2	2
	1
=STDEV(Q7:Q15)	0.8819171

Figure 57. Standard Deviation calculation

To work out a SD in Excel, select a cell and input =STDRV and then enter the range of data of interest and press return (Figure 57).

Statistical Significance

Statistical significance is a term used to describe the probability that the result is real and not simply a chance occurrence. This is expressed as a p-value, the lower the value, the lower the probability that the change is due to chance. The accepted level (called the alpha value) for a result to be seen as being significant is a score at or below .05. This number means there is a 95% probability that any difference found is real and not down to chance. It should be noted that this does not mean it could not have occurred by chance; there is still a 1 in 20 chance it may have with a p-value of .05.

Trustworthiness

A qualitative term referring to the quality and truthfulness of the findings.

T-tests

These are a set of parametric tests that compares the distribution to identify if there is a significant difference. A t-value is used to calculate the p-value.

Validity

Validity is a complex concept but some of the main points that need to be considered include:

Measure validity - Is the measure(s) used to produce the data measuring the correct concept? Are questionnaires accurate measures of behaviour?

Internal validity - Can any causal relationship be drawn from the results? For example, is the decrease in young driver casualties attributed to better driver education programmes or improvements in vehicle design?

External validity – Can the results be generalised to the wider population? For example, if the data is drawn from a single fire station would it be safe to say the finding would apply to all fire stations across the UK?

Validity testing is complex, but there is a range of statistical tests that can be used. More information about how to validate a scale can be found at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/

Variance

Variance is a measure of the dispersion of all the data points in a data set. The lower the variance the more the data is clustered around the mean. A Standard Deviation is a square root of the variance.

Z-scores

A measure of how far the data is from the mean.

Useful Links

Analysis of Variance (ANOVA)
https://www.scribbr.com/statistics/one-way-anova/

Cohen's Kappa test
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Content Analysis
https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis

Discourse Analysis
https://research.ncl.ac.uk/methodshub/methods/discourseanalysis/

Excel Data Analysis ToolPak
https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

Grounded Theory
https://www.personal.psu.edu/wxh139/grounded.htm

Measure reliability
https://www.simplypsychology.org/reliability.html

Narrative Analysis
https://www.tandfonline.com/doi/full/10.1080/21642850.2018.1515017

One-sided and two-sided tests
https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

Pivot tables
https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576

Trustworthiness in Qualitative research
https://journals.sagepub.com/doi/pdf/10.1177/1609406917733847

Validity testing
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/

Test Formulas

Kruskal-Wallis oneway analysis of variance

$$H = (N - 1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{r}_{ij} - \bar{r})^2}$$

Where:

- N is the total number of observations across all groups
- g is the number of groups
- n_i is the number of observations in group *i*
- r_{ij} is the rank (among all observations) of observation ***j*** from group ***i***
- $\bar{r}_{i\cdot} = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ is the average rank of all observations in group ***i***
- $\bar{r} = \frac{1}{2}(N + 1)$ is the average of all the r_{ij}

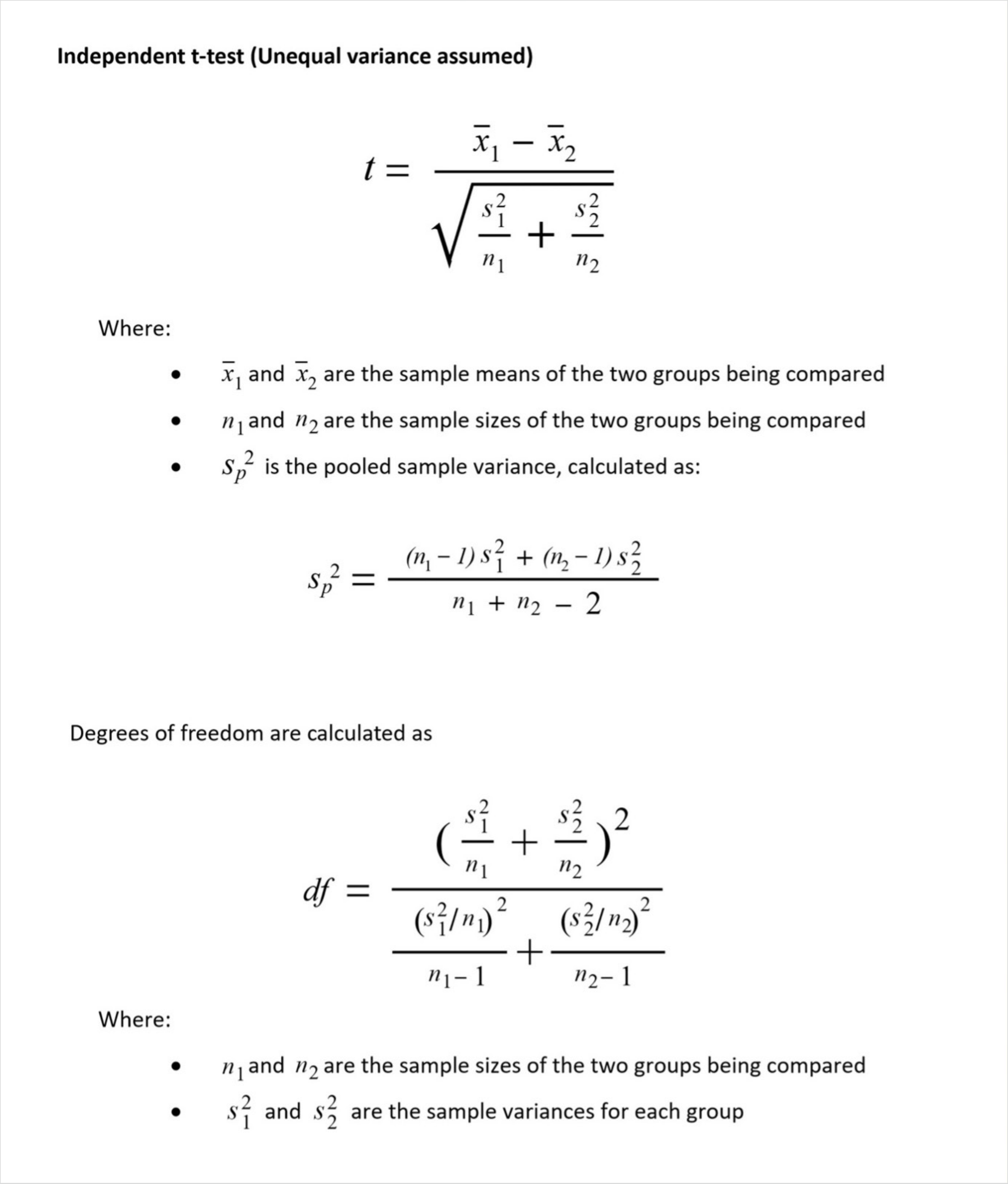
Pearson Correlation Coefficient test

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- r is the correlation coefficient
- x_i is the values of the x-variable in a sample
- $\bar{x}$ is the mean of the values of the x-variable
- y_i is the values of the y-variable in a sample
- $\bar{y}$ is the mean of the values of the y-variable

Test Formulas



Independent t-test (Unequal variance assumed)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups being compared
- n_1 and n_2 are the sample sizes of the two groups being compared
- s_p^2 is the pooled sample variance, calculated as:

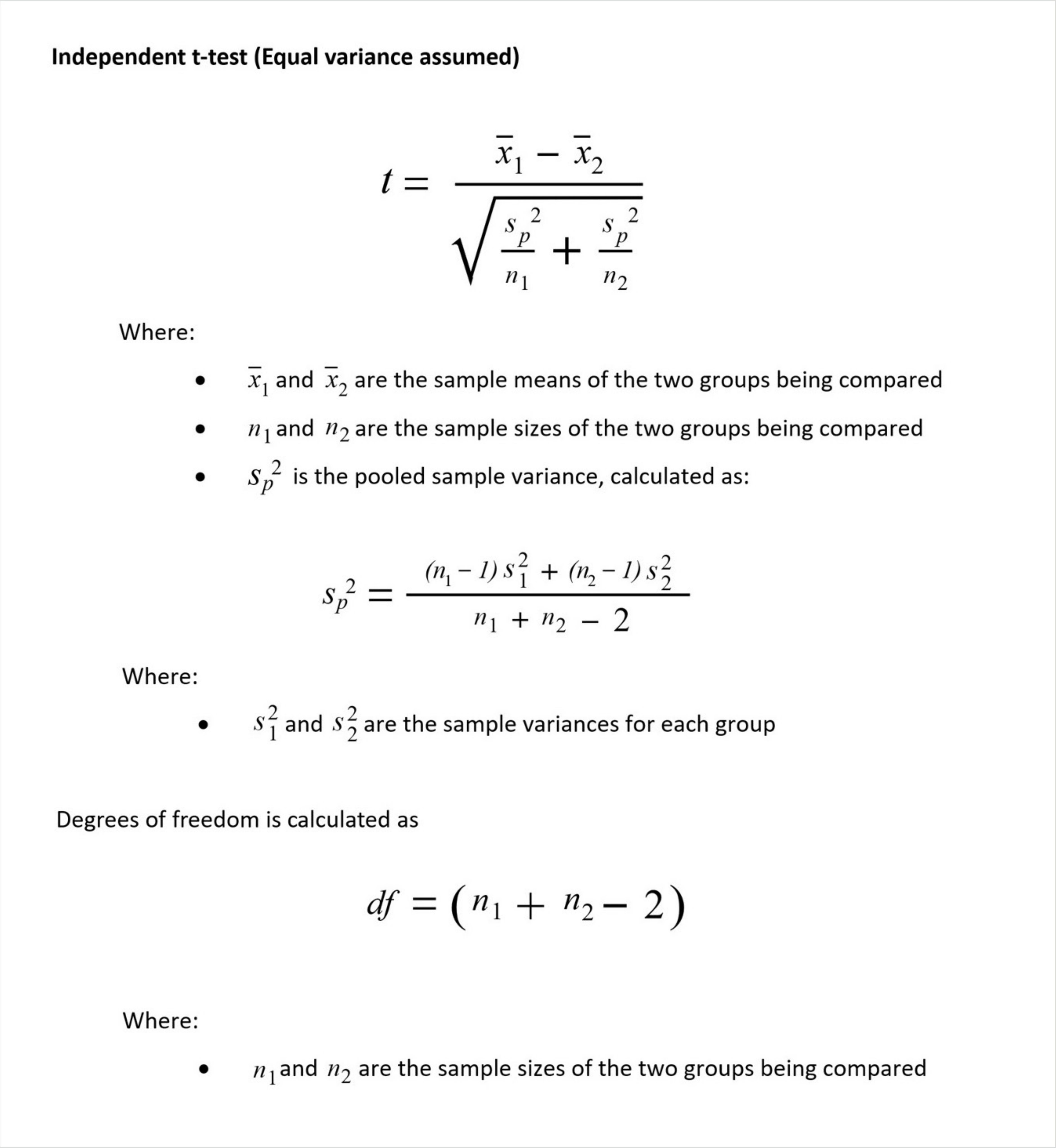
$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Degrees of freedom are calculated as

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Where:

- n_1 and n_2 are the sample sizes of the two groups being compared
- s_1^2 and s_2^2 are the sample variances for each group



Independent t-test (Equal variance assumed)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the two groups being compared
- n_1 and n_2 are the sample sizes of the two groups being compared
- s_p^2 is the pooled sample variance, calculated as:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Where:

- s_1^2 and s_2^2 are the sample variances for each group

Degrees of freedom is calculated as

$$df = (n_1 + n_2 - 2)$$

Where:

- n_1 and n_2 are the sample sizes of the two groups being compared

Test Formulas

Mann-Whitney U test

$$U = n_1 n_2 + \frac{n_2 \left(n_2 + 1 \right)}{2} - \sum_{i = n_1+1}^{n_2} R_{\mathrm{i}}$$

Where:

- U is Mann-Whitney U test
- n_1 is sample size one
- n_2 is sample size two
- R_{i} is the rank of the sample size

Wilcox signed-rank test

$$W = \sum_{i=1}^{N_{\mathrm{r}}} \left[\operatorname{sgn} \left(x_{2,i} - x_{1,i} \right) \cdot R_i \right]$$

Where:

- W is the test statistic
- N_{r} is the sample size, excluding pairs where x1 = x2
- sgn is sign function
- $x_{1,i}, x_{2,i}$ are corresponding ranked pairs from two distributions
- R_i is rank i

Paired sample t-test

The test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

Where:

- $\bar{d}$ is the mean difference
- s^2 is the sample variance
- n is the sample size
- t is a Student t quantile with n -1 degrees of freedom

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

Field, A. (2018). Discovering statistics using IBM SPSS statistics 5th ed.

Langdridge, D. (2004). Fundamentals of qualitative analysis. Introduction to research methods and data analysis in psychology, 249260.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. International journal of qualitative methods, 16(1), 1609406917733847.

Riessman, C. K. (2008). Narrative methods for the human sciences. Sage.