Statistical Inference for Data Science using SAS Studio

Contents

Where Do We Start?

The first thing we need to think about is what question we are being asked to answer, and how we can describe the data that we're working with.

Don't worry about whether it's parametric or non-parametric just yet, let's just consider:

(i) How many variables are there?

(ii) What are we comparing?

(iii) Are we looking for an association/relationship?

These small questions will help determine where to go from here.

ONE VARIABLE

There is only one set of tests for this case. Turn to page 4 (or 7 if your data is ranked/scored).

TWO VARIABLES

There are quite a few cases to consider here.

CASE 1: - If you've been asked to or want to determine if there's an ASSOCIATION or RELATIONSHIP between the two variables. Turn to page 24 (or page 29 if data is ranked/scored).

CASE 2: - If both samples are drawn INDEPENDENTLY (that is, a response from one sample won't affect the other, there are different participants/ subjects in one group to the other). Turn to page 9 (or 14 for ranked/scored data).

CASE 3: - If both samples are from the SAME SUBJECTS WITH THE SAME RESPONSE (that is, the same group of participants/subjects give a response/result for two different tests, results are matched/paired together for comparison). Turn to page 17 (or page 21 if data ranked/scored).

THREE OR MORE VARIABLES

There is only one set of tests for this case. Turn to page 32 (or 38 if data ranked/scored).

Once you've made a decision, read the next page on preliminary analysis and then move forward to the specified page you need.

Preliminary Analysis

Now that we know what type of test we want, it's time to consider whether the data meets the assumptions for using parametric methods and if not, using a non-parametric approach.

The absolute FIRST thing to think about which can save a ton of wasted time is whether the data is Ordinal or not. That is, is the data in the form of ranks or scores?

- **If the data is RANKED or in the form of SCORES this is NON-PARAMETRIC and assumed DISTRIBUTION FREE. You can skip straight to the non-parametric part of your specified section after this page, saving you time and confusion!**

- **If the data is NOT in a rank or score form (and this isn't mentioned anywhere in the brief/question) then we assume it's PARAMETRIC until proved otherwise through testing the assumptions. Go straight to the Parametric part of your specified section after this page if this is the case with you!**

The next thing we must consider is whether we require a ONE-TAILED or TWO-TAILED test. This will normally be specified in the context of the question and greatly affects the hypotheses that we will specify to test.

- **Are you being asked if something is more or less? Is something better than a given value or worse than a given value? Is one case better than another? Is one method better/ worse than another for the same participants? Is there a positive/negative association between variables etc. If so, this can be specified by a direction (< or >) and so is one-tailed.**
- **Are you just being asked is there is a difference between a hypothetical value and your data? Does a difference exist between two variables? Does a relationship exist between two variables etc. If so, this doesn't specify a direction (≠) and so is two-tailed.**

One Sample Testing: Parametric

Now let's get to the meat of the investigation and testing. The question we are looking to answer is:

"Is the Population Mean μ=c?"

Where c is a prespecified constant (it's the HYPOTHETICAL VALUE we are comparing our one variable too) and μ is the POPULATION MEAN for that same variable.

The Parametric approach for this case is the ONE-SAMPLE T-TEST and it has the following HYPOTHESES:

ONE-TAILED

H_0: μ=c

H_1: μ<c OR μ>c (Dependent on the context)

TWO-TAILED

H_0: μ=c

H_1: $\mu \neq$c (Where ≠ is the mathematical notation for NOT EQUAL TO)

The most important preliminary step before carrying out tests and analysing outputs in SAS Studio is to:

CHECK THE ASSUMPTIONS FOR A PARAMETRIC APPROACH!!

This is so important because:

IF THE DATA DOES NOT MEET THE ASSUMPTIONS, WE USE THE NON-PARAMETRIC APPROACH INSTEAD!!

This test has the following ASSUMPTIONS:

- **Sample size n is small (n<30).**
- **The Distribution of the population from which the sample is drawn is NORMALLY DISTRIBUTED. This is based on the HYPOTHESES:**

  H_0: There is no difference between the distribution of the variable and that of the Normal Distribution;

  H_1: There is a difference between the distribution of the variable and that of the Normal Distribution.

  So of course, to satisfy this assumption we want a P-value>α so that we DO NOT REJECT H_0.

Thankfully, the sample size can be checked in seconds by just looking at your data. The Normal assumption is tested in SAS and is part of the overall output.

IN SAS STUDIO:

1. **In TASKS, select T-TESTS.**
2. **Select your dataset and under ROLES choose One-sample test and an ANALYSIS variable (i.e. the variable that the mean will be calculated for and compared for the test). If you have correctly identified your data is one variable, there should be only one of these.**
3. **Under OPTIONS:**

   TAILS – choose based on your hypotheses

   NORMALITY ASSUMPTION – This must be ticked as it's so important.

   NONPARAMETRIC TESTS – For now let's ignore this, I'll mention why this is useful in the next section. It's worth noting however that if you did tick this, both tests get run at the same time, so you can rule out one and use the other! However, this produces a lot of output which is convoluted to interpret and sort through, which is why I split it up here.
4. **Under PLOTS, choose SELECTED PLOTS:**

   HISTORGRAM AND BOX PLOT – Will be ticked

   NORMALITY PLOT – Will be ticked

   CONFIDENCE INTERVAL PLOT – I would tick this as it's very useful to make sure it corroborates (agrees with) your P-value conclusion. The way it works is if the Confidence Interval (marked with two lines and

shaded a green-blue sort of colour) contains the value we are testing for (c in this case, would be 0 a lot of the time) , then 1- α of the time (95% usually as we tend to choose significance level 5% most often) the true value we are testing for will be in that confidence interval and we CANNOT be sure that the value ISNT c. So therefore, we can't reject the Null hypothesis if that's the case. However, if our value c is outside of the confidence interval, then there is evidence to reject the null hypothesis.

5. **Select the SAS RUN option (A running man icon).**

TIPS FOR ANALYSIS OF OUTPUT

1. **The first output to check will be the Test for Normality. Thankfully a lot of the tables that SAS runs will tell you in the labels and headings what the output is.**
   SAS runs a few different tests and compiles the P-values together. Thankfully whichever (one or two tailed) type of test you chose is considered and adjusted for in the output, so you just need to compare the P-values to α.
   To support the conclusion from this, use the Histogram and the QQ plot. QQ plots tell you if quantiles of your data are similar to the quantiles of a Standard Normal Distribution (whichever distribution you're testing for of course, usually normal), if all points are following that similar line then we likely have the same distribution.
   IF THERE IS EVIDENCE TO REJECT H_0, ASSUME NON-PARAMETRIC AND GO TO THE NEXT SECTION.
2. **The next output is summary statistics that may prove useful when interpreting the graphs or making a more informed conclusion.**
3. **Beneath these statistics is a small table with headings DF, t VALUE and PR>|t|. This is the table from which we draw our conclusions after assuring the assumptions are satisfied. Compare this P-value to α and draw your conclusion from there. It would be useful to discuss the confidence interval plot here too.**
4. **State the conclusion in the context of the problem.**

One Sample Testing: Non-Parametric

If the data you have is in the form of ranks/scores, or your data has failed the test for Normality in your SAS output you assume the data is DISTRIBUTION FREE and now we use a NON-PARAMETRIC APPROACH.

The question we're looking to answer here is:

"Is the Population Median m=c?"

Where c is a prespecified constant and m is the Population median. The reason this approach uses the median rather than the mean is due to the data being non-parametric. It does not follow a distribution and as a result could be highly skewed, therefore the mean will also be skewed and wouldn't be useful to draw conclusions from.

The Non-Parametric approach in this case is the WILCOXON ONE SAMPLE SIGNED RANKS TEST or the ONE SAMPLE SIGN TEST. Both are run in the SAS output. These methods have the following hypotheses:

ONE TAILED:

H_0: m=c

H_1: m>c or m<c

TWO TAILED:

H_0: m=c

H_1: m≠c

Thankfully with this test there are no assumptions to check and we can proceed with the SAS Studio.

IN SAS STUDIO

1. **In TASKS, select T-TESTS.**
2. **Select your dataset and under ROLES choose One-sample test and an ANALYSIS variable (i.e. the variable that the mean will be calculated for**

and compared for the test). If you have correctly identified your data is one variable, there should be only one of these.

3. **Under OPTIONS:**
   TAILS – choose based on your hypotheses
   NORMALITY ASSUMPTION – Since your data is ranked/scored and distribution free, you can untick this.
   NONPARAMETRIC TESTS – Now we must tick this as it will run both of the one sample tests and calculate the associated P-values.
4. **Under PLOTS, choose SELECTED PLOTS:**
   HISTORGRAM AND BOX PLOT – Will be ticked
   NORMALITY PLOT – I would tick this because it can be used to further clarify that the data is not normal and thus distribution free.
   CONFIDENCE INTERVAL PLOT – Unlike in the parametric case this is not useful for the non-parametric output as it runs a confidence interval on the mean of the data and not the median which we are testing here.
5. **Select the SAS RUN option (A running man icon).**

TIPS FOR ANALYSIS OF OUTPUT

1. **Summary statistics and a t-value will be printed first. Ignore the t-value as the data is non-parametric, the summary statistics could be used to further discuss the data if necessary.**
2. **The plots are printed next. The histogram and QQ plot can be used to support the conclusion that the data is distribution free and thus this approach was correct. I would only note this if your data failed the test of normality, however. If your data is ranked/scored, there is no need for this output to be examined. You can use the boxplot to also further comment on the skew of the data and comment on where the median is if this is useful.**
3. **The last output is a table of P-values, one for the t-statistic (ignore as it's not useful here, even if it agrees with the others), one for the Wilcoxon One Sample Signed Ranks Test and one for the One Sample Sign Test. Compare both P-values to α and draw conclusions.**
4. **State conclusion(s) in the context of the problem.**

Two Independent Sample Testing: Parametric

The question we are looking to answer in this case is:

"Do the Population Means μ_1 and μ_2 differ significantly?"

Where μ_1 is the POPULATION MEAN for the first sample and μ_2 is the POPULATION MEAN for the second sample.

The Parametric approach for two independent sample testing is the TWO SAMPLE T-TEST. This has the HYPOTHESES:

ONE TAILED

H_0: $\mu_1=\mu_2$

H_1: $\mu_1>\mu_2$ or $\mu_1<\mu_2$

TWO TAILED

H_0: $\mu_1=\mu_2$

H_1: $\mu_1\neq\mu_2$

The most important preliminary step before carrying out tests and analysing outputs in SAS Studio is to:

CHECK THE ASSUMPTIONS FOR A PARAMETRIC APPROACH!!

This is so important because:

IF THE DATA DOES NOT MEET THE ASSUMPTIONS, WE USE THE NON-PARAMETRIC APPROACH INSTEAD!!

The test has the following ASSUMPTIONS:

- **Both samples are drawn independently**
- **The Distribution of the population from which the samples are drawn is NORMALLY DISTRIBUTED. This is based on the HYPOTHESES:**
  H_0: There is no difference between the distribution of the variables and that of the Normal Distribution;

H_1: There is a difference between the distribution of the variables and that of the Normal Distribution.

So of course, to satisfy this assumption we want a P-value>α so that we DO NOT REJECT H_0.

- **The Variances of each group are equal, i.e. $\sigma^2_1=\sigma^2_2$ (Where σ^2_i is the population variance for group i). This is known as HOMOGENEITY OF VARIANCES.**

  This is based on the HYPOTHESES:

  H_0: $\sigma^2_1=\sigma^2_2$

  H_1: $\sigma^2_1\neq\sigma^2_2$

If you've chosen this method, it's safe to say that the first assumption is satisfied (remember, each group should contain different subjects, their responses aren't matched and have no effect on each other!).
The Normality assumption is tested for in the SAS output as well as the equality of variances as well as what to use if they are unequal.

<u>IN SAS STUDIO</u>

1. **In TASKS select T-TESTS**
2. **Select your dataset and under ROLES choose Two-sample test and an ANALYSIS variable (i.e. the variable that the mean will be calculated for and compared for the test) and a GROUPS variable (i.e. each sample, could be two countries, two nations, the overall group that the sample was drawn from).**
3. **Under OPTIONS:**

   TAILS: Choose based on your hypotheses.

   ALTERNATIVE HYPOTHESIS: This is where we set what the value that we are testing for is. Of course, since we have H_0: $\mu_1=\mu_2$ then a simple rearrangement gives $\mu_1-\mu_2=0$ and thus the value to input would be 0. This is the same for all of the tests in this section as we just want to know if the means are the same or not, which would mean the difference between them is 0.

COX AND COCHRAN PROBABILITY APPROXIMATION FOR UNEQUAL VARIANCE: Ignore this as we don't need to make use of it regardless if our variances are found not equal as there's another method in the original output we will get.

NON-PARAMETRIC TESTS: Ignore this for now, this will be explained in more detail in the next section. It's worth noting however that if you did tick this, both tests get run at the same time, so you can rule out one and use the other! However, this produces a lot of output which is convoluted to interpret and sort through, which is why I split it up here.

4. **Under PLOTS, choose SELECTED PLOTS**

   HISTOGRAM AND BOX PLOT – Ticked already

   NORMALITY PLOT – Ticked as it's very useful

   CONFIDENCE INTERVAL PLOT - I would tick this as it's very useful to make sure it corroborates (agrees with) your P-value conclusion. The way it works is if the Confidence Interval (marked with two lines and shaded a green-blue sort of colour) contains the value we are testing for (c in this case, would be 0 a lot of the time) , then 1- α of the time (95% usually as we tend to choose significance level 5% most often) the true value we are testing for will be in that confidence interval and we CANNOT be sure that the value ISNT c. So therefore, we can't reject the Null hypothesis if that's the case. However, if our value c is outside of the confidence interval, then there is evidence to reject the null hypothesis.

5. **Use the SAS RUN option (An icon of a running man) to run the test.**

TIPS FOR ANALYSIS OF OUTPUT

1. **The first output to check will be the Test for Normality. Thankfully a lot of the tables that SAS runs will tell you in the labels and headings what the output is. In this case it runs a test for each of the two variables so we will have two output tables. THE ONLY WAY WE CAN CONCLUDE THE DATA IS NORMAL IS IF BOTH VARIABLES CAUSE US TO ACCEPT H_0. ANY OTHER CASE IS ASSUMED NON-PARAMETRIC.**

SAS runs a few different tests and compiles the P-values together. Thankfully whichever (one or two tailed) type of test you chose is considered and adjusted for in the output, so you just need to compare the P-values to α.

To support the conclusion from this, use the Histogram and the QQ plot. QQ plots tell you if quantiles of your data are similar to the quantiles of a Standard Normal Distribution (whichever distribution you're testing for of course, usually normal), if all points are following that similar line then we likely have the same distribution.

IF THERE IS EVIDENCE TO REJECT H_0, ASSUME NON-PARAMETRIC AND GO TO THE NEXT SECTION.

2. **If the data was found to be distributed normally, we must now analyse the assumption of Homogeneity of Variances. Scroll down the output for a table with heading "Equality of Variances" where they calculated an F-statistic which led to a corresponding F-value which can be compared to α to see if there's a difference in the variances or not. Now if the variances are found EQUAL, then from now on refer to the POOLED estimate of the difference in variances row in any test output (this is taken during the test to make comparisons easier).**

   If the variances are found UNEQUAL, then from now on refer to the SATTERTHWAITE row in any test output. Keep referring back to this so that you don't confuse yourself and read the incorrect output.

3. **Now if we scroll up there's a descriptive statistics table that could be helpful with analysing plots if necessary. What's of most interest to us is the table involving the 95% Confidence Interval. Refer to the intervals calculated for the "Diff (1-2)" variables. Remember, EQUAL ⇒ POOLED, UNEQUAL ⇒ SATTERTHWAITE (where ⇒ means "implies"). Interpret this Confidence Interval (Remember, does it include 0, the value we're testing for? If so, we can't reject H_0, if it does, then we will reject H_0). This will be very helpful as our P-value conclusion should agree with this finding.**

4. **Now just below the Confidence Interval table there is a table of t values and associated P-values. Interpret whichever is appropriate for what the outcome for the Variances test and draw your conclusion.**
5. **State the conclusion(s) in the context of the problem.**

Two Independent Sample Testing: Non-Parametric

If the data you have is in the form of ranks/scores, or your data has failed the test for Normality in your SAS output you assume the data is DISTRIBUTION FREE and now we use a NON-PARAMETRIC APPROACH.

The question we're looking to answer here is:

"Do the PDFs of the samples differ significantly?"

Where the PDF is the Probability Density Function of Continuous Random Variables (The graph of this function describes the distribution, you'd be surprised how gross the PDF of the normal distribution is!).

The Non-Parametric approach in this case is the WILCOXON RANK-SUM TEST.

This method has the following hypotheses:

ONE TAILED:

H_0: $f(x_1)=f(x_2)$

H_1: $f(x_1)>f(x_2)$ or $f(x_1)<(x_2)$

TWO TAILED:

H_0: $f(x_1)=f(x_2)$

H_1: $f(x_1)\neq f(x_2)$

Where $f(x_i)$ is the PDF of sample i.

Thankfully, the only assumption for the Non-Parametric test is:

- **Independent random samples from two populations**

This shouldn't pose a problem as you assumed/deduced this when choosing this type of test.

<u>IN SAS STUDIO</u>

1. **In TASKS, select T-TESTS.**
2. **Select your dataset and under ROLES choose Two-sample test and an ANALYSIS variable (i.e. the variable that the mean will be calculated for**

and compared for the test) and a GROUPS variable (i.e. each sample, could be two countries, two nations, the overall group that the sample was drawn from).

3. **Under OPTIONS:**
   TAILS – choose based on your hypotheses
   NORMALITY ASSUMPTION – Since your data is ranked/scored and distribution free, you can untick this.
   NONPARAMETRIC TESTS – Now we must tick this as it will run the two-sample test and calculate the associate P-value.
4. **Under PLOTS, choose SELECTED PLOTS:**
   HISTORGRAM AND BOX PLOT – Will be ticked
   NORMALITY PLOT – I would tick this because it can be used to further clarify that the data is not normal and thus distribution free.
   CONFIDENCE INTERVAL PLOT – Unlike in the parametric case this is not useful for the non-parametric output as it runs a confidence interval on the mean of the data and not the median which we are testing here.
   WILCOXON BOXPLOT – Tick this as this can be useful for identifying possible skew and its corresponding direction. It'll also help to support the conclusion drawn from the P-value, as if they look completely different, it's reasonable to assume that their PDFs are different.
5. **Select the SAS RUN option (A running man icon).**

TIPS FOR ANALYSIS OF OUTPUT

1. **Summary statistics, 95% Confidence Intervals and the test for equality of variances are printed first. Ignore these as they only apply in the Parametric version of this test.**
2. **The plots are printed next. The histogram and QQ plot can be used to support the conclusion that the data is distribution free and thus this approach was correct. I would only note this if your data failed the test of normality, however. If your data is ranked/scored, there is no need for this output to be examined.**
3. **Scrolling past the plots gives us three tables. One is giving us some summary statistics using the Wilcoxon ranked sum method (don't worry**

about analysing this output), the second is the output for the Wilcoxon two sample test. Notice first that it says "Z includes a continuity correction of 0.5" this is just because a Discrete distribution was approximated by a continuous distribution to obtain the Z-statistic and corresponding P-value. There's a ton of theory behind this but that's beyond the scope of this course. Just interpret (as you normally would by comparing to α) the last value in the table which is under the t-approximation heading. The third is the output for the Kruskal Wallis test, we won't consider this in this section so ignore it for now.

4. **Support your conclusion from the test with the Wilcoxon boxplot, how different are the boxes in terms of placement, skew, where the mean and median are etc. Does this seem reasonable for your conclusion? If so, include it when drawing your conclusion.**
5. **State your conclusion(s) in the context of the problem.**

Two Matched/Paired Sample Testing: Parametric

The question we are looking to answer here is

"Is there a difference in the means between the matching data points?"

Remember here that the same group of subjects/participants will give the same response but for two different samples, this could be two brands of something for example. These samples aren't independent because they're paired with another similar response.

For the parametric approach the test we use is the PAIRED SAMPLES T-TEST. This has the following HYPOTHESES which use the DIFFERENCE OPERATOR which is defined as $D_0 = \mu_1 - \mu_2$ and is just an easier way to write the difference in POPULATION MEANS.

ONE TAILED

H_0: $D_0 = 0$

H_1: $D_0 > 0$ or $D_0 < 0$

TWO TAILED

H_0: $D_0 = 0$

H_1: $D_0 \neq 0$

The most important preliminary step before carrying out tests and analysing outputs in SAS Studio is to:

CHECK THE ASSUMPTIONS FOR A PARAMETRIC APPROACH!!

This is so important because:

IF THE DATA DOES NOT MEET THE ASSUMPTIONS, WE USE THE NON-PARAMETRIC APPROACH INSTEAD!!

This test has the following ASSUMPTIONS:

- **Samples are matched/paired with the same response in both groups**
- **The Distribution of the population from which the samples are drawn is NORMALLY DISTRIBUTED. This is based on the HYPOTHESES:**

  H_0: There is no difference between the distribution of the variables and that of the Normal Distribution;

  H_1: There is a difference between the distribution of the variables and that of the Normal Distribution.

  So of course, to satisfy this assumption we want a P-value>α so that we DO NOT REJECT H_0.

Thankfully, the first assumption will have been assumed/deduced when selecting this type of test. The Normality assumption is tested for when running the paired t-test in SAS Studio so we will discuss it in the next part.

IN SAS STUDIO

1. **In TASKS choose T-TESTS.**
2. **Select your dataset and then under T-TEST select PAIRED TEST. Next you must choose the Group 1 Variable and the Group 2 Variable. These would be like "Brand A", "Brand B" etc dependent on the context (The groups that you are comparing the means for in this test).**
3. **Under OPTIONS**

   TAILS: Input based on your hypotheses.

   ALTERNATIVE HYPOTHESIS: This is where we set what the value that we are testing for is. Of course, since we have H_0: D_0=0 which translates to $\mu_1 - \mu_2 = 0$, the value to input would be 0. This is the same for all of the tests in this section as we just want to know if the mean difference exists (i.e. is it 0?).

   NORMALITY ASSUMPTION: This is required for the assumptions

NON-PARAMETRIC TESTS: Ignore this for now, this will be explained in more detail in the next section. It's worth noting however that if you did tick this, both tests get run at the same time, so you can rule out one and use the other! However, this produces a lot of output which is convoluted to interpret and sort through, which is why I split it up here.

4. **Under PLOTS choose SELECTED PLOTS**
   HISTOGRAM AND BOXPLOT – Ticked automatically, very useful
   NORMALITY PLOT – Very useful for the test for Normality
   AGREEMENT PLOT – You can look up how to interpret this if necessary, but I would untick it.
   RESPONSE PROFILE PLOT – Again could be useful but is rather niche when it comes to drawing conclusions so I would untick it.
   CONFIDENCE INTERVAL PLOT - I would tick this as it's very useful to make sure it corroborates (agrees with) your P-value conclusion.
   The way it works is if the Confidence Interval (marked with two lines and shaded a green-blue sort of colour) contains the value we are testing for (c in this case, would be 0 a lot of the time) , then 1- α of the time (95% usually as we tend to choose significance level 5% most often) the true value we are testing for will be in that confidence interval and we CANNOT be sure that the value ISNT c. So therefore, we can't reject the Null hypothesis if that's the case. However, if our value c is outside of the confidence interval, then there is evidence to reject the null hypothesis.
5. **Select the SAS RUN option (Icon of a running man) to get your output.**

TIPS FOR ANALYSIS OF OUTPUT

1. **The first output is the test of normality on the difference operator D_0. SAS runs four different tests with four corresponding P-values from which you can draw conclusions. Thankfully a lot of the tables that SAS runs will tell you in the labels and headings what the output is.**

Thankfully whichever (one or two tailed) type of test you chose is considered and adjusted for in the output, so you just need to compare the P-values to α.

To support the conclusion from this, use the Histogram and the QQ plot.

QQ plots tell you if quantiles of your data are similar to the quantiles of a Standard Normal Distribution (whichever distribution you're testing for of course, usually normal), if all points are following that similar line then we likely have the same distribution.

IF THERE IS EVIDENCE TO REJECT H_0, ASSUME NON-PARAMETRIC AND GO TO THE NEXT SECTION.

2. **Next we have a table of summary statistics which could be useful to support a plot if necessary but the really important thing here is the 95% Confidence Interval table. Remember if this confidence interval contains the value we are testing for (in this case 0) then we cannot reject H_0. However, if the value lies outside the interval then there is evidence to suggest that we can reject H_0. This is very useful as we can check now to make sure the conclusion from our P-value falls in line with this conclusion.**
3. **Beneath the confidence interval table, we have a small table with headings DF, t VALUE, PR>|t|. This is the table with the all-important P-value for our test. Interpret this by comparing with α and draw conclusions.**
4. **State conclusions in the context of the problem.**

Two Matched/Paired Sample Testing: Non-Parametric

If the data you have is in the form of ranks/scores, or your data has failed the test for Normality in your SAS output you assume the data is DISTRIBUTION FREE and now we use a NON-PARAMETRIC APPROACH.

The question we're looking to answer here is:

"Does the difference between pairs follow a symmetric distribution around 0?"

Don't stress too much about what this means, you're just investigating whether the differences exist and are significant or not. Since this approach is non-parametric, we can't use the means as they tend to be skewed and therefore not useful in these situations.

The Tests we use for the non-parametric approach are the WILCOXON SIGNED RANK TEST and the PAIRED SAMPLE SIGN TEST. Both of these are run in the SAS output. These tests have the HYPOTHESES:

H_0: The difference between pairs follows a symmetric distribution around 0

H_1: The difference between pairs does not follow a symmetric distribution around 0

Thankfully when it comes to the non-parametric approach there is only one assumption to check and that is

- **The data is in the form of paired/matched samples**

Which would've been assumed/deduced when selecting this test.

IN SAS STUDIO

1. **In TASKS choose T-TESTS.**
2. **Select your dataset and then under T-TEST select PAIRED TEST. Next you must choose the Group 1 Variable and the Group 2 Variable. These would be like "Brand A", "Brand B" etc dependent on the context (The groups that you are comparing the means for in this test).**
3. **Under OPTIONS**
   TAILS: Input based on your hypotheses.
   ALTERNATIVE HYPOTHESIS: This is where we set what the value that we are testing for is. Of course, since we have hypotheses based on a symmetric distribution around 0 then in these cases, we will leave it as the base input of 0.
   NORMALITY ASSUMPTION: As your data is in the form of rank/scores or failed this assumption previously, untick it.
   NON-PARAMETRIC TESTS: Tick this as it will run both tests and is the most important thing in the options group for this section.
4. **Under PLOTS choose SELECTED PLOTS**
   HISTOGRAM AND BOXPLOT – Ticked automatically, you can use the boxplot.
   NORMALITY PLOT – Not necessary, untick.
   AGREEMENT PLOT – You can look up how to interpret this if necessary, but I would untick it.
   RESPONSE PROFILE PLOT – Again could be useful but is rather niche when it comes to drawing conclusions so I would untick it.
   CONFIDENCE INTERVAL PLOT – I would untick this here as it's based on the population means and we aren't testing those in the non-parametric case.
5. **Select the SAS RUN option (Icon of a running man) to get your output.**

TIPS FOR ANALYSIS OF OUTPUT

1. **The first output is the summary statistics which could be useful for discussing plots if that's necessary. The 95% confidence interval can be ignored as it's based on the means which would likely be skewed here. The next t-value table can also be ignored as we are interested in the non-parametric output, not the parametric one.**
2. **The plots are printed next. The histogram and QQ plot can be used to support the conclusion that the data is distribution free and thus this approach was correct. I would only note this if your data failed the test of normality, however. If your data is ranked/scored, there is no need for this output to be examined.**
3. **The last output which is the core of this approach is a table of P-values from multiple tests. Ignore the t-value one but take note of both the paired sample sign test P-value and the Wilcoxon Signed Ranks test P-value. Interpret these P-values by comparing with α. Remember they must ALL be significant to conclude that we can reject H_0. Now a conclusion can be drawn.**
4. **State conclusion in the context of the problem.**

Correlation Analysis: Parametric

The question we are looking to answer here is

"Is there a relationship/association between the two variables?"

We also have a subsequent question to answer here IF A RELATIONSHIP EXISTS

"Is this relationship a positive or negative association and how strong?"

Some quick theory on this before we get started. Overall this is one of the easiest tests but there is a little more set up involved interpreting terms and coefficients etc.

The CORRELATION COEFFICIENT r measures the STRENGTH OF ASSOCIATION BETWEEN TWO VARIABLES.

For the purposes of this course we have these guidelines:

Range of r values	Corresponding Relationship
-1<r<-0.6	STRONG NEGATIVE ASSOCIATION
-0.6<r<-0.3	WEAK NEGATIVE ASSOCIATION
-0.3<r<+0.3	LITTLE ASSOCIATION
+0.3<r<+0.6	WEAK POSITIVE ASSOCIATION
+0.6<r<+1	STRONG POSITIVE ASSOCIATION

It's useful to note that a correlation near or equal to 0 implies such little association that the variables can be assumed/concluded as independent.

Also, if the association is positive, this means AS ONE VARIABLE INCREASES SO DOES THE OTHER.

If the association is negative, this means AS ONE VARIABLE INCREASES THE OTHER DECREASES.

For the Parametric approach to correlation analysis we use PEARSON'S CORRELATION COEFFICIENT. This estimate (i.e. the sample correlation coefficient) is denoted as ρ. For this test we have the HYPOTHESES:

ONE TAILED

H_0: $\rho=0$

H_1: $\rho<0$ or $\rho>0$

TWO TAILED

H_0: $\rho=0$

H_1: $\rho \neq 0$

The most important preliminary step before carrying out tests and analysing outputs in SAS Studio is to:

CHECK THE ASSUMPTIONS FOR A PARAMETRIC APPROACH!!

This is so important because:

IF THE DATA DOES NOT MEET THE ASSUMPTIONS, WE USE THE NON-PARAMETRIC APPROACH INSTEAD!!

This test only has one very important ASSUMPTION

- **The Distribution of the population from which the variables are from is NORMALLY DISTRIBUTED. This is based on the HYPOTHESES:**
  H_0: There is no difference between the distribution of the variables and that of the Normal Distribution;
  H_1: There is a difference between the distribution of the variables and that of the Normal Distribution.
  So of course, to satisfy this assumption we want a P-value>α so that we DO NOT REJECT H_0.

<u>IN SAS STUDIO PART 1</u>

1. **We actually have two smaller processes to cover for the Parametric case. First, we must conduct a DISTRIBUTION ANALYSIS, as the next step CORRELATION ANALYSIS does not test for normality.**
2. **In TASKS → STATISTICS select DISTRIBUTION ANALYSIS.**

3. **Select your dataset and choose your ANALYSIS VARIABLES. These are the variables you wish to check the normality of (there could be many of course in your dataset).**
4. **In OPTIONS under EXPLORING DATA**
   This whole section is only necessary when you're wanting to get to grips with your data and checking its summary statistics etc, it's not as useful when checking for normality.
5. **In OPTIONS under CHECKING FOR NORMALITY**

- **HISTOGRAM AND GOODNESS OF FIT TESTS: Very useful for supporting conclusions drawn. You can add inset statistics if you wish. This is optional, you have a drop-down menu where you can tell SAS what to put as a box on the graph, e.g. N, Mean, Median etc. Tick whatever you feel is useful for your analysis.**
- **NORMAL PROBABILITY PLOT: I wouldn't bother with this as the following QQ plot will suffice.**
- **NORMAL QUANTILE-QUANTILE PLOT: Definitely tick this, as when combined with the histogram it provides excellent evidence to support the conclusion of your hypothesis test. You can once again add inset statistics if you wish.**

6. **In OPTIONS under FITTING DISTRIBUTIONS**
   Ignore this section entirely as we aren't testing for any other distributions to fit to the data. That is beyond the scope of the course.
7. **Select the SAS RUN option (The running man icon).**

<u>TIPS FOR ANALYSIS OF OUTPUT PART 1</u>

1. **Your first outputs are the histograms from the exploring data section if you ran that part of the analysis. If you didn't, skip to step 2.**
2. **Scroll down until you get to a table with heading "Goodness-of-Fit Tests for the Normal Distribution". There will be two of these tables, one for each variable. Remember that <u>BOTH</u> variables must conclude in favour of accepting H_0 to pass the test for normality. If they do not and at least one of them fails and we would reject H_0, SKIP TO THE NEXT SECTION AS THE DATA IS NON-PARAMETRIC. Interpreting these is the same as always, are the three P-values calculated < or > significance level α etc.**

3. **You can now use the Histograms with plotted Normal curve and QQ plots to support your conclusion from Step 2. QQ plots tell you if quantiles of your data are similar to the quantiles of a Standard Normal Distribution (whichever distribution you're testing for of course, usually normal), if all points are following that similar line then we likely have the same distribution. Following from that the histogram can show you how close the distribution of the data follows the plotted curve.**
4. **State your overall conclusion from the Distribution Analysis.**

IN SAS STUDIO PART 2

1. **Now that we have checked that the variables are normally distributed it's time to run a CORRELATION ANALYSIS.**
2. **IN TASKS → STATISTICS select CORRELATION ANALYSIS.**
3. **Select your dataset and provide your ANALYSIS VARIABLES, these are the variables you wish to analyse in order to find if there's a relationship or not (two out of a possible many).**
4. **Ignore the other sections here CORRELATE WITH, PARTIAL VARIABLES and ADDITIONAL ROLES.**
5. **In OPTIONS under STATISTICS → DISPLAY STATISTICS**

- **CORRELATIONS: Definitely as if there is an association, we will want to answer the question of it being positive or negative, and how strong!**
- **DISPLAY P-VALUES: Essential to our conclusion.**
- **The rest aren't necessary here, you can click DESCRIPTIVE STATISTICS to get a table of the mean, median, maximum, minimum etc if you wish.**
- **NON-PARAMETRIC CORRELATIONS: Ignore this for now as this will be explained in more detail in the next section. It's worth noting however that if you did tick this, both tests get run at the same time, so you can rule out one and use the other! However, this produces a lot of output which is convoluted to interpret and sort through, which is why I split it up here.**

6. **In OPTIONS under PLOTS**

- **TYPE OF PLOT: Pull down this menu and select MATRIX OF SCATTER PLOTS as this is very useful, will plot each variable against the other**

both ways (one on the y axis, one on the x axis and swapped etc) and is useful for the final conclusion.

- **Ignore the rest as it's not necessary for the analysis that we're doing here.**

7. **Select the SAS RUN option (The icon with the running man) to obtain the output.**

<u>TIPS FOR ANALYSIS OF OUTPUT PART 2</u>

1. **If you selected DESCRIPTIVE STATISTICS this will be the first output, use this however you wish but remember it's not as useful past the Exploratory Data Analysis stage.**
2. **The next output is the table of Pearson Correlation coefficients for each variable plotted against itself and then the other variable. Beneath these coefficient values there is a corresponding P-value as well. This P-value is what we will compare to α to determine our conclusion. What's great here is that is the Coefficient is above the "Little Association" range from the table we would expect the P-value to be significant at the 5% level. We can therefore kill two birds with one stone here, as we answer the initial question of the relationship existing, and if so, we also answer the follow up question of its sign and size.**
3. **The last output is the matrix of scatter plots which will be the perfect companion to the previous table to support conclusions drawn.**
4. **State your conclusions in the context of the problem. Don't forget to note the follow up question as well, as this will give you and the reader valuable insight.**

Correlation Analysis: Non-Parametric

The question we are looking to answer here doesn't change if the data is in the form of rank/scores or fails the normality assumption. It's

"Is there a relationship/association between the two variables?"

We also have a subsequent question to answer here IF A RELATIONSHIP EXISTS

"Is this relationship a positive or negative association and how strong?"

Some quick theory on this before we get started. Overall this is one of the easiest tests but there is a little more set up involved interpreting terms and coefficients etc.

The CORRELATION COEFFICIENT r measures the STRENGTH OF ASSOCIATION BETWEEN TWO VARIABLES.

For the purposes of this course we have these guidelines:

Range of r values	Corresponding Relationship
-1<r<-0.6	STRONG NEGATIVE ASSOCIATION
-0.6<r<-0.3	WEAK NEGATIVE ASSOCIATION
-0.3<r<+0.3	LITTLE ASSOCIATION
+0.3<r<+0.6	WEAK POSITIVE ASSOCIATION
+0.6<r<+1	STRONG POSITIVE ASSOCIATION

It's useful to note that a correlation near or equal to 0 implies such little association that the variables can be assumed/concluded as independent.

For the Non-Parametric approach to correlation analysis we use SPEARMAN'S RANK CORRELATION COEFFICIENT. This estimate (i.e. the sample correlation coefficient) is denoted as ρ. For this test we have the HYPOTHESES:

ONE TAILED

H_0: $\rho=0$

H_1: $\rho<0$ or $\rho>0$

TWO TAILED

H_0: ρ=0

H_1: $\rho \neq 0$

Thankfully with the non-parametric approach we have no assumptions to check. This is by FAR the shortest and least complicated test of them all.

<u>IN SAS STUDIO</u>

1. **We will be running a CORRELATION ANALYSIS.**
2. **IN TASKS → STATISTICS select CORRELATION ANALYSIS.**
3. **Select your dataset and provide your ANALYSIS VARIABLES, these are the variables you wish to analyse in order to find if there's a relationship or not (Two in this case out of possibly many).**
4. **Ignore the other sections here CORRELATE WITH, PARTIAL VARIABLES and ADDITIONAL ROLES.**
5. **In OPTIONS under STATISTICS → DISPLAY STATISTICS**

- **CORRELATIONS: Definitely as if there is an association, we will want to answer the question of it being positive or negative, and how strong!**
- **DISPLAY P-VALUES: Essential to our conclusion.**
- **The rest aren't necessary here, you can click DESCRIPTIVE STATISTICS to get a table of the mean, median, maximum, minimum etc if you wish.**
- **NON-PARAMETRIC CORRELATIONS: Absolutely essential that you tick this, specifically only the SPEARMANS RANK option, ignore the others.**

6. **In OPTIONS under PLOTS**

- **TYPE OF PLOT: Pull down this menu and select MATRIX OF SCATTER PLOTS as this is very useful, will plot each variable against the other both ways (one on the y axis, one on the x axis and swapped etc) and is useful for the final conclusion.**
- **Ignore the rest as it's not necessary for the analysis that we're doing here.**

7. **Select the SAS RUN option (The icon with the running man) to obtain the output.**

TIPS FOR ANALYSIS OF OUTPUT

1. **If you selected DESCRIPTIVE STATISTICS this will be the first output, use this however you wish but remember it's not as useful past the Exploratory Data Analysis stage.**
2. **The next output is the table of Pearson Correlation coefficients table and also the Spearman's Rank Correlation coefficients table for each variable plotted against itself and then the other variable. Beneath these coefficient values there is a corresponding P-value as well. This P-value is what we will compare to α to determine our conclusion. Remember this is the non-parametric case, so we ignore the Pearson table and just focus on the Spearman's Rank table. What's great here is that is the Coefficient is above the "Little Association" range from the table we would expect the P-value to be significant at the 5% level. We can therefore kill two birds with one stone here, as we answer the initial question of the relationship existing, and if so, we also answer the follow up question of its sign and size.**
3. **The last output is the matrix of scatter plots which will be the perfect companion to the previous table to support conclusions drawn.**
4. **State your conclusions in the context of the problem. Don't forget to note the follow up question as well, as this will give you and the reader valuable insight.**

Three or more Sample Testing: Parametric

The question we are looking to answer here is

"Are the population means of the m groups ($m \geq 3$) equal?"

There is also a more complicated subsequent question to answer

"If there are differences, where do they lie? Can we order the groups or compare them in a useful way knowing that differences exist?"

This will be a little more complex than the one or two sample case. The most common type of experiment we have is a randomised experiment where either

(i) Groups randomly assigned to experimental subjects

(ii) Random samples are drawn from each of the m populations

Note that the size of each sample (n) doesn't necessarily need to be equal.

The Parametric approach for this case is the ONE-WAY ANALYSIS OF VARIANCE (ANOVA) procedure. This test has the HYPOTHESES:

H_0: $\mu_1=\mu_2=\mu_3\ldots=\mu_m$ (Where the μ_i are the POPULATION MEANS)

H_1: The μ_i are not all equal

We tend to consider TWO TAILED here as we are considering many variables at once.

ANOVA discriminates between these groups via two measurements of variance

(i) Response WITHIN GROUPS

(ii) Differences in GROUP MEANS

Of course, you may be wondering why we don't do many individual Two sample t-tests for all of the means. This would not only be cumbersome, but each time a mean is re-compared with a different one, the probability of committing a TYPE 1 ERROR (i.e. rejecting the null hypothesis when it was in fact true!) increases beyond an acceptable region.

The most important preliminary step before carrying out tests and analysing outputs in SAS Studio is to:

CHECK THE ASSUMPTIONS FOR A PARAMETRIC APPROACH!!

This is so important because:

IF THE DATA DOES NOT MEET THE ASSUMPTIONS, WE USE THE NON-PARAMETRIC APPROACH INSTEAD!!

The key ASSUMPTIONS of the ANOVA are as follows:

- **Subjects are selected at random from the m groups**
- **The DEPENDENT VARIABLE (or response) in each group is NORMALLY DISTRIBUTED. This is based on the HYPOTHESES:**

  H_0: There is no difference between the distribution of the variables of the groups and that of the Normal Distribution;

  H_1: There is a difference between the distribution of the variables of the groups and that of the Normal Distribution.

  So of course, to satisfy this assumption we want a P-value>α so that we DO NOT REJECT H_0.
- **The DEPENDENT VARIABLE (or response) in each group has the same variance (i.e. we have HOMOGENEITY OF VARIANCES). If this isn't satisfied the data has HETEROSKEDASTICITY. This is based on the HYPOTHESES:**

  H_0: $\sigma_1^2=\sigma_2^2=\sigma_3^2\ldots=\sigma_m^2$

  H_1: σ_i^2 are not all equal

You can determine if the first assumption is satisfied via looking at the data and how it was collected. The Normality assumption and Homogeneity of Variances assumptions are checked in the SAS output.

<u>IN SAS STUDIO PART 1</u>

1. **We actually have two smaller processes to cover for the Parametric case. First, we must conduct a DISTRIBUTION ANALYSIS, as the next step ONE-WAY ANOVA does not test for normality.**

2. **In TASKS → STATISTICS select DISTRIBUTION ANALYSIS.**
3. **Select your dataset and choose your ANALYSIS VARIABLES. These are the variables you wish to check the normality of (there could be many of course in your dataset). Usually this will just be the dependent variable in each group.**
4. **In OPTIONS under EXPLORING DATA**
   This whole section is only necessary when you're wanting to get to grips with your data and checking its summary statistics etc, it's not as useful when checking for normality.
5. **In OPTIONS under CHECKING FOR NORMALITY**

- **HISTOGRAM AND GOODNESS OF FIT TESTS: Very useful for supporting conclusions drawn. You can add inset statistics if you wish. This is optional, you have a drop-down menu where you can tell SAS what to put as a box on the graph, e.g. N, Mean, Median etc. Tick whatever you feel is useful for your analysis.**
- **NORMAL PROBABILITY PLOT: I wouldn't bother with this as the following QQ plot will suffice.**
- **NORMAL QUANTILE-QUANTILE PLOT: Definitely tick this, as when combined with the histogram it provides excellent evidence to support the conclusion of your hypothesis test. You can once again add inset statistics if you wish.**

6. **In OPTIONS under FITTING DISTRIBUTIONS**
   Ignore this section entirely as we aren't testing for any other distributions to fit to the data. That is beyond the scope of the course.
7. **Select the SAS RUN option (The running man icon).**

<u>TIPS FOR ANALYSIS OF OUTPUT PART 1</u>

1. **Your first outputs are the histograms from the exploring data section if you ran that part of the analysis. If you didn't, skip to step 2.**
2. **Scroll down until you get to a table with heading "Goodness-of-Fit Tests for the Normal Distribution". This table must conclude in favour of accepting H_0 to pass the test for normality. If it does not, we would reject H_0, SKIP TO THE NEXT SECTION AS THE DATA IS NON-PARAMETRIC.**

Interpreting this is the same as always, as the three P-values calculated < or > significance level α etc.

3. **You can now use the Histogram with plotted Normal curve and QQ plot to support your conclusion from Step 2. QQ plots tell you if quantiles of your data are similar to the quantiles of a Standard Normal Distribution (whichever distribution you're testing for of course, usually normal), if all points are following that similar line then we likely have the same distribution. Following from that the histogram can show you how close the distribution of the data follows the plotted curve.**
4. **State your overall conclusion from the Distribution Analysis.**

IN SAS STUDIO PART 2

1. **In TASKS → STATISTICS select LINEAR MODELS**
2. **Choose ONE-WAY ANOVA and choose your dataset. You must now choose the DEPENDENT VARIABLE (the variable measured in the test, the response variable for each subject, participant etc) and the CATEGORICAL VARIABLE (i.e. the group variable, the groups that there is m of, could be country for example).**
3. **In OPTIONS**

- **HOMOGENEITY OF VARIANCES: There is a drop-down menu for TEST. The test we will be utilising is usually pre-selected and is LEVENE'S TEST. Untick the WELCH'S VARIANCE WEIGHTED ANOVA as this is beyond the scope of the course (this is where the ANOVA does not account for homogeneity of variances, we won't be utilising it here).**
- **COMPARISONS: There is a drop-down menu for COMPARISON METHOD. Now this is VERY IMPORTANT. Something necessary when it comes to answering the subsequent question on the differences (if they exist of course) is comparing them to make sense of the data. This method is called POST-HOC COMPARISON. There are different methods for Parametric and Non-Parametric approaches. More on how to interpret this in the Tips for analysis output section.**
  For the Parametric case we use the BONFERRONI TEST so select this from the menu and select your significance level.

- **PLOTS: There is a drop-down menu for DISPLAY PLOTS. Choose SELECTED PLOTS and tick the following:**
  BOX PLOT: Already ticked but very useful for comparisons
  MEANS PLOT: This is useful as it provides a visualisation of where the LS means are placed (LS means Least Squares, it's a linear regression term) and helps to solidify your conclusions in the post-hoc comparisons.
  LS-MEAN DIFFERENCES PLOT: This can also be useful, each comparison in the post-hoc section has a straight line plotted for it and is colour coded on whether it's significant or not, useful for just solidifying your conclusion with evidence.
  DIAGNOSTICS PLOT: I would leave this one unticked as it brings in more complexities that we don't need at this stage of the course.

4. **Select the SAS RUN option (The icon of the running man) and SAS will produce our output.**

<u>TIPS FOR ANALYSIS OF OUTPUT PART 2</u>

1. **The first output is just some information about the classes/groups and each of their names, along with the overall number of observations. This doesn't need to be included in the analysis in your report, but it is useful for getting to grips with the structure of the data.**
2. **Now scroll down until you get to a table with "LEVENE'S TEST FOR HOMOGENEITY OF VARIANCES". This is where we see if the third assumption for One-Way ANOVA is satisfied. There's a P-value to interpret (remember compare with significance level and draw conclusion).**
   If H_0 is rejected (so that there IS a difference between the variances) we can stop here for answering the first question about group differences and move on to the Bonferroni Test for Post-Hoc comparisons. This is because a difference in variances is a powerful enough conclusion to conclude OVERALL that the groups are indeed different, and we must carry on testing for individual differences. Skip to step 4.
   If H_0 is accepted (i.e. not rejected) so that the variances are the same, carry on from here as the P-value for the original hypothesis we set (not

the variances or normal one, the ANOVA one!) will now need to be interpreted. Carry on to step 3.

3. **Now scrolling back up past the Levene's test table, we have four confusing looking tables that all essentially say the same thing. We can now finally answer the first question since the assumptions have all been satisfied. All of these contain an F-value which is the P-value for our ANOVA hypothesis test (the one about the means of the m groups being equal!) Interpret this as normal and draw the initial conclusion. You can support this conclusion with the boxplot, as if there is evidence for differences then some should look considerably different in terms of location, skew etc than others.**

   If H_0 was rejected (So, there are differences in the means of the groups!) Carry on to step 4.

   If H_1 was accepted (i.e. not rejected) then there are no differences in the groups (of course support this with the boxplot) and we conclude our analysis here. Skip to step 5

4. **Now it's time to answer the second question on where the differences lie now that we know that they exist.**

   Scrolling down to the bottom we have the heading "Least Square Means Adjusted for Multiple Comparisons Bonferroni" and two tables underneath. Paste both of these into your analysis. The first just tells you what the adjusted means for this test was and what was used to calculate the P-values in the comparison test.

   Next we have a matrix of sorts where each group mean was compared to the others in turn and associated P-values calculated. Use these p-values to determine where there could be significant differences (i.e. comparisons that produced P-values less than the significance level) and use the means plot and LS means differences plot to come up with your final conclusions (it could be say that one group is significantly different from the rest, and looking at the means plot tells you its higher/lower etc) as this is very useful. I would also use the means difference plot as it uses a colour code for differences that are and aren't significant, just for extra evidence to support your conclusion.

5. **State your conclusion in the context of the problem.**

Three or more Sample Testing: Non-Parametric

If the data you have is in the form of ranks/scores, or your data has failed the test for Normality in your SAS output you assume the data is DISTRIBUTION FREE and now we use a NON-PARAMETRIC APPROACH.

The question we're looking to answer here is:

"Are the m distributions of the m groups identical?"

There is also a more complicated subsequent question to answer

"If there are differences, where do they lie? Can we order the groups or compare them in a useful way knowing that differences exist?"

This will be a little more complex than the one or two sample case. The most common type of experiment we have is a randomised experiment where either

(i) Groups randomly assigned to experimental subjects

(ii) Random samples are drawn from each of the m populations

Note that the size of each sample (n) doesn't necessarily need to be equal.

The Non-Parametric approach for this case is the KRUSKAL WALLIS TEST which is an extension of the Wilcoxon method and involves the CHI-SQUARE DISTRIBUTION (Denoted χ^2 Distribution).

This method has the HYPOTHESES:

H_0: The m distributions are identical

H_1: Not all of the m distributions are the same

We tend to consider TWO TAILED here as we are comparing lots of different variables at once.

The Non-Parametric approach does have some ASSUMPTIONS in this case so we must check that the data and groups satisfy them.

- **The m samples are INDEPENDENTLY and RANDOMLY selected from their respective populations.**
- **For the χ^2 Distribution we require that n>5 for each sample (i.e. there are 5 or more subjects/participants in each sample).**
- **Tied observations are assigned the value of ranks.**

Thankfully these assumptions can be checked really quickly and do not require tests in SAS. The first two can be checked off by looking at your data (you can easily see if n>5 for each group for example, the experiment that was carried out will tell you if they're random, sometimes if they're independent). Sometimes with independence it's worth noting that this means the responses in each of the m groups don't directly affect each other and each group only gives one response for one group (i.e. each group has different subjects who give one response that is completely separate from the subjects in the other groups).

Don't worry too much about the last assumption as this ranking procedure will be done in SAS when running the Non-Parametric ANOVA.

<u>IN SAS STUDIO</u>

1. **IN TASKS → LINEAR MODELS choose NONPARAMETRIC ONE-WAY-ANOVA.**
2. **Choose your dataset and then your DEPENDENT VARIABLE (the one that's being measured, i.e. the response variable in each group) and the CLASSIFICATION VARIABLE (the groups variable, could be country, just "group", school, etc etc).**
3. **In OPTIONS under PLOTS**

- **Ignore this option, not necessary for our testing here.**

In OPTIONS under TESTS

- **The drop-down menu has two options, leave it as the pre-selected ASYMPTOTIC TESTS.**
- **LOCATION DIFFERENCES: Leave WILCOXON SCORES ticked and leave the rest blank.**
- **SCALE DIFFERENCES: Leave these as we don't require them for our testing.**
- **LOCATION AND SCALE DIFFERENCES: We don't require these either.**
- **ADDITIONAL TESTS: Here is where we introduce how to answer the subsequent question if we find that a difference does in fact exist! Now this is VERY IMPORTANT. Something necessary when it comes to answering the subsequent question on the differences (if they exist of course) is comparing them to make sense of the data. This method is**

called POST-HOC COMPARISON. There are different methods for Parametric and Non-Parametric approaches. More on how to interpret this in the Tips for analysis output section.

For the Non-Parametric case we use PAIRWISE MULTIPLE COMPARISON ANALYSIS. Tick this option.

4. In OPTIONS under DETAILS

- Ignore this entire section as it's not needed for our analysis.

5. Select the SAS RUN option (The icon of the running man) to obtain the output.

TIPS FOR ANALYSIS OF OUTPUT

1. The first output is the table of Wilcoxon scores. I would include this in the report for completeness as the very last row says "Average score used for ties" which means the third assumption for Kruskal Wallis is therefore satisfied!
2. Next we have a small table for the Kruskal Wallis test output. The χ value is there as well as the associated P-value which you should interpret to draw your initial conclusion/answer to the first question.
   If H_0 is rejected (so that there are differences between the groups!) then use the boxplot to support this conclusion and then move forward to step 3.
   If H_0 is accepted (so that there are no differences between the groups!) we conclude our analysis here and just support your conclusion with the accompanying boxplot. Skip to step 4.
3. The last output is what we will use to answer the second question on determining what the individual differences are and where they lie. The groups are compared PAIRWISE (so in pairs) and the output is summarised in a table with P-values for each pairwise comparison. Interpreting these allows you to find where significant differences are (the tests where H_0 would be rejected) and then refer back to the boxplot to support your subsequent conclusions as these differences should be visible in the plot (i.e. if one group is found significantly different from all the others and on the box plot it's the lowest or the highest of

something when compared to the others, then you can conclude that it's the best/worst etc) which is very powerful as supportive evidence.

4. **State your conclusion in the context of the problem.**

Final Say

I just want to say thank-you for sticking with my incessant rambling and overly in-depth breakdown in this Statistical Inference shortcut document. This topic is very complicated and there are lots of cases to consider so I felt to really complete this document I had to include as much necessary information as I could whilst skimming over the underlying theory when it wasn't necessary to do the tests (you can feel free to look sections on Hypothesis testing terms up of course if necessary!).

I really hope that this guide helped you in some way, as helping those who need it is something I'm super passionate about, especially when it's about Statistics which is one of my favourite things that I always nerd out over but also love and hate the complexities of. With these tests and their little ins and outs, practice is the best way to get into your head how they work and why you're doing what you're doing and why this plot is useful etc. There are many resources online with examples for all types of tests that you can attempt to help with this.

Wishing you all the best

Jake Marshall