The following resources are associated:

‘Opening and labelling data in SAS’, ‘Checking normality in SAS’ and the dataset ’Birthweight’

**Independent t-test and Wilcoxon rank sum in SAS**

**Dependent variable:** Continuous (scale)

**Independent variable**: Binary (2 independent groups)

**Common Applications:** Comparing means of data from two unrelated groups on the same continuous, dependent variable; for example, you could use an independent t-test to test whether first year graduate salaries differed based on gender or whether there is a difference in test anxiety based on educational level which has two groups: "undergraduates" and "postgraduates".

**Data:** The data set ‘*Birthweight\_reduced’* contains details of 42 babies and their parents at birth. The research question is whether the mother smoking influences the birthweight of a baby, so the dependant variable is Birth weight (lbs) and the independent variable is whether the mother smokes (smoker).



0 = Non-smoker

1 = Smoker

Import the Birthweight dataset into SAS (see ‘*Opening and labelling data’* sheet). It is helpful to give the folder where the data is stored a short name e.g. here it is referred to as dat, using the libname statement at the beginning of each program.

libname dat "E:\SHUUSERs\!SharedData\Ellen\Statistics&Probability\Data";

Define and format the dataset to be used so that variables have the appropriate labels in output. The code below shows that the short name for the **data** in this analysis is birth and it is accessed from the birthweight dataset stored in the folder labelled dat in the previous command. The label command is used to show what each variable name represents e.g. mppwt is the mothers pre-pregnancy weight.

**data** birth; /\* The adjusted data has the name birth \*/

 set dat.birthweight; /\* The raw data is in the dat folder\*/

 label Birthweight='Birthweight (lbs)'

 smoker='Smoking status of the mother'

 mppwt='Prepregnancy weight of mother (lbs)';

The numeric values given to categories can also be defined by using the **proc** **format** procedure. These formats can then be applied to any variable within other procedures. Here a format called smokeS is defined where 0=Non-smoker and 1=smoker.

**proc** **format**; value smokeS **1**='Smoker' **0**='Non-smoker';

**run**;

## Assumptions

|  |  |  |
| --- | --- | --- |
| **Assumptions** | **How to check** | **What to do if the assumptions are not met** |
| The dependent variable should be approximately **normally distributed** for each group  | Use histograms produced as part of the t test outputFor more details see the *Checking normality in SAS* resource. | If the data for either group is very skewed, use the Wilcoxon rank sum test (see section below). |
| Homogeneity (equality) of variance: The variances (SD squared) should be similar for all the groups | The F-test produced tests the hypothesis ‘the variances are equal’. If p < 0.05, assumption is not met. The test is oversensitive for large sample sizes so a more reliable method is to compare the two standard deviations. If one SD is more than twice the other, the assumption has not been met.  | If one SD is more than twice the other, (or F-test p < 0.05), the assumption has not been met, so use the Satterthwaite row for the t-test. |

**Carrying out an independent t-test in SAS**

The independent t-test tests the null hypothesis ‘There is no difference in the mean birthweights of babies whose mothers smoke and don't smoke’. If the p-value produced within the output for the t-test is less than 0.05, the null is rejected and we conclude that there is a statistically significant difference between the groups.

To run an independent t-test, use the **proc** **ttest** procedure specifying the dataset of interest on the first line (here it is in the folder . The independent (binary grouping variable) goes on the class line and the continuous dependent variable goes on the var line. The format line tells SAS to apply the smokeS format to the variable smoker so that these labels appear instead of the values 0 and 1 in the output. Remember to turn the graphics on using ods graphics onso SAS produces histograms to check normality when a t test is run.

ods graphics on; run;

**proc** **ttest** data=birth;

class smoker;

var Birthweight;

format smoker smokeS.;

SAS produces summary statistics automatically for the two groups and the difference between the means within the **ttest** procedure. Assess whether you think there are big differences between the group means and standard deviations. This is particularly important for large sample sizes where small differences can be significant.

*\*Remember when reporting values from tables, 2 decimal places is usually enough*.

The table above shows that on average the birthweight of babies from mothers who smoke is higher by 0.81 lbs. We can also see that the standard deviations are similar.

The steps of an independent t test are…

# 1) Check for an approximate normal distribution by group

As long as you have turned ods graphics on, these histograms will be produced when the **ttest** procedure is run.



As long as the histograms show the data are approximately normally distributed the t-test can be carried out as normal.

If histograms are clearly very skewed, then t test results will be unreliable and a Mann-Whitney or Wilcoxon rank sum test should be run instead and medians rather than means compared. For more information on checking normality generally, see the *‘Checking normality in SAS’* resource.

2) **Check the assumption of equality of variance**

As the largest SD shown in the summary statistics table is not more than twice the smallest we can assume the variances are equal and use the standard ‘pooled row’ of the t test output.

The folded F-test is a different check testing the hypothesis ‘The variances are equal’. As the result is not significant, the assumption has been met. Remember that for large sample sizes you are likely to get a significant result when the difference is small just compare SD’s.

3) **Main t-test results**

Null hypothesis: ‘There is no difference in the mean birthweights of babies whose mothers smoke and don't smoke’. The main t-test output is in this table.

*Use the pooled row if the standard deviations are similar. If one is more than double the other use the Satterthwaite row.*



Here the pooled row is used. As the p-value of 0.0465 is less than 0.05, the null is rejected and we can conclude that there is statistically significant evidence to suggest a difference between the mean birthweights for the two groups. When reporting, you should quote the test statistic 2.05, the degrees of freedom of 40 and the p-value of 0.0465.

4) **Describe any differences:** If there is significant evidence of a difference, follow up by describing the difference and a confidence interval. The following table shows the means and confidence intervals for the two groups but also the mean difference.

A Confidence Interval acknowledges that different samples of babies would give different results so gives a range of values within which the population mean is expected to lie. Here the 95% Confidence interval for the difference between the means (Non-smoker – smoker) is (0.01, 0.61) so in the general population we would expect the mean difference to be somewhere between 0.01 lbs and 1.61 lbs.

# Reporting t-tests

# *An independent t-test showed significant evidence (t(40)=2.0545, p < 0.465) that there is a difference in the birthweight of babies whose mothers smoke and don't smoke. Smokers have babies who weigh 0.81 lbs less on average, 95% CI (0.01, 1.61).*

**Wilcoxon rank sum in SAS**

**Dependent variable:** Ordinal orcontinuous data which is very skewed

**Independent variable**: Binary (2 independent groups)

If the data are very skewed or the dependent variable is ordinal, a Wilcoxon rank sum test should be carried out instead of the independent t-test and the medians rather than the means used to compare the two groups. We will demonstrate here with the same data as above although in this case the test is unnecessary as the assumption of normality has been met.

The procedure for carrying out this test is **PROC** **NPAR1WAY** and wilcoxon needs to be specified in the top row. The other lines (class and var) are the same as the ttest procedure.

**PROC** **NPAR1WAY** data=birth wilcoxon correct=no;

class smoker;

var Birthweight;

format smoker smokeS.; **run**;

The Wilcoxon test ranks everyone in the data set from smallest to largest e.g. the lightest baby would have a rank of 1 and the largest a rank of 42. The ranks for the two groups are summed in the ‘*Sum of scores’* column and the average rank for each group in the *‘Mean Score’* column. It’s clear that the mean rank for smokers (18) is much smaller than the mean rank for Non-smokers (25).

A test statistic RS (502.5) can be compared to the special Wilcoxon tables but SAS also provides a test statistic which has been approximated to the normal distribution (Z=1.83) and corresponding two tailed p-value 0.068 which is not significant. If you have a small sample size (N < 30), use the t approximation p-value instead. For significant results, calculate the medians for the groups using **PROC** **MEANS**.

Report: *There is no significant evidence (p=0.068) to suggest a difference in the birthweight of babies born to smokers and non-smokers.*