statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-MannWhitR

The following resources are associated:
Independent samples t-test, Excel file 'Leg Ulcer.csv' and 'Summarising continuous variables by group'

Mann-Whitney U test in R
(Non-parametric equivalent to independent samples t-test)

The Mann-Whitney U test is used to compare whether there is a difference in the dependent variable for two independent groups. It compares whether the distribution of the dependent variable is the same for the two groups and therefore from the same population. The test ranks all of the dependent values i.e. lowest value gets a score of one and then uses the sum of the ranks for each group in the calculation of the test statistic.

Dependent: Numerical/continuous (skewed) or ordinal

Independent: Nominal (binary)

Data: Leg Ulcer data

This data was collected from a randomised controlled trial on patients with leg ulcers which aimed to compare a new treatment regime in the clinic with usual care at home. One of the variables of interest was the number of weeks patients remained ulcer free (UFW).

Research question: Is there a difference between the mean number of ulcer free weeks for the control and intervention groups?

Patients are in one of two groups (GROUP = Clinic or home) and the mean difference in ulcer free weeks (UFW) is of interest so an independent t-test is appropriate. However, one of the assumptions of an independent t-test is that the dependent variable needs to approximately normally distributed for both groups.

Open the leg ulcer comparison dataset from a csv file, call it ulcerR then attach the data so just the variable name is needed in commands.

© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield and Sheffield Hallam University

Reviewer: Jane Candlish
University of Sheffield

```
ulcerR<-read.csv("D:\\stcp-Rdataset-LegUlcer.csv",header=T,sep=",")

attach(ulcerR)
```

Tell R that 'Group' is a factor and attach labels to the categories e.g. 1 is an individual in the Clinic and 2 is a patient in the Home group.

```
GROUP<-factor(GROUP,c(1,2),labels=c('Clinic','Home'))
```

Before carrying any analysis, summarise the medians and interquartile range by group. When reporting differences between groups for skewed data, it is common to report the medians by group rather than the means and the interquartile range as a measure of spread.

Calculate medians and interquartile ranges for UFW by group using the `tapply(dependent, independent, summary statistic required, na.rm=T)` command e.g. `tapply(UFW,GROUP,median,na.rm=T)`. note: na.rm=T removes rows with missing values.

The median for the treatment group (Clinic) is considerably bigger than the 'Home' group. For more information on calculating summary statistics in R, check the '*Summarising continuous data by group in R'* resource.

```
        medians iqrs
Clinic    20.0 38.0
Home       3.1 31.6
```

Assumptions

The only assumptions for carrying out a Mann-Whitney test are that the two groups must be independent and that the dependent variable is ordinal or numerical (continuous). However, in order to report the difference between groups as medians, the shape of the distributions of the dependent variable by group must be similar. It doesn't matter if the distributions have a different location on the x-axis, they just have to be a similar shape.

To produce histograms of the dependent variable by the independent variable, first specify that two charts are needed in one graph window.

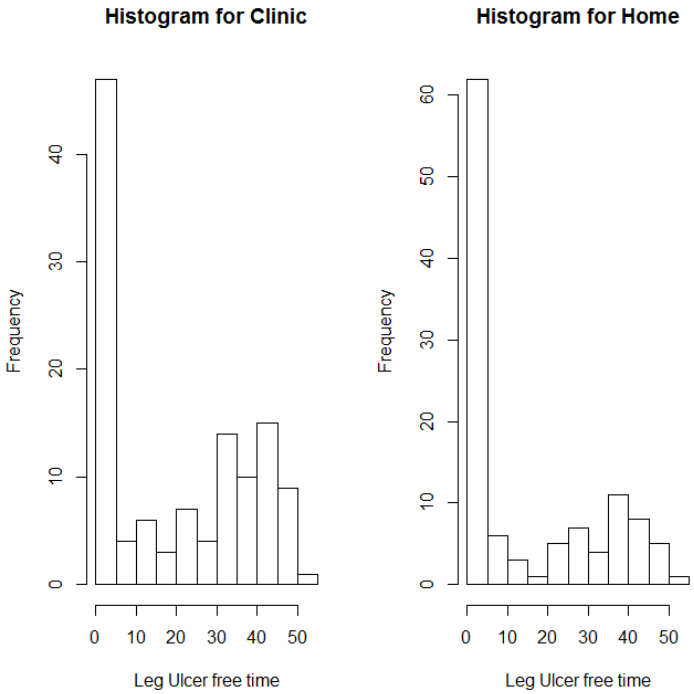
```
par(mfrow=c(1,2))
```

Plot histograms for the dependent ulcer free weeks by group

```
hist(UFW[GROUP=='Clinic'],main='Histogram for Clinic',xlab='Leg Ulcer free time')
```

```
hist(UFW[GROUP=='Home'],main='Histogram for Home',xlab='Leg Ulcer free time')
```

The histograms show that the two distributions have a similar pattern, they are both positively skewed, so the medians can be used to summarise the differences for the ulcer free weeks. If the two histograms looked different, differences in the mean ranks rather than medians would be summarised.



Conducting the Mann-Whitney U test in R

The Mann-Whitney U tests the null hypothesis 'There is no difference between the leg ulcer free weeks for the Clinic group compared to the group receiving the standard treatment'. The null is rejected if the p-value for the t-test is less than 0.05. Use the `wilcox.test(dependent~independent)`. By default it conducts the Mann Whitney U Test. On the left side of the formula place the continuous variable (leg ulcer free weeks) and place the group on the right.

```
> wilcox.test(UFW~GROUP)

        Wilcoxon rank sum test with continuity correction

data:  UFW by GROUP
W = 7964, p-value = 0.017
alternative hypothesis: true location shift is not equal to 0
```

The key bits of information in the table are the W-statistic, *W=7964* and the *p-value=0.017.* The Wilcoxon W is simply the lowest sum of ranks but in order to calculate the p-value (Asymp. Sig), R uses an approximation to the standard normal distribution and also makes a continuity correction. The approximation is less reliable for small sample sizes.

Reporting a Mann-Whitney test

A Mann-Whitney U test showed that there was a significant difference (W= 7964, p = 0.017) between the leg ulcer free weeks for the Clinic group compared to the group receiving the standard treatment. The median ulcer free weeks was 20 weeks for the Clinic group compared to 3.1 weeks for those receiving the standard treatment at home.