statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-indepR

The following resources are associated:
Independent t-test in R script, Checking normality in R and the dataset 'Birthweight reduced.csv'

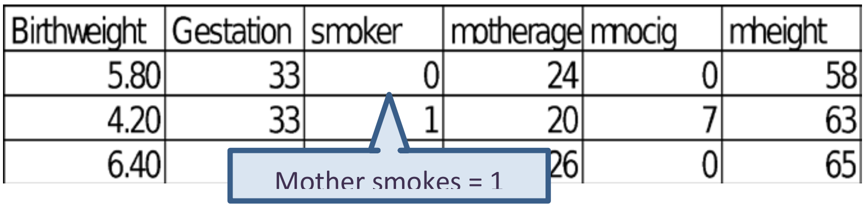
Independent t-test in R

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 has an effect on the birthweight of a baby so the dependant variable is Birth weight (lbs) and the independent variable is whether or not the mother smokes (smoker).



Birthweight	Gestation	smoker	motherage	mnocig	mheight
5.80	33	0	24	0	58
4.20	33	1	20	7	63
6.40			26	0	65

Mother smokes = 1

Download and save the '*Birthweight_reduced'* csv on your computer. Open the birthweight reduced dataset from a csv file, call it birthweightR then attach the data so just the variable name is needed in commands.

```
birthweightR<-
read.csv("D:\\Birthweight_reduced.csv",header=T,sep=",")
attach(birthweightR)
```

Tell R that 'smoker' is a factor and attach labels to the categories e.g. 1 is a smoker.

```
smoker<-factor(birthweightR$smoker,c(0,1),labels=c('Non-
smoker','Smoker'))
```

Before carrying any analysis, summarise birthweight (Birthweight) by smoking (smoker) using some summary statistics. Do the group means and standard deviations look similar or very different?

Calculate means and standard deviations for birthweight by smoker using the

```
tapply(dependent, independent, summary statistic required,
```

`na.rm=T)` command e.g. `tapply(Birthweight,smoker,mean,na.rm=T)`. na.rm=T removes rows with missing values.

```
mean<-tapply(Birthweight,smoker,mean,na.rm=T)
sd<-tapply(Birthweight,smoker,sd,na.rm=T)
```

Combine the results in one table and give it a name.

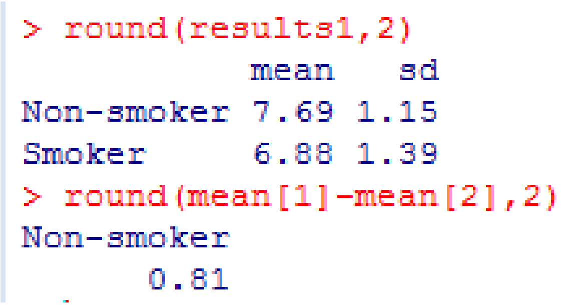
```
results1<-cbind(mean,sd)
```

Display the results rounding the statistics to 2 decimal places.

```
round(results1,2)
```

Use the results to calculate the difference between the means (Non-smoker – smoker) rounded to 2 decimal places.

```
round(mean[1]-mean[2],2)
```



```
> round(results1,2)
            mean   sd
Non-smoker  7.69 1.15
Smoker      6.88 1.39
> round(mean[1]-mean[2],2)
Non-smoker
      0.81
```

The mean birthweight for babies of smokers is 0.81 lbs lower than the mean for non-smokers. The standard deviations are similar so the groups are equally spread out.

Checking the assumptions

Assumptions	How to check	What to do if the assumptions is not met
The dependent variable should be approximately normally distributed for each group	Use histogram, QQ plots and normality tests by group For more details see the *Checking normality in R* resource.	If the data for either group is very skewed, use the Mann-Whitney test
Homogeneity (equality) of variance: The variances (SD squared) should be similar for all the groups	If one SD is more than twice the other, the assumption has not been met. Levene's test for equal variances could also be used through the package car `leveneTest()`. If p - value > 0.05 then equal variances can be assumed	If one SD is more than twice the other, the results of the t-test are less reliable. Use `var.equal=TRUE` in the t test command

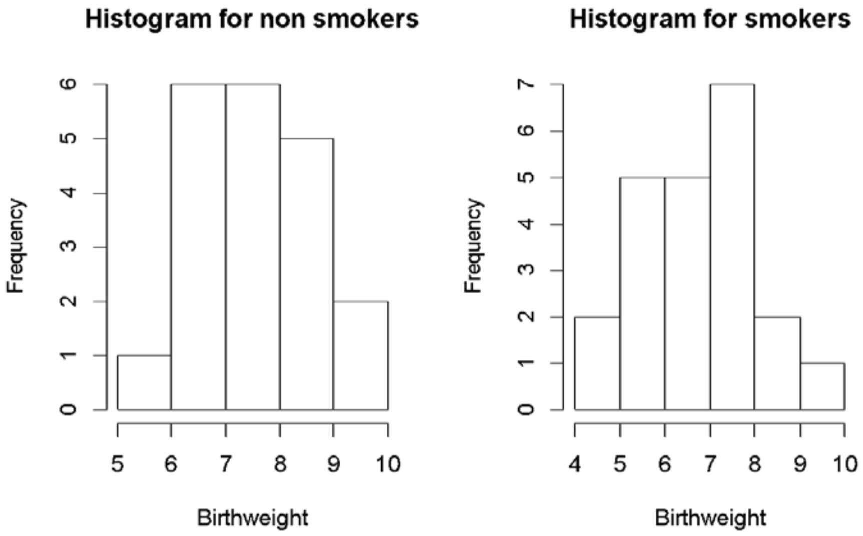
Checking the assumptions for this data

To check the assumption of normality, produce histograms of the dependent by the independent. First specify that two charts are needed in one graph window.

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

Plot histograms for the birthweight of babies of smoking and non-smoking mothers

```
hist(Birthweight[smoker=='Non-smoker'],main='Histogram for non
smokers', xlab='Birthweight')
hist(Birthweight[smoker=='Smoker'],main='Histogram for
smokers',xlab='Birthweight')
```



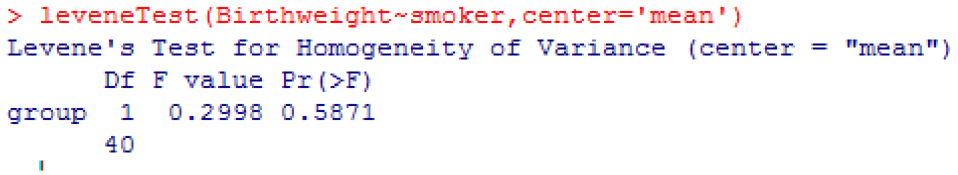
Both histograms are approximately normally distributed so the assumption has been met. QQplots or normality tests could also be used to assess normality (see *checking normality in R* resource).

The standard deviations are similar so the assumption of equal variances has been met. If you wish to perform Levene's test to check the same assumption, the Levene's test command `leveneTest` is part of the library car. In order to install the library go to *Packages>Install Packages>...* and find the library car. As soon as it finishes installing, load it using the command `library(car)` (remember that a library is only loaded for that session and will need to be loaded each time R is opened).

```
library(car)
```

Once loaded, carry out Levene's test

```
leveneTest(dependent, independent,location='mean').
```



```
> leveneTest(Birthweight~smoker,center='mean')
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.2998 0.5871
      40
```

The Levene's test p-value=0.58 which is greater than 0.05 and so equality of variances for the two independent samples can be assumed and the t-test used.

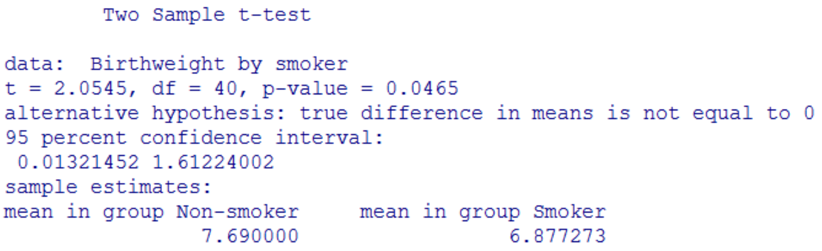
Note: Rstudio currently has some issues with not all commands in the car package working. An alternative is available in the lawstat package. Load through Tools --> install packages.

```
library(lawstat)
levene.test(Birthweight,smoker)
```

Steps for the independent t-test in R and output

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'. The null is rejected if the p-value for the t-test is less than 0.05 and statistically significant evidence of a difference concluded.

Use the `t.test(dependent~independent,var.equal=TRUE)` command if equal variances can be assumed. If the Levene's test is significant, use var.equal=FALSE.
`t.test(Birthweight~smoker,var.equal=TRUE)`



```
        Two Sample t-test

data:  Birthweight by smoker
t = 2.0545, df = 40, p-value = 0.0465
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01321452 1.61224002
sample estimates:
mean in group Non-smoker     mean in group Smoker 
                7.690000                 6.877273 
```

The key bits of information in the table are the t-statistic, *t=2.0545* and the p-value=*0.0465.* As the p-value < 0.05, the null is rejected and statistically significant difference in birth weight between the two groups is concluded. The mean difference between the groups was found to be 0.81lbs with smokers having the lighter babies.

Confidence intervals

Sometimes a confidence interval for this difference is reported. 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

There is 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. In this dataset, smokers have babies who weigh 0.81 lbs less on average, 95% CI (0.01, 1.61).

Meaningful differences and effect sizes

Another thing to consider is whether the difference in means is meaningful as with large samples very small differences will be classified as significant. Background knowledge of the subject is needed to decide this if a difference is meaningful. The average weight of a baby is about 7 lbs so a difference of 0.05lbs would not be meaningful but a difference of a pound would be. You may also wish to calculate an effect size to assess the magnitude of the difference between means. Eta-squared can be calculated using the tests statistic and degrees of freedom from the output:

$$\text{Eta squared} = \frac{\mathrm{t}^2}{\mathrm{t}^2 + df} = \frac{2.0545^2}{2.0545^2 + 40} = 0.31$$

Cohen (1988) using the following guidelines for interpretation: 0.01 (small effect), 0.06 (moderate) and 0.14 (large effect). Here the magnitude of the differences in the means was large (eta squared = 0.31).