statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-contR

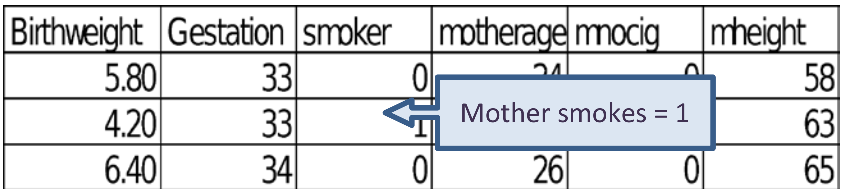
The following resources are associated: Birthweight_reduced.csv dataset, Summarising continuous variables in R script and Independent t-test in R resource

Summarising Continuous Variables in R

Dependent variable: Continuous

Independent variable: Categorical

Data: The data set 'Birthweight reduced.csv' contains details of 42 babies and their parents at birth e.g. birthweight, mothers age and whether or not the mother smokes (smoker). Download the '*birthweight_reduced.csv*' dataset and store it on your computer.

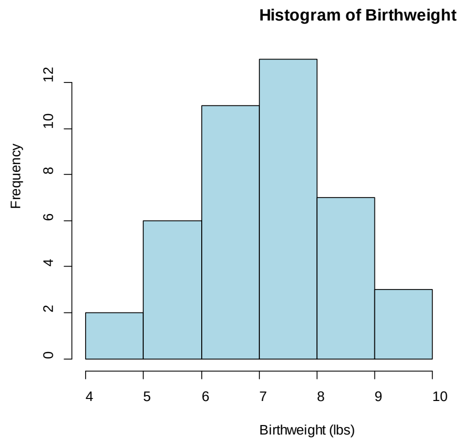
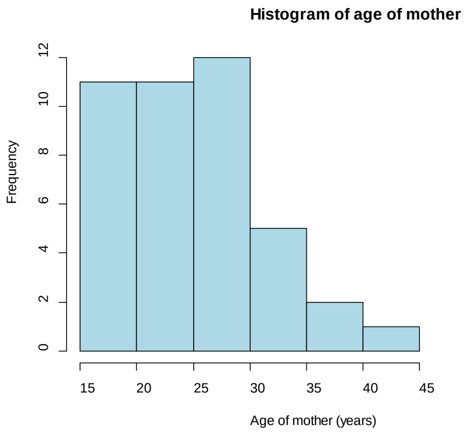


Birthweight	Gestation	smoker	motherage	mnocig	mheight
5.80	33	0	24	0	58
4.20	33	1	[illegible]	[illegible]	63
6.40	34	0	26	0	65

Mother smokes = 1

Choosing the right summary statistics

Different summary statistics are appropriate depending on the distribution of the data. Use histograms to check if the data is approximately normally distributed.

Normally distributed data	**Skewed data**
	
Average: Mean **Measure of spread**: Standard deviation Similar mean and median	**Average**: Median **Measure of spread**: Interquartile range Very different mean and median

Research question: Does a mother smoking have an effect on the birthweight of a baby? The dependant variable is Birth weight (lbs) and the independent variable is whether or not the mother smokes (smoker).

Open the birthweight reduced dataset from the place you have saved it, call it *birthweightR* then attach the data so just the variable name is needed in commands. Note: here it has been saved on a memory stick which is the D drive.

```
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(smoker,c(0,1),labels=c('Non-smoker','Smoker'))
```

Summary Statistics

There are several options for summarising continuous variables including the `summary()` and the `tapply()` command which are discussed here.

To calculate a range of summary statistics for birthweight use `summary(birthweight)`

To calculate these summary statistics by group

```
Smoking<-summary(Birthweight[smoker=='Smoker'])
Non_smoking<-summary(Birthweight[smoker=='Non-smoker'])
```

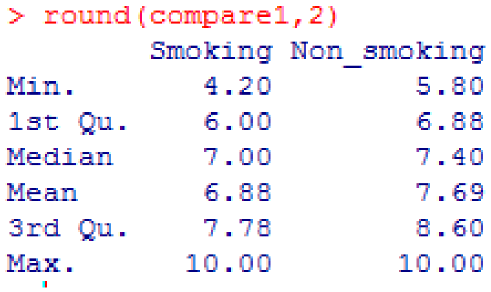
Combine the results into one table and give it a name

```
compare1<-cbind(Smoking,Non_smoking)
```

Then reduce the decimal places to 2.

```
round(compare1,2)
```

The means and medians for each group are similar suggesting that the data is approximately normal so the means and standard deviations are appropriate.



```
> round(compare1,2)
        Smoking Non_smoking
Min.       4.20        5.80
1st Qu.    6.00        6.88
Median     7.00        7.40
Mean       6.88        7.69
3rd Qu.    7.78        8.60
Max.      10.00       10.00
```

When data are skewed use the medians and interquartile range e.g. the median birthweight for mothers who smoked is 7 lbs (IQR: 6, 7.78). This means that the middle 50% of the data for smokers ranges from 6 to 7.78 lbs.

The summary method does not produce standard deviations automatically but customised summary statistics can be calculated using the `tapply(dependent, independent, summary statistic required, na.rm=T)` command where na.rm=T removes rows with missing values.

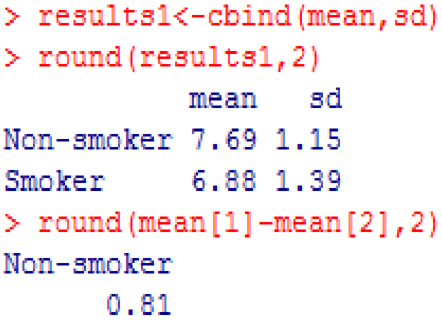
Calculate the means and standard deviations by group

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

Combine the results in one table and round the statistics to 2 decimal places.

```
results1<-cbind(mean,sd)
round(results1,2)
```

Use the results to calculate the difference between the means rounded to 2 decimal places.



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

Do the group means and standard deviations look similar or very different?

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. The variance can be derived with the `var()` command and for skewed data the `median()` and `IQR()` commands can be used.

Data Visualisation

To display the information graphically, use either histograms or boxplots per group. For a basic histogram with frequencies (`probability=F`) on the y-axis:

```
hist(variable,main='Title',xlab='x label',probability=F,col="colour")
```

To plot a histogram of with densities (probabilities) on the y-axis (needed if you are adding a normal curve), use `probability=T`. Colours are referred to by words e.g. “lightblue” .

To change the number of bars, use the breaks command. You can specify a number e.g. `breaks=5` or the break points e.g. `breaks =c(6,7,8,9,10).`
The scale on the x and y axis can be specified using `xlim=c(min,max) or ylim.`

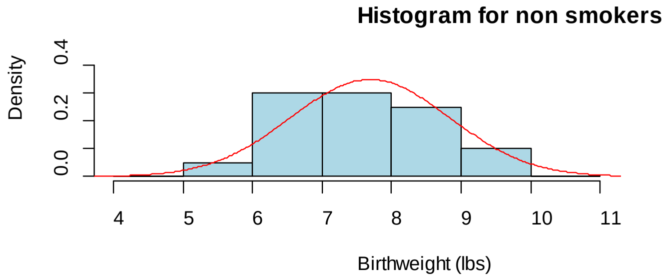
To produce histograms along with the normal curve for the two different groups, first specify that two charts are needed next to each other using `par(mfrow=c(2,1))`

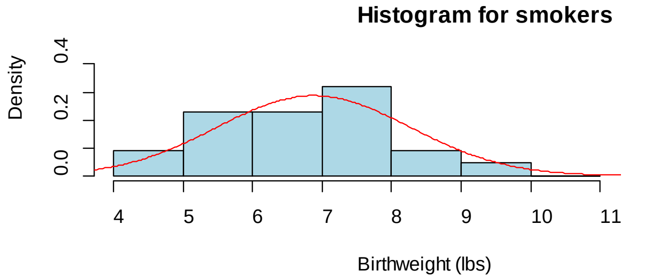
To produce a histogram of birthweights for non-smokers, with densities on the y-axis, light blue bars, a y-axis which ranges from 0 to 0.4 and breaks of 1 lb between 5 and 11 lbs:

```
hist(Birthweight[smoker=='Non-smoker'],main='Histogram for non smokers',xlab='Birthweight',probability=T,col="lightblue",ylim=c(0,0.4),breaks=c(5,6,7,8,9,10,11))
```

To help assess normality, a smooth version of the histogram called a density line can be added to the plot using `lines(density(Birthweight[smoker=='Non-smoker'].`Alternatively, a perfect normal curve can be added by generating many data points from a normal distribution with the same mean and variance as the variable.

```
lines(density(rnorm(n=10000000,mean=mean(Birthweight[smoker=='Non-smoker']),sd=sd(Birthweight[smoker=='Non-smoker']))),col=2)
```





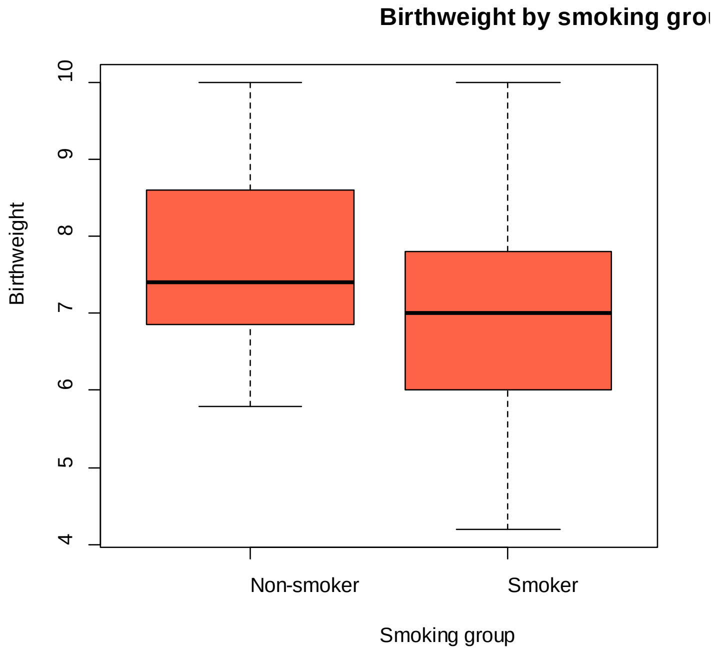
Repeat the histogram and density line procedure for the birthweight of babies whose mothers’ smoked.

Histograms can be helpful in order to assess normality and help choose the appropriate summary statistics and test.

The two histograms show that birthweight is approximately normally distributed for the babies of smokers and non-smokers. So it is appropriate to use means and standard deviations to describe the data and an independent t-test to look for a significant difference between the groups.

To visually show any comparison between the two groups, however, boxplots are more appropriate. They are derived using the command `boxplot(dependent~independent)`. Therefore the command that we used in this case is

```
boxplot(Birthweight~smoker,col='tomato',main='Birthweight by smoking
group of mother',xlab='Smoking group',ylab='Birthweight')
```



The boxplot represents the spread of birthweight for the smoker and non-smoker groups. The median is represented by the line in the middle of the box. The box limits represent the first quartile and the third quartile, and thus represents the middle 50% of the data called the “Inter Quartile Range”. The longer the box, the more spread out the data are. The bottom and top whiskers represent the minimum and maximum values or 1.5 times the Interquartile range if there are outliers.

Non-smoker mothers seem to have given birth to slightly heavier babies than smoker mothers. The smoker box contains a bigger range of values, which suggests a higher variation in the birth weight of babies born from a mother who smokes.

Tips on reporting

Do not include every possible summary from the `summaries()` command.

Think back to the key question of interest and answer this question.

Briefly talk about every chart and table you include but don’t discuss every number if the table is included.

Use the median and interquartile range if the data are very skewed.