Basic Statistics Using R – day3

Feb 7, 2020

Slide and R code File

https://github.com/npmldabook/IntroR

Look for

- Slides_Day3.pdf
- Day3_CODE.R

Day 2 : Several Questions about reading and writing R data sets

Method 1. Go to the Environment Pane/Tab => Click Import Datasets => Read Text data or Excel data

Method 2. Use R script to read and write data

See demo, code in "Day3_CODE.R"

Day 3 study topics

We use slightly different order : all R codes including updated ones in **Day3_CODE.R**

- DOE basics and statistical foundation/Inferences
- Correlation of Variables
- Regression
- Commonly used statistical tests

Design of experiment basics: Where did the data originally come from ?

- Statistics is often taught as though the design of the data collection and the data cleaning have already been done in advance.
- However, as most practicing statisticians quickly learn, typically problems that arise at the analysis stage, could have been avoided if the experimenter had consulted a statistician before the experiment was conducted to collect the data. A well designed studied is usually simple to analyze and interpret.
- Correlation (or association) is not causality. You've probably heard that before in any regression class. If you want to infer causality from data, then the best way is to use randomized experiments.

Type of Scientific Studies : Experimental vs. Observational studies

In a **comparative experimental study**, randomization is employed to assign a set of treatments to the experimental units, and the observed ourcomes among the treatment groups are compared to assess treatment effects.

- Cause-and-effect relationships between the experimental factors (predictors, Xs) and the outcome or response variable (Y) can be established in an experimental study.
- Ex 1: A experiment was conducted to study the effect o f baking temperature on the volume of a quick bread prepared from a package mix. Four oven temperatures-Iow. medium, high, and very high-were tested by **randomly** assigning each temperature to five package mixes. (n=20)

– Ex2: randomized clinical trials (RCTs): patients were randomized to be in the new treatment group vs. the control (e.g. standard care, placebo groups) to determine if the new treatmet is better than the control group; commonly used to establish one treatment is better than the control/placebo; gold standard for drug approval by FDA.

10,251 participants with type 2 diabetes and cardiovascular disease or additional risk factors for cardiovascular disease was randomly assigned to either standard therapy (n=5,123) targeting HbA1c levels of 7.0%-7.9% or intensive therapy (n=5,128) targeting HbA1c $< 6.0\%$.

Observational studies

Randomization of the treatments/study factors for the experimental units does not occur.

- Ex : In the quick bread example, if the four temperatures are not randomized, then it is an observational study.

Prospective observational study

- In a prospective observational study (cohort study), one or more groups are formed in a nonrandom manner according to the levels of a hypothesized causal factor, and then these groups are observed over time with respect to an outcome variable of interest. e.g. Small or large cohort studies. It answers the question: "What is going to happen?"
- Framingham Heart Study(FHS). Launched in 1948, these studies involve studying the health of various populations to uncover patterns, trends, and outcomes to identify common factors or characteristics that contribute to cardiovascular disease.
- Factors: high vs. low blood pressures, high vs. low lipids, diabetes or not.
- Outcome: cardiovascular disease, heart attack, cancer …

Retrospective observational study

Groups are defined on the basis of an observed outcome, and the differences among the groups at an earlier point in time are identified as potential causal effects. It answers the question: "What has happened?"

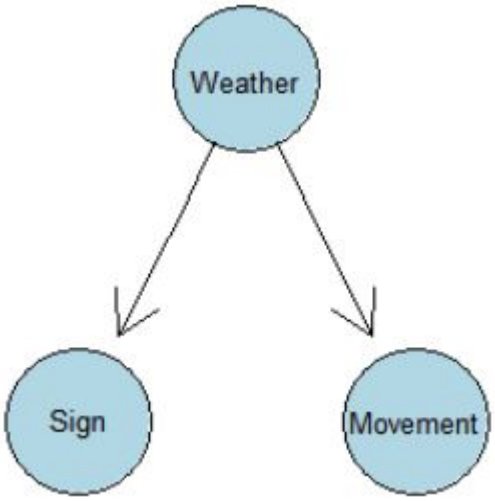
– Also known as the **case-control studies**: save study time when when an outcome of interest occurs infrequently.

– A retrospective of cancer study would identify persons who have a certain cancer (cases) and a matching persons who do not have this cancer (controls) and look back in time to assess differences in their risk exposure variable. For example, finding the lung cancer cases and matching nondisease controls, then collect their smoking history and environmental factors.

Mostly analysis can only establish the corelation or association

Unless we are comparing the outcome among the the randomization groups, most analysis including regressions are based on observational studies (factors are not randomized assigned) and can only establish association.

In such analysis, a potential danger is the existence of confounding variables (confounders). A confounder is a common cause for two variables.



- Is the seatbelt sign on an airplane causing a plane to shake? If we could switch it on ourselves, would the plane start shaking?
- Turbulent weather at the same time makes the pilot switch on the seatbelt sign and the plane shake. What we observe is an *association* between the appearance of the seatbelt sign and a shaking plane. The seatbelt sign is not a *cause* of the shaking plane.

Fundamental Statistical Concepts (p.46)

Statistical inference mostly include estimation and hypothesis testing.

Estimation of parameter and confidence Intervals: for model parameters, regression coefficients.

Hypothesis testing (H_0 vs. H_a): A statistical hypothesis is a statement either about the parameters of a probability distribution or the parameters of a model.

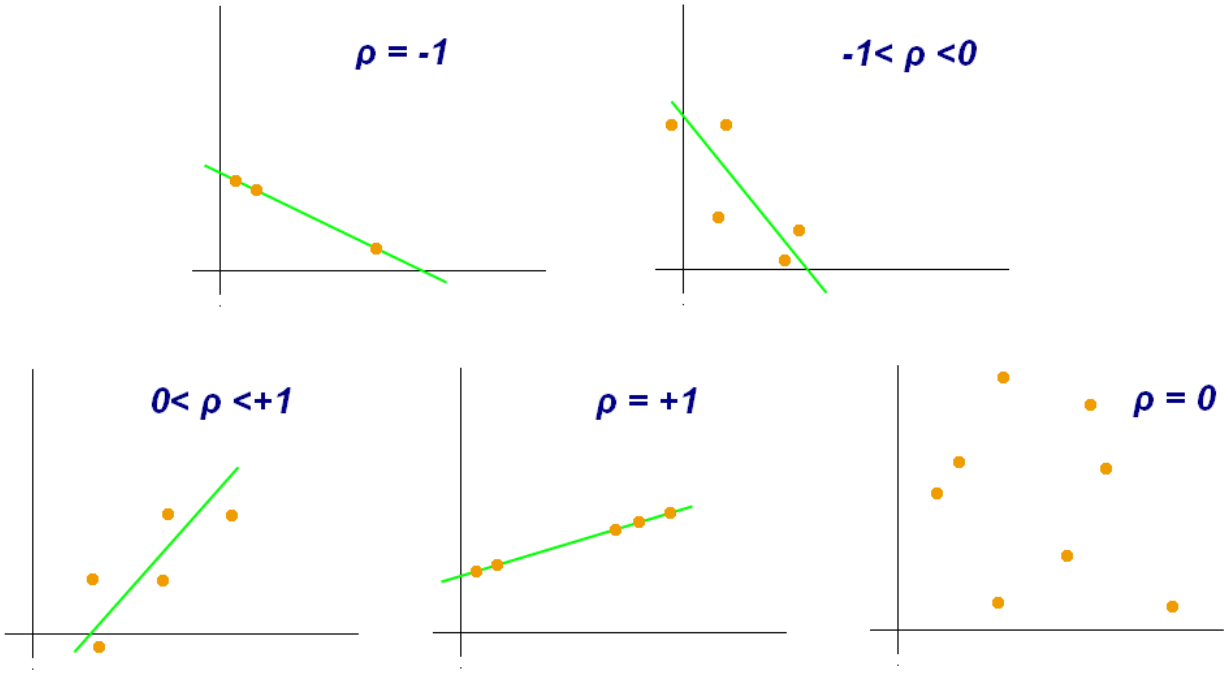
- Null vs. Alternative hypotheses (e.g. making initial assumptions of no difference in treatments vs. not)
- **Test statistic and decision rule (rejection rule)**: To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate test statistic, and then rejecting or failing to reject the null hypothesis H0 based on the computed value of the test statistic.

		Truth	
		Null Hypothesis	Alternative Hypothesis
Decision	Do not Reject Null	OK	Type II Error
	Reject Null	Type I Error	OK

- Type I (α) and type II error (β):
  - If the null hypothesis is rejected when it is true, a type I error has occurred (false positive).
  - If the null hypothesis is not rejected when it is false, a type II error has been made.
  - Power =1- β= probability of rejecting null when the alternative is true.
- **P-value** : is defined as the smallest level of significance that would lead to rejection of the null hypothesis H_0. Once the P-value has been determined, we can make decision if the test result is significant by comparing P-value with the given significance level α (such as 0.05, 0.01). A smaller α provides a stronger evidence against the null or more support for the alternative hypothesis.
  - If P-value $\leq \alpha$ => reject H0 at level α.
  - If P-value $> \alpha$ => do not reject H0 at level α.

Correlation (Sec 5)

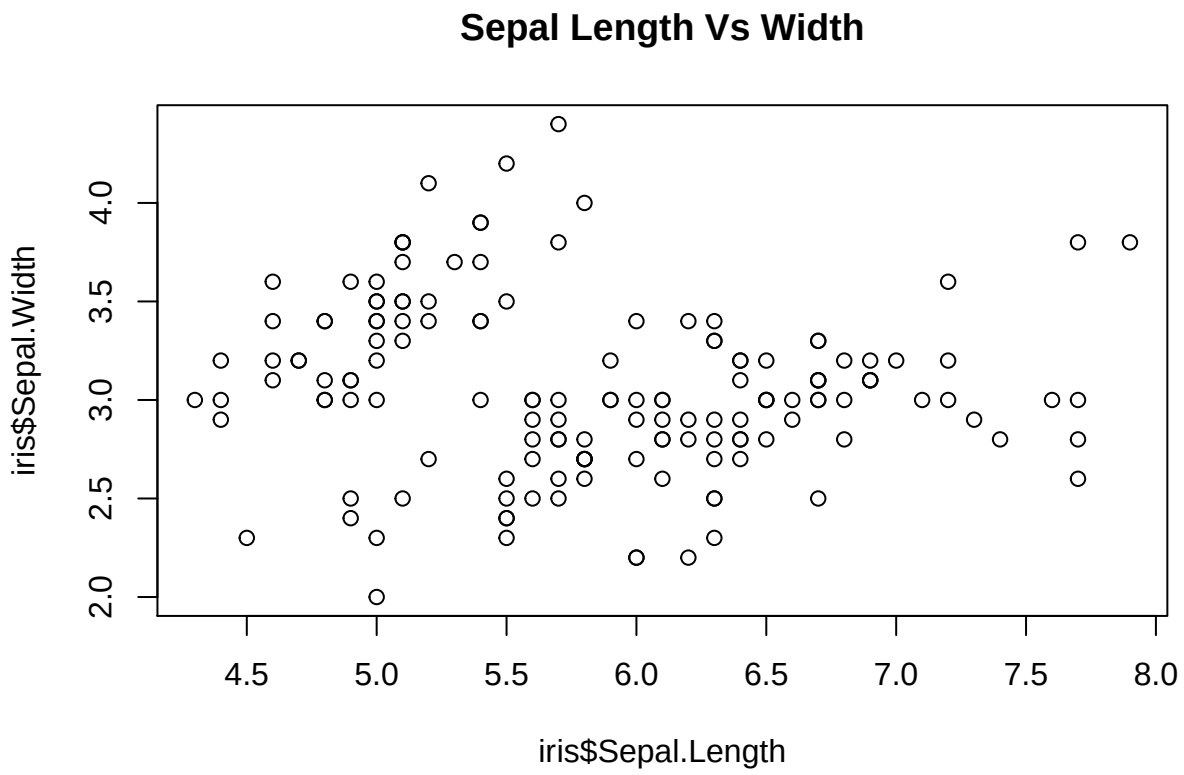
Correlation (range -1, 1) can be used to study the relationship between two variables.



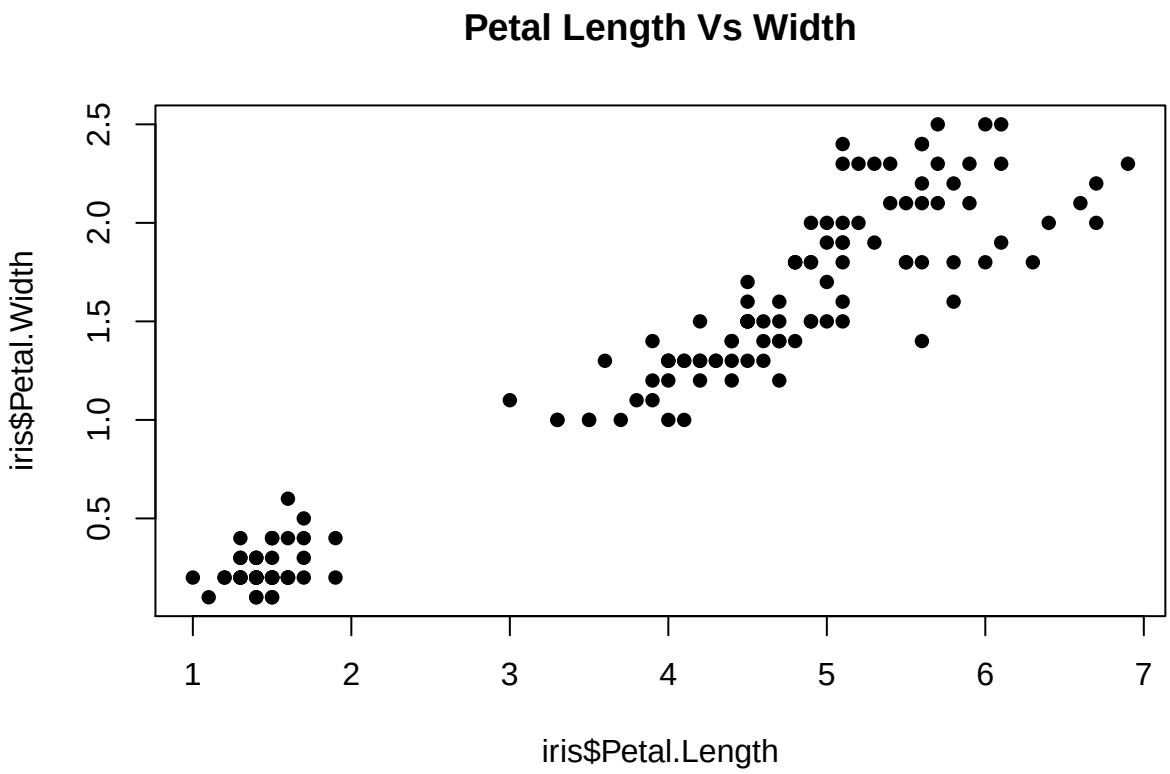
Visualization for correlation (R)

Scatterplots are commons ways to visually inspect for correlations between variables.

```r
#plot the length vs width
plot(iris$Sepal.Length, iris$Sepal.Width, main = "Sepal Length Vs Width")
```



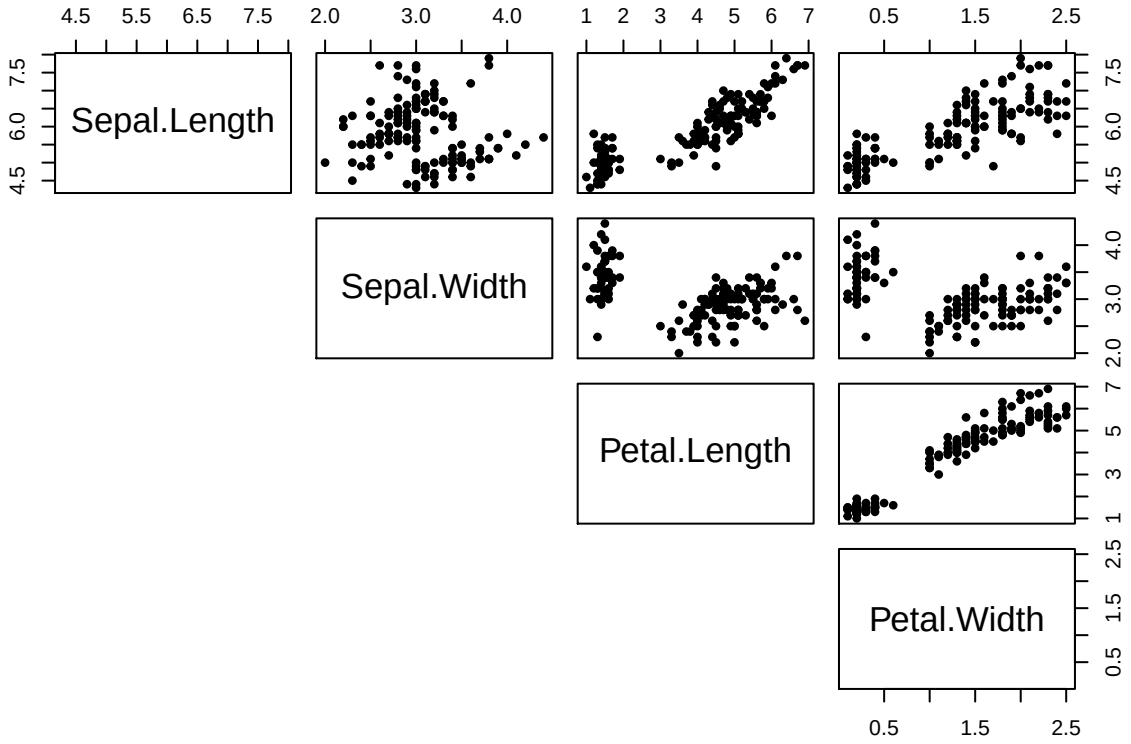

```r
plot(iris$Petal.Length, iris$Petal.Width, main = "Petal Length Vs Width", pch=16)
```




```
#does any of the above plots show any relationship ?
#do you see any linear relationship ?
#is the relationship positive or negative ?
```

We can use R to generate the scatter-plot matrices: pairs() to study >2 variables.

```
pairs(iris[,1:4], pch = 19, cex=0.7, lower.panel = NULL)
```




```
# Did you find some interesting relationship?
```

Correlation coefficients

There are several Correlation (coefficients): Pearson, Spearman, kendall.

Pearson r- common, product-moment, parametric, assumes linear relationship

spearman ρ - rank correlation, non-parametric, no assumption on relationship and distribution

kendall τ - rank correlation, non-parametric, no assumption on the relationship and distribution

The default correlation is Pearson's correlation, which is not robust to outliers.

```
#calculate the correlation coefficient between length vs width for sepal and petal
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.1175698
```

```
cor(iris$Petal.Length, iris$Petal.Width)
```

```
## [1] 0.9628654
```

```
#compare the correlation plots with the correlation coefficient values
```

We can specify " method= " to compute other more robust correlation.

```
cor(iris$Sepal.Length, iris$Sepal.Width, method="spearman")
```

```
## [1] -0.1667777
```

```
cor(iris$Sepal.Length, iris$Sepal.Width, method="kendall")
```

```
## [1] -0.07699679
```

Test for association (R)

We can use cor.test() to test if the correlation is significantly different from 0 (no correlation).

```
#test for association
cor.test(iris$Sepal.Length, iris$Sepal.Width)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Length and iris$Sepal.Width
## t = -1.4403, df = 148, p-value = 0.1519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27269325  0.04351158
## sample estimates:
##        cor
## -0.1175698
```

```
cor.test(iris$Petal.Length, iris$Petal.Width)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Petal.Width
## t = 43.387, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9490525 0.9729853
## sample estimates:
##       cor
## 0.9628654
```

```
cor.test(iris$Petal.Length, iris$Petal.Width, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  iris$Petal.Length and iris$Petal.Width
## z = 13.968, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.8068907
```

```r
cor.test(iris$Petal.Length, iris$Petal.Width, method = "spearman")
```

```
## Warning in cor.test.default(iris$Petal.Length, iris$Petal.Width, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  iris$Petal.Length and iris$Petal.Width
## S = 35061, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9376668
```

Regression Analysis

Linear Regression analysis is commonly used to study the relationship between the independent variable (s), $X_1, X_2, ..., X_n$ and a continuous response variable (Y). Independent variables are also called covariates or predictors, can be continuous or categorical.

- X (explanatory/independent variables, predictors, covariates, input)
- Y (outcome/response/dependent variable, output)

Simple linear regression ($Y \sim X$) :

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Multiple linear regression ($Y \sim X_1, X_2, ..., X_k$) :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i$$

- Usually, we assume the errors (residuals) in the model follow the normal distribution.

Simple linear regression (R)

Example: SLM for iris petal length vs width use **lm()**

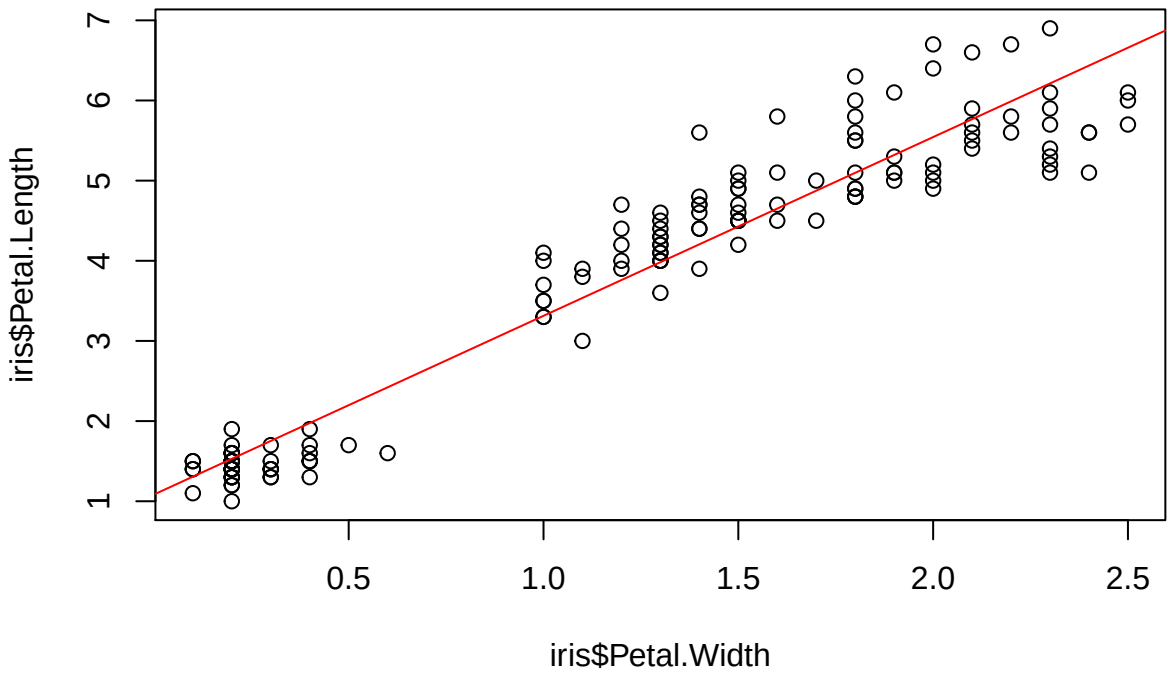
```r
#how width (indepednent variable) predicts length (dependent variable)
#length is response, width is predictor (width predicts length)

simplereg=lm(Petal.Length~Petal.Width, data = iris)

#model Length = a + b*Width
#a and b are predicted coefficients of the model
#lets look at the model and its contents

simplereg
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Coefficients:
## (Intercept)  Petal.Width
##       1.084        2.230
```

```r
summary(simplereg)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33542 -0.30347 -0.02955  0.25776  1.39453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.08356    0.07297   14.85   <2e-16 ***
## Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

```r
#draw the regression line on the plot
plot(iris$Petal.Width, iris$Petal.Length)
abline(simplereg, col=2)
```



We predict the length (Y), given width (X), based on our model.

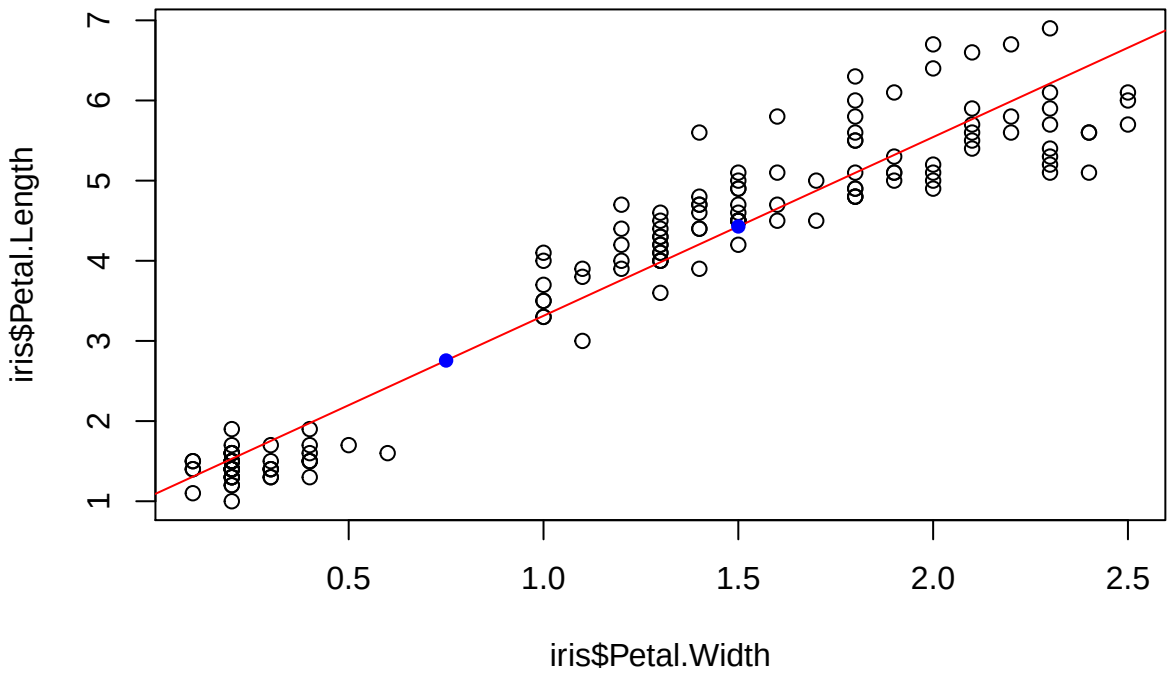
```
#create test width data frame
testwidth=data.frame(Petal.Width=c(0.75,1.5))
testwidth
```

```
##   Petal.Width
## 1        0.75
## 2        1.50
```

```
#predict the length for each of the width in our test dataframe
predict(simplereg, newdata=testwidth)
```

```
##        1        2
## 2.756013 4.428469
```

```
# Adding the predicted points to the linear
plot(iris$Petal.Width, iris$Petal.Length)
abline(simplereg, col=2)
points(c(0.75,1.5), predict(simplereg, newdata=testwidth), pch=16,  col=4)
```



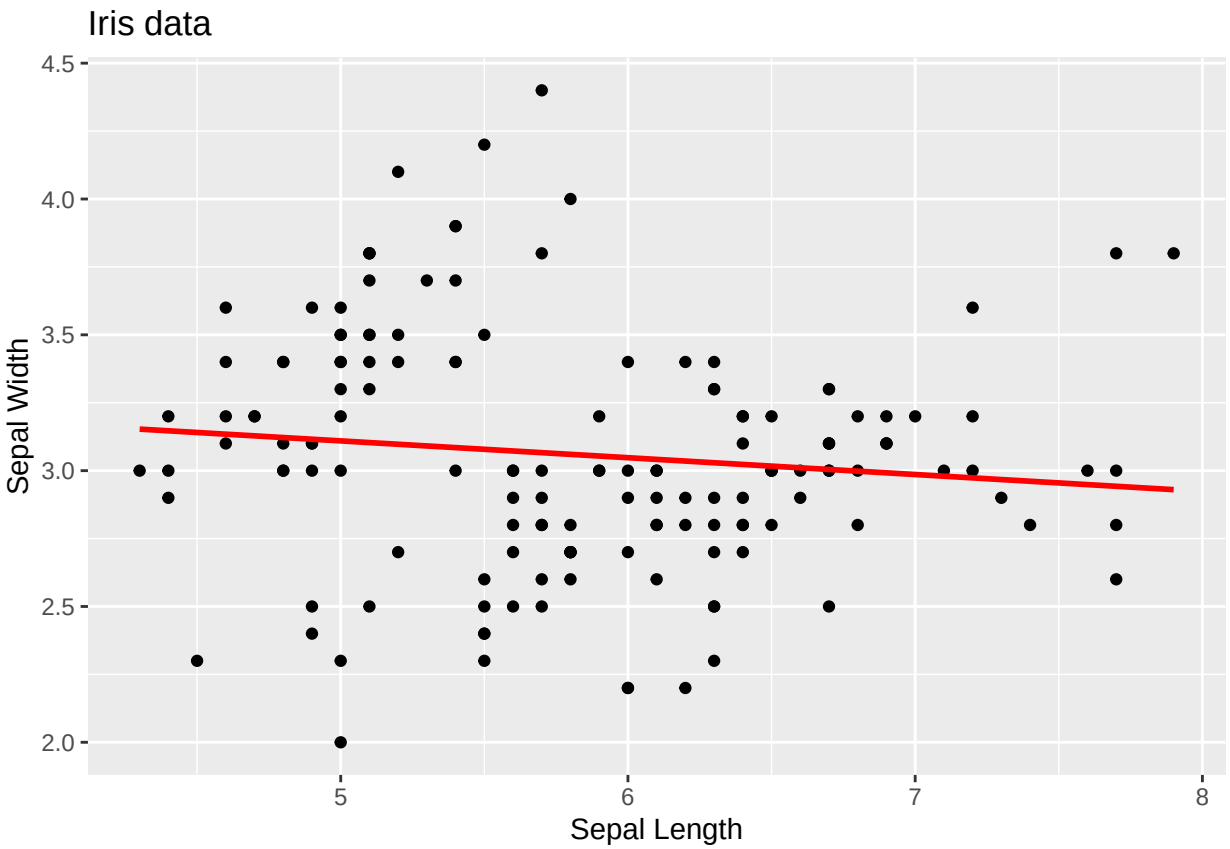

```
# Get prediction with confidence intervals
predict(simplereg, newdata=testwidth,  interval =  "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 2.756013 2.666369 2.845658
## 2 4.428469 4.345487 4.511450
```

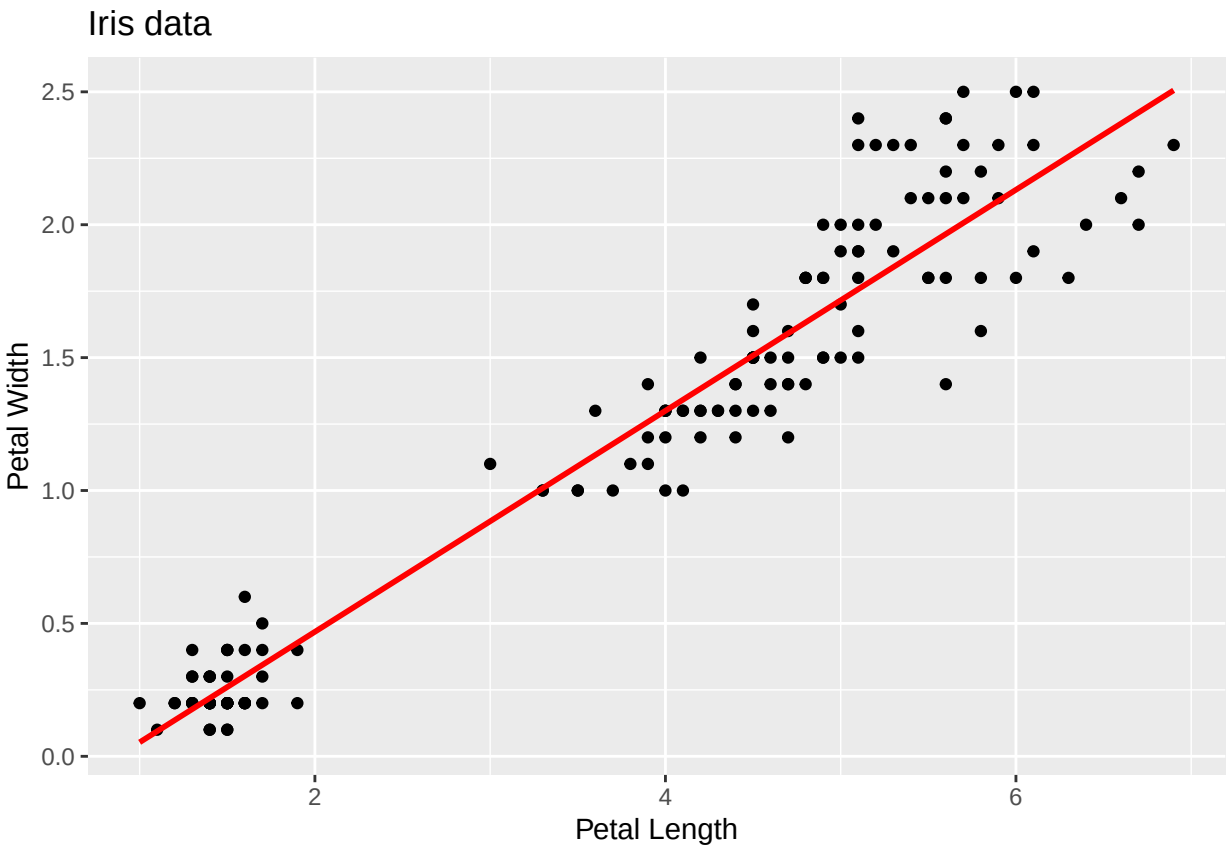
Making scatterplot using ggplot() (R)

The more advanced R graphic function based on **ggplot2** package, makes it much easier to generate scatter plot with a least suqare or simple linear regression line.

```
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width)) +
geom_point()+
stat_smooth(method="lm", col='red', se=FALSE)+
labs(title='Iris data', x="Sepal Length", y="Sepal Width")
```



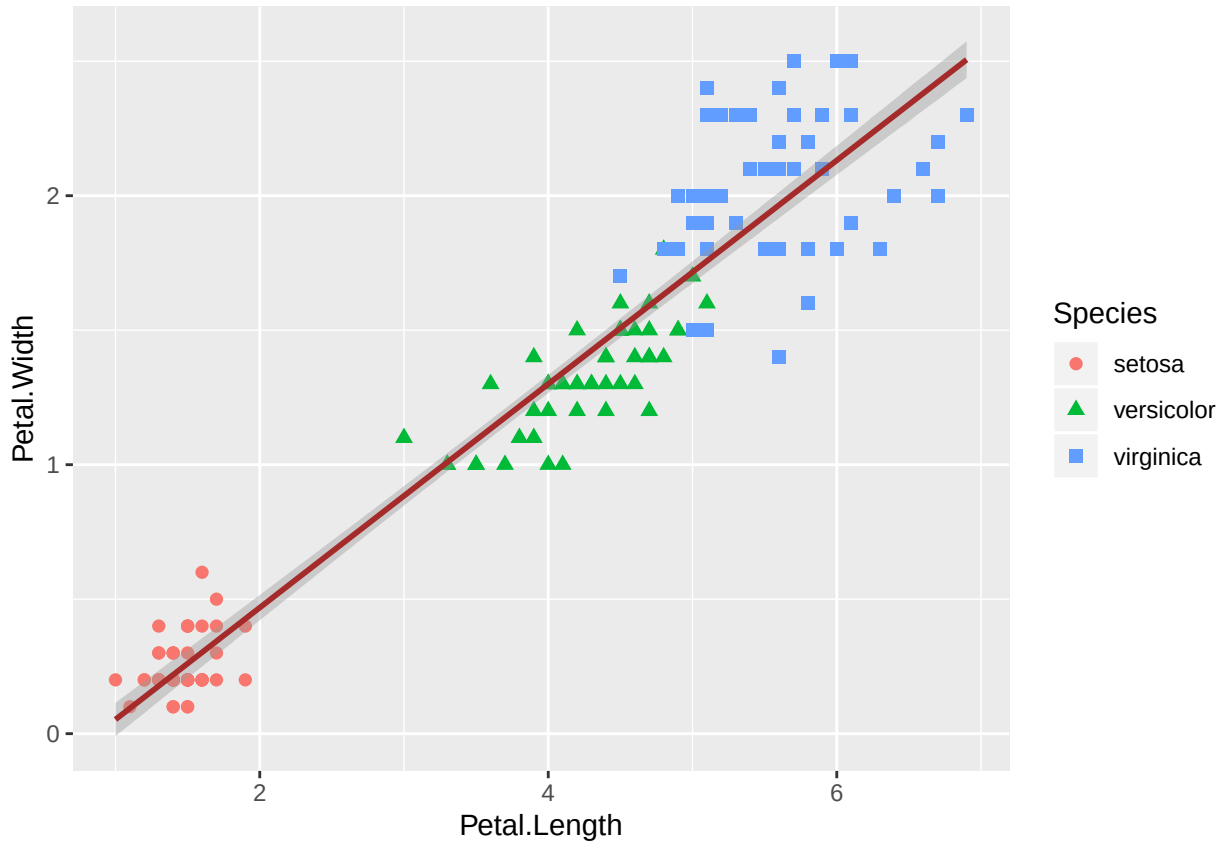

```
ggplot(iris, aes(x=Petal.Length, y=Petal.Width)) +
geom_point()+
stat_smooth(method="lm", col='red', se=FALSE)+
labs(title='Iris data', x="Petal Length", y="Petal Width")
```



Making scatterplot using ggplot() with subgroups (R)

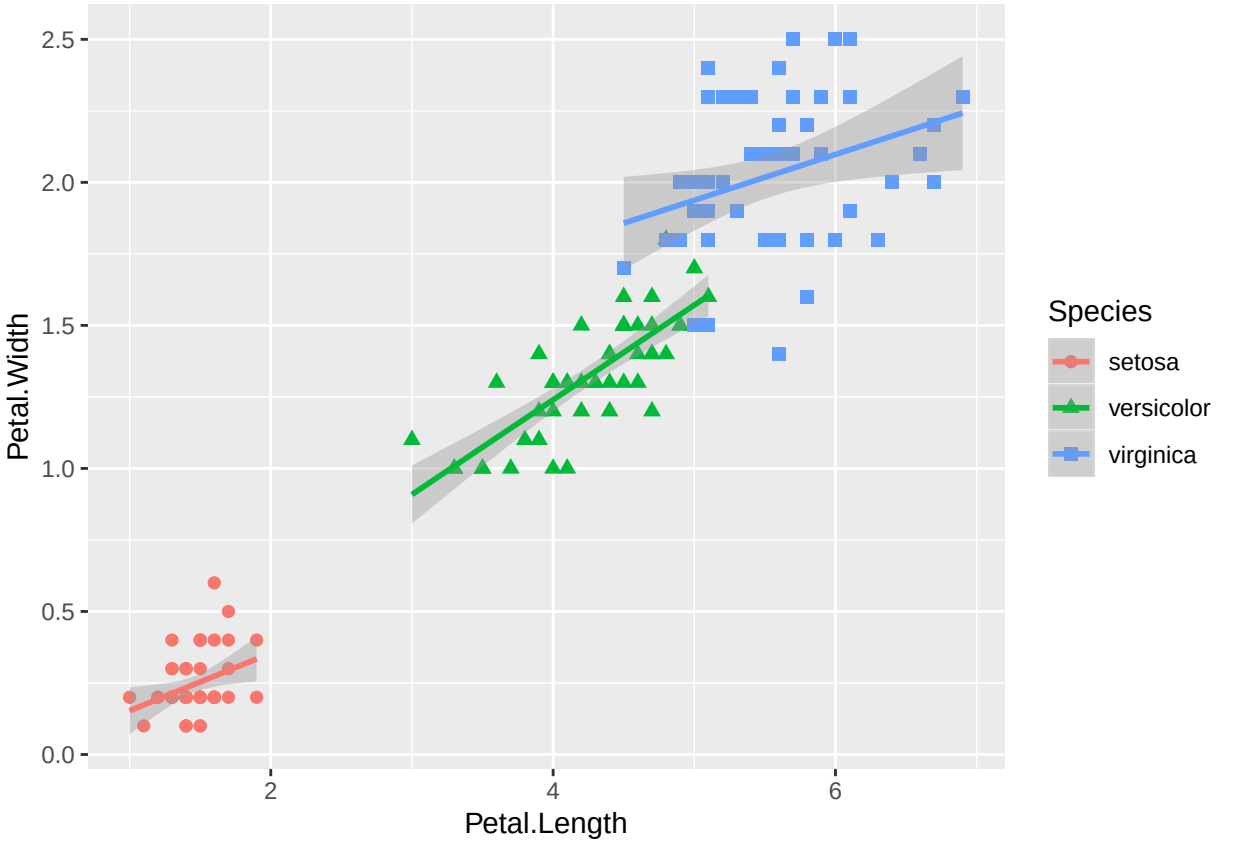
Color/shape by the species subgroups

```
ggplot(iris, aes(x=Petal.Length, y=Petal.Width)) +
geom_point(aes(color=Species, shape=Species), size=2)+
stat_smooth(method="lm", col='brown')
```



fit the lines within the species subgroups:

```
library(ggplot2)
ggplot(iris, aes(x=Petal.Length, y=Petal.Width, color=Species, shape=Species)) +
geom_point(size=2)+
stat_smooth(method="lm")
```



Mutiple linear regression: mpg example

If we are interested to study hwy (highway miles per gallon), we think year (year of manufacture), cyl (number of cylinders), cty (city miles per gallon) may be related to hwy.

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
LMfit <- lm(hwy~ cyl+ cty + year +drv, data=mpg  )
summary(LMfit)
```

```
##
```

```
## Call:
## lm(formula = hwy ~ cyl + cty + year + drv, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3066 -0.9527 -0.1976  0.7214  5.2657
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -118.04978   43.33017  -2.724  0.00694 **
## cyl           -0.15014    0.10929  -1.374  0.17087
## cty            1.15413    0.04392  26.279  < 2e-16 ***
## year           0.06072    0.02171   2.797  0.00560 **
## drvf           2.26833    0.26684   8.501 2.48e-15 ***
## drvr           2.27372    0.35022   6.492 5.22e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.472 on 228 degrees of freedom
## Multiple R-squared:  0.9402, Adjusted R-squared:  0.9389
## F-statistic: 716.5 on 5 and 228 DF,  p-value: < 2.2e-16
```

```
# Look at the coefficient and sign, and how do you intepret the results
```

Results: 1. hwy increases with cty and year (continuous var).

2. For the categorical variable: drv f = front-wheel drive, r = rear wheel drive, 4 = 4wd 4wd is the reference group. Meaning: drv f or drv r each had better hwy compared to 4wd.
3. hwy is not significantly associated with cyl,

Statistical Tests: One Sample T-test

We can use one-sample t-test to comparing the sample mean of a single vector, with a known value or a constant. Assumption of the test: i.i.d random samples from the normal distribution with unknown variance.

```
summary(iris$Sepal.Width)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.800   3.000   3.057   3.300   4.400
```

```
#lets assume that the mean of the petal widths is 2
#we can use the one sample t-test to find out how significantly different is the sample mean, compared

t.test(iris$Sepal.Width, mu=2)
```

```
##
##  One Sample t-test
##
## data:  iris$Sepal.Width
## t = 29.71, df = 149, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

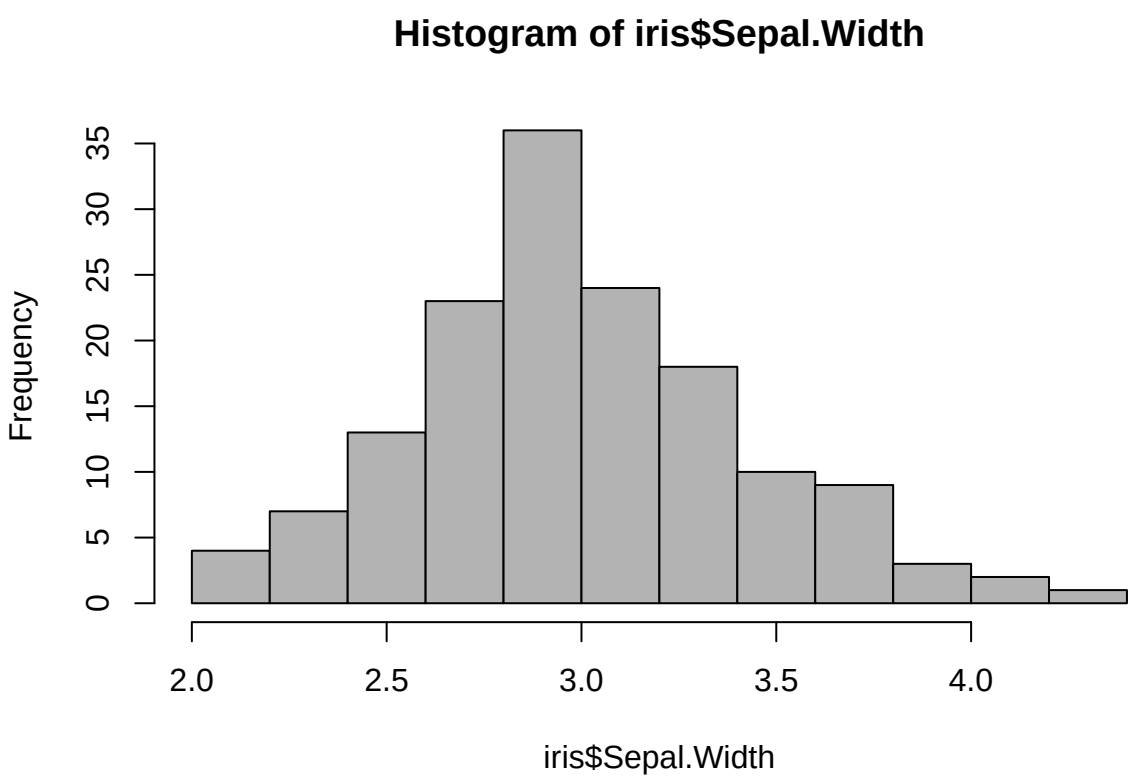
```
# p-value is really small, therefore we reject null. What is we test if mu=3?

t.test(iris$Sepal.Width, mu=3)
```

```
##
##  One Sample t-test
##
## data:  iris$Sepal.Width
## t = 1.611, df = 149, p-value = 0.1093
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  2.987010 3.127656
## sample estimates:
## mean of x
##  3.057333
```

```
# In this case we don't reject null.
```

Is it done? Remember the assumptions? We should check normality assumptions roughly using a histogram, QQplot or, run S-W test before we an trust the test results.

```
hist(iris$Sepal.Width, col='gray70')
```




```r
shapiro.test(iris$Sepal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

Unpaired Two Sample T-test (R)

Uses two vectors which are independent (unpaired). Asumption: both variables are normal distribution.

```r
#generate example data, for two independent groups
women_weight = c(38, 61, 73, 21, 63, 64, 48, 48, 48)
men_weight = c(67, 60, 63, 76, 89, 73, 67, 61, 62)
#create a data frame
weightdf = data.frame(group = rep(c("Woman", "Man"), each = 9), weight = c(women_weight,  men_weight))

#look at the data
weightdf
```

```
##    group weight
## 1  Woman     38
## 2  Woman     61
```

```
## 3  Woman     73
## 4  Woman     21
## 5  Woman     63
## 6  Woman     64
## 7  Woman     48
## 8  Woman     48
## 9  Woman     48
## 10   Man     67
## 11   Man     60
## 12   Man     63
## 13   Man     76
## 14   Man     89
## 15   Man     73
## 16   Man     67
## 17   Man     61
## 18   Man     62
```

```
#group summary
#dplyr is a popular package for grouping, summarising data.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

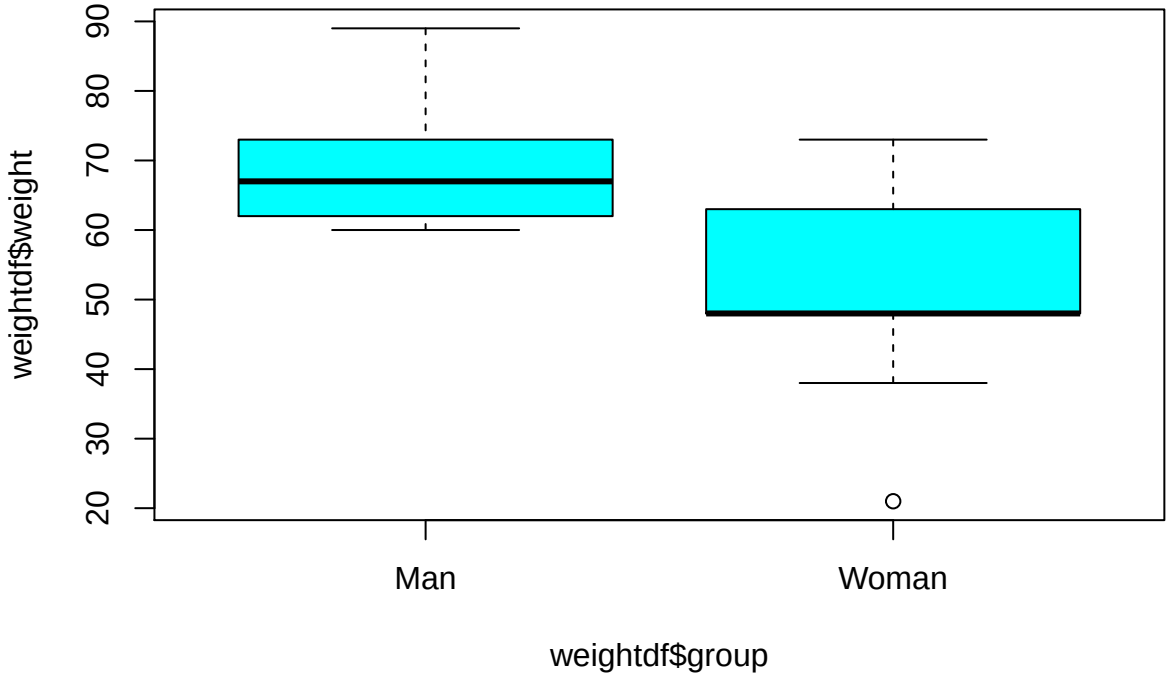
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#generate the average weight for the mean and women group
weightdf %>%
  group_by(group) %>%
    summarise(meanweight=mean(weight))
```

```
## # A tibble: 2 x 2
##   group meanweight
##   <fct>      <dbl>
## 1 Man         68.7
## 2 Woman       51.6
```

```
#plot groups
boxplot(weightdf$weight~weightdf$group, col=5)
```




```
#Unpaired t-test without assuming equal variance
weightresult = t.test(weight~group, data=weightdf, paired=F)

#interpreting the results
weightresult
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by group
## t = 2.7983, df = 13.019, p-value = 0.01506
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.902651 30.319571
## sample estimates:
##   mean in group Man mean in group Woman
##            68.66667            51.55556
```

Check normality assumptions.

```
shapiro.test(women_weight)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  women_weight
## W = 0.93991, p-value = 0.581
```

```
shapiro.test(men_weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  men_weight
## W = 0.85476, p-value = 0.08405
```

Paired T-test

Measurements are paired. For example, same individual before and after treatment.

```
#lets use an example data where extra sleep time after treatment with two different drugs is recorded
data(sleep)
#help(sleep)

str(sleep)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ extra: num  0.7 -1.6 -0.2 -1.2 -0.1 3.4 3.7 0.8 0 2 ...
##  $ group: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ID   : Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
```

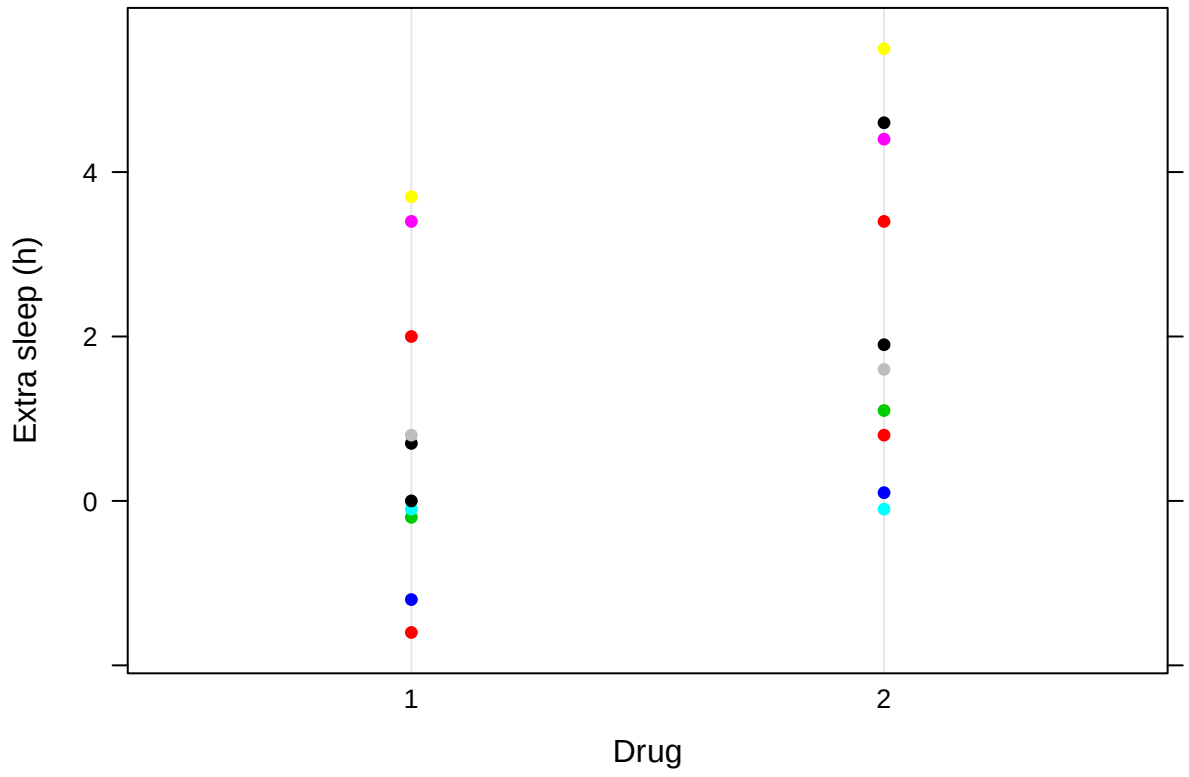
```
sleep
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

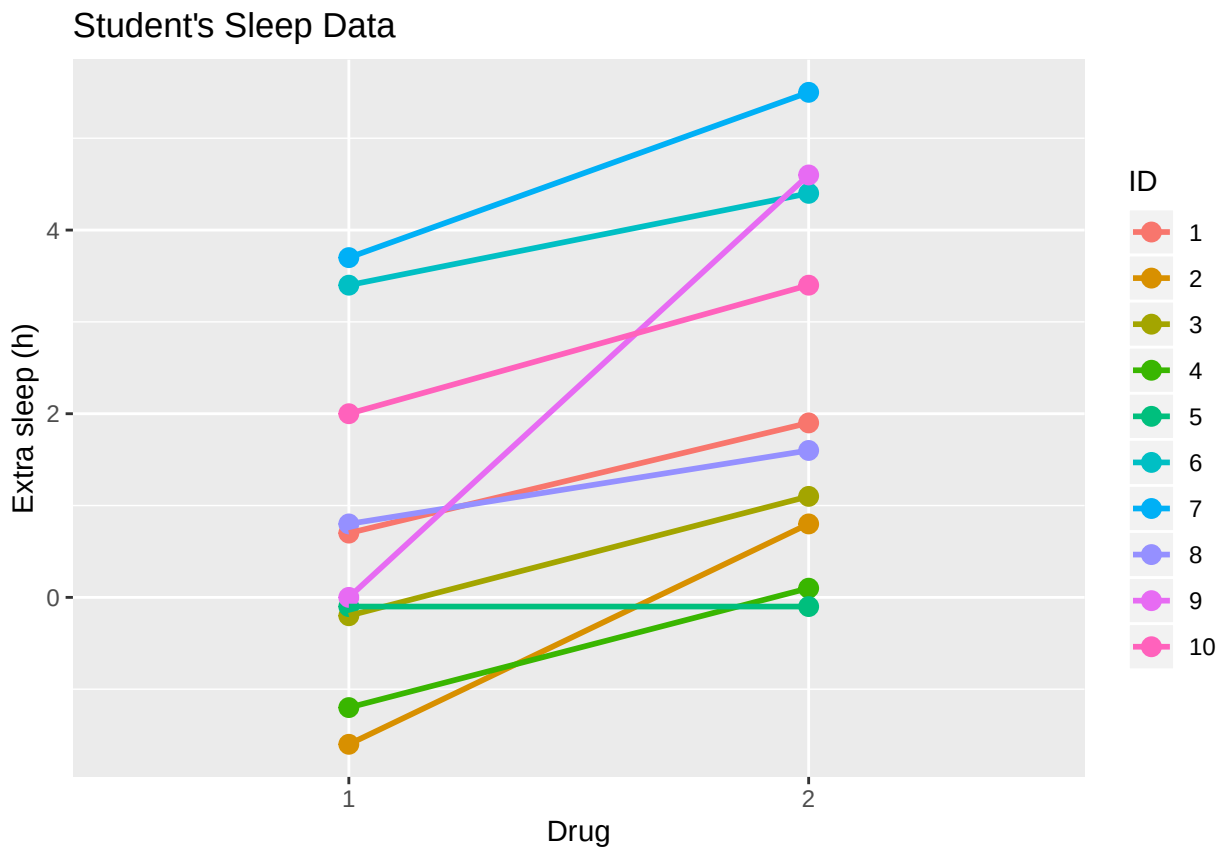
```r
# exam the data, each ID, 2 drugs, paired data on outcome "extra"

#plot the data groups using dotpot: same ID use sample color
require(lattice)
```

```
## Loading required package: lattice
```

```r
dotplot(extra~group, data=sleep, col=sleep$ID, ylab="Extra sleep (h)", xlab="Drug")
```



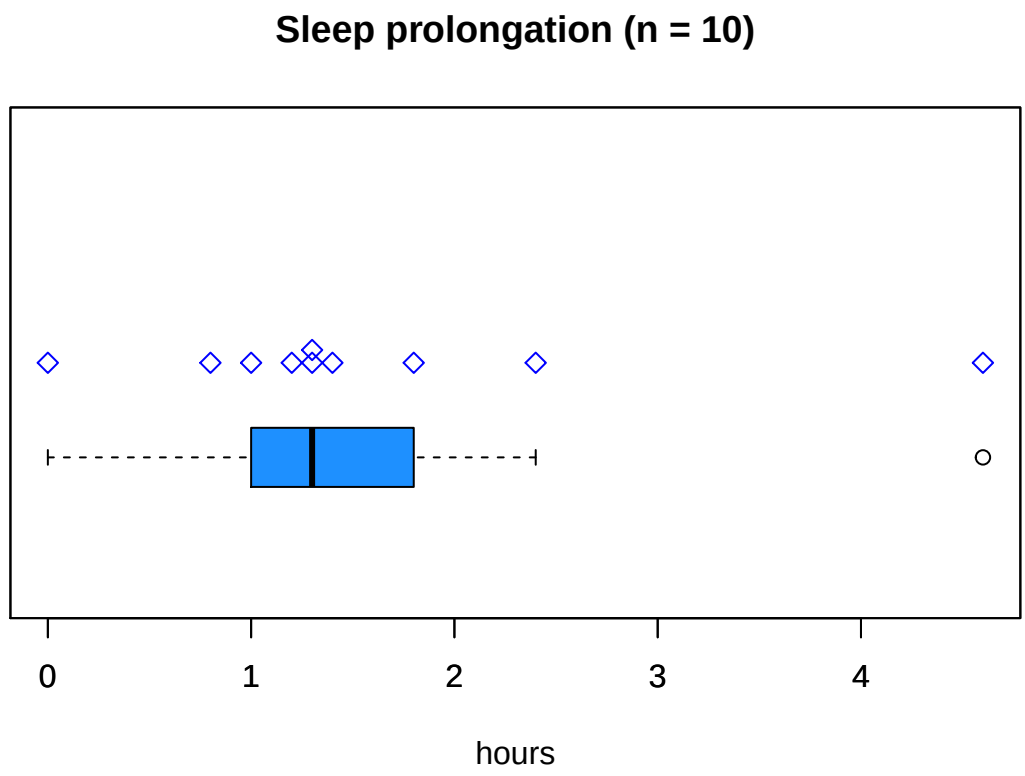

```r
# Other way to plot data: paired data should use connect lines:
# this shows much better from the effect comparing drug 1 with drug 2.

ggplot(sleep, aes(y = extra, x = group)) +
     geom_point( aes(group=ID, color=ID), size=3) +
      geom_line( aes(group=ID, color=ID), size=1)+
      labs( title = "Student's Sleep Data",
           x = "Drug",
           y = "Extra sleep (h)")
```




```
#Run paired t-test
t.test(extra~group, data=sleep, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

```
# The result is very significant. Can we display this results graphicaly
Diffsleep <- with(sleep, extra[group == 2] - extra[group == 1])
stripchart(Diffsleep, method = "stack", xlab = "hours",
           main = "Sleep prolongation (n = 10)", col='blue', pch=5, lwd=1)
boxplot(Diffsleep, horizontal = TRUE, add = TRUE,
        at = .6, pars = list(boxwex = 0.5, staplewex = 0.25), col='dodgerblue')
```



Test normality assumption

```
shapiro.test(sleep$extra)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sleep$extra
## W = 0.94607, p-value = 0.3114
```

Nonparametric test when data are not normally distributed

For two sample unpaired comparison, the Wilcoxon rank-sum test (Mann–Whitney U test) can be applied if the variables are not normally distributed. If data is normally distributed, then it is still OK to use Nonparametric test (lost slight efficiency).

```
weight.test2 = wilcox.test(weight~group, data=weightdf, paired=F)
```

```
## Warning in wilcox.test.default(x = c(67, 60, 63, 76, 89, 73, 67, 61, 62), :
## cannot compute exact p-value with ties
```

```
weight.test2
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  weight by group
## W = 65.5, p-value = 0.02983
## alternative hypothesis: true location shift is not equal to 0
```

For two sample paired comparison, the Wilcoxon signed-rank test can be applied if the variables are not normally distributioned.

```
wilcox.test(extra~group, data=sleep, paired=TRUE)
```

```
## Warning in wilcox.test.default(x = c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4,
## 3.7, : cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(x = c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4,
## 3.7, : cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  extra by group
## V = 0, p-value = 0.009091
## alternative hypothesis: true location shift is not equal to 0
```

Chi-squared Test for two categorical variables (contingency table)

We can use Chi-squared Test to test if two categorical variables are independent or not.

```
#From Agresti(2007)
Voters <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
dimnames(Voters) <- list(gender = c("F", "M"),
                         party = c("Democrat","Independent", "Republican"))
Voters
```

```
##       party
## gender Democrat Independent Republican
##      F      762         327        468
##      M      484         239        477
```

```
# To answer the questions if gender and party are indep, i.e. party affiliations are the same or differ

# we can check the % of voters within gender
prop.table(Voters[1,])
```

```
##    Democrat Independent  Republican
##   0.4894027   0.2100193   0.3005780
```

```r
prop.table(Voters[2,])
```

```
##    Democrat Independent  Republican
##   0.4033333   0.1991667   0.3975000
```

```r
# Run a chisq test.
chisq.test(Voters)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Voters
## X-squared = 30.07, df = 2, p-value = 2.954e-07
```

```r
#p-value is very small, not indepedent
```

One-way ANOVA (1)

Comparisons between two or more groups, where the grouping is based on only one variable (factor). If the data is comparing from an random experiment, this factor indicates the randomized treatment assignment.

```r
#example data - weight of plants under control and two different treatments - PlantGrowth

data("PlantGrowth")
#help("PlantGrowth")

str(PlantGrowth)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

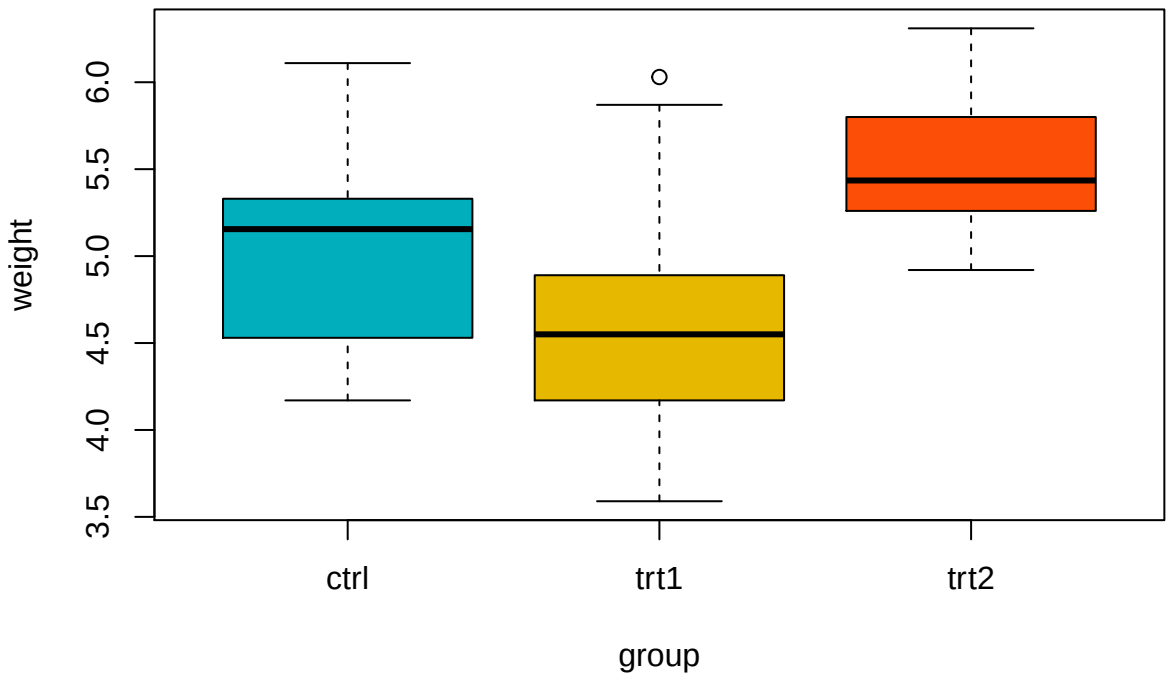
```r
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```r
levels(PlantGrowth$group)
```

```
## [1] "ctrl" "trt1" "trt2"
```

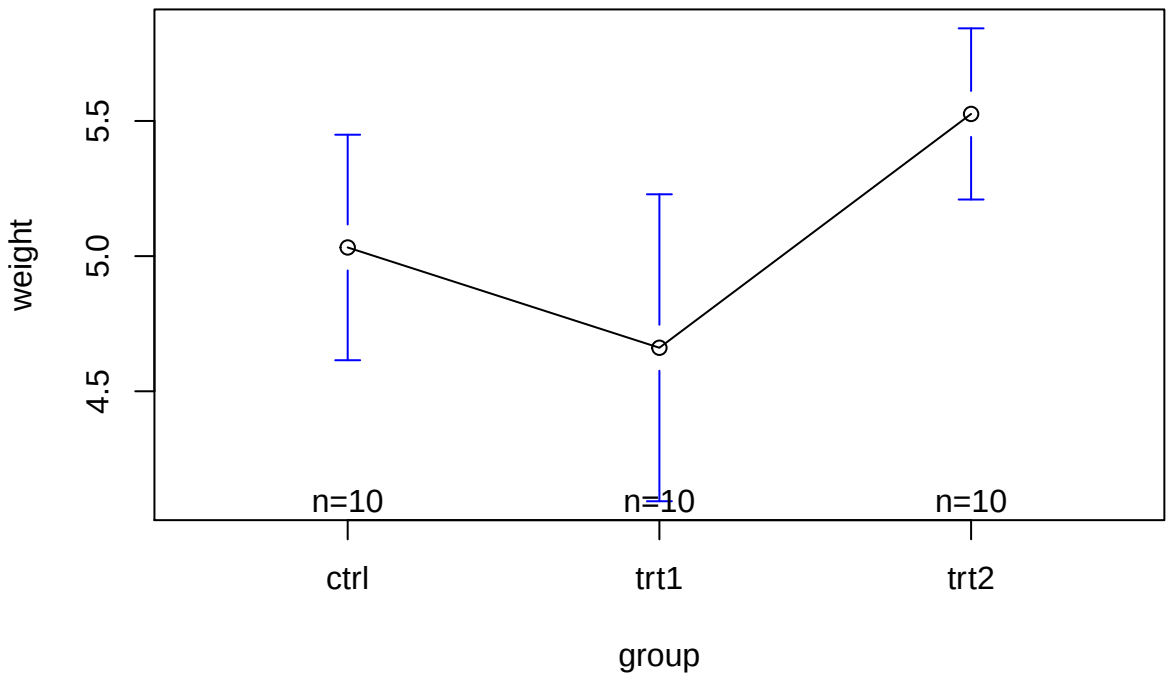
```r
#plot groups
#google for online color picker and pick your favorite color
boxplot(weight~group, data=PlantGrowth, col=c("#00AFBB", "#E7B800", "#FC4E07"))
```




```
#Plot group means and confidence intervals.
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```
plotmeans(weight~group, data=PlantGrowth)
```



One-way ANOVA (2)

```
#compute one-way anova using aov
oneanovares =  aov(weight~group, data=PlantGrowth)
summary(oneanovares)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#pvalue less than 0.05, shows significant differences between groups exist
#which of the group is better performing
```

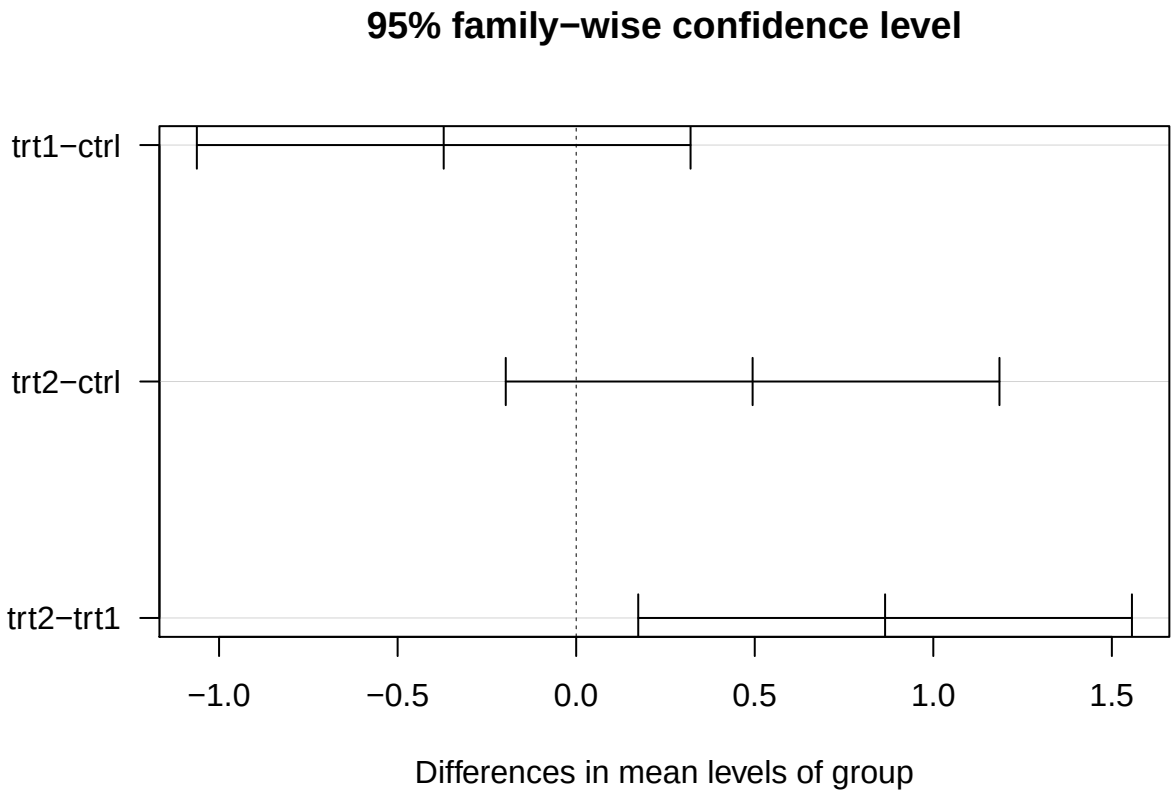
One-way ANOVA (3): Mutiple comparisons

```
TukeyHSD(oneanovares)
```

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ group, data = PlantGrowth)
##
## $group
##             diff        lwr       upr     p adj
## trt1-ctrl -0.371 -1.0622161 0.3202161 0.3908711
## trt2-ctrl  0.494 -0.1972161 1.1852161 0.1979960
## trt2-trt1  0.865  0.1737839 1.5562161 0.0120064
```

```
plot(TukeyHSD(oneanovares),las=1)
```

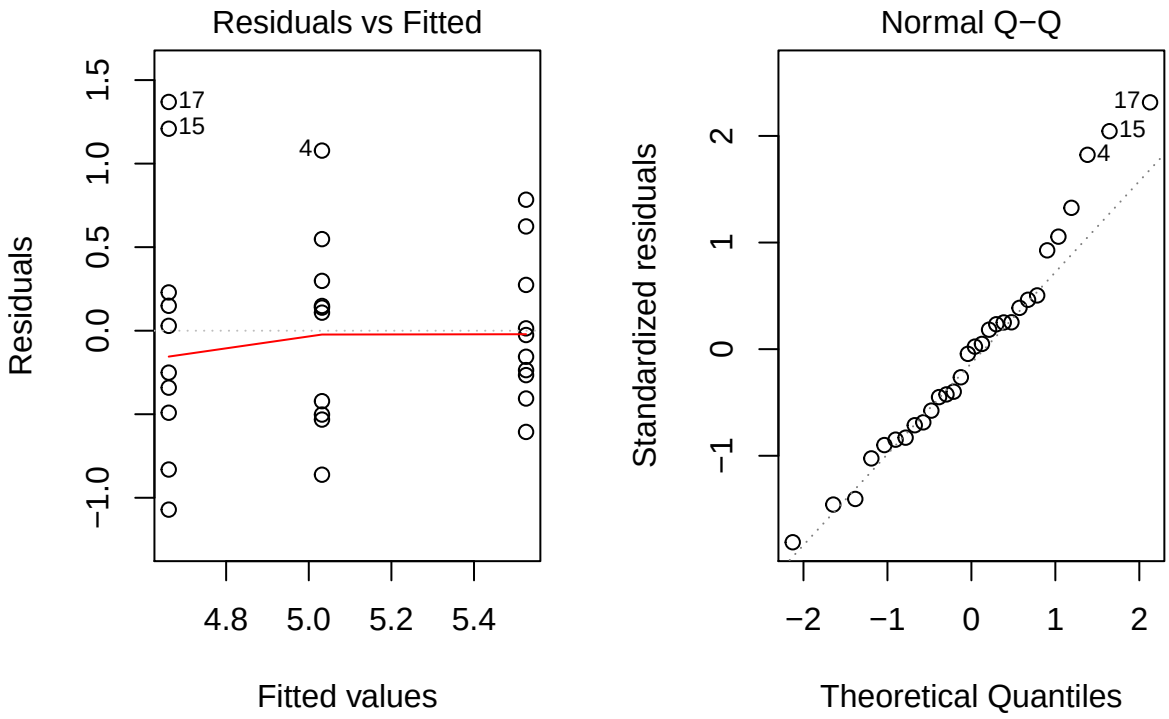


One-way ANOVA (4): stat assumption for ANOVA

Note, it is important to check the two assumptions for ANOVA test

1. the variances are equal for different groups
2. the data (or errors) are independent normally distributed.

```
#oneanovares =  aov(weight~group, data=PlantGrowth)
par(mfrow=c(1,2))
 plot(oneanovares,1)
 plot(oneanovares,2)
```



If two plots show a big departure from the ANOVA assumption, then the nonparametric test is prefered. Or we can apply a transformation on the outcome to improve it is normality, eg. using a log for a skewed data.

```
kruskal.test(weight~group, data=PlantGrowth)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by group
## Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842
```

3.6. Two-way ANOVA

Study the effect of two grouping variables (two experimental or study factors), simultaneously, on a reponse variable. Two-ANOVA also had the same two assumptions.

```
#ToothGrowth example data
#60 guinea  pigs, each treated one of three dose levels of vitamin C, by one of two deliver
#methods. Tooth lengh was measured.

data("ToothGrowth")
#? ToothGrowth
head(ToothGrowth, 10)
```

```
##     len supp dose
```

```
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
```

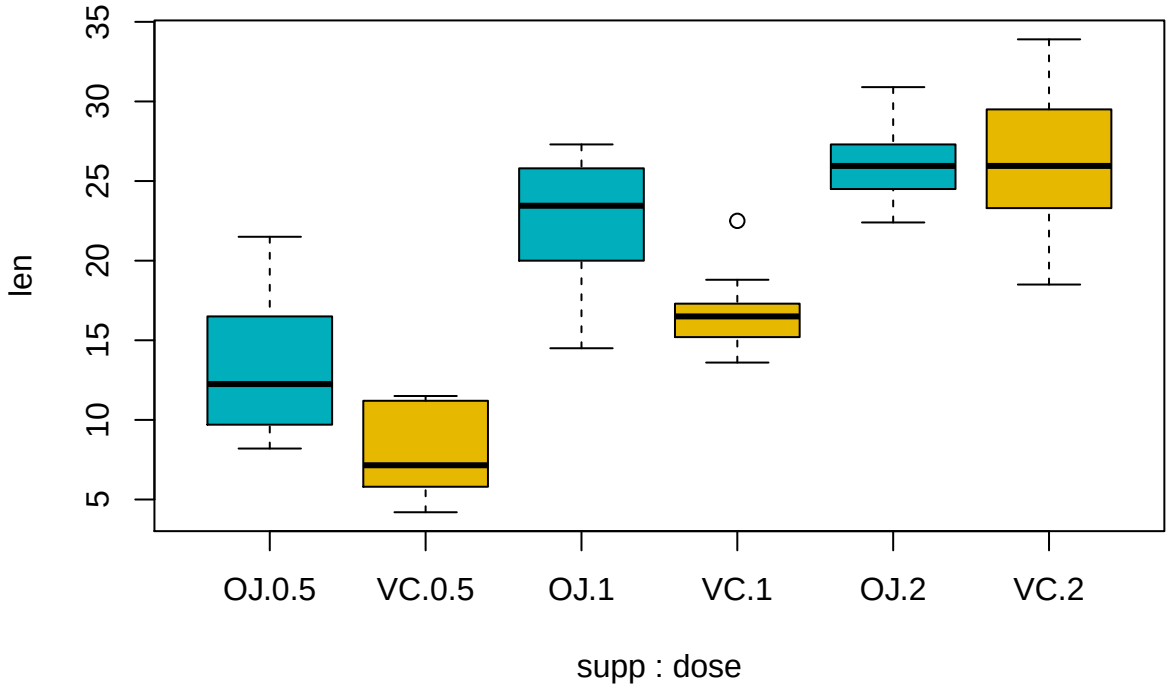
```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

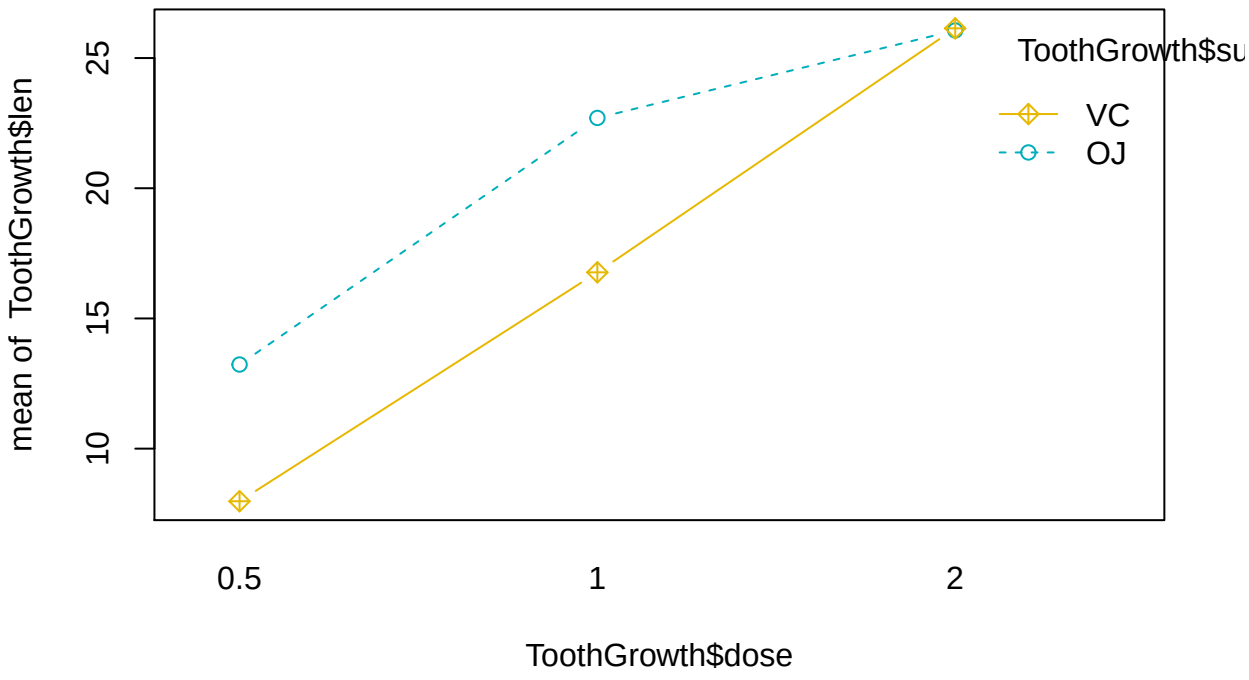
```
#convert dose column as factor
ToothGrowth$dose = factor(ToothGrowth$dose)

#generate frequency table
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

```
#visualize data
boxplot(len~supp * dose, data=ToothGrowth,
        col=c("#00AFBB", "#E7B800"))
```




```
#two-way interaction plot: parallel trends indicate a lack of interaction
interaction.plot(x.factor = ToothGrowth$dose, trace.factor = ToothGrowth$supp,
                 response = ToothGrowth$len, fun=mean, type="b",
                 legend=TRUE, pch=c(1,9), col=c("#00AFBB", "#E7B800"))
```




```
# here it does indicate certain degree of interaction
```

```
# Run two way ANOVA
twoanovares=aov(len~supp*dose, data = ToothGrowth)
summary(twoanovares)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## supp          1  205.4   205.4  15.572 0.000231 ***
## dose          2 2426.4  1213.2  92.000  < 2e-16 ***
## supp:dose     2  108.3    54.2   4.107 0.021860 *
## Residuals    54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# this indicate both main effects and their interactions are significant on the outcome.
```