Statistics

Annabel Belliard, Emily Beuken, Jan van Houten

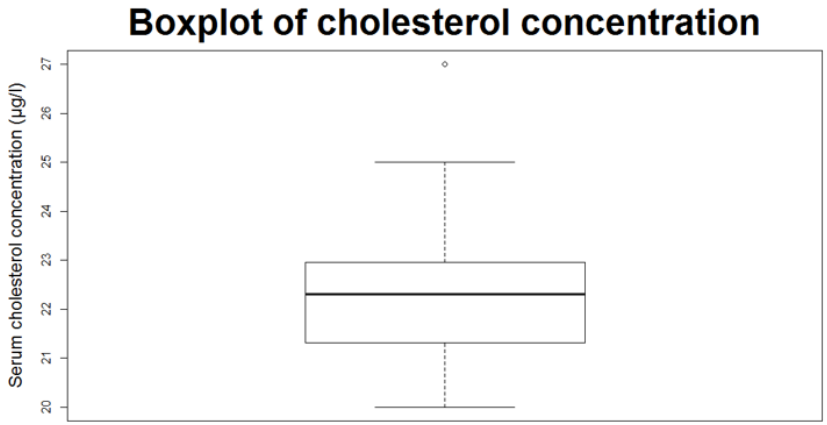
StudCee

October 26, 2017

Table of Contents

Numerical Data: One Variable

Serum cholesterol (C) concentration ($\mu g/l$) of 12 subjects

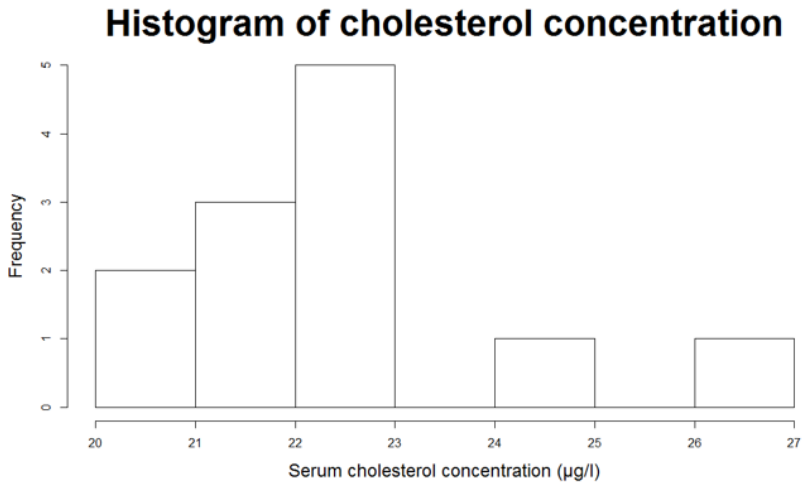


R Commands

```
y <- c(20,21, 21.52,21.1,21.9,23,22.9,22.2,22.4,22.7, 25, 27)
boxplot(y, ylab = "Serum cholesterol concentration (g/l)", main =
"Boxplot of cholesterol concentration")
```

Numerical Data: One Variable

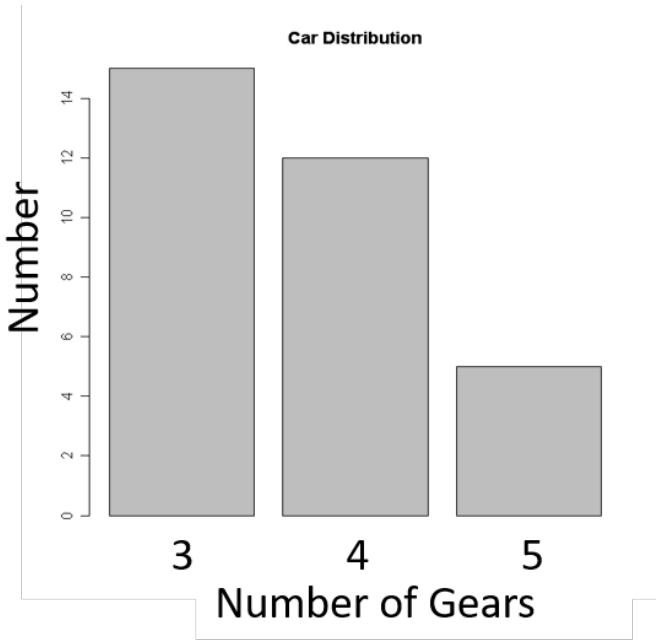
Serum cholesterol (C) concentration ($\mu g/l$) of 12 subjects



R Commands

```
y <- c(20,21, 21.52,21.1,21.9,23,22.9,22.2,22.4,22.7, 25, 27)
hist(y, xlab = "Serum cholesterol concentration (g/l)", main =
"Boxplot of cholesterol concentration")
```

Categorical Data: One Variable

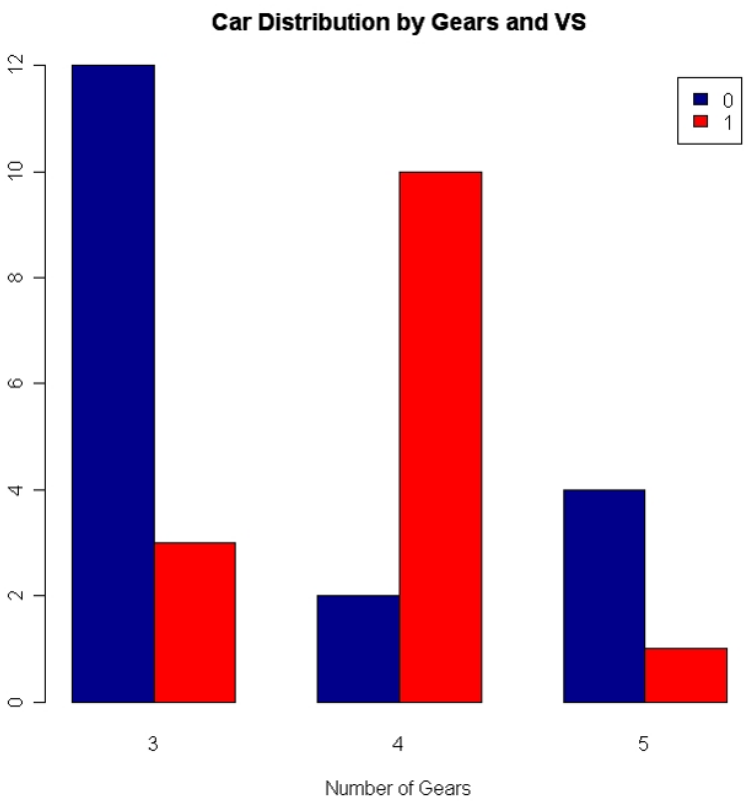


car	gear
MazdaRX4	4
MazdaRX4 Wag	4
Datsun710	4
Hornet4Drive	3
HornetSportabout	3
Valiant	3
Duster360	3
Merc240D	4
Merc230	4
Merc280	4

R Commands

```
counts <- table(mtcars$gear)
barplot(counts, main="Car Distribution", xlab="Number of Gears")
```

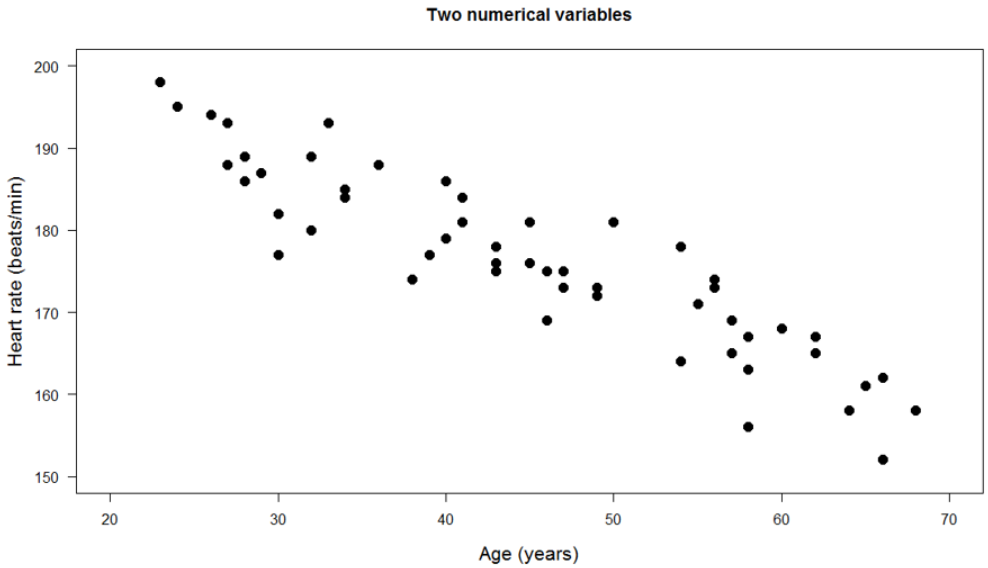
Categorical Data: Grouped Bar Plot



car	gear	vs
MazdaRX4	4	0
MazdaRX4 Wag	4	0
Datsun710	4	1
Hornet4Drive	3	1
HornetSportabout	3	0
Valiant	3	1
Duster360	3	0
Merc240D	4	1
Merc230	4	1
Merc280	4	1

Numerical Data: Two Variables

The relationship between heart rate and age.



R Commands

```
plot(Heartrate Age,pch=16,xlim=c(20,70),ylim=c(150,200),cex=1.5,
cex.lab=1.3,xlab="Age (years)", ylab="Heart rate
(beats/min),las=1, main = "Two numerical variables")
```

Tests to be discussed

- One-sample t-test
- Two-sample t-test
- Wilcoxon Signed-Rank test
- Wilcoxon Rank-Sum test
- Chi-Squared Test for given probabilities
- Chi-Squared test for independence
- ANOVA
- ANCOVA
- Regression

One Sample T-Test

Is the mean Serum cholesterol (C) concentration (g/l) of 12 subjects of this population equal to 22?

H0: The true mean is 22

HA: The true mean is not 22

R Commands

```
y <- c(20,21, 21.52,21.1,21.9,23,22.9,22.2,22.4,22.7, 25, 27)
t.test(y,mu=22)
```

One Sample T-Test

Is the mean Serum cholesterol (C) concentration (g/l) of 12 subjects of this population equal to 22?

H0: The true mean is 22

HA: The true mean is not 22

R Output

```
data:  y
t = 1.0331, df = 11, p-value = 0.3237
alternative hypothesis:  true mean is not equal to 22
95 percent confidence interval:  21.36691 23.75309
sample estimates:  mean of x 22.56
```

One Sample T-Test

Is the mean Serum cholesterol (C) concentration (g/l) of 12 subjects of this population equal to 22?

H0: The true mean is 22

HA: The true mean is not 22

R Output

```
data:  y
t = 1.0331, df = 11, p-value = 0.3237
alternative hypothesis:  true mean is not equal to 22
95 percent confidence interval:  21.36691 23.75309
sample estimates:  mean of x 22.56
```

Conclusion: We cannot reject the null hypothesis.

Two Sample T-Test

Is there a significant difference in height between Germans and Italians?

H0: There is no significant height difference

HA: Height between the Germans and Italians differ significantly

R Commands

```
it <- c(175, 168, 168, 180, 156, 181, 172, 165, 174, 179)
ge <- c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)
t.test(it,ge)
```

Two Sample T-Test

Is there a significant difference in height between Germans and Italians?

H0: There is no significant height difference

HA: Height between the Germans and Italians differ significantly

R Output

```
Welch Two Sample t-test data:  it and ge
t = -2.0048, df = 17.402, p-value = 0.0608
alternative hypothesis:  true difference in means not equal to 0
95 percent confidence interval:  -13.1234361 0.3234361
sample estimates:  mean of x:  171.8, mean of y:  178.2
```

Two Sample T-Test

Is there a significant difference in height between Germans and Italians?

H0: There is no significant height difference

HA: Height between the Germans and Italians differ significantly

R Output

```
Welch Two Sample t-test data:  it and ge
t = -2.0048, df = 17.402, p-value = 0.0608
alternative hypothesis:  true difference in means not equal to 0
95 percent confidence interval:  -13.1234361 0.3234361
sample estimates: mean of x: 171.8, mean of y: 178.2
```

Conclusion: We cannot reject the null hypothesis.

Paired T-Test

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

Individual	age5	age6
1	7.33	7.53
2	7.49	7.70
3	7.27	7.46
4	7.93	8.21
5	7.56	7.81
6	7.81	8.01
7	7.46	7.72
8	6.94	7.13
9	7.49	7.68
10	7.44	7.66
11	7.95	8.11
12	7.47	7.66
13	7.04	7.20
14	7.10	7.25
15	7.64	7.79

Paired T-Test

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Commands

```
t.test(age5,age6,paired=T)
```

Paired T-Test

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Output

```
data:  age5 and age6
t = -19.72, df = 14, p-value = 1.301e-11
alternative hypothesis:  true difference in means not equal to 0
95 percent confidence interval:  -0.2217521 -0.1782479
sample estimates:  mean of the differences -0.2
```

Paired T-Test

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Output

```
data:  age5 and age6
t = -19.72, df = 14, p-value = 1.301e-11
alternative hypothesis:  true difference in means not equal to 0
95 percent confidence interval:  -0.2217521 -0.1782479
sample estimates:  mean of the differences -0.2
```

Conclusion: We can reject the null hypothesis.

Wilcoxon

For non parametric data.

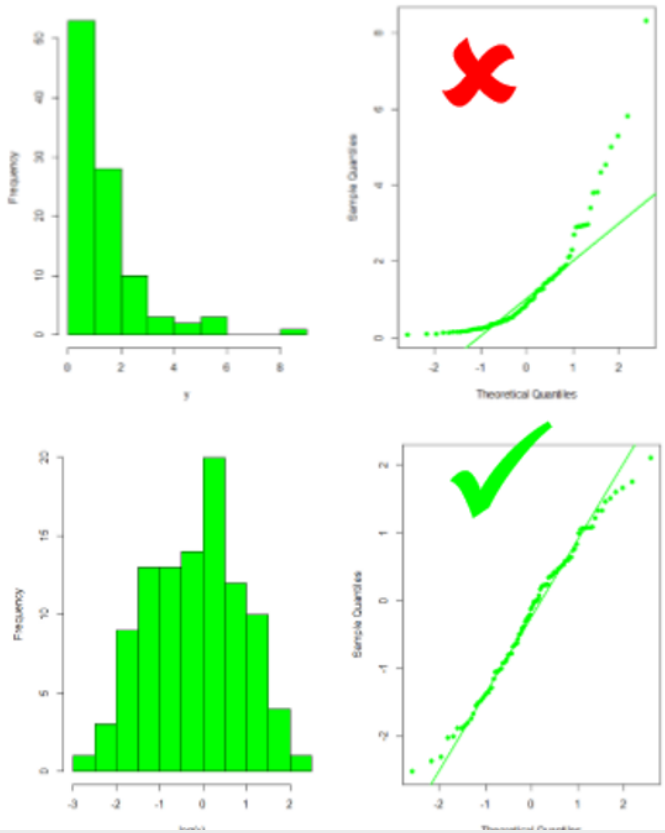
- Wilcoxon Signed Rank *Paired data*
- Wilcoxon Rank Sum *Non Paired data*

Wilcoxon Signed-Rank for Paired Data

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)



Wilcoxon Signed-Rank for Paired Data

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Commands

```
wilcox.test(age5,age6,paired=T)
```

Wilcoxon Signed-Rank for Paired Data

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Output

```
Wilcoxon signed rank test with continuity correction
data:  age5 and age6
V = 0, p-value = 0.0007193
alternative hypothesis:  true location shift is not equal to 0
```

Wilcoxon Signed-Rank for Paired Data

Question: Is there a significant difference in skull diameter in girls at ages 5 and 6?

H0: The diameter stayed the same (difference is 0)

HA: The diameter did not stay the same (difference is not 0)

R Output

```
Wilcoxon signed rank test with continuity correction
data:  age5 and age6
V = 0, p-value = 0.0007193
alternative hypothesis:  true location shift is not equal to 0
```

Conclusion: We can still reject the null hypothesis.

Kruskal-Wallis

Non-parametic one-way ANOVA

R Code

```
kruskal.test.(bodymass location)
```

R Output

```
kruskal wallis rank sum test
data:  bodymass by location
kruskal wallis chi squared =15.482 de=3
```

p-value=0.001448

ANOVA

Is there a significant effect of different 2% sugar solutions on the length (ocular units = 8.77*mm*) of peas

H0: There is no significant difference between the different treatments

HA: : There is a significant difference between at least 2 of the different treatments.

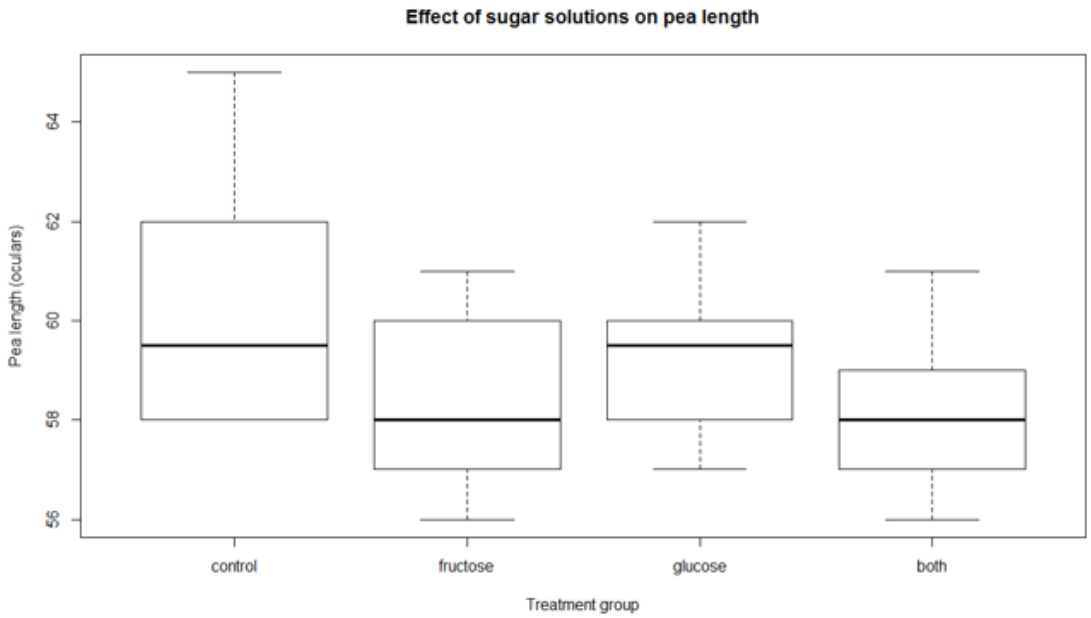
pealength	group
62	control
58	control
60	control
62	control
58	control
60	control
58	control
58	control
65	control
59	control
57	glucose
58	glucose
60	glucose
59	glucose
62	glucose
60	glucose
60	glucose
57	glucose
59	glucose
61	glucose
58	fructose
61	fructose
56	fructose
58	fructose

ANOVA

Is there a significant effect of different 2% sugar solutions on the length (ocular units = 8.77*mm*) of peas

H0: There is no significant difference between the different treatments

HA: : There is a significant difference between at least 2 of the different treatments.



ANOVA

Is there a significant effect of different 2% sugar solutions on the length (ocular units = 8.77mm) of peas

H0: There is no significant difference between the different treatments

HA: : There is a significant difference between at least 2 of the different treatments.

R Code

```
m1 <- lm(pealength group)
anova(m1)
```

ANOVA

Is there a significant effect of different 2% sugar solutions on the length (ocular units $= 8.77mm$) of peas

H0: There is no significant difference between the different treatments

HA: : There is a significant difference between at least 2 of the different treatments.

R Code

```
Analysis of Variance Table
Response:  pealength
             Df   Sum Sq   Mean Sq   F value   Pr(> F)
   group      3   26.675     8.892     2.588   0.06803   .
 Residuals   36  123.700     3.436
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

ANOVA

Is there a significant effect of different 2% sugar solutions on the length (ocular units = $8.77mm$) of peas

H0: There is no significant difference between the different treatments

HA: : There is a significant difference between at least 2 of the different treatments.

R Code

```
Analysis of Variance Table
Response:  pealength
             Df   Sum Sq   Mean Sq   F value   Pr(> F)
   group      3   26.675     8.892     2.588   0.06803   .
 Residuals   36  123.700     3.436
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Conclusion: the null hypothesis cannot be rejected. There is no significant difference between the sugar treatments.

But what if there was a difference?

Tukey Contrasts allows comparison of means.

R Code

```
m1 <- lm(pealength group)
anov <- anova(m1)
turkey <- TukeyHSD(x = anov)
```

ANCOVA

Is there a significant effect of different diets on weight of pigs?

H0: There is no significant difference between the different diets

HA: There is a significant difference between at least 2 of the different diets.

Diet	Weight	Leg
1	60.8	48
1	57	46.5
1	65	44.5
1	58.6	39.6
1	61.7	41.4
2	68.7	39.7
2	67.7	50.1
2	74	51.2
2	66.3	40.6
2	69.8	41.5
3	102.6	57.6
3	102.1	45.6
3	100.2	43.8
3	96.5	51.1
4	87.9	50
4	84.2	38.9
4	83.1	45.4
4	85.7	41.5

ANCOVA

Is there a significant effect of different diets on weight of pigs?

H0: There is no significant difference between the different diets

HA: There is a significant difference between at least 2 of the different diets.

R Code

```
m1 <-lm(Weight Leg.c*Diet,data=d)
anova(m1)
summary(m1)
```

ANCOVA

Is there a significant effect of different diets on weight of pigs?

H0: There is no significant difference between the different diets

HA: There is a significant difference between at least 2 of the different diets.

R Code

```
Analysis of Variance Table
Response:  Weight
              Df   Sum Sq   Mean Sq   F value   Pr(> F)
   Leg.c       1   813.7    813.67    95.032    9.531e-7   ***
   Diet        3   3437.5   1145.82   133.825   6.146e-9   ***
 Leg.c:Diet    3     9.4      3.13     0.3654    0.7794
 Residuals    11    94.2      8.56
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

ANCOVA

Is there a significant effect of different diets on weight of pigs?

H0: There is no significant difference between the different diets

HA: There is a significant difference between at least 2 of the different diets.

R Code

```
Analysis of Variance Table
Response:  Weight
              Df   Sum Sq   Mean Sq   F value   Pr(> F)
   Leg.c       1   813.7    813.67    95.032    9.531e-7   ***
   Diet        3   3437.5   1145.82   133.825   6.146e-9   ***
 Leg.c:Diet    3     9.4      3.13     0.3654    0.7794
 Residuals    11    94.2      8.56
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The interaction (Leg.c:Diet) is not significant. The model can been thrown out.

How do you find the best model?

- Choose the most significant variables yourself
- StepAIC

R Code

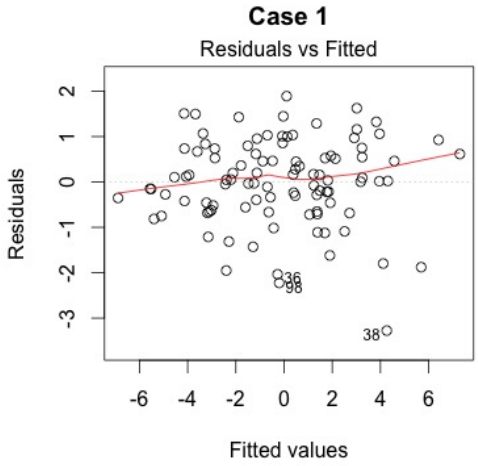
```
m1 <-lm(Weight Leg.c*Diet,data=d)
aic <- stepAIC(m1)
```

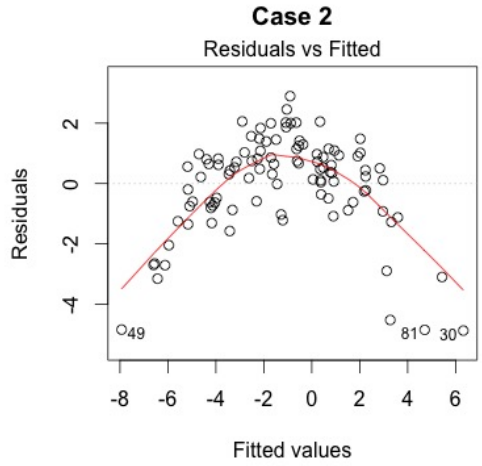
Regression

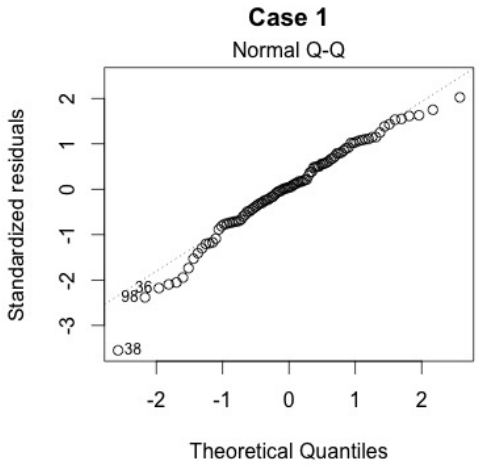
There are three main types of regression:

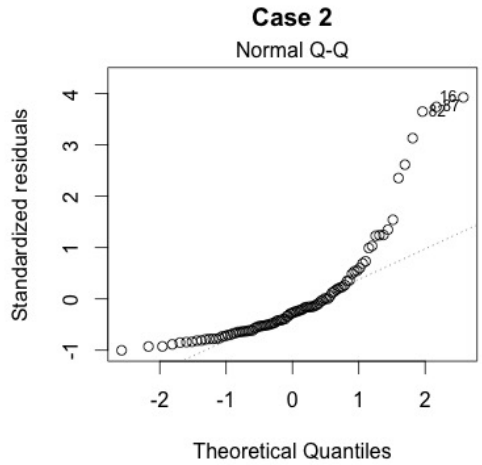
- Linear
- Nonlinear (quadratic)
- Logistic (used for categorical data)

How can you tell which type you need?









Linear Regression

Is there a significant effect of tannin concentration on caterpillar growth?

H0: There is no significant effect of growth on tannin

HA: There is a significant difference of growth on tannin R output.

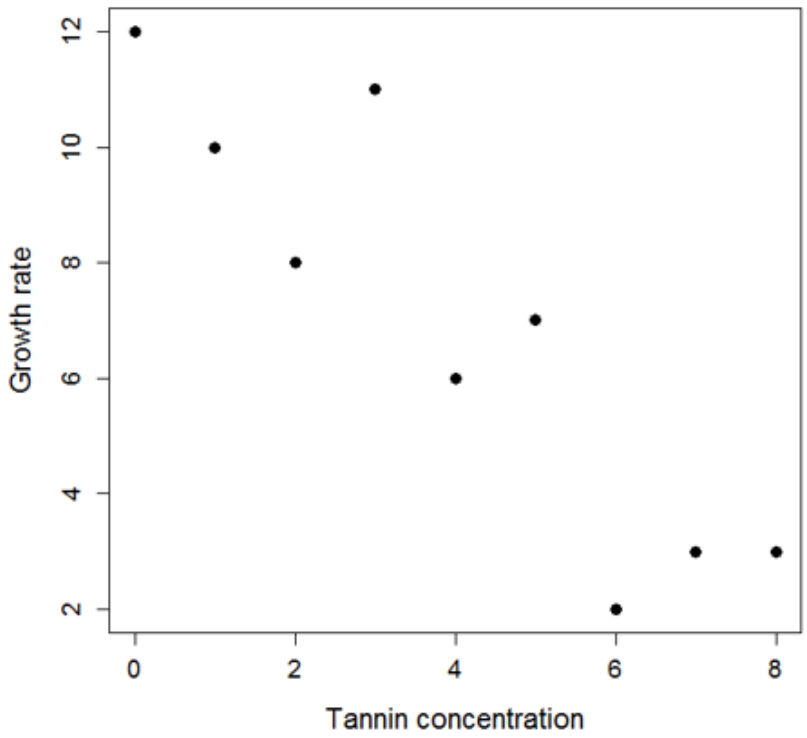
growth	tannin
12	0
10	1
8	2
11	3
6	4
7	5
2	6
3	7
3	8

Linear Regression

Is there a significant effect of tannin concentration on caterpillar growth?

H0: There is no significant effect of growth on tannin

HA: There is a significant difference of growth on tannin R output.



Linear Regression

Is there a significant effect of tannin concentration on caterpillar growth?

H0: There is no significant effect of growth on tannin

HA: There is a significant difference of growth on tannin R output.

R Code

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4556 -0.8889 -0.2389  0.9778  2.8944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept     11.755      1.041  11.295  9.54e-6 ***
tannin        -1.216      0.2186  -5.565 0.000846 ***
Residual standard error: 1.693 on 7 degrees of freedom Multiple
R-squared: 0.8157, Adjusted R-squared: 0.7893 F-statistic:
30.97 on 1 and 7 DF, p-value: 0.0008461
```

Chi-square for Independence

Are students smoking habit and exercise level independent of each other?

H0: The smoking habit is independent of the exercise level of the students.

HA: The smoking habit is NOT independent of the exercise level of the students.

		Smoking		
		Freq	None	Some
Exercise	Heavy	7	1	3
	Never	87	18	84
	Occas	12	3	4
	Regul	9	1	7

Chi-square for Independence

Are students smoking habit and exercise level independent of each other?

H0: The smoking habit is independent of the exercise level of the students.

HA: The smoking habit is NOT independent of the exercise level of the students.

R Code

```
tbl = table(smoke, exercise)
chisq.test(tbl)
```

Chi-square for Independence

Are students smoking habit and exercise level independent of each other?

H0: The smoking habit is independent of the exercise level of the students.

HA: The smoking habit is NOT independent of the exercise level of the students.

R Code

```
Pearson's Chi-squared test
data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
Warning message:
In chisq.test(tbl) :  Chi-squared approximation may be incorrect
```

Chi-square for Independence

Are students smoking habit and exercise level independent of each other?

H0: The smoking habit is independent of the exercise level of the students.

HA: The smoking habit is NOT independent of the exercise level of the students.

R Code

```
Pearson's Chi-squared test
data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
Warning message:
In chisq.test(tbl) :  Chi-squared approximation may be incorrect
```

Conclusion: We cannot reject the null hypothesis.

Chi-square for Goodness of Fit

Are all the tulip colours equally common?

H0: The tulip colours are equally common.

HA: The tulip colours are not equally common.

	Tulip Colours			
	Red	Yellow	White	Total
Counts	81	50	27	158
Hypothetical freq	0.33	0.33	0.33	0.33

Chi-square for Goodness of Fit

Are all the tulip colours equally common?

H0: The tulip colours are equally common.

HA: The tulip colours are not equally common.

R Code

```
tulip <- c(81, 50, 27)
res <- chisq.test(tulip, p = c(1/3, 1/3, 1/3))
res
```

Chi-square for Goodness of Fit

Are all the tulip colours equally common?

H0: The tulip colours are equally common.

HA: The tulip colours are not equally common.

R Code

```
Chi-squared test for given probabilities

data:  tulip

X-squared = 27.886, df = 2, p-value = 8.803e-07
```

Chi-square for Goodness of Fit

Are all the tulip colours equally common?

H0: The tulip colours are equally common.

HA: The tulip colours are not equally common.

R Code

```
Chi-squared test for given probabilities

data:  tulip

X-squared = 27.886, df = 2, p-value = 8.803e-07
```

Conclusion: We can reject the null hypothesis.

Which tests do you use?

- Is the mean equal to a number?

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?
  Use a *paired t-test* if the data are paired (two points per individual); else...

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?
  Use a *paired t-test* if the data are paired (two points per individual); else...use a *two-sample t-test*

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?
  Use a *paired t-test* if the data are paired (two points per individual); else...use a *two-sample t-test*
- Does the data in polulation 1 and those in population 2 come from the same distribution?

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?
  Use a *paired t-test* if the data are paired (two points per individual); else...use a *two-sample t-test*
- Does the data in polulation 1 and those in population 2 come from the same distribution?
  Use a *Wilcoxon Signed-Rank test* if the data are paired; else...

Which tests do you use?

- Is the mean equal to a number?
  Use a *one-sample t-test*
- Is the mean in population 1 equal to the mean in population 2?
  Use a *paired t-test* if the data are paired (two points per individual); else...use a *two-sample t-test*
- Does the data in polulation 1 and those in population 2 come from the same distribution?
  Use a *Wilcoxon Signed-Rank test* if the data are paired; else...use a *Wilcoxon Rank-Sum test*.

Which tests do you use?

- Is the variable spread according to given probabilities?

Which tests do you use?

- Is the variable spread according to given probabilities?
  Use a *chi-squared test for given probabilities*

Which tests do you use?

- Is the variable spread according to given probabilities?
  Use a *chi-squared test for given probabilities*
- Are the variables independent?

Which tests do you use?

- Is the variable spread according to given probabilities?
  Use a *chi-squared test for given probabilities*
- Are the variables independent?
  Use a *chi-square test for independence*.

How do you write your answers formally?

- Research Question
- Informal Analysis
- Formal Analysis
- Conclusion
- Discussion

How do you write your answers formally?

- **Research Question**
  - What information are you given?
  - What do you have to show?
  - What is your research question?
- Informal Analysis
- Formal Analysis
- Conclusion
- Discussion

How do you write your answers formally?

- Research Question
- **Informal Analysis**
  - Look at the distribution of the data
  - Analyse box plots, histograms, scatter plots
  - What is your hypothesis?
- Formal Analysis
- Conclusion
- Discussion

How do you write your answers formally?

- Research Question
- Informal Analysis
- **Formal Analysis**
  - State your hypothesis
  - State your null and alternative hypothesis
  - State when the null and alt. hypothesis will occur
  - Perform the statistical test
  - Can the null hypothesis be rejected?
- Conclusion
- Discussion

How do you write your answers formally?

- Research Question
- Informal Analysis
- Formal Analysis
- **Conclusion**
  - Relate your analysis back to your research question
  - Does the result from the analysis answer the question?
- Discussion

How do you write your answers formally?

- Research Question
- Informal Analysis
- Formal Analysis
- Conclusion
- **Discussion**
  - Were there flaws in the data provided?
  - Were there flaws in how the data was collected?
  - Did you use the best test for the data?

Questions?

Do you have any questions, either from lectures or tutorials, that you want answering?

Statistics Exam - 2016 Q1

Two professors in artificial intelligence developed a patient support system; one based on computational intelligence (CI), and the other based on traditional symbolic AI (TSAI). The two systems were implemented in two different units of a large hospital. Part of the resulting evaluations are expressed in patient satisfaction scores (PSS). An analysis of the patient characteristics revealed that the CI group was younger than the TSAI group.

Statistics Exam - 2016 Q1

50 patients: 25 for CI and 25 for TSAI

PSS	System
79	CI
76	CI
75	CI
...	...
75	TSAI
68	TSAI
73	TSAI

Numerical and Categorical Data

Statistics Exam - 2016 Q1

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment
- Comment on the population and on the sample
- Research question

Statistics Exam - 2016 Q1

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment

  Two patient support systems were implemented in separate units in a hospital; one based on CI, another based on TSAI.

- Comment on the population and on the sample

  There were 50 patients, 25 using each system. The CI group was younger and younger people are more exposed to technology can more easily adapt to change.

- Research question

  Which system has the highest patient satisfacton score (PSS)?

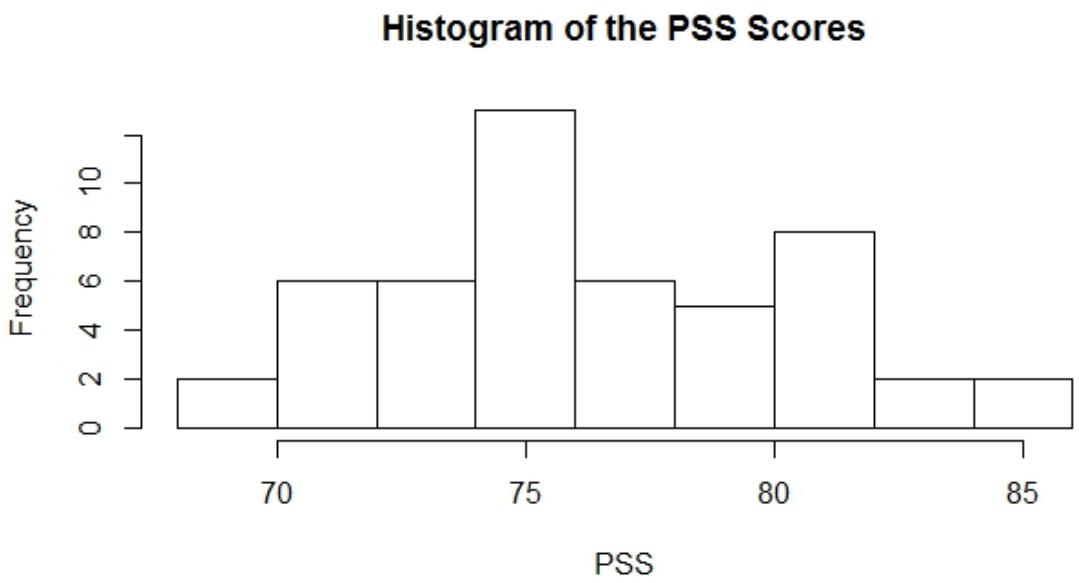
Statistics Exam - 2016 Q1

Perform an informal analysis to explore the research question. [6]

- Is the data normal?
- How can we compare the two groups?

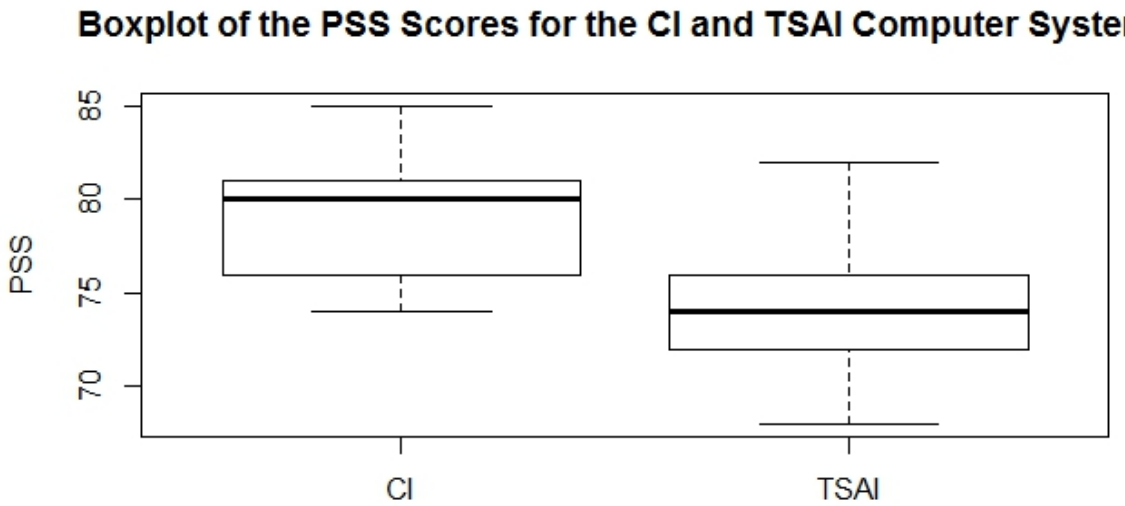
Statistics Exam - 2016 Q1

Perform an informal analysis to explore the research question. [6]



Statistics Exam - 2016 Q1

Perform an informal analysis to explore the research question. [6]



Can we make a hypothesis based on this analysis?

Statistics Exam - 2016 Q1

Perform a formal analysis testing a relevant hypothesis. [8]

What type of data do we have?

- PSS scores are not normally distributed
- Two computer systems
- 50 patients - split into two separate groups

What is our null and alternative hypotheses?

Statistics Exam - 2016 Q1

Perform a formal analysis testing a relevant hypothesis. [8]

H0: The means of the PSS scores are the same

HA: The means of the PSS scores are not the same

R Output

```
Welch Two Sample t-test
data:  systCIPSS and systTSAIPS
t = 5.7585, df = 47.904, p-value = 2.943e-07
alternative hypothesis:  true difference in means greater than 0
95 percent confidence interval:  3.628689 Inf
sample estimates:  mean of x:  79.36, mean of y:  74.24
```

Is this significant? Can the null hypothesis be rejected?

Statistics Exam - 2016 Q1

Give the conclusion including the answer on the research question. [5]

- What was our hypothesis?
- What test did we do? What did it show?
- Did the test help us to reject our hypothesis?
- Does it answer our research question?

Statistics Exam - 2016 Q1

Provide points of discussion about critical aspects of the analysis. [4]

Statistics Exam - 2016 Q1

Provide points of discussion about critical aspects of the analysis. [4]

- 50 patients, is it enough?
- The CI group had more younger patients.
- Was the data normally distributed?

Statistics Exam - 2016 Q2

A world wide survey among professors in artificial intelligence asked about academic background and the main type of tool used for research. The background was specified as computer science (CS), mathematics (MATH), psychology (PSY), and linguistics (LIN). The most important tool of investigation was specified as statistics (STAT), mathematical optimization (MATHOPT), and logic (LOGIC). About 40 percent of the professors responded to the survey.

Statistics Exam - 2016 Q2

797 data entries

Four different professions: Computer Science, Linguistics, Maths, Psychology

Three different fields: Logic, Maths, Statistics

Tool	Background
"MATHOPT"	"PSY"
"LOGIC"	"MATH"
"STAT"	"LIN"
"MATHOPT"	"LIN"
"MATHOPT"	"PSY"
"STAT"	"CS"
...	...

Categorical data

Statistics Exam - 2016 Q2

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment
- Comment on the population and on the sample
- Research question

Statistics Exam - 2016 Q2

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment

  AI professors were asked about their main tool for research...

- Comment on the population and on the sample

  797 professors responded, 40% of the total asked

- Research question

  For each field, which tool is preferred the most my professors?

Statistics Exam - 2016 Q2

Perform an informal analysis to explore the research question. [6]

	Logic	Maths	Statistics
Computer Science	52	37	38
Linguistics	38	39	115
Maths	112	122	33
Psychology	35	33	143

- Computer science professors prefer logic
- Linguistic professors prefer statistics
- Maths professors prefer maths
- Psychology professors prefer statistics

Statistics Exam - 2016 Q2

Perform a formal analysis testing a relevant hypothesis. [8]

`chisq.test(...)`

Statistics Exam - 2016 Q2

Perform a formal analysis testing a relevant hypothesis. [8]
Computer Science Professors

R Output

Chi-squared test for given probabilities data: AI[1,] X-squared = 3.3228, df = 2, p-value = 0.1899

Computer science professors prefer logic *FALSE*
Linguistics Professors

R Output

Chi-squared test for given probabilities data: AI[2,] X-squared = 60.969, df = 2, p-value = 5.765e-14

Linguistic professors prefer statistics *TRUE*

Statistics Exam - 2016 Q2

Perform a formal analysis testing a relevant hypothesis. [8]
Maths Professors

R Output

Chi-squared test for given probabilities data: AI[3,] X-squared = 53.416, df = 2, p-value = 2.517e-12

Maths professors prefer maths *TRUE*
Psychology Professors

R Output

Chi-squared test for given probabilities data: AI[4,] X-squared = 112.64, df = 2, p-value ¡ 2.2e-16

Psychology professors prefer statistics *TRUE*

Statistics Exam - 2016 Q2

Give the conclusion including the answer on the research question. [5]

- What was our hypothesis?
- What tests did we do? What did it show?
- Did the test help us to reject our hypothesis?
- Does it answer our research question?

Statistics Exam - 2016 Q2

Provide one point of discussion about critical aspects of the analysis. [4]

Statistics Exam - 2016 Q2

Provide one point of discussion about critical aspects of the analysis. [4]

- The survey was subjective - professors were asked to think
- 40% of the population responded, is this enough?

Statistics Exam - 2016 Q3

A large ICT company is testing the performance of new programmers. During an assessment several candidates write a programme during four hours solving an informatics problem. The quality of each programme is tested and marked as either successful (1) or unsuccessful (0).
A personnel consultant of the company wishes to find out which factors have an impact on the success of a particular programmer. She collects data with respect to Age of the Programmer (AoP), Weeks of Experience (WoE) with the same task, Mean Level of Performance on psychological tests (MLoP), and Weeks of Experience as a Consultant (WoEaC).

Statistics Exam - 2016 Q3

300 data entries

166 were successful (55.3%)

Quality	AOP	WoE	MLoP	WoEaC
1	46.950	97.144	227.958	187.514
0	49.795	98.391	211.000	187.385
1	54.341	98.662	220.610	172.988
0	45.822	92.169	217.519	177.773

Categorical and numerical data

Statistics Exam - 2016 Q2

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment
- Comment on the population and on the sample
- Research question

Statistics Exam - 2016 Q2

Give a research question corresponding to the situation at hand. In your answer comment on the population and on the sample. [5]

- Brief description of the experiment

  ICT company is testing performance...variables collected were Quality, AOP, WoE, MLoP, WoEaC...

- Comment on the population and on the sample

  The population is really specific, can this data be used elsewhere?

- Research question

  Which factors have an impact on a programmers success?

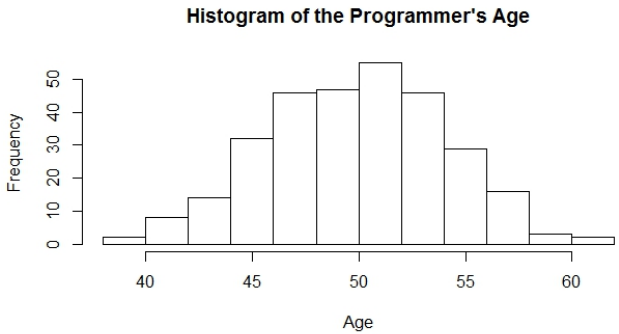
Statistics Exam - 2016 Q3

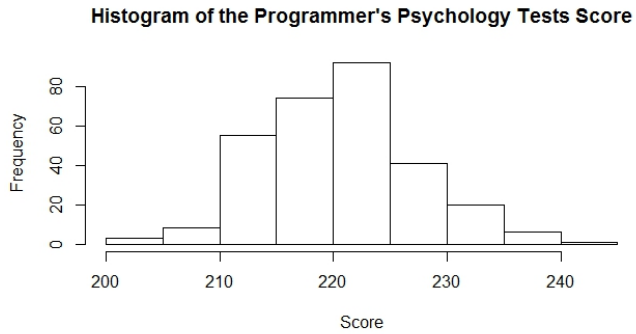
Perform an informal analysis to explore the research question. [6]

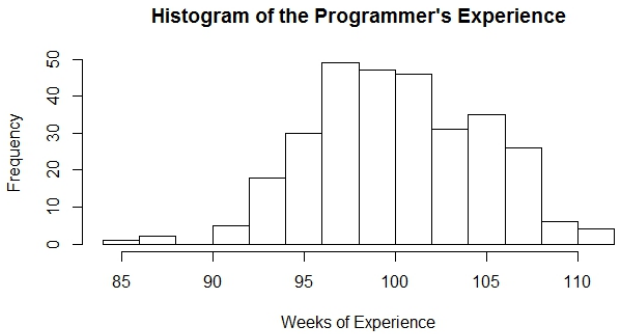
- Is the data normal?
- Is there a difference in the two groups?

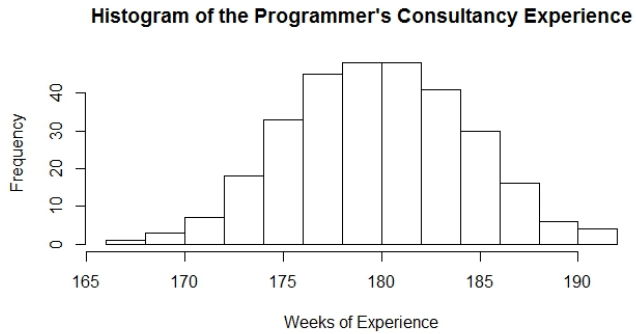
Statistics Exam - 2016 Q3

Perform an informal analysis to explore the research question. [6]



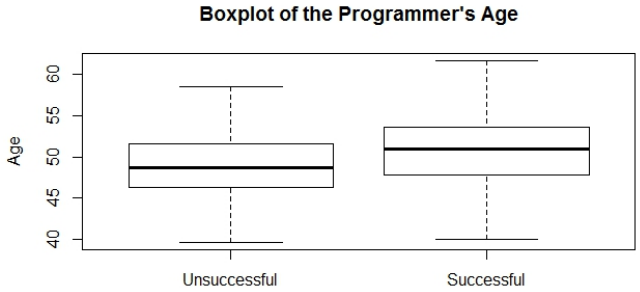


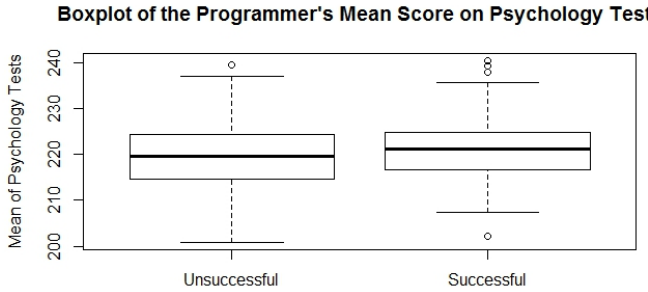


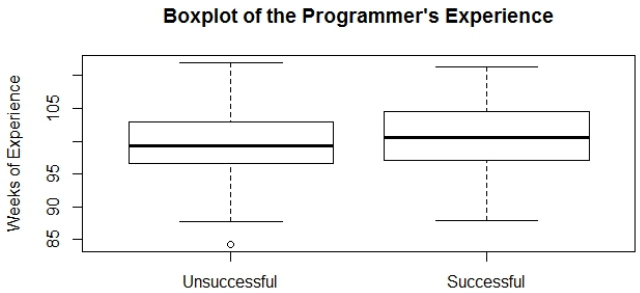


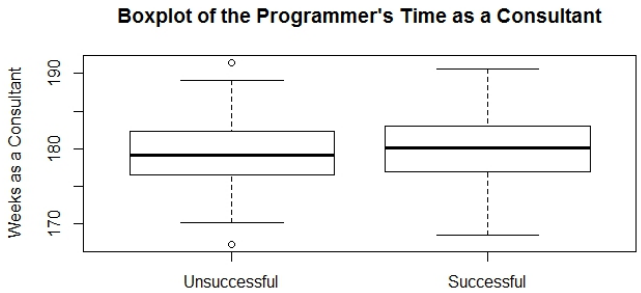
Statistics Exam - 2016 Q3

Perform an informal analysis to explore the research question. [6]









Statistics Exam - 2016 Q3

Perform a formal analysis testing a relevant hypothesis. [15]

- Lots of variables
- Quality is based on categorical data

Statistics Exam - 2016 Q3

Perform a formal analysis testing a relevant hypothesis. [15]

R Output

```
glm(formula = Quality   AoP + WoE + MLoP + WoEaC, family =
binomial(), data = dat3)
Coefficients:
             Estimate  Std.  Error  z value  Pr(>|z|)
 (Intercept)  -23.084      7.575     -3.047     0.002    **
      AoP       0.096      0.029      3.241     0.001    **
      WoE       0.054      0.026      2.078     0.037    *
     MLoP       0.038      0.018      2.114     0.034    *
    WoEaC       0.025      0.027      0.940     0.347
AIC: 402.88
```

How can we improve the model?

Statistics Exam - 2016 Q3

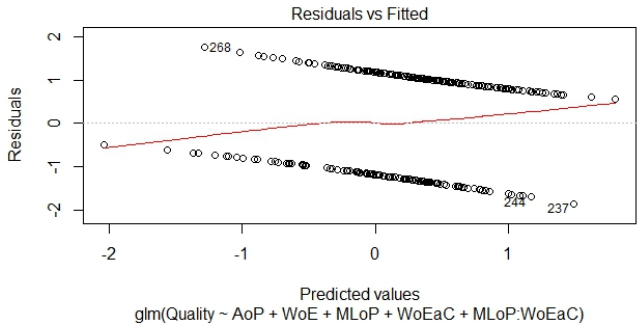
Perform a formal analysis testing a relevant hypothesis. [15]

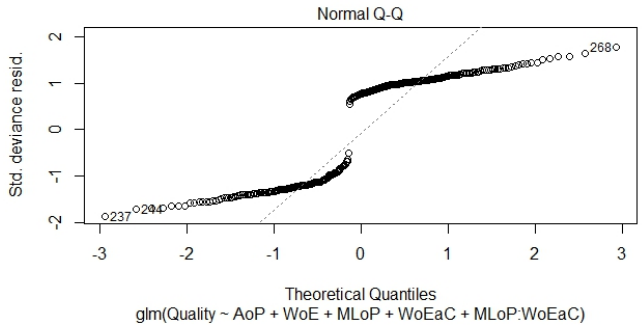
R Output

```
glm(formula = Quality   AoP + WoE + MLoP + WoEaC + MLoP:WoEaC,
family = binomial(), data = dat3)
Coefficients:
              Estimate   Std.  Error   z value   Pr(>|z|)
 (Intercept)  -334.688     168.975     -1.981     0.047     *
     AoP         0.100       0.030      3.353     0.000   ***
     WoE         0.055       0.026      2.123     0.033     *
    MLoP         1.448       0.763      1.897     0.057     .
   WoEaC         1.750       0.934      1.874     0.060     .
 MLoP:WoEaC     -0.007       0.004     -1.849     0.064     .
AIC: 401.37
```

Statistics Exam - 2016 Q3

Perform a formal analysis testing a relevant hypothesis. [15]





Statistics Exam - 2016 Q3

Give the conclusion including the answer on the research question. [5]

Statistics Exam - 2016 Q3

Provide one point of discussion about critical aspects of the analysis. [4]

Statistics Exam - 2016 Q3

Provide one point of discussion about critical aspects of the analysis. [4]

- Does the data fit the plots?
- Are there enough variables?
- Could they have chosen better variables? (length of time working, education level, etc.)

Where can you find the slides?

Slides can be found at
`https://studysupport.svcover.nl/`

Good Luck!


p = 0.02
!
p < 0.05
inhales
THUS THE NULL HYPOTHESIS IS REJECTED AND THE OBSERVED DIFFERENCES CANNOT BE ATTRIBUTED TO CHANCE ALONE