Lab 9

Goal: To gain experience with an ANOVA test, and use Rcmdr to run ANOVA tests as well as scatterplots.

Part 1 – ANOVA

A pharmaceutical company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment, 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 = extreme pain 30 minutes after taking the drug.

										$\bar{y}$	s
A:	4	5	4	3	2	4	3	4	4	3.6667	0.8660
B:	6	8	4	5	4	6	5	8	6	5.7778	1.4814
C:	6	7	6	6	7	5	6	5	5	5.8889	0.7817
Overall (all data pooled together)										5.1111	1.4763

Use Rcmdr to compute the overall sample average and standard deviation, as well as the average and standard deviation for each of the three drugs. You can import the data into Rcmdr using the **Ch28_Analgesics.xls** file on the course CMS website (see the Lab 9 assignment).

Q1] State the null and alternative hypotheses.

Perform an ANOVA by Hand.
Just to see where the number in our computer outcome come from, we'll do the ANOVA by hand and using the computer and compare the two. We need to fill in the table below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Treatment	2	28.2223	14.1111	11.91	< 0.01
Error	24	28.4444	1.1852		
Total	26	56.6667			

We have a total of 27 observations, so the total degrees of freedom is $n - 1 = 26$. There are 2 degrees of freedom for DRUG ($k - 1$) and 24 degrees of freedom for error.

Drug	Pain	Total Sum of Squares $(y_i - \bar{y})^2$	Error Sum of Squares $(y_i - \bar{y}_i)^2$
A	4	$(y_i - \bar{y})^2$ $= (4 - 5.1111)^2$ $= 1.2345$	$(y_i - \bar{y})^2$ $= (4 - 3.6667)^2$ $= 0.1111$
A	5	0.0123	1.7778
A	4	1.2346	0.1111
A	3	4.4568	0.4444
A	2	9.6790	2.7778
A	4	1.2346	0.1111
A	3	4.4568	0.4444
A	4	1.2346	0.1111
A	4	1.2346	0.1111
B	6	0.7901	0.0494
B	8	8.3457	4.9383
B	4	1.2346	3.1605
B	5	0.0123	0.6049
B	4	1.2346	3.1605
B	6	0.7901	0.0494
B	5	0.0123	0.6049
B	8	8.3457	4.9383
B	6	0.7901	0.0494
C	6	0.7901	0.0123
C	7	3.5679	1.2346
C	6	0.7901	0.0123
C	6	0.7901	0.0123
C	7	3.5679	1.2346
C	5	0.0123	0.7901
C	6	0.7901	0.0123
C	5	0.0123	0.7901
C	5	0.0123	0.7901

Compute the total sum of squares by taking each data point, subtracting off the overall mean, squaring the difference, and summing them all up.

56.6667

To get the error sum of squares, do the same thing, but instead subtract off the mean of the group each particular data point was in.

28.4444

The treatment sum of squares is the total sum of squares minus the error sum of squares.

28.2223

Fill in these quantities on the ANOVA table on the previous page.

Compute the mean square treatment (MS_{Tr}) by taking the treatment sum of squares and dividing by its degrees of freedom.

Computer the mean square error (MS_E) by taking the error sum of squares and dividing by its degrees of freedom.

Compute the *F*-statistic as the ratio of the two above quantities, and fill them in on the ANOVA table above.

P-values can be computed by hand using an F-table. To use it, note the degrees of freedom for the numerator (MS_{Tr}) and the denominator (MS_E). Here, we have 2 numerator and 24 denominator degrees of freedom. Go to the column corresponding to 2, and the row corresponding to 24. We have an F-value of 5.61. Note that the table says $\alpha = 0.01$ in the upper left corner. If our particular *F*-statistic is larger than 5.61, then it has a *p*-value < 0.01.

Table F   Numerator df

$\alpha = .01$	1	2	3	4	5	6	7	8	9	10	11	12
1	4052.2	4999.3	5403.5	5624.3	5764.0	5859.0	5928.3	5981.0	6022.4	6055.9	6083.4	6106.7
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.41	99.42
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	3.06	2.99

Denominator df

Q2] Based on your *F*-statistic and *p*-value, what do you conclude?

I conclude that we have evidence to reject the null hypothesis, and conclude that at least one of the means is different. Further testing could be done using multiple comparisons.

An ANOVA using Rcmdr
- Import the **Ch28_Analgesics.xls** file into R commander.
- Go to **Statistics → Means → One-way ANOVA**
- Select DRUG for the groups and PAIN for the response
- Click **OK**

Q3] Does the ANOVA table from R commander match the one you made by hand? It should!

Yes it does!

```
> summary(AnovaModel.1)
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
Drug         2  28.222  14.1111   11.906  0.0002559 ***
Residuals   24  28.444   1.1852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

Q4] Do the assumptions needed for ANOVA seem to be satisfied?

Not really. There's an outlier, and some of the boxes look very strange due to skewness.
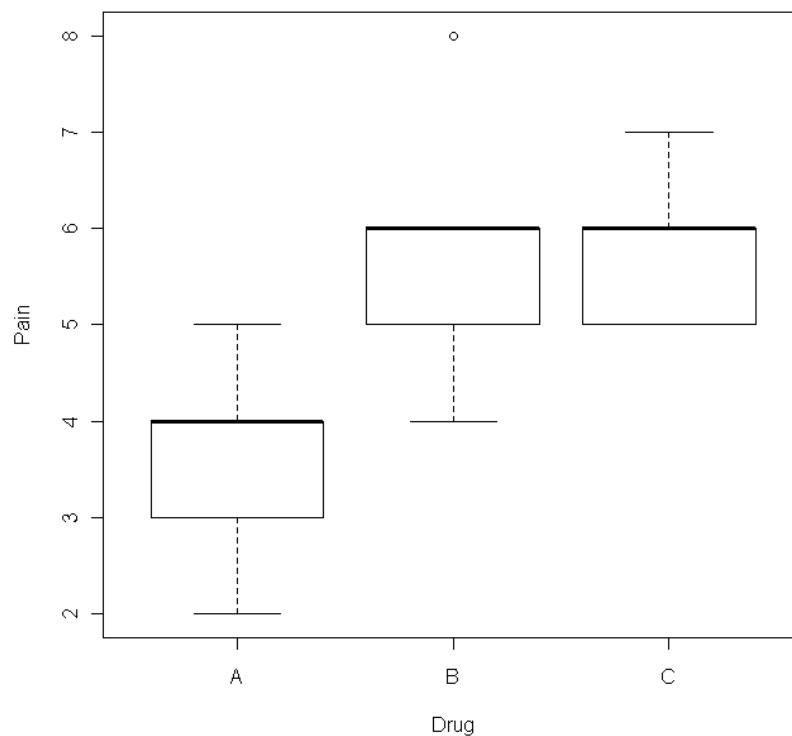
Part 2 – Scatterplots

In this part of the lab, we'll work with a scatterplot. Which mountain bike should we buy? Is there a relationship between the weight of a mountain bike and its price? A lighter bike is often preferred but do lighter bikes tend to be more expensive? The following table gives data on price and weight for 12 brands.

Brand and Model	Price ($)	Weight(lbs)
Trek VRX200	1000	32
Cannondale Super V400	1100	31
GT XCR-4000	940	34
Specialized FSR	1100	30
Trek 6500	700	29
Specialized Rockhopper	600	28
Haro Escape 7.1	440	29
Giant Yukon SE	450	29
Mongoose SX 6.5	550	30
Diamondback Sorrento	340	33
Motiv Rockbridge	180	34
Huffy Anorak 36789	140	37

We are interested in whether and how weight affects price.

Q5] Which variable is the logical choice for the explanatory variable?

Weight should be the explanatory variable.

Which variable is the logical choice for the response variable?

Price is the logical choice for the response variable.

Q6] Create a scatterplot using Rcmdr. To do this, import the bike data file from the CMS page. The file is **Lab9_Bikedata.xls**. Then, go to **Graphs → Scatterplot.** Select the x- and y-variables, and uncheck all the option checkboxes. Click on **OK**.

Does the relationship look linear?

No it does not.

In what way does it deviate from linearity?

It seems almost arced in shape.
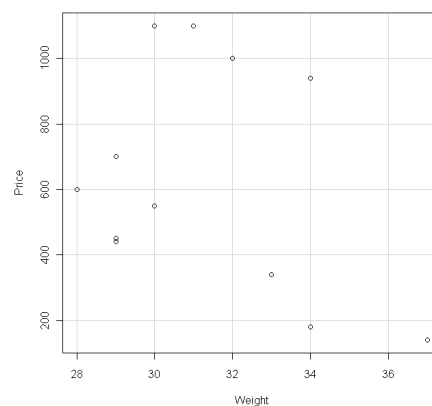
Part 3 –Correlation

In this part of the lab, we'll work with correlation. Suppose you want to study how well eyesight predicts your reaction time. You measure 10 people, measuring their eyesight in diopters and their reaction time in seconds. The data are shown below:

Subject	Eyesight	Reaction Time
1	0.20	0.11
2	1.30	0.16
3	2.20	0.12
4	2.70	0.23
5	3.30	0.27
6	4.80	0.28
7	5.10	0.31
8	6.40	0.36
9	8.10	0.32
10	8.21	0.42
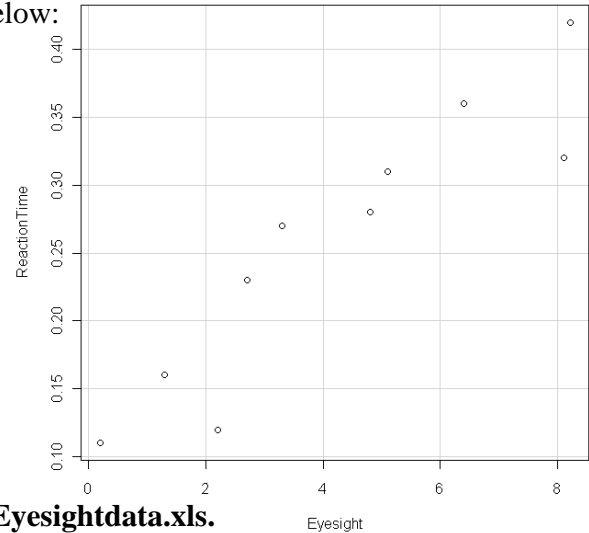
These data may be found on the CMS site as **Lab9_Eyesightdata.xls.**

Q7] Which variable is the explanatory variable? Which is the response variable?

   Eyesight is the explanatory variable and reaction time is the response.

Q8] Use Rcmdr to draw a scatterplot of the data is given above.
Do you think r will be close to –1, 0, or 1?   I think it will be close to 1.

Q9] Describe the direction, form, and strength of the association between eyesight and reaction time?

   We have a positive, linear association that is quite strong.

Q10] Calculate and interpret r? Go to **Statistics → Summaries → Correlation Matrix**.
Select both variables. You can hold down the shift key or ctrl key to select multiple variables.
Leave the "Type of Correlations" button set on "Pearson product-moment."

```
> cor(Eyesight[,c("Eyesight","ReactionTime")],
use="complete.obs")
            Eyesight ReactionTime
Eyesight    1.000000     0.922299
ReactionTime 0.922299    1.000000
```

   We have r = 0.92, indicating a strong, positive linear association.

Q11] Suppose we have an eleventh data point with an observation of 7.1 diopters for eyesight and a reaction time 0.15 seconds. What is the correlation (*r*) with this new observation included? You can click the "Edit data set" box in Rcmdr and type in this new observation. Then you can re-run the scatterplot and correlation computations. This should illustrate the impact of outliers on the correlation.

We have *r* = 0.73. The outlier changed things quite a bit!

```
> cor(Eyesight[,c("Eyesight","ReactionTime")],
use="complete.obs")
                Eyesight ReactionTime
Eyesight       1.0000000    0.7316096
ReactionTime 0.7316096    1.0000000
```