Statistics- SPSS Sample Assignment

1.) **Diets and lipid levels**: (Kutner, Nachtshiem, Neter and Li, 2004, page 913). A study was conducted to examine the effects of three experimental diets with different fat contents, which we will call diets 1, 2, and 3, on total lipid levels in plasma of human subjects. Total lipid level is a widely used predictor of coronary heart disease. Diet 1 has an extremely low fat content, Diet 2 has a fairly low fat content, and Diet 3 has a moderately low fat content. Fifteen male subjects who were within 20 percent of their ideal body weight were grouped into five blocks according to age. Within each block, one of the three men was randomly assigned to each of the three diets. Data on reduction in lipid levels (in grams per liter of plasma) were recorded after the subjects had all been on the assigned diets for the same fixed period of time. The data are shown in the following table and the data are posted in the file **lipid_levels.csv**. This file has one row for each observation and three columns corresponding to lipid levels, blocks, and diets, respectively.

Block	Diet 1	Diet 2	Diet 3
1. Ages 15-24	0.73	0.67	0.15
2. Ages 25-34	0.86	0.75	0.21
3. Ages 35-44	0.94	0.81	0.26
4. Ages 45-54	1.40	1.32	0.75
5. Ages 55-64	1.62	1.41	0.78

Consider the model $Y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}$, where Y_{ij} is the lipid concentration for the subject assigned to the j-th diet in the i-th block. Assume block effects are fixed effects when you do parts (a) through (g) of this problem. We will switch to random block effects in part (h). Assume the random errors are independent realizations from a normal distribution with mean zero and unknown variance σ^2_{error} . For this model, JMP will constrain the diet effects to add to zero (i.e., $\alpha_1 + \alpha_2 + \alpha_3 = 0$).

a) Use JMP to compute the ANOVA table and copy the output to your solution. Note that JMP reports the ANOVA table in two pieces. Report both pieces. (To do this part you will need to temporarily treat the block effects as fixed effects when running the "Fit Model" option in JMP.)

b) Clearly state the null and alternative hypotheses for each F-test in the two pieces of the ANOVA table reported in part (a). For each F-test state the null hypothesis in words and also in terms of parameters in the model presented above. Just state the alternative hypothesis in words (basically the alternative is that the null hypothesis is incorrect). Report the p-value for each F-test, and report your conclusion for each F-test.

c) List each conditions that must be checked to determine if an F-test can be accurately used to test the null hypothesis that the mean responses are the same for all three diets. For each condition, indicate if you believe that it is reasonably satisfied. Where possible, use information from the data, including appropriate graphs, to support your conclusions.

d) Using the Tukey HSD method for multiple comparisons, determine which diets have significantly different means for lipid concentrations. Use a family-wide type I error level of 0.05. Clearly state your conclusions.

e) Explain what μ and α_1 represent in the model displayed above. Also explain what $\mu+\alpha_1$ represents in the model displayed above.

f) Report the least squares estimates of μ and α_1, and report the corresponding standard errors. Also report the least squares estimate of $\mu+\alpha_1$ and its standard error.

g) Report the least squares estimates of α_3 and $\mu+\alpha_3$, and report the corresponding standard errors.

h) Now consider using random block effects in the model $Y_{ij}=\mu+\alpha_j+\beta_i+\varepsilon_{ij}$. The random block effects are modelled as a random sample of values from a normal population of possible block effects that has zero mean and variance σ^2_{blocks}. The random errors are still considered to be an independent random sample from another normal distribution with mean zero and variance σ^2_{error}. Use JMP to fit the model with random block effects to the data. Report the ANOVA table. Is the F-test of the null hypothesis that the mean responses are the same for all three diets affected by switching from fixed to random block effects?

i) For the model with random block effects, use JMP to obtain estimates of the variance of the block effects, σ^2_{blocks}, and the variance of the random errors, σ^2_{error}.

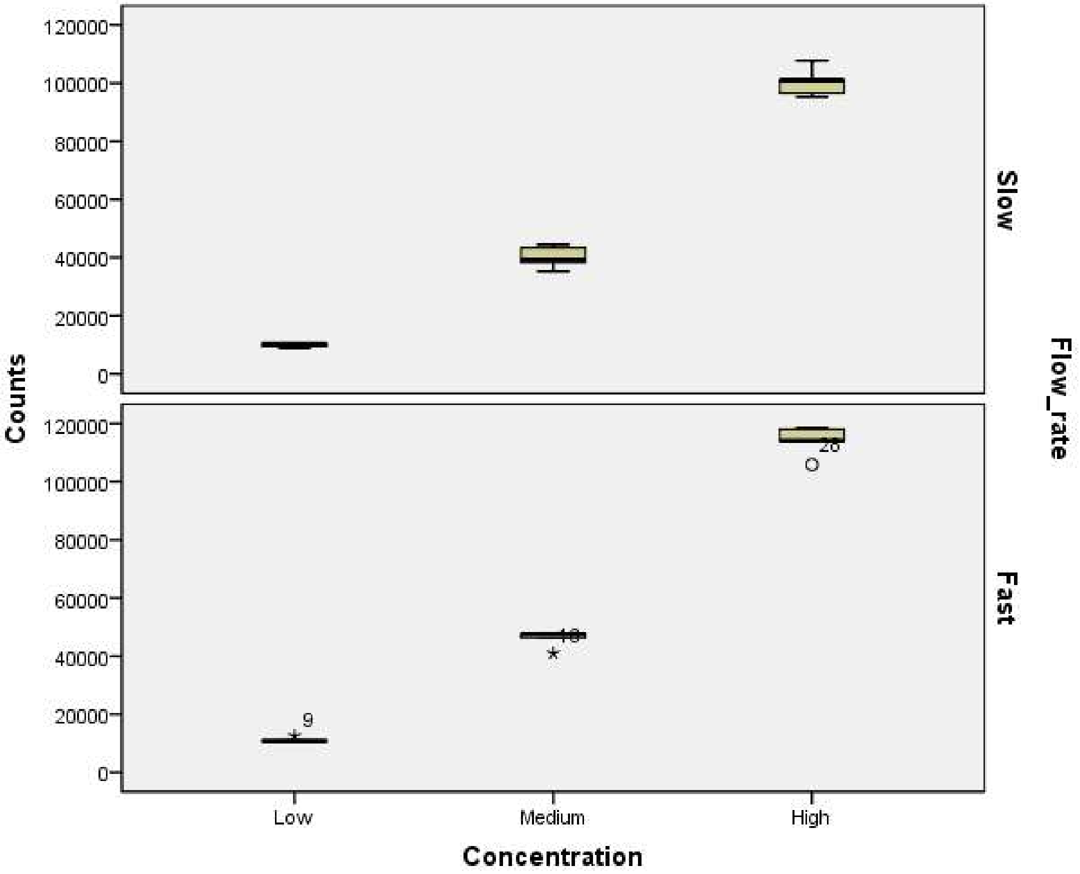
j) Does it appear that age is a good characteristic to use to create blocks? Explain.

2.) **Chromatography**: A gas chromatograph is an instrument that measures the amounts of various chemical compounds in samples of foods or other substances. Because different compounds travel through the system at different rates, chromatographers are able to identify and measure the level of different chemical compounds in a sample of a particular food or other substance. The instrument records counts that are proportional to the amounts of different chemical compounds in the sample. By looking at the counts at various times, a chemist is able to determine relative amounts of various compounds in the substance.

An experiment was performed to determine if changing the flow rate of the sample through the instrument would affect the observed counts. Mixtures were prepared with three different concentrations of a particular chemical compound (low, medium, and high concentrations). Two flow rates were used (slow and fast). Ten samples were taken from low concentration mixture and five were randomly selected to run through the instrument at a slow flow rate and the other five were run through the instrument at the fast flow rate. This was repeated for the mixtures with medium and high concentrations. Altogether, 30 samples were run through the instrument and a count was recorded for each of the 30 samples. The data are posted in the Data Files folder on Blackboard in the file These data are from problem 22 at the end of Chapter 27 in the fourth edition of the DeVeaux, Velleman, and Bock book. (This is also a problem at end of Chapter 29 in the third edition of the DeVeaux, Velleman, and Bock book.)

a) Use JMP to produce side-by-side box plots for the six treatments corresponding to the combinations of the two levels of the flow rate factor and the three levels of the concentration factor. Comment on what this display reveals.

There is an increasing trend from low through medium to high concentration of the count rates for both types of flow rate. However, the fast rate has slightly higher counts. For all six variants (concentration by flow rate) there is not much variability in the counts, which can be seen by the form of the boxes in the boxplots below. There are outliers in the fast rate boxplot, one in each concentration, while there are no outliers in the slow rate.



b) Compute the ANOVA table for the model containing main effects for the flow rate factor, main effects for the concentration factor, and interaction between the flow rate and concentration factors. Show your JMP output for the ANOVA table in this part of the assignment.

The ANOVA table below reveals that the model is in general significant at the 5% level (F=797.006 with p-value < 0.001). All the factors, concentration, flow rate and their interaction are significant at the 5% level (all p-values <0.001).

Tests of Between-Subjects Effects

Dependent Variable:Counts

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4,893E10	5	9,787E9	797,006	,000
Intercept	8,626E10	1	8,626E10	7024,918	,000
Concentration	4,837E10	2	2,418E10	1969,424	,000
Flow_rate	3,640E8	1	3,640E8	29,645	,000
Concentration * Flow_rate	2,030E8	2	1,015E8	8,267	,002
Error	2,947E8	24	12279085,000		
Total	1,355E11	30			
Corrected Total	4,923E10	29			

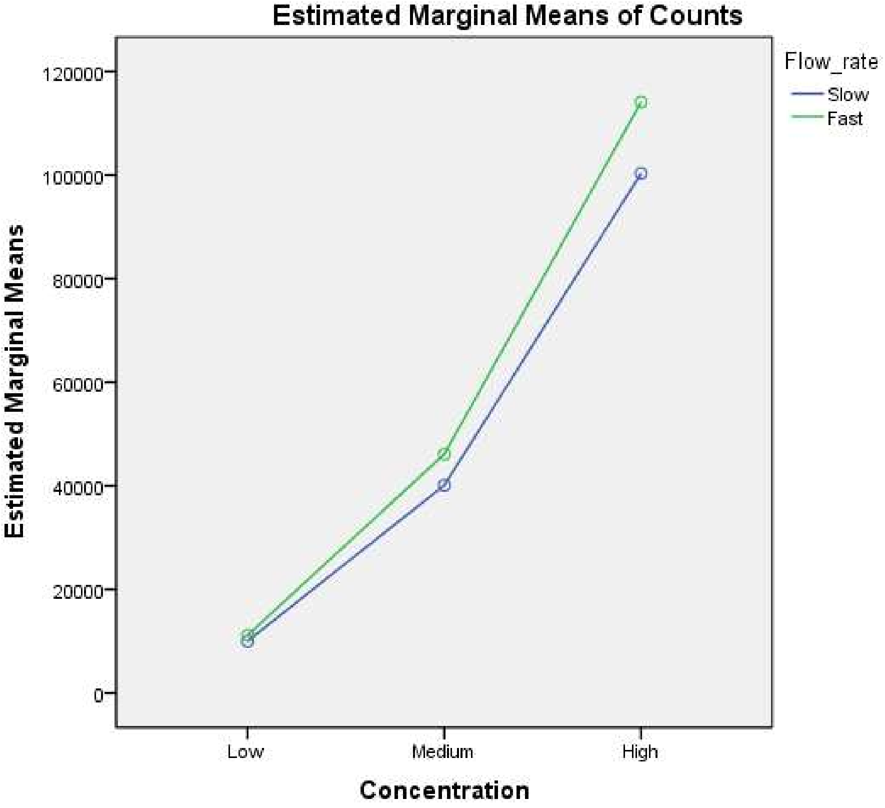
a. R Squared = .994 (Adjusted R Squared = .993)

c) Is there significant interaction between the flow rate and concentration factors? Report the value of an F-statistic, its degrees of freedom, and a p-value.

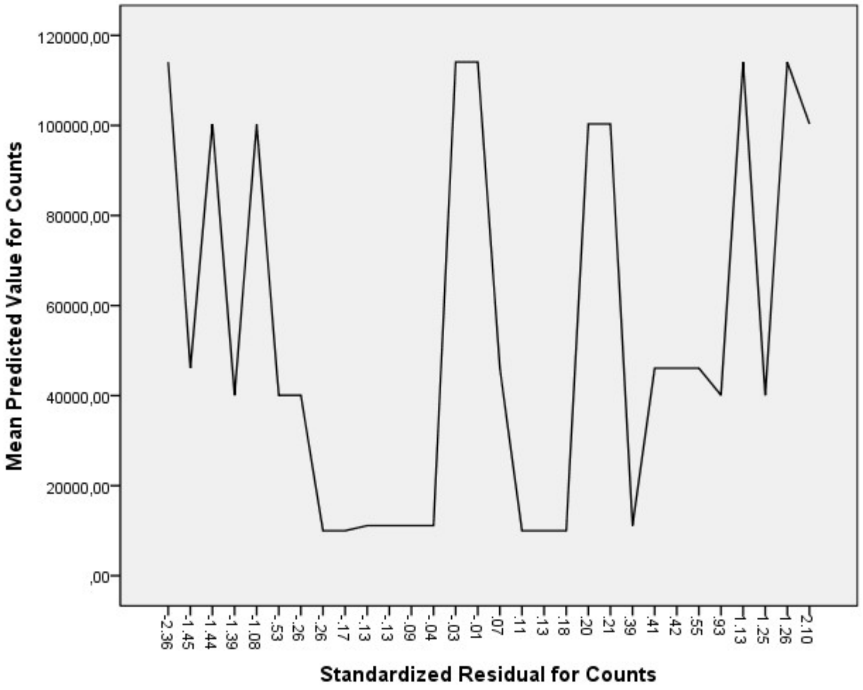
The interaction between the flow rate and the concentration is significant with F=8.267, df=2,24 and p-value =0.002.

d) Construct a profile plot. What does it indicate about the interaction between the flow rate and concentration factors?

The fast rate gives higher counts than those on the slow rate and on the higher levels of concentration on the fast rate the counts are higher than those on the lower levels of concentration.

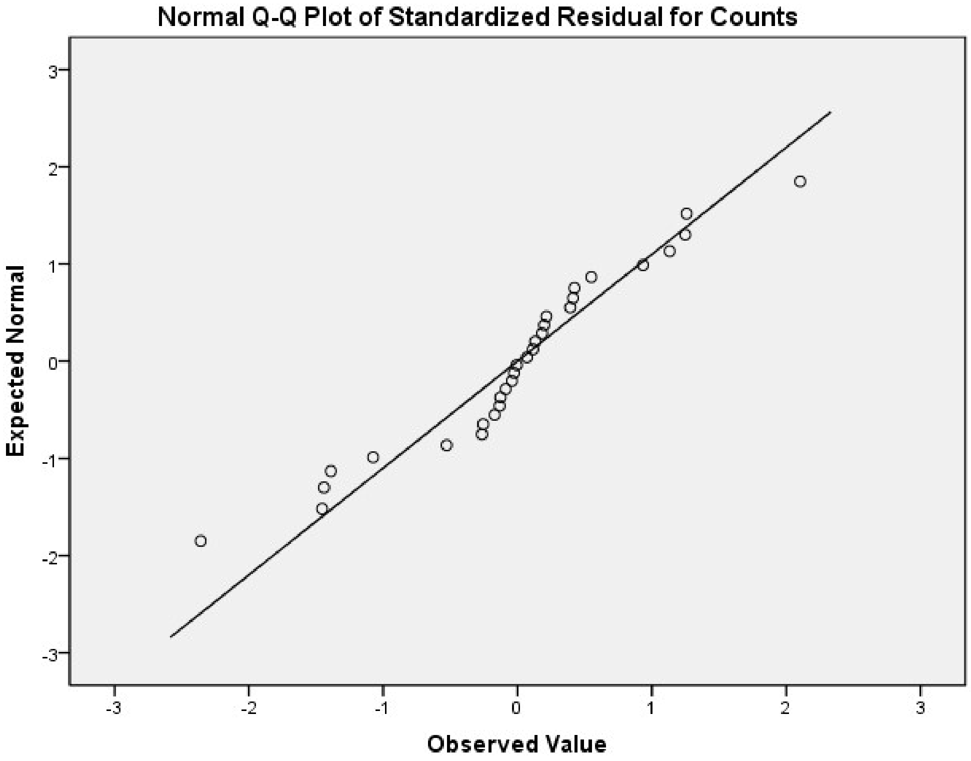


e) Plot the residuals against the predicted values (estimated mean responses). What does this plot indicate about homogeneity of variance condition? Are there any outliers or systematic patterns that might indicate some flaw in the model.



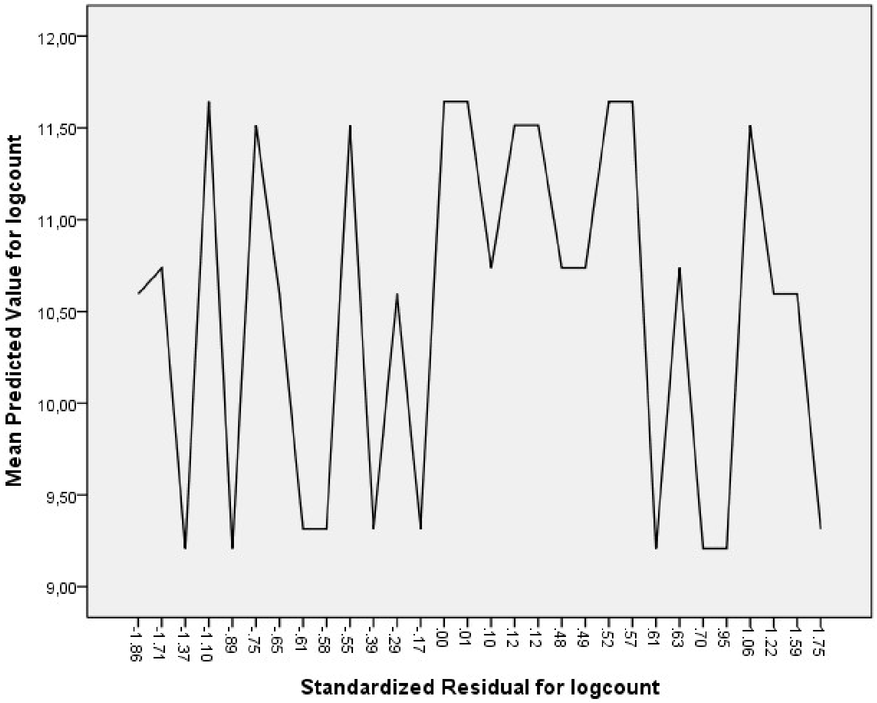
The plot indicates presence of heteroscedasticity. That is, the variance is not equal along all the residuals, there is some concentrations but no evident pattern.

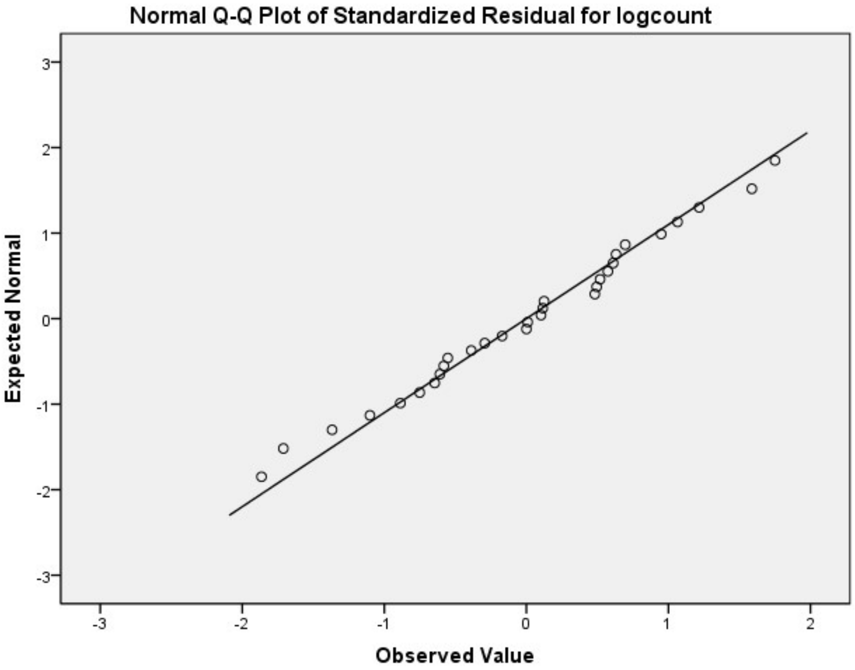
f) Use JMP to make a normal probability plot of the residuals. What does this plot indicate about the normality condition?



The assumption for normality of the residuals is violated as can be seen by the normality plot above. The observations are not over the normality diagonal line.

g) Apply the log transformation to the counts to promote homogeneity of variances (use the natural logarithm of each count as the new response). Re-fit the model with main effects and interaction using the log(count) as the response. Examine the plot of the residuals against the predicted values. Did the log transformation fix the non-homogeneity of variance problem? Also make a normal quantile plot of the residuals and comment on that plot.





Now, with the log of counts both issues with heteroscedasticity and normality of the residuals seem to be solved. The variance shows a random pattern with no concentrations just a slight deviation in the range 0:0.6, however the plot shows overall homoscedasticity. The residual normality plot seems to follow the diagonal line.

h) Report the ANOVA table for the log(count) responses. Does there appear to be a significant interaction between the levels of the flow rate factor and the levels of the concentration factor? Back up your response with the results of an F-test. Make a profile plot to examine the interaction and comment on that plot.

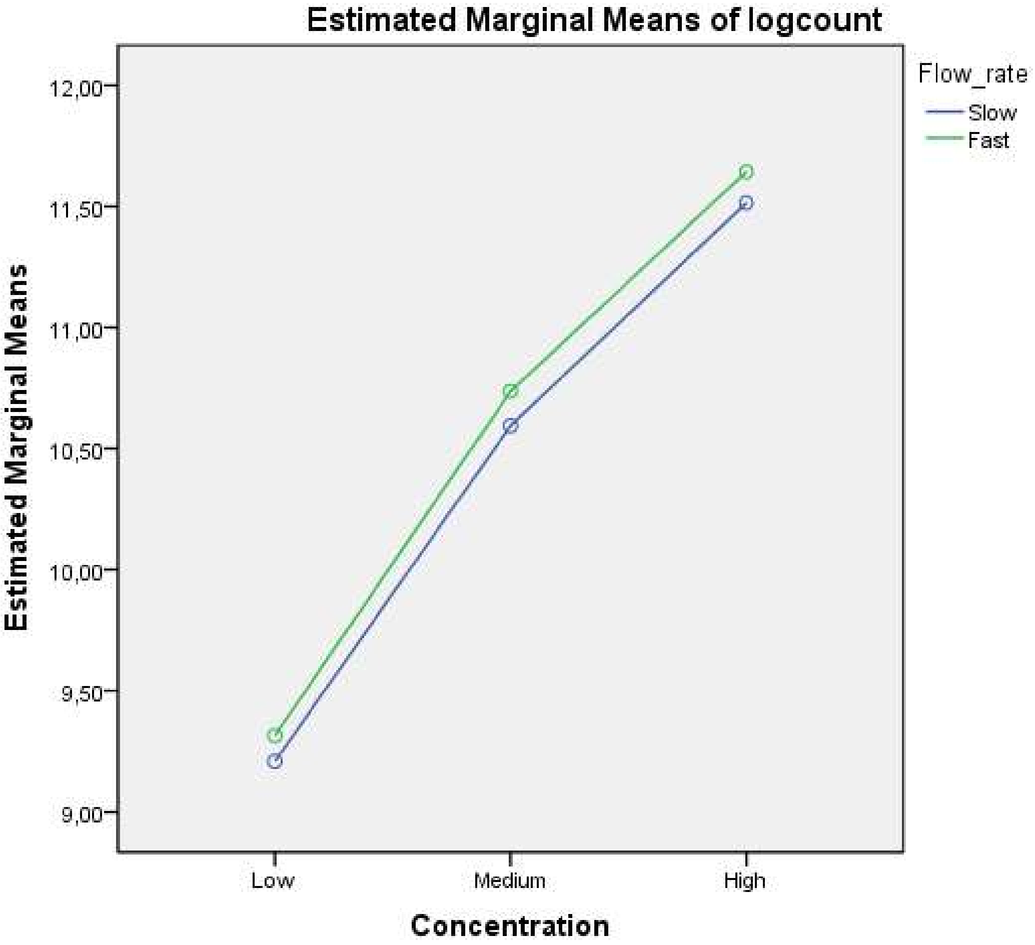
Now the interaction term is not significant with F=0.167 with p-value = 0.847.

Tests of Between-Subjects Effects

Dependent Variable:logcount

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	27,381[a]	5	5,476	1202,093	,000
Intercept	3309,068	1	3309,068	726382,103	,000
Concentration	27,261	2	13,630	2992,045	,000
Flow_rate	,119	1	,119	26,038	,000
Concentration * Flow_rate	,002	2	,001	,167	,847
Error	,109	24	,005		
Total	3336,559	30			
Corrected Total	27,490	29			

a. R Squared = .996 (Adjusted R Squared = .995)



i) Using the log(count) response, are there any significant main effects for flow rates? Report the results for an F-test, its degrees of freedom, and its p-value. State your conclusions.

The flow rate is significant at the 5% level with F=26.038 and p-value<0.001. Therefore, there is significant difference between the counts for slow and fast rates.

j) Using the log(count) response, are there any significant main effects for the concentration levels? Report the results for an F-test, its degrees of freedom, and its p-value. Use the Tukey HSD multiple testing procedure to determine which concentrations have different mean responses on the log(count) scale. State your conclusions.

The concentration is as well significant at the 5% level with F=2992.045 and p-value <0.001. That means that there is at least one significant difference between the counts for the three levels of concentration. Which exactly are the pairs for which the difference is significant the Tukey test shows. Apparently all the pairs differ significantly, between low and medium, low and high as well as between medium and high.

Multiple Comparisons

logcount
Tukey HSD

(I) Concentration	(J) Concentration	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Low	Medium	-1,4041*	,03018	,000	-1,4795	-1,3288
	High	-2,3177*	,03018	,000	-2,3931	-2,2424
Medium	Low	1,4041*	,03018	,000	1,3288	1,4795
	High	-,9136*	,03018	,000	-,9890	-,8382
High	Low	2,3177*	,03018	,000	2,2424	2,3931
	Medium	,9136*	,03018	,000	,8382	,9890

Based on observed means.
The error term is Mean Square(Error) = .005.
*. The mean difference is significant at the .05 level.

k) Summarize what you learned about the effects of flow rate and concentrations of the substance in the mixture on the counts recorded by the chromatography instrument.

An ANOVA was used to analyze the differences between the counts caused by two flow rates and three concentrations. The first model used the level form of the counts and revealed significant differences between the counts for each of the two flow rates, for each of the three concentrations and within the levels of concentration between the flow rates. However, those inferences were not reliable due to violations of the main assumptions of the model – presence of heteroscedasticity and lack of normality of the residuals. Therefore, a second ANOVA was performed with the log forms of the counts. That was the assumptions were met and it appeared that the flow rate and concentration alone have significant impact on the counts. The higher the concentration the higher counts and the faster the flow the more counts.

3) **Baldness and Heart Disease**: A retrospective study examined possible association between baldness and the incidence of heart disease. In the study, 1399 middle-aged men were selected at random and examined to see whether they shows signs of heart disease (or not) and what amount of baldness they exhibited (none, little, some, much). The counts are shown below.

Baldness	None	None	Little	Little	Some	Some	Much	Much
Heart Disease	Yes	No	Yes	No	Yes	No	Yes	No
Number of Men	241	322	154	224	195	184	53	26

The counts have also been entered into a csv file posted on Blackboard as **baldness.csv.**

(a) Perform an analysis of the data to determine if baldness and heart disease are related. (Use a 0,05 significance level for your test). You analysis should include the following steps:

Step 1: State the null and alternative hypotheses

Step 2: Check appropriate conditions for doing the test.

Step 3: Use JMP to compute the value of the test statistic and the corresponding degrees of freedom and p-value.

Step 4: State your conclusions in the context of the study.

(b) Do your conclusions support the claim that baldness is a cause of heart disease? Explain.