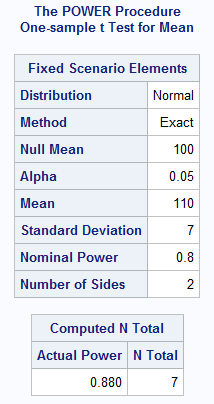
**Statistics 6371 Final Exam**

1. (8pts) Suppose a researcher wants to design a new study with a power of 0.8 and a significance of 0.05 to test whether the caffeine content for a brand of coffee is really 100mg. A previous study gave a mean caffeine level for this brand of 110 mg and a standard deviation of 7 mg. Use PROC POWER to determine how many cups of coffee need testing.

**The results and the code are shown below in Figure 1. You would need to include 7 cups of coffee in the sample.**

**proc power;**

**onesamplemeans**

**alpha = 0.05**

**nullmean = 100**

**mean = 110**

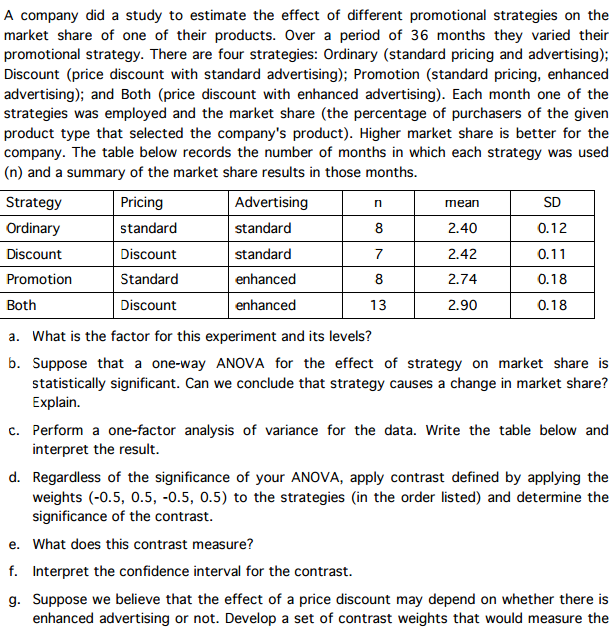
**stddev = 7**

**power = 0.8**

**ntotal = .;**

**run;**

**Figure 1 Proc Power SAS Code and Results**





* 1. The factor is strategy. It has four levels: Ordinary, Discount, Promotion, and Both.
  2. Because we do not know how the strategy was assigned each month (randomly or not), we cannot infer causality.

For completeness,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | Df | SS | MS | F | P-value |
| Model | 3 | 1.744 | 0.5813 | **23.6** | <0.00001 |
| Error | 32 | 0.789 | 0.024656 |  |  |
| Total | 35 | 1.5326 |  |  |  |

At p-value < 0.00001, (alpha = 0.05), there is sufficient evidence to reject the null hypothesis that all means are equal.

At significance level 0.05, there is insufficient evidence to reject the null hypothesis that the mean of the means of groups ordinary and promotion is significantly different from the mean of the means of groups discount and both.

* 1. The contrast measures the difference in the mean of the means of groups ordinary and promotion and the mean of the means of groups discount and both.

A 95% confidence interval for the difference in the mean of the means is:

At significance level 0.05, there is sufficient evidence to reject the null hypothesis that the mean of the means of groups ordinary and discount is significantly different from the mean of the means of groups promotion and both.

1. (12pts) It is suspected that an unnatural craving for substances such as paint might influence lead poisoning in children. A study was conducted to investigate this hypothesis. Ten of 20 rats were randomly assigned to a calcium-deficient diet (experimental group) and 10 to a regular diet (control group). Each of the rats occupied a separate cage and was monitored to determine the quantity of a 0.15% lead-acetate solution that they consumed during the study period. The amount each consumed is shown below. The researcher was interested in whether the calcium-deficient diet increased the consumption of the lead-acetate compared to the control group.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Rat*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** | ***9*** | ***10*** |
| ***Control*** | 5.4 | 6.2 | 3.1 | 3.8 | 6.5 | 5.8 | 6.4 | 4.5 | 4.9 | 4.0 |
| ***Experiment*** | 8.8 | 9.5 | 10.6 | 9.6 | 7.5 | 6.9 | 7.4 | 6.5 | 10.5 | 8.3 |

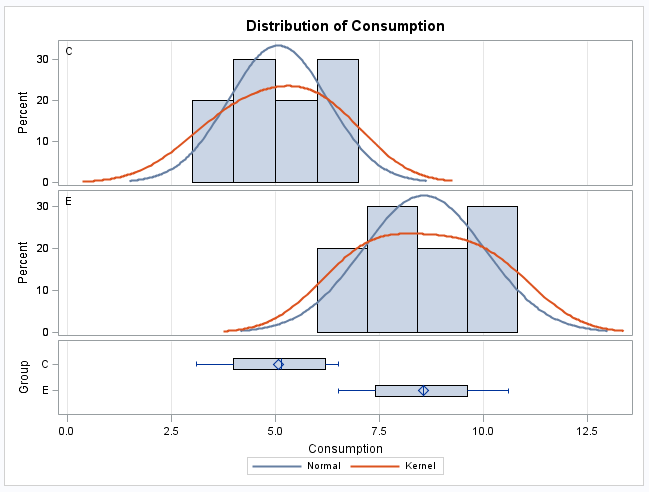
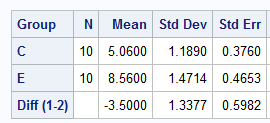
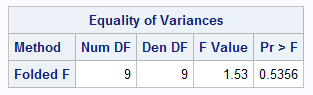
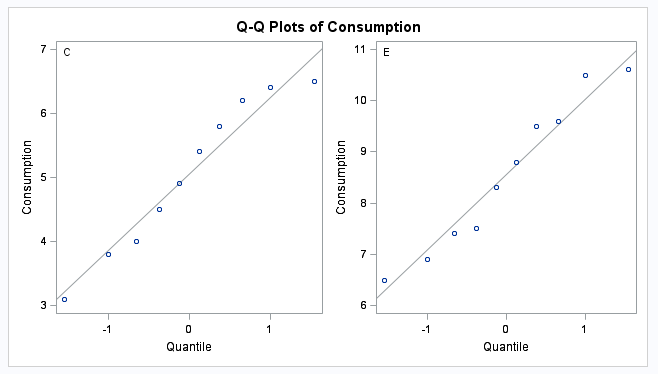
Perform a complete analysis of these data. (Show all 6 Steps.)

**Problem:** Does placing rats on a calcium-deficient diet cause an increase in the consumption of lead-acetate solution?

**Inferential Limitations:**

1. This is a randomized experiment, so casual inference can be drawn as a result of this analysis.

**Assumptions:** From an initial examination of the data, we find little evidence against normality in the two groups. The histograms and boxplots in Figure 1 are consistent with small sample size draws from a normally distributed population. Additionally, the QQ Plots in Figure 1 are fairly linear. The sample standard deviations are close to each other, but just to be sure since the normality assumption, is met we will examine the results of the Equality of Variances test. The test shows there is not enough evidence to reject the null hypothesis that the variances of the two samples are equal (p-value = 0.5356). We will assume the observations are independent and proceed with the test.

**Figure 2 Histograms, Boxplots, QQplots, Means and F-Test of Lead-Acetate Data**

**Test:** A two-sample t-test to compare the means of a dependent variable for two independent groups, conducted at the α = 0.05 level of significance.

**Step 1**: Null and Alternative Hypothesis are below

Ho : µE -µC = 0

Ha : µE -µC > 0

**Step 2**: Determine t-critical

t0.95, 20-2 = 1.729

**Step 3**: Calculate t-statistic based on the test statistic

The sample mean difference is 3.500

The pooled sample standard deviation (used because standard deviations were nearly equal) is 1.338.

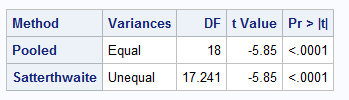
Compute the standard error of the average:

The standard error is 0.5984 with 19 degrees of freedom.

t-statistic = 5.85

**Step 4**:Calculate p-value

p-value = <0.0001 from Figure 3



**Figure 3 T-Test results**

**Step 5: Conclusion**

0.0001 < 0.05, Reject Ho

**Step 6: Relevant Conclusion:**

There is strong evidence to suggest at the α = .05 level of significance (p-value = <.0001, n = 20) that a calcium-deficient diet in rats caused an increase in the consumption of lead-acetate of 3.5 units. A 90% confidence interval for the increase is [2.465, 4.535].

1. (24 pts) As cheese ages, various chemical processes take place that determine the taste of the final product. This dataset contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample (variable taste – where a larger score is better). The variables "Acetic" and "H2S" are the natural logarithm of the concentration of acetic acid and hydrogen sulfide respectively.

data cheese;

input Case taste Acetic H2S;

datalines;

1 12.3 4.543 3.135

2 20.9 5.159 5.043

3 39 5.366 5.438

4 47.9 5.759 7.496

5 5.6 4.663 3.807

6 25.9 5.697 7.601

7 37.3 5.892 8.726

8 21.9 6.078 7.966

9 18.1 4.898 3.85

10 21 5.242 4.174

11 34.9 5.74 6.142

12 57.2 6.446 7.908

13 0.7 4.477 2.996

14 25.9 5.236 4.942

15 54.9 6.151 6.752

16 40.9 6.365 9.588

17 15.9 4.787 3.912

18 6.4 5.412 4.7

19 18 5.247 6.174

20 38.9 5.438 9.064

21 14 4.564 4.949

22 15.2 5.298 5.22

23 32 5.455 9.242

24 56.7 5.855 10.199

25 16.8 5.366 3.664

26 11.6 6.043 3.219

27 26.5 6.458 6.962

28 0.7 5.328 3.912

29 13.4 5.802 6.685

30 5.5 6.176 4.787

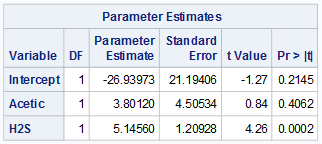
;

* 1. Which variable is the response variable?

The subjectively scored variable “Taste.” The larger score indicates better tasting cheese.

* 1. What is the equation of the regression line?

Regression Results from SAS are shown in Figure 4



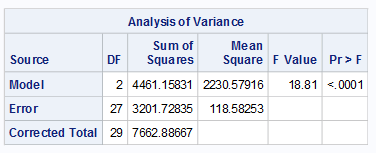
**Figure 4 Parameter Estimates from PROC REG**

µ{taste | Acetic, H2S} : taste = β0 + β1(Acetic) + β2(H2S)

µ{taste | Acetic, H2S} : taste = -26.94 + 3.80(Acetic) + 5.146(H2S)

* 1. Is the overall regression equation significant?

The overall regression equation is significant, based on the p-value shown in Figure 5. This shows that the model does a better job describing the variation in the response than the single parameter (equal means) model.

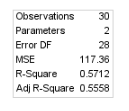
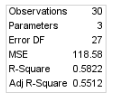


**Figure 5 P-value for Overall Regression Equation**

* 1. Are the slopes of the regression line statistically significant? Explain your answer.

Referring to Figure 4 the slope of the H2S variable is significant, with a p-value of 0.0002, meaning that the slope is significantly different than 0 at an α = 0.05. The slope of Acetic variable is not significant, with a p-value of 0.4062, meaning that the slope is not significantly different than 0 at an α = 0.05. These p-values are calculated by a hypothesis test where the null hypothesis is that the slope parameter is equal 0 and the alternative hypothesis is that the slope parameter is not equal to zero.

Practically, this means that we can eliminate the Acetic variable from the model without impacting the fit of the model. A quick comparison of R2 for both the 3 parameter model (with Acetic included) and the 2 parameter model (a simple linear regression based on H2S) in Figure 6 shows this is true.



**Figure 6 Comparison of R2 for 3 Parameter and 2 Parameter Model**

* 1. Interpret each slope.

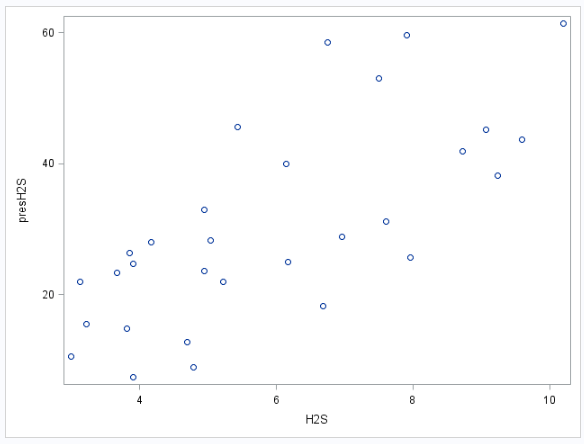
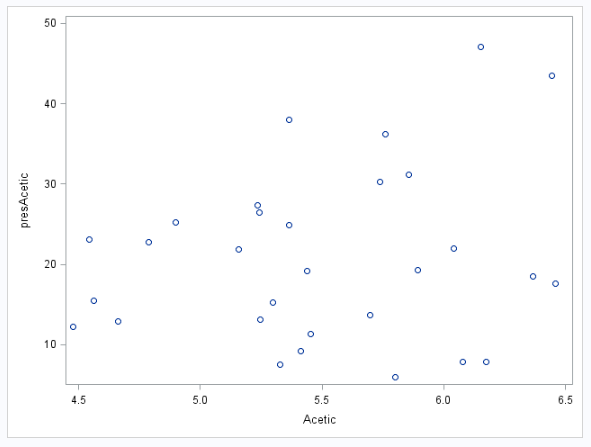
Of the two explanatory variables, the concentration of H2S is a statistically significant predictor of taste while the concentration of Acetic is not. Because the variable is log-transformed, holding the value of Acetic constant, a doubling of the concentration of H2S is associated with an increase of 3.567 in the mean of taste. While not statistically significant in this model, the best estimate of effect of increasing the Acetic concentration while holding the value of H2S constant is that a doubling of the concentration is associated with an increase of 2.634 in the mean of taste.

* 1. Obtain partial correlations for each variable and interpret their meanings.

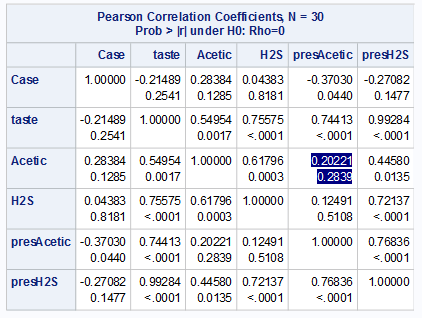
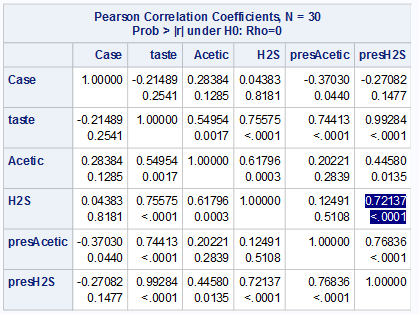
presH2S = taste - β0 - β1(Acetic) = taste + 26.94 - 3.80\*Acetic;

presacetic = taste - β0 - β2(H2S) = taste + 26.94 - 5.146\*H2S;

After computing the partial residual of both Acetic and H2S we are able to see clear evidence that the variation in taste is almost completely explained by the concentration of H2S, and that there is a much lower association between the concentration of Acetic and the taste response variable. The partial residual scatter plots show this in Figure 7, where the partial residual of Acetic on the left is a random cloud, while the partial residual for H2S on the right still demonstrates a linear relationship. The calculation of the correlation coefficient shown in Figure 8 between the partial residuals and the explanatory variables also shows this discrepancy, with values of 0.2022 and 0.7214 respectively. Clearly there is a much stronger relationship between the concentration of H2S than the concentration of Acetic.



**Figure 7 Partial Residuals for Acetic (left) and H2S (right) REALLY GOOD!**

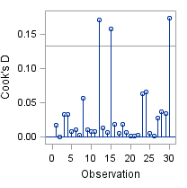
**Figure 8 Correlation Coefficients for Acetic (left) and H2S (right)**

* 1. Obtain a 90% confidence interval for population slope of the variable “H2S”. Interpret the interval.

A 90% CI for β1 is (-3.9, 11.5). A 90% CI for β2 is (3.1, 7.2).

* 1. Identify any leverage points in the data using any measure you like. Give a definition of the measure you chose. If you make your decision on the basis of several measures, then explain only one of them. If there are no leverage points, then explain why you came to that conclusion.

As shown in the Cook’s Distance graph in Figure 9, observations 12, 15, and 30 all have significant influence on the regression coefficients. Two of these are extremely large taste values, and one is an extremely low taste value. The threshold we are using for Cook’s Distance is 4/n, which equates to roughly 0.1333 for this 30 observation sample. Cook’s Distance is measure of overall influence on the regression coefficients, its numerator measures how much the fitted values change when the observation is deleted, and its denominator scales this difference based on the variance and the number of parameters in the model. Since each of these points is above the threshold, they could be problematic and should be investigated further. The fact that the taste variable was a subjective measure means possible that these unusually large positive and negative responses could be biased due to the influence of the taster.

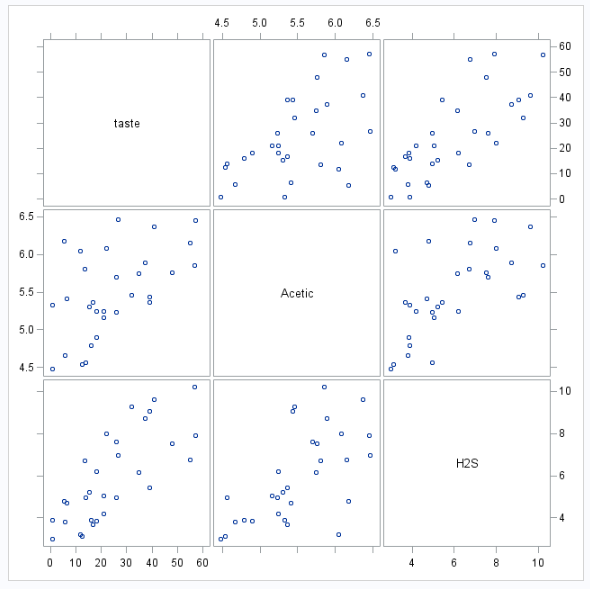
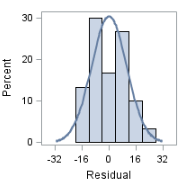
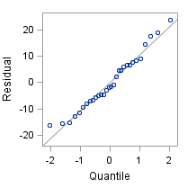
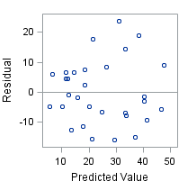
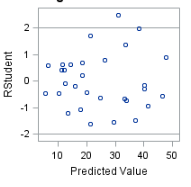


**Figure 9 Cook’s Distance for Cheese Study**

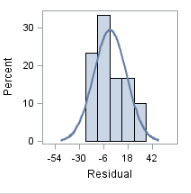
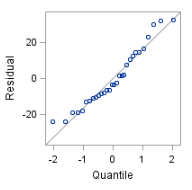
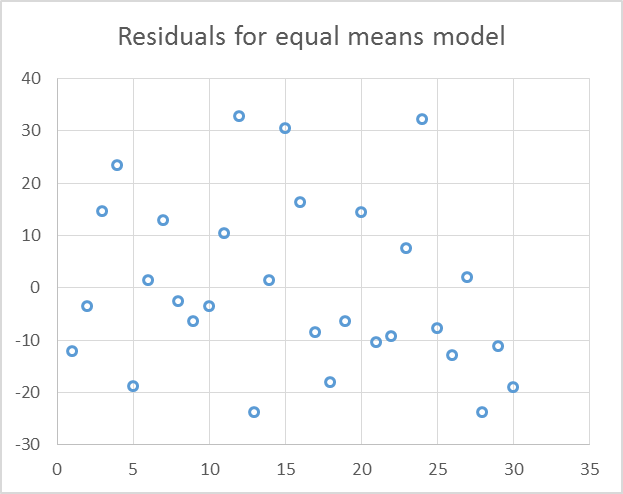
* 1. Assume we want to test the significance of Acetic and H2S. Perform an extra sum of squares F-test to simultaneously test the significance of these variables together.

**Problem:** Perform an extra sum of squares F-test to simultaneously test the significance of these variables together.

**Assumptions:** Based on the scatter matrix in Figure 10, there is not sufficient evidence against a linear relationship between taste and each of the explanatory variables Acetic and H2S. Based on the residual plots, residual histograms and residual QQ plots in Figures 10, we see little evidence against normality except the double peak of the residual histogram, which may be cause for concern. Similarly, the residuals appear to have constant variance, but there is an outlier with a predicted value of over 20 (and a studentized residual over 2) that we will investigate further. We will assume the residuals are independent of each other and proceed with an extra sum of squares test to determine which model is appropriate. For the reduced model (shown in Figure 11) we also find little evidence against the residuals being normally distributed with constant variance.

**Figure 10 Scatter Matrix and Residuals and QQ Plot/Histogram of Residuals for Full Model**



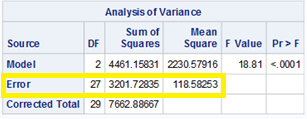
**Figure 11 Scatter Matrix and Residuals and QQ Plot/Histogram of Residuals for Reduced (Equal Means) Model**

**Step 1**: Null and Alternative Hypothesis are below

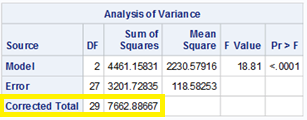
Ho : βAcetic = βH2S = 0;

Ha : βAcetic,βH2S ≠ 0;

**Step 2**: Full Model from SAS:



**Step 3:** Reduced Model from SAS:



**Step 4: F-Test Table (Identical to Tables in steps 2 and 3):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | MS | F | Pr > F |
| Model | 2 | 4461 | 2230 | 18.81 | <0.0001 |
| Error | 27 | 3201 | 118 |  |  |
| Total | 29 | 7662 |  |  |  |

**Step 5: Conclusion**

0.0001 < 0.05 Reject Ho

**Step 6: Conclusion in Context**

There is strong evidence to suggest at the α = .05 level of significance (p-value = <.0001, n = 30) that the variation in the model is better explained by the regression model that includes Acetic and H2S than the equal means model.

1. In one weight-loss study 89 sedentary men were randomly assigned to either a special diet or exercise for a year. Forty-two men were placed on a diet and they lost an average of 7.2 kg with a standard deviation of 3.7 kg. The other 47 men were put on an exercise program and they lost an average of 5.3 kg with a standard deviation of 3.9 kg. For these data, researchers calculated three confidence intervals. They are given in the table below.

|  |  |  |
| --- | --- | --- |
| Confidence Level | Lower Limit | Upper Limit |
| 90% | 0.6 | 3.2 |
| 95% | 0.3 | 3.5 |
| 99% | -0.2 | 4.0 |

* 1. Use the CI results to tell what p-value would be obtained in a two-sample t-test. (Hint: You can't give a precise value, only an interval. You can do the test to check but should report the answer to the question that is asked!)

**THE 90 AND 95% CONFIDENCE INTERVALS REJECTED AND THE 99% FTR. THEREFORE, THE TWO SIDED PVALUE SHOULD BE BETWEEN .05 AND .01.**

* 1. A critic complains that beginning weight is strongly associated with the amount of weight lost and that this experimental design does not control for this important factor. The physician conducting the study says that he doesn't need to worry about the weight of a subject at the beginning of the study. What feature of the experimental design guarantees that the conclusion is still valid even though the study did not account for beginning weight? Explain.

This was a randomized experiment, with each subject randomly assigned to the two treatment groups. We can assume that any confounding variables, such as starting weight, would be randomly distributed between the two groups and thus not an issue.