STAT 350 – Fall 2008
Final Exam

Your Name: ___________________________________ Your Seat: __________

Section Time (circle): 10:30 12:30 1:30

Note:

- You are responsible for upholding the Honor Code of Purdue University. This includes protecting your work from other students.
- Show your work on all questions. Unsupported work will not receive full credit. Credit will not be given for dumb luck. You do not need to show work for multiple choice questions.
- Decimal answers should be exact or to at least four significant digits.
- Unless otherwise stated, assume the significance level for any hypothesis test is 0.05.
- Standard Normal (*Z*) and values/probabilities must be taken from the tables provided. Probabilities, *p*-values, critical values, etc., for χ^2, *t*, and *F* distributions must also be taken from the tables provided, unless this information is available on SAS output provided, in which case the values from SAS should be used.
- You are allowed the following aids: a one-page (8.5×11 inch) cheat sheet, a scientific calculator, and pencils.
- Turn off and put away your cell phone before the exam begins!

Question	Points Possible	Points Missed
1	51	
2	10	
3	23	
4	16	
Total	100	

Final Score: _______________ / 100

1. Jason is a sociology major. For his senior thesis, Jason randomly selected a number of residents from his hometown to survey. He asked each subject a range of demographic questions. Among the questions he asked were: "How many years of schooling have you had?" and "What is your annual income?" Limiting his sample to just those 30 subjects who were no longer in school (that is, who had completed their schooling), the number of years of schooling ranged from 9 to 22 years (mean 15.4 years) and the annual incomes ranged from $28,984 to $61,267 (mean $44790). Using these 30 subjects, he conducted a regression analysis to explore whether the amount of schooling affects income. The SAS output from this analysis is given below.

```
                                           The REG Procedure
                                             Model: MODEL1
                                      Dependent Variable: income

                                Number of Observations Read          30
                                Number of Observations Used          30


                                         Analysis of Variance

                                                Sum of           Mean
        Source                   DF          Squares         Square    F Value    Pr > F

        Model                     1        297006095      297006095       7.05    0.0129
        Error                    28       1179094138       42110505
        Corrected Total          29       1476100233


                     Root MSE           6489.26074    R-Square     0.2012
                     Dependent Mean          44790    Adj R-Sq     0.1727
                     Coeff Var            14.48804


                                          Parameter Estimates

                                      Parameter       Standard
            Variable        DF         Estimate          Error    t Value    Pr > |t|

            Intercept        1            30207     5617.60177       5.38      <.0001
            years_school     1        946.97283      356.57432       2.66      0.0129
```

a. (2 pts) What percent of the variation in incomes is explained by the linear relationship between income and schooling?

b. (2 pts) What is the correlation between income and years of schooling?

c. (5 pts) Based on the above analysis, what is the income you would expect for an individual from this town who has had 17 years of schooling?

1 (continued)

d. (2 pts) The 5th subject in this analysis had 17 years of schooling and has an annual income of $41,019. What is the value of the 5th residual?

e. How much extra money should an individual in this town expect to earn for every additional year of school he or she has completed?

(i) (2 pts) Give a point estimate.

(ii) (5 pts) Give a 95% confidence interval

f. (8 pts) Jason wants to determine if the relationship between years of schooling and annual income is "statistically significant"?

(i) Give the value of the appropriate test statistic

(ii) Give the degrees of freedom for that test statistic

(iii) Give the *p*-value.

(iv) Based on this, is the relationship "statistically significant"? Just answer "yes" or "no".

g. (6 pts) Now assume that Jason wants to test the null hypothesis that years of schooling does not affect annual income (that is, average annual income does not change with an increase in the number of years of schooling) versus the alternative hypothesis that average annual income *increases* as the number of years of schooling increases.

(i) Give the value of the appropriate test statistic

(ii) Give the degrees of freedom for that test statistic

(iii) Give the *p*-value.

1 (continued)

h. According to the Bureau of Labor Statistics, nationwide, average income increased $1750 for each additional year of schooling. Jason wants to compare his town to the national average. He will test the null hypothesis that the trend in his town is the same as the national average against the alternative that the trend in his town is different than the national average.

(i) (4 pts) Give the value of the appropriate test statistic

(ii) (2 pts) Give the degrees of freedom for that test statistic

(iii) (2 pts) Give the *p*-value.

i. (6 pts) Give a 95% confidence interval for the true mean income of all residents of this town (who have completed their schooling).

j. (5 pts) Based solely on the preceding analysis, would it be appropriate for Jason to conclude that additional schooling causes increased income? Justify your answer.

2. Short Answer

a. (5 pts) Think back to the experiment done on mice we used in Lectures 18 and 19. Mice were randomly assigned to receive varying doses of alcohol and then timed running a maze. The correlation between the dose of alcohol and the time to run the maze was 0.91626. Would it be appropriate to conclude that increased doses of alcohol causes an increase in the time to run the maze? Justify your answer.

b. (5 pts) Define "*p*-value".

You shouldn't need more space than this!

3. *Background*: In Lab #3 we looked at a number of different ways to assess whether a sample may have come from a Normal distribution. Probability or QQ-plots are one approach. We also looked at a hypothesis test (the Anderson-Darling Test) provided by Minitab along with the probability plot. Another approach we used was to visually compare the sample histogram with a normal distribution curve with the $\mu = \bar{x}$ and $\sigma = s$ (for example, see Figure 1). Another approach is to *quantitatively* compare the sample histogram to the normal distribution curve. A chi-squared test can then be used to determine whether the observed frequencies in the various bins are close enough to the frequencies that would be expected if the sample did come from a normal distribution.

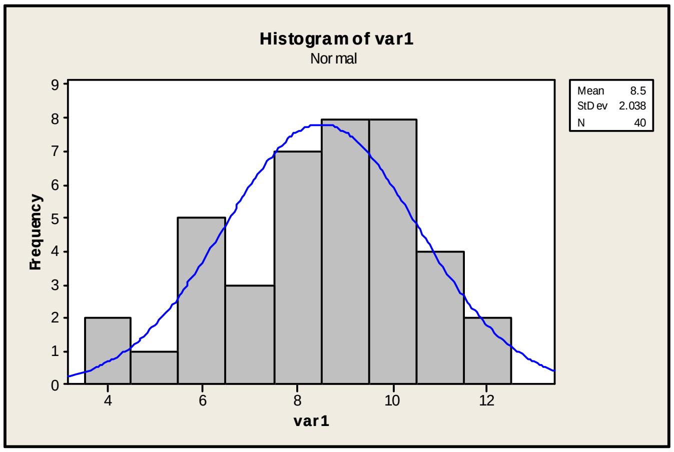


Figure 1. Histogram with Normal Distribution curve from Lab #3, Problem #1. Note, this is *not* the data for the problem below)

A sample of 35 observations was taken. You wish to assess whether these observations may have come from a Normal distribution. The sample had $\bar{x} = 100$ and $s = 10$. Of the 35 observations, 7 were less than 90, 10 were between 90 and 100, 10 were between 100 and 110, and the remaining 8 observations were greater than 110[*] (see table below).

bin	<90	90 to 100	100 to 110	>110
observed frequency	7	10	10	8

a. (8 pts) The null hypothesis is that the sample did come from a normal distribution with $\mu = 100$ and $\sigma = 10$. This null hypothesis can be restated in terms of the true proportions of each category, the π_i's. Give the values of the π_i's (to 4 decimal places)
Hint: $\pi_1 = P(X < 90)$, where $X \sim \text{Normal}(\mu = 100 \text{ and } \sigma = 10)$.
Another hint: $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$.

π_1 = ____________________

π_2 = ____________________

π_3 = ____________________

π_4 = ____________________

[*] Note: Ideally more bins would be used, but then the analysis would take a lot longer (this is me being nice). Also, the data were continuous, so there is no problem with observations falling on the cut-point for a bin (*e.g.*, no observation was exactly 100).

3 (continued)

b. (4 pts) Give the expected counts (to 4 decimal places) for each bin (fill in the table below).

bin	<90	90 to 100	100 to 110	>110
Expected Counts				

c. (4 pts) Give the value of the chi-square statistic for this analysis.

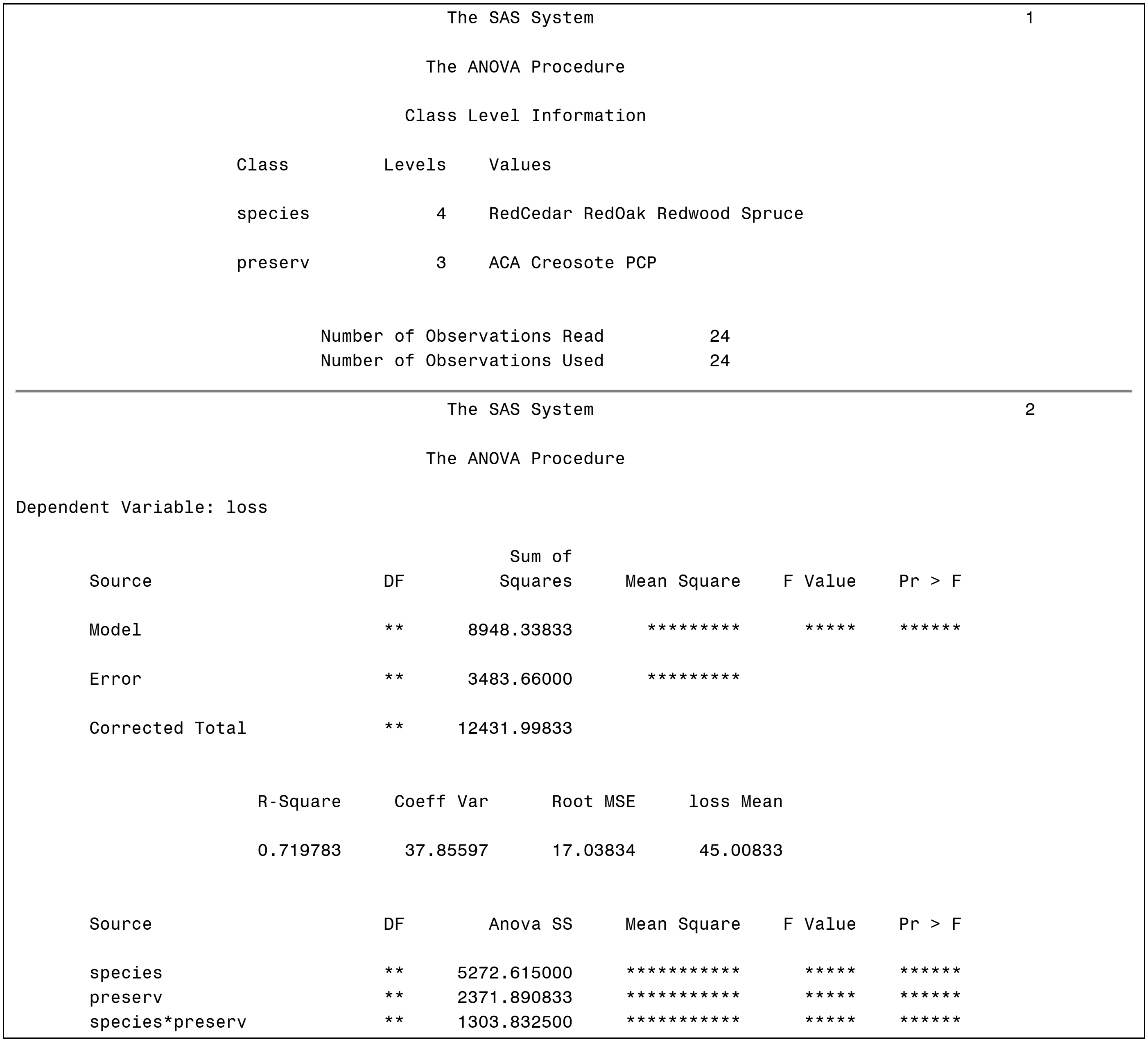
d. (2 pts) What is the critical value for this test ($\alpha = 0.05$)?

e. (5 pts) Based on your analysis above, what is your conclusion? Circle one of the options below.

① The sample did come from a normal distribution

② The sample did NOT come from a normal distribution

4. Susan is majoring in wood science. For her senior thesis, she wanted to examine the durability of wood used for decking. Deterioration of wood in use is commonly caused by decay fungi, certain insects (including termites) as well as other organisms, and weathering. She wanted to examine various species of wood and various preservatives. She used 4 species of wood: Red Oak, Spruce, Eastern Red Cedar, and Redwood. She used 3 types of preservatives: creosote, pentachlorophenol (PCP), and ammoniacal copper arsenate (ACA). She used 2 boards for each species-preservative combination. The boards were dried and weighed then placed outdoors (all in the same experimental plot) for 10 months. At the end of 10 months, the boards were collected, dried, and re-weighed. For each board, the dry weight-loss (in grams) was recorded. The SAS results from the ANOVA are given below. Some information has been omitted and replaced with asterisks (*).



```
                                   The SAS System                                1

                                 The ANOVA Procedure

                               Class Level Information

                    Class         Levels    Values

                    species            4    RedCedar RedOak Redwood Spruce

                    preserv            3    ACA Creosote PCP


                            Number of Observations Read          24
                            Number of Observations Used          24
----------------------------------------------------------------------------------------
                                   The SAS System                                2

                                 The ANOVA Procedure

Dependent Variable: loss

                                             Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F

       Model                       **       8948.33833       *********      *****    ******

       Error                       **       3483.66000       *********

       Corrected Total             **      12431.99833


                      R-Square     Coeff Var      Root MSE     loss Mean

                      0.719783      37.85597      17.03834      45.00833


       Source                      DF         Anova SS     Mean Square    F Value    Pr > F

       species                     **      5272.615000     ***********      *****    ******
       preserv                     **      2371.890833     ***********      *****    ******
       species*preserv             **      1303.832500     ***********      *****    ******
```

4 (continued)

a. Is there a statistically significant difference among the 3 preservatives?

(i) (4 pts) Give the value of the appropriate test statistic.

(ii) (2 pts) Give the degrees of freedom for this test statistic.

(iii)(2 pts) Give the critical value for this test ($\alpha = 0.05$).

(iv)(4 pts) Based on your analysis above, what is your conclusion? Circle one of the options below.

① There is NO statistically significant difference in the mean weight loss among the 3 types of preservatives studied

② The mean weight loss for each preservative is significantly different from each of the other preservatives.

③ At least one of the preservatives has a mean weight loss that is significantly different from the other preservatives.

b. (4 pts) Assume that the *p*-value for "`species`" had been large (*e.g.*, 0.80), but the *p*-value for "`species*preserv`" had been small (*e.g.*, 0.001)[**]. Would it then be reasonable to expect the same performance (loss) for all species of wood for a given preservative? Imagine you are a statistical consultant trying to explain the implication of these results to Susan.

[**] The *p*-values given for part (b) are not actually the *p*-values from the analysis above. But you are to assume they are for the sake of answering part (b)