# Midterm Exam

Friday the 29th of October, 2010

**Name:**_____

**General Comments:**

- This exam is closed book. However, you may use two pages, front and back, of notes and formulas. Write your answers on the exam sheets. If you need more space, continue your answer on the back of the page. A t table is attached at the end (it is the only one you need). Make sure you have all 14 pages!

- The exam is 110 minutes long. There are 4 questions, worth a total of 100 points. They are not equally weighted, nor are they of equal difficulty. The number of points each question is worth is printed with the problem. Read the questions carefully. If you are unsure of the interpretation, come ask.

- **You must show your work** to obtain full credit. If you use a result from class, state what result you are using. If you can't complete a problem for any reason, explain what concepts are at issue, and how you would attack the problem. If you can't work out a number you need for a later part of a problem give it a symbol and show how you would do the calculations with a symbol in place of the missing number. It is a good idea to explain your reasoning **briefly** in English. If I can't tell that you understood what you were doing, I can't give you credit, particularly if you get the wrong numerical answer. GOOD LUCK!

| Question Number | Total Points Possible | Score Received |
|---|---|---|
| 1 | 25 | |
| 2 | 20 | |
| 3 | 30 | |
| 4 | 25 | |
| Total | 100 | |

**THE STORY BEHIND THE EXAM:** Professor Urtha Green, an environmental health scientist at my favorite college, the University of Calculationally Literate Adults, is studying fine particle pollutants at elementary schools in the city of Los Seraphim. She is particularly interested in differences between indoor and outdoor pollutant levels and in how factors like distance from the freeway, temperature, location, weather and time of day affect the air quality. During the exam you will help her analyze some of her data and interpret the results.

## Question 1: Pollution Pile-ups (25 points, 25 minutes)

Dr. Green believes that pollution levels should go down the further you are from a major traffic source such as a freeway but she is not sure how rapidly they drop off. She has therefore conducted a small study in which she has recorded Y, the total particle concentration (in units of thousands of particles per cubic centimeter) at a variety of distances, X, (measured in meters) from the 47 freeway, the major road into Los Seraphim. A STATA printout of the simple linear regression for her data, and some corresponding confidence and prediction intervals for Y are shown below. Use them to answer the questions that follow.

```
. reg pollution distance

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  1,    24) =   21.47
       Model |  23400.0003     1  23400.0003           Prob > F      =  0.0001
    Residual |  26160.1198    24  1090.00499           R-squared     =  0.4722
-------------+------------------------------           Adj R-squared =  0.4502
       Total |  49560.1201    25   1982.4048           Root MSE      =  33.015

------------------------------------------------------------------------------
   pollution |      Coef.   Std. Err.      t     P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
    distance |        -.2   .0431655    -4.63   0.0001   -.2890891   -.1109109
       _cons |        150   12.58478    11.92   0.0000    124.0263    175.9737
------------------------------------------------------------------------------

. adjust distance = 725, se ci
------------------------------------------------------------
      All |         xb        stdp          lb          ub
----------+-------------------------------------------------
          |          5   (21.5016)   [-39.3772    49.3772]
------------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdp       =  Standard Error
           [lb , ub]  =  [95% Confidence Interval]

. adjust distance = 725, stdf ci
------------------------------------------------------------
      All |         xb        stdf          lb          ub
----------+-------------------------------------------------
          |          5   (39.3996)   [-76.3167    86.3167]
------------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdf       =  Standard Error (forecast)
           [lb , ub]  =  [95% Prediction Interval]
```

### Part a

Dr. Green believes there is a negative relationship between pollutant level and distance from the freeway. Perform the hypothesis test she would use to prove her theory. Give the null and alternative hypotheses mathematically and in words with a justification of your choice, obtain the p-value and give your real-world conclusions using $\alpha = .05$.

### Part b

Find the correlation between pollution level and distance from the freeway. Does there appear to be a strong relationship between these variables? Explain briefly.

### Part c

Public Health guidelines suggest that exposure to fine particle pollution levels of more than 50,000 particles per cubic centimeter is unhealthy. According to this model, what is your best estimate of how far you have to get from the freeway before the fine particle concentration drops to this level?

### Part d

The Los Seraphim Unified School District is about to build a large number of new elementary schools. It has decided that it will build them no closer than $X_0 = 725$ meters from the 47 freeway. The district Superintendent wants to be sure that the new schools schools will be under the critical exposure level from part (c). Can you be sure that 95% of schools at this distance will be under the threshold? Explain which interval from the printout you are using to address this question and why. (No calculations are required.)

### Part e

Cautious Connie suggests it would be even safer to build the new schools a distance of 1000 meters from the freeway. Find the estimated particle concentration at this distance. Does your answer make real-world sense? Discuss briefly why what has happened makes sense in the context of the problem.

## Question 2: In and Out Statistics (That's What This Math Test's All About...20 points, 20 minutes)

Next Dr. Green has decided to compare particle concentrations inside the students' classrooms to outside on the playground. She has taken $n_I = 11$ indoor measurements and $n_O = 11$ outdoor measurements and found $\bar{Y}_I = 25$, $s_I = 10$, $\bar{Y}_O = 30$, $s_O = 10$. (Note: As before the pollution levels are in thousands of particles per $cm^3$. You may assume for purposes of this problem that the various measurements are independent).

### Part a

What is the pooled estimate of the standard deviation for the particle level concentration? Show your work or explain your reasoning.

### Part b

Find a 95% confidence interval for the difference between pollution level inside and outside the classrooms. Based on this interval can you be sure there is a difference? If not, why not? If so, which location seems more polluted?

### Part c

Give an estimate for the effect size for the difference between inside and outside measurements. Is this a large effect size? Explain.

## Part d

Before carrying out her study, Dr. Green carried out the power calculation shown below based on the assumption that the true mean difference between inside and outside was $\mu_0 - \mu_I = 10$ and the standard deviation in both locations was $\sigma = 10$.

```
. sampsi 20 30, sd(10) n(11)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
            m1 =       20
            m2 =       30
           sd1 =       10
           sd2 =       10
sample size n1 =       11
            n2 =       11
         n2/n1 =     1.00

Estimated power:

         power =   0.6500
```

**(i)** Do you believe the study was adequately powered? Explain.
**(ii)** Name two things Dr. Green could practically have done to increase her power and briefly discuss their pros and cons.

**Part e (Optional Bonus) Note: This bonus part is long and probably shouldn't be tried until you have finished the rest of the exam!**

A t-test can be thought of as a simple linear regression where the only predictor is a single indicator variable. Let Y be particle concentration and let $X = 1$ if the measurement is taken outside and $X = 0$ if it is taken inside. Our model is $Y = \beta_0 + \beta_1 X + \epsilon$.

**(i)** Give brief real-world interpretations of $\beta_0$ and $\beta_1$ and what you think the best estimates of these quantities, $b_0$ and $b_1$, should be based on the given data.

**(ii)** Fill in the ANOVA table corresponding to this simple linear regression showing your work or explaining your reasoning.

Analysis of Variance

| Source | SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Regression | ------ | --- | ----- | ----- | ----- |
| Error | ------ | --- | ----- | | |
| Total | ------ | --- | | | |

## Question 3: Location, Location, Location (30 points, 35 minutes)

Dr. Green has decided to compare the pollution levels in several different locations and conditions. Specifically, she decides to take measurements in the parking (P) lot at morning drop-off, in the play yard at afternoon recess (R), in the cafeteria (C) at lunch time, and in two classrooms right before lunch, one with the windows closed and air conditioner running (A) and one with the windows open (W). An ANOVA printout for her data is shown below. Use it to answer the following questions. As before assume Y is in thousands of particles per $cm^3$ and measurements are independent.

```
. oneway particle location, tabulate bonferroni

                |         Summary of particle
       location |        Mean   Std. Dev.     Freq.
 ---------------+------------------------------------
 A (class, AC)  |        14.2      7.1         25
 C (cafeteria)  |        19.1      6.9         25
 P (parking lot)|        37.9      6.7         25
 W (window open)|        36.8      5.1         25
 R (play yard)  |        41.5      9.0         25
----------------+------------------------------------
          Total |        29.9     13.1        125

                      Analysis of Variance
    Source              SS      df       MS         F   Prob > F
------------------------------------------------------------------------
Between groups      15225.6      4   3806.4     76.13     0.0000
 Within groups       6000.0    120     50.0
------------------------------------------------------------------------
    Total           21225.6    124    171.2

           Comparison of particle by location   (Bonferroni)
Row Mean-Col Mean|
        p-value  |          A          C          P          W
        ---------+-------------------------------------------
              C |        4.9
                |      0.161
                |
              P |       23.7       18.8
                |      0.000      0.000
                |
              W |       22.6       17.7       -1.1
                |      0.000      0.000      1.000
                |
              R |       27.3       22.4        3.6        4.7
                |      0.000      0.000      0.723      0.192
```

### Part a

Is there overall evidence of a difference in pollution levels at the various locations? Justify your answer with an appropriate test, giving the mathematical hypotheses and your real-world conclusions using $\alpha = .05$.

### Part b

According to the printout, after using the Bonferroni adjustment for multiple comparisons, which pairs of locations do **not** have significantly different particle concentrations? Use this information to describe as precisely as you can the ordering of the locations in terms of air quality for the students.

### Part c

Which pairs of locations would have been significantly different **without adjusting for multiple testing.** Briefly explain your reasoning.

## Part d

Dr. Green is interested once again in comparing inside air quality to outside air quality.

**(i)** Write down the linear combination in which she is interested.
**(ii)** Give your best estimate for the linear combination and its standard error based on this data.
**(iii)** Perform the hypothesis test Dr. Green would use to show that the average air quality inside the buildings is at least 10,000 particles/$cm^3$ better than outside.

## Part e

Suppose that Dr. Green had wanted to include the test of the linear combination from part (d) in her correction for multiple comparisons. Name two approaches she could have used and say very briefly (no more than a sentence) why each is appropriate.

## Question 4: A Fine Model For Fine Particles (25 points, 30 minutes)

Dr. Green realized that many factors besides location might affect the air quality at the schools she was studying. Therefore in the data set collected for Question 3 she also recorded $X_1$, the distance of the measurement from the freeway or other major traffic source (in meters), $X_2$, the temperature at the time the measurement was taken (in degrees Fahrenheit, and an indicator for whether or not it was raining at the time of the measurement: $X_3 = 1$ for rain and $X_3 = 0$ for no rain. She has also used a set of 4 indicators for the different locations: $X_4 = 1$ if the measurement is in the parking lot in the morning, $X_5 = 1$ if the measurement is in the cafeteria at lunch, $X_6 = 1$ if the measurement is in a classroom before lunch with the windows closed and $X_7 = 1$ if the measurement is in a classroom with the windows open before lunch. The play yard at afternoon recess is the reference group with $X_4 = X_5 = X_6 = X_7 = 0$. Dr. Green's multiple regression printout is shown below. Use it to answer the following questions.

```
. regress particle distance temperature rain parking cafeteria closed open

      Source |       SS       df       MS              Number of obs =     125
-------------+------------------------------           F(  7,   117) =   52.43
       Model |  18350.6        7    2621.5             Prob > F      =  0.0000
    Residual |   2875.0      117      25.0             R-squared     =  0.8646
-------------+------------------------------           Adj R-squared =  0.8564
       Total |  21225.6      124     171.2             Root MSE      =   5.000

------------------------------------------------------------------------------
     particle |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     distance |     -0.05    0.0125       4.00   .0001     [-0.075,     -0.025]
  temperature |      0.40    0.1500       2.67   .0086     [ 0.103,      0.697]
         rain |    -15.00    5.0000      -3.00   .0033     [-24.90,     -5.098]
      parking |     10.00    2.0000       5.00   .0000     [ 6.039,     13.961]
    cafeteria |     -5.00    2.0000       2.50   .0138     [-9.951,     -0.049]
window closed |    -12.00    2.0000      -6.00   .0000     [-15.96,     -8.039]
  window open |      0.00    1.5000       0.00  1.0000     [-2.971,      2.971]
        _cons |     60.00    5.0000      12.00   .0000     [50.098,     69.902]
------------------------------------------------------------------------------
```

### Part a

Give units and real-world interpretations for $b_2$ and $b_4$, the coefficients of the temperature and rain variables. Make sure your answers incorporate the numerical values of the coefficients.

### Part b

Use the model to predict the pollution level in the parking lot on a rainy day when it is 60 degrees at a school that is 1000 meters from the nearest major freeway.

### Part c

The Los Seraphim Times reports that the temperature today is going to be 10 degrees hotter than yesterday but otherwise conditions around the city will be the same. Mrs. Jones wants to know how this is going to affect the air quality at her daughter's school. Give a 95% confidence interval for the average difference in air quality she should expect between today and yesterday, briefly explaining your reasoning.

### Part d

Based on this model, does there appear to be a significant difference (using $\alpha = .05$) between pollution levels in the parking lot and in the play yard? Briefly justify your answer.

### Part e

Your answer to part (d) is different to what was found in Question 3. Explain briefly what you believe has happened.

CONGRATULATIONS! YOU ARE DONE. HAVE A GREAT WEEKEND.....