**CPaT Stats Final Take Home –** The statistics problems here will not change (unless otherwise noted – but there will be 2 (maybe 3) other short answer questions (not yet finalized) and 2 problems (one STELLA and one 1kcs – not yet finalized) (by Monday). There is plenty to keep you busy here if you want to work over the weekend! Stay tuned….. Percentages for each question might change slightly so it adds up to 100.

**Hard Copy – due Tuesday, Week 10, in Lab (hand in your Portfolio at the same time).**

**You may hand in ONE portfolio per group – unless you worked alone, or swiched groups. Then you need to hand in a portfolio for you.**

There are xxx questions on this midterm (plus xxx optional). The last two problems are “open-ended”. That means you could spend many hours working on them, but you have limited time, so I’ve limited the number of pages for each question. If you have been keeping up with labs, you should be able to do this in way less than 6 hours.

This Take Home Exam should be an individual effort, but you are free to ask questions in class, lab, and help sessions, or email judyc if you get stuck. If there are clarifying questions that need to be answered I will email the whole class. The statistics tests (the first on the exam) are probably easier (and will be quicker than the Stella and 1kcs problems), so you should warm up on those.

**Question 1 (10%).** T-test.

In a study of cereal leaf beetle damage on oats, researchers measured the number of beetle larvae per stem in small plots of oats after randomly applying one of two treatments: no pesticide or Malathion at the rate of 0.25 pounds per acre. Is there significant evidence at the 1% level that the mean number of larvae per stem is reduced by Malathion?

1. Present the data graphically. Are there outliers or strong skewness that might prevent the use of the t-test?
2. State the null and alternative hypotheses for a statistical test of the claim that chicks fed high-lysine corn gain weight faster.
3. Carry out the test using JMP and report the result.
4. (optional) give a 95% confidence interval for the mean difference in number of larvae per stem.

Your answer to this question should be no more than 1 page!

**Question 2 (10%).** Resampling to Compare 2 Means. Test your hypotheses for the above question using Resampling.

**Question 3 (15%).** 1-way ANOVA.

A study of reading comprehension in children compared three methods of instruction (ref). As is common in such studies, several pretest variables were measured before any instruction was given. One purpose of the pretest was to see if the three groups of children were similar in their comprehension skills. One of the pretest variables was an “intruded sentences” measure, which measures one type of reading comprehension skill. The data for the 22 subjects in each group are given in the data sheet (Q4). The three methods of instruction are ‘basal’, ‘DRTA’, and strategies.

1. Determine whether the data meet the requirements for running a parametric ANOVA; if not, you will need to run a resampling ANOVA.
2. Make a table of means and standard errors
3. Present the above data graphically – as a histogram with error bars. Remember: error bars should be specific to each group.
4. State the null and alternative hypotheses for an ANOVA on these data, and explain in words what the ANOVA will test in this setting.
5. Run the ANOVA (in JMP or resampling) and report your results.
6. If the ANOVA showed that there is a difference among groups, determine which groups are different (and which not – if any), and report your results

Your answer to this question should be no more than 1 page!

**Question 4 (15%).** 1-way ANOVA.

How do nematodes (microscopic worms) affect plant growth? A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes into the pots. A tomato seedling is transplanted into each plot.

1. Determine whether the data meet the requirements for running a parametric ANOVA; if not, you will need to run a resampling ANOVA.
2. Make a table of means and standard errors and present the data graphically.
3. State the null and alternative hypotheses for an ANOVA on these data, and explain in words what the ANOVA will test in this setting.
4. Run the ANOVA (in JMP or resampling) and report your results.
5. If the ANOVA showed that there is a difference among groups, determine which groups are different (and which not – if any).

**Question 5 (10%).** Linear Regression

Airborne particles such as dust and smoke are an important part of air pollution. To measure particulate pollution, a vacuum motor draws air through a filter for 24 hours. Filters were weighted at the beginning and end of the period; weight gained is a measure of the concentration of particles in the air. A study of air pollution made measurements every 6 days with identical instruments in the center of a small city and at a rural location 10 miles southwest of the city. Because the prevailing winds blow from the west, we suspect that the rural readings will generally be lower than the city readings, but that the city readings can be predicted from the rural readings.

Given data on air pollution measurements in two locations, rural and city (note that some measurements are unavailable – these missing values are represented as ‘dots’, so you will have to make those cells blank before pasting into JMP, so JMP correctly types this variable as continuous.):

1. What is the equation of the line?
2. What is your R2 value?
3. What is your F-value?
4. What are your degrees of freedom? Model \_\_\_\_\_\_\_\_\_ Error \_\_\_\_\_\_\_\_
5. What is the p-value associated with this test?
6. (optional) Plot the residuals.
7. (optional) Make a normal quantile plot of the residuals (or a histogram). Is the distribution of the residuals nearly symmetric? Does it appear to be approximately normal?
8. Do you think these data satisfy the assumptions for regression? Say why or why not.
9. Interpret your results as if in a scientific paper (write a couple sentences and cite your *R2=*, *Fdf1,df2=*, and p= in parentheses at the end):
10. Make a scatterplot in JMP, and paste it and the equation just below the graph – titled “Linear Fit”.

**Question 6 (10%).** Chi-Square.

Nutrition and illness are related in a complex way. If the diet is inadequate, the abiliy to resist infection can be impaired andillness results. On the other hand, some illnesses cause lack of appetite, so that poor nutrition can be the result of illness.

In a study of morbidity and nutritional status in 1165 [rescjpp; cjo;drem ;voomg om [ppr cpmdotopms om De;jo, India, data were obtained on nutrition and illness. Nutrition was described by a standard method as normal or as one of 4 levels ofinadequate: I, II, III, and IV.

For purposes of analysis, the two most severly undernourished groups II and IV were combined. One part of the study examined four categories of illness during the past year: upper respiratory infection (URI), diarrhea, URI and diarrhea, and none.

Carry out an analysis of the association between nutritional status and type of illness. Describe the association numerically, assess its significance, and write a brief summary of your findings that refers to your analysis for substantiation.

You need not complete the following tasks in order; for example, you might choose to run a JMP chi-square and then develop a contingency table.

1. What is your scientific hypothesis for this study?
2. What are your null and alternative hypotheses?
3. Create a contingency table (with percentage in each cell and percentages and totals for rows and columns) for the data.
4. Run a chi-square analysis on the data and report your results.
5. What are your degrees of freedom? Say how these were calculated.
6. Run the ANOVA (in JMP or resampling) and report your results.
7. If the ANOVA showed that there is a difference among groups, determine which groups are different (and which not – if any). Represent these differences (and similarities) graphically.

**(Optional) Question 7 (10%).** CART analysis. A data set from ML to be selected.

1. Describe the result of your first split. What was your R-squared value?
2. How many splits did you decide to use?
3. What was the R-squared value for the tree?
4. How well were the xxx described by this model?
5. Paste your ‘final’ tree below:

**Question 8 (25%):** Running and interpreting a STELLA Model. On the fileshare, in /Handouts/Stats/Final you will find a sub-directory (zzz]). It contains xxx.

Populate the given Stella model with data (as provided), run the model and interpret your results. You will want to add graphs or aggregate flow counters to the model to prepare this report.

**Question 9 (25%):** Research Design. On the fileshare, in /Handouts/Stats/Final you will find not only this document, but also of the full 1kcs data set 1kcsStemMapAllforFinal.xlsx.

As you know, this rather large study (almost 6000 trees across 8 sites) was designed to determine whether forest structure varied according to age class.

You have just been hired as the data analyst on a large NSF project, and your first assignment is to prepare a 2-page summary of this data set. The data for this study have already been collected, but (alas!) not much work has been done to decide which statistical tests should be run to answer the major questions of the study: How does forest structure vary with age class?

Some background: It has been shown that certain at-risk species prefer an “old-growth” habitat. Old growth forest structure characteristics are observed in Pacific Northwest Forest sites aged 350 years and older. Forest structure can be defined as the characteristics of tree spacing, species and branch distribution, crown size, etc.

Your job is to define 1-3 statistical tests that you could perform using the existing data that might elicit differences between “old” and “new” forest structure.

In 2 pages, articulate these test(s) clearly. For each test, state the scientific, null and alternative hypotheses, which data you would use, which data are the independent and dependent variables.

(optional) Run one or more of these tests and interpret your results.

My advice is to keep this simple, especially if you are short on time! It’s better to do a good job with fewer tests than a slipshod job on three!

Question 10 TBA

Question 11 TBA