**CPaT Stats final Final TakeHome**

The statistics problems below (Questions 1-6) have (as promised) not changed. A dataset from ML for Question 7 (CART model) has been added (again as promised). There is no STELLA problem (sorry!). Percentages for each question have changed slightly so the total percent adds to 100.

**Hard Copy – due Tuesday, Week 10, in Lab (hand in your Statistics Portfolio at the same time). Do NOT put the exam in your portfolio! See notes on the Week 9 stats lab and on the Week 10 Schedule about the portfolio.**

There are 10 questions on this midterm. 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 post questions and answers on a FAQ - http://blogs.evergreen.edu/cpat/stats-2/stats-final-faq/. The statistics tests (the first on the exam) are probably easier (and will be quicker than the 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? Data for this problem are in the data sheet (Q1-t-test-beetles).

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 (10%).** 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 (Q3-ANOVA-reading). 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 (10%).** 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. The data for this problem are given in the data sheet (Q4-ANOVA-nematodes).

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. The data for this problem are given in the data sheet (Q5-regression-airborne).

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 ability to resist infection can be impaired and illness 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 children from Delhi, India, data were obtained on nutrition and illness. Nutrition was described by a standard method as normal or as one of 4 levels of inadequate: I, II, III, and IV.

For purposes of analysis, the two most severely undernourished groups III 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. The data for this problem are given in the data sheet (Q6-chiSq-nutrition).

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.

**Question 7 (10%).** CART analysis. See Week 7 lab (and key) for directions on how to do a CART model. The “glass” data set from ML has been chosen for this problem. Data for this problem are in the files ([programs/cpt/Handouts/Stats/exam\_final/](https://myfiles.evergreen.edu/academics/programs/cpt/Handouts/Stats/exam_final/) glass.arff and glass.csv).

The data (glass.csv) results from a study of classification of types of glass and was inspired by a forensics problem (see the arff file for details glass.arff). The scientific question you are being asked to answer is: Can glass of type “float” be accurately identified from the given attributes? The statistical question you are asked is whether CART analysis can be used to accurately classify glass based on the attributes provided.

Remember that you will need to add a header row with attribute names to the CSV file before pasting the file into JMP; attributes are given in the arff file.

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 final tree?
4. How well were the glass types described by your model?
5. Paste your ‘final’ tree below:
6. (optional) Perform a similar classification using WEKA. Say what classification method you used. How does the JMP CART model compare with the results you got in WEKA?

**Question 8old deleted:** Running and interpreting a STELLA Model.

**Question 8 (20%):** Research Design. On the fileshare, in /Handouts/Stats/exam\_final you will find not only this document, but also the full 1kcs data set 1kcsStemMapAll.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. <http://acdrupal.evergreen.edu/studycenter/1kcs> holds information about this study.

You are the data analyst on this large NSF project, and your first assignment was to prepare a 2-page summary of this data set (and a good job you did with that!). 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?* Forest structure can be defined as the characteristics of tree spacing, species and branch distribution, crown size, etc.

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. Jerry Franklin’s idea is that perhaps younger forests can be made to exhibit old growth characteristics and preserve species diversity – if we knew what those characteristics are.

Your next task as data wrangler and analyst is to suggest 1-3 statistical tests 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 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 9 (10% - short essay question – 1 page max): During the course of this quarter, you have used three different statistics packages: Excel, JMP, and R. Compare and contrast these three packages – which are best for which kinds of tasks? How would you advise research groups to choose which one(s) to use, and under which circumstances?

Question 10 (10% - optional - essay question – 1 page max – applicable only if you are doing both stats and ML!): During the quarter you have also been exposed to statistics and machine learning. Describe in your own words the overlap between the two fields. What of statistics do you think is relevant to being a good practitioner (and user!) of machine learning?