Statistics for Engineers, 4C3/6C3

Assignment 2

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 19 January 2011**

Assignment objectives

- A review of basic probability, histograms and sample statistics.
- Collect data from multiple sources, consolidate it, and analyze it.
- Deal with issues that are prevalent in real data sets.
- Improve your skills with R (if you are using R for the course).

Notes:

- I would normally expect you to spend between 3 and 5 hours outside of class on assignments. This assignment should take about that long. Answer with bullet points, not in full paragraphs.
- **Numbers in bold** next to the question are the grading points. Read more about the assignment grading system.
- 600-level students must complete all the question; 400-level students may attempt the 600 level question for extra credit. Also 600-level students must read the paper by PJ Rousseeuw, "Tutorial to Robust Statistics".

Question 1 [1]

Recall from class that $\mu = \mathcal{E}(x) = \frac{1}{N}\sum x$ and $\mathcal{V}\{x\} = \mathcal{E}\left\{(x-\mu)^2\right\} = \sigma^2 = \frac{1}{N}\sum (x-\mu)^2$.

1. What is the expected value thrown of a fair 12-sided dice?
2. What is the expected variance of a fair 12-sided dice?
3. Simulate 10,000 throws in R, MATLAB, or Python from this dice and see if your answers match those above. Record the average value from the 10,000 throws.
4. Repeat the simulation for the average value of the dice a total of 10 times. Calculate and report the mean and standard deviation of these 10 simulations and *comment* on the results.

Question 2 [1.5]

In the class last week I mentioned an example of independence. I said that if I take the grade for each question in an exam for a student, calculate the grade per question, then the average of those N grades will be normally distributed, even if the grades in individual question are not. For example: if there are 10 questions, and your grades for each question was 100%, 95%, 26%, 78%, ... 87%, then your average will be as if it came from a normal distribution.

1. This example was faulty: what was wrong with my reasoning?
2. 600-level students: However, when I look at the average grades from any exam, without fail they are always normally distributed. What's going on here?

Question 3 [1]

Write a few *bullet point* notes on the purpose of feedback control, and its effect on variability of process quality.

Question 4 [1.5]

The ammonia concentration in your wastewater treatment plant is measured every 6 hours. The data for one year are available from the dataset website.

1. Use a visualization plot to hypothesize from which distribution the data might come. Which distribution do you think is most likely?
2. Confirm your answer using a suitable plot.
3. Estimate location and spread statistics assuming the data are from a normal distribution.
4. What if I told you that these measured values are not independent. How does it affect your answer?
5. What is the probability of having an ammonia concentration greater than 40 mg/L when:
   - you may use only the data (do not use *any* estimated statistics)
   - you use the estimated statistics for the distribution?

   Note: Answer this entire question using computer software to calculate values from the normal distribution. But also make sure you can answer the last part of the question by hand, if given the mean and variance, and using the table of normal distributions. Print out the table and bring it with you to the midterm and final exam. The computer answer should agree with your hand-calculated value.

Question 5 [1]

One of the question we posed at the start of the course was: given the yields from a batch bioreactor system for the last 3 years (300 data points; we run a new batch about every 3 to 4 days).

1. What sort of distribution do the yield data have?
2. A recorded yield value was was less than 60%, what are the chances of that occurring? Express your answer as: *there's a 1 in X chance* of it occurring.
3. Which assumptions do you have to make for the second part of this question?

Question 6 [1]

Use the section on Historical data from Environment Canada's website and use the `Customized Search` option to obtain data for the `HAMILTON A` station from 1990 to 2000. Use the settings as `Year=1990`, and `Data interval=Monthly` and request the data for 1990, then click `Next year` to go to 1991 and so on.

- For each year from 1990 to 2000, record the total snowfall and the average of the `Mean temp` column over the 12 months (the sums and averages are reported at the bottom of the table).
- Plot these 2 variables against time.
- Perform a test to verify whether the 11 total snowfall values are possibly from a normal distribution.
- Now retrieve the long-term averages for these data from a different section of their website (use the same location, `HAMILTON A`, and check that the data range is 1961 to 1990). Superimpose the long-term average as a horizontal line on your previous plot.

Note: the purpose of this exercise is more for you to become comfortable with web-based data retrieval, which is common in most companies.

END