Statistics for Engineers, 4C3/6C3

Assignment 3

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 27 January 2011**

Assignment objectives

- Interpreting and using confidence intervals.
- Using tests of differences between two population samples.
- Simulating (creating) data and verifying whether theory matches the simulation.

Question 1 [1]

Your manager is asking for the average viscosity of a product that you produce in a batch process. Recorded below are the 12 most recent values, taken from consecutive batches. State any assumptions, and clearly show the calculations which are required to estimate a 95% confidence interval for the mean. Interpret that confidence interval for your manager, who is not sure what a confidence interval is.

Raw data:	[13.7, 14.9, 15.7, 16.1, 14.7, 15.2, 13.9, 13.9, 15.0, 13.0, 16.7, 13.2]
Mean:	14.67
Standard deviation:	1.16

You should the course statistical tables, rather than computer software, to calculate any limits.

Question 2 [1.5]

1. At the 95% confidence level, for a sample size of 7, compare and comment on the upper and lower bounds of the confidence interval that you would calculate if:
   (a) you know the population standard deviation
   (b) you have to estimate it for the sample.

   Assume that the calculated standard deviation from the sample, s matches the population $\sigma = 4.19$.
2. As a follow up, overlay the probability distribution curves for the normal and t-distribution that you would use for a sample of data of size $n = 7$.
3. Repeat part of this question, using larger sample sizes. At which point does the difference between the t- and normal distributions become *practically* indistinguishable?
4. What is the implication of this?

Question 3 [1]

You plan to run a series of 22 experiments to measure the economic advantage, if any, of switching to a corn-based raw material, rather than using your current sugar-based material. You can only run one experiment per day, and there is a high cost to change between raw material dispensing systems. Describe two important precautions you would implement when running these experiments, so you can be certain your results will be accurate.

Question 4 [1.5]

We have emphasized several times in class this week that engineering data often violate the assumption of independence. In this question you will create sequences of autocorrelated data, i.e. data that lack independence. The simplest form of autocorrelation is what is called lag-1 autocorrelation:

$$x_k = \phi x_{k-1} + a_k$$

For this question let $a_k \sim \mathcal{N}\left(\mu = 0, \sigma^2 = 25.0\right)$ and consider these 3 cases:

A: $\phi = +0.7$

B: $\phi = 0.0$

C: $\phi = -0.6$

For each case above perform the following analysis (if you normally submit code with your assignment, then only provide the code for one of the above cases):

1. Simulate the following $i = 1, 2, \ldots, 1000$ times:
   - Create a vector of 100 autocorrelated x_k values using the above formula, using the current level of ϕ
   - Calculate the mean of these 100 values, call it $\overline{x}_i$ and store it in a vector
2. Use this vector of 1000 means and answer:
   - Assuming independence, which is obviously not correct for 2 of the 3 cases, nevertheless, from which population should $\overline{x}$ be from, and what are the 2 parameters of that population?
   - Now, using your 1000 simulated means, estimate those two population parameters.
   - Compare your estimates to the theoretical values.

Comment on the results, and the implication of this regarding tests of significance (i.e. statistical tests to see if a significant change occurred or not).

Question 5 [2]

We emphasized in class that the best method of testing for a significant difference is to use an external reference data set. The data I used for the example in class are available on the dataset website, including the 10 new data points from feedback system B.

1. Use these data and repeat for yourself (in R, MATLAB, or Python) the calculations described in class. Reproduce the dot plot, but particularly, the risk value of 11%, from the above data. Note the last 10 values in the set of 300 values are the same as "group A" used in the course slides. The 10 yields from group B are: $[83.5, 78.9, 82.7, 93.2, 86.3, 74.7, 81.6, 92.4, 83.6, 72.4]$.
2. The risk factor of 11% seemed too high to reliably recommend system B to your manager. The vendor of the new feedback has given you an opportunity to run 5 more tests, and now you have 15 values in group B:

   $$[83.5, 78.9, 82.7, 93.2, 86.3, 74.7, 81.6, 92.4, 83.6, 72.4, \mathbf{79.1}, \mathbf{84.6}, \mathbf{86.9}, \mathbf{78.6}, \mathbf{77.1}]$$

   Recalculate the average difference between 2 groups of 15 samples, redraw the dot plot and calculate the new risk factor. Comment on these values and *make a recommendation to your manager*. Use bullet points to describe the factors you take into account in your recommendation.

Note: You can construct a dot plot by installing the `BHH2` package in R and using its `dotPlot` function. The `BHH2` name comes from Box, Hunter and Hunter, 2nd edition, and you can read about their case study with the dot plot on page 68 to 72 of their book. The case study in class was based on their example.