# Statistics for Engineering, 4C3/6C3

# Assignment 2

**Kevin Dunn, kevin.dunn@mcmaster.ca** **Due date: 30 January 2013**

---

**Assignment objectives: Univariate data analysis**

**Question 1 [5]**

Compute the mean, median, standard deviation and MAD for salt content for the various soy sauces given in this report (page 41) as described in this article from the Globe and Mail on 24 September 2009. Plot a box plot of the data and report the interquartile range (IQR). Comment on the 3 measures of spread you have calculated: standard deviation, MAD, and interquartile range.

**Question 2 [3]**

Give a reason why Statistics Canada reports the median income when reporting income by geographic area. Where would you expect the mean to lie, relative to the median? Use this table to look up the income for Hamilton. How does it compare to Toronto? And all of Canada?

Can you locate the data that shows the income for each riding in Hamilton? How are these data reported? [A "riding" is a political division of a city or region, used for election purposes].

**Question 3 [6]**

Use the data set on "Raw material properties".

- How many variables in the data set?
- How many observations?
- The data are properties of a plastic pellets. Plot each variable, one at a time, and locate any outliers. See how to use the `identify` function in part 9 of the tutorial.
- Compare values of the mean and median for a column contain strong outliers.
- Also compare the standard deviation against the MAD for the same column.

**Question 4 [8]**

A new wastewater treatment plant is being commissioned and part of the commissioning report requires a statement of the confidence interval of the biochemical oxygen demand (BOD). How many samples must you send to the lab to be sure the true BOD is within a range of 2 mg/L, centered about the sample average? If there isn't enough information given here, specify your own numbers and assumptions and work with them to answer the question.

**Question 5 [8]**

One of the questions we posed at the start of this chapter was: "Here are 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 today 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 [8 (required for 600-level students; extra credit for 400-level students)]**

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 7 [0]**

For additional R practice, and data interpretation practice, review question 4 from last year's Assignment 2 (attempt it without looking at the solutions).

**Question 8 [600-level: 0 points]**

*The solution appears as question 27 in* PID

The paper by PJ Rousseeuw, "Tutorial to Robust Statistics", *Journal of Chemometrics*, **5**, 1-20, 1991 discusses the breakdown point of a statistic.

1. Describe what the breakdown point is, and give two examples: one with a low breakdown point, and one with a high breakdown point. Use a vector of numbers to help illustrate your answer.
2. What is an advantage of using robust methods over their "classical" counterparts?

---

END