Statistics for Engineering, 4C3/6C3

Assignment 2

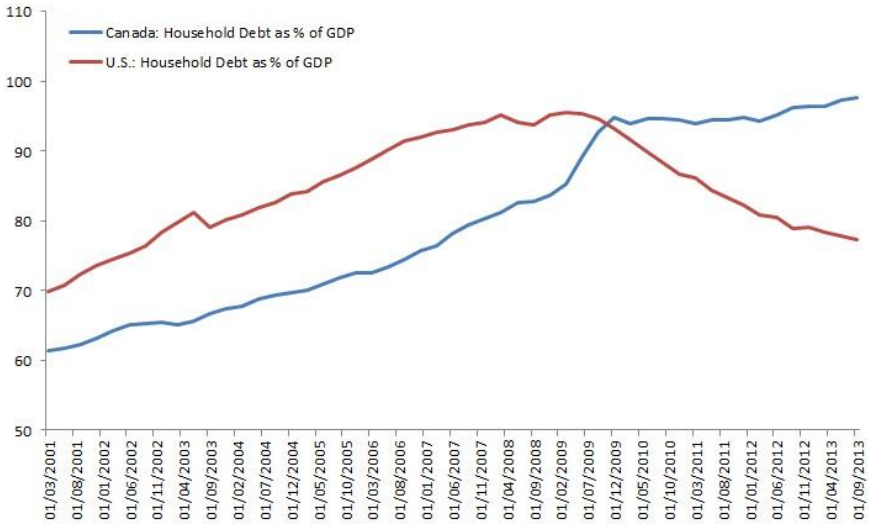
Kevin Dunn, kevin.dunn@mcmaster.ca **Due date: 23 January 2014**

Assignment objectives: interpreting data visualizations; univariate data analysis

Question 1 [5]

Similar to a question on the final exam, 2013:

The following visualization appeared in the Twitter feed of a Globe and Mail columnist, Scott Barlow.



1. The article itself is behind a paywall, so it is not accessible. What might have been the message Mr Barlow was wishing to convey with this plot?
2. Referring the principles of data visualization we learned about, what might he have improved on the graph?

Question 2 [3]

1. Why are robust statistics, such as the median or MAD, important in the analysis of modern data sets? Explain, using an example, if necessary.
2. What is meant by the break-down point of a robust statistic? Give an example to explain your answer.

Question 3 [8]

A food production facility fills bags with potato chips with an advertised bag weight of 50.0 grams.

1. The government's *Weights and Measures Act* requires that at most 1.5% of customers may receive a bag containing less than the advertised weight. At what setting should you put the target fill weight to meet this requirement exactly? The check-weigher on the bagging system shows the long-term standard deviation for weight is about 2.8 grams.
2. Out of 100 customers, how many are lucky enough to get 55.0 grams or more of potato chips in their bags?

Question 4 [20]

The following confidence interval is reported by our company for the amount of sulphur dioxide measured in parts per billion (ppb) that we send into the atmosphere.

$$123.6\,\text{ppb} \leq \mu \leq 240.2\,\text{ppb}$$

Only $n = 21$ raw data points (one data point measured per day) were used to calculate that 90% confidence interval. A z-value would have been calculated as an intermediate step to get the final confidence interval, where $z = \dfrac{\overline{x} - \mu}{s/\sqrt{n}}$.

1. What assumptions were made about those 21 raw data points to compute the above confidence interval?
2. Which lower and upper critical values would have been used for z? That is, which critical values are used before unpacking the final confidence interval as shown above.
3. What is the standard deviation, s, of the raw data?
4. Today's sulphur dioxide reading is 460 ppb and your manager wants to know what's going on; you can quickly calculate the probability of seeing a value of 460 ppb, or greater, to help judge the severity of the pollution. How many days in a 365 calendar-day year are expected to show a sulphur dioxide value of 460 ppb or higher?
5. Explain clearly why a wide confidence interval is not desirable, from an environmental perspective.

Question 5 [10]

Many students in the course expressed an interest in analyzing a large data set. Here's a baby-step: your manager has asked you to describe the flow rate characteristics of the overhead stream leaving the top of the distillation column at your plant. You download one month of data, available on the course website.

The data are from 1 March to 31 March, taken at one minute intervals.

Question 6 [600-level students only: 12]

In the course notes on the section on comparing differences between two groups we used, without proof, the fact that:

$$\mathcal{V}\left\{\bar{x}_B - \bar{x}_A\right\} = \mathcal{V}\left\{\bar{x}_B\right\} + \mathcal{V}\left\{\bar{x}_A\right\}$$

Using the fact that $\mathcal{V}\{cx\} = c^2\mathcal{V}\{x\}$, you can show that:

$$\mathcal{V}\left\{\bar{x}_B + \bar{x}_A\right\} = \mathcal{V}\left\{\bar{x}_B\right\} + \mathcal{V}\left\{\bar{x}_A\right\}$$

1. The first equation is only correct when an important assumption is true; what is that assumption?
2. *Based on an actual industrial problem*: A filling machine doses a drug to a canister. The patient will inhale the drug (imagine an asthma pump). The weight of the drug in the canister must be added as precisely and accurately as possible, to avoid patient over- or under-dosing.

   The weight filled will fluctuate with temperature in the building and is theoretically calculated as having a standard deviation of 32mg due to typical temperature variations. The filling line has 6 machines that fill the canisters and the variability from machine-to-machine is 40mg. The operators calibrate the machines at the start of each shift, and their estimated calibration accuracy is estimated at 15mg. The wear and tear on the machine parts over the year is estimated to only add an extra 10mg of variation.

   What is the expected long-term standard deviation of fill weights recorded from this process? What assumption(s) do you have to make to calculate this?

Question 7 [600-level: 0 points]

The solution appears in Process Improvement using Data.

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