Q1 (Modeled after WMS5 Exercise 1.1a* , p3)

For each of the two scenarios (a) and (b) described below,

- Define the population**, and the summary measure, of interest.
- Describe the inferential objective.
- Describe how you would go about meeting this objective.

(a) The city engineer has a 1000-page printout, listed by street name and address number, of the water consumption for the year 2000 for each of the 60,000 or so dwellings and businesses in the city. Beside each entry is a designation as to whether it is a single-family dwelling unit, a multi-family dwelling unit, a business or another entity. The city has recently switched computer systems, and the computer programmer who generated the listing has taken a higher paying job elsewhere. Because of these and other difficulties, it is not possible to extract the relevant data from the computerized database to answer the city engineer's question [what was average water consumption for single-family dwellings for the year 2000?] before the important city council decision to be taken next week. As a bright summer student, you are asked for your advice and help.

(b) Currently, the city does not use individual water meters, but wishes to estimate how much water single-dwelling units would consume consume in the next fiscal year if they all had meters (the fiscal year is just about to start). The city has purchased 100 water meters. Knowing that you are taking a course in statistics at McGill, the engineer asks you how best to use them to obtain an estimate, and some measure of the "goodness" of this estimate.

** The original question in the text was simply "Discuss the nature of the population of interest, the inferential objective, and how you might go about collecting a sample" in the situation where "a city engineer wants to estimate the average weekly water consumption for single-family dwelling units in the city"*

***Note: some statistical authors define "population" as the objects (items / subjects) in the universe ("complete set") of interest and consider <u>separately</u> the (quantitative or qualitative) features / properties / facts (i.e., "data") that characterize these objects. Other authors (e.g., WMS5, line 14, page 2) <u>combine</u> the objects and their properties and define the "population" as the "body of data" that is the target of interest.*

Q2 (Modeled loosely after WMS5 Exercises 1.2-1.4., p6, and Exercise 1.11, p10. Data are on class web page under 500_92_98)

a) Construct (back-to-back) histograms (using 2-week intervals) to describe and compare the "waiting times" for breast cancer surgery in Québec for the years 1992 and 1998 (samples of n=500 each). How much did they change from '92 to '98? [some ways to do this: use histogram from Data Analysis (If installed as an Add-In under Tools); use DCOUNT database function; .cf e.g. in Excel "homegrown" tips]

b) Calculate the mean and standard deviation for each year (you can use the inbuilt statistical functions AVERAGE and STDEV in Excel, or the formulae for $\bar{y}$ and s on pages 7 and 8). If you use the Excel database criteria (you don't have to), these functions start with the letter D, i.e. DAVERAGE and DSTDEV

c) How well does the "68/95/99.7 rule" at the top of page 9 of WMS do in describing the variation in the 1998 data? *[e.g. sort values, or use* RANK *or* COUNTIF *fn.]*

d) In their report, the authors reported the median (the middlemost of the n ranked values) rather than the mean. Calculate the median for 1998 (In Excel, you can either sort the durations, or use the MEDIAN function directly). Compute the mean for 1998 and compare it with median; explain why the mean and median differ.

e) Suppose you had to guess how long the interval for a randomly selected 1998 patient would be. Four choices of "measures of central tendency" are: (i) the mean; (ii) the mode (the most frequent value); (iii) the median; and (iv) 1/2 the distance between the minimum and the maximum. One criterion by which to rank these measures is to calculate, for each measure, the average absolute difference between the n [=500] observations and the proposed measure, i.e., to see how "close" the data are to the measure (or vice versa if you prefer to look at it the other way around)

- According to this criterion, and before doing any calculations, write down your thoughtful guess as to which of the above four measures of central tendency is "closest" to the 500 values for 1998. Then, carry out the calculations, rank the three measures (i) (iii) and (iv), and comment on how well your guessed-at measure performed. *[this issue is relevant for drag down menus!!]*

  [You might like to see the article by Hanley and Lippman under Notes/Resources for Chapter 1 on the web page for course 697]

f) What proportion of waits were longer than 90 days in 1992? 1998?

f) Refer to the distribution for all 27515 waiting times in the 8 years 1992-1998 (cf separate spreadsheet under all92-98 on the web page)

• What are the 5th, 25th, 50th, 75th and 95th percentiles?

• Can you think of a way to obtain the average waiting time via Excel or otherwise? *[you don't have to do it, just descrtibe what is involved]*

• Examine the frequency distribution for the waits of ≤168 days (24 weeks) in the all92-98 file or in the Fig. under freq_distn on the web page). Can you think of a reason for its "cyclical" pattern? *[We will consider this example theoretically in Chapter 6]*

• How does one go from this distribution to the type of curve in Fig. 1 on the 5th page of the CMAJ article -- and vice versa? *(imagine years were aggregated or segregated)*

• What *would* the histogram look like if we used log[wait], rather than wait itself?

3 (Empirical Studies of Variability; most taken from MRT2)

a) Take an ordinary deck of 52 playing cards containing 4 aces and shuffle thoroughly. Count from the top the number of cards down to and including the first ace, and record the count. Repeat the process.

• What is the average count? (without reading further, write down your thoughtful guess.)

• What is the probability that the first ace is on card 1? 2? ... 52?

• Within what number of cards will we find the first ace half of the time?

b) What is the average length of words, measured in letters, used in sports (or, if you prefer, music) reporting?

•Write down a thoughtful guess.

• (Independently of others) choose a report/article on sports or music in a newspaper or magazine or from the internet. Obtain a frequency distribution of the lengths of the first 50 words and compute the mean length. Hand in the passage of text (with source indicated) along with you work.

c) From a telephone book, find the frequency distribution of the last digits for 100 phone numbers. (Write down your guess for the frequency distribution.)

d) Find the frequency distribution of first (leftmost) digits in counts of votes from a given candidate or party by some unit of population, such as a county or constituency (or in physical measurements such as areas of states, heights of mountains, populations of countries, or the first significant digits in physical constants). In the number 345, the first significant digit is 3, as it is also in 0.00345.

• From an almanac, or other source, obtain thefrequency distribution of leftmost digits of areas (or populations) of states

• From a chemical or physical handbook, obtain the frequency distribution of the first significant digits of 50 physical constants

e) Open a novel to a page near the middle, and choose the first 10 full lines of text. Records the number of *e*'s in each line, and get the average number of *e*'s per line.Use the letter count from one line as a base, and estimate the percentage of letters that are *e*'s. One way to get your word processor to do this on a piece of electronic text is to do a 'global replace' of the letter *e* by the letter *e* and have it tell you how many replacements it made.

f) Record the numebr of rolls of a die before a 6 appears. Repeat the experiment 10 times, and obtain the average number of rolls required.

g) Record the total number of rolls of a die before every face has appeared. Repeat the experiment 5 times and obtain the average number of rolls.

h) *Server problem*: In a unit of time, there is a 50:50 chance that a customer appears at a counter to be served. If others are ahead of him at the counter, he lines up in the queue;

otherwise the server serves him and takes 2 units of time to complete the service. In the 10th unit, what is the average number in the queue if the the process starts with no customers at the counter? Use a coin or table of random numbers (or Excel) to carry out the experiment 5 times and get the average nummber. Also get the average number served at the close of the 10th unit of time. Example (a_i stands for a customer who arrived in the ith time interval):

Time Unit	1	2	3	4	5	6	7	8	9	10
Arrivals	a_1	a_2	–	–	–	a_6	a_7	a_8	a_9	a_{10}
Being served	a_1	a_1	a_2	a_2	–	a_6	a_6	a_7	a_7	a_8
In line	a_1	a_1, a_2	a_2	a_2	–	a_6	a_6, a_7	a_7, a_8	a_7, a_8 , a_9	a_8, a_9 , a_{10}

Total served: 4 Number in queue in 10th unit: 3

i) Toss a coin 50 times and carefully record the sequence of H's and T's.

j) *Simplified epidemic*: An infectious disease has a one-day infectious period, and after that day the patient is immune. Six hermits (numbered 1,2, 3, 4, 5, 6) live on an island, and if one has the disease he randomly visits another hermit for help during his infectious period. If the visited hermit has not had the disease, he catches it and is infectious the next day. Assume hermit 1 has the disease today, and the rest have not had it. Throw a die (or use a table of random numbers, or use Excel) to choose which hermit he visits (ignore face 1). That hermit is infectious tomorrow. Then throw again to see which hermit he visits, and so on. Continue the process until a sick hermit visits an immune one and the disease dies out. Repeat the experiment 5 times and find the average number who get the disease.