PART 1

Last Name:					S	O	L	U	T	I	O	N	S						
First Name:																			
Student #:																			

Instructors: M. Pivovarova/M.Tanaka

Duration: 80 minutes.

Allowed aids: A non-programmable calculator and aid sheets provided.

Format: This test consists of two parts and a SCANTRON form. For both parts combined there are a total of 80 possible points.

BEFORE we announce the end of the test, enter your name and student # on BOTH graded pieces:

(1) The pink SCANTRON form
(2) Part 1

Part 2: 14 multiple choice questions worth 2.5 points each for a total of 35 points

Part 1: 4 written questions worth a total of 45 points

- For each question we give a guide for your response in brackets. It indicates what is expected: a quantitative analysis, a graph, and/or a written response. For example, "Is sampling error a plausible explanation for the result? [Analysis & 2 – 3 sentences]"
- Make sure to focus on answering the specific questions asked. Extraneous analysis does not earn positive marks even if it is correct and may earn negative marks if incorrect.
- Show your work and answer clearly, concisely, and completely. You do not have to fill all of the blank space: a generous amount is provided for your convenience.

	Q1	Q2	Q3	Q4	Part 1	Part 2	Total	% Mark
Point Value	*14*	*10*	*9*	*12*	*45*	*35*	*80*	
Points Earned								

(1) [14 points] Below are the data on the daily temperatures collected by two research assistants for 6 days at 7 am, but in different units – in Farenheit and in Celsius.

Day	Celsius (C)	Farenheit (F)
1	0	32.5
2	-1	30
3	-2	28.5
4	1	33
5	2	35
6	-3	27

(a) [8 points] Find equation to convert Celsius into Farenheit based on the data provided (Note: it might not be the conventional equation used to convert Celsius in Farenheit F=9/5*C+32) [Your work and 1 equation]

Let Celsius be X and Farenheit be Y. First, compute the mean and standard deviations:

$$\overline{X} = \frac{0-1-2+1+2-3}{6} = -0.5 \qquad \overline{Y} = \frac{32.5+30+28.5+33+35+27}{6} = 31$$

$$s_X^2 = \frac{1}{5}\left[(0+0.5)^2 + (-1+0.5)^2 + (-2+0.5)^2 + (1+0.5)^2 + (2+0.5)^2 + (-3+0.5)^2\right] = 3.5$$

, $s_X = 1.87$

$$s_Y^2 = \frac{1}{5}\left[(32.5-31)^2 + (30-31)^2 + (28.5-31)^2 + (33-31)^2 + (35-31)^2 + (27-31)^2\right] = 9.1$$

$$s_Y = 3$$

$$COV(X,Y) = \frac{1}{5}\left[(0+0.5)(32.5-31) + (-1+0.5)(30-31) + (-2+0.5)(28.5-31) + \right.$$

$$\left.(1+0.5)(33-31) + (2+0.5)(35-31) + (-3+0.5)(27-31)\ \right] = 5.6$$

[1 point]

$$b = \frac{5.6}{3.5} = 1.6, \qquad a = \overline{Y} - b\overline{X} = 31 - 1.6*(-0.5) = 31.8$$

$$\boxed{\hat{y} = 31.8 + 1.6 * x}$$

(b) [3 points] What might explain the discrepancy between the standard equation to convert Farenheit into Celsius and the one you have obtained in part (a)? [1-2 sentences]

Measurement error or recording error (non-sampling error) might explain why the equation in part (a) is different from conventional equation.

(c) [3 points] What is the coefficient of determination? Interpret it. [1 number and 1-2 sentences]

Coefficient of determination is

$$r^2 = [corr(X,Y)]^2 = \left[\frac{COV(X,Y)}{s_X s_Y}\right]^2 = \left[\frac{5.6}{1.87 * 3.02}\right]^2 = 0.9833$$

Coefficient of determination tells us how much variation in dependent variable (Y) is explained by independent variable (X). Here, coefficient of determination is close to 1 indicates that all variation in Y is explained by X.

(2) [10 points] Below is the Stata summary of a random sample. Part of the output has been intentionally erased:

```
                          X
-------------------------------------------------------------
      Percentiles      Smallest
 1%           15              8
 5%           18             14
10%           25             15       Obs                 231
25%           30             16

50%           38                      Mean           37.64069
                        Largest       Std. Dev.      10.40593
75%           44             60
90%           52             61
95%           56             62
99%           61             63
```

(a) [3 points] Based on the summary above, what is the most reasonable statement you can make about distribution of the population? [1-2 sentences]

The most reasonable statement we can make is that the population is symmetric. You might also infer that it is bell-shaped (you may check whether Empirical rule holds for this sample).

(b) [7 points] Fill in the blanks with the appropriate statistics from the sample [7 numbers]:

$\overline{X}$	37.64
Median	38
Interquartile range	44-30=14
Range	63-8=55
s^2	108.16
99^{th} percentile	61
Coefficient of variation	0.276

(3) [9 points] A marketing manager of a ready-to-eat breakfast cereal company is planning a survey in order to learn about the relationship between the types of cereal and household composition in Canada.

(a) [3 points] One of the marketing team members proposed to collect the data by including a pre-paid postcard of the survey in each cereal boxes they produce. What can be a potential problem in this sampling design? [1-2 sentences]

There are two main problems with such design – one is the non-response bias and another one is selection bias. Non-response bias arises because the number of post-cards distributed (targeted sample) would be smaller than number of received responses. Since individuals are choosing whether respond or not respond, we can say that they "self-select" themselves into the sample and this will lead to a selection bias.

(b) [3 points] Another team member proposed to collect the data by running TV advertisements announcing the survey and asking viewers to access the company website and participate in the survey. What can be a potential problem in this sampling design? [1-2 sentences]

Selection bias is likely to be a problem with this survey design for the same reasoning as in part (a)

(c) [3 points] The survey received responses from 370 households. The mean number of household members is 3.3 and its standard deviation is 1.2. After the calculation of these statistics was complete, an analyst realized that there were 3 unrecorded observations in the dataset. These households had 1, 2, and 4 members, respectively. If she included them in the calculation, how would the mean change? [1-2 sentences]

Thtee possible solutions:

Three missing obs. are not included in 370

$$\overline{X} = \frac{\sum x_i}{n} \Rightarrow \sum x_i = \overline{X} \cdot n = 3.3 \cdot 370 = 1221$$

1221+1+2+4=1228

$$\overline{X} = \frac{\sum x_i}{n} = \frac{1228}{373} = 3.29$$

Three missing obs. are included in 370

$$\overline{X} = \frac{\sum x_i}{n} \Rightarrow \sum x_i = \overline{X} \cdot n = 3.3 \cdot 367 = 1211$$

1211+1+2+4=1218

$$\overline{X} = \frac{\sum x_i}{n} = \frac{1218}{370} = 3.29$$

The mean will go down by a little – from 3.3 to 3.29.

3) The mean will increase if an additional observation is greater than the previous mean while it will decrease if an additional observation is smaller than the previous mean. Since the mean of unrecorded observation is 2.33, the mean including these observations are smaller than before.

(4) [12 points] A small independent physicians' practice has three doctors. Dr. Doyle sees 41 percent of the patients, Dr. Holmes sees 32 percent of the patients, and Dr. Watson sees the rest. Dr. Doyle requests a blood test on 5 percent of his patients, Dr. Holmes requests it on 8 percent of his patients, and Dr. Watson requests it on 6 percent of his patient.

(a) [4 points] An auditor randomly selects a patient from the past week. What is the probability that the patient was given a blood test? What kind of probability is that? [Your work , 1 word and 1 number]

Construct a joint probability table:

	Doyle	**Holmes**	**Watson**	**Margnal**
Blood Test	0.05*0.41=0.0205	0.08*0.32=0.0256	0.06*0.27=0.0162	0.0623
No Blood Test	0.3895	0.2944	0.2538	0.9377
Marginal	0.41	0.32	0.27	1

In this question, we are being asked about the marginal probability that the patient is given a blood test.

We are given information on marginal probabilities:

P(Doyle)=.41, P(Holms)=.32, P(Watson)=.27,

and conditional probabilities (these are conditional probabilities because we know th fraction of each doctor's patients who were given the blood test, and we can compute joint probabilities as shown in the table):

P(Test|Doyle)=.05, P(Test|Holms)=.08, P(Test|Watson)=.06

We can find joint probabilities:

P(Doyle$\cap$ Test)=0.41*0.05=0.0205, P(Holms$\cap$ Test)=0.32*0.08=0.0256,
P(Watson$\cap$ Test)=0.27*0.06=0.0162

Therefore, the chance that a patient was given a blood test (or proportion of patient who were given a blood test) is:

P(Test)= P(Doyle$\cap$Test)+P(Holms$\cap$Test)+P(Watson$\cap$Test)= 0.0623.

This is marginal probability.

(b) [4 points] An auditor randomly selects a patient from the past week and found that he had done a blood test as a result of doctor's visit. Knowing this information, what is the probability that he is not Dr. Holmes' patient? What kind of probability is that? [Your work , 1 word and 1 number]

This question asks to compute conditional probability:

Prob(not Holmes| Test)=

$$\frac{P(notHolmes \cap Test)}{P(Test)} = \frac{P(Doyle \cap Test) + P(Watson \cap Test)}{P(Test)} = \frac{0.0205 + 0.0162}{0.0623} = 0.589$$

Or

$$\text{P(Holmes|Test)} = \frac{P(Holmes \cap Test)}{P(Test)} = \frac{0.0256}{0.0623} = 0.4109$$

P(not Holmes|Test)=1-P(Holmes|Test)=1-0.4109=0.5891

(c) [4 points] An auditor randomly selects a patient from the past week. What is the probability that the person is a patient of Dr. Watson or was not given a blood test? What type of probability is that? [Your work, 1 word and 1 number]

This question asks about the probability of the union of events or union probability:

P(No Test)=1-P(Test)=0.9377, P(Watson∩No Test)=0.27*(1-0.06)=0.2538

P(Watson or No Test)=Prob(Watson)+P(No Test)-Prob(Watson∩No Test)

=.27+(1- .0623)-.2538=0.9539

Extra Space: If you need to use this space, it is *your responsibility* to clearly indicate for which question(s) and which part(s) AND to clearly indicate at the end of the space specifically provided for that question and part that you have also used this extra space.

Extra Space: If you need to use this space, it is *your responsibility* to clearly indicate for which question(s) and which part(s) AND to clearly indicate at the end of the space specifically provided for that question and part that you have also used this extra space.