Please place your name on Each page now **NAME: SOLUTIONS**

Statistics I

Quiz 2 – SOLUTIONS

Problem 1: (35 percent)

Suppose you are in the business of producing milk. At the beginning of each hour you select bottles of milk to be tested for purity. Two different testing procedures (A and B) are available. Both procedures yield an unbiased estimate of the number of parts per million (PPM) of a particular contaminant (e.g., lead) that might be in the milk. However, the two procedures differ in their inherent variability and in the amount of time that it takes to do the test.

In particular, assume that the variance of the number of parts per million of the contaminant using method A is 100 (PPM^2) while the variance for method B is 300 (PPM^2). Method A takes 30 minutes per sample and method B takes only 15 minutes per sample.

Thus, every hour we can test 6 bottles of milk: 2 using method A and 4 using method B. Let the observations be $X_1^A, X_2^A, X_1^B, X_2^B, X_3^B$, and X_4^B where, for example, X_3^B denotes the third observation made using method B. Assume all observations are independent of each other.

Consider the following estimator of the actual number of parts per million in the milk in any hour:

$$\hat{X} = w_1 X_1^A + w_1 X_2^A + w_2 X_1^B + w_2 X_2^B + w_2 X_3^B + w_2 X_4^B$$

In other words, the two observations taken using method A are each weighted by w_1 while each of the 4 taken using method B are weighted by w_2. Clearly for $\hat{X}$ to be an unbiased estimator of the population level of contaminants, we require $2w_1 + 4w_2 = 1$.

Please place your name on Each page now **NAME:** **SOLUTIONS**

a) Find values of w_1 and w_2 so that the variance of $\hat{X}$ is minimized.

Hint: recall that the variance of $\hat{X}$ will be given by

$$Var(\hat{X}) = w_1^2 Var(X_1^A) + w_1^2 Var(X_2^A) + w_2^2 Var(X_1^B) + w_2^2 Var(X_2^B) + w_2^2 Var(X_3^B) + w_2^2 Var(X_4^B)$$

Minimize $Var(\hat{X}) = 2w_1^2(100) + 4w_2^2(300)$

Subject to $2w_1 + 4w_2 = 1$

OR substituting $w_1 = 0.5 - 2w_2$ into the objective function

Minimize $Var(\hat{X}) = 2(0.5-2w_2)^2(100) + 4w_2^2(300)$

SO

$$\frac{dVar(\hat{X})}{dw_2} = -8(0.5 - 2w_2)(100) + 2400w_2 = 0$$

SO

$$w_2 = \frac{400}{4000} = 0.1 \quad \text{AND} \quad w_1 = 0.5 - 2w_2 = 0.3$$

b) Using the values of w_1 and w_2 that you found in part (a), what is the variance of $\hat{X}$?

$$\begin{aligned} Var(\hat{X}) &= 2w_1^2(100) + 4w_2^2(300) \\ &= 2(0.09)(100) + 4(0.01)(300) \\ &= 30 \end{aligned}$$

c) What would be the variance of the unweighted average, $\bar{X} = \frac{1}{6}(X_1^A + X_2^A + X_1^B + X_2^B + X_3^B + X_4^B)$?

$$\begin{aligned} Var(\bar{X}) &= \frac{1}{36}\{Var(X_1^A) + Var(X_2^A) + Var(X_1^B) + Var(X_2^B) + Var(X_3^B) + Var(X_4^B)\} \\ &= \frac{1}{36}\{100 + 100 + 300 + 300 + 300 + 300\} \\ &= \frac{1}{36}\{1400\} \\ &= 38.8888 \end{aligned}$$

d) What is the percentage reduction in the variance of the estimator that results from using $\hat{X}$ instead of the unweighted average $\overline{X}$? (i.e. what is the value of the answer to (c) minus (a) all divided by (a) as a percentage?)

$$\frac{38.888-30}{30}\times 100\% = 29.63\%$$

Problem 2: (35 Percent)

Suppose you are interested in determining the mean household income in a city. To do so, you decide to sample 100 households randomly. Letting X_j be the income of the j^{th} household **measured in \$1000**, you obtain the following information.

$$\sum_{j=1}^{100} X_j = 4850 \qquad \sum_{j=1}^{100} X_j^2 = 241{,}561$$

a) Find the (1) sample average, the (2) sample variance and (3) sample standard deviation of household incomes measured in \$1000.

$$\overline{X} = \frac{\sum_{j=1}^{100} X_j}{100} = \frac{4850}{100} = 48.5$$

$$s^2 = \frac{\sum_{j=1}^{100} X_j^2 - \left(\sum_{j=1}^{100} X_j\right)^2 \Big/ 100}{99} = \frac{241561 - 4850^2/100}{99} = \frac{6336}{99} = 64$$

$$s = \sqrt{s^2} = \sqrt{64} = 8$$

b) Assuming that the household incomes are from a Normal distribution, find a two-sided 95% confidence interval for the true population mean household income. (Again, report your answer in terms of \$1000).

Since the data are Normally distributed and n is large we can use :

$$\overline{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le m \le \overline{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

giving

$$48.5 - 1.96\frac{8}{10} \le m \le 48.5 + 1.96\frac{8}{10}$$

OR

$$46.93 \le m \le 50.07$$

c) Consider the following hypothesis:

$$H_0: \quad m \geq 50$$
$$H_1: \quad m < 50$$

For what values of the sample average would you reject the null hypothesis at $\alpha = 0.05$? Do you reject the null hypothesis at this level of significance?

You will reject H_0 if $\overline{X} < 50 - z_\alpha \frac{s}{\sqrt{n}} = 50 - z_{0.05}\frac{8}{10} = 50 - 1.645 \times 0.8 = 48.684$

You do reject H_0 at the 0.05 significan ce level

d) Suppose the true mean is \$47.65 thousand dollars. What is the (approximate) probability of a type II error for the test outlined in part (c) above?

$$\begin{aligned} P(\textit{Type II error}) &= P(\text{do not reject } H_0 \,|\, H_1 \text{ is true}) \\ &= P(\overline{X} \geq 48.684 \,|\, m = 47.65) \\ &= P\left(\frac{\overline{X} - 47.65}{s/\sqrt{n}} \geq \frac{48.684 - 47.65}{s/\sqrt{n}}\right) \\ &= P\left(z \geq \frac{1.034}{8/10}\right) \\ &= P(z \geq 1.293) \\ &\approx 0.1 \text{ since the critical value for } 0.1 \text{ is } 1.282 \\ &\quad \text{and the actual value is about } 0.098 \end{aligned}$$

e) Find a 90% confidence interval for the population variance. You might find the following table useful.

	a									
d.f.	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
95	63.250	65.898	69.925	73.520	77.818	113.038	118.752	123.858	129.973	134.247
96	64.063	66.730	70.783	74.401	78.725	114.131	119.871	125.000	131.141	135.433
97	64.878	67.562	71.642	75.282	79.633	115.223	120.990	126.141	132.309	136.619
98	65.693	68.396	72.501	76.164	80.541	116.315	122.108	127.282	133.476	137.803
99	66.510	69.230	73.361	77.046	81.449	117.407	123.225	128.422	134.641	138.987
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

$$\frac{(n-1)s^2}{c^2_{n-1,\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{c^2_{n-1,1-\frac{\alpha}{2}}}$$

$$\frac{99\times 64}{123.225} \le \sigma^2 \le \frac{99\times 64}{77.046}$$

$$51.42 \le \sigma^2 \le 82.24$$

Problem 3: (30 Percent)

You are interested in determining whether a new spray-on coating can help skaters skate faster in speed skating. To determine the answer, you select 10 world-class athletes. Each athlete will skate the same course on two successive days. On one of those days (chosen at random for each athlete), the blades will be spayed with the new material. On the other day, the blades will be spayed with a simple air spray (like a placebo) so that the athlete does not know on which day she is skating with the new coating. The data are shown in the table below.

Skater	Without coating	With coating	Difference
Alice	145.31	144.10	1.21
Barbara	147.34	145.37	1.97
Carla	146.24	145.37	0.87
Diane	141.66	141.42	0.24
Faith	141.91	139.10	2.81
Gail	139.01	139.84	-0.83
Heather	144.29	145.68	-1.39
Joan	144.42	141.62	2.80
Keren	148.77	144.12	4.65
Tamar	143.64	139.40	4.24
TOTAL	**1,442.59**	**1,426.02**	**16.57**
SUM OF SQUARED VALUES	**208,181.70**	**203,415.20**	**64.15**
Average	**144.26**	**142.60**	**1.66**
Standard deviation	**2.889**	**2.623**	**2.019**
Variance	**8.345**	**6.878**	**4.077**

Throughout this problem you can assume that the times are Normally distributed.

a) Treat the observations of times as if they were independent samples. Test the following hypothesis using a significance level of α=0.05. Assume that the underlying variances of the times are equal.

Be sure to state clearly whether or not you reject the null hypothesis

$$H_0: \quad m_{without\ coating} \le m_{with\ coating}$$

$$H_1: \quad m_{without\ coating} > m_{with\ coating}$$

The test statistic is

$$T = \frac{\overline{X}_{without} - \overline{X}_{with}}{s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where $s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{9\times 8.345 + 9\times 6.878}{18} = 7.611$

so $T = \frac{144.26-142.60}{2.759\sqrt{\frac{1}{10}+\frac{1}{10}}} = 1.345$

which has a t distribution with 18 d.f.

The critical t value with 18 d.f. and a probability of 0.05 in the tail is 1.734

So we DO NOT reject H_0 at the 0.05 level of significance using this test

b) Now perform the same test, but do **not** assume that the underlying variances are the same.

Be sure to state clearly whether or not you reject the null hypothesis

we first calculate the new d.f. as follows

$$w_1 = \frac{8.345}{10} = 0.8345$$

$$w_2 = \frac{6.878}{10} = 0.6878$$

$$\boldsymbol{n} = \frac{(w_1+w_2)^2}{w_1^2/(n_1-1) + w_2^2/(n_2-1)} = \frac{(0.8345+0.6878)^2}{0.8345^2/9 + 0.6878^2/9} = \frac{2.317}{0.130} = 17.83$$

which we round down to 17

The new T statistic is $T = \dfrac{\overline{X}_{without} - \overline{X}_{with}}{\sqrt{\dfrac{s^2_{without}}{n_1} + \dfrac{s^2_{with}}{n_2}}} = \dfrac{144.26 - 142.60}{\sqrt{\dfrac{8.345}{10} + \dfrac{6.878}{10}}} = 1.345$

The critical value for 17 d.f. is 1.740

So we still DO NOT Reject H_0

c) Now perform the same hypothesis test treating the data as if they came from a matched pairs design (as they do since each woman skates both with and without the coating).

Be sure to state clearly whether or not you reject the null hypothesis

Now the T statistic is

$$T = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{1.66}{2.019/\sqrt{10}} = 2.60$$

Since the critical value of the t-distribution with 9 d.f. and 0.05 prob in the tail is 1.833 we DO reject H_0 using this test

d) Perform the following test using α=0.10.

$$H_0: \quad \sigma^2_{without\ coating} = \sigma^2_{with\ coating}$$

$$H_0: \quad \sigma^2_{without\ coating} \neq \sigma^2_{with\ coating}$$

Be sure to state clearly whether or not you reject the null hypothesis

$$F = \frac{s_1^2}{s_2^2} = \frac{8.345}{6.878} = 1.213$$

Since the critical value of $f_{9,9,0.05} = 3.18$

We do NOT reject H_0 in this case