Economics 224, Fall 2008 – Assignment 3

2. **Wages and salaries of Saskatchewan females and males**

(a) Let μ be the true mean wages and salaries. For this part of the question, it is not necessary to know the value of μ. The sample size is $n = 500$. Given this large sample size, by the Central Limit Theormen, $\bar{x}$ has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n} = 25.2/\sqrt{500} = 25.2/22.361 = 1.127$. The probability that $\bar{x}$ is within §1 thousand of the mean μ, is the area under the normal curve between $\mu - 1$ to $\mu + 1$. The Z value for $\bar{x} = \mu - 1$ is

$$\begin{aligned} Z &= \frac{(\mu - 1) - \mu}{\sigma/\sqrt{n}} \\ &= \frac{-1}{1.127} = -0.89 \end{aligned}$$

At $Z = -0.89$, the cumulative normal probability is 0.1867.

By symmetry, the Z value at $\bar{x} = \mu + 1$ is 0.89 and the cumulative probability is 0.8133.

The required probability is the difference between these two: $0.8133 - 0.1867 = 0.6266$.

The probability that $\bar{x}$ is within \$5 thousand of the population mean is the area under the normal curve between $\mu - 5$ to $\mu + 5$ and the Z value for the first is

$$\begin{aligned} Z &= \frac{(\mu - 5) - \mu}{\sigma/\sqrt{n}} \\ &= \frac{-5}{1.127} = -4.44 \end{aligned}$$

and the required probability is the area under the normal distribution between $Z = -4.44$ and $Z = 4.44$. Since this is such a large Z value, so large that it is not given in the table of the normal distribution, the probability is just under 1.000. The largest Z value in the table is -3.09 and there is less than 0.0010 of the area

to the left of this. by symmetry, there is also less than 0.0010 of the area to the right of $Z = 3.09$. As a result, the area between these two Z values is $1.000 - 0.002 = 0.998$. So the area between $Z = -4.44$ and $Z = 4.44$ must be even greater than this, perhaps 0.9999. It is almost certain that any sample mean will be within \$5 thousand of the population mean.

(b) For all these interval estimates, the Central Limit Theorem can again be used since the sample sizes are large. Even though the population standard deviation σ is unknown in each case, the sample standard deviation s provides a reasonable estimate of σ when n is large. As a result, in each case the sample mean $\bar{x}$ is normally distributed with mean μ for the group in question and a standard deviation of $s/\sqrt{n}$. The interval estimates are as follows.

i. For females with less than twelve years of education, $\bar{x} = 20.0$, $s = 11.4$, and $n = 270$. For 99% confidence, the Z value is 2.576 or 2.575 – the following uses the former but even 2.57 or 2.58 for Z would also be ok.

$$\begin{aligned} \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} &= \bar{X} \pm 2.576\frac{11.4}{\sqrt{270}} \\ &= \bar{X} \pm 2.576\frac{11.4}{16.432} \\ &= \bar{X} \pm (2.576 \times 0.694) \\ &= \bar{X} \pm 1.787 \\ &= 20.0 \pm 1.787 \end{aligned}$$

And the interval is 20.0 ± 1.8 or (18.2 , 21.8) thousand dollars.

ii. For females with 14-17 years of education, $\bar{x} = 35.7$, $s = 18.4$, and $n = 594$. For 90% confidence, the Z value is 1.645.

$$\begin{aligned} \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} &= \bar{X} \pm 1.645\frac{18.4}{\sqrt{594}} \\ &= \bar{X} \pm 1.242 \\ &= 35.7 \pm 1.242 \end{aligned}$$

And the interval is 35.7 ± 1.2 or (34.5 , 36.9) thousand dollars.

iii. For males with 14-17 years of education, $\bar{x} = 50.2$, $s = 29.4$, and $n = 647$. For 90% confidence, the Z value is 1.645.

$$\begin{aligned} \bar{X} \pm Z\frac{\sigma}{\sqrt{n}} &= \bar{X} \pm 1.645\frac{29.4}{\sqrt{647}} \\ &= 50.2 \pm 1.901 \end{aligned}$$

And the interval is 50.2 ± 1.9 or (48.3 , 52.1) thousand dollars.

(c) For 98% confidence, the Z values of -2.33 and 2.33 can be used since $n = 208$ is large, so the Central Limit Theorem applies. The margin of error is plus or minus

$$E = \frac{Z\sigma}{\sqrt{n}} = \frac{2.33 \times 36.2}{\sqrt{208}} = \frac{84.346}{14.422} = 5.848$$

and the margin of error is $E = 5.8$ thousand dollars. Note that again the sample standard deviation s is used as an estimate of the population standard deviation since the latter in unknown.

For an estimate to be accurate to within \$2 thousand, the margin of error is to be $E = 2$. Again σ, the population standard deviation is unknown, so the sample standard deviation s is used in the formula for sample size. For 98% confidence, $Z = 2.33$ and the required sample size is

$$\begin{aligned} n &= \left(\frac{Z\sigma}{E}\right)^2 \\ &= \left(\frac{2.33 \times 36.2}{2}\right)^2 \\ &= 42.173^2 \\ &= 1,778.6 \end{aligned}$$

or $n = 1,779$.

I would report that a random sample of 1,779 males would very likely achieve the required margin of error and that the probability is 0.98 that the sample mean would differ from the population mean by no more than \$2 thousand. However, I would caution the supervisor that we do not know the population standard deviation,

and if the population is more variable than indicated in Table 1, an even larger sample size would be necessary.

Also, I would report that this is a very large sample size and the cost of achieving a random sample of Saskatchewan males with 18 or more years of education could be very large. I would likely indicate that the requirement for a $2 thousand margin of error is very demanding and the Department might reconsider whether they need such a small margin of error. In addition, it would be difficult to obtain a random sample of these males.

I would probably recommend a margin of error somewhere between 2 and, say, 4 thousand dollars, and suggest that the confidence level be lowered a little, say to 95%. In the end, I might report that the sample in Table 1 is reasonably large with a margin of error that is not all that great. I would ask the Department to justify the need for more precise information than provided in Table 1. Only if the Department has strong reasons for obtaining more accuracy would I recommend further sampling.

(d) The table shows that for both males and females, income generally increases with years of education, but the average income for females at each education level is lower than for males. For females, each extra level of education adds between $7 and $10 thousand in average income. For males, it adds a little more, between $8 and $12 thousand in average income. The male-female gap for each of the three groups with the least education is very similar, about $15 thousand for each level. Then the gap increases to around $18 thousand for those with 18 or more years of education.

In terms of how precise these results are, I would report that the results are not entirely reliable, since they are based on samples. However, parts a and b show that the margin of error is not all that large, even with confidence levels at 98 and 99 per cent. While these data do not exactly represent the situation for all Saskatchewan males and females with each of these levels of education who work full-time, and who are ages 40 to 59, the reasonably small margins of error mean that the results likely come close to representing the situation.

I would also caution the supervisor that this sample does not

represent the situation for all Saskatchewan males and females. This is a sample of only those aged 40 to 59 who have full-time jobs in 2000. While the sample provides a good estimate for these individuals, it very likely does not represent the situation for those with part-time jobs or those who are in other age groups.

3. Here is the relevant part of the data set for this question. To select a random sample, enter =RAND() into the column to the right of the data and use sort.

ID	GPA	Online	DrinkAge
1	78	1	21
2	69	5	20
3	69	1	21
4	70	3	18
8	70	3	18
6	60	5	18
7	68	1	18
8	61	0	21
9	77	1	21
10	65	2	19
11	68	2	18
12	62	4	20
13	64.3	4	18
14	69.4	2	18
15	83	3	18
16	75	2	19
17	93	1	18
18	72	1	19
19	69	2	18
20	65	2	16
21	67	5	18
22	60.5	2	21
23	72.5	0.25	19
24	82	1	18
25	71	2	19
26	68	2	18
27	70	4.5	21
28	65	2	16
29	84	3	18
30	88	2	16
31	63	1	18
32	84	4	18
33	89.6	6	18
34	78	0	19
35	67	5	21
36	75	5	16
37	88	0.5	19
38	67	2	19
39	75	1	19
40	67	4	18
41	75	4	19
42	70	1	20

a. Here is the random sample I obtained, using the method =RAND() in the Excel worksheet. After the sorting, these were the first 7 cases.

Sample of size 7

ID	GPA	DrinkAge
32	84	18
13	64.3	18
19	69	18
30	88	16
20	65	16
41	75	19
10	65	19

i. Statistics for GPA:

GPA for n=7	
Mean	72.9
Standard Error	3.682714
Median	69
Mode	65
Standard Deviation	9.743545
Sample Variance	94.93667
Kurtosis	-1.22261
Skewness	0.791649
Range	23.7
Minimum	64.3
Maximum	88
Sum	510.3
Count	7
Confidence Level(90.0%)	7.156177
	90%

Column1	
Confidence Level(95.0%)	9.011276
	95%

Column1	
Confidence Level(99.0%)	13.6534
	99%

For this sample, the mean GPA is 72.9, the standard deviation of GPA is 9.74.
The 90% interval estimate is 72.9 ±7.2 or (65.7 , 80.1).
The 95% interval estimate is 72.9 ±9.0 or (63.9 , 81.9).
The 99% interval estimate is 72.9 ±13.7 or (59.2 , 86.6).

Statistics for DrinkAge

Drinkage for n=7					
Mean	17.71429				
Standard Error	0.473804				
Median	18				
Mode	18				
Standard Deviation	1.253566				
Sample Variance	1.571429				
Kurtosis	-1.09917				
Skewness	-0.68169				
Range	3				
Minimum	16				
Maximum	19				
Sum	124	*Column1*		*Column1*	
Count	7				
Confidence Level(90.0%)	0.920686	Confidence Level(95.0%)	1.159355	Confidence Level(99.0%)	1.756593
	90%		95%		99%

The mean for DrinkAge in this sample of size 7 is 17.71 and the standard deviation is 1.25.
The 90% interval estimate is 17.71 ± 0.92 or (16.79 , 18.63) pr 16.8 to 18.6 years.
The 95% interval estimate is 17.71 ± 1.16 or (16.55 , 18.87) pr 16.6 to 18.9 years.
The 99% interval estimate is 17.71 ± 1.76 or (15.95 , 19.47) pr 16.0 to 19.5 years.

ii. Statistics for the whole data set.

GPA		*Drinkage*	
Mean	72.24524	Mean	18.69048
Standard Error	1.29454	Standard Error	0.216851
Median	70	Median	18
Mode	75	Mode	18
Standard Deviation	8.389575	Standard Deviation	1.405357
Sample Variance	70.38498	Sample Variance	1.975029
Kurtosis	-0.06582	Kurtosis	-0.17975
Skewness	0.807333	Skewness	0.087078
Range	33	Range	5
Minimum	60	Minimum	16
Maximum	93	Maximum	21
Sum	3034.3	Sum	785
Count	42	Count	42

For the sample of size 7:
GPA, mean = 72.9 and sd = 9.74. This compares with an overall mean of 72.2 and sd of 8.39. The sample mean lies very close to the population mean and the sample standard deviation is reasonably close to the sd of the population – a difference of 9.74-8.39=1.35. Since the sample mean is so close to the population mean, the interval estimates all contain the population mean. But each random sample will yield different cases, so not all samples will be this close.

For DrinkAge, the sample of size 7 has a sample mean of 17.71 years and a sample standard deviation of 1.25 years. The population mean is 18.69, so the sample mean is low by 17.71-18.69=0.98 years. The sample standard deviation of 1.25 is quite close to the population standard deviation of 1.41. From the interval estimates, it can be seen that the 90% interval estimate does not contain the population mean but the 95% and 99% interval estimates of the mean contain the population mean.

iii. Since this is a small sample, even though the population standard deviation is known, it is best to use a t-distribution. The t value for the 95% interval estimate, with n-1=7-1=6 degrees of freedom is 2.447.

The margin of error is t times s divided by the square root of n, so this is (2.447 x 9.743545) / 2.645 = 9.0116 and this comes close to matching the Excel result of 9.0113. Rounding could account for the small difference.

b. Simulation.

i. and ii. In conducting the simulation, it is important to start the random selection process each time a new sample is selected. If you select the first ten rows for the first sample, then the second ten rows for the second sample, and the third ten rows for the third sample, the first sample has been randomly selected but the second and third are not. I obtained the following samples and report the ID and values of Online. The dexcriptive statistics for the 3 samples and the whole data set are as follows.

Sample 1

ID	Online
6	5
32	4
4	3
13	4
27	4.5
39	1
29	3
8	0
22	2
36	5

Column1	
Mean	3.15
Standard Error	0.537742
Median	3.5
Mode	5
Standard Deviation	1.70049
Sample Variance	2.891667
Kurtosis	-0.44061
Skewness	-0.74483
Range	5
Minimum	0
Maximum	5
Sum	31.5
Count	10
Confidence Level(95.0%)	1.216457

Sample2

ID	Online
41	4
27	4.5
20	2
21	5
16	2
15	3
22	2
31	1
29	3
35	5

Column1	
Mean	3.15
Standard Error	0.447524
Median	3
Mode	2
Standard Deviation	1.415195
Sample Variance	2.002778
Kurtosis	-1.40722
Skewness	0.060273
Range	4
Minimum	1
Maximum	5
Sum	31.5
Count	10
Confidence Level(95.0%)	1.01237

Sample 3

ID	Online
12	4
11	2
35	5
18	1
34	0
40	4
31	1
38	2
37	0.5
9	1

Column1	
Mean	2.05
Standard Error	0.539804
Median	1.5
Mode	1
Standard Deviation	1.707012
Sample Variance	2.913889
Kurtosis	-0.95315
Skewness	0.689412
Range	5
Minimum	0
Maximum	5
Sum	20.5
Count	10
Confidence Level(95.0%)	1.221123

All cases	
Mean	2.434524
Standard Error	0.248845
Median	2
Mode	2
Standard Deviation	1.612697
Sample Variance	2.600791
Kurtosis	-0.82549
Skewness	0.497273
Range	6
Minimum	0
Maximum	6
Sum	102.25
Count	42

iii. Variable is Online

	Sample 1	Sample 2	Sample 3	Population
Mean	3.15	3.15	2.05	2.43
sd	1.70	1.42	1.71	1.61
N	10	10	10	42
Margin of error at 95%	1.22	1.01	1.22	

None of the samples had sample means that were real close to the population mean – the first two were over 0.6 hours on the high side and Sample 3 has a mean about a third of an hour on the low side. However, each of the three margins of error, at 95% confidence, is sufficiently large to include the population mean.

The sample standard deviations are relatively close approximations of the population standard deviation. None is off by more than 0.2 hours. This gives us some confidence that the sample standard deviation often provides a reasonable estimate of the population standard deviation.

iv. The simulation demonstrates that sample means can vary a lot and are not always real close to the population mean, especially when the sample size is small. However, as noted above, the margins of error are sufficiently large in each case so the confidence interval for each of the three samples includes the population mean. Some of you may draw samples where the sample mean and the margin of error do not result in an interval that contains the population mean – but this should be the case in only about 5 of every 100 random samples drawn.

One value of a simulation of this sort is for researchers to see whether actual samples behave in the way that statistical theory predicts. The three samples here all result in sample means with the margin of error but, as noted in the last paragraph, there will be around 5 of 100 samples where the interval estimate does not contain the population mean.