MONT 107N – Understanding Randomness
Sample Question Solutions for Midterm Exam

(*Note:* The actual exam will be somewhat shorter than this; there are parts here to illustrate most of the different kinds of questions that might appear.)

I. Assume 125 draws with replacement are made from the following box:

1 3 5 7

A) What are the EV and SE for the sum of numbers on the drawn tickets?

Solution – The average of the box is 4 and the SD of the box is $\sqrt{5} \doteq 2.24$. So with 125 draws, EV $= 125 \times 4 = 500$ and the SE $= \sqrt{125} \times \sqrt{5} = 25$.

B) What are the chances that sum of the numbers will be 480 or larger?
Solution – In standard units, 480 corresponds to

$$z = \frac{480 - 500}{25} = -.8$$

The area under the normal curve with $z \geq -.8$ is the central area, plus the upper tail area:

$$A(.8) + \frac{1}{2}(100\% - A(.8)) = 50\% + 28.815\% \doteq 78.815\%.$$

II. Consider the following box.

0 1 2 2 2 3 3 4 4 4

A) If a single draw is made from the box, construct the probability histogram for the number that is drawn.

Solution – The probability histogram should have 5 boxes: the first from -.5 to .5 with height $1/10 \times 100\% = 10\%$ representing the chance of drawing a 0, second from .5 to 1.5 with height 10%, third from 1.5 to 2.5 with height 30%, fourth from 2.5 to 3.5 with height 20%, last from 4 to 4.5 with height 30% (chance of drawing a 4). Note the total area is 100% as always for us.

B) If 1000 draws with replacement were made from this box, draw an approximate probability histogram for the value of the sum. Show carefully the location of the EV and the SE for the sum relative to your histogram.

Solution – The average of this box is $\frac{25}{10} = 2.5$ and the SD is about 1.28. From the Central Limit Theorem (since every integer between 0 and 4000 can be obtained as a sum of tickets from this box), we would expect to see something very close to a shifted and scaled *normal curve* with peak at the EV for the sum: $1000 \times 2.5 = 2500$, and inflection points (concavity changes) at

$$2500 \pm \text{SE for sum} \doteq 2500 \pm (\sqrt{1000} \times 1.28) \doteq 2500 \pm 40.62.$$

(Note: A completely to-scale drawing should also show the correct heights so that the area is exactly 100%. We did not really discuss how those would be obtained. So, any recognizably normal graph would be OK here!)

C) Now, assume 200 draws with replacement are made from this box. What box model would you use in order to model *the number of 4s* that are chosen?

Solution – We create a 0/1 box by replacing all non-4s with 0s and all 4's with 1s:

$$0, 0, 0, 0, 0, 0, 0, 1, 1, 1.$$

D) Exactly 200 draws are made and 4s are drawn a total of 67 times. What is the difference between for the number of 4s for these 200 draws and the expected value of the number of 4s (that is, the *chance error* for this sample)?

Solution – The EV for the sum of 200 draws from the 0/1 box is $200 \times .3 = 60$. This is the same as the expected number of 4s. The chance error is $67 - 60 = 7$.

E) Would you consider the chance error from part D to be a large chance error, roughly an average chance error, or a small chance error? Explain your answer. (Hint: You may want to do some additional calculations to decide on your answer.)

Solution – To decide, we need to know the SE, which is $\sqrt{200} \times \sqrt{(.3)(.7)} = 6.48$ (using the "short-cut rule" which applies for 0/1 boxes). The chance error in part D is an *average chance error*, since we expect most sums will be within 1 SE of the EV.

III. On March 3, 2010, the *New York Times* reported results of a survey in which researchers contacted 600 people who graduated from a variety of high schools across the country in the last 4 to 12 years. The study was sponsored by the Bill and Melinda Gates Foundation, which has sought to shed light on low completion rates at both the high school and college levels. For the purposes of this question, assume the 600 survey respondents were a simple random sample of all recent high school graduates. In this sample, 402 people indicated that they thought their high school guidance counselor was "poor" or "fair" in helping them decide what colleges or technical schools would be good for them.

A) Based on this sample, construct a 95% confidence interval for the the percentage of all recent high school graduates who think their guidance counselor was "poor" or "fair" in helping them decide what colleges or technical schools would be good for them.

Solution – The observed percent is $402/600 \times 100\% = 67\%$. We use the "bootstrap" to estimate the SE for the percent:

$$\text{SE} \doteq \sqrt{\frac{(.67) \times (.33)}{600}} \times 100\% \doteq 1.9\%$$

So the approximate 95% confidence interval for the population percentage would be $67\% \pm 2 \times 1.9\%$, or 62.2% to 70.8%.

B) The *Times* reported that the survey had a sampling error of $\pm 5\%$. Does that match your confidence interval? If not, can you suggest where the 5% figure comes from?

Solution – Our estimate would say ±3.8%. But of course that used the bootstrap method, which depended on the observed percent from the sample. The "conservative" method we discussed in class would estimate the SE by $\sqrt{\frac{(.5)\times(.5)}{600}} \times 100\% \doteq 2.0\%$, so ±2SE would be ±4%, which is still less than the reported ±5% error bound. It might also be the case that the researchers were not using a simple random sample, and hence that they needed to use an even more conservative method to estimate the standard error for the percent.

IV. A sociologist takes a simple random sample of size 300 from the 27000 students at a large state university. In the sample, $222/300 \times 100\% = 74\%$ are undergraduates. The SE for the percentage is calculated as 2.5% and the sociologist writes down the 95% confidence interval for the percentage as $74\% \pm 5\%$. For each part, say whether the statement is true or false, and explain:

A) From the information in this sample it is believable that the actual percentage of undergraduate students is 78%.

Solution – *True* – any value in the confidence interval should be considered as a believable value for the population percentage.

B) The range from 69% to 79% is a 95% confidence interval for the percentage of undergraduates in the sample.

Solution – *False* – the percentage of undergradates *in the sample* is known (74%). There is no chance or uncertainty about that. This would be true if "sample" was replaced by "population." (Be sure you read slowly and carefully!)

C) The 95% figure comes from an area under the normal curve.

Solution – *True* – the bounds for the confidence interval come from the area between −2 and +2 standard units under the normal curve.

D) The probability that the true population percentage of undergraduates is in the interval 69% to 79% is 95%.

Solution – *False* – This is the incorrect interpretation of the confidence interval we saw in class and in the second lab. The chance is in the sampling process, not in the location of the true population percentage.

E) If a second sample of size 300 was taken in the same semester, and a 95% confidence interval was computed from that sample, then the result could be 67% to 79%.

Solution – *False* – (This is a slightly tricky question!) If this was the confidence interval then the observed percentage would have to be the midpoint, or 73%. That is certainly possible. But note then that the "bootstrap" formula for the estimated SE would give $\sqrt{.73 \times .27/300} \times 100\% \doteq 2.6\%$. The estimated 95% confidence interval would be narrower: $73\% \pm 5.2\%$.

F) If samples of size 1200 were taken in the same semester, then on average, we would expect the SE to decrease and the interval to be narrower.

Solution – *True* – On average and other things being equal (i.e. assuming the observed proportions p were the same), the larger number of samples will yield a narrower confidence interval:

$$\sqrt{\frac{(p)(1-p)}{1200}} < \sqrt{\frac{(p)(1-p)}{300}}.$$

G) The process of drawing the sample is like drawing 300 from a box with 27000 tickets and recording the sum, where the ticket for each first-year student has the value 1, the ticket for each second-year student has the value 2, the ticket for each third-year student has the value 3, the ticket for each fourth-year student has the value 4, and the ticket for each graduate student has the value 5.

Solution – *False* – This should be a 0/1 box, since we are just counting/classifying.