MONT 105N – Analyzing Environmental Data
Solutions for Midterm Exam Practice Problems
March 26, 2013

A. A. Waterbuck are large mammals in the deer family native to eastern South Africa. In 1976, biologists launched a three-year study of waterbuck ecology and determined the annual survival and fertility rates for females given in the following table:

	A = calves	B = yearlings	C = adults
fertility	0	0.048	0.081
survival	0.94	1.00	0.75

1. Express the information given above as a life-cycle graph.

   Solution: The life cycle graph should show the three groups A, B, C, with arrows
   - labeled 0.94 from A to B,
   - labeled 1.00 from B to C,
   - labeled .75 from C to itself,
   - labeled .048 from B back to A (new female births from yearling mothers)
   - labeled .081 from C back to A (new female births from adult mothers)

2. Express the information given above as a system of difference equations (time n in years).

   Solution: The corresponding difference equations would be

$$\begin{aligned} A(n) &= .048B(n-1) + .081C(n-1) \\ B(n) &= .94A(n-1) \\ C(n) &= B(n-1) + .75C(n-1) \end{aligned}$$

3. If the initial conditions are $A(0) = 100$, $B(0) = 2000$, $C(0) = 3000$, find the population of each group in years $n = 1, 2, 3, 4$.

   Solution: From the initial values, we compute

$$\begin{aligned} A(1) &= (.048)(2000) + (.081)(3000) = 339 \\ B(1) &= (.94)(100) = 94 \\ C(1) &= 2000 + (.75)(3000) = 4250 \end{aligned}$$

   Then the successive years are computed similarly yielding the following table of results (all values rounded to the nearest whole number):

year	A	B	C
0	100	2000	3000
1	339	94	4250
2	349	319	3282
3	281	328	2780
4	241	264	2413

(If you continue for several more years it becomes clear that the overall population is in a steep decline.)

B. The following data set has $n = 9$: $23, 28, 40, 44, 47, 50, 51, 54, 55$

1. Find the "5-number" summary for this data set.

   Solution: Min = 23, $Q1 = 40$ (median of "lower half" – including the 47 in the middle), Median = 47, $Q3 = 51$ (median of "upper half"), Max = 55

2. Draw the corresponding box plot.

   Solution: Show a box from $Q1$ to $Q3$, with the position of the Median marked by a vertical line, plus "whiskers" out to the Min and Max on either side.

3. Compute the (Bowley) measure of skewness. Does this seem reasonable from the box plot?

   Solution:

$$\text{skewness} = \frac{Q3 - 2 \times \text{Median} + Q1}{Q3 - Q1} = \frac{51 - 2 \cdot 47 + 40}{51 - 40} \doteq -0.27$$

   The boxplot shows the negative skew since the "whisker" on the left and the distance from $Q1$ to the Median are larger than the distance from the median to $Q3$ and the length of the right "whisker."

4. Compute the SD of the data set. How many of the points lie within two SD's of the mean? Is Chebyshev's Rule satisfied here? (Say what that rule says, and determine whether or not it is satisfied.)

   Solution: Mean: $\overline{x} \doteq 43.6$, so we will compute SD using the formula

$$SD = \sqrt{\frac{(23 - 43.6)^2 + (28 - 43.6)^2 + \cdots + (55 - 43.6)^2}{8}}$$

   (see the class notes for an explanation why it's *not* divided by 9 in the square root). This gives an approximate value $SD = 11.3$. The interval $\overline{x} \pm 2SD$ is $(20.9, 66.2)$. *All of the data points lie in this range.* That is consistent with Chebyshev's rule, because that says *at least 75%* of the data points should lie in this interval. (Note: Chebyshev's rule is an honest-to-God rule – it is *always* valid(!))

C. (Short answer) Suppose that a researcher collects 80 individuals of the Atlantic surf clam. These clams can be found at levels down to about a meter in the sand, and larger clams tend to live at deeper levels. The researcher finds an average shell width 10.2 cm. Think of this as a sampling process.

1. What is the population? What is the sample?

   Solution: The population is the collection of widths of all Atlantic surf clam shells. The sample is the $n = 80$ width measurements described in the problem.

2. Is the 10.2 a statistic or a parameter of the population?

   Solution: It is a *statistic* – something computed from the measured widths in the sample. The parameter would be the population mean – the average width of all the shells.

3. Would the researcher *know* the population mean in this circumstance?

   Solution: No. (Research like this is usually aimed at estimating population parameters like the mean shell width, but those values are never known exactly.)

4. What additional information would the researcher need in order to find a *confidence interval* for the population mean? Describe how that would be determined and how that would be interpreted.

   Solution: The additional information needed to derive the confidence interval is just the SD of the measured values. Since $n = 80 \geq 30$, the confidence interval would be computed using

   $$10.2 \pm 1.96 \cdot \frac{SD}{\sqrt{80}}$$

5. If you knew that reasearcher was being lazy about digging and the clams he collected were all taken from sand levels no deeper than 10cm, would that be a *simple random sample*?

   Solution: No. A simple random sample is produced by a process that would find any given collection of 80 shell widths from the population with equal probability. Here, the sampling process will be missing clams from the lower sand levels. (The results will be skewed toward smaller clams.)

D. Suppose that a large data set of air temperature readings is normally distributed with $\overline{x} = 18.6°\text{C}$ and $SD = .2°$ C.

1. What would be the z-score of a reading of 17.9°?

   Solution: $z = \frac{17.9-18.6}{.2} = -3.5$.

2. What temperature reading would correspond to a z-score of 1.4?

   Solution: $\frac{x-18.6}{.2} = 1.4$ when $x = 18.6 + (1.4)(.2) = 18.88$.

3. Based on this information, if a temperature reading T is selected at random from the data set, what is the probability that $18.2° \leq T \leq 18.9°$?
   Solution: $x = 18.2 \leftrightarrow z = -2$ and $x = 18.9 \leftrightarrow z = 1.5$. The probability we want is the area under the standard normal curve from $z = -2$ to $z = 1.5$. This can be found from the table and the symmetry of the normal curve: Area(2.00) + Area(1.50) = $.4772 + .4332 = .9104$ (about a 91% chance).

4. Based on this information, if a temperature reading T is selected at random from the data set, what is the probability that $T > 19.0°$?

Solution: Again we use the table, but since $x = 19.0 \leftrightarrow z = 2$, we want the area to *the right of* $z = 2$, which is $.5 - \text{Area}(2.00) = .5 - .4772 = .0228$.

E. Physicians measured the blood lead levels in 373 bridge workers employed by painting contractors in eight states. The lead levels had $\overline{x} = 27.2$ micrograms per liter of blood, with an SD of 16.1 micrograms per liter.

1. Determine a 95% confidence interval for the average lead level in bridge workers.

   Solution: Since $n = 373$ we use the value 1.96 in the margin of error formula and we don't need to consult the t-table. The confidence interval is

$$\overline{x} \pm 1.96 \cdot \frac{SD}{\sqrt{n}} = 27.2 \pm 1.96 \cdot \frac{16.1}{\sqrt{373}} = 27.2 \pm 1.6.$$

   In other words, we are "96% confident" that the mean lead level is between $27.2 - 1.6 = 25.6$ and $27.2 + 1.6 = 28.8$ micrograms per liter.

2. A health objective of a federal regulatory agency was the elimination of blood lead levels of 28 micrograms per liter or higher for these workers. From the evidence given by your confidence interval, does it seem that that objective was being met? Explain, by describing the way we interpret the meaning of a confidence interval of this sort.

   Solution: The interval we computed in part 1 contains both numbers less than the target level of 28 micrograms per liter and numbers greater than or equal to that. All of them should be taken as "believable" values for the population average lead level, based on the information in the sample. So we cannot say that the objective is being met. (We cannot say for sure that it is *not* being met either.)

F. A study shows that a 95% confidence interval for the average amount X of hazardous waste generated by a single hospital is $210 \leq X \leq 260$ (in units of kg/day). This interval was computed using the formulas we have discussed.

1. What was the sample mean $\overline{x}$ used to generate this confidence interval? What was the margin of error?

   Solution: Because of the way the formulas work, the sample mean is always the *midpoint* of the interval – here $(210 + 260)/2 = 235$ kg/day. The margin of error is then the distance from there to either endpoint: $260 - 235 = 25$ (and $235 - 210 = 25$), also in kg/day.

2. If the sample size was $n = 100$, what was the SD of the waste amounts in the sample?

   Solution: The SD can be found since we would compute the margin of error by the usual formula with $n \geq 30$: $1.96 \cdot \frac{SD}{\sqrt{n}}$, So

$$25 = 1.96 \cdot \frac{SD}{\sqrt{100}} \Rightarrow SD = \frac{250}{1.96} \doteq 125.6$$

3. If the sample size was $n = 16$, what was the SD of the waste amounts in the sample?

Solution: This is similar to part 2, except that now, since $n < 30$, we need to use the entry from the t-table for $n = 16$:

$$25 = 2.132 \cdot \frac{SD}{\sqrt{16}} \Rightarrow SD = \frac{100}{2.132} \doteq 46.9.$$