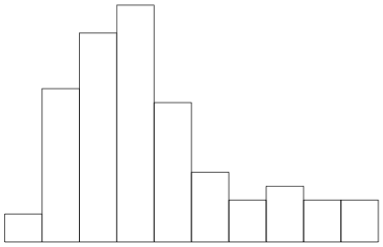
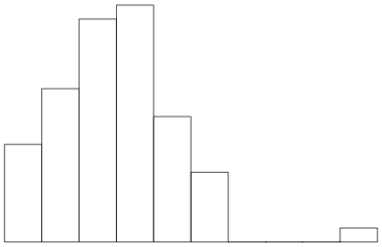
Math 216 Spring 2012
Problem Set 1 Answer Key

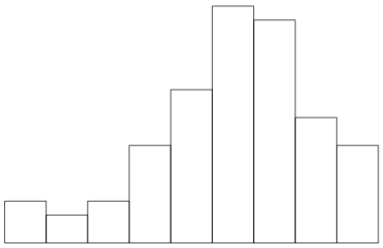
1. Sketch three histograms **[9pts]**

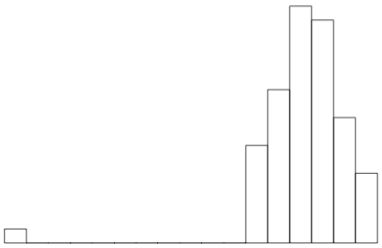
   (a) Extreme values affect the mean much more than the median.
   Examples: A right-skewed distribution or one with an outlier to the far right.



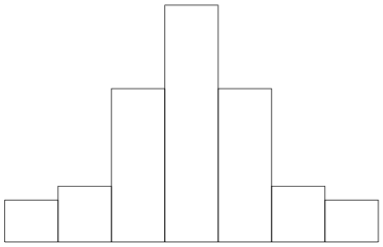


   (b) Examples: A left skewed distribution or one with an outlier to the far left.





   (c) Any symmetric distribution



2. The best sampling strategy is C, because it is the best approximation of a random sample of students. (Many of you picked B, because it is a larger sample size. But we cannot assume that the engineering majors are a good representation of all Vanderbilt students. Moreover, n=50 is large enough for the data to be meaningful.) **[5pts]**

3. **[20pts]**

   (a) Sample mean

   $$\bar{x} = \frac{1(0) + 3(1) + 2(2) + \cdots + 216(8) + 131(9) + 18(10))}{1000} \approx \mathbf{7.138}$$

   (b) Sample variance

   $$s^2 = \frac{1(0-\bar{x})^2 + 3(1-\bar{x})^2 + \cdots + 131(9-\bar{x})^2 + 18(10-\bar{x})^2}{1000-1} \approx 1.719$$

   Sample standard deviation

   $$s = \sqrt{s^2} \approx \mathbf{1.311}$$

   (c) $n = 1000$ so the median is halfway between the 500 and the 501st highest values. Simply add across the number of babies to find that both of those values are equal to 7. Hence, the median = **7**

   (d) First quartile is halfway between the 250 and 251st values. First quartile = **6**

   (e) Within one standard deviation of the mean is the interval $[5.828, 8.448]$. The only Apgar scores that fall within the interval are 6, 7, and 8. There are $198 + 367 + 216 = 781$ of these scores. Therefore, the proportion of scores within one standard deviation is **78.1%**

4. **[8pts]**

   (a) The **median** best describes the "typical" income of the patrons at the coffee shop. The addition of the two extremely high incomes significantly increases the mean, even though the sample population (the patronage of the coffee shop) has changed very little. On the other hand, the addition of two high values does not move the median by very much.
   Because the median is less affected by extreme values, it is the more robust measure.

   (b) Similarly, the **IQR** is the best indicator of variability in income among patrons of the coffee shop. It is less affected by the extreme values than standard deviation and is the more robust measure.

5. **No**, smoking is not independent of marital status based on the mosaic plot. If smoking was independent of marital status, we would expect the distribution of marital status within the smoker and non-smoker populations (heights of the bars within each column) to be the same. **[5pts]**

6. [**15pts**]

   (a) **False**

   (b) **True**. Compare the tip of the tail of the 4.0 sample to the bottom edge of the box on the 4.5 sample

   (c) **False**

   (d) **False**. Similar to a clicker question from class

   (e) **True**. Look at the tails because the boxes are approximately symmetric on both samples. Both samples have a longer tail on the right.

7. Matching histograms to box plots [**16pts**]

   (a) **(4)** Look for a right skewed box plot. (Long tail on top)

   (b) **(2)** Left skewed (Long bottom tail on box plot) and this is the only histogram that has a value that might be an outlier.

   (c) **(1)** Left skewed, but no clear outlier.

   (d) **(3)** Look for symmetry

8. **Bubble Chart 1** is correctly constructed.
   Chart 1 depicts the population of each country by the *area* of the circle, where as Chart 2 uses the *radius*. The problem with using the radius (Chart 2) is that the human eye is much better at seeing area than radius, so the former approach is much more visually meaningful. By using the radius, Chart 2 is in effect inordinately emphasizing the countries with larger populations. [**6pts**]

9. A heat map is useful for depicting relative values—in this case, the relative footprints of each country. The key here is that the color code depicts relative footprints in each column, *not* across the entire chart. In other words, we cannot compare the colors of cells across different columns in a meaningful way.
   So while the Norway Fishing Ground Footprint cell is one of the darkest red cells *on the entire chart*, it merely means that **Norway has a very large Fishing Ground Footprint relative to the Fishing Ground Footprints of other countries**, but it says nothing about the size of Norway's Fishing Ground Footprint relative to, for example, the Carbon Footprint of France. [**6pts**]

10. [**8pts**]

    (a) Chart 1 shows the *per capita* ecological footprint of each country by industry, while Chart 2 shows the actual ecological footprint of each country.

    (b) One way to think about this is that Chart 1 is useful for identifying room for improvement in a country's particular industry, while Chart 2 is more useful when trying to tackle ecological footprint on a global scale. For example, Denmark has

a high per capita carbon footprint relative to other countries (from Chart 1), but Denmark's carbon footprint is a tiny fraction of the overall carbon footprint across the seven depicted countries (Chart 2).