2.5 Problems

1. Follow the instructions in the posted \( \text{R} \) code folder (http://www.stat.wisc.edu/~ifischer/Intro_Stat/Lecture_Notes/Rcode/) for this problem, to reproduce the results that appear in the lecture notes for the “Memorial Union age” data.

2. A numismatist (coin collector) has a large collection of pennies minted between the years 1946-1962, when they were made of bronze: 95% copper, and 5% tin and zinc. (Today, pennies have a 97.5% zinc core; the remaining 2.5% is a very thin layer of copper plating.) The year the coin was minted appears on the obverse side (i.e., “heads”), sometimes with a letter below it, indicating the city where it was minted: D (Denver), S (San Francisco), or none (Philadelphia). Before 1959, a pair of wheat stalks was depicted on the reverse side (i.e., “tails”); starting from that year, this image was changed to the Lincoln Memorial. The overall condition of the coin follows a standard grading scale – Poor (PO or PR), Fair (FA or FR), About Good (AG), Good (G), Very Good (VG), Fine (F), Very Fine (VF), Extremely Fine (EF or XF), Almost Uncirculated (AU), and Uncirculated or Mint State (MS) – which determines the coin’s value.

(a) Using this information, classify each of the following variables as either numerical (specify continuous or discrete) or categorical (specify nominal: binary, nominal: not binary, or ordinal).

- Amount of zinc
- Image on reverse
- Year minted
- City minted
- Condition

(b) Suppose the collector accidentally drops 1000 pennies. Repeat the instructions in (a) for the variables

- Number of heads face-up
- Proportion of heads face-up

3. Sketch a dotplot (by hand) of the distribution of values for each of the data sets below, and calculate the mean, variance, and standard deviation of each.

- \( U \): 1, 2, 3, 4, 5, 6, 7
- \( X \): 2, 3, 4, 4, 4, 5, 6
- \( Y \): 3, 4, 4, 4, 4, 4, 5
- \( Z \): 4, 4, 4, 4, 4, 4, 4

What happens to the mean, variance, and standard deviation, as we progress from one data set to the next? What general observations can you make about the relationship between the standard deviation, and the overall shape of the corresponding distribution? **In simple terms, why should this be so?**

4. Useful Properties of Mean, Variance, Standard Deviation

(a) Suppose that a nonzero constant \( b \) is added to every value of a generic sample dataset \( \{x_1, x_2, x_3, \ldots, x_n\} \), to produce a new dataset \( \{x_1 + b, x_2 + b, x_3 + b, \ldots, x_n + b\} \). Provide a formal mathematical (i.e., algebraic) proof for what happens to the mean, variance, and standard deviation. (**Hint:** Think of the dotplot for informal motivation.)

(b) Suppose that every value in a generic dataset \( \{x_1, x_2, x_3, \ldots, x_n\} \) is multiplied by a nonzero constant \( a \) to produce a new dataset \( \{ax_1, ax_2, ax_3, \ldots, ax_n\} \). Provide a formal mathematical (i.e., algebraic) proof for what happens to the mean, variance, and standard deviation. (**Hint:** Think of the dotplot for informal motivation, and don’t forget that \( a \) can be negative!)
5. During a certain winter in Madison, the variable $X = \text{“Temperature at noon (°F)”}$ is measured every day over two consecutive weeks, as shown below.

<table>
<thead>
<tr>
<th></th>
<th>Sun</th>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>+8</td>
<td>+8</td>
<td>+8</td>
<td>+5</td>
<td>+3</td>
<td>+3</td>
<td>0</td>
</tr>
<tr>
<td>Week 2</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-5</td>
<td>-8</td>
<td>-8</td>
<td>-8</td>
</tr>
</tbody>
</table>

(a) Calculate the sample mean temperature $\bar{x}$ and sample variance $s^2$ for Week 1.

(b) Without performing any further calculations, determine the mean temperature $\bar{x}$ and sample variance $s^2$ for Week 2. [Hint: Compare the Week 2 temperatures with those of Week 1, and use the result found in 4(b).] Confirm by explicitly calculating.

6. A little practice using R: First, type the command `pop = 1:100` to generate a simulated “population” of integers from 1 to 100, and view them (read the intro to R to see how).

(a) Next, type the command `x.vals = sample(pop, 5, replace = T)` to generate a random sample of $n = 5$ values from this population, and view them. Calculate, without R, their sample mean $\bar{x}$, variance $s^2$, and standard deviation $s$. Show all work!

(b) Use R to calculate the sample mean in two ways: first, via the `sum` command, then via the `mean` command. Do the two answers agree with each other? If so, label this value `xbar`. Include a copy of the R output in your work.

(c) Use R to calculate the sample variance in two ways: first, via the `sum` command, then via the `var` command. Do the two answers agree with each other? Do they agree with (a)? If so, label this value `s.sqrd`. Include a copy of the R output in your work.

(d) Use R to calculate the sample standard deviation in two ways: first, via the `sqrt` command, then via the `sd` command. Do the two answers agree with each other? Do they agree with (a)? Include a copy of the R output in your work.

7. Note: You may want to run the program in the Rcode folder for this problem, as motivation. Consider a generic dataset $\{x_1, x_2, x_3, \ldots, x_n\}$, and their sample mean $\bar{x}$ and standard deviation $s_x$.

(a) Provide a formal mathematical (i.e., algebraic) proof that the deviations from the mean $x_i - \bar{x}$ sum to 0. (Hint: Use the formal definition of $\bar{x}$, and elementary properties of summations.)

(b) Now divide each of these individual deviations by the standard deviation $s_x$. These new values $\{z_1, z_2, z_3, \ldots, z_n\}$ are called “standardized” values, i.e., $z_i = \frac{x_i - \bar{x}}{s_x}$, for $i = 1, 2, 3, \ldots, n$.

Provide a formal mathematical (i.e., algebraic) proof for what happens to their mean $\bar{z}$ and standard deviation $s_z$. (Hint: Use problem 4.)
8. (a) The average score of a class of \( n_1 = 20 \) students on an exam is \( \bar{x}_1 = 90.0 \), while the average score of another class of \( n_2 = 30 \) students on the same exam is \( \bar{x}_2 = 80.0 \). If the two classes are pooled together, what is their combined average score on the exam?

(b) Suppose two other classes – one with \( n_1 = 24 \) students, the other with \( n_2 = 44 \) students – have the same mean score, but with standard deviations \( s_1 = 7.0 \) and \( s_2 = 10.0 \), respectively. If these two classes are pooled together, what is their combined standard deviation on the exam? (Hint: Think about how sample standard deviation is calculated.)

9. (Hint: See page 2.3-11) A random sample of \( n = 90 \) people is grouped according to age in the frequency table below:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 10)</td>
<td>9</td>
</tr>
<tr>
<td>[10, 25)</td>
<td>27</td>
</tr>
<tr>
<td>[25, 65]</td>
<td>54</td>
</tr>
</tbody>
</table>

(a) Calculate the group mean age and group standard deviation. Express in years and months.

(b) Construct a relative frequency histogram.

(c) Construct a density histogram.

(d) What percentage of the sample falls between 15 and 35 years old?

(e) Calculate the group quartile ages \( Q_1, Q_2, Q_3 \). Express in terms of years and months.

(f) Calculate the range and the interquartile range. Express in terms of years and months.

10. For any \( x = 0, 1, 2, \ldots, n \), consider a data set \( \{y_1, y_2, y_3, \ldots, y_n\} \) consisting entirely of \( x \) ones and \((n-x)\) zeros, in any order. For example, \( \{1, 1, \ldots, 1, 0, 0, \ldots, 0\} \). Also denote the sample proportion of ones by \( p = \frac{x}{n} \).

(a) How many such possible data sets can there be?

(b) Construct a relative frequency table for such a data set.

(c) Show that the sample mean \( \bar{y} = p \).

(d) Show that the sample variance \( s_y^2 = \frac{n}{n-1} p (1 - p) \).
11. (a) Consider the sample data \{10, 10, 10, ..., 10, 60, 60, 60, ..., 60\}, where half the values are 10 and half the values are 60. Complete the following relative frequency table for this sample, and calculate the sample mean \(\bar{x}\) and sample median.

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>(p(x_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

(b) Suppose the original dataset is unknown, and only given in grouped form, with each of the two class intervals shown below containing half the values.

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 20)</td>
<td></td>
</tr>
<tr>
<td>[20, 100)</td>
<td></td>
</tr>
</tbody>
</table>

- Complete this relative frequency table, and calculate the group sample mean \(\bar{x}_{\text{group}}\) and group sample median. How do these compare with the values found in (a)?
- Sketch the relative frequency histogram.
- Sketch the density histogram.
- Label the group sample mean and median in each of the two histograms. In which histogram does the mean more accurately represent the “balance point” of the data, and why?

12. By the end of the semester, Merriman forgets the scores he received on the four quizzes (each worth 100 points) he took in a certain course. He only remembers that their average score was 80 points, standard deviation 10 points, and that 3 out of the 4 scores were the same. From this information, compute all four missing quiz scores. [Hint: Recall that the \(i^{\text{th}}\) deviation of a value \(x_i\) from the mean \(\bar{x}\) is defined as \(d_i = x_i - \bar{x}\), so that \(x_i = \bar{x} + d_i\) for \(i = 1, 2, 3, 4\). Then use the given information.]

Note: There are two possible solutions to this problem. Find them both.
13. **Linear Interpolation** (A generalization of the method used on page 2.3-6.)

If software is unavailable for computations, this is an old technique to estimate values which are “in-between” tabulated entries. It is based on the idea that over a small interval, a continuous function can be approximated by a linear one, i.e., constant slope.

Given two successive entries \( a_1 \) and \( a_2 \) in the first column of a table, with corresponding successive entries \( b_1 \) and \( b_2 \), respectively, in the second column. For a given \( x \) value between \( a_1 \) and \( a_2 \), we wish to approximate the corresponding \( y \) value between \( b_1 \) and \( b_2 \), or vice versa. Then assuming equal proportions, we have

\[
\frac{y - b_1}{x - a_1} = \frac{b_2 - b_1}{a_2 - a_1}.
\]

Show that this relation implies that \( y \) can be written as a **weighted average** of \( b_1 \) and \( b_2 \). In particular,

\[
y = \frac{v_1 b_2 + v_2 b_1}{v_1 + v_2},
\]

where the weights are given by the differences \( v_1 = x - a_1 \) and \( v_2 = a_2 - x \). Similarly,

\[
x = \frac{w_1 a_2 + w_2 a_1}{w_1 + w_2},
\]

where the weights are given by the differences \( w_1 = y - b_1 \) and \( w_2 = b_2 - y \).