Chapter 1

Numerical Summaries of Sample Distributions

1.1 The "Average" of a Distribution

Suppose you have a *sample distribution*, that is, a list of n numbers $\{x_1, x_2, ..., x_n\}$, listed from smallest to largest.

- A **mode** of the distribution is the number that occurs most often. The mode is not unique. A distribution can have more than one mode.
- The **median** is the "middle" of the distribution. If the number of observations n is odd then the median is the center observation in the ordered list. If the number of observations n is even then the median is the arithmetic mean (see below) of the two center observations.
- The **arithmetic mean** $\bar{x}$ is what we will often times refer to simply as the **mean**. To find the arithmetic mean, you add all of the observations and divide the sum by the total number of observations. That is $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$.

Example 1. *Suppose you have taken a poll that asked all of the faculty members of the math department how many times they brush their teeth a day. The responses are as follows:*

$$0, 1, 2, 2, 2, 3, 7$$

- *The mode of this distribution is* 2.
- *The median of this distribution is* 2.
- *The arithmetic mean of this distribution is* $\frac{0+1+2+2+2+3+7}{7} =$ 2.43.

1.2 Exercises

Answer each of the questions below and **fully justify your answer** using complete sentences. If you answer "yes", explain why. If you answer "no", it may be appropriate to show a counterexample, that is, an example of a specific distribution where some property does not hold.

IMPORTANT: Using one example is not enough to show something is ALWAYS true; however, using one example can be enough to show something is NOT ALWAYS true.

1. Give an example of a distribution that has more than one mode.
2. What is the maximum number of modes a particular sample distribution can have?

   For numbers 3–7 below, suppose you are given a sample distribution $x_1, x_2, \ldots, x_n$.
3. Will the mode *always* take a value that is equal to x_i for some i?
4. Will the median *always* take a value that is equal to x_i for some i?
5. Will the arithmetic mean *always* take a value that is equal to x_i for some i?
6. Will the median *always* take a value that is equal to the mode?
7. Will the median *always* take a value that is equal to the arithmetic mean?
8. Suppose 6 students take an exam and the mean score is 80%. Five of the students scores are: 95,78,85,56, 96. What is the sixth student's score?
9. The number 15 is added to each of the biggest 150 numbers in a distribution of 301 numbers. (In other words, $n = 301$).
   (a) How does this addition affect the median of the distribution? [1, 4.2.7]
   (b) How does this addition affect the mode of the distribution?
   (c) How does this addition affect the arithmetic mean of the distribution?
10. Suppose there are n students in a class ($n \geq 4$). Some of the students appealed their scores on a certain test. The papers were reviewed and the scores of four students were raised from a 70 to a 90. How does this change affect the mean score of the class? [1, 4.2.8]

1.3 The Five-Number Summary of a Distribution

Suppose you have the distribution of a quantitative variable listed in order from smallest to largest with median Me.

- The **first quartile** Q_1 is the median of the observations smaller than Me.
- The **third quartile** Q_3 is the median of the observations larger than Me.
- The **five-number summary** of a distribution is a list containing: the smallest observation, Q_1, Me, Q_3, and the largest observation.

Example 2. *Suppose you are teaching a class of 11 third graders and have given them a quiz. Their scores out of 100 are as follows:*

$$55, 70, 70, 72, 85, 88, 90, 90, 92, 95, 98$$

Here $Me = 88, Q_1 = 70$, *and* $Q_3 = 92$.

The five-number summary is: 55, 70, 88, 92, 98.

1.4 Exercises

1. The five-number summary of a sample distribution of exam scores (in percentages) for 100 students is: $0, 0, 92, 94, 100$. Discuss the performance of the 100 students overall. Give as much detail as you can. Be sure to include in your discussion what you still do not know about the students' scores.
2. The 100 students from the previous question retake the exact same exam and the five-number summary of the new scores is: $55, 75, 92, 96, 100$. Discuss the performance of the students overall on this exam.
3. Billy Jo is one of the 100 students who took the exams described above. On the first exam, he got a 26% and on the second he got a 56%. Discuss Billy Jo's performance on each exam based on how his score compares to the rest of the class. If you were responsible for assigning grades to the students, what grade would you give to Billy Jo on each exam respectively?
4. Write down a sample distribution with $n = 20$ so that the five-number summary of the distribution is: $2, 4, 6, 8, 10$.

1.5 Variance and Standard Deviation

A **population** is a group of individuals or subjects that we want to learn something about. When we talk about a sample distribution, we typically consider the list of n numbers $\{x_1, x_2, ..., x_n\}$ as representing *some of* the individuals or subjects from a population.

- The **variance** s^2 of a sample distribution is the mean of the square of the distance each observation is away from the mean $\bar{x}$.

  To compute the variance:

  1. Compute the mean $\bar{x}$ of the distribution.
  2. Compute the distance each observation is away from $\bar{x}$.
  3. Square each of the distances found in 2.
  4. Find the mean of the list found in 3.

  The formal equation for this is:

$$\begin{aligned} s^2 &= \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+...+(x_n-\bar{x})^2}{n} \\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \end{aligned} \tag{1.1}$$

- The **standard deviation** s is the square root of the variance. In other words, $s = \sqrt{s^2}$.

Important Note: We are using the formulae above for variance and standard deviation. They are the "true" formulas. However, the terminology of *variance* and *standard deviation* are also used for what we will call **unbiased** estimator for a population's variance and standard deviation. In this case, the denominator in the formula (1.1) above is $n-1$ rather than n. Be careful. For now, we will use the formula given above but eventually, we will switch to the unbiased version.

Beware: When using software (like Exel, Minitab, or a calculator) to compute variance and standard deviation, they will give you the unbiased version.

Example 3. *Suppose you chart the number of green M&M's in each of 4 bags of M&M's. Your results are as follows:*

$$2, 6, 12, 13$$

- *To compute the variance:*

  1. *The mean* $\bar{x} = 8.25$.

2. *The distance each observation is away from the mean:*

$$\begin{aligned} 2-8.25 &= -6.25 \\ 6-8.25 &= -2.25 \\ 12-8.25 &= 3.75 \\ 13-8.25 &= 4.75 \end{aligned}$$

3. *Squaring each of the distances found in 2:*

$$\begin{aligned} (-6.25)^2 &= 39.0625 \\ (-2.25)^2 &= 5.0625 \\ (3.75)^2 &= 14.0625 \\ (4.75)^2 &= 22.5625 \end{aligned}$$

4. *Finding the mean of the numbers found in 3, we get*

$$s^2 = \frac{39.0625 + 5.0625 + 14.0625 + 22.5625}{4} = 20.1875.$$

- *The standard deviation* $s = \sqrt{20.1875} = 4.493$.

1.6 Notation

We have been using $\bar{x}$ and s to denote the mean and standard deviation of a sample distribution from a population. If we are talking about the mean and standard deviation of the entire population, we use μ and σ for the mean and standard deviation. (Greek letters mu and sigma.) At this point we will not worry too much about the distinction but it will be important later.

1.7 Exercises

1. Suppose you have a sample distribution 2,2, 6, 8, 10. Compute the mean, variance, and standard deviation of this distribution.

2. Students were asked to analyze a set of 50 nonnegative scores, not all of which were identical. The set included exactly three 0 (zero) scores. It also included two nonzero scores which were identical to the mean of all the scores.

   (a) One student decided not to include the three zero scores in his analysis, on the false assumption that zero is not a number. He

correctly calculated the following measures, based upon his altered data set. For each measure, would his calculation increase, decrease, or not change the original measure, or is it impossible to tell? Explain all of your answers.

i. mean
ii. median
iii. mode
iv. range (largest value minus the smallest value)
v. variance

(b) Another student decided not to include the two scores that were identical to the mean, arguing that most measures are based on deviations from the mean, whereas those two scores did not deviate from the mean. How would the measures (i – v above) that this student obtained change in relation to the correct answers of the complete set of scores? Answer by 'increase', 'decrease', 'no change', or 'impossible to know'. Explain your answers. [1, 1.1.3]

3. Write two numbers with mean 8 and variance 4.

4. (Extra credit) Is it possible to write two numbers different from those found in number 3 with mean 8 and variance 4?

5. Write three numbers with mean 5 and variance $\frac{8}{3}$.

6. (Extra credit) Is it possible to find three numbers different from those found in number 5 with mean 5 and variance $\frac{8}{3}$.

7. Suppose you have two numbers with variance 9. What is the range of this set? (Note: the range is the largest value minus the smallest value.)[1, 1.1.4]

8. In an educational research project it is necessary to construct a control group that shares some features with the experimental group. There are 12 children in the experimental group, and the investigator decides that the control group should be the same size.

   The mean and the variance of the variable x in the experimental group are 6.0 and 14.00, respectively. The investigator wishes to construct the control group so that both groups will have the same mean and variance.

   After ten children have been selected for the control group, the mean of their x-values is 5.8 and the variance is 16.36. Two additional children will be selected for the control group.

   What should the x-values of these children be so that the experimental group and the control group have equal means and variances? [1, 1.1.11]

9. Eight people took a test in which one can score only 1, 2, or 3.
   (a) You know that exactly two people scored 1 and that the distribution is *symmetric* about the mean. What is that variance of the set of scores?
   (b) Let the variance of the set be 1. List the eight scores.
   (c) Given that the mean of the scores is 3, what is the standard deviation of the set of scores?[1, 1.1.13]

10. The treasury department is considering several schemes for revising its salary and employment policies for government workers.

    The following three schemes are suggested. Determine, in each case, how the suggested revision would affect each of the following measures:

    (a) Each employee will get a raise of $125 per month.
    (b) The salaries will be increased by 15% across the board.
    (c) The number of employees at each salary level will be decreased to 90% of their original number.
        i. The mean monthly salary in dollars.
        ii. The variance of the monthly salaries.
        iii. The standard deviation of the monthly salaries.
        iv. The median monthly salary.
        v. The modal monthly salary.[1, 1.1.15]

11. The mean salary in a certain plant was $1500, and the standard deviation was $400. A year later each employee got a $100 raise. After another year each employee's salary (including the above mentioned raise) was increased by 20%. What are the mean and standard deviation of the current salaries in dollars? [1, 4.2.17]

12. For those of you who have been ignoring the sigma notation, that is the "Σ's" above, no longer!

    (a) Suppose $x_1 = 1, x_2 = 2, x_3 = 3, ..., x_{10} = 10$. In other words, $x_i = i$ where i takes all integer values from 1 to 10.

        Compute each of the following:
        i. $\sum_{i=1}^{10} i$
        ii. $\sum_{i=1}^{5} i^2$
    (b) Suppose $x_1 = 2, x_2 = 4$ and $x_3 = 7$. Compute $\sum_{i=1}^{3}(x_i - 2)$.
    (c) Suppose $x_i = 2^i$. Compute $\sum_{i=1}^{6}(x_i)^2$.
    (d) Suppose $x_1 = 100, x_2 = 95, x_3 = 82, x_4 = 60$. Compute $\sum_{i=1}^{4}(x_i - 60)^2$.

Chapter 4

Probability

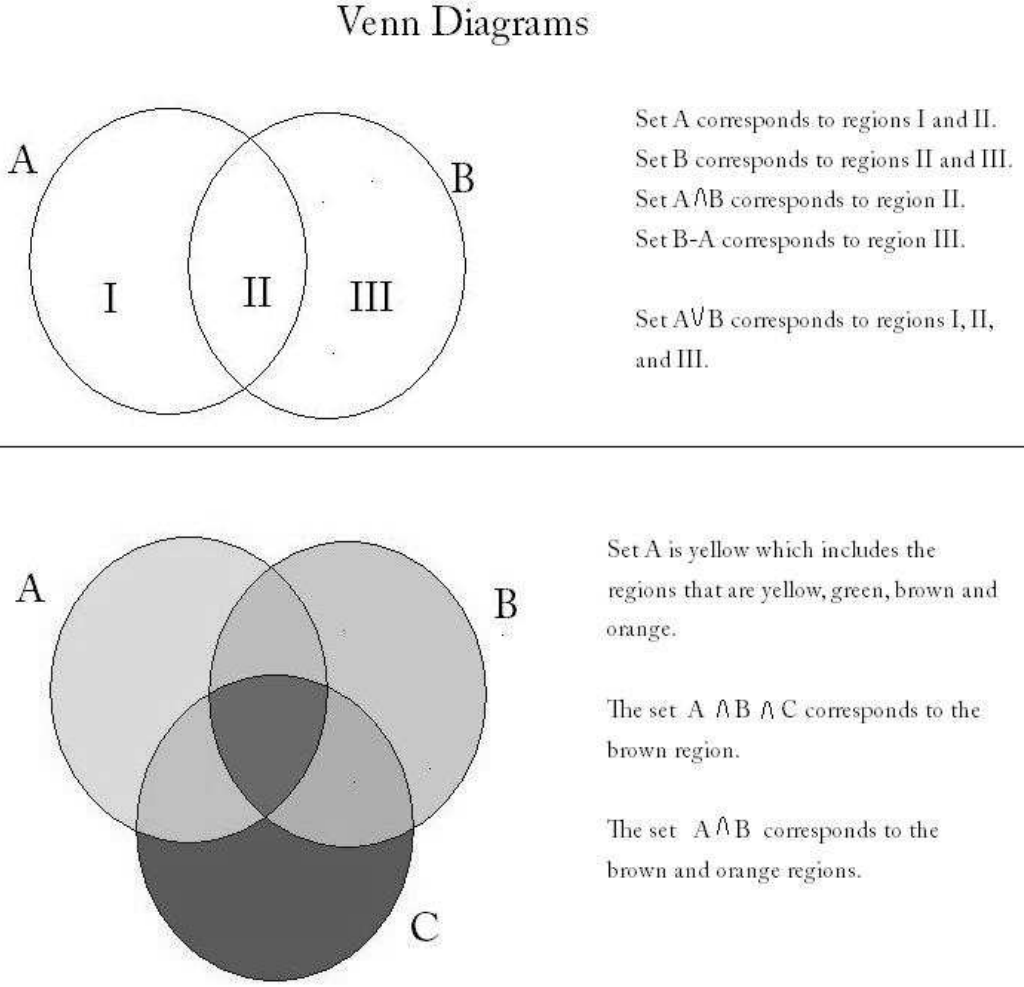
4.1 Set Theory

Here are some terminology and notation:

- $\emptyset$ denotes the empty set or the set that contains no elements. We also write this $\{\}$.
- A, B, and C denote sets.
- Ω denotes the universe of all possible elements in consideration.
- $\overline{A}$ denotes the set consisting of elements that are in Ω and not in the set A. We call this A *complement*.
- $A \cup B$ is the set consisting of all elements in the set A combined with all the elements in set B. We call this A *union* B.
- $A \cap B$ is the set that contains only the elements that are in both A and B. We call this A *intersect* B.
- We denote $A \subseteq B$ to say that "A is a subset of B". This means that every element of A is also an element of set B.
- We write $A - B$ to mean the set containing elements that are in A and not in B. Notice that $\overline{A} = \Omega - A$.
- We say that two sets are **disjoint** if they have no elements in common. In other words, A and B are disjoint if and only if $A \cap B = \emptyset$.

4.2 Exercises

1. Consider the sets $A = \{1,2,3,4,5\}$ and $B = \{2,4,6,8,10\}$ where $\Omega = \{1,2,3,4,5,6,7,8,9,10\}$. Compute each of the following sets:
   (a) $A \cup B$
   (b) $A \cap B$
   (c) $\overline{A}$
   (d) $\overline{B}$
   (e) $B - A$
   (f) $A - B$
   (g) $\overline{A \cup B}$
   (h) $\overline{A} \cup \overline{B}$
   (i) $(B - \overline{A}) \cap \overline{(A \cap B)}$
2. Suppose $\Omega = \{\text{red, orange, yellow, green, blue, indigo, violet}\}, A = \{\text{red}\}$ and $B = \{\text{red, orange, blue}\}$. Compute a–i from question 1 for this example.
3. Write down an example of a specific Ω, A, B, and C so that $A \subseteq B$ and C is disjoint from both A and B.
4. Write down an example of a specific Ω, A, B, and C so that A, B, and C are all disjoint and $A \cup B \cup C = \Omega$.
5. (Extra Credit) Consider the sets $A = \emptyset$ and $B = \{\emptyset\}$ and $C = \{1, \emptyset\}$. Find each of the following sets, if it is possible. If it isn't, state why.
   (a) $C \cup B$
   (b) $A \cap B$
   (c) $\overline{A}$
   (d) $B - A$
   (e) $C - B$



4.3 Venn Diagrams

A **Venn Diagram** is useful in illustrating sets and their relationships to each other. At the top of the page is an example of a Venn diagram with two sets. Below that is an example of a Venn diagram with three sets.

4.4 Exercises

1. Determine whether each of the following is true or false. If you say true, show that the Venn diagram of the left-hand side is the same as the Venn diagram of the right hand side. If you say false, come up with specific sets where the equality does not hold.[1, 2.1.1]
   (a) $\overline{A \cup B} = \overline{A} \cup \overline{B}$
   (b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
   (c) $\overline{A - B} = \overline{A} \cup B$
   (d) $A - \overline{A} = \emptyset$

(e) $\overline{A \cap \overline{A}} = \Omega$

(f) $A = (A \cap B) \cup (A \cap \overline{B})$

2. Let Ω be the set of all students currently enrolled in classes at Susquehanna University. Let A be the set of all students enrolled in intro stats this term, let B be the set of all students who play a varsity sport.

   Suppose there are 2,000 total students enrolled at SU and 120 are enrolled in a section of intro stats this term and 230 play a varsity sport. Note: These numbers are not the official counts.

   Interpret each of the sets below in terms of this example and, if possible, determine how many people are in each set.

   (a) A

   (b) B

   (c) Ω

   (d) $A \cap B$

   (e) $A \cup B$

   (f) $\overline{A}$

   (g) $\overline{B}$

   (h) $\overline{A \cup B}$

   (i) $\overline{A \cap B}$

3. Repeat number 2 with the added information that there are 102 students enrolled in Introductory Statistics this term who do not play a varsity sport.