Graphs, Charts, and Numbers

Once a sample is collected, there is a lot of data that is received. Some organization and management of the data will be required so that we can better understand it.

There are essentially two major strategies for describing a large data set:

1. describe the data using pictures (graphs and charts) and
2. use numerical summaries (averages, medians, quartiles, percentiles or five-number summaries).

Numerical summaries help us to identify where the data set sits in relation to the possible values that the data could take and are called measures of location.

15.1 Graphs and Charts

Let’s start with some basic terminology so that we can discuss different visual representations.

- **data set:**

- **data point:**

- **discrete variable:**

- **continuous variable:**

Bar Graphs, Pictograms, and Line Graphs

To get a understanding of these visual representations, let’s consider an example. (For space concerns, it begins on the next page.)
Example: Math 101 Midterm Scores
The following is a list of 75 scores (between 0 and 25) on a math midterm with their ID numbers provided for student confidentiality.

<table>
<thead>
<tr>
<th>ID</th>
<th>Score</th>
<th>ID</th>
<th>Score</th>
<th>ID</th>
<th>Score</th>
<th>ID</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>12</td>
<td>1235</td>
<td>13</td>
<td>1240</td>
<td>11</td>
<td>1241</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2345</td>
<td>10</td>
<td>2346</td>
<td>11</td>
<td>2351</td>
<td>9</td>
<td>2352</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3234</td>
<td>7</td>
<td>3236</td>
<td>7</td>
<td>3245</td>
<td>11</td>
<td>3246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4234</td>
<td>11</td>
<td>4238</td>
<td>13</td>
<td>4250</td>
<td>11</td>
<td>4251</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5345</td>
<td>13</td>
<td>5349</td>
<td>9</td>
<td>5372</td>
<td>8</td>
<td>5377</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1235</td>
<td>13</td>
<td>1236</td>
<td>9</td>
<td>1245</td>
<td>11</td>
<td>1246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2346</td>
<td>11</td>
<td>2347</td>
<td>12</td>
<td>2352</td>
<td>10</td>
<td>2353</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3234</td>
<td>7</td>
<td>3235</td>
<td>11</td>
<td>3245</td>
<td>11</td>
<td>3246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4234</td>
<td>11</td>
<td>4238</td>
<td>13</td>
<td>4250</td>
<td>11</td>
<td>4251</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5345</td>
<td>13</td>
<td>5349</td>
<td>9</td>
<td>5372</td>
<td>8</td>
<td>5377</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1235</td>
<td>13</td>
<td>1236</td>
<td>9</td>
<td>1245</td>
<td>11</td>
<td>1246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2346</td>
<td>11</td>
<td>2347</td>
<td>12</td>
<td>2352</td>
<td>10</td>
<td>2353</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3234</td>
<td>7</td>
<td>3235</td>
<td>11</td>
<td>3245</td>
<td>11</td>
<td>3246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4234</td>
<td>11</td>
<td>4238</td>
<td>13</td>
<td>4250</td>
<td>11</td>
<td>4251</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5345</td>
<td>13</td>
<td>5349</td>
<td>9</td>
<td>5372</td>
<td>8</td>
<td>5377</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1235</td>
<td>13</td>
<td>1236</td>
<td>9</td>
<td>1245</td>
<td>11</td>
<td>1246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2346</td>
<td>11</td>
<td>2347</td>
<td>12</td>
<td>2352</td>
<td>10</td>
<td>2353</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3234</td>
<td>7</td>
<td>3235</td>
<td>11</td>
<td>3245</td>
<td>11</td>
<td>3246</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4234</td>
<td>11</td>
<td>4238</td>
<td>13</td>
<td>4250</td>
<td>11</td>
<td>4251</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5345</td>
<td>13</td>
<td>5349</td>
<td>9</td>
<td>5372</td>
<td>8</td>
<td>5377</td>
<td>9</td>
</tr>
</tbody>
</table>

A frequency table is a table that lists the data points and how many times they occur in the data set. Create a frequency table for the above data set.

<table>
<thead>
<tr>
<th>Score</th>
<th>ID</th>
<th>Score</th>
<th>ID</th>
<th>Score</th>
<th>ID</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How many different actual values are there? (This number is usually denoted with an \( M \).)

Use this frequency table to make a bar graph:

Frequency charts that use icons or symbolic images instead of bars to display the frequencies are commonly referred to as pictograms. An alternative to the bar graph that eliminated the need for bars is the line graph. In a line graph, instead of bars we use points with adjacent points connected by lines. Line graphs are particularly useful when the number of values plotted is large or when the graph represents a time-series—the horizontal axis represents a time variable.
Categorical Variables and Pie Charts

Variables need not always represent a measurable quantity—variable can also represent non-numerical characteristics of a population such as gender, ethnicity, nationality, emotions, feelings, actions, etc. Variables of this type are called **categorical** variables.

Sometimes a variable starts out as a numerical variable but then, as a matter of convenience, it is converted into a categorical variable. Let’s return to our example.

**Example: Math 101 Midterm Grades**
The midterm score is a numerical variable that can take integer values between 0 and 25. The grading scale for the class is A=18-25, B=14-17, C=11-13, D=9-10 and F=0-8.

Make a frequency chart based on this grading scale.

Make a bar graph and pie chart representing this data.

Continuous Variables and Histograms

When a numerical variable is continuous, its possible values can vary by infinitesimally small increments. As a consequence, there are no gaps between the class intervals (data point ranges), the previous methods will not work. In this case we use a variation of a bar graph called a **histogram**.

Read and discuss **Example 15.7: Starting Salaries of TSU Graduates** on p.457.

What does the plus (i.e. $50^+ - 55$) mean in the histogram?

What is the main difference between a histogram and a bar graph?
15.2 Means, Medians, and Percentiles

One of the most convenient and commonly used devices for understanding data is to use numerical summaries of the data.

The Average

What is an average (mean)?

Compute the average of the scores in Example: Math 101 Midterm Scores.

Percentiles

What is a percentile?

Which is better: scoring in the 25th or 75th percentile on an exam?

There are several different ways to compute percentiles that will satisfy the definition and different statistics books describe different methods—we will discuss only one.

Finding the \( p \)th Percentile of a Data Set

Step 1: Sort the data set from smallest to largest. Let \( d_1, d_2, \ldots, d_N \) represent the sorted data.

Step 2: Find the locator \( L = \) ________________.

Step 3: Depending on whether \( L \) is a whole number or not, the \( p \)th percentile is given by

- \( d_{L,5} = \) ____________________________, if \( L \) is a whole number.
- \( d_{L^+} = \) ____________________________, if \( L \) is not a whole number.

Example: Scholarships by Percentile

To reward good academic performance from its athletes, Tasmania State University has a program in which athletes with GPAs in the 80th or higher percentile of their team’s GPAs get $5000 scholarship and athletes with GPAs in the 55th or higher percentile of their teams GPAs who did not get the $5000 scholarship get a $2000 scholarship.

The women’s volleyball team has \( n = 15 \) players on the roster. A list of their GPAs is as follows:


Which students will receive the $5000 scholarship?
Which students will receive the $2000 scholarship?

The most commonly used set of percentiles are the **quartiles**.

**The Median and the Quartiles**

The 50th percentile of a data set is known as the **median** and denoted by $M$. The median splits the data set into two halves. We can apply the percentile definition to compute it, but the bottom line comes down to this:

**Finding the Median of a Data Set**

- **Step 1**: Sort the data set from smallest to largest. Let $d_1, d_2, \ldots, d_N$ represent the sorted data.
- **Step 2**: Depending on whether $N$ is even or odd, the median is given by
  - $\frac{d_{\frac{N}{2}} + d_{\frac{N}{2} + 1}}{2}$, if $N$ is odd.
  - $d_{\frac{N}{2}}$, if $N$ is even.

Let’s return to our main example to practice computing these values:

Compute the median for the data in **Example: Math 101 Midterm Scores**.

Compute the quartile scores for the data in **Example: Math 101 Midterm Scores**.

**The Five-Number Summary**

A common way to summarize a large data set is by means of its five-number summary. The **five-number summary** is give by:

What is the five-number summary for the **Example: Math 101 Midterm**?
Box Plots

Invented in 1977 by statistician John Turkey, a box plot (a.k.a a box-and-whisker plot) is a picture of the five-number summary. The **box plot** consists of a rectangular box that sits above a number line representing the data values and extends from the first quartile $Q_1$ to the third quartile $Q_3$ on that number line.

\[
\begin{array}{c|c|c|c|c|c}
\text{Min} & Q_1 & M & Q_3 & \text{Max} \\
\end{array}
\]

Make a box-and-whisker plot for **Example: Math 101 Midterm**.

Box plots are extremely useful when comparing similar data for two or more populations.

### 15.3 Ranges and Standard Deviations

There are several different ways to describe the spread of a data set; in this section we will describe the three most commonly used ones.

- **range**:

- **interquartile range (IQR)**:

Compute the range *and* interquartile range for **Example: Math 101 Midterm**.

### The Standard Deviation

This is the most important and most commonly used measure of spread for a data. Before we can discuss it fully, we must address a few finer details.

If $A$ is the average of the data set and $x$ is an arbitrary data value, the difference $\ldots$ is $x$’s *deviation from the mean*.

What does the deviation from the mean tell us?

What kind of values are these?

The idea is to use this information to figure out how spread out the data is. There are still several steps before we get there.
Notice that the deviations from the mean are themselves a data set which we would like to summarize. One way to do that would be to average them. What would one potential problem with this method be?

How might we resolve this problem?

The average of the squared deviations is the **variance** $V$. Finally, taking the square root of the variance we get the **standard deviation** $\sigma$. More explicitly,

Finding the Standard Deviation of a Data Set

Step 1: Let $A$ denote the mean of the data set. For each number $x$ in the data set, compute its deviation from the mean ($x - A$) and square each of these numbers.

Step 2: Find the average of the squared deviations. This number is called the variance $V$.

Step 3: The standard deviation is the square root of the variance ($\sigma = \sqrt{V}$).

1. Compute the standard deviation for the following set of homework scores:  
   \[85, 86, 87, 88, 89, 91, 92, 93, 94, 95\]

2. Compute the standard deviation for the data in Example: Math 101 Midterm Scores.
Some notes about Standard Deviation:

Homework: p.469 # 5, 6, 7, 9, 11, 13, 24, 26, 36, 42, 45-46, 48, 56, 72, 75