2. Introduction to Statistics and Sampling

2.1 Introduction

2.1.1 Definition of statistics, statistic
2.2.2 Why study Statistics?

2.2 Definitions

2.2.1 Population
2.2.2 Sample
2.2.3 Probability
2.2.4 Continuous vs. discrete variables → we will concentrate on continuous variables

2.3 Measures of Central Tendency

2.3.1 Average
   - “Average” means the same as “central tendency,” a measure of the “middle” of a set of data.
   - Three most common measures of central tendency: (arithmetic) mean, median, and mode.

2.3.2 Mean
   - There are many types of means – arithmetic mean, geometric mean, harmonic mean, quadratic mean, to name a few. We will only use the arithmetic mean in this course.
   - The Arithmetic Mean is often referred to as just the mean, and sometimes referred to as the average (although we will not). It is defined as
     \[
     \mu = \frac{1}{N} \sum_{i=1}^{N} x_i \quad \text{(population mean),} \tag{2.1}
     \]
     where \(N\) is the number of measurements in the population.
   - If we are only looking at a sample from the population, the mean is written as
     \[
     \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{(sample mean),} \tag{2.2}
     \]
     where \(n\) is the number of measurements in the sample.

2.3.3 Median
   - Middle value (half of the values are below, half are above)
   - Sometimes given the symbol \(\bar{x}\)

2.3.4 Mode
   - The most frequently occurring value
**Example 2-1:** Determine the (a) mean, (b) median, and (c) mode in the following data: 1, 2, 2, 2, 3, 10.

**Solution:**
(a) You can verify by Equation (2.2) that the mean of the sample is 3.3.
(b) The median is the middle value, which, to be exact, is between the 3rd and 4th data point:

<table>
<thead>
<tr>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2.1. Data of Example 2-1.

It is customary to average the adjacent values (#3 and #4) to estimate the median; in this case, both values are the same, so the median of this set is 2.
(c) The mode is the most frequently occurring value, which in this case is 2.

2.3.5 Why don’t we just always use the arithmetic mean to describe central tendency? (AKA: How to choose an “honest” average)

**2.4 Measures of Spread**

2.4.1 One way to describe the spread of a set of data is to report the maximum ($x_{\text{max}}$), minimum ($x_{\text{min}}$), and range ($R = x_{\text{max}} - x_{\text{min}}$) of the set.
- Disadvantage: These measures are influenced by outliers (extreme values)

2.4.2 Quartiles
- Separates the data into 4 groups
- The first (lower) quartile $Q_1$ is value such that 25% of the data are smaller. In a sample of $n$ measurements, the location of the first quartile can be approximated as *the value corresponding to the 0.25($n+1$) ordered observation.*
- The second quartile $Q_2$ is the point below which 50% of the data fall. It can be approximated as *the value corresponding to the 0.5($n+1$) ordered observation.*
- The third (upper) quartile $Q_3$ is the point below which 75% of the data fall, and can be approximated as *the value corresponding to the 0.75($n+1$) ordered observation.*

- There are a number of more sophisticated ways of estimating quartiles (Excel has a *QUARTILE* function, for example), but the method described here is the simplest and will suffice for most applications.

**Q.** The second quartile is the same as another measure we’ve discussed. What is it?

**Q.** A “forth quartile” is not defined, but if it were, it would be the same as what measure we’ve previously defined?
2.4.3 Interquartile Range

- Interquartile Range is simply defined as

\[ IQR = Q_3 - Q_1 \]  

(2.3)

- Figure 2.1 illustrates quartiles and the interquartile range, and demonstrates their role in representing the spread of a set of data.

![Figure 2.1. Distribution of a set of data, and location of quartiles.](image)

2.4.4 Five-Number Summary

- The five-number summary is simply a listing of the following summary statistics:
  1. Minimum value \( x_{\text{min}} \)
  2. First (or Lower) Quartile \( Q_1 \)
  3. Median \( \tilde{x} \)
  4. Third (or Upper) Quartile \( Q_3 \)
  5. Maximum value \( x_{\text{max}} \)

2.4.5 Box Plot

- A graphical depiction of the Five-Number Summary
- Best illustrated by example

Example 2-2: Determine the five-number summary, and construct a box plot for the following data: 11, 32, 37, 68, 73, 154.

Solution: First, let’s arrange the data into a column, sorted from lowest to highest:

<table>
<thead>
<tr>
<th>No.</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
</tr>
<tr>
<td>5</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>154</td>
</tr>
</tbody>
</table>

Table 2.2. Data of Example 2-2, arranged in a column.
The minimum and maximum values can readily be identified as the values 1 and 10, respectively. The first quartile corresponds to measurement number 0.25(6+1) = 1.75, or somewhere between the values 11 and 32. Although some texts would interpolate between the values, it is usually adequate to simply choose the closest value, which in this case is the second value, 32.

The median is the same as the second quartile, and corresponds to measurement 3.5, exactly halfway between measurement numbers 3 and 4. In this case, it is best to simply average the two corresponding values, which for these data is 52.5.

The third, or upper quartile corresponds to measurement number 5.25, which we will round to 5. The corresponding value is 73.

Finally, a box plot is simply a plot of the five-number summary along a number line. The box has a width equal to the interquartile range (recall that this is the distance between the first and third quartiles), and a line in the middle represents the median value. Two lines on the outside of the box go to the maximum and minimum values in the data set (sometimes referred to as “whiskers”). The box plot for the data of this example is depicted in Figure 2.2 below.

![Box Plot Diagram](image)

**Figure 2.2. Box Plot of data of Example 2-2.**

**Comments:** Why do we locate the median using 0.5(n+1), instead of just 0.5n, which is even simpler? The answer is that 0.5n does not locate the center of the data set. For this data, the formula 0.5n would have located the median’s at position 3, which is not truly in the center of the data set.

- What’s the point of a box plot? It provides a quick interpretation of the data. The box represents the middle 50% of the data, giving a glimpse of the central tendency of the data set. The position of the median within the box gives an indication of how skewed the data are. Finally, the “whiskers” not only show the total range of the data, but also give some indication of the shape of the distribution of values. For example, the box plot below indicates a strongly skewed distribution, since the median is very close to one side of the box. The plot also suggests that the minimum value in the set could be an “outlier,” a perhaps erroneous data point.
2.4.6 Variance and Standard Deviation

- We want to develop a single value that represents the spread of a population of data. One way to do this could be to sum up the distance from the mean of every data point:
  \[
  \sum_{i=1}^{N} |x_i - \mu|
  \]
  where we sum the absolute values of the distances \(x_i - \mu\), because otherwise they would sum to zero – the negative distances will cancel the positive ones.
- Another way to define this deviation is to sum the squares of the distances, which has a similar effect as taking the absolute value.
  \[
  \sum_{i=1}^{N} (x_i - \mu)^2
  \]
- If we divide the above by \(N\), we obtain an average squared distance from the mean, which we call the **population variance**:
  \[
  \text{variance} = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.
  \] (2.4)
- The square root of the variance is called the **population standard deviation**,\[
  \sigma = \left[ \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \right]^{1/2} \text{ (population standard deviation).} \] (2.5)
- The population variance is given the symbol \(\sigma^2\).
- The standard deviation for a sample is similar, with the sample mean \(\bar{x}\) replacing \(\mu\), and the sample size \(n\) replacing \(N\):
  \[
  s = \left[ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} \text{ (sample standard deviation).} \] (2.6)
- The **sample variance** is \(s^2\).
- Again, the standard deviation can be considered a measure of the average distance an individual measurement is away from the mean value.
Again, we sum the squares, \((x_i - \bar{x})^2\), instead of just \((x_i - \bar{x})\), because the sum of \((x_i - \bar{x})\), even for a sample, is zero. The proof is a homework exercise.

**QUESTION:** If the standard deviation is the average distance the measurements are from the mean, why is \((n-1)\) used in the equation for the sample standard deviation \(s\), instead of simply using \(n\) like when we calculate the population standard deviation (Equation 2.5)?

**ANSWER:** If you look at the definition of population standard deviation, \(\sigma\), it makes sense that we divide the sum \(\sum_{i=1}^{N} (x_i - \mu)^2\) by \(N\), implying that the standard deviation has something to do with the average distance that any measurement \(x_i\) is from the mean \(\mu\). But if we use simply \(n\) in Equation (2.6), the resulting value of \(s\) is slightly incorrect.

What do we mean by this? Consider the sample mean, \(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\). Let’s say we take ten measurements from a population and calculate the mean of that sample. Then we take another ten, and another ten, and so on. As the number of samples of ten becomes large, we would find that the average of their means would equal the population mean. In other words, on average, the sample mean gives the same result as the population mean. Thus the sample mean \(\bar{x}\) is said to be an **unbiased estimator** of \(\mu\).

Now consider the sample standard deviation \(s\). If we defined \(s\) exactly as we defined \(\sigma\), we would find that, on average, \(s\) would not exactly equal \(\sigma\). In other words, the calculation of sample standard deviation would be slightly **biased**. The proof is beyond the scope of this course, but it turns out that defining the sample standard deviation \(s\) as we did in Equation (2.6) yields the desired result; that is, an **unbiased estimator** of \(\sigma\).

### 2.5 Other Statistical Measures

Other descriptive parameters include:

- Skewness (lack of symmetry)

\[
skewness = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}
\]

(2.7)

- Kurtosis ("peakedness")

\[
kurtosis = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4}
\]

(2.8)
2.6 Graphical Representation of a Sample: Histograms

- We will introduce the following concepts by example.

Example 2-3: Twenty ball bearings are taken off an assembly line, and their diameters measured, with the resulting values below.

Table 2.3. Ball bearing diameters (in.).

<table>
<thead>
<tr>
<th>Diameter</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>1.11</td>
<td></td>
</tr>
<tr>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

2.6.1 Frequency table

- $F_j$ = number of occurrences within a given range (the measurement’s frequency)
- $n$ = number of data points in sample
- $\Delta x_j$ = interval width
- Number of intervals (or bins) to use: Usually between 5 and 20. The right number depends on the number and spread of your data: too few bins, and you cannot see details of the distribution. Too many, and the details obscure the general trends you are trying to illustrate. Choosing the number of bins is actually more an art than a science, although much research has been done to try to develop a formula that can be used by statistical programs (to remove as much thinking as possible on the part of the user.) As a starting point, several equations exist to make a first guess. Figliola and Beasley suggest

$$k \approx 1.87(n - 1)^{0.40} + 1,$$

which seems to work well for small samples, say, below 50. Montgomery and Runger suggest

$$k \approx \sqrt{n},$$

which seems to work well with larger samples. Any method you use is just a rule of thumb, and you will likely need to adjust $k$ to make the data presentation as meaningful as possible.

In this example, $n = 20$, so we use Eq. (2.9) to calculate $k$ to be 7.072 (or 7, choosing the closest integer).

- How do we choose the interval width, $\Delta x_j$? Use

$$\Delta x_j = \frac{x_{\text{max}} - x_{\text{min}}}{k}.$$  

(2.11)

- In this example,
\[
\Delta x_j = \frac{x_{\text{max}} - x_{\text{min}}}{k} = \frac{1.34 - 0.68}{7.072} = 0.0933
\]

- We want to choose a simple bin width, one that could be read easily from a plot. We therefore choose 0.10 as the bin width.
- Knowing the bin width, we can now assign measurements to bins in a table.

2.6.2 Histogram

- A histogram is a frequency distribution in graphical form.
- Call \( F_j \) (defined above) the frequency.
- Can also plot a histogram of the relative frequency, which is the frequency divided by the total number of samples:

\[
f_j = \frac{F_j}{n} \quad (2.12)
\]

- Note that the relative frequency is the same as the probability.
- Note also that

\[
\sum_{j=1}^{k} f_j = 1 \quad (2.13)
\]

Figure 2.4. Histogram of ball bearing diameters from Example 2-3.

2.6.3 Cumulative Frequency or Relative Frequency Plots

- A plot in which each bar represents the total observation count to that point.
2.6.4 More definitions

- **Percentile**: the point below which a certain percentage of the data lie (Example, in Figure 2.5, a value of 1.15 inches is the 80th percentile, since 80% of the measurements are below 1.15 in.)

- **Fractile (or Quantile)**: the point below which a stated fraction of the measurements lie (1.15 inches is at the 0.80-fractile)

**Example 2-4**: Construct a five-number summary and box plot for the data of Example 2-3.

**Solution**: First, we arrange the data in ascending order and number them, as shown in Table 2.4. We can then identify the maximum and minimum values immediately; the quartiles are then calculated as shown in Table 2.4. Note that since there is an even number of data, there are two median values, data numbers 10 and 11. If these values were different, the average of these two values would represent the median. This is in fact done for the lower and upper quartile values.
Table 2.4. Data of Example 2-1, arranged in ascending order.

<table>
<thead>
<tr>
<th>number</th>
<th>diameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.68</td>
</tr>
<tr>
<td>2</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
</tr>
<tr>
<td>5</td>
<td>0.94</td>
</tr>
<tr>
<td>6</td>
<td>0.96</td>
</tr>
<tr>
<td>7</td>
<td>0.96</td>
</tr>
<tr>
<td>8</td>
<td>0.98</td>
</tr>
<tr>
<td>9</td>
<td>0.99</td>
</tr>
<tr>
<td>10</td>
<td>1.02</td>
</tr>
<tr>
<td>11</td>
<td>1.02</td>
</tr>
<tr>
<td>12</td>
<td>1.04</td>
</tr>
<tr>
<td>13</td>
<td>1.06</td>
</tr>
<tr>
<td>14</td>
<td>1.07</td>
</tr>
<tr>
<td>15</td>
<td>1.08</td>
</tr>
<tr>
<td>16</td>
<td>1.11</td>
</tr>
<tr>
<td>17</td>
<td>1.16</td>
</tr>
<tr>
<td>18</td>
<td>1.21</td>
</tr>
<tr>
<td>19</td>
<td>1.26</td>
</tr>
<tr>
<td>20</td>
<td>1.34</td>
</tr>
</tbody>
</table>

- Now, we can create a box plot of the data set as shown in Figure 2.6.

![Box Plot](image)

**Figure 2.6.** Box plot of the data of Example 2-4.

- As you can see from Figure 2.4, a box plot gives a quick visual “feel” for the data set, much like a histogram, but with additional information.
2.6.5 A note about histograms and box plots: they only depict overall sample behavior. You can (and probably should) plot the individual measurements in sequence, as shown below. The data of Example 2-3 in fact show a general downward trend, which in this case may indicate some kind of drift in the manufacturing process.

![Sequential plot of measured ball bearing diameters from Example 2-3.](image)

Figure 2.7. Sequential plot of measured ball bearing diameters from Example 2-3.

References


Homework

2-1 Consider the following table of data.

<table>
<thead>
<tr>
<th>d [mm]</th>
<th>d [mm]</th>
<th>d [mm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.90</td>
<td>4.32</td>
<td>4.60</td>
</tr>
<tr>
<td>5.02</td>
<td>4.98</td>
<td>4.71</td>
</tr>
<tr>
<td>4.90</td>
<td>4.89</td>
<td>4.78</td>
</tr>
<tr>
<td>5.10</td>
<td>5.01</td>
<td>4.32</td>
</tr>
<tr>
<td>5.02</td>
<td>4.18</td>
<td>4.77</td>
</tr>
<tr>
<td>4.74</td>
<td>5.00</td>
<td>4.85</td>
</tr>
<tr>
<td>4.78</td>
<td>4.66</td>
<td>5.12</td>
</tr>
<tr>
<td>4.94</td>
<td>4.74</td>
<td>5.09</td>
</tr>
<tr>
<td>4.98</td>
<td>4.92</td>
<td>5.20</td>
</tr>
<tr>
<td>4.50</td>
<td>5.01</td>
<td>5.20</td>
</tr>
</tbody>
</table>
a. Develop a frequency table for the data.
b. Draw a frequency distribution for the data. You may use Excel to do this.
c. Create a sequential plot of the data.

2-2 For the data of Problem 2-1, and the intervals determined there, construct a cumulative frequency plot.

2-3 Estimate the mean value of the data represented in the probability (relative frequency) distribution below.

2-4 The probability distribution below is missing one interval value, between 11.0 and 11.5.

a. From the plot, determine the probability of a value occurring between 11.0 and 11.5.
b. Determine the probability of a value occurring between 10.0 and 11.5.

2-5 For the cumulative frequency distribution shown below,

a. Estimate the probability of measurements being less than or equal to 40.
b. Find how many measurements occurred between $36 < x \leq 41$?
c. What value approximates the 2nd quartile?
d. What values approximate the following percentiles: $10^{th}$, $70^{th}$, $90^{th}$?

2-6 Calculate the mean and standard deviation for the following data sets. Calculate these by hand.
   a. (-10.0, 5.0, 15.0)
   b. (-23.5, 26.5)
   c. (1, 2)

2-7 Calculate the mean and standard deviation of the data set of Problem 2-1,
   a. Using Excel, but not using the embedded statistical functions AVERAGE() and STDEV().
   b. Using Excel, and the embedded functions listed in part a. Do the calculations match?

2-8 Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. 