Chapter 1
Introduction

Many people were at first surprised at my using the new words “Statistics” and “Statistical,” as it was supposed that some term in our own language might have expressed the same meaning. But in the course of a very extensive tour through the northern parts of Europe, which I happened to take in 1786, I found that in Germany they were engaged in a species of political inquiry to which they had given the name of “Statistics”… I resolved on adopting it, and I hope that it is now completely naturalised and incorporated with our language.

– Sinclair, 1791; Vol XX

**WHAT IS COVERED IN THIS CHAPTER**

- What is the subject of statistics?
- Population, sample, data
- Appetizer examples

The problems confronting health professionals today often involve fundamental aspects of device and system analysis, and their design and application. As such they are of extreme importance to engineers and scientists.

Due to many aspects of engineering and scientific practice involving nondeterministic outcomes, understanding and knowledge of statistics is important to any engineer and scientist. Statistics is a guide to the unknown. It is a science that deals with designing experimental protocols; collecting, summarizing, and presenting data; and, most important, making inferences and aiding decisions in the presence of variability and uncertainty.
For example, R. A. Fisher’s 1943 elucidation of the human blood-group system Rhesus in terms of the three linked loci C, D, and E, as described in Fisher (1947) or Edwards (2007), is a brilliant example of building a coherent structure of new knowledge guided by a statistical analysis of available experimental data.

The uncertainty that statistical science addresses derives mainly from two sources: (1) from observing only a part of an existing, fixed, but large population or (2) from having a process that results in nondeterministic outcomes. At least a part of the process needs to be either a black box or inherently stochastic, so the outcomes cannot be predicted with certainty.

A population is a statistical universe. It is defined as a collection of existing attributes of some natural phenomenon or a collection of potential attributes when a process is involved. In the case of a process, the underlying population is called hypothetical, for obvious reasons. Thus, populations can be either finite or infinite. A subset of a population selected by some relevant criteria is called a subpopulation.

Often we think about a population as an assembly of people, animals, items, events, times, etc., in which the attribute of interest is measurable. For example, the population of all US citizens older than 21 is an example of a population for which many attributes can be assessed. Attributes might be a history of heart disease, weight, political affiliation, level of blood sugar, etc.

A sample is an observed part of a population. Selection of a sample is a rich methodology in itself, but, unless otherwise specified, it is assumed that the sample is selected at random. The randomness ensures that the sample is representative of its population.

The sampling process depends on the nature of the problem and the population. For example, a sample may be obtained via a retrospective study (usually existing historical outcomes over some period of time), an observational study (an observer monitors the process or population in real time), a sample survey (a researcher administers a questionnaire to measure the characteristics and/or attitudes of subjects), or a designed study (a researcher makes deliberate changes in controllable variables to induce a cause/effect relationship), to name just a few.

Example 1.1. Ohm’s Law Measurements. A student constructed a simple electric circuit in which the resistance R and voltage E were controllable. The output of interest is current I, and according to Ohm’s law it is

\[ I = \frac{E}{R}. \]

This is a mechanistic, theoretical model. In a finite number of measurements under an identical R,E setting, the measured current varies. The population here is hypothetical – an infinite collection of all potentially obtainable measurements of its attribute, current I. The observed sample is finite. In the presence of sample variability, one establishes an empirical
(statistical) model for currents from the population as either

\[ I = \frac{E}{R} + \epsilon \quad \text{or} \quad I = \epsilon \frac{E}{R}. \]

On the basis of a sample, one may first select the model and then proceed with the inference about the nature of the discrepancy, \( \epsilon \).

**Example 1.2. Cell Counts.** In a quantitative engineering physiology laboratory, a team of four students was asked to make a LabVIEW© program to automatically count MC3T3-E1 cells in a hemocytometer (Fig. 1.1). This automatic count was to be compared with the manual count collected through an inverted bright field microscope. The manual count is considered the gold standard.

The experiment consisted of placing 10 \( \mu \)L of cell solutions at two levels of cell confluency: 20\% and 70\%. There were \( n_1 = 12 \) pairs of measurements (automatic and manual counts) at 20\% and \( n_2 = 10 \) pairs at 70\%, as in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Automated</th>
<th></th>
<th>Manual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20% confluency</td>
<td>34 44 40 62 53 51 30 33 38 51 26 48</td>
<td></td>
<td>30 43 34 53 49 39 37 42 30 50 35 54</td>
<td></td>
</tr>
<tr>
<td>70% confluency</td>
<td>72 82 100 94 83 94 73 87 107 102</td>
<td></td>
<td>76 51 92 77 74 81 72 87 100 104</td>
<td></td>
</tr>
</tbody>
</table>

The students wish to answer the following questions:

(a) Are the automated and manual counts significantly different for a fixed confluency level? What are the confidence intervals for the population differences if normality of the measurements is assumed?

(b) If the difference between automated and manual counts constitutes an error, are the errors comparable for the two confluency levels?

We will revisit this example later in the book (Exercise 10.20) and see that for the 20\% confluency level there is no significant difference between the automated and manual counts, whereas for the 70\% level the difference is significant. We will also see that the errors for the two confluency levels...
significantly differ. The statistical design for comparison of errors is called a difference in differences (DiD) and is quite common in biomedical data analysis.

**Example 1.3. Rana Pipiens.** Students in a quantitative engineering physiology laboratory were asked to expose the gastrocnemius muscle of the northern leopard frog (*Rana pipiens*), and stimulate the sciatic nerve to observe contractions in the skeletal muscle. Students were interested in modeling the length–tension relationship. The force used was the active force, calculated by subtracting the measured passive force (no stimulation) from the total force (with stimulation).

The active force represents the dependent variable. The length of the muscle begins at 35 mm and stretches in increments of 0.5 mm, until a maximum length of 42.5 mm is achieved. The velocity at which the muscle was stretched was held constant at 0.5 mm/s.

<table>
<thead>
<tr>
<th>Reading</th>
<th>Change in Length (in %)</th>
<th>Passive force</th>
<th>Total force</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.4</td>
<td>0.012</td>
<td>0.366</td>
</tr>
<tr>
<td>2</td>
<td>2.9</td>
<td>0.031</td>
<td>0.498</td>
</tr>
<tr>
<td>3</td>
<td>4.3</td>
<td>0.040</td>
<td>0.560</td>
</tr>
<tr>
<td>4</td>
<td>5.7</td>
<td>0.050</td>
<td>0.653</td>
</tr>
<tr>
<td>5</td>
<td>7.1</td>
<td>0.061</td>
<td>0.656</td>
</tr>
<tr>
<td>6</td>
<td>8.6</td>
<td>0.072</td>
<td>0.740</td>
</tr>
<tr>
<td>7</td>
<td>10.0</td>
<td>0.085</td>
<td>0.865</td>
</tr>
<tr>
<td>8</td>
<td>11.4</td>
<td>0.100</td>
<td>0.898</td>
</tr>
<tr>
<td>9</td>
<td>12.9</td>
<td>0.128</td>
<td>0.959</td>
</tr>
<tr>
<td>10</td>
<td>14.3</td>
<td>0.164</td>
<td>0.994</td>
</tr>
<tr>
<td>11</td>
<td>15.7</td>
<td>0.223</td>
<td>0.955</td>
</tr>
<tr>
<td>12</td>
<td>17.1</td>
<td>0.315</td>
<td>1.019</td>
</tr>
<tr>
<td>13</td>
<td>18.6</td>
<td>0.411</td>
<td>0.895</td>
</tr>
<tr>
<td>14</td>
<td>20.0</td>
<td>0.569</td>
<td>0.900</td>
</tr>
<tr>
<td>15</td>
<td>21.4</td>
<td>0.751</td>
<td>0.905</td>
</tr>
</tbody>
</table>

The correlation between the active force and the percent change in length from 35 mm is –0.0941. Why is this correlation so low?

For example, one possible model can be found using linear regression (least squares):

\[
\hat{F} = 0.0618 + 0.2084 \cdot \delta - 0.0163 \cdot \delta^2 + 0.0003 \cdot \delta^3
- 0.1732 \cdot \sin \left( \frac{\delta}{3} \right) + 0.1242 \cdot \cos \left( \frac{\delta}{3} \right),
\]

where \( \hat{F} \) is the fitted active force and \( \delta \) is the percent change. This model is nonlinear in variables but linear in coefficients, and standard linear regression methodology is applicable (Chapter [14]). The model achieves a coefficient of determination of \( R^2 = 87.16\% \).

A plot of the original data with superimposed model fit is shown in Figure [12a]. Figure [12b] shows the residuals \( F - \hat{F} \) plotted against \( \delta \).
Suppose that students are interested in estimating the active force for a change of 12%. The model prediction for $\delta = 12$ is 0.8183, with a 95% confidence interval of $[0.7867, 0.8498]$.

**Example 1.4. The 1954 Polio Vaccine Trial.** One of the largest and most publicized public health experiments was performed in 1954 when the benefits of the Salk vaccine for preventing paralytic poliomyelitis was assessed. To ensure that there was no bias in conducting and reporting, the trial was blind to doctors and patients. In boxes of 50 vials, 25 had active vaccines and 25 were placebo. Only the numerical code known to researchers distinguished the well-mixed vials in the box. The clinical trial involved a large number of first-, second-, and third-graders in the United States.

The results were convincing. While the numbers of children assigned to active vaccine and placebo were approximately equal, the incidence of polio in the active group was almost four times lower than that in the placebo group.

\[
\begin{array}{l|c|c}
 & \text{Inoculated with} & \text{Inoculated with} \\
 & \text{vaccine} & \text{placebo} \\
\hline
\text{Total number of children inoculated} & 200,745 & 201,229 \\
\text{Number of cases of paralytic polio} & 33 & 115 \\
\end{array}
\]

On the basis of this trial, health officials recommended that every child be vaccinated. Since the time of this clinical trial, the vaccine has improved; Salk’s vaccine was replaced by the superior Sabin preparation and polio is now virtually unknown in the United States. A complete account of this clinical trial can be found in Francis et al.’s (1955) article or Paul Meier’s essay in a popular book by Tanur et al. (1972).
The numbers are convincing, but was it possible that an ineffective vaccine produced such a result by chance?

In this example there are two hypothetical populations. The first consists of all first-, second-, and third-graders in the United States who would be inoculated with the active vaccine. The second population consists of US children of the same age who would receive the placebo. The attribute of interest is the presence/absence of paralytic polio. There are two samples from the two populations. Randomness of the samples was ensured by randomization of vials in the boxes and random selection of geographic regions for schools. Further analysis of this data can be found in Examples 10.17 and 18.11.

The term *statistics* has a plural form but is used in the singular when it relates to methodology. To avoid confusion, we note that *statistics* has another meaning and use. Any sample summary will be called a *statistic*. For example, a sample mean is a statistic, and sample mean and sample range are statistics. In this context, statistics is used in the plural.

The ultimate summary for quantifying a population attribute is a statistical model. The statistical model term is used in a broad sense here, but a component quantifying inherent uncertainty is always present. For example, random variables, discussed in Chapter 5, can be interpreted as basic statistical models when they model realizations of the attributes in a sample. The model is often indexed by one, several, or sometimes even an infinite number of unknown parameters. An inference about the model translates to an inference about its parameters.

Data are the specific values pertaining to a population attribute recorded from a sample. Often, the terms sample and data are used interchangeably. The term *data* is used as both singular and plural. The singular mode relates to a set, a collection of observations, while the plural is used when referring to the observations. A single observation is called a *datum*.

The following table summarizes the fundamental statistical notions that we discussed:

<table>
<thead>
<tr>
<th><strong>attribute</strong></th>
<th>Quantitative or qualitative property, feature(s) of interest</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>population</strong></td>
<td>Statistical universe; an existing or hypothetical totality of attributes</td>
</tr>
<tr>
<td><strong>sample</strong></td>
<td>A subset of a population</td>
</tr>
<tr>
<td><strong>data</strong></td>
<td>Recorded values/realizations of an attribute in a sample</td>
</tr>
<tr>
<td><strong>statistical model</strong></td>
<td>Mathematical description of a population attribute that incorporates incomplete information, variability, and the nondeterministic nature of the population</td>
</tr>
<tr>
<td><strong>population parameter</strong></td>
<td>A component (possibly multidimensional) in a statistical model; the models are typically specified up to a parameter that is left unknown</td>
</tr>
</tbody>
</table>


