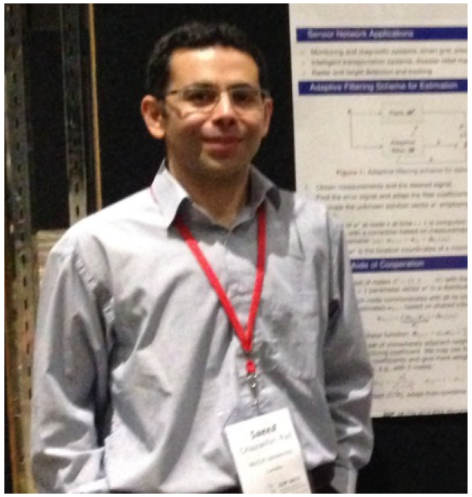
Probability and Statistics for Engineering



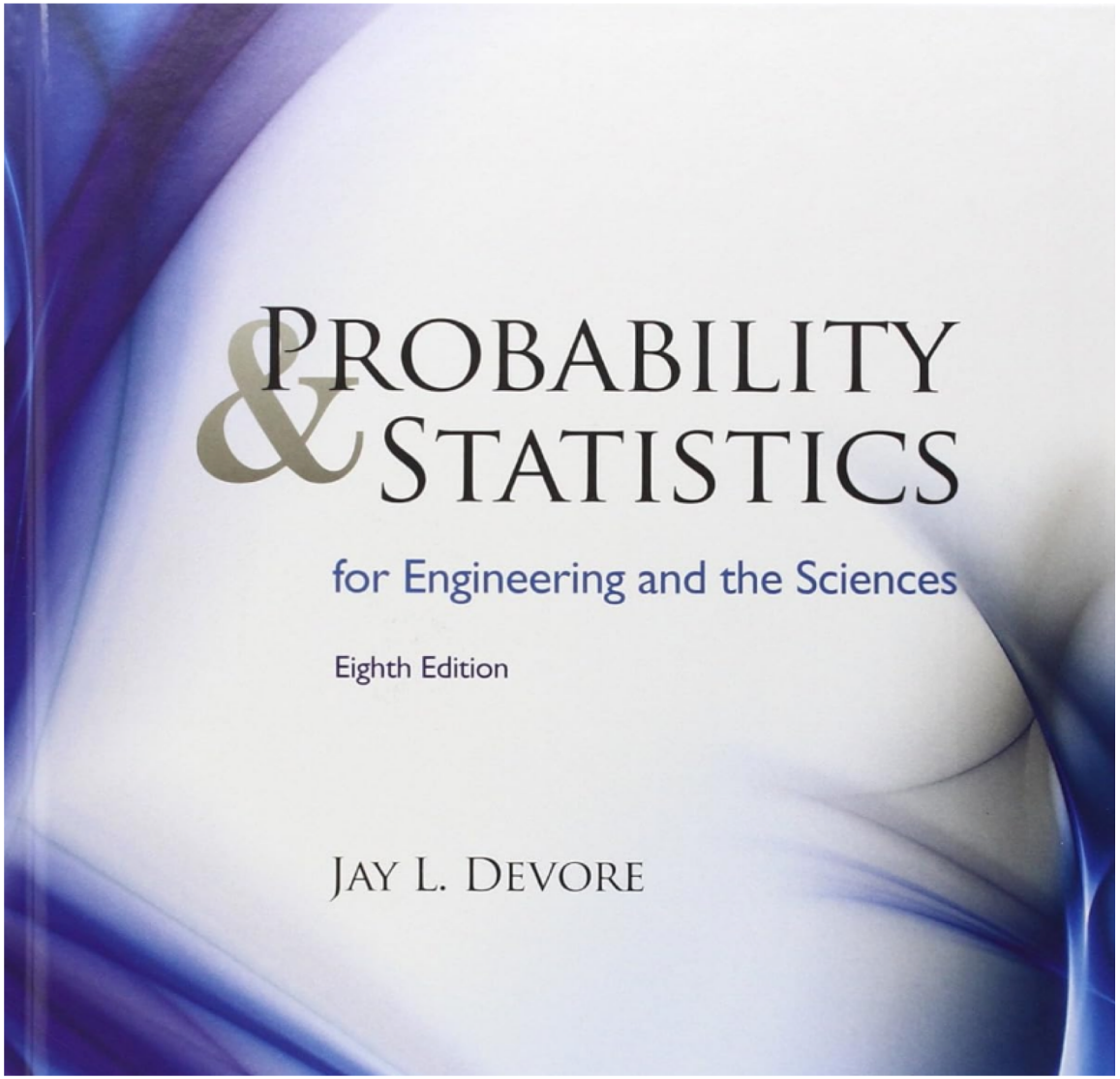
Prof. Saeed Ghazanfari-Rad

Email: grad.saeed@gmail.com

Github: github.com/saeedgrad

References

- **Probability and Statistics for Engineering and the Sciences,** J. Devore, (Cengage Learning).
- **Probability, Random Variables, and Stochastic Processes**, A. Papoulis, S. U. Pillai, (McGraw-Hill).
- **Probability, Statistics, and Random Processes For Electrical Engineering,** A. Leon-Garcia, (Pearson).
- **Introduction to Probability and Statistics for Engineers and Scientists,** S. M. Ross, (Academic Press).


PROBABILITY
&
STATISTICS
for Engineering and the Sciences
Eighth Edition
JAY L. DEVORE

1 Overview and Descriptive Statistics

Contents

Introduction

The discipline of statistics teaches us how to make **intelligent** judgments and **informed decisions** in the presence of **uncertainty and variation**.

If every **component of a particular type** had exactly the same **lifetim**e, if all **resistors** produced by a certain manufacturer had the same resistance **value**, if **pH determination**s for soil specimens from a particular locale gave **identical** results, and so on, then a single observation would reveal all desired information.

Introduction

An interesting manifestation of variation appeared in connection with determining the “greenest” way to travel. The article “Carbon Conundrum” (Consumer Reports, 2008: 9) identified organizations that help consumers calculate carbon output. The following results on output for a flight from New York to Los Angeles were reported:

Introduction

Carbon Calculator	CO_2 (lb)
Terra Pass	1924
Conservation International	3000
Cool It	3049
World Resources Institute/Safe Climate	3163
National Wildlife Federation	3465
Sustainable Travel International	3577
Native Energy	3960
Environmental Defense	4000
Carbonfund.org	4820
The Climate Trust/CarbonCounter.org	5860
Bonneville Environmental Foundation	6732

How to gather information and draw conclusions?

Suppose, for example, that a materials engineer has developed a coating for retarding corrosion in metal pipe under specified circumstances.

If this coating is applied to different segments of pipe, variation in environmental conditions and in the segments themselves will result in more substantial corrosion on some segments than on others.

Methods of statistical analysis could be used on data from such an experiment to decide whether the average amount of corrosion exceeds an upper specification limit of some sort or to predict how much corrosion will occur on a single piece of pipe.

How to gather information and draw conclusions?

suppose the engineer has developed the coating in the belief that it will be superior to the currently used coating.

A **comparative experiment** could be carried out to investigate this issue by applying the current coating to some segments of pipe and the new coating to other segments. This must be done with care lest the wrong conclusion emerge.

The investigator would likely observe a difference between the two coatings attributable not to the coatings themselves, but just to extraneous **variation**.

1.1 Populations, Samples, and Processes

Populations, Samples, and Processes

Engineers and scientists are constantly exposed to collections of facts, or **data,** both in their professional capacities and in everyday activities.

The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest.

Populations, Samples, and Processes

In one study, the population might consist of all gelatin capsules of a particular type produced during a specified period. Another investigation might involve the population consisting of all individuals who received a B.S. in engineering during the most recent academic year.

When desired information is available for all objects in the population, we have what is called a **census.**

Constraints on time, money, and other scarce resources usually make a census impractical or infeasible. Instead, a subset of the population—**a sample**—is selected in some prescribed manner.

Populations, Samples, and Processes

Thus we might obtain a sample of bearings from a particular production run as a basis for investigating whether bearings are conforming to manufacturing specifications, or we might select a sample of last year's engineering graduates to obtain feedback about the quality of the engineering curricula.

We are usually interested only in certain characteristics of the objects in a population: for example, the number of flaws on the surface of each casing, the thickness of each capsule wall, the gender of an engineering graduate, the age at which the individual graduated, and so on.

Populations, Samples, and Processes

A characteristic may be categorical, such as gender or type of malfunction, or it may be numerical in nature.

In the former case, the *value* of the characteristic is a category (e.g. female or insufficient solder), whereas in the latter case, the value is a number (e.g., age = 23 or diameter = .502 cm).

Populations, Samples, and Processes

A **variable** is any characteristic whose value may change from one object to another in the population. We shall initially denote variables by lowercase letters from the end of our alphabet. Examples include

x = brand of calculator owned by a student (**categorical**)

y = number of visits to a particular Web site during a specified period (**numerical, discrete**)

z = braking distance of an automobile under specified conditions (**numerical, continuous**)

Populations, Samples, and Processes

Data results from making observations either on a single variable or simultaneously on two or more variables.
A **univariate** data set consists of observations on a single variable.

For example, we might determine the type of transmission, automatic (A) or manual (M), on each of ten automobiles recently purchased at a certain dealership, resulting in the categorical data set

M A A A M A A M A A

Populations, Samples, and Processes

The following sample of lifetimes (hours) of brand D batteries put to a certain use is a numerical univariate data set:

5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5

We have **bivariate** data when observations are made on each of two variables. Our data set might consist of a (height, weight) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on.

Populations, Samples, and Processes

If an engineer determines the value of both x = component lifetime and y = reason for component failure, the resulting data set is bivariate with one variable numerical and the other categorical.

Multivariate data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

For example, a research physician might determine the systolic blood pressure, diastolic blood pressure, and serum cholesterol level for each patient participating in a study.

Populations, Samples, and Processes

Each observation would be a triple of numbers, such as (120, 80, 146). In many multivariate data sets, some variables are numerical and others are categorical.

Thus the annual automobile issue of *Consumer Reports* gives values of such variables as type of vehicle (small, sporty, compact, mid-size, large), city fuel efficiency (mpg), highway fuel efficiency (mpg), drivetrain type (rear wheel, front wheel, four wheel), and so on.

Branches of Statistics

Branches of Statistics

An investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from **descriptive statistics.**

Some of these methods are graphical in nature; the construction of histograms, boxplots, and scatter plots are primary examples.

Other descriptive methods involve calculation of numerical summary measures, such as means, standard deviations, and correlation coefficients. The wide availability of statistical computer software packages has made these tasks much easier to carry out than they used to be.

Branches of Statistics

Computers are much more efficient than human beings at calculation and the creation of pictures (once they have received appropriate instructions from the user!).

This means that the investigator doesn't have to expend much effort on "grunt work" and will have more time to study the data and extract important messages.

Throughout this book, we will present output from various packages such as Minitab, SAS, S-Plus, and R. The R software can be downloaded for free from the site **http://www.r-project.org**.

Example 1

cont'd

Here is data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Example 1

cont'd

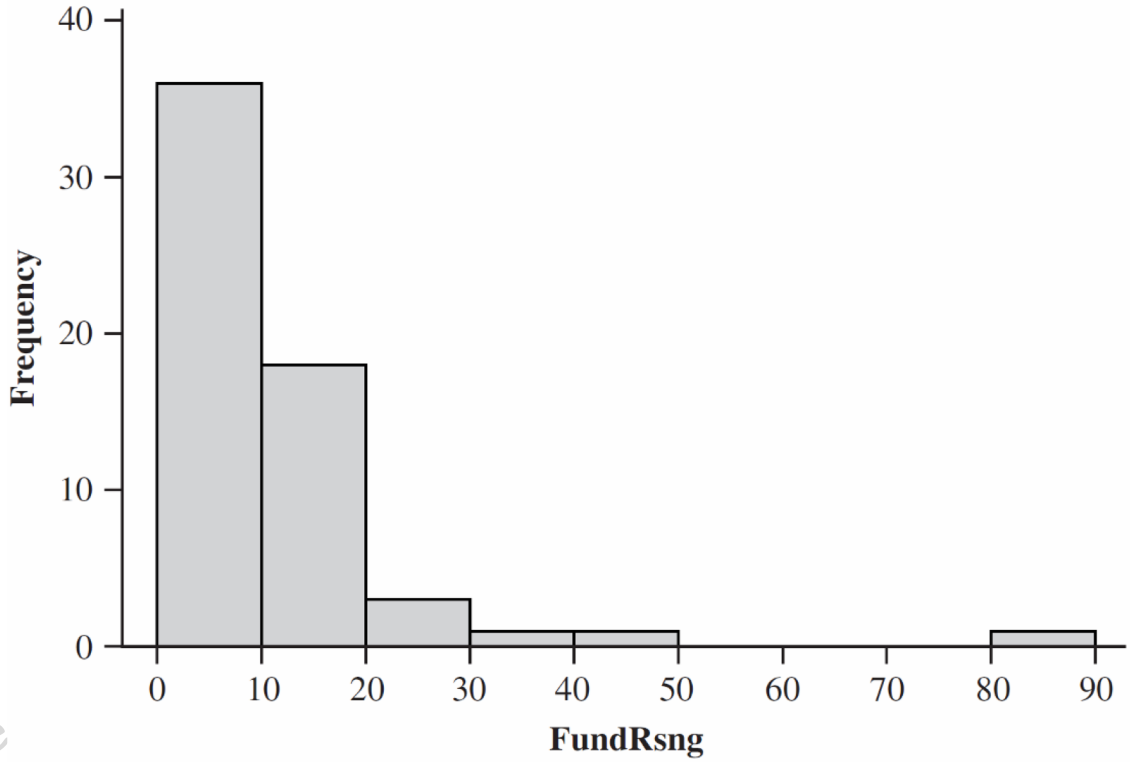
Without any organization, it is difficult to get a sense of the data's most prominent features—what a typical (i.e. representative) value might be, whether values are highly concentrated about a typical value or quite dispersed, whether there are any gaps in the data, what fraction of the values are less than 20%, and so on.

Example 1

cont'd

Figure 1.1 shows what is called a *stem-and-leaf display* as well as a *histogram.*

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
     0 | 01111122223333333344
     0 | 55556666666778888
     1 | 0001222244
     1 | 55666789
     2 | 01
     2 | 6
     3 | 4
     3 |
     4 |
     4 | 8
     5 |
     5 |
     6 |
     6 |
     7 |
     7 |
     8 | 3
```



A Minitab stem-and-leaf display (tenths digit truncated) and histogram for the charity fundraising percentage data

Figure 1.1

Branches of Statistics

Clearly a substantial majority of the charities in the sample spend less than 20% on fundraising, and only a few percentages might be viewed as beyond the bounds of sensible practice.

Having obtained a sample from a population, an investigator would frequently like to use sample information to draw some type of conclusion (make an inference) about the population.

That is, the sample is a means to an end rather than an end in itself. Techniques for generalizing from a sample to a population are gathered within the branch of our discipline called **inferential statistics.**

Probability vs. Statistics

Example:

As an example of the contrasting focus of probability and inferential statistics, consider drivers' use of manual lap belts in cars equipped with automatic shoulder belt systems. (The article "*Automobile Seat Belts: Usage Patterns in Automatic Belt Systems," Human Factors, 1998: 126–135*, summarizes usage data.)

Probability vs. Statistics

Example:

In probability, we might assume that 50% of all drivers of cars equipped in this way in a certain metropolitan area regularly use their lap belt (**an assumption about the population**).

we might ask, “How likely is it that a **sample** of 100 such drivers will include at least 70 who regularly use their lap belt?”

“How many of the drivers in a **sample** of size 100 can we expect to regularly use their lap belt?”

Probability vs. Statistics

Example:

In inferential statistics, we have sample information available; for example, a **sample of 100 drivers** of such cars revealed that 65 regularly use their lap belt.

We might then ask, “Does this provide substantial evidence for concluding that more than 50% of all such drivers in this area regularly use their lap belt?”

In this latter scenario, we are attempting to use sample information to answer a question about the structure of the entire population from which the sample was selected.

Probability vs. Statistics

In a probability problem, properties of the population under study are assumed known (e.g., in a numerical population, some specified distribution of the population values may be assumed), and questions regarding a sample taken from the population are posed and answered.

In a statistics problem, characteristics of a sample are available to the experimenter, and this information enables the experimenter to draw conclusions about the population.

Probability vs. Statistics

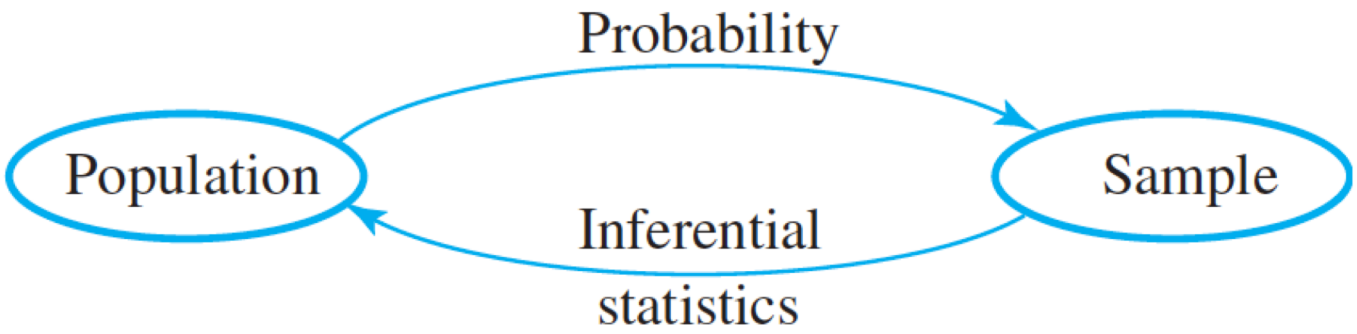


Figure 1.2 The relationship between probability and inferential statistics

The Scope of Modern Statistics (Recommended Reading)

The Scope of Modern Statistics

These days statistical methodology is employed by investigators in virtually all disciplines, including such areas as

- molecular biology (analysis of microarray data)
- ecology (describing quantitatively how individuals in various animal and plant populations are spatially distributed)
- materials engineering (studying properties of various treatments to retard corrosion)

The Scope of Modern Statistics

- marketing (developing market surveys and strategies for marketing new products)
- public health (identifying sources of diseases and ways to treat them)
- civil engineering (assessing the effects of stress on structural elements and the impacts of traffic flows on communities)

As you progress through the book, you'll encounter a wide spectrum of different scenarios in the examples and exercises that illustrate the application of techniques from probability and statistics.

The Scope of Modern Statistics

Many of these scenarios involve data or other material extracted from articles in engineering and science journals.

The methods presented herein have become established and trusted tools in the arsenal of those who work with data.

Meanwhile, statisticians continue to develop new models for describing randomness, and uncertainty and new methodology for analyzing data.

The Scope of Modern Statistics

It is our hope that you will become increasingly convinced of the importance and relevance of the discipline of statistics as you dig more deeply into the book and the subject. Hopefully you'll be turned on enough to want to continue your statistical education beyond your current course.

Collecting Data

Collecting Data

In order for a sample to give good *inference* about the population from which it was collected, this sample should resemble the population, it should be a **representative sample**

One good way to get a representative sample is by collecting it randomly.

Collecting Data

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of techniques for collecting the data. If data is not properly collected, an investigator may not be able to answer the questions under consideration with a reasonable degree of confidence.

One common problem is that the target population—the one about which conclusions are to be drawn—may be different from the population actually sampled. For example, advertisers would like various kinds of information about the television-viewing habits of potential customers.

Collecting Data

The most systematic information of this sort comes from placing monitoring devices in a small number of homes across the United States. It has been conjectured that placement of such devices in and of itself alters viewing behavior, so that characteristics of the sample may be different from those of the target population.

When data collection entails selecting individuals or objects from a frame, the simplest method for ensuring a representative selection is to take a ***simple random sample***. This is one for which any particular subset of the specified size (e.g., a sample of size 100) has the same chance of being selected.

Collecting Data

For example, if the frame consists of 1,000,000 serial numbers, the numbers 1, 2, . . . , up to 1,000,000 could be placed on identical slips of paper. After placing these slips in a box and thoroughly mixing, slips could be drawn one by one until the requisite sample size has been obtained.

Alternatively (and much to be preferred), a table of random numbers or a computer's random number generator could be employed.

Collecting Data

Sometimes alternative sampling methods can be used to make the selection process easier, to obtain extra information, or to increase the degree of confidence in conclusions. One such method, ***stratified sampling***, entails separating the population units into nonoverlapping groups and taking a sample from each one.

For example, a manufacturer of DVD players might want information about customer satisfaction for units produced during the previous year. If three different models were manufactured and sold, a separate sample could be selected from each of the three corresponding strata.

Collecting Data

This would result in information on all three models and ensure that no one model was over- or underrepresented in the entire sample.

Frequently a “convenience” sample is obtained by selecting individuals or objects without systematic randomization. As an example, a collection of bricks may be stacked in such a way that it is extremely difficult for those in the center to be selected.

Collecting Data

If the bricks on the top and sides of the stack were somehow different from the others, resulting sample data would not be representative of the population.

Often an investigator will assume that such a convenience sample approximates a random sample, in which case a statistician's repertoire of inferential methods can be used; however, this is a judgment call.

Collecting Data

Engineers and scientists often collect data by carrying out some sort of designed experiment. This may involve deciding how to allocate several different treatments (such as fertilizers or coatings for corrosion protection) to the various experimental units (plots of land or pieces of pipe).

Alternatively, an investigator may systematically vary the levels or categories of certain factors (e.g., pressure or type of insulating material) and observe the effect on some response variable (such as yield from a production process).

Example 4

An article in the *New York Times* (Jan. 27, 1987) reported that heart attack risk could be reduced by taking aspirin. This conclusion was based on a designed experiment involving both a **control group** of individuals that took a *placebo* having the appearance of aspirin but known to be inert and **a treatment group** that took aspirin according to a specified regimen.

Subjects were randomly assigned to the groups to protect against any biases and so that probability-based methods could be used to analyze the data.

Example 4

cont'd

Of the 11,034 individuals in the control group, 189 subsequently experienced heart attacks, whereas only 104 of the 11,037 in the aspirin group had a heart attack. The incidence rate of heart attacks in the treatment group was only about half that in the control group.

One possible explanation for this result is chance variation—that aspirin really doesn't have the desired effect and the observed difference is just typical variation in the same way that tossing two identical coins would usually produce different numbers of heads.

Example 4

cont'd

However, in this case, inferential methods suggest that chance variation by itself cannot adequately explain the magnitude of the observed difference.