Chapter 5
Statistical Inference

CHAPTER OUTLINE

Section 1 Why Do We Need Statistics?
Section 2 Inference Using a Probability Model
Section 3 Statistical Models
Section 4 Data Collection
Section 5 Some Basic Inferences

In this chapter, we begin our discussion of statistical inference. Probability theory is primarily concerned with calculating various quantities associated with a probability model. This requires that we *know* what the correct probability model is. In applications, this is often not the case, and the best we can say is that the correct probability measure to use is in a set of possible probability measures. We refer to this collection as the *statistical model*. So, in a sense, our uncertainty has increased; not only do we have the uncertainty associated with an outcome or response as described by a probability measure, but now we are also uncertain about what the probability measure is.

Statistical inference is concerned with making statements or inferences about characteristics of the true underlying probability measure. Of course, these inferences must be based on some kind of information; the statistical model makes up part of it. Another important part of the information will be given by an observed outcome or response, which we refer to as the *data*. Inferences then take the form of various statements about the true underlying probability measure from which the data were obtained. These take a variety of forms, which we refer to as *types of inferences*.

The role of this chapter is to introduce the basic concepts and ideas of statistical inference. The most prominent approaches to inference are discussed in Chapters 6, 7, and 8. Likelihood methods require the least structure as described in Chapter 6. Bayesian methods, discussed in Chapter 7, require some additional ingredients. Inference methods based on measures of performance and loss functions are described in Chapter 8.

5.1 Why Do We Need Statistics?

While we will spend much of our time discussing the theory of statistics, we should always remember that statistics is an applied subject. By this we mean that ultimately statistical theory will be applied to real-world situations to answer questions of practical importance.

What is it that characterizes those contexts in which statistical methods are useful? Perhaps the best way to answer this is to consider a practical example where statistical methodology plays an important role.

EXAMPLE 5.1.1 *Stanford Heart Transplant Study*
In the paper by Turnbull, Brown, and Hu entitled "Survivorship of Heart Transplant Data" (*Journal of the American Statistical Association*, March 1974, Volume 69, 74–80), an analysis is conducted to determine whether or not a heart transplant program, instituted at Stanford University, is in fact producing the intended outcome. In this case, the intended outcome is an increased length of life, namely, a patient who receives a new heart should live longer than if no new heart was received.

It is obviously important to ensure that a proposed medical treatment for a disease leads to an improvement in the condition. Clearly, we would not want it to lead to a deterioration in the condition. Also, if it only produced a small improvement, it may not be worth carrying out if it is very expensive or causes additional suffering.

We can never know whether a particular patient who received a new heart has lived longer *because* of the transplant. So our only hope in determining whether the treatment is working is to compare the lifelengths of patients who received new hearts with the lifelengths of patients who did not. There are many factors that influence a patient's lifelength, many of which will have nothing to do with the condition of the patient's heart. For example, lifestyle and the existence of other pathologies, which will vary greatly from patient to patient, will have a great influence. So how can we make this comparison?

One approach to this problem is to imagine that there are probability distributions that describe the lifelengths of the two groups. Let these be given by the densities f_T and f_C, where T denotes transplant and C denotes no transplant. Here we have used C as our label because this group is serving as a *control* in the study to provide some comparison to the treatment (a heart transplant). Then we consider the lifelength of a patient who received a transplant as a random observation from f_T and the lifelength of a patient who did not receive a transplant as a random observation from f_C. We want to compare f_T and f_C, in some fashion, to determine whether or not the transplant treatment is working. For example, we might compute the mean lifelengths of each distribution and compare these. If the mean lifelength of f_T is greater than f_C, then we can assert that the treatment is working. Of course, we would still have to judge whether the size of the improvement is enough to warrant the additional expense and patients' suffering.

If we could take an arbitrarily large number of observations from f_T and f_C, then we know, from the results in previous chapters, that we could determine these distributions with a great deal of accuracy. In practice, however, we are restricted to a relatively small number of observations. For example, in the cited study there were 30 patients

P	X	S	P	X	S	P	X	S
1	49	d	11	1400	a	21	2	d
2	5	d	12	5	d	22	148	d
3	17	d	13	34	d	23	1	d
4	2	d	14	15	d	24	68	d
5	39	d	15	11	d	25	31	d
6	84	d	16	2	d	26	1	d
7	7	d	17	1	d	27	20	d
8	0	d	18	39	d	28	118	a
9	35	d	19	8	d	29	91	a
10	36	d	20	101	d	30	427	a

Table 5.1: Survival times (X) in days and status (S) at the end of the study for each patient (P) in the control group.

in the control group (those who did not receive a transplant) and 52 patients in the treatment group (those who did receive a transplant).

For each control patient, the value of X — the number of days they were alive after the date they were determined to be a candidate for a heart transplant until the termination date of the study — was recorded. For various reasons, these patients did not receive new hearts, e.g., they died before a new heart could be found for them. These data, together with an indicator for the status of the patient at the termination date of the study, are presented in Table 5.1. The indicator value $S = a$ denotes that the patient was alive at the end of the study and $S = d$ denotes that the patient was dead.

For each treatment patient, the value of Y, the number of days they waited for the transplant after the date they were determined to be a candidate for a heart transplant, and the value of Z, the number of days they were alive after the date they received the heart transplant until the termination date of the study, were both recorded. The survival times for the treatment group are then given by the values of $Y + Z$. These data, together with an indicator for the status of the patient at the termination date of the study, are presented in Table 5.2.

We cannot compare f_T and f_C directly because we do not know these distributions. But we do have some information about them because we have obtained values from each, as presented in Tables 5.1 and 5.2. So how do we use these data to compare f_T and f_C to answer the question of central importance, concerning whether or not the treatment is effective? This is the realm of statistics and statistical theory, namely, providing methods for making inferences about unknown probability distributions based upon observations (samples) obtained from them.

We note that we have simplified this example somewhat, although our discussion presents the essence of the problem. The added complexity comes from the fact that typically statisticians will have available additional data on each patient, such as their age, gender, and disease history. As a particular example of this, in Table 5.2 we have the values of both Y and Z for each patient in the treatment group. As it turns out, this additional information, known as covariates, can be used to make our comparisons more accurate. This will be discussed in Chapter 10. ∎

P	Y	Z	S	P	Y	Z	S	P	Y	Z	S
1	0	15	d	19	50	1140	a	37	77	442	a
2	35	3	d	20	22	1153	a	38	2	65	d
3	50	624	d	21	45	54	d	39	26	419	a
4	11	46	d	22	18	47	d	40	32	362	a
5	25	127	d	23	4	0	d	41	13	64	d
6	16	61	d	24	1	43	d	42	56	228	d
7	36	1350	d	25	40	971	a	43	2	65	d
8	27	312	d	26	57	868	a	44	9	264	a
9	19	24	d	27	0	44	d	45	4	25	d
10	17	10	d	28	1	780	a	46	30	193	a
11	7	1024	d	29	20	51	d	47	3	196	a
12	11	39	d	30	35	710	a	48	26	63	d
13	2	730	d	31	82	663	a	49	4	12	d
14	82	136	d	32	31	253	d	50	45	103	a
15	24	1379	a	33	40	147	d	51	25	60	a
16	70	1	d	34	9	51	d	52	5	43	a
17	15	836	d	35	66	479	a				
18	16	60	d	36	20	322	d				

Table 5.2: The number of days until transplant (Y), survival times in days after transplant (Z), and status (S) at the end of the study for each patient (P) in the treatment group.

The previous example provides some evidence that questions of great practical importance require the use of statistical thinking and methodology. There are many situations in the physical and social sciences where statistics plays a key role, and the reasons are just like those found in Example 5.1.1. The central ingredient in all of these is that we are faced with uncertainty. This uncertainty is caused both by variation, which can be modeled via probability, and by the fact that we cannot collect enough observations to know the correct probability models precisely. The first four chapters have dealt with building, and using, a mathematical model to deal with the first source of uncertainty. In this chapter, we begin to discuss methods for dealing with the second source of uncertainty.

Summary of Section 5.1

- Statistics is applied to situations in which we have questions that cannot be answered definitively, typically because of variation in data.
- Probability is used to model the variation observed in the data. Statistical inference is concerned with using the observed data to help identify the true probability distribution (or distributions) producing this variation and thus gain insight into the answers to the questions of interest.

EXERCISES

5.1.1 Compute the mean survival times for the control group and for the treatment groups in Example 5.1.1. What do you conclude from these numbers? Do you think it is valid to base your conclusions about the effectiveness of the treatment on these numbers? Explain why or why not.

5.1.2 Are there any unusual observations in the data presented in Example 5.1.1? If so, what effect do you think these observations have on the mean survival times computed in Exercise 5.1.1?

5.1.3 In Example 5.1.1, we can use the status variable S as a covariate. What is the practical significance of this variable?

5.1.4 A student is uncertain about the mark that will be received in a statistics course. The course instructor has made available a database of marks in the course for a number of years. Can you identify a probability distribution that may be relevant to quantifying the student's uncertainty? What covariates might be relevant in this situation?

5.1.5 The following data were generated from an $N(\mu, 1)$ distribution by a student. Unfortunately, the student forgot which value of μ was used, so we are uncertain about the correct probability distribution to use to describe the variation in the data.

0.2	−0.7	0.0	−1.9	0.7	−0.3	0.3	0.4
0.3	−0.8	1.5	0.1	0.3	−0.7	−1.8	0.2

Can you suggest a plausible value for μ? Explain your reasoning.

5.1.6 Suppose you are interested in determining the average age of all male students at a particular college. The registrar of the college allows you access to a database that lists the age of every student at the college. Describe how you might answer your question. Is this a statistical problem in the sense that you are uncertain about anything and so will require the use of statistical methodology?

5.1.7 Suppose you are told that a characteristic X follows an $N(\mu_1, 1)$ distribution and a characteristic Y follows an $N(\mu_2, 1)$ distribution where μ_1 and μ_2 are unknown. In addition, you are given the results $x_1, \ldots, x_m$ of m independent measurements on X and $y_1, \ldots, y_n$ of n independent measurements on Y. Suggest a method for determining whether or not μ_1 and μ_2 are equal. Can you think of any problems with your approach?

5.1.8 Suppose we know that a characteristic X follows an Exponential(λ) distribution and you are required to determine λ based on i.i.d. observations $x_1, \ldots, x_n$ from this distribution. Suggest a method for doing this. Can you think of any problems with your approach?

PROBLEMS

5.1.9 Can you identify any potential problems with the method we have discussed in Example 5.1.1 for determining whether or not the heart transplant program is effective in extending life?

5.1.10 Suppose you are able to generate samples of any size from a probability distribution P for which it is very difficult to compute $P(C)$ for some set C. Explain how

you might estimate $P(C)$ based on a sample. What role does the size of the sample play in your uncertainty about how good your approximation is. Does the size of $P(C)$ play a role in this?

COMPUTER PROBLEMS

5.1.11 Suppose we want to obtain the distribution of the quantity $Y = X^4 + 2X^3 - 3$ when $X \sim N(0, 1)$. Here we are faced with a form of mathematical uncertainty because it is very difficult to determine the distribution of Y using mathematical methods. Propose a computer method for approximating the distribution function of Y and estimate $P(Y \in (1, 2))$. What is the relevance of statistical methodology to your approach?

DISCUSSION TOPICS

5.1.12 Sometimes it is claimed that all uncertainties can and should be modeled using probability. Discuss this issue in the context of Example 5.1.1, namely, indicate all the things you are uncertain about in this example and how you might propose probability distributions to quantify these uncertainties.

5.2 Inference Using a Probability Model

In the first four chapters, we have discussed probability theory, a good part of which has involved the mathematics of probability theory. This tells us how to carry out various calculations associated with the application of the theory. It is important to keep in mind, however, our reasons for introducing probability in the first place. As we discussed in Section 1.1, probability is concerned with measuring or quantifying uncertainty.

Of course, we are uncertain about many things, and we cannot claim that probability is applicable to all these situations. Let us assume, however, that we are in a situation in which we feel probability is applicable and that we have a probability measure P defined on a collection of subsets of a sample space S for a response s.

In an application of probability, we presume that we know P and are uncertain about a future, or concealed, response value $s \in S$. In such a context, we may be required, or may wish, to make an *inference* about the unknown value of s. This can take the form of a *prediction* or *estimate* of a plausible value for s, e.g., under suitable conditions, we might take the expected value of s as our prediction. In other contexts, we may be asked to construct a subset that has a high probability of containing s and is in some sense small, e.g., find the region that contains at least 95% of the probability and has the smallest size amongst all such regions. Alternatively, we might be asked to assess whether or not a stated value s_0 is an implausible value from the known P, e.g., assess whether or not s_0 lies in a region assigned low probability by P and so is implausible. These are examples of inferences that are relevant to applications of probability theory.

EXAMPLE 5.2.1
As a specific application, consider the lifelength X in years of a machine where it is known that $X \sim$ Exponential(1) (see Figure 5.2.1).

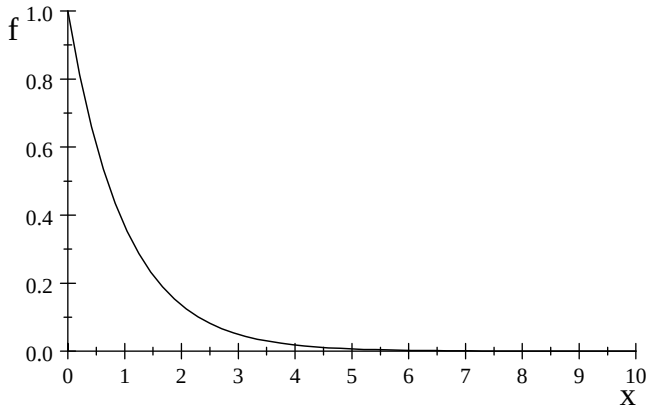

Figure 5.2.1: Plot of the Exponential(1) density f.

Then for a new machine, we might predict its lifelength by $E(X) = 1$ year. Furthermore, from the graph of the Exponential(1) density, it is clear that the smallest interval containing 95% of the probability for X is $(0, c)$, where c satisfies

$$0.95 = \int_0^c e^{-x}\,dx = 1 - e^{-c}$$

or $c = -\ln(0.05) = 2.9957$. This interval gives us a reasonable range of probable lifelengths for the new machine. Finally, if we wanted to assess whether or not $x_0 = 5$ is a plausible lifelength for a newly purchased machine, we might compute the tail probability as

$$P(X > 5) = \int_5^\infty e^{-x}\,dx = e^{-5} = 0.0067,$$

which, in this case, is very small and therefore indicates that $x_0 = 5$ is fairly far out in the tail. The right tail of this density is a region of low probability for this distribution, so $x_0 = 5$ can be considered implausible. It is thus unlikely that a machine will last 5 years, so a purchaser would have to plan to replace the machine before that period is over. ∎

In some applications, we receive some partial information about the unknown s taking the form $s \in C \subset S$. In such a case, we replace P by the conditional probability measure $P(\cdot \mid C)$ when deriving our inferences. Our reasons for doing this are many, and, in general, we can say that most statisticians agree that it is the right thing to do. It is important to recognize, however, that this step does not proceed from a mathematical theorem; rather it can be regarded as a basic axiom or principle of inference. We will refer to this as the *principle of conditional probability*, which will play a key role in some later developments.

EXAMPLE 5.2.2
Suppose we have a machine whose lifelength is distributed as in Example 5.2.1, and the machine has already been running for one year. Then inferences about the lifelength of the machine are based on the conditional distribution, given that $X > 1$. The density of this conditional distribution is given by $e^{-(x-1)}$ for $x > 1$. The predicted lifelength is now

$$E(X \mid X > 1) = \int_1^\infty x e^{-(x-1)}\, dx = -x e^{-(x-1)}\Big|_1^\infty + \int_1^\infty e^{-(x-1)}\, dx = 2.$$

The fact that the additional lifelength is the same as the predicted lifelength before the machine starts working is a special characteristic of the Exponential distribution. This will not be true in general (see Exercise 5.2.4).

The tail probability measuring the plausibility of the value $x_0 = 5$ is given by

$$P(X > 5 \mid X > 1) = \int_5^\infty e^{-(x-1)}\, dx = e^{-4} = 0.0183,$$

which indicates that $x_0 = 5$ is a little more plausible in light of the fact that the machine has already survived one year. The shortest interval containing 0.95 of the conditional probability is now of the form $(1, c)$, where c is the solution to

$$0.95 = \int_1^c e^{-(x-1)}\, dx = e(e^{-1} - e^{-c}),$$

which implies that $c = -\ln\left(e^{-1} - 0.95e^{-1}\right) = 3.9957$. ∎

Our main point in this section is simply that we are already somewhat familiar with inferential concepts. Furthermore, via the principle of conditional probability, we have a basic rule or axiom governing how we go about making inferences in the context where the probability measure P is known and s is not known.

Summary of Section 5.2

- Probability models are used to model uncertainty about future responses.
- We can use the probability distribution to predict a future response or assess whether or not a given value makes sense as a possible future value from the distribution.

EXERCISES

5.2.1 Sometimes the *mode of a density* (the point where the density takes its maximum value) is chosen as a predictor for a future value of a response. Determine this predictor in Examples 5.2.1 and 5.2.2 and comment on its suitability as a predictor.

5.2.2 Suppose it has been decided to use the mean of a distribution to predict a future response. In Example 5.2.1, compute the mean-squared error (expected value of the square of the error between a future value and its predictor) of this predictor, prior to

observing the value. To what characteristic of the distribution of the lifelength does this correspond?

5.2.3 Graph the density of the distribution obtained as a mixture of a normal distribution with mean 4 and variance 1 and a normal distribution with mean -4 and variance 1, where the mixture probability is 0.5. Explain why neither the mean nor the mode is a suitable predictor in this case. (Hint: Section 2.5.4.)

5.2.4 Repeat the calculations of Examples 5.2.1 and 5.2.2 when the lifelength of a machine is known to be distributed as $Y = 10X$, where $X \sim \text{Uniform}[0, 1]$.

5.2.5 Suppose that $X \sim N(10, 2)$. What value would you record as a prediction of a future value of X? How would you justify your choice?

5.2.6 Suppose that $X \sim N(10, 2)$. Record the smallest interval containing 0.95 of the probability for a future response. (Hint: Consider a plot of the density.)

5.2.7 Suppose that $X \sim \text{Gamma}(3, 6)$. What value would you record as a prediction of a future value of X? How would you justify your choice?

5.2.8 Suppose that $X \sim \text{Poisson}(5)$. What value would you record as a prediction of a future value of X? How would you justify your choice?

5.2.9 Suppose that $X \sim \text{Geometric}(1/3)$. What value would you record as a prediction of a future value of X?

5.2.10 Suppose that X follows the following probability distribution.

x	1	2	3	4
$P(X = x)$	1/2	1/4	1/8	1/8

(a) Record a prediction of a future value of X.

(b) Suppose you are then told that $X \geq 2$. Record a prediction of a future value of X that uses this information.

PROBLEMS

5.2.11 Suppose a fair coin is tossed 10 times and the response X measured is the number of times we observe a head.

(a) If you use the expected value of the response as a predictor, then what is the prediction of a future response X?

(b) Using Table D.6 (or a statistical package), compute a shortest interval containing at least 0.95 of the probability for X. Note that it might help to plot the probability function of X first.

(c) What region would you use to assess whether or not a value s_0 is a possible future value? (Hint: What are the regions of low probability for the distribution?) Assess whether or not $x = 8$ is plausible.

5.2.12 In Example 5.2.1, explain (intuitively) why the interval (0, 2.9957) is the shortest interval containing 0.95 of the probability for the lifelength.

5.2.13 (Problem 5.2.11 continued) Suppose we are told that the number of heads observed is an even number. Repeat parts (a), (b), and (c).

5.2.14 Suppose that a response X is distributed $\text{Beta}(a, b)$ with $a, b > 1$ fixed (see Problem 2.4.16). Determine the mean and the mode (point where density takes its

maximum) of this distribution and assess which is the most accurate predictor of a future X when using mean-squared error, i.e., the expected squared distance between X and the prediction.

5.2.15 Suppose that a response X is distributed $N(0, 1)$ and that we have decided to predict a future value using the mean of the distribution.

(a) Determine the prediction for a future X.

(b) Determine the prediction for a future $Y = X^2$.

(c) Comment on the relationship (or lack thereof) between the answers in parts (a) and (b).

5.2.16 Suppose that $X \sim$ Geometric$(1/3)$. Determine the shortest interval containing 0.95 of the probability for a future X. (Hint: Plot the probability function and record the distribution function.)

5.2.17 Suppose that $X \sim$ Geometric$(1/3)$ and we are told that $X > 5$. What value would you record as a prediction of a future value of X? Determine the shortest interval containing 0.95 of the probability for a future X. (Hint: Plot the probability function and record the distribution function.)

DISCUSSION TOPICS

5.2.18 Do you think it is realistic for a practitioner to proceed as if he knows the true probability distribution for a response in a problem?

5.3 Statistical Models

In a statistical problem, we are faced with uncertainty of a different character than that arising in Section 5.2. In a statistical context, we observe the *data* s, but we are uncertain about P. In such a situation, we want to construct inferences about P based on s. This is the inverse of the situation discussed in Section 5.2.

How we should go about making these *statistical inferences* is probably not at all obvious. In fact, there are several possible approaches that we will discuss in subsequent chapters. In this chapter, we will develop the basic ingredients of all the approaches.

Common to virtually all approaches to statistical inference is the concept of the *statistical model* for the data s. This takes the form of a set $\{P_\theta : \theta \in \Omega\}$ of probability measures, one of which corresponds to the true unknown probability measure P that produced the data s. In other words, we are asserting that there *is* a random mechanism generating s, and we *know* that the corresponding probability measure P is one of the probability measures in $\{P_\theta : \theta \in \Omega\}$.

The statistical model $\{P_\theta : \theta \in \Omega\}$ corresponds to the information a statistician brings to the application about what the true probability measure is, or at least what one is willing to assume about it. The variable θ is called the *parameter* of the model, and the set Ω is called the *parameter space.* Typically, we use models where $\theta \in \Omega$ indexes the probability measures in the model, i.e., $P_{\theta_1} = P_{\theta_2}$ if and only if $\theta_1 = \theta_2$. If the probability measures P_θ can all be presented via probability functions or density functions f_θ (for convenience we will not distinguish between the discrete and

continuous case in the notation), then it is common to write the statistical model as $\{f_\theta : \theta \in \Omega\}$.

From the definition of a statistical model, we see that there is a unique value $\theta \in \Omega$, such that P_θ is the true probability measure. We refer to this value as the *true parameter value*. It is obviously equivalent to talk about making inferences about the true parameter value rather than the true probability measure, i.e., an inference about the true value of θ is at once an inference about the true probability distribution. So, for example, we may wish to estimate the true value of θ, construct small regions in Ω that are likely to contain the true value, or assess whether or not the data are in agreement with some particular value θ_0, suggested as being the true value. These are types of inferences, just like those we discussed in Section 5.2, but the situation here is quite different.

EXAMPLE 5.3.1
Suppose we have an urn containing 100 chips, each colored either black B or white W. Suppose further that we are told there are either 50 or 60 black chips in the urn. The chips are thoroughly mixed, and then two chips are withdrawn without replacement. The goal is to make an inference about the true number of black chips in the urn, having observed the data $s = (s_1, s_2)$, where s_i is the color of the ith chip drawn.

In this case, we can take the statistical model to be $\{P_\theta : \theta \in \Omega\}$, where θ is the number of black chips in the urn, so that $\Omega = \{50, 60\}$, and P_θ is the probability measure on

$$S = \{(B, B), (B, W), (W, B), (W, W)\}$$

corresponding to θ. Therefore, P_{50} assigns the probability $50 \cdot 49/(100 \cdot 99)$ to each of the sequences (B, B) and (W, W) and the probability $50 \cdot 50/(100 \cdot 99)$ to each of the sequences (B, W) and (W, B), and P_{60} assigns the probability $60 \cdot 59/(100 \cdot 99)$ to the sequence (B, B), the probability $40 \cdot 39/(100 \cdot 99)$ to the sequence (W, W), and the probability $60 \cdot 40/(100 \cdot 99)$ to each of the sequences (B, W) and (W, B).

The choice of the parameter is somewhat arbitrary, as we could have easily labelled the possible probability measures as P_1 and P_2, respectively. The parameter is in essence only a label that allows us to distinguish amongst the possible candidates for the true probability measure. It is typical, however, to choose this label conveniently so that it means something in the problem under discussion. ∎

We note some additional terminology in common usage. If a single observed value for a response X has the statistical model $\{f_\theta : \theta \in \Omega\}$, then a sample $(X_1, \ldots, X_n)$ (recall that sample here means that the X_i are independent and identically distributed — see Definition 2.8.6) has joint density given by $f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$ for some $\theta \in \Omega$. This specifies the statistical model for the response $(X_1, \ldots, X_n)$. We refer to this as the *statistical model for a sample*. Of course, the true value of θ for the statistical model for a sample is the same as that for a single observation. Sometimes, rather than referring to the statistical model for a sample, we speak of a sample from the statistical model $\{f_\theta : \theta \in \Omega\}$.

Note that, wherever possible, we will use uppercase letters to denote an unobserved value of a random variable X and lowercase letters to denote the observed value. So an observed sample $(X_1, \ldots, X_n)$ will be denoted $(x_1, \ldots, x_n)$.

EXAMPLE 5.3.2
Suppose there are two manufacturing plants for machines. It is known that machines built by the first plant have lifelengths distributed Exponential(1), while machines manufactured by the second plant have lifelengths distributed Exponential(1.5). The densities of these distributions are depicted in Figure 5.3.1.

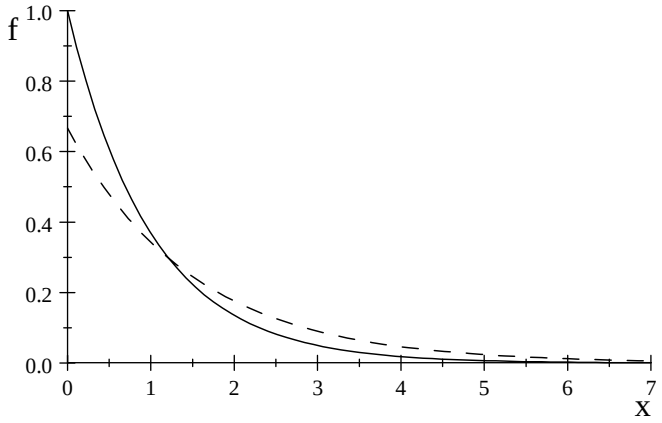

Figure 5.3.1: Plot of the Exponential(1) (solid line) and Exponential(1.5) (dashed line) densities.

You have purchased five of these machines knowing that all five came from the same plant, but you do not know which plant. Subsequently, you observe the lifelengths of these machines, obtaining the sample $(x_1, \ldots, x_5)$, and want to make inferences about the true P.

In this case, the statistical model for a single observation comprises two probability measures $\{P_1, P_2\}$, where P_1 is the Exponential(1) probability measure and P_2 is the Exponential(1.5) probability measure. Here we take the parameter to be $\theta \in \Omega = \{1, 2\}$.

Clearly, longer observed lifelengths favor $\theta = 2$. For example, if

$$(x_1, \ldots, x_5) = (5.0, 3.5, 3.3, 4.1, 2.8),$$

then intuitively we are more certain that $\theta = 2$ than if

$$(x_1, \ldots, x_5) = (2.0, 2.5, 3.0, 3.1, 1.8).$$

The subject of statistical inference is concerned with making statements like this more precise and quantifying our uncertainty concerning the validity of such assertions.

We note again that the quantity θ serves only as a label for the distributions in the model. The value of θ has no interpretation other than as a label and we could just as easily have used different values for the labels. In many applications, however, the parameter θ is taken to be some characteristic of the distribution that takes a unique

value for each distribution in the model. Here, we could have taken θ to be the mean and then the parameter space would be $\Omega = \{1, 1.5\}$. Notice that we could just as well have used the first quartile, or for that matter any other quantile, to have labelled the distributions, provided that each distribution in the family yields a unique value for the characteristic chosen. Generally, any 1–1 transformation of a parameter is acceptable as a parameterization of a statistical model. When we relabel, we refer to this as a *reparameterization* of the statistical model. ∎

We now consider two important examples of statistical models. These are important because they commonly arise in applications.

EXAMPLE 5.3.3 *Bernoulli Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) distribution with $\theta \in [0, 1]$ unknown. We could be observing the results of tossing a coin and recording X_i equal to 1 whenever a head is observed on the ith toss and equal to 0 otherwise. Alternatively, we could be observing items produced in an industrial process and recording X_i equal to 1 whenever the ith item is defective and 0 otherwise. In a biomedical application, the response $X_i = 1$ might indicate that a treatment on a patient has been successful, whereas $X_i = 0$ indicates a failure. In all these cases, we want to know the true value of θ, as this tells us something important about the coin we are tossing, the industrial process, or the medical treatment, respectively.

Now suppose we have no information whatsoever about the true probability θ. Accordingly, we take the parameter space to be $\Omega = [0, 1]$, the set of all possible values for θ. The probability function for the ith sample item is given by

$$f_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i},$$

and the probability function for the sample is given by

$$\prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}.$$

This specifies the model for a sample.

Note that we could parameterize this model by any 1–1 function of θ. For example, $\alpha = \theta^2$ would work (as it is 1–1 on Ω), as would $\psi = \ln\{\theta/(1-\theta)\}$. ∎

EXAMPLE 5.3.4 *Location-Scale Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution with $\theta = (\mu, \sigma^2) \in R^1 \times R^+$ unknown, where $R^+ = (0, \infty)$. For example, we may have observations of heights in centimeters of individuals in a population and feel that it is reasonable to assume that the distribution of heights in the population is normal with some unknown mean and standard deviation.

The density for the sample is then given by

$$\begin{aligned}\prod_{i=1}^{n} f_{(\mu,\sigma^2)}(x_i) &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{n-1}{2\sigma^2}s^2\right\},\end{aligned}$$

because (Problem 5.3.13)

$$\sum_{i=1}^{n}(x_i-\mu)^2 = n(\bar{x}-\mu)^2 + \sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{5.3.1}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the *sample mean*, and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

is the *sample variance*.

Alternative parameterizations for this model are commonly used. For example, rather than using (μ, σ^2), sometimes (μ, σ^{-2}) or (μ, σ) or $(\mu, \ln\sigma)$ are convenient choices. Note that $\ln\sigma$ ranges in R^1 as σ varies in R^+. ∎

Actually, we might wonder how appropriate the model of Example 5.3.4 is for the distribution of heights in a population, for in any finite population the true distribution is discrete (there are only finitely many students). Of course, a normal distribution may provide a good approximation to a discrete distribution, as in Example 4.4.9. So, in Example 5.3.4, we are also assuming that a continuous probability distribution can provide a close approximation to the true discrete distribution. As it turns out, such approximations can lead to great simplifications in the derivation of inferences, so we use them whenever feasible. Such an approximation is, of course, not applicable in Example 5.3.3.

Also note that heights will always be expressed in some specific unit, e.g., centimeters; based on this, we know that the population mean must be in a certain range of values, e.g., $\mu \in (0, 300)$, but the statistical model allows for any value for μ. So we often do have additional information about the true value of the parameter for a model, but it is somewhat imprecise, e.g., we also probably have $\mu \in (100, 300)$. In Chapter 7, we will discuss ways of incorporating such information into our analysis.

Where does the model information $\{P_\theta : \theta \in \Omega\}$ come from in an application? For example, how could we know that heights are approximately normally distributed in Example 5.3.4? Sometimes there is such information based upon previous experience with related applications, but often it is an *assumption* that requires checking before inference procedures can be used. Procedures designed to check such assumptions are referred to as *model-checking* procedures, which will be discussed in Chapter 9. In practice, model-checking procedures are required, or else inferences drawn from the data and statistical model can be erroneous if the model is wrong.

Summary of Section 5.3

- In a statistical application, we do not know the distribution of a response, but we know (or are willing to assume) that the true probability distribution is one of a

set of possible distributions $\{ f_\theta : \theta \in \Omega \}$, where f_θ is the density or probability function (whichever is relevant) for the response. The set of possible distributions is called the *statistical model*.

- The set Ω is called the *parameter space,* and the variable θ is called the *parameter* of the model. Because each value of θ corresponds to a distinct probability distribution in the model, we can talk about the *true value* of θ, as this gives the true distribution via f_θ.

EXERCISES

5.3.1 Suppose there are three coins — one is known to be fair, one has probability $1/3$ of yielding a head on a single toss, and one has probability $2/3$ for head on a single toss. A coin is selected (not randomly) and then tossed five times. The goal is to make an inference about which of the coins is being tossed, based on the sample. Fully describe a statistical model for a single response and for the sample.

5.3.2 Suppose that one face of a symmetrical six-sided die is duplicated but we do not know which one. We do know that if 1 is duplicated, then 2 does not appear; otherwise, 1 does not appear. Describe the statistical model for a single roll.

5.3.3 Suppose we have two populations (I and II) and that variable X is known to be distributed $N(10, 2)$ on population I and distributed $N(8, 3)$ on population II. A sample $(X_1, \ldots, X_n)$ is generated from one of the populations; you are not told which population the sample came from, but you are required to draw inferences about the true distribution based on the sample. Describe the statistical model for this problem. Could you parameterize this model by the population mean, by the population variance? Sometimes problems like this are called *classification problems* because making inferences about the true distribution is equivalent to classifying the sample as belonging to one of the populations.

5.3.4 Suppose the situation is as described in Exercise 5.3.3, but now the distribution for population I is $N(10, 2)$ and the distribution for population II is $N(10, 3)$. Could you parameterize the model by the population mean? By the population variance? Justify your answer.

5.3.5 Suppose that a manufacturing process produces batteries whose lifelengths are known to be exponentially distributed but with the mean of the distribution completely unknown. Describe the statistical model for a single observation. Is it possible to parameterize this model by the mean? Is it possible to parameterize this model by the variance? Is it possible to parameterize this model by the *coefficient of variation* (the coefficient of variation of a distribution equals the standard deviation divided by the mean)?

5.3.6 Suppose it is known that a response X is distributed Uniform$[0, \beta]$, where $\beta > 0$ is unknown. Is it possible to parameterize this model by the first quartile of the distribution? (The first quartile of the distribution of a random variable X is the point c satisfying $P(X \leq c) = 0.25$.) Explain why or why not.

5.3.7 Suppose it is known that a random variable X follows one of the following distributions.

θ	$P_\theta(X=1)$	$P_\theta(X=2)$	$P_\theta(X=3)$
A	$1/2$	$1/2$	0
B	0	$1/2$	$1/2$

(a) What is the parameter space Ω?

(b) Suppose we observe a value $X = 1$. What is the true value of the parameter? What is the true distribution of X?

(c) What could you say about the true value of the parameter if you had observed $X = 2$?

5.3.8 Suppose we have a statistical model $\{P_1, P_2\}$, where P_1 and P_2 are probability measures on a sample space S. Further suppose there is a subset $C \subset S$ such that $P_1(C) = 1$ while $P_2(C^c) = 1$. Discuss how you would make an inference about the true distribution of a response s after you have observed a single observation.

5.3.9 Suppose you know that the probability distribution of a variable X is either P_1 or P_2. If you observe $X = 1$ and $P_1(X = 1) = 0.75$ while $P_2(X = 1) = 0.001$, then what would you guess as the true distribution of X? Give your reasoning for this conclusion.

5.3.10 Suppose you are told that class #1 has 35 males and 65 females while class #2 has 45 males and 55 females. You are told that a particular student from one of these classes is female, but you are not told which class she came from.

(a) Construct a statistical model for this problem, identifying the parameter, the parameter space, and the family of distributions. Also identify the data.

(b) Based on the data, do you think a reliable inference is feasible about the true parameter value? Explain why or why not.

(c) If you had to make a guess about which distribution the data came from, what choice would you make? Explain why.

PROBLEMS

5.3.11 Suppose in Example 5.3.3 we parameterize the model by $\psi = \ln\{\theta/(1-\theta)\}$. Record the statistical model using this parameterization, i.e., record the probability function using ψ as the parameter and record the relevant parameter space.

5.3.12 Suppose in Example 5.3.4 we parameterize the model by $(\mu, \ln\sigma) = (\mu, \psi)$. Record the statistical model using this parameterization, i.e., record the density function using (μ, ψ) as the parameter and record the relevant parameter space.

5.3.13 Establish the identity (5.3.1).

5.3.14 A sample $(X_1, \ldots, X_n)$ is generated from a Bernoulli(θ) distribution with $\theta \in [0, 1]$ unknown, but only $T = \sum_{i=1}^{n} X_i$ is observed by the statistician. Describe the statistical model for the observed data.

5.3.15 Suppose it is known that a response X is distributed $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in R^1 \times R^+$ and θ is completely unknown. Show how to calculate the first

quartile of each distribution in this model from the values (μ, σ^2). Is it possible to parameterize the model by the first quartile? Explain your answer.

5.3.16 Suppose response X is known to be distributed $N(Y, \sigma^2)$, where $Y \sim N(0, \delta^2)$ and $\sigma^2, \delta^2 > 0$ are completely unknown. Describe the statistical model for an observation (X, Y). If Y is not observed, describe the statistical model for X.

5.3.17 Suppose we have a statistical model $\{P_1, P_2\}$, where P_1 is an $N(10, 1)$ distribution while P_2 is an $N(0, 1)$ distribution.

(a) Is it possible to make any kind of reliable inference about the true distribution based on a single observation? Why or why not?

(b) Repeat part (a) but now suppose that P_1 is a $N(1, 1)$ distribution.

5.3.18 Suppose we have a statistical model $\{P_1, P_2\}$, where P_1 is an $N(1, 1)$ distribution while P_2 is an $N(0, 1)$ distribution. Further suppose that we had a sample $x_1, \ldots, x_{100}$ from the true distribution. Discuss how you might go about making an inference about the true distribution based on the sample.

DISCUSSION TOPICS

5.3.19 Explain why you think it is important that statisticians state very clearly what they are assuming any time they carry out a statistical analysis.

5.3.20 Consider the statistical model given by the collection of $N(\mu, \sigma_0^2)$ distributions where $\mu \in R^1$ is considered completely unknown, but σ_0^2 is assumed known. Do you think this is a reasonable model to use in an application? Give your reasons why or why not.

5.4 Data Collection

The developments of Sections 5.2 and 5.3 are based on the observed response s being a realization from a probability measure P. In fact, in many applications, this is an assumption. We are often presented with data that could have been produced in this way, but we cannot always be sure.

When we cannot be sure that the data were produced by a random mechanism, then the statistical analysis of the data is known as an *observational study.* In an observational study, the statistician merely observes the data rather than intervening directly in generating the data, to ensure that the randomness assumption holds. For example, suppose a professor collects data from his students for a study that examines the relationship between grades and part-time employment. Is it reasonable to regard the data collected as having come from a probability distribution? If so, how would we justify this?

It is important for a statistician to distinguish carefully between situations that are observational studies and those that are not. As the following discussion illustrates, there are qualifications that must be applied to the analysis of an observational study. While statistical analyses of observational studies are valid and indeed important, we must be aware of their limitations when interpreting their results.

5.4.1 Finite Populations

Suppose we have a finite set Π of objects, called the *population*, and a real-valued function X (sometimes called a *measurement*) defined on Π. So for each $\pi \in \Pi$, we have a real-valued quantity $X(\pi)$ that measures some aspect or feature of π.

For example, Π could be the set of all students currently enrolled full-time at a particular university, with $X(\pi)$ the height of student π in centimeters. Or, for the same Π, we could take $X(\pi)$ to be the gender of student π, where $X(\pi) = 1$ denotes female and $X(\pi) = 2$ denotes male. Here, height is a *quantitative variable*, because its values mean something on a numerical scale, and we can perform arithmetic on these values, e.g., calculate a mean. On the other hand, gender is an example of a *categorical variable* because its values serve only to classify, and any other choice of unique real numbers would have served as well as the ones we chose. The first step in a statistical analysis is to determine the types of variables we are working with because the relevant statistical analysis techniques depend on this.

The population and the measurement together produce a *population distribution* over the population. This is specified by the *population cumulative distribution function* $F_X : R^1 \to [0, 1]$, where

$$F_X(x) = \frac{|\{\pi : X(\pi) \leq x\}|}{N},$$

with $|A|$ being the number of elements in the set A, and $N = |\Pi|$. Therefore, $F_X(x)$ is the proportion of elements in Π with their measurement less than or equal to x.

Consider the following simple example where we can calculate F_X exactly.

EXAMPLE 5.4.1
Suppose that Π is a population of $N = 20$ plots of land of the same size. Further suppose that $X(\pi)$ is a measure of the fertility of plot π on a 10-point scale and that the following measurements were obtained.

4	8	6	7	8	3	7	5	4	6
9	5	7	5	8	3	4	7	8	3

Then we have

$$F_X(x) = \begin{cases} 0 & x < 3 \\ 3/20 & 3 \leq x < 4 \\ 6/20 & 4 \leq x < 5 \\ 9/20 & 5 \leq x < 6 \\ 11/20 & 6 \leq x < 7 \\ 15/20 & 7 \leq x < 8 \\ 19/20 & 8 \leq x < 9 \\ 1 & 9 \leq x \end{cases}$$

because, for example, 6 out of the 20 plots have fertility measurements less than or equal to 4. ∎

The goal of a statistician in this context is to know the function F_X as precisely as possible. If we know F_X exactly, then we have identified the distribution of X

over Π. One way of knowing the distribution exactly is to conduct a *census*, wherein, the statistician goes out and observes $X(\pi)$ for every $\pi \in \Pi$ and then calculates F_X. Sometimes this is feasible, but often it is not possible or even desirable, due to the costs involved in the accurate accumulation of all the measurements — think of how difficult it might be to collect the heights of all the students at your school.

While sometimes a census is necessary, even mandated by law, often a very accurate approximation to F_X can be obtained by selecting a subset

$$\{\pi_1, \ldots, \pi_n\} \subset \Pi$$

for some $n < N$. We then approximate $F_X(x)$ by the *empirical distribution function* defined by

$$\begin{aligned} \hat{F}_X(x) &= \frac{|\{\pi_i : X(\pi_i) \leq x, i = 1, \ldots, n\}|}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X(\pi_i)) . \end{aligned}$$

We could also measure more than one aspect of π to produce a *multivariate measurement* $X : \Pi \to R^k$ for some k. For example, if Π is again the population of students, we might have $X(\pi) = (X_1(\pi), X_2(\pi))$, where $X_1(\pi)$ is the height in centimeters of student π and $X_2(\pi)$ is the weight of student π in kilograms. We will discuss multivariate measurements in Chapter 10, where our concern is the relationships amongst variables, but we focus on univariate measurements here.

There are two questions we need to answer now — namely, how should we select the subset $\{\pi_1, \ldots, \pi_n\}$ and how large should n be?

5.4.2 Simple Random Sampling

We will first address the issue of selecting $\{\pi_1, \ldots, \pi_n\}$. Suppose we select this subset according to some given rule based on the unique label that each $\pi \in \Pi$ possesses. For example, if the label is a number, we might order the numbers and then take the n elements with the smallest labels. Or we could order the numbers and take every other element until we have a subset of n, etc.

There are many such rules we could apply, and there is a basic problem with all of them. If we want $\hat{F}_X$ to approximate F_X for the full population, then, when we employ a rule, we run the risk of only selecting $\{\pi_1, \ldots, \pi_n\}$ from a subpopulation. For example, if we use student numbers to identify each element of a population of students, and more senior students have lower student numbers, then, when n is much smaller than N and we select the students with smallest student numbers, $\hat{F}_X$ is really only approximating the distribution of X in the population of senior students at best. This distribution could be very different from F_X. Similarly, for any other rule we employ, even if we cannot imagine what the subpopulation could be, there may be such a *selection effect*, or *bias*, induced that renders the estimate invalid.

This is the qualification we need to apply when analyzing the results of observational studies. In an observational study, the data are generated by some rule, typically

unknown to the statistician; this means that any conclusions drawn based on the data $X(\pi_1), , \ldots, X(\pi_n)$ may not be valid for the full population.

There seems to be only one way to guarantee that selection effects are avoided, namely, the set $\{\pi_1, \ldots, \pi_n\}$ must be selected using randomness. For *simple random sampling*, this means that a random mechanism is used to select the π_i in such a way that each subset of n has probability $1/\binom{N}{n}$ of being chosen. For example, we might place N chips in a bowl, each with a unique label corresponding to a population element, and then randomly draw n chips from the bowl without replacement. The labels on the drawn chips identify the individuals that have been selected from Π. Alternatively, for the randomization, we might use a table of random numbers, such as Table D.1 in Appendix D (see Table D.1 for a description of how it is used) or generate random values using a computer algorithm (see Section 2.10).

Note that with simple random sampling $(X(\pi_1), , \ldots, X(\pi_n))$ is random. In particular, when $n = 1$, we then have

$$P(X(\pi_1) \leq x) = F_X(x),$$

namely, the probability distribution of the random variable $X(\pi_1)$ is the same as the population distribution.

EXAMPLE 5.4.2
Consider the context of Example 5.4.1. When we randomly select the first plot from Π, it is clear that each plot has probability $1/20$ of being selected. Then we have

$$P(X(\pi_1) \leq x) = \frac{|\{\pi : X(\pi) \leq x\}|}{20} = F_X(x)$$

for every $x \in R^1$. ∎

Prior to observing the sample, we also have $P(X(\pi_2) \leq x) = F_X(x)$. Consider, however, the distribution of $X(\pi_2)$ given that $X(\pi_1) = x_1$. Because we have removed one population member, with measurement value x_1, then $NF_X(x) - 1$ is the number of individuals left in Π with $X(\pi) \leq x_1$. Therefore,

$$P\left(X(\pi_2) \leq x \mid X(\pi_1) = x_1\right) = \begin{cases} \frac{NF_X(x)-1}{N-1} & x \geq x_1 \\ \frac{NF_X(x)}{N-1} & x < x_1. \end{cases}$$

Note that this is not equal to $F_X(x)$.

So with simple random sampling, $X(\pi_1)$ and $X(\pi_2)$ are not independent. Observe, however, that when N is large, then

$$P\left(X(\pi_2) \leq x \mid X(\pi_1) = x_1\right) \approx F_X(x),$$

so that $X(\pi_1)$ and $X(\pi_2)$ are approximately independent and identically distributed (i.i.d.). Similar calculations lead to the conclusion that, when N is large and n is small relative to N, then with simple random sampling from the population, the random variables

$$X(\pi_1), \ldots, X(\pi_n)$$

are approximately i.i.d. and with distribution given by F_X. So we will treat the observed values $(x_1, \ldots, x_n)$ of $(X(\pi_1), \ldots, X(\pi_n))$ as a sample (in the sense of Definition 2.8.6) from F_X. In this text, unless we indicate otherwise, we will always assume that n is small relative to N so that this approximation makes sense.

Under the i.i.d. assumption, the weak law of large numbers (Theorem 4.2.1) implies that the empirical distribution function $\hat{F}_X$ satisfies

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X(\pi_i)) \xrightarrow{P} F_X(x)$$

as $n \to \infty$. So we see that $\hat{F}_X$ can be considered as an estimate of the population cumulative distribution function (cdf) F_X.

Whenever the data have been collected using simple random sampling, we will refer to the statistical investigation as a *sampling study*. It is a basic principle of good statistical practice that sampling studies are always preferred over observational studies, whenever they are feasible. This is because we can be sure that, with a sampling study, any conclusions we draw based on the sample $\pi_1, \ldots, \pi_n$ will apply to the population Π of interest. With observational studies, we can never be sure that the sample data have not actually been selected from some proper subset of Π. For example, if you were asked to make inferences about the distribution of student heights at your school but selected some of your friends as your sample, then it is clear that the estimated cdf may be very unlike the true cdf (possibly more of your friends are of one gender than the other).

Often, however, we have no choice but to use observational data for a statistical analysis. Sampling directly from the population of interest may be extremely difficult or even impossible. We can still treat the results of such analyses as a form of evidence, but we must be wary about possible selection effects and acknowledge this possibility. Sampling studies constitute a higher level of statistical evidence than observational studies, as they avoid the possibility of selection effects.

In Chapter 10, we will discuss *experiments* that constitute the highest level of statistical evidence. Experiments are appropriate when we are investigating the possibility of cause–effect relationships existing amongst variables defined on populations.

The second question we need to address concerns the choice of the sample size n. It seems natural that we would like to choose this as large as possible. On the other hand, there are always costs associated with sampling, and sometimes each sample value is very expensive to obtain. Furthermore, it is often the case that the more data we collect, the more difficulty we have in making sure that the data are not corrupted by various errors that can arise in the collection process. So our answer, concerning how large n need be, is that we want it chosen large enough so that we obtain the accuracy necessary but no larger. Accordingly, the statistician must specify what accuracy is required, and then n is determined.

We will see in the subsequent chapters that there are various methods for specifying the required accuracy in a problem and then determining an appropriate value for n. Determining n is a key component in the implementation of a sampling study and is often referred to as a *sample-size calculation*.

If we define

$$f_X(x) = \frac{|\{\pi : X(\pi) = x\}|}{N} = \frac{1}{N}\sum_{\pi \in \Pi} I_{\{x\}}(X(\pi)),$$

namely, $f_X(x)$ is the proportion of population members satisfying $X(\pi) = x$, then we see that f_X plays the role of the probability function because

$$F_X(x) = \sum_{z \le x} f_X(z).$$

We refer to f_X as the *population relative frequency function.* Now, $f_X(x)$ may be estimated, based on the sample $\{\pi_1, \ldots, \pi_n\}$, by

$$\hat{f}_X(x) = \frac{|\{\pi_i : X(\pi_i) = x, i = 1, \ldots, n\}|}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{\{x\}}(X(\pi_i)),$$

namely, the proportion of sample members π satisfying $X(\pi) = x$.

With categorical variables, $\hat{f}_X(x)$ estimates the population proportion $f_X(x)$ in the category specified by x. With some quantitative variables, however, f_X is not an appropriate quantity to estimate, and an alternative function must be considered.

5.4.3 Histograms

Quantitative variables can be further classified as either discrete or continuous variables. Continuous variables are those that we can measure to an arbitrary precision as we increase the accuracy of a measuring instrument. For example, the height of an individual could be considered a continuous variable, whereas the number of years of education an individual possesses would be considered a discrete quantitative variable. For discrete quantitative variables, f_X is an appropriate quantity to describe a population distribution, but we proceed differently with continuous quantitative variables.

Suppose that X is a continuous quantitative variable. In this case, it makes more sense to *group* values into intervals, given by

$$(h_1, h_2], (h_2, h_3], \ldots, (h_{m-1}, h_m],$$

where the h_i are chosen to satisfy $h_1 < h_2 < \cdots < h_m$ with (h_1, h_m) effectively covering the range of possible values for X. Then we define

$$h_X(x) = \begin{cases} \frac{|\{\pi : X(\pi) \in (h_i, h_{i+1}]\}|}{N(h_{i+1} - h_i)} & x \in (h_i, h_{i+1}] \\ 0 & \text{otherwise} \end{cases}$$

and refer to h_X as a *density histogram function.* Here, $h_X(x)$ is the proportion of population elements π that have their measurement $X(\pi)$ in the interval $(h_i, h_{i+1}]$ containing x, divided by the length of the interval.

In Figure 5.4.1, we have plotted a density histogram based on a sample of 10,000 from an $N(0, 1)$ distribution (we are treating this sample as the full population) and

using the values $h_1 = -5, h_2 = -4, \ldots, h_{11} = 5$. Note that the vertical lines are only artifacts of the plotting software and do not represent values of h_X, as these are given by the horizontal lines.

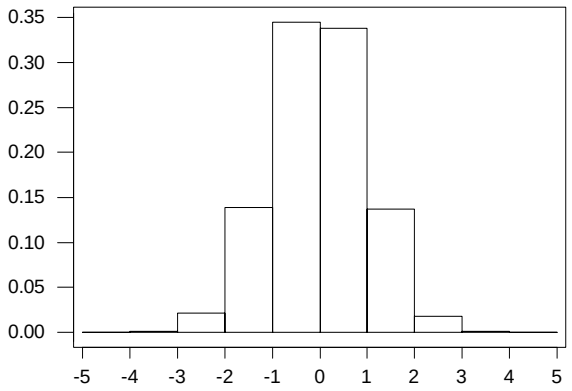

Figure 5.4.1: Density histogram function for a sample of 10,000 from an $N(0, 1)$ distribution using the values $h_1 = -5, h_2 = -4, \ldots, h_{11} = 5$.

If $x \in (h_i, h_{i+1}]$, then $h_X(x)\,(h_{i+1} - h_i)$ gives the proportion of individuals in the population that have their measurement X (π) in $(h_i, h_{i+1}]$. Furthermore, we have

$$F_X(h_j) = \int_{-\infty}^{h_j} h_X(x)\,dx$$

for each interval endpoint and

$$F_X(h_j) - F_X(h_i) = \int_{h_i}^{h_j} h_X(x)\,dx$$

when $h_i \leq h_j$. If the intervals $(h_i, h_{i+1}]$ are small, then we expect that

$$F_X(b) - F_X(a) \approx \int_a^b h_X(x)\,dx$$

for any choice of $a < b$.

Now suppose that the lengths $h_{i+1} - h_i$ are small and N is very large. Then it makes sense to imagine a smooth, continuous function f_X, e.g., perhaps a normal or gamma density function, that approximates h_X in the sense that

$$\int_a^b f_X(x)\,dx \approx \int_a^b h_X(x)\,dx$$

for every $a < b$. Then we will also have

$$\int_a^b f_X(x)\,dx \approx F_X(b) - F_X(a)$$

for every $a < b$. We will refer to such an f_X as a density function for the population distribution.

In essence, this is how many continuous distributions arise in practice. In Figure 5.4.2, we have plotted a density histogram for the same values used in Figure 5.4.1, but this time we used the interval endpoints $h_1 = -5, h_2 = -4.75, \ldots, h_{41} = 5$. We note that Figure 5.4.2 looks much more like a continuous function than does Figure 5.4.1.

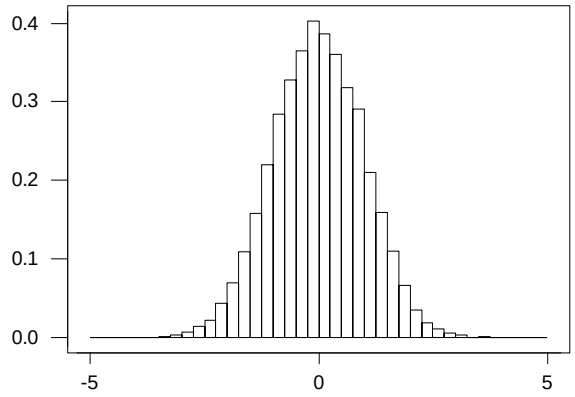

Figure 5.4.2: Density histogram function for a sample of 10,000 from an $N(0, 1)$ distribution using the values $h_1 = -5, h_2 = -4.75, \ldots, h_{41} = 5$.

5.4.4 Survey Sampling

Finite population sampling provides the formulation for a very important application of statistics, namely, *survey sampling* or *polling*. Typically, a survey consists of a set of questions that are asked of a sample $\{\pi_1, \ldots, \pi_n\}$ from a population Π. Each question corresponds to a measurement, so if there are m questions, the response from a respondent π is the m-dimensional vector $(X_1(\pi), X_2(\pi), \ldots, X_m(\pi))$. A very important example of survey sampling is the pre-election polling that is undertaken to predict the outcome of a vote. Also, many consumer product companies engage in extensive market surveys to try to learn what consumers want and so gain information that can lead to improved sales.

Typically, the analysis of the results will be concerned not only with the population distributions of the individual X_i over the population but also the joint population distributions. For example, the joint cumulative distribution function of (X_1, X_2) is given by

$$F_{(X_1,X_2)}(x_1, x_2) = \frac{|\{\pi : X_1(\pi) \le x_1, X_2(\pi) \le x_2\}|}{N},$$

namely, $F_{(X_1,X_2)}(x_1, x_2)$ is the proportion of the individuals in the population whose X_1 measurement is no greater than x_1 and whose X_2 measurement is no greater than x_2. Of course, we can also define the joint distributions of three or more measurements.

These joint distributions are what we use to answer questions like, is there a relationship between X_1 and X_2, and if so,what form does it take? This topic will be extensively discussed in Chapter 10. We can also define $f_{(X_1,X_2)}$ for the joint distribution, and joint density histograms are again useful when X_1 and X_2 are both continuous quantitative variables.

EXAMPLE 5.4.3
Suppose there are four candidates running for mayor in a particular city. A random sample of 1000 voters is selected; they are each asked if they will vote and, if so, which of the four candidates they will vote for. Additionally, the respondents are asked their age. We denote the answer to the question of whether or not they will vote by X_1, with $X_1(\pi) = 1$ meaning yes and $X_1(\pi) = 0$ meaning no. For those voting, we denote by X_2 the response concerning which candidate they will vote for, with $X_2(\pi) = i$ indicating candidate i. Finally, the age in years of the respondent is denoted by X_3. In addition to the distributions of X_1 and X_2, the pollster is also interested in the joint distributions of (X_1, X_3) and (X_2, X_3), as these tell us about the relationship between voter participation and age in the first case and candidate choice and age in the second case. ∎

There are many interesting and important aspects to survey sampling that go well beyond this book. For example, it is often the case with human populations that a randomly selected person will not respond to a survey. This is called *nonresponse error*, and it is a serious selection effect. The sampler must design the study carefully to try to mitigate the effects of nonresponse error. Furthermore, there are variants of simple random sampling (see Challenge 5.4.20) that can be preferable in certain contexts, as these increase the accuracy of the results. The design of the actual questionnaire used is also very important, as we must ensure that responses address the issues intended without biasing the results.

Summary of Section 5.4

- Simple random sampling from a population Π means that we randomly select a subset of size n from Π in such a way that each subset of n has the same probability — namely, $1/\binom{|\Pi|}{n}$ — of being selected.
- Data that arise from a sampling study are generated from the distribution of the measurement of interest X over the whole population Π rather than some subpopulation. This is why sampling studies are preferred to observational studies.
- When the sample size n is small relative to $|\Pi|$, we can treat the observed values of X as a sample from the distribution of X over the population.

EXERCISES

5.4.1 Suppose we have a population $\Pi = \{\pi_1, \ldots, \pi_{10}\}$ and quantitative measurement X given by:

i	1	2	3	4	5	6	7	8	9	10
$X(\pi_i)$	1	1	2	1	2	3	3	1	2	4

Calculate F_X, f_X, μ_X, and σ_X^2.

5.4.2 Suppose you take a sample of $n = 3$ (without replacement) from the population in Exercise 5.4.1.

(a) Can you consider this as an approximate i.i.d. sample from the population distribution? Why or why not?

(b) Explain how you would actually physically carry out the sampling from the population in this case. (Hint: Table D.1.)

(c) Using the method you outlined in part (b), generate three samples of size $n = 3$ and calculate $\bar{X}$ for each sample.

5.4.3 Suppose you take a sample of $n = 4$ (with replacement) from the population in Exercise 5.4.1.

(a) Can you consider this as an approximate i.i.d. sample from the population distribution? Why or why not?

(b) Explain how you would actually physically carry out the sampling in this case.

(c) Using the method you outlined in part (b), generate three samples of size $n = 3$ and calculate $\bar{X}$ for each sample.

5.4.4 Suppose we have a finite population Π and a measurement $X : \Pi \to \{0, 1\}$ where $|\Pi| = N$ and $|\{\pi : X(\pi) = 0\}| = a$.

(a) Determine $f_X(0)$ and $f_X(1)$. Can you identify this population distribution?

(b) For a simple random sample of size n, determine the probability that $n\hat{f}_X(0) = x$.

(c) Under the assumption of i.i.d. sampling, determine the probability that $n\hat{f}_X(0) = x$.

5.4.5 Suppose the following sample of size of $n = 20$ is obtained from an industrial process.

3.9	7.2	6.9	4.5	5.8	3.7	4.4	4.5	5.6	2.5
4.8	8.5	4.3	1.2	2.3	3.1	3.4	4.8	1.8	3.7

(a) Construct a density histogram for this data set using the intervals (1, 4.5], (4.5, 5.5], (5.5, 6.5](6.5, 10].

(b) Construct a density histogram for this data set using the intervals (1, 3.5], (3.5, 4.5], (4.5, 6.5], (6.5, 10].

(c) Based on the results of parts (a) and (b), what do you conclude about histograms?

5.4.6 Suppose it is known that in a population of 1000 students, 350 students will vote for party A, 550 students will vote for party B, and the remaining students will vote for party C.

(a) Explain how such information can be obtained.

(b) If we let $X : \Pi \to \{A, B, C\}$ be such that $X(\pi)$ is the party that π will vote for, then explain why we cannot represent the population distribution of X by F_X.

(c) Compute f_X.

(d) Explain how one might go about estimating f_X prior to the election.

(e) What is unrealistic about the population distribution specified via f_X? (Hint: Does it seem realistic, based on what you know about voting behavior?)

5.4.7 Consider the population Π to be files stored on a computer at a particular time. Suppose that $X(\pi)$ is the type of file as indicated by its extension, e.g., .mp3. Is X a categorical or quantitative variable?

5.4.8 Suppose that you are asked to estimate the proportion of students in a college of 15, 000 students who intend to work during the summer.

(a) Identify the population Π, the variable X, and f_X. What kind of variable is X?

(b) How could you determine f_X exactly?

(c) Why might you not be able to determine f_X exactly? Propose a procedure for estimating f_X in such a situation.

(d) Suppose you were also asked to estimate the proportion of students who intended to work but could not find a job. Repeat parts (a), (b), and (c) for this situation.

5.4.9 Sometimes participants in a poll do not respond truthfully to a question. For example, students who are asked "Have you ever illegally downloaded music?" may not respond truthfully even if they are assured that their responses are confidential. Suppose a simple random sample of students was chosen from a college and students were asked this question.

(a) If students were asked this question by a person, comment on how you think the results of the sampling study would be affected.

(b) If students were allowed to respond anonymously, perhaps by mailing in a questionnaire, comment on how you think the results would be affected.

(c) One technique for dealing with the respondent bias induced by such questions is to have students respond truthfully only when a certain random event occurs. For example, we might ask a student to toss a fair coin three times and lie whenever they obtain two heads. What is the probability that a student tells the truth? Once you have completed the study and have recorded the proportion of students who said they did cheat, what proportion would you record as your estimate of the proportion of students who actually did cheat?

5.4.10 A market research company is asked to determine how satisfied owners are with their purchase of a new car in the last 6 months. Satisfaction is to be measured by respondents choosing a point on a seven-point scale $\{1, 2, 3, 4, 5, 6, 7\}$, where 1 denotes completely dissatisfied and 7 denotes completely satisfied (such a scale is commonly called a *Likert scale*).

(a) Identify Π, the variable X, and f_X.

(b) It is common to treat a variable such as X as a quantitative variable. Do you think this is correct? Would it be correct to treat X as a categorical variable?

(c) A common criticism of using such a scale is that the interpretation of a statement such as $3 =$ "I'm somewhat dissatisfied" varies from one person to another. Comment on how this affects the validity of the study.

COMPUTER EXERCISES

5.4.11 Generate a sample of 1000 from an $N(3, 2)$ distribution.

(a) Calculate $\hat{F}_X$ for this sample.

(b) Plot a density histogram based on these data using the intervals of length 1 over the range $(-5, 10)$.

(c) Plot a density histogram based on these data using the intervals of length 0.1 over the range $(-5, 10)$.

(d) Comment on the difference in the look of the histograms in parts (b) and (c). To what do you attribute this?

(e) What limits the size of the intervals we use to group observations when we are plotting histograms?

5.4.12 Suppose we have a population of 10,000 elements, each with a unique label from the set $\{1, 2, 3, \ldots, 10, 000\}$.

(a) Generate a sample of 500 labels from this population using simple random sampling.

(b) Generate a sample of 500 labels from this population using i.i.d. sampling.

PROBLEMS

5.4.13 Suppose we have a finite population Π and a measurement $X : \Pi \to \{0, 1, 2\}$, where $|\Pi| = N$ and $|\{\pi : X(\pi) = 0\}| = a$ and $|\{\pi : X(\pi) = 1\}| = b$. This problem generalizes Exercise 5.4.4.

(a) Determine $f_X(0)$, $f_X(1)$, and $f_X(2)$.

(b) For a simple random sample of size n, determine the probability that $\hat{f}_X(0) = f_0$, $\hat{f}_X(1) = f_1$, and $\hat{f}_X(2) = f_2$.

(c) Under the assumption of i.i.d. sampling, determine the probability that $\hat{f}_X(0) = f_0$, $\hat{f}_X(1) = f_1$, and $\hat{f}_X(2) = f_2$.

5.4.14 Suppose X is a quantitative measurement defined on a finite population.

(a) Prove that the population mean equals $\mu_X = \sum_x x f_X(x)$, i.e., the average of $X(\pi)$ over all population elements π equals μ_X.

(b) Prove that the population variance is given by $\sigma_X^2 = \sum_x (x - \mu_X)^2 f_X(x)$, i.e., the average of $(X(\pi) - \mu_X)^2$ over all population elements π equals σ_X^2.

5.4.15 Suppose we have the situation described in Exercise 5.4.4, and we take a simple random sample of size n from Π where $|\Pi| = N$.

(a) Prove that the mean of $\hat{f}_X(0)$ is given by $f_X(0)$. (Hint: Note that we can write $\hat{f}_X(0) = n^{-1} \sum_{i=1}^n I_{\{0\}}(X(\pi_i))$ and $I_{\{0\}}(X(\pi_i)) \sim$ Bernoulli$(f_X(0))$.)

(b) Prove that the variance of $\hat{f}_X(0)$ is given by

$$\frac{f_X(0)\,(1 - f_X(0))}{n} \frac{N - n}{N - 1}. \tag{5.4.1}$$

(Hint: Use the hint in part (a), but note that the $I_{\{0\}}(X(\pi_i))$ are not independent. Use Theorem 3.3.4(b) and evaluate $\mathrm{Cov}\left(I_{\{0\}}(X(\pi_i)), I_{\{0\}}(X(\pi_i))\right)$ in terms of $f_X(0)$.)

(c) Repeat the calculations in parts (a) and (b), but this time assume that you take a sample of n with replacement. (Hint: Use Exercise 5.4.4(c).)

(d) Explain why the factor $(N - n)/(N - 1)$ in (5.4.1) is called the *finite sample correction factor*.

5.4.16 Suppose we have a finite population Π and we do not know $|\Pi| = N$. In addition, suppose we have a measurement variable $X : \Pi \to \{0, 1\}$ and we know that $N f_X(0) = a$ where a is known. Based on a simple random sample of n from Π, determine an estimator of N. (Hint: Use a function of $\hat{f}_X(0)$.)

5.4.17 Suppose that X is a quantitative variable defined on a population Π and that we take a simple random sample of size n from Π.

(a) If we estimate the population mean μ_X by the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X(\pi_i)$, prove that $E(\bar{X}) = \mu_X$ where μ_X is defined in Problem 5.4.14(a). (Hint: What is the distribution of each $X(\pi_i)$?)

(b) Under the assumption that i.i.d. sampling makes sense, show that the variance of $\bar{X}$ equals σ_X^2/n, where σ_X^2 is defined in Problem 5.4.14(b).

5.4.18 Suppose we have a finite population Π and we do not know $|\Pi| = N$. In addition, suppose we have a measurement variable $X : \Pi \to R^1$ and we know $T = \sum_{\pi} X(\pi)$. Based on a simple random sample of n from Π, determine an estimator of N. (Hint: Use a function of $\bar{X}$.)

5.4.19 Under i.i.d. sampling, prove that $\hat{f}_X(x) \xrightarrow{D} f_X(x)$ as $n \to \infty$. (Hint: $\hat{f}_X(x) = n^{-1}\sum_{i=1}^{n} I_{\{x\}}(X(\pi_i))$.)

CHALLENGES

5.4.20 (*Stratified sampling*) Suppose that X is a quantitative variable defined on a population Π and that we can partition Π into two subpopulations Π_1 and Π_2, such that a proportion p of the full population is in Π_1. Let f_{iX} denote the conditional population distribution of X on Π_i.

(a) Prove that $f_X(x) = p f_{1X}(x) + (1-p)\, f_{2X}(x)$.

(b) Establish that $\mu_X = p\mu_{1X} + (1-p)\,\mu_{2X}$, where μ_{iX} is the mean of X on Π_i.

(c) Establish that $\sigma_X^2 = p\sigma_{1X}^2 + (1-p)\,\sigma_{2X}^2 + p(1-p)\,(\mu_{1X} - \mu_{2X})^2$.

(d) Suppose that it makes sense to assume i.i.d. sampling whenever we take a sample from either the full population or either of the subpopulations, i.e., whenever the sample sizes we are considering are small relative to the sizes of these populations. We implement stratified sampling by taking a simple random sample of size n_i from subpopulation Π_i. We then estimate μ_X by $p\bar{X}_1 + (1-p)\,\bar{X}_2$, where $\bar{X}_i$ is the sample mean based on the sample from Π_i. Prove that $E(p\bar{X}_1 + (1-p)\,\bar{X}_2) = \mu_X$ and

$$\text{Var}\left(p\bar{X}_1 + (1-p)\,\bar{X}_2\right) = p^2\frac{\sigma_{1X}^2}{n_1} + (1-p)^2\,\frac{\sigma_{2X}^2}{n_2}.$$

(e) Under the assumptions of part (d), prove that

$$\text{Var}\left(p\bar{X}_1 + (1-p)\,\bar{X}_2\right) \leq \text{Var}\left(\bar{X}\right)$$

when $\bar{X}$ is based on a simple random sample of size n from the full population and $n_1 = pn, n_2 = (1-p)\,n$. This is called *proportional stratified sampling*.

(f) Under what conditions is there no benefit to proportional stratified sampling? What do you conclude about situations in which stratified sampling will be most beneficial?

DISCUSSION TOPICS

5.4.21 Sometimes it is argued that it is possible for a skilled practitioner to pick a more accurate representative sample of a population deterministically rather than by employing simple random sampling. This argument is based in part on the argument that it is always possible with simple random sampling that we could get a very unrepresentative sample through pure chance and that this can be avoided by an expert. Comment on this assertion.

5.4.22 Suppose it is claimed that a quantitative measurement X defined on a finite population Π is approximately distributed according to a normal distribution with unknown mean and unknown variance. Explain fully what this claim means.

5.5 Some Basic Inferences

Now suppose we are in a situation involving a measurement X, whose distribution is unknown, and we have obtained the data $(x_1, x_2, \ldots, x_n)$, i.e., observed n values of X. Hopefully, these data were the result of simple random sampling, but perhaps they were collected as part of an observational study. Denote the associated unknown population relative frequency function, or an approximating density, by f_X and the population distribution function by F_X.

What we do now with the data depends on two things. First, we have to determine what we want to know about the underlying population distribution. Typically, our interest is in only a few characteristics of this distribution — the mean and variance. Second, we have to use statistical theory to combine the data with the statistical model to make inferences about the characteristics of interest.

We now discuss some typical characteristics of interest and present some informal estimation methods for these characteristics, known as *descriptive statistics*. These are often used as a preliminary step before more formal inferences are drawn and are justified on simple intuitive grounds. They are called descriptive because they are estimating quantities that *describe* features of the underlying distribution.

5.5.1 Descriptive Statistics

Statisticians often focus on various characteristics of distributions. We present some of these in the following examples.

EXAMPLE 5.5.1 *Estimating Proportions and Cumulative Proportions*
Often we want to make inferences about the value $f_X(x)$ or the value $F_X(x)$ for a specific x. Recall that $f_X(x)$ is the proportion of population members whose X measurement equals x. In general, $F_X(x)$ is the proportion of population members whose X measurement is less than or equal to x.

Now suppose we have a sample $(x_1, x_2, \ldots, x_n)$ from f_X. A natural estimate of $f_X(x)$ is given by $\hat{f}_X(x)$, the proportion of sample values equal to x. A natural estimate of $F_X(x)$ is given by $\hat{F}_X(x) = n^{-1} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$, the proportion of sample values less than or equal to x, otherwise known as the empirical distribution function evaluated at x.

Suppose we obtained the following sample of $n = 10$ data values.

1.2	2.1	0.4	3.3	−2.1	4.0	−0.3	2.2	1.5	5.0

In this case, $\hat{f}_X(x) = 0.1$ whenever x is a data value and is 0 otherwise. To compute $\hat{F}_X(x)$, we simply count how many sample values are less than or equal to x and divide by $n = 10$. For example, $\hat{F}_X(-3) = 0/10 = 0$, $\hat{F}_X(0) = 2/10 = 0.2$, and $\hat{F}_X(4) = 9/10 = 0.9$. ∎

An important class of characteristics of the distribution of a quantitative variable X is given by the following definition.

Definition 5.5.1 For $p \in [0, 1]$, the *pth quantile* (or $100p$*th percentile*) x_p, for the distribution with cdf F_X, is defined to be the smallest number x_p satisfying $p \leq F_X(x_p)$.

For example, if your mark on a test placed you at the 90th percentile, then your mark equals $x_{0.9}$ and 90% of your fellow test takers achieved your mark or lower. Note that by the definition of the inverse cumulative distribution function (Definition 2.10.1), we can write $x_p = F_X^{-1}(p) = \min\{x : p \leq F_X(x)\}$.

When F_X is strictly increasing and continuous, then $F_X^{-1}(p)$ is the unique value x_p satisfying

$$F_X(x_p) = p. \tag{5.5.1}$$

Figure 5.5.1 illustrates the situation in which there is a unique solution to (5.5.1). When F_X is not strictly increasing or continuous (as when X is discrete), then there may be more than one, or no, solutions to (5.5.1). Figure 5.5.2 illustrates the situation in which there is no solution to (5.5.1).

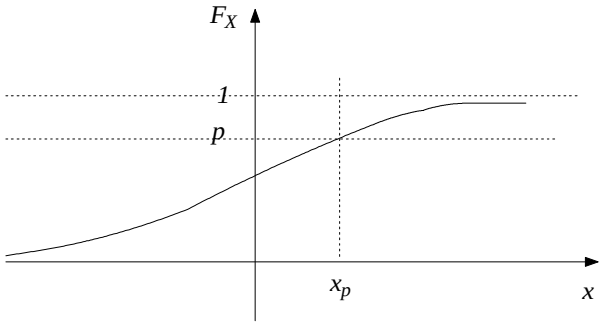

Figure 5.5.1: The pth quantile x_p when there is a unique solution to (5.5.1).

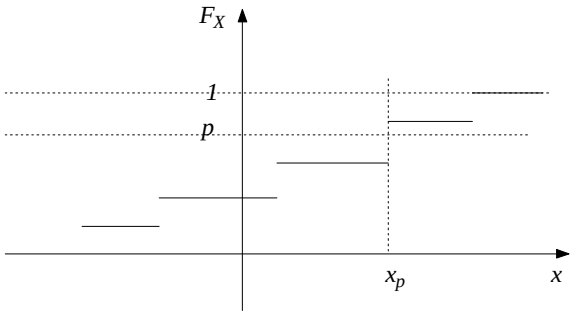

Figure 5.5.2: The pth quantile x_p determined by a cdf F_X when there is no solution to (5.5.1).

So, when X is a continuous measurement, a proportion p of the population have their X measurement less than or equal to x_p. As particular cases, $x_{0.5} = F_X^{-1}(0.5)$ is the *median*, while $x_{0.25} = F_X^{-1}(0.25)$ and $x_{0.75} = F_X^{-1}(0.75)$ are the first and third *quartiles*, respectively, of the distribution.

EXAMPLE 5.5.2 *Estimating Quantiles*
A natural estimate of a population quantile $x_p = F_X^{-1}(p)$ is to use $\hat{x}_p = \hat{F}_X^{-1}(p)$. Note, however, that $\hat{F}_X$ is not continuous, so there may not be a solution to (5.5.1) using $\hat{F}_X$.

Applying Definition 5.5.1, however, leads to the following estimate. First, order the observed sample values $(x_1, \ldots, x_n)$ to obtain the *order statistics* $x_{(1)} < \cdots < x_{(n)}$ (see Section 2.8.4). Then, note that $x_{(i)}$ is the (i/n)-th quantile of the empirical distribution, because

$$\hat{F}_X(x_{(i)}) = \frac{i}{n}$$

and $\hat{F}_X(x) < i/n$ whenever $x < x_{(i)}$. In general, we have that the *sample pth quantile* is $\hat{x}_p = x_{(i)}$ whenever

$$\frac{i-1}{n} < p \leq \frac{i}{n}. \tag{5.5.2}$$

A number of modifications to this estimate are sometimes used. For example, if we find i such that (5.5.2) is satisfied and put

$$\tilde{x}_p = x_{(i-1)} + n\left(x_{(i)} - x_{(i-1)}\right)\left(p - \frac{i-1}{n}\right), \tag{5.5.3}$$

then $\tilde{x}_p$ is the linear interpolation between $x_{(i-1)}$ and $x_{(i)}$. When n is even, this definition gives the *sample median* as $\tilde{x}_{0.5} = x_{(n/2)}$; a similar formula holds when n is odd (Problem 5.5.21). Also see Problem 5.5.22 for more discussion of (5.5.3).

Quite often the sample median is defined to be

$$\breve{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & n \text{ odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & n \text{ even,} \end{cases} \tag{5.5.4}$$

namely, the middle value when n is odd and the average of the two middle values when n is even. For n large enough, all these definitions will yield similar answers. The use of any of these is permissible in an application.

Consider the data in Example 5.5.1. Sorting the data from smallest to largest, the order statistics are given by the following table.

$x_{(1)} = -2.1$	$x_{(2)} = -0.3$	$x_{(3)} = 0.4$	$x_{(4)} = 1.2$	$x_{(5)} = 1.5$
$x_{(6)} = 2.1$	$x_{(7)} = 2.2$	$x_{(8)} = 3.3$	$x_{(9)} = 4.0$	$x_{(10)} = 5.0$

Then, using (5.5.3), the sample median is given by $\tilde{x}_{0.5} = x_{(5)} = 1.5$, while the sample quartiles are given by

$$\begin{aligned} \tilde{x}_{0.25} &= x_{(2)} + 10\left(x_{(3)} - x_{(2)}\right)(0.25 - 0.2) \\ &= -0.3 + 10\,(0.4 - (-0.3))\,(0.25 - 0.2) = 0.05 \end{aligned}$$

and

$$\begin{aligned} \tilde{x}_{0.75} &= x_{(7)} + 10\left(x_{(8)} - x_{(7)}\right)(0.75 - 0.7) \\ &= 2.2 + 10\,(3.3 - 2.2)\,(0.75 - 0.7) = 2.75. \end{aligned}$$

So in this case, we estimate that 25% of the population under study has an X measurement less than 0.05, etc. ∎

EXAMPLE 5.5.3 *Measuring Location and Scale of a Population Distribution*
Often we are asked to make inferences about the value of the *population mean*

$$\mu_X = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} X(\pi)$$

and the *population variance*

$$\sigma_X^2 = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} (X(\pi) - \mu_X)^2,$$

where Π is a finite population and X is a real-valued measurement defined on it. These are measures of the location and spread of the population distribution about the mean, respectively. Note that calculating a mean or variance makes sense only when X is a quantitative variable.

When X is discrete, we can also write

$$\mu_X = \sum_x x f_X(x)$$

because $|\Pi|\, f_X(x)$ equals the number of elements $\pi \in \Pi$ with $X(\pi) = x$. In the continuous case, using an approximating density f_X, we can write

$$\mu_X \approx \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Similar formulas exist for the population variance of X (see Problem 5.4.14).

It will probably occur to you that a natural estimate of the population mean μ_X is given by the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Also, a natural estimate of the population variance σ_X^2 is given by the *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 . \tag{5.5.5}$$

Later we will explain why we divided by $n-1$ in (5.5.5) rather than n. Actually, it makes little difference which we use, for even modest values of n. The *sample standard deviation* is given by s, the positive square root of s^2. For the data in Example 5.1.1, we obtain $\bar{x} = 1.73$ and $s = 2.097$.

The population mean μ_X and population standard deviation σ_X serve as a pair, in which μ_X measures where the distribution is located on the real line and σ_X measures how much spread there is in the distribution about μ_X. Clearly, the greater the value of σ_X, the more variability there is in the distribution.

Alternatively, we could use the population median $x_{0.5}$ as a measure of location of the distribution and the *population interquartile range* $x_{0.75} - x_{0.25}$ as a measure of the amount of variability in the distribution around the median. The median and interquartile range are the preferred choice to measure these aspects of the distribution whenever the distribution is *skewed*, i.e., not symmetrical. This is because the median is insensitive to very extreme values, while the mean is not. For example, house prices in an area are well known to exhibit a right-skewed distribution. A few houses selling for very high prices will not change the median price but could result in a big change in the mean price.

When we have a symmetric distribution, the mean and median will agree (provided the mean exists). The greater the skewness in a distribution, however, the greater will be the discrepancy between its mean and median. For example, in Figure 5.5.3 we have plotted the density of a χ^2 (4) distribution. This distribution is skewed to the right, and the mean is 4 while the median is 3.3567.

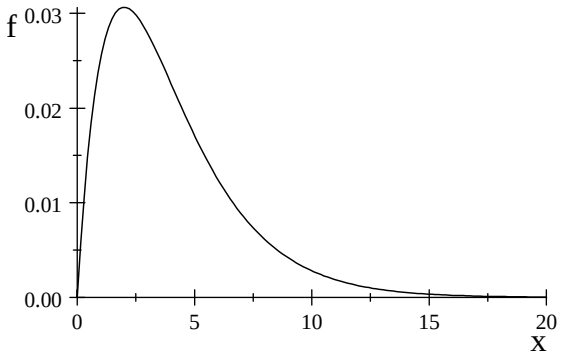

Figure 5.5.3: The density f of a χ^2 (4) distribution.

We estimate the population interquartile range by the *sample interquartile range* (IQR) given by $I\hat{Q}R = \tilde{x}_{0.75} - \tilde{x}_{0.25}$. For the data in Example 5.5.1, we obtain the sample median to be $\tilde{x}_{0.5} = 1.5$, while $I\hat{Q}R = 2.75 - 0.05 = 2.70$.

If we change the largest value in the sample from $x_{(10)} = 5.0$ to $x_{(10)} = 500.0$ the sample median remains $\tilde{x}_{0.5} = 1.5$, but note that the sample mean goes from 1.73 to 51.23! ∎

5.5.2 Plotting Data

It is always a good idea to plot the data. For discrete quantitative variables, we can plot $\hat{f}_X$, i.e., plot the sample proportions (relative frequencies). For continuous quantitative variables, we introduced the density histogram in section 5.4.3. These plots give us some idea of the shape of the distribution from which we are sampling. For example, we can see if there is any evidence that the distribution is strongly skewed.

We now consider another very useful plot for quantitative variables.

EXAMPLE 5.5.4 *Boxplots and Outliers*
Another useful plot for quantitative variables is known as a *boxplot*. For example, Figure 5.5.4 gives a boxplot for the data in Example 5.5.1. The line in the center of the box is the median. The line below the median is the first quartile, and the line above the median is the third quartile.

The vertical lines from the quartiles are called *whiskers*, which run from the quartiles to the *adjacent values*. The adjacent values are given by the greatest value less than or equal to the *upper limit* (the third quartile plus 1.5 times the $I\hat{Q}R$) and by the least value greater than or equal to the *lower limit* (the first quartile minus 1.5 times the $I\hat{Q}R$). Values beyond the adjacent values, when these exist, are plotted with a $*$; in this case, there are none. If we changed $x_{(10)} = 5.0$ to $x_{(10)} = 15.0$, however, we see this extreme value plotted as a $*$, as shown in Figure 5.5.5.

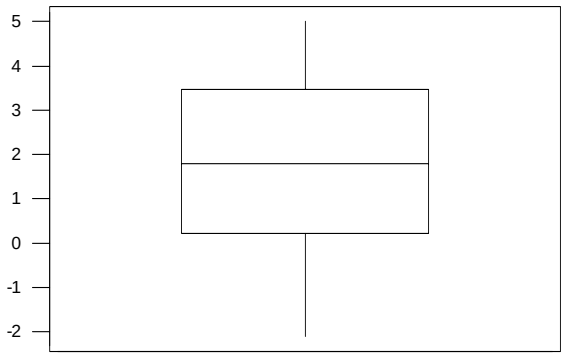

Figure 5.5.4: A boxplot of the data in Example 5.5.1.

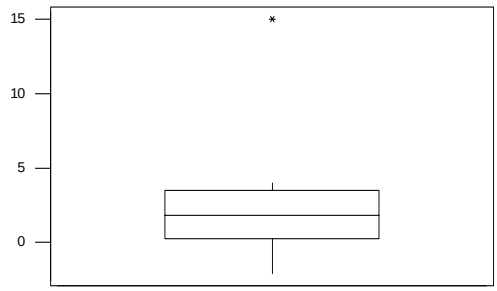

Figure 5.5.5: A boxplot of the data in Example 5.5.1, changing $x_{(10)} = 5.0$ to $x_{(10)} = 15.0$.

Points outside the upper and lower limits, and thus plotted by $*$, are commonly referred to as *outliers*. An outlier is a value that is extreme with respect to the rest of the observations. Sometimes outliers occur because a mistake has been made in collecting or recording the data, but they also occur simply because we are sampling from a long-tailed distribution. It is often difficult to ascertain which is the case in a particular application, but each such observation should be noted. We have seen in Example 5.5.3 that outliers can have a big impact on statistical analyses. Their effects should be recorded when reporting the results of a statistical analysis. ∎

For categorical variables, it is typical to plot the data in a bar chart, as described in the next example.

EXAMPLE 5.5.5 *Bar Charts*
For categorical variables, we code the values of the variable as equispaced numbers and then plot constant-width rectangles (the bars) over these values so that the height of the rectangle over a value equals the proportion of times that value is assumed. Such a plot is called a *bar chart*. Note that the values along the x-axis are only labels and not to be treated as numbers that we can do arithmetic on, etc.

For example, suppose we take a simple random sample of 100 students and record their favorite flavor of ice cream (from amongst four possibilities), obtaining the results given in the following table.

Flavor	Count	Proportion
Chocolate	42	0.42
Vanilla	28	0.28
Butterscotch	22	0.22
Strawberry	8	0.08

Coding Chocolate as 1, Vanilla as 2, Butterscotch as 3, and Strawberry as 4, Figure 5.5.6 presents a bar chart of these data. It is typical for the bars in these charts not to touch. ∎

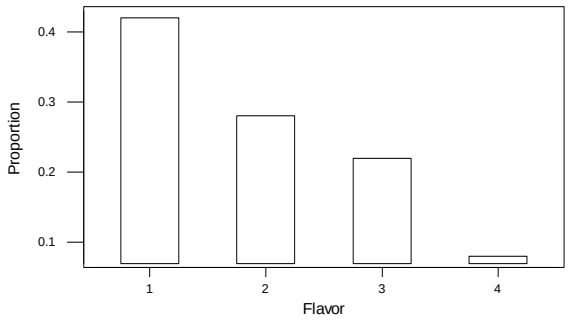

Figure 5.5.6: A bar chart for the data of Example 5.5.5.

5.5.3 Types of Inference

Certainly quoting descriptive statistics and plotting the data are methods used by a statistician to try to learn something about the underlying population distribution. There are difficulties with this approach, however, as we have just chosen these methods based on intuition. Often it is not clear which descriptive statistics we should use. Furthermore, these data summaries make no use of the information we have about the true population distribution as expressed by the statistical model, namely, $f_X \in \{f_\theta : \theta \in \Omega\}$. Taking account of this information leads us to develop a theory of statistical inference, i.e., to specify how we should combine the model information together with the data to make inferences about population quantities. We will do this in Chapters 6, 7, and 8, but first we discuss the types of inferences that are commonly used in applications.

In Section 5.2, we discussed three types of inference in the context of a known probability model as specified by some density or probability function f. We noted that we might want to do any of the following concerning an unobserved response value s.

(i) Predict an unknown response value s via a prediction t.

(ii) Construct a subset C of the sample space S that has a high probability of containing an unknown response value s.

(iii) Assess whether or not $s_0 \in S$ is a plausible value from the probability distribution specified by f.

We refer to (i), (ii), and (iii) as *inferences* about the unobserved s. The examples of Section 5.2 show that these are intuitively reasonable concepts.

In a statistical application, we do not know f; we know only that $f \in \{f_\theta : \theta \in \Omega\}$, and we observe the data s. We are uncertain about which candidate f_θ is correct, or, equivalently, which of the possible values of θ is correct.

As mentioned in Section 5.5.1, our primary goal may be to determine not the true f_θ, but some characteristic of the true distribution such as its mean, median, or the

value of the true distribution function F at a specified value. We will denote this characteristic of interest by $\psi(\theta)$. For example, when the characteristic of interest is the mean of the true distribution of a continuous random variable, then

$$\psi(\theta) = \int_{-\infty}^{\infty} x f_\theta(x)\, dx.$$

Alternatively, we might be interested in $\psi(\theta) = F_\theta^{-1}(0.5)$, the median of the distribution of a random variable with distribution function given by F_θ.

Different values of θ lead to possibly different values for the characteristic $\psi(\theta)$. After observing the data s, we want to make inferences about what the correct value is. We will consider the three types of inference for $\psi(\theta)$.

(i) Choose an estimate $T(s)$ of $\psi(\theta)$, referred to as the *problem of estimation.*

(ii) Construct a subset $C(s)$ of the set of possible values for $\psi(\theta)$ that we believe contains the true value, referred to as the problem of *credible region* or *confidence region* construction.

(iii) Assess whether or not ψ_0 is a plausible value for $\psi(\theta)$ after having observed s, referred to as the problem of *hypothesis assessment*.

So estimates, credible or confidence regions, and hypothesis assessment are examples of types of inference. In particular, we want to construct estimates $T(s)$ of $\psi(\theta)$, construct credible or confidence regions $C(s)$ for $\psi(\theta)$, and assess the plausibility of a hypothesized value ψ_0 for $\psi(\theta)$.

The *problem of statistical inference* entails determining how we should combine the information in the model $\{f_\theta : \theta \in \Omega\}$ and the data s to carry out these inferences about $\psi(\theta)$.

A very important statistical model for applications is the location-scale normal model introduced in Example 5.3.4. We illustrate some of the ideas discussed in this section via that model.

EXAMPLE 5.5.6 *Application of the Location-Scale Normal Model*
Suppose the following simple random sample of the heights (in inches) of 30 students has been collected.

64.9	61.4	66.3	64.3	65.1	64.4	59.8	63.6	66.5	65.0
64.9	64.3	62.5	63.1	65.0	65.8	63.4	61.9	66.6	60.9
61.6	64.0	61.5	64.2	66.8	66.4	65.8	71.4	67.8	66.3

The statistician believes that the distribution of heights in the population can be well approximated by a normal distribution with some unknown mean and variance, and she is unwilling to make any further assumptions about the true distribution. Accordingly, the statistical model is given by the family of $N(\mu, \sigma^2)$ distributions, where $\theta = (\mu, \sigma^2) \in \Omega = R^1 \times R^+$ is unknown.

Does this statistical model make sense, i.e., is the assumption of normality appropriate for this situation? The density histogram (based on 12 equal-length intervals from 59.5 to 71.5) in Figure 5.5.7 looks very roughly normal, but the extreme observation in the right tail might be some grounds for concern. In any case, we proceed as if

this assumption is reasonable. In Chapter 9, we will discuss more refined methods for assessing this assumption.

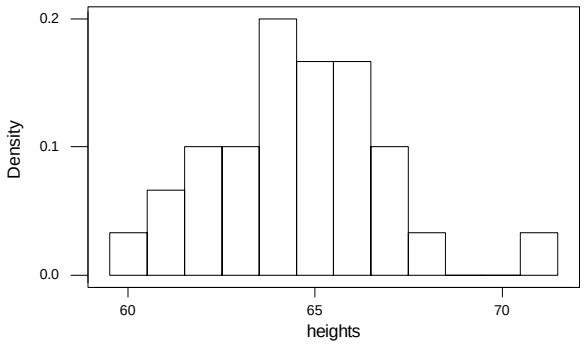

Figure 5.5.7: Density histogram of heights in Example 5.5.6.

Suppose we are interested in making inferences about the population mean height, namely, the characteristic of interest is $\psi(\mu, \sigma^2) = \mu$. Alternatively, we might want to make inferences about the 90th percentile of this distribution, i.e., $\psi(\mu, \sigma^2) = x_{0.90} = \mu + \sigma z_{0.90}$, where $z_{0.90}$ is the 90th percentile of the $N(0, 1)$ distribution (when $X \sim N(\mu, \sigma^2)$, then $P(X \leq \mu + \sigma z_{0.90}) = P((X - \mu)/\sigma \leq z_{0.90}) = \Phi(z_{0.90}) = 0.90$). So 90% of the population under study have height less than $x_{0.90}$, a value unknown to us because we do not know the value of (μ, σ^2). Obviously, there are many other characteristics of the true distribution about which we might want to make inferences.

Just using our intuition, $T(x_1, \ldots, x_n) = \bar{x}$ seems like a sensible estimate of μ and $T(x_1, \ldots, x_n) = \bar{x} + sz_{0.90}$ seems like a sensible estimate of $\mu + \sigma z_{0.90}$. To justify the choice of these estimates, we will need the theories developed in later chapters. In this case, we obtain $\bar{x} = 64.517$, and from (5.5.5) we compute $s = 2.379$. From Table D.2 we obtain $z_{0.90} = 1.2816$, so that

$$\bar{x} + sz_{0.90} = 64.517 + 2.379\,(1.2816) = 67.566.$$

How accurate is the estimate $\bar{x}$ of μ? A natural approach to answering this question is to construct a credible interval, based on the estimate, that we believe has a high probability of containing the true value of μ and is as short as possible. For example, the theory in Chapter 6 leads to using confidence intervals for μ, of the form

$$[\bar{x} - sc, \bar{x} + sc]$$

for some choice of the constant c. Notice that $\bar{x}$ is at the center of the interval. The theory in Chapter 6 will show that, in this case, choosing $c = 0.3734$ leads to what is known as a 0.95-confidence interval for μ. We then take the half-length of this interval, namely,

$$sc = 2.379\,(0.373\,4) = 0.888,$$

as a measure of the accuracy of the estimate $\bar{x} = 64.517$ of μ. In this case, we have enough information to say that we know the true value of μ to within one inch, at least with "confidence" equal to 0.95.

Finally, suppose we have a hypothesized value μ_0 for the population mean height. For example, we may believe that the mean height of the population of individuals under study is the same as the mean height of another population for which this quantity is known to equal $\mu_0 = 65$. Then, based on the observed sample of heights, we want to assess whether or not the value $\mu_0 = 65$ makes sense. If the sample mean height $\bar{x}$ is far from μ_0, this would seem to be evidence against the hypothesized value. In Chapter 6, we will show that we can base our assessment on the value of

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{64.517 - 65}{2.379/\sqrt{30}} = -1.112.$$

If the value of $|t|$ is very large, then we will conclude that we have evidence against the hypothesized value $\mu_0 = 65$. We have to prescribe what we mean by *large* here, and we will do this in Chapter 6. It turns out that $t = -1.112$ is a plausible value for t, when the true value of μ equals 65, so we have no evidence against the hypothesis. ∎

Summary of Section 5.5

- Descriptive statistics represent informal statistical methods that are used to make inferences about the distribution of a variable X of interest, based on an observed sample from this distribution. These quantities summarize characteristics of the observed sample and can be thought of as estimates of the corresponding unknown population quantities. More formal methods are required to assess the error in these estimates or even to replace them with estimates having greater accuracy.
- It is important to plot the data using relevant plots. These give us some idea of the shape of the population distribution from which we are sampling.
- There are three main types of inference: estimates, credible or confidence intervals, and hypothesis assessment.

EXERCISES

5.5.1 Suppose the following data are obtained by recording X, the number of customers that arrive at an automatic banking machine during 15 successive one-minute time intervals.

2	1	3	2	0	1	4	2
0	2	3	1	0	0	4	

(a) Record estimates of $f_X(0)$, $f_X(1)$, $f_X(2)$, $f_X(3)$, and $f_X(4)$.
(b) Record estimates of $F_X(0)$, $F_X(1)$, $F_X(2)$, $F_X(3)$, and $F_X(4)$.
(c) Plot $\hat{f}_X$.
(d) Record the mean and variance.

(e) Record the median and IQR and provide a boxplot. Using the rule prescribed in Example 5.5.4, decide whether there are any outliers.

5.5.2 Suppose the following sample of waiting times (in minutes) was obtained for customers in a queue at an automatic banking machine.

15	10	2	3	1	0	4	5
5	3	3	4	2	1	4	5

(a) Record the empirical distribution function.

(b) Plot $\hat{f}_X$.

(c) Record the mean and variance.

(d) Record the median and IQR and provide a boxplot. Using the rule given in Example 5.5.4, decide whether there are any outliers.

5.5.3 Suppose an experiment was conducted to see whether mosquitoes are attracted differentially to different colors. Three different colors of fabric were used and the number of mosquitoes landing on each piece was recorded over a 15-minute interval. The following data were obtained.

	Number of landings
Color 1	25
Color 2	35
Color 3	22

(a) Record estimates of $f_X(1)$, $f_X(2)$, and $f_X(3)$ where we use i for color i.

(b) Does it make sense to estimate $F_X(i)$? Explain why or why not.

(c) Plot a bar chart of these data.

5.5.4 A student is told that his score on a test was at the 90th percentile in the population of all students who took the test. Explain exactly what this means.

5.5.5 Determine the empirical distribution function based on the sample given below.

1.0	−1.2	0.4	1.3	−0.3
−1.4	0.4	−0.5	−0.2	−1.3
0.0	−1.0	−1.3	2.0	1.0
0.9	0.4	2.1	0.0	−1.3

Plot this function. Determine the sample median, the first and third quartiles, and the interquartile range. What is your estimate of $F(1)$?

5.5.6 Consider the density histogram in Figure 5.5.8. If you were asked to record measures of location and spread for the data corresponding to this plot, what would you choose? Justify your answer.

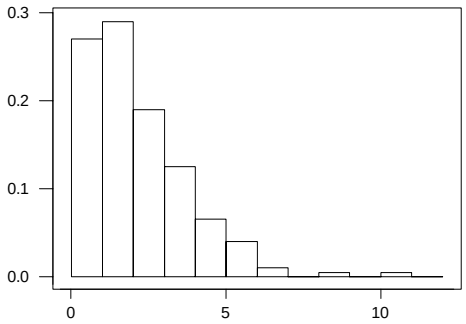

Figure 5.5.8: Density histogram for Exercise 5.5.6.

5.5.7 Suppose that a statistical model is given by the family of $N(\mu, \sigma_0^2)$ distributions where $\theta = \mu \in R^1$ is unknown, while σ_0^2 is known. If our interest is in making inferences about the first quartile of the true distribution, then determine $\psi(\mu)$.

5.5.8 Suppose that a statistical model is given by the family of $N(\mu, \sigma_0^2)$ distributions where $\theta = \mu \in R^1$ is unknown, while σ_0^2 is known. If our interest is in making inferences about the third moment of the distribution, then determine $\psi(\mu)$.

5.5.9 Suppose that a statistical model is given by the family of $N(\mu, \sigma_0^2)$ distributions where $\theta = \mu \in R^1$ is unknown, while σ_0^2 is known. If our interest is in making inferences about the distribution function evaluated at 3, then determine $\psi(\mu)$.

5.5.10 Suppose that a statistical model is given by the family of $N(\mu, \sigma^2)$ distributions where $\theta = (\mu, \sigma^2) \in R^1 \times R^+$ is unknown. If our interest is in making inferences about the first quartile of the true distribution, then determine $\psi(\mu, \sigma^2)$.

5.5.11 Suppose that a statistical model is given by the family of $N(\mu, \sigma^2)$ distributions where $\theta = (\mu, \sigma^2) \in R^1 \times R^+$ is unknown. If our interest is in making inferences about the distribution function evaluated at 3, then determine $\psi(\mu, \sigma^2)$.

5.5.12 Suppose that a statistical model is given by the family of Bernoulli(θ) distributions where $\theta \in \Omega = [0, 1]$. If our interest is in making inferences about the probability that two independent observations from this model are the same, then determine $\psi(\theta)$.

5.5.13 Suppose that a statistical model is given by the family of Bernoulli(θ) distributions where $\theta \in \Omega = [0, 1]$. If our interest is in making inferences about the probability that in two independent observations from this model we obtain a 0 and a 1, then determine $\psi(\theta)$.

5.5.14 Suppose that a statistical model is given by the family of Uniform$[0, \theta]$ distributions where $\theta \in \Omega = (0, \infty)$. If our interest is in making inferences about the coefficient of variation (see Exercise 5.3.5) of the true distribution, then determine $\psi(\theta)$. What do you notice about this characteristic?

5.5.15 Suppose that a statistical model is given by the family of Gamma(α_0, β) distributions where $\theta = \beta \in \Omega = (0, \infty)$. If our interest is in making inferences about the variance of the true distribution, then determine $\psi(\theta)$.

COMPUTER EXERCISES

5.5.16 Do the following based on the data in Exercise 5.4.5.

(a) Compute the order statistics for these data.

(b) Calculate the empirical distribution function at the data points.

(c) Calculate the sample mean and the sample standard deviation.

(d) Obtain the sample median and the sample interquartile range.

(e) Based on the histograms obtained in Exercise 5.4.5, which set of descriptive statistics do you feel are appropriate for measuring location and spread?

(f) Suppose the first data value was recorded incorrectly as 13.9 rather than as 3.9. Repeat parts (c) and (d) using this data set and compare your answers with those previously obtained. Can you draw any general conclusions about these measures? Justify your reasoning.

5.5.17 Do the following based on the data in Example 5.5.6.

(a) Compute the order statistics for these data.

(b) Plot the empirical distribution function (only at the sample points).

(c) Calculate the sample median and the sample interquartile range and obtain a boxplot. Are there any outliers?

(d) Based on the boxplot, which set of descriptive statistics do you feel are appropriate for measuring location and spread?

(e) Suppose the first data value was recorded incorrectly as 84.9 rather than as 64.9. Repeat parts (c) and (d) using this data set and see whether any observations are determined to be outliers.

5.5.18 Generate a sample of 30 from an $N(10, 2)$ distribution and a sample of 1 from an $N(30, 2)$ distribution. Combine these together to make a single sample of 31.

(a) Produce a boxplot of these data.

(b) What do you notice about this plot?

(c) Based on the boxplot, what characteristic do you think would be appropriate to measure the location and spread of the distribution? Explain why.

5.5.19 Generate a sample of 50 from a $\chi^2(1)$ distribution.

(a) Produce a boxplot of these data.

(b) What do you notice about this plot?

(c) Based on the boxplot, what characteristic do you think would be appropriate to measure the location and spread of the distribution? Explain why.

5.5.20 Generate a sample of 50 from an $N(4, 1)$ distribution. Suppose your interest is in estimating the 90th percentile $x_{0.9}$ of this distribution and we pretend that $\mu = 4$ and $\sigma = 1$ are unknown.

(a) Compute an estimate of $x_{0.9}$ based on the appropriate order statistic.

(b) Compute an estimate based on the fact that $x_{0.9} = \mu + \sigma z_{0.9}$ where $z_{0.9}$ is the 90th percentile of the $N(0, 1)$ distribution.

(c) If you knew, or at least were willing to assume, that the sample came from a normal distribution, which of the estimates in parts (a) or (b) would you prefer? Explain why.

PROBLEMS

5.5.21 Determine a formula for the sample median, based on interpolation (i.e., using (5.5.3)) when n is odd. (Hint: Use the *least integer function* or *ceiling* $\lceil x \rceil$ = smallest integer greater than or equal to x.)

5.5.22 An alternative to the empirical distribution function is to define a distribution function $\tilde{F}$ by $\tilde{F}(x) = 0$ if $x < x_{(1)}$, $\tilde{F}(x) = 1$ if $x \geq x_{(n)}$, $\tilde{F}(x) = \hat{F}(x_{(i)})$ if $x = x_{(i)}$, and

$$\tilde{F}(x) = \hat{F}(x_{(i)}) + \frac{\hat{F}(x_{(i+1)}) - \hat{F}(x_{(i)})}{x_{(i+1)} - x_{(i)}} \left(x - x_{(i)}\right)$$

if $x_{(i)} \leq x \leq x_{(i+1)}$ for $i = 1, \ldots, n$.

(a) Show that $\tilde{F}(x_{(i)}) = \hat{F}(x_{(i)})$ for $i = 1, \ldots, n$ and is increasing from 0 to 1.

(b) Prove that $\tilde{F}$ is continuous on $\left(x_{(1)}, \infty\right)$ and right continuous everywhere.

(c) Show that, for $p \in [1/n, 1)$, the value $\tilde{x}_p$ defined in (5.5.3) is the solution to $\tilde{F}(\tilde{x}_p) = p$.

DISCUSSION TOPICS

5.5.23 Sometimes it is argued that statistics does not need a formal theory to prescribe inferences. Rather, statistical practice is better left to the skilled practitioner to decide what is a sensible approach in each problem. Comment on these statements.

5.5.24 How reasonable do you think it is for an investigator to assume that a random variable is normally distributed? Discuss the role of assumptions in scientific modelling.