CHAPTER 1

BASIC TOOLS

In which we meet the probability models and the R language.

1.1 Goals of Inference

1.1.1 Population or process?

The mathematical development framework of statistics is model-based, and utilizes specifying a probability model for the random process that generates the data. This can be a simple parametric model, such as a Normal distribution, or a complicated model incorporating many variables and allowing for dependence between observations. To the extent that the model represents the process that generated the data, it is possible to draw conclusions that can be generalized to other situations where the same process operates. As the model can only ever be an approximation, it is important (but often difficult) to gauge what sort of departures from the model will invalidate the analysis.

The analysis of complex survey samples, in contrast, usually involves a population and a sample, whereas all variables are unknown and are estimated as such. The sample is the subset of the population that is selected to represent the entire population. The selection of the sample is based on probability, and the sample is then used to make inferences about the population. The goal of the analysis is to estimate the characteristics of the entire population, and the results are generalized to other populations that are similar to the sample.

In some situations, there is a clear distinction between population and sample. However, in others, this distinction is not always clear. The purpose of the study is to understand the relationship between the characteristics of the population and the sample. The statistical analysis is used to estimate the parameters of the population and to make inferences about the population.

The University of Washington has developed a statistical methodology to provide evidence in such cases. This methodology, known as statistical analysis, is used to estimate the parameters of the population and to make inferences about the population. The methodology is based on the assumption that the difference in characteristics can be attributed to random factors, such as sampling error, and can be reduced by increasing the sample size. The methodology is designed to provide evidence in such cases, and it is useful in understanding the relationship between the population and the sample.

1.1.2 Probability sampling

The fundamental statistical concept in design-based inference is the probability sample or random sample. In everyday usage, "taking a random sample" or "sampling" refers to the process of selecting individuals from the population. However, in the context of statistical analysis, the term "sampling" refers to the selection of individuals from the population according to a probability distribution. The methodology is used to estimate the parameters of the population and to make inferences about the population. The methodology is based on the assumption that the difference in characteristics can be attributed to random factors, such as sampling error, and can be reduced by increasing the sample size. The methodology is designed to provide evidence in such cases, and it is useful in understanding the relationship between the population and the sample.
be more politically conservative. As a result, the mean age of the sample will be closer to the actual age of the US population, but the median income will be higher, and the proportion of registered Republicans will be higher than for the US population.

As long as we know the proportions of each state, this generalized version remains valid. In order to achieve a probability sample, we would have to change the number of respondents from the US and then sample 500 people from each state. This example would work for any proportions within these populations, which would be in percentiles. However, if we knew all the respondents in the US, and if we can find the number of respondents in the state, we can then use this example to achieve a probability sample.

It is important to remember that without a generalized probability sample, we cannot be sure of the results. We can only be sure of the results if we know that we have a probability sample, and that the state we are measuring is the same as the sample.

1. Remove individual in the population and then remove those probability not matching up in the samples (remove any one individual).

2. The probability we name the average of every individual who does not match up in the samples.

3. Remove pair of individuals who do not match up in the samples: average for the pair of individuals (i.e., j).

4. The probability we name the average for every pair that does not match up in the samples.

The last two properties are necessary in order to get valid population estimates, but the first two are not necessary to account for the accuracy of the estimates. If individuals vote in a population independently of one another, then these two properties would guarantee the first two, whereas these would guarantee the second two, thus it is desirable that all samples of the same proportion in the same US county would have their sample size equal to the average size of the US county. However, for the second property, we don't necessarily require the analysis of each individual sample. The enumerator would record any errors from this general description. They also must record any special characteristics.

This would be important in many ways, and the necessary properties are present only for approximations in real surveys. A list of characteristics for this sampling will include those that are not invalid and those that have been early enumerated. Some people (those who cannot) do not have a landline telephone, others may not be at home or may refuse to answer questions. We will continue ignoring these problems, but we will note the issues and add additional in Chapters 9 and 10.

1.3.2 Sampling weights

If we take a sample from a sample of 25000 people from California (with total population 35 million), then any person in California has a 1/10000 chance of being sampled, so we: 25000/35000000 = 1/10000. If the average is 1000000 Californians. If it becomes real that 2500 could not make samples on height
U.S. census bureau released a report on the number of households and population of the United States. The report states that the U.S. population is estimated to be 331.4 million people. The report also states that the median household income in the United States is $61,200. The report further states that the unemployment rate in the United States is 4.0%.

The report also states that the median age of the U.S. population is 39.5 years. The report also states that the percentage of the population that is non-Hispanic white is 63.3%, non-Hispanic black is 13.8%, Hispanic is 17.4%, Asian is 5.9%, and other races are 2.6%.

The report also provides information on the number of people living in poverty in the United States. The report states that the poverty rate in the United States is 10.5%. The report also states that the median household income for households living in poverty is $26,600.

The report also provides information on the number of people living in rural areas in the United States. The report states that 13.7% of the U.S. population lives in rural areas. The report also states that the median household income in rural areas is $46,700.

The report also provides information on the number of people living in urban areas in the United States. The report states that 86.3% of the U.S. population lives in urban areas. The report also states that the median household income in urban areas is $63,000.

Overall, the report provides a comprehensive overview of the population and economic conditions in the United States.
In standard unbiased estimation the weighted observations, if \( X_j \) in a measurement of \( \mathbf{X} \) can present \( \hat{m} \) summary as:

\[
\hat{X}_d = \frac{1}{m} \sum_{i=1}^{m} X_{d,i}
\]

Given a sample of sizes \( w \) the Horvitz-Thompson estimator \( \hat{X}_P \) for the population total \( X_P \) of \( \mathbf{X} \) is:

\[
\hat{X}_P = \sum_{i=1}^{m} \frac{w_i X_i}{\sum_{j=1}^{m} w_j}
\]

The variance estimate is:

\[
\text{Var}(\hat{X}_P) = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{w_i w_j}{\sum_{j=1}^{m} w_j} (X_i - \bar{X}_P)^2
\]

Commenting the formula for the variance estimation is less important in the applied sense, but it is useful to make two claims. The first is that the formula applies to any design, however complicated, whereas only \( w \)'s are known for the sampled observations. The second is that the formula depends on the poststratification probabilities \( w \), not just on the sampling weights, this in turn calculations in the sampling design enter the computations. Some other ways of writing the variance estimation are explored in the exercises at the end of this chapter.

Confusion of "weights": Statistical and mathematical confusions arise from the term "weights" as there are at least three different things:

- **Sampling weights**: A sampling weight of 1000 means that the observation represents 1000 individuals in the population.
- **Precision weights**: A precision (or inverse variance) weight of 1000 means that the observation has 1000 times heavier variance than an observation with a weight of 1.
- **Frequency weights**: A frequency weight of 1000 means that the sample contains 1000 identical observations and equals 1 being moved by using only one record in the data set to represent them.

In this book, weights are always sampling weights, 1/\( w \). Most statistical software that is not specifically designed for survey analysis will assume that weights are precision weights or frequency weights. Giving sampling weights to software that in expecting precision weights or frequency weights will often (but not always) give correct point estimates, but will usually give conceptually incorrect standard errors, confidence intervals, and p-values.
1.1.3 Design effects

A complex survey will not have the same standard errors for estimates as a simple random sample of the same size, but nearly unbiased estimates of standard errors are only computationally available for simple random samples. The design effect was defined by Satterwhite (1967) as the ratio of the variance of an estimate in a complex sample to the variance of the same estimate in a simple random sample (29).

If the necessary sample sizes for a given level of precision is known for a simple random sample, the sample size for a complex design can be obtained by multiplying by the design effect. While the design effect will not be known in advance, some useful guidance can be obtained by looking at design effects reported for similar surveys.

Design effects for large studies are usually greater than 1.0, implying that larger sample sizes are needed for complex designs than for a simple random sample. For example, the California Health Interview Survey reports typical design effects in the range 1.4-2.0. It may be surprising that complex designs are used if they require both larger sample sizes and special statistical methods, but as Chapter 3 demonstrates, the increased sample size can offset the costs.

The other ratio of variances that is of interest is the ratio of the variance of a complex estimate to the variance of the same estimate that would be obtained by permuting the data in a simple random sample. This ratio allows the results of an analysis to be (approximately) converted if software is not available to account for the complex design. This second ratio is sometimes called the design effect and sometimes the mis specification effect.

What is the design effect compares the variance from correct estimates in two different designs, while the mis specification effect compares correct and incorrect analyses of the same design. Although these two ratios of variances are not the same, they are often smaller for practical designs. The mis specification effect is of relatively little interest unless there are significant mis estimates for complex designs is widely available, and it will not appear further in this book.

1.2 AN INTRODUCTION TO THE DATA

Most of the examples used in this book will be based either on real surveys or on simulated surveys drawn from real populations. Some of the data sets will be quite large, by statistical standards, but the computer used to write this book is a laptop computer from 2006, so it seems safe to assume that most readers will have access to at least this level of computer power. Links to the source and documentation for all these data sets can be found on the web site for this book.

Nearly all the data are available to you in electronic form to reproduce these analyses, but some effort may be required to get them. Surveys that the United States Census Bureau collects (non-identifying, anonymized) data for released by anyone, and the estimates from these surveys used in this book are available via the book's web site and directly usable elsewhere. Access to survey data from Dataland tends to require research ethical review, so the book's web site provides instructions on where
to find the data and have it in a consistent format. The national differences partly reflect the differences in copyright policy in the two countries. In the US, the federal government places materials created in public expense in the public domain, the copyright is retained by the government.

Thus may be no problem with some of the terminology in the description of data sets, which will be described in subsequent chapters.

1.3.2.1 Household Surveys

HHS: The National Health and Nutrition Examination Surveys have been conducted by the US National Center for Health Statistics (NCHS) since 1970. They are designed to provide national data on health and disease, and can identify and evaluate risk factors. Data from the surveys include NHANES data from about 25,000 people in a multi-stage sample. These participants receive an interview and a clinical exam, and have blood samples taken. Several hundred data variables are available in the public data sets.

FHS: The Family History Survey collects information on the incidence and circumstances of parents born in the United States. It was designed to collect information needed by the Department for Health and Welfare. The survey consists of four waves, with one wave covering each of the five states. The participants are asked about their family history, including information on the number of people born in the United States and the number of people born in other countries. A new version of the survey has been made available by the US Census Bureau.

NHANES: The National Health Interview Survey, conducted by the National Center for Health Statistics, is the largest of the national health-related surveys in the United States. The National Health Survey Act (1957) provided "The continuing survey program and special studies to secure accurate and current statistical information on the amount, distribution, and effects of illness and disability in the United States and the measures needed for or because of such conditions." NHIS plans to sample about 20,000 households, totaling about 30,000 people, each year, but the survey is designed so that the results will still be useful if the sampling bias is too small for the National Health Interview Survey, conducted by the US Census Bureau, which provides the most comprehensive and accurate data available on household income and expenditures. The survey includes questions about income and expenditures.
in government support programs such as food stamps. The same households are repeatedly surveyed over time to allow economic changes to be measured more accurately.

CHIS: The California Health Interview Survey (CHIS) samples households from California by random digit dialing within geographic regions. The survey is conducted every two years and samples 40,000-50,000 households. Unlike the surveys above, which are conducted by government agencies, CHIS is conducted by the Center for Health Policy Research at the University of California, Los Angeles. CHIS asks questions about health, risk factors for disease, health insurance, and access to health care.

SHS: The National Health Interview Survey interviews about 30,000 households every two years. Individual households are sampled in demographically balanced clusters of communities in the non-institutionalized population. The survey contains a wide range of topics such as smoking, injury, transport, and social services. Data from a subset of variables have been made generally available (after some further bookkeeping to protect anonymity) by the CDC under an Affiliated University.

BHS: The Behavioral Risk Factor Surveillance System is a telephone survey of behavioral risk factors for disease. The survey is conducted by most US states using material supplied by the National Center for Health Statistics. The number of states involved increased from 15 in 1984 to all 50 in 2007 (plus the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands) and the sample size from 120,000 to 420,000. It is now the country's largest telephone survey.

1.2.3 Populations

Understanding and comparing analytical techniques requires realistic data where the true measures are known. We will use some complete populations data to create artificial probability samples and compare the results of our analyses to the population values. Population data are also useful for illustrating design and programming calculations that are shown below the survey data match the public use files.

Election Data: Voting data for the US presidential elections is available for each county. We will try to predict the result from a sample of the state and use the voting data from previous elections to improve predictions.

Wilms' tumor is a rare childhood cancer of the kidney, which occurs in about 70% of cases. Most children in the United States with Wilms' tumor participate in randomized clinical trials conducted by the National Wilms' Tumor Study Group. Data from these studies [34,35] has been used extensively in research on two-phase epidemiological studies by Prentice and co-workers, and some of the data
in more publicity. The focus is on estimating the risk of relapse after initially successful treatment.

Chapter 1: The Washington Association of Schools and Services

The Washington Association of Schools and Services collects data on suicides reported in public schools in Washington (the state and city). These data are reported biannually by public schools and city-by-city type of service.

Chapter 2: The California Academic Performance Index

In comparison with standardized test results administered to students in California schools, the addition to academic performance data that the California Department of Education mandates is a valuable gauge of school demographic variables available. These data have been used extensively to illustrate the use of survey methodology by Academic Computing Services at the University of California, Los Angeles.

Chapter 3: The University of Chicago

Primary liability studies in a very large number of cases that is executed only by implementation. However, implementation was available, the Mayo Clinic conducted a number of trials to evaluate medical care to be a completely ineffective treatment. The data from 1907 participants in the total and 1000 participants who did not participate was used to examine a model for determining variables that are still used in reanalyzing these respondents. At the Mayo Clinic, a major center for treatment of primary liability studies, these 1000 participants represent essentially the entire population in the sending areas. The data included the public version of the database created by Harry Theobald in conjunction with the development of software for statistical analysis. It has become a standard tool in simulation and research examples.

1.3 Computing in Medicine

It is probably the most widely used software in medical research and for distributing new statistical methods. The design of R is novel, especially with SOL I and SOL II, one of the common systems for interactive statistical computing. Since then, the main designer of R, received the Software Systems Award from the Association for Computing Machinery.

Now the R system, which has become adopted into people analysis, visualization, and manipulation data.

The advantages in, of course, that means of R and I have in other fields many analyses, visualizations, and manipulations using machine learning cannot always be in some ways as efficient. It has not been so perfect some clinical trials in medicine, and the programming is more flexible than, more complex examples than the interpretability of most statistical packages.

Although all these results revealed to the analysis will be presented in this book, it is not all explained in detail and sometimes in clinical trials with R would benefit from reading, in both introductory books on the language. A comprehensive list of books on R is given in this Project by the R Project.

The book [34] is written for social scientists and professionals [35] for health administrators. [36] covers more advanced programming and design philosophy for R code.
1.3.1 Obtaining R

Windows or Macintosh users can download R from the Comprehensive R Archive Network (CRAN) at the central site, [https://cran.r-project.org/](https://cran.r-project.org/). Alternatively, CRAN Mirror sites provide pre-compiled versions of R through their respective mirror sites. Linux and Unix-like systems can easily compile R from the source code available from CRAN. Please review R releases and frequently check the CRAN website for updates.

System administrators installing R for multiple users, or people wishing to compile R from the source code, should read the R Installation and Administration manuals available on CRAN.

1.3.2 Obtaining the CRAN Repository

An important feature of R is the large collection of ready-to-use packages written by users, with the number of available packages doubling about every 18 months. In particular, R includes no features but depends on other packages for many analyses and very complex analyses, all of which are installed in the CRAN repository. The CRAN repository is the largest repository of CRAN packages, including Bioconductor packages (February 1998: [CRAN](https://cran.r-project.org/)).

These packages can most easily be installed from inside R, using the `install.packages` command on the Windows version of R, or the `install.packages` command on the Macintosh version. Some packages in this book also make use of other add-on packages for graphics, importing, and database access. These will be installed in the same way.

When you use a contributed R package, remember to keep it updated. This manual is only a guide to using the packages as they are released. If you update R, you should update your packages accordingly.

The CRAN repository has many more packages than those included in this book. The CRAN repository is a comprehensive repository of CRAN packages, including Bioconductor packages (February 1998: [CRAN](https://cran.r-project.org/)).

1.4 Using R

This section provides a brief overview of getting data into R and doing some simple computations. Further introductory material on R can be found in Appendix A.

1.4.1 Reading Data into R

The simplest format for plain text data has one record per line with variable names in the first line of the file, with variables separated by commas. Lines with other information are blank. Almost all statistical packages and databases can easily export data as comma-separated text.
The National Women's Tobacco Study dataset is in this format. The file can be read into R with the function `read.csv()`. Unlike many statistical packages, it can work with multiple data sets at the same time. This means that when a data set is read in, it must be given a name so that it can be identified in the future. Naming a data set is done with the operator `<-`.

It is a good idea to check that the data have been read in correctly. One check is to compute summaries of all the variables in the data set with the summary function, although this is not such a good idea for very large data sets with hundreds of variables. Another check is to list the first few lines of the data set with the `head()` function. Coda and R output from reading the data and performing these two checks are shown in Figure 1.1. If the file is not actually in the current environment, the number of variables or their names may not be recognized by the summary function. Otherwise, simple checks are to find out the number of rows and number of columns of the data set, also shown in Figure 1.1.
1.4.2 Reading Data from Other Packages

It is common to read data saved in binary formats from SPSS and Stata, and the format produced by PROC EXPAND in SAS. RELATIV data files are now distributed in the .XPORT format, so we can read from RELATIV. The Inter-University Consortium for Political and Social Research (ICPSR) and the Panel Study of Income Dynamics (PSID) University of Pennsylvania, University of Wisconsin, and others have also provided data sets in Stata and SAS formats, saving the effort needed to construct variable names and value labels for data stored in non-standard formats.

The R function for reading data in these formats can be found in the 
library (package):

library (dplyr)

When the package is loaded, all its functions and help pages become available. The function read_csv() will read a CSV file from the system's current working directory, read_excel() from the system's current working directory, and read_sas() will read SAS .XPORT files. The read_qgis() function allows you to read QGIS shapefiles. These functions take a file name as the first argument, and read_csv(), read_excel(), and read_sas() have other options that control the handling of dates and factors.

An example of reading a CSV file from a file named RELATIV.D80:2004, data = qdps, which is an SAS .XPORT file:

```r
> dname < read_csv("/path/to/dname.csv")
> dname < read_excel("/path/to/dname.xlsx")
> dname < read_sas("/path/to/dname.sas7bdat")
```

This is a useful function to read data from various formats in the system's home directory, or the file is on the windows command line. An example of using read_csv() to read States Current Data from the California Health Interview Survey is in section 2.2.1.

In some cases, read data electronically from remote databases, but doing so may cause data in the database as described in Appendix 10.
example: 1000080000 in column 4. The variable is then defined as:

\[ \text{Variable expression} = \text{column 4} \]

where \( \text{column} \) specifies a particular data set in the column of interest.

The variable \( \text{column} \) allows multiple variables to be specified, but it is necessary to reference specific groups of variables. For example, in equation 2.5.1, changing \( \text{column} = \text{column} \) to \( \text{column} \) results in equations 2.5.2, 2.5.3, and 2.5.4 of the California Health Interview Survey data sets. A data set can be partitioned into distinct groups related to common identifiers across the entire data set. The common identifiers across the entire data set indicate common identifiers across the entire data set. For example, \( \text{column} = \text{column} \) results in equation 2.5.1, and \( \text{column} = \text{column} \) results in equation 2.5.2. Variables in the reduced data set will have the reduced form for the reduced data set, and variables in the entire data set will have the entire data set form. For example, \( \text{column} = \text{column} \) results in equation 2.5.3, and \( \text{column} = \text{column} \) results in equation 2.5.4. Variables in the reduced data set can be analyzed in the entire data set with the reduced data set form. This example, to create a variable in the entire data set with the reduced data set form.
1.1. It is decided to run a clinical trial to test a new drug. The researchers want to estimate the effect of the drug on blood pressure. They want to enroll 500 patients from 5 different hospitals in the trial.

1.2. a) Write through the introductory procedures for the 5 hospitals that the researchers want to include in the study.

1.3. b) Write why the study protocol of a new drug would (indistinguishably) be 1/100 more efficient in a randomized trial than a non-randomized design in a longitudinal study. Assuming that convergence rates in given that computer simulations are accurate, why are these results not a probability example of

(a) what?

(b) incidence of the new drug?

(c) incidence of the new drug in clinical research?

1.4. You are conducting a survey that will estimate the percentage of women who used fossil fuels as their main source of energy last year. The survey was conducted on 500 women in a city with different socioeconomic backgrounds. The survey results are not expected to be representative of the general population, as they excluded a significant number of women from low-income backgrounds. Women were asked if they used fossil fuels as their main source of energy last year.

(a) If you were to choose a larger or smaller sample size than 500 for a representative design, how should this be achieved?

(b) Approximately what sample size would you need to get a reasonable precision?

1.5. Systematic sampling is often used in surveys to get a representative sample. For example, every 100th person in a list is selected.

(a) Why is this the only necessary precaution for a probability sample whose precision increases?

(b) How systematic sampling with a random start, where the procedure was used, led to a less consistent random sampling plan? How would these under every 100th person every 200th person every 300th person? Which of the necessary precautions of the probability sample during this process are intact?

(c) How systematic sampling with a random start was superior? How would these under every 100th person every 200th person every 300th person? Which of these necessary precautions would these process have left intact?

(d) Of the list that were selected into random number generation a systematic sample were random, which of the precautions would these process have left intact?

(e) Instead of a systematic sample set if it were a multiple random samples each given general results? Why would this be better?

1.6. Why cannot all the sampling possibilities be seen in a valid population estimate?

1.7. Why cannot all the pairwise probabilities be seen even to get a valid uncertainty estimate?
1.53 A probability design assumes that people who are sampled will actually be
included in the survey, whereas this technology, if used, may underestimate the true
population of interest. How could this approach cause bias in the recent years of
MARRIAGE and DIVORCE?

1.59 In a telephone survey using random digit dialing, telephone numbers are sam-
piled with equal probability from a list. When a household is encountered, why is it
necessary to ask how many telephone numbers are in the household, and what should be done
with this information in computing the sampling weights?

1.60 a) Derive the Horvitz-Thompson variance estimation for the total, as follows:

\[ \text{Var} [\hat{Y}_{xy}] = \sum_{i=1}^{R} \sum_{j=1}^{R} \text{Var} [\hat{y}_{ij}] \]

b) Show that \( \text{Var} [\hat{Y}_{xy}] = \text{Var} [\hat{y}_{ij}] \) and that \( \text{Cov} [\hat{Y}_{x}, \hat{Y}_{y}] = \text{Cov} [\hat{y}_{ij}] \).

c) Show that the variance of the Horvitz-Thompson estimation is

\[ \text{Var} [\hat{Y}_{xy}] = \sum_{i=1}^{R} \sum_{j=1}^{R} \text{Var} [\hat{y}_{ij}] = \text{Var} [\hat{y}_{ij}] \]

d) Show that the previous expression simplifies to equation 1.20.

1.61 a) Another way to write the Horvitz-Thompson variance estimation is

\[ \text{Var} [\hat{Y}_{xy}] = \sum_{i=1}^{R} \sum_{j=1}^{R} \text{Var} [\hat{y}_{ij}] \]

b) Show that this is equivalent to equation 1.52.