6.1: Introduction

The Basic Statistical Model

In the basic statistical model, we have a *population* of objects of interest. The objects could be persons, families, computer chips, acres of corn. In addition, we have various measurements or *variables* defined on the objects. We select a sample from the population and record the variables of interest for each object in the sample. Here are a few examples based on the data sets in this project:

- In the M&M data, the objects are bags of M&Ms of a specified size. The variables recorded for a bag of M&Ms are net weight and the counts for red, green, blue, orange, yellow, and brown candies.
- In the cicada data, the objects are cicadas from the middle Tennessee area. The variables recorded for a cicada are body weight, wing length, wing width, body length, gender, and species.
- In Fisher's iris data, the objects are irises. The variables recorded for an iris are petal length, petal width, sepal length, sepal width, and type.
- In the Polio data set, the objects are children. Although many variables were probably recorded for a child, the two crucial variables, both binary, were whether or not the child was vaccinated, and whether or not the child contracted Polio within a certain time period.
- In the Challenger data sets, the objects are Space Shuttle launches. The variables recorded are temperature at the time of launch and various measures of O-ring erosion of the solid rocket boosters.
- In Michelson's data set, the objects are beams of light and the variable recorded is speed.
- In Pearson's data set, the objects are father-son pairs. The variables are the height of the father and the height of the son.
- In Snow's data set, the objects are persons who died of cholera. The variables record the address of the person.
- In one of the SAT data sets, the objects are states and the variables are participation rate, average SAT Math score and average SAT Verbal score.

Thus, the observed outcome of a statistical experiment (the data) has the form $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where x_i is the vector of measurements for the ith object chosen from the population. The set S of possible values of $\boldsymbol{x}$ (before the experiment is conducted) is called the *sample space*. It is literally the space of samples. Thus, although the outcome of a statistical experiment can have quite a complicated structure (a vector of vectors), the hallmark of mathematical abstraction is the ability to gray out the features that are not relevant at any particular time, to treat a complex structure as a single object. This we do with the outcome $\boldsymbol{x}$ of the experiment.

The techniques of statistics have been enormously successful; these techniques are widely used in just about every subject that deals with quantification—the natural sciences, the social sciences, law, and medicine. On the other hand, statistics has a legalistic quality and a great deal of terminology and jargon that can make the subject a bit intimidating at first. In the rest of this section, we begin discussing some of this terminology.

The Empirical Distribution

Suppose again that the data have the form $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ where x_i is the vector of measurements for the ith object chosen. The *empirical distribution* associated with $\boldsymbol{x}$ is the probability distribution that places probability $1/n$ at each x_i. Thus, if the values are distinct, the empirical distribution is the discrete uniform distribution on $\{x_1, x_2, \ldots, x_n\}$. More generally, if x occurs k times in the data, then the empirical distribution assigns probability k/n to x. Thus, every finite data set defines a probability distribution.

Statistics

Technically, a *statistic* $w = w(\boldsymbol{x})$ is an observable function of the outcome $\boldsymbol{x}$ of the experiment. That is, a statistic is a computable function defined on the sample space S. The term *observable* means that the function should not contain any unknown quantities, because we need to be able to compute the value w of the statistic from the observed data $\boldsymbol{x}$. As with the data $\boldsymbol{x}$, a statistic w may have a complicated structure; typically, w is vector valued. Indeed, the outcome $\boldsymbol{x}$ of the experiment is itself is a statistic; all other statistics are derived from $\boldsymbol{x}$.

Statistics u and v are *equivalent* if there exists a one-to-one function r from the range of u onto the range of v such that $v = r(u)$. Equivalent statistics give equivalent information about $\boldsymbol{x}$.

Statistics u and v are equivalent if and only if the following condition holds: for any $\boldsymbol{x} \in S$ and $\boldsymbol{y} \in S$, $u(\boldsymbol{x}) = u(\boldsymbol{y})$ if and only if $v(\boldsymbol{x}) = v(\boldsymbol{y})$.

Equivalence really is an equivalence relation on the collection of statistics for a given statistical experiment. That is, if u, v, and w are arbitrary statistics then

1. u is equivalent to u (the *reflexive property*).
2. If u is equivalent to v then v is equivalent to u (the *symmetric property*).
3. If u is equivalent to v and v is equivalent to w then u is equivalent to w (the *transitive property*).

Descriptive and Inferential Statistics

There are two broad branches of statistics. The term *descriptive statistics* refers to methods for summarizing and displaying the observed data $\boldsymbol{x}$. As the name suggests, the methods of descriptive statistics usually involve computing various statistics (in the technical sense) that give useful information about the data: measures of center and spread, measures of association, and so forth. In the context of descriptive statistics, the term *parameter* refers to a characteristic of the entire population.

The deeper and more useful branch of statistics is known as *inferential statistics*. Our point of view in this branch is that the statistical experiment (before it is conducted) is a random experiment with a probability measure $\mathbb{P}$ on an underlying sample space. Thus, the outcome $\boldsymbol{x}$ of the experiment is an observed value of a random variable $\boldsymbol{X}$ defined on this probability space, with the distribution of $\boldsymbol{X}$ not completely known to us. Our goal is to draw inferences about the distribution of $\boldsymbol{X}$ from the observed value $\boldsymbol{x}$. Thus, in a sense, inferential statistics is the dual of probability. In probability, we try to *predict* the value of $\boldsymbol{X}$ *assuming* complete knowledge of the distribution. In statistics, by contrast, we *observe* the value of $\boldsymbol{x}$ of the random variable $\boldsymbol{X}$ and try to *infer* information about the underlying distribution of $\boldsymbol{X}$. In inferential statistics, a *statistic* (a function of $\boldsymbol{X}$) is itself a random variable with a distribution of its own. On the other hand, the term *parameter* refers to a characteristic of the distribution of $\boldsymbol{X}$. Often the inferential problem is to use various statistics to *estimate* or *test hypotheses* about a parameter. Another way to think of inferential statistics is that we are trying to infer from the *empirical* distribution associated with the observed data $\boldsymbol{x}$ to the *true* distribution associated with $\boldsymbol{X}$.

There are two basic types of random experiments in the general area of inferential statistics. A *designed experiment*, as the name suggests, is carefully designed to study a particular inferential question. The experimenter has considerable control over how the objects are selected, what variables are to be recorded for these objects, and the values of certain of the variables. In an *observational study*, by contrast, the researcher has little control over these factors. Often the researcher is simply given the data set and asked to make sense out of it. For example, the Polio field trials were designed experiments to study the effectiveness of the Salk vaccine. The researchers had considerable control over how the children were selected, and how the children were assigned to the treatment and control groups. By contrast, the Challenger data sets used to explore the relationship between temperature and O-ring erosion are observational studies. Of course, just because an experiment is designed does not mean that it is *well* designed.

Difficulties

A number of difficulties can arise when trying to explore an inferential question. Often, problems arise because of *confounding variables*, which are variables that (as the name suggests) interfere with our understanding of the inferential question. In the first Polio field trial design, for example, *age* and *parental consent* are two confounding variables that interfere with the determination of the effectiveness of the vaccine. The entire point of the Berkeley admissions data, to give another example, is to illustrate how a confounding variable (*department*) can create a spurious correlation between two other variables (*gender* and *admissions status*). When we correct for the interference caused by a confounding variable, we say that we have *controlled* for the variable.

Problems also frequently arise because of *measurement errors*. Some variables are inherently difficult to measure, and systematic bias in the measurements can interfere with our understanding of the inferential question. The first Polio field trial design again provides a good example. Knowledge of the vaccination status of the children led to systematic bias by doctors attempting to diagnose polio in these children. Measurement errors are sometimes caused by hidden confounding variables.

Confounding variables and measurement errors abound in political polling, where the inferential question is who will win an election. How do confounding variables such as race, income, age, and gender (to name just a few) influence how a person will vote? How do we know that a person will vote for whom she says she will, or if she will vote at all (measurement errors)? The Literary Digest poll in the 1936 presidential election and the professional polls in the 1948 presidential election illustrate these problems.

Confouding variables, measurement errors and other causes often lead to *selection bias*, which means that the sample does not represent the population with respect to the inferential question at hand. Often randomization is used to overcome the effects of confounding variables and measurement errors.

Random Samples

The most common and important special case of the inferential statistical model occurs when the observation variable

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \tag{6.1.1}$$

is a sequence of independent and identically distributed random variables. Again, in the standard sampling model, X_i is itself a vector of measurements for the ith object in the sample, and thus, we think of $(X_1, X_2, \ldots, X_n)$ as independent copies of an underlying measurement vector X. In this case, $(X_1, X_2, \ldots, X_n)$ is said to be a *random sample* of size n from the distribution of X.

Variables

The mathematical operations that make sense for variable in a statistical experiment depend on the *type* and *level of measurement* of the variable.

Type

Recall that a real variable x is *continuous* if the possible values form an interval of real numbers. For example, the weight variable in the M&M data set, and the length and width variables in Fisher's iris data are continuous. In contrast, a *discrete variable* is one whose set of possible values forms a discrete set. For example, the counting variables in the M&M data set, the type variable in Fisher's iris data, and the denomination and suit variables in the card experiment are discrete. Continuous variables represent quantities that can, in theory, be measured to any degree of accuracy. In practice, of course, measuring devices have limited accuracy so data collected from a continuous variable are necessarily discrete. That is, there is only a finite (but perhaps very large) set of possible values that can actually be measured. So, the distinction between a discrete and continuous variable is based on what is theoretically possible, not what is actually measured. Some additional examples may help:

- A person's *age* is usually given in years. However, one can imagine age being given in months, or weeks, or even (if the time of birth is known to a sufficient accuracy) in seconds. Age, whether of devices or persons, is usually considered to be a continuous variable.
- The *price* of an item is usually given (in the US) in dollars and cents, and of course, the smallest monetary object in circulation is the penny ($0.01). However, taxes are sometimes given in mills ($0.001), and one can imagine smaller divisions of a dollar, even if there are no coins to represent these divisions. Measures of wealth are usually thought of as continuous variables.
- On the other hand, the number of persons in a car at the time of an accident is a fundamentally discrete variable.

Levels of Measurement

A real variable x is also distinguished by its *level of measurement*.

Qualitative variables simply encode *types* or *names*, and thus few mathematical operations make sense, even if numbers are used for the encoding. Such variables have the *nominal level of measurement*. For example, the type variable in Fisher's iris data is qualitative. Gender, a common variable in many studies of persons and animals, is also qualitative. Qualitative variables are almost always discrete; it's hard to imagine a continuous infinity of names.

A variable for which only *order* is meaningful is said to have the *ordinal* level of measurement; differences are not meaningful even if numbers are used for the encoding. For example, in many card games, the suits are ranked, so the suit variable has the ordinal level of measurement. For another example, consider the standard 5-point scale (terrible, bad, average, good, excellent) used to rank teachers, movies, restaurants etc.

A quantitative variable for which *difference*, but not *ratios* are meaningful is said to have the *interval* level of measurement. Equivalently, a variable at this level has a relative, rather than absolute, zero value. Typical examples are temperature (in Fahrenheit or Celsius) or time (clock or calendar).

Finally, a quantitative variable for which ratios are meaningful is said to have the *ratio* level of measurement. A variable at this level has an absolute zero value. The count and weight variables in the M&M data set, and the length and width variables in Fisher's iris data are examples.

Subsamples

In the basic statistical model, subsamples corresponding to some of the variables can be constructed by *filtering* with respect to other variables. This is particularly common when the filtering variables are qualitative. Consider the cicada data for example. We might be interested in the quantitative variables body weight, body length, wing width, and wing length *by species*, that is, separately for species 0, 1, and 2. Or, we might be interested in these quantitative variables *by gender*, that is separately for males and females.

Exercises

Study Michelson's experiment to measure the velocity of light.

1. Is this a designed experiment or an observational study?
2. Classify the velocity of light variable in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. Designed experiment
2. Continuous, interval. The level of measurement is only interval because the recorded variable is the speed of light in km/hr minus $299\,000$(to make the numbers simpler). The actual speed in km/hr is a continuous, ratio variable.

Study Cavendish's experiment to measure the density of the earth.

1. Is this a designed experiment or an observational study?
2. Classify the density of earth variable in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. Designed experiment
2. Continuous, ratio.

Study Short's experiment to measure the parallax of the sun.

1. Is this a designed experiment or an observational study?
2. Classify the parallax of the sun variable in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. Observational study
2. Continuous, ratio.

In the M&M data, classify each variable in terms of type and level of measurement.

Answer

Each color count variable: discrete, ratio; Net weight: continuous, ratio

In the Cicada data, classify each variable in terms of type and level of measurement.

Answer

Body weight, wing length, wing width, body length: continuous, ratio. Gender, type: discrete, nominal

In Fisher's iris data, classify each variable in terms of type and level of measurement.

Answer

Petal width, petal length, sepal width, sepal length: continuous, ratio. Type: discrete, nominal

Study the Challenger experiment to explore the relationship between temperature and O-ring erosion.

1. Is this a designed experiment or an observational study?

2. Classify each variable in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. Observational study
2. Temperature: continuous, interval; Erosion: continuous, ratio; Damage index: discrete, ordinal

In the Vietnam draft data, classify each variable in terms of type and level of measurement.

Answer

Birth month: discrete, interval; Birth day: discrete, interval

In the two SAT data sets, classify each variable in terms of type and level of measurement.

Answer

SAT math and verbal scores: probably continuous, ratio; State: discrete, nominal; Year: discrete, interval

Study the Literary Digest experiment to to predict the outcome of the 1936 presidential election.

1. Is this a designed experiment or an observational study?
2. Classify each variable in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. designed experiment, although poorly designed
2. State: discrete, nominal; Electoral votes: discrete, ratio; Landon count: discrete, ratio; Roosevelt count: discrete, ratio

Study the 1948 polls to predict the outcome of the presidential election between Truman and Dewey. Are these designed experiments or an observational studies?

Answer

Designed experiments, but poorly designed

Study Pearson's experiment to explore the relationship between heights of fathers and heights of sons.

1. Is this a designed experiment or an observational study?
2. Classify each variable in terms of type and level of measurement.
3. Discuss possible confounding variables.

Answer

1. Observational study
2. height of the father: continuous ratio; height of the son: continuous ratio

Study the Polio field trials.

1. Are these designed experiments or observational studies?
2. Identify the essential variables and classify each in terms of type and level of measurement.
3. Discuss possible confounding variables and problems with measurement errors.

Answer

1. designed experiments
2. vacination status: discrete, nominal; Polio status: discrete, nominal

Identify the parameters in each of the following:

1. Buffon's Coin Experiment
2. Buffon's Needle Experiment
3. the Bernoulli trials model

4. the Poisson model

Answer

1. radius of the coin
2. length of the needle
3. probability of success
4. rate of arrivals

Note the parameters for each of the following families of special distributions:

1. the normal distribution
2. the gamma distribution
3. the beta distribution
4. the Pareto distribution
5. the Weibull distribution

Answer

1. mean μ and standard deviation σ
2. shape parameter k and scale parameter b
3. left parameter a and right parameter b
4. shape parameter a and scale parameter b
5. shape parameter k and scale parameter b

During World War II, the Allies recorded the serial numbers of captured German tanks. Classify the underlying *serial number* variable by type and level of measurement.

Answer

discrete, ordinal.

For a discussion of how the serial numbers were used to estimate the total number of tanks, see the section on Order Statistics in the chapter on Finite Sampling Models.