Lectures on Statistics

William G. Faris

December 1, 2003

Contents

Chapter 1

Expectation

1.1 Random variables and expectation

This chapter is a brief review of probability. We consider an experiment with a set Ω of *outcomes*. A random variable is a function from Ω to $\mathbf{R}$. Thus for each outcome ω in Ω there is a corresponding experimental number $X(\omega)$.

A probability model assigns to each positive random variable $X \geq 0$ an *expectation* (or *mean*) $E[X]$ with $0 \leq E[X] \leq \infty$. If X is a random variable that is not positive, then it is possible that the expectation is not defined. However, if $E[|X|] < \infty$, then $E[X]$ is defined, and $|E[X]| \leq E[|X|]$. In some circumstances the expectation will be called the *population mean*.

The expectation satisfies the following properties.

1. $E[aX] = aE[X]$.
2. $E[X + Y] = E[X] + E[Y]$
3. $X \leq Y$ implies $E[X] \leq E[Y]$.
4. $E[c] = c$.

The first two properties are called *linearity*. The third property is the *order* property, and the fourth property is *normalization*.

One more special but very useful class consists of the random variables for which $E[X^2] < \infty$. We shall see in a moment that for every such random variable $E[|X|]^2 \leq E[X^2]$, so this is included in the class of random variables for which $E[X]$ is defined. There is a fundamental inequality that is used over and over, the Schwarz inequality.

Theorem 1.1

$$|E[XY]| \leq \sqrt{E[X^2]}\sqrt{E[Y^2]}. \tag{1.1}$$

Proof:

Use the elementary inequality

$$\pm XY \leq \frac{1}{2}a^2 X^2 + \frac{1}{2}\frac{1}{a^2}Y^2. \tag{1.2}$$

By the order property and linearity

$$\pm E[XY] \leq \frac{1}{2}a^2 E[X^2] + \frac{1}{2}\frac{1}{a^2}E[Y^2]. \tag{1.3}$$

If $E[X^2] > 0$, then choose $a^2 = \sqrt{E[Y^2]}/\sqrt{E[X^2]}$. If $E[X^2] = 0$, then by taking a sufficiently large one sees that $\pm E[XY] = 0$.

Corollary 1.1

$$|E[X]| \leq \sqrt{E[X^2]}. \tag{1.4}$$

In probability it is common to use the *centered* random variable $X - E[X]$. This is the random variable that measures deviations from the expected value. There is a special terminology in this case. The *variance* of X is

$$\mathrm{Var}(X) = E[(X - E[X])^2]. \tag{1.5}$$

In the following we shall sometimes call this the *population variance*. Note the important identity

$$\mathrm{Var}(X) = E[X^2] - E[X]^2. \tag{1.6}$$

There is a special notation that is in standard use. The mean of X is written

$$\mu_X = E[X]. \tag{1.7}$$

The Greek mu reminds us that this is a mean. The variance of X is written

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{1.8}$$

The square root of the variance is the *standard deviation* of X. This is

$$\sigma_X = \sqrt{E[(X - \mu_X)^2]}. \tag{1.9}$$

The Greek sigma reminds us that this is a standard deviation. If we center the random variable and divided by its standard deviation, we get the *standardized* random variable

$$Z = \frac{X - \mu_X}{\sigma_X}. \tag{1.10}$$

The *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]. \tag{1.11}$$

Note the important identity

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]. \tag{1.12}$$

From the Schwarz inequality we have the following important theorem.

Theorem 1.2

$$|\operatorname{Cov}(X,Y)| \le \sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}. \tag{1.13}$$

Sometimes this is stated in terms of the *correlation coefficient*

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} \tag{1.14}$$

which is the covariance of the standardized random variables. In the following we shall sometimes call this the *population correlation coefficient*. The result is the following.

Corollary 1.2

$$|\rho(X,Y)| \le 1. \tag{1.15}$$

Perhaps the most important theorem in probability is the following. It is a trivial consequence of linearity, but it is the key to the law of large numbers.

Theorem 1.3

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X,Y) + \operatorname{Var}(Y). \tag{1.16}$$

Random variables X and Y are said to be *uncorrelated* if $\operatorname{Cov}(X,Y) = 0$. Note that this is equivalent to the identity $E[XY] = E[X]E[Y]$.

Corollary 1.3 *If X and Y are uncorrelated, then the variances add:*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y). \tag{1.17}$$

1.2 The sample mean

In statistics the *sample mean* is used to estimate the population mean.

Theorem 1.4 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{1.18}$$

be their sample mean. Then the expectation of $\bar{X}_n$ is

$$E[\bar{X}_n] = \mu. \tag{1.19}$$

Proof: The expectation of the sample mean $\bar{X}_n$ is

$$E[\bar{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}E[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}n\mu = \mu. \tag{1.20}$$

Theorem 1.5 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{1.21}$$

be their sample mean. Then the standard deviation of $\bar{X}_n$ is

$$\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{1.22}$$

Proof: The variance of the sample mean $\bar{X}_n$ is

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2. \tag{1.23}$$

We can think of these two results as a form of the weak law of large numbers. The law of large numbers is the "law of averages" that says that averaging uncorrelated random variable gives a result that is approximately constant. In this case the sample mean has expectation μ and standard deviation $\sigma/\sqrt{n}$. Thus if n is large enough, it is a random variable with expectation μ and with little variability.

The factor $1/\sqrt{n}$ is both the blessing and the curse of statistics. It is a wonderful fact, since it says that averaging reduces variability. The problem, of course, is that while $1/\sqrt{n}$ goes to zero as n gets larger, it does so rather slowly. So one must somehow obtain a quite large sample in order to ensure rather moderate variability.

The reason the law is called the weak law is that it gives a statement about a fixed large sample size n. There is another law called the strong law that gives a corresponding statement about what happens for all sample sizes n that are sufficiently large. Since in statistics one usually has a sample of a fixed size n and only looks at the sample mean for this n, it is the more elementary weak law that is relevant to most statistical situations.

1.3 The sample variance

The sample mean

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1.24}$$

is a random variable that may be used to estimate an unknown population mean μ. In the same way, the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{n-1} \tag{1.25}$$

may be used to estimate an unknown population variance σ^2.

The $n-1$ in the denominator seems strange. However it is due to the fact that while there are n observations X_i, their deviations from the sample mean $X_i - \bar{X}_n$ sum to zero, so there are only $n-1$ quantities that can vary independently. The following theorem shows how this choice of denominator makes the calculation of the expectation give a simple answer.

Theorem 1.6 *Let $X_1, X_2, X_3, \ldots X_n$ be random variables, each with mean μ and standard deviation $\sigma < \infty$. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let s^2 be the sample variance. Then the expectation of s^2 is*

$$E[s^2] = \sigma^2. \tag{1.26}$$

Proof: Compute

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X}_n+\bar{X}_n-\mu)^2 = \sum_{i=1}^{n}(X_i-\bar{X}_n)^2+\sum_{i=1}^{n}(\bar{X}_n-\mu)^2. \tag{1.27}$$

Notice that the cross terms sum to zero. Take expectations. This gives

$$n\sigma^2 = E[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2)] + n\frac{\sigma^2}{n}. \tag{1.28}$$

The result then follows from elementary algebra.

1.4 The central limit theorem

Random variables X, Y are called *independent* if for all functions g and h we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]. \tag{1.29}$$

Clearly independent random variables are uncorrelated. The notion of independence has an obvious generalization to more than two random variables.

Two random variables X and Y are said to have the same distribution if for all functions g we have $E[g(X)] = E[g(Y)]$. Thus all probability predictions about the two random variables, taken individually, are the same.

From now on we deal with a standard situation. We consider a sequence of random variables $X_1, X_2, X_3, \ldots, X_n$. They are assumed to be produced by repeating an experiment n times. The number n is called the sample size. Typically we shall assume that these random variables are independent. Further, we shall assume that they all have the same distribution, that is, they are *identically distributed*.

We say that a random variable Z has a *standard normal* distribution if for all bounded functions f we have

$$E[g(Z)] = \int_{-\infty}^{\infty} g(z)\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})\,dz. \tag{1.30}$$

It is called standard because it has mean zero and variance 1.

Theorem 1.7 *Let $X_1, \ldots, X_n$ be independent random variables, all with the same mean μ and standard deviation σ. Assume that they are identically distributed. Let $\bar{X}_n$ be the sample mean. The standardized sample mean is*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}. \tag{1.31}$$

Let g be a bounded piecewise continuous function. Then

$$E[g(Z_n)] \to E[g(Z)] \tag{1.32}$$

as $n \to \infty$, where Z is standard normal.

1.5 Joint distributions of random variables

Let $X_1, \ldots, X_n$ be random variables. Then the joint distribution of these random variables is specified by giving either a continuous density or a discrete density.

If for all functions g for which the expectation exists we have

$$E[g(X_1, \ldots, X_n)] = \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n, \tag{1.33}$$

then the random variables have a continuous probability density $f(x_1, \ldots, x_n \mid \theta)$ with parameter θ.

If for all functions g for which the expectation exists we have

$$E[g(X_1, \ldots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n \mid \theta), \tag{1.34}$$

then the random variables have a discrete probability density $f(x_1, \ldots, x_n \mid \theta)$ with parameter θ.

In the case of independent identically distributed random variables the joint probability density factors:

$$f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdots f(x_n \mid \theta). \tag{1.35}$$

Example: The exponential distribution is defined by the continuous density

$$f(x \mid \theta) = \theta e^{-\theta x} \tag{1.36}$$

for $x \geq 0$ and $f(x \mid \theta) = 0$ for $x < 0$. Here $\theta > 0$ is related to the mean by $\mu = 1/\theta$. It is a typical distribution for a waiting time (in continuous time) for the next jump. The distribution of $X_1 + \cdots + X_n$ is then Gamma with parameters n, θ. It is the waiting time for the nth jump.

Example: The normal distribution is defined by the continuous density

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.37}$$

The distribution of $X_1 + \cdots + X_n$ is then normal with mean $n\mu$ and variance $n\sigma^2$.

Example: The Bernoulli distribution is defined by the discrete density

$$f(x \mid p) = p^x(1-p)^{1-x} \tag{1.38}$$

for $x = 0, 1$. Here $\mu = p$. This counts the occurrence of a single success or failure. The distribution of $X_1 + \cdots + X_n$ is then binomial. It counts the number of successes in n independent trials.

Example: The geometric distribution is defined by the discrete density

$$f(x \mid p) = p(1-p)^x \tag{1.39}$$

for $x = 0, 1, 2, 3, \ldots$. Here $\mu = \frac{1}{p} - 1$. This is the distribution of the number of failures before the first success. The distribution of $X_1 + \cdots + X_n$ is then negative binomial. It is the number of failures before the nth success.

Example: The Poisson distribution with mean $\mu > 0$ is defined by the discrete density

$$f(x \mid \mu) = \frac{\mu^x}{x!} e^{-\mu} \tag{1.40}$$

for $x = 0, 1, 2, 3, \ldots$. It is a typical distribution of the number of successes in a fixed interval of continuous time. The distribution of $X_1 + \cdots + X_n$ is Poisson with mean $n\mu$. It counts the number of successes in n disjoint intervals.

1.6 Problems

1. Consider the experiment of throwing a die n times. The results are $X_1, \ldots, X_n$. Then $E[f(X_i)] = \frac{1}{6}\sum_{k=1}^{6} f(k)$, and the X_i are independent. Find the mean μ and standard deviation σ of each X_i.

2. Consider the dice experiment. Take $n = 25$. Find the mean $\mu_{\bar{X}}$ of the sample mean $\bar{X}$. Find the standard deviation $\sigma_{\bar{X}}$ of the sample mean $\bar{X}$.

3. Perform the dice experiment with $n = 25$ and get an outcome ω. Record the 25 numbers. Report the sample mean $\bar{X}(\omega)$. Report the sample standard deviation $s(\omega)$.

4. Consider independent random variables $X_1, \ldots, X_n$. For notational convenience, consider the centered random variables $Y_i = X_i - \mu$, so that $E[Y_i] = 0$. Let $\sigma^2 = E[Y_i^2]$ and $q^4 = E[Y_i^4]$. Prove that

$$E[\bar{Y}_n^4] = \frac{1}{n^4}[nq^4 + 3n(n-1)\sigma^4]. \tag{1.41}$$

5. In the proceeding problem, show that

$$E[\sum_{n=k}^{\infty} \bar{Y}_n^4] \leq \frac{1}{2} q^4 \frac{1}{(k-1)^2} + 3\sigma^4 \frac{1}{(k-1)}. \tag{1.42}$$

In terms of the original X_i this says that there is a constant C such that

$$E[\sum_{n=k}^{\infty}(\bar{X}_n - \mu)^4] \leq C\frac{1}{k}. \tag{1.43}$$

Thus if k is large, then all the sample means $\bar{X}_n$ for $n \geq k$ are likely to be close to μ, in some average sense. This is a form of the strong law of large numbers. Compare with the weak law

$$E[(\bar{X}_k - \mu)^2] = \sigma^2\frac{1}{k}, \tag{1.44}$$

which only shows that, for each fixed k, the sample mean $\bar{X}_k$ is very likely to be close to μ.

Chapter 2

Probability

2.1 Events and probability

Probability is a special case of expectation!

We consider an experiment with a set Ω of *outcomes*. An *event* is a subset of Ω. For each event A, there is a random variable 1_A called the *indicator* of this event. The value $1_A(\omega) = 1$ if the outcome ω belongs to A, and the value $1_A(\omega) = 0$ if the outcome ω does not belong to A. Thus one scores 1 if the outcome belongs to A, and one scores zero if the outcome does not belong to A. The *probability* of the event is defined by

$$P[A] = E[1_A]. \tag{2.1}$$

In the following we shall sometimes call the probability of an event the *population proportion*. Probability satisfies the following properties:

1. $P[\Omega] = 1$.
2. $P[A \cup B] + P[A \cap B] = P[A] + P[B]]$
3. $A \subset B$ implies $P[A] \leq P[B]$.
4. $P[\emptyset] = 0$.

The second properties is *additivity*. The third property is the *order* property. The first and fourth properties are normalizations. They say that the probability of the *sure event* Ω is one, while the probability of the *impossible event* $\emptyset$ is zero.

The additivity property is often used in the following form. Say that A, B are *exclusive* events if $A \cap B = \emptyset$. If A, B are exclusive, then $P[A \cup B] = P[A] + P[B]$.

Another useful concept is that of complementary event. The event A^c is defined to consist of all of the outcomes that are not in A. Then by additivity $P[A] + P[A^c] = 1$.

Events A, B are said to be *independent* if $P[A \cap B] = P[A]P[B]$.

In realistic probability problems the set theory language is often replaced by an equivalent terminology. An event is thought of as a property of outcomes. The notions of union, intersection, complement are replaced by the logical notations or, and, not. Thus additivity says that for *exclusive* events, the probability that one *or* the other occurs is the *sum* of the probabilities. Also, the probability that an event occurs plus the probability that an event does not occurs is one. The definition of *independent* events is that the probability that one event *and* the other event occur is the *product* of the probabilities.

Sometimes the conjunction (the intersection) is represented by a comma. Thus, for example, the definition of independent events could be written $P[A, B] = P[A]P[B]$.

If X is a random variable, and S is a set of real numbers, then the event $X \in S$ that X is in S consists of all the outcomes ω such that $X(\omega)$ is in S. Let 1_S be the indicator function defined on the set of real numbers that is 1 for a number in S and 0 for a number not in S. Then the indicator function of the event that X is in S is the random variable $1_S(X)$.

If X and Y are random variables with the same distribution, then for each set S of real numbers we have $P[X \in S] = P[Y \in S]$.

If X and Y are independent random variables, then for each pair of sets S, T of real numbers we have $P[X \in S, Y \in T] = P[X \in S]P[Y \in T]$.

2.2 The sample proportion

In statistics the *sample proportion* f_n is used to estimate the population proportion p.

Theorem 2.1 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Let $N_n = 1_{A_1} + \cdots + 1_{A_n}$ be the number of events that occur. Let*

$$f_n = \frac{N_n}{n} \tag{2.2}$$

be the sample frequency. Then the expectation of f_n is

$$E[f_n] = p. \tag{2.3}$$

Theorem 2.2 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Assume that each pair A_i, A_j of events $i \neq j$ are independent. Let $N_n = 1_{A_1} + \cdots + 1_{A_n}$ be the number of events that occur. Let*

$$f_n = \frac{N_n}{n} \tag{2.4}$$

be the sample frequency. Then the standard deviation of f_n is

$$\sigma_{f_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}. \tag{2.5}$$

Proof: The variance of 1_{A_1} is $p - p^2 = p(1-p)$.

In order to use this theorem, the following remark upper bound on the standard deviation for one observation is fundamental.

$$\sqrt{p(1-p)} \leq \frac{1}{2}. \tag{2.6}$$

This means that one can figure out an upper bound on the standard deviation of the sample frequency without knowing the population probability. Often, because of the central limit theorem, one wants to think of a reasonable bound on the probable error to be 2 standard deviations. Thus the memorable form of this result is

$$2\sigma_{f_n} \leq \frac{1}{\sqrt{n}}. \tag{2.7}$$

This is not quite as important, but for many practical problems there is also a useful lower bound for the standard deviation of one observation. If $1/10 \leq p \leq 9/10$, then $3/10 \leq \sqrt{p(1-p)}$.

2.3 The central limit theorem

If a random variable Z has a standard normal distribution, the we know the probability that Z belongs to a given interval. The distribution is symmetric about zero, so we only need to know the probabilities for intervals of positive numbers. In particular, $P[0 < Z < 1] = .341$, $P[1 < Z < 2] = .136$, while $P[2 < Z] = .023$. These sum to $P[0 < Z] = .5$. If we want to memorize these numbers to a rough approximation, we can think of them as 34 percent, 13 and a half percent, and 2 and a half percent. Sometimes it is also useful to know that $[1.645 < Z] = .05$ and $P[1.96 < Z] = .025$.

Recall the central limit theorem. It may be stated in terms of probabilities as follows.

Theorem 2.3 *Let $X_1, \ldots, X_n$ be independent random variables, all with the same mean μ and standard deviation $\sigma < \infty$. Assume that they are identically distributed. Let $\bar{X}_n$ be the sample mean. The standardized sample mean is*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}. \tag{2.8}$$

Then

$$P[a < Z_n < b] \to P[a < Z < b] \tag{2.9}$$

as $n \to \infty$, where Z is standard normal.

It is important to remember in connection with this theorem that $P[|Z| > 2] = 0.046$ which is approximately 5 percent. So two standard deviations is a kind of cutoff after which probability gets reasonably small. About 95 percent of the probability will be between two standard deviations of the mean.

Events A, B are independent if $P[A \cap B] = P[A]P[B]$. The notion of independence has a generalization to a set of more than two events: Every subset of the set must satisfy the product property.

If events A, B are independent, then so are the events A, B^c, and the events A^c, B, and the events A^c, B^c. So it does not matter whether one works with the original events or with their complements.

We consider a sequence of random variables $A_1, A_2, A_3, \ldots, A_n$. The are assumed to be produced by repeating an experiment n times. The number n is called the sample size. Typically we shall assume that these events are independent. Further, we shall assume that they all have the same probability.

Theorem 2.4 *Let $A_1, \ldots, A_n$ be independent events, all with the same probability p. Assume that they are identically distributed. Let f_n be the sample proportion. Let*

$$Z_n = \frac{f_n - p}{\sqrt{p(1-p)}/\sqrt{n}}. \tag{2.10}$$

Then

$$P[a < Z_n < b] \to P[a < Z < b] \tag{2.11}$$

as $n \to \infty$.

This theorem is proved by letting $X_i = 1_{A_i}$ and applying the general version of the central limit theorem. It was historically the first version of the central limit theorem.

2.4 Problems

1. Consider the experiment of throwing a die n times. Let A_i be the event that the ith throw gives a number in the range from 1 to 4. Find the mean μ and standard deviation σ of the indicator function of each A_i.

2. Consider the same dice experiment. Take $n = 50$. Find the mean μ_f of the sample proportion f. Find the standard deviation σ_f of the sample proportion f.

3. Perform the dice experiment with $n = 50$ and get an outcome ω. Record the 50 events A_i. Report the sample proportion $f(\omega)$.

4. Consider a random sample of size n from a very large population. The question is to find what proportion p of people in the population have a certain opinion. The proportion in the sample who have the opinion is f. How large must n be so that the standard deviation of f is guaranteed to be no larger than one percent?

5. Consider independent identically distributed random variables $X_1, \ldots, X_n$ with finite variance. We know from the weak law of large numbers that

$$E[(\bar{X}_n - \mu)^2] = \sigma^2 \frac{1}{n}, \tag{2.12}$$

   Use this to prove that

$$P[|\bar{X}_n - \mu| \geq t] \leq \frac{\sigma^2}{nt^2}. \tag{2.13}$$

   Thus for large n the probability that the sample mean $\bar{X}_n$ deviates from the population mean μ by t or more is small. This is the weak law of large numbers.

6. The result of the last problem can also be written

$$P[|\bar{X}_n - \mu| \geq \epsilon \frac{\sigma}{\sqrt{n}}] \leq \frac{1}{\epsilon^2}. \tag{2.14}$$

   Compare the result obtained from this for $\epsilon = 2$ with the result obtained from the central limit theorem. Which gives more useful information for large n?

7. Consider independent random variables $X_1, \ldots, X_n$ with finite fourth moments. We have seen that

$$E[\sum_{j=n}^{\infty} (\bar{X}_j - \mu)^4] \leq \frac{C}{n}. \tag{2.15}$$

   Show that

$$P[\max_{j \geq n} |\bar{X}_j - \mu| \geq t] \leq \frac{C}{nt^4}. \tag{2.16}$$

Thus for large n the probability that there exists some $j \geq n$ such that the sample mean $\bar{X}_j$ deviates from the population mean μ by t or more is small. This is the strong law of large numbers.

8. Let X_1 and X_2 be the numbers resulting from throwing two dice. Let A be the event that X_1 is odd, let B be the event that X_2 is odd, and let C be the event that $X_1 + X_2$ is odd. Show that A, B are independent, A, C are independent, and B, C are independent. Show that A, B, C are not independent.

Chapter 3

Estimation

3.1 Estimating means

Each member of a population has some characteristic quantity X. The mean of this quantity for the whole population is μ. The standard deviation of this quantity over the whole population is σ. One can think of taking a single random member of the population and measuring this quantity X_1. Then the expectation of the random variable X_1 is μ, and the standard deviation of X_1 is σ. Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$. These are to be independent random variables. (If the population is small, the sampling is taken to be with replacement. If the population is very large compared to the sample size, then it does not matter whether the sampling is with replacement or without replacement.)

Theorem 3.1 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{3.1}$$

be their sample mean. Then the expectation of $\bar{X}_n$ is

$$E[\bar{X}_n] = \mu. \tag{3.2}$$

Theorem 3.2 *(Weak Law of Large Numbers) Let $X_1, X_2, X_3, \ldots, X_n$ be independent random variables, each with mean μ and standard deviation σ. Let $\bar{X}_n$ be their sample mean. Then the standard deviation of $\bar{X}_n$ is*

$$\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{3.3}$$

For the statistician the μ and σ are unknown. The significance of the above theorems is that it is reasonable to use the experimental sample mean $\bar{X}_n$ to estimate the unknown population parameter μ, provided that n is large enough.

If n is large enough, then $\bar{X}_n$ is quite likely to be near μ. The central limit theorem gives a rather precise idea of how variable these sample means are.

Theorem 3.3 *(Central Limit Theorem) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and standard deviation $\sigma < \infty$. Let $\bar{X}_n$ be their sample mean. Then the standardized variable*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{3.4}$$

has the property that

$$\lim_{n\to\infty} P[a < Z_n < b] = P[a < Z < b], \tag{3.5}$$

where Z is standard normal.

To estimate the unknown population parameter σ^2, define the sample variance as

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1} \tag{3.6}$$

This is an experimental quantity used to estimate the unknown population variance σ^2.

Theorem 3.4 *Let $X_1, X_2, X_3, \ldots X_n$ be independent random variables, each with mean μ and standard deviation σ. Let s^2 be the sample variance. Then the expectation of s^2 is*

$$E[s^2] = \sigma^2. \tag{3.7}$$

If the sample size n is large, then s^2 will have small variance and be quite close to σ^2. However if the sample size is small, the random variable s^2 might be considerably larger or smaller than σ^2. If happens to be considerably smaller than σ^2, that is, unusually close to zero, then it is giving a misleadingly optimistic impression of how well sample means estimate the population mean.

Note: The central limit theorem gives normality of the sample means, but only for moderately large sample size n. Sometimes the statistician knows (or optimistically assumes) that the underlying population is normal. This means that the $(X_i - \mu)/\sigma$ are already standard normal random variables. In this case the random variable $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ is automatically normal for every n. In this case the distribution of s^2/σ^2 does not depend on μ or on σ. Its distribution is what is known as $\chi^2_{n-1}/(n-1)$. This is known to have mean 1 and variance $2/(n-1)$. If n is larger than about 20 the corresponding standard deviation is quite small, but if n is smaller than something like 10, the chance of an abnormally small value of s^2 becomes quite significant.

3.2 Two population means

A very common situation is when one is interested in the difference between two population means. Say that one population has mean μ_1 and the other population has mean μ_2. Then one is interested in $\mu_1 - \mu_2$. It seems reasonable to use the difference of sample means $\bar{X}_{1n_1} - \bar{X}_{2n_2}$ to estimate $\mu_1 - \mu_2$. In order to see how good a job one is doing with this method, it is useful to know the standard deviation of the difference of sample means. The variance of the difference is

$$\mathrm{Var}(\bar{X}_{1n_1} - \bar{X}_{2n_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{3.8}$$

The standard deviation, the quantity of real interest, is the square root of this variance. In order to use this result, one needs to know the two population variances σ_1^2 and σ_2^2. If the sample sizes n_1 and n_2 are large enough, these may be estimated by the corresponding sample variances for the two samples. Sometimes there is reason to believe (or hope) that the two population variances $\sigma_1^2 = \sigma_2^2$ are the same. In that case one can pool the sample variances. The usual formula for this is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - 1 + n_2 - 1}. \tag{3.9}$$

This should estimate the common variance σ^2.

3.3 Estimating population proportions

In statistics the *sample proportion* f_n is used to estimate the population proportion p.

Theorem 3.5 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Let $N_n = 1_{A_1} + \cdots + 1_{A_n}$ be the number of events that occur. Let*

$$f_n = \frac{N_n}{n} \tag{3.10}$$

be the sample frequency. Then the expectation of f_n is

$$E[f_n] = p. \tag{3.11}$$

Theorem 3.6 *Let $A_1, A_2, A_3, \ldots, A_n$ be events, each with probability p. Assume that each pair A_i, A_j of events $i \neq j$ are independent. Let f_n be the sample frequency. Then the standard deviation of f_n is*

$$\sigma_{f_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}. \tag{3.12}$$

Typically the p that occurs in the square root is unknown. So one estimates it with f_n.

Theorem 3.7 *Let $A_1, \ldots, A_n$ be independent events, all with the same probability p. Assume that they are identically distributed. Let f_n be the sample proportion. Let*

$$Z_n = \frac{f_n - p}{\sqrt{p(1-p)}/\sqrt{n}}. \tag{3.13}$$

Then

$$\lim_{n\to\infty} P[a < Z_n < b] = P[a < Z < b], \tag{3.14}$$

where Z is standard normal.

In some problems one wants to use this result in a situation where the p is unknown. There is nothing to be done about the p in the numerator, but the p in the denominator can be estimated by f_n. Thus for large n the statistic

$$Z_n = \frac{f_n - p}{\sqrt{f_n(1-f_n)}/\sqrt{n}} \tag{3.15}$$

is approximately standard normal.

3.4 Two population proportions

Say that one population has proportion p_1 and the other population has proportion p_2, and one is interested in $p_1 - p_2$. It seems reasonable to use the difference of sample proportions $f_{1n_1} - f_{2n_2}$ to estimate $p_1 - p_2$. In order to see how good a job one is doing with this method, it is useful to know the standard deviation of the difference of sample proportions. The variance of the difference is

$$\mathrm{Var}(f_{1n_1} - f_{2n_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \tag{3.16}$$

The standard deviation, the quantity of real interest, is the square root of this variance. In order to use this result, one needs to know the two population variances $p_1(1-p_1)$ and $p_2(1-p_2)$. If the sample sizes n_1 and n_2 are large enough, these may be estimated by using f_{1n_1} to estimate p_1 and f_{2n_2} to estimate p_2.

3.5 Supplement: Confidence intervals

Confidence intervals usually occur in the context of estimation. Suppose we have a population of numbers with unknown mean μ and unknown standard deviation σ. Consider a sample $X_1, \ldots, X_n$ of size n. Then the sample mean $\bar{X}$ estimates μ. One often wants to convey an idea of how good this estimate is. This information is conveyed by $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. However the σ that occurs in this formula is not known. One method is to use the sample standard deviation s to estimate σ. Then $s/\sqrt{n}$ estimates $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

Let us assume that the sample size n is reasonably large. Then by the central limit theorem,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{3.17}$$

is approximately standard normal. Since for n large s is probably quite close to σ, it is also true that

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{3.18}$$

has a distribution that is approximately standard normal. In that case

$$P[-2 < t < 2] = 0.95, \tag{3.19}$$

since this is correct for the standard normal distribution.

Let $L = \bar{X} - 2s/\sqrt{n}$ and $R = \bar{X} + 2s/\sqrt{n}$. Then this is equivalent to saying

$$P[L < \mu < R] = 0.95, \tag{3.20}$$

since $L < \mu < R$ is algebraically equivalent to $-2 < t < 2$. Thus these two conditions define the same event. The interval from L to R is called a 95% confidence interval for the unknown population mean. This is a random interval. This means that most of the time that statisticians use this procedure they get an interval that contains the unknown population mean. However about one time in twenty it does not contain the unknown population mean. The statistician is of course not aware that this has happened.

It is important to realize that the two numbers L and R contain exactly the same information as the two numbers $\bar{X}$ and $s/\sqrt{n}$. This is because we can go from the estimators to the confidence interval by $L = \bar{X} - 2s/\sqrt{n}$ and $R = \bar{X} + 2s/\sqrt{n}$, and we can go back from the confidence interval to the estimators by $\bar{X} = (R+L)/2$ and $s/\sqrt{n} = (R-L)/4$. The confidence interval is a packaging of the information that is more attractive to some people.

If the sample size n is small and the population is normal, then the statistic t has the t distribution with $n-1$ degrees of freedom. In this case to get a 95% confidence interval one has to replace the 2 with a slightly larger number a, depending on n. This is to compensate for the fact that there is an appreciable chance that s/σ may be considerably smaller than one. Since the s occurs in the denominator, this gives a chance for somewhat large t values. However this does not make a huge change. Even though $n = 8$ is a quite small sample size, the corresponding value of a is about 2.3, which is not all that different. In any case, with the proper determination of a the probability that $-a < t < a$ is 0.95. The corresponding 95% confidence interval goes from $L = \bar{X} - as/\sqrt{n}$ to $R = \bar{X} + as/\sqrt{n}$.

3.6 Problems

1. You are given Data Set 3.1 with sample size n drawn from a population with unknown mean μ and unknown standard deviation σ. The population

consists of stars, and the measurements are indices of brightness in a certain frequency range. Estimate μ. Estimate σ. Estimate the standard deviation of the sample means $\bar{X}_n$ in this kind of experiment.

2. You are given Data Set 3.2 with sample size n drawn from a population divided into successes and failures. The proportion p of successes is unknown. The population consists of people, and success is having knowledge of a certain public health measure. Estimate p. Estimate the standard deviation of the sample proportions f_n in this kind of experiment.

3. The chi square distribution is defined to be the distribution of the random variable $\chi^2_n = Z_1^2 + \cdots + Z_n^2$, where each Z_i is standard normal, and where the Z_i are independent. Show that the mean of χ^2_n is n and the variance of χ^2_n is $2n$.

4. For a normal population $(n-1)s^2/\sigma^2$ has the distribution χ^2_{n-1}. Find the mean and variance of s^2. Show that the variance is small for large n.

Chapter 4

Hypothesis testing

4.1 Null and alternative hypothesis

The philosophy of hypothesis testing is that there is some effect that may or may not be present. If the effect is not present, then the population parameters satisfy some special equations. If the effect is present, then these equations no longer hold. The first case is called the *null hypothesis* and the second case is called the *alternative hypothesis*.

Unfortunately, the experimenter does not know which hypothesis is true. So one tries to construct a procedure that will make an intelligent guess. The procedure should be constructed so that if the null hypothesis is true, then with high probability the experimenter guesses the null hypothesis. On the other hand, the procedure should also ensure that if the alternative hypothesis is true (and the equation defining the null hypothesis is rather far from being satisfied), then with high probability the experimenter guesses the alternative hypothesis.

In general, such a procedure can be found only if the sample size is fairly large. Otherwise there is an unpleasant tradeoff, so that a procedure that works well in one situation will work badly in the other situation.

The most common method of approaching this problem is to concentrate on the null hypothesis. This is largely a matter of convenience. One wants to pick a procedure that has the property that if the null hypothesis is true, then the experiment guesses the null hypothesis with at least reasonably high probability. Then one wants to try to finance a sample large enough so that there is also a good chance of making a correct decision if the alternative hypothesis is true.

4.2 Hypothesis on a mean

The picture is that there is a very large (theoretically infinite) population. Each member of the population has some characteristic quantity X. The mean of this quantity for the whole population is μ. The standard deviation of this quantity over the whole population is σ.

The null hypothesis is that $\mu = \mu_0$, where μ_0 is some standard value. In many situations $\mu_0 = 0$.

There are several common alternative hypotheses. The two sided alternative is $\mu \neq \mu_0$. The one-sided alternatives are $\mu > \mu_0$ and $\mu < \mu_0$.

Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$.

Theorem 4.1 *(Central Limit Theorem) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and standard deviation σ. Let $\bar{X}_n$ be their sample mean. Suppose the null-hypothesis $\mu = \mu_0$ is true. Then the standardized variable*

$$Z_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \tag{4.1}$$

has the property that for large n

$$P[s < Z_n < t] \approx P[s < Z < t], \tag{4.2}$$

where Z is standard normal.

Now the only trouble with this is that the σ is also an unknown population parameter. So to understand what one is doing, one has to estimate σ. The following result tells how to do this. This is done with the sample standard deviation s. When this is done, it is conventional to call the ratio t instead of Z. However, under the null hypothesis, for large n the random variable s will be nearly constant, and the ratio will continue to have the standard normal distribution. In any case, the test that results is to examine

$$t = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}} \tag{4.3}$$

For a two-sided test of $\mu = \mu_0$ against $\mu \neq \mu_0$, one guesses the null hypothesis if $|t| \leq a$. Here a is a cutoff, often taken to be somewhere near 2. For a one-sided test, for instance $\mu = \mu_0$ against $\mu > \mu_0$, one guesses the null hypothesis if $t \leq b$. Here again b is a chosen cutoff value. A common value is 1.65.

If the sample size is small, then this test has problems. The sample standard deviation s can then have a significant probability of being rather far from the population standard deviation. In particular, it can be unusually small. This makes the assumption that t is normal more dubious.

In this situation it is customary to assume that the individual random variables X_i are normal. That is, the population is normal. In this case one can do more explicit computations. Write the test statistic in the form

$$t = \frac{(\bar{X} - \mu_0)/(\sigma/\sqrt{n})}{s/\sigma}. \tag{4.4}$$

Under the null hypothesis $\mu = \mu_0$ the distribution of t is the distribution of $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ divided by s/σ. The numerator Z has the standard normal

distribution. The random variable $(n-1)s^2/\sigma^2$ has the χ^2_{n-1} distribution, the chi-squared distribution with $n-1$ degrees of freedom. These distributions only depend on the sample size n. So the distribution of t only depends on the sample size n. It is called the t distribution with $n-1$ degrees of freedom. The final result is that for small n this distribution is somewhat more spread out than the standard normal distribution. This is because the χ^2_{n-1} distribution can be considerably smaller than its expected value. This extra spread of the t distribution for small samples must be taken into account when one computes the cutoff values. Statistics books contain tables of these values.

But is the population normal? It is nice that even if the population is not normal, then for a large sample one can use the law of large numbers and the central limit theorem to compute everything.

4.3 Two means

A very common situation is when the null hypothesis involves two population means. Often the null hypothesis is $\mu_1 = \mu_2$. Then clearly $\mu_1 - \mu_2 = 0$. It seems reasonable to use the difference of sample means $\bar{X}_{1n_1} - \bar{X}_{2n_2}$ for the test.

In order to see how good a job one is doing with this method, it is useful to know the standard deviation of the difference of sample means. The variance of the difference is

$$\mathrm{Var}(\bar{X}_{1n_1} - \bar{X}_{2n_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{4.5}$$

The standard deviation, the quantity of real interest, is the square root of this variance.

In order to use this result, one needs to know the two population variances σ_1^2 and σ_2^2. If the sample sizes are reasonably large, then these can be estimated by the two sample variances. However in some circumstances one knows that these are the same number σ^2. Then there is only one number to estimate, and one uses the pooled sample variance s^2. In this latter case the test statistic is

$$t = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2}}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}. \tag{4.6}$$

Again there are one-sided and two-sided tests.

4.4 Hypothesis on a proportion

In statistics one possible null hypothesis is that the sample proportion $p = p_0$, some standard value. In many experimental situations $p_0 = 1/2$.

Theorem 4.2 *Let $A_1, \ldots, A_n$ be independent events, all with the same probability p. Let f_n be the sample proportion. Assume the null hypothesis $p = p_0$ is true. Let*

$$Z_n = \frac{f_n - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}. \tag{4.7}$$

Then for large n

$$P[s < Z_n < t] \approx P[s < Z < t] \tag{4.8}$$

where Z is standard normal.

4.5 Two proportions

A very common situation is when the null hypothesis is that two population proportions are equal. Say that one population has proportion p_1 and the other population has proportion p_2. Then one the null hypothesis is $p_1 = p_2$. It seems reasonable to use the difference of sample proportions $f_{1n_1} - f_{2n_2}$ to estimate $p_1 - p_2$, which is zero if the null hypothesis is true.

In order to construct the test, it is useful to know the standard deviation of the difference of sample proportions. The variance of the difference is

$$\mathrm{Var}(f_{1n_1} - f_{2n_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \tag{4.9}$$

The standard deviation, the quantity of real interest, is the square root of this variance.

In order to use this result, one needs to know the two population variances $p_1(1-p_1)$ and $p_2(1-p_2)$. Under the null hypothesis, these are the same. There is only one p. This may be estimated by the pooled sample proportion.

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}. \tag{4.10}$$

It follows that the appropriate test statistics is

$$Z = \frac{f_1 - f_2}{\sqrt{\frac{f(1-f)}{n_1} + \frac{f(1-f)}{n_2}}}. \tag{4.11}$$

4.6 Independence

The same formulas may be used in a somewhat different situation. This is when one has events A, B and one wants to test whether they are independent. The null hypothesis is

$$P[A, B] = P[A]P[B]. \tag{4.12}$$

The alternative hypothesis is a violation of this equation.

We take a sample of size n. Then the results are in the form of four numbers $N_{A,B}$, N_{A,B^c}, $N_{A^c,B}$, N_{A^c,B^c}. Their sum is n, so

$$N_{A,B} + N_{A,B^c} + N_{A^c,B} + N_{A^c,B^c} = n \tag{4.13}$$

and there are only three independent numbers.

Define the conditional probabilities $P[A \mid B] = P[A, B]/P[B]$ and $P[A \mid B^c] = P[A, B^c]/P[B^c]$. We have the following proposition:

Theorem 4.3 *The correlation between events is related to the difference in conditional probablities by*

$$P[A, B] - P[A]P[B] = (P[A \mid B] - P[A \mid B^c])P[B]P[B^c]. \tag{4.14}$$

This shows in particular that the null hypothesis can also be stated in the form

$$P[A \mid B] = P[A \mid B^c]. \tag{4.15}$$

In fact, under the original null hypothesis both sides are equal to the same number $P[A]$.

Thus we think instead of taking two samples of random size $N_B = N_{A,B} + N_{A^c,B}$ and $N_{B^c} = N_{A,B^c} + N_{A^c,B^c}$. The sample frequencies are for A, B and A, B^c are

$$f_{A|B} = \frac{N_{A,B}}{N_B} \tag{4.16}$$

and

$$f_{A|B^c} = \frac{N_{A,B^c}}{N_{B^c}} \tag{4.17}$$

Under the null hypothesis the estimator of the common conditional probability is the pooled frequency

$$f_A = \frac{N_A}{n}, \tag{4.18}$$

where $N_A = N_{A,B} + N_{A,B^c}$. The test statistic is

$$Z = \frac{f_{A|B} - f_{A|B^c}}{\sqrt{f_A(1 - f_A)}\sqrt{\frac{1}{N_B} + \frac{1}{N_{B^c}}}}. \tag{4.19}$$

Under the null hypothesis this should be approximately standard normal. Under the alternative hypothesis values far away from zero are more probable.

It might seem irritating that the test statistic appears not to be symmetric between A and B. After all the original null hypothesis treated them both on the same basis.

However the test statistic is symmetric! To see this, merely compute that it is equal to

$$Z = \frac{N_{A,B}N_{A^c,B^c} - N_{A,B^c}N_{A^c,B}}{\sqrt{\frac{N_A N_{A^c} N_B N_{B^c}}{n}}}. \tag{4.20}$$

The numerator is a determinant that measures the dependence between the columns, or, equally well, between the rows.

4.7 Power

In a test the statistician uses the data to decide for the null hypothesis or to decide for the alternative hypothesis. The *power* of the test is the probability of deciding for the alternative hypothesis. It is a function Power(θ) of the

unknown population parameter θ. The power is a function that can and should be computed before the experiment is conducted. It gives an almost complete story about how the test will generally perform.

In particular, the *level* of the test is the power when the null hypothesis $\theta = \theta_0$ is true. Thus the level is

$$\alpha = \text{Power}(\theta_0). \tag{4.21}$$

It is desirable to have the level be a fairly small number. On the other hand, when the alternative hypothesis is true, it is desirable to have a power near one. So the goal is to have the power small when $\theta = \theta_0$, but Power(θ) should rise rapidly as the parameter deviates from this value.

In general, calculations of power can be complicated, especially when there are several parameters. However computers can make such calculations feasible. A prudent statistician will think carefully about power when designing a test.

Example: For simplicity, take the situation when the population mean μ is unknown, but the population standard deviation σ is known. The null hypothesis is $\mu = \mu_0$, while the alternative hypothesis is $\mu > \mu_0$.

Thus there might be an established treatment where the population mean of the response is μ_0 and the population standard deviation is σ. Both these numbers are known from long experience. A new treatment is attempted, and this is supposed to raise the mean response to μ. It is assumed that the standard deviation stays the same. The null hypothesis is that the treatment is ineffective. The alternative hypothesis is that the treatment actually increases response. If the alternative hypothesis is true, then one should use the new treatment. If the null hypothesis is true, then it is not worth the extra expense.

An experiment is done with n individuals, and the sample mean $\bar{X}$ is measured. The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}. \tag{4.22}$$

This has a standard normal distribution when the null hypothesis is true. The test is to decide for the alternative hypothesis when the test statistic Z is in the critical region $Z > c$. This is equivalent to saying that the sample mean satisfies $\bar{X} > \mu_0 + c\sigma/\sqrt{n}$. The number c is chosen so that

$$P_{\mu_0}[Z > c] = P_{\mu_0}[\bar{X} > \mu_0 + c\sigma/\sqrt{n}] = \alpha. \tag{4.23}$$

For example, if $\alpha = 0.05$, then $c = 1.645$. The power of the test is

$$\text{Power}(\mu) = P_\mu[Z > c] = P_\mu[\bar{X} > \mu_0 + c\sigma/\sqrt{n}]. \tag{4.24}$$

This may be computed by considering

$$Z_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{4.25}$$

which has a standard normal distribution. The power is

$$\text{Power}(\mu) = P_\mu[\bar{X} > \mu + c\sigma/\sqrt{n} + (\mu_0 - \mu)]. \tag{4.26}$$

This can also be written

$$\text{Power}(\mu) = P_\mu[Z_1 > c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}]. \tag{4.27}$$

Yet another equivalent form is

$$\text{Power}(\mu) = P_\mu[-Z_1 < -c + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}] = F(-c + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}). \tag{4.28}$$

Here F is the cumulative distribution function of the standard normal distribution.

From this we see that the power starts with the value α when $\mu = \mu_0$. Thus, for instance, if $c = 1.645$, then $F(-c) = 0.05$. The power rises as rises as μ increases. Say, for instance, that we want a power of at least 0.8. Since $F(0.84) = 0.8$, this says that

$$-1.645 + \frac{\mu - \mu_0}{(\sigma/\sqrt{n})} = 0.84 \tag{4.29}$$

or

$$\mu - \mu_0 > 2.49\frac{\sigma}{\sqrt{n}}. \tag{4.30}$$

In other words, to get this level of power, the alternative μ must be about 2.5 times $\sigma/\sqrt{n}$ above the null μ_0.

One can think of this as a requirement of a large enough sample. If it is important to detect a certain level of effectiveness $\mu > \mu_0$, then the sample size must be large enough so that 2.5 times $\sigma/\sqrt{n}$ is less than $\mu - \mu_0$. If this size sample is not available, then the test is not adequate to its purpose.

Example: Again take the case when the population mean μ is unknown, but the population standard deviation σ is known. The null hypothesis is $\mu = \mu_0$, while the alternative hypothesis is $\mu \neq \mu_0$.

The test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}. \tag{4.31}$$

This has a standard normal distribution when the null hypothesis is true. The test is to decide for the alternative hypothesis when the test statistic Z is in the critical region $|Z| > c$. This is the same as $Z > c$ or $Z < -c$. In terms of the sample mean, it says that $\bar{X} > \mu_0 + c\sigma/\sqrt{n}$ or $\bar{X} < \mu_0 - c\sigma/\sqrt{n}$. The number c is chosen so that

$$P_{\mu_0}[|Z| > c] = \alpha. \tag{4.32}$$

For example, if $\alpha = 0.05$, then $c = 1.96$. The power of the test is

$$\text{Power}(\mu) = P_\mu[|Z| > c]]. \tag{4.33}$$

This may be computed by considering

$$Z_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{4.34}$$

which has a standard normal distribution. The power can be written

$$\text{Power}(\mu) = P_\mu[Z_1 > c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}] + P_\mu[Z_1 < -c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}]. \tag{4.35}$$

Yet another equivalent form is

$$\text{Power}(\mu) = F(-c + \frac{\mu - \mu_0}{\sigma/\sqrt{n}}) + F(-c + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}). \tag{4.36}$$

Here F is the cumulative distribution function of the standard normal distribution.

Say that we want a power of at least 0.8. Since $F(-1.96 + 2.8) + F(-1.96 - 2.8) = 0.8$, this says that $|\mu - \mu_0|$ must be at least 2.8 times $\sigma/\sqrt{n}$.

Example: Here is an example in the case of a population of successes and failures. The population proportion p is unknown. The null hypothesis is $p = 1/2$, while the alternative hypothesis is $p > 1/2$.

This could arise in a matched pair experiment where one subject in each pair gets the treatment, and the other is the control. A success is when the treated subject appears to respond better than the untreated subject. The null hypothesis is that the treatment is ineffective. The alternative hypothesis is that the treatment succeeds. An experiment is done with n individuals, and the sample proportion f is measured. The test statistic is

$$Z = \frac{f - \frac{1}{2}}{1/(2\sqrt{n})}. \tag{4.37}$$

If both np and $n(1-p)$ are considerably larger than one, then this has an approximate standard normal distribution when the null hypothesis is true. The test is to decide for the alternative hypothesis when the test statistic Z is in the critical region $Z > c$. This is equivalent to saying that the sample proportion satisfies $f > 1/2 + c/(2\sqrt{n})$. The number c is chosen so that

$$P_{\frac{1}{2}}[Z > c] = P_{\frac{1}{2}}[f > \frac{1}{2} + c\frac{1}{2\sqrt{n}}] = \alpha. \tag{4.38}$$

The power of the test is

$$\text{Power}(p) = P_p[Z > c] = P_p[f > \frac{1}{2} + c\frac{1}{2\sqrt{n}}]. \tag{4.39}$$

This may be computed by considering

$$Z_1 = \frac{f - p}{\sqrt{p(1-p)}/\sqrt{n}} \tag{4.40}$$

which has a standard normal distribution. The power is

$$\text{Power}(p) = P_p[f > p + c\frac{1}{2\sqrt{n}} + (\frac{1}{2} - p)]. \tag{4.41}$$

This can also be written

$$\text{Power}(p) = P_p[Z_1 > c\frac{1}{2\sqrt{p(1-p)}} + \frac{\frac{1}{2} - p}{\sqrt{p(1-p)}/\sqrt{n}}]. \tag{4.42}$$

Yet another equivalent form is

$$\text{Power}(p) = P_p[-Z_1 < (-c+\frac{p-\frac{1}{2}}{1/(2\sqrt{n})})\frac{1}{2\sqrt{p(1-p)}}] = F((-c+\frac{p-\frac{1}{2}}{1/(2\sqrt{n})})\frac{1}{2\sqrt{p(1-p)}}). \tag{4.43}$$

Here F is the cumulative distribution function of the standard normal distribution.

It turns out that the last factor involving the reciprocal of $2\sqrt{p(1-p)}$ ordinarily does not matter too much. This is because for the range of p of main interest this factor is close to one. Obviously one wants to stay away from $p = 0$ or $p = 1$.

To see this, take the example when $c = 1.645$. Then the power at $p = 1/2$ is 0.05. Say that we want a $p > 1/2$ that makes the power 0.8. Then $p - 1/2$ divided by $1/(2\sqrt{n})$ times $2\sqrt{p(1-p)}$ is about 2.5. Consider a sample of even moderate size, such as $n = 25$. Then $1/(2\sqrt{n})$ is 0.10, and so if we neglect the third factor the required p is $0.75 = 3/4$. Then $\sqrt{(p(1-p)} = \sqrt{3}/4 = 0.433$, which is not all that different from $1/2$. If the sample is larger, say $n = 50$, then $p = 0.67 = 2/3$, and $\sqrt{p(1-p)} = \sqrt{2}/3 = 0.47$, which is quite close to $1/2$. The conclusion is that for reasonable sample sizes the power may be approximated by

$$\text{Power}(p) = F(-c + \frac{p-\frac{1}{2}}{1/(2\sqrt{n})}). \tag{4.44}$$

In other words, to get a power of 0.8, the alternative p must be about 2.5 times $1/(2\sqrt{n})$ above the null $1/2$. A sample of 100 will have a good chance to detect a p of 0.625. If it is important to detect a certain proportion $p > 1/2$, then the sample size must be large enough so that 2.5 times $1/(2\sqrt{n})$ is less than $p - 1/2$. To detect a p of 0.55 requires a sample size at least 625.

4.8 Loss

Often hypothesis testing experiments are considered using the concept of power, without any explicit concept of loss. However if one is to act on the basis of a decision for the null or alternative hypothesis, then it may be worth considering the loss from making an inappropriate decision.

In a hypothesis testing experiment there are two *loss functions* to consider. The loss function $L(\theta, 1)$ is the loss due to deciding for the alternative hypothesis when the true parameter value is θ. When $\theta = \theta_0$ this is the loss from a type I error.

The loss function $L(\theta, 0)$ is the loss due to deciding for the null hypothesis when the true parameter value is θ. When $\theta \neq \theta_0$ this is the loss from a type II error.

The *risk function* $R(\theta)$ for a particular test is the expected loss when using that test, as a function of the unknown parameter θ. That is

$$R(\theta) = L(\theta, 1)\text{Power}(\theta) + L(\theta, 0)(1 - \text{Power}(\theta)). \tag{4.45}$$

Risk is to be avoided. That is, one would like the values of this function to be reasonably small. Sometimes the criterion is to choose the test to make the maximum risk (as θ varies) as small as possible, but this is not the only possible strategy.

What is the loss function? Statistics does not give the answer. One must think about the real world consequences of one actions. In writing a loss function there must be a balance between realism and simplicity. Here are some possibilities.

0. The simplest choice is a piecewise constant function. However this choice is not even continuous.

1. The next simplest choice is a piecewise linear function. This way one can at least get continuous functions.

2. Perhaps quadratic functions could be convenient. One could take $L(\theta, 0) = a(\theta - \theta_0)^2$ and $L(\theta, 1) = c - b(\theta - \theta_0)^2$.

Example: This example is too artificial, but it is instructive. Consider a test of the null hypothesis $\mu = \mu_0$ versus the alternative hypothesis $\mu > \mu_0$. Assume for convenience of analysis that the value of σ is known. Take the simplest kind of loss function, piecewise constant.

Thus $L(\mu, 0) = 0$ for $\mu_0 \leq \mu < \mu^*$ and $L(\mu, 0) = L_0$ for $\mu^* < \mu$. This says that a parameter value between μ_0 and μ^* is so lacking in practical importance that one does not even mind missing it. But an alternative larger than μ^* is vitally important, and missing it by incorrectly guessing the null hypothesis (type II error) is catastrophic.

Similarly, $L(\mu, 1) = L_1$ for $\mu_0 \leq \mu < \mu^*$ and $L(\mu, 0) = 0$ for $\mu^* < \mu$. This says that an alternative between μ_0 and μ^* is so lacking in practical importance that in acting on the alternative hypothesis one is making a costly mistake. However there is no loss to correctly deciding for the alternative when the parameter is larger than μ^*.

With this model the risk $R(\mu) = L_1\text{Power}(\mu)$ for $0 \leq \mu < \mu*$ and is $R(\mu) = L_0(1 - \text{Power}(\mu))$ when $\mu^* < \mu$.

Say that the test is to decide for the alternative if

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > c. \tag{4.46}$$

Equivalently, one decides for the alternative if

$$\bar{X} > \mu_0 + c\frac{\sigma}{\sqrt{n}} \tag{4.47}$$

The choice of test is determined by the choice of the critical value c. Then the power is given by

$$\text{Power}(\mu) = F(-c + \frac{\sqrt{n}}{\sigma}(\mu - \mu_0)). \tag{4.48}$$

The maximum risk occurs at the point μ^*. This maximum risk may be minimized by making the risk function continuous at this point. This says that

$$L_1 \text{Power}(\mu^*) = L_0(1 - \text{Power}(\mu^*)), \tag{4.49}$$

which may be solved to give

$$\text{Power}(\mu^*) = \frac{L_0}{L_0 + L_1}. \tag{4.50}$$

Say that $L_0 = 4$ and $L_1 = 1$. This says that one is particularly afraid of making a type II error, missing out on an important alternative. Then the right hand side of the last equation is $4/5 = 0.8$. Since $F(0.84) = 0.8$, the equation is equivalent to

$$-c + \frac{\sqrt{n}}{\sigma}(\mu^* - \mu_0) = 0.84 \tag{4.51}$$

which says that the cutoff for $\bar{X}$ is

$$\mu_0 + c\frac{\sigma}{\sqrt{n}} = \mu^* - 0.84\frac{\sigma}{\sqrt{n}}. \tag{4.52}$$

Thus when n is large, the cutoff for $\bar{X}$ is just a small amount under μ^*. This is sufficient to guard against the large loss $L_0 = 4$ that could be incurred if $\mu > \mu^*$, but it also takes into account the loss $L_1 = 1$ that could happen if $\mu < \mu^*$. On the other hand, when n is small, then one is not getting much information, and it may be better to be conservative and decide for the alternative in most situations. After all, one does not have as much to lose this way. It is making the best of a bad situation.

It is clear that this analysis is rather artificial. On the other hand, one does learn that the choice of critical point c and the corresponding level α may vary with sample size. For large sample size, the c is large and the level α is small. For small sample size it is just the opposite. This is supposed to guard against the worst possible risk.

4.9 Supplement: P-values

The P-value is a sample statistic that sometimes occurs in the context of hypothesis testing. Say that for instance we have a test of a null hypothesis $\mu = \mu_0$ against a one-sided alternative $\mu > \mu_0$. The test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \tag{4.53}$$

If the level of the test is to be α, then under the null hypothesis this statistics has the appropriate t distribution. So one can calculate a critical value c such that under the null hypothesis

$$P[t > c] = \alpha. \tag{4.54}$$

The test is then to decide for the null hypothesis if $t \leq c$ and to decide for the alternative hypothesis if $t > c$. If the null hypothesis is true, then the probability of making an error is α.

If the alternative hypothesis is true, then the t statistic as defined above does not have the t distribution. This is because the number that is subtracted from the sample mean is μ_0 which is not the true mean μ. If μ is considerably larger than μ_0 and if n is reasonably large, then the probability that $t > c$ is close to one. Thus under the alternative hypothesis the probability of making an error is small.

The P-value is a sample statistic defined as follows. Consider an independent repetition of the experiment, say to be done in the future. Let the test statistic for this experiment be denoted by t'. One regards this as an experiment for which the null hypothesis happens to be true. The P value is this probability under the null hypothesis of a value of t' in the future experiment that is greater than the present experimental value of t.

Another way of thinking of this is the following. Let F be the cumulative probability distribution function for the t distribution. Then for a one-sided test like this, the P-value is $1 - F(t)$, where the value of t in this formula comes from the result of the present experiment.

Example: Say n is large, so that under the null hypothesis t has an approximately standard normal distribution. Say that $\alpha = 0.05$, so the critical value $c = 1.645$. The test is to decide for the alternative if $t > 1.645$. Say that the experimental value of t is 3.18. Then the decision is for the alternative. The P-value is the probability that $t' > 2.85$, which is a very small number, about 0.002. Notice that the P-value is also an experimental number, since it depends on the experimental number 2.85. The fact that the P-value is smaller than α is another way of saying that one should decide for the alternative.

Some people think of a tiny P-value as an indication that an event of very small probability under the null hypothesis has occured. There are problems with this view. An event is a question about the experiment, and the question must be specified in advance. The event that $t' > t(\omega)$ for the future t' depends on the experimental value of $t(\omega)$, and so does not meet this condition. On the other hand, there are all sorts of events that have small probability under the null hypothesis. No matter how the experiment turns out, some of these improbable events are bound to occur. Example: For each integer k, consider the event A_k that $k/1,000,000 < t \leq (k+1)/1,000,000$. Surely one of these events will occur. Yet each of them has small probability.

Clearly it is necessary to specify events that have some importance. Events such as $t > c$ that have a better explanation under the alternative hypothesis are reasonable to consider. But then one is in a framework that goes beyond small probability under the null hypothesis. The alternative hypothesis comes explicitly into the picture.

The P-value has the property that a test of level α will decide for the alternative precisely when P is less than α. So some statisticians want to think of the P-value as a measure of the evidence in favor of the alternative. A smaller P-value is stronger evidence.

However this point of view is no longer hypothesis testing, at least if one considers hypothesis testing as making a decision between two courses of action. It does make some kind of sense, however, if one considers it as estimation.

Here is how to think of P-values in the context of estimation. Consider a quantity θ that is 1 if the null hypothesis $\mu = \mu_0$ is true and 0 if the alternative hypothesis $\mu > \mu_0$ is true. Say that one wants to use some statistic T to estimate θ. So θ is either 1 or 0, but which one is unknown. Obviously, if the null hypothesis is true, then one would like to have the statistic T tend to be near 1. If the alternative hypothesis is true, then one would like T to be near 0. One would like T to tend to be close to θ in any case. If the evidence is inconclusive, then a T somewhere near one half is a conservative compromise, since the distance from θ (which is either 0 or 1) will not be much more than one half. The P-value is a candidate for such a statistic T. When t is negative, then the alternative hypothesis begins to look less plausible, and then the P-value is not too far from 1. When t is very positive, then the alternative hypothesis begins to look convincing, and correspondingly the P-value gets very close to 0.

This use of the P value can be criticized, however, as naive. For one reason, it is not clear why one should use $1-F$, rather than some other decreasing function of the test statistic. This particular function does not use any information about the distribution of the test statistic under the alternative hypothesis. Thus it misses what would be an important part of any serious analysis.

In some cases it may be better to forget about P-values and deal directly with estimation. In other cases the proper framework may be a multiple decision problem. Thus one may have to decide between three actions: decide for the null hypothesis, decide for the alternative hypothesis, or report a failure to get conclusive evidence. Each of these actions may have practical consequences. It is not difficult to extend the ideas of hypothesis testing to this more general kind of multiple decision problem.

4.10 Problems

Each of these problems is a hypothesis test. For each situation, specify the null hypothesis and specify the alternative hypothesis. Set up the test so that if the null hypothesis is true, then the probability of incorrectly rejecting it is five percent. Specify the test. Perform the test with the data, and describe its conclusion.

1. Consider an experiment in which n pairs of individuals are carefully matched. One member of the pair is treated; the other is untreated. A blood sample in taken from each individual and a certain characteristic is measured. The data in Data Set 4.1 give the differences of these characteristics (treated minus untreated). Does the treatment make a difference, one way or the other?

2. Batteries without additives have average lifetime μ_1. Batteries with additives have average lifetime μ_2. Data Set 4.2 involves two samples of size

n_1 and n_2. Does this information indicate that $\mu_2 > \mu_1$? [It is natural to do this test with the usual pooled estimate of standard deviation. If, however, the lifetimes are regarded as exponentially distributed, then the population mean and the population standard deviation are equal, and so the pooled sample mean gives a natural estimate of the standard deviation.]

3. An experiment was conducted with n pairs of identical twins. One twin was treated, the other was untreated. A health trait was measured in each twin. Data Set 4.3 shows the twins for which the treated twin had the greater value of the health trait. Does the treatment improve health?

4. A television broadcast was intended to inform people of a certain issue. Let p_1 be the proportion of people informed about the issue in an area without the broadcast. Let p_2 be the proportion in an area with the broadcast. Data Set 4.4 gives samples of sizes n_1 and n_2. Do the data indicate that $p_2 > p_1$?

5. A survey studied smoking and abnormal hair loss. The sample size was 1200. The number of smokers with abnormal hair loss was 112. The number of smokers without abnormal hair loss was 325. The number of non-smokers with abnormal hair loss was 155. The number of non-smokers without abnormal hair loss was 608. Are smoking and abnormal hair loss independent?

Chapter 5

Order statistics

5.1 Sample median and population median

The picture is that there is a very large (theoretically infinite) population. Each member of the population has some characteristic quantity X. Consider a number α between zero and one. Then there is supposed to be a number t_α such that the proportion of the population for which the X is less than or equal to t_α is α.

One can think of taking a single random member of the population and measuring this quantity X_1. The assumption is that X_1 is a continuous random variable. Then the cumulative distribution function $F(t) = P[X \leq t]$ is continuous. It follows that there is a t_α such that $F(t_\alpha) = \alpha$.

There are several common examples. The most important is the value such that half the population is above this value and half the population is below this value. Thus when $\alpha = 1/2$ the corresponding $t_{\frac{1}{2}}$ is called the population median m.

Similarly, when $\alpha = 1/4$ the $t_{\frac{1}{4}}$ is called the lower population quartile. In the same way, when $\alpha = 3/4$ the $t_{\frac{3}{4}}$ is called the upper population quartile. In statistics the function F characterizing the population is unknown. Therefore all these t_α are unknown quantities associated with the population.

Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$. Thus again we have independent random variables all with the same distribution. We are assuming that the distribution is continuous. Thus the probability is one that for all $i \neq j$ the quantities $X_i \neq X_j$ are unequal.

The order statistics $X_{(1)}, \ldots, X_{(n)}$ are the quantities obtained by arranging the random variables $X_1, \ldots, X_n$ in increasing order. Thus by definition

$$X_{(1)} < X_{(2)} < \cdots < X_{(i)} < \cdots < X_{(n-1)} < X_{(n)}. \tag{5.1}$$

The order statistics are no longer independent, as we now see.

The joint density of $X_1, \ldots, X_n$ is $f(x_1) \cdots f(x_n)$. This product structure is equivalence to the independence of the random variables. On the other hand, the joint density of the order statistics $X_{(1)}, \ldots, X_{(n)}$ is $n! f(x_1) \cdots f(x_n)$ for $x_1 < x_2 < \cdots < x_n$ and zero otherwise. There is no way to factor this. The order statistics are far from independent.

The order statistics are quite useful for estimation. Take $\alpha = i/(n+1)$. Then it seems reasonable to use the order statistics $X_{(i)}$ to estimate t_α.

Thus, for instance, if n is odd and $i = (n+1)/2$ and $\alpha = 1/2$, then $X_{(i)}$ is the sample median. This estimates the population median $m = t_{\frac{1}{2}}$.

The fundamental theorem on order statistics is the following. It shows that questions about order statistics reduce to questions about binomial random variables.

Theorem 5.1 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. For each x, let $N_n(x)$ be the number of i such that $X_i \le x$. Then $N_n(x)$ is a binomial random variable with parameters n and $F(x)$. Furthermore,*

$$P[X_{(j)} \le x] = P[N_n(x) \ge j]. \tag{5.2}$$

This result can be stated even more explicitly in terms of the binomial probabilities. In this form it says that if $P[X_i \le x] = F(x)$, then

$$P[X_{(j)} \le x] = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}. \tag{5.3}$$

This theorem is remarkable, in that it gives a rather complete description of order statistics for large sample sizes. This is because one can use the central limit theorem for the corresponding binomial random variables.

Theorem 5.2 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Fix α and let $F(t_\alpha) = \alpha$. Let $n \to \infty$ and let $j \to \infty$ so that $\sqrt{n}(j/n - \alpha) \to 0$. Then the order statistics $X_{(j)}$ is approximately normally distributed with mean*

$$E[X_{(j)}] \approx t_\alpha \tag{5.4}$$

and standard deviation

$$\sigma_{X_{(j)}} \approx \frac{\sqrt{\alpha(1-\alpha)}}{f(t_\alpha)\sqrt{n}}. \tag{5.5}$$

Proof: Compute

$$P[X_{(j)} \le t_\alpha + \frac{a}{\sqrt{n}}] = P[\frac{N_n(t_\alpha + \frac{a}{\sqrt{n}})}{n} \ge \frac{j}{n}]. \tag{5.6}$$

The random variable has mean $\alpha_n' = F(t_\alpha + a/\sqrt{n})$ and variance $\alpha_n'(1-\alpha_n')/n$. So to standardize the random variable we need to subtract α_n' and multiply by

$\sqrt{n}/\sqrt{\alpha'_n(1-\alpha'_n)}$. So we can use the central limit theorem to write this as the probability that Z is greater than or equal to $j/n - \alpha'_n$ times $\sqrt{n}/\sqrt{\alpha'_n(1-\alpha'_n)}$. Write this as $(j/n - \alpha) + (\alpha - \alpha'_n)$ times $\sqrt{n}/\sqrt{\alpha'_n(1-\alpha'_n)}$. If we use the assumption that $\sqrt{n}(j/n - \alpha) \to 0$, we get

$$P[X_{(j)} \le t_\alpha + \frac{a}{\sqrt{n}}] \approx P[Z \ge \frac{\alpha - \alpha'_n}{\sqrt{\alpha'_n(1-\alpha'_n)}/\sqrt{n}}]. \tag{5.7}$$

However

$$\alpha'_n - \alpha = F(t_\alpha + \frac{a}{\sqrt{n}}) - F(t_\alpha) \approx f(t_\alpha)\frac{a}{\sqrt{n}}. \tag{5.8}$$

So we use this together with $\alpha_n \to \alpha$ to get

$$P[X_{(j)} \le t_\alpha + \frac{a}{\sqrt{n}}] \approx P[Z \ge -\frac{f(t_\alpha)a}{\sqrt{\alpha(1-\alpha)}}]. \tag{5.9}$$

In other words,

$$P[X_{(j)} \le t_\alpha + \frac{a}{\sqrt{n}}] \approx P[-\frac{\sqrt{\alpha(1-\alpha)}}{f(t_\alpha)}Z \le a]. \tag{5.10}$$

This gives the result.

Corollary 5.1 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Let m be the population median. Consider sample sizes n that are odd, so that the sample median $M_n = X_{(\frac{n+1}{2})}$ is defined. Let $n \to \infty$. Then the sample median M_n is approximately normally distributed with mean*

$$E[M_n] \approx m \tag{5.11}$$

and standard deviation

$$\sigma_{M_n} \approx \frac{1}{2f(m)\sqrt{n}}. \tag{5.12}$$

5.2 Comparison of sample mean and sample median

Say that each X_i has density f with population mean

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{5.13}$$

and population variance

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx. \tag{5.14}$$

Then our tendency is to use the sample mean $\bar{X}_n$ to estimate the population mean μ.

Say that

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \tag{5.15}$$

is the distribution function. Say that m is the population median, so $F(m) = \frac{1}{2}$. Then our tendency is to use the sample median M_n to estimate the population median m.

Now say that the density function is symmetric about some point. Then the population mean is the same as the population median. So the sample mean and the sample median are both trying to estimate the same quantity. Which is better?

The relative efficiency of the sample median to the sample mean is found by seeing which has smaller variance. The smaller the variance, the more efficient the estimation. The efficiency of the sample median with respect to the sample mean is the variance of the sample mean divided by the variance of the sample median. This ratio is

$$\frac{\sigma^2/n}{1/(4f(m)^2 n)} = 4f(m)^2\sigma^2. \tag{5.16}$$

If this ratio is less than one, then the sample mean is a better estimator. If this ration is greater than one, then the sample median is better.

For the normal distribution $f(m) = f(\mu) = 1/(\sqrt{2\pi}\sigma)$. Thus the relative efficiency of the sample median to the sample mean is $2/\pi$. This is not particularly good. If a statistician somehow knows that the population distribution is normal, then the sample mean is the better statistic to use.

However it is quite possible that the density value $f(m) > 0$ is well away from zero, but the distribution has long tails so that the σ is huge. Then the median may be much more efficient than the mean. So in many situations the median is the safer estimator to use.

Of course, maybe the statistician does not know the size of $f(m)$ or the size of σ. Some kind of preliminary analysis of the data is then necessary to establish a preference.

5.3 The Kolmogorov-Smirnov statistic

Say that one has a hypothesis that the population distribution is F and wants to check it. The following theorem gives a method.

Theorem 5.3 *Let X be a random variable with a continuous distribution F. Then $U = F(X)$ is a uniform random variable.*

Theorem 5.4 *Let $X_1, \ldots, X_n$ be independent random variables with the same continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Let $U_1 = F(X_1), \ldots, U_n = F(X_n)$. Then these are independent random variables, each uniformly distributed in the interval $[0, 1]$. Their order statistics are $U_{(1)} = F(X_{(1)}), \ldots, U_{(n)} = F(X_{(n)})$.*

The method is to compute the $U_{(1)}, \ldots, U_{(i)}, \ldots, U_{(n)}$ and compare them with $1/(n+1), \ldots, i/(n+1), \ldots, n/(n+1)$. If they are close, then this is a confirmation of the hypothesis.

There is a famous Kolmogorov-Smirnov statistic that is based on this general idea. This statistics gives a quantitative measure of the degree to which the order statistics $U_{(i)} = F(X_{(i)})$ behave as anticipated by the hypothesis.

In this statistic, the comparison is between the order statistics $U_{(i)}$ and the numbers $(i - 1/2)/n$ (rather than $i/(n+1)$). The reason for using $(i-1/2)/n$ seems technical. The following may provide some motivation. The proportion of order statistics less than the ith order statistic is $(i-1)/n$; the proportion of order statistics less than or equal to the ith order statistics is i/n. The average of these two numbers is $(i - 1/2)/n$. It may not matter all that much which definition is used, since the random errors are going to be proportional to $1/\sqrt{n}$, which is considerably bigger.

Thus the Kolmogorov-Smirnov statistic is defined to be

$$D = \frac{1}{2n} + \max_{1 \le i \le n} \left| F(X_{(i)}) - \frac{i - \frac{1}{2}}{n} \right|. \tag{5.17}$$

The first term is there because the Kolmogorov-Smirnov statistics is usually defined by a different and perhaps more natural formula, from which this term emerges after a calculation. Again it should not matter much for large n.

A typical result about the Kolmogorov-Smirnov statistics is that if F is the distribution function of the X_i, then for moderately large n

$$P[D > \frac{1.36}{\sqrt{n}}] \le 0.05. \tag{5.18}$$

The fact that this is proportional to $1/\sqrt{n}$ is not so surprising, as this is the usual amount of misalignment of order statistics. This provides a test of whether the data X_i really come from a population described by F. If so, then large values of D are unlikely.

5.4 Other goodness of fit statistics

Recall that the Kolmogorov-Smirnov statistic is

$$D = \frac{1}{2n} + \max_{1 \le i \le n} \left| U_{(i)} - p_i \right|, \tag{5.19}$$

where $p_i = (i - 1/2)/n$ and the random variable $U_{(i)} = F(X_{(i)})$.

Instead of this, one can use the Cramér-von Mises statistics

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} (U_{(i)} - p_i)^2 \tag{5.20}$$

for a test. This is even more convenient, and in some circumstances the resulting test has better power. Furthermore, calculations with this statistic tend to

be relatively simple. If the $U_{(i)}$ are indeed order statistics from a uniform distribution, then for large n the expectation of W^2 is about $1/6$. It may be shown that the top five percent cutoff is about 0.46. That is, if the order statistics are really order statistics from a uniform distribution, then $P[W^2 > 0.46] \approx 0.05$. This may be used for a test.

Yet another possible choice is the Anderson-Darling statistic. This may be even better, especially when one wants to detect large deviations in the tails. This statistic is

$$A^2 = -n - 2\sum_{i=1}^{n}[p_i \log(U_{(i)}) + (1-p_i)\log(1-U_{(i)})]. \tag{5.21}$$

Notice that the expression on the right is minimized when each $U_{(i)} = p_i$. Furthermore, when we expand to second order we see that

$$A^2 \approx -n - 2\sum_{i=1}^{n}[p_i \log(p_i) + (1-p_i)\log(1-p_i)] + \sum_{i=1}^{n}\frac{(U_{(i)}-p_i)^2}{p_i(1-p_i)}. \tag{5.22}$$

This shows that this is very much like the Cramér-von Mises statistic, except that the order statistics near the extremes receive greater emphasis. For large n the first two terms should cancel. This is because $-2\int_0^1 [p\log p + (1-p)\log(1-p)]\,dp = 1$. This gives the even more suggestive form

$$A^2 \approx \sum_{i=1}^{n}\frac{(U_{(i)}-p_i)^2}{p_i(1-p_i)}. \tag{5.23}$$

It may be shown that for large n the expectation of A^2 is about one. This makes it plausible that the top five percent cutoff is about 2.5. That is, if the order statistics are really order statistics from a uniform distribution, then $P[A^2 > 2.5] \approx 0.05$. This may be used for a test.

5.5 Comparison with a fitted distribution

The Kolmogorov-Smirnov statistic and its relatives are suited for the case when the null hypothesis is that the underlying population has a given distribution F. However, often the hypothesis is simply that the underlying population belongs to a parametric family. For instance, it might be normal, but with unknown μ and σ.

In such a case, it is tempting to use $\bar{X}$ and s to estimate μ and σ. Then one can perform the test with $\hat{F}$, the distribution obtained from the $\bar{X}$ and s parameters.

It is important to realize that this is going to make these statistics tend to be smaller, since the data itself was used to defined the distribution with which the data is compared. Thus one should use such a test with caution. It is necessary to use tables specially prepared for such a situation. On the other hand, if the

statistic is big, then this gives rather good evidence that the population does not belong to the parametric family.

There are many variations on the idea of the Kolmogorov-Smirnov statistic. As mentioned, it would be possible to compare $F(X_{(i)})$ with $i/(n+1)$ instead of with $(i-1/2)/n$. Or one could compare $X_{(i)}$ directly with $F^{-1}(i/(n+1))$. In either case one should get something rather close to a straight line.

5.6 Supplement: Uniform order statistics

For uniform random variables it is easy to do explicit computations with the order statistics. In this case the joint density is $n!$ for $x_1 < x_2 < \cdots < x_n$ and zero otherwise.

Theorem 5.5 *Let $U_1, U_2, U_3, \ldots, U_n$ be independent random variables each uniformly distributed on $[0,1]$. Let $U_{(1)}, U_{(2)}, U_{(3)}, \ldots, U_{(n)}$ be their order statistics. Then the expectation of $U_{(i)}$ is*

$$E[U_{(i)}] = \frac{i}{n+1}. \tag{5.24}$$

Theorem 5.6 *Let $U_1, U_2, U_3, \ldots, U_n$ be independent random variables each uniformly distributed on $[0,1$. Let $U_{(1)}, U_{(2)}, U_{(3)}, \ldots, U_{(n)}$ be their order statistics. Then the variance of $U_{(i)}$ is*

$$\mathrm{Var}(U_{(i)}) = \frac{1}{n+2}\frac{i}{n+1}\left(1 - \frac{i}{n+1}\right). \tag{5.25}$$

Note that if $n \to \infty$ is large and $i/(n+1) \to \alpha$, then the mean of $U_{(i)}$ converges to α and n times the variance of $U_{(i)}$ converges to $\alpha(1-\alpha)$. So this is consistent with the general picture given before. The nice thing is that one can do the calculations with fixed n.

We can use these results to obtain a better understanding of some of the statistics that we have been considering. Consider the Cramér-von Mises statistics

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}(U_{(i)} - p_i)^2. \tag{5.26}$$

We neglect the first term, which is small. Make the crude approximation that for large n each $U_{(i)}$ has variance $p_i(1-p_i)/n$. Then each term has expectation approximately equal to this quantity. The expectation of W^2 is the sum of these expectations. If we approximate the sum by an integral, then we get the answer $\int_0^1 p(1-p)\,dp = 1/6$.

The argument is similar for the Anderson-Darling statistic

$$A^2 = -n - 2\sum_{i=1}^{n}[p_i \log(U_{(i)}) + (1-p_i)\log(1-U_{(i)})]. \tag{5.27}$$

Use the large n approximation

$$A^2 \approx \sum_{i=1}^{n} \frac{(U_{(i)} - p_i)^2}{p_i(1 - p_i)}. \tag{5.28}$$

Then each term has variance about $1/n$. Thus the expectation of A^2 is about one.

5.7 Problems

1. For uniform random variables it is easy to do explicit computations with the order statistics. In this case the joint density is $n!$ for $x_1 < x_2 < \cdots < x_n$ and zero otherwise. Let $U_1, U_2, U_3, \ldots, U_n$ be independent random variables each uniformly distributed on $[0, 1]$. Let $U_{(1)}, U_{(2)}, U_{(3)}, \ldots, U_{(n)}$ be their order statistics. Show that the expectation of $U_{(i)}$ is

$$E[U_{(i)}] = \frac{i}{n + 1}. \tag{5.29}$$

2. Data Set 5.1 is a sample of size 49 from a Cauchy distribution with median $m = 100$ and scale $s = 10$. Thus the density is

$$f(x) = c\left(\frac{x - m}{s}\right)\frac{1}{s}, \tag{5.30}$$

where

$$c(z) = \frac{1}{\pi}\frac{1}{z^2 + 1}. \tag{5.31}$$

Compute the theoretically the standard deviation of the sample mean. Compute theoretically the standard deviation of the sample median. Compute the sample mean for the data. Compute the sample median for the data. How well do they do? How does this compare with the theory?

3. Data Set 5.2 is purported to be a random sample of size 35 from a normal population with mean 41 and standard deviation 5, but it might be that this is incorrect. Use the Kolmogorov-Smirnov statistic to test this hypothesis.

Chapter 6

The bootstrap

6.1 Bootstrap samples

Consider the situation of an independent sample $X_1, \ldots, X_n$ from some population. The statistic $\hat{\theta}$ computed from the sample that is an estimate of some parameter θ of the population. Thus θ could be the population mean, and $\hat{\theta}$ could be the sample mean. Or θ could be the population median, and $\hat{\theta}$ could be the sample median. In order to see how well $\hat{\theta}$ estimates θ, it is useful to know the variance of the random variable $\hat{\theta}$. However in order to do this, one has to know something about the population.

Nevertheless, there is a general method to estimate the variance of $\hat{\theta}$ from the sample. This method is known as the *bootstrap*. The idea of the bootstrap is to treat the sample as a new population. Thus let $X_1, \ldots, X_n$ be the sample values. We think of this as a finite population. We consider the experiment of taking an ordered random sample with replacement of size n from this finite population. Thus $X_1^*, \ldots, X_n^*$ are independent random variables, each of which can take on the value X_i with probability $1/n$. That is,

$$P[X_i^* = X_j] = \frac{1}{n}. \tag{6.1}$$

Once we have the bootstrap sample, then we can compute the corresponding random variable $\hat{\theta}^*$. Thus $\hat{\theta}^*$ could be the sample mean or the sample median of the bootstrap sample $X_1^*, \ldots, X_n^*$. The idea is to use the variance of $\hat{\theta}^*$ (with $X_1, \ldots, X_n$ fixed) as an estimate of the variance of $\hat{\theta}$. This variance is called the *ideal bootstrap estimator*.

How can one compute the ideal bootstrap estimator? In principle, it is simple. There are a finite number n^n of bootstrap samples, each with probability $1/n^n$. For each such bootstrap sample, compute the value of $\hat{\theta}^*$. Compute the mean of these n^n numbers. Then compute the mean of the squared deviations from this mean. This number is the ideal bootstrap estimator. The problem with this, of course, is that n^n is a huge number. Enumerating all the bootstrap samples in this way is completely impractical, even for moderate sized n.

6.2 The ideal bootstrap estimator

There is one case where one can carry out the computation explicitly. Say that we are interested in the population mean $\theta = \mu$. Consider the case when $\hat{\theta}$ is the sample mean

$$\hat{\theta} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{6.2}$$

This is the obvious estimate of μ. The variance of this $\bar{X}$ is known to be σ^2/n, but unfortunately we do not know σ^2.

In the context of the bootstrap this sample mean is considered as if it were a population mean for a finite population. The corresponding variance for this finite population is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}. \tag{6.3}$$

Then

$$\hat{\theta}^* = \frac{X_1^* + \cdots + X_n^*}{n} \tag{6.4}$$

is the bootstrap sample mean. In independent sampling, we know that the variance of such a sample mean is given by the variance of the population divided by the sample size. This shows that the ideal bootstrap estimator, the bootstrap variance of $\hat{\theta}^*$ is precisely

$$\mathrm{Var}(\hat{\theta}^*) = \frac{\hat{\sigma}^2}{n}. \tag{6.5}$$

This is almost exactly the usual estimate of variance of the sample means. The only modification is that the denominator is n instead of $n - 1$. Thus the bootstrap for this case has given us nothing essentially new.

6.3 The Monte Carlo bootstrap estimator

On the other hand, say that we are interested in the population median $\theta = m$. The estimator for that case is the sample median $\hat{\theta} = M$, the middle order statistic for the sample $X_1, \ldots, X_n$.

Each time we take a bootstrap sample $X_1^*, \ldots, X_n^*$ we get a bootstrap sample median $\hat{\theta}^* = M^*$. We would like to compute the variance of this random variable, but this would require summing over all n^n samples. There is no simple formula for the result.

The solution is to use the Monte Carlo method, that is, to take some number B of bootstrap samples (say a couple hundred), compute the statistic $\hat{\theta}^*$ for each sample, and then use the sample variance of these numbers to estimate the ideal bootstrap estimate of variance.

Thus the Monte Carlo bootstrap estimator is obtained as follows. Take the B bootstrap samples. Compute the sample mean

$$\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}. \tag{6.6}$$

and then the sample variance

$$\frac{\sum_{b=1}^{B}(\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2}{B-1}. \tag{6.7}$$

as an estimate of the ideal bootstrap estimator of variance.

This procedure needs a computer, but it is practical. Thus one takes B random bootstrap samples. For each bootstrap sample one computes its sample median. This gives B numbers. Then one computes the sample mean and sample variance of these B numbers. The latter is the Monte Carlo bootstrap estimator.

There are two sources of error in this procedure. The first was caused by using the sample $X_1, \ldots, X_n$ as a way of getting imperfect information about the population. There is nothing that can be done about this error if these are all the available data. The second is the error obtained by taking B bootstrap samples in the Monte Carlo method, rather than looking at all bootstrap samples. This error can be decreased at will, at the expense of more computer computation, simply by taking a larger value of B.

The actual computer simulation is fairly simple. You need the original data points $X_1, \ldots, X_n$. You need a random number generator that will generate the numbers $1, 2, 3, \ldots, n$ with equal probability. One repeats the following procedure B times. Run the random number generator n times, generating numbers $i_1, \ldots, i_n$. Pick the data points $X_1^* = X_{i_1}, \ldots, X_n^* = X_{i_n}$ corresponding to these n numbers. (Usually some data points X_i will be picked more than once, but this is inherent in sampling with replacement, and it is necessary for the method to work.) Calculate the statistic (median or whatever) from these $X_1^*, \ldots, X_n^*$. This calculated value of the statistic is recorded. Then at the end one uses these B calculated values as the input to a program that computes the sample mean and sample variance is the usual way. This final sample variance is the Monte-Carlo bootstrap estimator.

6.4 Supplement: Sampling from a finite population

This section records some basic facts about sampling from a finite population. Say that we have a sample of size n from a population of size m. The bootstrap is a special case in which we take $m = n$ and sample with replacement.

The first kind of sampling is sampling *with replacement*. In this case, an *ordered sample* is a function f from the set $\{1, 2, 3, \ldots, n\}$ to the population M of size m. There are m^n such samples. Each such sample has the same probability

$$P[f] = \frac{1}{m^n}. \tag{6.8}$$

In the case of sampling with replacement, an *unordered sample* consists of a function χ from M to the natural numbers such that $\sum_{p \in M} \chi(p) = n$. This

function is called an *occupation number* function. It measures the number of times each element of the population is selected. The number of occupation number functions is given by the binomial coefficient $\binom{m+n-1}{n}$. The number of ordered samples corresponding to occupation number function is given by the multinomial coefficient $\binom{n}{\chi}$, where χ is the list of occupation numbers. This multinomial coefficient may be expressed in terms of binomial coefficients as

$$\binom{n}{\chi} = \binom{n}{\chi(1)}\binom{n-\chi(1)}{\chi(2)}\binom{n-\chi(1)-\chi(2)}{\chi(3)}\cdots\binom{n-\chi(1)-\chi(2)-\cdots\chi(m-1)}{\chi(m)}. \tag{6.9}$$

This is because to choose a function with given occupation numbers, you first have to choose the $\chi(1)$ elements of the index set $\{1,\ldots,n\}$ that go to the first member of the population, then you have to choose among the remaining elements $n-\chi(1)$ elements of the index set the $\chi(2)$ elements that go to the second member of the population, and then among the remaining $n-\chi(1)-\chi(2)$ elements of the index set the $\chi(3)$ elements that go to the third member of the population, and so on. Thus the probability of an unordered sample is

$$P[\chi] = \binom{n}{\chi}\frac{1}{m^n}. \tag{6.10}$$

Notice that these probabilities are not all the same. For each unordered sample, one has to work out the corresponding multinomial coefficient.

Consider a random variable X for the sampling with replacement experiment. Then its expectation may be computed by the finite sum over ordered samples

$$E[X] = \sum_f X(f)\frac{1}{m^n}. \tag{6.11}$$

If the random variable does not depend on the order of the sampling, then it may also be computed by a finite sum over unordered samples

$$E[X] = \sum_\chi X(\chi)\binom{n}{\chi}\frac{1}{m^n}. \tag{6.12}$$

The second kind of sampling is sampling *without replacement*. In this case, an *ordered sample* is a one-to-one function f from the set $\{1,2,3,\ldots,n\}$ to the population M of size m. There are $(m)_n = m(m-1)(m-2)\cdots(m-n+1)$ such samples. Each such sample has the same probability

$$P[f] = \frac{1}{(m)_n}. \tag{6.13}$$

In the case of sampling without replacement, an *unordered sample* consists of a subset χ of M with n elements. This subset consists of the elements of the population that are selected. The number of subsets is given by the binomial

coefficient $\binom{m}{n}$. The number of ordered samples corresponding to each unordered sample is given by the $n!$. Thus the probability of an unordered sample is

$$P[\chi] = n! \frac{1}{(m_n)} = \frac{1}{\binom{m}{n}}. \tag{6.14}$$

Consider a random variable X for the sampling without replacement experiment. Then its expectation may be computed by the finite sum over ordered samples

$$E[X] = \sum_f X(f) \frac{1}{(m)_n}. \tag{6.15}$$

If the random variable does not depend on the order of the sampling, then it may also be computed by a finite sum over unordered samples

$$E[X] = \sum_\chi X(\chi) \frac{1}{\binom{m}{n}}. \tag{6.16}$$

The bootstrap is the case of sampling with replacement when $m = n$. Thus the number of ordered samples is n^n. This is huge. However most random variables that one would want to consider do not depend on the order. Therefore one can compute the expectation by summing over the unordered samples and weighting by the corresponding probabilities. There are $\binom{2n-1}{n}$ unordered samples, which seems considerably less. However this number is asymptotically $(1/\sqrt{\pi n})2^{2n-1}$ which is still huge even for moderate n.

Sampling without replacement can be used in a similar way for some statistics problems. One situation is when there are two populations. The null hypothesis is that they are the same. One takes samples of size n_1 and n_2 from the two populations. A test statistic is computed in terms of the two samples. If the null hypothesis were true, then it should not matter which of the $n_1 + n_2 = m$ values come from the first population. So to see how variable the test statistic is under the null hypothesis, one can consider samples without replacement of size n_1 from the m data points. For each such sample the test statistics has some value. The behavior of the test statistic in this new situation is supposed to give an indication of what would happen in the original situation, if the null hypothesis were true. This is sometimes called the permutation test.

6.5 Problems

1. A certain population consists of a large number of fuel containers with varying amounts of fuel. A sample of size $n = 3$ was taken and the sample values were 1, 3, 4. The sample median of 3 is the statistician's estimate of the population median for the population of fuel containers. The statistician wishes to get an estimate of the standard deviation of the sample medians from the fuel container population in this kind of experiment. The statistician instead considers random samples of size

$n = 3$ with replacement from the population consisting of the date points 1, 3, 4. For each such sample there is a median. Compute the expectation of the median and the standard deviation of the median. The latter is the ideal bootstrap estimator.

2. Take $B = 10$ random samples with replacement from the three sample values 1, 3, 4. For each such sample there is a median. Calculate the sample mean and the sample standard deviation of the medians from these samples. This is the Monte-Carlo bootstrap estimator. Compare the Monte-Carlo bootstrap estimator with the ideal bootstrap estimator.

Chapter 7

Variance and bias in estimation

7.1 Risk

A common estimation problem is when there is an unknown population parameter θ. One wants to estimate this using data $X_1, \ldots, X_n$. The probabilistic nature of the sample is described by a probability model that depends on θ. Therefore we want a random variable T that depends on the $X_1, \ldots, X_n$ such that no matter what θ is, T is likely to be close to θ.

Say that we measure the *loss* from making an estimate T that is not completely accurate by the squared error, so that the loss is $(T - \theta)^2$. Then when one does the experiment, this is a measure of the actual amount of pain inflicted by not being right on target.

It seems reasonable to try to use a procedure so that the *risk*, or expected loss, is small. Thus one would like the procedure to make $E_\theta[(T - \theta)^2]$ small. This quantity measures the amount of pain in the long run. Of course for any particular experiment one may have a greater or less loss, but as a measure of how the procedure is doing this quantity is reasonable.

The risk may be computed in terms of the variance and the *bias* $b(\theta) = E_\theta(T) - \theta$. The formula is

$$E_\theta[(T - \theta)^2] = \operatorname*{Var}_\theta(T) + (E_\theta[T] - \theta)^2. \tag{7.1}$$

So in general one wants small variance and small bias.

One method might be to use unbiased estimators. For such an estimator the risk is just the variance. We now look at some properties of unbiased estimators.

Example 1. Assume that the standard deviation σ is known, and we want to estimate the mean μ. We take a sample of size n. Say that we use the sample mean $T = \bar{X}$. Then $\bar{X}$ is an unbiased estimate of μ. The risk is the variance σ^2/n and does not depend on μ.

On the other hand, one could take $T = \mu_0$, independent of the data. Here μ_0 is some guess, or some dictate of authority. For this type of estimator the risk is all bias. In fact the risk is $(\mu - \mu_0)^2$. This is quite pleasant close to μ_0, but it is huge when μ is far from μ_0. Now the point is that one does not know μ_0. So this type of estimator seems unreasonable, though there is no strictly logical reason for ruling it out. Maybe someone with a strong hunch will ignore the evidence—and be right after all. But this is not usually a good strategy for people with a sense of how things can go wrong.

Example 2. Say that the standard deviation σ is known, and we want to estimate the square of the mean μ^2. The expectation of $\bar{X}^2$ is $\mu^2 + \sigma^2/n$. So this is a biased estimator. If we want an unbiased estimate, we would take $\bar{X}^2 - \sigma^2/n$ as the estimate of μ^2. However this has a certain absurd feature: It can give negative results. On the other hand, if we throw away negative results and replace them by zero, we get less risk, but a biased estimator again. The biased estimator is clearly better.

Example 3. Now take a normal population. Assume that the mean μ is known, but the variance σ^2 is to be estimated. For this purpose the obvious estimator is

$$T = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}. \tag{7.2}$$

Then nT/σ^2 is a χ^2_n random variable. A χ^2_n random variable has mean n and variance $2n$. So the mean of T is σ^2 and the variance of T is $2\sigma^4/n$. In particular, T is an unbiased estimate of σ^2. Also, the risk of T is $\mathrm{Var}(T) = 2\sigma^4/n$. So this would seem to be a reasonable estimator. This, however is wrong, because one can decrease the risk without any penalty. Say that one uses the estimator aT. Then the risk of this is

$$a^2 \,\mathrm{Var}(T) + (a\sigma^2 - \sigma^2)^2 = a^2 \frac{2\sigma^4}{n} + (a-1)^2\sigma^4. \tag{7.3}$$

This is minimized when $a = n/(n+2)$. The constant risk is $2/(n+2)\sigma^4$. This third example shows rather conclusively that unbiased estimators are not necessarily desirable!

7.2 Unbiased estimators

In spite of the fact that unbiased estimators may not be particularly good, we may choose to consider them. It may simply be convenient to do so.

The requirement that T be unbiased is that

$$E_\theta[T] = \theta. \tag{7.4}$$

Given an unbiased estimator, we want the variance

$$\operatorname*{Var}_\theta(T) = E_\theta[(T-\theta)^2] \tag{7.5}$$

is as small as possible.

Examples: The sample mean $\bar{X}$ is an unbiased estimator of μ. The sample variance s^2 is an unbiased estimator of σ^2.

Note, however, the following general fact: $E[T^2] = E[T]^2 + \mathrm{Var}(T)$. Therefore, for a non-constant random variable $E[T^2] > E[T]^2$. This says that if T is an unbiased estimator of some parameter, so that $E_\theta[T] = \theta$, then T^2 is not an unbiased estimator of θ^2. In fact, T^2 is biased high, $E_\theta[T^2] > \theta^2$.

On the other hand, if $T \geq 0$ and T^2 is an unbiased estimator of θ^2, so that $E[T^2] = \theta^2$, then T is a biased estimate of θ. In fact, it is biased low: $E_\theta[T] < \theta$.

Example: The sample standard deviation s is a biased estimator of the population standard deviation σ, in fact, $E[s] < \sigma$.

In any case, assuming that we have unbiased estimators, we can try to compare them in the basis of their variance.

Example: Let X be a random variable that is uniform on the interval given by $[\mu - c/2, \mu + c/2]$. We could use the sample mean $\bar{X}$ as an estimate of μ. Or we could use the average $(X_{(1)} + X_{(n)})/2$ of the two extreme order statistics, the minimum and the maximum. These are both unbiased estimators of μ.

First we need some computations with order statistics. The mean of the ith order statistics from a uniform $[0, 1]$ distribution is

$$E[U_{(i)}] = \frac{i}{n+1}. \tag{7.6}$$

The covariance is

$$\mathrm{Cov}(U_{(i)}, U_{(j)}) = \frac{1}{n+2}\ \frac{i}{n+1}\left(1 - \frac{j}{n+1}\right). \tag{7.7}$$

for $i \leq j$. Notice that if the mean $i/(n+1) = \alpha$, then the variance is $1/(n+2)$ times $\alpha(1-\alpha)$. It is thus small near the extreme order statistics.

We can write $X = \mu + c(U - \frac{1}{2})$ and use this to compute the mean and variance of X. It follows that the variance of the sample mean is $\sigma^2/n = c^2/(12n)$. On the other hand, the variance of each of the extreme order statistics is given by $c^2 n/((n+2)(n+1)^2)$. Their covariance is even smaller: $c^2/((n+2)(n+1)^2)$. So the variance of their average is $(1/2)c^2/((n+2)(n+1))$. This has an extra power of n in the denominator! Asymptotically, it is $(1/2)c^2/n^2$. So the method using the minimum and maximum has much smaller risk.

Example: Again, let X be a random variable that is uniform on the interval $[\mu - c/2, \mu + c/2]$. We could use the difference $X_{(n)} - X_{(1)}$ of the two extreme order statistics as an estimator of c. This is slightly biased, in that its expectation is $(n-1)/(n+1)c$. So the bias is $-2/(n+1)c$. The variance of the difference is $2c^2/((n+2)(n+1))$. The contributions to the risk from the variance and from the squared bias are both very small.

These example using the uniform distribution are rather singular, since they depend on the fact that the probability has a very sharp cutoff at a particular point. This is an unusual feature, and so one can get unusually good results.

7.3 The Cramér-Rao bound

Now we turn to a more normal situation, where the probability densities are nicely differentiable with respect to the parameter. The Cramér-Rao bound shows that in this case the variance of an unbiased estimator is always bounded below by a constant times $1/n$. In addition, one can figure out the constant.

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables each with density $f(x \mid \theta)$. This density depends on a population parameter θ. In many contexts we think of the density as a function of the value x with the parameter θ fixed. However now we want to think of this as a function of the parameter θ, with the value x fixed. When this point of view is taken, where $f(x \mid \theta)$ is considered a function of θ, this is called the *likelihood* function. The interest of the likelihood functions is that it shows how sensitive the probabilities of the data are to the population parameter θ.

Consider the function

$$t_1(x \mid \theta) = \frac{\partial \log f(x \mid \theta)}{\partial \theta} \tag{7.8}$$

This is the rate of change of the log likelihood as a function of θ. It measures how sensitive the probabilities are to a change in the parameter.

Let $T_1 = t_1(X \mid \theta)$. First note that

$$\int_{-\infty}^{\infty} f(x \mid \theta)\, dx = 1. \tag{7.9}$$

If we differentiate this with respect to θ, we obtain

$$E_\theta[T_1] = \int_{-\infty}^{\infty} t_1(x \mid \theta) f(x \mid \theta)\, dx = 0. \tag{7.10}$$

Define the *Fisher information*

$$I(\theta) = \operatorname*{Var}_\theta[T_1] = \int_{-\infty}^{\infty} \left(\frac{\partial \log f(x \mid \theta)}{\partial \theta} \right)^2 f(x \mid \theta)\, dx. \tag{7.11}$$

We have the following identity

$$I(\theta) = -E_\theta[\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2}] = -\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} f(x \mid \theta)\, dx. \tag{7.12}$$

Theorem 7.1 *The random variable T_1 obtained by inserting the data into the rate of change of the log likelihood with respect to the parameter has mean zero*

$$E_\theta[T_1] = 0 \tag{7.13}$$

and variance

$$\operatorname*{Var}_\theta(T_1) = I(\theta). \tag{7.14}$$

Example: Consider a normal population mean μ and standard deviation σ. Think of σ as known. The problem is to estimate μ. The log likelihood function is $\log f(x \mid \mu) = -\frac{1}{2}\log(2\pi\sigma^2) - (x-\mu)^2/(2\sigma^2)$. The first partial derivative with respect to μ is $(x-\mu)/\sigma^2$. The negative of the second partial derivative with respect to μ is $1/\sigma^2$. The variance of $(X-\mu)/\sigma^2$ is the same as the expectation of $1/\sigma^2$. This is the Fisher information $I(\mu) = 1/\sigma^2$.

Example: Consider a normal population mean μ and standard deviation σ. Think of μ as known. The problem is to estimate σ^2. The log likelihood function is $\log f(x \mid \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - (x-\mu)^2/(2\sigma^2)$. The first partial derivative with respect to σ^2 is $-\frac{1}{2}/\sigma^2 + (x-\mu)^2/(2\sigma^4)$. The negative of the second partial derivative with respect to σ^2 is $-\frac{1}{2}/\sigma^4 + (x-\mu)^2/\sigma^6$. The expectation of this is $1/(2\sigma^4)$. This is the Fisher information $I(\sigma^2) = 1/(2\sigma^4)$.

Now we can generalize this all to n variables. The *likelihood* function

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) \cdots f(x_n \mid \theta) \tag{7.15}$$

is the joint density of these independent random variables, considered as a function of the parameter θ.

Let

$$t_n(x_1, \ldots, x_n \mid \theta) = \frac{\partial \log f(x_1, \ldots, x_n \mid \theta)}{\partial \theta} \tag{7.16}$$

be the rate of change of the log likelihood as a function of θ.

Let

$$T_n = t_n(X_1, \ldots, X_n \mid \theta) \tag{7.17}$$

be the random variable obtained by inserting the data into the rate of change of the log likelihood function. Note that

$$T_n = t_1(X_1 \mid \theta) + \cdots + t_1(X_n \mid \theta) \tag{7.18}$$

is a sum of independent random variables. This gives the following theorem.

Theorem 7.2 *The random variable T_n obtained by inserting the data into the rate of change of the log likelihood function with respect to the parameter has mean zero*

$$E_\theta[T_n] = 0 \tag{7.19}$$

and variance

$$\operatorname*{Var}_\theta(T_n) = nI(\theta). \tag{7.20}$$

The Cramér-Rao lower bound is the following theorem.

Theorem 7.3 *Let*

$$\delta = d(X_1, \ldots, X_n) \tag{7.21}$$

be an unbiased estimator of θ, so that

$$E_\theta[\delta] = \theta. \tag{7.22}$$

Then

$$\operatorname*{Var}_\theta(\delta) \geq \frac{1}{nI(\theta)} \tag{7.23}$$

Proof: Consider the equation that says that δ is unbiased, and differentiate it with respect to θ. The result is that

$$E_\theta[\delta\, T_n] = 1. \tag{7.24}$$

Since T_n has mean zero, this is the covariance of δ and T_n. Since the correlation of δ and T_n is bounded by one, the covariance is bounded by the product of the standard deviations. Therefore 1 is bounded by the product of the standard deviations of δ and of T_n. That is, the variance of δ is bounded below by the variance of T_n.

This theorem shows that if we are interested in unbiased estimators, then we cannot do better than with an estimator whose variance is exactly $1/(nI(\theta))$. However it does not tell us how to find such an estimator. It may well not exist.

Example: For a normal population with known σ the best unbiased estimator of μ is the sample mean $\bar{X} = \sum_i X_i/n$. The variance is σ^2/n.

Example: For a normal population with known μ the best unbiased estimator of σ^2 is $\sum_i (X_i - \mu)^2/n$. The variance is $2\sigma^4/n$.

7.4 Functional invariance

Another principle that seems appealing is functional invariance. This principle says the method of estimation should be chosen so that if T is the estimate of θ, then $f(T)$ is the estimate of $f(\theta)$.

The following theorem (Jensen's inequality) will show that the principle of functional invariance and the principle of unbiased estimation are incompatible.

Theorem 7.4 *Let T be a random variable that is not constant. Let f be a function such that $f''(t) > 0$ for all t. Then*

$$E[f(T)] > f(E[T]). \tag{7.25}$$

Proof: Let $\theta = E[T]$. Then

$$f(T) = f(\theta) + f'(\theta)(T - \theta) + \frac{1}{2} f''(\tau)(T - \theta)^2, \tag{7.26}$$

where τ is between θ and T. Take expectations. This gives

$$E[f(T)] = f(\theta) + \frac{1}{2} E[f''(\tau)(T - \theta)^2]. \tag{7.27}$$

Since f is strictly convex and T is not constant, the expectation in the second term must be strictly positive.

This result shows that we need either to give up unbiased estimators or to give up the principle of functional invariance. We have already seen that unbiased estimates can give us estimates that have too much risk. We shall see that the principle of functional invariance can also lead us astray. So we shall give up both. However in the next chapter we shall see that for large sample size there is a sense in which one can get approximations to both at once.

7.5 Problems

1. Consider a population with known mean μ but unknown variance σ^2. Take an independent sample $X_1, \ldots, X_n$. Show that the statistic

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

is an unbiased estimator of σ^2.

2. Consider a uniform distribution on the interval from 0 to θ. Take a sample X of size 1. Find the estimator of the form cX that is an unbiased estimator of θ.

3. Consider a sequence of independent success-failure trials, with probability p of success. Let N_n be the number of successes in the first n trials. Let T_r be the trial on which the rth success occurs. Show that $P[T_r = k] = P[N_{k-1} = r - 1]p$. Show that $(r-1)/(T_r - 1)$ is an unbiased estimator of p. Use this to show that r/T_r is a biased estimator of p. Find the sign of the bias.

4. Let $X_1, \ldots, X_n$ be an independent sample from a Poisson population with mean θ. Show that the sample mean is an unbiased estimator of θ. Compute its variance. Use the Cramér-Rao bound to show that it there is no unbiased estimator of θ with a smaller variance.

5. Let

$$f(x \mid \theta) = \frac{1}{\pi}\frac{s}{s^2 + (x - \theta)^2} \tag{7.28}$$

be the Cauchy density with unknown center θ and known spread $s > 0$. Let $X_1, \ldots, X_n$ be an independent sample from this distribution. Find the lower bound for the variance of unbiased estimators of θ.

6. Show that if δ is an estimator of θ with bias $b(\theta)$, then its variance has a lower bound

$$\operatorname*{Var}_{\theta}(\delta) \geq \frac{(1 + b'(\theta))^2}{nI(\theta)}. \tag{7.29}$$

When can this lower bound be identically equal to zero?

Chapter 8

Maximum likelihood estimation

8.1 The likelihood function

We review the theory of the likelihood function. Let $X_1, \ldots, X_n$ be independent, identically distributed random variables each with density $f(x \mid \theta)$. This density depends on a population parameter θ.

Consider the function

$$t_1(x \mid \theta) = \frac{\partial \log f(x \mid \theta)}{\partial \theta}. \tag{8.1}$$

This is the rate of change of the log likelihood as a function of θ. It measures how sensitive the probabilities to a change in the parameter.

First note that

$$\int_{-\infty}^{\infty} f(x \mid \theta)\, dx = 1. \tag{8.2}$$

If we differentiate this with respect to θ, we obtain

$$E_\theta[t_1(X \mid \theta)] = \int_{-\infty}^{\infty} t_1(x \mid \theta) f(x \mid \theta)\, dx = 0. \tag{8.3}$$

Define the *Fisher information*

$$I(\theta) = \operatorname*{Var}_\theta[t_1(X \mid \theta)] = \int_{-\infty}^{\infty} \left(\frac{\partial \log f(x \mid \theta)}{\partial \theta} \right)^2 f(x \mid \theta)\, dx. \tag{8.4}$$

Let

$$t_1'(x \mid \theta) = \frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} \tag{8.5}$$

We have the following identity

$$I(\theta) = -E_\theta[t_1'(X \mid \theta)] = -\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x \mid \theta)}{\partial \theta^2} f(x \mid \theta)\, dx. \tag{8.6}$$

Theorem 8.1 *The random variable T_1 obtained by inserting the data into the rate of change of the log of the density with respect to the parameter has mean zero*

$$E_\theta[T_1] = 0 \tag{8.7}$$

and variance

$$\operatorname*{Var}_\theta(T_1) = I(\theta). \tag{8.8}$$

Theorem 8.2 *The random variable T_1' obtained by inserting the data into the second derivative of the log of the density with respect to the parameter has mean satisfying*

$$-E_\theta[T_1'] = I[\theta]. \tag{8.9}$$

Now we can generalize this all to n variables. Let

$$t_n(x_1, \ldots, x_n \mid \theta) = \frac{\partial \log f(x_1, \ldots, x_n \mid \theta)}{\partial \theta} \tag{8.10}$$

be the rate of change of the log likelihood as a function of θ. Let

$$t_n'(x_1, \ldots, x_n \mid \theta) = \frac{\partial^2 \log f(x_1, \ldots, x_n \mid \theta)}{\partial \theta^2} \tag{8.11}$$

Let

$$T_n = t(X_1, \ldots, X_n \mid \theta) \tag{8.12}$$

be the random variable obtained by inserting the data into the rate of change of the log likelihood function. Note that

$$T_n = t_1(X_1 \mid \theta) + \cdots + t_1(X_n \mid \theta) \tag{8.13}$$

is a sum of independent random variables. Similarly, let

$$T_n' = t_1'(X_1) + \cdots + t_1'(X_n) \tag{8.14}$$

This gives the following theorem.

Theorem 8.3 *The random variable T_n obtained by inserting the data into the rate of change of the log likelihood function with respect to the parameter is a sum of n independent random variables. It has mean zero*

$$E_\theta[T_n] = 0 \tag{8.15}$$

and variance

$$\operatorname*{Var}_\theta(T_n) = nI(\theta). \tag{8.16}$$

Theorem 8.4 *The random variable T_n' obtained by inserting the data into the second derivative of the log likelihood function with respect to the parameter is a sum of n independent random variables. It has mean satisfying*

$$-E_\theta[T_n'] = nI(\theta). \tag{8.17}$$

8.2 The maximum likelihood estimator

Let $X_1, \ldots, X_n$ be independent random variables, each with the distribution $f(x \mid \theta)$ depending on the parameter θ. Recall that the *likelihood* function

$$f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) \cdots f(x_n \mid \theta) \tag{8.18}$$

is the joint density of these random variables, considered as a function of the parameter θ.

For each fixed $x_1, \ldots, x_n$, let $\hat{\theta}(x_1, \ldots, x_n)$ be the value of θ that maximizes the likelihood function $f(x_1, \ldots, x_n \mid \theta)$. Of course it also maximizes the log likelihood function. Let us be optimistic and assume that this exists and is unique and is assumed at an interior point.

Recall that

$$t_n(x_1, \ldots, x_n \mid \theta) = \frac{\partial \log f(x_1, \ldots, x_n \mid \theta)}{\partial \theta} \tag{8.19}$$

is the rate of change of the log likelihood as a function of θ. Then the function $\hat{\theta}(x_1, \ldots, x_n)$ should satisfy

$$t_n(x_1, \ldots, x_n \mid \hat{\theta}(x_1, \ldots, x_n)) = 0. \tag{8.20}$$

This is because the derivative vanishes at a maximum.

The maximum likelihood estimator is

$$\hat{\Theta} = \hat{\theta}(X_1, \ldots, X_n). \tag{8.21}$$

Example: Consider a normal population with known σ. The log likelihood is $\log f(X_1, \ldots, X_n \mid \mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_i (X_i - \mu)^2/(2\sigma^2)$. Its derivative with respect to μ is $\sum_i (X_i - \mu)/\sigma^2$. If we set this equal to zero and solve for μ, we get the maximum likelihood estimator $\hat{\mu} = \sum_i X_i/n = \bar{X}$.

Example: Consider a normal population with known μ. The log likelihood is $\log f(X_1, \ldots, X_n \mid \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_i (X_i - \mu)^2/(2\sigma^2)$. Its derivative with respect to σ^2 is $-\frac{n}{2}/\sigma^2 + \sum_i (X_i - \mu)^2/(2\sigma^4)$. If we set this equal to zero and solve for σ^2, we get the maximum likelihood estimator $\hat{\sigma}^2 = \sum_i (X_i - \mu)^2/n$.

8.3 Asymptotic behavior of the maximum likelihood estimator

We want to argue that for large n the maximum likelihood estimator is about as good as one can get. So we assume that various technical conditions are satisfied, including that for each data result the maximum likelihood point is unique and is an interior point where the derivative is equal to zero. Then it may be shown that for large sample size the maximum likelihood estimator is close to the true value of the parameter, with high probability.

Expand the function

$$0 = t_n(x_1, \ldots, x_n \mid \hat{\theta}(x_1, \ldots, x_n)) \approx t(x_1, \ldots, x_n \mid \theta) + t'_n(x_1, \ldots, x_n \mid \theta)(\hat{\theta}(x_1, \ldots, x_n) - \theta). \tag{8.22}$$

for $\hat{\theta}(x_1, \ldots, x_n)$ close to θ.

We can solve this for $\hat{\theta}(x_1, \ldots, x_n) - \theta$. We get

$$\hat{\theta}(x_1, \ldots, x_n) - \theta \approx -\frac{t_n(x_1, \ldots, x_n \mid \theta)}{t'_n(x_1, \ldots, x_n \mid \theta)}. \tag{8.23}$$

Let T_n and T'_n be as above. Then we get

$$\hat{\Theta} - \theta \approx -\frac{T_n}{T'_n}. \tag{8.24}$$

It is more illuminating to write this equation as

$$\hat{\Theta} - \theta \approx -\frac{T_n/n}{T'_n/n}. \tag{8.25}$$

The denominator is a sum of independent random variables. Thus for large n we can use the law of large numbers and approximate $-T'_n/n \approx I(\theta)$. This gives

$$\hat{\Theta} - \theta \approx \frac{T_n/n}{I(\theta)}. \tag{8.26}$$

The numerator has mean zero and variance $I(\theta)/n$. So the whole expression has variance given by $I(\theta)/n$ divided by $I(n)^2$. Furthermore, the numerator is a sum of independent random variables, and so for large n is approximately normal. The conclusion is the following theorem.

Theorem 8.5 *Under appropriate technical assumptions, for large n the maximum liklihood estimator $\hat{\Theta}$ that maximizes $f(X_1, \ldots, X_n \mid \theta)$ is approximately normal, approximately unbiased, and has variance approximately equal to the Cramér-Rao lower bound $1/(nI(\theta))$.*

8.4 Asymptotic theory

The result of this chapter is that for large sample size the maximum likelihood estimator is approximately optimal. Does this mean that one should use a maximum likelihood estimator? Certainly not. Just because a method is good in the limit of large n, it does not mean that it is good for the moderate sample size n that one is stuck with.

Furthermore, there may be other estimators that are good in the limit of large n. Typically Baysian estimators have good properties in this limit, and they also have some desirable properties for small n as well. We shall study Baysian estimators in the next chapter. Admittedly, they are neither unbiased

nor do they satisfy the principle of invariance. But there is no fundamental reason to require either of these properties.

So why the theory of maximum likelihood estimators? There are several reasons. First, many of the traditional estimators are maximum likelihood. Second, if in a new situation you cannot think of an estimator, this gives something to try that in many cases is relatively simple to compute. The fact that the principle of invariance is satisfied can be quite convenient in finding a maximum likelihood estimator. But the main reason is a warning: If you are using an estimator that for large sample size has an asymptotic variance that is a large multiple of the asymptotic variance for the maximum likelihood estimator, then maybe you should evaluate what you are doing again.

Example: Consider the Cauchy distribution

$$f(x \mid \theta) = \frac{1}{\pi} \frac{\epsilon}{(x-\theta)^2 + \epsilon^2} \tag{8.27}$$

with unknown median θ and known spread $\epsilon > 0$. The Fisher information is $I(\theta) = 1/(2\epsilon^2)$. So the variance of an unbiased estimator must be bounded below by $2\epsilon^2/n$.

The asymptotic variance of the sample median is $1/(4f(\theta)^2 n) = (\pi^2/4)\epsilon^2/n$. Now $\pi^2/4$ is almost 2.47, so there is some loss in efficiency. Perhaps it would be better to find the maximum likelihood estimator.

This is obtained by differentiating the likelihood function for a sample from a Cauchy population. The resulting equation is

$$\frac{\partial f(x_1, \ldots, x_n \mid \theta)}{\partial \theta} = \frac{2(x_1 - \theta)}{(x_1 - \theta)^2 + \epsilon^2} + \cdots + \frac{2(x_n - \theta)}{(x_n - \theta)^2 + \epsilon^2} = 0. \tag{8.28}$$

This is a nasty equation to solve for θ. One could do it by implementing an iterative process for solving nonlinear equations on a computer. Imagine a θ that is a solution. Then the terms in the equation corresponding to data points x_i that are far from θ are close to zero. The terms in the equation corresponding to data points x_i that are close to θ then each have magnitude about $(x_i - \theta)/\epsilon^2$. So the θ solution is, roughly speaking, a sample mean of part of the data, leaving out the more extreme values. While this estimator is more efficient than the sample median, the sample median begins to look attractive from the point of view of convenience.

8.5 Maximum likelihood as a fundamental principle

If the method of maximum likelihood is to be a fundamental principle, then it should well in all circumstances. It is illuminating, then, to look at its behavior for small samples. This can be rather poor. This can be seen easily for an example with a sample size $n = 1$.

Let us take $f(x)$ to be a fixed probability density with mean zero and variance one. (For example it could be normal.) Let $a > 0$ be a fixed number. Let $f_0(x \mid \theta) = f((x-\theta+a)/\sigma_0 \mid \theta)$ to be a probability density peaked near $x = \theta - a$. The parameter $\sigma_0 > 0$ is taken small. Similarly, $f_1(x \mid \theta) = f((x-\theta-a)/\sigma_1 \mid \theta)$ is another probability density peaked near $x = \theta + a$. The parameter $\sigma_1 > 0$ is taken even smaller. The density $f_1(x \mid \theta)$ is narrower but with a higher peak than that of $f_0(x \mid \theta)$. Let

$$f(x \mid \theta) = (1-p)f_0(x \mid \theta) + pf_1(x \mid \theta). \tag{8.29}$$

Thus the population is a mixture of the 0 population and the 1 population, with probabilities $1-p$ and p.

Take $p > 0$ but very close to zero. Then take $\sigma_0 > 0$ so small that $f_0(x \mid \theta)$ is very narrow and peaked around $\theta - a$. Then take $\sigma_1 > 0$ so tiny that $f_1(x \mid \theta)$ is much more narrow and peaked about $\theta + a$, in fact so that the maximum of $pf_1(x \mid \theta)$ is larger than the maximum of $(1-p)f_0(x \mid \theta)$. Consider a single observation x. The maximum likelihood is assumed when $\theta = x - a$. So $d_-(x) = x - a$ is the maximum likelihood estimator. However we will see that in many circumstances a reasonable person would prefer to use the estimator $d_+(x) = x + a$.

Let us measure the loss by squared error. Then the loss using maximum likelihood is $(d_-(x)-\theta)^2 = (x-a-\theta)^2$. The loss using the competing estimator is $(d_+(x)-\theta)^2 = (x+a-\theta)^2$. The risk is the expected loss. So the risk for the maximum likelihood estimator is

$$r_- = \int (x-a-\theta)^2[(1-p)f_0(x \mid \theta)+pf_1(x \mid \theta)]\,dx \approx 4a^2(1-p)+0^2p = 4a^2(1-p). \tag{8.30}$$

The risk for the competing estimator is

$$r_+ = \int (x+a-\theta)^2[(1-p)f_0(x \mid \theta)+pf_1(x \mid \theta)]\,dx \approx 0^2(1-p)+4a^2p = 4a^2p \tag{8.31}$$

This is a much smaller risk. Certainly the competing estimator is to be prefered. The overwhelming probability is that the observation x is near $\theta - a$, and then $x + a$ is near θ.

Note that this example does not depend critically on the fact that the loss is measured by squared distance. It would also hold, with suitable modifications, if we used a loss $|d(x) - \theta|$ instead of $(d(x) - \theta)^2$.

Remark: The following remark concerns a somewhat unusual circumstance. In this circumstance the use of the maximum likelihood estimator for the above example might seem to be preferable. Take $\epsilon > 0$ to be such an incredibly small number that it is considerably less than the width of the peak of $f_1(x \mid \theta)$. Then let the loss be zero if $|d(x) - \theta| < \epsilon$ and the loss be one if $|d(x) - \theta| \geq \epsilon$. Then the risk for the maximum likelihood estimator satisfies

$$r_- = 1 - \int_{|x-a-\theta|<\epsilon} [(1-p)f_0(x \mid \theta)+pf_1(x \mid \theta)]\,dx \approx 1 - pf_1(\theta+a \mid \theta)\epsilon. \tag{8.32}$$

The risk for the competing estimator is

$$r_+ = 1 - \int_{|x+a-\theta|<\epsilon} [(1-p)f_0(x \mid \theta) + pf_1(x \mid \theta)]\, dx \approx 1 - (1-p)f_0(\theta - a \mid \theta)\epsilon. \tag{8.33}$$

In this case the competing estimator has larger risk. However this is due to a way of measuring loss that rewards only estimates that are extremely precise. The competing estimator is still making estimates in the right general vicinity, but usually gets no credit for it. The overwhelming probability is that the observation x is near $\theta - a$, and then $x + a$ is near θ. But it is not usually near enough to be rewarded. There is a tiny probability that the observation x is extremely close to $x + a$, but these infrequent observations are almost always rewarded. This situation seems rather artificial. With more typical measures of loss in estimation, the competing estimator does much better than maximum likelihood.

8.6 Problems

1. Consider the exponential density $f(x \mid \theta) = \theta e^{-\theta x}$ with parameter $\theta > 0$. Consider an independent sample $X_1, \ldots, X_n$ of size n. a. Find the maximum likelihood estimator of θ. b. Find the maximum likelihood estimator of $E[X_1]$.

2. Consider the geometric density $f(x \mid p) = p(1-p)^x$, where $x = 0, 1, 2, 3, \ldots$. Let $X_1, \ldots, X_n$ be a random sample of size n. a. Find the maximum likelihood estimator of p. b. Consider rolling a die until a six results. The number of times before the six results is a geometric random variable with $p = 1/6$. Perform the experiment with a sample size $n = 20$. Give the estimate of p based on the data.

3. The normal distribution with mean 0 and standard deviation σ is given by the density $f(x \mid \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$, a. Find the maximum likelihood estimator of σ based on an independent sample of size n. b. Find the maximum likelihood estimator of the variance $\theta = \sigma^2$ based on an independent sample of size n.

4. Let $X_1, \ldots, X_n$ be independent random variables, each uniform on the interval from 0 to θ. Find the maximum likelihood estimator of θ. Hint: Do not use calculus; go back to the definition of maximum.

Chapter 9

Bayesian theory

9.1 The Bayesian framework

In Bayesian theory we use a probability distribution on the space of possible parameters. This distribution $\pi(\theta)$ is called the Bayes *prior* distribution of the parameter. Then $f(x_1, \ldots, x_n \mid \theta)$ is regarded as the conditional distribution of the data, given the parameter. The Bayes joint distribution of the data and the parameters is random and given by $f(x_1, \ldots, x_n \mid \theta)\pi(\theta)$. Let

$$f(x_1, \ldots, x_n) = \int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta \tag{9.1}$$

be the Bayes distribution of the data. Finally,

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{f(x_1, \ldots, x_n)} \tag{9.2}$$

is the Bayes *posterior* conditional distribution of the parameter, given the data. The idea is that the data gives more information than is given by the prior distribution. The posterior distribution is the revised distribution when the data values are known.

From where does this prior distribution $\pi(\theta)$ arise? This is the big question about the Bayes theory. However it does no harm to imagine such a distribution and exploring the consequences of using it.

9.2 Baysian estimation for the mean of a normal population

The idea of Bayesian estimation is the following. You assign a prior distribution $\pi(\theta)$. Then you compute the posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$ given the data. The estimator is the mean of this distribution:

$$d(x_1, \ldots, x_n) = E_{x_1,\ldots,x_n}[\Theta] = \int \theta\pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{9.3}$$

Then you take the data $X_1, \ldots, X_n$, and your estimate is $d(X_1, \ldots, X_n)$. The main computational problem is to find the posterior distribution. One common case is that of a normal distribution.

Theorem 9.1 *Suppose that the population distribution is normal with mean θ and variance σ^2. Suppose that the prior distribution $\pi(\theta)$ is normal with mean μ_0 and variance α^2. Then the posterior distribution $\pi(\theta \mid x_1, \ldots, x_n)$ is normal with mean*

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\sigma^2/n}{\sigma^2/n + \alpha^2}\mu_0 + \frac{\alpha^2}{\sigma^2/n + \alpha^2}\bar{x} \tag{9.4}$$

and variance

$$\mathrm{Var}(\Theta \mid x_1, \ldots, x_n) = \frac{1}{\frac{1}{\sigma^2/n} + \frac{1}{\alpha^2}}. \tag{9.5}$$

This proves that if the prior distribution of Θ is normal with mean μ_0 and variance α^2, and if the population is normal with mean θ and variance σ^2, then the Bayes estimator is

$$d(X_1, \ldots, X_n) = \frac{\sigma^2/n}{\sigma^2/n + \alpha^2}\mu_0 + \frac{\alpha^2}{\sigma^2/n + \alpha^2}\bar{X}. \tag{9.6}$$

We see that if α^2 is very large or n is very large, then the Bayes estimator is very close to $\bar{X}$. That is, little prior information or a large sample make one trust the sample mean obtained from the data. On the other hand, if α^2 is small compared with the variance σ^2/n of the sample mean, then the prior information is very useful. One might as well use it and guess something close to μ_0, since the experiment is not doing much to refine your knowledge.

9.3 Probability distributions

There are a number of cases where it is easy to do Bayesian computations. This will be treated in the next section. First we review some basic probability distributions.

The normal (or Gaussian distribution) with parameters μ and σ^2 has density

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{9.7}$$

It has mean μ and variance σ^2.

The Poisson distribution with parameter $\lambda > 0$ is the distribution of natural numbers with discrete density

$$f(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{9.8}$$

for $x = 0, 1, 2, 3, \ldots$. It has mean λ and variance λ.

The Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ is the distribution of positive real numbers with density

$$f(x \mid \alpha, \beta) = \frac{(\beta x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x} \beta \tag{9.9}$$

for $x \geq 0$. It has mean α/β and variance α/β^2. Often $\alpha = n$ is a natural number. When $\alpha = n = 1$ the Gamma distribution is the exponential distribution with density $e^{-\beta x}\beta$. In general the Gamma with $\alpha = n$ is the distribution of the sum of n independent random variables, each with an exponential distribution. The constant in the denominator is then $\Gamma(n) = (n-1)!$. This case is important, since it is the case of sampling from an exponential population.

The Chi-squared distribution is a special case of the Gamma distribution. In fact, the Chi-squared distribution with m degrees of freedom is the Gamma distribution with parameters $\alpha = m/2$ and $\beta = 1/2$. When $\alpha = 1/2$ and $\beta = 1/(2\sigma^2)$ the Gamma distribution is the distribution of the square of a normal random variable with mean zero and variance σ^2. More generally the Gamma distribution with $\alpha = m/2$ and $\beta = 1/(2\sigma^2)$ is the distribution of the sum of m independent random variables, each the square of a mean zero normal.

The binomial distribution with parameters n and p has the discrete density

$$f(x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{9.10}$$

for $x = 0, 1, \ldots, n$. It has mean np and variance $np(1-p)$. When $n = 1$ this is the Bernoulli distribution with density $f(x \mid p) = p^x(1-p)^{1-x}$ with $x = 0, 1$. In general the binomial is the distribution of the sum of n independent Bernoulli random variables. So it is the case of sampling from a population of failures and successes, where one counts the successes with ones.

The negative binomial distribution with parameters n and p has the discrete density

$$f(x \mid p) = \binom{n+x-1}{n-1} p^m (1-p)^x \tag{9.11}$$

for $x = 0, 1, 2, \ldots$. It has mean $n(1/p)(1-p)$ and variance $n(1/p^2)(1-p)$. When $n = 1$ the negative binomial distribution is just the geometric distribution $f(x \mid p) = p(1-p)^x$ for $x = 0, 1, 2, \ldots$. In general the negative binomial is the distribution of the sum of n independent geometric random variables. So it covers sampling from a geometric population.

The Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ has density

$$f(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{9.12}$$

for $0 \leq x \leq 1$. It has mean $\alpha/(\alpha+\beta)$ and variance $1/(\alpha+\beta+1)$ times $\alpha/(\alpha+\beta)$ times $\beta/(\alpha+\beta)$. The normalization constant $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$.

9.4 Prior and posterior distributions

Consider a joint density $f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdots f(x_n \mid \theta)$ with parameter θ. Given a prior density $\pi(\theta)$, it is useful to be able to compute the posterior density

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{f(x_1, \ldots, x_n)}, \tag{9.13}$$

where $f(x_1, \ldots, x_n)$ is the integral of $f(x_1, \ldots, x_n \mid \theta)\pi(\theta)$ over θ. There are simple cases when this calculation is easy. Notice that once we have the posterior distribution, we also immediately have the posterior mean.

Mean of normal Consider the case when $f(x \mid \theta, \sigma^2)$ is the density of a normal distribution with mean θ and variance σ^2. The prior $\pi(\theta)$ is normal with mean μ_0 and variance α^2. Then the posterior density is normal with mean

$$E[\theta \mid x_1, \ldots, x_n] = \frac{\sigma^2/n}{\sigma^2/n + \alpha^2}\mu_0 + \frac{\alpha^2}{\sigma^2/n + \alpha^2}\bar{x}. \tag{9.14}$$

and variance

$$\mathrm{Var}(\theta \mid x_1, \ldots, x_n) = \frac{1}{\frac{1}{\sigma^2/n} + \frac{1}{\alpha^2}}. \tag{9.15}$$

Thus the posterior mean is a weighted mean of the prior mean and the sample mean. Similarly, the posterior variance is the half the harmonic mean of the prior variance and the variance of the sample mean.

Variance of normal Consider the case when $f(x \mid \mu, 1/\theta)$ is the density of a normal distribution with mean μ and variance $1/\theta$. The prior $\pi(\theta)$ is Gamma with parameters α and β. Then the posterior density is Gamma with parameters $\alpha + n/2$ and $\beta + ((x_1 - \mu)^2 + \cdots + (x_n - \mu)^2)/2$. This has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{2\alpha + n}{2\beta + (x_1 - \mu)^2 + \cdots + (x_n - \mu)^2} = \frac{\frac{2}{n} + \frac{1}{\alpha}}{\frac{2}{n}\frac{\beta}{\alpha} + \frac{1}{\alpha}\frac{\sum (x_i - \mu)^2}{n}}. \tag{9.16}$$

This is the Bayes estimator of the reciprocal of the variance of the normal. It is a weighed harmonic average of α/β and the estimator $n/\sum_i (x_i - \mu)^2$ from the sample.

Poisson Consider the case when $f(x \mid \theta)$ is the density of a Poisson distribution with mean θ. The prior $\pi(\theta)$ is Gamma with parameters α and β. Then the posterior density is Gamma with parameters $\alpha + x_1 + \cdots x_n$ and $\beta + n$. This has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\alpha + x_1 + \cdots + x_n}{\beta + n} = \frac{\frac{1}{n}\frac{\alpha}{\beta} + \frac{1}{\beta}\bar{x}}{\frac{1}{n} + \frac{1}{\beta}} \tag{9.17}$$

This is the Bayes estimator of the mean of the Poisson. It is a weighted average of the prior mean $\frac{\alpha}{\beta}$ and the sample mean $\bar{x}$ from the data.

Gamma Consider the case when $f(x \mid \nu, \theta)$ is the density of a Gamma distribution with parameters ν and θ. (This includes the case $\nu = 1$ of an exponential distribution with parameter θ. It also includes the case $\nu = 1/2$ of the distribution of the square of a normal with mean zero and variance $1/(2\theta)$.) The prior $\pi(\theta)$ is Gamma with parameters α and β. Then the posterior density is Gamma with parameters $\alpha + n\nu$ and $\beta + x_1 + \cdots + x_n$. This has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\alpha + n\nu}{\beta + x_1 + \cdots + x_n} = \frac{\frac{1}{n\nu} + \frac{1}{\alpha}}{\frac{1}{n\nu}\frac{\beta}{\alpha} + \frac{1}{\alpha}\frac{\bar{x}}{\nu}}. \tag{9.18}$$

This is the Bayes estimator of θ. It is a weighted harmonic average of the α/β and the estimator $\nu/\bar{x}$ from the sample.

Bernoulli Consider the case when $f(x \mid \theta) = \theta^x(1-\theta)^{1-x}$ is the density of a Bernoulli random variable with mean θ. The prior $\pi(\theta)$ is Beta with parameters α and β. Then the posterior density is Beta with parameters $\alpha + x_1 + \cdots + x_n$ and $\beta + n - (x_1 + \cdots + x_n)$. This has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\alpha + x_1 + \cdots + x_n}{\alpha + \beta + n} = \frac{\frac{1}{n}\frac{\alpha}{\alpha+\beta} + \frac{1}{\alpha+\beta}\bar{x}}{\frac{1}{n} + \frac{1}{\alpha+\beta}}. \tag{9.19}$$

This is the Bayes estimator of the parameter θ. It is a weighted average of the prior probability $\alpha/(\alpha+\beta)$ and the sample proportion $\bar{x}$.

Geometric Consider the case when $f(x \mid \theta) = \theta(1-\theta)^x$ is the density of a geometric distribution with parameter θ. The prior $\pi(\theta)$ is Beta with parameters α and β. Then the posterior density is Beta with parameters $\alpha + n$ and $\beta + x_1 + \cdots + x_n$. This has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\alpha + n}{\alpha + \beta + n + x_1 + \cdots + x_n} = \frac{\frac{1}{n} + \frac{1}{\alpha}}{\frac{1}{n}\frac{\alpha+\beta}{\alpha} + \frac{1}{\alpha}(1+\bar{x})}. \tag{9.20}$$

This is the Bayes estimator of the parameter θ. It is the weighted harmonic average of $\alpha/(\alpha+\beta)$ and the estimator $1/(1+\bar{x})$ based on the sample.

9.5 Bayesian estimation of a population proportion

Let us look at one of these examples in more detail. Consider the case when $f(x \mid \theta) = \theta^x(1-\theta)^{1-x}$ is the density of a Bernoulli random variable with mean θ. Thus the only possible values of x are 1 and 0, and the corresponding probabilities are θ and $1-\theta$. Thus there is a population where the proportion of successes is θ. The problem is to guess θ from the sample.

In the Bayesian analysis the prior $\pi(\theta)$ is Beta with parameters α and β. This says that

$$\pi(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{9.21}$$

Here $\alpha > 0$ and $\beta > 0$. The assumption that this is the distribution of θ is an assumption that one knows probabilistic information about the populations one will encounter. The parameter θ is taken to be a continuous random variable. The mean of this random variable is $\alpha/(\alpha + \beta)$. So this is prior information saying that the proportion is somewhere near this value. The variance of this random variable is $1/(\alpha + \beta + 1)$ times $\alpha/(\alpha + \beta)$ times $\beta/(\alpha + \beta)$. So as the sum $\alpha + \beta$ gets larger, the variance is getting smaller. This is more certain prior information.

The posterior density is gotten by taking the product

$$f(x_1 \mid \theta) \cdots f(x_n \mid \theta)\pi(\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \cdots \theta^{x_n}(1-\theta)^{1-x_n} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{9.22}$$

and normalizing. One can see what this has to be by looking at the exponents of θ and of $1-\theta$. The posterior density is Beta with parameters $\alpha + x_1 + \cdots + x_n$ and $\beta + n - (x_1 + \cdots + x_n)$. The effective sample size due to the prior knowledge and the new data is now $\alpha + \beta + n$.

The posterior density has mean

$$E[\Theta \mid x_1, \ldots, x_n] = \frac{\alpha + x_1 + \cdots + x_n}{\alpha + \beta + n} = \frac{\frac{1}{n}\frac{\alpha}{\alpha+\beta} + \frac{1}{\alpha+\beta}\bar{x}}{\frac{1}{n} + \frac{1}{\alpha+\beta}}. \tag{9.23}$$

This is the Bayes estimator of the parameter θ. It is a weighted average of the prior probability $\alpha/(\alpha + \beta)$ and the sample proportion $\bar{x}$. The weights are $1/n$ and $1/(\alpha + \beta)$. So the value of $\alpha + \beta$ is analogous to a sample size. It gives an idea of the amount of prior information.

What shall we do if we have no particular idea of what the prior $\pi(\theta)$ should be? One idea is to take $\pi(\theta) = 1$, which corresponds to $\alpha = \beta = 1$. In that case, the estimator is a weighed average of the sample proportion $\bar{x}$ and $1/2$. The weights are proportional to $1/2$ and to $1/n$. So for large n this is almost the usual sample proportion. However this is a significant difference for small n. In order to guard against the possibility that bad luck would make the sample frequency close to zero or to one, while the true value of θ was somewhere in the middle, the estimator avoid the end points. This minimizes the expected loss, as measured by squared distance.

It is tempting to try to take a Bayes prior that gives something close to the usual estimator, that is, the sample proportion itself. From the formula, it is clear that this corresponds to taking $\alpha = \beta$ and very close to zero. This give heavy prior weight to the extreme values near zero and one and makes it worth while making estimates near these values.

9.6 Problems

1. For each year the probability of x major earthquakes or a certain Island is given by
$$f(x \mid \theta) = \frac{\theta^x}{x!} e^{-\theta} \tag{9.24}$$
for $x = 0, 1, 2, 3, \ldots$. Here $\theta > 0$ is a parameter. Show that this is a probability density by explicit summation.

2. Find the mean number of earthquakes in a year. Find the variance of the number of earthquakes in a year.

3. Consider a sample of n years, treated as independent. Find the joint density
$$f(x_1, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \cdots f(x_n \mid \theta) \tag{9.25}$$
of the number of earthquakes in each of the years.

4. Find the mean and variance of the sample mean $\bar{x}$. [This is the frequentist result.]

5. A Bayesian statistician believes that the number θ is somewhere near a number θ^*. The statistician thinks that this prior belief is strong enough to be worth a total of β observations. So the statistician takes the prior distribution to be a Gamma distribution with parameters α and β, where $\alpha = \beta\theta^*$. Thus
$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \tag{9.26}$$
for $\theta \geq 0$. Find the mean of this distribution as a function of θ^*. Find its variance as a function of θ^* and β.

6. Calculate explicitly $f(x_1, \ldots, x_n \mid \theta)\pi(\theta)$ as a function of θ and show that it is a multiple of the density of a Gamma distribution with parameters $\beta\theta^* + x_1 + \cdots + x_n$ and $\beta + n$.

7. Find the mean of θ given by the posterior distribution
$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\int_0^\infty f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta} \tag{9.27}$$
as a weighted mean of θ^* and $\bar{x}$ with weights that depend on β and n. Also find the variance. [This is the Bayesian result.]

8. Consider the situation when n is very large. Then the Bayesian wants to say that θ is probably close to the experimental $\bar{x}$. Find an approximate expression for the posterior mean of θ in terms of $\bar{x}$. Find an approximate expression for the posterior variance of θ in terms of $\bar{x}$ and n. [This is the Bayesian result when there is a lot of data and the prior no longer matters.]

9. It is accepted that the number of earthquakes on the Island is a Poisson random variable with a certain mean. The statistician has prior beliefs about this mean based on experience about other islands and on theoretical ideas. These beliefs are represented by taking the mean of the Poisson distribution as being itself described by a Gamma distribution with mean 4.5. Furthermore, this belief is worth about two years of data. The statistician then collects five years of data. The data values are 8, 7, 3, 8, 5. What is the statistician's posterior distribution for the mean number of earthquakes, based on the prior beliefs and the data? Give the numerical values of the parameters.

10. Graph the density of this posterior distribution. [This distribution is the final product of the Baysian analysis.]

Chapter 10

Decision theory and Bayesian theory

10.1 Decision theory

We now want to look more generally about how one judges the performance of a statistical procedure. We consider a population parameter θ which is unknown. For each value of θ there is a probability distribution $f(x_1, \ldots, x_n \mid \theta)$. One is allowed to observe the values of random variables $X_1, \ldots, X_n$. On this basis one wants to make a decision.

A decision function is a function $d(x_1, \ldots, x_n)$ to a space of possible actions. The problem is to choose a suitable function. Then the action taken by the statistician is $d(X_1, \ldots, X_n)$. It is determined by the data $X_1, \ldots, X_n$ and by the decision function employed by the statistician.

Now to see what a good decision is, one must think of the consequences of a bad decision. Let $\mathcal{L}(\theta, a)$ be the *loss* when the parameter has the value θ and decision a is taken. One must think carefully about what this loss function.

The actual loss experienced by the statistician is random, and its value is

$$\mathcal{L}(\theta, d(X_1, \ldots, X_n)). \tag{10.1}$$

Since the statistician does not know θ, the statistician does not know the actual loss.

However the statistician can evaluate the performance of the procedure d in the long run by evaluating the *risk*, or expected loss. This is

$$\mathcal{R}(\theta, d) = E_\theta[\mathcal{L}(\theta, d(X_1, \ldots, X_n))]. \tag{10.2}$$

This can also be written

$$\mathcal{R}(\theta, d) = \int \cdots \int \mathcal{L}(\theta, d(x_1, \ldots, x_n) f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n. \tag{10.3}$$

For a given procedure d, this is a function of θ. While the actual value of θ is unknown, this function is known. The fundamental dogma of decision theory is that one should think hard about this function!

Example: In estimation the action is the value of the parameter that is guessed on the basis of the data. The loss function is a measure of how far the guessed value is from the actual value. It is traditional to take the loss function to be $\mathcal{L}(\theta, a) = (a-\theta)^2$. However this quadratic loss is mainly for mathematical convenience, since it makes the connection with ideas like variance that are easily computed. It might be better to try to get a more realistic idea of the loss function, even if this would complicate the mathematical theory. The decision function employed by the statistician is a function $d(x_1, \ldots, x_n)$ that takes data and uses it to estimate the parameter. The risk function for quadratic loss is of the form

$$\mathcal{R}(\theta, d) = E_\theta[(d(X_1, \ldots, X_n) - \theta)^2]. \tag{10.4}$$

We have seen previously how this may be decomposed into a variance part and a bias part.

Example: In hypothesis testing the actions are only two: guess the null hypothesis, guess the alternative hypothesis. So the loss function amounts to one function $\mathcal{L}(\theta, 0)$ that is the loss from guessing the null hypothesis and $\mathcal{L}(\theta, 1)$ that is the loss from guessing the alternative hypothesis. Say that the null hypothesis is that $\theta = \theta_0$. Then we might take $\mathcal{L}(\theta_0, 0) = 0$, but $\mathcal{L}(\theta, 0) > 0$ for $\theta \neq \theta_0$. On the other hand, we might take $\mathcal{L}(\theta_0, 1) > 0$, but $\mathcal{L}(\theta, 1) < \mathcal{L}(\theta_0, 1)$ for $\theta \neq \theta_0$. The decision function employed by the statistician is a function $d(x_1, \ldots, x_n)$ whose only possible values are 0 and 1. It divides the set of possible data into two complementary regions. The region where the decision function has the value 1 is called the critical region (or rejection region). The risk function is of the form

$$\mathcal{R}(\theta, d) = \mathcal{L}(\theta, 0)P_\theta[d(X_1, \ldots, X_n) = 0)] + \mathcal{L}(\theta, 1)P_\theta[d(X_1, \ldots, X_n) = 1]. \tag{10.5}$$

Notice that

$$P_\theta[d(X_1, \ldots, X_n) = 0)] + P_\theta[d(X_1, \ldots, X_n) = 1] = 1, \tag{10.6}$$

so that one needs to know just one of these functions of θ. The function

$$P_d(\theta) = P_\theta[d(X_1, \ldots, X_n) = 1] \tag{10.7}$$

is called the *power* function of d. The power function can also be written

$$P_d(\theta) = \int \cdots \int_{d(x_1,\ldots,x_n)=1} f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n. \tag{10.8}$$

We can write the risk in term of the power function by

$$\mathcal{R}(\theta, d) = \mathcal{L}(\theta, 0)(1 - P_d(\theta)) + \mathcal{L}(\theta, 1)P_d(\theta). \tag{10.9}$$

10.2 Bayesian decisions

The Bayes decision procedure corresponding to the prior distribution $\pi(\theta)$ is a procedure that minimizes the average prior risk

$$\int \mathcal{R}(\theta, d)\pi(\theta)\, d\theta = \int E_\theta[\mathcal{L}(\theta, d(X_1, \ldots, X_n))]\pi(\theta)\, d\theta. \tag{10.10}$$

In many cases this procedure will be unique.

Theorem 10.1 *A Bayes procedure corresponding to the prior $\pi(\theta)$ is obtained by defining for each $x_1, \ldots, x_n$ the decision $d(x_1, \ldots, x_n)$ that minimizes the average posterior loss*

$$\int \mathcal{L}(\theta, d(x_1, \ldots, x_n))\pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{10.11}$$

Note that one would get the same result by minimizing

$$\int \mathcal{L}(\theta, d(x_1, \ldots, x_n)) f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta \tag{10.12}$$

with fixed data values $x_1, \ldots, x_n$. Sometimes this form is convenient.

Proof: The average risk is

$$\int \mathcal{R}(\theta, d)\, \pi(\theta)\, d\theta. = \int \int \mathcal{L}(\theta, d(x_1, \ldots, x_n)) f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n\, \pi(\theta)\, d\theta. \tag{10.13}$$

This can also be written

$$\int \mathcal{R}(\theta, d)\, \pi(\theta)\, d\theta = \int \int \mathcal{L}(\theta, d(x_1, \ldots, x_n))\pi(\theta \mid x_1, \ldots, x_n)\, d\theta f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n. \tag{10.14}$$

To make the integral over the data as small as possible, one makes the integrand as small as possible at each data point.

Example: Consider estimation with quadratic loss. The Bayesian estimator is then the estimator $d(x_1, \ldots, x_n)$ that minimizes the average posterior loss

$$\int (d(x_1, \ldots, x_n) - \theta)^2 \pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{10.15}$$

However this is the mean with respect to the posterior distribution:

$$d(x_1, \ldots, x_n) = \int \theta\pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{10.16}$$

So to perform the estimate, first calculate the posterior distribution given the data. Then calculate the mean of this distribution.

Example: Let us consider the case of hypothesis testing, where the only decisions are between a null hypothesis and an alternative hypothesis. The decision

function $d(x_1, \ldots, x_n)$ has values 0 or 1. A Bayes procedure corresponding to the prior $\pi(\theta)$ is obtained by defining for each $x_1, \ldots, x_n$ the decision $d(x_1, \ldots, x_n)$ that minimizes the average posterior loss. Therefore, given the data, one computes

$$\int \mathcal{L}(\theta, 0)) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta. \tag{10.17}$$

$$\int \mathcal{L}(\theta, 1) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta. \tag{10.18}$$

and picks the smaller of the two. This defines the corresponding decision. Thus the critical region is where the second one is smaller. It is defined by an inequality

$$\int (\mathcal{L}(\theta, 0)) - \mathcal{L}(\theta, 1)) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta \geq 0. \tag{10.19}$$

10.3 Bayesian decisions and risk

Let $\mathcal{L}(\theta, a)$ be the loss from taking action a when the state of nature is θ. Let $X_1, \ldots, X_n$ be random variables whose joint distribution $f(x_1, \ldots, x_n \mid \theta)$ depends on θ. Let $d(x_1, \ldots, x_n)$ be a decision function. The action of the statistician using this decision function is then random; it is $d(X_1, \ldots, X_n)$. The loss the statistician incurs is also random; it is $\mathcal{L}(\theta, d(X_1, \ldots, X_n))$. The decision theory risk of this decision function is the expected loss given the parameter value

$$\mathcal{R}(\theta, d) = E_\theta[\mathcal{L}(\theta, d(X_1, \ldots, X_n))] = \int \cdots \int \mathcal{L}(\theta, d(x_1, \ldots, x_n)) f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n.$$

This is the risk of d as a function of θ.

We say that a decision function d cannot be improved upon if for every other decision function d' with $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ for all θ we have as a consequence $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$ for all θ. If d cannot be improved on, it does not necessarily mean that d is "good" in any sense. But it does mean that any attempt to improve on d that makes smaller risk for some parameter values will make larger risk for some other parameter values. [In the literature of decision theory a decision function that cannot be improved on is called "admissible".]

Let $\pi(\theta)$ be a Bayesian prior distribution on the parameters. Then

$$r(\pi, d) = \int E_\theta[\mathcal{L}(\theta, d(X_1, \ldots, X_n))]\, \pi(\theta)\, d\theta$$

is the Bayes risk of d with respect to π. The decision function d is Bayes with respect to π if it minimizes the Bayes risk with respect to π. This says that for every other decision function d'' we have $r(\pi, d) \leq r(\pi, d'')$.

Suppose that the decision function d is Bayes with respect to π. Suppose that for every decision function d' that is Bayes with respect to π and such that $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ for all θ it follows that $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$ for all θ. Then we say that d cannot be improved on by a Bayes decision function with the same prior.

Theorem 10.2 *Suppose that for some π the decision function d is Bayes with respect to the prior π. Suppose also that d cannot be improved on by a Bayes decision function with the same prior. Then the decision function d cannot be improved on.*

Proof: Suppose that d is Bayes with respect to π. Suppose that d' is some other decision function (not assumed to be Bayes!) with $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ for all θ. Then it follows by integration that

$$r(\pi, d') \leq r(\pi, d).$$

Since d is Bayes with respect to π, it follows that for every decision function d''

$$r(\pi, d) \leq r(\pi, d'').$$

It follows from the last two inequalities that

$$r(\pi, d') \leq r(\pi, d'').$$

Since d'' is arbitrary, it follows that d' is Bayes with respect to π. The hypothesis says that d cannot be improved on by a Bayes decision function with the same prior. Thus $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$ for all θ. This shows that d cannot be improved on.

Note 1. It is possible that for the given π there is only one Bayes decision function. Then the condition that d cannot be improved on by a Bayes decision function with the same prior is evident.

Note 2. Also, if the prior assigns strictly positive weight to each parameter value, so that $\pi(\theta) > 0$ for each θ, and if the risk functions are continuous in θ, then the Bayes decision function d with respect to π cannot be improved on by a Bayes decision function with the same prior. The reason is the following. Suppose that d' is another Bayes decision function with the same prior. Then $r(\pi, d') = r(\pi, d)$. Suppose also that $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ for all θ. If $\mathcal{R}(\theta, d') < \mathcal{R}(\theta, d)$ for some θ, then we could integrate $\mathcal{R}(\theta, d')\pi(\theta) < \mathcal{R}(\theta, d)\pi(\theta)$ for these θ and get $r(\pi, d') < r(\pi, d)$. This would be a contradiction. So $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$ for all θ.

The conclusion of the above discussion is that typically a a Bayes decision rule is not completely foolish. This is because the rule works very well against one possible situation: that in which nature has given us parameter values that obey the prior distribution $\pi(\theta)$. Of course a particular prior distribution may be somewhat stupid, but if one admits that it can occur, then one cannot do better in that situation.

In practice, one might want to try to think of a $\pi(\theta)$ that seems reasonable and find the corresponding Bayes decision function $d(x_1, \ldots, x_n)$. The one could look at the corresponding risk function $\mathcal{R}(\theta, d)$. If this looks satisfactory, then one might use this procedure, and hope for the best. At least one knows the risks, and one is not foolishly overlooking something better.

10.4 Problems

1. Consider a normal distribution $f(x \mid \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ with known standard deviation $\sigma > 0$ but unknown mean μ. Suppose a random value X is to be observed. Calculate the risk function as a function of μ for estimators of the form cX. Find a minimax estimator in this class of estimators. Hint: The task is to compute $R(\mu, c) = E_\mu[(cX - \mu)^2]$ where X is normal with mean μ and standard deviation σ. The minimax estimator is obtained by calculating for each fixed c the maximum value of $R(\mu, c)$. Then try to take c to make this small.

2. Suppose in the previous problem that μ itself is uniformly distributed on the interval from $-a$ to a. Calculate the mean risk for the estimators cX. Find the Bayes estimator. Hint: Compute $r(a, c) = E[R(\mu, c)]$ where the $R(\mu, c)$ is taken from problem 1 and where the expectation is an integral over μ. The density of the uniform is $1/(2a)$ in the interval from $-a$ to a. The Bayes estimator is obtained by taking c to minimize $r(a, c)$. If a is much larger than σ, then most of the information will come from the experiment. So you should get c close to one. On the other hand, if a is much smaller than σ, then your prior information is very informative, and you should more or less ignore the experimental data. So you should get c close to zero.

3. Say that $f(x \mid \theta) = 1/\theta$ is the uniform density on the interval $0 \leq x \leq \theta$. Say that θ itself is distributed according to a prior density $\pi(\theta) = \beta^2 \theta e^{-\beta\theta}$ for $\theta > 0$. a. Calculate the posterior density $f(\theta \mid x)$. b. Calculate the Bayes squared error estimator $E[\theta \mid x]$.

4. Consider a random sample of size n from a Poisson distribution with mean μ. Suppose we are using squared error as our loss function. Assume that the prior is $\pi(\mu) = \beta e^{-\beta\mu}$ for $\mu > 0$. Calculate the Bayes estimator of the unknown μ.

Chapter 11

Testing hypotheses

11.1 Null and alternative hypothesis

In hypothesis testing the actions are only two: guess the null hypothesis, guess the alternative hypothesis. So the loss function amounts to one function $\mathcal{L}(\theta, 0)$ that is the loss from guessing the null hypothesis and $\mathcal{L}(\theta, 1)$ that is the loss from guessing the alternative hypothesis. Say that the null hypothesis is that $\theta = \theta_0$. Then we might take $\mathcal{L}(\theta_0, 0) = 0$, but $\mathcal{L}(\theta, 0) > 0$ for $\theta \neq \theta_0$. On the other hand, we might take $\mathcal{L}(\theta_0, 1) > 0$, but $\mathcal{L}(\theta, 1) < \mathcal{L}(\theta_0, 1)$ for $\theta \neq \theta_0$.

Clearly there is a region of parameters θ where $\mathcal{L}(\theta, 0) < \mathcal{L}(\theta, 1)$ where it is better to guess the null hypothesis, and there is a region where $\mathcal{L}(\theta, 1) < \mathcal{L}(\theta, 0)$ where it is better to guess the alternative hypothesis.

The decision function employed by the statistician is a function $d(x_1, \ldots, x_n)$ whose only possible values are 0 and 1. It divides the set of possible data into two regions. The value where the decision function has the value 1 is called the critical region. The risk function is of the form

$$\mathcal{R}(\theta, d) = \mathcal{L}(\theta, 0) P_\theta[d(X_1, \ldots, X_n) = 0)] + \mathcal{L}(\theta, 1) P_\theta[d(X_1, \ldots, X_n) = 1]. \tag{11.1}$$

Notice that

$$P_\theta[d(X_1, \ldots, X_n) = 0] + P_\theta[d(X_1, \ldots, X_n) = 1] = 1, \tag{11.2}$$

so that one needs to know just one of these functions of θ. The function

$$P_d(\theta) = P_\theta[d(X_1, \ldots, X_n) = 1] \tag{11.3}$$

is called the *power* function of d. The power function can also be written

$$P_d(\theta) = \int \cdots \int_{d(x_1, \ldots, x_n)=1} f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n. \tag{11.4}$$

We can write the risk in term of the power function by

$$\mathcal{R}(\theta, d) = \mathcal{L}(\theta, 0)(1 - P_d(\theta)) + \mathcal{L}(\theta, 1) P_d(\theta). \tag{11.5}$$

Or, we can write it in the form

$$\mathcal{R}(\theta, d) = \mathcal{L}(\theta, 0) + (\mathcal{L}(\theta, 1) - \mathcal{L}(\theta, 0))P_d(\theta). \tag{11.6}$$

The power is the probability of guessing the alternative hypothesis. We see that if we want a test with small risk, then we want to make the power small in the region of θ near θ_0 where $\mathcal{L}(\theta, 1) - \mathcal{L}(\theta, 0) > 0$, that is, where it is better to guess the null hypothesis. On the other hand, we want to make the power large in the complementary region far from θ_0 where $\mathcal{L}(\theta, 1) - \mathcal{L}(\theta, 0) < 0$, that is, where it is better to guess the alternative hypothesis.

Even with a good test it is possible to have bad luck and guess wrong. If θ is near θ_0 and we guess the alternative hypothesis, then we have made a *type I error*. The probability of a type I error is denoted by $\alpha(\theta)$. Thus $\alpha(\theta) = P_d(\theta)$ when θ is near θ_0. Sometimes this is called the significance level or size of the test. If θ is far from θ_0 and we guess the null hypothesis, then we have made a *type II error*. The probability of a type II error is denoted by $\beta(\theta)$. Thus $\beta(\theta) = 1 - P_d(\theta)$ when θ is far from θ_0.

11.2 Simple null and alternative hypotheses

One particularly simple situation is when the null hypothesis is $\theta = \theta_0$ and the alternative hypothesis is $\theta = \theta_1$. Then $\alpha = P_d(\theta_0)$ and $\beta = 1 - P_d(\theta_1)$ are numbers. The risk function is given by $R(\theta_0, d) = \mathcal{L}(\theta_0, 0) + (\mathcal{L}(\theta_0, 1) - \mathcal{L}(\theta_0, 0))\alpha$ and by $R(\theta_1, d) = \mathcal{L}(\theta_1, 1) + (\mathcal{L}(\theta_1, 0) - \mathcal{L}(\theta_1, 1))\beta$. The coefficients in front of α and β are assumed to be each positive. So the real game is to try to make both α and β small.

This is of course impossible. However there is a systematic way of making tests that cannot be improved on. This is the content of the following Neyman-Pearson lemma.

Theorem 11.1 *Let $k > 0$ be a fixed constant. Let $d(x_1, \ldots, x_n)$ be such that*

$$f(x_1, \ldots, x_n \mid \theta_1) \geq k f(x_1, \ldots, x_n \mid \theta_0) \tag{11.7}$$

when $d(x_1, \ldots, x_n) = 1$ and

$$f(x_1, \ldots, x_n \mid \theta_1) \leq k f(x_1, \ldots, x_n \mid \theta_0) \tag{11.8}$$

when $d(x_1, \ldots, x_n) = 0$. Let $\alpha = P_{\theta_0}[d(X_1, \ldots, X_n) = 1]$ be the probability of a type I error, and let $\beta = P_{\theta_1}[d(X_1, \ldots, X_n) = 0]$ be the probability of a type II error. Then no other choice of d can decrease $k\alpha + \beta$. In particular, if d' is another procedure with $\alpha' \leq \alpha$, then $\beta \leq \beta'$.

Proof: The indicated choice of d makes

$$1_{d(x_1,\ldots,x_n)=0} f(x_1, \ldots, x_n \mid \theta_1) + 1_{d(x_1,\ldots,x_n)=1} k f(x_1, \ldots, x_n \mid \theta_0) \tag{11.9}$$

minimal. Therefore it makes the integral of this quantity minimal. However the integral is $\beta + k\alpha$.

Example: Consider a normal population with known σ. The null hypothesis is $\mu = \mu_0$ and the alternative hypothesis is $\mu = \mu_1 > \mu_0$. The condition on the likelihood ration is that

$$\exp(-\frac{\sum_{i=1}^n (x_i - \mu_1)^2}{2\sigma^2}) \geq k \exp(-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}). \tag{11.10}$$

This is equivalent to a condition $\bar{x} \geq a$, where a depends on k. So the critical region is where the sample mean is large.

Example: Consider a normal population with known μ. The null hypothesis is $\sigma = \sigma_0$ and the alternative hypothesis is $\sigma = \sigma_1 < \sigma_0$. The condition on the likelihood ration is that

$$\frac{1}{\sigma_1^n} \exp(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma_1^2}) \geq k \frac{1}{\sigma_0^n} \exp(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma_0^2}). \tag{11.11}$$

This is equivalent to a condition

$$\frac{\sum_{i=1}^n (x_i - \mu)^2}{n} \leq b, \tag{11.12}$$

where b depends on k. So the critical region is where this estimate of variance is small.

11.3 Minimax risk

The Neyman-Pearson lemma gives tests that cannot be bettered, but it does not tell us which one to use. This depends on the value of the parameter k. To clarify this issue, one must look at the risk. The risk of the test under the two hypotheses is

$$\mathcal{R}(\theta_0, d) = \mathcal{L}(\theta_0, 0)(1 - \alpha) + \mathcal{L}(\theta_0, 1)\alpha \tag{11.13}$$

for the null hypothesis and

$$\mathcal{R}(\theta_1, d) = \mathcal{L}(\theta_1, 0)\beta + \mathcal{L}(\theta_1, 1)(1 - \beta) \tag{11.14}$$

for the alternative hypothesis.

If we are pessimistic, then we may want to look at the maximum risk and try to minimize it. This is the *minimax* criterion for making a statistical decision.

Theorem 11.2 *Assume that $\mathcal{L}(\theta_0, 0) < \mathcal{L}(\theta_0, 1)$ and $\mathcal{L}(\theta_1, 0) > \mathcal{L}(\theta_1, 1)$. Consider the test described in the Neyman-Pearson lemma. Suppose the risks for the two possible hypotheses are equal: $\mathcal{R}(\theta_1, d) = \mathcal{R}(\theta_0, d)$. Then this test minimizes the maximum risk.*

Example: Say that one takes the situation where $0 = \mathcal{L}(\theta_0, 0) < \mathcal{L}(\theta_1, 0)$ and $0 = \mathcal{L}(\theta_1, 1) < \mathcal{L}(\theta_0, 1)$ Thus the loss from a correct decision is zero. Then the condition that the risks are equal is that

$$\mathcal{L}(\theta_0, 1)\alpha = \mathcal{L}(\theta_1, 0)\beta. \tag{11.15}$$

Say that $\mathcal{L}(\theta_0, 1)$ makes us look silly and gives a loss of 5, while $\mathcal{L}(\theta_1, 0)$ overlooks an important discovery and gives a loss of 10. Then we should choose $\alpha = 2\beta$. Now we may try various values of α and compute the corresponding value of β. Maybe when $\alpha = 0.3$ we have a $\beta = 0.15$. Then the risks are 1.5 versus 1.5. On the other hand, if $\alpha = 0.05$, then the corresponding $\beta = 0.50$. With this choice the risks are 0.5 and 2.5. Then clearly the traditional choice of $\alpha = 0.05$ is exposing one to a rather unpleasant risk of a type II error. A statistician wishing to guard against the worst that nature can provide would be better off using the level $\alpha = 0.3$.

11.4 One-sided tests

Now consider the more complicated situation with composite null hypothesis and alternative hypotheses. For practice we look at the case of a sample from a normal population with known variance σ^2 and unknown mean θ. Let us look at a somewhat artificial situation where the loss function for fixed decision has two values. The loss function $\mathcal{L}(\theta, 1)$ is equal to $\mathcal{L}(I) > 0$ for $\theta \leq \theta_0$, and zero elsewhere. The loss function $\mathcal{L}(\theta, 0)$ is equal to $\mathcal{L}(II) > 0$ for $\theta > \theta_0$ and zero elsewhere. Then the risk of the test is $\mathcal{L}(I)P_d(\theta)$ for $\theta \leq \theta_0$ and $\mathcal{L}(II)(1-P_d(\theta))$ for $\theta > \theta_0$. Say that we take the decision $d(x_1, \ldots, x_n) = 1$ when $\bar{x} > a$. Then $P_d(\theta) = P_\theta[\bar{X} > a]$. This is increasing with θ. So the greatest possible risk is either $\mathcal{L}(1)P_d(\theta_0)$ or $\mathcal{L}(II)(1-P_d(\theta_0))$. We can minimize this by taking choosing a to make these equal. This give

$$P_{\theta_0}[\bar{X} > a] = \frac{\mathcal{L}(II)}{\mathcal{L}(I) + \mathcal{L}(II)}. \tag{11.16}$$

Thus with this minimax procedure the size of the critical region is determined by the losses. If a type I error is 9 times as embarrassing as a type II error, then the level of the test should be taken to be 1/10. On the other hand, if one error looks as bad as the other, then the size of the test should be 1/2, which amounts to taking $a = \theta_0$ as the cutoff point, independent of the sample size. Minimax is a conservative policy. Consider a pessimist with no prior information about the parameter, but who must nevertheless make an important practical decision on the basis of the data. This individual should certainly consider minimax procedures.

11.5 Bayes tests for simple hypotheses

A Bayes procedure corresponding to the prior $\pi(\theta)$ is obtained by defining for each x the decision $d(x_1, \ldots, x_n)$ that minimizes the average posterior loss,

which is

$$\int \mathcal{L}(\theta, d(x_1, \ldots, x_n)) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta \tag{11.17}$$

divided by $f(x_1, \ldots, x_n)$. Therefore, given the data, one computes

$$\int \mathcal{L}(\theta, 0)) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta. \tag{11.18}$$

and

$$\int \mathcal{L}(\theta, 1) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta \tag{11.19}$$

and picks the smaller of the two. This defines the corresponding decision. Thus the critical region is where the second one is smaller. It is defined by an inequality

$$\int \mathcal{L}(\theta, 1) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta \leq \int \mathcal{L}(\theta, 0) f(x_1, \ldots, x_n \mid \theta) \pi(\theta)\, d\theta. \tag{11.20}$$

If we think of Θ as a random variable with the posterior distribution, this criterion says that $d(x_1, \ldots, x_n) = 1$ when

$$E_{x_1, \ldots, x_n}[\mathcal{L}(\Theta, 1)] \leq E_{x_1, \ldots, x_n}[\mathcal{L}(\Theta, 0)]. \tag{11.21}$$

We can look at all this in the special case of a simple hypothesis and simple alternative. In this case the prior probabilities $\pi(\theta_0)$ and $\pi(\theta_1)$ are two numbers that add to one. The posterior probabilities are then

$$\pi(\theta_0 \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0)}{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)} \tag{11.22}$$

and

$$\pi(\theta_1 \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)}{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)} \tag{11.23}$$

For the purpose of the hypothesis test one can ignore the denominator and compute

$$\mathcal{L}(\theta_0, 0) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + \mathcal{L}(\theta_1, 0) f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1) \tag{11.24}$$

and

$$\mathcal{L}(\theta_0, 1) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + \mathcal{L}(\theta_1, 1) f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1) \tag{11.25}$$

and the critical region is where the second one of these is smaller. This is equivalent to saying that

$$(\mathcal{L}(\theta_1, 0) - \mathcal{L}(\theta_1, 1)) f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1) \geq (\mathcal{L}(\theta_0, 1)) - \mathcal{L}(\theta_0, 0)) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0). \tag{11.26}$$

This is the same as the result given by the Neyman-Pearson lemma, in the case where k is the ratio of $(\mathcal{L}(\theta_0, 1)) - \mathcal{L}(\theta_0, 0))\pi(\theta_0)$ to $(\mathcal{L}(\theta_1, 0) - \mathcal{L}(\theta_1, 1))\pi(\theta_1)$. So the value of k is determined by the prior probabilities of the two hypotheses.

The Bayes risk of this decision is the risk averaged with the prior probabilities. This works out to be

$$r(\pi, d) = (\mathcal{L}(\theta_0, 0)(1 - \alpha) + \mathcal{L}(\theta_0, 1)\alpha)\pi(\theta_0) + (\mathcal{L}(\theta_1, 0)\beta + \mathcal{L}(\theta_1, 1)(1 - \beta))\pi(\theta_1) \tag{11.27}$$

Example: Say that there is no loss from a correct decision. However $\mathcal{L}(\theta_0, 1)$ makes us look silly and gives a loss of 5, while $\mathcal{L}(\theta_1, 0)$ overlooks an important discovery and gives a loss of 10. Now say that we think that the prior probability of the important discovery is quite low, say $\pi(1) = 0.1$. Then we should use a k equal to 5(0.9) divided by 10(0.1), that is, $k = 4.5$. This makes it not so likely that we will announce the alternative hypothesis, unless the evidence is quite strong. We remain more worried about looking silly, since we do not really much believe that there is a discovery there to be made in the first place. Only overwhelming evidence will convince us.

11.6 One-sided Bayes tests

For practice we look at the case of a sample from a normal population with known variance σ^2 and unknown mean θ. In keeping with the Bayes philosophy, let us assume that θ itself is normally distributed with mean μ_0 and variance α^2. Let us look at a somewhat artificial situation where the loss function for fixed decision has two values. The loss function $\mathcal{L}(\theta, 1)$ is equal to $\mathcal{L}(I) > 0$ for $\theta \leq \theta_0$, and zero elsewhere. The loss function $\mathcal{L}(\theta, 0)$ is equal to $\mathcal{L}(II) > 0$ for $\theta > \theta_0$ and zero elsewhere. Then the posterior risk of the test when $d(x_1, \ldots, x_n) = 1$ is

$$\mathcal{L}(I) \int_{-\infty}^{\theta_0} \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = \mathcal{L}(I) P_{x_1,\ldots,x_n}[\Theta \leq \theta_0], \tag{11.28}$$

where Θ is normal with the posterior mean $(\sigma^2/n\mu_0 + \alpha^2\bar{x})/(\sigma^2/n + \alpha^2)$ and with variance given by the reciprocal of the sum of the reciprocals of σ^2/n and α^2. Similarly, the posterior risk of the test when $d(x_1, \ldots, x_n) = 0$ is

$$\mathcal{L}(II) \int_{\theta_0}^{\infty} \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = \mathcal{L}(II) P_{x_1,\ldots,x_n}[\Theta > \theta_0]. \tag{11.29}$$

Therefore the critical region consists of all values of $x_1, \ldots, x_n$ such that

$$\mathcal{L}(I) P_{x_1,\ldots,x_n}[\Theta \leq \theta_0] \leq \mathcal{L}(II) P_{x_1,\ldots,x_n}[\Theta > \theta_0]. \tag{11.30}$$

This is the same as requiring that

$$P_{x_1,\ldots,x_n}[\Theta \leq \theta_0] \leq \frac{\mathcal{L}(II)}{\mathcal{L}(I) + \mathcal{L}(II)}. \tag{11.31}$$

For simplicity, take the case when α is very large. Then we can think of Θ as being normally distributed with mean $\bar{x}$ and variance σ^2/n. Then we can write the condition in the form

$$P[\bar{x} \leq \theta_0 - (\Theta - \bar{x})] \leq \frac{\mathcal{L}(II)}{\mathcal{L}(I) + \mathcal{L}(II)}. \tag{11.32}$$

Now take a such that

$$P[a \leq \theta_0 - (\Theta - \bar{x})] = \frac{\mathcal{L}(II)}{\mathcal{L}(I) + \mathcal{L}(II)}. \tag{11.33}$$

Since the random variable $\theta_0 - (\Theta - \bar{x})$ has mean θ_0 and variance σ^2/n, this is the critical a that was used before in the non-Bayesian theory. Thus to get a smaller probability we must have $a \leq \bar{x}$. So in the limit $\alpha \to \infty$ the Bayesian test is the minimax test. This is of course a special feature of the one-sided testing situation. But it shows that there can be very different ways of thinking about the same problem that lead to the same solution.

11.7 p values

Say that we have a one-sided hypothesis test situation, so that the decision is either for the null hypothesis $\theta \leq \theta_0$ or for the alternative hypothesis $\theta > \theta_0$. Then it is conventional to call the size (or level) α of a test to be the probability that the decision is for the alternative when $\theta = \theta_0$.

Let $X_1, \ldots, X_n$ be the data values from the sampling experiment. Say that there is a test statistic $t(x_1, \ldots, x_n)$ such that for each α there is a value a_α such that

$$P_{\theta_0}[a_\alpha < t(X_1, \ldots, X_n)] = \alpha. \tag{11.34}$$

Then we use $a_\alpha < t(x_1, \ldots, x_n)$ as the critical region in a test of level α.

Now let $X'_1, \ldots, X'_n$ be the random data values from an independent repetition of the sampling experiment. The p value function is defined by saying that

$$p(x_1, \ldots, x_n) = P'_{\theta_0}[t(x_1, \ldots, x_n) < t(X'_1, \ldots, X'_n)]. \tag{11.35}$$

Thus this is the probability of getting a larger value on another run.

Now we can plug the experimental values into the p-value function. This give the random variable

$$p(X_1, \ldots, X_n) = P'_{\theta_0}[t(X_1, \ldots, X_n) < t(X'_1, \ldots, X'_n)]. \tag{11.36}$$

From this it is easy to see that $p(X_1, \ldots, X_n)$ is a uniform random variable. In fact, $p(X_1, \ldots, X_n) \leq \alpha$ is equivalent to $a_\alpha \leq t(X_1, \ldots, X_n)$, which has probability α.

Sometimes this p value random variable is used to give an idea of how much one should believe the alternative hypothesis. This is contrary to the philosophy of hypothesis testing, where the idea is that the only question is a stark choice between two actions. However, in very special circumstances it is at least somewhat reasonable as the solution to an estimation problem.

Theorem 11.3 *Let θ range over the real line. Let $\theta \leq \theta_0$ be the parameter values corresponding to the null hypothesis, and let $\theta > \theta_0$ be the parameter values corresponding to the alternative hypothesis. Consider a normal distribution with mean θ and known variance σ^2. Take a sample of size n. Let $h_{\theta_0}(\theta)$ be 1 if $\theta \leq \theta_0$ and 0 if $\theta > \theta_0$. Consider the quadratic loss function. Let the parameter θ be itself normal with mean μ_0 and variance α^2. Then in the limit $\alpha^2 \to \infty$ the corresponding Bayes estimator of $h_{\theta_0}(\theta)$ is the p value function*

$$p(x_1, \ldots, x_n) = P_{\theta_0}[\bar{x} < \bar{X}']. \tag{11.37}$$

Proof: The Bayes estimator is the expectation

$$d(x_1, \ldots, x_n) = E_{x_1,\ldots,x_n}[h_{\theta_0}(\Theta)] = P_{x_1,\ldots,x_n}[\Theta \leq \theta_0]. \tag{11.38}$$

In the limit $\alpha \to \infty$ the random variable Θ has mean $\bar{x}$ and variance σ^2/n. Hence the random variable $\theta_0 - (\Theta - \bar{x})$ has mean θ_0 and variance σ^2/n. So

$$d(x_1, \ldots, x_n) = P[\bar{x} \leq \theta_0 - (\Theta - \bar{x})]. \tag{11.39}$$

However the right hand side is the same as the p value function in the statement of the theorem.

Of course this theorem does not say that the p value is the right thing to use. Once one has decided that one has an estimation problem, then one has to decide exactly what one wants to estimate. In the above theorem it was the function that is 1 for $\theta \leq \theta_0$ and 0 for $\theta > \theta_0$. But it could be some other function of the parameter. And the loss might not be quadratic.

11.8 Two-sided Bayes tests

As another exercise, let us look at hypothesis testing problems where the loss functions are quadratic. Again this is mostly for convenience. In fact, a legitimate criticism of the decision theory approach to statistics is that there are few cases where one is really sure what the loss function should be.

In any case, let us take the case of a normal distribution with unknown mean θ and known variance σ^2. The loss function is considered to be given by $\mathcal{L}(\theta, 0) = a(\theta - \theta_0)^2$ and $\mathcal{L}(\theta, 1) = c - b(\theta - \theta_0)^2$. Thus one would like to guess the alternative hypothesis if $c - b(\theta - \theta_0)^2 \leq a(\theta - \theta_0)^2$. This is equivalent to the condition that $c/(a + b) \leq (\theta - \theta_0)^2$. If this condition is not satisfied, then the effect is regarded as being so weak that it has no practical importance.

In keeping with the Bayes philosophy, we consider the prior distribution of θ to be normal with mean θ_0 and variance α^2. After the data is taken, the posterior distribution of θ is modified. The test is defined in terms of this posterior distribution. Then the rejection region is where

$$E_{x_1,\ldots,x_n}[c - b(\Theta - \theta_0)^2] \leq E_{x_1,\ldots,x_n}[a(\Theta - \theta_0)^2]. \tag{11.40}$$

This says that

$$\frac{c}{a+b} \leq E_{x_1,\ldots,x_n}[(\Theta - \theta_0)^2] = \operatorname*{Var}_{x_1,\ldots,x_n}(\Theta) + (E_{x_1,\ldots,x_n}[\Theta] - \theta_0)^2. \tag{11.41}$$

The variance on the right hand side is a constant. On the other hand, the difference

$$E_{x_1,\ldots,x_n}[\Theta] - \theta_0 = \frac{\alpha^2}{\sigma^2/n + \alpha^2}(\bar{x} - \theta_0). \tag{11.42}$$

So this test is based on the absolute value of the difference $\bar{x} - \theta_0$ between the sample mean and the θ_0 of the null hypothesis. If one knows the loss functions and the value of α^2, then one knows the proper cutoff for the test. The formula for the rejection region is given explicitly by

$$\frac{c}{a+b} \leq \frac{\alpha^2}{\sigma^2/n + \alpha^2}\left(\frac{\sigma^2}{n} + \frac{\alpha^2}{\sigma^2/n + \alpha^2}(\bar{x} - \theta_0)^2\right). \tag{11.43}$$

Notice how the cutoff depends on the Bayes prior. If α is very large, then it is regarded as a priori probable that the true value of θ is very far from θ_0. In that case, for small sample size the tendency is to guess the alternative hypothesis in any case. For large sample size the test says to guess the alternative hypothesis unless the sample mean is well within the region near θ_0 where one would prefer to guess the null hypothesis. On the other hand, if α is very small, then one gives large a priori probability to the null hypothesis. Then one does not guess the alternative hypothesis until the sample mean is very far from θ_0.

Perhaps this form of the prior, with just one peak at θ_0, is not what is desired. In this case one has to gear up for more complicated mathematics. However the general plan of the computation is always the same.

11.9 Lessons for hypothesis testing

The philosophy of hypothesis testing is that there is a decision to be made between two actions. There is no other choice. When we start doing calculations that seem to indicate a conflict with this goal, then perhaps we should be considering a multiple decision problem or even an estimation problem. Thus if someone asks a statistician to decide whether a treatment has little effect, great effect, or whether to continue with further study, then this is a request for a decision among three actions. Similarly, if someone asks for a statistic that gives the amount of evidence in favor of the effectiveness of a treatment, then this may be an estimation problem.

However, say that we have a genuine hypothesis testing problem. Then the decision depends on a choice of critical region. This critical region is sometimes defined by taking a 0.05 level of the probability of type I error, given that the null hypothesis is true. This is quite traditional, but it completely ignores the probability of type II error, given the alternative hypothesis. It also ignores the amount of loss due to a wrong decision.

So in choosing the level of a test, one should also look at the probabilities of both kinds of error. Presumably one should also look at the loss function. Perhaps one should guard against the possibility of a huge loss.

However even knowing this does not determine a test. One method of resolving the question is to single out prior probabilities and use a Bayes test.

This has the standard weakness of all Bayes procedures, namely, it may not be clear what objective basis there is for choosing a prior distribution. However if one uses a Bayes test and is convinced that the risks are acceptable, this may be the way to go.

Another method may be to use a minimax test. This may be appropriate for pessimists, but there is no reason to suppose that the nature is engaged in a conspiracy against statisticians. The assumption that this is so may be considered as somewhat like a Bayesian assumption, but rather than using a hunch about what is likely to be the case, one is assuming the worst possible scenario. It is possibly that worrying about the worst possible situation puts too much emphasis on a narrow part of parameter space. But who is to decide what is narrow? This brings us back to the Baysian problem. It seems that there is no best way to make decisions; the best one can hope for is to choose from a collection of reasonable procedures. In many cases these reasonable procedures will be among the Baysian procedures or their limits.

11.10 Problems

1. Statistician A is a consultant in a physics laboratory. This laboratory has samples of a certain radioactive substance. It is known that the decay time is governed by the law

$$f(x \mid \theta) = \theta e^{-\theta x} \tag{11.44}$$

for $x \geq 0$. Here $\theta > 0$ is an unknown parameter. However the substance is known to be of one of two kinds, each with a known decay rate. Thus either $\theta = \theta_0$ or $\theta = \theta_1$, where $\theta_0 < \theta_1$. Find the joint density $f(x_1, \ldots, x_n \mid \theta)$ for n independent observations.

2. Statistician A proposes to use a decision function d that has the values 1 and 0. The idea is to decide for θ_1 if $d(x_1, \ldots, x_n) = 1$ and to decide for θ_0 if $d(x_1, \ldots, x_n) = 0$. What is the probability of (incorrectly) deciding for θ_1 if $\theta = \theta_0$. What is the probability of (incorrectly) deciding for θ_0 if $\theta = \theta_1$?

3. Statistician A is under considerable pressure to make the best use of the data to make a correct decision. The loss from deciding for θ_0 when $\theta = \theta_1$ is $L(\theta_1, 0) > 0$. The loss from deciding for θ_1 when $\theta = \theta_0$ is $L(\theta_0, 1) > 0$. There is no loss for making a correct decision. What is the risk at θ_0? What is the risk at θ_1?

4. Statistician A is not a Bayesian, but Statistican A has a Bayesian friend B who is willing to dream up subjective prior probabilities in almost any circumstance. B assigns prior probabilities $\pi(\theta_0)$ and $\pi(\theta_1)$ with $\pi(\theta_0) + \pi(\theta_1) = 1$. Then B can calculate the posterior probabilities given the data,

which are

$$\pi(\theta_0 \mid x_1, \ldots x_n) = \frac{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0)}{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)} \tag{11.45}$$

and

$$\pi(\theta_1 \mid x_1, \ldots x_n) = \frac{f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)}{f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) + f(x_1, \ldots, x_n \mid \theta_1)\pi(\theta_1)} \tag{11.46}$$

These look like they might depend on the sample in a complicated way. Make B happy by writing them in terms of the sample mean $\bar{x}$.

5. Even though B has no particular interest in loss functions, it seems only reasonable to want to help his friend A. If the decision function $d(x_1, \ldots, x_n)$ is used, then the posterior loss is $L(\theta_0, 1)\pi(\theta_0 \mid x_1, \ldots, x_n)$ where $d(x_1, \ldots, x_n) = 1$ and is $L(\theta_1, 0)\pi(\theta_1 \mid x_1, \ldots, x_n)$ where $d(x_1, \ldots, x_n) = 0$. This loss is minimized by taking $d(x_1, \ldots, x_n) = 1$ if the first of these losses is less than the second one of the losses, and by taking $d(x_1, \ldots, x_n) = 0$ if the second is less than the first. (If the losses are equal, then B can choose either 1 or 0.) Find the condition on the relative sizes of the likelihood functions $f(x_1, \ldots, x_n \mid \theta_1)$ and $f(x_1, \ldots, x_n \mid \theta_0)$ that is equivalent to $d(x_1, \ldots, x_n) = 1$.

6. Express this condition in terms of a condition such as $\bar{x} \leq c$ on the sample mean. Find the constant c. This is the test that B recommends to A. Show that even though B could have recommended another Bayes decision function with the same prior, the risk function that A examines is uniquely determined by the prior provided by B.

7. Show that if B assigns much higher prior probability to θ_0 than to θ_1, then B will recommend ignoring the evidence and always deciding for θ_0. Statistician A finds this disturbing. A politely listens to B, but has private doubts.

8. Ultimately statistician A is not comfortable with Bayesian ideas and prefers to work with the probabilities $P_\theta[\bar{X} \leq c]$ for the two possible values of θ. What is are the mean and standard deviation of $\bar{X}$ in terms of θ? How could he use the central limit theorem to compute the probabilities?

9. Statistician A has at least learned from B that each test of the form $\bar{x} \leq c$ is a Bayes test with a uniquely determined risk function. A decision theory analysis shows that for each such Bayes test there is no way to find find another test that decreases risk at both θ_0 and θ_0. So even from a non-Bayesian point of view these tests are reasonable to consider. However A does not like to even think about the actual numerical values of the Bayesian prior probabilities. For want of a better idea, A decides to use the conservative minimax criterion $L(\theta_0, 1)P_{\theta_0}[\bar{X} \leq c] = L(\theta_1, 0)P_{\theta_1}[\bar{X} > c]$

with equal risks. An examination of the practical consequences of wrong decisions shows that $L(\theta_0, 1)$ is twice $L(\theta_1, 0)$. The values of θ_0 and θ_1 are $1/15$ and $1/10$. The sample size is $n = 25$. What is the cutoff c? [This may take some trial and error to find.] With this choice of c, what are the probabilities under each of the two hypotheses of an incorrect decision?

10. Statistician A finds experimentally that $\bar{X} = 12$. What is the ultimate decision?

Chapter 12

Bayes and likelihood procedures

12.1 Bayes decisions

Recall the ingredients of a statistical decision procedure. For each value θ of the unknown parameter there is a distribution $f(x_1, \ldots, x_n \mid \theta)$ of the data. Also, for each parameter value θ and each action a there is a loss $\mathcal{L}(\theta, a)$. The job of a statistician is to use the data to choose an action $a = d(x_1, \ldots, x_n)$. The risk of the procedure d is defined for each θ by

$$\mathcal{R}(\theta, d) = \int \mathcal{L}(\theta, d(x_1, \ldots, x_n)) f(x_1, \ldots, x_n \mid \theta)\, dx_1 \cdots dx_n. \tag{12.1}$$

It is the expected loss when this decision procedure is used on the data.

Let d be a decision procedure. Clearly, if there is another procedure d' such that for all θ we have $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ and for some θ also $\mathcal{R}(\theta, d') < \mathcal{R}(\theta, d)$, then the second procedure d' is an improvement over the first one. There is no reason to use the first one, except perhaps convenience.

If, on the other hand, whenever for all θ we have $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$ we also have $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$, then the procedure d cannot be improved on. This does not mean that we should use the procedure d. There may be other procedures that also cannot be improved on. However it shows that in some sense the procedure d is not completely stupid. Bayes procedures often have this desirable property. This is expressed in the following theorem.

Theorem 12.1 *Suppose that d is a Bayes procedure arising from a prior $\pi(\theta)$. Suppose that d cannot be improved on by a Bayes procedure with the same prior. The conclusion is that the Bayes procedure d cannot be improved on.*

Proof: A Bayes procedure with prior distribution π is obtained by finding a d that makes

$$r(\pi, d) = \int \mathcal{R}(\theta, d)\, \pi(\theta)\, d\theta \tag{12.2}$$

minimal. Thus if d' is any other procedure, then $r(\pi, d) \leq r(\pi, d')$. Suppose that d' is another procedure (not necessarily Bayes) such that for all θ we have $\mathcal{R}(\theta, d') \leq \mathcal{R}(\theta, d)$. Then

$$\int \mathcal{R}(\theta, d')\pi(\theta)\, d\theta \leq \int \mathcal{R}(\theta, d)\pi(\theta)\, d\theta. \tag{12.3}$$

This shows that d' is actually a Bayes procedure with respect to the prior distribution π. Hence by assumption $\mathcal{R}(\theta, d') = \mathcal{R}(\theta, d)$ for all θ. Thus d' cannot improve on d.

12.2 Estimation

There are not many general principles for making statistical decisions. The purpose of this chapter is to compare two of these: the likelihood principle and the use of a Bayesian prior.

The most common justification for the likelihood principle is that it has good properties in the limit of large sample sizes. But Bayes methods may also have good properties in the limit of large sample sizes. Other methods may share this desirable feature. So good large sample behavior is not a principle that gives a unique way of doing statistics.

One possible justification for using a Bayes method is that it cannot be improved on. In fact, in many cases it is true that the methods that cannot be improved on are the Bayes methods, or the limits of Bayes methods. This argues for the point of view that the Bayes methods form a natural and desirable class of statistical methods.

However the fact that a particular Bayes method cannot be improved on does not mean that it is the right thing to use. There are many Bayes methods, corresponding to many choices of prior distributions. Each of them cannot be improved on, since each is as good as possible in the situation where nature has chosen its own prior distribution. But this does not make clear which one is to be preferred. To pick a particular Bayes method, one must argue that the choice of prior $\pi(\theta)$ is natural, or that the resulting Bayes decision procedure d has a risk function $\mathcal{R}(\theta, d)$ that is acceptable.

As for the relation between the likelihood principle and the Bayes principle, we shall see that in some situations the likelihood principle is a special case of the Bayes principle. So this is also support for taking the Bayes point of view as more fundamental.

The joint distribution $f(x_1, \ldots, x_n \mid \theta)$ as a function of the unknown parameter θ, for fixed data values $x_1, \ldots, x_n$, is called the likelihood. In estimation the maximum likelihood principle says that given the data $x_1, \ldots, x_n$, the estimate of the unknown θ is that $\theta = \hat{\theta}$ that maximizes the likelihood $f(x_1, \ldots, x_n \mid \theta)$. In practice this is often computed by solving

$$\frac{\partial}{\partial \theta} \log f(x_1, \ldots, x_n \mid \theta) = 0 \tag{12.4}$$

for θ.

By contrast, the Bayes estimator needs extra information: the prior distribution $\pi(\theta)$. The posterior distribution is then

$$\pi(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{f(x_1, \ldots, x_n)}. \tag{12.5}$$

Here

$$f(x_1, \ldots, x_n) = \int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta \tag{12.6}$$

is the marginal distribution of the data. The Bayes estimator is calculated by taking $d(x_1, \ldots, x_n) = a$, where a minimizes the posterior loss

$$E_{x_1,\ldots,x_n}[\mathcal{L}(\Theta, a] = \int \mathcal{L}(\theta, a)\pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{12.7}$$

Let us look at various possible loss functions. We shall see that different loss functions lead to different estimators.

If the loss function is $\mathcal{L}(\theta, a) = (\theta - a)^2$, then the minimum is achieved when the derivative

$$\frac{d}{da}\int (\theta - a)^2 \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = 2\int (\theta - a)\pi(\theta \mid x_1, \ldots, x_n)\, d\theta = 0. \tag{12.8}$$

This minimum is clearly when a is the expectation of the posterior distribution

$$E_{x_1,\ldots,x_n}[\Theta] = \int \theta\pi(\theta \mid x_1, \ldots, x_n)\, d\theta. \tag{12.9}$$

If the loss function is $\mathcal{L}(\theta, a) = |\theta - a|$, then the minimum is achieved when the derivative

$$\frac{d}{da}\int |\theta - a|\pi(\theta \mid x_1, \ldots, x_n)\, d\theta = \int \operatorname{sign}(\theta - a)\pi(\theta \mid x_1, \ldots, x_n)\, d\theta = 0. \tag{12.10}$$

This minimum is clearly when

$$\int_{\theta > a} \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = \int_{\theta < a} \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = 0. \tag{12.11}$$

Thus a is the median of the posterior distribution

If the loss function is $\mathcal{L}(\theta, a) = 1$ for $|\theta - a| > \epsilon$, 0 otherwise, then the Bayes loss is the probability $P_{x_1,\ldots,x_n}[|\Theta - a| > \epsilon]$. To minimize the loss is the same as to maximize the probability $P_{x_1,\ldots,x_n}[|\Theta - a| \leq \epsilon]$. The maximum is achieved when the derivative

$$\frac{d}{da}\int_{|\theta - a| < \epsilon} \pi(\theta \mid x_1, \ldots, x_n)\, d\theta = \pi(a + \epsilon \mid x_1, \ldots, x_n) - \pi(a - \epsilon \mid x_1, \ldots, x_n) = 0. \tag{12.12}$$

If ϵ is small, then this quantity is approximately

$$\pi(a+\epsilon \mid x_1, \ldots, x_n) - \pi(a-\epsilon \mid x_1, \ldots, x_n) \approx 2 \frac{\partial}{\partial \theta} \pi(\theta \mid x_1, \ldots, x_n)|_{\theta=a} \, \epsilon. \quad (12.13)$$

Thus the condition is that

$$\frac{\partial}{\partial \theta} \pi(\theta \mid x_1, \ldots, x_n)|_{\theta=a} = 0. \quad (12.14)$$

The conclusion is that a is the mode (maximum) of the posterior distribution.

For this last case the Bayes procedure is to start with the data and maximize $\pi(\theta \mid x_1, \ldots, x_n)$. It is equivalent to maximize the numerator $f(x_1, \ldots, x_n \mid \theta)\pi(\theta)$. This makes apparent the relation to the maximum likelihood estimator. If we take $\pi(\theta) = 1$, then we get the maximum likelihood estimator.

Why not, then, simply take $\pi(\theta) = 1$ as the basic principle. For one thing, this is somewhat arbitrary. It is certainly not a hypothesis that is compatible with the principle of invariance. If we are interested in estimating a function of θ, then the corresponding probability density will not be constant, but will involve the absolute value of the derivative of the function. Perhaps in some circumstances the prior $\pi(\theta) = 1$ can be derived from a symmetry principle. But it is hardly fundamental.

The conclusion is that maximum likelihood estimation is at least somewhat related to Bayes estimation. In some circumstances this may provide a justification for this principle.

12.3 Testing

The likelihood ratio principle is used in testing. We consider a case when the null hypothesis is that the true value of θ is θ_0. From here on we shall think of θ_0 as a fixed constant. The alternative hypothesis is some other value of θ. The likelihood ratio principle attempts to use the likelihood ratio

$$k \leq \frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)}. \quad (12.15)$$

to define the critical region for the test, where one would guess the alternative hypothesis. If there is only value $\theta = \theta_1$ in the alternative hypothesis, then this test is well defined and is a Bayes test. The problem is that in general there can be several possibilities for the unknown θ.

One way around this is to consider a situation where there is a test statistic $t(x_1, \ldots, x_n)$ such that

$$\frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)} = F(t(x_1, \ldots, x_n), \theta). \quad (12.16)$$

We consider the one-sided situation where the null hypothesis is $\theta = \theta_0$ and the alternative hypothesis is $\theta > \theta_0$. The assumption of monotone likelihood ratio is that for each $\theta > \theta_0$ the function $F(z, \theta)$ is increasing in z.

Under the assumption of monotone likelihood ratio, the test with critical region

$$a \leq t(x_1, \ldots, x_n), \tag{12.17}$$

is equivalent to a test

$$F(a, \theta) \leq \frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)}. \tag{12.18}$$

So in this way of proceeding the ratio $k = F(a, \theta)$ for the test depends on the unknown θ. This may not matter so much; the test itself is based on a.

Example: Take the classical case when the population distribution is normal with unknown mean θ and known variance σ^2. The likelihood ration is easy to compute, and after some algebra it works out to be

$$\frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)} = \exp(\frac{n}{2\sigma^2}(-(\bar{x} - \theta)^2 + (\bar{x} - \theta_0)^2)). \tag{12.19}$$

The test statistic is $t(x_1, \ldots, x_n) = \bar{x}$, the sample mean. The function F is

$$F(z, \theta) = \exp(\frac{n}{2\sigma^2}(-(z - \theta)^2 + (z - \theta_0)^2)) = \exp(\frac{n}{2\sigma^2}(-\theta^2 + 2(\theta - \theta_0)z + \theta_0^2)). \tag{12.20}$$

Indeed, for each $\theta > \theta_0$ it is an increasing function of z. The likelihood ratio is $F(\bar{x}, \theta)$. So the critical region for this kind of test is given by a condition $a \leq \bar{x}$ on the sample mean.

We now show that this version of the likelihood ratio test may be interpreted in a Baysian framework. We take the null hypothesis to be $\theta = \theta_0$. The loss from guessing the alternative hypothesis is $\mathcal{L}(I)$. The alternative hypothesis is $\theta > \theta_0$. The loss from guessing the null hypothesis is $\mathcal{L}(II)$.

We take a Bayes prior that assigns probability $\pi(\theta_0) > 0$ to the null hypothesis. The prior probabilities associated with the alternative hypothesis are given by a density $\pi(\theta)$ for $\theta > \theta_0$ with total probability $1 - \pi(\theta_0)$. This means that

$$\pi(\theta_0) + \int_{\theta_0}^{\infty} \pi(\theta)\, d\theta = 1. \tag{12.21}$$

There are two actions. The posterior risk from the action of guessing the alternative hypothesis is proportional to $\mathcal{L}(I) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0)$. The posterior risk from the action of guessing the null hypothesis is proportional to $\mathcal{L}(II) \int_{\theta_0}^{\infty} f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta$. The critical region is where the first risk is less than the second risk. This is where

$$\mathcal{L}(I) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) \leq \mathcal{L}(II) \int_{\theta_0}^{\infty} f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta. \tag{12.22}$$

This can be written in the form

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq \int_{\theta_0}^{\infty} \frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)} \frac{\pi(\theta)}{\pi(\theta_0)}\, d\theta. \tag{12.23}$$

This integral involves the likelihood ratio. If we assume that the likelihood ration is a function of a test statistic, this can also be written in the form

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq \int_{\theta_0}^{\infty} F(t(x_1, \ldots, x_n), \theta) \frac{\pi(\theta)}{\pi(\theta_0)} \, d\theta. \tag{12.24}$$

Define

$$G(z) = \int_{\theta_0}^{\infty} F(z, \theta) \frac{\pi(\theta)}{\pi(\theta_0)} \, d\theta. \tag{12.25}$$

The critical region is

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq G(t(x_1, \ldots, x_n)). \tag{12.26}$$

Make the assumption of monotone likelihood ratio, so that the function $F(z, \theta)$ is increasing in z for fixed $\theta > \theta_0$. Then $G(z)$ is also an increasing function. So this is the same as the likelihood ratio test with critical region $a \leq t(x_1, \ldots, x_n)$. The constant a is determined by

$$G(a) = \frac{\mathcal{L}(I)}{\mathcal{L}(II)}. \tag{12.27}$$

This argument shows that in this special case of a one-sided test there an equivalence between Bayes ideas and likelihood methods. Since a Bayes method cannot be improved on, the same follows for the likelihood ratio test.

Example: Let us continue with the example of the normal population with mean θ. The function G is given by

$$G(z) = \int_{\theta_0}^{\infty} \exp(\frac{n}{2\sigma^2}(-\theta^2 + 2(\theta - \theta_0)z + \theta_0^2)) \frac{\pi(\theta)}{\pi(\theta_0)} \, d\theta. \tag{12.28}$$

The function $G(z)$ is given by a complicated integral that may have to be done numerically. However it is clear that it is increasing in z. So the critical region for the test is still of the form $a \leq \bar{x}$. This is equivalent to the Bayes test $G(a) \leq G(\bar{x})$.

There is another way to use the likelihood principle to perform a test. Again we consider the one-sided alternative $\theta > \theta_0$. The method is to insert the maximum likelihood estimator in the likelihood function used for the test. In the present case this is equivalent to using the test with critical region

$$k \leq \frac{\max_{\theta > \theta_0} f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)}. \tag{12.29}$$

If the likelihood ratio is a function of a test statistic, then this is

$$k \leq \max_{\theta > \theta_0} F(t(x_1, \ldots, x_n), \theta). \tag{12.30}$$

If we set $H(z) = \max_{\theta > \theta_0} F(z, \theta)$, then the critical region is

$$k \leq H(t(x_1, \ldots, x_n)). \tag{12.31}$$

Under the assumption of monotone likelihood ration, the function $H(z)$ is increasing in z. So the test is equivalent to the test $a \leq t(x_1, \ldots, x_n)$, where $H(a) = k$. However the function H may be quite different from the function G in the Baysian analysis. This procedure coincides with a Bayes procedure, but the ideas are not so close to Baysian ideas.

Example: Again take the example of the normal population with mean θ. The function $H(z)$ has a very simple expression:

$$H(z) = \max_{\theta > \theta_0} \exp(\frac{n}{2\sigma^2}(-(z-\theta)^2 + (z-\theta_0)^2)) = \exp(\frac{n}{2\sigma^2}(z-\theta_0)^2)). \quad (12.32)$$

Actually, this expression is correct only for $z \geq \theta_0$, since for $z < \theta_0$ we have $H(z) = 1$. Again it is clear that $H(z)$ is increasing in z. So the critical region for the test is still of the form $a \leq \bar{x}$. This is equivalent to the likelihood ratio test $k \leq H(\bar{x})$.

We can see the contrast better if we go to two-sided alternatives. Thus the null hypothesis is $\theta = \theta_0$ and the alternative hypothesis is $\theta \neq \theta_0$. The most common way to do a likelihood ratio test would be to make the critical region be

$$k \leq \frac{\max_\theta f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)}. \quad (12.33)$$

If the likelihood ratio is a function of a test statistic, then this is

$$k \leq \max_\theta F(t(x_1, \ldots, x_n), \theta). \quad (12.34)$$

If we set $H(z) = \max_\theta F(z, \theta)$, then the critical region is

$$k \leq H(t(x_1, \ldots, x_n)). \quad (12.35)$$

Example: Again take the example of the normal population with mean θ. The function $H(z)$ has a very simple expression:

$$H(z) = \max_{\theta} \exp(\frac{n}{2\sigma^2}(-(z-\theta)^2 + (z-\theta_0)^2)) = \exp(\frac{n}{2\sigma^2}(z-\theta_0)^2)). \quad (12.36)$$

Notice that this is not increasing in z. The critical region for the test is of the form $k \leq H(\bar{x})$. This produces a two sided test, where the critical region is where $\bar{x}$ is sufficiently far above or below θ_0.

We can also analyze the two-sided alternative in the Baysian framework. We take the null hypothesis to be $\theta = \theta_0$. The loss from guessing the alternative hypothesis is $\mathcal{L}(I)$. The alternative hypothesis is $\theta \neq \theta_0$. The loss from guessing the null hypothesis is $\mathcal{L}(II)$.

We take a Bayes prior that assigns probability $\pi(\theta_0) > 0$ to the null hypothesis. The prior probabilities associated with the alternative hypothesis are given by a density $\pi(\theta)$ with total probability $1 - \pi(\theta_0)$. This means that

$$\pi(\theta_0) + \int_{-\infty}^{\infty} \pi(\theta)\, d\theta = 1. \quad (12.37)$$

The critical region is where

$$\mathcal{L}(I) f(x_1, \ldots, x_n \mid \theta_0)\pi(\theta_0) \leq \mathcal{L}(II) \int_{-\infty}^{\infty} f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta. \tag{12.38}$$

This can be written in the form

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq \int_{-\infty}^{\infty} \frac{f(x_1, \ldots, x_n \mid \theta)}{f(x_1, \ldots, x_n \mid \theta_0)} \frac{\pi(\theta)}{\pi(\theta_0)}\, d\theta. \tag{12.39}$$

This integral involves the likelihood ratio. If we assume that the likelihood ration is a function of a test statistic, this can also be written in the form

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq \int_{-\infty}^{\infty} F(t(x_1, \ldots, x_n), \theta) \frac{\pi(\theta)}{\pi(\theta_0)}\, d\theta. \tag{12.40}$$

Define

$$G(z) = \int_{-\infty}^{\infty} F(z, \theta) \frac{\pi(\theta)}{\pi(\theta_0)}\, d\theta. \tag{12.41}$$

The critical region is

$$\frac{\mathcal{L}(I)}{\mathcal{L}(II)} \leq G(t(x_1, \ldots, x_n)). \tag{12.42}$$

The Bayes has roughly the same form as the likelihood ratio test, but it is hard to see the correspondence. A likelihood ratio test might well be a Bayes test, but this would be a kind of accident, and it would take some analysis to discover this fact. In any case, the Bayes test itself uses an integral of likelihood ratios, so the two method have something of the same spirit. It is the Bayes method that has a more fundamental justification.

12.4 Problems

1. Consider a random sample of size 3 from a Bernoulli population. The null hypothesis is that $p = 1/2$, while the alternative hypothesis is that $p = 2/3$. The test must be such that when $p = 1/2$ the probability of a type I error cannot exceed $1/8$. What critical region defines a test such that if $p = 2/3$ the probability of a type II error is as small as possible?

2. Say that $f(x \mid \theta) = \theta e^{-\theta x}$ for $x > 0$ is an exponential density. Let $\theta_1 < \theta_0$. What is the best critical region, based on a sample of size n, for testing the null hypothesis that $\theta = \theta_0$ against the alternative hypothesis that $\theta = \theta_1$?

3. Consider a normal population with unknown μ and with $\sigma = 1$ known. Say that a test is to discriminate between the null hypothesis that $\mu = 10$ and the alternative hypothesis that $\mu = 11$. A random sample of size 25 is available. The test is at level 0.1 and the problem is to find the best test and to calculate the probability of a type II error.

4. Consider a normal population with $\mu = 0$ and with σ^2 unknown. Say that a test is to discriminate between the null hypothesis that $\sigma^2 = 10$ and the alternative hypothesis that $\sigma^2 = 12$. A random sample of size 10 is available. The test is at level 0.1 and the problem is to calculate the probability of a type II error for the best test.

5. Consider a normal random variable with μ unknown and $\sigma = 1$. The null hypothesis is that $\mu = 0$ and the alternative hypothesis is two-sided. The sample size is 4. The test is the natural symmetric test at the 0.05 level. Calculate and graph the power function for this test.

6. Consider a sample of size 3 from a Bernoulli population. Let X be the number of successes, so that $X = 0, 1, 2, 3$ are all possible. The null hypothesis is that $p = 1/3$. The critical region of a test is $X = 3$. Calculate the power function. The critical region of another test is $X = 2, 3$. Calculate the power function. Compare these power functions graphically. Is one better than another?

Chapter 13

Regression and Correlation

13.1 Regression

In the regression model there are values x_i that are given and known. For each x_i there is a corresponding observation Y_i. Here i ranges from 1 to n. The model is

$$Y_i = \alpha + \beta x_i + \epsilon_i. \tag{13.1}$$

Here the ϵ_i are independent normal random variables with mean zero and variance σ^2. These are not observed. Thus the unknown parameters of the model are the α and β that convey the true linear relationship and the σ that conveys the size of the error. The true linear relationship is expressed by saying that the mean of Y_i is $\mu_i = \alpha + \beta x_i$. The scientific problem is to capture this relationship from the experimental data points Y_i.

Given the data Y_i, the estimate β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{13.2}$$

Also, the estimate of α is determined by

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{x}. \tag{13.3}$$

The estimate of μ_i is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i. \tag{13.4}$$

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}. \tag{13.5}$$

These are all unbiased estimators.

As usual, we want to know how good a job these estimators do. The variance of $\hat{\beta}$ is

$$\sigma_{\hat{\beta}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{13.6}$$

Therefore the natural estimate of the variance of $\hat{\beta}$ is

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{13.7}$$

For a test of the null hypothesis $\beta = \beta_0$ one can use the t statistic

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}}. \tag{13.8}$$

If the null hypothesis is true, this has a t distribution with $n - 2$ degrees of freedom.

Sometimes people prefer to do regression analysis in the context of a sum of squares identity. For this case the identity expresses the sum of squares of the observations about their mean as a sum of squares of the observations about the regression line plus another sum of squares term that is said to be explained by the regression line. The identity is

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2. \tag{13.9}$$

Since $\hat{Y}_i - \bar{Y} = \hat{\beta}(x_i - \bar{x})$, this can also be written

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 + \hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2. \tag{13.10}$$

If the null hypothesis $\beta = 0$ is true, then the sum of squares explained by the regression line should be small compared to the sum of squares about the regression line. Thus the F statistic

$$F = \frac{\hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y})^2/(n-2)} \tag{13.11}$$

should not be too large. It turns out that this is equivalent to the t test for $\beta = 0$ given above.

There is yet another language for the analysis of regression experiments. Define the sample correlation coefficient r by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{13.12}$$

Then there is an alternate expression for $\hat{\beta}$, namely

$$\hat{\beta} = r\frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{13.13}$$

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}{n-2}(1 - r^2). \tag{13.14}$$

The sum of squares identity becomes

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + r^2 \sum_{i=1}^n (Y_i - \bar{Y})^2. \tag{13.15}$$

From this we see that the F statistic for testing $\beta = 0$ is

$$F = \frac{r^2}{\frac{1-r^2}{n-2}}. \tag{13.16}$$

If one is interested in a one-sided test, one can use the t statistic for testing $\beta = 0$, which is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}. \tag{13.17}$$

This is an especially convenient form of the test. In this situation the relation between F and t is that $F = t^2$.

There is another way of looking at the test of the null hypothesis $\beta = \beta_0$. That is to change variables to reduce it to this special case. Let the mean of $\tilde{Y}_i = Y_i - \beta_0 x_i$ be given by $\tilde{\mu}_i = \alpha + \tilde{\beta} x_i$, where $\tilde{\beta} = \beta - \beta_0$. Then in these new variables the null hypothesis is $\tilde{\beta} = 0$. Calculate the sample correlation coefficient $\tilde{r}$ with the new variables $\tilde{Y}_i$. The test statistic is

$$t = \frac{\tilde{r}}{\sqrt{(1 - \tilde{r}^2)/(n-2)}}. \tag{13.18}$$

An algebraic calculation shows that this is the same as the test for $\beta = \beta_0$ given above.

This reduction to the special case $\tilde{\beta} = 0$ is theoretically important. It shows that it is sufficient to deal with the case when the null hypothesis is defined by a homogeneous equation. This is convenient from the point of view of linear algebra.

13.2 Correlation

In the correlation model the data points are random pairs X_i, Y_i. The model is that the mean and variance of X_i are μ_X and σ_X^2 and the mean and variance of Y_i are μ_Y and σ_Y^2. The covariance of X_i and Y_i is

$$\mathrm{cov}(X_i, Y_i) = \rho \sigma_X \sigma_Y. \tag{13.19}$$

Here ρ is the population correlation coefficient.

Given the data (X_i, Y_i), the estimates of μ_X and μ_Y are the usual sample means $\bar{X}$ and $\bar{Y}$. The estimates of σ_X^2 and σ_Y^2 are

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \tag{13.20}$$

and

$$s_Y^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}. \tag{13.21}$$

The estimator of ρ is the sample correlation coefficient r given by

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}. \tag{13.22}$$

Notice that up to now everything is symmetric between the two variables. Observe also that the correlation analysis does not specify a line in the data plane.

We can condition on the event that X_i has some value x_i. The conditional expectation is

$$E[Y_i \mid X_i = x_i] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x_i - \mu_X). \tag{13.23}$$

This can also be written

$$E[Y_i \mid X_i = x_i] = \alpha + \beta x_i, \tag{13.24}$$

where α and β are determined by

$$\beta = \rho\frac{\sigma_Y}{\sigma_X} \tag{13.25}$$

and

$$\mu_Y = \alpha + \beta\mu_X. \tag{13.26}$$

This is a regression model. The regression line for this model would have

$$\hat{\beta} = r\frac{s_Y}{s_X} \tag{13.27}$$

and

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{x}. \tag{13.28}$$

If the null hypothesis in the correlation model is $\rho = 0$, then the corresponding null hypothesis in the regression model is $\beta = 0$. So it is no surprise that the test statistic for this situation is also based on

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{13.29}$$

or on $F = t^2$.

We can also condition on the event that Y_i has some value y_i. The conditional expectation is

$$E[X_i \mid Y_i = y_i] = \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y_i - \mu_Y). \tag{13.30}$$

This can also be written

$$E[X_i \mid Y_i = y_i] = \alpha' + \beta' y_i, \tag{13.31}$$

where α' and β' are determined by

$$\beta' = \rho \frac{\sigma_X}{\sigma_Y} \tag{13.32}$$

and

$$\mu_X = \alpha' + \beta' \mu_Y. \tag{13.33}$$

This is a different regression model. The regression line for this model would have

$$\hat{\beta}' = r \frac{s_Y}{s_X} \tag{13.34}$$

and

$$\bar{X} = \hat{\alpha}' + \hat{\beta}' \bar{y}. \tag{13.35}$$

Notice that the regression lines with the two kinds of conditioning differ, except for the case of perfect correlation $r = 1$. For instance, take r very small, and plot the x variables horizontally and the y variables vertically. Then the line is almost horizontal in the first model (conditioning on $X_i = x_i$) and the line is almost vertical in the second model (conditioning on $Y_i = y_i$).

13.3 Principal component analysis

There is one circumstance in which it is reasonable to draw a line through the data in a correlation experiment. This is when the units of measurement in the two variables is the same.

The population covariance matrix is the matrix

$$\begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \tag{13.36}$$

If the units are the same, then the eigenvalues and eigenvectors of this symmetric matrix are meaningful. The eigenvalues are real, and the eigenvectors define two axes, orthogonal directions in the plane. There are two lines in these directions that pass through the mean point (μ_X, μ_Y). The eigenvalues are not equal, except for the case when $\rho = 0$ and $\sigma_X^2 = \sigma_Y^2$. The line corresponding to the largest eigenvalue is the principal axis.

The sample estimators for this case make up a matrix

$$\begin{bmatrix} s_X^2 & r s_X s_Y \\ r s_X s_Y & s_Y^2 \end{bmatrix}. \tag{13.37}$$

The eigenvalues are real, and the eigenvectors define two axes, orthogonal directions in the plane. One can draw two lines in these directions that pass through the mean point $(\bar{X}, \bar{Y})$. The eigenvalues are not equal, except for the case when $r = 0$ and $s_X^2 = s_Y^2$. The line corresponding to the largest eigenvalue is the principal axis.

What does this principal axis mean in terms of the data? Make the transformation to new variables by a rotation

$$(Z_i - \bar{Z}) = \cos(\theta)(X_i - \bar{X}) + \sin(\theta)(Y_i - \bar{Y}) \tag{13.38}$$

and

$$(W_i - \bar{W}) = -\sin(\theta)(X_i - \bar{X}) + \cos(\theta)(Y_i - \bar{Y}). \tag{13.39}$$

Let r' be the correlation of the new variables. Then

$$\begin{bmatrix} s_Z^2 & r' s_Z s_W \\ r' s_Z s_W & s_W^2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} s_X^2 & r s_X s_Y \\ r s_X s_Y & s_Y^2 \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}. \tag{13.40}$$

If the λ and μ are the eigenvalues of the original covariance matrix, and the two columns of the matrix on the right are the corresponding eigenvectors, then we have

$$\begin{bmatrix} s_X^2 & r s_X s_Y \\ r s_X s_Y & s_Y^2 \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}. \tag{13.41}$$

Hence

$$\begin{bmatrix} s_Z^2 & r' s_Z s_W \\ r' s_Z s_W & s_W^2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}. \tag{13.42}$$

Thus the new variables are uncorrelated, and their variances are the eigenvalues. The new variable that has larger variance is the one measured on the principal axis.

The formula for the θ that gives eigenvectors and hence makes the matrix diagonal is given by the following formula. Let

$$\tan(\chi) = \frac{s_Y}{s_X} \tag{13.43}$$

be the slope determined by the two standard deviations. Since they have the same units, this is dimensionless. Then the condition is that

$$\tan(2\theta) = r \tan(2\chi). \tag{13.44}$$

One can take χ between 0 and $\pi/2$. Then 2χ is between 0 and π. There are two solutions for 2θ between 0 and 2π that differ by π. So there are two solutions for θ between 0 and π that differ by $\pi/2$. These give the two axes corresponding to the two eigenvectors.

The relation between the slope $k = \tan(\theta)$ of such an axis and the slope $m = \tan(\chi)$ may also be expressed algebraically. In fact the double angle

formula for the tangent function shows that the identity $\tan(2\theta) = r\tan(2\chi)$ may be written in the form

$$\frac{1}{k} - k = \frac{1}{r}\left(\frac{1}{m} - m\right). \tag{13.45}$$

This is quite nonlinear. So changing the standard deviation in one direction by a constant factor does not change the slope of the principal axis by the same factor. This shows once again how crucial it is that the two axes have the same units.

This formula for the axes is useful computationally but not particularly nice. The formula for the corresponding eigenvalues is also awkward. However it is easy to see what the answer is in certain extreme cases.

When $r = 0$, the axes are along the directions of the x and y coordinate axes. The eigenvalues are s_X^2 and s_Y^2, and so the principal axis is the one with the largest variance.

When $r = \pm 1$ the eigenvalues are $s_X^2 + s_Y^2$ and 0. The principal axis is the one with slope $\pm s_Y/s_X$.

Perhaps the most interesting case is when $s_X^2 = s_Y^2$. In that case the eigenvalues are this common value times $1 \pm r$. The corresponding axes have slopes ± 1. The principal axis is the one with slope equal to the sign of r.

Sometimes people want to do a principal component analysis when the units on the two axes are not the same. This does not make sense. It is not clear what the dimensions of the rotated variables would be. One way to make the problem dramatic is to look at the case when the correlation is zero. Then which axis has the larger variance? It depends entirely on the units. You can change the units on one axis and not on the other. This will change what you think of as the principal axis.

One obvious idea for taking care of the case of unequal units is to normalize the data by dividing $(X_i - \bar{X})$ by s_X and dividing $(Y_i - \bar{Y})$ by s_Y. This gives dimensionless variables. The resulting covariance matrix is

$$\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}. \tag{13.46}$$

As we have seen, the eigenvalues are $1 \pm r$. The corresponding axes have slopes ± 1. The principal axis is the one with slope equal to the sign of r. So there is no new information in the principal component analysis aside from the value of r. In this situation one would do just as well to simply report this value.

Chapter 14

Linear models: Estimation

14.1 Estimation

We consider a very general situation. The model is random variables $Y_1, \ldots, Y_n$. These are independent normal random variables, each with variance σ^2. The covariance is thus $\mathrm{Cov}(Y_i, Y_j) = \sigma^2 I_{ij}$. That is, the covariance matrix is σ^2 times the identity matrix I.

The mean vector μ in $\mathcal{R}(n)$ is unknown. Also σ^2 is unknown. However the mean μ is known to be in a subspace $\mathcal{L}$ of dimension k. The problem is to use the observation vector Y in $\mathcal{R}(n)$ to estimate the vector μ and the number σ^2.

The solution is simple. One uses the projection $\hat{Y}$ of Y onto $\mathcal{L}$ as the estimate of the vector μ in $\mathcal{L}$. The estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{|Y - \hat{Y}|^2}{n - k}. \tag{14.1}$$

Say the subspace $\mathcal{L}$ consists of all vectors of the form $Y = X\beta$, where X is a fixed matrix in $\mathcal{M}(n, k)$, and the parameter β varies over $\mathcal{R}(k)$. Then one can try to estimate the parameter vector β. The appropriate estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{14.2}$$

Then the estimator of μ is

$$\hat{\mu} = \hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y. \tag{14.3}$$

We can also work out the covariance of the estimator $\hat{\beta}$. We use the following lemma.

Lemma 14.1 *If the random vector Y has covariance matrix C, and A is a fixed matrix, then the covariance of the random vector AY is ACA'.*

Proof: Suppose $\mathrm{Cov}(Y_p, Y_q) = C_{pq}$. Then

$$\mathrm{Cov}(\sum_p A_{ip}Y_p, \sum_q A_{jq}Y_q) = \sum_p \sum_q A_{ip}C_{pq}A_{jq} = \sum_p \sum_q A_{ip}C_{pq}A'_{qj}. \tag{14.4}$$

Theorem 14.1 *Suppose that Y has covariance matrix $\sigma^2 I$. Then the covariance matrix of the estimator*

$$\hat{\beta} = (X'X)^{-1}XY \tag{14.5}$$

is

$$\sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \tag{14.6}$$

14.2 Regression

The regression model is

$$\mu(x) = \sum_{j=1}^{k} \beta_j f_j(x). \tag{14.7}$$

This expresses the expected value as a linear combination of functions $f_j(x)$ with coefficients β_j. It is assumed that the functions are known but the coefficients are unknown.

An actual observation has errors. So if we take values $x_1, \ldots, x_n$, the observations are

$$Y_i = \sum_{j=1}^{k} \beta_j f_j(x_i) + E_i. \tag{14.8}$$

The mean of Y_i is $\mu(x_i)$. The errors E_i have mean zero and variance σ^2 and are independent.

We can write this in matrix form by defining the matrix

$$X_{ij} = f_j(x_i). \tag{14.9}$$

This matrix depends on the points x_i that are chosen for observation. Sometimes the matrix X is called the *design matrix*. It is assumed to be known. We assume that it has zero null space.

Then the model takes the form

$$Y_i = \sum_{j=1}^{k} X_{ij}\beta_j + E_i. \tag{14.10}$$

Here the assumption is that the errors E_i have mean zero and variance σ^2 and are independent.

The statistical problem is to look at the Y_i and use these to estimate the parameters β_j. We can see what to do if we write the problem in matrix form

$$Y = X\beta + E. \tag{14.11}$$

The solution is to take the estimate $\hat{\beta}$ that gives the orthogonal projection $\hat{Y} = X\hat{\beta}$ onto the range of X. This is given by

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{14.12}$$

The vector $\hat{Y}$ that is the actual orthogonal projection is called the vector of predicted values.

The estimate of variance is then given by

$$\hat{\sigma}^2 = \frac{|Y - \hat{Y}|^2}{n - k}. \tag{14.13}$$

Example: The simplest case is when $k = 2$ and the model is $Y_i = \alpha + \beta(x_i - \bar{x}) + E_i$. The matrix X has two columns. The first column consists of ones. The second column consists of the $x_i - \bar{x}$ values. Subtracting the mean value of the x_i points is convenient, since then the two columns of X are orthogonal. With this choice the matrix

$$X'X = \begin{bmatrix} n & 0 \\ 0 & \sum_i (x_i - \bar{x})^2 \end{bmatrix}. \tag{14.14}$$

Then

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = (X'X)^{-1} \begin{bmatrix} \sum_i Y_i \\ \sum_i Y_i (x_i - \bar{x}) \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \sum_i (x_i - \bar{x}) Y_i / \sum_i (x_i - \bar{x})^2 \end{bmatrix}. \tag{14.15}$$

Thus the regression line is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x}). \tag{14.16}$$

Here $\hat{\alpha} = \bar{Y}$. For greater symmetry we use the identity $\sum_i (x_i - \bar{x}) Y_i = \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$ and write

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{14.17}$$

The estimator of variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}. \tag{14.18}$$

14.3 Analysis of variance: one way

The model is

$$Y_{ij} = \mu_j + E_{ij} \tag{14.19}$$

for $j = 1$ to c and $i = 1$ to n_j. The parameters are the μ_j. So there are a total of c parameters. Note for later comparison that we could write $\mu_j = \mu + b_j$, where μ is the mean of the μ_j, and where $\sum_{j=1}^c b_j = 0$.

The idea is that there are c populations, corresponding to c different treatments. Untreated, the populations would have the same mean. But the treatments may make a difference. The task is to estimate the effect of the treatments. The fact that the populations are similar except for the effect of the treatments is reflected in the assumption that all the errors E_{ij} have the same variance σ^2.

We take a sample of size n_j from the jth population. These are the Y_{ij} for $i = 1, \ldots, n_j$. All these numbers together form a vector space of dimension $n_1 + \cdots + n_c$. The subspace $\mathcal{L}$ consists of the vectors that do not depend on the i index. This is a subspace of dimension c. Thus the orthogonal projection onto $\mathcal{L}$ is obtained by averaging over the i index. Thus the estimator of μ_j is the sample mean of the sample from population j. This sample mean is denoted $\bar{Y}_{\cdot j}$. Thus

$$\hat{\mu}_j = \bar{Y}_{\cdot j}. \tag{14.20}$$

Since the σ^2 is the same for all the populations, we pool them together for the purpose of estimating σ^2. The estimator of σ^2 is thus

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - Y_{\cdot j})^2}{n - c}. \tag{14.21}$$

14.4 Analysis of variance: two way

The model is

$$Y_{ij} = \mu + a_i + b_j + E_{ij} \tag{14.22}$$

for $i = 1$ to r and $j = 1$ to c. The parameters are the scalar μ and the vectors a_i and b_j. We assume that $\sum_{i=1}^{r} a_i = 0$ and $\sum_{j=1}^{c} b_j = 0$. So there are a total of $r + c - 1$ independent parameters.

The idea is that there are r experimental blocks, with the c members of each block to be as similar as possible. These members are subject to c different treatments. Untreated, the each block would have its own mean $\mu + a_i$, independent of the treatment. But the treatments may make a difference. The task is to estimate the effect of the blocks and the effect of the treatments. The fact that the populations are similar except for the effect of the treatments is reflected in the assumption that all the errors E_{ij} have the same variance σ^2.

The experimental numbers are the Y_{ij} for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. They may be summarized in the form of a data matrix. These matrices form a vector space of dimension rc. The subspace $\mathcal{L}$ consists of the matrices of the form $\mu + a_i + b_j$ as above. It has dimension $r + c - 1$. Thus the estimator of μ is the overall mean $\bar{Y}$. The estimator of a_i is $\bar{Y}_{i\cdot} - \bar{Y}$. The estimator of b_j is $\bar{Y}_{\cdot j} - \bar{Y}$. Thus the orthogonal projection of a data matrix with components Y_{ij} onto $\mathcal{L}$ is

$$\hat{Y}_{ij} = \bar{Y} + (\bar{Y}_{i\cdot} - \bar{Y}) + (\bar{Y}_{\cdot j} - \bar{Y}). \tag{14.23}$$

This can also be written in the equivalent form

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}. \tag{14.24}$$

Since the σ^2 is the same for all the populations, we pool them together for the purpose of estimating σ^2. The estimator of σ^2 is thus

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{r} \sum_{j=1}^{c} (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y})^2}{rc - (r + c - 1)}. \tag{14.25}$$

Notice that the denominator may also be written in the simple form $(r-1)(c-1)$.

14.5 Problems

1. Say that one is interested in finding the least squares estimates of the coefficients in the polynomial regression function $\hat{y} = a_0 + a_1 P_1(x) + a_2 P_2(x) + \cdots + a_r P_r(x)$. The polynomial $P_j(x)$ is of degree j. However, furthermore assume that the polynomials are such that for the points $x_1, \ldots, x_n$ the values of the polynomials form r orthogonal vectors. Show that the estimators $\hat{a}_j$ have a particularly simple form.

2. Prove that in two-way analysis of variance the estimator $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Chapter 15

Linear models: Hypothesis testing

15.1 Hypothesis testing

We consider a very general situation. The model is random variables $Y_1, \ldots, Y_n$. These are independent normal random variables, each with variance σ^2. The covariance is thus $\mathrm{Cov}(Y_i, Y_j) = \sigma^2 I_{ij}$. That is, the covariance matrix is σ^2 times the identity matrix I.

The mean vector μ in $\mathcal{R}(n)$ is unknown. Also σ^2 is unknown. However the mean μ is known to be in a subspace $\mathcal{L}$ of dimension k.

In the hypothesis testing context the null hypothesis is that the unknown mean vector μ is in a yet smaller subspace $\mathcal{L}_0$ of dimension $\ell < k$. The alternative hypothesis is that the unknown mean is not in this subspace. The problem is to use the observation vector Y in $\mathcal{R}(n)$ to make an appropriate decision in favor of the null hypothesis or the alternative hypothesis.

The projection $\hat{Y}$ of Y onto $\mathcal{L}$ is the estimate of the vector μ in $\mathcal{L}$. If the null hypothesis were known to be true, then it would be appropriate to use the projection $\hat{\hat{Y}}$ of Y onto $\mathcal{L}_0$ as the estimate of the vector μ in $\mathcal{L}_0$. Now consider the sum of squares identity

$$|Y - \hat{\hat{Y}}|^2 = |Y - \hat{Y}|^2 + |\hat{Y} - \hat{\hat{Y}}|^2. \tag{15.1}$$

The two terms on the right are independent. The second term on the right is a obvious indicator of a possible failure of the null hypothesis. The first term on the right enters into an estimate of the variance that is appropriate under either hypothesis. Under the null hypothesis, the second term on the right is σ^2 times a $\chi^2_{k-\ell}$ random variable. The first term on the right is σ^2 times a χ^2_{n-k} random variable. Thus under the null hypothesis

$$M = \frac{|\hat{Y} - \hat{\hat{Y}}|^2}{k - \ell} \tag{15.2}$$

estimates σ^2. In any case,

$$\hat{\sigma}^2 = \frac{|Y - \hat{Y}|^2}{n - k} \tag{15.3}$$

also estimates σ^2. Thus under the null hypothesis

$$F = \frac{M}{\hat{\sigma}^2} \tag{15.4}$$

estimates 1. Of course, under the alternative hypothesis M will tend to be a lot larger than σ^2, and F will be much larger than one. This is the basis for the test.

15.2 Chi-squared and F

Recall that a chi-squared random variable is one of the form

$$\chi^2_n = Z_1^2 + \cdots + Z_n^2, \tag{15.5}$$

where the Z_i are independent standard normal random variables. A χ^2_n random variable has mean n and variance $2n$. Its standard deviation is thus $\sqrt{2n}$. Therefore a χ^2_n/n random variable has mean 1 and standard deviation $\sqrt{2/n}$. If $n = 1$ this is Z^2, and we know that 95 percent of the probability is in the interval from 0 to 4. On the other hand, if n is reasonably large, say $n = 32$, then the central limit theory is a fair approximation. The standard deviation is $1/4$, and we would expect approximately 95 percent of the probability to be in the interval from 0 to $3/2$.

The F statistic is a ratio

$$F = \frac{\chi^2_{k-\ell}/(k-\ell)}{\chi^2_{n-k}/(n-k)}. \tag{15.6}$$

It is easy to guess what a cutoff for F at the five percent level might be. We reason on the basis of the null hypothesis. If $n - k$ is reasonably large, say 20 or so, then the denominator will be almost constant and fairly close to one. Its effect will be to make F tend to be a bit larger. So we then mainly have to worry about the numerator. If $k - \ell$ is one, then this is just the square of a normal random variable, so the cutoff for F is somewhat larger than $2^2 = 4$. On the other hand, as $k - \ell$ itself gets larger, the numerator also is more nearly constant. If $k - \ell = 32$, then the cutoff should be somewhat larger than $3/2$. Of course these are only rough guesses of the F distribution, so it is best to consult the tables. But if you have an even moderately large sample and an F value of 10, you would want to reject the null hypothesis.

15.3 Regression

The regression model is

$$Y_i = \sum_{j=1}^{k} X_{ij}\beta_j + E_i. \tag{15.7}$$

Here the assumption is that the errors E_i have mean zero and variance σ^2 and are independent. The space $\mathcal{L}$ is the span of the k columns of X.

The null hypothesis is that only the first ℓ of the parameters β_j are non-zero. Thus $\beta_j = 0$ for $j = \ell + 1, \ldots, k$. The space $\mathcal{L}_0$ is thus the span of the first ℓ columns of X.

Let $\hat{Y}$ and $\hat{\hat{Y}}$ be the projections of the data vector Y onto $\mathcal{L}$ and onto $\mathcal{L}_0$. Then the test uses the F statistic based on the corresponding sums of squares.

Example: The simplest case is when $k = 2$ and the model is

$$Y_i = \alpha + \beta(x_i - \bar{x}) + E_i. \tag{15.8}$$

The matrix X has two columns. The first column consists of ones. The second column consists of the $x_i - \bar{x}$ values. With this choice the two columns are orthogonal. Thus the regression line is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x}). \tag{15.9}$$

Thus $\hat{\alpha} = \bar{Y}$ and

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{15.10}$$

The estimator of variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}. \tag{15.11}$$

Under the null hypothesis $\beta = 0$. Then the regression line is simply

$$\hat{\hat{Y}}_i = \hat{\alpha} = \bar{Y}. \tag{15.12}$$

The difference is

$$\hat{Y} - \hat{\hat{Y}} = \hat{\beta}(x_i - \bar{x}). \tag{15.13}$$

The numerator is thus

$$M = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2. \tag{15.14}$$

So the test statistic is

$$F = \frac{M}{\hat{\sigma}^2} = \frac{\hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}^2}. \tag{15.15}$$

15.4 Analysis of variance: one way

The model is

$$Y_{ij} = \mu_j + E_{ij} \tag{15.16}$$

for $j = 1$ to c and $i = 1$ to n_j. The parameters are the μ_j. So there are a total of c parameters.

The idea is that there are c populations, corresponding to c different treatments. Untreated, the populations would have the same mean. But the treatments may make a difference. The task is to see whether this is so. Thus the null hypothesis is that all the means μ_j are the same.

We take a sample of size n_j from the jth population. These are the Y_{ij} for $i = 1, \ldots, n_j$. All these numbers together form a vector space of dimension $n = n_1 + \cdots + n_c$. The subspace $\mathcal{L}$ consists of the vectors that do not depend on the i index. This is a subspace of dimension c. Thus the orthogonal projection onto $\mathcal{L}$ is obtained by averaging over the i index. The estimator of μ_j is the sample mean of the sample from population j. This sample mean is denoted $\bar{Y}_{\cdot j}$. Thus

$$\hat{\mu}_j = \bar{Y}_{\cdot j}. \tag{15.17}$$

Since the σ^2 is the same for all the populations, we pool them together for the purpose of estimating σ^2. The estimator of σ^2 is thus

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - Y_{\cdot j})^2}{n - c}. \tag{15.18}$$

Under the null hypothesis the mean is in the subspace $\mathcal{L}_0$ of vectors that have constant entries. The estimator $\hat{\hat{Y}} = \bar{Y}$, the overall sample mean. This is a projection onto a one dimensional subspace. So the numerator in the F test is

$$M = \frac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{\cdot j} - \bar{Y})^2}{c - 1}. \tag{15.19}$$

As usual, the test statistic is $F = M/\hat{\sigma}^2$.

15.5 Analysis of variance: two way

The model is

$$Y_{ij} = \mu + a_i + b_j + E_{ij} \tag{15.20}$$

for $i = 1$ to r and $j = 1$ to c. The parameters are the scalar μ and the vectors a_i and b_j. We assume that $\sum_{i=1}^{r} a_i = 0$ and $\sum_{j=1}^{c} = 0$. So there are a total of $r + c - 1$ independent parameters.

The idea is that there are r experimental blocks, with the c members of each block to be as similar as possible. These members are subject to c different treatments. Untreated, the each block would have its own mean $\mu + a_i$, independent of the treatment. But the treatments may make a difference. The task is to see whether this is so.

The experimental numbers are the Y_{ij} for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. They may be summarized in the form of a data matrix. These matrices form a vector space of dimension rc. The subspace $\mathcal{L}$ consists of the matrices of the form $\mu + a_i + b_j$ as above. It has dimension $r + c - 1$. The orthogonal projection of a data matrix with components Y_{ij} onto $\mathcal{L}$ is

$$\hat{Y}_{ij} = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}. \tag{15.21}$$

Since the σ^2 is the same for all the populations, we pool them together for the purpose of estimating σ^2. The estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - Y_{i\cdot} - Y_{\cdot j} + \bar{Y})^2}{(r-1)(c-1)}. \tag{15.22}$$

Under the null hypothesis the treatments make no difference. Thus the $b_j = 0$. The subspace $\mathcal{L}_0$ consists of all matrices that do not depend on the j index. This is a subspace of dimension r. The projection onto $\mathcal{L}_0$ of Y_{ij} is

$$\hat{\hat{Y}}_{ij} = \bar{Y}_{i\cdot}. \tag{15.23}$$

The estimator in the numerator of the F test is

$$M = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\hat{Y}_{ij} - \hat{\hat{Y}}_{ij})^2}{(r+c-1)-r} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{\cdot j} - \bar{Y})^2}{c-1}. \tag{15.24}$$

The test statistic is $F = M/\hat{\sigma}^2$.

15.6 One way versus two way

In the two way analysis of variance the model is that $\mu_j = \mu + a_i + b_j$, where the a_i and the b_j each sum to zero. Each observation has variance σ^2. The estimate of the block effect a_i is $\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$. These estimates range over a space of dimension $r-1$. The estimate of the treatment effect b_j is $\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$. These estimates range over a space of dimension $c-1$. The estimate of the mean (the fit) is $\bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$. The estimate of the each error (the residual) is thus $Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}$. These estimates range over a space of dimension $(r-1)(c-1)$.

In the two way analysis of variance the null hypothesis is that the $b_j = 0$. Under this hypothesis the estimate of the mean is $\bar{Y}_{i\cdot}$. The test is based on the difference of the two estimates of the mean, that is, on the treatment effect $\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$.

The numerator in the F test is thus

$$M_2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2}{c-1}. \tag{15.25}$$

The denominator is

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2}{(r-1)(c-1)}. \tag{15.26}$$

Compare all this with the one way analysis of variance with c treatment groups, each of the same size r. The model is that $\mu_j = \mu + b_j$, where the b_j sum to zero. Each observation has variance σ^2. The estimate of the treatment effect b_j is $\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$. These estimates range over a space of dimension $c-1$. The estimate of the mean (the fit) is $\bar{Y}_{\cdot j}$. The estimate of the each error (the

residual) is thus $Y_{ij} - \bar{Y}_{\cdot j}$. These estimates range over a space of dimension $rc - c = (r-1)c$.

In the one way analysis of variance the null hypothesis is that the $b_j = 0$. Under this hypothesis the estimate of the mean is $\bar{Y}_{\cdot\cdot}$. The test is based on the difference of the two estimates of the mean, that is, on the treatment effect $\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$.

The numerator in the F test is thus

$$M_1 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2}{c-1}. \tag{15.27}$$

Notice that $M_1 = M_2$. The same statistic for the numerator is used in both models. The denominator is

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - \bar{Y}_{\cdot j})^2}{(r-1)c}. \tag{15.28}$$

The relation between these two denominators is clarified by looking at a sum of squares identity. The error in the one way model $Y_{ij} - \bar{Y}_{\cdot j}$ is the sum of the error in the two way model $Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}$ with the block effect in the two way model $Y_{i\cdot} - \bar{Y}_{\cdot\cdot}$. Furthermore, these two vectors are orthogonal. So the sums of squares are related by

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{ij} - \bar{Y}_{\cdot j})^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{c}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2. \tag{15.29}$$

If the true model is the two way model, with the block effects $a_i \neq 0$, then it is clearly wrong to use the one way model. It would be using the wrong estimate of the error, by confusing the block effects with the random error.

On the other hand, if the true model is the one way model, then the harm that is done by using the two way model is more subtle. Even if the estimate of variance is different, it is still an unbiased estimate of σ^2. The only difference is that there are $(r-1)(c-1)$ independent summands that contribute to the estimate, instead of the $(r-1)c$ in the other analysis. This is comparable to losing one column of data in the estimate of the variance. The effect of this on the test is explored in the problems.

15.7 Problems

1. Consider a regression model $Y_i = \alpha + \beta x + \gamma x^2$. Describe the hypothesis test where the null hypothesis is $\gamma = 0$.

2. Say that a statistician had a situation where the appropriate model was the one-way analysis of variance. However all $n_j = r$, and the rather confused statistician instead used the test for the two-way analysis of variance. How do the two tests differ in this situation? Has the statistician made a major blunder? Discuss. In particular, compare the power of the two tests.

Appendix A

Linear algebra review

A.1 Vector spaces

Let $\mathcal{M}(n,k)$ be the space of all n by k matrices. Thus such a matrix has n rows and k columns. If A is in $\mathcal{M}(n,k)$, then its entries A_{ij} are defined for $i = 1,\ldots,n$ and $j = 1,\ldots,k$ and are real numbers.

Matrices in $\mathcal{M}(n,k)$ may be combined by the *vector space* operations. These are addition of matrices and multiplication of a matrix by a scalar (a real number). Thus if A and B each belong to $M(n,k)$, and if c is a scalar, then so do $A+B$ and cA. They are defined by

$$(A+B)_{ij} = A_{ij} + B_{ij} \tag{A.1}$$

and

$$(cA)_{ij} = cA_{ij}. \tag{A.2}$$

In particular, we may repeat these operations and form a linear combination like $cA+dB$. There are even more general linear combinations. We also define $-B$ to be $(-1)B$ and $A-B = A+(-1)B$. Thus negation and subtraction are also defined by the vector space operations.

Among the matrices in $A(m,n)$ there is a special matrix, the zero matrix. This will be denoted 0.

Consider a subset $\mathcal{L}$ of $\mathcal{M}(n,k)$ that contains 0. It is called a *subspace* if whenever A and B are in the subset, then also $A+B$ and cA are in the subset. It follows that if matrices belong to a subspace, then linear combinations of these matrices belong to the subspace.

There is an important special case of matrices: column vectors. We let $\mathcal{R}(n)$ be the space $\mathcal{M}(n,1)$ of n by 1 matrices. Sometimes we will think of a column vector in $\mathcal{R}(n)$ as a vector of data values or as a vector of predicted data values. On other occasions we will think of a column vector in $\mathcal{R}(k)$ as a vector of parameter values.

Suppose that A is a matrix in $\mathcal{M}(n,k)$. Then the set of solutions b to $Ab = 0$ is a subspace of $\mathcal{R}(k)$, called the *null space* of A. Also, the set of all vectors $z = Ab$ is a subspace of $\mathcal{R}(n)$, called the *range* (or column space) of A.

A.2 Matrix multiplication

If A is a n by k matrix and B is a k by m matrix, then AB is a n by m matrix defined by

$$(AB)_{ij} = \sum_{p=1}^{k} A_{ip} B_{pj}. \tag{A.3}$$

Notices that the matrices do not have to have the same size, but when we multiply a matrix A in $\mathcal{M}(n,k)$ by a matrix B in $\mathcal{M}(k,m)$, it is important that the number of columns of A is equal to the number of rows of B.

For each n there is a square matrix in $\mathcal{M}(n,n)$ called the *identity matrix* and denoted by I. It has the values $I_{ij} =$ for $i = j$ and $I_{ij} = 0$ for $i \neq j$. If A is in $\mathcal{M}(n,k)$, then we have the identities $IA = A$ and $AI = A$, where the I matrix on the left is n by n, while the I matrix on the right is k by k.

Suppose that A is a matrix in $\mathcal{M}(n,k)$ and b is a column vector in $\mathcal{R}(k)$. Then $\hat{y} = Ab$ is a column vector in $\mathcal{R}(n)$. This can be thought of as a transformation that takes the parameter vector b and produces a predicted data vector $\hat{y}$.

Say that A is a square matrix in $\mathcal{M}(k,k)$. If there is another matrix A^{-1} with $AA^{-1} = A^{-1}A = I$, where I is the k by k identity matrix, then A is said to be invertible, and A^{-1} is its *inverse*. If B is another invertible square matrix in $\mathcal{M}(k,k)$, then $(AB)^{-1} = B^{-1}A^{-1}$. Note the reversal in order. Also note that $(A^{-1})^{-1} = A$.

If A is a square matrix in $\mathcal{M}(k,k)$, then the trace of A is the number

$$\mathrm{tr}(A) = \sum_{i=1}^{k} A_{ii}. \tag{A.4}$$

That is, the trace of A is the sum of the diagonal entries of A. Consider square matrices A, B in $\mathcal{M}(k,k)$. While in general it is not true that $AB = BA$, it is always true that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

A.3 The transpose

If A is a matrix in $\mathcal{M}(n,k)$, then its *transpose* is a matrix A' in $\mathcal{M}(k,n)$ defined by

$$A'_{ij} = A_{ji}. \tag{A.5}$$

If A is in $\mathcal{M}(n,k)$ and B is in $\mathcal{M}(k,m)$, then we have seen that the matrix product AB is in $\mathcal{M}(n,m)$. It is not hard to see that $(AB)'$ is in $\mathcal{M}(m,n)$ and

$(AB)' = B'A'$. Note the reversal in order. Furthermore, it is always true that $A'' = A$.

The matrix $A'A$ is a square matrix that is symmetric, that is, equal to its transpose. It is not difficult to show that A has trivial null space precisely in the case when $A'A$ is invertible. [Proof: Suppose $A'A$ is invertible. If $Ab = 0$, then $A'Ab = 0$, and so $b = 0$. Thus A has trivial null space. For the converse, suppose that A has trivial null space. If $A'Ab = 0$, then $b'A'Ab = 0$, and so $(Ab)'(Ab) = 0$, which implies that $Ab = 0$. It follows that $A'A$ has trivial null space. Since $A'A$ is square, this implies $A'A$ is invertible.]

There are also some useful facts for square matrices. We always have $\mathrm{tr}(A) = \mathrm{tr}(A')$. For an invertible matrix we have $(A')^{-1} = (A^{-1})'$.

Once we have the notion of transpose, we have the important notion of *inner product*. This is also called the *scalar product* or *dot product*. If A is in $\mathcal{M}(n,k)$ and B is in $\mathcal{M}(n,k)$, then the inner product $A \cdot B$ is defined by

$$A \cdot B = \mathrm{tr}(A'B). \tag{A.6}$$

Notice that $A \cdot B = B \cdot A$, since $\mathrm{tr}(A'B) = \mathrm{tr}((A'B)') = \mathrm{tr}(B'A'') = \mathrm{tr}(B'A)$. It is also useful to have the formula for $A \cdot B$ in terms of the matrix entries:

$$A \cdot B = \sum_{i=1}^{k} \sum_{j=1}^{n} A_{ji} B_{ji}. \tag{A.7}$$

If A is a matrix in $\mathcal{M}(n,k)$, then we define its *norm* (or Euclidean *length* to be

$$\|A\| = \sqrt{A \cdot A} = \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{n} A_{ji}^2}. \tag{A.8}$$

(Note: There are other notions of norm for matrices, but this is the natural notion of norm in the context of this inner product.)

If a and b are column vectors in $\mathcal{R}(n)$, then again we have the notion of inner product. However then we may write $a \cdot b = a'b$, the matrix product of a row on the left with a column on the right. Since this gives a 1 by 1 matrix, we may think of it as a number, and we do not have to bother to do a sum to compute the trace. The formula in terms of the entries of the vectors is:

$$a \cdot b = \sum_{j=1}^{n} a_j b_j. \tag{A.9}$$

The norm is

$$\|a\| = \sqrt{a \cdot a} = \sqrt{\sum_{j=1}^{n} a_{ji}^2}. \tag{A.10}$$

A.4 The theorem of Pythagoras

In this section we shall deal with column vectors in $\mathcal{R}(n)$. However all the same ideas work for matrices in $\mathcal{M}(n,k)$, since everything is defined in terms of the vector space operations and the inner product.

We say that vectors a, b are *orthogonal* or *perpendicular* if $a \cdot b = 0$.

Theorem A.1 *Theorem of Pythagoras. If $a \cdot b = 0$, then*

$$\|a+b\|^2 = \|a\|^2 + \|b\|^2. \tag{A.11}$$

Proof: This is easy:

$$\|a+b\|^2 = (a+b)\cdot(a+b) = a\cdot a + 2a\cdot b + b\cdot b = \|a\|^2 + 0 + \|b\|^2. \tag{A.12}$$

In statistics the theorem of Pythagoras is called a sum of squares identity. In components it says that if

$$\sum_{i=1}^{n} a_i b_i = 0. \tag{A.13}$$

then

$$\sum_{i=1}^{n} (a_i + b_i)^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} b_i^2. \tag{A.14}$$

Whenever you have orthogonality, then you have a sum of squares identity.

A.5 The projection theorem

We treat the case of column vectors in $\mathcal{R}(n)$. Again everything would work the same for the space of matrices $\mathcal{M}(n,k)$. The following result is fundamental. It is called the *projection theorem*.

Theorem A.2 *Let $\mathcal{L}$ be a subspace of $\mathcal{R}(n)$. Let y be a vector in $\mathcal{R}(n)$. Then there exists a unique vector $\hat{y}$ in $\mathcal{L}$ such that $y - \hat{y}$ is orthogonal to every vector in $\mathcal{L}$.*

The vector $\hat{y}$ described in the theorem is the *orthogonal projection* of y onto $\mathcal{L}$. Since $y = \hat{y} + (y - \hat{y})$, whenever we have an orthogonal projection we also have a sum of squares identity.

Theorem A.3 *Let $\mathcal{L}$ be a subspace of $\mathcal{R}(n)$. Let y be a vector in $\mathcal{R}(n)$. Let $\hat{y}$ be the orthogonal projection of y onto $\mathcal{L}$. Then*

$$\|y\|^2 = \|\hat{y}\|^2 + \|y - \hat{y}\|^2. \tag{A.15}$$

Example: Let $\mathcal{L}$ consist of the constant vectors in $\mathcal{R}(n)$. Then the orthogonal projection of y onto $\mathcal{R}(n)$ is the vector such that each entry is the sample mean $\bar{y}$. The corresponding sum of squares identity is

$$\sum_{i=1}^{n} y_i^2 = n(\bar{y})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{A.16}$$

Furthermore, we have the following characterization of the orthogonal projection of y onto $\mathcal{L}$ as the vector in $\mathcal{L}$ that is closest to y. For this reason, the orthogonal projection is also called the *least squares* vector.

Theorem A.4 *Let $\mathcal{L}$ be a subspace of $\mathcal{R}(n)$. Let y be a vector in $\mathcal{R}(n)$. Let $\hat{y}$ be the orthogonal projection of y onto $\mathcal{L}$. Let z be another vector in $\mathcal{L}$. Then*

$$\|y - \hat{y}\|^2 \leq \|y - z\|^2. \tag{A.17}$$

There is an equality only when $z = \hat{y}$.

In components this inequality says that $\hat{y}$ is in the subspace and

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \leq \sum_{i=1}^{n} (y_i - z_i)^2. \tag{A.18}$$

for all z in the subspace. Thus the sum of squares is least when $z = \hat{y}$.

A very important special case consists of the situation when $\mathcal{L}$ consists of all vector of the form $z = Ab$, where A is a fixed matrix in $\mathcal{M}(n, k)$ and b varies over $\mathcal{R}(k)$. Thus $\mathcal{L}$ is the range of A. It is also convenient to assume that the null space of A is trivial. Then the b vectors parameterize $\mathcal{L}$, in the sense that A sets up a one-to-one correspondence between the b vectors in $\mathcal{R}(k)$ and the subspace $\mathcal{L}$ that is the range of A.

Theorem A.5 *Fix A in $\mathcal{M}(n, k)$ with trivial null space. Let $\mathcal{L}$ be the range of A. Let y be a vector in $\mathcal{R}(n)$. Then the orthogonal projection of y onto $\mathcal{L}$ satisfies the equation $\hat{y} = A\hat{b}$, where $\hat{b}$ is a solution of*

$$(A'A)\hat{b} = A'y. \tag{A.19}$$

This equation may be solved in the form

$$\hat{b} = (A'A)^{-1}A'y. \tag{A.20}$$

Thus

$$\hat{y} = A(A'A)^{-1}A'y. \tag{A.21}$$

Proof: The condition that $\hat{y} = A\hat{b}$ is the orthogonal projection is that $y - \hat{y} = y - A\hat{b}$ is perpendicular to every vector Ab in $\mathcal{L}$. This says that

$$Ab \cdot (y - A\hat{b}) = 0 \tag{A.22}$$

for all b in $\mathcal{R}(k)$. This equation may also be written as

$$(Ab)'(y - A\hat{b}) = 0 \tag{A.23}$$

or

$$bA'(y - A\hat{b}) = 0 \tag{A.24}$$

or

$$b(A'y - A'A\hat{b}) = 0. \tag{A.25}$$

However this is true for all b precisely when $A'y - AA\hat{b} = 0$.

The only bad thing about this theorem is that one has to invert the matrix $A'A$. The most convenient case is when A has orthogonal columns, which is the same as saying that $A'A$ is a diagonal matrix. Then the inversion is easy.

A.6 Problems

1. Consider the matrix

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{bmatrix}. \tag{A.26}$$

   The column space $\mathcal{L}$ of X is a two-dimensional subspace of $\mathcal{R}^6$. Consider the vector

$$y = \begin{bmatrix} 11 \\ -9 \\ -8 \\ -7 \\ -6 \\ -5 \end{bmatrix}. \tag{A.27}$$

   Find the matrix $X'X$ and its inverse. Find the projection $\hat{y} = Xb$ onto the column space of X. Find the parameter vector b in $\mathcal{R}^2$.

2. In the preceding problem, verify the theorem of Pythagoras for $\hat{y}$ and $y - \hat{y}$.

3. Consider the matrix

$$Z = \begin{bmatrix} 1 & -5 \\ 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{bmatrix}. \tag{A.28}$$

   Show that the column space of Z is the same $\mathcal{L}$. However Z has orthogonal columns. Consider the same vector y. Find the matrix $Z'Z$ and its inverse. Find the projection $\hat{y} = Zc$ onto the column space of Z. Find the parameter vector c in $\mathcal{R}^2$.