choose to fit 10 given data pairs with a polynomial of degree 9. The resulting polynomial model will provide a perfect fit to the data, but will nevertheless be incorrect. As a rule of thumb, there should be at least five times (preferably ten times) more data points than there are parameters to be estimated.

(e) **Causality.** The discovery of a linear relation between two variables x and y should not be mistaken for a discovery of a causal relation. A tight fit may be due to the fact that variable x has a causal effect on y, but may be equally due to a causal effect of y on x. Alternatively, there may be some external effect, described by yet another variable z, that affects both x and y in similar ways. For a concrete example let x_i be the wealth of the first born child and y_i be the wealth of the second born child in the same family. We expect y_i to increase roughly linearly with x_i, but this can be traced on the effect of a common family and background rather than a causal effect of one child on the other.

9.3 BINARY HYPOTHESIS TESTING

In this section, we revisit the problem of choosing between two hypotheses, but unlike the Bayesian formulation of Section 8.2, we will assume no prior probabilities. We may view this as an inference problem where the parameter θ takes just two values, but consistent with historical usage, we will forgo the θ-notation and denote the two hypotheses as H_0 and H_1. In traditional statistical language, hypothesis H_0 is often called the **null** hypothesis and H_1 the **alternative** hypothesis. This indicates that H_0 plays the role of a default model, to be proved or disproved on the basis of available data.

The available observation is a vector $X = (X_1, \ldots, X_n)$ of random variables whose distribution depends on the hypothesis. We will use the notation $\mathbf{P}(X \in A; H_j)$ to denote the probability that the observation X belongs to a set A when hypothesis H_j is true. Note that consistent with the classical inference framework, these are not conditional probabilities, because the true hypothesis is not treated as a random variable. Similarly, we will use notation such as $p_X(x; H_j)$ or $f_X(x; H_j)$ to denote the PMF or PDF, respectively, of the vector X, under hypothesis H_j. We want to find a decision rule that maps the realized values x of the observation to one of the two hypotheses (see Fig. 9.7).

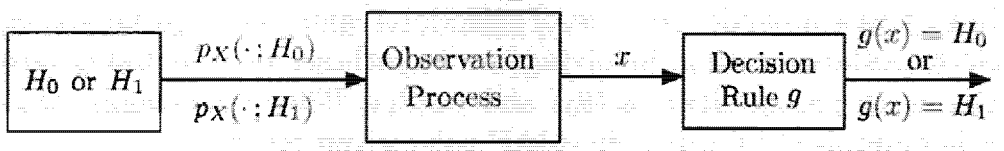

Figure 9.7: Classical inference framework for binary hypothesis testing.

Any decision rule can be represented by a partition of the set of all possible values of the observation vector $X = (X_1, \ldots, X_n)$ into two subsets: a set R, called the **rejection region**, and its complement, R^c, called the **acceptance region**. Hypothesis H_0 is **rejected** (declared to be false) when the observed data $X = (X_1, \ldots, X_n)$ happen to fall in the rejection region R and is **accepted** (declared to be true) otherwise; see Fig. 9.8. Thus, the choice of a decision rule is equivalent to choosing the rejection region.

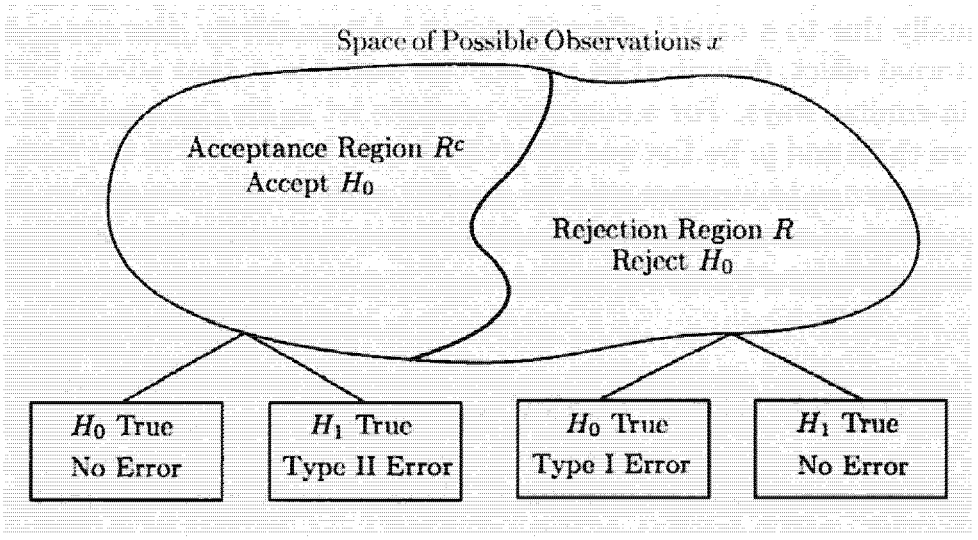

Figure 9.8: Structure of a decision rule for binary hypothesis testing. It is specified by a partition of the set of all possible observations into a set R and its complement R^c. The null hypothesis is rejected if the realized value of the observation falls in the rejection region.

For a particular choice of the rejection region R, there are two possible types of errors:

(a) Reject H_0 even though H_0 is true. This is called a **Type I** error, or a **false rejection**, and happens with probability

$$\alpha(R) = \mathbf{P}(X \in R; H_0).$$

(b) Accept H_0 even though H_0 is false. This is called a **Type II** error, or a **false acceptance**, and happens with probability

$$\beta(R) = \mathbf{P}(X \notin R; H_1).$$

To motivate a particular form of rejection region, we draw an analogy with Bayesian hypothesis testing, involving two hypotheses $\Theta = \theta_0$ and $\Theta = \theta_1$, with respective prior probabilities $p_\Theta(\theta_0)$ and $p_\Theta(\theta_1)$. Then, the overall probability of error is minimized by using the MAP rule: given the observed value x of X, declare $\Theta = \theta_1$ to be true if

$$p_\Theta(\theta_0)p_{X|\Theta}(x \mid \theta_0) < p_\Theta(\theta_1)p_{X|\Theta}(x \mid \theta_1)$$

(assuming that X is discrete).† This decision rule can be rewritten as follows: define the **likelihood ratio** $L(x)$ by

$$L(x) = \frac{p_{X|\Theta}(x \mid \theta_1)}{p_{X|\Theta}(x \mid \theta_0)},$$

and declare $\Theta = \theta_1$ to be true if the realized value x of the observation vector X satisfies

$$L(x) > \xi,$$

where the **critical value** ξ is

$$\xi = \frac{p_\Theta(\theta_0)}{p_\Theta(\theta_1)}.$$

If X is continuous, the approach is the same, except that the likelihood ratio is defined as a ratio of PDFs:

$$L(x) = \frac{f_{X|\Theta}(x \mid \theta_1)}{f_{X|\Theta}(x \mid \theta_0)}.$$

Motivated by the preceding form of the MAP rule, we are led to consider rejection regions of the form

$$R = \{x \mid L(x) > \xi\},$$

where the likelihood ratio $L(x)$ is defined similar to the Bayesian case:†

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)}, \qquad \text{or} \qquad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)}.$$

The critical value ξ remains free to be chosen on the basis of other considerations. The special case where $\xi = 1$ corresponds to the ML rule.

Example 9.10. We have a six-sided die that we want to test for fairness, and we formulate two hypotheses for the probabilities of the six faces:

$$H_0 \text{ (fair die):} \qquad p_X(x; H_0) = \frac{1}{6}, \qquad x = 1, \ldots, 6,$$

$$H_1 \text{ (loaded die):} \qquad p_X(x; H_1) = \begin{cases} \frac{1}{4}, & \text{if } x = 1, 2, \\ \frac{1}{8}, & \text{if } x = 3, 4, 5, 6. \end{cases}$$

† In this paragraph, we use conditional probability notation since we are dealing with a Bayesian framework.

† Note that we use $L(x)$ to denote the value of the likelihood ratio based on the observed value x of the random observation X. On the other hand, before the experiment is carried out, the likelihood ratio is best viewed as a random variable, a function of the observation X, in which case it is denoted by $L(X)$. The probability distribution of $L(X)$ depends, of course, on which hypothesis is true.

The likelihood ratio for a single roll x of the die is

$$L(x) = \begin{cases} \dfrac{1/4}{1/6} = \dfrac{3}{2}, & \text{if } x = 1, 2, \\ \dfrac{1/8}{1/6} = \dfrac{3}{4}, & \text{if } x = 3, 4, 5, 6. \end{cases}$$

Since the likelihood ratio takes only two distinct values, there are three possibilities to consider for the critical value ξ, with three corresponding rejection regions:

$$\begin{aligned} \xi < \frac{3}{4}: &\quad \text{reject } H_0 \text{ for all } x; \\ \frac{3}{4} < \xi < \frac{3}{2}: &\quad \text{accept } H_0 \text{ if } x = 3, 4, 5, 6; \text{ reject } H_0 \text{ if } x = 1, 2; \\ \frac{3}{2} < \xi: &\quad \text{accept } H_0 \text{ for all } x. \end{aligned}$$

Intuitively, a roll of 1 or 2 provides evidence that favors H_1, and we tend to reject H_0. On the other hand, if we set the critical value too high ($\xi > 3/2$), we never reject H_0. In fact, for a single roll of the die, the test makes sense only in the case $3/4 < \xi < 3/2$, since for other values of ξ, the decision does not depend on the observation.

The error probabilities can be calculated from the problem data for each critical value. In particular, the probability of false rejection $\mathbf{P}(\text{Reject } H_0 : H_0)$ is

$$\alpha(\xi) = \begin{cases} 1, & \text{if } \xi < \dfrac{3}{4}, \\ \mathbf{P}(X = 1, 2; H_0) = \dfrac{1}{3}, & \text{if } \dfrac{3}{4} < \xi < \dfrac{3}{2}, \\ 0, & \text{if } \dfrac{3}{2} < \xi. \end{cases}$$

and the probability of false acceptance $\mathbf{P}(\text{Accept } H_0 : H_1)$ is

$$\beta(\xi) = \begin{cases} 0, & \text{if } \xi < \dfrac{3}{4}, \\ \mathbf{P}(X = 3, 4, 5, 6; H_1) = \dfrac{1}{2}, & \text{if } \dfrac{3}{4} < \xi < \dfrac{3}{2}, \\ 1, & \text{if } \dfrac{3}{2} < \xi. \end{cases}$$

Note that choosing ξ trades off the probabilities of the two types of errors, as illustrated by the preceding example. Indeed, as ξ increases, the rejection region becomes smaller. As a result, the false rejection probability $\alpha(R)$ decreases, while the false acceptance probability $\beta(R)$ increases (see Fig. 9.9). Because of this tradeoff, there is no single best way of choosing the critical value. The most popular approach is as follows.

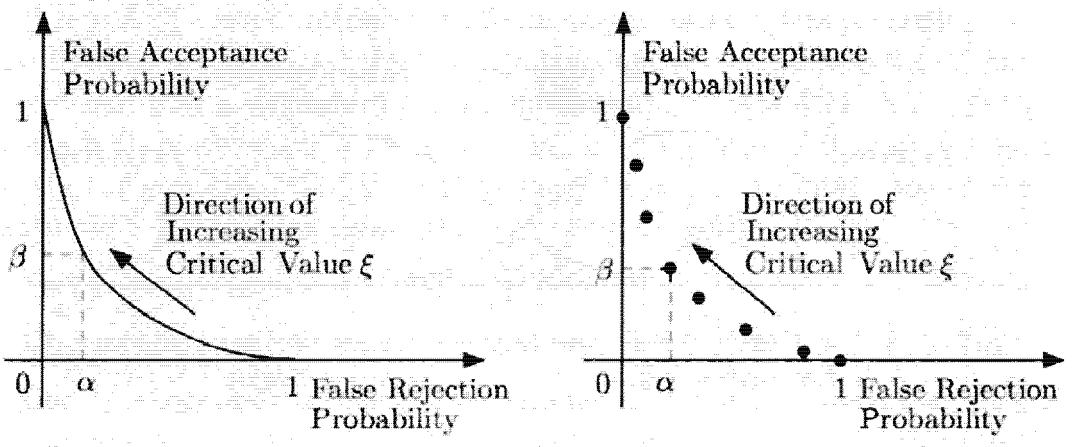

Figure 9.9: Error probabilities in a likelihood ratio test. As the critical value ξ increases, the rejection region becomes smaller. As a result, the false rejection probability α decreases. while the false acceptance probability β increases. When the dependence of α on ξ is continuous and strictly decreasing, there is a unique value of ξ that corresponds to a given α (see the figure on the left). However, the dependence of α on ξ may not be continuous. e.g., if the likelihood ratio $L(x)$ can only take finitely many different values (see the figure on the right).

Likelihood Ratio Test (LRT)

- Start with a target value α for the false rejection probability.
- Choose a value for ξ such that the false rejection probability is equal to α:

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha.$$

- Once the value x of X is observed, reject H_0 if $L(x) > \xi$.

Typical choices for α are $\alpha = 0.1$, $\alpha = 0.05$, or $\alpha = 0.01$, depending on the degree of undesirability of false rejection. Note that to be able to apply the LRT to a given problem, the following are required:

(a) We must be able to compute $L(x)$ for any given observation value x, so that we can compare it with the critical value ξ. Fortunately. this is always the case when the underlying PMFs or PDFs are given in closed form.

(b) We must either have a closed form expression for the distribution of $L(X)$ [or of a related random variable such as $\log L(X)$] or we must be able to approximate it analytically, computationally, or through simulation. This is needed to determine the critical value ξ that corresponds to a given false rejection probability α.

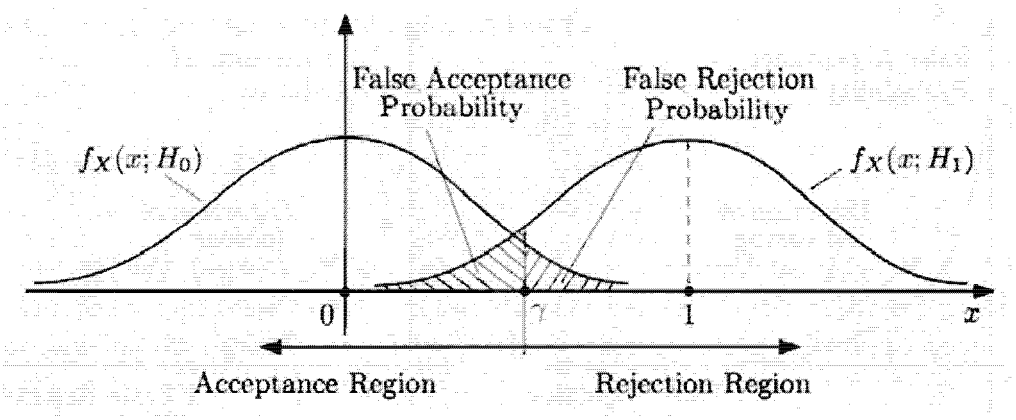

Figure 9.10: Rejection and acceptance regions in Example 9.11, and corresponding false rejection and false acceptance probabilities.

Example 9.11. A surveillance camera periodically checks a certain area and records a signal $X = W$ or $X = 1 + W$ depending on whether an intruder is absent or present (hypotheses H_0 or H_1, respectively). We assume that W is a normal random variable with mean 0 and known variance v. Since

$$f_X(x; H_0) = \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{x^2}{2v}\right\}, \qquad f_X(x; H_1) = \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{(x-1)^2}{2v}\right\},$$

the likelihood ratio is

$$L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} = \exp\left\{\frac{x^2 - (x-1)^2}{2v}\right\} = \exp\left\{\frac{2x-1}{2v}\right\}.$$

For a given critical value ξ, the LRT rejects H_0 if $L(x) > \xi$, or equivalently, after a straightforward calculation, if

$$x > v \log \xi + \frac{1}{2}.$$

Thus, the rejection region is of the form

$$R = \{x \mid x > \gamma\}$$

for some γ, which corresponds to ξ via the relation

$$\gamma = v \log \xi + \frac{1}{2};$$

see Fig. 9.10. We set a target value α for the false rejection probability, and we proceed to determine γ from the relation

$$\alpha = \mathbf{P}(X > \gamma; H_0) = \mathbf{P}(W > \gamma),$$

and the normal tables. For example, if $\alpha = 0.025$, then $\gamma = 1.96\sqrt{v}$. We may also calculate the false acceptance probability,

$$\beta = \mathbf{P}\big(X \le \gamma; H_1\big) = \mathbf{P}\left(1 + W \le \gamma\right) = \mathbf{P}\left(W \le \gamma - 1\right),$$

by using again the normal tables.

When $L(X)$ is a continuous random variable, as in the preceding example, the probability $\mathbf{P}\big(L(X) > \xi; H_0\big)$ moves continuously from 1 to 0 as ξ increases. Thus, we can find a value of ξ for which the requirement $\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha$ is satisfied. If, however, $L(X)$ is a discrete random variable, it may be impossible to satisfy the equality $\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha$ exactly, no matter how ξ is chosen; cf. Example 9.10. In such cases, there are several possibilities:

(a) Strive for approximate equality.

(b) Choose the smallest value of ξ that satisfies $\mathbf{P}\big(L(X) > \xi; H_0\big) \leq \alpha$.

(b) Use an exogenous source of randomness to choose between two alternative candidate critical values. This variant (known as a "randomized likelihood ratio test") is of some theoretical interest. However, it is not sufficiently important in practice to deserve further discussion in this book.

We have motivated so far the use of a LRT through an analogy with Bayesian inference. However, we will now provide a stronger justification: for a given false rejection probability, the LRT offers the smallest possible false acceptance probability.

Neyman-Pearson Lemma

Consider a particular choice of ξ in the LRT, which results in error probabilities

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha, \qquad \mathbf{P}\big(L(X) \leq \xi; H_1\big) = \beta.$$

Suppose that some other test, with rejection region R, achieves a smaller or equal false rejection probability:

$$\mathbf{P}(X \in R; H_0) \leq \alpha.$$

Then,

$$\mathbf{P}(X \notin R; H_1) \geq \beta,$$

with strict inequality $\mathbf{P}(X \notin R; H_1) > \beta$ when $\mathbf{P}(X \in R; H_0) < \alpha$.

For a justification of the Neyman-Pearson Lemma, consider a hypothetical Bayesian decision problem where the prior probabilities of H_0 and H_1 satisfy

$$\frac{p_\Theta(\theta_0)}{p_\Theta(\theta_1)} = \xi,$$

so that

$$p_\Theta(\theta_0) = \frac{\xi}{1+\xi}, \qquad p_\Theta(\theta_1) = \frac{1}{1+\xi}.$$

Then, the threshold used by the MAP rule is equal to ξ, as discussed in the beginning of this section, and the MAP rule is identical to the LRT rule. The

probability of error with the MAP rule is

$$e_{\text{MAP}} = \frac{\xi}{1+\xi}\alpha + \frac{1}{1+\xi}\beta,$$

and from Section 8.2, we know that it is smaller than or equal to the probability of error of any other Bayesian decision rule. This implies that for any choice of rejection region R, we have

$$e_{\text{MAP}} \leq \frac{\xi}{1+\xi}\mathbf{P}(X \in R; H_0) + \frac{1}{1+\xi}\mathbf{P}(X \notin R; H_1).$$

Comparing the preceding two relations, we see that if $\mathbf{P}(X \in R; H_0) \leq \alpha$, we must have $\mathbf{P}(X \notin R; H_1) \geq \beta$, and that if $\mathbf{P}(X \in R; H_0) < \alpha$, we must have $\mathbf{P}(X \notin R; H_1) > \beta$, which is the conclusion of the Neyman-Pearson Lemma.

The Neyman-Pearson Lemma can be interpreted geometrically as shown in Fig. 9.11. We illustrate the lemma with a few examples.

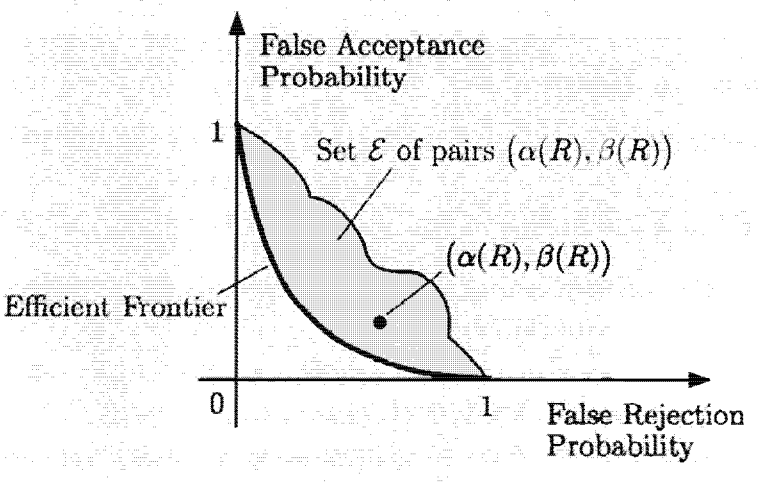

Figure 9.11: Interpretation of the Neyman-Pearson Lemma. Consider the set $\mathcal{E}$ of all error probability pairs $(\alpha(R), \beta(R))$, as R ranges over all possible rejection regions (subsets of the observation space). The **efficient frontier** of $\mathcal{E}$ is the set of all $(\alpha(R), \beta(R)) \in \mathcal{E}$ such that there is no $(\alpha, \beta) \in \mathcal{E}$ with $\alpha \leq \alpha(R)$ and $\beta < \beta(R)$, or $\alpha < \alpha(R)$ and $\beta \leq \beta(R)$. The Neyman-Pearson Lemma states that all pairs $(\alpha(\xi), \beta(\xi))$ corresponding to LRTs lie on the efficient frontier.

Example 9.12. Consider Example 9.10, where we roll a six-sided die once and test it for fairness. We consider the set $\mathcal{E}$ of all error probability pairs $(\alpha(R), \beta(R))$ as R ranges over all possible rejection regions (all subsets of the observation space $\{1, \ldots, 6\}$). The set $\mathcal{E}$ is shown in Fig. 9.12 and it can be seen that the error probability pairs $(1, 0)$, $(1/3, 1/2)$, and $(0, 1)$ associated with the LRTs have the property given by the Neyman-Pearson Lemma (i.e., lie on the efficient frontier, in the terminology of Fig. 9.11).

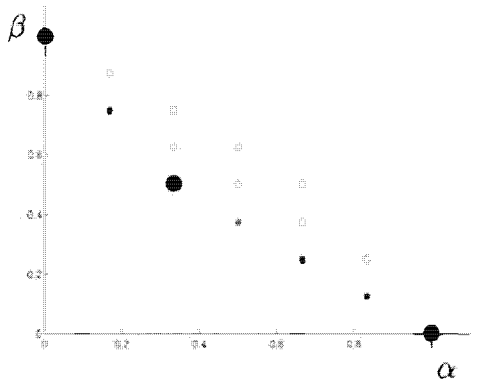

Figure 9.12: Set of pairs $\big(\alpha(R), \beta(R)\big)$ as the rejection region R ranges over all subsets of the observation space $\{1, \ldots, 6\}$ in Examples 9.10 and 9.12. The pairs $(1,0)$, $(0,1)$, and $(1/3, 1/2)$ are the ones that correspond to LRTs.

Example 9.13. Comparison of Different Rejection Regions. We observe two i.i.d. normal random variables X_1 and X_2, with unit variance. Under H_0 their common mean is 0; under H_1, their common mean is 2. We fix the false rejection probability to $\alpha = 0.05$.

We first derive the form of the LRT, and then calculate the resulting value of β. The likelihood ratio is of the form

$$L(x) = \frac{\frac{1}{\sqrt{2\pi}} \exp\Big\{ -\big((x_1-2)^2 + (x_2-2)^2\big)/2 \Big\}}{\frac{1}{\sqrt{2\pi}} \exp\Big\{ -(x_1^2 + x_2^2)/2 \Big\}} = \exp\Big\{ 2(x_1 + x_2) - 4 \Big\}.$$

Comparing $L(x)$ to a critical value ξ is equivalent to comparing $x_1 + x_2$ to $\gamma = (4 + \log \xi)/2$. Thus, under the LRT, we decide in favor of H_1 if $x_1 + x_2 > \gamma$, for some particular choice of γ. This determines the shape of the rejection region.

To determine the exact form of the rejection region, we need to find γ so that the false rejection probability $\mathbf{P}(X_1 + X_2 > \gamma; H_0)$ is equal to 0.05. We note that under H_0, $Z = (X_1 + X_2)/\sqrt{2}$ is a standard normal random variable. We have

$$0.05 = \mathbf{P}(X_1 + X_2 > \gamma; H_0) = \mathbf{P}\left(\frac{X_1 + X_2}{\sqrt{2}} > \frac{\gamma}{\sqrt{2}} ; H_0 \right) = \mathbf{P}\left(Z > \frac{\gamma}{\sqrt{2}} \right).$$

From the normal tables, we obtain $\mathbf{P}(Z > 1.645) = 0.05$, so we choose

$$\gamma = 1.645 \cdot \sqrt{2} = 2.33,$$

resulting in the rejection region

$$R = \big\{ (x_1, x_2) \,|\, x_1 + x_2 > 2.33 \big\}.$$

To evaluate the performance of this test, we calculate the resulting false acceptance probability. Note that under H_1, $X_1 + X_2$ is normal with mean equal to 4 and variance equal to 2, so that $Z = (X_1 + X_2 - 4)/\sqrt{2}$ is a standard normal random variable. Thus, using the normal tables, the false acceptance probability is

given by

$$\begin{aligned}\beta(R) &= \mathbf{P}(X_1 + X_2 \leq 2.33; H_1) \\ &= \mathbf{P}\left(\frac{X_1 + X_2 - 4}{\sqrt{2}} \leq \frac{2.33 - 4}{\sqrt{2}}; H_1\right) \\ &= \mathbf{P}(Z \leq -1.18) \\ &= \mathbf{P}(Z \geq 1.18) \\ &= 1 - \mathbf{P}(Z \leq 1.18) \\ &= 1 - 0.88 \\ &= 0.12.\end{aligned}$$

We now compare the performance of the LRT with that resulting from a different rejection region R'. For example, let us consider a rejection region of the form

$$R' = \{(x_1, x_2) \mid \max\{x_1, x_2\} > \zeta\},$$

where ζ is chosen so that the false rejection probability is again 0.05. To determine the value of ζ, we write

$$\begin{aligned}0.05 &= \mathbf{P}\big(\max\{X_1, X_2\} > \zeta; H_0\big) \\ &= 1 - \mathbf{P}\big(\max\{X_1, X_2\} \leq \zeta; H_0\big) \\ &= 1 - \mathbf{P}(X_1 \leq \zeta; H_0)\,\mathbf{P}(X_2 \leq \zeta; H_0) \\ &= 1 - \big(\mathbf{P}(Z \leq \zeta; H_0)\big)^2,\end{aligned}$$

where Z is a standard normal. This yields $\mathbf{P}(Z \leq \zeta; H_0) = \sqrt{1 - 0.05} \approx 0.975$. Using the normal tables, we conclude that $\zeta = 1.96$.

Let us now calculate the resulting false acceptance probability. Letting Z be again a standard normal, we have

$$\begin{aligned}\beta(R') &= \mathbf{P}\big(\max\{X_1, X_2\} \leq 1.96; H_1\big) \\ &= \big(\mathbf{P}(X_1 \leq 1.96; H_1)\big)^2 \\ &= \big(\mathbf{P}(X_1 - 2 \leq -0.04; H_1)\big)^2 \\ &= \big(\mathbf{P}(Z \leq -0.04)\big)^2 \\ &= (0.49)^2 \\ &= 0.24.\end{aligned}$$

We see that the false acceptance probability $\beta(R) = 0.12$ of the LRT is much better than the false acceptance probability $\beta(R') = 0.24$ of the alternative test.

Example 9.14. A Discrete Example. Consider $n = 25$ independent tosses of a coin. Under hypothesis H_0 (respectively, H_1), the probability of a head at each toss is equal to $\theta_0 = 1/2$ (respectively, $\theta_1 = 2/3$). Let X be the number of heads observed. If we set the false rejection probability to 0.1, what is the rejection region associated with the LRT?

We observe that when $X = k$, the likelihood ratio is of the form

$$L(k) = \frac{\binom{n}{k}\theta_1^k(1-\theta_1)^{n-k}}{\binom{n}{k}\theta_0^k(1-\theta_0)^{n-k}} = \left(\frac{\theta_1}{\theta_0}\cdot\frac{1-\theta_0}{1-\theta_1}\right)^k\cdot\left(\frac{1-\theta_1}{1-\theta_0}\right)^n = 2^k\left(\frac{2}{3}\right)^{25}.$$

Note that $L(k)$ is a monotonically increasing function of k. Thus, the rejection condition $L(k) > \xi$ is equivalent to a condition $k > \gamma$, for a suitable value of γ. We conclude that the LRT is of the form

$$\text{reject } H_0 \text{ if } X > \gamma.$$

To guarantee the requirement on the false rejection probability, we need to find the smallest possible value of γ for which $\mathbf{P}(X > \gamma; H_0) \leq 0.1$, or

$$\sum_{i=\gamma+1}^{25}\binom{25}{i}2^{-25} \leq 0.1.$$

By evaluating numerically the right-hand side above for different choices of γ, we find that the required value is $\gamma = 16$.

An alternative method for choosing γ involves an approximation based on the central limit theorem. Under H_0,

$$Z = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = \frac{X - 12.5}{\sqrt{25/4}}$$

is approximately a standard normal random variable. Therefore, we need

$$0.1 = \mathbf{P}(X > \gamma; H_0) = \mathbf{P}\left(\frac{X - 12.5}{\sqrt{25/4}} > \frac{\gamma - 12.5}{\sqrt{25/4}}; H_0\right) = \mathbf{P}\left(Z > \frac{2\gamma}{5} - 5\right).$$

From the normal tables, we have $\Phi(1.28) = 0.9$, and therefore, we should choose γ so that $(2\gamma/5) - 5 = 1.28$, or $\gamma = 15.7$. Since X is integer-valued, we find that the LRT should reject H_0 whenever $X > 15$.

9.4 SIGNIFICANCE TESTING

Hypothesis testing problems encountered in realistic settings do not always involve two well-specified alternatives, so the methodology in the preceding section cannot be applied. The purpose of this section is to introduce an approach to this more general class of problems. We caution, however, that a unique or universal methodology is not available, and that there is a significant element of judgment and art that comes into play.

For some motivation, consider problems such as the following:

(i) A coin is tossed repeatedly and independently. Is the coin fair?

(ii) A die is tossed repeatedly and independently. Is the die fair?

(iii) We observe a sequence of i.i.d. normal random variables $X_1, \ldots, X_n$. Are they standard normal?

(iv) Two different drug treatments are delivered to two different groups of patients with the same disease. Is the first treatment more effective than the second?

(v) On the basis of historical data (say, based on the last year), is the daily change of the Dow Jones Industrial Average normally distributed?

(vi) On the basis of several sample pairs (x_i, y_i) of two random variables X and Y, can we determine whether the two random variables are independent?

In all of the above cases, we are dealing with a phenomenon that involves uncertainty, presumably governed by a probabilistic model. We have a default hypothesis, usually called the **null hypothesis**, denoted by H_0, and we wish to determine on the basis of the observations $X = (X_1, \ldots . X_n)$ whether the null hypothesis should be rejected or not.

In order to avoid obscuring the key ideas, we will mostly restrict the scope of our discussion to situations with the following characteristics.

(a) **Parametric models:** We assume that the observations $X_1, \ldots . X_n$ have a distribution governed by a joint PMF (discrete case) or a joint PDF (continuous case), which is completely determined by an unknown parameter θ (scalar or vector), belonging to a given set $\mathcal{M}$ of possible parameters.

(b) **Simple null hypothesis:** The null hypothesis asserts that the true value of θ is equal to a given element θ_0 of $\mathcal{M}$.

(c) **Alternative hypothesis:** The alternative hypothesis, denoted by H_1, is just the statement that H_0 is not true, i.e., that $\theta \neq \theta_0$.

In reference to the motivating examples introduced earlier, notice that examples (i)-(ii) satisfy conditions (a)-(c) above. On the other hand, in examples (iv)-(vi), the null hypothesis is not simple. violating condition (b).

The General Approach

We introduce the general approach through a concrete example. We then summarize and comment on the various steps involved. Finally, we consider a few more examples that conform to the general approach.

Example 9.15. Is My Coin Fair? A coin is tossed independently $n = 1000$ times. Let θ be the unknown probability of heads at each toss. The set of all

possible parameters is $\mathcal{M} = [0, 1]$. The null hypothesis H_0 ("the coin is fair") is of the form $\theta = 1/2$. The alternative hypothesis is that $\theta \neq 1/2$.

The observed data is a sequence $X_1, \ldots, X_n$, where X_i equals 1 or 0, depending on whether the ith toss resulted in heads or tails. We choose to address the problem by considering the value of $S = X_1 + \cdots + X_n$, the number of heads observed, and using a decision rule of the form:

$$\text{reject } H_0 \text{ if } \left|S - \frac{n}{2}\right| > \xi,$$

where ξ is a suitable **critical value**, to be determined. We have so far defined the shape of the **rejection region** R (the set of data vectors that lead to rejection of the null hypothesis). We finally choose the critical value ξ so that the probability of false rejection is equal to a given value α:

$$\mathbf{P}(\text{reject } H_0 ; H_0) = \alpha,$$

Typically, α, called the **significance level**, is a small number; in this example, we use $\alpha = 0.05$.

The discussion so far involved only a sequence of intuitive choices. Some probabilistic calculations are now needed to determine the critical value ξ. Under the null hypothesis, the random variable S is binomial with parameters $n = 1000$ and $p = 1/2$. Using the normal approximation to the binomial and the normal tables, we find that an appropriate choice is $\xi = 31$. If, for example, the observed value of S turns out to be $s = 472$, we have

$$|s - 500| = |472 - 500| = 28 \leq 31$$

and the hypothesis H_0 **is not rejected at the 5% significance level**.

Our use of the language "not rejected" as opposed to "accepted," at the end of the preceding example is deliberate. We do not have any firm grounds to assert that θ equals $1/2$, as opposed to, say, 0.51. We can only assert that the observed value of S does not provide strong evidence against hypothesis H_0.

We can now summarize and generalize the essence of the preceding example, to obtain a generic methodology.

Significance Testing Methodology

A statistical test of a hypothesis "$H_0 : \theta = \theta^*$" is to be performed, based on the observations $X_1, \ldots, X_n$.

- The following steps are carried out before the data are observed.
  - (a) Choose a **statistic** S, that is, a scalar random variable that will summarize the data to be obtained. Mathematically, this involves the choice of a function $h : \Re^n \to \Re$, resulting in the statistic $S = h(X_1, \ldots, X_n)$.

(b) Determine the **shape of the rejection region** by specifying the set of values of S for which H_0 will be rejected as a function of a yet undetermined critical value ξ.

(c) Choose the **significance level,** i.e., the desired probability α of a false rejection of H_0.

(d) Choose the **critical value** ξ so that the probability of false rejection is equal (or approximately equal) to α. At this point, the rejection region is completely determined.

- Once the values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ are observed:

(i) Calculate the value $s = h(x_1, \ldots, x_n)$ of the statistic S.

(ii) Reject the hypothesis H_0 if s belongs to the rejection region.

Let us add some comments and interpretation for the various elements of the above methodology.

(i) There is no universal method for choosing the "right" statistic S. In some cases, as in Example 9.15, the choice is natural and can also be justified mathematically. In other cases, a meaningful choice of S involves a certain generalization of the likelihood ratio, to be touched upon later in this section. Finally, in many situations, the primary consideration is whether S is simple enough to enable the calculations needed in step (d) of the above methodology.

(ii) The set of values of S under which H_0 is not rejected is usually an interval surrounding the peak of the distribution of S under H_0 (see Fig. 9.13). In the limit of a large sample size n, the central limit theorem often applies to S, and the symmetry of the normal distribution suggests an interval which is symmetric around the mean value of S. Similarly, the symmetry of the rejection region in Example 9.15 is well-motivated by the fact that, under H_0, the distribution of S (binomial with parameter $1/2$) is symmetric around its mean. In other cases, however, nonsymmetric rejection regions are more appropriate. For example, if we are certain that the coin in Example 9.15 satisfies $\theta \geq 1/2$, a one-sided rejection region is natural:

$$\text{reject } H_0 \text{ if } S - \frac{n}{2} > \xi.$$

(iii) Typical choices for the false rejection probability α range between $\alpha = .10$ and $\alpha = 0.01$. Of course, one wishes false rejections to be rare, but in light of the tradeoff discussed in the context of simple binary hypotheses, a smaller value of α makes it more difficult to reject a false hypothesis, i.e., increases the probability of false acceptance.

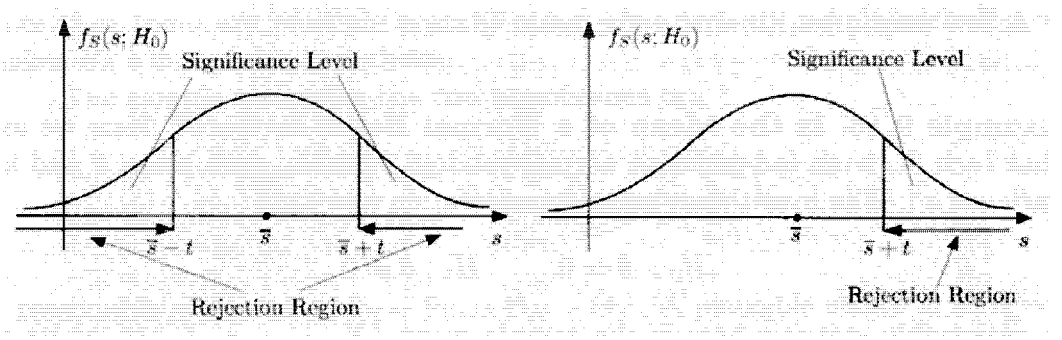

Figure 9.13: Two-sided and one-sided rejection regions for significance testing, based on a statistic S with mean $\overline{s}$ under the null hypothesis. The significance level is the probability of false rejection, i.e., the probability, under H_0, that the statistic S takes a value within the rejection region.

(iv) Step (d) is the only place where probabilistic calculations are used. It requires that the distribution of $L(X)$ [or of a related random variable such as $\log L(X)$] under the hypothesis H_0 be available, possibly approximately. In special cases, this is straightforward or involves an exercise in derived distributions. However, except for relatively simple situations, the distribution of S cannot be found in closed form. If n is large, one can often use well-justified approximations, e.g., based on the central limit theorem. On the other hand, if n is moderate, useful approximations may be difficult to obtain. For this reason, the choice of the statistic S is sometimes guided by the desire to obtain a tractable expression or approximation for the distribution of S. Alternatively, the distribution of S may be estimated by simulation, e.g., by generating many independent samples of X, and by using the resulting samples of $L(X)$ to build a histogram/estimated distribution.

Given the value of α, if the hypothesis H_0 ends up being rejected, one says that H_0 is **rejected at the α significance level.** This statement needs to be interpreted properly. It does not mean that the probability of H_0 being true is less than α. Instead, it means that when this particular methodology is used, we will have false rejections a fraction α of the time. Rejecting a hypothesis at the 1% significance level means that the observed data are highly unusual under the model associated with H_0; such data would arise only 1% of the time, and thus provide strong evidence that H_0 may be false.

Quite often, statisticians skip steps (c) and (d) in the above described methodology. Instead, once they calculate the realized value s of S, they determine and report an associated **p-value** defined by

$$p\text{-value} = \min\{\alpha \mid H_0 \text{ would be rejected at the } \alpha \text{ significance level}\}.$$

Equivalently, the p-value is the value of α for which s would be exactly at the threshold between rejection and non-rejection. Thus, for example, the null hypothesis would be rejected at the 5% significance level if and only if the p-value is smaller than 0.05.

A few examples illustrate the main ideas.

Example 9.16. Is the Mean of a Normal Equal to Zero? Here we assume that each X_i is an independent normal random variable, with mean θ and known variance σ^2. The hypotheses under consideration are:

$$H_0 : \theta = 0, \qquad H_1 : \theta \neq 0.$$

A reasonable statistic here is the sample mean $(X_1 + \cdots + X_n)/n$ or its scaled version

$$S = \frac{X_1 + \cdots + X_n}{\sigma\sqrt{n}}.$$

A natural choice for the shape of the rejection region is to reject H_0 if and only if $|S| > \xi$. Because S has a standard normal distribution, the value of ξ corresponding to any particular value of α is easily found from the normal tables. For example, if $\alpha = 0.05$, we use the fact that $\mathbf{P}(S \leq 1.96) = 0.975$ to obtain a rejection region of the form

$$\text{reject } H_0 \text{ if } \ |S| > 1.96,$$

or equivalently,

$$\text{reject } H_0 \text{ if } \ |X_1 + \cdots + X_n| > 1.96\,\sigma\sqrt{n}.$$

In a one-sided version of this problem, the alternative hypothesis is of the form $H_1 : \theta > 0$. In this case, the same statistic S can be used, but we will reject H_0 if $S > \xi$, where ξ is chosen so that $\mathbf{P}(S > \xi) = \alpha$. Once more, since S has a standard normal distribution, the value of ξ corresponding to any particular value of α is easily found from the normal tables.

Finally, if the variance σ^2 is unknown, we may replace it by an estimate such as

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{X_1 + \cdots + X_n}{n}\right)^2.$$

In this case, the resulting statistic has a t-distribution (as opposed to normal). If n is relatively small, the t-tables should be used instead of the normal tables (cf. Section 9.1).

Our next example involves a composite null hypothesis H_0, in the sense that there are multiple parameter choices that are compatible with H_0.

Example 9.17. Are the Means of Two Populations Equal? We want to test whether a certain medication is equally effective for two different population groups. We draw independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ from the two populations, where $X_i = 1$ (or $Y_i = 1$) if the medication is effective for the ith person in the first (respectively, the second) group, and $X_i = 0$ (or $Y_i = 0$) otherwise. We view each X_i (or Y_i) as a Bernoulli random variable with unknown mean θ_X (respectively, θ_Y), and we consider the hypotheses

$$H_0 : \theta_X = \theta_Y, \qquad H_1 : \theta_X \neq \theta_Y.$$

Note that there are multiple pairs (θ_X, θ_Y) that are compatible with H_0, which makes H_0 a composite hypothesis.

The sample means for the two populations are

$$\hat{\Theta}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \hat{\Theta}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i.$$

A reasonable estimator of $\theta_X - \theta_Y$ is $\hat{\Theta}_X - \hat{\Theta}_Y$. A plausible choice is to reject H_0 if and only if

$$|\hat{\Theta}_X - \hat{\Theta}_Y| > t,$$

for a suitable threshold t to be determined on the basis of the given false rejection probability α. However, an appropriate choice of t is made difficult by the fact that the distribution of $\hat{\Theta}_X - \hat{\Theta}_Y$ under H_0 depends on the unspecified parameters θ_X and θ_Y. This motivates a somewhat different statistic, as we discuss next.

For large n_1 and n_2, the sample means $\hat{\Theta}_X$ and $\hat{\Theta}_Y$ are approximately normal, and because they are independent, $\hat{\Theta}_X - \hat{\Theta}_Y$ is also approximately normal with mean $\theta_X - \theta_Y$ and variance

$$\text{var}(\hat{\Theta}_X - \hat{\Theta}_Y) = \text{var}(\hat{\Theta}_X) + \text{var}(\hat{\Theta}_Y) = \frac{\theta_X(1-\theta_X)}{n_1} + \frac{\theta_Y(1-\theta_Y)}{n_2}.$$

Under hypothesis H_0, the mean of $\hat{\Theta}_X - \hat{\Theta}_Y$ is known (equal to zero), but its variance is not, because the common value of θ_X and θ_Y is not known. On the other hand, under H_0, the common value of θ_X and θ_Y can be estimated by the overall sample mean

$$\hat{\Theta} = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2},$$

the variance $\text{var}(\hat{\Theta}_X - \hat{\Theta}_Y) = \text{var}(\hat{\Theta}_X) + \text{var}(\hat{\Theta}_Y)$ can be approximated by

$$\hat{\sigma}^2 = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{\Theta}(1-\hat{\Theta}),$$

and $(\hat{\Theta}_X - \hat{\Theta}_Y)/\hat{\sigma}$ is approximately a standard normal random variable. This leads us to consider a rejection region of the form

$$\text{reject } H_0 \text{ if } \frac{|\hat{\Theta}_X - \hat{\Theta}_Y|}{\hat{\sigma}} > \xi,$$

and to choose ξ so that $\Phi(\xi) = 1 - \alpha/2$, where Φ is the standard normal CDF. For example, if $\alpha = 0.05$, we obtain a rejection region of the form

$$\text{reject } H_0 \text{ if } \frac{|\hat{\Theta}_X - \hat{\Theta}_Y|}{\hat{\sigma}} > 1.96.$$

In a variant of the methodology in this example, we may consider the hypotheses

$$H_0 : \theta_X = \theta_Y, \qquad H_1 : \theta_X > \theta_Y,$$

which would be appropriate if we had reason to exclude the possibility $\theta_X < \theta_Y$. Then, the corresponding rejection region should be one-sided, of the form

$$\text{reject } H_0 \text{ if } \frac{\hat{\Theta}_X - \hat{\Theta}_Y}{\hat{\sigma}} > \xi,$$

where ξ is chosen so that $\Phi(\xi) = 1 - \alpha$.

The preceding example illustrates a generic issue that arises whenever the null hypothesis is composite. In order to be able to set the critical value appropriately, it is preferable to work with a statistic whose approximate distribution is available and is the same for all parameter values compatible with the null hypothesis, as was the case for the statistic $(\hat{\Theta}_X - \hat{\Theta}_Y)/\hat{\sigma}$ in Example 9.17.

Generalized Likelihood Ratio and Goodness of Fit Tests

Our last topic involves testing whether a given PMF conforms with observed data. This an important problem, known as testing for **goodness of fit**. We will also use it as an introduction to a general methodology for significance testing in the face of a composite alternative hypothesis.

Consider a random variable that takes values in the finite set $\{1, \ldots, m\}$, and let θ_k be the probability of outcome k. Thus, the distribution (PMF) of this random variable is described by the vector parameter $\theta = (\theta_1, \ldots, \theta_m)$. We consider the hypotheses

$$H_0 : \ \theta = (\theta_1^*, \ldots, \theta_m^*), \qquad H_1 : \ \theta \neq (\theta_1^*, \ldots, \theta_m^*),$$

where the θ_k^* are given nonnegative numbers that sum to 1. We draw n independent samples of the random variable of interest, and let N_k be the number of samples that result in outcome k. Thus, our observation is $X = (N_1, \ldots, N_m)$ and we denote its realized value by $x = (n_1, \ldots, n_m)$. Note that $N_1 + \cdots + N_m = n_1 + \cdots + n_m = n$.

As a concrete example, consider n independent rolls of a die and the hypothesis H_0 that the die is fair. In this case, $\theta_k^* = 1/6$, for $k = 1, \ldots, 6$, and N_k is the number of rolls whose result was equal to k. Note that the alternative hypothesis H_1 is composite, as it is compatible with multiple choices of θ.

The approach that we will follow is known as a **generalized likelihood ratio test** and involves two steps:

(a) Estimate a model by ML, i.e., determine a parameter vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$ that maximizes the likelihood function $p_X(x; \theta)$ over all vectors θ.

(b) Carry out a LRT that compares the likelihood $p_X(x; \theta^*)$ under H_0 to the likelihood $p_X(x; \hat{\theta})$ corresponding to the estimated model. More concretely, form the generalized likelihood ratio

$$\frac{p_X(x; \hat{\theta})}{p_X(x; \theta^*)},$$

and if it exceeds a critical value ξ, reject H_0. As in binary hypothesis testing, we choose ξ so that the probability of false rejection is (approximately) equal to a given significance level α.

In essence, this approach asks the following question: is there a model compatible with H_1 that provides a better explanation for the observed data than that provided by the model corresponding to H_0? To answer this question, we compare the likelihood under H_0 to the largest possible likelihood under models compatible with H_1.

The first step (ML estimation) involves a maximization over the set of probability distributions $(\theta_1, \ldots, \theta_m)$. The PMF of the observation vector X is multinomial (cf. Problem 27 in Chapter 2), and the likelihood function is

$$p_X(x;\, \theta) = c\, \theta_1^{n_1} \cdots \theta_m^{n_m},$$

where c is a normalizing constant. It is easier to work with the log-likelihood function, which takes the form

$$\log p_X(x;\, \theta) = \log c + n_1 \log \theta_1 + \cdots + n_{m-1} \log \theta_{m-1} + n_m \log(1 - \theta_1 - \cdots - \theta_{m-1}),$$

where we have also used the fact $\theta_1 + \cdots + \theta_m = 1$ to eliminate θ_m. Assuming that the vector $\hat{\theta}$ that maximizes the log-likelihood has positive components, it can be found by setting the derivatives with respect to $\theta_1, \ldots, \theta_{m-1}$ of the above expression to zero, which yields

$$\frac{n_k}{\hat{\theta}_k} = \frac{n_m}{1 - \hat{\theta}_1 - \cdots - \hat{\theta}_{m-1}}, \qquad \text{for } k = 1, \ldots, m-1.$$

Since the term on the right-hand side is equal to $n_m/\hat{\theta}_m$, we conclude that all ratios $n_k/\hat{\theta}_k$ must be equal. Using also the fact $n_1 + \cdots + n_m = n$, it follows that

$$\hat{\theta}_k = \frac{n_k}{n}, \qquad k = 1, \ldots, m.$$

It can be shown that these are the correct ML estimates even if some of the n_k happen to be zero, in which case the corresponding $\hat{\theta}_k$ are also zero.

The resulting generalized likelihood ratio test is of the form†

$$\text{reject } H_0 \text{ if } \quad \frac{p_X(x;\, \hat{\theta})}{p_X(x;\, \theta^*)} = \prod_{k=1}^{m} \frac{(n_k/n)^{n_k}}{(\theta_k^*)^{n_k}} > \xi,$$

where ξ is the critical value. By taking logarithms, the test reduces to

$$\text{reject } H_0 \text{ if } \quad \sum_{k=1}^{m} n_k \log\left(\frac{n_k}{n\theta_k^*}\right) > \log \xi.$$

† We adopt the convention that $0^0 = 1$ and $0 \cdot \log 0 = 0$

We need to determine ξ by taking into account the required significance level, that is,

$$\mathbf{P}(S > \log \xi;\ H_0) = \alpha,$$

where

$$S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta_k^*}\right).$$

This may be problematic because the distribution of S under H_0 is not readily available and can only be simulated.

Fortunately, major simplifications are possible when n is large. In this case, the observed frequencies $\hat{\theta}_k = n_k/n$ will be close to θ_k^* under H_0, with high probability. Then, a second order Taylor series expansion shows that our statistic S can be approximated well by $T/2$, where T is given by†

$$T = \sum_{k=1}^{m} \frac{(N_k - n\theta_k^*)^2}{n\theta_k^*}.$$

Furthermore, when n is large, it is known that under the hypothesis H_0, the distribution of T (and consequently the distribution of $2S$) approaches a so-called "χ^2 distribution with $m-1$ degrees of freedom."‡ The CDF of this distribution

---

† We note that the second order Taylor series expansion of the function $y \log(y/y^*)$ around any $y^* > 0$ is of the form

$$y \log\left(\frac{y}{y^*}\right) \approx y - y^* + \frac{1}{2}\frac{(y-y^*)^2}{y^*},$$

and is valid when $y/y^* \approx 1$. Thus,

$$\sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta_k^*}\right) \approx \sum_{k=1}^{m}(N_k - n\theta_k^*) + \frac{1}{2}\sum_{k=1}^{m}\frac{(N_k - n\theta_k^*)^2}{n\theta_k^*} = \frac{T}{2}.$$

‡ The χ^2 distribution with ℓ degrees of freedom is defined as the distribution of the random variable

$$\sum_{i=1}^{\ell} Z_i^2,$$

where $Z_1, \ldots, Z_\ell$ are independent standard normal random variables (zero mean and unit variance). Some intuition for why T is approximately χ^2 can be gained from the fact that as $n \to \infty$, N_k/n not only converges to θ_k^* but is also asymptotically normal. Thus, T is equal to the sum of the squares of m zero mean normal random variables, namely $(N_k - n\theta_k^*)/\sqrt{n\theta_k^*}$. The reason that T has $m-1$, instead of m, degrees of freedom is related to the fact that $\sum_{k=1}^{m} N_k = n$, so that these m random variables are actually dependent.

is available in tables, similar to the normal tables. Thus, approximately correct values of $\mathbf{P}(T > \gamma; H_0)$ or $\mathbf{P}(2S > \gamma; H_0)$ can be obtained from the χ^2 tables and can be used to determine a suitable critical value that corresponds to the given significance level α. Putting everything together, we have the following test for large values of n.

The Chi-Square Test:

- Use the statistic

$$S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta_k^*}\right)$$

(or possibly the related statistic T) and a rejection region of the form

$$\text{reject } H_0 \text{ if } 2S > \gamma$$

(or $T > \gamma$, respectively).

- The critical value γ is determined from the CDF tables for the χ^2 distribution with $m - 1$ degrees of freedom so that

$$\mathbf{P}(2S > \gamma; H_0) = \alpha,$$

where α is a given significance level.

Example 9.18. Is My Die Fair? A die is rolled independently 600 times and the number of times that the numbers $1, 2, 3, 4, 5, 6$ come up are

$$n_1 = 92, \quad n_2 = 120, \quad n_3 = 88, \quad n_4 = 98, \quad n_5 = 95, \quad n_6 = 107,$$

respectively. Let us test the hypothesis H_0 that the die is fair by using the chi-square test based on the statistic T, at a level of significance $\alpha = 0.05$. From the tables for the χ^2 with 5 degrees of freedom, we obtain that for $\mathbf{P}(T > \gamma; H_0) = 0.05$ we must have $\gamma = 11.1$.

With $\theta_1^* = \cdots = \theta_6^* = 1/6$, $n = 600$, $n\theta_k^* = 100$, and the given values n_k, the value of the statistic T is

$$\begin{aligned}\sum_{k=1}^{m} \frac{(n_k - n\theta_k^*)^2}{n\theta_k^*} &= \frac{(92-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(88-100)^2}{100} \\ &\quad + \frac{(98-100)^2}{100} + \frac{(95-100)^2}{100} + \frac{(107-100)^2}{100} \\ &= 6.86.\end{aligned}$$

Since $T = 6.86 < 11.1$, the hypothesis that the die is fair is not rejected. If we use instead the statistic S, then a calculation using the data yields $2S = 6.68$, which

is both close to T and also well below the critical value $\gamma = 11.1$. If the level of significance were $\alpha = 0.25$, the corresponding value of γ would be 6.63. In this case, the hypothesis that the die is fair would be rejected since $T = 6.86 > 6.63$ and $2S = 6.68 > 6.63$.

9.5 SUMMARY AND DISCUSSION

Classical inference methods, in contrast with Bayesian methods, treat θ as an unknown constant. Classical parameter estimation aims at estimators with favorable properties such as a small bias and a satisfactory confidence interval, for all possible values of θ. We first focused on ML estimation, which is related to the (Bayesian) MAP method and selects an estimate of θ that maximizes the likelihood function given x. It is a general estimation method and has several desirable characteristics, particularly when the number of observations is large. Then, we discussed the special but practically important case of estimating an unknown mean and constructing confidence intervals. Much of the methodology here relies on the central limit theorem. We finally discussed the linear regression method that aims to match a linear model to the observations in a least squares sense. It requires no probabilistic assumptions for its application, but it is also related to ML and Bayesian LMS estimation under certain conditions.

Classical hypothesis testing methods aim at small error probabilities, combined with simplicity and convenience of calculation. We have focused on tests that reject the null hypothesis when the observations fall within a simple type of rejection region. The likelihood ratio test is the primary approach for the case of two competing simple hypotheses, and derives strong theoretical support from the Neyman-Pearson Lemma. We also addressed significance testing, which applies when one (or both) of the competing hypotheses is composite. The main approach here involves a suitably chosen statistic that summarizes the observations, and a rejection region whose probability under the null hypothesis is set to a desired significance level.

In our brief introduction to statistics, we aimed at illustrating the central concepts and the most common methodologies, but we have barely touched the surface of a very rich subject. For example, we have not discussed important topics such as estimation in time-varying environments (time series analysis, and filtering), nonparametric estimation (e.g., the problem of estimating an unknown PDF on the basis of empirical data), further developments in linear and nonlinear regression (e.g., testing whether the assumptions underlying a regression model are valid), methods for designing statistical experiments, methods for validating the conclusions of a statistical study, computational methods, and many others. Yet, we hope to have kindled the reader's interest in the subject and to have provided some general understanding of the conceptual framework.

PROBLEMS

SECTION 9.1. Classical Parameter Estimation

Problem 1. Alice models the time that she spends each week on homework as an exponentially distributed random variable with unknown parameter θ. Homework times in different weeks are independent. After spending 10, 14, 18, 8, and 20 hours in the first 5 weeks of the semester, what is her ML estimate of θ?

Problem 2. Consider a sequence of independent coin tosses, and let θ be the probability of heads at each toss.

(a) Fix some k and let N be the number of tosses until the kth head occurs. Find the ML estimator of θ based on N.

(b) Fix some n and let K be the number of heads observed in n tosses. Find the ML estimator of θ based on K.

Problem 3. Sampling and estimation of sums. We have a box with k balls; $\overline{k}$ of them are white and $k - \overline{k}$ are red. Both k and $\overline{k}$ are assumed known. Each white ball has a nonzero number on it, and each red ball has zero on it. We want to calculate the sum of all the ball numbers, but because k is very large, we resort to estimating it by sampling. This problem aims to quantify the advantages of sampling only white balls/nonzero numbers and exploiting the knowledge of $\overline{k}$. In particular, we wish to compare the error variance when we sample n balls with the error variance when we sample a smaller number m of white balls.

(a) Suppose we draw balls sequentially and independently, according to a uniform distribution (with replacement). Denote by X_i the number on the ith ball drawn, and by Y_i the number on the ith white ball drawn. We fix two positive integers n and m, and denote

$$\hat{S} = \frac{k}{n}\sum_{i=1}^{n} X_i, \qquad \overline{S} = \frac{\overline{k}}{\overline{N}}\sum_{i=1}^{n} X_i, \qquad \tilde{S} = \frac{\overline{k}}{m}\sum_{i=1}^{m} Y_i,$$

where $\overline{N}$ is the (random) number of white balls drawn in the first n samples. Show that $\hat{S}$, $\overline{S}$, and $\tilde{S}$ are unbiased estimators of the sum of all the ball numbers.

(b) Calculate the variances of $\tilde{S}$ and $\hat{S}$, and show that in order for them to be approximately equal, we must have

$$m \approx \frac{np}{p + r(1-p)},$$

where $p = \overline{k}/k$ and $r = \mathbf{E}[Y_1^2]/\mathrm{var}(Y_1)$. Show also that when $m = n$,

$$\frac{\mathrm{var}(\tilde{S})}{\mathrm{var}(\hat{S})} = \frac{p}{p + r(1-p)}.$$

(c) Calculate the variance of $\overline{S}$, and show that for large n,

$$\frac{\text{var}(\overline{S})}{\text{var}(\hat{S})} \approx \frac{1}{p + r(1-p)}.$$

Problem 4. Mixture models. Let the PDF of a random variable X be the mixture of m components:

$$f_X(x) = \sum_{j=1}^{m} p_j f_{Y_j}(x).$$

where

$$\sum_{j=1}^{m} p_j = 1, \qquad p_j \geq 0, \quad \text{for } j = 1, \ldots, m.$$

Thus, X can be viewed as being generated by a two-step process: first draw j randomly according to probabilities p_j, then draw randomly according to the distribution of Y_j. Assume that each Y_j is normal with mean μ_j and variance σ_j^2, and that we have a set of i.i.d. observations $X_1, \ldots, X_n$, each with PDF f_X.

(a) Write down the likelihood and log-likelihood functions.

(b) Consider the case $m = 2$ and $n = 1$, and assume that μ_1, μ_2, σ_1, and σ_2 are known. Find the ML estimates of p_1 and p_2.

(c) Consider the case $m = 2$ and $n = 1$, and assume that p_1, p_2, σ_1, and σ_2 are known. Find the ML estimates of μ_1 and μ_2.

(d) Consider the case $m \geq 2$ and general n, and assume that all parameters are unknown. Show that the likelihood function can be made arbitrarily large by choosing $\mu_1 = x_1$ and letting σ_1^2 decrease to zero. *Note:* This is an example where the ML approach is problematic.

Problem 5. Unstable particles are emitted from a source and decay at a distance X, which is exponentially distributed with unknown parameter θ. A special device is used to detect the first n decay events that occur in the interval $[m_1, m_2]$. Suppose that these events are recorded at distances $X = (X_1, \ldots, X_n)$.

(a) Give the form of the likelihood and log-likelihood functions.

(b) Assume that $m_1 = 1$, $m_2 = 20$, $n = 6$, and $x = (1.5, 2, 3, 4, 5, 12)$. Plot the likelihood and log-likelihood as functions of θ. Find approximately the ML estimate of θ based on your plot.

Problem 6. Consider a study of student heights in a middle school. Assume that the height of a female student is normally distributed with mean μ_1 and variance σ_1^2, and that the height of a male student is normally distributed with mean μ_2 and variance σ_2^2. Assume that a student is equally likely to be male or female. A sample of size $n = 10$ was collected, and the following values were recorded (in centimeters):

164, 167, 163, 158, 170, 183, 176, 159, 170, 167.

(a) Assume that μ_1, μ_2, σ_1, and σ_2 are unknown. Write down the likelihood function.

(b) Assume we know that $\sigma_1^2 = 9$ and $\mu_1 = 164$. Find numerically the ML estimates of σ_2 and μ_2.

(c) Assume we know that $\sigma_1^2 = \sigma_2^2 = 9$. Find numerically the ML estimates of μ_1 and μ_2.

(d) Treating the estimates obtained in part (c) as exact values, describe the MAP rule for deciding a student's gender based on the student's height.

Problem 7. Estimating the parameter of a Poisson random variable. Derive the ML estimator of the parameter of a Poisson random variable based on i.i.d. observations $X_1, \ldots, X_n$. Is the estimator unbiased and consistent?

Problem 8. Estimating the parameter of a uniform random variable I. We are given i.i.d. observations $X_1, \ldots, X_n$ that are uniformly distributed over the interval $[0, \theta]$. What is the ML estimator of θ? Is it consistent? Is it unbiased or asymptotically unbiased? Can you construct alternative estimators that are unbiased?

Problem 9. Estimating the parameter of a uniform random variable II. We are given i.i.d. observations $X_1, \ldots, X_n$ that are uniformly distributed over the interval $[\theta, \theta + 1]$. Find a ML estimator of θ. Is it consistent? Is it unbiased or asymptotically unbiased?

Problem 10. A source emits a random number of photons K each time that it is triggered. We assume that the PMF of K is

$$p_K(k; \theta) = c(\theta) e^{-\theta k}, \qquad k = 0, 1, 2, \ldots,$$

where θ is the inverse of the temperature of the source and $c(\theta)$ is a normalization factor. We also assume that the photon emissions each time that the source is triggered are independent. We want to estimate the temperature of the source by triggering it repeatedly and counting the number of emitted photons.

(a) Determine the normalization factor $c(\theta)$.

(b) Find the expected value and the variance of the number K of photons emitted if the source is triggered once.

(c) Derive the ML estimator for the temperature $\psi = 1/\theta$, based on $K_1, \ldots, K_n$, the numbers of photons emitted when the source is triggered n times.

(d) Show that the ML estimator is consistent.

Problem 11.* Sufficient statistics – factorization criterion. Consider an observation model of the following type. Assuming for simplicity that all random variables are discrete, an initial observation T is generated according to a PMF $p_T(t; \theta)$. Having observed T, an additional observation Y is generated according to a conditional PMF $p_{Y|T}(y \mid t)$ that does not involve the unknown parameter θ. Intuition suggests that out of the overall observation vector $X = (T, Y)$, only T is useful for estimating θ. This problem formalizes this idea.

Given observations $X = (X_1, \ldots, X_n)$, we say that a (scalar or vector) function $q(X)$ is a **sufficient statistic** for the parameter θ if the conditional distribution of X

given the random variable $T = q(X)$ does not depend on θ, i.e., for every event D and possible value t of the random variable T,

$$\mathbf{P}_\theta\big(X \in D \,|\, T = t\big)$$

is the same for all θ for which the above conditional probability is well-defined [i.e., for all θ for which the PMF $p_T(t;\theta)$ or the PDF $f_T(t;\theta)$ is positive]. Assume that either X is discrete (in which case, T is also discrete), or that both X and T are continuous random variables.

(a) Show that $T = q(X)$ is a sufficient statistic for θ if and only if it satisfies the following **factorization criterion**: the likelihood function $p_X(x;\theta)$ (discrete case) or $f_X(x;\theta)$ (continuous case) can be written as $r\big(q(x),\theta\big)s(x)$ for some functions r and s.

(b) Show that if $q(X)$ is a sufficient statistic for θ, then for any function h of θ, $q(X)$ is a sufficient statistic for the parameter $\zeta = h(\theta)$.

(c) Show that if $q(X)$ is a sufficient statistic for θ, a ML estimate of θ can be written as $\hat{\Theta}_n = \phi\big(q(X)\big)$ for some function ϕ. *Note*: This supports the idea that a sufficient statistic captures all essential information about θ provided by X.

Solution. (a) We consider only the discrete case; the proof for the continuous case is similar. Assume that the likelihood function can be written as $r\big(q(x),\theta\big)s(x)$. We will show that $T = q(X)$ is a sufficient statistic.

Fix some t and consider some θ for which $\mathbf{P}_\theta(T = t) > 0$. For any x for which $q(x) \neq t$, we have $\mathbf{P}_\theta(X = x \,|\, T = t) = 0$, which is trivially the same for all θ. Consider now any x for which $q(x) = t$. Using the fact $\mathbf{P}_\theta\big(X = x, T = t\big) = \mathbf{P}_\theta\big(X = x, q(X) = q(x)\big) = \mathbf{P}_\theta(X = x)$, we have

$$\begin{aligned}
\mathbf{P}_\theta(X = x \,|\, T = t) &= \frac{\mathbf{P}_\theta(X = x, T = t)}{\mathbf{P}_\theta(T = t)} = \frac{\mathbf{P}_\theta(X = x)}{\mathbf{P}_\theta(T = t)} \\
&= \frac{r(t,\theta)s(x)}{\displaystyle\sum_{\{z \,|\, q(z)=t\}} r\big(q(z),\theta\big)s(z)} = \frac{r(t,\theta)s(x)}{r(t,\theta)\displaystyle\sum_{\{z \,|\, q(z)=t\}} s(z)} \\
&= \frac{s(x)}{\displaystyle\sum_{\{z \,|\, q(z)=t\}} s(z)},
\end{aligned}$$

so $\mathbf{P}_\theta(X = x \,|\, T = t)$ does not depend on θ. This implies that for any event D, the conditional probability $\mathbf{P}_\theta(X \in D \,|\, T = t)$ is the same for all θ for which $\mathbf{P}_\theta(T = t) > 0$, so T is a sufficient statistic.

Conversely, assume that $T = q(X)$ is a sufficient statistic. For any x with $p_X(x;\theta) > 0$, the likelihood function is

$$p_X(x;\theta) = \mathbf{P}_\theta\big(X = x \,|\, q(X) = q(x)\big)\,\mathbf{P}_\theta\big(q(X) = q(x)\big).$$

Since T is a sufficient statistic, the first term on the right-hand side does not depend on θ, and is of the form $s(x)$. The second term depends on x through $q(x)$, and is of

the form $r\big(q(x),\theta\big)$. This establishes that the likelihood function can be factored as claimed.

(b) This is evident from the definition of a sufficient statistic, since for $\zeta = h(\theta)$, we have

$$\mathbf{P}_\zeta\big(X \in D \,|\, T = t\big) = \mathbf{P}_\theta\big(X \in D \,|\, T = t\big),$$

so $\mathbf{P}_\zeta\big(X \in D \,|\, T = t\big)$ is the same for all ζ.

(c) By part (a), the likelihood function can be factored as $r\big(q(x),\theta\big)s(x)$. Thus, a ML estimate maximizes $r\big(q(x),\theta\big)$ over θ [if $s(x) > 0$] or minimizes $r\big(q(x),\theta\big)$ over θ [if $s(x) < 0$], and therefore depends on x only through $q(x)$.

Problem 12.* Examples of a sufficient statistic I. Show that $q(X) = \sum_{i=1}^n X_i$ is a sufficient statistic in the following cases:

(a) $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables with parameter θ.

(b) $X_1, \ldots, X_n$ are i.i.d. Poisson random variables with parameter θ.

Solution. (a) The likelihood function is

$$p_X(x;\theta) = \theta^{q(x)}(1-\theta)^{n-q(x)},$$

so it can be factored as the product of the function $\theta^{q(x)}(1-\theta)^{n-q(x)}$, which depends on x only through $q(x)$, and the constant function $s(x) \equiv 1$. The result follows from the factorization criterion for a sufficient statistic.

(b) The likelihood function is

$$p_X(x;\theta) = \prod_{i=1}^n p_{X_i}(x_i) = e^{-\theta}\prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} = e^{-\theta}\theta^{q(x)}\frac{1}{\prod_{i=1}^n x_i!},$$

so it can be factored as the product of the function $e^{-\theta}\theta^{q(x)}$, which depends on x only through $q(x)$, and the function $s(x) = 1/\prod_{i=1}^n x_i!$, which depends only on x. The result follows from the factorization criterion for a sufficient statistic.

Problem 13.* Examples of a sufficient statistic II. Let $X_1, \ldots, X_n$ be i.i.d. normal random variables with mean μ and variance σ^2. Show that:

(a) If σ^2 is known, $q(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for μ.

(b) If μ is known, $q(X) = \sum_{i=1}^n (X_i - \mu)^2$ is a sufficient statistic for σ^2.

(a) If both μ and σ^2 are unknown, $q(X) = \big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\big)$ is a sufficient statistic for (μ, σ^2).

Solution. Use the calculations in Example 9.4, and the factorization criterion.

Problem 14.* Rao-Blackwell theorem. This problem shows that a general estimator can be modified into one that only depends on a sufficient statistic, without loss of performance. Given observations $X = (X_1, \ldots, X_n)$, let $T = q(X)$ be a sufficient statistic for the parameter θ, and let $g(X)$ be an estimator for θ.

(a) Show that $\mathbf{E}_\theta\big[g(X)\,|\,T\big]$ is the same for all values of θ. Thus, we can suppress the subscript θ, and view

$$\hat{g}(X) = \mathbf{E}\big[g(X)\,|\,T\big]$$

as a new estimator of θ, which depends on X only through T.

(b) Show that the estimators $g(X)$ and $\hat{g}(X)$ have the same bias.

(c) Show that for any θ with $\mathrm{var}_\theta\big(g(X)\big) < \infty$,

$$\mathbf{E}_\theta\big[\big(\hat{g}(X) - \theta\big)^2\big] \leq \mathbf{E}_\theta\big[\big(g(X) - \theta\big)^2\big], \qquad \text{for all } \theta.$$

Furthermore, for a given θ, this inequality is strict if and only if

$$\mathbf{E}_\theta\Big[\mathrm{var}\big(g(X)\,|\,T\big)\Big] > 0.$$

Solution. (a) Since $T = q(X)$ is a sufficient statistic, the conditional distribution $\mathbf{P}_\theta(X = x\,|\,T = t)$ does not depend on θ, so the same is true for $\mathbf{E}_\theta\big[g(X)\,|\,T\big]$.

(b) We have by the law of iterated expectations

$$\mathbf{E}_\theta\big[g(X)\big] = \mathbf{E}_\theta\Big[\mathbf{E}\big[g(X)\,|\,T\big]\Big] = \mathbf{E}_\theta\big[\hat{g}(X)\big],$$

so the biases of $g(X)$ and $\hat{g}(X)$ are equal.

(c) Fix some θ and let b_θ denote the common bias of $g(X)$ and $\hat{g}(X)$. We have, using the law of total variance,

$$\begin{aligned}
\mathbf{E}_\theta\big[\big(g(X) - \theta\big)^2\big] &= \mathrm{var}_\theta\big(g(X)\big) + \mathrm{b}_\theta^2 \\
&= \mathbf{E}_\theta\Big[\mathrm{var}\big(g(X)\,|\,T\big)\Big] + \mathrm{var}_\theta\Big(\mathbf{E}\big[g(X)\,|\,T\big]\Big) + \mathrm{b}_\theta^2 \\
&= \mathbf{E}_\theta\Big[\mathrm{var}\big(g(X)\,|\,T\big)\Big] + \mathrm{var}_\theta\big(\hat{g}(X)\big) + \mathrm{b}_\theta^2 \\
&= \mathbf{E}_\theta\Big[\mathrm{var}\big(g(X)\,|\,T\big)\Big] + \mathbf{E}_\theta\big[\big(\hat{g}(X) - \theta\big)^2\big] \\
&\geq \mathbf{E}_\theta\big[\big(\hat{g}(X) - \theta\big)^2,
\end{aligned}$$

with the inequality being strict if and only if $\mathbf{E}_\theta\Big[\mathrm{var}\big(g(X)\,|\,T\big)\Big] > 0$.

Problem 15.* Let $X_1, \ldots, X_n$ be i.i.d. random variables that are uniformly distributed over the interval $[0, \theta]$.

(a) Show that $T = \max_{i=1,\ldots,n} X_i$ is a sufficient statistic.

(b) Show that $g(X) = (2/n)\sum_{i=1}^n X_i$ is an unbiased estimator.

(c) Find the form of the estimator $\hat{g}(X) = \mathbf{E}\big[g(X)\,|\,T\big]$, and then calculate and compare $\mathbf{E}_\theta\big[\big(\hat{g}(X) - \theta\big)^2\big]$ with $\mathbf{E}_\theta\big[\big(g(X) - \theta\big)^2\big]$.

Solution. (a) The likelihood function is

$$f_X(x_1,\ldots,x_n;\theta) = f_{X_1}(x_1;\theta)\cdots f_{X_n}(x_n;\theta) = \begin{cases} 1/\theta^n, & \text{if } 0 \le \max_{i=1,\ldots,n} x_i \le \theta \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

and depends on x only through $q(x) = \max_{i=1,\ldots,n} x_i$. The result follows from the factorization criterion for a sufficient statistic.

(b) We have

$$\mathbf{E}_\theta\big[g(X)\big] = \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_\theta[X_i] = \frac{2}{n}\sum_{i=1}^{n}\frac{\theta}{2} = \theta.$$

(c) Conditioned on the event $\{T = t\}$, one of the observations X_i is equal to t. The remaining $n-1$ observations are uniformly distributed over the interval $[0,t]$, and have a conditional expectation equal to $t/2$. Thus,

$$\mathbf{E}\big[g(X)\,|\,T=t\big] = \frac{2}{n}\mathbf{E}\Big[\sum_{i=1}^{n} X_i\,|\,T=t\Big] = \frac{2}{n}\Big(t + \frac{(n-1)t}{2}\Big) = \frac{n+1}{n}t,$$

and $\hat{g}(X) = \mathbf{E}\big[g(X)\,|\,T\big] = (n+1)T/n$.

We will now calculate the mean squared error of the two estimators $\hat{g}(X)$ and $g(X)$, as functions of θ. To this end, we evaluate the first and second moment of $\hat{g}(X)$. We have

$$\mathbf{E}_\theta\,[\hat{g}(X)] = \mathbf{E}_\theta\big[\mathbf{E}\,[g(X)\,|\,T]\,\big] = \mathbf{E}_\theta\big[g(X)\big] = \theta.$$

To find the second moment, we first determine the PDF of T. For $t \in [0,\theta]$, we have $\mathbf{P}_\theta(T \le t) = (t/\theta)^n$, and by differentiating, $f_T(t;\theta) = nt^{n-1}/\theta^n$. Thus,

$$\begin{aligned}\mathbf{E}_\theta\big[\big(\hat{g}(X)\big)^2\big] &= \Big(\frac{n+1}{n}\Big)^2\mathbf{E}\,[T^2] = \Big(\frac{n+1}{n}\Big)^2\int_0^\theta t^2 f_T(t;\theta)\,dt \\ &= \Big(\frac{n+1}{n}\Big)^2\int_0^\theta t^2\frac{nt^{n-1}}{\theta^n}\,dt = \frac{(n+1)^2}{n(n+2)}\theta^2.\end{aligned}$$

Since $\hat{g}(X)$ has mean θ, its mean squared error is equal to its variance, and

$$\mathbf{E}_\theta\big[\big(\hat{g}(X)-\theta\big)^2\big] = \mathbf{E}_\theta\big[\big(\hat{g}(X)\big)^2\big] - \theta^2 = \frac{(n+1)^2}{n(n+2)}\theta^2 - \theta^2 = \frac{1}{n(n+2)}\theta^2.$$

Similarly, the mean squared error of $g(X)$ is equal to its variance, and

$$\mathbf{E}_\theta\big[\big(g(X)-\theta\big)^2\big] = \frac{4}{n^2}\sum_{i=1}^{n}\mathrm{var}_\theta(X_i) = \frac{4}{n^2}\cdot n\cdot\frac{\theta^2}{12} = \frac{1}{3n}\theta^2.$$

It can be seen that $\dfrac{1}{3n} \ge \dfrac{1}{n(n+2)}$ for all positive integers n. It follows that

$$\mathbf{E}_\theta\big[\big(\hat{g}(X)-\theta\big)^2\big] \le \mathbf{E}_\theta\big[\big(g(X)-\theta\big)^2\big],$$

which is consistent with the Rao-Blackwell theorem.

SECTION 9.2. Linear Regression

Problem 16. An electric utility company tries to estimate the relation between the daily amount of electricity used by customers and the daily summer temperature. It has collected the data shown on the table below.

Temperature	96	89	81	86	83
Electricity	23.67	20.45	21.86	23.28	20.71
Temperature	73	78	74	76	78
Electricity	18.21	18.85	20.10	18.48	17.94

(a) Set up and estimate the parameters of a linear model that can be used to predict electricity consumption as a function of temperature.

(b) If the temperature on a given day is 90 degrees, predict the amount of electricity consumed on that day.

Problem 17. Given the five data pairs (x_i, y_i) in the table below,

x	0.798	2.546	5.005	7.261	9.131
y	−2.373	20.906	103.544	215.775	333.911

we want to construct a model relating x and y. We consider a linear model

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \qquad i = 1, \ldots, 5,$$

and a quadratic model

$$Y_i = \beta_0 + \beta_1 x_i^2 + V_i, \qquad i = 1, \ldots, 5,$$

where W_i and V_i represent additive noise terms, modeled by independent normal random variables with mean zero and variance σ_1^2 and σ_2^2, respectively.

(a) Find the ML estimates of the linear model parameters.

(b) Find the ML estimates of the quadratic model parameters.

(c) Assume that the two estimated models are equally likely to be true, and that the noise terms W_i and V_i have the same variance: $\sigma_1^2 = \sigma_2^2$. Use the MAP rule to choose between the two models.

Problem 18.* Unbiasedness and consistency in linear regression. In a probabilistic framework for regression, let us assume that $Y_i = \theta_0 + \theta_1 x_i + W_i$, $i = 1, \ldots, n$, where $W_1, \ldots, W_n$ are i.i.d. normal random variables with mean zero and variance σ^2. Then, given x_i and the realized values y_i of Y_i, $i = 1, \ldots, n$, the ML estimates of θ_0 and θ_1 are given by the linear regression formulas, as discussed in Section 9.2.

(a) Show that the ML estimators $\hat{\Theta}_0$ and $\hat{\Theta}_1$ are unbiased.

(b) Show that the variances of the estimators $\hat{\Theta}_0$ and $\hat{\Theta}_1$ are

$$\text{var}(\hat{\Theta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad \text{var}(\hat{\Theta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

respectively, and their covariance is

$$\text{cov}(\hat{\Theta}_0, \hat{\Theta}_1) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

(c) Show that if $\sum_{i=1}^{n} (x_i - \overline{x})^2 \to \infty$ and $\overline{x}^2$ is bounded by a constant as $n \to \infty$, we have $\text{var}(\hat{\Theta}_0) \to 0$ and $\text{var}(\hat{\Theta}_1) \to 0$. (This, together with Chebyshev's inequality, implies that the estimators $\hat{\Theta}_0$ and $\hat{\Theta}_1$ are consistent.)

Note: Although the assumption that the W_i are normal is needed for our estimators to be ML estimators, the argument below shows that these estimators remain unbiased and consistent without this assumption.

Solution. (a) Let the true values of θ_0 and θ_1 be θ_0^* and θ_1^*, respectively. We have

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \qquad \hat{\Theta}_0 = \overline{Y} - \hat{\Theta}_1 \overline{x},$$

where $\overline{Y} = \left(\sum_{i=1}^{n} Y_i\right)/n$, and where we treat $x_1, \ldots, x_n$ as constant. Denoting $\overline{W} = \left(\sum_{i=1}^{n} W_i\right)/n$, we have

$$Y_i = \theta_0^* + \theta_1^* x_i + W_i, \qquad \overline{Y} = \theta_0^* + \theta_1^* \overline{x} + \overline{W},$$

and

$$Y_i - \overline{Y} = \theta_1^* (x_i - \overline{x}) + (W_i - \overline{W}).$$

Thus,

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \left(\theta_1^* (x_i - \overline{x}) + W_i - \overline{W}\right)}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \theta_1^* + \frac{\sum_{i=1}^{n} (x_i - \overline{x})(W_i - \overline{W})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$= \theta_1^* + \frac{\sum_{i=1}^{n} (x_i - \overline{x}) W_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where we have used the fact $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Since $\mathbf{E}[W_i] = 0$, it follows that

$$\mathbf{E}[\hat{\Theta}_1] = \theta_1^*.$$

Also

$$\hat{\Theta}_0 = \overline{Y} - \hat{\Theta}_1\overline{x} = \theta_0^* + \theta_1^*\overline{x} + \overline{W} - \hat{\Theta}_1\overline{x} = \theta_0^* + (\theta_1^* - \hat{\Theta}_1)\overline{x} + \overline{W},$$

and using the facts $\mathbf{E}[\hat{\Theta}_1] = \theta_1^*$ and $\mathbf{E}[\overline{W}] = 0$, we obtain

$$\mathbf{E}[\hat{\Theta}_0] = \theta_0^*.$$

Thus, the estimators $\hat{\Theta}_0$ and $\hat{\Theta}_1$ are unbiased.

(b) We now calculate the variance of the estimators. Using the formula for $\hat{\Theta}_1$ derived in part (a) and the independence of the W_i, we have

$$\operatorname{var}(\hat{\Theta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2\operatorname{var}(W_i)}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Similarly, using the formula for $\hat{\Theta}_0$ derived in part (a),

$$\operatorname{var}(\hat{\Theta}_0) = \operatorname{var}(\overline{W} - \hat{\Theta}_1\overline{x}) = \operatorname{var}(\overline{W}) + \overline{x}^2\operatorname{var}(\hat{\Theta}_1) - 2\overline{x}\operatorname{cov}(\overline{W}, \hat{\Theta}_1).$$

Since $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ and $\mathbf{E}[\overline{W}W_i] = \sigma^2/n$ for all i, we obtain

$$\operatorname{cov}(\overline{W}, \hat{\Theta}_1) = \frac{\mathbf{E}\left[\overline{W}\sum_{i=1}^{n}(x_i - \overline{x})W_i\right]}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\frac{\sigma^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.$$

Combining the last three equations, we obtain

$$\operatorname{var}(\hat{\Theta}_0) = \operatorname{var}(\overline{W}) + \overline{x}^2\operatorname{var}(\hat{\Theta}_1) = \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

By expanding the quadratic forms $(x_i - \overline{x})^2$, we also have

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 + n\overline{x}^2 = \sum_{i=1}^{n} x_i^2.$$

By combining the preceding two equations,

$$\text{var}(\hat{\Theta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

We finally calculate the covariance of $\hat{\Theta}_0$ and $\hat{\Theta}_1$. We have

$$\text{cov}(\hat{\Theta}_0, \hat{\Theta}_1) = \mathbf{E}\big[(\hat{\Theta}_0 - \theta_0^*)(\hat{\Theta}_1 - \theta_1^*)\big] = \mathbf{E}\big[\big((\theta_1^* - \hat{\Theta}_1)\overline{x} + \overline{W}\big)(\hat{\Theta}_1 - \theta_1^*)\big],$$

or

$$\text{cov}(\hat{\Theta}_0, \hat{\Theta}_1) = -\overline{x}\,\text{var}(\hat{\Theta}_1) + \text{cov}(\overline{W}, \hat{\Theta}_1).$$

Since, as shown earlier, $\text{cov}(\overline{W}, \hat{\Theta}_1) = 0$, we finally obtain

$$\text{cov}(\hat{\Theta}_0, \hat{\Theta}_1) = -\frac{\overline{x}\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

(c) If $\sum_{i=1}^{n} (x_i - \overline{x})^2 \to \infty$, the expression for $\text{var}(\hat{\Theta}_1) \to 0$ derived in part (b) goes to zero. Then, the formula

$$\text{var}(\hat{\Theta}_0) = \text{var}(\overline{W}) + \overline{x}^2 \text{var}(\hat{\Theta}_1),$$

from part (b), together with the assumption that $\overline{x}^2$ is bounded by a constant, implies that $\text{var}(\hat{\Theta}_0) \to 0$.

Problem 19.* Variance estimate in linear regression. Under the same assumptions as in the preceding problem, show that

$$\hat{S}_n^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i)^2$$

is an unbiased estimator of σ^2.

Solution. Let $\hat{V}_n = \sum_{i=1}^{n} (Y_i - \hat{\Theta}_0 - \hat{\Theta}_1 x_i)^2$. Using the formula $\hat{\Theta}_0 = \overline{Y} - \hat{\Theta}_1 \overline{x}$ and the expression for $\hat{\Theta}_1$, we have

$$\begin{aligned}
\hat{V}_n &= \sum_{i=1}^{n} \big(Y_i - \overline{Y} - \hat{\Theta}_1 (x_i - \overline{x})\big)^2 \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - 2\hat{\Theta}_1 \sum_{i=1}^{n} (Y_i - \overline{Y})(x_i - \overline{x}) + \hat{\Theta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 - \hat{\Theta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 - \hat{\Theta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2.
\end{aligned}$$

Taking expectation of both sides, we obtain

$$\mathbf{E}[\hat{V}_n] = \sum_{i=1}^{n} \mathbf{E}\left[Y_i^2\right] - n\mathbf{E}\left[\overline{Y}^2\right] - \sum_{i=1}^{n}(x_i - \overline{x})^2\mathbf{E}\left[\hat{\Theta}_1^2\right].$$

We also have

$$\mathbf{E}\left[Y_i^2\right] = \text{var}(Y_i) + \left(\mathbf{E}[Y_i]\right)^2 = \sigma^2 + (\theta_0^* + \theta_1^* x_i)^2,$$

$$\mathbf{E}\left[\overline{Y}^2\right] = \text{var}(\overline{Y}) + \left(\mathbf{E}[\overline{Y}]\right)^2 = \frac{\sigma^2}{n} + (\theta_0^* + \theta_1^* \overline{x})^2,$$

$$\mathbf{E}\left[\hat{\Theta}_1^2\right] = \text{var}(\hat{\Theta}_1) + \left(\mathbf{E}[\hat{\Theta}_1]\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + (\theta_1^*)^2.$$

Combining the last four equations and simplifying, we obtain

$$\mathbf{E}[\hat{V}_n] = (n-2)\sigma^2.$$

SECTION 9.3. Binary Hypothesis Testing

Problem 20. A random variable X is characterized by a normal PDF with mean $\mu_0 = 20$, and a variance that is either $\sigma_0^2 = 16$ (hypothesis H_0) or $\sigma_1^2 = 25$ (hypothesis H_1). We want to test H_0 against H_1, using three sample values x_1, x_2, x_3, and a rejection region of the form

$$R = \{x \mid x_1 + x_2 + x_3 > \gamma\}$$

for some scalar γ. Determine the value of γ so that the probability of false rejection is 0.05. What is the corresponding probability of false acceptance?

Problem 21. A normal random variable X is known to have a mean of 60 and a standard deviation equal to 5 (hypothesis H_0) or 8 (hypothesis H_1).

(a) Consider a hypothesis test using a single sample x. Let the rejection region be of the form

$$R = \{\, x \mid |x - 60| > \gamma \}$$

for some scalar γ. Determine γ so that the probability of false rejection of H_0 is 0.1. What is the corresponding false acceptance probability? Would the rejection region change if we were to use the LRT with the same false rejection probability?

(b) Consider a hypothesis test using n independent samples $x_1, \ldots, x_n$. Let the rejection region be of the form

$$R = \left\{(x_1, \ldots, x_n) \mid \left|\frac{x_1 + \cdots + x_n}{n} - 60\right| > \gamma\right\},$$

where γ is chosen so that the probability of false rejection of H_0 is 0.1. How does the false acceptance probability change with n? What can you conclude about the appropriateness of this type of test?

(c) Derive the structure of the LRT using n independent samples $x_1, \ldots, x_n$.

Problem 22. There are two hypotheses about the probability of heads for a given coin: $\theta = 0.5$ (hypothesis H_0) and $\theta = 0.6$ (hypothesis H_1). Let X be the number of heads obtained in n tosses, where n is large enough so that normal approximations are appropriate. We test H_0 against H_1 by rejecting H_0 if X is greater than some suitably chosen threshold k_n.

(a) What should be the value of k_n so that the probability of false rejection is less than or equal to 0.05?

(b) What is the smallest value of n for which both probabilities of false rejection and false acceptance can be made less than or equal to 0.05?

(c) For the value of n found in part (b), what would be the probability of false acceptance if we were to use a LRT with the same probability of false rejection?

Problem 23. The number of phone calls received by a ticket agency on any one day is Poisson distributed. On an ordinary day, the expected value of the number of calls is λ_0, and on a day where there is a popular show in town, the expected value of the number of calls is λ_1, with $\lambda_1 > \lambda_0$. Describe the LRT for deciding whether there is a popular show in town based on the number of calls received. Assume a given probability of false rejection, and find an expression for the critical value ξ.

Problem 24. We have received a shipment of light bulbs whose lifetimes are modeled as independent, exponentially distributed random variables, with parameter equal to λ_0 (hypothesis H_0) or equal to λ_1 (hypothesis H_1). We measure the lifetimes of n light bulbs. Describe the LRT for selecting one of the two hypotheses. Assume a given probability of false rejection of H_0 and give an analytical expression for the critical value ξ.

SECTION 9.4. Significance Testing

Problem 25. Let X be a normal random variable with mean μ and unit variance. We want to test the hypothesis $\mu = 5$ at the 5% level of significance, using n independent samples of X.

(a) What is the range of values of the sample mean for which the hypothesis is accepted?

(b) Let $n = 10$. Calculate the probability of accepting the hypothesis $\mu = 5$ when the true value of μ is 4.

Problem 26. We have five observations drawn independently from a normal distribution with unknown mean μ and unknown variance σ^2.

(a) Estimate μ and σ^2 if the observation values are $8.47, 10.91, 10.87, 9.46, 10.40$.

(b) Use the t-distribution tables to test the hypothesis $\mu = 9$ at the 95% significance level, using the estimates of part (a).

Problem 27. A plant grows on two distant islands. Suppose that its life span (measured in days) on the first (or the second) island is normally distributed with unknown mean μ_X (or μ_Y) and known variance $\sigma_X^2 = 32$ (or $\sigma_Y^2 = 29$. respectively). We wish to test the hypothesis $\mu_X = \mu_Y$. based on 10 independent samples from each island. The corresponding sample means are $\overline{x} = 181$ and $\overline{y} = 177$. Do the data support the hypothesis at the 95% significance level?

Problem 28. A company considers buying a machine to manufacture a certain item. When tested, 28 out of 600 items produced by the machine were found defective. Do the data support the hypothesis that the defect rate of the machine is smaller than 3 percent, at the 5% significance level?

Problem 29. The values of five independent samples of a Poisson random variable turned out to be 34, 35, 29, 31, and 30. Test the hypothesis that the mean is equal to 35 at the 5% level of significance.

Problem 30. A surveillance camera periodically checks a certain area and records a signal $X = W$ if there is no intruder (this is the null hypothesis H_0). If there is an intruder the signal is $X = \theta + W$. where θ is unknown with $\theta > 0$. We assume that W is a normal random variable with mean 0 and known variance $v = 0.5$.

(a) We obtain a single signal value $X = 0.96$. Should H_0 be rejected at the 5% level of significance?

(b) We obtain five independent signal values $X = 0.96, -0.34, 0.85, 0.51, -0.24$. Should H_0 be rejected at the 5% level of significance?

(c) Repeat part (b), using the t-distribution, and assuming the variance v is unknown.