CHAPTER 1. DECISION THEORY AND TESTING SIMPLE HYPOTHESES

1.1 Deciding between two simple hypotheses: the Neyman-Pearson Lemma. Probability theory is reviewed in Appendix D. Suppose an experiment has a set X of possible outcomes. The outcome has some probability distribution μ defined on X. In statistics, we typically don't know what μ is, but we have hypotheses about what it may be. After making observations we'll try to make a decision between or among the hypotheses. In general there could be infinitely many possibilities for μ, but to begin with we're going to look at the case where there are just two possibilities, $\mu = P$ or $\mu = Q$, and we need to decide which it is. For example, a point x in X could give the outcome of a test for a certain disease, where P is the distribution of x for those who don't have the disease and Q is the distribution for those who do.

Often, we have n observations independent with distribution μ. Then X can be replaced by the set X^n of all ordered n-tuples $(x_1, \ldots, x_n)$ of points of X, and μ by the Cartesian product measure $\mu \times \cdots \times \mu$ of n copies of μ. In this way, the case of n observations $x_1, \ldots, x_n$ reduces to that of one "observation" $(x_1, \ldots, x_n)$.

The probability measures P and Q are each defined on some σ-algebra $\mathcal{B}$ of subsets of X, such as the Borel sets in case X is the real line $\mathbb{R}$ or a Euclidean space. A *test* of the hypothesis that $\mu = P$ will be given by a measurable set A, in other words a set A in $\mathcal{B}$. If we observe x in A, then we will reject the hypothesis that $\mu = P$ in favor of the alternative hypothesis that $\mu = Q$. Then $P(A)$ is called the *size* of the test A (at P). The size is the probability that we'll make the error of rejecting P when it's true, i.e. when $\mu = P$, sometimes called a *Type I error*. On the other hand, $Q(A)$ is called the *power* of the test A against the alternative Q. The power is the probability that when Q is true, the test correctly rejects P and prefers Q. The complementary probability $1 - Q(A)$ is sometimes called the probability of a *Type II error*. Given P and Q, for the test A to be as effective as possible, we'd like the size to be small and the power to be large. In the rest of this section, it will be shown how the choice of A can be made optimally.

Example 1.1.1. Let $X = \mathbb{R}$ and let P and Q be normal measures, both with variance 0.04, $P = N(0, 0.04)$ and $Q = N(1, 0.04)$. Larger values of x tend to favor Q, so it seems reasonable to take A as a half-line $[c, \infty)$ for some c. At $x = 1/2$, the densities of P and Q are equal. For $x < 1/2$, P has larger density. For $x > 1/2$, Q does. So if we have no reason in advance to prefer one of P and Q, we might take $c = 1/2$. Then the probabilities of the two types of errors are each about 0.0062 (from tables of the normal distribution). In other words the size is 0.0062 and the power is 0.9938. If the variances had been larger, so would the error probabilities.

It's not always best to prefer the distribution (P or Q) with larger density at the observation (or vector of observations). In testing for a disease, an error indicating a disease when the subject is actually healthy can lead to further, possibly expensive tests or inappropriate treatments. On the other hand the error of overlooking a disease when the patient has it could be much more serious, depending on the severity of the disease.

Numerical values called *losses* will be assigned to the consequences of mistaken decisions. Let $L_{\mu\nu}$ be the loss incurred when μ is true and we decide in favor of ν. A correct decision will be assumed to cause zero loss, so $L_{PP} = L_{QQ} = 0$. The losses L_{PQ} and L_{QP} will be positive and in general will be different.

Also, the statistician may have assigned some probabilities to P or Q in advance, called *prior* probabilities, say $\pi(P) = 1 - \pi(Q)$ with $0 < \pi(P) < 1$. For example it could be known from other data (approximately) what fraction of people in a population being tested have a disease. The part of statistics in which prior probabilities are assumed to exist is known as *Bayesian* statistics, as contrasted with *frequentist* statistics where priors are not assumed. In this book, both are treated. Later on, some pros and cons of the Bayesian and frequentist approaches will be mentioned.

It will turn out that the best tests between P and Q will be based on the ratio of densities of P and Q, called the likelihood ratio, defined as follows. In general, P or Q could have continuous or discrete parts, but P and Q are always absolutely continuous with respect to $P + Q$, so that there is a Radon-Nikodym derivative (RAP, 5.5.4) $h(x) = (dP/d(P+Q))(x)$. Then $dQ/d(P+Q) = 1-h$. The *likelihood ratio* $R_{Q/P}(x)$ of Q to P at x is defined as $(1 - h(x))/h(x)$, or $+\infty$ if $h(x) = 0$. The likelihood ratio, like h, is defined up to equality $(P + Q)$-almost everywhere.

If P and Q have densities f and g respectively with respect to some measure, for example Lebesgue measure on $\mathbb{R}$, then we can take $R_{Q/P}(x) = g(x)/f(x)$ if $f(x) > 0$, or $+\infty$ when $g(x) > 0 = f(x)$, or 0 when $g(x) = 0 = f(x)$. For a proof, see Appendix A.

In Example 1.1.1, $R_{Q/P}(x) \equiv e^{25(x-0.5)}$. Or, let P and Q both be Poisson distributions on the set $\mathbb{N}$ of nonnegative integers with $P(k) = P_\lambda(k) = e^{-\lambda}\lambda^k/k!$ and $Q = P_\rho$ for some ρ. Then $R_{Q/P}(k) = e^{\lambda-\rho}(\rho/\lambda)^k$ for all $k \in \mathbb{N}$.

The sizes $\alpha = 0.05$, 0.01 and 0.001 were chosen rather arbitrarily in the first half of the 20th century and used in selecting tests. So, if a test A has size $\alpha = 0.05$ or less at P, and the observation x is in A, the outcome is called "statistically significant" and the hypothesis P is rejected. If $\alpha \leq 0.001$ the outcome is called "highly significant." The levels 0.05 etc. are still in wide use in some applied fields, such as medicine and psychology, although they are no longer very popular among statisticians themselves. For discrete distributions, not many sizes of tests may be available, as in the following:

Example 1.1.2. Let $X = \{0, 1, 2\}$, $P(0) = 0.8$, $P(1) = 0.05$, $P(2) = 0.15$, $Q(0) = 0.008$, $Q(1) = 0.002$, $Q(2) = 0.99$. Then $R_{Q/P}(x)$ is 0.01, 0.04, and 6.6 for $x = 0, 1, 2$, respectively.

Example 1.1.2 suggests that, at least for discrete distributions, one not insist on conventional, specific sizes for tests. Another approach is to extend the definition of tests as follows. A *randomized test* is a measurable function f defined on X with $0 \leq f(x) \leq 1$ for all x. In such a test, Q is chosen with probability $f(x)$. Specifically, let Y be a random variable uniformly distributed in $[0, 1]$ and independent of x. If x is observed, then we decide in favor of Q if $Y \leq f(x)$ and P otherwise. The *size* of the randomized test f (at P) is defined as $\int f\, dP$ and the *power* of f (at Q) as $\int f\, dQ$. Again, the size is the probability of Type I error and the power is the probability of not making a Type II error. If f is the indicator function of a set A then a randomized test reduces to a test as previously defined.

In Example 1.1.2, the only non-empty non-randomized test of size ≤ 0.05 is $\{1\}$, but this test has very low power 0.002. The empty set gives a test with size 0 but power also 0. A randomized test of size 0.05 is $f(0) = f(1) = 0$ and $f(2) = 1/3$, which has power 0.33, much better than 0.002.

A (randomized) test f of P vs. Q will be said to *dominate* another test g, or to be *as good as* g, if both $\int f\,dP \leq \int g\,dP$ and $\int f\,dQ \geq \int g\,dQ$. If in addition at least one of the two inequalities is strict, $\int f\,dP < \int g\,dP$ or $\int f\,dQ > \int g\,dQ$, then f is said to dominate g *strictly*, or to be *better than* g, or to *improve* g. If a test f dominates a test g then f is as good as g or better, both as to size (at P) and power (at Q).

If P is true, then the average or expected loss in using the test f, called the *risk* $r(f,P)$, is $L_{PQ}\int f\,dP$. If Q is true, the average loss in using f is the risk $r(f,Q) = L_{QP}\int 1-f\,dQ$. Note that when P and Q are interchanged, so are f and $1-f$.

For any prior π, the overall average risk for f, called the *Bayes risk*, is

$$r(f) \;=\; \pi(P)r(f,P) + \pi(Q)r(f,Q).$$

If f dominates g, then $r(f) \leq r(g)$, and if the domination is strict, then $r(f) < r(g)$. So, although the notion of (strict) domination is defined without reference to losses or priors, the notion that a test f which dominates g is "better" than g does hold in terms of losses and priors.

Randomized tests are very rarely used in applications. An experimental scientist would not want to make a decision based on the random variable Y extraneous to (independent of) the actual experiment, if $0 < f(x) < 1$. But, randomized tests provide a way to formulate the merits of tests based on the likelihood ratio while avoiding the pitfalls indicated by Example 1.1.2. It turns out that for the best tests, randomization ($0 < f(x) < 1$) only occurs when the likelihood ratio $R_{Q/P}(x)$ has one value c, depending on the test (Theorem 1.1.3 below). Also, in the Bayes case, what happens when $R_{Q/P} = c$ doesn't affect the Bayes risk (Remark 1.1.9).

A randomized test g will be called *inadmissible* if some other randomized test f strictly dominates g. If there is no such f, then g is called *admissible*. The following main theorem characterizes admissible tests of P vs. Q:

1.1.3 Theorem (Neyman-Pearson Lemma). For any two different probability measures P and Q on $(X,\mathcal{B})$, a randomized test f of P vs. Q is admissible if and only if

$$\begin{gathered}\text{there is some } c \text{ with } 0 \leq c \leq \infty \text{ such that } f(x) = 1 \text{ for } (P+Q)\text{-almost} \\ \text{all } x \text{ satisfying } R_{Q/P}(x) > c \text{ or } R_{Q/P}(x) = +\infty, \text{ and } f(x) = 0 \\ \text{for } (P+Q)\text{-almost all } x \text{ satisfying } R_{Q/P}(x) < c \text{ or } R_{Q/P}(x) = 0.\end{gathered} \tag{1.1.4}$$

Every randomized test g is dominated by a randomized test f satisfying (1.1.4) where for some measurable function h, $f(x) \equiv h(R_{Q/P}(x))$ with $0 \leq h(y) \leq 1$ for all y.

From Theorem 1.1.3, it follows directly that:

1.1.5 Corollary. For any $P \neq Q$ on $(X,\mathcal{B})$, the non-randomized tests $\{R_{Q/P}(x) > b\}$ and $\{R_{Q/P}(x) \geq c\}$ are admissible whenever $0 \leq b < \infty$ and $0 < c \leq \infty$.

Proof of Theorem 1.1.3. For any c with $0 \le c \le +\infty$ and γ with $0 \le \gamma \le 1$, let $f_{c,\gamma}(x) := h_{c,\gamma}(R_{Q/P}(x))$ where $h_{c,\gamma}(y) = 1$ for $y > c$, γ for $y = c$, and 0 for $y < c$. Note that each such $f_{c,\gamma}$ is a randomized test. Then we have

1.1.6 Lemma. For any $P \neq Q$ on X and $0 \le \alpha \le 1$, there is some $f_{c,\gamma}$ with size α.

Proof. If $\alpha = 0$ take $c := c(0) := +\infty$ and $\gamma := \gamma(0) := 1$. (Any γ would give size 0, but $\gamma = 1$ will be useful later for admissibility.) If $\alpha = 1$ let $c := c(1) := 0$ and $\gamma := \gamma(1) := 1$. (The latter is needed to get size 1 if $P(R_{Q/P} = 0) > 0$, although then the test won't be admissible.)

If $0 < \alpha < 1$ let $c := \inf\{t : P(R_{Q/P} > t) \le \alpha\}$. Then $0 \le c \le +\infty$, $P\{R_{Q/P} > c\} \le \alpha$ and $P(R_{Q/P} > t) > \alpha$ for any $t < c$, so $P(R_{Q/P} \ge c) \ge \alpha$. If $P(R_{Q/P} = c) = 0$, any γ will work, say $\gamma = 1$. Otherwise let

$$\gamma := \gamma(\alpha) := (\alpha - P\{R_{Q/P} > c\})/P(R_{Q/P} = c),$$

noting that

$$0 \le \alpha - P(R_{Q/P} > c) \le P(R_{Q/P} \ge c) - P(R_{Q/P} > c) = P(R_{Q/P} = c),$$

so $0 \le \gamma \le 1$. In each case, $f_{c,\gamma}$ has size α. □

Let $c := c(\alpha) := c(\alpha, P, Q)$ as defined in the last proof. Let g be any randomized test and let G be the randomized test which equals 1 if $R_{Q/P} = +\infty$, g if $0 < R_{Q/P} < \infty$ and 0 if $R_{Q/P} = 0$. Any change from $g < 1$ to $G = 1$ on the set where $R_{Q/P} = +\infty$ can only increase the power of the test, without increasing the size since $P = 0$ on that set. Likewise, any change where $R_{Q/P} = 0$ from $g > 0$ to $G = 0$ can only decrease the size without loss of power. So G dominates g, strictly unless $g = G$ almost everywhere for $P + Q$.

1.1.7 Lemma. Let g be any randomized test of P vs. Q. Let its size be α. Then $f := f_{c,\gamma}$ dominates g for $c := c(\alpha)$ and $\gamma = \gamma(\alpha)$.

Proof. The two tests both have size α, so it needs to be shown that f has at least as large power as g. If $c = +\infty$, then $\alpha = 0$ and $\gamma = 1$. So $g = 0$ almost everywhere for P, and so almost everywhere for Q on the set where $R_{Q/P} < \infty$. So $g \le f$ almost everywhere for $P + Q$, and f does dominate g. So we can assume $c < \infty$. Now

$$\{f > g\} \subset \{f > 0\} \subset \{R_{Q/P} \ge c\} \quad \text{and} \quad \{f < g\} \subset \{f < 1\} \subset \{R_{Q/P} \le c\},$$

so $J(x) := (f - g)(x) \cdot (R_{Q/P}(x) - c) \ge 0$ for all x, where (in this case) $0 \cdot \infty$ is taken to be 0. Let $R := R_{Q/P}$. Then

$$0 \le \int_{R<\infty} J \, dP = \int_{R<\infty} (f - g)(dQ - c dP), \quad \text{so}$$

$$\int_{R<\infty} f - g \, dQ \ \ge \ c \int_{R<\infty} f - g \, dP.$$

On the set where $R = +\infty$, $f = 1 \geq g$, so

$$\int_{R=\infty} f - g\, dQ \;\geq\; 0 \;=\; c \int_{R=\infty} f - g\, dP \quad \text{and} \quad \int f - g\, dQ \;\geq\; c \int f - g\, dP \;=\; 0$$

since the two tests have the same size. So $\int f\, dQ \geq \int g\, dQ$. □

Now, to continue the proof of Theorem 1.1.3, let $J(\cdot)$ be defined as in the last proof. If $P(J > 0) > 0$, then since $P(R = \infty) = 0$, f would have strictly larger power than g and g would be inadmissible. So if g is admissible, then $g = f$ almost everywhere for P on the set where $R \neq c$. As shown above, $g = G = 0$ for P- and so $(P+Q)$-almost all x such that $R_{Q/P}(x) = 0$, and $g = G = 1$ for Q- and so $(P+Q)$-almost all x such that $R_{Q/P}(x) = +\infty$. It follows that admissible randomized tests all satisfy (1.1.4), while Lemma 1.1.7 proved the last statement in the Theorem. What is left is to show that any randomized test f satisfying (1.1.4) for some c is admissible (whether or not f is of the form $f_{c,\gamma}$).

If $c = +\infty$, then the size of f is 0 and by (1.1.4), $f = f_{\infty,1}$ almost everywhere for $P+Q$. By the proof of Lemma 1.1.7, f dominates any other test of size 0, so f is admissible. If $c = 0$, then f has power 1, and by (1.1.4), $f = 1_{R>0}$ almost everywhere for $P+Q$. A test g of smaller size must have smaller integral for P over $\{R > 0\}$, so $\mu(\{g < 1\} \cap \{R > 0\}) > 0$ for $\mu = P$ and thus for $\mu = Q$, so the power of g is less than 1. Thus g doesn't dominate f and f is admissible. So we can assume $0 < c < \infty$.

If f is not admissible, let F be a randomized test which dominates f strictly. Let α be the size of F and let $b = c(\alpha)$, $\gamma = \gamma(\alpha)$. So by Lemma 1.1.7, $f_{b,\gamma}$ is a test of size α which dominates F and f. If $P(R = c) > 0$, let $z \;:=\; E_P(f|R = c) \;:=\; \int_{R=c} f\, dP/P(R = c)$. Then also $z = E_Q(f|R = c)$ since on $\{R = c\}$, $Q = cP$. Otherwise let $z = 0$. Then f has the same size and power as $f_{c,z}$.

Say $f \leq g$ if $f(x) \leq g(x)$ for all x. Then the set of all functions $f_{a,t}$ is linearly ordered since $f_{a,t} \leq f_{d,u}$ if $d < a$ or if $d = a$ and $t \leq u$. If $f_{b,\gamma} \leq f_{c,z}$, then the power of $f_{b,\gamma}$ must equal that of $f_{c,z}$ since it is larger or equal. So $f_{b,\gamma} = f_{c,z}$ almost everywhere for Q, and so also for P since $b \geq c > 0$. On the other hand if $f_{c,z} \leq f_{b,\gamma}$ then the sizes of these two tests must be the same, so the functions must be equal almost everywhere for P, and so also for Q since $b \leq c < \infty$. In either case, the randomized tests $f_{b,\gamma}$ and $f_{c,z}$ are equal almost everywhere for $P+Q$, contradicting the assumption that F and so $f_{b,\gamma}$ dominated f and so $f_{c,z}$ strictly, proving Theorem 1.1.3. □

A randomized test f of P vs. Q will be called a *Bayes* test for a given prior π and losses L_{PQ}, L_{QP} if f minimizes the Bayes risk $r(f)$. Such tests are of the Neyman-Pearson form with c as follows:

1.1.8 Theorem. A randomized test f is Bayes for given losses $L_{PQ} > 0 < L_{QP}$ and prior π with $0 < \pi(P) = 1 - \pi(Q) < 1$ if and only if f satisfies (1.1.4) for

$$c \;:=\; c_{PQ} \;:=\; c_{\pi,L} \;:=\; \pi(P)L_{PQ}/(\pi(Q)L_{QP}).$$

Proof. It's easily seen that if some F strictly dominates f, then F has smaller Bayes risk than f. So a Bayes test f must be admissible, and so satisfy (1.1.4), which implies that

$f = 0$ almost surely for P where $R = 0$ and $f = 1$ almost surely for Q where $R = \infty$. Assuming f is Bayes, we need to find it on the set $C := \{0 < R < \infty\}$. A Bayes test f must minimize

$$\int_C \pi(P)L_{PQ}f(x) + (1 - \pi(P))L_{QP}(1 - f(x))R(x)dP(x)$$

or equivalently, minimize

$$\int_C f(x)\{\pi(P)L_{PQ} - \pi(Q)L_{QP}R(x)\}dP(x)$$

where $0 \le f(x) \le 1$. So we must have $f(x) = 0$ if $R(x) < c$ and $f(x) = 1$ if $R(x) > c$. (The behavior of f when $R = c$ is unrestricted.) So f satisfies (1.1.4) for $c = c_{PQ}$. The same calculation shows that conversely, any randomized test f satisfying (1.1.4) for $c = c_{PQ}$ is Bayes. □

1.1.9 Remark. It follows from Theorem 1.1.8 that under the given conditions there is no need for randomization to get a Bayes test for any given prior and losses: both $\{R_{Q/P} > c_{PQ}\}$ and $\{R_{Q/P} \ge c_{PQ}\}$ give non-randomized Bayes tests.

In Example 1.1.2, the following non-randomized tests are admissible, and are Bayes for the given values of $c_{\pi,L} = c_{PQ}$:
$\{0, 1, 2\}$ (always choose Q) is Bayes for $0 < c_{\pi,L} \le 0.01$;
$\{1, 2\}$ (choose P only if $x = 0$) is Bayes for $0.01 \le c_{\pi,L} \le 0.04$;
$\{2\}$ (choose Q only if $x = 2$) is Bayes for $0.04 \le c_{\pi,L} \le 6.6$;
The empty set (always choose P) is Bayes for $6.6 \le c_{\pi,L} < \infty$.

Thus, $\{2\}$ is the preferred test if the losses L_{PQ} and L_{QP} are not too different and neither are the priors $\pi(P)$ and $\pi(Q)$. If $c_{\pi,L}$ is < 0.01 or > 6.6 then we need more than one observation to have an actual choice between P and Q. If $c > 6.6$ and most of many observations equal 2 then we can choose Q. Likewise, if $c < 0.01$ and most of many observations equal 0 then we can choose P.

PROBLEMS

1. Suppose we have two independent observations X_1, X_2 in $\mathbb{R}$ with distribution μ where either $\mu = P = N(0,1)$ or $\mu = Q = N(0,4)$. In $\mathbb{R}^2$, find a region giving an admissible test of $P \times P$ vs. $Q \times Q$ for $c = 1$. What are the size and power of this test? Hint: use polar coordinates.

2. Let P have density e^{-x} on $[0, \infty)$ (so P is a standard exponential distribution) and let Q be the distribution of $X + 1$ where X has distribution P.

   (a) What is the maximum power of a test of P vs. Q with size $\alpha \le 0.05$?

   (b) Answer the same question if we have n independent observations each with distribution $\mu = P$ or Q, so that P, Q are replaced by P^n, Q^n.

3. In Example 1.1.2, if $0 < \pi(P) = 1 - \pi(Q) < 1$, $0 < L_{PQ} < \infty$, and $0 < L_{QP} < \infty$,

   (a) Under what conditions on $\pi(P)$, L_{PQ}, and L_{QP} is $\{0\}$ a better test of P vs. Q than $\{2\}$ is?

   (For a test set A, we choose Q and reject P if the observation is in A, otherwise we choose [don't reject] P.)

   (b) Under the conditions found in (a), what is the Bayes test or decision rule (i.e. the rule with minimum risk) for deciding between P and Q based on one observation in the given sample space $\{0, 1, 2\}$?

4. Prove or disprove: let the set A define an admissible, non-randomized test of P vs. Q. Then for any other set B also defining such a test, and of the same size as for A, the two sets A and B differ only by a set of measure 0 for $P + Q$.

5. A statistic x is used in testing for a disease D. Let $P = N(2, 0.25)$ be the distribution of x for those without D and $Q = N(5, 1)$ be the distribution for those with D. Suppose that the losses are $L_{QP} =$ \$600,000 and $L_{PQ} =$\$500. It is known that about 1 percent of the population being tested has D. Find a test which minimizes the Bayes risk. Hint: the answer is a union of two half-lines. One is negligible. Why?

NOTES TO SECTION 1.1

The original publication of the Neyman-Pearson Lemma was by Neyman and Pearson (1933). Egon Pearson was the son of Karl Pearson, an even more prominent statistician, inventor of the chi-squared test. The proof given above is based on Lehmann (1986, Sec. 3.2). Theorem 1.1.8 appears in Cox and Hinkley (1974, p. 419).

Ferguson (1967) uses the terms that f is "as good as" or "better than" g. Bickel and Doksum (1977) said f "improves" g; in their second edition (2001) I didn't find that terminology or any replacement for it.

REFERENCES

Bickel, Peter J., and Kjell A. Doksum (1977, 2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. 2d. ed. vol. 1 Prentice Hall, Upper Saddle River, NJ; 1st ed. Holden-Day, San Francisco.

Cox, David R., and David V. Hinkley (1974). *Theoretical Statistics.* Chapman and Hall, London.

Ferguson, Thomas S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York.

Lehmann, Erich L. (1986). *Testing Statistical Hypotheses*, 2d ed. Wiley, New York.

Neyman, Jerzy, and Egon S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. London, Ser. A* **231**, 289-337.

RAP = Dudley, R. M. (2002). *Real Analysis and Probability*, 2d edition. Cambridge University Press.