1.3 Bayes decision theory. The distinguishing feature of *Bayesian* statistics is that a probability distribution π, called a *prior*, is given on the parameter space $(\Theta, \mathcal{T})$. Sometimes, priors are also considered which may be infinite, such as Lebesgue measure on the whole real line, but such priors will not be treated here at least for the time being.

A Bayesian statistician chooses a prior π based on whatever information on the unknown θ is available in advance of making any observations in the current experiment. In general, no definite rules are prescribed for choosing π. Priors are often useful as technical tools in reaching non-Bayesian conclusions such as admissibility in Theorems 1.2.5 and 1.2.6.

Bayes decision rules were defined near the end of the last section as rules which minimize the Bayes risk and for which the risk is finite. Bayes tests of P vs. Q, treated in Theorem 1.1.8, are a special case of Bayes decision rules. We saw in that case that Bayes rules need not be randomized (Remark 1.1.9). The same is true quite generally in Bayes decision theory: if, in a given situation, it is Bayes to choose at random among two or more possible decisions, then the decisions must have equal risks (conditional on the observations) and we may as well just take one of them. Theorem 1.3.1 will give a more precise statement.

In game theory, randomization is needed to have a strategy that is optimal even if the opponent knows it and can choose a strategy accordingly. If one knows the opponent's strategy then it is not necessary to randomize. Sometimes, statistical decision theory is viewed as a game against an opponent called "Nature." Unlike an opponent in game theory, "Nature" is viewed as neutral, not trying to win the game. Assuming a prior, as in Bayes decision theory, is to assume in effect that "Nature" follows a certain strategy.

In showing that randomization isn't needed, it will be helpful to formulate randomization in a fuller way, where we not only choose a probability distribution over the possible actions, but then also choose an action according to that distribution, in a measurable way, as follows:

Definition. A randomized decision rule $d: X \to D_{\mathcal{E}}$ is *realizable* if there is a probability space $(\Omega, \mathcal{F}, \mu)$ and a jointly measurable function $\delta: X \times \Omega \to A$ such that for each x in X, $\delta(x, \cdot)$ has distribution $d(x)$, in other words $d(x)$ is the image measure of μ by $\delta(x, \cdot)$, $d(x) = \mu \circ \delta(x, \cdot)^{-1}$.

For example, a randomized test as in Sec. 1.1 is always a realizable rule, where we can take Ω as the interval $[0, 1]$ with Lebesgue measure and let $\delta(x, t) = d_Q$ if $t \leq f(x)$ and d_P otherwise.

It is shown in the next section that decision rules are realizable under conditions wide enough to cover a great many cases, for example whenever the action space is a subset of a space $\mathbb{R}^k$ with Borel σ-algebra. It will be shown next that randomization is unnecessary for realizable Bayes rules. The idea is that the Bayes risk of a realizable randomized Bayes rule $d(\cdot)$ is an average of Bayes risks of non-randomized rules $\delta(\cdot, \omega)$. Since a Bayes rule has minimum Bayes risk, the risks of $\delta(\cdot, \omega)$ are no smaller, so they must almost all be equal to that of $d(\cdot)$. Then such non-randomized $\delta(\cdot, \omega)$ for fixed ω are Bayes rules.

1.3.1 Theorem. For any decision problem for a measurable family $\{P_\theta,\ \theta \in \Theta\}$ and prior π, if there is a realizable Bayes randomized decision rule d, then there is a non-randomized Bayes decision rule.

Proof. First, here is a helpful technical fact:

1.3.2 Lemma. For any measurable family $\{P_\theta,\ \theta \in \Theta\}$ and nonnegative, jointly measurable function $f : \langle \theta, x, \omega \rangle \mapsto f(\theta, x, \omega)$, the function g defined by $g(\theta, \omega) := \int f(\theta, x, \omega) dP_\theta(x)$ is jointly measurable.

Proof. If $f(\theta, x, \omega) = 1_T(\theta) 1_B(x) 1_F(\omega)$ for some $T \in \mathcal{T}$, $B \in \mathcal{B}$ and $F \in \mathcal{F}$, then $g(\theta, \omega) = P_\theta(B) 1_T(\theta) 1_F(\omega)$ is measurable in $\langle \theta, \omega \rangle$ since $\theta \mapsto P_\theta(B)$ is measurable by assumption. The rest of the proof of the Lemma is like that of Prop. 1.2.4. □

Now to prove Theorem 1.3.1, take $(\Omega, \mathcal{F}, \mu)$ and $\delta(\cdot, \cdot)$ as in the definition that d is realizable. For each fixed $\omega \in \Omega$, $\delta(\cdot, \omega)$ is a non-randomized decision rule. So $r(\pi, \delta(\cdot, \omega)) \geq r(\pi, d)$ since d is Bayes for π. Also, writing $\nu(da) := d\nu(a)$ for a measure ν,

$$r(\pi, d) = \int r(\theta, d) d\pi(\theta) = \int \int r(\theta, d(x)) dP_\theta(x) d\pi(\theta) \quad \text{(by the definitions)}$$

$$= \int \int \int L(\theta, a) d(x)(da) dP_\theta(x) d\pi(\theta) = \int \int \int L(\theta, \delta(x, \omega)) d\mu(\omega) dP_\theta(x) d\pi(\theta)$$

by the image measure theorem, e.g. RAP, 4.1.11. So by the Tonelli-Fubini theorem for nonnegative measurable functions, twice, and the measurability shown in Lemma 1.3.2, we get

$$r(\pi, d) = \int \int \int L(\theta, \delta(x, \omega)) dP_\theta(x) d\pi(\theta) d\mu(\omega) = \int r(\pi, \delta(\cdot, \omega)) d\mu(\omega).$$

Thus, $r(\pi, \delta(\cdot, \omega)) = r(\pi, d)$ for μ-almost all ω, and so for some ω, providing a Bayes non-randomized decision rule $\delta(\cdot, \omega)$. □

If every randomized rule is realizable, as is shown in the next section under conditions given there, then Theorem 1.3.1 shows that the non-randomized rules form an essentially complete class, as defined in Sec. 1.2. It will also be shown in Sec. 2.2 below that non-randomized rules are (essentially) complete under some other conditions.

Definition. A family $\{P_\theta,\ \theta \in \Theta\}$ of laws on a measurable space $(X, \mathcal{B})$ will be called *dominated* if for some σ-finite measure v, each law P_θ is absolutely continuous with respect to v, in other words for any $A \in \mathcal{B}$, $v(A) = 0$ implies $P_\theta(A) = 0$ for all θ.

Often, v would be Lebesgue measure on $\mathbb{R}^k$; or, if the measures were all concentrated on a countable set such as the integers, v would be counting measure (the measure giving mass 1 to each point) on the set.

If P_θ is absolutely continuous with respect to v, then by the Radon-Nikodym theorem (RAP, 5.5.4), it has a density or Radon-Nikodym derivative $f(\theta, x) := (dP_\theta/dv)(x)$. A σ-algebra $\mathcal{B}$ is called *countably generated* if there is a countable subcollection $\mathcal{C} \subset \mathcal{B}$ such that $\mathcal{B}$ is the smallest σ-algebra including $\mathcal{C}$. In any separable metric space, the Borel σ-algebra is countably generated (taking $\mathcal{C}$ as the set of balls with rational radii and centers in a countable dense set). In the great majority of applications of statistics, sample spaces are separable metric spaces, in fact Euclidean spaces $\mathbb{R}^k$. At any rate, from here on it will be assumed that $\mathcal{B}$ *is countably generated*, unless something to the contrary is stated.

1.3.3 Theorem. If $\{P_\theta,\ \theta \in \Theta\}$ is a dominated, measurable family on a sample space $(X, \mathcal{B})$, for a parameter space $(\Theta, \mathcal{T})$ and a σ-finite measure v, then the density function $f(\theta, x) = (dP_\theta/dv)(x)$ can be taken to be jointly measurable in θ and x.

Proof. Let $\mathcal{B}_r,\ r = 1, 2, \ldots$, be an increasing sequence of finite Boolean algebras of subsets of X whose union generates $\mathcal{B}$. (Such algebras exist by the blanket assumption that $\mathcal{B}$ is countably generated.) There is a probability measure Q equivalent to (mutually absolutely continuous with) v: to see this, let X be a union of disjoint measurable sets A_j with $0 < v(A_j) < \infty$, and for any $B \in \mathcal{B}$ let $Q(B) = \sum_{j=1}^\infty v(B \cap A_j)/(2^j v(A_j))$. So we can assume that v is a probability measure.

For each θ, let $g(\theta, \cdot) := dP_\theta/dv$. Let $g_r(\theta, \cdot)$ be the conditional expectation of $g(\theta, \cdot)$ given $\mathcal{B}_r$ for v, $g_r(\theta, \cdot) := E(g(\theta, \cdot)|\mathcal{B}_r)$. This can be defined in either of two ways. One is that since P_θ remains absolutely continuous with respect to v if both are restricted to $\mathcal{B}_r$, and $g_r(\theta, \cdot) = dP_\theta/dv$ (Radon-Nikodym derivative) for these restrictions to $\mathcal{B}_r$. The other is that $\mathcal{B}_r$ is generated by a finite collection of *atoms*, which are non-empty sets $A \in \mathcal{B}_r$ of which no proper, non-empty subset belongs to $\mathcal{B}_r$. Then for x in such an atom A, $g_r(\theta, x) = P_\theta(A)/v(A)$, or if $v(A) = 0$ then let $g_r(\theta, x) = 0$. Let B_{ri} for $i = 1, \ldots, I(r)$ be the atoms of $\mathcal{B}_r$. Then since $\{P_\theta,\ \theta \in \Theta\}$ is measurable, for each fixed x, $g_r(\cdot, x)$ is measurable. There are only finitely many possibilities for this function, each for x in a measurable set B_{ri}, so g_r is jointly measurable in θ and x.

Here a fact from probability theory will be used: for each fixed θ, the sequence $g_r(\theta, \cdot)$ of functions on X is a right-closed martingale (RAP, p. 283), with $g_\infty := g$, and $g_r(\theta, x) \to g(\theta, x)$ as $r \to \infty$ for P_θ-almost all x.

The set on which a sequence of measurable real-valued functions converges is measurable (RAP, proof of Theorem 4.2.5). Let $f(\theta, x) := \lim_{r\to\infty} g_r(\theta, x)$ whenever the limit exists and $f(\theta, x) = 0$ otherwise. Then f is jointly measurable and for each θ, $f(\theta, x) = g(\theta, x)$ almost surely for P_θ, so $f(\theta, \cdot)$ is a density of P_θ with respect to v. □

Under the hypotheses of Theorem 1.3.3, it will be assumed from here on that $f(\theta, x)$ *is jointly measurable* in θ and x.

If $\{P_\theta,\ \theta \in \Theta\}$ is a dominated, measurable family of laws on x, with jointly measurable densities $q(\theta, x)$ with respect to some measure v, then the family of laws for n i.i.d. observations, $\{P_\theta^n :\ \theta \in \Theta\}$ on X^n, is clearly dominated and measurable, with jointly measurable densities $f(\theta, x) = \prod_{j=1}^n q(\theta, x_j)$. For a dominated, measurable family and for a fixed x, $f(\cdot, x)$ is a function on Θ called the *likelihood function*. The *posterior distribution* on Θ given x is the law π_x having density with respect to π given by $f(\cdot, x)/\int f(\theta, x)d\pi(\theta)$, provided that the integral in the denominator is strictly positive and finite. In other words,

for any measurable set C of parameters,

$$\pi_x(C) \;=\; \int_C f(\theta,x)d\pi(\theta)/\int f(\theta,x)d\pi(\theta). \tag{1.3.4}$$

Here the denominator is a measurable function of x by joint measurability. By the Tonelli-Fubini theorem, $\int\int f(\theta,x)d\pi(\theta)dv(x) = 1$, so $\int f(\theta,x)d\pi(\theta) < \infty$ for v-almost all x. If x is such that $\int f(\theta,x)d\pi(\theta) = 0$, then the posterior given x is not defined. Observing such an x indicates that the prior and/or likelihood function are incorrectly specified. If before taking the observation the (Bayesian) statistician believed that θ had the prior distribution π, then after observing x the distribution of θ becomes π_x.

Next, π and $\{P_\theta,\ \theta\in\Theta\}$ give a joint distribution for θ and x:

1.3.5 Proposition. For any measurable family $\{P_\theta,\ \theta\in\Theta\}$ and prior π on $(\Theta,\mathcal{T})$, there is a probability distribution $\Pr$ on $(\Theta\times X,\mathcal{T}\otimes\mathcal{B})$ for which the marginal distribution on Θ is π and for each θ, the conditional distribution of x is P_θ.

Proof. For any $A\in\mathcal{T}\otimes\mathcal{B}$, let $\Pr(A) := \int\int 1_A(\theta,x)dP_\theta(x)d\pi(\theta)$ if the integrals are defined. The collection of all sets A for which the integrals are defined contains all sets $C\times B$ for $C\in\mathcal{T}$ and $B\in\mathcal{B}$. Thus it contains all measurable sets, as in the construction of product measures (RAP, Sec. 4.4). So $\Pr$ is well-defined and by monotone convergence is a countably additive probability measure on $\mathcal{T}\otimes\mathcal{B}$. Clearly, π is the marginal distribution of θ for $\Pr$ and P_θ is a conditional distribution of x given θ. □

The marginal distribution of x for $\Pr$, namely the law γ on X having density $\int f(\theta,x)d\pi(\theta)$ with respect to v, is called the *predictive distribution* of x. For any $B\in\mathcal{B}$ we have

$$\gamma(B) = \int P_\theta(B)d\pi(\theta). \tag{1.3.6}$$

Next is an existence fact for posteriors:

1.3.7 Theorem. For any dominated, measurable family $\{P_\theta,\ \theta\in\Theta\}$ and prior π, we have $0 < \int f(\theta,x)d\pi(\theta) < \infty$ for γ-almost all x, and the posterior π_x is well-defined.

Proof. As noted above, $\int f(\theta,x)d\pi(\theta) < \infty$ for v-almost all x. If $B\in\mathcal{B}$ and $v(B) = 0$, then $P_\theta(B) = 0$ for π-almost all θ, so $\gamma(B) = 0$ by (1.3.6). So "v-almost" implies "γ-almost" all x.

Let $D := \{x:\ \int f(\theta,x)d\pi(\theta) = 0\}$. Then by (1.3.6),

$$\gamma(D) \;=\; \int\int_D f(\theta,x)dv(x)d\pi(\theta) \;=\; \int_D\int f(\theta,x)d\pi(\theta)dv(x) \;=\; 0.$$

So the given inequalities are proved. To finish the proof it will be shown that the posterior distribution π_x doesn't depend on the choice of the σ-finite dominating measure v. More precisely, if v and w are two such measures and $\pi_x^v,\ \pi_x^w$ the corresponding posteriors, it will be shown that $\pi_x^v = \pi_x^w$ for γ-almost all x.

Here $v + w$ will be another dominating measure, and $v + w$ is σ-finite since we can take $A_i \cap B_j$ with $v(A_i) < \infty$ and $w(B_j) < \infty$. So we can replace w by $v + w$. Applying the Radon-Nikodym theorem on sets where v and w are both finite, we get a measurable function $dv/dw := g \geq 0$ such that $v(A) = \int_A g dw$ for all $A \in \mathcal{B}$. We also have

$$\frac{dP_\theta}{dw} = \frac{dP_\theta}{dv} \cdot \frac{dv}{dw}$$

almost everywhere for w. Thus in the definition of posterior, for a given x, both numerator and denominator are multiplied by $g(x)$, so the posterior is unchanged if $g(x) > 0$. The set C where $g = 0$ has $v(C) = 0$ and so $\gamma(C) = 0$ as desired, finishing the proof. □

The *conditional risk* of an action $c \in A$, given x, is defined as

$$r_x(\pi, c) := \int L(\theta, c) d\pi_x(\theta)$$

if π_x exists, as we just saw it does for γ-almost all x. The next fact shows that decision rules are Bayes if they minimize the conditional risk for almost all observations.

1.3.8 Theorem. If for a given measurable family $\{P_\theta,\ \theta \in \Theta\}$, prior π and loss function L, $a(\cdot)$ is a decision rule such that for γ-almost all x,

$$r_x(\pi, a(x)) = \inf\{r_x(\pi, c) :\ c \in A\}, \tag{1.3.9}$$

and if there exists a rule $e(\cdot)$ with finite risk, then $a(\cdot)$ is a Bayes rule, and any Bayes rule $b(\cdot)$ in place of $a(\cdot)$ also satisfies (1.3.9).

Proof. Applying Theorem 1.3.7, let π_x exist and (1.3.9) hold for $x \notin B$ where $\gamma(B) = 0$. Then by (1.3.6), $P_\theta(B) = 0$ for π-almost all θ. From the definitions and the Tonelli-Fubini theorem, for any decision rule $b(\cdot)$,

(1.3.10)
$$r(\pi, b) = \int \int_{X \setminus B} L(\theta, b(x)) f(\theta, x) dv(x) d\pi(\theta) = \int_{X \setminus B} \int L(\theta, b(x)) f(\theta, x) d\pi(\theta) dv(x).$$

Given x, minimizing $\int L(\theta, c) f(\theta, x) d\pi(\theta)$ with respect to c is equivalent to minimizing

$$r_x(\pi, c) = \int L(\theta, c) f(\theta, x) d\pi(\theta) / \int f(\psi, x) d\pi(\psi),$$

since $\int f(\psi, x) d\pi(\psi)$ is strictly positive and finite and doesn't depend on c. So $a(x)$ achieves this minimum for $x \notin B$. Thus for any decision rule $b(\cdot)$,

$$\int L(\theta, a(x)) f(\theta, x) d\pi(\theta) \leq \int L(\theta, b(x)) f(\theta, x) d\pi(\theta) \tag{1.3.11}$$

for $x \notin B$. Taking $\int_{X \setminus B}$ of both sides and applying (1.3.10) gives $r(\theta, a(\cdot)) \leq r(\theta, b(\cdot))$. Taking $b(\cdot) = e(\cdot)$ we see that the minimum risk, which $a(\cdot)$ achieves, is finite. In other words, $a(\cdot)$ is Bayes. If $b(\cdot)$ is also Bayes, then $r(\pi, b(\cdot)) = r(\pi, a(\cdot))$. Thus the inequality in (1.3.11) must be an equality for v-almost and so γ-almost all x, and (1.3.9) must hold for $b(\cdot)$ in place of $a(\cdot)$ and for γ-almost all x. □

If A is finite, then any function on A attains its minimum, so Bayes rules always exist. They may not when A is infinite, as was mentioned in the last section for decision problems without a sample space:

1.3.12 Example. For A infinite, a Bayes rule need not exist even if X is a singleton, say $X = \{0\}$, so that an observation makes no difference and $\{P_\theta,\ \theta \in \Theta\}$ reduces to the single law $\{\delta_0\}$. For example let A be the set of positive integers and let $L(m) := L(\delta_0, m) := 1/m$ for $m = 1, 2, \ldots$. Then the infimum of risks is 0 but it is not attained by any decision rule.

The following will not be hard to prove:

1.3.13 Proposition. For any dominated measurable family $\{P_\theta,\ \theta \in \Theta\}$ of laws on a sample space $(X, \mathcal{B})$ and prior π on Θ, the posterior distribution π_x given x is a conditional distribution of θ for Pr (defined in Proposition 1.3.5) given x.

Proof. From the proof of Proposition 1.3.5, Pr has a density $f(\theta, x)$ with respect to $\pi \times v$. Thus for any $C \in \mathcal{T}$, $\pi(C) = \int_X \int_C f(\theta, x) d\pi(\theta) dv(x)$. Multiplying and dividing by $\int f(\psi, x) d\pi(\psi)$, which by Theorem 1.3.7 is strictly positive and finite for γ-almost all x, we get

$$\pi(C) = \int_X \pi_x(C) \int_\Theta f(\psi, x) d\pi(\psi) dv(x) = \int_X \pi_x(C) d\gamma(x)$$

since γ is the X marginal of Pr. It follows that a conditional distribution of θ given x for Pr is the posterior distribution π_x. □

PROBLEMS

1. If $\mathcal{P} = \{P, Q\}$ as in the Neyman-Pearson situation, and π is a prior with $\pi(P) = p = 1 - q$, find the posterior probabilities given x in terms of p, q and $R_{Q/P}(x)$.
2. Suppose the action space A is countable and there is a Bayes randomized decision rule d such that for each x, we have $d(x)(a) > 0$ for every $a \in A$. Then, show that every randomized decision rule is Bayes.
3. Let X be the Cartesian product of n copies of $\{0, 1\}$ (the vertices of the unit n-cube) and let P_θ be the product of n copies of the law with probability θ at 1 and $1 - \theta$ at 0 for $\theta \in \Theta = [0, 1]$. In other words, suppose we have n independent trials with probability θ of success. Suppose that the prior for θ is the uniform distribution on $0 \leq \theta \leq 1$. If the observations (in other words, the coordinates of "the" observation) consist of k 1's and $n - k$ 0's, find the posterior distribution of θ.