2

Belief, probability and exchangeability

We first discuss what properties a reasonable belief function should have, and show that probabilities have these properties. Then, we review the basic machinery of discrete and continuous random variables and probability distributions. Finally, we explore the link between independence and exchangeability.

2.1 Belief functions and probabilities

At the beginning of the last chapter we claimed that probabilities are a way to numerically express rational beliefs. We do not prove this claim here (see Chapter 2 of Jaynes (2003) or Chapters 2 and 3 of Savage (1972) for details), but we do show that several properties we would want our numerical beliefs to have are also properties of probabilities.

Belief functions

Let F, G, and H be three possibly overlapping statements about the world. For example:

$F =$ { a person votes for a left-of-center candidate }
$G =$ { a person's income is in the lowest 10% of the population }
$H =$ { a person lives in a large city }

Let Be() be a *belief function*, that is, a function that assigns numbers to statements such that the larger the number, the higher the degree of belief. Some philosophers have tried to make this more concrete by relating beliefs to preferences over bets:

- $\text{Be}(F) > \text{Be}(G)$ means we would prefer to bet F is true than G is true.

We also want Be() to describe our beliefs under certain conditions:

- $\text{Be}(F|H) > \text{Be}(G|H)$ means that if we knew that H were true, then we would prefer to bet that F is also true than bet G is also true.

P.D. Hoff, *A First Course in Bayesian Statistical Methods*,
Springer Texts in Statistics, DOI 10.1007/978-0-387-92407-6_2,

- $\text{Be}(F|G) > \text{Be}(F|H)$ means that if we were forced to bet on F, we would prefer to do it under the condition that G is true rather than H is true.

Axioms of beliefs

It has been argued by many that any function that is to numerically represent our beliefs should have the following properties:

B1 $\text{Be}(\text{not } H|H) \leq \text{Be}(F|H) \leq \text{Be}(H|H)$
B2 $\text{Be}(F \text{ or } G|H) \geq \max\{\text{Be}(F|H), \text{Be}(G|H)\}$
B3 $\text{Be}(F \text{ and } G|H)$ can be derived from $\text{Be}(G|H)$ and $\text{Be}(F|G \text{ and } H)$

How should we interpret these properties? Are they reasonable?

B1 says that the number we assign to $\text{Be}(F|H)$, our conditional belief in F given H, is bounded below and above by the numbers we assign to complete disbelief ($\text{Be}(\text{not } H|H)$) and complete belief ($\text{Be}(H|H)$).

B2 says that our belief that the truth lies in a given set of possibilities should not decrease as we add to the set of possibilities.

B3 is a bit trickier. To see why it makes sense, imagine you have to decide whether or not F and G are true, knowing that H is true. You could do this by first deciding whether or not G is true given H, and if so, then deciding whether or not F is true given G and H.

Axioms of probability

Now let's compare **B1, B2** and **B3** to the standard axioms of probability. Recall that $F \cup G$ means "F or G," $F \cap G$ means "F and G" and $\emptyset$ is the empty set.

P1 $0 = \Pr(\text{not } H|H) \leq \Pr(F|H) \leq \Pr(H|H) = 1$
P2 $\Pr(F \cup G|H) = \Pr(F|H) + \Pr(G|H)$ if $F \cap G = \emptyset$
P3 $\Pr(F \cap G|H) = \Pr(G|H) \Pr(F|G \cap H)$

You should convince yourself that a probability function, satisfying **P1, P2** and **P3**, also satisfies **B1, B2** and **B3**. Therefore if we use a probability function to describe our beliefs, we have satisfied the axioms of belief.

2.2 Events, partitions and Bayes' rule

Definition 1 (Partition) *A collection of sets $\{H_1, \ldots, H_K\}$ is a partition of another set $\mathcal{H}$ if*

1. *the events are disjoint, which we write as $H_i \cap H_j = \emptyset$ for $i \neq j$;*
2. *the union of the sets is $\mathcal{H}$, which we write as $\cup_{k=1}^{K} H_k = \mathcal{H}$.*

In the context of identifying which of several statements is true, if $\mathcal{H}$ is the set of all possible truths and $\{H_1, \ldots, H_K\}$ is a partition of $\mathcal{H}$, then exactly one out of $\{H_1, \ldots, H_K\}$ contains the truth.

Examples

- Let $\mathcal{H}$ be someone's religious orientation. Partitions include
  - {Protestant, Catholic, Jewish, other, none};
  - {Christian, non-Christian};
  - {atheist, monotheist, multitheist}.
- Let $\mathcal{H}$ be someone's number of children. Partitions include
  - {0, 1, 2, 3 or more};
  - {0, 1, 2, 3, 4, 5, 6, ...}.
- Let $\mathcal{H}$ be the relationship between smoking and hypertension in a given population. Partitions include
  - {some relationship, no relationship};
  - {negative correlation, zero correlation, positive correlation}.

Partitions and probability

Suppose $\{H_1, \ldots, H_K\}$ is a partition of $\mathcal{H}$, $\Pr(\mathcal{H}) = 1$, and E is some specific event. The axioms of probability imply the following:

$$\textbf{Rule of total probability}: \quad \sum_{k=1}^{K} \Pr(H_k) = 1$$

$$\begin{aligned}\textbf{Rule of marginal probability}: \quad \Pr(E) &= \sum_{k=1}^{K} \Pr(E \cap H_k) \\ &= \sum_{k=1}^{K} \Pr(E|H_k)\Pr(H_k)\end{aligned}$$

$$\begin{aligned}\textbf{Bayes' rule}: \quad \Pr(H_j|E) &= \frac{\Pr(E|H_j)\Pr(H_j)}{\Pr(E)} \\ &= \frac{\Pr(E|H_j)\Pr(H_j)}{\sum_{k=1}^{K}\Pr(E|H_k)\Pr(H_k)}\end{aligned}$$

Example

A subset of the 1996 General Social Survey includes data on the education level and income for a sample of males over 30 years of age. Let $\{H_1, H_2, H_3, H_4\}$ be the events that a randomly selected person in this sample is in, respectively, the lower 25th percentile, the second 25th percentile, the third 25th percentile and the upper 25th percentile in terms of income. By definition,

$$\{\Pr(H_1), \Pr(H_2), \Pr(H_3), \Pr(H_4)\} = \{.25, .25, .25, .25\}.$$

Note that $\{H_1, H_2, H_3, H_4\}$ is a partition and so these probabilities sum to 1. Let E be the event that a randomly sampled person from the survey has a college education. From the survey data, we have

$$\{\Pr(E|H_1), \Pr(E|H_2), \Pr(E|H_3), \Pr(E|H_4)\} = \{.11, .19, .31, .53\}.$$

These probabilities do not sum to 1 - they represent the proportions of people with college degrees in the four different income subpopulations H_1, H_2, H_3 and H_4. Now let's consider the income distribution of the college-educated population. Using Bayes' rule we can obtain

$$\{\Pr(H_1|E), \Pr(H_2|E), \Pr(H_3|E), \Pr(H_4|E)\} = \{.09, .17, .27, .47\},$$

and we see that the income distribution for people in the college-educated population differs markedly from $\{.25, .25, .25, .25\}$, the distribution for the general population. Note that these probabilities do sum to 1 - they are the conditional probabilities of the events in the partition, given E.

In Bayesian inference, $\{H_1, \ldots, H_K\}$ often refer to disjoint hypotheses or states of nature and E refers to the outcome of a survey, study or experiment. To compare hypotheses post-experimentally, we often calculate the following ratio:

$$\begin{aligned}
\frac{\Pr(H_i|E)}{\Pr(H_j|E)} &= \frac{\Pr(E|H_i)\Pr(H_i)/\Pr(E)}{\Pr(E|H_j)\Pr(H_j)/\Pr(E)} \\
&= \frac{\Pr(E|H_i)\Pr(H_i)}{\Pr(E|H_j)\Pr(H_j)} \\
&= \frac{\Pr(E|H_i)}{\Pr(E|H_j)} \times \frac{\Pr(H_i)}{\Pr(H_j)} \\
&= \text{“Bayes factor”} \times \text{“prior beliefs”}.
\end{aligned}$$

This calculation reminds us that Bayes' rule does not determine what our beliefs should be after seeing the data, it only tells us how they should change after seeing the data.

Example

Suppose we are interested in the rate of support for a particular candidate for public office. Let

$\mathcal{H} = \{$ all possible rates of support for candidate A $\}$;
$H_1 = \{$ more than half the voters support candidate A $\}$;
$H_2 = \{$ less than or equal to half the voters support candidate A $\}$;
$E = \{$ 54 out of 100 people surveyed said they support candidate A $\}$.

Then $\{H_1, H_2\}$ is a partition of $\mathcal{H}$. Of interest is $\Pr(H_1|E)$, or $\Pr(H_1|E)/\Pr(H_2|E)$. We will learn how to obtain these quantities in the next chapter.

2.3 Independence

Definition 2 (Independence) *Two events F and G are conditionally independent given H if* $\Pr(F \cap G|H) = \Pr(F|H)\Pr(G|H)$.

How do we interpret conditional independence? By Axiom **P3**, the following is always true:

$$\Pr(F \cap G|H) = \Pr(G|H)\Pr(F|H \cap G).$$

If F and G are conditionally independent given H, then we must have

$$\begin{aligned} \Pr(G|H)\Pr(F|H \cap G) &\overset{\text{always}}{=} \Pr(F \cap G|H) \overset{\text{independence}}{=} \Pr(F|H)\Pr(G|H) \\ \Pr(G|H)\Pr(F|H \cap G) &= \Pr(F|H)\Pr(G|H) \\ \Pr(F|H \cap G) &= \Pr(F|H). \end{aligned}$$

Conditional independence therefore implies that $\Pr(F|H \cap G) = \Pr(F|H)$. In other words, if we know H is true and F and G are conditionally independent given H, then knowing G does not change our belief about F.

Examples

Let's consider the conditional dependence of F and G when H is assumed to be true in the following two situations:

$F =$ { a hospital patient is a smoker }
$G =$ { a hospital patient has lung cancer }
$H =$ { smoking causes lung cancer}

$F =$ { you are thinking of the jack of hearts }
$G =$ { a mind reader claims you are thinking of the jack of hearts }
$H =$ { the mind reader has extrasensory perception }

In both of these situations, H being true implies a relationship between F and G. What about when H is not true?

2.4 Random variables

In Bayesian inference a random variable is defined as an unknown numerical quantity about which we make probability statements. For example, the quantitative outcome of a survey, experiment or study is a random variable before the study is performed. Additionally, a fixed but unknown population parameter is also a random variable.

2.4.1 Discrete random variables

Let Y be a random variable and let $\mathcal{Y}$ be the set of all possible values of Y. We say that Y is discrete if the set of possible outcomes is *countable*, meaning that $\mathcal{Y}$ can be expressed as $\mathcal{Y} = \{y_1, y_2, \ldots\}$.

Examples

- Y = number of churchgoers in a random sample from a population
- Y = number of children of a randomly sampled person
- Y = number of years of education of a randomly sampled person

Probability distributions and densities

The event that the outcome Y of our survey has the value y is expressed as $\{Y = y\}$. For each $y \in \mathcal{Y}$, our shorthand notation for $\Pr(Y = y)$ will be $p(y)$. This function of y is called the *probability density function* (pdf) of Y, and it has the following properties:

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$;
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$.

General probability statements about Y can be derived from the pdf. For example, $\Pr(Y \in A) = \sum_{y \in A} p(y)$. If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned}\Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \sum_{y \in A} p(y) + \sum_{y \in B} p(y).\end{aligned}$$

Example: Binomial distribution

Let $\mathcal{Y} = \{0, 1, 2, \ldots, n\}$ for some positive integer n. The uncertain quantity $Y \in \mathcal{Y}$ has a *binomial distribution with probability* θ if

$$\Pr(Y = y|\theta) = \text{dbinom}(y, n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

For example, if $\theta = .25$ and $n = 4$, we have:

$$\begin{aligned}\Pr(Y = 0|\theta = .25) &= \binom{4}{0} (.25)^0 (.75)^4 = .316 \\ \Pr(Y = 1|\theta = .25) &= \binom{4}{1} (.25)^1 (.75)^3 = .422 \\ \Pr(Y = 2|\theta = .25) &= \binom{4}{2} (.25)^2 (.75)^2 = .211 \\ \Pr(Y = 3|\theta = .25) &= \binom{4}{3} (.25)^3 (.75)^1 = .047 \\ \Pr(Y = 4|\theta = .25) &= \binom{4}{4} (.25)^4 (.75)^0 = .004\,.\end{aligned}$$

Example: Poisson distribution

Let $\mathcal{Y} = \{0, 1, 2, \ldots\}$. The uncertain quantity $Y \in \mathcal{Y}$ has a *Poisson distribution with mean* θ if

$$\Pr(Y = y|\theta) = \text{dpois}(y, \theta) = \theta^y e^{-\theta}/y!.$$

For example, if $\theta = 2.1$ (the 2006 U.S. fertility rate),

$$\begin{aligned}
\Pr(Y = 0|\theta = 2.1) &= (2.1)^0 e^{-2.1}/(0!) = .12 \\
\Pr(Y = 1|\theta = 2.1) &= (2.1)^1 e^{-2.1}/(1!) = .26 \\
\Pr(Y = 2|\theta = 2.1) &= (2.1)^2 e^{-2.1}/(2!) = .27 \\
\Pr(Y = 3|\theta = 2.1) &= (2.1)^3 e^{-2.1}/(3!) = .19 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots
\end{aligned}$$

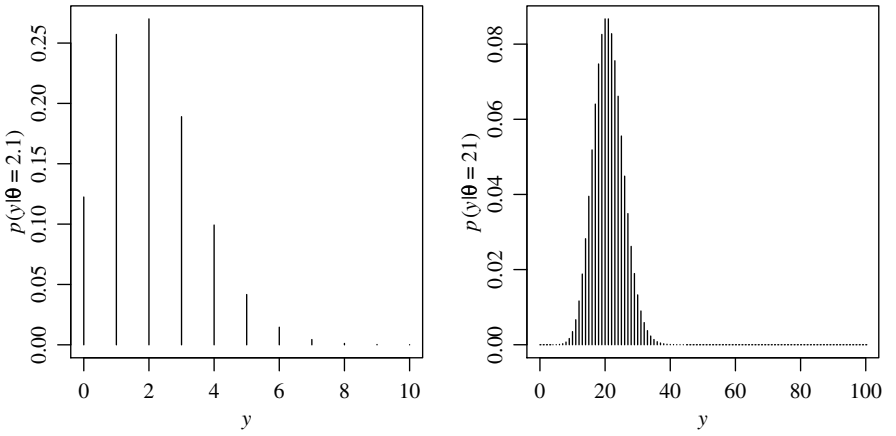

Fig. 2.1. Poisson distributions with means of 2.1 and 21.

2.4.2 Continuous random variables

Suppose that the sample space $\mathcal{Y}$ is roughly equal to $\mathbb{R}$, the set of all real numbers. We cannot define $\Pr(Y \leq 5)$ as equal to $\sum_{y \leq 5} p(y)$ because the sum does not make sense (the set of real numbers less than or equal to 5 is "uncountable"). So instead of defining probabilities of events in terms of a pdf $p(y)$, courses in mathematical statistics often define probability distributions for random variables in terms of something called a *cumulative distribution function*, or cdf:

$$F(y) = \Pr(Y \leq y)\,.$$

Note that $F(\infty) = 1$, $F(-\infty) = 0$, and $F(b) \leq F(a)$ if $b < a$. Probabilities of various events can be derived from the cdf:

- $\Pr(Y > a) = 1 - F(a)$
- $\Pr(a < Y \leq b) = F(b) - F(a)$

If F is continuous (i.e. lacking any "jumps"), we say that Y is a continuous random variable. A theorem from mathematics says that for every continuous cdf F there exists a positive function $p(y)$ such that

$$F(a) = \int_{-\infty}^{a} p(y)\ dy.$$

This function is called the probability density function of Y, and its properties are similar to those of a pdf for a discrete random variable:

1. $0 \leq p(y)$ for all $y \in \mathcal{Y}$;
2. $\int_{y \in \mathbb{R}} p(y)\ dy = 1$.

As in the discrete case, probability statements about Y can be derived from the pdf: $\Pr(Y \in A) = \int_{y \in A} p(y)\ dy$, and if A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned}\Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \int_{y \in A} p(y)\ dy + \int_{y \in B} p(y)\ dy.\end{aligned}$$

Comparing these properties to the analogous properties in the discrete case, we see that integration for continuous distributions behaves similarly to summation for discrete distributions. In fact, integration can be thought of as a generalization of summation for situations in which the sample space is not countable. However, unlike a pdf in the discrete case, the pdf for a continuous random variable is not necessarily less than 1, and $p(y)$ is not "the probability that $Y = y$." However, if $p(y_1) > p(y_2)$ we will sometimes informally say that y_1 "has a higher probability" than y_2.

Example: Normal distribution

Suppose we are sampling from a population on $\mathcal{Y} = (-\infty, \infty)$, and we know that the mean of the population is μ and the variance is σ^2. Among all probability distributions having a mean of μ and a variance of σ^2, the one that is the most "spread out" or "diffuse" (in terms of a measure called entropy), is the normal(μ, σ^2) distribution, having a cdf given by

$$\Pr(Y \leq y | \mu, \sigma^2) = F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}\ dy.$$

Evidently,

$$p(y | \mu, \sigma^2) = \text{dnorm}(y, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}.$$

Letting $\mu = 10.75$ and $\sigma = .8$ ($\sigma^2 = .64$) gives the cdf and density in Figure 2.2. This mean and standard deviation make the median value of e^Y equal to about 46,630, which is about the median U.S. household income in 2005. Additionally, $\Pr(e^Y > 100000) = \Pr(Y > \log 100000) = 0.17$, which roughly matches the fraction of households in 2005 with incomes exceeding $100,000.

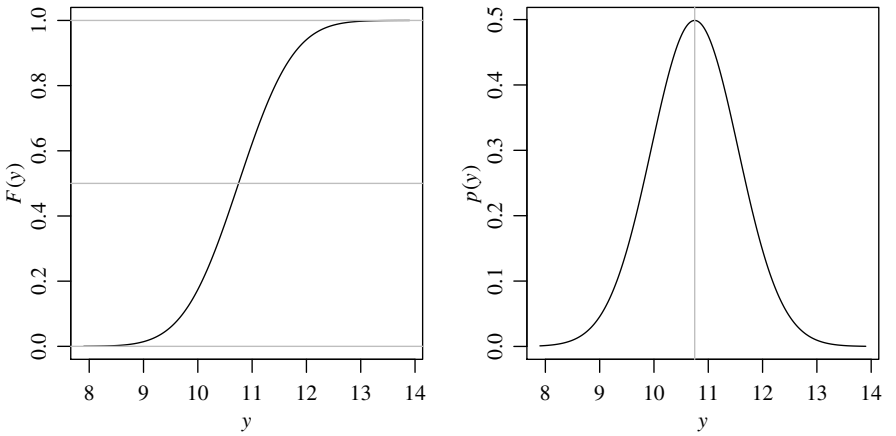

Fig. 2.2. Normal distribution with mean 10.75 and standard deviation 0.8.

2.4.3 Descriptions of distributions

The *mean* or *expectation* of an unknown quantity Y is given by

$\mathrm{E}[Y] = \sum_{y \in \mathcal{Y}} yp(y)$ if Y is discrete;
$\mathrm{E}[Y] = \int_{y \in \mathcal{Y}} yp(y)\, dy$ if Y is continuous.

The mean is the center of mass of the distribution. However, it is not in general equal to either of

the *mode*: "the most probable value of Y," or
the *median*: "the value of Y in the middle of the distribution."

In particular, for skewed distributions (like income distributions) the mean can be far from a "typical" sample value: see, for example, Figure 2.3. Still, the mean is a very popular description of the location of a distribution. Some justifications for reporting and studying the mean include the following:

1. The mean of $\{Y_1, \ldots, Y_n\}$ is a scaled version of the total, and the total is often a quantity of interest.
2. Suppose you are forced to guess what the value of Y is, and you are penalized by an amount $(Y - y_{\text{guess}})^2$. Then guessing $\mathrm{E}[Y]$ minimizes your expected penalty.

3. In some simple models that we shall see shortly, the sample mean contains all of the information about the population that can be obtained from the data.

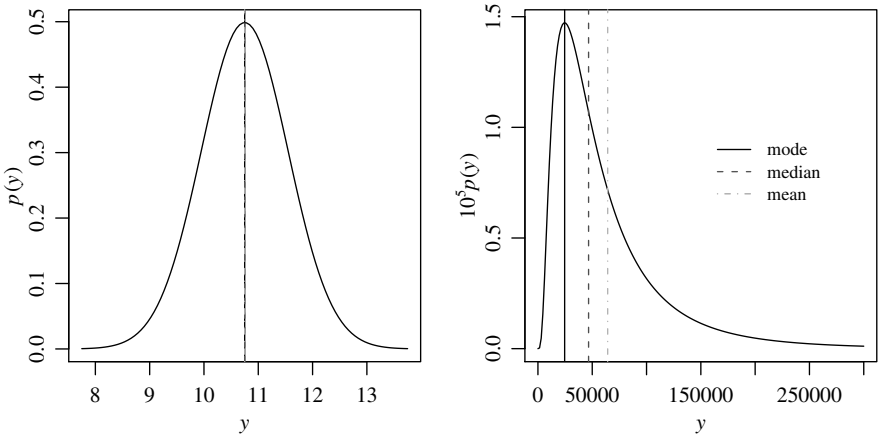

Fig. 2.3. Mode, median and mean of the normal and lognormal distributions, with parameters $\mu = 10.75$ and $\sigma = 0.8$.

In addition to the location of a distribution we are often interested in how spread out it is. The most popular measure of spread is the *variance* of a distribution:

$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{E}[(Y - \mathrm{E}[Y])^2] \\ &= \mathrm{E}[Y^2 - 2Y\mathrm{E}[Y] + \mathrm{E}[Y]^2] \\ &= \mathrm{E}[Y^2] - 2\mathrm{E}[Y]^2 + \mathrm{E}[Y]^2 \\ &= \mathrm{E}[Y^2] - \mathrm{E}[Y]^2.\end{aligned}$$

The variance is the average squared distance that a sample value Y will be from the population mean $\mathrm{E}[Y]$. The *standard deviation* is the square root of the variance, and is on the same scale as Y.

Alternative measures of spread are based on *quantiles*. For a continuous, strictly increasing cdf F, the α-quantile is the value y_α such that $F(y_\alpha) \equiv \Pr(Y \leq y_\alpha) = \alpha$. The *interquartile range* of a distribution is the interval $(y_{.25}, y_{.75})$, which contains 50% of the mass of the distribution. Similarly, the interval $(y_{.025}, y_{.975})$ contains 95% of the mass of the distribution.

2.5 Joint distributions

Discrete distributions

Let

- $\mathcal{Y}_1, \mathcal{Y}_2$ be two countable sample spaces;
- Y_1, Y_2 be two random variables, taking values in $\mathcal{Y}_1, \mathcal{Y}_2$ respectively.

Joint beliefs about Y_1 and Y_2 can be represented with probabilities. For example, for subsets $A \subset \mathcal{Y}_1$ and $B \subset \mathcal{Y}_2$, $\Pr(\{Y_1 \in A\} \cap \{Y_2 \in B\})$ represents our belief that Y_1 is in A and that Y_2 is in B. The *joint pdf* or *joint density* of Y_1 and Y_2 is defined as

$$p_{Y_1Y_2}(y_1, y_2) = \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \text{ for } y_1 \in \mathcal{Y}_1, \ y_2 \in \mathcal{Y}_2.$$

The *marginal density* of Y_1 can be computed from the joint density:

$$\begin{aligned} p_{Y_1}(y_1) &\equiv \Pr(Y_1 = y_1) \\ &= \sum_{y_2 \in \mathcal{Y}_2} \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &\equiv \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1Y_2}(y_1, y_2)\,. \end{aligned}$$

The *conditional density* of Y_2 given $\{Y_1 = y_1\}$ can be computed from the joint density and the marginal density:

$$\begin{aligned} p_{Y_2|Y_1}(y_2|y_1) &= \frac{\Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{\Pr(Y_1 = y_1)} \\ &= \frac{p_{Y_1Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}\,. \end{aligned}$$

You should convince yourself that

$\{p_{Y_1}, p_{Y_2|Y_1}\}$ can be derived from $p_{Y_1Y_2}$,
$\{p_{Y_2}, p_{Y_1|Y_2}\}$ can be derived from $p_{Y_1Y_2}$,
$p_{Y_1Y_2}$ can be derived from $\{p_{Y_1}, p_{Y_2|Y_1}\}$,
$p_{Y_1Y_2}$ can be derived from $\{p_{Y_2}, p_{Y_1|Y_2}\}$,

but

$p_{Y_1Y_2}$ cannot be derived from $\{p_{Y_1}, p_{Y_2}\}$.

The subscripts of density functions are often dropped, in which case the type of density function is determined from the function argument: $p(y_1)$ refers to $p_{Y_1}(y_1)$, $p(y_1, y_2)$ refers to $p_{Y_1Y_2}(y_1, y_2)$, $p(y_1|y_2)$ refers to $p_{Y_1|Y_2}(y_1|y_2)$, etc.

Example: Social mobility

Logan (1983) reports the following joint distribution of occupational categories of fathers and sons:

	son's occupation				
father's occupation	farm	operatives	craftsmen	sales	professional
farm	0.018	0.035	0.031	0.008	0.018
operatives	0.002	0.112	0.064	0.032	0.069
craftsmen	0.001	0.066	0.094	0.032	0.084
sales	0.001	0.018	0.019	0.010	0.051
professional	0.001	0.029	0.032	0.043	0.130

Suppose we are to sample a father-son pair from this population. Let Y_1 be the father's occupation and Y_2 the son's occupation. Then

$$\begin{aligned}\Pr(Y_2 = \text{professional}|Y_1 = \text{farm}) &= \frac{\Pr(Y_2 = \text{professional} \cap Y_1 = \text{farm})}{\Pr(Y_1 = \text{farm})} \\ &= \frac{.018}{.018 + .035 + .031 + .008 + .018} \\ &= .164\,.\end{aligned}$$

Continuous joint distributions

If Y_1 and Y_2 are continuous we start with a cumulative distribution function. Given a continuous joint cdf $F_{Y_1Y_2}(a,b) \equiv \Pr(\{Y_1 \le a\} \cap \{Y_2 \le b\})$, there is a function $p_{Y_1Y_2}$ such that

$$F_{Y_1Y_2}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} p_{Y_1Y_2}(y_1,y_2)\; dy_2 dy_1\,.$$

The function $p_{Y_1Y_2}$ is the joint density of Y_1 and Y_2. As in the discrete case, we have

- $p_{Y_1}(y_1) = \int_{-\infty}^{\infty} p_{Y_1Y_2}(y_2,y_2)\; dy_2$;
- $p_{Y_2|Y_1}(y_2|y_1) = p_{Y_1Y_2}(y_1,y_2)/p_{Y_1}(y_1)$.

You should convince yourself that $p_{Y_2|Y_1}(y_2|y_1)$ is an actual probability density, i.e. for each value of y_1 it is a probability density for Y_2.

Mixed continuous and discrete variables

Let Y_1 be discrete and Y_2 be continuous. For example, Y_1 could be occupational category and Y_2 could be personal income. Suppose we define

- a marginal density p_{Y_1} from our beliefs $\Pr(Y_1 = y_1)$;
- a conditional density $p_{Y_2|Y_1}(y_2|y_1)$ from $\Pr(Y_2 \le y_2|Y_1 = y_1) \equiv F_{Y_2|Y_1}(y_2|y_1)$ as above.

The joint density of Y_1 and Y_2 is then

$$p_{Y_1Y_2}(y_1, y_2) = p_{Y_1}(y_1) \times p_{Y_2|Y_1}(y_2|y_1),$$

and has the property that

$$\Pr(Y_1 \in A, Y_2 \in B) = \int_{y_2 \in B} \left\{ \sum_{y_1 \in A} p_{Y_1Y_2}(y_1, y_2) \right\} \, dy_2.$$

Bayes' rule and parameter estimation

Let

θ = proportion of people in a large population who have a certain characteristic.
Y = number of people in a small random sample from the population who have the characteristic.

Then we might treat θ as continuous and Y as discrete. Bayesian estimation of θ derives from the calculation of $p(\theta|y)$, where y is the observed value of Y. This calculation first requires that we have a joint density $p(y, \theta)$ representing our beliefs about θ and the survey outcome Y. Often it is natural to construct this joint density from

- $p(\theta)$, beliefs about θ;
- $p(y|\theta)$, beliefs about Y for each value of θ.

Having observed $\{Y = y\}$, we need to compute our updated beliefs about θ:

$$p(\theta|y) = p(\theta, y)/p(y) = p(\theta)p(y|\theta)/p(y)\,.$$

This conditional density is called the *posterior density* of θ. Suppose θ_a and θ_b are two possible numerical values of the true value of θ. The posterior probability (density) of θ_a relative to θ_b, conditional on $Y = y$, is

$$\begin{aligned} \frac{p(\theta_a|y)}{p(\theta_b|y)} &= \frac{p(\theta_a)p(y|\theta_a)/p(y)}{p(\theta_b)p(y|\theta_b)/p(y)} \\ &= \frac{p(\theta_a)p(y|\theta_a)}{p(\theta_b)p(y|\theta_b)}\,. \end{aligned}$$

This means that to evaluate the relative posterior probabilities of θ_a and θ_b, we do not need to compute $p(y)$. Another way to think about it is that, as a function of θ,

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

The constant of proportionality is $1/p(y)$, which *could* be computed from

$$p(y) = \int_\Theta p(y, \theta) \, d\theta = \int_\Theta p(y|\theta)p(\theta) \, d\theta$$

giving

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int_\theta p(\theta)p(y|\theta)\ d\theta}.$$

As we will see in later chapters, the numerator is the critical part.

2.6 Independent random variables

Suppose $Y_1, \ldots, Y_n$ are random variables and that θ is a parameter describing the conditions under which the random variables are generated. We say that $Y_1, \ldots, Y_n$ are conditionally independent given θ if for every collection of n sets $\{A_1, \ldots, A_n\}$ we have

$$\Pr(Y_1 \in A_1, \ldots, Y_n \in A_n|\theta) = \Pr(Y_1 \in A_1|\theta) \times \cdots \times \Pr(Y_n \in A_n|\theta).$$

Notice that this definition of independent random variables is based on our previous definition of independent events, where here each $\{Y_j \in A_j\}$ is an event. From our previous calculations, if independence holds, then

$$\Pr(Y_i \in A_i|\theta, Y_j \in A_j) = \Pr(Y_i \in A_i|\theta),$$

so conditional independence can be interpreted as meaning that Y_j gives no additional information about Y_i beyond that in knowing θ. Furthermore, under independence the joint density is given by

$$p(y_1, \ldots, y_n|\theta) = p_{Y_1}(y_1|\theta) \times \cdots \times p_{Y_n}(y_n|\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i|\theta),$$

the product of the marginal densities.

Suppose $Y_1, \ldots, Y_n$ are generated in similar ways from a common process. For example, they could all be samples from the same population, or runs of an experiment performed under similar conditions. This suggests that the marginal densities are all equal to some common density giving

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

In this case, we say that $Y_1, \ldots, Y_n$ are *conditionally independent and identically distributed* (i.i.d.). Mathematical shorthand for this is

$$Y_1, \ldots, Y_n|\theta \sim \text{i.i.d. } p(y|\theta).$$

2.7 Exchangeability

Example: Happiness

Participants in the 1998 General Social Survey were asked whether or not they were generally happy. Let Y_i be the random variable associated with this question, so that

$$Y_i = \begin{cases} 1 & \text{if participant } i \text{ says that they are generally happy,} \\ 0 & \text{otherwise.} \end{cases}$$

In this section we will consider the structure of our joint beliefs about $Y_1, \ldots, Y_{10}$, the outcomes of the first 10 randomly selected survey participants. As before, let $p(y_1, \ldots, y_{10})$ be our shorthand notation for $\Pr(Y_1 = y_1, \ldots, Y_{10} = y_{10})$, where each y_i is either 0 or 1.

Exchangeability

Suppose we are asked to assign probabilities to three different outcomes:

$$\begin{aligned} p(1,0,0,1,0,1,1,0,1,1) &= ? \\ p(1,0,1,0,1,1,0,1,1,0) &= ? \\ p(1,1,0,0,1,1,0,0,1,1) &= ? \end{aligned}$$

Is there an argument for assigning them the same numerical value? Notice that each sequence contains six ones and four zeros.

Definition 3 (Exchangeable) *Let $p(y_1, \ldots, y_n)$ be the joint density of Y_1, $\ldots$, Y_n. If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations π of $\{1, \ldots, n\}$, then $Y_1, \ldots, Y_n$ are* exchangeable.

Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels convey no information about the outcomes.

Independence versus dependence

Consider the following two probability assignments:

$$\begin{aligned} \Pr(Y_{10} = 1) &= a \\ \Pr(Y_{10} = 1 | Y_1 = Y_2 = \cdots = Y_8 = Y_9 = 1) &= b \end{aligned}$$

Should we have $a < b$, $a = b$, or $a > b$? If $a \neq b$ then Y_{10} is NOT independent of $Y_1, \ldots, Y_9$.

Conditional independence

Suppose someone told you the numerical value of θ, the rate of happiness among the 1,272 respondents to the question. Do the following probability assignments seem reasonable?

$$\begin{aligned}\Pr(Y_{10} = 1|\theta) &\stackrel{?}{\approx} \theta \\ \Pr(Y_{10} = 1|Y_1 = y_1, \ldots, Y_9 = y_9, \theta) &\stackrel{?}{\approx} \theta \\ \Pr(Y_9 = 1|Y_1 = y_1, \ldots, Y_8 = y_8, Y_{10} = y_{10}, \theta) &\stackrel{?}{\approx} \theta\end{aligned}$$

If these assignments are reasonable, then we can consider the Y_i's as conditionally independent and identically distributed given θ, or at least approximately so: The population size of 1,272 is much larger than the sample size of 10, in which case sampling without replacement is approximately the same as i.i.d. sampling with replacement. Assuming conditional independence,

$$\begin{aligned}\Pr(Y_i = y_i|\theta, Y_j = y_j, j \neq i) &= \theta^{y_i}(1-\theta)^{1-y_i} \\ \Pr(Y_1 = y_1, \ldots, Y_{10} = y_{10}|\theta) &= \prod_{i=1}^{10} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum y_i}(1-\theta)^{10-\sum y_i}.\end{aligned}$$

If θ is uncertain to us, we describe our beliefs about it with $p(\theta)$, a prior distribution. The marginal joint distribution of $Y_1, \ldots, Y_{10}$ is then

$$p(y_1, \ldots, y_{10}) = \int_0^1 p(y_1, \ldots, y_{10}|\theta)p(\theta)\ d\theta = \int_0^1 \theta^{\sum y_i}(1-\theta)^{10-\sum y_i}p(\theta)\ d\theta.$$

Now consider our probabilities for the three binary sequences given above:

$$\begin{aligned}p(1,0,0,1,0,1,1,0,1,1) &= \textstyle\int \theta^6(1-\theta)^4 p(\theta)\ d\theta \\ p(1,0,1,0,1,1,0,1,1,0) &= \textstyle\int \theta^6(1-\theta)^4 p(\theta)\ d\theta \\ p(1,1,0,0,1,1,0,0,1,1) &= \textstyle\int \theta^6(1-\theta)^4 p(\theta)\ d\theta\end{aligned}$$

It looks like $Y_1, \ldots, Y_n$ are exchangeable under this model of beliefs.

Claim:

If $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.

Proof:

Suppose $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some unknown parameter θ. Then for any permutation π of $\{1, \ldots, n\}$ and any set of values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$,

$$
\begin{aligned}
p(y_1,\ldots,y_n) &= \int p(y_1,\ldots,y_n|\theta)p(\theta)\ d\theta && \text{(definition of marginal probability)}\\
&= \int \left\{\prod_{i=1}^{n} p(y_i|\theta)\right\} p(\theta)\ d\theta && (Y_i\text{'s are conditionally i.i.d.})\\
&= \int \left\{\prod_{i=1}^{n} p(y_{\pi_i}|\theta)\right\} p(\theta)\ d\theta && \text{(product does not depend on order)}\\
&= p(y_{\pi_1},\ldots y_{\pi_n}) && \text{(definition of marginal probability)}\,.
\end{aligned}
$$

2.8 de Finetti's theorem

We have seen that

$$
\left.\begin{array}{l} Y_1,\ldots,Y_n|\theta \text{ i.i.d} \\ \theta \sim p(\theta) \end{array}\right\} \Rightarrow Y_1,\ldots,Y_n \text{ are exchangeable.}
$$

What about an arrow in the other direction? Let $\{Y_1, Y_2, \ldots\}$ be a potentially infinite sequence of random variables all having a common sample space $\mathcal{Y}$.

Theorem 1 (de Finetti) *Let $Y_i \in \mathcal{Y}$ for all $i \in \{1, 2, \ldots\}$. Suppose that, for any n, our belief model for $Y_1, \ldots, Y_n$ is exchangeable:*

$$
p(y_1,\ldots,y_n) = p(y_{\pi_1},\ldots,y_{\pi_n})
$$

for all permutations π of $\{1,\ldots,n\}$. Then our model can be written as

$$
p(y_1,\ldots,y_n) = \int \left\{\prod_{1}^{n} p(y_i|\theta)\right\} p(\theta)\ d\theta
$$

for some parameter θ, some prior distribution on θ and some sampling model $p(y|\theta)$. The prior and sampling model depend on the form of the belief model $p(y_1,\ldots,y_n)$.

The probability distribution $p(\theta)$ represents our beliefs about the outcomes of $\{Y_1, Y_2, \ldots\}$, induced by our belief model $p(y_1, y_2, \ldots)$. More precisely,

> $p(\theta)$ represents our beliefs about $\lim_{n\to\infty} \sum Y_i/n$ in the binary case;
> $p(\theta)$ represents our beliefs about $\lim_{n\to\infty} \sum(Y_i \le c)/n$ for each c in the general case.

The main ideas of this and the previous section can be summarized as follows:

$$
\left.\begin{array}{l} Y_1,\ldots,Y_n|\theta \text{ are i.i.d.} \\ \theta \sim p(\theta) \end{array}\right\} \Leftrightarrow Y_1,\ldots,Y_n \text{ are exchangeable for all } n\,.
$$

When is the condition "$Y_1,\ldots,Y_n$ are exchangeable for all n" reasonable? For this condition to hold, we must have exchangeability and repeatability. Exchangeability will hold if the labels convey no information. Situations in which repeatability is reasonable include the following:

$Y_1, \ldots, Y_n$ are outcomes of a repeatable experiment;
$Y_1, \ldots, Y_n$ are sampled from a finite population with replacement;
$Y_1, \ldots, Y_n$ are sampled from an infinite population without replacement.

If $Y_1, \ldots, Y_n$ are exchangeable and sampled from a finite population of size $N >> n$ without replacement, then they can be modeled as approximately being conditionally i.i.d. (Diaconis and Freedman, 1980).

2.9 Discussion and further references

The notion of subjective probability in terms of a coherent gambling strategy was developed by de Finetti, who is of course also responsible for de Finetti's theorem (de Finetti, 1931, 1937). Both of these topics were studied further by many others, including Savage (Savage, 1954; Hewitt and Savage, 1955).

The concept of exchangeability goes beyond just the concept of an infinitely exchangeable sequence considered in de Finetti's theorem. Diaconis and Freedman (1980) consider exchangeability for finite populations or sequences, and Diaconis (1988) surveys some other versions of exchangeability. Chapter 4 of Bernardo and Smith (1994) provides a guide to building statistical models based on various types of exchangeability. A very comprehensive and mathematical review of exchangeability is given in Aldous (1985), which in particular provides an excellent survey of exchangeability as applied to random matrices.

http://www.springer.com/978-0-387-92299-7

A First Course in Bayesian Statistical Methods
Hoff, P.D.
2009, IX, 271 p., Hardcover
ISBN: 978-0-387-92299-7