16.1: Conditional Independence, Given a Random Vector

In the unit on Conditional Independence , the concept of conditional independence of events is examined and used to model a variety of common situations. In this unit, we investigate a more general concept of conditional independence, based on the theory of conditional expectation. This concept lies at the foundations of Bayesian statistics, of many topics in decision theory, and of the theory of Markov systems. We examine in this unit, very briefly, the first of these. In the unit on Markov Sequences, we provide an introduction to the third.

The concept

The definition of conditional independence of events is based on a product rule which may be expressed in terms of conditional expectation, given an event. The pair $\{A, B\}$ is conditionally independent, given C, iff

$$E[I_A I_B|C] = P(AB|C) = P(A|C)P(B|C) = E[I_A|C]E[I_B|C]$$

If we let $A = X^{-1}(M)$ and $B = Y^{-1}(N)$, then $I_A = I_M(X)$ and $I_B = I_N(Y)$. It would be reasonable to consider the pair $\{X, Y\}$ conditionally independent, given event C, iff the product rule

$$E[I_M(X)I_N(Y)|C] = E[I_M(X)|C]E[I_N(Y)|C]$$

holds for all reasonable M and N (technically, all Borel M and N). This suggests a possible extension to conditional expectation, given a random vector. We examine the following concept.

Definition

The pair $\{X, Y\}$ is *conditionally independent, givenZ*, designated $\{X, Y\}$ ci $|Z$, iff

$$E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z]E[I_N(Y)|Z] \text{ for all Borel } M.\ N$$

Remark. Since it is not necessary that X, Y, or Z be real valued, we understand that the sets M and N are on the codomains for X and Y, respectively. For example, if X is a three dimensional random vector, then M is a subset of R^3.

As in the case of other concepts, it is useful to identify some key properties, which we refer to by the numbers used in the table in Appendix G. We note two kinds of equivalences. For example, the following are equivalent.

(CI1) $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z][E[I_N(Y)|Z]$ a.s. for all Borel sets M, N

(CI5) $E[g(X, Z)h(Y, Z)|Z] = E[g(X, Z)|Z]E[h(Y, Z)|Z]$ a.s. for all Borel functions g, h

Because the indicator functions are special Borel functions, (CI1) is a special case of (CI5). To show that (CI1) implies (CI5), we need to use linearity, monotonicity, and monotone convergence in a manner similar to that used in extending properties (CE1) to (CE6) for conditional expectation. A second kind of equivalence involves various patterns. The properties (CI1), (CI2), (CI3), and (CI4) are equivalent, with (CI1) being the defining condition for $\{X, Y\}$ ci $|Z$.

(CI1) $E[I_M(X)I_N(Y)|Z] = E[I_M(X)|Z][E[I_N(Y)|Z]$ a.s. for all Borel sets M, N

(CI2) $E[I_M(X)|Z, Y] = E[I_M(X)|Z]$ a.s. for all Borel sets M

(CI3) $E[I_M(X)I_Q(Z)|Z, Y] = E[I_M(X)I_Q(Z)|Z]$ a.s. for all Borel sets M, Q

(CI4) $E[I_M(X)I_Q(Z)|Y] = E\{E[I_M(X)I_Q(Z)|Z]|Y\}$ a.s. for all Borel sets M, Q

As an example of the kinds of argument needed to verify these equivalences, we show the equivalence of (CI1) and (CI2).

- (CI1) implies (CI2). Set $e_1(Y, Z) = E[I_M(X)|Z, Y]$ and $e_2(Y, Z) = E[I_M(X)|Z]$. If we show

  $$E[I_N(Y)I_Q(Z)e_1(Y, Z) = E[I_N(Y)I_Q(Z)e_2(Y, Z)] \text{ for all Borel } N, Q$$

  then by the uniqueness property (E5b) for expectation we may assert $e_1(Y, Z) = e_2(Y, Z)$ a.s. Using the defining property (CE1) for conditional expectation, we have

  $$E\{I_N(Y)I_Q(Z)E[I_M(X)|Z, Y]\} = E[I_N(Y)I_Q(Z)I_M(X)]$$

  On the other hand, use of (CE1), (CE8), (CI1), and (CE1) yields

  $$E\{I_N(Y)I_Q(Z)E[I_M(X)|Z]\} = E\{I_Q(Z)E[I_N(Y)E[I_M(X)|Z]|Z]\}$$

$$= E\{I_Q(Z)E[I_M(X)|Z]E[I_N(Y)|Z]\} = E\{I_Q(Z0E[I_M(X)I_N(Y)|Z]\}$$
$$= E[I_N(Y)I_Q(Z0I_M(X)$$

which establishes the desired equality.

- (CI2) implies (CI1). Using (CE9), (CE8), (CI2), and (CE8), we have

$$E[I_M(X)I_N(Y)|Z] = E\{E[I_M(X)I_N(Y)|Z,Y]|Z\}$$
$$= E[I_N(Y)E[I_M(X)|Z,Y]|Z\} = E\{I_N(Y)E[I_M(X)|Z]|Z\}$$
$$= E[I_M(X)|Z]E[I_N(Y)|Z]$$

Use of property (CE8) shows that (CI2) and (CI3) are equivalent. Now just as (CI1) extends to (CI5), so also (CI3) is equivalent to

(CI6) $E[g(X,Z)|Z,Y] = E[g(X,Z)|Z]$ a.s. for all Borel functions g

Property (CI6) provides an important *interpretation* of conditional independence:

$E[g(X,Z)|Z]$ is the best mean-square estimator for $g(X,Z)$, given knowledge of Z. The conditon $\{X,Y\}$ ci $|Z$ implies that additional knowledge about Y does not modify that best estimate. This interpretation is often the most useful as a modeling assumption.

Similarly, property (CI4) is equivalent to

(CI8) $E[g(X,Z)|Y] = E\{E[g(X,Z)|Z]|Y\}$ a.s. for all Borel functions g

The additional properties in Appendix G are useful in a variety of contexts, particularly in establishing properties of Markov systems. We refer to them as needed.

The Bayesian approach to statistics

In the *classical* approach to statistics, a fundamental problem is to obtain information about the population distribution from the distribution in a simple random sample. There is an inherent difficulty with this approach. Suppose it is desired to determine the population mean μ. Now μ is an unknown quantity about which there is uncertainty. However, since it is a constant, we cannot assign a probability such as $P(a < \mu \le b)$. This has no meaning.

The *Bayesian* approach makes a fundamental change of viewpoint. Since the population mean is a quantity about which there is uncertainty, it is *modeled as a random variable* whose value is to be determined by experiment. In this view, the population distribution is conceived as randomly selected from a class of such distributions. One way of expressing this idea is to refer to a *state of nature*. The population distribution has been "selected by nature" from a class of distributions. The mean value is thus a random variable whose value is determined by this selection. To implement this point of view, we assume

> The value of the parameter (say μ in the discussion above) is a "realization" of a *parameter random variable* H. If two or more parameters are sought (say the mean and variance), they may be considered components of a parameter random vector.
> The population distribution is a *conditional distribution*, given the value of H.

The Bayesian model

If X is a random variable whose distribution is the population distribution and H is the parameter random variable, then $\{X,H\}$ have a joint distribution.

> For each u in the range of H, we have a *conditional distribution* for X, given $H = u$.
> We assume a *prior distribution* for H. This is based on previous experience.
> We have a *random sampling process*, given H: i.e., $\{X_i : 1 \le i \le n\}$ is conditionally iid, given H. Let $W = (X_1, X_2, \cdots, X_n)$ and consider the joint conditional distribution function
>
> $$F_{W|H}(t_1,t_2,\cdots,t_n|u) = P(X_1 \le t_1, X_2 \le t_2, \cdots, X_n \le t_n|H = u)$$
> $$= E[\textstyle\prod_{i=1}^n I_{(-\infty,t_i]}(X_i)|H = u]] = \prod_{i=1}^n E[I_{(-\infty,t_i]}(X_i)|H = u] = \prod_{i=1}^n F_{X|H}(t_i|u)$$
>
> If X has conditional density, given H, then a similar product rule holds.

Population proportion

We illustrate these ideas with one of the simplest, but most important, statistical problems: that of determining the proportion of a population which has a particular characteristic. Examples abound. We mention only a few to indicate the importance.

The proportion of a population of voters who plan to vote for a certain candidate.
The proportion of a given population which has a certain disease.
The fraction of items from a production line which meet specifications.
The fraction of women between the ages eighteen and fifty five who hold full time jobs.

The parameter in this case is the proportion p who meet the criterion. If sampling is at random, then the sampling process is equivalent to a sequence of Bernoulli trials. If H is the parameter random variable and S_n is the number of "successes" in a sample of size n, then the conditional distribution for S_n, given $H = u$, is binomial (n, u). To see this, consider

$$X_i = I_{E_i}, \text{ with } P(E_i|H = u) = E[X_i|H = u] = e(u) = u$$

Anaysis is carried out for each fixed u as in the ordinary Bernoulli case. If

$$S_n = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} I_{E_i}$$

We have the result

$$E[I_{\{k\}}(S_i)|H = u] = P(S_n = k|H = u) = C(n, k)u^k(1 - u)^{n-k} \quad \text{and } E[S_n|H = u] = nu$$

The objective

We seek to determine the best mean-square estimate of H, given $S_n = k$.
If $H = u$, we know $E[S_n|H] = nu$. Sampling gives $S_n = k$. We make a *Bayesian reversal* to get an exression for $E[H|S_n = k]$.
To complete the task, we must assume a prior distribution for H on the basis of prior knowledge, if any.

The Bayesian reversal

Since $\{S_n = k\}$ is an event with positive probability, we use the definition of the conditional expectation, given an event, and the law of total probability (CE1b) to obtain

$$E[H|S_n = k] = \frac{E[HI_{\{k\}}(S_n)]}{E[I_{\{k\}}(S_n)]} = \frac{E\{HE[I_{\{k\}}(S_n)|H]\}}{E\{E[I_{\{k\}}(S_n)|H]\}} = \frac{\int uE[I_{\{k\}}(S_n)|H = u]f_H(u)\,du}{\int E[I_{\{k\}}(S_n)|H = u]f_H(u)\,du}$$
$$= \frac{C(n, k)\int u^{k+1}(1 - u)^{n-k}f_H(u)\,du}{C(n, k)\int u^k(1 - u)^{n-k}f_H(u)\,du}$$

A prior distribution for H

The beta (r, s) distribution (see Appendix G), proves to be a "natural" choice for this purpose. Its range is the unit interval, and by proper choice of parameters r, s, the density function can be given a variety of forms (see Figures 16.1.1 and 16.2.2).

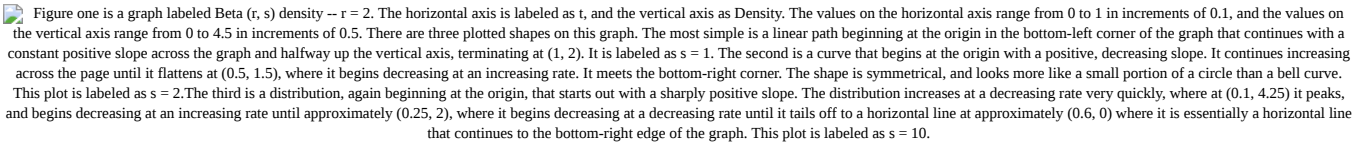
Figure one is a graph labeled Beta (r, s) density -- r = 2. The horizontal axis is labeled as t, and the vertical axis as Density. The values on the horizontal axis range from 0 to 1 in increments of 0.1, and the values on the vertical axis range from 0 to 4.5 in increments of 0.5. There are three plotted shapes on this graph. The most simple is a linear path beginning at the origin in the bottom-left corner of the graph that continues with a constant positive slope across the graph and halfway up the vertical axis, terminating at (1, 2). It is labeled as s = 1. The second is a curve that begins at the origin with a positive, decreasing slope. It continues increasing across the page until it flattens at (0.5, 1.5), where it begins decreasing at an increasing rate. It meets the bottom-right corner. The shape is symmetrical, and looks more like a small portion of a circle than a bell curve. This plot is labeled as s = 2.The third is a distribution, again beginning at the origin, that starts out with a sharply positive slope. The distribution increases at a decreasing rate very quickly, where at (0.1, 4.25) it peaks, and begins decreasing at an increasing rate until approximately (0.25, 2), where it begins decreasing at a decreasing rate until it tails off to a horizontal line at approximately (0.6, 0) where it is essentially a horizontal line that continues to the bottom-right edge of the graph. This plot is labeled as s = 10.

Figure 16.1.1. The Beta(r,s) density for $r = 2$, $s = 1, 2, 10$.

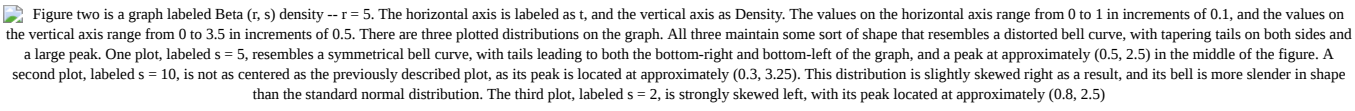
Figure two is a graph labeled Beta (r, s) density -- r = 5. The horizontal axis is labeled as t, and the vertical axis as Density. The values on the horizontal axis range from 0 to 1 in increments of 0.1, and the values on the vertical axis range from 0 to 3.5 in increments of 0.5. There are three plotted distributions on the graph. All three maintain some sort of shape that resembles a distorted bell curve, with tapering tails on both sides and a large peak. One plot, labeled s = 5, resembles a symmetrical bell curve, with tails leading to both the bottom-right and bottom-left of the graph, and a peak at approximately (0.5, 2.5) in the middle of the figure. A second plot, labeled s = 10, is not as centered as the previously described plot, as its peak is located at approximately (0.3, 3.25). This distribution is slightly skewed right as a result, and its bell is more slender in shape than the standard normal distribution. The third plot, labeled s = 2, is strongly skewed left, with its peak located at approximately (0.8, 2.5)

Figure 16.1.2. The Beta(r,s) density for $r = 5$, $s = 2, 5, 10$.

Its analysis is based on the integrals

$$\int_0^1 u^{r-1}(1 - u)^{s-1}\,du = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r + s)} \quad \text{with } \Gamma(a + 1) = a\Gamma(a)$$

For H ~ beta (r, s), the density is given by

$$f_H(t) = \frac{\Gamma(r + s)}{\Gamma(r)\Gamma(s)}t^{r-1}(1 - t)^{s-1} = A(r, s)t^{r-1}(1 - t)^{s-1} \quad 0 < t < 1$$

For $r \geq 2$, $s \geq 2$, f_H has a maximum at $(r-1)/(r+s-2)$. For r, s positive integers, f_H is a polynomial on [0, 1], so that determination of the distribution function is easy. In any case, straightforward integration, using the integral formula above, shows

$$E[H] = \frac{r}{r+s} \text{ and } \text{Var}[H] = \frac{rs}{(r+s)^2(r+s+1)}$$

If the prior distribution for H is beta (r, s), we may complete the determination of $E[H|S_n = k]$ as follows.

$$E[H|S_n = k] = \frac{A(r,s)\int_0^1 u^{k+1}(1-u)^{n-k}u^{r-1}(1-u)^{s-1}\,du}{A(r,s)\int_0^1 u^{k}(1-u)^{n-k}u^{r-1}(1-u)^{s-1}\,du} = \frac{\int_0^1 u^{k+r}(1-u)^{n+s-k-1}\,du}{\int_0^1 u^{k+r-1}(1-u)^{n+s-k-1}\,du}$$

$$= \frac{\Gamma(r+k+1)\Gamma(n+s-k)}{\Gamma(r+s+n+1)} \cdot \frac{\Gamma(r+s+n)}{\Gamma(r+k)\Gamma(n+s-k)} = \frac{k+r}{n+r+s}$$

We may adapt the analysis above to show that H is conditionally beta $(r+k, s+n-k)$, given $S_n = k$.

$$F_{H|S}(t|k) = \frac{E[I_t(H)I_{\{k\}}(S_n)}{E[I_{\{k\}}(S_n)} \text{ where } I_t(H) = I_{[0,t]}(H)$$

The analysis goes through exactly as for $E[H|S_n = k]$, except that H is replaced by $I_t(H)$. In the integral expression for the numerator, one factor u is replaced by $I_t(u)$. For $H \sim$ beta (r, s), we get

$$F_{H|S}(t|k) = \frac{\Gamma(r+s+n)}{\Gamma(r+k)\Gamma(n+s-k)} \int_0^t u^{k+r-1}(1-u)^{n+s-k-1}\,du = \int_0^t f_{H|S}(u|k)\,du$$

The integrand is the density for beta $(r+k, n+s-k)$.

Any prior information on the distribution for H can be utilized to select suitable r, s. If there is no prior information, we simply take $r = 1$, $s = 1$, which corresponds to

$H \sim$ uniform on (0, 1). The value is as likely to be in any subinterval of a given length as in any other of the same length. The information in the sample serves to modify the distribution for H, conditional upon that information.

Example 16.1.1 Population proportion with a beta prior

It is desired to estimate the portion of the student body which favors a proposed increase in the student blanket tax to fund the campus radio station. A sample of size $n = 20$ is taken. Fourteen respond in favor of the increase. Assuming prior ignorance (i.e., that H beta (1,1)), what is the conditional distribution given $s_{20} = 14$? After the first sample is taken, a second sample of size $n = 20$ is taken, with thirteen favorable responses. Analysis is made using the conditional distribution for the first sample as the prior for the second. Make a new estimate of H.

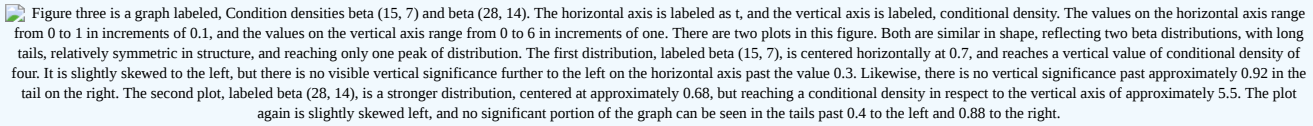

Figure 16.1.3. Conditional densities for repeated sampling, Example 16.1.1.

Solution

For the *first sample* the parameters are $r = s = 1$. According the treatment above, H is conditionally beta $(k+r, n+s-k) = (15, 7)$. The density has a maximum at $(r+k-1)/(r+k+n+s-k-2) = k/n$. The conditional expectation, however, is $(r+k)/(r+s+n) = 15/22 \approx 0.6818$.

For the *second sample*, with the conditional distribution as the new prior, we should expect more sharpening of the density about the new mean-square estimate. For the new sample, $n = 20$, $k = 13$, and the prior $H \sim$ beta (15, 7). The new conditional distribution has parameters $r^* = (28-1)/(28+14-2) = 27/40 = 0.6750$. The best estimate of H is 28/(28 + 14) = 2/3. The conditonal densities in the two cases may be plotted with MATLAB (see Figure 16.1.1).

```
t = 0:0.01:1;
plot(t,beta(15,7,t),'k-',t,beta(28,14,t),'k--')
```

As expected, the maximum for the second is somewhat larger and occurs at a slightly smaller t, reflecting the smaller k. And the density in the second case shows less spread, resulting from the fact that prior information from the first sample is incorporated into the analysis of the second sample.

The same result is obtained if the two samples are combined into one sample of size 40.

It may be well to compare the result of Bayesian analysis with that for classical statistics. Since, in the latter, case prior information is not utilized, we make the comparison with the case of no prior knowledge $(r = s = 1)$. For the classical case, the estimator for μ is the sample average; for the Bayesian case with beta prior, the estimate is the conditional expectation of H, given S_n.

$$\text{If } S_n = k\text{: Classical estimate = } k/n \text{ Bayesian estimate = } (k+1)/(n+2)$$

For large sample size n, these do not differ significantly. For small samples, the difference may be quite important. The Bayesian estimate is often referred to as the *small sample* estimate, although there is nothing in the Bayesian procedure which calls for small samples. In any event, the Bayesian estimate seems preferable for small samples, and it has the advantage that *prior information may be utilized*. The sampling procedure upgrades the prior distribution.

The essential *idea* of the Bayesian approach is the view that an unknown parameter about which there is uncertainty is modeled as the value of a random variable. The *name* Bayesian comes from the role of Bayesian reversal in the analysis.

The application of Bayesian analysis to the population proportion required Bayesian reversal in the case of discrete S_n. We consider, next, this reversal process when all random variables are absolutely continuous.

The Bayesian reversal for a joint absolutely continuous pair

In the treatment above, we utilize the fact that the conditioning random variable S_n is discrete. Suppose the pair $\{W, H\}$ is jointly absolutely continuous, and $f_{W|H}(t|u)$ and $f_H(u)$ are specified. To determine

$$E[H|W = t] = \int u f_{H|W}(u|t)\, du$$

we need $f_{H|W}(u|t)$. This requires a Bayesian reversal of the conditional densities. Now by definition

$$f_{H|W}(u|t) = \frac{f_{WH}(t,u)}{f_W(t)} \text{ and } f_{WH}(t,u) = f_{W|H}(t|u) f_H(u)$$

Since by the rule for determining the marginal density

$$f_W(t) = \int f_{WH}(t,u)\, du = \int f_{W|H}(t|u) f_H(u)\, du$$

we have

$$f_{H|W}(u|t) = \frac{f_{W|H}(t|u) f_H(u)}{\int f_{W|H}(t|u) f_H(u)\, du} \text{ and } E[H|W = t] = \frac{\int u f_{W|H}(t|u) f_H(u)\, du}{\int f_{W|H}(t|u) f_H(u)\, du}$$

Example 16.1.2 A Bayesian reversal

Suppose $H \sim$ exponential (λ) and the X_i are conditionally iid, exponential (u), given $H = u$. A sample of size n is taken. Put $W = (X_1, X_2, \cdots, X_n)$, and $t^* = t_1 + t_2 + \cdots + t_n$. Determine the best mean-square estimate of H, given $W = t$.

Solution

$$f_{X|H}(t_i|u) = ue^{-ut_i} \text{ so that } f_{W|H}(t|u) = \prod_{i=1}^n ue^{-ut_i} = u^n e^{-ut^*}$$

Hence

$$E[H|W = t] = \int u f_{H|W}(u|t)\, du = \frac{\int_0^\infty u^{n+1} e^{-ut^*} \lambda e^{-\lambda u}\, du}{\int_0^\infty u^n e^{-ut^*} \lambda e^{-\lambda u}\, du}$$

$$= \frac{\int_0^\infty u^{n+1} e^{-(\lambda + t^*)u}\, du}{\int_0^\infty u^n e^{-(\lambda + t^*)u}\, du} = \frac{(n+1)!}{(\lambda + t^*)^{n+2}} \cdot \frac{(\lambda + t^*)^{n+1}}{n!} = \frac{n+1}{(\lambda + t^*)} \quad \text{where } t^* = \sum_{i=1}^n t_i$$