STAT 830

Independence, conditional distributions and modelling

When analyzing data statisticians need to specify a statistical model for the data. That is, we regard the data as random variables and specify possible joint distributions for the data. Sometimes the modelling proceeds by modelling the joint density of the data explicitly. More commonly, however, modelling amounts to a specification in terms of marginal and conditional distributions.

We begin by describing independence. Our description is formal, mathematical and precise. It should be said however that the definitions work two ways. Often we will assume that events or random variables are independent. We will argue that such an assumption is justified by a lack of causal connection between the events – in such a case knowledge of whether or not one event happens should not affect the probability the other happens. This is more subtle than it sounds, though, as we will see when we discuss Bayesian ideas.

Definition: Events A and B are independent if

$$P(AB) = P(A)P(B)\,.$$

(Notation: we often shorten the notation for intersections by omitting the intersection sign. ThusAB is the event that both A and B happen, which is also written $A \cap B$.)

Definition: A sequence of events A_i, $i = 1, \ldots, p$ are **independent** if

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any $1 \le i_1 < \cdots < i_r \le p$.

Example: If we have $p = 3$ independent events then the following equations hold:

$$\begin{aligned} P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1A_2) &= P(A_1)P(A_2) \\ P(A_1A_3) &= P(A_1)P(A_3) \\ P(A_2A_3) &= P(A_2)P(A_3) \end{aligned}$$

All these equations are needed for independence! If you have 4 events there are 11 equations; for general p there are $2^p - p - 1$.

Example: Here is a small example to illustrate the fact that all these equations are really needed. In the example there are three events any two of which are independent but where it is not true that all three are independent. Toss a fair coin twice and define the following events.

$$\begin{aligned} A_1 &= \{\text{first toss is a Head}\} \\ A_2 &= \{\text{second toss is a Head}\} \\ A_3 &= \{\text{first toss and second toss different}\} \end{aligned}$$

Then $P(A_i) = 1/2$ for each i and for $i \neq j$

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)\,.$$

Definition: We say that two random variables X and Y are **independent** if

$$P(X \in A; Y \in B) = P(X \in A)P(Y \in B)$$

for all A and B.

Definition: We say that a set of random variables $X_1, \ldots, X_p$ are **independent** if, for any $A_1, \ldots, A_p$, we have

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod_{i=1}^{p} P(X_i \in A_i).$$

Theorem 1 *1. If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are independent then for all x, y*

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)\,.$$

2. If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are independent with joint density $f_{X,Y}(x, y)$ then X and Y have densities f_X and f_Y, and (for almost all, in the sense of Lebesgue measure) x and y we have

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)\,.$$

3. *If X and Y independent with marginal densities f_X and f_Y then (X, Y) has a joint density given by*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \,.$$

4. *If $F_{X,Y}(x, y) = F_X(x) F_Y(y)$ for* **all** *x, y then X and Y are independent.*

5. *If (X, Y) has joint density $f(x, y)$ and there exist $g(x)$ and $h(y)$ st $f(x, y) = g(x)h(y)$ for (almost)* **all** *(x, y) then X and Y are independent with densities given by*

$$f_X(x) = g(x) / \int_{-\infty}^{\infty} g(u) du$$

$$f_Y(y) = h(y) / \int_{-\infty}^{\infty} h(u) du \,.$$

6. *If the pair (X, Y) is discrete with joint probability mass function $f(x, y)$ and there exist functions $g(x)$ and $h(y)$ such that $f(x, y) = g(x)h(y)$ for* **all** *(x, y) then X and Y are independent with probability mass functions given by*

$$f_X(x) = g(x) / \sum_u g(u)$$

and

$$f_Y(y) = h(y) / \sum_u h(u) \,.$$

Proof: Some of these assertions are quite technical – primarily those involving densities. My class notes provide only the direct proofs. Here I give more detailed proofs but note that they are based on ideas which are not really part of the course most years.

1. Since X and Y are independent so are the events $X \leq x$ and $Y \leq y$; hence

$$P(X \leq x, Y \leq y) = P(X \leq x) P(Y \leq y) \,.$$

2. It is notationally simpler to suppose X and Y real valued. General dimensions are not really much harder, however. In assignment 2 I ask you to show that existence of the joint density $f_{X,Y}$ implies the

existence of marginal densities f_X and f_Y. Since X, Y have a joint density, we have, for any sets A and B

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y)dydx \\ P(X \in A)P(Y \in B) &= \int_A f_X(x)dx \int_B f_Y(y)dy \\ &= \int_A \int_B f_X(x)f_Y(y)dydx\,. \end{aligned}$$

Since $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

$$\int_A \int_B [f_{X,Y}(x,y) - f_X(x)f_Y(y)]dydx = 0\,.$$

It follows (using ideas from measure theory) that the quantity in [] is 0 for almost every pair (x, y).

3. For any A and B we have

$$\begin{aligned} &P(X \in A, Y \in B) \\ &\qquad = P(X \in A)P(Y \in B) \\ &\qquad = \int_A f_X(x)dx \int_B f_Y(y)dy \\ &\qquad = \int_A \int_B f_X(x)f_Y(y)dydx\,. \end{aligned}$$

If we **define** $g(x, y) = f_X(x)f_Y(y)$ then we have proved that for $C = A \times B$ (the Cartesian product of A and B)

$$P((X, Y) \in C) = \int_C g(x,y)dydx\,.$$

To prove that g is $f_{X,Y}$ we need only prove that this integral formula is valid for an arbitrary Borel set C, not just a rectangle $A \times B$.

This is proved via a *monotone class* argument. The collection of sets C for which identity holds has closure properties which guarantee that this collection includes the Borel sets. Here are some details.

Definition: A collection $\mathcal{M}$ of subsets of some set E is called a *monotone class* if, whenever $A_1, A_2, \ldots$ all belong to $\mathcal{M}$ and either

$$A_1 \subseteq A_2 \subseteq \cdots$$

or

$$A_1 \supseteq A_2 \supseteq \cdots$$

then, in the first case,

$$\cup_{i=1}^{\infty} A_i \in \mathcal{M}$$

and, in the second case,

$$\cap_{i=1}^{\infty} A_i \in \mathcal{M}.$$

Definition: A collection $\mathcal{F}$ of subsets of some set E is called a *field* if:

$$\begin{aligned} \emptyset &\in \mathcal{F} \\ A \in \mathcal{F} &\implies A^c \in \mathcal{F} \\ A_1, \ldots, A_p \in \mathcal{F} &\implies \cup_{i=1}^{p} A_i \in \mathcal{F}. \end{aligned}$$

This definition is simply the definition of a σ field but with the weaker requirement of closure under finite rather than countable unions.

Lemma 1 *The smallest monotone class containing a field $\mathcal{F}$ is the smallest σ-field containing $\mathcal{F}$.*

Proof: The power set of E (the collection of all subsets of E) is both a σ-field and a monotone class containing $\mathcal{F}$. By "smallest" σ-field containing $\mathcal{F}$ we mean the intersection of all σ-fields containing $\mathcal{F}$; the previous sentence says this is not an empty intersection. The meaning of "smallest" monotone class is analogous. Let $\mathcal{H}$ denote the smallest σ-field and $\mathcal{M}$ the smallest monotone class containing $\mathcal{F}$.

Any σ field containing $\mathcal{F}$ is a monotone class so the smallest monotone class containing $\mathcal{F}$ is a subset of the smallest σ-field containing $\mathcal{F}$. That is, $\mathcal{H} \supseteq \mathcal{M}$. It remains to prove the other direction. Let $\mathcal{G}$ be the collection of all sets $A \in \mathcal{M}$ such that $A^c \in \mathcal{M}$. If $A \in calF$ then

$A^c \in \mathcal{F}$ so $\mathcal{G}$ includes $\mathcal{F}$. If $A_1 \subseteq A_2 \subseteq \cdots$ are all sets in $\mathcal{G} \subseteq \mathcal{M}$ then $A \equiv \cup_n A_n \in \mathcal{M}$. On the other hand

$$A_1^c \supseteq A_2^c \supseteq \cdots$$

are all sets in $\mathcal{M}$. Since $\mathcal{M}$ is a monotone class we must have

$$\cap_n A_n^c \in \mathcal{M}$$

but $\cap_n A_n^c = A^c$ so $A^c \in \mathcal{M}$. That is, $\mathcal{G}$ is closed under monotone increasing unions (one of the two properties of a monotone class.

Similarly if

$$A_1 \supseteq A_2 \supseteq \cdots$$

are all sets in $\mathcal{G}$ then $A \equiv \cap_n A_n \in \mathcal{M}$ and

$$A_1^c \subseteq A_2^c \subseteq \cdots$$

are all sets in $\mathcal{M}$. Since $\mathcal{M}$ is a monotone class we must have

$$\cup_n A_n^c \in \mathcal{M}.$$

But $\cup_n A_n^c = A^c$ so $A^c \in \mathcal{M}$. Again we see that $\mathcal{G}$ is closed under monotone decreasing unions. Thus $\mathcal{G}$ is a monotone class containing $\mathcal{F}$. Since it was defined by taking only sets from $\mathcal{M}$ we must have $\mathcal{G} = \mathcal{M}$. That is:

$$A \in \mathcal{M} \implies A^c \in \mathcal{M}.$$

Next I am going to show that $\mathcal{M}$ is closed under countable unions, that is, if $A_1, A_2, \ldots$ are all in $\mathcal{M}$ then so is their union. (Notice that this union might not be a monotone union.) If I can establish this assertion then I will have proved that $\mathcal{M}$ is a σ-field containing $\mathcal{F}$ so $\mathcal{M} \supseteq \mathcal{H}$. This would finish the proof that $\mathcal{M} = \mathcal{H}$.

First fix a $B \in \mathcal{F}$ and let now $\mathcal{G}$ be the collection of all $A \in \mathcal{M}$ such that $A \cup B \in \mathcal{M}$. Just as in the previous part of the argument prove that this new $\mathcal{G}$ is a monotone class containing $\mathcal{F}$. This shows $\mathcal{G} = \mathcal{M}$ and that for every $A \in \mathcal{M}$ and every $B \in \mathcal{F}$ we have $A \cup B \in \mathcal{M}$. Now let $\mathcal{G}$ be the collection of all $B \in \mathcal{M}$ such that for all $A \in \mathcal{M}$ we have $A \cup B \in \mathcal{M}$. Again $\mathcal{G}$ contains $\mathcal{F}$. Check that this third $\mathcal{G}$ is a monotone class and deduce that for every $A \in \mathcal{M}$ and every $B \in \mathcal{M}$

we have $A \cup B \in \mathcal{M}$. In other words: $\mathcal{M}$ is closed under finite unions (by induction on the number of sets in the union).

We have now proved that $\mathcal{M}$ is a field and a monotone class. If $A_1, A_2, \ldots$ are all in $\mathcal{M}$ define $B_n = \cup_{i=1}^n A_i$. Then

(a) $B_1 \subseteq B_2 \subseteq \cdots$.

(b) Each $B_i \in \mathcal{M}$.

(c) $A \equiv \cup_n A_n = \cup_n B_n$

Since $\mathcal{M}$ is a monotone class this last union must be in $\mathcal{M}$. That is $\cup_n A_n \in \mathcal{M}$. This proves $\mathcal{M}$ is a σ-field. •

4. Another monotone class argument.

5.

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B g(x)h(y)dydx \\ &= \int_A g(x)dx \int_B h(y)dy\,. \end{aligned}$$

Take $B = \mathbb{R}^1$ to see that

$$P(X \in A) = c_1 \int_A g(x)dx$$

where $c_1 = \int h(y)dy$. So $c_1 g$ is the density of X. Since $\int\int f_{X,Y}(xy)dxdy = 1$ we see that $\int g(x)dx \int h(y)dy = 1$ so that $c_1 = 1/\int g(x)dx$. A similar argument works for Y.

6. The discrete case is easier.

Our next theorem asserts something students think is nearly obvious. It is proved by another monotone class argument but the proof is less important than the meaning. The idea is that if U, V, W, X, Y and Z are independent then, for instance U/V, $W + X$ and Ye^Z are independent.

Theorem 2 *If $X_1, \ldots, X_p$ are independent and $Y_i = g_i(X_i)$ then $Y_1, \ldots, Y_p$ are independent. Moreover, $(X_1, \ldots, X_q)$ and $(X_{q+1}, \ldots, X_p)$ are independent. Similarly $X_1, \ldots, X_{q_1}$, $X_{q_1+1}, \ldots, X_{q_2}$ and so on are independent (provided $q_1 < q_2 < \cdots$).*

Example: Suppose X and Y are independent standard exponential random variables. That is, X and Y have joint density

$$f_{X,Y}(x,y) = e^{-x}1(x>0)e^{-y}1y>0.$$

Let

$$U = \min\{X,Y\} \text{ and } W = \max\{X,Y\}$$

I will find the joint cdf and joint density of U and W. Begin by considering the event $\{U \le u, W \le w\}$. If $u \le 0$ or $w \le 0$ then the probability is 0 so now assume $u > 0$ and $w > 0$. We then have

$$\begin{aligned}\{U \le u, W \le w\} &= \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w\} \\ &= \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X < Y\} \\ &\quad \cup \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X > Y\} \\ &\quad \cup \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X = Y\}\end{aligned}$$

The first of these three events is

$$\{X \le u, X < Y \le X + w\}$$

while the second is

$$\{Y \le u, Y < X \le Y + w\}.$$

The third event is a subset of $\{X = Y\}$ which has probability 0. Thus

$$F_{U,W}(u,w) = P(X \le u, X < Y \le X + w) + P(Y \le u, Y < X \le Y + w).$$

Since X and Y are independent and have the same distribution the two probabilities on the right hand side are equal and we compute only the first. To do so we integrate the joint density of the random variables over the set

$$\{(x,y) : 0 < x \le u, x < y < x + w\}.$$

The second restriction makes it natural to integrate in the y direction first then in the x direction second. We get

$$P(X \le u, X < Y \le X + w) = \int_0^u \int_x^{x+w} e^{-x}e^{-y}\, dy\, dx.$$

The inside integral is just

$$e^{-x}\left(e^{-x} - e^{-(x+w)}\right) = e^{-2x}\left(1 - e^{-w}\right)$$

so

$$P(X \le u, X < Y \le X+w) = \left(1 - e^{-w}\right) \int_0^u e^{-2x}\, dx = \left(1 - e^{-w}\right)\left(1 - e^{-2u}\right)/2.$$

Assembling the results we get

$$F_{U,W}(u,w) = \begin{cases} \left(1 - e^{-w}\right)\left(1 - e^{-2u}\right) & u, w > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This function can be rewritten using indicators

$$F_{U,W}(u,w) = \left(1 - e^{-w}\right) 1(w > 0) \left(1 - e^{-2u}\right) 1(u > 0).$$

This evidently factors as the product $F_U(u)F_W(w)$ where

$$\begin{aligned} F_U(u) &= \left(1 - e^{-2u}\right) 1(u > 0) \\ F_W(w) &= \left(1 - e^{-w}\right) 1(w > 0). \end{aligned}$$

Thus we find $U \perp\!\!\!\perp W$ and that U has an exponential distribution with mean $1/2$ while W has an exponential distribution with mean 1.

Conditional probability

The interpretation of probability as long run relative frequency motivates the following definitions of conditional probability. Suppose we have an experiment in which two events A and B are defined and suppose that $P(B) > 0$. Imagine an infinite sequence of independent repetitions of the experiment. Amongst the first n repetitions there must be close to $nP(B)$ occasions where event B occurs in the sense that the ratio number of occurrences divided by n gets close to (B). That is

$$\frac{\#\ B\text{s in first } n \text{ trials}}{n} \to P(B).$$

Also

$$\frac{\#\text{ times both } A \text{ and } B \text{ occur in first } n \text{ trials}}{n} \to P(AB).$$

So if we just pick out of the first n trials those trials where B occur and then see what fraction of these *also* have A occurring we get

$$\frac{\#\text{ times both } A \text{ and } B \text{ occur in first } n \text{ trials}}{\#\ B\text{s in first } n \text{ trials}} \to \frac{P(AB}{P(B)}.$$

This leads to our basic definition.

Definition: We define the conditional probability of an event A given an event B with $P(B) > 0$ by

$$P(A|B) = P(AB)/P(B).$$

Definition: For discrete random variables X and Y the conditional probability mass function of Y given X is

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= f_{X,Y}(x, y)/f_X(x) \\ &= f_{X,Y}(x, y)/\sum_t f_{X,Y}(x, t) \end{aligned}$$

For an absolutely continuous random variable X we have $P(X = x) = 0$ for all x. So what is $P(A|X = x)$ or $f_{Y|X}(y|x)$ since we may not divide by 0? As is usual in mathematics we define the ratio 0/0 by taking a suitable limit:

$$P(A|X = x) = \lim_{\delta x \to 0} P(A|x \leq X \leq x + \delta x)$$

If, e.g., X, Y have joint density $f_{X,Y}$ then with $A = \{Y \leq y\}$ we have

$$\begin{aligned} &P(A|x \leq X \leq x + \delta x) \\ &\qquad = \frac{P(A \cap \{x \leq X \leq x + \delta x\})}{P(x \leq X \leq x + \delta x)} \\ &\qquad = \frac{\int_{-\infty}^{y} \int_{x}^{x+\delta x} f_{X,Y}(u, v)dudv}{\int_{x}^{x+\delta x} f_X(u)du} \end{aligned}$$

Divide the top and bottom by δx and let $\delta x \to 0$. The denominator converges to $f_X(x)$; the numerator converges to

$$\int_{-\infty}^{y} f_{X,Y}(x, v)dv$$

We now define the conditional cumulative distribution function of Y given $X = x$ by

$$P(Y \leq y|X = x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x, v)dv}{f_X(x)}$$

If we differentiate this formula by y we get the undergraduate definition of the conditional density of Y given $X = x$, namely,

$$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)\,;$$

in words we find "conditional = joint/marginal".

Example: The 3 cards problem revisited. This is the problem where we have 3 cards – red on both sides, green on both sides and red on one / green on the other. We draw a card and see the colour on the side which is face up. Suppose we see Red. What is the chance the side face down is Red?

Students sometimes think the answer is 1/2. They say: either I am looking at the all red card or the red/green card. These are equally likely so this conditional probability is 1/2. This is wrong – the two cards are not equally likely given that the side facing up is Red.

To see this clearly we should go back to the basics. Let A be the event that we see a red side. In terms of the elementary outcomes in the example at the start of Chapter 2 we have

$$A = \{\omega_1, \omega_2, \omega_3\}.$$

Let B be the event that the side face down is red. Then

$$B = \{\omega_1, \omega_2, \omega_4\}.$$

We then have

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{2/6}{3/6} = \frac{2}{3}.$$

It is also possible to do this more intuitively but to do so you have to be careful. You are conditioning on the event that you are looking at 1 of the 3 red sides – all equally likely. Of these three sides two have the property that the other side is red. That makes the conditional probability 2/3.

Bayes Theorem

The definition of conditional probability shows that if $P(A) > 0$ and $P(B) > 0$ then we have

$$P(AB) = P(A|B)P(B) = P(B|A)P(A).$$

The crucial point about this observation is that one formula conditions on B and the other on A. Bayes theorem just rewrites this formula to emphasize the change in order of conditioning:

Theorem 3 *If A and B are two events with $P(A) > 0$ and $P(B) > 0$ then*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

It seems to me to be useful to relate this to some reasoning ideas. If a certain statement P implies a statement Q then Q is always true whenever P is true. Of course if Q is not true then neither is P. That is, the statement "not Q" implies the statement "not P". In terms of probabilities the analogy is that if $P(B|A) = 1$ then $P(A^c|B^c) = 1$ (assuming that $P(B^c) \neq 0$). This follows from

$$\begin{aligned} P(A^c|B^c) &= \frac{P(A^cB^c)}{P(B^c)} \\ &= \frac{1 - P(A \cup B)}{P(B^c)} \\ &= \frac{1 - P(A) - P(B) + P(B|A)P(A)}{1 - P(B)} \\ &= \frac{1 - P(A) - P(B) + P(A)}{1 - P(B)} \\ &= \frac{1 - P(B)}{1 - P(B)} = 1. \end{aligned}$$

It is NOT a theorem of logic that if P implies Q then Q implies P. But there is a sense in which if P usually happens and usually when P happens so does Q then Q usually happens and when Q happens usually P does too. Let's look at the formula with statements P and Q replaced by events A and B. Imagine that P is "A happens" and Q is "B happens".

Then

$$P(B|A)P(A) = P(A|B)P(B)$$

so if both terms on the left are nearly 1 ("usually happens") then both terms on the right must be nearly 1 (because if either were small the product would be too small to equal the thing on the left which is nearly 1).

The idea underlying Bayes' Theorem can be translated into the language of conditional densities:

$$f_{X|Y} = \frac{f_{Y|X} f_X}{f_Y}$$

Nowadays Bayesians like to write

$$(x|y) = (y|x)(x)/(y)$$

with the parentheses indicating densities and the letters indicating variables. This notation uses the letter in the argument of a function to indicate which function is being discussed and is at least a bit dangerous since

$$(1|2) = (2|1)(1)/(2)$$

doesn't really tell you which variables are under discussion even though it a special case of the formula above with $x = 1$ and $y = 2$.

More general formulas arise like

$$P(ABCD) = P(A|BCD)P(B|CD)P(C|D)P(D)$$

This formula can be rewritten in many orders to get a variety of equivalent expressions which, divided by some of the terms involved give theorems like that of Bayes. Also, if $A_1, \ldots, A_k$ are *mutually exclusive and exhaustive* then

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_j)P(A_j)}$$

Bayes theorem is often written in this form. Of course the denominator is just $P(B)$. I remark that *mutually exclusive* means pairwise disjoint and *exhaustive* means

$$\cup_1^k A_i = \Omega.$$

The density formula is really analogous to this more general looking version of Bayes' theorem since integrals are limits of sums and

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_u f_{XY}(u,y)du}.$$