Conditional Distributions and Asymptotics

Devesh Raval

1 Conditional Distributions

Typically, the data we observe in the real world involves more than one variable. For this reason, it is useful to consider random vectors. A random vector is just a function from the sample space to $\mathbf{R}^d$. Each component of the random vector is itself a random variable. When $d = 2$, we call this a bivariate random vector. For ease of exposition, we will specialize to this case for the time being, but the ideas here extend naturally to the case where $d > 2$.

1.1 Joint Distributions

The *joint distribution* of a random vector is just a description of the probability with which it takes different values. We now describe two large classes of distributions for random vectors.

If each component of a random vector (X, Y) takes on finitely many (or at most countably many) different values, then (X, Y) is a discrete random vector. If we let $x_1, \ldots, x_k$ denote the possible values for X and $y_1, \ldots, y_\ell$ denote the possible values for Y, then the joint distribution of (X, Y) is completely described by the probability with which it takes each possible pair of values (x_i, y_j). The function

$$p(x, y) = P\{X = x, Y = y\}$$

is again referred to as the p.m.f. and it satisfies the following properties:

(i) $p(x, y) \geq 0$ for all (x, y);

(ii) $\sum_{i=1}^{k} \sum_{j=1}^{\ell} p(x_i, y_j) = 1$;

(iii) $P\{(X, Y) \in A\} = \sum_{i=1}^{k} \sum_{j=1}^{\ell} I\{(x_i, y_j) \in A\} p(x_i, y_j)$ for all $A \subseteq \mathbf{R}^2$.

If each component of a random vector takes on a continuum of values, then (X, Y) is a continuous random vector. Its distribution is completely described by its p.d.f. $f(x, y)$, which satisfies the following properties:

(i) $f(x, y) \geq 0$ for all (x, y);

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$;

(iii) $P\{(X,Y) \in A\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I\{(x,y) \in A\} f(x,y)dxdy$ for all $A \subseteq \mathbf{R}^2$.

There are, of course, other types of distributions for random vectors. In particular, it is possible for X to be a discrete random variable and for Y to be a continuous random variable. For example, X may be age (measured in years) and Y may be wages.

As with functions of random variables, a function of a random vector is also a random variable, that is, if (X,Y) is a random vector and $g : \mathbf{R}^2 \to \mathbf{R}$ is a function, then $g(X,Y)$ is a random variable. It is a random variable instead of a random vector because the resulting quantity is real-valued.

1.2 Marginal Distributions

The *marginal distribution* of X is just another name for the distribution of X, but when used in the context of a random vector (X,Y) it is used to emphasize the difference between the distribution of X and the joint distribution of (X,Y).

It is possible to compute the marginal distribution of X from the joint distribution of (X,Y). If (X,Y) is a continuous random vector with p.d.f $f(x,y)$, then

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy \ .$$

1.3 Conditional Distributions

The conditional distribution of Y given X is the distribution of Y when X takes on a particular value.

If (X,Y) is a discrete random vector where X takes on values $x_1, \ldots, x_k$ and Y takes on values $y_1, \ldots, y_\ell$, then for any x_i such that $P\{X = x_i\} > 0$

$$P\{Y = y_j | X = x_i\} = \frac{P\{X = x_i, Y = y_j\}}{P\{X = x_i\}} \ .$$

If (X,Y) is a continuous random vector with p.d.f $f(x,y)$, then for any x such that $f(x) > 0$

$$f(y|x) = \frac{f(x,y)}{f(x)} \ .$$

1.4 Conditional Expectation

The conditional expectation of Y given X, denoted by $E[Y|X]$, is just the expectation of the conditional distribution of Y given X. When X takes on a particular value x, it is denoted by $E[Y|X = x]$. Intuitively, one should have in mind the average value of Y over many repeated trials where $X = x$.

It is important to understand that the conditional expectation of Y given X, $E[Y|X]$, is a random variable because it is a function of X, which is a random variable. On the other hand, the conditional expectation of Y given X evaluated at a particular value x, $E[Y|X = x]$, is a constant.

If $E[Y|X]$ does not depend on X we say that Y is *mean independent* of X.

Example 1. Let (W, S) be a random vector where W is wages and S is an indicator variable for male (i.e., gender). Then, $E[W|S = 1]$ is the expected wage conditional on being male, and $E[W|S = 0]$ is the expected wage conditional on being female. They are constants, whereas $E[W|S]$, the expected wage conditional on the random variable for gender, is a function of a random variable and thus a random variable itself. If $E[W|S = 1] = E[W|S = 0]$, then $E[W|S]$ does not depend on S, so W is mean independent of S.

If (X, Y) is a continuous random vector with p.d.f $f(x, y)$, then for any x such that $f(x) > 0$

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f(y|x) dy \ .$$

Like the regular expectation, we have that "the conditional expectation of a sum is the sum of the conditional expectations." Moreover, anything that is a function of the conditioning variable essentially behaves like a constant with respect to the conditional expectation.

For any random vector (X, Y) and functions $g : \mathbf{R} \to \mathbf{R}$ and $h : \mathbf{R} \to \mathbf{R}$,

$$E[g(X) + h(X)Y|X] = g(X) + h(X)E[Y|X] \ .$$

1.5 Law of Iterated Expectations

It is perhaps not surprising that the weighted average of $E[Y|X]$ (weighted by the distribution of X) is simply $E[Y]$. The following theorem, which is known as the *Law of Iterated Expectations*, states this important relationship between the random variable $E[Y|X]$ and $E[Y]$.

Theorem 2. *For any random vector* (X, Y)*,*

$$E[Y] = E[E[Y|X]] \ .$$

Proof. First suppose (X, Y) is a discrete random vector where X takes on values $x_1, \ldots, x_k$ and Y takes on values $y_1, \ldots, y_\ell$. In this case,

$$\begin{aligned}
E[E[Y|X]] &= \sum_{i=1}^{k} E[Y|X = x_i] P\{X = x_i\} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} y_j P\{Y = y_j | X = x_i\} P\{X = x_i\} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} y_j P\{Y = y_j, X = x_i\} \\
&= \sum_{j=1}^{\ell} y_j \sum_{i=1}^{k} P\{Y = y_j, X = x_i\} \\
&= \sum_{j=1}^{\ell} y_j P\{Y = y_j\} = E[Y] \ .
\end{aligned}$$

You can modify these arguments for continuous random vectors. The result is true more generally for any random vector. □ □

The Law of Iterated Expectations is of fundamental importance and has many applications. Among others, we can use it to conclude that if $E[Y|X] = c$, then $E[Y] = E[E[Y|X]] = E[c] = c$. Hence, if $E[Y|X]$ does not depend on X, that is, $E[Y|X]$ equals a constant, then this constant must be equal to $E[Y]$.

1.6 Conditional Variance

The *variance of* Y *given* X is the variance of the conditional distribution of Y given X. It is denoted $\text{Var}[Y|X]$ and is defined to be

$$\text{Var}[Y|X] = E[(Y - E[Y|X])^2|X] \ .$$

As with conditional expectations, it is important to understand that $\text{Var}[Y|X]$ is a random variable, whereas $\text{Var}[Y|X = x]$ is a constant.

Example 3. Recall the setup of Example 1. $\text{Var}[Y|X = 1]$ is the variance of wages conditional on being male and $\text{Var}[Y|X = 0]$ is the variance of wages conditional on being female.

As with conditional expectations, anything that is a function of the conditioning variable essentially behaves like a constant with respect to the conditional variance.

For any random vector (X, Y) and functions $g : \mathbf{R} \to \mathbf{R}$ and $h : \mathbf{R} \to \mathbf{R}$,

$$\text{Var}[g(X) + h(X)Y|X] = h^2(X)\text{Var}[Y|X] \ .$$

1.7 Independence

It is sensible to ask the extent to which two random variables are related. The strongest notion of two random variables being unrelated is the notion of independence.

Two random variables X and Y are *independently distributed* or, more succinctly, *independent* if for any $A \subseteq \mathbf{R}$ and $B \subseteq \mathbf{R}$ we have that

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \ .$$

In this sense, knowing something about X reveals nothing about Y and vice versa.

Obviously, if X is a function of Y, then knowledge of X is completely informative of Y, so X is not independent of Y. On the other hand, if X is independent of Y, then any function of X is also independent of Y.

If (X, Y) is a continuous random vector with p.d.f $f(x, y)$, then independence is equivalent to saying that

$$f(x, y) = f(x)f(y)$$

for all $(x, y) \in \mathbf{R}^2$. Hence, for any y such that $f(y) > 0$,

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x)f(y)}{f(y)} = f(x) \ .$$

Since the conditional distribution of X given Y does not depend on Y when X and Y are independent, X is mean independent of Y, that is, $E[X|Y] = E[X]$. So, independence implies mean independence, but the reverse may not be true. Take the wages and sex example. Even if the average wage is the same for males and females, males could have a higher variance in wages in which case wages would not be independent of sex.

For any two random variables X and Y that are independently distributed,

$$E[XY] = E[X]E[Y] \ .$$

You should be able to prove this using the Law of Iterated Expectations.

It is important to note that independence is not a transitive property. Let (X, Y, Z) be a random vector. It is possible for X to be independent of Y, Y to be independent of Z, but X to not be independent of Z.

2 Asymptotic Theory

Asymptotic theory studies the large-sample properties of estimators- properties of the distribution of the estimator that hold approximately for large enough sample sizes, that is, for large enough values of n. In many cases, the asymptotic properties of estimators are much easier to derive than finite sample properties so we use asymptotic approximations when the sample size is large- even if in real life we only have finite samples. The sampling distribution of an estimator is difficult to compute in general as it depends on the distribution of $X_1, \ldots, X_n$.

Take X_i i.i.d (independent and identically distributed) from a population with mean μ and variance σ^2.

After building up some mathematical results on convergence, we will show that as the sample size gets large:

$$\begin{aligned} \bar{X}_n &\rightarrow \mu \\ Z = \sqrt{n}(\bar{X}_n - \mu)/\sigma &\rightarrow N(0,1) \\ \bar{X}_n &\approx N(\mu, \frac{\sigma^2}{n}) \end{aligned}$$

2.1 Convergence

What happens to random variables that depend on a sample as the sample size goes to infinity? If the random variable A_n becomes extremely close to some other random variable A that does not depend on the sample size, we say that A_n has converged to A. To make this concept more precise, we can define four modes of convergence- convergence in probability, in distribution, in mean square, and absolute surely. We will use convergence in probability and convergence in distribution in this class.

2.1.1 Convergence in probability

Recall that a sequence of real numbers $a_n, n \geq 1$ converges to another real number a if for all $\epsilon > 0$ there exists $N = N(\epsilon)$ such that for all $n > N$ we have

$$|a_n - a| < \epsilon \ .$$

But we are not dealing with a sequence of real numbers, but rather a sequence of random variables. Convergence in probability generalizes the idea of convergence of sequences of real numbers to sequences of random variables.

Let $A_n, n \geq 1$ be a sequence of random variables and let A be another random variable. We say that A_n converges in probability to A if for every $\epsilon > 0$ we have that

$$P\{|A_n - A| > \epsilon\} \to 0 \ .$$

In other words, as the sample size gets large, A_n no more than ϵ away from A with high probability. We may write A_n converges in probability to A as

$$A_n \xrightarrow{P} A \ .$$

2.1.2 Convergence in Distribution

Let $A_n, n \geq 1$ be a sequence of random variables and let A be a continuous random variable. We say that A_n *converges in distribution* to A if their c.d.f.s converge, that is, if the

$$P\{A_n \leq t\} \to P\{A \leq t\}$$

for every $t \in \mathbf{R}$. In this case, we write

$$A_n \xrightarrow{d} A \ .$$

If A has a special distribution, such as $A \sim N(0,1)$, then we may write instead

$$A_n \xrightarrow{d} N(0,1) \ .$$

If $A_n \xrightarrow{p} A$ then $A_n \xrightarrow{d} A$ but the reverse is not true.

2.1.3 CMT and Slutsky Thms:

Another useful technical tool is the continuous mapping theorem (CMT). It holds for random vectors, but we will only state a version of it for $d = 2$.

Theorem 4. *If $A_n, n \geq 1$ and $B_n, n \geq 1$ are sequence of random variables and a and b are constants that satisfy*

$$\begin{aligned} A_n &\overset{P}{\to} a \\ B_n &\overset{P}{\to} b \ , \end{aligned}$$

and $g : \mathbf{R}^2 \to \mathbf{R}$ is continuous at (a, b), then

$$g(A_n, B_n) \overset{P}{\to} g(a, b) \ .$$

The following result, known as Slutsky's Theorem, is frequently useful as well:

Theorem 5. *Let $A_n, n \geq 1$ and $B_n, n \geq 1$ be sequences of random variables, let A be another random variable, and let b be a constant. If $A_n \overset{d}{\to} A$ and $B_n \overset{P}{\to} b$, then*

$$\begin{aligned} B_n + A_n &\overset{d}{\to} b + A \\ B_n A_n &\overset{d}{\to} bA \\ A_n / B_n &\overset{d}{\to} A/b \textit{ if } b \neq 0 \ . \end{aligned}$$

These results are extremely useful as expectations do not have all of these properties. For example,

$$\begin{aligned} E(\frac{A_n}{B_n}) &\neq \frac{E(A_n)}{E(B_n)} \\ E(\log(A_n)) &\neq \log E(A_n) \end{aligned}$$

The second equation is due to Jensen's inequality:

If $g(.) : \Re \to \Re$ is convex, then for any random variable x for which $E|x| < \infty$ and $E|g(x)| < \infty$,

$$g(E(x)) \quad \leq \quad E(g(x))$$

Thus finite sample results can become much harder to show than asymptotic results.

2.2 LLN and CLT

2.2.1 LLN

The following result is known as the (weak) law of large numbers (WLLN). It formalizes the intuitive idea of the expectation of a random variable.

Theorem 6. *Let $X_1, \ldots, X_n$ be i.i.d. sample of size n from X. Suppose Var$[X] < \infty$. Then,*

$$\bar{X}_n \xrightarrow{P} E[X] \ .$$

Proof. We will need the following auxillary result, which is known as Chebychev's Inequality: For any $\epsilon > 0$ and any random variable A,

$$P\{|A| > \epsilon\} \leq \frac{E[A^2]}{\epsilon^2} \ .$$

□

To show that $\bar{X}_n \xrightarrow{P} E[X]$, we must show that for any $\epsilon > 0$ we have that

$$P\{|\bar{X}_n - E[X]| > \epsilon\} \to 0 \ .$$

To this end, apply Chebychev's inequality to $A = \bar{X}_n - E[X]$ to see that

$$P\{|\bar{X}_n - E[X]| > \epsilon\} \leq \frac{E[(\bar{X}_n - E[X])^2]}{\epsilon^2} \ .$$

But $\bar{X}_n - E[X]$ is a mean-zero random variable, so

$$E[(\bar{X}_n - E[X])^2] = \text{Var}[\bar{X}_n - E[X]] = \text{Var}[\bar{X}_n] = \frac{\text{Var}[X]}{n} \ .$$

Therefore,

$$P\{|\bar{X}_n - E[X]| > \epsilon\} \leq \frac{\text{Var}[X]}{n\epsilon^2} \to 0 \ ,$$

which completes the argument. □

Professor Heckman had asked for this proof on the graduate core exams a few years ago!

We call $E(X)$ the probability limit or plim of $\bar{X_n}$.

It is in fact possible to weaken the requirement in the theorem that Var$[X] < \infty$ to only $E[|X|] < \infty$, but it is much harder to prove this result. Another way to prove this is using convergence in mean square.

In general, we will use the LLN to show that sample averages converge to their expectations.

An estimator $\hat{\theta}_n$ of a parameter θ is said to be *consistent* if

$$\hat{\theta}_n \xrightarrow{P} \theta \ .$$

We have just shown that the sample mean is a consistent estimator of the population mean, as it converges to the population mean.

Example 7. Let $X_1, \ldots, X_n$ be i.i.d. sample of size n from X. Suppose Var$[X] < \infty$. Then, the WLLN implies that the sample mean, $\bar{X}_n$, is consistent for $E[X]$.

Let $X_1, \ldots, X_n$ be i.i.d. sample of size n from X. Suppose $E[X^4] < \infty$. Remember that this implies that $E[X^k] < \infty$ for $1 \leq k \leq 4$. Under this condition, the WLLN implies that $\hat{\sigma}_X^2$ is consistent for σ_X^2. To see this, write

$$\begin{aligned}
\hat{\sigma}_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}((X_i - \mu_X) + (\mu_X - \bar{X}_n))^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)^2 + 2(X_i - \mu_X)(\mu_X - \bar{X}_n) + (\mu_X - \bar{X}_n)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)^2 \\
&\qquad + \frac{1}{n-1}\sum_{i=1}^{n}2(X_i - \mu_X)(\mu_X - \bar{X}_n) + \frac{1}{n-1}\sum_{i=1}^{n}(\mu_X - \bar{X}_n)^2
\end{aligned}$$

By the WLLN,

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_X)^2 \xrightarrow{P} E[(X_i - \mu_X)^2] = \sigma_X^2 \ .$$

Also,

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}2(X_i - \mu_X)(\mu_X - \bar{X}_n) &= (\mu_X - \bar{X}_n)\frac{2}{n-1}\sum_{i=1}^{n}(X_i - \mu_X) \\
&= (\mu_X - \bar{X}_n)\frac{2n}{n-1}(\bar{X}_n - \mu_X) \\
&= -\frac{2n}{n-1}(\bar{X}_n - \mu_X)^2 \ ,
\end{aligned}$$

which tends to zero in probability by the CMT. Likewise,

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mu_X - \bar{X}_n)^2 = \frac{n}{n-1}(\mu_X - \bar{X}_n)^2$$

also tends to zero in probability by the CMT. Thus, by the CMT,

$$\hat{\sigma}_X^2 \xrightarrow{P} \sigma_X^2 \ .$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of size n from (X, Y). Suppose $E[X^4] < \infty$ and $E[Y^4] < \infty$. Then, the sample covariance, $\hat{\sigma}_{X,Y}$, is a consistent estimator of the covariance $\sigma_{X,Y}$. This can be proven using an argument similar to the one used to show that the sample variance is a consistent estimator of the variance. Moreover, by the CMT, we have that if $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$, then

$$\hat{\rho}_{X,Y} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y} \xrightarrow{P} \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \text{Corr}[X, Y] \ .$$

2.2.2 CLT

The following result, known as the Central Limit Theorem (CLT), gives us the result that we want.

Theorem 8. *Let $X_1, \ldots, X_n$ be i.i.d. sample of size n from X. Suppose $0 < Var[X] < \infty$. Then,*

$$\frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \xrightarrow{d} N(0,1) \ .$$

In other words, for every $t \in \mathbf{R}$,

$$P\{\frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \leq t\} \to \Phi(t) \ ,$$

where $\Phi(t)$ is the c.d.f. of the standard normal distribution.

We call $N(0,1)$ the limiting distribution of $\frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$.

Example 9. Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from X. Suppose $E[X^4] < \infty$ and $0 < \text{Var}[X] < \infty$. We know from the CLT

$$\frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \xrightarrow{d} Z \ ,$$

where $Z \sim N(0,1)$. We know from the WLLN,

$$\hat{\sigma}_X^2 \xrightarrow{P} \sigma_X^2 > 0 \ .$$

From the CMT,

$$\frac{\sigma_X}{\hat{\sigma}_X} \xrightarrow{P} 1 \ .$$

By Slutsky's Theorem,

$$\frac{\bar{X}_n - \mu_X}{\frac{\hat{\sigma}_X}{\sqrt{n}}} = \frac{\sigma_X}{\hat{\sigma}_X} \frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \xrightarrow{d} Z \ .$$

We can also derive results on functions of $\bar{X}_n$ as well through the Delta Method (or any function of a random variable with a limiting distribution):

Theorem 10. *If $\sqrt{n}(\theta_n - \theta) \to^d N(0, \Sigma)$ where $\theta$is m x 1 and Σ is m x m; and $g() : \Re^m \to \Re^k$, then:*

$$\sqrt{n}(g(\theta_n) - g(\theta_0)) \quad \to^d \quad N(0, \frac{d}{d\theta'} g(\theta_0) \Sigma \frac{d}{d\theta'} g(\theta_0)')$$

or, for $m = 1$:

$$\sqrt{n}(g(\theta_n) - g(\theta_0)) \quad \to^d \quad N(0, g'(\theta_0)^2 \sigma^2)$$

The proof of this depends on taking a Taylor expansion.