POL 571: Random Variables and Probability Distributions

Kosuke Imai
Department of Politics, Princeton University

February 22, 2006

1 Random Variables and Distribution Functions

Often, we are more interested in some consequences of experiments than experiments themselves. For example, a gambler is more interested in how much they win or lose than the games they play. Formally, a random variable is a function which maps the sample space into $\mathbf{R}$ or its subset.

Definition 1 *A random variable is a function $X : \Omega \mapsto \mathbf{R}$ satisfying $A(x) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbf{R}$. Such a function is said to be $\mathcal{F}$-measurable.*

After an experiment is done, the outcome $\omega \in \Omega$ is revealed and a random variable $X(\omega)$ takes some value in $\mathbf{R}$. For example, in a coin toss experiment, you may assign the value of 1 to a head and 0 to a tail. The reason for the technical requirement will become clear when we define the distribution function of a random variable, which describes how likely it is for X to take at least as large as a particular value.

Definition 2 *The (cumulative) distribution function of a random variable X is the function $F : \mathbf{R} \mapsto [0, 1]$ given by $F(x) = P(A(x))$ where $A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$ or equivalently $F(x) = P(X \leq x)$.*

We sometimes write $F_X(x)$ to emphasize this function is defined for the random variable X. The (cumulative) distribution function is also often called "CDF."

Definition 3 *The two random variables, X and Y, are said to be identically distributed if $P(X \in A) = P(Y \in A)$ for any $A \in \mathcal{F}$ or equivalently $F_X(x) = F_Y(x)$ for any x.*

Finally, a distribution function has the following important properties.

Theorem 1 (Distribution Function I) *Let x and y be real numbers. A distribution function $F(x)$ of a random variable X satisfies the following properties.*

1. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.
2. *F is increasing, i.e., if $x < y$, then $F(x) \leq F(y)$.*
3. *F is right-continuous, i.e., $\lim_{x \downarrow c} F(x) = F(c)$ for any $c \in \mathbf{R}$.*

Similarly, one can also prove the following additional properties.

Theorem 2 (Distribution Function II) *A distribution function $F(x)$ of a random variable X satisfies the following properties.*

1. $P(X > x) = 1 - F(x)$.
2. $P(x < X \leq y) = F(y) - F(x)$.
3. $P(X = x) = F(x) - \lim_{y \uparrow x} F(y)$.

Let's look at some examples of random variable and their distribution functions.

Example 1

1. **Bernoulli distribution.** *In a coin toss experiment, a Bernoulli random variable can be defined as $X(head) = 1$ and $X(tail) = 0$. What is the distribution function?*
2. **Geometric distribution.** *This random variable represents the number of Bernoulli trials required before the first success.*
3. **Logistic distribution.** *The distribution function of a logistic random variable is given by $F(x) = \frac{1}{1+e^{-x}}$. Confirm that this satisfies Theorem 1.*

Random variables can be classified into two classes based on their distribution functions.

Definition 4 *Let X be a random variable.*

1. *X is said to be discrete if its distribution function is a step function.*
2. *X is said to be continuous if its distribution function is a continuous function.*

From the materials we learned in POL 502, you should be able to show that the distribution function of a uniform random variable as well as that of a logistic random variable is continuous.

If a uniform random number generator is available (e.g., runif() in R), one can simulate a continuous random variable using the inverse of its distribution function. This is called the inverse CDF method where CDF stands for the cumulative distribution function.

Theorem 3 (Inverse CDF Method) *Let F be a distribution function. If U is a uniform random variable, then the distribution function of the random variable $F^{-1}(U)$ is given by F where $F^{-1}(u) = \inf\{x : F(x) \geq u\}$ with $0 \leq u \leq 1$ is called the generalized inverse of F (or the quantile function of X).*

The reason for this cumbersome definition of F^{-1} is that the distribution function is in general not one-to-one. However, for continuous distributions with a strictly increasing distribution function, F^{-1} equals the ordinary inverse function.

Example 2 *Write R functions that simulate a random variable from the following distributions via the inverse CDF method.*

1. **Bernoulli distribution.** *Its CDF is given in Example 1.*
2. **Logistic distribution.** *Its CDF is given in Example 1.*
3. **Exponential distribution.** *The probability function is given by $F(x) = 1 - e^{-\lambda x}$.*

Of course, the uniform random variable is a theoretical construct, and only a "pseudo-random" number is available to us, which is basically a deterministic sequence of values based on the *random seed* mimicking a random number. Von Neumann once said "There is no such thing as a random number – there are only methods of producing random numbers."

2 Probability Density and Mass Functions

While the distribution function defines the distribution of a random variable, we are often interested in the likelihood of a random variable taking a particular value. This is given by the probability density and mass functions for continuous and discrete random variables, respectively.

Definition 5 *Let X be a random variable and $x \in \mathbf{R}$.*

1. *If X is discrete, then it has the probability mass function $f : \mathbf{R} \mapsto [0, 1]$ defined by*
$$f(x) = P(X = x).$$
2. *If X is continuous, then it has the probability density function, $f : \mathbf{R} \mapsto [0, \infty)$, which satisfies*
$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$
where $F(x)$ is the distribution function of X.

We may write $f_X(x)$ to stress that the probability function is for the random variable X. Although the mass function corresponds to the probability, the density function does not. In particular, the latter is not bounded by 1. Fist, let's consider discrete random variables. The following theorem is a corollary of Theorems 1 and 2.

Theorem 4 (Probability Mass Function) *Let X be a discrete random variable and $\mathcal{X}$ be a set of all possible values X can take. Then, its probability mass function $f(x)$ and distribution function $F(x)$ have the following relationships.*

$$F(x) = \sum_{\{t \in \mathcal{X} : t \leq x\}} f(t), \quad f(x) = F(x) - \lim_{t \uparrow x} F(t), \quad \text{and} \quad \sum_{t \in \mathcal{X}} f(t) = 1.$$

We consider commonly used discrete random variables and their probability mass functions.

Example 3

1. **Binomial distribution.** *The sum of n identically distributed Bernoulli random variables with probability of success p is a Binomial random variable, whose probability mass function is*
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for} \quad x = 0, 1, \ldots, n.$$
2. **Bernoulli distribution.** *This is a special case of Binomial distribution with $n = 1$. The probability mass function is $f(x) = p^x(1-p)^{1-x}$.*
3. **Negative binomial distribution.** *A negative binomial random variable represents the number of failures required before the rth success occurs in Bernoulli trials. The probability mass function is given by*
$$f(x) = \binom{r+x-1}{x} p^r (1-p)^x, \quad \text{for} \quad x = 0, 1, 2 \ldots$$
4. **Geometric distribution.** *This is a special case of negative binomial distribution with $r = 1$. The probability mass function is given by $f(x) = p(1-p)^x$ for $x = 0, 1, 2, \ldots$.*

5. **Poisson distribution.** *A Poisson random variable X has the following probability mass function and the parameter λ*

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad \text{for} \quad x = 0, 1, 2, \ldots.$$

There is an interesting relationship between Poisson and Binomial distributions.

Theorem 5 (Poisson approximation to Binomial) *If n is large and p is small, Poisson probability mass function can approximate Binomial probability mass function.*

For continuous distributions, the probability density function has the following properties.

Theorem 6 (Probability Density Function) *Let X be a continuous random variable.*

1. *Its probability density function $f(x)$ has the following properties,*

$$P(X = x) = 0, \quad P(a \leq X \leq b) = \int_a^b f(x)\, dx, \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

2. *If the distribution function, $F(x)$, is differentiable at x, then $f(x) = F'(x)$.*

Now, look at some examples of continuous random variables.

Example 4

1. **Logistic distribution.** *What is the probability density function of Logistic distribution?*

2. **Gamma distribution.** *A Gamma random variable takes non-negative values and has the following density function with the parameters $\alpha > 0$ (shape parameter), $\beta > 0$ (scale parameter),*

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ is called the Gamma function.

3. **Exponential distribution.** *This is a special case of Gamma distribution with $\alpha = 1$, i.e., $f(x) = \beta e^{-\beta x}$. This distribution has the "memoryless" property.*

4. **χ^2 distribution.** *This is another special case of Gamma distribution with $\alpha = \nu/2$ and $\beta = 1/2$ where ν is called the degrees of freedom parameter.*

5. **Beta distribution.** *A Beta random variable takes values in $[0, 1]$ and has the following density function with the parameters $\alpha, \beta > 0$*

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1},$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}\, dx$ is called the Beta function.

6. **Uniform distribution.** *This is a special case of a Beta random variable with $\alpha = \beta = 1$. The density function is given by $f(x) = 1$. More generally, a uniform random variable X takes values in a closed interval, $[a, b]$, with the density function $f(x) = \frac{1}{b-a}$.*

7. **Normal (Gaussian) distribution.** *Normal distribution has two parameters, mean μ and variance σ^2,*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

If $\mu = 0$ and $\sigma^2 = 1$, then it is called the standard Normal distribution.

We now consider the "truncation" of a probability distribution where some values cannot be observed and hence are eliminated from the sample space.

Theorem 7 (Truncated Distribution) *Let X be a discrete (continuous) random variable and denote its probability function and probability mass (density) function by $F(x)$ and $f(x)$, respectively. If the distribution is truncated so that only the values in $\mathcal{X}$ are observed, then the probability mass (density) function of the truncated random variable is given by,*

$$g(x) = \frac{1\{x \in \mathcal{X}\} f(x)}{P(X \in \mathcal{X})}$$

Let's consider some examples of truncated distributions.

Example 5

1. **0-truncated Poisson distribution.** *What is the probability mass function of the 0-truncated Poisson distribution where 0 cannot be observed?*
2. **Right-truncated Normal distribution.** *What is the probability density function of the right-truncated Normal distribution where any value that is greater than t cannot be observed?*

Once we have a random sample from a distribution, we can define the "empirical counterpart" of the distribution function.

Definition 6 *Let X be a random variable with the distribution function $F(x)$, and $x_1, x_2, \ldots, x_n$ be a random sample from the distribution. The empirical (cumulative) distribution function is defined by*

$$\widetilde{F}(x) = \frac{\sum_{i=1}^{n} 1\{x_i \leq x\}}{n}.$$

Example 6 *Plot the empirical distribution function using a random sample you generated for the logistic distribution. Compare it with the true distribution function.*

Given a random sample from a probability distribution, we can also come up with a reasonable guess (i.e., estimation, the term used in statistics about which we will be learning soon) of the underlying probability density function. The oldest and most widely used method is the histogram.

Definition 7 *Let $x_1, x_2, \ldots, x_n$ a random sample from a distribution whose probability density function is $f(x)$. Given an origin x_0 and a bin width h, we define the bins of the histogram to be the intervals $[x_0 + mh, x_0 + (m+1)h)$ for positive and negative integers m. Then, the histogram is defined by,*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} 1\{x_i \in \text{the same bin as x}\}.$$

Example 7 *In R, obtain a random sample from a distribution of your choice (use a continuous distribution). Compare the histogram with the true probability density function. Let the bin width of the histogram vary to see how sensitive the graph is to this variable.*

3 Random Vector and Joint Distributions

So far, we have considered a single random variable. However, the results can be readily extended to a random vector, a vector of multiple random variables. For example, we can think about an experiment where we throw two dice instead of one at each time. Then, we need to define the *joint* probability (mass or density) function for a random vector. For simplicity, we only consider bivariate distributions, but the same principle applies to multivariate distributions in general.

Definition 8 *Let X and Y be random variables. The joint distribution function of X and Y is $F : \mathbf{R}^2 \mapsto [0, 1]$ defined by*

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

1. *If (X, Y) is a discrete random vector, the joint probability mass function $f : \mathbf{R}^2 \mapsto \mathbf{R}$ is defined by*

$$f(x, y) = P(X = x, Y = y).$$

2. *If (X, Y) is a continuous random vector, the joint probability density function $f : \mathbf{R}^2 \mapsto \mathbf{R}$ is defined by*

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)\, ds\, dt.$$

If $F(x, y)$ is differentiable at (x, y), then we have $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$.

Note that for a continuous distribution, we have

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x, y)\, dx\, dy.$$

Let's look at a couple of examples.

Example 8

1. **Multinomial distribution.** *An k-dimensional multinomial random vector $X = (X_1, \ldots, X_k)$ has the following probability mass function.*

$$f(x_1, x_2, \ldots, x_k) = \binom{n}{x_1\ x_2\ \ldots\ x_k} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_k}$$

where $p = (p_1, p_2, \ldots, p_n)$ is an k-dimensional vector of probabilities with $\sum_{i=1}^k p_i = 1$ and $n = \sum_{i=1}^k x_i$. A special case of this distribution is Binomial distribution.

2. **Multivariate Normal distribution.** *An k-dimensional multivariate normal random vector $X = (X_1, \ldots, X_k)$ with the following density function*

$$f(\mathbf{x}) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

where μ is an k-dimensional vector of mean and Σ is an $k \times k$ positive definite covariance matrix.

One can easily obtain the *marginal* distribution function from the joint distribution function.

Theorem 8 (Marginal and Joint Distribution Functions) *Let X and Y be random variables and $F(x, y)$ be their joint distribution function. If $F_X(x)$ and $F_Y(y)$ are the marginal distribution functions of X and Y, respectively, then*

$$F_X(x) = \lim_{y \to \infty} F(x, y), \quad \text{and} \quad F_Y(y) = \lim_{x \to \infty} F(x, y).$$

We now define the Independence of random variables.

Definition 9 *Two random variables, X and Y, are said to be independent if*

$$F(x, y) = F_X(x) F_Y(y),$$

where F is the joint distribution function, F_X and F_Y are the marginal distribution functions for X and Y, respectively.

It is also possible to obtain the *marginal* probability mass (density) function from the joint probability mass (density) function. If we have the joint density function of the two random variables X and Y, we can obtain the marginal density function of X by "integrating out" Y.

Theorem 9 (Marginal and Joint Density (Mass) Functions) *Let X and Y be random variables where $f_X(x)$ and $f_Y(y)$ are the marginal probability mass (density) functions. Let $f(x, y)$ be their joint probability mass (density) function. $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : f_Y(y) > 0\}$ are called the support of the marginal distributions of X and Y, respectively.*

1. *If both X and Y are discrete, then*

$$f_X(x) = \sum_{y \in \mathcal{Y}} f(x, y), \quad \text{and} \quad f_Y(y) = \sum_{x \in \mathcal{X}} f(x, y).$$

2. *If both X and Y are continuous, then*

$$f_X(x) = \int_{y \in \mathcal{Y}} f(x, y)\, dy, \quad \text{and} \quad f_Y(y) = \int_{x \in \mathcal{X}} f(x, y)\, dx.$$

In addition to joint and marginal distributions, *conditional* distributions are often of interest.

Definition 10 *Let X and Y be random variables with the marginal probability mass (density) function, $f_X(x)$ and $f_Y(y)$, and the joint probability mass (density) function, $f(x, y)$. The conditional mass (density) functions of X given Y and of Y given X is defined by*

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \quad \text{and} \quad f(y \mid x) = \frac{f(x, y)}{f_X(x)},$$

respectively.

Now, we show that the independence can be checked by decomposing the joint density (mass) function.

Theorem 10 (Independence) *Let X and Y be random variables with the joint probability mass (density) function, $f(x, y)$. X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that*

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

for every $x \in \mathbf{R}$ and $y \in \mathbf{R}$.

From this theorem, it is immediate that if random variables X and Y are independent, then $f(x \mid y) = f_X(x)$ and $f(y \mid x) = f_Y(y)$. Let's consider a couple of examples.

Example 9

1. **Multinomial and Binomial distributions.** *Let $(X_1, X_2, \ldots, X_k)$ be a multinomial random vector. Show that the marginal distribution of x_i for any $i \in \{1, \ldots, k\}$ is a Binomial distribution. Also, show that $(X_1, X_2, \ldots, X_{k-1})$ conditional on $X_k = x_k$ follows a multinomial distribution.*

2. **Bivariate Normal distribution.** *Rewrite the bivariate Normal density function using means μ_1, μ_2, variances σ_1, σ_2 and the correlation ρ. Write the joint density function as the product of the marginal and conditional density functions. Confirm that the correlation is zero if and only if the two random variables are independent.*

Many distributions can be derived hierarchically by combining conditional and marginal distributions. Here is one important example.

Example 10 Student t distribution. *A Student t random variable, X, with ν degrees of freedom, mean μ, and variance σ^2 can be simulated in the following manner,*

$$\begin{aligned} X \mid \nu, \mu, \sigma^2, Z &\sim \text{Normal}\left(\mu, \frac{\sigma^2 \nu}{Z}\right), \\ Z \mid \nu &\sim \chi^2_\nu. \end{aligned}$$

Use this fact to show that the density function of Student t distribution is equal to,

$$f(x \mid \mu, \sigma^2) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma^2}\left[1 + \frac{(X-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2}.$$