Probability:
Basic Notions

Vincenzo Bonifaci

April 10, 2013

1 Probability Spaces and Random Variables

Definition 1.1. A *probability space* has three components:

- a *sample space* Ω, the set of "outcomes";
- a family $\mathcal{F}$ of subsets of Ω, the *events*;
- a *probability function* $\Pr : \mathcal{F} \to \mathbb{R}$.

The *probability function* is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies:

- $0 \leq \Pr(E) \leq 1$ for any event E;
- $\Pr(\Omega) = 1$;
- for any finite or countable sequence of mutually disjoint events $E_1, E_2, \ldots$,

$$\Pr\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} \Pr(E_i).$$

When Ω has finite or countable cardinality we say that the probability space is *discrete*. In that case, $\mathcal{F}$ can be taken to be the set of all subsets of Ω.

Proposition 1.1. *For any two events E_1, E_2,*

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

Lemma 1.2 (Union Bound)**.** *For any finite or countable sequence of events $E_1, E_2, \ldots$,*

$$\Pr\left(\bigcup_{i \geq 1} E_i\right) \leq \sum_{i \geq 1} \Pr(E_i).$$

Definition 1.2. Two events E and F are *independent* when $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$. In general, events $E_1, E_2, \ldots, E_k$ are *mutually independent* if for any subset $I \subseteq \{1, \ldots, k\}$,

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$

Definition 1.3. The *conditional probability* that event E occurs given that event F occurs is

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

Theorem 1.3 (Law of Total Probability). *If $E_1, E_2, \ldots, E_n$ are mutually disjoint events in Ω such that $\cup_{i=1}^n E_i = \Omega$, then for any event B,*

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B|E_i) \cdot \Pr(E_i).$$

Theorem 1.4 (Bayes' Law). *If $E_1, E_2, \ldots, E_n$ are mutually disjoint events in Ω such that $\cup_{i=1}^n E_i = \Omega$, then for any event B,*

$$\Pr(E_j|B) = \frac{\Pr(B|E_j) \cdot \Pr(E_j)}{\sum_{i=1}^n \Pr(B|E_i) \cdot \Pr(E_i)}.$$

Definition 1.4. A *random variable* X on sample space Ω is a function $X : \Omega \to \mathbb{R}$. When the codomain of X is finite or countable, then X is a *discrete* random variable.

For a discrete random variable, we use the notation

$$\Pr(X = a) = \sum_{s \in \Omega : X(s) = a} \Pr(s).$$

Clearly, $\sum_a \Pr(X = a) = 1$ since X must take some value.

Definition 1.5. Two discrete random variables X, Y are *independent* if for all values x, y,

$$\Pr((X = x) \cap (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y).$$

Definition 1.6. The *expectation* of a discrete random variable X, denoted by $\mathbb{E}[X]$, is defined by $\mathbb{E}[X] = \sum_i i \cdot \Pr(X = i)$.

For a *continuous* random variable X there must exist a function $p(x)$ (*probability density function* or *pdf*) such that for any set $B \subseteq \mathbb{R}$,

$$\Pr(X \in B) = \int_B p(x)dx.$$

Similarly to the discrete case, the probability density function must satisfy $\int_{-\infty}^{\infty} p(x)dx = 1$.

The probability density function can be used to find out the probability of certain events. For example,

$$\Pr(a \leq X \leq b) = \int_a^b p(x)dx.$$

Definition 1.7. The *expectation* of a continuous random variable X with pdf f is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx.$$

Theorem 1.5 (Linearity of Expectations)**.** *For a finite collection of random variables $X_1, X_2, \ldots, X_n$ (with finite expectations),*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Moreover, for any constant c, $\mathbb{E}[cX] = c\,\mathbb{E}[X]$.

Theorem 1.6 (Jensen's Inequality)**.** *If f is a convex function, then $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)])$.*

Proof. We prove the theorem assuming that f has a Taylor expansion. Let $\mu = \mathbb{E}[X]$. By Taylor's theorem there is a $c \in \mathbb{R}$ such that

$$\begin{aligned} f(x) &= f(\mu) + f'(\mu)(x-\mu) + \frac{f''(c)(x-\mu)}{2} \\ &\ge f(\mu) + f'(\mu)(x-\mu), \end{aligned}$$

since $f''(c) \ge 0$. Taking expectations,

$$\begin{aligned} \mathbb{E}[f(x)] &\ge \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[x] - \mu) \\ &= \mathbb{E}[f(\mu)] = f(\mu) = f(\mathbb{E}[X]). \end{aligned}$$ □

Proposition 1.7. *If X, Y are independent random variables, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.*

Definition 1.8. The *variance* of a random variable X is $\mathbf{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

2 Fundamental Probability Distributions

Recall that $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Definition 2.1. A *Bernoulli* random variable Y with parameter p is defined by $\Pr(Y = 1) = p$, $\Pr(Y = 0) = 1 - p$.

Proposition 2.1. *The expectation of a Bernoulli random variable with parameter p is p.*

Proof. $\mathbb{E}[Y] = 1 \cdot p + 0 \cdot (1-p) = p$. □

Definition 2.2. A *binomial* random variable X with parameters n and p is defined by

$$\Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$

Proposition 2.2. *The expectation of a binomial random variable with parameters n and p is np.*

Proof. X can be written as $\sum_{i=1}^{n} X_i$ where each X_i is a Bernoulli random variable with parameter p. So

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np.$$

□

The *Euler beta function* is related to the binomial distribution. It is defined for $x > 0$, $y > 0$ by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt.$$

Proposition 2.3. *When x and y are positive integers,*

$$B(x, y) = \frac{(x-1)!(y-1)!}{(x+y-1)!}.$$

Definition 2.3. A *geometric* random variable X with parameter p is defined by

$$\Pr(X = n) = (1-p)^{n-1}p.$$

Proposition 2.4. *The expectation of a geometric random variable with parameter p is $1/p$.*

Proof. Define $q = 1 - p$. From the definition of expectation we get

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{\infty} kpq^{k-1} = \frac{p}{q}\sum_{k=1}^{\infty} kq^k \\
&= \frac{p}{q}\sum_{k=1}^{\infty}\sum_{j=1}^{k} q^k = \frac{p}{q}\sum_{j=1}^{\infty}\sum_{k=j}^{\infty} q^k \\
&= \frac{p}{q}\sum_{j=1}^{\infty} q^j \frac{1}{1-q} = \frac{1}{q}\sum_{j=1}^{\infty} q^j \\
&= \frac{1}{q} \cdot \frac{q}{1-q} = \frac{1}{p}.
\end{aligned}$$

□

Definition 2.4. A discrete *power-law* random variable X with parameter $\alpha > 1$ is defined by

$$\Pr(X = x) = C \cdot x^{-\alpha},$$

for all $x \geq 1$, where C is a normalization constant. Similarly, a *continuous power-law* random variable is defined by a pdf

$$p(x) = C \cdot x^{-\alpha}$$

for all $x \geq 1$ and an appropriate constant C.

Notice that $C = (\sum_{x\geq 1}^{\infty} x^{-\alpha})^{-1}$ in the discrete case. In the continuous case, it can be shown that $C = \alpha - 1$.

Note: The power-law distribution could be defined also when $\alpha \leq 1$, but in that case it is necessary to truncate the distribution at some large value M (so all values of X would be in $[1, M]$).

Proposition 2.5. *The expectation of a discrete power-law random variable with parameter* α *is* $C \sum_{x=1}^{\infty} \frac{1}{x^{\alpha-1}}$.

Proof. The expectation is

$$\sum_{x=1}^{\infty} xCx^{-\alpha} = C \sum_{x=1}^{\infty} \frac{1}{x^{\alpha-1}}.$$

Notice that this expectation is finite only when $\alpha > 2$, since the series $\sum_{x\geq 1} \frac{1}{x^{\gamma}}$ has finite value only for $\gamma > 1$. $\square$

Proposition 2.6. *The expectation of a continuous power-law random variable with parameter* α *is* $(\alpha - 1)/(\alpha - 2)$.

Proof. The expectation is

$$\int_1^{\infty} xCx^{-\alpha} dx = C \int_1^{\infty} \frac{1}{x^{\alpha-1}} dx = \frac{C}{\alpha - 2} = \frac{\alpha - 1}{\alpha - 2}. \quad \square$$

Power-law distributions are also called *scale-free* because their density function is scale-free, that is, for every constant a there is a constant b such that $p(ax) = bp(x)$ for all $x \geq 1$. Indeed,

$$p(ax) = \frac{C}{(ax)^{\alpha}} = \frac{1}{a^{\alpha}} \frac{C}{x^{\alpha}} = \frac{1}{a^{\alpha}} p(x).$$

The power-law distribution is the only scale-free distribution.

Notice that $\ln p(x) = \ln C - \alpha \ln x$, therefore if we draw the density function of a power-law distribution on a log-log scale plot, its graph is a straight line, with slope $-\alpha$.

A famous example of power-law distribution is the distribution of occurrences of words in texts.