A Quick Review of Basic Probability and Statistics

Peter Glynn (glynn@stanford.edu)
LaTeX set by Nick West
Modified for ME470 by Paul Constantine

April 7, 2011

Most of this review is from the CME308 course notes taught by Peter Glynn and scribed by Nick West. This course presumes knowledge of Chapters 1 to 3 of "Introduction to Probability Models" by Sheldon M. Ross. This material is also largely covered in the course text by P. Bremaud.

1.1 Probability: The Basics

$$\begin{aligned}
\Omega &: \text{sample space} \\
\omega \in \Omega &: \text{sample outcome} \\
A \subseteq \Omega &: \text{event} \\
X : \Omega \to S &: \text{“S-valued random variable”} \\
P &: \text{a probability (distribution / measure) on } \Omega
\end{aligned}$$

A probability has the following properties:

1. $0 \leq P\{A\} \leq 1$ for each event A.
2. $P\{\Omega\} = 1$
3. for each sequence $A_1, A_2, \ldots$ of mutually disjoint events

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\left\{A_i\right\}$$

1.2 Conditional Probability

The conditional probability of A, given B, written as $P\{A|B\}$, is defined to be

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

It is a probability on the new sample space $\Omega_B \subset \Omega$; $P\{A|B\}$ is interpreted as the likelihood / probability that A occurs given knowledge that B has occurred.

Conditional probability is fundamental to stochastic modeling. In particular in modeling "causality" in a stochastic setting, a causal connection between B and A means:

$$P\{A|B\} \geq P\{A\}.$$

1.3 Independence

Two events A and B are independent of one another if

$$P\{A|B\} = P\{A\}$$

i.e. $P\{A \cap B\} = P\{A\}P\{B\}$. Knowledge of B's occurrence has no effect on the likelihood that A will occur.

1.4 Continuous Random Variables

Given a continuous rv X taking values in $\mathbb{R}$, its probability density function $f_X(\cdot)$ is the function satisfying:

$$P\{X \leq x\} = \int_{-\infty}^{x} f_X(t)dt.$$

We interpret $f_X(x)$ as the "likelihood" that X takes on a value x. However, we need to exercise care in that interpretation. Note that

$$P\{X = x\} = \int_{x}^{x} f_X(t)dt = 0,$$

so the probability that X takes on precisely the value x (to infinite precision) is zero. The "likelihood interpretation" comes from the fact that

$$\frac{P\{X \in [a-\epsilon, a+\epsilon]\}}{P\{X \in [b-\epsilon, b+\epsilon]\}} = \frac{\int_{a-\epsilon}^{a+\epsilon} f_X(t)dt}{\int_{b-\epsilon}^{b+\epsilon} f_X(t)dt} \xrightarrow{\epsilon \to 0} \frac{f_X(a)}{f_X(b)}$$

so that $f_X(a)$ does indeed measure the relative likelihood that X takes on a value a (as opposed, say, to b).

Given a collection $X_1, X_2, \ldots, X_n$ of real-valued continuous rvs its joint probability density function (pdf) is defined as the function $f_{(X_1,X_2,\ldots,X_n)}(\cdot)$ satisfying

$$P\{X_1 \leq x_1, \ldots, X_n \leq x_n\} = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{(X_1,X_2,\ldots,X_n)}(t_1, t_2, \ldots, t_n)dt_1 \cdots dt_n.$$

Again, $f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)$ can be given a likelihood interpretation. The collection $X_1, X_2, \ldots$ is independent if

$$f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

Finally, the conditional pdf of X given $Y = y$ is given by

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

1.5 Expectations

If X is a continuous rv, its expectation is just

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

(assuming the integral exists).

Suppose that we wish to compute the expectation of $Y = g(X_1, \ldots, X_n)$, where $(X_1, \ldots, X_n)$ is a jointly distributed collection of continuous rvs. The above definition requires that we first compute the pdf of Y and then calculate $E[Y]$ via the integral

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy.$$

Fortunately, there is an alternative approach to computing $E[Y]$ that is often easier to implement.

Result 1.1: In the above setting, $E[Y]$ can be compute as:

$$E[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Remark 1.1: In older editions of his book, Sheldon Ross referred Result 1.1 as the "Law of the Unconscious Statistician"!.

Example 1.1: Suppose X is a uniformly distributed rv on $[0, 1]$, so that

$$f_X(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{o.w.} \end{cases}$$

Let $Y = X^2$.
Approach 1 to computing $E[Y]$: Note that $P\{Y \le y\} = P\{X^2 \le y\} = P\{X \le \sqrt{y}\} = \sqrt{y}$. So,

$$f_Y(y) = \frac{d}{dy} y^{\frac{1}{2}} = \frac{1}{2} y^{-\frac{1}{2}}.$$

Hence,

$$E[Y] = \int_0^1 y f_Y(y) dy = \frac{1}{2} \int_0^1 y^{\frac{1}{2}} dy = \frac{1}{2} \left[\frac{2}{3} y^{\frac{3}{2}} \right]_0^1 = \frac{1}{3}$$

Approach 2 to computing $E[Y]$:

$$E[Y] = \int_0^1 g(x) f_X(x) dy = \int_0^1 x^2 dx = \frac{1}{3}.$$

The expectation of a random variable is interpreted as a measure of a rv's "central tendency ." It is one of several summary statistics that are widely used in communicating the essential features of a probability distribution.
Finally, the expectation operator is a linear functional. Let $Y = \sum_i a_i X_i$. Then

$$E[Y] = \sum_i a_i E[X_i].$$

1.6 Commonly Used Summary Statistics

Given a rv X, the following are the most commonly used "summary statistics."

1. *Mean of X:* The mean of X is just its expectation $E[X]$. We will see later, in our discussion of the law of large numbers, why $E[X]$ is a key characteristic of X's distribution.

2. *Variance of X:*

$$\text{var}(X) = E\left[(X - E[X])^2\right]$$

   This is a measure of X's variability.

3. *Standard Deviation of X:*

$$\sigma(X) = \sqrt{\text{var}(X)}$$

This is a measure of variability that scales appropriately under a change in the units used to measure X (e.g. if X is a length, changing units from feet to inches multiplies the variance by 144, but the standard deviation by 12).

4. *Squared Coefficient of Variation*:

$$c^2(X) = \frac{\text{var}(X)}{\mathrm{E}[X]^2}$$

This is a dimensionless measure of variability that is widely used when characterizing the variation that is present in a non-negative rv X (e.g. task durations, component lifetimes, etc).

5. *kth Moment of X:* The kth moment of a random variable X is $E[X^k]$.

6. The probability that a random variable exceeds a given value x, $P(X \leq x)$ can also be written as an expectation of an indicator function,

$$P(X \leq x) = E[I(X \leq x)]$$

where $I = 1$ if $X \leq x$ and $I = 0$ otherwise.

1.7 Covariance and Correlation

The *covariance* of two random variables X and Y is given by

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

If $X = (X_1, \ldots, X_n)$ is a vector of random variables, then its covariance matrix is

$$C \;=\; E[(X - E[X])(X - E[X])^T] \;=\; E[XX^T] - E[X]E[X]^T.$$

A covariance matrix is always symmetric and positive semi-definite. The correlation coefficient is

$$\rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

The random variables X and Y are *uncorrelated* if $E[XY] = E[X]E[Y]$. If X is a vector of uncorrelated random variables, then C is diagonal.

1.8 Important Continuous Random Variables

1. *Uniform(a,b) rv:* $X \sim \text{Unif}\,(a, b)$, $a < b$ if

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{o.w.} \end{cases}$$

Applications: Arises in random number generation, etc.

Statistics:

$$\mathrm{E}[X] = \frac{a+b}{2} \qquad \text{var}\,(X) = \frac{(b-a)^2}{12}$$

2. *Beta(α, β) rv:* $X \sim \mathrm{Beta}(\alpha, \beta)$, $\alpha, \beta > 0$, if

$$f_X(x) = \begin{cases} \frac{x^\alpha (1-x)^\beta}{\mathrm{B}(\alpha,\beta)} & 0 \leq x \leq 1 \\ 0 & \text{o.w.} \end{cases}$$

where $\mathrm{B}(\alpha, \beta)$ is the "normalization factor" chosen to ensure that $f_X(\cdot)$ integrates to one, i.e.

$$\mathrm{B}(\alpha, \beta) = \int_0^1 y^\alpha (1-y)^\beta dy.$$

Applications: The Beta distribution is a commonly used "prior" on the Bernoulli parameter p.

Exercise 1.1: Compute the mean and variance of a $\mathrm{Beta}(\alpha, \beta)$ rv in terms of the function $\mathrm{B}(\alpha, \beta)$.

3. *Exponential(λ) rv:* $X \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Applications: Component lifetime, task duration, etc.

Statistics:

$$\mathrm{E}[X] = \frac{1}{\lambda} \qquad \mathrm{var}(X) = \frac{1}{\lambda^2}$$

4. *Gamma(λ, α) rv:* $X \sim \mathrm{Gamma}(\lambda, \alpha)$, $\lambda, \alpha > 0$, if

$$f_X(x) = \begin{cases} \frac{\lambda(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

is the "gamma function."

Applications: Component lifetime, task duration, etc.

Statistics:

$$\mathrm{E}[X] = \frac{\alpha}{\lambda} \qquad \mathrm{var}(X) = \frac{\alpha}{\lambda^2}$$

5. *Gaussian / Normal rv:* $X \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Applications: Arises all over probability and statistics (as a result of the "central limit theorem").

Statistics:

$$\mathrm{E}[X] = \mu \qquad \mathrm{var}(X) = \sigma^2$$

Note that $\mathrm{N}(\mu, \sigma^2) \stackrel{\mathcal{D}}{=} \mu + \sigma\,\mathrm{N}(0, 1)$, where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. (In other words, if one takes a $\mathrm{N}(0, 1)$ rv, scales it by σ and adds μ on to it, we end up with a $\mathrm{N}(\mu, \sigma^2)$ rv.