Lecture 8

BASIC PROBABILITY THEORY
(PRELUDE TO STATISTICAL DECISION THEORY)

Harrison H. Barrett
University of Arizona

OUTLINE

- Random events and random variables
- Probability and probability density functions
- Joint, conditional and marginal probabilities and PDFs
- Expectation values
- Moments and moment-generating functions
- Variance and covariance
- Some specific probability laws (as time permits)

References for this week:

Barrett and Myers, Foundations of Image Science

Appendix C, Probability

Chap. 8, Stochastic Properties of Objects and Images

Additional reference for next week:

Chap. 13, Statistical Decision Theory

Note: Appendix C in B&M deals only with scalar random variables; random vectors and random processes are introduced in Chap. 8 and used for statistical inference in Chap. 13. Poisson random variables, vectors and processes are in Chap. 11.

Events and random variables

An *event* is the outcome of some experiment

A *random variable* is a numerical quantity that depends on the outcome of the experiment

Random variables can be *continuous* (taking on values in some continuous range), *discrete* (taking only values from a discrete set), or both.

Examples of discrete random variables:

Outcome of a coin flip (heads $= 1$, tails $= 0$)

Outcome of a photon-counting experiment (N photons, $N = 1, ..., \infty$)

Examples of continuous random variables:

Electronic noise ($-\infty < V < \infty$)

Brightness of a light source used in a photon-counting experiment

Scalar and vector random variables

If the outcome of an experiment can be specified by a single number, that number is a *scalar* or *univariate* random variable

If it requires a set of numbers to describe the outcome, elements of that set can be regarded as components of a *random vector* or *multivariate random variable*.

Examples:

- A set of test scores
- Average and standard deviation of a set of test scores
- The set of counts from an array of photon counters
- A digital image, $\mathbf{g} = \{g_m, m = 1, ..., M\}$

Frequentist Definition of Probability

If event A occurs $m(A)$ times in M trials, relative frequency is number of occurrences of the event divided by the number of trials, $m(A)/M$. The probability of the event is defined by:

$$\Pr(A) = \lim_{M \to \infty} \frac{m(A)}{M} \,. \tag{C.6}$$

Criticisms of the frequentist view:

- Difficult to do large number of replications
- Difficult to hold experimental conditions fixed
- Fails in principle for complex events (e.g., images)

Alternatives to the frequentist view (see App. C):

- Ensemble definition (Gibbs)
- Classical definition (Laplace)
- Axiomatic definition (Kolmogorov)
- Bayesian definition (Rev. Thomas Bayes)

Probability Density Function (PDF)

Consider a continuous scalar random variable x that takes on values in $(-\infty, \infty)$. Probability of any particular value, $x = x_0$, is zero (set of measure zero), but we can define the probability that it takes on some value in a finite range, say $a < x < b$, in terms of its *probability density function* $\mathrm{pr}(x)$:

$$\Pr(a < x < b) = \int_a^b dx \; \mathrm{pr}(x) \, .$$

Consider a small range of values and a smooth PDF. If $a = x_0 - \frac{1}{2}\Delta x$ and $b = x_0 + \frac{1}{2}\Delta x$, we have

$$\Pr(x_0 - \tfrac{1}{2}\Delta x < x < x_0 + \tfrac{1}{2}\Delta x) = \int_{x_0 - \frac{1}{2}\Delta x}^{x_0 + \frac{1}{2}\Delta x} \mathrm{pr}(x) \approx \Delta x \; \mathrm{pr}(x_0)$$

Turn this argument around and *define*

$$\mathrm{pr}(x_0) \equiv \lim_{\Delta x \to 0} \frac{1}{\Delta x} \Pr(x_0 - \tfrac{1}{2}\Delta x \leq x \leq x_0 + \tfrac{1}{2}\Delta x) \, . \qquad \text{(C.21)}$$

Watch the notation: $\Pr(\,\cdot\,)$ means probability, $\mathrm{pr}(\,\cdot\,)$ means PDF.

NORMALIZATION

An event that is certain to occur is assigned probability 1. An impossible event has probability 0. Thus,

$$0 \leq \mathrm{Pr}(A) \leq 1\,.$$

PDFs, on the other hand, must be nonnegative but have no upper bound:

$$0 \leq \mathrm{pr}(x) \leq \infty\,.$$

For a discrete random variable that can take on values $x_n, n = 1, ..., N$,

$$\sum_{n=1}^{N} \mathrm{Pr}(x_n) = 1\,.$$

For a continuous scalar random variable that can take values in $(-\infty, \infty)$,

$$\int_{-\infty}^{\infty} dx\ \mathrm{pr}(x) = 1\,.$$

Joint probabilities

Often we can observe two or more events in one experiment, or equivalently two or more random variables.

Example: Flip two (distinguishable) coins simultaneously. Event A could be heads on the first coin and B could be heads on the second coin. Or, event C could be two heads, while event D could be two tails.

The *joint probability* of A and B, denoted $\Pr(A, B)$, is the relative frequency with which both A *and* B occur in the same *pair* of flips:

$$\Pr(A, B) = \lim_{M \to \infty} \frac{m(A, B)}{M} .$$

Here, the comma is to be read as "AND"; $m(A, B)$ is the number of times that both A AND B occur in one trial.

Note that $\Pr(A, B) = \Pr(B, A)$. Also, if two events cannot occur simultaneously, as in C and D, they are *mutually exclusive*, and $\Pr(C, D) = 0$.

Joint PDFs

Consider two continuous random variables X and Y. The joint probability density function for X and Y evaluated at $X = x$ AND $Y = y$ is given by

$$\mathrm{pr}_{X,Y}(x, y) \equiv$$

$$\lim_{\Delta x, \Delta y \to 0} \frac{1}{\Delta x \, \Delta y} \Pr(x - \tfrac{1}{2}\Delta x \leq X \leq x + \tfrac{1}{2}\Delta x), y - \tfrac{1}{2}\Delta y \leq Y \leq y + \tfrac{1}{2}\Delta y) \, .$$

Often we shall omit the subscripts and write $\mathrm{pr}(x, y)$ for $\mathrm{pr}_{X,Y}(x, y)$. Then, at the risk of confusion, we also drop the typographic distinction between the random variable X and a particular value x.

This concept and notation extends to random vectors with any finite number of components; thus $\mathrm{pr}(\mathrm{g})$, where g is an MD vector, is the same as the joint PDF $\mathrm{pr}(g_1, g_2, ..., g_M)$.

Normalization of joint probabilities and PDFs

If the discrete random variable x can take on values $x_n, n = 1, ..., N$ and the discrete random variable y can take on values $y_m, m = 1, ..., M$, then their joint probability must satisify:

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \mathsf{Pr}(x_n, y_m) = 1$$

For two continuous scalar random variables x and y that can take values in $(-\infty, \infty)$,

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \ \mathsf{pr}(x, y) = 1$$

For a continuous random vector g with M components, each taking values in $(-\infty, \infty)$,

$$\int_{-\infty}^{\infty} dg_1 \int_{-\infty}^{\infty} dg_2 \cdots \int_{-\infty}^{\infty} dg_M \ \mathsf{pr}(g_1, g_2, ..., g_M) = 1$$

or, in a useful shorthand,

$$\int_{\infty} d^M g \ \mathsf{pr}(\mathbf{g}) = 1$$

Conditional probabilities

Given that event B has occurred, what is the probability that event A has occurred? In other words, what is the probability of the event A *conditioned* on the occurrence of event B? We denote this probability $\Pr(A|B)$, where the vertical bar is read "given" or "conditioned on".

The conditional probability $\Pr(A|B)$ is found by dividing the joint probability $\Pr(A, B)$ of events A and B by the probability that event B has occurred:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}\,. \tag{C.7}$$

If A and B are mutually exclusive, then $\Pr(A|B) = 0$, and if B implies A then $\Pr(A|B) = 1$.

Equation (C.7) can be generalized to more than two events as follows:

$$\Pr(A, B, C) = \Pr(A)\Pr(B|A)\Pr(C|A, B) \tag{C.10}$$

Conditional PDFs

For two continuous scalar random variables x and y, the conditional PDF of x given y is defined by

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

The definition carries over to vector random variables also:

$$\mathrm{pr}(\mathbf{f}|\mathbf{g}) = \frac{\mathrm{pr}(\mathbf{f}, \mathbf{g})}{\mathrm{pr}(\mathbf{g})} \,.$$

Since $\mathrm{pr}(\mathbf{f}, \mathbf{g}) = \mathrm{pr}(\mathbf{g}, \mathbf{f})$, we immediately get *Bayes' rule*:

$$\mathrm{pr}(\mathbf{f}|\mathbf{g})\ \mathrm{pr}(\mathbf{g}) = \mathrm{pr}(\mathbf{g}|\mathbf{f})\,\mathrm{pr}(\mathbf{f}) \,.$$

From joint and conditional probabilities and PDFs to marginals

Suppose we know the joint PDF of two scalar random variables, $\mathsf{pr}(x, y)$, and we want to know the *marginal* PDF of one of them; we just have to integrate out the unwanted variable:

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

From the definition of the conditional, we can also write

$$\mathsf{pr}(x) = \int_{-\infty}^{\infty} dy \; \mathsf{pr}(x|y) \; \mathsf{pr}(y) \, .$$

Similar results hold for discrete random variables and for random vectors.

Statistical independence

Two random variables x and y are said to be *statistically independent* if

$$\mathsf{pr}(x, y) = \mathsf{pr}(x)\ \mathsf{pr}(y)\,.$$

An equivalent statement is that

$$\mathsf{pr}(x|y) = \mathsf{pr}(x)$$

Thus knowledge that y has been observed does not affect the PDF on observations of x.

Expected values

The *expected value* of a random variable x, also referred to as the *mean*, will be written as $\mathsf{E}\{x\}$, $\langle x \rangle$ or $\overline{x}$; these notations will be used interchangeably.

A discrete random variable x has an expected value defined by

$$\mathsf{E}\{x\} = \sum_i x_i \, \mathsf{Pr}(x_i) \,, \tag{C.32}$$

where the sum is over all possible values of x_i. We see that the expected value is a weighted average of the possible values of x_i, where the weights are the probabilities that the random variable takes on each of those values.

A continuous random variable x has an expected value defined by

$$\mathsf{E}\{x\} = \int_{-\infty}^{\infty} x \, \mathsf{pr}(x) \; dx \,, \tag{C.33}$$

where $\mathsf{pr}(x)$ is the PDF for x.

Moments and variance

The n^{th} moment of a scalar random variable x is defined as

$$m_n \equiv \langle x^n \rangle \,,$$

and the n^{th} central moment is

$$\mu_n \equiv \langle (x - \bar{x})^n \rangle \,.$$

If x is continuous, then

$$m_n = \langle x^n \rangle = \int_{-\infty}^{\infty} dx\, x^n \,\mathrm{pr}(x)\,, \qquad \mu_n = \int_{-\infty}^{\infty} dx\, (x - \bar{x})^n \,\mathrm{pr}(x)\,.$$

The *variance* of a scalar random variable is a measure of the spread of the variable about its mean. For a continuous RV, the variance is defined by

$$\sigma^2 = \mu_2 = \mathrm{Var}\{x\} = \mathrm{E}\{(x - \overline{x})^2\} = \int_{-\infty}^{\infty} (x - \overline{x})^2 \,\mathrm{pr}(x)\; dx\,. \qquad \text{(C.34)}$$

Note that

$$\sigma^2 = \mathrm{E}\{x^2\} - \overline{x}^2\,. \qquad \text{(C.35)}$$

The positive square root σ is referred to as the *standard deviation*.

Covariance matrix

The *covariance matrix* is a generalization of the variance to random vectors.

For an MD random vector g, the covariance matrix $\mathbf{K}$ is an $M \times M$ matrix with elements given by

$$K_{ij} = \left\langle (g_i - \overline{g}_i)(g_j - \overline{g}_j)^* \right\rangle \,, \tag{8.16}$$

where the asterisk indicates complex conjugate, allowing for the possibility that components of g might be complex. It follows from this definition that $\mathbf{K}$ is Hermitian, *i.e.*, $K_{ij} = K_{ji}^*$.

Any random variable covaries with itself. The diagonal elements of the covariance matrix are the variances of the components:

$$K_{jj} = \mathrm{Var}\{g_j\} \,. \tag{8.18}$$

Characteristic functions

For any random variable or vector, the *characteristic function* is the expectation of the appropriate Fourier kernel:

$$\psi(\xi) \equiv \mathsf{E}\left\{e^{-2\pi i \xi x}\right\} .$$

If x is scalar and real-valued, then

$$\psi(\xi) = \int_{-\infty}^{\infty} dx \; \mathsf{pr}(x)\, e^{-2\pi i \xi x} , \tag{C.53}$$

and the PDF and characteristic function form a Fourier transform pair:

$$\mathsf{pr}(x) = \int_{-\infty}^{\infty} d\xi \; \psi(\xi)\, e^{2\pi i \xi x} . \tag{C.54}$$

Caution: Do not confuse ξ with a spatial frequency.

Moments of the random variable x can be derived through differentiation of $\psi(\xi)$:

$$\mathsf{E}\left\{x^k\right\} = (-2\pi i)^{-k} \left.\frac{\partial^k}{\partial \xi^k}\psi(\xi)\right|_{\xi=0} . \tag{C.55}$$

Characteristic function of a random vector

For a real $M \times 1$ random vector g (column vector), the characteristic function is defined as

$$\psi_{\mathrm{g}}(\boldsymbol{\xi}) = \left\langle \exp(-2\pi i \boldsymbol{\xi}^t \mathrm{g}) \right\rangle \,, \tag{8.26}$$

where $\boldsymbol{\xi}^t$ is a real $1 \times M$ vector.

For the case of a continuous-valued random vector, $\psi_{\mathrm{g}}(\boldsymbol{\xi})$ can be written as

$$\psi_{\mathrm{g}}(\boldsymbol{\xi}) = \int_{\infty} d^M g \; \mathrm{pr}(\mathrm{g}) \exp(-2\pi i \boldsymbol{\xi}^t \mathrm{g}) \,. \tag{8.27}$$

This integral is the MD Fourier transform of the PDF, so

$$\mathrm{pr}(\mathrm{g}) = \int_{\infty} d^M \xi \; \psi_{\mathrm{g}}(\boldsymbol{\xi}) \exp(2\pi i \boldsymbol{\xi}^t \mathrm{g}) \,. \tag{8.28}$$

Two important probability laws

- The Poisson law
  - Applies to integer-valued discrete random variables
  - Often arises from independence assumptions
- The normal or Gaussian law
  - Applies to continuous random variables
  - Ofen arises as result of central-limit theorem

Both are crucial in imaging

Both are readily generalized to random vectors or random processes.

The univariate Poisson law

If one counts identical, indistinguishable events (e.g. photons, it is found that the probability of observing n events is given by

$$\mathsf{Pr}(n) = \frac{e^{-\lambda}\,\lambda^n}{n!}\,, \qquad n = 0, 1, 2, \ldots\,. \tag{C.165}$$

Note that this probability law is specified entirely by the single number λ, often called *the parameter* of the Poisson distribution.

The parameter λ is also the mean value of n:

$$\langle n \rangle = \sum_{n=0}^{\infty} n\,\mathsf{Pr}(n) = \lambda\,. \tag{C.167}$$

Moreover, λ is also the variance.

The multivariate Poisson law

Ref.: B&M Chap. 11

If an array of J detectors counts independent photons, the multivariate probability law of the outputs must have the form,

$$\Pr(\{g_j\}) = \prod_{j=1}^{J} \exp(-\overline{g}_j) \, \frac{(\overline{g}_j)^{g_j}}{g_j!} \, . \tag{11.40}$$

Because of this product form, counts in different elements are statistically independent and hence uncorrelated.

Also, the variance of the counts in one element is equal to its mean, so we can write the covariance matrix elements as

$$K_{jk} = \left\langle [g_j - \overline{g}_j] \, [g_k - \overline{g}_k] \right\rangle = \overline{g}_j \, \delta_{jk} \, . \tag{11.41}$$

Univariate and multivariate normal or Gaussian PDFs

The standard form of the normal PDF is

$$\text{pr}(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left[-\frac{(x-\overline{x})^2}{2\sigma^2}\right] . \qquad \text{(C.108)}$$

This is a two-parameter PDF, with $\overline{x}$ being the mean and σ^2 being the variance, as the notation implies.

The corresponding multivariate form is

$$\text{pr}(\mathbf{g}) = \left[(2\pi)^M \det(\mathbf{K})\right]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{g}-\overline{g})^t\, \mathbf{K}^{-1}(\mathbf{g}-\overline{g})\right] , \qquad \text{(8.185)}$$

where $\overline{g}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of $\mathbf{g}$.

Central limit theorem

Univariate form:

The PDF of a sum of N independent scalar random variables of (almost) arbitrary PDF tends to a univariate normal as $N \to \infty$. (In practice $N \simeq 5$ is often "good enough".)

Since the variables are independent, the mean of the sum is the sum of the means, and the variance of the sum is the sum of the variances, so the PDF on the sum is fully determined in the limit.

Multivariate form:

The PDF of a sum of N independent $M \times 1$ random vectors of (almost) arbitrary PDF tends to a multivariate normal as $N \to \infty$.