Probability Review for **fMRI** Data Analysis

1 Basics

Informally, we can think of a *random variable* as a quantity which takes on various values with probabilities specified by an associated function. This function is called a *probability density* if the set of values which the random variable can assume include an interval, and a *probability mass function* if this set of values is discrete. In both the continuous and discrete case, the term *probability distribution* is also often used. Random variables provide a mathematical model for answers to questions which are not deterministic, such as "Will it rain tomorrow?" or "What is the BOLD activity in a certain brain region under given conditions?"

A *probability* is a value between 0 and 1 (inclusive) which can be thought of either as the long-run average occurrence of some event, or else as subjective belief that an event will occur. (For some events [e.g., the event that Mitt Romney is elected Governor] the long-run frequency interpretation doesn't make sense, so one must think in terms of subjective probabilities.) The mathematics of probability theory is the same, whether one chooses to think in terms of long-run frequency or subjectivity.

There is a mathematical theory of probability, based on a handful of axioms. Many empirical characteristics of random phenomena are very commonly observed, such as the fact that a long-run average of measurements tend to "settle down" to some value, and that measurement noise often follows a "bell-shaped curve" (or *Gaussian distribution).* Beginning with the axioms of mathematical probability, one can model the behavior of random phenomena in great detail. In particular, conditions under which the two empirical properties mentioned above obtain are *theorems* in probability theory.

This is one indication that probability theory is a successful tool for explaining random phenomena in nature. Regarding the ubiquity of bell-shaped curves, which statisticians refer to as the *Central Limit Theorem*, the 19th century mathematician Poincaré once said "Everyone believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact" (quoted in Körner (1988), p. 347).

We will try to consistently write random variables using capital letters, and observed values (*realizations* of random variables) using lower-case letters. When it is desirable to emphasize the random variable(s) associated with a particular probability distribution, the random variable(s) will be indicated by subscripts.

As much as possible, we will illustrate results using the Gaussian distribution as an example. This approach has two advantages. Familiarity with Gaussian random variables will be very helpful when we discuss statistical methods for fMRI data. Since we have to introduce these properties anyway, it makes sense to have them do "double duty" as examples of more general results. Also, compendia of properties of various distributions are readily found. To systematically discuss various distributions would make these notes unacceptably long, and would also risk obscuring the basic ideas.

2 Gaussian Random Variables: Probability and Probability Density

We say that X has a Gaussian distribution if X takes on values according to the probability density

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

The *parameters* of this distribution are the *mean* μ and *standard deviation* σ.

The probability that X takes on values in any interval is obtained by integrating the probability density over that interval, for example the probability that X takes on values in $[a, b]$ is given by

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a),$$

where

$$F(x) \equiv \int_{-\infty}^{x} f_X(t)dt = \Pr(X \leq x) \tag{1}$$

is called the *cummulative distribution function* of X. Note that, where the derivative exists, $f_X(x) = F'(x)$. A cummulative distribution function exists for *any* random variable, continuous or discrete (not all random variables have probability densities). For the Gaussian, Student-t, and many other commonly used random variables, $f_X(x)$ is widely available in tables.

One small technical point needs to be emphasized here: the probability that a continuous random variable X takes on any particular value is *zero*. This is because there are an uncountable infinity of such values in any interval, however small. And the total probability that X takes on *any* value must equal one. So it is only meaningful to discuss probabilities of continuous random variables on intervals, and to obtain these probabilities by integration. Note that the probability density can exceed 1 (for example, evaluate f_X above at $x = \mu$), so one clearly cannot obtain probabilities at discrete points by evaluating this function.

Mathematicians and Statisticians almost universally call the Gaussian distribution the *normal* distribution; we will use "Gaussian" and "normal" interchangeably, with some preference for the former. A common notation for "X is a Gaussian random variable with mean μ and standard deviation σ" is $X \sim \mathrm{N}(\mu, \sigma^2)$. The notation is from Statistics, so it employs a "N".

3 Expectation, Mean and Variance

The *expectation* of a function $g(\cdot)$ of a random variable X, denoted $\mathrm{E}[g(X)]$, is the average value of this function over infinitely many realizations. To determine the expectation, weight the function of the random variable by the probability density of X and integrate.

The are two important special cases of expectation. The simplest case is $g(X) = X$; $\mathrm{E}(X)$ is called the *mean* of X and is often denoted μ. With this definition of μ, we can consider the squared deviation from the mean, $g(X) = (X - \mu)^2$; $\mathrm{E}[(X - \mu)^2] \equiv \sigma^2$ is called the *variance* of X. The square root of the variance has the units of X, and is called the *standard deviation*. It's not hard to show that, for our Gaussian example,

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} X f_X(x)dx = \mu,$$

and

$$\mathrm{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{\infty} (X-\mu)^2 f_X(x)dx = \sigma^2.$$

The variance of X is sometimes written $\mathrm{Var}\,(X)$.

The $\mathrm{E}\,(X^r)$ is called the rth *moment* of X, and $\mathrm{E}\left[(X-\mu)^r\right]$ is called the rth *central moment* of X. (Alternatively, one can refer to the moments as properties of the *distribution* of X.) If X is Gaussian, then $\mathrm{E}\,(X^r)$ is finite for all $r \geq 0$. But this is not be the case for all random variables, particularly for random variables for which the probability density fall off less rapidly with increasing $|x|$. A well-known example of a random variable for which even $E(X)$ doesn't exist is the Cauchy (or Lorentzian) distribution, which is the same as a t-distibution with 1 degree of freedom.

4 Random Vectors and the Multivariate Gaussian Distribution

In real applications, we usually have to deal with lots of random variables (e.g., a random variable for each voxel in an image). As an important example, consider the vector

$$X = \left[X_1, X_2, \ldots, X_p\right]^T,$$

where the superscript "T" denotes a transpose. Each component random variable X_i has a univariate distribution, called a *marginal* distribution. Assume that each X_i is Gaussian, i.e., that $X_i \sim \mathrm{N}\,(\mu_i, \sigma_i^2)$. The expectation of X is simply the vector of the component means μ_i

$$E(X) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

For the generalization of $\mathrm{Var}\,(X)$ to this multivariate case, however, the $p\,\sigma_i^2$ are not sufficient. In additional to these marginal variances, we need also to specify $p(p-1)/2$ *covariances*

$$\mathrm{Cov}\,(X_i, X_j) = \mathrm{Cov}\,(X_j, X_i) = E[(X_i - \mu_i)(X_j - \mu_j)] \equiv \sigma_{ij}.$$

Hence $\mathrm{Var}\,(X)$ can be written as a symmetric $p \times p$ matrix with the marginal variances on the diagonal and the covariances off the diagonal

$$\mathrm{Var}\,(X) \equiv \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}.$$

If each X_i has a Gaussian distribution, then we say that the vector X has a multivariate (here, p-variate) Gaussian distribution, with mean (vector) μ and covariance matrix Σ. The probability density of X is

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|} e^{-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)} \tag{2}$$

where $|\Sigma|$ denotes the determinant of the covariance matrix. If X is a p-variate Gaussian random vector with mean μ and covariance matrix Σ, then we sometimes write $X \sim \mathrm{N}_p\,(\mu, \Sigma)$.

5 Linear Combinations of Random Variables

Let X_i, $i = 1, 2, \ldots, p$ denote p random variables, with means μ_i and covariance matrix Σ. For arbitray constants a_i, we can form a linear combination of the X_i:

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p.$$

Expectation is a linear operator; this follows from the linearity of integration (and, in the case of discrete random variables, the linearity of summation). So the expectation of a linear combination of random variables is *always* the corresponding linear combination of the expectations:

$$E(Y) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_p E(X_p) = a_1\mu_1 + a_2\mu_2 + \cdots + a_p\mu_p \equiv \mu_Y. \tag{3}$$

The variance of Y requires a bit more work. Below we expand out the quadratic function within the expectation, and then use the linearity property that we've just introduced:

$$\mathrm{Var}\,(Y) = E[(Y - \mu_Y)^2] \;\; = \;\; E\left[\sum_{i=1}^{p}\sum_{j=1}^{p}(a_i X_i - a_i\mu_i)(a_j X_j - a_j\mu_j)\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j E\left[(X_i-\mu_i)(X_j-\mu_j)\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j \sigma_{ij} = a^T \Sigma a, \qquad (4)
\end{aligned}
$$

where $a \equiv [a_1, a_2, \ldots, a_p]^T$.

A variance must be non-negative, in particular $\mathrm{Var}\,(Y) \geq 0$. From the expression for $\mathrm{Var}\,(Y)$ above, we conclude that any covariance matrix Σ must be non-negative definite. Since Σ is both non-negative definite and symmetric, the eigenvalues of Σ are real and non-negative. One conequence of this is that the contours of equal probability of a multivariate Gaussian random variable are p-dimensional ellipsoids (this might not be immediately obvious, but it's not hard to demonstrate).

6 Independence and Correlation

Consider two random variables X and Y. If

$$E(XY) = E(X)E(Y),$$

then we say that X and Y are *uncorrelated*. Note that this implies that $\mathrm{Cov}\,(X,Y) = 0$, since, with obvious notation for $E(X)$ and $E(Y)$,

$$
\begin{aligned}
\mathrm{Cov}\,(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) \\
&= (\mu_X)(\mu_Y) - (\mu_X)(\mu_Y) - (\mu_Y)(\mu_X) + (\mu_X)(\mu_Y) = 0
\end{aligned}
$$

Denote the joint density of X and Y by $f_{X,Y}(x,y)$. The random variables X and Y are said to be *statistically independent* (or, simply *independent*) if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y). \qquad (5)$$

Random variables which are independent are uncorrelated, but the converse is not necessarily true. However, it's easy to show that uncorrelated *Gaussian* random variables are independent.

The *correlation* between X and Y is defined to be

$$\mathrm{Cor}\,(X,Y) \equiv \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}}. \qquad (6)$$

Correlation is a measure of how closely two random variables are to being linearly related. One can show (use the Cauchy-Schwartz inequality, or the definition of an inner product) that $|\mathrm{Cor}\,(X,Y)| \leq 1$, with equality holding if and only if X and Y are perfectly linearly related. Corresponding to each off-diagonal element of a covariance matrix is a correlation, so one can define a *correlation matrix* R in terms of Σ, with diagonal elements 1 and off-diagonal elements $R_{ij} \equiv \sigma_{ij}/(\sigma_i \sigma_j)$.

7 Conditional Distribution

To fix ideas, consider two random variables X and Y which have a bivariate Gaussian distribution. So $f_{X,Y}(x,y)$ is a "bell-shaped" surface in the (x,y) plane, from which one can determine the probability of X and Y taking on values in any set by integration. The volume under this surface equals the probability that X and Y jointly take on any real values, i.e: 1. Imagine that you are given the value of Y, say $Y = y$. Then you know that (X,Y) must fall on the horizontal line through $Y = y$. Hence, the probability desity of X given $Y = y$ (written $f_{X|Y}(x|y)$) is zero unless $Y = y$, and is proportional to the area under the surface for points (x,y) when $Y = y$. In order for the area above the line $Y = y$ and under the surface to equal 1, we must normalize approriately. Nowhere in the above discussion have we used properties specific to a Gaussian distribution, thus this argument motivates the following general definition of a conditional density:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{7}$$

You should satisfy yourself that this function integrates to 1 for any y. We have discussed a bivariate example above for simplicity only, the notions of conditional probability and conditional distribution generalize to multivariate situations of arbitrary dimension.

Conditional probability provides another way of describing independence. Random variables X and Y are independent if, and only if,

$$f_{X|Y}(x|y) = f_X(x) \tag{8}$$

Or, equivalently,

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

An informal way of expressing this is to say that knowing that the value of Y is some y does not change the probability distribution of X, and vice-versa.

8 Linear Transformations of Random Variables and Quadratic Forms

Given random variables $X_1, X_2, \ldots, X_p$, one can define another set of random variables $Y_1, Y_2, \ldots, Y_p$ by a linear transformation. For any j, let

$$Y_j \equiv a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p.$$

Using matrices, we write the system of equations defining the Ys in terms of the Xs as $Y = AX$, where the typical element of A is a_{ij}.

It's easy to show that

$$E(Y) = AE(X), \tag{9}$$

and

$$\text{Var}\,(Y) = A\text{Var}\,(X)A^T. \tag{10}$$

Let X denote a $p \times 1$ vector of random variables with mean μ and covariance matrix Σ, and let A be a $p \times p$ symmetric matrix, with ijth element a_{ij}. The function

$$Q = X^T AX \tag{11}$$

is an example of a *quadratic form*. It's easy to show that

$$E(Q) = \text{tr}\,(A\Sigma) + \mu^T A\mu, \tag{12}$$

where $\text{tr}\,(\cdot)$ denotes the *trace* of a matrix (the sum of the diagonal elements). (One can expand the quadratic and take expectations term-by-term, as we did above for the variance of a linear combination of random variables.) The $\text{Var}\,(Q)$ is more complicated, since it involves sums in which each term is the product of four random variables. Some basic results on $\text{Var}\,(Q)$ can be found in Seber (1977, Section 1.3). If X is multivariate Gaussian, then it can be shown that

$$\text{Var}\,(Q) = 2\text{tr}\,\left[(A\Sigma)^2\right] + 4\mu^T A\Sigma\mu. \tag{13}$$

References

1 DeGroot, M.H. and Schervish, M.J. (2002). *Probability and Statistics*, 3rd Edition, Addison-Wesley.

2 Körner, T.W. (1988). *Fourier Analysis*, Cambridge University Press.

3 Seber, G.A.F.(1977). *Linear Regression Analysis*. John Wiley and Sons, New York.