

A SYNOPSIS OF PROBABILITY, STATISTICS, AND STOCHASTIC PROCESSES

I. Probability, Random Variables, and Distributions

Defn I.1. Probability Space consists of a universal set Ω , a class of subsets, F , of Ω (containing Ω , closed under countable unions and complementation) and a countable additive function, P , on F such that $P(\Omega) = 1$, and $0 \leq P(A) \leq 1$ for all A in F . In applied work one hardly ever works directly in a probability space. However, it is useful to know a probability space is always lurking somewhere in the background - especially when we deal with stochastic processes.

Defn I.2. A Random Variable, X , is a function on a probability space to the real numbers such that the event $(\{\omega: X(\omega) \leq a\})$ is read as: "the set of ω such that $X(\omega)$ is less than or equal to a ".)

When discussing random variables we generally omit the argument ω and use $(X \leq a)$ in place of $\{\omega: X(\omega) \leq a\}$. It is conventional to use capital letters, X, Y, Z, \dots to designate random variables, lower case letters to designate numbers.

Defn I.3. $F(x) = P(X \leq x)$ is called the Distribution Function of X (Cumulative Distribution Function). (Note: $P(A)$ means the probability of A .) The distribution function contains most of the probabilistic information about X . If F is differentiable we call $f(x) = dF(x)/dx$ the density function of X and we say X is a continuous random variable. If X takes on only countable many values we call X a discrete random variable. In this case probabilities are obtained by summing $P(X = x_j) = f(x_j)$ rather than integrating. To avoid repetition, only continuous random variables will be used here - for discrete random variables, replace integrals by sums, $f(x)dx$ by $f(x_j)$ etc.

Defn I.4a. If X_1 and X_2 are two random variables we say their Joint Distribution Function is $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$. $F(x_1, x_2)$ tells us how, probabilistically, the two random variables behave together. Of course we could also have more than two random variables.

Defn I.4b. If X_1, X_2, \dots, X_n are n random variables ($F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$) is called their Joint Distribution Function. If $F(x_1, \dots, x_n)$ is differentiable in each x_i variable,

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

is called the joint density of X_1, \dots, X_n .

Given the joint density of X_1, \dots, X_n we can obtain the joint density of any subset of these by integrating over the excluded variables. Thus,

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

is the density of X_1 alone (also called the marginal density of X_1). Generally one can't get the joint densities from the marginal densities - that is only possible if the random variables are statistically independent.

Defn I.5.a. X_1 and X_2 are Statistically Independent if and only if $f(x_1, x_2) = f_1(x_1) f_2(x_2)$ for all x_1, x_2 .

If two random variables are statistically independent then knowing the value of one of them is useless for predicting the value of the other.

Defn I.5.b. X_1, X_2, \dots, X_n are Statistically Independent if and only if $f(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n)$. It is of interest to note that two random variables might be functionally dependent but statistically independent - e.g., we might have $Z = h(X_1, X_2)$ and $W = k(X_1, X_2)$ with Z and W statistically independent. (This is the case with the sample mean and sample variance when we are sampling from a normal population.)

In many cases we need to discuss distributions of functions of random variables - in principle we can do this directly. If, for example, we know the joint distribution of X_1 and X_2 and if $Z = h(X_1, X_2)$ and $W = k(X_1, X_2)$, we can calculate $P(Z \leq z, W \leq w)$ by using the joint distribution of X_1 and X_2 .

II. Expected Values

Defn II.1. If X is a continuous random variable with density $f(x)$ and if $\int_{-\infty}^{\infty} |x| f(x) dx$

is finite we define the expected value of x (denoted $E(X)$) as follows:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

$E(X)$ is also called the mean of the random variable. (Some authors use \bar{X} in place of $E(X)$, others $\langle X \rangle$. This can be confusing since \bar{X} is used to designate the sample mean (a random variable) while $E(X)$ is a pure number. $E(X)$ is also preferable because it indicates that E is an operator.)

Defn II.2. If $f(x_1, \dots, x_n)$ is the joint density of X_1, \dots, X_n and if $h(X_1, \dots, X_n)$ is a function of these variables we define:

$$E(h(X_1, \dots, X_n)) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1, \dots, dx_n$$

as the Expected Value of $h(X_1, \dots, X_n)$ provided:

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h(x_1, \dots, x_n)| f(x_1, \dots, x_n) dx_1, \dots, dx_n$$

is finite.

Note that there is a seeming inconsistency: $Y = h(X_1, \dots, X_n)$ is a random variable in its own right with a density $g(y)$, say, so we should use Defn II.1 to calculate

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

However, it is a theorem (called the theorem of the unconscious statistician by some authors) that $E(Y)$ and $E(h(X_1, \dots, X_n))$ as defined above always agree - one simply uses the most convenient one to do the calculation. Throughout the rest of this review I will assume that the indicated expected values always exist.

Defn II.3. $E(X^n)$, n a positive integer, is called the n^{th} moment of X .

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

Defn II.4. If $E(X) = \mu$ then we call $\sigma^2 \equiv \text{Var}(X) \equiv E[(X - \mu)^2]$ the Variance of X .

Defn II.5. If X_1 and X_2 are two random variables with $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$ we call $\text{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$ the Covariance of X_1 and X_2 .

The following simple properties of expected values are easy to establish and extremely useful. Throughout, a, b, c , etc. are constants, X_1, X_2, \dots, X_n are random variables, $\mu_i = E(X_i)$, $i = 1, \dots, n$.

Properties of Expected Values

1. $E(aX_1 + b) = aE(X_1) + b$
2. $E(X_1 + X_2 + \dots + X_n) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$
3. If X_1, \dots, X_n are statistically independent and $h_i(X_i), i = 1 \dots n$ are ~functions of the X_i 's, then $E(h_1(X_1) \cdot h_2(X_2), \dots, h_n(X_n)) = E(h_1(X_1)) E(h_2(X_2)), \dots, E(h_n(X_n))$
4. $\text{Var}(X_1) = E(X_1^2) - \mu_1^2$
5. $\text{Var}(aX_1 + b) = a^2 \text{Var}(X_1)$
6. $\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j>1}^n a_i a_j \text{Cov}(X_i X_j)$
7. $\text{Cov}(X_1, X_2) = E(X_1, X_2) - \mu_1 \mu_2$
8. If X_i and X_j are statistically independent, $\text{Cov}(X_i, X_j) = 0$
9. If X_1, \dots, X_n are statistically independent, $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$
10. $\text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_j b_j \text{Cov}(X_i, X_j)$ where
 $\text{Cov}(X_i, X_j) = \text{Var}(X_i)$.

Defn II.6 $\sigma \equiv \sqrt{\text{Var}(X)}$ is called the standard deviation of X.

Defn II.7

$$\rho(X_1, X_2) \equiv \frac{\text{Cov}(X_1, X_2)}{[\text{Var}(X_1) \text{Var}(X_2)]^{1/2}}$$

is called the correlation coefficient between X_1 and X_2 . One can show that $-1 \leq \rho(X_1, X_2) \leq 1$ for all pairs X_1 and X_2 . The covariance and correlation are indicators of the degree of relationship between X_1 and X_2 . Notice, however, that while independence implies a covariance of 0 the converse is false, and that while $\rho = \pm 1$ implies perfect correlation, two variables could be perfectly correlated even through $\rho = 0$. (e.g., Let U be uniform on $[-1, 1]$, $Y = U^2$. $\text{Cov}(U, Y) = 0$ but clearly knowing U we can perfectly predict Y .) This is because the covariance is basically a linear statistical analysis. In other words, if $\rho=0$ for two random variables, it implies that the two random variables are not linearly correlated.

III. Miscellaneous Facts

Generally, distributions of functions of random variables can be hard to obtain. However, one important result concerns the distribution of

$$\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$$

when the X_i 's are independent. Note by properties of expected values, if

$$E(X_i) = \mu_i, \text{Var}(X_i) = \sigma_i^2 \text{ then } E(\bar{X}) = \sum_{i=1}^n \mu_i / n, \text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2.$$

In certain cases we can say more.

The Simple Central Limit Theorem: If X_1, X_2, \dots, X_n are independent random variables each with the same distribution, $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$ then for n large \bar{X} is approximately normally distributed with mean μ and variance σ^2/n . This Theorem can even be extended to cases where the X_i 's have different distributions and even to some restricted cases where the X_i 's aren't independent - i.e., \bar{X} still has a normal distribution.

In Monte Carlo studies we often need to create realizations of random variables with a prescribed distribution. An interesting fact is that if we can create realizations from a uniform distribution, we can, in principle, get realizations from any desired distribution. (i.e., We only need a uniform random number generator.) This is the content of the next theorem.

The Fundamental Theorem of Simulations: If $y = F(x)$ is continuous distribution function with inverse function $x = F^{-1}(y)$ and if U is a uniform random variable then $X = F^{-1}(U)$ is a random

variable with distribution function $F(x)$.

IV. Common Random Variables

1. Z is a standard normal random variable if it has density

$$f(Z) = \frac{e^{-1/2 z^2}}{\sqrt{2\pi}} \quad -\infty < Z \leq \infty$$

2. X is a normal random variable with mean μ and variance σ^2 if $Z = (X - \mu)/\sigma$ is a normal random variable. If $Y = zx + b$, (a and b constants) then Y is also a normal random variable.

3. If X_1, X_2, \dots, X_n are independent normal random variables and a_1, a_2, \dots, a_n are constants then $\sum_{i=1}^n a_i X_i$ is a normal random variable.

4. If $X = \ln W$ is a normal random variable we say W has a log-normal distribution. If $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ then

$$E(W) = e^{\mu + \sigma^2/2}$$

$$\text{Var}(W) = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1)$$

If W is a log-normal then so is $1/W$, since $\ln(1/W) = -\ln W$ and by 2, with $a = -1$, $b = 0$, $-\ln W$ is normal.

5. U is a uniform random variable on $[a, b]$ if U has density₁

$$f(u) = \begin{cases} 1/(b - a) & a \leq u \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$E(U) = (a + b)/2, \quad \text{Var}(U) = (b - a)^2/12.$$

6. X is a gamma random variable if it has density

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \frac{e^{-x/\beta}}{\beta^\alpha} x^{\alpha-1} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where} \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

$$E(X) = \alpha\beta, \quad \text{Var}(X) = \alpha\beta^2.$$

If $\alpha = 1$ we have a negative exponential random variable.

7. X_1 and X_2 have a bivariate normal distribution with $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $\text{Var}(X_1) = \sigma_1^2$, $\text{Var}(X_2) = \sigma_2^2$ if they have joint density

$$f(x_1, x_2) = \frac{e^{-Q(x_1, x_2)}}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}}$$

$$\text{where } Q(x_1, x_2) = \frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right]$$

8. $\vec{X} = (X_1, \dots, X_n)$ has a multivariate normal distribution with mean vector $\vec{\mu}$ and covariance matrix Σ (the elements of Σ are $\text{Cov}(X_i, X_j)$) if it has joint density

$$f(\vec{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}) \Sigma^{-1} (\vec{x} - \vec{\mu}) \right\}.$$

Here $|\Sigma|$ = the determination of Σ , $(\vec{x} - \vec{\mu})'$ is the transpose of $(\vec{x} - \vec{\mu})$ (hence a column vector) and Σ^{-1} is the inverse matrix of Σ . A simpler way to characterize a multivariate

normal is to say $\sum_{i=1}^n a_i X_i$ is a one-dimensional normal for all constants

a_1, \dots, a_n .

9. If Z_1, \dots, Z_n are independent standard normal random variables then Z_i^2 , for each i , has chi-square distribution with 1 degree of freedom and $\sum_{i=1}^n Z_i^2$ has a chi-square distribution with n degrees of freedom (d.f.). (A chi-square with m d.f. is the same as a gamma with $\beta = 2$ and $\alpha = m/2$.)

V. Some Classical Statistical Results

X_1, X_2, \dots, X_n represents independent, identically distributed random variables each with $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$ in what follows.

Defn V.1 A 95% confidence interval for a parameter θ is a random interval $[A(X_1, \dots, X_n), B(X_1, \dots, X_n)]$ such that $P[A(X_1, \dots, X_n) \leq \theta \leq B(X_1, \dots, X_n)] = .95$. The probability is on the X_i 's not θ in the classical interpretation.

$\bar{X} = \sum_{i=1}^n X_i/n$ is called the sample mean and $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n - 1$ the sample variance. If the X_i 's are normal random variables then $\sqrt{n}(\bar{X} - \mu)/S$ has a t-distribution with $n-1$ degrees of freedom and $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $(n-1)$ d.f.

The t and chi-square distributions can be used to test hypotheses about μ and σ^2 , respectively. If the X_i 's are not normal but have relatively symmetric distributions then, for n large enough, tests and confidence intervals based on the t and chi-square are still reasonably good.

Another use of the chi-square which differs from its use in testing hypotheses about σ^2 is for goodness-of-fit tests. Suppose the X_i 's come from a population that we hypothesize has distribution $F(x)$. To test whether this is the case we can divide the axis into k subintervals, say

$(-\infty, \infty) = \bigcup_{i=1}^k (a_i, a_{i+1})$. Let N_i = the number of X_i 's in (a_i, a_{i+1}) and let $e_i = F(a_{i+1}) - F(a_i)$ be

the probability X_i lands in this interval if our hypotheses is correct. Then $\sum_{i=1}^k \frac{(N_i - e_i)^2}{e_i}$ has a chi-

square distribution with $k-1$ d.f. Suppose we let $X_{.05}$ be that value such that $P(W > X_{.05}^2) = .05$ where W has a chi-square distribution with $k-1$ d.f. Then we have a test of level .05 (Type 1 error). This means if we use this procedure many times in our lifetime, then about 5% of the time we reject the hypothesis when in fact it is true. To apply the chi-square $N_i \leq 5$ for each i . If $F(x)$ depends on some unknown parameters that are estimated from the data, one degree of freedom is lost for each estimated parameter.

Another procedure for goodness of fit tests which can be applied if the X_i 's are know to be continuous random variables is the Kolmogorov-Smirnov test. This isn't as well known as the chi-square test but is actually nicer and easier to apply.

First form the Empirical Distribution Function,

$$\hat{H}(x) = \frac{\text{The Number of } X_i\text{'s} \leq x}{n} .$$

(n = total number of observations.)

$$\text{Then let } D = \max_{-\infty \leq x \leq \infty} |F(x) - \hat{H}(x)| .$$

Then critical values for a .05 test have been tabulated based on D. For example, if $n \leq 50$ the critical value of D is $1.36 / \sqrt{n}$ for a test of level .05. That is, we reject the hypothesis that F(x) is the true distribution for the X_i 's if $D \leq 1.36 / \sqrt{n}$. Note no grouping is needed here and the empirical distribution function is useful to have in any case.