

FYI Handout #3: More on Probability and Expectations

A random variable X allows you to characterize the outcomes of a random phenomenon in terms of a probability distribution.

A discrete random variable takes only a finite number of values x , each of which has a certain probability $p(x)$. The probabilities satisfy the conditions (1) $p(x) \geq 0, \forall x$ and (2) $\sum_x p(x) = 1$. An example is the random variable $X =$ number of heads when flipping three coins. Its probability distribution is the following.

Values of X	0	1	2	3
Probability	1/8	3/8	3/8	1/8

A continuous random variable takes all the numerical values within a certain range. Its probability distribution is characterized by a density function, or pdf, $f(x)$. The probability that the random variable X takes values between x_1 and x_2 is represented by

the area under the pdf: $P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x) dx$. The pdf satisfies the conditions (1)

$f(x) \geq 0, \forall x$ and $\int_{-\infty}^{+\infty} f(x) dx = 1$. Two common pdfs are the uniform distribution $f(x) = 1$

for $0 \leq x \leq 1$ ($= 0$ otherwise) and the normal distribution $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

A probability distribution can be characterized by its *moments*, the most important of which are the mean and the variance.

Define the mathematical expectation or mean of a random variable X as

$E(X) = \mu_X = \int_{-\infty}^{+\infty} xf(x) dx$ (for a continuous r.v.) or $E(X) = \mu_X = \sum_x xp(x)$ (for a discrete

r.v.). This is simply a weighted average of the values of X , where the weights are the probabilities of each value of X .

Properties of the mathematical expectation or mean of a r.v: (1) if a is a constant, $E(a) = a$; (2) if b is a constant and X is a r.v., $E(bX) = bE(X)$; (3) if X and Y are two r.v., $E(X + Y) = E(X) + E(Y)$.

The variance of a r.v. is defined as $\sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x) dx$ (for a continuous r.v.) or $\sigma_X^2 = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 p(x)$ (for a discrete r.v.).

Properties of the variance of a r.v: (1) if a is a constant, $\sigma_a^2 = 0$; (2) if b is a constant and X is a r.v., $\sigma_{bX}^2 = b^2\sigma_X^2$; (3) if X and Y are two independent r.v., $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$.

Often, a random phenomenon of interest is characterized by more than one variable. For example, a medical study might be interested in the relationship between blood level of cholesterol and daily consumption of saturated fat. In this case the main interest is not about the individual probability distributions of these two variables; the main question is whether and how these variables are related. To study this type of random phenomenon, we need to introduce the concept of multivariate probability distribution.

A multivariate or joint probability distribution is a probability distribution for two or more r.v. For simplicity, let us consider the case of two random variables, X and Y . If the variables are discrete, each pair of values (x,y) is characterized by a probability $p(x,y)$. The probability distribution satisfies (1) $p(x,y) \geq 0, \forall(x,y)$ and (2) $\sum_x \sum_y p(x,y) = 1$.

For an hypothetical example, consider the joint distribution of $X =$ level of happiness and $Y =$ level of education:

Y = level of education	X = level of happiness		
	0	1	2
1	0.02	0.08	0.05
2	0.02	0.28	0.25
3	0.01	0.13	0.16

If the two r.v. are continuous, their joint distribution is represented by a pdf $f(x,y)$, that satisfies (1) $f(x,y) \geq 0, \forall(x,y)$ and $\int_x \int_y f(x,y) dx dy = 1$.

The joint probability distribution of X and Y embeds the individual probability distributions of X and Y . They are called *marginal* probability distributions and can be derived from the joint distribution. For the case of two continuous r.v. with pdf $f(x,y)$, the marginal pdf of X is obtained as $f(x) = \int_y f(x,y) dy$, and the marginal pdf of Y is

$$f(y) = \int_x f(x,y) dx .$$

In the discrete case, the marginal probability distributions of X and Y are obtained, respectively, as $p(x) = \sum_y p(x,y)$ and $p(y) = \sum_x p(x,y)$. If in the above example, we can obtain the marginal distributions of happiness and education as

Values of X (happiness)	0	1	2
Probability	0.05	0.49	0.46

Values of Y (education)	0	1	2
Probability	0.15	0.55	0.30

Now we can define the concept of statistical independence more precisely: Two random variables X and Y with joint pdf $f(x,y)$ and marginal distributions $f(x)$ and $f(y)$ are statistically independent if and only if $f(x,y) = f(x)f(y)$ for all (x,y) .

In the above example it can be readily verified that happiness and education are not independent because, for example, $p(X=0) * p(Y=1) = 0.05 * 0.15 = 0.0075$, which is different from $p(X=0, Y=1) = 0.02$.

The joint probability distribution of X and Y not only contains information of the individual (marginal) probability distributions of X and Y : it also contains information on how each random variable is distributed for different values of the other random variable. This additional information is summarized by the conditional probability distributions.

If X and Y are discrete r.v., the conditional distribution of X given $Y = y$ is defined as

$p(x | Y = y) = \frac{p(x, Y = y)}{p(Y = y)}$, in words, the joint probability of X and Y when $Y = y$ divided

by the marginal probability of Y when $Y = y$. This notation emphasizes the fact that there is a different conditional distribution of X for each value of Y . In the above example, the conditional probability of happiness for the less and the most educated are, respectively,

Values of X (happiness)	0	1	2
Conditional probability for Y (education) = 1	0.1333	0.5333	0.3333

Values of X (happiness)	0	1	2
Conditional probability for Y (education) = 3	0.0333	0.4333	0.5333

Notice that these conditional distributions of X are different from the marginal distribution of X : the distribution of happiness varies according to the level of education. As a matter of fact, finding differences between conditional and marginal distributions is another way of verifying that the two random variables are not independent.

If X and Y are continuous r.v., the conditional probability distribution of X given that $Y = y$ is represented by the conditional pdf $f(x | y)$, which is defined as $f(x | y) = \frac{f(x,y)}{f(y)}$.

Because the conditional distributions may differ from the marginal distribution, their respective means and variances will also vary. To account for these differences, we can compute conditional means and variances. For example, the conditional mean of X given

$Y = y$ is defined as $E(X | y) = \mu_{x|y} = \int_{-\infty}^{+\infty} xf(x | y)dx$.

A last point to consider is the mathematical expectation of a function of random variables $g(x,y)$. If X and Y are continuous r.v. with joint pdf $f(x,y)$ we can compute

$E[g(x,y)] = \int \int g(x,y)f(x,y)dxdy$. If the two random variables are discrete, we can

compute $E[g(x,y)] = \sum_x \sum_y g(x,y)p(x,y)$. This formula is useful to derive several

properties of expectations, including the ones we have already seen. For example if $g(x,y) = x + y$, the formula allows us to find the mean of the sum of two random variables. If $g(x,y) = [(x + y) - (\mu_X + \mu_Y)]^2$, the formula allows us to find the variance of the sum of two random variables X and Y .

Finally, if we define $g(x,y) = (x - \mu_X)(y - \mu_Y)$, the formula defines the *covariance* of X and Y . The covariance, which is denoted as σ_{XY} , is closely related to the correlation coefficient. As a matter of fact, the correlation of two random variables X and Y is

defined as $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$, where σ_X and σ_Y are, respectively, the standard deviations of X

and Y . (Notice that in class we have studied the *sample* correlation coefficient. What is the difference between the sample correlation coefficient and the correlation coefficient for two random variables?)

Of the many propositions that can be proved about covariances, two are particularly important. (1) If X and Y are two independent r.v., then its covariance σ_{XY} is zero. (2) If X and Y are two r.v. (not necessarily independent), then $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$ or, using the definition of correlation coefficient, $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_X \sigma_Y \rho_{XY}$. Combining (1) and (2) we derive the previous result about the sum of the variances of two independent random variables.

This last result can be applied, for example, to study the variance of the return of a portfolio of stocks. If X and Y are the returns of two individual stocks, and b and c are the weights of each stock in the portfolio ($b + c = 1$), we can derive using the above result along with the other properties of variances that $\sigma_{bX+cY}^2 = b^2 \sigma_X^2 + c^2 \sigma_Y^2 + 2bc \sigma_X \sigma_Y \rho_{XY}$. This formula computes the variance of the portfolio's return. Because individual stocks tend to be positively correlated ($\rho_{XY} > 0$), the assumption of independence would lead us to underestimate the risk of a portfolio, so the more general formula is appropriate in this context.

It should be pointed out that the formulas for expectations and variances of functions of random variables presented above are valid only for linear functions, but not for nonlinear functions. So while we can say, for example, that $E(aX) = aE(X)$, it is not true that $E[\log(X)] = \log(E[X])$. In this case we can still compute $E[\log(X)] = \sum \log(x)p(x)$ if X is a discrete random variable, or $E[\log(X)] = \int \log(x)f(x)dx$ if X is a continuous random variable.

-Application of expectations: Properties of an estimator

Let V represent an estimator of the population parameter θ . Because estimators vary randomly in repeated sampling, V is a random variable, while on the other hand, the population parameter θ is a fixed number.

Definition: The estimator V is said to be unbiased if and only if $E(V) = \theta$. If $E(V) \neq \theta$, V is a biased estimator of θ . Its bias is defined as $Bias(V) = E(V) - \theta$.

Example 1: Suppose a random variable X is distributed with a mean of $E(X) = \mu$. Now consider you take a simple random sample of size $n = 2$. Let X_1 and X_2 denote, respectively, the first and second element of the sample. In a simple random sample each element of the sample has the same probability of being selected. Therefore, X_1 and X_2 are independent random variables (Why?). Moreover, the two random variables have the same mean: $E(X_1) = E(X_2) = \mu$ (Why?).

Now consider the sample mean $\bar{x} = (X_1 + X_2)/2$. Because \bar{x} is a linear function of random variables, the sample mean is itself a random variable. Is \bar{x} an unbiased estimator of the population mean? The answer is yes. Applying the formula for the expectation of a linear function of random variables we get $E(\bar{x}) = [E(X_1) + E(X_2)]/2 = (\mu + \mu)/2 = \mu$.

As an example of a biased estimator of the population mean, consider the sum $U = X_1 + X_2$. Because $E(U) = [E(X_1) + E(X_2)] = 2\mu$, U is a biased estimator of μ . The bias is $2\mu - \mu = \mu$. Because the bias is positive, this estimator tends to *overestimate* the population mean.

Example 2: Ever wondered why the sample standard deviation is defined with $n-1$ in the denominator? The reason is that $s^2 = \frac{1}{n-1} \sum (X_i - \bar{x})^2$ is an unbiased estimator of the population variance: $E(s^2) = \sigma^2$ (The proof is beyond the level of this course: because this estimator is a quadratic function of random variables, we cannot apply the formula for expectations of linear functions of random variables we have learned.)

Accepting that s^2 is unbiased, what would be the bias of the mean squared deviation, defined as $MSD = \frac{1}{n} \sum (X_i - \bar{x})^2$? Multiplying and dividing by $n-1$, we get

$$MSD = \frac{n-1}{n} s^2 \Rightarrow Bias(MSD) = E(MSD) - \sigma^2 = \frac{n-1}{n} E(s^2) - \sigma^2 = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2.$$

Because the bias is negative, the MSD tends to *underestimate* the population

variance. Notice, though, that the bias gets smaller and smaller as the sample size increases.

Definition: If V and U are two unbiased estimators of the parameter θ , V is said to more efficient than U if V has a smaller variance than U : $Var(V) < Var(U)$.

Example: Consider the following two estimators of the population mean, the sample mean $\bar{x} = (X_1 + X_2)/2$ and the weighted average $U = (1/3)X_1 + (2/3)X_2$. It is easy to verify that the second estimator is also unbiased (try it). Which one is more efficient? Let σ^2 denote the population variance of X . Because X_1 and X_2 are independent random variables with the same variance (σ^2), it follows from the formula for the variance of a linear function of independent random variables that $Var(\bar{x}) = [Var(X_1) + Var(X_2)]/2^2 = (\sigma^2 + \sigma^2)/4 = \sigma^2/2$. Similarly, we get $Var(U) = (1/3)^2 Var(X_1) + (2/3)^2 Var(X_2) = (1/9)\sigma^2 + (4/9)\sigma^2 = (5/9)\sigma^2$. Since $(1/2)\sigma^2 < (5/9)\sigma^2$, \bar{x} is more efficient than U .

When comparing two unbiased estimators, the most efficient is preferable, but how do we compare two biased estimators?

Definition: The mean squared error of an estimator V of the parameter θ is defined as $MSE(V) = E(V - \theta)^2$.

If V is an unbiased estimator of θ , the mean squared error is equal to the variance. If V is biased, we add and subtract $E(V)$ in the definition: $MSE(V) = E[V - E(V) + E(V) - \theta]^2 = E\{[V - E(V)]^2 + 2[V - E(V)][E(V) - \theta] + [E(V) - \theta]^2\}$. First notice that by definition $V - E(V) = 0$, so the second term goes away. Second, notice that $[E(V) - \theta]^2$ is not a random variable but a constant: the bias of V squared (recall definition of bias above). Therefore, we can write $MSE(V) = E[V - E(V)]^2 + [E(V) - \theta]^2 = Var(V) + [Bias(V)]^2$.

The mean squared error criterion allows you to compare two biased estimators. The one with the lowest MSE is preferable.

Definition: An estimator is consistent if its MSE approaches zero as the sample size increases.

Example: \bar{x} is consistent. Because it is an unbiased estimator, its MSE equals its variance. When the sample size is n , the variance of \bar{x} is $Var(\bar{x}) = (1/n)^2 \sum Var(X_i) = (1/n)^2 \sum \sigma^2 = (1/n)^2 n\sigma^2 = \sigma^2/n$. Clearly, this expression goes to 0 as $n \rightarrow \infty$.

In the case of a biased estimator, consistency requires that both the variance and the bias converge to zero as the sample size increases.