Probability - Many Random Variables

Many Random Variables

We already mentioned that we will have to work with more than one random variable at a time. That is because the model we will use treats a sample of n observations as the observations of n random variables

Joint Distribution

In general, the model will consider these variables as a group, so that events will have the form

$$\{\omega | X_1(\omega) \in A_1, X_2(\omega) \in A_2, ... X_n(\omega) \in A_n\}$$

and the corresponding probabilities are often called the *joint distribution of the variables* X_1, X_2, ...X_n. Problems involving several variables are more complicated, and it is thus a welcome situation when the joint distribution does not require more than the individual distributions to be determined. This is a *very special case*, when the variables are said to be *independent*, and cannot be taken for granted as the right model, unless we make sure the circumstances warrant it.

Independent Random Variables

In general, the joint distribution of our observations could be difficult to calculate from the distribution of X. In a special case, though, the relation is extremely simple. This is also the special case where we obtain the most information about the unknown distribution from the observed sample. The special case we just mentioned is when the joint distribution is such that it is always true that

$$P[X_1 \in A_1, X_2 \in A_2, ... X_n \in A_n] = P[X_1 \in A_1] P[X_2 \in A_2] ... P[X_n \in A_n]$$

If that's the case, we say that the Variables are *Independent*.

The intuitive idea behind Independence is the following: the observation of any of the variable(s) does not affect the observation(s) of any other(s): whether we observe them one at a time, or together, the probabilities are not affected at all.

We generally assume, for example, that rolling two dice results in two outcomes that are independent, but if, say, the two dice were loaded with magnets, the outcome of one would definitely affect the other, and independence would not be a reasonable assumption.

Independence is not an obvious feature, and should be assumed in a given situation only when it is clear that it is a fair description. We will return in a little more detail in the next module to this question.

More About Independence

There is another way to look at independence, and it goes through the useful concept of *conditional probability*. The idea is the following: suppose we are interested in the value of some quantity (a simple example would be the points coming out of the toss of a die – let's call this outcome, a random variable, X). Suppose now that we are told the result of some *other* random variable, Y (for example, we are told whether the point was even or odd, but not what it was precisely). We are now armed with some more information than before (note that probability is relevant only *when we don't know*: we can talk about the probability of winning a lottery with our numbers, but once the winning numbers are known, there is no "probability": we either won or we lost), hence, we should be ready to modify our previous assessment. Indeed, it turns out that the following is a good choice for a modified probability: if we were interested in $P[X \in A]$, and are told that $Y \in B$, then the *conditional probability of the first event, given the second is defined as*

$$P[X \in A | Y \in B] = \frac{P[X \in A, Y \in B]}{P[Y \in B]}$$

The numerator is the probability of the joint events involving X and Y. In our die example, we would have that the probability of, say, tossing a 6, is, without any other information, $\frac{1}{6}$ (assuming the die is "fair"). The probability of, say, tossing an even number is $\frac{1}{2}$, since there are as many even as odd points. The probability of tossing a 6 **and** an even number is still $\frac{1}{6}$, since 6 is an even number. Hence

$$P[X=6|Y=\text{even}]=\frac{\frac{1}{6}}{\frac{1}{2}}=\frac{2}{6}=\frac{1}{3}$$

On the other hand, if we had been told that the point was odd, the joint probability of tossing a 6 **and** an odd number would be zero, and so would the conditional probability.

There are many ways to justify this choice, the simplest one being the observation that frequencies do exactly that.

Remark 1. Conditional probabilities may refer to events that are both yet to come, and we use this approach (intuitively) any time we play "what if". For example, we might want to assess the probability that our team will win tomorrow's game. Given the opponent is known to play worse in bad weather, we may want to break down our estimation in two: probability of win in good weather, and probability of win in bad weather.

Looking at the conditional probability formula, you can see right away that if X and Y are independent, then $P[X \in A|Y \in B] = P[X \in A]$, that is, the additional information has no effect on our assessment of the likelihood of the event we are interested in.

From the same formula, you can also see that if we know that $P[X \in A|Y \in B] = P[X \in A]$, then the two events are independent. The definition we gave of independence of random variables can also be stated as "two random variables are independent if $P[X \in A|Y \in B] = P[X \in A]$ for all A and B". In other words, independent variables carry no information about each other.

Remark 2. Please, note that *independent* is quite different – in fact almost the opposite – of *incompatible*. Two events are incompatible if they cannot occur together, for example the toss of a 6 and of an odd point. As we observed in the example above, the probability of tossing a 6 changes dramatically, if we know that the point was odd!

Indices for Sums of Random Variables

Expectation

Here is an important property of the expectation (it is easy to prove it from the formula above): for any collection of random variables $X_1, X_2, ..., X_n$, and real numbers $a_1, a_2, ..., a_n$,

$$E\left[\sum_{k=1}^{n} a_k X_k\right] = \sum_{k=1}^{n} a_k E[X_k]$$

Thus, for example

$$E[2X-3Y] = 2E[X] - 3E[Y]$$

More interestingly (think of the case when $a=\frac{1}{n}$)

$$E\left[\sum_{k=1}^{n} a X_k\right] = a\sum_{k=1}^{n} E[X_k]$$

Variance

Things don't work as simply for the variance. We can check, for example, for two random variables,

$$E\Big[(X+Y)^2\Big] - (E[X]+E[Y])^2 = E\big[X^2\big] + 2E[XY] + E\big[Y^2\big] - (E[X])^2 - 2E[X]E[Y] - (E[Y])^2 =$$
$$= \Big\{E\big[X^2\big] - (E[X])^2\Big\} + \Big\{E\big[Y^2\big] - (E[Y])^2\Big\} + 2\{E[XY] - E[X]E[Y]\}$$

The first two terms are the variances of X, and Y, respectively. However there is a third term, called the *covariance* of X and Y, $\mathrm{Cov}[X,Y]$, which has to be added. Incidentally, if you repeat the calculation for $X-Y$, you will find a similar formula:

$$\mathrm{Var}[X-Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\mathrm{Cov}[X,Y]$$

As far as the effect of multiplying by a constant, we can quickly see that

$$E\Big[(aX - E[aX])^2\Big] = a^2 E\Big[(X - E[X])^2\Big]$$

that is, it causes a multiplication of the variance by the square of the constant.

Variables with zero covariance are called *uncorrelated*, and, for them, variances just add (note that you still have an addition, even if the variables are subtracted – if you want to reduce the variance of a sum, you need the covariance term to provide a negative contribution, a classic observation in the theory of financial investments). It so happens that **independent variables are uncorrelated, but uncorrelated variables need not be independent**. Please, remember this statement, especially late in the next part, when we work on estimating correlations.

But, Wait A Minute

It is easy to set up examples where uncorrelated variables are very far from independent (the classic one is to take X, a random variable such that $E[X] = E\big[X^3\big] = 0$, and $Y = X^2$), so, indeed, un-correlation is not a reliable marker for independence. You might wonder why, nonetheless, this does not seem to be recognized as it should (you will notice this problem if you scan technical literature). Fact is, **for very few special distributions**, if two variables are not correlated, they are independent. This is, in particular, the case for two variables X, Y that have a **joint Gaussian (normal) distribution**. This is a much stronger condition that, X and Y being Gaussian: it is a condition on **joint** probabilities like $P[a < X < b, c < Y < d]$, and should be validated in any specific circumstance. As you will notice, going over much of the literature, assumptions such as this are often made, explicitly or not, without worrying too much about verifications. It is true that there is a "multi-dimensional Central Limit Theorem" that allows us to assume, when it is applicable, that variables are (approximately) jointly normal, but, as always, there are conditions for this to be a solid statement, and it is always a good idea to check whether these conditions are realistic for the situation at hand.

Sampling

We still need to specify how a sample can help us describe the distribution in more practical terms. To do this, we need to specify a mathematical model for the sampling operation, and look at the connections that this model establishes. The following is a strict definition of *simple random sampling*, that we may have, in certain circumstances, to weaken. However, weakening these conditions complicates the problem considerably, and we will not pursue this direction in our course.

As suggested above, we assume there is a random variable, let's call it X here, that models the process of observation: you are going to observe a value between a and b, for example, with a "likelihood", measured by the probability $P[a < X < b]$. A random sampling of this variable is a sequence of observations, *independent of each other, and in conditions such that each one is governed by the same probability distribution*.

Formally, we are looking at a set of random variables, X_1, X_2, ..., X_n, all independent of each other, and all with the same distribution, the distribution of X, which we are interested in. Note that it may not be easy to implement these conditions, and the failure of an opinion poll to predict accurately is always a good reminder that a lot of things can go wrong. As examples,

- The observations have to be *independent*: none should have any effect on any other. That is why a poll will never use two individuals from the same household, but there are more subtle ways in which independence might not hold.

- The observations have to be governed by the same distribution, and this distribution is the one we are trying to observe. This would work if individuals were chosen from a big urn containing all Americans, by a blindfolded child. This is not practical, and there are various methods to try to approximate that situation, but they all have some problem. This is an extremely tricky issue, and, in fact, it can be almost impossible to be totally sure we have managed it. Incidentally, this is the point of failure of "phone-in" or "Internet" polls: the people who call are not a "random sample", in that they choose to call, and will do so because they have some particular motive to do so – they are not a sample from the American population, but are a biased sample from the limited population that is following the program or the Internet site.

- In some situations, simple random sampling may not be the best practical choice. A popular modification of simple sampling is a more elaborate method, *stratified* random sampling. The idea is to break up the total population into groups, to be sampled separately. This should help manage the complexities of sampling large populations, but brings in a number of new challenges, from the determination of the subgroups, to the best way to weigh the resulting separate outcomes.

- More generally, there is a variety of sampling methods that have been devised to overcome some of the difficulties in performing a simple random sampling. We will go back to this issue, and mention a few, at the end of our course, but it should be kept in mind that any simplification attained by other sampling methods comes with a seriously enhanced risk of introducing uncontrolled *bias* in the result.

Sample and Population

There are several results that help us get information on the distribution of X, given our sample. First of all, if we look at the sample as a collection of independent variables (that is, we are considering the specific observations we made), we have that all have the same distribution as X, so that they have the same expectation, the same variance, and so on. Hence,

$$E\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} E[X_k] = nE[X]$$

In particular,

$$E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = E[X]$$

That is the arithmetical mean of the sample has the same expectation of the random variable we are interested in.

However, our variables are also independent, and, hence, uncorrelated. That means that their variances just add:

$$\mathrm{Var}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathrm{Var}[X_k] = n\,\mathrm{Var}[X]$$

Now, remembering the effect of multiplying our variables by a constant, if we multiply each by $\frac{1}{n}$ (that is we look at the variance of their arithmetic mean), we find

$$\mathrm{Var}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\cdot n\,\mathrm{Var}[X] = \frac{\mathrm{Var}[X]}{n}$$

That is, the variance of the mean is reduced by a factor of n. Taking n larger and larger, we have a smaller and smaller variance. Now, the variance of a random variable is a rough measure of its dispersion, hence, if we take large enough samples, they will distributed with the same mean as our original variable, with a tiny variance. In other words, it will be extremely likely that the mean of a large sample will be very close to the expectation of the distribution!

Remark 3. This observation has a few important consequences. The main one is discussed in the next section. A less precise, but useful, usage is the fact that, whatever distribution you are sampling from, a single observation will have a distribution with a certain spread. If the variance is a useful measure of spread (and it often is, especially for reasonably symmetric, not-too-dispersed distribution), observing a good sample (n independent observations, all with the same distribution), and computing its arithmetic mean (the *sample mean*), results in a value which will be expected to be much closer to the theoretical mean (the *expectation* of the distribution) than a single observation.

An index that you may encounter is the "standard error": that's a best guess at $\sqrt{\frac{\sigma^2}{n}}$ (the standard deviation of the mean), obtained by taking the "sample variance" as a proxy for σ^2. You will notice that it is significantly smaller than the "sample standard deviation".

A Special Case: Sums of Independent Identically Distributed Normal Random Variables

If the variables X_k are distributed according to $N(\mu,\, \sigma^2)$, then $\sum_{k=1}^{n} X_k$ is distributed according to $N(n\mu,\, n\sigma^2)$, and $\frac{1}{n}\sum_{k=1}^{n} X_k$ is distributed according to $N\left(\mu,\, \frac{\sigma^2}{n}\right)$. That follows from the discussion above, and is a peculiarity of the normal distribution.

Note that, still denoting $Z_k = \frac{X_k - E[X_k]}{\sqrt{\mathrm{Var}[X_k]}}$, since $\mathrm{Var}(Z_k) = 1 = E\left[Z_k^2\right]$ (the expectations are 0), we have that the expectation of a χ_n^2 random variable (as you may recall, it is the distribution of the sum of the squares of n independent, identically distributed, standard normal random variables) is equal to n.

Limit Theorems

The Law of Large Numbers

The comment on the variance of the mean is the basis for much of our statistical tools, and goes, in a more precise form, by the name of *Law of Large Numbers*. We can state it, somewhat imprecisely, as

Theorem 4. *Given a sequence of independent identically distributed random variables, X_1, X_2, ..., X_n, their arithmetic mean, $\frac{1}{n}\sum_{k=1}^{n} X_k$ is as close to the common expectation, $E[X_k]$, as we wish, with as high a probability as we wish, provided we choose n high enough.*

A more precise discussion is to be found in the file on limit theorems in the "Special Topics" section. Since getting hold of the mean of a distribution is one of the main goals in statistics, this theorem gives the sample mean a prominent role.

The Central Limit Theorem

The arithmetic mean of our sample has an even more remarkable property. The proof of the following statement is not nearly as simple as the proof of the Law of Large Numbers, but the statement is the cornerstone of a huge portion of the statistical toolbox. For brevity of notation, we will write

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

and we denote by X a random variable with the same distribution as the X_k.

Theorem 5. *(Central Limit Theorem) For a sequence of independent, identically distributed random variables, we have that, if n is large enough, the difference*

$$P\left[a < \frac{\bar{X}_n - E[X]}{\sqrt{\mathrm{Var}[X]/n}} < b\right] - (\Phi(b) - \Phi(a))$$

is as small as we wish.

Note that the denominator in the first term is the standard deviation of the arithmetic mean. The function Φ in the formula is the cumulative distribution function of the standard normal distribution, which we already met before, as an example.

Variables that follow a normal distribution (also called a *Gaussian* distribution), are called, fairly enough, *normal*, or *Gaussian*. Thus, if we look at the arithmetic mean of a sample, **provided the sample is large enough**, it will have a distribution that is almost Gaussian. This is one reason for the dominance of methods based on the normal distribution in statistics.

There are two caveats we might want to notice. One is that all of the above makes sense only if the variance (and the mean) exists. Distributions that have no variance cannot be handled in this way. We will not worry about this case in our course, but it needs to be considered in general. The second, much more general, regards the vague phrase "if n is large enough". What is "enough"? Unfortunately, this depends heavily on the distribution we are starting from.

As a rule of thumb, the more our original distribution is symmetric around the expectation, the faster we will reach "normal territory". However, a badly skewed distribution might need really large samples, possibly too large, for a normal approximation to be viable.

For example, IBM has long relied, in its computer simulation software, on the fact that the *uniform distribution*, that is a distribution such that X can only take values in an interval $[a, b]$, with density $\frac{1}{b-a}$ (hence the probability of falling in a given interval is proportional to the length of the interval), is very symmetric around its expectation, $\frac{a+b}{2}$. A computation will show that choosing the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$, the corresponding uniform variable has expectation 0, and variance $\frac{1}{12}$. Hence, the sum of 12 such variables, if they are independent, will have expectation 0, and variance 1, and be approximately, normally distributed. This rough algorithm has worked very well over the many years it has been employed.

An important special case is the distribution of n Bernoulli random variables (as in the response to a 2-outcome question in a poll). As mentioned before, the sum of these variables has a binomial distribution. However, being the sum of independent variables, if n is large enough, treating the sum as a normal variable is perfectly acceptable. The "large enough" depends on the symmetry of the distribution, which, in this case, means that p has to be not too close to 0 or 1. In other words, the normal approximation should work fine for values of n that are fairly small, provided the "population" is not extremely biased one way or another. A popular rule of thumb is to require $np(1-p) > 10$ (remember that it is only a rule of thumb).