CHAPTER 3: Basics Of Probability and Statistics

The evaluation of weather forecasts involves the application of probability and statistics. In this chapter, we will briefly cover some basics of probability and statistics that will be useful in reading the rest of the book. This is by no means meant to be a complete treatment, but it will serve to provide basic information and clarify notation for the text. For more details, we encourage you to consult statistics textbooks, such as Hays and Winkler (1971) or Wilks (1995).

3.1 Probability

Probability represents a way to quantify uncertainty. It provides a powerful way to describe the likelihood of different events and allows us to show the degree of confidence we have in those events. We are particularly interested in probability because, as we will see later, it provides a convenient way to express the relationships between forecasts and events.

An *event* is one of a group of possible outcomes of a process. The process can be simple, such as whether a rolled die will come up with a particular number, or it can be very complex, such as whether it will rain tomorrow at a particular location. We can consider two kinds of events-*elementary* and *compound*. Elementary events cannot be decomposed into two or more subevents, while compound events can be. For instance, “it rained today” may be an elementary event, while “it rained today and the forecast for today was that it would rain” would be a compound event. The set of all possible elementary events is called the *sample space* or *event space*. A collection of all possible elementary events is called *mutually exclusive and collectively exhaustive* (MECE). No more than one of the events can occur (*mutually exclusive*) and at least one of the events must occur (*collectively exhaustive*). A set of MECE events completely fills the sample space.

3.1.1 Axioms of probability and their consequences

Three fundamental axioms define the rules of probability and allow us to associate probabilities with each of the events within the sample space.

1. The probability of any event is nonnegative.
2. The sum of the probability of MECE events is 1.
3. The probability of a set of mutually exclusive events occurring is the sum of their individual probabilities.

The first and second axioms imply that the range of probabilities for an event, E, is

$$0 \le p(E) \le 1 . \tag{3.1}$$

If $p(E) = 0$, the event will not occur and if it is 1, the event is certain. The second axiom can be simply written as

$$\sum_{i=1}^{I} p(E_i) = 1 , \tag{3.2}$$

where there are I MECE events. From the first and third axioms, we can see that, if event E_1 only occurs if E_2 occurs, then

$$p(E_1) \le p(E_2) . \tag{3.3}$$

It is useful to consider the union of the probability of two events, $p(E_1 \cup E_2)$. The union is when either E_1 or E_2 or both events occur. Thus, from the third axiom, we can write

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2) \tag{3.4}$$

where $p(E_1 \cap E_2)$ is the probability that both E_1 and E_2 occur. This is the intersection of E_1 and E_2. It is also known as the *joint probability* and can be written as

$$p(E_1 \cap E_2) = p(E_1, E_2) \tag{3.5}$$

Frequently, we are interested in the probability of an event, given that some other event has occurred or will occur. For instance, we will look at the probability of precipitation, given that a forecast was made that precipitation would occur, or conversely, we may want to know the probability that the forecast was for precipitation, given that precipitation was observed. These are examples of a *conditional probability*, where the conditioning event is the "given". Notationally, we will use the following convention:

$$p(E_1 | E_2) = p(E_1 \text{ occurring given that } E_2 \text{ has occurred or will occur}) \tag{3.6}$$

If E_2 has occurred or will occur, the probability of E_1 is $p(E_1|E_2)$. Using the intersection notation, we can write (3.6) as

$$p(E_1|E_2) = \frac{p(E_1 \cap E_2)}{p(E_2)} . \tag{3.7}$$

We can rearrange (3.7) to find the *multiplicative law of probability*

$$p(E_1 \cap E_2) = p(E_1|E_2)p(E_2) = p(E_2|E_1)p(E_1). \quad (3.8)$$

Two events are said to be *independent* if the occurrence or nonoccurrence of one does not affect the probability of the other. In other words, if E_1 and E_2 are independent of each other, $p(E_1|E_2)=p(E_1)$ and $p(E_2|E_1)=p(E_2)$. For that special case, the multiplicative law, (3.8), reduces to

$$p(E_1 \cap E_2) = p(E_1)p(E_2). \quad (3.9)$$

In general, (3.9) can be extended to more than two independent events as

$$p(E_1 \cap E_2 \cap E_3 \ldots E_N) = \prod_{i=1}^{I} p(E_i). \quad (3.10)$$

Another powerful relationship that can be derived from the fundamental axioms of probability is the *law of total probability*. If we have a set of events, E_i, for i=1,…,I, that is MECE and an event, X, that is also defined on the sample space, we can calculate the probability of X by summing the conditional probabilities,

$$p(X) = \sum_{i=1}^{I} p(X \cap E_i). \quad (3.11)$$

Substituting from the multiplicative law, (3.11) becomes

$$p(X) = \sum_{i=1}^{I} p(X|E_i)p(E_i). \quad (3.12)$$

This expression allows to calculate probabilities indirectly in the face of limited information. Note that (3.12) only holds if the events are MECE.

3.1.2 Bayes' Theorem

The multiplicative law and law of total probability allow us to develop Bayes' Theorem, a powerful tool to invert conditional probabilities and to develop a framework to update estimates of subjective probability when new information is received. Bayes' Theorem can be derived by again considering a set of events, E_i, for i=1,…,I, that is MECE and an event, X, that is also defined on the sample space. From (3.8), we can write two expressions for the relationship between X and any E_i, using the joint probability $p(X,E_i)$

$$p(X, E_i) = p(X|E_i)p(E_i) = p(E_i|X)p(X). \quad (3.13)$$

Combining the two right equations, rearranging and applying (3.12), we get

$$p(E_i|X) = \frac{p(X|E_i)p(E_i)}{p(X)} = \frac{p(X|E_i)p(E_i)}{\sum_{j=1}^{J} p(X|E_j)p(E_j)}. \quad (3.14)$$

Bayes' Theorem can be applied from either of the two primary approaches to the interpretation of probabilities. The first is the so-called *frequentist* approach. It takes the view that probabilities represent the long-run frequency of an event. In other words, in a very large number of repeated trials, the probability is identical to the observed frequency. If a fair six-sided die is tossed 1,000,000 times, the probability of any side coming up will approach 1/6.

The second approach to interpreting probabilities is the *subjective* interpretation. In a sense, we don't get identical repeated trials for weather events (as well as many other things), so that applying the frequentist approach exactly is difficult. The subjective interpretation believes that the probability represents the degree of belief by a particular individual about the occurrence of an event. Many areas, including weather forecasting, have people who are capable of reliably estimating the probabilities of events in the future. It is possible that two well-trained, experienced people may come up with different subjective probabilities for a particular event on a particular day. That does not mean that one or the other is necessarily wrong (or right!). This, however, does not mean that individuals are free to select any values they choose. Subjective probabilities must obey the axioms of probability and the principles that are derived from them.

Although either approach can be used with Bayes' Theorem, the subjective interpretation has a particularly rich history of application, to the extent that it is sometimes called the *Bayesian* interpretation. Consider the problem of forecasting whether precipitation will occur or not. Since there are two states in this problem, I=J=2 in (3.14). In the absence of any other information, a weather forecaster may well use the climatological probability of precipitation [$p(E_1)$] on a day as an initial estimate of the probability. This estimate is referred to as the *prior probability* in the Bayesian approach. Now consider what happens if the forecaster knows the probability of the dewpoint temperature being above a certain threshold, say 15 °C, given that precipitation occurs on a day. In the notation of (3.14), we can write the conditional probabilities, $p(X|E_i)$ for i=1 (precipitation) or 2 (no precipitation). In the Bayesian approach, these conditional probabilities are referred to as *likelihoods*. If, now, the forecaster observes dewpoint temperatures above 15 °C, (3.14) can be used to revise the estimate of the probability of precipitation. The modified probability is referred to as a *posterior probability.* The amount of modification depends on the strength of the association between the precipitation and dewpoint temperature (through the likelihoods). If other factors that are associated with the likelihood of precipitation, such as numerical model guidance, also become available to the forecaster, additional revision of probabilities can take place.

3.2 Statistical Basics

3.2.1 Describing data distributions

We will often need to describe distributions of data of various kinds. For large data sets, it is impractical and unenlightening to list all values. Instead, we will use summary measures of the data to describe its important properties. Among these properties that we will want to use are the *location*, *spread*, and *symmetry* of the data set. In simple terms, the location refers to an estimate of some central tendency of the data. The spread refers to how tightly bunched the data are around the center, and the symmetry to how evenly the data are distributed around the center.

There are two approaches that are used generally to estimate these values. The first is called the *parametric* approach. In it, we will use the values of every member of the data sample to estimate the properties describing the sample. The parametric estimate of location is the *sample mean*, $\bar{x}$. It is simply the sum of all of the values divided by the number of values and is given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i , \tag{3.15}$$

where N is the number of elements in the sample and x_i is the value of the i-th sample. The *sample variance*, s^2, is given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2 . \tag{3.16}$$

The *sample standard deviation*, s, is simply the square root of (3.16) and is the parametric descriptor of the spread. Symmetry is measured by the *sample skewness coefficient*, γ,

$$\gamma = \frac{\frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^3}{s^3} . \tag{3.17}$$

The sample mean, standard deviation, and skewness coefficient are referred to as *moments-based* measures, since they include expressions of the form $(x_i - \bar{x})^{k+1}$, representing the k-th moment of the distribution. Similar expressions can, in principle, be generated for any value of k. The parametric approach is very powerful and can be used to describe any distribution of data. The calculations are easy to carry out, even without powerful computers. To the extent that the estimates are good, they can effectively reduce the number of values that one needs to consider to describe the data set down to a very small number.

The parametric approach, however, suffers in that it is neither *robust* nor *resistant*. Robust measures are ones that perform well in most circumstances. They do not depend on assumptions about the data distribution. The sample mean, (3.15), is a good estimate of the center of a data set that is described by the Gaussian distribution (the familiar bell-shaped curve), but is not a good estimate for highly skewed data. Resistant statistics are not sensitive to outliers in the data set. Again, the sample mean, standard deviation, and skewness are sensitive to outliers. This is of importance to us not just because of the possibility of extreme values that are real, but also because of the possibility of incorrect data. For instance, if a set of precipitation observations is supposed to be recorded in centimeters, but one of the values actually represents precipitation in millimeters, the values of the mean, standard deviation, and skewness from the sample will be dramatically affected by the presence of the one bad observation. This is particularly true for the standard deviation and skewness since they use higher-order exponents on the departures from the mean.

The non-parametric approach to describing data distributions makes no assumptions about the nature of the distribution and produces estimates that are both robust and resistant. It is computationally more intensive, since it requires the sorting of the data set, but the development of sorting algorithms for computers (Press et al. 1986) means that even large data sets can be described using the non-parametric approach. In order to see the values that can be used as the non-parametric equivalents of the mean, standard deviation, and skewness, let us start by consider a list of N data values in a sample sorted from largest to smallest. The sample *quantile*, q_p, that exceeds the fraction, p, of the data set is given by the value in the data set associated with

$$p = \frac{(i-1)}{(N-1)}. \tag{3.18}$$

For instance, the value

3.2.2 Probability distribution functions

The PDF of the beta distribution, which depends on the two parameters, a and b, is given by

$$f(x) = \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] x^{a-1}(1-x)^{b-1} \tag{3.19}$$

for $0 \le x \le 1$, with a and b greater than 0. The mean and variance of the beta distributions are given by

$$\mu = \frac{a}{a+b} \tag{3.20}$$

and

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.21}$$

The parameters, a and b, can be estimated by

$$\hat{a} = \frac{\bar{x}^2(1-\bar{x})}{s^2} - \bar{x} \tag{3.22}$$

and

$$\hat{b} = \frac{\hat{a}(1-\bar{x})}{\bar{x}}. \tag{3.23}$$