More on Probability

Laws of Probability

Set Notation

As mentioned earlier, determining the probability to actual events is not always easy. Nevertheless, it is often useful to seek an understanding of the rules of probability. To illustrate some of these rules, we introduce *set notation*. In so doing, let the *space* is a list of all possible events. Let *an event* represent any occurrence in a sample space.

For illustrative purposes, let us consider the sample space of pet ownership. Let A represent the event of dog ownership. Let B represent the event of cat ownership. We denote the probability of dog ownership as *Pr*(A) and the probability of cat ownership as *Pr*(B).

The logical term *and* represents the *joint occurrence* of events, indicating that both conditions must be true. For instance, in our illustrative example, "A and B" would represent ownership of both a dog and a cat. The probability of dog and cat ownership (together) is denoted *Pr*(A and B).

The logical term *or* represents the *union* of events, indicating that one or the other condition is true. For our illustrative example "A or B" represents ownership of either a dog or a cat (or both).

To illustrate the laws of probability, let *Pr*(A) = 0.3 and *Pr*(B) = 0.35. Also, let us assume that the ownership of both a dog and cat, *Pr*(A and B) = 0.15. From this information, all other occurrences can be deterimed.

Complements

The *complement* of an event represents that it does *not* occur (the logical "not"). In our illustrative example, "not A" represents that the unit does not own a dog. The *rule of complements* states:

$$Pr(\text{not A}) = 1 \;!\; Pr(\text{A})$$

Since *Pr*(A) = 0.3 in our illustrative example, *Pr*(not A) = 1 ! 0.3 = 0.7. Notice that the sum of an event and its complement adds up to 1.

Independence

Events are said to be *independent* if the occurrence of one event *fails* to predicts the likelihood of the other. In other words, events are independent when the outcome of the first event does not influence nor is influenced by the outcome of the second event.

The *law of independence* states that *A* and *B* are independent events *if and only if:*

$$Pr(\text{A and B}) = Pr(\text{A}) \times Pr(\text{B})$$

For instance, the probability of drawing King at random from a fair deck in 4 / 52. If we replace the first King and draw at random once again, the probability of drawing another King is still 4 / 52, and the probability of drawign two Kings in a row = (4/52)(4/52) = 0.005917. Therefore, the events are independent.

Furthermore, if dog and cat ownership were independent (which they are not), then the probability joint ownership would be $Pr(\text{A and B}) = Pr(\text{A}) \times Pr(\text{B}) = 0.3 \times 0.35 = 0.105$. However, since we know that $Pr(\text{A and B}) = 0.15$, we can say that events are *not* independent.

When events are not independent, they are said to be associated. In general:

! When $Pr(\text{A and B}) = Pr(\text{A}) \times Pr(\text{B})$, then events are independent and **no association** is said to exist.
! When $Pr(\text{A and B}) > Pr(\text{A}) \times Pr(\text{B})$, then events are not independent and a **positive association** is said to exist.
! If $Pr(\text{A and B}) < Pr(\text{A}) \times Pr(\text{B})$, then events are not independent and a **negative association** is said to exist.

Since the probability of ownership of a dog and cat is greater than expected under the assumption of independence, a positive association is said to exist. That is, if a person owns a dog, they are more likely on average to also own a cat.

Mutual Exclusivity

Events are **mutually exclusive** if $Pr(\text{A and B}) = 0$. In other words, events are mutually exclusive if the occurrence of one event prohibits the occurrence of the other event. For example, if A represented "male" and B represented "female," A and B would be mutually exclusive.

Addition

Recall that the logical "or" represent the union of events. This indicates the one or the other event occurs. For example, "A or B" represents "dog or cat ownership" in our illustrative example.

The *law of addition* states that:

$$Pr(\text{A or B}) = Pr(\text{A}) + Pr(\text{B}) \ ! \ Pr(\text{A and B})$$

In the dog/cat illustrative example, *Pr*(A) = 0.3, *Pr*(B) = 0.35, and *Pr*(A and B) = 0.15. Therefore, the probability of dog or cat ownership is *Pr*(*A* or *B*) = 0.3 + 0.35 ! 0.15 = 0.5; 50% of households would own either a dog or cat.

When events are mutually exclusive, the addition law states that the probability of the occurrence of two or more events is found by adding their separate probabilities. For example, we might ask what is the probability of drawing a King or a Queen from a fair deck of cards. The probability of drawing a King is 4 / 52. Also, the probability of drawing a Queen is 4 / 52. These events are mutually exclusive. Therefore, the probability of drawing a King or Queen = 4 /52 + 4 / 52 = 8 / 52, or about 15%.

Conditional Probability

Let *Pr*(B|A) represent the probability of *B* conditional on the occurrence of A. That is, the conditional probability of *B* given *A* is evident. By definition,

$$Pr(\text{A}|\text{B}) = Pr(\text{A and B}) \,/\, Pr(\text{A})$$

For example, the probability of a person owning a cat *given* that they own a dog can be calculated *Pr*(B|A) = *Pr*(A and B) / *Pr*(A) = 0.15 / 0.30 = 0.5 (but we already knew that).

From the definition of conditional probability, we see that:

$$\text{Pr(A and B)} = Pr(\text{A}) \times Pr(\text{A}|\text{B})$$

if *Pr*(A) is *not* equal to 0.

For our illustrative example, *Pr*(A and B) = *Pr*(A) *Pr*(B|A) = (.3)(.5) = 0.15.

Bayes' Theorum

Bayes' Law is an extension of the conditional probability which states:

$$Pr(\text{A}|\text{B}) = [Pr(\text{B}|\text{A})Pr(\text{A})] \,/\, [Pr(\text{B}|\text{A})Pr(\text{A}) + Pr(\text{B}|\text{not A})Pr(\text{not A})]$$

For example, the probability of dog ownership (A) given cat ownership (B) = [(.5)(.3)] / [(.5)(.3) + (.2857)(.7)] = 0.4286.

Notice that this probability could have been calculated directly with the definitional formula for conditional probability, as *Pr*(*A*|*B*) = *Pr*(*A* and *B*) / *Pr*(*B*) = (0.15) / (0.35) = 0.4286. However, Bayes' law allows us to calculate probabilities in a "backward direction," extending the rules of inference backward. This interpretation of Bayes' Law can take on profound philosophical implications when interpreting data, and has precipitated a schism (of sorts) among statisticians: there are practicing Baysians and non-Baysian statisticians (and neither the twain shall meet.) Let us, however, restrict our use of Bayes' theorem to its original purpose: as a way to evaluate conditional probabilities.

Summary of the "Dog/Cat" Illustrative Example Probabilities

Using the “dog/cat” illustrative example, we can now see how the various laws of probability fit together. Recall that event A represents dog ownership and event B represents cat ownership. We are given that *Pr*(*A*) = 0.3, *Pr*(*B*) = 0.35, and *Pr*(*A* and *B*) = 0.15. According to the laws of probability, it the follows:

For dog or cat ownership, *Pr*(*A* or *B*) = *Pr*(A) + *Pr*(B) ! *Pr*(A and B) = 0.3 + 0.35 ! 0.15 = 0.5.

For cat ownship given that we know the person owns a dog, *Pr*(*B*|*A*) = *Pr*(A and B) / *Pr*(A) = 0.15 / 0.3 = 0.5.

For dog ownership, given that we know the person owns a cat, *Pr*(*A*|B) = *Pr*(A and B) / *Pr*(*B*) = 0.15 / 0.35 = 0.4286.

For dog ownship, given that we know the person does not own a cat, *Pr*(A|not B) = *Pr*(A not B) / *Pr*(not B) = 0.15 / 0.65 = 0.2308.

For cat ownership, given that we know that the person does not own a dog, *Pr*(B|not A) = *Pr*(B not A) / P (not A) = 0.2 / 0.7 = 0.2857.

Risk

Epidemiologists define the **risk** as probability of occurrence over a specified time interval. For example, the risk of pulmonary thromboembolism for a person over a year might be 1 in 10,000 (0.0001). In practice, we use the *incidence* of an event as an *estimate* of its *average risk*.

Often, risks are compared in the form of a *risk ratio* ("*relative risk*"). For example, if the risk of pulmonary thromboembolism in a women using oral contraceptive users is 0.0003 per year and the risk of pulmonary thromboembolism in the same women not using oral contraceptives is 0.0001, we might then say that the relative risk of pulmonary thromboembolism associated with oral contraceptive use = 0.0003 / 0.0001 = 3. This indicates that oral contraceptive use triples the risk of pulmonary thromboembolism. Notice that the relative risk is an expression of two conditional probabilities. If we let D represent pulmonary thromboembolism and E represent oral contraceptive use, then the risk of thromboembolism given oral contraceptive use = $Pr(D|E)$; the risk of thromboembolism given non-use is $Pr(D|\text{not } E)$. Given this notation,

$$RR = \frac{Pr(D|E)}{Pr(D|\text{not } E)}$$

Also notice that **when *D* and *E* are independent,** the relative risk is equal to 1, suggests no association between *A* and *B*.

$$RR = \frac{Pr(D|E)}{Pr(D|\text{not } E)} = \frac{Pr(D \text{ and } E)/Pr(E)}{Pr(D \text{ and not } E)/Pr(\text{not } E)} = \frac{Pr(D)Pr(E)/Pr(E)}{Pr(D)Pr(\text{not } E)/Pr(\text{not } E)} = \frac{Pr(D)}{Pr(D)} = 1$$

Empirically, we *estimate* the *relative risk* of an outcome by calculating incidence proportions in two groups and then comparing them. Suppose, for example, that frequent alcohol consumption is denoted as D and male gender is denoted as E. Further suppose the following data are observed:

	Male (E)	Not male (not E)	Total
Alcohol Abuse Present (D)	100	50	150
Alc. Abuse Absent (not D)	700	800	1500
Total	800	850	1650

Although the actual risk of events is unknown, we can estimate $Pr(\text{D}|\text{E}) = 100 / 800 = 0.1250$ and $Pr(\text{D}|\text{not E}) = 50 / 850 = 0.0588$. Thus, and estimate of the $RR = 0.1250 / 0.0588 = 2.13$. This suggests an approximate doubling of frequent alcohol use in men compared with women. Thus, *D* and *E* are *not* independent, so a positive association is said to exist between Alcohol Abuse and Gender.

A Different Way to Conceptualize Probability

Probability can be conceptualized in a different way. Consider the outcome of a flip of a coin (heads or tails). This is usually considered a chance event, and the usual laws of chance (probability) apply. However, in classical mechanics, the outcome of the coin flip is actually determined completely by the application of physical laws and a sufficient description of the starting conditions of the toss. We could, in theory, measure the starting position of the coin, the angle and velocity of the toss, the rate of the coin's rotation, and so on, and predict the final position of the coin; all without the benefit of the laws of probability. In fact, the outcome of the toss is determine by physical forces. However, it is often impossible to measure all these complex factors. Therefore, "randomness" can be conceived of as a way to state likelihoods when addressing imperfect or unobservable knowledge of a complex system with multiple causes.

To put this in epidemiologic terms, consider the explanation for why an individual gets lung cancer. One hundred years ago, when little was known about the causes of lung cancer, a person might have said that it was a matter of chance. Today, we say that risk depends on how much the individual smokes and how much radon the individual has been exposed to. One might then ask what determines whether an individual who has smoked a specific amount and has a specified amount of exposure to all other known risk factors will get lung cancer. Today's answer might well be that it is a matter of chance. We can explain much more of the variability in lung cancer occurrence nowadays than we formerly could by taking into account factors known to cause it, but at the limits of our knowledge, we ascribe the remaining variability to what we call chance. In this way, "chance" is seen as a catchall term for our ignorance about causal explanations.

The Poisson Distribution

The Poisson distribution describes the probability of rare, discrete events that occur randomly in time and space. This distribution is especially useful when dealing with rates and concentrations when the total number of possible events is unknown. For example, we may apply the Poisson distribution to the number of cancer cases in a geographic region in a year.

The Poisson distribution is defined by a single parameter, μ, representing both the expected value *and* variance of the distribution. The formula for calculating Poisson probabilities is:

$$Pr(X = i) = e^{-\mu}\mu^{i}/i!$$

where *e* is the universal constant (2.71828...), μ is the expected value of *X*, and *i*! is "*i* factorial". We may represent a particular Poisson random variable with the notation *X*~p(μ). For example, a Poisson random variable with μ = 1 is denoted *X*~p(1).

Let us consider a Poisson random variable with an expected value of 1. If this is the case, the probability of observing no (0) cases is: $Pr(X = 0) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{0}/0! = 0.3679$.

The probability of observing 1 case is $Pr(X = 1) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{1}/1! = 0.3679$. Continuing the *pdf*:
$Pr(X = 2) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{2}/2! = 0.1839$
$Pr(X = 3) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{3}/3! = 0.0613$
$Pr(X = 4) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{4}/4! = 0.0153$
$Pr(X = 5) = e^{\mu}\mu^{x}/x! = e^{1}\mu^{5}/5! = 0.0031$

To complete the *pdf*, we calculate *Pr*(*X* $ 6), noting that Pr(*X* $ 6) = 1 ! *Pr*(X # 5) = 1 ! 0.9994 = 0.0006.

> *Pr*(*X* # 5) = *Pr*(*X* = 0) + *Pr*(*X* = 1) + *Pr*(*X* = 2) + *Pr*(*X* = 3) + *Pr*(*X* = 4) = 0.3679 + 0.3679 + 0.1839 + 0.0613 + 0.0153 + 0.0031 = 0.9994.

The Poisson distribution can be used to help infer whether a given number of cases is likely under am assumed set of circumstance. For example, the observation of 6 cases under the above *pmf* would indeed be very unusual, since *Pr*(*X* $ 6) = 0.0006. This would give us reason to question the assumption that μ is actually equal to 1, and is perhaps greater than expected.