Fundamentals of Clinical Research for Radiologists

Lawrence Joseph[1,2]
Caroline Reinhold[3]

Introduction to Probability Theory and Sampling Distributions

Received July 9, 2002; accepted after revision August 1, 2002.

Series editors: Craig A. Beam, C. Craig Blackmore, Stephen Karlik, and Caroline Reinhold.

This is the tenth in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the American Journal of Roentgenology. *The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous clinical research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).*

Project coordinator: Bruce J. Hillman, Chair, ACR Commission on Reseach and Technology Assessment

[1]Department of Medicine, Division of Clinical Epidemiology, Montreal General Hospital, 1650 Cedar Ave., Montreal, Quebec, H3G 1A4, Canada. Address correspondence to L. Joseph.

[2]Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Ave. W., Montreal, Quebec, H3A 1A2, Canada.

[3]Department of Diagnostic Radiology, Montreal General Hospital, McGill University Health Centre, 1650 Cedar Ave., Montreal, Quebec, H3G 1A4, Canada.

AJR 2003;180:917–923

0361–803X/03/1804–917

Statistical inference allows one to draw conclusions about the characteristics of a population on the basis of data collected from a sample of subjects from that population. Almost all the statistical inferences typically seen in the medical literature are based on probability models that connect summary statistics calculated using the observed data to estimates of parameter values in the population. This article will cover the basic principles behind probability theory and examine a few simple probability models that are commonly used, including the binomial, normal, and Poisson distributions. We will then see how sampling distributions are used as the basis for statistical inference and how they are related to simple probability models. Thus, this article forms the foundation for future articles in the series that will present the details of statistical inference in particular clinical situations.

Making medical decisions on the basis of findings from various radiologic diagnostic tools is an everyday occurrence in clinical practice. In radiologic research, one often needs to draw conclusions about the relative performance of one diagnostic tool compared with another for the detection of a given condition of interest. Both of these tasks depend, in large part, on probability theory and its applications. In diagnosis, we are interested in calculating the probability that the condition of interest is present on the basis of results of a radiologic test. This probability depends on how sensitive and specific that test is in diagnosing the condition and on the background rate of the condition in the population.

This calculation largely depends on a result from probability called Bayes' theorem. Similarly, all statistical inferences, whether comparisons of two proportions representing diagnostic accuracies from two instruments or inferences from a more complex model, are based on probabilistic reasoning. Therefore, a thorough understanding of the meaning and proper interpretation of statistical inferences, crucial to daily decision making in a radiology department, depends on an understanding of probability and probability models.

This article is composed of three main parts. We begin with an introduction to probability, including the definitions of probability, the different schools of thought about the interpretation of probabilities, and some simple examples. We continue by defining conditional probabilities and present Bayes' theorem, which is used to manipulate conditional probabilities. The most common simple probability models, including the binomial, normal, and Poisson distributions, are presented next, along with the types of situations in which we would be most likely to use them. Finally, sampling strategies are examined. Armed with these basics of probability and sampling, we conclude with a discussion of how the outcome of interest defines the model parameter on which to focus inferences and how the sampling distribution of the estimator of that parameter enables valid inferences from the data collected in the sample about the population at large.

Probability

Definitions of Probability

Last's *Dictionary of Epidemiology* [1] presents two main definitions for probability. The

first definition, which represents the view of the frequentist school of statistics, defines the probability of an event as the number of times the event occurs divided by the number of trials in which it could have occurred, *n*, as *n* approaches infinity. For example, the probability that a coin will come up heads is 0.5 because, assuming the coin is fair, as the number of trials (flips of the coin) gets larger and larger, the observed proportion will be, on average, closer and closer to 0.5. Similarly, the probability that an intervention for back pain is successful would be defined as the number of times it is observed to be successful in a large (theoretically infinite) number of trials in patients with back pain.

Although this definition has a certain logic, it has some problems. For example, what is the probability that team A will beat team B in their game tonight? Because this is a unique event that will not happen an infinite number of times, the definition cannot be applied. Nevertheless, we often hear statements such as "There is a 60% chance that team A will win tonight." Similarly, suppose that a new intervention for back pain has just been developed, and a radiologist is debating whether to apply it to his or her next patient. Surely the probability of success of the new intervention compared with the probability of success of the standard procedure for back pain will play a large role in the decision. However, no trials (and certainly not an infinite number of trials) as yet exist on which to define the probability. Although we can conceptualize an infinite number of trials that may occur in the future, this projection does not help in defining a probability for today's decision. Clearly, this definition is limited, not only because some events can happen only once, but also because one cannot observe an infinite number of like events.

The second definition, often referred to as the Bayesian school, defines the probability of any event occurring as the personal degree of belief that the event will occur. Therefore, if I personally believe that there is a 70% chance that team A will win tonight's game, then that is my probability for this event. In coin tossing, a Bayesian may assert, on the basis of the physics of the problem and perhaps a number of test flips, that the probability of a coin flip coming up heads should be close to 0.5. Similarly, on the basis of an assessment that may include both previously available data and subjective beliefs about the new technique, a radiologist may assert that the probability that a procedure will be successful is 85%.

The obvious objection to Bayesian probability statements is that they are subjective, and thus different radiologists may state different probabilities for the success rate of the new technique. In general, no single "correct" probability statement may be made about any event, because such statements reflect personal subjective beliefs. Supporters of the Bayesian viewpoint counter that the frequentist definition of probability is difficult to apply in practice and does not pertain to many important situations. Furthermore, the possible lack of agreement as to the correct probability for any given event can be viewed as an advantage, because it will correctly mirror the range of beliefs that may exist about any event that does not have a large amount of data from which to accurately estimate its probability. Hence, having a range of probabilities depending on the personal beliefs of a community of clinicians is a useful reflection of reality. As more data accumulate, Bayesian and frequentists probabilities tend to agree, each essentially converging to the mean of the data. When this occurs, similar inferences will be reached from either viewpoint.

Discussion of these two ways of defining probability may seem to be of little relevance to radiologists but, later in this series, it will become apparent that it has direct implications for the type of statistical analysis to be performed. Different definitions of probability lead to different schools of statistical inference and, most importantly, often to different conclusions based on the same set of data. Any given statistical problem can be approached from either a frequentist or a Bayesian viewpoint, and the choice often depends on the experience of the user more than it does on one or the other approach being more appropriate for a given situation. In general, Bayesian analyses are more informative and allow one to place results into the context of previous results in the area [2], whereas frequentist methods are often easier to carry out, especially with currently available commercial statistical packages. Although most analyses in medical journals currently follow the frequentist definition, the Bayesian school is increasingly present, and it will be important for readers of medical journals to understand both.

The lack of a single definition of probability may be disconcerting, but it is reassuring to know that whichever definition one chooses, the basic rules of probability are the same.

Rules of Probability

Four basic rules of probability exist. These rules are usually expressed more rigorously than is necessary for the purposes of this article, through the use of set theory and probability notation.

The first rule states that, by convention, all probabilities are numbers between 0 and 1. A probability of 0 indicates an impossible event, and a probability of 1 indicates an event certain to happen. Most events of interest have probabilities that fall between these extremes.

The second rule is that events are termed "disjoint" if they have no outcomes in common. For example, the event of a patient having cancer is disjoint from the event of the same patient not having cancer, because both cannot happen simultaneously. On the other hand, the event of cancer is not disjoint from the event that the patient has cancer with metastases because in both cases the outcome of cancer is present. If events are disjoint, then the probability that one or the other of these events occurs is given by the sum of the individual probabilities of these events. For example, in looking at an MR image of the liver, if the probability that the diagnosis is a hepatoma is 0.5 (meaning 50%) and the probability of a metastases is 0.3, then the probability of either hepatoma or metastases must be 0.8, or 80%.

The third rule is expressed as follows: If one could list the set of all possible disjoint events of an experiment, then the probability of one of these events happening is 1. For example, if a patient is diagnosed according to a 5-point scale in which 1 is defined as no disease; 2, as probably no disease; 3, as uncertain disease status; 4, as probably diseased; and 5, as definitely diseased, then the probability that one of these states is chosen is 1.

The fourth rule states that, if two events are independent (i.e., knowing the outcome of one provides no information concerning the likelihood that the other will occur), then the probability that both events will occur is given by the product of their individual probabilities. Thus, if the probability that findings on an MR image will result in a diagnosis of a malignant tumor is 0.1, and the probability that it will rain today is 0.3 (an independent event, presumably, from the results of the MR imaging), then the probability of a malignant tumor and rain today is $0.1 \times 0.3 = 0.03$, or 3%.

In summary, probabilities for events always follow these four rules, which are compatible with common sense. Such probability calculations can be useful clinically, for example, in deriving the probability of a certain diagnosis given one or more diagnostic test results. Many probability calculations used in clinical research involve conditional probabilities. These are explained next.

Conditional Probabilities and Bayes' Theorem

What is the probability that a given patient has endometrial cancer? Clearly, this depends on a number of factors, including age, the presence or absence of postmenopausal bleeding, and others. In addition, our assessment of this probability may drastically change between the time of the patient's initial clinic visit and the point at which diagnostic test results become known. Thus, the probability of endometrial cancer is conditional on other factors and is not a single constant number by itself. Such probabilities are known as conditional probabilities. Notationally, if unconditional probabilities can be denoted by *Pr*(cancer), then conditional probabilities can be denoted by *Pr*(cancer | diagnostic test is positive), read as "the probability of cancer given or conditional on a positive diagnostic test result," and, similarly, *Pr*(cancer | diagnostic test is negative), read as "the probability of cancer given a negative diagnostic test result." These probabilities are highly relevant to radiologic practice and clinical research in radiology.

Because they are a form of probability, conditional probabilities must follow all rules as outlined in the previous section. In addition, however, there is an important result that links conditional probabilities to unconditional probability statements. In general, if we denote one event by *A*, and a second event by *B*, then we can write

$$Pr(A \mid B) = \frac{Pr(A \text{ and } B)}{Pr(B).}$$

In words, the probability that event *A* occurs, given that we already know that event *B* has occurred, denoted by *Pr* (*A* | *B*), is given by dividing the unconditional probability that these two events occur together by the unconditional probability that *B* occurs. Of course, this formula can be algebraically manipulated, so that it must also be true that

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

For example, suppose that in a clinic dedicated to evaluating patients with postmenopausal bleeding, endovaginal sonography is often used for the detection of endometrial cancer. Assume that the overall probability of a patient in the clinic having endometrial cancer is 10%. This probability is unconditional, that is, it is calculated from the overall prevalence in the clinic; before any test results are known. Furthermore, suppose that the sensitivity of endovaginal sonography for diagnosing endometrial cancer is 90%. If we let *A* represent the event that the patient has a positive endovaginal sonography, and let *B* represent the probability of endometrial cancer in this patient population, then we can summarize the above information as $Pr(B) = 0.1$ and $Pr(A \mid B) = 0.9$. By using the formula described, we can deduce that the probability that a patient in this clinic has both endometrial cancer and positive results on endovaginal sonography is $0.1 \times 0.9 = 0.09$ or 9%.

In typical clinical situations, we may know the background rate of the disease in question in the population referred to a particular clinic (which may differ from clinic to clinic), and we may have some idea of the sensitivity and specificity of the test. Notice that in the terms used, sensitivity and specificity may be considered conditional probabilities because they provide the probability of testing positive given a subject who truly has the condition of interest (i.e., *Pr* [*A* | *B*], which is the sensitivity), and the probability of not testing positive given the absence of the condition of interest (i.e., the specificity, *Pr* [not *A* | not *B*]). What should a clinician conclude if a patient walks through the door with a "positive" test result in hand? In this case, one would like to know the probability of the patient's being truly positive for the condition, given that he or she has just had a test with positive findings. Of course, if the diagnostic test is a perfect gold standard, one can simply look at the test result and be confident of the conclusion.

However, most tests do not have perfect sensitivity and specificity, and thus a probability calculation is needed to find the probability of a true-positive, given the positive test result. In our notation, we know the prevalence of the condition in our population, *Pr*(*B*), and we know the sensitivity and specificity of our test, given by *Pr* (*A* | *B*) and *Pr* (not *A* | not *B*), but we want to know *Pr*(*B* | *A*), which is opposite in terms of what is being conditioned on. How does one reverse the conditioning argument, in effect making statements about *Pr*(*B* | *A*) when we only know *Pr* (*A* | *B*)? The answer is to use a general result from probability theory, called Bayes' theorem, which states

$$Pr(B \mid A) = \frac{Pr(B) \times Pr(A \mid B)}{Pr(B) \times Pr(A \mid B) + Pr(\text{not } B) \times Pr(A \mid \text{not } B).}$$

Suppose that the background rate of endometrial cancer seen in patients referred to a particular radiology clinic is 10% and that a diagnostic test is applied that has *Pr*(*A* | *B*) = 90% sensitivity and *Pr*(not *A* | not *B*) = 80% specificity. What is the probability that a patient with positive test results in fact has endometrial cancer? According to Bayes' theorem, we calculate

$$Pr(B \mid A) = \frac{Pr(B) \times Pr(A \mid B)}{Pr(B) \times Pr(A \mid B) + Pr(\text{not } B) \times Pr(A \mid \text{not B})}$$

$$= \frac{0.1 \times 0.9}{0.1 \times 0.9 + 0.9 \times 0.2}$$

$$= 0.33$$

or about 33%. In this case, even when a patient has a positive test result, the chances that the disease is present are less than 50%.

Similarly, what is the probability that a subject testing negative has endometrial cancer? Again using Bayes' theorem,

$$Pr(B \mid \text{not } A) = \frac{Pr(B) \times Pr(\text{not } A \mid B)}{Pr(B) \times Pr(\text{not } A \mid B) + Pr(\text{not } B) \times Pr(\text{not } A \mid \text{not } B)}$$

$$= \frac{0.1 \times 0.1}{0.1 \times 0.1 + 0.9 \times 0.8}$$

$$= 0.013.$$

Thus, starting from a background rate of 10% (pretest probability), the probability of cancer rises to 33% after a positive diagnosis and falls to approximately 1% after a negative test (posttest probabilities). Thus, Bayes' theorem allows us to update our probabilities after learning the test result, and it is thus of great usefulness to practicing radiologists. The next module in this series covers Bayes' theorem and diagnostic tests in more detail.

Probability Models

Rather than working out all problems involving probabilities by first principles using the basic probability rules as we have discussed, it is possible to use short cuts that have been devised for common situations, leading to probability functions and probability densities. Here we review three of the most common distributions: the binomial, the normal, and the Poisson. Which distribution to use depends on many situation-specific factors, but we provide some general guidelines for the appropriate use of each.

The Binomial Distribution

One of the most commonly used probability functions is the binomial. The binomial distribution allows one to calculate the probability of obtaining a given number of "successes" in a given number of trials, wherein the probability of a success on each trial is assumed to be p. In general, the formula for the binomial probability function is

$$Pr(\text{x successes in } n \text{ trials}) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},$$

where $n!$ is read "n factorial" and is shorthand for

$$n\times(n-1)\times(n-2)\times(n-3)\times\ldots\times3\times2\times1.$$

For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, and so on. By convention, $0! = 1$. Suppose we wish to calculate the probability of $x = 8$ successful angioplasty procedures in $n = 10$ patients with unilateral renal artery stenosis, wherein the probability of a successful angioplasty each time is 70%. From the binomial formula, we can calculate

$$\frac{10!}{8!2!}0.7^8(1-0.7)^2 = 0.2335,$$

so that there is slightly less than a one-in-four chance of getting eight successful angioplasty procedures in 10 trials. Of course, these days such calculations are usually done by computer, but seeing the formula and calculating a probability using it at least once helps to avoid that "black box" feeling one can often get when using a computer and adds to the understanding of the basic principles behind statistical inference. Similarly, the probability of getting eight or more (that is, eight or nine or 10) successful angioplasty procedures is found by adding three probabilities of the type shown, using the second probability rule because these events are disjoint. As an exercise, one can check that this probability is 0.3829. See Figure 1 for a look at all probabilities for this problem, in which x varies from zero to 10 successes for $n = 10$ and $p = 0.7$.

The binomial distribution has a theoretic mean of $n \times p$, which is a nice intuitive result. For example, if one performs $n = 100$ trials, and on each trial the probability of success is $p = 0.4$ or 40%, then one would intuitively expect $100 \times 0.4 = 40$ successes. The variance, σ^2, of a binomial distribution is $n \times p \times (1 - p)$, so that in the example just given it would be $100 \times 0.4 \times 0.6 = 24$. Thus, the SD is

$$\sqrt{\sigma^2} = \sigma = \sqrt{24} = 4.90,$$

roughly meaning that although on average one expects approximately 40 successes, one also expects each result to deviate from 40 by an average of approximately five successes.

The binomial distribution can be used any time one has a series of independent trials (different patients in any trial can usually be considered as independent) wherein the probability of success remains the same for each patient. For example, suppose that one has a series of 100 patients, all with known endometrial cancer. If each patient is asked to undergo MR imaging, for example, and if the true sensitivity of this test is 80%, what is the probability that 80 of them will in fact test positive? By plugging $p = 0.8$, $n = 100$, and $x = 80$ into the binomial probability formula as discussed, one finds that this probability is 0.0993, or about 10%. (One would probably want to do this calculation on a computer because 100!, for example, would be a tedious calculation.)

Normal Distribution

Perhaps the most common distribution used in statistical practice is the normal distribution, the familiar bell-shaped curve, as seen in Figure 2. Many clinical measurements follow normal or approximately normal distributions (e.g., tumor sizes). Technically, the curve is traced out by the normal density function

$$\frac{1}{\sqrt{2\pi\sigma}}\ \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\},$$

where "exp" denotes the exponential function to the base $e = 2.71828$. The Greek letter μ is the mean of the normal distribution set to zero in the SD curve of Figure 2, and the SD is σ, set to 1 in the standard normal curve. Although Figure 2 presents the standard version of the normal curve ($\mu = 0$, $\sigma^2 = \sigma = 1$), more generally, the mean μ can be any real number and the SD can be any number greater than zero. Changing the mean shifts the curve depicted in Figure 2 to the left or right so that it remains centered at the mean, whereas changing the SD stretches or shrinks the curve around the mean, all while keeping its bell shape. Note that the mean (usual arithmetic average), median (middle value, i.e., point at which 50% of the area under the curve lies above and below), and mode (most likely value, i.e., highest point on the curve) of a normal distribution are always the same and equal to μ. Approximately 95% of the area under the curve falls

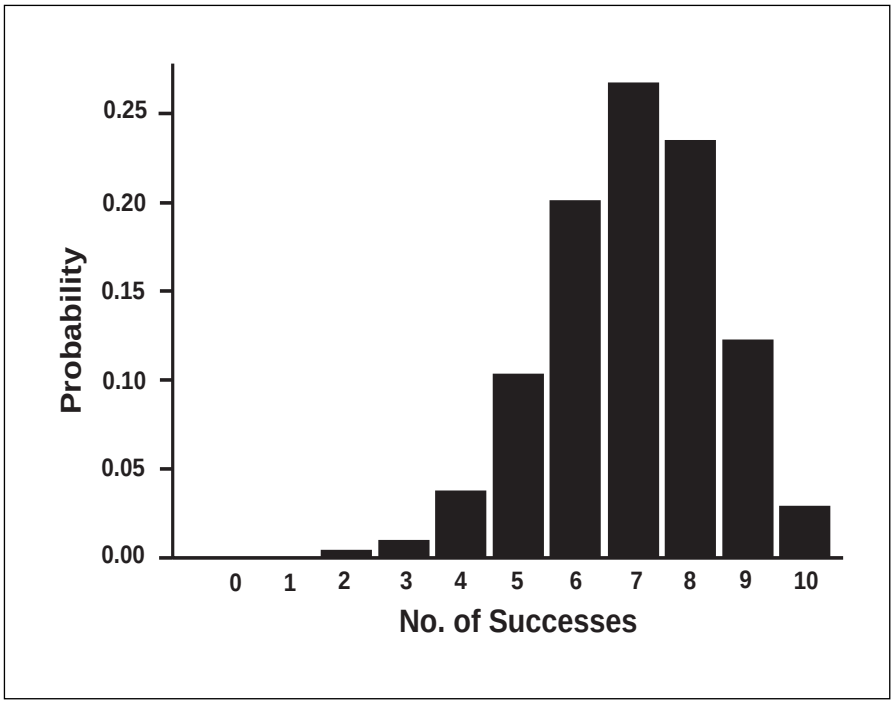

Fig. 1.—Graph shows binomial distribution with sample size of 10 and probability of success $p = 0.7$.

within 2 SDs on either side of the mean, and approximately 68% of the area falls within 1 SD of the mean.

The normal density function has been used to represent the distribution of many measures in medicine. For example, tumor size, biparietal diameter, or bone mineral density in a given population may be said to follow a normal distribution with a given mean and SD. It is highly unlikely that any of these or other quantities exactly follow a normal distribution. For instance, none of these quantities can have negative numbers, whereas the range of the normal distribution always includes all negative (and all positive) numbers. Nevertheless, for appropriately chosen mean and SD, the probability of out-of-range numbers will be vanishingly small, so that this may be of little concern in practice. We may say, for example, that tumor size in a given population follows a normal distribution with a mean of 20 mm and an SD of 10 mm, so that the probability of a value less than zero is only approximately 2.5%. In the words of statistician George Box [3], "All models are wrong, but some are useful."

To calculate probabilities associated with the normal distribution, one must find the area under the normal curve. Because doing so is mathematically difficult, normal tables or a computer program are usually used. For example, the area under the standard normal curve between –1 and 2 is 0.8186, as calculated via normal tables or via a computer package for statistics.

The normal distribution is central to statistical inference for an additional reason. Consider taking a random sample of 500 patients visiting their family physicians for periodic health examinations. If the blood pressure of each patient were recorded and an average were taken, one could use this value as an estimate of the average in the population of all patients who might visit their family physicians for routine checkups. However, if the experiment were repeated, it would be unexpected for the second average of 500 patients to be identical to the first average, although one could expect it to be close.

How these averages vary from one sample to another is given by the central limit theorem, which in its simplest form is explained as follows. Suppose that a population characteristic has true (but possibly unknown) mean μ and standard deviation σ. The distribution of the sample average, $\bar{x}$, based on a sample of size n, approaches a normal distribution as the sample size grows large, with mean μ and SD $\sigma / \sqrt{n}$. As will be explained in future articles, the sample average, $\bar{x}$, is used to estimate the true (but unknown) population mean μ. The SD about a sample mean, $\sigma / \sqrt{n}$, is often called the standard error (SE).

This useful theorem has two immediate consequences. First, it accounts for the popularity of the normal distribution in statistical practice. Even if an underlying distribution in a population is nonnormal (e.g., if it is skewed or binomial), the distribution of the sample average from this population becomes close to normal if the sample size is large enough. Thus, statistical inferences can often be based on the normal distribution, even if the underlying population distribution is nonnormal. Second, the result connects the sample mean to the population mean, forming the basis for much of the statistical inference. In particular, notice that as the sample size n increases, the SD (SE) $\sigma / \sqrt{n}$ of the sample mean around the true mean decreases so that on average the sample mean $\bar{x}$ gets closer and closer to μ. We return to this important point later, but first look at our last distribution, the Poisson.

Poisson Distribution

Suppose that we would like to calculate probabilities relating to numbers of cancers over a given period of time in a given population. In principle, we can consider using a binomial distribution because we are talking about numbers of events in a given number of trials. However, the numbers of events may be enormous (number of persons in the population times the num-

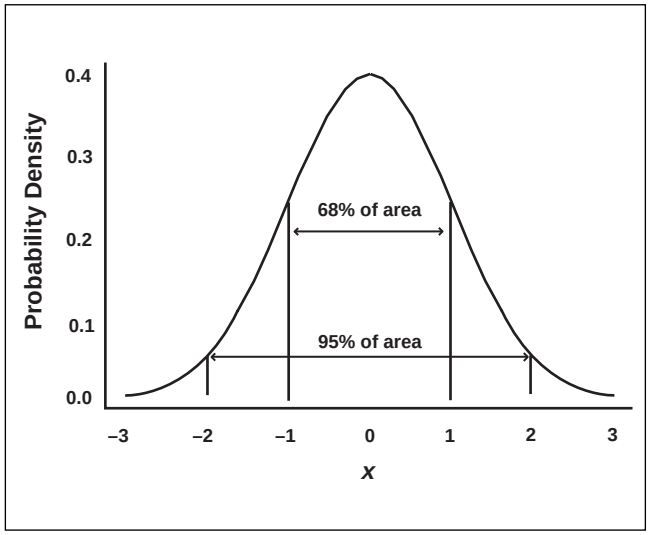

Fig. 2.—Graph shows the standard normal distribution with mean μ = 0 and SD σ = 1. Approximately 95% of area under curve falls within 2 SDs on either side of mean, and about 68% of area falls within 1 SD from mean.

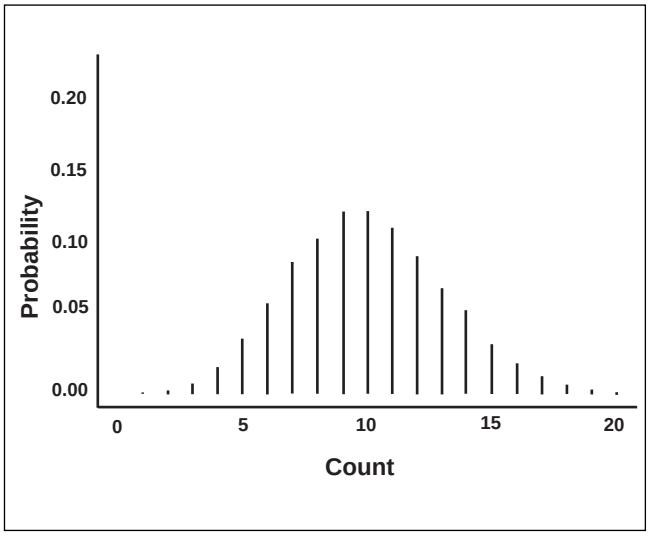

Fig. 3.—Chart shows the Poisson distribution with μ (mean = 10).

ber of time periods). Furthermore, we may not even be certain of the denominator but may have some idea of the rate (e.g., per year) of cancers in this population from previous data. In such cases in which we have counts of events through time rather than counts of successes in a given number of trials, we can consider using the Poisson distribution. More precisely, we make the following assumptions:

First, we assume that the probability of an event (e.g., a cancer) is proportional to the time of observation. We can notate this as *Pr* (cancer occurs in time t) = $\lambda \times t$, wherein λ is the rate parameter, indicating the event rate in units of events per time. Second, we assume that the time t is small enough that two events cannot occur in time t. For cancer in a population, t may be, for example, 1 min. The event rate λ is assumed to be constant through time (homogeneous Poisson process). Finally, we assume that events (cancers) occur independently.

If all of these assumptions are true, then we can derive the distribution of the number of counts in any given period of time. Let $\mu = \lambda \times t$ be the rate multiplied by time, which is the Poisson mean number of events in time t. Then the Poisson distribution is given by

$$Pr(x \text{ events occur in time } t) = \frac{e^{-\mu}\mu^{x}}{x!},$$

where $e = 2.71828\ldots$, and x denotes factorial of x (the same as in the binomial distribution). Both the mean and the variance of the Poisson distribution are equal to μ. The graph of the Poisson distribution for $\mu = 10$ is given in Figure 3.

As an example of the use of the Poisson distribution, suppose that the incidence of a certain type of cancer in a given region is 250 cases per year. What is the probability that there will be exactly 135 cancer cases in the next 6 months? Let t =1 year, then $\mu = 250$ cancers per year. We are interested, however, in $t = 0.5$, which means that $\mu = 125$ cancers per 6-month period. Using the Poisson distribution, we can calculate

$$Pr(135 \text{ cancers} \mid \mu = 125) = \frac{e^{-125}125^{135}}{135!} = 0.0232.$$

Therefore, approximately 2.3% of a chance exists of observing 135 cancers in the next 6 months.

Summary

The binomial distribution is used for yes/no or success/fail dichotomous variables, the normal distribution is often used for probabilities concerning continuous variables, and the Poisson distribution is used for outcomes arising from counts. These three distributions, of course, are by no means the only ones available, but they are among the most commonly used in practice. Deciding whether they are appropriate in any given situation requires careful consideration of many factors and verification of the assumptions behind each distribution and its use.

This ends our brief tour of the world of probability and probability distributions. Armed with these basics, we are now ready to consider some simple statistical inferences.

Sampling Distributions

So far, we have seen the definitions of probability, the rules probabilities must follow, and three probability distributions. These ideas form the basis for statistical inferences, but how? The key is sampling distributions.

First, we must distinguish sampling distributions from probability distributions and population distributions, which can be explained through an example: Suppose we would like to measure the average tumor size on detection at MR imaging for a certain type of cancer. If we were able to collect the tumor size for all patients with this disease (i.e., a complete census) and create a histogram of these values, then these data would represent the population distribution. The mean of this distribution would represent the true average tumor size in this population.

It is rare, if not impossible, for anyone to perform a complete census, however. One will usually have the opportunity to observe only a subset of the subjects in the target population (i.e., a sample). Suppose that we are able to take a random sample of subjects from this population, of, for example, $n = 100$ patients. In each case, we observe the tumor size and record the average value. Suppose this average value is $\bar{x}$ = 20 mm, with a SD of σ = 10 mm. We can thus conclude that 20 mm, the average value in our sample, is a reasonable (unbiased) point estimate of the average tumor value in our population, but how accurate is it? How does this accuracy vary if we change the sample size to only 10 patients? What about if we increase it to 1000 patients?

The answer to these questions lies in the sampling distribution of the estimator, $\bar{x}$. First of all, what is the sampling distribution of $\bar{x}$? Suppose we were to take a second random sample of size 100 and record its mean. It would not likely be exactly 20 mm but perhaps be close to that value, for example, 18 mm. If we repeated this process for a third sample, we might get a mean of 21 mm, and so on. Now imagine the thought experiment in which we would repeat this process an infinite number of times and draw the histogram of these means of 100 subjects. The resulting histogram would represent the sampling distribution of $\bar{x}$ for this problem.

According to the central limit theorem, the sampling distribution of $\bar{x}$ is a normal distribution, with mean μ representing the true but unknown mean tumor size (available only if a complete census is taken), and with an SE $\sigma / \sqrt{n}$. Therefore, the SE in our example is $10 / \sqrt{100} = 1$ mm. So the sampling distribution of $\bar{x}$ is normal, with unknown mean μ, and SE of 1. Although we do not know the mean of the sampling distribution, we do know, from our facts about the normal distribution, that 95% of all $\bar{x}$'s sampled in this experiment will be within $\pm 2 \times 1 = 2$ SEs from μ. Thus, although μ remains unknown, we do expect it to be near $\bar{x}$ in this sense. Chances are very good that $\bar{x}$ will be within 2 mm of μ, allowing statements called confidence intervals about μ that we will examine more closely in subsequent articles in this series. If we observed only 10 tumors rather than 100, our SE would have been $10 / \sqrt{10} = 3.2$ mm, leading to less accuracy in estimating μ, whereas a sample size of 1000 would lead to an SE of 0.32, leading to increased accuracy compared with a size of 100.

To summarize, population distributions represent the spread of values of the variable of interest across individuals in the target population, whereas sampling distributions show how the estimate of the population mean varies from one sample to the next if the experiment were to be repeated and the mean calculated each time. The sampling distribution connects the estimator, here $\bar{x}$, to the parameter of interest, here μ, the mean tumor size in the population. Larger sample sizes lead to more accurate estimation.

Similar inferences can be made from observations that are dichotomous using the binomial distribution or for count data using the Poisson distribution. Again, these topics are relegated to a future article in this series.

Notice that we had to make various assumptions in the previous discussion—for example, that the distribution of tumor sizes in the population is approximately normal and, most importantly, that the subjects are representative of the population to whom we wish to make infer-

ences. The easiest way to ensure representativeness is through random selection, but this may not be possible in some situations for practical reasons. For true random selection to occur, one must have a list of all members of the population and select subjects to form the study sample by random number generation or another random process. Lists of all members of the target population are rare, however, so that different mechanisms of subject selection are often necessary. Case series, or consecutive patients in a clinic, may or may not be representative, depending on the particularities of the selection process. Similarly, convenience samples—taking the subjects most easily available—are often not completely representative, because the very fact that subjects are easily available often tends to make them younger, less sick, and living near the clinic.

Because many outcomes of interest may differ between, for example, young and old or urban and rural patients, convenience samples and often case series are always suspect in terms of selection bias. In other words, although a tumor size of 20 mm may in fact be the average in your sample, this estimate is biased if patients with smaller or larger tumors are systematically left out. For example, subjects with preclinical symptoms may not visit your clinic, even if their tumors might have been detectable on MR imaging, resulting in 20 mm being an overestimate of the true average tumor size detectable on MR imaging in the clinic. Similarly, if patients with advanced disease do not visit the clinic because their tumors were clinically detected by other means, 20 mm may in fact be an underestimate of the true average. Selection bias should always be kept in mind when reading the medical literature.

Conclusion

This brief tour of probability, distributions, and the roots of statistical inferences barely scratches the surface. Many of these ideas will be amplified in future articles of this series. For the impatient, or those who want more detailed explanations of the concepts presented here, countless books explain basic statistical concepts—dozens with a focus on biostatistics. Among them are the works of Armitage and Berry [4], Colton [5], Rosenberg et al. [6], and Rosner [7].

References

1. Last J. *A dictionary of epidemiology,* 2nd ed. New York: Oxford University Press, **1988**:xx
2. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *JAMA* **1995**;273:871–875
3. Box G. *Statistics for experimenters: an introduction to design, data analysis, and model building.* New York: Wiley, **1978**
4. Armitage P, Berry G. *Statistical methods in medical research,* 3rd ed. Oxford: Blackwell Scientific Publications, **1994**
5. Colton T. *Statistics in medicine.* Boston: Little, Brown, **1974**
6. Rosenberg L, Joseph L, Barkun A. *Surgical arithmetic: epidemiological, statistical and outcome-based approach to surgical practice.* Austin, TX: Landes Biosciences, **2000**
7. Rosner B. *Fundamentals of biostatistics.* Belmont, CA: Duxbury, **1994**:105

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003