Using Probability Distributions to Make Decisions

Welcome JVT readers to “Statistical Viewpoint.” This column presents statistical principles useful to practitioners in compliance and validation. We hope it will become a useful resource. Barriers to the understanding and use of statistical principles include complex mathematics, abstract ideas, relationship to real applications, required assumptions, unfamiliar terminology, and Greek symbols. This column presents statistical principles with these barriers in mind. It will be our challenge to explain statistical concepts in a meaningful way so that our readers understand these important topics and are able to apply statistical concepts to work situations.

Reader comments, questions, and suggestions are needed to help us fulfill our objective to make this column useful. Suggestions for future discussion topics or questions to be addressed are requested. Readers are invited to participate and contribute manuscripts for this column. Case studies sharing regulatory strategies are most welcome. We need your help to make “Statistical Viewpoint” a useful resource. Please e-mail your comments, suggestions, or manuscripts for publication to the managing editor at shaigney@advanstar.com.

KEY POINTS
This installment of “Statistical Viewpoint” presents eight example scenarios that demonstrate how a probability approach can help with resource planning and decision-making. All examples utilize various discrete or continuous probability distribution calculations available in Microsoft (MS) Excel. This program is readily available and provides an easy method for calculations. A list of respective discrete distributions and continuous distributions including their corresponding applications is provided in Tables I and II, respectively. Selection of the appropriate distribution model for calculations is based on the model’s application “story.” These tools enable quantifiable predictions for common pharmaceutical problems.

INTRODUCTION
The first installment of “Statistical Viewpoint” was published in the Spring 2008 issue of the Journal of GXP Compliance (1) and presented basic probability distribution concepts. Data should not be considered definitive, but thought of as random variables because on repeated measurement, they exhibit different values. Imagine that sample data are drawn from a hypothetical population of all possible values. Underlying physical parameters that describe the data generation process are known only with uncertainty and may also be considered random variables. All repeat data and underlying parameters can be thought of as samples from probability distributions. In choosing a probability distribution, we must understand whether the data are inherently discrete (i.e., take on only specific values) or continuous (i.e., take on infinitely many possible values within some defined range). Discrete data should be modeled using discrete distributions (e.g., binomial). Continuous data should be modeled using continuous distributions (e.g., normal). Probability distributions are a way to understand uncertainty and variation and thus make predictions, manage risk, and make optimal decisions.

Before the advent of the modern personal computer, use of probability distributions required the use of complex tables that often were not readily available. Many complicated probability challenges still require the use of specialized statistical packages, but simple probability problems can often be solved with MS

ABOUT THE AUTHOR
David LeBlond, Ph.D., has 29 years experience in the pharmaceutical and medical diagnostics fields. He is currently a principal research statistician supporting analytical and pharmaceutical development at Abbott. David can be reached at David.LeBlond@abbott.com.
Based on his prior knowledge and experience, Dick believes that the new design will support a 24-month shelf life. Jane’s prior knowledge and experience suggest to her, however, that the new product needs to have a shorter shelf life. Recently, progress has stalled over the stability issue with Jane and Dick locked in unproductive discussions over their differing subjective opinions.

Before proceeding with a lengthy stability study to resolve the impasse, the project director wants to probe deeper and leverage the prior knowledge and experience of his two senior team members. He feels a better study plan will result if the team takes a more quantitative point of view. He meets with Dick and Jane and reminds them of the following advice from Lord Kelvin (2).

> "... when you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot ... express it in numbers, your knowledge is of a meager and unsatisfactory kind ... "

Both chemists agree that the active component is sensitive to oxidation over time and a solution is to include an anti-oxidant in the formulation; however, this measure is not completely effective and the degree of protection seems to vary from lot to lot. Both chemists agree that the variability is due to many small additive effects. The director, therefore, suggests that they express
Table II: Some common continuous distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Excel function</th>
<th>Distribution “Story”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal or Gaussian</td>
<td>D or P =NORMDIST(X, μ, σ)</td>
<td>In the manufacture of a drug lot, an amount of API exactly needed to achieve the desired potency, μ, is added. However a large (i.e., infinite) number of subsequent processing steps (e.g., weighing and adding various excipients) each produce small, random increases or decreases in final lot potency. The net effect produces a standard deviation of σ about the mean of μ. Make a lot and measure the observed potency X.</td>
</tr>
<tr>
<td></td>
<td>X = NORMINV(P, μ, σ) = + σ *NORMINV(P,0,1)</td>
<td></td>
</tr>
<tr>
<td>Lognormal</td>
<td>P=LOGNORMDIST(X,μ,σ) X=LOGINV(P,μ,σ)</td>
<td>The measured response, X, of an analytical test is the multiplicative product of the effects of many small, independent factors (volumes, path-length, etc). The natural log of X follows a normal distribution with mean μ and standard deviation σ.</td>
</tr>
<tr>
<td>Exponential</td>
<td>D or P=EXPONDIST(X,λ)</td>
<td>An electrical supply system experiences occasional random blackouts. The long term mean blackout frequency is known to be constant, λ. Measure X, the time between blackouts.</td>
</tr>
<tr>
<td></td>
<td>X= -LN(1-P)/λ</td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td>D or P =GAMMADIST(X,K,1/λ,*) X= GAMMAINV(P,K,1/λ)</td>
<td>Same as the Exponential above, except measure X, the time required for k &gt; 1 blackouts to occur.</td>
</tr>
<tr>
<td>Weibull</td>
<td>D or P =WEIBULL(X,α,1/α,*)</td>
<td>A very large lot of tablets is known to contain V defects. Tablets are sampled randomly from the lot until W defective tablets have been found. Measure X, the fraction of the lot tested.</td>
</tr>
<tr>
<td></td>
<td>X = -(LN(1-P)/λ)^α/(1/α)</td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>P = BETADIST(X,ω,v−ω+1,0,1) X= BETAINV(P,ω,v−ω+1,0,1)</td>
<td>A very large lot of tablets is known to contain V defects. Tablets are sampled randomly from the lot until W defective tablets have been found. Measure X, the fraction of the lot tested.</td>
</tr>
<tr>
<td>Rectangular or Uniform</td>
<td>P = X D = 1</td>
<td>Same as the Beta, with ω = v = 1.</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>P = 1-CHIDIST(X,v−1) D = GAMMADIST(X,(v−1)/2,2,TRUE) X = CHIINV(1−P,(v−1)/2)</td>
<td>A sample of size v is taken from a normal random variable with known standard deviation = σ. The sample variance (V=square of the standard deviation) is calculated. Measure X=(v−1)·V/σ^2.</td>
</tr>
<tr>
<td>Students-t</td>
<td>if X&lt;0 P = TDIST(ABS(X),v−1,1)</td>
<td>A sample of size v is taken from a normal population with mean μ = 0. The sample average, X , and standard deviation, s, are calculated. Measure X = \frac{X}{s/\sqrt{v}} .</td>
</tr>
<tr>
<td></td>
<td>if X&gt;0 P = 1-TDIST(X,v−1,1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>if 0.5 ≤ P &lt; 1 X = TINV(2·(1-P),v−1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>if 0 &lt; P &lt; 0.5 X = -TINV(2·P, v−1)</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>P = 1-FDIST(X, v−1,ω−1)</td>
<td>A sample of size v is taken from a normal population and the variance (V1) is calculated. A second sample of size ω is also taken from the same population and the variance (V2) is calculated. Measure X = \frac{V_1}{V_2} .</td>
</tr>
<tr>
<td></td>
<td>X = FINV(1−P,v−1,ω−1)</td>
<td></td>
</tr>
</tbody>
</table>

*FALSE produces the probability of density, D, of observing exactly X. TRUE produces the probability, P, of observing x or less (cumulative probability density function).
their opinions about the instability in the form of a continuous normal probability distribution (see the normal distribution story in Table II). He asks each chemist to provide their best estimate of: 1) The expected mean potency loss after 24 months storage, and 2) The expected 90th percentile of the mean potency loss.

After careful thought, both chemists agree that, on average, product lots should lose about 6 units of their potency after 24 months of storage. But they disagree on the 90th percentile. Jane feels that 90% of lots would lose 9 units of potency or less, while Dick feels that 90% of lots would lose no more than 7 units of potency. The director takes 6 units as the normal mean, \( \mu \), and uses the following relationship

\[
\text{NORMINV}(P, \mu, \sigma) = \frac{P - \mu}{\sigma} \]

in Excel to estimate the normal standard deviations in Excel as follows:

For Jane: \( \sigma = \frac{(9 - 6)}{\text{NORMINV}(0.9,0,1)} = 2.34 \)

For Dick: \( \sigma = \frac{(7 - 6)}{\text{NORMINV}(0.9,0,1)} = 0.78 \)

Armed with these prior estimates of the mean and standard deviations, the manager generates Figure 1 using the Excel function \( \text{NORMDIST}(\text{Loss}, 6, \sigma, \text{FALSE}) \). He explains that these distributions capture the prior expectations of each chemist quantitatively. It becomes clear that the difference of opinion is due to opinions about the magnitude of lot to lot variation, with Jane’s prior distribution showing greater spread than Dick’s.

To express the expected failure rate quantitatively, the director uses the equation

\[
\text{Probability that observed loss is greater than 10 units} = 1 - \text{NORMDIST}(10, 6, \sigma, \text{TRUE})
\]

in Excel to estimate the prior probability of making a lot that could not meet the stability requirement of less than 10 units loss on storage. The resulting prior failure rates for Dick and Jane are 0.0000001 and 0.044, respectively. Clearly, based on her prior beliefs, Jane’s concern is warranted. She predicted that 4.4% of future lots could not meet the stability specification. The chemists agree that tighter control over the level of antioxidant would be needed.

While statistics alone cannot resolve a chemical dilemma, this example demonstrates how the use of probability tools can focus discussion by forcing quantitative thinking. By re-expressing a verbal disagreement as a difference in probability distributions, lot to lot variation was identified as the real issue. This recognition might lead to fruitful discussions about process capability improvement that could yield a viable product.

PREDICTING CONTROL CHART TREND TEST DECISION ERROR RATES USING THE NORMAL AND BINOMIAL DISTRIBUTIONS

The Western Electric Handbook suggests a set of decision rules for detecting nonrandom patterns in control charts (3, 4). Such control chart rules act like smoke alarms. They alert us to possible changes or drifts in our methods and processes; the so-called "special causes" that can be investigated. They prevent us from responding to the "white-noise," also known as "common cause" variation that is always present but formidable to investigate. However, even control chart rules are imperfect. What are their false positive alarm rates? If we decide to use these rules, how often can we expect to start an investigation with little hope of uncovering a special cause? We can address these questions by using some simple probability distribution tools.

Four common rules are illustrated in Figure 2. To apply these rules, divide the chart into three zones, A, B, and C that denote the distance in standard deviation (\( \sigma \)) units from the mean (\( \mu \)). In practice, \( \sigma \) and \( \mu \) would be estimated from the control chart data. In estimating the false alarm rate for each rule, we will assume that \( \sigma \) and \( \mu \) are known precisely. We will also assume that there are no special causes present and that the common cause variation is distributed randomly and follows a continuous normal distribution (see Table II). The rules apply to only one side of the center line at a time. We state the rules as follows:

1. One point beyond zone A
2. Two out of three consecutive points beyond zone B
3. Four out of five consecutive points beyond zone C
4. Eight consecutive points on the same side of the center line.

The false alarm rate for rule 1 is simply the probability of observing a single point at least \( 3\sigma \) from the mean. Since the continuous normal distribution is symmetrical, we can focus on the side below the mean.
This probability can be obtained using the following Excel formula:

Rule 1 false alarm rate = NORMDIST(-3,0,1,TRUE)  
= 0.00135

This means that rule 1 can be expected to fail on 1 out of 1350 control chart points or 0.135% of the time. We have to keep in mind that the rule might also fail on the upper side of the chart and so the combined false alarm rate is 0.0027.

Rule 2 is a little more complicated. First note that the probability of exceeding $-2\sigma$ on the lower side of the chart can be obtained using NORMDIST(-2,0,1,TRUE) = 0.0227. Next note that the false alarm rate will be equal to the binomial probability of randomly selecting 2 out of 3 defective points from a large population of points in which 2.27% of the points are “defective.” See the “story” associated with the categorical binomial distribution in Table I. Thus, the required probability can be obtained as follows:

Rule 2 false alarm rate = BINOMDIST(2,3,0.0227,FALSE) = 0.00152

The error probabilities for rules 3 and 4 may be obtained similarly as follows:

Rule 3 false alarm rate = BINOMDIST(4,5, NORMDIST(-1,0,1,TRUE),FALSE)  
= 0.00267

Rule 4 false alarm rate = BINOMDIST(8,8,0.5,FALSE)  
= 0.00391

It can be seen that the rules have been designed so that the one-sided false alarm rate for each is below 0.004 (4 times in a thousand separate sets of points examined). As stated with rule 1, it must be kept in mind that failures can occur on either side of the center line, so the two-sided false alarm rates are double the values calculated above. Another important consideration is the problem of multiplicity. When multiple statistical tests are actively employed, the false alarm rate that one or more will accidentally “sound” may be correspondingly higher. If the tests are independent with very low false alarm rates, the use of 4 tests may increase the overall false alarm rate nearly 4 fold. Thus caution is advised when employing multiple control chart rules together.

**PREDICTING IN-PROCESS CONTENT UNIFORMITY FAILURE RISK USING THE NORMAL AND BINOMIAL DISTRIBUTIONS**

A particular new pharmaceutical tablet product can be manufactured so that, on average, each batch contains 100mg of active drug. However, due to the small additive effects of many process variables, there is within batch, tablet to tablet variation. The within batch standard deviation in tablet drug content is reproducibly 5mg.

Each batch of tablets must pass an in-process unit dose content uniformity test. The test requires a potency assay on each of 10 randomly selected tablets from the batch. The potency assay is very accurate and precise. The test fails and the batch must be discarded if the measured potency of any of the 10 tablets exceeds the range 85 – 115 mg. Failure means that valuable drug product is lost and manufacturing efficiency is reduced. For planning and financing purposes, operations needs to know how often future batches will fail the in-process content uniformity.

The chief engineer knows that, assuming the process is behaving as expected, tablet potencies should be distributed according to a continuous normal distribution (see Table II) with a mean, $\mu$, of 100 mg and standard deviation, $\sigma$, of 5 mg. Thus the probability that a random tablet will exceed the 85 – 115 mg range can be calculated as follows:

$\text{NORMDIST}(85,100,5,\text{TRUE}) + 1 - \text{NOR}$
Similar to the control chart example given above, the probability of obtaining 1 or more out of range tablets in a random sample of 10 will follow a categorical binomial distribution (see Table I) and we can obtain the batch failure rate as

\[
1 - \text{BINOMDIST}(0, 10, 0.0027, \text{TRUE}) = 0.0267
\]

Thus future batches can be expected to fail the in-process content uniformity test 2.67% of the time.

**ASSESSING THE SUITABILITY OF ANALYTICAL METHODS USING THE CHI-SQUARED DISTRIBUTION**

In this scenario, two high-pressure liquid chromatography (HPLC) analytical methods are being developed. It is decided that six control samples will be run with each analytical run. To meet the system suitability requirements (i.e., for the results of unknown samples in that run to be reportable), the observed standard deviation of these \(\nu = 6\) replicate tests must be no more than 1% of the nominal control concentration. A system suitability failure is costly in terms of analyst time and reagents. Further, analytical equipment is limited. Consequently, proper resource management requires the ability to anticipate the rate of system suitability failures.

The two analytical methods are very accurate, but have within run standard deviations (\(\sigma\)) of 0.5 and 0.7%, respectively. In both cases, the within run deviations from the expected control value are distributed normally. This suggests that a continuous Chi-squared distribution might be used to describe the distribution of the within run variance (i.e., \(V\), the square of the observed standard deviation).

The Chi-squared "story" describes sampling from a normal population whose standard deviation is \(\sigma\) (see Table II). This parameter corresponds to the within standard deviations of 0.5 or 0.7%. Thus the probability of exceeding the 1% limit can be calculated for the two analytical methods as follows:  

\[
\text{Failure rate for first method} = \text{CHIDIST}((6-1)^*1^2/0.5^2, 6-1) = 0.0013
\]

\[
\text{Failure rate for second method} = \text{CHIDIST}((6-1)^*1^2/0.7^2, 6-1) = 0.0697
\]

Surprisingly, the small increase in true standard deviation from 0.5 to 0.7% results in nearly a 70-fold increase in system suitability failure rate. The second assay would fail nearly 7 out of every 100 analytical runs.

It is often instructive to actually plot the probability density function to see the shape of the sampling distribution. Unfortunately, Excel does not supply a Chi-squared density function. Such functions are available in software such as SAS, JMP, or R. Direct calculation of the Chi-squared density is possible but quite complicated. Fortunately, the Chi-squared distribution is a special case of the continuous Gamma distribution (see Table II) such that

\[
\text{Chi-squared probability density} = \text{GAMMADIST}((\nu-1)V/\sigma^2, (\nu-1)/2, 2, \text{TRUE})
\]

This equation was used to generate the scaled probability density plots shown in Figure 3, with \(V\) plotted on the horizontal axis. The excessive failure rate with the second assay is clearly due to the fact that the Chi-squared density tends to have a long right tail (i.e., is right skewed) when \(\nu\) is small.

**SETTING A PRECISION SPECIFICATION LIMIT FOR AN ANALYTICAL METHOD TRANSFER USING THE F DISTRIBUTION**

When an analytical method is transferred from one laboratory to another, its use in the second laboratory must first be validated. This is to assure that different analysts, using different equipment and reagents (and possibly slightly different procedures) can achieve the
same analytical performance. A key element of performance for quantitative assays is precision as measured by method standard deviation. One goal of a method transfer protocol is, therefore, to make a statistical comparison of standard deviations. The performance metric might be a ratio of the two observed standard deviations (test/control).

If this performance metric is squared, the resulting statistic is a ratio of two variances. This statistic is denoted $F$ after the English scientist Sir Ronald Fisher who first tabulated its sampling distribution. If the variances produced by the test and control laboratories are in fact truly equal and if the variation of measurements is distributed normally, then the situation matches the F-distribution “story” (see Table II) and we may use the F-distribution to set a statistical acceptance limit for the precision criterion in the method transfer protocol.

Consider a very simple protocol for assessment of precision. Each laboratory tests 10 replicates of a certain control and determines the variance of these 10 replicates. Then the F-variance ratio is calculated (test/control). The precision criteria pass as long as the ratio is at or below some limit. If the limit is set too low, the criterion would fail too often simply by random chance. The acceptance criterion can be reduced by increasing the number of replicates for each laboratory run. Thus the F distribution can be used to select both the required sample size and the decision criterion.

Note that the F test described above is a one-sided test. We have clearly defined our F ratio (test/control) and we only are concerned about failures on the high side (large F values). If we were concerned about testing for equality of the variances we would also be concerned with failures on the low side (small F values). In the latter case, we would need to construct a two-sided test to detect significant differences of either kind.

The cumulative F distribution for a ratio of variances, each obtained from replicates of 10 is given as follows:

$$1 - \text{FDIST}(5.4, 10, 1, 10) = 0.9903.$$  

If one uses an acceptance limit of 5.4, corresponding to a standard deviation ratio of 2.32, the protocol criterion for precision will have a false alarm rate of less than 1% because of random chance alone. The acceptance criterion can be reduced by increasing the number of replicates for each laboratory run. Thus the F distribution can be used to select both the required sample size and the decision criterion.

### SETTING A SPECIFICATION LIMIT FOR A POWDER BLEND VALIDATION USING THE POISSON DISTRIBUTION

A powder blending process must be validated to assure that the process distributes the ingredients evenly throughout the batch. The blend consists of two white ingredients whose particles are of equal, well-controlled size and have very similar physical properties. An engineer has decided to use his newfound understanding of probability distributions to simplify the validation trials. He plans to include a trace amount of a third inert ingredient, also of equal particle size and physical properties, except that the particles are black. They can be easily identified and counted in a gram of powder by eye using a magnifying glass.

The engineer reasons that if the blending is even, the distribution of black “specks” in the final blend should follow a categorical Poisson distribution (see the Poisson “story” in Table I). He plans to introduce...
enough of the trace ingredient so that a gram of the final powder will contain, on average, 20 black specks. The test for each batch will consist of carefully removing exactly 1 gram of powder from a random location in the blended bed using a thief, and counting the number of black specks observed in the sample. If much fewer or much greater than 20 specks are seen, the batch will be judged to fail the blend uniformity validation. Thus the test will be two-sided. The question is what lower and upper limits should be placed on the black speck count. The limit should be set wide enough to avoid failing the criterion by random chance alone, yet not so wide that important levels of stratification in the blend are missed. In choosing the limits, it can be helpful to look at the expected Poisson distribution for this situation. From Table I, the probability mass function is obtained as follows:

\[ \text{Probability of observing exactly } X \text{ black specks} = \text{POISSON}(X,20,\text{FALSE}) \]

The probability mass function is shown in Figure 5 with \( X \) plotted on the horizontal axis. It would appear that normal random variation could easily produce a range of \( X \) values between 10 and 32. The engineer feels that the test should not fail by chance more than 1% of the time. He calculates failure rates on the low and high sides respectively as follows:

- Probability that \( X < 10 = \text{POISSON}(10 - 1, 20, \text{TRUE}) = 0.00499 \)
- Probability that \( X > 32 = 1 - \text{POISSON}(32, 20, \text{TRUE}) = 0.00472 \)

Thus an acceptance criterion of 10 to 32 inclusive would satisfy the required false alarm rate of 1%.

**SETTING A PROCESS MEAN SPECIFICATION IN A COMPARABILITY STUDY USING THE \( t \)-DISTRIBUTION**

The relative accuracy of a new (test) brand of laboratory balance is being compared to the existing (control) brand. The true (population) balance to balance standard deviation can be assumed equal for both brands with normal errors, although the true means and standard deviations are unknown. Ten \((=N)\) balances of each brand will be tested by weighing the same standard once on each balance by the same operator. The 20 balances will be tested in random order. The observed (sample) mean \( (= \bar{X}) \) and standard deviation \( (=S) \) of each brand will be calculated.

In the previous examples, a statistic, whose sampling distribution was constructed under the assumption that the processes being compared were the same, was known; then the acceptance criteria was set for this statistic that gave a low risk of failure by random variation alone. However, we do not know the distribution of either the means or standard deviations. How do we proceed? A solution to this problem was not known before the 20th century. Then in 1908 William Gossett, a brewer working at the Guinness brewery in Dublin, published the appropriate statistic under the pseudonym “Student” (5). His ingenious solution suggests that we form the following “Students \( t \)” statistic as a decision criterion.

\[
t = \frac{\bar{X}_{\text{test}} - \bar{X}_{\text{control}}}{\sqrt{\frac{(N_{\text{test}} - 1)S^2_{\text{test}} + (N_{\text{control}} - 1)S^2_{\text{control}}}{N_{\text{test}} + N_{\text{control}} - 2} \left(\frac{1}{N_{\text{test}}} + \frac{1}{N_{\text{control}}}\right)}}
\]

where, in our case, \( N_{\text{test}} = N_{\text{control}} = 10 \). The numerator of this statistic is simply the estimated mean difference and the denominator is the estimated pooled standard deviation of the mean difference. The trick of
normalizing a statistic by its estimated standard deviation is called “Studentizing.” The resulting t statistic follows a continuous Students-t distribution (see Table II) with \( \nu - 1 = N_{\text{test}} + N_{\text{control}} - 2 \) degrees of freedom.

The cumulative distribution function of t can be obtained as follows:

Probability that \( t \) will be less than \( X \) = 

\[ \text{TDIST}(\text{ABS}(X), 10 + 10 - 2, 1) \text{ for } X \leq 0, \text{ and} \]

\[ 1 - \text{TDIST}(X, 10 + 10 - 2, 1) \text{ for } X > 0. \]

The cumulative distribution function of this statistic is shown in Figure 6 with \( X \) given on the horizontal axis. It would appear that the acceptance range for this statistic of –3 to +3 ought to be considered. The actual risk of failing the test by random chance can be calculated as follows:

Probability that \( t \leq -3 = \text{TDIST}(\text{ABS}(-3), 10 + 10 - 2, 1) \]

= 0.00384

Probability that \( t \geq 3 = 1 - \text{TDIST}(3, 10 + 10 - 2, 1) \]

= 0.00384

So the total probability of failing the test by random chance will be 0.00768, or under 1%.

The above t-test is called a test of equality. For such a test, the only concerns are about the Type I error or false alarm rate when designing the test. However the probability of failing to detect an important difference (also called a Type II error, or broken alarm rate) is a consideration when we are trying to gather evidence of equivalence. Both equality and equivalence tests will be discussed in more detail in future installments of “Statistical Viewpoint.”

ASSESSING THE PREDICTIVE VALUE OF A SCREENING TEST USING THE BERNOULLI DISTRIBUTION AND BAYES’ RULE

A diagnostic screening test for a disease such as cancer or hepatitis is meant to be a fast, safe, and relatively inexpensive procedure. Such tests often have an appreciable false positive rate. Consequently a positive result is usually followed up with a more invasive and costly, but more discriminating confirmatory test. However, receiving a positive result on such a screening test can be a stressful experience for a patient until the more discriminating confirmatory test is able (hopefully) to discount the original screening result.

Just how much new information does a screening test provide? How does receiving a positive screening result change the probability that a patient has the disease? Is the screening test cost-effective? Such questions are not only important to patients and doctors, but also to manufacturers concerned with the quality and appropriateness of such products, and to the regulators who assure the safety and effectiveness of these devices for the public.
In 1764 the work of Reverend Thomas Bayes, a humble English Presbyterian minister, was published posthumously by his friend Richard Price (6). His method is worth considering because it illustrates the value of including prior knowledge (e.g., disease prevalence in a population being screened) when interpreting new data (e.g., results of a recent screening test).

A screening test generally has only two outcomes: + or -. The state of health of a person has only two levels: H or D. These are analogous to a categorical binomial variable with a sample size of 1 (see Table I). This special case of the binomial distribution is referred to as the Bernoulli distribution after the 17th-century Italian mathematician Jacob Bernoulli. Working with the Bernoulli distribution is deceptively simple because the probability mass of a + result is equal to the population proportion, \( \pi \) (see the Bernoulli “story” in Table I). A special Excel function is not needed. The proportion as a probability can be worked with directly.

For example, James is a patient who may have a disease. James has not yet been screened for the disease. What is the probability that he has the disease? James might do an Internet search and learn the prevalence of the disease in people like himself. Prevalence is defined as the probability of having the disease, P(D), for members of a well-defined population. Let’s pretend that this entire population had been screened and that the true disease state is known. Representing the population as the area of a unit square, as shown in Figure 7, is useful because probability can be manipulated as an area or a length. The area of any region within the square is taken as a measure of the fraction of the population that meet certain criteria. This leads to a geometrical solution to James’ question. From the base of the unit square we divide the population with a dashed vertical line separating the diseased subpopulation (on the left) from the healthy subpopulation (on the right). The probability that James is healthy is then

\[
P(H) = 1 - P(D) \tag{Equation 1}
\]

Now James receives his screening result. Unfortunately the result is positive (+). Now, what is the probability that he has the disease, given this positive result? We denote this probability P(D|+), where the vertical bar reads “conditional on” or “given that.” What this means is that we are restricting James to a subpopulation who would have a + screening result. James wants to know P(D|+) and to compare it to the prior probability P(D).

To find P(D|+), we must further divide the diseased (D) and healthy (H) subpopulations in the unit square of Figure 7 into two sub-subpopulations each depending on the proportion who would receive a + (upper) or – (lower) screening test result. This generates four total sub-subpopulations whose joint proportions in the original unit square are P(D+), P(D&-), P(H&+), and P(H&-). The & symbol reads “and.” It is commutative so that P(D&+) = P(+&D). Along the left and right vertical sides of the unit square are four new vertical quantities whose lengths correspond to the conditional probabilities P(+|D), P(-|D), P(+|H), and P(-|H). The first and second probabilities refer to proportions of the diseased subpopulation and the third and fourth probabilities refer to proportions of the healthy subpopulation.

Think of P(D&+) as the proportion of the area of the unit square equal to the proportion of the population that is both diseased and has a + screening result. Inspection of the upper left area Figure 7 shows that

\[
P(D&+) = P(+|D)*P(D) \tag{Equation 2}
\]

Similarly, the total proportion of + test results in the whole unit square population is seen as

\[
P(+) = P(+|D)*P(D) + P(+|H)*P(H) \tag{Equation 3}
\]
Considering \( P(+|H) \) and \( P(-|H) \) as proportions of a unit line that add to 1, we see from the right vertical side of Figure 7 that

\[
P(+|H) = 1 - P(-|H). \quad \text{[Equation 4]}
\]

The same population may also be subdivided differently, as shown in Figure 8, into four subgroups having the same proportions as those in Figure 7, but with different quantities along the horizontal and vertical axes of the unit square. In Figure 8 we first divide the population according to screening test result (+ on left, - on right) then according to disease state (D above and H below). An examination of Figure 8 areas will show that

\[
P(D|+) = \frac{P(D\&+)}{P(+)} \quad \text{[Equation 5]}
\]

Notice from equation 5 that we are making progress in obtaining the quantity \( P(D|+) \) for James. Without boring the reader with the algebra, we simply state that equations 1 through 5 may be combined to obtain the following solution for \( P(D|+) \)

\[
P(D|+) = \frac{P(D|+) \cdot P(D)}{P(D|+) \cdot P(D) + (1 - P(-|H)) \cdot (1 - P(D))}
\]

\[ \text{[Equation 6]} \]

where

\[
P(+|D) = \text{sensitivity of the screening test (probability of a + given disease)}
\]

\[
P(-|H) = \text{specificity of the screening test (probability of a - given health)}
\]

Equation 6, called “Bayes’ Rule” in honor of Thomas Bayes, is finding more applications in today’s era of fast computers. \( P(D) \) is referred to as the prior probability (of disease) before data is available, and \( P(D|+) \) is referred to as the posterior probability (of disease) given that data is available. We can now use Bayes’ Rule to answer James question. Let’s take the following as input to equation 6:

\[
P(D) = 0.01, \text{ that is 1% of the population has the disease}
\]

\[
P(+|D) = 0.99 = \text{test sensitivity}
\]

\[
P(-|H) = 0.98 = \text{test specificity}
\]
Most people would consider a screening test with a 99% sensitivity and 98% specificity a highly discriminating test. However, we can show from Bayes’ Rule that when applied to a population in which only 1% of the people have the disease, the predictive value of a positive result, \( P(D|+) \) is only 0.33. Thus the probability that James has the disease is less than half; it is more likely that he is healthy than diseased. Notice however that this new data (a positive screening result) increased the probability that James has the disease from 0.01 (the prevalence) to 0.33, a 33-fold increase. So the new data does provide significant information. Had James received a negative screening test result, we could similarly obtain his \( P(H|-) \), the predictive value of a negative result (7). Other applications of Bayes’ Rule will be addressed in future columns of “Statistical Viewpoint.”

**CONCLUSIONS**

Decision-making is often facilitated by transforming qualitative descriptions and discussions into quantitative probability models. The previous eight examples were given to illustrate a framework for using probability calculations to make quantitative predictions, plan resources, and assess risk from data and/or prior information for some common pharmaceutical decision-making situations. This approach may be summarized as follows:

- **Know your objective.** If this involves predicting the values of random variables (variable data/statistics, or uncertain parameters) then using probability distribution methods will be useful.
- **Understand the sampling distribution of the data/statistics you are collecting.** Identify the type of data (discrete or continuous) and the distribution story and assumptions that apply to your situation.
- **Try (if possible) to find a match between your own data generation “story” and that of a common distribution, such as those in Tables I or II.**
- **Become familiar with probability functions in readily available software such as Microsoft Excel.** Some of these are presented in Tables I and II.
- **Consider whether it is appropriate to include prior information (e.g., expert opinion, prior relevant data, theory) as part of your predictions.** If so, try to reduce that subjective prior information to probability distributions with appropriate shape and spread. A Bayesian approach may prove fruitful.
- **Do not fail to consult a trained statistician in cases where the objectives, appropriate distribution, assumptions, or the right approach appear unclear.**

Statistics and probability are mature disciplines with at least a 300-year development history. Whatever your particular risk assessment or prediction problem, it is highly likely that a good approach to meeting your objective is well known. In the interest of promoting best practices, a wide range of such approaches will be presented in future installments of “Statistical Viewpoint.”

**GLOSSARY**

**Bayes’ rule:** A process for combining the information in a data set with relevant prior information (theory, past data, expert opinion, and knowledge) to obtain posterior information. Prior and posterior information are expressed in the form of prior and posterior probability distributions, respectively, of the underlying physical parameters.

**Bernoulli distribution:** A special case of the Binomial distribution in which the sample size is equal to 1. A variable following the Bernoulli distribution has only two possible levels: 0 or 1. The probability mass associated with the “1” level is equal to the proportion parameter, \( \pi \).

**Broken alarm rate:** The ability of a statistical test to detect a difference that is large enough to be practically important (e.g., the difference between two means being greater than is acceptable) is a consideration when gathering evidence for equivalence of two processes. The probability that a given test will fail to detect such a difference is called the broken alarm rate, the false negative rate, a Type II error rate, or the beta error rate.

**Conditional probability:** The probability of an event (say A) occurring, given that some other relevant event (say B) has occurred, symbolized \( P(A|B) \). For example, let \( A = “\text{house is on fire}” \) and \( B = “\text{smoke alarm is sounding}.” \)

**False alarm rate:** Because all statistical tests (e.g., control chart rules, comparison with acceptance limits, t- or F-tests) are based on a comparison of a random variable (e.g., some categorical or continuous measure calculated from data) to some fixed decision limit, there is generally a non-zero probability of exceeding the limit (i.e., failing the test) when, in fact, the process
is performing acceptably. This probability is called the false alarm rate (also referred to as the false positive, the “Type I” or “alpha” error rate).

**Multiplicity:** When multiple statistical tests (e.g., control chart rules) are applied to different aspects (e.g., trending patterns) of a data set, the overall false alarm rate of any one test failing may be greater than that of any single test when applied alone. This statistical phenomenon is referred to as multiplicity.

**One-sided or two-sided test:** Statistical tests often involve the comparison of some random statistic, calculated from data (such as an F-ratio, a t value, a count, or other measurement) to some “critical failure range.” In a one-sided test, the test can fail on only one side (i.e., statistic > an upper limit or statistic < a lower limit, but not both). In a two-sided test, both an upper and lower limit are used and thus the test can fail on either side.

**Pooled variance:** An estimate of a population variance obtained from two or more separate samples obtained from the same population or at least from populations having the same population variance. The square root of the pooled variance is called the pooled standard deviation.

**Posterior probability:** An improved estimate of a proportion that is obtained by using Bayes’ rule to combine information from a subjective prior probability with that available from experimental data.

**Predictive value of a failing test result:** The conditional probability that an object under test (e.g., a person) is defective (e.g., diseased), given that the object failed the test (e.g., a “+” result on a diagnostic test).

**Prior probability:** An estimate of a proportion that is based on theory, past data, expert opinion, and knowledge (e.g., prevalence of disease in a population, false positive or negative rates of a screening test). Often the proportion is a parameter of a probability distribution such as the binomial or Bernoulli distribution.

**Predictive value of a passing test result:** The conditional probability that an object under test (e.g., a person) is acceptable (e.g., healthy), given that the object passed the test (e.g., a “-” result on a diagnostic test).

**Sensitivity of a test:** The conditional probability of failing a test (e.g., a “+” result on a diagnostic test), given that the object under test (e.g., a person) is truly defective (e.g., diseased).

**Skewness:** One characteristic of a continuous probability distribution that indicates the degree of symmetry of the distribution about its center (i.e., mean, mode, or median). Distributions showing longer tails at lower or higher values are said to be left or right skewed, respectively. The normal distribution has no skewness and is perfectly symmetrical about its mean. The log-normal distribution is skewed to the right.

**Specificity of a test:** The conditional probability of passing a test (e.g., a “-” result on a diagnostic test), given that the object under test (e.g., a person) is truly acceptable (e.g., healthy).

**Statistic:** A numeric value calculated from data. Often a statistic serves as an estimate of some underlying parameter of an appropriate probability distribution (e.g., the sample average and standard deviation are estimates of the μ and σ parameters, respectively, of the normal probability distribution from which the measurements are drawn).

**Subpopulation:** a subset of a population.

**Sub-Subpopulation:** A subset of a subpopulation.

**REFERENCES**