# Chapter 3

# Introduction to Probability and Correlations

# [H1] Learning Objectives

# [INSERT NL]

1. Define conditional probability as a reduction in the universe of possibilities
2. Use likelihood ratios to measure the impact of a predictor
3. Use standard query language to calculate likelihood ratios
4. Use Bayes odds to revise the chances of the occurrence of an event

**[END NL]**

# [H1] Key Concepts

# [INSERT BL]

* Conditional probability
* Shrinking universe of possibilities
* Prior and posterior odds
* Likelihood ratio
* Contingency table
* Correlation
* Chi-square test of independence
* Bayes’s formula

**[END BL]**

# [H1] Chapter at a Glance

This chapter introduces the concept of probability and relationship, or dependence, among a set of variables. Both probability and conditional probability can be best understood in relation to the universe of possibilities. In massive electronic health record databases, the universe of possible values is available. Probability is easily calculated as the ratio of frequency of the event to all possible events. Conditional probability can be understood as shrinking the universe of possibilities. The universe of possibilities can be shrunk through filtering the data. So, calculation of conditional probabilities translates into repeated shrinking of the universe of possibilities, a task easily done through standard query language. The chapter also introduces Bayes’s formula, independence assumptions, likelihood ratios, contingency tables, and correlations.

# [H1] Probability

Probability is a quantitative measure of uncertainty. When we are sure that an event will *occur*, we say that it has a probability of 1. When we are certain that an event will *not* *occur*, we assign it a probability of zero. When we are completely in the dark, we give it a probability of 0.5—it has a 50 percent chance of occurrence. Values between 0 and 1 measure how uncertain the analyst is about the occurrence of an event. More specifically, the probability of an event is the likelihood or chance that such event will occur. There are three basic types of probability:

**[INSERT BL]**

* Theoreticalprobability
* Subjective or personal probability
* Empirical probability

**[END BL]**

*Theoretical probabilities* are rooted in theory and can be calculated without actually observing data. This is possible if the process responsible for generating an event is known. Theoretical probabilities are typically associated with games of chance such as a coin toss, roll of a die, or drawing a card from a standard deck of 52 playing cards, and they involve mutually exclusive and equally likely events. Events are mutually exclusive if only one of all possible events can occur at any given point in time. Events are equally likely if the probability of occurrence of each event is the same. If there are a total of *n* mutually exclusive and equally likely possible events and *m* of these are event *A*, then the probability of event *A*, denoted by *p*(*A*), is given by the following equation:

**[INSERT EQUATION]**

.

**[END EQUATION]**

For example, three hospitals are competing for a health maintenance organization (HMO) contract. What is the chance that one of them will be awarded the contract? In this scenario, the total number of possible events is three, any of the three participants can be awarded the contract, and one contract must be awarded. Not knowing anything else about each hospital’s bid, we can assume that they have equal chances to get the award. Therefore, the probability is

**[INSERT EQUATION]**

**[END EQUATION]**

We have derived this probability from our knowledge of the process of bidding, making assumptions without actually sampling data—hence, it is theoretical.

*Subjective probabilities* are based on personal experience and consequently may vary across individuals. One can assess the subjective probability that a group of experts assigns to an event and use this consensus to plan the future expansion of a service. Subjective probabilities can be used when theoretical probabilities are not applicable and the observed data are insufficient to form empirical probabilities. If experts are consulted, subjective probabilities may be more accurate than sampled data, especially if the sample is not representative of the population examined.

*Empirical probabilities* are rooted in observed data. Typically *frequency distributions* and *contingency tables* are used to organize observed data, which in turn are used to obtain probability estimates. Empirical probabilities are relative frequencies associated with the occurrence of one or more events. For an event *A* of interest, the probability of *A*, written as *p*(*A*), is given by

**[INSERT EQUATION]**

 ,

**[END EQUATION]**

where *f* is the frequency with which outcome *A* occurs in the observed data, and *n* is the sum of frequencies corresponding to all possible outcomes.

For example, an administrator observed data related to newly admitted patients over a period of 12 months, as seen in exhibit 3.1. He wants to know the probability that the next patient will have a cost overrun.

**[INSERT EXHIBIT]**

**Exhibit 3.1** Cost Overruns

|  |  |
| --- | --- |
| *Cost Overrun* | *Frequency* |
| Over | 399 |
| Under | 665 |
| Total | 1,064 |

**[END EXHIBIT]**

In this example, past data suggest that 399 patients out of a total of 1,064 had cost overruns. The corresponding relative frequency is 399 ÷ 1,064 = 0.375. Thus, the next patient has a 37.5 percent chance of causing a cost overrun.

All probabilities, whether subjective, theoretical, or empirical follow three simple rules. Furthermore, any set of scores that follows these rules is a probability function:

**[INSERT NL]**

1. The probability of an event is a positive number between 0 and 1.
2. One event certainly will happen, so the sum of the probabilities of all events is 1.
3. The probability of any two mutually exclusive events occurring equals the sum of the probability of each occurring.

**[END NL]**

A probability function can be defined by these rules, but doing so provides little insight into what a probability function actually is. One way to think of probability is the frequency of observing an event among all possible events. This is called the *frequentist* definition of probability. There are also other ways of defining probabilities. Probabilities can express strength of opinion. A frequentist definition is not possible for events that have not occurred or do not have repeated observations. In these situations, probabilities are better understood as strength of opinions about the occurrence of an event.

In this book, we rely on the frequentist definition of probability: the prevalence of the event among all possible events. The probability of a small business failing is the number of business failures divided by the number of small businesses. The probability of an iatrogenic infection (a mistake made in a healthcare setting) in the last month is the number of patients who had an iatrogenic infection in our hospital last month divided by the number of patients in our hospital during last month. If we can count the events of interest, then the probability of the event can be calculated, even for rare adverse events. The daily probability of wrong-side surgery in our facility is the number of days in which a wrong-side surgery was reported divided by the number of days in which the facility had surgeries.

We can show a probability graphically in exhibit 3.2 by defining the square to be proportional to the number of possible events and the circle as all the ways in which event *A* occurs; the ratio of the circle to the square is the probability of *A*.

**[INSERT EXHIBIT; render in gray scale; make text inside shapes white; make yellow light gray; make blue medium gray]**

**Exhibit 3.2** A Graphical Representation of Probability



## [END EXHIBIT]

# [H1] Probability Calculus

Notice that for any event of interest *A*, the probability of occurrence of *A* is a complement of the probability that event *A* will not occur and is calculated as

**[INSERT EQUATION]**

 and

,

**[END EQUATION]**

where is the probability that event *A* will not occur. The probability of two mutually exclusive and exhaustive events can be calculated from the same set of relationships, where *A* and *A*ʹ are replaced with the two mutually exclusive and exhaustive events.

There are two basic rules related to probability: the *multiplication rule* and the *addition rule*. They can be used to calculate the joint probability of two or more events. The exact probability formula depends on whether the events are independent. Events are independent when the probability of occurrence of one event does not affect the probability of occurrence of the other event. On the other hand, events are not independent, or their probabilities are conditional, when the probability of occurrence of one event does affect the probability of occurrence of the other event. For any two events *A* and *B* that are independent, the multiplication rule is

**[INSERT EQUATION]**
,

**[END EQUATION]**

where indicates intersection and is the probability that both *A* and *B* will occur. For two events *A* and *B* that are not independent, the multiplication rule is

**[INSERT EQUATION]**

**[END EQUATION]**

where is the conditional probability of the occurrence of event *B* given that event *A* has occurred. If *B* occurred first, then the multiplication rules changes to

**[INSERT EQUATION]**

.

**[END EQUATION]**

The addition rule of probability can be used to calculate the probability that one or another event will occur. The probability of one of two events *A* and *B* occurring is calculated by first summing all the possible ways in which event *A* will occur plus all the ways in which event *B* will occur, minus all the possible ways in which both event *A* and *B* will occur (this term is subtracted because it is double counted in the previous sums). This sum is divided by all possible ways that any event will occur. This is represented in mathematical terms as

**[INSERT EQUATION]**

.

**[END EQUATION]**

The concept is shown graphically in exhibit 3.3.

**[INSERT EXHIBIT; render in gray scale; make text inside shapes white; make yellow light gray; make blue medium gray and add a solid black border to that left-hand light gray rectangle so that I matches the similar rectangle on the right ; make red black. Replace ampersand with “and”]**

**Exhibit 3.3** Graphical Representation of Probability of A or B



**[END EXHIBIT]**

Similarly, the probability of events *A* and *B* co-occurring corresponds to the overlap between *A* and *B* and can be shown graphically as the intersection area divided by all possible events (the rectangular area) in exhibit 3.4.

**[INSERT EXHIBIT; render in gray scale; make blue medium gray; make red black.]**

**Exhibit 3.4** Graphical Representation of Probability A and B



**[END EXHIBIT]**

Repeated application of the definition of probability gives us a simple calculus for combining the uncertainty of two or more events. We can now ask questions such as “What is the probability that the frail elderly (age > 75) or infants will join our HMO?” According to our formulas, this can be calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

Because the chance of being both elderly and an infant is zero (i.e., the two events are mutually exclusive), we can rewrite the formula as

**[INSERT EQUATION]**

.

**[END EQUATION]**

# [H1] Conditional Probability

The definition of probability also helps us calculate the probability of an event conditioned on the occurrence of other events. In these circumstances, we know something has happened and we are calculating the probability of another event. In mathematical terms, we show this as *p*(A|B) and read it as the probability of *A* given *B*. When an event occurs, it reduces the remaining list of possibilities; we no longer need to track the possibility that the event may not occur. We can use our definition of probabilities to calculate conditional probabilities by restricting the possibilities to only events that we know have occurred. Graphically, this is shown as in exhibit 3.5.

**[INSERT EXHIBIT; render in gray scale; make text inside shapes white; make blue medium gray; make red black. Replace ampersands with “and”]**

**Exhibit 3.5** Conditional Probability Is Calculated by Reducing Possibilities



**[END EXHIBIT]**

For example, we can now calculate the probability that a frail elder joins the HMO and is hospitalized. Instead of looking at hospitalization rate among the frail elderly, we need to restrict the possibilities to frail elderly who have joined the HMO. Then the probability is calculated as the ratio of hospitalization among frail elderly HMO members to number of frail elderly HMO members.

Conditional probabilities are a very useful concept. They allow us to think through an uncertain sequence of events. If each event can be conditioned on its predecessor, a chain of events can be examined. Then if one component of the chain changes, we can calculate the impact of the change throughout the chain. In this sense, conditional probabilities show how a series of clues might forecast a future event.

# [H1] Odds

Some people prefer to describe their uncertainty about an event in terms of odds for the event occurring and not use the concept of probability. The two concepts are related. Odds are expressed as ratios, while probabilities are expressed as decimals between 0 and 1. The odds of an event are related to its probability by the formula

**[INSERT EQUATION]**

.

**[END EQUATION]**

For example, if the probability of being in hospice is 90 percent, then the odds of being in hospice are 0.9 ÷ (1 − 0.9) = 9 to 1. If the odds of an event are 2 to 1, the probability of the event is 2 ÷ (1 + 2) = 0.66. Odds and probabilities are always positive numbers. There is no upper limit to an odds ratio, but the maximum probability is 1. Odds of 1 to 1, implies a 50 percent chance or a probability of 0.50. This shows that the person is completely uncertain about the event. Odds of 2 to 1 increase the probability of the event to 0.66. Odds of 3 to 1 increase the probability of the event to 0.75 and odds of 4 to 1 increase the probability of the event to 0.8.

# [H1] Bayes’s Formula

We can derive Bayes’s formula from the definition of conditional probability. Bayes’s formula is an optimal model for revising an existing opinion (sometimes called prior opinion) in the light of new evidence or clues. The theorem states:

**[INSERT EQUATION]**

**[END EQUATION]**

where:

**[INSERT BL]**

* designates probability of the event in the parentheses,
* marks a target event or hypothesis occurring,
* designates the same target event not occurring,
* mark the clues 1 through *n*,
* is the probability of hypothesis *H* occurring given clues 1 through *n*,
* is the probability of hypothesis *H* not happening given clues 1 through *n*,
* is the prevalence of the clues in situations in which hypothesis *H* has occurred, and
* is the prevalence of the clues in situations in which hypothesis *H* has not occurred.

**[END BL]**

Bayes’s theorem states that the posterior odds of the event are calculated as the product of a likelihood ratio of the event times the prior odds:

**[INSERT EQUATION]**

Posterior odds after review of clues = Likelihood ratio associated with the clues × Prior odds.

**[END EQUATION]**

The difference between the left and right side of this equation is the knowledge of clues. Thus, the theorem shows how data (the likelihood ratio) should change prior odds of an event. We are claiming that prior odds of an event are multiplied by the likelihood ratio associated with various clues to obtain the posterior odds for the event. At first glance, this may seem strange. Why multiply, not add? Why not include some other probabilities in addition to prior odds and likelihood ratios? Bayes’s theoremwas first proven mathematically by Thomas Bayes, a British mathematician, although he never submitted his paper for publication. Using Bayes’s notes, fellow British mathematician Richard Price presented a proof of Bayes’s theorem (Bayes and Price 1763). Bayes’s formula follows from very reasonable assumptions regarding how partitions work.

## [H1] Independence Simplifies Bayes’s Formula

When two events are independent, we can calculate the probability of both co-occurring from the marginal probabilities of each event occurring:

**[INSERT EQUATION]**

.

**[END EQUATION]**

Thus, we can calculate the probability of a diabetic having a car accident as the product of the probability of being diabetic and the probability of a car accident in this way:

**[INSERT EQUATION]**

.

**[END EQUATION]**

A related concept is conditional independence. Conditional independence means that for a specific population, the presence of one clue does not change the probability of another. Mathematically, this is shown as

**[INSERT EQUATION]**

.

**[END EQUATION]**

We know that *C* has occurred; telling us that *B* has occurred does not add any new information to the estimate of the probability of event *A*. In other words, in population *C*, knowing *B* does not tell us much about the chance of *A*. As before, conditional independence allows us to calculate joint probabilities from marginal probabilities using the formula

**[INSERT EQUATION]**

.

**[END EQUATION]**

The formula says that among the population *C*, the probability of *A* and *B* occurring is equal to the product of the probability of each event occurring. Independence and conditional independence are often invoked to simplify the calculation of complex likelihoods involving multiple events. We have already shown how independence facilitates the calculation of joint probabilities. The advantage of verifying independence becomes even more pronounced when examining more than two events. Recall that the use of Bayes’s odds requires the estimation of the likelihood ratio. When multiple events are considered before revising the prior odds, the estimation of the likelihood ratio involves conditioning future events on all prior events:

**[INSERT EQUATION]**

*p*(*C1,C2,C3, ...,Cn|H1*) = *p*(*C1|H1*) × *p*(*C2|H1,C1*) ×
*p*(*C3|H1,C1,C2*) × *p*(*C4|H1,C1,C2,C3*) × ... × *p*(*Cn|H1,C1,C2,C3,...,Cn-1*).

**[END EQUATION]**

Note that each term in the formula is conditioned on previous events. When events are considered, the posterior odds are modified, then used to condition all subsequent events. All terms are conditioned on the hypothesis. The first term is conditioned on no additional event; the second term is conditioned on the first event; the third term is conditioned on the first and second events, and so on, until the last term, which is conditioned on all subsequent *n−*1 events. If we stay with our analogy that conditioning is reducing the sample size to the portion of the sample that has the condition, then the formula suggests a sequence for reducing the sample size. Because there are many events, the data have to be portioned in increasingly smaller sizes. Obviously, for data to be partitioned so many times, one needs a large database.

Conditional independence allows us to calculate likelihood ratios associated with a series of events without needing large databases. Instead of conditioning the event on the hypothesis and all prior events, we can now ignore all prior events:

**[INSERT EQUATION]**

**[END EQUATION]**

Conditional independence simplifies the calculation of the likelihood ratios. Now we can rewrite the Bayes’s odds form in terms of the likelihood ratio associated with each event.

**[INSERT EQUATION]**

.

**[END EQUATION]**

To state the concept in words, the posterior odds of an event are the product of the likelihood ratios of each clue times the prior odds of the event. The formula has many applications. It is often used to estimate how various clues (events) may help revise prior probability of a target event. For example, we might use the formula to predict the posterior odds of hospitalization for a frail elderly female patient, if we accept that age and gender are conditionally independent of each other. Suppose the likelihood ratio associated with frail elders is 5:2, meaning that knowing the patient is frail and elderly will increase the odds of hospitalization by 2.5 times. Also suppose that knowing the patient is female reduces the odds for hospitalization by 9:10. Now, if the prior odds for hospitalization were 1:2, the posterior odds for hospitalization can be calculated using the following formula:

**[INSERT EQUATION]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Posterior odds of hospitalization | = | Likelihood ratio associated with being frail and elderly | × | Likelihood ratio associated with being female | ×  | Prior odds of hospitalization. |

**[END EQUATION]**

The posterior odds of hospitalization can now be calculated as

**[INSERT EQUATION]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Posterior odds  | = (5 ÷ 2) | × | (9 ÷ 10) | × | (1 ÷ 2) = 1.125. |

**[END EQUATION]**

For mutually exclusive and exhaustive events, the probability of the event can be calculated from its odds:

**[INSERT EQUATION]**

**[END EQUATION]**

Using the formula, we can calculate the probability of hospitalization as

**[INSERT EQUATION]**

Probability of hospitalization = 1.125 ÷ (1 + 1.125) = 0.53.

**[END EXHIBIT]**

# [H1] Contingency Tables and Likelihood Ratios

A contingency table is a count of how two or more variables are co-occurring. The counts of joint observations are provided in the cells of the table. Exhibit 3.6 shows a hypothetical contingency table. The total number of times that both variables occur is given on the left upper cell as *a*. The number of times neither variable is present is given as *d*. In the margins, the row and column totals are provided. Thus the total number of times that variable 1 is present is a + b. The total number of times that variable 2 is present is a + c.

**[INSERT EXHIBIT]**

**Exhibit 3.6** Sample Contingency Table

|  |  |
| --- | --- |
|  | *Variable 2* |
| *Present* | *Absent* | *Total* |
| Variable 1 | Present | a | b | a + b |
| Absent | c | d | c + d |
| Total | a + c | b + d | a + b + c +d |

**[END EXHIBIT]**

In the contingency table, the universe of possibilities is given by a + b + c + d. The probability of variable 1 occurring is given by the number of times it occurs, a + b, divided by the universe of possibilities, a + b + c + d, written in equation form as

**[INSERT EQUATION]**

.

**[END EQUATION]**

Similarly, the probability of variable 2 occurring can be calculated as the number of times variable 2 occurs divided by the universe of possibilities. The probability of variable 2 not occurring is the number of times it does not occur, or b + d, divided by the universe of possibilities. The odds of variable 1 occurring is calculated as the ratio of its probability of occurring divided by its probability of not occurring, or

**[INSERT EQUATION]**

.

**[END EQUATION]**

The conditional probability is calculated by shrinking the universe of possibilities to occasions where the condition has occurred. For example, to estimate the probability of variable 1 occurring after we know that variable 2 has occurred, the universe is shrunk from a + b + c + d to all situations in which the condition—that is, variable 2—has occurred: a + c. Now the conditional probability of variable 1 is calculated by the frequency of variable 1 in the reduced universe (i.e., a, divided by the reduced universe) in this way:

**[INSERT EQUATION]**

.

**[END EQUATION]**

Many students make a mistake in calculating the conditional probability from a contingency table. The easiest way to avoid making a mistake is to remember to first calculate the reduced universe, then the conditional probability. The conditional probability for variable 2 occurring, given that variable 1 has occurred, is

**[INSERT EQUATION]**

.

**[END EQUATION]**

Note again that first we reduce the universe to all occasions in which variable 1 occurs, and then, in this reduced sample, we look for occurrences of variable 2. If you follow this procedure, then you will not have any difficulty in calculating conditional probabilities involving double negatives, such as, “What is the probability of variable 2 not occurring given that variable 1 has not occurred?”, in this manner:

**[INSERT EQUATION]**

.

**[END EQUATION]**

Bayes’s formula predicts the odds of a variable occurring from the likelihood ratio associated with other variables. Let us call the variable we want to predict the dependent variable and refer to other variables as predictors. The relationship between each predictor and the dependent variable is measured through the concept of likelihood ratios. The best way to think of the likelihood ratio is as a ratio of two conditional probabilities. A likelihood ratio is the ratio of the probability of the predictor when the dependent variable occurs divided by the probability of the predictor when the dependent variable does not occur, written as

**[INSERT EQUATION]**

.

**[END EQUATION]**

Likelihood ratios measure how informative a predictor is. A ratio of 1 indicates that the predictor is not informative. A ratio of .5 means that the predictor is twice as likely to be seen in patients who do not have the dependent variable. A ratio of 10 means the predictor increases the odds of the dependent variable tenfold. The likelihood ratio plays an important role in statistical modeling. It is a concept that we will repeatedly refer to in several places in this book.

# [H1] Contingency Tables in Excel

Contingency tables are useful in examining the relationship between variables. For example, the hypothetical data in exhibit 3.7 provide caloric intake and blood pressure information for 20 patients. Suppose the statistician wishes to know whether there is an association between caloric intake and blood pressure. In real life, many other factors could influence the relationship between caloric intake and blood pressure, but at this point we are simply interested in these two variables.

**[INSERT EXHIBIT]**

**Exhibit 3.7** Data on Caloric Intake and Blood Pressure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Caloric Intake* | *Blood Pressure* |  | *Caloric Intake* | *Blood Pressure* |
| 2,300 | 120/78 |  | 2,900 | 138/89 |
| 2,500 | 124/80 |  | 2,300 | 120/79 |
| 2,200 | 118/75 |  | 2,000 | 112/75 |
| 2,800 | 130/85 |  | 2,100 | 115/78 |
| 3,000 | 136/88 |  | 3,200 | 142/90 |
| 3,500 | 135/86 |  | 3,100 | 140/91 |
| 2,200 | 120/80 |  | 2,450 | 125/80 |
| 2,450 | 125/80 |  | 2,670 | 135/82 |
| 2,600 | 127/87 |  | 2,820 | 138/86 |
| 2,700 | 129/85 |  | 2,900 | 139/89 |

**[END EXHIBIT]**

The first step in building a contingency table is to dichotomize the continuous variables into discrete or categorical variables. This process, called *creating dummy variables*, requires a conceptual understanding of the variables. In this case, using the standard cutoffs for blood pressure and caloric intake, we can come up with the following categories: caloric intake of up to 2,500 kilocalories (kcal) per day (normal) and caloric intake of more than 2,500 kcal (overindulgence). Similarly, blood pressure up to 130/80 will be considered nonhypertensive; over 130/80 will be considered hypertensive. Exhibit 3.8 shows the resulting categories.

**[INSERT EXHIBIT]**

**Exhibit 3.8** Categorical Data for Blood Pressure and Caloric Intake

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Caloric Intake*  | *Blood Pressure*  |  | *Caloric Intake*  | *Blood Pressure*  |
| Normal | Nonhypertensive |  | Over indulgence | Hypertensive  |
| Normal | Hypertensive  |  | Normal | Nonhypertensive |
| Normal | Nonhypertensive |  | Normal | Nonhypertensive |
| Overindulgence | Hypertensive  |  | Normal | Nonhypertensive |
| Overindulgence | Hypertensive  |  | Overindulgence | Hypertensive  |
| Overindulgence | Hypertensive  |  | Overindulgence | Hypertensive  |
| Normal | Nonhypertensive |  | Normal | Hypertensive  |
| Normal | Hypertensive  |  | Overindulgence | Hypertensive  |
| Overindulgence | Hypertensive  |  | Overindulgence | Hypertensive  |
| Overindulgence | Hypertensive  |  | Overindulgence | Hypertensive |

**[END EXHIBIT]**

Using Microsoft Excel’s pivot table feature, the analyst can create a contingency table by choosing blood pressure as the column variable and caloric intake as the row variable. Exhibit 3.9 shows the initial step—indicating the range of the variables in Excel. Once the data range has been specified, the variables are dragged to the column or row of the table. The cell values are specified by dragging a variable into the cell area and asking for a count of the variable.

**[INSERT EXHIBIT]**

**Exhibit 3.9** Specifying the Range of Pivot Table

**[END EXHIBIT]**

Exhibit 3.10 is a contingency table. It shows the count of co-occurrence of caloric intake and hypertension. You might notice that patients who eat more than 2,500 calories per day are more likely to also have hypertension. The likelihood ratio associated with overindulgence is calculated as

**[INSERT EQUATION]**

, thus**[END EQUATION]**

**[INSERT EQUATION]**

, which translates to

**[END EQUATION]**

**[INSERT EQUATION]**

.

**[END EQUATION]**

Note that the likelihood ratio cannot be calculated in this case because of division by zero. In a later chapter, we discuss how to estimate likelihood ratios when dividing by zero.

**[INSERT EXHIBIT]**

**Exhibit 3.10** Contingency Table Produced by Excel

|  |  |  |
| --- | --- | --- |
|  | *Blood Pressure* |  |
| *Row Labels* | *Normal* | *Overindulgence* | *Grand Total* |
| Hypertensive  | 3 | 10 | 13 |
| Nonhypertensive | 6 | 0 | 6 |
| Grand total | 9 | 10 | 19 |

**[END EXHIBIT]**

# [H1] The Chi-Square Test

Even in large samples, some random variations exist in the data. Two variables in the sample, no matter how independent, may not have exactly the same probabilities. Small differences in the calculated probabilities may exist. The real question is whether the differences in the calculated probabilities are large enough that they could not be the result of random variations. The chi‑square test answers this question. The chi-square test allows us to examine whether two events are independent, and whether small differences in the calculated probability could be ignored. To test a hypothesis, statisticians have devised a general method that can also be used for conducting the chi-square test. Throughout *Statistical Analysis of Electronic Health Records*, we repeatedly return to hypothesis testing in different contexts using different methods. The principles across these tests are always the same. This chapter introduces these principles and shows how the chi-square test would be carried out. All hypothesis tests go through the following three steps:

**[INSERT NL]**

1. Make a hypothesis that could be rejected by the available data. Data can reject a hypothesis, but no data can confirm that the hypothesis is correct; the best we can say is that the data did not reject the hypothesis.
2. Calculate a statistic that compares the extent to which data differ from the hypothesized situation.
3. Reject the hypothesis if the probability of observing the statistic is not small—say, less than 0.05. This probability is often referred to as alpha (written . It reflects the chance that we could mistakenly reject the hypothesis when it is true. If the probability is small, then we have little chance of arriving at erroneous conclusions.

**[END NL]**

These steps can be applied to the chi-square test with these three steps:

**[INSERT NL]**

1. The hypothesis is that the two variables are independent.
2. The chi-square test statistic is calculated from the difference between observed and expected probabilities, where expected probability is calculated as if independence assumptions were met. If events are independent, then the expected count of the combinations of two or more events is calculated as
**[INSERT EQUATION]**

**[END EQUATION]**

In this formula, the symbol indicates that the items that follow it should be multiplied by each other. Thus, the formula says that the expected count of two or more independent events is the product of the count of each event divided by the total number of cases in the sample. Once the expected count has been calculated, we can calculate the chi-square statistic using the following formula:
**[INSERT EQUATION]**

.

**[END EQUATION]**

In the above equation, is the observed cell values, is the expected cell values (if there was no association and variables were independent), *c* denotes the degrees of freedom, and the sign ∑ indicates summation across all cell values in a contingency table.

The formula, in essence, relies on the difference between observed and expected values if there was no association. The chi-square test, therefore, assesses the probability of independence of any given distribution by testing the null hypothesis, which in this case states that the variables are independent. In other words, the chi‑square test is used to determine whether an observed distribution is the result of chance (hence the phrase *goodness of fit*) as it compares the observed distribution of the data to the distribution that would have been expected if we were sure that the variables were independent.

The chi-square test is sensitive to sample size—in sufficiently large samples, even small differences are more likely to be statistically significant. It is also highly sensitive to distribution in the cells, so collapsing small categories (those with only a few observations) into one category is advantageous, where possible. The chi-square test may not be accurate if the number of observations that fall in the cell is fewer than five.

3. The chi-square statistic has a standard distribution, from which we can read the probability of observing different values of the statistic. A search on the web provides these probabilities. The information is organized by degrees of freedom, which is the total number of observations minus parameters estimated. As degree of freedom increases, variance decreases, and one expects to see larger values of the chi-square statistic.

**[END NL]**

# [H1] Relationship Among Continuous Variables

Thus far, in this chapter, we have been discussing how binary or discrete variables can be related to each other. In the remaining parts of this chapter, we focus on how a pair of continuous variables might be related to each other. A scatter plot is a plot of two variables, *x* and *y*, at least one of which is continuous. Every time *x* is measured, *y* is also measured. Thus, one has a number of paired readings of the data. These paired data are displayed in a grid; *x* usually defines the horizontal axis and *y* describes the vertical axis. Each pair of measures represents one point on the grid. The order of presentation of these paired measures is irrelevant. *x* can have a high value in one measure, a low value in the next, and a high value after that. The measurements are not taken in order of *x* increasing or decreasing. Thus, making a line plot of these paired measures will be misleading, as line plots show the values of *x* in order.

The convention is that the horizontal axis is reserved for the independent variable and the vertical axis for the dependent variable. A scatter plot can suggest various kinds of relations between the two variables. The relationship is called *rising* if increases in the value of *x* lead to increases in the value of *y*; if increases in the value of *x* lead to decline in the value of *y*, it is referred to as *falling*. If the scatter plot shows a wide dispersion of the data, then there is no relationship between the two variables.

To create a scatter diagram in Excel, first select a range of paired data, then select the Insert tab. In the Charts group, choose Scatter. If you wish to overlay multiple scatter diagrams on top of each other, make separate plots and then select the series in one of the plots and insert it into the other. This procedure ensures that both scatter plots are made on the same range of *x* and *y* values, though different sets of points are shown. Label the points appropriately so the reader can understand their differences.

## [H1] Correlation

Correlation calculates the strength of the relationship between two variables, *x* and *y*. There are different types of correlations. The Pearson correlation calculates for two continuous variables. If there are *n* observations of the pair of variables *x* and *y*, then the Pearson correlation coefficient *r* is calculated as

**[INSERT EQUATION]**

,

**[END EQUATION]**

where is the average of and is the mean value of .

The value of Pearson *r* always lies between −1 and +1. A value of 0 indicates that no linear relationship exists between the two variables (see exhibit 3.11, lower right corner); a nonlinear relationship can still exist (see exhibit 3.11, upper left corner). A value near +1 indicates a strong positive linear relationship (as in exhibit 3.11, lower left corner). In this situation, the two variables are positively related to each other. A value close to −1 indicates a strong negative linear relationship (see exhibit 3.11, upper right corner). In this case, the two variables are inversely related to each other.
**[INSERT EXHIBIT]**

**Exhibit 3.11** Correlation Coefficients for Various Simulated Data

****

**[END EXHIBIT]**

The stronger the correlation, the closer the correlation coefficient comes to either +1 or −1. Hinkle, Wiersma, and Jurs (2003) suggest the following interpretation of various ranges in correlations: greater than 0.7 or less than −0.7 is high; 0.5 to 0.7 or −0.5 to −0.7 is moderate; and 0.5 to −0.5 is low.

The analyst can also test the statistical significance of a correlation coefficient when she hypothesizes that there is no relationship between *x* and *y* (i.e., *r* = 0.0). Correlation can be calculated in Excel using the CORREL function. Free software named R is available for statistical analysis. Different statistical procedures are provided in different R packages. The R code for estimating correlation is given in package COR:

**[LIST FORMAT]**

> cor(dat[,1],dat[,2])

**[END LIST]**

This R command will correlate column 1 of data file “dat” to column 2 of data file “dat.” The two columns must be in the same data frame. If they are not, the bind command can be used.

It is easy to misinterpret correlations. A common misinterpretation is that a correlation of zero means that there is no relationship between the two variables. A nonlinear relationship may still exist. Correlation of zero only says that there is no linear relationship—we cannot assume that nonlinear relationships do not exist. In the upper left corner of exhibit 3.11, for example, we see a nonlinear variable that has a low correlation. If the data were transformed so that the relationship between the two variables were linear, higher correlations would be observed.

# [H1] Summary

This chapter introduced the concept of likelihood ratio and provided an intuitive way of thinking about conditional probabilities as a reduction in the universe of possibilities.

# [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL, and other related material are in the course website.

**[H1] References**

Bayes, T., and R. Price. 1763. [“An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.](http://www.stat.ucla.edu/history/essay.pdf)” Philosophical Transactions of the Royal Society of London 53: 370–418.

Hinkle, D. E., W. Wiersma, and S. G. Jurs. 2003. *Applied Statistics for the Behavioral Sciences*, 5th ed. Boston: Houghton Mifflin.

Bottom of Form