

## **Probability and Probability Distributions - Revised 2/26**

### **Overview**

The Descriptive Statistics module showed how to describe data by computing certain measures of location and certain measures of variation. You also learned how to construct certain graphical presentations to give us a quick look at how the data is distributed. Eventually, you will want to go beyond this and actually make some informed decisions about the behavior of data. The framework for making such decisions is probability. Chapter 3 of Rosner defines probability and a basic set of rules for working with probabilities are introduced. If you are going to understand what is presented in the remainder of this course, you really need to have a basic understanding of the concept of probability. Module 3 also includes the material covered in Chapters 4 and 5 of Rosner. You now are going to be engaged in the process of: looking at data that you or someone else has collected; applying some of the probabilistic framework you have learned; considering specific probability models that may fit our own data. Chapter 4 of Rosner covers discrete random variables and discrete probability distributions. Chapter 5 of Rosner moves from discrete random variables and discrete probability distributions to continuous random variables and continuous probability distributions.

Chapters 3, 4 and 5 of the Cartoon Guide cover much of the same material as Chapters 3, 4, and 5 of Rosner and is recommended as additional reading.

### **Learning Objectives**

1. Understand the concept of probability and the basic laws governing probability: the multiplication law and the addition law.
2. Understand conditional probability.
3. Know what the sensitivity and specificity of a screening test are and the importance of false positives and false negatives in the testing for diseases.
4. Know what the incidence and prevalence of disease are.
5. Understand the definitions of both discrete and continuous random variable.
6. Understand the basic properties of the binomial, Poisson and normal distributions.
7. Know the importance of the standard normal distribution and how it can be used to describe any normal distribution and how it can be used to approximate other distributions.

### **Readings**

Article: Breast MRI may find missed cancer - Specificity , Cost at Issue: 2/7

Chapter 3: 2/2

Chapter 4: 2/9

Chapter 5: 3/2

### **Problems**

3.14 - 3.16, 3.17 - 3.23, 3.113 - 3.115 and the Breast Cancer Article Question: 2/7

4.1- 4.13, 4.36- 4.38, 4.63 - 4.65: 2/14

5.6-5.9, 5.41-5.45, 5.72-5.74, 5.99-5.101: 3/4,3/7

---

### **What You Should Know**

#### **Chapter 3**

Probability is an important concept in statistical work. It is the conceptual framework used to make decisions about data and how it should be interpreted. It is important to understand the definition of probability and a basic set of rules for working with probabilities. Definition 3.1 in Rosner defines probability in terms of a sample space and an event. Make sure you understand this definition. Note in Example 3.3 that the sample space is the set of three possible outcomes of the skin test, and each possible outcome has a probability associated with it. In this particular example they give us only one of the three probabilities but the other two outcomes also have probabilities associated with them.

Probability always falls somewhere between 0 and 1 with the endpoints of 0 and 1 possible outcomes. If two events cannot both happen at the same time, then the probability of one or the other of these two events occurring is the sum of the two individual probabilities (Equation 3.1). Take a look at Example 3.6 in the text to make sure you understand this concept. Two events that cannot both happen at the same time are called "mutually exclusive" events (Definition 3.2). The hypertension examples (3.6 and 3.7) should provide you with a clear understanding about when two events are either mutually exclusive or not mutually exclusive. Also, Figure 3.1 shows the difference graphically.

There are two important rules that you need to know when working with probabilities. If you are asked to compute probabilities you will almost always rely on one or both of these rules. If you understand the difference between independent events and dependent events (Definitions 3.7 and 3.8), then you should also understand the Multiplication Law of Probability (Equation 3.2). The second important law is the Addition Law of Probability. Equation 3.3 provides the general form of this law, while Equation 3.4 provides the Addition Law for independent events. Examples 3.15, 3.16 and 3.17 should help you understand this material.

An additional important topic in probability is the concept of Conditional Probability. It is particularly important in public health and medical studies because of its use in defining relative risk (Definition 3.10). Go through Examples 3.19 and 3.20 carefully to insure you understand how to compute conditional probabilities and relative risk.

Chapter 3 also introduces us to several other important topics that are specific to the health field. Screening tests for diseases are of varying value, depending on how well they identify those who are really ill and how well they identify those who are really free of the disease being screened for. You need to understand the concepts of predictive values (both positive and negative), sensitivity, specificity, and false positive and false negative readings. One of your written assignments later is to read an article on breast cancer screening and make some observations about the screening tests being discussed.

Finally, the definitions of the prevalence and incidence of disease (Definitions 3.19 and 3.20) are given. These are important definitions as these terms will be used often in this course, both in the text and in extra reading assignments.

Please note that you do not have to cover the material in Sections 3.8 (Bayesian Inference) or 3.9 (ROC Curves).

#### *Chapter 4*

Chapter 4 of Rosner introduces us to discrete probability distributions in general and two specific discrete distributions that are very important in statistics, the binomial distribution and the Poisson distribution. First you need to know what a random variable is (Definition 4.1) and then what a discrete random variable is (Definition 4.2). If you look carefully at Examples 4.3 and 4.4 you should get a pretty good idea of what is meant by discrete random variables. This is contrasted with continuous random variables (Definition 4.3 and Example 4.5) which will be covered in Chapter 5. Make sure you can see the difference between discrete and continuous random variables.

Since discrete random variables can take on only a countable number of possible values, by some method you will be able to assign probabilities to each of these possible outcomes. Note that the sum of the probabilities for all possible outcomes must equal 1 and that every individual probability must be between 0 and 1. This concept was covered in Chapter 3. The probability mass function or probability distribution for a discrete random variable (Definition 4.4) is nothing more than the mathematical rule that assigns a probability of occurrence to all possible values of the random variable. Given this information, you can now compare this probability distribution to a sample distribution as described on Pages 84 and 85 of Rosner. Example 4.8 provides an excellent example of how one can compare the frequency distribution from a sample of data with the probability distribution derived from the binomial distribution. This example illustrates what much of what statistics is all about. You have a sample of data, categorize our sample data into a distribution and then want to compare that distribution with some theoretical distribution or some distribution based upon previous experience.

In Chapter 2 of Rosner, you learned how to calculate measures of location (like the arithmetic mean) and measures of variation (like the variance and standard deviation) for a sample of data. In much the same way, you can calculate the expected value of a discrete random variable (analogous to the arithmetic mean for a sample) and the variance of a discrete random variable (analogous to the variance for a sample). Definitions 4.5 and 4.6 provide the formulas for the expected value and variance of a discrete random variable. Equation 4.2 tells us that for many random variables (but certainly not all) approximately 95% of the probability mass will fall within two standard deviations of the mean of a random variable. This will become one of the most important relationships you will use throughout the remainder of the course.

Section 4.7 describes Permutations and Combinations. Definitions 4.8 through 4.11 need to be understood. This basic understanding of factorials, permutations and combinations is needed when you cover the binomial and Poisson distributions.

Finally, you need to know the definition for both the binomial distribution (Equation 4.5) and the Poisson distribution (Equation 4.8). The binomial distribution is applicable to situations where only two outcomes are possible and the outcomes of different trials are independent. This is the classic "success/ failure", "heads/tail" or "yes/no" situation. The Poisson distribution is most useful in describing the distribution of rare events. You need to know how to calculate the expected value and variance for both of these distributions and understand how to use Tables 1 and 2 in the Appendix to compute the probabilities. Carefully go over all the examples in the text in sections 4.8 through 4.13 of the chapter.

#### *Chapter 5*

You now move on to continuous probability distributions in general and the normal distribution in particular. The normal distribution is easily the most important and most widely used distribution in statistics. Success in this course will depend on a good understanding of the material in Chapter 5. The beginning of the chapter talks about all the same things talked about in Chapter 4 for discrete probability distributions, only now for continuous probability distributions. You need to completely understand the concept of "area under the curve" as summarized in Definition 5.1 and demonstrated in Examples 5.3, 5.4 and 5.5.

Because so many things follow a normal distribution or tend toward a normal distribution, it assumes an important place in the field of statistics. Definitions 5.6 and 5.7 introduce us to the terminology used to denote a normal distribution in general and a standard normal distribution in particular. Students often get confused about the standard normal and why it is important. Think of the standard normal as a normal distribution that has a mean

value of zero and a variance of one. When you have data that follows a normal distribution it will have a mean and a variance. But, to analyze the data more easily, you need to convert it into a standard format with a mean zero and variance of one. Once you do that conversion, then you have many tools available to assist us in the analysis. Sections 5.4 through 5.5 takes us through the most important topics thus far covered. You learn about the properties of the standard normal distribution in Section 5.4. Equation 5.2 contains most important information about how the standard normal distribution is shaped. Take a look at Figure 5.9 and you should understand what this figure is depicting. This is the basis for construction of the Standard Normal Table (Table 3 in the Appendix). Do not move on until you feel completely comfortable using Table 3. You will need to understand how to convert any normal distribution into a standard normal distribution so you can use the known properties of the standard normal to compute probabilities. Unless you know how to do this conversion, you will not be able to use Table 3 because its use assumes you are working with a standard normal distribution.

You can skip Section 5.6 of this chapter.

Finally Chapter 5 covers the topics of using the normal distribution to approximate both the binomial and Poisson distributions. The use of the normal distribution as an approximation to the binomial distribution is important when the number of observations is large and the binomial distribution becomes cumbersome to work with. Equations 5.14 and 5.15 describe how the normal approximation to the binomial distribution works and under what circumstances it works the best. The discussion related to figures 5.20 and 5.21 are particularly illustrative of when this approximation works well. Section 5.8 covers the use of the normal distribution to approximate the Poisson distribution. This is particularly useful when dealing with a Poisson distribution with a large value for the expected number of events occurring over the time period  $t$ .

---

## **Problems**

### *Chapter 3*

The assigned problems for Chapter 3 of Rosner and for the Cancer article do not require the use of a data set or the use of any software.

#### *Problems 3.14 -3.16 and 3.17 - 3.23 ---- Computing Probabilities*

These problems will help you to understand whether or not you have a good grasp on how to compute basic probabilities. You may also want to look at Problems 3.1 through 3.11 before doing the assigned problems. The answers to Problems 3.1 through 3.11 are given at the end of Rosner and you can easily check to see how well you understand the basic concepts.

#### *Problems 3.113 - 3.115 ---- Prevalence and Incidence*

These problems focus on the concepts of prevalence rates and incidence rates. These are important public health statistics that you need to understand. Some of these problems require verbal answers to questions.

#### *Breast Cancer Article*

After reading the article "Breast MRI may find missed cancer - Specificity , Cost at Issue" answer the following question: If you were the physician treating a women of high risk for breast cancer and she had no current symptoms or signs of breast cancer, would you recommend screening by MRI or by mammogram? Why?

There is no correct answer to this question. It is important that you understand the concepts of sensitivity and specificity of screening tests. Please justify your answer to this question in terms of your judgement about the importance of false negative results, false positive results, cost considerations, etc.

### *Chapter 4*

As with the problems related to Chapter 3, the following problems do not require the use of a data set or the use of the any software.

#### *Problems 4.1 - 4.13 ---- Discrete Probability Distributions*

These problems address the binomial distribution and the Poisson distribution. Please note that some of the answers to these problems are given in the Appendix to Rosner.

#### *Problems 4.36 - 4.38 ---- Binomial Distribution*

These problems address the binomial distribution and some of them require you to use Table 1 in the Appendix of Rosner.

#### *Problems 4.63 - 4.65 ---- Poisson Distribution*

These problems address the Poisson distribution and some of them require you to use Table 2 in the Appendix of Rosner.

*Chapter 5*

All these problems can be done using a calculator and tables in the Appendix.

*Problems 5.6 - 5.9 --- Nutrition*

*Problems 5.41 - 5.45 --- Blood Chemistry*

*Problems 5.72 - 5.74 --- Cancer , Neurology*

*Problems 5.99 - 5.101 --- Obstetrics*

To answer these problems, you need to convert the data provided to a standard normal distribution and use tables in the Appendix of Rosner. You also have to decide when to use the binomial, Poisson, or normal distribution and when to use the normal approximation to either the binomial or Poisson distribution.