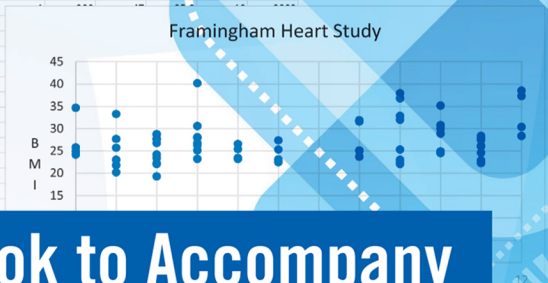
Robert P. Hirsch

sex	sbp	dbp	scl	chdfate	followup	age	bmi	month	id
1	120	80	267	1	18	55	25	8	2642
1	130	78	192	1	35	53	28.4	12	4627
1	144	90	207	1	109	61	25.1	8	2568
1	92	66	231	1	147	48	26.2	11	4192
1	162	98	271	1	169	39	28.4	11	3977
2	212	118	182	1	199	61	33.3	2	659
1	140	85	276	1	201	44	25.3	6	2290
1	174	102	259	1	209	39	27.9	11	4267
1	142	94	242	1	265	47	26.6	5	2035
1	115	70	242	1	278	60	30.8	10	3587
1	202	124	260	1	290	58	28.7	3	1046
2	130	94	326						
1	136	88	185						
2	108	72	255						
2	164	102	232						
2	152	96	221						
1	195	112	192						
1	152	98	265						
1	138	76	265						
2	168	96	309						
1	138	88	228						
1	136	90	232						
1	120	80	221						
1	138	92	225						
2	160	100	263						



Workbook to Accompany

Introduction to

BIOSTATISTICAL APPLICATIONS

in Health Research with Microsoft® Office Excel®

WILEY

WORKBOOK TO ACCOMPANY INTRODUCTION TO BIOSTATISTICAL APPLICATIONS IN HEALTH RESEARCH WITH MICROSOFT® OFFICE EXCEL®

WORKBOOK TO ACCOMPANY INTRODUCTION TO BIOSTATISTICAL APPLICATIONS IN HEALTH RESEARCH WITH MICROSOFT® OFFICE EXCEL®

ROBERT P. HIRSCH
Foundation for the Advanced Education in the Sciences

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

***Library of Congress Cataloging-in-Publication Data*:**

Name: Hirsch, Robert P., author.
Title: Introduction to biostatistical applications in health research with Microsoft Office Excel / Robert P. Hirsch.
Description: New York, NY : John Wiley & Sons Inc., 2016.
Identifiers: LCCN 2015039977 | ISBN: 9781119089650 (cloth) | ISBN: 9781119089865 (workbook)
Subjects: | MESH: Microsoft Excel (Computer file) | Biostatistics–methods. | Data Interpretation, Statistical. | Mathematical Computing.
Classification: LCC R858 | NLM WA 950 | DDC 610.285–dc23 LC record available at http://lccn.loc.gov/2015039977

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

For many students of statistics, working out problems helps their understanding. For those students, I have written this workbook. For each chapter of the textbook there are several "Examples" with detailed answers. In addition, I have provided a summary of each chapter, a glossary, and a list of equations. These are intended to provide readers with a quick reference and as a means of review. Also for each chapter there are problems ("Exercises" with answers for only the odd exercises. The even exercises are problems that might be assigned for grading by your instructor.

The workbook has 13 chapters, but the text has only 12 chapters. The 13^{th} chapter in the workbook addresses using the flowchart to select statistical methods for a given set of data. This chapter provides an overview of the textbook and its flowcharts.

Robert P. Hirsch

Acknowledgements

I gratefully acknowledge the role my students play in challenging me to help them to really understand statistical methods using more than just mathematical explanations.

Notices

The examples in this text are not intended to be a reflection of good clinical practice nor are they intended to provide information on use of medications. Some of the examples use data that have been created or modified to illustrate statistical methods.

CHAPTER 1

Thinking About Chance

CHAPTER SUMMARY

In Chapter 1 we learn that probabilities are useful in thinking about events (things that occur or characteristics that exist) relative to observations (opportunities for things to occur or characteristics to exist). When thinking about probabilities, we use literary, graphic, or mathematic language. In literary language, a probability is the frequency of events relative to the number of observations. In graphic language, we use Venn diagrams to think about probabilities. In Venn diagrams, a circular area usually represents occurrences of the event and a rectangle represents the observations. Then, the probability of the event is reflected by the area of the circle relative to the area of the rectangle. Mathematically, probabilities are proportions, because the number of events is part of the number of observations.

Workbook to Accompany Introduction to Biostatistical Applications in Health Research with Microsoft® Office Excel®, First Edition. Robert P. Hirsch.

Regardless of which language we use to think about probabilities, we notice that probabilities have certain properties. One of these is that a probability must have a value within the range of zero to one. A probability of zero tells us that the event never occurs. A probability of one tells us that the event always occurs. Probabilities between zero and one tell us that the event sometimes occurs.

In addition to thinking about single events, we can use probabilities to think about collections of events. The first collection of events considered in Chapter 1 includes the event and its complement. The complement of an event includes everything that could happen in an observation except the event. Events and their complements always have two characteristics. One is that they are always collectively exhaustive. Events are collectively exhaustive when at least one of the events must occur in every observation. Another characteristic of events and their complements is that they are mutually exclusive. Being mutually exclusive means that, at most, only one of the events can occur in a particular observation.

We can have collections of events other than just a particular event and its complement. Other collections of events might be collectively exhaustive and/or mutually exclusive. An event and its complement, however, are always collectively exhaustive and mutually exclusive.

With other collections of events, we can be interested in two types relationships of the events. These are the intersection and the union of those events. In an intersection of events, we are interested in those observations in which all of the events occur. In a union of events, we are interested in those observations in which at least one of the events occurs.

When we are interested in the intersection of events, we can use the multiplication rule to calculate the probability of the intersection. There are two versions of the multiplication rule. The simplified version involves multiplying the probabilities of each event together. This simplified version is appropriate if the events are statistically independent (from each other). That is, if the probability of each event is the same regardless of whether the other event(s) occur(s). The full version of the multiplication rule uses conditional probabilities.

Many of the probabilities we encounter in health research and practice are conditional probabilities. What distinguishes conditional probabilities from other probabilities is the fact that conditional probabilities address a subset of the observations, rather than all of the observations. That subset of observations is specified by the conditioning event(s). The event(s) addressed by the conditional probability is specified by the conditional event(s). If events are statistically independent, then the probability of the conditional event occurring is the same regardless of whether the conditioning event occurs.

When we are interested in the union of events, we use the addition rule to calculate the probability of the union. There are two versions of the addition rule. In the simplified version, the probabilities of the events in the union are added together. This simplified version of the addition rule can be used when the events are mutually exclusive. If the events are not mutually exclusive, the probabilities of their intersections need to be taken into account.

The conditional and conditioning events in conditional probabilities have very different functions. A conditional probability addresses the probability that the conditional event(s) will occur under the assumption that the conditioning event(s) has occurred. Often we find we are interested in the probability that the conditioning event will occur assuming the conditional event has occurred. One example of this situation is the relationship among the probabilities used in interpreting diagnostic tests. Tests are characterized by their sensitivities and specificities. The conditioning events in sensitivity and specificity are whether or not a person has the disease. To interpret the result of a diagnostic test, however, we want to consider the probability that a person has the disease. In other words, we want to change whether someone has the disease from being the conditioning event to be the conditional event. The way in which we exchange conditional and conditioning events is by using Bayes' theorem.

GLOSSARY

Addition Rule – the method of calculating the union of two or more events. The simplified version involves adding the probabilities of the individual events together. The simplified version can be used if the events are mutually exclusive. The full version for the union of two events involves adding the probabilities of the events together and subtracting the probability of their intersection.

Bayes' Theorem – the mathematical description of the relationship between the probability of one event conditional on the occurrence of another event and the conditional probability of the second event conditional on the occurrence of the first event.

Chance – a process of producing events that has no apparent cause. For example, to say that chance affects whether or not a particular person in the population will be selected to be included in a sample, implies that we know of no characteristics of that person that make it more or less likely they will be selected. See Probability.

Collectively Exhaustive – a collection of events that includes all possible observations. For instance, being male or female is a collectively exhaustive set of genders.

Complement – (of an event) the occurrence of anything except the event. For example, if the event is being exposed to a particular carcinogen, the complement of that event is not being exposed to that carcinogen. An event and its complement are always mutually exclusive and collectively exhaustive.

Conditional Event – the particular event addressed by a conditional probability.

Conditional Probability – the chance that that a particular event will occur in an observation in which another event (or events) has occurred. For example, if we are interested in comparing the chance of getting a disease between persons who are either exposed or unexposed, a conditional probability of interest would be the probability of getting the disease given that a person is exposed.

Conditioning Event – the event that has occurred in an observation that influences the chance that the conditional event will occur in that same observation. For example, if we are interested in comparing the chance of getting a disease between persons who are either exposed or unexposed, the conditioning event would be exposure.

Event – something that happens or exists in an observation. Examples of events include being exposed, having a disease, being a woman, and being selected to be in a sample.

Frequency – how often something happens. For example, if there are 10 persons with a particular disease in a group of persons, the frequency of disease in the group is 10.

Independence – see Statistical Independence.

Intersection – observations that include all the events. For example, the intersection between being exposed and having a disease includes those persons who are both exposed and have the disease.

Multiplication Rule – the method of calculating the intersection of two or more events. The simplified version involves multiplying the probabilities of the individual events together. The simplified version can be used if the events are statistically independent. The full version for the intersection of two events involves multiplying the probability of one of the events by the probability of the other event, given that the first event occurs.

Mutually Exclusive – a collection of events in which only one event can occur in a given observation. For example, suppose that a disease has five stages and each person with the disease can only be in one of those stages. Then, the stages of that disease are mutually exclusive.

Observation – an opportunity for an event to occur. In health research, the most common observations are persons.

Probability – the chance that an event will occur. For example, the probability of someone developing a particular disease might be equal to 0.1. That implies that one-tenth of the observations will develop the disease. A probability is a proportion, so it can have values in the range of 0-1. See Chance and Proportion.

Proportion – a fraction in which the number in the numerator is also included in the denominator. A proportion has a discrete range of possible values ranging from zero to one.

Statistical Independence – a property of two (or more) events in which the probability of one event occurring is not affected by whether the other event (or events) has occurred. For example, if developing a disease and being exposed are statistically independent, the probability of developing the disease is the same regardless of whether a person is exposed.

Unconditional Probability – the probability of an event occurring regardless of whether another event has occurred. For example, the unconditional probability of developing a disease does not separate exposed from unexposed persons.

Union – observations which include one or more of a collection of events. For example, if we consider risk factors for breast cancer as a collection of events, the union of those risk factors includes persons who have at least one risk factor.

Venn Equations – a method of representing probabilities that combines graphic language (Venn diagram) and mathematic language (equation).

Venn Diagram – a graphic representation of the relationship between events and observations in which the area of a figure (usually a circle for events and a rectangle for observations) corresponds to the frequency of an event or observation. For example, a Venn diagram representing the occurrence of a disease would have a circle (representing persons with the disease) inside a rectangle (representing persons). The area of the circle relative to the area of the rectangle would be the same as the number of persons with the disease relative to the total number of persons.

EQUATIONS

$p(\overline{A}) = 1 - p(A)$	the probability of the complement of event A as it relates to the probability of event A. (see Equation{1.2})
$p(A \text{ and } B) = p(A) \cdot p(B\|A)$ $= p(B) \cdot p(A\|B)$	the probability of the intersection of events A and B. Either of the events can be represented by an unconditional probability. Then, the other event is represented by a conditional probability. This is the multiplication rule. (see Equation {1.4})
$p(A\|B) = p(A\|\overline{B}) = p(A)$	three probabilities that are equal to the same value if event A is statistically independent of event B. (see Equation {1.8})
$p(B\|A) = \frac{p(A \text{ and } B)}{p(A)}$	the probability of event B given that event A occurs. (see Equation {1.9})
$p(A\|B) = \frac{p(A \text{ and } B)}{p(B)}$	the probability of event A given that event B occurs.
$p(A \text{ and/or } B) =$ $p(A) + p(B) - p(A \text{ and } B)$	the probability of the union of events A and B. This is the addition rule. (see Equation {1.13})
$p(B\|A) = \frac{p(B) \cdot p(A\|B)}{\left[p(B) \cdot p(A\|B)\right] + \left[p(\overline{B}) \cdot p(A\|\overline{B})\right]}$	the relationship between the probability of event B occurring given that event A occurs and the probability of event A occurring given that event B occurs. This is Bayes' theorem. (see Equation {1.18})

EXAMPLES

Suppose 25 persons who ate a buffet lunch at a particular restaurant developed salmonella infections (a type of food poisoning). As epidemiologists, we are interested in finding out what food from the buffet was associated with becoming ill. To investigate this, we ask the 25 persons who became ill (the cases), and another 100 persons who ate at the buffet, but did not become ill (the controls), what they ate. Imagine we observe the following results for the items offered that are most likely to be the source of the infection:

Table 1.1 Frequencies of eating different foods for cases and for controls.

FOOD	CASES	CONTROLS
Potato salad	10	40
Chicken salad	5	20
Egg salad	2	8
Seafood salad	5	20
Cole slaw	4	16
Deviled eggs	8	32
Turkey	12	18
Dressing	12	24
Chicken	10	40

Eating a particular food is considered an event.

1.1. Are these events mutually exclusive?

To be mutually exclusive, the probability of one event occurring given that another event has occurred must be equal to zero. In this context, mutual exclusion would mean that a person could not eat both turkey and dressing, for example. This is illustrated in Figure 1.1.

This is not likely to be true. Further, we can tell that at least some of the people ate more than one item. Among cases, the total number of items eaten is 68, but there are only 25 cases. Among the controls, the total number of items eaten is 218, but there are only 100 controls. The only explanation for those frequencies is that some people ate more than one item.

1.2. Are these events collectively exhaustive?

To be collectively exhaustive, all of the possible events must be listed. We are told that the listed items were "most likely to be the source of infection." This implies that there were other items on the buffet that were unlikely to cause a salmonella

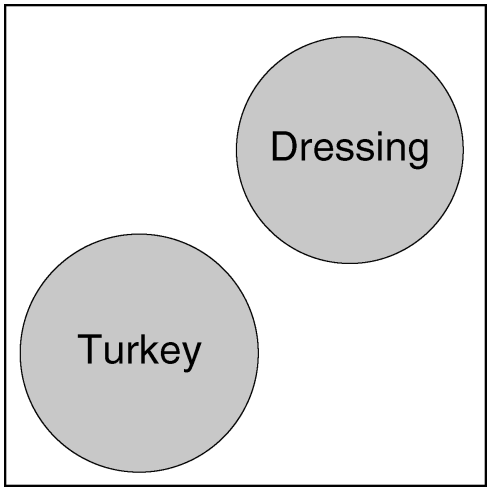


Figure 1.1. Venn diagram illustrating mutual exclusion between eating turkey and eating dressing.

infection. If there were other items on the buffet, then the listed foods are not collectively exhaustive.

To determine which foods might be the source of the infection, we want look at the associations between each type of food eaten and becoming ill. An association between events is the same as the events not being statistically independent.

1.3. To look for statistical independence, what type of probabilities are going to be of primary interest? Why?

Statistical independence is defined by the relationship between conditional probabilities. If two events are statistically independent, then the probability of one event is the same regardless of whether another event occurs. Here, statistical independence means the probability of eating any particular item is the same for cases as it is for controls.

Next, we will change the data in Table 1 so that they reflect probabilities of eating each of the foods for cases and for controls.

1.4. What will be the numerator of those probabilities?

The numerator will be the number of cases eating a particular food or the number of controls eating a particular food.

1.5. What will be the denominator of those probabilities?

The denominator will be the number of cases or the number of controls.

1.6. Put the results of these calculations in a table.

Table 1.2 Probabilities of cases and controls eating particular foods.

FOOD	CASES	CONTROLS
Potato salad	$^{10}/_{25} = 0.40$	$^{40}/_{100} = 0.40$
Chicken salad	$^{5}/_{25} = 0.20$	$^{20}/_{100} = 0.20$
Egg salad	$^{2}/_{25} = 0.08$	$^{8}/_{100} = 0.08$
Seafood salad	$^{5}/_{25} = 0.20$	$^{20}/_{100} = 0.20$
Cole slaw	$^{4}/_{25} = 0.16$	$^{16}/_{100} = 0.16$
Deviled eggs	$^{8}/_{25} = 0.32$	$^{32}/_{100} = 0.32$
Turkey	$^{12}/_{25} = 0.48$	$^{18}/_{100} = 0.18$
Dressing	$^{12}/_{25} = 0.48$	$^{24}/_{100} = 0.24$
Chicken	$^{10}/_{25} = 0.40$	$^{40}/_{100} = 0.40$

1.7. What kind of probabilities are those in Table 1.2?

They are conditional probabilities with case or control as the conditioning event and eating a particular food as the conditional event.

1.8. If there is an association between eating a particular type of food and becoming ill, what relationship would we expect to see between the pairs of probabilities in Table 1.2?

If there is association between eating a particular type of food and becoming ill, then the conditional probabilities for that item will be unequal. Equal conditional probabilities are a sign of statistical independence, which is the same as no association.

1.9. For which food(s) is there an association with becoming ill?

The conditional probabilities for turkey and dressing are equal to different values for cases compared to controls. Cases had a higher probability of eating either of those foods than did controls.

1.10. Now, suppose no one ate both chicken and turkey. What do we call the relationship between those two events?

When the probability of one event is equal to zero if another event occurs, we call the events mutually exclusive.

Suppose we are interested in the probability that someone ate either turkey or chicken.

1.11. In what type of relationship between those events are we interested?

When we are interested in the probability of one and/or more events occurring, we are interested in the union of those events.

1.12. What is the probability that a case ate either turkey or chicken if no one ate both?

We find the union of two or more events by using the addition rule. When the events are mutually exclusive, we can use the simplified version of the addition rule.

$$p(\text{turkey or chicken}|\text{case}) = p(\text{turkey}|\text{case}) + p(\text{chicken}|\text{case}) = 0.48 + 0.40 = 0.88$$

Therefore, 88% of the cases ate either turkey or chicken.

1.13. What is the probability that a control ate either turkey or chicken if no one ate both?

Again, we are interested in the union of eating turkey and/or chicken. Since no one ate both, we can use the simplified version of the addition rule.

$$p(\text{turkey or chicken}|\text{control}) = p(\text{turkey}|\text{control}) + p(\text{chicken}|\text{control}) = 0.18 + 0.40 = 0.58$$

Therefore, 58% of the controls ate either turkey or chicken.

Now, suppose we are interested in how many people ate both potato salad and dressing.

1.14. In what type of relationship between those events are we interested?

When we are interested in both (all) events occurring in the same observation, we are interested in the intersection of the events.

1.15. If the probability of eating potato salad given that someone ate dressing is equal to 0.2 for both cases and controls, what is the probability that a case ate both potato salad and dressing?

Here, we are asked to calculate the probability of the intersection of eating potato salad and eating dressing for cases. We are told the probability of eating potato salad given that a case ate dressing is equal to 0.2. We know the events are not statistically independent since the conditional probability of eating potato salad is equal to 0.2, but the unconditional probability of eating potato salad among cases