A Probability and Statistics Companion

John J. Kinney

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

A Probability and Statistics Companion

A Probability and Statistics Companion

John J. Kinney

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

Copyright ©2009 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Kinney, John J.
A probability and statistics companion / John J. Kinney.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-47195-1 (pbk.)
1. Probabilities. I. Title.

QA273.K494 2009
519.2–dc22
2009009329

Typeset in 10/12pt Times by Thomson Digital, Noida, India.
Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Cherry, Kaylyn, and James

Contents

Preface

Courses in probability and statistics are becoming very popular, both at the college and at the high school level, primarily because they are crucial in the analysis of data derived from samples and designed experiments and in statistical process control in manufacturing. Curiously, while these topics have put statistics at the forefront of scientific investigation, they are given very little emphasis in textbooks for these courses.

This book has been written to provide instructors with material on these important topics so that they may be included in introductory courses. In addition, it provides instructors with examples that go beyond those commonly used. I have developed these examples from my own long experience with students and with teachers in teacher enrichment programs. It is hoped that these examples will be of interest in themselves, thus increasing student motivation in the subjects and providing topics students can investigate in individual projects.

Although some of these examples can be regarded as advanced, they are presented here in ways to make them accessible at the introductory level. Examples include a problem involving a run of defeats in baseball, a method of selecting the best candidate from a group of applicants for a position, and an interesting set of problems involving the waiting time for an event to occur.

Connections with geometry are frequent. The fact that the medians of a triangle meet at a point becomes an extremely useful fact in the analysis of bivariate data; problems in conditional probability, often a challenge for students, are solved using only the area of a rectangle. Graphs allow us to see many solutions visually, and the computer makes graphic illustrations and heretofore exceedingly difficult computations quick and easy.

Students searching for topics to investigate will find many examples in this book.

I think then of the book as providing both supplemental applications and novel explanations of some significant topics, and trust it will prove a useful resource for both teachers and students.

It is a pleasure to acknowledge the many contributions of Susanne Steitz-Filler, my editor at John Wiley & Sons. I am most deeply grateful to my wife, Cherry; again, she has been indispensable.

John Kinney
Colorado Springs
April 2009

Chapter 1

Probability and Sample Spaces

CHAPTER OBJECTIVES:

- to introduce the theory of probability
- to introduce sample spaces
- to show connections with geometric series, including a way to add them without a formula
- to show a use of the Fibonacci sequence
- to use the binomial theorem
- to introduce the basic theorems of probability.

WHY STUDY PROBABILITY?

There are two reasons to study probability. One reason is that this branch of mathematics contains many interesting problems, some of which have very surprising solutions. Part of its fascination is that some problems that appear to be easy are, in fact, very difficult, whereas some problems that appear to be difficult are, in fact, easy to solve. We will show examples of each of these types of problems in this book. Some problems have very beautiful solutions.

The second, and compelling, reason to study probability is that it is the mathematical basis for the statistical analysis of experimental data and the analysis of sample survey data. Statistics, although relatively new in the history of mathematics, has become a central part of science. Statistics can tell experimenters what observations to take so that conclusions to be drawn from the data are as broad as possible. In sample surveys, statistics tells us how many observations to take (usually, and counter-intuitively, relatively small samples) and what kinds of conclusions can be taken from the sample data.

A Probability and Statistics Companion, John J. Kinney
Copyright © 2009 by John Wiley & Sons, Inc.

Each of these areas of statistics is discussed in this book, but first we must establish the probabilistic basis for statistics.

Some of the examples at the beginning may appear to have little or no practical application, but these are needed to establish ideas since understanding problems involving actual data can be very challenging without doing some simple problems first.

PROBABILITY

A brief introduction to probability is given here with an emphasis on some unusual problems to consider for the classroom. We follow this chapter with chapters on permutations and combinations, conditional probability, geometric probability, and then with a chapter on random variables and probability distributions.

We begin with a framework for thinking about problems that involve randomness or chance.

SAMPLE SPACES

An experimenter has four doses of a drug under testing and four doses of an inert placebo. If the drugs are randomly allocated to eight patients, what is the probability that the experimental drug is given to the first four patients?

This problem appears to be very difficult. One of the reasons for this is that we lack a framework in which to think about the problem. Most students lack a structure for thinking about probability problems in general and so one must be created. We will see that the problem above is in reality not as difficult as one might presume.

Probability refers to the *relative frequency* with which events occur where there is some element or randomness or chance. We begin by enumerating, or showing, the set of all the possible outcomes when an experiment involving randomness is performed. This set is called a *sample space*.

We will not solve the problem involving the experimental drug here but instead will show other examples involving a sample space.

EXAMPLE 1.1 *A Production Line*

Items coming off a production line can be classified as either good (G) or defective (D). We observe the next item produced.

Here the set of all possible outcomes is

$$S = \{G, D\}$$

since one of these sample points must occur.

Now suppose we inspect the next five items that are produced. There are now 32 sample points that are shown in Table 1.1.

Table 1.1

Point	Good	Runs	Point	Good	Runs
GGGGG	5	1	*GGDDD*	2	2
GGGGD	4	2	*GDGDD*	2	4
GGGDG	4	3	*DGGDD*	2	3
GGDGG	4	3	*DGDGD*	2	5
GDGGG	4	3	*DDGGD*	2	3
DGGGG	4	2	*DDGDG*	2	4
DGGGD	3	3	*DDDGG*	2	2
DGGDG	3	4	*GDDGD*	2	4
DGDGG	3	4	*GDDDG*	2	3
DDGGG	3	2	*GDDGD*	2	4
GDDGG	3	3	*GDDDD*	1	2
GDGDG	3	5	*DGDDD*	1	3
GDGGD	3	4	*DDGDD*	1	3
GGDDG	3	3	*DDDGD*	1	3
GGDGD	3	4	*DDDDG*	1	2
GGGDD	3	2	*DDDDD*	0	1

We have shown in the second column the number of good items that occur with each sample point. If we collect these points together we find the distribution of the number of good items in Table 1.2.

It is interesting to see that these frequencies are exactly those that occur in the binomial expansion of

$$2^5 = (1 + 1)^5 = 1 + 5 + 10 + 10 + 5 + 1 = 32$$

This is not coincidental; we will explain this subsequently.

The sample space also shows the number of *runs* that occur. A *run* is a sequence of like adjacent results of length 1 or more, so the sample point $GGDGG$ contains three runs while the sample point $GDGDD$ has four runs.

It is also interesting to see, in Table 1.3, the frequencies with which various numbers of runs occur.

Table 1.2

Good	Frequency
0	1
1	5
2	10
3	10
4	5
5	1

Table 1.3

Runs	Frequency
1	2
2	8
3	12
4	8
5	2

We see a pattern but not one as simple as the binomial expansion we saw previously. So we see that like adjacent results are almost certain to occur somewhere in the sequence that is the sample point. The mean number of runs is 3. If a group is asked to write down a sequence of, say, *G's* and *D's*, they are likely to write down too many runs; like symbols are very likely to occur together. In a baseball season of 162 games, it is virtually certain that runs of several wins or losses will occur. These might be noted as remarkable in the press; they are not. We will explore the topic of runs more thoroughly in Chapter 12.

One usually has a number of choices for the sample space. In this example, we could choose the sample space that has 32 points or the sample space {0, 1, 2, 3, 4, 5} indicating the number of good items or the set {1, 2, 3, 4, 5} showing the number of runs. So we have three possible useful sample spaces.

Is there a "correct" sample space? The answer is "no". The sample space chosen for an experiment depends upon the *probabilities* one wishes to calculate. Very often one sample space will be much easier to deal with than another for a problem, so alternative sample spaces provide different ways for viewing the same problem. As we will see, the probabilities assigned to these sample points are quite different.

We should also note that good and defective items usually do not come off production lines at random. Items of the same sort are likely to occur together. The frequency of defective items is usually extremely small, so the sample points are by no means equally likely. We will return to this when we consider *acceptance sampling in* Chapter 2 and *statistical process control* in Chapter 11. ■

EXAMPLE 1.2 *Random Arrangements*

The numbers 1, 2, 3, and 4 are arranged in a line at random.

The sample space here consists of all the possible orders, as shown below.

$$S = \begin{Bmatrix} 1234^* & 2134^* & 3124^* & 4123 \\ 1243^* & 2143 & 3142 & 4132^* \\ 1324^* & 2314^* & 3214^* & 4231^* \\ 1342^* & 2341 & 3241^* & 4213^* \\ 1423^* & 2413 & 3412 & 4312 \\ 1432^* & 2431^* & 3421 & 4321 \end{Bmatrix}$$

S here contains 24 elements, the number of possible linear orders, or arrangements of 4 distinct items. These arrangements are called *permutations*. We will consider permutations more generally in Chapter 2.

A well-known probability problem arises from the above permutations. Suppose the "natural" order of the four integers is 1234. If the four integers are arranged randomly, how many of the integers occupy their own place? For example, in the order 3214, the integers 2 and 4 are in their own place. By examining the sample space above, it is easy to count the permutations in which at least one of the integers is in its own place. These are marked with an asterisk in S. We find 15 such permutations, so $15/24 = 0.625$ of the permutations has at least one integer in its own place.

Now what happens as we increase the number of integers? This leads to the well-known "hat check" problem that involves n people who visit a restaurant and each check a hat, receiving a numbered receipt. Upon leaving, however, the hats are distributed at random. So the hats are distributed according to a random permutation of the integers $1, 2, \ldots, n$. What proportion of the diners gets his own hat?

If there are four diners, we see that 62.5% of the diners receive their own hats. Increasing the number of diners complicates the problem greatly if one is thinking of listing all the orders and counting the appropriate orders as we have done here. It is possible, however, to find the answer without proceeding in this way. We will show this in Chapter 2.

It is perhaps surprising, and counterintuitive, to learn that the proportion for 100 people differs little from that for 4 people! In fact, the proportion approaches $1 - 1/e = 0.632121$ as n increases. (To six decimal places, this is the exact result for 10 diners.) This is our first, but by no means our last, encounter with $e = 2.71828\ldots$, the base of the system of natural logarithms. The occurrence of e in probability, however, has little to do with natural logarithms. ■

The next example also involves e.

EXAMPLE 1.3 *Running Sums*

A box contains slips of paper numbered 1, 2, and 3, respectively. Slips are drawn one at a time, replaced, and a cumulative running sum is kept until the sum equals or exceeds 4.

This is an example of a *waiting time* problem; we wait until an event occurs. The event can occur in two, three, or four drawings. (It must occur no later than the fourth drawing.)

The sample space is shown in Table 1.4, where n is the number of drawings and the sample points show the order in which the integers were selected.

Table 1.4

n	Orders
2	(1,3),(3,1),(2,2) (2,3),(3,2),(3,3)
3	(1,1,2),(1,1,3),(1,2,1),(1,2,2) (1,2,3),(2,1,1),(2,1,2),(2,1,3)
4	(1,1,1,1),(1,1,1,2),(1,1,1,3)

Table 1.5

n	Expected value
1	2.00
2	2.25
3	2.37
4	2.44
5	2.49

We will show later that the *expected number* of drawings is 2.37.

What happens as the number of slips of paper increases? The approach used here becomes increasingly difficult. Table 1.5 shows exact results for small values of n, where we draw until the sum equals or exeeds $n + 1$.

While the value of n increases, the expected length of the game increases, but at a decreasing rate. It is too difficult to show here, but the expected length of the game approaches $e = 2.71828\ldots$ as n increases.

This does, however, make a very interesting classroom exercise either by generating random numbers within the specified range or by a computer simulation. The result will probably surprise students of calculus and be an interesting introduction to e for other students. ■

EXAMPLE 1.4 *An Infinite Sample Space*

Examples 1.1, 1.2, and 1.3 are examples of finite sample spaces, since they contain a finite number of elements. We now consider an infinite sample space.

We observe a production line until a defective (D) item appears. The sample space now is infinite since the event may never occur. The sample space is shown below (where G denotes a good item).

$$S = \left\{ \begin{array}{c} D \\ GD \\ GGD \\ GGGD \\ \cdot \\ \cdot \\ \cdot \end{array} \right\}$$

We note that S in this case is a countable set, that is, a set that can be put in one-to-one correspondence with the set of positive integers. Countable sample spaces often behave as if they were finite. Uncountable infinite sample spaces are also encountered in probability, but we will not consider these here. ■

EXAMPLE 1.5 ***Tossing a Coin***

We toss a coin five times and record the tosses in order. Since there are two possibilities on each toss, there are $2^5 = 32$ sample points. A sample space is shown below.

$$S = \left\{\begin{matrix} TTTTT & TTTTH & TTTHT & TTHTT & THTTT & HTTTT \\ HHTTT & HTHTT & HTTHT & TTTHH & THTHT & TTHHT \\ HTTHH & HTTTH & THTTH & TTHTH & HHHHT & THTHH \\ THHTH & THHHT & TTHHH & HTHTH & THHTT & HHHTT \\ HTHHT & HHTHT & HHHHT & HHHTH & HHTHH & HTHHH \\ THHHH & HHHHH & & & & \end{matrix}\right\}$$

It is also possible in this example simply to count the number of heads, say, that occur. In that case, the sample space is

$$S_1 = \{0, 1, 2, 3, 4, 5\}$$

Both S and S_1 are sets that contain all the possibilities when the experiment is performed and so are sample spaces. So we see that the sample space is not uniquely defined. Perhaps one can think of other sets that describe the sample space in this case. ■

EXAMPLE 1.6 ***AP Statistics***

A class in advanced placement statistics consists of three juniors (J) and four seniors (S). It is desired to select a committee of size two. An appropriate sample space is

$$S = \{JJ, JS, SJ, SS\}$$

where we have shown the class of the students selected in order. One might also simply count the number of juniors on the committee and use the sample space

$$S_1 = \{0, 1, 2\}$$

Alternatively, one might consider the individual students selected so that the sample space, shown below, becomes

$$\begin{aligned} S_2 = \{&J_1J_2, J_1J_3, J_2J_3, S_1S_2, S_1S_3, S_1S_4, S_2S_3, S_2S_4, S_3S_4, \\ &J_1S_1, J_1S_2, J_1S_3, J_1S_4, J_2S_1, J_2S_2, J_2S_3, J_2S_4, J_3S_1, \\ &J_3S_2, J_3S_3, J_3S_4\} \end{aligned}$$

S_2 is as detailed a sample space one can think of, if order of selection is disregarded, so one might think that these 21 sample points are equally likely to occur provided no priority is given to any of the particular individuals. So we would expect that each of the points in S_2 would occur about $1/21$ of the time. We will return to assigning probabilities to the sample points in S and S_2 later in this chapter. ■

EXAMPLE 1.7 ***Let's Make a Deal***

On the television program Let's Make a Deal, a contestant is shown three doors, only one of which hides a valuable prize. The contestant chooses one of the doors and the host then opens one of the remaining doors to show that it is empty. The host then asks the contestant if she wishes to change her choice of doors from the one she selected to the remaining door.

Let W denote a door with the prize and E_1 and E_2 the empty doors. Supposing that the contestant switches choices of doors (which, as we will see in a later chapter, she should do), and we write the contestant's initial choice and then the door she finally ends up with, the sample space is

$$S = \{(W, E_1), (W, E_2), (E_1, W), (E_2, W)\}$$

■

EXAMPLE 1.8 ***A Birthday Problem***

A class in calculus has 10 students. We are interested in whether or not at least two of the students share the same birthday. Here the sample space, showing all possible birthdays, might consist of components with 10 items each. We can only show part of the sample space since it contains $365^{10} = 4.1969 \times 10^{25}$ points! Here

$$S = \{(\text{March 10, June 15, April 24, } \ldots), (\text{May 5, August 2, September 9, } \ldots)\}$$

It may seem counterintuitive, but we can calculate the probability that at least two of the students share the same birthday without enumerating all the points in S. We will return to this problem later. ■

Now we continue to develop the theory of probability.

SOME PROPERTIES OF PROBABILITIES

Any subset of a sample space is called an *event*. In Example 1.1, the occurrence of a good item is an event. In Example 1.2, the sample point where the number 3 is to the left of the number 2 is an event. In Example 1.3, the sample point where the first defective item occurs in an even number of items is an event. In Example 1.4, the sample point where exactly four heads occur is an event.

We wish to calculate the *relative likelihood*, or *probability*, of these events. If we try an experiment n times and an event occurs t times, then the relative likelihood of the event is t/n. We see that relative likelihoods, or probabilities, are numbers between 0 and 1. If A is an event in a sample space, we write $P(A)$ to denote the probability of the event A.

Probabilities are governed by these three axioms:

1. $P(S) = 1$.

2. $0 \leq P(A) \leq 1$.

3. If events A and B are disjoint, so that $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.

Axioms 1 and 2 are fairly obvious; the probability assigned to the entire sample space must be 1 since by definition of the sample space some point in the sample space must occur and the probability of an event must be between 0 and 1. Now if an event A occurs with probability $P(A)$ and an event B occurs with probability $P(B)$ and if the events cannot occur together, then the relative frequency with which one or the other occurs is $P(A) + P(B)$. For example, if a prospective student decides to attend University A with probability $2/5$ and to attend University B with probability $1/5$, she will attend one or the other (but not both) with probability $2/5 + 1/5 = 3/5$. This explains Axiom 3.

It is also very useful to consider an event, say A, as being composed of distinct points, say a_i,with probabilities $p(a_i)$. By Axiom 3 we can add these individual probabilities to compute $P(A)$ so

$$P(A) = \sum_{i=1}^{n} p(a_i)$$

It is perhaps easiest to consider a finite sample space, but our conclusions also apply to a countably infinite sample space. Example 1.4 involved a countable infinite sample space; we will encounter several more examples of these sample spaces in Chapter 7.

Disjoint events are also called *mutually exclusive* events.

Let $\overline{A}$ denote the points in the sample space where event A does not occur. Note that A and $\overline{A}$ are mutually exclusive so

$$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A}) = 1$$

and so we have

Fact 1. $P(\overline{A}) = 1 - P(A)$.

Axiom 3 concerns events that are mutually exclusive. What if they are not mutually exclusive?

Refer to Figure 1.1.

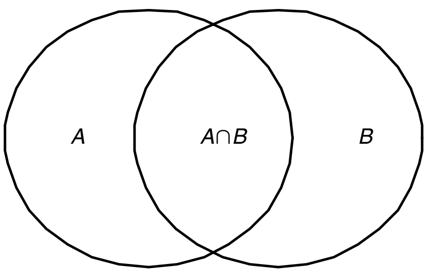

Figure 1.1