**Statistical Inference By Means of Computer Simulations**
**By Howard G. Tucker**

## Introductory Remarks

This text is based on the lecture notes for a one quarter course, Mathematics 7, that I gave during the Spring Quarter in 2004 and Winter Quarter in 2005 at the University of California at Irvine. The students in the first class were mostly majors in the biological sciences. This course is a required course for their major, and in requiring it, the understanding is that sometime in the future they will need to do some statistical analyses and will possibly remember some of the statistical procedures developed in this course. The mathematical preparation needed for this course is high school algebra, and this is perhaps the only mathematical preparation most of the students have. This course is elementary, where the word elementary is used in the sense that mathematicians use the word. However, this course is not easy. It is not easy because it cannot be memorized. It requires a lot of thought, deep thought, logical thought. It requires a certain amount of memory. However, the main theme of this course is this: **don't memorize, but remember**.

I have departed considerably from what is usually presented in this course. Essentially, a text used for this course at present is a fifth or seventh edition of a 650 page book (with small type) on data analysis and statistical inference from which most or all of the topics are expected to be covered or at least mentioned. The book is usually expensive. Contentwise, the statistical inference portion of the text could have been written 50 years ago. In general, it is equivalent to the old Dixon and Massey text that was very popular a half century ago. However, Dixon and Massey was used for a one year course, not a one quarter course.

What exactly do present day texts cover? In general, they offer lots of topics and lots of details. They dwell on all the minutiae of calculations of indices of one sort or another. They dwell mostly on parametric inference with perhaps some elementary nonparametric inference thrown in. They come with extensive tables of distributions and also the attached CD which takes a course of study just to learn how to use it. Enter the profit motive. Every two or three years a new edition is brought out which essentially contains nothing new. It costs plenty, it does away with the second hand market in

text books already in existence, and the author and the publisher make lots of money.

The question that I am concerned with is: does the student really begin to learn how to do statistics, and does he or she understand what he or she is doing, even when doing it correctly? I sincerely doubt it. So what is the real purpose of such a course? It appears to me to be to satisfy a statistics course requirement for a biology major or computer science major, who is pushed into taking the course kicking and screaming. If it is taught in a mathematics department, the course is given to a part-time lecturer or to a visiting postdoc who doesn't know the material and who doesn't know how to select material to fit the time frame of one quarter. Somehow the course is completed, the students receive credit for having satisfied a statistics requirement, the students who have not received grades below their expectations are happy, and the unwilling instructor is paid. This is not a happy situation.

Is there a better way? I feel so. Somehow, most authors of the books to date do not realize that more realistic procedures are possible using simulation techniques. The old fashioned permutation test, originally due to R. A. Fisher sometime around 1936, but unusable at that time because the modern desk top computer was not available then, is a method whose time has come. And it is not difficult to explain to students who are willing to think. So I have assembled here just a few basic statistical procedures, most of which are nonparametric and based on the principle of the permutation test. Only I call them simulation tests. Does one need any fancy software for them? No. They are not too difficult to program using a BASIC compiler. Others can be done using one or two distributions found on EXCEL which is on most computers. Unexpectedly, as I developed the material, I found that there are unifying themes in this course. My conscious aim in developing the material was to teach just a little bit of mathematics and to develop in a plausible manner sound statistical procedures that do not rely on unrealistic assumptions concerning the data. In particular, the normal distribution is used only as an approximation of the binomial distribution, $Bin(n, p)$, when $n$ is large. I do not bring in the standard $t_n$-distribution, nor the $\chi^2_n$-distribution nor the $F_{m,n}$-distribution, nor the Poisson distribution, nor the hypergeometric distribution, let alone some of the more esoteric distributions that creep into the modern encyclopedic texts.

Another important development that I am contributing to here is that this text is free. Yes, free. It is posted on my web page along with the several BASIC programs that I have compiled. This does not cover very

much of statistics. Really, how much can be covered in one academic quarter for students who do not have a more mature mathematical background. But it does cover what are perhaps the most frequently used statistical methods needed and used in the biomedical sciences, and it does cover the reasoning behind them. There is not much to read. But there are many opportunities to think.

## Chapter 1. Representing a Data Set by One Number

**1. Working with Summations.** This chapter contains what one might call methods for handling data. These methods require only a knowledge of high school algebra. They are best learned and understood by writing them as you read on. Writing as you read is a best way to learn mathematics. Indeed, mathematics comes out of the point of a pencil.

**Notation.** If $x_1, \cdots, x_n$ are numbers, then when we write

$$\sum_{i=1}^{n} x_i,$$

we shall mean by this the sum $x_1 + x_2 + \cdots + x_n$. (Start at $i = 1$ and stop when $i = n$.)

For example, suppose you wanted to evaluate

$$\sum_{i=2}^{4} \frac{i}{2i+3}.$$

Note that the sum starts with $i = 2$. So for the first term, you replace $i$ in the expression $\frac{i}{2i+3}$ by 2 to obtain $\frac{2}{(2\times 2)+3}$. Then you add the second term in which 3 replaces $i$ in $\frac{i}{2i+3}$ to obtain $\frac{3}{(2\times 3)+3}$. You stop at the third term where $i = 4$, where you obtain $\frac{4}{(2\times 4)+3}$. Putting it all together, we obtain.

$$\begin{aligned} \sum_{i=2}^{4} \frac{i}{2i+3} &= \frac{2}{(2\times 2)+3} + \frac{3}{(2\times 3)+3} + \frac{4}{(2\times 4)+3} \\ &= \frac{2}{7} + \frac{3}{9} + \frac{4}{11} \\ &= 0.9827 \end{aligned} .$$

There are basic properties of summations of this sort that will be used throughout this text. We gather some of them here along with their proofs.

**Proposition 1.** If $x_0, x_1, \cdots, x_n, y_0, y_1, \cdots, y_n$ and $c$ are numbers, then the following distributive and commutative laws of real numbers,

$$\sum_{i=0}^{n} cx_i = c \sum_{i=0}^{n} x_i$$

and

$$\sum_{i=0}^{n} x_i + \sum_{i=0}^{n} y_i = \sum_{i=0}^{n}(x_i + y_i),$$

are true.

Proof: You could prove this by mathematical induction. First check if the two equations are true for $n = 1$. Yes, they are true because, in the first equation, $cx_0 + cx_1 = c(x_0 + x_1)$ and, in the second equation, $(x_0 + x_1) + (y_0 + y_1) = (x_0 + y_0) + (x_1 + y_1)$. Then, if $n$ is a positive integer for which both equations are true, we shall prove both are true for $n + 1$. Indeed, by associativity, distributiovity and induction hypothesis, the first equation is true for $n + 1$, since

$$\begin{aligned} c\sum_{j=0}^{n+1} x_j &= c\left(\sum_{j=0}^{n} x_j + x_{n+1}\right) \\ &= c\sum_{j=0}^{n} x_j + cx_{n+1} \\ &= \sum_{j=0}^{n} cx_j + cx_{n+1} \\ &= \sum_{j=0}^{n+1} cx_j. \end{aligned}$$

As for the second equation, we again observe that it is true for $n = 1$ by noting that

$$(x_0 + x_1) + (y_0 + y_1) = (x_0 + y_0) + (x_1 + y_1).$$

Then for every value of $n$ for which it is true, we have, by commutativity and by induction hypothesis,

$$\begin{aligned} \sum_{i=0}^{n+1} x_i + \sum_{i=0}^{n+1} y_i &= \sum_{i=0}^{n} x_i + x_{n+1} + \sum_{i=0}^{n} y_i + y_{n+1} \\ &= \sum_{i=0}^{n} x_i + \sum_{i=0}^{n} y_i + (x_{n+1} + y_{n+1}) \\ &= \sum_{i=0}^{n}(x_i + y_i) + (x_{n+1} + y_{n+1}) \\ &= \sum_{i=0}^{n+1}(x_i + y_i). \end{aligned}$$

This proves the proposition.

**Proposition** 2. If $x_1, \cdots, x_n$ and $c$ are numbers, then

$$\sum_{j=1}^{n}(x_j + c) = \sum_{i=1}^{n} x_i + nc.$$

Proof: By proposition 1, if we let $y_i = c$ for $1 \leq i \leq n$, we have

$$\sum_{j=1}^{n}(x_j + c) = \sum_{j=1}^{n} x_j + \sum_{j=1}^{n} c = \sum_{i=1}^{n} x_i + nc.$$

**Proposition** 3. **(The Telescoping Procedure)** If $x_0, \cdots, x_n$ are numbers, then

$$\sum_{j=1}^{n}(x_j - x_{j-1}) = x_n - x_0.$$

Proof: Again we have an opportunity to do a proof by induction. Clearly, for $n = 1$, we have $x_1 - x_0 = x_1 - x_0$. Now for any value of $n$ for which the proposition is true, that is, for any value of $n$ for which

$$\sum_{j=1}^{n}(x_j - x_{j-1}) = x_n - x_0$$

is true, we have, by induction hypothesis,

$$\begin{aligned}\sum_{j=1}^{n+1}(x_j - x_{j-1}) &= \sum_{j=1}^{n}(x_j - x_{j-1}) + x_{n+1} - x_n \\ &= x_n - x_0 + x_{n+1} - x_n \\ &= x_{n+1} - x_0,\end{aligned}$$

so it is true for $n + 1$. This proves the theorem.

Another result that will be needed from time to time is this:

**Proposition** 4**.** For every positive integer $n$, the following is true:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

Proof: Note that we may add these numbers from the largest, $n$, down to the smallest, 1. Thus, if $S$ denotes the sum of the first $n$ positive integers, we have

$$S = \sum_{i=1}^{n} i \text{ and } S = \sum_{i=1}^{n}(n - i + 1).$$

Now by the second conclusion of Proposition 1, we have

$$2S = \sum_{i=1}^{n}(i + (n - i + 1)) = \sum_{i=1}^{n}(n + 1) = n(n + 1).$$

Solving for $S$, we obtain $S = \frac{n(n+1)}{2}$, which proves the proposition.

Thus, if one has to add the numbers from 1 to 100, all one has to do is multiply 100 by 101 and divide the prooduct by 2. The answer is 5050.

**Proposition** 5. If $x_0, x_1, \cdots, x_n, y_0, y_1, \cdots, y_n$ are numbers such that $x_i \leq y_i$ for $0 \leq i \leq n$, then

$$\sum_{i=0}^{n} x_i \leq \sum_{i=0}^{n} y_i.$$

Proof: The hypothesis implies that $x_i - y_i \leq 0$ for $0 \leq i \leq n$. Since the sum of negative numbers is negative, we have, by proposition 1, that

$$\sum_{i=0}^{n}(x_i - y_i) = \sum_{i=0}^{n} x_i - \sum_{i=0}^{n} y_i \leq 0,$$

from which we obtain the conclusion of the proposition.

**Exercises**

1. Prove: If $c$ is a real number, then $\sum_{j=1}^{n} c = nc$.
2. Evaluate $\sum_{j=1}^{n} 1$.
3. Evaluate $\sum_{j=6}^{25} j$. (Hint: Use the formula $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ that was established in Proposition 4 combined with the fact that $\sum_{j=1}^{25} j = \sum_{j=1}^{5} j + \sum_{j=6}^{25} j$ .)
4. Evaluate $\sum_{j=26}^{250} \left(\frac{j+1}{j+2} - \frac{j}{j+1}\right)$. (Do not work too hard.)
5. Evaluate $\sum_{h=6}^{13} 3$.
6. Evaluate $\sum_{i=1}^{20}(3i + 2)$.
7. Evaluate $\sum_{j=15}^{50}(4j + 1)$.

**2. The Sample Mean.** The word "data" in this text generally refers to some numbers recorded as a result of some experiment. Note that the word "data" is the plural of the Latin noun "datum". Thus we shall write, "These data are ...", rather than, "This data is...". We shall also, from time to time, refer to a collection of data as a data set. We shall consider in this section one of the ways of using just one number to "represent" a set of data, $x_1, \cdots, x_n$.

**Notation**: If $x_1, \cdots, x_n$ are data, then we define the **sample mean**, $\overline{x}$, of these data by

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

This is nothing other than the average. In high school physics you learned something about the center of gravity, or centroid, of a spread of mass. If one unit of mass is placed at each of the points $x_1, \cdots, x_n$ on an $x$-axis, then if a fulcrum is placed at the point $\overline{x}$, the system of masses will be in perfect balance. In statistics, the sample mean is used as one of the numbers that in some sense summarizes a data set. In order to justify this notion, the next

proposition will give us a feeling that $\overline{x}$ is somewhere between the smallest number and the largest number in a data set.

**Notation.** If $x_1, \cdots, x_n$ are data, then $\max\{x_1, \cdots, x_n\}$ will denote a largest of the numbers, and $\min\{x_1, \cdots, x_n\}$ will denote a smallest of the numbers.

**Proposition** 1. If $x_1, \cdots, x_n$ are data, then $\min\{x_1, \cdots, x_n\} \leq \overline{x} \leq \max\{x_1, \cdots, x_n\}$.

Proof. Let $c = \min\{x_1, \cdots, x_n\}$, and let $d = \max\{x_1, \cdots, x_n\}$. Then, $c \leq x_i \leq d$ for $1 \leq i \leq n$, so, by Proposition 5 at the end of section 1,

$$\sum_{j=1}^{n} c \leq \sum_{j=1}^{n} x_j \leq \sum_{j=1}^{n} d.$$

The lower end sum adds $c$ exactly $n$ times to obtain $nc$, and the upper end sum adds $d$ exactly $n$ times to obtain $nd$. Thus

$$nc \leq \sum_{j=1}^{n} x_j \leq nd.$$

Dividing through by $n$, we obtain

$$c \leq \frac{1}{n} \sum_{j=1}^{n} x_j \leq d, \text{ or } c \leq \overline{x} \leq d,$$

which is the conclusion desired.

If we were to use a number $c$ to represent a set of data, $x_1, \cdots, x_n$ in some best way, we should want to make all of the positive differences, or errors, $|x_1 - c|$, $|x_2 - c|$, $\cdots$, $|x_n - c|$, as small as possible. One way of doing this is by considering the squares of these errors, or distances from $c$ to each of the numbers, and finding a value of $c$ that minimizes this sum of squares of errors.

**Proposition** 2. If $x_1, \cdots, x_n$ are data with sample mean $\overline{x}_n$, then the value of $c$ that minimizes the sum of the squares of the errors, or distances, from $c$ to each of the numbers,

$$\sum_{i=1}^{n} (x_i - c)^2,$$

is $c = \overline{x}$.

Proof: First we should have clearly in mind what we mean by "the value of $c$ that minimizes $\sum_{i=1}^{n} (x_i - c)^2$." It means a number $c_0$ that satisfies the

inequality $\sum_{i=1}^n (x_i - c_0)^2 \leq \sum_{i=1}^n (x_i - c)^2$ for all possible values of $c$. We may write

$$\begin{aligned} \sum_{i=1}^n (x_i - c)^2 &= \sum_{i=1}^n ((x_i - \overline{x}) + (\overline{x} - c))^2 \\ &= \sum_{i=1}^n (x_i - \overline{x})^2 + 2\sum_{i=1}^n (x_i - \overline{x})(\overline{x} - c) + n(\overline{x} - c)^2. \end{aligned}$$

But, factoring out $(\overline{x} - c)$ in the second sum on the right hand side, we have

$$\sum_{i=1}^n (x_i - \overline{x})(\overline{x} - c) = (\overline{x} - c)\sum_{i=1}^n (x_i - \overline{x}).$$

But, by Proposition 2 in section 1 and the definition of $\overline{x}$,

$$\sum_{i=1}^n (x_i - \overline{x}) = (\sum_{i=1}^n x_i) - n\overline{x} = 0.$$

Thus

$$\sum_{i=1}^n (x_i - c)^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - c)^2.$$

Since $(\overline{x} - c)^2 \geq 0$, the smallest value that the right side can achieve is when and only when $c$ is chosen to satisfy $n(\overline{x} - c)^2 = 0$, i.e., when $c = \overline{x}$. This proves the proposition.

Proposition 2 allows us to say that $\overline{x}$ is the number that "best summarizes the data in the sense of least squares". We observe the following useful property of the sample mean.

**Proposition** 3. If $x_1, \cdots, x_n$ are data with sample mean $\overline{x}_n$, if $y_1, \cdots, y_n$ are numbers that are defined by $y_i = ax_i + b$ for $1 \leq i \leq n$, and if $\overline{y}$ denotes the sample mean of the data $y_1, \cdots, y_n$, then

$$\overline{y} = a\overline{x} + b.$$

Proof: We observe that

$$\begin{aligned} \overline{y} &= \tfrac{1}{n}\sum_{i=1}^n y_i \\ &= \tfrac{1}{n}\sum_{i=1}^n (ax_i + b) \\ &= a\tfrac{1}{n}\sum_{i=1}^n x_i + b = a\overline{x} + b. \end{aligned}$$

**Definition**: If $x_1, \cdots, x_n$ are data, we define the **sample variance**, $s_x^2$, of this data set by

$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^n (x_j - \overline{x})^2.$$

Notice that if all of the data are close together, the value of the sample variance is small, but if the data are spread out, the sample variance will be

large. Thus, the sample variance can be considered as a measure of the spread or variability of the data. This is brought out by the following proposition.

**Proposition 4**. If $x_1, \cdots, x_n$ are data, and if $y_1, \cdots, y_n$ are numbers that are defined by $y_i = ax_i + b$ for $1 \leq i \leq n$, if $s_x^2$ denotes the sample variance of the $x$-data, and if $s_y^2$ denotes the sample variance of the $y$-data, then $s_y^2 = a^2 s_x^2$.

Proof: By proposition 3, $\overline{y} = a\overline{x} + b$, so

$$\begin{aligned} s_y^2 &= \tfrac{1}{n-1}\textstyle\sum_{j=1}^n (y_i - \overline{y})^2 \\ &= \tfrac{1}{n-1}\textstyle\sum_{j=1}^n ((ax_i + b) - (a\overline{x} + b))^2 \\ &= \tfrac{a^2}{n-1}\textstyle\sum_{j=1}^n (x_i - \overline{x})^2 = a^2 s_x^2. \end{aligned}$$

Let us look at the above proposition a little longer. If $a > 1$, then the transformed data, $y_1, \cdots, y_n$, have larger spaces between them than the original data, while if $0 < a < 1$, the values of the transformed data are closer together. Notice also that the sample variance of the $y$-data does not depend on the value of $b$. This is a little confirmation that the sample variance is a measure only of spead or variability only.

## Exercises

1. Suppose the numbers in the data set $x_1, \cdots, x_5$ are $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, and $x_5 = 5$, i.e., $x_i = i$ for $1 \leq i \leq 5$. Suppose $y_i = 3x_i + 10$. Then

($i$) Display the data set $\{y_i, 1 \leq i \leq 5\}$.

($ii$) Compute $\overline{x}$, $\overline{y}$, $s_x^2$ and $s_y^2$.

($iii$) Check that propositions 3 and 4 are true for these two data sets.

2. Prove: If $x_1, \cdots, x_n$ are data, and if $c$ is any constant, then

$$\frac{1}{n}\textstyle\sum_{j=1}^n (x_j - c) = \overline{x} - c.$$

3. Prove: If $x_1, \cdots, x_n$ are data, then $\sum_{j=1}^n (x_j - \overline{x}) = 0$.

4. Prove: If $x_1, \cdots, x_n$ are data, then $s_x^2 = \frac{1}{n-1}(\sum_{j=1}^n x_j^2 - n\overline{x}^2)$.

5. If $x_1, \cdots, x_n$ are data for which $\overline{x} = 7$ and $s_x^2 = 23$, and if $y_1, \cdots, y_n$ are numbers defined by $y_i = 3x_i + 2$, evaluate $\overline{y}$ and $s_y^2$.

6. Prove: If $x_1, \cdots, x_n$ are data, and if $y_i = -x_i$ for $1 \leq i \leq n$, then $\overline{x} = -\overline{y}$ and $s_x^2 = s_y^2$.

7. Draw an accurate graph of $y = f(x)$, where $f(x) = \sum_{i=1}^{4} |i - x|$, for $-1 \leq x \leq 6$.

**3. The Sample Median.** There is more than one way to represent a data set by one number. Another way is by means of the sample median. The sample median of a data set, $x_1, \cdots, x_n$, will be of great use in the methods of statistical inference subsequently developed.

When a data set is obtained, the numbers in it are not necessarily arranged in any particular order. For some purposes, we shall wish to have the data represented in nondecreasing order. Here is an example of this. Consider the following data set:

25.1, 23.6, 27.2, 25.1, 21.3, 26.4, 25.3.

The arrangement of these data in nondecreasing order is as follows:

21.3, 23.6, 25.1, 25.1, 25.3, 26.4 , 27.2.

Note that we must use the word "nondecreasing" rather than "increasing", because there may be repetitions, or ties, in the data set, as in this set.

If $x_1, \cdots, x_n$ are data, then we shall denote their rearrangement in nondecreasing order by $x_{(1)}, \cdots, x_{(n)}$. In the example given above, the original set of data are $x_1 = 25.1$, $x_2 = 23.6$, $x_3 = 27.2$, $x_4 = 25.1$, $x_5 = 21.3$, $x_6 = 26.4$, $x_7 = 25.3$, but $x_{(1)} = 21.3$, $x_{(2)} = 23.6$, $x_{(3)} = 25.1$, $x_{(4)} = 25.1$, $x_{(5)} = 25.3$, $x_{(6)} = 26.4$, $x_{(7)} = 27.2$. Notice that the subscrpts of this reordered set of numbers are enclosed by parentheses. This reordered set of numbers will be referred to as the **order statistics** of the sample. In particular, $x_{(1)}$ is called the first order statistic of the sample, $x_{(2)}$ is called the second order statistic of the sample, etc.

**Definition.** If $x_1, \cdots, x_n$ are data, we shall have slightly different definitions for sample median, depending on whether $n$ is odd or even. If $n$ is odd, in other words, if $n = 2m + 1$ for some positive integer $m$, then we define the sample median as the number $x_{(m+1)}$. If $n$ is an even number, say, $n = 2r$, then we shall define the sample median as $(x_{(r)} + x_{(r+1)})/2$.

Note that in the example given above, the sample median is $x_{(4)} = 25.1$.

When we developed the notion of the sample mean in section 2, we showed that it had a very nice property, namely, that it minimized the sum of squares of the errors. We shall now show that the sample median has an optimal property too in that it minimizes the sum of the absolute values, or sizes, of the errors.

**Proposition 1.** If $x_1, \cdots, x_n$ are data, then the sample median of the data is a number $c$ that minimizes the sum of the absolute values, or distances, of the errors,

$$\sum_{i=1}^{n} |x_i - c|.$$

More precisely, if a number $M$ is the sample median of $x_1, \cdots, x_n$, then

$$\sum_{i=1}^{n} |x_i - M| \leq \sum_{i=1}^{n} |x_i - c|$$

for every number $c$.

**Proof in two stages:** We shall first do a proof of this proposition in a very special case and then do a formal proof. You should read this very slowly. You should not go on to the next sentence until the sentence you are reading makes complete sense. We shall from time to time refer to the absolute value of the difference of two numbers as the distance between the two numbers. Thus the distance between $-2$ and 13 is $|-2-13| = |13-(-2)| = 15$. We shall refer to numbers sometimes as points and shall visualize them as points on a line.

Suppose at first that the data consist of just three numbers, say,

$$x_1 = 40,\ x_2 = 10 \text{ and } x_3 = 30.$$

The order statistics for this data set are

$$x_{(1)} = 10,\ x_{(2)} = 30 \text{ and } x_{(3)} = 40.$$

If you reread the proposition above, you note that we are trying to prove that the sample median of this data set, $M = 30$, is a value of a number $c$ that minimizes the sum of the distances from $c$ to each of these three numbers. Let us see why this conclusion is true in this example. If we take $c = 30$, the sum of its distances to the three order statistics is $20 + 0 + 10 = 40$. Now let us take any number $c$ less than 30 but equal to or greater than 10. We have decreased the former distance to 10 by the positive amount $30 - c$, but we have increased each of the distances from $c$ to 30 and from $c$ to 40 by the amount $30 - c$. Thus the sum of the distances from $c$ to the three points has increased by the amount $30 - c$. Now suppose one considers the value of $c = 10$. The sum of the distances from $c = 10$ to the three numbers is $0 + 20 + 30 = 50$. Now suppose one moves the point $c$ to the left of 10. In this case the sum of the distances from $c$ to the three points is the following sum of sums of positive numbers

$$(10 - c) + (20 + (10 - c)) + (30 + (10 - c)),$$

which has increased the previous sum of the distances by $3(10-c)$. Thus we see that if we start with $c = 30$ and let $c$ equal to any value smaller, we have only increased the sum of the distances from $c$ to the points of the data set.

You can show that if you start the value of $c$ again at 30 and move it to the right, until you reach the number 40, you are subtracting the positive number $c - 30$ from the distance from 30 to 40 , but you are adding twice that positive number to each of the distances from $c$ to 10 and 30. Thus it follows that only when $c$ is at the middle number, in other words, the sample median of these three numbers, only then the sum of the distances to the three points is at its smallest.

If you were to try this with four numbers, you would discover that any number between the second and third order statistics makes the sum of these distances smallest. Thus the sample median of a set of four numbers minimizes the sum of distances from itself to the four numbers of the data set.

Here is how one would do this proof in general, even when the $n$ numbers are not necessarily distinct. There are three cases, namely, when $n$ is odd, when $n$ is even (and thus $n = 2m$ for some positive integer $m$) but $x_{(m)} = x_{(m+1)}$, and when $n$ is even but $x_{(m)} < x_{(m+1)}$ where again $n = 2m$. Consider first the case when $n$ is odd. In this case there is a positive integer $m$ such that $n = 2m+1$. In this case, the sample median is by definition $x_{(m+1)}$. If we were to consider any number $c < x_{(m+1)}$, we see that there are at most $m$ numbers of the data set that are less than $c$, but there are at least $m + 1$ numbers greater than $c$. Thus the sum of the distances from $c$ to all of the numbers in the set has increased over the sum of the distances from the sample median by an amount that is as large or larger than $(m+1)(x_{(m+1)}-c)-m(x_{(m+1)}-c) = x_{(m+1)}-c > 0$. The same is true if one takes the value of $c$ to be greater than $x_{(m+1)}$, except in this case the sum of the distances is increased by at least $c - x_{(m+1)} > 0$. Now consider the case when $n$ is an even number, i.e., when there is a positive integer $m$ that satisfies $n = 2m$. If $x_{(m)} = x_{(m+1)}$, then the sample median is this common value. If one were to take $c$ as a number less than $x_{(m)} = x_{(m+1)}$, then the sum of the distances from $c$ to each number less than it would be decreased at most by $(m-1)(x_{(m)} - c)$. However the sum of the distances from $c$ to the numbers greater than it would be increased by at least $(m + 1)(x_{(m)} - c) > 0$, and thus the sum of the distances from $c$ to each of the $n$ numbers increases by at least $2(x_{(m)} - c) > 0$. And again, taking $c > x_{(m+1)}$ in this case, the sum of the distances from $c$ to all numbers greater than it is at most $(m-1)(c - x_{(m)})$ while the sum of the distances from $c$ to the numbers less than it has increased by an amount that is at least

$(m+1)(c-x_{(m)})$. Thus the sum of the distances from $c$ to all of the numbers of the set have increased by an amount not less than $2(c-x_{(m)}) > 0$. In the third case, if $c$ remains anywhere $x_{(m)}$ and $x_{(m+1)}$, then it is easily seen that the sum of the distances from $c$ to the $n$ numbers remains constant, But if $c$ is is decreased from $x_{(m)}$ to any number less than $x_{(m)}$, then the sum of the distances from $c$ to the numbers less than it has decreased by an amount that is at most $(m-1)(x_{(m)}-c)$ while the sum of the distances from $c$ to the numbers that are larger than it has increased by an amount that is at least $(m+1)(x_{(m)}-c)$, and thus the sum of the distances has increased by at least $2(x_{(m)}-c)$. Similarly, if $c$ is is increased from $x_{(m+1)}$ to any number greater than $x_{(m+1)}$, then the sum of the distances from $c$ to the numbers greater than it has decreased by an amount that is at most $(m-1)(c-x_{(m+1)})$ while the sum of the distances from $c$ to the numbers that are smaller than it has increased by an amount that is at least $(m+1)(x_{(m)}-c)$, and thus the total distance has increased by the amount $2(c-x_{(m+1)})$. Thus we see that in every case, as soon as we move the number $c$ away from the value of the sample median, the sum of the distances from $c$ to all of the numbers in the data set can only increase. This completes the formal proof of this proposition.

**Exercises**

1. Consider the following set of data:

3.2, 4.5, 1.3,4.5, 4.5, 3.7, 4.8, 5.6, 5.3.

($i$) What are the numbers $x_i, 1 \leq i \leq 9$?

($ii$) Find the sample mean and sample variance of these data.

($iii$) Find the sum of squares of errors of the data about the sample mean.

($iv$) Find the sum of squares of errors of the data about the number $c = 5.1$.

($v$) Find the sum of squares of errors of the data about the number $c = 3.6$.

2. Consider the same set of data as given in problem 1 :

3.2, 4.5, 1.3, 4.5, 4.5, 3.7, 4.8, 5.6, 5.3.

($i$) Evaluate the order statistics $x_{(i)}, 1 \leq i \leq 9$.

($ii$) Find the sample median of these data.

($iii$) Find the sum of the absolute values of the errors of the data about the sample median.

($iv$) Find the sum of the absolute values of the errors of the data about the number $c = 5.1$.

($v$) Find the sum of the absolute values of the errors of the data about the number $c = 3.6$.

3. Consider this data set:

$$3.2,\ 4.5,\ 1.3,\ 4.6,\ 4.2,\ 3.7,\ 4.8,\ 5.6,\ 5.3,\ 5.4.$$

($i$) Evaluate the order statistics $x_{(i)}, 1 \leq i \leq 10$.

($ii$) Find the sample median of these data.

($iii$) Find the sum of the absolute values of the errors of the data about the sample median.

($iv$) Find the sum of the absolute values of the errors of the data about the number $c = 4.5$.

($v$) Find the sum of the absolute values of the errors of the data about the number $c = 4.6$.

($iv$) Find the sum of the absolute values of the errors of the data about the number $c = 4.8$.

4. Compute both the sample mean and the sample median of the following set of data::

$$3.2, 4.1, 3.6, 236.1, 5.2, 4.3, 1.3, 5.1, 2.8.$$

Which of the two "centers" do you think represents the data in a more realistic manner?

5. We shall call a data set of even size $n = 2m$ symmetric if $x_{(m+1+i)} = x_{(m-i)}$ for $1 \leq i \leq m-1$. A data set of odd size $n = 2m+1$ is said to be symmetric if $x_{(m+1+i)} = x_{(m+1-i)}$ for $1 \leq i \leq m$. Prove that if $x_1, \cdots, x_n$ is any symmetric data set, then its sample mean and sample median are equal.

## Chapter 2. Events and Their Probabilities

**1. Events.** For the purposes of this course, we shall define a game or experiment or trial as an activity that consists of a number of equally likely individual outcomes. An event, $A$, is simply a defined subset or collection of some of these individual outcomes. For example, a game might be the outcomes of two tosses of an unbiased coin. There are four equally likely individual outcomes of such a game. They are $HH$, $HT$, $TH$, and $TT$. The individual outcome $HH$ denotes the outcome in which the coin comes up

heads on the first toss and heads on the second toss. The individual outcome $HT$ denotes the outcome in which the coin comes up heads on the first toss and comes up tails on the second toss. Similar interpretations can be given for elementary events $TH$ and $TT$. In playing this game one might define an event $A$ by the following: in two tosses of a coin the number of times that it comes up a head is 1. Thus the event $A$ consists of two individual outcomes: $HT$ and $TH$. In this case we might write $A = \{HT, TH\}$.

For any game, experiment or trial, two events, $A$ and $B$, might share some of the individual outcomes, or they might have none in common. For example consider the game where a fair die is tossed twice. The equally likely individual outcomes of such a game may be represented by 36 pairs of numbers between 1 and 6:

$$\{(1,1),1,2),\cdots,(6,5),(6,6)\}.$$

This set of individual outcomes can also by represented by

$$\{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}.$$

The number pair $(5,3)$ refers to the individual outcome in which the die comes up 5 on the first toss and comes up 3 on the second toss. Note that the equally likely individual outcomes $(5,3)$ and $(3,5)$ are different individual outcomes and each is as likely as $(3,3)$ or as $(5,5)$. We might consider an event $A$ in this game to be defined as this: the sum of the two numbers in an individual outcome is equal to or greater then 10. Then we might define an event $B$ by: the sum of the two numbers in an individual outcome is equal to or greater than 8 and is equal to or less than 10. A third event $C$ might be defined to be the set of equally likely outcomes for which each number in the pair is equal to or less than 3. Then we may exhibit these events by

$$A = \{(4,6),(5,6),(6,6),(5,5),(5,6),(6,4),(6,5),(6,6)\},$$

$$B = \{(6,2),(6,3),(6,4),(5,3),(5,4),(5,5),(4,4),(4,5),(4,6),(3,5),(3,6),(2,6)\}$$

and

$$C = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}.$$

Note that $A$ and $B$ have elements in common; they are

$$\{(6,4),(5,5),(4,6)\}.$$

However, events $A$ and $C$ have no individual outcomes in common. For this reason these two events, $A$ and $C$ are called **disjoint.**

**Definition:** Two events are said to be disjoint if there is no individual outcome that is in both of them. Three or more events are also called disjoint if the events of every pair of them are disjoint.

If you encounter two events, call them $C$ and $D$, and if you wish to prove that they are disjoint, you might take any individual outcome $\omega$ in $C$ and prove that it is not in $D$. Then no individual outcome is in both.

There is some notation that will be used from time to time. If $A$ and $B$ are events, then $A \cup B$ will denote the event that at least one of the two events, $A$, $B$ occurs. This means: upon playing the game, if an individual outcome occurs that is in at least one of these two events, then we say that the event $A \cup B$ occurs. (Note the expression "at least".) This means $A \cup B$ is defined to be an event composed of those individual outcomes, each of which is in at least one of the events $A, B$. For example, consider the event that a fair coin is tossed 3 times. The collection of the 8 equally likely individual outcomes for this game are as follows:

$$HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.$$

The individual outcome $TTH$ is the individual outcome in which tails comes up on the first toss, tails comes up on the second toss, and heads comes up on the third toss. Suppose the event $A$ is defined as the collection of individual outcomes in which heads occurs exactly once. Then

$$A = \{HTT, THT, TTH\}.$$

Now suppose we consider an event $B$ defined by: heads occurs at least once but not before the second toss. Then

$$B = \{THT, TTH, THH\}.$$

Then by the definition given above,

$$A \cup B = \{HTT, THT, TTH, THH\}.$$

The Greek letter $\Omega$ will denote the set of all equally likely individual outcomes of a game. It is called "the sure event". An individual outcome in $\Omega$ will be denoted by $\omega$. The expression "$\omega$ is an individual outcome in the

event $B$", will be written as: $\omega \in B$. For example, if the game is to toss a fair coin three times, then for this game,

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

In the example given above, since the individual outcome $HTT$ is one of the individual outcomes in the event $A$, we write $HTT \in A$. A general way of defining $A \cup B$ is

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}.$$

(N.B. The "or" used in this definition is the logical "or", and it means that at least one of the properties $\omega \in A$, $\omega \in B$ is true.) This displayed equation reads as follows: $A \cup B$ is (or equals) the set of those individual outcomes $\omega$ in $\Omega$ such that $\omega \in A$ or $\omega \in B$.

If an individual outcome, $\omega$, is not in an event $A$, we shall denote this by $\omega \notin A$.

If $A$ is an event, its negation, denoted by $A^c$, is defined to be the event consisting of those individual outcomes that are not in $A$. In precise notation, $A^c = \{\omega \in \Omega : \omega \notin A\}$. Thus, if the event $A$ does not occur, then the event $A^c$ occurs. It should be noted that if $A$ is an event, then $A$ and $A^c$ are disjoint. Further, if $\Omega$ denotes the set of all individual outcomes of some game, and if $A$ is an event, then $\Omega = A \cup A^c$.

Again, let $\Omega$ denote the set of all equally likely individual outcomes of some game. If $A$ and $B$ are possible events for this game, and if every individual outcome in $A$ is also an individual outcome of $B$, then we shall write $A \subset B$, and so that if the event $A$ occurs, then event $B$ occurs. If you encounter a problem where you wish to prove that for two events, $A$ and $B$, $A \subset B$, you should take an arbitrary individual outcome $\omega$ in $A$ and show that it is also an individual outcome in $B$. If $A \subset B$ and if $B \subset A$, then we write $A = B$.

Finally, if $A$ and $B$ are events, then $A \cap B$ will denote the event that both $A$ and $B$ occur; in other words, $A \cap B = \{\omega \in \Omega : \omega \in A \textbf{ and } \omega \in B\}$. We readily see that two events $A$ and $B$ are disjoint if and only if $A \cap B$ contains no individual outcomes. In this case, we write $A \cap B = \emptyset$, where $\emptyset$ denotes the empty set or impossible event or the set that contains no individual outcomes.

**Proposition 1.** If $A$ and $B$ are events, then $A \cap B \subset A \subset A \cup B$.

Proof: We first wish to prove that $A \cap B \subset A$. As indicated above, let us take any individual outcome $\omega$ in $A \cap B$; we must show that $\omega \in A$. Since by hypothesis, $\omega \in A \cap B$, then by the above definition, we know that $\omega$ is in $A$ and $\omega$ is in $B$. This certainly implies that $\omega \in A$, which proves $A \cap B \subset A$. Next we must prove that $A \subset A \cup B$. So take any $\omega$ in $A$; we must prove that $\omega \in A \cup B$. Since by hypothesis, $\omega$ is in $A$, then it is in at least one of the events $A, B$, namely, in $A$. Thus we have proved $A \subset A \cup B$. Since we have proved that both $A \cap B \subset A$ and $A \subset A \cup B$ are true, we have thus proved that $A \cap B \subset A \subset A \cup B$, and the proposition is proved.

## Exercises

**1**. Consider a game in which a fair coin is tossed four times. One typical outcome could be represented by $HTHT$, which could denote that heads occurred on the first and third tosses and tails occurred on the second and fourth tosses.

($i$) How many equally likely outcomes are there in this game?

($ii$) Make a list of all the equally likely outcomes for this game.

($iii$) Let $A$ denote the event that the coin comes up heads on the third toss. How many individual outcomes are in this event?

($iv$) Let $B$ denote the event that the coin comes up tails in the fourth toss. Make a list of all the individual outcomes in $B$.

($v$) Let $C$ denote the event that the second time that the coin comes up $H$ is on the fourth toss. Make a list of the individual outcomes in $C$.

($vi$) Make a list of the individual outcomes in $C^c$.

**2**. Suppose a bowl contains 5 tags, numbered $1, 2, 3, 4, 5$. The tags are mixed up thoroughly and, without looking, two tags are drawn from the bowl at random, leaving the bowl with only three tags in it. One of the equally likely outcomes is that the two numbers $\{3, 5\}$ are drawn.

($i$) There are nine more pairs of numbers that are possible. List them.

($ii$) Let $A$ denote the event that the sum of the two numbers is an even number. List all the equally likely outcomes in the event $A$.

($iii$) Let $B$ denote the event that a larger of the two numbers is at most 4. List all the equally likely outcomes in the event $B$.

($iii$) List all of the equally likely outcomes in the following events: $A \cup B$, $A \cap B$, $A^c$ and $B^c$.

3. A bowl contains 10 tags numbered 1 through 10. The first five tags are colored red, and the remaining tags are some other color. One selects a

tag at random. Let $A$ denote the event that he selects a red tag, and let $B$ denote the event that he selects in this same selection an even numbered tag.

($i$) List the numbers in each of these events: $A$, $B$, $A \cap B$ and $A \cap B^c$.

($ii$) Verify in this case that $A = (A \cap B) \cup (A \cap B^c)$.

4. Prove: if $A$ and $B$ are events, then $A = (A \cap B) \cup (A \cap B^c)$.

5. Prove: if $A$ and $B$ are events, then $A \cup B = A \cup (A^c \cap B)$.

6. A die is tossed 10 times. Let $A_i$ denote the event that it comes up 1 on the $i$th trial. Match the following events with the statements that follow them. The events are:

1. $\cup_{j=1}^{3} A_j$
2. $A_4 \cap \bigcap_{i=1}^{3} A_i^c$
3. $A_4 \cap \bigcap_{j=5}^{10} A_j^c$
4. $(A_1 \cap A_2^c \cap A_3^c) \cup (A_1^c \cap A_2 \cap A_3^c) \cup (A_1^c \cap A_2^c \cap A_3)$

The statements are:

($i$) The last time that the die comes up 1 is on the 4th toss.

($ii$) The first time in which the die comes up 1 is on the fourth toss.

($iii$) The die comes up 1 in at least one of the first three tosses.

($iv$) The die comes up 1 exactly once during the first three tosses.

**2. Counting.** We shall need to know how to determine the number of individual outcomes in some of the events that we shall encounter. We now develop some techniques that will be of help.

**Definition.** If $n$ is a positive integer, we define $n!$ (pronounced: n factorial) by

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1.$$

For example, $1! = 1$, $2! = 2 \times 1 = 2$, $3! = 3 \times 2 \times 1 = 6$, and so on. As $n$ gets large, $n!$ gets larger much faster. It will be convenient to define $0! = 1$. It should be noticed that whenever $n \geq 1$, then

$$n \times (n-1)! = n! \text{ or } \frac{n!}{n} = (n-1)!.$$

Factorials are used to compute the number of arrangements, or permutations, for arranging $n$ distinct objects in a line. If there are $n$ distinct or different objects, and if one wished to determine the number of ways they can be arranged in order in a line, a person might reason as follows. There

are $n$ ways in which the first object for the first place in the line can be selected. For each way this first object is selected, there are $n-1$ ways in which the second object may be selected from the remaining $n-1$ objects for the second place in the line. Thus there are $n \times (n-1)$ ways in which the first two objects can be selected. Then for each way in which the first two objects can be selected, there are $n-2$ ways of selecting the third object from the remaining $n-2$ objects. Thus there are $n \times (n-1) \times (n-2)$ ways of selecting the first three objects. Note from the numbering used so far that there are

$$n \times (n-1) \times \cdots \times (n-(k-1))$$

ways of selecting the first $k$ objects, where $1 \leq k \leq n$. It should be noted that this is the same as the following number:

$$\frac{n!}{(n-k)!}.$$

We call this the number of arrangements, or permutations, of $n$ objects taken $k$ at a time. Also, letting $k = n$, the equality of the above two expressions looks like this:

$$n \times (n-1) \times \cdots \times 2 \times 1 = \frac{n!}{0!},$$

which is one explaination why it is convenient to take $0! = 1$. Each such arrangement is called a **permutation.**

Associated with the concept of a permutation is that of the number of combinations of $n$ distinct objects taken $k$ at a time. In a combination, one is not interested in the order in which things happen but only in which objects have been selected. For example, if you are asked for the number of permutations of $n$ things taken $k$ at a time, you have from the above that this is

$$\frac{n!}{(n-k)!}.$$

However, this number is way too high; it gives the number of rearrangements for every subset or combination of $k$ objects. This number of rearrangements is $k!$. So we must divide the above number by $k!$ to get

$$\frac{n!}{k!(n-k)!}.$$

The number you get from this is what is referred to as a **binomial coefficient**, and it is denoted as follows:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

This binomial coefficient is usually referred to as "$n$ choose $k$". Here is a simple consequence of the above definition.

**Proposition.** If $0 \leq k \leq n$, then

$$\binom{n}{k} = \binom{n}{n-k}.$$

Proof: If one uses the definition of $\binom{n}{k}$ on both sides of the equation, the conclusion is immediate.$\binom{n}{k}$

This leads us to the binomial theorem which is basic to this course.

**The Binomial Theorem**. For any numbers $a$ and $b$ and for every positive integer, $n$,

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}.$$

Proof: We may write

$$(a+b)^n = (a+b)(a+b)\cdots(a+b),$$

where $(a+b)$ is laid out $n$ times on the right hand side. In order to multiply the right hand side, we take one of the numbers $a, b$ out of each of the $n$ terms, multiply them and continue doing this until we have made all possible $2^n$ selections. (Note that there are $2^n$ ways in which one can pick one summand out of each of the $n$ sums.) Let us consider the selections from which we get $k$ $a$'s and thus $n-k$ $b$'s. These are $\binom{n}{k}$ in number. So the product $a^k b^{n-k}$ will be obtained $\binom{n}{k}$ times. Thus, to make a long story short,

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j},$$

which proves the theorem.

## Exercises

1. In how many ways can 6 boys be arranged in a line?
2. In how many ways can 6 boys be arranged in a circle?

3. In how many ways can a lookout committee of 3 boys be appointed out of a collection of 6 boys?

4. Suppose 6 boys decide to form a club. In how many ways can a president, a vice president and a secretary be elected?

5. Evaluate $\binom{n}{k}$, when $n = 12$ and $k = 10$. (Do not leave this as a fraction. Obseve how much cancellation you can do.)

6. Evaluate $\binom{9}{3}$ and $\binom{9}{6}$. (Did you notice anything?)

7. Write a proof of the proposition: If $0 \leq k \leq n$, then $\binom{n}{k} = \binom{n}{n-k}$.

8. Prove the binomial theorem by using mathematical induction.

9. A coin is tossed $n$ times, during which it comes up as a head $k$ of those times. In how many different patterns of the $n$ times can this occur?

10. In how many ways can a group of 8 children form a circle if no pair of Mary, Betty and Alice can be next to each other?

**3. Probability.** In this course we shall only need the definition of probability in the case of a game with a finite number of equally likely individual outcomes.

**Definition.** If $A$ is an event that can occur in a game or experiment, if the number of equally likely individual outcomes of the game or experiment is $N$, and if $S$ is the number of equally likely individual outcomes in which the event $A$ can occur, then we define the probability of $A$, $P(A)$, by

$$P(A) = \frac{S}{N}.$$

For example, in the game of tossing a fair coin 4 times, the number of individual outcomes is $2^4 = 16$, so $N = 16$. If $A$ is the event that the coin comes up heads at most once, it is easy to see that $A$ contains 5 individual outcomes in it, so $S = 5$. Using the definition, the probability that a fair coin comes up heads at most once is $\frac{5}{16}$.

**Proposition** 1. If $\Omega$ denotes the set of all possible equally likely individual outcomes of some game or experiment, then $P(\Omega) = 1$.

Proof: In this case, if $N$ denotes the number of equally likely outcomes of the game, and if $A = \Omega$, then $S = N$, so $P(\Omega) = \frac{N}{N} = 1$.

**Proposition 2.** If $A$ is an event, then $0 \leq P(A) \leq 1$.

Proof: Since $0 \leq S \leq N$, we have

$$\frac{0}{N} \leq \frac{S}{N} \leq \frac{N}{N},$$

from which we obtain that $0 \leq P(A) \leq 1$.

**Proposition** 3. If $A$ and $B$ are events for some game, and if $A \subset B$, then $P(A) \leq P(B)$.

Proof. Since by hypothesis $A \subset B$, this implies that the number, $S_A$, of individual outcomes in $A$ is equal to or less than the number, $S_B$, of individual outcomes in $B$. Thus, if $N$ denotes the total number of individual outcomes of the game,

$$\frac{S_A}{N} \leq \frac{S_B}{N}, \text{ or } P(A) \leq P(B).$$

**Proposition** 4.: If $A$ and $B$ are disjoint events for some game or experiment, then $P(A \cup B) = P(A) + P(B)$.

Proof:: Using the notation above, since $A$ and $B$ are disjoint events it follows that $S_{A \cup B} = S_A + S_B$. Thus

$$P(A \cup B) = \frac{S_{A \cup B}}{N} = \frac{S_A}{N} + \frac{S_B}{N} = P(A) + P(B).$$

**Corollary to Proposition** 4. **If** $A$ is an event, then $P(A^c) = 1 - P(A)$.

Proof: Since $A$ and $A^c$ are disjoint, since $\Omega = A \cup A^c$, since $P(\Omega) = 1$ and since by proposition 4, $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$, it follows that $P(A^c) = 1 - P(A)$.

**Definition.** Events $A_1, A_2, \cdots, A_n$ are said to be independent if for all integers $r$ and $k_1 < k_2 < \cdots < k_r$ that satisfy $2 \leq r \leq n$ and $1 \leq k_1 < k_2 < \cdots < k_r \leq n$, the following is true:

$$P(A_{k_1} \cap \cdots \cap A_{k_r}) = P(A_{k_1}) \cdots P(A_{k_r}),$$

or,

$$P\left(\bigcap_{j=1}^{r} A_{k_j}\right) = \prod_{j=1}^{r} P(A_{k_j}).$$

As an example, suppose a fair coin is tossed 4 times. Let $A_i$ denote the event that the coin comes up heads on the $i$th toss, $1 \leq i \leq 4$. There are 16 equally likely individual outcomes for the game. The event $A_2$ occurs in 8 equally likely individualk outcomes (list them!), the event $A_4$ occurs in 8 equally likely individual outcomes (list them!), and the event $A_2 \cap A_4$ occurs in 4 individual outcomes (verify this!). Thus

$$P(A_2 \cap A_4) = \frac{4}{16} = \frac{8}{16}\frac{8}{16} = P(A_2)P(A_4).$$

This is but one of the 11 equations that these events satisfy, and thus they are independent.

**Proposition** 5: For every positive integer $n$,

$$\sum_{j=0}^{n} \binom{n}{j} = 2^n.$$

Proof: This follows from the fact that $1 + 1 = 2$ and from the binomial theorem given above.

**Proposition** 6. The number of equations that events $A_1, A_2, \cdots, A_n$ must satisfy in order that they be independent is $2^n - n - 1$.

Proof: Clearly the number of equations that must be satisfied is equal to the number of ways in which 2 objects out of $n$ objects can be selected plus the number of ways that 3 objects out of $n$ objects can be selected plus $\cdots$ plus the number of ways $n$ objects can be selected out of $n$ objects, which turns out, using the previous proposition, to be

$$\sum_{j=2}^{n} \binom{n}{j} = \sum_{j=0}^{n} \binom{n}{j} - n - 1 = 2^n - n - 1.$$

**Proposition** 7**.** If events $A$ and $B$ are independent, then so are the events $A$ and $B^c$.

Proof: First note that $A = (A \cap B) \cup (A \cap B^c)$. Since the right side is a disjoint union, we have $P(A) = P(A \cap B) + P(A \cap B^c)$. Thus by independence and the fact that $P(B) + P(B^c) = 1$, we have

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

**Definition.**. If $A$ and $B$ are events in some game, and if $P(B) > 0$, we define the conditional probability of $A$ given the event $B$, $P(A|B)$, by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

**Proposition** 8. If $A$ and $B$ are events, and if $P(B) > 0$ and $P(B^c) > 0$, then $A$ and $B$ are independent events if and only if $P(A|B) = P(A)$, which is true if and only if $P(A|B^c) = P(A)$.

Proof: This follows from the definition of conditional probability and Proposition 7.

**Remark**. Independent events generally occur when a game is played several times under the same conditions, or, more generally, when several games are played, where no outcome of any game influences the outcome of any of the others. In other words, one looks at several plays of a game as a game itself. For example, suppose three games $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are played, and suppose that $n_1$ is the number of equally likely outcomes from game $\mathcal{A}$, suppose that $n_2$ is the number of equally likely outcomes of game $\mathcal{B}$, and suppose that $n_3$ is the number of equally likely outcomes for game $\mathcal{C}$. Further, suppose that $A$ is an event that can occur in one of $a$ of the $n_1$ equally likely outcomes when $\mathcal{A}$ is played, suppose that $B$ is an event that can occur in one of $b$ of the $n_2$ equally likely outcomes when $\mathcal{B}$ is played, and suppose that $C$ is an event that can occur in one of $c$ of the $n_3$ equally likely outcomes when $\mathcal{C}$ is played. Then the probability of $A$ and $B$ and $C$ occurring is

$$P(A \cap B \cap C) = \frac{abc}{n_1 n_2 n_3} = P(A)P(B)P(C).$$

Also, the probability of $A$ occurring in the first game and $C$ occurring in the third game (the outcome of the second game is left unsaid) is

$$P(A \cap C) = \frac{an_2c}{n_1 n_2 n_3} = P(A)P(C).$$

Also, in the same way,

$$P(A \cap B) = P(A)P(B) \text{ and } P(B \cap C) = P(B)P(C).$$

Thus, all the $2^3 - 3 - 1$ equations for independence are satisfied.

## Exercises

1. An unbiased coin is tossed 4 times. Evaluate:

($i$) The number of equally likely outcomes of the game.

($i$) $P(A)$, where $A$ is the event that the number of times the coin comes up heads is 2.

($ii$) $P(B)$, where $B$ is the event that that the coin comes up heads at most two times.

($iii$) $P(C)$, where $C$ is the event that the coin comes up heads at least 3 times.

2. A box contains 4 red balls and 2 white balls. A trial consists of mixing up the balls, selecting a ball at random, noting its color and then returning it to the box. So:

($i$) What is the probability of obtaining a red ball in one trial?

($ii$) For two trials, find the probability that a red ball is selected on the first trial and a white ball is selected on the second trial.

($iii$) For two trials, find the probability that a white ball is selected in the first trial and a red ball is selected in the second trial.

($iv$) For two trials, find that probability that the number of red balls selected is 1.

($v$) For three trials, find the probability that 2 red balls are selected.

($vi$) For three trials, find the probability that the number of red balls selected is not greater than 2.

3. Twenty numbered tags are in a hat. The number 1 is on seven of the tags, the number 2 is on five of the tags and the number 3 is on eight of the tags. The experiment is to stir the tags without looking at them and to select one tag "at random".

($i$) What is the total number of equally likely individual outcomes of the experiment?

($ii$) From among these twenty equally likely individual outcomes what is the total number of equally likely ways in which the outcome is the number 1?

($iii$) From among the total number of equally likely outcomes of the experiment, what is the total number of equally likely outcomes in which one draws the number 3?

($iv$) Compute the probability of selecting a tag numbered 1 or 3.

($v$) What is the sum of the probabilities of the outcomes obtained in ($ii$) and ($iii$)?

4. Prove: if each of two events, $A$ and $B$, has a positive probability, and if $A$ and $B$ are independent events, then they are not disjoint. (I do not know why it is, but in my experience many students confuse independence with disjointness. Do not ever make this mistake.)

5. Prove: if events $A$ and $B$ are independent, then so are the events $A^c$ and $B^c$.

6. Use the binomial theorem and the fact that for every real number $p$, $p + (1 - p) = 1$, to prove that

$$\sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} = 1.$$

7. Prove: If $A$, $B$ and $C$ are independent events, then so are $A$ and $B \cap C$.

8. Prove: If $A$, $B$ and $C$ are independent events, then so are $A$ and $B \cup C$.

9. In the example following the first definition given in this section, list all of the individual outcomes in the event $A$.

## Chapter 3. Random Variables.

**1. Random Variables and Their Distributions.** The data that are obtained in situations where a statistical analysis is called for are obtained from observations on something called random variables. In brief, a random variable is the mechanism by which data are produced in a given game or experiment. We shall be more explicit shortly. But in the meantime, let us consider a game in which an event $A$ whose probability is $p$ can occur. Now suppose that one decides to play this game $n$ times. Let $X$ denote the number of times that the event $A$ actually occurs in $n$ plays. If $p \neq 0$ and $p \neq 1$, then $X$ could possibly be any value between 0 and $n$. This chance numerical outcome is but one example of what is called a random variable. If the value that $X$ equals or takes is $k$, where $k$ is some integer between 0 and $n$, then we say that the event $[X = k]$ has occurred. This event has a probability which we now determine.

**Theorem 1.** If, in each of $n$ independent plays of a game, the probability of a certain event $A$ occurring in a play of the game is $p$, then the probability that $A$ occurs $k$ times during the $n$ plays is given by

$$P([X = k]) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } 0 \leq k \leq n.$$

Proof: Suppose that the game has $r$ equally likely outcomes, and that $A$ occurs in $s$ of these. Thus the probability of $A$ occurring in a particular game is $p = \frac{s}{r}$. The probability of the event that $A$ occurs in all of the first $k$ plays of the game and not in any of the last $n - k$ plays of the game is

$$\frac{s^k}{r^k} \frac{(r-s)^{n-k}}{r^{n-k}} = p^k (1-p)^{n-k}.$$

But this is the same probability as that of $A$ occurring in just any one particular combination (or pattern) of $k$ out of the $n$ plays. There are $\binom{n}{k}$ of such combinations or patterns. Thus we obtain the conclusion.

This collection of $n + 1$ probabilities, $\left\{ \binom{n}{k} p^k (1-p)^{n-k}, 0 \leq k \leq n \right\}$ is called the **Binomial Distribution**. By the binomial theorem, they add up to

1. When a random numerical outcome, $X$, has this particular distribution of probabilities, we say that $X$ has the binomial distribution, and we sometimes write: $X$ is $Bin(n,p)$ or $X$ has the $Bin(n,p)$ distribution.

For example, suppose the game is a roll of a fair die, and the event, $A$, of concern is: the outcome is a 1 or 2. Thus, $P(A) = \frac{2}{6} = \frac{1}{3}$. Next suppose the die is tossed 4 times, and suppose that $X$ denotes the number of times that the event $A$ occurs. The value of $X$ can be any number from 0 to 4. Thus, the probability that the event $A$ occurs 2 times in the 4 tosses of the die is

$$P([X=2]) = \binom{4}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{4-2} = \frac{4\cdot 3}{1\cdot 2}\cdot\frac{1}{9}\cdot\frac{4}{9} = \frac{8}{27}.$$

For relatively small values of $n$, values of these probabilities can be obtained on any desktop computer in which EXCEL is installed. Go to EXCEL, click on $f_x$, then in the Function Category, click on Statistical, and in the Function Name, click on BINOMDIST. The rest you should be able to figure out for yourself.

We now are able to present a more rigorous definition of a random variable.

**Definition 1**: A random variable $X$ defined relative to a game or experiment is a function whose domain is the set of all equally likely individual outcomes of that game or experiment. Expressed otherwise, a random variable is a numerical outcome of a game. It assigns a number to each equally likely individual outcome.

As an example, consider a game where an unbiased coin is tossed 3 times. The eight equally likely individual outcomes of this game are

$$HHH, HHT, HTH, HTT, THH, THT, TTH, TTT.$$

A random variable $X$ that one might consider is "the number of heads in a play of the game", i.e., in three tosses of the coin. Thus $X$ assigns the number 3 to the individual outcome $HHH$, it assigns the number 2 to the individual outcome $HHT$, and it assigns the number 0 to the individual outcome $TTT$. This can be spelled out completely as follows:

$$X(HHH) = 3,\ X(HHT) = 2,\ X(HTH) = 2,\ X(HTT) = 1$$
$$X(THH) = 2,\ X(THT) = 1,\ X(TTH) = 1,\ X(TTT) = 0.$$

**Definition 2.** The range of a random variable $X$, denoted by $range(X)$, is the set of all distinct values that the random variable takes or is equal to,

i.e., $range(X) = \{X(\omega) : \omega \in \Omega\}$, where $\Omega$ is the set of all equally likely outcomes. We shall consider only random variables whose ranges are of finite size.

In the example given above, the range of the random variable, $X$, is $\{0, 1, 2, 3\}$, and this is written as

$$range(X) = \{0, 1, 2, 3\}.$$

Be careful to note that the numbers in the range of a random variable are not necessarily equally likely.

**Definition 3.** If $X$ is a random variable with range $S$, then for each $x \in S$, we denote $[X = x]$ to be the set of all individual outcomes $\omega$ in $\Omega$ for which the value of $X(\omega)$ is $x$, and we write this as

$$[X = x] = \{\omega \in \Omega : X(\omega) = x\}.$$

In this last example, this means

$$\begin{array}{l} [X = 0] = \{(TTT)\}, \\ [X = 1] = \{(HTT), (THT), (TTH)\}, \\ [X = 2] = \{(HHT), (HTH), (THH)\}, \text{ and} \\ [X = 3] = \{(HHH)\}. \end{array}$$

**Theorem 2**: If $X$ is a random variable, then the events $\{[X = x] : x \in range(X)\}$ are disjoint and their probabilities add up to 1, i.e.,

$$\sum\{P([X = x] : x \in range(X)\} = 1.$$

Proof: Suppose $x_1$ and $x_2$ are two distinct numbers in $range(X)$. If $\omega \in [X = x_1]$, then $X(\omega) = x_1$. Since $x_1 \neq x_2$, then $\omega \notin [X = x_2]$. Thus every pair of events, $[X = x_1]$, $[X = x_2]$, are disjoint. In order to prove the second conclusion, let $n(x)$ denote the number of $\omega$'s in $\Omega$ such that $X(\omega) = x$, and let $n$ denote the total number of individual outcomes in $\Omega$. Then $P([X = x]) = \frac{n(x)}{n}$. But $\sum\{n(x) : x \in range(X)\} = n$, so

$$\begin{array}{ll} \sum\{P([X = x] : x \in range(X)\} & = \sum\{\frac{n(x)}{n} : x \in range(X)\} \\ & = \frac{1}{n}\sum\{n(x) : x \in range(X)\} \\ & = 1, \end{array}$$

which proves the theorem.

Here is another example of a game and random variables. Suppose you have an urn with 4 tags in it numbered $1, 2, 3, 4$. The game is to select 2 tags at random from the urn without replacement. In this case the set of equally likely outcomes may be represented as

$$\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}.$$

The probability of each of these outcomes is $\frac{1}{6}$. We shall define two random variables, $X$ and $Y$, with respect to this game. The random variable $X$ assigns to each individual outcome the sum of the two numbers, and the random variable $Y$ assigns to each individual outcome the smaller of the two numbers. Thus $[X = 3] = \{(1,2)\}$, $[X = 4] = \{(1,3)\}$, $[X = 5] = \{(1,4),(2,3)\}$, $[X = 6] = \{(2,4)\}$ and $[X = 7] = \{(3,4)\}$, from which we compute $P([X = 3]) = \frac{1}{6}$, $P([X = 4]) = \frac{1}{6}$, $P([X = 5]) = \frac{2}{6} = \frac{1}{3}$, $P([X = 6]) = \frac{1}{6}$ and $P([X = 7]) = \frac{1}{6}$. Also $[Y = 1] = \{(1,2),(1,3),(1,4)\}$,

$[Y = 2] = \{(2,3),(2,4)\}$, and $[Y = 3] = \{(3,4)\}$, from which we compute $P([Y = 1]) = \frac{1}{2}, P([Y = 2]) = \frac{1}{3}$ and $P([Y = 3]) = \frac{1}{6}$.

**Exercises**

1. Suppose an urn contains 5 tags, numbered 1, 2, 3, 4 and 5. Three tags are drawn from it at random without replacement. The set of all individual outcomes of this game is the set of all subsets consisting of 3 of the numbers.

($i$) Make a list of all 10 of these equally likely individual outcomes.

($ii$) Let the random variable $Y$ denote the sum of three numbers selected at random without replacement from the urn. What value does $Y$ assign to each of the 10 individual outcomes?

($iv$) List the members of $range(Y)$.

($v$) Find $\{P([Y = y]), y \in range(Y)\}$.

($vi$) Think: Is the number of distinct members of $range(Y)$ the same as the number of the 10 equally likely outcomes obtained in part ($i$)?

($vii$) Let $Z$ denote the middle number among the three numbers selected. Find the range of $Z$ and the values of $\{P([Z = z]), z \in range(Z)\}$.

2. Prove that the sum of probabilities, $\left\{\binom{n}{k}p^k(1-p)^{n-k},\ 0 \le k \le n\right\}$, is equal to 1.

3. Prove: If $Z$ is $Bin(n,p)$, and if $0 \le k < n$, then $\Omega = [Z \le k] \cup [Z \ge k+1]$.

4. Prove: If $Z$ is $Bin(n,p)$, then $P([Z \geq k]) = 1 - P([Z \leq k-1])$ for $1 \leq k \leq n$.

**2. Expectation of a Random Variable.** The notion of expectation of a random variable, $X$, is the same as that of center of gravity, or centroid, of a spread of point masses along the $x$-axis, the point masses being the values of $P([X = x])$ at each point $x$ in the range of the random variable.

**Definition 1**: If $X$ is a random variable, we define its expectation, $E(X)$, by

$$E(X) = \sum\{xP([X = x]) : x \in range(X)\}.$$

As an example, consider an urn with 10 tags, in which 5 tags have the number 1 on each of them, 3 tags have the number 12 written on each of them, and 2 tags have the number 5 written on each of them. Let $X$ denote the number on a tag picked at random from the urn. Then

$$range(X) = \{1, 12, 5\}.$$

Since $P([X = 1]) = \frac{5}{10}$, $P([X = 12]) = \frac{3}{10}$ and $P([X = 5]) = \frac{2}{10}$, the expectation of $X$ is computed by

$$\begin{aligned} E(X) &= 1P([X = 1]) + 5P([X = 5]) + 12P([X = 12]) \\ &= 0.5 + 1.0 + 3.6 \\ &= 5.1. \end{aligned}$$

(In terms of mechanics, if you have a weightless rod of length 12, if you place a weight of amount 0.5 at a distance 1 from the left end, a weight of amount 0.2 at a distance 5 from the left end, and a weight of amount 0.3 at the right end, and if you then place your finger under the rod at a distance 5.1 from the left end of the rod and lift it, the weights will be in balance. (Recall the teeter-totter principle mentioned in chapter 1.)

**Theorem 1.** If $X$ is is a random variable with the $Bin(n,p)$ distribution, i.e., if

$$P([X = k]) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } 0 \leq k \leq n,$$

then $E(X) = np$.

Proof: Using the definition of expectation and also using the binomial

theorem,

$$\begin{aligned} E(X) &= \sum_{j=0}^{n} j\binom{n}{j}p^j(1-p)^{n-j} \\ &= \sum_{j=1}^{n} j\frac{n!}{j!(n-j)!}p^j(1-p)^{n-j} \\ &= np\sum_{j=1}^{n}\frac{(n-1)!}{(j-1)!((n-1)-(j-1))!}p^{j-1}(1-p)^{(n-1)-(j-1)} \\ &= np\sum_{j-1=0}^{n-1}\frac{(n-1)!}{(j-1)!((n-1)-(j-1))!}p^{j-1}(1-p)^{(n-1)-(j-1)} \\ &= np(p+(1-p))^{n-1} = np, \end{aligned}$$

which proves the theorem.

If $X$ is a random variable, then so is $X + c$ for any constant $c$, and

$$range(X + c) = \{x + c : x \in range(X)\}.$$

If $Z$ is defined by $Z = X + c$, then

$$P([Z = x + c]) = P([X + c = x + c]) = P([X = x]).$$

Also, if $Y$ is a random variable, then so is $Y^2$, and

$$range(Y^2) = \{y^2 : y \in range(Y)\}.$$

**Theorem 2.** If $X$ is a random variable, and if $c$ is any constant, then $E(X + c) = E(X) + c$.

Proof: First note that $range(X + c) = \{x + c : x \in range(X)\}$. By the definition of expectation and by theorem 2 above,

$$\begin{aligned} E(X+c) &= \sum_{x+c\in range(X+c)}(x+c)P([X+c=x+c]) \\ &= \sum_{x\in range(X)} xP([X=x]) + c\sum_{x\in range(X)} P([X=x]) \\ &= E(X)+c. \quad \text{Q.E.D.} \end{aligned}$$

**Theorem 3**. If $V$ is a random variable, and if $c$ is a constant, then $E(cV) = cE(V)$.

Proof: If $c = 0$, the the conclusion is immediate. If $c \neq 0$, then by the definition of expectation, we have

$$\begin{aligned} E(cV) &= \sum_{cv\in range(cV)} cvP([cV=cv]) \\ &= c\sum_{v\in range(V)} vP([X=v]) \\ &= cE(V). \text{ Q.E.D.} \end{aligned}$$

**Theorem 4**. If $V$ is a random variable, and if $c$ and $d$ are constants, then $E(cV + d) = cE(V) + d$.

Proof: This follows immediately from the previous two theorems.

**Definition 5**. If $X$ is a random variable, then we define the variance of $X$, $Var(X)$, by

$$Var(X) = E((X - E(X))^2).$$

**Theorem 5**. If $V$ is a random variable, and if $c$ and $d$ are constants, then $Var(cV + d) = c^2 Var(V)$.

Proof: Note that $range(cV + d) = \{cv + d : v \in range(V)\}$. Thus, by the previous proposition,

$$\begin{aligned} Var(cV + d) &= \sum\{((cv + d) - cE(V) - d)^2 P([V = v]) : v \in range(V)\} \\ &= c^2 \sum\{(v - E(V))^2 P([V = v]) : v \in range(V)\} \\ &= c^2 Var(V). \end{aligned}$$

It should be noted that $Var(cV + d)$ does not depend on the value of $d$. Thus, the variance of a random variable depends only on the "spread" of the range. If the value of $c$ is greater than 1 or less than $-1$, then the the numbers in $range\{cV + d\}$ are more "spread out" than the number in $range(V)$, and if $|c| < 1$, then the numbers in $range(V)$ are more "bunched up".

**Theorem 5**. If $X$ is a random variable whose distribution is $Bin(n, p)$, then

$$Var(X) = np(1 - p).$$

Proof: Since $E(X) = np$, and since the easily verified identity, $k^2 = k(k - 1) + k$, helps with the fifth line below, we have

$$\begin{aligned} Var(X) &= E((X - E(X))^2) = E((X - np)^2) \\ &= \textstyle\sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \textstyle\sum_{k=0}^{n} (k^2 - 2knp + n^2p^2) \binom{n}{k} p^k (1 - p)^{n-k} \\ &= \textstyle\sum_{k=1}^{n} k^2 \binom{n}{k} p^k (1 - p)^{n-k} - 2n^2p^2 + n^2p^2 \\ &= \textstyle\sum_{k=2}^{n} k(k - 1) \binom{n}{k} p^k (1 - p)^{n-k} + \sum_{k=1}^{n} k \binom{n}{k} p^k (1 - p)^{n-k} - n^2p^2 \\ &= n(n - 1)p^2 \textstyle\sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!((n-2)-(k-2))!} p^{k-2} (1 - p)^{(n-2)-(k-2)} + np - n^2p^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1 - p). \end{aligned}$$

## Exercises

1. A box contains 4 red balls and 2 black balls. A ball is selected at random, its color is noted, and it is replaced in the box.

$(i)$ What is the probability of getting a red ball in one trial.

$(ii)$ Suppose a sample of size 4 with replacement is taken (as described above, do it 4 times). Let $W$ denote the number of times a red ball is drawn. Compute the following 5 numbers:

$$\{P([W = k]),\, 0 \leq k \leq 4\}.$$

$(iii)$ Compute $P([W \leq 2])$, $P([W \geq 2])$, and $P([|W - 2| \leq 1])$, where $W$ is as in part $(ii)$.

$(iv)$ Compute $E(W)$, where $W$ is as in part $(ii)$.

$(v)$ Compute $Var(W)$.

2. Prove: if a fair coin is tossed $n$ times, and if $Z$ denotes the number of times that it comes up heads, then

$$P([Z = k]) = \frac{\binom{n}{k}}{2^n} \text{ for } 0 \leq k \leq n.$$

3. A fair coin is tossed 3 times. Let $Y$ denote the toss at which the first head appears, and if no head occurs, let $Y = 0$.

$(i)$ Compute $P([Y = k])$, $0 \leq k \leq 3$.

$(ii)$ Compute $E(Y)$ and $Var(Y)$.

4. Suppose $X$ is a random variable, and suppose $E(X) = 10$. Evaluate $E(X + 3)$, $E(X - 15)$, $E(13X)$, and $E(2X - 13)$.

5. Suppose $Y$ is a random variable whose distribution is given by $P([Y = -4]) = 0.5$, $P([Y = 12]) = 0.2$ and $P([Y = 10]) = 0.3$. Compute $E(Y)$ and $E(Y^2)$.

6. A box contains 4 tags numbered $-2$, 2 tags numbered 0, 1 tag numbered 1, and 3 tags numbered 2. A tag is selected at random from the box. Let $W$ denote the number on the tag. Compute:

$(i)$ $E(W)$,

$(iii)$ $Var(W)$,

$(iii)$ $E(2W + 15)$,

$(iv)$ $Var(-2W)$, and

$(v)$ $Var(3W + 21)$.

7. If $T$ is a random variable whose distribution is $Bin(6, \frac{2}{3})$, compute $P([T = 6])$, $P([T = 5])$, and $P([T \geq 5])$.

8. If $W$ is the random variable given in problem 6, compute $P([W \geq 2])$, $P([W \geq 0])$ and $P([W \geq 1])$.

9. Prove: If $X$ is a random variable, then $E(-X) = -E(X)$. (Hint: $-X = (-1)X$.)

10. Prove: If $Y$ is a random variable, then $Var(-Y) = Var(Y)$.

11. In a certain game, an outcome or event $A$ can occur with probability 0.25. Let $Z$ denote the number of times that $A$ occurs if you play this game 60 times.

$(i)$ What is the distribution of $Z$?

$(ii)$ Compute $E(Z)$.

$(iii)$.Compute $Var(Z)$.

$(iv)$.What are the smallest and largest values that the random variable $Z$ can take?

$(v)$ Compute $E(\frac{Z}{60})$.

$(vi)$ Compute $Var(\frac{Z}{60})$.

$(vii)$ What are the smallest and largest values that the random variable $\frac{Z}{60}$ can take?

12. Prove: If $X$ is a random variable with at least two distinct numbers in its range, then $Var(X) > 0$.

**3. Limit Theorems.** The fundamental theorem for this course, Bernoulli's Theorem, is proved in this section. It connects a certain amount of theory with a certain amount of reality. Also of concern is how to obtain at least a good approximation for probabilities of events like $P([X \leq k])$, when $X$ has the $Bin(n, p)$ distribution and when $n$ is large, say, over 100. This is accomplished by stating without proof the Laplace-DeMoivre theorem; its proof is simply beyond the scope of this course. However, for our uses, Bernoulli's theorem can be proved. But first we need a lemma.

**Lemma 1. (Chebishev's inequality)**. If $X$ is a random variable, then, for every $\epsilon > 0$,

$$P([|X - E(X)| < \epsilon]) \geq 1 - \frac{Var(X)}{\epsilon^2}.$$

Proof: For any random variable $Y$,

$$\begin{aligned} E(Y^2) &= \sum\{y^2 P([Y = y]) : y \in range(Y)\} \\ &\geq \sum\{y^2 P([Y = y]) : y \in range(Y), |y| \geq \epsilon\} \\ &\geq \sum\{\epsilon^2 P([Y = y]) : y \in range(Y), |y| \geq \epsilon\} \\ &= \epsilon^2 P([|Y| \geq \epsilon]), \text{ or} \\ P([|Y| \geq \epsilon]) &\leq \tfrac{E(Y^2)}{\epsilon^2}, \end{aligned}$$

and, subtracting both sides from 1, we obtain

$$P([|Y| < \epsilon]) \geq 1 - \frac{E(Y^2)}{\epsilon^2}.$$

Replacing $Y$ by $X - E(X)$, we obtain the conclusion.

**Theorem 1. (Fundamental Theorem for This Course: Bernoulli's Theorem).** Suppose $S_n$ is $Bin(n, p)$. Then, for every $\epsilon > 0$, no matter how small $\epsilon$ might be,

$$P\left(\left[\left|\frac{S_n}{n} - p\right| < \epsilon\right]\right) \to 1 \text{ as } n \to \infty.$$

Proof: By the propositions in section 2, $E(\frac{S_n}{n}) = \frac{1}{n}E(S_n) = \frac{1}{n}np = p$, and

$$Var(\frac{S_n}{n}) = \frac{1}{n^2}Var(S_n) = \frac{1}{n^2}np(1-p) = \frac{1}{n}p(1-p)$$

Thus, by Chebishev's inequality,

$$1 \geq P\left(\left[\left|\frac{S_n}{n} - p\right| < \epsilon\right]\right) \geq 1 - \frac{1}{n\epsilon^2}p(1-p) \to 1$$

as $n \to \infty$, which concludes the proof.

**Important:** Bernoulli's theorem is central to this course as presented in this treatise. First, Bernoulli's theorem gives a physical meaning to the notion of probability. Namely, for any game, the probability of an event means that if the game is played a large number of times, then the relative frequency with which a certain event occurs becomes close to the probability of the event. From a practical point of view, suppose you have a game, and you wish to know the probability of a certain event $E$. The mathematics of computing the number of equally likely outcomes and the number of these equally likely outcomes for which $E$ occurs might be too difficult or complicated to compute $P(E)$. However, if you can play the game again and again, many times, you can approximate the probability of $E$. Suppose that during $1,000$ plays, the event occurs $S_n = 631$ times. Thus, you are observing $\frac{S_n}{n} = \frac{631}{1,000}$, and you can rest assured that $P(E)$ is somewhere near the value $0.631$. Of course, if you play the game more than $1,000$ times, you can obtain greater accuracy. This greater accuracy can be achieved quickly by playing the game with a computer. This course is based in large part on being able to approximate a probability by playing a game many thousands of times

and using the relative frequency of the occurence of the event in question to approximate the probability. This is referred to here as **simulating the game.**

Consider this primitive example. Suppose you have a box with 10 tags in it, numbered from 1 to 10, and you play the following game. A tag is drawn at random from the box. The number on it is noted. It is replaced, and a second tag is drawn at random, and its number is noted. The problem is to find the probability that the sum of the two numbers is less than 13. There is a combinatorial way in which to solve this problem, but if you do not know about this, you can get an approximation of the probability in this manner. On your calculator there is a button that serves as a random number generator. Simply put, if the random number generator only gives you random numbers with accuracy to four decimal places, this means that it is in essence pulling a number from $10,000$ tags in a box numbered from $0.0000$ to $0.9999$. So after entering your seed number, look at the first random number that you get. Suppose it is of the form $0.uvxy$. Pretend that the first tag that the first number that you pull out of the box is $u$ and that the second number is $v$. If either of these two number is 0, replace it with the number 10. Then compute the sum of these two numbers. If the sum is less than 13, then you have played your original game just once, and the event that the sum is less than 13 has occurred. So be patient, and play this realization of the game perhaps 100 times. By Bernoulli's theorem, the ratio of the number of times that the outcome is less than 13 to the number of times the game is played should start to get somewhat close to the probability that you seek.

In addition to Bernoulli's theorem, we shall need a theorem that approximates the binomial distribution when the number of trials is too large. After stating the theorem, we shall illustrate its use.

**Theorem 2 (Laplace-DeMoivre Theorem).** If $S_n$ has the $Bin(n,p)$ distribution, then, for every real $x$,

$$P\left(\left[\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right]\right) \to \Phi(x)$$

as $n \to \infty$, where (and the symbol in the following display formula will be explained below)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

As stated above, a proof of this theorem is beyond the scope of this

course. The function $\Phi(x)$ is usually referred to as the normal distribution or the standard normal distribution.

The value of $\Phi(c)$ is the area enclosed by the curve $y = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, the $t-$axis and to the left of the line $t = c$. The total area enclosed by the curve $y = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ and the $t$-axis is 1.

**Proposition.** The curve $y = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ is symmetric with respect to the $y$-axis.

Proof: Recall that a function $y = f(x)$ is said to be symmetric with respect to the $y$-axis if $(-x, f(x))$ is a point on the curve for every real $x$ in the domain of $f$. Thus, $\frac{1}{\sqrt{2\pi}} e^{-(-t)^2/2} = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$, which proves the proposition.

Values of $\Phi(x)$ for various values of $x$ can be found in tables at the end of any statistics book or in some of the popular spreadsheets. In particular, $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$ are convenient numbers to remember. Because of the symmetry proved in the previous proposition, $\Phi(-1.645) = 0.05$ and $\Phi(-1.96) = 0.025$. These values of $\Phi$ should be remembered. (For other values of $\Phi(x)$, go to EXCEL, click on $f_x$, in the Function Category, click on Statistical, and in the Function Name, click on NORMDIST.)

A most frequent application of the Laplace-DeMoivre theorem occurs when one wishes to evaluate $P([X \leq k])$ when $X$ is $B(n, p)$ and when $n$ is so large that software like EXCEL will not provide it. In this case, note that

$$P([X \leq k]) = P\left(\left[\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{k - np}{\sqrt{np(1-p)}}\right]\right).$$

Looking at the statement of the Laplace-DeMoivre theorem, one sees that for $n$ very large, the above probability is close to

$$\Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right).$$

Actually, a slightly better approximation that can be justified is by evaluating

$$\Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right),$$

and, not going into details to prove this, we shall use this correction (called the integer correction) from here on.

Here is an example of the use of the Laplace-DeMoivre theorem to approximate a particular value of the binomial distribution. Suppose $X$ is a random variable whose distribution is $Bin(2000, 0.4)$, and we wish to determine $P([X \leq 1030])$. Then, since the inequality $X \leq 1030$ is the same as the inequality

$$\frac{X - 2000 \times 0.5}{\sqrt{2000 \times 0.5 \times 0.5}} \leq \frac{1030 - 2000 \times 0.5}{\sqrt{2000 \times 0.5 \times 0.5}},$$

we may write

$$P([X \leq 1030]) = P\left(\left[\frac{X - 2000 \times 0.5}{\sqrt{2000 \times 0.5 \times 0.5}} \leq \frac{1030 - 2000 \times 0.5}{\sqrt{2000 \times 0.5 \times 0.5}}\right]\right).$$

Then we may use the Laplace-DeMoivre theorem (and the integer correction referred to above) to approximate this probability by

$$\Phi\left(\frac{1030.5 - 2000 \times 0.5}{\sqrt{2000 \times 0.5 \times 0.5}}\right), \text{ which equals } \Phi\left(1.3640\right).$$

Using EXCEL, we find that $\Phi\left(1.3640\right) = 0.9137$.

Another example is this. Suppose $X$ is a random variable whose distribution is $Bin(625, 0.75)$, and we wish to find a positive number $k$ that satisfies $P([X \leq k]) = 0.35$. According to the manipulations performed above, we wish to find the value of $k$ such that

$$\Phi\left(\frac{k + 0.5 - 625 \times 0.75}{\sqrt{625 \times 0.75 \times 0.25}}\right) = 0.35.$$

Using the NORMINV program in EXCEL we find that $k$ must satisfy

$$\frac{k + 0.5 - 625 \times 0.75}{\sqrt{625 \times 0.75 \times 0.25}} = -0.3853.$$

Solving for $k$, we obtain $k = 464.08$ or $k = 464$. But remember, this is an approximation. Actually, there is probably no number $k$ for which that probability is exact.

## Exercises

1.A game was played 312 times under identical initial conditions. A certain event $E$ occurred in 211 of these games. Determine an approximate value for $P(E)$.

2. A pollster wished to find an estimate of what proportion of the voting population in a certain city was in favor of voting the scoundrels out of office. She determined that this proportion of the population was none other than the probability of selecting a person at random who was in favor of such an action. She selected 600 citizens at random and discovered that 420 of them were in favor of this action. Determine an approximation of the unknown proportion.

3. If $X$ is a random variable whose distribution is $Bin(576, 0.37)$, find the value of $P([X \leq 200])$ (using EXCEL), and find the approximation of it by applying the Laplace-DeMoivre theorem.

4.If $Y$ is a random variable whose distribution is $Bin(1600, 0.5)$, use the Laplace-DeMoivre theorem to approximate the value of $P([Y \leq 790])$.

5.If $Z$ is a random variable whose distribution is $Bin(200, 0.45)$, find the value of $P([87 \leq Z \leq 94])$. (Hint: Make use of the fact that

$$[Z \leq 94] = [Z \leq 86] \cup [87 \leq Z \leq 94],$$

which is the union of disjoint events.)

6. Suppose $W$ is a random variable whose distribution is $Bin(100, 0.63)$. Determine $P([W \leq 60])$ first by using BINOMDIST in EXCEL and then by using the approximation provided by the Laplace-DeMoivre theorem.

7. In a previous era, before the advent of desktop computers, the LaPlace-DeMoivre approximation was used to aproximate $P([X \leq k])$ when $X$ was $Bin(n, p)$ when $n$ was much smaller than 100 but $p$ was not too close to either 0 or 1. Fill out the following table to see how much the approximation improves with larger values of $n$:

| – | – | – | BINOMDIST | LaPlace-DeMoivre |
|---|---|---|---|---|
| $n$ | $p$ | $k$ | $P([X \leq k])$ | $P([X \leq k])$ |
| 20 | 0.5 | 11 | | |
| 40 | 0.5 | 22 | | |
| 60 | 0.5 | 33 | | |
| 80 | 0.5 | 44 | | |
| 100 | 0.5 | 55 | | |

8. Prove: if $0 < m < n$, then

$$\sum_{k=m}^{n} k = (n - m + 1)\frac{n+m}{2}.$$

(Hint: Start out by writing

$$\sum_{k=m}^{n} k = (m) + (m+1) + (m+2) + \cdots).)$$

(Another hint: Or, start out by writing

$$\sum_{k=1}^{m-1} k + \sum_{k=m}^{n} k = \sum_{k=1}^{n} k.)$$

9. Let $S_5$ be a random variable with the $Bin(5, 0.5)$ distribution. Find the smallest positive integer $k$ such that $P([S_5 \geq k]) \leq 0.10$. (Hint: You know that $range(S_5) = \{0, 1, 2, 3, 4, 5\}$. So I would start out with $k = 5$ and then proceed to $k = 4$, etc.)

10. (You might wish to use EXCEL for this and experiment with different values of $p$ until you find the right answer.) Suppose $Z$ is a random variable whose distribution is $Bin(10, p)$. Find the largest value of $p$, accurate up to two decimal points, so that $P([Z \leq 8]) \geq 0.95$.

11. If $U$ is a random variable whose distribution is $Bin(10, p)$ find the largest value of $p$ (up to two decimal point accuracy) such that $P(U \geq 10]) \leq 0.01$.

12. Prove: If $a$ and $b$ are real numbers, and if $a \geq b$, Then $1 - a \leq 1 - b$.

13. If $X$ is a random variable with the binomial distribution $Bin(n, p)$, prove that $Var(\frac{X}{n}) \leq \frac{1}{4n}$.

14. If $X$ is $Bin(500, 0.4)$, find the largest value of $k$ such that $P([X \leq k]) \leq 0.6$.

## Chapter 4. Binomial Tests.

**1. Binomial Test for Small Sample Sizes**. The first three chapters dealt with the basic material in probability needed for the rest of this course. We shall use all the results established there for the balance of this text. In all, we shall develop quite a few of the statistical procedures most frequently used in research and development in the biomedical and social sciences. We begin here with the simplest test that illustrates the basic ideas.

Suppose that a medical research scientist has developed a new treatment for a certain disease. Further, suppose that she wishes to determine if it is

better than a traditional treatment, but a treatment that is not always effective. In such a case there might have been gathered data on a large number of people who have had the disease in the past and were given the traditional treatment. These data yield for us two numbers: $(i)$ the total number, $N$, of people with the disease who were treated by this traditional treatment in the past, and $(ii)$ the number, $S$, of these patients who were treated and who responded favorably to this treatment. If we consider the observation of patient after patient so treated as the playing and replaying of the same game under identical conditions, then the event "favorable response" has a probability which can be estimated, using Bernoulli's theorem, by $S/N$.

Let us suppose for concreteness that out of many thousands treated in the past by this traditional treatment, 60% of them responded favorably to it. This means, by Bernoulli's theorem, that the probability that an individual with this disease will respond favorably, if given the traditional treatment, or a treatment that is only as good as it, is about $p = 0.60$. Suppose that this medical research scientist wants to determine whether her new treatment is better than the traditional treatment. So she selects 30 patients at random from among new people who at present have the disease, and she tries the treatment on each of them. This is essentially playing a game 30 times and observing the number of those times in which the event, which we denote by [favorable response], occurs. Suppose that when the results are in, she finds that 22 of these 30 patients have responded favorably. Note that out of the 30, the proportion of patients that respond favorably is $\frac{22}{30} = 0.7333$, or the percentage that respond favorably is 73.33%. So the problem arises whether she can make a claim that her treatment is an improvement over the traditional treatment.

If her treatment is no better than the traditional treatment, then she has observed the value of a random variable whose distribution is $Bin(30, 0.6)$. The question then arises on whether a random variable whose distribution is $Bin(30, 0.6)$ can achieve a value **as extreme as** 22. If the probability of this happening is **unbelievably small**, then we might conclude that the probability of obtaining a favorable response with this treatment on a patient with this disease is larger than 0.60. In such a case, she might conclude that her treatment is superior to the traditional treatment.

But what is meant by "as extreme as" 22? By Bernoulli's theorem, we would expect the values of a $Bin(30, 0.6)$ random variable to be reasonably close to its expectation, $30 \times 0.6 = 18$. So "as extreme as" 22 means "is equal to or greater than 22". She must compute $P([X \geq 22])$ where $X$ has

the $Bin(30, 0.60)$ distribution. In this case

$$P([X \geq 22]) = 1 - P([X \leq 21]) = 1 - \sum_{j=0}^{21} \binom{30}{j} (0.6)^j (0.4)^{30-j}.$$

Performing this computation, she obtained $P([X \geq 22]) = 0.0941$. Is this unbelievably small? Not quite. Research scientists generally hold to the definition of "unbelievably small" as being 0.05 or less. This is reasonable if the research project is just in the exploratory stage. However, if one wanted to announce a major breakthrough, then one might think of a probability as small as 0.01. We shall refer to a computation of the sort in the displayed formula when $n$ is small, say, less than 100 or 150, as the **Binomial Test for Small Sample Sizes**. The probability, $P([X \geq 22])$, is called the **P-value** of the test.

This question might arise in the above problem: What is the minimum number of favorable responses out of 30 that one would have needed in order for the P-value to be less than 0.05, or less than 0.01, or less than 0.001? Using BINOMDIST in EXCEL, we find that $P([X \leq 22]) = 0.9565$, $P([X \leq 23]) = 0.9828$, $P([X \leq 24]) = 0.9943$, $P([X \leq 25]) = 0.9984$ and $P([X \leq 26]) = 0.9997$. Thus one would have had to have at least 23 favorable responses for the P-value to be less tha 0.05, at least 25 favorable responses for the P-value to be less than 0.01, and at least 27 favorable responses for the P-value to be less than 0.001.

Recall from section 1 of chapter 3 that in order to compute probabilities of the form

$$\sum_{j=0}^{21} \binom{30}{j} (0.6)^j (0.4)^{30-j}$$

that we obtained above, use BINOMDIST as found in EXCEL.

## Exercises

1. Suppose that you want to test a coin to see if it is an unbiased, or fair, coin. So you toss it 45 times and discover that it came up heads 17 times. If 0.05 is for you an unbelievably small probability, determine whether the coin is fair or not. Note that 17 is less than the expectation, so the number 17 is extreme in the sense that it might be too small. So one would wish to compute $P([X \leq 17])$ when $X$ has the $Bin(45, \frac{1}{2})$ distribution. What is the value of $P([X \leq 17])$?

2. In problem 1, what is the largest number of heads in order for the P-value to be smaller that 0.05, 0.01 or 0.001?

3. Two cleaning detergents, A and B, are being tested on garments to see which detergent, if any, is better than the other. The design of the trial is what is usually referred to as **matched pairs**. You assemble various patches of cloth of various degrees of dirtiness, cut each in half, randomize their order within each pair, apply A to one halfpatch and B to the other halfpatch, and finally conclude with each pair of halfpatches which detergent did the better job. In trying this on 80 matched pairs, A did a better job than B in 48 pairs. Is A really better than B?

4. In problem 3, what is the smallest number of halfpatch pairs needed for A to do a better job than B in order for the P-value to be smaller than 0.05, 0.01 or 0.001?

5. A standard treatment for a certain ailment is successful in treating it 60% of the time. A new treatment is proposed and is tried on 20 subjects. If the new treatment is no better than the standard one, what is the probability that it will be successful on at least 15 of these patients?

6. In Problem 5, what is the probability that the new treatment will be successful in at most 7 patients?

**2. Binomial Test for Large Samples.** In the same spirit as above, we now indicate the application of the Laplace-DeMoivre theorem when sample sizes are too large to use the summation formula provided in the previous section. Let us suppose you are a sample survey statistician who is asked to find out if candidate A is leading candidate B in the upcoming election. Suppose that there are hundreds of thousands of voters in this particular area, and you cannot ask each one whom he or she intends to vote for. So let us say that you take a random sample of size $1,600$ from the population and ask them concerning their preference. This is equivalent to playing a game $1,600$ times and determining for each play whether the event [VOTER for A] occurs or does not occur. The problem here is to determine whether $P([\text{VOTER for A}]) > \frac{1}{2}$ or whether this probability is $\leq \frac{1}{2}$. So suppose that 830 out of your $1,600$ to whom you put the question state that they are for Candidate A. The question that arises is whether this extreme a value is possible if the largest value that $P([\text{VOTER for A}])$ can be is $\frac{1}{2}$. So we assume that at most, A is just even with B, i.e., that $P([\text{VOTER for A}]) = \frac{1}{2}$, and we wonder whether, under this assumption, the probability of obtaining as extreme a value as 830 is unbelievably small. If it is, then we know that the assumption of $P([\text{VOTER for A}]) \leq \frac{1}{2}$ is incorrect and that this probability is larger than 0.5. (And thus the proportion of the population for candidate A is greater than one half.) So let $X$ be a random variable whose

distribution is $Bin(1600, \frac{1}{2})$. We must determine the value of $P([X \geq 830])$. For large sample sizes, the practical difficulties of computing this probability by evaluating

$$\sum_{j=830}^{1600} P([X = j]) = 1 - \sum_{j=0}^{829} P([X = j])$$

are overwhelming. (Try it and you will see.) But we can approximate this probability by using the Laplace-DeMoivre theorem given in chapter 3. The reasoning goes like this. The number $1,600$ is rather close to infinity, so we may use this approximation which starts on the third line:

$$\begin{aligned}
\sum_{j=0}^{829} P([X = j]) &= P([X \leq 829]) \\
&= P\left(\left[\frac{X - 1600 \times \frac{1}{2}}{\sqrt{1600 \times \frac{1}{2} \times (1 - \frac{1}{2})}} \leq \frac{829 - 1600 \times \frac{1}{2}}{\sqrt{1600 \times \frac{1}{2} \times (1 - \frac{1}{2})}}\right]\right) \\
&= \Phi\left(\frac{829.5 - 1600 \times \frac{1}{2}}{\sqrt{1600 \times \frac{1}{2} \times (1 - \frac{1}{2})}}\right) \\
&= \Phi(1.4750) = 0.9299.
\end{aligned}$$

Thus, the probability of getting a value as extreme as 830 is approximately $1 - 0.9299$, or $0.0701$, a probability that is not unbelievably small. We do not have convincing evidence that over half the voters will vote for candidate A.

As mentioned before, for values of $\Phi(x)$ that you might need, go to NORMDIST on EXCEL.

## Exercises

1. In a study of a certain group of 900 people with a common health problem, 475 of them had temperatures greater than standard temperature of 98.6 degrees Fahrenheit and 425 had temperatures less than 98.6. Are you able to conclude that the median temperature of all people with this health problem is higher than 98.6? (By median temperature, we mean in this case a number such that the probability of being greater than it is $\frac{1}{2}$ and the probability of being less than it is $\frac{1}{2}$.)

2. A claim has been made that 10% of all golfers are left-handed. In a random sample of size 260 golfers, it was determined that 38 of them were left-handed. This is more than 10% of the number of golfers in the sample. Do you have reason to doubt the claim?

3. In testing a new drug on 500 people for lowering systolic blood pressure, their blood pressure was measured before taking the drug, and after taking

the drug for a week, their blood pressure was measured again. In 270 cases their systolic blood pressure decreased, and in 230 cases it increased. We might assume that if this new drug has no effect, then after a week the systolic blood pressure has an equal chance of going up as going down. Is there any reason to believe that the new drug is somewhat effective?

**4**. One week before a hotly contested election between candidate A and candidate B for the post of mayor of a large city, a simple random sample of size 600 was taken of the registered voters. Each individual in the sample was asked whether he or she intended to vote for candidate A or candidate B. It turned out that 320 of those in the sample declared their intention to vote for candidate A. If 0.05 is consisered to be an unbelievably small probability, is candidate A able to confortably assume that he will win?

**5**. Management considers that a defective rate of 2% is the largest they can endure in order to remain competative in the market for this product. Among the first 200 units of the product made and then tested, 5 of them were found to be defective. Does this necessarily mean that the defective rate is more than 2%?

**3. The Sign Test.** Sometimes the data consist of $n$ pairs of numbers on $n$ individuals, call one such pair $(x, y)$, in which you wish to test whether one treatment is better than the other, where by "better" we might mean "has a larger measurement". In such a pair, $x$ might be a measurement on treatment $A$, and $y$ might be a measurement on treatment $B$, Depending on the circumstances, if the first measurement is always greater than the second measurement, and if $n$ is large, then we might conclude that the first treatment is better than the second, while if the second is always greater than the first, we would tend to believe that the second treatment is better than the first. But such clear-cut results are not usually forthcoming. In a usual case, if there is no difference between the two treatments, it is equally likely that $x > y$ and $y > x$. Thus, if there is no difference between the two treatments, and if $x \neq y$ for all pairs of individuals involved in the clinical trial, we might use as a model that the number of pairs of individuals for which $x > y$ is a random variable whose distribution is $Bin(n, \frac{1}{2})$. A count is made on the number of individuals among the $n$ individuals for whom $x > y$. Suppose that for $k$ of these individuals, the result is $x > y$, while for the remaining $n - k$ individuals, $y \geq x$. Thus, for a random variable $X$, we wish to find the probability $P([X \leq k])$ if $k < n/2$, and we wish to find the probability $P([X \geq k])$ if $k > n/2$. If the probability of obtaining a value **as extreme as** $k$ is **unbelievably small**, then we would wish to reject the null

hypothesis that there is no difference between the two treatments in favor of treatment B in the first case or in favor of treatment A in the second case. It should be easily recognized that this is just a special case of the binomial test.

As an example, let us consider a design known as a **crossover trial.** We suppose that two headache remedies are to be tested, call them remedy A and remedy B. The numerical outcome of our product test is to measure how long the headache continues after taking the remedy. We assemble a sample of $n$ individuals who are subject to headaches. Suppose $n$ is an even number. Select $n/2$ of these individuals at random to give remedy A to for the first test, and give to the remaining $n/2$ individuals remedy B first. For each, record the time that it takes to recover from the headache. When you are finished, wait for a period of time (called the washout period), and then when each individual in the trial gets a headache again, have that person take the other remedy, and, for each such individual, record the time it takes to recover from the headache. For each individual, let $x$ denote the length of time it takes to recover from the headache after taking remedy A, and let $y$ denote the same measurement for remedy B. For concreteness, suppose that $n = 32$, and suppose that in 12 of these cases, $x > y$.

At this point we should not hastily jump to the conclusion that remedy B is better than remedy A. Instead, we consider a random variable $X$ whose distribution is $Bin(32, \frac{1}{2})$ and find the probability $P([X \leq 12])$. In this case we must evaluate

$$P([X \leq 12]) = \sum_{i=0}^{12} \binom{32}{i} \frac{1}{2^{32}}.$$

This turns out to be 0.10766, which by most standards is not unbelievably small. So we cannot conclude that there is a difference between the two remedies. (The largest number $k$ for which $P([X \leq k]) \leq .05$ is true is 10.)

Similar treatment can be made for matched pairs. In this case, you select pairs of people that match each other in health, age, gender, etc., as much as possible. It should be obvious how such data are treated in this case.

For appropriate software, go to BINOMDIST in EXCEL if $n$ is not too large, or go to NORMDIST on EXCEL in order to apply the Laplace-DeMoivre theorem for large values of $n$.

### Exercises

1. Verify the statement made in parenthesis above: the largest value of $k$ for which $P([X \leq k]) \leq .05$ is true is 10.

2. Suppose $X$ is a random variable whose distribution is $Bin(1220, \frac{1}{2})$. Find the largest value of $k$ such that $P([X \leq k]) \leq 0.05$.

3. Suppose $X$ is $Bin(35, 0.6)$. Find $P([X \leq 19])$ and $P([X \geq 20])$. Do they add up to 1?

4. Suppose $Y$ is $Bin(1420, 0.3)$. Show how you would use the Laplace-DeMoivre theorem to evaluate (with reasonable accuracy) $P([Y \geq 434])$.

5. An investigator wishes to determine if sitting upright in a chair versus lying down in bed will affect a person's systolic blood pressure. Here are the data obtained by experimenting with 10 subjects.

| Patient # | Upright | Lying down |
|---|---|---|
| 1 | 142 | 154 |
| 2 | 100 | 106 |
| 3 | 112 | 110 |
| 4 | 92 | 100 |
| 5 | 104 | 112 |
| 6 | 100 | 101 |
| 7 | 108 | 120 |
| 8 | 94 | 90 |
| 9 | 104 | 105 |
| 10 | 98 | 114. |

Well, does it?

**4. Comparison of Two Populations for Rare Events.** This deals with comparing the rates of occurrence of rare events in two populations. Here is an example.

Let us suppose that there are two cities, which we shall denote by A and B. The population of A is $100,000$, and the population of B is $150,000$. City A has nothing unusual about it. City B just happens to be next to a toxic waste dump. The state public health authorities have noticed a higher rate of a certain cancer in city B than in city A, and they are beginning to wonder whether this is by chance or if there is a real difference, with city B's larger than that of city A's. If so, then one might begin to wonder whether this particular difference, if it exists, is due to that toxic waste dump nearby. A study is to be made during a particular year to see if there is any difference in the incidence of new cancer cases.

We may consider each new cancer case diagnosed in either city as a play of a game, in which the outcome of each play of the game is either the event [the cancer occurred in city A] or [the cancer occurred in city B]. Let us

suppose that there is no difference in cancer rate between the two cities. It is reasonable to assume that the probability that a new cancer case comes from city A is

$$p = \frac{100,000}{100,000 + 150,000},$$

or the probability is $p = 0.40$. Then at the end of the year, if there were no difference in the cancer rates of the two cities, the number of cases occurring in city A should be an observation on a random variable $S$ whose distribution is $Bin(r, p)$, where $r$ denotes the total number of cases of cancer in both cities, and $p = 0.40$. In this case suppose $r = 65$ and the number of those cancer cases that occur in city A is 20. Under our null hypothesis that both cities' cancer rates are the same, we would expect that the number of cancer cases in city A should be $65 \times 0.40 = 26$. Indeed, the number in city A is lower than expected. If we let $X$ denote a random variable whose distribution is $Bin(65, 0.40)$, and if we assume that there is no difference between the two cities with respect to the incidence of cancer, then we wish to compute $P([X \leq 20])$, which is the probability of observing a value **as extreme as** 20. If this probability is **unbelievably small**, then we should wish to reject our null hypothesis that there is no difference in favor of the alternative that there is a difference, and city B's is larger.

It turns out in this example that $P([X \leq 20]) = 0.08$. This is not an unbelievably small probability, and so we cannot conclude that the cancer rate in city B is greater than that of city A.

For appropriate software, go to BINOMDIST or NORMDIST on EXCEL as the value of $n$ warrants.

## Exercises

1. Two industrial processes are being investigated to see if there is a difference between them in producing defective items. In process A, out of $8,000$ items produced, 26 were defective. In process B, out of $6,000$ items produced, 27 were found to be defective. Does management have any reason to suspect process B? Explain in full, and find the P-value.

2. In 1954, nationwide clinical trials were conducted on the new Salk vaccine to determine if it prevented polio myelitis. The trials consisted of taking two groups at random from the general population of children in the U.S.A., giving one group (called the treatment group) the Salk Vaccine, and giving a placebo (an inert imitation of the Salk Vaccine) to the other group

(called the control group). Each group contained about 200,000 children. After a fixed period of time, 35 children in the treatment group came down with crippling polio, while in the control group, 120 children came down with crippling polio. So what do you get for a P-value here? (The numbers in this problem are close but not exact.)

3. In a certain large city, there were 55 gang crimes of a certain category committed during the year 2006. In 2007 with new law enforcement procedures in use, there were 35 such crimes committed. Did the crime rate for this category actually decrease?

**5. Test for Median Value.** On some very few occasions, someone has a set of data taken on some phenomenon, say the numbers are $x_1, x_2, \cdots, x_n$, and he or she wishes to test whether "the median" for the phenomenon producing these data is an already known number, which we shall denote by $\mu$. The first question that we must ask is: what is meant by "the median"? In this course, we shall consider only those cases where the definition of "the median" is a number for which $(i)$ the probability that any observation taken is greater than this number is $\frac{1}{2}$, $(ii)$ the probability of any observation being less than it is also $\frac{1}{2}$, and $(iii)$ the probability of any observation being equal to it is zero. You already have these observations, which constitute a history of playing the game $n$ times. The event that we are considering is the event that the outcome is less than $\mu$. It is clear that one should count the number of data values that are less than $\mu$ and the number that are greater than $\mu$. We could then consider the smaller of these two numbers, call it $k$, and determine the value of $P([X \leq k])$, where $X$ has the $Bin(n, \frac{1}{2})$ distribution. If the probability of getting a number this extreme is unbelievably small, we should reject the null hypothesis that the median is $\mu$. Naturally, if the probability $P([X \leq k])$ is unbelievably small and we reject the null hypothesis, we should state that the true value of the median is greater than $\mu$.

For appropriate software, go to BINOMDIST or NORMDIST on EXCEL as the value of $n$ warrants.

A far more frequently encountered problem is to use the data to find a range of the possible values of $\mu$. This will be encountered in the chapter on confidence intervals.

## Exercises

1. In problem 5 in section 3, the problem could be reformulated to ask:

for the data provided by taking the ten differences between sitting upright and lying down, is zero the median? Verify that this is the case.

2. Suppose you are given a data set consisting of 30 distinct numbers. These are arranged from smallest to largest. The smallest is called the first order statistic, the next to smallest is called the second order statistic, ..., and the largest is called the 30th order statistic. What is the largest order statistic that is less than the median with probability equal to or less than 0.05?

3. Suppose that you are given a data set of 160 distinct numbers. What is the smallest order statistic that is greater than the median with probability 0.95?

## Chapter 5. Two Sample Simulation Tests.

**1. Two Sample Simulation Tests Based on the Difference of Sample Means and Difference of Sample Medians.** The central idea of a permutation test or simulation test can be introduced through a special application that occurs frequently in biomedical research and development. Let us consider a certain ailment or disease for which there already is a standard treatment, whose effectiveness can be measured and recorded as some number. For example, the response to such a treatment might be the weight loss, or it might be the time between the administration of the treatment and recovery, or it might be the change in level of blood sugar. In any case the response for each patient is some number which we are able to observe and record.

Now suppose that a new treatment has been developed, and suppose that those who developed it claim it is a better treatment. The problem faced is to determine in as objective a manner as possible whether it is indeed a better treatment. We shall assume, for this ailment or disease, that a large measured response indicates a better treatment than a small measured response would indicate. In order to carry out a test for this new treatment, one would wish to have two groups of people who are initially suffering from the disease and at the same severity. Patients in one group would receive the traditional treatment; this group would be called the **control group**. The patients in the other group would receive the new treatment; this group is called the **treatment group**. Thus, if we were to have two identical, large groups of people, all of whom have the same symptoms and severity of the disease,

and if after receiving treatments everyone in the **control group** ended up with the same small measurement, and if everyone in the **treatment group** ended up with the same larger measurement, then we might conclude that the new treatment is better than the traditional treatment.

But results of clinical trials are not as clean cut as this, and a rigorous protocol must be followed for a meaningful analysis. What usually happens in practice is that $n$ people at a certain stage of the disease are given the traditional treatment, and thus they become the control group. Then $m$ others are given the new treatment, and thus they become the treatment group. The two groups are usually determined as follows in order to rule out bias. First there is a method of selecting $m + n$ people who are all at the same stage of the disease and who agree to participate in the clinical trial. Each knows ahead of time that he or she will receive one of the treatments but will not know which treatment he or she is receiving. Also each knows that the medical personnel administering the treatment are unaware of which treatment is being administered to which patient. These last two sentences define what is called a *double blind study*; it insures unbiasedness. Next, the patients are numbered from 1 to $m + n$. This might be done by numbering them in the order in which they arrive to participate in the study, or they might be numbered in their alphabetical order. Then a sample of size $m$ is selected at random without replacement from this group of $m + n$ patients by selecting $m$ numbers at random without replacement from the numbers 1 to $m + n$. This chosen group will become the treatment group, and the remaining $n$ patients will become the control group. Persons not connected with participating in or administering the trials will be the only ones who know who is getting the new treatment and who is receiving the traditional treatment. At the conclusion of the study, the quantity $z$ is measured and recorded for each patient. If all the measurements of those in the treatment group are well above all the measurements of those in the control group, we would be tempted to conclude that the new treatment is better than the old one.

But here is what usually happens. The first thing that one notices in practice is that for each group the values of $z$ are not all the same. There is usually considerable scatter for the $z$-values of each group, and there is usually some overlap in which some of the $z$-values of the control group might be higher than some of the largest $z$-values of the treatment group. And so, although most values of the treatment group could be higher than most values of the control group, we hesitate to draw a conclusion. Thus the

question arises: is there really a difference between the two groups? How can we tell? We might agree that they are really different if the arithmetic or sample mean $\overline{x}$ of the $z$-values of the treatment group is substantially larger than the arithmetic or sample mean $\overline{y}$ of the $z$-values of the control group. However, what do we mean by "substantially larger"?

One reasonable approach is as follows. Suppose we observe these two arithmetic means, $\overline{x}$ and $\overline{y}$, and suppose we observe that $\overline{x} > \overline{y}$. We ask ourselves: if there were no difference between these two groups, i.e., if there were no difference between the two treatments, is it possible for the difference $\overline{x} - \overline{y}$ to be as large as we observe it to be? If it is not possible for this difference to be so large (under our assumption of no difference), then we would say that the new treatment is better than the standard treatment or control. Looking at the problem and **pretending** that there is no difference between the two treatments is tantamount to stating that we gave all $m+n$ patients the same treatment, and then selected $m$ patients at random out of the $m+n$, observed the mean $\overline{x}$ of their $z$-scores, and then observed the mean $\overline{y}$ of the $z$-scores among those remaining. Thus we may correctly ask, since we feel that the number $\overline{x}-\overline{y}$ is large, if we were to select $m$ numbers at random out of these $m+n$ numbers and denote their arithmetic mean of the $m$ patients by $\overline{X}$, with $\overline{Y}$ denoting the arithmetic mean of those $n$ patients not selected, where both $\overline{X}$ and $\overline{Y}$ are now random variables, what is the probability that the value of $\overline{X} - \overline{Y}$ is as extreme as $\overline{x} - \overline{y}$? In other words, what is the value of $P([\,\overline{X} - \overline{Y} \geq \overline{x} - \overline{y}])$? If this probability were $0.000,001$, our response would be this: here we are, conducting an important trial, and if there were no difference, then the probability of observing a difference as large as the one we observed is "one in a million"! Unbelievable! And thus we would reject our null hypothesis that there is no difference between the two treatments in favor of the alternative that the measurements on the treatment group are significantly larger than those of the control group. On the other hand, if this probability were 0.18, then we would say, "An event of probability 0.18 can certainly occur. Maybe it did. In which case we would have no overwhelming reason to reject our null hypothesis that there is no difference.

Thus, the problem becomes that of computing $P([\,\overline{X} - \overline{Y} \geq \overline{x} - \overline{y}])$. A correct but difficult method of computing this probability is this. First find the number of ways in which one can select $m$ objects out of $m+n$; this turns out to be $\binom{m+n}{m}$. Next, one must look at all of these $\binom{m+n}{m}$ outcomes;

for each of them, one would compute the sample mean $\overline{X}$ of the $m$ numbers selected and the sample mean $\overline{Y}$ of the remaining $n$ numbers, and then one determines the number of equally likely outcomes among the $\binom{m+n}{m}$ for which the event $[\overline{X} - \overline{Y} \geq \overline{x} - \overline{y}]$ occurs. Call this number $N$. Thus, **under the assumption of no difference between the two treatments,**

$$P([\overline{X} - \overline{Y} \geq \overline{x} - \overline{y}]) = \frac{N}{\binom{m+n}{m}}.$$

Now, for large values of $m$ and $n$, this is terribly difficult or impossible to compute in this manner.

However, Bernoulli's theorem in chapter 4 suggests that a next best procedure is the following, which can be done using just about any desktop computer. After having computed $\overline{x}$ and $\overline{y}$ from the data, let $z_1, z_2, \cdots, z_{m+n}$ denote the pooled sample of numbers. Then take a random permutation, $i_1, i_2, \cdots, i_{m+n}$, of the numbers $1, 2, \cdots, m+n$, and observe the reordered data set, $z_{i_1}, z_{i_2}, \cdots, z_{i_{m+n}}$. Let $\overline{x}'$ denote the arithmetic mean of the first $m$ of these numbers, i.e., $\overline{x}' = \frac{1}{m}\sum_{k=1}^{m} z_{i_k}$ and let $\overline{y}'$ denote the arithmetic mean of the remaining data, i.e., $\overline{y}' = \frac{1}{n}\sum_{k=m+1}^{m+n} z_{i_k}$. Evaluate $\overline{x}' - \overline{y}'$ for this outcome, record the number 1 if the inequality $\overline{x}' - \overline{y}' \geq \overline{x} - \overline{y}$ is observed, and record the number 0 if the inequality $\overline{x}' - \overline{y}' < \overline{x} - \overline{y}$ is observed. Repeat this $24,999$ more times. Keep track of the number of 1's, and then divide the number of times 1 occurred by the total number of trials (which is $25,000$). According to Bernoulli's theorem, this ratio should be very close to $P([\ \overline{X} - \overline{Y} \geq \overline{x} - \overline{y}])$, **under the assumption of no difference between the two treatments.**

Again, there is no absolute rule for determining which value of $P([\ \overline{X} - \overline{Y} \geq \overline{x} - \overline{y}])$ is so small as to conclude that there must be a difference between the treatment and the control groups. In a preliminary study like the one just described, if the value of this probability is less than 0.05, then further development and testing are certainly warranted. If it is less than 0.001, then there is strong evidence that there is a difference. It should be noted above all that this simulation or permutation test was suitable because of the way in which the original question was phrased. For appropriate software, use the program PERMMEAN.EXE.

It is also possible to perform a two-sample test on the data obtained as just described by doing a permutation test based on the differences of sample medians. Recall that we defined the sample median for a data set in

section 3 of chapter 1. Let us denote the sample median of the treatment group by $med(x)$ and the sample median of the control group by $med(y)$. Again we wish to know if the observed positive difference, $med(x) - med(y)$ is too large, **under the assumption of no difference between the two treatments.** After having computed $med(x)$ and $med(y)$ from the data, let $z_1, z_2, \cdots, z_{m+n}$ denote the pooled sample. Then take a sample of size $m$ of these numbers, and call the value of their sample median $med(X)$. Let $med(Y)$ denote the sample median of the remaining numbers. We wish now, as above, to find the value of

$$P([med(X) - med(Y) \geq med(x) - med(y)]).$$

We proceed as above, but this time for sample medians. Take a random permutation $z_{i_1}, z_{i_2}, \cdots, z_{i_{m+n}}$ of $z_1, z_2, \cdots, z_{m+n}$. Let $med(X)$ denote the sample median of the first $m$ numbers, and let $med(Y)$ denote the sample median of the remaining $n$ numbers. Then evaluate the number $med(X) - med(Y)$ for this outcome, record the number 1 if the inequality $med(X) - med(Y) \geq med(x) - med(y)$ is observed, and record the number 0 if the inequality $med(X) - med(Y) < med(x) - med(y)$ is observed. Repeat this $24,999$ more times. Keep track of the number of 1's, and then divide the number of times that 1 occurred by the total number of trials (which is $25,000$). According to Bernoulli's theorem, this ratio should be very close to $P([med(X) - med(Y) \geq med(x) - med(y)])$. If this probability is unbelievably small, then one may conclude that the new treatment is better than the old treatment. Appropriate software for this permutation test based on the difference of medians is PERMMED.EXE.

So which test should one use when the two sample test is called for. One might prefer to do both tests. If both tests indicate a significant difference, then one might conclude that there is one.

## Exercises

1. Consider the following set of data:

3.1, 4.2, 5.6, 3.7, 4.6, 4.3, 4.2.

Find the probability that the sample mean of three numbers picked at random from these without replacement is equal to or greater than 4.166666.

2. In problem 1, find the probability that the sample median of 4 numbers picked at random without replacement is equal to or less than 3.5.

3. Here is a walk-through of a simulation test rendered in slow-enough motion so that you can see what is going on. Let us suppose that there are two treatments, the standard treatment and a newly proposed treatment. It is desired to conduct a clinical trial with three patients in each group. While someone is out scouting for six patients who are all suffering from the same ailment and to the same degree, the statistician is selecting three numbers at random from $\{1, 2, 3, 4, 5, 6\}$, with it being decided ahead of time that the six patients are to be numbered from 1 to 6 according to the order in which their last names are alphabetized. So three numbers are selected at random without replacement from $\{1, 2, 3, 4, 5, 6\}$, and suppose they turn out to be 4, 5 and 1. The patients arrive and are numbered, and then patients with numbers 4, 5 and 1 are given the standard treatment, thus becoming the control group. The remaining patients, those numbered 2, 3 and 6, are given the new treatment and are called the treatment group. At the conclusion of the treatments, the $x$-values are measured; they turn out to be

| | |
|---|---|
| control group | 42, 35, 51 |
| treatment group | 29, 41, 36 |

$(i)$ Compute $\overline{x}$, the arithmetic mean of the treatment group, then compute $\overline{y}$, the arithmetic mean of the control group, and finally compute their difference $\overline{y} - \overline{x}$. Set these values aside for the moment.

$(ii)$ List all $\binom{6}{3}$ choices of combinations of $x$-values for the treatment group with the corresponding $y$-values for those remaining as the control group.

$(iii)$ For each individual outcome listed in $(ii)$, let $\overline{X}$ denote the arithmetic mean of the three numbers selected, and let $\overline{Y}$ denote the arithmetic mean of those remaining. Compute $\overline{X}$, $\overline{Y}$ and $\overline{Y} - \overline{X}$ for each of the 20 individual outcomes.

$(iv)$ Compute

$$P\left(\left[\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}\right]\right) = \frac{\#\{\omega : \overline{Y}(\omega) - \overline{X}(\omega) \geq \overline{y} - \overline{x}\}}{\binom{6}{3}} \ .$$

$(v)$ Now we verify the value of this probability experimentally. Use the random number generator on your hand-held calculator to select three numbers at random without replacement from $\{42, 35, 51, 29, 41, 36\}$, and compute their average, $\overline{X}$. Then compute the average, $\overline{Y}$, of the remaining numbers that are not in the sample. If $\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}$, then count 1; if

$\overline{Y} - \overline{X} < \overline{y} - \overline{x}$, then count 0. Repeat this 100 times, and compute the number of $1's$ divided by 100. (Here is a hint: choose 42 if the random number generated is between 0 and 0.1666, choose 35 if the random number generated is between .1667 and .3333, choose 51 if the random number generated is between 0.3334 and 0.4999, etc. If you get a repetition, do not pay any attention to it.)

$(vi)$ Using a desktop computer, simulate this $10,000$ times. (Try PERMMEAN.EXE.)

4. In problem 3, prove: if $n = \frac{N}{2}$, then $\overline{Y} - \overline{X}$ has a symmetric distribution, i.e., the densities of $\overline{Y} - \overline{X}$ and $\overline{X} - \overline{Y}$ are the same.

5. Two teaching methods were tried on two different groups of ten students per group, the same teacher doing the teaching to both groups. At the conclusion of the interval of instruction, they were given the same test at the same time. Here are the two sets of grades in percentages:

Method A: $50, 70, 90, 80, 70, 90, 100, 60, 80, 90$

Method B: $60, 90, 100, 80, 70, 80, 90, 80, 100, 90$ .

Is there any significant difference between the two teaching methods?

6. Here is a challenging problem. Let $x_1, \cdots, x_N$ denote real numbers, let $\overline{X}$ denote the arithmetic mean of a simple random sample of size $n$ taken from them without replacement, where $n < N$, and let $\overline{Y}$ denote the arithmetic mean of those remaining. Prove that $E(\overline{X}) = E(\overline{Y})$.

7. From a group of 9 rats available for a study of transfer of learning, five were selected at random without replacement and were taught to imitate leader rats in a maze. They were then placed together with the four untrained rats in a situatiion where imitation of the leaders enabled them to avoid receiving an electric shock. The results (the number of trials required to obtain 10 correct resposes in 10 consecutive trials were as follows:

| | | | | | |
|---|---|---|---|---|---|
| Trained Rats | 78 | 64 | 75 | 45 | 82 |
| Controls | 110 | 70 | 53 | 51 | |

.

Is there any difference between the two groups, and what is the P-value of the test?

8. The effectiveness of vitamin C in orange juice and in synthetic ascorbic acid was compared on 20 guinea pigs divided at random into two groups of ten each in terms in the length of the odontoblasts after 6 weeks with the following results:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Orange juice | 8.4 | 9.4 | 9.6 | 9.7 | 10.0 | 14.5 | 15.2 | 16.1 | 17.6 | 21.5 |
| Ascorbic acid | 4.2 | 5.2 | 5.8 | 6.4 | 7.0 | 7.3 | 10.1 | 11.2 | 11.3 | 11.5 |

.

Is there any difference between the effectiveness of vitamin C between orange juice and synthetic ascorbic acid at the 0.05 level of significance?

9. To test the effectiveness of vitamin $B_1$ in stimulating the growth of mushrooms, vitamin $B_1$ was applied to 13 mushrooms selected at random from a group of 24, while the remaining 11 mushroom received no such treatment. The weights of the muchrooms at the end of the period of observation were:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Controls | 18 | 14.5 | 13.5 | 12.5 | 23 | 24 | 21 | 17 | 18.5 | 9.5 | 14 | | |
| Treated | 27 | 34 | 20.5 | 29.5 | 20 | 28 | 20 | 26.5 | 22 | 24.5 | 34 | 35.5 | 19 |

Is Vitamin $B_1$ effective at stimulating growth in mushrooms? What is the P-value?

**2. The Irwin-Fisher Test.** Here we are concerned with the following general problem. Suppose there are two games, and in each there is an event $A$ that might or might not occur. Suppose that for the first game the probability of $A$ occurring is $p'$, and in the second game the probability of $A$ occurring is $p$". Let $X$ denote the number of times that the event $A$ occurs in $m$ plays of the first game, and let $Y$ denote the number of times that the event $A$ occurs when the second game is played $n$ times. In both cases the values of $m$ and $n$ are known, but the values of $p'$ and $p$" are unknown.. Thus $X$ is $Bin(m, p')$, and $Y$ is $Bin(n, p")$. The values of $X$ and $Y$ are observed, and suppose that you are interested in is deciding whether $p' = p$" or is $p' < p$"?

A problem of this sort arises frequently in testing a new medical treatment to see if it is as effective or more effective than a traditional treatment. The use of historical data is not the best way to test this new treatment in clinical trials involving humans. This is due to what is called the "placebo effect". Thus one must take two samples and try the traditional treatment on one sample of $m$ patients and the new treatment on the second sample of $n$ patients. The way one does this is to effectively assemble $m + n$ patients, then select $m$ of them at random without replacement. These are to be the control patients, i.e., the patients to receive the traditional treatment. The remaining $n$ patients serve as the treatment group, i.e., they receive the new treatment. However, no one knows which treatment he or she is getting, i.e., the trial is double blinded. So far we are no different from the two sample permutation test developed in the last section. However, here the data are categorical; this means in this case that the outcome on each patient is either success or failure. If we let $A$ denote the event that

a patient responds favorably to a treatment, then the number of patients who respond favorably to the traditional treatment, $X$, and the number that respond favorably to the new treatment , $Y$, are independent, (that is, the events $[X = x]$ and $[Y = y]$ are independent) and these random variables have distributions that are $Bin(m, p')$ and $Bin(n, p")$ respectively. In this case, if $\frac{X}{m}$ and $\frac{Y}{n}$ are close to each other, we would conclude that there is no difference between treatment and control and that $p' = p"$. This is because of Bernoulli's theorem. But if they differ substantially, we might wish to conclude that they are different and, if $\frac{Y}{n} > \frac{X}{m}$, then we might wish to conclude that the new treatment is better than the traditional treatment.

So suppose that we observe the values of $X$ and $Y$ and discover them to be $x$ and $y$. Further, suppose that the difference

$$\frac{y}{n} - \frac{x}{m}$$

is positive and somewhat large. If it is true that there is no difference, i.e., that $p' = p"$, then we might as well assume that we are playing just one game $m + n$ times. For each play, we shall record the number 1 if the event $A$ occurs and the number 0 if $A$ does not occur. So the data resulting from our plays are $x + y$ 1's and $m + n - x - y$ 0's.We continue in the same way as in the two sample simulation test of the previous section. Now again, we wish to take a simple random sample of size $m$ without replacement from the 1's and 0's, compute the number of 1's in it, call it $x'$, then let $y'$ denote the number of 1's in the $n$ remaining numbers, and compute

$$\frac{y'}{n} - \frac{x'}{m}.$$

If

$$\frac{y'}{n} - \frac{x'}{m} \geq \frac{y}{n} - \frac{x}{m},$$

then we shall count 1, otherwise 0. After doing this a large number of times, say $25,000$ times, we shall compute

$$\frac{\#\text{ of 1's}}{25,000}.$$

According to Bernoulli's theorem, this ratio should be close to

$$P([\frac{y'}{n} - \frac{x'}{m} \geq \frac{y}{n} - \frac{x}{m}])$$

**when the null hypothesis** that $p' = p$" is true. If this estimate of the probability is unbelievably small, we shall reject the null hypothesis that $p" = p'$ in favor of $p" > p'$. Again, the definition of an unbelievably small probability is subjective. It could be 0.05 or 0.03 or 0.01 or even smaller.

For appropriate software, use the program SIMIRFIS.EXE.

## Exercises

1. Mannitol and Decadron are drugs that are often administered to patients with severe head injury when they are admitted to the emergency room of a hospital. One suggestion is that this type of treatment would be more beneficial if administered by paramedics to patients in the field before they are transported to the hospital. So in order to plan such a study, a pilot study is performed with 10 patients receiving the drugs in the field before they are hospitalized and 10 patients receiving the drug only after they arrive at the hospital. After the study is carried out, it was observed that 4 out of 10 patients with field treatment and 6 out of another 10 with no field treatment die before being discharged from the hospital. Does this indicate that a larger study might show that the suggestion might have some merit?

2. In a 1985 study of the effectiveness of streptokinase in treatment of patients who have been hospitalized after myocardial infarction, 9 out of 199 males receiving streptokinase and 13 out of 97 males in the control group died within 12 months. Is there a difference in their mortality rates?

3. Heart patients routinely take aspirin because it reduces blood clots. Many doctors take their patients off the drug before surgery because of concerns of excess bleeding. A study at the Mayo Clinic between the years 2000 and 2002 was made on $1,636$ patients who underwent open heart surgery; $1,316$ had taken aspirin up to the time of surgery, and there were 320 patients who had stopped taking aspirin within 5 days of undergoing surgery.

($i$) It was reported that the mortality rate in the hospital for those taking aspirin was 1.7 percent, while for those not aking aspirin within 5 days of surgery was 4.4 percent. Was there a difference?

($ii$) Those taking aspirin up to the time of surgery "had a 3.5 percent risk of excess bleeding" compared to 3.4 percent of those not on aspirin. Any difference?

($iii$) Patients on aspirin "had a 2.7 percent chance of having a stroke after surgery, compared to 3.8 percent who were not on the drug". Was there a difference on the possibility of a stroke after surgery?

($iv$) How would you possibly fault the study?

($v$) If one stretches things, one could analyze these data by doing a binomial test for rare events. Do the three tests.

4. Out of 156 people suffering from carpal tunnel syndrome, 73 were selected at random and treated by surgery, and 83 were treated by splinting. In the surgery group, 67 were considered to have benefitted from the treatment, while in the splint group, 60 were considered to have benefitted. Is there a decisive difference between the two treatments?

5. In a study on malaria prevention in Kenya, among 343 infants who were provided with bednets, 15 caught malaria, while among 294 infants who were not provided with bednets, 27 caught malaria. What is the P-value of the test for no difference against the alternative that the malaria rate for no bednets is greater?

6. The Centers for Disease Control reported that out of a sample of $1,012$ men who were 65 years of age or older, 411 of them suffered from some form of arthritis. Out of a random sample of $1,062$ women 65 years of age or older, 511 suffered from some form of arthritis. Is the rate for men less than the rate for women?

**3. Test for Independence in a $2 \times 2$ Contingency Table.** We now consider a game in which there are two events, $A$ and $B$, each of which might or might not occur. We are concerned whether these two events are independent of each other. Recall that we developed the notion of independent events in section 3 of chapter 2, in which we began with the definition that these events are independent if $P(A \cap B) = P(A)P(B)$. We do not know any of these probabilities, but we are able to play this game $n$ times and observe with each play whether each event occurred or not. Suppose that after playing the game $n$ times, it is observed that in $X_{11}$ of these plays, both $A$ and $B$ occurred, that in $X_{21}$ of these plays of the game, $A$ occurred but $B$ did not occur, in $X_{12}$ of these games $A$ did not occur but $B$ occurred, and in $X_{22}$ of these $n$ games, neither $A$ nor $B$ occurred. This can be expressed by the following table:

$$\begin{array}{cccc} & A & A^c & \\ B & X_{11} & X_{12} & X_{1\cdot} \\ B^c & X_{21} & X_{22} & X_{2\cdot} \\ & X_{\cdot 1} & X_{\cdot 2} & n \end{array} .$$

Clearly , $\sum_{i=1}^{2} \sum_{j=1}^{2} X_{ij} = n$. Also, let $X_{i\cdot} = \sum_{j=1}^{2} X_{ij}$ denote the sum of

the $i$th row, $1 \leq i \leq 2$, and let $X_{\cdot j} = \sum_{i=1}^{2} X_{ij}$ denote the sum of the $j$th column, $1 \leq j \leq 2$, so that $\sum_{i=1}^{2} X_{i\cdot} = \sum_{i=1}^{2} X_{\cdot j} = n$.

At this point it is a good idea to go back and review the notion of independence in chapter 2. In the treatment of independence of two events in chapter 2, we first gave the formal definition, and then we proved a more intuitive condition that is equivalent to our definition for independence of two events. This stated that events $A$ and $B$ are independent if and only if $P(A|B) = P(A)$, which in turn is true if and only if $P(A|B^c) = P(A)$, where it is assumed that $P(A) > 0$ and $P(B) > 0$. All this should be kept in mind as we develop the test for independence.

Now suppose that this game is played $n$ times, and the numbers of times for the four disjoint outcomes are $x_{11}$, $x_{12}$, $x_{21}$ and $x_{22}$ respectively. The row sums become $x_{1\cdot} = x_{11} + x_{12}$ and $x_{2\cdot} = x_{21} + x_{22}$ for the first and second rows respectively. We reason loosely this way. Consider the sequence of $x_{1\cdot}$ plays of the game in which the event $B$ occurs. In each play of this sequence of plays, the event $A$ may or may not occur, and it occurs with some probability $p'$. Also, in the sequence of plays of the game when $B^c$ occurs, the event $A$ may or may not occur at each play, and the probability of it occurring is $p$". Now, if $A$ and $B$ occur independently of each other, the probability of $A$ occurring should not depend on whether it occurred when $B$ occurred or when $B$ did not occur. (Now look back at the previous paragraph.) If $A$ and $B$ are independent, we should have $p' = p$". So we are essentially provided with the number of times, $x_{11}$, that the event $A$ occurred in the $x_{1\cdot} = x_{11} + x_{12}$ plays in which $B$ occurred and the number of times, $x_{21}$, that the event $A$ occurred in the $x_{2\cdot} = x_{21} + x_{22}$ plays when $B$ did not occur. In order to test the hypothesis that $p' = p$", we shall need to use the Irwin-Fisher test given in section 2. Thus our data consist of $x_{11} + x_{21}$ 1's and $x_{12} + x_{22}$ 0's, and we do a two sample permutation or simulation test explained in section 1 of this chapter and as specialized in section 2 as the Irwin-Fisher test.

For appropriate software, one could use PERMMEAN.EXE or SIMIRFIS.EXE. Note that if you use the PERMMEAN.EXE program, you must create of data file with $n$ 0's and 1's. So it is better to use the program for the Irwin-Fisher test.

## Exercises

1. A 1979 study investigated the relationship between cigarette smoking and subsequent mortality in men with prior history of coronary disease. It

was found that 264 out of 1731 nonsmokers and 208 out of 1058 smokers had died in the five year period after the study began. Compare the mortality rates of the two groups to see if smoking and coronary disease are independent or not.

2. A study was conducted in Wales relating blood pressure and blood-lead levels. It was reported that out of 455 men with blood-lead levels $\leq 11\mu g/100$ ml , 4 had elevated systolic blood pressure levels, while 16 out of 410 men with blood-lead levels $\geq 11\mu g/100$ ml also had systolic blood pressure levels. So what can you conclude?

**3**. The Consumers Union led a study for the United States Department of Agriculture on the presence or absence of pesticide residues in foods. In the $19,514$ samples in which pesticide presence was determined, 29 were organically grown, while among the $7,184$ samples in which pesticide waws not found, 98 were organically grown. Is the presence or absence of pesticide independent or whether the food was organically grown?

**4**. In a study published in 2002, 450 patients who suffered from cardiac arrest were sampled. Among the 361 of these who did not suffer from depression, 67 died within 4 years. Among the remaining 89 patients, 26 died within 4 years. Are depression and early death independent?

**5**. In 1998 a San Diego reproductive clinic reported that among 245 patients treated, there were 42 live births among the 157 women under the age of 38, but there were 7 live births among those 38 years of age or older. Does success in the treatment depend on age?

**4. The Wilcoxon Rank-Sum Test.** This test attacks the same problem as that treated in section 1. Again there are two samples of data, say of sizes $m$ and $n$, and let us suppose that all the $m+n$ measurements are distinct, that is, no two are equal. The problem here is the same as the problem in section 1. Namely, if these observed numbers are from two treatments of two groups of people, do the treatments differ, or are the results the same.

Let us denote the outcomes of the first sample (on the control group) by $x_1, x_2, \cdots, x_m$ and the outcomes for the second sample (on the treatment group) by $y_1, y_2, \cdots, y_n$. Finally, let

$$z_1, z_2, \cdots, z_{m+n}$$

denote the set of all $m+n$ distinct numbers from the two samples lined up from smallest on the left end to largest on the right end, that is,

$$z_1 < z_2 < \cdots < z_{m+n}.$$

This is referred to as the pooled ordered sample. Thus there is exactly one number from one of the two the samples whose value is $z_1$, there is exactly one number from the two samples whose value is $z_2$, $\cdots$, and there is exactly one number from the two samples whose value is $z_{m+n}$. We shall say that $r$ is the **rank** of $x_1$ in the pooled ordered sample if $x_1 = z_r$. In other words, $x_1$ is the $r$th number from the left or smallest in the pooled ordered sample. Thus, if the two samples are

$$10, 25, 20, 30 \text{ and } 15, 21, 27, 19, 22,$$

the pooled ordered sample is

$$10, 15, 19, 20, 21, 22, 25, 27, 30.$$

The ranks of the numbers in the first sample are $1, 7, 4, 9$, and the ranks of the numbers in the second sample are $2, 5, 8, 3, 6$ respectively. (For example, the number 25 in the first sample has rank 7 because it is seventh from smallest in the pooled ordered sample.) If the the numbers in the first sample were smaller than the numbers of the second sample, it is clear that the ranks of these numbers in the pooled ordered sample would be small too, and the average of their ranks would be small; similarly, if all of the $x$'s in the first sample were larger than all the numbers in the $y$-sample, then their ranks would be larger. and the average of their ranks would be larger. If there is no difference between the two treatments that produce the two samples, we would expect that the $x$-values would be sprinkled uniformly within the pooled ordered sample, and the same statement can be made for their ranks.

Just for intellectual curiosity, we shall first consider the range of values of the sum of the ranks of the of $m$ numbers within a set of $m + n$ distinct numbers. The smallest that the sum of the ranks of the $x$-values can be is $1 + 2 + \cdots + m$, which from section 1 of chapter 1 is equal to

$$\frac{m(m+1)}{2}.$$

Thus their average rank is $\frac{1}{m}(1 + 2 + \cdots + m) = \frac{m+1}{2}$. The largest that the sum of the ranks of the $x$-values in the pooled, ordered sample could be is

$$(n + 1) + (n + 2) + \cdots + (n + m),$$

which is easily seen to be equal to

$$mn + \frac{m(m+1)}{2},$$

and the largest average rank of the $x$-values is $n + \frac{m+1}{2}$.

So let us assume that there is no difference between the two treatments. Then it is reasonable to believe that if the two treatments or populations were not different, the average rank of the $x$-values, denoted by $\overline{rank(x)}$, would be in the middle of the numbers $1, 2, \cdots, m+n$, and the same should hold true for the average rank of the $y$-values, denoted by $\overline{rank(y)}$. But if $\overline{rank(x)} - \overline{rank(y)}$ is sufficiently negative, we should want to find the probability that the difference of the average of $m$ numbers selected at random without replacement from $\{1, 2, \cdots, m+n\}$ and the average of the remaining numbers is equal to or less than $\overline{rank(x)} - \overline{rank(y)}$. If this probability is unbelievably small, then we would wish to state that there is a difference between the two populations from which the samples were taken. It again looks as if we have an opportunity to use the simulation test outlined in section 1 of this chapter but this time on the ranks of the observations.

It often occurs in data sets that there are ties, namely, two or more numbers in the pooled ordered sample are equal. So how shall we compute the ranks of each $x_i$ and each $y_j$ from the two samples. It seems reasonable to replace all those numbers that are equal by their average rank among those tied; this average rank is usually called a midrank or tied midrank. Here is an example. Suppose that the two samples are

21, 24, 26, 24 and 19, 24, 26.

Then the pooled ordered sample is

19, 21, 24, 24, 24, 26, 26.

Their tied midranks of the pooled ordered sample are

1, 2, 4, 4, 4, 6.5, 6.5.

Thus the midranks in the two samples are

2, 4, 5.5, 4 and 1, 4, 5.5.

The use of the Wilcoxon rank sum test as developed above might not be as accurate as the permutation test developed in section 1, but it can be of some use in the case where numbers are not available but where different qualities or classes of results can be observed. Let us consider the following type of problem to illustrate this. In a test of the effect of special psychological

counseling compared with ordinary psychological counseling, 80 boys were divided at random into two groups of 40 each. For the control group of 40 boys, only ordinary counseling was available, but for the treatment group, special counseling was made available. At the end of the study, a careful evaluation of the adjustment of each boy was made according to the following classifications: poor, fairly poor, fairly good and good. The data obtained were as follows:

| | Poor | Fairly poor | Fairly good | Good |
|---|---|---|---|---|
| Treatment | 5 | 7 | 16 | 12 |
| Control | 7 | 9 | 15 | 9 |

Since we have no numerical scores, we cannot do the permutation test of section 1, but we can use the tied ranks of the boys in the two groups.

First consider the treatment group. The ranks of the 12 boys of the pooled ordered sample among these lowest scorers of 5 and 7 boys would be $1, 2, \cdots, 12$, provided we could distinguish among them But we cannot, so we give each boy the value of the tied midrank

$$\frac{1 + 2 + \cdots + 12}{12} = 6.5.$$

Among the total of 16 in the fairly poor column, their average midrank is

$$\frac{13 + 14 + \cdots + 28}{16} = 20.5.$$

Among the 31 judged to be fairly good, the tied midrank turns out to be 44, and, among the 21 judged to be good , the tied midrank for this group is 70. Thus, in our data set that we provide for our permutation test, the treatment group will contain 40 numbers, 5 of them being the number 6.5, then 7 of them will be the number 20.5, then 16 of them will be the number 44, and 12 of them will be the number 70. Likewise, in our control group, 7 of them being the number 6.5, then 9 of them will be the number 20.5, then 15 of them will be the number 44, and 4 of them will be the number 70. Performing the two sample permutation or simulation test of section 1 of this chapter on these data gives us a P-value near 0.16.

So now a question arises. It looks as if we have three tests that can be performed for the two sample problem treated in section 1 of this chapter. Which test should be used? In section 1 this question was discussed for the

two tests given there. There is nothing wrong with doing all three tests on the data. If they all agree, then you have ample evidence for a decision. If they do not agree, then there are problems. For samples that are not too small in size, if the Wilcoxon rank sum test results in a determination of no difference, while the other two tests indicate a significant difference between the two samples, this might be an indication that the data in one of the samples might contain contaminated data.

For appropriate software, use SIMWLCXN.EXE, which is the same as PERMMEAN.EXE but it computes within the program the midranks of the observations.

**Exercises**

1. Here is a walk-through of the Wilcoxon rank-sum test performed on the problem at the end of section 1 that allows you to see what is going on. Let us suppose that there are two treatments, the standard treatment and a newly proposed treatment. It is desired to conduct a clinical trial with three patients in each group. While someone is out scouting for six patients who are all suffering from the same ailment and to the same degree, the statistician is selecting three numbers at random from $\{1, 2, 3, 4, 5, 6\}$, with it being decided ahead of time that the six patients are to be numbered from 1 to 6 according to the order in which their last names are alphabetized. So three numbers are selected at random without replacement from $\{1, 2, 3, 4, 5, 6\}$, and suppose they turn out to be 4, 5 and 1. The patients arrive and are numbered, and then patients with numbers 4, 5 and 1 are given the standard treatment, thus becoming the control group. The remaining patients, those numbered 2, 3 and 6, are given the new treatment and are called the treatment group. At the conclusion of the treatments, the $x$-values are measured; they turn out to be

| | |
|---|---|
| control group | 29, 41, 36 |
| treatment group | 42, 35, 51 |

($i$) Find the ranks of the control group data in the pooled ordered sample, and the ranks of the three observations in the treatment group.

($ii$) Compute $\overline{x}$, the arithmetic mean of the ranks of the control, then compute $\overline{y}$, the arithmetic mean of the ranks of the treatment group, and finally compute their difference $\overline{y} - \overline{x}$. Set these values aside for the moment.

(*iii*) List all $\binom{6}{3}$ choices of combinations of ranks of the $x$-values for the treatment group with the corresponding ranks of $y$-values for those remaining as the control group.

(*iv*) For each individual outcome listed in (*ii*), let $\overline{X}$ denote the arithmetic mean of the three numbers selected, and let $\overline{Y}$ denote the arithmetic mean of those remaining. Compute $\overline{X}$, $\overline{Y}$ and $\overline{Y} - \overline{X}$ for each of the 20 individual outcomes.

(*v*) Compute

$$P\left(\left[\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}\right]\right) = \frac{\#\{\text{i.o.'s} : \overline{Y} - \overline{X} \geq \overline{y} - \overline{x}\}}{\binom{6}{3}} \ .$$

(*v*) Now we verify the value of this probability experimentally. Use the random number generator on your hand-held calculator to select three numbers at random without replacement from $\{1, 2, 3, 4, 5, 6\}$, and compute their average, $\overline{X}$. Then compute the average, $\overline{Y}$, of the remaining numbers that are not in the sample. If $\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}$, then count 1; if $\overline{Y} - \overline{X} < \overline{y} - \overline{x}$, then count 0. Repeat this 100 times, and compute the number of $1'$s divided by 100. (Here is a hint: choose 1 if the random number generated is between 0 and 0.1666, choose 2 if the random number generated is between 0.1667 and 0.3333, choose 3 if the random number generated is between 0.3334 and 0.4999, etc. If you get a repetition, do not pay any attention to it.)

(*vi*) Using a desktop computer, simulate this 10,000 times. (Try PERMMEAN.EXE.)

2. Two teaching methods, A and B, were tried on two different groups of ten students per group, the same teacher doing the teaching to both groups. At the conclusion of the interval of instruction, they were given the same test at the same time. Here are the two sets of grades in percentages:

Method A: 50, 68, 90, 80, 72, 88, 66, 58, 81, 92

Method B: 60, 94, 100, 79, 83, 85, 64, 84, 96, 93.

Is there any significant difference between the two teaching methods? (Try SIMWLCXN.EXE.)

3. Prove: if $\overline{X}$ is the average of a sample of size $m$ taken at random without replacement from the numbers $\{1, 2, \cdots, m+n\}$, then $E(\overline{X}) = (m+n+1)/2$.

4. Suppose that in a two sample problem that the $(k+1)$th through the $(k+r)$th ranks are tied. Prove that the tied midrank for each of these $r$ observations is $\frac{1}{r}(kr + \frac{r(r+1)}{2})$.

5. Suppose that 20 treatment patients are being compared with 20 controls and that the progress of each patient is classified as very poor, poor, indifferent, good or very good. If the data are as given in the following table,

| | Very poor | Poor | Indifferent | Good | Very good |
|---|---|---|---|---|---|
| Control | 2 | 2 | 11 | 4 | 1 |
| Treatment | 0 | 1 | 9 | 7 | 3 |

find their tied midranks and test the null hypothesis that there is no difference between the treatment group and the control group against the alternative that there is a difference.

**5. Test for Change of Rate in a Sequence of Categorical Responses.** There are situations in which, in effect, a sequence of games is played in which a certain event might or might not occur. The probability of the occurence of this event might be constant or it might increase or decrease as one goes from one game to the next. As an example, consider a sequence of products as they come off an assembly line. Occasionally there will be a defective item. After a considerable period of time, the question might arise whether the rate at which defectives occur is increasing or not. Thus, items come off the production line in sequence, and after $m+n$ items have passed through inspection, it has been noted that $m$ of them tested as being defective. If these $m$ defectives occurred more frequently among the later items that were inspected, one would say that the incidence of defectives is increasing for some reason, so the production line would be shut down so as to determine the cause of the increased rate of defectives.

In order to analyze a problem such as this, one would consider a model of $m+n$ games being played independently of each other, in which a certain event $E$ occurred $m$ times. Let us denote $x_1, \cdots, x_m$ as the trial or game numbers at which the event occurred. If the sum, $\sum_{j=1}^{m} x_j$, of these trial numbers is "too large", then we might conclude that the probability was increasing from trial to trial, and similarly if the sum were "too small", we would tend to believe that the probabilities were decreasing. If the probability of $E$ in the $j$th trial did not change from trial to trial but remains a constant $p$, then the trial numbers at which $E$ occurred should be more uniformly spaced throughout the $m+n$ trial numbers. In this case the arithmetic mean (i.e., the average) of the trial numbers at which the event $E$ occurs is close to the arithmetic mean of the trial numbers at which the event $E$ does not occur. Thus it appears that we are back in the situation of the Wilcoxon rank-sum test. So applying that test here, we can test the null hypothesis

that the probability of $E$ does not change from trial to trial against the alternative that $P(E)$ increases in value during the $m+n$ plays or decreases in value during the $m+n$ plays.

### Exercises

1. An urn contains 10 tags numbered 1 through 10. Three tags are selected at random without replacement. What is the probability that the sum of the three numbers is not greater than 10?

2. In the 22 working days of a month in a large factory, there were accidents on days numbered 8, 16, 19 and 21. If 0.10 is an unbelievably small probability for you, can you conclude that the rate of accidents occurring was increasing?

3. At a certain large industrial plant that employs several thousand workers, a record is kept of all serious accidents that occur. The problem that confronted management at the beginning of a year was whether the rate or incidence of serious accidents increased during the course of the previous year or not. The accidents occured on the following dates: January 28, March 2, April 23, May 2, July 3, August 25, September 5, October 3, October 19, November 1, November 20, November 30, December 5 and December 20.(It was not a leap year.) So analyze the data to determine if the accident rate was increasing during the year.

## CHAPTER 6. CONFIDENCE INTERVALS

**1. Confidence Interval For $p$ in $Bin(n,p)$ For Large Values of $n$.** Suppose you observe the value $k$ of a random variable $X_n$ whose distribution is $Bin(n,p)$, where you know the value of $n$ but do not know the value of $p$. The problem is: based on the values of $k$ and $n$, one wishes to provide two numbers, $a<b$, so that you can declare that the value of $p$ is between $a$ and $b$. Also, one wishes to be able to explain how good this interval is in estimating $p$. Of course, you would want these two numbers to be as close together as possible.

Let us recall the Laplace-DeMoivre theorem that was given in section 3 of Chapter 3 which states that if $X_n$ is $Bin(n,p)$, then

$$P\left(\left[\frac{X_n-np}{\sqrt{np(1-p)}}\le x\right]\right)\to\Phi(x) \text{ as } n\to\infty.$$

If $n$ is large, which we now assume, then we may take the two quantities above as (approximately) equal. So, for simplicity, set, approximately,

$$P\left(\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leq x\right]\right) = \Phi(x).$$

Now we also observe that

$$P\left(\left[\frac{X_n - np}{\sqrt{np(1-p)}} \leq -x\right]\right) = \Phi(-x)\ .$$

Hence by an argument based on the symmetry of $\Phi(x)$, namely, upon observing that $\Phi(-x) = 1 - \Phi(x)$, we may write

$$P\left(\left[-x \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq x\right]\right) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1.$$

This is the important identity. Suppose we take $x = 1.96$. Then since $\Phi(1.96) = 0.975$, we obtain $2\Phi(x) - 1 = 0.95$, and thus we may state (approximately)

$$P\left(\left[-1.96 \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq 1.96\right]\right) = 0.95.$$

Now let us look at the double inequality

$$-1.96 \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq 1.96,$$

which occurs with probability 0.95. This is the same as

$$-1.96\sqrt{np(1-p)} \leq X_n - np \leq 1.96\sqrt{np(1-p)},$$

or, upon further arranging,

$$\frac{X_n}{n} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \frac{X_n}{n} + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

So we may take

$$a = \frac{X_n}{n} - 1.96\sqrt{\frac{p(1-p)}{n}} \text{ and } b = \frac{X_n}{n} + 1.96\sqrt{\frac{p(1-p)}{n}},$$

we may safely say (approximately) that $P([a \leq p \leq b]) = 0.95$, where $a$ and $b$ are both random variables that depend on $X_n$ (fortunately) and on $p$ (unfortunately). If we replace $a$ by something smaller and $b$ by something larger, say, by $a^* \leq a$ and $b^* \geq b$ respectively, and which do not depend on $p$, then we may state that $P([a^* \leq p \leq b^*]) \geq 0.95$.

We accomplish this as follows. The quantity $p(1-p)$ is equal to 0 when $p = 0$ and when $p = 1$, but between 0 and 1 it is positive. We wish to find its largest value. Notice that

$$p(1-p) = \frac{1}{4} - (\frac{1}{2} - p)^2.$$

The largest that $p(1-p)$ can be is when $(\frac{1}{2} - p)^2$ is the smallest. Since this latter quantity is nonnegative, it achieves its smallest value when $p = \frac{1}{2}$. Thus the largest value that $p(1-p)$ can achieve is $\frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4}$. So if we take

$$a^* = \frac{X_n}{n} - 1.96\sqrt{\frac{1}{4n}} \text{ and } b^* = \frac{X_n}{n} + 1.96\sqrt{\frac{1}{4n}},$$

we can now state that $P([a^* \leq p \leq b^*]) \geq 0.95$, as promised above.

So the statistician computes the values of $a^*$ and $b^*$ for someone who brings to him or her a value of $n$ and an observed value, $k$, of $X_n$. The explanation that should be given is that the value of $p$ for this particular set of data might or might not be between these two values $a^*$ and $b^*$. However, it is comforting to know that among the many, many times he or she uses these formulas, the correct value of $p$ is between $a^*$ and $b^*$ that is computed at each occasion at least 95 percent of the time; this is according to Bernoulli's theorem.

If one is interested in any other coverage probability than 0.95, one can go to the normal distribution and find the appropriate value of $x$. For suitable software, go to EXCEL and use NORMINV. For example, suppose one is interested in estimating the proportion $p$ of voters in a well-defined large population who favor candidate A in a forthcoming election. If one selects a voter at random from this population and observes that the voter selected is or is not for candidate A, this is the same as playing a game in which an event [favors candidate A] can occur with probability $p$. Now suppose we sample this population $1,000$ times; that is, we select $1,000$ voters at random, and observe a value of $X_{1000}$, which has the $Bin(1000, p)$ distribution. And suppose we wish to find a confidence interval for $p$ which covers $p$ with

probabliity 0.96, in other words, to find what statisticians call a 96% confidnece interval for $p$. So we go to EXCEL and find NORMINV. For the first entry, where it asks for probability, we enter 0.98 in accord with what was explained above. For mean, we enter 0 always, and for standard_dev, we enter 1 always. The value of $x$ that we get is 2.054. Thus the 96% confidence interval for $p$ that we shall use is

$$a^* = \frac{X_{1000}}{1000} - 2.054\sqrt{\frac{1}{4000}} = 0.42 \text{ and } b^* = \frac{X_{1000}}{1000} + 2.054\sqrt{\frac{1}{4000}} = 0.48.$$

So we go out and sample the voting population 1000 times, and suppose 450 of our respondents declare that they are for candidate A. Substituting 450 for $X_{1000}$ in the two formulas obtained above, we may state that the value of $p$ is between 0.48 and 0.42.

**Exercises**

1. In a sample survey, a sample of size 900 was taken, and it turned out that 550 of the respondents said YES to a certain issue that was presented to them. Find the 95 percent confidence interval for the proportion, $p$, of the entire population who would say YES to this issue.

2. In problem 1, find the 92 percent confidence interval for $p$.

3. (Pre-problem explanation: In using NORMDIST in EXCEL, in order to determine the value of $\Phi(x)$ for any number $x$, enter the value of $x$ in the top box, enter 0 in the box labelled "mean", enter 1 in the third box labelled "Standard_dev", and enter TRUE in the remaining box.) Evaluate $\Phi(1.00)$, $\Phi(0)$, $\Phi(2)$, $\Phi(-1)$ and $\Phi(-2)$.

3. Use Excel to evaluate $x$ when $\Phi(x) = 0.995$.

4. Use Excel to evaluate $x$ when $\Phi(x) = 0.90$.

5. Use Excel to evaluate $x$ when $\Phi(x) = 0.80$.

6. In problem 1, find the 90 percent confidence interval for the proportion $p$.

7. In problem 1, find the 99 percent confidence interval for the proportion $p$.

8. In problem 1, when you give the sponsor of the survey the two end values of the confidence interval for $p$, and if he or she asks what those two numbers mean, what should be your response?

9. In a sample survey involving $1,500$ people selected at random, 820 responded favorably to a certain issue. Find a 90 percent confidence interval

for the proportion of all people in the defined population who would respond favorably to this issue.

**2. Confidence Interval For $p$ in $Bin(n,p)$ for Small Values of $n$.** We consider the same problem now as we did in section 1, but here we have small samples, say, of size 1000 or less. Again, we may observe the value of a random variable $X$ whose distribution is $Bin(n,p)$. It is assumed that we know the value of $n$ and are able to observe the value of $X$. The problem is to use these two pieces of information to find an interval of numbers which should contain the value of $p$ according to some method of defining our confidence. In this case, when the sample is small, in particular, less than 100 or perhaps a little more, one sacrifices accuracy if one were to use the Laplace-DeMoivre theorem. There is a more exact and conservative method in the case of small values of $n$.

The method of attack is this. We first decide on a rather small number, $\alpha$, which for concreteness can be 0.025 or 0.05. We observe the value of $X$ to be $k$, where $k$ is an integer between 0 and $n$ (which we know). So we believe that $k$ should be close to the expectation of $X$ which is $np$. Hence we would like to think that the value of $p$ is near$\frac{k}{n}$. Suppose we consider a candidate value of $p$, call it $p'$, which is less than $\frac{k}{n}$ and ask the following question: if $Z$ is a random variable whose distribution is $Bin(n,p')$, what is the probability that $Z$ takes a value as extreme as $k$, i.e., $P([Z \geq k])$? This probability will either be larger than $\alpha$ or it will be smaller than $\alpha$. If it is larger than $\alpha$, then we should wish to try again with a value of $p'$ that is smaller than the one we just chose. Similarly, if it is smaller than $\alpha$, it means that we selected our value of $p'$ to be too small, and we should try again with a slightly larger value of $p'$, and repeat the process. After a few tries, one should be able to zero in on the value of $p'$ with any degree of accuracy that we wish so that we could truly state that $P([Z \geq k]) = \alpha$. Then we do almost the same thing by selecting a candidate value $p''$ of $p$ that is greater than $\frac{k}{n}$ and ask ourselves this question: if $W$ is a random variable whose distribution is $Bin(n,p'')$, what is the probability that $W$ takes a value as extreme as $k$, i.e., $P([W \leq k])$? If this probability is larger than $\alpha$, then we have taken the value of $p$" to be too small, and we try again with a larger value for $p''$; if it is smaller than $\alpha$, then we try again with a smaller value of $p''$. After a few tries, we should be able to zero in on the value of $p''$ with any degree of accuracy we wish so that we could state that $P([W \leq k]) = \alpha$ .

Having found the values of $p'$ and $p$", we then would like to state that whatever the correct value of $p$ might be, the value of $P([p' \leq p \leq p"]) \geq$

$1-2\alpha$. A word of interpretation is due here. First of all, the quantities $p'$ and $p$" are random variables that are functions of $X$ whose value we observed to be $k$. Second, this procedure has been known for over a half century and has been called by some "the statistical method". However, the statement that

$$P([p' \leq p \leq p"]) \geq 1-2\alpha$$

has not (to my knowledge) been proved until recently. It appears in a research paper by M. Finkelstein, H. G. Tucker and J. A. Veeh that was published in the journal Communications in Statistics-Theory and Methods, Vol 29(8), pages 1911-1928, (2000). The proof of the above inequality is beyond the scope of this course.

An interpretation of this is again as follows. The statistician can make this claim: Among the many, many cases in which I am asked to provide what is called a $100(1-2\alpha)$ percent confidence interval, and if each time I compute $p'$ and $p$" as outlined above, then the event $[p' \leq p \leq p"]$ will occur in at least $100(1-2\alpha)$ percent of those cases.

For appropriate software, use BINOMDIST in EXCEL.

Here is an example of how to find such a confidence interval. Suppose you play a certain game 85 times, and a possible outcome of this game, an event $E$, occurs in 29 of these 85 plays. The problem is to find a 90 percent confidence interval for the probability $p = P(E)$ of this game. Since EXCEL does not have a statistical program which might be called BINOMDISTINV or BINOMINV, we shall have to do some guessing. Since the relative freqency of the occurrence of $E$ is $\frac{29}{85} = 0.34$, Bernoulli's theorem suggests that we start with this number. What we are after is a value of $p' < p$ such that if $Z$ is $Bin(85, p')$, then $P([Z \geq 29])$ is close to, but equal to or less than, 0.05, or what amounts to the same thing, $P([Z \leq 28])$ is close to but equal to or greater than 0.95. Using what was suggested above in order to find a smallest possible value for $p$, let us select as a candidate $p' = 0.30$. So go to BINOMDIST in EXCEL. In the entry after Number_s, type 28. For Trials, enter 85. After Probability_s, enter 0.30, and after Cumulative, enter TRUE. The answer we get is $P([Z \leq 28]) = 0.764$, which is not large enough. So we lower the entry after Probability_s to $p' = 0.27$, and we get $P([Z \leq 28]) = 0.91$, which is still not quite large enough. After several more up and down guesses, we obtain: if $p' = 0.252$, then $P([Z \leq 28]) = 0.950$, rounded off. Similarly, starting with a guess of $p'' = 0.42$, we get $P([Z \leq 29]) = 0.037$. After a few more trials we get: if $p'' = 0.422$, then $P([Z \leq 29]) = 0.051$.

The confidence interval that one announces for $p$ here is: $0.252 \leq p \leq 0.422$.

At this point the person who requested the 90% confidence interval will sometimes ask for an interpretation and will suggest that he or she may state that he or she is 90% confident that $p$ lies between the values of 0.252 and 0.422. The answer that should be provided to them is that such a statement is totally meaningless. Another person might even be so bold as to suggest, "May I now state that $P([0.252 \leq p \leq 0.422]) \geq 0.90$?" The answer here is NO, since the inequality $0.252 \leq p \leq 0.422$ is either true with probability 1 or is false with probability 1. So then you should explain, keeping Bernoulli's theorem in mind, that as you go through life computing 90% confidence intervals for people, in about 90% of those times the inequality that you provide will be correct, and in about 10% of the cases, you will be incorrect. This is all that can be said.

## Exercises

1. Suppose a game is played 80 times in which a certain event $E$ occurred 47 times. The probability of $E$ is some unknown number $p$, and the problem is to find a 95 percent confidence interval for $p$. Use EXCEL to find the interval. Here are some hints. If I guess that $p'' = .75$, then if $X$ is $Bin(80, 0.75)$, we obtain $P([X \leq 47]) < 0.025$, and if $X$ is $Bin(80, 0.60)$, then $P([X \leq 47]) > 0.025$. Try it out and see if this is right. So first find $p''$ so that $P([X \leq 47]) = 0.025$ (accurate to three decimal places). Then to find $p'$, it should satisfy the equation

$$P([X \geq 47]) = 1 - P([X \leq 46]) = 0.025$$

to three place decimal accuracy, where in this case $X$ is $Bin(80, p'')$. Check out $p'' = 0.4$ and $p'' = 0.5$ first, and then aim for greater accuracy.

2. Suppose you observe that when a game is played 75 times, a certain event occurs 50 times. Find a 90 percent confidence interval for the probability of the event.

3. Suppose you observe that when a game is played 65 times, a certain event occurs 50 times. Find an 80 percent confidence interval for the probability of the event.

4. In problem 2, find an 80% confidence interval, then a 70% confidence interval, and then a 60% confidence interval for the probability of the event.

**3. Confidence Interval for the Median of a Population.** A median of some data producing phenomenon is a number $\mu$ such that if $X$ is an

observation on this data producing phenomenon, then $P([X \geq \mu]) \geq \frac{1}{2}$ and $P([X \leq \mu]) \geq \frac{1}{2}$. A median therefore is, in a way, a center for the data producing phenomena. From time to time the occasion arises where one has data $x_1, x_2, \cdots, x_n$ on a sample taken from a data producing phenomenon, and he or she wishes to find a $100(1-\alpha)$ percent confidence interval for $\mu$, where $\alpha$ is a small, positive number like 0.05, 0.10, or even in some cases as large as 0.20. We shall obtain a method for doing so for the time being in a simpler case, namely, when $X$ is a random variable for which there is a number $\mu$ such that $P([X > \mu]) = \frac{1}{2}$ and $P([X < \mu]) = \frac{1}{2}$ and where the data are distinct, so that for all intents and purposes, $P([X_i = x]) = 0$ for every real number $x$.

It should be obvious that the number of observations that are less than $\mu$ is a random variable whose distribution is $Bin(n, 0.5)$. We shall need the following proposition.

**Proposition 1**: Let $Z$ be a random variable whose distribution is $Bin(n, 0.5)$. If $k$ is as defined as the largest integer $k$ such that $P([Z \leq k]) \leq \frac{\alpha}{2}$.above, then $P([k < Z < n-k]) \geq 1-\alpha$.

Proof: Referring to the formula for $P([X = k])$ where $Z$ has the $Bin(n, \frac{1}{2})$ distribution, we readily observe, for $0 \leq i \leq n$, that

$$P([Z = i]) = P([Z = n-i]).$$

Thus, by the definition of $k$, it follows that $P([Z > n-k]) = P([Z < k])$, and each is equal to or less than $\frac{\alpha}{2}$. Now

$$P([Z \leq k-1]) + P([k < Z < n-k]) + P([\dot{Z} \geq n-k+1]) = 1,$$

so

$$P([k < Z < n-k]) \geq 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha,$$

which proves the proposition.

We are now able to find a confidence interval for $\mu$. If the sample is $x_1, x_2, \cdots, x_n$, and if $x_{(i)}$ denotes the $i$th number from the smallest, then the $100(1-\alpha)$ percent confidence interval for $\mu$ is the interval of numbers greater than $x_{(k)}$ and less than $x_{(n-k)}$, where $k$ is as obtained in the above proposition, and $x_{(k)}$ and $x_{(n-k)}$ are the $k$th and $(n-k)$th order statistics respectively for the data set. This can be verified as follows. Let $x_0 = -\infty$. Then for $0 \leq i \leq n-1$, the event

$$[x_{(i)} < \mu < x_{(i+1)}]$$

is the event that exactly $i$ of the $n$ observations are less than the median $\mu$. Thus if $k$ is as defined, we have

$$P([x_{(k)} < \mu < x_{(n-k)}]) = \sum_{i=k}^{n-k-1} P([x_{(i)} < \mu < x_{(i+1)}]).$$

If we let $x_0 = -\infty$, then by the proposition above, we have

$$1 - 2\sum_{i=0}^{k-1} P([x_{(i)} < \mu < x_{(i+1)}]) \geq 1 - \alpha.$$

For example, suppose our set of data is

$$\begin{array}{lllll} x_1 = 15.2 & x_2 = 16.1 & x_3 = 15.3 & x_4 = 16.4 & x_5 = 15.7 \\ x_6 = 14.7 & x_7 = 14.3 & x_8 = 14.5 & x_9 = 15.6 & \end{array},$$

and we wish to find an 80 percent confidence interval of the median $\mu$. The values of the order statistics are

$$\begin{array}{lllll} x_{(1)} = 14.3 & x_{(2)} = 14.5 & x_{(3)} = 14.7 & x_{(4)} = 15.2 & x_{(5)} = 15.3 \\ x_{(6)} = 15.6 & x_{(7)} = 15.7 & x_{(8)} = 16.1 & x_{(9)} = 16.4 & \end{array}.$$

Since

$$\sum_{j=0}^{2} \binom{9}{j} (\frac{1}{2})^9 = 0.0898 \text{ and } \sum_{j=0}^{3} \binom{9}{j} (\frac{1}{2})^9 = 0.2539$$

and

$$\sum_{j=7}^{9} \binom{9}{j} (\frac{1}{2})^9 = 0.0898 \text{ and } \sum_{j=6}^{9} \binom{9}{j} (\frac{1}{2})^9 = 0.2539,$$

we may say that the interval from $x_{(2)}$ to $x_{(7)}$, i.e., all numbers strictly greater that 14.5 but strictly less than 15.7, is the 82 percent confidence interval.

When the value of $n$ is large so that $k$ cannot be calculated directly, we obtain the value of $k$ as follows. We first find the value of the number $x$ so that

$$\Phi(x) = \frac{\alpha}{2}$$

as follows. Then let $X_n$ be a random variable whose distribution is $Bin(n, \frac{1}{2})$. Our problem is to find the number $k$ such that

$$P([X_n \leq k]) = \frac{\alpha}{2}.$$

This is the same as finding the value of $k$ such that

$$P\left(\left[\frac{X_n - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \leq \frac{k - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right]\right) = \frac{\alpha}{2}.$$

By the Laplace-DeMoivre theorem, for large values of $n$, the left hand side is very close to the value of $\Phi(t)$ where

$$t = \frac{k + 0.5 - \frac{n}{2}}{\sqrt{\frac{n}{4}}}.$$

So let $x$ denote the number that satisfies $\Phi(x) = \frac{\alpha}{2}$. Then set

$$\frac{k + 0.5 - \frac{n}{2}}{\sqrt{\frac{n}{4}}} = x,$$

and, solving for $k$, we obtain

$$k = 0.5 + \frac{1}{2}(x\sqrt{n} + n).$$

Since the number on the right is rarely an integer, we take $k$ to be $[0.5 + (x\sqrt{n} + n)/2]$, where $[y]$ denotes the largest integer less than or equal to $y$.

For appropriate software, go to EXCEL and use either BINOMINV or NORMDIST, depending on how large $n$ is.

## Exercises

1. Prove: If $0 \leq k < n$, then

$$\binom{n}{k} = \binom{n}{n-k}.$$

2. Find the 80% confidence interval for the median of a random variable in the case that 15 independent observations on it are as follows:

21.3, 24.7, 26.2, 22.7, 20.6, 21.8, 23.9, 24.5, 26.7, 25.3, 28.3, 20.9, 25.1, 26.3, 27.1.

3. From what has been developed in this section, the problem of obtaining a confidence interval for the median only secondarily depends on the values of the observations. The most important problem is to find which order statistics are the left and right ends of the confidence interval. So suppose you have a data set consisting of $1,000$ observations. Then find the pairs of order

statistics that determine the endpoints of the 90 percent confidence interval of the median and the endpoints that determine the 80 percent confidence interval of the median.

4.. Repeat problem 3 when the number of observations is 90.

## CHAPTER 7. HYPOTHESIS TESTING FOR MORE THAN TWO SAMPLES

**1. A Simulation Alternative to the One Way Analysis of Variance Test.** In chapter 5 we dealt with two samples that usually arise when one wishes to determine whether two treatments are the same or are different. It sometimes occurs that one has data for more than two treatments, and one wishes to test whether there are are no differences among them or whether there are differences among them, and if so, at least to determine for which pairs of samples differences do occur. The following treatment is an alternative to the one way analysis of variance test which is also applied to such a problem.

Let us suppose we have the outcomes of three samples for three treatments. Let us say that for $n_1$ observations on treatment number 1 the data are $x_{11}, x_{12}, \cdots, x_{1n_1}$, for another treatment they are $x_{21}, x_{22,} \cdots, x_{2n_2}$ and for treatment number 3 they are $x_{31}, x_{32}, \cdots, x_{3n_3}$. If there were no differences among them, then the three sample means, $\overline{x}_{1\cdot}$, $\overline{x}_{2\cdot}$ and $\overline{x}_{3\cdot}$ should all be close to each other and also close to the arithmetic mean, $\overline{x}_{\cdot\cdot}$, defined by

$$\overline{x}_{\cdot\cdot} = \frac{1}{n_1 + n_2 + n_3} \sum_{i=1}^{3} \sum_{j=1}^{n_i} x_{ij},$$

of the pooled sample,

$$x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, x_{22,} \cdots, x_{2n_2}, x_{31}, x_{32}, \cdots, x_{3n_3}.$$

Again, if there were no differences among the three treatments, we would expect the absolute values of the three differences $\overline{x}_{1\cdot} - \overline{x}_{\cdot\cdot}$ and $\overline{x}_{2\cdot} - \overline{x}_{\cdot\cdot}$ and $\overline{x}_{3\cdot} - \overline{x}_{\cdot\cdot}$ to be small, so their squares would be small, and even the sum of their squares would be small. This sum of squares of differences appears as

$$SS = \sum_{i=1}^{3} (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2.$$

In order to test whether this value of $SS$ is not unusual or extreme, we shall take a random permutation,

$$x_{11}^{'}, x_{12}^{'}, \cdots, x_{1n_1}^{'}, x_{21}^{'}, x_{22,}^{'} \cdots, x_{2n_2}^{'}, x_{31}^{'}, x_{32}^{'}, \cdots, x_{3n_3}^{'},$$

of the pooled set of data and shall compute

$$SS' = \sum\nolimits_{i=1}^{3} (\overline{x}_{i\cdot}' - \overline{x}_{\cdot\cdot})^2,$$

where

$$\overline{x}_{i\cdot}' = \frac{1}{n_i} \sum\nolimits_{j=1}^{n_1} x_{ij}', 1 \leq i \leq 3,$$

and where $n = n_1 + n_2 + n_3$. In this case of no pairwise differences, the value of the sum of squares $SS'$should take a wide variety of values, of which the value of $SS$ should be some "usual value". As with our two-sample simulation test, we might wish to see if $SS$ is too large. So we wish to evaluate

$$P([SS' \geq SS]).$$

This can be done by simulation in much the same way as we did for the two sample simulation test. Simply take a random permutation of the pooled data set,

$$x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, x_{22,} \cdots, x_{2n_2}, x_{31}, x_{32}, \cdots, x_{3n_3},$$

a large number of times. Each time this is done it should be noted whether the value of $SS'$ is as extreme as the value of $SS$, i.e., whether the event $[SS' \geq SS]$ occured. After a large number of simulations, one should find the ratio of the number of times that this event occurred divided by the total number of simulations. If this ratio (which, by Bernoulli's theorem, is just about equal to the probability $P([SS' \geq SS])$ if there were no differences) is unbelievably small, then we would reject the null hypothesis that there are no differences among the treatments in favor of the conclusion that the treatments of at least one pair of treatments differ from each other.

Unfortunately, if one rejects the null hypothesis, one cannot completely determine which data pairs come from populations that do not differ and those that do. But one may conclude that the pair whose arithmetic means differ the most are different.

For appropriate software, use SIMANOVA.EXE.

## Exercises

1. Three different detergent brands, A, B and C, are tested for the amount of dirt removed from a standard household load of laundry. Here are the results:

A: $11, 13, 17, 17, 15, 16, 14, 10, 12, 14$

B: $12, 14, 17, 19, 21, 18, 19, 18, 16, 18$

C: $18, 16, 18, 20, 22, 15, 17, 21, 16, 20$.

the P-value in testing for differences.

2. A sputtering machine is used for metalization on wafers in the semiconductor industry. The reflectances of different sputtering machines (larger numbers are better) were compared, with the following results:

Machine 1: $88.80, 90.20, 91.30, 89.50, 90.30$

Machine 2: $90.20, 91.70, 90.00, 90.90, 92.50$

Machine 3: $94.80, 93.50, 90.90, 94.20, 94.10$. Are

there any differences among the three machines if 0.05 is considered an unbelievably small probability?

3. A dietician wishes to determine which factors in the diet lead to increased weight. Seven different diets are fed to mice, yielding the following weights in grams:

$$
\begin{aligned}
&\textit{Diet } 1: 110, 113, 108, 103, 119\\
&\textit{Diet } 2: \ 90, 109, 98, 95, 115\\
&\textit{Diet } 3: \ 104, 122, 121, 116, 188, 140, 115\\
&\textit{Diet } 4: \ 108, 109, 118, 123 \qquad .\\
&\textit{Diet } 5: \ 114, 131, 111, 130, 134, 121\\
&\textit{Diet } 6: \ 127, 114, 119, 122, 132\\
&\textit{Diet } 7: \ 108, 109, 118, 117, 124
\end{aligned}
$$

Are there any significant differences among the diets with respect to weight?

4. In the following there are four sets of eight measurements each of the smoothness of a certain type of paper , obtained from four different laboratories:

| Laboratory | |
|---|---|
| A | $38.7, 41.5, 43.8, 44.5, 45.5, 46.0, 47.7, 58.0$ |
| B | $39,2, 39.3, 39.7, 41.4, 41.8, 42.9, 43.3, 45.8$ |
| C | $34.0, 35.0, 39.0, 40.0, 43.0, 43.0, 44.0, 45.0$ |
| D | $34.0, 34.8, 34.8, 35.4, 37.2, 37.8, 41.2, 42.8$ |

Are there any differences among the four laboratories if your largest probability of an unlikely event is $0,03$?

**2. Multiple Comparisons: The Tukey Decider.** In section 1 we obtained a test for equality of several treatments. But a new question arises: if we reject the null hypothesis that there are no differences, which pairs are unequal and which are essentially the same? For example, in the case of three samples considered in section **1**, if we reject the null hypothesis, it might be due to the fact that the response to treatment #2 is larger that either of #1 or of #3, but that the responses to #1 and #3 are indistinguishable from each other. Thus we need a decision method by which to order the responses of the treatment. At the same time we shall obtain a rather different test of hypothesis.

Again,let us suppose we have the outcomes of three samples for three treatments. Let us say that for $n_1$ observations on treatment number 1 the data are $x_{11}, x_{12}, \cdots, x_{1n_1}$, for treatment number 2 they are $x_{21}, x_{22,} \cdots, x_{2n_2}$ and for treatment number 3 they are $x_{31}, x_{32}, \cdots, x_{3n_3}$. If there were no differences among them, then the three sample means, $\overline{x}_{1\cdot}$, $\overline{x}_{2\cdot}$ and $\overline{x}_{3\cdot}$ should all be close to each other, and the absolute values of all differences, $|\overline{x}_{1\cdot} - \overline{x}_{2\cdot}|$, $|\overline{x}_{1\cdot} - \overline{x}_{3\cdot}|$ and $|\overline{x}_{2\cdot} - \overline{x}_{3\cdot}|$ should be small. Thus, the maximum of all these absolute values,

$$M = \max\{|\overline{x}_{1\cdot} - \overline{x}_{2\cdot}|, |\overline{x}_{1\cdot} - \overline{x}_{3\cdot}|, |\overline{x}_{2\cdot} - \overline{x}_{3\cdot}|\},$$

should be small. The value of $M$ should be relatively large if and only if the null hypothesis of no differences among them is false.

Let us recall a usual way in which the candidates for the treatments are chosen. One might decide in advance the number $n_i$ of subjects for treatment $i$, $1 \leq i \leq 3$. After obtaining all the subjects and pooling them to form a set of $n = n_1 + n_2 + n_3$ subjects, one selects $n_1$ of them at random without replacement to receive treatment #1. Then, from the remaining $n_2 + n_3$ subjects, one selects $n_2$ of them at random without replacement for treatment #2, and those not so far selected receive treatment #3. The value of $M$ as defined above is then computed. It is a number, and if the null hypothesis is true, the value of $M$ should be relatively small.

The protocol presented above suggests doing the following. Let

$$x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, x_{22,} \cdots, x_{2n_2}, x_{31}, x_{32}, \cdots, x_{3n_3}.$$

denote the pooled sample of measurements. Then let

$$x_{11}^{'}, x_{12}^{'}, \cdots, x_{1n_1}^{'}, x_{21}^{'}, x_{22,}^{'} \cdots, x_{2n_2}^{'}, x_{31}^{'}, x_{32}^{'}, \cdots, x_{3n_3}^{'},$$

denote a random permutation of the above numbers, and let

$$M' = \max\{|\overline{x}'_{1.} - \overline{x}'_{2.}|, |\overline{x}'_{1.} - \overline{x}'_{3.}|, |\overline{x}'_{2.} - \overline{x}_{3.'}|\}$$

denote the random variable formed by taking the maximum value of the absolute values of all differences of the arithmetic means. If the null hypothesis is true, then all these permutations are equally likely, and $M$ should be a usual value of the random variable $M'$.

So we need to find the value of $P([M' \geq M])$. This is the P-value of the test. If this probability is unbelievably small, then we would reject the null hypothesis that the set of all permutations of

$$x_{11}, x_{12}, \cdots, x_{1n_1}, x_{21}, x_{22,} \cdots, x_{2n_2}, x_{31}, x_{32}, \cdots, x_{3n_3}.$$

are equally likely and would tend to believe that at least one treatment is better than all the others. Suppose that this is the case. In order to compare the treatments, we should select a small probability, $\alpha$, to be the largest probability with which we are willing to reject the null hypothesis **if the null hypothesis is true**. So if $\alpha > 0$ is a small number (like 0.05 or 0.01), we would wish to find the smallest number $x_\alpha > 0$ such that

$$P([M' \geq x_\alpha]) \leq \alpha.$$

We then find those pairs of treatments for which the absolute values of the difference of sample means is as large or larger than $x_\alpha$,

$$\{(i,j) : 1 \leq i \ < j \leq 3, |\overline{x}_{i.} - \overline{x}_{j.}| \geq x_\alpha\},$$

and declare that these are the pairs of treatments that differ. (The direction in which they differ is obvious.) A computer program for doing this is TUKEYSIM.EXE.

One disconcerting note is that one might not get a consistent ordering of the treatment effects. In this case, one must judge the decisiveness of any difference in means indicated by the Tukey test outlined above.

Let us consider an example. It is wished to determine if there are any differences among three different methods of teaching fifth grade mathematics. A group of 15 typical fifth graders was assembled and randomly divided into

three groups with five children in each group. Each group was taught by a different one of the methods, and at the end of a fixed period of time they were all given the same examination. Their examination grades are tabled as follows.

| | | | | | |
|---|---|---|---|---|---|
| Method 1 | 48 | 73 | 51 | 65 | 87 |
| Method 2 | 55 | 85 | 70 | 69 | 90 |
| Method 3 | 84 | 68 | 95 | 74 | 67 |

The sample means for the three groups are, respectively, $\overline{x}_{1\cdot} = 64.8$, $\overline{x}_{2\cdot} = 73.8$ and $\overline{x}_{3\cdot} = 77.6$. Using the program SIMANOVA.EXE as outlined in section 1 of this chapter, we obtain a P-value of approximately 0.35. If one uses TUKEYSIM.EXE as outlined above, one gets a P-value of approximately 0.348. If we choose $\alpha = 0.05$ and simulate the value of $M'$ about 5000 times, one obtains an approximate value of $x_\alpha$ as defined above to be approximately 20.2. Since the absolute values of all the pairwise differences of the sample means are all less than 20.2, we have a confirmation that there are no differences in the three different teaching methods.

**Exercises**

1. In an investigation into 3 different treatments, 8 individuals were selected. In a random manner, 3 were selected from the 8 for treatment #1, 2 were selected at random from the remaining 5 for treatment #2, and the remaining 3 were given treatment #3. In how many ways can this be done?

2. In a two sample problem (as in Chapter 5) with $m$ individuals in the $x$-sample and $n$ in the $y$-sample, suppose that $K > 0$ is the number in the range of $\overline{x}' - \overline{y}'$ that satisfies

$$P([\overline{x}' - \overline{y}' \geq K]) = 0.05.$$

However, in the original data, suppose that $\overline{x} - \overline{y} > K$. Which of the following does the P-value of the permutation test, PERMMEAN.EXE, satisfy: $P > 0.05$. $P = 0.05$ or $P < 0.05$, and why?

3. Suppose you are dealing with a four sample problem. Let $K > 0$ be the number obtained from TUKEY.EXE that allows you to state roughly that $P([M' \geq K]) = 0.05$, and let $\overline{x}_{1\cdot}$, $\overline{x}_{2\cdot}$, $\overline{x}_{3\cdot}$ and $\overline{x}_{4\cdot}$ denote the sample means observed from the data of the four treatments. Suppose that one observes that $\overline{x}_{1\cdot} - \overline{x}_{3\cdot}$, $\overline{x}_{2\cdot} - \overline{x}_{3\cdot}$, $\overline{x}_{1\cdot} - \overline{x}_{4\cdot}$ and $\overline{x}_{2\cdot} - \overline{y}_{4\cdot}$ are the only differences that are greater than $K$. How would you order the magnitudes of the four samples, and why?

4. Do problem 1 at the end of section 1 of this chapter using both SIMANOVA.EXE and TUKEYSIM.EXE, and if there are differences, indicate them.

5. Do problem 2 at the end of section 1 of this chapter using both SIMANOVA.EXE and TUKEY.EXE, and if there are differences, indicate them.

6. Do problem 3 at the end of section 1 of this chapter using both SIMANOVA.EXE and TUKEYSIM.EXE, and if there are differences, indicate them.

7. Do problem 4 at the end of section 1 of this chapter using both SIMANOVA.EXE and TUKEYSIM.EXE, and if there are differences, indicate them.

**3. Test for Equality of $p$'s for Several Binomial Observations.** Suppose we have three treatments in which the outcome on a patient is either success $S$ or failure $F$. We try treatment 1 on $n_1$ patients, treatment 2 on $n_2$ patients and treatment 3 on $n_3$ patients. These are three different sets of patients, so they are acting independently of each other. Let $p_1$ denote the probability of a success, $S$, for a patient undergoing treatment 1, let $p_2$ denote the probability of success, $S$, for a patient undergoing treatment 2, and let $p_3$ denote the probability of success, $S$, for a patient undergoing treatment 3. The number of successes, $N_1$, among the the $n_1$ patients undergoing treatment 1 has the $Bin(n_1, p_1)$ distribution, the number of successes, $N_2$, among the the $n_2$ patients undergoing treatment 2 is $Bin(n_2, p_2)$, and the number of successes, $N_3$, among the the $n_3$ patients undergoing treatment 3 is $Bin(n_3, p_3)$. Clearly the value of $p_i$ in treatment $i$ is a measure of how good treatment $i$ is; larger values are considered better than smaller values. One might wish to compare the treatments to see if there are any differences among them or if they are all of equal effectiveness. Thus we wish to test the null hypothesis that $p_1 = p_2 = p_3$ against the alternative that the two probabilities of at least for one pair of treatments are unequal. This looks just like the simulation alternative to the one way analysis of variance test that was developed in section 1.

However, in this case our $x_{ij}$ values are zeros and ones, 1 if the treatment on a patient turns out to be a success and 0 if the treatment on a patient turns out to be a failure. In particular, if one observes the value of $N_1$ to be $k_1$, the value of $N_2$ to be $k_2$ and the value of $N_3$ to be $k_3$, then our data consist of three samples, the first being a sample consisting of $k_1$ 1's and $n_1 - k_1$ 0's, the second being a sample consisting of $k_2$ 1's and $n_2 - k_2$ 0's, and the

third being a sample consisting of $k_3$ 1's and $n_3 - k_3$ 0's. Thus, the pooled sample consists of $s = k_1 + k_2 + k_3$ numbers that are1's and $\sum_{i=1}^{3} n_i - \sum_{i=1}^{3} k_i$ numbers that are 0's. Let $n = n_1 + n_2 + n_3$ and $s = k_1 + k_2 + k_3$, and let

$$SS = \sum_{i=1}^{3} \left( \frac{k_i}{n_i} - \frac{s}{n} \right)^2$$

denote the sum of squares of differences of the sample relative frequency with the pooled sample relative frequency. If the null hypothesis that $p_1 = p_2 = p_3 =$ (some common but unknown) $p$ is true, then by Bernoulli's theorem, $SS$ should not be large. So under this hypothesis, we select $n_1$ of these $n$ numbers at random, observe $k_1'$ of these to be 1's, select $n_2$ at random from the remaining $n_2 + n_3$ numbers, and observe $k_2'$ of them to be 1's, and then observe $k_3'$ 1's among the remaining $n_3$ numbers. We then define

$$SS' = \sum_{i=1}^{3} \left( \frac{k_i'}{n_i} - \frac{s}{n} \right)^2$$

and compute

$$P([SS' \geq SS]).$$

If this probability is unbelievably small, then we reject the null hypothesis.

SIMBINAN.EXE is an appropriate program for this test.

## Exercises

1. The campaign headquarters for mayor of a big city looked over five weekly polls to see if their underdog candidate was making any progress against the entrenched incumbent. For a 10 week campaign, these were the results of small polls taken during those first five weeks:

Week 1: 50 out of a sample of size 150 were for their candidate.

Week 2: 52 out of a sample of size 150 were for their candidate.

Week 3: 53 out of a sample of size 150 were for their candidate.

Week 4: 57 out of a sample of size 150 were for their candidate.

Week 5: 60 out of a sample of size 150 were for their candidate.

there any increase in the proportion of voters for candidate A over the five week period?

**4. Test for Independence in a $2 \times t$ Contingency Table.** Let us consider the following voter preference problem in which we are interested in

whether there is a difference between the two sexes with respect to a certain issue. We are sampling voter preferences from $n$ voters, who are selected at random from the population at large, with respect to a certain issue, and the data we are collecting from each voter sampled are the following:

$(i)$ Is the individual **FOR** the issue, **OPPOSED** to this issue or **UNDECIDED**?

$(ii)$ Is the individual **MALE** or **FEMALE**?

These data can be organized in a table as follows:

| | FOR | OPPOSED | UNDECIDED |
|---|---|---|---|
| MALE | $x_{11}$ | $x_{12}$ | $x_{13}$ |
| FEMALE | $x_{21}$ | $x_{22}$ | $x_{23}$ |

.

In this table, for example, $x_{11}$ denotes the number of voters among these $n$ voters who are for the issue and are male. We denote

$$x_{1\cdot} = \sum_{j=1}^{3} x_{1j},$$

which is the number of voters in the entire sample who are males, and we denote

$$x_{\cdot 2} = \sum_{i=1}^{2} x_{i2},$$

as the number of voters in the entire sample who are opposed to this issue.

The problem here is to see if positions on this particular issue differ between the sexes. Another way of stating this is to ask whether gender is independent of the stand one takes on the issue. Thus, if being male or female does not depend on the stand an individual takes on the issue, the test is clearly that of considering the playing of three games called FOR, OPPOSED and UNDECIDED. These three games are played $x_{\cdot 1}$ times, $x_{\cdot 2}$ times and $x_{\cdot 3}$ times respectively, the outcome of each time a game is played being MALE or FEMALE. Thus the gender does not influence the stand that one takes if the probability of obtaining a male on each issue remains the same. We are therefore back to the case of testing for equality of $p$'s in three independent binomially distributed distributions. We are now in the same situation as for the test in the previous section.

The appropriate software for this is SIMBINAN.EXE.

## Exercises

1. In a very large freshman calculus class, a record was kept of those who always attended the lecture and arrived on time and those who rarely attended the lecture or who usually arrived late. Below are the data for the final grades:

| | $A$ | $B$ | $C$ | $D$ | $F$ |
|---|---|---|---|---|---|
| Always | 40 | 88 | 95 | 40 | 10 |
| Rarely | 3 | 10 | 17 | 22 | 27 |

($i$) Are grades and attendance habits independent?

($ii$) What moral do you draw from these data?

2. The following data were gathered in a study of the relation between socioeconomic status of parents and cheating in school by their children. Four hundred students were picked at random. The results are given in the following table, where the columns lower, middle and upper refer to socioeconomic status, and rows labelled yes and no refer to those who cheated and those who did not.

| | Lower | Middle | Upper |
|---|---|---|---|
| Did | 28 | 72 | 37 |
| Did not | 16 | 71 | 176 |

What conclusions can you draw?

## Chapter 8. On the Problem of Sample Size.

**1. Sample Size for the Binomial Test.** This is a tale of two experimental scientists, one wise and one foolish. Both are working independently on a new treatment for a certain ailment. There already is one treatment for this ailment, and it has been in use for a long time. According to the records, this already existing treatment is not too bad, but it could and should be improved upon. According to historical records, 60 percent of the patients who have this ailment and are given this already existing treatment exhibit a favorable response. As new patients are treated with this traditional treatment, one may consider each person so treated to be a play of a game; if the person responds favorably, we shall say that the event "Favorable Response" occurs. Thus we may say that if a person suffering from this ailment is selected at random and treated with this traditional treatment, the probability that he or she will respond favorably is 0.6. At this point one can wax even more mathematical and state that if $n$ patients suffering from this ailment are

selected at random and given this traditional treatment, then the number, $S_n$, who respond favorably has the $Bin(n, 0.6)$ distribution, i.e.,

$$P([S_n = k]) = \binom{n}{k} 0.6^k \times (1 - 0.6)^{n-k} \text{ for } 0 \leq k \leq n.$$

Each of these scientists decides to select some number $n$ of patients with this ailment, each of whom will be provided with the scientist's newly developed treatment. Each hopes see a significant increase in the cure rate.

**The foolish** experimental scientist is anxious to get started. He thinks he will try his treatment on perhaps 20 patients. He feels that if he gets considerably more favorable responses than 12, which is 60% of 20, then his treatment is superior to the traditional treatment. So he finds 20 patients with the ailment and administers his treatment. And indeed, he does have more than 12 favorable responses. It turns out that he has 15 of them, which leads him to believe that he has an improvement over the traditional treatment. However, no one will believe him. Why? Because if his treatment is no better than the traditional treatment and is only just as effective as the traditional treatment, then the probability getting at least 15 favorable responses out of a sample of size 20 is

$$P([S_n \geq 15]) = \sum_{k=15}^{20} \binom{20}{k} 0.6^k \times (1 - 0.6)^{n-k} = 0.1256.$$

An event with this probability can happen, and it is not usually considered an event of small probability. If the probability of obtaining a value as extreme as the one obtained were less than 0.05, then some credence might be given to his claim. But he was also foolish in that if the probability of a favorable response was really higher, he would have wanted to detect this with a large probability. Now consider the following case.

**The wise** experimental scientist was not overly anxious to get started. She wanted to make sure that she was doing things correctly. So she visited a friendly statistician. She told him what she expected to do, which was fine with the statistician, but then she asked this question: "What should the size of my treatment group be in order to draw a valid conclusion?" The statistician said he could not answer that question until he got from her the answers to three questions. So here are his questions and her answers.

The first question was: if the new treatment is no better than the traditional treatment, what is the largest probability with which she was willing to make a mistake, in other words, by concluding that the new treatment was better? She responded that she did not want that to be very likely to

happen, so she wanted the probability of this to be small, not larger than 0.05. So our statistician then observed: assuming the new treatment was no different from the old, then for a sample of size $n$, yet to be determined, a conclusion that her treatment was better than the traditional one could be made if the number of the cures was the smallest integer $k$ that satisfied

$$P([S_n \geq k]) = \sum_{j=k}^{n} \binom{n}{j} 0.6^j \times (1-0.6)^{n-j} \leq 0.05,$$

where $S_n$ is $Bin(n, 0.6)$. This was the first step.

The second question was: since he knew that she did not wish to make any wild claims about the new treatment if it raised the cure rate to only 61% or 62%, what is the smallest cure rate that she would really like to detect? Her answer was that she would really like to detect and announce an improvement if the cure rate was at least 75%.

The third question was: if the cure rate were at least 75%, with what probability would she wish to detect this fact experimentally? She replied that she would like a large probability of detecting this fact, at least 0.8.

So the statistician said: the problem appears to be to find a sample size $n$ and from this calculate the value of $k$ according to the answer of the first question and such that the values of $n$ and $k$ also satisfy

$$P([S_n \geq k]) = \sum_{j=k}^{n} \binom{n}{j} 0.75^j \times (1-0.75)^{n-j} \geq 0.80,$$

where now $S_n$ is $Bin(n, 0.75)$. The problem remained to find $n$ and $k$.

So here is what the statistician did next. He first assumed that maybe $n = 30$ would work. He then found that the value of $k$ should be $k = 23$ in order for $P([S_{30} \geq 23]) \leq 0.05$ when $p = 0.60$. Indeed, 23 is the smallest integer such that this inequality is satisfied; actually $P([S_{30} \geq 23]) = 0.043$ when $p = 0.60$. Then he computed $P([S_{30} \geq 23])$ when $S_{30}$ is $Bin(30, 0.75)$ and discovered that in this case $P([S_{30} \geq 23]) = 0.514$. Clearly the sample size was not large enough, since she wanted this last probability to be at least 0.80.

He tried several larger values for $n$, but always this last probability was less than 0.80. He knew he was close when he tried $n = 60$. In this case he found that if $k = 43$, then $P([S_{60} \geq 43]) \leq .05$ when $p = 0.60$. (In other words, 43 is the smallest integer such that $P([S_{60} \geq 43]) \leq .05$ when $p = 0.60$; actually, $P([S_{60} \geq 43]) = 0.041$ when $p = 0.60$.) But he found that when $p = 0.75$, then $P([S_{60} \geq 43]) = 0.775$. Finally, selecting $n = 65$ did the trick. For this value of $n$, $P([S_{65} \geq 46]) = 0.048$ when $p = 0.60$, and

$P([S_{65} \geq 46]) = 0.825$ when $p = 0.75$. His final advice to her was to make use of at least 65 patients; if the number of cured patients from among 65 them was 46 or larger, then conclude that her new treatment had a cure rate that was higher than 0.60. Of course, if she used more than 65 patients and computed $k$ according to the observation made after the answer to the first question, then the probability of rejecting the null hypothesis when $p = 0.75$ is larger than 0.80.

The moral to the above story is that an experimental scientist should consult a professional statistician before embarking on experimentation that is costly both in time and money.

Now let us consider the more general case of finding sample size in the above problem when the difference between the two rates is smaller, thus requiring possibly much larger sample sizes. Let $X$ be a random variable whose distribution is $Bin(n, p_0)$, and let $Y$ be a random variable whose distribution is $Bin(n, p_1)$, where $0 < p_0 < p_1 < 1$. Let $\alpha > 0$ be a very small probability, like 0.05, and let $\beta > \alpha$ be a large probability, like 0.80 or 0.90. The problem is to find values of positive integers $k$ and $n$ such that these two requirements are satisfied:

$$P([X \geq k]) = \alpha \text{ and } P([Y \geq k]) = \beta,$$

where, of course, the equalities are only approximate. We observe that, using the approximation provided by the Laplace-DeMoivre theorem that

$$\begin{aligned} P([X \geq k]) &= 1 - P([X \leq k-1]) \\ &= 1 - \Phi\left(\frac{k-0.5-np_0}{\sqrt{np_0(1-p_0)}}\right) \\ &= \alpha. \end{aligned}$$

Thus, $k$ and $n$ must satisfy

$$\Phi\left(\frac{k-0.5-np_0}{\sqrt{np_0(1-p_0)}}\right) = 1 - \alpha.$$

Now let $x_0$ satisfy $\Phi(x_0) = 1 - \alpha$. Then

$$\frac{k-0.5-np_0}{\sqrt{np_0(1-p_0)}} = x_0$$

or

$$k = 0.5 + np_0 + x_0\sqrt{np_0(1-p_0)}.$$

But we also want $n$ and $k$ to satisfy

$$1-\beta = P([Y \leq k-1]) = \Phi\left(\frac{k-0.5-np_1}{\sqrt{np_1(1-p_1)}}\right).$$

Let $x_1$ satisfy

$$\Phi(x_1) = 1-\beta.$$

Note that $x_1 < 0 < x_0$. Thus $k$ and $n$ must also satisfy

$$k = 0.5 + np_1 + x_1\sqrt{np_1(1-p_1)}.$$

Upon equating these two expressions for $k$, we get

$$0.5 + np_0 + x_0\sqrt{np_0(1-p_0)} = 0.5 + np_1 + x_1\sqrt{np_1(1-p_1)},$$

or

$$n(p_1-p_0) = \sqrt{n}(x_0\sqrt{p_0(1-p_0)} - x_1\sqrt{p_1(1-p_1)}).$$

Solving for $\sqrt{n}$ and squaring both sides, we obtain

$$n = \frac{1}{(p_1-p_0)^2}\left(x_0\sqrt{p_0(1-p_0)} - x_1\sqrt{p_1(1-p_1)}\right)^2.$$

Of course, this solution for $n$ is not necessarily an integer, so the required sample size should be the smallest integer that is equal to or greater than it.

If needed, appropriate software for this is BINOMDIST and NORMDIST on EXEL.

## Exercises

1. In a certain medical testing procedure, 20% of those tested had to be called back for retesting because of results that were not clear. A new procedure has been worked out which, in theory, should require a much lower percentage of recalls than the traditional procedure. Dr. X has been asked to try the new procedure and to report on whether it significantly lowered the recall rate. Dr. X is wise, so she visits a friendly statistician concerning the design of this clinical trial.

($i$) What are the three questions that the statistician should ask?

($ii$) If Dr. X does not wish to say that the new procedure lessens the recall rate, when indeed the recall rate is not significantly lower, with probability

not more than 0.05, and if Dr. X does not wish to announce an improvement unless the recall rate is at most 10%, and if Dr. X wishes to state that the recall rate has been lessened, when indeed it has lessened, with probability at least 0.80, what is the smallest number of patients that this new procedure should be tried on?

(*iii*) After obtaining the sample size in part (*ii*), how will Dr. X decide on the announcement she will make to the world.

## CHAPTER 9. LINEAR REGRESSION

**1. The Least Squares Regression Line.** Frequently in experimental work, one obtains a set of $n$ pairs of numbers,

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n),$$

and the need arises to find the equation of a line, $y = mx + b$, that "best fits the data". If all of these points do lie on the same line, then the problem is easy in that one can take two distinct points, call the $(x_i, y_i)$ and $(x_j, y_j)$; the equation of the line that is determined by these two points is

$$\frac{y - y_i}{x - x_i} = \frac{y_j - y_i}{x_j - x_i} \text{ or } \; y = y_i + \frac{y_j - y_i}{x_j - x_i}(x - x_i),$$

a line whose slope is $\frac{y_j - y_i}{x_j - x_i}$ and whose $y-$intercept is $y_i - \frac{y_j - y_i}{x_j - x_i} x_i$. But almost always, not all points lie on the same line. We might graph these $n$ points and draw a line so that some of the points lie above the line and some below it, but this is not specifying the line exactly. So we need a definition of a line that best fits the data. One definition that is used the most frequently is: a line that best fits the data is a line that minimizes the sum of squares of the vertical differences between the $n$ points and the line. We shall use the following result that was proved in section 1 of chapter 1.

**Proposition**. If $x_1, x_2, \cdots, x_n$ is a set of data, the value of $c$ that minimizes $\sum_{i=1}^{n} (x_i - c)^2$ is the sample mean,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Let us consider the equation of any nonvertical line, $y = mx + b$. The point $(x_i, y_i)$ is on the line if and only if $y_i = mx_i + b$. If the point $(x_i, y_i)$ is not on the line, the point $(x_i, mx_i + b)$ is the point on the line that is

directly above it or below it. The distance between the two points, or the amount of error, is $|y_i - mx_i - b|$, and $(y_i - mx_i - b)^2$ is the square of this vertical distance or error. The smaller we can make the sum of squares of these errors,

$$\sum_{i=1}^{n}(y_i - mx_i - b)^2$$

by adjusting the values of $m$ and $b$, the better we should feel that we are getting a line that best fits the data. Thus we would like to find values of $m$ and $b$ which minimize this sum of squares of errors. This is the same as stating that we would like to find values $\widehat{m}$ and $\widehat{b}$ of $m$ and $b$ respectively such that

$$\sum_{i=1}^{n}(y_i - \widehat{m}x_i - \widehat{b})^2 \leq \sum_{i=1}^{n}(y_i - mx_i - b)^2$$

for all pairs of numbers $m, b$. By the above recalled proposition, whatever value one may assign to $m$, the quantity $SS$, defined by

$$SS = \sum_{i=1}^{n}(y_i - mx_i - b)^2,$$

is minimized when

$$\widehat{b} = \frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i) = \overline{y} - m\overline{x}.$$

Substituting this for $b$ in the formula for $SS$, we obtain

$$\begin{aligned} SS &= \textstyle\sum_{i=1}^{n}((y_i - \overline{y}) - m(x_i - \overline{x}))^2 \\ &= m^2 \textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2 - 2m\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=1}^{n}(y_i - \overline{y})^2. \end{aligned}$$

Now let us denote $A = \sum_{i=1}^{n}(x_i - \overline{x})^2$, $B = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ and $C = \sum_{i=1}^{n}(y_i - \overline{y})^2$. Since $SS$ is a sum of squares, it is nonnegative, i.e., $SS = Am^2 - 2Bm + C \geq 0$. After some algebraic rearranging, and since $A > 0$ (because not all the $x_i$'s are equal), we obtain

$$SS = A(m - \frac{B}{A})^2 + \frac{AC - B^2}{A}.$$

Thus $SS$ is minimized when $m = \widehat{m} = \frac{B}{A}$, i.e.,

$$\widehat{m} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

The solutions for $\widehat{m}, \widehat{b}$ given above determine the equation of the line $y = \widehat{m}x + \widehat{b}$ as the line that "best fits" the data in the sense of least squares.

The value of $\widehat{m}$ is sometimes written in different form. We may rewrite the above solution for the slope of the line that best fits the data as

$$\widehat{m} = \frac{s_y}{s_x}\widehat{\rho}_{x,y},$$

where

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

and

$$\widehat{\rho}_{x,y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

The quantity $\widehat{\rho}_{x,y}$ is called the **sample correlation coefficient** of $x$ and $y$, the quantity $s_x$ is called the **sample standard deviation** of the $x$-data, and $s_y$ is called the **sample standard deviation** of the $y$-data. Recall that from chapter 1, the numbers $s_x^2$ and $s_y^2$ are called the sample variances of the corresponding sets of data. Something that should be known about $\widehat{\rho}_{x,y}$ is that the data all lie on a straight line if and only if $|\widehat{\rho}_{x,y}| = .1$, i.e., $\widehat{\rho}_{x,y} = 1$ or $\widehat{\rho}_{x,y} = -1$. But first we need an important lemma.

**Lemma: (Cauchy-Schwarz inequality).** If $u_1, \cdots, u_n, v_1, \cdots, v_n$ are real numbers, then

$$\left(\sum_{k=1}^{n} u_k v_k\right)^2 \leq \left(\sum_{r=1}^{n} u_r^2\right)\left(\sum_{s=1}^{n} v_s^2\right),$$

and equallity holds if and only if there exists a number $t$ such that $v_i = tu_i$ for $1 \leq i \leq n$.

Proof: We observe that for every real number $t$,

$$\sum_{i=1}^{n}(u_i t - v)^2 = \sum_{i=1}^{n}(u_i^2 t^2 - 2u_i v_i + v_i^2)$$

$$= \left(\sum_{i=1}^{n} u_i^2\right)t^2 - 2\left(\sum_{k=1}^{n} u_k v_k\right)t + \left(\sum_{k=1}^{n} v_k^2\right).$$

Let us denote

$$\begin{array}{rl} A & = \sum_{i=1}^{n} u_i^2 \\ B & = \sum_{k=1}^{n} u_k v_k \\ C & = \sum_{k=1}^{n} v_k^2 \end{array}.$$

Then the previous display may be written

$$\sum_{i=1}^{n}(u_i t - v)^2 = At^2 - 2Bt + C.$$

Now $A > 0$, and we may write

$$\begin{aligned}\sum_{i=1}^{n}(u_i t - v_i)^2 &= At^2 - 2Bt + C \\ &= A\left(t^2 - 2\tfrac{B}{A}t + \tfrac{C}{A}\right) \\ &= A\left(\left(t^2 - 2\tfrac{B}{A}t + \tfrac{B^2}{A^2}\right) + \left(\tfrac{C}{A} - \tfrac{B^2}{A^2}\right)\right),\end{aligned}$$

from which we obtain that for all real numbers $t$,

$$\sum_{i=1}^{n}(u_i t - v_i)^2 = A\left(t - \frac{B}{A}\right)^2 + \left(\frac{AC - B^2}{A}\right).$$

Now let $t = \frac{B}{A}$. Since $A > 0$, and since the left hand side is nonnegative, it follows that $AC - B^2 \geq 0$. From this same equation it follows that $AC - B^2 = 0$ if and only if $v_i = \frac{B}{A}u_i$ for $1 \leq i \leq n$.

We now prove a key property of the correlation coefficient.

**Proposition. If** $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ are as above, then $-1 \leq \widehat{\rho}_{x,y} \leq 1$. The correlation coefficient $\widehat{\rho}_{x,y} = 1$ if and only if $y_i - \overline{y} = c(x_i - \overline{x})$ for all $i$ and some $c > 0$, and the correlation coefficient $\widehat{\rho}_{x,y} = -1$ if and only if $y_i - \overline{y} = c(x_i - \overline{x})$ for all $i$ and some $c < 0$. In each case,

$$c = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i-1}^{n}(x_i - \overline{x})^2}.$$

Proof: This follows immediately from the Cauchy-Schwarz inequality by letting $u_i = x_i - \overline{x}$ and $v_i = y_i - \overline{y}$.

The moral to this section is that if the number pairs in a data set are positively or negatively correlated with sample correlation coefficients that are close to 1 or $-1$ respectively, then you are able to provide an approximate $y$-value if you are presented with the $x$-value. This means: if the equation of your regression line is, say, $y = 22.1 + 2.3x$, based on a data set, and if someone provides you with a value of $x$, say $x = 2$, then you are able to predict (somewhat) that the value of $y$ that should have accompanied that value of $x$ is $y = 22.1 + 2.3 \times 2 = 26.7$.

One may compute $\widehat{\rho}_{xy}$ using either EXCEL or RGRNLINE.EXE.

## Exercises

1. For a set of 16 male babies aged 48 weeks, both the height and head circumference measurements were taken with the following results:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Height | 77.3 | 73.0 | 73.9 | 71.7 | 79.6 | 75.4 | 77.6 | 72.0 |
| Circumference | 47.5 | 46.9 | 45.9 | 46.3 | 47.5 | 47.4 | 47.1 | 47.3 |
| Height | 76.4 | 75.6 | 74.9 | 70.5 | 71.6 | 73.3 | 70.9 | 75.0 |
| Circumference | 48.2 | 46.5 | 46.4 | 48.2 | 48.2 | 45.0 | 46.1 | 47.4 |

($i$) Find the equation of the line $y = mx + b$ that best fits these data in the sense of least squares, with the $x$-coordinate taken as height and the $y$-coordinate taken as circumference, and construct a nice, labelled graph of it.

($ii$) Compute the sample correlation coefficient for these data.

2. Given the following $x, y$ data, ($i$) plot these data carefully, ($ii$) find the least squares regression line for the data, and ($iii$) compute the sample correlation coefficient. The data are: $(2.4, 1.27)$, $(3.7, 1.87)$, $(4.6, 2.4)$, $(3.1, 1.6)$, $(7.6, 4.1)$.

3. Do the same as in problem 2 but with the following set of data: $(10.1, 2.4)$, $(8.2, 4.1)$, $(4.3, 8.1)$, $(6.5, 6.1)$, $(12.4, 1.1)$.

4. Do the same as in problem 2 but with the following set of data: $(9.3, 4.7)$, $(10.7, 5.2)$, $(3.9, 4.9)$, $(12.6, 5.0)$, $(7.3, 5.2)$.

5. Prove that the slope of the sample regression line is positive if and only if the sample correlation coefficient is positive, is negative if and only if the slope of the sample regression line is negative and is zero if and only if the sample correlation coefficient is 0.

**2. Test for Zero Slope for Bivariate Data.** The estimates for slope and $y$-intercept obtained in section 1 might be called computational statistics in the sense of chapter 1. But now the question arises as to whether the $x-$measurement and $y$-measurement are correlated or uncorrelated. What this means is whether one may conclude that there are increases in the value of $y$ when there are increases in the value of $x$. More formally, we wish to test the null hypothesis that $m = 0$ against the alternative that $m \neq 0$.

One must first determine a meaning to this question. One interpretation is that the $y$-values do not depend at all on the $x$-values and that for these given $x$-values arranged in order, all permutations of the order of the matching $y$-values are equally likely. Let us be concrete about this. Suppose

we have a(n unrealistic) data set consisting of four pairs of numbers, say, $(2.3, 4.5), (6.1, 3.2), (4.1, 4.0)$ and $(3.1, 4.2)$. If there were no correlation between the $x$- and $y$-values, then, keeping the first coordinates fixed in their places, all permutations of the second coordinates would be equally likely, each with probability $1/4!$, or $1/24$. Thus, all 24 sets of 4 number pairs would be equally likely when the null hypothesis is true. What we would then want to do is to compute $\widehat{m}$ for the above data set. Then we would wish to compute the value of $\widehat{m}'$ for each of the 24 data sets obtained by permuting the second coordinates. Now suppose that the value of $\widehat{m}$ is positive. Then we would determine the number of possible $\widehat{m}'$ values among the 24 that are as extreme as $\widehat{m}$, i.e., equal to or greater than $\widehat{m}$, and divide this by 24. This gives us the probability under the null hypothesis of $m = 0$ of the value of $\widehat{m}'$ being as large or as extreme as the observed value $\widehat{m}$. If this probability is unbelievably small, we would reject the null hypothesis that there is zero correlation in favor of there being a positive correlation.

For larger samples, the above procedure would be tedious or impossible. However, we can simulate the value of this probability as follows. One would take a random permutation, $j_1, j_2, \cdots, j_n$, of $1, 2, \cdots n$, and compute the value for $\widehat{m}$, call it $\widehat{m}'$, for the number pairs

$$(x_1, y_{j_1}), (x_2, y_{j_2}), \cdots, (x_n, y_{j_n}).$$

Repeat this many times, say, $10,000$ or $25,000$ times. If $\widehat{m} > 0$ in the original pairing of data, then we find the relative frequency with which we observe the event $[\widehat{m}' \geq \widehat{m}]$. In other words, we would divide the number of times that we observe the occurrence of the event $[\widehat{m}' \geq \widehat{m}]$ by the number of times we took a new random permutation of $1, \cdots, n$. On the other hand, if $\widehat{m} < 0$, we would calculate the relative frequency of the event $[\widehat{m}' \leq \widehat{m}]$. If either relative frequency is unbelievably small, then we reject the null hypothesis that $x$ and $y$ are uncorrelated in favor of positive correlation in the first case and negative correlation in the second case.

A frequently used test for correlation is the **Spearman test.** This test is based on ranks, which were discussed in the section devoted to the Wilcoxon rank-sum test.. Suppose we let $r_i$ denote the rank of $x_i$ among the first coordinates, $x_1, x_2, \cdots, x_n$, and let $s_i$ denote the rank of $y_i$ among the second coordinates $y_1, y_2, \cdots, y_n$. So apply the test just obtained above to the sample of pairs of ranks,

$$(r_1, s_1), (r_2, s_2), \cdots, (r_n, s_n)$$

by computing the value of $\widehat{m}$ for this set, treating them in exactly the same way as the raw data were treated above.

For appropriate software, use TST4TRND.EXE.

Before leaving the subject of this section, let us revisit the problem encountered in section 5 of chapter 5. There we had a sequence of $n$ games where a certain event $E$ occurs with probability $p$. The problem we were concerned with was whether the value of $P(E)$ remained constant from game to game or in general increased as one played the game. We reasoned that if $p = P(E)$ remained constant, and if $E$ ccurred $k$ times during those $n$ plays, then the game numbers at which $E$ occurred should be evenly spaced among the $n$ trials, but if the game numbers at which the $k$ times that $E$ occurs are not evenly distributed but are all close to $n$, then we would say that $p$ is increasing. We concluded that a two sample test was adequate for this, and, to make a long story short, we used the Wilcoxon rank-sum test to test the null hypothesis of constant $p$ against the alternative that $p$ was possibly increasing as one continues to play the game. We can also treat this problem by the method outlined earlier in this section, which we now show.

Suppose the data are considered as order pairs $\{(i, y_i) : 1 \leq i \leq n\}$ where each $y_i$ is 1 or 0 according as the event $E$ did or did not occur at time or trial $i$. If the value of $p$ does not increase with increasing $i$, then we would expect the slope of the regression line obtained from these data to be close to 0. But if the frequency of 1's among the $y_i$'s increases as $i$ increases, we would expect the slope of the regression line constructed for these data. to be positive. So assume that the slope of the regression line computed for the data set $\{(i, y_i) : 1 \leq i \leq n\}$ to be positive, i.e.,

$$m = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x^2} > 0.$$

But

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - n\overline{xy}.$$

So the slope of the regression line can be expressed as

$$m = \frac{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy})\right)}{s_x^2}.$$

Now observe that if we take $x_i = i$ and take a random permutation $y_1', \cdots, y_n'$

of $y_1, \cdots, y_n$ and compute the slope

$$m' = \frac{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i' - n\overline{xy}'\right)}{s_x^2},$$

we would wish to determine $P([m' \geq m])$. If $m > 0$, and if $P([m' \geq m])$ is unbelievably small, then we would reject the null hypothesis that the value of $p$ remains constant from game to game in favor of the alternative that the value of $p$ increases from game to game.

At this point we have a happy coincidence. It is that, as we shall now prove, that the test just outlined is exactly the same test as the Wilcoxon rank sum test as applied to the same problem in section 5 of Chapter 5. A proof goes as follows. First note that for any random permutation $y_1', \cdots, y_n'$ of $y_1, \cdots, y_n$, then $\overline{y}' = \overline{y}$, so that the event $[m' \geq m]$ is the same event as the event

$$\left[\sum_{i=1}^{n} x_i y_i' \geq \sum_{i=1}^{n} x_i y_i\right].$$

But if $k$ of the $y_i$'s are 1's, then it is easily observed that this is exactly the same as the Wilcoxon rank sum test.

## Exercises

1. Do a test for zero correlation with the data of problem 1 at the end of section 1 above.

2. Do a test for zero correlation with the data in problem 2 at the end of section 1 above.

3. Suppose that for $1 \leq i \leq 25$, $x_i = i$ and

$$y_i = \begin{cases} 1 \text{ if } i \in \{5, 10, 15, 20\} \\ 0 \text{ if } i \notin \{5, 10, 15, 20\}. \end{cases}$$

($i$) Does it appear that among the 25 trials that the 1's are "evenly spaced?

($ii$) Compute the slope of the line that best fits the data, $\{(x_i, y_i), 1 \leq i \leq 25\}$.

($iii$) Quote the first three sentences in a pargraph in this section that your answer in part ($ii$) illustrates.

4. Suppose that under a new program of control of defective items, among the first 100 items that come off an assembly line, the following items are defective: #7, #15, #25, #50 and #85. Is this sufficient evidence to infer

that the rate of production of defectives is decreasing? (You might try both SIMWLCXN.EXE and TST4TRND.EXE in working out your answer, but note that the two data files are different.)

5. During the first 93 days of the last 100 days of a hotly fought political campaign, Candidate A has been caught telling blatant lies on the following days: day 20, day 40, day 60, day 70, day 80, day 85, day 91. It is now day 93. Is there any evidence that Candidate A has a tendency to embroider more on the facts as election day approaches?

6. Let us look at problem 1 of section 2 of chapter 7 critically. Notice that each week the proportion of voters in a sample of size 150 increased. Now suppose that there was no upward trend. Then all permutations would be equally likely. Perhaps you can use the program TST4TRND.EXE on the ranking to see if there is hope.

**3. Test for Equality of Two Regression Lines.** On occasion, a statistician is presented with two sets of data pairs and is asked if their regression lines are the same or not. What does this mean?

The first sample of size $m$ consists of independent observations $(X_1, Y_1), \cdots, (X_m, Y_m)$ as in sections 1 and 2, for which it was assumed that these observations were taken on pairs of data where the underlying relationship is assumed to be a line with equation $y = ax + b$ plus a small possible random error. In section 1 we determined estimates $\widehat{a}$ and $\widehat{b}$ of $a$ and $b$ respectively that were best in the sense of least squares. The second sample of size $n$ consists of independent observations $(U_1, V_1), \cdots, (U_n, V_n)$ like, and idependent of, the above sample of size $m$. This second sample comes from some experiment or activity where the underlying relationshiup between the first and second entry in each pair is assumed to be that of a line of the form $v = cu + d$ plus some small possible random error, assumed to be roughly of the same order of magnitude as the error for the first sample. And again, as in section 1, we are able to determine estimates $\widehat{c}$ and $\widehat{d}$ of $c$ and $d$ respectively that are best in the sense of least squares. The problem is to determine if both lines are the same line, i.e., does $a = c$, and does $b = d$? More accurately, the problem is to determine if there is any reason to reject the null hypothesis that $a = c$ and $b = d$.

Let us denote the pairs of numbers that are observed in the first sample by $(x_1, y_1), \cdots, (x_m, y_m)$ and the numbers observed for the second sample by $(u_1, v_1), \cdots, (u_n, v_n)$. Second, let us assume that there is no difference between the two lines so that $a = c$, and $b = d$. In such a case, we may assume that all $m+n$ number pairs are in reality one sample of size $m+n$ on

some phenomenon for which there is a linear relationship between the first and second coordinates of an observation. Let us denote these $m+n$ number pairs by

$$(w_1, z_1), (w_2, z_2), \cdots, (w_{m+n}, z_{m+n}),$$

and let us denote the common linear relationship of this pooled sample by

$$z = ew + f + \text{a random error},$$

of which the random error might vary about as much as the random errors for the two samples. Thus, as in section 1, we may compute least squares estimates $\widehat{e}$ and $\widehat{f}$ for $e$ and $f$. Now, if the null hypothesis is true that the two samples come from populations or experiments in which the two lines are the same, then

$$a = c = e \text{ and } b = d = f,$$

or,

$$a - e = 0,\ c - e = 0,\ b - f = 0, \text{ and } d - f = 0.$$

Accordingly, we compute

$$F = (\widehat{a} - \widehat{e})^2 + (\widehat{c} - \widehat{e})^2 + (\widehat{b} - \widehat{f})^2 + (\widehat{d} - \widehat{f})^2.$$

If the null hypothesis is true, then this sum of squares of differences of least squares estimates should be small. Now let

$$\pi(1), \pi(2), \cdots, \pi(m+n)$$

denote a random permutation of $1, 2, \cdots, m+n$, and consider the following

$$(w_1, z_{\pi(1)}), (w_2, z_{\pi(2)}), \cdots, (w_{m+n}, z_{\pi(m+n)}).$$

Compute $\widehat{a}'$ and $\widehat{b}'$ from

$$(w_1, z_{\pi(1)}), \cdots, (w_m, z_{\pi(m)}),$$

and $\widehat{c}'$ and $\widehat{d}'$ from

$$(w_{m+1}, z_{\pi(m+1)}), \cdots, (w_{m+n}, z_{\pi(m+n)}).$$

We then consider the random variable

$$F' = (\widehat{a}' - \widehat{e})^2 + (\widehat{c}' - \widehat{e})^2 + (\widehat{b}' - \widehat{f})^2 + (\widehat{d}' - \widehat{f})^2.$$

If $P([F' \geq F])$ is unbelievably small, then we shall reject the null hypothesis that both regression lines are the same. As before, we can simulate this probability by repeatedly taking a random permutation of $1, 2, \cdots, m+n$ many times and calculating the relative frequency with which the event $[F' \geq F]$ occurs.

A BASIC copmuter program for this is TWOLINES.EXE.

## Exercises

1. Consider the following two sets of bivariate data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ : | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $y$ : | 4.8 | 7.3 | 8.5 | 12 | 12.5 | 16 | 17 |

and

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $u$ : | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $v$ : | 5 | 6.7 | 0.5 | 10 | 12 | 14 | 15 |

.

Test the null hypothesis that both have the same regression line.

## Closing Remarks

This is enough. This text is about 100 pages, which is just about right for an undergraduate one quarter general statistics course for students who are not mathematics majors. It contains enough ideas so that if one comes upon an applied problem not covered above, he or she can figure out some way to determine a simulation test for it, if such exists and is not too difficult to obtain. I feel happy about what I covered and about what I did not cover. The course has a nice cohesiveness with the coincidences that occur. There is a lot more that one can do about contingency tables and regression models. But this is just about right.