SOME THEORY AND PRACTICE OF STATISTICS
by Howard G. Tucker

CHAPTER 4. BEGINNING STATISTICAL INFERENCE

4.1 Statistical Inference Involving Historical Data. 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 a treatment for this ailment, and it has been in use for a long time. According to the records, this treatment is not too bad, but it should be improved upon. According to historical records, this traditional treatment cures 60% of the patients who have ths ailment that it is tried on. 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 recover 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 treatment, then the number who are cured, S_n, 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 scientists 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 cures 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 cures. It turns out that he has 15 cures, 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 cures is

$$P([S_{20} \geq 15]) = \sum_{k=15}^{20} \binom{20}{k} 0.6^k \times (1 - 0.6)^{n-k} = 0.1256.$$

An event with this probability can happen, and 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 percentage of cures 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 replied: assuming the new treatment was no different from the old, then for a sample of size n, a conclusion that her treatment was better than the traditional one could be made if the number of the cures was a number k that satisfied

$$P([S_n \geq k]) = \sum_{j=k}^{n} \binom{n}{j} 0.6^j \times (1-0.6)^{n-j} \leq 0.05.$$

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 questions 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 second question such that the value of n satisfies

$$P([S_n \geq k]) = \sum_{j=k}^{n} \binom{n}{j} 0.75^j \times (1-0.75)^{n-j} \geq 0.80.$$

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 number 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 $Bin(30, .75)$ and discovered that in this case $P([S_{30} \geq 23]) = 0.514$. Clearly the sample size was not large enough.

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 $P([S_{60} \geq 46]) = 0.825$. His final advice to her was to make use of 65 patients; if the number of cured patients from among them was 46 or larger, then conclude that her new treatment had a cure rate that was higher than 0.60.

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.

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 statisticain 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 Dr. X will decide on the announcement she will make to the world.

4.2 Randomization Procedures. Before we begin our study of permutation tests, we should make sure that we know how to use a random number generator in a mathematical sense. A random number generator is a program that is on many hand-held calculators and in programming software for table top computers, and it generates what are called random numbers. However, these random number generators do not generate actual random numbers but pseudo-random numbers. In practice we pretend that they are truly random. The actual random numbers are generated in the following manner. Consider a bowl with ten tags in it that are numbered $0, 1, 2, \cdots, 9$. Let X_1, X_2,$\cdots$, X_k denote a simple random sample of size k taken with replacement. This means: take a tag at random, let X_1 denote the number observed on it, replace the tag, and repeat this process $k-1$ more times. The random variable, X_i, will denote the number observed on the ith tag drawn.. The outcomes of these k trials are independent events. Clearly, each of these random variables is uniformly distributed over the numbers $\{0, 1, \cdots, 9\}$, i.e., the discrete density of each X_i is

$$f_{X_i}(x) = \begin{cases} \frac{1}{10} \text{ if } x = 0, 1, \cdots, 9 \\ 0 \text{ otherwise.} \end{cases}$$

The random number obtained is

$$\frac{X_1}{10} + \frac{X_2}{10^2} + \cdots + \frac{X_k}{10^k} \ .$$

Then the random number appears in decimal form as $0.X_1X_2 \cdots X_k$ just as the constant number $\frac{3}{10} + \frac{8}{10^2} + \frac{5}{10^3}$ appears as 0.385. If $k = 3$, then the above procedure yields a number selected at random from $\{.000, .001, \cdots, .998, .999\}$. The probability of selecting any particular number, such as .729, is .001, and the probability of selecting any number less than 0.729 is 0.729. (Do not forget .000.) In general, if one obtains a random number with k decimal places, then the probability of selecting any fixed constant x, expressed as

$$x = .x_1x_2 \cdots x_k \ ,$$

is $1/10^k$, and the probability of selecting a number less than $x \neq 0$ in the interval $[0, 1)$ is the very same number $x = .x_1x_2 \cdots x_k$. This shows how to select a number at random in the unit interval $[0, 1)$.

The problem we now address is how to select a number at random from the set of integers $\{1, 2, \cdots, N\}$. If N is a power of 10, the above procedure provides us with an answer. For example, if $N = 10^4$, and if the random number generator produced a random number $0.X_1X_2 \cdots X_8$, we might decide (ahead of time) always to select the first four digits $X_1, X_2X_3X_4$, except, when the outcome is $0,000$; we instead denote it as $10,000$. However, if the random number generator on one's hand-held calculator produces random numbers only to three decimal places, then one needs to take two random numbers $.X_1X_2X_3$ and $.Y_1Y_2Y_3$ and conclude that the number selected at random from $\{1, 2, \cdots, 10^4\}$ is $X_1, X_2X_3Y_1$ if at least one of the four digits is not zero, and is $10,000$ if all four digits are zeros.

So the problem we face is that of selecting a number at random from $\{1, 2, \cdots, N\}$, when N is not a power of ten, using a random number generator that generates a random number in $[0, 1)$ to n decimal places. We shall show, with proper reservations, that the answer is to select an n-digit random number $X^{(n)}$, multiply it by N, take the largest integer equal to or less than $X^{(n)}N$ and add 1 to it, i.e., select

$$[X^{(n)}N] + 1,$$

where $[x]$ means the largest integer equal to or less than x.

Theorem 1. *If $c = 0.c_1c_2 \cdots$ is a number (in decimal form) in $[0, 1)$, and if $X^{(n)}$ is an n-digit random number in $[0, 1)$, then $P([X^{(n)} < c]) \to c$ as $n \to \infty$.*

Proof: It is easy to verify that

$$[X^{(n)} < 0.c_1c_2 \cdots c_n] \subset [X^{(n)} < c] \subset [X^{(n)} < 0.c_1c_2 \cdots c_n + \frac{1}{10^n}].$$

Taking probabilities of the three events we have

$$0.c_1c_2 \cdots c_n \leq P([X^{(n)} < c]) \leq 0.c_1c_2 \cdots c_n + \frac{1}{10^n}.$$

Now $0 \leq c - 0.c_1c_2 \cdots c_n \leq \frac{1}{10^n}$ for every positive integer n. This implies that $0.c_1c_2 \cdots c_n \to c$ as $n \to \infty$. Taking limits in the above displayed inequality as $n \to \infty$ yields $c \leq lim_{n\to\infty}P([X^{(n)} < c]) \leq c$, which implies the conclusion.

So far we have dealt only with a finite digit random number $X^{(n)}$ which in practice is all we can deal with or observe. But theoretically we can deal with a "purely random number in $(0, 1)$" as follows. Consider a denumerable sequence of draws from the urn above. This yields an infinite sequence of random variables, $X_1, X_2, \cdots, X_n, \cdots$, of which the outcomes are independent of each other, each X_i being uniformly distributed over the positive integers from 0 through 9. Thus, an n-digit random number may still be represented by

$$X^{(n)} = \sum_{j=1}^{n} \frac{X_j}{10^j} = 0.X_1 X_2 \cdots X_n \ ,$$

where the right-most expression is the usual decimal expression of such a number. Note that Ω is the set of all sequences of integers between 0 and 9.

Definition. We define a random number X by

$$X = \sum_{j=1}^{\infty} \frac{X_j}{10^j} \ .$$

In order for this definition to have any meaning, we must show that this series exists (i.e., converges) for every possible sequence of outcomes obtained, that it is indeed a random variable and that its distribution is the uniform distribution.

Theorem 2. *The infinite series above converges at all $\omega \in \Omega$, its limit X is a random variable, and the distribution function of X is the uniform distribution, i.e.,*

$$F_X(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } 0 \le x < 1 \\ 1 \text{ if } x \ge 1 \ . \end{cases}$$

Proof: For every $\omega \in \Omega$ and for every positive integer n, $X_n(\omega)$ is a nonnegative integer in the set $\{0, 1, \cdots, 9\}$. Hence, over Ω,

$$\sum_{j=1}^{\infty} \frac{X_j}{10^j} \le \sum_{j=1}^{\infty} \frac{1}{10^{j-1}} = \frac{10}{9} < \infty \ .$$

Thus the series converges absolutely. In order to show that X is a random variable, we must prove that for every real number c, the relation

$$[X \le c] \in \mathcal{A}$$

is true. We shall make use of the fact that the X_j's are all random variables and that, as we proved earlier, finite linear combinations of random variables are random variables. We first prove a claim.

Claim 1. The following is true:

$$\bigcap_{n=1}^{\infty}\left[\sum_{j=1}^{n}\frac{X_j}{10^j}\le c\right]=[X\le c].$$

Proof of claim 1: Since $\sum_{j=1}^{n}\frac{X_j}{10^j}\le\sum_{j=1}^{\infty}\frac{X_j}{10^j}=X$, it follows that the right hand side is a subset of the left hand side. Now let ω be any elementary event in the left hand side. Then $\sum_{j=1}^{n}\frac{X_j(\omega)}{10^j}\le c$ for all n. Taking the limit of both sides as $n\to\infty$, we obtain $X(\omega)\le c$, i.e., ω belongs also to the right hand side, which proves the claim.

Since $\left[\sum_{j=1}^{n}\frac{X_j}{10^j}\le c\right]\in\mathcal{A}$ for all n, and since $\mathcal{A}$ is closed under countable intersections, it follows that $[X\le c]\in\mathcal{A}$, which proves the second conclusion of the theorem. We now wish to prove the third conclusion of the theorem that X is uniformly;distributed over the unit interval.

Claim 2. $X^{(n)}\le X\le X^{(n)}+\frac{1}{10^n}$ and $X^{(n)}\le X^{(n+1)}$ for all positive integers n.

Proof of claim 2: The first inequality follows from

$$X^{(n)}=\sum_{j=1}^{n}\frac{X_j}{10^j}\le\sum_{j=1}^{\infty}\frac{X_j}{10^j}=X\ ,$$

and the second inequality follows from the fact that

$$X-X^{(n)}=\sum_{j=n+1}^{\infty}\frac{X_j}{10^j}\le\sum_{j=n+1}^{\infty}\frac{9}{10^j}=\frac{9}{10^{n+1}}\frac{1}{1-\frac{1}{10}}=\frac{1}{10}\ ,$$

which proves claim 2. The third inequality follows because $X_{n+1}\ge 0$.

Continuing our proof of the third conclusion of the theorem, we have, by claim 2, the inequality $X\le X^{(m+n)}+\frac{1}{10^n}$ for all positive integers m,n . Hence

$$[X\le x]\subset[X^{(m+n)}\le x]\subset[X^{(m+n)}-\frac{1}{10^n}<x]\ ,$$

from which we obtain

$$P([X\le x])\le P([X^{(m+n)}-\frac{1}{10^n}<x]).$$

First take the limit of both sides as $m \to \infty$ and apply theorem 1 to obtain $P([X \leq x]) \leq x + \frac{1}{10^n}$ for all n. Then take the limit of both sides as $n \to \infty$ to obtain $P([X \leq x]) \leq x$. Now since $0 < x < 1$, consider only those values of n that satisfy $0 < \frac{1}{10^n} < x$. Applying claim 2 again, we obtain

$$[X^{(m+n)} + \frac{1}{10^n} < x] \subset [X^{(n)} + \frac{1}{10^n} < x] \subset [X \leq x] \ .$$

Again, taking probabilities and then limits as $m \to \infty$ (with n held constant) and applying theorem 1, we obtain $x - \frac{1}{10^n} \leq P([X \leq x])$ for all sufficiently large values of n. Now let $n \to \infty$ to obtain $x \leq P([X \leq x])$. This combined with the reverse inequality just proved concludes the proof of the third conclusion, and hence of the theorem.

Theorem 3. *If N is a positive integer, if $1 \leq k \leq N$, and if $Y^{(n)} = [NX^{(n)}] + 1$, then $P([Y^{(n)} = k]) \to \frac{1}{N}$ as $n \to \infty$.*

Proof: We observe that

$$\begin{aligned}[][Y^{(n)} = k] &= [[NX^{(n)}] = k - 1] \\ &= [k - 1 \leq NX^{(n)} < k] \\ &= [\tfrac{k-1}{N} \leq X^{(n)} < \tfrac{k}{N}] \ ,\end{aligned}$$

so that

$$P([Y^{(n)} = k]) = P([\frac{k-1}{N} \leq X^{(n)} < \frac{k}{N}]) \ .$$

Now one easily verifies that

$$P([X^{(n)} < \frac{k}{N}]) = P([X^{(n)} < \frac{k-1}{N}]) + P([\frac{k-1}{N} \leq X^{(n)} < \frac{k}{N}]) \ ,$$

or

$$P([\frac{k-1}{N} \leq X^{(n)} < \frac{k}{N}]) = P([X^{(n)} < \frac{k}{N}]) - P([X^{(n)} < \frac{k-1}{N}]).$$

Substituting this into the expression above for $P([Y^{(n)} = k])$ and applying Theorem 1, we obtain the conclusion.

Theorem 3 shows that if one wishes to select a number at random from $\{1, 2, \cdots, N\}$ one should select an n-digit random number $X^{(n)}$ with n as large as possible, multiply it by N, take the integer part of it and add 1. Thus, if we wished to take a simple random sample of size m *with replacement* from $\{1, 2, \cdots, N\}$, we would perform the above procedure m times to get an ordered m-tuple of integers $\{k_1, \cdots, k_m\}$. The probability of obtaining any particular m-tuple of units (and repetitions are possible) is $(1/N)^m$, because of the independence of the successive draws of the random number.

Now suppose we wish to take a sample of size m from $\{1, 2, \cdots, N\}$ *without replacement.* If these were just numbered tags in a bowl, and if we selected m of them at random without replacement, then we know that the probability of selecting the m distinct integers $k_1, \cdots, k_m$ *in this order* is $1/(N!/(N-m)!)$. The problem arises on how one could use a random number generator to select a simple random sample of size m without replacement. Let us consider this procedure. Select a number at random from $\{1, 2, \cdots, N\}$ using the random number generator, and suppose this number is k_1. Now select a second number at random by the same method. If it is different from k_1, call it k_2. If it is the same as k_1, disregard it, and select another number at random from $\{1, 2, \cdots, N\}$. If it is unequal to k_1, call it k_2. If it is equal to k_1, keep sampling until a number unequal to k_1 is obtained. One continues selecting random numbers from $\{1, 2, \cdots, N\}$ until n distinct integers $k_1, k_2, \cdots, k_n$ are obtained., $(n \leq N)$.

Theorem 4. *By sampling with the above procedure, the probability of obtaining the ordered n-tuple of distinct numbers $k_1, k_2, \cdots, k_n$ from $\{1, 2, \cdots, N\}$ in this order is*

$$\frac{1}{N!/(N-n)!} \ .$$

Proof: Let $[k_1, k_2, \cdots, k_n]$ denote the event that in continued sampling from $\{1, 2, \cdots, N\}$ by simple random sampling with replacement that the first integer selected is k_1, the second integer obtained is k_2, $\cdots$, and the nth integer selected is k_n, where as stated in the hypothesis, $k_1, k_2, \cdots, k_n$ are distinct integers selected from $\{1, 2, \cdots, N\}$. One easily sees that the event $[k_1, k_2, \cdots, k_n]$ can be represented as the following countable union of disjoint events:

$$[k_1, k_2, \cdots, k_n] = \bigcup_{m_1=0}^{\infty} \cdots \bigcup_{m_n=0}^{\infty} A(m_1, \cdots, m_{n-1}) \ ,$$

where $A(m_1, \cdots, m_{n-1})$ denotes this event: k_1 is selected in the first trial; it is followed by m_1 k_1's; these are followed by k_2, which is followed by m_2 numbers consisting of only k_1's and k_2's, followed by k_3, followed by m_3 k_1's,

k_2's and k_3's, followed by $\cdots$, followed by k_n. Because of independence,

$$\begin{aligned} P(A(m_1, \cdots, m_{n-1})) &= \tfrac{1}{N}\left(\tfrac{1}{N}\right)^{m_1} \tfrac{1}{N}\left(\tfrac{2}{N}\right)^{m_2} \cdots \left(\tfrac{n-1}{N}\right)^{m_{n-1}} \tfrac{1}{N} \\ &= \tfrac{1}{N^n} \prod_{q=1}^{n-1} \left(\tfrac{q}{N}\right)^{m_q} . \end{aligned}$$

Since the events $\{A(m_1, \cdots, m_{n-1}) : m_1 \geq 0, \cdots, m_{n-1} \geq 0\}$ are disjoint it follows that

$$\begin{aligned} P([k_1, k_2, \cdots, k_n]) &= \tfrac{1}{N^n} \sum_{m_1=0}^{\infty} \cdots \sum_{m_{n-1}=0}^{\infty} \prod_{q=1}^{n-1} \left(\tfrac{q}{N}\right)^{m_q} \\ &= \tfrac{1}{N^n} \prod_{q=1}^{n-1} \sum_{m_q=0}^{\infty} \left(\tfrac{q}{N}\right)^{m_q} \\ &= \tfrac{1}{N^n} \prod_{q=1}^{n-1} \tfrac{1}{1-q/N} \\ &= \tfrac{1}{N} \prod_{q=1}^{n-1} \tfrac{1}{N-q} = \tfrac{1}{N!/(N-n)!} . \end{aligned}$$

Theorem 4 validates the algorithm laid out before it for taking a simple random sample without replacement.

EXERCISES

1. Prove: if x is any real number, then $[x] + 1 = [x + 1]$.

2. In using a random number generator to take a simple random sample of size three without replacement from $\{1, 2, 3, 4, 5\}$, find the probability of being able to do this by selecting just three random numbers.

3. In using a random number generator to obtain a sample of size three without replacement from $\{1, 2, 3, 4, 5\}$, find the probability of selecting the last number of this sample with the fourth random number.

4. Prove: if $| p_q | < 1$ for $1 \leq q \leq n$, where n is a positive integer, then

$$\sum_{m_1=0}^{\infty} \cdots \sum_{m_n=0}^{\infty} \prod_{q=1}^{n} p_q^{m_q} = \prod_{q=1}^{n} \frac{1}{1 - p_q}.$$

4.3 Bernoulli's Theorem. Here we give some theoretical meaning to the methods of statistical inference that follow, tying together the notion of the probability of an event and the notion of relative frequency of the occurrence of that event under repeated independent trials.

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

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

Proof: Let Y be a random variable with finite second moment. We may write

$$\begin{aligned} Y^2 &= Y^2 I_{[|Y|<\epsilon]} + Y^2 I_{[|Y|\geq\epsilon]} \\ &\geq Y^2 I_{[|Y|\geq\epsilon]} \geq \epsilon^2 I_{[|Y|\geq\epsilon]} \ . \end{aligned}$$

Thus $E(Y^2) \geq \epsilon^2 P([|Y| \geq \epsilon])$. Replacing Y by $X - E(X)$ yields the conclusion.

Chebishev's inequality is not sharp; this is illustrated in problem 1 in the set of exercises at the end of this section. However, it has important uses as a lemma and for statistical interpretation. We first use it as a lemma to prove Bernoulli's theorem.

If $\{X_n\}$ is a sequence of $Bernoulli(p)$ trials, and if we denote

$$S_n = \sum_{k=1}^{n} X_k \ ,$$

then S_n is the number of successes (or 1's) in n trials. If the probability of success is p, then the distribution of S_n is $Bin(n, p)$. Thus, S_n/n is the proportion or relative frequency of the number of successes in the first n trials. In chapter 3 we obtained $E(S_n) = np$ and $Var(S_n) = np(1-p)$. Thus

$$E(S_n/n) = p \text{ and } Var(S_n/n) = p(1-p)/n \ .$$

Theorem 1. (Bernoulli's theorem) *If S_n denotes the number of successes in the first n Bernoulli trials in an infinite sequence of Bernoulli trials, then for every $\epsilon > 0$,*

$$P\left(\left[\mid \frac{S_n}{n} - p \mid \geq \epsilon\right]\right) \to 0 \text{ as } n \to \infty \ .$$

Proof: By Chebishev's inequality,

$$0 \leq P\left(\left[\mid \frac{S_n}{n} - p \mid \geq \epsilon\right]\right) \leq \frac{Var(S_n/n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \to 0 \text{ as } n \to \infty \ .$$

The moral to Bernoulli's theorem is clear. For all practical purposes, it relates relative frequency to probability, helping to provide a real-life interpretation to the applications we make of purely mathematical results. In order to illustrate this, we present in the next section an application to a statistical test of hypothesis, an application called a *permutation test.*

EXERCISES

1. Chebishev's inequality is not very sharp. To illustrate this, suppose X is a random variable with uniform distribution over the unit interval $[0, 1]$.
(i) Compute $E(X)$ and $Var(X)$.
(ii) Find the exact value of $P([|X - E(X)| \geq \frac{1}{8}])$.
(iii) Compute $Var(X)/\epsilon^2$, where $\epsilon = \frac{1}{8}$.
2. Prove: if Y is a random variable, and if $\epsilon > 0$, then

$$Y^2 \quad = Y^2 I_{[|Y|<\epsilon]} + Y^2 I_{[|Y|\geq\epsilon]}.$$

3. Prove: if Y is a random variable with finite nth moment, then, for every $\epsilon > 0$,

$$P([|X| \geq \epsilon]) \leq \frac{E(|X|^n)}{\epsilon^n}.$$

4. Prove: if X is a random variable with finite second moment, and if $\epsilon > 0$, then

$$P([|X - E(X)| \geq \frac{s.d.(X)}{\sqrt{\epsilon}}]) \leq \epsilon \ .$$

5. Prove that the largest value of $p(1 - p)$ is $\frac{1}{4}$.

4.4 A Permutation test Based on the Difference of Means. The central idea for a permutation test can be introduced through a special application that occurs frequently in biomedical research and development. Suppose there is a certain ailment or disease for which there is a standard treatment. 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 is some number. Now suppose that a new treatment has been developed, and suppose that those who developed it claim it is a better treatment. The problem that is faced is to determine in as objective a manner as possible whether it is indeed a better treatment. Let us suppose in what follows that there is some measurement x made on such a patient that indicates the severity of the disease: a large value of x indicating a severe case, and a small value of x indicating a mild case. In order to carry out a test for this new treatment, one would wish to have two groups of people suffering from the disease. 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 everyone in the *control group* ends up with the same large measurement, and if everyone in the *treatment group* ends up with the same smaller 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 m people at a certain stage of the disease are given the traditional treatment and thus become the control group, and n others are given the new treatment and thus 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 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 each 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 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 x is measured and recorded for each patient. If all the scores of those in the treatment group are well below all the scores of those in the control group, we would be tempted to conclude that the new treatment is better than the old one.

But it does not always happen like this. Here is what happens. The first thing that one notices in practice is that for each group the values of x are not all the same. There is usually considerable scatter for the x-values of each group, and there is usually some overlap, where in this case some of the x-values of the control group might be lower than some of the largest x-values of the treatment group. And so, although most values of the treatment group could be lower 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 mean $\overline{x}$ of the treatment group is substantially smaller than the arithmetic mean $\overline{y}$ of the control group. However, what do we mean by "substantially smaller"?

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 $\overline{y} - \overline{x}$ to be as large as we observe it to be? If it is not possible (under our assumption of no difference) then we would say that the new treatment is better than the standard treatment. 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 x-scores, and then observed the mean $\overline{y}$ of the x-scores among those remaining. Thus we may correctly ask, since we feel that the number $\overline{y} - \overline{x}$ is large, if we were to select m numbers at random out of these $m+n$ numbers and denote their arithmetic mean by $\overline{X}$, with $\overline{Y}$ denoting the arithmetic mean of those not selected, where both $\overline{X}$ and $\overline{Y}$ are now random variables, what is the probability that the value of $\overline{Y} - \overline{X}$ is as extreme as $\overline{y} - \overline{x}$? In other words, what is the value of $P([\,\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}])$? 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! On the other hand, if this probability were 0.18, then we would say, "Well, 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 $P([\,\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}])$. 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 these $\binom{m+n}{m}$ outcomes; for each of them (i.e., for each ω) one computes $Y(\omega)$, then $X(\omega)$, and then one determines the number of ω's for which $\overline{Y}(\omega) - \overline{X}(\omega) > \overline{y} - \overline{x}$. Call this number N. Thus

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

Now, for large values of m and n, this is terribly difficult or impossible to compute. However, a next best procedure is the following, which can be done on any desktop computer (of recent vintage). Take a simple random sample of size m from the integers from 1 to $m+n$ by the method outlined in section 4.1. Evaluate $\overline{Y} - \overline{X}$ for this outcome, record the number 1 if the inequality $\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}$ is observed, and record the number 0 if the inequality $\overline{Y} - \overline{X} < \overline{y} - \overline{x}$ is observed. Repeat this $9,999$ 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 $10,000$). According to Bernoulli's theorem, this ratio should be very close to $P([\,\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}])$.

There is no absolute rule for determining which value of $P([\,\overline{Y} - \overline{X} \geq \overline{y} - \overline{x}])$ 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 .05, then further development and testing are certainly warranted. If it is less than .001, then there is strong evidence that there is a difference. It should be noted above all that this permutation test was suitable because of the way in which the original question was phrased.

EXERCISES

1. Here is a walk-through of a permutation 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 $\overline{y} - \overline{x}$.

(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 elementary outcome ω listed in (ii), let $\overline{X}(\omega)$ denote the arithmetic mean of the three numbers selected, and let $\overline{Y}(\omega)$ denote the arithmetic mean of those remaining. Compute $\overline{X}(\omega)$, $\overline{Y}(\omega)$ and $\overline{Y}(\omega) - \overline{X}(\omega)$.

(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.

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

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

3. 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 $E(\overline{X}) = E(\overline{Y})$.

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.