SOME THEORY AND PRACTICE OF STATISTICS
by Howard G. Tucker

Chapter 6. Statistical Inference Involving Discrete Distributions

6.1 A Basic Statistical Method in Hypothesis Testing. In chapter 4 we encountered our first problems in statistical inference. One dealt with the problem of sample size. The other was a problem of a statistical test of hypothesis. It dealt with the problem of whether two samples arose from the same distribution function or whether there was a shift and essentially a difference in means between the two parent populations. What we did with that other problem was to assume that both parent distributions were the same, and, based on this assumption, we wondered whether the sample means could differ as much as they did without violating this assumption. If the separation or difference observed was so large that, under this assumption of there being no difference, the probability of the difference being at least as large as that observed was "unbelievably small," then we rejected the null hypothesis of there being no difference in favor of the alternative that there was a difference and in the direction indicated. We shall develop this notion further in this chapter but shall apply it only to discrete distributions. We shall develop this notion in two types of statistical problems: hypothesis testing and confidence intervals.

In the problem of hypothesis testing we consider a set of observable random variables $\mathbf{X} = (X_1, \cdots, X_n)$ whose joint distribution function depends upon some unknown constant (or scalar parameter) $\theta \in \Theta$. There might be a particular value of θ that we are interested in, call it θ_0. Based on our observations on the values of these random variables, our big problem will be if these data exhibit enough evidence to lead us to reject the idea that the true value of θ is θ_0. The traditional manner of stating this is to state that we are testing the null hypothesis that $\theta = \theta_0$ against the alternative that $\theta \neq \theta_0$. We shall only reject the null hypothesis if it is too unreasonable, based on our observed values of the random variables. Symbolically, statisticians write that they are testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$ or against the alternative $H_1 : \theta > \theta_0$ or against the alternative $H_1 : \theta < \theta_0$. An important thing to remember in all this is that we reject the null hypothesis in favor of an alternative only when there is overwhelming evidence to do so. A definition of "overwhelming evidence" will be introduced shortly.

The problem of confidence intervals for an unknown parameter involves no *a priori* idea of what the value θ_0 of θ is. Instead one wishes to find two functions $\mathcal{L}(\mathbf{X})$ and $\mathcal{U}(\mathbf{X})$ of the set of observable random variables such that whatever the true value of θ is, the probability $P_\theta([\mathcal{L}(\mathbf{X}) \leq \theta \leq \mathcal{U}(\mathbf{X})])$ is a large pre-set value, like .90 or .95 or .99. What is important here is that this probability must be computed under the assumption of θ being the true value of the parameter, and this is the reason we write P_θ for the probability above. We shall render the definitions of these two types of problems more precise later.

The inference with which we shall be concerned in this section is when we have a known function $M(\mathbf{X})$ of the observable random variables $\mathbf{X}$ that obeys a *monotonicity condition*, a condition that states that the following function of θ,

$$P_\theta([M(\mathbf{X}) \geq x]),$$

is either a monotone nondecreasing or q monotone nonincreasing in θ over Θ.

For example, suppose the observable random variables $X_1, \cdots, X_n$ are a simple random sample taken *with replacement* from the set of positive integers $\{1, 2, \cdots, N\}$, where the largest integer N is not known. Thus the joint density of $X_1, \cdots, X_n$ depends on the one parameter N. The observable random variables $X_1, \cdots, X_n$ are independent and identically distributed, each with common discrete density given by

$$f_{X_1}(k) = \begin{cases} \frac{1}{N} \text{ if } 1 \leq k \leq N \\ 0 \text{ otherwise.} \end{cases}$$

The function (called a statistic) $M(\mathbf{X})$ defined by $M(\mathbf{X}) = \max\{X_i : 1 \leq i \leq n\}$ can easily be shown to satisfy the monotonicity property stated above. Indeed,

$$\begin{aligned} P_N([\max\{X_i : 1 \leq i \leq n\} \geq j]) &= 1 - P_N([\max\{X_i : 1 \leq i \leq n\} \leq j-1]) \\ &= 1 - \left(\tfrac{j-1}{N}\right)^n \end{aligned}$$

which is easily seen to be an increasing function of N. We now illustrate how the test of hypothesis is executed with this example.

Suppose in the example just given we have a particular value of N, call it N_0, and we wish to test the null hypothesis that $N = N_0$. For good reason, we decide to choose as our test statistic $M(\mathbf{X}) = \max\{X_i : 1 \leq i \leq n\}$. Now we observe the values of the sample and compute their maximum value,

m. At this point we ask ourselves whether this observed value of $M(\mathbf{X})$ is believable when the value of N is this known number N_0. So we ask ourselves two questions: what is the probability that the maximum of a simple random sample taken with replacement from $\{1, 2, \cdots, N_0\}$ is equal to or less than m, and what is the probability of it being equal to or greater than m? Let the random variables in question be $U_1, \cdots, U_n$. We find the first probability to be $(m/N_0)^n$, and we find the second probability to be $1 - ((m-1)/N_0)^n$. Let us take a small probability α, for example, $\alpha = .05$ or $\alpha = .01$. If the inequality $(m/N_0)^n \leq \alpha$ is satisfied, then we might be willing to state that the value of $N = N_0$ is too large, and we reject it in favor of a value of N less than N_0. If the inequality $1 - ((m-1)/N_0)^n \leq \alpha$ is satisfied, then we must conclude that the value N_0 of N is too small, and we reject it in favor a value that is larger. If neither inequality is satisfied, then we are satisfied that there is no overwhelming evidence that disqualifies N_0 from being the true value of N.

A few remarks are in order here. In statistics, nothing is certain. A statistician is called to work on a problem in the outside or real world when the means of a certain decision are exhausted. The statistician looks at a problem and asks what is possible. An event of probability .5 is always possible. An event of probability .1 is also possible. Upon observing an event of probability .05, the statistician begins to think that the hypothesis that he or she is willing to accept might be in doubt. If the experiment is one that is preliminary to more experimentation, a probability as small as this gives one encouragement to continue with more experimentation. If the probability is down to .01, then one can be fairly sure that the assumption of the mathematical model that produced this event is not correct, and thus one would reject the null hypothesis. As you can see, the notion of a probability being "unbelievably small" is rather a subjective thing. Most social scientists use anything less than .05 as unbelievably small. This also holds in biomedical research and development, especially if the investigation is a preliminary one. If one has data on a proposed cure for the common cold, one should be more conservative.

EXERCISES

1. Suppose you are presented with an urn that contains an unknown number of tags but are told that they are numbered from 1 to some number N. You want to test the null hypothesis that $N = 1,000$. So you select a tag

at random and find that it has the number 997 on it. Why or why not would you cling to your original hypothesis?

2. Do the two probabilities in the last paragraph of this section add up to 1 and why or why not?

3. In a production line, there is a probability p of producing a defective item. You need to keep that probability(or proportion) of defective items below some number p_0. So you wait all day until you observe the first defective item; it is the rth item to be produced since you started looking. Based on this observation of a random variable with a geometric distribution, we wish to see if the process is under control, i.e., whether the true value of p is equal to or less than p_0.

(i) Suppose X has the $geom(p)$ distribution. Prove that $P_p([X \geq n])$ decreases as p increases.

(ii) Suppose α is a very small probability that you have decided upon ahead of time that if $P_p([X \leq r]) \leq \alpha$, then you would reject the null hypothesis that $p \leq p_0$. Find the smallest value of r that you could observe and still remain confident that the process was under control.

4. Suppose $X_1, \cdots, X_n$ is a simple random sample taken with replacement from $\{1, 2, \cdots, N\}$. Find the density of $M = \max\{X_i : 1 \leq i \leq n\}$.

5. In problem 4, find the density of M when the sampling is done without replacement.

6. Work out the test of hypothesis that $N = N_0$ in the example worked out in the text when the observable random variables are a simple random sample taken without replacement.

7. Prove in the example worked out in section 6.1 that $1 - \left(\frac{j-1}{N}\right)^n$ for fixed j and n is a monotone increasing function of N.

6.2 Binomial Test for Small Samples. McNemar's Test. The Sign Test. There are many occasions where one observes a value of a random variable X whose distribution is known to be $Bin(n, p)$ for some known value of n and unknown value of p. In many applications, one wishes to know whether the value of p is some known number, p_0, or whether the value of p is, say, greater than p_0 or perhaps less than p_0. We shall give in this section three examples of such tests. For a theoretical justification of some of our interpretation, we shall need the following theorem.

Theorem 1. *If U and V are random variables whose distributions are $Bin(n, p)$ and $Bin(n, p')$ respectively, if $0 < k \leq n$, and if $p < p'$, then the inequality $P([U \geq k]) \leq P([V \geq k])$ is true.*

Proof: This proof illustrates a method known as *coupling*. Let $W_1, \cdots, W_n$ denote n random numbers, i.e., n numbers selected at random in $[0, 1)$ as described in the previous chapter. Let us define random variables U and V by

$$U = \sum_{j=1}^{n} I_{[W_j \leq p]} \text{ and } V = \sum_{j=1}^{n} I_{[W_j \leq p']} \ .$$

Let us note that

$$V = \sum_{j=1}^{n} I_{[W_j \leq p]} + \sum_{j=1}^{n} I_{[p < W_j \leq p']} = U + W \ ,$$

where U is $Bin(n, p)$, V is $Bin(n, p)$ and W is $Bin(n, p' - p)$. The conclusion now follows from the easily verified fact that $[U \geq k] \subset [V \geq k]$.

Thus we see that for a known value of n, the binomial distribution satisfies the monotonicity condition discussed in the previous section. Thus we may apply it directly to hypothesis testing problems where we wish to test a null hypothesis that $p = p_0$, where p_0 is some known probability. We present now three examples of such an application: a test against historical data, McNemar's test and the sign test.

As mentioned in section 4.2, a test against historical data assumes the existence of a sizable amount of data over a considerable time along with the notion that the relative frequency of success remains constant. Here is an example of such a thing, laid out previously in section 4.2. Consider a treatment for a certain ailment which over the course of time has proved to be effective in a certain percentage of the cases, say, $100p_0\%$ of the cases, where $0 < p_0 < 1$. If we can assume this percentage of success will continue to hold, then we would conclude that using this treatment on a new subject would be a play of a game with a probability p_0 of success. Now suppose that someone has come up with a new treatment which might be better. In order to test whether it is better, it is tried on n independent subjects afflicted with this ailment. This will result in success for X of these subjects. Thus, if there is no difference between the new treatment and the old treatment, the distribution of X will be $Bin(n, p_0)$, while if the new treatment is better than the old treatment, the distribution of X will be $Bin(n, p')$ for some unknown probability p',where $p' > p_0$. In order to determine if this new treatment is better, we conduct the trial on n subjects, and let us say that we observe it to be successful on k of the subjects. Thus, k is the value of X that we

observe. Now, as was explained toward the end of the previous section, let us calculate these two probabilities:

$$P[W \leq k]) = \sum_{j=0}^{k} \binom{n}{j} p_0^j (1-p_0)^{n-j},$$

and

$$P[W \geq k]) = \sum_{j=k}^{k} \binom{n}{j} p_0^j (1-p_0)^{n-j},$$

where W is a random variable whose distribution is $Bin(n, p_0)$. If the first probability is unbelievably small, then we should wish to reject the null hypothesis that $p = p_0$ in favor of $p < p_0$; this follows from theorem 1 above. If the second probability is unbelievably small, we reject the null hypothesis that $p = p_0$ in favor of $p > p_0$; this also follows from theorem 1 above. If neither is unbelievably small, we do not reject the null hypothesis. It is impossible for both of them to be unbelievably small since their sum is *greater than* 1. This is a test against historical data.

A binomial test, called McNemar's test, also arises in testing two treatments for matched pairs. This is in connection with testing the effect of two treatments on human subjects suffering from the same ailment. Let us denote these treatments by T_1 and T_2. By comparing these two treatments we mean we wish to compare probabilities of a favorable or positive response to each of the two treatments. It might well occur that for each of the two treatments a very young person of one gender who has had this particular ailment for a very short time will have a different rate or probability of response than an older person of the opposite gender who has had this ailment for a long time. Thus there is a need for a test for *matched pairs*, where the two people of each pair are matched for age, gender, duration of the ailment, etc. In the model we shall treat here, the responses of the two individuals of each pair need not be independent. However, we are assuming that the univariate marginals are the same, i.e., for each individual in a matched pair, the probability of individual positive responses are equal. Thus our model here includes that of the *crossover trial*.

Having obtained n matched pairs, we then use a random mechanism, such as an unbiased coin, for each pair to determine which one receives treatment T_1 and which one receives T_2. Our basic assumption is that, for each matched pair, there is either no difference within each pair between the probabilities

of a favorable response to T_1 or T_2, or, if there is for one pair, then it exists for all pairs, and the direction of the inequality is the same for all pairs. It should be noted that whether the hypothesis of no difference for all pairs is true or not, these probabilities might change from pair to pair.

Let us consider the above model for n pairs. For each pair we observe the response of each member of the pair to the treatment administered. We wish to use these data to test the null hypothesis H_0 that there is no difference between treatments against the alternative H_1 that for each pair the probability of a favorable response to T_1 is greater than that for T_2. First we need a lemma.

Lemma 1. *If A and B are events, and if $P(A) = P(B)$, then*

$$P(A \cap B^c) = P(A^c \cap B).$$

Proof: Using the fact that $P(A) = P(A \cap B) + P(A \cap B^c)$ and also $P(B) = P(A \cap B) + P(A^c \cap B)$ and the hypothesis, the conclusion easily follows.

Let $[T_i \ + \text{ in } j\text{th }]$ denote the event that in the jth matched pair the response to T_i is favorable, and let $[T_i \ - \text{ in } j\text{th }]$ denote the event that in the jth matched pair the response to T_i is not favorable. Under the null hypothesis

$$P_{H_0}([T_i \ + \text{ in } j\text{th }]) = p_j$$

does not depend on i. Each of the following events

$$[T_1 \ + \text{ in } j\text{th }] \cap [T_2 \ - \text{ in } j\text{th }],$$
$$[T_1 \ - \text{ in } j\text{th }] \cap [T_2 \ + \text{ in } j\text{th }],$$

is called a *discordant outcome* of the jth trial or the jth pair. Thus, under H_0 and by lemma 1, it follows that the probabilities of the discordant outcomes of any trial are equal, i.e.,

$$P([T_1 + \text{ in } j\text{th }] \cap [T_2 \ - \text{ in } j\text{th }]) = P([T_1 \ - \text{ in } j\text{th }] \cap [T_2 \ + \text{ in } j\text{th }]).$$

We shall let q_j denote this common probability. We shall let X denote the number of pairs for which the event

$$[T_1 + \text{in } j\text{th}] \cap [T_2 - \text{in } j\text{th}]$$

occurs and shall let Y denote the number of pairs for which the outcome is

$$[T_1 - \text{in } jth] \cap [T_2 + \text{in } j\text{th}].$$

Note that if not all of the q_i's are equal, then (X, Y) does not necessarily have the multinomial distribution, but the random vector (X, Y) is the sum of n independent random vectors, (X_1, Y_1), (X_2, Y_2),$\cdots$, (X_n, Y_n), where (X_i, Y_i) is $\mathcal{MN}(1; q_i, q_i)$ when H_0 is true. If we denote $X = \sum_{j=1}^{n} X_j$ and $Y = \sum_{j=1}^{n} Y_j$, then we should wish to reject H_0 in favor of H_1 if X is too large compared to $X + Y$.

Theorem 2. *If* $0 \le k \le r \le n$, *then*

$$P_{H_0}([X = k] \mid [X + Y = r]) = \binom{r}{k} \frac{1}{2^r} \ .$$

Proof: We observe that

$$\begin{aligned} P_{H_0}([X = k] \mid [X + Y = r]) &= \frac{P_{H_0}([X=k] \cap [Y=r-k])}{P_{H_0}([X+Y=r])} \\ &= \frac{\sum\limits_{1 \le i_1 < \cdots < i_r \le n} \binom{r}{k} \left\{ \prod\limits_{j=1}^{r} q_{i_j} \right\} \prod\{(1-2q_t) : t \notin \{i_1, \cdots, i_r\}\}}{\sum\limits_{1 \le i_1 < \cdots < i_r \le n} \left\{ \prod\limits_{j=1}^{r} (2q_{i_j}) \right\} \prod\{(1-2q_s) : s \notin \{i_1, \cdots, i_r\}\}} \ . \end{aligned}$$

The reasoning behind the expression in the numerator is as follows. From every selection of r trial numbers for a discordant pair of responses to occur, there are $\binom{r}{k}$ ways of selecting the k trial numbers in which there is a favorable response to T_1 but none to T_2. In the i_jth trial (i.e., pair), no matter which one of the pair of discordant responses occurs, the probability of it occurring when H_0 is true is q_{i_j}. The probability of neither discordant pair occurring in the tth trial, where $t \notin \{i_1, \cdots, i_r\}$ is $1-2q_t$. This explains the numerator. In the denominator, for every r trial numbers $\{i_1, \cdots, i_r\}$, $1 \le i_1 < \cdots < i_r \le n$, at which a discordant pair occurs, the probability of a discordant pair occurring at the i_jth trial is $2q_{i_j}$ and, as above, the probability of neither discordant pair occurring in the tth trial, where $t \notin \{i_1, \cdots, i_r\}$ is $1 - 2q_t$. Simplification of the above yields the conclusion of the theorem.

Thus the test is as follows. Observe that value of $X + Y$; suppose it is r. Then observe the value of X; suppose it is k. Then compute$_i$

$$\sum_{j=k}^{r} \binom{r}{j} \frac{1}{2^r} \text{ and } \sum_{j=0}^{k} \binom{r}{j} \frac{1}{2^r} \ .$$

If the first inequality is "too small", we shall have observed that among the r discordant pairs, the number of times that T_1 had a favorable response

and T_2 had a negative response was too large, and thus the probability of T_1 having a favorable response is larger than the probability of T_2 having a favorable response.

There is yet another binomial test that is, in a sense, really not another test. It is called the *sign test.* In brief, it is used in the case when two different treatments are tried on each of n individuals selected at random in order to answer the question: which treatment, if any, is better? Accordingly, the mathematical model for this is a sequence of independent, identically distributed random vectors, $\{(X_i, Y_i), 1 \leq i \leq n\}$, where X_i denotes the numerical response of the first treatment on the ith individual, and Y_i denotes the magnitude of the response of the second treatment on the ith individual. Under the null hypothesis of no difference between the two treatments, one assumes that $P([X > Y]) = P([X < Y])$ and $P([X = Y]) = 0$. We see that the outcomes of these n independent trials form a sequence of Bernoulli trials with common probability $p = \frac{1}{2}$. If we denote S to be the number of individuals for which the first treatment was better than the second treatment, then under the null hypothesis of no difference, the distribution of S is $Bin(n, \frac{1}{2})$. However, by theorem 1, if $P([X > Y]) > \frac{1}{2} > P([X < Y])$, then the value of S can be larger. Thus we observe the value of S, call it s_0, and we compute $P([X \geq s_0])$ where the distribution of X is $Bin(n, \frac{1}{2})$. If this probability is unbelievably small, like .01, we reject the null hypothesis that there is no difference between the two treatments in favor of the alternative that the first treatment is better than the second treatment.

It might occur that, for some individuals, one observes the event $[X_i = Y_i]$, in other words, one has ties. It should be noted that in this case we are in the same situation in which McNemar's test applies, except that now the number of ties is equal to the total number of concordant pairs, a number that never figured into this test.

EXERCISES

1. Let $\{(X_i, Y_i, Z_i), 1 \leq i \leq n\}$ be n independent random vectors, where (X_i, Y_i, Z_i) is $\mathcal{MN}(1; p_i, p_i, p_i)$, $1 \leq i \leq n$, $0 < p_i < \frac{1}{3}$. Let $X = \sum_{i=1}^{n} X_i$, $Y = \sum_{i=1}^{n} Y_i$ and $Z = \sum_{i=1}^{n} Z_i$. Prove: if $0 < r \leq n$, then the conditional joint density of X, Y given $[X + Y + Z = r]$ is $\mathcal{MN}(r; \frac{1}{3}, \frac{1}{3})$.

2. Prove that for $0 \leq k \leq n$ and for $0 < p < 1$,

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

3. Two treatments, T_1 and T_2, are being considered as cures for the smoking habit. Each involves one week's treatment, and a cure is pronounced if the person who underwent the treatment does not return to smoking within a month's time. It is obvious that the problem of curing a smoker of his or her habit is different for a sixty year old man who has smoked three packs a day for the last 45 years and a twenty five year old woman who has smoked one pack per day for the last five years. The need for a matched pairs design is apparent. Suppose that among 50 matched pairs, 30 of these pairs were discordant, i.e., one treatment cured the individual of smoking, and the other did not. Suppose that among these 30 discordant pairs there were 18 pairs in which T_1 cured the smoker of the smoking habit according to the above definition and T_2 did not. Is there any evidence that T_1 is better than T_2?

4. In a preliminary test on a new drug that will possibly replace an old drug that over the years has had a sixty percent success rate when applied, it is found that in 30 trials, the drug was successful in 22 cases. The scientists involved congratulated themselves on achieving a higher success rate of 73.3%. Do they really have cause to congratulate themselves?

6.3 The Irwin-Fisher Test. Because of the so-called placebo effect in the one sample trial design outlined at the beginning of the previous section, doing a two sample test even in spite of the availability historical data is preferable to doing the one sample test. So the two treatments are tried on two different groups in the way outlined in chapter 5. Here the responses considered will only be "yes" or "no", "true" or "false" or 0 or 1. Thus, we shall be presented with the observations on two independent random variables with distributions $Bin(m, p')$ and $Bin(n, , p'')$, and we shall be concerned whether $p' = p''$ or $p' < p''$ or $p' > p''$. We shall discover another application of the monotonicity theorem from section 1.

Theorem 1. *If X and Y are independent random variables, if X is $Bin(m, p)$, and Y if is $Bin(n, p)$, then the conditional distribution of X, given the event $[X + Y = r]$, is the hypergeometric distribution:*

$$P([X = k]|[X + Y = r]) = \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}}$$

if $max\{0, r-n\} \leq k \leq min\{m, r\}$.

Proof: Because of independence of X and Y, we have

$$\begin{aligned} P([X=k] \mid [X+Y=r]) &= \frac{P([X=k]\cap[X+Y=r])}{P([X+Y=r])} \\ &= \frac{P([X=k]\cap[Y=r-k])}{P([X+Y=r])} \\ &= \frac{\binom{m}{k}p^k(1-p)^{m-k}\binom{n}{r-k}p^{r-k}(1-p)^{n-r+k}}{\binom{m+n}{r}p^r(1-p)^{m+n-r}} \\ &= \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}} \text{ if } max\{0, r-n\} \leq k \leq min\{m, r\} \ . \end{aligned}$$

One thing to notice about theorem 1 is that the conditional density of (X, Y) given the event $[X+Y=r]$ does not depend on p. For this reason, $X+Y$ is said to be a *sufficient statistic* for p. We shall encounter this notion again later.

Theorem 2. *If X and Y are independent random variables whose distributions are $Bin(m, p)$ and $Bin(n, p_0)$ respectively, and if $1 \leq k \leq \min\{m, r\}$, then*

$$P([X \geq k] \mid [X+Y=r]),$$

as a function of p with p_0 held constant, is a montone increasing function of p.

Proof: (Due to Clifford Qualls) Let us denote

$$b(x; n, p) = \binom{n}{x}p^x(1-p)^{n-x},\ 1 \leq x \leq n-1.$$

Then one easily verifies that, for $0 \leq x \leq n$,

$$\frac{\partial}{\partial p}b(x; n, p) = \frac{x-np}{p(1-p)}b(x; n, p).$$

Let $g(p)$ be defined by $g(p) = P([X \geq k]|[X+Y=r])$. Since by hypothesis X and Y are independent, it follows that, for any fixed k satisfying $\max\{1, r-n\} \leq k \leq \min\{m, r\}$,

$$P([X \geq k][X+Y=r]) = \sum_{x=k}^{\min\{m,r\}} P([X=x])P([Y=r-x]),$$

Letting

$$A(p) = \sum_{x=k}^{\min\{m,r\}} b(x; m, p)b(r-x; n, p_0)$$

and

$$B(p) = \sum_{x=\max\{1,r-n\}}^{\min\{m,r\}} b(x;m,p)b(r-x;n,p_0)$$

we obtain

$$g(p) = \frac{A(p)}{B(p)}.$$

Using the partial derivative computed above, we find that

$$A'(p) = \sum_{x=k}^{\min\{m,r\}} \frac{x-mp}{p(1-p)} b(x;m,p)b(r-x;n,p_0)$$

and

$$B'(p) = \sum_{x=\max\{1,r-n\}}^{\min\{m,r\}} \frac{x-mp}{p(1-p)} b(x;m,p)b(r-x;n,p_0).$$

Thus

$$g'(p) = \frac{B(p)A'(p) - A(p)B'(p)}{B(p)^2}.$$

We wish to prove $g'(p) > 0$. The numerator, $f(p)$, of $g'(p)$ is

$$f(p) = B(p)A'(p) - A(p)B'(p).$$

It is sufficient to prove $f(p) > 0$. Disregarding the term $p(1-p)$ in the denominators, since it is positive, and denoting

$$\sum_{z=a}^{b} = \sum_{z=a}^{b} b(z;m,p)b(r-z;n,p_0)$$

and

$$\sum_{z=a}^{b} z = \sum_{z=a}^{b} (z-mp)b(z;m,p)b(r-z;n,p_0),$$

we have

$$\begin{aligned} f(p) &= \left(\sum_{y=\max\{1,r-n)}^{k-1} + \sum_{y=k}^{\min\{m,r\}} \right) \sum_{x=k}^{\min\{r,m\}} x - \left(\sum_{x=k}^{\min\{r,m\}} \right) \left(\sum_{y=\max\{1,r-n)}^{k-1} y + \sum_{y=k}^{\min\{m,r\}} y \right) \\ &= \sum_{y=\max\{1,r-n)}^{k-1} \sum_{x=k}^{\min\{r,m\}} x - \sum_{y=\max\{1,r-n)}^{k-1} y \sum_{x=k}^{\min\{m,r\}} \\ &= \sum_{y=\max\{1,r-n)}^{k-1} \sum_{x=k}^{\min\{m,r\}} (x-y)b(x;m,p)b(r-x,n,p_0)b(y;m,p)b(r-y;n,p_0). \end{aligned}$$

But since $x > y$ in all summands, then $f(p) > 0$.

The above two theorems allow us to develop the Irwin-Fisher test. Let X and Y be two observable random variables, where X is $Bin(m, p')$, and Y is $Bin(n, p'')$, where m and n are known, but p' and p'' are not known. We wish to test the composite hypothesis that $p' = p''$ (i.e., we wish to test the composite hypothesis or null hypothesis $H_0 : p' = p''$) against the composite alternative that $p' \neq p''$ (i.e., against the alternative $H_1 : p' \neq p''$). Be sure to note that the null hypothesis, H_0, is composite in that it is the set of densities $\{f_{X,Y}(x, y|p), 0 < p < 1\}$, where

$$\begin{aligned} f(x, y|p) &= P([X = x] \cap [Y = y]) \text{ when } p' = p'' = p \\ &= P([X = x])P([Y = y]) \text{ (because } X \text{ and } Y \text{ are independent)} \\ &= \tbinom{m}{x} p^x (1-p)^{m-x} \tbinom{n}{y} p^y (1-p)^{n-y} \text{ for } 0 \leq x \leq m, 0 \leq y \leq n \ . \end{aligned}$$

Let $P_{H_0}(\cdot)$ denote "the probability of $\cdot$ when the null hypothesis H_0 is true". By theorem 1 above, we have: if the null hypothesis is true, then

$$P_{H_0}([X = k]|[X + Y = r]) = \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}} \text{ for } i_0 \leq k \leq i_1 \ ,$$

where $i_0 = max\{0, r-n\}$ and $i_1 = min\{r, m\}$. Under the null hypothesis (or assumption) that $p' = p''$, and having already observed the event $[X+Y = r]$, we know that the conditional density of X given this event is the same as that of the number of red balls in a simple random sample of size r selected without replacement from an urn with m red balls and n black balls. Thus, if the null hypothesis is true, we might desire to compute the probability that the value of X observed is as extreme as k. In other words, if W is the number of red balls in a simple random sample of size r taken without replacement from an urn containing m red balls and n black balls, what are the values of $P([W \leq k])$ and $P([W \geq k])$? If $P([W \leq k])$ is unbelievably small, like .05 or .01 or .001, then we might conclude that our assumption (i.e., the null hypothesis) is not true, and we might conclude that $p' < p''$. Note that this follows from theorem 2 above, which implies that if, instead of $p' = p''$, the inequality $p' < p''$ is true, then

$$P_{p'}([X \leq k]|[X + Y = r]) > P_{p''}([X \leq k]|[X + Y = r]).$$

The same thing is true if $P([W \geq k])$ is unbelievably small. In this case theorem 2 above tells us that if p'

$$P_{p'>p''}([X \geq k]|[X + Y = r]) > P_{p''}([X \geq k]|[X + Y = r]),$$

in which case we would reject the null hypothesis that $p' = p''$ in favor of $p' > p''$.

Let us therefore outline the procedure.

1. Observe the values of X and Y, call them k and $r - k$ respectively.

2. Letting W denote the number of red balls in a simple random sample without replacement of size r from an urn that contains m red balls and n black balls, compute

$$P([W \leq k]) = \sum_{j=\max\{0,r-n\}}^{k} \frac{\binom{m}{j}\binom{n}{r-j}}{\binom{m+n}{r}}$$

and

$$P([W \geq k]) = \sum_{j=k}^{\min\{r,m\}} \frac{\binom{m}{j}\binom{n}{r-j}}{\binom{m+n}{r}}.$$

3. If the first probability is unbelievably small, reject the null hypothesis in favor of the alternative $p' < p$". If the second probability is unbelievably small, reject the null hypothesis in favor of the alternative $p' > p$". If neither probability is unbelievably small, then one fails to reject the null hypothesis. Since the sum of these two probabilities is greater than 1, it is never possible for both of them to be unbelievably small. The smaller of the two probabilities is called the P-value of the test.

4. Note that with all the factorials involved in computing the P-value, it is apparent that a program written for this will give you problems for large values of m and n. Treatment of this test for larger values will come later.

Example: A survey is taken of two communities to determine if they both have the same rates of unemployment. A random sample of size thirty is taken of each community, and it may be assumed that this sampling is done with replacement. In community A, 8 out of the 30 were unemployed, while in community B, 4 out of 30 were unemployed. So we have taken observations on two independent binomially distributed random variables. We shall refer to the 30 in community B as the red balls and to the 30 in community A as the black balls. Performing the Irwin-Fisher test outlined above, we compute

$$\sum_{j=0}^{4} \frac{\binom{30}{j}\binom{30}{12-j}}{\binom{60}{12}} = 0.1667\ .$$

An event of probability 0.1667 is not a rare or unexpected event; it is comparable to a fair die coming up a 6.

EXERCISES

1. An experiment in educational research involves two teaching methods, call them C and D. In a class of 20 students, 14 students passed a standard examination, while in a class of 30 students taught by method D, 25 students passed the same examination. Is there statistical evidence that method D is superior to method C?

2. Sixty patients suffering from the same ailment are divided at random into two groups of patients of 30 patients in each group. The patients of group I are given the traditional treatment, and 15 of them recover. The patients in group II are given a newly developed treatment, and 20 of them recover. Is there enough evidence here to conclude that the new treatment is better than the old treatment?

3. Prove: If X and Y are integer-valued random variables, then

$$[X = j] \cap [X + Y = n] = [[X = j] \cap [Y = n - j] \ .$$

6.4 The Irwin-Fisher Test Adapted to Small Probabilities. In the example given in the previous section, the probabilities of being unemployed in each of the communities were of considerable size, and the sample sizes were relatively small, so that the binomial distribution easily applied. However there are cases where the sample sizes are exceedingly large and the probabilities of occurrence are exceedingly small, and it might be difficult to compute that hypergeometric distribution. For example, suppose one is interested in comparing death rates between two communities. Suppose community A has $50,000$ souls, and the number of deaths during a fixed period of time was 12, while community B has $70,000$ souls with 11 deaths during that same period of time. In each community each individual constitutes a Bernoulli trial with a very small but constant probability of dying during a particular year. So the number 11 is the observed value of a random variable X whose distribution is $Bin(70,000, p')$, and 12 is the observed value of a random variable Y whose distribution is $Bin(50,000, p'')$. If the problem is to determine whether the death rates p' and p'' are equal or unequal, we have the same problem as in the previous section and wish to apply the Irwin-Fisher test. It should be intuitively clear that p' and p'' are very small. By theorem 5 in section 2.2 we are able to approximate the distributions of X and Y by Poisson distributions. In particular, we may assume that X is

$\mathcal{P}(70,000p')$ and that Y is $\mathcal{P}(50,000p'')$, and our problem is to test the null hypothesis that $p' = p''$ against the alternative that $p' \neq p''$. In order to do this we need the following theorem.

Theorem 1. *If X and Y are independent random variables, if X is $\mathcal{P}(\lambda)$, and Y if is $\mathcal{P}(\mu)$, then*

$$P([X = k] \mid [X + Y = r]) = \binom{r}{k} \left(\frac{\lambda}{\lambda + \mu}\right)^k \left(1 - \frac{\lambda}{\lambda + \mu}\right)^{r-k}$$

for $0 \leq k \leq r$.

Proof: By the definition of conditional probability we have

$$P([X = k] \mid [X + Y = r]) = \frac{P([X = k] \cap [X + Y = r])}{P([X + Y = r])}.$$

Now by theorem 6 in section 2.2 we have that $X + Y$ is $\mathcal{P}(\lambda + \mu)$ Then by the easily proved identity

$$[X = k] \cap [X + Y = r] = [X = k] \cap [Y = r - k] \ ,$$

and continuing to remember that X and Y are independent, we have, for $0 \leq k \leq r$,

$$\begin{aligned} P([X = k] \mid [X + Y = r]) &= \tfrac{P([X=k] \cap [Y=r-k])}{P([X+Y=r])} \\ &= \tfrac{e^{-\lambda} \frac{\lambda^k}{k!} e^{-\mu} \frac{\mu^{r-k}}{(r-k)!}}{e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^r}{r!}} \\ &= \tbinom{r}{k} \left(\tfrac{\lambda}{\lambda+\mu}\right)^k \left(1 - \tfrac{\lambda}{\lambda+\mu}\right)^{r-k} \ . \end{aligned}$$

Theorem 2. *If X and Y are independent random variables, if X is $\mathcal{P}(\lambda)$, if Y is $\mathcal{P}$ (λ_0), and if $1 \leq k \leq r$, then $P([X \geq k] | [X + Y = r])$ is an increasing function of λ.*

Proof: We first observe that $\lambda/(\lambda + \lambda_0)$ is an increasing function of λ (by noticing that the derivative with respect to λ is positive). It is sufficient to prove that $f(\lambda)$, defined by

$$f(\lambda) = \frac{1}{P([X \geq k] | [X + Y = r])}$$

is decreasing in λ. Note that, because of independence,

$$\begin{aligned} f(\lambda) &= \frac{P([X+Y=r])}{P([X\geq k]\cap[X+Y=r])} \\ &= \frac{\sum_{j=0}^{r} P([X=j])P([Y=r-j])}{\sum_{j=k}^{r} P([X=j])P([Y=r-j])} \\ &= \frac{\sum_{j=0}^{k-1} P([X=j])P([Y=r-j])}{\sum_{j=k}^{r} P([X=j])P([Y=r-j])} + 1 \\ &= \frac{\sum_{j=0}^{k-1} e^{-\lambda}\frac{\lambda^j}{j!}e^{-\lambda_0}\frac{\lambda_0^{r-j}}{(r-j)!}}{\sum_{j=k}^{r} e^{-\lambda}\frac{\lambda^j}{j!}e^{-\lambda_0}\frac{\lambda_0^{r-j}}{(r-j)!}} + 1 \\ &= \frac{\sum_{j=0}^{k-1}\binom{r}{j}\left(\frac{\lambda}{\lambda+\lambda_0}\right)^j\left(\frac{\lambda_0}{\lambda+\lambda_0}\right)^{r-j}}{\sum_{j=k}^{r}\binom{r}{j}\left(\frac{\lambda}{\lambda+\lambda_0}\right)^j\left(\frac{\lambda_0}{\lambda+\lambda_0}\right)^{r-j}} + 1. \end{aligned}$$

If U is a random variable whoxe distribution is $Bin(r, \lambda/(\lambda+\lambda_0))$, then by the observation above and by theorem 1 in section 6.2, $f(\lambda)$ is decreasing as λ increases.

Now let us consider the general statistical problem where we can observe independent random variables, X, whose distribution is $Bin(m, p')$, and Y, whose distribution is $Bin(n, p'')$, where m and n are very large, where p' and p'' are very small, and where the products mp' and np'' are "just right". We wish to test the null hypothesis, $p' = p''$, against the alternative $p' \neq p''$. Let us assume now that the null hypothesis, $p' = p''$, is true and try to see if the outcome that we observe is possible with this assumption. Suppose we observe the value of X to be the nonnegative integer k and the value of Y to be the positive integer $r-k$, and thus the observed value of $X+Y$ is r. Letting p be the common value of p' and p'' under the above assumption, we may apply theorem 5 in section 2.2 to approximate the distribution function of X as $\mathcal{P}(mp)$ and of Y as $\mathcal{P}(np)$. We may now apply theorem 1 above to obtain

$$P([X=k] \mid [X+Y=r]) = \binom{r}{k}\left(\frac{m}{m+n}\right)^k\left(1-\frac{m}{m+n}\right)^{r-k}.$$

Since this is the (conditional) binomial distribution, $Bin(r, \frac{m}{m+n})$ whose expectation is $r\frac{m}{m+n}$, then if our observed value of X, namely k, is much smaller that this expectation, we would want to find the probability of the value of X being as extreme as that, i.e., being as small or smaller than k. In order to find this probability, we would consider a random variable W whose distribution is $Bin(r, \frac{m}{m+n})$ and compute $P([W \leq k])$. In other words, we would

compute

$$P([W \leq k]) = \sum_{j=0}^{k} \binom{r}{j} \left(\frac{m}{m+n}\right)^j \left(1 - \frac{m}{m+n}\right)^{r-j} .$$

If this P-value is too small, we would reject the null hypothesis that $p' = p''$. But then what would we decide upon rejection? We would decide that $p' < p''$ since clearly the observed value of X is too small relative to the sum of the two values $X + Y$.

So now go back to the example in the first paragraph of this section. In this case

$$P([W \leq 11]) = \sum_{j=0}^{11} \binom{23}{j} \left(\frac{7}{12}\right)^j \left(1 - \frac{7}{12}\right)^{23-j} = 0.208.$$

This probability is not small enough to reject the null hypothesis, and thus for this example we would say that unemployment rates are equal.

EXERCISES

1. In a mass production process, 11 items were found to be defective out of 1,231 that were produced that day. That night the production machinery was readjusted, and on the following day 8 items were found to be defective out of the 1,403 that were produced. Has the defective rate been reduced by the readjustment?

6.5 The Irwin-Fisher Test Adapted to 2×2 Contingency Tables. We consider now a trial or an experiment and two events A and B. These events are not disjoint, and it is assumed that one is able to perform the experiment or trial n times with the outcome of each trial independent of the outcomes of all the other trials. The problem is to determine whether these two events that are outcomes of a trial are independent, based on the outcomes of these n trials. Accordingly, let X_1 denote the number of trials from among the n trials in which both events A and B occur, let X_2 denote the number of these trials in which A does not occur but B does occur, let Y_1 denote the number of trials in which A occurs but B does not occur, and let Y_2 denote the number of trials in which neither A nor B occurs. We see

immediately that $X_1 + X_2 + Y_1 + Y_2 = n$. The data are usually arranged in a table as follows:

$$\begin{array}{ccc} & A & A^c \\ B & X_1 & X_2 \\ B^c & Y_1 & Y_2 \end{array}$$

and is usually referred to as a 2×2 contingency table. Before we derive a test for independence, we develop the notion of conditioning with a conditional probability.

Definition. If E is an event with positive probability, we define P_E over $\mathcal{A}$ by $P_E(A) = P(A|E)$ for all $A \in \mathcal{A}$.

We recall from Theorem 1 in section 1.5 that P_E is a probability. Hence if $D \in \mathcal{A}$ and $P_E(D) > 0$, we are able to define the conditional probability of an event C given D relative to the probability P_E by

$$P_E(C|D) = \frac{P_E(C \cap D)}{P_E(D)} \ .$$

Theorem 1. *If C, D and E are events such that $P(C \cap E) > 0$, then $P_C(D|E) = P(D|E \cap C)$.*

Proof: Since $C \cap E \subset C$ and $P(C \cap E) > 0$, it follows that $P(C) > 0$. Thus we may consider the conditional probability P_C. We now observe that

$$P_C(D|E) = \frac{P_C(D \cap E)}{P_C(E)} = \frac{P(D \cap E \cap C)}{P(E \cap C)} = P(D|E \cap C) \ ,$$

which concludes the proof.

The test for independence is done here within the framework of the Irwin-Fisher test, based on the sufficient statistic $(X_1 + Y_1, X_1 + X_2)$. Let $P(A) = p_1, P(A^c) = p_2 = 1 - p_1$, $P(B) = q_1$ and $P(B^c) = q_2 = 1 - p_1$.

Theorem 2. *Under the null hypothesis that events A and B are independent, we have*

$$P([X_1 = k]|[X_1 + Y_1 = r] \cap [X_1 + X_2 = a]) = \frac{\binom{a}{k}\binom{n-a}{r-k}}{\binom{n}{r}} \ ,$$

for all values k of that satisfy $max\{0, r - n + a\} \leq k \leq min\{a, r\}$.

Proof: When the null hypothesis of independence is true, then $P(A \cap B) = p_1 q_1$, $P(A^c \cap B) = p_2 q_1$ $P(A \cap B^c) = p_1 q_2$ and $P(A^c \cap B^c) = p_2 q_2$. Thus

$$P([X_1 = x_1] \cap [Y_1 = y_1]|[X_1 + X_2 = a])$$

can be expressed as

$$= \frac{P([X_1=x_1]\cap[Y_1=y_1][X_2=a-x_1])}{P([X_1+X_2=a])}$$
$$= \frac{\frac{n!}{x_1!(a-x_1)!y_1!(n-a-y_1)!}(p_1q_1)^{x_1}(p_2q_1)^{a-x_1}(p_1q_2)^{y_1}(p_2q_2)^{n-a-y_1}}{\frac{n!}{a!(n-a)!}q_1^a q_2^{n-a}}$$
$$= \frac{a!}{x_1!(a-x_1)!}p_1^{x_1}p_2^{a-x_1}\frac{(n-a)!}{y_1!(n-a-y_1)!}p_1^{y_1}p_2^{n-a-y_1} \quad ,$$

i.e., the joint distribution of X_1 and Y_2 conditioned on the event $[X_1 + X_2 = a]$ is that of two independent random variables whose distributions are $Bin(a, p_1)$ and $Bin(n-a, p_1)$ respectively. Thus by theorem 1 in section 6.3 and by theorem 1 above, we have the conclusion of the theorem.

The test now follows along the lines of the Irwin-Fisher test.

EXERCISES

1. Among 119 students examined at a university health center, 20 of them had infectious mononucleosis. Among these 20, there were 5 students who had a tonsillectomy, while among the 99 students who did not have infectious mononucleosis, 41 of them had a tonsillectomy. Is there any stochastic dependence between infectious mononucleosis and a previous tonsillectomy?

2. Prove: if C and D are events, and if $P(C \cap D) > 0$, then $P(C) > 0$ and $P(D) > 0$.

6.6 Conservative Confidence Intervals. In Section 6.1, we explained the general notion of a confidence interval. In section 5.3, we actually constructed a confidence interval for the unknown probability p upon observing the value of a random variable S_n whose distribution is $Bin(n, p)$ and where n is known. However, such a confidence interval is not completely satisfactory. All we knew about it was that the probability of it containing the unknown parameter p was approximately .95, with its accuracy improving as n gets very large. However, in that problem we had no idea about what values of n are large enough to trust such an estimate. In this section we shall develop a method of obtaining what are called "conservative" confidence intervals for an unknown parameter. By this we shall mean the following. Suppose one has some observable random variables $\mathbf{X} = (X_1, \cdots, X_n)$, whose joint distribution depends on one unknown parameter θ. The general problem is: are we able to find functions $L(\mathbf{X})$ and $U(\mathbf{X})$ such that for any value of θ, $P_\theta([L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})]) \geq 1 - \alpha$, where α is some small number, for example,

like .1 or .05 or .01. In this section we shall outline a general condition under which this can be done and shall apply it to the binomial and Poisson distributions.

We shall assume here that we have some observable random variables $\mathbf{X} = (X_1, \cdots, X_n)$ whose distribution depends on the single parameter θ. We shall assume that there exists a function $M = M(\mathbf{X})$ of the random variables that has the following monotonicity property: for every $x \in \mathbf{R}^1$, the function $P_\theta([M \geq x])$ is nondecreasing in θ.

Theorem 1. *Under the monotonicity hypothesis that $P_\theta([M \geq x])$ is nondecreasing in θ and if, for $\alpha \in (0,1)$ and $r \in \mathbf{R}^1$, $U(r)$ is defined by*

$$U(r) = sup\{\theta : P_\theta([M \leq r]) > \alpha\},$$

then $P_\theta([U(M) \geq \theta]) \geq 1 - \alpha$.

Proof: We shall prove this by a sequence of claims.

Claim 1: If $r' < r''$, then $U(r') \leq U(r'')$. To prove this we make use of the monotonicity hypothesis to observe that

$$\{\theta : P_\theta([M \leq r']) > \alpha\} \subset \{\theta : P_\theta([M \leq r'']) > \alpha\}.$$

This set inclusion implies

$$sup\{\theta : P_\theta([M \leq r']) > \alpha\} \leq sup\{\theta : P_\theta([M \leq r'']) > \alpha\},$$

i.e., $U(r') \leq U(r'')$.

Claim 2: For every value of θ, the set $\{r : U(r) < \theta\}$ is an interval of the form $(-\infty, r_0]$ or $.(-\infty, r_0)$. The proof of this follows immediately from the claim 1 by noticing that for every r in $\{r : U(r) < \theta\}$, all real numbers less than it is in the set.

Claim 3.If $\{r : U(r) < \theta\} = (-\infty, r_0]$, then $P_\theta([U(M) \geq \theta]) \geq 1 - \alpha$. Under the hypothesis of this claim,

$$P_\theta([M \in \{r : U(r) < \theta\}]) = P_\theta([M \in (-\infty, r_0]]),$$

in other words,

$$P_\theta([U(M) < \theta\}]) = P_\theta([M \leq r_0]).$$

By hypothesis, $r_0 \in \{r : U(r) < \theta\}$. Therefore, $U(r_0) < \theta$, so by the definition of $U(r)$,

$$\theta \notin \{\theta' : P_{\theta'}([M \leq r_0]) > \alpha\}.$$

From this we obtain $P_\theta([M \leq r_0]) \leq \alpha$. Thus, by the second display equation in the proof of this claim,

$$P_\theta([U(M) \geq \theta\}]) = P_\theta([M > r_0]) \geq 1 - \alpha,$$

which proves the claim.

Claim 4: *If* $\{r : U(r) < \theta\} = (-\infty, r_0)$, *then* $P([U(M) \geq \theta]) \geq 1 - \alpha$. For every positive integer n, we first observe that $r_0 - \frac{1}{n} \in (-\infty, r_0)$. Thus, by the hypothesis of this claim, $r_0 - \frac{1}{n} \in \{r : U(r) < \theta\}$ for all positive integers n, i.e.,

$$U(r_0 - \frac{1}{n}) < \theta \text{ for all } n.$$

By this inequality and the definition of $U(r)$, $P_\theta([M \leq r_0 - \frac{1}{n}]) \leq \alpha$ for all n. (Why? By the definition of $U(r)$,

$$U(r_0 - \frac{1}{n}) = sup\{\theta' : P_{\theta'}([M \leq r_0 - \frac{1}{n}]) > \alpha\},$$

from which it follows that

$$\theta \notin \{\theta' : P_{\theta'}([M \leq r_0 - \frac{1}{n}]) > \alpha\},$$

i.e., $P_\theta([M \leq r_0 - \frac{1}{n}]) \leq \alpha$.) Letting $n \to \infty$, we obtain $P_\theta([M < r_0]) \leq \alpha$. But recall that in this case, $\{r : U(r) < \theta\} = (-\infty, r_0)$. This and our last conclusion imply that

$$P_\theta([U(M) < \theta]) = P_\theta([M < r_0]) \leq \alpha,$$

from which we obtain the conclusion of the claim and the conclusion of the proof of the theorem.

Theorem 2. *Under the monotonicity hypothesis that* $P_\theta([M \geq x])$ *is nondecreasing in* θ, *and if, for any* $\alpha \in (0, 1)$ *and* $r \in \mathbf{R}^1$, $L(r)$ is *defined by*

$$L(r) = inf\{\theta : P_\theta([M \geq r]) > \alpha\},$$

then $P_\theta([L(M) \leq \theta]) \geq 1 - \alpha$.

Proof: The proof of this theorem is similar to the proof of theorem 1. We shall accomplish this proof with a sequence of four claims.

Claim 1: If $r' < r''$, *then* $L(r') \leq L(r'')$. For each fixed θ, $P_\theta([M \geq r']) \geq P_\theta([M \geq r''])$. This implies that

$$\{\theta : P_\theta([M \geq r'']) > \alpha\} \subset \{\theta : P_\theta([M \geq r']) > \alpha\},$$

from which it follows that

$$inf\{\theta : P_\theta([M \geq r'']) > \alpha\} \geq inf\{\theta : P_\theta([M \geq r']) > \alpha\},$$

which proves the claim.

Claim 2: $\{r : L(r) > \theta\}$ *is an interval of the form* $[r_0, \infty)$ *or* (r_0, ∞). This is an immediate consequence of claim 1.

Claim 3: The theorem is true when $\{r : L(r) > \theta\} = [r_0, \infty)$. In this case, $r_0 \in \{r : L(r) > \theta\}$, so $L(r_0) > \theta$. Hence, by the definition of $L(r)$, $\theta < inf\{\theta' : P_{\theta'}([M \geq r_0]) > \alpha\}$. Hence $P_\theta([M \geq r_0]) \leq \alpha$. Since in this case, $[L(M) > \theta] = [M \geq r_0]$, then

$$P_\theta([L(M) > \theta]) = P_\theta([M \geq r_0]) \leq \alpha,$$

from which we obtain $P_\theta([L(M) \leq \theta]) \geq 1 - \alpha$, the conclusion of the claim.

Claim 4: The theorem is true when $\{r : L(r) > \theta\} = (r_0, \infty)$. In this case, $r_0 \notin \{r : L(r) > \theta\}$. Hence $L(r_0) \leq \theta$, but for every positive integer n, $L(r_0 + \frac{1}{n}) > \theta$. Recall that $L(r) = inf\{\theta' : P_{\theta'}([M \geq r]) > \alpha\}$, so if $\theta \notin \{\theta' : P_{\theta'}([M \geq r_0 + \frac{1}{n}]) > \alpha\}$, then $P_\theta([M \geq r_0 + \frac{1}{n}]) \leq \alpha$. As in the proof of theorem 1, let $n \to \infty$, and we get $P_\theta([M > r_0]) \leq \alpha$. Thus, by the hypothesis in this claim,

$$\alpha \geq P_\theta([M > r_0]) = P_\theta([L(M) > \theta]),$$

from which we obtain

$$1 - \alpha \leq P_\theta([M \leq r_0]) = P_\theta([L(M) \leq \theta]),$$

which concludes the proof of the claim and of the theorem.

The above two theorems give one-sided confidence intervals. We may now obtain a solution to the problem originally posed.

Theorem 3. *Under the monotonicity hypothesis that* $P_\theta([M \geq x])$ *is nondecreasing in* θ *and if, for any* $\alpha \in (0, 1)$ *and* $r \in \mathbf{R}^1$, $U(M)$ *and* $L(M)$ *are as defined in theorems 1 and 2, then, for every* θ,

$$P_\theta([L(M) \leq \theta \leq U(M)]) \geq 1 - 2\alpha.$$

Proof: Indeed, by theorems 1 and 2,

$$\begin{aligned}
&1 - P_\theta([L(M) \leq \theta \leq U(M)]) \\
&= 1 - P_\theta([L(M) \leq \theta] \cap [\theta \leq U(M)]) \\
&= P_\theta([L(M) > \theta] \cup [\theta < U(M)]) \\
&\leq P_\theta([L(M) > \theta]) + P([\theta < U(M)]) \leq 2\alpha,
\end{aligned}$$

which proves the theorem.

We present two immediate applications of this method of obtaining conservative confidence intervals. As a first application consider the case where one observes a value of a random variable S_n, whose distribution is $Bin(n, p)$, where n is known but p is unknown. Suppose one wishes to obtain a confidence interval $L(S_n), U(S_n)$ such that whatever the value of p is, then $P_p([L(S_n) \leq p \leq U(S_n)]) \geq 1 - 2\alpha$. Suppose one observes that the value of S_n is r, i.e., one observes the event $[S_n = r]$. One then selects the value of α that one wishes to use; usually it is .025, but sometimes it is as large as .05 or even .10. Then one need only compute

$$U(r) = max\{p : \sum_{j=0}^{r} \binom{n}{j} p^j (1-p)^{n-j} > \alpha\}$$

and

$$L(r) = min\{p : \sum_{j=r}^{n} \binom{n}{j} p^j (1-p)^{n-j} > \alpha.$$

These two numbers supply a $100(1-2\alpha)\%$ confidence interval. An interpretation of this is important. If p_0 is the correct value of p, it is *absolutely not correct* to state that the probability that $L(r) \leq p_0 \leq U(r)$ is equal to or greater than $1 - 2\alpha$; the probability of this inequality is either zero or one. What is a correct interpretation is a statement that among the many independent occasions in which this interval is computed, the inequality is correct in about $100(1-2\alpha)\%$ of the cases. The evaluation of $U(r)$ and $L(r)$ can be obtained upon suitable programming by a desktop computer. It should be noted that the binomial coefficients involved here are unwieldy for values of n that are too large. When n is too large, one has to be satisfied with the approximation developed earlier using the Laplace-DeMoivre theorem.

Another immediate application is in connection with the hypergeometric distribution. Suppose we consider the case where we have an urn with a known number N of balls in it, R of them being red. Suppose you know the

number N but do not know the number $\dot{R}$. One wishes to find a $100(1-2\alpha)\%$ confidence interval for R by taking a sample of size n without replacement. The above procedure applies here, but first we must prove a theorem.

Theorem 4. *If X is the number of red balls in a sample of size n selected at random without replacement from an urn containing R red balls and B black balls then $P_R([X \geq k])$ is a nondecreasing function of R.*

Proof: This is another proof by the method of coupling. Let X denote the number of numbers taken from $\{1, 2, \cdots, N\}$ by a simple random sample taken without replacement of size n that are equal to or less than R, and let Y denote the number of numbers taken from $\{1, 2, \cdots, N\}$ by the same simple random sample taken without replacement of size n that are equal to or less than $R + 1$. Then it is clear that $X \leq Y$, which implies that $P([X \geq k]) < P([Y \geq k])$.

Theorem 4 verifies that the distribution of the number of red balls in the sample satisfies the same monotonicity condition we have been dealing with. We may now obtain a $100(1 - 2\alpha)\%$ confidence interval for R. As before, select the α that you want. You know the value of N, and you observe the value X of red balls in the sample of size n taken without replacement. Suppose you observe the value of X to be r, i.e., you observe the event $[X = r]$. Thus you compute

$$U(r) = \max\{R : \quad \sum_{j=0}^{r} \frac{\binom{R}{j}\binom{N-R}{n-j}}{\binom{N}{n}} > \alpha\}$$

and

$$L(r) = \min\{R : \quad \sum_{j=r}^{n} \frac{\binom{R}{j}\binom{N-R}{n-j}}{\binom{N}{n}} > \alpha\}.$$

The same remarks hold as with the binomial treated above.

EXERCISES

1. In the proof of theorem 4, one knows that $P([X \geq k]) \leq P([Y \geq k])$ because $X \leq Y$. However, the strict inequality is also true. So prove it.

2. Let $X_1, \cdots, X_n$ be a sample of size n taken from $\{1, 2, \cdots, N\}$ without replacement. The problem here is to find a $100(1 - 2\alpha)\%$ conservative confidence interval of N which is unknown. Let $M(\ X_1, \cdots, X_n) = max\{X_1, \cdots, X_n\}$.

(i) Prove that the distribution of $M(\ X_1, \cdots, X_n)$ satisfies the monotonicity condition.

(ii) What are the formulas for the two ends of the $100(1-2\alpha)\%$ confidence interval for N.

3. Suppose that in the setup of the conservative confidence interval problem at the beginning of this section the monotonicity condition is reversed, i.e., suppose that $P_\theta([M \geq k])$ is nonincreasing in θ. Find the formulas for $U(M)$ and $L(M)$ for a $100(1-2\alpha)\%$ confidence interval for θ.

4. Sometimes in practice one observes a random variable X that is $Bin(n,p)$ in which the investigator knows the value of p but does not know the value of n. This occurs when one is using the transect method to estimate wildlife populations. Prove that the distribution of X satisfies a monotonicity condition, and find the formulas for $U(M)$ and $L(M)$ for a $100(1-2\alpha)\%$ confidence interval for n.

5. Prove: if $X_1, \cdots, X_n$ are independent and identically distributed random variables whose common distribution function is absolutely continuous and uniform over the interval $(0,\theta)$, and where $\Theta = (0,\infty)$, and if

$$M(\mathbf{X}) = \max\{X_i : 1 \leq i \leq n\},$$

then

$$P_\theta([M(\mathbf{X}) \geq x])$$

is nondecreasing in θ for every fixed x.