

THE UNIVERSITY OF TEXAS AT AUSTIN
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
AUSTIN, TEXAS 78712-1276
SZB 564 TEL: (512) 471-2742; (800) 551-0294 FAX: (512) 471-3971

LIS 397.1
INTRODUCTION TO RESEARCH IN LIBRARY AND INFORMATION SCIENCE
IN-CLASS MIDTERM, 1999 April 8, QUESTIONS AND ANSWERS

You will have 50 minutes for this part of the exam. Pace yourself accordingly. Keep this part of the exam; you will need it for the take-home part.

1. (10 points) State a definition of a general hypothesis.

A (general) hypothesis is a statement about the relationship between two or more variables. The stated relationship must be testable; i.e., there must be, at least in principal, some way of checking the hypothesis against reality to determine whether the hypothesis is true (scil., appears consistent with reality) or is false (scil., appears inconsistent with reality).

2. (10 points) State a definition of a statistical hypothesis.

A statistical hypothesis is either (1) a statement about the value of a population parameter or (2) a statement about the probability distribution that a random variable obeys.

3. (20 points each) For each of the following situations, describe succinctly what statistical procedure(s) might be appropriate. This includes some or all of the following: saying what statistical technique would be appropriate; if the technique is one that includes a built-in null hypothesis, stating that hypothesis in terms of the problem situation (i.e., not merely stating the abstract null hypothesis); and indicating briefly what kind of sample(s) and sampling procedure would be involved.

A. You have just been appointed Head of Technical Services in a large academic library in a state university. In appointing you, the library's director stressed the general expectation that, in its next session, the state's legislature was likely to go crazy with the excitement of making budget cuts--rational or not. Therefore, she told you, you must be constantly on the lookout for any and all possible ways of reducing expenses without affecting your library's services to its users.

You decide that one of the first things you need to do is to establish some benchline data for what your current operations are costing you. One of the more costly services your library provides is its OPAC, i.e., its Online Public Access Catalog. Among the benchline data that you would like are data on the numbers of OPAC users and the frequency of uses of (i.e., queries of or accesses to) the OPAC.

The operating system on the mainframe computer that supports your OPAC is capable of reporting the number of current users of the OPAC (i.e., the number of persons currently logged into the OPAC) at any particular time that you ask the system for that datum. How could you use this system capability to provide you with an idea of the typical number of simultaneous OPAC users?

The essence of the problem is that you need to determine the mean number of users logged in at any one time. A pertinent technique is that of constructing confidence intervals for the mean of a population. (Alternatively, you could use the t-test procedure to test the null hypothesis that the population mean has a certain value, i.e., a hypothesis of the form $H_0: \mu = \mu_0$, for some arbitrary value of μ_0 . In this case, the null hypothesis states that the mean number of simultaneous OPAC users has some specified value.)

To construct a confidence interval for the population mean, you would take a random sample of the numbers of OPAC users at various times during the OPAC's hours of operation. Using the observed sample mean

number of users, \bar{x} , the observed sample standard deviation in the number of users, s , and the sample size, n , you would construct the confidence interval for whatever level of confidence you wished to use, e.g., 95% or 99%.

B. After obtaining the benchline data for the current typical number of simultaneous OPAC users, you realize that this number may change over time. You suspect that the usage of the OPAC is increasing fairly rapidly, both because your university's enrollment is increasing and because an ever higher proportion of students, faculty, and staff are users of microcomputers for a variety of purposes and are therefore likely to view using the OPAC as their preferred way of beginning any particular occasion of their using the library.

(In fact, you recently read a *Library Journal* article that is relevant. It points out that although many librarians have assumed that library patrons would be reluctant to shift from manual to automated catalogs, what often happens in fact is that the great majority of patrons are eager and delighted to make the shift. The author reports that in his library, patrons regularly line up to use the OPAC terminals, leaning against the unused card catalog while they wait.)

You decide that for the purpose of long-range planning, you need to know whether OPAC use is increasing in your library and, if so, at what rate. After all, if the rate of use is going up fast, your library will need acquire some more expensive hardware soon, and will need to get a request for the funds into its budget request.

Fortunately (actually, because of good system design), the mainframe's operating system is capable of recording the total number of uses of the OPAC during any given period, e.g., during each 24-hour day. (The total number of uses in a given day is the total number of times during that day that any and all users log into the OPAC for a session of using it, regardless of whether a particular user has used the OPAC earlier that day.) How could you employ this system capability to find out whether the number of OPAC uses appears to be increasing and, if so, what the rate of increase is?

The essence of the problem is that you would like to find out whether there is a trend over time in the number of OPAC uses. The pertinent technique is that of linear regression. To determine whether there appears to be a trend in the number of OPAC uses, you need first to see whether there is a correlation between time and the number of OPAC uses. (Note: you will recall that we discussed in class the fact that this is a widespread and widely accepted use of correlation, despite the fact that it is technically questionable because time is not a Gaussianly distributed variable, contrary to the assumptions on which Pearson correlation is based.)

Your first step is to use the Pearson correlation procedure to test the null hypothesis $H_0: r_{XY} = 0$, where X represents time and Y represents the number of OPAC uses at the time of observation, i.e., the hypothesis that the number of OPAC uses is not correlated with time.

For example, you could take a sample of days over a year and observe the total number of OPAC uses recorded on each day in the sample. In such a sample, X would represent an observed day's number in the year and Y would be the number of uses on day X. If your sample data allowed you to reject the null hypothesis, then (1) you would be able to conclude that there is a real trend over time in the number of uses, and (2) you would be able to use your sample data to construct the regression equation for the data and thus be able to predict the expected number of uses in the future (subject, of course, to the usual qualifications and doubts about predictions into the future).

C. As you continue pursuing your close examination of how your library's OPAC is being used, you start to worry about possible problems that might be caused by peak loads on the mainframe computer that supports the OPAC. This computer also supports the library's circulation-control system. It occurs to you that if peak loads on the OPAC and the circulation-control system tend to occur at almost the same time (e.g., when there is a heavy influx of students into the library), this could cause operational difficulties.

The well designed operating system of your library's mainframe computer can report the number of circulation transactions that occur in any given period and, as noted earlier, can also report the number of OPAC uses in a given period. How could you use these capabilities to try to determine whether peak loads tend to occur concurrently on the OPAC and the circulation-control system?

The essence of this problem is that you wish to determine whether the numbers of uses of the circulation system and of the OPAC during short intervals of time (e.g., 5-minute periods) tend to be correlated, i.e., whether high numbers of circulation-system uses tend to co-occur along with high numbers of OPAC uses.

You need to use the Pearson correlation procedure to test the null hypothesis, $H_0: r_{XY} = 0$, where X represents where X represents numbers of circulation-system uses and Y represents the number of OPAC uses during the same short intervals of observation, i.e., the hypothesis that the number of circulation-system uses is not correlated with the number of OPAC uses.

For example, you could take a random sample of 5-minute intervals during the hours of operation of the circulation system over a period of a week, or better, a month. (There would be no need to make observations when the circulation system is not in use.) In such a sample, X would represent the number of circulation system uses during the 5-minute interval and Y would represent the (maximum) number of simultaneous OPAC uses during the same interval. If your sample data allowed you to reject the null hypothesis, then you would be able to conclude that circulation-system uses and concurrent OPAC uses are correlated, which would tell you that peak loads, and low loads, would tend to co-occur on the circulation and OPAC systems.

D. In the South Texas independent school district for which you work as a school librarian-media specialist, several PTA groups have helped their schools to acquire microcomputer software and even some smaller pieces of hardware, but your school's PTA has not done so—at least, not yet. The incoming president of your school's PTA has, in past years, shown himself to be genuinely interested in your library-media center. He is also concerned about your school's difficulties in raising the reading-skill level of a large fraction of the children.

One day you read in a professional journal about a piece of Macintosh software, *Read 'Em, Cowpoke*, whose vendor claims that it has shown unusual success in improving the reading skills of low-literacy second- and third-grade children in some schools in the Texas Panhandle. An idea strikes you: you will tell the new PTA president about *Read 'Em, Cowpoke*, and try to enlist his aid in persuading the PTA to undertake a fund-raising campaign to acquire two Macintoshes and appropriate software, including *Read 'Em, Cowpoke*, to be placed in your library and under your control. He agrees, the PTA agrees, the campaign is successful, and the hardware and software are acquired and installed in your library, in time for the start of the spring semester.

Next, you enlist the aid of the second- and third-grade teachers in your school. Through their cooperation, you will be able to give some 40 low-reading-skill children in those grades a chance to spend two half-hours each week using *Read 'Em, Cowpoke*. You can measure the reading-level scores of these children before they start using *Read 'Em, Cowpoke* and again after they have spent a semester coming to the library for their hour a week with *Read 'Em, Cowpoke*. How could you use such data to determine whether *Read 'Em, Cowpoke* appears to work for poor readers in your school's second and third grades, in the sense of improving their reading skills?

The essence of this problem is that you want to determine whether the use of the *Read 'Em, Cowpoke* software helps low-literacy second- and third-grade children improve their reading skills. Fortunately, it is clearly possible to observe the variable of interest, reading skill, in each child *before and after* the child uses the software, i.e., it is possible to collect your data in the form of pairs of observations. Hence, the t-test for dependent samples (or, alternatively, the ANOVA procedure for paired samples, also known as ANOVA with repeated measures) is the preferred procedure. It is preferred because by using it, we have to cope with *only* the random variability of each student with himself or herself, rather than having to cope with the *total* variability among the whole set of students in the sample.

You should use the t-test for dependent samples procedure to test your hypothesis that the use of the *Read 'Em, Cowpoke* makes a difference in the reading scores of low-literacy second- and third-grade children. The appropriate null hypothesis is $H_0: \Delta = 0$, where Δ represents the mean difference between pre-Cowpoke and post-Cowpoke reading scores for low-literacy second- and third-grade children. In other words, the appropriate null hypothesis is that the use of the *Read 'Em, Cowpoke* makes no difference in the reading scores of low-literacy second- and third-grade children. You would use the available group of 40 low-reading-skill children, assuming them to be an appropriate random sample; you would administer an appropriate test of reading skill to each child before that child makes use of the *Read 'Em, Cowpoke* software and again after the

child has used that software. You would take the pairs of pre-Cowpoke and post-Cowpoke scores for each child and use them in the t-test for dependent samples. If the result enabled you to reject the null hypothesis that *Read 'Em, Cowpoke* makes no difference in reading skills, you would conclude that *Read 'Em, Cowpoke* appears to be a useful tool for helping low-reading-skill second- and third-grade children learn to read better.