

**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  
TAKE-HOME MIDTERM, 1999 April 8, QUESTIONS AND ANSWERS

The following problems (25 points each) are keyed to Part 3 of the In-Class Midterm, to which you will need to refer for descriptions of the situations. Shortly after the In-Class part of the exam is over, data for problems A, B, C, and D will be available, in Microsoft Excel format, from the course Website, as files 3971p99adata.xls, 3971p99bdata.xls, 3971p99cdata.xls, and 3971p99ddata.xls. Please let me know if you have any problems in accessing these data files.

Please note that, as is explained in the course outline, you are expected to work these problems by yourself. That means you are expected to work them without aid from other students, from the Teaching Assistant for the course, or from the staff of the Information Technology Laboratory. Because I recognize that I may, unintentionally, fail to state problems unambiguously, you should feel free to ask me any questions you wish (but I reserve the right not to answer them fully).

A. You arrange with one of the library's computer programmers to have the mainframe operating system report, at randomly chosen times, during the OPAC's hours of operation, the number of OPAC users at each such reporting time. By thus sampling the numbers of users at 200 randomly chosen times during the next month you obtain the following sample of numbers of users:

1	57	34	74	12	65	10	70	24	58
24	86	38	73	100	0	121	122	117	57
27	85	22	22	29	28	108	6	31	121
29	90	11	24	38	42	45	83	83	35
12	4	37	46	24	57	32	57	120	92
23	73	17	30	15	47	65	12	17	8
0	36	22	63	39	22	12	63	31	34
35	38	7	3	121	53	46	110	91	34
39	44	17	43	92	42	99	142	104	2
18	64	1	73	17	22	46	126	64	63
7	82	35	10	53	74	74	36	1	42
23	8	39	43	84	29	101	110	37	37
3	91	8	39	28	8	68	70	5	29
43	19	32	74	116	40	14	119	34	67
34	16	36	30	4	74	50	59	38	53
37	75	38	29	20	59	98	39	9	74
21	78	28	50	13	2	92	95	22	82
14	6	21	67	16	32	121	141	24	20
12	62	18	58	19	1	123	123	35	41
46	26	29	40	12	84	118	136	1	38

A.1 (15 points) What is the best conclusion you can reach about the value of the mean number of users?

**Here is the result of entering the above data into Microsoft Excel and using the Descriptive Statistics tool on them:**

<i>Numbers of Users</i>	
Mean	47.695
Standard Error	2.491021
Median	38
Mode	29
Standard Deviation	35.22836
Sample Variance	1241.037
Kurtosis	-0.22198
Skewness	0.799269
Range	142
Minimum	0
Maximum	142
Sum	9539
Count	200
Confidence Level(95.0%)	4.912187

The *best* way to report a value for a population mean (here, the mean number of users) is in terms of a confidence interval, rather than merely a point estimate. We conclude that the point estimate of the population mean is  $m = 47.695$ , and that the 95% confidence interval for the population is: We are 95% confident that  $m$  is in the interval  $47.695 \pm 4.912 = (42.783, 52.607)$ . (Note that Excel provides a value for the confidence level, i.e., the half-width of the confidence interval, that is calculated by multiplying the standard error by the exact value of the confidence factor, in this case 1.972, that is provided by the Student's t distribution [calculated at the 0.05 level with 199 degrees of freedom], rather than by the Gaussian value, 1.96. Values of Student's t can be calculated in Excel by the TINV function.)

The above results used the default value of 95% for the confidence interval with the Descriptive Statistics tool. Had we chosen to work at the 99% level, we could have specified this level in the Descriptive Statistics window, and we would have obtained the following bottom row in the Descriptive Statistics display (the other rows would have remained the same as above):

Confidence Level(99.0%)	6.478535
-------------------------	----------

We can thus calculate that: We are 99% confident that  $m$  is in the interval  $47.695 \pm 6.478 = (41.217, 54.173)$ .

A.2 (10 points) What number of users appears to be exceeded just 2% of the time?

To find a number  $U$  such that the number of users can be expected to exceed  $U$  just 2% of the time, we employ the properties of the Gaussian distribution. From a table of the Gaussian distribution, we find that the value 2% or 0.02 for the area under the right tail of the Gaussian curve corresponds to  $z = 2.054$ . This tells us that if we go out to the right of the mean, by 2.054 standard deviations, we will have the desired value of  $U$ . In short,  $U = \bar{x} + 2.054s = 47.695 + 2.054 \cdot 35.228 = 120.053$ .

An alternative to looking up the tabled value corresponding to the 2% right tail is to use the Excel function, NORMINV(probability, mean, standard\_dev) where the "probability" parameter specifies the cumulative Gaussian probability. Since we are seeking the value corresponding to the 2% right tail, the pertinent cumulative Gaussian probability value is 98% or 0.98; and you can easily confirm that NORMINV(0.98, 0, 1) = 2.0537.

A less sophisticated, but acceptable, estimate of  $U$  can be obtained by using Excel's Rank and Percentile tool. Applying this tool to the sample data shows that the two sample points 123 correspond to percentile 97.4 and sample point 126 corresponds to percentile 98.4. We can interpolate between 123 and 126 by noting that 98.0

is 6/10 of the way up from 97.4 to 98.4; analogously,  $U$  should be 6/10 of the way up from 123 to 126. Since the distance between 123 and 126 is 3, we use  $0.6 \cdot 3 = 1.8$  to yield  $123 + 1.8 = 124.8$  as the point estimate of  $U$ .

B. After thinking about the problem, you decide that looking at the sequence of weekly totals of numbers of OPAC uses (i.e., the total number of times that the OPAC was accessed during each week) is an appropriate way to proceed. Again you arrange with one of the library's programmers to have the total number of uses during the past seven days stored at the end of each week. At the end of a year, you have obtained the following sample of data:

WEEK	USES	WEEK	USES	WEEK	USES	WEEK	USES
1	917	14	784	27	1047	40	1231
2	910	15	1090	28	1104	41	1056
3	904	16	1024	29	861	42	1270
4	769	17	1104	30	867	43	988
5	805	18	1026	31	1096	44	1179
6	783	19	963	32	1049	45	1142
7	998	20	739	33	881	46	815
8	892	21	804	34	1227	47	951
9	981	22	1174	35	823	48	882
10	762	23	1125	36	964	49	1007
11	1016	24	1050	37	959	50	1192
12	1048	25	959	38	1008	51	1090
13	1028	26	898	39	1247	52	1063

B.1 (15 points) What can you conclude about the rate of increase, if any?

Here is the result of entering the above data into Microsoft Excel and using the Regression tool on them:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.415803
R Square	0.172892
Adjusted R Square	0.15635
Standard Error	126.6758
Observations	52

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	167714.4783	167714.4783	10.45161228	0.002172126
Residual	50	802337.8294	16046.75659		
Total	51	970052.3077			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	891.1085973	35.64646926	24.99850941	6.44446E-30	819.5105274	962.7066671	819.5105274	962.7066671
Week	3.784000683	1.170468043	3.232895341	0.002172126	1.43304547	6.134955896	1.43304547	6.134955896

From the above data we see that the sample Pearson correlation coefficient (which Excel calls "Multiple R") is 0.4158, well above the tabled threshold value, 0.273, for  $df = 50$  at the 5% level of significance with  $df=50$ ; hence, we can easily reject the null hypothesis that  $r = 0.0$ . Alternatively, we see that Significance F = 0.002,

which is lower than the 5% level of significance at which we choose to run this test. Therefore, we are entitled to believe that there really is a non-zero trend.

As still another alternative way of deciding whether you are entitled to reject the null hypothesis, you could use the formula for converting an observed value of the sample Pearson product-moment correlation coefficient into a Student's  $t$  value that can be checked against a table of Student's  $t$  or by the use of the TINV function in Excel. This formula is

$$t = r_{XY} \sqrt{\frac{n-2}{1-r_{XY}^2}}$$

The number of degrees of freedom to be used in checking the resulting value of  $t$  against a table of Student's  $t$  is  $n-2$ . Here we find  $t = 3.232$ , which is well above the tabled threshold of 2.009 for  $df = 50$  at the 5% level of significance, so that we can reject the null hypothesis of no correlation.

B.2 (10 points) On the basis of your observations, what can you predict as the probable number of uses halfway through the second year (to be specific, for the twenty-seventh week in the second year)?

The regression equation is  $\hat{Y} = B_0 + B_1X = 891.109 + 3.784X$ , so for the twenty-seventh week in the second year, we have  $\hat{Y} = 891.109 + 3.784X = 891.109 + 3.784(52 + 27) = 1,190.045$ . That is, we can predict that halfway through the second year there will be an average of approximately 1,190 uses of the OPAC each week.

Alternatively, you can use the Excel function FORECAST(X, known\_y's, known\_x's), where X is the data point for which you want to predict a value, the known\_y's are the range of observed values of the dependent (or predicted) variable, and the known\_x's are the range of observed values of the independent (or predictor) variable. Assuming that you loaded the data into cells A2:A53 for the weeks and B2:B53 for the numbers of OPAC uses (and used cells A1 and B1 for labels), you would find FORECAST(79, B2:B53, A2:A53) = 1190.045.

C. With the cooperation of your friendly computer programmer, you arrange for the operating system to record the following data. It will note (a) the maximum number of OPAC users who are on line at any given second during each 5-minute period of observation and (b) the number of circulation transactions occurring during the same 5-minute period. These 5-minute periods of observation are to be chosen at random from all 5-minute periods occurring during the hours that the library is open. (After all, circulation transactions can occur only during open hours, unlike OPAC accesses, which can occur 24 hours a day--or, at least, whenever the computer is running.) Again, you set things up so that some 200 observations of pairs of values of OPAC accesses and circulation-control transactions will be recorded during the next month. You obtain the following data, in which the OPAC columns contain the maximum numbers of OPAC users who are on line at any given time during each 5-minute period of observation, and the CCT columns contain the numbers of circulation-control transactions carried out during the same 5-minute period:

OPAC	CCT	OPAC	CCT	OPAC	CCT	OPAC	CCT	OPAC	CCT
117	10	111	4	89	57	123	63	78	35
68	45	137	59	117	9	36	12	122	55
120	44	57	46	88	31	106	4	104	62
27	62	70	39	37	28	149	47	73	24
36	36	43	41	108	40	31	61	40	34

22	70	26	35	89	45	148	50	8	5
37	32	93	6	44	38	89	13	87	35
119	34	133	13	75	11	139	60	130	11
38	50	119	50	62	66	0	55	106	27
149	48	16	22	91	13	62	30	146	63
21	31	36	1	79	30	38	19	87	55
76	7	13	29	38	26	7	17	120	42
18	2	13	22	15	15	32	46	40	11
55	31	2	11	84	41	49	52	17	55
78	3	17	1	62	4	33	18	122	19
1	32	10	21	80	30	11	30	26	37
12	34	18	25	76	59	1	50	52	21
77	48	34	23	62	40	43	44	104	0
37	36	10	22	36	21	41	1	18	16
21	49	10	12	29	1	34	46	69	36
24	13	25	37	144	35	27	58	73	19
19	65	79	51	12	0	2	54	46	60
20	65	110	41	94	16	99	3	56	46
79	55	100	57	141	70	81	23	38	25
22	16	147	17	44	55	145	61	118	47
146	51	51	45	119	71	132	14	97	34
140	56	106	44	97	20	92	72	31	21
65	65	58	25	41	29	64	3	112	74
69	54	12	64	138	48	0	34	144	39
51	56	38	13	136	9	60	8	26	71
96	1	29	3	26	46	35	24	25	41
87	11	13	14	52	38	51	15	109	34
36	35	3	2	0	53	27	48	44	57
32	6	19	10	77	35	50	52	3	39
93	32	4	28	37	60	15	37	80	3
1	29	8	9	63	41	56	19	38	19
63	37	4	6	76	26	27	6	28	44
91	37	40	14	73	44	12	33	91	11
80	36	11	16	43	47	6	26	78	16
79	18	16	24	20	5	31	18	62	2

What can you conclude?

Here is the result of entering the above data into Microsoft Excel and using the Correlation tool on them:

	OPAC	CCT
OPAC	1	
CCT	0.190828	1

This tells us that the sample Pearson correlation coefficient is  $r = 0.19$ . Since there are 200 pairs of observations, the number of degrees of freedom,  $df = 198$ . The value  $r = 0.19$  is above the tabled threshold value of 0.138 for the 5% level of significance, so we can reject the null hypothesis of no correlation in the population. However, though the correlation is real, it is not strong, so we conclude that there is only a modest tendency for peak values of circulation-system use and OPAC use to coincide.

Alternatively, you could use the  $r$  to  $t$  transformation mentioned above in the answer to Part B.1. Using that transformation, we find  $t = 2.735$ , which is well above the tabled threshold value of 1.972 for  $df = 198$  at the 5% level of significance, so that we can reject the null hypothesis of no correlation in the population.

D. You measure the reading-level scores of each of the 40 children at the start of the spring semester, and again at the end of the semester. Here are your data:

CHILD	STARTING SCORE	ENDING SCORE	CHILD	STARTING SCORE	ENDING SCORE
1	32	43	21	49	62
2	32	42	22	21	33
3	17	44	23	25	55
4	38	38	24	12	30
5	33	49	25	46	20
6	35	40	26	34	27
7	37	21	27	38	49
8	27	64	28	30	21
9	10	40	29	36	37
10	10	47	30	45	33
11	15	26	31	15	52
12	29	51	32	41	54
13	35	57	33	26	56
14	34	68	34	44	52
15	24	45	35	46	31
16	49	61	36	19	43
17	14	31	37	15	47
18	47	20	38	15	61
19	49	34	39	18	66
20	26	67	40	24	35

What can you conclude?

Here is the result of entering the above data into Microsoft Excel and using the t-test paired two-sample tool on them:

t-Test: Paired Two Sample for Means

	STARTING SCORE	ENDING SCORE
Mean	29.8	43.8
Variance	145.3948718	194.1641026
Observations	40	40
Pearson Correlation	-0.097150047	
Hypothesized Mean Difference	0	
Df	39	
t Stat	-4.58951565	
P(T<=t) one-tail	2.26522E-05	
t Critical one-tail	1.684875315	
P(T<=t) two-tail	4.53044E-05	
t Critical two-tail	2.022688932	

We see that Excel reports the probability of our observed result,  $P(T \leq t)$  two-tail, as 0.000045, which is far below our preferred level of significance,  $\alpha = 0.05$  (and, indeed, well below  $\alpha = 0.01$ ). So we reject the null hypothesis that the *Read 'Em, Cowpoke* software has no effect on the reading skills of low-literacy second- and third-grade students. Since the mean pre-*Cowpoke* reading level in the sample was 29.8, and the mean post-*Cowpoke* reading level was 43.8, we can conclude that this software program is effective in helping such students raise their reading skills.

Alternatively, we could note (since Excel calls the observed value of the  $t$  statistic “t Stat”) that  $t_{obs} = -4.5895$ , which in absolute value is well above the tabled threshold value of the Student's  $t$  distribution for 39 degrees of freedom at the 5% level of significance, which Excel reports as 2.0227. Hence, at the 5% level of significance we reject the null hypothesis of no difference between the means.

Had we chosen to work at the 1% level of significance, we would still reject the null hypothesis, since the reported value of  $P(T \leq t)$  two-tail is well below 0.01, as noted above. Of course, if we had started by working at the 1% level of significance, we would have chosen to set the alpha level at 0.01 when we initially employed the  $t$ -test paired two-sample tool, and Excel would have provided a report that contained values corresponding to the 1% level of significance.

Thus, working at either the 5% or the 1% levels of significance, we reject the null hypothesis and conclude that the observed difference in the pre-*Cowpoke* and post-*Cowpoke* sample means is due to a real difference in the corresponding population means, i.e., that *Read 'Em Cowpoke* really does help low-literacy students.

Alternatively, we could use a tool for carrying out the ANOVA procedure for repeated measures if we had such a tool available. Unfortunately, Microsoft Excel 97 does not include such a tool; perhaps Excel 2000 will.

**Note:** To carry out an investigation like the above in the real world, one should employ also a control group of low-literacy second- and third-grade students who are taught by conventional reading-instruction methods. That consideration is ignored in this problem, because the focus of this exam is basic statistical techniques rather than design of experiments as a whole.