SIMPLE STATISTICS FOR MANUAL ANALYSIS OF FARM SURVEY DATA

L.W. Harrington*

Economics Training Note
1981

* Economist at CIMMYT, Mexico. The views expressed are not necessarily those of CIMMYT.

Training Note

SIMPLE STATISTICS FOR MANUAL ANALYSIS OF FARM SURVEY DATA

L.W. Harrington

1.0 Introduction

When conducting formal surveys for the purpose of helping focus on-farm agronomic experiments, experience indicates that a relatively small sample of farmers is normally sufficient for a given recommendation domain or "homogeneous" group of farmers. A sample size of 40-60 farmers is normally large enough for reasonably precise estimates of variables measuring farmer practices and problems in the production of a target crop, and the cropping system in which that crop is grown (Byerlee, Collinson et al, 1980). As one adds domains, of course, minimum sample size for formal surveys increases. Nonetheless, researchers involved in on-farm research will frequently find themselves conducting and analyzing surveys with sample sizes of less than one hundred farmers.

In these cases, manual analysis of survey data can be more efficient than computer analysis. This is because there is a high "fixed cost" associated with setting-up analysis via computer. Before beginning the actual analysis, researchers must formulate a code-book that describes the codes corresponding to each possible answer for all questions, code the data, have the coding-sheets key-punched, formulate

an input statement and run and edit a data listing. In the time this takes, the researcher might well have finished the entire analysis if performed manually. 1/

Analysis of survey data to help focus on-farm experiments rarely requires sophisticated procedures. Cross-tabulations and comparisons of means for sub-populations are used to test hypotheses on the delineation of recommendation domains. Means and simple frequencies are employed to describe the current farming system and the management of the target crop within that system, for each domain. Cross-tabulations and comparisons of means for sub-populations are then used to identify factors that limit production or income, and the interactions between these limiting factors and the farmers' current practice.

The basic analytical tools used, then are: (1) means, (2) simple frequencies, (3) cross-tabulations, and (4) comparisons of the means of two sub-populations. Of these four tools, the first two require no further attention. The purpose of this note is to examine two statistics associated with cross-tabulations (chi-square) and comparisons of sub-population means (Student's t), emphasizing their manual calculation and interpretation.

2.0 Cross-Tabulations and Chi-Square

Researchers often wish to arrange survey data by two variables, or criteria of classification, wishing to know if the two variables are

1/ On the other hand, if the researcher is well-acquainted with computer analysis and expects to conduct a large number of statistical tests due to study area complexity, computer analysis can be more rapid.

independent of one another. For example, it may be of some interest to determine whether or not the planted maize variety is related to the cropping cycle (dry vs wet), or if the tillage system used is independent of the land type. The first step, then, is to cross-tabulate the data by noting the frequencies associated with each possible combination of the two classification variables.

An example is given in Figure 1 for maize variety by land type. From Figure 1 it appears that farmers plant hybrids on flat clays and the local variety on other land types.

At times, however, an apparent relationship between two variables will be only due to random chance. A common way to formally test for a relationship between two variables is to calculate a chi-square χ^2 for the cross-tabulation between those variables.

Two cautions are in order. First, cross-tabulations and chi-square tests should be used with discrete rather than continuous variables. Second, the results of a chi-square only test the probability that a relationship exists, but does not indicate either the strength or the direction of that relationship (Loether and Metovish, 1974).

The calculation of a chi-square for a given cross-tabulation is based on the difference, for each "cell", between the observed frequency and the frequency that would be expected if no relation existed between the two variables.

Figure 1: Cross-Tabulation of Maize Variety by Land Type

Land Type	Maize Variety		
	Hybrids	Local Variety	Sum
Flat Clay	20	10	30
Flat Sand	3	7	10
Hilly Gravel	7	13	20
Sum	30	30	Sample Size = 60

For the calculation of a chi-square, the following formula is used (Huntsburger and Billingsley, 1973):

$$\chi^2 = \sum_{ij} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}}$$

where Y_{ij} = observed frequency for the "cell" in row i and column j.

where E_{ij} = expected frequency for the "cell" in row i and column j.

and $E_{ij} = \frac{R_i C_i}{n}$

where R_i = sum for row i

C_j = sum for column j

and n = sample size = $R_i = C_j$

Using the data from Figure 1, we find, for example, that $R_1 = 30$, $R_2 = 10$, $R_3 = 20$, $C_1 = 30$, and $C_2 = 30$, n = 60, $Y_{11} = 20$, $Y_{12} = 10$, so on. The calculations for the corresponding chi-square may be seen in Figure 2. The calculated chi-square for the cross-tabulation of maize variety by land type is $\chi^2 = 6.74$.[1/]

How is this number interpreted? The larger the chi-square, the lower the probability that the relationship between two

---

[1/] According to Yamane, 1964, the expected frequency for each cell (E_{ij}) should be greater or equal to five for X^2 tests to be accurate. If several cases of E_{ij} less than 5 occur, the data should be re-grouped into fewer groups.

Figure 2: Calculating a Chi-Square for the Cross-Tabulation of Maize Variety by Land Type

Cell	Y_{ij}	E_{ij}	$\frac{(Y_{ij} - E_{ij})^2}{E_{ij}}$
Y_{11}	20	15	1.67
Y_{12}	10	15	1.67
Y_{21}	3	5	0.80
Y_{22}	7	5	0.80
Y_{31}	7	10	0.90
Y_{32}	13	10	0.90
		χ^2 =	6.74

1/ $\frac{R_1 \times C_1}{n} = \frac{30 \times 30}{60} = 15$, etc.

2/ $\frac{(20-15)^2}{15} = 1.67$, etc.

variables is due to chance alone. But how large is "large enough"? To answer this question, one must compare the calculated chi-square with the "tabular" chi-square of the corresponding degrees of freedom and desired level of significance (normally 5% or 10%). The degrees of freedom for a given cross-tabulation are found by the relation:

$$df = (r - 1)\ (c - 1)$$

where r = number of rows and c = number of columns.

In our example (maize variety by land type) we are dealing with a three by two table, so degrees of freedom equals :

$$df = (3 - 1)\ (2 - 1) = 2$$

Using a 5% level of significance, and two degrees of freedom, a glance at a table of the chi-square distribution (see Appendix) informs us that the tabular chi-square of interest is 5.99. As our calculated chi-square (6.74) is greater than the appropriate tabular value (5.99), we may conclude that at the 5% significance level there does exist a relation between maize variety and land type.

3.0 Comparisons of Means for Sub-Populations: Student's t

Researchers often wish to compare sample means for the same variable, for two different sub-populations. For example, it may be of some interest to determine whether farmers using official bank financing use the same or a different dose of nitrogen on their maize,

when compared with farmers who do not use this source of funds. Equally, researchers may wish to use their survey data to determine if farmers plant different maize varieties at the same or at different densities.

A "large" difference between sample means leads one to suspect that a "significant" difference does, in fact, exist for the sub-populations themselves. But how "large" must this difference be in order to be "significant"?

In Figure 3, sample data are shown for N dose for two sub-populations: users and non-users of official bank financing. It appears that bank users use a bit more N ($\bar{X}$ = 92 kg/ha N) than non-users ($\bar{X}$ = 76 kg/ha N). Is this apparent difference real or is it only due to chance?

A formal test using the "t" statistic is of help in this situation. When calculating a "t" for the purpose of comparing the means of two sub-populations, the following formula is used (Huntsberger and Billingsley, 1973):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_d}$$

where $\bar{X}_1$ = sample mean of sub-population 1

$\bar{X}_2$ = sample mean of sub-population 2

S_d = standard deviation of the difference between the two means, which is found by:

Figure 3: N Dose (kg/ha) by Source of Financing

	Bank Users	Self-Financed
	87	60
	90	120
	120	110
	88	40
	70	70
	90	50
	87	70
	95	80
	60	110
	110	50
	120	
	90	
$\bar{X}$	92.2	76.0
S	17.8	28.4
N	12	10

$$S_d^2 = S^2 \left(\frac{1}{n_i} + \frac{1}{n_2}\right)$$

where

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and

S_1^2 = sample variance sub-population 1

S_2^2 = sample variance sub-population 2

n_1 = sample size sub-population 1

n_2 = sample size sub-population 2

The corresponding calculations are relatively easy because most modern hand calculators will automatically compute $\overline{X}$ and S_i^2 for a series of observations.

For the data presented in Figure 3, a t statistic is calculated as follows (bank users = sub-population 1; non-users = sub-population 2):

$$S^2 = \frac{(12-1)\ 17.8^2 + (10 - 1)\ 28.4^2}{12 + 10 - 2}$$

$$= 537.2$$

Therefore,

$$S_d^2 = 537.2 \left(\frac{1}{12} + \frac{1}{10}\right)$$
$$= 98.5$$

and $S_d = 9.9$

The corresponding t equals

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_d}$$
$$= \frac{92.2 - 76}{9.9}$$
$$= 1.64$$

How is this number interpreted? First, the larger the calculated t value, the more likely the apparent difference between sample means indicates a real difference in means between the two sub-populations. But how large must the calculated t value be? One must compare the calculated t value with a "tabular" t of the correct degrees of freedom and the desired level of significance (normally 5% or 10%).

The degrees of freedom of the t statistic calculated above are readily obtained:

$$df = n_1 + n_2 - 2$$

In our example, $n_1 = 12$ and $n_2 = 10$, so $df = 12 + 10 - 2 = 20$.

Setting up a one-tailed t test, the null hypotheses should be:

$$HO: U_1 \leq U_2$$

where U_1 and U_2 are population means (not sample means) for subpopulations 1 and 2.

This indicates that the sub-population of bank users apply the same amount of or less N than non-users. The alternative hypothesis should be:

$$Ha: \quad U_1 > U_2$$

(bank users apply more N than non-users). In general, in comparing means of two sub-populations we expect one sub-population mean to be larger than the other: it is this expected relation that should be expressed in the alternative hypothesis.

If the calculated t is greater than the tabular t, we reject Ho.

In our example, the calculated t = 1.64. The tabular t for df = 20 and a 10% significance level = 1.325 (see Appendix), so we may conclude that the subpopulation of bank users does apply more N than self-financed farmers, at a 10% significance level.

A final caution is in order: The above t test that may only be conducted when one assumes that two independent samples came from normal populations with the same standard deviation.

REFERENCES

Byerlee, D., M. Collinson, et al, 1980. Planning Technologies Appropriate to Farmers: Concepts and Procedures. CIMMYT.

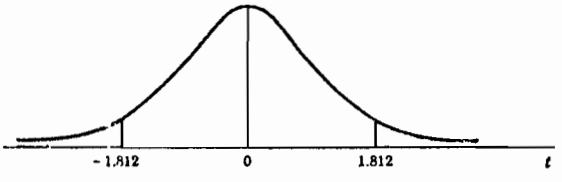
Huntsburger, D. and P. Billingley, 1973. Elements of Statistical Inference. Boston: Allyn and Bacon, Inc.

Loether, H. and D. McTavish, 1974. Descriptive Statistics for Sociologists. Boston: Allyn and Bacon, Inc.

Yamane, T., 1964. Statistics: An Introductory Analysis. New York: Harper and Row.

APPENDIX

Percentage Points of the t Distribution



Example

For 10 degrees of freedom:

$P\,[t > 1.812] = 0.05$

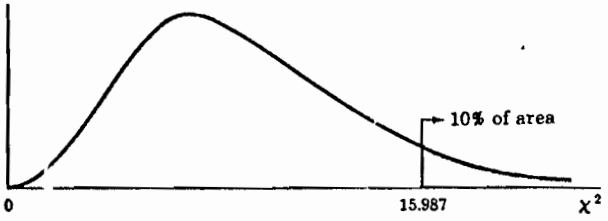
$P\,[t < -1.812] = 0.05$

n \ α	.25	.20	.15	.10	.05	.025	.01	.005	.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.685	.857	1.059	1.318	1.711	2.064	2.492	2.397	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

SOURCE: This table is abridged from Table III of Fisher & Yates: *Statistical Tables for Biological, Agricultural and Medical Research* published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.

APPENDIX

Percentage Points of the χ^2 Distribution



Example

For 10 degrees of freedom:

$P[\chi^2 > 15.987] = .10$

n	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.0³157	.0³628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.033	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit variance remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve.

SOURCE: This table is abridged from Table IV of Fisher & Yates: *Statistical Tables for Biological, Agricultural and Medical Research* published by Oliver & Boyd Ltd., Edinburgh, and by permission of the authors and publishers.