DOCUMENT RESUME

ED 415 701 FL 025 009

AUTHOR Hiroto, Nagato
TITLE Bare Minimum Knowledge for Understanding Statistical Research Studies.
PUB DATE 1997-00-00
NOTE 14p.; In: Classroom Teachers and Classroom Research; see FL 024 999.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Research; Classroom Techniques; Comparative Analysis; *Data Interpretation; Foreign Countries; *Information Utilization; Language Teachers; *Research Methodology; Research Utilization; Second Language Instruction; Second Languages; Statistical Analysis

ABSTRACT

Designed for language teachers who find reading statistical research difficult but necessary, this article focuses on the minimal knowledge needed about statistical techniques for interpreting research findings. It first examines the statistical reasoning underlying quantitative, empirical studies, including normal distribution, standard deviation, and three indices of central tendency (mean, median, and mode). It then looks briefly at three kinds of comparisons most frequently reported in quantitative studies: comparisons of means, frequencies, and correlational coefficients. Statistical tools typically used for each comparison are touched upon, and useful checkpoints that readers of statistical studies should be equipped with are also discussed. (Contains six references.) (MSE)

Chapter 10

Bare Minimum Knowledge for Understanding Statistical Research Studies

Nagata Hiroto

Yokohama National University

This article is for those language teachers who find reading statistical research troublesome, but necessary. First it will examine the statistical reasoning underlying quantitative, empirical studies. Normal distribution, standard deviation, three indices of central tendency (mean, median, and mode), and skewness will be discussed. Then, I will briefly look at the three different kinds of comparisons most frequently reported in quantitative studies, namely, comparisons of means, frequencies, and correlation coefficients. Statistical tools typically used for each comparison will be touched upon. In the course of the discussion, useful checkpoints will be provided that readers of statistical research studies—not to mention researchers themselves—should be equipped with.

Normal Distribution

One of the most interesting phenomena that occurs and recurs in nature, is a pattern called normal distribution (also called bell or normal curve). Suppose, for instance, if you decided to plot on a chart the lengths of a bagful of bean sprouts, the result might look something like the one in Figure 1.1. The shape is not yet "normal," but the larger the number of sprouts you measure, the more "normal" the form of the scatterplot becomes, and it will eventually look some-

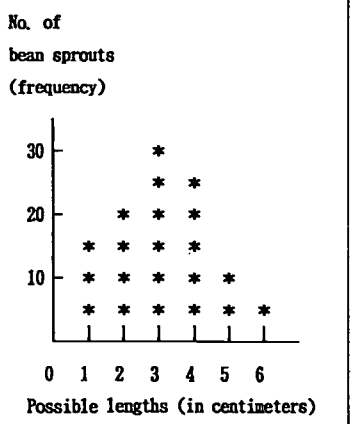


Figure 1.1
Length of bean sprouts
(Histogram)

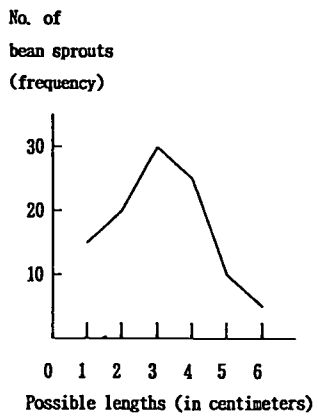


Figure 1.2
Length of bean sprouts
(Frequency polygon)

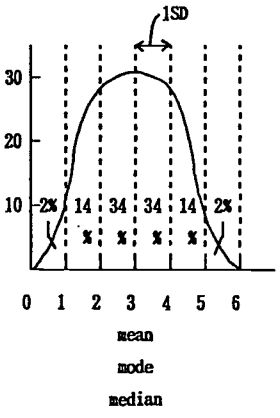


Figure 1.3
The approximate percentages falling
through the normal distribution

thing like the frequency polygon shown in Figure 1.2. As a language teacher, you might want to plot your students' test scores. In this case, most probably, you would also find something close to a normal distribution among your students when the number of them is sufficiently high (usually, at least 30).

Central Tendency

Let us now briefly review three indices of central tendency: the mean, the mode, and the median. The mean is the arithmetic average, and is the most commonly reported indicator of central tendency. The mode is the most frequent score in a set of data. The median is the score which divides the entire data in half. If the mean and the median coincide, and if such a distribution has a single mode, the frequency distribution will show an ideal symmetrical curve.

What is intriguing about such normal distribution is that we would expect about 68% of all the scores to fall within one standard deviation of the mean, and 95% to fall within two standard deviations (see Figure 1.3 above). Standard deviation shows dispersion, or in other words, how much the scores vary away from, or spread out around, the mean (Brown, 1991, p. 574).

Figure 2.1 A Symmetrical Distribution

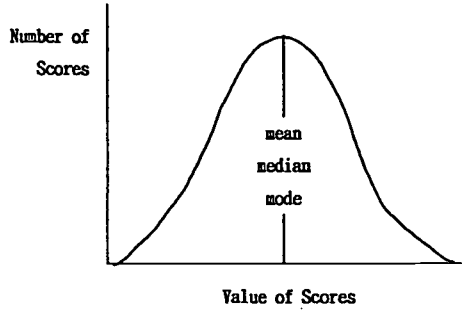


Figure 2.2 A Nagatively Skewed Distribution

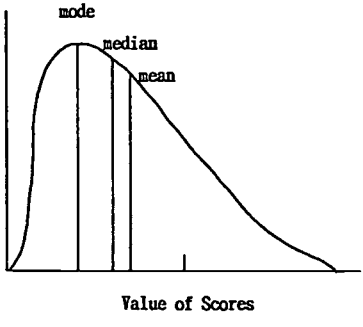


Figure 2.3 A Positively Skewed Distribution

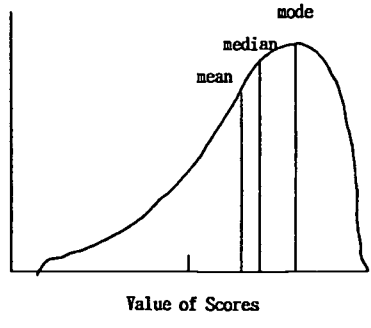


Table 1 Descriptive Statistics for Comprehension

Group	m	SD	min	max	n
G1	3.80	2.20	.0	10.0	75
G2	2.76	1.67	.0	9.0	75
G3	4.79	2.47	.0	13.0	75
G4	4.04	1.88	.0	10.0	75

This normal distribution, in its perfect form, however, rarely occurs in the field of language teaching unless the number of subjects is extremely large. The values of the measures of central tendency usually differ. When the value of mean is influenced by the size of extreme scores, it is pulled toward either end of the distribution in which the extreme scores lie, as shown in Figures 2.2 and 2.3, and the distribution will show an asymmetrical curve (described as being skewed).

When we read statistical studies, we should look to see how far the distribution is skewed from the normal curve, because statistical analyses are based on normal distribution. If you find the distribution is extremely skewed (not close to normal), you should look for the kind of interpretation that is made by the researcher of the study. The skew of a distribution can be identified by comparing the mean and the median without necessarily constructing a histogram or frequency polygon (see Figures 1.1. and 1.2 for examples). When the distribution is skewed toward the lower end, or negatively skewed, the mean is always smaller than the median, and the median is usually smaller than the mode (see Figure 2.2). When a distribution is skewed toward the higher end, or positively skewed, the mean is always greater than the median and the median is usually greater than the mode (see Figure 2.3). Unfortunately, most statistical studies do not provide information on the median. Thus, we usually have to turn to other indices usually provided in the table called "descriptive statistics." Table 1 is an abbreviated version of one such example. It displays such vital information as the number of students involved in the study (n), the mean (m), the standard deviation (SD), and minimum (min) and maximum (max) scores.

As is shown, four groups (G1 to G4) each consisting of 75 subjects took a comprehension test of some kind, and their scores ranged from zero to 13. The means (m) for all groups are pulled slightly towards the lower ends. We know this because all mean scores are lower than the halfway point scores between the lowest possible scores of zero and the highest possible scores of ten, nine,

thirteen, and ten for each group, respectively. However, the scores of each group are spread out to a reasonable degree (we know this because there is room for two standard deviations above and below the mean within the range of the lowest possible and highest scores), therefore, we can conclude that the skewness is not a serious problem here. The table also shows that in terms of means, Group 3 comes first, and Groups 4, 1, and 2 follow. The standard deviation (SD) for Groups 1 and 3 are higher, however, indicating more spread or greater dispersion.

What we would like to know, then, is whether these mean differences are significant. Can we confidently say that they are really (statistically) different? In other words, did the differences we observed happen accidentally, or did they manifest themselves because of some other systematic factors? In order to answer this question, researchers conduct statistical analyses. (For a clear and concise account on the logic of statistical inference behind these analyses, see Nunan, 1992, pp. 28–37).

Three Different Kinds of Comparisons

The three most common types of statistical tests typically reported are mean comparisons, frequency comparisons, and comparisons of correlation coefficients to zero. Let us look briefly at each of them.

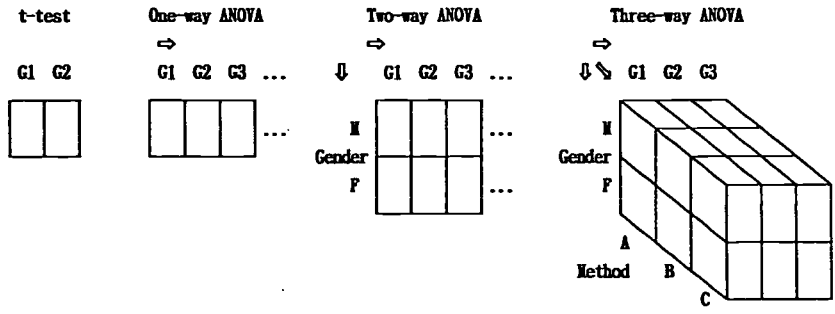
Mean Comparisons

A mean comparison is comparing two or more groups by comparing their average scores on some test. There are four kinds of statistical tests typically used within this category. They are t-test, one-way ANOVA, n-way ANOVA (factorial design), and MANOVA.

The t-test is used to compare the means of two groups, whereas one-way ANOVA is used to examine the differences in more than two groups. N-way ANOVA enables us to analyze the effect of different treatments in more complex conditions, such as different proficiency levels, or different types of learners (e.g., learners with different learning styles). Therefore, the comparisons are drawn in more than two directions (actually, N directions) in n-way ANOVA. In other words, there are more than two (N) independent variables. However, ANOVA is just an exploratory tool, which only tells the researcher that there are some significant differences somewhere in the data but does not specify where they are. Subsequent analyses are needed for the researcher to identify where the significant differences lie (see the one-way ANOVA section below). Still, however, no information can be obtained about the magnitude of the effect.

Diagramatic representations of each of the tests are on the following page.

Now, let us use the data in Table 1 and actually run these statistics to see if we can make any sense out of the computer printouts.



Diagramatic representations of types of designs appropriate for t-test, one-way ANOVA, and n-way ANOVA

T-test: For mean comparison between two groups. For the sake of simplicity, let us assume there are only two randomly chosen groups (groups 2 and 3 in Table 1) learning a set of new vocabulary, and we are comparing the performances of these two groups. Group 2 learned a set of vocabulary using only a listening task, whereas Group 3 learned the same set through the same listening task, but with the help of written input. After the task, we investigate the effect of the written input by giving both groups the same vocabulary test.

Below is a computer printout (SPSS/PC+) of the t-test results. It shows the t-value to be -5.89, which is significant at the .05 level. This means that Group 3 subjects who learned a set of vocabulary through a listening task with the help of written input learned better than those in Group 2. This will be reported as: $t = -5.89$; $p \leq .05$. This significance level (called alpha decision level) should be set at the beginning of the study. Usually, it is set at the conservative $\alpha < .01$, or at the more liberal $\alpha < .05$.

One-way ANOVA: For comparing differences in more than two groups. Suppose there were four groups (groups 1, 2, 3, and 4 in Table 1). Each group was given a different kind of treatment, for instance, they were taught by using the G-T method, TPR, Silent Way, and Suggestopedia. Now, we want to know if these different treatments had any effect. Since the comparisons are made in only one direction, that of the independent variable, which is "method," this ANOVA is called a "One-way" ANOVA.

Below is a computer printout of the one-way ANOVA. F probability (.0000) indicates that there were significant differences between the four groups, but it

Independent samples of GRP

Group 1: GRP EQ 2.0 Group 2: GRP EQ 3.0

t-test for: COMP

	Number of Cases	Mean	Standard Deviation	Standard Error
Group 1	75	2.7600	1.667	.193
Group 2	75	4.7867	2.468	.285

		Pooled Variance Estimate			Separate Variance Estimate		
F Value	2-Tail Prob.	t Value	Degrees of Freedom	2-Tail Prob.	t Value	Degrees of Freedom	2-Tail Prob.
2.19	.001	-5.89	148	.000	-5.89	129.91	.000

[SPSS/PC+ printout 1]

does not specify where those differences lie. Also, although descriptive statistics (mean and standard deviation) provides an indication of where the differences are likely to exist (e.g., of all the mean differences, that between G1 and G4 is the smallest), these insights must be checked statistically. The post hoc Scheffe contrast test does this job. As shown at the bottom of the printout, where an asterisk (*) denotes pairs of groups significantly different at the .05 level, differences exist between Groups 1 and 2, 2 and 3, 2 and 4, and 1 and 3, but not 1 and 4.

Readers are advised not to use multiple t-tests here. Although there have been several studies using multiple t-tests published even in prestigious journals in our field, using t-tests repeatedly (for multiple comparisons) is not recommended because the more comparisons you make, the more chances of creating spurious results.

N-way ANOVA: For examining the effect of several variables studied simultaneously, as well as the interactions among the variables. As is shown diagramatically above, the comparisons are made in more than two directions (i.e., in N ways) in N-way ANOVA. (For a more detailed account, see Hatch & Lazaraton, 1991, pp. 301–331.)

MANOVA: For research designs including more than one dependent variable. Language studies often include two or more dependent variables which are related to each other. For example, if three different kinds of tests were given to

_____ ONEWAY _____

Variable COMP
By Variable GRP

Analysis of Variance

Source	D.F.	Sum of Squares	Mean Squares	F Ratio	F Prob.
Between Groups	3	157.8000	52.6000	12.1909	.0000
Within Groups	296	1277.1467	4.3147		
Total	299	1434,9467			

_____ ONEWAY _____

Variable COMP
By Variable GRP

Multiple Range Test

Scheffe Procedure
Ranges for the .050 level -
3.98 3.98 3.98
The ranges above are table ranges.
The value actually compared with Mean(J) - Mean (I) is ..
1.4688 * Range * Sqrt (1/N(I) + 1/N(J))

(*) Denotes pairs of groups significantly different at the .050 level

---

_____ ONEWAY _____

Variable COMP
(Continued)

Mean	Group	Grp 2	Grp 1	Grp 4	Grp 3
2.7600	Grp 2				
3.8000	Grp 1	*			
4.0400	Grp 4	*			
4.7867	Grp 3	*	*		

[SPSS/PC+ printout 2]

LIFE Is life exciting or dull? by SEX Respondents's sex

SEX Page 1 of 1

Count / Exp Val / Residual	Male 1	Female 2	Row Total
LIFE			
1 Excited	300 279.0 21.0	384 405.0 -21.0	684 46.8%
2 Not excited	296 317.0 -21.0	481 460.0 21.0	777 53.2%
Column Total	596 40.8%	865 59.2%	1461 100.0%

Chi-Square	Value	DF	Significance
Pearson	5.00467	1	.02528
Continuity Correction	4.76884	1	.02898
Likelihood Ratio	5.00327	1	.02530
Mantel-Haenszel test for linear association	5.00124	1	.02533

Minimum Expected Frequency = 279.031

Number of Missing Observations: 12

[Chi-square printout]

the same subjects in two different groups, the statistical test called for is not ANOVA, but MANOVA, because the scores from the three different kinds of tests administered to the same subjects should somehow be related. These

Correlations:	FIN	MLAT 5	PLAB5	PLAB6	MLAT4	COMP
FIN						
MLAT5	.1735*					
PLAB5	.1233	.1346*				
PLAB6	-.0231	.0795	.2317**			
MLAT4	.0326	.2718**	.1071	.2111**		
COMP	.0033	.3876**	.1032	.0083	.2126**	
PT1	.1424*	.2582**	.1615*	-.0656	.1241	.4303**
PT2	.0656	.2191**	.0791	.0250	.1188	.4474**
PT3	.1100	.1199	.1047	-.0738	.1787**	.3083**

N of cases: 300 1-tailed Signif: * - .01 ** - .001

Correlations:	PT1	PT2	PT3
FIN			
MLAT5			
PLAB5			
PLAB6			
MLAT4			
COMP			
PT1			
PT2	.6597**		
PT3	.6137**	.6634**	

N of cases: 300 1-tailed Signif: * - .01 ** - .001

[SPSS/PC+ printout 3]

related variables should be analyzed together, not separately. (For further details, refer to Hatch & Lazaraton, 1991, pp. 386–387).

Frequency Comparisons

Chi-square: For comparing two nominal (frequency) data. Frequency means counting the number of times something happens. When examining relations between frequencies, the Chi-square analysis is the usual procedure employed. For example, suppose you conducted a survey asking people whether or not they think life is exciting. One thousand four hundred and sixty one persons

responded, and 50.3% of all the men (300 out of 596) and 44.4% of all the women (384 out of 865) described life as exciting. Is this evidence sufficient enough to believe that men differ from women in finding life exciting? To answer a question like this, the Chi-square test is calculated.

As shown in the computer printout, this procedure compares the observed frequencies and expected frequencies to see if the former are greater than chance alone. The observed significance level of a Chi-square of 5.005 is 0.025. This indicates that a discrepancy this large between the observed and expected frequencies would occur only 2.5% of the time if, in the population, men and women are equally excited with their lives. Since the observed significance level is quite small, we can conclude that men and women are not equally likely to find life exciting.

One important fact we should bear in mind here is that this family of analyses should be used to research where only nominal variables are included and frequencies are compared, and that, more important, the Chi-square procedure does not allow us to make cause–effect claims.

Comparisons of Relationship

Pearson correlation: For examining existing relationships between variables without any manipulation of the variables (e.g., no treatment). For example, if we wonder whether a good language aptitude is related to success in vocabulary learning and acquisition, we might want to know the relationship of the subjects' scores on aptitude tests as well as their scores on a vocabulary comprehension and retention test. Below is the computer printout of one such study (Nagata & Ellis, 1996, Forthcoming). FIN, MLAT4, PLAB5, PLAB6, and MLAT4 are subtests of aptitude batteries. COMP, PT1, PT2, and PT3 are vocabulary comprehension and retention tests. The asterisk (*) shows where significant relationships were detected.

Here again, as with the Chi-square analysis, we should be cautioned that the correlation tells us only that there is some degree of relationship between the two variables. We cannot, therefore, make any cause–effect claims.

Statistical Assumptions to Remember

Selecting an appropriate statistical test is a crucial point in any statistical study. Major questions we should ask ourselves are: (a) How many variables are there and what are their functions in the study (independent vs. dependent variables)?; (b) What types of measurement tools (scales) were used?, and; (c) Where do the data come from, i.e., from two different groups (a between-groups design), or two or more measures taken from the same group (a repeated-measures design)?

Even after an appropriate test has been selected, we should not forget the fact that all statistical tests have certain assumptions underlying their formulation. When

even one of these assumptions is violated, the probability gets distorted, and it becomes difficult to know how much confidence to place in the findings. Check to see if these assumptions are met. If you are the reader of a statistical study, see if the researcher checked whether the basic assumptions of the specific test(s) employed were met. (For a compact list of assumptions and solutions when these assumptions cannot be met, see Hatch & Lazaraton, 1991, pp. 546–554.)

By Way of Conclusion

I have taken you quickly through the winding mountain path of basic statistical manipulations often used in our field. Just to recapitulate some of the checkpoints which have so often been neglected in the statistical studies published in our field:

1. Check to see if the appropriate statistical tests are used.
2. Check to see if all the statistical assumptions are met.
3. See if the number of subjects is sufficient. (Rule of thumb says more than 30.)
4. Check to see if multiple comparisons are being made erroneously. Especially watch out for the use of the ordinary t-test for making multiple comparisons).
5. See if the nominal, frequency data are properly dealt with, in other words, by the Chi-square analyses for independent data.
6. Check to see if the researcher's interpretation of the findings stays within the statistical logic. For example, the Chi-square procedure is a non-parametric test. It does not allow us to make cause–effect claims. Neither does the Pearson correlation.
7. Check to see if the dependent variables in the study are related. If so, use MANOVA, instead of ANOVA.
8. See if the data are from two different groups or are two or more measures taken from the same group. In other words, be careful about the repeated-measures designs. Data, then, should be examined in terms of within-subjects differences rather than between-subjects differences. If, for instance, a pretest and a posttest are administered, the study is repeated-measures in design.

References

I would like to use this section not only as a conventional reference section, but an introductory resource section. Therefore, the books and articles are not in conventional alphabetical order. The following are books and articles which I have found accessible, understandable, and extremely readable. If you are new to statistics you might like to start reading in the order they are presented here.

1. Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.
2. Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. Especially, Chapter 9: Analyzing the Data (pp. 201–242). Oxford: Oxford University Press.
3. Brown, J. D. (1991). Statistics as a foreign language—Part 1: What to look for in reading statistical language studies. *TESOL Quarterly, 25* (4), 569–586.
4. Nunan, D. (1992). *Research methods in language learning*. Especially, Chapter 2: The Experimental Method (pp. 24–51). Cambridge: Cambridge University Press.
5. Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied Linguistic*. Rowley, MA: Newbury House.
6. Brown, J. D. (1992). Statistics as a foreign language—Part 2: More things to consider in reading statistical language studies. *TESOL Quarterly, 26* (4), 629–634.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

FL024999 – FL025011

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").