

THE UNIVERSITY OF TEXAS AT AUSTIN SCHOOL OF INFORMATION

MATHEMATICAL NOTES FOR LIS 397.1 INTRODUCTION TO RESEARCH IN LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys
Last revised: 2003 Jan 15

WHICH STANDARDIZED STATISTICAL PROCEDURE SHOULD I USE?

1 Introduction

In this note we provide a brief summary of how the various standardized statistical procedures can be used to test statistical hypotheses. We list the most commonly employed types of statistical hypotheses and outline which of the procedures should be used to test which of the different types of hypotheses. First, we review the basic types of statistical hypotheses that you are most likely to encounter or to need to employ, and how hypotheses concerning them can be formally stated.

2 Common Types of Statistical Hypotheses

2.1 Hypotheses about the Mean(s) of a Population or Populations

2.1.1. *The population mean is some specified number*, which we can represent by μ_0 .

$$H_0: \mu = \mu_0$$

Example: "The average daily circulation total is 123."

2.1.2. *The means of two different populations are equal*. If we call the populations 1 and 2, respectively, we can write the hypothesis as

$$H_0: \mu_1 = \mu_2$$

Example: "The average [mean] cost per online search using Service A is the same as the average [mean] cost per online search using Service B."

2.1.3. *The means of three or more different populations are equal*. If we call the populations 1, 2, 3, etc., we can write the hypothesis as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots \text{ etc.}$$

Example: "Among high-school students, the average [mean] number of library books borrowed per student each semester is the same for sophomores, juniors, and seniors."

2.2 Hypotheses about the (Pearson) Correlation between Two Variables

Variables X and Y are not correlated. It is assumed that variables X and Y are both Gaussianly distributed.

$$H_0: \rho_{XY} = 0$$

Example: "There is no correlation between the age and the salary of a typical librarian."

2.3 Hypotheses about the Association between Two Variables

Categorical variables X and Y are not associated.

H₀: Variables X and Y are not associated.

Example: "There is no association between the sex of a library patron and the type of book the patron prefers."

3 Which Standardized Statistical Procedure Should I Use for Which Situation?

In the following discussion, $Prob(t_{obs})$ represents the **probability of getting, when the null hypothesis is true, a value of t that is equal to or greater than the value actually observed.** This kind of probability is calculated by many computerized statistical procedures, and is typically reported by the program with a label like *Prob* or *P*.

3.1 Procedures for Hypotheses about the Mean(s) of a Population or Populations

In dealing with a hypothesis concerning the mean(s) of a population or populations, you should recall that whenever you take a sample of observations from some population, you will be able to determine the numerical values of the sample mean, \bar{x} , the sample standard deviation, s , and the sample size, n .

3.1.1. *The population mean is some specified number, which we can represent by μ_0 .*

H₀: $\mu = \mu_0$

The pertinent procedure is the t-test. The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

The observed value of the test statistic is compared against a threshold value from (a) a table of the Student's t-distribution if $n \leq 30$, or (b) if $n \geq 31$, it is permissible to use a table of the Gaussian distribution (although it is always preferable to use a table of the t-distribution if one is available for the sample size being used). The relevant number of degrees of freedom is $n - 1$. The null hypothesis is accepted if the absolute value of t_{obs} is equal to or less than the threshold; it is rejected if the absolute value of t_{obs} exceeds the threshold. Alternatively, the null hypothesis is accepted if $Prob(t_{obs}) \geq \alpha$; it is rejected if $Prob(t_{obs}) < \alpha$.

3.1.2. *The means of two different populations are equal.*

H₀: $\mu_1 = \mu_2$

The pertinent procedure depends on what kind of sampling process is possible. If it is possible to take a pair of dependent samples, then the procedure to be used is the t-test for dependent samples. If the samples are independent, then the procedure to be used is either the t-test for independent samples or the ANOVA procedure (with independent samples).

3.1.2.1 The t-Test for Dependent Samples

In the case of the t-test for dependent samples, we work with the pairwise differences in the values observed for each sample pair. If the values for the i -th pair in the sample are represented by x_i and y_i , then the corresponding pairwise difference can be represented by $d_i = x_i - y_i$; the mean of the pairwise differences, by \bar{d} ; and the standard deviation of the pairwise differences, by s_d . The test statistic is

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

The observed value of the test statistic is compared against a threshold value from (a) a table of the Student's t-distribution if $n \leq 30$, or (b) if $n \geq 31$, it is permissible to use a table of the Gaussian distribution (although it is always preferable to use a table of the t-distribution if one is available for the sample size being used). The relevant number of degrees of freedom is $n - 1$. The null hypothesis is accepted if the absolute value of t_{obs} is equal to or less than the threshold; it is rejected if the absolute value of t_{obs} exceeds the threshold. Alternatively, the null hypothesis is accepted if $Prob(t_{obs}) \geq \alpha$; it is rejected if $Prob(t_{obs}) < \alpha$.

3.1.2.2 The t-Test for Independent Samples

In the case of the t-test for independent samples, we work with the two samples, x_{1i} and x_{2i} . We begin by getting the two sample standard deviations, s_1 and s_2 , and calculating what is called the pooled estimate of population variance, s_p^2 , as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 - 1 + n_2 - 1}$$

Next, we calculate the standard error of the difference of the means,

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Finally, we form the test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

The observed value of the test statistic is compared against a threshold value from (a) a table of the Student's t-distribution if $n_1 + n_2 \leq 31$, or (b) if $n_1 + n_2 \geq 32$, it is permissible to use a table of the Gaussian distribution (although it is always preferable to use a table of the t-distribution if one is available for the sample size being used). The relevant number of degrees of freedom is $n_1 + n_2 - 2$. The null hypothesis is accepted if the absolute value of t_{obs} is equal to or less than the threshold; it is rejected if the absolute value of t_{obs} exceeds the threshold. Alternatively, the null hypothesis is accepted if $Prob(t_{obs}) \geq \alpha$; it is rejected if $Prob(t_{obs}) < \alpha$.

3.1.2.3 The ANOVA Procedure with Independent Samples

In using the ANOVA procedure for testing a hypothesis that the means of two (or more) populations are equal, you will enter the observations into a computer program that runs the ANOVA procedure. The program will respond with a report as to what it finds for the value of the test statistic,

$$F_{obs} = \frac{\hat{\sigma}_{bg}}{\hat{\sigma}_{wg}}$$

You can compare this observed value of the test statistics with the threshold from a table of the F-distribution, using the appropriate values for the within-group degrees of freedom and the between-group degrees of freedom. As usual, if the observed value of the test statistic exceeds the tabled threshold value, you reject the null hypothesis; otherwise, you accept the null hypothesis. However, nowadays the easier way of making the decision is to examine the probability that the computer program reports. If this probability is less than the level of significance at which you chose to conduct your test of the null hypothesis, you reject the hypothesis; otherwise, you accept it.

3.1.3. *The means of three or more different populations are equal.* If we call the populations 1, 2, 3, etc., we can write the hypothesis as

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots \text{ etc.}$$

The way to test such a hypothesis is to use the ANOVA procedure, as outlined above in paragraph 3.1.2.3.

3.2 Procedures for a Hypothesis about the (Pearson) Correlation Coefficient

Variables X and Y are not correlated. It is assumed that variables X and Y are both Gaussianly distributed.

$$H_0: \rho_{XY} = 0$$

The test statistic is

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(n-1)s_X s_Y}} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{(n-1)s_X s_Y}}$$

The observed value of the test statistic is to be compared with the threshold value found in a table of the distribution of Pearson's correlation coefficient, with degrees of freedom = sample size - 2. As usual, if the observed value of the test statistic exceeds the tabled threshold value, you reject the null hypothesis; otherwise, you accept the null hypothesis. Unfortunately, most computer programs that calculate the correlation coefficient do not report the probability of getting a sample value equal to or larger than the observed value in the situation when the null hypothesis of zero correlation is true. If you use a program that does report this probability, then if this probability is less than the level of significance at which you chose to conduct your test of the null hypothesis, you reject the hypothesis; otherwise, you accept it.

3.3 Procedures for a Hypothesis about the Association between Two Variables

Categorical variables X and Y are not associated.

$$H_0: \text{Variables X and Y are not associated.}$$

The test statistic is

$$\chi^2_{obs} = \sum \frac{O_i - E_i}{E_i}^2 = \sum \frac{O_i^2}{E_i} - n$$

where O_i and E_i represent, respectively, the observed and the expected number of occurrences in each of the categories in the cross-tabulation table of the categories being studied, and n represents the total size of the sample (scil., the sum of the O_i s). The observed value of the test statistic is compared with the threshold value taken from a table of the chi-square distribution, with $df = (\text{number of rows} - 1) + (\text{number of columns} - 1)$.