

# 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

## QUESTIONS ABOUT POPULATION MEANS AND PROPORTIONS

This discussion concerns how questions about population means are handled with the aid of the Student's t distribution and how questions concerning population proportions are answered.

### USES OF THE t-TEST

There are two main uses of the t-test in inferential statistics.

**Use I:** To test the null hypothesis that the mean of a population, denoted here by  $\mu$ , has a particular numerical value, which we shall represent by  $\mu_0$ . This hypothesis can be written symbolically as

$$H_0: \mu = \mu_0$$

A common occasion for this use is when you, the investigator, want to see if the previously known average value of some variable has stayed the same following a change in the situation in which the variable occurs. For example, if Book-Dealer B has just installed a computer and claims that its average response time to orders placed by your acquisitions department is now shorter than it used to be, you might want to test B's claim against what you know B's average response time has been in the past.

**Use II:** To test the null hypothesis that two populations, which we can call "Pop<sub>1</sub>" and "Pop<sub>2</sub>", have equal means (or equivalently, that they have the same mean). Using  $\mu_1$  and  $\mu_2$  to represent the means, we can express this hypothesis symbolically as

$$H_0: \mu_1 = \mu_2$$

### Independent vs. Dependent Samples

We must distinguish two different circumstances in which you can use the t-test to see whether two populations have the same mean. The first is when you get your information about the two populations through *two* samples that are drawn *independently*, one from each of the two populations. The second circumstance is when you get your information about the two populations through *one* sample, by making *two observations* (for example, under different conditions or at different times) *on each element* in the sample.

In the first circumstance we speak of using "the t-test for independent samples"; in the second, of using "the t-test for dependent samples," since the two measurements on each sample element obviously bear a relation to each other (and hence are not independent of each other).

Among the occasions for using the t-test for comparing the means of two populations is the type of circumstance in which you have an overall (a superordinate) population that can be subdivided into two sub-populations, and in which you want to compare the average values that some variable takes on in the two sub-populations.

For example, the population that consists of numbers of books borrowed from a college library by individual history majors could be subdivided into two sub-populations: numbers of books borrowed by female history majors, and numbers of books borrowed by male history majors. Assuming your circulation department has kept the requisite records, you could take a random sample of female history majors and find out how many books each of them borrowed last semester; you could do the same for male history majors; and thus you could find out whether female history majors borrowed more books on the average than male history majors.

Note that in this example (except for certain extremely rare exceptions) you would not be able to make a pair of observations of, say, first, the numbers of books borrowed by a student while he was a male and, second, the number of books borrowed by the same student while she was a female. Aside from such exceptions, your samples would have to be independent.

With other examples, however, it can be possible for the elements of a given sample to belong to two different sub-populations. For example, suppose you wanted to study the effects on book-borrowing of a "how to use the library" workshop that all students were required to take sometime during their freshman year. You would have a choice of methods. One method would be to draw two independent samples, one from the population of freshmen who had already taken the "how to use the library" workshop, and the other from the population of freshmen who had not yet done so. (Taking such samples early in the spring semester should yield a reasonably good comparison of the book-borrowing practices of the two groups of freshmen.) The other method would be to draw a single sample of freshmen who took the library-use workshop at the start of the spring semester, and to compare each student's borrowing of books during the fall semester with that same student's book-borrowing during the spring semester.

The point of the second method is that each student belongs, at different times, to each of the two sub-populations. Thus the two sub-populations can be compared by means of two observations made on each sample element, one observation when the element is in one of the sub-populations, and the other observation when the element is in the other sub-population. When the nature of the sub-populations you are investigating permits elements of the superordinate population to *change* from one of the sub-populations to the other, it may be possible to make paired observations, as in this second method.

If it is possible to make paired observations, it will usually be desirable to do so. The reason is that a test employing pairs of observations on  $n$  elements in a sample has a higher power (i.e., a better chance, when the null hypothesis is false, of correctly detecting that fact) than does a test of the same populations using two independent samples of  $n$  elements each.

In short, tests of a hypothesis that two populations have the same, or equal, means may, depending on the circumstances, be handled by two independent samples, or by one (dependent) sample that provides pairs of observations. Taking into account also tests of the first kind of hypothesis mentioned in this handout (viz., that a population has a particular numerical value), we have three distinct situations whose details need to be discussed.

### **The Essence of Making a Test of a Statistical Hypothesis**

The procedures for handling the three situations share, with other procedures for testing statistical hypotheses, the following features: (a) the pertinent test statistic is calculated from the evidence available in the sample or samples, and (b) this observed value of the test statistic is compared with a threshold value for the statistic obtained from a table. The tabled threshold value always depends on the significance level at which you want to test the hypothesis; often, it also depends on the number of degrees of freedom, denoted by "df", associated with the test procedure and with the sample size; and, further, it can also depend on whether you are making a *one-tailed* or a *two-tailed* test. In LIS 397.1 we discuss only two-tailed tests.

The basic criterion for reaching a decision whether to accept or reject a null hypothesis is the following: If the magnitude (i.e., the size with the minus sign, if any, replaced by a plus sign) of the observed value of the test statistic *exceeds* the pertinent tabled threshold value, you can *reject* the null hypothesis; otherwise, you cannot reject it. Non-rejection of a null hypothesis is usually referred to as *accepting* the null hypothesis (although purist prefer to call it *failing to reject* the null hypothesis).

Many computer programs that do inferential statistics now provide an alternative way of reaching a decision on a statistical hypothesis. These programs not only report the observed value of the test statistic, but also calculate the probability that this observed value will occur, by chance alone, when the null hypothesis is true. A large value for this prob-

ability means that what was observed is the kind of thing that is *likely* to occur when the null hypothesis is true; hence, its occurrence suggests strongly that the null hypothesis is true. On the other hand, a small value for this probability means that what was observed is something that is quite *unlikely* to occur when the null hypothesis is true; hence, its occurrence suggests strongly that the null hypothesis is false.

In particular, if this probability (viz., the probability of observing the observed value of the test statistic by chance alone when the null hypothesis is true) is small enough to be less than what you chose as the level of significance,  $\alpha$ , of the test procedure, then the decision rule is: Reject  $H_0$ . If, however, this probability is equal to or greater than the level of significance, then the decision rule is: Accept  $H_0$ .

### Three Situations for Using the t-Test

Here is how these general considerations apply to the three situations in which the t-test is used.

**Situation A.**  $H_0: \mu = \mu_0$

This hypothesis asserts that the population mean,  $\mu$ , is equal to a particular number, denoted by  $\mu_0$ .

#### Procedure for Situation A

The evidence available from the sample is the sample mean,  $\bar{X}$ , the sample standard deviation,  $s$ , and the sample size,  $n$ . The standard error of the mean, is calculated from

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

and the test statistic is

$$t_{obs} = \frac{(\bar{X} - \mu_0)}{s_{\bar{x}}} = \frac{(\bar{X} - \mu_0)}{\frac{s}{\sqrt{n}}}$$

When you have calculated the observed value of the test statistic, you compare its (absolute) value with the tabled threshold value of Student's t that is determined by the level of significance you have chosen and by the number of degrees of freedom. In this situation,  $df = n - 1$ .

(Note: You may notice that the numerator of the above test statistic shows the signed distance [i.e., the distance with a plus or minus sign attached]. In LIS 397.1 you are going to be working only with the absolute value of  $t_{obs}$ , and thus you can ignore the sign for the purposes of the discussion here.)

When you calculate the test statistic, what you are doing can be expressed as follows. The numerator in the above expression,  $\bar{X} - \mu_0$ , is simply the distance between the observed sample mean  $\bar{X}$  and the particular number  $\mu_0$ . That is, you are studying how far apart the sample mean and the particular number are.

Specifically, you are interested in how far apart the sample mean and the particular number are when that distance is measured in units of the standard error. Let us examine the reasoning, using initially a sample size large enough so that we may use, in our discussion, the Gaussian factors of 1.96 (corresponding to a 95% confidence interval or a 5% level of significance) and 2.58 (corresponding to a 99% confidence interval or a 1% level of significance). There are two possible cases.

**Case 1.** If the observed sample mean  $\bar{X}$  and the particular number  $\mu_0$  are no more than 1.96 standard-error units apart, then you can conclude that the *observed* sample mean appears to be one of the 95% of *all possible* sample means that would lie within 1.96 standard-error units on either side of the hypothesized population mean  $\mu_0$ .

Hence, you can conclude that the evidence from the sample is consistent with the idea that the population mean  $\mu$  is equal to  $\mu_0$ . Therefore, your decision will be to accept the null hypothesis. (If you are working at a 1% level of significance, you would use a factor 2.58 in the above reasoning instead of 1.96.)

**Case 2.** If the observed sample mean  $\bar{X}$  and the particular number  $\mu_0$  are more than 1.96 standard-error units apart, then you need to reason as follows:

The null hypothesis is either true or false. If the null hypothesis is *true*, then the sample mean that you have observed must be one of the 5% of all sample means that lie outside the central interval around the population mean. (Remember that the central interval contains 95% of the sample means.)

It is *possible* that the null hypothesis is true and that you have observed a rare, atypical sample, but it is *unlikely* that that is what happened. What is *likely* is that the null hypothesis is not true and that the sample that you observed is a quite typical sample drawn from a population for which the null hypothesis is false.

Thus you should conclude that the evidence from the sample is not consistent with the idea that the population mean  $\mu$  is equal to  $\mu_0$ . Therefore, your decision will be to reject the null hypothesis. (Again, you would use 2.58 instead of 1.96 if you are working with a 1% level of significance.)

What we have just said holds if the sample size is large enough (i.e., if the sample size is 31 or more, according to the usual rule of thumb). If the sample size is 30 or under, then instead of the Gaussian factors of 1.96 or 2.58, you must use a t-table factor that will depend on the level of significance you are using and the number of degrees of freedom for your sample. (As usual, however, the best advice is to use the t-table factors whenever possible, even for sample sizes above 30.)

**An Example of Situation A.** Suppose you are the librarian of a small private college. You have been buying books through a particular wholesaler, Vendor A, with whose performance you have been reasonably, but not entirely, satisfied. Today you are approached by a sales representative from another wholesaler, Vendor B, who would like you to switch over to using her company. The Vendor B salesrep claims that her company can give you faster delivery than your present vendor, i.e., that her company can average a shorter delivery time than the 38 days (as measured from the day you send the company an order) that you know is Vendor A's mean delivery time.

Naturally, you are interested in the possibility of improving the delivery time on your book orders, since that possibility, if a reality, would ultimately provide better service to your patrons, to whom new or newly ordered books would become available sooner. But, also naturally, you are skeptical, having had experience in the past with overly encouraging promises by salespersons. You conclude that you should make a test of Vendor B's delivery performance.

The most natural way to make such a test might seem to be to place some book orders with Vendor B and observe the delivery times. This method is open to the objection that Vendor B would be likely to spot such orders as coming from a prospective customer and give them special attention, which would defeat the purpose of the test. Fortunately, you happen to be aware that Enormous State University, located in the same city as your college, has been using Vendor B for some time for many of its acquisitions. You enlist the cooperation of the director of the ESU Library and make the following test.

You make a random selection of 60 books that you ordered between 12 and 6 months ago from Vendor A. Checking with the ESU Library, you find that all but 11 of the 60 books were ordered by ESU from Vendor B, and, upon further examination, you find yourself feeling reasonably sure that these 11 books are not different from the others in any way that is likely to affect the delivery time. For the 49 books that ESU ordered from Vendor B, a check of the ESU Library's records shows a mean delivery time of 36 days, with a sample standard deviation of 8.

The fundamental question is whether, on the average, Vendor B does a faster job than Vendor A of getting books to a customer. At this point, you can deal with this question only in terms of whether, on the basis of the information in this sample, Vendor B appears to be doing a faster job.

**Two Approaches to Answering the Question.** There are two logically equivalent approaches to using the information in the sample to find out whether Vendor B appears to deliver books faster than Vendor A.

One approach is to use the information in the sample to construct a confidence interval for the population mean,  $\mu$ , the average time that Vendor B takes to deliver a book. That is, we can use the sample information to provide an interval within which we can be reasonably sure that the mean delivery time of Vendor B lies. If it turns out that the confidence interval for  $\mu$  includes the known mean delivery time of Vendor A (viz., 38 days), then we have to conclude that Vendor B's mean delivery time is not really distinguishable from Vendor A's mean delivery time; in other words, we have to conclude that so far as we can tell from the sample, the two vendors deliver books equally fast (or slowly) on the average. If, on the other hand, it turns out that the confidence interval does not include the known mean delivery time of Vendor A, then we can conclude that the two vendors have definitely different mean delivery times.

The other way of using the information in the sample is to test the null hypothesis that the mean delivery time for Vendor B equals the known 38-day mean delivery time for Vendor A (in symbols,  $H_0: \mu_0 = \mu_0 = 38$ ). If the sample information leads us to accept this null hypothesis, then we have to conclude that the information in the sample is consistent with the idea that Vendor B's mean delivery time is the same as Vendor A's; in other words, we have to conclude that so far as we can tell from the sample, the two vendors deliver books equally fast (or slowly) on the average. If, on the other hand, it turns out that the information in the sample leads us to reject the null hypothesis, then we can conclude that the two vendors have definitely different mean delivery times.

**The Confidence-Interval Approach.** Here is how you can work things out the confidence-interval way. From your sample, your data on Vendor B's performance are these:  $\bar{X} = 36, s = 8, n = 49$ . Using them in the usual formula for a confidence interval, and choosing to form a 95% confidence interval, we know that

$$\mu \text{ is in the interval, } \bar{X} \pm C \left( \frac{s}{\sqrt{n}} \right), \text{ with 95\% confidence.}$$

I.e., using the sample data, we can say with 95% confidence that  $\mu$  is in the interval

$$36 \pm 1.96 \left( \frac{8}{\sqrt{49}} \right) = 36 \pm 1.96 \left( \frac{8}{7} \right) = 36 \pm 2.24 = (33.76, 38.25)$$

Thus the 95% confidence interval for Vendor B's mean delivery time,  $\mu$ , includes the value, 38, that is known to be Vendor A's mean delivery time. Thus, on the basis of the observed behavior of Vendor B, we have to conclude that Vendor B's mean delivery time is not distinguishably different from that of Vendor A.

**The Hypothesis-Testing Approach.** Here is how you can work things out the hypothesis-testing way. From your sample, your data on Vendor B's performance are these:  $\bar{X} = 36, s = 8, n = 49$ . You use them to test the null hypothesis,  $H_0: \mu_0 = 38$ . You compute the test statistic,

$$t_{obs} = \frac{(\bar{X} - \mu_0)}{s_{\bar{X}}} = \frac{(\bar{X} - \mu_0)}{\frac{s}{\sqrt{n}}} = \frac{(36 - 38)}{\frac{8}{\sqrt{49}}} = -\frac{2}{\frac{8}{7}} = -\frac{2}{1.1428} = -1.75$$

You next compare the absolute value of the test statistic,  $|t_{obs}| = 1.75$ , with the tabled threshold value, at a 5% level of significance, for  $t$  with 48 df. From a table of the  $t$  distribution, you find an entry for the 5% level of significance with 50 df, viz., 2.0086, and you realize that the value of  $t$  for 48 df will be approximately the same. Since  $t_{obs}$  is considerably smaller, you can conclude immediately that your decision must be to accept  $H_0$ . (An exact value of the 5%-level-of-significance threshold for  $t$  with 48 df can be calculated by interpolation between the values in the table for 40 df and 50 df [it works out to be 2.0111], but that is not necessary here since 1.75 is clearly much less than the threshold.) Since the sample size exceeds 30, you could also simply use the Gaussian threshold, 1.96, in the comparison; again, 1.75 is clearly less than the threshold.

Whichever way you obtain an appropriate threshold value, you find that the comparison of the observed value of the test statistic,  $t$ , with the threshold value leads to the decision to accept the null hypothesis,  $H_0: \mu_0 = 38$ . In other words, the data in your sample of Vendor B's performance lead you to conclude that Vendor B's mean delivery time is not distinguishably different from that of Vendor A.

**Comparing the Confidence-Interval and Hypothesis-Testing Approaches.** What these two ways of working out the meaning of your sample data illustrate is this: When you have a null hypothesis of the type  $H_0: \mu = \mu_0$ , i.e., when you are in Situation A, the hypothesis-testing approach and the confidence-interval-construction approach are logically equivalent. If one of them leads to a decision to accept the null hypothesis, so will the other approach; if one of them leads to a decision to reject the null hypothesis, so will the other approach.

**When Can Situation A Occur in Practice?** You may be wondering when you might encounter Situation A in practice. One possible way of encountering it has already been suggested by the preceding example: a situation in which a vendor asserts that a certain cost has some average amount, and you are interested in seeing whether that average cost is true in your particular circumstances. Other easily encountered situations include: figures from national statistics quoted by professional groups, assertions in articles published in the professional literature, and historical records from earlier times in your own institution (or elsewhere) when only summary data were recorded.

**Situation B.**  $H_0: \mu_1 = \mu_2$  with independent samples.

This hypothesis asserts that the means of two different populations are equal to each other. The populations are to be sampled independently of each other. The evidence available in the sample drawn from the first population,  $\text{Pop}_1$ , is the sample mean,  $\bar{X}_1$ , the sample standard deviation,  $s_1$ , and the sample size,  $n_1$ . For the sample drawn from  $\text{Pop}_2$  the analogous data are  $\bar{X}_2$ ,  $s_2$ , and  $n_2$ .

In this case the test statistic is the following:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

As you can see the numerator of this expression consists of the difference of the two sample means; i.e., the numerator shows the distance between the sample means. The denominator of the expression contains a new concept,  $s_{\bar{X}_1 - \bar{X}_2}$ , called the *standard error of the difference of means*.

(Note: As with Situation A earlier, you may notice that the numerator shows the signed distance [i.e., the distance with a plus or minus sign attached], but since you are going to be working with the absolute value of  $t_{obs}$ , you can again ignore the sign for the purposes of the discussion here.)

One way to think of the standard error of the difference of means is this: It is approximately the average distance between the sample means that you would find if you drew many pairs of samples from two populations whose means are equal. More precisely, the standard error of the difference of means has the following property:

If we drew all possible pairs of samples from a population and for each pair we computed the distance between their sample means, then we would find that there is a Gaussian distribution of these distances. This distribution would be centered around a mean difference of zero, and the standard deviation of the distribution would be the standard error of the difference of means. That is, there would be a central interval around zero within which 95% of all the differences would lie, and this interval would extend from 1.96 units of the standard error of the difference of means below zero to 1.96 such units above zero. A similar interval extending from 2.58 such standard-error units below zero to 2.58 standard-error units above zero would contain 99% of all the differences in sample means.

We use the standard error of the difference of means in much the same way as we used the ordinary standard error in Situation A. That is, we use the standard error of the difference of means as the reference unit of measure by which we judge the separation of the observed sample means. As with Situation A, there are two possible cases.

**Case 1.** If the observed sample means are no more than 1.96 units apart, as measured in units of the standard error of the difference of means, then you can conclude that the *observed pair* of sample means appears to be one of the 95% of *all possible such pairs* that would lie within 1.96 standard-error units on either side of the hypothesized mean difference of zero (hypothesized as zero because your hypothesis is that the two population means are equal, which equates to a hypothesis that the difference of the population means will be zero). Hence, you can conclude that the evidence from the sample is consistent with the idea that the population means are equal. Therefore, your decision will be to accept the null hypothesis. (If you are working at a 1% level of significance, you would use a factor 2.58 in the above reasoning instead of 1.96.)

**Case 2.** If the observed sample means are more than 1.96 standard-error units apart, you have to conclude that the evidence from the sample is not consistent with the idea that the difference between the two population means is zero. Therefore, your decision will be to reject the null hypothesis. (Again, you would use 2.58 instead of 1.96 if you are working with a 1% level of significance.)

The standard error of the difference of means is calculated in either of two ways. The simpler way is to set

This called the "separate estimate of variance" method.

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

A slightly more complicated, but generally preferable, way is called the *pooled estimate of variance* method and involves one extra step. In this step you calculate  $s_p^2$ , the pooled estimate of variance:

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

Here you have pooled the information in the two separate sample variances into a single number--their weighted average--, with weights based on the size of each sample, or more precisely, on the number of degrees of freedom associated with each sample variance.

Once you have found  $s_p^2$ , you calculate

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

In this form you can see that  $s_p^2$  is being substituted for the two original sample variances. An easier and more direct way to perform the calculation is

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

When you have calculated the standard error of the difference of means, by either the separate variance or the pooled variance method, you next work out the value of the test statistic:

$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

and compare its (absolute) value with the tabled threshold determined by the level of significance you have chosen and by the number of degrees of freedom.

In this case, the number of degrees of freedom must take into account the fact that there are two samples involved, one from each population. For this case, we have  $df = n_1 + n_2 - 2$ , since  $n_1 + n_2 - 2 = (n_1 - 1) + (n_2 - 1)$ . Thus, the overall df is just the sum of the dfs of the two samples.

**An Example of Situation B.** Suppose you are the head of public services in a public library and are concerned over ways of making your patrons feel easier about asking questions of the reference librarians and more willing to do so. As you ponder this problem, it occurs to you that one factor may lie in the answer to the following question: Is the reference librarian on duty usually sitting behind the massive antique desk that someone gave to the library long ago (in order to obtain an income-tax deduction), or is the reference librarian usually not sitting behind that desk, so that he or she thus appears more accessible.

As you think further on this matter, you realize that you could test a hypothesis dealing with the question of whether patrons seem to be less willing to approach the reference librarian on duty when he or she is seated behind the massive antique than when the situation is different. You decide that you will make such a test, and that your criterion variable will be "numbers of questions (other than mere direction-seeking questions) asked per day" under the two conditions, "A: reference librarian seated behind massive antique desk set parallel to the wall" and "B: reference librarian seated at small modern desk set perpendicular to the wall." Specifically, you will test the null hypothesis,  $H_0: \mu_1 = \mu_2$ , which says that the mean numbers of questions asked per day under the two conditions will be equal.

Enlisting the aid of the head janitor, you arrange that on randomly selected days over the next few weeks you will alternate between conditions A and B. You collect the following data, i.e., numbers of questions asked under the two conditions:

A: 19, 10, 10, 17, 23, 11, 17, 18, 21, 24, 9, 12, 13, 12, 17, 15, 22, 18, 16, 10, 14, 19, 20, 11, 8

B: 18, 20, 17, 14, 22, 23, 14, 21, 16, 19, 15, 23, 18, 22, 13, 21, 12, 13, 19, 22

From these data you obtain the following sample values:  $\bar{X}_A = 15.4400$ ,  $s_A = 4.6911$ ,  $s_A^2 = 22.0067$ ,  $n_A = 25$ ;  $\bar{X}_B = 18.1000$ ,  $s_B = 3.6548$ ,  $s_B^2 = 13.3579$ ,  $n_B = 20$ . Working with these data and using the generally preferred pooled-estimate-of-variance method, you obtain

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{(24)(22.0067) + (19)(13.3579)}{25 + 20 - 2} = 18.1851$$

Taking the square root of the left and right sides you find  $s_p = 4.2644$ . Then

$$\begin{aligned} s_{\bar{X}_A - \bar{X}_B} &= s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 4.2644 \sqrt{\frac{1}{25} + \frac{1}{20}} = 4.2644 \sqrt{.04 + .05} \\ &= 4.2644 \sqrt{.09} = 4.2644 (.3) = 1.2793 \end{aligned}$$

Next you compute the test statistic,

$$t_{obs} = \frac{\bar{X}_A - \bar{X}_B}{s_{\bar{X}_A - \bar{X}_B}} = \frac{15.4400 - 18.1000}{1.2793} = \frac{-2.6680}{1.2793} = -2.0793$$

Your next step is to compare this observed value of the test statistic (or, more exactly, its magnitude, i.e., its value ignoring the negative sign) with the tabled threshold value for  $t$  at the 5% level of significance with 43 df. From a table of the  $t$  distribution, you can find entries for 40 df, 2.021, and for 60 df, 2.000. Clearly,  $|t_{obs}|$  is larger than either of these and hence will be larger than the threshold for 43 df, which must lie between them. (An exact value for the 43-df threshold can be worked out by interpolation; it is 2.017.)

Since in magnitude  $|t_{obs}|$  exceeds the appropriate tabled threshold value, you reject the null hypothesis of equality, and you conclude that the mean numbers of questions asked per day are different under the two conditions, A and B. Specifically, you conclude that, on the average, more reference questions are asked when the reference librarian is sitting at the small modern desk, placed at a right angle to the wall, than when the librarian is sitting behind the massive antique desk, placed parallel to the wall.

**Situation C.**  $H_0: \mu_1 = \mu_2$  with one sample consisting of pairs of observations.

This hypothesis asserts that the means of two different populations are equal to each other. Unlike Situation B, however, the populations are to be sampled in a dependent manner.

Here the key pieces of evidence are the *differences* between the first and second observations on each element in the sample. For the  $i$ -th sample element, we can write these observations as  $X_{i1}$  and  $X_{i2}$ , respectively, and their difference as  $D_i = X_{i1} - X_{i2}$ . Thus a sample consisting of  $n$  pairs of observations will yield a set of  $n$  observed differences:  $D_1, D_2, \dots, D_n$ . This set of differences, like any other set of numbers, has a mean and a standard deviation, which we shall represent by  $\bar{D}$  and  $s$ , respectively; and it has a *standard error of the mean sample difference*,  $s_{\bar{D}}$ , given by  $s_{\bar{D}} = s / \sqrt{n}$ .

The null hypothesis that the means of the populations are equal can be rephrased as, "The mean difference,  $\Delta$ , between the populations is zero":

$$H_0: \Delta = 0$$

This hypothesis should remind you of the null hypothesis of Situation A, and the procedure for the test is very similar to that of Situation A.

The test statistic is

$$t_{obs} = \frac{(\bar{D} - 0)}{s_{\bar{D}}} = \frac{\bar{D}}{s_{\bar{D}}} = \frac{\bar{D}}{s / \sqrt{n}}$$

The (absolute) value of the observed test statistic is compared with the tabled threshold value determined by the level of significance you have chosen and by the number of degrees of freedom. Here,  $df = n - 1$ .

**An Example of Situation C.** Suppose that you are director of an academic library and that you have become curious about the question of whether the dress style (formal vs. casual) of the reference librarians seems to affect the willingness of students in your library to make inquiries of them. In this case you have a situation in which it will be appropriate and easy to make pairs of observations, by observing the numbers of inquiries put to each reference librarian when he or she is dressed formally and when dressed casually.

Enlisting the cooperation of the 10 people who, at various times, staff the reference desk, you make a random selection, for each person, of 5 days in the coming month when that person will be staffing the reference desk and will be dressed formally, and of another 5 days of reference duty when he or she will be dressed casually. On each of the selected days the person is to keep count of the number of (non-directional) inquiries he or she receives. You will be testing the null hypothesis,  $H_0: \Delta = 0$ , meaning that there is, on the average, a difference of zero (i.e., no difference) between the numbers of inquiries asked of formally dressed reference librarians and of casually dressed reference librarians.

From the experiment you obtain the following data for the total numbers of inquiries on the 5 days that each of the 10 librarians is dressed each way.

<u>Librarian</u>	<u>Casual Dress</u>	<u>Formal Dress</u>	<u>Difference</u>
Adams	91	95	-4
Burns	102	87	15
Clark	113	101	12
Dunn	83	89	-6
Evans	64	53	11
Frome	77	67	10
Gann	130	124	6
Hall	82	73	9
Innis	68	72	-4
Jones	93	81	12

The differences in the rightmost column are the key elements of your observations. From this sample of differences you obtain  $\bar{D} = 6.1000$ ,  $s = 7.7953$ , and  $n = 10$  as the three basic sample values. You form the test statistic

$$t_{obs} = \frac{\bar{D}}{s_{\bar{D}}} = \frac{\bar{D}}{s / \sqrt{n}} = \frac{6.1000}{7.7923 / \sqrt{10}} = \frac{6.1000}{7.7953 / 3.1623} = \frac{6.1000}{2.4651} = 2.4745$$

Since the observed value of the test statistic exceeds the tabled threshold value of  $t$  at the 5% level of significance with 9 df, viz., 2.2622, you reject the null hypothesis that there is no difference between mean numbers of inquiries asked of formally dressed librarians and of casually dressed librarians. You conclude that students tend to ask more questions when the reference librarians are dressed casually.

**Comparing the Dependent- and Independent-Sample Tests.** It is worth noting that if you had obtained the data you used in the foregoing example from two *independent* samples, you would have had different results.

Suppose that that had been the case; i.e., suppose that you had gathered the same observations of numbers of inquiries as above, but from two *independent* samples of 10 staff members each (i.e., from 20 staff members in all). In this case you would need to work with the data in the same way as in the example of a situation with independent samples.

Using C to indicate the casual-dress data and F the formal-dress data, you would obtain the following sample values:

$$\bar{X}_C = 90.3000, s_C^2 = 413.7889, n_C = 10; \bar{X}_F = 90.3000, s_F^2 = 413.7889, n_F = 10.$$

From these data you obtain

$$s_p^2 = \frac{10(413.7889) + 10(413.7889)}{10 + 10 - 2} = \frac{7291.7000}{18} = 405.0944; \text{ hence } s_p = 20.1270$$

Using this value for  $s_p$ , we find

$$s_{\bar{X}_C - \bar{X}_F} = 20.1270 \sqrt{\frac{1}{10} + \frac{1}{10}} = 20.1270 \sqrt{0.2} = 20.1270 (0.4472) = 9.0011$$

and hence

$$t_{obs} = \frac{\bar{X}_C - \bar{X}_F}{s_{\bar{X}_C - \bar{X}_F}} = \frac{90.3000 - 84.2000}{9.0011} = \frac{6.1000}{9.0011} = 0.6777$$

This observed value, 0.6777, of the test statistic is considerably less than the tabled threshold value of  $t$  at the 5% level of significance with 18 df, viz., 2.1009. Hence, you conclude that you must accept the null hypothesis that the mean numbers of inquiries are the same for casually and for formally dressed reference librarians.

The reason for the difference between this test result and the previous one for the dependent-sample situation lies in the obscuration caused by the big differences from one reference librarian to another in the numbers of inquiries. One could conjecture whether these differences arise from the different hours on duty of the various librarians (e.g., daytimes vs. evenings), or from differences in other aspects of the personal appearance of the librarians, or some other cause. Regardless of whether such conjectures are correct, the inter-personal differences in the numbers of inquiries are so great that they thoroughly obscure any differences in the numbers of inquiries that may be due to differences in dress style. In the dependent-sample case, on the other hand, the inter-personal differences do not matter, since all that is involved in the test is the set of differences, person by person, between each person dressed casually and that same person dressed formally.

### Large Samples and the z-Table

We have seen how to calculate, in various situations, a test statistic called the "Student's  $t$  statistic," whose essence is the comparison of means (i.e., averages). We have also seen that the process of using this statistic consists of calculating it and then comparing it with a threshold value, which is obtained from a table of the Student's  $t$  distribution.

The comparison of observed and tabled values depends not only on the level of significance but also on a function of the sample size, the degrees of freedom. Since most tables of the  $t$  distribution go no higher than 100 or 120 in providing numbers of degrees of freedom, the question arises: What should you do when the number of degrees of freedom in a particular problem is larger than what the available  $t$ -distribution tables cover?

The answer is that in such cases there will be only a slight difference between (a) the unavailable threshold value of Student's  $t$  and (b) the corresponding value of the Gaussian distribution. Hence, you can simply use the Gaussian table (i.e., the normal, or  $z$ -distribution, table) to provide the threshold. For the most frequently used levels of significance, 5% and 1%, the Gaussian thresholds are 1.96 and 2.58, respectively.

### Estimating Population Proportions

From time to time problems arise concerning what proportion of a given population has a certain characteristic. One frequent example is the question of what percentage of the population (in the statistical sense) of voters in the next election will vote for Candidate A or will vote in favor of Proposition X. In the realm of libraries, one might be interested in such matters as what proportion of the population of library users would back their library's director in resisting attempts by special-interest groups to remove an "obscene" book from the library.

A suitable procedure in such cases is to take a sample and observe the proportion  $P$  of elements in the sample that possess the characteristic of interest (e.g., the intention of voting for a library-bond issue). You can then form a confidence interval (CI) for the population proportion, which we shall denote by  $\pi$ . To do so, you will need to use the evidence in the sample. This evidence consists of  $P$  and the size,  $n$ , of the sample.

As you may recall, the pattern for constructing CIs for a population mean is

(sample mean)  $\pm$  (confidence factor)(standard error)

In the case of proportions, the rôle of the sample mean is played by the sample proportion,  $P$ . The pertinent standard error is the *standard error of the sample proportion*, denoted  $s_{\bar{p}}$ , which plays the analogous rôle here to that of  $s_{\bar{x}}$ , the standard error of the mean. To derive the standard error of the sample proportion, we begin with the standard deviation of a sample proportion. Omitting the details, we simply state that this standard deviation is most easily computed as

$$s = \sqrt{P(1 - P)}$$

and that the corresponding standard error of the sample proportion is given by

$$s_{\bar{p}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{P(1 - P)}}{\sqrt{n}} = \sqrt{\frac{P(1 - P)}{n}}$$

Carrying out the same processes as we used for calculating confidence intervals for population means, we can find the 95% CI for the population proportion,  $\pi$ , by calculating

$$P \pm 1.96s_{\bar{p}} = P \pm 1.96\sqrt{\frac{P(1 - P)}{n}}$$

and the 99% CI by calculating

$$P \pm 2.58s_{\bar{p}} = P \pm 2.58\sqrt{\frac{P(1 - P)}{n}}$$

The foregoing discussion gives you the basic ideas about how to estimate a population proportion.

However, an important detail remains to be mentioned: the size of the sample. In light of previous discussions of how to estimate population parameters, you are probably expecting a rule-of-thumb for distinguishing between large samples, for which you would use the Gaussian distribution to provide the confidence factor, and small samples, for which you would use the Student's  $t$  distribution. Unfortunately, the estimation of population proportions from small samples requires using, not the  $t$  distribution, but another distribution, the binomial. Lacking the time in course LIS 397.1 to treat the use of the binomial distribution for this purpose, we shall dodge the question of estimating population proportions from small samples, and shall confine ourselves to samples large enough for the Gaussian distribution to be used.

But what is a "large" sample in this situation. The usual rule-of-thumb is that if both  $nP$  and  $n(1-P)$  exceed 5, then the sample size is large enough for you to use the Gaussian distribution in constructing the confidence intervals (e.g., for you to use 1.96 for 95% CIs and 2.58 for 99% CIs).

For example, if you took a sample of size  $n = 15$  and found  $P = 0.4$  to be the proportion possessing the characteristic of interest, you would be justified in using the Gaussian distribution; for you would have  $nP = 15(.4) = 6$  and  $n(1 - P) = 15(1 - .4) = 15(.6) = 9$ , both of which exceed 5. (This example shows you that it is not hard for a sample to be "large" in the sense indicated above.) The 95% CI in this case would be

$$0.4 \pm 1.96\sqrt{\frac{0.4(0.6)}{15}} = 0.4 \pm 1.96(0.13) = 0.4 \pm 0.25 = (0.15, 0.65)$$

## Tests of Hypotheses about Population Proportions

A problem related to the foregoing discussion is that of testing a hypothesis that the population proportion,  $\pi$ , is equal to a particular numerical value, which we shall represent by " $\pi_0$ ". This null hypothesis can be written symbolically as

$$H_0: \pi = \pi_0$$

The test statistic is

$$z_{obs} = \frac{P - \pi_0}{s_{\bar{P}}} = \frac{P - \pi_0}{\sqrt{\frac{P(1-P)}{n}}}$$

When you have calculated the observed value of the test statistic, you compare it with the tabled threshold value of  $z$  determined by the level of significance at which you have chosen to work (e.g., 1.96 for a 5% level of significance or 2.58 for a 1% level of significance). As usual in hypothesis testing, if the (absolute) value of the observed test statistic exceeds the tabled threshold value, you can reject the null hypothesis; otherwise, you cannot reject it.

Quite similar is the problem of testing a hypothesis that two populations, from each of which you have drawn a sample, have equal proportions of the characteristic of interest (or equivalently, that they have the same proportion). Using " $\pi_1$ " and " $\pi_2$ " to represent the population proportions, we can write this hypothesis symbolically as

$$H_0: \pi_1 = \pi_2$$

The evidence available from the two samples is the observed sample proportions,  $P_1$  and  $P_2$ , and the sample sizes,  $n_1$  and  $n_2$ . The rule of thumb is that if each of the products  $n_1 P_1$ ,  $n_2 P_2$ ,  $n_1(1 - P_1)$ , and  $n_2(1 - P_2)$  exceeds 5, you can consider the samples large enough to use the Gaussian distribution in testing the hypothesis. (Again we omit discussion of what to do with samples too small to satisfy this criterion, noting only that it is easy for samples to be large enough to satisfy it.)

The preferred procedure is to pool the information in the two samples concerning the population proportions, under the assumption that the null hypothesis is true. (This is analogous to the pooled estimate of variance procedure in testing a hypothesis that two population means are equal.)

To obtain the pooled proportion,  $P_p$ , which is the weighted average of the observed sample proportions with their respective sample sizes taken as the weights, you form

$$P_p = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

As you will recall, in the case of testing a hypothesis that the means of two populations are equal, one uses the standard error of the difference of means. In similar fashion in the present case, testing whether two population proportions are equal, we use the *standard error of the difference in proportions*, which is represented by  $s_{\bar{P}_1 - \bar{P}_2}$ . We can write this as

$$s_{\bar{P}_1 - \bar{P}_2} = \sqrt{[P_p(1 - P_p)] \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

which should remind you of both the standard error of the sample proportion and the standard error of the difference of means.

The test statistic has a form that should, by now, be familiar to you. It is the difference between the observed sample proportions, as compared to (i.e., as divided by) the standard error of such differences:

$$z_{obs} = \frac{p_1 - p_2}{s_{\bar{p}_1 - \bar{p}_2}}$$

When you have calculated the observed value of the test statistic, you compare it with the tabled threshold value of  $z$  determined by the level of significance at which you have chosen to work (e.g., 1.96 for a 5% level of significance or 2.58 for a 1% level of significance). If the (absolute) value of the observed test statistic exceeds the tabled threshold value, you can reject the null hypothesis; otherwise, you cannot reject it.