10 Inferences about Population Means

In this chapter we are going to discuss ways for making inferences about means, first for a single population and then for a difference between the means of two populations. The procedures and conventions of significance testing were emphasized in the previous chapter, and now we are going to apply these procedures to tests of hypotheses about means. Remember that a significant result will always be one falling among those that are extremely deviant from expectation and that are improbable if the null hypothesis were true, but that agree relatively well with expectation and have relatively higher probability if some situation covered by the alternative hypothesis were true. Before the test is carried out, some α level for significance is chosen as a specification of "improbable event, given H_0," for this situation. The conventional rules of the game determine which of the values of α are chosen.

10.1 Large-sample problems with unknown population σ^2

In most of the examples of hypothesis testing up to this point we have actually "fudged" a bit on the usual situation: we have assumed that σ^2 is somehow known, so that the standard error of the mean is also known exactly. In these examples the author did not explain how σ^2 became known, largely because he could not think up a good reason. Now we must face the cold facts of the matter: for inferences about the population mean, σ^2 is seldom known. Instead, we must use the only substitute available for σ^2, which is our unbiased estimate s^2, calculated from the sample.

Notice that this problem does not exist for hypotheses about a population proportion p, since the existence of an exact hypothesis about p specifies

what the value of the standard error of P, the sample proportion, must be. Therefore, the special techniques of this chapter apply only to inferences about means, and not to inferences about proportions.

From what we have already seen of the relation between sample size and accuracy of estimation, it makes sense that for large samples s^2 should be a very good estimate of σ^2. *In general, for very large samples, there is rather little risk of a sizable error when one uses s in place of σ in estimating the standard error of the mean.*

Hence, when the sample size is quite large, tests of hypotheses about a single mean are carried out in the same way as when σ is known, except that the standard error of the mean is estimated from the sample:

$$\text{est. } \sigma_M = \frac{s}{\sqrt{N}} = \frac{S}{\sqrt{N-1}}.$$

The standardized score corresponding to the sample mean is then referred to the normal distribution. This step is justified by the central limit theorem when N is large, regardless of the population distribution's form.

For example, consider the following problem. A small rodent characteristically shows hoarding behavior for certain kinds of foodstuffs when the environmental temperature drops to a certain point. Numerous previous experiments have shown that in a fixed period of time, and given a fixed food supply, the mean amount of food hoarded by an animal is 9 grams. The experimenter is currently interested in possible effects that early food deprivation may have upon such hoarding behavior in the animal as an adult. So, the experimenter takes a random sample of 175 infant animals and keeps them on survival rations for a fixed period while they are at a certain age, and on regular rations thereafter. When the animals are adults he puts each one in an experimental situation where the lowered temperature condition is introduced. The amount of food each hoards is recorded, and a score is assigned to each animal.

What is the null hypothesis implied here? The basic experimental question is "Does the experimental treatment (deprivation) tend to affect the amount of food hoarded?" The experimenter has no special reason to expect either an increase or a decrease in amount, but is interested only in finding out if a difference from normal behavior occurs. This question may be put into the form of a null and an alternative hypothesis:

$$H_0\colon \mu_0 = 9 \text{ grams}$$
$$H_1\colon \mu \neq 9 \text{ grams}.$$

Suppose that the conventional level chosen for α is .01, so that the experimenter will say that the result is significant only if the sample mean falls among either the upper .005 or the lower .005 of all possible results, given H_0. Reference to Table I shows that .005 is the probability of z score in a normal distribution falling at or below -2.58, and the probability is likewise .005 for a z equal to or exceeding $+2.58$. Accordingly, the sample result will be significant

only if

$$z_M = \frac{M - E(M)}{\text{est. } \sigma_M}$$

equals or exceeds 2.58 in absolute magnitude (disregarding sign). When the null hypothesis is true, $E(M) = 9$, and for a sample this large the value of the standard error of the mean should be reasonably close to $\frac{s}{\sqrt{N}}$ or $\frac{S}{\sqrt{N-1}}$, the value of the sample estimate.

Everything is now set for a significance test except for the sample results. The sample shows a mean of 8.8 grams of food hoarded, with a standard deviation, S, of 2.3. The estimated standard error of the mean is thus

$$\text{est. } \sigma_M = \frac{2.3}{\sqrt{175 - 1}} = \frac{2.3}{13.23} = .1738$$

The standardized score of the mean is found to be

$$z_M = \frac{8.8 - 9}{.174} = \frac{-.2}{.174} = -1.149.$$

This result does not qualify for the region of rejection for $\alpha = .01$. Since the experimenter feels that he can afford to reject H_0 only if the α probability of error is no more than .01, then he cannot do so on the basis of this sample. On the other hand, the risk run in accepting H_0 is unknown, so he might well suspend judgment, pending more evidence.

10.2 Confidence Intervals for Large Samples with Unknown σ^2

Confidence intervals may also be found by the methods of Chapter 9. However, either when σ^2 is unknown, or when the population distribution has unknown form, a normal sampling distribution is assumed only for large samples. Just as in significance tests, the estimated standard error of the mean can be used in place of σ_M in finding confidence limits when the sample is relatively large.

For example, the experimenter studying hoarding behavior computes the approximate 99 percent confidence limits in the following way:

$$M - 2.58 \text{ (est. } \sigma_M)$$

and

$$M + 2.58 \text{ (est. } \sigma_M)$$

so that for this problem, the numerical confidence limits are

$$8.8 - 2.58(.174) \text{ or } 8.35$$

and

$$8.8 + 2.58(.174) \text{ or } 9.25.$$

The experimenter can say that the probability is approximately .99 that the true value of μ is covered by an interval such as that between 8.35 and 9.25.

Notice that the value $\mu_0 = 9$ falls between these limits, reflecting the fact that the hypothesis H_0 cannot be rejected if α is set at .01 (two-tailed).

10.3 The problem of unknown σ^2 when sample size is small

Just as for any statistic used to estimate a parameter value, the estimated standard error of the mean will very likely not be exactly equal to σ_M. This is not a particular problem when sample size is large, since we can at least be sure that est. σ_M is very likely to be close to the true σ_M in value.

On the other hand, we simply cannot have this confidence in our estimate of the standard error when sample size is small. Our estimate is almost bound to be in error to some extent, and if the sample size is very small, we can expect the size of this error to be substantial in any given sample. This necessitates a different approach to the problem of testing hypotheses and establishing confidence intervals for the population mean for small samples.

In inferences about μ, the ratio we would like to evaluate and refer to a normal sampling distribution is the standardized score

$$z_M = \frac{M - E(M)}{\sigma_M}. \qquad [10.3.1^*]$$

However, when we have only an *estimate* of σ_M, then the ratio we really compute and use is not a normal standardized score at all, although it has much the same form. The ratio actually used is

$$t = \frac{M - E(M)}{\text{est. } \sigma_M}. \qquad [10.3.2^*]$$

There is an extremely important difference between the two ratios, z_M and t. For z_M, the numerator $(M - E(M))$ is a random variable, the value of which depends upon the particular sample drawn from a given population situation; on the other hand, the denominator is a constant, σ_M, which is the same regardless of the particular sample of size N we observe. Now contrast this ratio with the ratio t: just as before, the numerator of t is a random variable, but the denominator is also a random variable, since the particular value of s— and hence the estimate of σ_M— is a sample quantity. Over several different samples, the same value of M must give us precisely the same value of z_M; however, over different samples, the same value of M will give us different t values. Similar intervals of t and z_M values should have different probabilities of occurrence. For this reason it is risky to use the ratio t as though it were z_M unless the sample size is very large.

10.4 The distribution of t

The solution to the problem of the nonequivalence of t and z_M rests on the study of t itself as a random variable. That is, suppose that the t ratio were

computed for each conceivable sample of N independent observations drawn from some normal population distribution with true mean μ. Each sample would have some t value,

$$t = \frac{M - E(M)}{\text{est. } \sigma_M} = \frac{M - \mu}{s/\sqrt{N-1}}. \qquad [10.4.1^*]$$

Over the different samples the value of t would vary, of course, and the different possible values would each have some probability-density. A random variable such as t is an example of a *test*-statistic, so called to distinguish it from an ordinary descriptive statistic or estimator, such as M or s^2. The t value depends on other sample statistics, but is not itself an estimate of a population value. Nevertheless, such test-statistics have sampling distributions just as ordinary sample statistics do, and these sampling distributions have been studied extensively.

In order to find the exact distribution of t, one must assume that the basic population distribution is normal. The main reason for the necessity of this assumption is that only for a normal distribution will the basic random variables in numerator and denominator, sample M and s, be statistically independent; this is a use of the important fact mentioned in Section 8.8. Unless M and s are independent, the sampling distribution of t is extremely difficult to specify exactly. On the other hand, for the special case of normal populations, the distribution of the ratio t is quite well known. In order to learn what this distribution is like, let us take a look at the rule for the density function associated with this random variable.

The density function for t is given by the rule:

$$f(t;\, \nu) = G(\nu)\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2} \quad \begin{matrix} -\infty < t < \infty \\ 0 < \nu \end{matrix} \qquad [10.4.2^*]$$

Here, $G(\nu)$ stands for a constant number which depends *only* on the parameter ν (Greek nu), and how this number is found need not really concern us. Let us focus our attention on only the "working part" of the rule, which involves only ν and the value of t. This looks very different from the normal distribution function rule in Section 8.1. As with the normal function rule, however, a quick look at this mathematical expression tells us much about the distribution of t (for $\nu > 1$).

First of all, notice that the particular value of t enters this rule only as a squared quantity, showing that the distribution of sample t values must be symmetric, since a positive and a negative value having the same absolute size must be assigned the same probability-density by this rule. Second, since all the constants in the function rule are positive numbers, and the entire term involving t is raised to a negative power, the largest possible density value is assigned to $t = 0$. Thus $t = 0$ is the distribution mode. Furthermore, although it is not quite so apparent from an examination of the function rule, the distribution is unimodal and "bell-shaped." If we inferred from the symmetry and unimodality of this distribution that the mean of t is also 0, we should be quite correct. In short, the t distribution is a unimodal, symmetric, bell-shaped distribution having a graphic form much like a normal distribution, even though the two function rules are quite

dissimilar. Loosely speaking, the curve for a t distribution differs from the standardized normal in being "plumper" in extreme regions and "flatter" in the central region, as Figure 10.4.1 shows. (Note that both t and the standardized normal distribution have a mean of zero, $\nu > 1$.)

The most important feature of the t distribution will appear if we return for a look at the function rule. Notice that the only unspecified constants in the rule are those represented in **10.4.2** by ν and $G(\nu)$, which depends only on ν. This is a **one-parameter distribution:** the single parameter is ν, called *the degrees of freedom. Ordinarily, in most applications of the t distribution to problems involving a single sample, ν is equal to $N - 1$, one less than the number of independent observations in the sample.* For samples of N independent observations from any normal population distribution, the exact distribution of sample t values depends only on the degrees of freedom, $N - 1$. Remember, however, that the value of $E(M)$ or μ must be specified when a t ratio is computed, although the true value of σ need not be known.

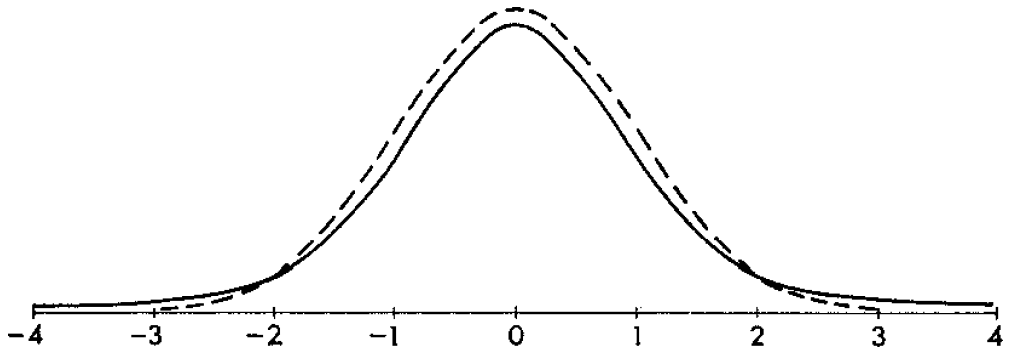

FIG. 10.4.1. Distribution of t with $\nu = 4$, and standardized normal distribution

In principle, the value of ν can be any positive number, and it just happens that $\nu = N - 1$ is the value for the degrees of freedom for the particular t distributions we will use first. Later we will encounter problems calling for t distributions with other numbers of degrees of freedom. Like most theoretical distributions, the t distribution is actually a family of distributions, with general form determined by the function rule, but with particular probabilities dictated by the parameter ν. For any value of $\nu > 1$, the mean of the distribution of t is 0.For $\nu > 2$ the variance of the t distribution is $\nu/(\nu - 2)$, so that the smaller the value of ν the larger the variance. As ν becomes large the variance of the t distribution approaches 1.00, which is the variance of the standardized normal distribution.

Incidentally, the random variable t is often called "Student's t," and the distribution of t, "Student's distribution." This name comes from the statistician W. S. Gosset, who was the first to use this distribution in an important problem, and who first published his results in 1908 under the pen-name "Student." Distributions of the general "Student" form have a number of important applications in statistics. One such application will occur in Chapter 19. It should also be noted that Student distributions are closely related to the beta family of distributions discussed in Chapter 8. This connection will be developed more fully in the next chapter.

10.5 The t and the Standardized Normal Distribution

As we have seen, the "shape" of the t distribution is not unlike that of the normal distribution. Just as for the standardized normal, the mean of the distribution of t is 0 for $\nu > 1$ although the variance of t is greater than 1.00 for finite $\nu > 2$. Given any extreme interval of fixed size on either tail of the t distribution, the probability associated with this interval in the t distribution is larger than that for the corresponding normal distribution of z_M. The smaller the value of ν, the larger is this discrepancy between t and normal probabilities at the extreme ends of each distribution. This reflects and partly explains the danger of using a t ratio as though it were a z ratio: extreme values of t are relatively more likely than comparable values of z_M. A small sample size corresponds to a small value of ν, or $N - 1$, and thus there is serious danger of underestimating the probability of an extreme deviation from expectation when sample size is small. This is apparent in the illustration (Figure 10.4.1) showing the distribution of t together with the standardized normal function.

Suppose that a sample of 5 observations is drawn, and from this sample we compute a ratio, t, using the estimate of σ_M from the sample. Furthermore, suppose that

$$t = \frac{M - E(M)}{\text{est. } \sigma_M} \geq 2.13.$$

That is, we obtain a value for t greater than or equal to 2.13. In the t distribution for samples of size 5 ($\nu = 4$), this interval has probability of .05. That is, when sample size is 5, so that degrees of freedom are 4, the probability of obtaining a ratio in this interval of values is 1/20. However, if the ratio is interpreted as a z_M variable, then the normal probability for this interval is .0166. Incorrectly considering a t ratio as a standardized normal variable leads one to underestimate the probability of values in extreme intervals, which are really the only intervals of interest in significance tests.

On the other hand, notice what should happen to the distribution of t as ν becomes large (sample size grows large), as suggested both by Figure 10.4.1 and by the variance of a t distribution. *As sample size N grows large, the distribution of t approaches the standardized normal distribution. For large numbers of degrees of freedom, the exact probabilities of intervals in the t distribution can be approximated closely by normal probabilities.*

The practical result of this convergence of the t and the normal probabilities is that the t ratio *can* be treated as a z_M ratio, provided that the sample size is substantial. The normal probabilities are quite close to—though not identical with—the exact t probabilities for large ν. On the other hand, when sample size is small the normal probabilities cannot safely be used, and instead one uses a special table based on the t distribution.

How large is "large enough" to permit use of the normal tables? If the population distribution is truly normal, even forty or so cases permit a fairly

accurate use of the normal tables in confidence intervals or tests for a mean. If really good accuracy is desired in determining interval probabilities, the t distribution should be used even when the sample size is around 100 cases. Beyond this number of cases, the normal probabilities are extremely close to the exact t probabilities. For example, in the "hoarding" experiment just discussed, use of t rather than z values would have given confidence limits of $M \pm 2.6(\text{est}\,\sigma_M)$ instead of $M \pm 2.58(\text{est}\,\sigma_M)$, a very slight difference.

Recall that the stipulation is made that the population distribution be *normal* when the t distribution is used, even when the normal approximations are substituted for the exact t-distribution probabilities. As we have already seen, for a normal population the distribution of sample means must be normal anyway; the difficulty with the use of a normal sampling distribution for small N comes solely from the fact that our estimate of the standard error is a random variable rather than a constant over samples, and this is the reason we must use the t distribution. Thus the t distribution is related to the normal distribution in two distinct ways: the parent distribution must be normal if t probabilities are to be found exactly, and for sufficiently large N, the distribution of t approaches the normal sampling distribution in form.

10.6 The application of the t distribution when the population is not normal

It is apparent that the requirement that the population be normal limits the usefulness of the t distribution, since this is an assumption that we can seldom really justify in practical situations. Fortunately, when sample size is fairly large, and provided that the parent distribution is roughly unimodal and symmetric, the t distribution apparently still gives an adequate approximation to the exact (and often unknown) probabilities of intervals for t ratios under these circumstances. However, one should insist on a relatively *larger* sample size the *less* confident that he is that the normal rule holds for the population, if he plans to use the t distribution. In effect, if the sample size is large enough so that the normal probabilities are good approximations to the t probabilities anyway, then the form of the parent distribution is more or less irrelevant. However, often the sample size is so small that the t distribution must be used and here it *is* somewhat risky to make inferences from t ratios unless the population is more or less normally distributed. This is an especially serious problem when one-tailed tests of hypotheses are made, since a very skewed population distribution can make the t probabilities for one-tailed tests considerably in error. Once again, it is wise to plan on somewhat larger samples when one is considering a one-tailed test using the t distribution and the population is not assumed normal.

10.7 Tables of the t distribution

Unlike the table of the standardized normal function, which suffices for all possible normal distributions, tables of the t distribution must actually in-

clude many distributions each depending on the value of ν, the degrees of freedom. Consequently, tables of t are usually given only in abbreviated form; otherwise, a whole volume would be required to show all the different t distributions one might need.

Table III in Appendix C shows selected percentage points of the distribution of t, in terms of the value of ν. Different ν values appear along the left-hand margin of the table. The top margin gives values of Q, which is $1 - p(t \leq a)$, one minus the cumulative probability that t is less than or equal to a specific value a, for a distribution within the given value for ν. A cell of the table then shows the value of t cutting off the upper Q proportion of cases in a distribution for ν degrees of freedom.

This sounds rather complicated, but an example will clarify matters considerably: suppose that $N = 10$, and we want to know the value *beyond which* only 10 percent of all sample t values should lie. That is, for the distribution of t shown in Figure 10.7.1, we want the t value that cuts off the shaded area in the curve, the upper 10 percent:

First of all, since $N = 10$, $\nu = N - 1 = 9$. So, we enter the table for the row marked 9. Now since we want the upper 10 percent, we find the column for which $Q = .1$. The corresponding cell in the table is the value of t we are looking for, $t = 1.383$. We can say that in a t distribution with 9 degrees of freedom,

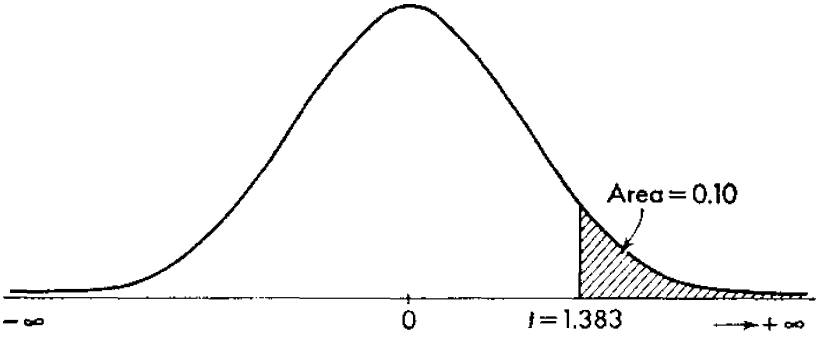

FIG. 10.7.1

the probability is .10 that a t value *equals or exceeds* 1.383. Since the distribution is symmetric, we also know that the probability is .10 that a t value *equals or falls below* −1.383. If we wanted to know the probability that t equals or exceeds 1.383 in absolute value, then this must be .10 + .10 = .20, or $2Q$.

Suppose that in a sample of 21 cases, we get a t value of 1.98. We want to see if this value falls into the upper .05 of all values in the distribution. We enter row $\nu = 21 - 1 = 20$, and column $Q = .05$; the t value in the cell is 1.725. Our obtained value of t is larger than this, and so the obtained value does fall among the top 5 percent of all such values. On the other hand, suppose that the obtained t had been −3. Does this fall either in the top .001 or the bottom .001 of all such sample values? Again with $\nu = 20$, but this time with $Q = .001$, we find a t value of 3.552. This means that at or above 3.552 lie .001 of all sample values, and also at or below −3.552 lie .001 of all sample values. Hence our sample value does not fall into either of these intervals; we can say that the sample value does not fall into the rejection region for $\alpha = .002$.

The very last row, marked ∞, shows the z scores that cut off various areas in a normal distribution curve. If you trace down any given column, you find that as ν gets larger the t value bounding the area specified by the column comes closer and closer to this normal deviate value, until finally, for an infinite sample size, the required value of t is the same as that for z.

For one-tailed tests of hypotheses, the column Q values are used to find the t value which exactly bounds the rejection region. If the region of rejection is on the upper tail of the distribution, then Q is the probability of a sample value's falling into the region greater than or equal to the tabled value of t. If the region is on the lower tail, the t value in the table is given a negative sign, and Q is the probability of a sample's falling at or below the negative t value. If a two-tailed region is to be used, then the total α probability of error is $2Q$, and the number in the table shows the absolute value of t that bounds the rejection region on *either* tail.

10.8 The concept of degrees of freedom

Before we proceed to the uses of the t distribution, it is well to examine the notion of degrees of freedom. The degrees of freedom parameter reflects the fact that a t ratio involves a sample standard deviation as the basis for estimating σ_M. Recall the basic definitions of the sample variance and the sample standard deviation:

$$S^2 = \frac{\Sigma(x - M)^2}{N}$$

and

$$S = \sqrt{\frac{\Sigma(x - M)^2}{N}}.$$

The sample variance and standard deviation are both based upon a sum of squared deviations from the sample mean. However, recall another fact of importance about deviations from a mean: in Section 6.6 it was shown that

$$\sum_i (x_i - M) = 0,$$

the sum of deviations about the mean must be zero.

These two facts have an important consequence: Suppose that you are told that $N = 4$ in some sample, and that you are to guess the four deviations from the mean M. For the first deviation you can guess any number, and suppose you say

$$d_1 = 6.$$

Similarly, quite at will, you could assign values to two more deviations, say

$$d_2 = -9$$
$$d_3 = -7.$$

However, when you come to the fourth deviation value, you are *no longer free* to guess any number you please. The value of d_4 *must* be

$$d_4 = 0 - d_1 - d_2 - d_3$$

or

$$d_4 = 0 - 6 + 9 + 7 = 10.$$

In short, given the values of any $N - 1$ deviations from the mean, which could be any set of numbers, the value of the last deviation is completely determined. Thus we say that there are $N - 1$ degrees of freedom for a sample variance, reflecting the fact that only $N - 1$ deviations are "free" to be any number, but that given these free values, the last deviation is completely determined. It is not the sample size per se that dictates the distribution of t, but rather the number of degrees of freedom in the variance (and standard deviation) estimate. We will consider the degrees of freedom again in the next chapter, where the variance will be studied in more detail, and also in Chapter 14.

10.9 Significance tests for single means using the t distribution

For the moment you can relax; there is really nothing new to learn! When the null hypothesis concerns a single mean, then the test is carried out just as before, except that the table of t (Appendix C, Table III) is used instead of the normal table (Appendix C, Table I). The α level is chosen, and the value (or values) of t corresponding to the region of rejection can be determined from the t table. The number of degrees of freedom used is simply $\nu = N - 1$. Then, the ratio

$$t = \frac{M - E(M)}{\text{est. } \sigma_M} \qquad [10.9.1\dagger]$$

obtained from the sample is compared with values in the rejection region specified by Table III. If the obtained t ratio falls into the rejection region chosen, the sample result is said to be significant beyond the α level.

If the sample size is large, then the only difference in procedure is in the use of the normal tables to establish the region of rejection. Naturally, all the considerations hitherto discussed, especially the assumed normal distribution of the population, should be faced before the sample size and rejection region are decided upon. If large samples are available then the assumption of a normal population is relatively unimportant; on the other hand, this matter should be given some serious thought if you are limited to a very small sample size.

10.10 Confidence limits for the mean using t distributions

The t distribution may also be used to establish confidence limits for the mean. For some fixed percentage representing the confidence level, $100(1 - \alpha)$ percent, the sample confidence limits depend upon three things: the sample value of M, the estimated standard error, or est. σ_M, and the number of degrees of freedom, ν. For some specified value of ν, then the $100(1 - \alpha)$ percent

confidence limits are found from

$$M - t_{(\alpha/2;\,\nu)}\,(\text{est. } \sigma_M)$$
$$M + t_{(\alpha/2;\,\nu)}\,(\text{est. } \sigma_M). \qquad [10.10.1\dagger]$$

Here, $t_{(\alpha/2;\,\nu)}$ represents the value of t that bounds the upper $\alpha/2$ proportion of cases in a t distribution with ν degrees of freedom. In Table II this is the value listed for $Q = \alpha/2$ and ν. Thus, if one wants the 99 percent confidence limits, the value of $\alpha = .01$, and one looks in the table for $Q = .005$.

For example, imagine a study using 8 independent observations drawn from a normal population. The sample mean is 49 and the estimated standard error of the mean is 3.7. Now we want to find the 95 percent confidence limits. First of all, $\alpha = .05$, so that $Q = .025$. The value of ν is $N - 1$, or 7. The table shows a t value of 2.365 for $Q = .025$ and $\nu = 7$, so that $t_{(\alpha/2;\,\nu)} = 2.365$. The confidence limits are

$$49 - (2.365)(3.7) = 40.25$$
and $$49 + (2.365)(3.7) = 57.75.$$

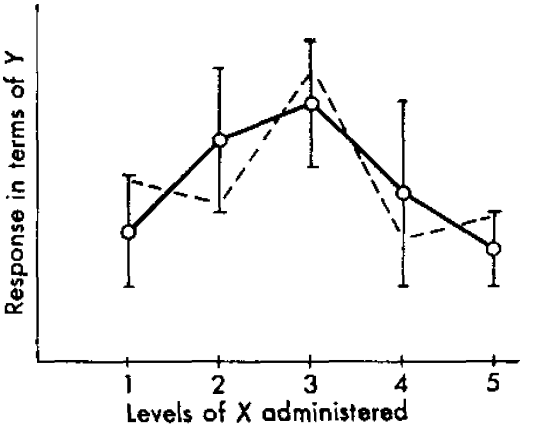

FIG. 10.10.1. Confidence intervals for means based on five independent samples

Over all random samples, the probability is .95 that the true value of μ is covered by an interval such as that between 40.25 and 57.75, the confidence interval calculated for this sample.

In summary, confidence limits are calculated in much the same way, and have the same general interpretation, when based on the t distribution as for the normal distribution. The essential difference is that values of t corresponding to $\alpha/2$ and ν must be used instead of normal z values.

One important application of confidence intervals in psychology occurs when there is a set of some J independent means, each based on a different sample given exactly one of a set of J experimental treatments. In particular, the experimental treatments may represent some quantitative experimental variable (represented here by X), and the experimenter may be trying to infer the general form of relationship between the amount of treatment applied and the average or expected response of a subject in terms of variable Y. Here, he may choose to construct a confidence interval around *each* of the sample means on variable Y. Figure 10.10.1 represents a set of such means, with the 95 percent confidence interval shown for each. In this figure, the horizontal axis represents the different levels or quantities of the treatment administered, and the open circles the corresponding means of the samples on the dependent variable, Y. The vertical bars extending to either side of a mean point symbolize the 95 percent confidence interval based on that sample's mean. The experimenter's best guess about the general form of the function relating the experimental variable to the dependent variable is symbolized by the heavy line in the figure: this is simply a plot joining the sample

means, since his best guess about the population mean under any given treatment is the sample mean. Nevertheless, it may well be true that the form of relationship in the population is something like that shown by the broken line. The experimenter has no basis at all for discounting the possibility of some such true relation on the basis of the obtained relation alone.

How sure can the experimenter be that the set of J confidence intervals based on independent means all *simultaneously* cover the population values? In other words, how confident can he be that he has narrowed the possible relationships between the experimental and dependent variables to those symbolized by graphs joining points *within* the various intervals of Figure 10.10.1? A little thought should convince you that the probability is *not* .95 that all J of the confidence intervals simultaneously cover the true means; we can, however, work out an approximation to the value of this probability.

Suppose that both the means and the estimated σ_M values, and thus the obtained values for the confidence limits themselves, are independent across samples. We can consider the event "confidence interval covers true μ" as though it were a "success" in a binomial experiment for any of the samples. The probability of any such success is .95 for a 95 percent confidence interval. Then the probability that all J independent confidence intervals simultaneously cover the true means is simply the probability of exactly J out of J possible successes in a binomial experiment:

$$\text{prob.(all } J \text{ of the 95 percent confidence intervals cover true values simultaneously)} = \binom{J}{J} (.95)^J (.05)^0 = (.95)^J.$$

For the example in Figure 10.10.1, $J = 5$, so that the probability that the true means all are covered by the indicated intervals is

$$(.95)^5 = .77.$$

The experimenter can have considerably less "confidence" in the statement that *all* of the confidence intervals simultaneously cover the true means than in the statement that *any one* confidence interval covers the true mean.

The probability of .77 calculated for this example was based on the assumption that each of the confidence limits obtained for a sample is independent of the corresponding limits obtained for the other samples. However, this is not a reasonable assumption in a great many instances, because the same estimated value of σ or of σ_M may be used for determining each of the confidence intervals. Nevertheless, even when the confidence limits for the various samples are not independent, the probability that all J of the $100(1 - \alpha)$ percent confidence intervals simultaneously cover the true values must lie between $1 - \alpha$ and $1 - J\alpha$. Conversely, given any J such confidence intervals, the probability that *at least one* fails to cover the true value is between α and $J\alpha$. For J independent confidence intervals, the probability that at least one of the set fails to cover the true value is

exactly $1 - (1 - \alpha)^J$. The practical implication is clear: given enough confidence intervals calculated from a set of data the probability can be quite high that *at least one* fails to cover the true parameter value. For similar reasons, given enough significance tests carried out on a set of data, each with some conventional value for α, the probability can be much greater than α that *at least one* of these tests results in a Type I error. This point is an important one, and will recur in Chapters 12 and 14.

10.11 Questions about differences between population means

Examples of hypotheses about single means often sound rather "phony" in their experimental contexts, and the reason for this is not hard to find. In most experimental work it is not true that the experimenter knows about one particular population in advance and then draws a single sample for the purpose of comparing some experimental population to the known population. Rather, it is far more common to draw two samples, to only one of which the experimental treatment is applied; the other sample is given no treatment, and stands as a control group for comparison with the treated group. In other situations, two different treatments may be compared. The advantages of this method over the single sample procedure are obvious; the experimenter can exercise the same experimental controls on both samples, making sure that insofar as possible they are treated in exactly the same way, with the only systematic experimental difference being in the fact that something was done to representatives of one sample which was not done to members of the other. Then, if a very large difference appears between the two samples he can rest assured that the difference is a product of the experimental treatments and not just a peculiarity introduced by the way in which his data were gathered.

Each treatment group is a sample from a potential population of observations made under that treatment. A difference between the treatment populations should exist if the treatment is having an effect; but what can the experimenter infer from a sample difference? *His best estimate (based on these data alone) is that the population means are different to the same extent as the sample means. Regardless of the significance level given by any test he may apply, the actual difference obtained is always the best estimate he can make of the true difference between the population means.*

As always, this estimate is in error to some unknown extent, and although the obtained difference between the sample means is the best guess the experimenter can make, there is absolutely no guarantee that this estimate is exactly correct. It could well be true that the difference the experimenter observes has no real connection with the treatment administered, and is purely a chance result.

What is needed is a way of applying statistical inference to differences between means of samples representing two populations. First, large sample distributions of *differences* between sample means will be studied. Then, the application of the t distribution to small sample differences will be introduced.

10.12 The sampling distribution of differences between means

Suppose that we wished to test a hypothesis that two populations have means which differ by some specified amount, say 20 points. This is tested against the hypothesis that the population means do not differ by that amount. In our more formal notation:

$$H_0: \mu_1 - \mu_2 = 20$$
$$H_1: \mu_1 - \mu_2 \neq 20.$$

We draw a sample of size N_1 from population 1, and an *independent* sample of size N_2 from population 2, and consider the difference between their means, $M_1 - M_2$. Now suppose that we kept on drawing pairs of independent samples of these sizes from these populations. For each pair of samples drawn, the difference $M_1 - M_2$ is recorded. What is the distribution of such sample *differences* that we should expect in the long run? In other words, what is the sampling distribution of the difference between two means?

You may already have anticipated the form of the sampling distribution of the difference between two means, since all the groundwork for this distribution has been laid in Section 8.9. The difference between sample means drawn from independent samples is actually a linear combination:

$$(1)M_1 + (-1)M_2.$$

Let us apply the results of Sections 8.9 and 8.10 to this problem. In the first place,

$$E(M_1 - M_2) = E(M_1) - E(M_2) = \mu_1 - \mu_2, \qquad [10.12.1^*]$$

which accords with principle **8.10.1** for any linear combination. Second, what is the standard error of the difference between two independent sample means? By principle **8.10.3**,

$$\begin{aligned} \text{var.}(M_1 - M_2) &= (1)^2\sigma^2_{M_1} + (-1)^2\sigma^2_{M_2} \\ &= \sigma^2_{M_1} + \sigma^2_{M_2}. \end{aligned} \qquad [10.12.2^*]$$

Hence, the standard error of the difference, $\sigma_{\text{diff.}}$, is

$$\sigma_{\text{diff.}} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \qquad [10.12.3^*]$$

provided that samples 1 and 2 are completely independent.

Actually, we could have found this last result quite easily without invoking principle **8.10.2**. It may be instructive to do so.

By definition:

$$\begin{aligned} \text{var.}(M_1 - M_2) &= E[(M_1 - M_2) - (\mu_1 - \mu_2)]^2 \\ &= E[(M_1 - \mu_1) - (M_2 - \mu_2)]^2. \end{aligned}$$

For any given pair of samples, expanding the square gives

$$[(M_1 - \mu_1) - (M_2 - \mu_2)]^2 = (M_1 - \mu_1)^2 + (M_2 - \mu_2)^2 - 2(M_1 - \mu_1)(M_2 - \mu_2).$$

Now let us take the expectation of each of these terms separately:

$$E(M_1 - \mu_1)^2 = \sigma^2_{M_1}$$

and

$$E(M_2 - \mu_2)^2 = \sigma^2_{M_2}$$

by the definition of the variance of a sampling distribution of the mean. Furthermore,

$$E[(M_1 - \mu_1)(M_2 - \mu_2)] = 0$$

by rule 6, Appendix B, since M_1 and M_2 are independent. Thus, combining these results, we find that

$$\text{var.}(M_1 - M_2) = \sigma^2_{M_1} + \sigma^2_{M_2}$$

or

$$\sigma_{\text{diff.}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}.$$

Notice that there is no requirement at all that the samples be of equal size. Regardless of the sample sizes, the expectation of the difference between two means is always the difference between their expectations, and the variance of the difference between two *independent* means is the *sum* of the separate sampling variances.

Furthermore, these statements about the mean and the standard error of a difference between means are true regardless of the form of the parent distributions. However, the form of the sampling distribution can also be specified under either of two conditions:

If the distribution for each of two populations is normal then the distribution of differences between sample means is normal.

This follows quite simply from principle 8.9 for linear combinations. When we can assume both populations normal, the form of the sampling distribution is known to be *exactly* normal.

On the other hand, one or both of the original distributions may not be normal; in this case the central limit theorem comes to our aid:

As both N_1 and N_2 grow infinitely large, the sampling distribution of the difference between means approaches a normal distribution, regardless of the form of the original distributions.

In short, when we are dealing with two very large samples, then the question of the form of the original distributions becomes irrelevant, and we can approximate the sampling distribution of the difference between means by a normal distribution.

10.13 An example of a large-sample significance test for a difference between means

An experimenter working in the area of motivational factors in perception was interested in the effects of deprivation upon the perceived size of

objects. Among the studies carried out was one done with orphans, who were compared with nonorphaned children on the basis of the judged size of parental figures viewed at a distance. Each child was seated at a viewing apparatus in which cut-out figures appeared. Each figure was actually of the same size and at the same distance from the viewer, although he was not told that the figures had the same size. A device was provided on which the child could actually judge the apparent sizes of the different figures in numerical terms. Several of the figures in the set viewed were obviously parents, whereas others were more or less neutral, such as milkmen, postmen, nurses, and so on. Each child was given a score, which was itself a difference in average judged size of parental and nonparental figures.

Now two independent randomly selected groups were used. Sample 1 was a group of orphaned children without foster parents. Sample 2 was a group of children having a normal family with both parents. Both populations of children sampled showed the same age level, sex distribution, educational level, and so forth.

The question asked by the experimenter was, "Do deprived children tend to judge the parental figures relatively larger than do the nondeprived?" In terms of a null and alternative hypothesis,

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_1: \mu_1 - \mu_2 > 0.$$

The α level for significance decided upon was .05. The actual results were

Sample 1	*Sample 2*
$M_1 = 1.8$	$M_2 = 1.6$
$s_1 = .7$	$s_2 = .9$
$N_1 = 125$	$N_2 = 150$

These sample sizes are rather large, and the experimenter felt safe in using the normal approximation to the sampling distribution, even though he had no idea about the distribution form for the two populations sampled. The t ratio used was

$$t = \frac{(M_1 - M_2) - E(M_1 - M_2)}{\text{est. } \sigma_{\text{diff.}}} \qquad [10.13.1^*]$$

In this problem, $E(M_1 - M_2) = 0$, under the hypothesis tested. It was obviously necessary for the experimenter to estimate the standard error of the difference, since both σ_1 and σ_2 were unknown to him. This estimate was found by first estimating $\sigma^2_{M_1}$ and $\sigma^2_{M_2}$:

$$\text{est. } \sigma^2_{M_1} = \frac{s_1^2}{N_1} = \frac{.49}{125} = .004 \qquad [10.13.2\dagger]$$

$$\text{est. } \sigma^2_{M_2} = \frac{s_2^2}{N_2} = \frac{.81}{150} = .005. \qquad [10.13.3\dagger]$$

Then,

$$\text{est. } \sigma_{\text{diff.}} = \sqrt{\text{est. } \sigma^2_{M_1} + \text{est. } \sigma^2_{M_2}} = \sqrt{.004 + .005} = .095. \qquad [10.13.4\dagger]$$

On making these substitutions, the experimenter found

$$t = \frac{1.8 - 1.6}{.095} = 2.11.$$

The rejection region implied by the alternative hypothesis is on the *upper* tail of the sampling distribution. For a normal distribution the upper 5 percent is bounded by $z = 1.65$. Thus, the result is significant; deviations this far from zero have a probability of less than .05 of occurring by chance alone when the true difference is zero.

The experimenter may conclude that a difference exists between these two populations, *if* an α value less than .05 is a small enough probability of error to warrant this decision. However, the experimenter does not necessarily conclude that parental deprivation causes an increase in perceived size. The statistical conclusion suggests that it *might* be safe to assert that a particular direction of numerical difference exists between the mean scores of the two populations of children, but the statistical result is absolutely noncommittal about the reason for this difference, if such exists. The experimenter takes the step of advancing a reason at his own peril. The statistical test as a mathematical tool is absolutely neutral about what these numbers measure, the level of measurement, what was or was not represented in the experiment, and, most of all the cause of the experimenter's particular finding. As always, the test takes the numerical values as given, and cranks out a conclusion about the conditional probability of such numbers, given certain statistical conditions.

The general procedure for hypotheses about two means when sample size is quite large is represented by this example. The test statistic is

$$t = \frac{(M_1 - M_2) - E(M_1 - M_2)}{\text{est. } \sigma_{\text{diff.}}}.$$

This t value may be referred to a normal distribution. The expected difference depends upon the hypothesis tested, and the estimated $\sigma_{\text{diff.}}$ is found directly from the estimate σ_M^2 for each sample by **10.13.4.**

The exact hypothesis actually tested is of the form

$$H_0\colon \mu_1 - \mu_2 = k,$$

where k is any difference of interest. Quite often, as in the example, the experimenter is interested only in $k = 0$, but it is entirely possible to test any other meaningful difference value. The alternative hypothesis may be directional,

$$H_1\colon \mu_1 - \mu_2 > k$$

or

$$H_1\colon \mu_1 - \mu_2 < k,$$

or nondirectional,

$$H_1\colon \mu_1 - \mu_2 \neq k,$$

depending on the form of the original question.

As an illustration of a situation where some value other than zero figures in the null hypothesis, and also as an illustration of a one-tailed test, take the following example: a manufacturer is considering introducing a change in training procedure for his new employees. However, it is more expensive than the

old, and he feels that he cannot afford it unless the average output of a man trained in the new way is more than 50 units per hour better than that of a man trained under the old procedure. The null hypothesis is

$$H_0: \mu_1 - \mu_2 \leq 50,$$

since the exact value that the null hypothesis requires is given by 50 units per hour. The alternative hypothesis states

$$H_1: \mu_1 - \mu_2 > 50.$$

Notice how the null and the alternative hypotheses are framed so as to correspond to the alternative practical decisions that our manufacturer may make: if the null hypothesis is true, he will not adopt the new training procedure since it does not meet the requirement he set up. He has no interest in the training procedure if it is less than 50 units better than the old. If, however, the alternative hypothesis is true, then he will adopt the new procedure. In this instance, where clear-cut courses of action depend on the evidence, the one-tailed test of a nonzero hypothesis makes sense. Subjects are assigned at random to two groups, one getting the new and the other the old training. Given large samples, the t ratio is computed just as in the previous example, except that here $E(M_1 - M_2) = 50$. A significant result gives the manufacturer considerable assurance in saying that one procedure is on the average more than 50 units better than the other.

10.14 Large-sample confidence limits for a difference

When both samples are large, as in the example in Section 10.13, confidence limits are found exactly as for a single mean, except that $(M_1 - M_2)$ and est. $\sigma_{\text{diff.}}$ are substituted for M and est. σ_M respectively. Thus, 95 percent confidence limits for a difference with large samples are

$$\begin{aligned} &M_1 - M_2 - 1.96 \text{ (est. } \sigma_{\text{diff.}}) \\ &M_1 - M_2 + 1.96 \text{ (est. } \sigma_{\text{diff.}}). \end{aligned} \qquad [10.14.1^*]$$

For the example in the preceding section, the 95 percent limits are

$$\begin{aligned} &.2 - 1.96(.095) \\ &.2 + 1.96(.095) \end{aligned}$$

or .014 and .386. Notice that since the value $\mu_1 - \mu_2 = 0$ does not fall within these values this value can be rejected as a hypothesis beyond the .05 level (two-tailed).

10.15 Using the t distribution to test hypotheses about differences

Given the assumption that both populations sampled have normal distributions, any hypothesis about a difference can be tested using the t distribu-

tion, regardless of sample size. However, one additional assumption becomes necessary: *in order to use the t distribution for tests based on two (or more) samples, one must assume that the standard deviations of both (or all) populations are equal.* The basis for this assumption will be discussed in the next chapter.

Given these assumptions, then the distribution of t for a difference has the same form as for a single mean, except that the degrees of freedom are

$$\nu = N_1 - 1 + N_2 - 1 = N_1 + N_2 - 2.$$

When samples are drawn from populations with equal variance, then the estimated standard error of a difference takes a somewhat different form. First of all, when $\sigma_1 = \sigma_2 = \sigma$,

$$\sigma_{\text{diff.}} = \sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}} = \sqrt{\sigma^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}. \qquad [10.15.1^*]$$

Now, as we showed in Section 7.17, when one has two or more estimates of the same parameter σ^2, the *pooled* estimate is actually better than either one taken separately. From 7.17.5 it follows that

$$\text{est. } \sigma^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

is our best estimate of σ^2 based on the two samples. Hence

$$\begin{aligned}\text{est. } \sigma_{\text{diff.}} &= \sqrt{\text{est. } \sigma^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \\ &= \sqrt{\left(\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)} \qquad [10.15.2^\dagger]\end{aligned}$$

This estimate of the standard error of the difference ordinarily forms the denominator of the t ratio when the t distribution is used for hypotheses about a difference.

10.16 An example of inferences about a difference for small samples

Two random samples of subjects are being compared on the basis of their scores on a motor learning task. The subjects are allotted to two experimental groups, with five subjects in the first and seven in the second. In the first group a subject is rewarded for each correct move made, and in the second each incorrect move is punished. The score is the number of trials to reach a specific criterion of performance. The experimenter wishes to find evidence for the question, "Does the kind of motivation employed, reward or punishment, affect the performance?" This question implies the null and alternative hypotheses:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0.$$

The experimenter is willing to assume that the population distributions of scores

are normal, and that the population variances are equal. The probability of Type I error decided upon is .01. Since this is a two-tailed test, a glance at Table II shows that for $N_1 + N_2 - 2$ or $5 + 7 - 2 = 10$ degrees of freedom, and for $2Q = .01$, the required t value is 3.169. Thus an obtained t ratio equaling or exceeding 3.169 in absolute value is grounds for rejecting the hypothesis of no difference between population means.

The sample results are

$$M_1 = 18 \qquad M_2 = 20$$
$$s_1^2 = 6.00 \qquad s_2^2 = 5.83$$

The estimated standard error of the difference is found by the pooling procedure given in the last section:

$$\begin{aligned}\text{est. } \sigma_{\text{diff.}} &= \sqrt{\text{est. } \sigma^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \\ &= \sqrt{\frac{(4)(6) + (6)(5.83)}{10}\left(\frac{12}{35}\right)} \\ &= \sqrt{2.02} \\ &= 1.42.\end{aligned}$$

Thus, the t ratio is

$$t = \frac{(M_1 - M_2) - E(M_1 - M_2)}{\text{est. } \sigma_{\text{diff.}}} = \frac{-2}{1.42} = -1.41.$$

This value comes nowhere close to that required for rejection, and thus if α must be no more than .01 the experimenter does not reject the null hypothesis. His best choice may be to suspend judgement, pending more evidence.

Confidence intervals are found just as for a single small sample mean: the limits are

$$\begin{aligned}&(M_1 - M_2) - t_{(\alpha/2;\,\nu)}\,(\text{est. } \sigma_{\text{diff.}}) \\ &(M_1 - M_2) + t_{(\alpha/2;\,\nu)}\,(\text{est. } \sigma_{\text{diff.}}).\end{aligned} \qquad [10.16.1^*]$$

For this example, the 99 percent limits are

$$-2 - (3.169)(1.42)$$

and

$$-2 + (3.169)(1.42)$$

or approximately -6.5 and 2.5. The probability is .99 that the true *difference*, $\mu_1 - \mu_2$, is covered by an interval such as this. Once again, notice that this interval *does* contain the value 0, indicating the hypothesis entertained above is not rejected.

10.17 The Importance of the Assumptions in a t Test of a Difference

In order to justify the use of the t distribution in problems involving a difference between means, one must make two assumptions: the populations

sampled are normal, and the population variances are homogeneous, σ^2 having the same value for each population. Formally, these two assumptions are essential if the t probabilities given by the table are to be exact. On the other hand, in practical situations these assumptions are sometimes violated with rather small effect on the conclusions.

The first assumption, that of a normal distribution in the populations, is apparently the less important of the two. So long as the sample size is even moderate for each group quite severe departures from normality seem to make little practical difference in the conclusions reached. Naturally, the results are more accurate the more nearly unimodal and symmetric the population distributions are, and thus if one suspects radical departures from a generally normal form then he should plan on larger samples. Furthermore, the departure from normality can make more difference in a one-tailed than in a two-tailed result, and once again some special thought should be given to sample size when one-tailed tests are contemplated for such populations. By and large, however, this assumption may be violated almost with impunity provided that sample size is not extremely small.

On the other hand, the assumption of homogeneity of variance is more important. In older work it was often suggested that a separate test for homogenity of variance be carried out before the t test itself, in order to see if this assumption were at all reasonable. However, the most modern authorities suggest that this is not really worth the trouble involved. In circumstances where they are needed most (small samples), the tests for homogeneity are poorest. Furthermore, for samples of equal size relatively big differences in the population variances seem to have relatively small consequences for the conclusions derived from a t test. On the other hand, when the variances are quite unequal the use of different sample sizes can have serious effects on the conclusions. The moral should be plain: given the usual freedom about sample size in experimental work, *when in doubt use samples of the same size.*

However, sometimes it is not possible to obtain an equal number in each group. Then one way out of this problem is by the use of a correction in the value for degrees of freedom. This is useful when one cannot assume equal population variances and samples are of different size. In this situation, however, the t ratio is calculated as in Section 10.13, where the separate standard errors are computed from each sample and the pooled estimate is not made. Then the corrected number of degrees of freedom is found from

$$\nu = \frac{(\text{est. } \sigma^2_{M_1} + \text{est. } \sigma^2_{M_2})^2}{(\text{est. } \sigma^2_{M_1})^2/(N_1 + 1) + (\text{est. } \sigma^2_{M_2})^2/(N_2 + 1)} - 2. \qquad [10.17.1\dagger]$$

This need not result in a whole value for ν, in which case the use of the nearest whole value for ν is sufficiently accurate for most purposes. When somewhat greater accuracy is desired, the approximate formula for critical values of t given in Section 14.17 is useful. When both samples are quite large, then both the assumptions of normality and of homogeneous variances become relatively unimportant, and the method of Section 10.13 can be used.

10.18 The power of t tests

The idea of the power of a statistical test was discussed in the preceding section only in terms of the normal distribution. Nevertheless, the same general considerations apply to the power of tests based on the t distribution. Thus, the power of a t test increases with sample size, increases with the discrepancy between the null hypothesis value and the true value of a mean or a difference, increases with any reduction in the true value of σ, and increases with any increase in the size of α, given a true value covered by H_1.

Unfortunately, the actual determination of the power for a t test against any given true alternative is more complicated than for the normal distribution. The reason is that when the null hypothesis is false, each t ratio computed involves $E(M)$ or $E(M_1 - M_2)$, which is the exact value given by the null (and false) hypothesis. If the true value of the expectation could be calculated into each t ratio, then the distribution would follow the t function tabled in the appendix. However, when H_0 is false, each t value involves a false expectation; this results in a somewhat different distribution, called the **noncentral t distribution.** The probabilities of the various t's cannot be known unless one more parameter, δ, is specified beside ν. This is the so-called noncentrality parameter, defined by

$$\delta^2 = \left(\frac{\mu - \mu_0}{\sigma_M}\right)^2. \qquad [10.18.1]$$

The parameter δ^2 expresses the squared difference between the true expectation μ and that given by the null hypothesis, or μ_0, in terms of σ_M. For a hypothesis about a difference and for samples of equal size,

$$\delta^2 = \left[\frac{(\mu_1 - \mu_2) - (\mu_{0_1} - \mu_{0_2})}{\sigma_{\text{diff}}}\right]^2. \qquad [10.18.2]$$

The value of the parameter δ is then the positive square root of δ^2.

The matter is made even more complex by the fact that a noncentral t distribution not only has an additional parameter that must be specified; the form of a noncentral t' distribution differs from that of a central t distribution. Hence, rather detailed tables become necessary for each pair of parameter values ν and δ if exact determinations of power are to be made. Such tables are provided in some advanced texts on statistics.

Fortunately, when great accuracy is not required, an approximation based upon the normal distribution can be used. This approximation, given by Scheffé (1959), provides the cumulative probability that the variable t' is less than or equal to some value x, given the noncentral distribution with parameters ν and δ. This is found by use of the expression

$$\Pr(t'_{(\nu,\delta)} \leq x) = \Pr\left\{z \leq (x - \delta)\left(1 + \frac{x^2}{2\nu}\right)^{-1/2}\right\},$$

where z is a value in a normal distribution with mean 0 and variance 1.00.

The use of this approximation can be demonstrated in terms of the preceding problem (Section 10.16). There, the null hypothesis was that of no

difference between the two population means. Let us determine the power of this test against the alternative that $\mu_1 - \mu_2 = 4$. That is, given that $\mu_1 - \mu_2 = 4$, what is the probability that the obtained t' value would fall outside the interval with limits -3.169 to 3.169?

We start off by calculating the value of the noncentrality parameter δ. We really need to know the true value of the standard error of the difference, $\sigma_{\text{diff.}}$, but in the absence of this information we will use the estimate from the samples. This was found to be 1.42. Then the value of δ corresponding to a difference of 4 is given by

$$\delta = \left| \frac{4}{\sigma_{\text{diff.}}} \right| = \frac{4}{1.42} = 2.816.$$

Then

$$\begin{aligned} \Pr(t'_{(\nu,\delta)} \leq 3.169) &= \Pr\left\{ z \leq (3.169 - 2.816) \left[1 + \frac{(3.169)^2}{2(10)} \right]^{-1/2} \right\} \\ &= \Pr\left(z \leq \frac{.353}{\sqrt{1.5}} \right) \\ &= \Pr(z \leq .288) \\ &= .614, \quad \text{approximately.} \end{aligned}$$

Thus, we have found that if the true difference between the means is 4, the probability of an obtained t' value less than 3.169 is approximately the same as the probability of a normal z value less than .288. This probability is about .614.

Since this is a two-tailed test, we must also consider the possibility of an obtained t value that is less than -3.169. Then the probability of a Type II error will be the probability that $t \leq 3.169$ minus the probability that $t \leq -3.169$ (i.e., the probability that t falls in the region of nonrejection for H_0, even though the true difference is 4). Hence we take

$$\begin{aligned} \Pr(t'_{(\nu,\delta)} \leq -3.169) &= \Pr\left\{ z \leq \frac{(-3.169 - 2.816)}{\sqrt{1.5}} \right\} \\ &= \Pr(z \leq -4.88). \end{aligned}$$

This probability is virtually zero in a normal distribution. We then take the probability that $-3.169 \leq t' \leq 3.169$ to be approximately .614, and this is the probability of a Type II error when $\mu_1 - \mu_2 = 4$. The power of the t test against this alternative is then approximately $1 - .614$ or .386. If we desired, we could keep applying this method and construct the entire power function of the test for the various alternatives to the null hypothesis.

It should be kept in mind that this method depends upon the usual assumptions underlying the use of a t distribution being satisfied. That is, one still assumes that the parent distributions underlying the data are normal, that the observations are made independently and at random, and that, if two distributions are involved, each has the same variance. The noncentral variable t' differs from the central t variable only in that its distribution depends upon the new parameter δ. All of the other requirements for the use of a t distribution must be met.

10.19 Testmanship, or how big is a difference?

When an experimenter assigns subjects at random to two experimental groups, giving a different treatment to subjects in each group, he is usually looking for evidence of a statistical relation. Here, the independent variable represents the various experimental treatments and the dependent variable is the score of any subject within a group. Each treatment group is a random sample of all potential subjects given that treatment. The sample space is conceived as the set of all possible treatment-subject combinations, and the statistical relation itself is defined in terms of this sample space.

As we saw in Chapter 4, the complete absence of a statistical relation, or no association, occurs only when the conditional distribution of the dependent variable is the same regardless of which treatment is administered. Thus if the independent variable is not associated at all with the dependent variable the population distributions must be identical over the treatments. If, on the other hand, the means of the different treatment populations *are* different, the conditional distributions themselves must be different and the independent and dependent variables must be associated. The rejection of the hypothesis of no difference between population means is tantamount to the assertion that the treatment given does have some statistical association with the dependent variable score.

However, the occurrence of a significant result says nothing at all about the strength of the association between treatment and score. A significant result leads to the inference that some association exists, but in no sense does this mean that an important degree of association necessarily exists. Conversely, evidence of a strong statistical association can occur in data even when the results are not significant. The game of inferring the true degree of statistical association has a joker: this is the sample size. The time has come to define the notion of the strength of a statistical association more sharply, and to link this idea with that of the true difference between population means.

Just as in our discussion of relations in Chapters 1 and 4, let us call the experimental variable (or the independent variable) X once again. Here, X may symbolize a number standing for a quantity of some treatment, or it may simply represent any one of a set of qualitatively different treatments. In either circumstance, X stands for the status of the individual observation on the experimental factor, the condition manipulated by the experimenter. The dependent variable is Y, which here stands for a numerical score. If we conceive the sample space as comprising the outcomes of our observing all of a population of individuals under each of the possible set of treatments X, then each possible observation in the experiment is some (x, y) event. Furthermore, if individuals from the population of potential subjects are sampled at random, and assigned at random to the various possible treatments X in the experiment, then the occurrence of any individual in the treatment x has a probability $p(x)$. For our purposes, it will be convenient to assume that

$$p(x) = \frac{\text{number of individuals observed under treatment } x}{\text{total number of individuals observed}}.$$

When does it seem appropriate to say that a strong association exists between the experimental factor X and the dependent variable Y? Over all of the different possibilities for X there is a probability distribution of Y values, which is the **marginal** distribution of Y over (x,y) events. The existence of this distribution implies that we do not know exactly what the Y value for any observation will be; we are always uncertain about Y to some extent. However, given any particular X, there is also a **conditional** distribution of Y, and it may be that in this conditional distribution the highly probable values of Y tend to "shrink" within a much narrower range than in the marginal distribution. If so, we can say that the information about X tends to *reduce uncertainty* about Y. *In general we will say that the strength of a statistical relation is reflected by the extent to which knowing X reduces uncertainty about Y.*

One of the best indicators of our uncertainty about the value of a variable is σ^2, the variance of its distribution. The marginal distribution of Y has variance σ^2_Y, and given any X, the conditional distribution has variance $\sigma^2_{Y|X}$. For the time being, let us assume that $\sigma^2_{Y|X}$ is the same regardless of which X we specify. This is exactly the assumption of equal variances made in the t test, since each population distribution is actually a conditional distribution, given some treatment specification. The reduction in uncertainty provided by X is then proportional to

$$\sigma^2_Y - \sigma^2_{Y|X}, \qquad \text{[10.19.1*]}$$

the difference between the marginal and the conditional variance of Y.

It is convenient to turn this reduction in uncertainty into a **relative reduction** by dividing by σ^2_Y, giving

$$\omega^2 = \frac{\sigma^2_Y - \sigma^2_{Y|X}}{\sigma^2_Y}. \qquad \text{[10.19.2*]}$$

The relative reduction in uncertainty about Y given by X is shown by the index ω^2 (Greek omega, squared). Sometimes the value ω^2 is called **the proportion of variance in Y accounted for by X.** Viewed either as a relative reduction in uncertainty, or as a proportion of variance accounted for, the index ω^2 represents the strength of association between independent and dependent variables. (The index ω^2 is almost identical to two other indices to be introduced later, *the intraclass correlation* and the *correlation ratio*, usually represented by the symbols ρ_I and η^2 respectively. However, since these indices were developed for and are used in somewhat different contexts, it seems better to use the relatively neutral symbol ω^2 here, to avoid later confusion.)

This index reflects the predictive power afforded by a relationship: when ω^2 is zero, then X does not aid us at all in predicting the value of Y. On the other hand, when ω^2 is 1.00, this tells us that X lets us know Y exactly. All intermediate values of the index represent different degrees of predictive ability. Notice that for any functional relation, $\omega^2 = 1.00$, since there can be only one Y for each possible X. A value less than unity tells us that precise prediction is not possible, although X nevertheless gives *some* information about Y unless $\omega^2 = 0$.

About now you should be wondering what the index ω^2 has to do

with the difference between population means. It can be shown, by methods we shall use in Chapter 12, that when $p(x_1) = p(x_2) = 1/2$,

$$\sigma_Y^2 = \sigma_{Y|X}^2 + \frac{(\mu_1 - \mu_2)^2}{4} \qquad [10.19.3^*]$$

where μ_1 is the mean of population 1, μ_2 that of population 2, and

$$\frac{\mu_1 + \mu_2}{2} = \mu,$$

the mean of the marginal distribution.

On substituting into 10.19.2, we find

$$\omega^2 = \frac{(\mu_1 - \mu_2)^2}{4\sigma_Y^2}.$$

For two treatment-populations with equal variances the strength of the statistical association between treatment and dependent variable varies directly with the squared difference between the population means, relative to the unconditional, marginal, variance of Y.

When the difference $\mu_1 - \mu_2$ is zero, then ω^2 must be zero. In the usual t test for a difference, the hypothesis of no difference between means is equivalent to the hypothesis that $\omega^2 = 0$. On the other hand, when there is any difference at all between population means, the value of ω^2 must be greater than 0. In short, a true difference is "big" in the sense of predictive power only if the square of that difference is large relative to σ_Y^2. However, in significance tests such as t, we compare the difference we get with an estimate of $\sigma_{diff.}$. The standard error of the difference can be made almost as small as we choose if we are given a free choice of sample size. Unless sample size is specified, there is no *necessary* connection between significance and the true strength of association.

This points up the fallacy of evaluating the "goodness" of a result in terms of statistical significance alone, without allowing for the sample size used. All significant results do not imply the same degree of true association between independent and dependent variables.

It is sad but true that researchers have been known to capitalize on this fact. There is a certain amount of "testmanship" involved in using inferential statistics. *Virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be.* There is surely nothing on earth that is completely independent of anything else. The strength of an association may approach zero, but it should seldom or never be exactly zero. If one applies a large enough sample of the study of any relation, trivial or meaningless as it may be, sooner or later he is almost certain to achieve a significant result. Such a result may be a valid finding, but only in the sense that one can say with assurance that some association is not exactly zero. The degree to which such a finding enhances our knowledge is debatable. If the criterion of strength of association is applied to such a result, it becomes obvious that little or nothing is actually contributed to our ability to predict one thing from another.

For example, suppose that two methods of teaching first grade children to read are being compared. A random sample of 1000 children are taught to read by method I, another sample of 1000 children by method II. The results of the instruction are evaluated by a test that provides a score, in whole units, for each child. Suppose that the results turned out as follows:

Method I	*Method II*
$M_1 = 147.21$	$M_2 = 147.64$
$s_1^2 = 10$	$s_2^2 = 11$
$N_1 = 1000$	$N_2 = 1000$

Then, the estimated standard error of the difference is about .145, and the z value is

$$z = \frac{147.21 - 147.64}{.145} = -2.96.$$

This certainly permits rejection of the null hypothesis of no difference between the groups. However, does it really tell us very much about what to expect of an individual child's score on the test, given the information that he was taught by method I or method II? If we look at the group of children taught by method II, and assume that the distribution of their scores is approximately normal, we find that about 45 percent of these children fall *below* the mean score for children in group I. Similarly, about 45 percent of children in group I fall above the mean score for group II. Although the difference between the two groups is significant, the two groups actually overlap a great deal in terms of their performances on the test. In this sense, the two groups are really not very different at all, even though the difference between the means is quite significant in a purely statistical sense.

Putting the matter in a slightly different way, we note that the grand mean of the two groups is 147.425. Thus, our best bet about the score of any child, not knowing the method of his training, is 147.425. If we guessed that any child drawn at random from the combined group should have a score above 147.425, we should be wrong about half the time. However, among the original groups, according to method I and method II, the proportions falling above and below this grand mean are approximately as follows:

	Below 147.425	*Above 147.425*
Method I	.51	.49
Method II	.49	.51

This implies that if we know a child is from group I, and we guess that his score is below the grand mean, then we will be wrong about 49 percent of the time. Similarly, if a child is from group II, and we guess his score to be above the grand mean, we will be wrong about 49 percent of the time. If we are not given the group to which the child belongs, and we guess either above or below the

grand mean, we will be wrong about 50 percent of the time. Knowing the group does reduce the probability of error in such a guess, but it does not reduce it very much. The method by which the child was trained simply doesn't tell us a great deal about what the child's score will be, even though the difference in mean scores is significant in the statistical sense.

This kind of testmanship flourishes best when people pay too much attention to the significance test and too little to the degree of statistical association the finding represents. This clutters up the literature with findings that are often not worth pursuing, and which serve only to obscure the really important predictive relations that occasionally appear. The serious scientist owes it to himself and his readers to ask not only, "Is there any association between X and Y?" but also, "How much does my finding suggest about the power to predict Y from X?" Much too much emphasis is paid to the former, at the expense of the latter, question.

10.20 Estimating the strength of a statistical association from data

It is quite possible to estimate the amount of statistical association implied by any obtained difference between means. The ingredients for this kind of estimation are essentially those used in a t test. The problems connected with the sampling distribution of this estimate will be deferred until Chapter 12, and for the moment we shall consider only how this estimate is made and used.

(A number of ways have been proposed for estimating the strength of a statistical association from obtained differences between means. For reasons to be elaborated later, none of these methods is entirely satisfactory. The method to be introduced here is thus only one of the ways that may be encountered in the statistical literature, but it seems to have as much to recommend it as any other.)

For samples from two populations, each of which has the same true variance, $\sigma^2_{Y|X}$, a rough estimate of ω^2 is provided by

$$\text{est. } \omega^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}. \qquad [10.20.1]$$

(A more general form for estimating ω^2 will be given in Chapter 12.) Notice that if t^2 is less than 1.00, then this estimate is negative, although ω^2 cannot assume negative values. In this situation the estimate of ω^2 is set equal to zero.

Let us consider an example using this estimate. Imagine a study involving two groups of 30 cases each. Subjects are assigned at random to these two groups, and each set of subjects is given a different treatment. The results are

Group 1	*Group 2*
$M_1 = 65.5$	$M_2 = 69$
$s_1^2 = 20.69$	$s_2^2 = 28.96$
$N_1 = 30$	$N_2 = 30$

First of all the t ratio is computed in the usual way (Section 10.16):

$$\text{est. } \sigma^2 = \frac{(29)(20.69 + 28.96)}{58} = 24.83$$

and

$$\text{est. } \sigma_{\text{diff.}} = \sqrt{\frac{24.83(2)}{30}} = 1.29.$$

Thus,

$$t = \frac{65.5 - 69}{1.29} = -2.71.$$

For a two-tailed test with 58 degrees of freedom, this value is significant beyond the .01 level. Thus, we are fairly safe in concluding that some association exists.

What do we estimate the true degree of association to be? Substituting into **10.20.1**, we find

$$\text{est. } \omega^2 = \frac{(2.71)^2 - 1}{(2.71)^2 + (60 - 1)} = .096.$$

Our rough estimate is that X (the treatment administered) accounts for about 10 percent of the variance of Y (the obtained score).

Suppose, however, that the groups had contained only 10 cases each, and that the results had been:

Group 1	*Group 2*
$M_1 = 65.5$	$M_2 = 69$
$s_1^2 = 5.55$	$s_2^2 = 7.78$
$N_1 = 10$	$N_2 = 10$

Here

$$\text{est. } \sigma^2 = \frac{9(5.55 + 7.78)}{18} = 6.67$$

$$\text{est. } \sigma_{\text{diff.}} = \sqrt{6.67\left(\frac{2}{10}\right)} = 1.15$$

so that

$$t = \frac{-3.5}{1.15} = -3.04.$$

For 18 degrees of freedom, this value is also significant beyond the .01 level (two-tailed), and once again we can assert with confidence that some association exists.

Again, we estimate the degree of association represented by this finding:

$$\text{est. } \omega^2 = \frac{(3.04)^2 - 1}{(3.04)^2 + 19} = .29.$$

Here, our rough estimate is that X accounts for about 29 percent of the variance in Y. Even though the difference between the sample means is the same in these two examples, and both results are significant beyond the .01 level, the second experiment gives a much higher estimate of the true association than the first.

The point of this discussion should be evident by now: statistical significance is not the only, or even the best, evidence for a strong statistical association. A significant result implies that it is safe to say some association exists, but the estimate of ω^2 tells how strong that association appears to be. It

seems far more reasonable to decide to follow up a finding that is *both* significant *and* indicates a strong degree of association than to tie this course of action to significance level alone. Conversely, when a result fails to attain significance and there is no ready way to estimate the β probability, the experimenter really has at least two courses of action: he can suspend judgment temporarily and actually collect more data, or he can suspend judgment permanently by forgetting the whole business. If the estimated strength of association is relatively small it may not be worthwhile to spend more time and effort in this direction. Regardless of the courses of action open to the experimenter, on the whole it is reasonable that a better decision can be made in terms of both significance level and estimated strength of relation than by either taken alone. In most experimental problems we want to find and refine relationships that "pay off," that actually increase our ability to predict behavior. When the results of an experiment suggest that the strength of an association is very low, then perhaps the experimenter should ask himself whether this matter is worth pursuing after all, regardless of the statistical significance he may attain by increased sample size or other refinements of the experiment.

10.21 Strength of association and sample size

Of all the questions that psychologists carry to statisticians, surely the most frequently heard is, "How many subjects do I need in this experiment?" The response of the statistician is very likely to be unsatisfactory, the gist being "How big is a difference that you consider important?" This is a question that can be answered only by the psychologist, and then only if he has given the matter some serious thought. If the experimenter cannot answer this question the statistician really cannot help him. Perhaps framing the essential point in terms of strength of a relation rather than the size of a difference will make it a little easier to grasp.

Basically, the question of sample size depends upon the strength of association the experimenter *wants* to detect as significant. Actually, this matter is properly discussed in terms of the t distribution, but for the kinds of rough determinations most psychologists need to make, the normal approximation will suffice.

Recall the basic definition of ω^2 in terms of two population means:

$$\omega^2 = \frac{(\mu_1 - \mu_2)^2}{4\sigma_Y^2}.$$

From this definition, we can derive the fact that

$$\frac{|\mu_1 - \mu_2|}{\sigma_{Y|X}} = 2\sqrt{\frac{\omega^2}{1 - \omega^2}} = \Delta. \qquad [10.21.1^*]$$

Given any value of ω^2, we can find the ratio of the absolute difference between population means to the standard deviation of either population. The symbol Δ

(capital Greek delta) will stand for this absolute difference between means in standard deviation units.

If we want to discuss the difference between means in units of the standard error of the difference, then for samples of the same size, n,

$$\frac{|\mu_1 - \mu_2|}{\sigma_{\text{diff.}}} = \Delta\sqrt{\frac{n}{2}}.$$

Now suppose that an experiment is being planned which involves two groups, each of size n. The experimenter wants to be very sure that he will detect a significant difference if the true degree of association ω^2 is k or more in value. How large should n be in each sample?

Several things must be specified: the value of $k = \omega^2$, the α probability, and the probability $1 - \beta$, which is the power of the test when the true degree of association is equal to k. Given these three specifications, then one can approximate the required size of n by taking

$$\sqrt{\frac{n}{2}} = \frac{[z_{(1-\alpha/2)} - z_{(\beta)}]}{\Delta} \qquad \text{[10.21.2†]}$$

or

$$n = \frac{2[z_{(1-\alpha/2)} - z_{(\beta)}]^2}{\Delta^2} \qquad \text{[10.21.3†]}$$

where $z_{(1-\alpha/2)}$ is the value of a standardized score in a normal distribution cutting off the lower $(1 - \alpha/2)$ proportion of cases, and $z_{(\beta)}$ is the standardized score cutting off the lower β proportion of cases. The value of Δ is found from the required value of ω^2 by substitution into **10.21.1**.

For example, suppose that the experimenter wants to be very sure to detect a true association when X actually accounts for 25 percent or more of the variance of Y, so that ω^2 is .25 or more. He wants the test to have a power of .99 when $\omega^2 = .25$, and he has already decided that α must be .01. How many cases should he include in each sample?

First of all, solving from **10.21.1** for Δ, we find

$$\Delta = 2\sqrt{\frac{.25}{1 - .25}}$$

or

$$\Delta = 1.15.$$

The value of $z_{(1-\alpha/2)}$ is 2.58, and that for $z_{(\beta)} = -2.33$. Thus,

$$n = \frac{2(2.58 + 2.33)^2}{(1.15)^2} = 36.5.$$

In order to have $\alpha = .01$, and to have a test with power of about .99 for a significant result when $\omega^2 = .25$, the experimenter should plan on about 37 subjects in each sample, a total of 74 subjects in all.

This may be more subjects than the experimenter can manage to obtain. He can reduce his estimate of the required number either by *lowering* his

requirements for the power of the test, making the power, say, .95 for $\omega^2 = .25$, or by *raising* the probability α to, say, .05. Suppose that he adopts the latter course, making $\alpha = .05$. In this instance, his revised estimate of sample size is given by

$$n = \frac{2(1.96 + 2.33)^2}{(1.15)^2} = 27.9$$

showing that these requirements are approximately satisfied if he takes around 28 subjects in each group, for a total of about 56 cases in all.

These estimates of required sample size are only approximate, since we use the normal rather than the t distribution. They are to be regarded only as rough guides to the general sample sizes required. Unless the sample size estimates turn out rather large as in the example, and if it is important that the experimenter fulfill the requirements he has set himself about ω^2, α, and $1 - \beta$, he is very wise to take samples somewhat larger than his estimate suggests.

It is remarkable how few studies reported in psychology seem to be based on sample sizes chosen in any systematic way. Certainly there are situations where real limits exist about how large a sample can be, and here the experimenter merely does the best he can. However, there is usually some freedom of choice within fairly broad limits. One does not have to look very far to find the reason this question is often ignored: all too seldom is the experimenter prepared to state the strength of association that he feels he *must* be sure to detect as a significant result. To decide this requires a great deal of thought about the potential applications, or the experimental follow-up that should be implied by a significant finding. On the other hand, unless this thought is expended in planning a study there is simply no way to determine required sample size. If psychologists are going to use conventions for deciding significance of results then perhaps a few conventions are called for about the strength of association it is desirable to *detect* as significant.

Regardless of the sample size actually chosen, the experimenter can form a rough estimate of the sensitivity of the experiment for detecting statistical association. For two relatively large samples of equal size, the expression

$$\Delta^2 = \frac{2[z_{(1-\alpha/2)}]^2}{n} \qquad [10.21.4^*]$$

can be solved for Δ^2. Then by the relation

$$\omega^2 = \frac{\Delta^2}{\Delta^2 + 4} \qquad [10.21.5^*]$$

one finds the strength of association for which the power is approximately .50. One can be reasonably sure that if the true degree of association is greater than the value of ω^2 found by this procedure, then he has better than a fifty-fifty chance of detecting this fact as a significant result. Conversely, if the true association is less than the value of ω^2 found the chances are about .50 or better that he will not detect this as a significant result.

For example, suppose that 25 subjects are used in each of two experimental groups. The α level chosen is .01, two-tailed. Then

$$\Delta^2 = \frac{2(2.58)^2}{25} = .53$$

so that

$$\omega^2 = \frac{.53}{.53 + 4} = .12.$$

The experimenter can say that if the true degree of association is about .12 or more he has at least a fifty-fifty chance of detecting this as a significant result.

Had the experimenter used 100 subjects per group, then the value of ω^2 in **10.21.4** would have been about .03. For this relatively large sample size the experimenter has about a fifty-fifty chance of detecting a significant difference when the experimental variable X accounts for no more than about 3 percent of the variance of Y. The larger the sample size, the smaller the proportion of variance accounted for that we can safely expect to be detected as a significant result.

The discussion of sample size in this section has been conducted in terms of two-tailed tests, since these are most comon in psychological research. However, the same idea applies in approximating the required sample size for a one-tailed test as well. Instead of using $z_{(1-\alpha/2)}$ in the computations, one simply substitutes $z_{(1-\alpha)}$ where α is the chosen error probability in the one-tailed region. For one-tailed tests the statements made about degrees of association and their detection are valid only for differences in the direction of the region of rejection, of course.

10.22 Can a sample size be too large?

In one sense, even posing this question sounds like heresy! Psychologists are often trained to think that large samples are *good things*, and we have seen that the most elegant features within theoretical statistics actually are the limit theorems, each implying a connection between sample size and the goodness of inferences.

Nevertheless, it seems reasonable that sample size can never really be discussed apart from what the experimenter is trying to do, and the stakes that he has in the experiment. *So long as the experimenter's primary interest is in precise estimation, then the larger the sample the better.* When he wants to come as close as he possibly can to the true parameter values, he can always do better by increasing sample size.

This is not, however, the main purpose of some experiments. These studies are, in the strict sense, exploratory. The experimenter is trying to map out the main relationships in some area. His study serves as a guide for directions that he will pursue in further, more refined, studies. He wants to find those statistical associations that are relatively large and that give considerable promise that a more or less precise relationship is there to be discovered and refined. He does not want to waste his time and effort by concluding an association exists when the degree of prediction actually afforded by that association is negligible. In short,

the experimenter would like a significant result to represent not only a nonzero association, but an association of considerable size.

When this is the situation it is advisable to look into the effects of sample size on the probability of finding a significant result given a *weak* association. For example, an experimenter has decided to use 30 subjects in each of two experimental groups. However, he does not want to waste his time with a significant result when the true degree of association is .01 or less. He decides that he wants the probability of a significant result to be .05 or less when the true ω^2 is .01 or less. He has already decided that α must be .01, and he knows that this must be the probability of a significant result when $\omega^2 = 0$ (the true difference is zero). For a two-tailed test, he cannot make the power of the test less than α. However, the experimenter's requirement is that his test have power of only .05 or less when ω^2 is less than or equal to .01.

Using 10.21.1 from the previous section,

$$\Delta = 2\sqrt{\frac{.01}{1 - .01}}$$
$$= 2(.10) = .2, \text{ approximately.}$$

In this problem, $1 - \beta = .05$, and so, by 10.21.3,

$$n = \frac{2(2.58 - 1.65)^2}{(.2)^2}$$
$$= 43.$$

If he uses *no more* than about 43 subjects in each group then the experimenter can be quite sure that a significant result is not likely to occur when the true degree of association is .01 or less. However, if he uses more subjects, he cannot be this confident of *not* detecting a very small association.

What does setting maximum sample size at 43 dictate about the ω^2 values he *will* detect as significant? How large a ω^2 will he detect as a significant result 95 percent of the time? This is found from

$$\Delta^2 = \frac{2(2.58 + 1.65)^2}{43}$$
$$= .83$$

so that

$$\omega_1^2 = \frac{(.83)^2}{(.83)^2 + 4}$$
$$= .15.$$

Even with the sample size of 43 cases per group, the experimenter knows that there is a probability of about .95 of finding a significant result when the true proportion of variance accounted for is as small as .15. This is not necessarily a negligible degree of association, and in some contexts it may be very important to account for as much as .15 proportion of variance. Nevertheless, in this instance the experimenter is interested in large degrees of association, and is content to rule out of consideration proportions of variance accounted for as small or smaller than .15.

Trivial associations may well show up as significant results when the sample size is very large. If the experimenter wants significance to be very likely to reflect a sizable association in his data, and also wants to be sure that he will not be led by a significant result into some blind alley, then he should pay attention to both aspects of sample size. Is the sample size *large* enough to give confidence that the big associations will indeed show up, while being *small* enough so that trivial associations will be excluded from significance?

10.23 Paired observations

Sometimes it happens that subjects are actually sampled in pairs. Even though each subject is experimentally different in one respect (nominally, the independent variable) from his pair-mate and each has some distinct dependent variable score, the scores of the members of a pair are not necessarily independent. For instance, one may be comparing scores of husbands and wives; a husband is "naturally" matched with his wife, and it makes sense that knowing the husband's score gives us some information about his wife's, and vice versa. Or individuals may be matched on some basis by the experimenter, and within each matched pair the members are assigned at random to experimental treatments. This matching of pairs is one form of experimental control, since each member of each experimental group must be identical (or nearly so) to his pair-mate in the other group with respect to the matching factor or factors, and thus the factor or factors used to match pairs is less likely to be responsible for any observed difference in the groups than if two unmatched groups are used.

Given two groups matched in this pairwise way, either by the experimenter or otherwise, it is still true that the difference between the means is an unbiased estimate of the population difference (in two matched populations):

$$E(M_1 - M_2) = \mu_1 - \mu_2.$$

However, the matching, and the consequent *dependence* within the pairs, changes the standard error of the difference. This can be shown quite simply: By definition, the variance of the difference between two sample means is

$$\sigma^2_{\text{diff.}} = E(M_1 - M_2 - \mu_1 + \mu_2)^2$$

which is the same as

$$E[(M_1 - \mu_1) - (M_2 - \mu_2)]^2.$$

Expanding the square, we have

$$E(M_1 - \mu_1)^2 + E(M_2 - \mu_2)^2 - 2E(M_1 - \mu_1)(M_2 - \mu_2).$$

The first of these terms is just $\sigma^2_{M_1}$, and the second is $\sigma^2_{M_2}$. However, what of the third term? From rule 6 Appendix B we find that the expectation of this product must be zero when the variables are independent. On the other hand, when variables are dependent the expectation is *not* ordinarily zero. Let us denote this last

term above as cov.(M_1,M_2), the **covariance** of the means. Then, for matched groups,

$$\sigma^2_{\text{diff.}} = \sigma^2_{M_1} + \sigma^2_{M_2} - 2 \text{ cov.}(M_1,M_2).$$

In general, for groups matched by pairs, this covariance is a positive number, and thus the variance and standard error of a difference between means will usually be *less* for matched than for unmatched groups. This fact accords with the experimenter's purpose in matching in the first place: to remove one or more sources of variability, and thus to lower the sampling error.

On the other hand, some caution must be exercised in this matching process. In the first place, it can be true that the factor on which subject-pairs are matched is such that the means are *negatively* related. Thus, for example, suppose that one had an effective measure of the dominance of personality of an individual. It just might be that highly dominant women tend to marry men with low dominance, and vice versa, so that among husband-wife pairs, dominance scores are negatively related. Then, if our interest is basically that of comparing men and women generally on such scores, it would be a mistake to match, since the negative relationship would lead to a larger, rather than a smaller, standard error of the difference than would a comparison of unmatched groups.

Furthermore, such matching may be less efficient than the comparison of unmatched random groups, unless the factor used in matching introduces a relatively strong positive relationship between the means. While a positive relationship, reflected in a positive covariance term, does reduce the standard error of the difference, this procedure also *halves* the number of degrees of freedom. Dealing with a sample of N pairs gives only half the number of degrees of freedom available when we deal with two independent groups of N cases each. Thus, if the factor entering into the matching is only slightly relevant to the differences between the groups, or is even irrelevant to such differences, matching is not a desirable procedure. The experimenter should have quite good reasons for matching before he adopts this procedure in preference to the simple comparison of two randomly selected groups.

Let us leave these considerations for a moment and return to the actual procedure for matched groups. The unknown value of cov. (M_1,M_2) could be something of a problem, but actually it is quite easy to bypass this difficulty altogether. Instead of regarding this as two samples, we simply think of the data coming from one sample of *pairs*. Associated with each pair i is a difference

$$D_i = (y_{i1} - y_{i2}),$$

where Y_{i1} is the score of the member of pair i who is in group 1, and Y_{i2} is the score of the member of pair i who is in group 2. Then an ordinary t test for a *single* mean is carried out using the scores D_i. That is,

$$M_D = \frac{\sum_i D_i}{N}$$

and
$$s_D^2 = \frac{\sum_i D_i^2}{N-1} - \frac{N(M_D)^2}{N-1}.$$

Then
$$\text{est. } \sigma_{M_D} = \frac{s_D}{\sqrt{N}}$$

and t is found from
$$t = \frac{M_D - E(M_D)}{\text{est. } \sigma_{M_D}}$$

with $N - 1$ degrees of freedom. *Be sure to notice that here N stands for the number of differences, which is the number of pairs.*

Naturally, the hypothesis is about the true value of $E(M_D)$, which is always $\mu_1 - \mu_2$. Thus any hypothesis about a difference can be tested in this way, provided that the groups used are matched *pairwise*. Similarly, confidence limits are found just as for a single mean, using M_D and σ_{M_D} in place of M and σ_M.

An example will now be given of this method of computation for a test of the difference in means of two matched groups. Not only will this example illustrate the method; the data have also been "rigged" to illustrate the point made above, that in some situations it may be less efficient to match than to simply take two randomly selected groups for comparison. This example will involve matched groups where a negative relationship exists among the pairs.

Consider once again the question of scores on a test of dominance. The basic question has to do with the mean score for men as opposed to the mean score for women. In carrying out the experiment, the investigator decided to sample eight husband-wife pairs at random. The members of each pair were given the test of dominance separately, and the data turned out as follows:

Pair	*Husband*	*Wife*	*D*	D^2
1	26	30	−4	16
2	28	29	−1	1
3	28	28	0	0
4	29	27	2	4
5	30	26	4	16
6	31	25	6	36
7	34	24	10	100
8	37	23	14	196
			31	369

Then
$$M_D = \frac{31}{8} = 3.87, \qquad s_D^2 = \frac{369}{7} - \frac{(31)^2}{7(8)} = 35.59,$$
$$\text{est.}\sigma_{M_D} = \sqrt{\frac{35.59}{8}} = \sqrt{4.45} = 2.109.$$

The t test is thus given by

$$t = \frac{3.87 - 0}{2.109} = 1.835.$$

For 7 degrees of freedom, this result is not significant (two-tailed test) for $\alpha = .05$ or less.

Now let us change our frame of reference slightly. Suppose that it had been true that these data came from two independent groups, one of men and one of women, each drawn at random. In this case, the men's group (formerly the husbands) would show a mean of 30.37, and an unbiased estimate of the variance s^2 of 13.00. The women's group would have a mean of 26.50, with an s^2 value of 6.02. Then the standard error of the difference would be estimated from

$$\text{est.}\sigma_{\text{diff}} = \sqrt{\frac{7(13 + 6.02)}{16 - 2}\left(\frac{2}{16}\right)}$$
$$= \sqrt{1.189} = 1.09, \quad \text{approximately.}$$

Then, in this instance the t would be given by

$$t = \frac{(30.37 - 26.50) - 0}{1.09} = 3.55,$$

which, for 14 degrees of freedom, is significant well beyond the .01 level, two-tailed. Why this very different result from the same set of numbers? The answer is given by the relationship of the scores when they are regarded as paired. Notice that high scores for husbands are paired with low scores for wives, and vice versa. This implies a negative relationship of such scores, leading to a negative covariance term in the estimate of the standard error. This, in turn, actually *increases* the size of the standard error relative to that for unmatched groups. When such a situation actually exists, there is a distinct disadvantage to matching.

Now do not get the idea that the option open to the author here is open to an experimenter. The sample is either drawn from a population of pairs or from two independent populations, and one does not have the right to change his mind about the nature of the sample after the fact. This procedure was strictly for illustrative purposes! Furthermore, when such matching or pairing is used, prior evidence or sheer common sense should suggest whether or not a positive relationship among the scores should exist or not. If such a positive relationship should exist, then the matching procedure may well reduce the sampling variance sufficiently to offset the loss in degrees of freedom, and a matching procedure may thus be desirable. The point is that this is not an automatic consequence of such matching. As with any question of experimental design, no routine procedure is advantageous for all situations, and the experimenter must bring his judgment and knowledge to bear on such decisions.

10.24 Significance Testing in More Complicated Experiments

Only in the very simplest experimental problems does the experimenter confine himself to two treatment groups. It is far more usual to find

experiments that involve a number of qualitatively or quantitatively different treatments. However, the basic conception of what the experimenter is doing remains the same: he is looking for evidence of a statistical relation between experimental and dependent variables. When there are several groups it is no longer possible to make a simple and direct connection between the degree of statistical association and the difference between any pair of means; here there are any number of pairs of means that may be different and thus imply association, and the mechanism of the simple t test breaks down. Thus, we will introduce this problem once again in somewhat different terms in Chapter 12. However, before we can discuss methods general enough to handle multi-group data, we need to study two more theoretical sampling distributions, both of which grow out of problems of inference about population variances. The next chapter is devoted to the study of these two distributions.

Exercises

1. A random sample of 300 American women were asked to record their body temperatures twice a day for a full month. From their records an average value was found for each woman. The mean of these values was 98.7 with a standard deviation S of .95. Test the hypothesis that the mean body temperature of such American women is 98.6 against the alternative that the mean is some other value.
2. Find the 99 percent confidence interval for the mean in problem 1.
3. In a study of truth in advertising, a government agency opened 500 boxes selected at random of a well-known brand of raisin bran. For each box the actual number of raisins was counted. The mean number of raisins was 32.4, with a standard deviation $S = 4.1$. Evaluate the company's claim that each box contains 34 raisins on the average, against the alternative of fewer raisins than claimed.
4. Find the 95 percent confidence interval for the mean in problem 3.
5. Suppose that the body weight at birth of normal children (single births) within the United States is approximately normally distributed and has a mean of 115.2 ounces. A pediatrician believes that the birth weights of normal children born of mothers who are habitual smokers may be lower on the average than for the population as a whole. In order to test this hypothesis, he secures records of the birth weights of a random sample of 20 children from mothers who are heavy smokers. The mean of this sample is 114.0 with $S = 4.3$. Evaluate the pediatrician's hunch.
6. Reevaluate the data of problem 5 on the assumption that a sample of 80 children had been used.
7. For the results of problem 5, find the 99 percent confidence interval for the mean birth weight of normal children from smoking mothers.
8. Suppose that in a certain large community the number of hours that a TV set is turned on in a given home during a given week is approximately normally distributed. A sample of 26 homes was selected, and careful logs were kept of how many hours per week the TV set was on. The mean number of hours per week in the sample turned out to be 36.1 with a standard deviation S of 3.3 hours. Find the 95 percent confidence interval for the mean number of hours that TV sets are played in the homes of this community.
9. For the data of problem 8, test the hypothesis that the true mean number of hours is 35. Test the hypothesis that the mean number of hours is 30.

10. Four random samples are taken independently from a population. For each random sample, the ninety percent confidence interval for the mean is found. What is the probability that at least one of those confidence intervals fails to cover the population mean? What is the probability that two or more confidence intervals fail to cover the population value?
11. The same government agency referred to in problem 3 has decided to compare two well-known brands of raisin bran with respect to the numbers of raisins each contain on the average. Some 100 boxes of Brand A were taken at random, and the same number of boxes of Brand B were randomly selected. On the average the Brand A boxes contained 38.7 raisins, with $S = 3.9$, and Brand B contained an average of 36 raisins with $S = 4$. Test the hypothesis that the two brands are actually identical in the average number of raisins that their boxes contain. Let H_1 be "not H_0."
12. For problem 11, find the 99 percent confidence interval for the difference in average number of raisins for Brands A and B.
13. The editor of a journal in Psychology tends to believe that the contributors to that journal now use shorter sentences on the average than they did a few years ago. In order to test this hunch, he takes a random sample of 150 sentences from journal articles written ten years ago and a random sample of 150 sentences from articles published within the last two years. The first sample showed a mean length of 127 type spaces per sentence, whereas the second sample showed a mean length of 113 type spaces. The first standard deviation $S = 41$, and the second standard deviation $S = 45$. Should he conclude that the recent articles do tend to have shorter sentences?
14. Find the 95 percent confidence interval for difference in sentence length from problem 13.
15. In an experiment, subjects were assigned at random between two conditions, five to each. Their scores turned out as follows:

CONDITION A	CONDITION B
128	123
115	115
120	130
110	135
103	113

Can one say that there is a significant difference between these two conditions? What must one assume in carrying out this test?

16. Find the 99 percent confidence interval for the difference between Conditions A and B in problem 8. On the evidence of this confidence interval, could one reject the hypothesis that the true mean of Condition B is five points higher than that of condition A?
17. In an experiment, the null hypothesis is that two means will be equal. The variance of each population is believed to be equal to 16. If $\alpha = .05$, two-tailed, and the test is to have a power of .90 against the alternative that $M_1 - M_2 = 3$, about how many cases should one take in each experimental group?
18. Suppose that two brands of gasoline were being compared for mileage. Samples of each brand were taken and used in identical cars under identical conditions. Nine tests were made of Brand I and six tests of Brand II. The following miles per gallon were found.

BRAND I	BRAND II
16	13
18	15
15	11
23	17
17	12
14	13
19	
21	
16	

Are the two brands significantly different? What must be assumed here in order to carry out the test?

19. An experimenter was interested in dieting and weight losses among men and among women. He believed that in the first two weeks of a standard dieting program, women would tend to lose more weight than men. As a check on this notion, a random sample of 15 husband-wife pairs were put on the same strenuous diet. Their weight losses after two weeks showed the following:

PAIR	HUSBANDS	WIVES
1	5.0 lbs	2.7 lbs
2	3.3	4.4
3	4.3	3.5
4	6.1	3.7
5	2.5	5.6
6	1.9	5.1
7	3.2	3.8
8	4.1	3.5
9	4.5	5.6
10	2.7	4.2
11	7.0	6.3
12	1.5	4.4
13	3.7	3.9
14	5.2	5.1
15	1.9	3.4

Did wives lose significantly more than husbands? What are we assuming here?