A Unified Approach to Some Standard Statistical Tests

James H. Steiger

November 13, 2006

Standard introductory texts on statistical methods in psychology deal with tests on means, variances, correlations, and proportions prior to treating the analysis of variance. Generally, excluding ANOVA, they cover a maximum of 12 testing situations, dealing with, for each of the above four parameters, a “1-Sample,” “2-Sample Independent,” and “Two-Sample Dependent” hypothesis test.

Many books present only a subset of these twelve situations. For example, standard texts seldom present procedures for comparing correlations from two dependent samples. In this chapter, we shall develop a general theoretical approach which will handle (with some minor adjustments for individual cases) all of the standard testing situations for means, proportions, and correlations. (Variances will require a special treatment later.) The method will also allow us to generate test statistics for a variety of circumstances not always covered in psychological statistics texts. Later, as a convenient by-product, we will also develop a general procedure for constructing confidence intervals, using a slight modification of the material developed in this section.

Our approach will be as follows. First, we shall develop general theory which applies to all of the relevant testing situations. Then we will apply the theory, adding modifications and/or improvements where applicable, to testing means, proportions, and correlations.

1. A General Model for Linear Combination Hypotheses

In all of the situations covered in this chapter, we will be interested in a single numerical value which is expressable as a linear combination of a group of *J* parameters. We will refer to these *J* parameters as θ_j, and the linear combination of interest can be written

$$\kappa = \sum_{j=1}^{J} c_j \theta_j \tag{1.1}$$

In the two-tailed testing situation, our statistical null hypothesis H_0 will be of the form $H_0 : \kappa = a$ where *a* is numerical constant (often zero). In one-tailed testing situations, the hypothesis could be of the form $\kappa \geq a$ for example. Since the generalization of our procedure from 2-tailed to 1-tailed situations is rather obvious, we will, for compactness, not emphasize it during our general discussion.

Instead, we will concentrate on the two-tailed test in our examples. The general form $\kappa = a$ covers a variety of interesting cases, including the traditional 1- and 2-sample tests. For example, consider tests on means, in which the most common tests are for the following null hypotheses: $\mu = a$, $\mu_1 = \mu_2$ (for independent samples) and $\mu_1 = \mu_2$ (for dependent, or "matched" samples). Since they can, respectively, be written $\mu = a$, $\mu_1 - \mu_2 = 0$, and $\mu_1 - \mu_2 = 0$, they are all of the form $\kappa = a$.

In what follows, we will assume that, for each θ_j, we have an unbiased, normally distributed estimator $\hat{\theta}_j$. Furthermore, we also assume that the sampling variance γ^2 of $\hat{\theta}$ is somehow known (later we will relax this assumption somewhat, and assume only that a consistent estimator $\hat{\gamma}^2$ of this variance is available). Symbolically, we will express the fact that "$\hat{\theta}_j$ is normally distributed with mean θ_j and sampling variance γ_j^2" with notation

$$\hat{\theta}_j \approx N\left(\theta_j, \gamma_j^2\right) \tag{1.2}$$

In the case where the $\hat{\theta}_j$ have been computed on dependent samples (for example, repeated measures on the same subjects), we will assume further that, for any two estimates $\hat{\theta}_i$ and $\hat{\theta}_j$, their covariance γ_{ij} is also known or estimable.

1.1 Independent Samples, Sampling Variances Known

In this section, we assume that *J independent* samples of size n_j are available to test hypotheses about *J* parameters. We can easily construct a normally distributed, unbiased estimator *K* of κ as

$$K = \sum_{j=1}^{J} c_j \hat{\theta}_j \tag{1.3}$$

Since the $\hat{\theta}_j$ are based on independent samples, they will, themselves, be independent. Consequently, the expected value and variance of *K* may be calculated easily from the general theory of linear composites.

$$\begin{aligned} E(K) &= E\left(\sum_{j=1}^{J} c_j \hat{\theta}_j\right) \\ &= \sum_{j=1}^{J}\left[E\left(c_j \hat{\theta}_j\right)\right] \\ &= \sum_{j=1}^{J} c_j E\left(\hat{\theta}_j\right) \\ &= \sum_{j=1}^{J} c_j \theta_j = \kappa \end{aligned} \tag{1.4}$$

$$Var(K) = \sigma_K^2 = \sum_{j=1}^{J} c_j^2 \gamma_j^2 \tag{1.5}$$

We know that linear composites of independent, normal random variables are themselves normally distributed, so that we can, consequently state that

$$K \approx N\left(\kappa, \sigma_K^2\right) \tag{1.6}$$

From this, it immediately follows that

$$Z = \frac{K - \kappa}{\sigma_K} \tag{1.7}$$

will have a $N(0,1)$ distribution.

Equations 1.4 through 1.7 thus provide a general form for constructing a test statistic in cases where a linear composite κ of parameters is of interest, and unbiased, normally distributed estimators with known variance are available for each parameter.

Example. Consider the mean μ of a single population *P* with a $N\left(\mu, \sigma^2\right)$ distribution. We estimate μ with $\overline{x}_\bullet$, a sample mean based on a random sample of n observations from *P*. We wish to test the statistical null hypothesis

$$H_0\!: \mu = a\ .$$

This example is particularly interesting, because we can apply the general theory given above twice. First, it is important to realize that a sample of n independent observations, all taken from the same population *P*, may be conceptualized as either (1) a single sample of size *n* or (2) *n* independent samples of size 1. Suppose we adopt the latter conceptualization first, and apply our theory to derive the expected value and sampling variance of the statistic $\overline{x}_\bullet$. Since $\overline{x}_\bullet$ may be written as

$$\overline{x}_\bullet = \sum_{i=1}^{n} \frac{1}{n} x_i\ ,$$

where all the x_i have the same mean μ and variance σ^2, it follows directly from Equations 1.4 through 1.7 that

$$\overline{x}_\bullet \approx N\left(\mu, \frac{\sigma^2}{n}\right) \tag{1.8}$$

Specifically,

$$\begin{aligned} E(\bar{x}_{\bullet}) &= E\left(\sum_{i=1}^{n} \frac{1}{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(x_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \ \mu \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned} \tag{1.9}$$

and

$$\begin{aligned} \sigma^2_{\bar{x}_{\bullet}} &= \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 \sigma^2 \\ &= n\left(\frac{1}{n}\right)^2 \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{1.10}$$

Now that we have established the expected value and sampling variance of $\bar{x}_{\bullet}$, we may use Equations 1.4 through 1.7 directly, to write a test statistic for H_0. In this case, $E(\kappa) = \mu$, $K = \bar{x}_{\bullet}$, $\gamma^2 = \sigma^2/n$, whence, substituting in Equations 1.4 through 1.7, we obtain

$$Z = \frac{\bar{x}_{\bullet} - a}{\sqrt{\sigma^2/n}} \tag{1.11}$$

This is the familiar “single-sample” test for a single mean, “when the population variance is known” given in many introductory psychological statistics textbooks.

1.2 Independent Samples, Sampling Variances Unknown but Estimable

Case 1.1 is of theoretical interest, but it seldom arises in practice. Needless to say, it is not likely that we will know enough about a population to have precise knowledge of its variance without also knowing its mean. It might seem, then, that the theory we have developed so far would not be of much use. Actually, however, the assumptions in 1.1 can be relaxed considerably, and an asymptotically normal test statistic can still be obtained. Specifically, we now require only that the *asymptotic* distribution of

$$n^{1/2}\left(\hat{\theta}_i - \theta_i\right)$$

be $N\left(0, \beta_i^2\right)$ for some finite β_i^2, and that a consistent estimator of β_i^2 be available. This will be true when the following situations are satisfied:

1. The $\hat{\theta}_i$ have asymptotic distributions which are $\mathrm{N}\left(\theta_i, \gamma_i^2\right)$.

2. The γ_i^2 can be written in the form β_i^2 / n for some finite β_i^2.

3. The β_i^2 can be estimated (consistently with estimates $\hat{\beta}_i^2$) from sample data.

If 1–3 are met, then the statistic

$$Z = \frac{K - a}{\sqrt{\hat{\sigma}_K^2}}, \tag{1.12}$$

where

$$\hat{\sigma}_K^2 = \sum_{j=1}^{J} c_j^2 \hat{\gamma}_j^2 = \sum_{j=1}^{J} c_j^2 \frac{\hat{\beta}_j^2}{n_j} \tag{1.13}$$

will have an asymptotic distribution which, if the null hypothesis $\kappa = a$ is true, will be $N(0, 1)$.

Example. The Central Limit Theorem, in the form usually given in psychology statistics texts, states that, so long as the population has an arbitrary distribution with finite variance, a sample mean $\bar{x}_\bullet$ will have an asymptotic sampling distribution which, as n becomes large, approaches a normal distribution with mean μ and variance σ^2/n. Clearly, then, the sample mean $\bar{x}_\bullet$ meets the minimal requirements of this section, since, in the above notation, $\beta^2 = \sigma^2$, and σ^2 can be estimated with a sample variance s^2. Hence, for any population distribution having a finite variance σ^2 and mean μ, the test statistic

$$Z = \frac{\bar{x}_\bullet - a}{\sqrt{s^2/n}} \tag{1.14}$$

will be asymptotically distributed $N(0,1)$ if $\mu = a$.

1.3 Dependent Samples, Sampling Variances and Covariances Unknown but Estimable

In some situations, subjects are observed more than once, or, alternatively, samples are matched (e.g., husband-wife pairs are used in two experimental groups) so that it is no longer tenable to assume that observations in the *J* groups are independent. In this case, one can still construct an asymptotically normal test statistic, so long as consistent estimators for the sampling covariances of the estimators $\hat{\theta}_j$ are available. The test statistic will be similar in form to Equation 1.12 above, except now (assuming $\hat{\theta}_j$ and $\hat{\theta}_k$ have covariance $\gamma_{jk} = \beta_{jk}/n$

$$\hat{\sigma}_K^2 = (1/n)\left(\sum_{j=1}^{J} c_j^2 \hat{\beta}_j^2 + 2\sum_{j>k} c_j c_k \hat{\beta}_{jk}\right) \tag{1.15}$$

This modified equation takes into account the fact that the individual estimators are no longer independent.

In what follows, we will draw upon the general theory presented above to construct test statistics to handle the most common tests on means, proportions, and correlations. Special cases will involve minor improvements on and extensions to the general theory, and these will be made explicit where necessary.

2. Specific Theory for Tests on Means

2.1 Variances and/or Covariances Known

The theory we present here is given in many of the more advanced psychological statistics texts. It is obtained very easily from the general theory, by simply substituting known facts about the sampling variance and covariance of sample means into relevant expressions in Chapter 1.

Specifically, suppose the sample mean $\bar{x}_{\bullet j}$ is based on a sample of size n_j from a population having mean μ_j and variance σ_j^2. As we have seen already, $\bar{x}_{\bullet j}$ has expected value μ_j, and sampling variance σ_j^2/n_j . Moreover, it can be shown (in a manner similar to the way we proved the result for a sampling variance) that, if two means are based on non-independent samples on random variables having covariance σ_{ij} , that the covariance between the means is σ_{ij}/n . Substituting these results, i.e.,

$$\gamma_j^2 = \sigma_j^2/n_j \tag{2.1}$$

$$\theta_j = \mu_j \tag{2.2}$$

$$\hat{\theta}_j = \bar{x}_{\bullet j} \tag{2.3}$$

$$\gamma_{ij} = \sigma_{ij}/n \tag{2.4}$$

we have the following general theory for testing means when variances and covariances are known:

2.1.1 Independent Samples

For a null hypothesis of the form

$$H_0\colon\ \kappa = \sum_{j=1}^{J} c_j \mu_j = a\ , \tag{2.5}$$

the test statistic will be of the form

$$Z = \frac{K - a}{\sigma_K}, \tag{2.6}$$

where

$$K = \sum_{j=1}^{J} c_j \bar{x}_{\bullet j}\,, \tag{2.7}$$

and

$$\sigma_K^2 = \sum_{j=1}^{J} c_j^2 \frac{\sigma_j^2}{n_j}\,. \tag{2.8}$$

Example. An experimenter hypothesizes that the mean “Depression Score” of first-year students at her university is 85. The standard deviation of such scores is known to be 15. The experimenter takes a random sample of 36 independent observations from the population of first year students, and finds a sample mean depression score of 90. What is the probability of obtaining an $\bar{x}_\bullet$ of 90 or higher if the experimenter's hypothesis is true?

In this case, the null hypothesis may be expressed as

$$H_0\colon \mu = 85\,.$$

This is a linear combination of the form $\kappa = a$, where only one population mean is involved in κ and the single linear weight c_1 is $+1$. Hence, $K = (+1)\bar{x}_\bullet = \bar{x}_\bullet$ and $\sigma_K^2 = (+1)^2\, \sigma^2 \big/ n = \sigma^2 \big/ n$. The test statistic is therefore

$$\begin{aligned} Z &= \frac{\bar{x}_\bullet - a}{\sqrt{\sigma^2 \big/ n}} \\ &= \frac{\bar{x}_\bullet - a}{\sigma \big/ \sqrt{n}} \\ &= \frac{90 - 85}{15/6} \\ &= 2.00 \end{aligned} \tag{2.9}$$

Consulting the normal curve table, we see that the probability of a Z less than or equal to 2.00 is .9772, and the probability of a Z higher than 2.00 is only .0228. In the face of this evidence, we might well decide to reject the experimenter's hypothesis.

Example. A professor has a hypothesis that vitamin E, consumed for two days prior to a statistics exam, will have absolutely no effect on a statistics student's test performance. His graduate student believes otherwise and conducts the following experiment. Two groups of 25 students each are randomly selected from the statistics classes at the university. Prior to the statistics exams, one group consumes 400 I.U. of Vitamin E, while the other group consumes placebo gelatin capsules. Exam results are: $\overline{x}_{\bullet 1}\left(\text{Vitamin E group}\right) = 47$; $\overline{x}_{\bullet 2}$ (Placebo group) $= 86$. It is known that $\sigma^2 = 50$ for both populations. In our general theoretical formulation, this is a hypothesis of the form

$$H_0\text{: } \mu_1 - \mu_2 = 0\,.$$

Hence it is a linear combination hypothesis in which there are two linear weights, $+1$ and -1. Consequently, $K = \overline{x}_{\bullet 1} - \overline{x}_{\bullet 2}$, and $\sigma_K^2 = \sigma_1^2 \big/ n_1 + \sigma_2^2 \big/ n_2$. The test statistic is

$$Z = \frac{\overline{x}_{\bullet 1} - \overline{x}_{\bullet 2}}{\sigma_K}\,,$$

where

$$\sigma_K = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,.$$

We have

$$\begin{aligned} Z &= \frac{47 - 86}{\sqrt{\frac{50}{25} + \frac{50}{25}}} \\ &= \frac{-39}{2} \\ &= -19.5 \end{aligned}$$

If the professor's hypothesis is true, the probability of obtaining a Z-statistic this small or smaller is virtually zero. The experimental results would thus lead us to suspect very strongly that the professor's hypothesis is incorrect.

2.1.2 Dependent Samples

In this case, we assume that covariances as well as variances can be estimated. The test statistic would be the same as in the independent sample case, except that σ_K^2 would now be calculated as

$$\sigma_K^2 = (1/n)\sum_{j=1}^{J} c_j^2 \sigma_j^2 + 2\sum_{j>k} c_j c_k \sigma_{jk}$$

No examples will be given, since this theory is rarely ever applied directly.

2.2 Variances and Covariances Unknown, Populations Normally Distributed

2.2.1 Independent Samples

In most cases of practical interest involving tests on population means, it is unreasonable to assume that variances are somehow known. As we have seen in Section 1.1, an asymptotically normal statistic could be obtained by simply substituting sample variances and covariances in the formulae given in sections 1.1, 2.1.1 and 2.1.2 above.

However, in 1908, writing under the pen name “Student,” the statistician W.S. Gossett produced results which imply that, if certain assumptions are met, and if a particular consistent estimation process is used in implementing the theory in Section 1.1, then the exact distribution of the resulting test statistic can be determined. This statistic, which is of the same general form as the test statistics given in Section 2.1.1, is called “Student's t ” statistic in honor of its creator.

For Gossett's results to be (strictly) applicable, the following assumptions about the statistical populations must hold.

- The populations must have a multivariate normal distribution. (If samples are independent, this simplifies to an assumption that individual populations are normally distributed.)
- The populations must have equal variances, if the test is to be performed on independent samples. (This assumption is referred to by a variety of names, such as the “homoscedasticity assumption,” or “homogeneity of variances” assumption.)

For the statistic to have a t distribution, the assumption of equal variances must be incorporated into the denominator of the test statistic. When this is done, the term $\hat{\sigma}^2$ can be factored out of the resulting expression. To obtain a *t*-statistic, a particular kind of estimator must be used. Thus, for independent samples, the formula for $\hat{\sigma}_K^2$ is simplified to

$$\hat{\sigma}_K^2 = \hat{\sigma}^2 \sum_{j=1}^{J} \frac{c_j^2}{n_j} \tag{2.10}$$

where $\hat{\sigma}^2$ is the pooled, unbiased estimator of σ^2 (also known as "mean square within" in the analysis of variance), computed as

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{J}(n_j - 1)s_j^2}{\sum_{j=1}^{J}(n_j - 1)} = \frac{\sum_{j=1}^{J}(n_j - 1)s_j^2}{n_\bullet - J} \tag{2.11}$$

where

$$n_\bullet = \sum_{j=1}^{J} n_j \tag{2.12}$$

is the total number of observations in the *J* groups. (Note that, when sample sizes are equal, $\hat{\sigma}^2$ is simply the arithmetic average of the *J* sample variances.) Consequently, the general formulae for constructing a *t*-statistic for testing a linear combination hypothesis about means, when variances are not known, but can assumed to be equal, and populations are normal, is

$$t_\nu = \frac{K - a}{\sqrt{\hat{\sigma}_K^2}} \tag{2.13}$$

where

$$\nu = n_\bullet - J \tag{2.14}$$

is the number of degrees of freedom for the *t*-statistic,

$$K = \sum_{j=1}^{J} c_j \overline{x}_{\bullet j}\,, \tag{2.15}$$

and $\hat{\sigma}_K^2$ is as given by Equations 2.10–2.12.

Example. The Müller-Lyer Illusion in Memory. You may already be familiar with the Müller-Lyer illusion. In this illusion, one horizontal line appears to be shorter than it actually is, while another line appears to be longer than it actually is. Suppose, in a variant of the typical M-L task, we ask a subject to study one part of the M-L illusion (i.e., the one with the “outward arrows” which make a line appear longer than it actually is) and then draw the horizontal line from memory. Suppose the horizontal line is actually 5 cm. The question is, will the illusion persist in memory? The null hypothesis in this case is

$$H_0\colon\ \mu = 5\,,$$

that is, if there is no illusion effect, the average “reproduced” line segment should be 5 cm. Suppose that an experiment is run in which 9 subjects all give their judgements. The following data are obtained: $\bar{x}_\bullet = 9,\ s^2 = 4,\ n = 9$. Application of the general formulae in this section yield the following test statistic:

$$\begin{aligned} t_{n-1} &= \frac{\bar{x}_\bullet - 5}{s/\sqrt{n}} \\ t_8 &= \frac{9-5}{2/3} \\ &= 6.00 \end{aligned}$$

Apparently, there is a significant illusion effect, even in memory. The rejection point for a *t* statistic with 8 degrees of freedom for a hypothesis test with $\alpha = .05$, two-tailed, is 2.306.

Example. Transcendental Meditation and Memory. Transcendental Meditation (TM) is a meditation technique which was publicized widely in the 1970's. Many benefits were claimed by its adherents, and improved memory was one of them. To test the hypothesis that TM improves memory, we conduct a two group experiment, in which one group receives TM training, the other group a similar type of training which is claimed by the skeptical to be equivalent to TM, but which TM adherents claim is clearly an inferior meditation technique. Ten subjects are randomly selected for each group. All subjects receive training, followed by a standard memory recall task. The null hypothesis is

$$H_0\colon\ \mu_1 = \mu_2\,.$$

Using the general equations, we construct a test statistic of the form

$$t_{n_1+n_2-2} = \frac{\bar{x}_{\bullet 1} - \bar{x}_{\bullet 2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{\sigma}^2}}\,, \tag{2.16}$$

where

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

This is, of course, the well-known “Two-sample Student's *t*-statistic”.

Suppose we obtained the following data from our experiment:

$n_1 = n_2 = 10,\ \bar{x}_{\bullet 1} = 23,\ \bar{x}_{\bullet 2} = 19.8,\ s_1^2 = 23,\ s_2^2 = 27$

In this case, since sample sizes are equal, we have

$$\begin{aligned}\hat{\sigma}^2 &= \frac{s_1^2 + s_2^2}{2}\\ &= \frac{27+23}{2}\\ &= \frac{50}{2}\\ &= 25\end{aligned}$$

and, consequently,

$$\begin{aligned}t_{18} &= \frac{23-19.8}{\sqrt{\left(\frac{1}{10}+\frac{1}{10}\right)25}}\\ &= \frac{3.2}{\sqrt{5}}\\ &= 1.43\end{aligned}$$

Assuming, again, a two-tailed test with $\alpha = .05$, the rejection point for a *t*-statistic with 18 degrees of freedom is 2.101, and so our current result is “not significant.” We would conclude that the performance of the TM group is not significantly different from that of the control group.

2.2.2 Dependent Samples

When samples are dependent, a particularly simple form of test statistic is available for testing linear combination hypotheses about means. Suppose that each of *n* subjects is observed *J* times, and you wish to test a linear combination hypothesis about the resulting *J* population means. The basic strategy of the simplified technique is to compute, for each subject, a linear combination score k_i. If x_{ij} is the score of the *i*th subject on the *j*th repeated measure, then

$$k_i = \sum_{j=1}^{J} c_j x_{ij} \tag{2.17}$$

By combining, for each subject, the *J dependent* scores into one score, we have eliminated the need to estimate covariances, and we have converted a *J*-Sample problem into a 1-Sample problem. Specifically, since the mean of the k_i scores, $\bar{k}_\bullet$, has an expected value of κ, we can test the null hypothesis that $\kappa = a$ with a 1-Sample *t*-statistic of the form

$$t_{n-1} = \frac{\overline{k}_\bullet - a}{\sqrt{s_k^2 / n}} \tag{2.18}$$

The most familiar application of this general technique is for the “two-sample, correlated sample *t*-test,” or “matched sample *t*-test”, in which each subject is measured twice, and the null hypothesis is of the form

$$H_0\text{: } \mu_1 = \mu_2 \text{ .}$$

In this case, we compute, for each subject, a difference score k_i , which is simply the difference between his/her scores on the two measures. We then perform a 1-sample test on these difference scores, as if we were testing a hypothesis that $\mu = 0$.

Example. Suppose you have a hypothesis that, because of the interesting metabolic characteristics of statistics students, consumption of beer has absolutely no effect on their cognitive capacities. You decide to test this hypothesis by having each of 10 randomly selected students play games of “Night-Mission Pinball” either sober, or immediately after consuming 3 beers. (To control for practice effects, order is counterbalanced.) The raw data for the 10 subjects are as follows:

Beer	No Beer	Difference
65	54	+11
60	187	−127
102	99	+3
143	265	−122
97	119	−22
234	445	−211
254	354	−100
45	65	−20
89	111	−22
123	167	−44

Here are the summary statistics for the difference scores:

$\overline{k}_\bullet = -65.4$

$n = 10$

$s_k^2 = 5132.29$

We have

$$
\begin{aligned}
t_9 &= \frac{\overline{k}_{\bullet}}{\sqrt{s_k^2/n}} \\
&= \frac{-65.4}{\sqrt{5132.29/10}} \\
&= \frac{-65.4}{22.65} = -2.89
\end{aligned}
$$

The resulting t-statistic is -2.89.

If we were performing the two-tailed hypothesis at the $\alpha = .05$ level, the critical value of *t*, with 9 degrees of freedom, is 2.262, and so we would decide that there is a significant decrement in performance when beer is consumed.

The above "dependent-sample procedure" is so simple, one might be tempted to use it all the time when sample sizes for the *J* groups are equal, even if the groups were independent samples. However, the statistic (though it would be valid) would suffer a loss of power (relative to the independent sample procedure), since degrees of freedom are sharply reduced, and (if sample sizes are small) rejection points for the resultant statistics would differ by an appreciable amount. Needless to say, this procedure definitely should be used if samples are dependent.

3. Specific Theory for Tests on Proportions

In this section, we consider the general class of situations in which data are binary (i.e., the outcome is in one of two distinct classes, and hence can be scored either 0 or 1), and the parameter of interest is the population proportion π. In such situations, we can frequently model the sample proportion *p* as resulting from *n* independent observations of a Bernoulli (i.e., binomial) process, and the resulting proportion *p* can be expressed as $p = X/n$, where *X* is a $B(n, \pi)$ random variable (i.e., a binomial random variable with *n* trials and π probability of success). In this case, it is clear from our previous work on the binomial distribution that

$$E(p) = \pi \tag{3.1}$$

and

$$\sigma_p^2 = \frac{\pi(1-\pi)}{n} \,. \tag{3.2}$$

3.1 Independent Samples

Furthermore, since *p* is also the sample mean of the binary data, the Central Limit Theorem applies, and *p* is, asymptotically, normally distributed. In this case, we can apply the theory from Section 1.1 to develop general theory for testing linear combination hypotheses on independent sample proportions. Specifically, for any hypothesis of the form

$$H_0\colon \kappa = \sum_{j=1}^{J} c_j \pi_j = a$$

we could construct a test statistic of the form

$$Z = \frac{K-a}{\sqrt{\hat{\sigma}_K^2}} \tag{3.3}$$

where

$$\hat{\sigma}_K^2 = \sum_{j=1}^{J} c_j^2 \frac{p_j(1-p_j)}{n_j} \,, \tag{3.4}$$

and

$$K = \sum_{j=1}^{J} c_j p_j \tag{3.5}$$

The statistics described in Equations 3.3–3.5 are valid, asymptotically normal test statistics. Below we consider two simple applications of the theory, and discover that the statistics generated by these equations are, in practice, “improved” slightly by minor modifications.

Example. The one-sample proportions test. Suppose that, on the basis of a random sample of size *n* from some population, we wish to test the hypothesis that a proportion *a* of the population favors a particular political point of view. The null hypothesis is

$$H_0\colon \pi = a \,.$$

According to Equations 3.3–3.5, the test statistic would be

$$Z = \frac{p-a}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.6}$$

Actually, in practice, the null hypothesis is “incorporated” into the denominator. Since if H_0 is true, $\pi = a$, then, to control Type I Error, we do not need to estimate π in the denominator. Rather, we may simply use *a*, its assumed value under the null hypothesis.

This leads to an “improved” statistic,

$$Z = \frac{p-a}{\sqrt{\frac{a(1-a)}{n}}} \tag{3.7}$$

Most texts recommend the statistic in Equation 3.7 as superior to the statistic given in Equation 3.6. However, their recommendations are based, implicitly, on an overemphasis of Type I Error rate performance. Clearly, when the null hypothesis is false, the denominator in Equation 3.7 can be a biased estimator of the asymptotic standard error of *p*. When the denominator tends to be too large, power for this statistic can suffer relative to the statistic in Equation 3.6. Suppose, for example, the null hypothesis posits an *a* of .5, when the true value of π is .3. Clearly the quantity $p(1-p)$ converges in the limit to .21, while $a(1-a)$ is .25. In such a situation, the statistic in Equation 3.6 will have superior power.

Example. The Two-Sample Independent Sample Test for Equal Proportions. In this case, two independent samples of possibly unequal size are taken, and sample proportions observed, in order to determine whether the population proportions are the same. A classic application would be a two group experiment in which the major question of interest is whether a treatment (possibly a persuasive message) affects the proportion of the population agreeing with some position. The statistical null hypothesis is

$$H_0\!: \pi_1 - \pi_2 = 0$$

Direct application of Equations 3.3–3.5 lead to the following statistic for testing this hypothesis:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{3.8}$$

In practice, a somewhat different statistic, which incorporates the null hypothesis into the variance estimate in the denominator, is employed. This statistic, taking into account the assumed equality of proportions π_1 and π_2 uses a pooled estimator $\overline{p}$, in place of p_1 and p_2 in its denominator. The resulting statistic is

$$Z = \frac{p_1 - p_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\overline{p}\left(1 - \overline{p}\right)}} \tag{3.9}$$

where $\overline{p}$ is the proportion in the group produced by combining the two experimental samples, i.e.,

$$\overline{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \tag{3.10}$$

3.2 Dependent Samples

The theory in this section is primarily useful when subjects are observed several times, (perhaps under several different experimental conditions), and a binary measure of behavior is recorded. In this situation, the sample proportions obtained on the two occasions *i* and *j* will have a covariance given by

$$\sigma_{p_i, p_j} = \frac{\pi_{ij} - \pi_i \pi_j}{n} \tag{3.11}$$

where π_{ij} refers to the population proportion of subjects who produce the behavior scored “1” on both occasions *i* and *j*. Theory from Section 1.3, combined with Equations 3.3–3.5 and 3.11, would allow us to construct test statistics for any linear combination hypothesis on dependent proportions. Here, we will consider the simplest special case.

Example*. *Comparing Two Correlated Proportions. In this case, a group of *n* subjects is observed on two occasions. The statistical null hypothesis is that the proportion of subjects performing a behavior of interest has not changed, i.e.,

$$H_0\colon \pi_1 - \pi_2 = 0 \tag{3.12}$$

Direct application of the standard theory for asymptotically normal test statistics, given in Section 1.3, together with the special results for proportions given above, would yield the following test statistic:

$$Z = \sqrt{n}\,\frac{p_1 - p_2}{\sqrt{p_1\left(1 - p_1\right) + p_2\left(1 - p_2\right) - 2\left(p_{12} - p_1 p_2\right)}}$$

In practice, the pooled estimator $\overline{p}$ is substituted for p_1 and p_2 in the denominator of the preceding equation, yielding

$$Z = \sqrt{n}\,\frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p}) + \bar{p}(1-\bar{p}) - 2\left(p_{12} - \bar{p}^2\right)}}$$

$$= \sqrt{n}\,\frac{p_1 - p_2}{\sqrt{2(\bar{p} - p_{12})}}$$

Now, in the repeated measures case,

$$\bar{p} = \frac{p_1 + p_2}{2}.$$

Substituting in the above, and rearranging, we obtain

$$\begin{aligned} Z &= \frac{n(p_1 - p_2)}{\sqrt{n}\sqrt{(p_1 - p_{12}) + (p_2 - p_{12})}} \\ &= \frac{n(p_1 - p_2)}{\sqrt{n(p_1 - p_{12}) + n(p_2 - p_{12})}} \end{aligned} \qquad (3.13)$$

The psychometrician Quinn McNemar found an ingeniously convenient notation which allows further simplification. Define n_{01} as the number of subjects who did not perform the behavior of interest on the first occasion, but did on the second. Similarly, define n_{10} as the number who performed the behavior on the first occasion, but not on the second, n_{00} as the number who perform the behavior on neither occasion, and n_{11} as the number who perform the behavior on both occasions.

The number of subjects who perform the behavior on the first occasion is $n_{10} + n_{11}$. Similarly, the number who perform the behavior on the second occasion is $n_{01} + n_{11}$. Hence, $np_1 = n_{10} + n_{11}$, $np_2 = n_{01} + n_{11}$, and $np_{12} = n_{11}$. Substituting in 3.13, we obtain

$$Z = \frac{n_{10} - n_{01}}{\sqrt{n_{10} + n_{01}}} \qquad (3.14)$$

Equation 3.14 is referred to as “McNemar's test for correlated proportions.” The test appears in several variations in a number of books. Keep in mind that, when the test is two-tailed, and, as in this case, the distribution of the test statistic is symmetric with respect to the $\alpha/2$ and $1-\alpha/2$ probability points, the sign of the statistic has no effect on the decision. Hence, it does not matter whether the numerator is written $n_{01} - n_{10}$, or as it appears in Equation 3.14. Also, either variant of the entire formula can be squared, yielding a chi-square statistic.

4. Specific Theory for Tests on Correlations

There are a huge variety of correlational tests available for comparing one, two, or several correlations. Traditional psychological statistics tests emphasize only a tiny subset of them, and often ignore key facts about them, such as the effect of violation of statistical assumptions.

4.1 Independent Samples

In 1898, Pearson and Filon published a paper which contained a key result, i.e., the asymptotic multivariate distribution of sample correlation coefficients. They showed that, *if the population distribution is multivariate normal*, the large sample distribution of a single sample correlation r_{ij} is approximately

$$N\left(\rho,\ \frac{\left(1-\rho^2\right)^2}{n}\right).$$

This fact can be used, along with the general theory in Section 1.2, to construct asymptotic tests on independent correlation coefficients. For example, suppose we wished to test the hypothesis that

$$H_0\!:\ \rho = a\ .$$

One statistic that comes immediately to mind is obtained by substituting in our general form:

$$Z = \frac{r-a}{\sqrt{\dfrac{\left(1-r^2\right)^2}{n}}}$$

$$= \frac{\sqrt{n}\left(r-a\right)}{1-r^2}$$

However, we could also incorporate the null hypothesis into the denominator, obtaining

$$Z = \sqrt{n}\left(\frac{r-a}{1-a^2}\right) \qquad (4.1)$$

Notice that, when $a = 0$, i.e., we are testing the hypothesis that $\rho = 0$, the above equation takes on a particularly simple form, i.e.,

$$Z = \sqrt{n}\ r \qquad (4.2)$$

There are several reasons one does not find this equation in textbooks. First, although the sample correlation *r* has an asymptotic distribution that is normal, it is, unfortunately, the case that convergence to this asymptotic distribution is rather slow, that is, *r* has a distribution which departs appreciably from normality at small sample sizes. This departure from normality becomes more severe as ρ approaches 1 in absolute value. Consequently, the test statistic in Equation 4.1 is only useful if *a* is close to 0, or if *n* is very large.

R. A. Fisher, the famous statistician, introduced the "Fisher transform" as a solution to these problems. The Fisher transform, which we will denote as $\phi(r)$, is a monotonic functional transform of *r* (actually its inverse hyperbolic tangent), which is usually computed as

$$\phi(r) = \tfrac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$\phi(r)$ has some extremely useful statistical properties. Succinctly, we can simply say that, with a fairly high degree of accuracy

$$\phi(r) \approx N\left(\phi(\rho), 1/(n-3)\right)$$

$\phi(r)$ is almost exactly normally distributed, regardless of the value of ρ, and, it has the added virtue of having a variance which, for a given sample size, is known! The statistical problems associated with *r* can be bypassed, to a great extent, by using $\phi(r)$.

The way we shall do this is as follows. A hypothesis about ρ will be rephrased as a hypothesis about $\phi(\rho)$. Then, the standard normal theory in Section 1.1 will be applied to construct test statistics, using $\phi(r)$ as the estimator for $\phi(\rho)$.

Example. The Fisher Z test for a single correlation. Suppose we wish to test the hypothesis of the form

$$H_0 \colon \rho = a$$

This hypothesis is true if and only if $\phi(\rho) = \phi(a)$. Hence, we can test the former hypothesis, indirectly, by testing the latter. The test statistic is of the form

$$Z = \frac{\phi(r) - \phi(a)}{\sqrt{1/(n-3)}} \tag{4.3}$$

Example. Comparison of two independent correlations. Frequently, we wish to test whether two correlation coefficients, measured on independent samples, are equal. For example, we might wish to test whether the correlation between anxiety and smoking is the same for men as it is for women. The hypothesis, of the form

$$H_0\colon \rho_1 = \rho_2$$

would be tested by taking samples of (possibly different) sizes n_1 and n_2, computing the correlations r_1 and r_2, and using the test statistic

$$Z = \frac{\phi(r_1) - \phi(r_2)}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \tag{4.4}$$

Unfortunately, the use of $\phi(r)$ for linear combination hypotheses is limited to the abovementioned two special cases. Fortunately, these are two fairly important special cases. To see why the use of $\phi(r)$ is limited, remember that, although it is a monotonic transform, it is not a linear transform. Looking back on our discussion of permissible transforms at the beginning of the year, we recall that $\phi(r)$ preserves equality-inequality relationships, and preserves an ordering, but does not preserve the relative size relationships of intervals. Consequently, we could not, for example, use the Fisher Transform to test the hypothesis that $\rho_1 - \rho_2 = .4$, because the quantity $\phi(\rho_1) - \phi(\rho_2)$ can take on infinitely many values, depending on the precise values of ρ_1 and ρ_2, and is generally not equal to $\phi(.4)$.

4.2 Dependent Samples

There are many circumstances in which psychologists wish to compare correlations measured on the same person. For example, we might wish to test which of two personality measures is a better predictor of smoking behavior. We take 3 measures, say, smoking, anxiety, and neuroticism. The statistical null hypothesis would be of the form

$$H_0\colon \rho_{12} = \rho_{13}$$

We cannot use Equation 4.4 to test this hypothesis, because the two sample correlation coefficients, computed on observations from the same population, would not, in general, be independent. They would, like means based on repeated measures, have a sampling covariance, which must be taken into account in a test statistic. Pearson and Filon, in their 1898 paper, showed that the large sample distribution of two sample correlation coefficients, r_{jk} and r_{hm}, measured on the same subjects, is approximately bivariate normal, with r_{jk} and r_{hm} having a sampling covariance of

$$\sigma_{r_{jk}r_{hm}} = \frac{1}{2n}\left(\begin{array}{r} \left(\rho_{jh} - \rho_{jk}\rho_{kh}\right)\left(\rho_{km} - \rho_{kh}\rho_{hm}\right) + \left(\rho_{jm} - \rho_{jh}\rho_{hm}\right)\left(\rho_{kh} - \rho_{kj}\rho_{jh}\right) \\ + \left(\rho_{jh} - \rho_{jm}\rho_{mh}\right)\left(\rho_{km} - \rho_{kj}\rho_{jm}\right) + \left(\rho_{jm} - \rho_{jk}\rho_{km}\right)\left(\rho_{kh} - \rho_{km}\rho_{mh}\right) \end{array}\right). \quad (4.5)$$

In typical applications, we could estimate this covariance by substituting sample correlations (but possibly also incorporating the null hypothesis) in the above equation. It is well-known (see, for example, my paper “Tests for Comparing Elements of a Correlation Matrix,” in the 1980 *Psychological Bulletin* that the covariance of two Fisher-transformed correlations

$$\varphi_{jk,hm} = \left(\frac{n}{n-3}\right)\frac{\sigma_{r_{jk},r_{hm}}}{\left(1-\rho_{jk}^2\right)\left(1-\rho_{hm}^2\right)} \quad (4.6)$$

The variance of the Fisher-transformed correlations is, as we noted above, $1/(n-3)$. In practice, if we were testing the hypothesis that $\rho_{jk} = \rho_{hm}$, we would incorporate the null hypothesis into the denominator of the test statistic. We would do this by obtaining a pooled ("ordinary least squares") estimate of ρ_{jk} and ρ_{hm} by averaging r_{jk} and r_{hm}. We would then obtain an estimate of $\varphi_{jk,hm}$ by substituting sample correlations in Equations 4.5 and 4.6, except that the pooled estimate would be inserted in place of ρ_{jk} and ρ_{hm}. The resulting statistic, Equation 15 in my 1980 *Psychological Bulletin* article, is illustrated with examples on page 249 of that article. The article discusses a wide range of statistics for comparing correlation coefficients.

4.3 ADF Tests

One aspect of correlational tests which receives virtually no coverage in standard textbooks is the robustness of the standard procedures to violations of assumptions. In particular, although standard “Normal Theory” (NT) correlational tests are *not* robust to violations of the assumption of multivariate normality, one finds no mention at all of this fact in most texts! This is surprising, because most books *do* give prominent mention to the fact that the Student's *t* tests are robust to violations of the assumption of normality, and mention rather routinely the non-robustness of some of the traditional tests for comparing variances.

A variety of confusing, and sometimes conflicting reports about the “robustness of the Pearson *r*” have appeared over the years in places like *Psychological Bulletin*. The confusion in these articles stemmed, apparently, from the reliance by some authors on Monte Carlo methods. The basic facts about the effect of non-normality on the distribution of *r* are easily inferred from the asymptotic distribution theory. This theory is summarized by Steiger & Hakstian (1982, *British Journal of Mathematical and Statistical Psychology*). Here are the key points:

- Generally, the further ρ is from zero, the less robust the NT statistical test will be. If ρ is precisely zero, the test will be very robust.
- The effect of non-normality is a complicated effect of 4th order moments of the multivariate distribution. Where a single correlation coefficient is involved, the *kurtosis* of the marginal variates is the most important factor. The primary effect of kurtosis is on the variance of the correlation coefficient.
- Skewness, by itself, has relatively little effect. Monte Carlo studies which have appeared to determine otherwise have frequently confounded skewness with kurtosis, i.e., have chosen a “skewed” distribution which is also leptokurtic.

It is possible to correct NT tests for kurtosis, by estimating the fourth order moment structure from sample data, and, using formulae in Steiger and Hakstian (1982), correcting the NT test statistic. These corrected tests, often termed “Asymptotically Distribution Free,” or ADF tests, apparently work reasonably well with moderate to large sample sizes.