Given $s_{b_1-b_2}$, we can now solve for t:

$$t = \frac{b_1 - b_2}{s_{b_1-b_2}} = \frac{(-0.40) - (-0.20)}{0.192} = -1.04$$

on 198 *df*. Because $t_{.025}(198) = \pm 1.97$, we would fail to reject H_0 and would therefore conclude that we have no reason to doubt that life expectancy decreases as a function of smoking at the same rate for males as for females.

It is worth noting that although $H_0: b^* = 0$ is equivalent to $H_0: \rho = 0$, it does not follow that $H_0: b_1^* - b_2^* = 0$ is equivalent to $H_0: \rho_1 - \rho_2 = 0$. If you think about it for a moment, it should be apparent that two scatter diagrams could have the same regression line ($b_1^* = b_2^*$) but different degrees of scatter around that line, hence, $\rho_1 \neq \rho_2$. The reverse also holds—two different regression lines could fit their respective sets of data equally well.

Testing the Difference Between Two Independent *r*s

When we test the difference between two independent *r*s, a minor difficulty arises. When $\rho \neq 0$, the sampling distribution of r is not approximately normal (it becomes more and more skewed as $\rho \Rightarrow \pm 1.00$), and its standard error is not easily estimated. The same holds for the difference $r_1 - r_2$. This raises an obvious problem because, as you can imagine, we will need to know the standard error of a difference between correlations if we are to create a t test on that difference. Fortunately, R. A. Fisher provided the solution.

Fisher (1921) showed that if we transform r to

$$r' = (0.5) \log_e \left| \frac{1 + r}{1 - r} \right|$$

then r' is approximately normally distributed around ρ' (the transformed value of ρ) with standard error

$$s_{r'} = \frac{1}{\sqrt{N - 3}}$$

(Fisher labeled his statistic "z," but "r'" is often used to avoid confusion with the standard normal deviate.) Because we know the standard error, we can now test the null hypothesis that $\rho_1 - \rho_2 = 0$ by converting each r to r' and solving for

$$z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

Note that our test statistic is z rather than t because our standard error does not rely on statistics computed from the sample (other than N) and is therefore a parameter.

Appendix r' tabulates the values of r' for different values of r, which eliminates the need to solve the equation for r'.

To take a simple example, assume that for a sample of 53 males, the correlation between number of packs of cigarettes smoked per day and life expectancy was .50. For a sample of 43 females, the correlation was .40. (These are unrealistically high values for r, but they better illustrate the effects of the transformation.) The question of interest is, Are these two coefficients significantly different, or are the differences in line with what we would expect when sampling from the same bivariate population of X, Y pairs?

	Males	Females
r	.50	.40
r'	.549	.424
N	53	53

$$z = \frac{.549 - .424}{\sqrt{\frac{1}{53-3} + \frac{1}{53-3}}} = \frac{.125}{\sqrt{\frac{2}{50}}} = \frac{.125}{\frac{1}{5}} = 0.625$$

Because $z_{obt} = 0.625$ is less than $z_{.025} = \pm 1.96$, we fail to reject H_0 and conclude that with a two-tailed test at $\alpha = .05$, we have no reason to doubt that the correlation between smoking and life expectancy is the same for males as it is for females.

Testing the Hypothesis That ρ Equals Any Specified Value

Now that we have discussed the concept of r', we are in a position to test the null hypothesis that ρ is equal to any value, not just to zero. You probably can't think of many situations in which you would like to do that, and neither can I. But the ability to do so allows us to establish confidence limits on ρ, a more interesting procedure.

As we have seen, for any value of ρ, the sampling distribution of r' is approximately normally distributed around ρ′ (the transformed value of ρ) with a standard error of $\frac{1}{\sqrt{N-3}}$. From this it follows that

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{N-3}}}$$

is a standard normal deviate. Thus, if we want to test the null hypothesis that a sample r of .30 (with $N = 103$) came from a population where $\rho = .50$, we proceed as follows:

$$r = .30 \qquad r' = .310$$

$$\rho = .50 \qquad \rho' = .549$$

$$N = 103 \qquad s_{r'} = 1/\sqrt{N-3} = 0.10$$

$$z = \frac{.310 - .549}{0.10} = -0.239/0.10 = -2.39$$

Because $z_{obt} = -2.39$ is more extreme than $z_{.025} = \pm 1.96$, we reject H_0 at $\alpha = .05$ (two-tailed) and conclude that our sample did not come from a population where $\rho = .50$.

Confidence Limits on ρ

We can easily establish confidence limits on ρ by solving the previous equation for ρ instead of z. To do this, we first solve for confidence limits on ρ′, and then convert ρ′ to ρ.

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{N-3}}}$$

therefore

$$\sqrt{\frac{1}{N-3}}(\pm z) = r' - \rho'$$

and thus

$$\text{CI}(\rho') = r' \pm z_{\alpha/2}\sqrt{\frac{1}{N-3}}$$

For our stress example, $r = .506$ ($r' = .557$) and $N = 107$, so the 95% confidence limits are

$$\begin{aligned}\text{CI}(\rho') &= .557 \pm 1.96\sqrt{\frac{1}{104}} \\ &= .557 \pm 1.96(0.098) = .557 \pm 0.192 \\ &= .365 \leq \rho' \leq .749\end{aligned}$$

Converting from ρ' back to ρ and rounding,

$$.350 \leq \rho \leq .635$$

Thus, the limits are $\rho = .350$ and $\rho = .635$. The probability is .95 that limits obtained in this way encompass the true value of ρ. Note that $\rho = 0$ is not included within our limits, thus offering a simultaneous test of H_0: $\rho = 0$, should we be interested in that information.

Confidence Limits versus Tests of Significance

At least in the behavioral sciences, most textbooks, courses, and published research have focused on tests of significance, and paid scant attention to confidence limits. In some cases that really is probably appropriate, but in other cases, it leaves the reader short.

In this chapter, we have repeatedly referred to an example on stress and psychological symptoms. For the first few people who investigated this issue, it really was an important question whether there was a significant relationship between these two variables. But now that everyone believes it, a more appropriate question becomes how large the relationship is. And for that question, a suitable answer is provided by a statement such as the correlation between the two variables was .506, with a 95% confidence interval of $.350 \leq \rho \leq .635$. (A comparable statement from the public opinion polling field would be something like $r = .506$ with a 95% margin of error of $\pm .15$(approx.).[13]

Testing the Difference Between Two Nonindependent *r*s

Occasionally, we come across a situation in which we want to test the difference between two correlations that are not independent. (I am probably asked this question a couple of times per year.) One case arises when two correlations share one variable in common. We will see such an example later. Another case arises when we correlate two variables at Time 1 and then again at some later point (Time 2), and we want to ask whether there has been a significant change in the correlation over time. I will not cover that case, but a very good discussion of that particular issue can be found at core.ecu.edu/psyc/wuenschk/StatHelp/ZPF.doc and in a paper by Raghunathan, Rosenthal, and Rubin (1996).

As an example of correlations that share a common variable, Reilly, Drudge, Rosen, Loew, and Fischer (1985) administered two intelligence tests (the WISC-R and the McCarthy) to first-grade children, and then administered the Wide Range Achievement Test (WRAT) to

[13] I had to insert the label "approx." here because the limits, as we saw earlier, are not exactly symmetrical around *r*.

those same children 2 years later. Reilly et al. obtained, among other findings, the following correlations:

	WRAT	WISC-R	McCarthy
WRAT	1.00	.80	.72
WISC-R		1.00	.89
McCarthy			1.00

Note that the WISC-R and the McCarthy are highly correlated but that the WISC-R correlates somewhat more highly with the WRAT (reading) than does the McCarthy. It is of interest to ask whether this difference between the WISC-R–WRAT correlation (.80) and the McCarthy–WRAT correlation (.72) is significant, but to answer that question requires a test on nonindependent correlations because they both have the WRAT in common and they are based on the same sample.

When we have two correlations that are not independent—as these are not, because the tests were based on the same 26 children—we must take into account this lack of independence. Specifically, we must incorporate a term representing the degree to which the two tests are themselves correlated. Hotelling (1931) proposed the traditional solution, but a better test was devised by Williams (1959) and endorsed by Steiger (1980). This latter test takes the form

$$t = (r_{12} - r_{13}) \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}$$

where

$$|R| = \left(1 - r_{12}^2 - r_{13}^2 - r_{23}^2\right) + (2r_{12}r_{13}r_{23})$$

This ratio is distributed as t on $N - 3$ *df*. In this equation, r_{12} and r_{13} refer to the correlation coefficients whose difference is to be tested, and r_{23} refers to the correlation between the two predictors. $|R|$ is the determinant of the 3×3 matrix of intercorrelations, but you can calculate it as shown without knowing anything about determinants.

For our example, let

r_{12} = correlation between the WISC-R and the WRAT = .80

r_{13} = correlation between the McCarthy and the WRAT = .72

r_{23} = correlation between the WISC-R and the McCarthy = .89

$N = 26$

then

$$|R| = (1 - .80^2 - .72^2 - .89^2) + (2)(.80)(.72)(.89) = .075$$

$$t = (.80 - .72) \sqrt{\frac{(25)(1+.89)}{2\left(\frac{25}{23}\right)(.075) + \frac{(.80+.72)^2}{4}(1-.89)^3}}$$

$$= 1.36$$

A value of $t_{obt} = 1.36$ on 23 *df* is not significant. Although this does not prove the argument that the tests are equally effective in predicting third-grade children's performance on the reading scale of the WRAT, because you cannot prove the null hypothesis, it is consistent with that argument and thus supports it.