Covariance and Correlation

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Introduction

- In this module, we discuss a pair of extremely important statistical concepts — *covariance* and *correlation*.
- We begin by defining covariance, and then extend the concept to a special kind of covariance known as correlation.

Bivariate Distributions and Scatterplots

- Through most of the course, we have dealt with data sets in which each person (or, more generally, *unit of observation*) was represented by a score on just one variable.
- In the module on the correlated sample t test, we extended our work to cover two repeated measures on the same individuals.
- When we have two measures on the same individuals, it is common to plot each individual's data in a two dimensional plot called a *scatterplot*.
- The scatterplot often allows us to see a functional relationship between the two variables.

Bivariate Distributions and Covariance

- Here's a question that you've thought of informally, but probably have never been tempted to assess quantitatively: "What is the relationship between shoe size and height?"
- We'll examine the question with a data set from an article by Constance McLaren in the 2012 *Journal of Statistics Education.*

Bivariate Distributions and Covariance

- The data file is available in several places on the course website. You may download the file by right-clicking on it (it is next to the lecture slides).
- These data were gathered from a group of volunteer students in a business statistics course.
- If you place it in your working directory, you can then load it with the command

```
> all.heights <- read.csv("shoesize.csv")
```

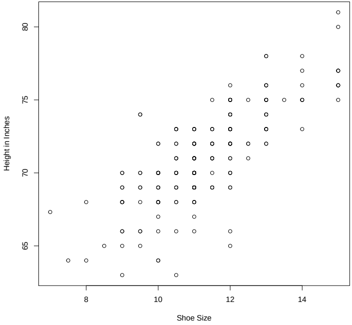
- Alternatively, you can download directly from a web repository with the command

```
> all.heights <- read.csv(
+   "http://www.statpower.net/R2101/shoesize.csv")
```

Bivariate Distributions and Scatterplots

- Here is the scatterplot for the male data.

```
> male.data <- all.heights[all.heights$Gender=="M",] #Select males
> attach(male.data)#Make Variables Available
> # Draw scatterplot
> plot(Size,Height,xlab="Shoe Size",ylab="Height in Inches")
```



Bivariate Distributions and Scatterplots

- This scatterplot shows a clear connection between shoe size and height.
- Traditionally, the variable to be predicted (the dependent variable) is plotted on the vertical axis, while the variable to be predicted from (the independednt variable) is plotted on the horizontal axis.
- Note that, because height is measured only to the nearest inch, and shoe size to the nearest half-size, a number of points overlap. The scaterplot indicates this by making some points darker than others.
- But how can we characterize this relationship accurately?
- We notice that shoe size and height vary together.
- A statistician might say they "covary."
- This notion is operationalized in a statistic called *covariance.*

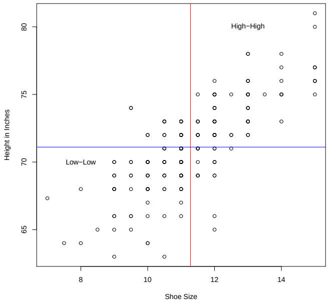
Bivariate Distributions and Scatterplots

- Let's compute the average height and shoe size, and then draw lines of demarcation on the scatterplot.

```
> mean(Height)
[1] 71.10552
> mean(Size)
[1] 11.28054
```

Bivariate Distributions and Scatterplots

```
> plot(Size,Height,xlab="Shoe Size",ylab="Height in Inches")
> abline(v=mean(Size),col="red")
> abline(h=mean(Height),col="blue")
> text(13,80,"High-High")
> text(8,70,"Low-Low")
```



Bivariate Distributions and Scatterplots

- The upper right ("High-High") quadrant of the plot represents men whose heights and shoe sizes were both above average.
- The lower left ("Low-Low") quadrant of the plot represents men whose heights and shoe sizes were both below average.
- Notice that there are far more data points in these two quadrants than in the other two: This is because, when there is a direct (positive) relationship between two variables, the scores tend to be on the same sides of their respective means.
- On the other hand, when there is an inverse (negative) relationship between two variables, the scores tend to be on the opposite sides of their respective means.
- This fact is behind the statistic we call *covariance*.

Covariance
The Concept

- What is *covariance*?
- We convert each variable into deviation score form by subtracting the respective means.
- If scores tend to be on the same sides of their respective means, then
  1. Positive deviations will tend to be matched with positive deviations, and
  2. Negative deviations will tend to be matched with negative deviations
- To capture this trend, we sum the cross-product of the deviation scores, then divide by $n - 1$.
- So, essentially, the sample covariance between X and Y is an estimate of the average cross-product of deviation scores in the population.

Covariance
Computations

- The *sample covariance of X and Y* is defined as

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_x)(Y_i - M_y) \tag{1}$$

- An alternate, more computationally convenient formula, is

$$s_{x,y} = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} \right) \tag{2}$$

- An important fact is that *the variance of a variable is its covariance with itself*, that is, if we substitute x for y in Equation 1, we obtain

$$s_x^2 = s_{x,x} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_x)(X_i - M_x) \tag{3}$$

Covariance
Computations

- Computing the covariance between two variables "by hand" is tedious though straightforward and, not surprisingly (because the variance of a variable *is* a covariance), follows much the same path as computation of a variance:
    1. If the data are very simple, and especially if n is small and the sample mean a simple number, one can convert X and Y scores to deviation score form and use Equation 1.
    2. More generally, one can compute $\sum X$, $\sum Y$, $\sum XY$, and n and use Equation 2.

Covariance
Computations

Example (Computing Covariance)

Suppose you were interested in examining the relationship between cigarette smoking and lung capacity. You asked 5 people how many cigarettes they smoke in an average day, and you then measure their lung capacities, which are corrected for age, height, weight, and gender. Here are the data:

	Cigarettes	Lung.Capacity
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

(... Continued on the next slide)

Covariance
Computations

Example (Computing Covariance)

In this case, it is easy to compute the mean for both Cigarettes (X) and Lung Capacity (Y), i.e., $M_{cigarettes} = M_x = 10$, $M_{lung.capacity} = M_y = 36$, then convert to deviation scores and use Equation 1 as shown below:

	X	dX	dXdY	dY	Y	XY
1	0	-10	-90	9	45	0
2	5	-5	-30	6	42	210
3	10	0	0	-3	33	330
4	15	5	-25	-5	31	465
5	20	10	-70	-7	29	580

The sum of the dXdY column is -225, and then compute the covariance as

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n} dX_i dY_i = \frac{-215}{4} = -53.75$$

(...Continued on the next slide)

Covariance
Computations

Example (Computing Covariance)

Alternatively, one might compute $\sum X = 50$, $\sum Y = 180$, $\sum XY = 1585$, and n, and use Equation 2.

$$
\begin{aligned}
s_{x,y} &= \frac{1}{n-1}\left(\sum XY - \frac{\sum X \sum Y}{n}\right) \\
&= \frac{1}{5-1}\left(\sum 1585 - \frac{50 \times 180}{5}\right) \\
&= \frac{1}{4}\left(\sum 1585 - \frac{9000}{5}\right) \\
&= \frac{1}{4}\left(\sum 1585 - 1800\right) \\
&= \frac{1}{4}(-215) \\
&= -53.75
\end{aligned}
$$

Of course, there is a much easier way, using R. (...Continued on the next slide)

Covariance

Computations

Example (Computing Covariance)

Here is how to compute covariance using R's `cov` command. In the case of really simple textbook examples, you can copy the numbers right off the screen and enter them into R, using the following approach.

```
> Cigarettes <- c(0,5,10,15,20)
> Lung.Capacity <- c(45,42,33,31,29)
> cov(Cigarettes,Lung.Capacity)
[1] -53.75
```

Covariance
Limitations

- Covariance is an extremely important concept in advanced statistics.
- Indeed, there is a statistical method called *Analysis of Covariance Structures* that is one of the most widely used methodologies in Psychology and Education.
- However, in its ability to convey information about the nature of a relationship between two variables, covariance is not particularly useful as a single descriptive statistic, and is not discussed much in elementary textbooks.
- What is the problem with covariance?

Covariance

Limitations

- We saw that the covariance between smoking and lung capacity in our tiny sample is -53.75.
- The problem is, this statistic is not invariant under a change of scale.
- As a measure on deviation scores, we know that adding or subtracting a constant from every X or every Y will not change the covariance between X and Y.
- However, multiplying every X or Y by a constant will multiply the covariance by that constant.
- It is easy to see that from the covariance formula, because if you multiply every raw score by a constant, you multiply the corresponding deviation score by that same constant.
- We can also verify that in R. Suppose we change the smoking measure to packs per day instead of cigarettes per day by dividing X by 20. This will divide the covariance by 20.

Covariance
Limitations

- Here is the R calculation:

```
> cov(Cigarettes, Lung.Capacity)
[1] -53.75
> cov(Cigarettes, Lung.Capacity) / 20
[1] -2.6875
> cov(Cigarettes/20,Lung.Capacity)
[1] -2.6875
```

- The problem, in a nutshell, is that the sign of a covariance tells you whether the relationship is positive or negative, but the absolute value is, in a sense, "polluted by the metric of the numbers."
- Depending on the scale of the data, the absolute value of the covariance can be very large or very small.
- So how can we fix this?
- Easy — we take the metric out of the numbers.
- How do we do that?

The (Pearson) Correlation Coefficient

Definition

- To take the metric out of covariance, we compute it on the Z-scores instead of the deviation scores. (Remember that Z-scores *are also deviation scores*, but they have the standard deviation divided out.)
- The sample correlation coefficient $r_{x,y}$, sometimes called the Pearson correlation, but generally referred to as "the correlation" is simply the sum of cross-products of Z-scores divided by $n-1$:

$$r_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} Zx_i Zy_i \tag{4}$$

- The population correlation $\rho_{x,y}$ is the average cross-product of Z-scores for the two variables.

The (Pearson) Correlation Coefficient

Definition

- One may also define the correlation in terms of the covariance, i.e.,

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} \tag{5}$$

- Equation 5 shows us that we may think of a correlation coefficient as a covariance with the standard deviations factored out.
- Alternatively, since we may turn the equation around and write

$$s_{x,y} = r_{x,y} s_x s_y \tag{6}$$

we may think of a covariance as a correlation with the standard deviations put back in.

The (Pearson) Correlation Coefficient

Computing the Correlation

- Most textbooks give computational formulas for the correlation coefficient. This is probably the most common version.

$$r_{x,y} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \tag{7}$$

If we compute the quantities n,$\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, $\sum XY$, and substitute them into Equation 7, we can calculate the correlation as shown on the next slide.

The (Pearson) Correlation Coefficient

Computing the Correlation

Example (Computing a Correlation)

$$
\begin{aligned}
r_{xy} &= \frac{(5)(1585) - (50)(180)}{\sqrt{\left[(5)(750) - 50^2\right]\left[(5)(6680) - 180^2\right]}} \\
&= \frac{7925 - 9000}{\sqrt{(3750 - 2500)(33400 - 32400)}} \\
&= \frac{-1075}{\sqrt{(1250)\,(1000)}} \\
&= -.9615
\end{aligned}
$$

(Continued on the next slide . . .)

The (Pearson) Correlation Coefficient

Computing the Correlation

Example (Computing a Correlation)

In general, you should *never* compute a correlation by hand if you can possibly avoid it. If n is more than a very small number, your chances of successfully computing the correlation would not be that high. Better to use R.Computing a correlation with R is very simple. If the data are in two variables, you just type

```
> cor(Cigarettes,Lung.Capacity)
[1] -0.9615092
```

By the way, the correlation between height and shoe size in our example data set is

```
> cor(Size,Height)
[1] 0.7677094
```

The (Pearson) Correlation Coefficient

Interpreting a Correlation

- What does a correlation coefficient *mean*? How do we interpret it?
- There are many answers to this. There are more than a dozen different ways of viewing a correlation. Professor Joe Rodgers in our department co-authored an article on the subject titled *Thirteen Ways to Look at the Correlation Coefficient*.
- We'll stick with the basics here.

The (Pearson) Correlation Coefficient

Interpreting a Correlation

- There are three fundamental aspects of a correlation:
  1. *The sign.* A positive sign indicates a direct (positive) relationship, a negative sign indicates an inverse (negative) relationship.
  2. *The absolute value.* As the absolute value approaches 1, the data points in the scatterplot get closer and closer to falling in a straight line, indicating a strong linear relationship. So the absolute value is an indicator of the strength of the linear relationship between the variables.
  3. *The square of the correlation.* $r^2_{x,y}$ can be interpreted as the "proportion of the variance of Y accounted for by X."

The (Pearson) Correlation Coefficient

Interpreting a Correlation

Example (Interpreting a Correlation)

Suppose $r_{x,y} = 0.50$ in one study, and $r_{a,b} = -.55$ in another. What do these statistics tell us?

Answer. They tell us that the relationship between X and Y in the first study is positive, while that between A and B in the second study is negative. However, the linear relationship is actually slightly stronger between A and B than it is between X and Y.

The (Pearson) Correlation Coefficient

Interpreting a Correlation

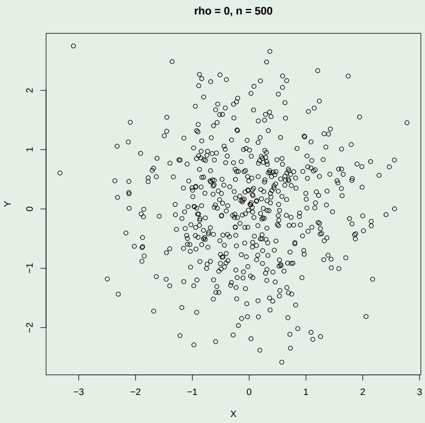
Example (Some Typical Scatterplots)

Let's examine some bivariate normal scatterplots in which the data come from populations with means of 0 and variances of 1. These will give you a feel for how correlations are reflected in a scatterplot.

The (Pearson) Correlation Coefficient

Interpreting a Correlation

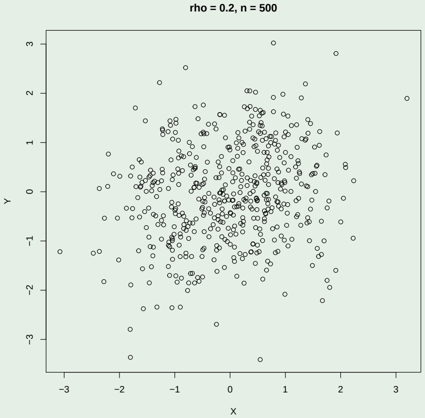
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

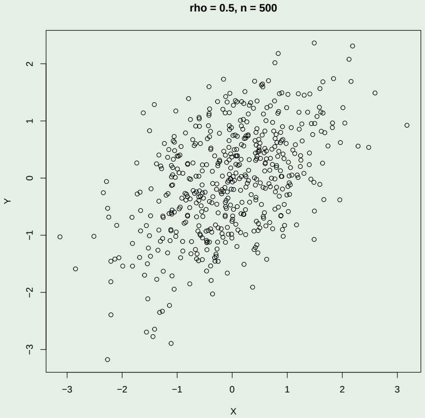
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

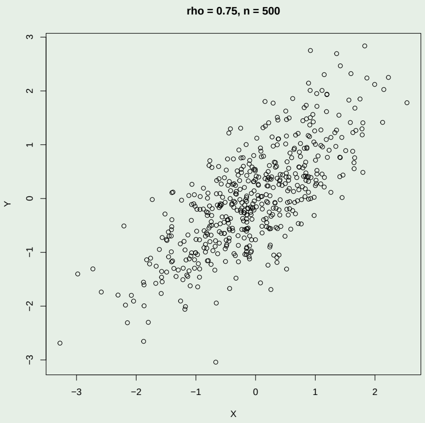
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

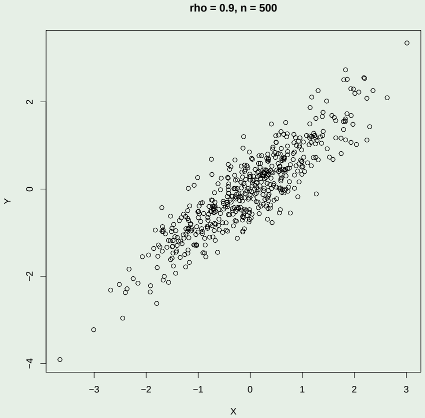
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

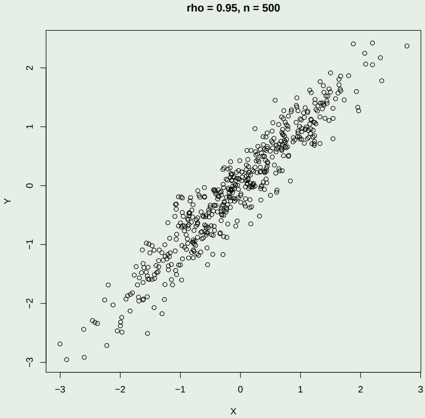
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

Example (Some Typical Scatterplots)



Some Other Correlation Coefficients

Introduction

- The Pearson correlation coefficient is by far the most commonly computed measure of relationship between two variables.
- If someone refers to "the correlation between X and Y," they are almost certainly referring to the Pearson correlation unless some other coefficient has been specified.
- In this section, we review the other commonly employed correlation coefficients that are discussed in your text.

Some Other Correlation Coefficients

The Spearman Rank-Order Correlation

- In a situation in which the data are only ordinal, or in which there are severe outliers that strongly affect a correlation, the Spearman rank-order correlation can be very useful.
- Recall that, when data are merely ordinal, *any monotonic increasing function* can be applied to the data without destroying the ordinal information.
- When the data are ordinal, converting to ranks reduces the extraneous (and meaningless) information in the data, and reduces the data to its essentials.
- In general, computers are extremely fast at sorting data and converting them to ranks. The only complication occurs if two scores are tied. What do you do then?
- The common solution is this: If two or more scores are tied, you assign to each of the tied scores the *arithmetic average* of the ranks that the scores would have received had they not been tied.
- For example, if a set of scores is $3, 3, 4, 7, 7, 7, 9$, the corresponding ranks would be $1.5, 1.5, 3, 5, 5, 5, 7$.

Some Other Correlation Coefficients

The Spearman Rank-Order Correlation

Example (The Spearman Rank-Order Correlation)

```
   x  y rank.x rank.y
1  0 31      1    1.0
2  5 40      2    4.0
3 10 33      3    2.5
4 15 33      4    2.5
5 20 50      5    5.0
```

Given the above data, compute the Pearson correlation *and* the Spearman correlation.

Some Other Correlation Coefficients
The Spearman Rank-Order Correlation

Example (The Spearman Rank-Order Correlation)

Answer.

```
> # The Pearson Correlation is
> cor(x,y)
[1] 0.6260391
> # The Spearman correlation is
> cor(rank.x,rank.y)
[1] 0.6668859
> # However, R will do all the work for you!!
> cor(rank.x,rank.y,method="spearman")
[1] 0.6668859
```

Some Other Correlation Coefficients
The Phi Coefficient

- In some situations, the data can be reduced to a binary variable.
- Examples are True-False, Pass-Fail, Alive-Dead, Male-Female, Experimental-Control.
- If *both* variables are reduced to $0 - 1$ binary variables, then the Pearson correlation between the resulting variables is called a *Phi Coefficient*.

Some Other Correlation Coefficients

The Phi Coefficient

Example (The Phi Coefficient)

In this example, a random sample of participants is obtained, and each individual is classified in terms of birth-order position as first-born versus later-born. Then, each individual's personality is classified as either introvert or extrovert. Here are the resulting data from Gravetter and Walnau. Notice how the original data on birth order are dichotomized into a $0-1$ variable:

Some Other Correlation Coefficients
The Phi Coefficient

Example (The Phi Coefficient)

Original Data		Converted Scores	
Birth Order (X)	*Personality (Y)*	*Birth Order (X)*	*Personality (Y)*
1st	Introvert	0	0
3rd	Extrovert	1	1
1st	Extrovert	0	1
2nd	Extrovert	1	1
4th	Extrovert	1	1
2nd	Introvert	1	0
1st	Introvert	0	0
3rd	Extrovert	1	1

Some Other Correlation Coefficients

The Phi Coefficient

Example (The Phi Coefficient)

To process the problem in R, we simply enter the $0-1$ data for each variable and compute the Pearson correlation with the `cor` function.

```
> x <- c(0,1,0,1,1,1,0,1)
> y <- c(0,1,1,1,1,0,0,1)
> cor(x,y)
[1] 0.4666667
```

Some Other Correlation Coefficients
The Point-Biserial Correlation

- If only one of the two variables is a binary $0-1$ variable, and the other is a variable measured on an interval scale of measurement, then the Pearson correlation coefficient calculated on the two variables is known as a *point-biserial correlation.*
- We already encountered this correlation when discussing measures of effect size in connection with the two-sample, independent sample t-test.
- Recall that the relationship between the *coefficient of determination* r^2 and the two-sample t statistic is

$$r^2 = \frac{t^2}{t^2 + df}$$

Some Other Correlation Coefficients
The Point-Biserial Correlation

Example (The Point-Biserial Correlation)

Consider the following data:

	Group	Score
1	1	5
2	1	7
3	1	3
4	1	11
5	1	7
6	0	14
7	0	14
8	0	20
9	0	15
10	0	16

Some Other Correlation Coefficients
The Point-Biserial Correlation

Example (The Point-Biserial Correlation)

In this case, the Experimental Group (Group $= 1$) has a mean Score of 6.6 and a standard deviation of 2.9665.

The Control Group (Group $= 0$) has a mean Score of 15.8 and a standard deviation of 2.49.

Given that both sample sizes are $n = 5$, we can use the simplified formula for the 2-sample t, or load the R routine from the course website. I'll take the easy way out and use the routine from the website.

Some Other Correlation Coefficients

The Point-Biserial Correlation

Example (The Point-Biserial Correlation)

```
> results <- t.2.sample(6.6,15.8,2.9665,2.49,5,
+                        5,alpha=0.05,tails=2)
> results
$t.statistic
[1] -5.311583

$df
[1] 8

$alpha
[1] 0.05

$critical.t.values
[1] -2.306004  2.306004
```

Some Other Correlation Coefficients

The Point-Biserial Correlation

Example (The Point-Biserial Correlation)

The t statistic is way beyond the rejection point. But what is the effect size. We can compute it directly in R as

```
> ## Grab t-statistic from results
> ##   of previous calculation
> t <- results$t.statistic
> df <- results$df
> ## Compute the r.squared
> t^2/(t^2 + df)
[1] 0.7790843
```

Of course, we can get the same result by computing the point-biserial correlation as the ordinary Pearson correlation between `Group` and `Score` and then squaring it. The slight difference in results is due to my rounding off some of the statistics input to the t routine.

```
> r <- cor(Group,Score)
> r^2
[1] 0.7790869
```

Significance Test for r

- To test whether Pearson correlation r is significantly different from zero, use the following t statistic, which has $n-2$ degrees of freedom. Of course, the statistical null hypothesis is that the population correlation $\rho = 0$.

$$t_{n-2} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \tag{8}$$

Significance Test for r

Example

Suppose you observe a correlation coefficient of 0.2371 with a sample of $n = 93$. Can you reject the null hypothesis that $\rho = 0$? Use $\alpha = 0.05$.

Significance Test for r

Example

Answer. We compute the t statistic with R.

```
> df <- 93 - 2
> t <- sqrt(df)*0.2371 / sqrt(1-0.2371^2)
> t
[1] 2.328177
> df
[1] 91
> t.crit <- qt(0.975,df)
> t.crit
[1] 1.986377
```

Since the observed t exceeds the critical value, we can reject the null hypothesis and declare the correlation statistically significant at the 0.05 level, two-tailed.