Bivariate Relationships, Covariance and Correlation (Part Un)
or ...

$r$ You Gonna $\rho$ My Way?

Ryne Estabrook

Department of Psychology
University of Virginia

October 1, 2008
PSYC 771
Housekeeping

- Lab is tonight.
- Midterm is in two weeks. A good chunk of the grading is done.
- Monica Erbacher is giving DADA this week, dealing with an application of Item Response Theory to longitudinal Likert-style data.
Previously on...
Deep down, you love this stuff.

- Previously on...
  - Surprisingness
  - Describing Data & Distributions (particularly, the normal and \( t \))
  - Null Hypothesis Testing, T-tests, Power.

- Today: Bivariate Relationships
Bivariate Relationships

- To this point, we’ve relied on two relatively simple presentations of data.
- Univariate Data:
  - Measures of central tendency and spread.
  - The graphical tools we’ve used so far have displayed only univariate data.
  - The distributions we’ve talked about thus far have all described only a single variable.
- “Group” Data:
  - We have one continuous dependent variable, and some binary predictor or group variable.
  - Really, we’ve been comparing a univariate statistic from one group to the same statistic from another group.
Bivariate Relationships

- It is very unlikely that binary group differences in univariate data is the most sophisticated tool that you’ll need.
- As your questions and accompanying data grow more complex, you’ll need multivariate statistics to describe them.
  - I’m using multivariate to describe multiple variables of any type, rather than multivariate dependent variables as it is often used.
- The first step to multivariate data is bivariate or two-variable data.
  - The first tool we’ll use to describe bivariate data is the scatter plot.
Scatter Plot
Our First Scatter–Plot
Bivariate Relationships

- A plot is the simplest way to visually describe bivariate data.
  - For every observation, we have a value for variable 1 and a value for variable 2. Those values become the (Cartesian) coordinates in our plot.
- Note that which variable takes the horizontal axis and which one the vertical does not matter.
- Unlike t-tests and future analyses in this course, the topics of today’s lecture won’t differentiate between independent and dependent variables.
- We’re describing relationships.
As nice as scatter plots are, we can’t use them exclusively.
We use descriptive statistics to summarize vast amounts of (univariate) data, but we need need a bivariate version of those statistics.
Typically, we talk about bivariate relationships by indicating the strength of the association.
The simplest way is to describe relationships is in terms of linear relationships.
How about a slope?
Let’s Try A Slope

Our First Scatter-Plot
Let’s Try A Slope
Let's Try A Slope

Our First Scatter-Plot
Let’s Try A Slope
Covariance

- Slope doesn’t seem to work, because it’s dependent on which way we draw the graph.
- If we’re looking for a statistic, what if we looked to variance as an example?
- By using two variables to compute variance instead of one, we create a statistic called **covariance**.

\[
\text{Var}(x) = \sigma_x^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{\sum(x - \bar{x})(x - \bar{x})}{n-1}
\]

\[
\text{Var}(y) = \sigma_y^2 = \frac{\sum(y - \bar{y})^2}{n-1} = \frac{\sum(y - \bar{y})(y - \bar{y})}{n-1}
\]

\[
\text{Cov}(x, y) = \sigma_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n-1}
\]
Covariance

What are the properties of a covariance?

- This solved our slope problem, as it is equivalent regardless of the order of the variables.
  - $\text{Cov}(x,y) = \text{Cov}(y,x)$.
- As a matter of fact, the slopes we got in the examples before were directly related to the covariance and the variance of the variable on the horizontal ($x$) axis.
  - $\text{Cov}(V_1, V_2) = 1.2$, $\text{Var}(V_1) = 1$, $\text{Var}(V_2) = 4$.

\[
\text{Slope} = \frac{\Delta Y}{X} = \frac{\text{Cov}(x,y)}{\text{Var}(x)}
\]
\[
\text{Slope}_{X=\text{Var}_1} = \frac{\text{Cov}(V_1, V_2)}{\text{Var}(V_1)} = \frac{1.2}{1.0} = 1.20
\]
\[
\text{Slope}_{X=\text{Var}_2} = \frac{\text{Cov}(V_1, V_2)}{\text{Var}(V_2)} = \frac{1.2}{4.0} = 0.30
\]
Covariance

What are the properties of a covariance?

▶ Variances are just special kinds of covariances.
  ▶ $\text{Cov}(X,X) = \text{Var}(X)$. That’s actually pretty useful.
▶ It can be positive or negative, and that’s a good thing!
  ▶ Variances are restricted to $[0, \infty)$, but covariances have no restrictions.
  ▶ Positive covariances mean higher values of $x$ are associated with higher values of $y$.
  ▶ Negative covariances mean higher values of $x$ are associated with lower values of $y$.
  ▶ Covariances of zero mean no relationship (i.e. slopes of zero).
Covariance

What are the properties of a covariance?

- Covariances have units, but they’re not terribly useful.
  - Covariances are the mean product of $x$ and $y$.
  - Variances have equally obtuse units.
- For by hand or matrix calculation, you can calculate covariances another way:

$$\text{Cov}(x, y) = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N - 1}$$
Covariance

 Degrees of Freedom Question

\[ \text{Cov}(x, y) = \sigma_{xy} = \frac{(x - \bar{x})(y - \bar{y})}{n - 1} \]

- Why \( n-1 \) and not \( n-2 \)?
  - The means of \( x \) and \( y \) can be thought of as a single joint mean (vector) of length 2.
  - The covariance is thought of as the mean cross-product, centered on a vector \([\text{mean}(x), \text{mean}(y)]\).
  - If this means nothing to you, that’s fine.

- For a number of techniques based on these formulas, having variances and covariances use the same denominator is very handy.
  - We’ll discuss a sampling correction a little later.
Problems with Covariance

What Else You Got?

- The units suck.
- There isn’t a standard deviation analog.
  - We’re already working in “raw" units, just complex raw units.
- The scale is dependent on the variances of the two input variables.
  - A covariance of 1 could be a very strong or very weak relationship.
  - You have to interpret it in terms of the variances of your two variables.
What if we scaled the covariance in terms of the variances of both the input variables?

\[ r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \]

This scaled covariance is known as the Pearson Product-Moment Correlation, and is what most people talk about when they say “correlation.”

\[ r_{V1V2} = \frac{\text{Cov}(V1, V2)}{\sqrt{\text{Var}(V1) \text{Var}(V2)}} = \frac{1.2}{\sqrt{1.0 \times 4.0}} = 0.60 \]
Properties of Correlation

General

- Just like covariance, it can be positive or negative.
  - Zero correlation still means no relationship between the two variables.
- Unlike covariance, it is constrained to be between -1 and 1.
  - This constraint makes it easier to understand strength.
  - This also can lead to misunderstandings: a correlation of .50 does not mean a 50% relationship.
  - A perfect (±1) correlation indicates a perfect relationship; if you know one variable, the other variable is determined.
- You can turn it back into covariance when needed.

\[ \text{Cov}(x, y) = r_{xy} \sigma_x \sigma_y \]
You may have been taught correlation as the covariance or mean cross-product of two standardized variables.

While you don’t have to standardize to calculate correlations, they can be a useful way to approach them.

Returning to our slope analog, the correlation is the slope of the best-fitting line between two standardized variables, regardless which one is on which axis.

\[ Cor(z_x, z_y) = \frac{Cov(z_x, z_y)}{\sigma_{z_x}\sigma_{z_y}} = \frac{Cov(z_x, z_y)}{1 \times 1} = Cov(z_x, z_y) \]
Standardized Variables

![Standardized Solutions](image)

The scatter plot shows the standardized solutions for two standard variables. The correlation coefficient, $r = 0.60$, indicates a moderate positive relationship between the variables.
Standardized Variables

Standardized Solutions

$r = 0.60$
Linearity of $r$

- The slopes we’re referring to are the “best-fitting" lines for $x$ and $y$, such that the variance around the line is minimized by definition for either variable.
  - That is, if you picked another slope and calculated the mean squared errors around that line, those MSEs would be higher for any other line you choose.
- Even if you’re not concerned about lines, covariance and correlation has a two strong assumptions:
  - That the relationship in question is linear.
  - That variance is a good way to account for your data.
Linearity
Linearity
Variance Appropriate?

Two Studies

r = .86
Variance Appropriate?
Other Approaches to $r$

- How many of you have heard the terms “sums of squares and cross products?”
  - They’re very useful in calculating ANOVAs by hand.
  - As we’ll get to later, correlation and ANOVA just aren’t that different.
- To calculate SS and CP, you just have to center or subtract the mean from every variable.

\[
\text{Var}(x) = \frac{SS_x}{df_x}
\]
\[
\text{Cov}(x, y) = \frac{CP_{xy}}{df_{xy}}
\]
Other Approaches to $r$

Covariance and Correlation Matrices

Because covariances and correlations are easy to read, you can put them all in a special kind of table called a matrix.

- A matrix is a rectangular table of numbers.
- Sets of matrices can be analyzed using matrix or linear algebra, allowing you to add or multiply tables together.

Covariance and correlation matrices are just tools for looking at a large amount of data all at once.
Other Approaches to $r$

Covariance Matrices

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>Cov(w,w)</td>
<td>Cov(w,x)</td>
<td>Cov(w,y)</td>
<td>Cov(w,z)</td>
</tr>
<tr>
<td>x</td>
<td>Cov(x,w)</td>
<td>Cov(x,x)</td>
<td>Cov(x,y)</td>
<td>Cov(x,z)</td>
</tr>
<tr>
<td>y</td>
<td>Cov(y,w)</td>
<td>Cov(y,x)</td>
<td>Cov(y,y)</td>
<td>Cov(y,z)</td>
</tr>
<tr>
<td>z</td>
<td>Cov(z,w)</td>
<td>Cov(z,x)</td>
<td>Cov(z,y)</td>
<td>Cov(z,z)</td>
</tr>
</tbody>
</table>
Other Approaches to $r$

Covariance Matrices

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>Var(w)</td>
<td>Cov(w,x)</td>
<td>Cov(w,y)</td>
<td>Cov(w,z)</td>
</tr>
<tr>
<td>x</td>
<td>Cov(x,w)</td>
<td>Var(x)</td>
<td>Cov(x,y)</td>
<td>Cov(x,z)</td>
</tr>
<tr>
<td>y</td>
<td>Cov(y,w)</td>
<td>Cov(y,x)</td>
<td>Var(y)</td>
<td>Cov(y,z)</td>
</tr>
<tr>
<td>z</td>
<td>Cov(z,w)</td>
<td>Cov(z,x)</td>
<td>Cov(z,y)</td>
<td>Var(z)</td>
</tr>
</tbody>
</table>
Other Approaches to \( r \)

Correlation Matrices

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>Var(w)</td>
<td>Cov(w,x)</td>
<td>Cov(w,y)</td>
<td>Cov(w,z)</td>
</tr>
<tr>
<td>x</td>
<td>Cov(x,w)</td>
<td>Var(x)</td>
<td>Cov(x,y)</td>
<td>Cov(x,z)</td>
</tr>
<tr>
<td>y</td>
<td>Cov(y,w)</td>
<td>Cov(y,x)</td>
<td>Var(y)</td>
<td>Cov(y,z)</td>
</tr>
<tr>
<td>z</td>
<td>Cov(z,w)</td>
<td>Cov(z,x)</td>
<td>Cov(z,y)</td>
<td>Var(z)</td>
</tr>
</tbody>
</table>

\[
\text{Var}(w) = \text{Var}(x) = \text{Var}(y) = \text{Var}(z) = 1
\]
### Other Approaches to $r$

**Correlation Matricies**

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1</td>
<td>$r_{wx}$</td>
<td>$r_{wy}$</td>
<td>$r_{wz}$</td>
</tr>
<tr>
<td>x</td>
<td>$r_{wx}$</td>
<td>1</td>
<td>$r_{xy}$</td>
<td>$r_{xz}$</td>
</tr>
<tr>
<td>y</td>
<td>$r_{wy}$</td>
<td>$r_{xy}$</td>
<td>1</td>
<td>$r_{yz}$</td>
</tr>
<tr>
<td>z</td>
<td>$r_{wz}$</td>
<td>$r_{xz}$</td>
<td>$r_{yz}$</td>
<td>1</td>
</tr>
</tbody>
</table>
Other Approaches to $r$

Correlation Matrices

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>$r_{wx}$</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>$r_{wy}$</td>
<td>$r_{xy}$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>$r_{wz}$</td>
<td>$r_{xz}$</td>
<td>$r_{yz}$</td>
<td>1</td>
</tr>
</tbody>
</table>
Other Approaches to $r$

**Correlation Matrices**

<table>
<thead>
<tr>
<th>Var</th>
<th>w</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>$r_{wx}$</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>$r_{wy}$</td>
<td>$r_{xy}$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>$r_{wz}$</td>
<td>$r_{xz}$</td>
<td>$r_{yz}$</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>$\bar{w}$</th>
<th>$\bar{x}$</th>
<th>$\bar{y}$</th>
<th>$\bar{z}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>$\sigma_w$</td>
<td>$\sigma_x$</td>
<td>$\sigma_y$</td>
<td>$\sigma_z$</td>
</tr>
</tbody>
</table>
Other Features of $r$

- $r$ can be extended to a multivariate applications:
  - With one variable being predicted by a set of variables, it’s the ANOVA/regression statistic called $R$.
  - With two sets of variables, it’s called canonical correlation.
- With an adjustment, Pearson’s $r$ is an unbiased predictor of the population correlation, called $\rho$.
  - It’s not the covariance that’s the problem, remember that the population standard deviations tend to be underestimates (just like the $t$-test).
  - In multiple correlation extensions, $N-2$ is replaced by $N-p-1$, with $p$ being the number of predictor variables.
  - Strangely, this correction is rarely used in descriptive statistics.

$$\rho = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{(N - 2)}}$$
\( R^2 \) (that is, \( r^2 \)) is a very common and very flexible statistical measure.

Along with Cohen’s \( d \), is one of the two most common measures of effect size.

We’ll talk about it in a few different ways:

- \( R^2 \) as effect size and variance reduction.
- The relationships between \( R^2 \) and other measures.
$R^2$, also known as the coefficient of determination, is discussed as “variance accounted for” as an effect-size estimate.

- The squared correlation coefficient describes the proportion of variance in either variable that is “shared” with the other variable.

$$R^2 = r_{xy}^2 = \frac{(\sigma_{xy})^2}{\sigma_x^2 \sigma_y^2}$$
While we haven’t discussed model fitting yet, it has an important relationship with $R^2$.

We can say that variable $Y$ can be decomposed or split into two components: what is shares with $x$, which we’ll call $b^*x$, and what is unique, which we’ll call $e$ for error.

- We’ll assume that $x$ and $e$ are uncorrelated, which is true by definition, and that $y$ and $x$ are standardized for convenience sake.
- $b$ will be the slope of that line we drew through the plot of $x$ and $y$, which we’ll specify in terms of the correlation coefficient $r$ and the variances of $x$ and $y$

$$y = b^* x + e = r \sqrt{\frac{Var(y)}{Var(x)}} x + e$$
Because $x$ and $e$ are uncorrelated, we can use the Variance Sum Law ($\text{Var}(A+B)=\text{Var}(A)+\text{Var}(B)$) to find the variance of $y$:

\[
\text{Var}(y) = \text{Var}(b \times x) + \text{Var}(e) = b^2 \times \text{Var}(x) + \text{Var}(e)
\]

\[
= \left( r \sqrt{\frac{\text{Var}(y)}{\text{Var}(x)}} \right)^2 \text{Var}(x) + \text{Var}(e)
\]

\[
= r^2 \frac{\text{Var}(y)}{\text{Var}(x)} \text{Var}(x) + \text{Var}(e) = r^2 \text{Var}(y) + \text{Var}(e)
\]

$R^2$ describes the amount of variance in $y$ that is composed of covariance with $x$.

The variance we assigned to $e$ is the variance in $y$ that’s left over.
$R^2$ from Model Fitting

- $R^2$ describes the amount of variance in $y$ that is accounted for or determined by $x$.
  - This is not a statement of causality.
- This is why we can talk about $R^2$ as a measure of effect size.
- Sometimes we’ll talk about $1-R^2$ or $(1-R^2)\times \text{Var}(y)$ as the amount of unexplained variation in $y$. 

R² and Sampling Corrections

- While the sampling correction for \( r \) isn’t commonly used, the correction is applied to R².
- Whereas the population correlation was termed \( \rho \), the population R² is often termed \( \eta^2 \).

\[
\eta^2 = 1 - (1 - R^2) \frac{N - 1}{N - 2}
\]

- Why the distinction?
  - Why the distinction? Correlations are rarely analyzed as raw units, but their raw unit forms have many handy transformations that \( \rho \) doesn’t have.
  - R² has handy transformations as well, but is more often used in modeling, thus having the more direct link to sampling.
Why are $R^2$ and $d$ used relatively interchangeably as measures of the same concept, effect size?

- They’re nonlinearly but perfectly related.

$$R^2 = \frac{d^2}{1 + d^2}$$

- We’ll discuss this a little more on Monday, but remember this.
We’ve dealt with linear combinations a little bit thus far.

We generally like adding and subtracting things at will, but sometimes we don’t have all of the information we need.

For example, let’s take the problem of power analysis for the matched sample $t$:

- I’m interested in a given treatment changes people from pre-test to post-test (we’ll ignore the control group for simplicity).
- From a literature review, I expect a mean change of 2 units with the variances of both pre-test and post-test at 25.
- How would I compute power?
Linear Combinations

To calculate power, I would need $d$, which requires the mean difference (post-pre) and the variance of the change score.

I don’t have the variance of the change score. What do I do?

The only thing I have is the variance sum law, which states that the variance of the sum of two independent variables is the sum of their variances.

- Independent means that the correlation between them is zero.
- Whether you add or subtract, you still sum the variances.

$$Var(post - pre | r_{pre,post} = 0) = Var(post) + Var(pre)$$
Linear Combinations

- Will that work?

No, because we hate that assumption: If pre-test and post-test aren't correlated, then we probably shouldn't be analyzing them as a matched-sample t. This is related to that invalidity-unreliability dilemma; the unreliability part has been disproved, but the invalidity part still holds.

So how can we figure out what the variance of the change score is?
Will that work?

No, because we hate that assumption:

- If pre-test and post-test aren’t correlated, then we probably shouldn’t be analyzing them as a matched-sample $t$.
- This is related to that invalidity-unreliability dilemma; the unreliability part has been disproved, but the invalidity part still holds.

So how can we figure out what the variance of the change score is?
Linear Combinations

Definitions

- The variance of a linear combination is defined by:

\[ \text{Var}(A + B) = \text{Var}(A) + \text{Var}(B) + 2 \times \text{Cov}(A, B) \]

- The only “trick” to this is when you use it on differences:

\[ \text{Var}(A - B) = \text{Var}(A + (-B)) = \text{Var}(A) + \text{Var}(-B) + 2 \times \text{Cov}(A, -B) \]
\[ = \text{Var}(A) + \text{Var}(B) - 2 \times \text{Cov}(A, B) \]
Linear Combinations

Examples

- What if we just double A, effectively combining A+A?

\[
\text{Var}(A + A) = \text{Var}(A) + \text{Var}(A) + 2 \times \text{Cov}(A, A)
\]
\[
= \text{Var}(A) + \text{Var}(A) + 2 \times \text{Var}(A) = 4 \times \text{Var}(A) = \text{Var}(2A)
\]

- What about A-A?

\[
\text{Var}(A - A) = \text{Var}(A + (-A)) = \text{Var}(A) + \text{Var}(-A) + 2 \times \text{Cov}(A, -A)
\]
\[
= \text{Var}(A) + \text{Var}(A) - 2 \times \text{Cov}(A, A)
\]
\[
= \text{Var}(A) + \text{Var}(A) - 2 \times \text{Var}(A) = 0
\]
Linear Combinations

Examples

Let’s go back to our pre-test/post-test example, and say we found a correlation between pre and post of 0.80. How would we run the analysis?

First, we’d have to convert the correlation back into a covariance:

\[ \text{Cov}(Pre, Post) = r_{Pre,Post} \times \sqrt{\text{Var}(Pre) \times \text{Var}(Post)} \]
\[ = 0.80 \times \sqrt{625} = 0.80 \times 25 = 20 \]

Then, we just do the math:

\[ \text{Var}(Post - Pre) = \text{Var}(Post) + \text{Var}(Pre) - 2 \times \text{Cov}(Post - Pre) \]
\[ = 25 + 25 - 2 \times 20 = 50 - 40 = 10 \]
Linear Combinations

- We can go back and calculate our effect size (Cohen’s $d$) for each of the two variances ($\text{Var(Sum)}=50$, $\text{Var(Right Way)}=10$), with our predicted mean difference of 2.

- The variance-sum law version, which we now know is wrong:

$$d = \frac{\bar{\text{post}} - \bar{\text{pre}}}{\sigma} = \frac{2}{\sqrt{50}} = 0.283$$

- The linear combination version, which we now know is right:

$$d = \frac{\bar{\text{post}} - \bar{\text{pre}}}{\sigma} = \frac{3}{\sqrt{10}} = 0.632$$

- Does the difference between a small effect and a medium effect really make that much of a difference?
Power

Right Way vs. Wrong Way

▶ It matters because power is drastically affected.
▶ If you wanted 90% power and used the wrong technique, you’d run 105 more subjects than you really needed.
  ▶ If you actually ran the study like this, you spent way too much money and time.
  ▶ If you were begging for grant money, you vastly underreported your power with 75 subjects (.676 vs .999).
▶ Seriously, 99.9% power.

<table>
<thead>
<tr>
<th>Power</th>
<th>Subjects Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>.70</td>
<td>80</td>
</tr>
<tr>
<td>.80</td>
<td>100</td>
</tr>
<tr>
<td>.90</td>
<td>134</td>
</tr>
</tbody>
</table>

Subjects Needed

Estabrook & Brick University of Virginia

Correlation Part Un
Dealing with uncertainty in the correlation coefficient.

- With our other statistics, we’ve always had some way of describing the precision of our estimates.
  - For means, we have the standard error.
- The good news is that we don’t have to worry about variance; the standard error of $r$ is entirely dependent on sample size.
- There’s bad news.
  - The easy way isn’t good.
  - The good way isn’t easy.
An Simple Test of the Significance of $r$

- A direct way of dealing with significance testing of $r$, provided you’re only interested in the null hypothesis of $r=0$ and you have sufficiently large N.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

- This statistic is $t$ distributed with $N-2$ degrees of freedom.
- When $\rho$ is zero (null is true), the sampling distribution of $r$ is symmetric around zero, making tests in raw units possible.
Problems

\[ t = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}} \]

- What is “sufficiently large N?” Good luck with that.
  - Large enough you don’t need a sampling correction on \( r \).
  - Large enough for the CLT to help you out.
  - Large enough a reviewer can’t say “your N was too small.”

- How do I get a confidence interval, or compare samples, or test different null hypotheses?
  - You don’t.

- We’ll have to think of another way...
Dealing with uncertainty in the correlation coefficient.

- The correlation coefficient is restricted between -1 and 1.
- Having such a strong restriction on the range of $r$ means that we can’t construct confidence intervals in the same way as before.
  - As the population correlation $\rho$ approaches $\pm 1$, the sampling distribution of $r$ gets increasingly skewed.
- To correct this, we need to transform $r$ so that its sampling distribution is approximately normal across the entire range.
- The transformation can’t be one of those easy linear transformations, because that wouldn’t do that.

$$z = r' = .5 \times \ln\left(\frac{1 + r}{1 - r}\right)$$
Dealing with uncertainty in the correlation coefficient.

\[ z = r' = 0.5 \times \ln\left(\frac{1 + r}{1 - r}\right) = 0.5 \times (\ln(1 + r) - \ln(1 - r)) \]

- This transformation is known as Fisher’s z transformation, sometimes called \( r' \) to distinguish it from the z score.
- The text uses vertical lines to enclose the vertical log: these aren’t (or shouldn’t be) absolute value lines.
- Fisher’s z is unbiased and normal around Fisher-transformed \( \rho \), with a standard error of:

\[ SE_z = \frac{1}{\sqrt{N - 3}} \]
Fisher’s z

![Histogram of Raw R values ranging from 0.5 to 0.9. The frequency peaks at around 0.8, with lower frequencies at other values.]

Estabrook & Brick University of Virginia
Correlation Part Un
Fisher’s z

Fisher Transformed

Frequency

0.6 0.8 1.0 1.2 1.4 1.6

0 50 100 150 200 250

z
Fisher’s z

Because Fisher’s z is normally distributed regardless of location (mean), we can construct confidence intervals using the normal distribution.

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

Where \( r' \) is the Fisher-transformed correlation and \( z \) is the \( \frac{\alpha}{2} \)th quantile of the normal distribution.
Fisher’s z

- We can also run z-tests (they are appropriate here, because \( r \) is a normally distributed sample statistic).
- We’ll see if our correlation coefficient is likely given the null that \( \rho \) is some value.

\[
z = \frac{r' - \rho}{\sigma_r'} = \frac{r - \rho}{\sqrt{\frac{1}{N-3}}}
\]

- Let’s practice one, shall we?
Example

- My study found that shoe size and reading ability share a 0.60 correlation in a sample of 52 kids. What’s the probability of getting a result this extreme if there’s actually no relationship?

\[
r' = 0.5 \times \log\left(\frac{1 + r}{1 - r}\right) = 0.5 \times \log\left(\frac{1 + 0.6}{1 - 0.6}\right) = 0.5 \times \log\left(\frac{1.6}{0.4}\right) = 0.693
\]

\[
\rho' = 0.5 \times \log\left(\frac{1 + \rho}{1 - \rho}\right) = 0.5 \times \log\left(\frac{1 + 0}{1 - 0}\right) = 0.5 \times \log(1) = 0
\]

\[
z = \frac{r' - \rho}{\sigma_r'} = \frac{0.693 - 0}{\sqrt{\frac{1}{52 - 3}}} = \frac{0.693 - 0}{0.143} = 4.85
\]

- \( P(z > |4.85|) = 0.000001 \). It’s unlikely that this sample came from a population with \( \rho = 0 \).
What’s the confidence interval around my sample statistic?

\[ CI(\rho) = r' \pm z_{\alpha/2} \sqrt{\frac{1}{N-3}} = 0.693 \pm 1.96 \times 0.143 = [0.413, 0.973] \]

So my confidence interval is from [0.413, 0.973]?

No, our confidence interval in Fisher’s z units is from [0.413, 0.973]. We’ve got to change it back.

\[ r = \frac{e^{2r'} - 1}{e^{2r'} + 1} \]
Example

\[ r = \frac{e^{2r'} - 1}{e^{2r'} + 1} \]

\[ Cl_{lower} = \frac{e^{2r'} - 1}{e^{2r'} + 1} = \frac{e^{2(.413)} - 1}{e^{2(.413)} + 1} = .391 \]

\[ Cl_{lower} = \frac{e^{2r'} - 1}{e^{2r'} + 1} = \frac{e^{2(.973)} - 1}{e^{2(.973)} + 1} = .749 \]

- Our 95% confidence interval for the correlation between shoe size and reading ability is from [.391,.749], with an estimate of \( r = 0.60 \).
How can shoe size and reading ability be related?

- ???

Older kids have bigger feet and more reading experience.

This is not just a problem of correlation: Any statistical model would have found the same relationship. It's up to you to look at graphs, consider assumptions, etc, etc.

Some might say "Correlation Does Not Equal Causation." Some might say "Causation is a Design Issue."
How can shoe size and reading ability be related?

- Older kids have bigger feet and more reading experience.
- This is not just a problem of correlation:
  - Any statistical model would have found the same relationship.
  - It's up to you to look at graphs, consider assumptions, etc, etc.
  - Some might say “Correlation Does Not Equal Causation.”
  - Some might say “Causation is a Design Issue.”
Next Time

- More Correlations: Chapters 9 and 10.