Module E: Advanced Inference

Slide Deck E8:

A History of Correlation

The section in which we examine how statisticians arrive at certain measures by seeing how different measures produce different results. Each measurement has advantages. The key is to understand what the measurements actually mean.

Today's Objectives

By the end of this slidedeck, you should

1. understand the history of correlation calculations
2. understand the theory behind testing...
   - relationship between two numeric variables
3. better understand the p-value and how to test hypotheses

Framing Example

Example

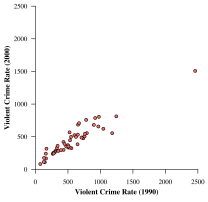
Understanding the persistence of violent crime in cities can help determine which cities are being successful in reducing it and which cities are not. To start the research, we will need to determine whether there is a relationship between the violent crime rate in 1990 and in 2000.

Note: At this point, we only want to know whether there is a significant linear relationship. How could we determine this?

Great question! Let's look at a graph to see if that can give any hints.

Framing Example

Questions: From this graphic, does there appear to be a correlation between these two variables? What is it about the graphic makes you think there is a relationship?



The Theory

That is the key question:

- What is it about the graphic tells us that there is a relationship between these two variables??

This was the question that led us — eventually — to our modern measure of correlation.

The next several slides walk through the history of many answers to this question.

Concordance

The first attempt compared the number of **concordant points** to **discordant points**.

Definition (Concordant Point)

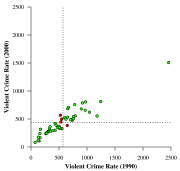
A **concordant point** is one that supports a *positive* relationship between the two variables.

Definition (Discordant Point)

A **discordant point** is one that supports a *negative* relationship between the two variables.

Concordance Ratio

The original correlation measure was the number of concordant points minus the number of discordant points, divided by the sample size.



This "concordance ratio" measure for this data is $(47 - 4)/51 = 0.84$.

Concordance Ratio

This measure had the strengths

- easily calculated
- ranged between -1 and $+1$

This measure had the weaknesses

- ignores information about the points (specifically, how far from the center)

To fix this weakness, we created the "covariance" measure.

Covariance

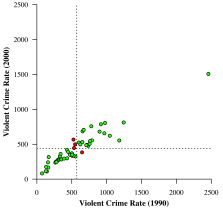
Covariance, $cov(x, y)$ or $s_{x,y}$, is a measure of how much the two variables vary *together*, taking into consideration the data values.

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Check that this measure takes into consideration both concordance *and the distance* to the center.

Covariance

The covariance is `cov(vcrime90,vcrime00)` = 88,047.41.



Covariance

This measure had the strengths

- based on concordance
- based on how far the point is from the center

This measure had the weakness

- it does not range between -1 and $+1$
  - as such, it cannot be compared across data sets

To fix this weakness, the "correlation" was devised.

Correlation

Correlation is a *standardized* covariance (Bravais, 1844). Check that the following measure takes into consideration these two things:

1. the concordance
2. the distance to the center

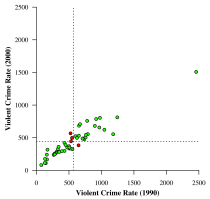
all the while ensuring that it ranges between -1 and $+1$:

$$cor(x,y) = r_{x,y} = \frac{s_{x,y}}{s_x\ s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

This is the formula for the "sample correlation." It is a good estimator of the population correlation $\rho_{x,y}$, which is what we like to understand.

Correlation

The correlation is `cor(vcrime90,vcrime00)` $= 0.936\,469$.



Correlation

Since correlation ranges between -1 and $+1$,

- we can compare correlations between different pairs of variables

Correlation Test

Rahman (1944) showed, for correlations near 0, that

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

With this, we have our test statistic, from which we can test hypotheses concerning hypotheses of the forms $\rho \leq 0$, $\rho = 0$, or $\rho \geq 0$:

$$\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

Confidence intervals when r is far from 0 are a bit trickier, however.

In R, the command to perform the correlation test is `cor.test(x,y)`.

Example 1: Property Crime Rate

Example

Are the property crime rate in 1990 and in 2000 significantly correlated?

Here is the code to answer this question:

```
source("http://rfs.kvasaheim.com/stat200.R")

dt = read.csv("http://rfs.kvasaheim.com/data/crime.csv")
attach(dt)

cor.test(pcrime90,pcrime00)
```

Example 1: Property Crime Rate

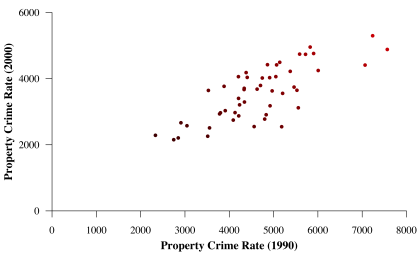
The resulting output is

```
Pearson's product-moment correlation

data:  pcrime90 and pcrime00
t = 9.3781, df = 49, p-value = 1.623e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6748213 0.8821406
sample estimates:
      cor
0.8013737
```

Brief Conclusion: Because the p-value is much less than our $\alpha = 0.05$, we reject the null hypothesis that there is no correlation between the property crime rates in 1990 and in 2000. In fact, we are 95% confident that the true correlation is between 0.67 and 0.88. The graphic on the next page illustrates this.

Example 1: Property Crime Rate



Example 2: Wealth and Professionalism

Example

Are the average wealth in the state and the level of professionalism in its legislature correlated?

Here is the code to answer this question:

```
cor.test(gspcap00, profleg)
```


Example 2: Wealth and Professionalism

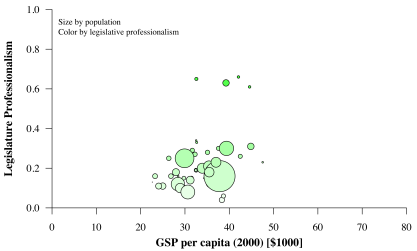
The resulting output is

```
Pearson's product-moment correlation

data:  gspcap00 and profleg
t = 2.3585, df = 48, p-value = 0.02247
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04823515 0.55116484
sample estimates:
      cor
0.3222562
```

Brief Conclusion: Because the p-value of 0.0225 is less than our $\alpha = 0.05$, we reject the null hypothesis that there is no correlation between the average wealth in the state and the level of professionalism in its legislature. In fact, we are 95% confident that the true correlation is between 0.05 and 0.55. The graphic on the next page illustrates this.

Example 2: Wealth and Professionalism



Example 3: Wealth and Education

Example

Are the average wealth in the state and the school enrollment correlated?

Here is the code to answer this question:

```
cor.test(gspcap00, enroll00)
```

Example 3: Wealth and Education

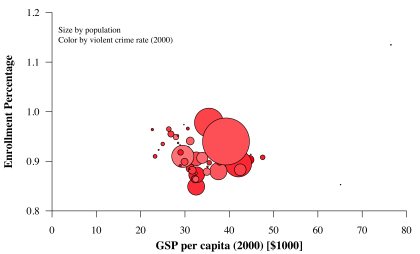
The resulting output is

```
Pearson's product-moment correlation

data:  gspcap00 and enroll00
t = 1.767, df = 49, p-value = 0.08346
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03306579  0.48745373
sample estimates:
      cor
0.2447481
```

Brief Conclusion: Because the p-value of 0.0835 is greater than our $\alpha = 0.05$, we should not reject the null hypothesis. We did not detect a relationship between the average wealth in the state and the enrollment percentage. In fact, we are 95% confident that the true correlation is between -0.03 and 0.49. The graphic on the next page illustrates this.

Example 3: Wealth and Education



Example 4: Wealth and Crime

Example

Are the average wealth in the state and the 2000 violent crime rate correlated?

Here is the code to answer this question:

```
cor.test(gspcap00, vcrime00)
```


Example 4: Wealth and Crime

The resulting output is

```
Pearson's product-moment correlation

data:  gspcap00 and vcrime00
t = 3.5628, df = 49, p-value = 0.0008289
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2034498 0.6481614
sample estimates:
      cor
0.4536012
```

Brief Conclusion: Because the p-value of 0.0008 is less than our $\alpha = 0.05$, we should reject the null hypothesis. We did detect a relationship between the average wealth in the state and the violent crime rate. In fact, we are 95% confident that the true correlation is between 0.20 and 0.65.

Example 4: Wealth and Crime

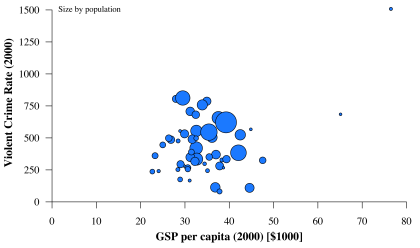
This is what a full conclusion would look like:

Conclusion:

We are asked to determine if there is a relationship between the average wealth and the violent crime rate in the state. Since both of these variables are numeric, we will use the Pearson's product-moment correlation test.

Because the p-value of 0.0008 is less than our $\alpha = 0.05$, we should reject the null hypothesis. We did detect a relationship between the average wealth in the state and the violent crime rate. In fact, we are 95% confident that the true correlation is between 0.20 and 0.65.

Example 4: Wealth and Crime



Today's Objectives

Now that we have concluded this lecture, you should be able to

1. understand the history of correlation calculations
2. understand the theory behind testing...
   - relationship between two numeric variables
3. better understand the p-value and how to test hypotheses

At this point, we can only test if a linear relationship exists...

- non-linear relationships are beyond this test
- we cannot predict values of the dependent variable

Linear regression can help with both weaknesses. However, we will only look to it to solve the second in this course.

Today's R Functions

Here are R functions we used in our exploration of relationships between numeric variables

- `concordanceRatio(x,y)`
  This calculates the concordance ratio between two variables
- `cov(x,y)`
  This calculates the covariance between two variables
- `cor.test(x,y)`
  This performs the correlation test

Supplemental Activities

The following activity is currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning correlation.

- SCA 42a

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding the theory of hypothesis testing:

- Hawkes Learning: Section 12.1
- Intro to Modern Statistics: Section 7.1
- `R` for Starters: Section 12.1

- Wikipedia: Correlation

Please do not forget to use the `allProcedures` document that lists all of the procedures we will use in `R`.