Review of Basic Statistical Concepts

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Review of Basic Statistical Concepts

1. Introduction
2. The Mean and the Expected Value
3. Listwise Operations and Linear Transformations in R
4. Deviation Scores, Variance, and Standard Deviation
5. *Z*-Scores
6. Covariance and Correlation
7. Covariance
   - The Concept of Covariance
   - Computing Covariance
   - Limitations of Covariance
8. The (Pearson) Correlation Coefficient
   - Definition
   - Computing
   - Interpretation
9. Some Other Correlation Coefficients
   - Introduction
10. Population Variance, Covariance and Correlation
11. Laws of Linear Combination
12. Significance Test for the Correlation Coefficient

Introduction

- In this module, we will quickly review key statistical concepts and their algebraic properties.
- These concepts are taken for granted (more or less) in all graduate level discussions of regression analysis.
- There are extensive review chapters available to help you gain/recover familiarity with the concepts.

The Mean

- The mean of a list of numbers is the arithmetic average of the list, i.e., the sum divided by n.

$$\overline{X}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The Expected Value

- The expected value of a random variable is the long run arithmetic average of the values taken on by the random variable.
- The expected value of a random variable X is denoted $E(X)$, and is also often simply referred to as the mean of the random variable X.

Algebraic Properties of Linear Transformation

- A listwise operation is a mathematical transformation applied uniformly to every number in a list.
- A key fact discussed extensively in Psychology 310 is that addition, subtraction, multiplication, and division of all the values in a list (or, alternatively, all the values taken on by a random variable) comes “straight through” in the mean.
- A linear transformation of the form $Y = aX + b$ includes all 4 basic listwise operations as special cases.

Algebraic Properties of Linear Transformation

Theorem (Mean of a Linear Transform)

Suppose Y and X are random variables, and $Y = aX + b$ for constants a and b. Then

$$E(Y) = aE(X) + b$$

If Y and X are lists of numbers and $Y_i = aX_i + b$, then a similar rule holds, i.e.,

$$\overline{Y}_\bullet = a\overline{X}_\bullet + b$$

Algebraic Properties of Linear Transformation

Example (Listwise Transformation and the Sample Mean)

Suppose you have a list of numbers X with a mean of 5.

If you multiply all the X values by 2 and then add 3 to all those values, you have transformed X into a new variable Y by the *listwise operation* $Y = 2X + 3$.

In that case, the means of Y and X will be related by the same formula, i.e., $\overline{Y}_{\bullet} = 2\overline{X}_{\bullet} + 3 = 2(5) + 3 = 13$.

Algebraic Properties of Linear Transformation

Example (Listwise Transformation and the Population Mean)

Suppose you have a random variable X with an expected value of $E(X) = 10$. Define the random variable $Y = 2X - 4$. Then $E(Y) = 2E(X) - 4 = 20 - 4 = 16$.

Elementary Listwise Operations

- Getting a short list of data into R is straightforward with an assignment statement.
- Here we create an X list with the integer values 1 through 5.

```
> X <- c(1,2,3,4,5)
```

Elementary Listwise Operations

- Creating a new variable that is a linear transformation of the old one is easy:

```
> Y = 2*X + 5
> Y

[1]  7  9 11 13 15
```

- And, the means of X and Y obey the linear transformation rule.

```
> mean(X)

[1] 3

> 2 * mean(X) + 5

[1] 11

> mean(Y)

[1] 11
```

Deviation Scores, Variance, and Standard Deviation

- If we re-express a list of numbers in terms of where they are relative to their mean, we have created deviation scores.
- Deviation scores are calculated as

$$dx_i = X_i - \overline{X}\bullet$$

- This is done easily in R as

```
> dx = X - mean(X)
> X

[1] 1 2 3 4 5

> dx

[1] -2 -1  0  1  2
```

Deviation Scores, Variance, and Standard Deviation

- If we want to measure how spread out a list of numbers is, we can look at the size of deviation scores.
- Bigger spread means bigger deviations around the mean.
- One might be tempted to use the average deviation score as a measure of spread, or variability.
- But that won't work.

Deviation Scores, Variance, and Standard Deviation

Why Not?

Deviation Scores, Variance, and Standard Deviation

- A better idea is the average squared deviation.
- An even better idea, if you are estimating the average squared deviation in a large population from the information in the sample, is to use the sample variance

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_\bullet)^2$$

- The sample standard deviation is simply the square root of the sample variance, i.e.,

$$S_X = \sqrt{S_X^2}$$

Deviation Scores, Variance, and Standard Deviation

- Computing the variance or standard deviation in R is very easy.

```
> var(X)
[1] 2.5
> sd(X)
[1] 1.581139
```

Linear Transformation Rules for Variances and Standard Deviations

- Multiplication or division comes straight through in the standard deviation if the multiplier is positive — otherwise the absolute value of the multiplier comes straight through.
- This makes sense if you recall that there is no such thing as a negative variance or standard deviation!
- Additive constants have no effect on deviation scores, and so have no effect on the standard deviation or variance.

Linear Transformation Rules for Variances and Standard Deviations

INVESTIGATE! IN R!!

Linear Transformation Rules for Variances and Standard Deviations

```
> X

[1] 1 2 3 4 5

> X - mean(X)

[1] -2 -1  0  1  2

> sd(X)

[1] 1.581139

> Y <- X + 5
> Y - mean(Y)

[1] -2 -1  0  1  2

> sd(Y)

[1] 1.581139
```

Linear Transformation Rules for Variances and Standard Deviations

```
> Y <- 2*X + 5
> Y - mean(Y)

[1] -4 -2  0  2  4

> sd(Y)

[1] 3.162278

> var(Y)

[1] 10
```

Linear Transformation Rules for Variances and Standard Deviations

- Unless stated otherwise, we will generally assume that linear transformations are "positive," i.e., the multiplier is a positive number.
- With that assumption, we can say the following:

Theorem

Let Y and X represent lists of numbers, and a and b be constants. Then if

$$Y = aX + b \quad \textit{and } a > 0$$

$$S_Y = aS_X$$

and

$$S_Y^2 = a^2 S_X^2$$

In analogous fashion, if Y and X are random variables, then

$$\sigma_Y = a\sigma_X$$

and

$$\sigma_Y^2 = a^2\sigma_X^2$$

Z-Scores

- In Psychology 310, we go into quite a bit of detail explaining how any list of numbers can be thought of as having
  1. Shape
  2. Metric, comprised of a mean and a standard deviation.

Z-Scores

- Shape, the pattern of relative interval sizes moving from left to right on the number line, is *invariant under positive linear transformation*.
- It can be thought of as the information in a list that “transcends scaling.”

Z-Scores

- Metric, the mean and standard deviation of the numbers, can be thought of as the information in a list that “reflects scaling.”
- In a lot of situations, “metric can be thought of as arbitrary.”

Z-Scores

What does THAT mean??

Z-Scores

If metric is arbitrary, do we need it??

Z-Scores

- Consider the Z score transformation, which transforms a list of X values as

$$Z_i = \frac{X_i - \overline{X}_\bullet}{S_x}$$

- If we do this to a list of numbers, what will their mean and standard deviation (i.e., their metric) become?

Z-Scores

Did your mind go blank??

Z-Scores

If it did — a helpful strategy

Z-Scores

- Create a “random” list of numbers.
- Not too small, not too large, call it X
- Now, convert to Z scores and see what happens.

```
> X <- c(16.2,33,13.9,12.8,3.3)
> X
[1] 16.2 33.0 13.9 12.8  3.3
> Z <- (X - mean(X))/sd(X)
> mean(Z)
[1] 2.502339e-17
> sd(Z)
[1] 1
```

Z-Scores

Now YOU try it.

Z-Scores

- It seems like, no matter what list of numbers we generate, the Z-transform converts them so that they have a mean of 0 (ignoring round-off error) and a standard deviation of 1.
- Now that we suspect we know the answer, we can perhaps be more confident as we set out to *prove* that, in fact, this suspicion is correct.

Z-Scores

- Let's "track" what happens to a list of numbers X as we apply the Z-score transformation.

$$Z = \frac{X - \overline{X}_{\bullet}}{S_X}$$

Z-Scores

We start in the numerator with the original scores in X. What happens to the scores when we subtract $\overline{X}_\bullet$?

$$Z = \frac{X - \overline{X}_\bullet}{S_X}$$

We recall from our linear transformation rules that subtracting the constant $\overline{X}_\bullet$ has no effect on the standard deviation of the scores, so the scores will still have a standard deviation of S_X. However, subtracting $\overline{X}_\bullet$ reduces the mean of the scores by $\overline{X}_\bullet$, so the mean has been changed to 0.

So at this stage of the transformation, we have scores with a mean of zero and a standard deviation of S_X.

Z-Scores

Moving on to the next stage of the transformation, we realize that dividing by S_X divides the standard deviation by S_X, and so the standard deviation becomes $S_X/S_X = 1$.

The mean is $0/S_X = 0$, and remains unchanged.

We now see that what R demonstrated to us numerically is mathematically inevitable.

$$Z = \frac{X - \overline{X}_\bullet}{S_X}$$

Z-Scores

- In an important sense, Z-scoring removes the metric from a list of numbers by rendering any list with the same, simple metric.
- We say that scores are in Z-score form if they have a mean of 0 and a standard deviation of 1.
- Once scores are in Z-score form, we can convert them into any other desired metric by just mulitplying by the desired standard deviation, then adding the desired mean.

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

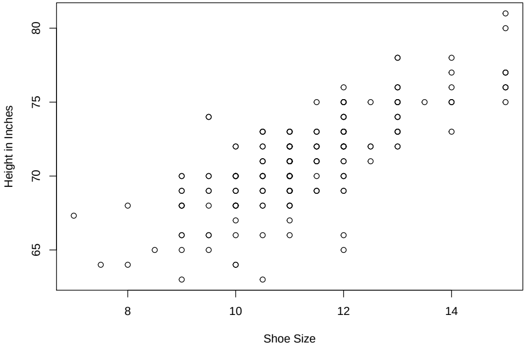
- We can isolate the male data from all the data with the following command:

  ```
  > rm(X,Y) # remove old X,Y variables
  > male.data <- all.heights[all.heights$Gender=="M",] #Select males
  > attach(male.data)#Make Variables Available
  ```

Bivariate Distributions and Scatterplots

- Let's draw a scatterplot:

```
> # 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 independent 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.

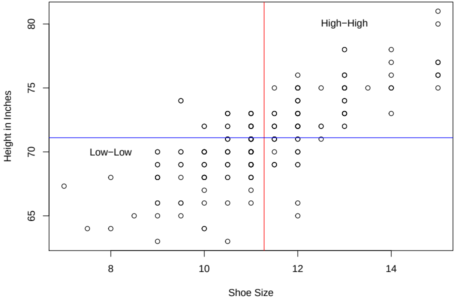
Bivariate Distributions and Scatterplots

- 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

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_\bullet)(Y_i - \overline{Y}_\bullet) \tag{1}$$

- 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

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

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., $\overline{X}_\bullet = 10$, $\overline{Y}_\bullet = 36$, 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 we 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.

Covariance

Limitations

How can we fix this?

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

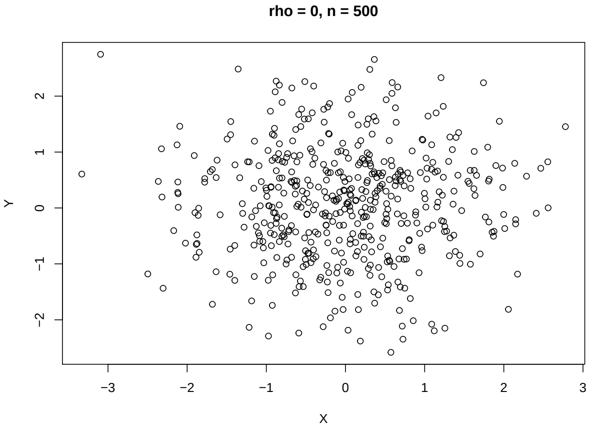
Example (Some Typical Scatterplots)

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

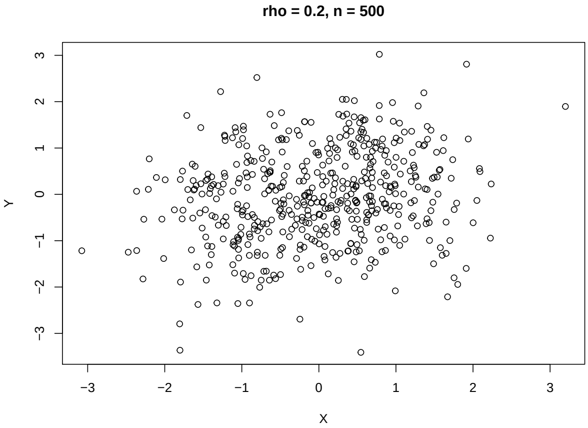
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

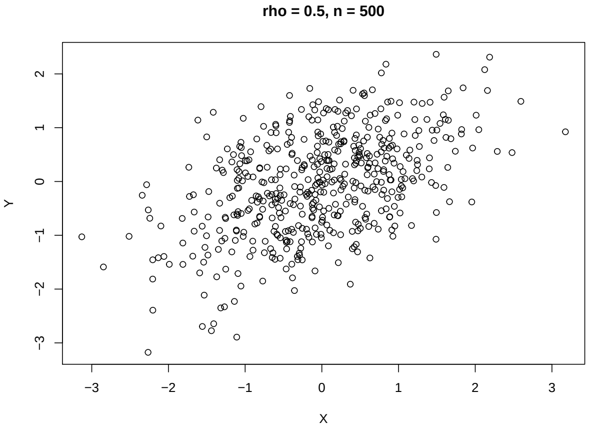
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

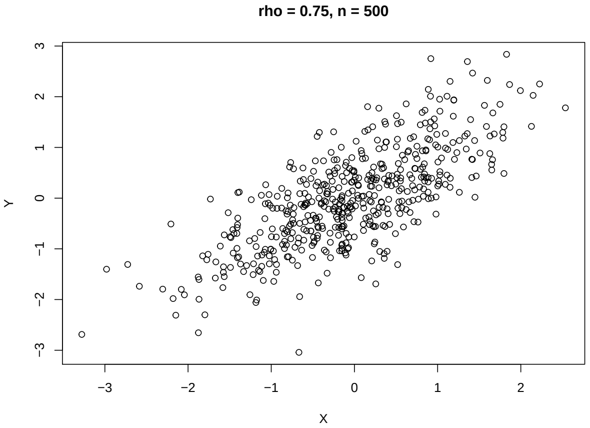
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

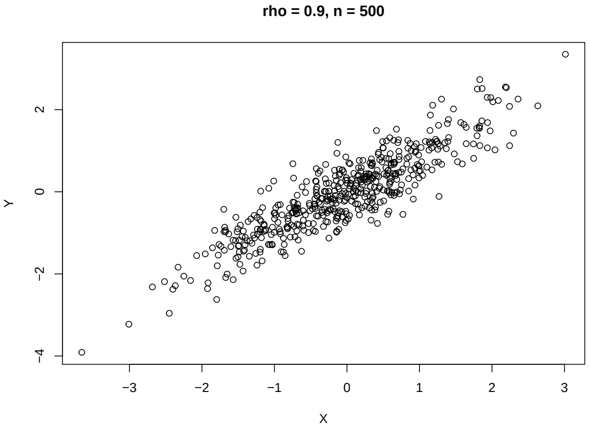
Example (Some Typical Scatterplots)



The (Pearson) Correlation Coefficient

Interpreting a Correlation

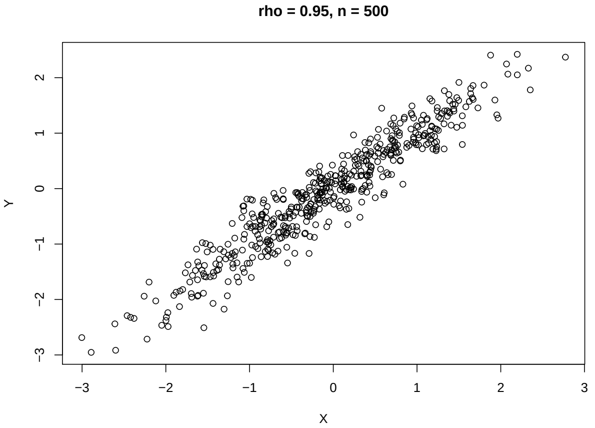
Example (Some Typical Scatterplots)



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.

Population Variance, Covariance and Correlation

Introduction

- Each of the sample quantities, variance, covariance, and correlation has a corresponding population quantity that is usually described in terms of expected value theory.
- In this section we will review some important aspects of the algebra of expected values.

Population Variance, Covariance and Correlation

Expected Value Algebra

- Recall that the expected value of a random variable X, denoted $E(X)$, is the long run average of values taken on by the random variable.
- In general, functions of random variables are themselves random variables. For example, if X is a random variable, X^2 is a random variables, as is $2X + 4$.

Population Variance, Covariance and Correlation

Expected Value Algebra

For random variables X and Y, and constants a and b, we have the following results.

$$
\begin{aligned}
E(a) &= a && (8)\\
E(aX+b) &= aE(X)+b && (9)\\
E(X+Y) &= E(X)+E(Y) && (10)
\end{aligned}
$$

Population Variance, Covariance and Correlation

Population Variance

Definition (Population Variance and Standard Deviation)

The variance of a random variable X is defined as the long run average squared deviation score, i.e.,

$$\mathrm{Var}(X) = \sigma_X^2 = E((X - E(X))^2) \tag{11}$$

The standard deviation σ_X of a random variable X is the square root of the variance of X.

The variance of a random variable may also be computed with the important formula

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 \tag{12}$$

Population Variance, Covariance and Correlation

Population Covariance

Definition (Population Covariance)

The covariance of the random variables X and Y is defined as the long run average cross-product of deviation scores, i.e.,

$$\mathrm{Cov}(X, Y) = \sigma_{X,Y} = E((X - E(X))(Y - E(Y))) \tag{13}$$

The covariance of X and Y may also be computed as

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) \tag{14}$$

Population Variance, Covariance and Correlation

Z-Score Random Variables

Definition (Z-score Random Variable)

A random variable is said to be in deviation score form if it has a mean of zero. It is said to be in Z-score form if it has a mean of zero and a standard deviation of 1. Any random variable X with positive variance may be converted to Z score form with the formula

$$Z_X = \frac{X - E(X)}{\sigma_X} = \frac{X - \mu_X}{\sigma_X}$$

Population Variance, Covariance and Correlation

Population Correlation

Definition (Population Correlation)

The correlation of random variables X and Y is defined as the long run average cross-product of Z scores, i.e.,

$$\rho_{X,Y} = E(Z_Y Z_Y) \tag{15}$$

The correlation of X and Y may also be computed as

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{16}$$

Laws of Linear Combination

Definition (Linear Combination)

A linear combination of two random variables X and Y is any expression of the form $aX + bY$ where a and b are constants called linear weights.

Laws of Linear Combination

Mean of a Linear Combination

Theorem (Mean of a Linear Combination)

If random variables X and Y have means $E(X)$ and $E(Y)$, respectively, then the linear combination $aX + bY$ has mean $E(aX + bY) = aE(X) + bE(Y)$.

A similar result holds for linear combinations with sample data. That is, if X and Y represent lists of numbers, and $W_i = aX_i + bY_i$, then $\overline{W}_\bullet = a\overline{X}_\bullet + b\overline{Y}_\bullet$.

Laws of Linear Combination

Variance of a Linear Combination

Theorem (Variance of a Linear Combination)

For random variables W,X,and Y, if $W = aX + bY$, then

$$\sigma^2_W = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{X,Y}$$

In a similar vein, for lists of numbers X and Y, if $W_i = aX_i + bY_i$, then

$$S^2_W = a^2S^2_X + b^2S^2_Y + 2abS_{X,Y}$$

Laws of Linear Combination

The General Heuristic Rule

Theorem (The General Heuristic Rule)

A general rule that allows computation of the variance of any linear combination or transformation, as well as the covariance between any two linear transformations or combinations, is the following:

- *For the variance of a single expression, write the expression, square it, and apply the simple mnemonic conversion rule described below.*
- *For the covariance of any two expressions, write the two expressions, compute their algebraic product, then apply the conversion rule described below.*

The conversion rule is as follows:

- *All constants are carried forward.*
- *If a term has the product of two variables, replace the product with the covariance of the two variables.*
- *If a term has the square of a single variable, replace the squared variable with its variance.*
- *Any term without the product of two variables or the square of a variable is deleted.*

Laws of Linear Combination

The General Heuristic Rule

Example (The General Heuristic Rule)

Suppose X and Y are random variables, and you compute the following new random variables:

- $W = X - Y$
- $M = 2X + 5$

Construct formulas for

1. σ^2_W
2. σ^2_M
3. $\sigma_{W,M}$

(Answers on next slide . . .)

Laws of Linear Combination

The General Heuristic Rule

Example (The General Heuristic Rule)

Answers.

1. To get σ^2_W, we square $X - Y$, obtaining $X^2 + Y^2 - 2XY$, and apply the conversion rule to get $\sigma^2_W = \sigma^2_X + \sigma^2_Y - 2\sigma_{X,Y}$.
2. To get σ^2_M, we square $2X + 5$, obtaining $4X^2 + 20X + 25$. Applying the conversion rule, we drop the last two terms, neither of which have the square of a variable or the product of two variables. We are left with the first term, which yields $\sigma^2_M = 4\sigma^2_X$.
3. To get $\sigma_{W,M}$, we begin by computing $(X - Y)(2X + 5) = 2X^2 - 2XY + 5X - 5Y$. We drop the last two terms, and obtain $\sigma_{W,M} = 2\sigma^2_X - 2\sigma_{X,Y}$.

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 $\rho = 0$.

$$t_{n-2} = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \tag{17}$$

Significance Test for r

Example (Significance Test for r)

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) ## this command gets the 0.975 quantile of t
> 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.