Stats 401 Lab 6

Ed Wu

10/12/2018

Announcements

- Homework 5 is due next Friday (Oct 19)

Outline

- Bivariate Random Variables
- Correlation and Covariance
- The Bivariate Normal Distribution

Bivariate Random Variables

- Recall: a random variable X is a value whose outcome is determined by a random process
  - For example, X might be the value of a roll of a die
  - X takes a value in $\{1, 2, 3, 4, 5, 6\}$, each with probability $1/6$
- We might be interested in vector valued random variables instead
- A bivariate random variable (X, Y) is a vector of length 2 whose values are each random variables

Bivariate Random Variables

- One reason to consider the bivariate random variable (X, Y) jointly is that the outcomes for X and Y may be related
- Suppose we roll two dice. We let X be the value of the first die and Y be the sum of the two dice. Then

  $$\mathrm{P}((X, Y) = (1, 7)) = \mathrm{P}(\text{First Die is 1, Second Die is 6}) = 1/36$$

Measuring Association

- Correlation is a measure of the linear association between two random variables
- If X tends to be large when Y tends to be large, X and Y are positively correlated
- If X tends to be small when Y tends to be large, X and Y are negatively correlated
- If X and Y have no linear association, then the correlation between X and Y is zero

Correlation

- Correlation is always between -1 and 1 (inclusive)
- $\text{Cor}(X, Y) = \pm 1$ implies linear dependence between X and Y
  - This means we can write a linear equation to express the value of X in terms of Y (and vice versa)
  - Let X be the value of the roll of a die and Y be one plus twice the value of X
  - Since we can write Y as $2X + 1$, $\text{Cor}(X, Y) = 1$
- Correlation is symmetric: $\text{Cor}(X, Y) = \text{Cor}(Y, X)$

Covariance

- Covariance is the unscaled version of correlation
- While correlation has been scaled to remove units, covariance depends on the original units
  - If $\text{Cov}(X, Y) = 2$ and $\text{Cor}(X, Y) = 0.5$, then $\text{Cov}(2X, Y)$ will be 4, while $\text{Cor}(X, Y)$ remains 0.5
- Covariance is less interpretable than correlation because the size of the covariance depends on the units of the random variables – correlation is always between -1 and 1
- Covariance is often more useful for calculations

Formulas

Covariance

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y)\right]$$

Correlation

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Formulas

Suppose we have n measurements $(x_1, y_1), \ldots, (x_n, y_n)$. Let $\bar{x}$ and $\bar{y}$ be the sample means of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$.
Then we have the sample covariance:

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

and sample correlation:

$$\text{cor}(\mathbf{x}, \mathbf{y}) = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}(\mathbf{x})\text{var}(\mathbf{y})}}$$

(reminder var($\mathbf{x}$) is the sample variance of $(x_1, \ldots, x_n)$)

Relationship between Covariance and Sample Covariance Formulas

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y)\right]$$

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

Example

Suppose we flip two coins, each with a 1 on one side and a 2 on the other. Let X be the value of the first coin and Y the sum of the two flips

What is the covariance of X and Y?

Example

Let's use R to see how close the sample covariance is to the covariance

The following function takes *n* samples from the bivariate random variable described

```
my_bivrv = function(n){
  flips = replicate(n,sample(1:2,2,replace = TRUE))
  x = flips[1,]
  y = apply(flips,2,sum)
  return(cbind(x,y))
}
```

Example

We use my_bivrv() to draw 10 samples

```
xy = my_bivrv(10)
head(xy)
```

```
##      x y
## [1,] 1 3
## [2,] 2 4
## [3,] 2 3
## [4,] 1 2
## [5,] 2 3
## [6,] 2 3
```

Then we calculate the sample covariance

```
cov(xy[,1], xy[,2])
```

```
## [1] 0.2
```

Example

We generate samples of size $10, 50, 100, 500, 1000$ and calculate the covariance each time using the function above. We can see that the sample covariance is generally close to the true value of 0.25 for larger sample sizes.

```
##      sample size sample covariance
## [1,]          10         0.2000000
## [2,]          50         0.2959184
## [3,]         100         0.2581818
## [4,]         500         0.2493026
## [5,]        1000         0.2497047
```

Lab Activity (Part 1)

1. If $\text{Cor}(W, Z) = 0.5$, what is the correlation of $\text{Cor}(2W, Z + 1)$?
2. Let (X, Y) take the values $(0, 1), (1, 1), (1, 2)$, each with probability $1/3$
   - What is the covariance of X and Y?
   - We take a sample of size 5: $(0, 1), (0, 1), (1, 2), (1, 1), (1, 2)$. What is sample covariance?

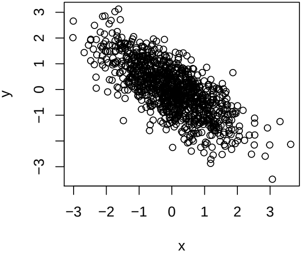
Bivariate Normal Distribution

- Suppose $X \sim \mathsf{N}(\mu_x, \sigma_x)$ and $Y \sim \mathsf{N}(\mu_y, \sigma_y)$ are normal random variables
- (X, Y) is a bivariate normal random variable
- We can characterize a bivariate normal random variable with its mean vector (μ_x, μ_y) and its variance-covariance matrix

$$\mathbb{V} = \begin{bmatrix} \sigma_x^2 & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \sigma_y^2 \end{bmatrix}$$

Lab Activity (Part 2)

The scatterplot below was generated from a bivariate normal distribution with mean vector $(0, 0)$

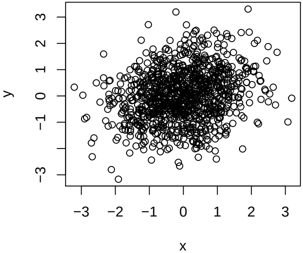


Which of the following is the variance-covariance matrix?

1. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; 2. $\begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$; 3. $\begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$

Lab Activity (Part 2)

The scatterplot below was generated from a bivariate normal distribution with mean vector $(0, 0)$

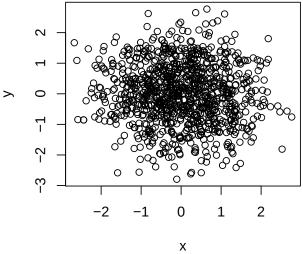


Which of the following is the variance-covariance matrix?

1. $\begin{bmatrix} 1 & -0.2 \\ -0.2 & 1 \end{bmatrix}$; 2. $\begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$; 3. $\begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$

Lab Activity (Part 2)

The scatterplot below was generated from a bivariate normal distribution with mean vector $(0, 0)$



Which of the following is the variance-covariance matrix?

1. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; 2. $\begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$; 3. $\begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}$

Multivariate Random Variables

In this lab, we discussed bivariate random variables and the bivariate normal distribution. We can extend these concepts to multivariate random variables

- For example, we might have the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$

Multivariate Random Variables

- Summary statistics for a multivariate random variable include the expected value vector and the variance-covariance matrix
- The expected value vector $\mathrm{E}(\mathbf{X}) = (\mathrm{E}(X_1), \ldots, \mathrm{E}(X_p))$ tells us the means for each component of $\mathbf{X}$
- The variance-covariance matrix gives the variances for each component along the diagonal and the pairwise covariances in the other entries:

$$\mathbb{V} = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}(X_p) \end{bmatrix}$$

Lab Ticket

1. Why is $\begin{bmatrix} 4 & 0 \\ 0.25 & 4 \end{bmatrix}$ not a valid variance-covariance matrix?
2. Let (X, Y) be bivariate normal with mean $(6, 4)$ and variance-covariance matrix $\mathbb{V} = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}$.
   - What are the mean and standard deviation of Y?
   - What is the covariance of X and Y?