Distributions for Pairs of Random Variables

Chapter 3, Lab 4: Template

OpenIntro Biostatistics

Topics

- Marginal, joint, and conditional distributions
- Correlated random variables
- Variance of the sum or difference of two correlated random variables

There are many examples of correlated random variables, such as height and weight in a population of individuals. This lab discusses distributions for pairs of random variables, in addition to the correlation of two random variables.

The material in this lab corresponds to Section 3.6 of *OpenIntro Biostatistics*.

Marginal, joint, and conditional distributions

The **joint distribution** $p_{X,Y}(x, y)$ for a pair of random variables (X, Y) is the collection of probabilities

$$p(x_i, y_j) = P(X = x_i, Y = y_j) \text{ , for all pairs of values } (x_i, y_j).$$

Joint distributions are often displayed in tabular form: If X and Y have k_1 and k_2 possible values respectively, there will be $(k_1)(k_2)$ possible (x, y) pairs.

When two variables X and Y have a joint distribution, the **marginal distribution** of X is the collection of probabilities for X when Y is ignored. The marginal distribution of X can be written as $p_X(x)$, and a specific value in the marginal distribution written as $p_X(x_i)$.

The **conditional distribution** $p_{Y|X}(y|x)$ for a pair of random variables (X, Y) is the collection of probabilities

$$P(Y = y_j | X = x_i) = \frac{P(Y = y_j, X = x_i)}{P(X = x_j)} \text{ , for all pairs of values } (x_i, y_j).$$

Unlike the marginal distribution of Y, the conditional distribution of Y given X accounts for information from X.

Correlated random variables

Two random variables X and Y are **independent** if the probabilities

$$P(Y = y_j | X = x_i) = P(Y = y_j), \text{ for all pairs of values } (x_i, y_j).$$

Equivalently, X and Y are independent if the probabilities

$$P(Y = y_j \text{ and } X = x_i) = P(Y = y_j)P(X = x_i) \text{ , for all pairs of values } (x_i, y_j).$$

Two random variables that are not independent are called **correlated random variables**. The correlation between two random variables, ρ, is a measure of the strength of the relationship between them. When two random variables are positively correlated, they tend to increase or decrease together. If one of the variables increases while the other decreases (or vice versa) they are negatively correlated.

$$\rho_{X,Y} = \sum_i \sum_j p(i,j) \frac{(x_i - \mu_X)}{\text{sd}(X)} \frac{(y_j - \mu_Y)}{\text{sd}(Y)}$$

Variance of the sum or difference of two correlated random variables

When two random variables X and Y are correlated:

$$\text{Variance}(X + Y) = \text{Variance}(X) + \text{Variance}(Y) + 2\sigma_X \sigma_Y \rho_{X,Y}$$

$$\text{Variance}(X - Y) = \text{Variance}(X) + \text{Variance}(Y) - 2\sigma_X \sigma_Y \rho_{X,Y}$$

When random variables are positively correlated the variance of the sum or the difference of two variables will be larger than the sum of the two variances. When they are negatively correlated the variance of the sum or difference will be smaller than the sum of the two variances.

The standard deviation for the sum or difference is the square root of the variance.

1. Let X represent the outcome from a roll of a fair six-sided die. Then, toss a fair coin X times and let Y denote the number of tails observed.
   a) Consider the joint probability table of X and Y. How many entries are there total? How many entries equal 0?
   b) Compute the joint probability $P(X = 1, Y = 0)$.
   c) Compute the joint probability $P(X = 1, Y = 2)$.
   d) Compute the joint probability $P(X = 6, Y = 3)$.
   e) Compute the marginal probability $P(Y = 5)$.

2. Consider the following joint probability distribution:

	Y		
X	-1	0	1
-1	0.10	0	0.35
0	0	0.10	0.10
1	0.15	0.10	0.10

a) Compute $P(X > Y)$.

b) Calculate the marginal distributions.

c) Compute $E(X)$.

d) Compute $\text{Var}(Y)$.

e) Compute $\rho_{X,Y}$.

f) Calculate $P(X|Y = 0)$.

g) Compute $E(Y|X = 1)$.

h) Calculate $\text{Var}(X - Y)$.

3. *Simulating from a Joint Probability Distribution.* The code shown in the template writes a function called `rcat()` that simulates n random pairs from a joint probability distribution specified as a table. The output is a $n \times 2$ matrix of the randomly generated pairs. You do not need to know the details of the code that creates the `rcat()` function.

```
#load reshape package
library(reshape)

#write function
rcat = function(n, ptable) {
  pmatrix <- melt(ptable)
  rows <- which(rmultinom(n, 1, pmatrix$value) == 1, arr.ind = TRUE)[ ,'row']
  indices <- pmatrix[rows, c('X1','X2')]
  colnames(indices) <- c('i','j')
  rownames(indices) <- seq(1, nrow(indices))
  return(indices)
}
```

Run the following code to enter the joint probability distribution used in the previous problem, then simulate 10,000 draws from the joint probability distribution and summarize the results.

```
#enter joint probability distribution
X = -1:1; Y = -1:1
ptable = matrix(nrow = 3, byrow = TRUE, c(0.1, 0, 0.35,
                            0, 0.1, 0.1,
                            0.15, 0.1, 0.1))
rownames(ptable) = X; colnames(ptable) = Y

#check entries
ptable
```

```
##      -1   0    1
## -1 0.10 0.0 0.35
## 0  0.00 0.1 0.10
## 1  0.15 0.1 0.10
```

```
#set seed for pseudo-random sampling
set.seed(2019)
results = rcat(10000, ptable)
table(results)
```

```
##     j
## i      -1    0    1
##   -1  978    0 3564
##   0     0 1028  976
##   1  1445 1010  999
```

a) Based on the simulation results, what is $P(X = -1, Y = 1)$? How does this compare to the theoretical value?

b) Calculate $P(X > Y)$ from the simulated data and compare this to the theoretical value.