Data 605 Final Problem#1

Jean Jimenez

2023-12-04

Problem #1

Question

Using R, set a random seed equal to 1234 (i.e., set.seed(1234)). Generate a random variable X that has 10,000 continuous random uniform values between 5 and 15.Then generate a random variable Y that has 10,000 random *normal* values with a mean of 10 and a standard deviation of 2.89.

Probability. Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the median of the X variable, and the small letter "y" is estimated as the median of the Y variable. Interpret the meaning of all probabilities.

5 points a. P(X>x | X>y) b. P(X>x & Y>y) c. P(X<x | X>y)

5 points. Investigate whether P(X>x & Y>y)=P(X>x)P(Y>y) by building a table and evaluating the marginal and joint probabilities.

5 points. Check to see if independence holds by using Fisher's Exact Test and the Chi Square Test. What is the difference between the two? Which is most appropriate? Are you surprised at the results? Why or why not?

Work and Answer

```
library(stats)
set.seed(1234)

X = runif(10000, min = 5, max = 15)
Y = rnorm(10000, mean = 10, sd = 2.89)
```

a. P(X>x | X>y)

```
x=median(X)
y=median(Y)

pa=sum(X>x & X>y)/sum(X>y)

cat("a. \n Probability that X is greater than its median given that X is greater than Y's median is: \n
```

```
## a.
##  Probability that X is greater than its median given that X is greater than Y's median is:
##
##  1 or 100 %
```

b. P(X>x & Y>y)

```
pb=sum(X > x & Y > y) / length(X)

cat("b. \n Probability that both X is greater than its median and Y is greater than its median is: \n \
```

```
## b.
##  Probability that both X is greater than its median and Y is greater than its median is:
##
##  0.2507 or 25.07 %
```

c. P(X<x | X>y)

```
pc = sum(X < x & X > y) / sum(X > y)

cat("b. \n Probability that X is less than its median given that X is greater than Y's median.: \n \n",
```

```
## b.
##  Probability that X is less than its median given that X is greater than Y's median.:
##
##  0 or 0 %
```

P(X>x & Y>y) = P(X>x)P(Y>y)

```
PXx=sum(X>x)/length(X)
PYy=sum(Y>y)/length(Y)

table_of_p=data.frame(
  `P(X>x)`=PXx,
  `P(Y>y)`=PYy,
  `joint_probability`= pb,
  `Marginals_product`=PXx*PYy
)

print(table_of_p)
```

```
##   P.X.x. P.Y.y. joint_probability Marginals_product
## 1    0.5    0.5            0.2507              0.25
```

```
cat("Joint Probability is:", pb, "\n",
    "Marginals Probability is:", PXx*PYy, "\n \n ",
    "Since the Joint and Marginal probability are really close (essentially almost the same, we can say
```

```
## Joint Probability is: 0.2507
##  Marginals Probability is: 0.25
##
##   Since the Joint and Marginal probability are really close (essentially almost the same, we can say
```

Chi-Square vs. Fisher's Exact Test

Null Hypothesis: Observed values will NOT significantly deviate from expected.

Alt. Hypothesis: Observed values will NOT significantly deviate from expected.

```r
X_bin = ifelse(X > x, 1, 0)
Y_bin = ifelse(Y > y, 1, 0)

#contingency table

c_table = table(X_bin, Y_bin)

cs_test = chisq.test(c_table)
cat("\nChi-Square Test:\n")
```

Chi-Square

```
##
## Chi-Square Test:
```

```r
print(cs_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  c_table
## X-squared = 0.0676, df = 1, p-value = 0.7949
```

The low chi-squared value means that the observed values do not significantly deviate than the expected. Since the Odds Ratio is 1,the probability of one of the events occurring is different than another event occurring. The large P-value of 0.7949 means that we have a higher probability of observing what we did if the two variables are independent. The Chi-Squared Test supports the null hypothesis.

```r
fe_test = fisher.test(c_table)
cat("\nFisher's Exact Test:\n")
```

Fisher's Exact Test

```
##
## Fisher's Exact Test:
```

```r
print(fe_test)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  c_table
## p-value = 0.7949
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9342763 1.0946016
## sample estimates:
## odds ratio
##   1.011264
```

The large P-value of 0.7949 means that we have a higher probability of observing what we did if the two variables are independent. Since the Odds Ratio is 1,the probability of one of the events occurring is different than another event occurring. The Fisher's Exact Test supports the null hypothesis.

Comparing between the Two The results of both tests support the null hypothesis. The observed does not significantly deviate from the expected. We fail to reject the null hypothesis. Observed values will **NOT** significantly deviate from expected.

Since we have a large sample size with 10,000 observations, the chi-square is more appropriate. The Fisher's test gave the same p-value, supporting our conclusion even more.

I am not surprise by these results. Since we observed that $P(X>x \ \& \ Y>y) = P(X>x)P(Y>y)$, we know that both events occur independently and do not influence one another.