Homework 8 - Stat 495

Justin Papagelis

Due Wednesday, Nov. 16th by midnight (11:59 pm)

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle/Git repo, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

PROBLEMS TO TURN IN: Add 1

Additional 1

The goal for this problem is to practice using simulation to explore a potentially unfamiliar setting, and to then communicate your results and overall process.

We are all familiar with the usual Pearson correlation coefficient, r, which is computed pairwise between quantitative variables as a measure of the strength of their linear relationship. Other correlation type statistics exist - two nonparametric ones that are notable are Kendall’s Tau, τ and Spearman’s rho. We will focus on Kendall’s tau and Pearson’s r here. An example of how to obtain test output checking to see if the population parameters estimated by these values are statistically significantly different from 0 is shown below.

```
data(iris)
with(iris, cor.test(Sepal.Length, Sepal.Width))
```

```
##
##  Pearson’s product-moment correlation
##
## data:  x and y
## t = -1.44, df = 148, p-value = 0.152
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2726932  0.0435116
## sample estimates:
##      cor
## -0.11757
```

```
with(iris, cor.test(Sepal.Length, Sepal.Width, method = "kendall"))
```

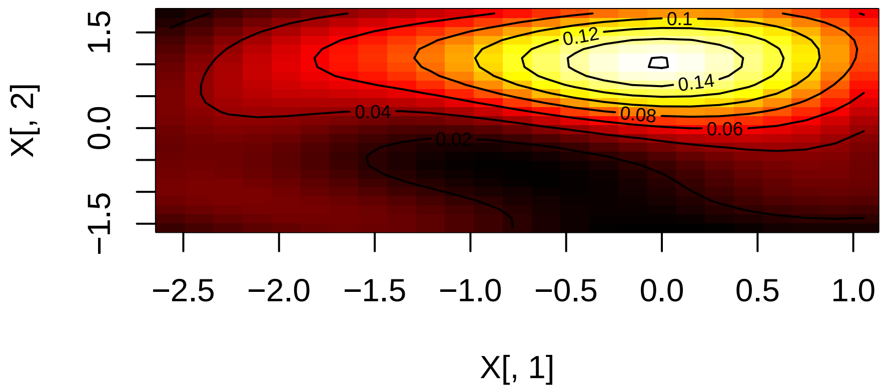
```
##
##  Kendall’s rank correlation tau
##
## data:  x and y
## z = -1.332, p-value = 0.183
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.0769968
```

Your task is to compare how the two procedures perform in terms of identifying significant relationships (significantly non-zero correlations) between the two variables by writing simulations for the following three settings:

- independent observations from two different distributions (your choice, but not say, two normals with different means; two different distributions, like a beta and a normal (pick another pair!))
- bivariate normal observations with a moderate correlation (say something between $0.3 < \rho < 0.5$)
- double exponentially distributed observations with a moderate correlation (covariance between 0.3 and 0.5)

In each setting, you should use 15 observations for a run (as one setting). You should use variances of 1 for the bivariate normal and double exponential. An example of generating from the double exponential is shown below. It requires a package that you may need to install.

```r
X <- rmvl(15, c(0,1), matrix(c(1, 0.2, 0.2, 1), nrow = 2)) #look at the help file!
joint.density.plot(X[, 1], X[, 2], color = TRUE)
```



If you want to adjust either n or the variances, feel free to do that after completing the requested case. This is not required, but some of you get curious and might want to check out other behaviors.

An outline for the solution is provided below for you to fill in. Please write in complete sentences. Note that you can understand what is going on here with 10,000 or fewer runs in each setting. Referring to the lab on Coding and Simulation may be useful, and remember that your submission should be reproducible. You can delete these instructions and prompts below in your submission, or keep for your reference.

For the sections below, these are potential outlines / suggestions of questions to answer you could follow. But be sure you craft the responses into sentences and paragraphs. Short answers to the bullet-points don't demonstrate your communication skills.

Simulation Comparing Kendall's tau and Pearson's r

Intro and Simulation Overview

- What is the task, as you understand it?
- What do you plan to do for your simulation (generally speaking) to compare the performance of the methods?
- Some questions to consider answering: Are you storing correlation values? p-values? Storing nothing? How many runs are you doing? Is a single run sufficient? What seems reasonable? What will you compare to assess the performance of the methods?
- You should aim to convey the general structure of your simulation in a paragraph or two, so that a reader can follow along, and could think about pseudocode steps based on your description.
- What do you expect to happen in the simulations? There are three settings. Describe what you expect to happen in each setting. You can check this with your simulations, but you should be able to make some predictions based on your statistical background knowledge.

The task of the simulation is to compare how Kendall's tau and Pearson's r perform in terms of identifying significant relationships between two variables from varying distributions. For the simulation, we will generate two variables, each from a specified distribution and then once we have that, we will store the p-value that shows if the population parameters estimated by these values are statistically significantly different than 0 for each method. Once we run this about 10,000 and get a large number of p-values, we can see how many of the p-values from each method are significant. A small proportion of significant p-values would tell us that most of the correlation tests failed to reject the null that the true correlation between the two distributions is 0.

We expect that for the independent variables from two different distributions, most of the correlation tests will fail to reject the null because the variables were independently chosen from two different distributions. For the variables from the bivariate normal distribution with a moderate correlation, we expect there to be a larger amount of tests that reject the null because the variables have a moderate correlation. The same goes for the variables from the double exponential distribution with moderate correlation. We would expect to see a larger amount of tests reject the null because the two variables are moderately correlated.

Independent Variables from Two Different Distributions

- independent observations from two different distributions (your choice, but not say, two normals with different means; two different distributions, like a beta and a normal (pick another pair!))

- What distributions did you choose? Report appropriate parameters.

- Simulation (R code)

- Summarize your results. How do the methods compare?

- How do your findings match your expectations?

We chose to use a Normal distribution and Gamma distribution. The Normal distribution has a mean of 40 and a standard deviation of 3. The Gamma distribution has a shape of 8 and a rate of 2.

```r
set.seed(1)

reps <- 10000
n <- 15

r_pvals <- rep(0, reps)
```

```r
tau_pvals <- rep(0, reps)

for(i in 1: reps) {
  x <- rnorm(n, 40, 3)
  y <- rgamma(n, 8, 2)

  values <- data.frame(x,y)

  r_pvals[i] <- with(values, cor.test(x, y))$p.value
  tau_pvals[i] <- with(values, cor.test(x, y, method = "kendall"))$p.value
}

sum(r_pvals <= 0.05)/reps
```

```
## [1] 0.0463
```

```r
sum(tau_pvals <= 0.05)/reps
```

```
## [1] 0.0436
```

It appears that Pearson's r rejects the null in 4.63% of the correlation tests and Kendall's tau rejects the null in 4.36% of the correlation tests. These are similar rejection rates which means that both of the methods perform in similar manners for independent observations from two different distributions. The findings match our expectations because we expected both of the methods to fail to reject a large number of the null hypotheses and that's what occurred.

Bivariate Normal with Moderate Correlation

- What correlation did you choose? (i.e. Report appropriate parameters)
- Simulation (R code)
- Summarize your results. How do the methods compare?
- How do your findings match your expectations?

Since we used a covariance matrix of $\begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$ we have a moderate correlation of 0.4.

```r
set.seed(1)

reps <- 10000
n <- 15

r_pvals <- rep(0, reps)
tau_pvals <- rep(0, reps)

for(i in 1: reps) {

  b <- rmvnorm(n, mean = c(0,1), sigma = matrix(c(1, 0.4, 0.4, 1), nrow = 2)) #correlation of 0.4
  values <- data.frame(b)

  r_pvals[i] <- with(values, cor.test(X1, X2))$p.value
  tau_pvals[i] <- with(values, cor.test(X1, X2, method = "kendall"))$p.value
}

sum(r_pvals <= 0.05)/reps
```

```
## [1] 0.3318
```

```
sum(tau_pvals <= 0.05)/reps
```

```
## [1] 0.268
```

It appears that the Pearson's r rejects the null hypotheses in 33.18% of the correlation tests while Kendall's tau rejects the null in 26.8% of the cases. Pearson's r performs better because we know that there is a moderate correlation between the two variables.. It seems that Pearson's r performs better than Kendall's tau with this distribution. This matches what our prediction was for this distribution. A larger amount of tests rejected the null hypothesis that the true correlation is 0.

Double Exponential with Moderate Correlation

- What correlation (or covariance) did you choose? (i.e. Report appropriate parameters)
- Simulation (R code)
- Summarize your results. How do the methods compare?
- How do your findings match your expectations?

Since we used a covariance matrix of $\begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$ we have a moderate correlation of 0.4.

```
set.seed(1)

reps <- 10000
n <- 15

r_pvals <- rep(0, reps)
tau_pvals <- rep(0, reps)

for(i in 1: reps) {

  b <- rmvl(15, c(0,1), matrix(c(1, 0.4, 0.4, 1), nrow = 2)) #correlation of 0.4
  values <- data.frame(b)

  r_pvals[i] <- with(values, cor.test(X1, X2))$p.value
  tau_pvals[i] <- with(values, cor.test(X1, X2, method = "kendall"))$p.value
}

sum(r_pvals <= 0.05)/reps
```

```
## [1] 0.3778
```

```
sum(tau_pvals <= 0.05)/reps
```

```
## [1] 0.291
```

The Pearson's r rejects the null in 37.78% of the correlation tests while Kendall's tau rejects the null in 29.1% of the cases. Pearson's r performs better because we know there is a moderate correlation between the two variables. As well as before, Pearson's r performs better than Kendall's tau. This matches what our prediction was for this distribution.

Conclusion

- Sum up what you found.
- What takeaway would you want a reader to get from your submission?

Overall, in the simulation of variables from independent distributions, Pearson's p and Kendall's tau performed similarly in rejecting the null hypothesis, however for the next two simulations where there was a known moderate correlation, Pearson's p outperformed Kendall's tau in rejecting the null hypothesis that the true correlation between the variables is 0.