Homework 8 - Stat 495

Cassandra Jin

Due Wednesday, Nov. 29th 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 use a simulation study 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)
cor.test(Sepal.Length ~ Sepal.Width, data = iris)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Sepal.Length and Sepal.Width
## 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
```

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

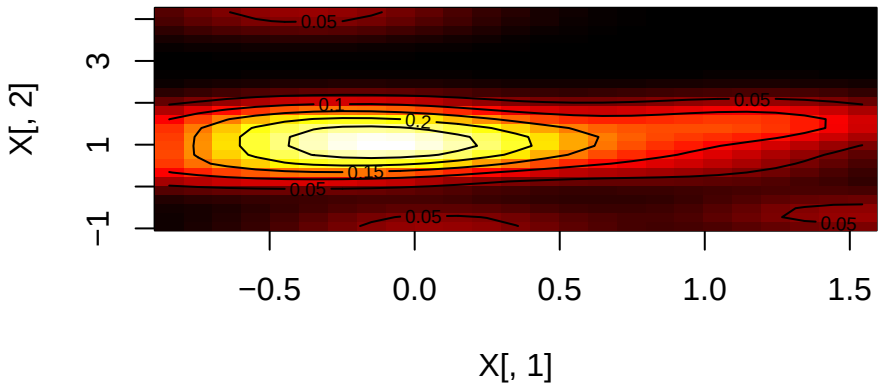
```
##
##  Kendall's rank correlation tau
##
## data:  Sepal.Length and Sepal.Width
## z = -1.332, p-value = 0.183
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.0769968
```

Your task is to compare how these two procedures perform in terms of identifying significant relationships (significantly non-zero correlations) or the lack thereof, between the two variables by performing a simulation study covering 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, code below)

In each setting, you should use 15 observations for a single 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.

```
library(LaplacesDemon) #move up to normal package chunk in submission
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. The submission must cover those 3 settings with n = 15 and the specified variances, at a minimum.

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 Simulation Studies 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 τ and Pearson's r

Intro and Simulation Overview

Kendall's τ and Pearson's r are both measures of strength of the linear relationship between two quantitative variables, and our task is to compare how these two procedures perform in terms of identifying whether or not a significant relationship is present by conducting a simulation study. The simulation study will be run under three settings: independent observations from two different distributions, bivariate normal observations with a moderate correlation, and double exponentially distributed observations with a moderate correlation.

We will conduct 10,000 runs of each setting of the simulation (sufficient repetitions to obtain a comprehensive understanding of the comparison in all three cases, considerably small variance), with $n = 15$ observations for each run and a variance of 1 for the bivariate normal and double exponential settings. Throughout the simulations, we will store the correlation values and their associated p-values, and afterwards we will count the frequency of correlations with significantly small p-values (0.05 cutoff) for both procedures in each setting. Comparing these frequencies will show us the relative performances of the methods.

In the setting of observing independent observations from two different distributions, we expect to find (close to) no significant correlations. In the settings of bivariate normal observations with a moderate correlation and double exponentially distributed observations with a moderate correlation, we expect to find moderately frequent significant correlation values.

Independent Variables from Two Different Distributions

We will observe the relationship between the data from a Cauchy distribution (location = 40, scale = 3) and an exponential distribution (rate = 1). Since the variables are chosen to be independent and continuously distributed, we expect both procedures to rarely identify significant correlations. So the simulation should yield very small proportions of demonstrated significant correlation values for both methods.

```
sim_diff <- function(locationinput, scaleinput, rateinput, ninput = 15, repsinput = 10000){
# Set Useful Values
reps <- repsinput #number of repetitions
n <- ninput #sample size
location <- locationinput
scale <- scaleinput
rate <- rateinput

#Initialize storage vectors
kendall_pvals <- rep(0,reps)
pearson_pvals <- rep(0,reps)

#Generate random data, do tests, save p-values
for(i in 1:reps){
  cauchy_x <- rcauchy(n, location, scale)
  exponential_x <- rexp(n, rate)
  kendall_pvals[i] <- cor.test(cauchy_x, exponential_x, method = "kendall")$p.value
  pearson_pvals[i] <- cor.test(cauchy_x, exponential_x)$p.value
}

output <- c(sum(kendall_pvals <= 0.05)/reps, sum(pearson_pvals <= 0.05)/reps)
names(output) <- c("Kendall's tau", "Pearson's r")
output
}
```

```
set.seed(495)
sim_diff(locationinput = 40, scaleinput = 3, rateinput = 1)
```

```
## Kendall's tau   Pearson's r
##        0.0457        0.0582
```

In this setting of the simulation study for the two correlation test performances between independent variables from a Cauchy distribution and an exponential distribution, we find that both methods resulted in very small proportions of runs yielding significant correlations (p-value < 0.05), which matches our expectation. However, the proportion of significant correlation values obtained through Pearson's r, 0.0582, is slightly larger than that from Kendall's τ, 0.0457. Thus, we see that, in this setting with the specific seed we set, the Pearson's r method produced slightly more false positives than Kendall's τ.

Bivariate Normal with Moderate Correlation (hw on chapters 4 + 5)

For a moderate correlation setting, we will choose a correlation value of 0.4, and we expect both procedures to identify moderate proportions of significant correlations.

```
sim_bivar <- function(correlationinput, ninput = 15, repsinput = 10000){
# Set Useful Values
reps <- repsinput #number of repetitions
n <- ninput #sample size
correlation <- correlationinput #target correlation for bivariate normal distribution

#Initialize storage vectors
kendall_pvals <- rep(0,reps)
pearson_pvals <- rep(0,reps)

#Generate random data, do tests, save p-values
for(i in 1:reps){
  bivar_x <- mvrnorm(n = ninput, mu = c(0, 0),
                           Sigma = matrix(c(1, correlation, correlation, 1), nrow = 2))
  kendall_pvals[i] <- cor.test(bivar_x[,1], bivar_x[,2], method = "kendall")$p.value
  pearson_pvals[i] <- cor.test(bivar_x[,1], bivar_x[,2])$p.value
}

output <- c(sum(kendall_pvals <= 0.05)/reps, sum(pearson_pvals <= 0.05)/reps)
names(output) <- c("Kendall's tau", "Pearson's r")
output
}
```

```
set.seed(495)
sim_bivar(correlationinput = 0.4)
```

```
## Kendall's tau   Pearson's r
##        0.2707        0.3286
```

In this setting of the simulation study for the two correlation test performances for bivariate normal variables with moderate correlation, we find that both methods found moderate proportions of runs yielding significant correlations (p-value < 0.05), which matches our expectation. Similar to the simulation for two independent variables from different distributions, the proportion of significant correlation values obtained through Pearson's r, 0.3286, is larger than that from Kendall's τ, 0.2707. In this setting, the difference between the two demonstrated proportions (≈ 0.05) is larger than that from the first setting (≈ 0.01), where both proportions were very small and quite close. We see that, when testing for significant correlation values in the bivariate normal setting, the Pearson's r method was slightly more likely to detect significant correlations than Kendall's τ and closer to the true correlation value.

Double Exponential with Moderate Correlation

- What correlation (or covariance) did you choose? (i.e. Report appropriate parameters)

```
sim_doubexp <- function(covarianceinput, ninput = 15, repsinput = 10000){
# Set Useful Values
reps <- repsinput #number of repetitions
n <- ninput #sample size
covariance <- covarianceinput #target covariance value for double exponential distribution

#Initialize storage vectors
kendall_pvals <- rep(0,reps)
pearson_pvals <- rep(0,reps)

#Generate random data, do tests, save p-values
for(i in 1:reps){
  doubexp_x <- rmvl(n, mu = c(0, 0),
                    Sigma = matrix(c(1, covariance, covariance, 1), nrow = 2))
  kendall_pvals[i] <- cor.test(doubexp_x[,1], doubexp_x[,2], method = "kendall")$p.value
  pearson_pvals[i] <- cor.test(doubexp_x[,1], doubexp_x[,2])$p.value
}

output <- c(sum(kendall_pvals <= 0.05)/reps, sum(pearson_pvals <= 0.05)/reps)
names(output) <- c("Kendall's tau", "Pearson's r")
output
}
```

```
set.seed(495)
sim_doubexp(covarianceinput = 0.4)
```

```
## Kendall's tau   Pearson's r
##        0.2961        0.3756
```

In this setting of the simulation study for the two correlation test performances for double exponential variables with moderate correlation, we find that both methods again found moderate proportions of runs yielding significant correlations (p-value < 0.05), which matches our expectation. Similar to both previous simulations, the proportion of significant correlation values obtained through Pearson's r, 0.3756, is larger than that from Kendall's τ, 0.2961. In this setting, the difference between the two demonstrated proportions (≈ 0.08) is larger than that from the bivariate normal setting (≈ 0.05). Overall, when testing for significant correlation values in the double exponential setting, the Pearson's r method was slightly more likely to detect significant correlations than Kendall's τ and closer to the true correlation value.

Conclusion

In order to evaluate the performance of two measures of strength of the linear relationship between two quantitative variables, Kendall's τ and Pearson's r, we conducted a simulation study with three settings: independent observations from two different distributions, bivariate normal observations with a moderate correlation, and double exponentially distributed observations with a moderate correlation. Each setting was run 10,000 times with $n = 15$ observations per run, and we observed the proportion of significant correlation values that the two methods identified in each setting.

Our expectations were met for all three cases: both Kendall's τ and Pearson's r methods detected very few instances of significant correlations between independent observations from two different distributions, and the two procedures identified moderately frequent significant correlation values for bivariate normal observations with a moderate correlation and double exponentially distributed observations with a moderate correlation. More importantly, in all three settings, the Pearson's r method was slightly more likely to detect significant

correlations than Kendall's τ and found proportions closer to the true values, meaning that Kendall's τ is more conservative than Pearson's r, and Pearson's r performs better.