Homework9

Akhil Havaldar

11/12/2022

Question 1

a

```
library(readr)
dat1 <- read_csv("data1 (1).csv")
cor1 <- cor(dat1$V1, dat1$V2)
cor1
```

```
## [1] 0.524066
```

b

```
set.seed(05202001)
paired <- function(size) {
  rows <- dat1[sample(nrow(dat1), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans1 <- sapply(c(13), test)
bans1
```

```
## [1] 0.0493
```

c

```
ttest2 <- function(size) {
  rows <- dat1[sample(nrow(dat1), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans1 <- sapply(c(13), test2)
cans1
```

```
## [1] 0.0079
```

Question 2

a

```
dat2 <- read_csv("data2.csv")
cor2 <- cor(dat2$V1, dat2$V2)
cor2
```

```
## [1] -0.52036
```

b

```
set.seed(05202001)
paired <- function(size) {
  rows <- dat2[sample(nrow(dat2), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans2 <- sapply(c(13), test)
bans2
```

```
## [1] 0.052
```

c

```
ttest2 <- function(size) {
  rows <- dat2[sample(nrow(dat2), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans2 <- sapply(c(13), test2)
cans2
```

```
## [1] 0.1057
```

Question 3

a

```
dat3 <- read_csv("data3.csv")
cor3 <- cor(dat3$V1, dat3$V2)
cor3
```

```
## [1] 0.002426237
```

b

```
set.seed(05202001)
paired <- function(size) {
  rows <- dat3[sample(nrow(dat3), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans3 <- sapply(c(13), test)
bans3
```

```
## [1] 0.0493
```

c

```
ttest2 <- function(size) {
  rows <- dat3[sample(nrow(dat3), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans3 <- sapply(c(13), test2)
cans3
```

```
## [1] 0.0521
```

Question 4

```
q1 <- c(cor1, bans1, cans1)
q2 <- c(cor2, bans2, cans2)
q3 <- c(cor3, bans3, cans3)

df1 <- data.frame(q1, q2, q3)
rownames(df1) <- c("cor", "paired", "2-samp")
df1 <- t(df1)
df1
```

```
##             cor paired 2-samp
## q1  0.524066001 0.0493 0.0079
## q2 -0.520360008 0.0520 0.1057
## q3  0.002426237 0.0493 0.0521
```

- We can see that for each dataset, the paired t-test gives us similar values for type 1 error. For the 2-sample t test, the type 1 error values become much more varied.

Question 5

a

```
dat4 <- read_csv("data4.csv")
cor4 <- cor(dat4$V1, dat4$V2)
cor4
```

```
## [1] 0.5906402
```

b

```
set.seed(05202001)
paired <- function(size) {
  rows <- dat4[sample(nrow(dat4), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans4 <- sapply(c(13), test)
bans4
```

```
## [1] 0.0554
```

c

```
ttest2 <- function(size) {
  rows <- dat4[sample(nrow(dat4), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans4 <- sapply(c(13), test2)
cans4
```

```
## [1] 0.0134
```

Question 6

a

```r
dat5 <- read_csv("data5.csv")
cor5 <- cor(dat5$V1, dat5$V2)
cor5
```

```
## [1] -0.5721193
```

b

```r
set.seed(05202001)
paired <- function(size) {
  rows <- dat5[sample(nrow(dat5), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans5 <- sapply(c(13), test)
bans5
```

```
## [1] 0.0678
```

c

```r
ttest2 <- function(size) {
  rows <- dat5[sample(nrow(dat5), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans5 <- sapply(c(13), test2)
cans5
```

```
## [1] 0.1128
```

Question 7

a

```r
dat6 <- read_csv("data6.csv")
cor6 <- cor(dat6$V1, dat6$V2)
cor6
```

```
## [1] -0.007297158
```

b

```r
set.seed(05202001)
paired <- function(size) {
  rows <- dat6[sample(nrow(dat6), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided", paired=TRUE)$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test <- function(size){
  samps <- replicate(K, paired(size))
  length(which(samps == TRUE)) / 10000
}

bans6 <- sapply(c(13), test)
bans6
```

```
## [1] 0.0348
```

c

```r
ttest2 <- function(size) {
  rows <- dat6[sample(nrow(dat6), size), ]
  pval <- t.test(rows$V1, rows$V2, mu=0, alternative="two.sided")$p.value
  if(pval < 0.05){TRUE}
  else{FALSE}
}

K <- 10000
test2 <- function(size){
  samps <- replicate(K, ttest2(size))
  length(which(samps == TRUE)) / 10000
}

cans6 <- sapply(c(13), test2)
cans6
```

```
## [1] 0.0374
```

Question 8

```
q5 <- c(cor4, bans4, cans4)
q6 <- c(cor5, bans5, cans5)
q7 <- c(cor6, bans6, cans6)

df2 <- data.frame(q5, q6, q7)
rownames(df2) <- c("cor", "paired", "2-samp")
df2 <- t(df2)
df2
```

```
##              cor paired 2-samp
## q5  0.590640226 0.0554 0.0134
## q6 -0.572119341 0.0678 0.1128
## q7 -0.007297158 0.0348 0.0374
```

- The skewed data gives us much more variability in the paired t test type 1 error value, but the 2-samp t test type 1 error value remains varied.

```
df1
```

```
##              cor paired 2-samp
## q1  0.524066001 0.0493 0.0079
## q2 -0.520360008 0.0520 0.1057
## q3  0.002426237 0.0493 0.0521
```

```
df2
```

```
##              cor paired 2-samp
## q5  0.590640226 0.0554 0.0134
## q6 -0.572119341 0.0678 0.1128
## q7 -0.007297158 0.0348 0.0374
```