STA104 - Homework 4

Xiaowei Zeng

2023-12-08

Problem 1

a. Exercise 1 (P141)

1. Compute all pairwise differences D_i.

Trt1	Trt2	difference
250	240	-10
50	48	-2
80	72	-8
55	47	-8
188	230	42

2. Obtain all possible assignments of plus and minus to the D_i and calculate the corresponding D_i.

D_1	D_2	D_3	D_4	D_5	$\bar{D}$
10	2	8	8	42	14
10	2	8	8	-42	-2.8
10	2	8	-8	42	10.8
10	2	8	-8	-42	-6
10	2	-8	8	42	10.8
10	2	-8	8	-42	-6
10	2	-8	-8	42	7.6
10	2	-8	-8	-42	-9.2
10	-2	8	8	42	13.2
10	-2	8	8	-42	-3.6
10	-2	8	-8	42	10
10	-2	8	-8	-42	-6.8
10	-2	-8	8	42	10
10	-2	-8	8	-42	-6.8
10	-2	-8	-8	42	6.8
10	-2	-8	-8	-42	-10
-10	2	8	8	42	10
-10	2	8	8	-42	-6.8
-10	2	8	-8	42	6.8
-10	2	8	-8	-42	-10
-10	2	-8	8	42	6.8
-10	2	-8	8	-42	-10

D_1	D_2	D_3	D_4	D_5	$\bar{D}$
-10	2	-8	-8	42	3.6
-10	2	-8	-8	-42	-13.2
-10	-2	8	8	42	9.2
-10	-2	8	8	-42	-7.6
-10	-2	8	-8	42	6
-10	-2	8	-8	-42	-10.8
-10	-2	-8	8	42	6
-10	-2	-8	8	-42	-10.8
-10	-2	-8	-8	42	2.8
-10	-2	-8	-8	-42	-14

3. Obtain the permutation distribution of the mean of the differences.

$\bar{D}$	-14	-13.2	-10.8	-10	-9.2	-7.6	-6.8	-6	-3.6	-2.8
Pr	0.03125	0.03125	0.0625	0.09375	0.03125	0.03125	0.09375	0.0625	0.03125	0.03125

$\bar{D}$	2.8	3.6	6	6.8	7.6	9.2	10	10.8	13.2	14
Pr	0.03125	0.03125	0.0625	0.09375	0.03125	0.03125	0.09375	0.0625	0.03125	0.03125

The permutation distribution has heavy tails. It may be attributed to the unexpectedly large difference in the fifth pair.

b. Exercise 3 (P142)

1. Give $|D_i|$ the ranks.

Trt1	Trt2	difference	rank
250	240	-10	4
50	48	-2	1
80	72	-8	2.5
55	47	-8	2.5
188	230	42	5

2. Obtain all possible assignments of plus and minus to the R_i and calculate the corresponding SR_+.

R_1	R_2	R_3	R_4	R_5	SR_+
4	1	2.5	2.5	5	15
4	1	2.5	2.5	-5	10
4	1	2.5	-2.5	5	12.5
4	1	2.5	-2.5	-5	7.5
4	1	-2.5	2.5	5	12.5
4	1	-2.5	2.5	-5	7.5
4	1	-2.5	-2.5	5	10

R_1	R_2	R_3	R_4	R_5	SR_+
4	1	-2.5	-2.5	-5	5
4	-1	2.5	2.5	5	14
4	-1	2.5	2.5	-5	9
4	-1	2.5	-2.5	5	11.5
4	-1	2.5	-2.5	-5	6.5
4	-1	-2.5	2.5	5	11.5
4	-1	-2.5	2.5	-5	6.5
4	-1	-2.5	-2.5	5	9
4	-1	-2.5	-2.5	-5	4
-4	1	2.5	2.5	5	11
-4	1	2.5	2.5	-5	6
-4	1	2.5	-2.5	5	8.5
-4	1	2.5	-2.5	-5	3.5
-4	1	-2.5	2.5	5	8.5
-4	1	-2.5	2.5	-5	3.5
-4	1	-2.5	-2.5	5	6
-4	1	-2.5	-2.5	-5	1
-4	-1	2.5	2.5	5	10
-4	-1	2.5	2.5	-5	5
-4	-1	2.5	-2.5	5	7.5
-4	-1	2.5	-2.5	-5	2.5
-4	-1	-2.5	2.5	5	7.5
-4	-1	-2.5	2.5	-5	2.5
-4	-1	-2.5	-2.5	5	5
-4	-1	-2.5	-2.5	-5	0

3. Obtain the permutation distribution of the signed-rank statistic SR_+ for the data.

SR_+	0	1	2.5	3.5	4	5	6	6.5	7.5
Pr	0.03125	0.03125	0.0625	0.0625	0.03125	0.09375	0.0625	0.0625	0.125

SR_+	8.5	9	10	11	11.5	12.5	14	15
Pr	0.0625	0.0625	0.09375	0.03125	0.0625	0.0625	0.03125	0.03125

The permutation distribution of signed rank sum statistics seems more likely to be normally distributed. It is robust against the outlier to some extent.

Problem 2

Load the dataset for Problem 2.

```
data = data.frame(Subject = 1:8,
                  Before = c(89, 90, 87, 98, 120, 85, 97, 110),
                  After = c(76, 101, 84, 86, 105, 84, 93, 115))
data$Difference = data$After - data$Before
head(data)
```

```
##   Subject Before After Difference
## 1       1     89    76        -13
## 2       2     90   101         11
## 3       3     87    84         -3
## 4       4     98    86        -12
## 5       5    120   105        -15
## 6       6     85    84         -1
```

Use Wilcoxon's signed-rank statistic to test whether there is a significant difference between the LDH readings before and after fasting. Suppose that we want to conduct a left-tail test and the significance level is 0.05.

$$H_0 : \theta_{\text{after}} = \theta_{\text{before}} \quad \text{v.s.} \quad H_1 : \theta_{\text{after}} < \theta_{\text{before}}.$$

First check that there are no ties in the data.

```
duplicated(abs(data$Difference))
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Compute the signed-rank statistic for the original data.

```
SR_obs = sum((sign(data$Difference) + 1) / 2 * rank(abs(data$Difference)))
SR_obs
```

```
## [1] 9
```

a. Exact p-value

Define the function of obtaining the permutations of the signs of the differences.

```
permute_sign = function(n, rank = FALSE){
  out = matrix(ncol = n, nrow = 2 ^ n)
  for (i in 1:n){
    out[, i] = rep(c(1, ifelse(rank, 0, -1)), each = 2 ^ (n - i))
  }
  return (out)
}
```

Obtain the permutation distribution of the signed-rank statistic SR_+ and calculate the exact p-value.

```
n = nrow(data)
signs = permute_sign(n, rank = TRUE)
perm_dist = rowSums(signs * matrix(rank(abs(data$Difference)),
                                   nrow = nrow(signs),
                                   ncol = ncol(signs),
                                   byrow = TRUE))
p_value1 = sum(perm_dist <= SR_obs) / length(perm_dist)
p_value1
```

```
## [1] 0.125
```

The p-value of the exact signed-rank test is $0.125 > 0.05$, indicating that we cannot reject the null hypothesis and the difference is not statistically significant.

b. Approximate p-values

Use normal approximation to obtain the p-value.

$$Z = \frac{\text{SR}_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

```
Z1 = (SR_obs - n * (n + 1) / 4) / sqrt(n * (n + 1) * (2 * n + 1) / 24)
p_value2 = pnorm(Z1)
p_value2
```

```
## [1] 0.1037892
```

The approximate p-value without continuity correction is 0.104, a bit smaller than the exact p-value. $0.104 > 0.05$, so the conclusion is the same as the exact test, that is we cannot reject the null hypothesis and the difference is not statistically significant.

Try using continuity correction.

```
Z2 = (SR_obs + 0.5 - n * (n + 1) / 4) / sqrt(n * (n + 1) * (2 * n + 1) / 24)
p_value3 = pnorm(Z2)
p_value3
```

```
## [1] 0.1169764
```

The approximate p-value with continuity correction is 0.117, which is very close to the exact p-value of 0.125. $0.117 > 0.05$, so the conclusion is the same as the exact test, that is we cannot reject the null hypothesis and the difference is not statistically significant.

Problem 3

a. Exercise 1 (P189)

1. Permute the weight data. There are 24 possible orders of weight.

W1	W2	W3	W4
120	145	153	162
120	145	162	153
120	162	145	153
162	120	145	153
162	120	153	145
120	162	153	145
120	153	162	145
120	153	145	162
153	120	145	162
153	120	162	145
153	162	120	145
162	153	120	145
162	153	145	120
153	162	145	120

W1	W2	W3	W4
153	145	162	120
153	145	120	162
145	153	120	162
145	153	162	120
145	162	153	120
162	145	153	120
162	145	120	153
145	162	120	153
145	120	162	153
145	120	153	162

2. For height and each permuted weight, we calculate the slope of the least squares line.

$$\text{slope}_{\text{OLS}} = \frac{\sum_{i=1}^{n} \left(H_i - \bar{H}\right)\left(W_i - \bar{W}\right)}{\sum_{i=1}^{n} \left(H_i - \bar{H}\right)^2}$$

$$\begin{aligned}\sum_{i=1}^{n} \left(H_i - \bar{H}\right)^2 &= (57-67.5)^2 + (65-67.5)^2 + (70-67.5)^2 + (78-67.5)^2 \\ &= 233\end{aligned}$$

W1	W2	W3	W4	$\text{slope}_{\text{OLS}}$
120	145	153	162	1.98
120	145	162	153	1.67
120	162	145	153	1.3
162	120	145	153	-0.14
162	120	153	145	-0.41
120	162	153	145	1.03
120	153	162	145	1.22
120	153	145	162	1.81
153	120	145	162	0.67
153	120	162	145	0.09
153	162	120	145	-0.81
162	153	120	145	-1.12
162	153	145	120	-1.98
153	162	145	120	-1.67
153	145	162	120	-1.3
153	145	120	162	0.14
145	153	120	162	0.41
145	153	162	120	-1.03
145	162	153	120	-1.22
162	145	153	120	-1.81
162	145	120	153	-0.67
145	162	120	153	-0.09
145	120	162	153	0.81
145	120	153	162	1.12

3. Obtain the permutation distribution of the slope of the least squares line for the height and weight data.

slope	-1.98	-1.81	-1.67	-1.3	-1.22	-1.12	-1.03	-0.81	-0.67	-0.41	-0.14	-0.09
Pr	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417

slope	0.09	0.14	0.41	0.67	0.81	1.03	1.12	1.22	1.3	1.67	1.81	1.98
Pr	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417	0.0417

b. Exercise 2 (P189)

i. Spearman Rank Correlation

1. Permute the rank of weight data. There are 24 possible orders of ranks.

R(W1)	R(W2)	R(W3)	R(W4)
1	2	3	4
1	2	4	3
1	4	2	3
4	1	2	3
4	1	3	2
1	4	3	2
1	3	4	2
1	3	2	4
3	1	2	4
3	1	4	2
3	4	1	2
4	3	1	2
4	3	2	1
3	4	2	1
3	2	4	1
3	2	1	4
2	3	1	4
2	3	4	1
2	4	3	1
4	2	3	1
4	2	1	3
2	4	1	3
2	1	4	3
2	1	3	4

2. For the rank of height and each permuted rank of weight, we calculate Spearman's r_S.

$$r_S = 1 - \frac{\sum_{i=1}^{n}[R(Y_i) - R(W_i)]^2}{n(n^2-1)/6}$$

R(W1)	R(W2)	R(W3)	R(W4)	r_S
1	2	4	3	0.8
1	4	2	3	0.4

R(W1)	R(W2)	R(W3)	R(W4)	r_S
4	1	2	3	-0.2
4	1	3	2	-0.4
1	4	3	2	0.2
1	3	4	2	0.4
1	3	2	4	0.8
3	1	2	4	0.4
3	1	4	2	0
3	4	1	2	-0.6
4	3	1	2	-0.8
4	3	2	1	-1
3	4	2	1	-0.8
3	2	4	1	-0.4
3	2	1	4	0.2
2	3	1	4	0.4
2	3	4	1	-0.2
2	4	3	1	-0.4
4	2	3	1	-0.8
4	2	1	3	-0.4
2	4	1	3	0
2	1	4	3	0.6
2	1	3	4	0.8

3. Obtain the permutation distribution of Spearman's r_s. (Obviously there are no ties in the dataset.)

r_S	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
Pr	0.0417	0.125	0.0417	0.1667	0.0833	0.0833	0.0833	0.1667	0.0417	0.125	0.0417

ii. Kendall's Tau

1. For height and each permuted weight, we calculate the slope of the least squares line.

$$\tau = 2P\left[\left(X_i - X_j\right)\left(Y_i - Y_j\right) > 0\right] - 1$$

W1	W2	W3	W4	τ
120	145	153	162	1
120	145	162	153	0.67
120	162	145	153	0.33
162	120	145	153	0
162	120	153	145	-0.33
120	162	153	145	0
120	153	162	145	0.33
120	153	145	162	0.67
153	120	145	162	0.33
153	120	162	145	0
153	162	120	145	-0.33
162	153	120	145	-0.67

W1	W2	W3	W4	τ
162	153	145	120	-1
153	162	145	120	-0.67
153	145	162	120	-0.33
153	145	120	162	0
145	153	120	162	0.33
145	153	162	120	0
145	162	153	120	-0.33
162	145	153	120	-0.67
162	145	120	153	-0.33
145	162	120	153	0
145	120	162	153	0.33
145	120	153	162	0.67

2. Obtain the permutation distribution of Kendall's τ.

τ	-1	-0.67	-0.33	0	0.33	0.67	1
Pr	0.0417	0.125	0.2083	0.25	0.2083	0.125	0.0417

Problem 4

Load the dataset for Problem 4.

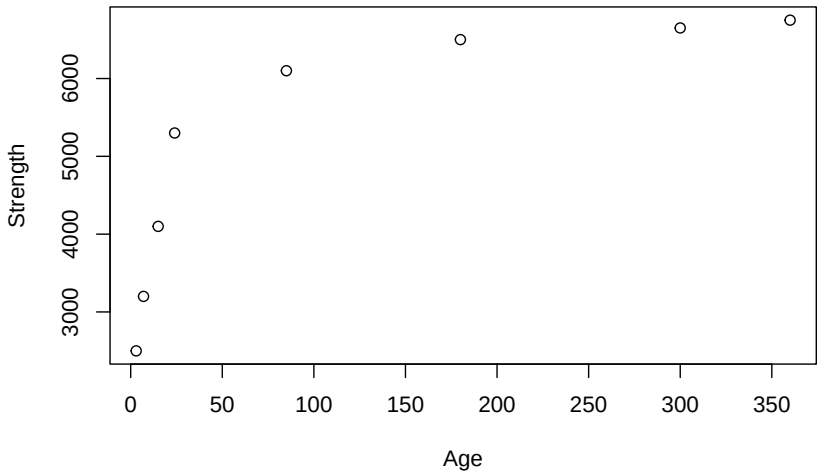
```
data = data.frame(Age = c(3, 7, 15, 24, 85, 180, 300, 360),
                  Strength = c(2500, 3200, 4100, 5300,
                               6100, 6500, 6650, 6750))
head(data)
```

```
##   Age Strength
## 1   3     2500
## 2   7     3200
## 3  15     4100
## 4  24     5300
## 5  85     6100
## 6 180     6500
```

a. Plot the data to show a nonlinear relationship. Compute Pearson's correlation, Spearman's correlation, and Kendall's tau.

Plot the data points.

```
attach(data)
plot(Age, Strength)
```



Obviously, the relationship between age and strength is nonlinear.

i. Pearson's correlation The Pearson's correlation is 0.80.

```
r_p = sum((Age - mean(Age)) * (Strength - mean(Strength))) /
  sqrt(sum((Age - mean(Age)) ^ 2) * sum((Strength - mean(Strength)) ^ 2))
r_p
```

```
## [1] 0.7999108
```

ii. Spearman's correlation The Spearman's correlation is 1.

```
R_Age = rank(Age)
R_Str = rank(Strength)
r_s = sum((R_Age - mean(R_Age)) * (R_Str - mean(R_Str))) /
  sqrt(sum((R_Age - mean(R_Age)) ^ 2) * sum((R_Str - mean(R_Str)) ^ 2))
r_s
```

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

iii. Kendall's tau Define a function of calculating the pairwise difference.

```
calc_diff = function(x){
  n = length(x)
  grid = matrix(x, nrow = n, ncol = n) -
    matrix(x, nrow = n, ncol = n, byrow = T)
  return (grid[upper.tri(grid)])
}
```

The Kendall's tau is 1.

```
tau = 2 * mean(calc_diff(Age) * calc_diff(Strength) > 0) - 1
tau
```

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

b. Test for significant association using each of the measures of association in part a.

Suppose that we want to conduct a two-sided test and the significance level is 0.05.

i. Pearson's correlation Define the function of obtaining permutations.

```
permute_x = function(x){
  n <- length(x)
  out <- matrix(nrow = factorial(n), ncol = n)
  p <- ip <- seqn <- 1:n
  d <- rep(-1, n)
  d[1] <- 0
  m <- n + 1
  p <- c(m, p, m)
  i <- 1
  use <- -c(1, n + 2)
  while (m != 1) {
    out[i, ] <- x[p[use]]
    i <- i + 1
    m <- n
    chk <- (p[ip + d + 1] > seqn)
    m <- max(seqn[!chk])
    if (m < n) {d[(m + 1):n] <- -d[(m + 1):n]}
    index1 <- ip[m] + 1
    index2 <- p[index1] <- p[index1 + d[m]]
    p[index1 + d[m]] <- m
    tmp <- ip[index2]
    ip[index2] <- ip[m]
    ip[m] <- tmp
  }
  return (out)
}
```

The p-value for Pearson's correlation is $0.0185 < 0.05$, indicating that the correlation of age and strength is significant.

```
perm_Str = permute_x(Strength)
perm_Age = matrix(Age,
                  nrow = nrow(perm_Str),
                  ncol = ncol(perm_Str),
                  byrow = TRUE)
perm_dist = rowSums((perm_Age - rowMeans(perm_Age)) *
  (perm_Str - rowMeans(perm_Str))) /
  sqrt(rowSums((perm_Age - rowMeans(perm_Age)) ^ 2) *
         rowSums((perm_Str - rowMeans(perm_Str)) ^ 2))
p_value1 = mean(abs(perm_dist) >= abs(r_p))
p_value1
```

```
## [1] 0.01850198
```

Or we can use the t_{n-2} approximation to obtain the approximate p-value.

```
n = nrow(data)
t_obs = r_p * sqrt((n - 2) / (1 - r_p ^ 2))
p_value1 = 2 * min(1 - pt(t_obs, df = n - 2), pt(t_obs, df = n - 2))
p_value1
```

```
## [1] 0.01714168
```

$0.0171 < 0.05$, indicating that the correlation of age and strength is significant.

ii. Spearman's correlation The p-value for Spearman's correlation is $0.00005 < 0.05$, indicating that the correlation of age and strength is significant.

```
perm_Str = permute_x(R_Str)
perm_Age = matrix(R_Age,
                  nrow = nrow(perm_Str),
                  ncol = ncol(perm_Str),
                  byrow = TRUE)
perm_dist = rowSums((perm_Age - rowMeans(perm_Age)) *
  (perm_Str - rowMeans(perm_Str))) /
  sqrt(rowSums((perm_Age - rowMeans(perm_Age)) ^ 2) *
         rowSums((perm_Str - rowMeans(perm_Str)) ^ 2))
p_value2 = mean(abs(perm_dist) >= abs(r_s))
p_value2
```

```
## [1] 4.960317e-05
```

iii. Kendall's tau The p-value for Kendall's tau is also $0.00005 < 0.05$, indicating that the correlation of age and strength is significant.

```
perm_Str = permute_x(Strength)
perm_dist = 2 * colMeans(calc_diff(Age) *
                           apply(perm_Str, 1, calc_diff) > 0) - 1
p_value3 = mean(abs(perm_dist) >= abs(tau))
p_value3
```

```
## [1] 4.960317e-05
```

Problem 5

Load the dataset for Problem 5.

```
data = data.frame(Nearby = c(4, 7), Not_Nearby = c(3, 2))
rownames(data) = c('Low', 'High')
data = as.matrix(data)
data
```

```
##      Nearby Not_Nearby
## Low       4          3
## High      7          2
```

It can be shown that X has a hypergeometric distribution with probability

$$P(X = x) = \frac{\binom{4+7}{x}\binom{3+2}{4+3-x}}{\binom{4+7+3+2}{4+3}} = \frac{\binom{11}{x}\binom{5}{7-x}}{\binom{16}{7}}.$$

Suppose that we want to conduct a left-tail test and the significance level is 0.05.

```
Pr_x = function(x){choose(11, x) * choose(5, 7 - x) / choose(16, 7)}
p_value = sum(apply(matrix(2:4), 1, Pr_x))
p_value
```

```
## [1] 0.3653846
```

The p-value for Fisher's exact test is $0.365 > 0.05$, indicating that the association between contamination and distance is not statistically significant.

Problem 6

Load the dataset for Problem 6.

```
data = data.frame(Missed_Second = c(5, 3),
                  Made_Second = c(14, 8))
rownames(data) = c('Missed_first', 'Made_First')
data = as.matrix(data)
data
```

```
##              Missed_Second Made_Second
## Missed_first             5          14
## Made_First               3           8
```

Thirty recreational basketball players were asked to shoot two free throws. Data on whether they made or missed their shots are shown in the table. The question of interest is whether the probability of making a shot on the first attempt is different from the probability of making a shot on the second attempt.

a. McNemar's test

Suppose that we want to conduct a two-sided test and the significance level is 0.05. The test statistic is $T_3 = \frac{X_{AB} - X_{BA}}{\sqrt{X_{AB} + X_{BA}}}$, which has an approximate standard normal distribution under H_0.

```
X_AA = data[1, 1]
X_AB = data[1, 2]
X_BA = data[2, 1]
X_BB = data[2, 2]
T_3 = (X_AB - X_BA) / sqrt(X_AB + X_BA)
p_value1 = 2 * (1 - pnorm(T_3))
p_value1
```

```
## [1] 0.007632882
```

It is equivalent to using $T_4 = \frac{(X_{AB} - X_{BA})^2}{X_{AB} + X_{BA}}$ as the test statistic.

```r
T_4 = (X_AB - X_BA) ^ 2 / (X_AB + X_BA)
p_value2 = 1 - pchisq(T_4, df = 1)
p_value2
```

```
## [1] 0.007632882
```

Therefore, the p-value for McNemar's test is $0.0076 < 0.05$, indicating that the probability of making a shot on the first attempt is statistically significantly different from the probability of making a shot on the second attempt.

b.

Suppose the pairings of the observations were ignored, and we analyzed the data using a permutation chi-square test as discussed in Section 5.4.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$
$$e_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Compute e.

```r
e = data
for (i in 1:nrow(e)){
  for (j in 1:ncol(e)){
    e[i, j] = rowSums(data)[i] * colSums(data)[j] / sum(data)}}
e
```

```
##              Missed_Second Made_Second
## Missed_first      5.066667   13.933333
## Made_First        2.933333    8.066667
```

Calculate the χ^2_{obs} test statistic for the original data.

```r
chisq_obs = sum((data - e) ^ 2 / e)
chisq_obs
```

```
## [1] 0.003262288
```

Conduct the permutation chi-square test.

```r
perm_dist = c()
for (i in 1:9){
  perm_n <- matrix(c(i - 1, 20 - i, 9 - i, 2 + i), nrow = nrow(data), byrow = TRUE)
  perm_e <- perm_n
  for (j in 1:2){
    for (k in 1:2){
      perm_e[j, k] <- rowSums(perm_n)[j] * colSums(perm_n)[k] / sum(perm_n)
    }}
  freq = choose(8, perm_n[1, 1]) * choose(22, perm_n[2, 2])
```

```
  perm_chisq = sum((perm_n - perm_e) ^ 2 / perm_e)
  perm_dist = c(perm_dist, rep(perm_chisq, freq))
}
p_value3 = mean(perm_dist >= chisq_obs)
p_value3
```

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

Here, the p-value is 1, which is totally different from the p-value in part a, indicating that the probability of making a shot on the first attempt is not statistically different from the probability of making a shot on the second attempt.