**STAT 40001/STAT 50001 Statistical Computing Fall 2024**

**Lab-10**

1. An experiment was conducted to measure the pulse rate of students in a class. The students took their own pulse rate. They were then asked to flip a coin. If the coin came up heads, they were to run in place for one minute Otherwise they sat for one minute. Then everyone took their pulse again. The pulse rates and other physiological and lifestyle data are given in the data.

<http://www.statsci.org/data/oz/ms212.txt>

|  |  |  |
| --- | --- | --- |
|  | | |
|  | Variable |  | Description |
|  |  | | |
|  | Height |  | Height (cm) |
|  | Weight |  | Weight (kg) |
|  | Age |  | Age (years) |
|  | Gender |  | Sex (1 = male, 2 = female) |
|  | Smokes |  | Regular smoker? (1 = yes, 2 = no) |
|  | Alcohol |  | Regular drinker? (1 = yes, 2 = no) |
|  | Exercise |  | Frequency of exercise (1 = high, 2 = moderate, 3 = low) |
|  | Ran |  | Whether the student ran or sat between the first and second pulse measurements (1 = ran, 2 = sat) |
|  | Pulse1 |  | First pulse measurement (rate per minute) |
|  | Pulse2 |  | Second pulse measurement (rate per minute) |
|  | Year |  | Year of class (93 - 98) |
|  |  | | |

a) Test the hypothesis whether there is a difference in pulse rate if the students were sitting.

b) Test the hypothesis whether the average pulse rate for running students is increased by 10 units after they ran.

> Q1 <- read.table("http://www.statsci.org/data/oz/ms212.txt", header = T)

> head(Q1)

Height Weight Age Gender Smokes Alcohol Exercise Ran Pulse1 Pulse2 Year

1 173 57 18 2 2 1 2 2 86 88 93

2 179 58 19 2 2 1 2 1 82 150 93

3 167 62 18 2 2 1 1 1 96 176 93

4 195 84 18 1 2 1 1 2 71 73 93

5 173 64 18 2 2 1 3 2 90 88 93

6 184 74 22 1 2 1 3 1 78 141 93

> dim(Q1)

[1] 110 11

> attach(Q1)

> table(Ran)

Ran

1 2

46 64

> S <- subset(Q1, Ran == 2)

> head(S)

Height Weight Age Gender Smokes Alcohol Exercise Ran Pulse1 Pulse2 Year

1 173 57 18 2 2 1 2 2 86 88 93

4 195 84 18 1 2 1 1 2 71 73 93

5 173 64 18 2 2 1 3 2 90 88 93

7 162 57 20 2 2 1 2 2 68 72 93

8 169 55 18 2 2 1 2 2 71 77 93

9 164 56 19 2 2 1 1 2 68 68 93

> t.test(S$Pulse1, S$Pulse2, paired = T) # since the same set of students we are using, paired = TRUE

Paired t-test

data: S$Pulse1 and S$Pulse2

t = 2.0234, df = 62, p-value = 0.04734

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

0.01209616 1.98790384

sample estimates:

mean difference

1

> cat("Since p-value < 0.05, we reject the null hypothesis. There is sufficient evidence to conclude that there is a significant difference in pulse rates before and after sitting.")

Since p-value < 0.05, we reject the null hypothesis. There is sufficient evidence to conclude that there is a significant difference in pulse rates before and after sitting.

> R <- subset(Q1, Ran == 1)

> head(R)

Height Weight Age Gender Smokes Alcohol Exercise Ran Pulse1 Pulse2 Year

2 179 58 19 2 2 1 2 1 82 150 93

3 167 62 18 2 2 1 1 1 96 176 93

6 184 74 22 1 2 1 3 1 78 141 93

10 168 60 23 1 2 1 2 1 88 150 93

11 170 75 20 1 2 1 1 1 76 88 93

18 180 70 18 1 2 1 2 1 80 146 93

> dim(R)

[1] 46 11

> t.test(R$Pulse2, R$Pulse1, mu = 10, alt = "greater", paired = T)

Paired t-test

data: R$Pulse2 and R$Pulse1

t = 13.311, df = 45, p-value < 2.2e-16

alternative hypothesis: true mean difference is greater than 10

95 percent confidence interval:

46.16911 Inf

sample estimates:

mean difference

51.3913

> cat("Since p-value < 0.05, we reject the null hypothesis. There is sufficient evidence to conclude that the average pulse rate increase after running is greater than 10 bpm.")

Since p-value < 0.05, we reject the null hypothesis. There is sufficient evidence to conclude that the average pulse rate increase after running is greater than 10 bpm.

1. A bottled water company acquires its water from two independent sources, X and Y. Data set *Water* in *PASWR* contains the sodium content in each brand of water. Is there statistical evidence to suggest that the variance of sodium content in the water from source X is different than the variance of sodium content in water from source Y?

> install.packages("PASWR")

> library(PASWR)

> data("Water")

> head(Water)

X Y Sodium Source

1 84 78 84 X

2 73 79 73 X

3 92 84 92 X

4 84 82 84 X

5 95 80 95 X

6 74 85 74 X

> attach(Water)

> str(Water)

'data.frame': 30 obs. of 4 variables:

$ X : int 84 73 92 84 95 74 80 86 80 77 ...

$ Y : int 78 79 84 82 80 85 81 83 79 81 ...

$ Sodium: int 84 73 92 84 95 74 80 86 80 77 ...

$ Source: Factor w/ 2 levels "X","Y": 1 1 1 1 1 1 1 1 1 1 ...

> boxplot(Sodium~Source)

> var.test(Sodium~Source)

F test to compare two variances

data: Sodium by Source

F = 23.215, num df = 19, denom df = 9, p-value = 3.921e-05

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

6.302573 66.858988

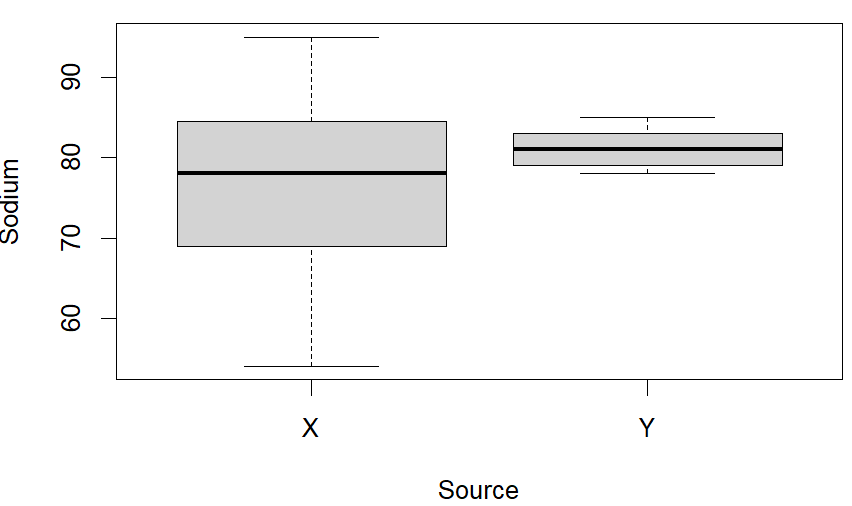
sample estimates:

ratio of variances

23.21451

> cat("Since the p-value is less than 0.05, we reject the null hypothesis. This means that there is strong evidence to suggest that the variances of sodium content from Source X and Source Y are different.")

Since the p-value is less than 0.05, we reject the null hypothesis. This means that there is strong evidence to suggest that the variances of sodium content from Source X and Source Y are different.



1. The Pew Research Group conducted a poll in which they asked, “Are you in favor of, or opposed to, executing persons as a general policy when the crime was committed while under the age of 18?” Of the 580 Catholics surveyed, 180 indicated they favored capital punishment; of the 600 seculars (those who do not associate with a religion) surveyed, 238 favored capital punishment. Is there a significant difference in the proportion of individuals in these groups in favor of capital punishment for persons under the age of 18?

> prop.test(x = c(180,238), n = c(580,600))

2-sample test for equality of proportions with continuity correction

data: c(180, 238) out of c(580, 600)

X-squared = 9.233, df = 1, p-value = 0.002377

alternative hypothesis: two.sided

95 percent confidence interval:

-0.14232944 -0.03031424

sample estimates:

prop 1 prop 2

0.3103448 0.3966667

> cat("Since the p-value is less than 0.05, we reject the null hypothesis. This means that there is significant statistical evidence that the proportions of Catholics and secular individuals in favor of capital punishment for persons.")

Since the p-value is less than 0.05, we reject the null hypothesis. This means that there is significant statistical evidence that the proportions of Catholics and secular individuals in favor of capital punishment for persons.

1. Furness and Bryant (1996) compared the metabolic rates of male and female breeding northern fulmars (data described in Logan (2010) and Quinn (2002)).

|  |  |
| --- | --- |
| Sex | Metabolic rate |
| Female | 728 |
| Female | 1087 |
| Female | 1091 |
| Female | 1361 |
| Female | 1491 |
| Female | 1956 |
| Male | 526 |
| Male | 606 |
| Male | 843 |
| Male | 1196 |
| Male | 1946 |
| Male | 2136 |
| Male | 2309 |
| Male | 2950 |

1. Display the metabolic rate of Female and Male group using side-by-side boxplot
2. Test the hypothesis whether there is a difference in Metabolic rate based on gender.

> Q4 <- read.csv("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 10\\q4.csv", header = T)

> head(Q4)

Sex Metabolic.rate

1 Female 728

2 Female 1087

3 Female 1091

4 Female 1361

5 Female 1491

6 Female 1956

> dim(Q4)

[1] 14 2

> boxplot(Q4$Metabolic.rate~Q4$Sex)

> t.test(Q4$Metabolic.rate~Q4$Sex)

Welch Two Sample t-test

data: Q4$Metabolic.rate by Q4$Sex

t = -0.77341, df = 10.466, p-value = 0.4564

alternative hypothesis: true difference in means between group Female and group Male is not equal to 0

95 percent confidence interval:

-1075.375 518.708

sample estimates:

mean in group Female mean in group Male

1285.667 1564.000

> cat("Fail to reject the null hypothesis: There is no significant difference in the metabolic rates between genders (p-value = 0.4564).")

Fail to reject the null hypothesis: There is no significant difference in the metabolic rates between genders (p-value = 0.4564).

> # also

> # if variance\_equal, more appropriate is as below

> var.test(Q4$Metabolic.rate~Q4$Sex)

F test to compare two variances

data: Q4$Metabolic.rate by Q4$Sex

F = 0.2214, num df = 5, denom df = 7, p-value = 0.1161

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.04189038 1.51727484

sample estimates:

ratio of variances

0.2214006

> cat("Fail to reject the null hypothesis: There is no significant difference in variances between the groups (p-value = 0.1161).")

Fail to reject the null hypothesis: There is no significant difference in variances between the groups (p-value = 0.1161).

> # if var.test p value > 0.05, then

> t.test(Q4$Metabolic.rate~Q4$Sex, var.equal = T)

Two Sample t-test

data: Q4$Metabolic.rate by Q4$Sex

t = -0.70106, df = 12, p-value = 0.4966

alternative hypothesis: true difference in means between group Female and group Male is not equal to 0

95 percent confidence interval:

-1143.3658 586.6992

sample estimates:

mean in group Female mean in group Male

1285.667 1564.000

> cat("Fail to reject the null hypothesis: There is no significant difference in the mean metabolic rates between genders (p-value = 0.4966).")

Fail to reject the null hypothesis: There is no significant difference in the mean metabolic rates between genders (p-value = 0.4966).



1. The website below provides the dataset related to the study of the maternal smoking and infant health.

<http://www.stat.berkeley.edu/~statlabs/data/babiesI.data>

There are two variables

bwt=Birth weight in ounces

smoke=Smoking status of mother 0=not now, 1=yes now, 9=unknown

1. Import the data set in R
2. How many observations have smoking status unknown?
3. CLEAN data set by removing subjects whose smoking status is unknown.
4. Do we have evidence to prove that the newborn baby will have significantly low weight for a smoker mom than for a non-smoker mom?

> Q5 <- read.table("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 10\\babiesI.data", header = T)

> attach(Q5)

> dim(Q5)

[1] 1236 2

> head(Q5)

bwt smoke

1 120 0

2 113 0

3 128 1

4 123 0

5 108 1

6 136 0

> sum(Q5$smoke == "9")

[1] 10

> head(Q5,20)

bwt smoke

1 120 0

2 113 0

3 128 1

4 123 0

5 108 1

6 136 0

7 138 0

8 132 0

9 120 0

10 143 1

11 140 0

12 144 1

13 141 1

14 110 1

15 114 0

16 115 0

17 92 1

18 115 1

19 144 0

20 119 1

> clean\_data <- subset(Q5,smoke!=9)

> dim(clean\_data)

[1] 1226 2

> t.test(bwt~smoke, data = clean\_data, alt = "greater")

Welch Two Sample t-test

data: bwt by smoke

t = 8.5813, df = 1003.2, p-value < 2.2e-16

alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0

95 percent confidence interval:

7.222928 Inf

sample estimates:

mean in group 0 mean in group 1

123.0472 114.1095

> cat("Reject the null hypothesis: There is significant evidence (p-value < 2.2e-16) that newborn babies of smoker moms have significantly lower birth weights than those of non-smoker moms.")

Reject the null hypothesis: There is significant evidence (p-value < 2.2e-16) that newborn babies of smoker moms have significantly lower birth weights than those of non-smoker moms.

1. Are Female Hurricanes Deadlier than Male Hurricanes? The data set contains archival data on actual fatalities caused by hurricanes in the United States between 1950 and 2012 is provided with this Lab. Please note that two deadliest hurricanes (hurricane Katrina in 2005 (1833 deaths) and Audrey in 1957 (416 deaths) were removed from the data set). Perform the appropriate test.

> # install.packages("xlsx") -> range = "A1:D47" not working

> # readxl package -> read\_excel

> install.packages(readxl)

Error in install.packages : object 'readxl' not found

> library(readxl)

> data1 <- read\_excel("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 10\\Hurricane.xlsx", range = "A1:D47")

> data2 <- read\_excel("C:\\Users\\PNW\_checkout\\Downloads\\vaishak\\PNW\_COURSE-WORK\\FALL24\\STATISTICAL COMPUTING\\Assignment\\Assignment 10\\Hurricane.xlsx", range = "E1:H47")

> head(data1)

# A tibble: 6 × 4

Hurricane Year Gender Death

*<chr>* *<dbl>* *<chr>* *<dbl>*

1 Easy 1950 Female 2

2 King 1950 Male 4

3 Able 1952 Male 3

4 Barbara 1953 Female 1

5 Florence 1953 Female 0

6 Carol 1954 Female 60

> dim(data1)

[1] 46 4

> head(data2)

# A tibble: 6 × 4

Hurricane Year Gender Death

*<chr>* *<dbl>* *<chr>* *<dbl>*

1 Elena 1985 Female 4

2 Gloria 1985 Female 8

3 Juan 1985 Male 12

4 Kate 1985 Female 5

5 Bonnie 1986 Female 3

6 Charley 1986 Male 5

> dim(data2)

[1] 46 4

> data <- rbind(data1, data2)

> head(data)

# A tibble: 6 × 4

Hurricane Year Gender Death

*<chr>* *<dbl>* *<chr>* *<dbl>*

1 Easy 1950 Female 2

2 King 1950 Male 4

3 Able 1952 Male 3

4 Barbara 1953 Female 1

5 Florence 1953 Female 0

6 Carol 1954 Female 60

> dim(data)

[1] 92 4

> attach(data)

> boxplot(Death~Gender)

> t.test(Death~Gender, alt = "greater")

Welch Two Sample t-test

data: Death by Gender

t = 1.9022, df = 86.161, p-value = 0.03024

alternative hypothesis: true difference in means between group Female and group Male is greater than 0

95 percent confidence interval:

1.622587 Inf

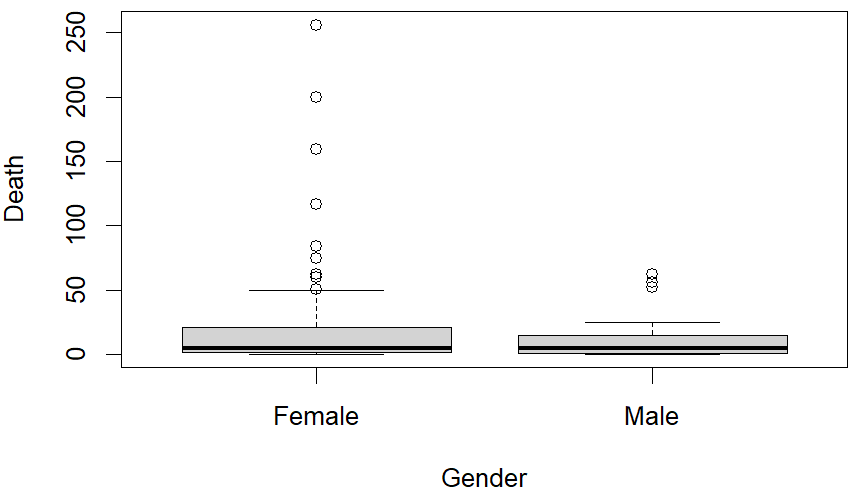
sample estimates:

mean in group Female mean in group Male

24.71429 11.82759

> cat("Reject the null hypothesis. This suggests that female-named hurricanes, on average, cause more deaths than male-named hurricanes.")

Reject the null hypothesis. This suggests that female-named hurricanes, on average, cause more deaths than male-named hurricanes.



1. The *birthwt* data in MASS package were collected at Baystate Medical Center, Springfield, Massachusetts. Perform a two-sample proportion test to determine whether smoking gives a higher fraction of low-weight births.

Variables provided in the data

Low: indicator of birth weight less than 2.5 kg.

Age:mother's age in years.

Lwt: mother's weight in pounds at last menstrual period.

Race:mother's race (1 = white, 2 = black, 3 = other).

Smoke:smoking status during pregnancy.

Pt1:number of previous premature labours.

Ht:history of hypertension.

Ui:presence of uterine irritability.

Ftv:number of physician visits during the first trimester.

Bwt:birth weight in grams.

> install.packages("MASS")

> library(MASS)

> data("birthwt")

> tab <- table(birthwt$smoke, birthwt$low)

>

> prop.test(tab)

2-sample test for equality of proportions with continuity correction

data: tab

X-squared = 4.2359, df = 1, p-value = 0.03958

alternative hypothesis: two.sided

95 percent confidence interval:

0.004967192 0.301495793

sample estimates:

prop 1 prop 2

0.7478261 0.5945946

> cat("p value is 0.03958 < 0,05, There is a significant higher fraction of low-weight births among smokers.")

p value is 0.03958 < 0,05, There is a significant higher fraction of low-weight births among smokers.