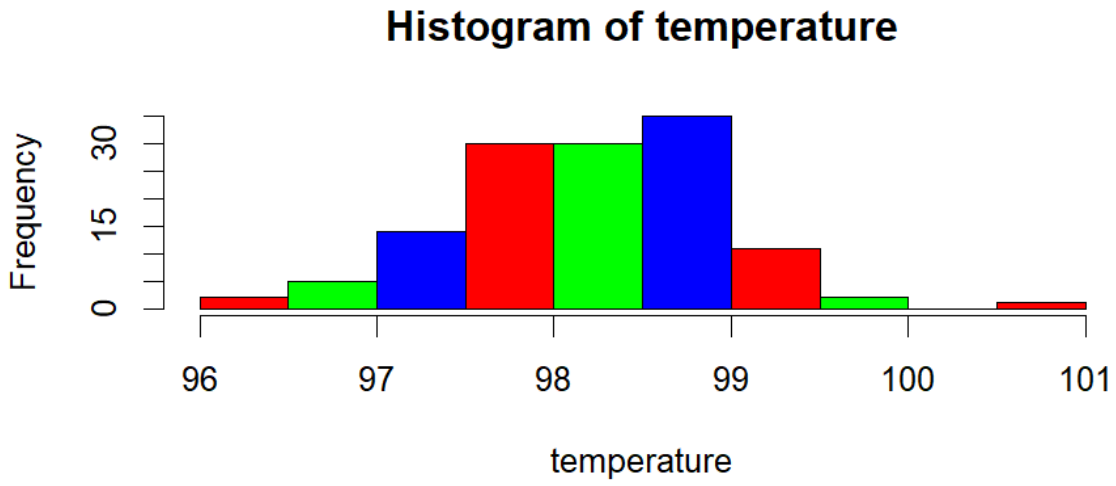
STAT 40001/STAT 50001 Statistical Computing Fall 2024

Lab-9

[Consider **p value 0.05**, as a reference line, to make decision]

1. The data set normtemp in UsingR package contains measurements of 130 healthy randomly selected individuals. The variable temperature contains normal body temperature. Does the data support that the average body temperature is 98.6^0 F. Perform the hypothesis test.

```
> install.packages("UsingR")
> library(UsingR)
> data("normtemp")
> head(normtemp, 6)
  temperature gender hr
1        96.3      1 70
2        96.7      1 71
3        96.9      1 74
4        97.0      1 80
5        97.1      1 73
6        97.1      1 75
> attach(normtemp)
> t.test(temperature, mu = 98.6)

        One Sample t-test

data:  temperature
t = -5.4548, df = 129, p-value = 2.411e-07
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.12200 98.37646
sample estimates:
mean of x
 98.24923

> cat("Since p value is 2.411e-07 < typical alpha value 0.05, we reject the null hypothesis, and
conclude that the average body temperature is not 98.6 C")
Since p value is 2.411e-07 < typical alpha value 0.05, we reject the null hypothesis, and conclud
e that the average body temperature is not 98.6 C
> hist(temperature, col = rainbow(3))
```



2. ***faraway*** package in R contains a data set ***prostate*** which describes 97 men with prostate cancer who were due to receive a radical prostatectomy. Test whether the participants are younger than 65 years.

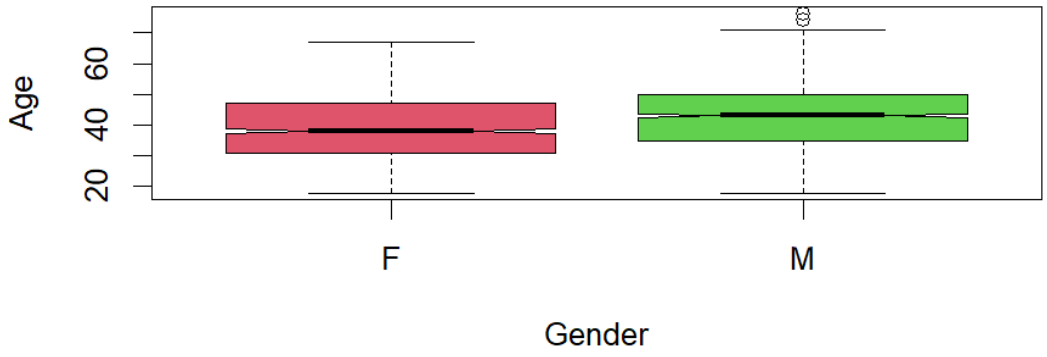
```
> install.packages("faraway")
> library(faraway)
> data("prostate")
> head(prostate)
      lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
> dim(prostate)
[1] 97  9
> attach(prostate)
> t.test(age, mu = 65, alt = "less")

        One Sample t-test

data:  age
t = -1.5002, df = 96, p-value = 0.06843
alternative hypothesis: true mean is less than 65
95 percent confidence interval:
    -Inf 65.1215
sample estimates:
mean of x
 63.86598

> cat("Since the p-value is 0.06843, which is greater than the significance level of 0.05, we fai
l to reject the null hypothesis. This suggests that the average age is not significantly less tha
n 65.")
Since the p-value is 0.06843, which is greater than the significance level of 0.05, we fail to re
ject the null hypothesis. This suggests that the average age is not significantly less than 65.
```

3. Napa Valley Marathon Times by Age and Gender for 2015 are provided with this assignment.

   a. Import the data in R
   b. How many runners are older than 50 years of age?
   c. Display the age distributions of the runner by gender.
   d. Are men older than women?
   e. The average completion time for all runners is 4.361 hours. Test whether the completion time for men is lower than 4.361 hours.
   f. The average age for all runners is 41.33 years. Test whether women are younger than 41.33 years.

```
> Q3 = read.csv("C:\\Users\\PNW_checkout\\Downloads\\vaishak\\PNW_COURSE-WORK\\FALL24\\STATISTICA
L COMPUTING\\Assignment\\Assignment 9\\NAPA.csv")
> head(Q3)
  Gender Age  Hours    mph finish_all
1      F  32 2.7625 9.4910          8
2      F  27 2.8619 9.1612         15
3      F  30 2.9031 9.0314         19
4      F  32 2.9264 8.9594         26
5      F  28 3.0181 8.6873         45
6      F  44 3.0289 8.6562         46
> dim(Q3)
[1] 1882    5
> length(Q3)
[1] 5
> names(Q3)
[1] "Gender"     "Age"        "Hours"      "mph"        "finish_all"
> attach(Q3)
> length(Q3$Age[Age>50])
[1] 383
> sum(Age > 50)
```

```
[1] 383
> boxplot(Age~Gender, col = c(2,3), notch = TRUE)$out
[1] 74 76 74
> t.test(Age~Gender, alt = "less")
Welch Two Sample t-test

data:  Age by Gender
t = -9.0319, df = 1878.1, p-value < 2.2e-16
alternative hypothesis: true difference in means between group F and group M is less than 0
95 percent confidence interval:
      -Inf -3.598041
sample estimates:
mean in group F mean in group M
       39.04862        43.44831

> cat("Since the p-value is less than 2.2e-16, which is much smaller than 0.05, we reject the nul
l hypothesis. This suggests that the average age of males is significantly less than females.")
Since the p-value is less than 2.2e-16, which is much smaller than 0.05, we reject the null hypot
hesis. This suggests that the average age of males is significantly less than females.
> Male = subset(Q3, Gender == "M")
> head(Male)
    Gender Age  Hours       mph finish_all
906      M  28 2.3850 10.993187          1
907      M  32 2.3950 10.947286          2
908      M  30 2.5217 10.397389          3
909      M  38 2.6794  9.785144          4
910      M  45 2.6906  9.744735          5
911      M  24 2.7367  9.580542          6
> t.test(Male$Hours, mu = 4.361, alt = "less")

         One Sample t-test

data:  Male$Hours
t = -6.548, df = 976, p-value = 4.707e-11
alternative hypothesis: true mean is less than 4.361
95 percent confidence interval:
     -Inf 4.239651
sample estimates:
mean of x
  4.19889

> cat("Since the p-value is 4.707e-11, which is smaller than 0.05, we reject the null hypothesis.
This suggests that the average hours for males is significantly less than 4.361.")
Since the p-value is 4.707e-11, which is smaller than 0.05, we reject the null hypothesis. This s
uggests that the average hours for males is significantly less than 4.361.
> Female = subset(Q3, Gender == "F")
> head(Female)
  Gender Age  Hours    mph finish_all
1      F  32 2.7625 9.4910          8
2      F  27 2.8619 9.1612         15
3      F  30 2.9031 9.0314         19
4      F  32 2.9264 8.9594         26
5      F  28 3.0181 8.6873         45
6      F  44 3.0289 8.6562         46
> t.test(Female$Age, mu = 41.33, alt = "less")

         One Sample t-test

data:  Female$Age
t = -6.645, df = 904, p-value = 2.617e-11
alternative hypothesis: true mean is less than 41.33
95 percent confidence interval:
     -Inf 39.61391
sample estimates:
mean of x
 39.04862

> cat("Since the p-value is 2.617e-11, which is smaller than 0.05, we reject the null hypothesis.
This suggests that the average age of females is significantly less than 41.33.")
Since the p-value is 2.617e-11, which is smaller than 0.05, we reject the null hypothesis. This s
uggests that the average age of females is significantly less than 41.33.
```



4. The *birthwt* data in MASS package were collected at Baystate Medical Center, Springfield, Massachusetts. Import the data and identify the proportion of low birthweights based on the Race.

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.
`Ptl`: 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")
> head(birthwt)
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
88   0  21 108    1     1   0  0  1   2 2594
89   0  18 107    1     1   0  0  1   0 2600
91   0  21 124    3     0   0  0  0   0 2622
> names(birthwt)
 [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"
[10] "bwt"
> dim(birthwt)
[1] 189  10
> attach(birthwt)
> # xtabs(~low+race)
> table(low,race)
   race
low  1  2  3
  0 73 15 42
  1 23 11 25
> prop = c(23/96, 11/26, 25/67)
> prop
[1] 0.2395833 0.4230769 0.3731343
```

5. The waiting time (mins) of 100 bank customers before service is being rendered are provided below:

0.8, 0.8, 1.3, 1.5, 1.8, 1.9, 1.9, 2.1, 2.6, 2.7, 2.9, 3.1, 3.2, 3.3, 3.5, 3.6, 4.0, 4.1, 4.2, 4.2, 4.3, 4.3, 4.4, 4.4, 4.6, 4.7, 4.7, 4.8, 4.9, 4.9, 5, 5.3, 5.5, 5.7, 5.7, 6.1, 6.2, 6.2, 6.2, 6.3, 6.7, 6.9, 7.1, 7.1, 7.1, 7.1, 7.4, 7.6, 7.7, 8, 8.2, 8.6, 8.6, 8.6, 8.8, 8.8, 8.9, 8.9, 9.5, 9.6, 9.7, 9.8, 10.7, 10.9, 11, 11, 11.1, 11.2, 11.2, 11.5, 11.9, 12.4, 12.5, 12.9, 13, 13.1, 13.3, 13.6, 13.7, 13.9, 14.1, 15.4, 15.4, 17.3, 17.3, 18.1, 18.2, 18.4, 18.9, 19, 19.9, 20.6, 21.3, 21.4, 21.9, 23.0, 27, 31.6, 33.1, 38.5

a) Construct a 95% confidence interval for waiting time for the bank customers.

b) Construct a 99% confidence interval for waiting time for the bank customers.

c) Do you have enough evidence to conclude that it takes on average more than 8 minutes before you are served?

```
> Q5 = c(0.8, 0.8, 1.3, 1.5, 1.8, 1.9, 1.9, 2.1, 2.6, 2.7, 2.9, 3.1, 3.2, 3.3, 3.5, 3.6, 4.0, 4.
1, 4.2, 4.2, 4.3, 4.3, 4.4, 4.4, 4.6, 4.7, 4.7, 4.8, 4.9, 4.9, 5, 5.3, 5.5, 5.7, 5.7, 6.1, 6.2,
6.2, 6.2, 6.3, 6.7, 6.9, 7.1, 7.1, 7.1, 7.1, 7.4, 7.6, 7.7, 8, 8.2, 8.6, 8.6, 8.6, 8.8, 8.8, 8.9
, 8.9, 9.5, 9.6, 9.7, 9.8, 10.7, 10.9, 11, 11, 11.1, 11.2, 11.2, 11.5, 11.9, 12.4, 12
.5, 12.9, 13, 13.1, 13.3, 13.6, 13.7, 13.9, 14.1, 15.4, 15.4, 17.3, 17.3, 18.1, 18.2, 18.4, 18.9
, 19, 19.9, 20.6, 21.3, 21.4, 21.9, 23.0, 27, 31.6, 33.1, 38.5)
> t.test(Q5)$conf.int
[1]  8.441023 11.312977
attr(,"conf.level")
[1] 0.95
> t.test(Q5, conf.level = 0.99)$conf.int
[1]  7.976271 11.777729
attr(,"conf.level")
[1] 0.99
> t.test(Q5, mu = 8, alt = "greater")

         One Sample t-test

data:  Q5
t = 2.5936, df = 99, p-value = 0.005468
alternative hypothesis: true mean is greater than 8
95 percent confidence interval:
 8.675376      Inf
sample estimates:
mean of x
    9.877

> cat("Since the p-value is 0.005468, which is less than 0.05, we reject the null hypothesis and
conclude that the true mean is significantly greater than 8.")
Since the p-value is 0.005468, which is less than 0.05, we reject the null hypothesis and conclud
e that the true mean is significantly greater than 8.
```

6. In order to investigate the possible relationship between marijuana smoking and a deficit in performance on a task measuring short term memory--the digit span task from the Wechsler Adult Intelligence Scale two groups of ten subjects were tested. One group, the "nonsmokers," claimed not to smoke marijuana. A second group, the "smokers," claimed to smoke marijuana regularly. Below are the scores.

nonsmokers : 18,22,21,17,20,17,23,20,22,21

smokers : 16,20,14,21,20,18,13,15,17,21

Do we have enough evidence that the nonsmoker has higher score than smokers?

```
> NS = c(18,22,21,17,20,17,23,20,22,21)
> S = c(16,20,14,21,20,18,13,15,17,21)
> t.test(NS, S, alt = "greater")

        Welch Two Sample t-test

data:  NS and S
t = 2.2573, df = 16.376, p-value = 0.01899
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5919017       Inf
sample estimates:
mean of x mean of y
     20.1      17.5

> cat("Since the p-value is 0.01899, which is less than 0.05, we reject the null hypothesis and c
onclude that the mean of NS is significantly greater than the mean of S.")
Since the p-value is 0.01899, which is less than 0.05, we reject the null hypothesis and conclude
that the mean of NS is significantly greater than the mean of S.
```