Practical Exercise 4 | Statistics for Premasters DSS/CSAI

Hilal Caliskan Egilli_u927185

Part A - Chi-Squared Tests

For this part of the practical exercise, we are going to load the data using R. You can do so by installing and loading the package "DAAG", and then using the function data(ais). Your data will be stored under the name "ais".

We will be using statistical tests from the "lsr" package, so let's also load lsr here using library(lsr).

Task 1. Load and inspect the dataset.

```
#install.packages("DAAG")
library(DAAG)
```

```
## Warning: package 'DAAG' was built under R version 4.3.3
```

```
library(lsr)

#Read in the data
data(ais)

#We inspect the data:
head(ais)
```

```
##    rcc wcc   hc   hg ferr   bmi   ssf pcBfat   lbm    ht   wt sex  sport
## 1 3.96 7.5 37.5 12.3   60 20.56 109.1  19.75 63.32 195.9 78.9   f B_Ball
## 2 4.41 8.3 38.2 12.7   68 20.67 102.8  21.30 58.55 189.7 74.4   f B_Ball
## 3 4.14 5.0 36.4 11.6   21 21.86 104.6  19.88 55.36 177.8 69.1   f B_Ball
## 4 4.11 5.3 37.3 12.6   69 21.88 126.4  23.66 57.18 185.0 74.9   f B_Ball
## 5 4.45 6.8 41.5 14.0   29 18.96  80.3  17.64 53.20 184.6 64.6   f B_Ball
## 6 4.10 4.4 37.4 12.5   42 21.04  75.2  15.58 53.77 174.0 63.7   f B_Ball
```

Task 2. Compute a general test for Goodness of Fit for the sport variable from the "ais" data set.

```
#General tests for Goodness of Fit
goodnessOfFitTest(ais$sport)
```

```
##
##      Chi-square test against specified probabilities
##
## Data variable:   ais$sport
##
## Hypotheses:
```

```
##    null:        true probabilities are as specified
##    alternative: true probabilities differ from those specified
##
## Descriptives:
##         observed freq. expected freq. specified prob.
## B_Ball              25           20.2             0.1
## Field               19           20.2             0.1
## Gym                  4           20.2             0.1
## Netball             23           20.2             0.1
## Row                 37           20.2             0.1
## Swim                22           20.2             0.1
## T_400m              29           20.2             0.1
## T_Sprnt             15           20.2             0.1
## Tennis              11           20.2             0.1
## W_Polo              17           20.2             0.1
##
## Test results:
##    X-squared statistic:  38.594
##    degrees of freedom:  9
##    p-value:  <.001
```

```
#Quite significant results, when expecting equal probabilities
```

Task 3. Print a table of the sports in the dataset.

```
table(ais$sport)
```

```
##
##  B_Ball   Field     Gym Netball     Row    Swim  T_400m T_Sprnt  Tennis  W_Polo
##      25      19       4      23      37      22      29      15      11      17
```

Task 4. We are interested in a subset of the data which includes all sports except for "Gym", "W_Polo" and "Netball". Do the following to prepare the data frame:

- Make a subset of your data frame that excludes the three factor levels of "sport" we do not need, you can call the new data frame "excludeSport".
- Use droplevels() on the sport variable in "excludeSport" to remove the now empty levels.
- Inspect the "excludeSport" data frame to see if everything went as expected.

HINT: *You can use the subset() function with a != operator to exclude specific values of a variable. You can exclude all of the values at once by adding & between each conidition.*

```
# Get factors of sports out which we do not need
excludeSport <- subset(ais, sport != "Gym" & sport != "W_Polo" & sport != "Netball")

# The %in% operator can also be used to create this subset:
# excludeSport <- subset(ais, !sport %in% c("Gym", "W_Polo", "Netball"))

# We drop the levels (otherwise this would impact our df, and test statistic)
excludeSport$sport <- droplevels(excludeSport$sport)

# We re-check the new dataset where factors are dropped
table(excludeSport$sport)
```

```
##
##   B_Ball    Field      Row     Swim   T_400m  T_Sprnt   Tennis
##       25       19       37       22       29       15       11
```

Task 5. Now, create two subsets in order to do individual tests for males and females.

HINT: *You would have to subset the "excludeSport" data frame again. After that, check the female and male data frames to see if they are smaller now.*

```
#Make two subset to do invidual tests
Females <- excludeSport[excludeSport$sex == "f", ]
Males <- excludeSport[excludeSport$sex == "m", ]

#Check if our datasets are smaller now
nrow(Males)
```

```
## [1] 85
```

```
nrow(Females)
```

```
## [1] 73
```

Task 6a. Write down the null and alternative hypotheses for the male and female data sets.

- H0: No difference between distributions of males and females
- H1: Difference between distributions of males and females

Task 6b. Calculate the critical value, which would aid you in determining whether you can reject the null hypothesis or not (you could use the qchisq() function for this).

```
qchisq(0.95, df = 6)
```

```
## [1] 12.59159
```

Task 7. Do individual tests on the male and female subsets to determine whether each sport has an equal probability of being played.

HINT: *The goodnessOfFitTest() tests for equal probabilities by default.*

```
goodnessOfFitTest(Males$sport)
```

```
##
##      Chi-square test against specified probabilities
##
## Data variable:   Males$sport
##
## Hypotheses:
##    null:        true probabilities are as specified
##    alternative: true probabilities differ from those specified
##
## Descriptives:
```

```
##         observed freq. expected freq. specified prob.
## B_Ball              12       12.14286       0.1428571
## Field               12       12.14286       0.1428571
## Row                 15       12.14286       0.1428571
## Swim                13       12.14286       0.1428571
## T_400m              18       12.14286       0.1428571
## T_Sprnt             11       12.14286       0.1428571
## Tennis               4       12.14286       0.1428571
##
## Test results:
##    X-squared statistic:  9.129
##    degrees of freedom:  6
##    p-value:  0.166
```

```r
goodnessOfFitTest(Females$sport)
```

```
##
##      Chi-square test against specified probabilities
##
## Data variable:   Females$sport
##
## Hypotheses:
##    null:        true probabilities are as specified
##    alternative: true probabilities differ from those specified
##
## Descriptives:
##         observed freq. expected freq. specified prob.
## B_Ball              13       10.42857       0.1428571
## Field                7       10.42857       0.1428571
## Row                 22       10.42857       0.1428571
## Swim                 9       10.42857       0.1428571
## T_400m              11       10.42857       0.1428571
## T_Sprnt              4       10.42857       0.1428571
## Tennis               7       10.42857       0.1428571
##
## Test results:
##    X-squared statistic:  19.918
##    degrees of freedom:  6
##    p-value:  0.003
```

Task 8. Check if the female distribution is similar to the male distribution using the chi-squared test of association.

```r
associationTest(~sport + sex, excludeSport)
```

```
##
##      Chi-square test of categorical association
##
## Variables:   sport, sex
##
## Hypotheses:
##    null:        variables are independent of one another
```

```
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##          sex
## sport      f  m
##   B_Ball  13 12
##   Field    7 12
##   Row     22 15
##   Swim     9 13
##   T_400m  11 18
##   T_Sprnt  4 11
##   Tennis   7  4
##
## Expected contingency table under the null hypothesis:
##          sex
## sport         f     m
##   B_Ball  11.55 13.45
##   Field    8.78 10.22
##   Row     17.09 19.91
##   Swim    10.16 11.84
##   T_400m  13.40 15.60
##   T_Sprnt  6.93  8.07
##   Tennis   5.08  5.92
##
## Test results:
##    X-squared statistic:  8.318
##    degrees of freedom:  6
##    p-value:  0.216
##
## Other information:
##    estimated effect size (Cramer's v):  0.229
```

Task 9a. Load and inspect the "Salaries" data set from "carData" package. Assign the data set "Salaries" to the "salaries" variable.

```r
library(lsr)

#Let's load in the data
salaries <- carData::Salaries

#Look at the variables in the data set
colnames(salaries)
```

```
## [1] "rank"          "discipline"    "yrs.since.phd" "yrs.service"
## [5] "sex"           "salary"
```

Task 9b. Test the following hypotheses:

- H0: There is no difference in rank between males and females.
- H1: There is a difference in rank between males and females

```r
table(salaries$rank, salaries$sex)
```

```
##
##             Female Male
##   AsstProf      11   56
##   AssocProf     10   54
##   Prof          18  248
```

```r
#We put both variables after the ~, as we are not dealing with dependent variables.
associationTest(~ rank + sex, data = salaries)
```

```
##
##      Chi-square test of categorical association
##
## Variables:   rank, sex
##
## Hypotheses:
##    null:        variables are independent of one another
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##            sex
## rank        Female Male
##   AsstProf      11   56
##   AssocProf     10   54
##   Prof          18  248
##
## Expected contingency table under the null hypothesis:
##            sex
## rank        Female  Male
##   AsstProf    6.58  60.4
##   AssocProf   6.29  57.7
##   Prof       26.13 239.9
##
## Test results:
##    X-squared statistic:  8.526
##    degrees of freedom:  2
##    p-value:  0.014
##
## Other information:
##    estimated effect size (Cramer's v):  0.147
```

Task 10. Explain whether we reject, or fail to reject H0. Report the relevant statistics in APA format.

An Association Test was conducted, based on which the null was rejected. $\chi^2(2) = 8.526$, $p = 0.01$. The effect size of 0.147 is considered small.

Part B: Analyzing Results

For this part of the assignment, consider the results from:

a) Task 7
b) Task 8
c) Task 10

Task 11. Analyze the results you obtained based on the factors listed under "Don't List" from the "Where do we go from here?" lecture slides in Module 6. Be sure to include elements from the article "Moving to a world beyond p<0.05" in your analysis (such as compatibility intervals).

For this part of the task, multiple elements could have been analyzed.

A possible interpretation could have been centered around the p-values obtained from the tasks. For the p-values that were close to the threshold of 0.05, it could have been pointed out that it is in general not a good practice to solely base conclusions on whether an association/effect was found to be "statistically significant" based on an arbitrary threshold (p<.05).

Task 12. Analyze the results based on the ATOMIC (Wasserstein et al., 2019) factors.

For the ATOMIC factors, a good response would have taken one or more of them and briefly interpret the results according to the principles stated. For instance, if the "Thoughtful" ATOMIC factor would have been addressed, a few answers could have been added to questions such as "What are the practical implications of the estimate?", "How precise is the estimate?", "Is the model correctly specified?". A brief response to these questions based on the limited available information of the statistical test results would have sufficed.

Part C: Various Tests

For this part we will be using a dataset which represents a sample of 397 University Professors in the U.S. (https://www.rdocumentation.org/packages/carData/versions/3.0-4/topics/Salaries).

In order to load in the data, you have to run the following line of code: "salaries <- carData::Salaries"

Make sure you have the carData package installed!

Task 13. Load and inspect the data.

```
salaries <- carData::Salaries
summary(salaries)
```

```
##         rank     discipline yrs.since.phd    yrs.service       sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266              Median :21.00   Median :16.00
##                             Mean   :22.31   Mean   :17.61
##                             3rd Qu.:32.00   3rd Qu.:27.00
##                             Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

```
head(salaries)
```

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```r
tail(salaries)
```

```
##         rank discipline yrs.since.phd yrs.service  sex salary
## 392     Prof          A            30          19 Male 151292
## 393     Prof          A            33          30 Male 103106
## 394     Prof          A            31          19 Male 150564
## 395     Prof          A            42          25 Male 101738
## 396     Prof          A            25          15 Male  95329
## 397 AsstProf          A             8           4 Male  81035
```

Task 14a. Check the yrs.since.phd variable. What is the difference in salary of the professor (you can include assistant and associate professors) with the highest yrs.since.phd and lowest yrs.since.phd?

```r
youngest_prof <- salaries[salaries$yrs.since.phd == min(salaries$yrs.since.phd), ]

oldest_prof <- salaries[salaries$yrs.since.phd == max(salaries$yrs.since.phd), ]

#There seem to be more professors with a similar z-score (due to rounding of years)
nrow(youngest_prof)
```

```
## [1] 4
```

```r
nrow(oldest_prof)
```

```
## [1] 2
```

```r
difference_salary <- abs(mean(oldest_prof$salary) - mean(youngest_prof$salary))

print(difference_salary)
```

```
## [1] 50584.25
```

```r
#We see that there is a wage gap of $50584.25 between the oldest and youngest professors
```

Task 14b. What is the range between the highest- and lowest salary in the dataset?

```r
diff_salary <- max(salaries$salary) - min(salaries$salary)
print(diff_salary)
```

```
## [1] 173745
```

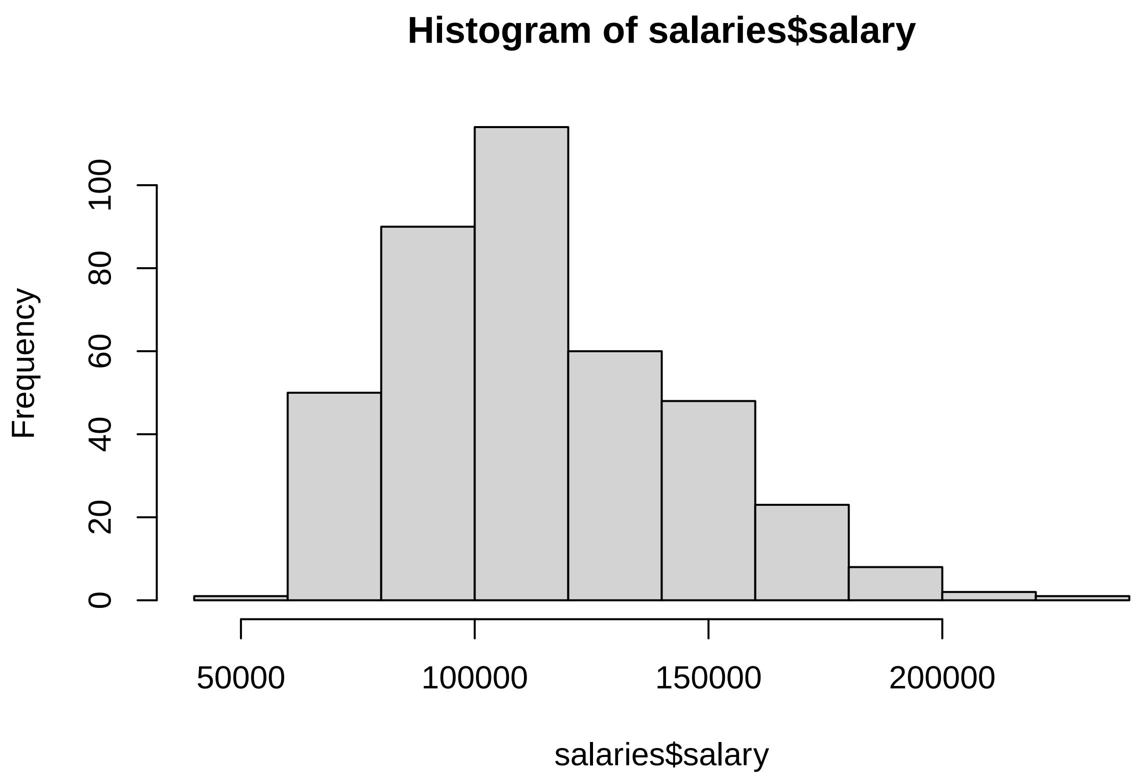
Task 15a. Check the assumption of normality for the salary variable, first by using visual inspection with a histogram and a Q-Q plot, and then by using the Shapiro-Wilk significance test.

HINT: *The functions you need besides the histogram are car::qqPlot() and shapiro.test().*

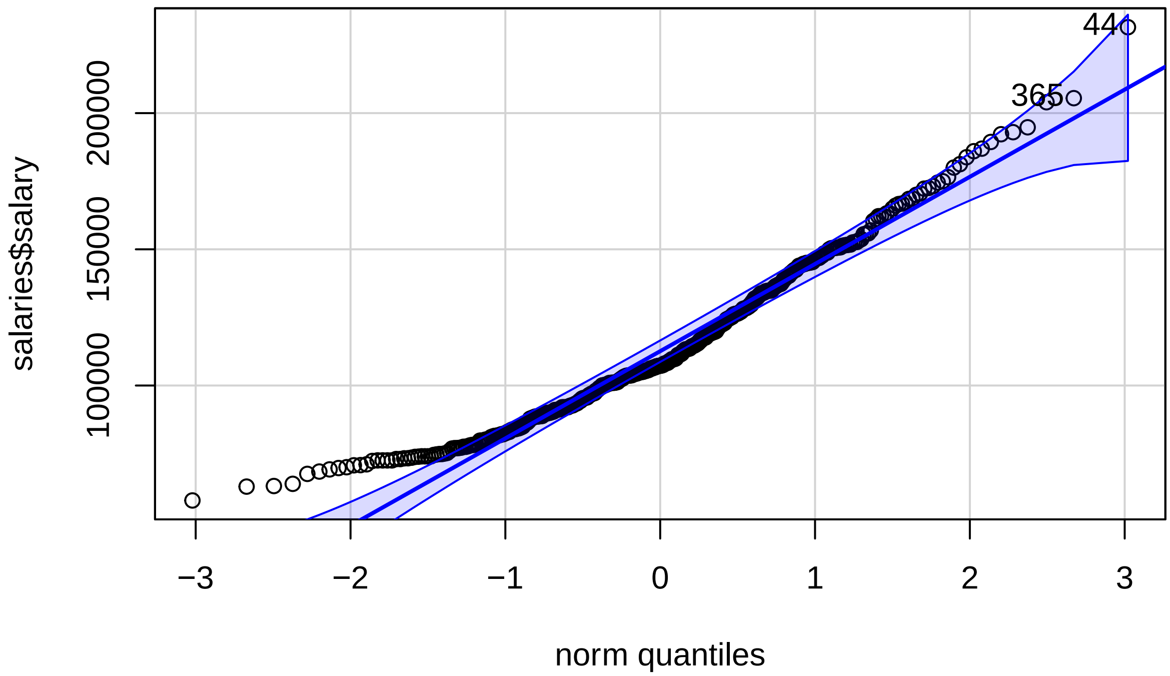
The null hypothesis for shapiro.test() is that the data is normally distributed.

A Q-Q plot for normally distributed data would have most of the quantiles lined up along the guiding diagonal line in the center.

```r
#Histogram looks quite normal
hist(salaries$salary)
```




```r
#QQ plot looks quite normal as well
car::qqPlot(salaries$salary)
```




```
## [1]  44 365
```

```
#Shapiro-Wilk test indicates non-normality
shapiro.test(salaries$salary)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salaries$salary
## W = 0.95988, p-value = 6.076e-09
```

Task 15b. Would you say that the assumption of normality is met?

The histogram and Q-Q plot both appear relatively normal, however the Shapiro-Wilk test indicates non-normality. The Shapiro-Wilk test can sometimes be overly sensitive, so if the visual tests indicate normality, we can still proceed with a parametric test. If in doubt, we can use a non-parametric test as well and compare the results.

Task 16. (Performing t-test) - Use a t-test to test the following hypotheses:

- H0: The salary between male professors and female professors is equal.
- H1: The salary between male professors and female professors is not equal

```
#Welch's
independentSamplesTTest(salary ~ sex, data = salaries)
```

```
##
##    Welch's independent samples t-test
##
## Outcome variable:   salary
## Grouping variable:  sex
```

```
##
## Descriptive statistics:
##                Female       Male
##    mean     101002.410 115090.419
##    std dev.  25952.127  30436.927
##
## Hypotheses:
##    null:        population means equal for both groups
##    alternative: different population means in each group
##
## Test results:
##    t-statistic:  -3.161
##    degrees of freedom:  50.122
##    p-value:  0.003
##
## Other information:
##    two-sided 95% confidence interval:  [-23037.916, -5138.102]
##    estimated effect size (Cohen's d):  0.498
```

```r
#Student's
independentSamplesTTest(salary ~ sex, data = salaries, var.equal = TRUE)
```

```
##
##    Student's independent samples t-test
##
## Outcome variable:   salary
## Grouping variable:  sex
##
## Descriptive statistics:
##                Female       Male
##    mean     101002.410 115090.419
##    std dev.  25952.127  30436.927
##
## Hypotheses:
##    null:        population means equal for both groups
##    alternative: different population means in each group
##
## Test results:
##    t-statistic:  -2.782
##    degrees of freedom:  395
##    p-value:  0.006
##
## Other information:
##    two-sided 95% confidence interval:  [-24044.91, -4131.107]
##    estimated effect size (Cohen's d):  0.469
```

Task 16a. Which t-test did you use, and why?

Preferably independentSamplestTTest (Welch's). We prefer Welch's over Student's as Welch's does not assume homogeneity of variance.

Task 16b. Explain whether you have to accept or reject H0 based on your sample. Report the relevant statistics in APA format (max 100 words)

An Independent Samples T-test (Welch's) was conducted in which the null has been evaluated. T-test results show that there is a difference between the salary of males (M = 115,090.42, SD = 30,436.93) and the salaries of females (M = 101,002.41, SD = 25,952.13), $t(50.12) = 3.16$, $p = 0.003$, CI95 = [-23,037.92, -5,138.1]. The Cohen's D (0.5) effect size appeared to be medium sized.

Task 16c. What are the assumptions of the different t-tests we mentioned during class (max 200 words)?

One Sample T-Test

- The dependent variable must be continuous
- The dependent variable must be normally distributed
- The observations must be independent from each other

Independent Samples T-Test (Student's)

- The dependent variable must be continuous
- The dependent variable must be normally distributed
- The observations must be independent from each other
- The variance is expected equal in both groups

Independent Samples T-Test (Welch's)

- The dependent variable must be continuous
- The dependent variable must be normally distributed
- The observations must be independent from each other

Paired Samples T-Test

- The differences between matched pairs must be normally distributed