STAT5002 Lab11 Question Sheet

Introduction to Statistics

STAT5002

1 Does Psychiatric Diagnosis Depend on Social Class?

The table below (taken from *The Analysis of Contingency Tables, B. Everitt*, Table 3.13, p56) shows how 284 consecutively admitted patients to a psychiatric hospital were classified with respect to social class (the social class is categorised as 1, 2, 3) and diagnosis:

Diagnosis:	Neurotic	Depressed	Personality disorder	Schizophrenic
1	45	25	21	18
2	10	45	24	22
3	17	21	18	18

We want to use R to perform a chi-square test of independence on this. We use the following code to create a matrix `Oi` contains the data in the above table.

- Create 4 vectors **`neur`**, **`depr`**, **`pdis`** and **`schz`** containing the values of each column.

```
neur = c(45, 10, 17)
depr = c(25, 45, 21)
pdis = c(21, 24, 18)
schz = c(18, 22, 18)
```

- Use `cbind()` to form these into a matrix called `Oi`, and give the rows nice names/labels

```
Oi = cbind(neur, depr, pdis, schz)
rownames(Oi) = c("SC1", "SC2", "SC3")
Oi
```

```
    neur depr pdis schz
SC1   45   25   21   18
SC2   10   45   24   22
SC3   17   21   18   18
```

1.1 State the null and alternative hypotheses

1.2 Compute a matrix consisting of expected frequencies

You need to compute row sums, column sums and hence a matrix `Ei` of corresponding expected frequencies *under the assumption that diagnosis and social class are independent.* \

Hint: You can use the functions `rowSums()` and `colSums()` for row sums and column sums, respectively.

1.3 Check assumptions, you may need to use the exptected frequencies

1.4 Compute the value of Pearson's chi-squared statistic for testing independence.

1.5 Obtain an (approximate) P-value and the critical region of rejection for 1% level of significance.

What is the correct degrees of freedom to be used here?

1.6 What is your conclusion based on the P-value

1.7 Check your work above using the R function `chisq.test()`.

2 More T-tests: Tied Ridging in Ethiopia

Background

It is believed that micro basin technology, such as tied ridging, can increase crop yield by concentrating runoff around the rootzone of the crop. However, tied ridging requires more effort on behalf of the farmer than traditional tillage.

The data in the data file `TiedRidging.csv` is from experiments conducted in Ethiopia. Maize was planted into equally sized tied ridge plots and the crop yield was measured (tons/hectare) at the end of the growing season.

See the detail of the experiment here.

```r
TiedRidging <- read.csv("data/TiedRidging.csv")
head(TiedRidging)
```

```
  Variety Yield
1       A   4.6
2       A   4.3
3       A   3.8
4       A   3.4
5       A   3.9
6       A   3.9
```

```r
dim(TiedRidging)
```

```
[1] 29  2
```

```r
str(TiedRidging)
```

```
'data.frame':   29 obs. of  2 variables:
 $ Variety: chr  "A" "A" "A" "A" ...
 $ Yield  : num  4.6 4.3 3.8 3.4 3.9 3.9 3.9 4.4 3.6 3.6 ...
```

```r
table(TiedRidging$Variety)
```

```

 A  B 
17 12 
```

2.1 1-Sample t-test

We are interested in finding out whether the yield of the tied ridge plot (Variety A) is significantly greater than that of a traditional yield which is 2.6 tons/hectare for this area. Use a hypothesis test to provide a recommendation to Ethiopian farmers.

Hint: use `yieldA = TiedRidging$Yield[TiedRidging$Variety=="A"]` *to extract the yields for variety A.*

```r
yieldA =  TiedRidging$Yield[TiedRidging$Variety=="A"]
nA = length(yieldA)
nA
```

```
[1] 17
```

2.1.1 State the null and alternative hypotheses

Introduce a parameter and express your null and alternative hypotheses in terms of this parameter.

2.1.2 Check assumptions using graphical summaries

2.1.3 What is the observed value of the test statistic? What values of test statistic argue against H_0?

2.1.4 Calculate P-value and the critical region of rejection for the 5% level of significance

2.1.5 What is your conclusion?

2.1.6 Calculate the two-sided 95% confidence interval

Considering the consistency defined by the two-sided 95% prediction interval, calculate the two-sided 95% confidence interval.

2.1.7 Check your working above using the R function `t.test()`.

You can use `t.test(x, mu=2.6, alt="greater")` to perform a one-sample t-test with $H_0 : \mu = 2.6$ and one-sided alternative $H_1 : \mu > 2.6$. Compare your calculation above with the result of `t.test()`. Why the confidence interval of `t.test()` is different from the two-sided one?

2.2 2-Sample t-test

Now test whether the yields for the 2 varieties (A and B) are different from each other. Check all your assumptions so that you can choose which t-test is appropriate. What would be your recommendation to Ethiopian farmers? **Note**: you may use `t.test()`.

Hint: use `yieldB = TiedRidging$Yield[TiedRidging$Variety=="B"]` *to extract the yields for variety B.*

```r
yieldB =  TiedRidging$Yield[TiedRidging$Variety=="B"]
```

2.2.1 State the null and alternative hypotheses

Introduce parameters and express your null and alternative hypotheses in terms of these parameters.

2.2.2 Check assumptions using graphical summaries and numerical summaries

You should also use numerical summaries to detect which t-test should be used here.

2.2.3 Calculate the test Statistic

2.2.4 Use `t.test()` to obtain P-value and draw conclusion at the 5% level of significance. What is your conclusion based on them?

2.2.5 Calculate the P-value and 95% confidence interval using simulation.

- Simulate 10,000 Welch statistics.
- Calculate the P-value based on simulated Welch statistics and the observed value of Welch statistic
- Calculate the 95% confidence interval based on the simulated Welch statistics and the observed difference of sample means.