STAT5002 Lab11 Solution 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

Solution: - H_0 : the social class is independent of the diagnosis - H_0 : the social class and the diagnosis are dependent

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. **Solution:** see code below

```
# Write your code here
rsums = rowSums(Oi)
rsums
```

```
SC1 SC2 SC3
109 101  74
```

```
csums = colSums(Oi)
csums
```

```
neur depr pdis schz
  72   91   63   58
```

```
Ei = outer(rsums, csums)/sum(Oi)
Ei
```

```
        neur     depr     pdis     schz
SC1 27.63380 34.92606 24.17958 22.26056
SC2 25.60563 32.36268 22.40493 20.62676
SC3 18.76056 23.71127 16.41549 15.11268
```

Note that without specifying any operations, `outer` uses the multiplication `*` as its default operation. `outer(rsums, csums, FUN="*")/sum(Oi)` gives the same results as above.

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

Solution: All expected counts are above 5 and the total sample size is 284. So it is appropriate to apply the chi-square test if the observations are independent (which we cannot check given the context).

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

```
# Write your code here
stat = sum(((Oi-Ei)^2)/Ei)
stat
```

```
[1] 30.79876
```

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?

Solution: Degrees of freedom are $(r-1)(c-1) = 2 \times 3 = 6$. The critical region of rejection is $(16.81, \infty)$, as we have a one-sided test, only large values of the test statistics arguing against H_0.

```
# Write your code here
pval = pchisq(stat, df=6, lower.tail=F)
pval
```

```
[1] 2.769301e-05
```

```
qchisq(0.99, df=6)
```

```
[1] 16.81189
```

1.6 What is your conclusion based on the P-value

Solution: The P-value is very small, thus providing strong evidence against the null hypothesis of independence. We reject H_0 at the 1% level of significance. This (indirectly) suggests a dependence between Diagnosis and Social Class. Similarly, the observed test statistic is in the crtical region.

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

```
# Write your code here
chisq.test(Oi)
```

```
    Pearson's Chi-squared test

data:  Oi
X-squared = 30.799, df = 6, p-value = 2.769e-05
```

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.

```
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
```

```
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.

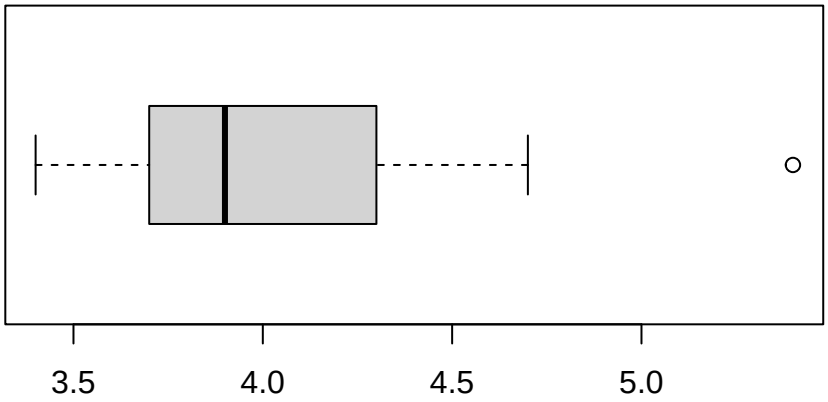
Solution: Let μ denote the mean yield under Variety A.

- H_0: $\mu = 2.6$
- H_1: $\mu > 2.6$.

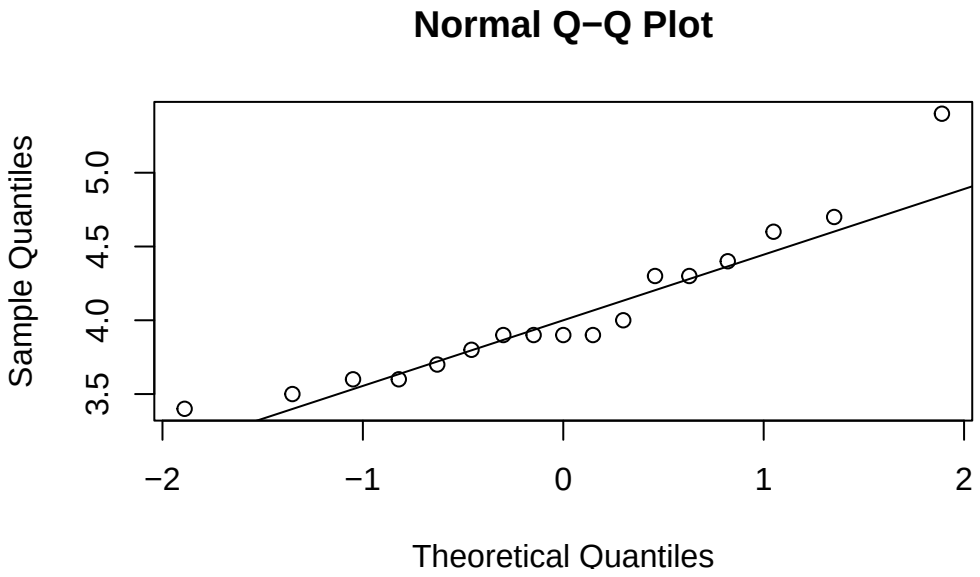
2.1.2 Check assumptions using graphical summaries

Solution: The boxplot is reasonably symmetric, with a single outlier. The QQ plot suggests that the quantiles of the data approximately follow the normal curve. However, in the upper tail (right end), the outlier has an impact on the normality as the QQ plot deviates away from the QQ line in the end of the upper tail area.

```
# Write your code here
boxplot(yieldA, horizontal=T)
```




```
qqnorm(yieldA)
qqline(yieldA)
```



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

Solution: The observed value of the test statistic is given in the following code. Here large values of test statistic argue against H_0, as suggested by the alternative hypothesis.

```
# Write your code here
mA = mean(yieldA)
mA
```

```
[1] 4.052941
```

```
est.se = sd(yieldA)/sqrt(nA)
est.se
```

```
[1] 0.1242973
```

```
stat = (mA-2.6)/est.se
stat
```

```
[1] 11.68924
```

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

Solution: The P-value is given below. The critical region is $T > 1.746$ by only considering the 5% upper tail area, as only large values of test statistic arguing against H_0.

```
# Write your code here
pt(stat, df=nA-1, lower.tail=F)
```

```
[1] 1.503073e-09
```

```
qt(0.95, df=nA-1)
```

```
[1] 1.745884
```

2.1.5 What is your conclusion?

Solution: The p-value is much smaller than the 5% level of significance, thus indicating "overwhelming" evidence against the null hypothesis that the mean under Variety A is 2.6 (the mean yield under traditional tillage). This (indirectly) suggests that Variety A gives a higher yield than traditional tillage.

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.

Solution: see the code below.

```
# Write your code here
qt(.975, df=nA-1)
```

```
[1] 2.119905
```

```
mA + c(-1,1)*qt(.975, df=nA-1)*est.se
```

```
[1] 3.789443 4.316440
```

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 there are differences in the confidence interval?

Solution: see the code below for `t.test()`.

```
# Write your code here
t.test(yieldA, mu=2.6, alt="greater")
```

```
    One Sample t-test

data:  yieldA
t = 11.689, df = 16, p-value = 1.503e-09
alternative hypothesis: true mean is greater than 2.6
```

```
95 percent confidence interval:
 3.835932      Inf
sample estimates:
mean of x
 4.052941
```

Since one-sided test use a different definition of consistency, so the confidence interval is different. The following is not for examination.

Let's verify this: we define a one-sided prediction interval $P(T < u) = 0.95$ such that 95% of T statistics falling below the multiplier u. Any test statistics exceeding u are considered as inconsistent with the underlying population mean for defining the statistic T (and thus arguing against H_0 in our one-sided test). The value of u is

```
u = qt(0.95, df=nA-1)
round(u, 3)
```

```
[1] 1.746
```

Since we have

$$P(T < u) = P\left(\frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{n}} < u\right) = 0.95$$

which gives the one-sided 95% confidence interval in this case

$$P\left(\mu > \bar{X} - u \times \frac{\hat{\sigma}}{n}\right) = 0.95$$

which is an interval from the left boundary $\bar{X} - u \times \frac{\hat{\sigma}}{n}$ to ∞,. For the current data set, we have the left boundary

```
mA - u*est.se
```

```
[1] 3.835932
```

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.*

```
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.

Solution: Denoting the yields for the 2 varieties (A and B) by μ_A and μ_B, respectively, we have

- H_0: $\mu_A = \mu_B$
- H_1:$\mu_A \neq \mu_B$

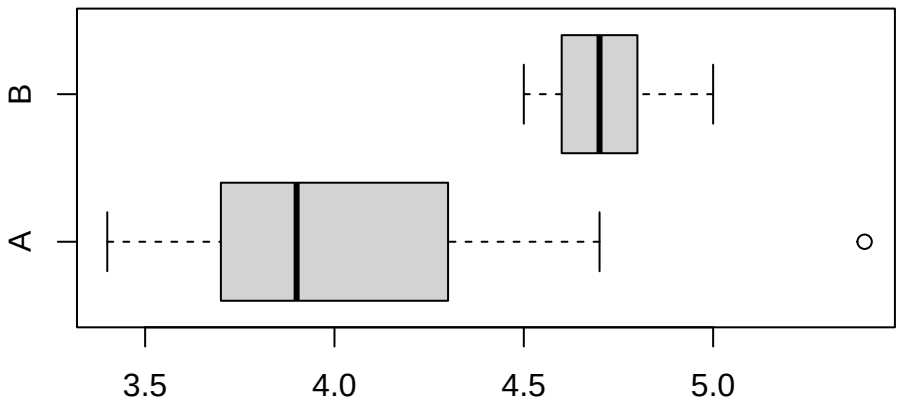
This should be a **two-sided** test: we can only reasonably use a one-sided test if there was some reason to expect a difference in a particular direction **before we saw the data**.

2.2.2 Check assumptions using graphical summaries and numerical summaries

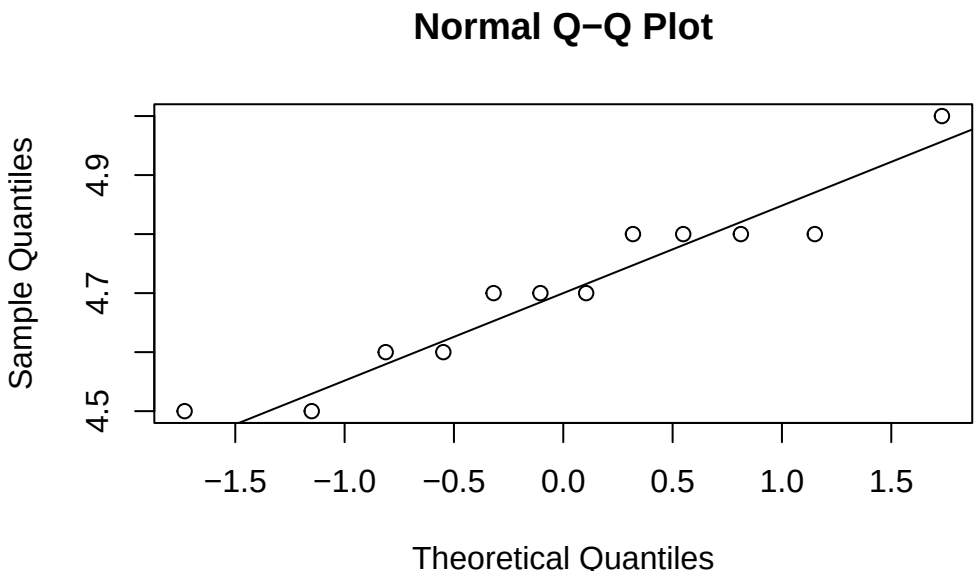
You should also use numerical summaries to detect which t-test should be used here.

Solution: The QQ plot suggests that `yieldB` is normal shaped. Its boxplot is also reasonably symmetric. However, the spreads of those two variables are very different, which is confirmed by their sample SDs. This suggests that it is not ok to assume a equal variance. Thus a Welch test is the appropriate two-sample t-test to perform.

```
# Write your code here
boxplot(yieldA, yieldB, names=c("A","B"), horizontal=T)
```



```
qqnorm(yieldB)
qqline(yieldB)
```




```
sd(yieldA)
```

```
[1] 0.512491
```

```
sd(yieldB)
```

```
[1] 0.1443376
```

2.2.3 Calculate the test Statistic

Solution: We can calculate the Welch test statistic as follows

```
# Write your code here

mB = mean(yieldB)
mB
```

```
[1] 4.708333
```

```r
nB = length(yieldB)
nB
```

```
[1] 12
```

```r
est.se.dif = sqrt((sd(yieldA)^2)/nA + (sd(yieldB)^2)/nB)
est.se.dif
```

```
[1] 0.1310951
```

```r
stat = (mA-mB)/est.se.dif
stat
```

```
[1] -4.999362
```

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?

Solution: The very small p-value provides very strong evidence against the null hypothesis that the mean yield is the same for both Varieties. This indirectly suggests there is a difference between Varieties. The confidence interval does not include 0, which also suggests that the data are not consistent with the null hypothesis.

```r
# Write your code here
t.test(yieldA, yieldB)
```

```
    Welch Two Sample t-test

data:  yieldA and yieldB
t = -4.9994, df = 19.441, p-value = 7.458e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9293569 -0.3814274
sample estimates:
mean of x mean of y
 4.052941  4.708333
```

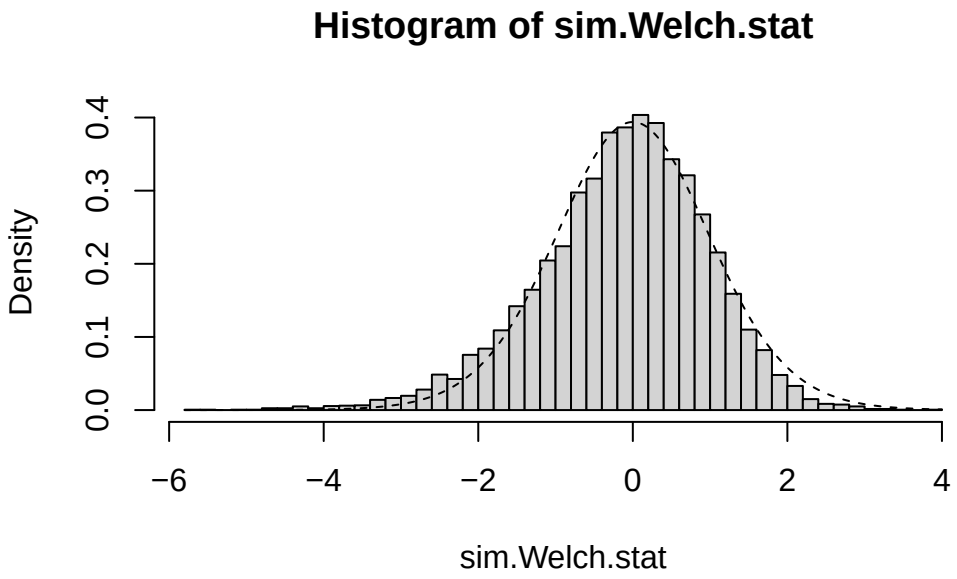
2.2.5 Calculate the P-value and 95% confidence interval using simulation.

- 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.

Solution:

```
boxA = yieldA-mean(yieldA)
boxB = yieldB-mean(yieldB)
sim.Welch.stat=0

for(i in 1:10000) {
    sampA = sample(boxA, size=nA, replace=T)
    sampB = sample(boxB, size=nB, replace=T)
    sim.Welch.stat[i] = t.test(sampA, sampB)$stat
}
hist(sim.Welch.stat, pr=T, n=50)
curve(dt(x, df=19.441), add=T, lty=2, n=1001)
```



The distribution of the simulated Welch test statistics is a bit skewed (possibly due to the outlier in `yieldA`).

The simulated P-value is

```r
mean(abs(sim.Welch.stat)>=abs(stat))
```

```
[1] 3e-04
```

which is larger than that of the Welch test. The CI is

```r
u.l = quantile(sim.Welch.stat, prob=c(.975, .025))
mA-mB - u.l*est.se.dif
```

```
     97.5%       2.5%
-0.8900815 -0.3249288
```

It is shifted a bit to the right, compared to the `t.test()` interval, perhaps due to the outlier in `yieldA`.