MAST20005/MAST90058: Week 8 Lab

Goals: (i) Practice various parametric and non-parametric tests for independent samples; (ii) Goodness-of-fit test; (iii) Association tests between categorical variables.

Data for Section 2: Butterfat data (`butterfat.txt`). Measurements of the amount of butterfat (in pounds) produced by 90 cows during a 305-day milk production period following their first calf. The data file can be obtained from the shared folder in the computer labs, or from the LMS.

1 Two-sample tests

Let X and Y be, respectively, the blood volumes in milliliters of a male who is a paraplegic and participates in vigorous physical activities and of a male who is able-bodied and participates in everyday, ordinary activities. Assume $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. The following are observations on X and Y:

```
x <- c(1612, 1352, 1256, 922, 1560, 1456, 2324)
y <- c(1082, 1300, 1092, 1040, 910 , 1248, 1092, 1040, 1092, 1288)
```

1. Test the null hypothesis $H_0 \colon \mu_X = \mu_Y$ against the two-sided alternative $H_1 \colon \mu_X \neq \mu_Y$ at the $\alpha = 0.05$ level of significance.

```
t.test(x, y, var.equal = TRUE)


##
##  Two Sample t-test
##
## data:  x and y
## t = 2.6641, df = 15, p-value = 0.01769
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   75.78143 682.27572
## sample estimates:
## mean of x mean of y
##  1497.429  1118.400
```

Since the p-value is smaller than $\alpha = 0.05$, we reject H_0 and conclude that there is sufficient evidence to show that the population means are the different.

2. Let us check the assumptions for the above test. Note that by using the option `var.equal = TRUE` we are assuming $\sigma_X = \sigma_Y$. We can check this assumption by testing $H_0 \colon \sigma_X = \sigma_Y$ against the alternative $H_1 \colon \sigma_X \neq \sigma_Y$ as follows:

```
var.test(x, y)


##
##  F test to compare two variances
```

```
##
## data:  x and y
## F = 12.121, num df = 6, denom df = 9, p-value = 0.001445
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##    2.805969 66.949239
## sample estimates:
## ratio of variances
##              12.121
```

We reject H_0 at the usual significance level of 5%. We can see that there is substantial evidence that the two variances differ.

3. A modification of the t-test known as Welch's test can help us correct for this problem by estimating both variances, and adjusting the degrees of freedom to use in the test. (This is the default procedure, if we do not specify `var.equal = TRUE` argument.)

```
t.test(x, y)
```

```
##
##   Welch Two Sample t-test
##
## data:  x and y
## t = 2.2644, df = 6.6981, p-value = 0.05963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -20.42556 778.48271
## sample estimates:
## mean of x mean of y
##   1497.429  1118.400
```

Now the p-value is larger than $\alpha = 0.05$, so we cannot reject $H_0 \colon \mu_X = \mu_Y$. It turns out that our earlier test was too optimistic: once we factor in the additional uncertainty from estimating two unknown variances it is harder to reject the null hypothesis.

4. Another assumption used in the t-test is normality. The following creates two normal QQ-plots:

```
par(mfrow = c(1, 2))
qqnorm(x, main = "X", col = 4)
qqline(x, lty = 2)
qqnorm(y, main = "Y", col = 4)
qqline(y, lty = 2)
```
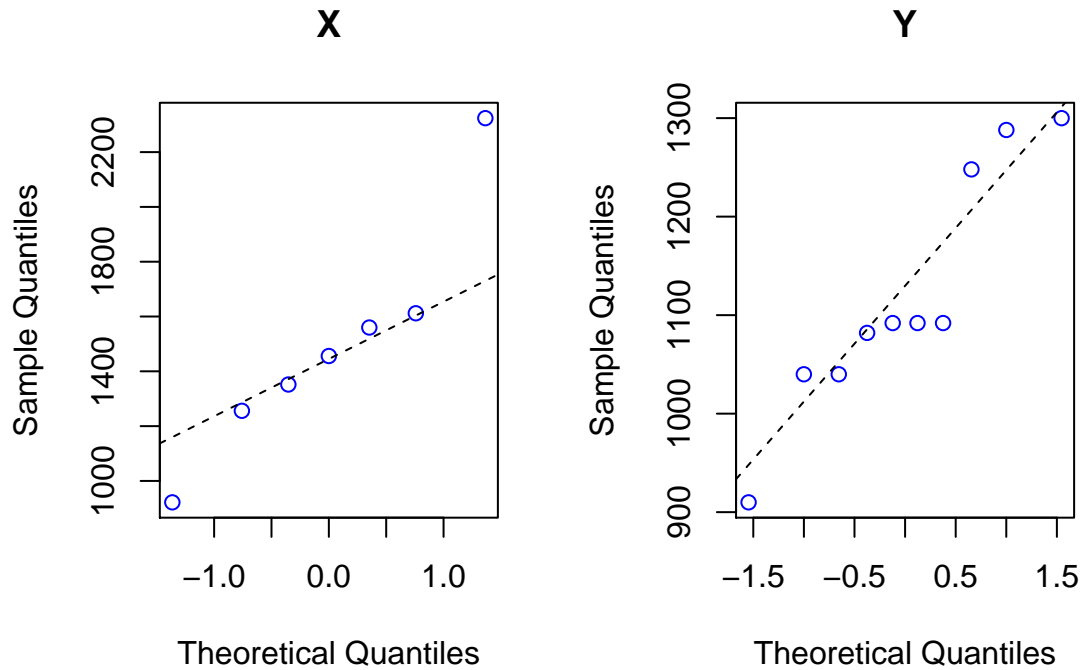
In the X sample there is a potential outlier in the right tail of the distribution, suggesting some degree of incompatibility with the normality assumption. The Y sample shows other deviations from the straight line.

5. The previous analyses suggest that a distribution-free test could suit these data better than the t-test. Consider a two-sample Wilcoxon test for the null hypothesis $H_0 \colon m_X = m_Y$ (population medians are equal) against the alternative $H_0 \colon m_X \neq m_Y$.

```
# Wilcoxon test with normal approximation
wilcox.test(x, y, exact = FALSE)


##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 59, p-value = 0.02142
## alternative hypothesis: true location shift is not equal to 0
```

The option `exact` specifies whether we wish to use a normal approximation for the distribution of the test statistic instead of the exact distribution.

6. The Wilcoxon test is said to be a robust test because outliers do not affect its results very much. For example, replace the smallest observation in the X sample by some more extreme number. What happens to the p-value?

```
y2 <- y
y2[5] <- 470
wilcox.test(x, y2, exact = FALSE)
```

3

2 Goodness-of-fit test

Consider the butterfat dataset. Let X be the amount of butterfat. Test the hypothesis that $X \sim \mathrm{N}(\mu, \sigma^2)$.

1. We first need to fit the model. To this end, compute the sample mean $\bar{x}$ and standard deviation s_x.

```
butterfat <- read.table("butterfat.txt")[, 1]
x.bar <- mean(butterfat)
s <- sd(butterfat)
```

2. Partition the data into a few discrete classes and compute observed frequencies.

```
b <- c(0, seq(374, 624, 50), 1000)      # class boundaries
O <- table(cut(butterfat, breaks = b))  # observed counts
O

##
##     (0,374]   (374,424]   (424,474]   (474,524]   (524,574]   (574,624]
##           5           9          16          22          15          13
## (624,1e+03]
##          10
```

3. We could compute expected frequencies using:

```
> prob <- rep(0, 7)
> prob[1] <- pnorm(b[2], x.bar, s) - pnorm(b[1], x.bar, s)
...
> prob[7] <- pnorm(b[8], x.bar, s) - pnorm(b[7], x.bar, s)
> E <- prob * length(butterfat)
```

Let's package this up into a loop:

```
prob <- rep(0, 7)
for (k in 1:7)
    prob[k] <- pnorm(b[k + 1], x.bar, s) - pnorm(b[k], x.bar, s)
E <- prob * length(butterfat)  # expected frequencies
```

and then carry out the test:

```
C <- sum((O - E)^2 / E)  # chi-squared statistic
d1 <- length(O)          # number of classes
qchisq(0.95,  d1 - 3)    # critical value

## [1] 9.487729

1 - pchisq(C, d1 - 3)    # p-value
```

```
## [1] 0.9076048
```

```
cbind(O, E)
```

```
##                O          E
## (0,374]        5   5.222232
## (374,424]      9   9.042687
## (424,474]     16  15.768114
## (474,524]     22  20.020264
## (524,574]     15  18.509602
## (574,624]     13  12.460972
## (624,1e+03]   10   8.976128
```

4. You could also conduct the chi-squared test using:

```
chisq.test(O, p = prob)
```

```
##
## 	Chi-squared test for given probabilities
##
## data:  O
## X-squared = 1.0144, df = 6, p-value = 0.9851
```

but remember that the final degrees of freedom should be $k - p - 1$, where k is the number of classes and p the number of estimated parameters. That means we need to recalculate the p-value using the given value of the statistic and the correct degrees of freedom:

```
1 - pchisq(1.0144, d1 - 3)
```

```
## [1] 0.9076047
```

5. Use the `qqnorm` command to illustrate your result.

3 Chi-squared test for association

The following data describe the state of grief for a sample of mothers who had suffered a neonatal death. The table relates this to the amount of support given to these women:

	Support		
Grief state	Good	Adequate	Poor
I	171	93	82
II	62	54	11
III	31	55	43
IV	11	27	57

5

1. We enter these data in R as follows:

```
X.raw <- c(171, 93, 82, 62, 54, 11, 31, 55, 43, 11, 27, 57)
X      <- matrix(X.raw, nrow = 4, byrow = TRUE)
test   <- chisq.test(X)
test


##
##  Pearson's Chi-squared test
##
## data:  X
## X-squared = 105.62, df = 6, p-value < 2.2e-16
```

The p-value is very small so we reject the null hypothesis that the rows and columns are independent. Hence, we conclude that we have strong evidence that the amount of support is associated with the grief state of mothers.

2. The following shows the expected frequencies:

```
test$expected

##             [,1]      [,2]     [,3]
## [1,] 136.51363 113.67862 95.80775
## [2,]  50.10760  41.72597 35.16643
## [3,]  50.89670  42.38307 35.72023
## [4,]  37.48207  31.21234 26.30560
```

In R, show how to compute the first cell in the table of expected frequencies and give the first term in the chi-squared statistic:

```
E.1 <- sum(X[, 1]) * sum(X[1, ]) / sum(X)
E.1

## [1] 136.5136


O.1 <- 171
(O.1 - E.1)^2 / E.1

## [1] 8.712022
```

Exercises

Do each of the tutorial problems but now using R.