MAST20005/MAST90058: Assignment 3 Solutions

1.
```r
x <- c(134, 146, 104, 119, 124, 161, 112, 83, 113, 129, 97, 123)
y <- c(70, 118, 101, 85, 107, 132, 94)
```

   (a) i. $H_0\colon m_x = 110$ versus $H_1\colon m_x < 110$.

```r
binom.test(sum(x > 110), length(x), alternative = "less")
##
##  Exact binomial test
##
## data:  sum(x > 110) and length(x)
## number of successes = 9, number of trials = 12, p-value =
## 0.9807
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.9281297
## sample estimates:
## probability of success
##                   0.75
```

The p-value is 0.98, so we cannot reject H_0.

ii. $H_0\colon m_x = m_y$ versus $H_1\colon m_x > m_y$.

```r
wilcox.test(x, y, alternative = "greater", exact = TRUE)
##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 63, p-value = 0.04156
## alternative hypothesis: true location shift is greater than 0
```

The p-value is 0.042, so we reject H_0.

iii. $H_0\colon \mu_x = \mu_y$ versus $H_1\colon \mu_x > \mu_y$.

```r
t.test(x, y, alternative = "greater")
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 1.9607, df = 12.964, p-value = 0.03588
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.875212      Inf
## sample estimates:
## mean of x mean of y
##  120.4167  101.0000
```

The p-value is 0.036, so we reject H_0.

2. (a) H_0: Poisson versus H_1: Not Poisson.
First, we need to estimate the rate parameter, λ, assuming a Poisson distribution.

```
hours <- 0:4
count <- c(10, 24, 10, 6, 3)
n <- sum(count)
lambda.hat <- sum(count * hours) / n
lambda.hat

## [1] 1.396226
```

Next, we calculate the expected counts under this assumption.

```
p <- c(dpois(0:3, lambda.hat), 1 - ppois(3, lambda.hat))
e <- n * p
e

## [1] 13.119052 18.317166 12.787456  5.951395  2.824932
```

The expected number for the last group (4+ hours) is less than 5, therefore we merge the last two groups together.

```
expected <- c(    e[1:3],     e[4] +     e[5])
observed <- c(count[1:3], count[4] + count[5])
Q <- sum((expected - observed)^2 / expected)
Q

## [1] 3.117952

1 - pchisq(Q, 2)

## [1] 0.2103513
```

The value of test statistic is 3.118 and it follows a χ^2 distribution with $4 - 1 - 1 = 2$ degrees of freedom. The p-value is 0.21, therefore we cannot reject H_0.
Note that we can also calculate the test statistic using, for example:

```
chisq.test(observed, p = expected, rescale.p = TRUE)

##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 3.118, df = 3, p-value = 0.3738
```

However, note that the p-value given by this is incorrect since it doesn't adjust the degrees for freedom appropriately.

(b) H_0: Age and exercise are independent vs H_1: Age and exercise are not independent.
Similar to part (a), we merge the last two columns (note that R will otherwise warn you that the chi-squared approximation may be inaccurate):

	0 hours	1 hour	2 hours	3+ hours
Younger than 40 years	10	24	10	9
40 years or older	7	22	18	15

```r
x <- rbind(younger = c(10, 24, 10, 9), older = c(7, 22, 18, 15))
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 3.7205, df = 3, p-value = 0.2933
```

The p-value from the test is 0.29, and therefore we cannot reject H_0.

3. The cdf of X is $F(x) = \int_\theta^x e^{-(y-\theta)}\,dy = \left[-e^{-(y-\theta)}\right]_\theta^x = 1 - e^{-(x-\theta)}$ if $x \geqslant \theta$. If $x < \theta$, $F(x) = 0$.

   (a) $\Pr(X_{(1)} > x) = (1 - F(x))^n = e^{-n(x-\theta)}$ if $x \geqslant \theta$, and 1 if $x < \theta$. Therefore,
   $$F_1(x) = \Pr(X_{(1)} \leqslant x) = (1 - e^{-n(x-\theta)})I(x \geqslant \theta).$$

   (b) By definition, $p = F(\pi_p) = 1 - e^{-(\pi_p - \theta)}$. Solving for π_p,
   $$\begin{aligned} e^{-(\pi_p-\theta)} &= 1 - p \\ -(\pi_p - \theta) &= \log(1-p) \\ \pi_p &= \theta - \log(1-p). \end{aligned}$$

   (c) The median of X is $m = \pi_{0.5} = \theta + \log 2$. To find the asymptotic variance of $\hat{M}$, we first need to find $f(m)$,
   $$f(m) = e^{-(m-\theta)} = e^{-\log 2} = 0.5.$$
   Using the asymptotic distribution of sample quantiles, we deduce that,
   $$\mathrm{var}(\hat{M}) \to \frac{1}{4nf(m)^2} = \frac{1}{n}.$$

4. (a) We have $X_{ij} \sim \mathrm{N}(\alpha_i, \sigma_j^2)$ and therefore by independence we deduce that $\sum_{j=1}^n X_{ij} \sim \mathrm{N}(n\alpha_i, \sum_{j=1}^n \sigma_j^2)$. Therefore, $\bar{X}_{i\cdot} \sim \mathrm{N}(\alpha_i, n^{-2}\sum_{j=1}^n \sigma_j^2)$.

   (b) By expansion and simplification, it is straightforward to show that:
   $$\sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot})^2 = \sum_{j=1}^n X_{ij}^2 - n\bar{X}_{i\cdot}^2.$$
   Then, using the identity $\mathbb{E}(A^2) = \mathrm{var}(A) + \mathbb{E}(A)^2$, we have:
   $$\begin{aligned} \mathbb{E}(X_{ij}^2) &= \alpha_i^2 + \sigma_j^2, \\ \mathbb{E}(\bar{X}_{i\cdot}^2) &= \alpha_i^2 + n^{-2}\sum_{j=1}^n \sigma_j^2. \end{aligned}$$
   Putting these together gives,
   $$\begin{aligned} \mathbb{E}\left\{\sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot})^2\right\} &= \sum_{j=1}^n (\alpha_i^2 + \sigma_j^2) - n(\alpha_i^2 + n^{-2}\sum_{j=1}^n \sigma_j^2) \\ &= \frac{n-1}{n}\sum_{j=1}^n \sigma_j^2. \end{aligned}$$

5.
```r
# Set up the data.
y <- c(270, 310, 220, 290, 350, 305, 446, 487, 500, 440, 428, 530,
       410, 305, 450, 382, 320, 380, 598, 480, 510, 470, 415, 400,
       180, 290, 330, 220, 170, 260, 290, 283, 260, 246, 275, 330)
loc  <- rep(c("OuterSurburb", "InnerSurburb", "CBD"), each = 12)
comp <- rep(3:0, times = 3, each = 3)

# Quick check that the factors are structured correctly.
loc

##  [1] "OuterSurburb" "OuterSurburb" "OuterSurburb" "OuterSurburb"
##  [5] "OuterSurburb" "OuterSurburb" "OuterSurburb" "OuterSurburb"
##  [9] "OuterSurburb" "OuterSurburb" "OuterSurburb" "OuterSurburb"
## [13] "InnerSurburb" "InnerSurburb" "InnerSurburb" "InnerSurburb"
## [17] "InnerSurburb" "InnerSurburb" "InnerSurburb" "InnerSurburb"
## [21] "InnerSurburb" "InnerSurburb" "InnerSurburb" "InnerSurburb"
## [25] "CBD"          "CBD"          "CBD"          "CBD"
## [29] "CBD"          "CBD"          "CBD"          "CBD"
## [33] "CBD"          "CBD"          "CBD"          "CBD"

comp

##  [1] 3 3 3 2 2 2 1 1 1 0 0 0 3 3 3 2 2 2 1 1 1 0 0 0 3 3 3 2 2 2 1 1 1
## [34] 0 0 0

# Two-way ANOVA.
anova(lm(y ~ factor(loc) + factor(comp)))

## Analysis of Variance Table
##
## Response: y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## factor(loc)   2 175542   87771  24.799 4.399e-07 ***
## factor(comp)  3 111311   37104  10.483 7.032e-05 ***
## Residuals    30 106180    3539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The outcome variable is modelled as $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$, with $i = 1, \dots, 3$ and $j = 1, \dots, 4$ and $k = 1, \dots, 3$, where ε_{ijk} denote independent errors such that $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$. Further, we assume that $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$. The null hypothesis of interest is $H_0\colon \alpha_i = 0$ for all i, while the alternative hypothesis is that at least one of the α_i is non-zero.

Let factor A denote be store locations, and factor B denote number of competitiors. The observed test statistic is $F = \frac{SS(A)/2}{SS(E)/30} = 24.8$, which will follow an $\mathrm{F}_{2,30}$ distribution under the null. Since the p-value is $4.4 \times 10^{-7} < 0.05$, we can reject H_0 at a 5% significance level. We have strong evidence that the retail sales are affected by store locations.

Since we have multiple observations for each factor combination, it **is** possible to test for interaction. Let's do that:

```
anova(lm(y ~ factor(loc) * factor(comp)))

## Analysis of Variance Table
##
## Response: y
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## factor(loc)               2 175542   87771 36.3023 5.528e-08 ***
## factor(comp)              3 111311   37104 15.3462 8.732e-06 ***
## factor(loc):factor(comp)  6  48153    8026  3.3194   0.01596 *
## Residuals                24  58027    2418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The outcome variable is modelled as $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$. The null hypothesis of interest is $H_0\colon \gamma_{ij} = 0$ for all i and j, while the alternative hypothesis is that at least one of the γ_{ij} is non-zero. We obtain a p-value of 0.016 and so we can reject null hypothesis. We have reasonable evidence of an interaction between store locations and the number of competitors.

```
par(mar = c(4, 4, 1, 1))   # tighter margins
interaction.plot(factor(loc), factor(comp), y, col = "blue")
```