# MAST20005/MAST90058: Assignment 2 Solutions

1. (a) $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$.

   (b)
```
prop.test(x = c(520, 600), n = c(800, 1000))

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(520, 600) out of c(800, 1000)
## X-squared = 4.5166, df = 1, p-value = 0.03357
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.003993323 0.096006677
## sample estimates:
## prop 1 prop 2
##   0.65   0.60
```

   The p-value is less than our significance level (0.05) so we reject the null hypothesis in this case. We have enough evidence to suggest that the rates differ between the cities.

   (c) From the R output, a 95% confidence interval for $p_1 - p_2$ is **(0.0040, 0.096)**.

   *Note:* Similar answers are obtained if you don't use continuity correction:

```
prop.test(x = c(520, 600), n = c(800, 1000))


##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(520, 600) out of c(800, 1000)
## X-squared = 4.5166, df = 1, p-value = 0.03357
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.003993323 0.096006677
## sample estimates:
## prop 1 prop 2
##   0.65   0.60
```

2. (a) The mean of sample is $\bar{x} = 15.06$, the standard error is $s_X = 2.26$, and the 97.5% percentile of $t_9(0.975) = 2.26$. Therefore, a 95% CI for $\mu_1$ is $\bar{x}_1 \pm t_9(0.975)s_X/\sqrt{n} = (12.79, 17.33)$.

   (b) The 97.5% percentile of the standard normal distribution is $c = 1.96$, and the sample size is

$$n = \left(\frac{c\sigma_X}{\epsilon}\right)^2 = \left(\frac{1.96 \times 3}{1}\right)^2 = 34.57.$$

   Therefore, we need a sample size of 35.

(c) Using the Welch approximation we get 95% confidence intervals for $\mu_X - \mu_Y$ of the form:

$$\bar{x} - \bar{y} \pm t_r^{-1}(0.975)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

with the value of $r$ given by the Welch approximation formula

$$r = \frac{(\frac{s_X^2}{n} + \frac{s_Y^2}{m})^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}} = \frac{(\frac{3.17^2}{10} + \frac{0.81^2}{8})^2}{\frac{3.17^4}{900} + \frac{0.81^4}{448}} = 10.44.$$

Therefore, $t_r^{-1}(0.975) = 2.21$, and a 95% confidence interval for $\mu_X - \mu_Y$ is

$$15.06 - 20.33 \pm 2.21\sqrt{\frac{3.17^2}{10} + \frac{0.81^2}{8}} = (-7.58, 2.96).$$

(d) The 0.025 and 0.975 quantiles of the $F_{7,9}$ distribution are 0.207 and 4.197 respectively. Therefore, a 95% confidence interval for $\sigma_X^2/\sigma_Y^2$ is $(0.207 \cdot s_X^2/s_Y^2, 4.197 \cdot s_X^2/s_Y^2) = (3.16, 64.13)$.

(e)
```
x = c(12.1, 12.2, 17.4, 13.1, 17.8, 19.8, 13.0, 10.8, 18.4, 16.0)
y = c(20.1, 21.3, 20.4, 21.7, 20.3, 19.5, 19.4, 19.9)
var.test(x,y)

##
##  F test to compare two variances
##
## data:  x and y
## F = 15.28, num df = 9, denom df = 7, p-value = 0.00163
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   3.167976 64.130185
## sample estimates:
## ratio of variances
##           15.27984
```

A 95% confidence interval for $\sigma_X^2/\sigma_Y^2$ is $(3.17, 64.13)$.

3. First, load the data in R:

```
coffee <- read.table("coffee.txt", header = TRUE)
```

(a) The following code fits the model, $\text{sales}_i = \alpha + \beta \times \text{numberOfCustomers}_i + \varepsilon_i$:

```
m1 = lm(sales ~ customer, data = coffee)
summary(m1)

##
## Call:
## lm(formula = sales ~ customer, data = coffee)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -116.51  -47.18    1.25   36.21  136.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.3442    62.2027   -0.52    0.609
## customer      6.4005     0.6345   10.09 7.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.6 on 18 degrees of freedom
## Multiple R-squared:  0.8497,Adjusted R-squared:  0.8413
## F-statistic: 101.7 on 1 and 18 DF,  p-value: 7.816e-09
```

Estimates: $\hat{\alpha} = -32.3$, $\hat{\beta} = 6.4$ and $\hat{\sigma} = 64.6$ ($\hat{\sigma}^2 = 4173$).

(b)
```
confint(m1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -163.027250 98.338933
## customer       5.067368  7.733558
```

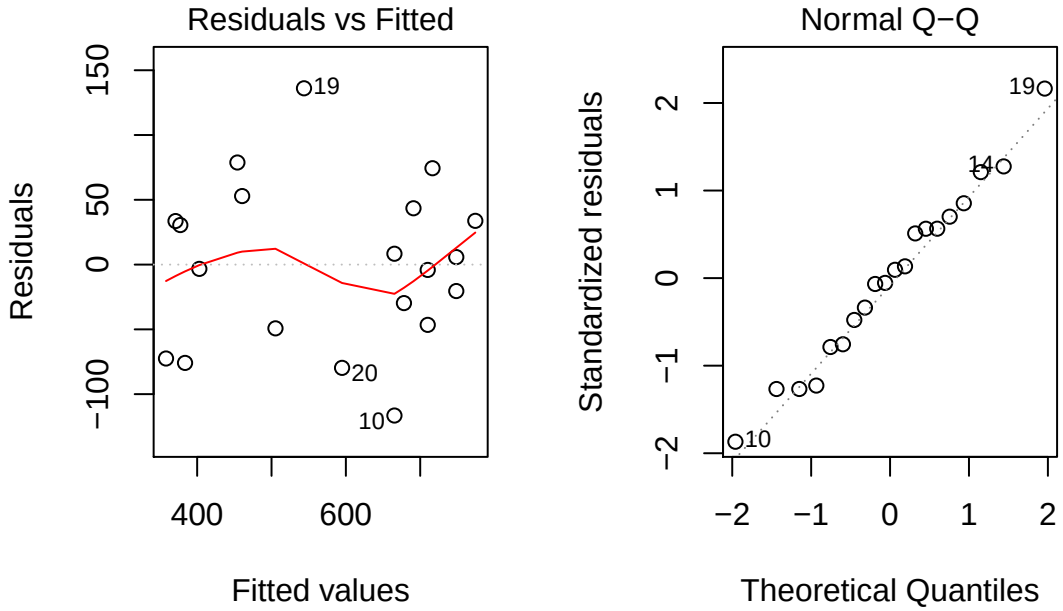
(c)
```
newdata = data.frame(customer = 100)
predict(m1, newdata, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 607.7022 576.7281 638.6762
```

(d)
```
predict(m1, newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 607.7022 468.4949 746.9094
```

(e)
```
par(mfrow = c(1, 2))
plot(m1, 1:2)
```




The model assumes that the error variance $\text{var}(\varepsilon_i) = \sigma^2$ is constant. If so, the plot of residuals (left) should show points randomly scattered around the zero line. The residuals plot does indeed look like this, and is thus consistent with the assumption of equal variance.

The QQ plot of the residuals (right) shows points being near the reference line, which is consistent with the assumption of Gaussian errors.

Overall, the model seems a resonably good fit to the data. If we wanted a simple model to explain the relationship, this one looks to be adequate.

4. (a) $T_1 \sim \text{N}(0, \sigma^2/n)$ is not a pivot, since the distribution of $T_1$ depends on the unknown parameter $\sigma^2$.

   (b) $T_2 \sim \text{N}(0, 1)$ is a pivot, since the distribution of $T_2$ does not depend on the unknown parameters.

   (c) $T_3 \sim t_{n-1}$ is a pivot, since the distribution of $T_3$ does not depend on the unknown parameters.

   (d) $T_4$ is a pivot, since $T_4 = T_3/\sqrt{n}$, and thus it is a deterministic function of $T_3$. Since $T_3$ is a pivot, $T_4$ must also be a pivot; both will have probability distributions that do not depend on the unknown parameters.

5. The cdf of $X$ is $F(x \mid \theta) = 1 - e^{-x/\theta}$, for $x > 0$.

   (a) $\alpha = \Pr(X > 4 \mid \theta = 2) = e^{-2} = 0.14$

   (b) $\beta = \Pr(X < 4 \mid \theta = 5) = 1 - e^{-4/5} = 0.55$

   (c) Power $= 1 - \beta = 0.45$

   (d) Solve $0.05 = \Pr(X > c \mid \theta = 2) = e^{-c/2}$, which gives $c = -2\ln(0.05) = 5.99$. Therefore, a test with rejection rule $X > 5.99$ has significance level 0.05.