MAST20005 - Assignment 2

Lucas Fern

21/09/2020

Question 1

1a)

$$H_0 : p_1 = p_2 \tag{1}$$
$$H_1 : p_1 \neq p_2 \tag{2}$$

1b)

```r
y <- c(520, 600)
n <- c(800, 1000)

prop.test(y, n, alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  y out of n
## 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 for this test is 0.03357, therefore there is enough evidence to reject H_0 at the 0.05 significance level.

1c)

```r
prop.test(y, n, alternative = "two.sided", conf.level=0.95)$conf.int
```

```
## [1] 0.003993323 0.096006677
## attr(,"conf.level")
## [1] 0.95
```

$(0.00399, 0.0960)$ is a 95% confidence interval for $p_1 - p_2$.

Question 2

2a)

Let the travel time be X

$$X \sim N(\mu_x, \sigma_x^2)$$

$$\Rightarrow \bar{X} \sim N(\mu_x, \sigma_x^2/n)$$

$$\frac{\bar{X}-\mu_x}{S/\sqrt{n}} \sim t_{n-1}$$

$$\bar{x} = 15.06$$

$$s = \tfrac{1}{9}\sum_{i=1}^{10}(x_i - 15.06)$$
$$= 3.17$$

With the CDF of t_9 as $F(x)$:

$$F^{-1}(0.025) = -2.26$$
$$F^{-1}(0.975) = 2.26$$

$$\Pr\left(-2.26 < \frac{\bar{X}-\mu_x}{S/\sqrt{n}} < 2.26\right) = 0.95$$

$$\Longrightarrow \left((-2.26)\left(\frac{3.17}{\sqrt{10}}\right)+15.06,\ (2.26)\left(\frac{3.17}{\sqrt{10}}\right)+15.06\right)$$

$\approx (12.79, 17.33)$ is a 95% CI for μ_x.

2b)

The CI is given by $\bar{x} \pm c\frac{\sigma_x}{\sqrt{n}}$ so $c\frac{\sigma_x}{\sqrt{n}} = 1$ is required.

$$c = \Phi^{-1}(0.975)$$

$$\Rightarrow 1.96\left(\frac{3}{\sqrt{n}}\right) = 1$$

$$n = (3 \times 1.96)^2$$

$$= 34.57$$

so a sample of at least 35 is needed.

2c)

Using Welch's approximation:

$$W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \sim t_r$$

$$r = \frac{\left(S_X^2/n + S_Y^2/m \right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

$$s_y^2 = 0.812$$
$$m = 8$$
$$\Longrightarrow r \approx 10.443$$

Let $F(x)$ be the CDF of $t_{10.443}$

$$c = F^{-1}(0.975) = 2.215$$

So $\bar{x} - \bar{y} \pm 2.215 \sqrt{\frac{3.17^2}{10} + \frac{0.812^2}{8}}$

$$= 15.06 - 20.325 \pm 2.215(1.044)$$
$$= (-7.58, -2.95) \text{ is a 95\% CI for } \mu_x - \mu_y$$

2d)

Let $F(x)$ be the CDF of $F_{m-1,n-1} = F_{7,9}$

$$c = F^{-1}(0.025) = 0.207$$
$$d = F^{-1}(0.975) = 4.197$$

A 95% CI for $\frac{\sigma_X^2}{\sigma_Y^2}$ is $\left(c\,\frac{S_X^2}{S_Y^2},\ d\,\frac{S_X^2}{S_Y^2}\right)$

$$= (3.168, 64.130)$$

2e)

```r
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)$conf.int
```

```
## [1]  3.167976 64.130185
## attr(,"conf.level")
## [1] 0.95
```

This shows the same confidence interval as 2d; $\approx (3.168, 64.130)$.

Question 3

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

3a)

```r
s.c.reg <- lm(sales ~ customer, data=coffee)
s.c.reg
```

```
##
## Call:
## lm(formula = sales ~ customer, data = coffee)
##
## Coefficients:
## (Intercept)     customer
##      -32.34         6.40
```

So the coefficients are $\alpha = -32.34$ and $\beta = 6.40$.

3b)

To find a 95% CI for the coefficients of the model:

```r
confint(s.c.reg)
```

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

So a 95% CI for α is $(-163.027, 98.339)$, and a 95% CI for β is $(5.067, 7.734)$.

3c)

```r
cust.100 <- data.frame("customer" = 100)
predict(s.c.reg, newdata = cust.100, interval = "confidence")
```

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

So $(576.73, 638.68)$ is a 95% CI for the mean when the number of customers is 100.

3d)

```r
predict(s.c.reg, newdata = cust.100, interval = "prediction")
```

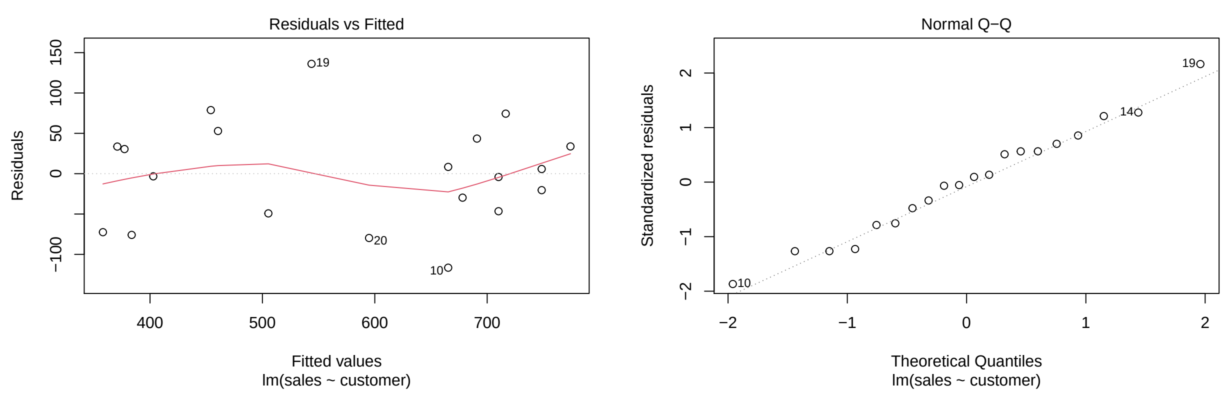
```
##        fit      lwr      upr
## 1 607.7022 468.4949 746.9094
```

So $(468.49, 746.91)$ is a 95% prediction interval for the number of sales when the number of customers is 100.

3e)

The regular regression model assumptions are that the magnitude of the residuals is independent of the fitted value, and that the residuals are normally distributed.

```r
plot(s.c.reg, 1:2)
```



The left plot shows that the size of the residuals does not seem to be dependent on the fitted value, so the assumption of homoscedasticity is valid, though it's hard to be confident with the relatively small data set. The QQ-plot on the right shows that the residuals do indeed seem to be normally distributed. Therefore the usual regression model assumptions are appropriate.

Question 4

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

4a)

$$(\bar{X} - \mu) \sim \mathrm{N}\left(0, \frac{\sigma^2}{n}\right)$$

So $\bar{X} - \mu$ is not a pivot since it's distribution depends on σ.

4b)

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}\left(0, 1\right)$$

So $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a pivot since it's distribution does not depend on μ or σ.

4c)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

So $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is a pivot since it's distribution does not depend on μ or σ.

4d)

$$\frac{\bar{X} - \mu}{S} = \frac{1}{\sqrt{n}} \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Since it has been shown that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is a pivot, and multiplying by $\frac{1}{\sqrt{n}}$ simply dilates the distribution by a factor of $\sqrt{n}$ (and doesn't reintroduce either of the parameters), $\frac{\bar{X}-\mu}{S}$ is also a pivot.

Question 5

5a)

Type 1 error : Reject H_0 when true.

$$= Pr(X > 4 \mid \theta = 2)$$
$$= 1 - Pr(X \leqslant 4 \mid \theta = 2)$$
$$= 1 - \int_0^4 f(x \mid \theta = 2)\, dx$$
$$= 1 - 0.8647$$
$$= 0.1353$$

So there is an ~13.53% chance of Type I error.

5b)

Type 2 error : Failing to reject H_0 when it is false.

$$= Pr(X \leqslant 4 \mid \theta = 5)$$
$$= \int_0^4 f(x \mid \theta = 5)\, dx$$
$$= 0.5507$$

So there is an ~55.07% chance of Type II error.

5c)

$$\begin{aligned} \text{Power} &= 1 - \Pr(\text{Type II error}) \\ &= 1 - 0.5507 \\ &= 0.4493 \end{aligned}$$

5d)

$E(X) = \lambda$

So we expect a realisation from $f(x|\theta=2)$ to be smaller than one from $f(x|\theta=5)$.

$\therefore$ Reject H_0 if $X > c$

$\Pr(X > c \mid \theta=2) = 0.05$

$\Rightarrow \Pr(X \leqslant c \mid \theta=2) = 0.95$

With $F(x)$ as the CDF of $f(x|\theta=2)$

$F^{-1}(0.95) = 5.99$

So with test statistic x we reject H_0 if x is in the critical region $(6, \infty)$.