Assignment 2 MAST20005 S2 2020 Michael Le (998211)

Question 1:

Let p1 be the proportion of residents who support more bicycle lanes in the City of Yarra and p2 be the proportion in the City of Moreland. Respective random samples of residents from each city, of size n1 = 800 and n2 = 1000, gave y1 = 520 and y2 = 600 respondents who support more bicycle lanes. Is there evidence that the proportions differ between the two cities? Set this up as a hypothesis test.

(a) State appropriate null and alternative hypotheses

```
x = c(520, 600) #respondents who support more bicycle lanes from 1 and 2
respectively
n = c(800, 1000) # sample sizes of residents from each city 1 and 2
respectively
prop.test(x,n,alternative = "greater", correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 4.7269, df = 1, p-value = 0.01485
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01233411 1.00000000
## sample estimates:
## prop 1 prop 2
##   0.65   0.60

#How can we tell our alternative hypothesis?

#H0: p1 = p2
#H1: p1 > p2
```

(b) Carry out a test that has significance level α = 0.05. What is your conclusion?

The output above shows the value of the chi-square statistic defined as Z^2. The observed this test is |z(obs)| = sqrt(4.5166) = 2.1252. When alpha = 0.05, the rejection region for this test is |z| > 1.67 (1 tail). (Or p-value = 0.01678 < 0.05 = alpha) Therefore, we reject H0 and conclude there is evidence that the difference of proportions of respondents who support more bicycle lanes between the two cities.

With continuity

The output above shows the value of the chi-square statistic defined as Z^2. The observed this test is |z(obs)| = sqrt(4.7269) = 2.1741. When alpha = 0.05, the rejection region for this test is |z| > 1.67 (1 tail). (or p-value = 0.01485 < 0.05 = alpha) Therefore, we reject H0 and conclude that the proportions differ between the two cities.

without continutiy

(c) Give a 95% confidence interval for the difference in proportions between the two cities. Given in the R-code the 95% confidence interval is

[0.01120911, 1.00000000] #with continuity

[0.01233411, 1.00000000] #without continuity

Q2a)

Sol^n

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{9}\left((12.1)^2 + (12.2)^2 + (17.4)^2 + (13.1)^2 + (17.8)^2 + (19.8)^2 + (13)^2 + (10.8)^2 + (18.4)^2 + (16)^2\right)$$

$$- \frac{1}{9}(10 \times 15.06^2) = 10.07$$

95 CI for μ:

$$\mu_1 = \bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n}x_i = \frac{1}{10}(12.1 + 12.2 + 17.4 + 13.1 + 17.8 + 19.8 + 13 + 10.8 + 18.4 + 16)$$

$$= 15.06$$

$$\mu_1 \pm s_1 t_{9}/\sqrt{n} = 15.06 \pm \frac{\sqrt{10.07} \times 2.262}{\sqrt{10}} = (12.79, 17.33)$$

(9 = degrees of freedom, n−1)

Q2b)

$$1.96 \times 3/\sqrt{n} = 1 \Rightarrow \frac{5.88}{\sqrt{n}} = 1 \Rightarrow (\sqrt{n})^2 = (5.88)^2$$

$$n = 34.5744$$

$$\simeq 35 \text{ experiments.}$$

Q2c) $\bar{x}_2 = \mu_2 = \frac{1}{m}\sum_{i=1}^{m}x_i = \frac{1}{8}(20.1 + 21.3 + 20.4 + 21.7 + 20.3 + 19.5 + 19.4 + 19.9)$

$$= 20.325 = \mu_2$$

$$s_2^2 = \frac{1}{m-1}\left(\sum_{i=1}^{m}x_i^2 - n\bar{x}^2\right) = \frac{1}{7}\left((20.1)^2 + (21.3)^2 + (20.4)^2 + (21.7)^2 + (20.3)^2 + (19.5)^2 + (19.4)^2 + (19.9)^2 - 8 \times 20.325^2\right)$$

$$= 0.6593$$

Normal 2 means, unknown σ, common variance

$$\mu_1 - \mu_2 \pm c \times s_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

$$S_p = \sqrt{\frac{S_1^2(n-1) + S_2^2(m-1)}{n+m-2}} = \sqrt{\frac{10.07 \times 9 + 0.6593 \times 7}{16}}$$

$c = t_{n+m-2} = t_{16} = 2.12$ (Assume it is 2-tailed)

$15.06 - 20.325 \pm 2.12 \times 2.44 \times \sqrt{\frac{1}{10} + \frac{1}{8}} = (-7.72, -2.81)$

Q2d) $\left[c\frac{S_1^2}{S_2^2}, d\frac{S_1^2}{S_2^2}\right]$

$F_{9,7}$:

$c = F^{-1}(0.05) = 0.3037$ ~~0.2383~~ 0.2075

$d = F^{-1}(0.95) = 3.6767$ ~~4.823~~ 4.1998

$\frac{S_1^2}{S_2^2} = \frac{10.07}{0.6593} = 15.27$

$$\left[0.3037 \times 15.27, 3.6767 \times 15.27\right]$$

$$= [4.64, 56.14]$$

Question 2e Calculate a 95% confidence interval for σ2X/σ2Y, using the R function var.test().

First we must refer from lab 4

```
x = c(12.1, 12.2, 17.4, 13.1, 17.8, 19.8, 13, 10.8, 18.4, 16)
#for the normal disrtibution for N(mew1,simga1^2)
var(x)

## [1] 10.07378

y = c(20.1, 21.3, 20.4, 21.7, 20.3, 19.5, 19.4, 19.9)
#for the normal disrtibution for N(mew2,simga2^2)
var(y)

## [1] 0.6592857

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
```

The 95 percent confidence interval is [3.167976, 64.130185]

Question 3: This question refers to the data in the file coffee.txt from a coffee shop. It shows the sales (in dollars) and the number of customers each day for twenty days.

(a) Fit a simple linear regression model for the sales given the number of customers. #State point estimates for all parameters.

```r
setwd("~/Desktop/mast20005-mast90058-semester2-2020 (4)/labs/data")
coffee = read.csv('coffee.csv')
m1 = lm(sales~customer, data = coffee) #Recall given refers to the
conditional probabilty where sales is the response variable and customers is
the predictor variable.

summary(m1)
```

```
##
## Call:
## lm(formula = sales ~ customer, data = coffee)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.494  -47.168    1.275   36.212  136.132
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.4582    62.1702  -0.522    0.608
## customer      6.4014     0.6342  10.094 7.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.57 on 18 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8415
## F-statistic: 101.9 on 1 and 18 DF,  p-value: 7.735e-09
```

```r
#Point estimates
beta.hat = 6.4014
alpha.hat =  -32.458
sigma.hat = 64.57
```

(b) Find 95% confidence intervals for the regression coefficients.

```r
confint(m1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -163.072882 98.156552
## customer       5.068999  7.733794
```

```r
#Case 1 95% Confidence Interval for alpha.hat:

##[-163.072882, 98.156552]

#Case 2 95% Confidence Interval for beta.hat:
##[5.068999, 7.733794]
```

(c) Give a 95% confidence interval for the mean sales if the number of customers is 100.

```r
# 95% Confidence Interval for the mean sales when
newdata <- data.frame(customer = 100)
predict(m1, newdata, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 607.6815 576.7237 638.6393
```

[576.7237, 638.6393]

(d) Give a 95% prediction interval for the sales if the number of customers is 100

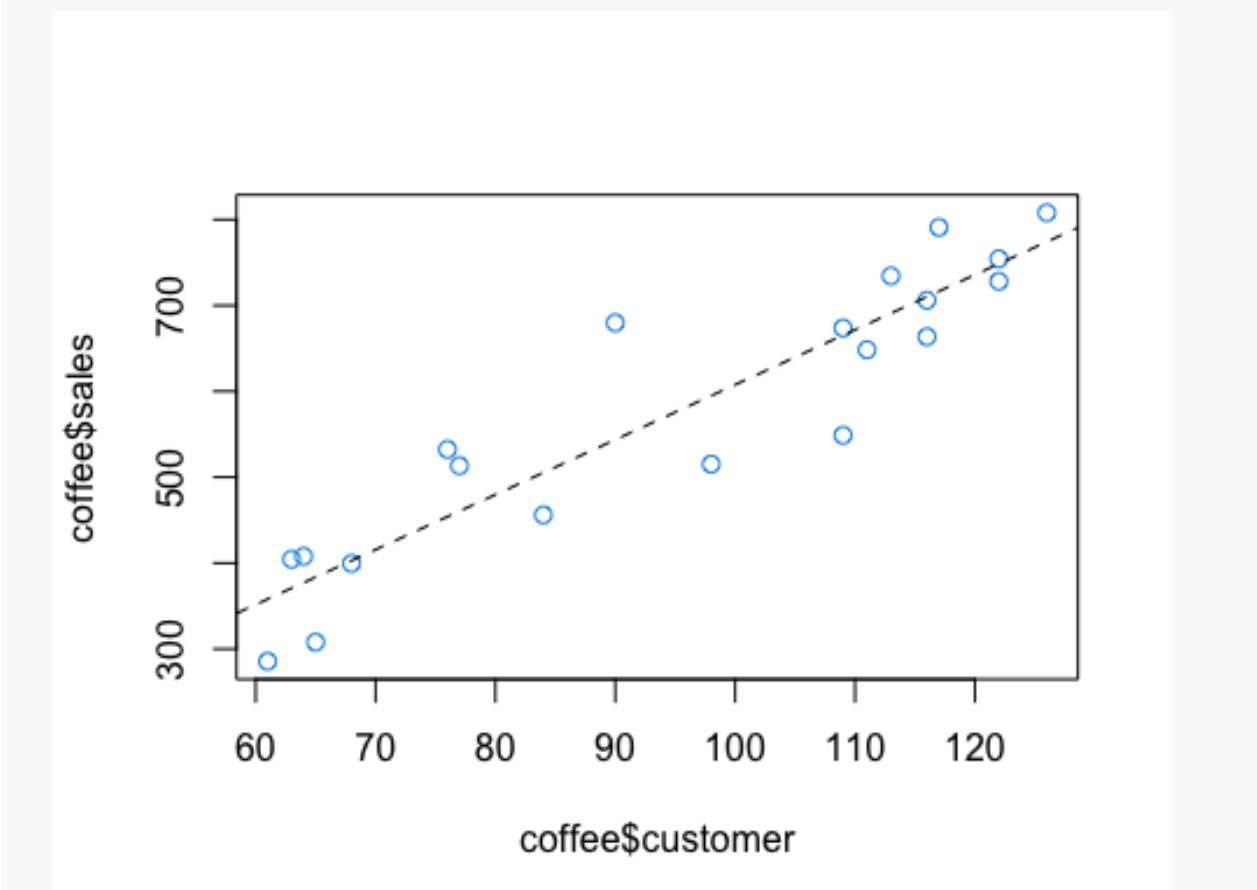
```r
predict(m1, newdata, interval = "prediction")
```

```
##        fit     lwr     upr
## 1 607.6815 468.547 746.816
```

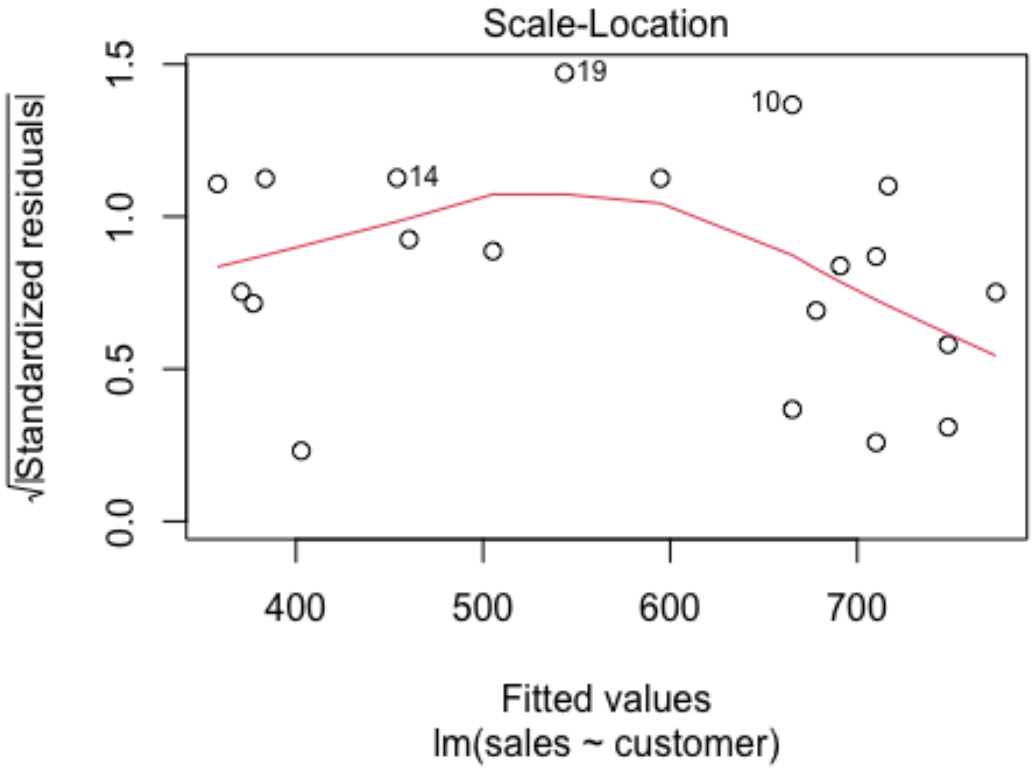
[468.547, 746.816]

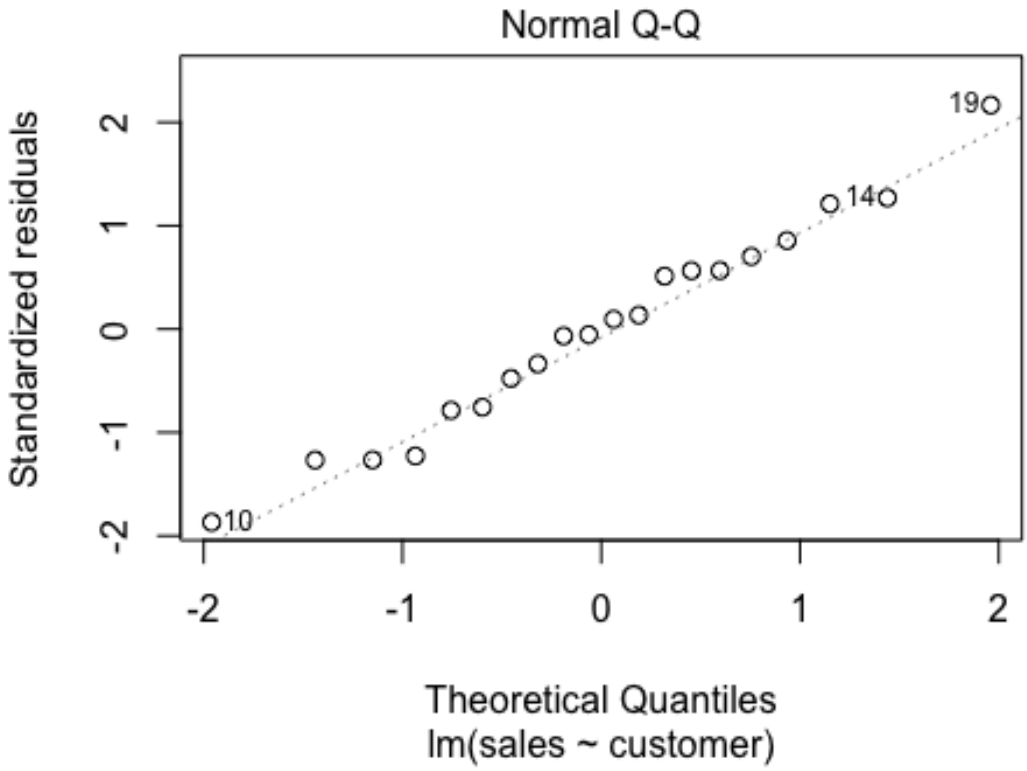
(e) Are the usual regression model assumptions appropriate? #Answer by reporting two appropriate diagnostic plots and commenting on them

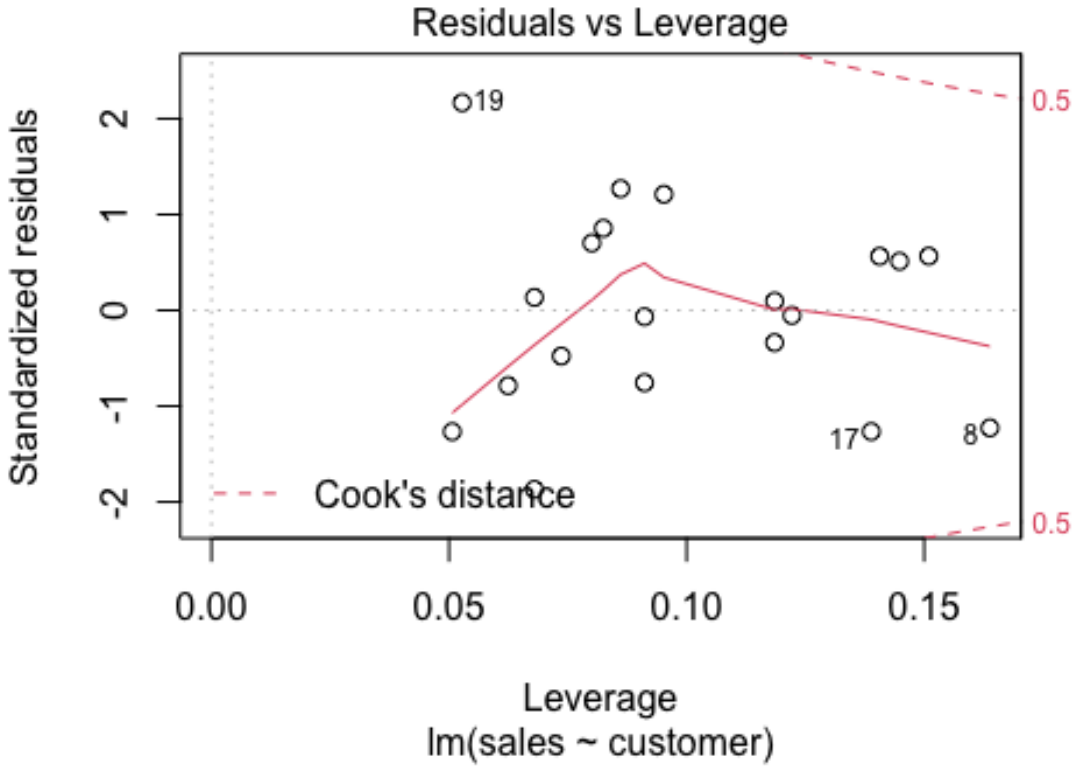
```r
plot(m1)
```

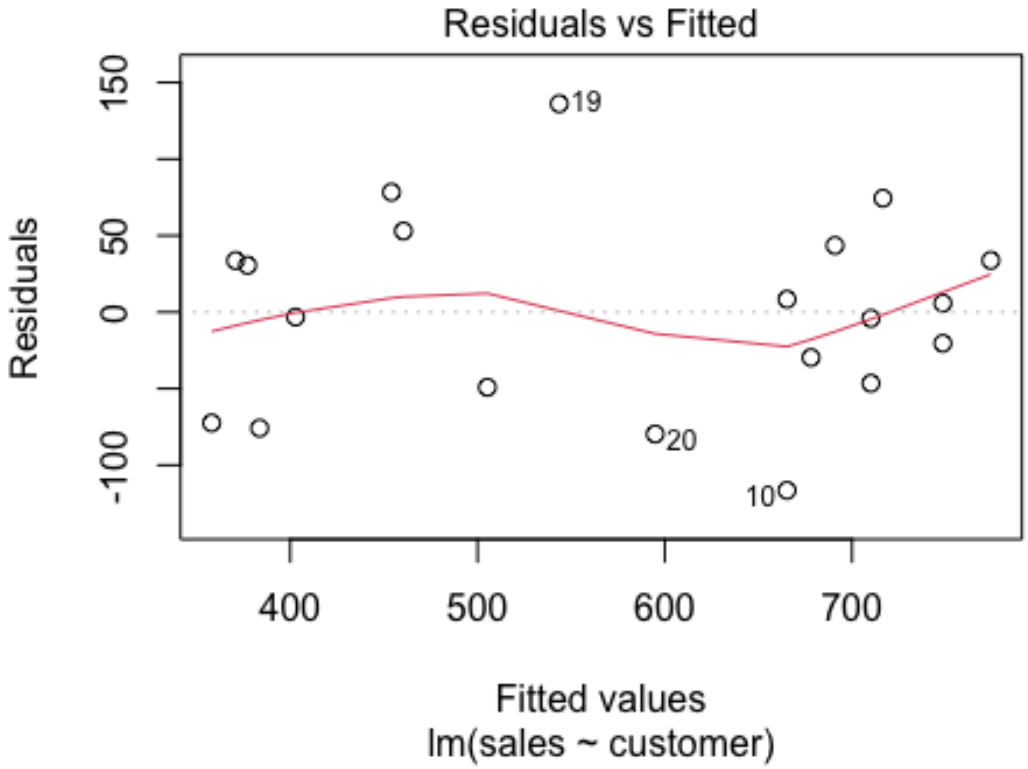


```r
plot(coffee$customer,coffee$sales,col=4)
abline(m1,lty=2)
```









From above plots, the linear model seems appropriate. Where the residuals are positibe in between roughly 400-550. There is also a slight increase of the number of residuals after 715.

Q49) $\bar{X} \sim N(\mu, \sigma^2/n)$, ~~$\bar{X}_n \sim N(\mu, 0)$~~

$T_1 = \bar{X} - \mu \neq^{d} N(0, \sigma^2/n)$ ← not a pivot, since it depends on distribution, μ is not parameter independent. (final answer)

$\sum (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2$

$= \sum [(X_i - \bar{X}) + (\bar{X} - \mu)]^2$

$= \sum_i (X_i - \bar{X})^2 + \sum (\bar{X} - \mu)^2 + 2\sum (X_i - \bar{X})(\bar{X} - \mu)$ (this $= 0$)

↳ Attempt from the lecture slides. (ignore this)

Q46) $T_2 = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{d}{=} N(0,1)$. $\bar{X} \sim N(\mu, \sigma^2/n)$, $\bar{X} - \mu \sim N(\mu, 0)$

$\bar{X} \sim (N, \sigma/\sqrt{n})$

⇒ pivot (see lecture slide 44) → this is independent of the parameter

Q4c) $T_3 = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ Standard normal.

fl. 40

$$T_3 = \frac{\overbrace{\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}}^{\sim N(0,1)}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$\dfrac{(n-1)S^2}{\sigma^2}$ ← chi-squared.

$$\chi^2_{n-1} = \underbrace{T_{2(1)}^2}_{\sim N(0,1)\ \chi^2_{N(0,1)}} + \underbrace{T_{2(2)}^2}_{\sim N(0,1)\ \chi^2_{N(0,1)}} + \ldots + \underbrace{T_{2(n)}^2}_{\sim N(0,1)\ \chi^2_{N(0,1)}}$$

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

$$\sqrt{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 + \ldots + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2}$$

⇒ σ^2 is unknown, this is ~~not~~ a pivot ⇒ since where given a sample variance, distribution does not change.

4d)

$T_4 = \frac{\bar{X} - \mu}{S}$ ← pivot:

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma/\sqrt{n}}{S} = \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$= \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - n\bar{X}^2)}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}\sqrt{\frac{1}{n-1}}\sqrt{\sum X_i^2 - n\bar{X}^2}}$$

$$= \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n(n-1)}\sum(X_i^2 - n\bar{X}^2)}}$$

$$\frac{n^2-n}{(n-\frac{1}{2})^2} \sim$$

$$= \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n^2 - n(n-\frac{1}{2})^2 - \frac{1}{4}}\sum(X_i^2 - n\bar{X}^2)}}$$

n

Thm $T_4 = \frac{\bar{X} - \mu}{S} \sim \frac{1}{\sqrt{n}} t_{n-1}$

⇒ Pivot: can depend on parameters
→ its distribution can not
→ i.e.

Q5 $H_0: \theta = 2$ $\quad \therefore f(x|\theta) = \theta^{-1} e^{-x/\theta}, \; x > 0$

$H_1: \theta = 5$ $\quad$ we reject the null if $X > 4$

A). $\alpha = Pr(\text{Type I error}) = Pr(\text{reject the } H_0 | H_0 \text{ is true})$

$= Pr(X > 4 | \theta = 2)$

or $\theta > 5$

$P(X>4) = \int_4^{\infty} \frac{1}{\theta} e^{-x/\theta} dx$

$= \frac{1}{\theta} \int_4^{\infty} e^{-x/\theta} dx = \frac{1}{\theta} \times \theta \left[-e^{-x/\theta} \right]_{4=x}^{\infty=x}$

$= e^{-4/\theta}$

sub $\theta = 2$ into $e^{-4/\theta} \rightarrow e^{-2} = 0.1353$ ✓

B). $\beta = Pr(\text{Type II error}) = Pr(\text{fail to reject } H_0 | H_0 \text{ is false})$

$= Pr(X \leq 4 | \theta = 5)$

$P(X \leq 3) = \int_0^4 \frac{1}{5} e^{-x/5} dx$

$x \geq 0$ constraint.
integrating from 0 to 3!!!

$= \left[-e^{-x/5} \right]_0^4 = -e^{-4/5} + 1 \rightarrow$ cdf Recall probability.

$= 0.5507$

C). $K = 1 - \beta = 1 - 0.5507$

$= 0.4493$

d) $X \geq 4$ is the critical region,

First we need to find the variance & mean as well.

2-statistic., $X \geq 4$ critical region.

$$\Rightarrow E(x) = \int_0^\infty x f(x)\,dx = \int_0^\infty \frac{x}{\theta} e^{-x/\theta}\,dx$$

Recall

$\alpha = Pr(\text{type I error}) = Pr(X \geq ? \mid \theta = 2) = 0.05$

? = what value of $X = x$ that gives the probability of the type I error of 0.05 (i.e. $\alpha = 0.05$ (5%)).

$H_0: \theta = 2$ and $H_1: \theta = 5$

$$\Rightarrow \frac{\frac{1}{2} e^{-x/2}}{\frac{1}{5} e^{-x/5}} \leq K$$

$$\Rightarrow \frac{5}{2} e^{-x(3/10)} \leq K$$

$$\Rightarrow K_1 \geq e^{-3x/10}$$

$$\Rightarrow K_2 \geq -\frac{3}{10} x$$

$$\Rightarrow x > K_3$$

let ? = a

$$0.05 = \int_a^\infty \frac{1}{2} e^{-x/2}\,dx$$

$$\left[-e^{-x/2}\right]_a^\infty = 0.05$$

$$e^{-a/2} = 0.05$$

$$-a/2 = \ln(0.05)$$

$$a = -2\ln(0.05) = 5.992$$

$a \geq 5.9915$

we reject the H_0 region.