Discrete Analysis Assignment 1

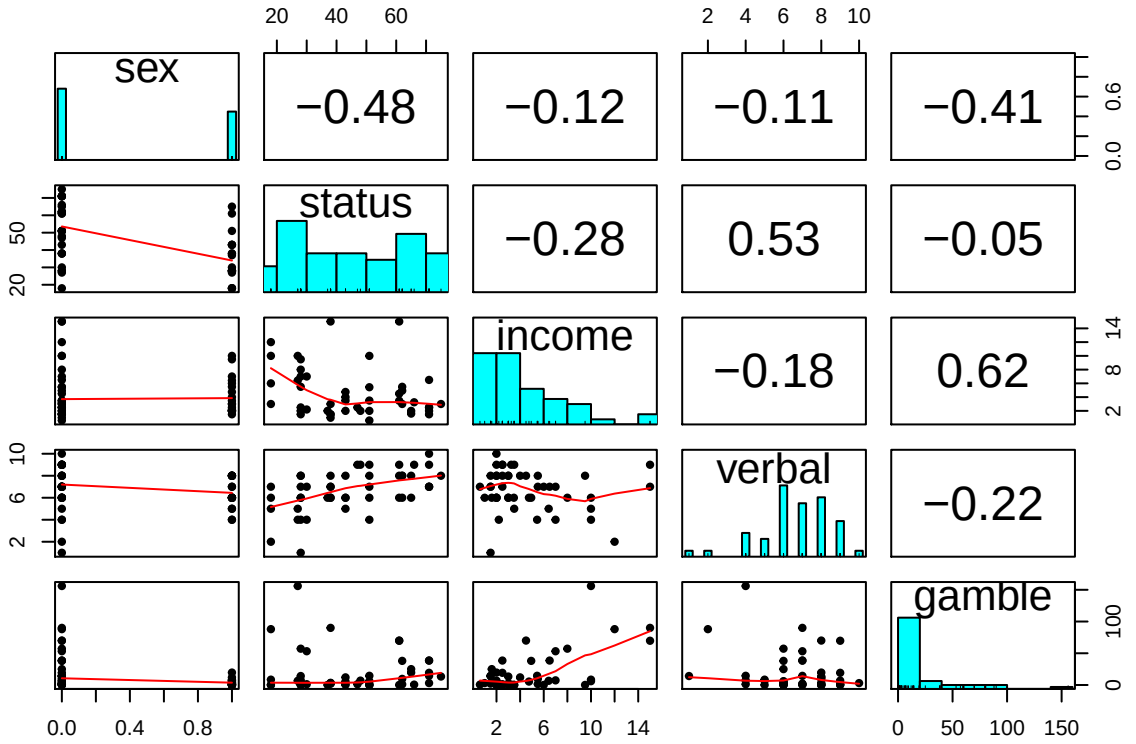
110024516 邱繼賢

Problem 1.

首先觀察各變數的數值和圖形特徵：

變數名稱	變數類型	變數範圍
sex	qulitative (nominal)	0=male ; 1=female
status	quantitative (approximately continuous)	18 ~ 75
income	quantitative (approximately continuous)	0.6 ~ 15
verbal	quantitative (discrete)	1,2,…,10
gamble	quantitative (approximately continuous)	0 ~ 156

```
##  sex        status          income          verbal          gamble      
##  0:28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0  
##  1:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1  
##         Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0  
##         Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3  
##         3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4  
##         Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```



- 性別男多於女
- 變數 *gamble* 和 *income* 皆有著右偏分佈
- 變數 *gamble* 有大量的數值為零
- 男性在變數 *status* 和 *gamble* 平均皆大於女性
- 變數 *income* 和 *gamble* 呈現正相關，可推測收入較高的人可能會投入較多的錢在賭博
- *income* 和 *gamble* 的散佈圖資料點多集中在左下角，較不易觀察
- 變數 *status* 和 *verbal* 呈現正相關，可推測父母社經地位較高者，語言能力也可能較高

將變數 *gamble+0.1* 以確保反應變數的數值皆大於零，方便未來進行 Box-Cox transformation，建構模型：

$$model_1 \ : \ gamble + 0.1 \ \sim \ sex \ + \ status \ + \ income \ + \ verbal$$

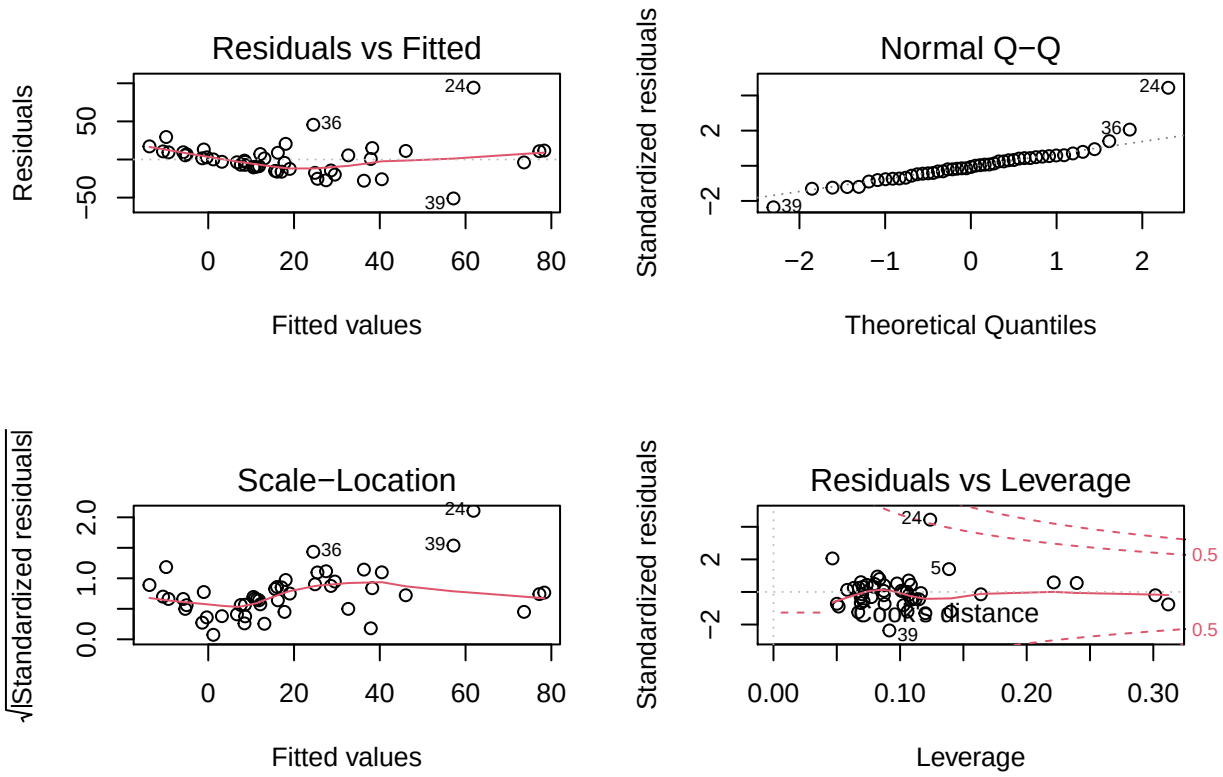
##

```
## Call:
## lm(formula = gamble + 0.1 ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.65565   17.19680   1.317   0.1948
## sex1        -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

- 變數 *sex* 和 *income* 呈現顯著

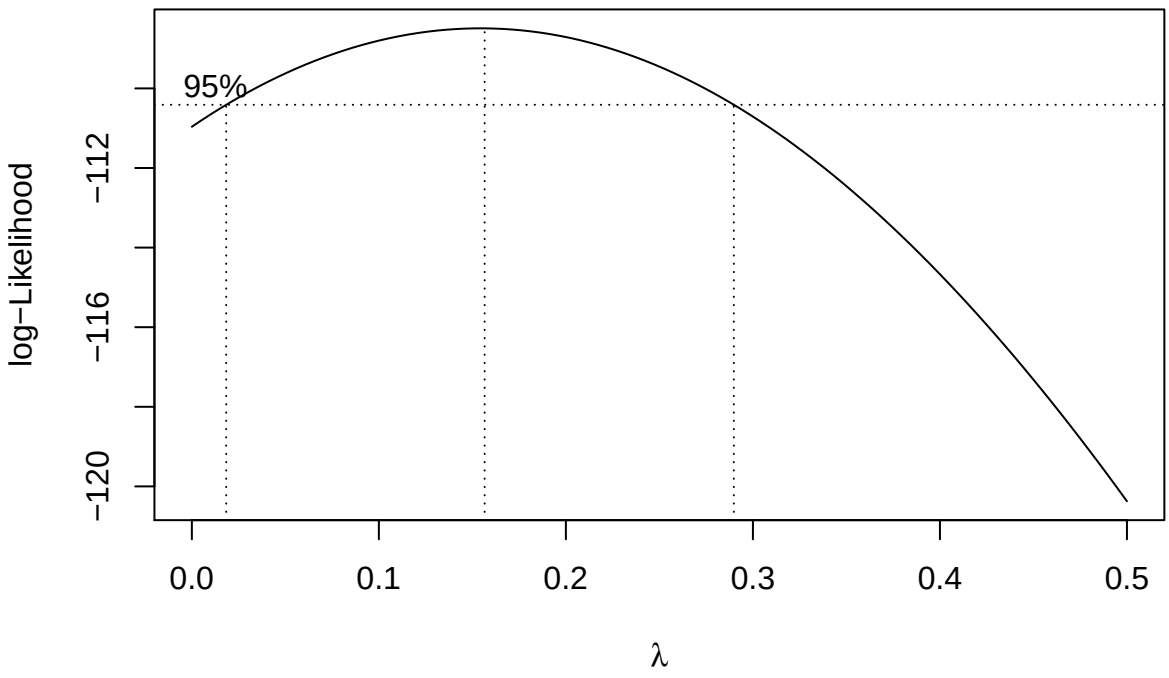
- $R^2 = 52.67\%$

接下來對模型做 diagnostic



- 第 24 個觀測值的 Cook's statistic 數值偏大，有可能為 influential observation

對模型做 Box- Cox transformation，檢定是否需要對反應變數做變換



- λ 的 95% 信賴區間並沒有包含 1，應對反應變數做 transformation
- 取 $\hat{\lambda} = \frac{1}{5}$ 來做變數變換

建構模型：

$$model_2 \ : \ (gamble + 0.1)^{\frac{1}{5}} \ \sim \ sex \ + \ status \ + \ income \ + \ verbal$$

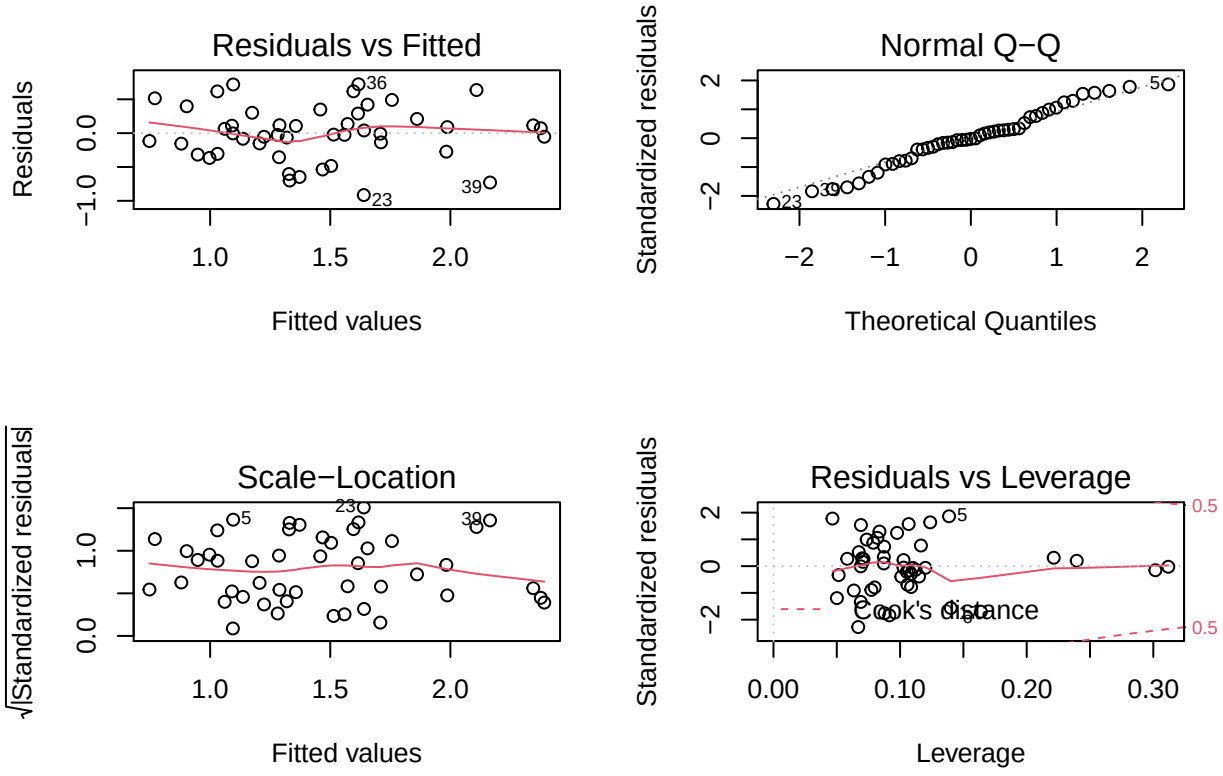
```
##
## Call:
## lm(formula = (gamble + 0.1)^0.2 ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91482 -0.21447 -0.00845  0.24979  0.72204
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.387125   0.315287   4.400 7.28e-05 ***
## sex1        -0.356334   0.150543  -2.367   0.0226 *
## status       0.010539   0.005154   2.045   0.0472 *
## income       0.083002   0.018800   4.415 6.93e-05 ***
## verbal      -0.098244   0.039824  -2.467   0.0178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.416 on 42 degrees of freedom
## Multiple R-squared:  0.525,  Adjusted R-squared:  0.4798
## F-statistic: 11.61 on 4 and 42 DF,  p-value: 1.95e-06
```

- 所有解釋變數皆呈現顯著

- $R^2 = 52.5\%$ 略小於 $model_1$

一樣對此模型做 diagnostic

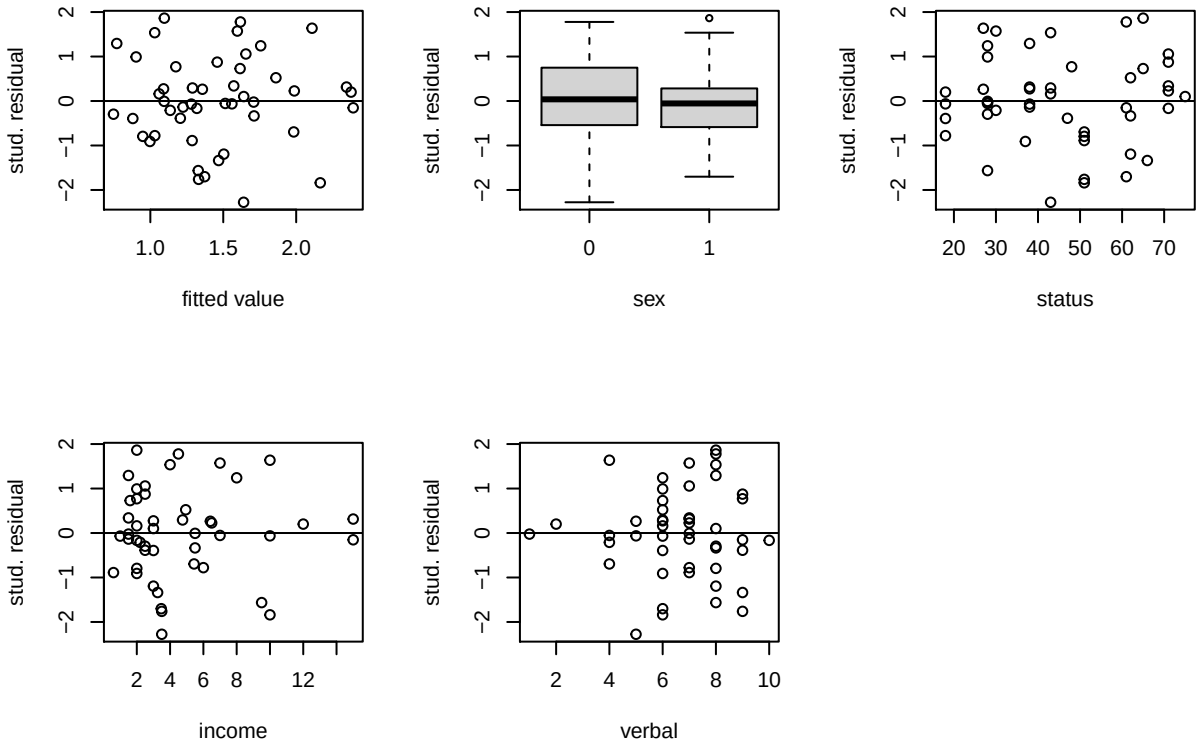


- 基本上都沒有觀測值為 influential observation

- studentized residual 也符合 normal assumption

- residual plot 也沒有 mean curvature 和 non-constant variance

再進一步對模型的 fitted value 和各變數繪製 residual plot



- fitted value 和各變數對 studentized residual 繪圖大致上也都沒有異常

雖然執行了 Box-Cox transformation 會對反應變數進行變數變換，進而造成模型係數解釋的不易，但由於此模型在各個解釋變數都呈現顯著，以及在 diagnostic 表現皆比 $model_1$ 來得優秀，故決定最終配適模型：

$$\widehat{\sqrt[5]{gamble + 0.1}} \;=\; \hat{Z} \;=\; 1.3871 - 0.3563\ sex + 0.0105\ status + 0.083\ income - 0.0982\ verbal$$

- 性別由男性變為女性，$Z(= \sqrt[5]{gamble + 0.1})$ 的預測值會隨之減少 *0.3563* 單位

- *status* 每上升一單位，Z 的預測值會隨之上升 *0.0105* 單位

- *income* 每上升一單位，Z 的預測值會隨之上升 *0.083* 單位

- *verbal* 每上升一單位，Z 的預測值會隨之下降 *0.0982* 單位

選擇一組變數

$$sex = 0\ ,\ status = 45\ ,\ income = 10\ ,\ verval = 7$$

代入模型中求得 Z 的預測值和預測區間上下界，在透過計算 $Z^5 - 0.1 = gamble$ 回推求得變數 $gamble$ 的預測值和預測區間

```
##        fit      lwr      upr
## 1 32.19719 1.741825 197.1693
```

- 此為資料內差的預測 $(0.6 \leq income \leq 15)$

- *gamble* 預測值為 32.2

- 預測區間大小為 $197.17 - 1.75 = 195.42$

選擇另一組變數

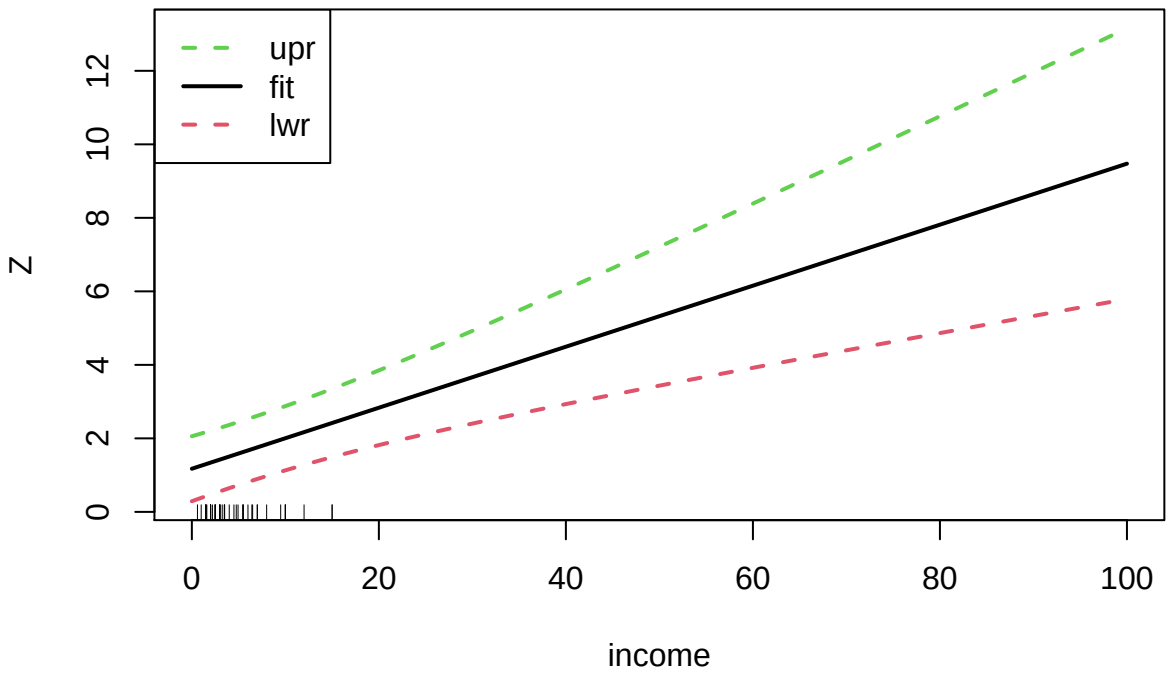
$$sex = 0\ ,\ status = 45\ ,\ income = 20\ ,\ verval = 7$$

```
##        fit      lwr    upr
## 1 182.6212 19.71766 846.01
```

- 此為資料外差的預測 $(income \geq 15)$

- *gamble* 預測值為 182.62

- 預測區間大小為 $846.01 - 19.72 = 826.29$
  $\Rightarrow$ 外差的預測區間寬度明顯大於內差的



- 此為 Z 在不同的 $income$ 下的預測值和預測區間

- 距離資料中心點越遠，預測區間的寬度越大

Problem 2.

a. nominal variable：政黨之間並沒有根據一個順序大小排列

b. ordinal variable：焦慮程度依照其嚴重度遞增排列

c. interval variable：病人存活月數為數組已知邊界的區間

d. nominal variable：診所地點並沒有根據一個順序大小排列

e. ordinal variable：腫瘤對化療的反應依照其根除的程度排列

f. nominal variable：喜歡的雜貨店並沒有根據一個順序大小排列

Problem 3.

a. Let X be the random variable of the number of correct answer in 100 questions, where $X \sim bin(n = 100, p = \frac{1}{4})$

b. $E(X) = np = 25$, $\sigma_X = \sqrt{Var(X)} = \sqrt{np(1-p)} = 4.3301$

It will be really surprising because $P(X \geq 50) = \sum_{x=50}^{100} C_x^{100} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{100-x} \approx 0$

c. $(n_1\ ,\ n_2\ ,\ n_3\ ,\ n_4) \sim multinomial\left(n = 100\ ,\ p_1 = \frac{1}{4}\ ,\ p_2 = \frac{1}{4}\ ,\ p_3 = \frac{1}{4}\ ,\ p_4 = \frac{1}{4}\right)$

d.

$$E(n_j) = np_j = 25\ ,\ Var(n_j) = np_j(1-p_j) = 18.75$$

$$Cov(n_j, n_k) = -np_ip_k = -6.25\ ,\ Cor(n_j, n_k) = \frac{Cov(n_j, n_k)}{\sqrt{Var(n_j)\ Var(n_k)}} = -0.333$$