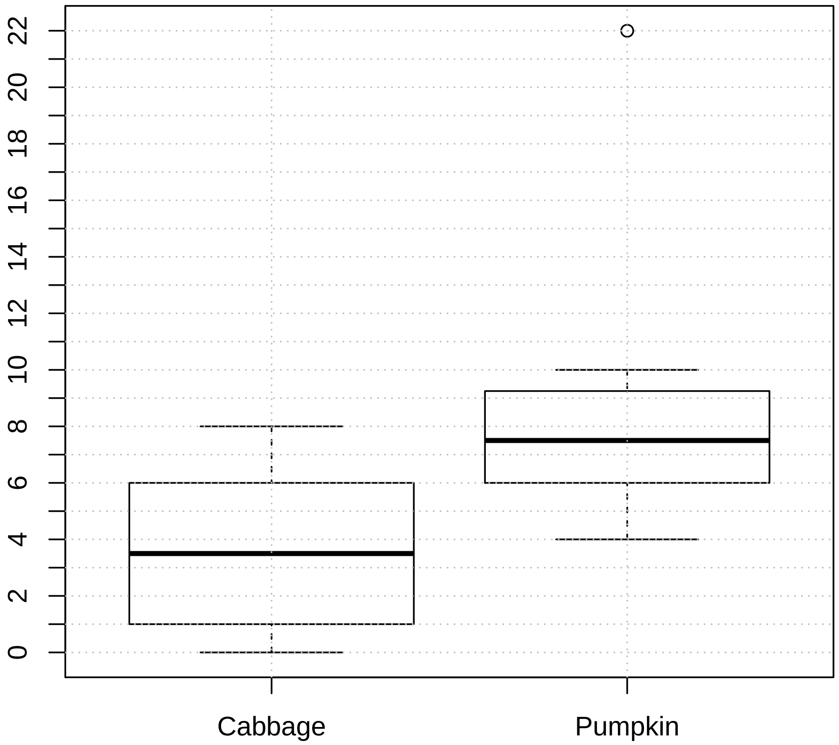
1.(a)Below there are two box plots that are shown.

Comparison of weights in Pounds



(i) Which vegetable has a higher median weight ?
Answer: Pumpkin. □

(ii) What is the approximate median weight of pumpkins ?
Answer: 7.5. □

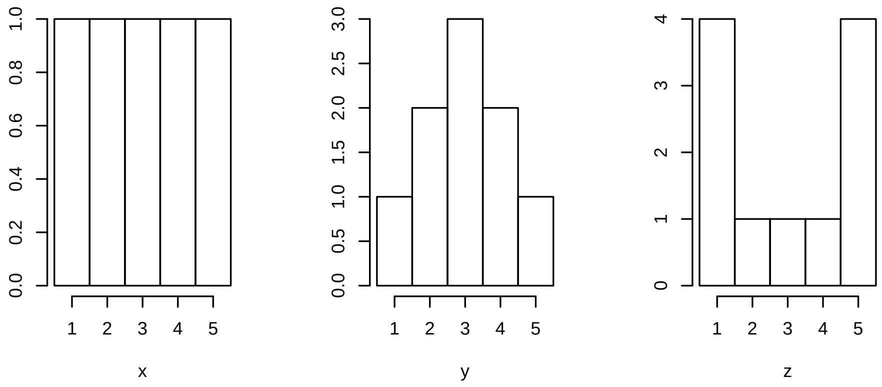
(iii) What is the `IQR` for Cabbages ?
Answer: Q3 = 6, Q1 =1 and therefore IQR=6-1 = 5. □

(iv) How many outliers are there in this data set ?
Answer: One for Pumpkin and 0 for Cabbage. One in this data set. □

(v) Which vegetable has a larger range of weights ?
Answer: Range of Pumpkin = 22-4= 18, Range of Cabbage = 8-0 =8. Pumpkin has larger range. □

1(b). In each of the following below, please circle the correct choice. No justification is required.

(I): Consider the three histograms given below of datasets x, y and z



The ordering of the dataset from the smallest to biggest standard deviations is given by:

$(i)(\mathtt{x},\mathtt{y},\mathtt{z})$ $(ii)(\mathtt{x},\mathtt{z},\mathtt{y})$ $(iii)(\mathtt{y},\mathtt{x},\mathtt{z})$ $(iv)(\mathtt{y},\mathtt{z},\mathtt{x})$ $(v)(\mathtt{z},\mathtt{y},\mathtt{x})$ $(vi)(\mathtt{z},\mathtt{x},\mathtt{y})$

Answer: The answer is (iii). □

(II): Bivariate data $\{(x_i, y_i)\}_{i=1}^{n}$ is given to us and is assumed to arise from the model, $y_i = bx_i + \varepsilon_i$, where ε_i are random variables. For simple linear regression to be appropriate, it is sufficient to assume that

(i)ε_i are all Normal (μ_i, σ_i^2). (ii) ε_i are all Normal (μ_i, σ).

(iii)ε_i are all independent. (iv) ε_i are all independent with mean 0 and Variance σ^2.

Answer: The answer is (iv). □

(III): Sambhavi has just finished a two sample t-test for equality in means between populations x and y. She concludes that the null hypothesis can be rejected at a level of significance 0.05. A best possible estimate for the probability that the two datasets came from distributions having the same mean is :

(i) 16.5 (ii) $\frac{1}{19}$ (iii)$\frac{1}{21}$ (iv)$\frac{1}{20}$

Answer: The answer is (iv). □

2. At the ISI co-rec basketball league in the 10 games played team *Unit-disc* scored:

$$59, 62, 59, 74, 70, 61, 62, 66, 62, 75$$

Assume that the number of points scored by *Unit-disc* is Normally distributed.

(a) Compute a 95% confidence interval for the mean, μ.

(b) We want to test the null hypothesis that $\mu = 63$ versus the alternative hypothesis that $\mu \neq 63$. Decide and execute a test that can check if there is enough evidence whether one can reject the null hypothesis at 5% level of significance.

Answer: 2(a) Let $X_1, X_2, X_3, \ldots, X_{10}$ represent the data set. Then

$$\bar{X} = \frac{1}{9}\sum_{i=1}^{10} X_i = 65 \text{ and } S = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X})^2} = 5.98.$$

Further by Normality assumption we know that

$$\sqrt{10} \cdot \frac{\bar{X} - \mu}{S} \sim t_9,$$

where t_9 follows t-distribution with 9 degrees of freedom. Now, suppose $t_{9,0.975}$ is 97.5% quantile of t_9 then by the symmetry of the t-distribution we have

$$\mathbb{P}(\left|\sqrt{10} \cdot \frac{\bar{X} - \mu}{S}\right| \leq t_{9,0.975}) = \mathbb{P}(|t_9| \leq t_{9,0.975}) = 0.95.$$

From the Table 3, we have $t_{9,0.975} = 2.26$. Therefore

$$\mathbb{P}(\left|\sqrt{10} \cdot \frac{\bar{X} - \mu}{S}\right| \leq 2.26) = 0.95.$$

Consequently a 95% confidence interval for μ is given by

$$\left(\bar{X} - \frac{t_{9,0.975}\, S}{\sqrt{10}}, \bar{X} + \frac{t_{9,0.975}\, S}{\sqrt{10}}\right) = \left(65 - \frac{((2.26)(5.98)}{\sqrt{10}}, 65 + \frac{((2.26)(5.98)}{\sqrt{10}}\right) \approx (60.72, 69.27)$$

□

Answer: 2(b) Under the null hypothesis, $\mu = 63$. Given that the alternative hypothesis is $\mu \neq 63$. Then the p-value is given by

$$\mathbb{P}(| t_9 | \geq \sqrt{10}\frac{\bar{X} - 63}{S})$$

which is equal to

$$\begin{aligned}\mathbb{P}(\mid t_9 \mid\geq \sqrt{10}\frac{65-63}{5.98}) &= \mathbb{P}(\mid t_9 \mid\geq 1.057) \qquad \text{(by symmetry of } t_9 \text{ distribution)}\\ &= 2\mathbb{P}(t_9 \geq 1.057)\\ &= 2(1-\mathbb{P}(t_9 \leq 1.057)) \qquad \text{(from Table 1)}\\ &\approx 2(1-(0.828)) = 0.344\end{aligned}$$

As the p-value is larger than 0.05 there is not enough evidence to reject the null hypothesis at 5% level of significance.

□

3. Gobarkanth collects $X_1, X_2, X_3, \ldots, X_n$ of i.i.d measurements of radiation from Canteen's Gobar Gas plant. He assumes that the observations follow a Rayleigh distribution with parameter α, with p.d.f. given by

$$f(x) = \begin{cases} \alpha x \exp(-\frac{1}{2}\alpha x^2) & \text{if } x \geq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Find the maximum likelihood estimate for α.

Answer: We observe that $\alpha \in (0,\infty)$ and the likelihood function from $X_1, X_2, X_3, \ldots, X_n$ is given by

$$L(\alpha; X_1, X_2, X_3, \ldots, X_n) = \begin{cases} \alpha^n \prod_{i=1}^n X_i \exp(-\frac{1}{2}\alpha \sum_{i=1}^n X_i^2) & \text{if } X_i \geq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Suppose any of the observations are 0 then the likelihood is a constant function with av Given the assumptions on the sample collected by Gobarkanth, the log-likelihood is $LL : (0,\infty) \to \mathbb{R}$ given by

$$LL(\alpha) \equiv LL(\alpha; X_1, X_2, X_3, \ldots, X_n) = n\ln(\alpha) + \ln(\sum_{i=1}^n X_i^2) - \frac{1}{2}\alpha\sum_{i=1}^n X_i^2$$

It is clear that on $(0,\infty)$

$$\frac{\partial}{\partial\alpha}LL(\alpha) = \frac{n}{\alpha} - \frac{1}{2}\sum_{i=1}^n X_i^2 \qquad \text{and} \qquad \frac{\partial^2}{\partial^2\alpha}LL(\alpha) = -\frac{n}{\alpha^2}$$

As the second derivate is always negative, the critical point

$$\hat{\alpha} = \frac{2n}{\sum_{i=1}^n X_i^2}$$

is a global maximum for LL and as ln is an increasing function it is a global maximum for L as well. Therefore the M.L.E is given by $\hat{\alpha}$.

□

4. The following `R` code simulates a random variable `X`

```
> L = 10
> i = 0
> U = runif(1, min=0, max =1)
> Y = -log(U)/L
> Sum = Y
> while (Sum<1) {
+     U = runif(1, min=0, max =1)
+     Y = -log(U)/L
+     Sum = Sum +Y
+     i = i + 1
+ }
> X = i
```

Find the distribution of X (*Other than p.d.f. or p.m.f. of standard distribution functions please provide adequate justification of any result that you are using*).

Answer: Let $\{U_n\}_{n\geq 1}$ be Uniform $(0,1)$ random variables. Then the output X in the above algorithm is given by

$$X = \begin{cases} 0 & \text{if } \frac{-1}{10}\ln(U_1) > 1 \\ \max\{j \geq 1 : \frac{-1}{10}\ln(U_1 U_2 \ldots U_j) \leq 1\} & \text{otherwise} \end{cases}$$

First note that,

$$\text{Range}\{X\} = \{0\} \cup \mathbb{N}. \tag{1}$$

Observe that

$$\mathbb{P}(X = 0) = \mathbb{P}(\frac{-1}{10}\ln(U_1) > 1) = \mathbb{P}(U_1 \leq e^{-10}) = e^{-10} \tag{2}$$

For $i \geq 1$, let $T_i = -\frac{1}{10}\ln(U_i)$. Then $T_i \sim \text{Exp}(10)$ and consequently

$$\frac{-1}{10}\ln(U_1 U_2 \ldots U_j) = \frac{-1}{10}\sum_{i=1}^{j}\ln(U_i) = \sum_{i=1}^{j} T_i \sim \text{Gamma}(j, 10).$$

For $n \geq 1$, observe that

$$\{X \geq n\} = \{\frac{-1}{10}\ln(U_1 U_2 \ldots U_n) \leq 1\} = \{\sum_{i=1}^{n} T_i \leq 1\}$$

Using this, for $n \geq 1$,

$$\begin{aligned}
\mathbb{P}(X = n) &= \mathbb{P}(X \geq n) - \mathbb{P}(X \geq n+1) \\
&= \mathbb{P}(\sum_{i=1}^{n} T_i \leq 1) - \mathbb{P}(\sum_{i=1}^{n+1} T_i \leq 1) \\
&= \frac{10^n}{n-1!} \int_0^1 z^{n-1} e^{-10z} dz - \frac{10^{n+1}}{n!} \int_0^1 z^n e^{-10z} dz \\
&\quad \text{(Using the p.d.f of Gamma distribution)} \\
&= \frac{10^n}{n-1!} \int_0^1 z^{n-1} e^{-10z} dz - \frac{10^{n+1}}{n!} \left[-\frac{e^{-10z} z^n}{10} \Big|_0^1 + \frac{n}{10} \int_0^1 z^{n-1} e^{-10z} dz \right] \\
&\quad \text{(Integration by parts)} \\
&= e^{-10} \frac{10^n}{n!}. \qquad \text{(As the above of limit of partial sums)} \qquad (3)
\end{aligned}$$

From (1),(2), and (3) we conclude that $X \sim \text{Poisson}(10)$ □

5. In an experiment in breeding plants, a geneticist has obtained 219 brown wrinkled seeds, 81 brown roundseeds, 69 white wrinkled seeds and 31 white round seeds. Theory predicts that these types of seeds should be obtained in the ratios $9:3:3:1$. Assuming that the null hypothesis is given by the theory, execute a test that can check if there is enough evidence to reject the null hypothesis at 5% level of significance.

Answer: The theory predicts ratio of seeds to be $9:3:3:1$. Hence the probability for each variety occuring is given by

$$\left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

From the experiments it was observed that there were 219 brown wrinkled seeds, 81 brown roundseeds, 69 white wrinkled seeds and 31 white round seeds. Hence the total number of seeds under consideration

$$219 + 81 + 69 + 31 = 400.$$

Therefore the expected counts from the population of 400 for each variety are given by

$$\left(\frac{9}{16} \times 400, \frac{3}{16} \times 400, \frac{3}{16} \times 400, \frac{1}{16} \times 400 \right) = (225, 75, 75, 25).$$

The chi-square statistic is then given by

$$\begin{aligned}
X^2 &= \tfrac{(219-225)^2}{225} + \tfrac{(81-75)^2}{75} + \tfrac{(69-75)^2}{75} + \tfrac{(31-25)^2}{25} \\
&= \tfrac{64}{25} = 2.56
\end{aligned}$$

Applying the chi-square goodness of fit test, we would take the null hypothesis to be as theory has predicted. The $p-$value being calculated, using Table 7,

$$\mathbb{P}(\chi_3^2 \geq X^2) = \mathbb{P}(\chi_3^2 \geq 2.56) = 1 - \mathbb{P}(\chi_3^2 \leq 2.56) = 1 - 0.535 = 0.465$$

Since the p-value is larger than 0.05 we do not reject the null hypothesis.

6. The responses for three treatments `A`, `B`, `C` to a population of mice are given below. We wish to verify if the treatments are different or not. The data is entered and following test is performed by `R`.

```
> A = c(37, 39, 90, 92, 51)
> B = c(13, 17, 46, 30, 23)
> C= c(52, 25, 23, 43, 52)
> y= c(A,B,C)
> x = c(rep("A",5),rep("B",5),rep("C",5))
> oneway.test(y~x, var.equal=TRUE)
```

```
        One-way analysis of means

data:  y and x
F = 4.4826, num df = 2, denom df = 12, p-value = 0.03516
```

(a) Describe what test is being performed and what is the conclusion you can infer.

(b) Denote the data set by $y := (y_{ij})$ with $1 \leq i \leq I, 1 \leq j \leq J$. In terms of (y_{ij}), explain what are: `F`, `num df`, `denom df` and `p-value` in the above output.

Answer: 6(a) `One way analysis of means` is test that can be used to check if the treatments (three above) have the same mean responses (under normality assumption). Specifically the null hypothesis is that all the treatments have the same mean response. In the above the p-value is 0.03516 and we would reject the null hypothesis at 5% level of siginifcance.

Answer: 6(b) Let I be the number of treatments, in the above $I = 3$. Let J be population size, in the above $J = 5$. We assume that the responses y_{ij} are modeled as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

with errors ϵ_{ij} being independent Normal $(0, \sigma^2)$ and differential effect $\sum_{i=1}^{I} \alpha_i = 0$.

We propose:
Null Hypothesis: $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$ and
Alternative Hypothesis: One of the α_i differ from 0.

Observe that

$$\text{Total Sum of squares} := \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y_{i\cdot}})^2 + J\sum_{i=1}^{I}(\bar{y_{i\cdot}} - \bar{y})^2$$

$$:= \text{Sum of squares within treatment groups}$$

$$+ \text{Sum of squares between treatment groups}$$

In Short:

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{W}} + \mathrm{SS}_{\mathrm{B}}$$

We consider the Test Statistic:

$$F := \frac{\mathrm{SS}_{\mathrm{B}}/(I-1)}{\mathrm{SS}_{\mathrm{W}}/(I(J-1)}.$$

It can be shown that $F \sim F(I-1, I(J-1))$. Then we decide on α and calculate

$$p-\text{value } := P\left(F(I-1, I(J-1)) > \frac{\mathrm{SS}_{\mathrm{B}}/(I-1)}{\mathrm{SS}_{\mathrm{W}}/(I(J-1)}\right)$$

We reject the Null Hypothesis if p-value is less that α.

Relating the above to the given code:

$I = 3, J = 5$ and `F` $= \frac{\mathrm{SS}_{\mathrm{B}}/(I-1)}{\mathrm{SS}_{\mathrm{W}}/(I(J-1)} = 4.4826$
`num df` $=$ I-1 $= 2$
`denom df` $=$ I(J-1) $= 12$
`p-value` gives $P\left(F(I-1, I(J-1)) > \texttt{F}\right)) = P\left(F(2, 12) > 4.4826\right)) = 1 - 0.965 = 0.035$ from Table 8. □