Q1

Let $\{X_i : 1 \leq i \leq n\}$ be a random sample from a population $X \sim F(\mu, \sigma^2)$, a distributional model with finite but unknown mean μ ∈ R and variance $\sigma^2 > 0$. Let

$$\hat{\mu} = \sum_{i=1}^{n} w_i X_i$$

be a general linear estimator of μ, where w_i's are (general) weights.

(a)

Show that μ_b is unbiased if and only if $\sum_{i=1}^{n} w_i = 1$

Proof:
Definition of an unbiased estimator:
An estimator $\hat{\mu}$ is unbiased for μ if:

$$\mathbb{E}[\hat{\mu}] = \mu$$

Expected value of $\hat{\mu}$:

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\sum_{i=1}^{n} w_i X_i\right]$$

Using the linearity of expectation:

$$\mathbb{E}[\hat{\mu}] = \sum_{i=1}^{n} w_i \mathbb{E}[X_i]$$

Since X_i are random samples from a population with mean μ, we know:

$$\mathbb{E}[X_i] = \mu$$

Substituting this into the equation:

$$\mathbb{E}[\hat{\mu}] = \sum_{i=1}^{n} w_i \mu$$

$$\mathbb{E}[\hat{\mu}] = \mu \sum_{i=1}^{n} w_i$$

For $\hat{\mu}$ to be unbiased,

$$\mathbb{E}[\hat{\mu}] = \mu$$

. This implies:

$$\mu \sum_{i=1}^{n} w_i = \mu$$

Dividing through by μ (assuming $\mu \neq 0$):

$$\sum_{i=1}^{n} w_i = 1$$

Conclusion:
$\hat{\mu}$ is unbiased if and only if $\sum_{i=1}^{n} w_i = 1$.

(b)

Show that $\hat{\mu}$ is the Best Linear Unbiased Estimator (BLUE) if and only if $w_i \equiv 1/n$ for all i.

Proof:

Recall from part (a) that $\hat{\mu} = \sum_{i=1}^{n} w_i X_i$ is unbiased if and only if $\sum_{i=1}^{n} w_i = 1$.

Variance of the estimator:
Assume the X_i are i.i.d. with variance σ^2:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i^2 \mathrm{Var}(X_i) = \sigma^2 \sum_{i=1}^{n} w_i^2$$

Minimize variance subject to unbiasedness:
We want to minimize $\sum_{i=1}^{n} w_i^2$ subject to $\sum_{i=1}^{n} w_i = 1$.

This is a constrained optimization problem, solved using Lagrange multipliers:

Let

$$L(w_1, \ldots, w_n, \lambda) = \sum_{i=1}^{n} w_i^2 - \lambda \left(\sum_{i=1}^{n} w_i - 1\right)$$

Take partial derivatives and set to zero:

$$\frac{\partial L}{\partial w_i} = 2w_i - \lambda = 0 \implies w_i = \frac{\lambda}{2}$$

$$\sum_{i=1}^{n} w_i = n \cdot \frac{\lambda}{2} = 1 \implies \lambda = \frac{2}{n}$$

Therefore,

$$w_i = \frac{1}{n} \quad \text{for all } i$$

Q2

Let $X = (X_1, \cdots, X_n)'$ be a random sample from a population $X \sim F(\mu, \sigma^2)$ with finite but unknown mean μ > 0 and variance $\sigma^2 > 0$. Define

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Show that $\hat{\sigma}^2$ is a biased estimator of σ^2. Find its bias.

Proof

According to definition:

$$bias(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$$

Recall:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

We want to compute $E[\hat{\sigma}^2]$:

$$E[\hat{\sigma}^2] = E[\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2]$$

$$E[\hat{\sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}\left(E[X_i^2] - 2E[X_i\bar{X}] + E[\bar{X}^2]\right)$$

Since all X_i are i.i.d.:

- $E[X_i^2] = \sigma^2 + \mu^2$
- $E[\bar{X}] = \mu$
- $E[\bar{X}^2] = \mathrm{Var}(\bar{X}) + (E[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2$
- $E[X_i\bar{X}] = E[X_i] \cdot E[\bar{X}] = \mu^2$ for $i \neq j$, but for $i = j$, $E[X_i^2]$

After simplifying the expression, the expectation is:

$$E[\hat{\sigma}^2] = \frac{n+1}{n}\sigma^2$$

Bias:

$$\mathrm{Bias}(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2 = \frac{n+1}{n}\sigma^2 - \sigma^2 = \frac{\sigma^2}{n}$$

Q3

Read the iris data into R and define the Sepal Length of Setosa and Versicolor as variables X and Y respectively. Refer to the following R codes.

```
data('iris')
X <- iris[iris[, 5]=='setosa', 1]
Y <- iris[iris[, 5]=='versicolor', 1]
```

Denote the mean and variance of X and Y as μ_x, μ_y, σ_x^2 and σ_y^2,respectively. Do the following tests at the 5% significance level.

(a) Test the normality of X.

```
shapiro.test(X)
```

I got the result:

```
W = 0.9777, p-value = 0.4595
```

Conclusion

At the 5% significance level, since the p-value (0.4595) is greater than 0.05, we fail to reject the null hypothesis.

Therefore, the Sepal Length of Setosa (X) is consistent with being normally distributed.

(b) Assuming $\sigma_x = \sigma_y$, test $\mu_x = \mu_y$ against $\mu_x \neq \mu_y$.

Hypotheses:

- Null hypothesis (H_0): $\mu_x = \mu_y$
- Alternative hypothesis (H_1): $\mu_x \neq \mu_y$

```
t.test(X, Y, var.equal = TRUE)
```

```
> t.test(X, Y, var.equal = TRUE)

        Two Sample t-test

data:  X and Y
t = -10.521, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1054165 -0.7545835
sample estimates:
mean of x mean of y
    5.006     5.936
```

Conclusion:

At the 5% significance level, the p-value is much less than 0.05, so we reject the null hypothesis.

Therefore, there is strong evidence that the mean Sepal Length of Setosa and Versicolor are different.

(c) Assuming $\sigma_x \neq \sigma_y$, test $\mu_x \leq \mu_y$ against $\mu_x > \mu_y$.

Hypotheses:

- Null hypothesis (H_0): $\mu_x \leq \mu_y$
- Alternative hypothesis (H_1): $\mu_x > \mu_y$

```r
t.test(X, Y, var.equal = FALSE, alternative = "greater")
```

```
> t.test(X, Y, var.equal = FALSE, alternative = "greater")

        Welch Two Sample t-test

data:  X and Y
t = -10.521, df = 86.538, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.07697      Inf
sample estimates:
mean of x mean of y
    5.006     5.936
```

Conclusion:

At the 5% significance level, the p-value is 1, which is much greater than 0.05.
Therefore, we fail to reject the null hypothesis. There is no evidence that the mean Sepal Length of Setosa is greater than that of Versicolor.

(d) Test $\sigma_x^2 \geq \sigma_y^2/2$ against $\sigma_x^2 < \sigma_y^2/2$.

Hypotheses:

- Null hypothesis (H_0): $\sigma_x^2 \geq \sigma_y^2/2$ (i.e., $\sigma_x^2/\sigma_y^2 \geq 0.5$)
- Alternative hypothesis (H_1): $\sigma_x^2 < \sigma_y^2/2$ (i.e., $\sigma_x^2/\sigma_y^2 < 0.5$)

We use an F-test in R:

```r
var.test(X, Y, alternative = "less", ratio = 0.5)
```

```
> var.test(X, Y, alternative = "less", ratio = 0.5)

        F test to compare two variances

data:  X and Y
F = 0.93269, num df = 49, denom df = 49, p-value = 0.4041
alternative hypothesis: true ratio of variances is less than 0.5
95 percent confidence interval:
 0.0000000 0.7495481
sample estimates:
ratio of variances
         0.4663429
```

Conclusion:

At the 5% significance level, the p-value (0.4041) is greater than 0.05, so we fail to reject the null hypothesis.

Therefore, there is no evidence that the variance of Setosa is less than half the variance of Versicolor.