1. Let X be the weight of a randomly chosen individual from a population of adult men. In order to estimate the mean and variance of X, we observe a random sample X_1,X_2,$\cdots$,X_{10}. Thus, the X_i's are i.i.d. and have the same distribution as X. We obtain the following values (in pounds):

   165.5, 175.4, 144.1, 178.5, 168.0, 157.9, 170.1, 202.5, 145.5, 135.7

   Find the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

   Solution: The sample mean is

$$\begin{aligned}\overline{X} &= \frac{X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8+X_9+X_{10}}{10}\\ &= \frac{165.5+175.4+144.1+178.5+168.0+157.9+170.1+202.5+145.5+135.7}{10}\\ &= 164.32\end{aligned}$$

   The sample variance is given by

$$S^2 = \frac{1}{10-1}\sum_{k=1}^{10}(X_k - 164.32)^2 = 383.70,$$

   and the sample standard deviation is given by

$$S = \sqrt{S^2} = 19.59.$$

   You can use the following MATLAB code to compute the above values:

```
x=[165.5, 175.4, 144.1, 178.5, 168.0, 157.9, 170.1, 202.5,
145.5, 135.7];
m=mean(x);
v=var(x);
s=std(x);
```

2. Let X_1, X_2, X_3, ..., X_n be a random sample with unknown mean $EX_i = \mu$, and unknown variance $\text{Var}(X_i) = \sigma^2$. Suppose that we would like to estimate $\theta = \mu^2$. We define the estimator $\hat{\Theta}$ as

$$\hat{\Theta} = \left(\overline{X}\right)^2 = \left[\frac{1}{n}\sum_{k=1}^{n} X_k\right]^2$$

to estimate θ. Is $\hat{\Theta}$ an unbiased estimator of θ? Why?

Solution:

We have

$$\begin{aligned} E[\hat{\Theta}] &= E[\overline{X}^2] \\ &= \text{Var}(\overline{X}) + \big(E[\overline{X}]\big)^2 \\ &= \frac{\sigma^2}{n} + \mu^2. \end{aligned}$$

Therefore, the bias of this is estimator is

$$\begin{aligned} B(\hat{\Theta}) &= E[\hat{\Theta}] - \theta \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Thus, $\hat{\Theta}_n$ is NOT an unbiased estimator of θ

3. Let X_1, X_2, X_3, ..., X_n be a random sample from the following distribution

$$f_X(x) = \begin{cases} \theta\left(x - \frac{1}{2}\right) + 1 & \text{for } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

where $\theta \in [-2, 2]$ is an unknown parameter. We define the estimator $\hat{\Theta}_n$ as

$$\hat{\Theta}_n = 12\overline{X} - 6$$

to estimate θ.

(a) Is $\hat{\Theta}_n$ an unbiased estimator of θ?

(b) Is $\hat{\Theta}_n$ a consistent estimator of θ?

(c) Find the mean squared error (MSE) of $\hat{\Theta}_n$.

Solution: Let's first EX and $\text{Var}(X)$ in terms of θ. We have

$$\begin{aligned} EX &= \int_0^1 x \left[\theta\left(x - \frac{1}{2}\right) + 1\right] \; dx \\ &= \frac{\theta + 6}{12}, \end{aligned}$$

$$\begin{aligned} EX^2 &= \int_0^1 x^2 \left[\theta\left(x - \frac{1}{2}\right) + 1\right] \; dx \\ &= \frac{\theta + 4}{12}, \end{aligned}$$

$$\begin{aligned}\text{Var}(X) &= EX^2 - EX^2 \\ &= \frac{12-\theta^2}{144}.\end{aligned}$$

(a) Is $\hat{\Theta}_n$ an unbiased estimator of θ? To see this, we write

$$\begin{aligned}E[\hat{\Theta}_n] &= E[12\overline{X} - 6] \\ &= 12E[\overline{X}] - 6 \\ &= 12 \cdot \frac{\theta+6}{12} - 6 \\ &= \theta.\end{aligned}$$

Thus, $\hat{\Theta}_n$ IS an unbiased estimator of θ.

(b) To show that $\hat{\Theta}_n$ is a consistent estimator of θ, we need to show

$$\lim_{n\to\infty} P\big(|\hat{\Theta}_n - \theta| \geq \epsilon\big) = 0, \qquad \text{for all } \epsilon > 0.$$

Since $\hat{\Theta}_n = 12\overline{X} - 6$ and $\theta = 12EX - 6$, we conclude

$$\begin{aligned}P\big(|\hat{\Theta}_n - \theta| \geq \epsilon\big) &= P\big(12|\overline{X} - EX| \geq \epsilon\big) \\ &= P\big(|\overline{X} - EX| \geq \frac{\epsilon}{12}\big)\end{aligned}$$

which goes to zero as $n \to \infty$ by the law of large numbers. Therefore, $\hat{\Theta}_n$ is a consistent estimator of θ.

(c) To find the mean squared error (MSE) of $\hat{\Theta}_n$, we write

$$\begin{aligned}MSE(\hat{\Theta}_n) &= \text{Var}(\hat{\Theta}_n) + B(\hat{\Theta}_n)^2 \\ &= \text{Var}(\hat{\Theta}_n) \\ &= \text{Var}(12\overline{X} - 6) \\ &= 144\text{Var}(\overline{X}) \\ &= 144\frac{\text{Var}(X)}{n} \\ &= 144 \cdot \frac{12-\theta^2}{144n} \\ &= \frac{12-\theta^2}{n}.\end{aligned}$$

Note that this gives us another way to argue that $\hat{\Theta}_n$ is a consistent estimator of θ. In particular, since

$$\lim_{n\to\infty} MSE(\hat{\Theta}_n) = 0,$$

we conclude that $\hat{\Theta}_n$ is a consistent estimator of θ.

4. Let $X_1, \ldots, X_4$ be a random sample from a *Geometric*(p) distribution. Suppose we observed $(x_1, x_2, x_3, x_4) = (2, 3, 3, 5)$. Find the likelihood function using $P_{X_i}(x_i; p) = p(1-p)^{x_i - 1}$ as the PMF.

   Solution: The likelihood function is

$$\begin{aligned} L(x_1, \ldots, x_4; p) &= P_{X_1, \ldots, X_4}(x_1, \ldots, x_4; p) \\ &= P_{X_1}(x_1; p) \cdots P_{X_4}(x_4; p) \\ &= p^4 (1-p)^{\sum_{i=1}^{4} x_i - 4} \end{aligned}$$

   Now if we plug in our data,

$$p^4 (1-p)^{(2+3+3+5)-4} = p^4 (1-p)^9.$$

5. Let $X_1, \ldots, X_4$ be a random sample from an *Exponential*(θ) distribution. Suppose we observed $(x_1, x_2, x_3, x_4) = (2.35, 1.55, 3.25, 2.65)$. Find the likelihood function using

$$f_{X_i}(x_i; p) = \theta e^{-\theta x_i}, \qquad \text{for } x_i \geq 0$$

   as the PDF.

   Solution: If $X_i \sim$ *Exponential*(θ), then

$$f_{X_i}(x; \theta) = \theta e^{-\theta x}$$

   Thus, for $x_i \geq 0$, we can write

$$\begin{aligned} L(x_1, x_2, x_3, x_4; \theta) &= f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4; \theta) \\ &= f_{X_1}(x_1; \theta) f_{X_2}(x_2; \theta) f_{X_3}(x_3; \theta) f_{X_4}(x_4; \theta) \\ &= \theta^4 e^{-(x_1 + x_2 + x_3 + x_4)\theta}. \end{aligned}$$

   Since we have observed $(x_1, x_2, x_3, x_4) = (2.35, 1.55, 3.25, 2.65)$, we have

$$L(2.35, 1.55, 3.25, 2.65; \theta) = \theta^4 e^{-9.8\theta}.$$

6. Often when working with maximum likelihood functions, out of ease, we maximize the log-likelihood rather than the likelihood to find the maximum likelihood estimator. Why is maximizing $L(\mathbf{x}; \theta)$ as a function of θ equivalent to maximizing log $L(\mathbf{x}; \theta)$?

   Solution: Since the log function is a monotone increasing function, the estimate, $\hat{\theta}$, that maximizes log $L(\mathbf{x}; \theta)$ also maximizes $L(\mathbf{x}; \theta)$.

7. Let X be one observation from a $N(0, \sigma^2)$ distribution.

(a) Find an unbiased estimator of σ^2.

(b) Find the log likelihood, $\log(L(x;\sigma^2))$, using

$$f_X(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

as the PDF.

(c) Find the Maximum Likelihood Estimate (MLE) for the standard deviation σ, $\hat{\sigma}_{ML}$.

Solution:

(a) Note that

$$E(X^2) = \text{Var}(X) + (EX)^2 = \sigma^2 + \mu^2 = \sigma^2.$$

Therefore x^2 is an unbiased estimator of σ^2.

(b) The likelihood function is

$$L(x;\sigma^2) = f_X(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x)^2}.$$

The log-likelihood function is

$$\ln L(x;\sigma^2) = -\ln(2\pi)^{\frac{1}{2}} - \ln\sigma - \frac{x^2}{2\sigma^2}.$$

(c) To find the MLE for σ we differentiate $\ln L(x;\sigma^2)$ with respect to σ and set it equal to zero.

$$\begin{aligned} \frac{\partial}{\partial\sigma}\ln L &= -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} \\ &= -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} \overset{set}{=} 0 \to \hat{\sigma}X^2 = \hat{\sigma}^3 \to \hat{\sigma} = |X|. \\ \frac{\partial^2}{\partial\sigma^2}\ln L &= \frac{1}{\sigma^2} - \frac{3x^2}{\sigma^4} < 0 \text{ when } \hat{\sigma} = |x|. \end{aligned}$$

8. Let $X_1, \ldots, X_n$ be a random sample from a $Poisson(\lambda)$ distribution.

(a) Find the likelihood equation, $L(x_1, \ldots, x_n; \lambda)$, using

$$P_{X_i}(x_1, \ldots, x_n; \lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

as the PMF.

(b) Find the log likelihood function and use that to obtain the MLE for λ, $\hat{\lambda}_{ML}$.

Solution:

(a) The likelihood function is

$$\begin{aligned} L(x_1, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} P_{X_i}(x_1, \ldots, x_n; \lambda) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$

(b) To obtain the MLE for λ, we first find the log-likelihood function

$$\ln(x_1, \ldots, x_n; \lambda) = -n\lambda + \ln(\lambda)\sum_{i=1}^{n} x_i - \ln(\prod_{i=1}^{n} x_i!).$$

Now we take the derivative of this with respect to λ and set it to 0.

$$\begin{aligned} \frac{d}{d\lambda} \ln L(x_1, \ldots, x_n; \lambda) &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \overset{set}{=} 0 \\ \hat{\lambda} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \text{observed value of } \overline{X}. \end{aligned}$$

Thus, the MLE can be written as

$$\hat{\Lambda} = \overline{X}.$$

Note that this MLE is unbiased, $E\hat{\Lambda} = \lambda$.

9. In this problem, we would like to find the CDFs of the order statistics. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Define $X_{(1)}, \ldots, X_{(n)}$ as the order statistics and show that

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} \binom{n}{k} \big[F_X(x)\big]^k \big[1 - F_X(x)\big]^{n-k}.$$

Hint: Fix $x \in \mathbb{R}$. Let Y be a random variable that counts the number of $X_j's \leq x$. Define $\{X_j \leq x\}$ as a "success" and $\{X_j > x\}$ as a "failure", and show that $Y \sim Binomial(n, p = F_X(x))$.

Solution:

Let Y be a random variable thats counts the number of $X_1, \dots, X_n \leq x$ where x is fixed. Now if we define $\{X_j \leq x\}$ as a “success,” $Y \sim \textit{binomial}(n, F_X(x))$. The event $\{X_{(i)} \leq x\}$ is equivalent to the event $\{Y \geq i\}$, so

$$F_{X_{(i)}}(x) = P(Y \geq i) = \sum_{k=i}^{n} \binom{n}{k} \big[F_X(x)\big]^k \big[1 - F_X(x)\big]^{n-k}.$$

10. In this problem, we would like to find the PDFs of order statistics. Let $X_1, \dots, X_n$ be a random sample from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Define $X_{(1)}, \dots, X_{(n)}$ as the order statistics. Our goal here is to show that

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x) \big[F_X(x)\big]^{i-1} \big[1 - F_X(x)\big]^{n-i}.$$

One way to do this is to differentiate the CDF (found in Problem 9). However, here, we would like to derive the PDF directly. Let $f_{X_{(i)}}(x)$ be the PDF of $X_{(i)}$. By definition of the PDF, for small δ, we can write

$$f_{X_{(i)}}(x)\delta \approx P(x \leq X_{(i)} \leq x + \delta)\delta.$$

Note that the event $\{x \leq X_{(i)} \leq x + \delta\}$ occurs if $i - 1$ of the $X_j's$ are less than x, one of them is in $[x, x + \delta]$, and $n - i$ of them are larger than $x + \delta$. Using this, find $f_{X_{(i)}}(x)$.

Hint: Remember the multinomial distribution. More specifically, suppose that an experiment has 3 possible outcomes, so the sample space is given by

$$S = \{s_1, s_2, s_3\}.$$

Also, suppose that $P(s_i) = p_i$ for $i = 1, 2, 3$. Then for $n = n_1 + n_2 + n_3$ independent trials of this experiment, the probability that each s_i appears n_i times is given by

$$\binom{n}{n_1, n_2, n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} = \frac{n!}{n_1! n_2! n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}.$$

Solution: For each X_j we have three possibilities: $X_j < x$, $X_j \in [x, x + \delta]$, or $X_j > x + \delta$. Note that

$$\begin{aligned}
&P(X_j < x) = F_X(x), \\
&P(x < X_j < x + \delta) \approx f_X(x)\delta, \\
&P(X_j > x + \delta) = F_X(x + \delta) \approx 1 - F_X(x).
\end{aligned}$$

Therefore, we have a multinomial experiment with 3 possible outcomes. The event $\{x \leq X_{(i)} \leq x+\delta\}$ occurs if $i-1$ of the $X_j's$ are less than x, one of them is in $[x, x+\delta]$, and $n-i$ of them are larger than $x+\delta$. Using this, find $f_{X_{(i)}}(x)$. Thus,

$$\begin{aligned} f_{X_{(i)}}(x)\delta &\approx P(x \leq X_{(i)} \leq x+\delta)\delta \\ &= \frac{n!}{(i-1)!1!(n-i)!} F_X(x)^{i-1} f_X(x)\delta[1-F_X(x)]^{n-i}, \end{aligned}$$

Thus,

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x) \big[F_X(x)\big]^{i-1} \big[1-F_X(x)\big]^{n-i}.$$

11. A random sample X_1, X_2, X_3, ..., X_{100} is given from a distribution with known variance $\mathrm{Var}(X_i) = 81$. For the observed sample, the sample mean is $\overline{X} = 50.1$. Find an approximate 95% confidence interval for $\theta = EX_i$.

    Solution: Since n is large, a 95% CI can be expressed as given by

    $$\left[\overline{X} - z_{0.025}\sqrt{\frac{\mathrm{Var}(X_i)}{n}}, \overline{X} + z_{0.025}\sqrt{\frac{\mathrm{Var}(X_i)}{n}}\right].$$

    If we plug in known values, the 95% CI is (48.3, 51.9).

12. To estimate the portion of voters who plan to vote for Candidate A in an election, a random sample of size n from the voters is chosen. The sampling is done with replacement. Let θ be the portion of voters who plan to vote for Candidate A among all voters.

    (a) How large does n need to be so that we can obtain a 90% confidence interval with 3% margin of error?

    (b) How large does n need to be so that we can obtain a 99% confidence interval with 3% margin of error?

    Solution:

    (a) Here,

    $$\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$$

    is a valid $(1-\alpha)100\%$ confidence interval for θ. Therefore, we need to have

    $$\frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} = 0.03$$

Here $\alpha = 0.10$, so $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$. Therefore, we obtain

$$n = \left(\frac{1.645}{2 \times 0.03}\right)^2.$$

We conclude $n \geq 752$ is enough.

(b) Here,

$$\left[\overline{X} - \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}\right]$$

is a valid $(1-\alpha)100\%$ confidence interval for θ. Therefore, we need to have

$$\frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} = 0.03$$

Here $\alpha = 0.01$, so $z_{\frac{\alpha}{2}} = z_{0.005} = 2.576$. Therefore, we obtain

$$n = \left(\frac{2.576}{2 \times 0.03}\right)^2.$$

We conclude $n \geq 1844$ is enough.

13. Let X_1, X_2, X_3, ..., X_{100} be a random sample from a distribution with unknown variance $\text{Var}(X_i) = \sigma^2 < \infty$. For the observed sample, the sample mean is $\overline{X} = 110.5$, and the sample variance is $S^2 = 45.6$. Find a 95% confidence interval for $\theta = EX_i$.

Solution: Since n is relatively large, the interval

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

is approximately a $(1-\alpha)100\%$ confidence interval for θ. Here, $n = 100$, $\alpha = .05$, so we need

$$z_{\frac{\alpha}{2}} = z_{0.025} = \Phi^{-1}(1 - 0.025) = 1.96.$$

Thus, we can obtain a 95% confidence interval for μ as

$$\begin{aligned}\left[\overline{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right] &= \left[110.5 - 1.96 \cdot \frac{\sqrt{45.6}}{10}, 110.5 + 1.96 \cdot \frac{\sqrt{45.6}}{10}\right] \\ &\approx [109.18, 111.82]\end{aligned}$$

Therefore, $[109.18, 111.82]$ is an approximate 95% confidence interval for μ.

14. A random sample X_1, X_2, X_3, ..., X_{36} is given from a normal distribution with unknown mean $\mu = EX_i$ and unknown variance $\mathrm{Var}(X_i) = \sigma^2$. For the observed sample, the sample mean is $\overline{X} = 35.8$, and the sample variance is $S^2 = 12.5$.

    (a) Find and compare 90%, 95%, and 99% confidence interval for μ.

    (b) Find and compare 90%, 95%, and 99% confidence interval for σ^2.

    Solution:

    (a) Here, the interval

$$\left[\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right]$$

is a $(1-\alpha)100\%$ confidence interval for μ.

$$90\%\text{ CI}: \left[35.8 - 1.69\frac{\sqrt{12.5}}{6}, 35.8 + 1.69\frac{\sqrt{12.5}}{6}\right] \approx [34.80, 36.80],$$

$$95\%\text{ CI}: \left[35.8 - 2.03\frac{\sqrt{12.5}}{6}, 35.8 + 2.03\frac{\sqrt{12.5}}{6}\right] \approx [34.60, 37.00],$$

$$99\%\text{ CI}: \left[35.8 - 2.72\frac{\sqrt{12.5}}{6}, 35.8 + 2.72\frac{\sqrt{12.5}}{6}\right] \approx [34.20, 37.40].$$

    (b) Here, the CI is given by

$$90\%: \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = \left[\frac{35 \times 12.5}{49.80}, \frac{35 \times 12.5}{22.47}\right] \approx [8.79, 19.47].$$

$$95\%: \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = \left[\frac{35 \times 12.5}{53.20}, \frac{35 \times 12.5}{20.57}\right] \approx [8.22, 21.27].$$

$$99\%: \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = \left[\frac{35 \times 12.5}{60.27}, \frac{35 \times 12.5}{17.19}\right] \approx [7.26, 25.45].$$

15. Let X_1, X_2, X_3, X_4, X_5 be a random sample from a $N(\mu, 1)$ distribution, where μ is unknown. Suppose that we have observed the following values

$$5.45, \quad 4.23, \quad 7.22, \quad 6.94, \quad 5.98$$

    We would like to decide between

H_0: $\mu = \mu_0 = 5$,

H_1: $\mu \neq 5$.

(a) Define a test statistic to test the hypotheses and draw a conclusion assuming $\alpha = 0.05$.

(b) Find a 95% confidence interval around $\overline{X}$. Is μ_0 included in the interval? How does the exclusion of μ_0 in the interval relate to the hypotheses we are testing?

Solution:

(a) Here we define the test statistic as

$$\begin{aligned} W &= \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{5.96 - 5}{1/\sqrt{5}} \\ &\approx 2.15. \end{aligned}$$

Here, $\alpha = .05$, so $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. Since $|W| > z_{\frac{\alpha}{2}}$, we reject H_0 and accept H_1.

(b) The 95% CI is give by

$$\left(5.96 - 1.96 * \frac{1}{\sqrt{(5)}}, 5.96 + 1.96 * \frac{1}{\sqrt{(5)}}\right) = (5.09, 6.84).$$

Since μ_0 is not included in the interval, we are able to reject the null hypothesis and conclude that μ is not 5.

16. Let $X_1, \ldots, X_9$ be a random sample from a $N(\mu, 1)$ distribution, where μ is unknown. Suppose that we have observed the following values

16.34, 18.57, 18.22, 16.94, 15.98, 15.23, 17.22, 16.54, 17.54

We would like to decide between

H_0: $\mu = \mu_0 = 16$,

H_1: $\mu \neq 16$.

(a) Find a 90% confidence interval around $\overline{X}$. Is μ_0 included in the interval? How does this relate to our hypothesis test?

(b) Define a test statistic to test the hypotheses and draw a conclusion assuming $\alpha = 0.1$.

Solution:

(a) The 90% CI is give by

$$(16.95 - 1.645 * \frac{1}{\sqrt{(9)}}, 16.95 + 1.645 * \frac{1}{\sqrt{(9)}}) = (16.40, 17.50).$$

Since μ_0 is NOT included in the interval, we reject the null hypothesis.

(b) Here we define the test statistic as

$$\begin{aligned} W &= \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{16.95 - 16}{1/\sqrt{9}} \\ &\approx 2.85. \end{aligned}$$

Here, $\alpha = .1$, so $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$. Since $|W| > z_{\frac{\alpha}{2}}$, we reject H_0.

17. Let X_1, X_2 ,..., X_{150} be a random sample from an unknown distribution. After observing this sample, the sample mean and the sample variance are calculated to be

$$\overline{X} = 52.28, \qquad S^2 = 30.9$$

Design a level 0.05 test to choose between

H_0: $\mu = 50$,

H_1: $\mu > 50$.

Do you accept or reject H_0?

Solution:

$$\begin{aligned} W &= \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \\ &= \frac{52.28 - 50}{\sqrt{30.9/150}} \\ &= 5.03 \end{aligned}$$

Since $5.03 > 1.96$, we reject H_0.

18. Let X_1, X_2, X_3, X_4, X_5 be a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ are both unknown. Suppose that we have observed the following values

$$27.72, \qquad 22.24, \qquad 32.86, \qquad 19.66, \qquad 35.34$$

We would like to decide between

H_0: $\mu \geq 30$,

H_1: $\mu < 30$.

Assuming $\alpha = 0.05$, what do you conclude?

Solution:

$$\begin{aligned} W &= \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \\ &= \frac{27.6 - 30}{6.70/\sqrt{5}} \\ &= -0.801 \end{aligned}$$

Here, $\alpha = 0.05$ and $n = 5$. Since $W \geq -t_{\alpha,n-1} = -2.13$, we accept H_0.

19. Let X_1, X_2 ,..., X_{121} be a random sample from an unknown distribution. After observing this sample, the sample mean and the sample variance are calculated to be

$$\overline{X} = 29.25, \qquad S^2 = 20.7$$

Design a test decide between

H_0: $\mu = 30$,

H_1: $\mu < 30$,

and calculate the P-value for the observed data.

Solution: We define the test statistic as

$$\begin{aligned} W &= \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \\ &= \frac{29.25 - 30}{\sqrt{20.7}/\sqrt{121}} \\ &= -1.81 \end{aligned}$$

The P-value is P(type I error) when the test threshold c is chosen to be $c = -1.81$.

$$-z_\alpha = 1.81$$

Noting that by definition $z_\alpha = \Phi^{-1}(1 - \alpha)$, we obtain P(type I error) as

$$\alpha = 1 - \Phi(1.81) \approx 0.035$$

Therefore,

$$P - \text{value} = 0.035$$

20. Suppose we would like to test the hypothesis that at least 10% of students suffer from allergies. We collect a random sample of 225 students and 21 of them suffer from allergies.

    (a) State the null and alternative hypotheses.

    (b) Obtain a test statistic and a P-value.

    (c) State the conclusion at the $\alpha = 0.05$ level.

    Solution: Let θ be the portion of students suffering from allergies.

    (a) H_0: $\theta \geq \theta_0 = 0.1$,
    H_1: $\theta < 0.1$,

    Note that as it is described in the text, we can consider the following hypotheses instead:

    H_0: $\theta = \theta_0 = 0.1$,
    H_1: $\theta < 0.1$,

    (b) Let X be the random variable showing the number of observed heads, so in our experiment we observed $X = 21$. Since $n = 225$ is relatively large, assuming H_0 is true, the random variable

    $$W = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = \frac{X - 22.5}{4.5}$$

    is (approximately) a standard normal random variable, $N(0,1)$. Thus, we can suggest the following test: We choose a threshold c. If $W \geq c$, we accept H_0; otherwise, we accept H_1. To calculate P(type I error), we can write

    $$\begin{aligned} P(\text{type I error}) &= P(\text{Reject } H_0 \mid H_0) \\ &= P(W < c \mid H_0). \end{aligned}$$

    The observed value of W is

    $$w_1 = \frac{21 - 22.5}{4.5} = -0.3333$$

    Thus,

    $$\begin{aligned} P - \text{value} &= P(\text{type I error when } c = -0.333) \\ &= P(W < -.3333) \\ &= \Phi(-.3333) \approx 0.37 \end{aligned}$$

    (c) Since our $P - value = 0.37 > 0.05 = \alpha$, we accept H_0.

21. Consider the following observed values of (x_i, y_i):

$$(-5,-2), \qquad (-3,1), \qquad (0,4), \qquad (2,6), \qquad (1,3).$$

(a) Find the estimated regression line

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$$

based on the observed data.

(b) For each x_i, compute the fitted value of y_i using

$$\hat{y}_i = \hat{\beta_0} + \hat{\beta_1} x_i.$$

(c) Compute the residuals, $e_i = y_i - \hat{y}_i$.

(d) Calculate R-squared.

Solution:

(a) We have

$$\begin{aligned}
\overline{x} &= \frac{-5-3+0+2+1}{5} = -1 \\
\overline{y} &= \frac{-2+1+4+6+3}{5} = 2.4 \\
s_{xx} &= (-5+1)^2 + (-3+1)^2 + (0+1)^2 + (2+1)^2 + (1+1)^2 = 34 \\
s_{xy} &= (-5+1)(-2-2.4) + (-3+1)(1-2.4) + (0+1)(4-2.4) \\
&\quad + (2+1)(6-2.4) + (1+1)(3-2.4) = 34.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\hat{\beta_1} &= \frac{s_{xy}}{s_{xx}} = \frac{34}{34} = 1 \\
\hat{\beta_0} &= 2.4 - (1)(-1) = 3.4.
\end{aligned}$$

(b) The fitted values are given by

$$\hat{y}_i = 3.4 + 1x_i,$$

so we obtain

$$\hat{y}_1 = -1.6, \qquad \hat{y}_2 = 0.4, \qquad \hat{y}_3 = 3.4, \qquad \hat{y}_4 = 5.4, \qquad \hat{y}_4 = 4.4.$$

(c) We have

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 = -2 + 1.6 = -0.4, \\ e_2 &= y_2 - \hat{y}_2 = 1 - 0.4 = 0.6, \\ e_3 &= y_3 - \hat{y}_3 = 4 - 3.4 = 0.6, \\ e_4 &= y_4 - \hat{y}_4 = 6 - 5.4 = 0.6 \\ e_4 &= y_4 - \hat{y}_4 = 3 - 4.4 = -1.4. \end{aligned}$$

(d) We have

$$s_{yy} = (-2 - 2.4)^2 + (1 - 2.4)^2 + (4 - 2.4)^2 + (6 - 2.4)^2 + (3 - 2.4)^2 = 37.2.$$

We conclude

$$r^2 = \frac{(34)^2}{34 \times 37.2} \approx 0.914.$$

22. Consider the following observed values of (x_i, y_i):

$$(1, 3), \qquad (3, 7).$$

(a) Find the estimated regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

based on the observed data.

(b) For each x_i, compute the fitted value of y_i using

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

(c) Compute the residuals, $e_i = y_i - \hat{y}_i$.

(d) Calculate R-squared.

(e) Explain the above results. In particular, can you conclude that the obtained regression line is a good model here?

Solution:

(a) We have

$$\begin{aligned} \overline{x} &= \frac{1+3}{2} = 2 \\ \overline{y} &= \frac{3+7}{2} = 5 \\ s_{xx} &= (1-2)^2 + (3-2)^2 = 2 \\ s_{xy} &= (1-2)(3-5) + (3-2)(7-5) = 4. \end{aligned}$$

Therefore, we obtain

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{4}{2} = 2$$
$$\hat{\beta}_0 = 5 - (2)(2) = 1.$$

(b) The fitted values are given by

$$\hat{y}_i = 1 + 2x_i,$$

so we obtain

$$\hat{y}_1 = 3, \qquad \hat{y}_2 = 7.$$

(c) We have

$$e_1 = y_1 - \hat{y}_1 = 3 - 3 = 0,$$
$$e_2 = y_2 - \hat{y}_2 = 7 - 7 = 0.$$

(d) We have

$$s_{yy} = (3-5)^2 + (7-5)^2 = 8.$$

We conclude

$$r^2 = \frac{4^2}{2 \times 8} = 1$$

(e) Here, the residuals are zero and R-squared is equal to 1. However, this is due to the fact that we have only two pairs of data. In other words, the regression line is simply the line that passes through the two given points. Based on the available information, you CANNOT conclude that the obtained regression line is a good model or not. All we can say is that we need to observe more data points.

23. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i's are independent $N(0, \sigma^2)$ random variables. Therefore, Y_i is a normal random variable with mean $\beta_0 + \beta_1 x_i$ and variance σ^2. Moreover, Y_i's are independent. As usual, we have the observed data pairs (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) from which we would like to estimate β_0 and β_1. In this chapter, we found the following estimators

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}},$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}.$$

where

$$\begin{aligned} s_{xx} &= \sum_{i=1}^{n} (x_i - \overline{x})^2, \\ s_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}). \end{aligned}$$

(a) Show that $\hat{\beta}_1$ is a normal random variable.

(b) Show that $\hat{\beta}_1$ is an unbiased estimator of β_1, i.e.,

$$E[\hat{\beta}_1] = \beta_1.$$

(c) Show that

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}.$$

Solution:

(a) Note that

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{s_{xx}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})Y_i}{s_{xx}} - \frac{\overline{Y} \sum_{i=1}^{n} (x_i - \overline{x})}{s_{xx}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})Y_i}{s_{xx}}. \end{aligned}$$

Thus, $\hat{\beta}_1$ can be written as a linear combination of Y_i's, i.e.,

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i.$$

Since the Y_i's are normal and independent, we conclude that $\hat{\beta}_1$ is a normal random variable.

(b) Note that

$$\begin{aligned} Y_i - \overline{Y} &= (\beta_0 + \beta_1 x_i + \epsilon_i) - (\beta_0 + \beta_1 \overline{x} + \bar{\epsilon}) \\ &= \beta_1 (x_i - \overline{x}) + (\epsilon_i - \bar{\epsilon}). \end{aligned}$$

Therefore,

$$\begin{aligned} E[Y_i - \overline{Y}] &= \beta_1(x_i - \overline{x}) + E[\epsilon_i - \bar{\epsilon}] \\ &= \beta_1(x_i - \overline{x}). \end{aligned}$$

Thus,

$$\begin{aligned} E[\hat{\beta}_1] &= \frac{\sum_{i=1}^n (x_i - \overline{x}) E[Y_i - \overline{Y}]}{s_{xx}} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x}) \beta_1 (x_i - \overline{x})}{s_{xx}} \\ &= \beta_1. \end{aligned}$$

(c) We have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{s_{xx}},$$

where the Y_i's are independent, so

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \text{Var}(Y_i)}{s_{xx}^2} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \sigma^2}{s_{xx}^2} \\ &= \frac{\sigma^2}{s_{xx}}. \end{aligned}$$

24. Again consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i's are independent $N(0, \sigma^2)$ random variables, and

$$\begin{aligned} \hat{\beta}_1 &= \frac{s_{xy}}{s_{xx}}, \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{x}. \end{aligned}$$

(a) Show that $\hat{\beta}_0$ is a normal random variable.

(b) Show that $\hat{\beta}_0$ is an unbiased estimator of β_0, i.e.,

$$E[\hat{\beta}_0] = \beta_0.$$

(c) For any $i = 1, 2, 3, ..., n$, show that

$$\text{Cov}(\hat{\beta}_1, Y_i) = \frac{x_i - \overline{x}}{s_{xx}} \sigma^2.$$

(d) Show that

$$\mathrm{Cov}(\hat{\beta}_1, \overline{Y}) = 0.$$

(e) Show that

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{ns_{xx}}\sigma^2.$$

Solution:

(a) Since both $\overline{Y}$ and $\hat{\beta}_1$ can be written as linear combinations of the Y_i's, and $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$, we conclude that $\hat{\beta}_0$ is also a linear combination of the Y_i's so it is a normal random variable.

(b) Note that

$$\overline{Y} = \beta_0 + \beta_1\overline{x} + \overline{\epsilon}.$$

Thus

$$\begin{aligned} E[\hat{\beta}_0] &= E[\overline{Y} - \hat{\beta}_1\overline{x}] \\ &= \beta_0 + \beta_1\overline{x} + E[\overline{\epsilon}] - E[\hat{\beta}_1]\overline{x} \\ &= \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} \\ &= \beta_0. \end{aligned}$$

(c) For any $i = 1, 2, 3, ..., n$, we have

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}_1, Y_i) &= \mathrm{Cov}\left(\frac{\sum_{k=1}^n (x_k - \overline{x})Y_k}{s_{xx}}, Y_i\right) \\ &= \frac{\sum_{k=1}^n (x_k - \overline{x})}{s_{xx}}\mathrm{Cov}(Y_k, Y_i) \\ &= \frac{x_i - \overline{x}}{s_{xx}}\sigma^2. \end{aligned}$$

(d) We have

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}_1, \overline{Y}) &= \mathrm{Cov}\left(\hat{\beta}_1, \frac{1}{n}\sum_{k=1}^n Y_i\right) \\ &= \frac{1}{n}\sum_{k=1}^n \mathrm{Cov}(\hat{\beta}_1, Y_i) \\ &= \frac{1}{n}\sum_{k=1}^n \frac{x_i - \overline{x}}{s_{xx}}\sigma^2 \\ &= 0. \end{aligned}$$

(e) We have

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}.$$

Thus,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_0) &= \mathrm{Var}(\overline{Y}) + \overline{x}^2 \mathrm{Var}(\hat{\beta}_1) - 2\overline{x}\mathrm{Cov}(\hat{\beta}_1, \overline{Y}) \\
&= \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{s_{xx}} \\
&= \left[\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}\right] \sigma^2 \\
& \frac{\sum_{i=1}^n x_i^2}{n s_{xx}} \sigma^2.
\end{aligned}$$