ASSIGNMENT 2 :-

PROBLEM 3.6:

Consider the partial datasets extracted from body fat prediction dataset on kaggle.

Only body fat column is extracted. The dataset is partitioned into two sets, Xla and Xny . Suppose that first 100 instances belong to LA and the remaining 152 instances belong to NYC. Let the subscripts ‘la’ and ‘ny’ indicate two cities. Two datasets are exclusive:

n = nla +nny and X = Xla ∪Xny .

Following parameters for each city are already computed:

X_{la}
12.3
6.1
25.3
10.4
⋮
11.3
17.8
22.2

X_{ny}
21.2
20.4
20.1
22.3
⋮
33.6
29.3
26.0
31.9

$n_{la} = 100$
$\mu(X_{la}) = 18.191$
$\sigma^2(X_{la}) = 80.763$
$\eta_3(X_{la}) = 0.2055$
$\eta_4(X_{la}) = 2.0279$

$n_{ny} = 152$
$\mu(X_{ny}) = 19.7822$
$\sigma^2(X_{ny}) = 61.513$
$\eta_3(X_{ny}) = 0.1641$
$\eta_4(X_{ny}) = 3.2783$

Find the parameter values in real time without accessing the original dataset.

a). Find the mean value, μ(X).

-> $\mu(X) = \dfrac{na\mu(Xa) + nb\mu(Xb)}{na + nb}$

$\mu(X) = \dfrac{100 \times 18.191 + 152 \times 19.7822}{252}$

$= 19.1508$

b) Find the variance value, σ2(X).

$$\sigma^2(X) = \frac{na(\sigma^2(Xa) + \mu(Xa)^2) + nb(\sigma^2(Xb) + \mu(Xb)^2)}{na + nb} - \mu(X)^2$$

σ^2(X) =

$$\frac{100 \times (80.763 + 18.191^2) + 152 \times (61.513 + 19.7822^2)}{252} - 19.151^2$$

= 69.7579

c). Find $E(X^2)$.

$$E(X^2) = \sigma^2(X) + \mu^2(X)$$

= 69.7579 + 19.1508²
= 436.51
≈ 436.5108

d). Find the Pearson's moment coefficient of skewness, $\eta_3(X)$.

$$\eta_p(X) = \frac{M_p(X)}{\sigma p(X)}$$

$M_3(X_{Ia}) = \eta_3(X_{Ia}) (\sqrt{\sigma^2(X_{Ia})})^3$

$M_3(X_{Ia}) = 0.2055 \times (\sqrt{80.763})^3$

= 149.1525 ≈ 149.1882

$M_3(_{n\gamma}) = 0.1641 \times (\sqrt{61.513})^3$

= 79.1697 ≈ 79.1709

$$M3(X) = \frac{n_{Ia}(M_3(X_{Ia}) - B1(X_{Ia})) + n_{n\gamma}(M3(X_{n\gamma}) - B1(X_{n\gamma}))}{n_{Ia} + n_{n\gamma}} + B_1(X)$$

$B1(X) = -\mu3(X) - 3\mu(X)\sigma^2(X)$

$B1(X) = -19.15083 - 3 \times 19.1508 \times 69.7579 \approx -11031$
$B1(Xla) = -18.1913 - 3 \times 18.191 \times 80.763 \approx -10427$
$B1(Xny) = -19.78223 - 3 \times 19.7822 \times 61.513 \approx -11392$

$$M3(X) = \frac{(100 \times (149.1525 + 10427) + 152 \times (79.1697 + 11392))}{(100 + 152 - 11031)}$$

$= 85.0041 \approx 84.7608$

$$\eta 3(X) = \frac{85.0041}{\sqrt{69.7579^3}}$$

$= 0.1459 \approx 0.1455$

e) Find $E(X^3)$.
$E(X^3) = M_3(X) + \mu_3(X) + 3\mu(X)\sigma^2(X)$
$E(X^3) = 85.0041 + 19.15083 + 3 \times 19.1508 \times 69.7579$
$E(X^3) = 11116$

f) Find the fourth standardized moment coefficient of kurtosis, $\eta_4(X)$.

$$\eta_p(X) = \frac{M_p(X)}{\sigma_p(X)}$$

$M_4(Xla) = \eta_4(Xla) \times \sigma^4(Xla)$
$M_4(Xla) = 2.0279 \times 80.763^2 = 13227 \approx 13228$

$M_4(Xny) = 3.2783 \times 61.513^2 = 12405 \approx 12404$

$$M_4(X) = \frac{(nla \times (M_4(Xla) - 4B_1(Xla)) + nny \times (M_4(Xny) - 4B_1(Xny)))}{(nla + nny) + 4B_1(X)} \quad \text{(Equation 3.68)}$$

$pBk(X) = \sum$ (p choose k) $E(X^{(p-i)}) (-\mu(X))^i$ where $0 \le k \le p$

$4B_1(X) = -4E(X^3)\mu(X) + 6E(X^2)\mu^2(X) - 3\mu^4(X)$

$4B_1(X) = -4 \times 11116 \times 19.1508 + 6 \times 436.51 \times 19.1508^2 - 3 \times 19.1508^4 = -294500$

$E(X^3_{Ia}) = 79.1697 + 18.1913 + 3 \times 18.191 \times 80.763 = 10576$

$E(X^2_{Ia}) = \sigma^2(XIa) + \mu^2(XIa) = 80.763 + 18.191^2$
$= 411.6755$

$4B_1(XIa) = -4 \times 10576 \times 18.191 + 6 \times 411.6755 \times 18.191^2 - 3 \times 18.191^4$
$= -280690$

$E(X^3_{n\gamma}) = 79.1709 + 19.78223 + 3 \times 19.7822 \times 61.513 = 11471$

$E(X^2_{n\gamma}) = \sigma^2(X_{n\gamma}) + \mu^2(X_{n\gamma}) = 61.513 + 19.7822^2 = 452.8484 \approx 452.8497$

$4B_1(X_{n\gamma}) = -4 \times 11471 \times 19.7822 + 6 \times 452.8484 \times 19.7822^2 - 3 \times 19.7822^4 = -303820$

$$M_4(X) = \frac{(100 \times (13227 + 280690) + 152 \times (12405 + 303820))}{(100 + 152 - 294500)}$$

$= 12873 \approx 12891$

$\eta_4(X) = 12873 / 69.7579^2$
$= 2.6454 \approx 2.6491$

g) Find $E(X^4)$.

$E(X^4) = M_4(X) - 4B_1(X)$
$E(X^4) = 12873 + 303820$
$E(X^4) = 316693$

h) Validate your answers in a) ~ g) using the full dataset and discuss the differences.

Parameter	Real-Time Value	By Definition
$\mu(X)$	19.1508	19.1508
$\sigma^2(X)$	69.7579	69.7579
$E(X^2)$	436.51	436.5108
$M_3(X)$	85.0041	84.7608
$\eta_3(X)$	0.1459	0.1455
$E(X^3)$	11116	11116
$M_4(X)$	12873	12891
$\eta_4(X)$	2.6454	2.6491
$E(X^4)$	316693	?

For questions j) ~ m), suppose that a person ($x_{la,100}$ = 22.2) in LA moved to NYC.

j) Find the following mean values:
$\mu(X_{la} - \{22.2\})$ and $\mu(X_{n\gamma} \cup \{22.2\})$.

Let $X_{n\gamma+} = X_{n\gamma} \cup \{22.2\}$ and $X_{n\gamma-} = X_{n\gamma} - \{22.2\}$ for simplicity.

$$\mu(X_{1\sim n}) = \frac{((n-1)\mu(X_{1\sim n-1}) + x_n)}{n}$$

$$\mu(X_{n\gamma+}) = \frac{(152 \times 19.7822 + 22.2)}{153}$$

$$\mu(X_{n\gamma+}) = 19.798$$

$$\mu(X_{1\sim n-1}) = (n\mu(X_{1\sim n}) - x_n) / (n-1)$$

$$\mu(X_{la-}) = (100 \times 19.1508 - 22.2) / 99$$
$$\mu(X_{la-}) = 18.150$$

k) Find the following variance values:

$\sigma^2(X_{Ia} - \{22.2\})$ and $\sigma^2(X_{n\gamma} \cup \{22.2\})$.

$$\sigma^2(X_{1\sim n}) = \frac{((n-1)(\sigma^2(X_{1\sim n-1}) + \mu^2(X_{1\sim n-1})) + x_n^2)}{n} - \mu^2(X_{1\sim n})$$

$$\sigma^2(X_{n\gamma+}) = \frac{(152 \times (61.513 + 19.7822^2) + 22.2^2)}{153} - 19.798^2$$

$$\sigma^2(X_{n\gamma+}) = 61.149 \approx 61.1487$$

$$\sigma^2(X_{1\sim n-1}) = (n(\sigma^2(X_{1\sim n}) + \mu^2(X_{1\sim n})) - x_n^2) / (n-1) - \mu^2(X_{1\sim n-1})$$

$$\sigma^2(X_{Ia-}) = (100 \times (80.763 + 18.191^2) - 22.2^2) / 99 - 18.1505^2$$
$$\sigma^2(X_{Ia-}) = 81.41$$

l) Find the following Pearson's moment coefficients of skewness:
$\eta_3(X_{Ia} - \{22.2\})$ and $\eta_3(X_{n\gamma} \cup \{22.2\})$.

To find $\eta_3(X_{n\gamma} \cup \{22.2\})$:

$$M_3(X_{1\sim n}) = \frac{((n-1)(M_3(X_{1\sim n-1}) - B_1(X_{1\sim n-1})) + x_n^3)}{n + B_1(X_{1\sim n})}$$

$$M_3(X_{n\gamma+}) = \frac{(152(M_3(X_{n\gamma}) - B_1(X_{n\gamma})) + 22.2^3)}{153 + B_1(X_{n\gamma+})}$$

$$B_1(X_{n\gamma+}) = -\mu_3(X_{n\gamma+}) - 3\mu(X_{n\gamma+})\sigma^2(X_{n\gamma+})$$
$$= -19.798^3 - 3 \times 19.798 \times 61.149$$
$$= -11392$$

$$M_3(X_{n\gamma+}) = \frac{(152 \times (79.1697 + 11392) + 22.2^3)}{153 - 11392}$$
$$= 75.705$$

$\eta_3(X_{n\gamma+}) = M_3(X_{n\gamma+}) / \sigma^3(X_{n\gamma+})$
$= 75.705 / \sqrt{(61.149^3)}$
$= 0.1583 \approx 0.1586$

To find $\eta_3(X_{Ia} - \{22.2\})$:

$M_3(X_{1\sim n-1}) =$
$\underline{(n(M_3(X_{1\sim n}) - B_1(X_{1\sim n})) - x_n^3)}$
$(n - 1) + B_1(X_{1\sim n-1})$

$M_3(X_{Ia-}) = (100(M_3(X_{Ia}) - B_1(X_{Ia})) - 22.2^3) / 99 + B_1(X_{Ia-})$

$B_1(X_{Ia-}) = -\mu_3(X_{Ia-}) - 3\mu(X_{Ia-})\sigma^2(X_{Ia-})$
$= -18.1505^3 - 3 \times 18.1505 \times 81.415$
$= -10413$

$M_3(X_{Ia-}) = (100 \times (149.1525 + 10427) - 22.2^3) / 99 - 10413$
$= 159.47$

$\eta_3(X_{Ia-}) = \underline{M_3(X_{Ia-})}$
$\sigma^3(X_{Ia-})$
$= 159.47 / \sqrt{(81.415^3)}$
$= 0.2171 \approx 0.2177$

m) Find the following fourth standardized moment coefficients of kurtosis:
$\eta_4(X_{Ia} - \{22.2\})$ and $\eta_4(X_{n\gamma} \cup \{22.2\})$.

To find $\eta_4(X_{n\gamma+})$,

$M_p(X_{1\sim n}) = ((n - 1)(M_p(X_{1\sim n-1}) - pB_1(X_{1\sim n-1})) + x_n^p) / n + pB_1(X_{1\sim n})$
(Equation 3.67)

$M_4(X_{n\gamma+}) = (152 \times (M_4(X_{n\gamma}) - 4B_1(X_{n\gamma})) + 22.2^4) / 153 + 4B_1(X_{n\gamma+})$

$4B_1(X_{nγ+}) = -4E(X^3_{nγ+})μ(X) + 6E(X^2_{nγ+})μ^2(X_{nγ+}) - 3μ^4(X_{nγ+})$

$E(X^3_{nγ+}) = M_3(X_{nγ+}) + μ^3(X_{nγ+}) + 3μ(X_{nγ+})σ^2(X_{nγ+})$
$= 75.705 + 19.798^3 + 3 \times 19.798 \times 61.149$
≈ 11467.63

$E(X^2_{nγ+}) = σ^2(X_{nγ+}) + μ^2(X_{nγ+})$
$= 61.149 + 19.798^2$
$= 453.11$

n) Validate your answers in j) ~ m) using the full dataset and discuss the differences.

➔ For NYC dataset ($n_{nγ} = 152$):

$μ(X_{nγ}) = 19.7822$
$σ^2(X_{nγ}) = 61.513$
$η_3(X_{nγ}) = 0.1641$
$η_4(X_{nγ}) = 3.2783$
$M_3(X_{nγ}) = 79.1697$
$M_4(X_{nγ}) = 12405$
$4B_1(X_{nγ}) = -303820$

➔ For updated dataset after adding 22.2 (n = 252)

$μ(X \cup \{22.2\}) = 19.798$
$σ^2(X \cup \{22.2\}) = 61.149$
$η_3(X \cup \{22.2\}) = 0.1583$
$η_4(X \cup \{22.2\}) = 2.0064$
$M_3(X \cup \{22.2\}) = 75.705$
$M_4(X \cup \{22.2\}) = 7502.16$
$4B_1(X \cup \{22.2\}) = -303436.57$

➔ For LA dataset ($n_{la} = 100$)

$\mu(X_{la}) = 18.191$

$\sigma^2(X_{la}) = 80.763$

$\eta_3(X_{la}) = 0.2055$

$\eta_4(X_{la}) = 2.0279$

$M_3(X_{la}) = 149.1525$

$M_4(X_{la}) = 13227$

$4B_1(X_{la}) = -280690$

➔ For the full dataset (n = 252)

$\mu(X) = 19.1508$

$\sigma^2(X) = 69.7579$

$\eta_3(X) = 0.1459$

$\eta_4(X) = -6.0038$

$M_3(X) = 85.0041$

$M_4(X) = -22449.37$

$4B_1(X) = -303436.57$

PROBLEM 3.7:

In order to investigate the inheritance of traits in pea plants, Sir Francis Galton conducted a scientific experiment. Galton's pea data table, which is originally from [6, p.226] contains a list of frequencies of daughter seeds of various sizes according to the size of their parent seeds. Parent pea seeds of various diameter, 15 ~ 21. The unit of the diameter of seed is 0.01 inch. Each parent seed has 100 filial seeds and their diameter frequency distribution is given in the following table.

Parent seeds \ Filial seed	< 15	15	16	17	18	19	20	21
15	46	14	9	11	14	4	2	0
16	34	15	18	16	13	3	1	0
17	37	16	13	16	13	4	1	0
18	34	12	13	17	16	6	2	0
19	35	16	12	13	11	10	2	1
20	23	10	12	17	20	13	3	2
21	22	8	10	18	21	13	6	2

Let X15 be the set of daughter seeds who parent seed's diameter is 15. Let Xx,15∼18 is a subset of daughter seeds whose diameters range from 15 to 18 and their parent seed diameter is x. Let Xx,<15 is a subset of daughter seeds whose diameter is less than 15 and they are worthless as a commodity because they are too small and their exact diameter value is unknown. Let F15,17 = 11 denote the frequency of case where parent and daughter seed diameters are 15 and 17. Let F15,15∼18 = ⟨14, 9, 11, 14⟩ be the subset frequency.

A) Find the median values of each X_x for x ∈ {15, · · · , 21}.

$_w\mu_i(X_{15}, P_{15}) = 15$
because $\sum^{15}_{i=-\infty} P(X_{15}, i) =$
$\frac{46 + 14}{100} = 0.6 \geq 0.5$

And $\sum^{21}_{i=15} =$
$\frac{14 + 9 + 11 + 14 + 4 + 2 + 0}{100} = 0.54 \geq 0.5$

Parent seeds diameter x	15	16	17	18	19	20	21
$_w\mu_i(X_x, P_x)$	15	16	16	16	15	17	17

b). Find the median values of each $X_{x,15\sim21}$ for $x \in \{15, \cdots, 21\}$.
For example, $\mu_i(X_{15,15\sim21}) = 17$
$_w\mu_i(X_{15,15\sim21}, P_{15,15\sim21}) = 17$
because $\sum^{18}_{i=15} P(X_{15,i}) =$

$$\frac{14 + 9 + 11}{54} = 0.62963 \geq 0.5$$

$\text{AND} \sum^{21}_{i=18} P(X_{15,i}) =$

$$\frac{14 + 9 + 11}{54} = 0.54 \geq 0.5$$

$_w\mu_i(X_{16,15\sim21}, P_{16,15\sim21}) = 16.5$
because $\sum^{16}_{i=15} P(X_{16,i}) =$

$$\frac{15 + 18}{66} = 0.5$$

$\text{And} \sum^{21}_{i=17} P(X_{16,i}) =$

$$\frac{16 + 13 + 3 + 1 + 0}{66} = 0.5 = 0.5$$

Parent seeds diameter x	15	16	17	18	19	20	21
$_w\mu_i(X_{x,15\sim21}, P_{x,15\sim21})$	17	16.5	17	17	17	17	18

c) Find the mean values of each $X_{x,15\sim21}$ for $x \in \{15, \cdots, 21\}$.

$$\mu_w(X_{15,15\sim21}, P_{15,15\sim21}) = \frac{14 \times 15 + 9 \times 16 + \cdots + 0 \times 21}{54}$$

$$= 16.833$$

Parent seeds diameter x	16	17	18	19	20	21
$\mu_w(X_{x,15\sim21}, P_{x,15\sim21})$	16.606	16.667	16.955	16.954	17.403	17.603

d). Find the variance values of each $X_{x,15\sim21}$ for $x \in \{15, \cdots, 21\}$.

$$\sigma^2_w(X_{15,15\sim21}, P_{15,15\sim21}) = \frac{14 \times (15 - 16.833)^2 + 9 \times (16 - 16.833)^2 + \cdots + 0 \times (21 - 16.833)^2}{54}$$

$$= 2.0648$$

Parent seeds diameter x	16	17	18	19	20	21
$\mu_w(X_{x,15\sim21}, P_{x,15\sim21})$	1.5418	1.7143	1.801	2.4748	2.2145	2.2138

e) Find the MAD around median values of each $X_{x,15\sim21}$ for $x \in \{15, \cdots, 21\}$.

$$W\delta_i(X_{15,15\sim21}, P_{15,15\sim21}) = \frac{14 \times |15 - 17| + 9 \times |16 - 17| + \cdots + 0 \times |21 - 17|}{54}$$

$$= 1.2037$$

Parent seeds diameter x	16	17	18	19	20	21
$\mu_w(X_{x,15\sim21}, P_{x,15\sim21})$	1.0606	1.0952	1.0758	1.3077	1.2338	1.1923

f) Find the Groeneveld & Meeden's coefficient values of each $X_{x,15\sim21}$ for $x \in \{15, \cdots, 21\}$.

For example, $\kappa_{mi}\delta_i\ (X_{15,15\sim21}) = -12.408$.

$\kappa_{mi}\delta_i\ (X_{15,15\sim21}, P_{15,15\sim21}) =$

$$\frac{\mu w(X_{15,15\sim21}, P_{15,15\sim21}) - w\mu i(X_{15,15\sim21}, P_{15,15\sim21})}{w\delta i(X_{15,15\sim21}, P_{15,15\sim21})}$$

$$= \frac{16.833 - 17}{1.2037}$$

$$= -0.13874$$

Parent seeds diameter x	16	17	18	19	20	21
$\mu_w(X_{x,15\sim21}, P_{x,15\sim21})$	0.2	−0.3044	−0.0423	-0.0353	0.3263	-0.3333

g) Suppose $\mu(X_{20}) = 16.33$ and $\mu(X_{21}) = 16.5$. Find the mean value, $\mu(X_{20} \cup X_{21})$?

$$\mu(X_{20} \cup X_{21}) = \frac{100 \times \mu(X_{20}) + 100 \times \mu(X_{21})}{100 + 100}$$

$$= \frac{1633 + 1650}{200}$$

$$= 16.415$$

h) Suppose $\mu(X_{20}) = 16.33$, $\mu(X_{21}) = 16.5$, $\sigma^2(X_{20}) = 5.8811$, and $\sigma^2(X_{21}) = 6.41$. Find the variance value, $\sigma^2(X_{20} \cup X_{21})$?

$$\sigma^2(X_{20} \cup X_{21}) = \frac{100 \times (\sigma^2(X_{20}) + \mu(X_{20})^2) + 100 \times (\sigma^2(X_{21}) + \mu(X_{21})^2)}{100 + 100} - \mu(X_{21} \cup X_{21})^2$$

$$= \frac{100 \times (5.8811 + 16.332) + 100 \times (6.41 + 16.52)}{100 + 100} - 16.4152$$

$$= 6.1528$$

I) Suppose $\mu(X_{15}) = 14.9$ and $\mu(X15,15\sim21) = 16.8333$. Find the mean value, $\mu(X15,<15)$?

$\mu(X15) = \mu(X15,<15) \cup \mu(X15,15\sim21)$,

$$\mu(X15) = \frac{46 \times \mu(X15,<15) + 54 \times \mu(X15,15\sim21)}{100}$$

$$14.9 = \frac{46 \times \mu(X15,<15) + 54 \times 16.8333}{100}$$

$$\mu(X15,<15) = \frac{14.9 \times 100 - 54 \times 16.8333}{46}$$

$$= 12.63$$

j) Suppose μ(X15) = 14.9, μ(X15,15~21) = 16.8333, σ²(X15) = 6.23, and σ2(X15,15~21) = 2.0648. Find the variance value, σ²(X15,<15)?

$$\sigma^2(X15) + \mu(X15)^2 = \frac{46(\sigma^2(X15,<15) + \mu(X15,<15)2) + 54(\sigma^2(X15,15\sim21) + \mu(X15,15\sim21)^2)}{100}$$

$$6.23 + 14.92 = \frac{46 \times (\sigma2(X15,<15) + 12.632) + 54 \times (2.0648 + 16.83332)}{100}$$

$$\sigma^2(X_{15}, <_{15}) = \frac{100 \times (6.23 + 14.92) - 54 \times (2.0648 + 16.8333^2)}{46} - 12.63^2$$

= 1.59