Statistics for Data Science - 2

Week 7 Graded assignment

Use the following values of standard normal distribution if needed.

$F_Z(0.71) = 0.76115, F_Z(1.26) = 0.89617, F_Z(1.58) = 0.94295, F_Z(2.58) = 0.99506, F_Z(1.96) = 0.975$

1. Let X_1, X_2, X_3 are three independent and identically distributed random variables with mean μ and variance σ^2. Given below are 3 different formulations of sample mean. (Observe that $E[A] = E[B] = E[C]$).

$$A = \frac{X_1 + X_2 + X_3}{3}$$
$$B = 0.1X_1 + 0.3X_2 + 0.6X_3$$
$$C = 0.2X_1 + 0.3X_2 + 0.5X_3$$

Choose the correct option from the following:

(a) $\text{Var}(A) = \text{Var}(B) = \text{Var}(C)$

(b) $\text{Var}(A) \geq \text{Var}(B) \geq \text{Var}(C)$

(c) $\text{Var}(A) \leq \text{Var}(B) \leq \text{Var}(C)$

(d) $\text{Var}(A) \leq \text{Var}(C) \leq \text{Var}(B)$

Solution:

Let $X_1, X_2, X_3 \sim$ i.i.d.X, where $E[X] = \mu, \text{Var}(X) = \sigma^2$

$$\begin{aligned}\text{Var}(A) &= \text{Var}\left(\frac{X_1 + X_2 + X_3}{3}\right)\\ &= \frac{1}{9}\left(\text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3]\right)\\ &= \frac{1}{9}(3\sigma^2) = \frac{\sigma^2}{3}\end{aligned}$$

$$\begin{aligned}\text{Var}(B) &= \text{Var}\left(0.1X_1 + 0.3X_2 + 0.6X_3\right)\\ &= 0.01\text{Var}[X_1] + 0.09\text{Var}[X_2] + 0.36\text{Var}[X_3]\\ &= 0.46(3\sigma^2)\\ &= 1.38\sigma^2\end{aligned}$$

$$
\begin{aligned}
\text{Var}(C) =& \text{Var}\left(0.2X_1 + 0.3X_2 + 0.5X_3\right)\\
=&0.04\text{Var}[X_1] + 0.09\text{Var}[X_2] + 0.25\text{Var}[X_3]\\
=&0.38(3\sigma^2)\\
=&1.14\sigma^2
\end{aligned}
$$

Therefore, $\text{Var}(B) \geq \text{Var}(C) \geq \text{Var}(A)$.

---

2. A random sample of size 25 is collected from a normal population with mean of 50 and standard deviation of 5. Find the variance of the sample mean.

Solution:
We know that variance of the sample mean $\overline{X}$ is given by

$$
\begin{aligned}
\text{Var}[\overline{X}] &= \frac{\sigma^2}{n}\\
&= \frac{5^2}{25} = 1
\end{aligned}
$$

---

3. Let $X_1, X_2, \ldots, X_{50} \sim$ i.i.d. Poisson(0.04) and let $Y = \sum_{i=1}^{50} X_i$. Use Central Limit theorem to find $P(Y > 3)$. Enter the answer correct to 2 decimal places.

Solution:

Let $X \sim \text{Poisson}(0.04)$.
Consider the samples $X_1, X_2, \ldots, X_{50}$ from X.
$E[X] = \text{Var}[X] = 0.04$
$E[Y] = E\left[\sum_{i=1}^{50} X_i\right] = 50 \times 0.04 = 2$, $\text{Var}[Y] = \text{Var}\left[\sum_{i=1}^{50} X_i\right] = 50 \times 0.04 = 2$
To find: $P(Y > 3)$.
By CLT, we know that

$$
\begin{aligned}
\frac{Y - n\mu}{\sigma\sqrt{n}} &\sim \text{Normal}(0, 1)\\
\Rightarrow \left(\frac{Y - 2}{\sqrt{2}}\right) &\sim \text{Normal}(0, 1)
\end{aligned}
$$

Now,

$$
\begin{aligned}
P(Y > 3) &= P(Y - 2 > 1)\\
&= P\left(\frac{Y - 2}{\sqrt{2}} > \frac{3 - 2}{\sqrt{2}}\right)\\
&= P(Z > 0.707)\\
&= 1 - F_Z(0.707) = 1 - 0.76 = 0.24
\end{aligned}
$$

4. Let the moment generating function of a random variable X be given by

$$M_X(\lambda) = \left(\frac{1}{4}\right) e^{-2\lambda} + \left(\frac{1}{40}\right) + \left(\frac{3}{10}\right) e^{-\lambda} + \left(\frac{3}{40}\right) e^{2\lambda} + \left(\frac{7}{20}\right) e^{\lambda}$$

Find the distribution of X.

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{3}{40}$	$\frac{3}{10}$	$\frac{1}{40}$	$\frac{7}{20}$

(a)

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{40}$	$\frac{3}{10}$	$\frac{3}{40}$	$\frac{7}{20}$

(b)

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{3}{10}$	$\frac{1}{40}$	$\frac{7}{20}$	$\frac{3}{40}$

(c)

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{3}{40}$	$\frac{1}{40}$	$\frac{7}{20}$	$\frac{3}{10}$

(d)

Solution:

The MGF of a discrete random variable X with the PMF $f_X(x) = P(X=x)$, $x \in T_X$ is given by

$$\begin{aligned} M_X(\lambda) &= E[e^{\lambda X}] \\ &= \sum_{x \in T_X} P(X=x)e^{\lambda x} \end{aligned}$$

Now, MGF of a random variable X is given as

$$M_X(\lambda) = \left(\frac{1}{4}\right) e^{-2\lambda} + \left(\frac{1}{40}\right) + \left(\frac{3}{10}\right) e^{-\lambda} + \left(\frac{3}{40}\right) e^{2\lambda} + \left(\frac{7}{20}\right) e^{\lambda}$$

Therefore, distribution of X is given by

X	-2	-1	0	1	2
$P(X=x)$	$\frac{1}{4}$	$\frac{3}{10}$	$\frac{1}{40}$	$\frac{7}{20}$	$\frac{3}{40}$

5. A fair coin is tossed 1000 times. Use CLT to compute the probability that head appears at most 520 times. Enter the answer correct to 3 decimal places.

Solution:

Define a random variable X such that

$$X = \begin{cases} 1 & \text{if head appears on tossing a fair coin} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $E[X] = \mu = \frac{1}{2}$ and
$\text{Var}(X) = \sigma^2 = \frac{1}{2}.\frac{1}{2} = \frac{1}{4}$

Let $X_1, X_2, \ldots, X_{1000}$ be outcomes on tossing the fair coin 1000 times.
Notice that $X_1 + X_2 + \ldots + X_{1000}$ will denote the number of times head appears in 1000 tosses.

Let $S = X_1 + X_2 + \ldots + X_{1000}$

To find: $P(S \leq 520)$

By CLT, we know that

$$\frac{S - 1000\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0,1)$$
$$\Rightarrow \frac{S - 500}{5\sqrt{10}} \sim \text{Normal}(0,1)$$

Now,

$$\begin{aligned} P(S \leq 520) &= P(S - 500 \leq 20) \\ &= P\left(\frac{S-500}{5\sqrt{10}} \leq \frac{20}{5\sqrt{10}}\right) \\ &= P(Z \leq 1.26) \\ &= 0.896 \end{aligned}$$

---

6. A fair die is rolled 100 times. Let X denote the number of times six is obtained. Find a bound for the probability that $\frac{X}{100}$ differs from $\frac{1}{6}$ by less than 0.1 using weak law of large numbers.

(a) at least $\frac{5}{36}$

(b) at least $\frac{31}{36}$

(c) at most $\frac{5}{36}$

(d) at most $\frac{31}{36}$

Solution:

X denotes the number of times six is obtained on rolling a fair die 100 times.
Let $X_1, X_2, \ldots, X_{100}$ be 100 i.i.d. samples such that

$$X_i = \begin{cases} 1 & \text{if six appears on rolling a fair die} \\ 0 & \text{otherwise} \end{cases}$$

$E[X_i] = \mu = \frac{1}{6}$ and
$\text{Var}(X_i) = \sigma^2 = \frac{5}{36}$
Notice that $X = X_1 + X_2 + X_3 + \ldots + X_{100}$

To find: Bound on $P\left(\left|\frac{X}{100} - \frac{1}{6}\right| < 0.1\right)$.

By weak law of large numbers, we have

$$\begin{aligned} &P(|\overline{X} - \mu| < \delta) \geq 1 - \frac{\sigma^2}{n\delta^2} \\ \Rightarrow &P\left(\left|\frac{X}{100} - \frac{1}{6}\right| < 0.1\right) \geq 1 - \frac{5}{36 \times 100 \times 0.01} \\ \Rightarrow &P\left(\left|\frac{X}{100} - \frac{1}{6}\right| < 0.1\right) \geq 1 - \frac{5}{36} = \frac{31}{36} \end{aligned}$$

---

7. Let $X_1, X_2, \ldots, X_{500} \sim$ i.i.d Normal$(0, 1)$. Evaluate $P(X_1^2 + X_2^2 + \ldots + X_{500}^2 > 550)$ using Central Limit theorem. Enter the answer correct to 2 decimal places.
Hint: $(X_1^2 + X_2^2 + \ldots + X_{500}^2) \sim$ Gamma $(250, 0.5)$.

Solution:

Given $X_1, \ldots, X_{500} \sim$ i.i.d. Normal$(0, 1)$.

We know that if $X \sim \text{Normal}(0, 1) \implies X^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$

Also, Sum of n independent Gamma(α, β) is Gamma$(n\alpha, \beta)$.

Therefore, $X_i^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, for all i.

and $(X_1^2 + X_2^2 + \ldots + X_{500}^2) \sim$ Gamma $(250, 0.5)$

Let $Y = Y_1 + Y_2 + \ldots + Y_{500}$, where $Y_i = X_i^2$ for all $i : 1 \to 500$

$E[Y_i] = \dfrac{0.5}{0.5} = 1$ and $\text{Var}[Y_i] = \dfrac{0.5}{0.25} = 2$, for $i : 1 \to 500$

$E[Y] = \dfrac{250}{0.5} = 500$ and $\text{Var}[Y] = \dfrac{250}{0.5^2} = 1000$

To find: $P(Y > 550)$

By CLT, we know that

$$\frac{Y - 500\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0, 1)$$
$$\Rightarrow \frac{Y - 500}{10\sqrt{10}} \sim \text{Normal}(0, 1)$$

Now,

$$\begin{aligned} P(Y > 550) &= P(Y - 500 > 50) \\ &= P\left(\frac{Y - 550}{10\sqrt{10}} > \frac{5}{\sqrt{10}}\right) \\ &= P(Z > 1.58) \\ &= 1 - F_Z(1.58) = 1 - 0.94 = 0.06 \end{aligned}$$

Use the below information to answer questions 8 and 9.

Let X be a random variable having the gamma distribution with the parameters $\alpha = 2n$ and $\beta = 1$.

Hint:

- If $X \sim \text{Gamma}(\alpha, \beta), E[X] = \dfrac{\alpha}{\beta}$ and $\text{Var}[X] = \dfrac{\alpha}{\beta^2}$
- Sum of n independent $\text{Gamma}(\alpha, \beta)$ is $\text{Gamma}(n\alpha, \beta)$

8. Use the Weak Law of Large number to find the value of n such that

$$P\left(\left|\frac{X}{2n} - 1\right| > 0.01\right) < 0.01$$

(a) 505000

(b) 470000

(c) 498000

(d) 482000

Solution:

Given $X \sim \text{Gamma}(2n, 1)$
Let $X = X_1 + X_2 + X_3 + \ldots + X_{2n}$, where $X_i \sim \text{Gamma}(1, 1)$.

$E[X] = \mu = 1$ and $\text{Var}(X) = \sigma^2 = 1$
$E[\bar{X}] = 1$ and $\text{Var}[\bar{X}] = \dfrac{1}{2n}$

To find: The value of n such that $P\left(\left|\dfrac{X}{2n} - 1\right| > 0.01\right) < 0.01$.

By weak law of large numbers, we have

$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$
$$\Rightarrow P\left(\left|\frac{X}{2n} - 1\right| > 0.01\right) \leq \frac{1}{2n \times 0.01^2}$$

Therefore, $\dfrac{1}{2n \times 0.01^2} < 0.01 \implies 2n > \dfrac{1}{0.01^3} \implies n > 500000$.

---

9. Use CLT to find the value of n such that

$$P\left(\left|\frac{X}{2n} - 1\right| > 0.01\right) < 0.01$$

Hint: Use $F_Z(2.58) = 0.995, F_Z(1.96) = 0.975$ if needed.

(a) 34570
(b) 33500
(c) 32500
(d) 30000

Solution:

$E[X_1 + \ldots + X_{2n}] = 2n$ and $\text{Var}[X_1 + \ldots + X_{2n}] = 2n$

To find: The value of n such that $P\left(\left|\dfrac{X}{2n} - 1\right| > 0.01\right) < 0.01$.

By CLT, we know that

$$\frac{X - 2n\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0, 1)$$

$$\Rightarrow \frac{X - 2n}{\sqrt{2n}} \sim \text{Normal}(0, 1)$$

Now,

$$\begin{aligned}
P\left(\left|\frac{X}{2n} - 1\right| > 0.01\right) &< 0.01 \\
\implies P\left(\left|\frac{X_1 + \ldots + X_n}{2n} - 1\right| > 0.01\right) &< 0.01 \\
\implies P\left(\left|\frac{X_1 + \ldots + X_n - 2n}{\sqrt{2n}}\right| > 0.01\sqrt{2n}\right) &< 0.01 \\
\implies P(\mid Z \mid > 0.01\sqrt{2n}) &< 0.01 \\
\implies 2P(Z > 0.01\sqrt{2n}) &< 0.01 \\
\implies 1 - F_Z(0.01\sqrt{2n}) &< \frac{0.01}{2} \\
\implies F_Z(0.01\sqrt{2n}) &> 0.995 \\
\implies F_Z(0.01\sqrt{2n}) &> F_Z(2.58) \\
\implies n &> 33282
\end{aligned}$$

---

10. Let the time taken (in hours) for failure of an electric bulb follow the exponential distribution with the parameter 0.05. Suppose that 100 such light bulbs say $L_1, L_2, \ldots, L_{100}$ are used in the following manner: For every i, as soon as the light L_i fails, L_{i+1} becomes operative, where $i : 1 \to 99$ (i.e. If L_1 fails, L_2 becomes operative, if L_2 fails, L_3 becomes operative, and so on). Let the total time of operation of 100 bulbs be denoted by T. Using CLT, compute the probability that T exceeds 2500 hours.

    (a) $F_Z(1.5)$
    (b) $1 - F_Z(1.5)$
    (c) $F_Z(2.5)$
    (d) $1 - F_Z(2.5)$

    Solution:

    Given, time to failure (in hours) of an electric bulb has the exponential distribution with the parameter $\lambda = 0.05$.
    Since, the bulbs are used in such a way, that as soon as light L_1 fails, L_2 becomes operative, L_2 fails, L_3 becomes operative, and so on.

    We know that if $X \sim \text{Gamma}(\alpha, \beta)$ with parameter $\alpha = 1$, then $X \sim \text{Exp}(\beta)$.
    Also, sum of n i.i.d. $\text{Exp}(\lambda)$ is $\text{Gamma}(n, \lambda)$.

Since each of the L_i's are exponentially distributed with parameter $= 0.05$, therefore

$$L_1 + \ldots + L_{100} \sim \text{Gamma}(n\alpha, \beta) = \text{Gamma}(100, 0.05)$$

Let $T = L_1 + \ldots + L_{100}$

$E[L_i] = \mu = \dfrac{1}{0.05} = 20$ and $\text{SD}[L_i] = \sigma = \dfrac{1}{0.05} = 20$

To find: $P(T \geq 2500)$

By CLT, we know that

$$\frac{T - 100\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0, 1)$$

$$\Rightarrow \frac{T - 2000}{20\sqrt{100}} \sim \text{Normal}(0, 1)$$

Now,

$$\begin{aligned} P(T \geq 2500) &= P(T - 2000 \geq 500) \\ &= P\left(\frac{T - 2000}{200} \geq \frac{500}{200}\right) \\ &= P(Z \geq 2.5) \\ &= 1 - F_Z(2.5) \end{aligned}$$

---

11. Suppose speeds of vehicles on a particular road are normally distributed with mean 36 mph and standard deviation 2 mph. Find the probability that the mean speed $\overline{X}$ of 20 randomly selected vehicles is between 35 and 38 mph.

(a) $F_Z(\sqrt{5}) - F_Z(-\sqrt{5})$
(b) $F_Z(\sqrt{20}) - F_Z(-\sqrt{20})$
(c) $F_Z(\sqrt{38}) - F_Z(-\sqrt{35})$
(d) $F_Z(\sqrt{20}) - F_Z(-\sqrt{5})$

Solution:

Let X denote the speed of a vehicle on a particular road.
Given that $X \sim \text{Normal}(36, 2^2)$.
Therefore, $\mu = 36$ and $\sigma = 2$
Select $X_1, X_2, \ldots X_{20}$ samples such that $X_1, X_2, \ldots X_{20} \sim$ iid X

Let $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_{20}}{20}$ and $S = X_1 + X_2 + \ldots + X_{20}$

To find: $P(35 < \overline{X} < 38)$ From CLT, we know that

$$
\begin{aligned}
&\frac{X_1 + X_2 \ldots + X_n - nE[X]}{\sqrt{n}\sigma} \sim \text{Normal}(0,1) \\
\Rightarrow &\frac{S - n\mu}{\sqrt{n}\sigma} \sim \text{Normal}(0,1) \\
\Rightarrow &\frac{(S - 36(20))}{(2\sqrt{20})} \sim \text{Normal}(0,1)
\end{aligned}
$$

Now,

$$
\begin{aligned}
P(35 < \overline{X} < 38) &= P(35 < \frac{S}{20} < 38) \\
&= P(-1 < \frac{S}{20} - 36 < 2) \\
&= P(-1 < \frac{S - 36(20)}{20} < 2) \\
&= P(\frac{-\sqrt{20}}{2} < \frac{S - 36(20)}{2\sqrt{20}} < \sqrt{20}) \\
&= P(-\sqrt{5} < \frac{S - 36(20)}{2\sqrt{20}} < \sqrt{20}) \\
&= F_Z(\sqrt{20}) - F_Z(-\sqrt{5})
\end{aligned}
$$