Statistics for Data Science - 2

Week 7 practice Assignment

Statistics from samples and Limit theorems

1. If $X, Y \sim$ i.i.d. Normal$(0, 4)$, what will be the variance of $\frac{X}{Y}$?

   (a) 4

   (b) 2

   (c) 1

   (d) Undefined

   Solution:
   We know that if $X, Y \sim$ i.i.d. Normal$(0, \sigma^2)$, $\frac{X}{Y} \sim$ Cauchy$(0, 1)$ and variance of Cauchy distribution is undefined.
   Therefore, option(d) is correct.

---

2. A population has mean 60 and standard deviation 6. Random samples of size 100 from this population are collected independently. Find the expected value of the sample mean.
   Solution:
   We know that expected value of the sample mean $\overline{X}$ is given by

$$\begin{aligned} E[\overline{X}] &= \mu \\ &= 60 \end{aligned}$$

---

3. Let X_1, X_2, X_3, X_4 and $X_5 \sim$ i.i.d. Normal$(2, 25)$. Calculate $P(2X_1 + X_2 + 3X_3 + X_4 + X_5 \geq 10)$.

   1. $F_Z(0.3)$
   2. $1 - F_Z(0.3)$
   3. $F_Z(-0.3)$
   4. $1 - F_Z(-0.3)$

   Solution:

   We know that linear combination of independent Normal distributions is again a normal distribution.

Hence, $2X_1 + X_2 + 3X_3 + X_4 + X_5$ will follow a Normal distribution.
Let $Y = 2X_1 + X_2 + 3X_3 + X_4 + X_5$
$E[Y] = E[2X_1 + X_2 + 3X_3 + X_4 + X_5] = (2 + 1 + 3 + 1 + 1)E[X] = 16$
$\text{Var}(Y) = \text{Var}(2X_1 + X_2 + 3X_3 + X_4 + X_5) = (4 + 1 + 9 + 1 + 1)\text{Var}(X) = 400$

It implies that $Y \sim \text{Normal}(16, 20^2)$.
To find: $P(Y \geq 10)$

Now,

$$\begin{aligned} P(Y \geq 10) &= P(Y - 16 \geq -6) \\ &= P(\frac{Y - 16}{20} \geq \frac{-6}{20}) \\ &= P(\frac{Y - 16}{20} \geq -0.3) \\ &= P(Z \geq -0.3) \\ &= 1 - P(Z < -0.3) \\ &= 1 - F_Z(-0.3) \end{aligned}$$

---

4. Random samples of size 100 are collected from a population of unknown parameters. If the variance of the sample mean is 36, what will be the standard deviation of the actual population?
Solution:
We know that variance of the sample mean is given by $\frac{\sigma^2}{n}$ where σ is the standard deviation of the actual population and n is the sample size.

By the given information, we have

$$\begin{aligned} &\frac{\sigma^2}{n} = 36 \\ \Rightarrow &\frac{\sigma^2}{100} = 36 \\ \Rightarrow &\sigma^2 = 3600 \\ \Rightarrow &\sigma = 60 \end{aligned}$$

Therefore, standard deviation of the actual population is 60.

---

5. A random sample of size 50 is collected from a population with a standard deviation of 5. Find the upper bound on the probability that the sample mean will be at least 10 away from the actual mean using the weak law of large numbers. Write your answer correct to three decimal places.

Solution:
Given: standard deviation of the population, $\sigma = 5$
Sample size, $n = 50$

To find: upper bound on $P(|\overline{X} - \mu| \geq 10)$ where $\overline{X}$ and μ are sample mean and population mean, respectively.

Now, by weak law of large number, we have

$$\begin{aligned} P(|\overline{X} - \mu| \geq \delta) &\leq \frac{\sigma^2}{n\delta^2} \\ \Rightarrow P(|\overline{X} - \mu| \geq 10) &\leq \frac{25}{100 \times 50} \\ \Rightarrow P(|\overline{X} - \mu| \geq 10) &\leq 0.005 \end{aligned}$$

---

6. A study shows that the average daily sleeping hours of teenagers is ten hours with a standard deviation of two hours. If a sample of 100 teenagers is collected, what will be the probability that the mean of the sleeping hours of these 100 teenagers is at least 0.4 hours away from the population mean? Assume that each observation in the sample is independent. Assume that F_Z denotes the CDF of standard normal distribution.

   (a) $1 + F_Z(-2) - F_Z(2)$
   (b) $1 - F_Z(-2) + F_Z(2)$
   (c) $F_Z(2) - F_Z(-2)$
   (d) $F_Z(2)$

Solution:
let X denote the average daily sleeping hours of teenagers.
Given: standard deviation of X, $\sigma = 2$
Sample size, $n = 100$

To find: $P(|\overline{X} - \mu| \geq 0.4)$ where $\overline{X}$ and μ are sample mean and population mean, respectively.

Let $S = X_1 + X_2 + \ldots X_{100}$ where X_i denotes the ith sample.
By CLT, we know that $\frac{S - n\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0, 1) \Rightarrow \frac{S - 100\mu}{20} \sim Z$ (Standard Normal)

Now,

$$\begin{aligned}
P(|\overline{X} - \mu| \geq 0.4) &= P(\left|\frac{S}{n} - \mu\right| \geq 0.4) \\
&= P(\left|\frac{S - n\mu}{n}\right| \geq 0.4) \\
&= P(\left|\frac{S - n\mu}{\sigma\sqrt{n}}\right| \geq \frac{0.4\sqrt{n}}{\sigma}) \\
&= P(|Z| \geq 2) \\
&= P(Z \geq 2) + P(Z \leq -2) \\
&= 1 - P(Z \leq 2) + P(Z \leq -2) \\
&= 1 - F_Z(2) + F_Z(-2)
\end{aligned}$$

---

7. What is the fourth moment of the Normal$(0, 4)$ distribution?
Solution:

$$\begin{aligned}
M_X(\lambda) = E[e^{\lambda X}] &= E[1 + \lambda X + \frac{\lambda^2 X^2}{2!} + \frac{\lambda^3 X^3}{3!} + ...] \\
&= 1 + \lambda E[X] + \frac{\lambda^2 E[X^2]}{2!} + \frac{\lambda^3 E[X^3]}{3!} + ...
\end{aligned}$$

In the moment generating function, coefficient of λ will give first moment ($E[X]$), coefficient of $\frac{\lambda^2}{2!}$ will give the second moment ($E[X^2]$) and similarly, coefficient of $\frac{\lambda^k}{k!}$ will give the kth moment ($E[X^k]$).

Moment generating function of Normal$(0, \sigma^2)$ is given by $e^{\lambda^2\sigma^2/2}$.
Let $N \sim$ Normal$(0, 2^2)$

$$\begin{aligned}
M_N(\lambda) &= e^{\lambda^2 2^2/2} \\
&= 1 + \frac{\lambda^2 2^2}{2} + \frac{\lambda^4 2^4}{2!(4)} + \dots \\
&= 1 + \frac{\lambda^2 2^2}{2} + 48\frac{\lambda^2}{4!} + \dots
\end{aligned}$$

Therefore, 4th moment of Normal$(0, 2^2)$ = coefficient of $\frac{\lambda^4}{4!} = 48$

---

8. Let $X \sim$ Gamma$(2, \frac{1}{2})$ and $Y \sim$ Gamma$(5, \frac{1}{2})$ be two independent random variables. What will be the expected value of $\frac{X}{X+Y}$? Write your answer correct to two decimal

places.

Solution:

We know that if $X \sim \text{Gamma}(\alpha, k)$ and $Y \sim \text{Gamma}(\beta, k)$ be two independent random variables, then $\dfrac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$.

Given that $X \sim \text{Gamma}(2, \frac{1}{2})$ and $Y \sim \text{Gamma}(5, \frac{1}{2})$ are two independent random variables. It implies that

$$\frac{X}{X+Y} \sim \text{Beta}(2, 5)$$

Therefore, $E\left[\dfrac{X}{X+Y}\right] = \dfrac{2}{2+5} = 0.28$

---

9. A study says that the delivery time of pizzas has a standard deviation of 10 minutes. A pizza shop collected the data of some deliveries and their delivery time. The probability that the mean delivery time of this sample is at least $\sqrt{5}$ minutes away from the actual mean delivery time is at most $\frac{1}{5}$ as per the weak law of large numbers. What is the size of the sample?

   Solution:

   Let X denote the delivery time of pizzas.

   Given that $\sigma = 10$

   To find: size of the sample such that $P(|\overline{X} - \mu| \geq \sqrt{5}) \leq \frac{1}{5}$ ...(1).

   By the weak law of large numbers, we have

$$P(|\overline{X} - \mu| \geq \delta) \leq \frac{\sigma^2}{n\delta^2}$$
$$\Rightarrow P(|\overline{X} - \mu| \geq \sqrt{5}) \leq \frac{100}{n \times 5} \quad \text{...(1)}$$

   By equation (1) and (2), we have

$$\frac{1}{5} = \frac{100}{5n}$$
$$\Rightarrow n = 100$$

---

10. A company sells eggs whose weights are normally distributed with a mean of 70g and a standard deviation of 2g. Suppose that these eggs are sold in packages that each contain four eggs. Assume that the weight of each egg is independent. What is the probability that the mean weight of the four eggs in a package is greater than 68.5g? Write your answer correct to two decimal places.

    (Hint: Use the fact that linear combination of normal distributions is again a normal distribution. $F_Z(-1.5) = 0.066$)

Solution:
Let X denote the weight of an egg.
Given that $E[X] = \mu = 70$
$\text{SD}(X) = \sigma = 2$
$X \sim \text{Normal}(70, 2^2)$ Let X_1, X_2, X_3 and X_4 denote the weights of four eggs in a package.

Suppose that
$\overline{X} = \dfrac{X_1 + X_2 + X_3 + X_4}{4}$

To find: $P(\overline{X} > 68.5)$

We know that linear combination of independent Normal distribution is again a Normal distribution.
It implies that $\overline{X}$ is a Normal distribution.

$E[\overline{X}] = \mu = 70$ and
$\text{Var}(\overline{X}) = \dfrac{\sigma^2}{n} = \dfrac{4}{4} = 1$

It implies that $\overline{X} \sim \text{Normal}(70, 1) \Rightarrow \overline{X} - 70 \sim \text{Normal}(0, 1)$

Now,

$$\begin{aligned} P(\overline{X} > 68.5) &= P(\overline{X} - 70 > -1.5) \\ &= P(Z > -1.5) \\ &= 1 - F_Z(-1.5) \\ &= 1 - 0.066 = 0.93 \end{aligned}$$

---

11. Let $X_1, X_2, X_3, \ldots X_n$ be i.i.d. Poisson(4). What should be the value of n such that $P(3.8 \leq \overline{X} \leq 4.2) \geq 0.95$? [2 marks]
(Hint: Use $F_Z(1.96) = 0.975$)

    1. at least 200
    2. at least 385
    3. at least 450
    4. at least 585

Solution:
Given that $X_1, X_2, X_3, \ldots X_n \sim$ i.i.d. Poisson(4)

Mean of the distribution $= \mu = 4$
Variance of the distribution $= \sigma^2 = 4$
Let $S = X_1 + X_2 + \ldots + X_n$ and
$\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$

To find: value of n such that $P(3.8 \leq \overline{X} \leq 4.2) \geq 0.95$
By CLT, we know that
$\dfrac{S - n\mu}{\sqrt{n}\sigma} \sim \text{Normal}(0, 1)$
$\Rightarrow \dfrac{S - 4n}{2\sqrt{n}} \sim \text{Normal}(0, 1)$ ...(1)

$$
\begin{aligned}
& P(3.8 \leq \overline{X} \leq 4.2) \geq 0.95 \\
\Rightarrow & P(3.8 \leq \frac{S}{n} \leq 4.2) \geq 0.95 \\
\Rightarrow & P(-0.2 \leq \frac{S}{n} - 4 \leq 0.2) \geq 0.95 \\
\Rightarrow & P(-0.2 \leq \frac{S - 4n}{n} \leq 0.2) \geq 0.95 \\
\Rightarrow & P(-0.1 \leq \frac{S - 4n}{2n} \leq 0.1) \geq 0.95 \\
\Rightarrow & P(-0.1\sqrt{n} \leq \frac{S - 2n}{2\sqrt{n}} \leq 0.1\sqrt{n}) \geq 0.95 \\
\Rightarrow & F_Z(0.1\sqrt{n}) - F_Z(-0.1\sqrt{n}) \geq 0.95 \\
\Rightarrow & F_Z(0.1\sqrt{n}) - (1 - F_Z(0.1\sqrt{n})) \geq 0.95 \\
\Rightarrow & 2F_Z(0.1\sqrt{n}) - 1 \geq 0.95 \\
\Rightarrow & F_z(0.1\sqrt{n}) \geq 0.975 \\
\Rightarrow & 0.1\sqrt{n} \geq 1.96 \\
\Rightarrow & n \geq 384.16
\end{aligned}
$$

---

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

$$M_X(\lambda) = \left(\frac{1}{8}\right) e^{-4\lambda} + \left(\frac{1}{6}\right) e^{-2\lambda} + \left(\frac{1}{6}\right) e^{2\lambda} + \left(\frac{1}{8}\right) e^{4\lambda} + \left(\frac{5}{12}\right)$$

Find the distribution of X. [1 mark]

X	-4	-2	0	2	4
$P(X = x)$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{5}{12}$

1.

X	-4	-2	0	2	4
$P(X=x)$	$\frac{5}{12}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{8}$

2.

X	-4	-2	0	2	4
$P(X=x)$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{1}{6}$	$\frac{1}{8}$

3.

X	-4	-2	0	2	4
$P(X=x)$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{1}{8}$	$\frac{1}{6}$

4.

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 be given by

$$M_X(\lambda) = \left(\frac{1}{8}\right) e^{-4\lambda} + \left(\frac{1}{6}\right) e^{-2\lambda} + \left(\frac{1}{6}\right) e^{2\lambda} + \left(\frac{1}{8}\right) e^{4\lambda} + \left(\frac{5}{12}\right)$$

Therefore, distribution of X is given by

X	-4	-2	0	2	4
$P(X=x)$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{1}{6}$	$\frac{1}{8}$

---

13. A fair die is rolled 3600 times. Use CLT to compute the probability that six appears at most 630 times. Enter the answer correct to two decimal places.
(Hint: Use $F_Z(1.341) = 0.91$)
Solution:
Define a random variable X such that

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

Therefore, $E[X] = \mu = \frac{1}{6}$ and
$\text{Var}(X) = \sigma^2 = \frac{1}{6}.\frac{5}{6} = \frac{5}{36}$

Let $X_1, X_2, \ldots, X_{3600}$ be outcomes on rolling the fair die 3600 times.
Notice that $X_1+X_2+\ldots+X_{3600}$ will denote the number of times six appears in 3600 rolls.

Let $S = X_1 + X_2 + \ldots + X_{3600}$

To find: $P(S \leq 630)$

By CLT, we know that

$$\frac{S - 3600\mu}{\sigma\sqrt{n}} \sim \text{Normal}(0, 1)$$
$$\Rightarrow \frac{S - 600}{10\sqrt{5}} \sim \text{Normal}(0, 1)$$

Now,

$$\begin{aligned} P(S \leq 630) &= P(S - 600 \leq 30) \\ &= P(\frac{S - 600}{10\sqrt{5}} \leq \frac{30}{10\sqrt{5}}) \\ &= P(Z \leq 1.34) \\ &= 0.91 \end{aligned}$$

---

14. A fair die is rolled 1000 times. Let X denote the number of times six is obtained. Find a bound for the probability that $\frac{X}{1000}$ differs from $\frac{1}{6}$ by more than 0.2 using weak law of large numbers.
    1. at least $\frac{5}{1440}$
    2. at least $\frac{1436}{1440}$
    3. at most $\frac{5}{1440}$
    4. at most $\frac{1436}{1440}$

Solution:
X denotes the number of times six is obtained on rolling the die 1000 times.
Let $X_1, X_2, \ldots, X_{1000}$ be 1000 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 $X = X_1 + X_2 + X_3 + \ldots + X_{1000}$

To find: Bound on $P\left(\left|\frac{X}{1000} - \frac{1}{6}\right| > 0.2\right)$.

By weak law of large numbers, we have

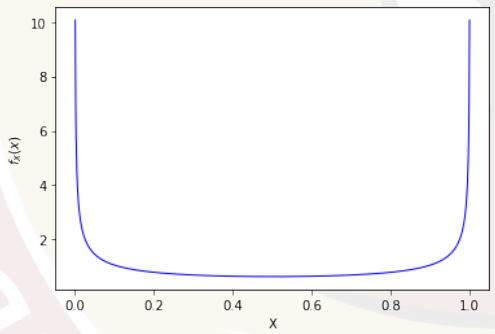
$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$

$$\Rightarrow P\left(\left|\frac{X}{1000} - \frac{1}{6}\right| > 0.2\right) \leq \frac{5}{36 \times 1000 \times 0.04}$$

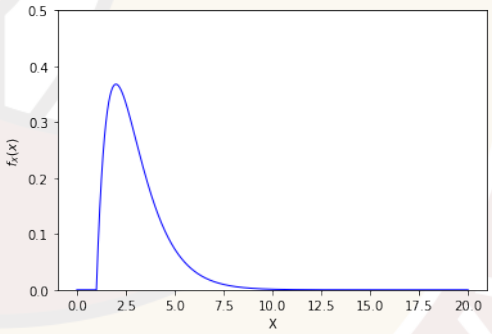
$$\Rightarrow P\left(\left|\frac{X}{1000} - \frac{1}{6}\right| > 0.2\right) \leq \frac{5}{1440}$$

---

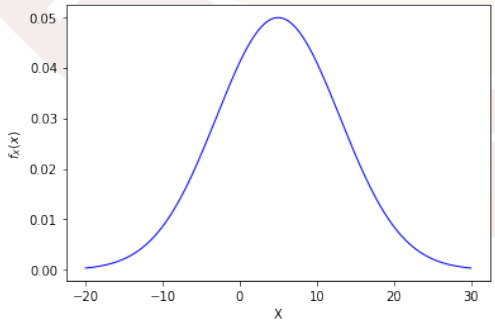
15. Consider the following PDF curves and match them with the correct distribution. [1 mark]



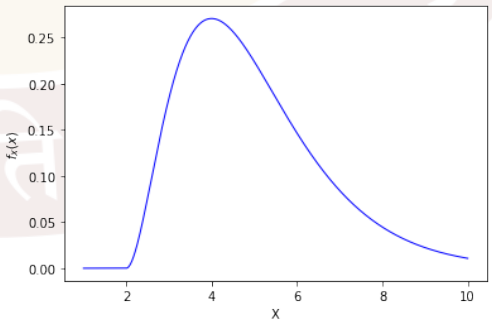
Graph 1



Graph 2



Graph 3



Graph 4

(a) Graph 1 → Gamma, Graph 2 → Normal, Graph 3 → Gamma, Graph 4 → Beta.

(b) Graph 1 → Beta, Graph 2 → Gamma, Graph 3 → Normal, Graph 4 → Gamma.

(c) Graph 1 → Beta, Graph 2 → Normal, Graph 3 → Normal, Graph 4 → Gamma.

(d) Graph 1 → Gamma, Graph 2 → Normal, Graph 3 → Normal, Graph 4 → Beta.

Solution:
Graph 1: Range of the distribution is [0, 1] and shape of the graph resembles to the Beta distribution.

Graph 2: PDF curve is not symmetric about mean and shape of the graph resembles to the Gamma distribution.

Graph 3: PDF curve is symmetric about mean and shape of the graph resembles to the Normal distribution.

Graph 4: PDF curve is not symmetric about mean and shape of the graph resembles to the Gamma distribution.
Therefore, Graph 1 → Beta, Graph 2 → Gamma, Graph 3 → Normal, Graph 4 → Gamma.

---

16. Let X_1, X_2 and $X_3 \sim$ i.i.d. X where X has the following probability mass function:

x	-1	2
$f_X(x)$	$\frac{2}{3}$	$\frac{1}{3}$

Table 7.1.P: PMF of X

Find the distribution of $Y = X_1 + X_2 + X_3$. [1 mark]

(a)

Y	-3	0	3	6
$P(Y = y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$

(b)

Y	-3	0	3	6
$P(Y = y)$	$\frac{8}{27}$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{1}{27}$

(c)

Y	-3	0	3	6
$P(Y = y)$	$\frac{8}{27}$	$\frac{1}{27}$	$\frac{4}{9}$	$\frac{2}{9}$

(d)

Y	-3	0	3	6
$P(Y = y)$	$\frac{2}{9}$	$\frac{8}{27}$	$\frac{1}{27}$	$\frac{4}{9}$

Solution:
The PMF of X is given by

x	-1	2
$f_X(x)$	$\frac{2}{3}$	$\frac{1}{3}$

Given that $Y = X_1 + X_2 + X_3$ where X_1, X_2 and $X_3 \sim$ i.i.d. X.
To find: Distribution of Y.

We will find the distribution of X by finding the MGF of Y.

$$\begin{aligned}
M_Y(\lambda) &= E[e^{\lambda Y}] \\
&= E[e^{\lambda(X_1+X_2+X_3)}] \\
&= E[e^{\lambda X_1} e^{\lambda X_2} e^{\lambda X_3}] \\
&= E[e^{\lambda X_1}]E[e^{\lambda X_2}]E[e^{\lambda X_3}] \qquad \text{(Since, } X_1, X_2 \text{ and } X_3 \text{ are independent)} \\
&= E[e^{\lambda X}]E[e^{\lambda X}]E[e^{\lambda X}] \qquad \text{(Since, } X_1, X_2 \text{ and } X_3 \sim \text{i.i.d. } X\text{)} \\
&= [M_X(\lambda)]^3 \qquad \ldots(1)
\end{aligned}$$

Now,

$$\begin{aligned}
M_X(\lambda) &= E[e^{\lambda X}] \\
&= e^{-1\lambda}.P(X = -1) + e^{2\lambda}.P(X = 2) \\
&= \frac{2e^{-\lambda}}{3} + \frac{e^{2\lambda}}{3} \qquad \ldots(2)
\end{aligned}$$

From equation (1) and (2), we have

$$\begin{aligned}
M_Y(\lambda) &= \left(\frac{2e^{-\lambda}}{3} + \frac{e^{2\lambda}}{3}\right)^3 \\
&= \frac{1}{27}(2e^{-\lambda} + e^{2\lambda})^3 \\
&= \frac{1}{27}(8e^{-3\lambda} + e^{6\lambda} + 12e^{-2\lambda}e^{2\lambda} + 6e^{-\lambda}e^{4\lambda}) \qquad \text{(since, } (a+b)^3 = a^3 + b^3 + 3a^2b + 3ab^2\text{)} \\
&= \frac{8}{27}e^{-3\lambda} + \frac{1}{27}e^{6\lambda} + \frac{4}{9} + \frac{2}{9}e^{3\lambda}
\end{aligned}$$

Therefore, distribution of Y is given by

Y	-3	0	3	6
$P(Y = y)$	$\frac{8}{27}$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{1}{27}$