Statistics and data analysis I

Week 8
“Random variable(1): Random variable and expectation”

Takashi Sano, Hirotada Honda

Lecture plan

Week1: Introduction of the course and some mathematical preliminaries
Week2: Overview of statistics, One dimensional data(1): frequency and histogram
Week3: One dimensional data(2): basic statistical measures
Week4: Two dimensional data(1): scatter plot and contingency table
Week5: Two dimensional data(2): correlation coefficients, simple linear regression and concepts of Probability /
Probability(1):randomness and probability, sample space and probabilistic events
Week6:Probability(2): definition of probability, additive theorem, conditional probability and independency
Week7:Review and exam(i)
Week8: Random variable(1): random variable and expectation
Week9: Random variable(2): Chebyshev's inequality, Probability distribution(1):binomial and Poisson distributions
Week10: Probability distribution(2): normal and exponential distributions
Week11: From descriptive statistics to inferential statistics -z-table and confidwncw interval-
Week12: Hypothesis test(1) -Introduction, and distributions of test statistic (t-distribution)-
Week13: Hypothesis test(2) -Test for mean-
Week14: Hypothesis test(3) -Test for difference of mean-
Week15: Review and exam(2)

※ Might be changed!

Agenda

1. Random variable and distribution
2. Expectation and variance
3. Chebyshev inequality

1. Random variable and distribution

Random variable and distribution

  - Random variable (R.V.) is a number whose values are determined according to a certain probability

- Probability distribution is a relationship between numbers of a random variable and the corresponding probability.

Probability distribution of discrete random variables

- In case the values of a random variable is discrete (integer, for instance), its distribution is shown by a table below.

R.V. X	x_1	x_2	x_3	…	x_n
Probability P	p_1	p_2	p_3	…	p_n

□ $P(X=x_i)=p_i$

□ $p_i \geqq 0$

□ $\sum p_i=1$

Probability distribution of discrete random variables

- As an example, if we regard the pips of a dice as a R.V., we have :

R.V. *X*	1	2	3	…	6
Probability *P*	1/6	1/6	1/6	…	1/6

□ $P(X=x_i)=1/6$

□ $p_i \geqq 0$

□ $\sum p_i = 1/6+1/6+\ldots+1/6 = 1$

Probability distribution of discrete random variables

- If we regard the probability of each number of a R.V. as a function , i.e., we define$P(X=x_k)=f(x_k)$, then this function *f* is called *discrete probability distribution*.

□ $f(x_k) \geqq 0$,　　$k=1, 2, \ldots$

□ $\sum f(x_k)=1$

□ f is called the disctere probability distribution.

Probability distribution of discrete random variables

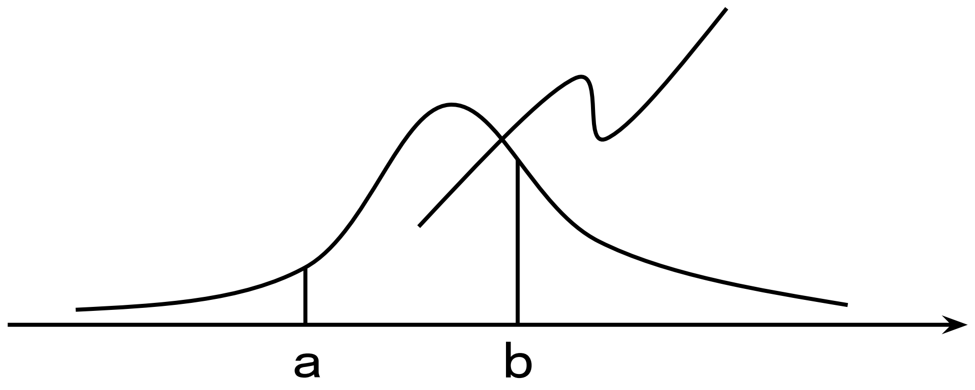
- Ex) If we regard the sum of pips of two dices as a R.V., then we have

X	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/34	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Probability distribution of discrete random variables

- Continuous R.V. takes continuous values (for instance, time / error in length or weight).
- The probability is define on an interval in its range, by using a certain function *f(x)*.

$$P(a \leqq X \leqq b) = \int_a^b f(x)\,dx$$



- This f(x) is called the *probability density* of X.
- The probability density has to satisfy

$$f(x) \geqq 0,\ \text{and},\ \int f(x)\,dx = 1$$

Probability distribution of discrete random variables

- Ex) Waiting time subject to the exponential distribution.

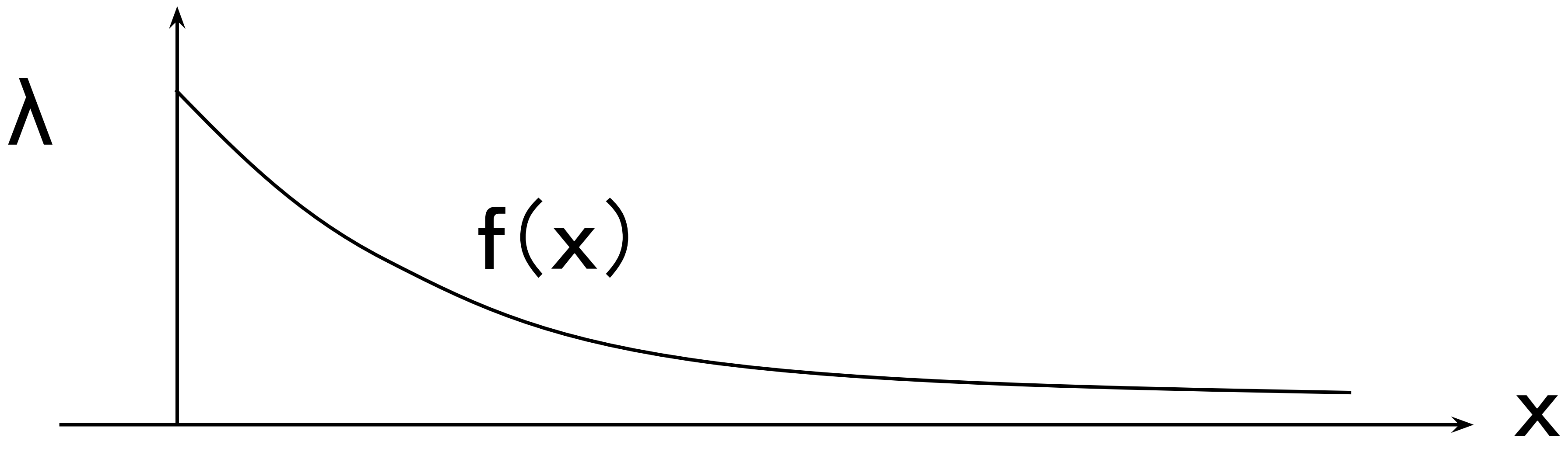
□ The intervals between large disasters;

□ The lifetime of a light bulb;

□ The intervals between the calls, and so forth.

- Exponential distribution.

□ $f(x) = \lambda e^{-\lambda x} \quad (x \geqq 0)$



Cumulative distribution

- The probability that a R.V. takes a value x or less.

□$F(x)=P(X\leqq x)$

- The discrete R.V.

□$F(x)=\sum f(u)$

- The continuous R.V.

□$F(x)=\int f(u)\,du$

□$F'(x)=f(x)$

2. Expectation and variance

Expected value

- The mean of possible values of a R.V, weighted with the probability of each value. It’s denoted as E(X).
- Ex)
- If we regard the pips of a dice as a R.V., its expected value is

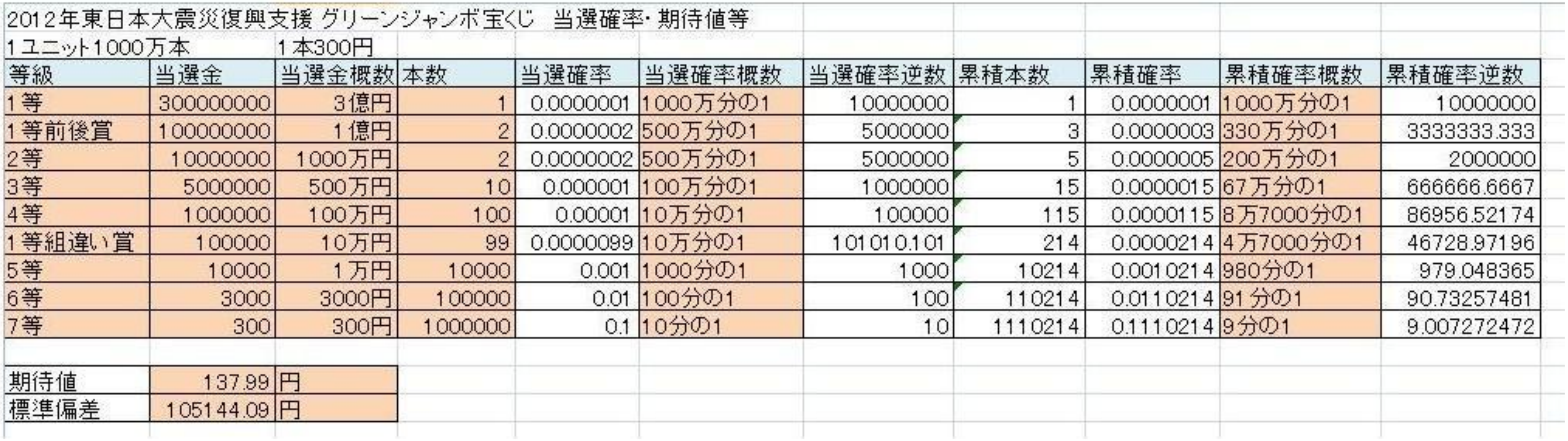
  □$E(X)=1\cdot(1/6)+\ldots+6\cdot(1/6)=3.5$
- Discrete R.V.

  □$E(X)=\sum x\cdot f(x)$
- Continuous R.V.

  □$E(X)=\int x\cdot f(x)\,dx$

Example of expected value

- Expected value of lottery



2012年東日本大震災復興支援 グリーンジャンボ宝くじ　当選確率･期待値等

1ユニット1000万本　1本300円

等級	当選金	当選金概数	本数	当選確率	当選確率概数	当選確率逆数	累積本数	累積確率	累積確率概数	累積確率逆数
1等	300000000	3億円	1	0.0000001	1000万分の1	10000000	1	0.0000001	1000万分の1	10000000
1等前後賞	100000000	1億円	2	0.0000002	500万分の1	5000000	3	0.0000003	330万分の1	3333333.333
2等	10000000	1000万円	2	0.0000002	500万分の1	5000000	5	0.0000005	200万分の1	2000000
3等	5000000	500万円	10	0.000001	100万分の1	1000000	15	0.0000015	67万分の1	666666.6667
4等	1000000	100万円	100	0.00001	10万分の1	100000	115	0.0000115	8万7000分の1	86956.52174
1等組違い賞	100000	10万円	99	0.0000099	10万分の1	101010.101	214	0.0000214	4万7000分の1	46728.97196
5等	10000	1万円	10000	0.001	1000分の1	1000	10214	0.0010214	980分の1	979.048365
6等	3000	3000円	100000	0.01	100分の1	100	110214	0.0110214	91分の1	90.73257481
7等	300	300円	1000000	0.1	10分の1	10	1110214	0.1110214	9分の1	9.007272472

期待値	137.99	円
標準偏差	105144.09	円

Calculation of expected value

- $E(c)=c$
- $E(X+c)=E(X)+c$
- $E(cX)=cE(X)$
- $E(X+Y)=E(X)+E(Y)$: Addition formula

□ Now, let us compare the expected values of the pip of a dice and the mean of the pips of two dices.

□ $E(X)=3.5$

□ $E(Y)=E\{(X_1+X_2)/2\}=\{E(X_1)+E(X_2)\}/2=3.5$

Calculation of expected value (simple example)

- Distribution of pips of dice
  - f(x)＝1/6,　(x＝1，2，...，6)
  - Can be generalized as...
- Suppose there are N balls numbered from 1 to N. They are placed in a box.
- Suppose you take out a box from the box, and repeat such events. Then, consider the distribution of the numbers of balls.
  - f(x)＝1/N,　(x＝1，2，...，N)

Calculation of expected value (simple example)

R.V. (X)	1	2	3	…	N
Probability	1/N	1/N	1/N	…	1/N

$$\begin{aligned} E[X] &= 1 \times \frac{1}{N} + 2 \times \frac{1}{N} + \ldots + N \times \frac{1}{N} \\ &= \frac{1}{N} \times \left\{ 1 + 2 + \ldots + N \right\} \\ &= \frac{1}{N} \times \frac{N(N+1)}{2} \\ &= \frac{(N+1)}{2}. \end{aligned}$$

- Expected value: $E[X] = 1 \times \frac{1}{N} + 2 \times \frac{1}{N} + \ldots + N \times \frac{1}{N} = \sum_{i=1}^{N} x_i f(x_i) = \frac{N+1}{2}.$

Variance

- You cannot capture the characteristics of R.V.s. For instance, two R.V.s with different distributions may have the same expected values.
  - Let X be the pip of a dice, and Y, the mean of pips of two dices:$Y=(X_1+X_2)/2$. Here X1 and X2 are the pips of a dice.
  - Let us compare the expected values of X and Y.

- Variance : the scale of variation of a R.V. around its expected value.

Variance

- Let us denote the expected value and variance as $\mu=E(X)$ and $V(X)$， respectively.

  □$V(X)=E\{(X-\mu)^2\}$

- For discrete R.V.s,

  □$V(X)=\sum (x-\mu)^2 f(x)$

- For continuous R.V.s,

  □$V(X)=\int (x-\mu)^2 f(x)\,dx$

  The following formula is frequently used.

  □$V(X)=E(X^2)-\{E(X)\}^2$

（Expected value of X^2）
-（ squared expected value
）

Exercise

- Let us regard the pip of a dice as a random variable X. Then, find its variance.

Exercise【Answer】

- Let us regard the pip of a dice as a random variable *X*. Then, find its variance.
- By using the formula we have seen before (for N in general), if we apply N=6, we have
- E[X] =(6+1)/2 = 7/2.
- Next, let us consider $E[X^2]$.

Exercise【Answer】

- Let us regard the pip of a dice as a random variable *X*. Then, find its variance.
- Next, let us consider $E[X^2]$.

X^2	1^2	2^2	3^2	…	6^2
Probability	1/6	1/6	1/6	…	1/6

$$E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + \ldots + 6^2 \times \frac{1}{6}$$

$$= (1^2 + 2^2 + \ldots + 6^2) \times \frac{1}{6} \quad = \frac{91}{6}$$

Exercise【Answer】

- Let us regard the pip of a dice as a random variable *X*. Then, find its variance.
- Then, by using the formula below, we have

$$V[X] = E[X^2] - (E[X])^2 = \frac{91}{6} - (\frac{7}{2})^2 = \boxed{\frac{35}{12}}$$

Calculation of variance

- $V(c)=0$
- $V(X+c)=V(X)$
- $V(cX)=c^2V(X)$

Standard deviation and z-variable

- Standard deviation is the square root of variance.
- It is denoted as D[X].

$$D[X] = \sqrt{V[X]}$$

- Normalization of R.V.

$$Z = \frac{(X - E[X])}{D[X]}$$

  - Every R.V. can be transformed to another R.V. Z that satisfies
  - E[Z]＝0, V[Z]＝1
  - This Z is called as the normalized R.V.

3. Chebyshev inequality

Chebyshev inequality

- Shows the relationship between the distribution and S.D. It holds for arbitrary random variable as far as its expected value and standard deviation are finite.

- The probability of a set of values of a r.v. X, that are apart from the expected value by n×S.D., is less than $1/n^2$.

$$P(|X-\mu|\geq k\sigma)\leq\frac{1}{k^2}$$

Here, $\mu=E(X)$, $\sigma^2=V(X)$

Chebyshev inequality

Suppose there are a large amount of sentences, and the mean length of them is 1000 strings, and S.D. is 200.
Then, we can conclude that the sentences of 600-1400 strings account for at least 75%.

Chebyshev inequality

Suppose there are a large amount of sentences, and the mean length of them is 1000 strings, and S.D. is 200.
Then, we can conclude that the sentences of 600-1400 strings account for at least 75%.

$$P(|X-1000| \geqq 200k) \leqq 1/k^2$$

$$P(|X-1000| \geqq 200*2) \leqq 1/2^2 = 0.25$$

$$P(X \leqq 600 \text{ or } X \geqq 1400) \leqq 0.25$$

$$P(600 < X < 1400) = 1-P(X \leqq 600 \text{ or } X \geqq 1400))$$
$$\geqq 1-0.25 = 0.75$$

Summary : Chebyshev's inequality.

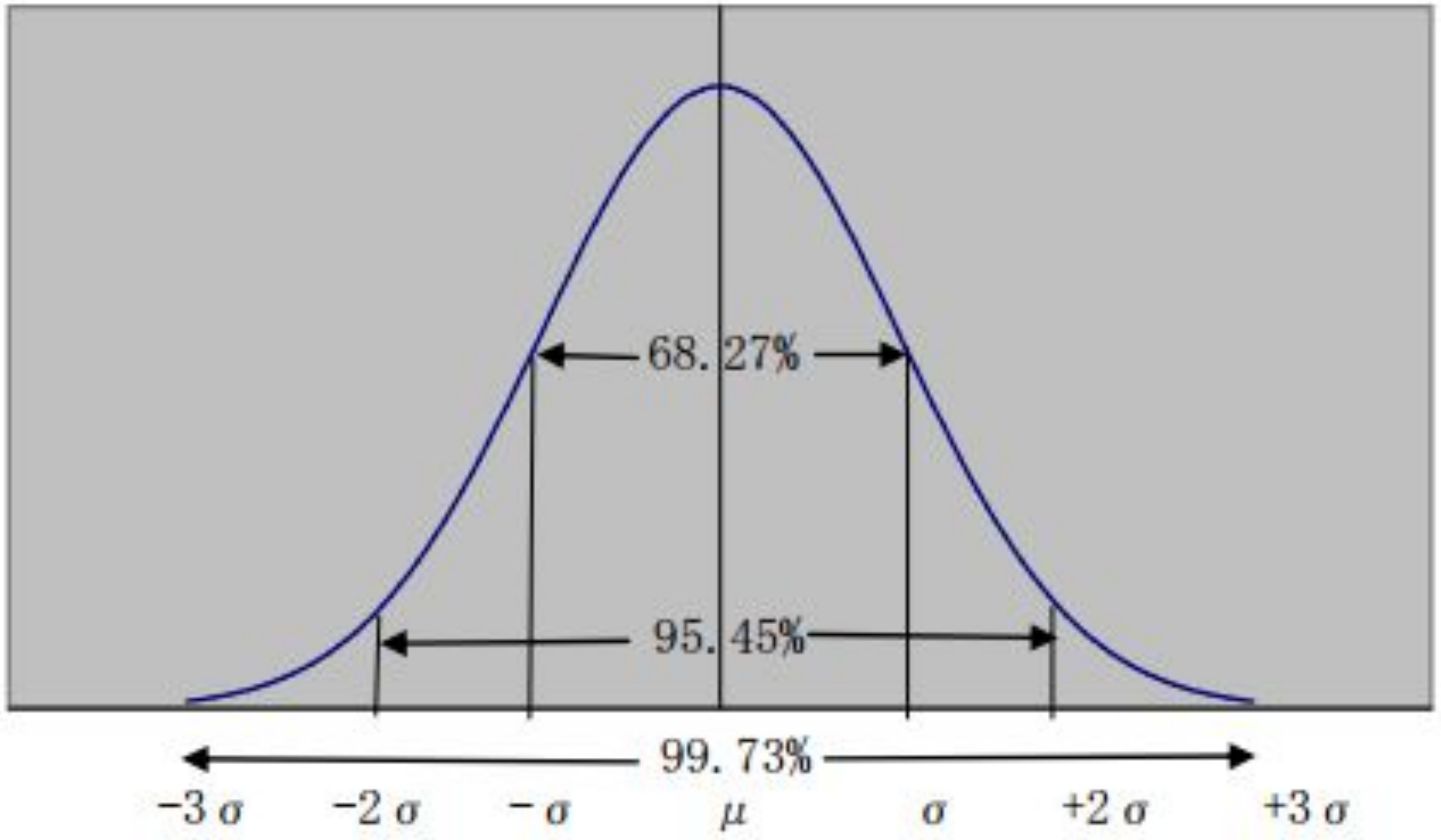
Assume the expected value and SD of a certain r.v. X are μ and σ, resp.
Then, for arbitrary k(>0), we have

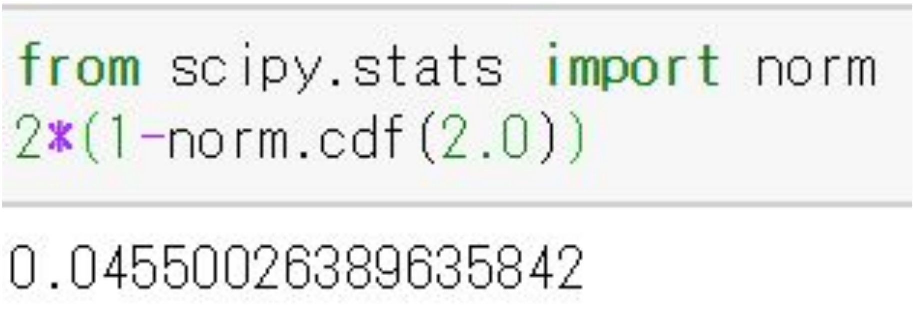
$$P(|X-\mu|\geq k\sigma)\leq \frac{1}{k^2}$$

【Ref.】Normal dits.

In case of normal dist. we know (can calc.) (see, Week3)

$$P\Big(|X-\mu| \geq 2\sigma\Big) = 0.0455$$





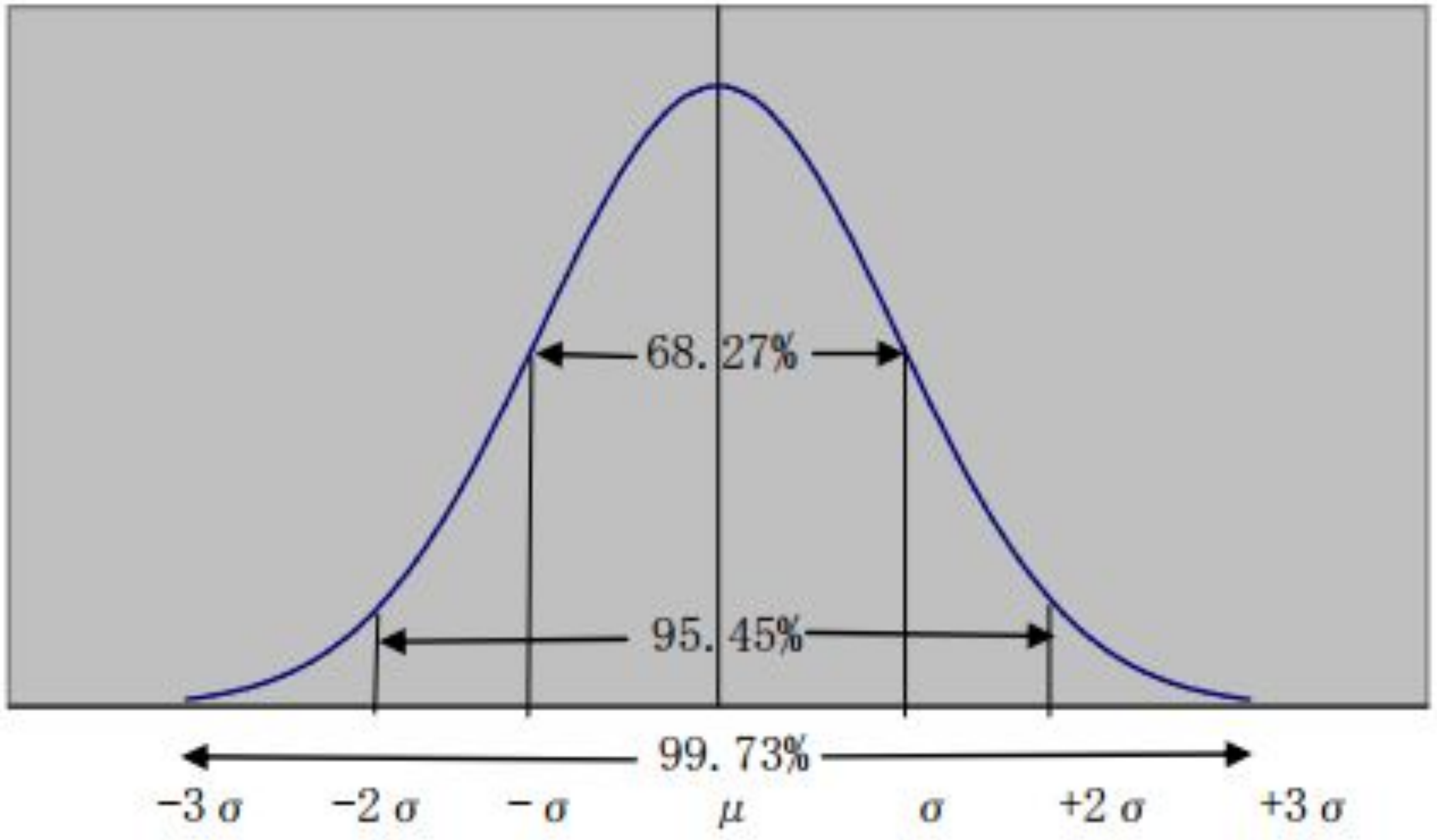
```
from scipy.stats import norm
2*(1-norm.cdf(2.0))
```

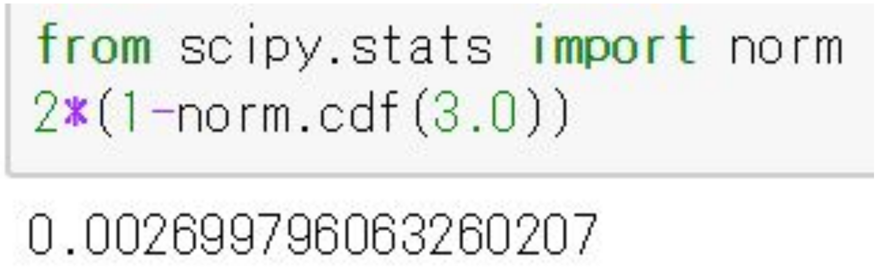
```
0.04550026389635842
```

【Ref.】Normal dits.

In case of normal dist. we know (can calc.) (see, Week3)

$$P\Big(|X-\mu| \geq 3\sigma\Big) = 0.0027$$





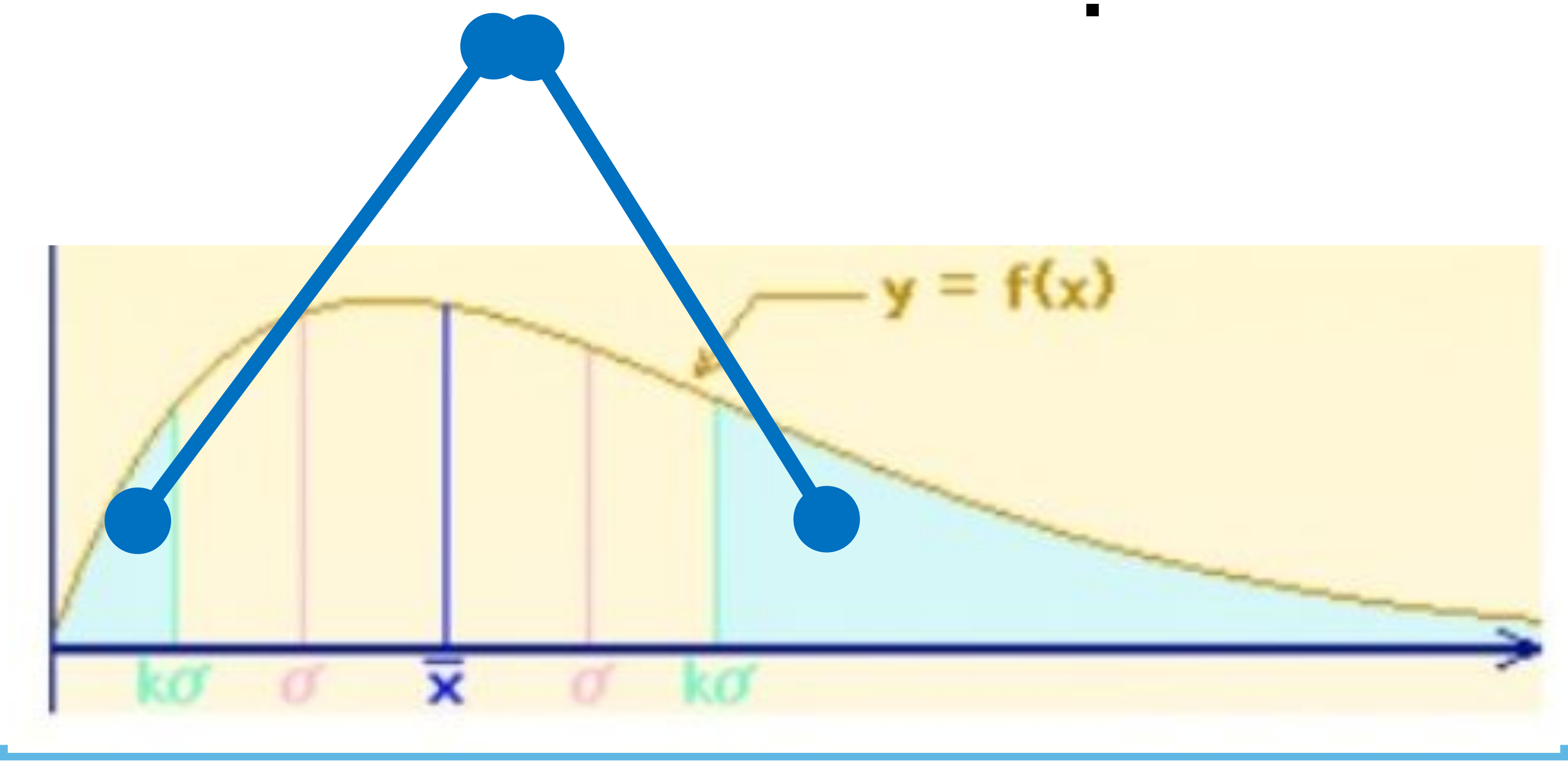
```
from scipy.stats import norm
2*(1-norm.cdf(3.0))
```

```
0.0026997960632602071
```

Estimate of prob.

What about the arbitrary r.v.?
“Under the assumption of non-normality, find the probability
 that the value lies outside of μ by the distance of
2 SDs or more.”

$$P\left(|X-\mu| \geq k\sigma\right)$$ =??
?

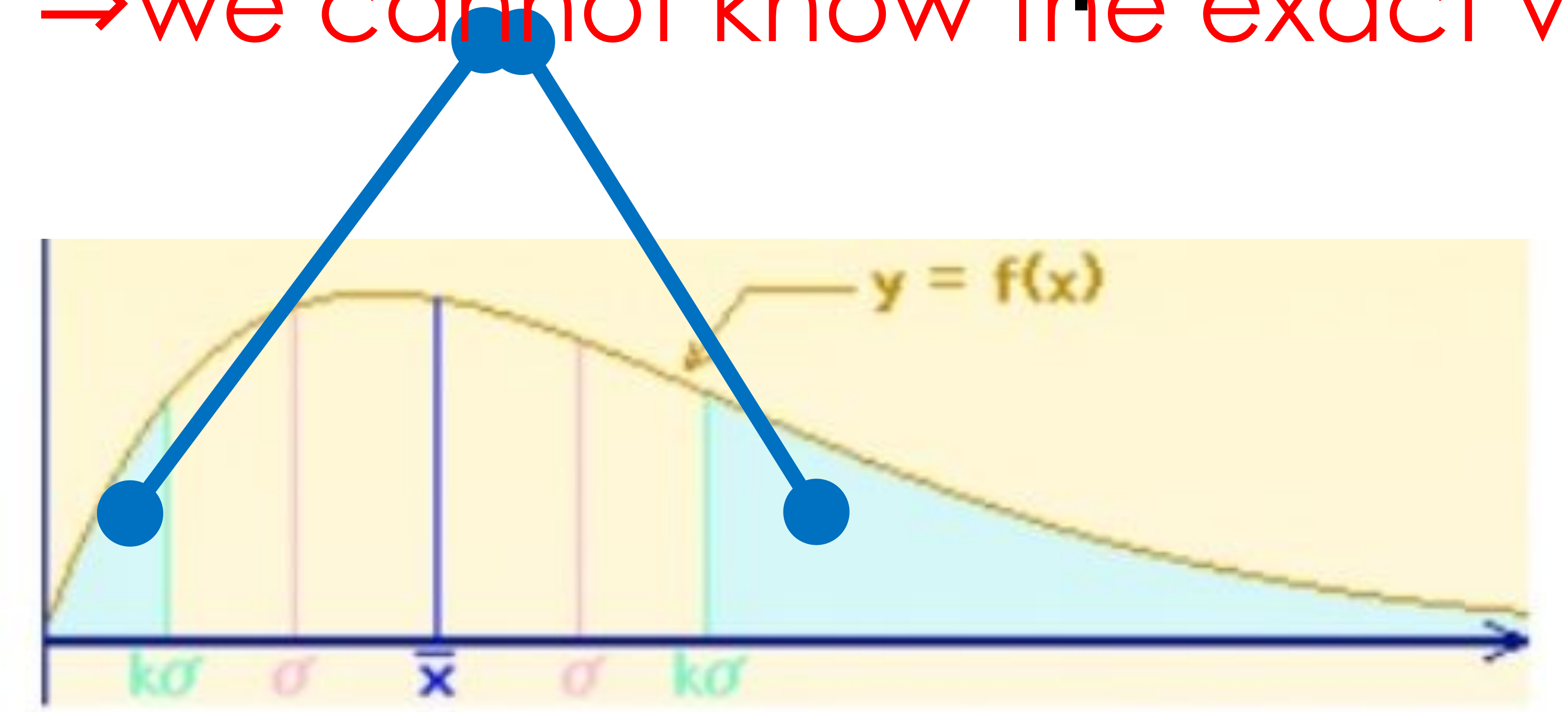


Estimate of prob.

What about the arbitrary r.v.?
“Under the assumption of non-normality, find the probability
 that the value lies outside of μ by the distance of
2 SDs or more.”

$$P\left(|X-\mu| \geq k\sigma\right)$$ =??
?

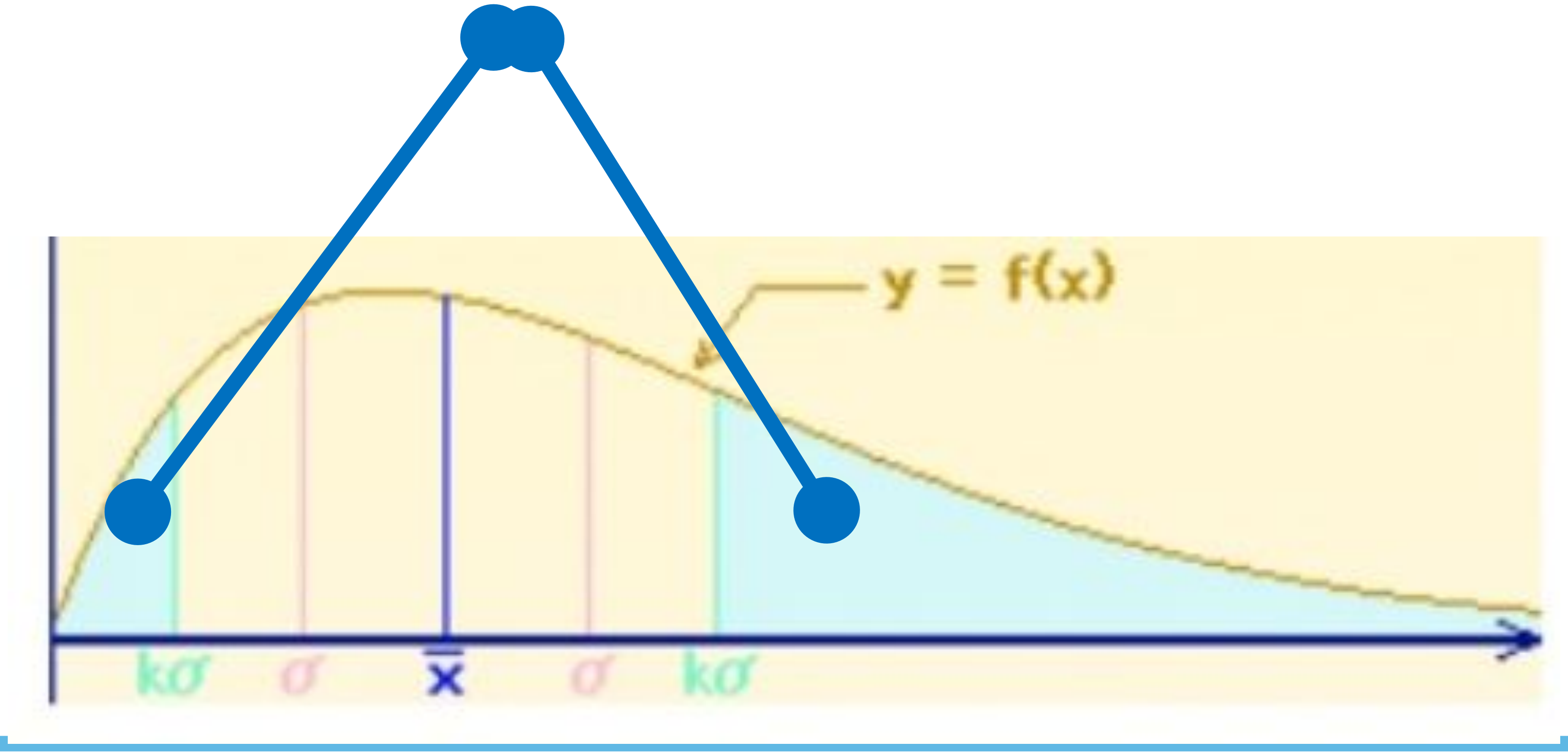
⇒we cannot know the exact value as in case of normal dist.



Estimate of prob.

However, thanks to the **Chebyshev's inequality,**
we can estimate the desired prob. from above (or below)

$$P(|X-\mu|\geq k\sigma)\leq\frac{1}{k^2}$$

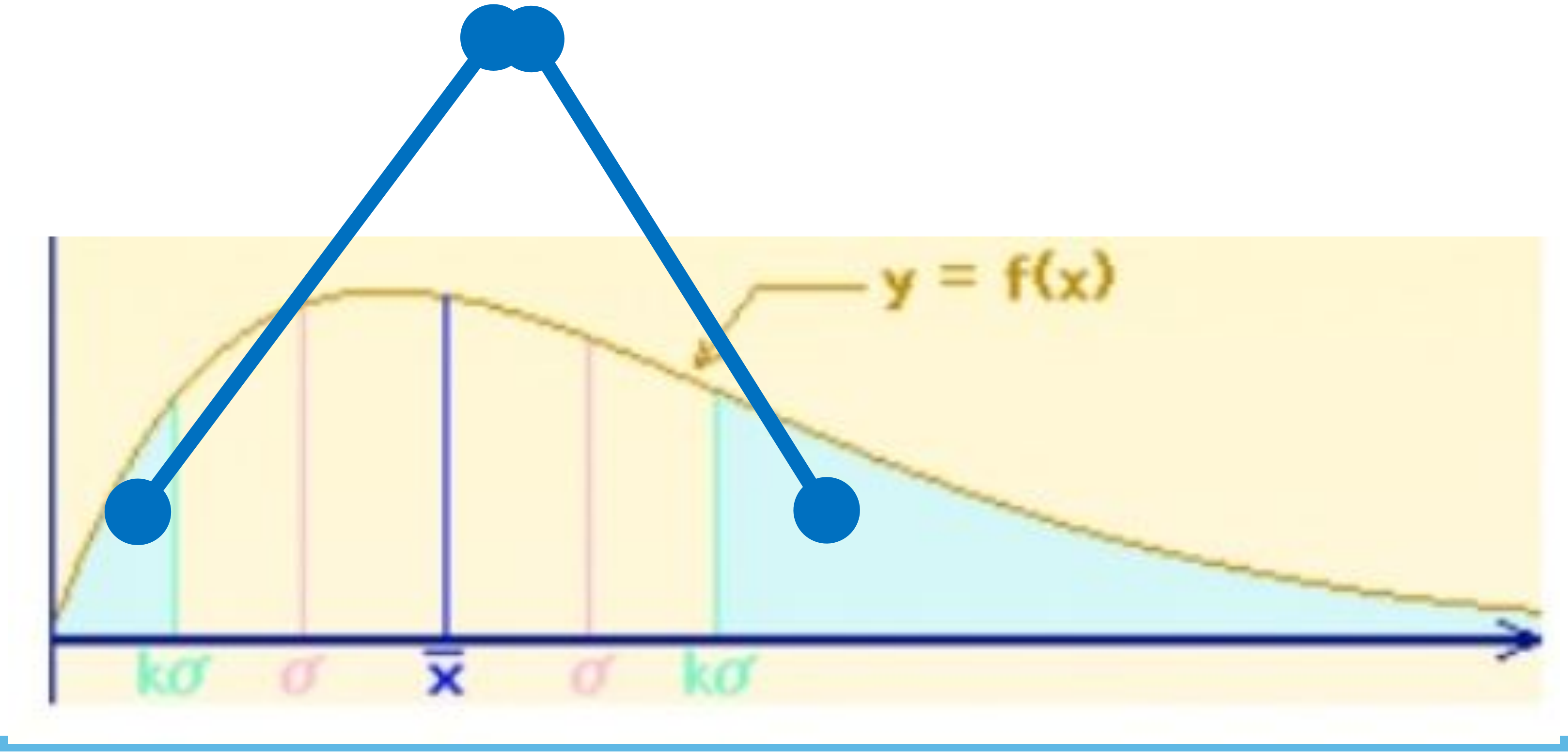


$$P\Big(|X-\mu|\geq 2\sigma\Big)\leq\frac{1}{4}=0.25.$$

Estimate of prob.

For instance, when k=2,

$$P\Big(|X-\mu| \geq 2\sigma\Big) \leq \frac{1}{4} = 0.25.$$



Exercise ①

For a certain r.v. X, let us assume that μ=E[X] = 10, and V[X]=3. Then, estimate P(|X-10|≧2) by using the Chebyshev inequality.

Exercise ① 【Answers】

For a certain r.v. X, let us assume that μ=E[X] = 10, and V[X]=3.
Then, estimate P(|X-10|≧2) by using the Chebyshev inequality.

□P(|X−10|≧1.732k)≦1／k^2

Take k so that 1.732k＝2 . Then, k=1.1547, k^2=4/3

P(|X−10|≧2)
≦1／k^2=<u>3/4</u>

Exercise ②

For a certain r.v. X, let us assume that μ=E[X] = 5000, and V[X]=2500.

Then, estimate P(|X-5000|<400) by using the Chebyshev inequality.

Exercise ②【Answer】

For a certain r.v. X, let us assume that $\mu=E[X] = 5000$, and $V[X]=2500$.
Then, estimate $P(|X-5000|<400)$ by using the Chebyshev inequality.

□$P(|X-5000|\geqq 50k)\leqq 1/k^2$

Take k so that $50k=400$, i.e. $k=8$.

$P(|X-5000|\geqq 400)\leqq 1/k^2=1/64$

$$P(|X-5000|<400) = 1-P(|X-5000|\geqq 400)$$
$$\geqq 1-1/64=\underline{63/64}$$

Exercise ③

For a certain r.v. X, let us assume that $\mu=E[X] = 0$, and $V[X]=1/5$.
Then, estimate $P(|X|<3/4)$ by using the Chebyshev inequality.

Exercise ③【Answer】

For a certain r.v. X, let us assume that μ=E[X] = 0,
and V[X]=1/5.
Then, estimate P(|X|<3/4) by using the Chebyshev inequality.

□$P(|X-0|\geqq k/4)\leqq 1/k^2$

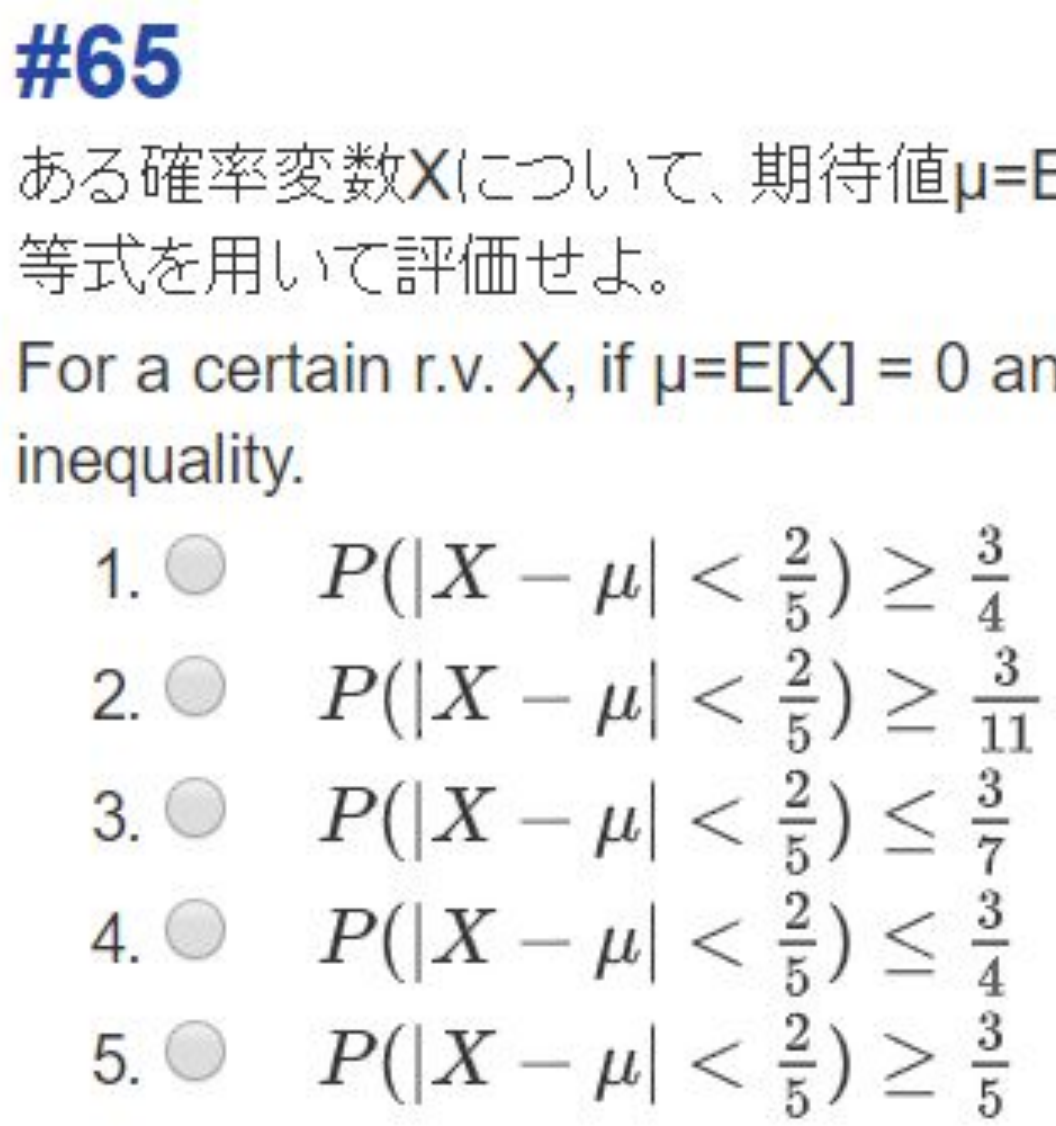
$k/4=3/4$ i.e., $k=3$

$P(|X|\geqq 3/4)$

$\leqq 1/k^2=1/9$

$P(|X|<3/4)=1-P(|X|\geqq 3/4)$

$\geqq 1-1/9=8/9.$

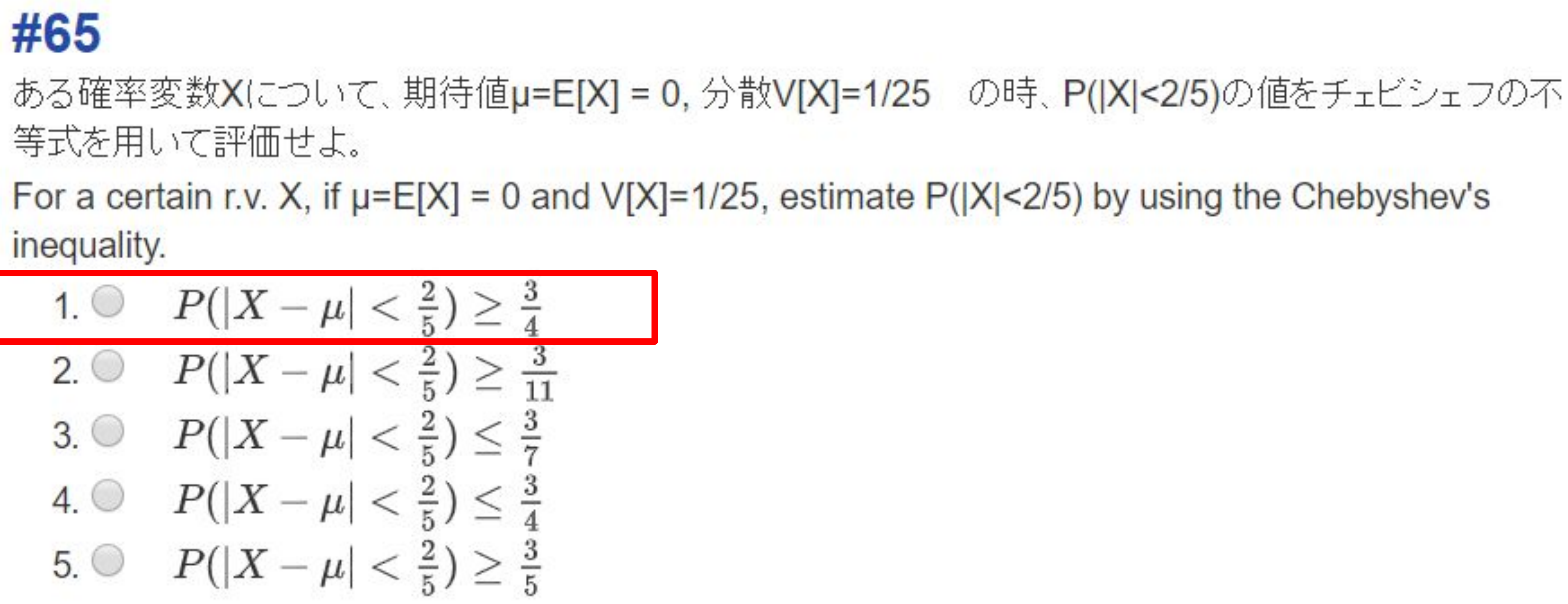


#65

ある確率変数Xについて、期待値μ=E[X] = 0, 分散V[X]=1/25　の時、P(|X|<2/5)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 0 and V[X]=1/25, estimate P(|X|<2/5) by using the Chebyshev's inequality.

1. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{4}$
2. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{11}$
3. ○ $P(|X-\mu| < \frac{2}{5}) \leq \frac{3}{7}$
4. ○ $P(|X-\mu| < \frac{2}{5}) \leq \frac{3}{4}$
5. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{5}$



#65

ある確率変数Xについて、期待値μ=E[X] = 0, 分散V[X]=1/25　の時、P(|X|<2/5)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 0 and V[X]=1/25, estimate P(|X|<2/5) by using the Chebyshev's inequality.

[✓ 1. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{4}$]

2. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{11}$

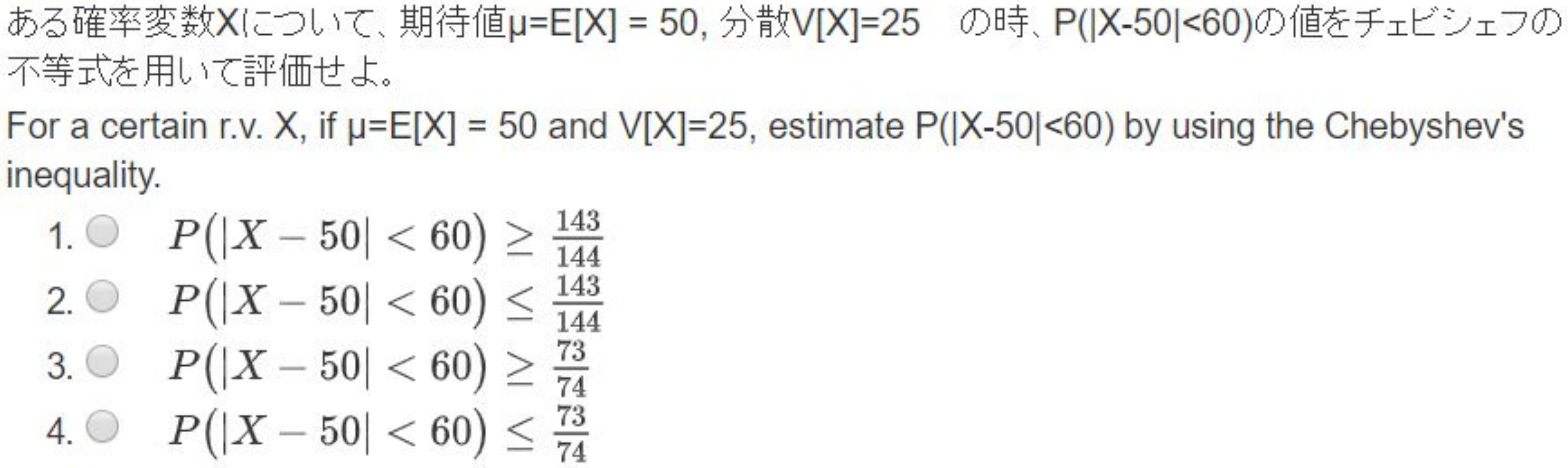
3. ○ $P(|X-\mu| < \frac{2}{5}) \leq \frac{3}{7}$

4. ○ $P(|X-\mu| < \frac{2}{5}) \leq \frac{3}{4}$

5. ○ $P(|X-\mu| < \frac{2}{5}) \geq \frac{3}{5}$

σ=1/5, ⇒　k=2

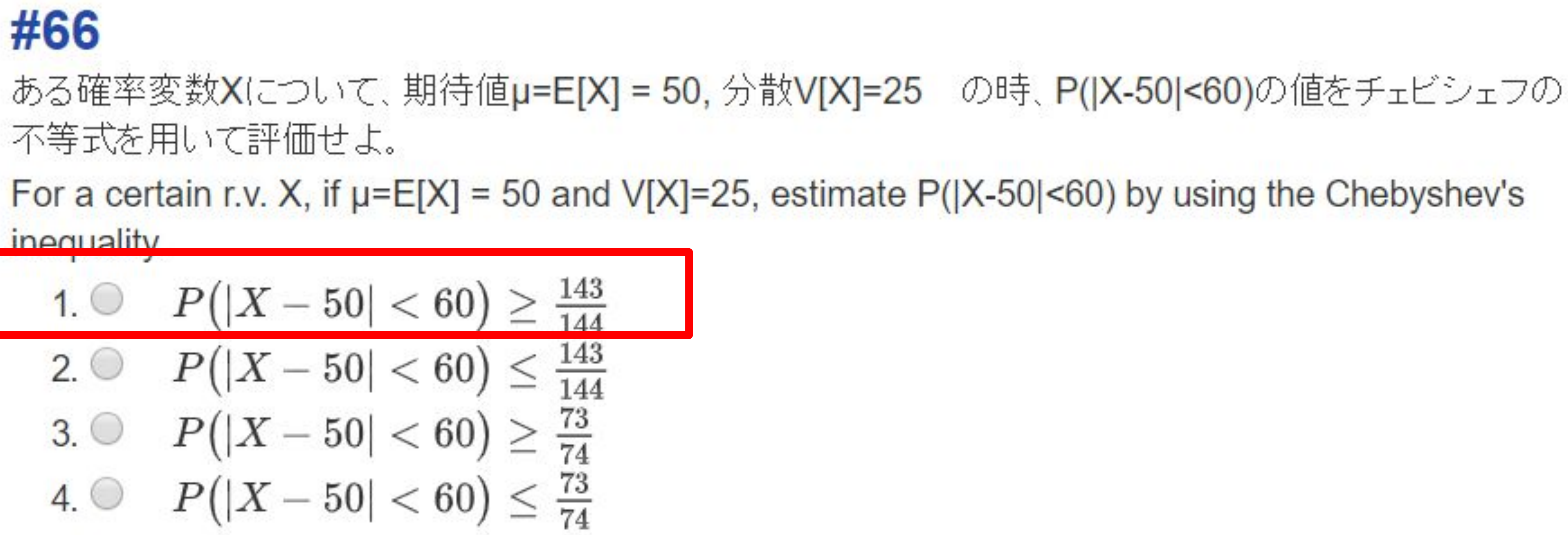
#66



ある確率変数Xについて、期待値μ=E[X] = 50, 分散V[X]=25　の時、P(|X-50|<60)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 50 and V[X]=25, estimate P(|X-50|<60) by using the Chebyshev's inequality.

1. ○ $P(|X-50|<60) \geq \frac{143}{144}$
2. ○ $P(|X-50|<60) \leq \frac{143}{144}$
3. ○ $P(|X-50|<60) \geq \frac{73}{74}$
4. ○ $P(|X-50|<60) \leq \frac{73}{74}$



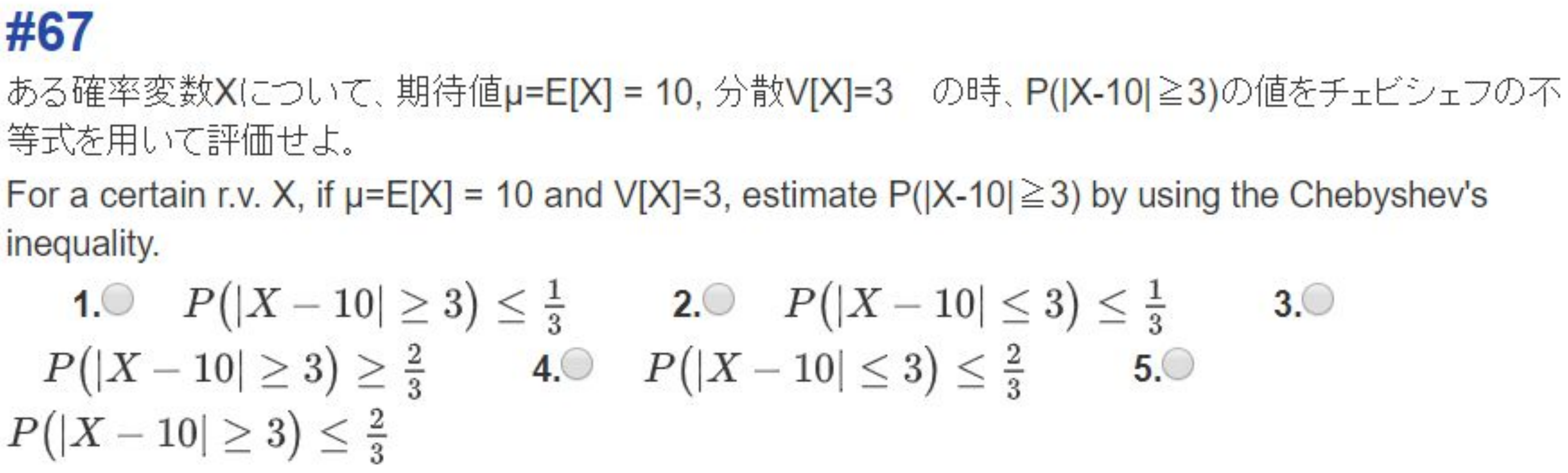
#66

ある確率変数Xについて、期待値μ=E[X] = 50, 分散V[X]=25　の時、P(|X-50|<60)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 50 and V[X]=25, estimate P(|X-50|<60) by using the Chebyshev's inequality

1. [✓ ○ $P(|X-50| < 60) \geq \frac{143}{144}$]
2. ○ $P(|X-50| < 60) \leq \frac{143}{144}$
3. ○ $P(|X-50| < 60) \geq \frac{73}{74}$
4. ○ $P(|X-50| < 60) \leq \frac{73}{74}$

σ=5, ⇒　k=12

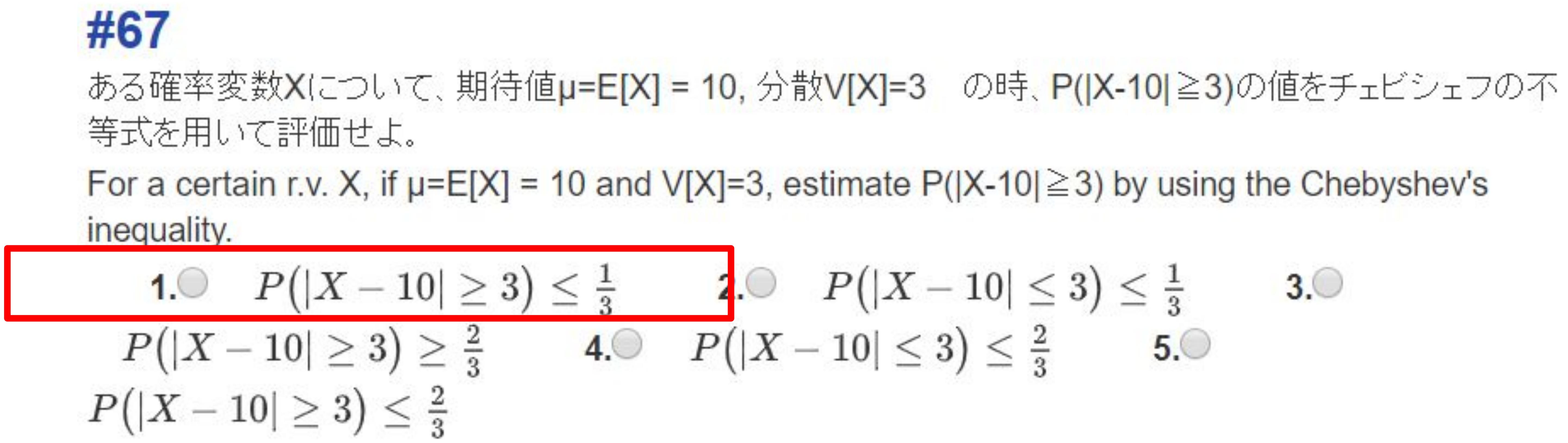


#67

ある確率変数Xについて、期待値μ=E[X] = 10, 分散V[X]=3　の時、P(|X-10|≧3)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 10 and V[X]=3, estimate P(|X-10|≧3) by using the Chebyshev's inequality.

1. ○ $P(|X-10| \geq 3) \leq \frac{1}{3}$
2. ○ $P(|X-10| \leq 3) \leq \frac{1}{3}$
3. ○ $P(|X-10| \geq 3) \geq \frac{2}{3}$
4. ○ $P(|X-10| \leq 3) \leq \frac{2}{3}$
5. ○ $P(|X-10| \geq 3) \leq \frac{2}{3}$



#67

ある確率変数Xについて、期待値μ=E[X] = 10, 分散V[X]=3　の時、P(|X-10|≧3)の値をチェビシェフの不等式を用いて評価せよ。

For a certain r.v. X, if μ=E[X] = 10 and V[X]=3, estimate P(|X-10|≧3) by using the Chebyshev's inequality.

[✓ 1.○ $P(|X-10| \geq 3) \leq \frac{1}{3}$] 2.○ $P(|X-10| \leq 3) \leq \frac{1}{3}$ 3.○ $P(|X-10| \geq 3) \geq \frac{2}{3}$ 4.○ $P(|X-10| \leq 3) \leq \frac{2}{3}$ 5.○ $P(|X-10| \geq 3) \leq \frac{2}{3}$

σ= $\sqrt{3}$ ⇒ k= $\sqrt{3}$

Summary

- You studied the discrete and continuous R.V.s.
- You also studied the expected value, variance (S.D.) and their features.

Summary(Checklist)

- You can state the difference between the discrete and continuous probability distributions?
- You can explain the cumulative distribution?
- Can you state the elementary calculations of expected value and variance of random variables?
- You can make the normalized R.V?