

Statistics and data analysis I

Week 9

“Probability Distribution (I): Binomial and Poisson Distributions”

2019.11th June

Hirotsada 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: Review and exam(ii)
 Week12: From descriptive statistics to inferential statistics -z-table and confidence interval-
 Week13: Hypothesis test(1) -Introduction, and distributions of test statistic (t-distribution)-
 Week14: Hypothesis test(2) -Test for mean-
 Week15: Hypothesis test(3) -Test for difference of mean-
 Week16: Review and exam(3)

※ Might be changed!

On Exam-2

- Exam-2(Week11, 6/25)
- Coverage : Week 8-10
- Question sheet is given as a pdf file, and you should answer via MOOCS test site.

Review : Chebyshev's inequality.

Assume the expected value and SD of a certain r.v. X are μ and σ , resp.

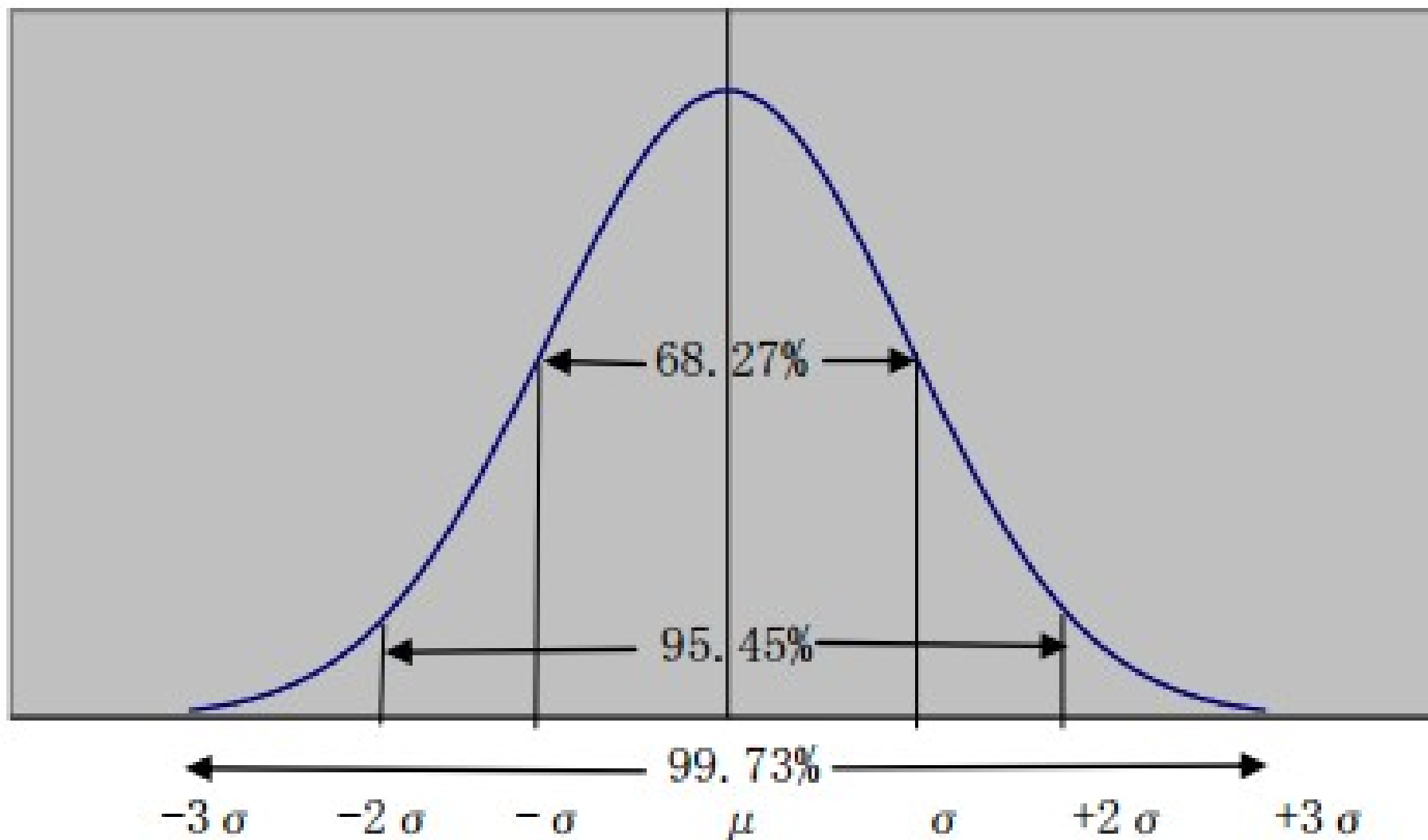
Then, for arbitrary $k(>0)$, we have

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

【Ref.】 Normal dits.

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

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



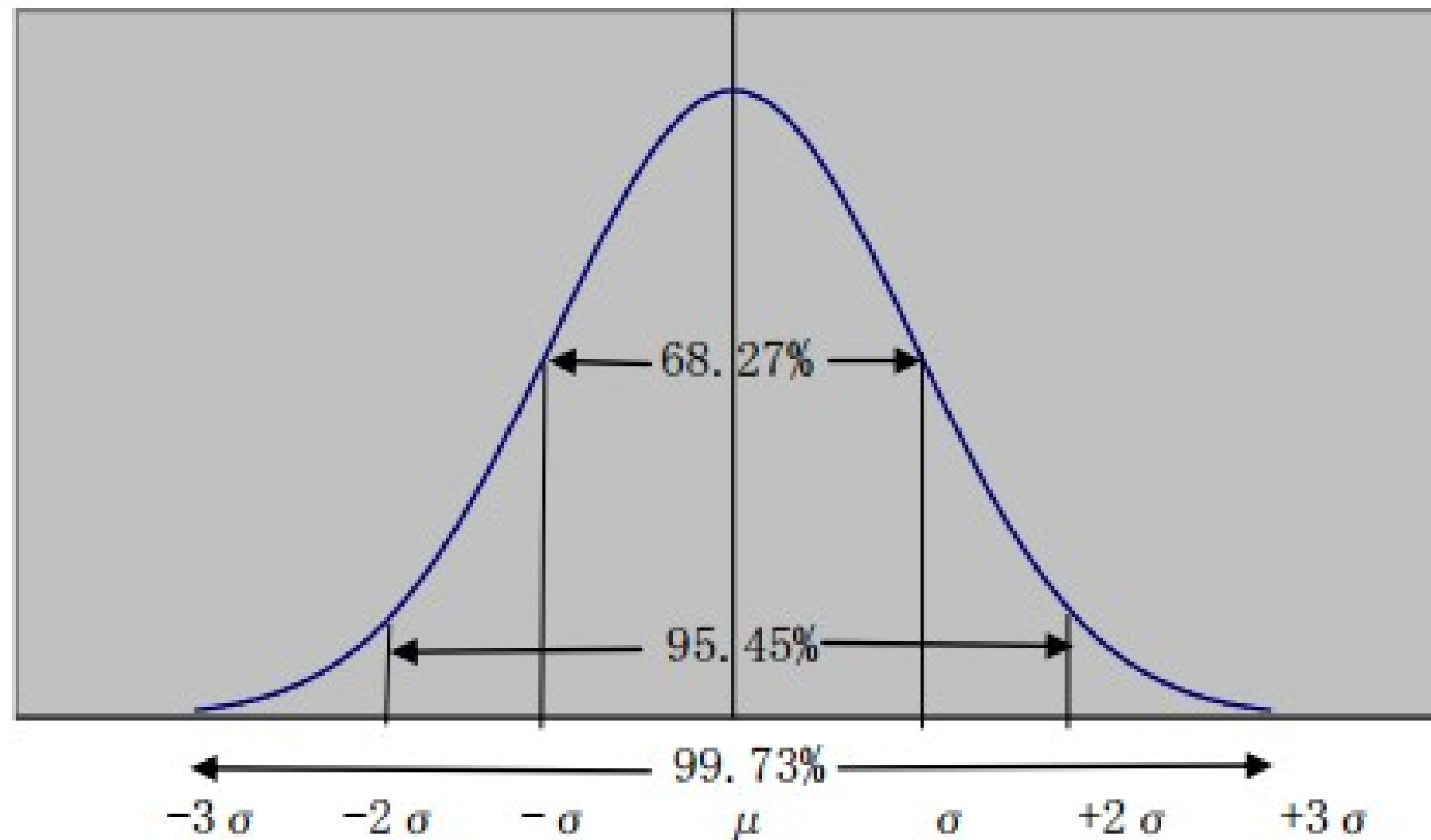
```
from scipy.stats import norm  
2*(1-norm.cdf(2.0))
```

0.04550026389635842

【Ref.】 Normal dits.

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

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



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

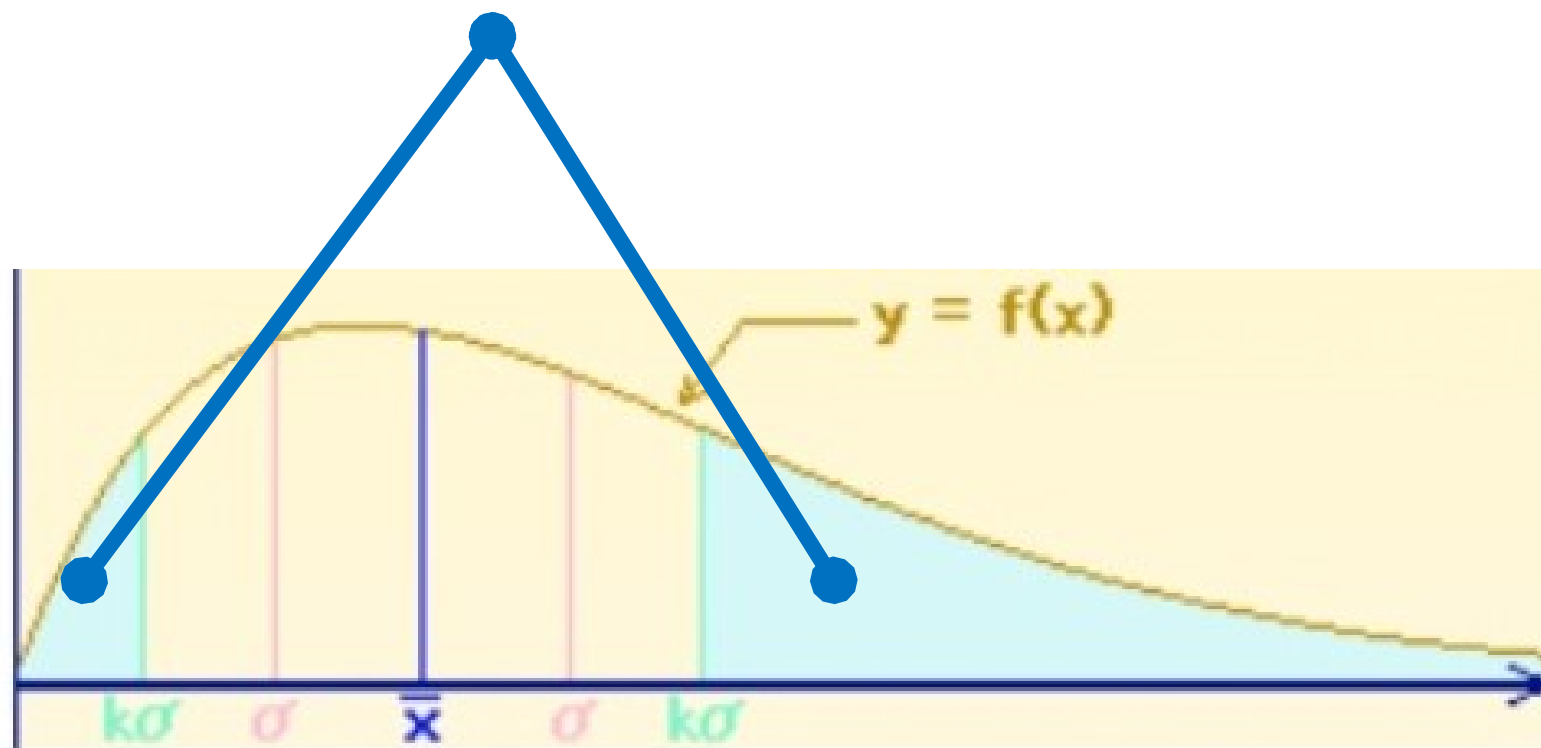
0.002699796063260207

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.”

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



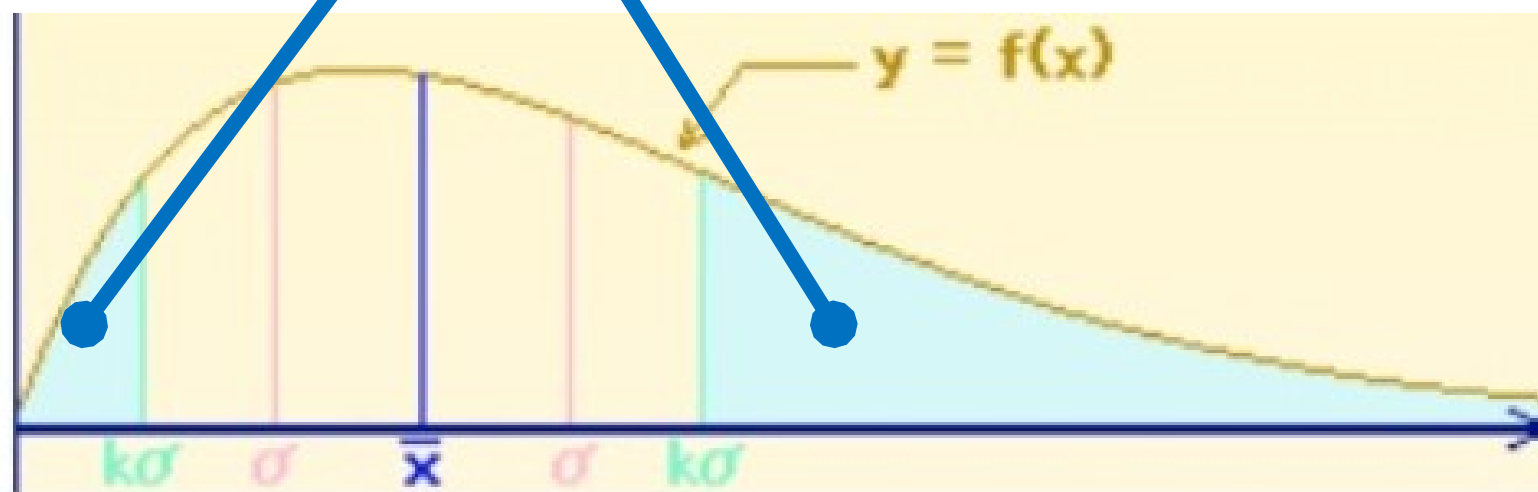
Estimate of prob.

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.”

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

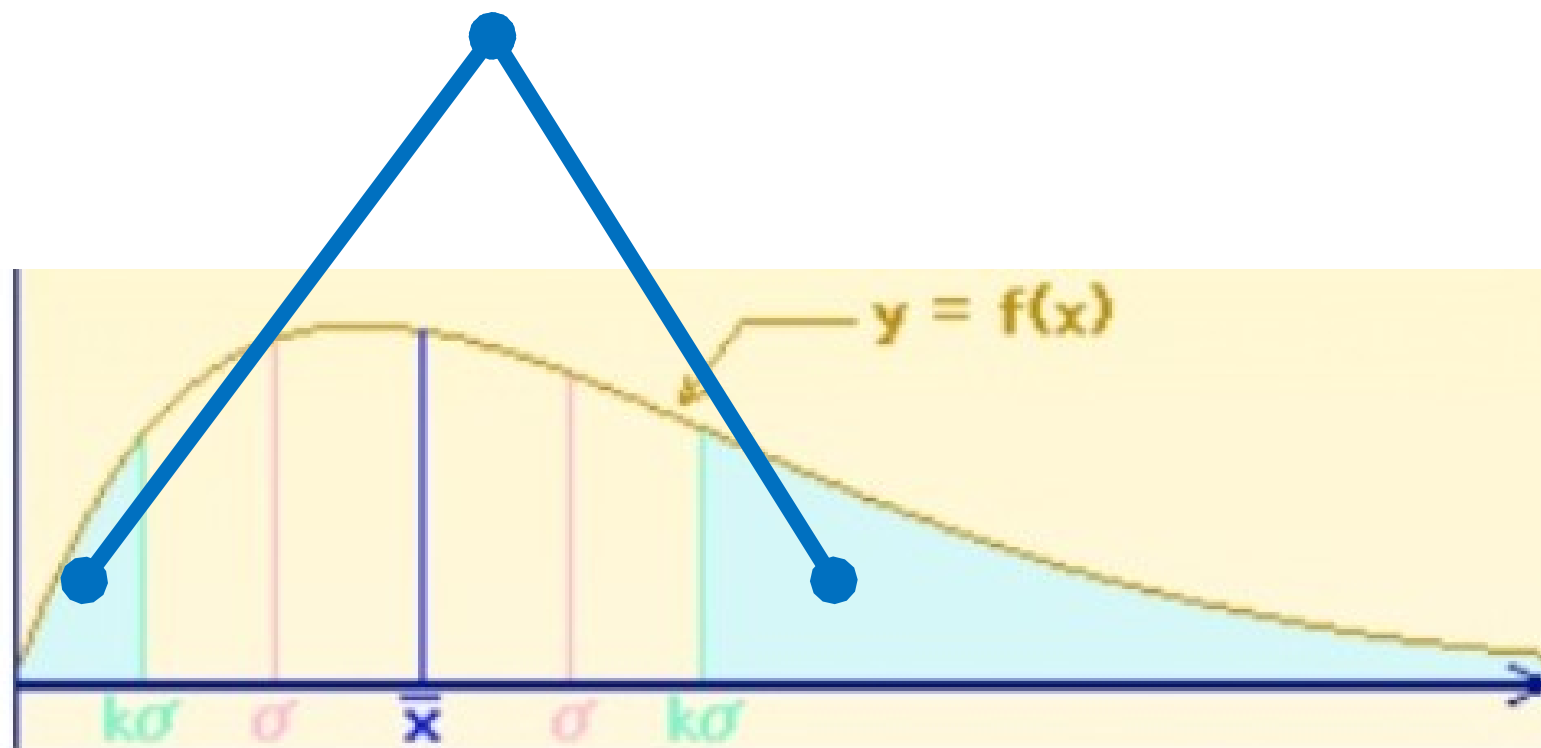
⇒ we cannot know the exact value as in case of normal dist.



Estimate of prob.

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

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

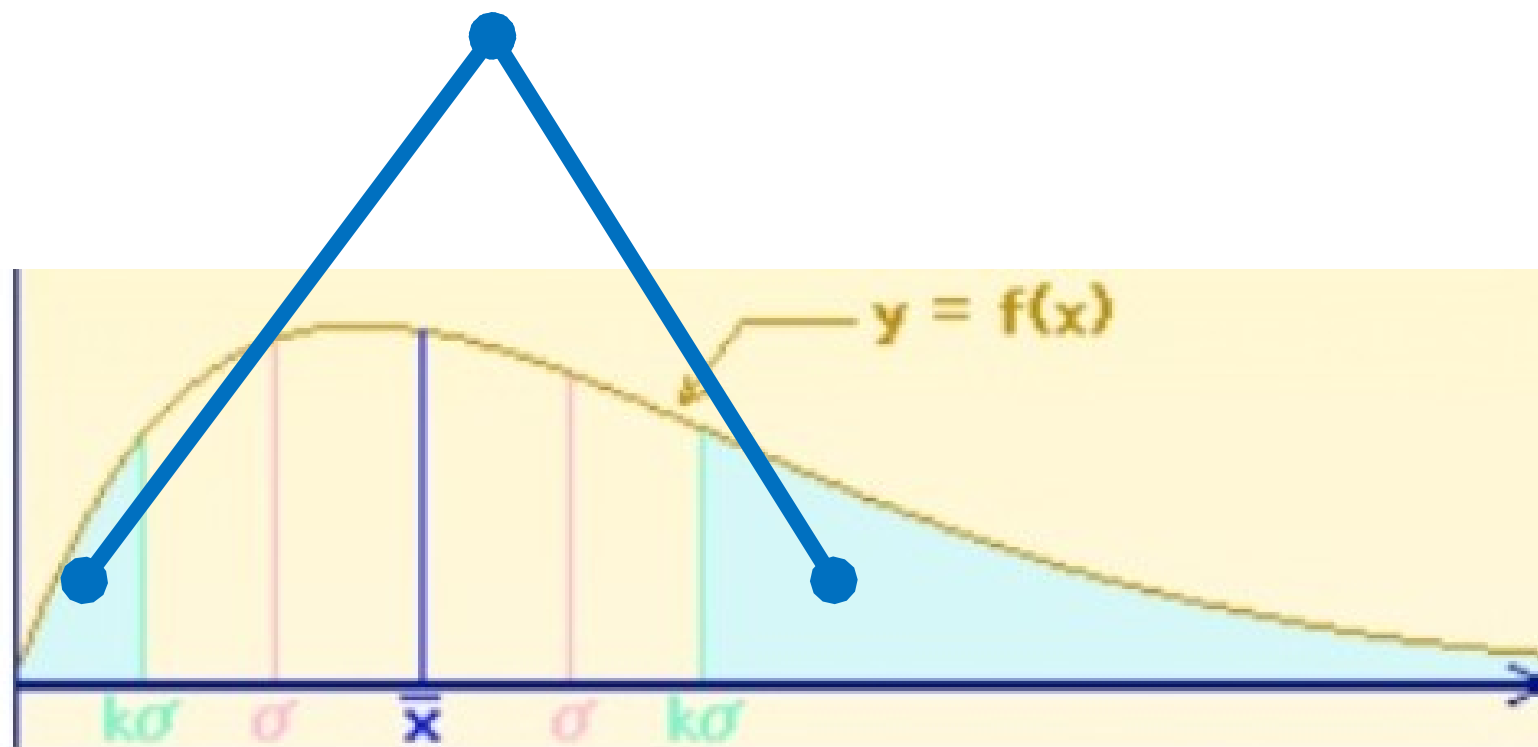


$$P(|X - \mu| \geq 2\sigma) \leq \frac{1}{4} = 0.25.$$

Estimate of prob.

For instance, when $k=2$,

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



Q. A

If a certain r.v. X satisfies $E(X) = 50$ and $V(X) = 25$, estimate $P(|X - 50| < 20)$
By using the Chebyshev's inequality.

(i) 0.8275

(ii) 0.8375

(iii) 0.925

(iv) 0.9375

Q. A 【Ans】

If a certain r.v. X satisfies $E(X) = 50$ and $V(X) = 25$, estimate $P(|X - 50| < 20)$
By using the Chebyshev's inequality.

(i) 0.8275

(ii) 0.8375

(iii) 0.925

(iv) 0.9375

I_10_E-A

The tables below show the temporal stock prices of two banks
For 5 days. Now, find the correlation of them.

Bank-A	Bank-B
847.1	5013
866.8	5095
863.4	5090
876.8	5177
881.9	5228

- (i) 0.853
- (ii) 0.912
- (iii) 0.933
- (iv) 0.979

I_10_E-A 【Ans】

The tables below show the temporal stock prices of two banks For 5 days. Now, find the correlation of them.

Bank-A	Bank-B
847.1	5013
866.8	5095
863.4	5090
876.8	5177
881.9	5228

SD of A:12.06

SD of B:74.68

Cov.:881.9

(i) 0.853

(ii) 0.912

(iii) 0.933

(iv) 0.979

I_10_E-B

Find the 3-days-moving average of following data for each.

Bank-A	Bank-B
847.1	5013
866.8	5095
863.4	5090
876.8	5177
881.9	5228

- (i) A 銀行: [859.1, 869, 874] B 銀行: [5066, 5120.67, 5165]
(ii) A 銀行: [879.1, 889, 874] B 銀行: [5069, 5220.67, 5165]
(iii) A 銀行: [859.1, 889, 894] B 銀行: [5066, 5220.67, 5265]
(iv) A 銀行: [859.1, 859, 854] B 銀行: [5066, 5320.67, 5365]

I_10_E-B

Find the 3-days-moving average of following data for each.

Bank-A	Bank-B
847.1	5013
866.8	5095
863.4	5090
876.8	5177
881.9	5228

- (i) A 銀行: [859.1, 869, 874] B 銀行: [5066, 5120.67, 5165]
- (ii) A 銀行: [879.1, 889, 874] B 銀行: [5069, 5220.67, 5165]
- (iii) A 銀行: [859.1, 889, 894] B 銀行: [5066, 5220.67, 5265]
- (iv) A 銀行: [859.1, 859, 854] B 銀行: [5066, 5320.67, 5365]

Agenda

- Examples of probability distribution or probability density
- Discrete probability distribution
 - Uniform distribution
 - Hypergeometric distribution
 - Binomial distribution
 - Poisson distribution
- Exercises

Examples of probability distribution / density

- Well known probability distributions for natural or socio-physical phenomena

- Pip of dice: Uniform distribution
- Resource survey: Hypergeometric distribution
- Summarization of questionnaire : Binomial distribution
- Accidents of airplanes
- Measurements of something: Normal distribution
- Waiting time: Exponential distribution
- Lifetime of a system: Gamma distribution
- Income and saving : Logarithmic normal distribution

Discrete

Continuous

Uniform distribution

- The probability distribution of a pip of a dice

- $f(x) = 1/6, \quad (x = 1, 2, \dots, 6)$

- More generally, suppose that you take out a ball from a box, which contains N balls, each of which has its own number ranging from 1 to N . Then, consider the probability distribution of the number of the ball that you have taken out.

- $f(x) = 1/N, \quad (x = 1, 2, \dots, N)$

Uniform distribution

- More generally, suppose that you take out a ball from a box, which contains N balls, each of which has its own number ranging from 1 to N . Then, consider the probability distribution of the number of the ball that you have taken out.

➤ $f(x) = 1/N, \quad (x = 1, 2, \dots, N)$

- Expected value : $E(X) = \sum x \cdot f(x) = (N+1) / 2$

- Variance : $V(X) = E(X^2) - \{E(X)\}^2$
 $= (N+1)(2N+1)/6 - \{(N+1)/2\}^2$
 $= (N^2 - 1)/12$

Proof of expected value and variance

r.v. (X)	1	2	3	...	N
Probability	1/N	1/N	1/N	...	1/N

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

Proof of expected value and variance

- Proof of $E(X^2) = (N+1)(2N+1)/6$

r.v. (X^2)	1^2	2^2	3^2	...	N^2
Probability	$1/N$	$1/N$	$1/N$...	$1/N$

$$\begin{aligned}
 E[X^2] &= 1^2 \times \frac{1}{N} + 2^2 \times \frac{1}{N} + \dots + N^2 \times \frac{1}{N} \\
 &= \frac{1}{N} \times \left\{ 1^2 + 2^2 + \dots + N^2 \right\} \\
 &= \frac{1}{N} \times \frac{N(N+1)(2N+1)}{6} \\
 &= \frac{(N+1)(2N+1)}{6}.
 \end{aligned}$$

Hypergeometric distribution

- There are two attributes, A and B. You have N materials that consist of M materials of attribute A, and (N-M) materials of attribute B. Now, suppose that you take out n materials from this population, and regard the number x of materials of attribute A (Ofcourse, then the number of attribute B is $(n-x)$).
- Expected value : $E(X) = n (M / N)$
- Variance : $V(X) = n \{M (N - M) / N^2\} \{ (N - n) / (N - 1) \}$
- The ratio of attribute A is $p = M / N$. Then, as $N \rightarrow \infty$,
 - Expected value : $E(X) = n p$
 - Variance : $V(X) = n p (1 - p)$

Hypergeometric distribution

- Suppose that there are 1000 fish in a lake, 200 of which has red marks. Now, if you catch 5 fish from this lake, find the probability that the number of marked fish is
- (i) 0 (ii) 1.

Hypergeometric distribution 【Answer】

- Suppose that there are 1000 fish in a lake, 200 of which has red marks. Now, if you catch 5 fish from this lake, find the probability that the number of marked fish is
- (i) 0 (ii) 1.

$$f(x) = \frac{{}_M C_x \cdot {}_{N-M} C_{n-x}}{{}_N C_n}$$

$$N = 1000, M = 200, n = 5$$

$$f(0) = \frac{{}_{800} C_5}{{}_{1000} C_5} = 0.32686$$

$$f(1) = \frac{{}_{200} C_1 \cdot {}_{800} C_4}{{}_{1000} C_5} = 0.41063$$

Used for the resource survey.

Binomial distribution

- Suppose that a certain trial has two results (say, S and F, for instance). Each result occurs with probability p and $(1-p)$.
- If you repeat such trials independently n times under the same condition, it is called as the

Bernoulli trial.

➤ The probability of S happens x times , and also F happens $(n - x)$ times :

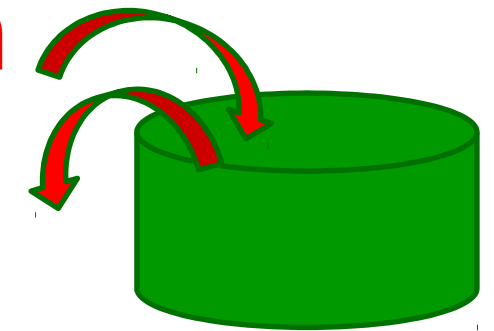
➤ $f(x) = {}_n C_x p^x (1-p)^{n-x}$

- $E(X) = np$

- $V(X) = np(1-p)$

Difference of hypergeometric and binomial distributions

- Similar (ratio of specific attributes in a large population)
- Hypergeometric distribution corresponds to **sampling without replacement**.
- Binomial distribution corresponds to **sampling with replacement**.



- ※Sampling without replacement : taking out N samples from population without returning them into the population.
- ※Sampling with replacement : taking out N samples from population returning them into the population every time

Difference of hypergeometric and binomial distributions

- However, the larger the population becomes, the closer the results of both samplings (and consequently, the difference of hypergeometric and binomial distributions tends to zero).

Example of binomial distribution

- Suppose that there are a large amount of fish in a lake.

The fish with red marks account for the ratio of 0.2 among the whole fish.

Now, if you catch 5 fish from this lake, find the probability that the number of marked fish is

- (i) 0 (ii) 1.

Example of binomial distribution 【Answer】

- Suppose that there are a large amount of fish in a lake.

The fish with red marks account for the ratio of 0.2 among the whole fish. Now, if you catch 5 fish from this lake, find the probability that the number of marked fish is

- (i) 0 (ii) 1.

$$f_p(x) = {}_n C_x p^x (1-p)^{n-x} = {}_5 C_0 (0.2)^0 (0.8)^5 = 0.32768$$

$$f(1) = {}_5 C_1 (0.2)^1 (0.8)^4 = 0.40960$$

$$f(2) = {}_5 C_2 (0.2)^2 (0.8)^3 = 0.20480$$

$$f(3) = {}_5 C_3 (0.2)^3 (0.8)^2 = 0.05120$$

$$f(4) = {}_5 C_4 (0.2)^4 (0.8)^1 = 0.00640$$

$$f(5) = {}_5 C_5 (0.2)^5 (0.8)^0 = 0.00032$$

Close to the values of hypergeometric distribution.

Binomial dist. with python

```
import numpy as np
from scipy.stats import binom

p=0.2
N=5
binom.pmf(0, N, p)
```

0.32768

```
import numpy as np
from scipy.stats import binom

p=0.2
N=5
binom.pmf(1, N, p)
```

0.4095999999999999

```
import numpy as np
from scipy.stats import binom

p=0.2
N=5
binom.pmf(2, N, p)
```

0.20479999999999998

Exercise

- A lot of electronic components are placed in a box, which include defective goods with the ratio of 1%. Now, if you take out 5 components through the sampling with replacement, find the probability that the number of defective goods are 2 or less.

Exercise 【Answer】

- A lot of electronic components are placed in a box, which include defective goods with the ratio of 1%. Now, if you take out 5 components through the sampling with replacement, find the probability that the number of defective goods are 2 or less.

You should apply the binomial distribution here.

Then, the probability that x defective goods are found in the sample of n is $f(x) = {}_n C_x p^x (1-p)^{n-x}$

$$f(0) = {}_5 C_0 (0.01)^0 (0.99)^5$$

$$f(1) = {}_5 C_1 (0.01)^1 (0.99)^4$$

$$f(2) = {}_5 C_2 (0.01)^2 (0.99)^3$$

$$f(0) + f(1) + f(2) = \underline{0.9999901}$$

Python code

```
import numpy as np
from scipy.stats import binom

p=0.01
N=5
binom.pmf(0, N, p)+binom.pmf(1, N, p)+binom.pmf(2, N, p)

0.9999901494000001
```

【 reference 】

In the practical situation, you will be required to estimate the ratio of the defective goods, rather than the number of them in samples.

In that case, the desired ratio is denoted as θ , and then you will estimate the value or conduct hypothesis testing through the results of samplings.

Maximum likelihood estimate / hypothesis testing

Exercise

In a certain elementary school, the ratio of 0.6 students suffer from tooth sick among the 715 students.

Now, if you capture 10 students randomly, find the probability that just 2 students of them suffer from tooth sick.

Exercise 【Answer】

In this case, you should apply the hypergeometric distribution.

Let the set of tooth sick students be “set-A”, and others be “set-B”. Then, the size of set-A is $M=715*0.6 = 429$,
And the total students $N=715$. Of course, the size of set-B is

$$N-M = 286.$$

The sicked students in the sample is $x=2$.

$$\begin{aligned}
 \triangleright f(x) &= {}_M C_x \cdot {}_{N-M} C_{n-x} / {}_N C_n \\
 &= {}_{429} C_2 \cdot {}_{286} C_8 / {}_{715} C_{10} = \underline{0.01}
 \end{aligned}$$

Python code

```
import numpy as np  
from scipy.stats import hypergeom
```

```
m=715*0.6  
N=715  
k=10  
hypergeom.pmf(2, N, m, k)
```

```
0.01022213405993778
```

【 reference 】

In the practical situation, you will be required to estimate the ratio of students suffering from tooth sick.

Then, from the results of the sampling survey, you can estimate the desired ratio.

-> Maximum likelihood estimate / hypothesis testing

Poisson distribution

- In the binomial distribution, suppose that n is large and p is small at the same time.

➤ Ex) Probability of a contract formation concerning the real estate.

Suppose $p = 0.002$, and find the probability that 3 contracts hold:

For $f(x) = {}_n C_x p^x (1-p)^{n-x}$,

set $n = 1000$, $p = 0.002$, but the calculation of ${}_{1000} C_3 (0.002)^3 (0.998)^{997}$ is inefficient.

Since $E(X) = np = 2$, the probability of $x = 0, 1, 2, 3$ may not be so small.

Poisson distribution

Consider the case $n \rightarrow \infty$ and $p \rightarrow 0$ so that $np \rightarrow \lambda$.

Then, for each x , the following statement holds
(Poisson's Law of Small Numbers).

$${}_nC_x p^x (1-p)^{n-x} \rightarrow e^{-\lambda} \cdot \lambda^x / x!$$

$$f(x) = e^{-\lambda} \cdot \lambda^x / x! \quad (\lambda > 0, x = 0, 1, 2, \dots)$$

: Poisson distribution denoted as $Po(\lambda)$.

$$f(3) = e^{-2} \cdot 2^3 / 3! = 0.180447$$

- $E(X) = \lambda$ ($\doteq np$)
- $V(X) = \lambda$ ($\doteq np(1-p)$)

Poisson distribution depends **only on λ** .

【 reference 】 Expected value of Poisson distribution

Below, we use x instead of k for simplicity.

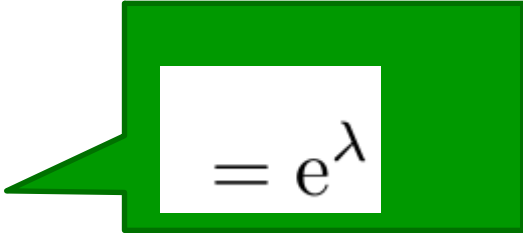
Applying Taylor's expansion of e^λ , we have:

$$E[X] = \sum_{k=1}^{\infty} k \times e^{-\lambda} \frac{\lambda^k}{k!}$$

$$= \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-1)!}$$

$$= \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^{k+1}}{k!}$$

$$= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$


$$= e^\lambda$$

$$= \lambda.$$

Examples of Poisson distribution

- Applied for the risk management and security issues.
 - Number of traffic incidents,
 - Ratio of defective goods,
 - Number of huge disasters,
 - ...

Exercise

In a certain store, 4 customers arrive every one hour on average. Under the assumption that the arrival of customers is subject to the Poisson distribution, find the probability that 3 customers will arrive in one hour.

Exercise 【Answer】

In a certain store, 4 customers arrive every one hour in average. Under the assumption that the arrival of customers is subject to the Poisson distribution, find the probability that 3 customers will arrive in one hour.

$\lambda=4$ [customers/hour] is known.

Thereby,

$$f(3) = e^{-4} \cdot 4^3 / 3! = \underline{0.195}$$

```
from scipy.stats import poisson  
lamb=4  
poisson.pmf(3,lamb)
```

0.19536681481316454

Summary

- You have learned the discrete distributions.
 - Uniform distribution,
 - Hypergeometric distribution,
 - Binomial distribution, and
 - Poisson distribution.
- You have also learned the derivations of expected value and variance.

Points (checklist)

- You can state the probability distributions of the following discrete distributions?
 - Uniform distribution,
 - Hypergeometric distribution,
 - Binomial distribution, and
 - Poisson distribution.
- You can also state the expected values and variance of the distributions above?