Statistics and data analysis I

Week 11

“From descriptive statistics to inferential statistics”

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-
$\boxed{\text{Week12: Hypothesis test(1) -Introduction, and distributions of test statistic (t-distribution)-}}$
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. Law of large numbers
2. Population mean and sample mean
3. Interval estimation on population mean

Law of large numbers

Example

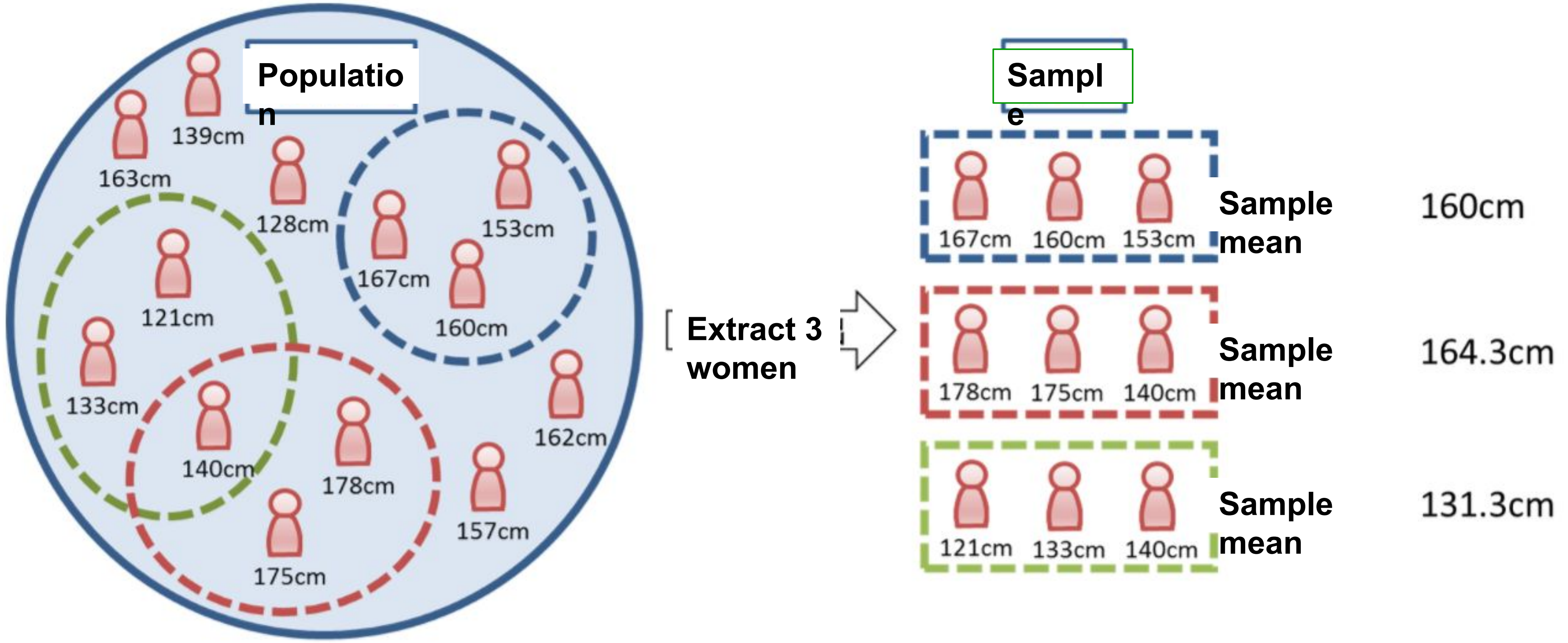
- Consider : "find the mean height of women in Japan."

- ⇒ Unable to measure the height of all women in Japan.
- Estimation based on sample(s)

- But… the sample mean matches with the population mean?

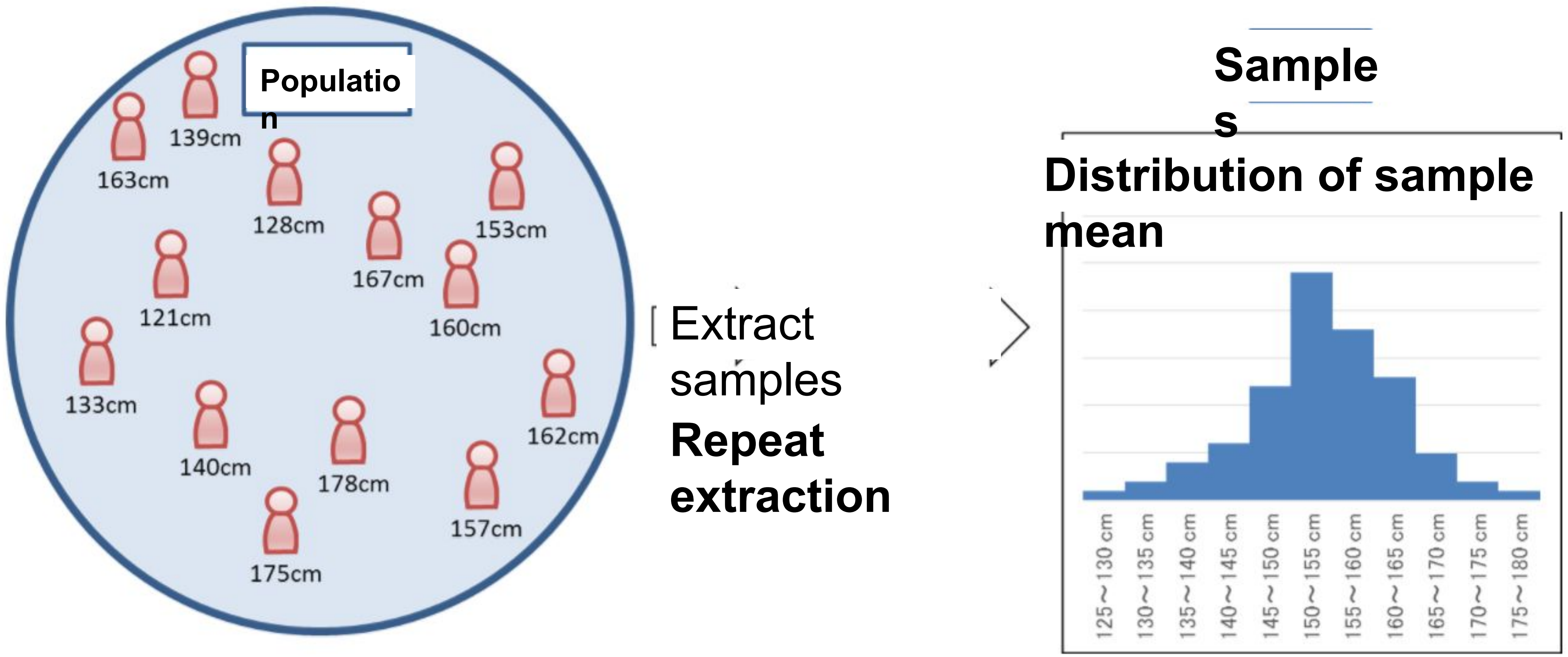
Example

Ex) Now, consider the population of 13 women.
You extract 3 women from them, and find the sample mean. ⇒ Depends on samples! Various sample means!



Population
mean :152cm

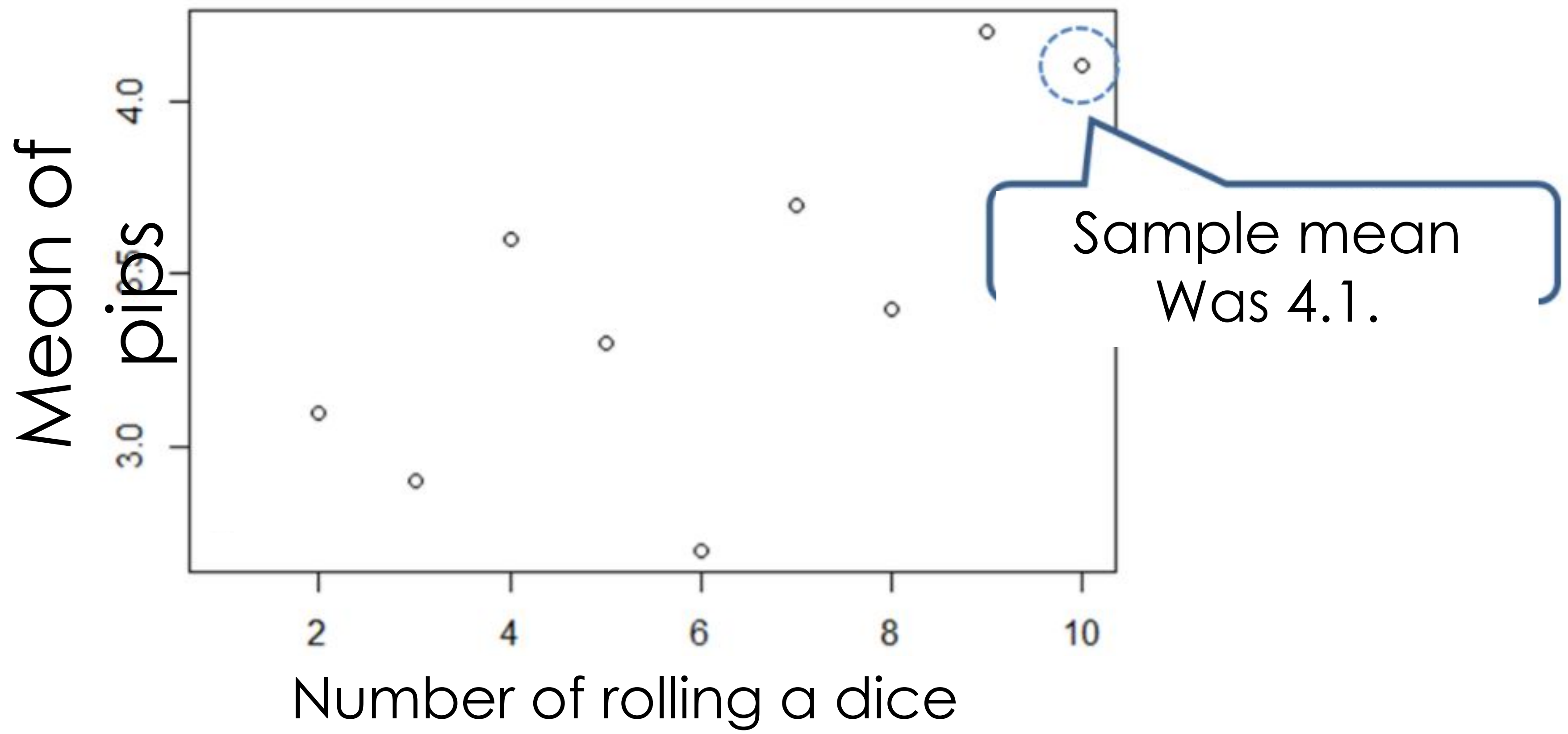
Ex. Sample mean does not match the population mean in general. It distributes around the population mean.
= Sample distribution.
Sample distribution has some features, from which we can estimate the population mean.



Law of large numbers

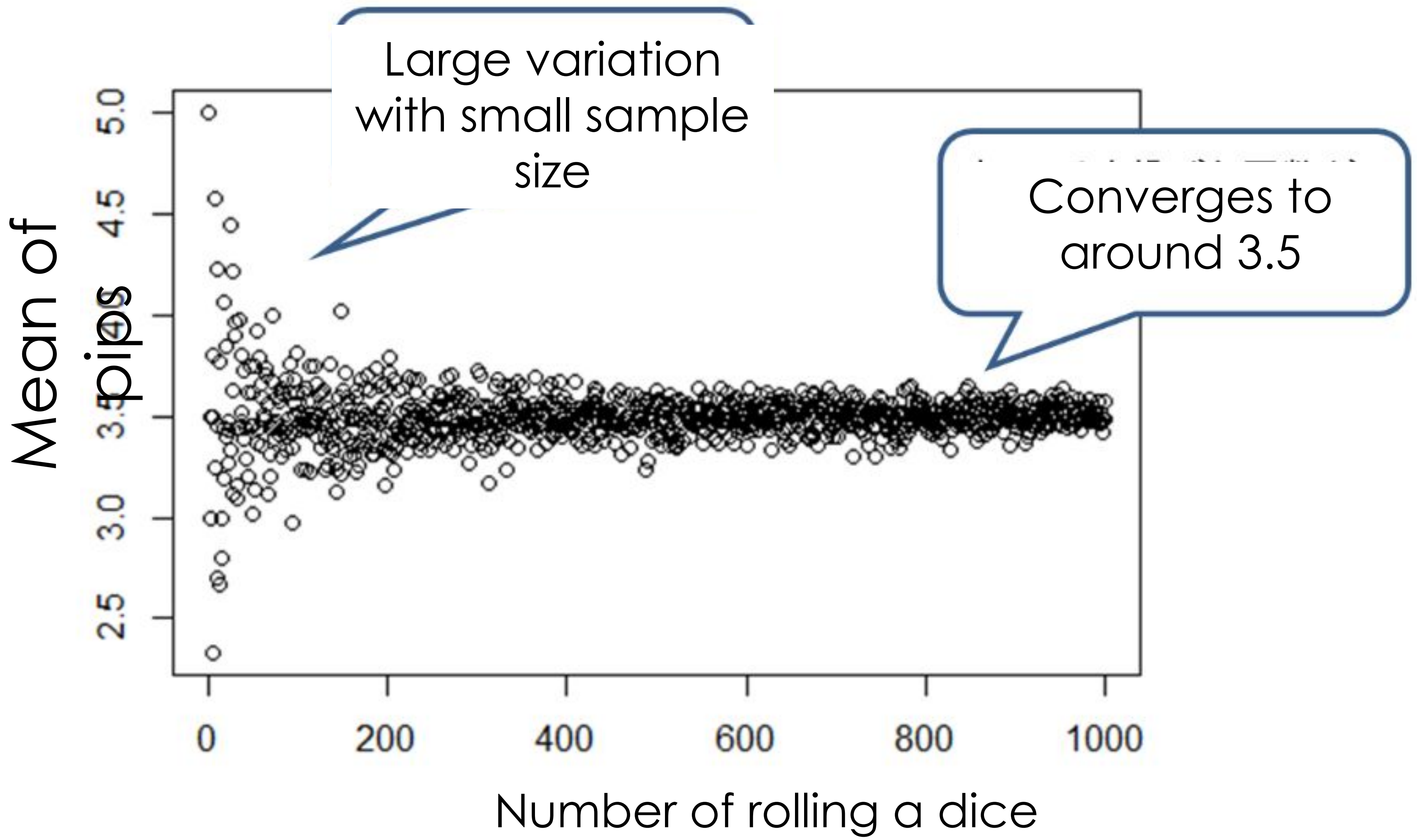
Consider the mean of the pips when you roll a dice many times. The expected value should be 3.5.
But how about the “sample mean”..?

The mean of 10 samples was 4.1。。。



Making sample size larger

Make the sample size larger. If you roll the dice up to 1000 times, the sample means were plotted as below.



Law of large numbers

「If you extract samples from the population with mean μ, the sample mean converges to μ as the sample size gets larger.」

In case of coin tossing:
「In an event occurs with probability *p*, then the ratio of the occurrence of that event tends to *p* as the number of trials gets larger.」

The larger the sample size is, the more accurately you can estimate the population.

Population and sample

Population mean / Population variance / Population SD

- The mean of population, noted as μ, is the population mean

- The measure of variation of population is population variance / population SD
- Population SD $\sigma = (\text{Population variance } \sigma^2)^{1/2}$

Sample mean

- Sample mean is the mean of sample elements.
- Distinguished from population mean.
  - □ Sample mean＝(Sum of samples)÷(sample size)
- Sample size means the number of sample elements.
- By taking sample means, you can remove the bias, and can obtain the value closer to the actual population mean.

⇒ Law of large numbers

Simulation with R

- Sample means of samples that follow the normal dist. with expected value and SD of unity.

```
> mean(rnorm(10,1,1))
[1] 0.9790969
> mean(rnorm(10,1,1))
[1] 1.08963
> mean(rnorm(10,1,1))
[1] 0.7131717
> mean(rnorm(10,1,1))
[1] 0.5656154
> mean(rnorm(10,1,1))
[1] 1.424441
> mean(rnorm(10,1,1))
[1] 1.407049
> mean(rnorm(10,1,1))
[1] 1.30749
> mean(rnorm(10,1,1))
[1] 0.9489847
```

Quiz for mean

- Based on the observed values that follow the normal dist., estimate the configuration of the expected value of R.

```
> rnorm()
 [1]  0.4443899  0.2807700  1.0972959 -0.5151003 -0.9606174  0.7029789
 [7]  0.5633878 -0.2720555  0.1963411 -0.6494952  0.4248215 -0.3468013
[13] -0.2301140  0.2116248 -1.9480825
```

Quiz for mean

- Not pointwise, but interval estimation is allowed for now.

Quiz for mean

- Mr. A : "The expected value lies in (-1,1)."

Quiz for mean

- Mr. B : "The expected value lies in (-100,100)."

Quiz for mean

- Which seems likely to be correct?

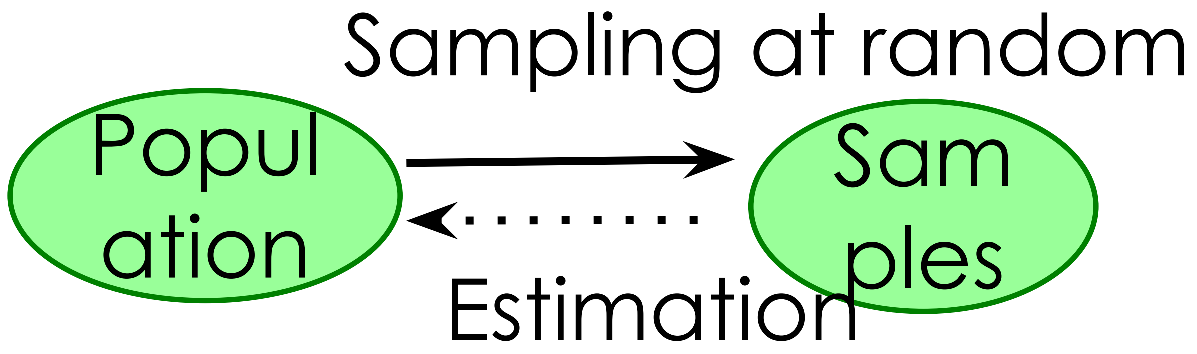
Quiz for mean

- The wider, the more likely to include the actual mean.
- But not effective, if too wide.
- → ‘Suitable’ interval estimation？

2. Interval estimate of population mean

Interval estimation on population mean

Based on the observed samples, estimate the ‘actual population mean’.



Ex) Suppose that you’ve observed 5 sample elements of random variables that follow the normal dist., ”1.4, 2.2, 3.0, 4.2, 5.3” generated by a random generator.

Now, find the 95% CI (=confidence interval) of population mean of this random generator.

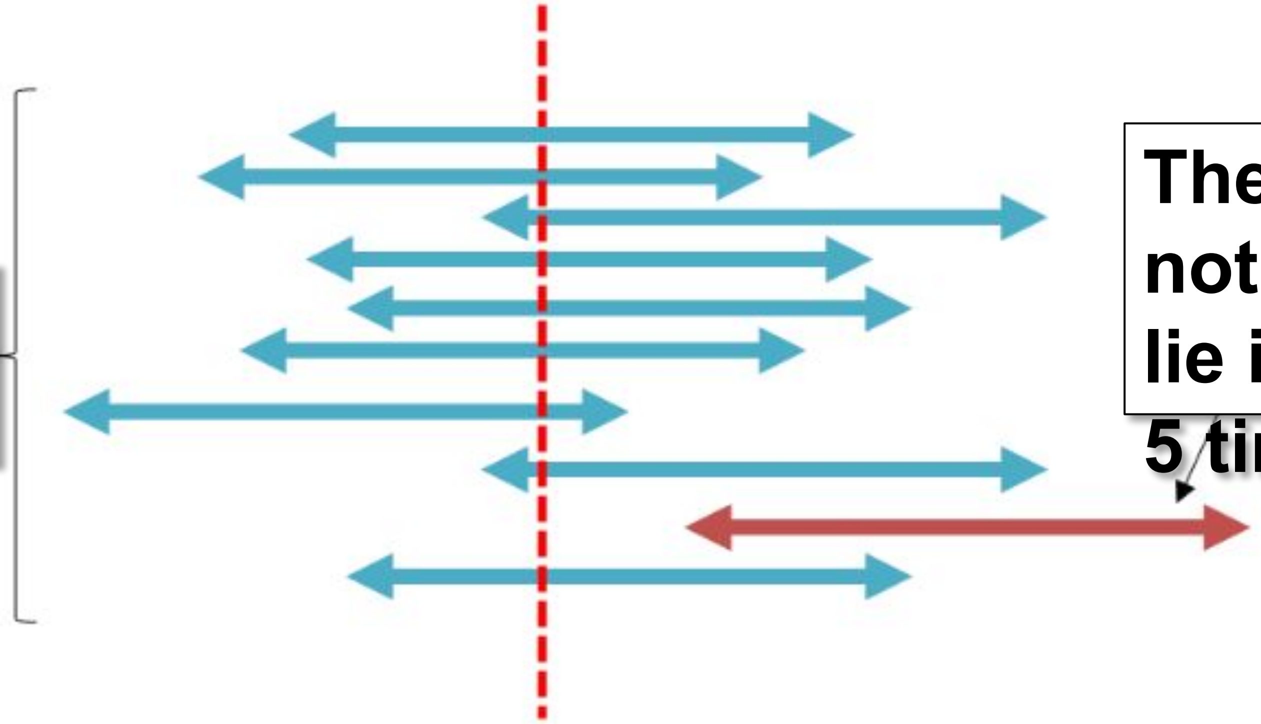
95%CI

Not possible to estimate with 100% accuracy.
Then, you can answer with the form of the interval

“the population mean lies in (1.2, 5.4)“

CI

Extract samples from population, and then find the CI. If you repeat this procedure 100 times, then, the population mean lies
In the estimated interval 95 times.



95% CI (in case population SD is known)

- Solve below.

$$-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96$$

$$\Downarrow$$

$$\bar{x} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{1.96\sigma}{\sqrt{n}}$$

CI for population mean

2 cases.

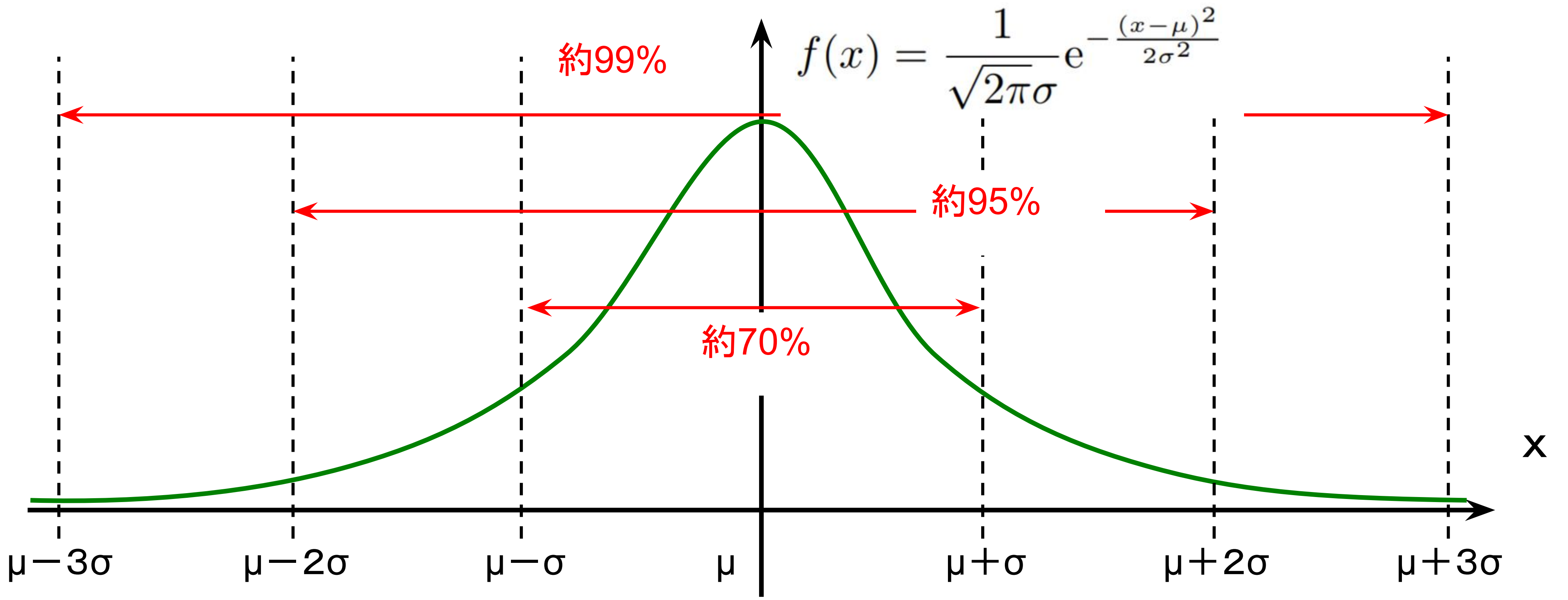
①In case population SD is known.
⇒ Normal disdt.

②In case population SD is unknown.
⇒ t-dist.

In case population SD is known.

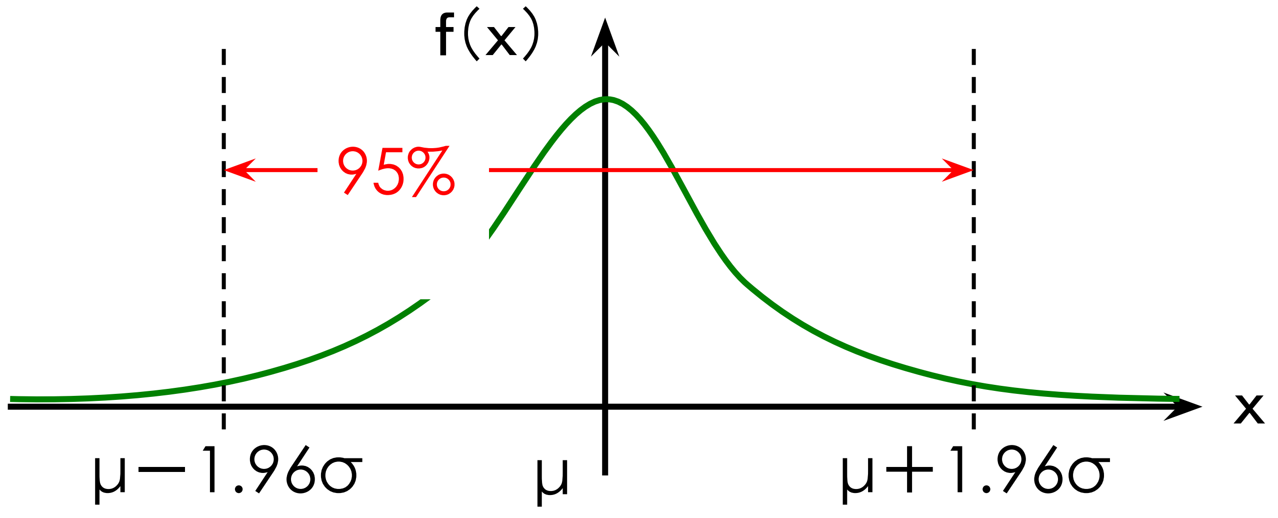
Estimate on the basis of normal distribution

- Estimate the "correct answer" by using the probability distribution
  - If it's subject to the normal distribution…
  - You may use the sample mean as the estimate of the population mean.
  - With the interval "xxx or more , yyy or less"



95% interval of normal distribution

- In statistics, the accuracy of “95%” or “99%” are used frequently.
  - Conversely, the estimation might be wrong with the probability of 5% (or 1%, respectively).
  - Set the 95% interval around μ



$$\mu - 1.96\sigma \leqq x \leqq \mu + 1.96\sigma$$

$$-1.96 \leqq \frac{x-\mu}{\sigma} \leqq +1.96$$

95% Confidence interval

- Ex) Let us measure the temperature of a certain liquid with a thermometer, that is not so accurate. It is known that the measured value is subject to N(μ,5^2). Now, estimate the 95% confidence interval under the situation the measured temperature is 20℃.

To be estimated.
(Actual temperature : unknown)

N(μ,5^2)

Observed temperature: 20℃

For actual μ,

$$-1.96 \leq z = \frac{x-\mu}{\sigma} \leq 1.96$$

$$-1.96 \leqq \frac{x-\mu}{\sigma} \leqq +1.96$$

$$10.2 \leqq \mu \leqq 29.8$$

Features of sample mean from the normal population

- Let $\bar{X}$ be the sample mean from the normal population (∼N(μ,σ)). Note that $\bar{X}$ is again a r.v., and is still subject to the normal distribution.
- The expected value of $\bar{X}$ is μ. But the S.D. becomes $\frac{\sigma}{\sqrt{n}}$

- 95% confidence interval based on the sample mean from the normal population:

$$-1.96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +1.96$$

Interval estimation of population mean (In case σ is known)

- You should use:

$$-1.96 \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq +1.96$$

- Rewriting this, we have

$$\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Interval estimation of population mean (In case σ is known)

$$\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

- The width of CI is $\dfrac{3.92\sigma}{\sqrt{n}}$

If you like to reduce the width W or less,

$$\frac{3.92\sigma}{\sqrt{n}} \leq W$$

i.e.,

$$n \geq \left(\frac{3.92\sigma}{W}\right)^2$$

should hold (take such n).

In case population SD is unknown

Sample mean from normal dist.(In case population SD σ is unknown)

- We have :

$$\bar{x} - \frac{S t_{n-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{S t_{n-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}}$$

- Here, S is the unbiased SD. Squared root of :

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{x})^2}{n-1}$$

- $t_{n-1}\left(\frac{\alpha}{2}\right)$: upper a/2*100-percentile of t-dist. with degree of freedom (=df) n-1 (a=0.05 for now.)

CI (In case population SD σ is unknown)

- 95% CI:

$$\bar{X} - t_{n-1}\left(\frac{0.05}{2}\right) \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1}\left(\frac{0.05}{2}\right) \times \frac{S}{\sqrt{n}}$$

Upper2.5-percentile of t-dist.

- In case σ^2 is known:

$$\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Upper2.5-percentile of z-dist.

- 2 diff. ; upper 2.5-percentile, and SD.

Example.

- A certain sphygmomanometer returns the measured value that follows the normal dist. With mean of the actual blood pressure, and SD of 6.
-
- Now, given the measured value of this sphygmomanometer 120, find the 95% CI of your blood pressure.

Example【ans】

$$-1.96 \leqq \frac{x-\mu}{\sigma} \leqq +1.96$$

$120+1.96\times 6 \geqq \mu \geqq 120-1.96\times 6$ So,

$108.24 \leqq \mu \leqq 131.76$

Python

Week12_Exercise4.ip

```
import numpy as np
import scipy.stats as st

x=np.array([120])

#Sample size.
n=x.size

#Sample mean.
x_mean=x.mean()

#Population SD.
x_sd=6

st.norm.interval(alpha=0.95,loc=x_mean,scale=x_sd/np.sqrt(n))
```

```
(108.24021609275968, 131.75978390724032)
```

Exercise-5

- A certain sphygmomanometer returns the measured value that follows the normal dist. With mean of the actual blood pressure, and SD of 6.
-
- Now, given the measured value of this sphygmomanometer 120 and 130, find the 95% CI of your blood pressure.

Exercise-5【Ans】

Week12_Exercise5.ipynb

```
import numpy as np
import scipy.stats as st

x=np.array([120,130])

#Sample size.
n=x.size

#Sample mean.
x_mean=x.mean()

#Population SD.
x_sd=6

st.norm.interval(alpha=0.95,loc=x_mean,scale=x_sd/np.sqrt(n))
```

```
(116.68457705390193, 133.31542294609807)
```

Example-2 : Ci under unknown population SD

- In a certain laboratory, they like to know the actual PH-value of a certain solution. Now, the results of 5 times' measurements were:
- 7.86, 7.89, 7.84, 7.90, 7.82.
- The population SD is unknown. Then, find the 95% CI.

Example-2【ans】

測定回	1	2	3	4	5
pH	7.86	7.89	7.84	7.90	7.82
偏差	-0.002	0.028	-0.022	0.038	-0.042
偏差^2	0.000004	0.000784	0.000484	0.001444	0.001764

Mean：7.862

Sum of squared diff.：0.00448

Example-2【ans】

Thus, the unbiased variance is

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}$$ =0.00448/4 = 0.00112

S=0.033

Sample size is N=5, so t-dist. of df=4.

$$t_4(0.025) = 2.776$$

$$7.862 - t_4(0.025) \times 0.033/\sqrt{5} \leq \mu \leq 7.862 + t_4(0.025) \times 0.033/\sqrt{5}$$

7.82≦μ≦7.90

Example-2【ans】

- Look at the column of '2.5%' (colored column)

	有意確率								
	0.10	0.05	0.01	0.001	両側	0.10	0.05	0.01	0.001
df	0.05	0.025	0.005	0.0005	片側	0.05	0.025	0.005	0.0005
1	6.3138	12.706	63.657	636.62	18	1.7341	2.1009	2.8784	3.922
2	2.9200	4.3027	9.9248	31.598	19	1.7291	2.0930	2.8609	3.883
3	2.3534	3.1825	5.8409	12.941	20	1.7247	2.0860	2.8453	3.850
4	2.1318	$\boxed{2.7764}$	4.6041	8.610	21	1.7207	2.0796	2.8314	3.819
5	2.0150	2.5706	4.0321	6.859	22	1.7171	2.0739	2.8188	3.792
6	1.9432	2.4469	3.7074	5.959	23	1.7139	2.0687	2.8073	3.767
7	1.8946	2.3646	3.4995	5.405	24	1.7109	2.0639	2.7969	3.745
8	1.8595	2.3060	3.3554	5.041	25	1.7081	2.0595	2.7874	3.725
9	1.8331	2.2622	3.2498	4.781	26	1.7056	2.0555	2.7787	3.707
10	1.8125	2.2281	3.1693	4.587	27	1.7033	2.0518	2.7707	3.690
11	1.7959	2.2010	3.1058	4.437	28	1.7011	2.0484	2.7633	3.674
12	1.7823	2.1788	3.0545	4.318	29	1.6991	2.0452	2.7564	3.659
13	1.7709	2.1604	3.0123	4.221	30	1.6973	2.0423	2.7500	3.646

Example-2【ans】

Pytho

Week12_Example2.ip

```
import numpy as np
import scipy.stats as st

x=np.array([7.86, 7.89, 7.84, 7.90, 7.82])

#Sample size.
n=x.size

#Sample mean.
x_mean=x.mean()

#Unknown SD.
x_sd=np.std(x,ddof=1)

st.t.interval(alpha=0.95,df=n-1,loc=x_mean,scale=x_sd/np.sqrt(n))
```

```
(7.820445974652658, 7.903554025347343)
```

Summary

- Introduction to the inferential statistics:
  - Hypothesis testing
  - 95% confidence interval（interval estimation）
  - Estimation of population mean from sample mean

Summary【checklist】

- You can state the law of large numbers?
- Can state the nature of sample mean and variance quantitatively?
- You can construct the 95% confidence interval under known/unknown population SDs?

【Appendix】On t-distribution

T-distribution

- 【Ex】
- Let r.v.s X_1, X_2, …,X_n be independent with each other, and subject to N(μ, σ^2). Then, the quantity

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

is subject to t-distribution of (n-1) degree of freedom. Here,

$$s = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1}}$$

t-distribution

- In other words…
- Let Z∼N(0, 1) and W be subject to χ^2-distribution of n degree of freedom. We also assume that they are independent of each other. Then, the following quantity is subject to t-distribution of n degree of freedom.

$$t = \frac{Z}{\sqrt{\frac{W}{n}}}$$

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S^2}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2}}{n-1}}} = \frac{Z}{\sqrt{\frac{W}{n-1}}}$$

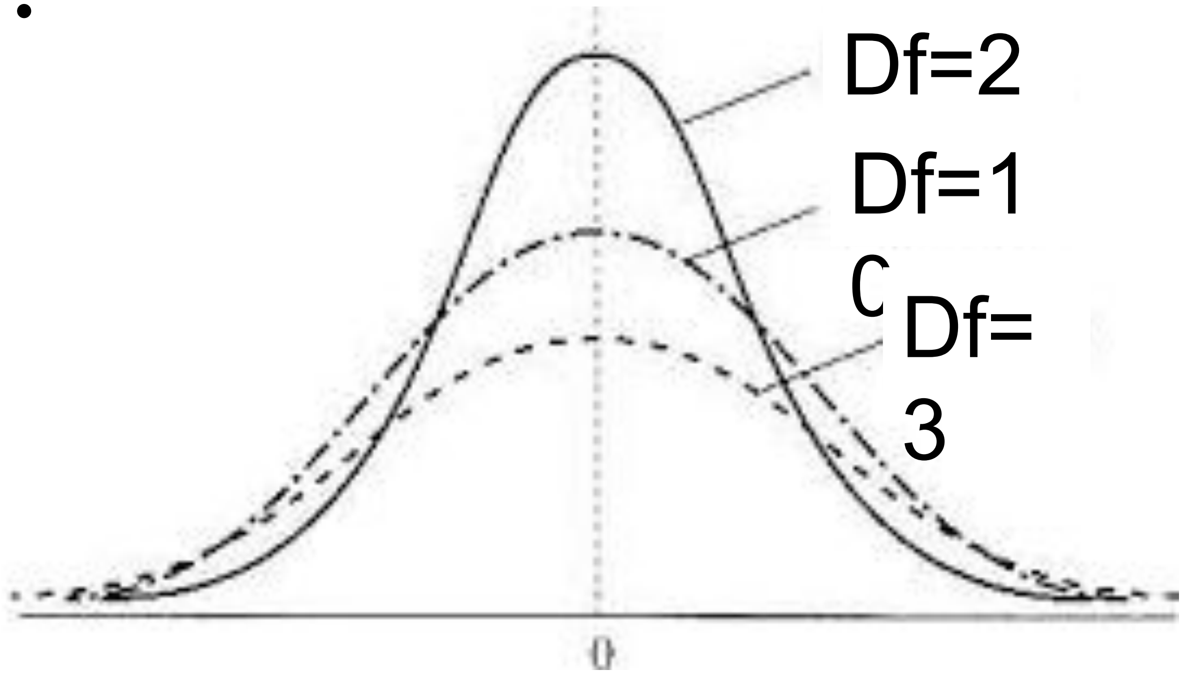
t-distribution

- T-distribution has degree of freedom.
- Used for the interval estimation / hypothesis testing of population mean.
- 【Probability density】
- The probability density of t-distribution of n degree of freedom is

$$f(x;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}\Gamma\left(\frac{n}{2}\right)}$$

t-distribution

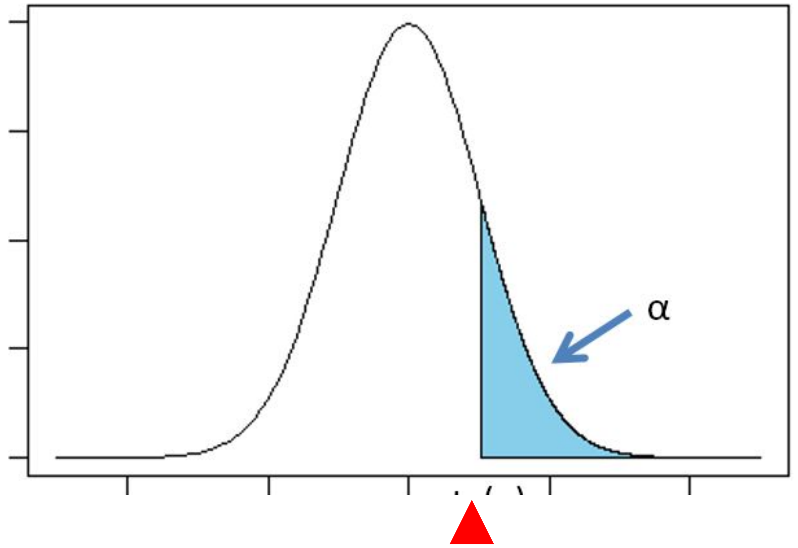
- Probability density of t-distributions of various degree of freedom (df).



- Symmetric with respect to x=0 (as z-dits.) !
- Asymptotically tends to z-dist. as df=n→∞.
- If df=n is n large (n≧30, for instance), can be regarded as z-dist.

Percentile of t-distribution

- We denote t-distribution of n degree of freedom as t_n hereafter.
- It's upper a-percentile is denoted as $t_n(a)$.
  - Ex:
  - Upper 5-percentile of t-distribution of df=5.



5%点 $t_5(0.05)=2.015$

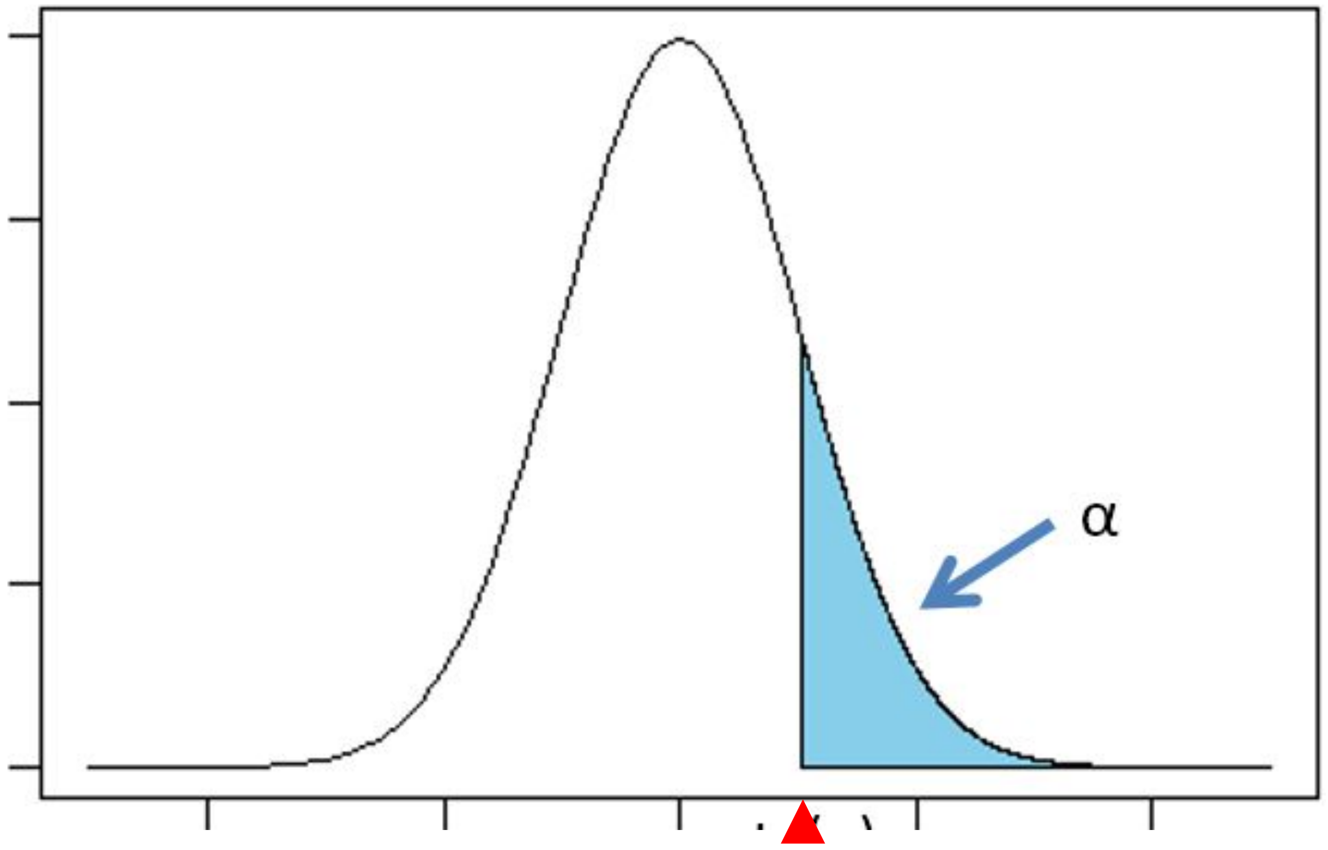
How to find percentile?

- Tables(z-table / t-table)

- Python

Percentile of t-distribution

- For instance, the upper 2.5-percentile of t-distribution of df=5 is about 2.57.



$t_5(0.025)=2.57$

- Using python;
  - E.g.) Find the upper 2.5-percentile of t-dist. With df=9.

```
from scipy.stats import t
t.ppf(0.975,9)
```

```
2.262157162740915
```

T-table

- In making 95% confidence interval, upper 2.5-percentile is needed.
- Therefore, you should look into the pink column of "2.5%" in "one-side(片側)"

	有意確率								
	0.10	0.05	0.01	0.001	両側	0.10	0.05	0.01	0.001
df	0.05	0.025	0.005	0.0005	片側	0.05	0.025	0.005	0.0005
1	6.3138	12.706	63.657	636.62	18	1.7341	2.1009	2.8784	3.922
2	2.9200	4.3027	9.9248	31.598	19	1.7291	2.0930	2.8609	3.883
3	2.3534	3.1825	5.8409	12.941	20	1.7247	2.0860	2.8453	3.850
4	2.1318	2.7764	4.6041	8.610	21	1.7207	2.0796	2.8314	3.819
5	2.0150	2.5706	4.0321	6.859	22	1.7171	2.0739	2.8188	3.792
6	1.9432	2.4469	3.7074	5.959	23	1.7139	2.0687	2.8073	3.767
7	1.8946	2.3646	3.4995	5.405	24	1.7109	2.0639	2.7969	3.745
8	1.8595	2.3060	3.3554	5.041	25	1.7081	2.0595	2.7874	3.725
9	1.8331	2.2622	3.2498	4.781	26	1.7056	2.0555	2.7787	3.707
10	1.8125	2.2281	3.1693	4.587	27	1.7033	2.0518	2.7707	3.690
11	1.7959	2.2010	3.1058	4.437	28	1.7011	2.0484	2.7633	3.674
12	1.7823	2.1788	3.0545	4.318	29	1.6991	2.0452	2.7564	3.659
13	1.7709	2.1604	3.0123	4.221	30	1.6973	2.0423	2.7500	3.646