Visually Do Statistics
(1)What is the advantage of studying statistics ?
(2)Visualize Singular Value Decomposition

2019/02/26

Prof. Yukari SHIROTA (Gakushuin University)
Prof. Basabi Chakraborty (Iwate Prefectural University)

Why do you have to study the statistical comparison method ?

- Student: I am not interested in the statistics.
- Teacher: The comparison method is needed when you write a paper.
- Student: Then when we need to analyze the data, following the teacher's instruction, I can just input the data to SAS or Matlab or so.
- Teacher: If there is no teacher, you can never conduct the analysis. Without the knowledge of the statistical comparison method, you can not analyze your result.

To become a person who can analyze your data,
why don't you study the statistical comparison methods ?
We will tell you that using the visualization materials.

Frequently Seen Error

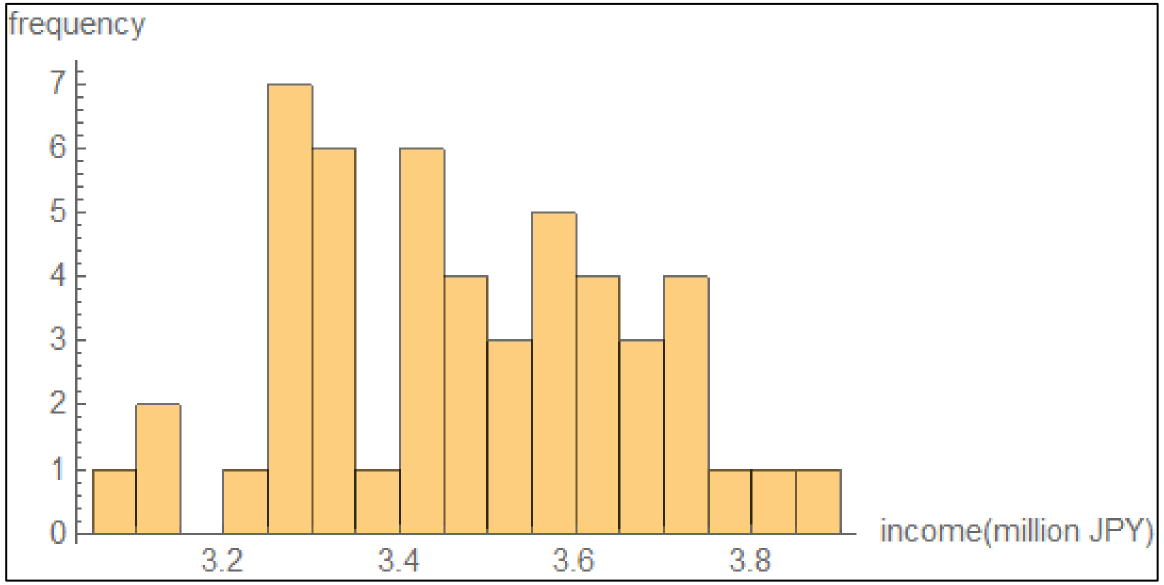
You do not notice that the sample average changes every time you measure.

MAYBE

Frequently Seen Error

The sample average changes every time you measure

- Mr BEAN surveyed the income among the city just once owing to the budget problem.
  The sample size was 50 persons.
- Boss: How much is the income average ?
- Mr BEAN: The average is **3.47** million JPY.
- Boss: I found that you had not studied statistics.
- Mr BEAN was so shocked to hear that.
  How does a statistics studying person answer in this case ?
- Answer example:
  The 95% confidence interval of the population average concerning income is 3.42 million JPY to 3.53 million JPY



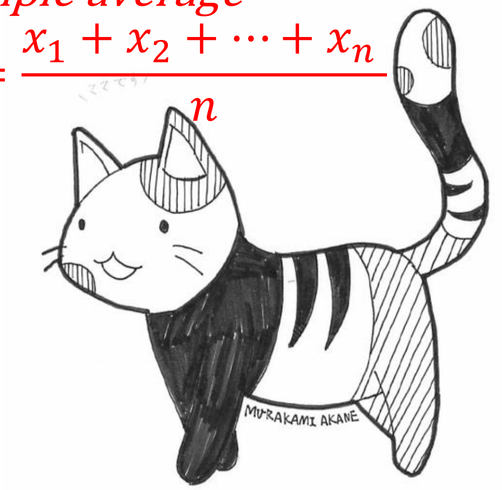
What is the confidence interval ?

Frequently Seen Error
The sample average changes every time you measure

- You should not forget that, every time you measure. For example, 3.47, 3.51, 3.33,⋯. The histogram of the sample income average changes as shown here.
- We **cannot know the true population income average**. Then we guess that by statistics.

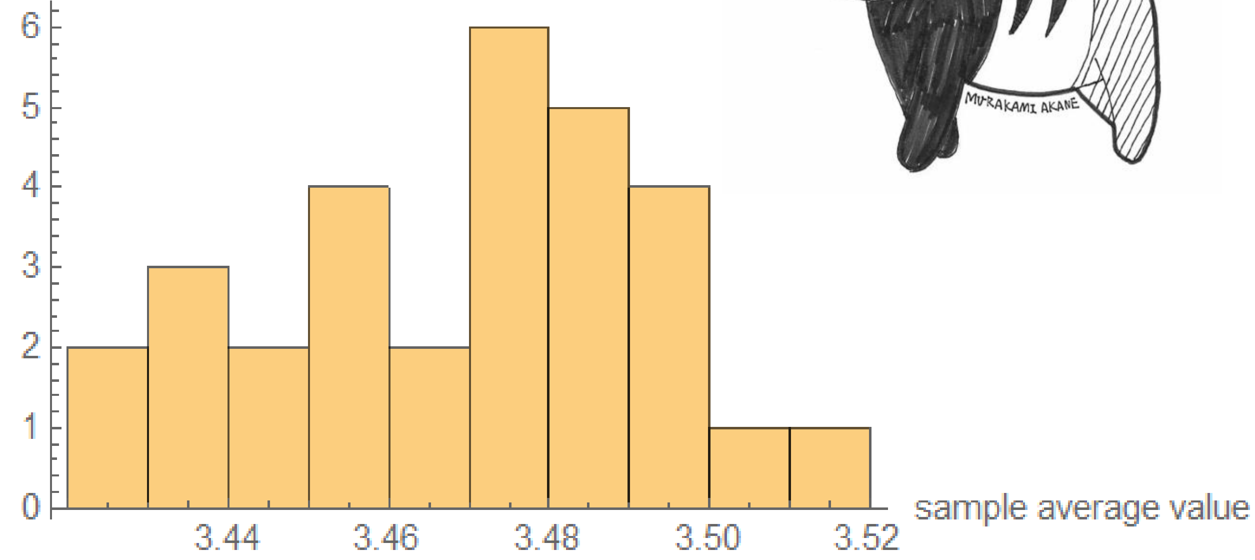
Sample average

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



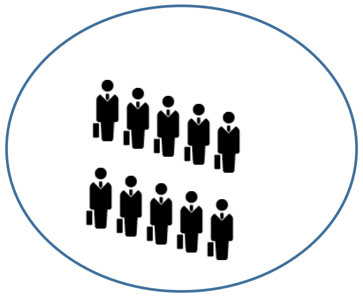
30 sampling results

frequency



Descriptive statistics VS Inferential statistics

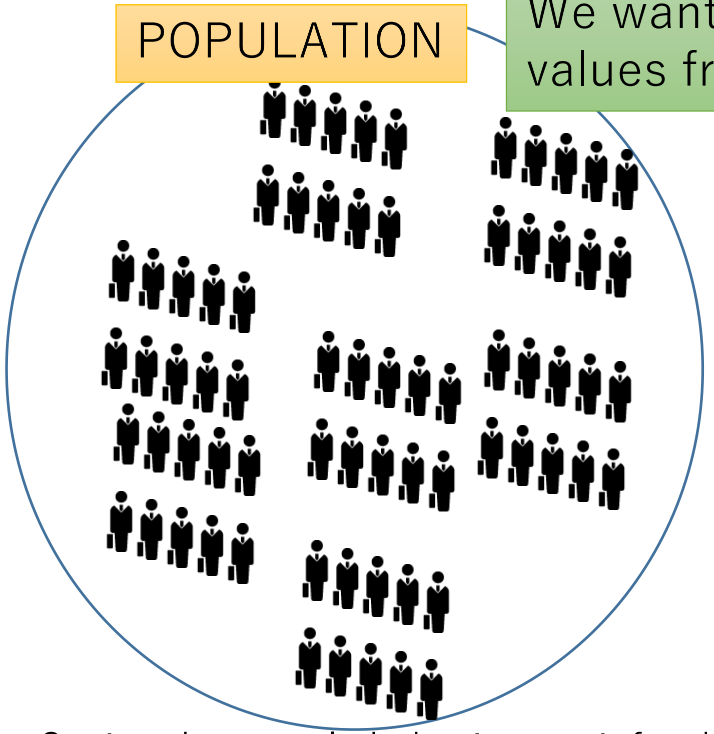
- # of data =10. That's all.



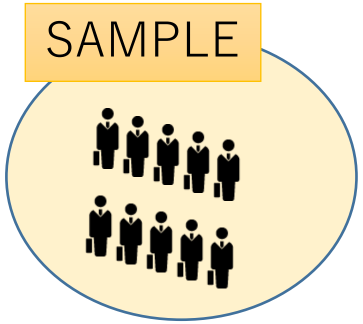
Sum of squares of deviations

$$\bar{X} = \frac{1}{10}\sum x_i \quad Var(X) = \frac{1}{10}\sum (x_i - \bar{X})^2$$

average of independent 10 data



We want to infer the population values from the sampling.



- Seeing the sample behavior, we infer the population behavior.
- The average is not the true population average. Although we want to infer the population variance, we have to use this constrained average.
- $\bar{X} = \frac{1}{10}\sum x_i \quad Var(X) = \frac{1}{10-1}\sum (x_i - \bar{X})^2$

Degrees of freedom is 9

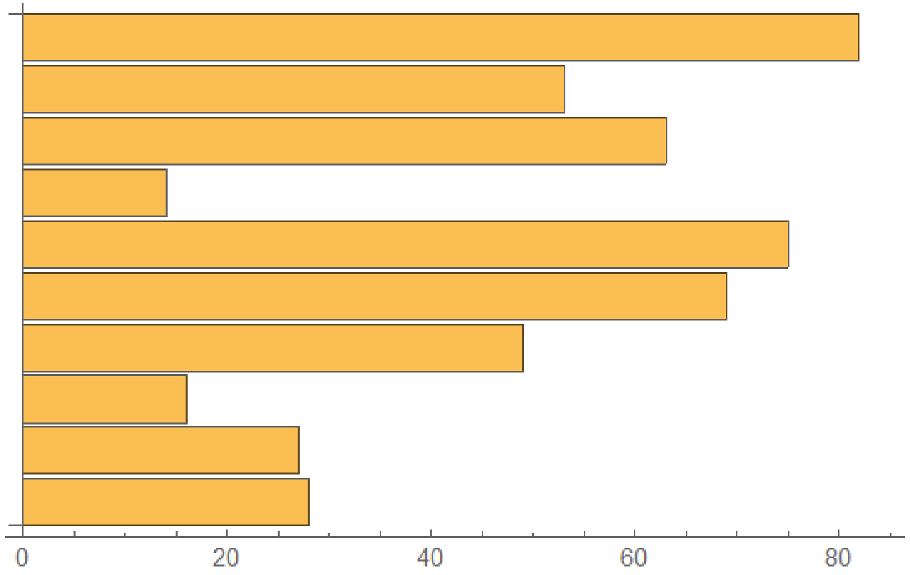
Using ***Sample average***

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

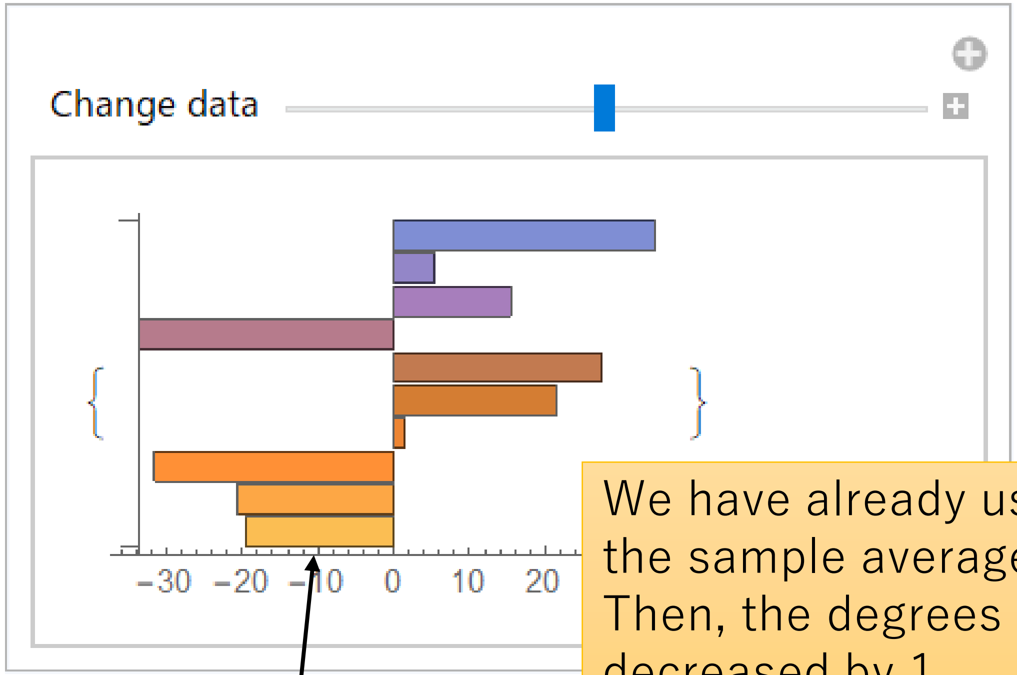
$$\sum (x_i - \bar{X}) = 0$$

decreases freedom by 1

- Given 10 sample data



- Deviation $x_i - \bar{X}$



We have already used the sample average. Then, the degrees of freedom decreased by 1.

The last one **cannot** move freely.

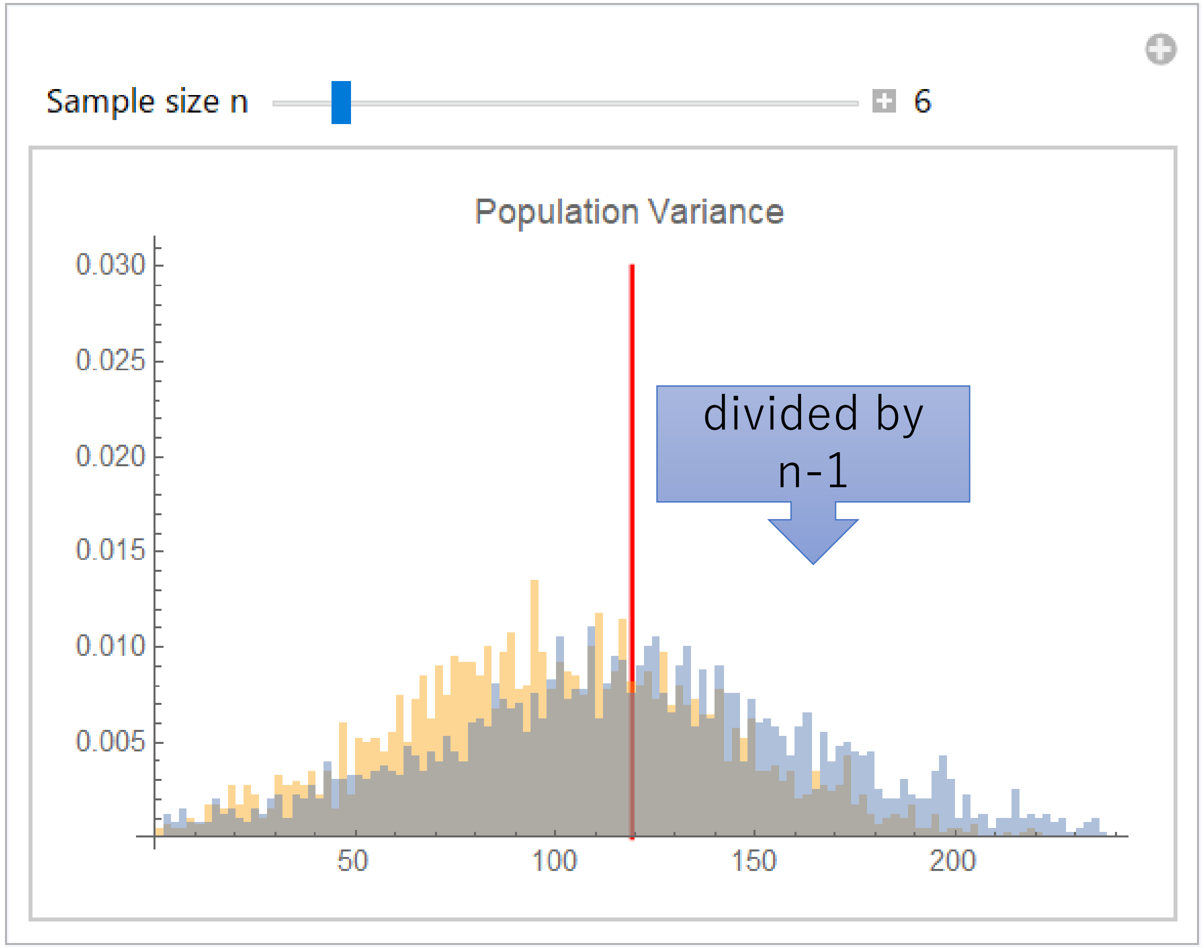
Degrees of freedom is 9

Simulation to see a sample variance

https://www-cc.gakushuin.ac.jp/~20010570/VDStat/unviasedVar1.cdf

- Red line: the true population variance that nobody knows
- Blue histogram: sample variance definition used
- Yellow histogram: population variance definition used
- In inferential statistics, a sample variance (divided by (n-1) offers us **a better value of the population variance**.

- Please install Wolfram CDF player which is free software.



Remember

- *sample variance*

$$U^2 = \frac{1}{n-1}\sum(x_i - \bar{X})^2$$

- $N(\mu,\ \sigma^2) \cong N(\mu,\ U^2)$

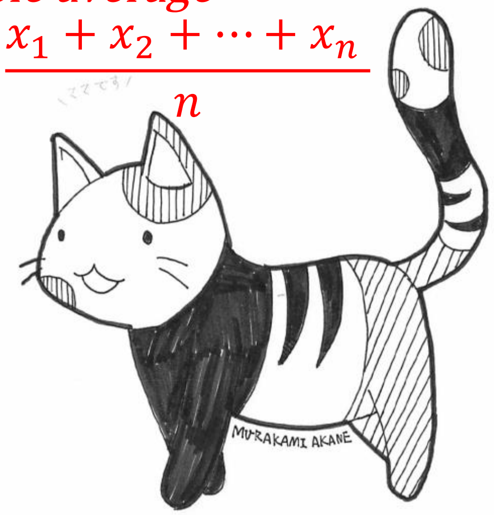
- Distribution of sample averages

- $N(\mu, \frac{\sigma^2}{n}) \cong N(\mu, \frac{U^2}{n})$

(from Central Limit Theorem)

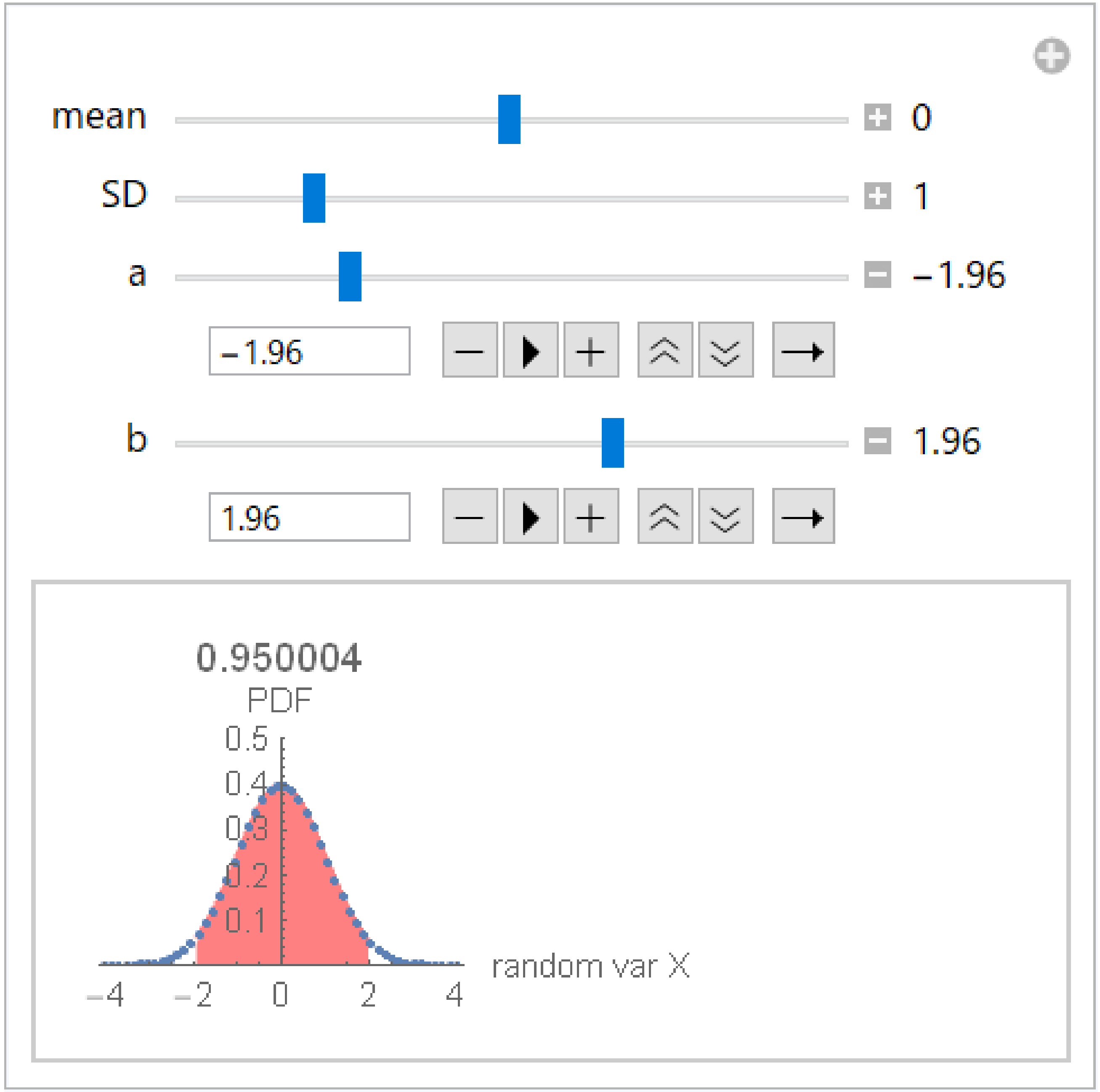
Sample average

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Remember

- N(0, 1)
- Prob(-1.96<X<1.96)=95%



95% Confidential Interval

1. Suppose that the sample average distribution follows the normal distribution.
2. Calculate the standard deviation sigma of $\bar{X}$ from the income variance.

σ Standard_deviation= $\sqrt{Variance}$

$Normal\ Distribution\ N(\mu, \frac{\sigma^2}{n})$

we can get this from Central Limit Theorem

$$\sigma = \sqrt{\frac{0.0354}{n}} = \sqrt{\frac{0.0354}{50}} = 0.02662$$

$$N(3.473, \frac{0.0354}{n})$$

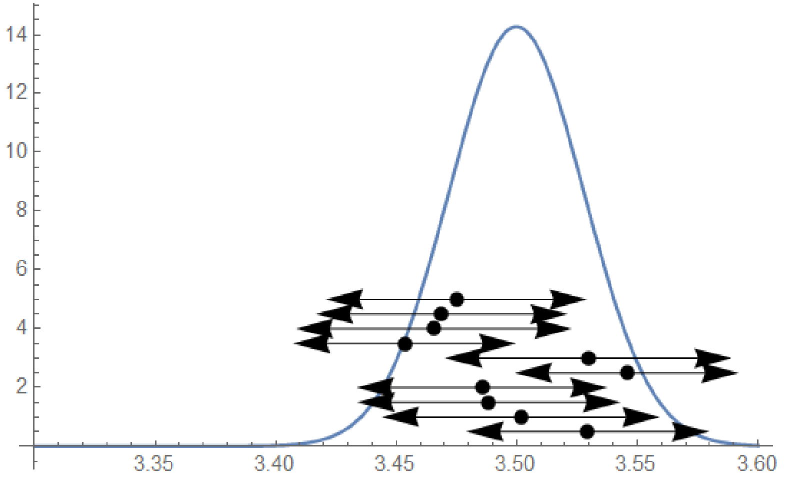
Sample average

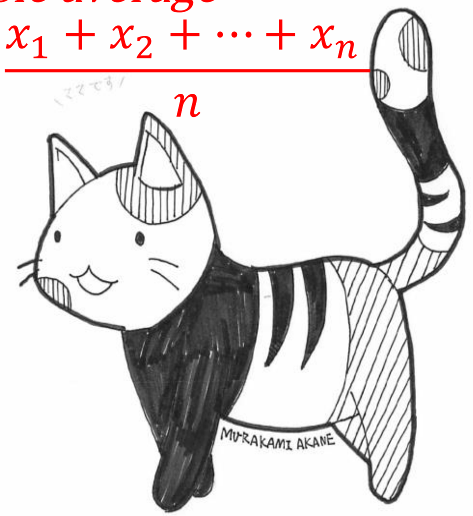
$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

5. Find the interval

$$\bar{X} - 1.96\,\sigma \ to \ \bar{X} + 1.96\,\sigma$$

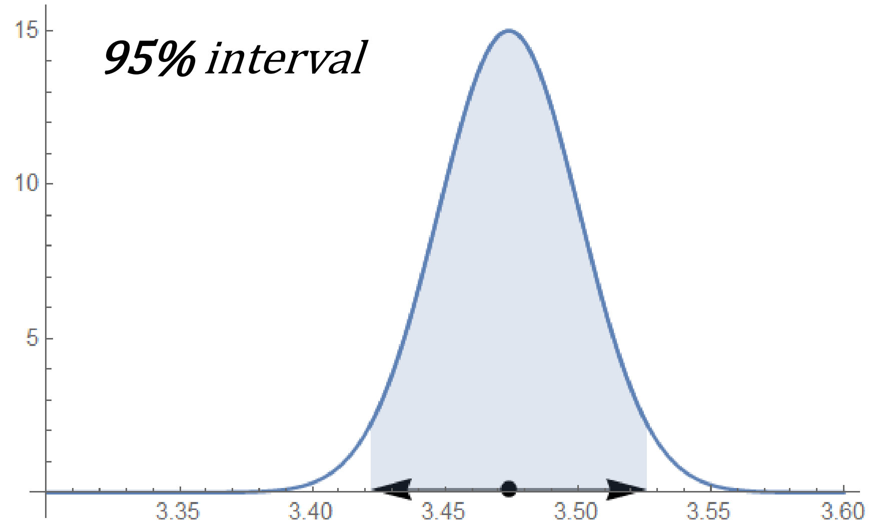
3.42 to 3.53 (million JPY)



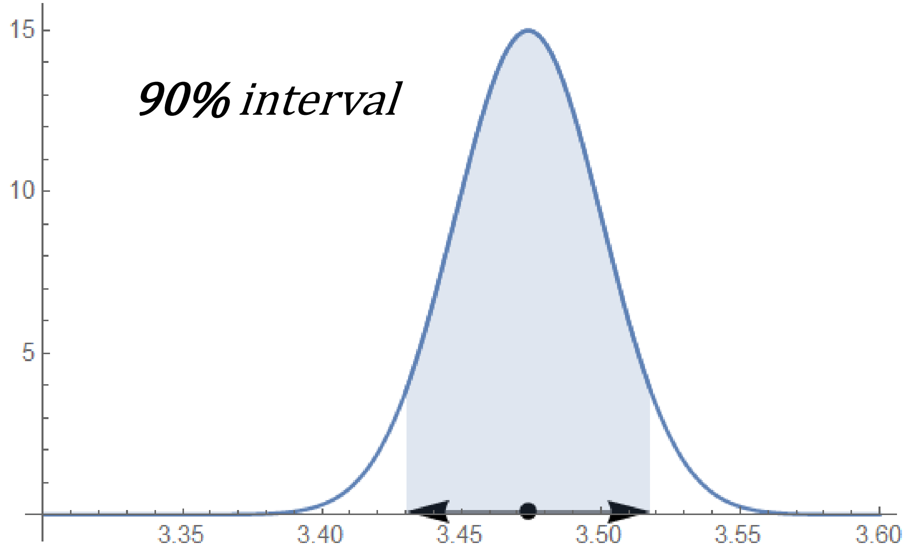


95% Confidential Interval

- ***95%* interval**
  $\bar{X} - 1.96\,\sigma$ *to* $\bar{X} + 1.96\,\sigma$



- ***90%* interval**
  $\bar{X} - 1.65\,\sigma$ *to* $\bar{X} + 1.65\,\sigma$



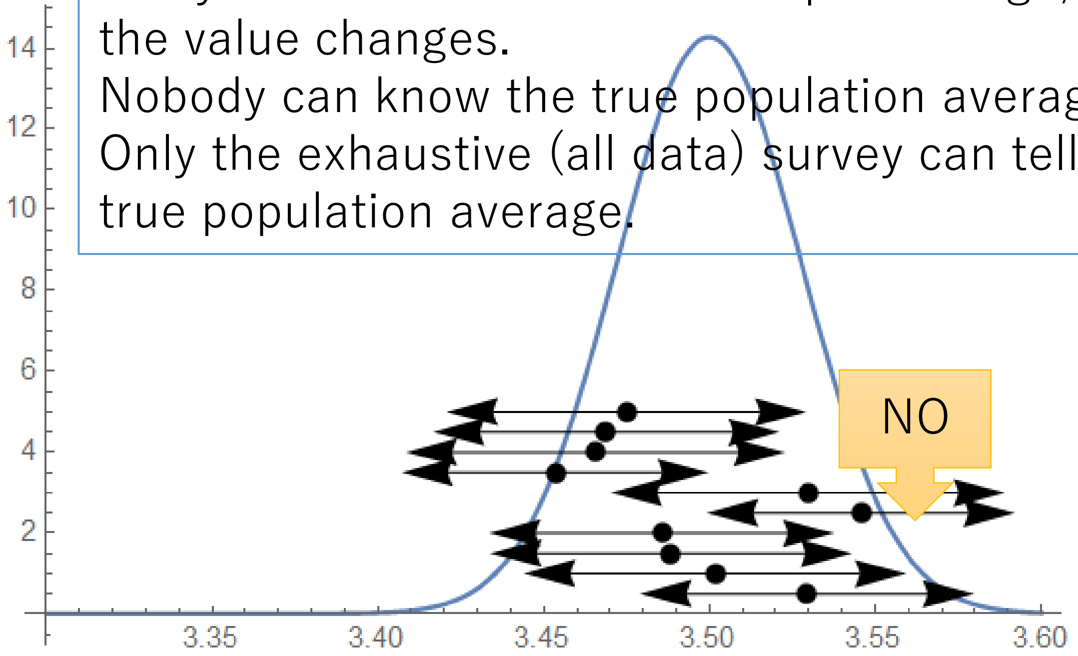
- Drill: Find the probability from $\bar{X} - 1\sigma$ *to* $\bar{X} + 1\sigma$.

95% Confidential Interval

- *95% interval*
  $\bar{X} - 1.96\,\sigma$ *to* $\bar{X} + 1.96\,\sigma$
- We can say only that the true population average exists in the interval with 95 % confidence level.
- We cannot know even which side the population average exists.

- *95% interval*
  $\bar{X} - 1.96\,\sigma$ *to* $\bar{X} + 1.96\,\sigma$
- We can say only that the true population average exists in the interval with 95 % confidence level.
- We cannot know even which side the population average exists.

Every time we measure the sample average,
the value changes.
Nobody can know the true population average.
Only the exhaustive (all data) survey can tell us the true population average.

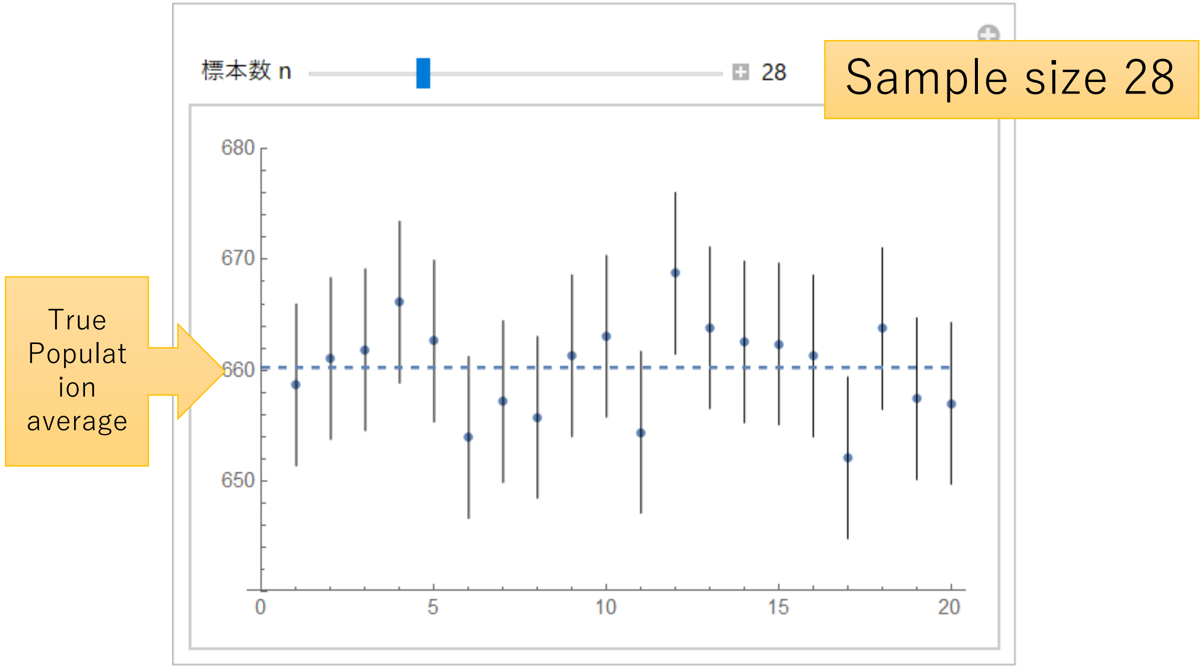


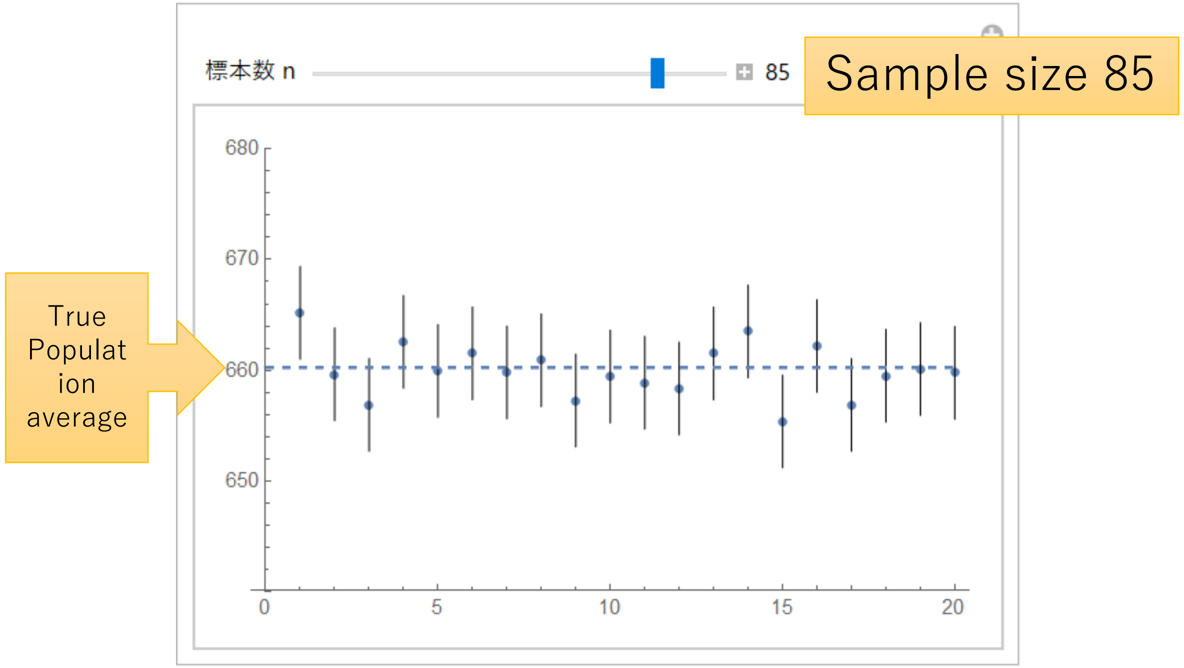
Graphical materials for 95% confidential interval

Please install Wolfram CDF player which is free.

https://shirotaabc.sakura.ne.jp/usefulMath/ABC/7-25.cdf

- How many tries can include the true population average ?
- Answer: 95% and 95 % (1 failure per 20 tries)

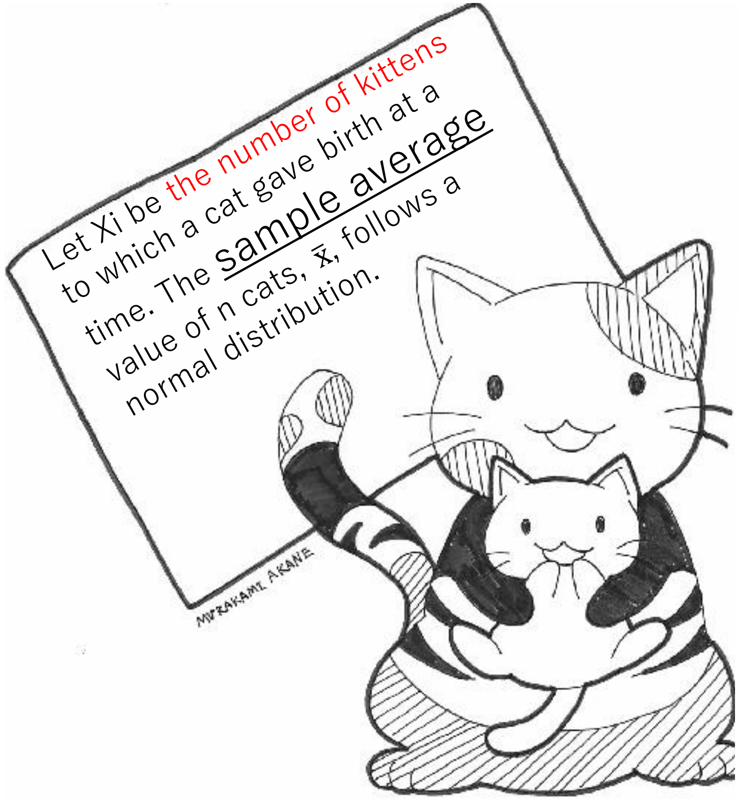


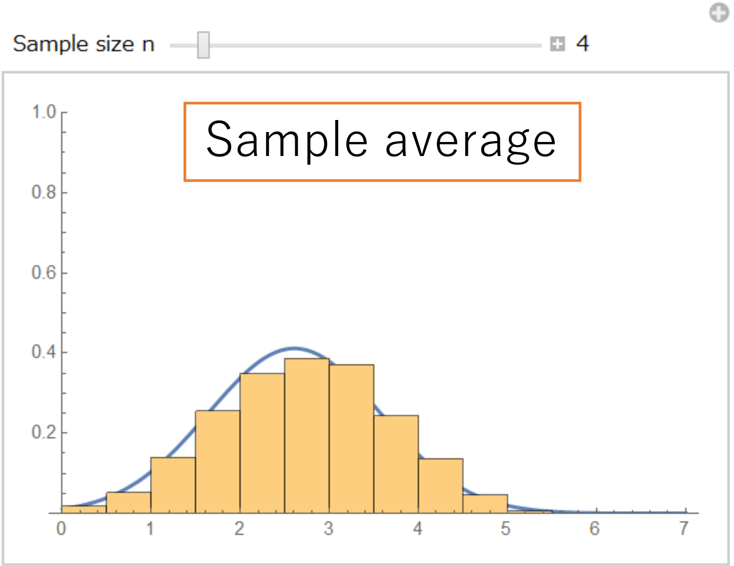


Cited from Y. Shirota et al., 「大学生のための役に立つ数学」, p.144, Kyoritu, Tokyo, 2014.

Central Limit Theorem

$$Normal\ Distribution\ N(\mu, \frac{Var(X)}{n})$$

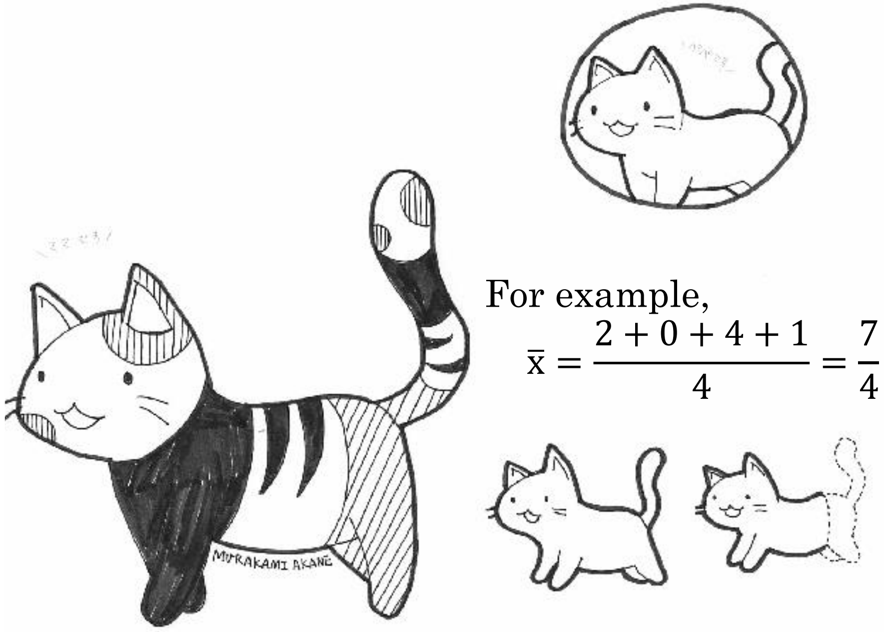




Sample average

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Father



For example,

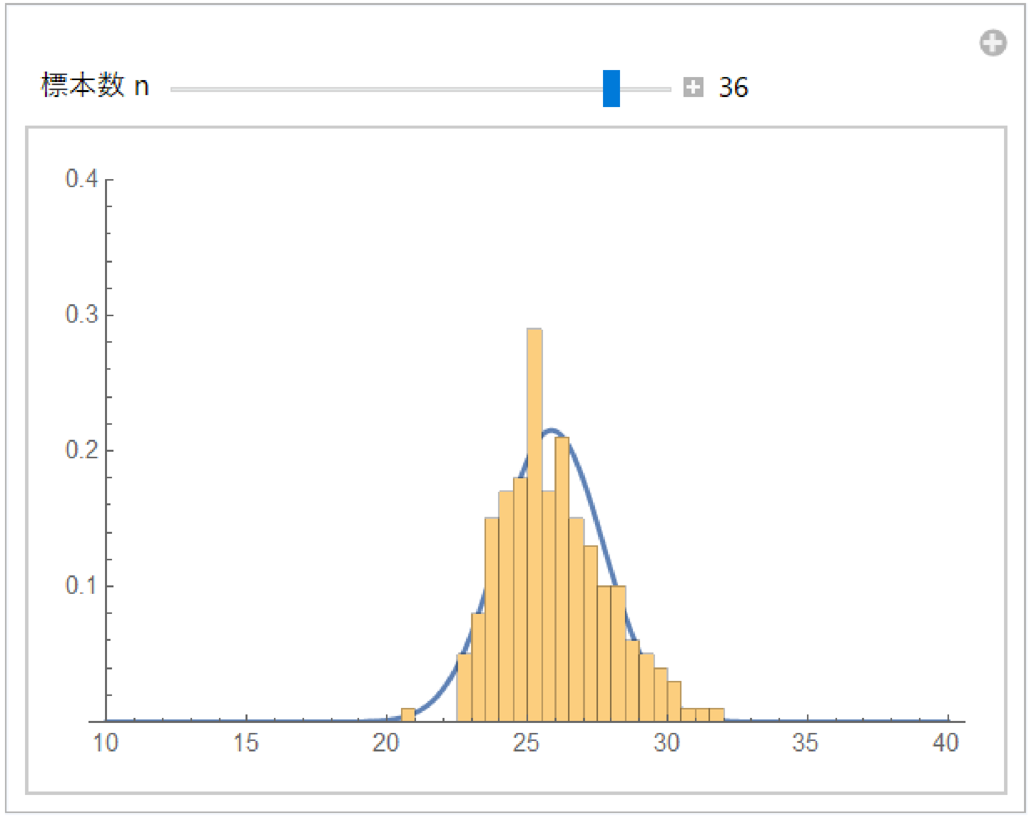
$$\bar{x} = \frac{2+0+4+1}{4} = \frac{7}{4}$$

Graphical materials for CLT

Please install Wolfram CDF player which is free.

https://shirotaabc.sakura.ne.jp/usefulMath/ABC/7-19.cdf

- Let's operate and look at the distribution, getting near to the normal distribution.



Cited from Y. Shirota et al.,「大学生のための役に立つ数学」, p.143, Kyoritu, Tokyo, 2014.

Summary

- 95% confidence interval
- Central Limit Theorem

Frequently Seen Error

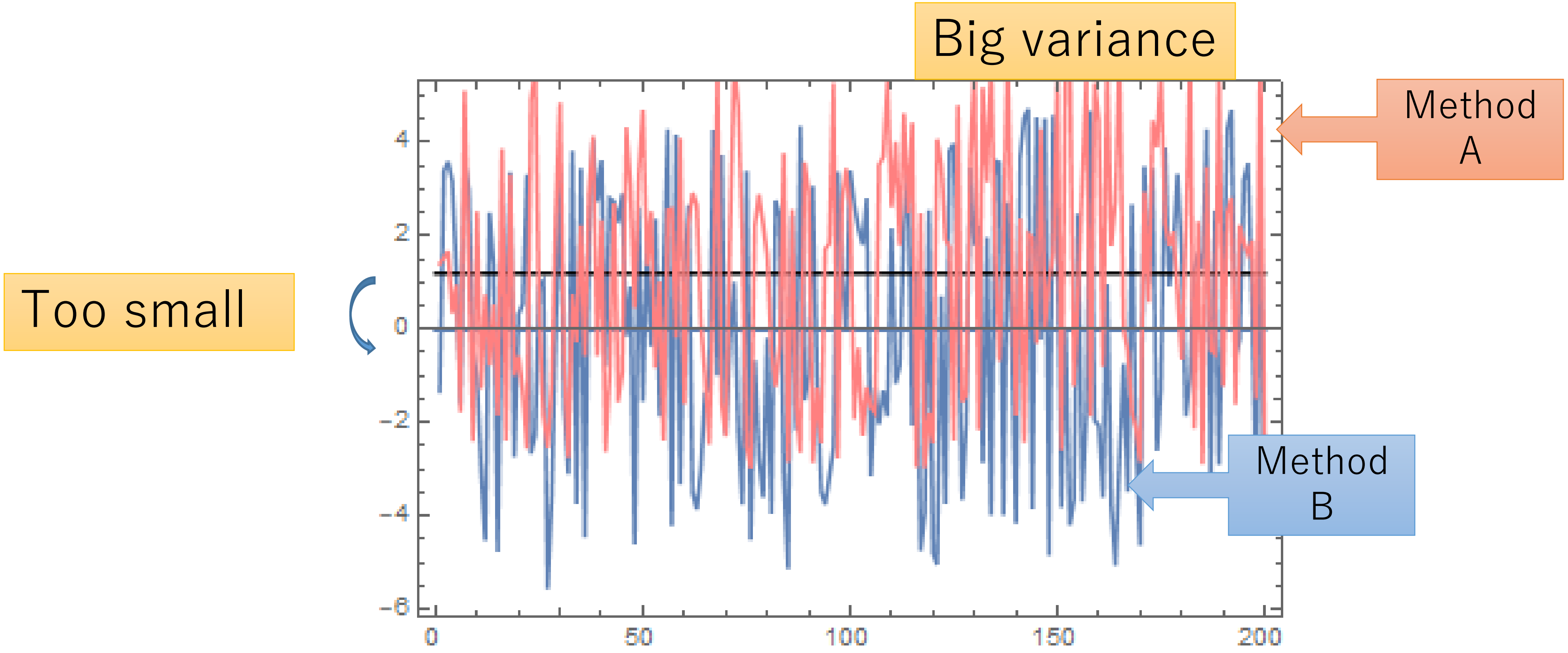
In comparison of average values,
the higher average value does not always mean the better methods.

(1) Comparison of two methods→ Hypothesis testing about $\mu_1 - \mu_2$

(2) Comparison of three methods→ Analysis of variance (ANOVA)

See the variances

- If the average difference is too small, compared to variances, we cannot say that the method A is superior.

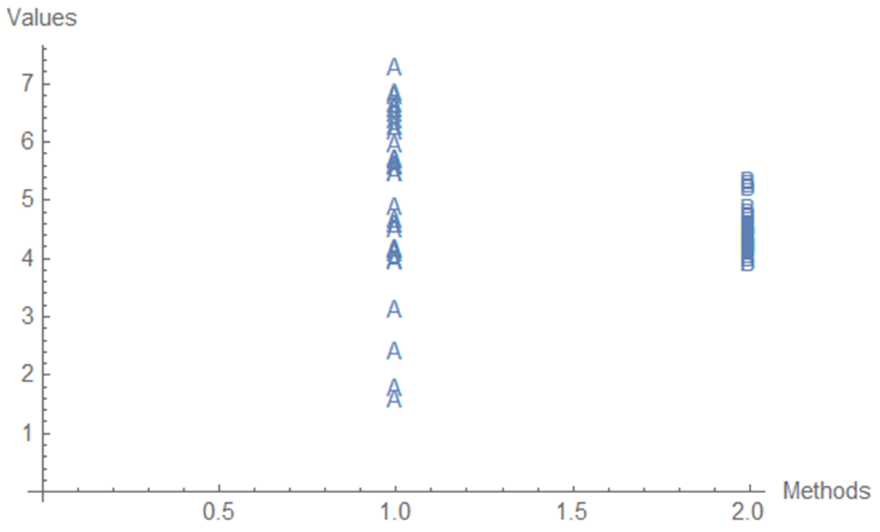


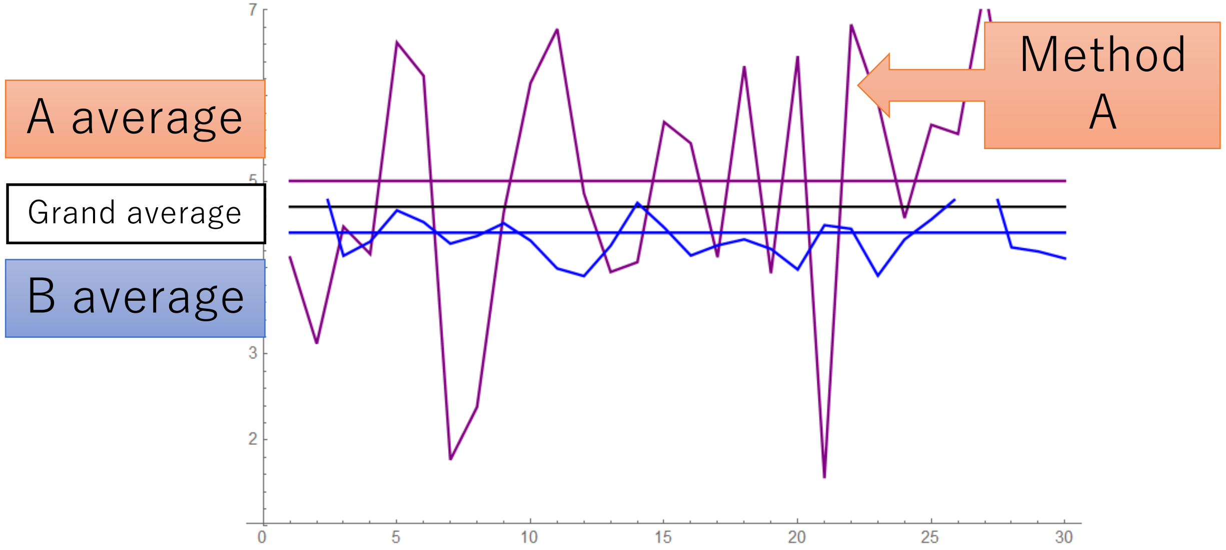
Comparison of two methods→ Hypothesis testing about $\mu_1 - \mu_2$

Frequently Seen Error

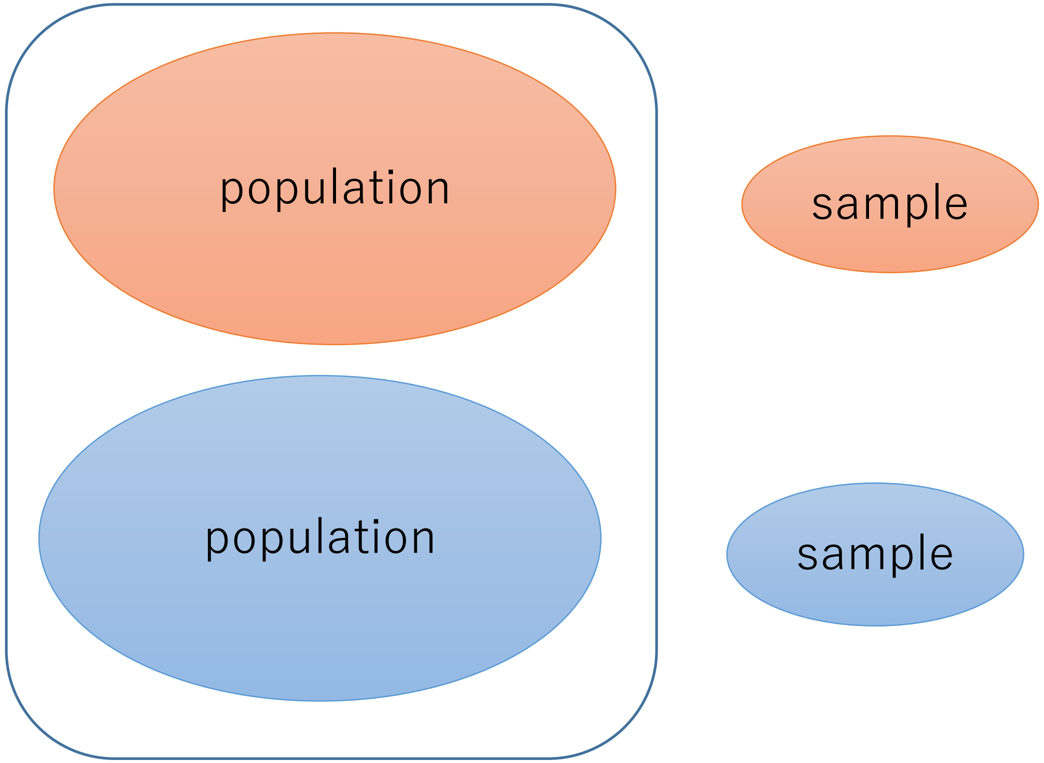
The higher average value does not mean the better methods.

- There are two kinds of methods(treatments) A and B which you can take.
- Two samples of the effects are as shown here where each sample size is 30.
- Higher value, the better effect the method has.
- Mr BEAN's remark: The method A is better than B, because the average is higher than B.
- The remark is INCORRECT.
- You should use a hypothesis test for the comparison to say whether A is superior or not.





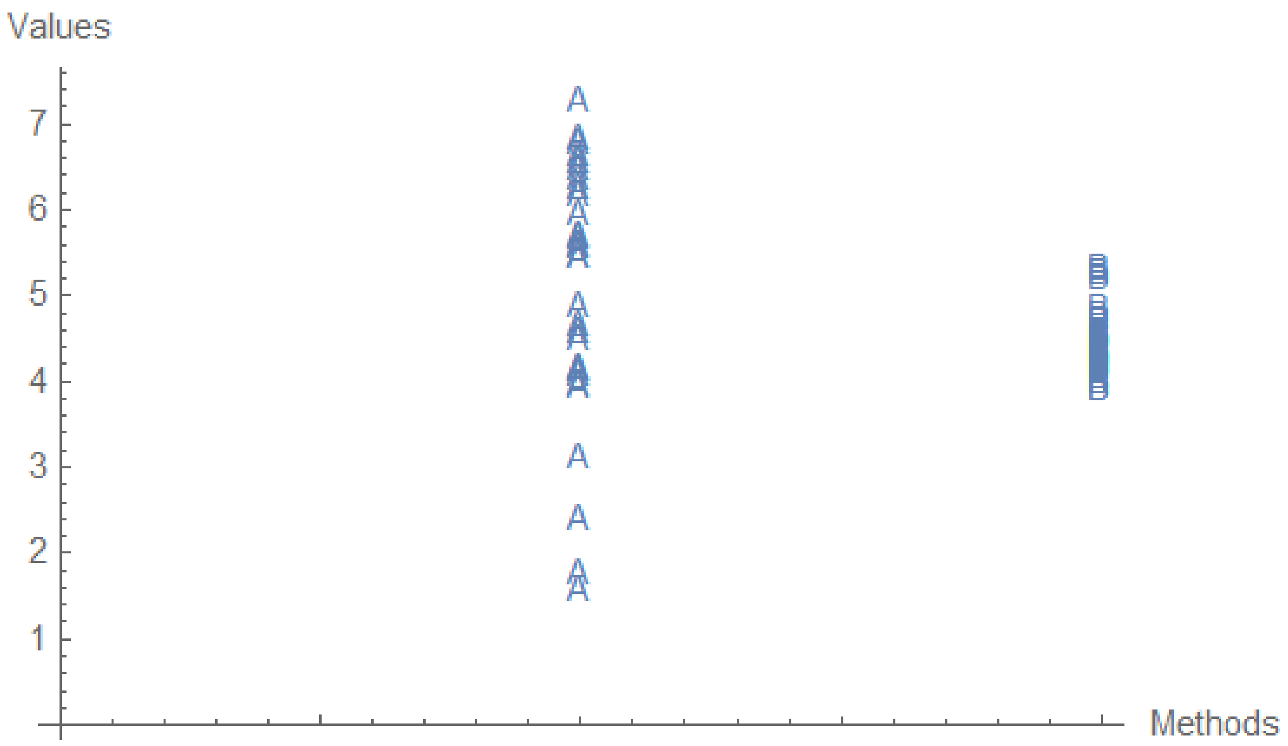
Calculate the variance between samples and the variance within samples



We want to infer these behaviours.

- This is inferential statistics.

The variance of sample A is still larger than one of sample B.

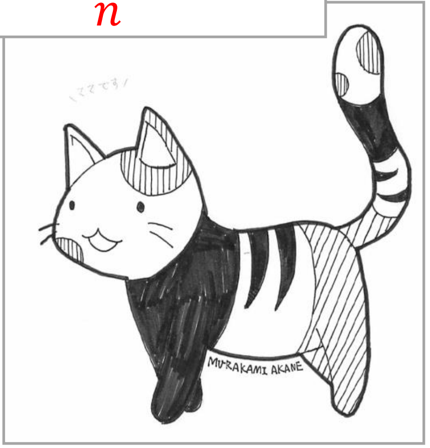


- The difference of two average $\overline{X_1}$ and $\overline{X_2}$ $is\ not\ so\ large,$ $compared\ to\ the\ variance\ of\ A$

Sampling distribution of $\overline{X_1} - \overline{X_2}$

Sample average

$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_N}{n}$$

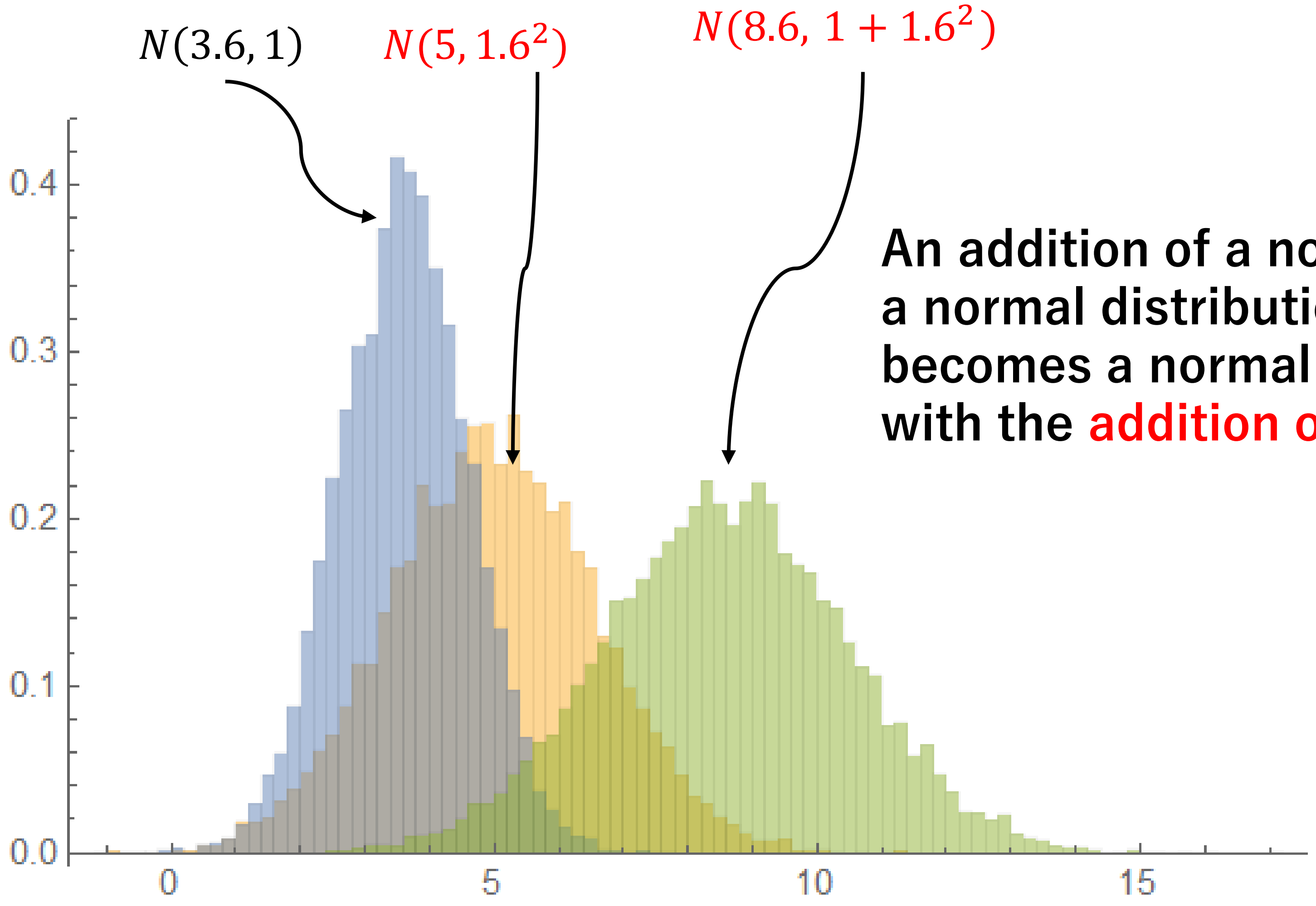


- $\overline{X_1} - \overline{X_2}$ follows a normal distribution

$$N\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

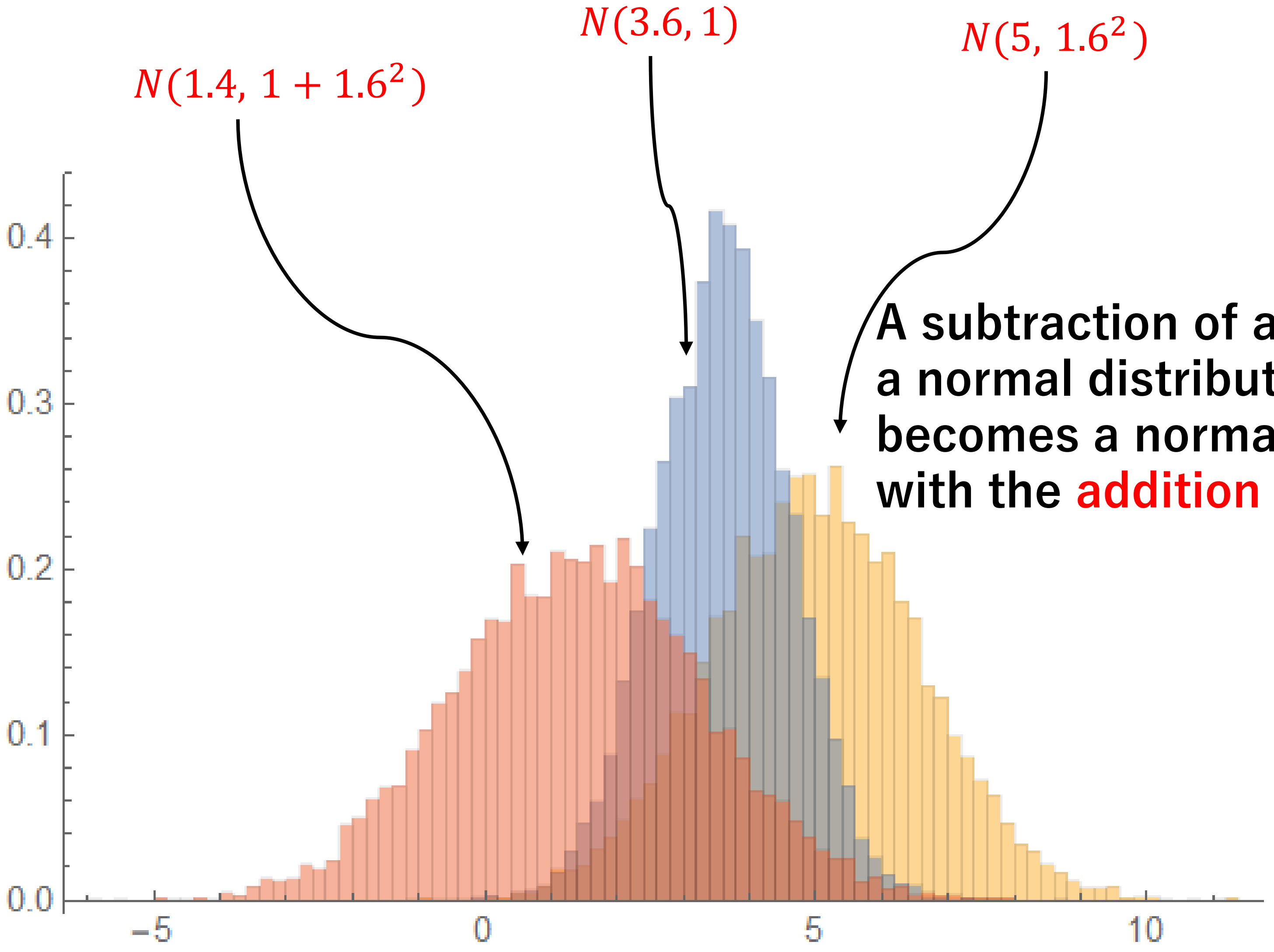
This is the theory.

Distribution of $X_1 + X_2$ *follows* $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$



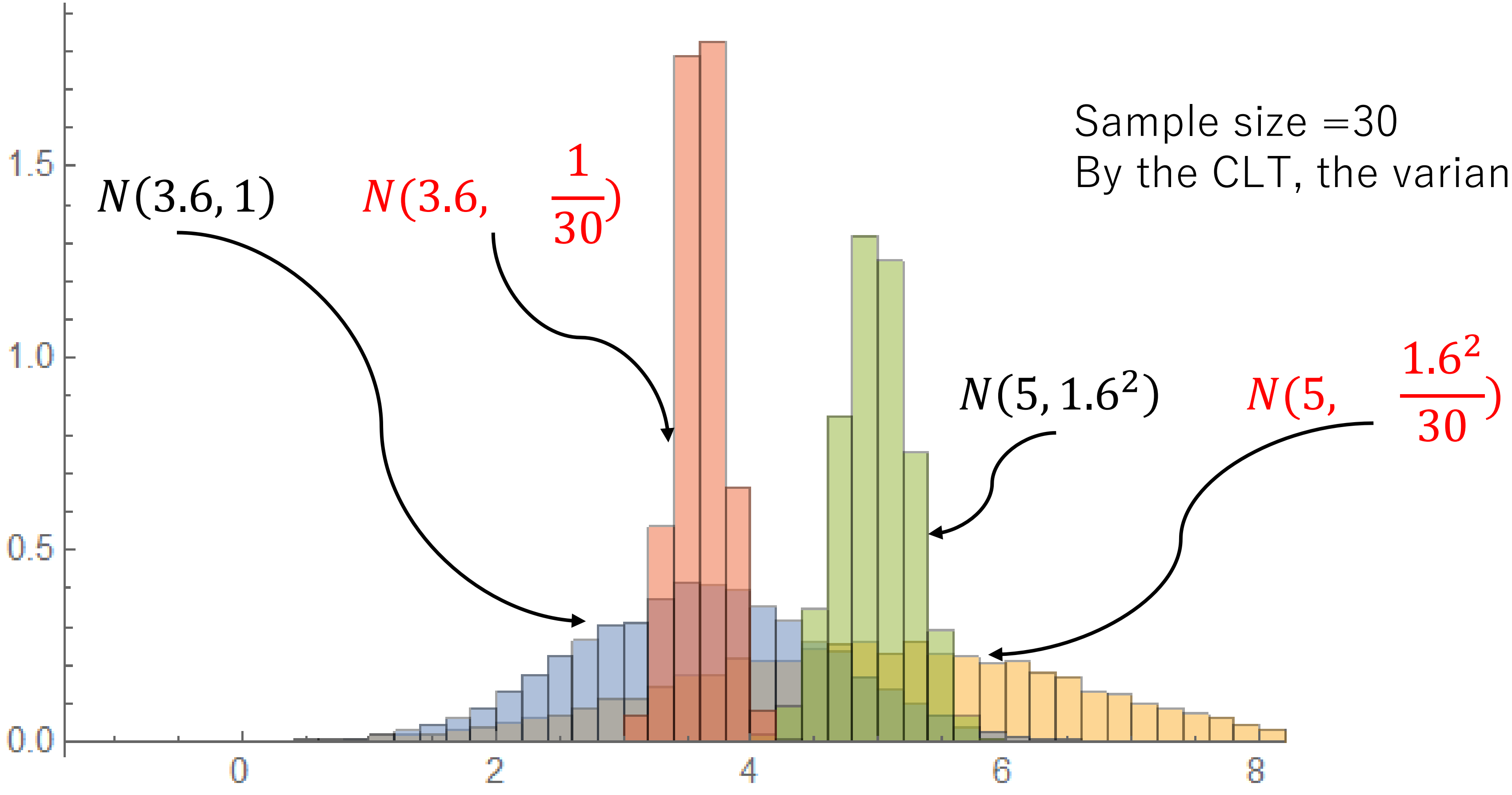
An addition of a normal distribution variable and a normal distribution variable becomes a normal distribution with the addition of the two variances

Distribution of $X_1 - X_2$ *follows* $\color{red}{N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)}$



A subtraction of a normal distribution variable from a normal distribution variable becomes a normal distribution with the addition of the two variances

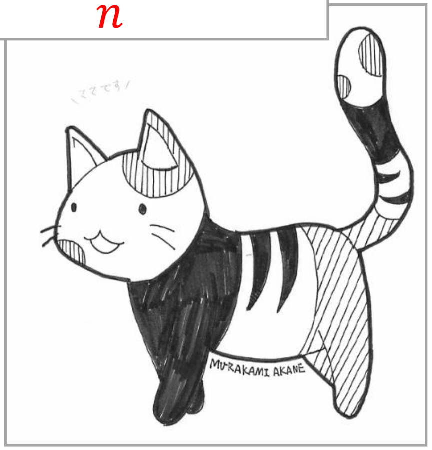
Sampling distribution $\overline{X_1}$ follows $N(\mu_1, \frac{\sigma_1^2}{n_1})$



Sample size =30
By the CLT, the variance becomes small.

Sampling distribution of $\overline{X_1} - \overline{X_2}$

Sample average
$\bar{X} = \dfrac{x_1 + x_2 + \cdots + x_N}{n}$



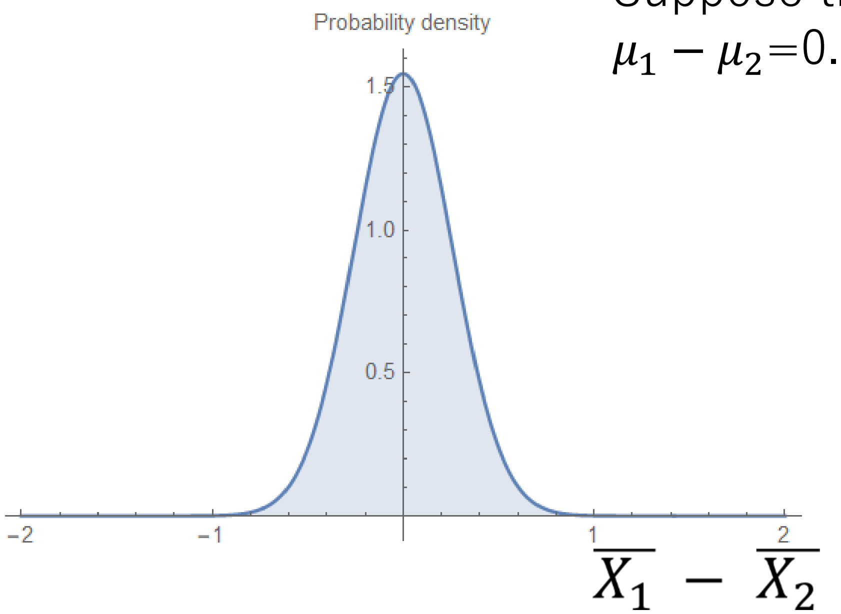
- $\overline{X_1} - \overline{X_2}$ follows a normal distribution

$$N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

This is from the theory CLT.

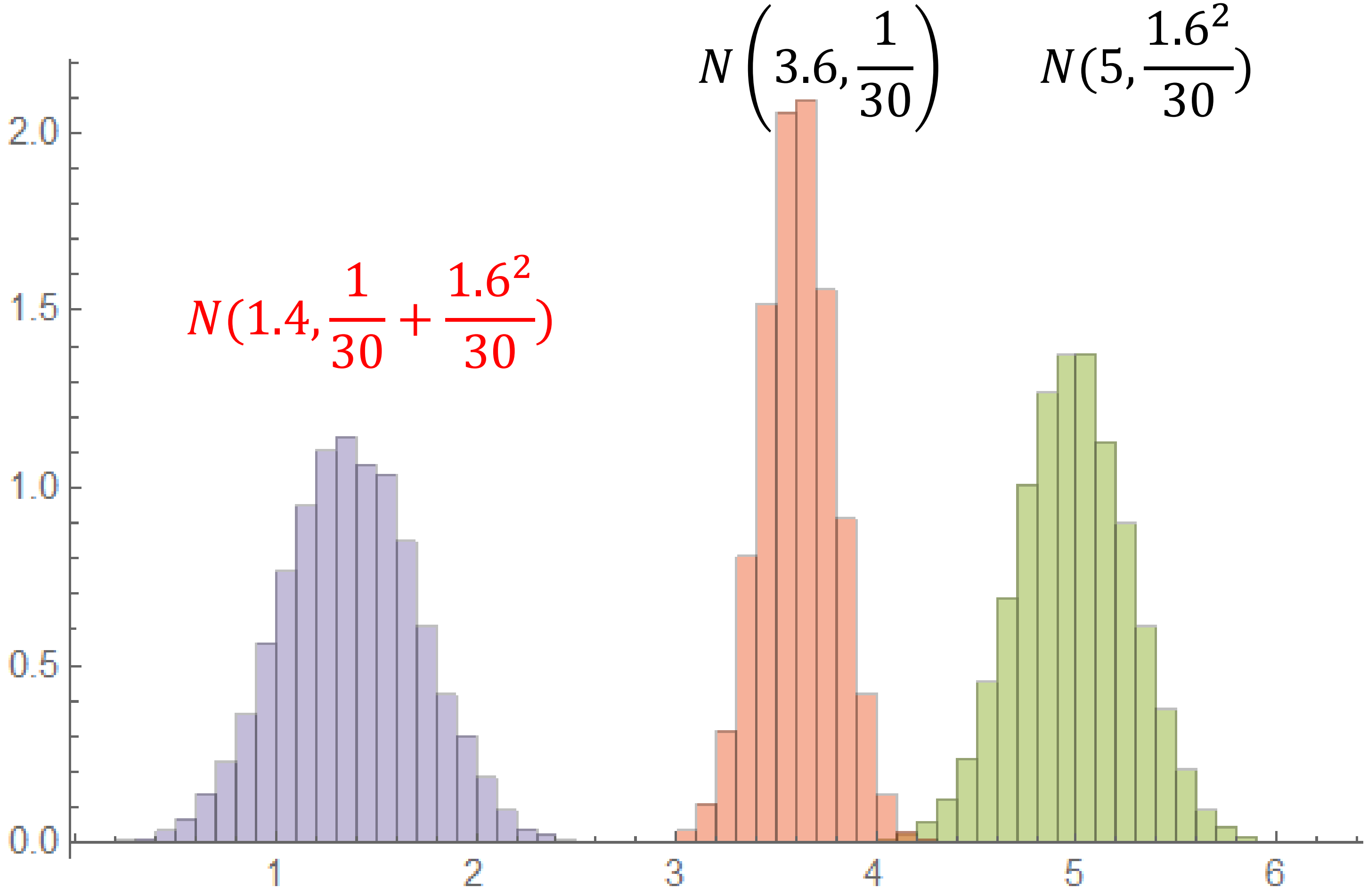
- Conditions to be required:
  - Two samples are independent
  - The standard deviations σ_1 and σ_2 are known.
  - Both sample sizes are large (≧30)

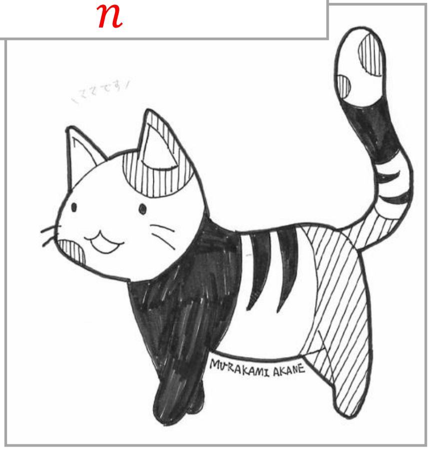
Suppose that $\mu_1 - \mu_2 = 0$.



Sampling distribution of $\overline{X_1} - \overline{X_2}$

Sample average
$\bar{X} = \frac{x_1 + x_2 + \cdots + x_N}{n}$





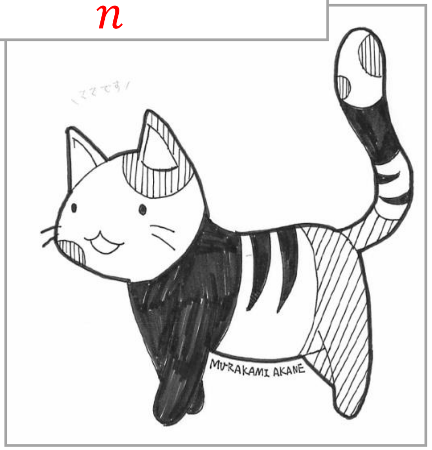
Hypothesis testing

- **Null hypothesis**
  - H_0: $\mu_1 - \mu_2 = 0$ (The two population averages are not different.)

- **Alternative hypothesis**
  - H_1: : $\mu_1 > \mu_2$ (The population average of A is greater than one of B.)

The obtained value

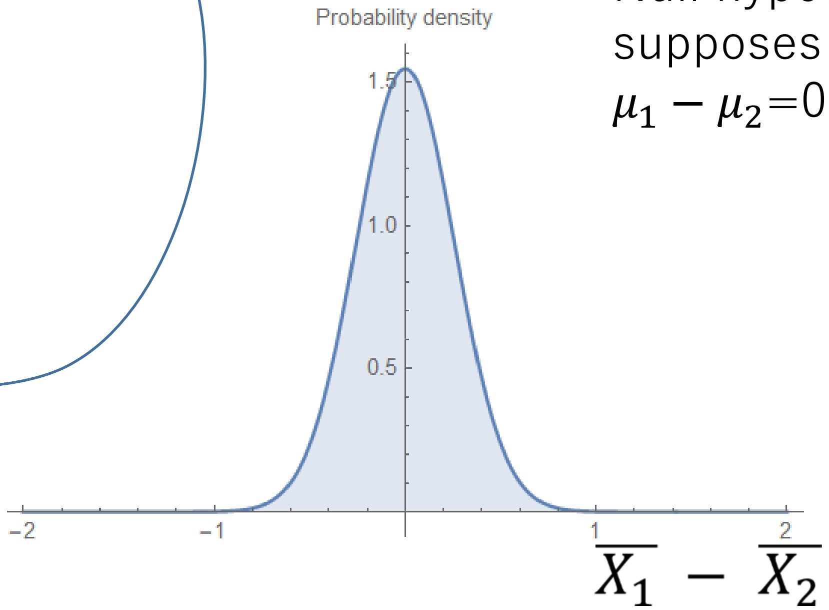
Sample average
$$\bar{X} = \frac{x_1 + x_2 + \cdots + x_N}{n}$$



- $\overline{X_1} - \overline{X_2}$ was 1.3 → $\frac{1.3}{\sigma} = \frac{1.3}{0.344} = 3.77$

$N(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

- Conditions to be required:
  - The standard deviations σ_1 *and* σ_2 are known. $\sigma_1 = 1.6^2, \sigma_2 = 1$
  - $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ = 1.6^2/30 + 1/30 = 0.1187
  - $\sigma = (\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$^0.5 = 0.344



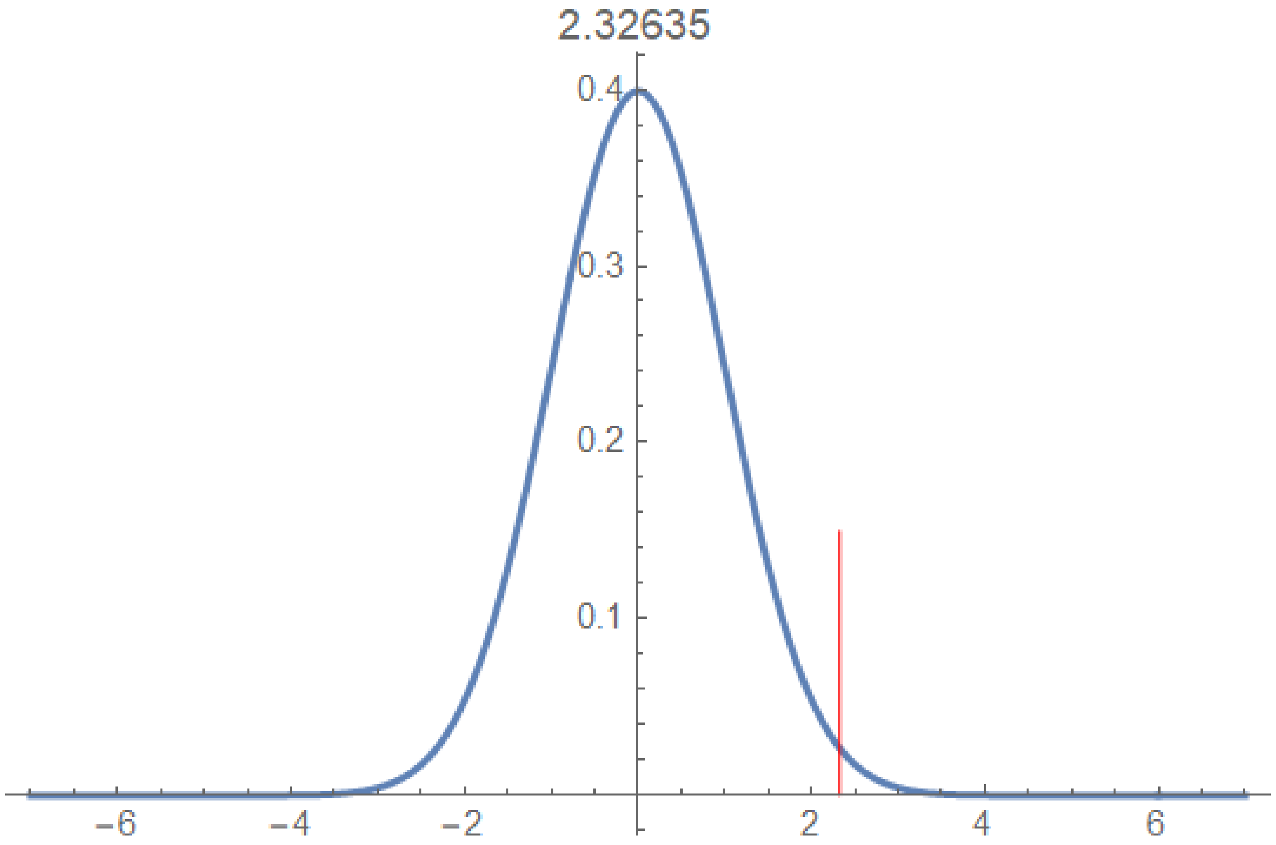
Null hypothesis supposes that $\mu_1 - \mu_2 = 0$.

Two-tailed test
significance level 1%

- Z=3.77 falls in the rejection region

Because the 1% boundary is 2.32

- The null hypothesis was rejected.
- Making the decision: We conclude that $\mu 1 > \mu 2$

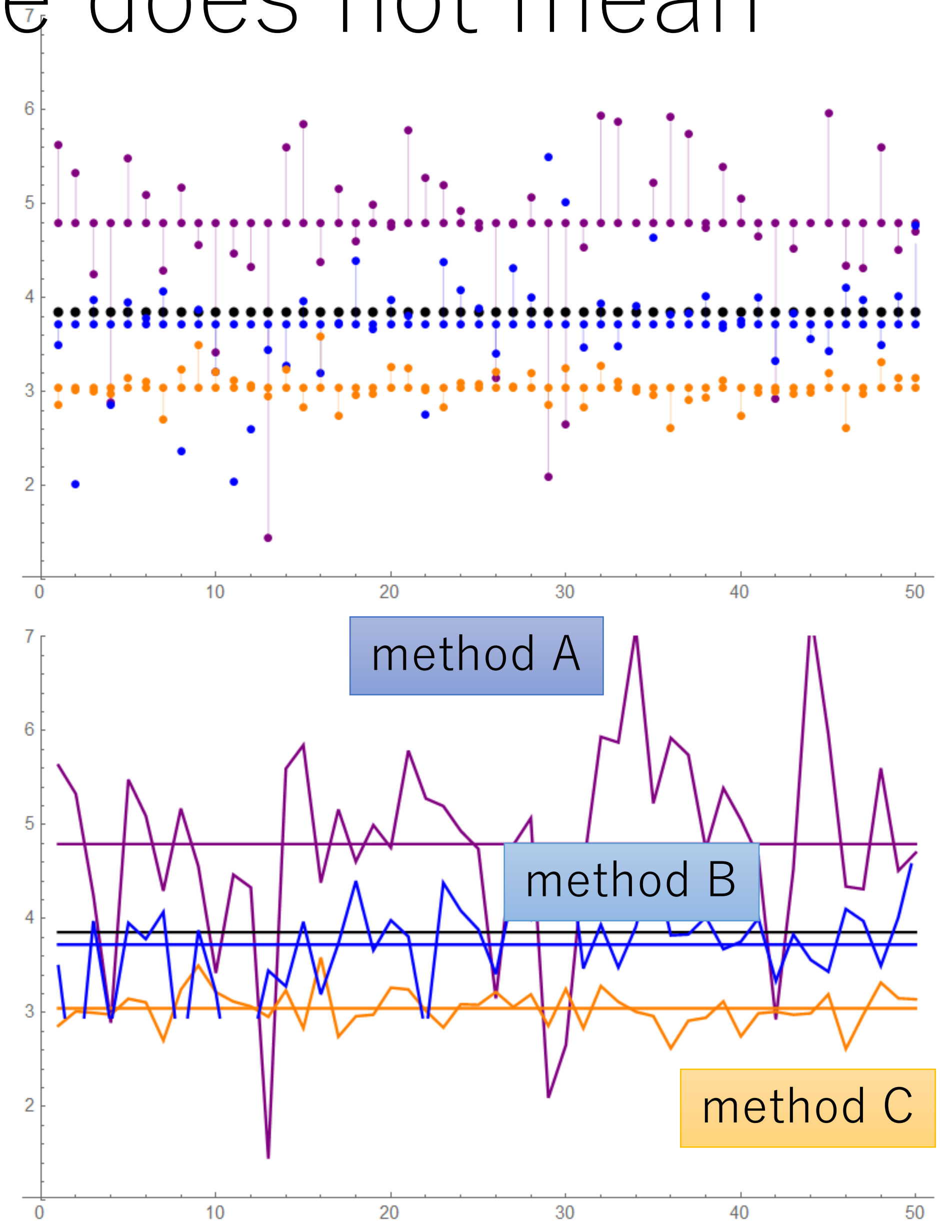


Comparison of three methods→ Analysis of variance (ANOVA)

Frequent Error

The higher average value does not mean better method.

- There are three kinds of methods(treatments) A, B and C which you can take.
- Three samples of the effects is as shown here where each sample size is 50.
- Higher the value, better the effect the method has.
- Mr BEAN's remark: The method A is better than B or C, because the average is higher than others.
- The remark is INCORRECT.
- You should use ANOVA for the comparison to say whether A is superior or not.



Hypothesis testing

- **Null hypothesis**
  - H_0: $\mu_1 = \mu_2 = \mu_3$ (All three methods population averages are equal.)

- **Alternative hypothesis**
  - H_1: Not all three methods population averages are equal.

ANOVA is a procedure that is used to test the null hypothesis.

ANOVA

The ratio of effect of treatment variance and effect of noise variance follows F-distribution.

- $grand\ average\quad \bar{X} = \frac{1}{\{50\times3\}}\sum_{i=1}^{3}\sum_{j=1}^{50} x_{i,j}$
- $sample(treatment)average\ \ \bar{X}_i = \frac{1}{50}\sum_{j=1}^{50} x_{i,j}$
- Let's calculate the total sum of squares of deviations.

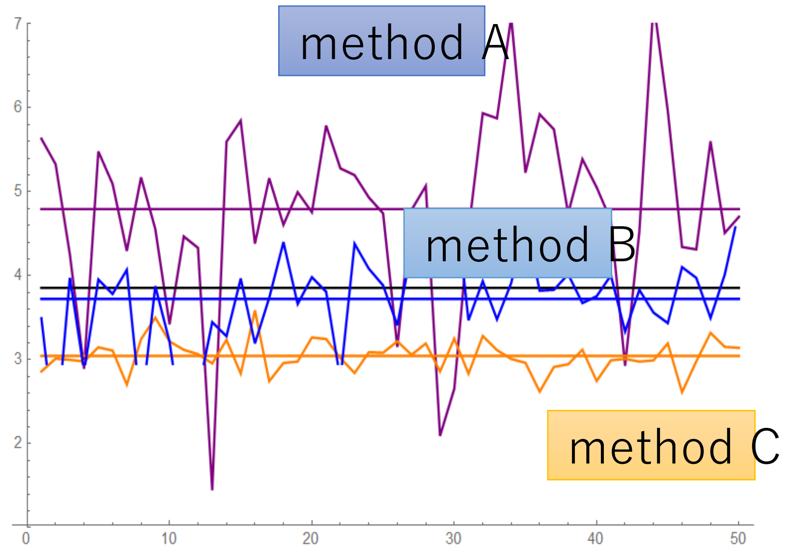
$$x_{i,j} - \bar{X} = \boxed{(\bar{X}_i - \bar{X})} + \boxed{(x_{i,j} - \bar{X}_i)}$$

effect of treatment effect of noises

$$\sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X})^2 = \sum_{i=1}^{3}\sum_{j=1}^{50}(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X}_i)^2$$

Total sum of squares is between-samples sum of squares +within-samples sum of squares

Calculation is a bit troublesome.



If between-samples sum of squares >> within-samples sum of squares, the null hypothesis is rejected.

ANOVA

The ratio of effect of treatment variance and effect of noise variance follows F-distribution.

$$x_{i,j} - \bar{X} = (\bar{X}_i - \bar{X}) + (x_{i,j} - \bar{X}_i)$$

effect of treatment   effect of noises

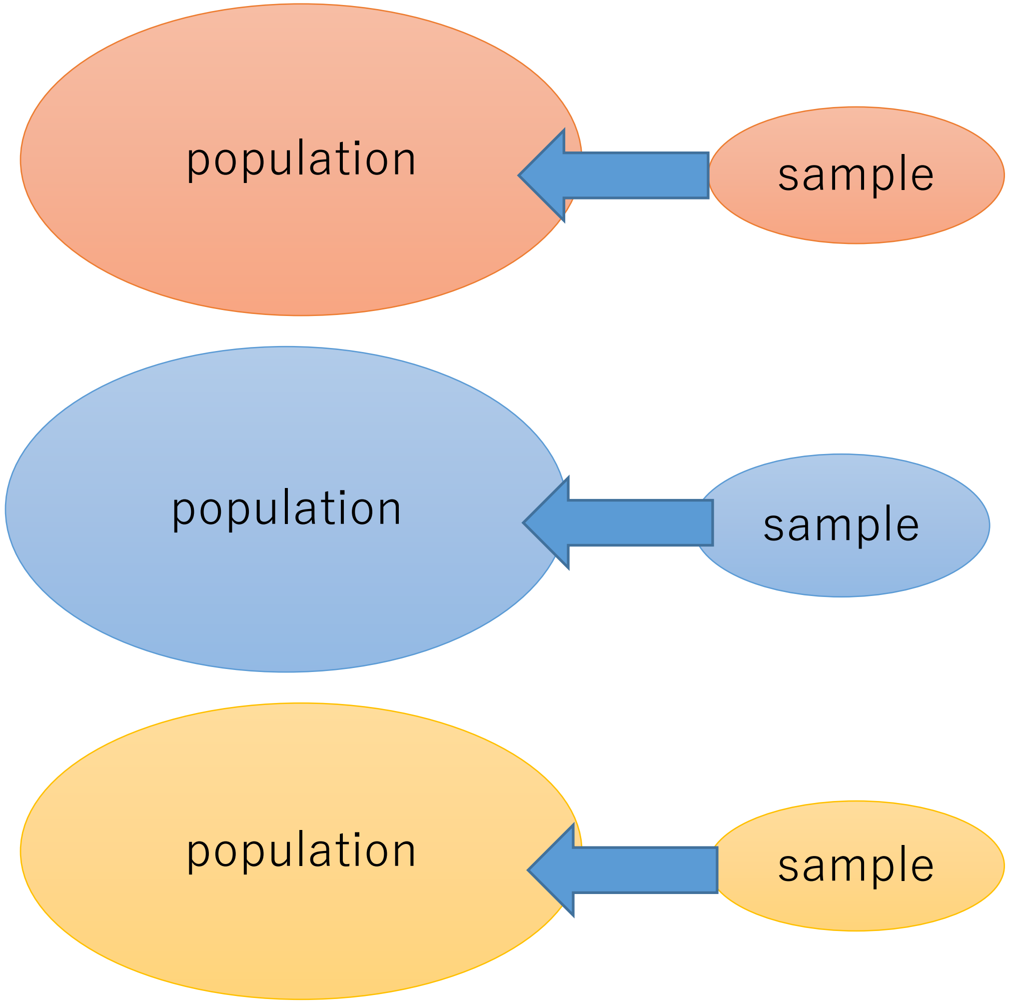
$$\sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X})^2 = \sum_{i=1}^{3}\sum_{j=1}^{50}(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X}_i)^2$$

Total sum of squares is between-samples sum of squares +within-samples sum of squares

If between-samples sum of squares >> within-samples sum of squares,
then the null hypothesis is rejected.

Some treatment exists.

Calculate the variance between samples and the variance within samples



- This is inferential statistics.
- THEORY:
  Variance is defined as

$$\frac{\{sum\ of\ squares\ of\ deviations\}}{\{degrees\ of\ freedom\}}$$

Ratio of
the variance of between-samples and
the variance of within-samples

Degrees of freedom of each term

$$\sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X})^2 = \sum_{i=1}^{3}\sum_{j=1}^{50}(\bar{X}_i-\bar{X})^2 + \sum_{i=1}^{3}\sum_{j=1}^{50}(x_{i,j}-\bar{X}_i)^2$$

Total sum of squares is between-samples sum of squares +within-samples sum of squares

- 3*50-1=149
- 3-1=2
- 3*(50-1)=147

k: # of methods(treatments)
n: # of data within the method

- k*n - 1
- k – 1
- k*(n-1)

- If every time the average value changes, the average is a constraint which decreases the degrees of freedom.

Calculate the variance between samples and the variance within samples

$$\sum_{i=1}^{3}\sum_{j=1}^{50}\left(x_{i,j}-\bar{X}\right)^2 = \sum_{i=1}^{3}\sum_{j=1}^{50}\left(\bar{X}_i-\bar{X}\right)^2 + \sum_{i=1}^{3}\sum_{j=1}^{50}\left(x_{i,j}-\bar{X}_i\right)^2$$

Total sum of squares is between-samples sum of squares +within-samples sum of squares

- $\dfrac{\sum_{i=1}^{3}\sum_{j=1}^{50}\left(\bar{X}_i-\bar{X}\right)^2}{2} = 38.8333$
- $\dfrac{\sum_{i=1}^{3}\sum_{j=1}^{50}\left(x_{i,j}-\bar{X}_i\right)^2}{147} = 0.566189$

- Ratio of the variance between samples and the variance within samples is called the test statistics F 68.5872

- F follows F distribution(α, β)

α: *degrees of freedom for the numerator*

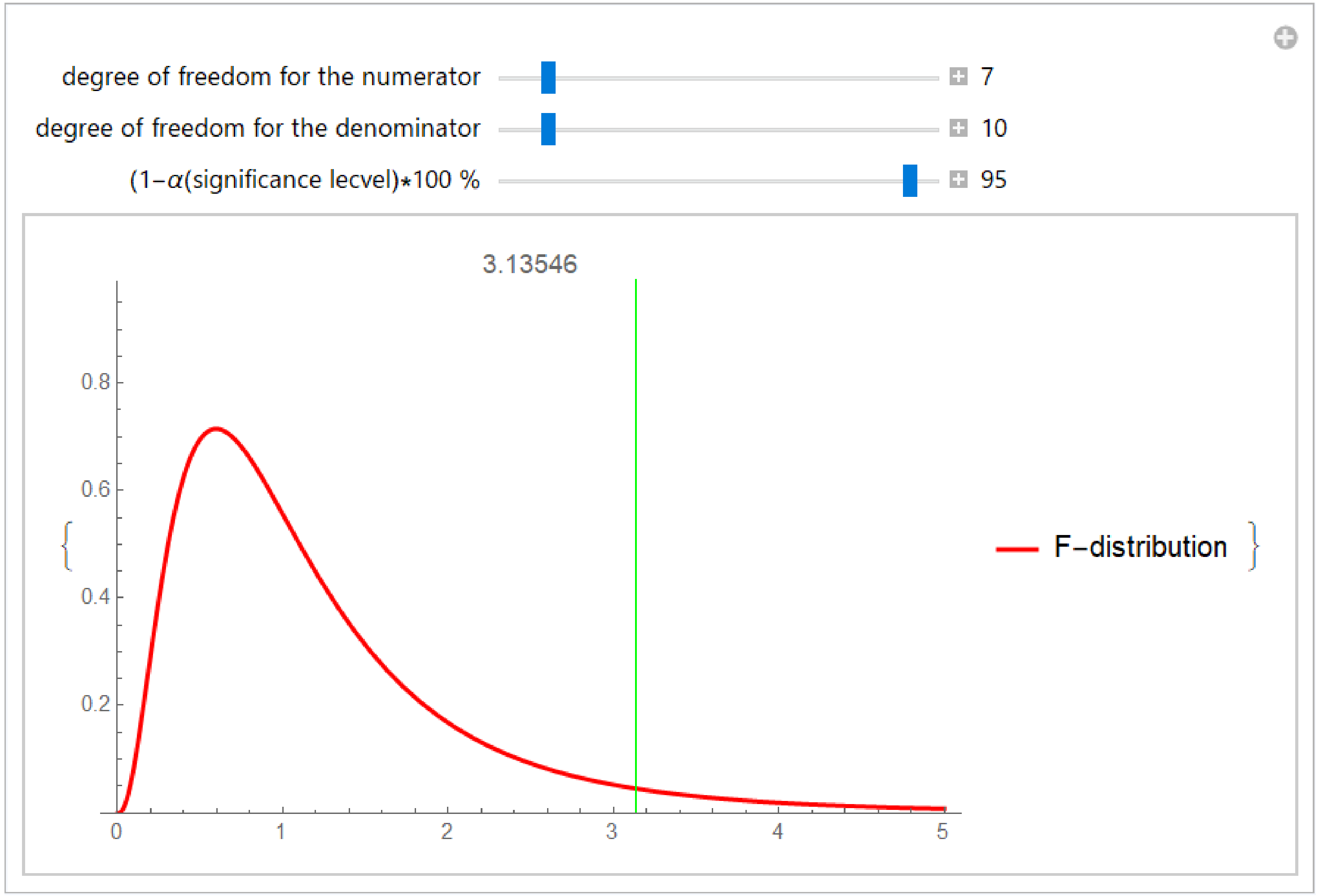
β: *degrees of freedom for the denominator*

Sample of F-distribution

Please install Wolfram CDF player which is free.

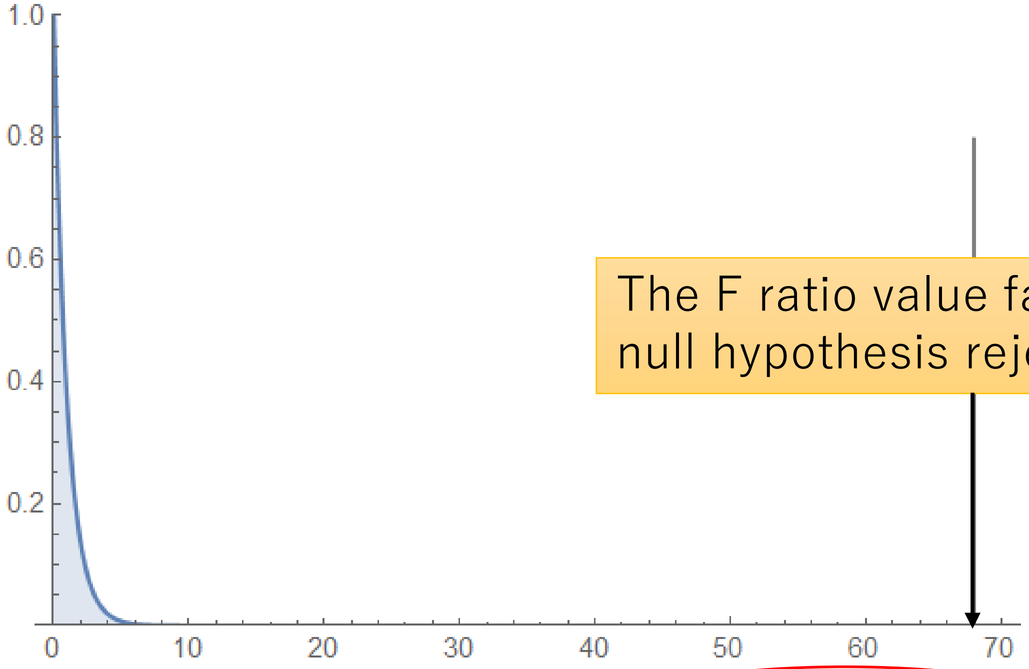
https://www-cc.gakushuin.ac.jp/~20010570/VDStat/Fdist.cdf

- F distribution with 7 and 10



Compare the F-distribution for df(2, 147)

Probability density



The F ratio value falls in the null hypothesis rejection region.

{ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	77.6666	38.8333	68.5872	9.10621×10^{-22}
Error	147	83.2298	0.566189		
Total	149	160.896			

Almost 0

, CellMeans →

All	3.85639
Model[1]	4.79577
Model[2]	3.72557
Model[3]	3.04782

}

The effect of treatments is larger than the noise effect.

Make a decision

- Not all three methods population averages are equal.

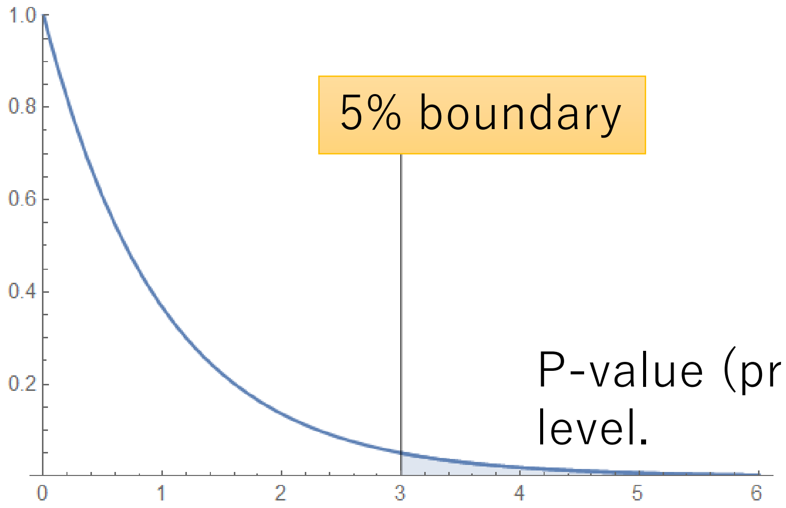
{ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	77.6666	38.8333	68.5872	9.10621×10^{-22}
Error	147	83.2298	0.566189		
Total	149	160.896			

, CellMeans →

All	3.85639
Model[1]	4.79577
Model[2]	3.72557
Model[3]	3.04782

}



P-value (probability of the F ratio) is smaller than the given significant level.

Comparisons A vs C, A vs B, and B vs C

- A vs C

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	76.383	76.383	120.365	9.57598×10^{-19}
Error	98	62.1902	0.634594		
Total	99	138.573			

, CellMeans →

All	3.9218
Model[1]	4.79577
Model[3]	3.04782

- A is superior to C
- A vs B

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	28.6335	28.6335	34.5199	5.80839×10^{-8}
Error	98	81.2889	0.829478		
Total	99	109.922			

, CellMeans →

All	4.26067
Model[1]	4.79577
Model[2]	3.72557

- A is superior to B
- B vs C

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	11.4833	11.4833	48.9706	3.24597×10^{-10}
Error	98	22.9805	0.234495		
Total	99	34.4638			

, CellMeans →

All	3.3867
Model[2]	3.72557
Model[3]	3.04782

Summary of ANOVA

- Please do not say method A is better, only because the average is higher than others.
- Remember
  - Variance ratio follows the F-distribution
  - Variance definition

$$\frac{\{sum\ of\ squares\ of\ deviations\}}{\{degrees\ of\ freedom\}}$$

Conclusion of this talk

- Let's infer the population average/variance more precisely, with the power of statistics.
- KEYWORD here appeared:
  - Normal distribution
  - 95% confidential interval
  - Inferential statistics
  - Degrees of freedom
  - Hypothesis testing, Null hypothesis, Alternative hypothesis
  - Significant level, Rejection region
  - ANOVA
  - Test statistics F
  - F-distribution