

Research Methods in Political Science I

2. Statistical Computing with R

Yuki Yanai

October 14, 2015



KOBE UNIVERSITY

Today's Menu



- 1 Basics of Quantitative Methods
 - What Are Data?
 - Summarizing Data
- 2 Probability and Probability Distribution
 - Probability
 - A Variety of Probability Distributions
- 3 Inferential Statistics
 - Essentials of Statistical Inference
 - Sampling Distribution
 - Central Limit Theorem

R Codes and Explanation



Course Website :

URL <http://yukiyanai.com>

- Classes
- Research Methods in Political Science I
- Class Materials
- Statistics with R

What Are Data?

Data



Data: data (pl) - datum (s)

Information collected by survey or observation

- Quantitative data: age, income, population, GDP, etc.
 - Qualitative data: gender, party ID, electoral participation, etc.

What Are Data?

What Kind of Data Are We Interested in?



- data the values of which change by unit: data containing **variables**
 - different income for different people
 - Not interested in constants
 - Occupation of students, sex of girl's high school students, etc.

Variable!



- What are variables (変数) ?
 - a variable has values that vary by unit
 - a variable takes a variety of (at least two different) values: a variable has a **distribution**
 - a variable's value does not have to be a number
- a constant has a single value for all units

Types of Variables



- Quantitative variables
 - ratio scale
 - interval scale
- Qualitative variables
 - ordinal scale
 - nominal scale

What Are Data?

Characteristics of Variables



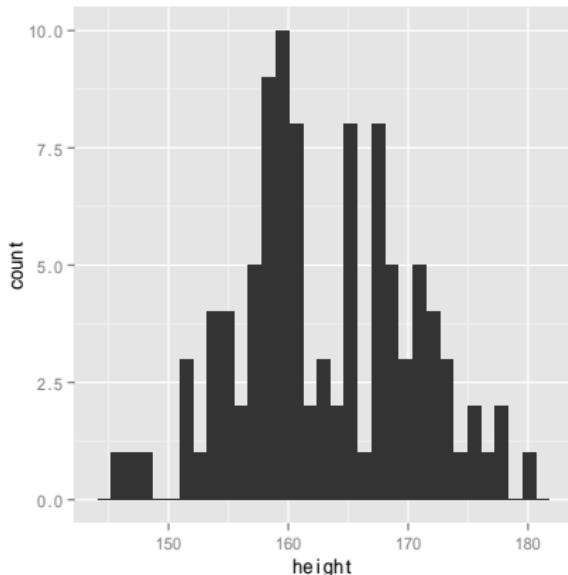
		We can tell			
	Type	identity	order	difference	ratio
Qualitative	nominal	<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
	ordinal	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
Quantitative	interval	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
	ratio	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Summarizing Data

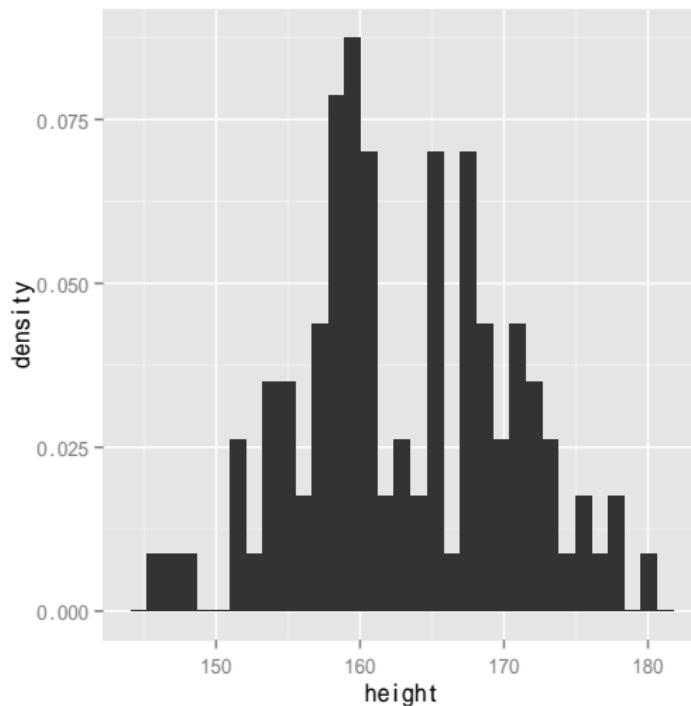
Visualize Data: Histograms



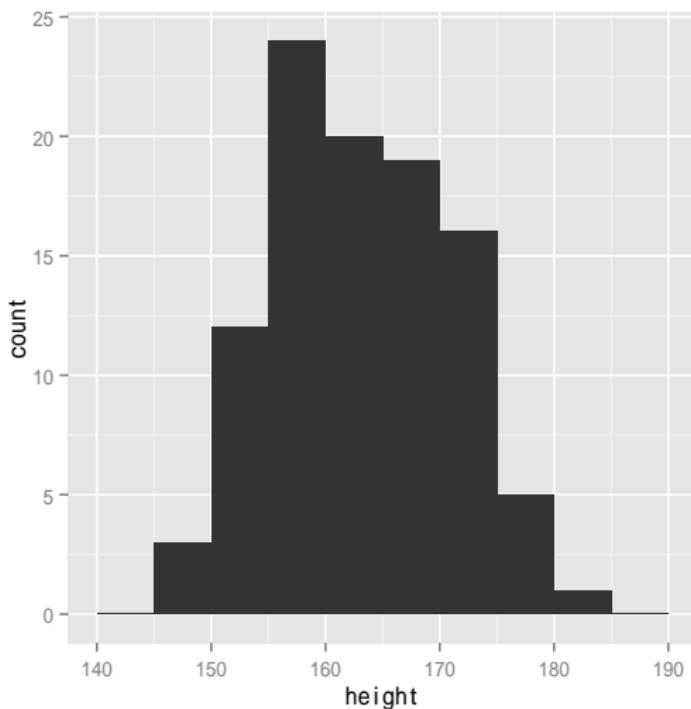
- Examine a variable's distribution by histograms
 - center of distribution?
 - symmetric?
 - how many peaks?
 - range?
 - Figure: distribution of height



Histogram: Density for Vertical Axis

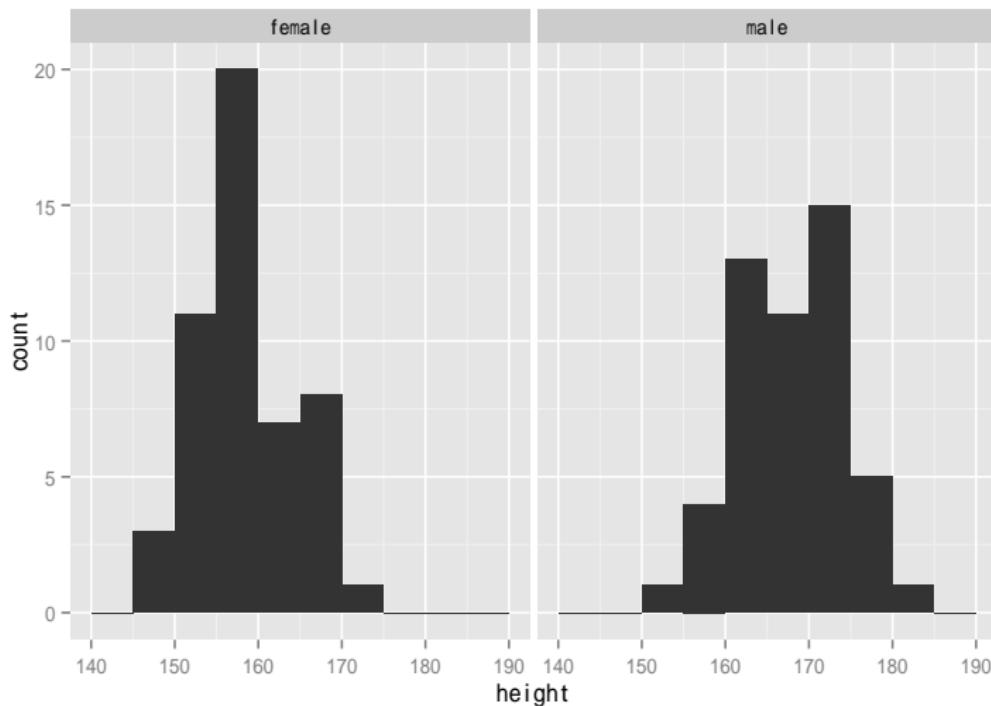


Histogram: Modify Bin Width





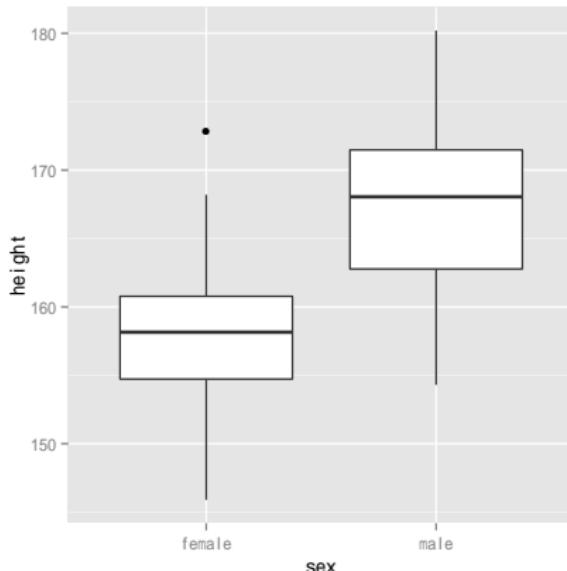
Histogram: Grouping by Gender



Visualize Data: Box (Box-and-Whisker) Plots



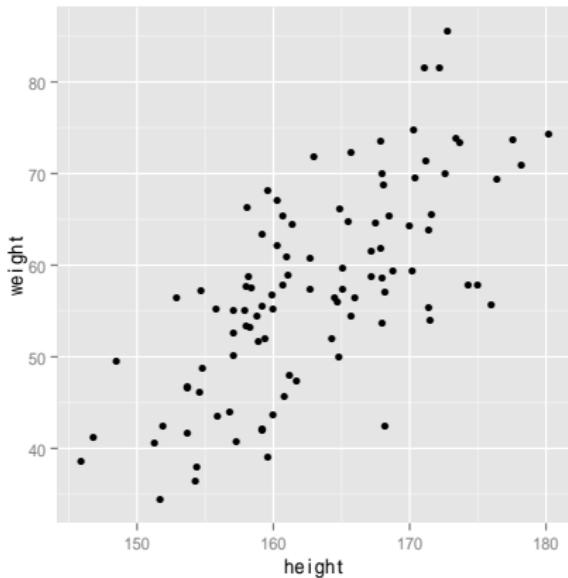
- Compare distributions of variables (or groups in a variable)
 - Shows the minimum (except outliers), the first quartile, the median, the third quartile, and the maximum (except outliers)
 - Figure: Distribution of height by gender



Visualize Data: Scatter Plots

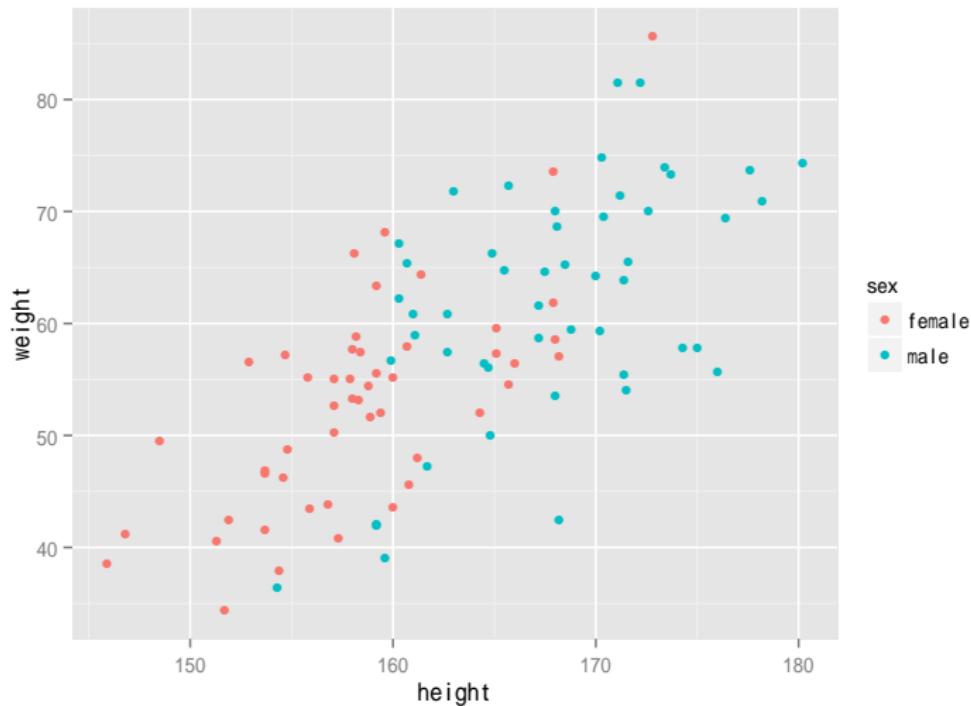


- Examine the relationship between two variables by scatter plots
 - Note that we detect some patterns even if there are no relationship between two variables
 - Figure: Scatter plot of weight vs. height



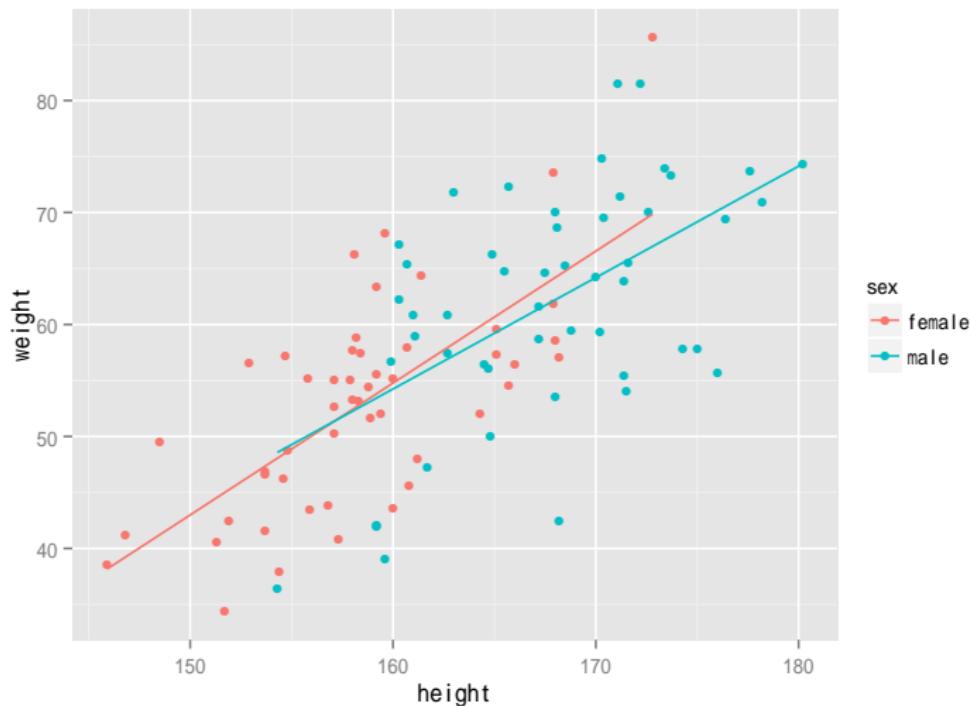
Summarizing Data

3D Scatter Plot: Grouping by Color



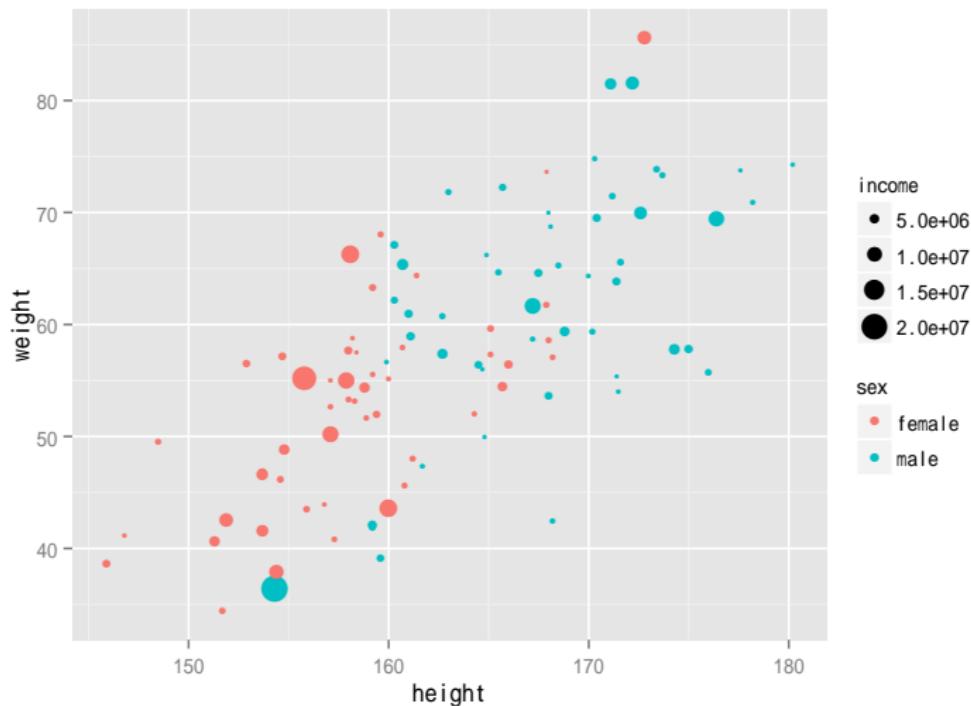
Summarizing Data

Scatter Plot: Impose a line



Summarizing Data

4D Scatter Plot



5D Scatter Plot!



Watch a YouTube Video:

<https://youtu.be/jbkSRLYSojo>

Statistics



What are statistics? – statistic (s)

- a formula to show a characteristic of variables
- can be obtained by some algorithms
- variety of statistics

Summarizing Data

A Measure of Central Tendency: Mean



- The **mean** of a variable x (算術平均, 相加平均) : \bar{x} (reads “x bar”)

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^n x_i}{n} \\ &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n)\end{aligned}$$

- Means locate the gravity center of histograms
- Means are sensitive to outliers

A Measure of Central Tendency: Median



- The **median** of a variable x (中央値, 中位値) m satisfies

$$\int_{-\infty}^m dF(x) \geq \frac{1}{2} \quad \text{and} \quad \int_m^{\infty} dF(x) \geq \frac{1}{2},$$

where $F(x)$ is CDF of x

- the middle value when we order the variable from the smallest (largest) to the largest (smallest)
- If there is no value in *the* middle (if the number of observations n is even), use the mean of two middle values
- Insensitive to outliers
- The median equals the mean for a variable with symmetric distribution

A Measure of Variability: IQR



- The inter-quartile range (**IQR**) of a variable x is the difference between the third quartile point ($Q_{3/4}$) and the first ($Q_{1/4}$)

$$\text{IQR} = Q_{3/4} - Q_{1/4}$$

- Quartile points are the 0th (i.e. min), 25th, 50th (median), 75th, and 100th (max) percentiles.
 - E.g., $x = \{0, 1, 1, 2, 4, 8, 9, 10\}$
 - The first and third quartiles are 1 and 8.5, respectively (*not* 2.5 and 7.5)
- IQR is the height of the box in a box plot
- IQR is used to detect outliers
 - Values outside $[Q_{1/4} - 1.5\text{IQR}, Q_{3/4} + 1.5\text{IQR}]$ are considered as outliers

Summarizing Data

A Measure of Variability: Variance



- The unbiased **variance** of a variable x (不偏分散) : u_x^2

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

- Shows variability of a variable: more variable as the variance gets larger
- $u_x \leq 0$ ($u_x = 0$ is x is constant)

Replace $n - 1$ with n in the formula, and you get (biased) variance. Some call the unbiased one the variance

A Measure of Variability: Standard Deviation



- The standard deviation (sd ; 標準偏差) of a variable x is the square root of the variance
- Let u_x denote the sd of a variable x :

$$u_x = \sqrt{u_x^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

- Keeps the measuring unit of a variable while the variance squares the unit

Some call u_x the square root of the unbiased variance and define the sd by the formula where $n-1$ replaced with n .

Definition of Probability



Assign a value $\Pr(A)$ to an event A in the sample space S . We call $\Pr(A)$ a probability of A if it has the following three properties.

- ① Any event has a non-negative probability.

$$\Pr(A) \geq 0$$

- ② The probability of the sample space is 1.

$$\Pr(S) = 1$$

- ③ For events that are mutually exclusive, the probability satisfies the following summation rule:

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$$

Probability Distribution



- A in $\Pr(A)$ can be one of a number of different things (A take a variety of values)
- Give a probability to each event: **probability distribution**
 - E.g. Flip a “fair” coin twice, and count how many heads you get
 - $S = \{0, 1, 2\}$
 - $\Pr(0) = 1/4, \Pr(1) = 1/2, \Pr(2) = 1/4,$
- A variety of distributions: describe them by probability mass function or probability density function and cumulative distribution function.

Probability Mass Function (PMF)



- For a *discrete* random variable X , we use **PMF** (確率質量関数) to describe its distribution.
- When the sample space is defined as $S = \{x_1, x_2, \dots\}$, PMF of X can be written as:

$$f_X(x_i) = \Pr(X = x_i) = \Pr(x_i).$$

Probability Density Function (PDF)

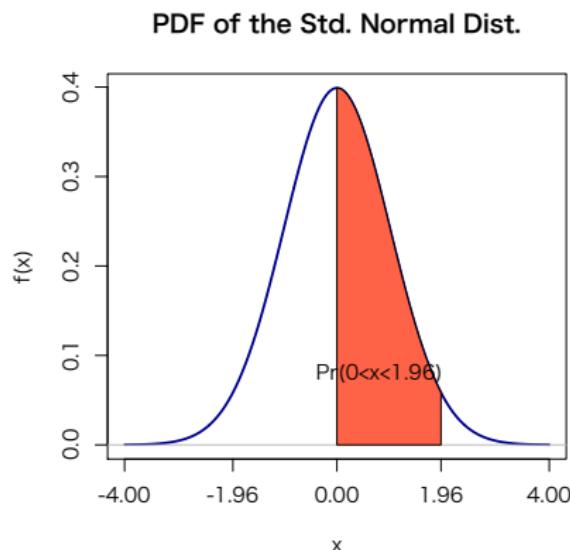


- For a *continuous* random variable X , we use **PDF (確率密度関数)** to describe its distribution
- The probability of X taking a specific value x_i is 0:
 $\Pr(X = x_i) = 0$
- We consider the probability of an interval: the probability of X taking a value in $[a, b]$:

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x) dx$$

Example of PDF: Standard Normal Distribution

- PDF of the standard normal distribution (figure)
- Horizontal axis: values that the variable can take
- Vertical axis: **probability density**
- Probability: Area under the PDF
 - Probability of x taking a value in $[0, 1.96]$: the colored area in the figure



Cumulative Distribution Function (CDF)



- We also use **CDF** (累積分布関数) to describe probability distributions
- CDF is usually denoted by F_X :

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f(x)dx$$

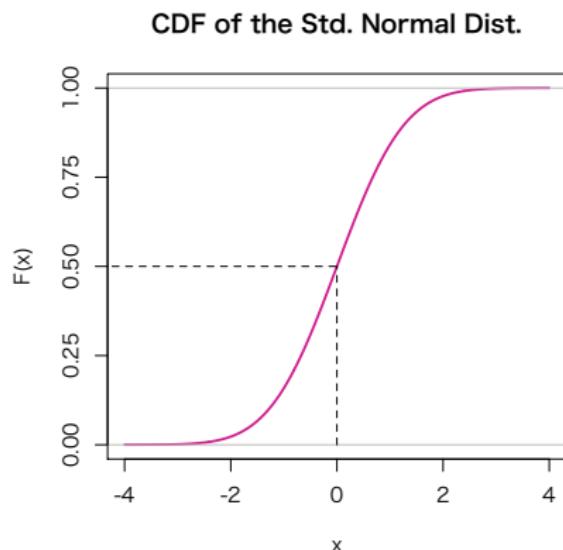
- PDF is the derivative of a CDF

Probability

Example of CDF: Standard Normal Distribution



- CDF of the standard normal distribution (figure)
- Horizontal axis: the values the variable can take
- Vertical axis: **probability** – $\Pr(X \leq x)$
 - Dashed line shows the probability of X taking **0 or less**
 - *Not* the probability of X being 0



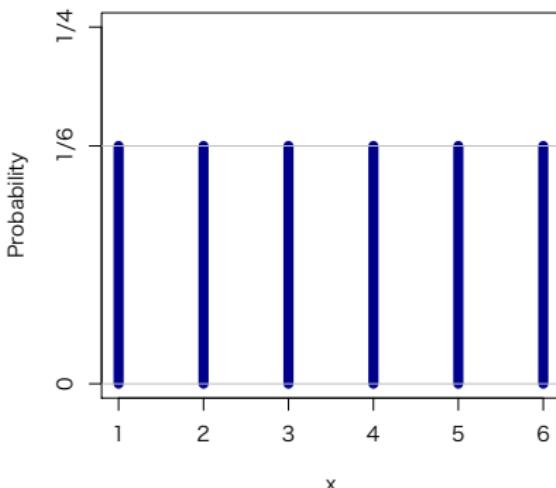
A Variety of Probability Distributions

Discrete uniform distribution (離散一様分布)



- A random variable X takes n different values with the same probabilities
- E.g. Probability distribution of rolling a “fair” die once (figure)

PMF of a Discrete Uniform Dist.: A Die



A Variety of Probability Distributions

Continuous Uniform Distribution (連續一樣分布) : $U(a, b)$

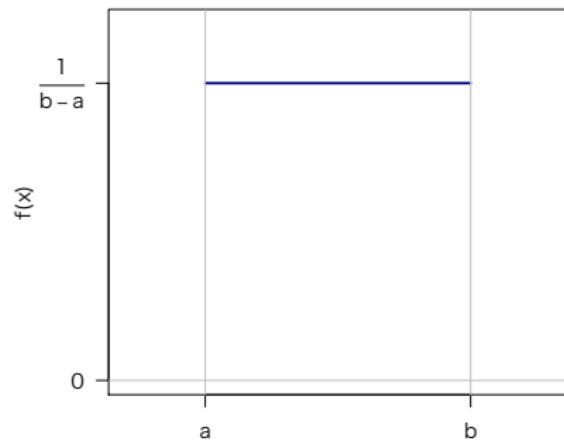


$$X \sim \mathsf{U}(a, b)$$

$$f(x) = \frac{1}{b-a} \quad (a \leq x \leq b)$$

- Parameters: the minimum a and the maximum b
 - Constant (i.e. uniform) density in the interval $[a, b]$

PDF of a Continuous Uniform Dist.



A Variety of Probability Distributions

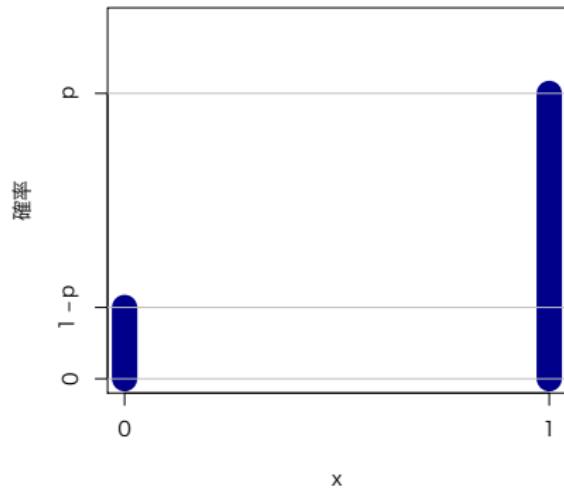
Bernoulli distribution (ベルヌーイ分布) : $\text{Ber}(p)$ 

$$X \sim \text{Ber}(p)$$

$$f(x) = p^x(1-p)^{1-x}$$

PMF of Bernoulli ($p = 0.8$)

- X takes either 1 (success) or 0 (failure)
- Takes the value of 1 with probability p and 0 with $1 - p$
- E.g., Flipping a fair coin: Bernoulli with $p = 1/2$
- Parameter: probability of success p



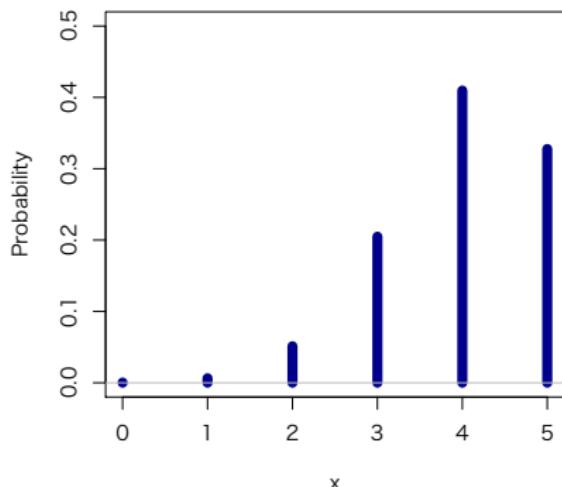
A Variety of Probability Distributions

Binomial Distribution (二項分布) : $\text{Bin}(n, p)$ 

$$X \sim \text{Bin}(n, p)$$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Distribution of the number of successes for n Bernoulli trials with p
- Parameters: the number of trials n and the probability of success for a trial p
- The possible numbers of successes $k = 0, 1, \dots, n$

PMF of Binomial ($p=0.8, n=5$)

A Variety of Probability Distributions

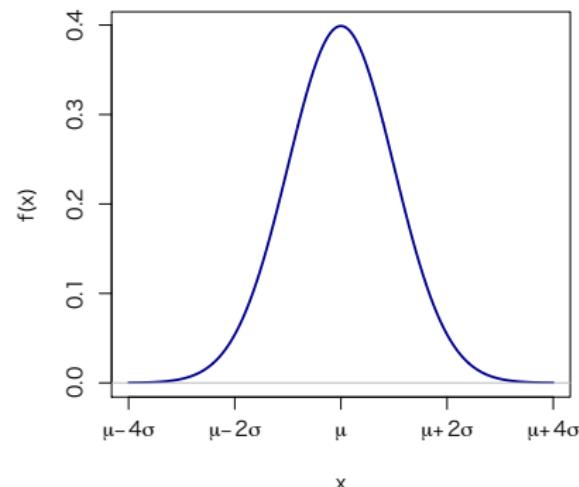
Normal Distribution (正規分布) : $N(\mu, \sigma^2)$ 

$$X \sim N(\mu, \sigma^2) \text{ (or } N(\mu, \sigma))$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- Parameters: the mean μ and the variance σ^2
- Symmetry about the center μ
- About 68% of the values are in $\mu \pm \sigma$
- About 95% of the values are in $\mu \pm 1.96\sigma$

PDF of the Std. Normal Dist.



A Variety of Probability Distributions

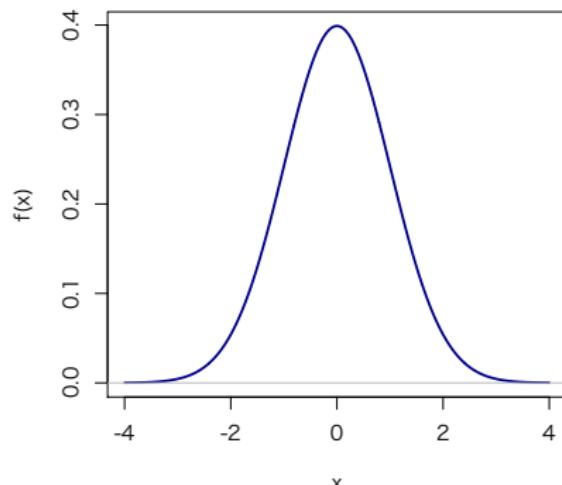
Standard Normal Distribution (標準正規分布) : $N(0, 1)$ 

$$X \sim N(0, 1)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

- Special case of $N(\mu, \sigma^2)$ where $\mu = 0$ and $\sigma^2 = 1$
- Symmetric about the center 0
- About 68% of the values are in $[-1, 1]$
- About 95% of the values are in $[-1.96, 1.96]$

PDF of the Std. Normal Dist.



A Variety of Probability Distributions

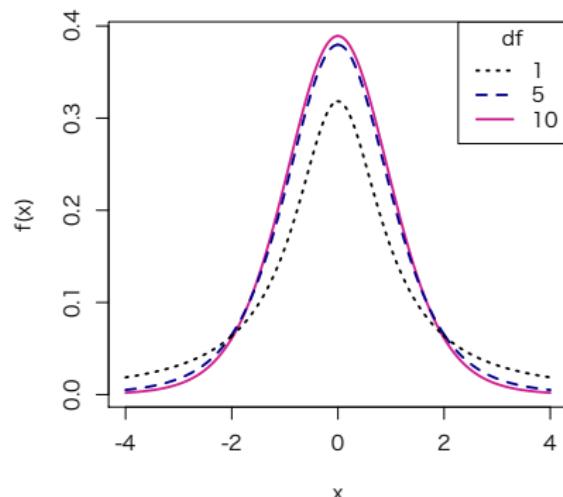
Student's t distribution (t 分布) : $t(v)$ 

$$X \sim t(v)$$

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

- Parameter: the degree of freedom (df) v (a positive real number)
- If n is not large enough, the sampling distributions will follow a t distribution
- Converge to the standard normal as $v \rightarrow \infty$

PDF of Student's t Dist.



A Variety of Probability Distributions

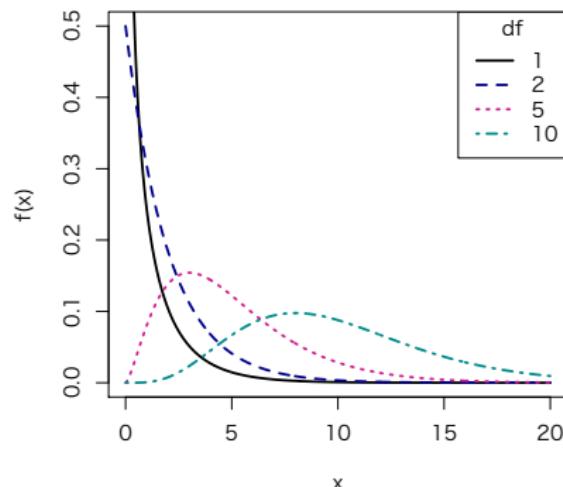
 χ^2 Distribution (カイ二乗分布) : $\chi^2(k)$ 

$$X \sim \chi^2(k)$$

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \quad (x \geq 0)$$

- Parameter: df k (a natural number)
- Used for the test of independence for cross tabulated data
- Approach (slowly) to a normal distribution as $k \rightarrow \infty$

PDF of Chi-squared Dist



Other Important Distributions



- F Distribution (F 分布)
- Gamma Distribution (ガンマ分布)
- Negative Binomial Distribution (負の二項分布)
- Poisson Distribution (ポアソン分布)
- Dirichlet Distribution (ディリクレ分布)
- Cauchy Distribution (コーチー分布)
- etc.

Population and Samples



- Population (母集団) : target of study – usually not observable or data not available
- Sample (標本) : A part of the population – observable and data available
- Sampling (標本抽出) : to choose a part of the population

The Number of Samples and the Sample Size



Don't confuse the number of samples (標本数) with the sample size (標本サイズ) – For Japanese speakers: 標本サイズのことを標本数と呼ばない！

Sample size (標本サイズ)

The number of observations for each variable contained in a sample: normally denoted as n

The number of samples (標本数)

The count of sets of n -observation samples

E.g., Take two samples from the population of 100 million people. Call the first sample containing 2,000 people S1 and the second 1,500 S2. In this situation, the number of samples is 2 (S1 and S2), and the sample sizes are 2,000 and 1,500 for S1 and S2, respectively.

Infer a Whole from Its Parts



We usually obtain a sample(s) only

To investigate if Japanese voters (population) support the Abe Administration

Sample 2,000 people from the population, and ask if they support or not

Use information obtained by the sample (a part) to infer the population (the whole)

Japanese voters support Abe or not?

Using the answers by 2,000 respondents, infer the support rate of Abe gov't: infer the population rate by the sample rate

Statistical inference

Some Caveats



- We never know the population fully from the sample
- Inference always comes with **errors**
- We try to minimize the inferential errors
- We need to report the degree of errors when we publish results

Example of Statistical Inference



Investigate if Japanese voters support Abe govt

Sampled 1,000 people from the Japanese electorate, and asked them if they support Abe govt or not. As a result, 542 answered "Yes".

Infer the population support rate by the sample support rate

Infer the Population

Because 542 out of 1,000 people support the govt, 54.2% of Japanese voters should support the govt.

The best guess (point estimate) is 54.2, but how about the error?

Example of Statistical Inference (cont.)



Investigate if Japanese voters support Abe govt

Sampled 1,000 people from the Japanese electorate, and asked them if they support Abe govt or not. As a result, 542 answered "Yes".

- Best guess (point estimate): 54.2%
- Standard error (SE): 0.016
- The (rough) 95% confidence interval: [sample rate -2SE , sample rate $+2\text{SE}$] = [51.0%, 57.4%]
- What is **SE**?

Sampling Distribution



- There are a number of ways to sample n units from the population of size N_{pop}
- Take every possible combination of n units (i.e., take $\binom{N_{\text{pop}}}{n}$ different samples of size n)
- Calculate the sample mean for each sample. Then, the sample means distribute.
- We call this distribution the **sampling distribution** (of sample means).

Sampling Distribution

Example of Sampling Distribution



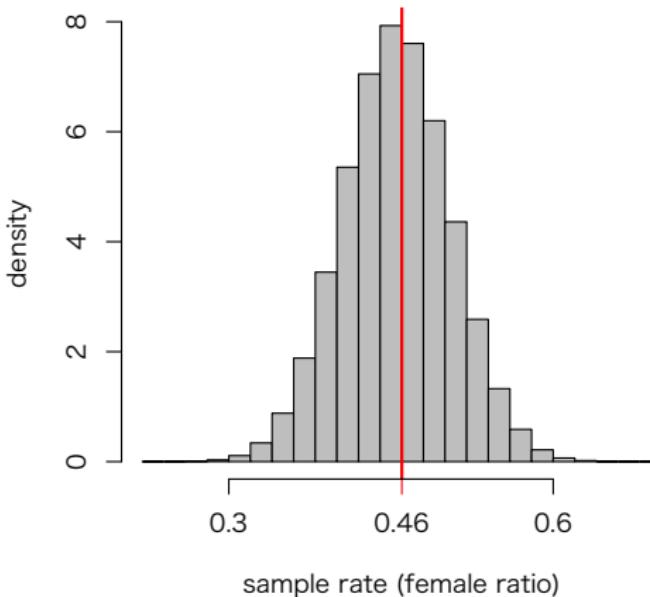
- Choose 100 people from 4,600 females and 5,400 males
- There are 6.5×10^{241} different ways to choose 100 people
- Distribution of the female ratio for each of 6.5×10^{241} :
Sampling distribution (of female ratio)
- In practice, we can't obtain the sampling distribution because there are too many ways to choose a sample
- Let's run a simulation in R and get a pseudo sampling distribution

Sampling Distribution

Pseudo Sampling Distribution: Sampling Distribution of Sample Rates Where the Population Rate $\pi = 0.46$



Pseudo Sampling Distribution: n=100



Standard Errors (SE)



- **SE (標準誤差)** : variability in sampling distributions – the standard deviation of a statistic
- When the population is large enough (about 100 times larger than the sample size n),

$$\text{SE} = \frac{u}{\sqrt{n}},$$

where n is the square root of the unbiased variance.

- **SE gets smaller as n gets larger**

Population Size and the Sample Size



Survey

Investigate if people are for or against a tax hike in the prefectures of Tokyo (population: 13 million) and Iwate (1.3 million). Should we set the sample size of Tokyo ten times larger than that of Iwate?

- Not necessarily.
- If the population ratios are the same between two, we get the same accuracy with the same sample sizes
- Why?: **SE doesn't depend on the population size**

Central Limit Theorem

Central Limit Theorem (CLT)



CLT (中心極限定理)

The sampling distribution of sample means is approximately normal if n is large enough, regardless of the distribution of the population.

We can use the properties of the normal distribution to infer a non-normal population, if n is large enough.

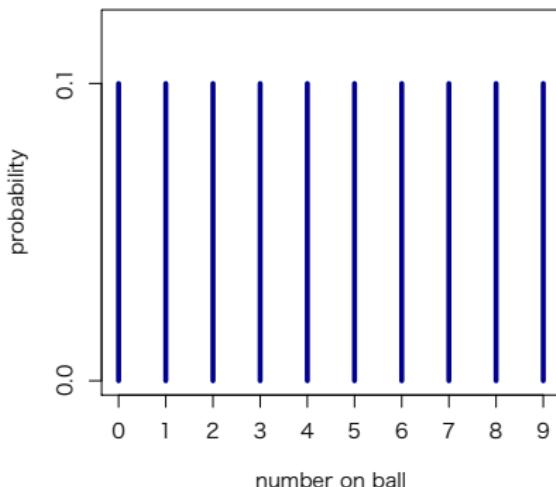
Central Limit Theorem

CLT Simulation: Discrete Uniform Distribution (1)



- A bag containing 10 balls
- Balls are numbered 0, 1, 2, ..., 9
- The mean of the ball number: $(9-0)/2 = 4.5$

Distribution of the ball numbers



CLT Simulation: Discrete Uniform Distribution (2)



- Suppose we don't know the numbers on the balls
- Try to estimate the population (true) mean by inference using some samples
- Randomly draw a ball from the bag n times, and use the sample mean as the point estimate
- Conduct sampling with replacement

CLT Simulation: Discrete Uniform Distribution (3)

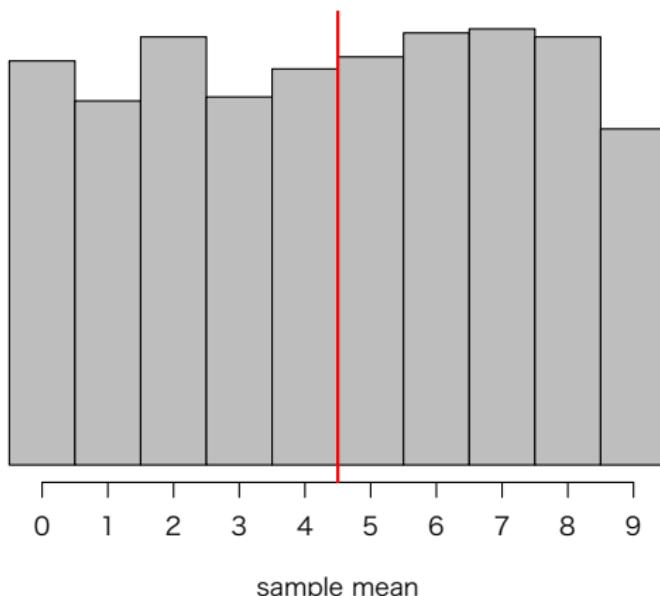


- Run a simulation in R
- Repeat the task of “sampling n balls and calculating the mean” 1,000 (or more) times
- Draw a histogram of 1,000 (or more) sample means!
- What happens if you gradually increase n from a small value to a large one

Simulation: $n = 1$



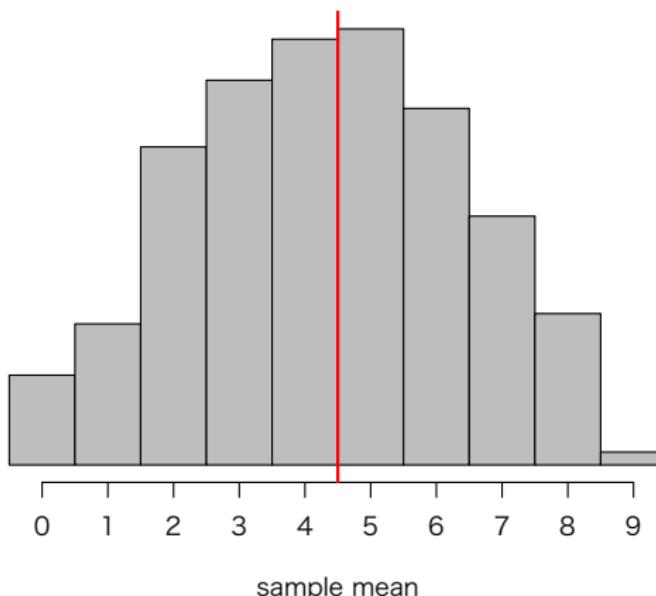
Sampling distribution from Uniform: $n=1$



Simulation: $n = 2$



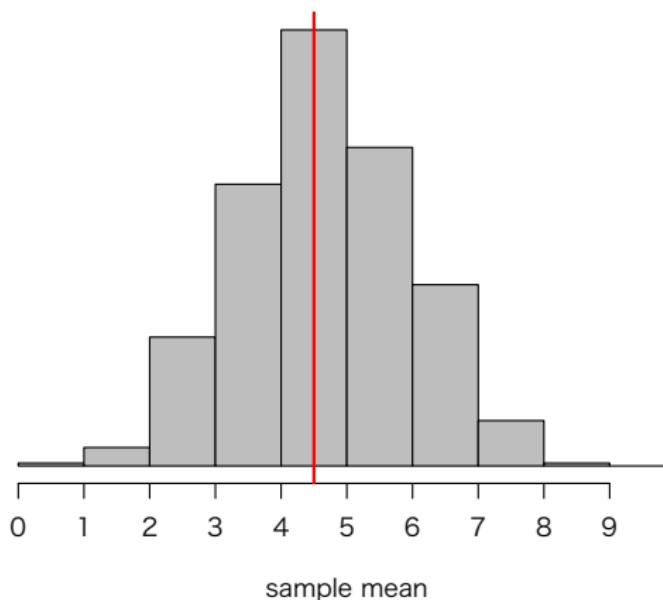
Sampling distribution from Uniform: n=2



Simulation: $n = 5$



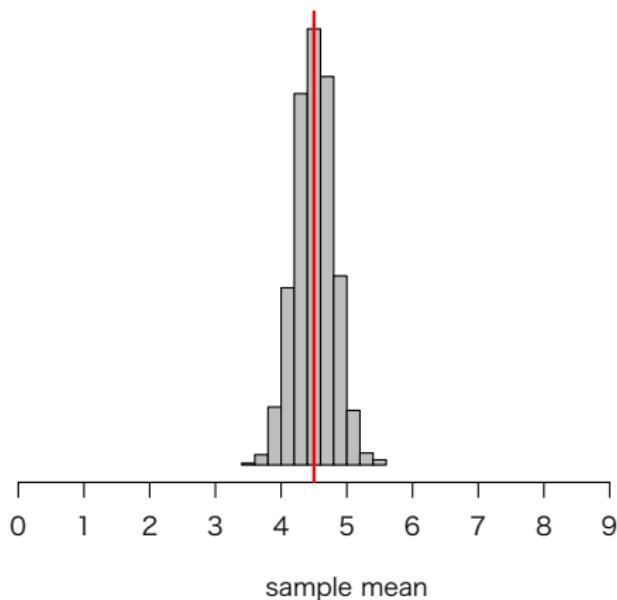
Sampling distribution from Uniform: n=5



Simulation: $n = 100$



Sampling distribution from Uniform: n=100



CLT Simulation from U(0, 1): $n = 1$



Central Limit Theorem

CLT Simulation from $U(0, 1)$: $n = 2$



Central Limit Theorem

CLT Simulation from U(0, 1): $n = 10$



Approximately Normal (if n is large enough)!



- Sampling distributions can be approximated by the normal if n is large enough, regardless of the population distribution
- The expected value of the sample means equals the population mean: **unbiasedness** (不偏性)
- Normal: 95% of the values are in $[\mu - 1.96\sigma, \mu + 1.96\sigma]$
- Sampling distribution approximated by the normal: 95% of the values are in $[\bar{x} - 1.96SE, \bar{x} + 1.96SE]$
- We can use this fact for statistical inference and test
- What if n is *not* large enough?: Use t distribution instead of the normal

Next Class



Reproducible Research

- What is reproducible research?
- Introduction to literate programming: R Markdown