Essential Mathematics for Global Leaders I

Spring 2019

Statistics

Lecture 5: 2019 May 27 – June 3rd

Xavier DAHAN
Ochanomizu Graduate Leading Promotion Center

Office:理学部2号館503
mail: dahan.xavier@ocha.ac.jp

Where are we ? Today's plan

PART I. Notions of Probability 必要な確率論

3. Sampling distribution and Central Limit Theorem 標本分布と中心極限定理

3.1 Introduction to Sampling　標本調査の概念

3.2 Law of Large Numbers (LoLN) and Central Limit Theorem (CLT) 大数の法則と中心極限定理

3.3 Application of CLT to infer the mean (CLTを用いて母平均を推測する)

3.4 More on sample statistics

Chapter 3: Sampling Distribution and Central Limit Theorem

3.1 Introduction to sampling

- Later in this course we learn how to make inferences about the population based on information in the sample. 本講義の後ほど、標本のデータから母集団について推測する。
- This is often done under the assumption that the population is approximately normally distributed. その推測は、正規分布と近似した母集団という仮定の下でしばしば行われる。

Parameters vs Sample Statistics 母数 vs 統計量

- In practice, sample statistics are used to estimate population parameters.
  実際には標本統計量は母集団の母数を推定することである。
  - A **parameter** is a numerical descriptive measure of a population. Its value is almost always unknown.
    母数とは母集団を記述する値である。ほとんどの場合においては未知である。
  - A **sample statistic** is a numerical descriptive measure of a sample. It can be calculated from the observations (=measurements=data)
    統計量とは母集団から抽出される標本の測定である。標本の観測を用いて計算できる値である。
    Example: mean, variance, median, quantile etc.

Sampling as random variables

- Sampling is assumed to be a random process
  ☞ generally modeled by random variables
  標本は無作為の過程だと仮定する
  ☛大抵、確率変数によってモデル化される
- Each element in the sample is *often* chosen independently from the others ☛ independent random variables
  標本の元が*しばしば*独立に抽出される
  ☞ 独立確率変数によってモデル化される
- A sample of size n is one subset of size n of the whole population (among all possible subsets of size n).
  抽出したサイズnの標本は、母集団のサイズnの部分集合の中で一つである。
  ☛ input: sample space of the sample is: Ω^n (if Ω is much larger than the sample☞ see slides 13-14)
  ☞ 入力：標本の標本空間：Ω^n (Ωは標本の大きさに比べて充分に大きいのとき☞ ページ13-14を参照する）

Sample statistics as random variables

- Measurement (or observations or data): 観測
- Example: Ω =a population
  - Example of measurements: Height, Weight, Gender, etc.
    想定の例：身長、体重、性別、。。。
  - Given a sample of 100 persons 100人からなる標本
  - Height: 100 random variables: H_1, H_2,…,H_{100} that gives the height of the 1st ,2nd ,…,100th person
    H_i: 第i人目の身長
  - Sample mean: $\overline{H} = \frac{1}{100}(H_1 + H_2 + \cdots + H_{100})$
    標本平均 (H_iの算術平均)
  - Sample space of $\overline{H}$: all subsets of 100 persons among Ω
  - The distribution of the sample mean $\overline{H}$ is a sample distribution.
    標本平均$\overline{H}$の(累積）分布は標本分布の例の一つである。

A (Very) Simple Example: sample mean

- Population is very small. Only $\{1,2,3\}$
- Sample Statistic: sample mean $\bar{x}$ of the sample.
  - Samples of size 1: $\bar{x}$ $E(\bar{X})$
    - $X_1 = 1$ mean: 1
    - $X_1 = 2$ mean: 2
    - $X_1 = 3$ mean: 3

    $$\frac{\sum \bar{x}}{3} = \frac{1+2+3}{3} = 2$$
  - Samples of size 2: $\bar{x}$
    - $X_1, X_2 = 1,2$ mean: 1.5
    - $X_1, X_2 = 1,3$ mean: 2
    - $X_1, X_2 = 2,3$ mean: 2.5

    $$\frac{\sum \bar{x}}{3} = \frac{1.5+2+2.5}{3} = 2$$
  - Sample of size 3: $\bar{x}$
    - $X_1, X_2, X_3 = 1,2,3$ mean: 2

    $$\frac{\sum \bar{x}}{1} = 2$$
- Sample mean $\bar{X}$ is a random variable on the set of samples.
  標本平均は標本の全集合による確率変数

Back 9
Back 12

Independent Identically Distributed Random Variable (i.i.d.r.v.) 独立同分布に従う確立変数

- All measurement are identically distributed: it is the same random variable on the same population!
  観測はすべて同じ母集団上の同様な分布に従う。
- Very often the measurements are independent
  測定は互いに独立である。
- In the example page 6, the $H_1, H_2, \ldots, H_{100}$ are i.i.d.r.v. (日本語でも使われている省略)。

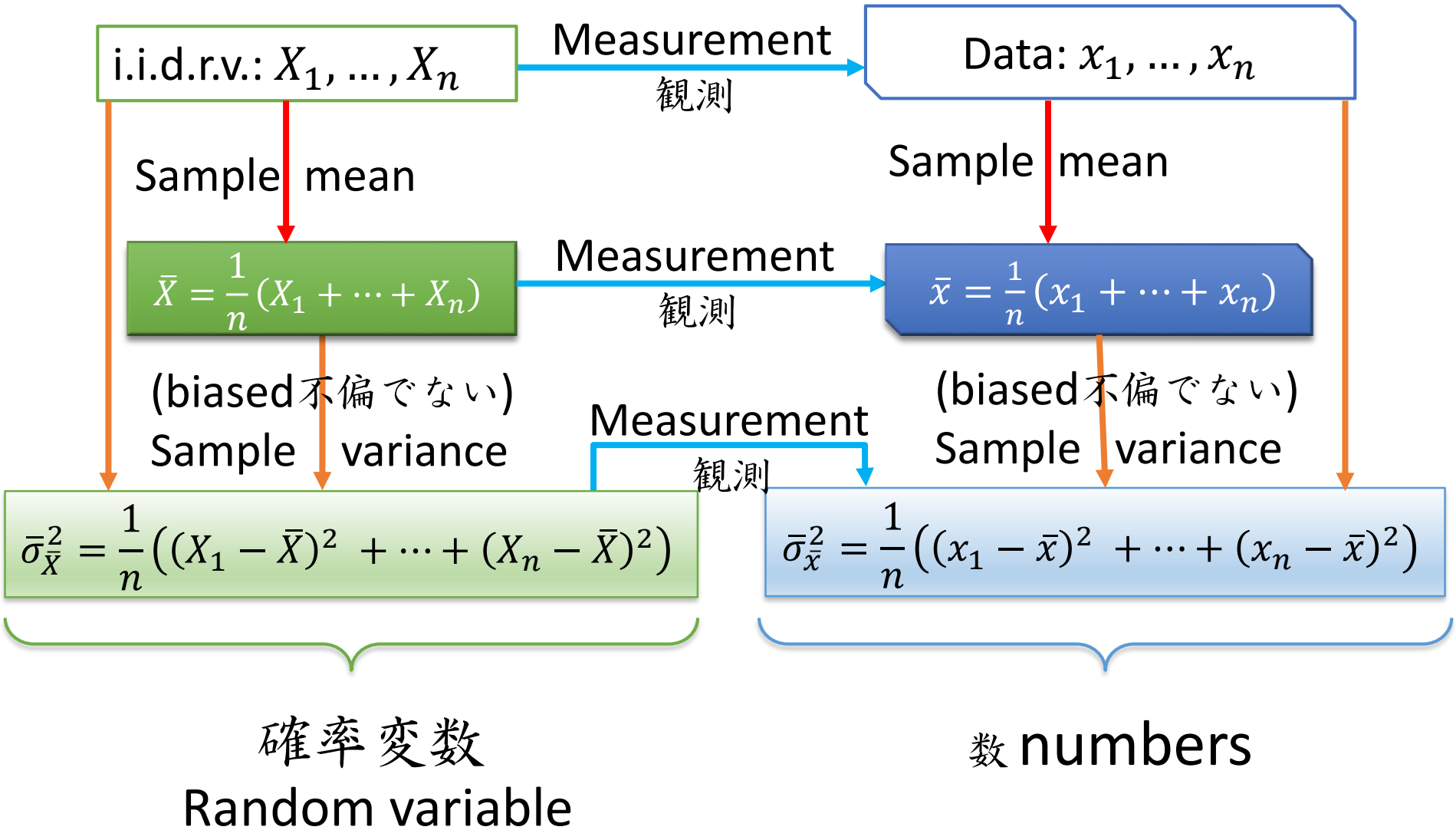
- Very often: sample of n measurements
  ⇔ n i.i.id random variables $X_1, X_2, \ldots, X_n$.

Point estimation, sampling distribution
点推定。標本分布。

- A **point estimator** $\widehat{\theta_n}$ of a parameter θ (attached to a population), obtained from a sample of size n is a number that estimates the parameter θ.
  母集団に付随する母数θの**点推定**$\widehat{\theta_n}$とは、θを推定するサイズn標本に付随する統計量である。
  Example: the sample mean (of size n) is a point estimator of the mean of the population (slide 7)

- The cumulative distribution function (cdf) of the point estimator $\widehat{\theta_n}$ is called the **sampling distribution.** (this is a function of the set of samples of size n).
  点推定$\widehat{\theta_n}$は、サイズnの標本による関数を見なしたら（統計量として）累積分布関数をもつ。これは**標本分布**という。

Sampling → i.i.d.r.v. or data ? Both



Standard error of the mean (SEM)
平均値の標本誤差

- Sample of n i.i.d measurements $X_1, X_2, \ldots, X_n$.

$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n) \quad (sample\ mean)$$

- Let $Var(X_i) = \sigma^2$ (variance of the population)
- $Var(\overline{X_n}) = \frac{1}{n^2}\left(Var(X_1) + \cdots + Var(X_n)\right)$
- $Var(\overline{X_n}) = \frac{\sigma^2}{n}$ $\qquad \sigma_{\overline{X_n}} = \frac{\sigma}{\sqrt{n}}$
- **Definition**: the standard deviation $\sigma_{\bar{X}_n}$ of (the sample distribution) of the sample mean $\bar{X}_n$ is called the standard error of the mean 標本平均の標本分布の標準偏差とは平均の標本誤差という。

Example slide 7 $\Omega = \{1,2,3\}$

Mean: $\mu_X = E(X) = \frac{1+2+3}{3} = 2$ and $Var(X) = \frac{1+1}{3} = \frac{2}{3}$

- Samples of size 1: $\bar{x}$ | $E(\bar{X})$ | $Var(\bar{X})$

  $X_1 = 1$ mean: 1
  $X_1 = 2$ mean: 2
  $X_1 = 3$ mean: 3

  $\frac{\sum \bar{x}}{3} = 2$

  $Var(\bar{X}) = \frac{1}{3}(1^2 + 0^2 + 1^2) = 2/3$

- Samples of size 2: $\bar{x}$

  $X_1, X_2 = 1,2$ mean: 1.5
  $X_1, X_2 = 1,3$ mean: 2
  $X_1, X_2 = 2,3$ mean: 2.5

  $\frac{\sum \bar{x}}{3} = 2$

  $Var(\bar{X}) = \frac{1}{3}\left(\frac{1}{4} + \frac{1}{4}\right) = 1/6$

- Sample of size 3: $\bar{x}$

  $X_1, X_2, X_3 = 1,2,3$ mean: 2

  $\frac{\sum \bar{x}}{1} = 2$

  $Var(\bar{X}) = 0$

Correction factor for finite populations
有限母集団の修正項

- Size 1: $Var(\overline{x_1}) = \frac{2}{3}$ Slide 11: $Var(\bar{x}_n) = \frac{\sigma^2}{n}$
  - σ^2/n=2/3 /1 = $\frac{2}{3}$
- Size 2: $Var(\bar{x}_2) = \frac{1}{6}$
  - $\sigma^2/n = 2/3\ /2 = \frac{1}{3} \neq \frac{1}{6}$????
- Size 3: $Var(\overline{x_3}) = 0$
  - $\sigma^2/n = 2/3\ /3 = \frac{2}{9} \neq 0$????

Correction factor
修正項 (slide ??)
$(N - n/N - 1)$

$\frac{1}{3} \cdot \frac{3-2}{3-1} = \frac{1}{6}$

$\frac{2}{9} \cdot \frac{3-3}{3-1} = 0$

Correction factor for finite populations
有限母集団の修正項 (続き)

- What happens??
- The population is finite (and small) and the samples are without replacement:
  母集団は有限で小さくて、復元のない抽出だから:
  - ☛ Samples are not independent.
  - ☛ 抽出は互いに独立ではない.
- Correction factor （有限母集団修正項）

$$Var(\bar{X}) = (\sigma^2/n)(N - n/N - 1)$$

- 母集団は大抵大きくて、あるいは無限で、この問題が起こらない。
  In general, the population is so big that this never happens. $N \gg n$.

Chapter 3: Sampling Distribution and Central Limit Theorem 3.2 LoLN and CLT

Motivation モチベション

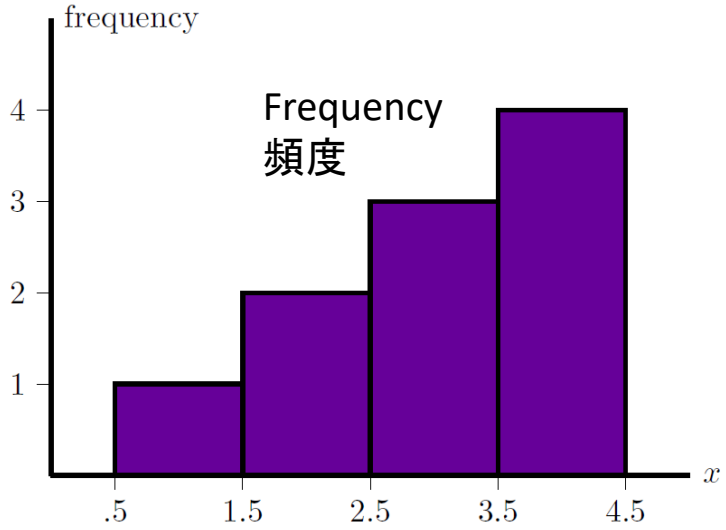
- When sample size $n \to \infty$, what is the distribution of the sample mean or other statistics ?
  標本の大きさnが、大きくなるとき、標本平均の分布または別の統計量の分布はどうなるか。

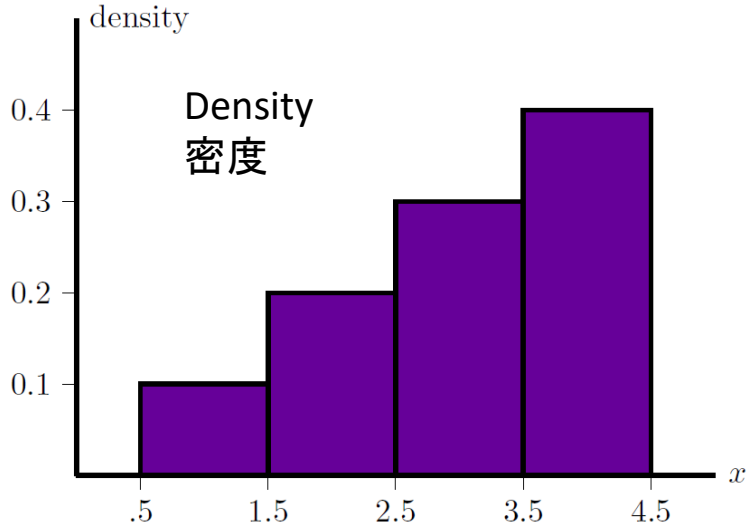
- The LoLN and CLT are fundamental tools in Statistics to derive asymptotic distribution.
  大数の法則と中心極限定理は統計学にわたって基本原理で、漸近分布を導出するために使われている。

Histograms ヒストグラム

- Made by “binning” data (　英：”to bin” 箱に入れる)
- Frequency (頻度) : height of bar over bin = number of data points in the bin　柱の高さ＝柱の中にあるデータ点の個数。
- Density (密度): area (面積) of bar is the fraction of all data points that lie in the bin. So total area is 1. 柱の面積は、柱の中にあるすべてのデータ点の分数。よって、全面積は１である。

$n = 10$ data





Example of binning data in a histogram

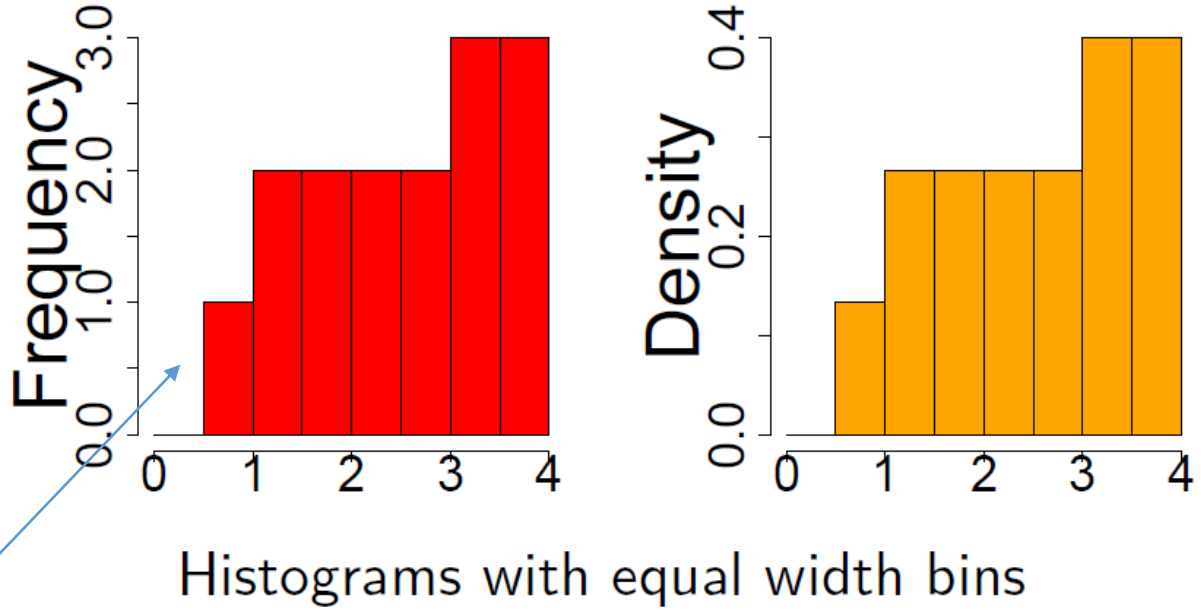
1. Make both a frequency and density histogram from the data below.
2. Use bins of width 0.5 starting at 0. The bins should be right closed.
   - 1 , 1.2 , 1.3 , 1.6 , 1.6
   - 2.1 , 2.2 , 2.6 , 2.7
   - 3.1, 3.2 , 3.4 , 3.8 , 3.9 , 3.9

3. Same question using unequal width bins with edges 0, 1, 3, 4.

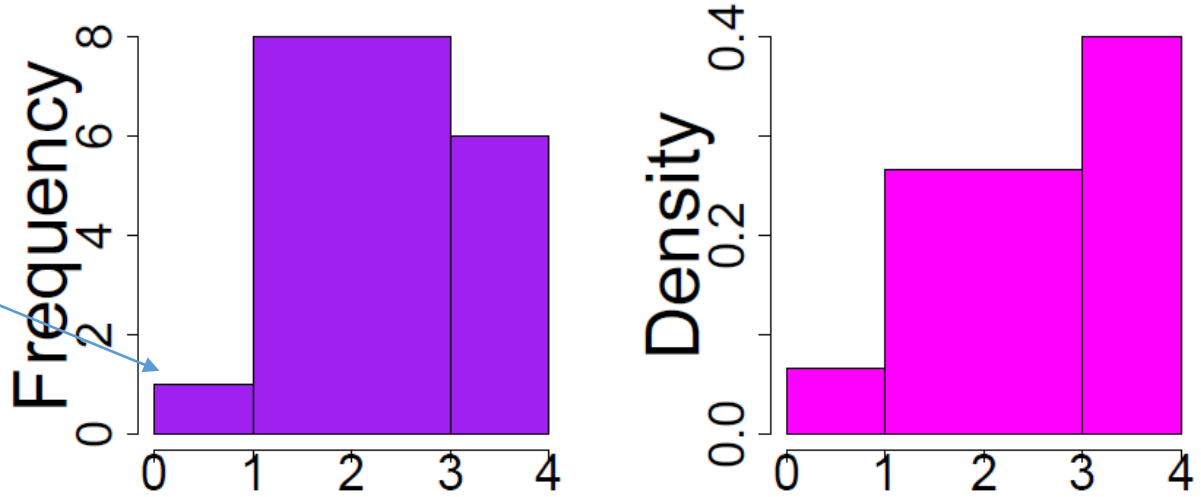
Result

Right closed bins:

Data point 1 is in the bin (0.5,1] or (0,1]



Histograms with equal width bins



Histograms with unequal width bins

Law of Large Numbers (LoLN) 大数の法則 数学

- **Informally**: an average of many measurements is more accurate than a single measurement.
  略式に：さまざまな観測の平均値のほうがただの一つ観測よりも精度である。
- **Formally**: Let $X_1, X_2, \ldots$ be i.i.d random variables all with expected value μ and standard deviation σ. Let $\overline{X_n}$ be the sample mean random variable:

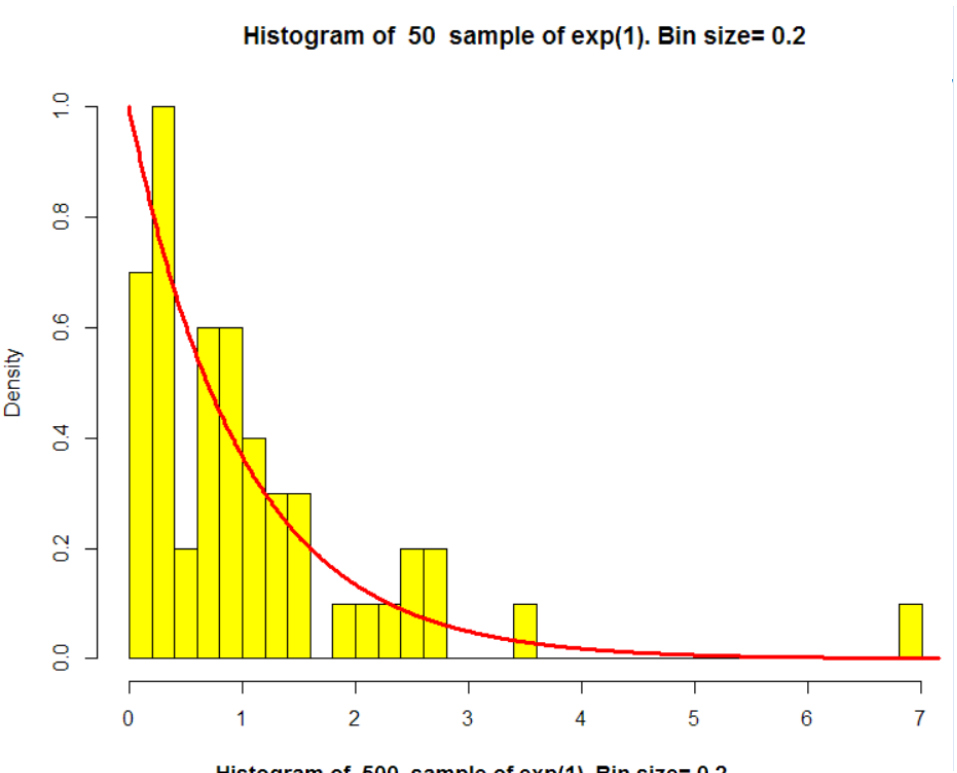
$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

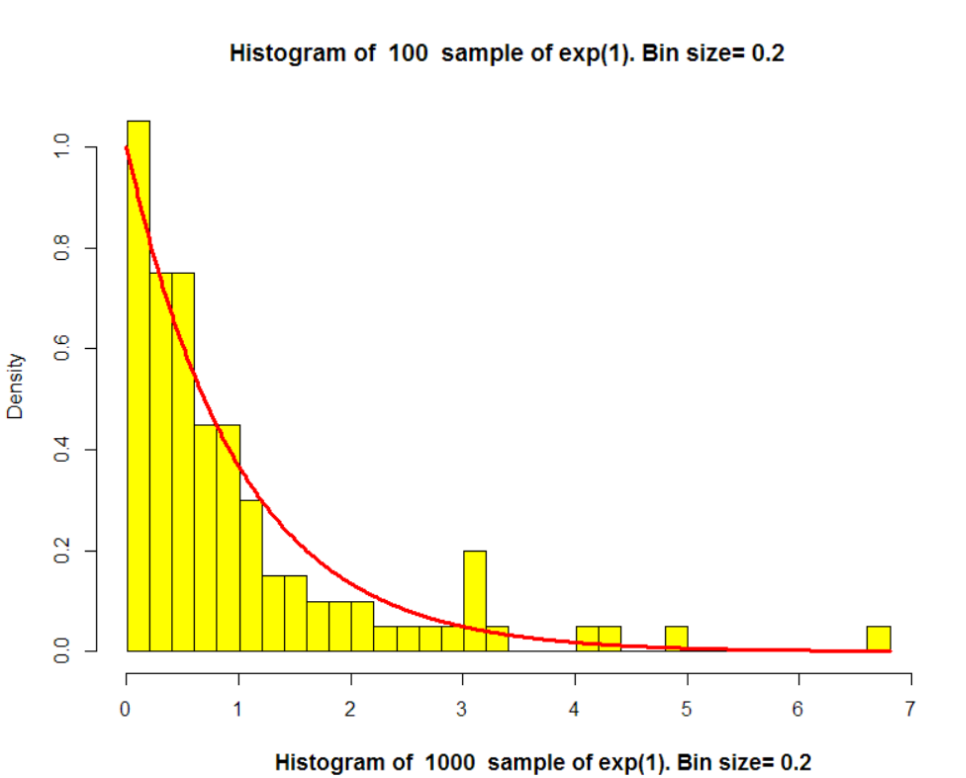
$$P\left(\lim_{n\to\infty} \overline{X_n} = \mu\right) = 1$$

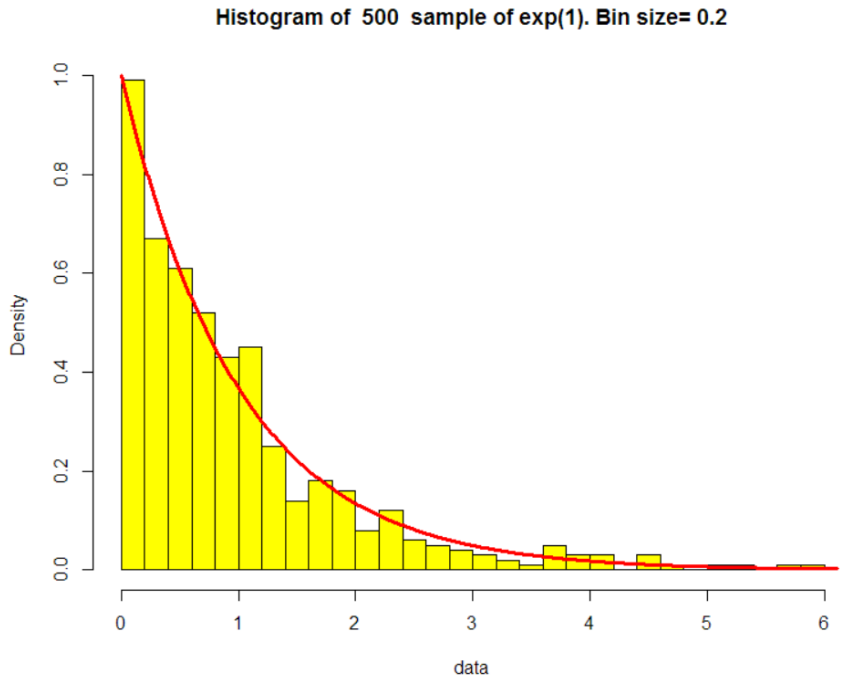
Only few samples give a sample mean that is far away from the mean .平均値μから離れている標本平均の標本は少数だけがある。

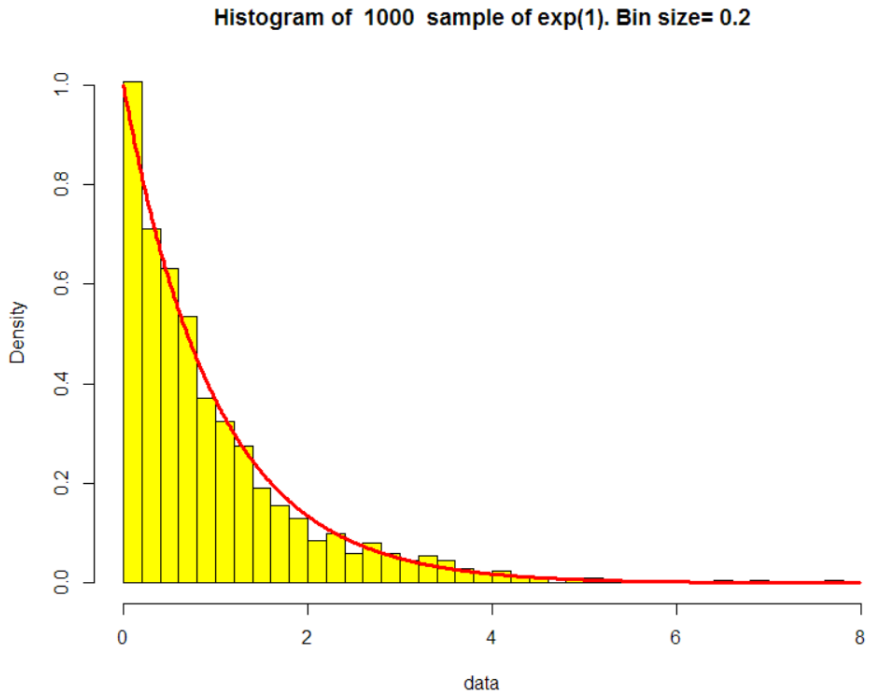
Visualizing the LoLN with histograms

- Empirical distribution (経験分布)
  - Sample from an (infinite) population whose distribution has density function $f(x)$. 密度関数$f(x)$を持つ分布に従う母集団からサンプルする。
  - Measurements are put in bins and we look at the histogram: 得られる測定は柱（びん）に入れてヒストグラムを立てる。
- **Consequence of the LoLN:**
  - As the sample size increases, the histogram gets closer and closer to the curve of the pdf $f(x)$.
    サンプルサイズが増えるにつれて、ヒストグラムは確率密度関数$f(x)$の曲線にますます近寄っていく。
- Example next slide: red curve: $X \sim Exp(1)$


Histogram of 50 sample of exp(1). Bin size= 0.2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7

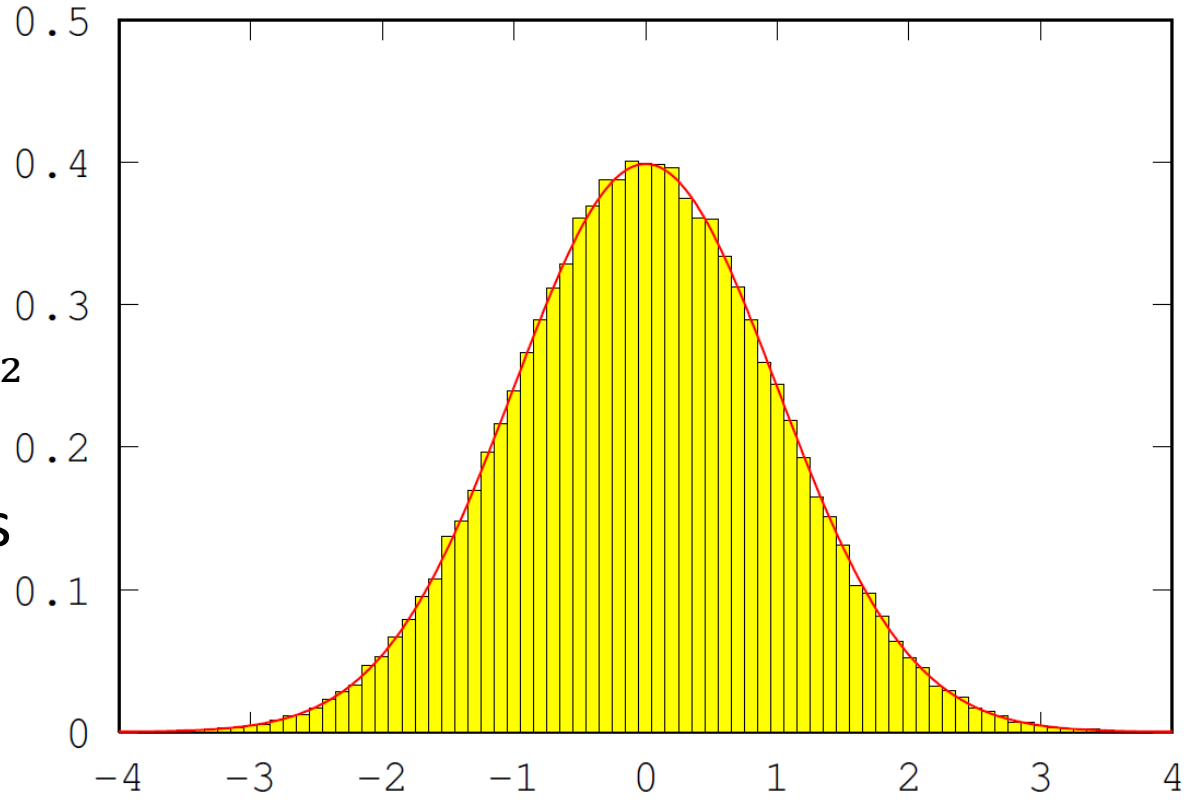

Histogram of 100 sample of exp(1). Bin size= 0.2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
7


Histogram of 500 sample of exp(1). Bin size= 0.2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
data


Histogram of 1000 sample of exp(1). Bin size= 0.2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
data

Visualizing the LoLN with histograms

- Sample from an (infinite) population whose distribution has density function $f(x)$. (example: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$)
- Measurements are put in bins and we look at the histogram:
- Red: Standard Normal curve $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$
- 100000 samples put in bins of width 0.1



Central Limit Theorem (CLT) 中心極限定理

- Setting: $X_1, X_2, \ldots,$ i.i.d.r.v. with expected value μ and stand. dev. σ.
- For each n,

$$\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$$
$$S_n = X_1 + X_2 + \cdots + X_n$$

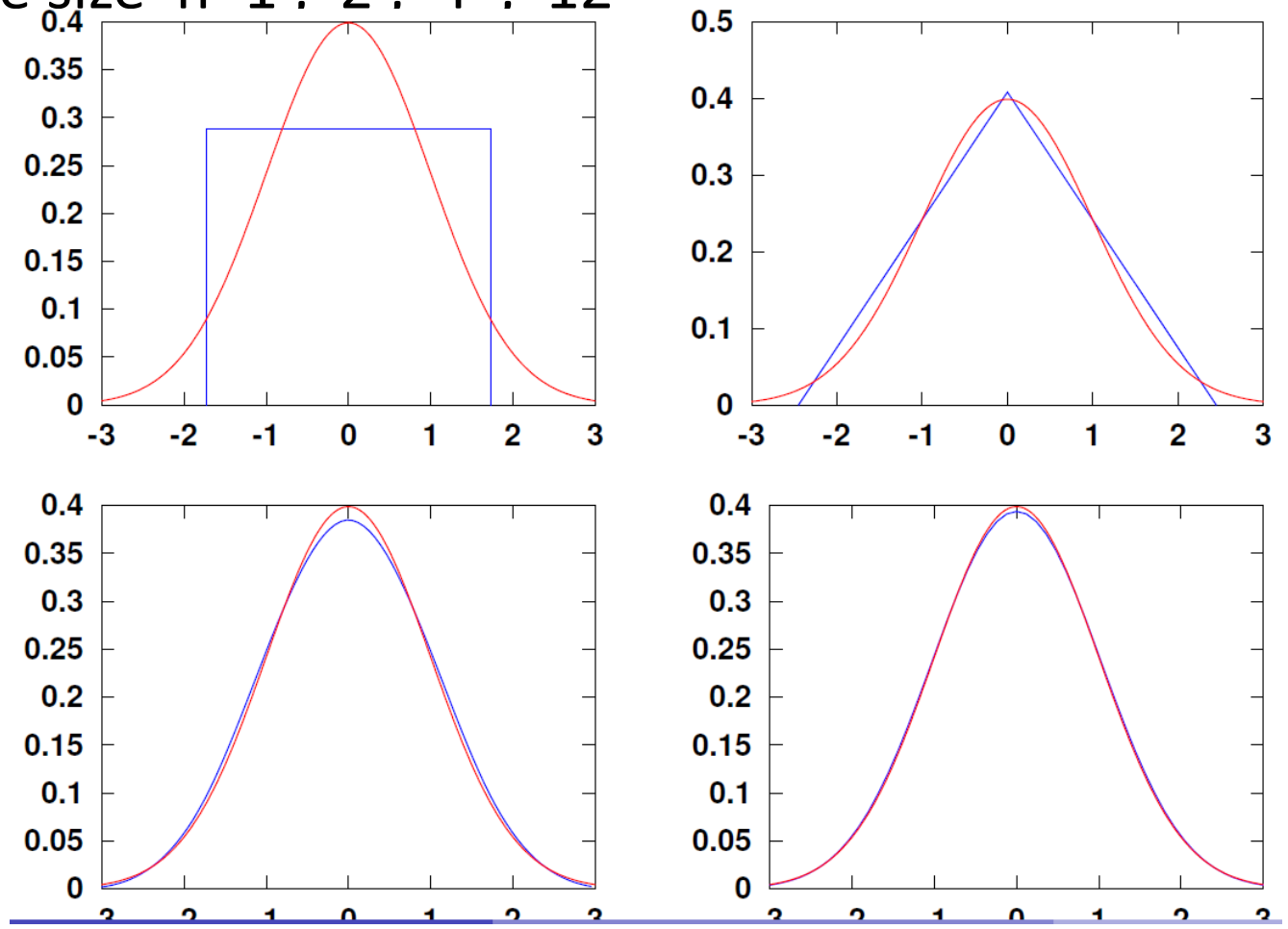
- Conclusion: For "large" n, $(n > 30 \; or \; 50)$

$$\overline{X_n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$
$$S_n \approx N(n\mu, n\sigma^2)$$

- Standardized S_n or $\overline{X_n} \approx N(0,1)$.

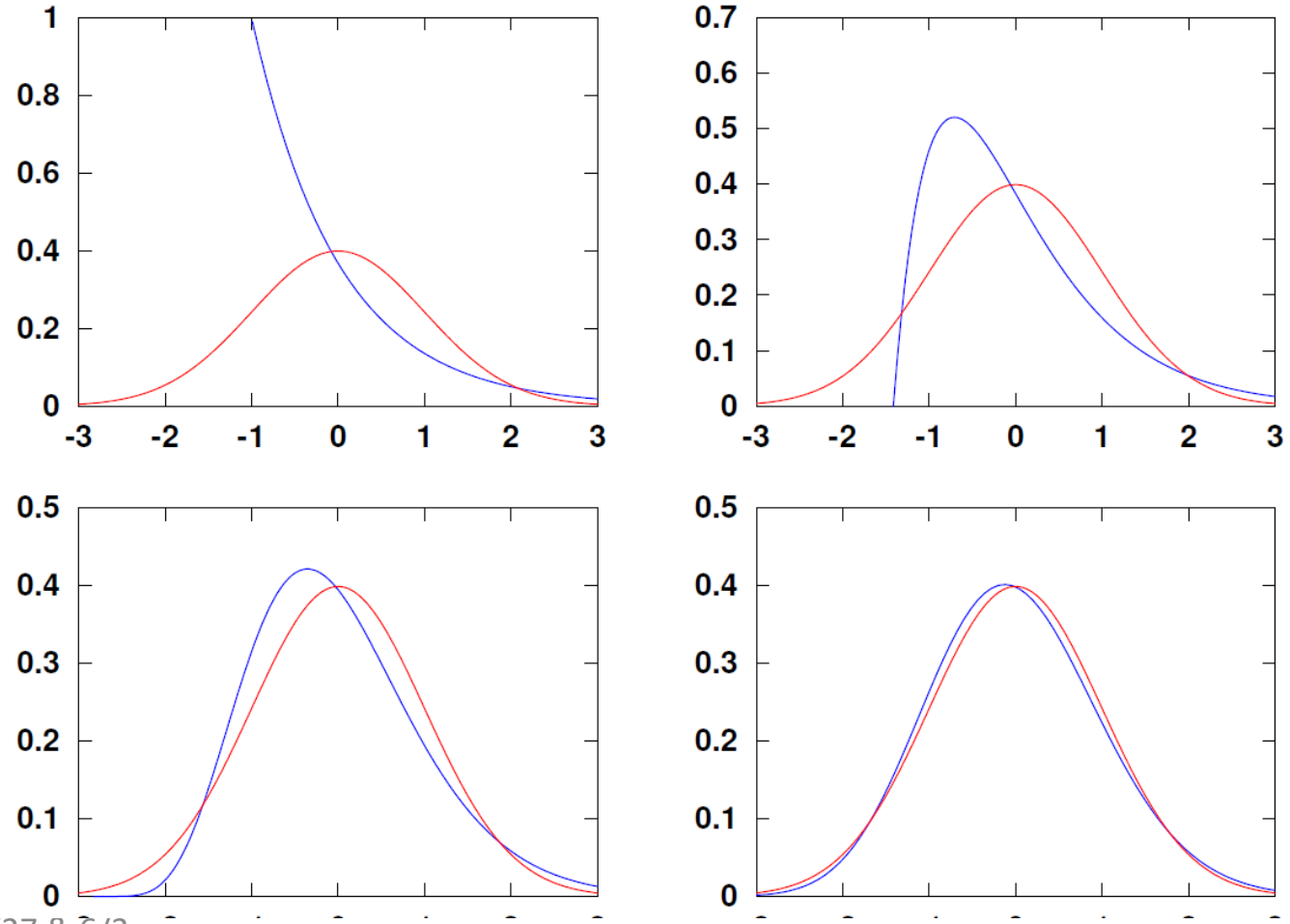
CLT: pictures 1

- Standardized (標準化) average of n i.i.d. uniform random variables. (一様分布に従う独立同分布n個). Sample size n=1 . 2 . 4 . 12



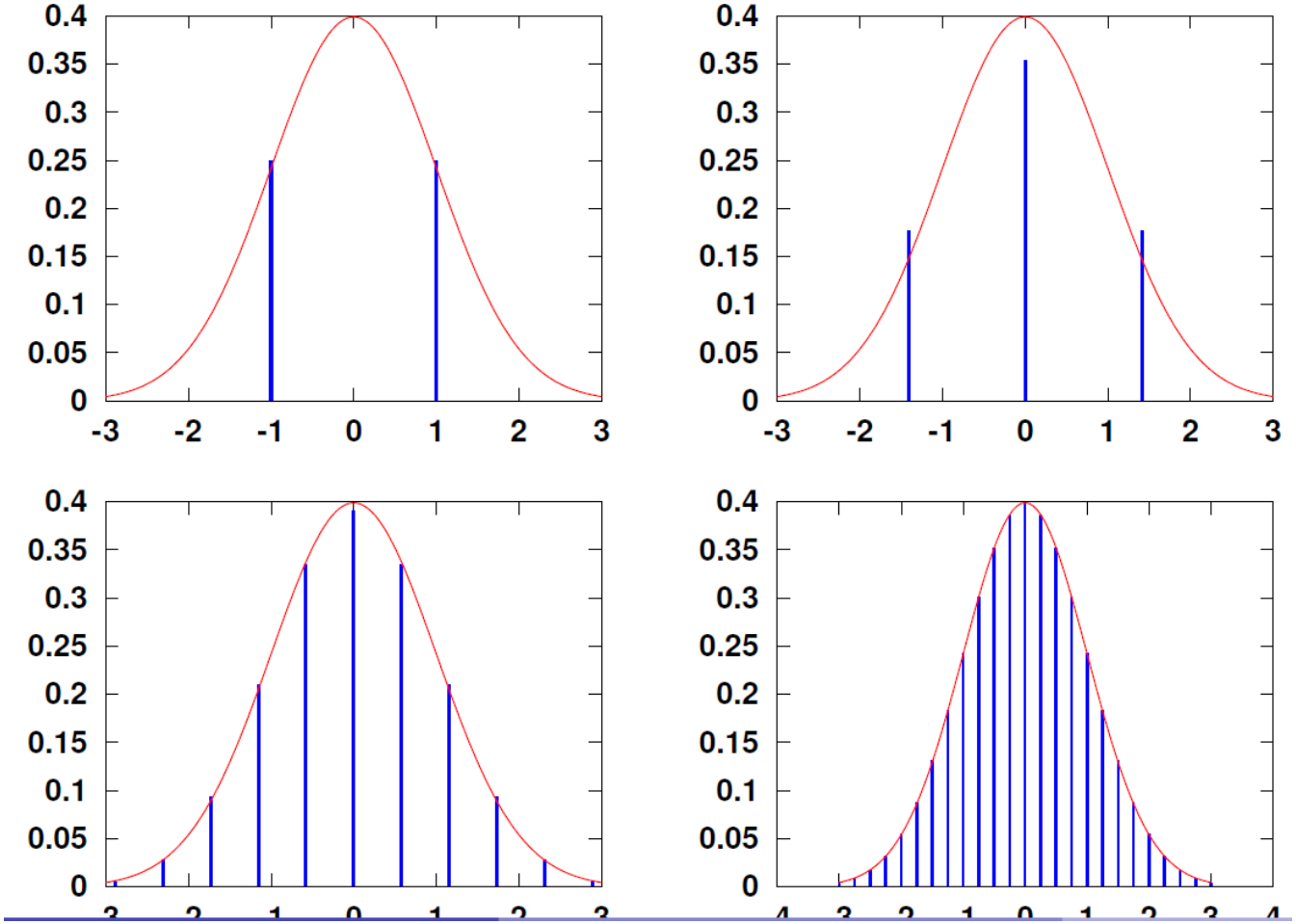
CLT: pictures 2

- Standardized (標準化) average of n i.i.d. exponential random variables. (指数分布に従う独立同分布n個). Sample size n=1, 2 , 8 , 64



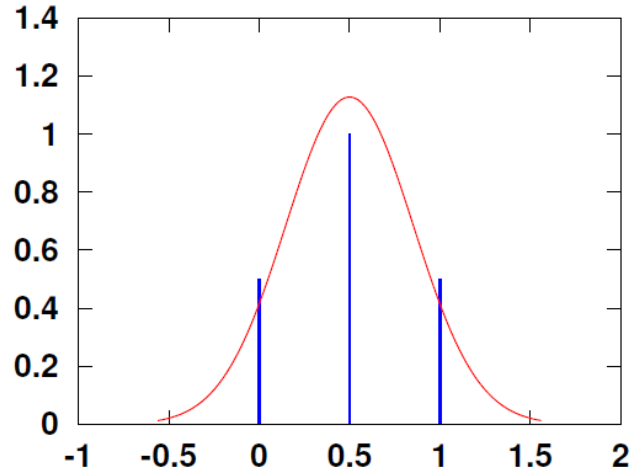
CLT: pictures 3

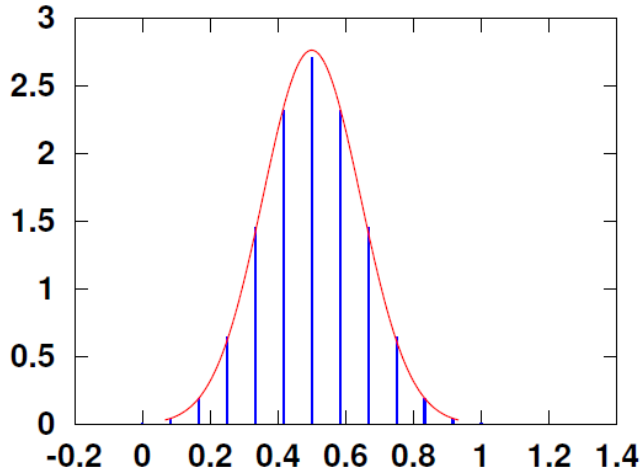
- Standardized (標準化) average of n i.i.d. Bernoulli(1/2) random variables. (ベルヌーイ分布に従う独立同分布n個). Sample size n=1 , 2 , 12 , 64

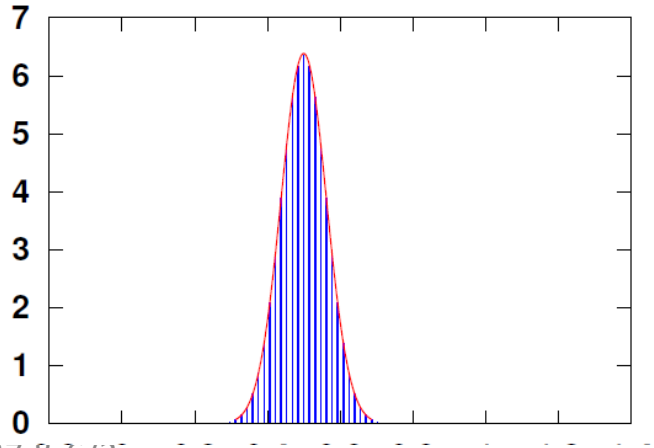


CLT: pictures 4

- (non-standardized 標準化してない) average of n i.i.d. Bernoulli(1/2) random variables. (ベルヌーイ分布に従う独立同分布n個). Sample size n=4, 12, 64







3. Sampling distribution and Central Limit Theorem

3.3 Application of CLT to infer the mean (CLTを用いて母平均を推測する)

- If the sample size is large enough ($n \geq 30$ *or* 50), CLT tells that $\bar{X} \sim N(\mu, \sigma^2/n)$.
  標本の大きさは充分に大きければ　$\bar{X} \sim N(\mu, \sigma^2/n)$

☛ If the standard deviation σ is known, it suffices to use CLT !

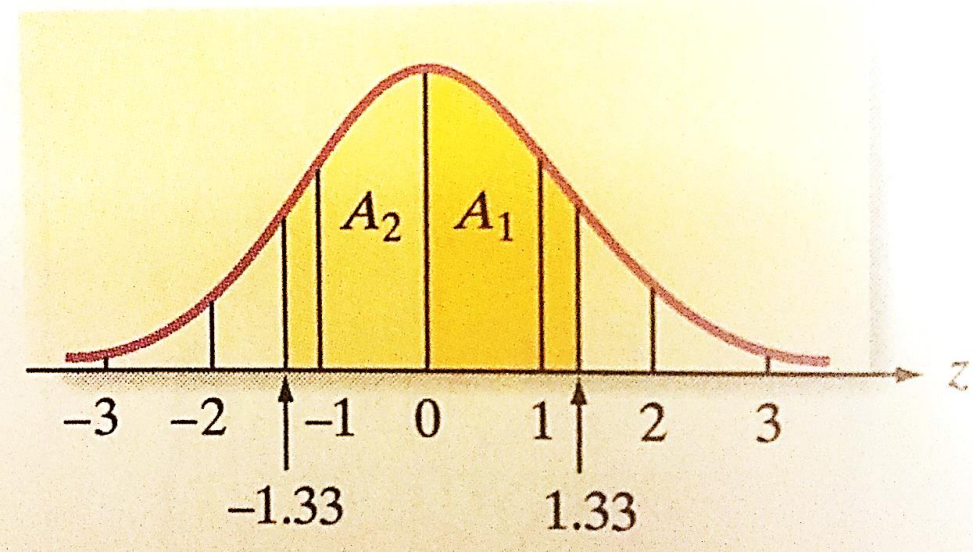
Example: approximating binomial with CLT

Flip a fair coin 100 times. Estimate the probability of more 55 heads. 公正なコインを100回なげ、表が55回以上表れたという事象の確率を評価せよ。

1. Let X_j be the result H=heads or T=tail at the j-th toss
   第j回目のコイン投げの結果をX_jとする。
   What is the probability law of X_j?
   What is $E(X_j)$ and $\sigma_{X_j}^2$?
2. Let $S_{100} = X_1 + \cdots + X_{100}$. Translate the question using the random variable S_{100}.
3. What is $E(S_{100})$ and $\sigma_{S_{100}}^2$?
   What the CLT says about S_{100}?
4. Approximate the probability by a normal probability.
   求めたい確率を正規分布に関わる確率で近似せよ。

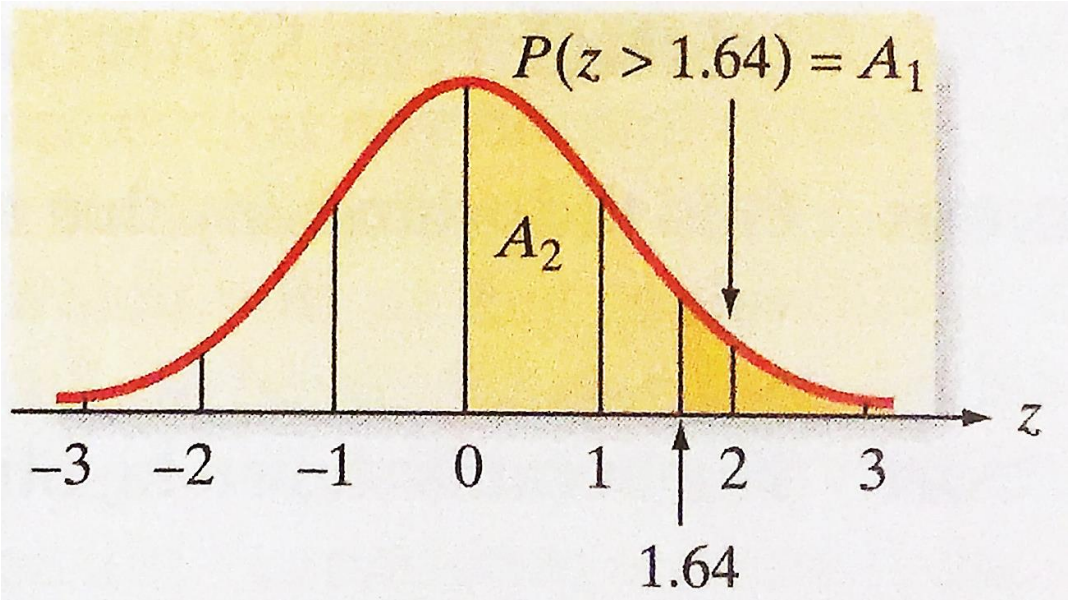
Using the standard normal table I
標準正規分布の表を使うI

- Find the probability that the standard normal variable *z* falls between -1.33 and +1.33.
  標準正規確率変数*z* が
  -1.33と+1.33の間に
  ある確率は何か。



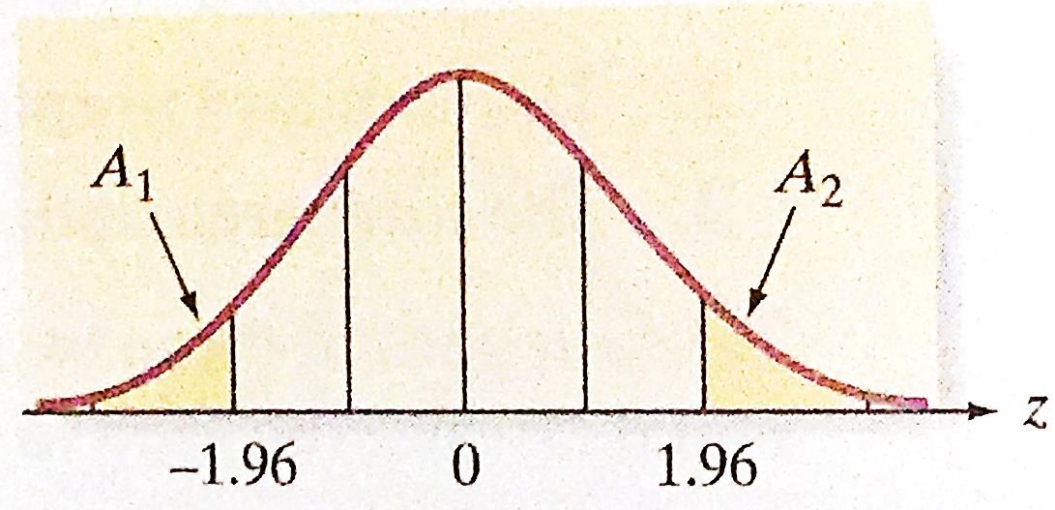
Using the standard normal table II
標準正規分布の表を使う II

- Find the probability that the standard normal variable *z* exceeds 1.64; that is find $P(z > 1.64)$.
  標準正規確率変数*z* が1.64を超える確率は何か。



Using the standard normal table III
標準正規分布の表を使う III

- Find the probability that the standard normal variable *z* exceeds 1.96 in absolute value.
  標準正規確率変数*z* の絶対値が1.96を超える確率は何か。



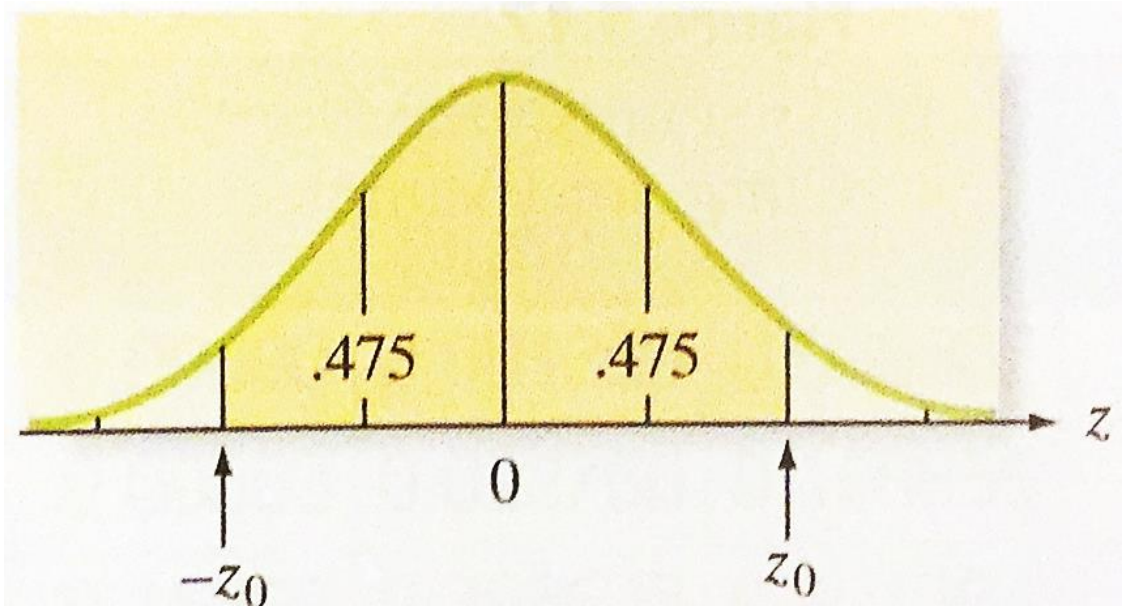
Finding a (non-standard) normal probability
（標準でない）正規確率を求める

- The time X between two charges of a cell phone is normally distributed with a mean of 10hours and a standard deviation of 1.5hours.
  携帯電話の充電間隔をX時間とし、Xは期待値10時間と標準偏差1.5時間の正規分布に従う。
- What is the probability that you can use the cell phone between 8h and 12h without charging it.
  8時間と12時間の間に、充電せずに携帯電話を使える確率は何か。

- Answer: $P(-1.33 \leq z \leq 1.33) = 0.8164$

Using the normal table in reverse
標準正規分布の表を逆利用する

- Find the value of z_0 such that 95% of the standard normal z values lie between $-z_0$ and $+z_0$.



Application of the normal table in reverse
標準正規分布の表を逆利用する

- Entrance examination test: mean score 550 (σ=100).
  入学試験：平均点数　550 点、標準偏差100点.
- Scores follow a normal distribution $N(550,100)$.
  成績は正規分布$N(550,100)$に従う。
- Prestigious university admits only the 90th- percentile of the distribution (=the best 10% examinees).
  有名な大学は分布の9分位数だけを認める。(最も高い点数を得た10%の受験者).

1. What is the minimal score to obtain in order to enter the prestigious university?
   有名大学に入学するために、最低成績は何か？

- *(Answer:*　678)