Lesson 8

Statistics I

Todays topics

- **estimating population mean**
- **estimating population variance**
- **consistent estimator (一致推定量)**
- **unbiased estimator (不偏推定量)**

来嶋 秀治 (Shuji Kijima)
Dept. Informatics,
Graduate School of ISEE

Statistical Inference

母分布(の特徴量を)を推定する

Natural Science

Strawberries in X farm is said to be sweeter than other farms.

Check degrees Brix (糖度) some samples using a machine.

	1	2	3	4	5	6	7	8	9
A	7.2								
B									

Operations Research

We want to know the coefficient of restitution (反発係数) of a ball made in a factory Y, where it is required to be between 0.38 and 0.42.

We have checked 1000 random samples.

What is the expectation and variance of coeff. rest.

	1	2	3	4	5	6	7	8	9
	0.406	0.402	0.403	0.397	0.401	0.389	0.396	0.402	0.411

不良品の出現分布の推定

母集団：工場で作られるボール（10^6個/年）

Engineering

We have devised a new matter Z which is energy efficient.

To know the efficiency, we made trial productions.

Voting

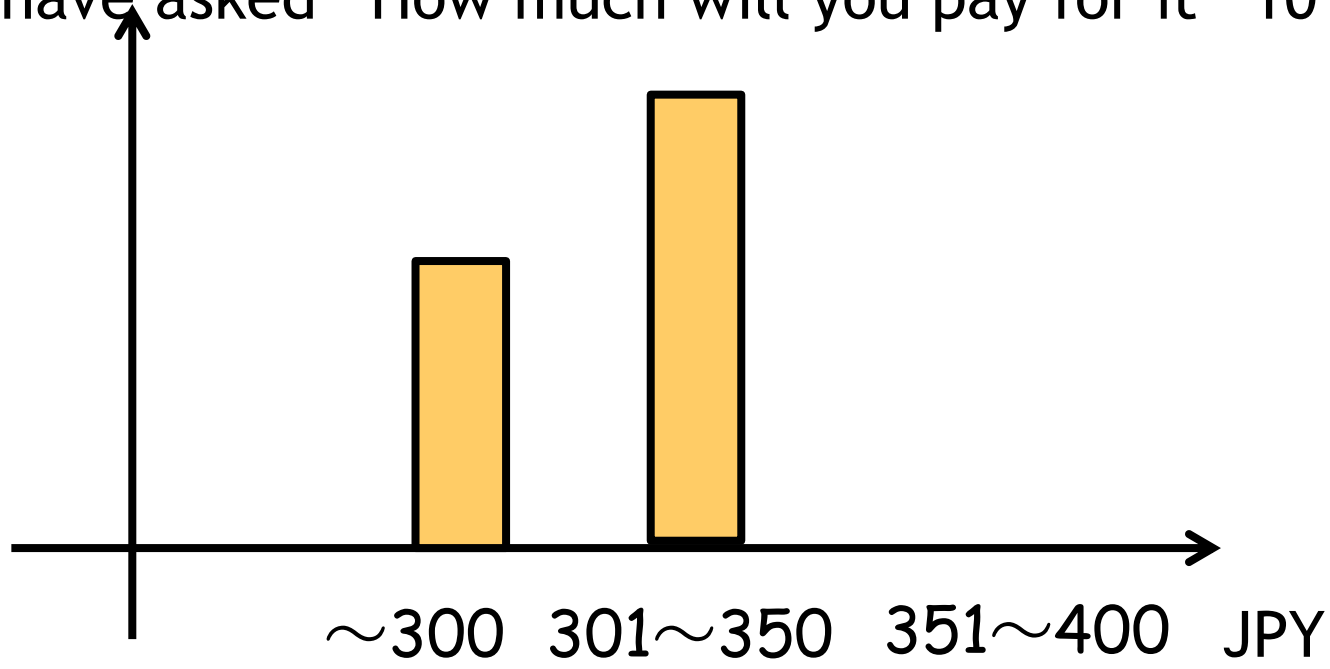
Candidate T gets 47% votes of 1000 samples.

Sample test

We have developed new potato chips.

We want to decide its price.

We have asked “How much will you pay for it” 10 testers.



Statistics / Data science

Statistics Inference (統計的推論)

- ✓ Estimation (推定)
- ✓ Statistical test (統計検定)
- ✓ Regression (回帰)
- ✓ Applications
  - ➢ Machine learning (機械学習),
  - ➢ Pattern recognition (パターン認識),
  - ➢ Data mining (データマイニング), etc.

Statistical inference

統計的推論

sample vs population vs stochastic model

- ✓ population (母集団)
- ✓ sample (標本)
- ✓ stochastic model (確率モデル)

Example 1

We sample 6 accounts of twixxer at random. The following table shows the numbers of followers.

	1	2	3	4	5	6
#followers	372	623	89	781	3219	152

Suppose that the number of followers follows some distribution (e.g., exponential distribution, Poisson distribution, Zipf's distribution etc.).

Terminology (Statistics)

Statistics (統計学)

- ✓ population (母集団)
  - ➢ population distribution (母集団分布)
- ✓ random sample (無作為標本)
  - ➢ sample value (標本値)
  - ➢ sample distribution (標本分布)
- ✓ statistics (統計量)
  - ➢ sample mean (標本平均)
  - ➢ sample variance (標本分散)
  - ➢ etc.

statistical inference (統計的推論)

Estimating the population mean

母平均の推定

Population, sample, stochastic mo

- ✓ population (母集団)
- ✓ sample (標本)
- ✓ stochastic model (確率モデル)

Example 1

We sample 6 accounts of twixxer at random. The following table shows the numbers of followers.

	1	2	3	4	5	6
#followers	372	623	89	781	3219	152

Q. How large is the population mean of followers?

Population, sample, stochastic mo

- ✓ population (母集団)
- ✓ sample (標本)
- ✓ stochastic model (確率モデル)

Example 1

We sample 6 accounts of twixxer at random. The following table shows the numbers of followers.

	1	2	3	4	5	6
#followers	372	623	89	781	3219	152

Q. How large is the population mean of followers?

Suppose that the number of followers follows some distribution (e.g., $\mathrm{Ex}(\lambda)$) with expectation μ.

Population, sample, stochastic mo

- ✓ population (母集団)
- ✓ sample (標本)
- ✓ stochastic model (確率モデル)

Example 1

We sample 6 accounts of twixxer at random. The following table shows the numbers of followers.

	1	2	3	4	5	6
#followers	372	623	89	781	3219	152

Q. How large is the population mean of followers?

Suppose that the number of followers follows some distribution (e.g., $\mathrm{Ex}(\lambda)$) with expectation μ.

=> Sample mean $\bar{X} = \frac{X_1 + \cdots + X_n}{n} = 872.7$

Statistics, Data science

Statistics Inference (統計的推論)

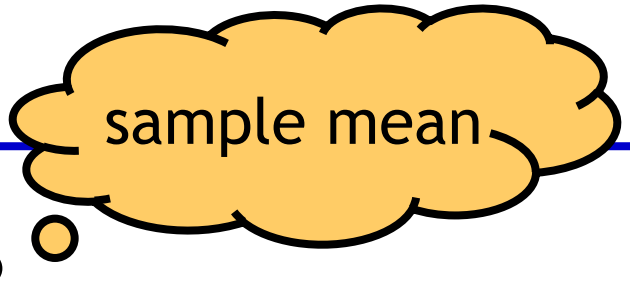
- ✓ estimation (推定)
  - ➢ How good is the estimator?
    - ☐ consistent estimator (一致推定量)
    - ☐ unbiased estimator (不偏推定量)
    - ☐ Minimum mean square error (最小二乗誤差推定)
  - ➢ Techniques of estimation
    - ☐ Maximum likelihood (最尤推定)
    - ☐ Bayesian inference (ベイズ推定)
- ✓ Statistical test (統計検定)
- ✓ Regression
- ✓ Advanced
  - ➢ Machine learning, Pattern recognition, Data mining, etc.

Consistent estimator

Def.

T is a consistent estimator of θ if $\lim_{n\to\infty} \Pr[T = \theta] = 1$.

Sample mean



Proposition

$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is a consistent estimator of μ.

Proof.

By the law of large numbers.

Unbiased estimator

Definition

Let X_i be i.i.d. F_θ.

Let $T(X)$ denote an estimator of a parameter $g(\theta)$ of F_θ, then we call $E_\theta[T(X)] - g(\theta)$ *bias*.

Definition

$T(X)$ is an *unbiased estimator* of $g(\theta)$ if $E_\theta[T(X)] - g(\theta) = 0$ holds.

Sample mean

sample mean

Proposition

$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is an unbiased estimator of μ.

Proof.

$$\begin{aligned} \mathrm{E}[\bar{X}] &= \mathrm{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] \\ &= \frac{1}{n} \cdot n \cdot \mu \\ &= \mu \end{aligned}$$

Estimating the population variance

母分散の推定

Population, sample, stochastic mo

- ✓ population (母集団)
- ✓ sample (標本)
- ✓ stochastic model (確率モデル)

Example 1

We sample 6 accounts of twixxer at random. The following table shows the numbers of followers.

	1	2	3	4	5	6
#followers	372	623	89	781	3219	152

Q. How large is the population variance of #followers?

Suppose that the number of followers follows some distribution (e.g., $\mathrm{Ex}(\lambda)$) with expectation μ and variance σ^2

Recall $\mathrm{Var}[X] := \mathrm{E}[(X-\mu)^2]$

Variance

$\sigma^2 = \mathrm{E}[(X-\mu)^2]$

Proposition

$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ is NOT an unbiased estimator of σ^2 (in general)

$$\mathrm{E}\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)-\left(\overline{X}-\mu\right)\right)^2\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left((X_i-\mu)^2-2(X_i-\mu)\left(\overline{X}-\mu\right)+\left(\overline{X}-\mu\right)^2\right)\right]$$

$$= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]-2\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\left(\overline{X}-\mu\right)\right]+\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2\right]$$

Variance

$\sigma^2 = \mathrm{E}[(X-\mu)^2]$

Proposition

$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ is NOT an unbiased estimator of σ^2 (in general)

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[(X_i-\mu)^2] = \frac{1}{n}n\mathrm{E}[(X_i-\mu)^2] = \sigma^2$$

$$-2\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)\right] = -2\mathrm{E}\left[(\bar{X}-\mu)\frac{\sum_{i=1}^{n}(X_i-\mu)}{n}\right] = -2\mathrm{E}\left[(\bar{X}-\mu)\left(\frac{\sum_{i=1}^{n}X_i}{n}-n\frac{\mu}{n}\right)\right] = -2\mathrm{E}\left[(\bar{X}-\mu)^2\right]$$

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{X}-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(\bar{X}-\mu)^2\right] = \frac{1}{n}n\mathrm{E}\left[(\bar{X}-\mu)^2\right] = \mathrm{E}\left[(\bar{X}-\mu)^2\right]$$

$$\mathrm{E}\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}\right] = \sigma^2 - \mathrm{E}\left[(\bar{X}-\mu)^2\right] < \sigma^2$$

unless $\mathrm{E}[Z^2] = 0$.

($\mathrm{E}[Z^2] = 0$ only when $\Pr[Z=0] = 1$)

Unbiased sample variance

Proposition

$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ is an unbiased estimator of σ^2 (in general)

Good estimator

Def.

T: estimator of a population parameter θ.

$E[(T-\theta)^2]$ is called mean square error (平均二乗誤差)

Proposition

If T is an unbiased estimator, then $E[(T-\theta)^2] = \mathrm{Var}[T]$.

$$\begin{aligned}
\mathrm{E}[(T-\theta)^2] &= \mathrm{E}\left[\left((T-\mathrm{E}[T]) + (\mathrm{E}[T-\theta])\right)^2\right] \\
&= \mathrm{E}[(T-\mathrm{E}[T])^2 + 2(T-\mathrm{E}[T])(\mathrm{E}[T-\theta]) + (\mathrm{E}[T-\theta])^2] \\
&= \mathrm{E}[(T-\mathrm{E}[T])^2] + 2\mathrm{E}[(T-\mathrm{E}[T])(\mathrm{E}[T-\theta])] + \mathrm{E}[(\mathrm{E}[T-\theta])^2] \\
&= \mathrm{Var}[T] + (\mathrm{E}[T-\theta])^2
\end{aligned}$$

$$\begin{aligned}
2\mathrm{E}[(T-\mathrm{E}[T])(\mathrm{E}[T-\theta])] &= 2(\mathrm{E}[T-\theta])\mathrm{E}[(T-\mathrm{E}[T])] \\
&= 2(\mathrm{E}[T-\theta])(\mathrm{E}[T]-\mathrm{E}[T]) \\
&= 0
\end{aligned}$$

Remind: Probability III

i.i.d. (multivariate distribution)

Distribution of random variables X and Y of $(\Omega, \mathcal{F}, P)$.

Ex1. two dice.

$\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$

X = sum of casts

Y = product of casts

例2. poker

choose five cards,

X = # of A's

Y = # of spades

Prop.

$$\Pr[(X_1 = x_1)\&(X_2 = x_2)\&\cdots\&(X_n = x_n)] = \Pr[X_1 = x_1]\Pr[X_2 = x_2]\cdots\Pr[X_n = x_n]$$

Prop.

Suppose ***discrete*** r.v. $X_1, X_2, \ldots X_n$ are i.i.d. w/pmf f.

Then $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$.

Prop.

Suppose ***continuous*** r.v. $X_1, X_2, \ldots X_n$ are i.i.d. w/pdf f.

Then $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$.

Prop.

$\Pr[(X \le x)\&(Y \le y)] = \Pr[X \le x]\Pr[Y \le y]$
i.e., $F_{XY}(x,y) = F_X(x)F_Y(y)$

Prop.

Suppose ***continuous*** r.v. X and Y are **independent**.

Then, $f_{XY}(x,y) = f_X(x)f_Y(y)$.