Essential Mathematics for Global Leaders I

Spring 2019

Statistics

Lecture 6: 2019 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

LECTURE 5

- 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

3. Sampling distribution and Central Limit Theorem 標本分布と中心極限定理 3.4 More on sample statistics

You believe that the lifetimes of a certain type of lightbulb follow an exponential distribution with parameter λ. ある電球の寿命はλ-指数分布に従うと思われる。

To test this hypothesis you measure the lifetime of 5 bulbs and get data $x_1, \ldots, x_5$. この仮定を検定するために電球の五つの寿命を測定しデータを得る$x_1, \ldots, x_5$：

- Which of the following are statistics? 統計量は何か？
  a. The sample average $\bar{x} = \frac{x_1+x_2+x_3+x_4+x_5}{5}$
  b. The expected value of a sample, namely $1/\lambda$.
  c. The difference between $\bar{x}$ and $1/\lambda$.

1. (a) 2. (b) 3. (c) 4. (a) and (b)
5. (a) and (c) 6. (b) and (c) 7. all three 8. none of them

Examples of (point) statistics

$X_1, X_2, \ldots, X_n$: i.i.d.r.v. (sample of size n)

- Sample mean: $\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$
- (biased) Sample variance: $\bar{\sigma}_{\bar{X}}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$
- Standard error of the mean (SEM): $\text{se}_{\bar{X}} = \frac{\bar{\sigma}_{\bar{X}}}{\sqrt{n}}$
- $Median(X_1, \ldots, X_n)$ for the sample $x_1, \ldots, x_n$ is:
  - X_t where $|\{i : X_i \le X_t\}| = \frac{n+1}{2}$ if n is odd (奇数)
  - $\frac{X_{t-1}+X_t}{2}$ where $|\{i : X_i \le X_t\}| = \frac{n}{2}$ if n is even (偶数)
    - Example: $X_1 = 6,\ X_2 = 3, X_3 = 3,\ X_4 = 9,\ X_5 = 7$
      (odd) Median=
      (even) $X_1 = 6,\ X_2 = 3,\ X_3 = 3,\ X_4 = 9$ Median=
- Quantile (分位数), $Max(X_1, \ldots, X_n)$ etc…

Sample space of sampling distribution 標本分布の標本空間

- A sample (statistic) of a sample $X_1, \dots, X_n$ is a function $Y = g(X_1, \dots, X_n)$
  - Example: $Y = \overline{X_n}$ is the sample mean
    $$g(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n)$$
- Sample space of the sampling distribution of Y: set of all possible samples of size n : Ω^n
- Need multidimensional probability, called here "joint distribution". 同時分布 or 結合分布

$$E(Y) = \sum_{x_1,\dots,x_n} p(X_1 = x_1, \dots, X_n = x_n)\, g(x_1, \dots, x_n)$$

(sum over all samples $x_1, \dots, x_n$ of size n)

- Because $X_1, \dots, X_n$ are independent:
  $$p(X_1 = x_1, \dots, X_n = x_n) = p(X_1 = x_1) \cdots p(X_n = x_n)$$

Unbiased estimator 不偏推定量

- Estimator (=statistic) $\widehat{\theta_n}$ of the parameter of the population θ is unbiased if $E\left(\widehat{\theta_n}\right) = \theta$
(expected value over the sampling distribution☞ all possible samples of size n).
統計量$\widehat{\theta_n}$の（すべてのサイズn標本上の）期待値と対象の母数パラメターθは、等しければ、$\widehat{\theta_n}$は不偏推定量という。

- The sample mean $\overline{X_n}$of n i.i.d. measurements $X_1, \ldots, X_n$ is an unbiased estimator of the mean of the population
標本平均の期待値は母平均の不偏推定量である。
$$E(\overline{X_n}) = \mu$$
(μ　母平均)

Proof that sample mean is unbiased (sample of size 2, discrete case)

$$E(\overline{X_2}) = \sum_{x_1, x_2 \in \Omega} p(X_1 = x_1, X_2 = x_2)\ \overline{X_2}(x_1, x_2)$$

$$= \sum_{x_1, x_2} p(x_1)p(x_2)\frac{(x_1 + x_2)}{2}$$

$$= \frac{1}{2}\sum_{x_1} x_1 p(x_1)\left(\sum_{x_2} p(x_2)\right) + \frac{1}{2}\sum_{x_2} x_2 p(x_2)\left(\sum_{x_1} p(x_1)\right)$$

Total law of probability : $\sum_{x_1} p(x_1) = 1 = \sum_{x_2} p(x_2)$

$$E(\bar{X}) = \frac{1}{2}\sum_{x_1} x_1 p(x_1) + \frac{1}{2}\sum_{x_2} x_2 p(x_2)$$

$$= \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

Continuous case 連続確率変数の場合

- In case each i.i.d.r.v are continuous, then we replace $P(X_1 = x_1, \ldots, X_n = x_n)$ by the joint pdf of $X_1, \ldots, X_n$ and the summation symbol $\sum$ by an integration one $\int$.
- 独立同分布に従う確率変数は連続であるとき、各確率$P(X_1 = x_1, \ldots, X_n = x_n)$、総和記号$\sum$、積分記号$\int$の変わりに、 $X_1, \ldots, X_n$ の結合分布を書くと正しい。

$$E(Y) = \int_\Omega \cdots \int_\Omega f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) g(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

- $X_1, \ldots, X_n$ are independent, so the joint distribution is:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_X(x_1) \cdots f_X(x_n)$$

$$E(Y) = \int_\Omega \cdots \int_\Omega f_X(x_1) \cdots f_X(x_n) g(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

Full Proof that the sample mean is unbiased (**Continuous case**, sample size n=2)

$$E(\bar{X}_2) = \int_{\Omega^2} f_{(X_1,X_2)}(x_1,x_2) g(x_1,x_2) dx_1 dx_2$$

$$= \frac{1}{2}\int_{\Omega^2} f_{(X_1,X_2)}(x_1,x_2)(x_1+x_2) dx_1 dx_2$$

$$= \frac{1}{2}\int_{\Omega^2} f_{X_1}(x_1) f_{X_2}(x_2)(x_1+x_2) dx_1 dx_2$$

$$= \frac{1}{2}\int_{\Omega} f_{X_1}(x_1) x_1 \left(\int_{\Omega} f_{X_2}(x_2) dx_2\right) dx_1$$

$$+ \frac{1}{2}\int_{\Omega} f_{X_2}(x_2) x_2 \left(\int_{\Omega} f_{X_1}(x_1) dx_1\right) dx_2$$

$$= \frac{1}{2}\left(\int_{\Omega} f_{X_1}(x_1) x_1 dx_1\right) + \frac{1}{2}\left(\int_{\Omega} f_{X_2}(x_2) x_2 dx_2\right)$$

$$= \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$$

Unbiased sample variance 不偏標本分散

- The (biased) sample variance (Lecture 5, Slide 10)

$$\bar{\sigma}_{\bar{X}}^2 = \frac{1}{n}\left((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right)$$

is *biased*: $\mathrm{E}\left(\bar{\sigma}_{\bar{X}}^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ (σ population's variance)

☛ **Definition (定義）** Unbiased sample variance:

$$\mathrm{s}_{\bar{X}}^2 = \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \cdots (X_n - \bar{X})^2\right)$$

It is easy to check that: $E\left(\mathrm{s}_{\bar{X}}^2\right) = \sigma^2$ (*unbiased 不偏*)

Example: sample median. Biased or not ?

- $\Omega = \{\,0\,,3\,,12\,\}$.Take 3 samples X_1, X_2, X_3 (with repetition) from Ω.
  復元抽出
- How many possible samples are there ?
- Up to permutations (置換を考慮しないと) possible samples are:

000 | 003 |033 | 333 | 0 0 12|0 3 12|0 12 12| 3 3 12| 3 12 12| 12 12 12

- Take the permutation into account to fill in the probability table of the sample mean: $\bar{x} = 1/3\,(x_1 + x_2 + x_3)$

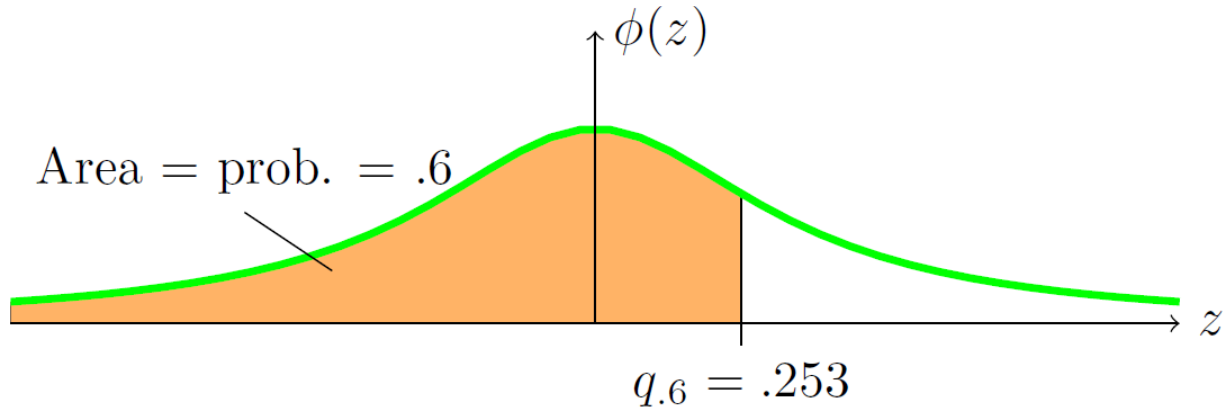
$\bar{x}$	0	1	2	3	4	5	6	7	8	9	10	11	12
$p(\bar{x})$													

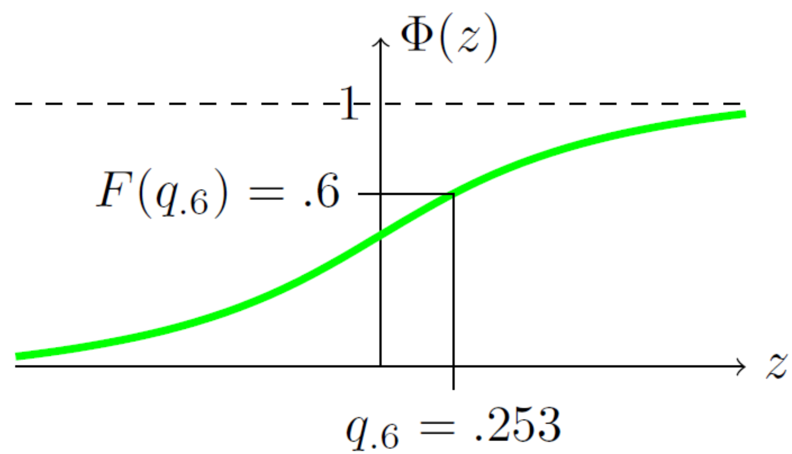
- Compute $E(\bar{X}) = \sum_{\bar{x}} p(\bar{x})\,\bar{x}$ the expected value of the sample mean (overall size 3 samples in Ω).
- Fill in the probability table of the sample median M. Deduce $E(M)$

m	0	1	2	3	4	5	6	7	8	9	10	11	12
$p(M = m)$													

Quantile 分位数 (q-values q-値)

- Quantiles give a measure of location.
  分位数は中心傾向を測るものである。
- Cdf is increasing function so it is invertible. 可逆関数。
- F cdf , f pdf ($F'(x) = f(x)$) if X is continuous and f pmf if X is discrete.
- $q_{0.6}$ is such that: $F(q_{0.6}) = 0.6$



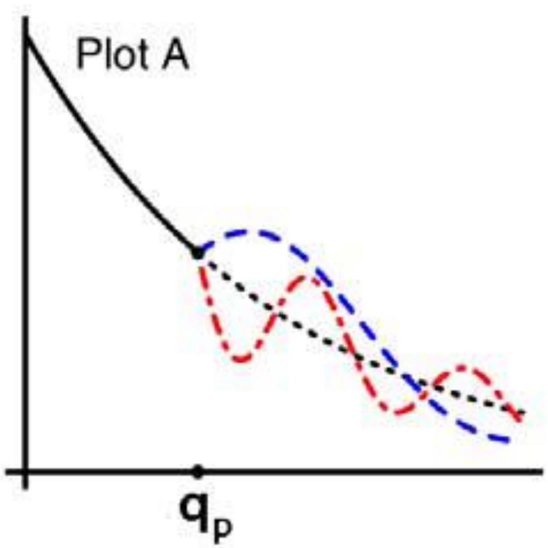


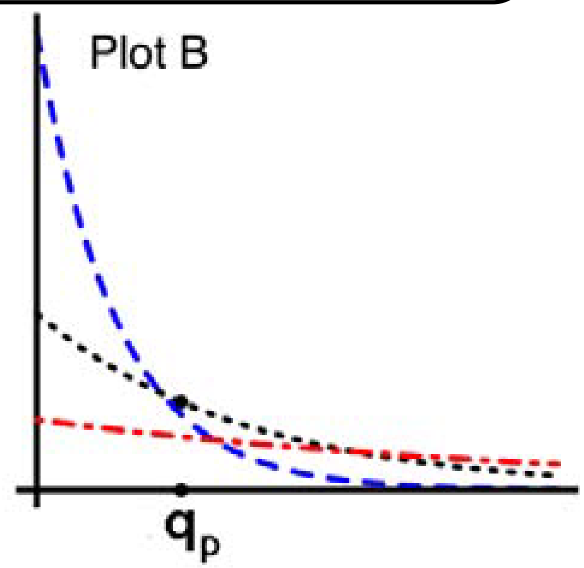
For $0 \leq t \leq 1$, q_t is such that $q_t = F^{-1}(t)$.
$F(q_t) = P(X \leq q_t) = t$

Median of a continuous random variable 連続確率変数の中位部 (中央値)

- $P(X \leq q_{0.5}) = 0.5 = P(X > q_{0.5})$
- Question: Three pdf are plotted (black, red, blue).
- The median of the black density is at q_p.
- Which density has the greatest median?

1 All the same	2 Red	3 Blue
4 Black	5 Impossible to tell	

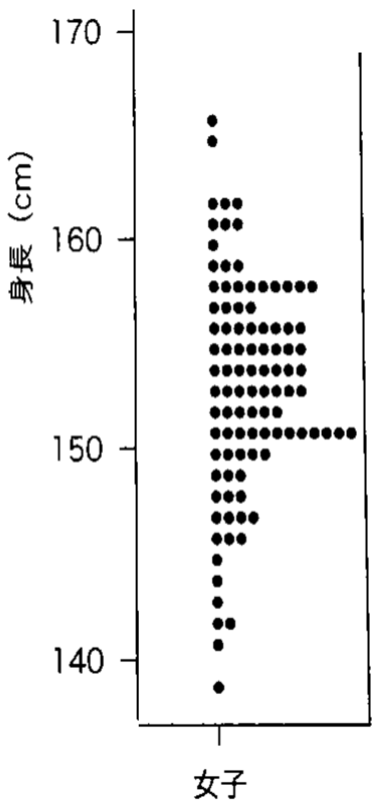




Quartile, Inter-quartile, Box & Whisker plot
四分位点、四分意範囲、箱ひげ図

- $Q_1 = q_{0.25}$, 1st quartile $Q_3 = q_{0.75}$, 3rd quartile
- $IQR = \frac{Q_3 - Q_1}{2}$ interquartile

ドットプロット　dot-plot
(=縦ヒストグラム =
vertical histogram)



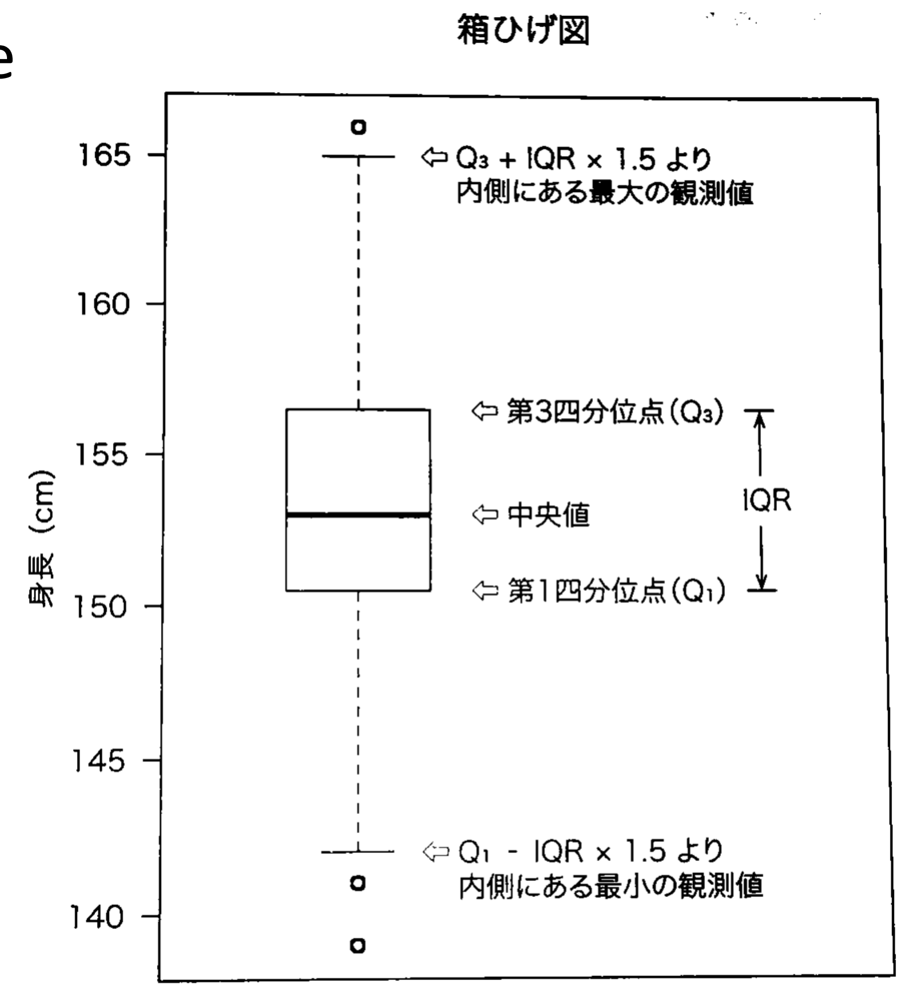


図 3. 女子生徒 100 人の身長