Essential Mathematics for Global Leaders I

Spring 2019

Statistics

Lecture 10: 2019 July 8-July. 15

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 II: Statistical inference (推計統計学)

5 Confidence Intervals (CI) 信頼区間

Chapter 5: Confidence Intervals 信頼区間

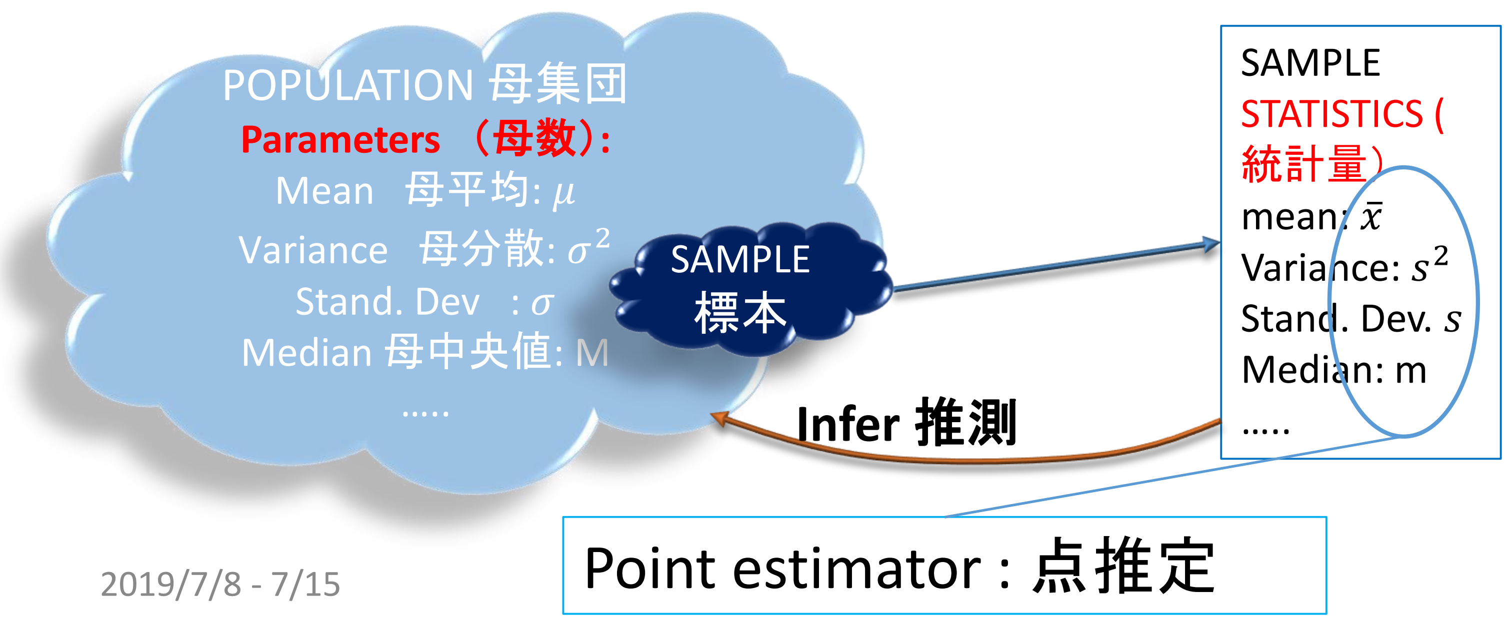
5.1 Introduction and CI normal data

信頼区間への入門と正規母集団の信頼区間

From a sample $x_1, \ldots, x_n$ drawn from a population of mean μ and variance σ^2,

母平均μと母分散σ^2から抽出された標本$x_1, \ldots, x_n$

we can find estimator (Lect. 6, p. 6) for the mean μ (like the sample mean $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ or for the variance σ^2 (like the sample variance $s^2 = \frac{1}{n-1}\left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right)$.



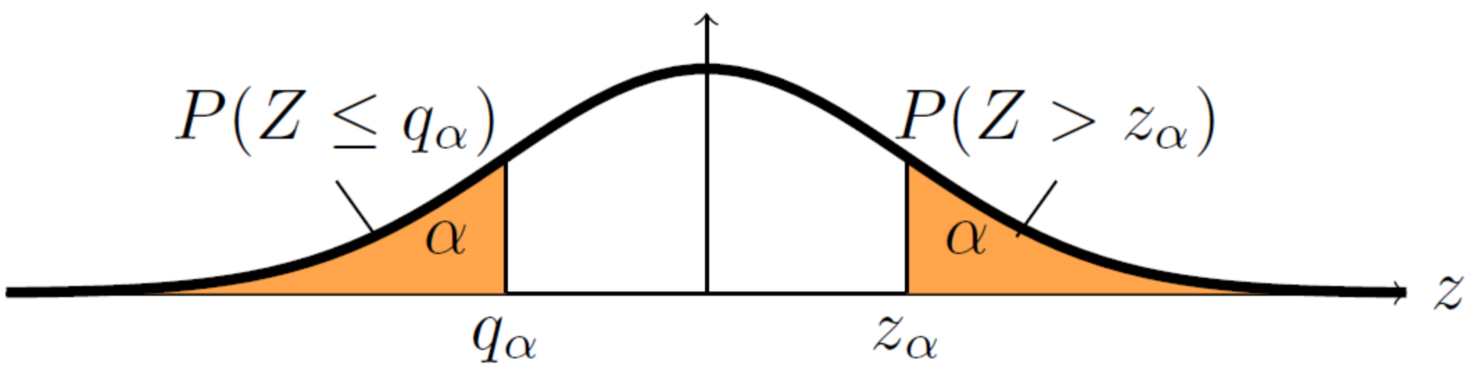
Introduction to confidence intervals
信頼区間への入門

- Point estimator alone does not provide a good enough inference
  点推定だけで、十分に有意義な推定を与えない。
- Better is a "Confidence Interval":
  - Point estimator and an "interval statistics"
    点推定　+「統計区間」
  - Example of point estimator (sample mean) $\bar{x} = 2.2$
    Confidence interval: $\bar{x} \pm 1 = [1.2\,,\ 3.2]$
  - Computed from data and known parameters (never use an unknown parameter)
    データと既知のパラメータによって計算されるもの
    (未知のパラメータを使ってはいけない)

Review of critical values and quantiles 分位数と臨界値（棄却値）の復習

- Quantile 分位数 (Lecture 6, Chapter 3, page 12):
  - Left tail $P(X < q_\alpha) = \alpha$
- Critical values 臨界値 (Lecture 7, Chapter 4.1, page 27)
  - Right tail $P(X > c_\alpha) = \alpha$

- Letters (文字) used for critical values
  - z_α for normal standard distribution $N(0,1)$
  - $t_\alpha(n)$ or t_α for t-distribution T_n (d-o-f n Student)
  - c_α, x_α all purpose (汎用)



Computing confidence intervals from normal data 正規母集団から信頼区間を求める (I)

- Suppose the data $x_1, \ldots, x_n$ is drawn from $N(\mu, \sigma^2)$
- Confidence level = $\mathbf{1-\alpha}$ 信頼水準
  (corresponds to significance level $\boldsymbol{\alpha}$ of NHST !)
  (仮説検定における有意水準$\boldsymbol{\alpha}$に相似する)
- z confidence interval for the mean μ (σ known)
  母平均のためのz信頼区間(既知分散)

$$\left[\bar{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right]$$

- t confidence interval for the mean μ (σ unknown)
  母平均のためのt信頼区間(未知分散)

$$\left[\bar{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \bar{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}\right]$$

- $t_{\alpha/2}$: $n-1$ degrees of freedom (自由度)
- s^2 : sample variance (標本分散)

Computing confidence intervals from normal data 正規母集団から信頼区間を求める (II)

- χ^2 confidence interval for σ^2 (Lecture 8 , page 7)

$$\left[\frac{n-1}{c_{\alpha/2}} s^2, \frac{n-1}{c_{1-\alpha/2}} s^2\right]$$

χ^2 is not symmetric
カイ二乗分布は対称ではない

  - χ^2 have $n-1$ degrees of freedom (自由度)
  - s^2 is the sample variance (標本分散)

- Similar to 2-sided non-critical region in NHST but here we do not set any (null) hypothesis.
仮説検定における両側の棄却しない域と相似するが、ここで仮説をしない。

Practice: computing confidence intervals

Data are 1,2,3,4 drawn from $N(\mu, \sigma^2)$ with μ unknown.

1. Find a 90% z confidence interval for μ given that $\sigma = 2$.

Below, suppose that σ is unknown.

2. Find a 90% confidence interval for μ
3. Find a 90% χ^2 confidence interval for σ^2
4. Find a 90% χ^2 confidence interval for σ.

5. Given a normal sample with $n = 100$, $\bar{x} = 12$ and $s = 5$ find a 95% confidence interval for μ

Solution:

1) $\bar{x} = 2.5$ α= 0.1 , $\alpha/2$=0.05 $z_{.05} = 1.64$ 2.5 ± 1.64

2) $s^2 = \frac{1}{3}\left(1.5^2 + \frac{1}{4} + \frac{1}{4} + 1.5^2\right) = \frac{5}{3} = 1.667$ $\mathrm{s} = 1.29$

$t_{.05}(3) = 2.353\,,\quad \bar{x} \pm 2.353 \cdot 1.29 \cdot \frac{1}{2} = 2.5 \pm\ 1.517$

3) $\left[3 \cdot \frac{1.667}{7.85}\ \ ,3 \cdot \frac{1.667}{0.352}\right] = [0.696\,,14.207]$

4) $\left[\sqrt{0.696}\,,\sqrt{14.207}\right] = [.593\,,\ 3.769]$

5) $\sqrt{n} = 10\ ,\ t_{0.025}(99) = 1.984$

$12 \pm 1.984 \cdot \frac{1}{10} \cdot 5 = 12 \pm 0.992$

Chapter 5: Confidence Intervals 信頼区間

5.2 Confidence Intervals and Null Hypothesis Significance Test 信頼区間と仮説検定の関係

Conceptual view of Confidence intervals 信頼区間の概念視点

- Computed from known data ☛ Interval Statistic
  既知のデータから求められる：統計区間
- Estimates a parameter ☛ interval estimate
  推定しようとするパラメータ:推定区間
- Confidence level: performance (in NHST: significance level)
  信頼水準：性能(仮説検定において：有意水準)
  Width: precision (in NHST: power of a test)
  幅：正確性 (仮説検定において：検定力)

A common misunderstanding concerning CI
信頼区間に対してよく起こる誤解

- A 95% confidence interval (random interval depending on the random sample) of [1.2 , 3.4] for μ does **not** mean that $P(1.2 \leq \mu \leq 3.4) = 0.95$
μの95%信頼区間 [1.2 , 3.4]とは, $\mu \in [1.2\,, 3.4]$の確率は0.95だという意味**ではない**。

- If we draw 100 samples of size n and compute a 95% confidence interval, then around 95 of them will contain μ
サイズn標本を１００回引き取って95%信頼区間を求めると、その100の中、ほぼ95は母平均μを含んでいる。

A Common Sense question (常識の問)

- How does the width of a confidence interval for the mean μ change:
  母平均μの信頼区間の幅はどんなふうに変わるか：

1. If we increase the sample size n B
2. If we increase the confidence level c A
3. We increase μ C
4. We increase σ A

(A) It gets wider
幅広くなる

(B) it gets narrower
幅の狭くなる

(C) it stays the same
同じくなる

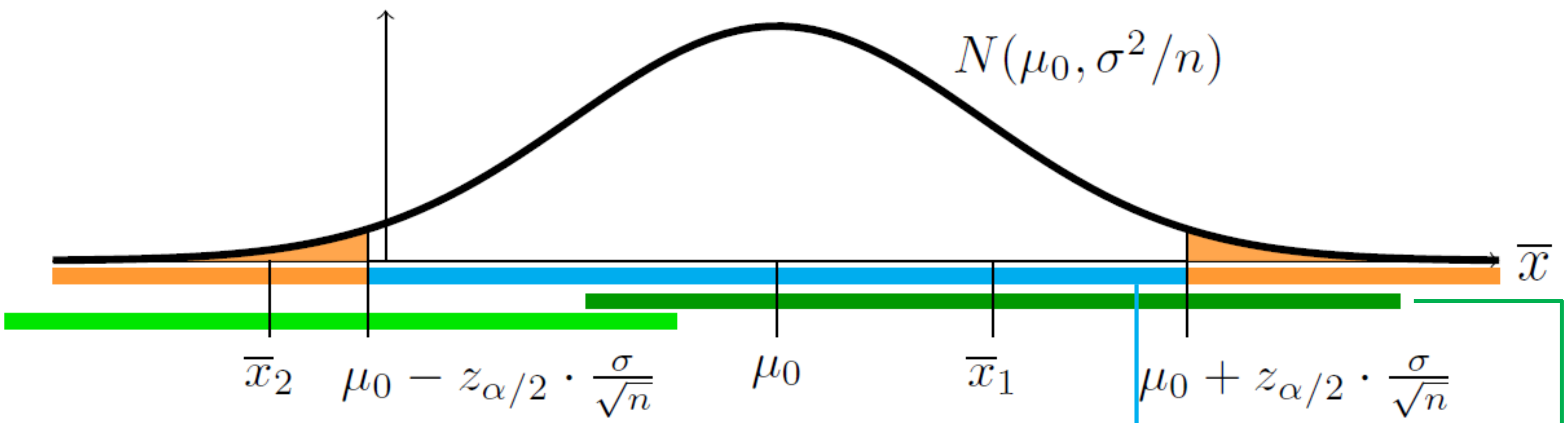
Confidence interval and non-rejection region
棄却しない域と信頼区間

- Suppose that $x_1, \dots, x_n \sim N(\mu, \sigma^2)$ with σ known.
- Consider two intervals: 以下の二つ区間を考える。

1. The z confidence interval around $\bar{x}$ at confidence level $1-\alpha$.
   $\bar{x}$周りの信頼水準$1-\alpha$のz信頼区間。
2. The z non-rejection region for the null hypothesis $H_0: \mu = \mu_0$ at significance level α.
   有意水準αの帰無仮説$H_0: \mu = \mu_0$のz棄却しない域

μ_0 is in the 1st interval $\Leftrightarrow$ $\bar{x}$ is in the 2nd interval

Confidence Interval and Non-rejection region



- Rejection region 棄卻域
- Confidence interval 1 : $\bar{x}_1 \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$
- Confidence interval 2: $\bar{x}_2 \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$
- Non-rejection region: $\mu_0 \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$

They all have **same** width !

CI: Using hypothesis testing

Unknown parameter θ of the population. Estimator: x.
θ:未知の母パラメータ。統計量：x

- For any value θ_0, we can run a NHST with null hypothesis:
  各θ_0の値に対して、下記の帰無仮説下での仮説検定を行える。

  H_0 : $\theta = \theta_0$ at significance level α.

- **Fact**: Given x, the (1 – α) confidence interval contains all θ_0 which are not rejected when they are the null hypothesis.
  $(1-\alpha)$信頼区間は棄却されない帰無仮説のθ_0からなる区間である。

- **Definition**. A type 1 CI error occurs when the confidence interval does not contain the true value of θ.
  信頼区間が実のθの値を含めないときに、第1種信頼区間過誤は起こすという。

- For a 1 – α confidence interval, the type 1 CI error rate is α.
  $(1-\alpha)$信頼区間に対して、第1種信頼区間過誤の割合はαである。

5.3 Non normal data and polling

Non-normal data 正規でない母集団から抽出されたデータ

back 20

- Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a distribution $f(x)$ with mean μ and variance σ^2 that may not be normal.
- A (version) of the CLT (Lect. 8, p. 12) says that for large n, the sampling distribution of the studentized mean is approximately standard normal:

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0,1)$$

- So for large n the $(1 - \alpha)$ confidence interval for μ is approximately.

$$\left[\bar{x} - \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}, \quad \bar{x} + \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}\right]$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value for $N(0, 1)$.

- This is called the large sample confidence interval

Large sample CI: how large ?

- How large the sample size n must be so that the approximate interval $\left[\bar{x} - \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}, \bar{x} + \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}\right]$ is a at least a 95% CI.
  サンプルサイズnは近似区間....が95%信頼区間となるようにどのくらい大きいか。
- Equivalently: Probability that the interval does not contain the true parameter is ≤ 0.05.
  同等に：区間が実のパラメータを含めない確率≤ 0.05
- How to proceed? ☛ back to definition
  どうする？ 定義を振り返る。
- If we repeat the sampling many times and compute a CI,

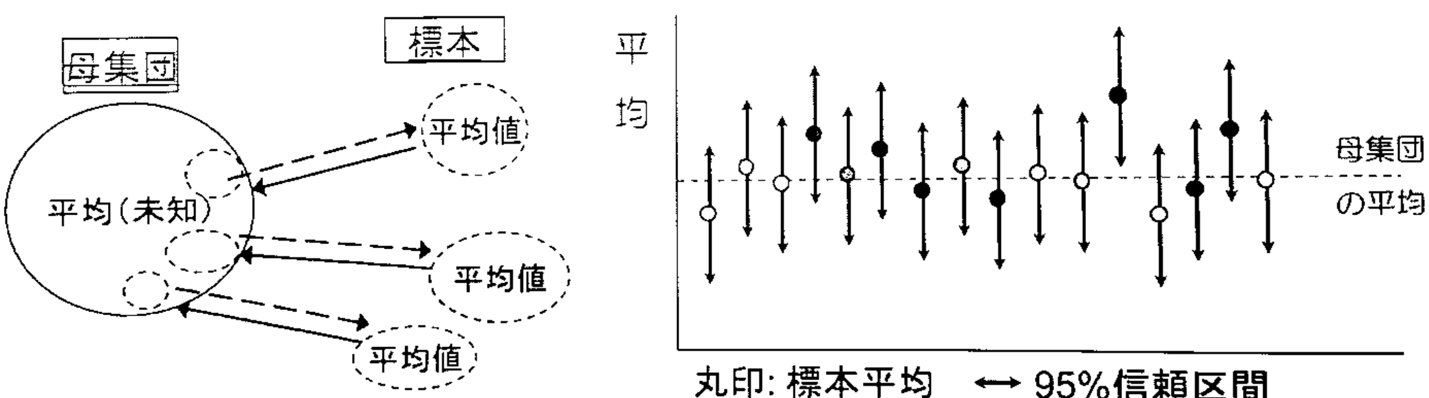


図 3. 95% 信頼区間のイメージ

then 5% of all the CI won't contain the true parameter.

Large sample CI: how large ? ☛Simulation

- Simulation. Repeat 100,000 times the experiment:

1. draw n samples from $Exponential(1)$.

2. compute the sample mean $\bar{x}$ and sample standard deviation s.

3. construct the large sample confidence interval:

$$\bar{x} \pm \frac{z_{\alpha/2} \cdot s}{\sqrt{n}}$$

4. check for a type 1 CI error, i.e. see if the true mean $\mu = 1$ is not in the interval.

- For sample size $n = 20,\ 50,\ 100,\ 400$ and confidence intervals $1 - \alpha = 0.95,\ 0.9,\ 0.8$

Simulation results with $Exp(1)$ and $N(0,1)$

n	$1-\alpha$	simulated conf.
20	0.95	0.905
20	0.90	0.856
20	0.80	0.762
50	0.95	0.930
50	0.90	0.879
50	0.80	0.784
100	0.95	0.938
100	0.90	0.889
100	0.80	0.792
400	0.95	0.947
400	0.90	0.897
400	0.80	0.798

Simulation for $\boldsymbol{Exp(1)}$

n	$1-\alpha$	simulated conf.
20	0.95	0.936
20	0.90	0.885
20	0.80	0.785
50	0.95	0.944
50	0.90	0.894
50	0.80	0.796
100	0.95	0.947
100	0.900	0.896
100	0.800	0.797
400	0.950	0.949
400	0.900	0.898
400	0.800	0.798

Simulation for $\boldsymbol{Normal(0,1)}$

- When $n \gtrsim 50-100$ usually the large sample confidence interval is precise enough for any distribution.
- **Question**: Why when $n = 20$, the simulated CI for $N(0,1)$ is lower smaller than $1-\alpha$?

Answer: small sample is more precise with t rather than z.

$z_{0.025} = 1.96$ $t_{0.025} = 2.09$

Polling: binomial proportion confidence interval 投票：二項 比率の信頼区間

- Polling with 2 candidates: 候補者が二人いるとき：
- Voters are random variables following a Bernoulli distribution
  投票者はベルヌーイ分布に従う分布とみなす.
- Data are the outcome of a sample of n voters
  データは投票者n人の標本に得られる結果である。

$$x_1, \ldots, x_n \sim Ber(p)$$

- Proportion of voters voting for A is p
  候補者Aに投票しようと考える投票者の比率はp
- A "**conservative normal**" $(1-\alpha)$ confidence interval for p is given by: $\left[\overline{\boldsymbol{x}} - \frac{\boldsymbol{z}_{\alpha/2}}{2\sqrt{\boldsymbol{n}}}, \overline{\boldsymbol{x}} + \frac{\boldsymbol{z}_{\alpha/2}}{2\sqrt{\boldsymbol{n}}}\right]$
- Proof: Use the CLT (or large sample confidence interval slide 16) and $\sigma = \sqrt{p(1-p)} \leq 1/2$.

Question about polling

- During the presidential election season, pollsters often do polls and report a 'margin of error' of about ±5%.
  大統領選挙の時期に、世論調査員は調査を行い、±5%誤差をよく報告している。
- The number of people polled is in which of the following ranges?
  必要な調査された人の人数の範囲はどれか。

a) 0 – 50

b) 50 – 100

c) 100 – 300

d) 300 – 600

e) 600 – 1000

- Answer: 5% $= 1/20$. So $20 = \sqrt{n} \Rightarrow n = 400$.

Question II:

- A $(1-\alpha)$ confidence interval for p is given by

$$\left[\bar{x}-\frac{z_{\alpha/2}}{2\sqrt{n}},\bar{x}+\frac{z_{\alpha/2}}{2\sqrt{n}}\right]$$

1. How many people would you have to poll to have a margin of error of .01 with 95% confidence?
2. How many people would you have to poll to have a margin of error of .01 with 80% confidence. (Use a table here.)

3. If $n = 900$, compute the 95% and 80% confidence intervals for p.

1) n=10,000 **2)** a=0.2, a/2=0.1 table gives 1.28 so $1.28/2\cdot\sqrt{n}$ =0.01 n=4106 **3)** 95% : $\bar{x}\pm\frac{1}{\sqrt{n}}=\bar{x}\pm\frac{1}{30}=\bar{x}\pm 0.333$.

80% : $\bar{x}\pm\frac{z_{0.1}}{2\sqrt{n}}=\bar{x}\pm 1.28\cdot\frac{1}{60}=\bar{x}\pm 0.021$

Exact binomial Confidence Interval
正確な二項信頼区間

- Exact ≠ Conservative normal ☛ exact instead of CLT.

(Slide 15) **Fact**: Given x, the (1 − α) confidence interval contains all θ_0 which are not rejected when they are the null hypothesis. $(1-\alpha)$信頼区間は棄却されない帰無仮説のθ_0からなる区間である。

Use this and table next page of **binomial(8,$\boldsymbol{\theta}$)** probabilities below to:

1. Find the rejection region with significance level 0.1 for each value θ
2. Given $x = 7$, find the 90% confidence interval for θ
3. Repeat 2. for $x = 4$.

Exact binomial Confidence Interval Solution

- For each θ, put the non-rejection region in blue, and the rejection region in red. 各θに対して，棄却しない域は青に、棄却域は赤にしなさい。

1. In each row, the rejection region has probability at most $\alpha = .10$.

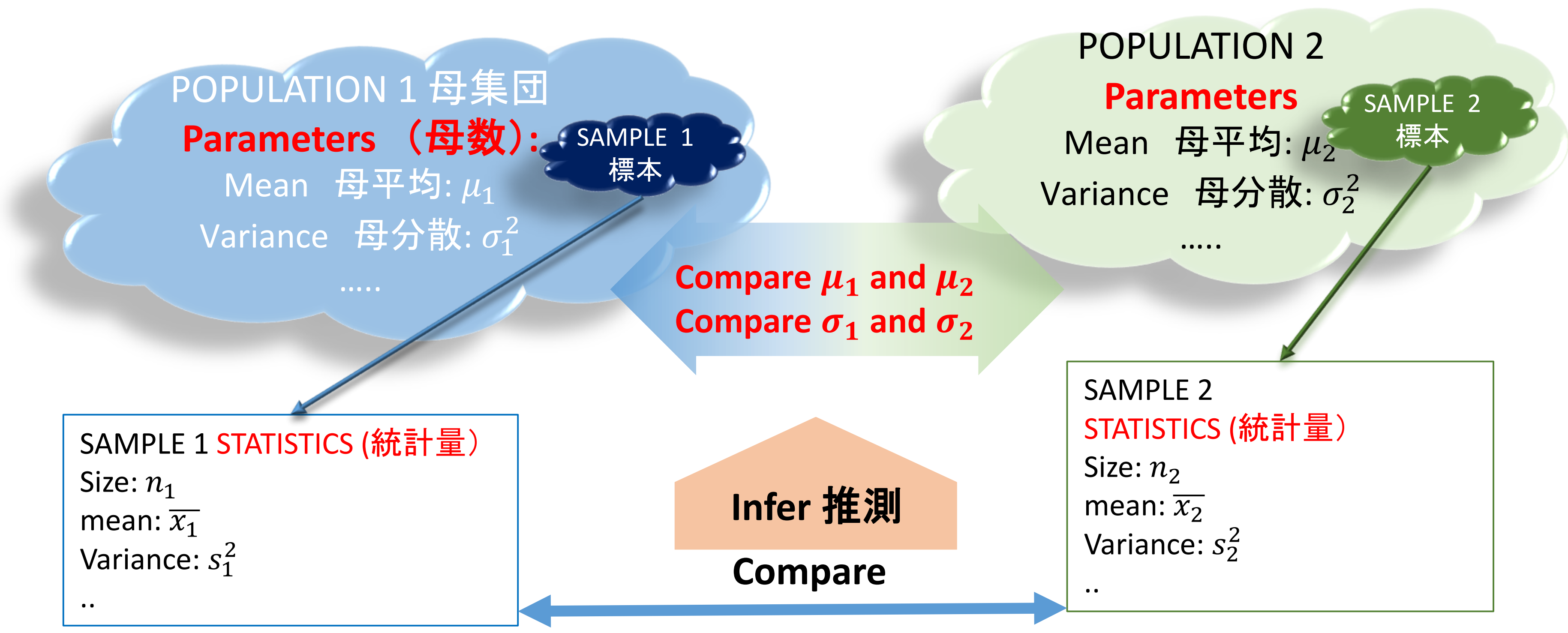
θ/x	0	1	2	3	4	5	6	7	8
.1	0.430	0.383	0.149	0.033	0.005	0.000	0.000	0.000	0.000
.3	0.058	0.198	0.296	0.254	0.136	0.047	0.010	0.001	0.000
.5	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
.7	0.000	0.001	0.010	0.047	0.136	0.254	0.296	0.198	0.058
.9	0.000	0.000	0.000	0.000	0.005	0.033	0.149	0.383	0.430

2. For x = 7 the 90% confidence interval for p is $[.7, .9]$, These are the blue entries in the x = 7 column.
3. For x = 4 the 90% confidence interval for p is [.3,.7].

Chapter 5: Confidence Intervals

5.4: CI for comparing two populations

comparing the means (independent sampling)
母平均を比する（独立な無作為抽出）
comparing the means (paired data)
平均を比する（対応のあるデータ）
comparing variances　母分散を比する



Comparing two popuations means Confidence interval for $\mu_1 - \mu_2$

- Review of two-samples t-test: Lect. 8, Sec. 4.4 (p. 15-19)
- Population 1: $\approx N(\mu_1, \sigma_1)$ Sample sizes: n_1, n_2
- Population 2: $\approx N(\mu_2, \sigma_2)$ Equal variance $\sigma_1 = \sigma_2$
- Sample means: $\overline{x_1}, \overline{x_2}$, sample variances: s_1^2, s_2^2.
- **Pooled sample variance**: $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
- **Target parameter**: $\mu_1 - \mu_2$
- **Point estimator**: $\overline{x_1} - \overline{x_2}$
- $\frac{\overline{x_1}-\overline{x_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim T_{n_1+n_2-1}$ (t-distribution, df: $n_1 + n_2 - 1$)
- **Confidence level $1 - \alpha$ CI**: $\overline{x_1} - \overline{x_2} \pm t_{\alpha/2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Practice: Bulimic students

- Practice Problem NHST II (Problem 2)

BULIMIA

Bulimic students	21	13	10	20	25	19	16	21	24	13	14			
Normal students	13	6	16	13	8	19	23	18	11	19	7	10	15	20

Source: Randles, R. H. "On neutral responses (zeros) in the sign test and ties in the Wilcoxon-Mann-Whitney test." ***The American Statistician,*** Vol. 55, No. 2, May 2001 (Figure 3).

- Infer about a difference in the average "Fear of Negative Evaluation" index between group Bulimic and group Normal.
- *(Review)* $\bar{x}_B = 17.82,\ \bar{x}_N = 14.143,\ s_B^2 = 24.16,$ $s_N^2 = 27.98$
- $s_P^2 = 26.3197$ $\sqrt{\frac{1}{n_B} + \frac{1}{n_N}} \approx 2.1$
- **Q?:** Find the 95% CI for the difference mean $\mu_B - \mu_N$.

Comparing means of paired data (対応のあるデータ)　Confidence interval for μ_D

- Review of two-samples test for paired data:
  ☛ Lecture 8, Sec. 4.5 (p. 26)
- **Two populations**: two "paired" samples of size n
  $x_1, \dots, x_n$ and $y_1, \dots, y_n$ (same size n)
  (no independent sampling)
- **Sample of differences** $x_1 - y_1, \dots, x_n - y_n$ has mean x_D and sample variance s_D^2
- **Assumption**: population of differences $\approx N(\mu_D, \sigma_D^2)$
- **Target parameter**: $\mu_D = \mu_1 - \mu_2$
- **Point estimator**: $\overline{x_D}$
- $1 - \alpha$ CI (Large n): $\overline{x_D} \pm z_{\alpha/2} \frac{s_D}{\sqrt{n}}$
- $1 - \alpha$ CI (Small n): → $\overline{x_D} \pm t_{\alpha/2} \frac{s_D}{\sqrt{n}}$

Practice paired data: Male/Female graduates salaries

- Practice Problems NHST II: Problem 3

TABLE 9.5 Data on Annual Salaries for Matched Pairs of College Graduates

Pair	Male	Female	Difference Male − Female	Pair	Male	Female	Difference Male − Female
1	$29,300	$28,800	$ 500	6	$37,800	$38,000	$ −200
2	41,500	41,600	−100	7	69,500	69,200	300
3	40,400	39,800	600	8	41,200	40,100	1,100
4	38,500	38,500	0	9	38,400	38,200	200
5	43,500	42,600	900	10	59,200	58,500	700

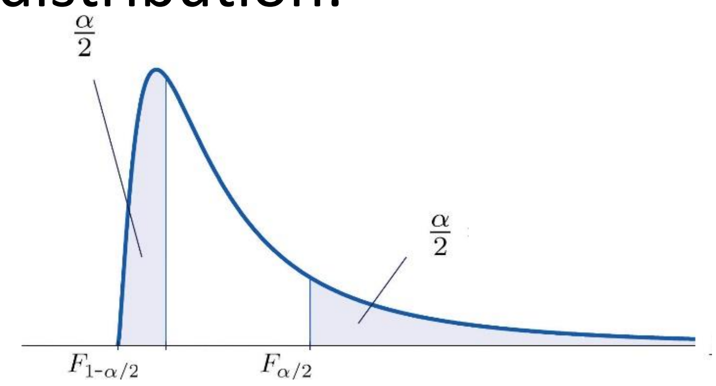
- Sample salary difference M/F.

$$\overline{x_D} = 400, \qquad s_D = 434.613, \qquad n_D = 10.$$

- **Q?** Find the 95% CI for the difference mean μ_D.

Comparing two populations variance: Confidence intervals for σ_1^2/σ_2^2

- Review of F-test ☛ Lect. 8, Section 4.6
- Population 1: $Normal(\mu_1, \sigma_1)$ Sample size 1: n_1
- Population 2: $Normal(\mu_2, \sigma_2)$ Sample size 2: n_2
- **Target parameter**: σ_1^2/σ_2^2
- **Point estimator**: s_1^2/s_2^2 (÷ of sample variances)
- Interval estimation with F_{n_1-1, n_2-1} distribution:
  - $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$.
  - $F_{1-\alpha/2} \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq F_{\alpha/2}$
  - **Confidence interval at confidence level $1-\alpha$,**



$$\frac{s_1^2/s_2^2}{F_{\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{1-\alpha/2}}$$

Practice Problem: Mice' weight variability

- Practice Problem NHST I (Problem 2):
  Find the supplier that sells mice who have the most homogeneous weight.

TABLE 9.9 Weights (in ounces) of Experimental Mice

Supplier 1					
4.23	4.35	4.05	3.75	4.41	4.37
4.01	4.06	4.15	4.19	4.52	4.21
4.29					
Supplier 2					
4.14	4.26	4.05	4.11	4.31	4.12
4.17	4.35	4.25	4.21	4.05	4.28
4.15	4.20	4.32	4.25	4.02	4.14

- $\overline{x_1} = 4.2, \overline{x_2} = 4.19, \ \ s_1^2 = 0.0408, \ \ s_2^2 = 0.00964$
- $f = s_1^2 / s_2^2 \approx 4.237$
- **Q?:** Find a 95% CI for σ_1^2 / σ_2^2.