Essential Mathematics for Global Leaders I

Spring 2019

Statistics

Lecture 9: 2019 July 1st-8th

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)

4. Null Hypothesis Significant Test (NHST) 帰無仮説検定

4.1 Concepts and 1st example: z-test
4.2 chi2 and sample variance （カイ二乗と標本分散）
4.3 the Student t-test （Student t-検定)
4.4 Two-sample t-test (=variance) 対応のないt検定
4.5 Paired difference sampling 対応のあるデータ
4.6 Comparing two population variances: F-test

4.7 chi2-test (goodness-of-fit) カイ二乗(簡単な適合度検定)
4.8 chi2-test of independence カイ二乗検定による独立性
4.9 chi2-test for Homogeneity 同質性の検定
4.10 One-way ANOVA (F-test) 一元配置分散分析 (F検定)

Chapter 4: NHST

Section 4.7 chi2-test カイ二乗検定

Background 背景: Categorical data

- $X_1, \ldots, X_k$ i.i.d.r.v. coming from an unknown distribution
  未知の独立同分布に従う確率変数$X_1, \ldots, X_k$とする。
- Each measurement of the r.v. X_i falls into k possible categories: $\forall\omega \in \Omega, X_i(\omega) \in \{v_1, \ldots, v_k\}$
  カテゴリー$v_1, \ldots, v_k$に属する確率変数の観測$X_i(\omega) \in \{v_1, \ldots, v_k\}$
  - Example: $X_i(\omega) \in \{Heads, Tails\}$ for the Bernoulli distribution
  - Example: Polling (投票): $X_i(\omega) \in$ {Candidate候補者A, Candidate候補者B, Candidate候補者C,...}

- **Categorical Data. Want to test:**
  1) **Goodness-of-fit (GOF)適合度**
  2) **independence 独立性**
  3) **Homogeneity 同質性**

Chi2-test of "goodness-of-fit (GOF)" (適合度検定)

- Aim:
  - from a sample of data $X_1(\omega_1), \dots, X_n(\omega_n)$
    標本値$X_1(\omega_1), \dots, X_n(\omega_n)$から，
    ☛ infer $p_i = P(X = v_i)$ the probability that one random sample X is in category v_i,
    各iに対し一つの標本Xがカテゴリーv_iに属する確率
    $p_i = P(X = v_i)$を推定すると目的する。
- Usual notation 通常の記号
  - O="observed" (観測)
    - $O_i = |\{\omega_r : X_r(\omega_r) = v_i\}|$ number of measurements that falls in category v_i
  - E="expected" (期待)
    - expected number (=null hypothesis) of data in each category
      帰無仮説の下で各々カテゴリーに属するデータの位数。

Chi2-test of "GOF" (適合度検定) in practice

- $E_i = |\{r: X(\omega_r) = v_i\}|$ expected number under $\boldsymbol{H_0}$ of data in category v_i 帰無仮説の下でカテゴリーv_iに属するデータの位数。
- $p_i = P(X = v_i)$ 未知 unknown
- $\boldsymbol{H_0}$:未知 $(p_1, \dots, p_k) = (p_{01}, p_{02}, \dots, p_{0k})$ 仮定
- Under $\boldsymbol{H_0}$ ($\boldsymbol{H_0}$の下で) thus ☛ $E_i = np_{0i}$
- $\boldsymbol{H_A}$: at least one $p_{0i} \neq p_i$
- **Test statistic**: $x^2 = \frac{(O_1 - E_1)^2}{E_1} + \cdots + \frac{(O_k - E_k)^2}{E_k}$
- **Theorem**: Under $\boldsymbol{H_0}$, $x^2 \sim \chi^2_{n-1}$ (n-1 自由度)
- P-value: $p = P(X > x^2)$ $X \sim \chi^2_{n-1}$
  (always 1-sided 常に右側検定: since $\chi^2 \geq 0$)
- Reject H_0 if $p < \alpha$ (α the significance level)

Example: Restaurant’s customers

- Ellen is thinking of buying a restaurant and asks about the distribution of lunch costumers.
  Ellenさんはレストランを買おうと考えている。ランチのときの客様数の分布を聞いてみる。
- The owner provides the row 1 below.
  オーナーは以下の行１を挙げる。
- Ellen records data in row 2 herself one week.
  1周間にEllenさんは自分でデータを収集して行２に記す.

	M	T	W	R	F	S
Owner’s distribution	.1	.1	.15	.2	.3	.15
Observed # of cust.	30	14	34	45	57	20

(It is a 1-way table: categories are only in one line)

- Run a chi2 goodness-of-fit test with hypotheses:
    - H_0:the owner's distribution is correct.
    - H_A: the owner's distribution is not correct.

1. The total number of observed costumers is 200.
2. Under H_0 the expected counts are: 20 20 30 40 60 30
3. $x^2 = \frac{(30-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(34-30)^2}{30} + \frac{(45-40)^2}{40} + \frac{(60-57)^2}{60} + \frac{(20-30)^2}{30}$
   $= 5 + 1.8 + 0.533 + 0.625 + 0.15 + 3.33 = 11.441$
4. df=6-1=5, $P(X > x^2) = P(X > 11.441) \leq 0.05,\ X \sim \chi_5^2$
5. Reject H_0 at significance level 0.05 in favor of the alternative hypothesis H_A: the owner's distribution is wrong.　0.05の有意水準でH_0を棄却し、対立仮説H_Aの支持に回る。

Chapter 4: NHST

Section 4.8 chi2-test for independence
分割表分析　カイ二乗検定による独立性

Example

- Admissions data (入学者) at Berkeley university.
- Is choice of major independent of gender?
  主専攻の選択は性別に依存するか？

分割法
Contingency table

	Male	Female
Easy Major	1385	133
Difficult Major	1306	1702

	B	B^c
A	1385	133
A^c	1306	1702

- Let A the event: 'choice of Easy Major'
  A^c is the 'choice of Difficult Major'
- $p = P(A)$
- Let B be the event: 'student is a male'
  and B^c be the event 'student is a female'
- $q = P(B)$

Category	$A \cap B$	$A \cap B^c$	$A^c \cap B$	$A^c \cap B^c$
Probabilty	pq	$p(1-q)$	$(1-p)q$	$(1-p)(1-q)$

- Then we are in the previous case (Section 4.7) with the one-way table above, **if we know p and q.**
  確率pとqが既存のとき、前のカイ二乗検定に帰着した
- However p and q are not known in general.
  しかし、pもqも未知である。
- We must estimate them. これらを評価しないといけない。

	B	B^c
A	X_{11}	X_{12}
A^c	X_{21}	X_{22}

Row A → $\hat{p}$; column B → $\hat{q}$

- $\hat{p} = \dfrac{X_{11}+X_{12}}{n}$
- $\hat{q} = \dfrac{X_{11}+X_{21}}{n}$

- $x^2 = \frac{(X_{11} - n\hat{p}\hat{q})^2}{n\hat{p}\hat{q}} + \frac{(X_{12} - n\hat{p}(1-\hat{q}))^2}{n\hat{p}(1-\hat{q})} + \frac{(X_{21} - n\hat{q}(1-\hat{p}))^2}{n(1-\hat{p})\hat{q}} + \frac{(X_{22} - n(1-\hat{p})(1-\hat{q}))^2}{n(1-\hat{p})(1-\hat{q})}$

"sum of 4 squares"

- Back to Berkeley admission
- $n = 4526$
- $\hat{p} = \frac{1385+133}{4526} \approx 0.34$
- $\hat{q} = \frac{1385+1306}{4526} \approx 0.59$
- $x^2 = \cdots \approx 947$ df=2 $(P(X > 947)) \approx 0$ $X \sim \chi_2^2$
- So we reject H_0 : gender and choice of major are indeed dependent. H_0を棄却する：性別と主専攻の選択は独立ではない.

	Male	Female
Easy	1385	133
Difficult	1306	1702

General chi2-test for independence

	$\boldsymbol{A_1}$	$\cdots$	$\boldsymbol{A_i}$	$\cdots$	$\boldsymbol{A_{n_1}}$	
$\boldsymbol{B_1}$	X_{11}		X_{1i}		X_{1n_1}	$\widehat{q_1}$
	$\vdots$	$\ddots$				
$\boldsymbol{B_j}$	X_{j1}		X_{ji}		X_{jn_1}	$\widehat{q_j}$
	$\vdots$			$\ddots$		
$\boldsymbol{B_{n_2}}$	$X_{n_2 1}$		$X_{n_1 i}$		$X_{n_1 n_2}$	$\widehat{q_{n_2}}$
	$\widehat{p_1}$		$\widehat{p_i}$		$\widehat{p_{n_1}}$	

- Data have two characteristics A and B 二つの分類カテゴリー
  - Characteristic A has n_1 categories: $A_1, \ldots, A_{n_1}$ with probabilities $p_1, p_2, \ldots, p_{n_1}$ 第1の分類カテゴリー
  - Characteristic B has n_2 categories $B_1, \ldots, B_{n_2}$ with probabilities $q_1, \ldots, q_{n_2}$ 第2の分類カテゴリー
- All in all, there are $n_1 n_2$ categories.
- H_0: A_i and B_j are independent for all i and j.

Chi2 independence: estimate probabilities

- $O_{ij} = X_{ij}$ observations 観測
- Under null hypothesis H_0: $E_{ij} = np_i q_j$
- We don't know p_i and q_j
- Estimate them with: $\hat{p}_i$ and $\hat{q}_j$

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n_2} X_{ji}, \qquad \hat{q}_j = \frac{1}{n}\sum_{i=1}^{n_1} X_{ji}$$

- χ^2-statistic:

$$x^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{\left(X_{ij} - n\hat{q}_j\hat{p}_i\right)^2}{n\hat{p}_i\hat{q}_j}$$

- Degrees of freedom: df=$(n_1 - 1)(n_2 - 1)$

Independence:
$P(A_i \cap B_j)$
$= P(A_i)P(B_j)$
$= p_i q_j$

Example: 血液型

	A	B	AB	O	
$Rh+$	320	96	40	412	$\widehat{p_1} =$
$Rh-$	66	23	9	65	$\widehat{p_2} =$
	$\widehat{q_1} =$	$\widehat{q_2} =$	$\widehat{q_3} =$	$\widehat{q_4} =$	

- Are the type and Rh factor independent? Test at 0.05 significance level.
  Hint: *n=1031*, $n_1 = 4, n_2 = 2$. $\hat{p}_1 = 0.84$, $\widehat{p_2} = 0.16$
  $\widehat{q_1} = 0.37$, $\widehat{q_2} = 0.12$, $\widehat{q_3} = 0.05$, $\widehat{q_4} = 0.46$
  Answer: $x^2 = 3.54$........

Chapter 4: NHST
Section 4.9 chi2-test for Homogeneity (同質性の検定) Example

- Three treatments for a disease are compared in a clinical trial, yielding the following data:
  臨床試験によって病気の三つの治療が比べられて、以下のデータが出る：

	Treatment 1	Treatment 2	Treatment 3
Cured	50	30	12
Not cured	100	80	18

(2-way table: 6 categories from 2 lines/3 columns)

- Use the chi2-test to compare the cure rate of the treatment.
  カイ二乗検定を使って治療の治癒率を比べよ。

- H_0: all the treatments have the same cure rate
- H_A: the 3 treatments have different cure rates.

- Total cure rates:
  $(\text{total cured})/(\text{total treated}) = 92/290 = 0.317$

- H_0: $(p_1, p_2, p_3) = (0.317\ ,\ 0.317\ ,\ 0.317)$
- H_A: at least one $p_i \neq 0.317$
- Under H_0, we add the expected data E_i besides O_i

	Treatment 1	Treatment 2	Treatment 3	
Cured	50, 47.6	30, 34.9	12, 9.5	92
Not cured	100, 102.4	80, 75.1	18, 20.5	198
	150	110	30	290

Example (end)

- Test statistic is thus:

$$x^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_6 - E_6)^2}{E_6}$$

$$= \frac{2.4^2}{47.6} + \frac{2.4^2}{102.4} + \frac{4.9^2}{35.9} + \frac{4.9^2}{75.1} + \frac{2.5^2}{9.5} + \frac{2.5^2}{9.5} = 2.1477$$

- Degree of freedom:
  - If we fill any two values in the table, all the other cells can be deduced ☞ df =2
    もし表の任意の成分の二つを記入したら、ほかの成分に従うため、ここで自由度は2である。

- $P(X > x^2) = P(X > 2.1477) > 0.1 \qquad X \sim \chi^2_2$
  so we do not reject H_0 (at significance level 有意水準 0.1)

Chapter 4: NHST
Section 4.10 One-way ANOVA (F-test)
Example

- Like t-test "$\mu_1 = \mu_2$" to compare the means of two populations but with n groups here.
  二つの母集団の平均値を比べるt検定と同じだが、母集団の個数は任意nとする。
- Sample Data:

Group 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,m}$
Group 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,m}$
	$\cdots$	$\cdots$	$\cdots$	
Group n	$x_{n,1}$	$x_{n,2}$		$x_{n,m}$

One-way ANOVA: preliminaries

- **Assumption:**
  - $x_{1,j} \sim Normal(\mu_1, \sigma^2)$
  - $x_{2,j} \sim Normal(\mu_2, \sigma^2)$
  - $\dots$
  - $x_{n,j} \sim Normal(\mu_n, \sigma^2)$

  - Variance σ is unknown but the same. 分散の**均一性**
  - Group means μ_i are unknown and maybe different.

- H_0: all the means are equal $\mu_1 = \mu_2 = \cdots = \mu_n$
- H_A: at least two means are not equal

Test statistic $w = MS_B / MS_W$

- $\overline{x_i} = (x_{i,1} + \cdots + x_{i,m}) \cdot 1/m$ = mean of group i
- $\bar{x}$ = grand mean among all data (総平均)
- s_i^2=sample variance of group i

$$s_i^2 = \left(\frac{1}{m-1}\right) \sum_{j=1}^{m} \left(x_{i,j} - \overline{x_i}\right)^2$$

Group 1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,m}$	$\overline{x_1}$	s_1^2
Group 2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,m}$	$\overline{x_2}$	s_2^2
	$\cdots$	$\cdots$	$\cdots$		$\vdots$	$\vdots$
Group n	$x_{n,1}$	$x_{n,2}$		$x_{n,m}$	$\overline{x_n}$	s_n^2
					$\bar{x}$	

統計量　MS_B　と　MS_W

- MS_B=between group variance
  (or Mean square treatment = 平均処理平方)
  =m × sample variance of group means
  = $(m/n-1) \cdot \sum_{i=1}^{n}(\overline{x_i} - \bar{x})^2$

- MS_W=average within group variance
  (or Mean Square Error = 平均誤差平方)
  =sample mean of $s_1^2, \dots, s_n^2$ = $\frac{1}{n}(s_1^2 + \cdots s_n^2)$

- Expected value over the sampling distribution
- $E(MS_B) = \sigma^2, \quad E(MS_W) = \sigma^2 + \frac{n\,m}{n-1}\sum_i(\mu_i - \mu)^2$
  - $\mu = \frac{1}{n}(\mu_1 + \cdots + \mu_n)$ mean over all groups

$w = MS_B/MS_W$ and F-distribution

- **Theorem**: Under H_0 *(all means are equal)*
$$w = \frac{MS_B}{MS_W} \sim F_{n-1,n(m-1)}$$

Where $F_{n-1,n(m-1)}$ is the F distribution with $n-1$ and $n(m-1)$ degrees of freedom.

- p-value: $P(W > w)$ where $W \sim F_{n-1,n(m-1)}$
  - Right-sided because $MS_B \geq MS_W$

- Reject H_0 if $p \leq \alpha$ at significance level α.
  有意水準αで$p \leq \alpha$ならばH_0を棄却する。

Example: Level of pain of treatments

The table shows patients' perceived level of pain (on a scale of 1 to 6) after 3 different medical procedures.
表は医療処置によってもたらす痛みの程度(1から6まで)を表す。

T_1	T_2	T_3
2	3	2
4	4	1
1	6	3
5	1	3
3	4	5

- Set up and run a F-test comparing the means of these 3 treatments.
- What can we say about the treatments?
- *Hint:* $\overline{x_1} = 3, \overline{x_2} = \frac{18}{5}, \overline{x_3} = \frac{14}{5}, \bar{x} = \frac{47}{15}$

Hint: $s_1^2 = \frac{5}{2}, s_2^2 = 3.3\,, s_3^2 = \frac{11}{5}$

More about ANOVA

- We assumed that all groups have same size
  ☛ Possible to generalize when each group has different size $m_1, m_2, \ldots, m_n$.
- If H_0 is rejected, (all group means are not equal) then more analysis is often necessary (*post-hoc analysis)*
  帰無仮説を棄却したら、さらなる分析が必要。
  - Tukey's HSD (Honestly Significant Differences) test.
    チューキーの母平均の対比較検定
  - Experimental Design (実験計画）、Blocking Design ブロック計画
- There are 2-way anova 二元配置分散分析 and more, called Multivariate ANOVA or MANOVA