Two-way Contingency Table (列聯表)

nominal/ordinal/discrete interval

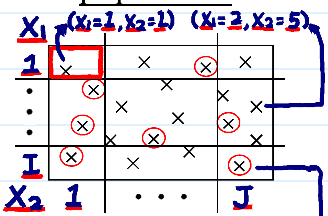
- Two *cross-classified* categorical variables X_1 and X_2

crossing cf. nesting

  - ➢ X_1 has I categories, denoted by $i = 1, 2, \ldots, I$
  - ➢ X_2 has J categories, denoted by $j = 1, 2, \ldots, J$

might be symbols only, not numerical

population



- Classifications of subjects in some *population* on X_1 and X_2 have IJ possible combinations. Define the population parameters: unknown

a sample → data
o: subjects in the sample
sampling scheme

  - ➢ π_{ij} = the proportion of the subjects in the population with $X_1=i$ and $X_2=j$

$0 \le \pi_{ij} \le 1$
$\sum_{i,j} \pi_{ij} = 1$

P(getting a subject with $X_1=i, X_2=j$) under simple random sampling

  - ▪ arrange π_{ij}'s in the *cells* of a rectangular table having I rows for categories of X_1 and J columns for categories of X_2 to display the population distribution

joint distribution of X_1 & X_2

a cell

X_1 \ X_2	1	…	J	
1	π_{11}	…	π_{1J}	π_{1+}
…	…	…	…	…
I	π_{I1}	…	π_{IJ}	π_{I+}
	π_{+1}	…	π_{+J}	$\pi_{++}=1$

a distribution on X_1

a distribution on X_2

$P(X_2=j)$: column marginal

  - ➢ $\pi_{i+} \equiv \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{+j} \equiv \sum_{i=1}^{I} \pi_{ij} \Rightarrow$ *marginal proportion*

$P(X_1=i)$: row marginal

  - ➢ $\pi_{++} \equiv \sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = \sum_{i=1}^{I} \pi_{i+} = \sum_{j=1}^{J} \pi_{+j} = 1$

a distribution

  - ➢ $\pi_{i|j} \equiv \pi_{ij}/\pi_{+j}$ and $\pi_{j|i} \equiv \pi_{ij}/\pi_{i+} \Rightarrow$ *conditional proportion*

a distribution

$\Rightarrow \sum_{i=1}^{I} \pi_{i|j} = 1, \forall j$ and $\sum_{j=1}^{J} \pi_{j|i} = 1, \forall i$

- π_{ij}: joint pmf
- π_{i+}, π_{+j}: marginal pmf
- $\pi_{i|j}, \pi_{j|i}$: conditional pmf

- **Q**: For the population, what questions might be of interest?

Note. These questions are defined in terms of parameters (i.e., π_{ij}'s)

marginal dist. is a uniform dist.

  - ➢ $\pi_{1+} = \cdots = \pi_{I+}$? or $\pi_{+1} = \cdots = \pi_{+J}$?

e.g., change of X_1 has no impact on the prob. $\pi_{i+} = P(X_1=i)$, X_1 has no main effects

  - ➢ Are X_1 and X_2 observed from a randomly sampled subject independent, i.e., does X_1 affect X_2 and vice versa? → or association

e.g., $P(X_2=1|X_1=1) \gg P(X_2=1|X_1=2)$ or $X_1=1 \Rightarrow$ often $X_2=1$, $X_1=2 \Rightarrow$ often $X_2=2$

If X_1 and X_2 are independent, then

check the table in LNp.5-1

$\pi_{ij} = P(X_1=i, X_2=j) = P(X_1=i)P(X_2=j) = \pi_{i+}\pi_{+j}$

$\pi_{i|j} = P(X_1=i|X_2=j) = P(X_1=i) = \pi_{i+}, \forall j$

$\pi_{j|i} = P(X_2=j|X_1=i) = P(X_2=j) = \pi_{+j}, \forall i$

Check LNp.3-18~19 & LNp.3-21

$\pi_{11} : \cdots : \pi_{1J} = \pi_{21} : \cdots : \pi_{2J} = \cdots = \pi_{I1} : \cdots : \pi_{IJ}$

$\pi_{11} : \cdots : \pi_{I1} = \pi_{12} : \cdots : \pi_{I2} = \cdots = \pi_{1J} : \cdots : \pi_{IJ}$

$I=2$, $J=2$

For 2×2 table, odd ratio

$X_1=1$, $\pi_{11} : \pi_{12} = \pi_{1|1} : (1-\pi_{1|1})$
$X_1=2$, $\pi_{21} : \pi_{22} = \pi_{1|2} : (1-\pi_{1|2})$

$\pi_{11} : \pi_{12}$
$\pi_{21} : \pi_{22}$

$$\Delta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\pi_{11}\times\pi_{22}}{\pi_{12}\times\pi_{21}} = 1$$

X_1 \ X_2	1	2
1	π_{11}	π_{12}
2	π_{21}	π_{22}

$1 \le i_1 < i_2 \le I$, $1 \le j_1 < j_2 \le J$

For $I \times J$ table and any $1 \le i < I$ and $1 \le j < J$,

$$1 = \frac{\Delta_{i_1 j_1} \times \Delta_{i_2 j_2}}{\Delta_{i_1 j_2} \times \Delta_{i_2 j_1}} \stackrel{?}{=} \frac{\pi_{i_1 j_1} \times \pi_{i_2 j_2}}{\pi_{i_1 j_2} \times \pi_{i_2 j_1}} \quad \text{(exercise)}$$

$$\Delta_{ij} \equiv \frac{\pi_{ij} \times \pi_{IJ}}{\pi_{iJ} \times \pi_{Ij}} = 1$$

fixed, reference

# of constraints $= (I-1)\times(J-1)$ = d.f. for interaction X_1, X_2

i \ j	j	J
i	π_{ij}	π_{iJ}
I	π_{Ij}	π_{IJ}

fixed

- For a *sample* drawn from the population, let
  - ➢ y_{ij} = total number of subjects in the sample with $X_1=i$ and $X_2=j$ (count data; π_{ij})

    carries information about the population parameters
  - ➢ marginal totals (row totals or column totals)

    $y_{i+} \equiv \sum_{j=1}^{J} y_{ij}$ and $y_{+j} \equiv \sum_{i=1}^{I} y_{ij}$ (π_{i+}, π_{+j})

X_1 \ X_2	1	…	J	
1	y_{11}		y_{1J}	y_{1+}
…	…	…	…	…
I	y_{I1}		y_{IJ}	y_{I+}
	y_{+1}	…	y_{+J}	y_{++}

a cell; counts

  - ➢ grand total $y_{++} \equiv \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} = \sum_{i=1}^{I} y_{i+} = \sum_{j=1}^{J} y_{+j}$

    carries what information? suff. stat. check examples in LNp.2-8~15

- When the cells of the rectangular table contain y_{ij}'s, it is called a $I \times J$ *contingency table* (2-way)

Data in the sample

	X_1	X_2
subject 1	2	3
subject 2	1	5
⋮	⋮	⋮

- The above treatments for π's and y's can be generalized to more than two categorical variables (e.g. 3-way table; X_1, X_2, X_3)
- **Q**: how to model the data (i.e., what's the joint distribution of y_{ij}'s)?
  - ➢ The statistical modeling of the data depends on the *sampling schemes*. Recall. Simple random sampling (SRS) on population or sub-populations; build the connection btwn parameters π & observations y

➢ consider an example of wafer data:

X_1 Quality	X_2 No Particles	Particles	ⓒ
Good	320	14	334
Bad	80	36	116
ⓑ	400	50	450 ⓐ

➢ Consider the sampling schemes

Data from different sampling schemes would carry different amounts & types of information

1. Observe the manufacturing process for a certain period of time (1-dim)
2. Decide to sample 450 wafers
3. Decide to sample 400 wafers without particles and 50 wafers with particles
4. Scheme 3 and the 450 wafers must also include, by design, 334 good wafers and 116 bad ones

1. • ⓐⓑⓒ: random • SRS from whole population

2. • ⓐ: fixed, ⓑⓒ: random • SRS from whole population

3. • ⓐⓑ: fixed, ⓒ: random • SRS from 2 sub-populations

4. • ⓐⓑⓒ: fixed • SRS from 450 wafers if X_1, X_2 independent

1, 2, 3, 4: more restrictive, lesser information

▪ Note 1: the first three schemes are all plausible

▪ Note 2: scheme 4 seems less likely in this example; such a scheme is more attractive when one level of each variable is relatively rare and we choose to over-sample both levels to ensure some representation (e.g., 50 & 116 in this case)

Recall. Fisher's Tea-Tasting experiment (LNp.2-11)

Recall. case-control study

X_1 \ X_2	1	2	
1	Y_{11}	Y_{12}	Y_{1+}
2	Y_{21}	Y_{22}	Y_{2+}
	Y_{+1}	Y_{+2}	Y_{++}

- Scheme 1
  - ➢ Model: y: fixed; Y: random; red square: free (related to dim($(Y_{ij}$'s$)$))

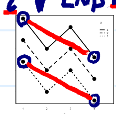
Recall. Poisson Process (LNp. 4-3)
△ $Y_{++} \sim$ Poisson(t)
π_{11} π_{12} π_{21} π_{22}
Y_{11} Y_{12} Y_{21} Y_{22}

- Response: $Y_{ij} \overset{\text{indep.}}{\sim}$ Poisson(μ_{ij}), i=1, 2; j=1, 2 → $\mu_{i+}, \mu_{+j}, \mu_{++}$
- For a *random sample* (SRS), can assume $\mu_{ij}= t\times\pi_{ij}$, (e.g. average yield → t) where t is an unknown value of a size variable ← a parameter. Note. $Y_{++}=\sum_{ij} Y_{ij} \sim$ Poisson(t)

All Y_{ij}'s have same value of the size variable (cf. LNp.4-12)

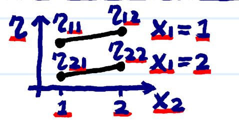
- X_1 (=i) and X_2 (=j) are covariates

check LNp 1-31



➢ Suppose the data (from an $I\times J$ table, 2×2) is fitted with a Poisson GLM with log link

$Y_{ij}\sim$ Poisson(μ_{ij}) $\xrightarrow{\log}$ $X\beta = \eta_{ij}$



$0 = (\eta_{11}-\eta_{12})-(\eta_{21}-\eta_{22}) = \log\left(\frac{\pi_{11}\times\pi_{22}}{\pi_{12}\times\pi_{21}}\right)$, $=1$

- When π_{ij}=$\pi_{i+}\pi_{+j}$ (X_1 and X_2 independent),

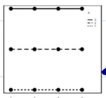
Recall. ANOVA model

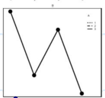
$=\gamma+\alpha_i+\beta_j$, $\alpha_I=0$, $\beta_J=0$

$$\eta_{ij} = \log(\mu_{ij}) = \log(t\,\pi_{ij}) = \log(t\,\pi_{i+}\,\pi_{+j})$$
$$= \log(t) + \log(\pi_{i+}) + \log(\pi_{+j})$$

- Intercept (1 parameter)
- Main effects of X_1 (I-1 parameters ∵ $\sum_i \pi_{i+}=1$)
- Main effects of X_2 (J-1 parameters ∵ $\sum_j \pi_{+j}=1$)

I-1 dummy var. J-1 dummy var.





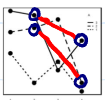
$\Rightarrow$ corresponds to a main-effect model, i.e., $Y_{ij} \sim X_1$+$X_2 \overset{*}{\equiv} S$

- When π_{ij}=$\pi_{i+}\pi_{+j}$ and $\pi_{1+} = \cdots = \pi_{I+}$ (or $\pi_{+1} = \cdots = \pi_{+J}$) — uniform dist.

$$\eta_{ij} = \log(t) + \log(\pi_{+j}) \quad (\text{or } \eta_{ij} = \log(t) + \log(\pi_{i+}))$$

$=\gamma+\beta_j$, $\beta_J=0$ $\qquad$ $=\gamma+\alpha_i$, $\alpha_I=0$

$\Rightarrow$ corresponds to the model $Y_{ij} \sim X_2$ (or $Y_{ij} \sim X_1$)



- When π_{ij}≠$\pi_{i+}\pi_{+j}$ (X_1 and X_2 not independent) → check LNp.5-3: at least one $\Delta_{ij} \neq 1$ for a pair of (i, j)

$\Rightarrow$ add interaction X_1:X_2 — (I-1)(J-1) dummy var.

$\eta_{ij}=\gamma+\alpha_i+\beta_j+(\alpha\beta)_{ij}$, $\alpha_I=0$, $\beta_J=0$, $(\alpha\beta)_{iJ}=0, \forall i$, $(\alpha\beta)_{Ij}=0, \forall j$ → $IJ-I-J+1$ (4/21)

$\hat{\mu}_{ij} \overset{\square}{=} Y_{ij}$

$\Rightarrow$ may consider $Y_{ij} \sim X_1$+X_2+X_1:$X_2 \equiv L$ (saturated model)

- **Q**: what type of π's corresponds to the following models?

- without or with interactions ⇔ independent or dependent
- without or with main effects ⇔? uniform or non-uniform marginal dist.

$Y_{ij} \sim 1 \qquad Y_{ij} \sim X_1$+$X_1$:$X_2 \qquad Y_{ij} \sim X_2$+$X_1$:$X_2$

exp $\hat{\eta} = X\hat{\beta}_{MLE}$

➢ Recall. For a Poisson GLM with log link (canonical link), $X^T Y = X^T \hat{\mu}$

$=Y_{1+}+Y_{2+}$, $=Y_{+1}+Y_{+2}$

For models without interactions,

Sufficient statistics of β

$\Rightarrow$ $X^T Y$ is only related to marginal totals (Y_{i+}'s, Y_{+j}'s)

e.g., 2×2 table
$X=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\1&0&0\end{bmatrix}$, $Y=\begin{bmatrix}Y_{11}\\Y_{12}\\Y_{21}\\Y_{22}\end{bmatrix}$, $X^TY=\begin{bmatrix}Y_{++}\\Y_{1+}\\Y_{+1}\end{bmatrix}$

$\Rightarrow$ the fitted values $\hat{\mu}$ is a function of marginal totals → $\hat{\mu}_{ij}=\exp(\hat{\eta}_{ij})=\exp(\hat{\gamma}+\hat{\alpha}_i+\hat{\beta}_j)$ (Y_{++}, Y_{i+}'s, Y_{+j}'s)

In general, consider 2-way ANOVA model in DOE

$\mu_{ij}=t\,\pi_{ij} = t\,\pi_{i+}\pi_{+j}$

$\Rightarrow$ for example, for main-effect model $Y_{ij} \sim X_1 + X_2$

$$= t\,\frac{\mu_{i+}}{\mu_{++}}\times\frac{\mu_{+j}}{\mu_{++}} = \frac{\mu_{i+}\times\mu_{+j}}{\mu_{++}} \rightarrow \hat{\mu}_{ij} = Y_{++}\,\hat{\pi}_{i+}\,\hat{\pi}_{+j} \overset{\triangle}{=} Y_{i+}\,Y_{+j}/Y_{++}$$

($t = \mu_{++}$; cf. $\square$)

$\hat{\pi}_{i+}=Y_{i+}/Y_{++}$, $\hat{\pi}_{+j}=Y_{+j}/Y_{++}$

➢ To test whether π_{ij}=$\pi_{i+}\pi_{+j}$ (H_0) $\Rightarrow H_0$: S vs. H_1: $L\backslash S$ (★ (LNp.5-5))

It's the goodness-of-fit test for S

using Poisson GLM

- Deviance based (also called G^2_S): $D_S - D_L \overset{a}{\sim} \chi^2_{(I-1)(J-1)}$ ← ∵ L is saturated ⇒ $D_L=0$