STAT 426

Author: Wenxiao Yang

Institute: Department of Mathematics, University of Illinois at Urbana-Champaign

All models are wrong, but some are useful.

Contents

Chapter 1 Basic of Categorical Data

1.1 Variable Measurement

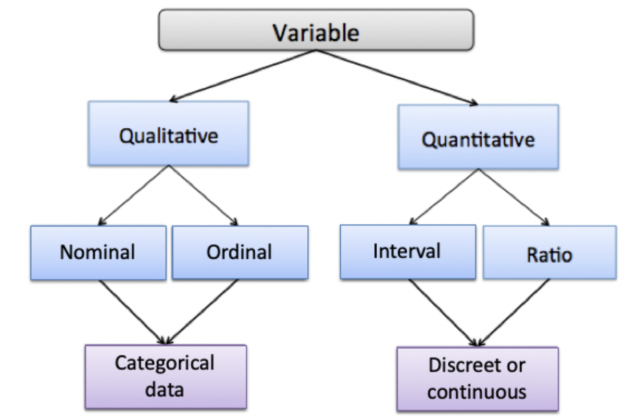

Figure 1.1: Variable Type

a) Nominal: Categories do not have a natural order. Ex. blood type, gender.

b) Ordinal: Categories have a natural order. Ex. low/middle/high education level

c) Interval: There is a numerical distance (difference between two different values is meaningful) between any two values. Ex. blood pressure level, 100 blood pressure doesn't mean the double degree of 50 pressure.

d) Ratio: An interval variable where ratios are valid (presence of absolute zero, i.e. zero is meaningful). Ex. weight, 4g is double degree of 2g, distance run by an athlete.

Levels of measurements

A variable's level of measurement determines the statistical methods to be used for its analysis.

Variables hierarchy: Ratio $>$ Interval $>$ Ordinal $>$ Nominal

Statistical methods applied to variables at a lower level can be used with variables at a higher level, but the contrary is not true.

1.2 Statistical Inference for Categorical Data

There is a distribution $F(\beta)$ with p.d.f. (p.m.f.) $f(x \mid \beta)$, where β a generic unknown parameter and $\hat{\beta}$ the parameter estimate.

1.2.1 Maximum likelihood Estimation (MLE)

Given a set of observations $\vec{x} = (x_1, ..., x_n)$, the likelihood function of these observations with parameter β is $l(\vec{x} \mid \beta)$. We want to find parameter $\hat{\beta}$ that maximizes the likelihood function,

$$\hat{\beta} = \arg\max_{\beta} l(\vec{x} \mid \beta)$$

which is also equivalent to maximizing the logarithm of the likelihood function $L(\vec{x} \mid \beta) = \log(l(\vec{x} \mid \beta))$,

$$\hat{\beta} = \arg\max_{\beta} L(\vec{x} \mid \beta)$$

Definition 1.1 (score function)

The score function is

$$u(\beta, \vec{x}) = \nabla_\beta L(\vec{x} \mid \beta) = \frac{\nabla_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)}$$

♣

Lemma 1.1 (mean of score function)

The mean of score function is 0,

$$\mathbb{E}_{\vec{x}} u(\beta, \vec{x}) = 0$$

♡

Proof 1.1

$$\begin{aligned}
\mathbb{E}_{\vec{x}} u(\beta, \vec{x}) &= \int_{\vec{x}} l(\vec{x} \mid \beta) \frac{\nabla_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)} d\vec{x} \\
&= \int_{\vec{x}} \nabla_\beta l(\vec{x} \mid \beta) d\vec{x} \\
&= \nabla_\beta \left(\int_{\vec{x}} l(\vec{x} \mid \beta) d\vec{x} \right) \\
&= \nabla_\beta 1 = 0
\end{aligned}$$

Lemma 1.2 (variance of score function)

The variance of the score function is

$$\mathrm{Var}_{\vec{x}}(u(\beta, \vec{x})) = \mathbb{E}_{\vec{x}} \left(u(\beta, \vec{x}) u(\beta, \vec{x})^T \right)$$

♡

Proof 1.2

Prove by the zero mean.

Definition 1.2 (Fisher information)

The (Fisher) information is

$$\iota(\beta) = -\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right]$$

♣

Lemma 1.3

The Fisher information is equal to the variance of score function.

$$\text{Var}_{\vec{x}}(u(\beta,\vec{x})) = \mathbb{E}_{\vec{x}}\left(u(\beta,\vec{x})u(\beta,\vec{x})^T\right) = -\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = \iota(\beta)$$

♡

Proof 1.3

$$\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = \mathbb{E}_{\vec{x}}\left(\frac{\partial \frac{\nabla_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}}{\partial \beta}\right) = \mathbb{E}_{\vec{x}}\left(\frac{\nabla^2_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)} - \frac{\nabla_\beta l(\vec{x} \mid \beta)\nabla_\beta l(\vec{x} \mid \beta)^T}{(l(\vec{x} \mid \beta))^2}\right)$$

where $\mathbb{E}_{\vec{x}}\left(\frac{\nabla^2_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}\right) = \int_{\vec{x}} l(\vec{x} \mid \beta)\frac{\nabla^2_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}d\vec{x} = \int_{\vec{x}} \nabla^2_\beta l(\vec{x} \mid \beta)d\vec{x} = \nabla^2_\beta \int_{\vec{x}} l(\vec{x} \mid \beta)d\vec{x} = \nabla^2_\beta 1 = 0$

Hence,

$$\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = -\mathbb{E}_{\vec{x}}\left(\frac{\nabla_\beta l(\vec{x} \mid \beta)\nabla_\beta l(\vec{x} \mid \beta)^T}{(l(\vec{x} \mid \beta))^2}\right) = -\mathbb{E}_{\vec{x}}\left(u(\beta,\vec{x})u(\beta,\vec{x})^T\right)$$

Proposition 1.1

When the sample x is made up of i.i.d. observations, the covariance matrix of the maximum likelihood estimator $\hat{\beta}$ is approximately equal to the inverse of the information matrix.

$$\text{Cov}(\hat{\beta}) \approx (\iota(\beta))^{-1}$$

♠

Hence, the covariance matrix can be estimated as $(\iota(\hat{\beta}))^{-1}$. Similarly, *SE* is estimated by $\sqrt{(\iota(\hat{\beta}))^{-1}}$.

1.2.2 Likelihood Inference (Wald, Likelihood-Ratio, Score)

We want to test

$$H_0 : \beta = \beta_0 \qquad H_\alpha : \beta \neq \beta_0$$

or form a confidence interval (CI) for β.

Definition 1.3 (Wald Test)

The Wald statistic:

$$z_W = \frac{\hat{\beta} - \beta_0}{SE} = \frac{\hat{\beta} - \beta_0}{\sqrt{(\iota(\hat{\beta}))^{-1}}}$$

where SE $= \sqrt{(\iota(\hat{\beta}))^{-1}}$.

Usually, as $n \to \infty$, $z_W \xrightarrow{d} N(0,1)$ under $H_0 : \beta = \beta_0$.

(1) We reject the H_0 if $|z_W| \geq z_{\frac{\alpha}{2}}$ for a two-sided level α test.

(2) The $(1-\alpha)100\%$ Wald (confidence) interval is

$$\{\beta_0 : |z_W| = \frac{|\hat{\beta} - \beta_0|}{SE} < z_{\frac{\alpha}{2}}\} = (\hat{\beta} - z_{\frac{\alpha}{2}} SE, \hat{\beta} + z_{\frac{\alpha}{2}} SE)$$

(3) The Wald test also has a chi-squared form, using

$$z_W^2 = \frac{(\hat{\beta} - \beta_0)^2}{(\iota(\hat{\beta}))^{-1}} \sim \chi_1^2 \quad \text{(under } H_0)$$

♣

Definition 1.4 (Likelihood Ratio Test)

Let

$$\Lambda = \frac{l(\vec{x} \mid \beta_0)}{l(\vec{x} \mid \hat{\beta})}$$

where $l(\vec{x} \mid \hat{\beta}) = \max_\beta l(\vec{x} \mid \beta)$, so the ratio $\Lambda \in [0, 1]$.

*The **likelihood-ratio test (LRT) chi-squared statistic**:*

$$-2 \ln \Lambda = -2 \left(L(\beta_0) - L(\hat{\beta}) \right)$$

It has an approximate χ_1^2 distribution under $H_0 : \beta = \beta_0$, and otherwise tends to be larger.

(1) Thus, reject H_0 if

$$-2 \ln \Lambda \geq \chi_1^2(\alpha)$$

(2) The $(1-\alpha)100\%$ likelihood-ratio (confidence) interval is

$$\{\beta_0 : -2 \ln \Lambda = -2 \left(L(\beta_0) - L(\hat{\beta}) \right) < \chi_1^2(\alpha)\}$$

Unlike Wald, this interval is not degenerate. (i.e., For general case, the interval does not have an explicit form.)

♣

Definition 1.5 (Score Test)

*The **score statistic**:*

$$z_S = \frac{u(\beta_0)}{\sqrt{\iota(\beta_0)}}$$

As $n \to \infty$, $z_S \xrightarrow{d} N(0, 1)$ under $H_0 : \beta = \beta_0$. Otherwise, it tends to be further from zero.

(1) Thus, reject H_0 if $|z_S| \geq z_{\frac{\alpha}{2}}$ for a two-sided level α test.

(2) The $(1-\alpha)100\%$ score (confidence) interval is

$$\{\beta_0 : |z_S| = \frac{|u(\beta_0)|}{\sqrt{\iota(\beta_0)}} < z_{\frac{\alpha}{2}}\}$$

Unlike Wald, it is not degenerate for some distributions.

(3) There is also a chi-squared form:

$$z_S^2 = \frac{u(\beta_0)^2}{\iota(\beta_0)} \sim \chi_1^2 \quad \text{(under } H_0\text{)}$$

♣

We can also use P-value to measure the probability of the statistic is more extreme under the H_0. We can reject H_0 if the P-value is $\leq \alpha$.

All three kinds tend to be "asymptotically equivalent" as $n \to \infty$. For smaller n, the likelihood-ratio and score methods are preferred.

Chapter 2 Association in Contingency Tables

2.1 Association in Two-Way Contingency Tables

Consider joint observations of two categorical variables: X with I categories, Y with J categories.

We can summarize data in an $I \times J$ **contingency table**:

		Y		
		1	$\cdots$	J
	1			
X	$\vdots$			
	I			

Each **cell** contains a count.

2.1.1 Distribution

If both X and Y are random, let

$$\pi_{ij} = P(X \text{ in row } i, Y \text{ in col } j)$$

be the **joint** distribution of X and Y.

The **marginal** distribution of X is defined by

$$\pi_{i+} = P(X \text{ in row } i)$$

and similarly for Y:

$$\pi_{+j} = P(Y \text{ in col } j)$$

The **conditional** distribution of Y given that X is in row i is defined by

$$\pi_{j|i} = P(Y \text{ in col } j \mid X \text{ in row } i) = \frac{\pi_{ij}}{\pi_{i+}}$$

2.1.2 Independent / Homogeneity

Definition 2.1 (independent)

*If both X and Y are random, they are **independent** if*

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \ \forall i, j$$

which implies $\pi_{j|i} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}, \forall i, j$. That is, $\pi_{j|i}$ doesn't depend on i and is the same as the *marginal distribution of Y. (Intuitively, knowing X tells nothing about Y.)* ♣

Definition 2.2 (homogeneity)

*Even if X is not really random, the condition that $\pi_{j|i} = \pi_{+j}, \forall i, j$ is called **homogeneity**. This might still be relevant in a situation where X is deliberately chosen and Y is observed as a response.* ♣

2.1.3 Descriptive Statistics

Let $n_{ij} =$ *count in row i and col j* and $n = \sum_i \sum_j n_{ij}$.

The **margins** of the table:

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}$$

Natural Estimation

1. Natural estimate of π_{ij}: $p_{ij} = \frac{n_{ij}}{n}$
2. Similarly marginals: $p_{i+} = \sum_j p_{ij} \quad p_{+j} = \sum_i p_{ij}$
3. And conditionals: $p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}}$

2.1.4 Sampling Models (Examples)

Possible joint distributions for counts in $I \times J$ table:

1. Poisson (random total): $Y_{ij} =$ count in cell (i, j),

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

   and the Y_{ij}s are independent.

2. Multinomial (fixed total n): $N_{ij} =$ count in cell (i, j),

$$\{N_{ij}\} \sim \text{multinomial}(n, \{\pi_{ij}\})$$

3. Independent Multinomial: Assume n_{i+} (row totals n_i) are fixed,

$$\left.\begin{array}{c} \{N_{1j}\}_{j=1}^J \sim \text{multinomial}(n_1, \{\pi_{j|1}\}_{j=1}^J) \\ \vdots \\ \{N_{Ij}\}_{j=1}^J \sim \text{multinomial}(n_I, \{\pi_{j|I}\}_{j=1}^J) \end{array}\right\}$$

   (When $J = 2$, this is independent binomial sampling, for which we may just write π_i for $\pi_{1|i}$.)

2.1.5 Measuring Inhomogeneity

Homogeneity is the condition $\pi_1 = \pi_2$. We can measure inhomogeneity by:

1. **difference of proportions**:

$$\pi_1 - \pi_2$$

2. **relative risk**:

$$RR = \frac{\pi_1}{\pi_2}$$

3. **odds ratio**:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

When $\theta = 1$, we can say there is no association.

The **odds** for a probability π is $\Omega = \frac{\pi}{1-\pi}$. Note $\pi = \frac{\Omega}{1+\Omega}$.

(In the multinomial model: $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$ ("cross-product ratio"); in Poisson model: $\theta = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}$)

The usual (unrestricted) estimates

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Useful properties of odds ratio:

(1) Interchanging rows (or cols) changes θ to $\frac{1}{\theta}$.

(2) Interchanging X and Y doesn't change θ.

(3) Multiplying a row (or col) by a factor doesn't change $\hat{\theta}$.

(4) Relationship to relative risk: $\theta = RR \cdot \frac{1-\pi_2}{1-\pi_1}$. ($\theta$ and RR are similar if both π_1 and π_2 are small.)

2.2 Conditional Association in Three-Way Tables

Add a third categorical variable Z.

Example 2.1 Is a drug more effective at curing a disease among younger patients than among older? X = drug or placebo; Y = disease cured or not; Z = age group (young, old).

2.2.1 Conditional Association

Z may be called a **stratification variable**. We are interested in the distribution of (X, Y) *conditional* on Z.

Definition 2.3 (partial table)

*Each Z category defines a **partial table** for X and Y.* ♣

Example 2.2 When $Z = 1, 2$ and X, Y are binary ($2 \times 2 \times 2$ table):

$$Z=1:\quad X \quad \overset{Y}{\begin{array}{|c|c|}\hline n_{111} & n_{121} \\ \hline n_{211} & n_{221} \\ \hline\end{array}} \qquad Z=2:\quad X \quad \overset{Y}{\begin{array}{|c|c|}\hline n_{112} & n_{122} \\ \hline n_{212} & n_{222} \\ \hline\end{array}}$$

These represent **conditional associations**.

Definition 2.4 (marginal table)

*The **marginal table** sums the partial tables:* ♣

$$X \quad \overset{Y}{\begin{array}{|c|c|}\hline n_{11+} & n_{12+} \\ \hline n_{21+} & n_{22+} \\ \hline\end{array}}$$

This represents the **marginal association** (ignoring Z).

In general, let $\mu_{ijk} =$ *expected count in row i, col j, table k.*

The **conditional odds ratios**,

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

which are estimated by

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

The **marginal odds ratio**

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

is estimated from the marginal table.

2.2.2 Simpson's Paradox

Some counter-intuitive but possible situations:

1. There are conditional associations ($\theta_{XY(k)} \neq 1$) but no marginal association ($\theta_{XY} = 1$)
2. There is a marginal association ($\theta_{XY} \neq 1$) but no conditional associations ($\theta_{XY(k)} = 1$)
3. **Simpson's paradox**: The conditional associations are in the opposite direction from the marginal, e.g. $\theta_{XY(k)} > 1, \theta_{XY} < 1$

	Full Population, N = 52			Men (M), N = 20			Women (¬M), N = 32		
	Success (S)	Failure (¬S)	Success Rate	Success	Failure	Success Rate	Success	Failure	Success Rate
Treatment (T)	20	20	50%	8	5	$\approx$ 61%	12	15	$\approx$ 44%
Control (¬T)	6	6	50%	4	3	$\approx$ 57%	2	3	$\approx$ 40%

TABLE 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).

Figure 2.1: Simpson's paradox

2.2.3 Conditional Independence, Marginal Independence

Definition 2.5 (conditionally independent given Z, marginal independent)

We also call X and Y are ***conditionally independent given*** *$Z = k$ if $\theta_{XY(k)} = 1$. If this is true for all k, X and Y are* ***conditionally independent given*** *Z. Not the same to "X and Y are* ***marginal independent*** *if $\theta_{XY} = 1$".*

Proposition 2.1

For multinomial sampling, can show that conditional independence is

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad \forall i, j, k$$

2.2.4 Homogeneous Association

Definition 2.6

Let Z have K categories. X and Y have ***homogeneous association*** *over Z if*

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

(Conditional independence is a special case.)