STA4030 Categorical Data Analysis Instructor 白胜

Lecture 1. Preliminaries Part I.

§1.1. Categorical Response Data

Definition 1. Categorical Variable → A variable has a measurement scale consisting of a set of categories is called a categorical variable. e.g. Grade → A, B, C, D, F

Definition 2. Data Set → A Data set consists of frequency counts for the categories.

(*)(*) Categorical variables can be classified into some basic classes.

Nominal variables. Ordinal variables. Interval variables

1°. Nominal → Variables having categories without a natural ordering. e.g. Gende

2°. Ordinal → Variables having ordered categories. e.g. Grades.

3°. Interval → Variables having numerical distances between any two values.

(*) The levels of categorical variables depend on the amount of information they include: Normal Variables (lowest level) → Ordinal Variables → Interval variables (Highest level)

Remark: Tests designed for low level variables can be applied to high level variables. But tests for higher level variables should not be applied to lower level variables.

§1.2. Some Important Distribution

1. Bernoulli Distribution $P(Y=y)=\pi^y(1-\pi)^{1-y},\ y=0,1 \quad \mu=E(Y)=\pi,\ Var(Y)=\pi(1-\pi)$

2. Binomial Distribution $P(Y=y)=\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y},\ y=0,1,\cdots,n.$

$$\mu=E(Y)=n\pi \quad \sigma^2=Var(Y)=n\pi(1-\pi)$$

Remark: $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi) \Rightarrow \sum_{i=1}^{n} Y_i \sim N(n\pi, n\pi(1-\pi))$ as n grows large

3. Multinomial Distribution: For a multinomial random variable N with n trails and c possible outcomes with probabilities $\pi=(\pi_1,\pi_2,\cdots,\pi_c)$. $N\sim \text{Multi}(n,\pi)$

$$P(N_1=n_1,N_2=n_2,\cdots,N_c=n_c)=P(n_1,n_2,\cdots,n_c)=\frac{n!}{n_1!n_2!\cdots n_c!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c} \quad \sum_{j=1}^{c}n_j=n$$

$$\mu_j=E(N_j)=n\pi_j \quad j=1,2,\cdots,c \quad Var(N_j)=n\pi_j(1-\pi_j) \quad Cov(N_j,N_h)=-n\pi_j\pi_h$$

Remark: 1°. Multinomial(n, π) with $c=2$ is equivalent to the binomial distributio

2°. The marginal distribution of each N_i is binomial. $N_j \sim \text{Binomial}(n, \pi)$.

4. Poisson Distribution → Describe the counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent,

$P(Y=y) = \frac{e^{-\mu}\mu^y}{y!}$, $y=0,1,2,\cdots$ $E(Y)=\mu$, $Var(Y)=\mu$.

Remark: 1°. The poisson distribution approaches the normal distribution $N(\mu,\mu)$ as $\mu \to \infty$

2°. If $Y_i \sim \text{Poisson}(\mu_i)$, $i=1,2,\cdots,c$ are independent, then $\sum_{i=1}^{c} Y_i \sim \text{Poisson}(\sum_{i=1}^{c}\mu_i)$

3°. Consider c independent Poisson variables, $Y_1, Y_2, \cdots, Y_c$ with parameters $\mu_1, \mu_2, \cdots, \mu_c$. Then the distribution of $Y := (Y_1, Y_2, \cdots, Y_c)$ conditioned on the event $\sum_{i=1}^{c} Y_i = n$ is Multinomial(n, π), where $\pi = (\pi_1, \pi_2, \cdots, \pi_c)$ and $\pi_j = \frac{\mu_i}{\sum_{i=1}^{c}\mu_i}$ $i=1,2,\cdots,c$

5. Negative Binomial Distribution

Form 1. $P(Y=y) = \binom{y-1}{r-1}\pi^r(1-\pi)^{y-r}$, $y=r, r+1, \cdots$ $r=$ # of successes, $Y=$ # of trails until r successes.

Form 2. $P(Y=y) = \binom{y+r-1}{y}\pi^y(1-\pi)^r$, $y=0,1,\cdots$, $r=$ # of failures, $Y=$ # of successes until r failure

Form 3. $P(Y=y) = \frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}\left(\frac{r}{\mu+r}\right)^r\left(1-\frac{r}{\mu+r}\right)^y$, $y=0,1,\cdots$; $r=$ # of successes, $Y=$ # of failures until r successes

$E(Y)=\mu$, $Var(Y)=\mu+\frac{\mu^2}{r}$

§1.3. Likelihood and Maximum-Likelihood Estimation

The overall likelihood is a product of the individual likelihoods:

$$L(\theta; y) = L(\theta; y_1) \times L(\theta; y_2) \cdots \times L(\theta; y_n) = \prod_{i=1}^{n} L(\theta; y_i) = \prod_{i=1}^{n} p(y_i; \theta)$$

e.g. $Y = \text{Poisson}(\mu)$, $y = (y_1, y_2, \cdots, y_n) \Rightarrow L(\mu; y) = e^{-n\mu}\frac{\mu^{\sum_{i=1}^{n} y_i}}{y_1! y_2! \cdots y_n!}$

(*) Loglikelihood function $l(\theta; y) = \log(L(\theta; y)) = \sum_{i=1}^{n} l(\theta; y_i)$ (for computational reasons)

(**) Maximum-Likelihood Estimation $\hat{\theta} = \arg\max_{\theta} l(\theta; y)$

e.g. $Y = \text{Poisson}(\mu)$, $l(\mu, y) = -n\mu + \sum_{i=1}^{n} y_i \log(\mu) - \sum_{i=1}^{n}\log(y!)$ Kernal

$$\frac{\partial l(\mu; y)}{\partial \mu} = -n + \frac{1}{\mu}\sum_{i=1}^{n} y_i \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ (sample mean)}$$

§1.4 Large Sample Inference

The asymptotic properties of maximum likelihood estimators provide ways for us to make large sample inference on the parameters of discrete distributions.

(*)(*)(*) Three significance tests of a null hypothesis: $H_0: \theta=\theta_0$ versus $H_1: \theta\neq\theta_0$

1°. Wald test and CI. Fisher Information $I(\theta)=\int_x \left(\frac{d}{d\theta}\log f(x|\theta)\right)^2 f(x|\theta)\,dx$

$\hat{\theta}$ is the unrestricted MLE. $\iota(\hat{\theta})$ is the fisher information evaluated at $\hat{\theta}$.

the Wald test statistic: $z=\frac{(\hat{\theta}-\theta_0)}{SE}$, $SE=\frac{1}{\sqrt{\iota(\hat{\theta})}}$. $z \overset{\text{Approximate}}{\sim} N(0,1)$ when $\theta=\theta_0$.

example. Given a sample of n i.i.d. Bernoulli random variables with probability of success π. Consider $H_0: \pi=\pi_0$ vs $H_1: \pi\neq\pi_0$. Use wald test.

Solution: Fisher Information $I(\theta)=\sum_x \left(\frac{d}{d\theta}\log f(x|\theta)\right)^2 f(x|\theta)$ $\quad f(x|\theta)=\theta^x(1-\theta)^{1-x}$

$$\text{Then } I_{x_i}(\theta)=\sum_x\left(\frac{d}{d\theta}\left(x\ln\theta+(1-x)\ln(1-\theta)\right)\right)^2\theta^x(1-\theta)^{1-x}=\sum_{x=0}^{1}\left(\frac{x}{\theta}-\frac{1-x}{1-\theta}\right)^2\theta^x(1-\theta)^{1-x}$$

$$=\left(\frac{1}{1-\theta}\right)^2(1-\theta)+\frac{1}{\theta^2}\theta=\frac{1}{1-\theta}+\frac{1}{\theta}=\frac{1}{\theta(1-\theta)} \qquad I_x(\theta)=nI_{x_i}(\theta)=\frac{n}{\theta(1-\theta)}$$

(unrestricted)

MLE. $L(x;\pi)=\prod_{i=1}^{n}\pi^{x_i}(1-\pi)^{1-x_i}=\pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i}$

$$\frac{\partial}{\partial\pi}l(x;\pi)=\frac{\partial}{\partial\pi}\log L(x;\pi)=\frac{\partial}{\partial\pi}\left(\sum_{i=1}^n x_i\ln\pi+\left(n-\sum_{i=1}^n x_i\right)\ln(1-\pi)\right)$$

$$=\frac{1}{\pi}\sum_{i=1}^n x_i-\frac{1}{1-\pi}\left(n-\sum_{i=1}^n x_i\right)\overset{\text{set}}{=}0 \Rightarrow \hat{\pi}=\frac{\sum_{i=1}^n x_i}{n}=\bar{x} \Rightarrow \iota(\hat{\pi})=\frac{n}{\bar{x}(1-\bar{x})}$$

Hence, under H_0, Wald test statistic $z=\frac{\bar{x}-\pi_0}{\sqrt{\bar{x}(1-\bar{x})/n}}\sim N(0,1)$

Then the related $100(1-\alpha)\%$ confidence interval is $\hat{\pi}\pm z_{\frac{\alpha}{2}}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$

2°. Score test and CI. Score function: $\boxed{u(\theta)=\frac{\partial l(\theta;y)}{\partial\theta}}$ → loglikelihood function

Generally speaking, the larger the absolute value of $u(\theta_0)$, $(u(\hat{\theta})=0)$
the less the data supports the null hypothesis H_0.

test statistic: $z=\frac{u(\theta_0)}{\sqrt{\iota(\theta_0)}}$ → Null SE. Does not require to compute MLE.

the test statistic $z\sim N(0,1)$ Approximately.

Example: n i.i.d Bernoulli random variables. $z=\frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\sim N(0,1)$

$H_0: \pi=\pi_0 \quad H_1: \pi\neq\pi_0$

3°. Likelihood Ratio Test and CI.

Define the ratio: $\boxed{\Lambda = \frac{l_0}{l_1}}$ → maximized likelihood under H_0. → maximized likelihood under $H_0 \cup H_1$.

Define the LR test statistic. $\boxed{-2\log(\Lambda) = 2(\log(L_1) - \log(L_0))}$

Has a chi-squared distribution in the limit as $n \to \infty$.

(*) DF is the difference between the dimensions of the parameter spaces under $H_0 \cup H_1$ and H_0

(*)(*)(*) Remark: The three tests are asymptotically equivalent, which means that in the limit, their test statistics will follow a chi-square distribution with the same df. if H_0 is true.

(*) The Wald test is the most commonly used, because it is the simplest.

Lecture 3. One-way Tables.

1. Notation: 1°. Represent a one-way table with c categories by a vector $X = (X_1, X_2, \cdots, X_c)$, where X_j is the count/frequency ($X_j \to RV$) (*)

2°. Represent the observed counts/frequencies in cell j by n_j. $n = \sum_{j=1}^{c} n_j$

3°. Let $\pi = (\pi_1, \pi_2, \cdots, \pi_c)$ be the joint distribution of $(X_1, X_2, \cdots, X_n)$ $\sum_{j=1}^{c} \pi_j = 1$

4°. Estimate π_j by $p_j = \frac{n_j}{n}$ → sample version of π_j.

2. Important Question about how the data is generated.

Question: Did sampling occur with a fixed sample size or not?

$$\begin{cases} \text{Fixed} \to \text{Binomial / Multinomial sampling} \\ \text{Not fixed} \to \text{Poisson sampling (Perhaps)} \end{cases}$$

3. Binomial Sampling

3.1. Characterisation. ① n is fixed

② Each observation is a "trail" with only two possible outcomes.

③ The trails are IID.

3.2. Inference ① MLE: $\hat{\pi} = \frac{n_s}{n}$, where category s is the "success" category.

② "large" sample size n is needed to use the Wald, Score, LR test for π.

A good rule: $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

③ If proportions are extreme, e.g. π or $1-\pi < 0.2$, the Score and LR can be preferred to the Wald.

4. Multinomial Sampling 多项抽样.

4.1. Characterisation. ① n is fixed ② The trails are IID.

③ Each observation is a trial with only c possible outcomes.

4.2. Compute MLE $X \sim Mult(n, \pi) \Rightarrow p(X_1 = n_1, X_2 = n_2, \cdots, X_c = n_c) = \frac{n!}{n_1! \cdots n_c!} \pi_1^{n_1} \cdots \pi_c^{n_c}$

Loglikelihood function: $l(\pi; (n_1, n_2, \cdots, n_c)) = \sum_{j=1}^{c} n_j \log(\pi_j) + \text{Const}$.

for $j = 1, 2, \cdots, c-1$, $\frac{\partial l(\pi; (n_1, n_2, \cdots, n_c))}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{1 - \sum_{j=1}^{c-1} \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c}$

Setting these c-1 equations to zero $\Rightarrow \boxed{\hat{\pi}_j = \frac{n_j}{n_c} \hat{\pi}_c}$. $j = 1, 2, \cdots, c-1$.

$\sum_{j=1}^{c-1} \hat{\pi}_j = (n - n_c) \frac{\hat{\pi}_c}{n_c} = 1 - \hat{\pi}_c \Rightarrow \hat{\pi}_c = \frac{n_c}{n} \Rightarrow \hat{\pi}_j = \frac{n_j}{n}$, $j = 1, 2, \cdots, c$.

4.3. Hypothesis $H_0: \pi = \pi_0 = (\pi_{10}, \pi_{20}, \cdots, \pi_{c0})$ $H_1: \pi \neq \pi_0$

(vector)

where π_0 is a completely specified distribution.

Remark: The LR is best placed to test hypotheses like this.

(*) Hypothesis testing using LR test statistic G^2

$$G^2 = -2\log\Lambda = -2\log \prod_{j=1}^{c} \left(\frac{n\pi_{j0}}{n_j}\right)^{n_j} = 2\sum_{j=1}^{c} n_j \log\left(\frac{n_j}{n\pi_{j0}}\right) \sim \chi^2_{c-1}$$

Remark: Because under H_0 no parameter were estimated and under $H_0 \vee H_1$, we need to estimate c-1 of the π_j's ($\hat{\pi}_c = 1 - \hat{\pi}_1 - \hat{\pi}_2 - \cdots - \hat{\pi}_{c-1}$).

Table:

Categories	X_1	X_2	X_3	······	X_c
$O_j = \text{Count } n_j$	n_1	n_2	n_3	······	n_c
$E_j = n\pi_{j0}$	$n\pi_{10}$	$n\pi_{20}$	$n\pi_{30}$	······	$n\pi_{c0}$
$O_j \log\left(\frac{O_j}{E_j}\right)$	$O_1 \log\left(\frac{n_1}{n\pi_{10}}\right)$		······		$O_c \log\left(\frac{O_c}{E_c}\right)$

$H_0: \pi = \pi_0 = (\pi_{10}, \pi_{20}, \cdots, \pi_{c0})$

$H_1: \pi \neq \pi_0$

$G^2 = 2\sum_{j=1}^{c} O_j \log\left(\frac{O_j}{E_j}\right) \sim \chi^2_{c-1}$

(*)(*)(*) Remark: If π_0 is unknown, then we need to estimate it using the data we have. Still, test statistic $\boxed{G^2 = 2\sum_{j=1}^{c} O_j \log\left(\frac{O_j}{E_j}\right) \sim \chi^2_r}$

where $r = c-1-\#$ parameters estimated under H_0.

4.4. Pearson's Chi-squared Test → to see if models fit table data.

①. Hypothesis: H_0: Model M_0 fits. H_1: Model M_0 does not fit

②. Test Statistic: $\boxed{X^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_r}$, $r = c-1-\#$ parameters estimated under H_0

Table:

(*)(*)(*)

Categories	x_1	x_2	x_3	......	x_c
O_j = Count n_j	n_1	n_2	n_3	......	n_c
$E_j = n\hat{\pi}_{j0}$	$n\hat{\pi}_{10}$	$n\hat{\pi}_{20}$	$n\hat{\pi}_{30}$	......	$n\hat{\pi}_{c0}$
$\frac{(O_j-E_j)^2}{E_j}$	$\frac{(n_1-n\hat{\pi}_{10})^2}{n\hat{\pi}_{10}}$	......			$\frac{(n_c-n\hat{\pi}_{c0})^2}{n\hat{\pi}_{c0}}$

Conclusion: If X^2 & G^2 are similar, we can be confident that the large-sample approximation to normality has worked.

Remark: O_j → observed count of category j.

E_j → the expected count of category j if H_0 were true.

(*)(*)(*) Small expected cell counts: 1°. the rule of thumb used to be $E_j \geq 5$

2°. We can have E_j for at most 20% cells, none of the E_js can be smaller 1

3°. If some of the E_js are too small, combining (*) categories.

5. Poisson Sampling / Distribution

5.1. Characterisation: ① The total sample size n is not fixed.

② The counts $X_1, X_2, \cdots, X_c$ are independent Poisson variables, with rate $\mu_1, \mu_2, \cdots, \mu$

③ The poisson distribution itself requires independence of events.

5.1. Inference: (*) Given a sample $y_1, y_2, \cdots, y_n$ from a Poisson(μ) distribution.

MLE: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (sample mean). $n\uparrow$ $\hat{\mu} \sim N(\mu, \frac{\mu}{n})$

Wald test statistic: $\boxed{\frac{\hat{\mu}-\mu_0}{\sqrt{\hat{\mu}/n}}}$

Score test statistic: $\boxed{(\hat{\mu}-\mu_0)/\sqrt{\mu_0/n}}$

LR test statistic: $\boxed{2n(\mu_0 - \hat{\mu}) + 2n\hat{\mu}\log(\hat{\mu}/\mu_0)}$

Lecture 4. Two-way tables: Tests for Independence and Homogeneity

1. Two-Way Tables.

1.1. Response and explanatory variables.

1°. Two-way tables involve two categorical variables. X with r categories (row variable) and Y with c (column variable)

2°. Both X and Y are response variables → talk about joint distribution

3°. Y → response variable X → explanatory variable → talk about conditional distribution (Y given X)

1.2. Notation.

1°. n → total number of observations (sample size).

2°. n_{ij} → number of observations in row i and column j.

3°. $p_{ij} = \frac{n_{ij}}{n}$ → proportion of the total sample falling in the (i,j) $\sum_i \sum_j p_{ij} = 1$

4°. $\{p_{ij}\}$ → joint distribution

5°. $\{p_{i+}\}, \{p_{+j}\}$ → marginal distribution

6°. $\{p_{j(i)}\}, \{p_{i(j)}\}$ → conditional distribution

1.3. An (r×c) contigency table is (example):

X \ Y	1	2	3	⋯	c	Total
1	n_{11}	n_{12}	n_{13}	⋯	n_{1c}	n_{1+}
2	n_{21}	n_{22}	n_{23}	⋯	n_{2c}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	⋯	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	n_{+3}	⋯	n_{+c}	n

2. Sampling Models

2.1. Poisson Sampling:

Each cell frequency n_{ij} has an independent Poisson distribution with mean μ_{ij}. The joint probability mass function is $\boxed{\prod_{i,j} \frac{\mu_{ij}^{n_{ij}} e^{-\mu_{ij}}}{n_{ij}!}}$

2.2. Multinomial Sampling:

If the total sample size n is fixed and each element of the sample is classified according to two categories X and Y, the joint distribution of n_{ij} is $Mult(n, \pi)$, with $\pi = (\pi_{11}, \pi_{12}, \cdots, \pi_{1c}, \pi_{21}, \pi_{22}, \cdots, \pi_{rc})$.

Probability mass function is $\frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}}$

2.3. Product Multinomial Sampling

Usually the sampling scheme when one of the variables is the response.

And the other is the explanatory variable.

Treat row total as fixed, and using the notation n_{i+}. Suppose independent:

Multinomial form:

$Y\mid X=i$	1	2	3	⋯	c	Total
n_{ij}	n_{i1}	n_{i2}	n_{i3}	⋯	n_{ic}	n_{i+}
$\pi_{j(i)}$	$\pi_{1(i)}$	$\pi_{2(i)}$	$\pi_{3(i)}$	⋯	$\pi_{c(i)}$	π_{i+}

$$\boxed{\frac{n_{i+}!}{\prod_{j=1}^{c} n_{ij}!}\prod_{j=1}^{c}\pi_{j(i)}^{n_{ij}}}$$

2.4. Hypergeometric Sampling 超几何抽样

fixed by design

Studies in which both marginal totals (row and column) of the contingency table are

Example: $r=c=2$. We have N balls, of which K are red and the rest are blue. We draw a sample of n balls without replacement. P(sample contains k red balls)?

Solution: $p(k;N,K,n)=\dfrac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$ for $k=0,1,2,\cdots,n$ 超几何分布.

distribution

Remark: when the table is larger than 2x2, we have multivariate hypergeometric

3. Test of Independence and test of Homogeneity.

3.1. Definition

A principal aim of many studies is to compare conditional distribution of Y at various levels of explanatory variables.

Independence $\to$ $p(X=i,Y=j)=p(X=i)p(Y=j) \Leftrightarrow \pi_{ij}=\pi_{i+}\pi_{+j}$ for $\begin{cases} i=1,2,\cdots,r \\ j=1,2,\cdots,c \end{cases}$

3.2. Test of Independence.

Observed frequencies:

$X\setminus Y$	1	2	⋯	c	Total
1	n_{11}	n_{12}	⋯	n_{1c}	n_{1+}
2	n_{21}	n_{22}	⋯	n_{2c}	n_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
r	n_{r1}	n_{r2}	⋯	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	⋯	n_{+c}	n

Sampling model: Multinomial model with size n

X and Y responses

Frequent table:

$X\setminus Y$	1	2	⋯	c	Total
1	π_{11}	π_{12}	⋯	π_{1c}	π_{1+}
2	π_{21}	π_{22}	⋯	π_{2c}	π_{2+}
⋮	⋮	⋮	⋮	⋮	⋮
r	π_{r1}	π_{r2}	⋯	π_{rc}	π_{r+}
Total	π_{+1}	π_{+2}	⋯	π_{+c}	1

$$\sum_i\sum_j \pi_{ij} = \sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$$

Under H_0, the MLEs of π_{i+} and π_{+j} are $\hat{\pi}_{i+} = \frac{n_{i+}}{n}$, $\hat{\pi}_{+j} = \frac{n_{+j}}{n}$. $\begin{cases} i=1,2,\cdots,r \\ j=1,2,\cdots,c \end{cases}$

Then the estimated expected frequencies are $E_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$

Then LR test statistic is $G^2 = 2\sum_{cells} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{n_{ij}n}{n_{i+}n_{+j}}\right)$

Pearson's Chi-squared test statistic is $X^2 = \sum_{cells}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$

Both X^2 and G^2 have the χ^2_{df} distribution.

$df =$ # parameters estimated under $H_0 \cup H_1$ − # parameters estimated under H_0

$= rc-1-((r-1)+(c-1)) = (r-1)(c-1)$

3.3. Test of Homogeneity

Population distribution:

X \ Y	1	2	⋯	c	Total
1	$\pi_{1(1)}$	$\pi_{2(1)}$	⋯	$\pi_{c(1)}$	1.0
2	$\pi_{1(2)}$	$\pi_{2(2)}$	⋯	$\pi_{c(2)}$	1.0
⋮	⋮	⋮	⋮	⋮	⋮
r	$\pi_{1(r)}$	$\pi_{2(r)}$	⋯	$\pi_{c(r)}$	1.0

$\sum_{j=1}^{c}\pi_{j(i)} = 1 \quad \forall i=1,2,\cdots,r$

Hypothesis: $H_0: \pi_{j(1)} = \pi_{j(2)} = \cdots \pi_{j(r)} = \pi_j$, $j=1,2,\cdots,c$

The MLE of π_j under H_0 is $\hat{\pi}_j = \frac{n_{+j}}{n}$

Then the estimated expected frequencies are

$E_{ij} = n_{i+}\hat{\pi}_{j(i)} = \frac{n_{i+}n_{+j}}{n}$

$X^2, G^2 \sim \chi^2_{df} \quad df = (r-1)(c-1)$

3.4. Fisher's Exact Test

Both marginal totals are fixed, the LR and Pearson Chi-squared tests are not appropriate for this kind of data.

Lecture 5. Two-way-table: Measures of Association

1. Introduction: If we apply the LR and Pearson's Chi-squared tests to some data and reject the null hypothesis of "independence". G^2 and X^2 tell us nothing about the direction of the dependence, nor the magnitude of it. They only tell if it is statistically significant or not. Need statistics which tell about the size & direction of the independence

2. (*)(*)(*) Difference of Proportions

2.1. Assumption: Restrict ourselves to 2×2 tables.

Use generic terms success and failure for the response categories of a binary variable

(*) Compare the probability of a successful response in row 1 with the probability of a successful response in row 2.

2.2. Difference of proportions: $\boxed{\delta = \pi_{1(1)} - \pi_{1(2)} = \frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}}}$ $\quad -1 \le \delta \le 1$

Remark: When the sampling is product binomial and the data exhibits homogeneity $\delta = 0$

When the sampling is Poisson or Multinomial and the data exhibits independence. $\delta = 0$

(*)(*)(*) Inference: 1°. MLE of δ is $\boxed{\hat{\delta} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}}$

2°. The large sample $100(1-\alpha)\%$ Wald confidence interval for δ is $\boxed{\hat{\delta} \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{\delta})}$

where the estimated standard error of $\hat{\delta}$ is (treating two rows as independent binomial samples).

$$\boxed{\hat{\sigma}(\hat{\delta}) = \sqrt{\frac{\frac{n_{11}}{n_{1+}}\left(1-\frac{n_{11}}{n_{1+}}\right)}{n_{1+}} + \frac{\frac{n_{21}}{n_{2+}}\left(1-\frac{n_{21}}{n_{2+}}\right)}{n_{2+}}} = \sqrt{\frac{n_{11}n_{12}}{n_{1+}^3} + \frac{n_{21}n_{22}}{n_{2+}^3}}}$$

3. (*)(*)(*) Relative Risk

3.1. Introduction: The ratio of proportions of success in each row is called relative risk.

$$\boxed{RR = \frac{\pi_{1(1)}}{\pi_{1(2)}} = \frac{\pi_{11}/\pi_{1+}}{\pi_{21}/\pi_{2+}}} \rightarrow \text{Can be any nonnegative real number.}$$

Remark: 1°. The relative risk can have different interpretations to the difference of proportions.

2°. A relative risk of 1 means that response is independent of group.

3°. RR is probably a better measure of association the δ when proportions are extreme.

3.2. Inference. 1°. MLE for RR is $\boxed{\hat{RR} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}}$

2°. The large sample $100(1-\alpha)\%$ Wald confidence interval for $\log(RR)$ is

$$\boxed{\log(\hat{RR}) \pm z_{\frac{\alpha}{2}}\hat{\sigma}(\log(\hat{RR}))}, \quad \boxed{\hat{\sigma}(\log(\hat{RR})) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}}$$

4. ⊛⊛⊛ Odds Ratios

4.1. Introduction:

Odds: The odds of an event is the ratio of the probability of the event occurring to the probability the event does not occur. $\boxed{\text{odds} = \Omega = \frac{\pi}{1-\pi}}$

X\Y	S	F
1	π_{11}	π_{12}
2	π_{21}	π_{22}

$\Omega_1 = \frac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}}$: Odds of S to F for Y given X=1

$\Omega_2 = \frac{\pi_{21}/\pi_{2+}}{\pi_{22}/\pi_{2+}}$: Odds of S to F for Y given X=2.

$$\boxed{\text{Odds Ratio} = \theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \text{ (Cross-product ratio)}}$$

4.2. Properties:

1°. $\theta = 1 \Leftrightarrow$ X and Y are independent.

2°. $\theta \in [0,1) \Rightarrow$ Individuals in row 2 are less likely to fall in column 2 than are individuals in row 1, more when $\theta \in (1, +\infty)$

3°. Odds Ratio = Relative risk $\times \left(\frac{\pi_{22}/\pi_{2+}}{\pi_{12}/\pi_{1+}}\right)$

4.3. Inference:

1°. MLE of θ is $\boxed{\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}}$

2°. $\boxed{\log\hat{\theta} \xrightarrow{\text{Large Sample}} N(\log\theta, \sigma^2(\log\hat{\theta})), \quad \sigma^2(\log\hat{\theta}) = \frac{1}{n\pi_{11}} + \frac{1}{n\pi_{12}} + \frac{1}{n\pi_{21}} + \frac{1}{n\pi_{22}}}$

Can estimate $\sigma^2(\log\hat{\theta})$ by $\hat{\sigma}^2(\log\hat{\theta}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \Rightarrow \boxed{\log\hat{\theta} \pm z_{\frac{\alpha}{2}}\hat{\sigma}(\log(\hat{\theta}))}$

4.4. For r×c Tables

$\boxed{\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}}$ $i = 1, 2, \cdots, r-1$, $j = 1, 2, \cdots, c-1$.

These $(r-1)(c-1)$ odds ratios determine all odds ratios formed from pairs of rows and columns

4.5. Row Fractions

For each row, divide the observed frequency in every cell by the sum of the frequencies of the row

Compare the row fractions among the rows and if the row fractions are identical among all rows, the variables are not associated.

Lecture 6. Two-way Tables: Ordinal Data

Deal with ordering properties.

1. Introduction:

When we tested for Independence before, X^2 and G^2 allowed for any kind of statistical dependence. They need $(r-1)(c-1)$ degrees of freedom.

Most ordinal tests require <u>only one degree of freedom</u>, because they are testing a particular type of association that can be summarized in <u>one parameter</u>.

(*) X^2 and G^2 ignore the ordering of rows and columns. → $\boxed{\text{Ordinal Data}}$

2. Ordinal Measure of Association: Gamma

2.1. Concordant & Discordant pairs.

(X_i, Y_i) (X_j, Y_j) 2 points

A pair is <u>concordant</u> if the subject ranked higher on X also ranks higher on Y.

A pair is <u>discordant</u> if the subject ranking higher on X ranks lower on Y.

The pair is tied if the subjects have the same classification on X and on Y.

2.2. Gamma.

Denote the <u>total number of</u> concordant pairs by C.

Denote the <u>total number of</u> discordant pairs by D.

Given that a pair is untied on both variables:

$\pi_c/(\pi_c+\pi_d)$ → the probability of ~~cord~~ concordance

$\pi_d/(\pi_c+\pi_d)$ → the probability of discordance

Gamma: $\boxed{\gamma = \frac{\pi_c-\pi_d}{\pi_c+\pi_d}}$ → the difference between these probabilities.

Sample version: $\hat{\gamma} = \frac{C-D}{C+D}$ $\gamma \in [-1, 1]$

Remark: Gamma treats the variables symmetrically. i.e. It is unnecessary to identify one classification as a response variable.

3. Ordinal Measure of Association: Correlation

3.1. Pearson's ρ → describes the strength of a linear trend in the population.

Assign scores to categories. Let $u_1 \le u_2 \le \cdots \le u_I$ and $v_1 \le v_2 \le \cdots \le v_J$ denote scores for X and Y, respectively.

Remark: the scores should reflect the distances between categories, with greater distances between categories regarded as farther apart.

Let $\boxed{\bar{u} = \sum_{i=1}^{r} u_i p_{i+}}$ → the sample mean of the row scores.

$\boxed{\bar{v} = \sum_j v_j p_{+j}}$ → the sample mean of the columns scores.

$\boxed{\sum_{i,j} (u_i - \bar{u})(v_j - \bar{v}) p_{ij}}$ → sample covariance of X and Y.

Sample correlation: $\boxed{r = \dfrac{\sum_{i,j} (u_i - \bar{u})(v_j - \bar{v}) p_{ij}}{\sqrt{[\sum_i (u_i - \bar{u})^2 p_{i+}][\sum_j (v_j - \bar{v})^2 p_{+j}]}}}$

dimension

Remark: the larger $|\rho|$ is, the farther the data fall from independence in the linea

3.2. Inference.

1°. Estimating ρ via r is informative.

(Two-sided)

2°. Necessary Hypothesis are H_0: X and Y are independent vs $H_1: \rho \neq 0$.

Appropriate test statistic is $\boxed{M^2 = (n-1) r^2 \sim \chi^2_{(1)}}$ → $\begin{cases} n \text{ large} \\ \text{Approximately} \end{cases}$

(Large values contradict independence.

3°. For a one-sided test, H_0: X and Y are independent vs $H_1: \rho > 0$

Appropriate test statistic is $\boxed{M = \sqrt{n-1}\, r \sim N(0,1)}$ → $\begin{cases} n \text{ Large} \\ \text{Approximately} \end{cases}$

Interpretation

(*)

Lecture 7. Three-way Tables

1. Introduction.

1.1. Definition: Let X, Y and Z denote three categorical response variables.

X → I categories, Y → J categories, Z → K categories. I×J×K table.

Remark: In most studies it is important to investigate how variables are interrelated.

1.2. Notation: n_{ijk} → number of units belonging to $X=i, Y=j, Z=k$

$n_{i+k} = \sum_{j=1}^{J} n_{ijk}$ → "+" indicates summing over the categories of a variable

π_{ijk} → the probability that a randomly selected member of the population belongs to $X=i, Y=j, Z=k$. $\sum_i \sum_j \sum_k \pi_{ijk} = 1$ $\boxed{\hat{\pi}_{ijk} = \frac{n_{ijk}}{n}}$ MLE

1.3. Sampling schemes

(1) Poisson Sampling → n random, each cell count a Poisson random variable

(2) Multinomial sampling → n fixed. sample allocated to cells according to $\{\pi_{ijk}\}$

(3) Product-Multinomial Sampling → Happen in several different ways.

2. Partial and Marginal Associations

2.1. partial tables and Marginal tables

Partial table: Cross-sections of the three-way table. Display the relationship between X and Y while holding the level of the third variable constant

Marginal table: Obtained by summing counts in the partial tables. Display the relationship between two variables, Y and Z.

2.2. Conditional and Marginal Odds Ratios

Conditional/partial Association: Association obtained from a partial table.

Marginal Association: Association obtained from a marginal table.

Can be measured by the appropriate odds ratios.

Consider a 2×2×K table, where K denotes the number of categories of the control variable.

Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling model.

Then for each level of Z, $\boxed{\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \quad k=1,2,\cdots,K}$ → XY conditional odds ratios

Marginalizing over $Z \Rightarrow \boxed{\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}} \to XY$ marginal odds ratio

Then the sample analogues are $\boxed{\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}}$ $k=1,2,\cdots,K$ $\boxed{\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}}$

3. Race and the Death Penalty Example. Simpson's Paradox

Usually caused by the association between variable X and Z or between Y and Z.

4. Types of Independence

4.1 Conditional Independence

For three-way tables, we say Y and X are conditionally independent at level k of Z if $P(Y=j|X=i, Z=k) = P(Y=j|Z=k)$ for all levels i, j.

We say Y and X are conditionally independent given Z if Y and X are conditionally independent at every level of Z.

X and Y conditionally independent given Z $\Leftrightarrow \boxed{\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}}$ for $\forall i, j, k$

For 2×2×K tables, X and Y conditionally independent $\Leftrightarrow \boxed{\theta_{XY(k)} = 1}$ for $k=1,2,\cdots,K$

4.2. Marginal Independence

In terms of the population proportions, X and Y are marginal independent if $\boxed{\pi_{ij+} = \pi_{i++}\pi_{+j+}}$

X and Y are marginally independent $\Leftrightarrow \theta_{XY} = 1$

Remark: Conditional independence does not imply marginal independence.

4.3. Homogeneous Association

A 2×2×K table has homogeneous XY association when:

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

Conditional independence of X and Y is the special case: $\theta_{XY(k)} = 1.0$, $k=1,2,\cdots,K$.

Remark: When homogeneous XY association occurs, there is no interaction between X and Y in their effects on Z.

5. Cochran-Mantel-Haenszel Methods

5.1. Introduction:

(1) Tests of conditional independence and homogeneous association with the K conditional odds ratios in 2×2×K tables.

② Combine the sample odds ratios from the K partial tables into a single summary of measure of partial association.

5.2. The Cochran-Mantel-Haenszel (CMH) test

Null Hypothesis: $H_0: \theta_{XY(k)} = 1,\ k = 1, 2, \cdots, K$ (X and Y are conditionally independent).

Under H_0: $\boxed{\mu_{11k} = E(n_{11k}) = \frac{n_{1+k} n_{+1k}}{n_{++k}}}$ $\boxed{Var(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n_{++k}^2 (n_{++k} - 1)}}$

test statistic: $\boxed{CMH = \frac{\left(\sum_k (n_{11k} - \mu_{11k})\right)^2}{\sum_k Var(n_{11k})}} \sim \chi^2_{(1)}$ (large sample)

Remark: CMH statistic takes larger values when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative for all partial tables.

(*)(*)(*) 1. This test is inappropriate when the association varies dramatically among the partial tables. It works best when the XY association is similar in each partial table.

(*)(*)(*) 2. The ~~CH~~ CMH statistic combines information accross partial tables

(*)(*)(*) 3. It is improper to combine results by adding the partial tables together to form 2x2 table for test. Simpson's paradox might occur.

5.3. Estimation of common odds ratio

More informative to estimate the strength of association

In a 2x2xK table. suppose that $\boxed{\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}}$

Then Mantel-Haenszel estimator $\boxed{\hat{\theta}_{MH} = \frac{\sum_k \left(\frac{n_{11k} n_{22k}}{n_{++k}}\right)}{\sum_k \left(\frac{n_{12k} n_{21k}}{n_{++k}}\right)}}$

Remark: If the true odds ratios are not identical but do not vary drastically, $\hat{\theta}_{MH}$ still provides a ~~sa~~ useful summary of the K conditional associations.