ELEMENTS OF THE ANALYSIS OF DISCRETE DATA

M. Zelen
Harvard University

July, 2000

Slide 1

Slide 2

Lecture 3. Independence and Urn Sampling (Slide 59)

8. Test for Independence
9. Urn Sampling Model
10. Wilcoxon Rank Sum Test

Slide 3

Lecture 4. Correlated Outcomes

(Slide 94)

11. Independence of Binary Outcomes
12. Matched Pairs

Slide 4

Slide 5

Slide 6

Lecture 7. Multivariate Problems

(Slide 151)

19. Comparison of k Binomial Populations
20. Testing Two Multinomial Populations
21. Analogue Between Logistic Regression and Polychotomous Regression (one covariate)

Slide 7

Lecture 8: Polychotomous Regression

(Slide 161)

22. Review (Testing k binomials)
23. Testing Two Multinomial Distributions
24. Logistic Regression and Polychotomous Regression (one-covariate)
    - Urn Sampling Model and Asymptotics
    - Examples
      Testing two Multinomials
      Generalized Wilcoxon Test
      (Rank Sum Test)

Slide 8

Lecture 1: Examples and Elementary Theory

1. Introduction: Some examples
2. Theory
3. Logistic Regression

Slide 9

1. Introduction: Some examples

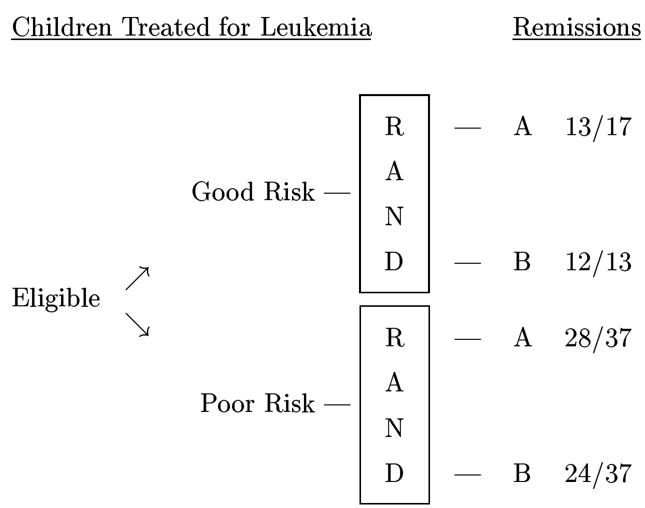
1.1 One 2×2 Contingency Tables

Treatment for Diabetes		Death Rate (5 yrs)
RAND — Placebo	11/205	= .054
RAND — Tolbutamide	26/204	= .127

	Deaths	Survival	Total
Placebo	11	194	205
Tolbutamide	26	178	204
	37	372	409

Slide 10

1.2 Two Contingency Tables



A: MTX → 6MP

B: 6MP → MTX

Slide 11

Children Treated for Leukemia

Good Risk

	Remissions	Failure	
A	13	4	17
B	12	1	13
	25	5	30

Poor Risk

	Remissions	Failure	
A	28	9	37
B	24	13	37
	52	22	74

Slide 12

1.3 Retrospective Studies

Lung Cancer Patients — Random Sample (n=63) — Are you a heavy smoker? 60/63

Controls (No Lung Cancer) — Random Sample (n=43) — Are you a heavy smoker? 32/43

<u>Data</u>: P(Smoker|Lung Cancer) =60/63

P(Smoker|Control)=32/43

Slide 13

Retrospective Studies (continued)

Can one say something about:

P(Lung Cancer|Smoker)

P(Lung Cancer|Non-Smoker)

	Smoker	Non-Smoker	
Lung Cancer	60	3	63
Control	32	11	43
	92	14	106

Slide 14

1.4 Multinomial Classification

Suppose individuals are classified according to two variables; e.g., Hodgkin's disease patients were classified according to whether the mediastinum had disease or not and by cell type.

	Mediastinum		
	Disease	No Disease	
Nodular			
Sclerosing Cells	15	9	24
Mixed Cells	6	21	27
	21	30	51

Is disease involvement in mediastinum and cell type independent?

Slide 15

1.5 Order:

Is there a relationship between birth order and the frequency of mongoloid children?

	Birth Order				
	1	2	3	4	5 or greater
Frequency	$\frac{248}{559{,}510}$	$\frac{175}{398{,}903}$	$\frac{74}{190{,}252}$	$\frac{27}{71{,}093}$	$\frac{8}{34{,}993}$
	44×10^{-5}	44×10^{-5}	39×10^{-5}	38×10^{-5}	26×10^{-5}

Slide 16

1.6 Independence of Binary Outcomes

A sequence of 0's and 1's are observed. Is the sequence independent?

Example: Presidential Elections in U. S.

1 ⇒ Democratic Elected, 0 ⇒ Republican Elected

1912	1	Wilson	44	1	Truman	76	1	Carter
16	1	”	48	1	”	80	0	Reagan
20	0	Hardy	52	0	Eisenhower	84	0	”
24	0	”	56	0	”	88	0	Bush
28	0	Hoover	60	1	Kennedy	92	1	Clinton
32	1	Roosevelt	64	1	Johnson	96	1	”
36	1	”	68	0	Nixon			
40	1	”	72	0	”			

Slide 17

1.7 Wilcoxon Two Sample Rank Test

Consider Two Groups of Observations

A: 10, 18, 22, 33 $n_A = 4$

B: 12, 35, 40, 45, 48 $n_B = 5$

Arrange Data as an Ordered Sample

Rank	1	2	3	4	5	6	7	8	9
	10	12	18	22	35	36	40	45	48
A	1	0	1	1	0	1	0	0	0
B	0	1	0	0	1	0	1	1	1

Is there a trend or are the 0's and 1's random.

Slide 18

2. Theory

Binary Random Variable $Y = \begin{cases} 1 \\ 0 \end{cases}$

$$\begin{aligned} \theta &= P\{Y = 1\} \\ 1 - \theta &= P\{Y = 0\} \end{aligned}$$

$$f(y) = P\{Y = y\} = \theta^y (1-\theta)^{1-y} \quad \text{for } y = 0, 1 \qquad (1)$$

Suppose $y_1, y_2, \cdots, y_n$ is a sequence of *iid* random variables following (1).

Joint distribution:

$$\begin{aligned} f(y_1, y_2 \cdots, y_n)) &= \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^{\Sigma y_i} (1-\theta)^{n - \Sigma y_i} = \theta^s (1-\theta)^{n-s} \end{aligned}$$

where

$$s = \sum_{1}^{n} y_i = \text{ number of 1's}$$

$$f(y_1, \cdots, y_n) = \theta^s (1-\theta)^{n-s}.$$

Slide 19

$$f(y_1, \cdots, y_n) = \theta^s (1-\theta)^{n-s}.$$

$S = \sum_{i=1}^{n} y_i$ is sufficient statistic for θ

i.e., If $f(y_1, \cdots, y_n | \theta)$ is joint distribution and $t(\boldsymbol{y})$ is the function of $\boldsymbol{y} = (y_1, \cdots, y_n)$ so that $f(\boldsymbol{y} | t(\boldsymbol{y}))$ is independent of θ, then $t(y)$ is sufficient for θ.

Slide 20

Joint Distribution $\boxed{f(\boldsymbol{y}) = \theta^s(1-\theta)^{n-s}}$

Since S is sufficient, consider distribution of S; i.e.,

$$\begin{aligned} f(s) &= P\{S=s\} = \sum_{y_1+\cdots+y_n=s} f(y_1, y_2, \cdots, y_n) \\ &= \theta^s(1-\theta)^{n-s} \sum_{y_1+\cdots+y_n=s} 1 = \binom{n}{s}\theta^s(1-\theta)^{n-s} \end{aligned}$$

as $\sum_{y_1+\cdots+y_n=s} 1 =$ number of ways of arranging $y_1, \cdots, y_n$ so that they always sum to s.

Binomial Distribution:

$f(s) = \binom{n}{s}\theta^s\,(1-\theta)^{n-s}$

$E(S) = n\theta, \quad V(S) = n\theta(1-\theta)$

Slide 21

Non-identically distributed Random Variables

Consider $Y_1, Y_2, \cdots, Y_n$ independent r.v.

$$P\{Y_i = y_i\} = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

$$\boxed{f(y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}}$$

$$S = \sum_{i=1}^{n} Y_i \quad E(S) = \sum_{1}^{n} \theta_i, \quad V(S) = \sum_{1}^{n} \theta_i(1-\theta_i)$$

Two-Populations

Suppose

$$\begin{aligned} \theta_i &= \theta_1 \quad \text{for } i = 1, 2, \cdots, n_1 \\ \theta_i &= \theta_2 \quad \text{for } i = n_1+1, n_1+2. \cdots, n \end{aligned}$$

$$n = n_1 + n_2$$

Slide 22

Then

$$f(y_1,\cdots,y_n) = \prod_{i=1}^{n_1}\theta_1^{y_i}(1-\theta_1)^{1-y_i} \times \prod_{i=n_1+1}^{n}\theta_2^{y_i}(1-\theta_i)^{1-y_i}$$

$$= f(y_1,\cdots,y_{n_1})f(y_{n_1+1},\cdots,y_n) \quad = \theta_1^{s_1}(1-\theta_1)^{n_1-s_1} \times\theta_2^{s_2}(1-\theta_2)^{n_2-s_2}$$

$$s_1 = \sum_{1}^{n_1} y_i, \quad s_2 = \sum_{n_1+1}^{n} y_i, \quad n = n_1 + n_2$$

Since (S_1, S_2) are sufficient for (θ_1, θ_2)

$$f(s_1,s_2) = \binom{n_1}{s_1}\theta_1^{s_1}(1-\theta_1)^{n_1-s_1}\binom{n_2}{s_2}\theta_2^{s_2}\,(1-\theta_2)^{n_2-s_2}$$

Slide 23

3. Logistic Regression

Consider $Y_1,\cdots,Y_n$ to be ind. binary r.v. with

$$\theta_i = e^{\alpha+\beta x_i}\big/\,1+e^{\alpha+\beta x_i}$$

$$1-\theta_i = 1\,\big/\,1+e^{\alpha+\beta x_i}$$

Note

$$\theta_i/(1-\theta_i) = e^{\alpha+\beta x_i}$$

$$\boxed{\log[\theta_i/(1-\theta_1)] = \alpha+\beta x_i}$$ Logistic Regression

Observations: (x_i, Y_i) $\quad$ Y_i: binary r.v.

$i = 1, 2, \cdots, n$

Usually inference is made on β; i.e.,

$$H_0 : \beta = 0 \ \ vs. \ \ H_i : \beta \neq 0$$

The parameter α is a nuisance parameter.

Slide 24

Examples

Two Population Problem

(2×2 Contingency Tables)

(x_i, Y_i): Observations

$\lambda_i = \log \theta_i/(1-\theta_i) = \alpha + \beta x_i$

$x_i = 1$ for $i = 1, 2, \cdots, n_1$

$x_i = 0$ for $i = n_1 + 1, \cdots, n_1 + n_2$

$\theta_i = \frac{e^{\lambda_i}}{1+e^{\lambda_i}} = e^{\alpha+\beta x_i} \Big/ 1 + e^{\alpha+\beta x_i}$

$\theta_1 = e^{\alpha+\beta}/1 + e^{\alpha+\beta}$ for $x_i = 1$

$\theta_2 = e^{\alpha}/1 + e^{\alpha}$ for $x_i = 0$

If $\beta = 0 \Longrightarrow \theta_1 = \theta_2$

$$\lambda_1 = \log \frac{\theta_1}{1-\theta_1} = \alpha + \beta, \lambda_2 = \log \frac{\theta_2}{1-\theta_2} = \alpha$$

$$\begin{aligned} \lambda_1 - \lambda_2 = \beta \quad &= \log \tfrac{\theta_1}{1-\theta_1} - \log \tfrac{\theta_2}{1-\theta_2} = \log \left\{ \tfrac{\theta_1/1-\theta_1}{\theta_2/1-\theta_2} \right\} \\ &= logodds \end{aligned}$$

Slide 25

Trends Tests

$$\lambda_i = \alpha + \beta x_i$$

Suppose $x_i = i$

$\lambda_i = \alpha + \beta i$

Order:	1	2	3	$\cdots$	n
Y_i:	1	0	0	$\cdots$	1

Is the order of the $\{Y_i\}$ a random sequence?

If random sequence $\beta = 0$

Slide 26

Mongoloid Children

	Birth Order				
	1	2	3	4	5
# mongoloids	s_1	s_2	s_3	s_4	s_5
Total no. births	n_1	n_2	n_3	n_4	n_5

Define

$$
\begin{aligned}
x_i &= 1 \quad \text{for} \quad i = 1, 2, \cdots, n_1 \\
& \quad\; 2 \quad \text{for} \quad i = n_1 + 1, \cdots, n_1 + n_2 \\
& \quad\; 3 \quad \text{for} \quad i = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3 \\
& \quad\; 4 \quad \text{for} \quad i = n_1 + n_2 + n_3 + 1, \cdots, n_1 + n_2 + n_3 + n_4 \\
& \quad\; 5 \quad \text{for} \quad i = n_1 + n_2 + n_3 + n_4 + n_5 + 1, \cdots, n_1 + \cdots + n_5
\end{aligned}
$$

Slide 27

Markovian Sequence (Presidents)

$$Y_1, Y_2, \cdots, Y_n$$

Assume $P(Y_i|Y_1, Y_2, \cdots, Y_{i-1}) = P(Y_i|Y_{i-1})$.

Conditional distribution only depends on previous observation.

Markovian Sequence: $P(Y_i|Y_{i-1})$

Independent Sequence: $P(Y_i|Y_{i-1}) = P(Y_i)$

Model: $P(Y_i|Y_{i-1}) = e^{\alpha+\beta Y_{i-1}} \big/ i + e^{\alpha+\beta Y_{i-1}}$

$$
\begin{aligned}
\lambda_i &= \alpha + \beta Y_{i-1} \\
&= \begin{cases} \alpha & if \quad Y_{i-1} = 0 \\ \alpha + \beta & \text{if} \quad Y_{i-1} = 1 \end{cases}
\end{aligned}
$$

If $\beta = 0 \Rightarrow$ Independent Sequence

$\beta \neq 0 \Rightarrow$ Markovian Sequence

Slide 28

Multinomial Classification (4 groups)

Example:

Mediastinal Disease	(Y_1)
Cell Type	(Y_2)

$$Y_1 = \begin{cases} 1 & \text{if disease present} \\ 0 & \text{if disease absent} \end{cases}$$

$$Y_2 = \begin{cases} 1 & \text{if nodular sclerosis} \\ 0 & \text{if mixed cell} \end{cases}$$

Slide 29

Observations: $(Y_1, Y_2) = (0,0), (0,1), (1,0), (1,1)$

$$\begin{aligned} \theta_1 &= P\{Y_1 = 1\} = e^{\alpha_1} \big/ 1 + e^{\alpha_1} \\ \theta_2(Y_1) &= P\{Y_2 = 1|y_1\} = e^{\alpha_2+\beta y_1} \big/ 1 + e^{\alpha_2+\beta y_1} \end{aligned}$$

$$\begin{aligned} P(Y_1, Y_2) &= P(Y_1)P(Y_2|y_1) = \theta_1^{y_1}(1-\theta_1)^{1-y_1}\theta_2(y_1)^{y_2}(1-\theta_2(y_1))^{1-y_2} \\ &= \frac{e^{\alpha_1 y_1}}{1+e^{\alpha_1}} \cdot \frac{e^{(\alpha_2+\beta y_1)y_2}}{1+e^{(\alpha_2+\beta y_1)}} \\ &= \frac{e^{\alpha_1 y_1+\alpha_2 y_2+\beta y_1 y_2}}{(1+e^{\alpha_1})(1+e^{(\alpha_2+\beta y_1)})} \end{aligned}$$

If $\beta = 0$

$\Rightarrow$ $P(Y_1, Y_2) = \dfrac{e^{\alpha_1 y_1+\alpha_2 y_2}}{(1+e^{\alpha_1})(1+e^{\alpha_2})} = \left(\dfrac{e^{\alpha_1 y_1}}{1+e^{\alpha_1}}\right)\left(\dfrac{e^{\alpha_2 y_2}}{1+e^{\alpha_2}}\right)$

$\Rightarrow$ Independence

Slide 30

Lecture 2: Logistic Regression

- Review

4. Sampling Theory of Logistic Regression
5. Conditional Distributions
6. Comparing Two Binomial Populations

Slide 31

Lecture 2
Review

$$\begin{aligned} Y &= \begin{cases} 1 \\ 0 \end{cases} \quad \text{Binary Random Variable} \\ \theta &= P\{Y=1\}, \quad 1-\theta = P\{Y=0\} \\ E(Y) &= \theta, V(Y) = \theta(1-\theta) \\ f(y) &= \theta^y(1-\theta)^{1-y} \quad \text{for } y = 0,1 \\ &= 1-\theta \qquad \text{for } y=0 \\ &= \theta \qquad\quad \text{for } y=1 \end{aligned}$$

Let $Y_1, \cdots, Y_n$ be independent binary random variables such that $\theta_i = \{Y_i = 1\}$

Slide 32

Then the joint distribution is

$$\boxed{f(y_1, \cdots, y_n) = \prod_{i=1}^{n} \theta_i^{y_i} (1 - \theta_i)^{1-y_i}}$$

If $\theta_i = \theta \Rightarrow f(y_1, \cdots, y_n) = \theta^s (1-\theta)^{n-s}$

$$s = \sum_{1}^{n} y_i$$

$P\{S = s\} = f(s) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$

Binomial Distribution

Slide 33

$$S = \sum_{1}^{n} Y_i$$

$$E(S) = n\theta, \quad V(S) = n\theta(1-\theta)$$

Logistic Regression

x = Independent Variable

Y = Binary Random Variable

$$P\{Y = 1|x\} = e^{\alpha+\beta x}/1 + e^{\alpha+\beta x} = \theta$$

$$\frac{\theta}{1-\theta} = e^{\alpha+\beta x}$$

$$\boxed{\text{logit } \theta = \log \frac{\theta}{1-\theta} = \alpha + \beta x}$$

Slide 34

Nearly all inference problems test

$$H_0 : \beta = 0 \ \textit{vs.} \ H_1 : \beta \neq 0$$

The parameter α is a “nuisance parameter”

Settings

2 population problem

Several 2×2 Tables

Does $\{Y_i\}$ depend on order

Are $\{Y_i\}$ independent

Independence of Two Binary R. V.

Slide 35

4. Sampling Theory of Logistic Regression

Observations: $(x_i, Y_i) \quad i = 1, 2, \cdots, n$

$$\begin{aligned} \lambda_i &= \log \tfrac{\theta_i}{1-\theta_i} = \alpha + \beta x_i \\ \theta_i &= e^{\lambda_i}/1 + e^{\lambda_i} \quad 1 - \theta_i = 1/1 + e^{\lambda_i} \end{aligned}$$

Joint Distribution:

$$f(y_1, \cdots, y_n) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

$$\begin{aligned} f(y_1, \cdots, y_n) &= \prod_{i=1}^{n} \left\{ \left(\frac{e^{\lambda_i}}{1+e^{\lambda_i}} \right)^{y_i} \left(\frac{1}{1+e^{\lambda_i}} \right)^{1-y_i} \right\} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\lambda_i y_i}}{1+e^{\lambda_i}} \right) = \frac{\exp \Sigma_{i=1}^{n} \lambda_i y_i}{\prod_i (1+e^{\lambda_i})} \end{aligned}$$

$$\sum_i \lambda_i y_i = \sum_i (\alpha + \beta x_i) y_i = \alpha \sum_i y_i + \beta \sum_i x_i y_i$$

Slide 36

Define $t_0 = \sum_i y_i, \quad t_1 = \sum_i x_i y_i$

$$\therefore \boxed{f(y_1, \cdots, y_n) = e^{\alpha t_0 + \beta t_1} \Big/ \prod_i (1 + e^{\lambda_i})}$$

(t_0, t_1) are sufficient statistics for (α, β)

Slide 37

$$\boxed{f(y_1, \cdots, y_n) = e^{\alpha t_0 + \beta t_1} \Big/ \prod_i (1 + e^{\lambda_i})}$$

$$t_0 = \sum_i y_i, \quad t_1 = \sum_i x_i y_i, \quad \lambda_i = \alpha + \beta x_i$$

Since t_0 and t_1 are sufficient statistics, it is only necessary to consider the distribution of

$$T_0 = \sum_i Y_i \quad , T_1 = \sum_i x_i Y_i \quad ;$$

i.e., $$P\{T_0 = t_0, T_1 = t_1\}$$

$$= \sum_{\substack{y_1 + \cdots + y_n = t_0 \\ x_1 y_1 + \cdots + x_n y_n = t_1}} f(y_1, \cdots, y_n) \quad = \frac{e^{\alpha t_0 + \beta t_1}}{\prod_i (1 + e^{\lambda_i})} \sum_{\substack{y_1 + \cdots + y_n = t_0 \\ x_1 y_1 + \cdots + x_n y_n = t_1}} (1)$$

Slide 38

$$C(t_0,t_1) = \sum_{\substack{\Sigma y_i = t_0 \\ \Sigma x_i y_i = t_1}} (1) = \text{No. of ways of arranging } y_1, \cdots, y_n \text{ which satisfy the two conditions.}$$

Example:

$n = 4$, $\Sigma y_i = 3$, $\Sigma x_i y_i = 2$,

$x_1 = 0$, $x_2 = 0$, $x_3 = 1$, $x_4 = 1$

$$x_i = \begin{array}{cccc} \underline{0} & \underline{0} & \underline{1} & \underline{1} \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{array} \left.\vphantom{\begin{array}{c}0\\0\\0\end{array}}\right\} \text{only 2 ways}$$

$C(t_0 = 3, t_1 = 2) = 2$

Slide 39

$$f(t_0,t_1) = P\{T_0 = t_0, T_1 = t_1\} = \boxed{\frac{C(t_0,t_1)e^{\alpha t_0 + \beta t_1}}{\prod_{i=1}^{n}(1+e^{\lambda_i})}}$$

$t_0 = \sum_i y_i, \quad t_1 = \sum_i x_i y_i$

$C(t_0,t_1) =$ number of ways of arranging $(y_1, \cdots, y_n)$ so that

$$\sum_i y_i = t_0, \quad \sum_i x_i y_i = t_1$$

Slide 40

5. Conditional Distribution of T_1

Given $T_0 = t_0$

Consider

$$f(t_1|t_0) = P\{T_1 = t_1 | T_0 = t_0\} = \frac{f(t_0,t_1)}{f(t_0)}$$

Since

$$f(t_0) = \sum_{t_1} f(t_0,t_1) = \frac{e^{\alpha t_0} \sum_{t_1} C(t_0,t_1) e^{\beta t_1}}{\prod_i (1+e^{\lambda_i})}$$

$$f(t_1|t_0) = \frac{e^{\alpha t_0} C(t_0,t_1) e^{\beta t_1} \Big/ \prod_i (1+e^{\lambda_i})}{e^{\alpha t_0} \sum_{t_1} C(t_0,t_1) e^{\beta t_1} \Big/ \prod_i (1+e^{\lambda_i})}$$

$$\boxed{f(t_1|t_0) = C(t_0,t_1) e^{\beta t_1} \Big/ \sum_z C(t_0,z) e^{\beta z}}$$

Slide 41

$$\boxed{f(t_1|t_0) = \frac{C(t_0,t_1) e^{\beta t_1}}{\sum_z C(t_0,z) e^{\beta z}}}$$

Note that α has been eliminated

Distribution only depends on $C(t_0,t_1)$ and (β, t_0, t_1).

$t_0 = \sum_i y_i$, $t_1 = \sum_i x_i y_i$

Suppose $H_0 : \beta = 0$ is true

$$\boxed{f_0(t_1|t_0) = C(t_0,t_1) \Big/ \sum_z C(t_0,z)}$$

is a distribution which is parameter free.

All that is necessary is to evaluate $C(t_0,t_1)$ and then distribution is completely known.

Slide 42

Example

Learning Situation

A person is doing a repetitive task. Does the person's success rate tend to increase with experience?

Order

<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>
0	1	0	1	1	1

$$y = \begin{cases} 1 & \text{if } S \\ 0 & \text{if } F \end{cases}, \quad \lambda_i = \alpha + \beta i \quad (x_i = i)$$

$\Sigma_i \, y_i = 4, \; \Sigma_i \, x_i y_i = \Sigma_i \, i y_i = 2\cdot1+4\cdot1+5\cdot1+6\cdot1 = 17$

Slide 43

To ease calculations define

$y_i' = 1 - y_i, t_0' = \Sigma_i y_i' = 2, \quad t_1' = \Sigma x_i y_i' = 4$

Note

$$\textstyle\sum_i i y_i + \Sigma i(1-y_i) = \sum_1^6 i = 21 = t_1 + t_1'$$

$$\sum_i y_i + \sum_i (1-y_i) = t_0 + t_0' = n = 6$$

$\therefore f(t_1'|t_0) = f(t_1 - 21|6 - t_0') = f(t_1'|t_0')$

Slide 44

Order

	1	2	3	4	5	6	
y_i:	0	1	0	1	1	1	$t_0 = 4,\ t_1 = 17$
y_i'	1	0	1	0	0	0	$t_0' = 2,\ t_1' = 4$

Set	t_1'	t_1'	$C(t_0', t_1')$	$t_1 = 21 - t_1'$	$f(t_1 \mid t_0)$
1,2	3	3	1	18	1/15
1,3	4	4	1	17	1/15
1,4	5	5	2	16	2/15
1,5	6	6	2	15	2/15
1,6	7	7	3	14	3/15
2,3	5	8	2	13	2/15
2,4	6	9	2	12	2/15
2,5	7	10	1	11	1/15
2,6	8	11	1	10	1/15
3,4	7		15		
3,5	8				
3,6	9				
4,5	9				
4,6	10				
5,6	11				

$$f(t_1 \mid t_0) = \frac{C(t_0, t_1)}{\sum_{t_1} C(t_0, t_1)}$$

Slide 45

Large values (or small values) of $t_1 = \sum_i iy$ are evidence of a trend.

$P\{T_1 = 17 \mid T_0 = 4\} = 1/15$,

Critical Region: $T_1 = 17, 18, 11, 10$

$P\{T_1 \geq 17 \mid T_0 = 4\} + P\{T_1 \leq 11 \mid T_0 = 4\} = \frac{2}{15} + \frac{2}{15} = \frac{4}{15}$

Slide 46

$$\theta = P\{Y=1\} = e^{\alpha+\beta x} \big/ 1 + e^{\alpha+\beta x}$$

If $x > 0, \beta > 0 \Longrightarrow \theta \uparrow$ as $\beta x \uparrow$

If $x > 0, \beta < 0 \Longrightarrow \theta \downarrow$ as $\beta x \downarrow$

e.g.,

$x_i = 1, \beta > 0$: increasing trend of successes.

$x_i = 1, \beta < 0$: decreasing trend of successes.

Example:

<u>Learning</u>

$H_0 : \beta = 0$ *vs.* $H_1 : \beta > 0$

$t_1 = 17$, $t_0 = 4$

$$\begin{aligned} P\{T_1 \geq 17 | T_0 = 4\} &= P\{T_1 = 17 | t_0 = 4\} \\ &+ P\{T_1 = 18 | t_0 = 4\} \\ &= = \tfrac{1}{15} + \tfrac{1}{15} = \tfrac{2}{15} \end{aligned}$$

Slide 47

6. Comparing Two Binomial Distributions

Consider two groups referred to as group "0" and group "1". Let there be n_0 observations in group 0 and n_1 observations in group 1.

$$\begin{aligned} \theta_i &= \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \\ x_i &= 0 \text{ for } i = 1, 2, \cdots, n_0 \\ &\quad \theta_i = e^{\alpha} \big/ 1 + e^{\alpha} \\ x_i &= 1 \text{ for } i = n_0 + 1, \cdots, n_0 + n_1 \\ &\quad \theta_i = e^{\alpha+\beta} \big/ 1 + e^{\alpha+\beta} \\ t_0 &= \sum_{i=1}^{n_0+n_1} y_i = \sum_0 y_i + \sum_1 y_i = s_0 + s_1 \\ t_1 &= \sum_{i=1}^{n_0+n_1} x_i y_i = \sum_1 y_i = s_1 \end{aligned}$$

Slide 48

Group	Number of Successes	Number of Failures	Totals
1	s_1	$n_1 - s_1$	n_1
0	s_0	$n_0 - s_0$	n_0
	t_0	$n - t_0$	$n_0 + n_1 = n$

Logistic Regression Theory says condition on t_0

($t_0 = s_0 + s_1 =$ Total number of successes).

Since the sample sizes (n_0, n_1) are fixed and t_0 is fixed, $n - t_0$ is fixed. Hence all marginal totals are fixed. There is only one free entry in the 2×2 table; once s_1 is assigned, the remaining three entries can be calculated.

We know (binomial distribution)

$$\begin{aligned} f(s_1) &= \binom{n_1}{s_1}\theta_1^{s_1}(1-\theta_1)^{n_1-s_1} \\ f(s_0) &= \binom{n_0}{s_0}\theta_0^{s_0}(1-\theta_0)^{n_0-s_0} \end{aligned}$$

Slide 49

$$f(s_i) = \binom{n_i}{s_i}\theta_i^{s_i}(1-\theta_i)^{n_i-s_i} \quad i = 0, 1$$

We wish to find

$$f(s_1|t_0 = s_0 + s_1) = \frac{f(s_0, s_1)}{f(t_0)} = \frac{f(s_0)f(s_1)}{f(t_0)}$$

$$f(t_0) = \sum_{s_0+s_1=t_0} f(s_0, s_1) = \sum_{s_1=1}^{t_0} f(t_0 - s_1)f(s_1)$$

Consider

$$f(s_0)f(s_1) = f(t_0 - s_1)f(s_1)$$

$$= \binom{n_0}{t_0-s_1}\theta_0^{t_0-s_1}(1-\theta_0)^{n_0-t_0+s_1}$$
$$\times \binom{n_1}{s_1}\theta_1^{s_1}(1-\theta_1)^{n_1-s_1}$$
$$= \binom{n_0}{t_0-s_1}\binom{n_1}{s_1}\left[\frac{\theta_1/1-\theta_1}{\theta_0/1-\theta_0}\right]^{s_1}\left[\frac{\theta_0}{1-\theta_0}\right]^{t_0}(1-\theta_0)^{n_0}(1-\theta_1)^{n_1}$$

$$\boxed{f(s_0)f(s_1) = C(t_0, t_1)e^{\beta s_1}g(t_0), \quad t_0 = s_0 + s_1}$$

Slide 50

$$f(s_1|t_0 = s_0 + s_1) = \frac{f(s_0,s_1)}{f(t_0)} = \frac{f(s_0)f(s_1)}{f(t_0)} = \frac{f(t_0-s_1)f(s_1)}{f(t_0)}$$

where

$$f(t_0) = \sum_{s_1} f(t_0 - s_1)f(s_1).$$

Since $f(t_0 - s_0)f(s_1) = C(t_0, s_1)e^{\beta s_1}g(t_0)$

$$f(s_1|t_0) = \frac{C(t_0, s_1)e^{\beta s_1}g(t_0)}{\sum_{s_1} C(t_0, s_1)e^{\beta s_1}g(t_0)} = C(t_0, s_1)e^{\beta s_1} \Big/ \sum_{s_1} C(t_0, s_1)e^{\beta s_1}$$

where

$$C(t_0, s_1) = \binom{n_0}{t_0 - s_1}\binom{n_1}{s_1}, \beta = \log\left[\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right]$$

$$\boxed{f(s_1|t_0 = s_0 + s_1) = \frac{\binom{n_0}{t_0 - s_1}\binom{n_1}{s_1}e^{\beta s_1}}{\sum_{s_1}\binom{n_0}{t_0 - s_1}\binom{n_1}{s_1}e^{\beta s_1}}}$$

Slide 51

$$f(s_1|t_0) = \frac{C(t_0, s_1)e^{\beta s_1}}{\sum_{s_1} C(t_0, s_1)e^{\beta s_1}}$$

Suppose $\beta = 0 \Longleftrightarrow \theta_0 = \theta_1$

$$f_0(s_1|t_0 = s_0 + s_1) = \frac{\binom{n_0}{t_0 - s_1}\binom{n_1}{s_1}}{\sum_{s_1}\binom{n_0}{t_0 - s_1}\binom{n_1}{s_1}}$$

$$\boxed{f(s_1, t_0) = \binom{n_0}{t_0 - s_1}\binom{n_1}{s_1} \Big/ \binom{n_0 + n_1}{t_0}}$$

The above distribution is the Hypergeometric distribution.

Slide 52

7. Urn Sampling Model.

$n = n_0 + n_1$ = number of balls in urn

t_0 balls are red and $n - t_0$ are white

n_1 balls are randomly drawn without replacement

s_1 of these balls are red.

	Red	White	Total
Sample	s_1	$n_1 - s_1$	n_1
Remaining	X	X	X
	t_0	$n - t_0$	n

(X values are calculated)

Slide 53

Urn Sampling Model (Two-Sample Problem)

n=no of balls	• ○ •		Sample n_1
t_0=no. red balls	○ • •	→	balls without
$n - t_0$=no. white balls	○ •		replacement

	Red	White	Totals
Sample	s_1	$n_1 - s_1$	n_1
Remaining in Urn	s_0	$n_0 - s_0$	n_0
	t_0	$n - t_0$	n

s_1= no. of red balls

$n_1 - s_1$ = no. of white balls

Slide 54

$\binom{n_1}{s_1}$ = Number of ways of drawing s_1 red balls and $n_1 - s$ white balls

$\binom{n_0}{s_0}$ = Number of ways of drawing s_0 red balls and $n_0 - s_0$ white balls

$\binom{n_0}{s_0}\binom{n_1}{s_1}$ = Number of ways of drawing the two samples

$\Rightarrow \binom{n_0+n_1}{s_0+s_1}$ = Number of ways of drawing (s_0+s_1) red balls from among (n_0+n_1) balls

Hypergeometric Distribution

$$f(s_1|t_0 = s_0+s_1) = \binom{n_0}{s_0}\binom{n_1}{s_1}\Big/\binom{n_0+n_1}{s_0+s_1}$$

$$= \binom{n-n_1}{t_0-s_1}\binom{n_1}{s_1}\Big/\binom{n}{t_0}$$

Slide 55

Example:

$$f(s_1|t_0 = s_1+s_0) = \binom{n_1}{s_1}\binom{n_0}{s_0}\Big/\binom{n}{t_0}$$

Group	S	F	
1	$4 = s_1$	1	$5 = n_1$
0	4	3	$7 = n_0$
	$t_0 = 8$	4	$12 = n$

$$f(s_1|t_0 = 8) = \frac{\binom{n_1}{s_1}\binom{n_0}{s_0}}{\binom{n}{t_0}}$$

$$f(4|t_0 = 8) = \frac{\binom{5}{4}\binom{7}{4}}{\binom{12}{8}} = \frac{175}{495}$$

Slide 56

Other Tables with Marginal Totals Fixed

5	0	5
3	4	7
8	4	12

$$f(5|8) = \frac{\binom{5}{5}\binom{7}{3}}{\binom{12}{8}} = \frac{35}{495}$$

1	4	5
7	0	7
8	4	12

$$f(1|8) = \frac{\binom{5}{1}\binom{7}{7}}{\binom{12}{8}} = \frac{5}{495}$$

2	3	5
6	1	7
8	4	12

$$f(2|8) = \frac{\binom{5}{2}\binom{7}{6}}{\binom{12}{8}} = \frac{70}{495}$$

3	2	5
5	2	7
8	4	12

$$f(3|8) = \frac{\binom{5}{3}\binom{7}{5}}{\binom{12}{8}} = \frac{210}{495}$$

Slide 57

Summary:

	s_1	$f(s_1 \vert t_0 = 8)$
	1	$5/495 = .010$
	2	$70/495 = .141$
	3	$210/495 = .424$
Observed Table →	4	$175/495 = .354$
	5	$35/495 = .071$

To carry out a 2-sided test we calculate the probability associated with table having a lower probability of occurring.

$$P = \frac{175 + 35 + 70 + 5}{495} = \frac{285}{495} = .576$$

To carry out a 1-sided test $H_1 : \beta > 0$

$$P = \frac{175 + 35}{495} = \frac{210}{495} = .425$$

Slide 58

Lecture 3.
Independence and Urn Sampling

8. Test for Independence
9. Urn Sampling Model
10. Wilcoxon Rank Sum Test

Slide 59

8. Tests for Independence

Example: Leukemia patients recovering from bone marrow transplant graft *vs.* Host disease.

Donor for marrow may be matched or mismatched with respect to person receiving transplant (MHC Status).

Severity of GVHD Toxicity

MHC	Minor	Major	Totals
Mismatched	11	7	18
Matched	15	4	19
	26	11	37

Is severity of GVHD independent of whether MHC is matched?

Slide 60

Theory: Consider two random variables, Y_1, Y_2, having the joint distribution $P(Y_1, Y_1)$; i.e.

$$P\{Y_1 = y_1, Y_2 = y_2\} = P(y_1, y_2).$$

If (Y_1, Y_2) are independent, then

$$P(Y_1, Y_2) = P(y_1)P(y_2) = P(Y_1 = y_1)P(Y_2 = y_2)$$

Conversely, if $P(y_1, y_2) = P(y_1)P(y_2)$ then the two random variables are independent.

Our case: Y_1, Y_2 are binary

Y_1: Severity of toxicity

Y_2: MHC Status

Slide 61

Test for Independence

Consider Y_1, Y_2. Both binary random variables.

Let

$P\{Y_1 = 0\} = 1/1 + e^{\alpha_1}, P\{Y_1 = 1\} = e^{\alpha_1}/1 + e^{\alpha_1}$.

(1) or $\boxed{f(y_1) = P\{Y_1 = y_1\} = e^{\alpha_1 y_1}/1 + e^{\alpha_1}}$

Let $P\{Y_2 = 0|Y_1 = 0\} = 1/1 + e^{\alpha_2}$,

$P\{Y_2 = 1|Y_1 = 0\} = e^{\alpha_2}/1 + e^{\alpha_2}$

(2) $\boxed{f(y_2|y_1 = 0) = e^{\alpha_2 y_2}/1 + e^{\alpha_2}}$

$$P\{Y_2 = 0|Y_1 = 1\} = 1/1 + e^{\alpha_2+\beta},$$

$$P\{Y_2 = 1|Y_1 = 1\} = \frac{e^{\alpha_2+\beta}}{1 + e^{\alpha_2+\beta}}$$

Slide 62

(2) $\boxed{f(y_2|y_1=0) = e^{\alpha_2 y_2} \big/ 1 + e^{\alpha_2}}$

(3) $\boxed{f(y_2|y_1=1) = e^{(\alpha_2+\beta) y_2} \big/ 1 + e^{\alpha_2+\beta}}$

(2) and (3) may be written

$$\boxed{f(y_2|y_1) = e^{(\alpha_2+\beta y_1) y_2} \big/ 1 + e^{\alpha_2+\beta y_1}}$$

<u>Note:</u>

$$\frac{f(y_2=1|y_1)}{f(y_2=0|y_1)} = \frac{e^{\alpha_2+\beta y_1}}{1} = e^{\alpha_2+\beta y_1}$$

Slide 63

$$\boxed{\log \tfrac{f(y_2=1|y_1)}{f(y_2=0|y_1)} = \alpha_2 + \beta y_1}$$

$$\begin{aligned} f(y_1) &= e^{\alpha_1 y_1} \big/ 1 + e^{\alpha_1} \\ f(y_2|y_1) &= e^{(\alpha_2+\beta y_1) y_2} \big/ 1 + e^{\alpha_2+\beta y_1} \end{aligned}$$

Parameters are: $(\alpha_1, \alpha_2, \beta)$

$$\begin{aligned} f(y_1, y_2) &= f(y_2|y_1)(f(y_1) = f(y_1) f(y_2|y_1) \\ &= \frac{e^{\alpha_1 y_1} \cdot e^{(\alpha_2+\beta y_1) y_2}}{(1+e^{\alpha_1})\left(1+e^{\alpha_2+\beta y_1}\right)} \end{aligned}$$

$$\boxed{f(y_1, y_2) = \tfrac{e^{\alpha_1 y_1 + \alpha_2 y_2 + \beta y_1 y_2}}{(1+e^{\alpha_1})\left(1+e^{\alpha_2+\beta y_1}\right)}}$$

Outcomes are: $(y_1, y_2) = (0,0), (1,0), (0,1), 1, 1)$

Multinomial (4 categories)

$$\boxed{f(y_1, y_2) = \theta_{00}^{(1-y_1)(1-y_2)} \theta_{10}^{y_1(1-y_2)} \theta_{01}^{(1-y_1) y_2} \theta_{11}^{y_1 y_2}}$$

$\theta_{00} + \theta_{10} + \theta_{01} + \theta_{11} = 1$ (3 parameters)

Slide 64

$$\boxed{f(y_1, y_2) = \frac{e^{\alpha_1 y_1 + \alpha_2 y_2 + \beta y_1 y_2}}{(1 + e^{\alpha_1})\left(1 + e^{\alpha_2 + \beta y_1}\right)}}$$

Note if $\beta = 0$,

$$\begin{aligned} f(y_1, y_2) &= e^{\alpha_1 y_1 + \alpha_2 y_2} / (1 + e^{\alpha_1})(1 + e^{\alpha_2}) \\ &= \left(\frac{e^{\alpha_1 y_1}}{1 + e^{\alpha_1}}\right)\left(\frac{e^{\alpha_2 y_2}}{1 + e^{\alpha_2}}\right) \Rightarrow \text{ independence} \end{aligned}$$

Hence $H_0 : \beta = 0$ corresponds to test of independence.

(α_1, α_2) are nuisance parameters.

Slide 65

Consider a sample of n (Y_{1j}, Y_{2j}) $j = 1, 2, \cdots, n$

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2) = \prod_{j=1}^{n} f(y_{1j}, y_{2j})$$

$$\boxed{f(\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{e^{\alpha_1 \sum_j y_{1j} + \alpha_2 \sum_j y_{2j} + \beta \sum_j y_{1j} y_{2j}}}{(1 + e^{\alpha_1})^n \prod_{j=1}^{n} \left(1 + e^{\alpha_2 + \beta y_{1j}}\right)}}$$

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{e^{\alpha_1 \sum_j y_{1j} + \alpha_2 \sum_j y_{2j} + \beta \sum_j y_{1j} y_{2j}}}{(1 + e^{\alpha_1})^n \prod_{j=1}^{n} \left(1 + e^{\alpha_2 + \beta y_{1j}}\right)}$$

Let $t_1 = \Sigma_j y_{1j}, t_2 = \Sigma_j y_{2j}, t_3 = \Sigma_j y_{i1} y_{2j}$

Slide 66

Note:

$$\prod_{j=1}^{n} \left(1 + e^{\alpha_2 + \beta y_{1j}}\right) = \left(1 + e^{\alpha_2}\right)^{\Sigma_j (1 - y_{1j})} \left(1 + e^{\alpha_2 + \beta}\right)^{\Sigma_j y_{1j}}$$

$$= \left(1 + e^{\alpha_2}\right)^{n - t_1} \left(1 + e^{\alpha_2 + \beta}\right)^{t_1}$$

$$\therefore \boxed{f(y_1, y_2) = \frac{e^{\alpha_1 t_1 + \alpha_2 t_2 + \beta t_3}}{(1+e^{\alpha_2})^{n-t_1}\left(1+e^{\alpha_2+\beta}\right)^{t_1}}}$$

(t_1, t_2, t_3) are sufficient statistics.

$$\boxed{f(t_1, t_2, t_3) = \frac{C(t_1,t_2,t_3) e^{\alpha_1 t_1 + \alpha_2 t_2 + \beta t_3}}{D}}$$

$C(t_1, t_2, t_3) =$ Number of ways of permuting (y_{1j}, y_{2j}) such that t_1, t_2, t_3 are fixed.

Slide 67

$$f(t_1, t_2, t_3) = \frac{C(t_1, t_2, t_3) e^{\alpha_1 t_1 + \alpha_2 t_2 + \beta t_3}}{D}$$

$$D = (1 + e^{\alpha_2})^{n - t_1} (1 + e^{\alpha_2 + \beta})^{t_1}$$

Since (α_1, α_2) are nuisance parameters, consider

$$f(t_3 | t_1, t_2) = \frac{C(\boldsymbol{t}) e^{\alpha_1 t_1 + \alpha_2 t_2 + \beta t_3}}{\sum_{t_3} C(\boldsymbol{t}) e^{\alpha_1 t_1 + \alpha_2 t_2 + \beta t_3}}$$

$$\boxed{f(t_3 | t_1, t_2) = C(\boldsymbol{t}) e^{\beta t_3} \Big/ \sum_{t_3} C(\boldsymbol{t}) e^{\beta t_3}}$$

and if $\beta = 0$

$$\boxed{f_0(t_3 | t_1, t_2) = C(\boldsymbol{t}) \Big/ \sum_{t_3} C(\boldsymbol{t})}$$

Slide 68

Data Display

		Factor 2: 1	Factor 2: 0	Totals
Factor 1	1	s_{11}	s_{10}	t_1
	0	s_{01}	s_{00}	$n - t_1$
Totals:		t_2	$n - t_2$	n

Slide 69

$$
\begin{aligned}
s_{00} &= \text{No. of observations for which } y_{1j} = y_{2j} = 0 \\
s_{00} &= \sum_{j=1}^{n} (1 - y_{1j})(1 - y_{2j}) \\
s_{01} &= \sum_{j=1}^{n} (1 - y_{1j}) y_{2j} = \text{no. of observations for which } y_{1j} = 0, y_{2j} = 1 \\
s_{10} &= \sum_{j} y_{1j}(1 - y_{2j}) \\
s_{11} &= \sum_{j} y_{1j} y_{2j} \\
s_{10} + s_{11} &= \Sigma y_{1j}(1 - y_{2j}) + \sum_{j} y_{1j} y_{2j} \\
&= \sum_{j} y_{1j}(1 - y_{2j} + y_{2j}) = \Sigma y_{1j} = t_1
\end{aligned}
$$

Slide 70

		Factor 2		
		1	0	
Factor 1	1	s_{11}	s_{10}	t_1
	0	s_{01}	s_{00}	$n-t_1$
		t_2	$n-t_2$	n

But this is a 2×2 table with all margins fixed.

Slide 71

In comparing two binomials we had

	S	F	
A	s_a	$n_a - s_a$	n_a
B	s_b	$n_a - n_b$	n_b
	t	$n-t$	n

s_{11}	s_{10}	t_1
s_{01}	s_{00}	$n-t_1$
t_2	$n-t_2$	n

$$f(s_a|t) = \frac{\binom{n_a}{s_a}\binom{n_b}{s_b}e^{\beta s_a}}{\sum_{s_a}\binom{n_a}{s_a}\binom{n-n_a}{t-s_a}e^{\beta s_a}}$$

or

$$\boxed{f(s_a|t) = \binom{n_a}{s_a}\binom{n-n_a}{t-s_a}e^{\beta s_a} \Big/ D}$$

Slide 72

$$f(s_a|t) = \binom{n_a}{s_a}\binom{n-n_a}{t-s_a} e^{\beta s_a} \Big/ D$$

and if $\beta = 0$

$$f_0(s_a|t) = \binom{n_a}{s_a}\binom{n-n_a}{t-s_a} \Big/ \binom{n}{t}$$

Slide 73

Comparison of Notation

Proportions	Independence
s_a	s_{11}
n_a	t_1
t	t_2
n	n

$$\boxed{f_0(s_{11}|t_1, t_2) = \binom{t_1}{s_{11}}\binom{n-t_1}{t_2-s_{11}} \Big/ \binom{n}{t_2}}$$

Slide 74

What is β?

Recall for $n = 1$ (Multinomial distribution)

$$f(y_1, y_2) = \theta_{00}^{(1-y_1)(1-y_2)} \theta_{10}^{y_1(1-y_2)} \theta_{01}^{1-y_1)y_2} \theta_{11}^{y_1 y_2}$$

$$= \theta_{00} \left(\frac{\theta_{10}}{\theta_{00}}\right)^{y_1} \left(\frac{\theta_{01}}{\theta_{00}}\right)^{y_2} \left(\frac{\theta_{11}\theta_{00}}{\theta_{10}\theta_{01}}\right)^{y_1 y_2}$$

But

$$f(y_1, y_2) = \frac{e^{\alpha_1 y_1 + \alpha_2 y_2 + \beta y_1 y_2}}{(1 + e^{\alpha_1})(1 + e^{\alpha_2})^{1-y_1}(1 + e^{\alpha_2+\beta})^{y_1}}$$

$$= \frac{1}{(1 + e^{\alpha_1})(1 + e^{\alpha_2})} \cdot \left[\frac{(1 + e^{\alpha_2})e^{\alpha_1}}{1 + e^{\alpha_2+\beta}}\right]^{y_1} e^{\alpha_2 y_2} e^{\beta y_1 y_2}$$

Slide 75

$$\therefore \theta_{00} = 1/(1 + e^{\alpha_1})(1 + e^{\alpha_2}), \quad \frac{\theta_{01}}{\theta_{00}} = e^{\alpha_2}$$

$$\frac{\theta_{10}}{\theta_{00}} = \frac{e^{\alpha_1}(1+e^{\alpha_2})}{1+e^{\alpha_2+\beta}}, \quad \frac{\theta_{11}\theta_{00}}{\theta_{10}\theta_{01}} = e^{\beta}$$

Suppose factors are independent $\Rightarrow \theta_{ij} = a_i b_j$

$$\therefore e^{\beta} = \frac{\theta_{11}\theta_{00}}{\theta_{10}\theta_{01}} = \frac{a_1 b_1 a_0 b_0}{a_1 b_0 a_0 b_1} = 1$$

$\Rightarrow \beta = 0$

Slide 76

MHC	Severity of GVHD: Minor	Major	
Mismatched	11	7	18
Matched	15	4	19
	26	11	37

$P = 0.295$

Severity of GVHD

	None	Mild	Moderate	Severe	Extreme	
Mismatch	4	4	3	3	4	18
Matched	6	8	1	4	0	19
	10	12	4	7	4	37

$P = .09$

Slide 77

First Protocol

	None	Mild	Moderate	Severe	Extreme	
Mismatch	2	2	2	1	1	8
Matched	3	4	1	2	0	10
	5	6	3	3	1	18

Second Protocol

	None	Mild	Moderate	Severe	Extreme	
Mismatch	2	2	1	1	3	10
Matched	3	4	0	2	0	9
	5	6	1	4	3	19

Slide 78

Review

$Y_1, Y_2, \cdots, Y_n$ independent random variables with

$$\theta_i = P\{Y_i = 1\}, \quad 1 - \theta_i = P\{Y_i = 0\}$$

Model: $\theta_i = e^{\alpha+\beta x_i} / 1 + e^{\alpha+\beta x_i}$

where (α, β) are unknown parameters and $\{x_i\}$ is a covariate.

$$\boxed{\text{logit } \theta_i = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta x_i}$$

The joint distribution of $y_1, \cdots, y_n$ is:

$$f(y_1, \cdots, y_n) = e^{\alpha t_0 + \beta t_1} \Big/ \sum_{i=1}^{n} (1 + e^{\alpha+\beta x_i})$$

$$t_0 = \sum_{i=1}^{n} y_i \quad , \quad t_1 = \sum_{i=1}^{n} x_i y_i$$

(t_0, t_1) are sufficient statistics for (α, β)

Slide 79

$$f(y_1, \cdots, y_n) = e^{\alpha t_0 + \beta t_1} \Big/ \prod_{1}^{n} (1 + e^{\alpha+\beta x_i})$$

$t_0 = \Sigma y_i \quad , \quad t_1 = \sum_i x_i y_i$

Hence the distribution of $T_0 = \sum_i Y_i, T_1 = \sum_i x_i Y_i$ is

$$\boxed{f(t_0, t_1) = P\{T_0 = t_0, T_1 = t_1\} = \frac{C(t_0, t_1) e^{\alpha t_0 + \beta t_1}}{\prod_{i=1}^{n} (1 + e^{\alpha+\beta x_i})}}$$

$C(t_0, t_1)$ = number of ways of arranging the binary variables $(y_1, y_2, \cdots, y_n)$ such that

$$\sum_{i=1}^{n} y_i = t_0, \quad \sum_{1}^{n} x_i y_i = t_1$$

In nearly all testing situations we are concerned with $H_0 : \beta = 0$ *vs.* a two-sided alternative $H_1 : \beta \neq 0$ or a one-slded alternative $H_1 : \beta > 0$ or $H_1 : \beta < 0$

Slide 80

$$f(t_0, t_1) = C(t_0, t_1) e^{\alpha t_0 + \beta t_1} \Big/ \prod_{i=1}^{n} \left(1 + e^{\alpha + \beta x_i}\right)$$

$$t_0 = \sum_i y_i, \quad t_1 = \sum_i x_i y_i$$

The conditional distribution of $T_1 = \sum_i x_i Y_i$ conditional on $T_0 = t_0$ results in the elimination of the nuisance parameter α; i.e.,

$$f(t_1|t_0) = P\{T_1 = t_1 | T_0 = t_0\} = \frac{C(t_0, t_1) e^{\beta t_1}}{\sum_{t_1} C(t_0, t_1) e^{\beta t_1}}$$

Under $H_0 : \beta = 0$ and $f(t_1|t_0)$ becomes

$$f_0(t_1|t_0) = C(t_0, t_1) \Big/ \sum_{t_1} C(t_0, t_1)$$

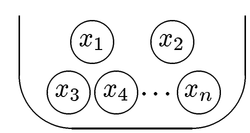
Note $f_0(t_1|t_0)$ is parameter-free and serves as the basis of "exact tests."

Slide 81

9. General Urn Sampling Model

The asymptotic distribution of $T_1 = t_1$ conditional on $T_0 = t_0$ can be calculated using an urn sampling model. This serves as a convenient approximation to all testing procedures.

Urn Model:



Slide 82

Notation: Urn Sampling

$$n = \text{number of balls in urn}$$
$$x_i = \text{value of } i^{th} \text{ ball}$$
$$Y_i = \begin{cases} 1 & \text{if } i^{th} \text{ drawn ball in sample} \\ 0 & \text{otherwise} \end{cases}$$
$$t_0 = \sum_1^n y_i = \text{ size of sample}$$
$$t_1 = \sum_{\imath=1}^n x_i y_i = \text{ value of sample}$$

Since sampling is random, each ball has same probability of being drawn

If

$$\theta = P\{Y_i = 1\}, \quad E(Y_i) = \theta$$
$$E(T_0) = t_0 = \sum_1^n E(Y_i) = n\theta$$
$$\Rightarrow \theta = \frac{t_0}{n} = \frac{\text{size of sample}}{\text{Total number of balls}}$$
$$\boxed{E(Y_i) = t_0/n}$$
$$V(Y_i) = \theta(1-\theta) = \frac{t_0}{n}\left(1 - \frac{t_0}{n}\right)$$

Slide 83

What are $\text{Cov}(Y_i, Y_j)$ for $i \neq j$?

$$V(T_0) = 0 = V\left(\sum_1^n Y_i\right) = nV(Y_i) + 2\tbinom{n}{2}\text{Cov}$$
$$0 = n\tfrac{t_0}{n}\left(1 - \tfrac{t_0}{n}\right) + 2\tfrac{n(n-1)}{2}\text{Cov}$$

where $\text{Cov} = \text{Cov}(Y_i, Y_j), i \neq j$

$$\boxed{\text{Cov} = -\tfrac{1}{n-1}\theta(1-\theta), \quad \theta = t_0/n}$$

Slide 84

$$
\begin{aligned}
T_0 &= \sum_{i=1}^{n} Y_i \quad , T_1 = \Sigma x_i Y_i \\
E(Y_i) &= \theta = t_0/n \\
V(Y_i) &= \theta(1-\theta) \\
\mathrm{Cov}(Y_i, Y_j) &= -\tfrac{1}{n-1}\theta(1-\theta) \text{ for } i \neq j \\
\therefore E(T_1|t_0) &= \Sigma x_i \tfrac{t_0}{n} = t_0 \overline{x} \\
V(T_1|t_0) &= \sum_{i=1}^{n} x_i^2 V(Y_i) + \sum_{i \neq j} x_i x_j \mathrm{Cov}(Y_i, Y_j) \\
&= \theta(1-\theta) \sum_{1}^{n} x_i^2 - \frac{\theta(1-\theta)}{n-1} \sum_{i \neq j} x_i x_j \\
&= \theta(1-\theta) \left\{ \sum_{i=1}^{n} x_i^2 - \sum_{i \neq j} \frac{x_i x_j}{n-1} \right\}
\end{aligned}
$$

Slide 85

Recall

$$
\begin{aligned}
\left(\sum_{1}^{n} x_i \right)^2 &= \sum_{i} x_i^2 + \sum_{i \neq j} x_i x_j \\
\sum_{i \neq j} x_i x_j &= (\textstyle\sum x_i)^2 - \sum_{1}^{n} x_i^2
\end{aligned}
$$

$$
\begin{aligned}
V(T_1|t_0) &= \theta(1-\theta) \left\{ \sum x_i^2 - \tfrac{1}{n-1} \left[\left(\sum_{1}^{n} x_i \right)^2 - \sum_{1}^{n} x_i^2 \right] \right\} \\
&= \theta(1-\theta) \left\{ \left(1 + \tfrac{1}{n-1} \right) \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n-1} \right\} \\
&= \theta(1-\theta) \left\{ \frac{n \Sigma x_i^2 - (\Sigma x_i)^2}{n-1} \right\} \\
&= \frac{\theta(1-\theta)n}{n-1} \left\{ \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \right\}
\end{aligned}
$$

Slide 86

$$V(T_1|t_0) = \frac{\theta(1-\theta)n}{n-1}\left\{\sum x_i^2 - \frac{(\Sigma x_i)^2}{n}\right\}$$

$$\boxed{V(T_1|t_0) = n\theta(1-\theta)s^2}$$

$$\theta = t_0/n, \quad s^2 = \frac{\sum_1^n x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \sum_1^n \frac{(x_i - \overline{x})^2}{n-1}$$

Thus

$$\boxed{\begin{aligned} E(T_1|t_0) &= t_0\overline{x} \\ V(T_1|t_0) &= t_0\left(1 - \tfrac{t_0}{n}\right)s^2 \end{aligned}}$$

$T_1 = \sum_1^n x_i y_i$ is asymptotically normal with above mean and variance.

Slide 87

Two Population Problem

$$x_i = \begin{cases} 0 & \text{for} \quad i = 1, 2, \cdots, n_0 \\ 1 & \text{for} \quad i = n_0 + 1, \cdots, n_0 + n_1 = n \end{cases}$$

$$\overline{x} = \frac{\sum_1^n x_i}{n} = \frac{n_1}{n}$$

$$\begin{aligned} s^2 &= \tfrac{1}{n-1}\left\{\sum_1^n x_i^2 - \frac{(\Sigma x_i)^2}{n}\right\} \\ &= \tfrac{1}{n-1}\left\{n_1 - \frac{n_1^2}{n}\right\} = \frac{n_1}{n-1}\left\{\left(1 - \frac{n_1}{n}\right)\right\} = \frac{n_1 n_0}{n(n-1)} \end{aligned}$$

Slide 88

Since

$$E(T_1|t_0) = t_0\overline{x} = t_0\frac{n_1}{n}, \quad t_0 = \sum_1^n y_i, \quad t_0 = s_0 + s_1$$

$$\begin{aligned} V(T_1|t_0) &= t_0\left(1 - \tfrac{t_0}{n}\right)s^2 \\ &= t_0\left(1 - \tfrac{t_0}{n}\right)\tfrac{n_1 n_0}{n(n-1)} \end{aligned}$$

$$T_1 = \Sigma_i x_i Y_i = S_1$$

$$\frac{S_1 - \frac{t_0 n_1}{n}}{\sqrt{t_0\left(1 - \frac{t_0}{n}\right)\frac{n_1 n_0}{n(n-1)}}} \sim N(0,1)$$

Slide 89

$$Z = \frac{S_1 - \frac{t_0 n_1}{n}}{\sqrt{t_0\left(1 - \frac{t_0}{n}\right)\frac{n_1 n_0}{n(n-1)}}} \quad \sim \quad N(0,1)$$

Population	Successes	Failures	
1	s_1	f_1	n_1
0	s_0	f_0	n_0
	t_0	$n - t_0$	n

$$z^2 = \chi^2 = \frac{(n-1)(s_1 f_0 - s_0 f_1)^2}{n_0 n_1 t_0 (n - t_0)}$$ chi-square with 1.d.f.

Slide 90

Two Sample
10. Wilcoxon Rank Sum Test

Population		
A	10, 18, 22, 36	$n_a = 4$
B	12, 35, 40, 45, 48	$n_b = 5$

Arrange data as ordered sample

	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>
	10	12	18	22	35	36	40	45	48
A	1	0	1	1	0	1	0	0	0

Consider the A Sample 0's and 1's.

Are 0's and 1's random or is there a trend; i.e.

$\lambda_i = \alpha + \beta x_i = \alpha + \beta i \quad (x_i = i)$

$$t_0 = \sum_1^n y_i = n_a = 4$$

$$t_1 = \sum_1^n x_i y_i = \Sigma i y_i \quad = \text{Sum of ranks of } A$$

$$= 1 + 3 + 4 + 6 = 14$$

Slide 91

$$t_0 = \sum y_i = n_a \quad , t_1 = \sum_1^n i y_i = \text{rank sum of } A \text{ sample}$$

$$E(T_1|t_0) = t_0 \overline{x} \quad , V(T_1|t_0) = t_0 \left(1 - \frac{t_0}{n}\right) s^2$$

$$\overline{x} = \sum_1^n \frac{x_i}{n} = \frac{\sum_1^n i}{n} = \frac{n(n+1)/2}{n} = \frac{n+1}{2}$$

$$s^2 = \frac{1}{n-1} \left\{ \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \right\}$$

$$\sum_1^n x_i^2 = \sum_1^n i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$s^2 = n(n+1)/12$$

Slide 92

$$n = n_a + n_b, \quad 1 - \frac{n_a}{n} = \frac{n_b}{n}$$

$$V(T_1|t_0) = n_a \left(1 - \frac{n_a}{n}\right) \frac{n(n+1)}{12}$$

$$\boxed{V(T_1|t_0) = \frac{n_a n_b}{12}(n+1)}$$

$$E(T_1|t_0) = t_0 \overline{x} = n_a \left(\frac{n+1}{2}\right)$$

Our example: $n_a = 4, n = 9, t_1 = 14$

$$E(T_1|t_0) = \tfrac{4 \cdot 10}{2} = 20$$
$$V(T_1|t_0) = \tfrac{4 \cdot 5 \cdot 10}{12} = 16.67$$

$$Z = \frac{T_1 - E(T_1|t_0)}{\sqrt{V(T_1|t_0)}} = \frac{14 - 20}{\sqrt{16.67}} = 1.47$$

$P = 0.14$ (2 sided test)

Slide 93

Lecture 4. Correlated Outcomes

11. Independence of binary outcomes
12. Matched pairs

Slide 94

11. Independence of Binary Outcomes

A sequence of 0's and 1's are observed. Is the sequence independent?

Example: U.S. Presidential Elections (1912–1996)

Year 1912 ↘ ⟷ ↙ 1996

1 1 0 0 0 1 1 1 1 1 0 0 1 1 0 0 1 0 0 0 1 1

$$Y = \begin{cases} 1 & \text{if Democrat elected} \\ 0 & \text{if Republican elected} \end{cases}$$

Slide 95

Joint Distribution

$$\begin{aligned} f(y_1, \cdots, y_n) &= f(y_1) f(y_2|y_1) f(y_2|y_1, y_2) \\ &\quad \ldots f(y_n|y_1, y_2, \cdots, y_{n-1}) \end{aligned}$$

Assume

$$f(y_i|y_1, y_2, \cdots, y_{i-1}) = f(y_i|y_{i-1})$$

Markovian Assumption

$$\Rightarrow \boxed{f(y_1, \cdots, y_n) = f(y_1) f(y_2|y_1) f(y_3|y_2) \cdots f(y_n|y_{n-1})}$$

Slide 96

$$\boxed{f(y_1,\cdots,y_n) = f(y_1)f(y_2|y_1)\cdots f(y_n|y_{n-1})}$$

Let

$$\theta_i = P\{Y_i = 1|Y_{i-1} = y_{i-1}\} = e^{\alpha+\beta y_{i-1}} \big/ 1 + e^{\alpha+\beta y_{i-1}}$$

$$= \begin{cases} e^{\alpha} \big/ 1 + e^{\alpha} & \text{if} \quad y_{i-1} = 0 \\ e^{\alpha+\beta} \big/ 1 + e^{\alpha+\beta} & \text{if} \quad y_{i-1} = 1 \end{cases}$$

$$\lambda = \text{logit } \theta_i = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta y_{i-1}$$

Slide 97

This is logistic regression model with $x_i = y_{i-1}$

$$t_0 = \sum_{i=1}^{n} y_i \quad , t_1 = \sum_{1}^{n} x_i y_i = \sum_{1}^{n} y_{i-1} y_i$$

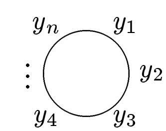
Note: $t_1 = y_0 y_1 + y_1 y_2 + \cdots + y_{n-1} y_n$

What is y_0? — Undefined.

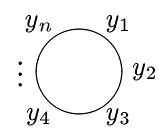
Arbitrarily define $y_0 = y_n$

Instead of considering a sequence on a line, $y_1 \cdots, y_n$

Consider a sequence on a circle



Slide 98



Then it is natural to consider $y_0 = y_n$.

With $y_0 = y_n$ $\quad t_0 = \sum_1^n y_i \;, t_1 = \sum_1^n y_{i-1} y_i$

are defined.

Consider

$$t_0 - t_1 = \sum_1^n y_i - \sum_1^n y_{i-1} y_i$$

$$t_0 - t_1 = \sum_1^n y_i (1 - y_{i-1})$$

$$y_i(1 - y_{i-1}) = \begin{cases} 1 & \text{if } y_i = 1 \text{ and } y_{i-1} = 0 \\ 0 & \text{otherwise} \end{cases}$$

Slide 99

Example:

$t_0 = 6, t_1 = 4$

0 1
1 1 $t_0 - t_1 = 2$
1 1 4 runs
0 1

$t_0 = 4, t_1 = 1$

0 0
1 1 $t_0 - t_1 = 3$
1 0 6 runs
0 1

A run is a consecutive sequence of 0's and 1's. If W = number of runs,

$$\boxed{W = 2(t_0 - t_1)}$$

Slide 100

$$W = 2(t_0 - t_1)$$

We know

$$E(T_1|t_0) = t_0\overline{x}, \quad V(T_1|t_0) = t_0(1 - \frac{t_0}{n})s^2$$

Conditional on $T_0 = t_0$ being fixed. Hence if we find $E(T_1|t_0)$ and $V(T_1|t_0)$ we have the mean and variance of W.

$$t_0 = \sum_1^n y_i, \overline{x} = \sum_1^n \frac{y_{i-1}}{n} = \sum_1^n \frac{y_i}{n} = \frac{t_0}{n} = p$$

$$\begin{aligned} E(T_1|t_0) &= t_0\overline{x} = t_0 p = np^2 \\ V(T_1|t_0) &= t_0(1 - \tfrac{t_0}{n})s^2 = np(1-p)s^2 \end{aligned}$$

Slide 101

$$\begin{aligned} s^2 &= \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1} = \frac{\Sigma_1^n y_i - \frac{(\Sigma_1^n y_i)^2}{n}}{n-1} = \frac{t_0 - \frac{t_0^2}{n}}{n-1} \\ &= \frac{t_0\left(1 - \frac{t_0}{n}\right)}{n-1} = \frac{npq}{n-1} \quad , q = 1 - \frac{t_0}{n} \end{aligned}$$

$$\therefore V(T_1|t_0) = np(1-p) \cdot \frac{npq}{n-1} = \frac{(npq)^2}{n-1}$$

$$\therefore E(W) = 2(t_0 - np^2) = 2(np - np^2) = 2npq$$

$$V(W) = V\left(2(t_0 - t_1)\right) = 4V(T_1|t_0) = \frac{4(npq)^2}{n-1}$$

Slide 102

$$E(W) = 2npq \quad , V(W) = \frac{4(npq)^2}{n-1} = \frac{[E(W)]^2}{n-1}$$

$$\therefore Z = \frac{W - E(W)}{\sqrt{V(W)}} = \frac{W - E(W)}{\sqrt{\frac{E(W)^2}{n-1}}} = \sqrt{n-1}\frac{(W - E(W))}{E(W)}$$

Z is approx. $N(0,1)$

Slide 103

Presidential example: $W = 8$, $n = 22$,

$t_0 = \sum_1^n y_i = 12$

$$p = \frac{t_0}{n} = \frac{12}{22} = .545 \quad , q = .455$$

$$E(W) = 2npq = 2(22)(.545)(.455) = 10.911$$

$$Z = \frac{|(8 - 10.911)|\sqrt{21}}{10.911} = 1.22 \quad , P = 0.22$$

Slide 104

12. Matched Pairs

Example:

The ECOG (Eastern Cooperative Oncology Group) carries out multi-center clinical trials evaluating cancer therapies. ECOG is composed of large treatment centers (members) and community hospitals from which patients are entered in some protocol. Consider clinical trials in which outcome is response (significant reduction in tumor size).

Do patients from community hospitals have the same response as patients from member institutions?

To answer this question, we will extract data from database—many hospitals and protocols.

Slide 105

Experimental Design: Match a community hospital patient with a member institution patient having: same protocol, same treatment, same gender which were entered into a study within 90 days of each other.

A: Community Hospital
B: Member Institution

$$Y = \begin{cases} 1 & \text{if response} \\ 0 & \text{otherwise} \end{cases}$$

Outcome:	(Y_a, Y_b)	No. of pairs
	(1, 1)	132
	(1, 0)	146
	(0, 1)	157
	(0, 0)	501
	Total	936

Slide 106

Data can be summarized in a 2×2 table

		Member Response	Non-response	
Community	Response	132	146	278
Hospital	Non-response	157	501	658
		289	647	936

This is not an ordinary 2×2 table, but represents an aggregate of 936 2×2 tables.

Slide 107

A Typical Table (one pair) (Y_a, Y_b) is

	Response	Non-response	
A	y_a	$1 - y_a$	1
B	y_b	$1 - y_b$	1
	t	$2 - t$	2

Recall in our study of 2×2 tables.

A	s	$n_a - s_a$	n_a
B	$t - s$	$n_b - s_b$	n_b
	t	$n - t$	n

$$\begin{aligned} \theta_a &= e^{\alpha+\beta}/1 + e^{\alpha+\beta} \\ \theta_b &= e^{\alpha}/1 + e^{\alpha} \end{aligned}$$

$$f(s|t) = \frac{\binom{n_a}{s}\binom{n_b}{t-s} e^{\beta s}}{\sum_s \binom{n_a}{s}\binom{n_b}{t-s} e^{\beta s}}, \beta = \log \frac{\theta_a/1-\theta_a}{\theta_b/1-\theta_b}$$

Slide 108

A	$y_a = s$	$1 - y_a$	$1 = n_a$	$y_a, y_b = 0, 1$
B	y_b	$1 - y_b$	$1 = n_b$	
	t	$2 - t$	2	

$$f(s|t) = \binom{n_a}{s}\binom{n_b}{t-s}e^{\beta s} \Big/ \sum_s \binom{n_a}{s}\binom{n_b}{t-s}e^{\beta s}$$

Hence

$$n_a = n_b = 1, s = y_a, t = y_a + y_b$$

$$\therefore \quad \boxed{P\{y_a = y_a|t\} = \binom{1}{y_a}\binom{1}{t-y_a}e^{\beta y_a} \Big/ \binom{1}{0}\binom{1}{t-0}e^{\beta\cdot 0} + \binom{1}{1}\binom{1}{t-1}e^{\beta}}$$

Slide 109

<u>Suppose</u> $t = 0 \Rightarrow P\{Y_a = 1|t = 0\} = 0$

<u>Suppose</u> $t = 2 \Rightarrow P\{Y_a = 1|t = 2\} = 1$

Hence when $t = 0$ or $t = 2$, the value of Y_a is completely determined and carries no uncertainty.

Suppose $t = 1$

$$\begin{aligned} P\{Y_a = y_a|t = 1\} &= \binom{1}{y_a}\binom{1}{1-y_a}e^{\beta y_a} \Big/ \binom{1}{0}\binom{1}{1} + \binom{1}{1}\binom{1}{0}e^{\beta} \\ &= \boxed{\binom{1}{y_a}\binom{1}{1-y_a}e^{\beta y_a} \Big/ 1 + e^{\beta}} \\ &= \begin{matrix} 1/1 + e^{\beta} & \text{if} \quad y_a = 0 \\ e^{\beta}/1 + e^{\beta} & \text{if} \quad y_a = 1 \end{matrix} \end{aligned}$$

Slide 110

$$
\begin{aligned}
P\{Y_a = y_a | t = 1\} &= \binom{1}{y_a}\binom{1}{1-y_a} e^{\beta y_a} \big/ 1 + e^{\beta} \\
&= \boxed{\begin{array}{lll} 1 \big/ 1 + e^{\beta} & \text{if} & y_a = 0 \\ e^{\beta} \big/ 1 + e^{\beta} & \text{if} & y_a = 1 \end{array}}
\end{aligned}
$$

Consider the i^{th} pair: (i refers to treatment, gender pair)

$$\theta_{ai} = e^{\alpha_i + \beta} \big/ 1 + e^{\alpha_i + \beta} = \begin{array}{l}\text{Probability of response} \\ \text{of community hospital} \\ \text{Patient for } i^{th} \text{ pair}\end{array}$$

$$\theta_{bi} = e^{\alpha_i} \big/ 1 + e^{\alpha_i} = \begin{array}{l}\text{Probability of response} \\ \text{of member hospital} \\ \text{Patient for } i^{th} \text{ pair}\end{array}$$

$$\lambda_i = \log \frac{\theta_{ai}/1 - \theta_{ai}}{\theta_{bi}/1 - \theta_{bi}} = \alpha_i + \beta \quad i = 1, 2, \cdots, N$$

$$N = \text{ no. of pairs}$$

$\Rightarrow$ Every pair has a unique parameter, but

$$\boxed{P\{Y_{ai} = y_{ai} | t_i = 1\} = \begin{cases} 1 \big/ 1 + e^{\beta} & \text{if} \quad y_{ai} = 0 \\ e^{\beta} \big/ 1 + e^{\beta} & \text{if} \quad y_{ai} = 1 \end{cases}}$$

Slide 111

We only consider pairs for which $t_i = 1$.

The outcomes are dissimilar; i.e.,

$$(y_a, y_b) = (0, 1) \text{ or } (1, 0)$$

Pairs which are like outcomes

$$(y_a, y_b) = (0, 0) \text{ or } (1, 1)$$

do not depend on β and hence cannot be used for an inference on β.

Consider all pairs for which $t_i = 1$.

$$
\begin{aligned}
f(y_{ai} | t_i = 1) &= \left(\frac{1}{1+e^{\beta}}\right)^{1-y_{ai}} \left(\frac{e^{\beta}}{1+e^{\beta}}\right)^{y_{ai}} \\
&= e^{\beta y_{ai}} \big/ (1 + e^{\beta})
\end{aligned}
$$

Slide 112

If there are n pairs for which $t_i = 1$

$$f(y_{a1}, y_{a2}, \cdots, y_{an} | t_1 = 1, \cdots, t_n = n) = e^{\beta \Sigma_1^n y_{ai}} \Big/ (1 + e^{\beta})^n$$

$$s = \sum_{i=1}^{n} y_{ai} = \text{sufficient statistic for } \beta$$

$$f(\boldsymbol{y}_a | \boldsymbol{t} = 1) = e^{\beta s} \,/ (1 + e^{\beta})^n$$

Slide 113

$$f(\boldsymbol{y}_a | \boldsymbol{t} = 1) = e^{\beta s} \,/ (1 + e)^n \,, s = \sum_{i=1}^{n} y_{ai}$$

Note that this joint distribution is the same as having n independent Bernouilli Trials with success probability $\theta = e^{\beta} \,/ 1 + e^{\beta}$. Hence the distribution of $S = \sum_{1}^{n} Y_{ai}$ is

Binomial Distribution

$$f(s | \boldsymbol{t} = 1) = P\{S = s | \boldsymbol{t} = 1\} = \binom{n}{s} \frac{e^{\beta s}}{(1 + e^{\beta})^n}$$

or

$$f(s | \boldsymbol{t} = 1) = \tbinom{n}{s} \theta^s (1 - \theta)^{n-s}, \quad \begin{array}{l} \theta = e^{\beta} / 1 + e^{\beta} \\ 1 - \theta = 1 / 1 + e^{\beta} \end{array}$$

The statistical inference is $H_0 : \beta = 0$ *vs.* $H_1 : \beta \neq 0$

Slide 114

$$f(s|\mathbf{t}=1) = \binom{n}{s}\theta^s(1-\theta)^{n-s}, \theta = e^\beta/1+e^\beta$$

Note:

$$\begin{aligned} E(S) &= n\theta = ne^\beta/1+e^\beta \\ V(S) &= n\theta(1-\theta) = ne^\beta/(1+e^\beta)^2 \end{aligned}$$

If $\beta = 0 \quad \theta = 1/2$

$\Rightarrow \quad E(S) = n/2$

$V(S) = n/4$

Slide 115

If $H_0 : \beta = 0$

$$E(S) = n/2 \quad , V(S) = n/4$$

$$\boxed{f(s) = \binom{n}{s}\left(\frac{1}{2}\right)^n, \text{ exact distribution}}$$

Large Sample Distribution

$$S \sim N\left(\frac{n}{2}, \frac{n}{4}\right)$$

$$Z = \frac{S - E(S)}{\sqrt{V(S)}} = \frac{S - \frac{n}{2}}{\sqrt{n/4}} = \boxed{\frac{2s-n}{\sqrt{n}}} \sim N(0,1)$$

$$2s - n = s - (n - s)$$

Example:

		Members R	Members NR
Community	R	132	146
Hospital	NR	157	501

R=Response
NR=Non-response

Only off diagonal terms correspond to $t = 1$ outcomes

Slide 116

Therefore, define $s = 157$, $n = 157 + 146 = 303$

$$Z = \frac{2s - n}{\sqrt{n}} = \frac{s - (n - s)}{\sqrt{n}} = \frac{157 - 146}{\sqrt{303}} = .633$$

$$P\{|Z| > .632\} = .53$$

This test on matched pairs in which like pairs are discarded is called McNemars Test.

The test is simply the **sign test** on dissimilar pairs.

Slide 117

Lecture 5: Proportional Hazards Models and Urn Sampling

13. Proportional Hazards Models
14. Proportional Hazards Models and Urn Sampling

Slide 118

13. Proportional Hazard Models

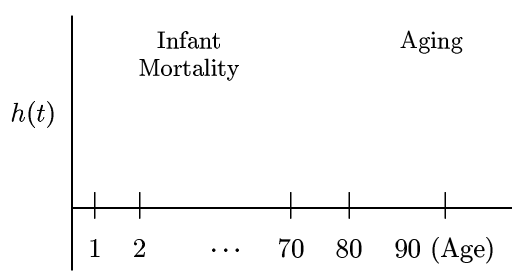
Definitions: Proportional Hazard Models

T= Random Variable Denoting Survival Time

$$h(t)\Delta t = P\{t < T \le t + \Delta t | t > t\}$$

$$\boxed{h(t)\Delta t = P\{t < T \le t + \Delta t\} / P\{T > t\}}$$

Human Population



Slide 119

Proportion Hazard Assumption

$$h(t|x) = h_0(t)e^{\beta x} \quad x\text{: Covariate}$$

Consider $(x_{(i)}, t_{(i)})$ $i = 1, 2, \cdots, n$ where $t_{(i)}$: i^{th} ordered failure; i.e.,

$$t_{(1)} \le t_{(2)} \le t_{(3)} \le \cdots \le t_{(n)}$$

Define $X_{(i)}$ = Covariate for i^{th} failure time.

$$P\{X_{(i)} = x_j\} \propto h_0(t_{(i)})e^{\beta x_j}$$

Slide 120

$$P\{X_{(i)} = x_j\} \propto h_0(t_{(i)})e^{\beta x_j}$$

$$P\{X_{(i)} = x_j\} = \frac{h_0(t_{(i)})e^{\beta x_j}}{\sum_j h_0(t_{(i)})e^{\beta x_j}} = \frac{e^{\beta x_j}}{\sum_j e^{\beta x_j}}$$

j ranges over the $(n - i + i)$ patients who have not failed.

Suppose $\beta = 0$, then

$$P\{X_{(i)} = x_j\} = 1/n$$

i.e. Every x_j has the same probability of being associated with the i^{th} ordered failure time.

Note: More generally

$$P\{X_{(i)} = x_j\} = h_0(t_{(i)})\delta(x_j, \beta)$$

where $\delta(x_j, \beta)$ is a non-negative function of β such that $\delta(x_j, \beta = 0) = 1$ for all x_j.

Slide 121

14. Proportional Hazards Models and Urn Sampling

Observations: $(x_i, t_i) \quad i = 1, 2, \cdots, n$

$\{t_i\}$ represents survival.

Question: It there a relationship between $\{x_i\}$ and $\{t_i\}$.

Case A: No censoring.

Procedure: order $\{t_i\}$

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$$

i.e.; $t_{(i)}$: i^{th} smallest survival time

$x_{(i)}$: covariate associated with $t_{(i)}$.

$x_2 \quad x_3 \cdot x_n$

$x_1 \quad \cdots \quad x_4$

First draw corresponds to covariate associated with $t_{(1)}$. Second draw is covariate associated with $t_{(2)}$, etc.

Slide 122

Draw	Possible Balls (Those at Risk)	Value Drawn	Expected Value
1	$x_{(1)}, \cdots, x_{(n)}$	$x_{(1)}$	$\frac{x_{(1)} + x_{(2)} + \cdots + x_{(n)}}{n}$
2	$x_{(2)}, \cdots, x_{(n)}$	$x_{(2)}$	$\frac{x_{(2)} + x_{(3)} + \cdots + x_{(n)}}{n-1}$
3	$x_{(3)}, \cdots, x_{(n)}$	$x_{(3)}$	$\frac{x_{(3)} + x_{(4)} + \cdots + x_{(n)}}{n-2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{(n)}$	$x_{(n)}$	$x_{(n)}$

$$\text{Total} = x_{(1)} \left[\tfrac{1}{n}\right] + x_{(2)} \left[\tfrac{1}{n} + \tfrac{1}{n-1}\right] + \cdots + x_{(n)} \left[\tfrac{1}{n} + \tfrac{1}{n-1} + \cdots + 1\right]$$

Slide 123

Sum Expected Values

$$T = \frac{x_{(1)}}{n} + x_{(2)} \left[\tfrac{1}{n} + \tfrac{1}{n-1}\right] + x_{(3)} \left[\tfrac{1}{n} + \tfrac{1}{n-1} + \tfrac{1}{n-2}\right] + \cdots + x_{(n)} \left[\tfrac{1}{n} + \tfrac{1}{n-1} + \cdots + 1\right]$$

$$\boxed{T = \sum_{i=1}^{n} x_{(i)} \xi_i}$$

$$\xi_i = \frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots \frac{1}{n-i+1} \quad i = 1, 2, \cdots, n$$

$$\xi_i = \sum_{j=1}^{i} \frac{1}{n-j+1}$$

ξ_i are expected values of order statistics from unit exponential.

Slide 124

$$\xi_i = \frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{n-i+1}$$

$$\boxed{T = \sum_{i=1}^{n} x_{(i)} \xi_i}$$

The distribution of T under the null hypothesis is equivalent to every ball having the same opportunity to be drawn at every draw.

$\Rightarrow$ Exact Distribution of T is found by permuting x's.

Asymptotic Distribution

$$E(T) = \overline{x}\sum_{1}^{n}\xi_i = n\overline{x}, \quad \Sigma\xi_i = n$$

$$V(T) = s^2\left[n - \xi_n\right], \quad s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n-1}$$

$$Z = \frac{T - E(T)}{\sqrt{V(T)}} \quad \text{is asymptotic} \quad N(0,1).$$

Slide 125

Case B: Censored Observations

Right censored observations arise when at the time of analysis a person is still alive; i.e., times: $1, 5+, 6, 8+$.

Among these four observations 5 and 8 are censored (Right censored).

Slide 126

Modification of Urn Sampling Model with Censored Observations

Only draw ball if observation is complete (non-censored).

Example: $(x_{(1)}, 1), (x_{(2)}, 5+), (x_{(3)}, 6), (x_{(4)}, 8+)$

Draw	At Risk	Ball Drawn	Expected Value
1	$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$	$x_{(1)}$	$\left(x_{(1)} + x_{(2)} + x_{(3)} + x_{(4)}\right)/4$
2	$(x_{(3)}, x_{(4)}$	$x_{(3)}$	$x_{(3)} + x_{(4)})/2$

$$T = (x_{(1)} + x_{(2)})(\frac{1}{4}) + (x_{(3)} + x_{(4)})(\frac{1}{4} + \frac{1}{2})$$

Slide 127

Modification with Censoring

Data: (x_i, t_i, δ_i) for $i = 1, 2, \cdots, n$

$$\delta_i = \begin{cases} 1 & \text{if non-censored} \\ 0 & \text{if censored} \end{cases}$$

If $t_{(i)}$ are ordered, then we have $\left((x_{(i)}, t_{(i)}, \delta_{(i)}\right)$.

Slide 128

Define

$$\begin{aligned} \xi_{(j)} &= \sum_{i=1}^{j} \delta_{(i)} / (n-i+1) \\ T &= \sum_{j} \left(\delta_{(j)} - \xi_{(j)}\right) x_{(j)} = \sum_{j} \delta_{(j)} x_{(j)} - \sum_{j} \xi_{(j)} x_{(j)} \end{aligned}$$

Since

$$\begin{aligned} \sum_{j=1}^{n} \xi_{(j)} x_{(j)} &= \sum_{j=1}^{n} \sum_{i=1}^{j} \left[\delta_{(i)}/(n-i+1)\right] x_{(j)} \\ &= \sum_{j=1}^{n} \delta_{(j)} \sum_{i=j}^{n} x_{(i)}/(n-i+1) = \sum_{j=1}^{n} \delta_{(j)} E(x_{(j)}) \\ T &= \sum_{j=1}^{n} \delta_{(j)} [x_{(j)} - E(x_{(j)})] \end{aligned}$$

Slide 129

Then if there is no relation between t_i and x_i.

$$\begin{aligned} E(T) &= 0 \\ V(T) &= \theta n - \xi_n, \quad \theta = \sum_i \delta_i / n, \quad \xi_n = \sum_{i=1}^{n} (n-i+1)^{-1} \end{aligned}$$

Hence $Z = T/\sqrt{\theta n - \xi_n}$ is approximately $N(0,1)$.

Slide 130

Lecture 6: Multiple Logistic Regression

1. Example
2. Comparison of theoretical results from 2×2 tables
3. Joint distribution (Conditional)
4. Asymptotic Distribution

Slide 131

Review: Logistic Regression (one covariate)

Observations: (x_i, y_i) $i = 1, 2, \cdots, n$

Y_i: Binary random variables

$\theta_i = P\{Y_i = 1|x_i\} = e^{\alpha+\beta x_i} / 1 + e^{\alpha+\beta x_i}$

$$\lambda_i = \log \frac{\theta_i}{1-\theta_i} = \text{logit } \theta_i = \alpha + \beta x_i$$

Examples:

2×2 tables

Wilcoxon Two Sample Test

Markovian Models and Run Test

Distribution:

Exact

Asymptotic (Urn Sampling)

Other:

Proportional Hazards Models

Matched Pairs

Slide 132

15. Model and Examples

$$Y_i = \begin{cases} 1 \\ 0 \end{cases} \quad \boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \text{ p-covariates}$$

$$\theta_i = P\{Y_i = 1 | x_i\} \quad \text{Observations } (y_i, x_i)$$

$$\lambda_i = \text{ logit } \theta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\boxed{p = 1 \quad \lambda_i = \beta_0 + \beta_1 x_{i1}}$$

Tests are made on $(\beta_1, \beta_2, \cdots, \beta_p)$.

Examples:

1. Recurrence and Breast Cancer

RAND	
R	— A
A	
N	
D	— B

End Point is Recurrence

Slide 133

Stratify Outcomes by Menopausal Status

Pre-Menopause		Recurrence	No Recurrence	
	A	1	29	30
	B	11	26	37
		12	55	67

Post-Menopause		Recurrence	No Recurrence	
	A	7	59	66
	B	13	50	63
		20	109	129

Slide 134

Example 2:
Deaths and Prognostic Factors
For Breast Cancer

Treatment: A, B

Prognostic Factors

Number of Nodes with Cancer: 1-3, 4+

Size of Tumor: $\leq$ 3cm, $>$ 3 cm

Deaths/Totals

Nodes	Size	A	B
1-3	≤ 3	4/21	1/21
4+	≤ 3	4/11	4/20
1-3	> 3	3/13	2/15
4+	> 3	9/15	4/12

Four 2×2 Tables (Factorial Structure)

Slide 135

Example 3:
Testing k Binomial Populations

Example: $k = 3$

Group	S	F	
A	0	4	4
B	2	2	4
C	1	5	6
	3	11	14

Slide 136

Example 4:

Multi-Center Trial (Two Drugs)

	A		B	
Test Site	S	F	S	F
1	0	15	0	15
2	0	39	6	32
3	1	20	3	18
4	1	14	2	15
5	1	20	2	19
6	0	12	2	10
7	3	49	10	42
8	0	19	2	17
9	1	14	0	15
10	2	26	2	27
	9	228	29	210

Slide 137

16. Several Contingency Tables

Models

Consider k 2×2 tables

i^{th}Table		S	F		
	A	s_i	–	n_i	$i = 1, 2, \cdots, k$
	B	–	–	m_i	
		t_i	$n_i - t_i$	N_i	

θ_{ai}, θ_{bi} : Success Probabilities

$$\lambda_{ai} = \log\left(\frac{\theta_{ai}}{1-\theta_{ai}}\right) \quad , \lambda_{bi} = \log\left(\frac{\theta_{bi}}{1-\theta_{bi}}\right)$$

Common Odds Ratio: $\lambda_{ai} = \alpha_i + \beta \quad , \lambda_{bi} = \alpha_i$

$$\lambda_{ai} - \lambda_{bi} = \beta = \log\left(\frac{\theta_{ai}/1-\theta_{ai}}{\theta_{bi}/1-\theta_{bi}}\right)$$

Slide 138

Note on Model with Interaction

$$\lambda_{ai} - \lambda_{bi} = \beta_i = \beta + \beta_i - \beta = \beta + \delta_i$$

where $\delta_i = \beta_i - \beta$.

If $\beta_1 = \beta_2 = \ldots = \beta_k$ then $\delta_i = 0$ for all i.

$$\beta = \sum_1^k \beta_i / k \quad , \therefore \Sigma\delta_i = 0$$

Hence since

$$\sum_{i=1}^k \delta_i = 0 \Rightarrow k - 1 \text{ independent deviations}$$

Slide 139

Alternatively

$$\lambda_{ai} - \lambda_{bi} = \beta_i = \beta + \delta_i \qquad i = 1, 2, \cdots, k-1$$

$$(\lambda_{ai} - \lambda_{bi)} - (\lambda_{ak} - \lambda_{bk}) = \beta_i - \beta_k$$

$\beta_i - \beta_k = \delta_i$ $i = 1, 2, \cdots, k-1$ deviations.

In both cases there are only k β_i.

If $\sum_{i=1}^k \delta_i = 0 \Rightarrow k - 1$ of δ_i are linearly independent.

If $\delta_k = 0 \Rightarrow \delta_1, \delta_2, \cdots, \delta_{k-1}$ are independent.

Slide 140

Reminder: Single 2×2 Table

	$\underline{S}$	$\underline{F}$	
A	s	–	n
B	–	–	m
	t	$N-t$	N

$\lambda_a - \lambda_b = \beta$

$$f(s|t) = C(s,t)e^{\beta s} \Big/ \sum_s C(s,t)e^{\beta s}$$

$$C(s,t) = \binom{n}{s}\binom{m}{t-s} \Big/ \binom{N}{t}$$

Also, if $\beta = 0$

$$E(S|t) = \tfrac{nt}{N},$$
$$V(S|t) = \tfrac{t(N-t)nm}{N^2(N-1)}$$

or if $p = t/N$, under $H_0 : \beta = 0$

$$E(S|t) = np, \quad V(S|t) = pqn\frac{(N-n)}{N-1}$$

Slide 141

17. Theoretical Development: (Several Contingency Tables, Constant Odds Ratio)

For i^{th} Table assume $\beta_i = \beta$ i.e.,

$$\begin{aligned} f(s_i|t_i) &= C(s_i,t_i)e^{\beta s_i} \Big/ \sum_{s_i} C(s_i,t_0)e^{\beta s_i} \\ &= C(s_i,t_i)\ e^{\beta s_i} / D_i \end{aligned}$$

Hence joint distribution of $(s_1, s_2, \cdots, s_k)$ conditional on $t_1, \cdots, t_k$ is

$$f(s_1, \cdots, s_k | t_1, \cdots, t_k) = \prod_1^k f(s_i|t_i)$$

$$f(s_1, \cdots, s_k | t_1, \cdots, t_k) = \prod_{i=1}^k C(s_i,t_i)\ e^{\beta \Sigma_1^k s_i} \Big/ \prod_1^k D_i$$

Note that conditional on $t_1, \cdots, t_k$, $s = \Sigma s_i$ is a sufficient statistic. Hence we only need the distribution of $s = \sum_1^k s_i$.

Slide 142

$$f(s|t_1,\cdots,t_k) = \sum_{s_1+\cdots+s_k=s} f(s_1,\cdots,s_k|t_1,\cdots,t_k)$$
$$= e^{\beta s} \sum_{s_1+\cdots+s_k=s} \prod_1^k C(s_i,t_i) \Big/ \prod_1^k D_i$$

Slide 143

$$f(\boldsymbol{s}|\boldsymbol{t}) = e^{\beta s} \sum_{s_1+\cdots+s_k=s} \prod_1^k C(s_i,t_i) \Big/ \prod_1^k D_i$$

Define

$$C(\boldsymbol{s},\boldsymbol{t}) = \sum_{s_1+\cdots+s_k=s} \prod_1^k C(s_i,t_i)$$

$$\boxed{f(\boldsymbol{s}|\boldsymbol{t}) = \frac{e^{\beta S} C(s,\boldsymbol{t})}{\sum_s e^{\beta S C(s,\boldsymbol{t})}}}$$

Since

$$C(s_i,t_i) = \binom{n_i}{s_i}\binom{m_i}{t_i - s_i}$$

$$\boxed{C(\boldsymbol{s},\boldsymbol{t}) = \sum_{s_1+\cdots+s_k=s} \prod_1^k \binom{n_i}{s_i}\binom{m_i}{t_i - s_i}}$$

Slide 144

If $\beta = 0 \Rightarrow \theta_{ai} = \theta_{bi}$ for all i

$$f_0(\boldsymbol{s}|\boldsymbol{t}) = C(s, \boldsymbol{t}) \Big/ \sum_s C(\boldsymbol{s}, \boldsymbol{t})$$

StatXact computes the exact test for

$$H_0 : \beta = 0 \ vs. \ H_1 : \beta \neq 0.$$

Slide 145

Asymptotics: If $\beta = 0$

$$E(S_i) = n_i p_i \quad , V(S_i) = n_i p_i q_i \left(\frac{N_i - n_i}{N_i - 1} \right)$$

$$p_i = t_i / N_i$$

Since the test statistic is $S = \sum_1^k S_i$

$$E(S) = \sum_{i=1}^k n_i p_i$$
$$V(S) = \sum_{i=1}^k n_i p_i q_i \left(\frac{N_i - n_i}{N_i - 1} \right)$$

and

$$Z = \frac{S - E(S)}{\sqrt{V(S)}} \sim N(0, 1)$$

The test using the approximation Z is called the "Mantel-Haenzel Test." First proposed by W. C. Cochran, but with $N_i - 1$ replaced by N_i.

Slide 146

18. Theoretical Development (General Case). Test for Constant Odds Ratio.

Consider $\lambda_{ai} - \lambda_{bi} = \beta_i = \beta_k + \delta_i \quad \delta_i = \beta_i - \beta_k$

$$f(s_i|t_i) = C(s_i, t_i)e^{(\beta_k+\delta_i)s_i} / D_i$$

Hence

$$f(s_1, \cdots, s_k|t_1, \cdots, t_k) = f(\boldsymbol{s}, \boldsymbol{t})$$

$$= \prod_1^k C(s_i, t_i)e^{\beta_k s + \Sigma_1^{k-1}\delta_i s_i} \Big/ \prod_1^k D_i$$

where $s = \sum_1^k s_i$.

A test of $H_0 : \delta_1 = \delta_2 = \cdots = \delta_{k-1} = 0$ corresponds to the test for interaction. Note that $s, s_1, \cdots, s_{k-1}$ are sufficient statistics conditional on $\boldsymbol{t} = (t_1, \cdots, t_k)$.

Slide 147

Since β is a nuisance parameter, condition on s and $\boldsymbol{t}$.

$$f(s_1, \cdots, s_k|\boldsymbol{t}, s) = \frac{C(\boldsymbol{s}, \boldsymbol{t})e^{\Sigma_1^{k-1}\delta_i s_i}}{\sum_{s_1+\cdots+s_k=s} C(\boldsymbol{s}, \boldsymbol{t})e^{\Sigma_1^{i-1}\delta_i s_i}}$$

where

$$C(\boldsymbol{s}, \boldsymbol{t}) = \prod_{i=1}^k \binom{n_i}{s_i}\binom{m_i}{t_i - s_i} \quad , s_1 + \cdots + s_k = s$$

If $H_0 : \delta_1 = \cdots = \delta_{k-1} = 0$, then

$$\boxed{f(s_1, \cdots, s_k|\boldsymbol{t}, s) = C(\boldsymbol{s}, \boldsymbol{t}) \Big/ \sum_{s_1+\cdots+s_k=s} C(\boldsymbol{s}, \boldsymbol{t})}$$

Slide 148

Define

$$\mu_i = n_i p_i \quad , \sigma_i^2 = n_i p_i q_i \left(\frac{N_i - n_i}{N_i - 1} \right), p_i = t_i / N_i$$

Then if $\delta_i = 0$ for $i = 1, 2, \cdots, k-1$

$$E(S_i | t_i) = \mu_i \quad V(S_i | t_i) = \sigma_i^2$$

If $S_i \sim N(\mu_i, \sigma_i^2)$

$$f(s_1, \cdots, s_k | \boldsymbol{t}) \propto \exp - \sum_1^k (s_i - \mu_i)^2 / 2\sigma_i^2$$

and since $S = \sum_1^k (S_i)$

$$f(s|\boldsymbol{t}) \propto e^{-\frac{1}{2}(s-\mu)^2/\sigma^2}$$

$$\mu = \sum_1^k \mu_i \quad , \sigma^2 = \sum_1^k \sigma_i^2$$

$$f(s_1 \cdots, s_k | \boldsymbol{t}, s) = \frac{f(s_1, \cdots, s_k | \boldsymbol{t})}{f(s|\boldsymbol{t})}$$

Slide 149

$$f(s_1 \cdots, s_k | \boldsymbol{t}, s) = \frac{f(s_1, \cdots, s_k | \boldsymbol{t})}{f(s|\boldsymbol{t})}$$

$$\propto \quad \exp^{-\frac{1}{2}\Sigma_1^k (s_i - \mu_i)^2 / \sigma_i^2} \Big/ \exp^{-\frac{1}{2}(s-\mu)^2/\sigma^2}$$

$$= e^{-\frac{1}{2}\left\{ \Sigma (s_i - \mu_i)^2 / \sigma_i^2 - \frac{(s-\mu)^2}{\sigma^2} \right\}}$$

$$\boxed{\chi_{k-1}^2 = \sum_1^k \frac{(s_i - \mu_i)^2}{\sigma_i^2} - \frac{(s-\mu)^2}{\sigma^2}}$$

Slide 150

$$\chi^2_{k-1} = \sum_1^k \frac{(s_i-\mu_i)^2}{\sigma_i^2} - \frac{(s-\mu)^2}{\sigma^2}$$

is asymptotically chi-square with d.f.=$k-1$

Note:

If $Z_i = (S_i - \mu)/\sigma_i$ then

$$\frac{S-\mu}{\sigma} = \sum_1^k \frac{\sigma_i}{\sigma} Z_i = \sum_{i=1}^k w_i Z_i \quad , w_i = \sigma_i/\sigma$$

$$\begin{aligned} \chi^2_{k-1} &= \textstyle\sum_{i=1}^k Z_i^2 - \left(\sum_{i=1}^k w_i Z_i\right)^2 \\ &= Z'AZ, \quad A = I - ww' \\ & \qquad w' = \left(\tfrac{\sigma_1}{\sigma}, \tfrac{\sigma_2}{\sigma}, \cdots, \tfrac{\sigma_n}{\sigma}\right) \end{aligned}$$

$A^2 = A$ (idempotent)

trace $A = k-1$

$\Rightarrow Z'AZ$ is χ^2_{k-1} if $Z \sim N(0, I)$

Since Z is asymptotically normal, $Z'AZ$ is asymptotically chi-square.

Slide 151

Lecture 7. Multivariate Problems

(Slide 155)

19. Comparison of k Binomial Populations
20. Testing Two Multinomial Populations
21. Analogue Between Logistic Regression and Polychotomous Regression (one covariate)

Slide 152

19. Comparison of k Binomial Populations

Example: 3 binomials

Groups	S	F	n
1	0	4	4
2	2	2	4
3	1	5	6
	3	11	14

General Problem

Groups	S	F	Totals	$\theta = P(S)$
1	s_1	$n_1 - s_1$	n_1	θ_1
2	s_2	$n_2 - s_2$	n_2	θ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
k	s_k	$n_k - s_k$	n_k	θ_k
	t	$N - t$	N	

$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$

$$\lambda_i = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta_i \quad i = 1, 2, \cdots, k$$

Slide 153

In the problem formulation it is convenient to set $\beta_k = 0$ as there are only k independent parameters.

If $\beta_1 = \cdots = \beta_{k-1} = 0 \Rightarrow \theta_1 = \theta_2 = \cdots = \theta_k$

Slide 154

$$f(s) = \binom{n}{s}\theta^s(1-\theta)^{n-s}$$

However if

$$\lambda = \log\frac{\theta}{1-\theta} = \alpha + \beta$$

$$\theta = e^{\alpha+\beta} \big/ 1 + e^{\alpha+\beta}$$

$$\boxed{f(s) = \tbinom{n}{s} e^{(\alpha+\beta)s} \big/ \left(1 + e^{\alpha+\beta}\right)^n}$$

Hence for i^{th} table

$$f(s_i) = \binom{n_i}{s_i} e^{(\alpha+\beta_i)s_i} \Big/ \left(1 + e^{\alpha+\beta_i}\right)^{n_i}$$

$\therefore$ **Joint distribution**

$$\begin{aligned} f(s_1, \cdots, s_k) &= \prod_{i=1}^{k} \binom{n_i}{s_i} e^{(\alpha+\beta_i)s_i} \Big/ \left(1 + e^{\alpha+\beta_i}\right)^{n_i} \\ &= \prod_{1}^{k} \binom{n_i}{s_i} e^{\alpha\Sigma s_i + \Sigma_1^{k-1}\beta_i s_i} \Big/ D \end{aligned}$$

Slide 155

Sufficient Statistics are: $t = \Sigma_i s_i,\ s_1, \cdots, s_{k-1}$

α is nuisance parameter $\Rightarrow$ condition on t

$$f(s_1, \cdots, s_{k-1}|t) = \frac{C(\boldsymbol{s})\, e^{\alpha t + \sum_1^{k-1}\beta_i s_i} \Big/ D}{\sum_{s_1+\cdots+s_k=t} C(\boldsymbol{s})\, e^{\alpha t + \sum_1^{k-1}\beta_i s_i} \Big/ D}$$

where

$$C(\boldsymbol{s}) = \prod_{1}^{k} \binom{n_i}{s_i}$$

Hence

$$f(s_1, \cdots, s_{k-1}|t) = \frac{C(\boldsymbol{s}) e^{\sum_1^{k-1}\beta_i s_i}}{\sum_{s_1+\cdots+s_k=t} C(\boldsymbol{s}) e^{\sum_1^{k-1}\beta_i s_i}}$$

Slide 156

$$f(s_1, \cdots, s_{k-1}|t) = \frac{C(\boldsymbol{s}) e^{\Sigma_1^{k-1} \beta_i s_i}}{\sum_{s_1,+\cdots+s_k=t} C(\boldsymbol{s}) e^{\Sigma \beta_i s_i}} \quad \left(\begin{array}{l} \alpha \text{ has} \\ \text{dropped} \\ \text{out} \end{array} \right)$$

$$C(\boldsymbol{s}) = \prod_1^k \binom{n_i}{s_i}$$

If $\beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0 \Longleftrightarrow \theta_1 = \theta_2 = \cdots = \theta_k$

$$f_0(\boldsymbol{s}|t) = \prod_1^k \binom{n_i}{s_i} \Big/ \sum_{s_1+\cdots+s_k=t} \prod_1^k \binom{n_i}{s_i}$$

$$\boxed{f_0(\boldsymbol{s}|t) = \prod_1^k \binom{n_i}{s_i} \Big/ \binom{N}{t}}$$ Multivariate Hypergeometric Distribution

$$N = \sum_i n_i, \quad t = \sum_i s_i$$

Slide 157

Consider Data Set

Groups	S	F	
1	0	4	4
2	2	2	4
3	1	5	6
	$t = 3$	11	14= n

$$f(s_1, s_2) = \prod_1^3 \binom{n_i}{s_i} \Big/ \binom{N}{t}$$

$$= \frac{\binom{4}{s_1}\binom{4}{s_i}\binom{6}{s_3}}{\binom{14}{3}} \quad , s_1 + s_2 + s_3 = 3$$

Note that all marginal totals are fixed.

Slide 158

Permutation Distribution

	s_1	s_2	s_3	$\frac{\binom{4}{s_1}\binom{4}{s_2}\binom{6}{3-s_1-s_2}}{}$
	0	0	3	$1\cdot 1\cdot\binom{6}{3}$=20
	0	1	2	=60
	1	0	2	=60
Observed	0	2	1	=36
	1	1	1	=96
	1	2	0	=24
	2	1	0	=24
	3	0	0	=4
	0	3	0	=4
	2	0	1	=36
				364

Probability of outcome = or more extreme than observed

$$P = \frac{20+36+36+24+24+4+4}{\binom{14}{3}} = .407$$

Slide 159

Urn Sample Model

$N = n_1 + n_2 + \cdots + n_k$ = number of balls in urn

Let there be k types of balls (Red, blue, yellow, etc.

n_i = number of balls of i^{th} type

Draw a random sample of t balls without replacement

Define: s_i = number of balls of Type i in sample

$$t = \sum_1^k s_i$$

We can show

$$\begin{aligned} E(S_i) &= t\tfrac{n_i}{N} = tp_i \quad, p_i = n_i/N \\ V(S_i) &= \tfrac{t(N-t)}{N-1} p_i(1-p_i) = \sigma_{ii} \\ \mathrm{Cov}(S_i, S_j) &= -\tfrac{t(N-t)}{N-1} p_i p_j = \sigma_{ij} \end{aligned}$$

Asymptotically $S_1, \cdots, S_{k-1}$ has multivariate normal distribution with mean $E(S_i) = tp_i$ and variance-covariance matrix $V = (\sigma_{ij})$

$\Rightarrow [S - E(S)]' V^{-1} [S - E(S)] = \chi^2_{k-1}$

Slide 160

$$Q = \frac{N-1}{N-t} \sum_{1}^{k} \frac{(s_i - tp_i)^2}{tp_i} \quad , p_i = n_i/N$$

Asymptotic χ^2 with d.f. $= k-1$

This statistic is like $\frac{\text{(observed-expected)}^2}{\text{Expected}}$

Data Set				Expected $= tp_i = \frac{tn_i}{N}$
1	0	4	4	$3 \cdot \frac{4}{14}$
2	2	2	4	$3 \cdot \frac{4}{14}$
3	1	5	6	$3 \cdot \frac{6}{14}$
	3	11	14	

$t = 3, N = 14$

$$Q = \frac{13}{11} \left[\frac{(0 - \frac{12}{14})^2}{12/14} + \frac{(2 - \frac{12}{14})^2}{12/14} + \frac{(1 - \frac{18}{14})^2}{18/14} \right]$$

$= 2.89$ d.f. $= k - 1 = 2$

$\boxed{P = 0.41}$

Slide 161

Lecture 8: Polychotomous Regression

22. Review (Testing k binomials)
23. Testing Two Multinomial Distributions
24. Logistic Regression and Polychotomous Regression (one-covariate)
    - Urn Sampling Model and Asymptotics
    - Examples
      Testing two Multinomials
      Generalized Wilcoxon Test
      (Rank Sum Test)

Slide 162

Review

Testing k Binomial Populations

$$f(s_i = \binom{n_i}{s_i}\theta_i^{s_i}(1-\theta_i)^{n_i-s_i} \quad i = 1, 2, \cdots, k$$

<u>$k \times 2$ Table</u>

Populations	S	F	
1	s_1	$n_1 - s_1$	n_1
2	s_2	$n_2 - s_2$	n_2
$\vdots$	$\vdots$		
k	s_k	$n_k - s_k$	n_k
	t	$N - t$	N

$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$

$$\lambda_i = \log \frac{\theta_i}{1-\theta_i} = \alpha_i + \beta_i \quad i = 1, 2, \cdots, k, \quad \beta_k = 0$$

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$

Slide 163

Distribution under null

$$f_0(s_1, \cdots, s_{k-1}|t) = \prod_1^k \binom{n_i}{s_i} \Big/ \binom{N}{t}, \quad t = \sum_1^k s_i$$

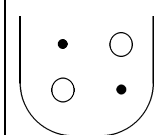
Multivariate Hypergeometric Distribution

Urn Sampling Model

$N = n_1 + n_2 + \cdots + n_k =$ No. of balls in urn.

Let there be k types of balls (different colors)

$n_i =$ No. of balls of type i



→ Draw sample of t balls without replacement → $s_i =$ no. of balls in sample of Type i

Slide 164

20. Testing Two Multinomial Distributions

$$f_a(s_{a1}, s_{a2}, \cdots, s_{ak}) = \frac{n_a!}{\prod_{i=1}^{k} s_{ai!}} \theta_{a1}^{s_{a1}} \theta_{a2}^{s_{a2}} \cdots \theta_{ak}^{s_{ak}}$$

$$\sum_1^k \theta_{ai} = 1$$

$$f_b(s_{b1}, s_{b2}, \cdots, s_{bk}) = \frac{n_b!}{\prod_{i=1}^{k} s_{bi!}} \theta_{b1}^{s_{b1}} \theta_{b2}^{s_{b2}} \cdots \theta_{bk}^{s_{bk}}$$

$$\sum_1^k \theta_{bi} = 1$$

$$H_0 : \theta_{a1} = \theta_{b1}, \theta_{a2} = \theta_{b2}, \cdots, \theta_{ak} = \theta_{bk}$$

Slide 165

Re-parameterize

$$\lambda_{ai} = \log \tfrac{\theta_{ai}}{\theta_{ak}} = \alpha_i + \beta_i \quad i = 1, 2, \cdots, k-1$$

$$\lambda_{bi} = \log \tfrac{\theta_{bi}}{\theta_{bk}} = \alpha_i$$

$$\frac{\theta_{ai}}{\theta_{ak}} = e^{\alpha_i + \beta_i}, \theta_{ai} = \theta_{ak} e^{\alpha_i + \beta_i}$$

But

$$\sum_1^k \theta_{ai} = \theta_{ak} \left[\sum_1^{k-1} e^{\alpha_i + \beta_i} + 1 \right] = 1$$

$$\Rightarrow \boxed{\theta_{ak} = \left[1 + \sum_1^{k-1} e^{\alpha_i + \beta_i} \right]^{-1}}$$

Slide 166

$$\theta_{ai} = \theta_{ak} e^{(\alpha_i+\beta_i)} \quad i-1,2,\cdots,k-1$$

$$\theta_{ak} = 1 \Big/ 1+\sum_1^{k-1} e^{\alpha_i+\beta_i}$$

$$\therefore \boxed{\theta_{ai} = e^{\alpha_i+\beta_i} \Big/ 1+\sum_1^{k-1} e^{\alpha_i+\beta_i}} \quad i=1,2,\cdots,k-1$$

Similarly

$$\boxed{\theta_{ai} = e^{\alpha_i} \Big/ 1+\sum_1^{k-1} e^{\alpha_1}} \quad i=1,2,\cdots,k-1$$

Slide 167

Likelihood:

$$f_a(\boldsymbol{s}_a)f_b(\boldsymbol{s}_b) = \frac{n_a!}{\prod s_{ai}!}\prod_1^k \theta_{ai}^{s_{ai}} \times \frac{n_b!}{\prod s_{bi}!}\prod_1^n \theta_{bi}^{s_{bi}}$$

$$= C(\boldsymbol{s}_b,\boldsymbol{s}_b) \quad \prod_{i=1}^{k-1}\left[\frac{e^{\alpha_i+\beta_i}}{1+\Sigma_1^{k-1}e^{\alpha_i+\beta_i}}\right]^{s_{ai}} \theta_{ak}^{s_{ak}}$$

$$\times \prod_{i=1}^{k-1}\left[\frac{e^{\alpha_i}}{1+\Sigma_1^{k-1}e^{\alpha_i}}\right]^{s_{bi}} \theta_{bk}^{s_{bk}}$$

$$f_a(\boldsymbol{s}_a)f_b(\boldsymbol{s}_b) = C(\boldsymbol{s}_a,\boldsymbol{s}_b)\frac{e^{\Sigma_1^{k-1}\alpha_i(s_{ai}+s_{bi})+\Sigma_1^{k-1}\beta_i s_{ai}}}{\left(1+\Sigma_1^{k-1}e^{\alpha_i+\beta_i}\right)^{n_a}\left(1+\Sigma_1^{k-1}e^{\alpha_i}\right)^{n_b}}$$

Slide 168

$$f_a(\boldsymbol{s}_a)f_b(\boldsymbol{s}_b) = C(\boldsymbol{s}_a, \boldsymbol{s}_b)\frac{e^{\Sigma_1^{k-1}\alpha_i(s_{ai}+s_{bi})+\Sigma_1^{k-1}\beta_i s_{ai}}}{\left(1+\Sigma_1^{k-1}e^{\alpha_i+\beta_i}\right)^{n_a}\left(1+\Sigma_1^{k-1}e^{\alpha_i}\right)^{n_b}}$$

Sufficient Statistics are $t_i = s_{ai} + s_{bi},\ s_{ai}$ for $i = 1, 2, \cdots, k-1$.

The α_i are nuisance parameters. Consider

$$f(\boldsymbol{s}_a|t_1, \cdots, t_{k-1}) = \frac{Ce^{\Sigma_1^{k-1}\alpha_i t_i + \Sigma_1^{k-1}\beta_i s_{ai}} \Big/ D}{\sum_{s_{a1}+\cdots+s_{ak}=n_a} Ce^{\Sigma\alpha_i t_i + \Sigma\beta_i s_{ai}} \Big/ D}$$

$$= \frac{Ce^{\Sigma_1^{k-1}\beta_i s_{ai}}}{\sum_{s_{a1}+\cdots+s_{ak}=n_a} Ce^{\Sigma_1^{k-1}\beta_i s_{ai}}}$$

Slide 169

If $H_0 : \beta_1 = \cdots = \beta_{k-1} = 0$ then

$$f_0(\boldsymbol{s}_a, \boldsymbol{s}_b|\boldsymbol{t}) = \frac{\frac{n_a!}{s_{a1}!\cdots s_{ak}!}\cdot\frac{n_b!}{s_{b1}!\cdots s_{bk}!}}{\sum_{s_{a1}+\cdots+s_{ak}=n_a}\frac{n_a!}{s_{a1}!\cdots s_{ak}!}\frac{n_b!}{s_{b1}!\cdots s_{bk}!}}$$

Slide 170

Note:

$$\frac{n_{a!}}{s_{a1!}\cdots s_{ak!}}\frac{n_{b!}}{(t_1 - s_{a1!})!\cdots(t_k - s_{ak!})!}$$

$$= \binom{t_1}{s_{a1}}\binom{t_2}{s_{a2}}\cdots\binom{t_k}{s_{ak}}\frac{n_a!n_b!}{\prod t_i!}$$

and

$$\sum_{s_{a1}+\cdots+s_{ak}=n_a}\binom{t_1}{s_{a1}}\binom{t_2}{s_{a2}}\cdots\binom{t_k}{s_{ak}} = \binom{\Sigma_1^n t_j}{\Sigma_1^n s_{ai}} = \binom{t}{n_a}$$

$$\therefore f_0(\boldsymbol{s}_a|t) = \prod_1^n \binom{t_i}{s_{ai}} \Big/ \binom{t}{n_a}, t = \sum_1^n t_j, \sum_1^k s_{ai} = n_a$$

Slide 171

i.e.,

Population	1	2	$\cdots$	k	
A	s_{a1}	s_{a2}	$\cdots$	s_{ak}	n_a
B	s_{b1}	s_{b2}	$\cdots$	s_{bk}	n_b
	t_1	t_2	$\cdots$	t_k	N

$2 \times k$ table with all marginals fixed. Thus test for 2 multinomials and k binomials is exactly the same.

Slide 172

21. Analogy Between Logistic Regression and Polychotomous Regression (one covariate)

Recall for logistic regression

Observations: $(x_i, Y_i) \quad i = 1, 2, \cdots, n$

$$f(y_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \quad \theta = P\{Y_i = 1\}$$

$$\lambda_i = \log \frac{\theta_i}{1-\theta_i} = \alpha + \beta x_i$$

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(Y_i)$$

Sufficient Statistics: $s = \sum_1^n y_i, \quad t = \sum_1^n x_i y_i$

$\boxed{f_0(t|s) = C(s,t)/\Sigma C(s,t)}$ Exact Test

Slide 173

Consider the k random variables

$Y_1, Y_2, \cdots, Y_k$ such that

$$\theta_i = P\{Y_i = 1\}, \quad \sum_1^k Y_i = 1, \quad \sum_1^k \theta_i = 1$$

$$\boxed{f(y_1, \cdots, y_k) = \theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k}}$$

Note: Above is generalization of binary random variable with $k = 2$,

$$f(y_1, y_2) = \theta_1^{y_1}\theta_2^{y_2} = \theta_1^{y_1}(1-\theta_1)^{1-y_1}$$

Slide 174

Define $\theta_i = P\{Y_i = 1\}$

$$\frac{\theta_i}{\theta_k} = e^{\alpha_i+\beta_i x} \quad i = 1, 2, \cdots, k-1$$

Since

$$\sum_1^k \theta_i = 1 = \sum_1^{k-1} \left(\theta_k e^{\alpha_i+\beta_i x}\right) + \theta_k$$

$$\theta_k = \frac{1}{1 + \sum_1^{k-1} e^{\alpha_i+\beta_i x}}$$

$$\therefore \boxed{f(\boldsymbol{y}) = \prod_1^k \theta_i^{y_i} = \frac{\prod_{i=1}^{k-1} e^{(\alpha_i+\beta_i x) y_i}}{\left(1+\sum_1^{k-1} e^{\alpha_i+\beta_i x}\right)}}$$

Compare for $k = 2$

$$f(\boldsymbol{y}) = \theta_1^{y_1}(1-\theta_1)^{1-y_1} = e^{(\alpha+\beta x)y_1} / 1 + e^{\alpha+\beta x}$$

Binary Random Variable

Slide 175

$$f(\boldsymbol{y}) = \prod_{i=1}^{k-1} e^{(\alpha_i+\beta_i x)y_i} \Bigg/ \left[1 + \sum_{i=1}^{k-1} e^{\alpha_i+\beta_i x_i}\right]$$

Suppose there were n observations from multinomials of the form $(y_1, y_2, \ldots, y_k, x)$; i.e.,

j^{th} Sample: $(y_{1j}, y_{2j}, \cdots, y_{kj}, x_j) \quad j = 1, 2, \cdots, n$

$$f_j(\boldsymbol{y}_j|x_j) = \prod_1^{k-1} \left[e^{(\alpha_i+\beta_i x_j)y_{ij}} \Bigg/ \left(1 + \sum_1^{k-1} e^{\alpha_i+\beta_i x_j}\right)\right]$$

Slide 176

Then the joint distribution is

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, y_n | x_1, \cdots, x_n) = \prod_{j=1}^{n} f_j(\boldsymbol{y}_j | x_j)$$

$$= \frac{\exp\left\{ \sum_{i=1}^{k-1} \alpha_i \sum_{j=1}^{n} y_{ij} + \sum_{i=1}^{k-1} \beta_i \left[\sum_{j=1}^{n} x_j y_{ij} \right] \right\}}{\prod_{j=1}^{n} \left[1 + \sum_{1}^{k-1} e^{\alpha_i + \beta_i x_j} \right]}$$

$$t_{i0} = \sum_{j=1}^{n} y_{ij}, \quad t_i = \sum_{j=1}^{n} x_j y_{ij} \quad i = 1, \cdots, k-1$$

are sufficient statistics for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$

Slide 177

$$f(\boldsymbol{y}|\boldsymbol{x}) = e^{\Sigma_1^{k-1} \alpha_i t_{i0} + \Sigma_1^{k-1} \beta_i t_i} / D$$

$$t_{i0} = \sum_{j=1}^{n} y_{ij}, \quad t_i = \sum_{j=1}^{n} x_j y_{ij}$$

Hence

$$f(\boldsymbol{t}_0, \boldsymbol{t}) = \frac{C(\boldsymbol{t}_0, \boldsymbol{t}) e^{\Sigma_1^{k-1} \alpha_i t_{i0} + \Sigma_1^{k-1} \beta_i t_i}}{D}$$

where

$$C(\boldsymbol{t}_0, \boldsymbol{t}) = \underset{\substack{y_{1j}, \cdots, y_{kj} \quad (j=1,2,\cdots,n) \\ t_{i0}, t_i \text{ fixed}}}{\Sigma \cdots \Sigma} (1)$$

= Number of ways of permuting $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$

So that t_{i0}, t_i are fixed.

Slide 178

Example:

$k = 3$:(Failure, Partial Success, Success) $n = 5$

j	x	Y_1 Failure	Y_2 PS	Y_3 S
1	$x_1 = 0$	1	0	0
2	$x_2 = 0$	0	0	1
3	$x_3 = 0$	1	0	0
4	$x_4 = 1$	0	1	0
5	$x_5 = 1$	1	0	0
	$t_{i0} =$	3	1	1

$$t_{10} = 3, t_{20} = 1, t_{30} = 1$$

$$t_1 = \sum_j y_j x_j = x_1 + x_3 + x_5 = 1$$

$$t_2 = \Sigma y_{2j} x_j = x_4 = 1, \quad t_3 = \Sigma y_{3j} x_j = x_2 = 0$$

$\Rightarrow t_1 = 1, t_2 = 1, t_3 = 0$

Slide 179

We have to permute observations so that

$t_{10} = 3, t_{20} = 1, t_{30} = 1, t_1 = t_2 = 1, t_3 = 0.$

Below under F there are three ways of permuting $(1, 1, 0)$ and two ways of permuting $(1, 0)$ keeping the restrictions. Similarly, under PS, only two ways of permuting $(1, 0)$. Similarly, under S, we can permute the first three. Hence number of permutations is $3 \times 3 \times 2 \times 2 = 36$.

x	F	PS	S
0	1 ⎫	0	1 ⎫
0	1 ⎬ 3	0	0 ⎬ 3
0	0 ⎭	0	0 ⎭
1	1 ⎫ 2	1 ⎫ 2	0
1	0 ⎭	0 ⎭	0

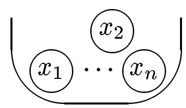
Number of ways $= 3 \times 2 \times 2 \times 3 = 36$

$C(t_{10} = 3, t_{20} = 1, t_{30} = 1, t_1 = 1, t_2 = 1, t_3 = 0) = 36$

Slide 180

Urn Sampling Model

n balls in urn having values $x_1, x_2, \cdots, x_n$



$$n_i = t_{i0} = \sum_{j=1}^{n} y_{ij}, \quad t_i = \sum_{j=1}^{n} x_j y_{ij} = s_i$$

Sample $t_{10} = n_1$ balls, $t_1 = s_1 =$ sum of x's drawn

Sample $t_{20} = n_2$ balls, $t_2 = s_2 =$ sum of x's drawn

$\vdots \qquad \vdots \qquad \vdots$

Sample $t_{k0} = n_k$ balls, $t_k = s_k =$ sum of x's drawn

Define

$$y_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ ball is drawn on } i^{th} \text{ sample} \\ 0 & \text{otherwise} \end{cases}$$

Slide 181

We can show if $H_0 : \beta_1 = \cdots = \beta_{k-1} = 0$

$$\begin{aligned} E(S_i) &= n_i \overline{x} \quad (n_i = t_{i0}) \;\; (s_i = t_i) \\ \text{Var } S_i &= n_i \tfrac{(n-n_i)}{n} s^2, \quad s^2 = \sum_1^n \frac{(x_i - \overline{x})^2}{n-1} \\ \text{Cov}(S_i, S_j) &= -\tfrac{n_i n_j s^2}{n} = \sigma_{ij}, \quad i \neq j \end{aligned}$$

$$\begin{aligned} \Rightarrow \chi^2_{k-1} &= \textstyle\sum_1^k \frac{(s_i - n_i \overline{x})^2}{n_i s^2} \left(\begin{array}{l} \text{Asymptotic chi-square} \\ \text{distribution with } (k-1) \\ \text{degrees of freedom.} \end{array} \right) \\ &= [S - E(S)]' V^{-1} [S - E(S)] \end{aligned}$$

where $S = (S_1, S_2, \ldots, S_k)$

$V = (\sigma_{ij}) \quad i, j = 1, 2, \ldots, k-1$

Slide 182