Fisher’s exact test

More examples

07/22/2021

Berkeley
UNIVERSITY OF CALIFORNIA

In the previous lecture,

- LRT for the multinomial distribution:
  - For $H_0: (p_1, \ldots, p_m) = (p_1(\theta), \ldots, p_m(\theta))$,

$$-2\log\lambda(\mathbf{X}_n) = 2\sum_{i=1}^{m} O_i \log\frac{O_i}{E_i} \xrightarrow{d} \chi^2_{m-2}, \text{ as } n \to \infty.$$

  - Goodness-of-fit test:
    To divide up the interval of possible values into m "cells" or "categories".

- Fisher's exact test:
  - Looks at a 2x2 table:

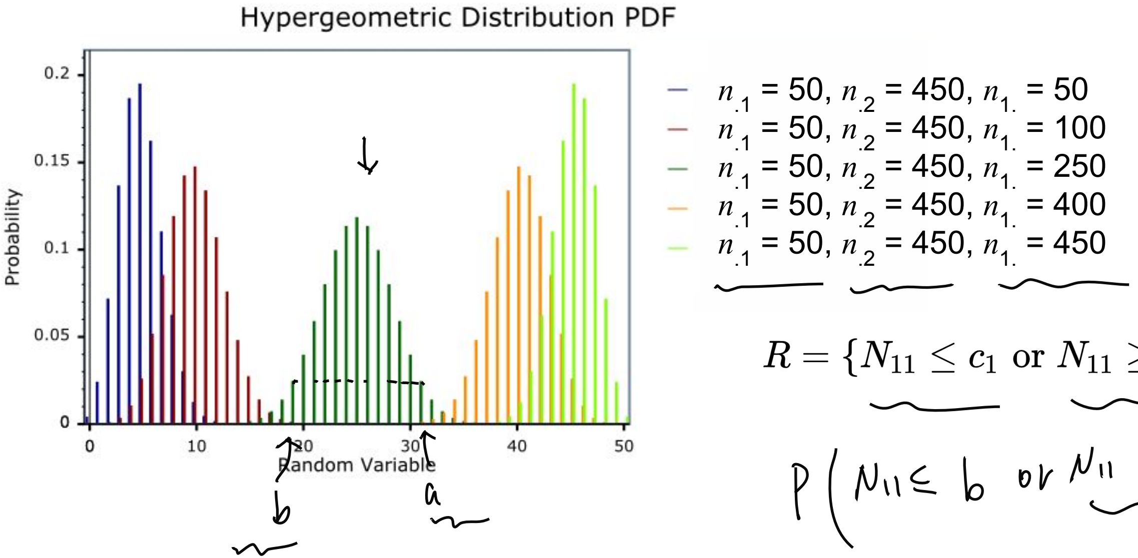
n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

  - If there is no relation between the row and column classifications:

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{1.}-n_{11}}}{\binom{n_{..}}{n_{1.}}} \sim \text{Hypergeometric distribution}$$

  - Only feasible for 2x2 table and small samples.

Hypergeometric distribution



$$R = \{N_{11} \leq c_1 \text{ or } N_{11} \geq c_2\}$$

$$P(N_{11} \leq b \text{ or } N_{11} \geq a \mid H_0)$$

$$\approx 2 * P(N_{11} \geq a \mid H_0)$$

Tea-tasting experiment

H_0 : Bristol has no skills in determining the order.

	Milk first	Tea first	
Milk first	*4*	*0*	*4*
Tea first	*0*	*4*	*4*
	4	*4*	*8*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{4}{n_{11}}\binom{4}{4-n_{11}}}{\binom{8}{4}}$$

N_{11}	0	1	2	3	4
p	0.014	0.229	0.514	0.229	0.014

p-value $= P(N_{11} \leq 0$ or $N_{11} \geq 4 \mid H_0)$

$= 2 \times 0.014 = 0.028 < 0.05$

We reject H_0.

pnorm pbinom pgamma $\to P(X \le x)$ lower.tail = FALSE $1 - P(X \le x) = P(X > x)$

Dependencies between row and column classifications

Example 6. A group of supervisors each examined a personnel file to decide whether to promote the employee or not. The files are identical except for the gender label.

H_0 : There is no gender bias. Any imbalance is due to randomization. $N_{11}: 0, 1, \cdots, 24$

	Male	Female	
Promote	21	14	35
Hold file	3	10	13
	24	24	24+24

* From 13.2 of Rice

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{24}{n_{11}}\binom{24}{35-n_{11}}}{\binom{24+24}{35}}$$

$\Updownarrow$

```
dhyper(n11, m=24, n=24, k=35)
```

$2 * P(N_{11} \ge 21 \mid H_0)$

```
p-value = 2*phyper(21-1, m=24, n=24, k=35,
lower.tail = FALSE)
```

```
[1] 0.04899141
```

< 0.05

$2 * P(N > 20 \mid H_0)$

$2 * P(N > 21 \mid H_0)$

We reject H_0 that there is no gender bias.

Dependencies between row and column classifications

Example 5 *cont'd*. During phase 3 trial, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination.

H_0 : There is no relation. Any imbalance is due to randomization.

$n = 0, 1, \cdots, 341$

	Pfizer / *BNT162b2*	Placebo	
Fever ≥ 38.0°C	331	10	341
No fever	1,767	2,093	3860
	2,098	2,103	n

* Systemic reactions in persons aged 18–55 years

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{2098}{n_{11}}\binom{2103}{341-n_{11}}}{\binom{2098+2103}{341}}$$

$\Updownarrow$

```
dhyper(n11, m=2098, n=2103, k=341)
```

Benefits of more trials and repeated tests
⇒ More significant results

```
p-value = 2*phyper(331-1, m=2098, n=2103, k=341,
lower.tail = FALSE)
```

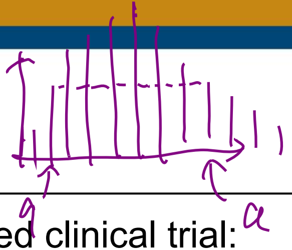
```
[1] 2.548842e-90
```

$\ll 0.05$

$2 * P(N_{11} \geq 331 \mid H_0) =$

Reject H_0

Dependencies between row and column classifications



Example 7. Phase 3 trial was a large, randomized, double-blind, placebo-controlled clinical trial:

H_0 : Infection rate is not related to vaccine/placebo treatment. $\leftrightarrow$ H_1 : It is related.

	Pfizer / *BNT162b2*	Placebo	
SARS-CoV-2 infected	9	169	178
No infection	21,711	21,559	43,439
	21,720	21,728	

* *Age ≥ 16, infections observed with onset at least 7 days after the second dose*

N_{11} : 0, 1, $\cdots$ 178

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{21720}{n_{11}}\binom{21728}{178-n_{11}}}{\binom{21720+21728}{178}}$$

```
dhyper(n11, m=21720, n=21728, k=178)
```

```
p-value = 2*phyper(9, m=21720, n=21728, k=178,
lower.tail = TRUE)
[1] 1.702187e-39
```

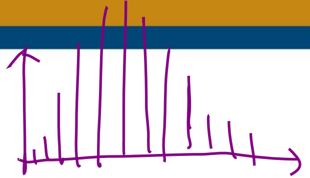
< 0.05

$P(N_{11} \le 9 \text{ or } N_{11} \ge a \mid H_0)$

$\approx 2 * P(N_{11} \le 9 \mid H_0)$

$P(N_{11} \le n_{11})$

Dependencies between row and column classifications



Example 7 *cont'd*. Phase 3 trial was a large, randomized, double-blind, placebo-controlled clinical trial:

H_0 : Infection rate is not related to vaccine/placebo treatment. $\leftrightarrow$ H_1 : Infection rate is $\boxed{\text{lower}}$ in the vaccine group.

	Pfizer / *BNT162b2*	Placebo	
SARS-CoV-2 infected	9	169	178
No infection	21,711	21,559	43,439
	21,720	21,728	

* *Age ≥ 16, infections observed with onset at least 7 days after the second dose*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{21720}{n_{11}}\binom{21728}{178-n_{11}}}{\binom{21720+21728}{178}}$$

$\Updownarrow$

```
dhyper(n11, m=21720, n=21728, k=178)
```

```
p-value = phyper(9, m=21720, n=21728, k=178,
lower.tail = TRUE)
```

```
[1] 8.510933e-40
```

< 0.05

Reject H_0

$R = \{ N_{11} \le c \}$

p-value $= P(N_{11} \le 9 \mid H_0)$

χ^2 test of independence

13.4 of Rice

07/22/2021

Dependencies between row and column classifications

Example 8. A random sample of 650 residents of the city is taken. Their occupations and neighborhoods are recorded.

H_0 : Neighborhood of residence is independent of occupational classification $\quad \leftrightarrow \quad H_1$: It is not.

	A	*B*	*C*	*D*	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	**650**

Assume that Neighborhood $(N) \perp\!\!\!\perp$ Occupation (O) under H_0.

$P(N=A,\ O=W) = \pi_{AW}$

$P(N=B,\ O=W) = \pi_{BW}$

$\vdots$

4x3 joint probabilities

$P(X=x) = \sum_{y \in Y} P(X=x, Y=y)$

$P(N=A)$

$= \sum_{i=\{W,B,N\}} P(N=A, O=i)$

$\pi_{AW} = P(N=A)$

$P(O=W)$

$X \perp\!\!\!\perp Y \iff P(X=x, Y=y)$

$= P(X=x)\, P(Y=y)$

Dependencies between row and column classifications

Example 8 *generalized*. A sample of size n is cross-classified in a table with I rows and J columns.

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \quad \leftrightarrow \quad H_1 : H_0 \text{ is not true.}$$

	1	2	$\cdots$	J
1	π_{11}	π_{12}	$\cdots$	π_{1J}
2	π_{21}	π_{22}	$\cdots$	π_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}
	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$		$\pi_{\cdot j}$

sum $\rightarrow \sum_{j=1}^{J} \pi_{1j} = \pi_{1\cdot}$

$\rightarrow \sum_{j=1}^{J} \pi_{Ij} = \pi_{I\cdot}$

$\pi_{11} = P(R=1, C=1)$

$\pi_{12} = P(R=1, C=2)$

$\pi_{ij} = P(R=i, C=j)$

$\pi_{1\cdot} = P(R=1) = \sum_{j=1}^{J} P(R=1, C=j)$

$\pi_{\cdot j} = P(C=j) = \sum_{i=1}^{I} \pi_{ij}$

$I \times J$ cells $\rightarrow$ multinomial

$\Theta_0 = \{ \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \sum_{i=1}^{I} \pi_{i\cdot} = 1, \sum_{j=1}^{J} \pi_{\cdot j} = 1 \}$

$\Theta = \Theta_0 \cup \Theta_1 = \{ \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_{ij} = 1 \}$

$\dim \Theta_0 = I - 1 + J - 1$

$\dim \Theta = I \times J - 1$

$\nu = \dim \Theta - \dim \Theta_0 = IJ - 1 - I + 1 - J + 1$

$= IJ - I - J + 1 = (I-1)(J-1).$

$\sup_{\Theta} L(\underbrace{\pi_{ij}, i=1,\ldots I, j=1,\ldots J}\mid \underbrace{n_{ij}, i=1,\ldots I, j=1\ldots J})$

$= L(\hat{\pi}_{ij} \mid n_{ij}) = \dfrac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} \hat{\pi}_{ij}^{n_{ij}}.$

$\uparrow$

multinomial $(\pi_{11}, \pi_{12}, \ldots \pi_{IJ})$

$\leftarrow \hat{\pi}_{ij} = \dfrac{n_{ij}}{n}$

Dependencies between row and column classifications

$$\sup_{H_0} L(\pi_{ij} \mid n_{ij}) = \sup_{\substack{\pi_{i\cdot} \\ \pi_{\cdot j}}} \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}} \underbrace{\prod_{i=1}^{I}\prod_{j=1}^{J} (\pi_{i\cdot}\,\pi_{\cdot j})^{n_{ij}}} = \sup_{\substack{\pi_{i\cdot} \\ \pi_{\cdot j}}} \frac{n!}{\prod_i \prod_j n_{ij}} \left[\underbrace{\prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{i\cdot}^{n_{ij}}}\right] \left[\prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{\cdot j}^{n_{ij}}\right]$$

column sum $= n_{i\cdot}$ row sum $= n_{\cdot j}$

$$= \sup_{\substack{\pi_{i\cdot} \\ \pi_{\cdot j}}} \frac{n!}{\prod_i \prod_j n_{ij}} \left[\prod_{i=1}^{I} \pi_{i\cdot}^{\sum_{j=1}^{J} n_{ij}}\right] \left[\prod_{j=1}^{J} \pi_{\cdot j}^{\sum_{i=1}^{I} n_{ij}}\right]$$

$$= \sup_{\substack{\pi_{i\cdot} \\ \pi_{\cdot j}}} \frac{n!}{\prod_i \prod_j n_{ij}} \underbrace{\left[\prod_{i=1}^{I} \pi_{i\cdot}^{n_{i\cdot}}\right]} \underbrace{\left[\prod_{j=1}^{J} \pi_{\cdot j}^{n_{\cdot j}}\right]}.$$

$$= \frac{n!}{\prod_i \prod_j n_{ij}} \underbrace{\sup_{\pi_{i\cdot}} \left[\prod_{i=1}^{n} \pi_{i\cdot}^{n_{i\cdot}}\right]}_{L_0(\pi_{1\cdot}, \cdots \pi_{I\cdot})} \underbrace{\sup_{\pi_{\cdot j}} \left[\prod_{j=1}^{J} \pi_{\cdot j}^{n_{\cdot j}}\right]} \Rightarrow \hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

$$l_0(\pi_{1\cdot}, \cdots \pi_{I\cdot}) = \sum_{i=1}^{I} n_{i\cdot} \log \pi_{i\cdot} = \sum_{i=1}^{I-1} n_{i\cdot} \log \pi_{i\cdot} + n_I \log(1 - \pi_{1\cdot} - \pi_{(I-1)\cdot})$$

$$I-1 \begin{cases} \dfrac{\partial l_0}{\partial \pi_{1\cdot}} = \dfrac{n_{1\cdot}}{\pi_{1\cdot}} - \dfrac{n_{I\cdot}}{1 - \pi_{1\cdot} - \cdots - \pi_{(I-1)\cdot}} = 0 \\ \vdots \\ \dfrac{\partial l_0}{\partial \pi_{(I-1)\cdot}} = \dfrac{n_{(I-1)\cdot}}{\pi_{(I-1)\cdot}} - \dfrac{n_{I\cdot}}{1 - \pi_{1\cdot} - \cdots - \pi_{(I-1)\cdot}} \end{cases} \Rightarrow \hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}$$

$$\lambda(\underline{X}_n) = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L} = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\hat{\hat{\pi}}_{ij}}{\hat{\pi}_{ij}} \right)^{n_{ij}} \quad \text{where} \quad \hat{\hat{\pi}}_{ij} = \hat{\pi}_{i\cdot}\, \hat{\pi}_{\cdot j} = \frac{n_{i\cdot}\, n_{\cdot j}}{n^2}$$

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}.$$

$$-2\log \lambda(\underline{X}_n) = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} - n_{ij} \log\left(\frac{\hat{\hat{\pi}}_{ij}}{\hat{\pi}_{ij}} \right) \xrightarrow{d} \underline{\chi^2_\nu = \chi^2_{(I-1)(J-1)}}$$

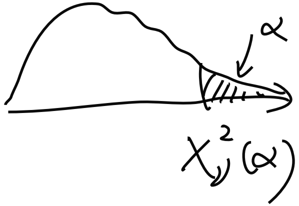
$$= 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log\left(\frac{n\hat{\pi}_{ij}}{n\hat{\hat{\pi}}_{ij}} \right)$$

$$= 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{n_{i\cdot}\, n_{\cdot j}/n} \right)$$

$$= \boxed{2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}} \right) \quad \text{with} \quad E_{ij} = n_{i\cdot}\, n_{\cdot j}/n}.$$

$$\approx \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

$$R = \{ \lambda(\underline{X}_n) \le c \} = \{ -2\log\lambda(\underline{X}_n) \ge \chi^2_\nu(\alpha) \}$$



Dependencies between row and column classifications

Example 8 *cont'd*. A random sample of 650 residents of the city is taken. Their occupations and neighborhoods are recorded.

H_0 : Neighborhood of residence is independent of occupational classification $\leftrightarrow$ H_1 : It is not.

	A	*B*	*C*	*D*	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	**650**

```
> row_sum <- c(349, 151, 150)
> column_sum <- c(150, 150, 200, 150)
> I = length(row_sum); J = length(column_sum); n = sum(row_sum)
>
> Expected <- matrix(NA, nrow = I, ncol=J)
> for (i in 1:I){
+   for (j in 1:J){
+     Expected[i,j] = row_sum[i]*column_sum[j]/n
+   }
+ }
> Expected
         [,1]     [,2]      [,3]     [,4]
[1,] 80.53846 80.53846 107.38462 80.53846
[2,] 34.84615 34.84615  46.46154 34.84615
[3,] 34.61538 34.61538  46.15385 34.61538
```

Dependencies between row and column classifications

Example 8 *cont'd*. A random sample of 650 residents of the city is taken. Their occupations and neighborhoods are recorded.

H_0 : Neighborhood of residence is independent of occupational classification $\leftrightarrow$ H_1 : It is not.

```
> row_sum <- c(349, 151, 150)
> column_sum <- c(150, 150, 200, 150)
> I = length(row_sum); J = length(column_sum); n = sum(row_sum)
>
> Expected <- matrix(NA, nrow = I, ncol=J)
> for (i in 1:I){
+   for (j in 1:J){
+     Expected[i,j] = row_sum[i]*column_sum[j]/n
+   }
+ }
> Expected
         [,1]     [,2]      [,3]     [,4]
[1,] 80.53846 80.53846 107.38462 80.53846
[2,] 34.84615 34.84615  46.46154 34.84615
[3,] 34.61538 34.61538  46.15385 34.61538
```

```
> Observed <- matrix(c(90, 30, 30, 60, 50, 40,
                       104, 51, 45, 95, 20, 35), ncol=4)

> sum((Observed-Expected)^2/Expected)
[1] 24.5712

> qchisq(0.05, df = (I-1)*(J-1), lower.tail=FALSE)
[1] 12.59159
```

```
> pchisq(24.5712, df = (I-1)*(J-1), lower.tail = FALSE)
[1] 0.0004098431
```

Dependencies between row and column classifications

Example 8 *cont'd*. A random sample of 650 residents of the city is taken. Their occupations and neighborhoods are recorded.

H_0 : Neighborhood of residence is independent of occupational classification $\quad \leftrightarrow \quad H_1$: It is not.

	A	B	C	D	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	**150**	**150**	**200**	**150**	**650**

```
> chisq.test(Observed)

        Pearson's Chi-squared test

data:  Observed
X-squared = 24.571, df = 6, p-value = 0.0004098
```

Dependencies between row and column classifications

Example 7 *cont'd*. Phase 3 trial was a large, randomized, double-blind, placebo-controlled clinical trial:

H_0 : Infection rate is not related to vaccine/placebo treatment. $\leftrightarrow$ H_1 : It is related.

	Pfizer / *BNT162b2*	Placebo	
SARS-CoV-2 infected	9	169	178
No infection	21,711	21,559	43,439
	21,720	21,728	

* *Age ≥ 16, infections observed with onset at least 7 days after the second dose*

```
p-value = 2*phyper(9, m=21720, n=21728,
k=178, lower.tail = TRUE)
```

```
[1] 1.702187e-39
```

```
> Observed <- matrix(c(9, 21711, 169, 21559), ncol=2)
> chisq.test(Observed, correct = FALSE)

        Pearson's Chi-squared test

data:  Observed
X-squared = 144.35, df = 1, p-value < 2.2e-16
```

Dependencies between row and column classifications

χ^2 test of independence:

1. The population concerns two categorical variables;
2. Tickets in each cell are independent;
3. Large sample size n so that no more than 20% of expected counts less than 5.

χ^2 test of homogeneity

13.4 of Rice

07/22/2021

Homogeneity among J multinomial distributions

Example 9. Jane Austen left the novel *Sandition* unfinished when she died. An admirer completed the novel while emulating her style. Morton counted the occurrences of various words in several works.

H_0 : Admirer's usage of the words is consistent with Austin's. $\leftrightarrow$ H_1 : It is not.

Word	*Sense and Sensibility*	*Emma*	*Sanditon* I	*Sanditon* II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94 π_{i1}	105 π_{i2}	37 π_{i3}	22 π_{i4}
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196

$I-1$ $\quad I-1$ $\quad I-1$ $\quad I-1$

$H_0: \pi_{i1} = \cdots = \pi_{iJ}$ versus H_1: H_0 is not true

$\Theta_0 = \{\pi_{i1} = \cdots = \pi_{iJ}, \sum_{i=1}^{I} \pi_{ij} = 1\}$, $\Theta = \{\sum_{i=1}^{I} \pi_{ij} = 1\}$

$\dim \Theta_0 = I-1$ $\qquad \dim \Theta = J(I-1)$

$\nu = \dim \Theta - \dim \Theta_0 = (J-1)(I-1)$.

$$\sup_{\Theta} L(\pi_{ij} \mid n_{ij}) = \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!} \prod_{i=1}^{I} \pi_{ij}^{n_{ij}}$$

$$\Rightarrow \hat{\pi}_{ij} = \frac{n_{ij}}{n_{\cdot j}}.$$

Homogeneity among J multinomial distributions

Example 8 *generalized*. Compare J independent multinomial distributions each having I categories.

$$H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ}, \quad i = 1, \ldots, I \quad \leftrightarrow \quad H_1 : H_0 \text{ is not true.}$$

$$\sup_{\Theta_0} L(\pi_{ij} \mid n_{ij}) = \sup_{\pi_{i1}} \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!} \prod_{i=1}^{I} (\pi_{i1})^{n_{ij}} = \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!} \cdot \sup_{\pi_{i1}} \prod_{j=1}^{J}\prod_{i=1}^{I} (\pi_{i1})^{n_{ij}}$$

$$= \prod_{j=1}^{J} \frac{n_{\cdot j}!}{n_{1j}! \cdots n_{Ij}!} \sup_{\pi_{i1}} \prod_{i=1}^{I} \pi_{i1}^{n_{i\cdot}}$$

$$\Rightarrow \tilde{\pi}_{i1} = \frac{n_{i\cdot}}{n} = \hat{\pi}_{i2} = \cdots = \hat{\pi}_{iJ}.$$

$$\lambda(\mathcal{E}_n) = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L} = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{ij}} \right)^{n_{ij}}$$

$$-2\log\lambda(\mathcal{E}_n) = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{\cdot j}\hat{\pi}_{ij}}{n_{\cdot j}\hat{\pi}_{ij}} = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}}{n_{i\cdot} n_{\cdot j}/n} \xrightarrow{d} \chi^2_{(I-1)(J-1)}$$

$$= \boxed{2 \sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij} \log \frac{O_{ij}}{E_{ij}} \quad \text{with} \quad E_{ij} = n_{i\cdot} n_{\cdot j}/n}.$$

$$\approx \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_i - E_i)^2}{E_i}$$

$$R = \{-2\lambda(\mathcal{E}_n) \geq \chi^{2\,(\alpha)}_{(I-1)(J-1)}\}.$$

Homogeneity among J multinomial distributions

Example 9 *cont'd*. Jane Austen left the novel *Sandition* unfinished when she died. An admirer completed the novel while emulating her style. Morton counted the occurrences of various words in several works.

$$H_0: \text{ Admirer's usage of the words is consistent with Austin's.} \quad \leftrightarrow \quad H_1: \text{It is not.}$$

Word	*Sense and Sensibility*	*Emma*	*Sanditon* I	*Sanditon* II	
a	147	186	101	83	517
an	25	26	11	29	91
this	32	39	15	15	101
that	94	105	37	22	258
with	59	74	28	43	204
without	18	10	10	4	42
Total	375	440	202	196	**1,213**

```
> I = length(row_sum); J = length(column_sum); n = sum(row_sum)
>
> Expected <- matrix(NA, nrow = I, ncol=J)
> for (i in 1:I){
+   for (j in 1:J){
+     Expected[i,j] = row_sum[i]*column_sum[j]/n
+   }
+ }
> Expected
          [,1]      [,2]      [,3]     [,4]
[1,] 159.83100 187.53504 86.095631 83.53833
[2,]  28.13273  33.00907 15.154163 14.70404
[3,]  31.22424  36.63644 16.819456 16.31987
[4,]  79.76092  93.58615 42.964551 41.68838
[5,]  63.06678  73.99835 33.971970 32.96290
[6,]  12.98434  15.23495  6.994229  6.78648
```

Homogeneity among J multinomial distributions

Example 9 *cont'd*. Jane Austen left the novel *Sandition* unfinished when she died. An admirer completed the novel while emulating her style. Morton counted the occurrences of various words in several works.

$$H_0: \text{ Admirer's usage of the words is consistent with Austin's.} \quad \leftrightarrow \quad H_1: \text{It is not.}$$

```
> I = length(row_sum); J = length(column_sum); n = sum(row_sum)
>
> Expected <- matrix(NA, nrow = I, ncol=J)
> for (i in 1:I){
+   for (j in 1:J){
+     Expected[i,j] = row_sum[i]*column_sum[j]/n
+   }
+ }
> Expected
          [,1]      [,2]      [,3]     [,4]
[1,] 159.83100 187.53504 86.095631 83.53833
[2,]  28.13273  33.00907 15.154163 14.70404
[3,]  31.22424  36.63644 16.819456 16.31987
[4,]  79.76092  93.58615 42.964551 41.68838
[5,]  63.06678  73.99835 33.971970 32.96290
[6,]  12.98434  15.23495  6.994229  6.78648
```

```
> Observed <- matrix(c(147, 25, 32, 94, 59, 18, 186, 26, 39, 105,
                       74, 10, 101, 11, 15, 37, 28, 10, 83, 29,
                       15, 22, 43, 4), ncol=4)

> sum((Observed-Expected)^2/Expected)
[1] 45.57751

> qchisq(0.05, df = (I-1)*(J-1), lower.tail=FALSE)
[1] 24.99579
```

```
> pchisq(45.57751, df = (I-1)*(J-1), lower.tail = FALSE)
[1] 6.204958e-05
```

Homogeneity among J multinomial distributions

Example 9 *cont'd*. Jane Austen left the novel *Sandition* unfinished when she died. An admirer completed the novel while emulating her style. Morton counted the occurrences of various words in several works.

H_0 : Admirer's usage of the words is consistent with Austin's. $\leftrightarrow$ H_1 : It is not.

Word	*Sense and Sensibility*	*Emma*	*Sanditon* I	*Sanditon* II	
a	147	186	101	83	517
an	25	26	11	29	91
this	32	39	15	15	101
that	94	105	37	22	258
with	59	74	28	43	204
without	18	10	10	4	42
Total	375	440	202	196	**1,213**

```
> chisq.test(Observed)

        Pearson's Chi-squared test

data:  Observed
X-squared = 45.578, df = 15, p-value = 6.205e-05
```

Homogeneity among J multinomial distributions

Example 9 *cont'd*. Jane Austen left the novel *Sandition* unfinished when she died. An admirer completed the novel while emulating her style. Morton counted the occurrences of various words in several works.

$$H_0: \text{ Admirer's usage of the words is consistent with Austin's.} \quad \leftrightarrow \quad H_1: \text{It is not.}$$

Examine the contributions to the chi-square statistic (or relative frequencies) cell by cell:

```
> (Observed-Expected)^2/Expected
      [,1]  [,2]  [,3]   [,4]
[1,] 1.030 0.013 2.580  0.003
[2,] 0.349 1.488 1.139 13.899
[3,] 0.019 0.152 0.197  0.107
[4,] 2.542 1.392 0.828  9.298
[5,] 0.262 0.000 1.050  3.056
[6,] 1.937 1.799 1.292  1.144
```

Homogeneity among J multinomial distributions

χ^2 test of homogeneity:

1. The population concerns J multinomial variables;
2. Tickets in each cell and across columns are independent;
3. Large sample size n so that no more than 20% of expected counts less than 5.

Next Tuesday ...

- Analysis of variance (ANOVA)
- Kruskal-Wallis test (Non-parametric)