To Do

Read Sections 7.1 – 7.3.

Do Problems 1-12

Assignment 5 is due Monday December 5.

Last Class

(1) Multinomial Likelihood Function

(2) Likelihood Ratio Goodness of Fit Test

(3) Pearson Goodness of Fit Test

Likelihood Ratio Goodness of Fit Test Statistic

$$\Lambda(\theta_0) = -2\log\left[\frac{L(\theta_0)}{L(\tilde{\theta})}\right]$$

$$= -2\log\left[\prod_{j=1}^{k}\left(\frac{E_j}{Y_j}\right)^{Y_j}\right]$$

$$= 2\sum_{j=1}^{k} Y_j \log\left(\frac{Y_j}{E_j}\right)$$

= 2*{sum [observed * log (observed /expected)]}

Observed Value

The observed value of the likelihood ratio test statistic is

$$\lambda(\theta_0) = 2\sum_{j=1}^{k} y_j \log\left(\frac{y_j}{e_j}\right)$$

If $e_j \geq 5$ for all j then *p*-value ≈ P($W \geq \lambda(\theta_0)$)

where

$W \backsim \chi^2$(k - 1- (no. parameters estimated assuming H_0 is true))

Today's Class

1) Two-Way Tables and Multinomial Models

2) Two-Way Tables and Testing for Independence of Two Variates

Bivariate Categorical Data

Data collected in January

PROGRAM/ HOMETOWN	Canadian Hometown	Non-Canadian Home town	Total
Computer Science	35	43	78
Non-Computer Science	18	69	87
Total	53	112	165

Is there a relationship between hometown and program?

Bivariate Categorical Data

Previously we summarized these data using relative risk as a numerical summary.

Proportion of CS students with Canadian hometown = 35/78 = 0.448.

Proportion of Non-CS students with Canadian hometown = 18/87 = 0.206.

The relative risk of a Canadian hometown among CS students as compared to non-CS students = (35/78)/ (18/87) = 2.17.

Two-Way Tables and Testing for Independence of Two Variates

Suppose *n* individuals are classified according to two different variates which have two possible values. The data can be displayed in a two way table:

	B	$\bar{B}$	Total
A	y_{11}	y_{12}	$r_1 = y_{11}+y_{12}$
$\bar{A}$	y_{21}	y_{22}	$n - r_1$
Total	$c_1 = y_{11}+y_{21}$	$n - c_1$	n

The row and column totals will be useful in calculating expected frequencies.

Model

Let Y_{11} = number of $A \cap B$ outcomes,
Y_{12} = number of $A \cap \bar{B}$ outcomes,
Y_{21} = number of $\bar{A} \cap B$ outcomes and
Y_{22} = number of $\bar{A} \cap \bar{B}$ outcomes.

Then

$(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \backsim \text{Multinomial}(n, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$

where $\theta_{11} = P(A \cap B)$, $\theta_{12} = P(A \cap \bar{B})$,
$\theta_{21} = P(\bar{A} \cap B)$, and $\theta_{21} = P(\bar{A} \cap \bar{B})$.

Hypothesis of Independence

The null hypothesis is that the variates *A* and *B* are independent, or

$$H_0: \mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

Let P(*A*) = α and P(*B*) = β then the null hypothesis may be written as

$$H_0: \theta_{11} = \alpha\beta.$$

Likelihood Function

The likelihood function (ignoring constants) is

$$L(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = \theta_{11}^{y_{11}} \theta_{12}^{y_{12}} \theta_{21}^{y_{21}} \theta_{22}^{y_{22}}$$

The maximum likelihood estimates are

$$\hat{\theta}_{ij} = \frac{y_{ij}}{n}, \ i = 1, 2; \ \ j = 1, 2$$

and the maximum likelihood estimators are

$$\tilde{\theta}_{ij} = \frac{Y_{ij}}{n}, \ i = 1, 2; \ \ j = 1, 2$$

Parameter Estimation under the Null Hypothesis

If H_0: $\theta_{11} = \alpha\beta$ is true, then the likelihood function under H_0 is

$$L(\alpha,\beta) = (\alpha\beta)^{y_{11}}\left[\alpha(1-\beta)\right]^{y_{12}}\left[(1-\alpha)\beta\right]^{y_{21}}\left[(1-\alpha)(1-\beta)\right]^{y_{22}}$$

$$= \alpha^{y_{11}+y_{12}}(1-\alpha)^{y_{21}+y_{22}}\,\beta^{y_{11}+y_{21}}(1-\beta)^{y_{y_{12}+y_{22}}}$$

$$0<\alpha<1,\ 0<\beta<1$$

Parameter Estimation under the Null Hypothesis

The maximum likelihood estimates under H_0: $\theta_{11} = \alpha\beta$ are

$$\hat{\alpha} = \frac{y_{11} + y_{12}}{n}, \quad \hat{\beta} = \frac{y_{11} + y_{21}}{n}$$

and the maximum likelihood estimators are

$$\tilde{\alpha} = \frac{Y_{11} + Y_{12}}{n}, \quad \tilde{\beta} = \frac{Y_{11} + Y_{21}}{n}$$

Parameter Estimation under the Null Hypothesis

Why do these estimates make sense?

The estimate of P(*A*) = *α* is

$$\hat{\alpha} = \frac{y_{11} + y_{12}}{n} = \frac{\text{no. of times outcome A occurred}}{n}$$

The estimate of P(*B*) = *β* is

$$\hat{\beta} = \frac{y_{11} + y_{21}}{n} = \frac{\text{no. of times outcome B occurred}}{n}$$

Likelihood Ratio Test Statistic

The likelihood ratio test for testing

H_0: θ_{11} = P($A \cap B$) = P(A)P(B) = $\alpha\beta$ is

$$\Lambda = -2\log\left[\frac{L(\tilde{\alpha},\tilde{\beta})}{L(\tilde{\theta}_{11},\tilde{\theta}_{12},\tilde{\theta}_{21},\tilde{\theta}_{22})}\right]$$

$$= 2\left[Y_{11}\log\left(\frac{Y_{11}}{E_{11}}\right) + Y_{12}\log\left(\frac{Y_{12}}{E_{12}}\right) + Y_{21}\log\left(\frac{Y_{21}}{E_{21}}\right) + Y_{22}\log\left(\frac{Y_{22}}{E_{22}}\right)\right]$$

which is of the form

2*{sum [observed * log (observed /expected)]}

Observed value of likelihood ratio test statistic

The observed value of the likelihood ratio test statistic is

$$\lambda = 2\left[y_{11}\log\left(\frac{y_{11}}{e_{11}}\right) + y_{12}\log\left(\frac{y_{12}}{e_{12}}\right) + y_{21}\log\left(\frac{y_{21}}{e_{21}}\right) + y_{22}\log\left(\frac{y_{22}}{e_{22}}\right) \right]$$

Note that $e_{11} = n\hat{\alpha}\hat{\beta} = n\left(\frac{r_1}{n}\right)\left(\frac{c_1}{n}\right) = \frac{r_1 c_1}{n}$

and the other expected frequencies can be obtained by subtraction from the row and column totals.

Observed [Expected]

	B	$\bar{B}$	Total
A	y_{11} [$e_{11} = r_1 \bullet c_1/n$]	y_{12} [$e_{12} = r_1 - e_{11}$]	$r_1 = y_{11}+y_{12}$
$\bar{A}$	y_{21} [$e_{21} = c_1 - e_{11}$]	y_{22} [$e_{22} = r_2 - e_{21}$]	$r_2 = y_{21}+y_{22}$
Total	$c_1 = y_{11}+y_{21}$	$n - c_1$	n

Degrees of Freedom for the Chi-squared Approximation

What are the degrees of freedom for the Chi-squared approximation?

How many parameters in the original model?

$(Y_{11}, Y_{12}, Y_{21}, Y_{22}) \backsim$ **Multinomial**$(n;\ \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$

How many parameters in the model assuming H_0: $\theta_{11} = \alpha\beta$ is true?

Why do the degrees of freedom make sense?

Approximate p-value

$$
\begin{aligned}
p-value &\approx P\left(W \geq \lambda\right) \quad \text{where } W \sim \chi^2(1) \\
&= 2\left[1 - P\left(Z \leq \sqrt{\lambda}\right)\right] \quad \text{where } Z \sim G(0,1)
\end{aligned}
$$

Example

Expected values in [brackets].

PROGRAM/ HOMETOWN	Canadian Hometown	Non-Canadian Home town	Total
Computer Science	**35** [(78*53)/165 = 25.05]	**43** [78 - 25.05 = 52.95]	**78**
Non-Computer Science	**18** [53 - 25.05 = 27.95]	**69** [87 – 27.95 = 59.05]	**87**
Total	**53**	**112**	**165**

Example

$$\lambda = 2[35\log\left(\frac{35}{25.05}\right) + 43\log\left(\frac{43}{52.95}\right)$$
$$+18\log\left(\frac{18}{27.95}\right) + 69\log\left(\frac{69}{57.05}\right)] = 11.15$$

$$p - value \approx 2\left[1 - P\left(Z \leq \sqrt{11.15}\right)\right] = 2[1 - P(Z \leq 3.34)]$$
$$= 0.00084$$

Since the *p*-value is less than 0.001 there is very strong evidence based on the data against the hypothesis that the variates hometown and program are independent.

Data Collected Last Week

	Canadian Hometown	Non-Canadian Hometown	Total
CS/Bioinfomatics	20	20	40
ACTSC/STAT/ FARM/DD/BUS	7	17	24
OTHER	8	16	24
Total	35	53	88

Is there relationship between hometown and program?

Larger Two-Way Tables

Individuals are classified according to each of two variates *A* and *B*.

For *A*, an individual can be any of *a* mutually exclusive types $A_1, A_2, \ldots, A_a$.

For B, an individual can be any of *b* mutually exclusive types $B_1, B_2, \ldots, B_b$.

Larger Two-Way Tables

Let Y_{ij} = the number that have A-type A_i and B-type B_j in a random sample of size n.

Let θ_{ij} be the probability a randomly selected individual is of type A_i and B_j.

$(Y_{11}, Y_{12}, \ldots, Y_{ab})$ has a

Multinomial$(n; \theta_{11}, \theta_{12}, \ldots, \theta_{ab})$ distribution.

Larger Two-Way Tables

The observed data can be arranged into an *a*×*b* two-way table:

	B_1	B_2	$\cdots$	B_b	Total
A_1	y_{11}	y_{12}	$\cdots$	y_{1b}	r_1
A_2	y_{21}	y_{22}	$\cdots$	y_{2b}	r_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_a	y_{a1}	y_{a2}	$\cdots$	y_{ab}	r_a
Total	c_1	c_2	$\cdots$	c_b	n

Hypothesis of Independence for a Two Way Table

Let α_i = P(an individual is type A_i)

β_j = P(an individual is type B_j)

To test whether *A* and *B* are independent variates we test

H$_0$: $\theta_{ij} = \alpha_i \beta_j$

for all i = 1,2,…,a and j = 1,2,…,b.

Expected Frequencies under Hypothesis of Independence

It can be shown that the expected frequencies under *H_0* are:

$$e_{ij} = \frac{r_i c_j}{n} \quad i = 1, 2, \ldots, a; \;\; j = 1, 2, \ldots, b$$

Likelihood Ratio Test Statistic

The likelihood ratio test statistic for testing the hypothesis of independence is

$$\Lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij} \log\left(\frac{Y_{ij}}{E_{ij}}\right)$$

with observed value

$$\lambda = 2\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} \log\left(\frac{y_{ij}}{e_{ij}}\right)$$

Data Collected Last Week

Expected values in [brackets].

	Canadian Hometown	Non-Canadian Hometown	Total
CS/Bioinfomatics	20 [40x35/88 = 15.91]	20 [40 - 15.91 = 24.09]	40
ACTSC/STAT/ FARM/DD/BUS	7 [40x35/88 = 9.55]	17 [24 – 9.55 = 14.45]	24
OTHER	8 [35 – 15.91- 9.55 = 9.55]	16 [24 – 9.55 = 14.45]	24
Total	35	53	88

Data Collected Last Week

Expected values in [brackets].

	Canadian Hometown	Non-Canadian Hometown	Total
CS/Bioinfomatics	20 [15.91]	20 [24.09]	40
ACTSC/STAT/ FARM/DD/BUS	7 [9.55]	17 [14.45]	24
OTHER	8 [9.55]	16 [14.45]	24
Total	35	53	88

$$\lambda = 2\sum_{i=1}^{3}\sum_{j=1}^{2} y_{ij}\log\left(\frac{y_{ij}}{e_{ij}}\right) = 3.3069$$

Data Collected Last Week

Expected values in [brackets].

	Canadian Hometown	Non-Canadian Hometown	Total
CS/Bioinfomatics	20 [15.91]	20 [24.09]	40
ACTSC/STAT/ FARM/DD/BUS	7 [9.55]	17 [14.45]	24
OTHER	8 [9.55]	16 [14.45]	24
Total	35	53	88

What are the degrees of freedom for the Chi-squared approximation?

Approximate Chi-squared Distribution and p-value

Under the hypothesis of independence Λ has approximately a

$\chi^2((a\text{-}1)\cdot(b\text{-}1))$ distribution

if *n* is reasonably large and the expected frequencies are all at least five.

Approximate Chi-squared Distribution and p-value

REMINDER:

1) If (*a*-1)•(*b*-1) = 1 then

$$p - value \approx P(W \geq \lambda) \quad \text{where } W \sim \chi^2(1)$$
$$= 2\left[1 - P\left(Z \leq \sqrt{\lambda}\right)\right] \text{ where } Z \sim G(0,1)$$

2) If (*a*-1) •(*b*-1) = 2 then

$$p - value \approx P(W \geq \lambda) = e^{-\lambda/2}$$

$$\text{where } W \sim \chi^2(2) = Exponential(2)$$

Data Collected Last Week

Expected values in [brackets].

	Canadian Hometown	Non-Canadian Hometown	Total
CS/Bioinfomatics	20 [15.91]	20 [24.09]	40
ACTSC/STAT/ FARM/DD/BUS	7 [9.55]	17 [14.45]	24
OTHER	8 [9.55]	16 [14.45]	24
Total	35	53	88

$$p-value \approx e^{-3.3069/2} = 0.1914$$

Conclusion

Since the *p*-value = 0.1914 there is no evidence based on the data against the null hypothesis of independence between the two variates hometown and program.