STAT 24400 Lecture 19
Section 9.5 Chi-Squared Tests for Multinomial Data

Yibi Huang
Department of Statistics
University of Chicago

Multinomial Distributions

Multinomial Distributions (Generalized Binomial)

If n trials are performed:

- in each trial there are $m > 2$ possible outcomes (categories)
- $p_i = \mathrm{P}(\text{category } i)$, for each trial, $\sum_{i=1}^{m} p_i = 1$
- trials are **independent**
- $X_i =$ number of trials fall in category i out of n trials

$(X_1, X_2, \ldots, X_m)$ is said to have the *multinomial distribution*, denoted as

$$(X_1, X_2, \ldots, X_m) \sim \text{Multinom}(n, p_1, p_2, \ldots, p_m).$$

with the joint PMF below

$$\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = \frac{n!}{x_1!\, x_2! \cdots x_m!}\, p_1^{x_1}\, p_2^{x_2} \cdots p_m^{x_m}$$

where $0 \le x_i \le n$ for all i and $\sum_{i=1}^{m} x_i = n$.

Example

Suppose proportions of individuals with genotypes AA, Aa, and aa in a large population are

$$(p_{AA}, p_{Aa}, p_{aa}) = (0.25, 0.5, 0.25).$$

Randomly sample $n = 5$ individuals from the population.

Example

Suppose proportions of individuals with genotypes AA, Aa, and aa in a large population are

$$(p_{AA}, p_{Aa}, p_{aa}) = (0.25, 0.5, 0.25).$$

Randomly sample $n = 5$ individuals from the population.

The chance of getting 2 AA's, 2 Aa's, and 1 aa is

$$\begin{aligned} \mathrm{P}(X_{AA} = 2, X_{Aa} = 2, X_{aa} = 1) &= \frac{5!}{2!\,2!\,1!} p_{AA}^2 p_{Aa}^2 p_{aa}^1 \\ &= \frac{5!}{2!\,2!\,1!} (0.25)^2 (0.5)^2 (0.25)^1 \approx 0.117 \end{aligned}$$

and the chance of getting no AA, 3 Aa's, and 2 aa's is

$$\mathrm{P}(X_{AA} = 0, X_{Aa} = 3, X_{aa} = 2) = \frac{5!}{0!\,3!\,2!} (0.25)^0 (0.5)^3 (0.25)^2 \approx 0.078$$

Properties of Multinomial Distributions

If $(X_1, X_2, \ldots, X_m)$ has a multinomial distribution with n trials and the category probabilities $(p_1, p_2, \cdots, p_m)$, then

- Each $X_i \sim \text{Binomial}(n, p_i)$
- $\text{E}(X_i) = np_i$ for $i = 1, 2, \ldots, m$
- $\text{Var}(X_i) = np_i(1 - p_i)$,
- X_i, X_j are not independent since $\sum_{i=1}^{m} X_i = n$.
- $\text{Cov}(X_i, X_j) = -np_ip_j$ $\leftarrow$ **Why negative?**

Properties of Multinomial Distributions

If $(X_1, X_2, \ldots, X_m)$ has a multinomial distribution with n trials and the category probabilities $(p_1, p_2, \cdots, p_m)$, then

- Each $X_i \sim \text{Binomial}(n, p_i)$
- $\text{E}(X_i) = np_i$ for $i = 1, 2, \ldots, m$
- $\text{Var}(X_i) = np_i(1 - p_i)$,
- X_i, X_j are not independent since $\sum_{i=1}^{m} X_i = n$.
- $\text{Cov}(X_i, X_j) = -np_ip_j$ $\leftarrow$ **Why negative?**
  - $X_1, \ldots, X_m$ are dependent since they must add up to n.

Likelihood Ratio Tests of Multinomial Data

MLE for Multinomial

Observe $(X_1, X_2, \ldots, X_m) \sim \text{Multinom}(n, p_1, p_2, \ldots, p_m)$, where n is known, but p_i's are unknown. What's the MLE for $(p_1, \ldots, p_m)$?

- likelihood:

$$L(p_1, \ldots, p_m) = \frac{n!}{X_1!\, X_2! \cdots X_m!}\, p_1^{X_1}\, p_2^{X_2} \cdots p_m^{X_m}$$

- log-likelihood:

$$\ell(p_1, \ldots, p_m) = C + \sum_{i=1}^{m} X_i \log(p_i)$$

  where $C = \log(\frac{n!}{X_1!\, X_2! \cdots X_m!})$ includes terms not involving p_i's.

- However, we CANNOT find the MLE as usual by solving

$$0 = \frac{\partial \ell}{\partial p_i} = \frac{X_i}{p_i}, \quad i = 1, \ldots, m,$$

  due to the constraint $\sum_{i=1}^{m} p_i = 1$.

MLE for Multinomial (Lagrange Multiplier)

To maximize this likelihood subject to the constraint $\sum_{i=1}^m p_i = 1$, we introduce a *Lagrange multiplier*

$$\ell(p_1, \ldots, p_m; \lambda) = C + \sum_{i=1}^m X_i \log(p_i) - \lambda \left(\sum_{i=1}^m p_i - 1 \right).$$

Then we find $(p_1, \ldots, p_m, \lambda)$ that maximize the Lagrange multiplier by taking its partial derivative with respect to each p_i and to λ

$$\begin{cases} 0 = \frac{\partial \ell}{\partial p_i} = \frac{X_i}{p_i} - \lambda, \quad i = 1, \ldots, m \\ 0 = \frac{\partial \ell}{\partial \lambda} = -(\sum_{i=1}^m p_i) + 1 \end{cases}$$

The first m equations give $p_i = X_i/\lambda$. Plugging $p_i = X_i/\lambda$ into the last equation, we get,

$$1 = \sum_{i=1}^m p_i = \frac{\sum_{i=1}^m X_i}{\lambda} \Rightarrow \lambda = \sum_{i=1}^m X_i = n \Rightarrow p_i = \frac{X_i}{n}.$$

$\Rightarrow$ The MLE is $(\widehat{p}_1, \ldots, \widehat{p}_m) = \left(\frac{X_1}{n}, \ldots, \frac{X_m}{n} \right)$

Likelihood Ratio Tests of Multinomial Data

Suppose we observe $(X_1, \ldots, X_m) \sim \text{Multinom}(n, p_1, \ldots, p_m)$ and wish to test

- H_0: $(p_1, \ldots, p_m) = (p_{10}, \ldots, p_{m0})$ against
- H_1: $(p_1, \ldots, p_m) \neq (p_{10}, \ldots, p_{m0})$

We can conduct a generalized likelihood ratio test.

- Likelihood: $L(p_1, \ldots, p_m) = C \prod_{i=1}^m p_i^{X_i}$
  where $C = \frac{n!}{X_1! X_2! \cdots X_m!}$ includes terms not involving p_i's.
- Under H_0: max $L(p_1, \ldots, p_m)$ is simply $L(p_{10}, \ldots, p_{m0})$
- Under H_0 or H_1, max $L(p_1, \ldots, p_m)$ is $L(\frac{X_1}{n}, \ldots, \frac{X_m}{n})$.
- The GLR is thus

$$\Lambda = \frac{L(p_{10}, \ldots, p_{m0})}{L(\frac{X_1}{n}, \ldots, \frac{X_m}{n})} = \frac{C \prod_{i=1}^m p_{i0}^{X_i}}{C \prod_{i=1}^m (X_i/n)^{X_i}} = \prod_{i=1}^m \left(\frac{np_{i0}}{X_i} \right)^{X_i}.$$

Likelihood Ratio Tests of Multinomial (2)

According to the large-sample theory of GLR, when n is large,

$$-2\log\Lambda = -2\sum_{i=1}^{m} X_i \log\left(\frac{np_{i0}}{X_i}\right) = 2\sum_{i=1}^{m} X_i \log\left(\frac{X_i}{np_{i0}}\right)$$
$$= 2\sum_{i=1}^{m} O_i \times \log\left(\frac{O_i}{E_i}\right) \text{ is approx. } \sim \chi^2_{m-1} \text{ under } \mathrm{H}_0,$$

where $O_i = X_i =$ observed count in the ith category

$E_i = np_{i0} =$ expected count in the ith category under H_0

There are $m-1$ degrees of freedom since

- under H_1, there are $m-1$ free parameters $(p_1, \ldots, p_m)$, subject to the constraint $\sum_i p_i = 1$;
- under H_0, there is no free parameters as all parameters are specified.

Pearson's Chi-Squared Statistic for Multinomial

For multinomial data, the likelihood ratio test statistic is usually refer to as G^2

$$G^2 = 2\sum_{i=1}^{m} O_i \log\left(\frac{O_i}{E_i}\right)$$

Another (more) commonly used test statistic for multinomial data is *Pearson's Chi-Squared statistic*,

$$\text{Pearson's } X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \sim \text{approx. } \chi^2_{m-1}.$$

- When $\frac{O_i - E_i}{E_i} \approx 0$, Pearson's X^2 and G^2 are usually close.
- The sampling distribution of Pearson's X^2 converges to **chi-square** faster than that of G^2. Hence, Pearson's X^2 is more commonly used than G^2.
- The larger the value of Pearson's X^2 or G^2, the stronger the evidence against H_0
- If $O_i = E_i$ for all i, then Pearson's $X^2 = 0$, $G^2 = 0$.

Proof of $G^2 \approx X^2$

By Taylor expansion, $\log(1+x) \approx x - \frac{x^2}{2}$ when $x \approx 0$, we have

$$\log\left(\frac{O_i}{E_i}\right) = \log\left(1 + \frac{O_i - E_i}{E_i}\right) \approx \frac{O_i - E_i}{E_i} - \frac{1}{2}\frac{(O_i - E_i)^2}{E_i^2} \quad \text{when } \frac{O_i - E_i}{E_i} \approx 0.$$

and thus

$$\begin{aligned} O_i \log\left(\frac{O_i}{E_i}\right) &\approx [E_i + (O_i - E_i)]\left(\frac{O_i - E_i}{E_i} - \frac{1}{2}\frac{(O_i - E_i)^2}{E_i^2}\right) \\ &= (O_i - E_i) - \frac{1}{2}\frac{(O_i - E_i)^2}{E_i} + \frac{(O_i - E_i)^2}{E_i} - \frac{1}{2}\frac{(O_i - E_i)^3}{E_i^2} \\ &= (O_i - E_i) + \frac{1}{2}\frac{(O_i - E_i)^2}{E_i}\left(1 - \frac{(O_i - E_i)}{E_i}\right) \end{aligned}$$

Summing over i, we get

$$G^2 = 2\sum_i O_i \log\left(\frac{O_i}{E_i}\right) \approx 2\underbrace{\sum_i (O_i - E_i)}_{=0} + \sum_i \frac{(O_i - E_i)^2}{E_i}\left(1 - \underbrace{\frac{(O_i - E_i)}{E_i}}_{\approx 0}\right)$$

$$\approx \sum_i \frac{(O_i - E_i)^2}{E_i} = X^2$$

Example: Seasonal Variation of Suicide Rates

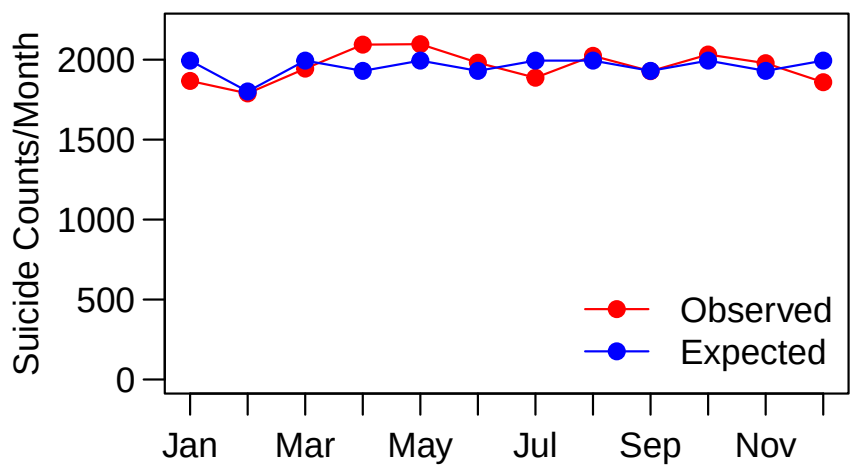
US Monthly Suicide Counts (1970)

Month	Num. of Suicides	Days/ Month	Expected count
Jan	1867	31	
Feb	1789	28	1801.205
Mar	1944	31	1994.192
Apr	2094	30	1929.863
May	2097	31	1994.192
Jun	1981	30	1929.863
July	1887	31	1994.192
Aug	2024	31	1994.192
Sept	1928	30	1929.863
Oct	2032	31	1994.192
Nov	1978	30	1929.863
Dec	1859	31	1994.192
Total	23480	365	23480

Does the suicide rate vary seasonally, or is it uniform from day to day?

If uniform from day to day, we expect 31/365 of the suicides to occur in January.

$$\text{(total number of suicides)} \times \frac{31}{365} = 1994.192.$$



Example: Seasonal Variation of Suicide Rates

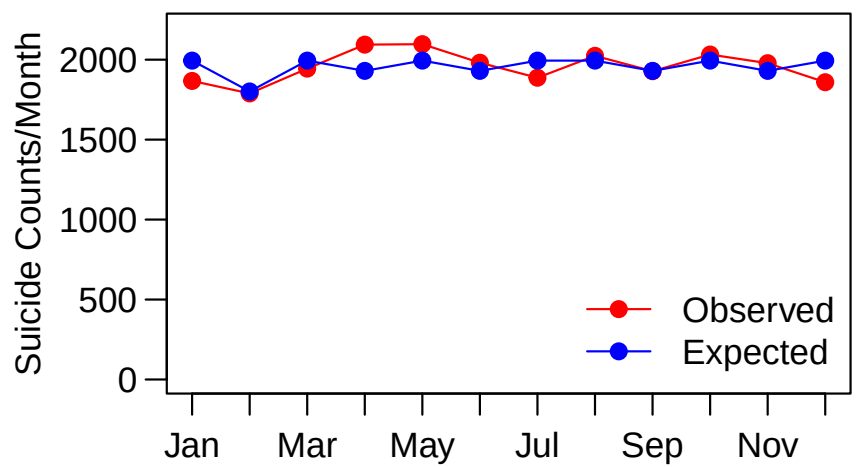
US Monthly Suicide Counts (1970)

Month	Num. of Suicides	Days/ Month	Expected count
Jan	1867	31	1994.192
Feb	1789	28	1801.205
Mar	1944	31	1994.192
Apr	2094	30	1929.863
May	2097	31	1994.192
Jun	1981	30	1929.863
July	1887	31	1994.192
Aug	2024	31	1994.192
Sept	1928	30	1929.863
Oct	2032	31	1994.192
Nov	1978	30	1929.863
Dec	1859	31	1994.192
Total	23480	365	23480

Does the suicide rate vary seasonally, or is it uniform from day to day?

If uniform from day to day, we expect 31/365 of the suicides to occur in January.

$$\text{(total number of suicides)} \times \frac{31}{365} = 1994.192.$$



Example: Seasonal Variation of Suicide Rates

Likelihood Ratio test statistic is

$$2\left[1867 \log\left(\frac{1867}{1994.192}\right) + 1789 \log\left(\frac{1789}{1801.205}\right) + \ldots + 1859 \log\left(\frac{1859}{1994.192}\right)\right] \approx 47.378$$

Pearson's X^2-statistic is

$$X^2 = \frac{(1867 - 1994.192)^2}{1994.192} + \frac{(1789 - 1801.205)^2}{1801.205} + \ldots + \frac{(1859 - 1994.192)^2}{1994.192} \approx 47.365$$

Both have $12 - 1 = 11$ degrees of freedom. Both P-values are ≈ 0.00000185, meaning the suicide rate is not uniform from day to day.

```
pchisq(47.378, df=11, lower.tail=FALSE)
[1] 0.000001842
pchisq(47.365, df=11, lower.tail=FALSE)
[1] 0.000001852
```

Example: Hardy-Weinberg Equilibrium

In fact, we don't have to fully specify p_i in the H_0. We can specify p_i's with a few parameter(s) θ like

- H_0: $(p_1, \ldots, p_m) = (p_1(\theta), \ldots, p_m(\theta))$

For example, Hardy-Weinberg Equilibrium assumes proportions of individuals with genotypes AA, Aa, and aa in a large population are

$$(p_{AA}, p_{Aa}, p_{aa}) = ((1-\theta)^2, 2\theta(1-\theta), \theta^2),$$

where θ is an unknown constant.

Let (X_1, X_2, X_3) be the counts of the genotypes AA, Aa, and aa in a random sample of n individuals. Then

$$(X_1, X_2, X_3) \sim \text{Multinom}((1-\theta)^2, 2\theta(1-\theta), \theta^2)$$

What's the MLE for θ?

Example: Hardy-Weinberg Equilibrium

Likelihood:

$$L(\theta) = C[(1-\theta)^2]^{X_1}[2\theta(1-\theta)]^{X_2}[\theta^2]^{X_3} = C2^{X_2}(1-\theta)^{2X_1+X_2}.\theta^{X_2+2X_3}.$$

Log-likelihood:

$$\ell(\theta) = \text{const} + (2X_1 + X_2)\log(1-\theta) + (X_2 + 2X_3)\log\theta$$

Solve for MLE

$$0 = \frac{\partial \ell}{\partial \theta} = -\frac{2X_1 + X_2}{1-\theta} + \frac{X_2 + 2X_3}{\theta}.$$

We can obtain the MLE

$$\widehat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3} = \frac{X_2 + 2X_3}{2n}.$$

Chi-Square Test of Multinomial

To test

- H_0: $(p_1, \ldots, p_m) = (p_1(\theta), \ldots, p_m(\theta))$, against
- H_1: $(p_1, \ldots, p_m)$ is not as specified in H_0,

we can also test using

$$-2\log\Lambda = 2\sum_{i=1}^{m} O_i \times \log\left(\frac{O_i}{E_i}\right) \text{ or Pearson's } X^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

where

$O_i = X_i =$ observed count in the ith category

$E_i = np_i(\widehat{\theta}) =$ expected count in the ith category under H_0

and $\widehat{\theta}$ is the MLE of θ.
Under H_0, both $-2\log\Lambda$ and X^2 are approx. $\sim \chi^2_k$ where

$k =$ (# of free parameters in H_1) $-$ (# of free parameters in H_0).

Example: Hardy-Weinberg Equilibrium

In a sample from the Chinese population of Hong Kong in 1937, blood types occurred with the following frequencies, where M and N are erythrocyte antigens:

	Blood Type			
	M	MN	N	Total
Count	342	500	187	1029

The MLE for θ is thus

$$\widehat{\theta} = \frac{X_2 + 2X_3}{2n} = \frac{500 + 2 \cdot 187}{2 \cdot 1029} = \frac{874}{2058} \approx 0.4247.$$

The expected counts for the 3 blood types are thus $np_i(\widehat{\theta})$

$$E_1 = n(1 - \widehat{\theta})^2 \approx 1029(1 - 0.4247)^2 \approx 340.6$$

$$E_2 = n2\widehat{\theta}(1 - \widehat{\theta}) \approx 1029(2)(0.4247)(1 - 0.4247) \approx 502.8$$

$$E_3 = n\widehat{\theta}^2 \approx 1029(0.4247)^2 \approx 185.6$$

Example: Hardy-Weinberg Equilibrium

	Blood Type			
	M	MN	N	Total
Observed Count	342	500	187	1029
Expected Count	340.6	502.8	185.6	

Likelihood Ratio and Pearson's X^2-statistic are respectively

$$-2\log\Lambda = 2\left[342\log\left(\frac{342}{340.6}\right) + 500\log\left(\frac{500}{502.8}\right) + 187\log\left(\frac{187}{185.6}\right)\right]$$
$$\approx 0.0319,$$
$$X^2 = \frac{(342-340.6)^2}{340.6} + \frac{(500-502.8)^2}{502.8} + \frac{(187-185.6)^2}{185.6} \approx 0.0319$$

They have $2-1=1$ degree of freedom since

- under H_1, there are $3-1=2$ free parameters
- under H_0, there is 1 free parameter θ

Example — Fatalities From Horse Kicks (p.45, Textbook)

The # of deaths in a year resulted from being kicked by a horse or mule was recorded for each of 10 corps of Prussian cavalry over a period of 20 years, giving 200 corps-years worth of data.

# of Deaths (in a corp in a year)	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

The count of deaths due to horse kicks in a corp in a given year may have a Poisson distribution because

- $p = P(\text{a soldier died from horsekicks in a given year}) \approx 0$;
- $n = \#$ of soldiers in a corp was large (100's or 1000's);
- whether a soldier was kicked was (at least nearly) independent of whether others were kicked

Example (Fatalities From Horse Kicks — Cont'd)

The fitted Poisson probability to have k deaths from horsekicks for $\widehat{\lambda} = 0.61$ is

$$\mathrm{P}(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-0.61}\frac{(0.61)^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

k	Observed Frequency	Relative Frequency	Poisson Probability
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003
Total	200	1	0.999

Recall the MLE $\widehat{\lambda} = 0.61$ is the sample mean, i.e., the average of the 200 counts

$$\frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61.$$

Example (Fatalities From Horse Kicks — Cont'd)

Can check accuracy of Poisson fit by LRT or Pearson's X^2.

- H_0: $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, 3$ and $p_4 = \sum_{i=4}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}$
- H_1: p_k is not as specified above.

k	Observed Count	Expected Count
0	109	108.67
1	65	66.29
2	22	20.22
3	3	4.11
4	1	0.71
Total	200	200

$\Rightarrow$ Pearson's $X^2 \approx 0.60$, $-2\log\Lambda \approx 0.61$.

They have $4 - 1 = 3$ degrees of freedom since

- under H_1, there are $5 - 1 = 4$ free parameters
- under H_0, there is 1 free parameter λ

The P-values are both ≈ 0.89, meaning the Poisson fit is good.

Test For Independence for Two-Way Contigency Tables

Setting

Suppose X and Y are two discrete random variables.

- X takes values $1, 2, \ldots, r$, and
- Y takes values $1, 2, \ldots, c$.

Denote the joint PMF of (X, Y) as

$$p_{ij} = \mathrm{P}(X = i, Y = j), \quad \text{for } i = 1, \ldots, r, \text{ and } j = 1, \ldots, c.$$

The marginal PMF of X and of Y are then respectively

$$\mathrm{P}(X = i) = \sum_j \mathrm{P}(X = i, Y = j) = \sum_j p_{ij} \overset{\text{define}}{=} p_{i+}, \quad i = 1, \ldots, r$$

$$\mathrm{P}(Y = j) = \sum_i \mathrm{P}(X = i, Y = j) = \sum_i p_{ij} \overset{\text{define}}{=} p_{+j}, \quad j = 1, \ldots, c.$$

In this notation, X and Y are independent if and only if

$$p_{ij} = p_{i+} p_{+j} \quad \text{for all } i, j.$$

Two-Way Contigency Tables

Suppose we observe n i.i.d. pairs of (X, Y) variables from the joint distribution on the previous slide.

$$(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n).$$

Let N_{ij} = the number of XY-pairs such that $(X = i, Y = j)$.
The data are usually summarized as a *2-way contingency table*.

count	$Y = 1$	$Y = 2$	$\cdots$	$Y = c$	row total
$X = 1$	N_{11}	N_{12}	$\cdots$	N_{1c}	N_{1+}
$X = 2$	N_{21}	N_{22}	$\cdots$	N_{2c}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X = r$	N_{r1}	N_{r2}	$\cdots$	N_{rc}	N_{r+}
column total	N_{+1}	N_{+2}	$\cdots$	N_{+c}	$N_{++} = n$

Here N_{ij}'s are refer to as the *cell counts*, while $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$ are respectively the *row total* and the *column total*.

Distribution of Counts in a Two-Way Contigency Table

Observe that the cell counts have the multinomial distribution.

$$\begin{pmatrix} N_{11} & N_{12} & \cdots & N_{1c} \\ N_{21} & N_{22} & \cdots & N_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ N_{r1} & N_{r2} & \cdots & N_{rc} \end{pmatrix} \sim \text{Multinom}\left(n, \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rc} \end{pmatrix} \right)$$

The row totals and the column totals also have multinomial distributions.

$$(N_{1+}, N_{2+}, \ldots, N_{r+}) \sim \text{Multinom}(n, p_{1+}, p_{2+}, \ldots, p_{r+}),$$
$$(N_{+1}, N_{+2}, \ldots, N_{+c}) \sim \text{Multinom}(n, p_{+1}, p_{+2}, \ldots, p_{+c}).$$

Likelihood Ratio Test Statistic of Independence

Our goal is to find the likelihood ratio test statistic for testing

H_0: X and Y are independent vs H_1: X and Y are dependent

As shown earlier, the MLE for p_{ij} in general is $\widehat{p}_{ij} = N_{ij}/n$. The likelihood for $(p_{11}, \ldots, p_{rc})$ is

$$L(p_{11}, \ldots, p_{rc}) = C \prod_{i=1}^{r} \prod_{j=1}^{c} p_{ij}^{N_{ij}}.$$

The maximized likelihood in general is thus

$$L(\frac{N_{11}}{n}, \ldots, \frac{N_{rc}}{n}) = C \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{N_{ij}}{n}\right)^{N_{ij}}.$$

It remains to find the max likelihood under the H_0 of independence.

Under H_0, we know $p_{ij} = p_{i+}p_{+j}$ for all i, j.
There are only $r + c$ parameters. Their likelihood is

$$\begin{aligned} L(p_{i+}, \ldots, p_{+j}, \ldots) &= C \prod_{i=1}^{r} \prod_{j=1}^{c} (p_{i+}p_{+j})^{N_{ij}} \\ &= C \Big(\prod_{i=1}^{r} \prod_{j=1}^{c} p_{i+}^{N_{ij}} \Big) \Big(\prod_{j=1}^{c} \prod_{i=1}^{r} p_{+j}^{N_{ij}} \Big) \\ &= C \Big(\prod_{i=1}^{r} p_{i+}^{\sum_{j=1}^{c} N_{ij}} \Big) \Big(\prod_{j=1}^{c} p_{+j}^{\sum_{i=1}^{r} N_{ij}} \Big) \\ &= C \Big(\prod_{i=1}^{r} p_{i+}^{N_{i+}} \Big) \Big(\prod_{j=1}^{c} p_{+j}^{N_{+j}} \Big), \end{aligned}$$

which is the product of a function of $(p_{1+}, \ldots, p_{r+})$ and a function of $(p_{+1}, \ldots, p_{+r})$. We can maximize them separately with the MLE's

$$\widehat{p}_{i+} = N_{i+}/n, \quad \widehat{p}_{+j} = N_{+j}/n.$$

The max likelihood under H_0 is thus

$$L(\widehat{p}_{i+}, \ldots, \widehat{p}_{+j}, \ldots) = C \prod_{i=1}^{r} \prod_{j=1}^{c} (\widehat{p}_{i+}\widehat{p}_{+j})^{N_{ij}} = C \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{N_{+j}N_{+j}}{n^2} \right)^{N_{ij}}.$$

The generalized likelihood ratio is thus

$$\Lambda = \frac{C \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{N_{+j} N_{+j}}{n^2} \right)^{N_{ij}}}{C \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{N_{ij}}{n} \right)^{N_{ij}}} = \prod_{i=1}^{r} \prod_{j=1}^{c} \left(\frac{N_{+j} N_{+j}}{n N_{ij}} \right)^{N_{ij}},$$

and the generalized likelihood ratio test statistic of independence if

$$\begin{aligned} G^2 = -2 \log \Lambda &= 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \left(\frac{n N_{ij}}{N_{+j} N_{+j}} \right) \\ &= 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \log \left(\frac{O_{ij}}{E_{ij}} \right) \end{aligned}$$

where $O_{ij} = N_{ij} =$ observed count in the (i, j) cell, and

$$\begin{aligned} E_{ij} = \frac{N_{+j} N_{+j}}{n} &= \frac{(\text{row total})(\text{column total})}{\text{overall total}} \\ &= \text{expected count in the } (i, j) \text{ cell under } \mathrm{H}_0 \end{aligned}$$

Degrees of Freedom for LR Test Statistic of Independence

According to the large-sample theory of GLR, when n is large,

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right) \text{ is approx. } \sim \chi^2_{(r-1)(c-1)} \text{ under } \mathrm{H}_0.$$

It has $(r-1)(c-1)$ degrees of freedom since

- under H_1, there are $rc-1$ free parameters (p_{ij}), subject to the constraint $\sum_{ij} p_{ij} = 1$;
- under H_0, the joint PMF $\{p_{ij}\}$ are completely determined by the marginal PMF $\{p_{i+}\}$ and $\{p_{+j}\}$ since $p_{ij} = p_{i+}p_{+j}$.
  - there are $(r-1)$ free parameters for $\{p_{i+}\}$ subject to the constraint $\sum_i p_{i+} = 1$;
  - there are $(c-1)$ free parameters for $\{p_{+j}\}$ subject to the constraint $\sum_j p_{+i} = 1$;

Thus $df = (rc-1) - [(r-1)+(c-1)] = (r-1)(c-1)$.

Pearson's X^2-Statistic

Similarly, we can also test the H_0 of independence using

$$\text{Pearson's } X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

where O_{ij} and E_{ij} are as defined for G^2.

- When $\frac{O_{ij}-E_{ij}}{E_{ij}} \approx 0$ for all cell, X^2 and G^2 are usually close.
- The sampling distribution of Pearson's X^2 converges to **chi-square** faster than that of G^2. Hence, Pearson's X^2 is more commonly used than G^2.
- The larger the value of Pearson's X^2 or G^2, the stronger the evidence against H_0
- If $O_i = E_i$ for all i, then Pearson's $X^2 = 0$, $G^2 = 0$.

Example: Age & Source of News

A question asked in the 2008 General Social Survey is "*Where do you get most of your information about current news events?*" Possible answers included TV, Internet, and newspapers, as well as other possibilities such as radio, family, and friends. The table below summarizes the results by age group.

	Source (Y)				
Age (X)	TV	Internet	Newspapers	Other	Total
18-29	109	92	25	36	262
30-49	272	157	88	63	580
50+	345	59	165	63	632
Total	726	308	278	162	1474

Question: Did the way people get news change with age?

Expected Counts

The expected counts for the Age and News data are

Age (X)	Source (Y) TV	Internet	Newspapers	Other	Total
18-29	$\frac{262\times726}{1474}=129.04$	$\frac{262\times308}{1474}=54.75$	$\frac{262\times278}{1474}=49.41$	$\frac{262\times162}{1474}=28.8$	262
30-49	$\frac{580\times726}{1474}=285.67$	$\frac{580\times308}{1474}=121.19$	$\frac{580\times278}{1474}=109.39$	$\frac{580\times162}{1474}=63.74$	580
50+	$\frac{632\times726}{1474}=311.28$	$\frac{632\times308}{1474}=132.06$	$\frac{632\times278}{1474}=119.2$	$\frac{632\times162}{1474}=69.46$	632
Total	726	308	278	162	1474

Note the expected cell counts need NOT be **whole numbers**.
Do NOT round the expected counts to integers.

G^2- and X^2-Statistic — Age News Example

The observed counts and the expected counts (in parentheses)

Age (X)	Source (Y) TV	Internet	Newspapers	Other	Total
18-29	109 (129.04)	92 (54.75)	25 (49.41)	36 (28.8)	262
30-49	272 (285.67)	157 (121.19)	88 (109.39)	63 (63.74)	580
50+	345 (311.28)	59 (132.06)	165 (119.2)	63 (69.46)	632
Total	726	308	278	162	1474

The likelihood ratio G^2 and Pearson's X^2 are respectively

$$G^2 = 2\left[109\log\left(\frac{109}{129.04}\right) + 92\log\left(\frac{92}{54.75}\right) + \cdots + 63\log\left(\frac{63}{69.46}\right)\right]$$
$$\approx 126.4471$$

$$X^2 = \frac{(109-129.04)^2}{129.04} + \frac{(92-54.75)^2}{54.75} + \cdots + \frac{(63-69.46)^2}{69.46}$$
$$\approx 120.0253$$

P-value of Pearson's X^2-Test — Age News Example

The table is 3×4, so

$$df = (r-1)(c-1) = (3-1)(4-1) = 6$$

The P-value is

- $\mathrm{P}(G^2 > 126.4471) \approx 7.192 \times 10^{-25}$ for $G^2 = 126.4471$
- $\mathrm{P}(X^2 > 120.0253) \approx 1.62 \times 10^{-23}$ for $X^2 = 120.0253$.

```
pchisq(126.4471, df=6, lower.tail=FALSE)
[1] 7.192e-25
pchisq(120.0253, df=6, lower.tail=FALSE)
[1] 1.61e-23
```

There is strong evidence against H_0:

people in different age groups had significantly different preference in ways of getting news.