Inference for Multinomial Parameters

Dipankar Bandyopadhyay, Ph.D.

Department of Biostatistics,
Virginia Commonwealth University

The Multinomial Distribution

Suppose we are sampling from a population Ω which contains c types of objects for which π_i equals the probability an object selected at random is of type i for $i = 1, 2, \ldots, c$.

Now, suppose we draw a simple random sample of size n from Ω and classify the objects into the c categories.

Then, we could summarize our sample using the following table.

	Population Categories				
	1	2	$\cdots$	c	Totals
Cell Probabilities	π_1	π_2	$\cdots$	π_c	1
Obs. Frequencies	n_1	n_2	$\cdots$	n_c	n

We will want to develop statistical tests to draw inference on the parameters π_i.

Inference for a Multinomial Parameter

- Suppose n observations are classified into c categories according to the probability vector $\vec{\pi} = (\pi_1, \pi_2, \cdots, \pi_c)$.
- The joint distribution for $n_1, n_2, \cdots, n_c$ is given by

$$P(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

subject to the following constraints

$$\sum_{i=1}^{c} n_i = n$$

$$\sum_{i=1}^{c} \pi_i = 1$$

- We want to find the MLE of $\vec{\pi}$.

Multinomial Coefficient

- The coefficient $\left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)$ is the number of ways to group n objects into c categories.
- You can easily "prove" this coefficient by considering the following:

$$
\begin{aligned}
P\left(\begin{array}{l}\text{Arranging n}\\ \text{objects into}\\ \text{ c categories}\end{array}\right) &= P\left(\begin{array}{l}\text{Selecting}\\ n_1 \text{ objects}\\ \text{from n}\end{array}\right) \times P\left(\begin{array}{l}\text{Selecting}\\ n_2 \text{ objects}\\ \text{from } n-n_1\end{array}\right) \times \\
&\quad \cdots \times P\left(\begin{array}{l}\text{Selecting}\\ n_c \text{ objects}\\ \text{from } n-n_1-\cdots-n_{c-1}\end{array}\right) \\
&= \binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\cdots-n_{c-1}}{n_c} \\
&= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-(n_1+n_2))}\cdots\frac{(n-n_1-n_2-\cdots n_{c-1})!}{n_c!(n-n)!} \\
&= \frac{n!}{n_1!n_2!\cdots n_c!}
\end{aligned}
$$

Multinomial Likelihood

Let the multinomial likelihood be defined as

$$L(n_1, n_2, \cdots, n_{c-1}, \pi_1, \cdots, \pi_c) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

with a log likelihood of

$$\begin{aligned} l(\cdot) &= \log\{\left(\tfrac{n!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}\} \\ &= k + \sum_{i=1}^{c} n_i \log\{\pi_i\} \end{aligned}$$

To maximize $l(\cdot)$ subject to the constraint $\sum \pi_i = 1$, we will use Lagrange's multiplier.

Lagrange's Multiplier in a nut shell

- Suppose you want to maximize function $f(n, y)$ subject to the constraint $h(n, y) = 0$
- You can define a new function $G(n, y, \lambda)$ to be

$$G(n, y, \lambda) = f(n, y) - \lambda h(n, y)$$

- λ is called Lagrange's Multiplier
- You take differentials of G w.r.t. both the π and λ.

Lagrange's Applied to the Multinomial

Let

$$G = \sum_{i=1}^{c} n_i log\{\pi_i\} - \lambda(\sum_{i=1}^{n} \pi_i - 1)$$

where the first part of G represents the kernel of the likelihood and λ is the Lagrange multiplier.

To maximize G, we will take the partial derivatives and set them to zero.

$$\frac{\partial G}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda$$

$$\frac{\partial G}{\partial \lambda} = -(\sum_{i=1}^{n} \pi_i - 1)$$

Setting

$$\frac{\partial G}{\partial \pi_j} = \frac{\partial G}{\partial \lambda} = 0$$

yields

$$\frac{n_j}{\widehat{\pi}_j} - \widehat{\lambda} = 0 \qquad\qquad (\sum \widehat{\pi}_i - 1) = 0$$

$$\widehat{\pi}_j = \frac{n_j}{\widehat{\lambda}} \qquad\qquad \sum \widehat{\pi}_i = 1$$

or $n_j = \widehat{\pi}_j \widehat{\lambda}$

Since $\sum n_i = n$ and $n_j = \widehat{\pi}_j \widehat{\lambda}$,

$$\sum_{i=1}^{c} n_i = \sum_{i=1}^{c} \widehat{\pi}_j \widehat{\lambda} = n$$
$$\widehat{\lambda} \sum_{i=1}^{c} \widehat{\pi}_j = n$$
$$\widehat{\lambda} = n$$

$$\therefore \widehat{\pi}_j = \frac{n_j}{n}$$

Exact Multinomial Test (EMT)

Suppose you want to test the hypothesis

$$H_0 : \pi_j = \pi_{j0}, \forall j \in \{1, 2, \ldots, c\}$$

where $\sum \pi_j = 1$.
Let $\vec{n}$ be the vector of observed counts. To calculate the exact probability of observing this configuration, use the multinomial PDF.

That is,

$$P(\vec{n}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

The exact P-value is then defined as the sum of all of the probabilities as extreme or more extreme than the observed sample when all possible configurations are enumerated.

Example EMT

- Suppose you have a population with 3 categories ($c = 3$)
- Let the true population probabilities be $\vec{\pi} = \{0.1, 0.2, 0.7\}$
- We want to test $H_0 : \vec{\pi} = \{0.1, 0.2, 0.7\}$ by drawing a random sample of size 3 ($n = 3$).

Let $\vec{n} = \{2,\ 0\ ,\ 1\}$, then the $P(\vec{n}) = 0.0210$

We will want to calculate the probabilities of the other configurations.

You can calculate all of these by hand, but the following SAS program can help.

SAS Code

```
DATA MULT3;
N=3;
P1=.1;  P2=.2; P3=.7;
DO N1=0 TO N;
DO N2=0 TO (N-N1);
N3=N-(N1+N2);
DEN=LGAMMA(N1+1)+LGAMMA(N2+1)+LGAMMA(N3+1);
NUM=(N1*LOG(P1))+(N2*LOG(P2))+(N3*LOG(P3))+LGAMMA(N+1);
PRO=NUM-DEN;
PROB=EXP(PRO);
OUTPUT;
END;
END;
```

```
PROC SORT; BY PROB; RUN;
DATA NEW;
SET MULT3;
CUM+PROB;
RUN;

PROC PRINT NOOBS;
VAR N1 N2 N3 PROB CUM;
FORMAT PROB CUM 7.4;
RUN;
```

N1	N2	N3	PROB	CUM	
3	0	0	0.0010	0.0010	
2	1	0	0.0060	0.0070	
0	3	0	0.0080	0.0150	
1	2	0	0.0120	0.0270	
2	0	1	0.0210	0.0480	<--- Observed Sample
0	2	1	0.0840	0.1320	
1	1	1	0.0840	0.2160	
1	0	2	0.1470	0.3630	
0	1	2	0.2940	0.6570	
0	0	3	0.3430	1.0000	

Therefore, the calculated exact probability is 0.048 and at the $\alpha = .05$ level of significance, we would reject H_0.

Limitations of EMT

Enumeration of the permutations of the sample size can be cumbersome for large n or c.

In general, there are

$$M = \binom{n+c-1}{c-1}$$

possible configurations.

Table of Possible Configurations

	Sample Size (n)			
c	5	10	20	50
3	21	66	231	1326
5	126	1001	10,626	316,251
10	2002	92,378	100,015,005	$> 10^9$
20	42,504	20,030,010	¿6 x 10^{10}	(too many to count)

The conclusion:

Unless *n* and *c* are small, we will need to consider large sample approximations.

Pearson Statistic

Suppose you want to test the hypothesis

$$H_0 : \pi_j = \pi_{j0}, \forall j \in \{1, 2, \ldots, c\}$$

where $\sum \pi_j = 1$.

Let μ_j be the expected count based on the null probability.

That is,

$$\mu_j = n\pi_{j0}$$

Then Pearson's Statistic is defined as

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}$$

Notes about X^2

- Let X^2_{obs} be the observed value of X^2
- When the Null Hypothesis is true, $(n_j - \mu_j)$ should be small. That is, the expected counts (μ_j) are similar to the observed counts (n_j).
- Greater differences in $(n_j - \mu_j)$ support the alternative hypothesis.
- For large samples, $X^2 \dot{\sim} \chi^2$ with $c - 1$ degrees of freedom.
- The large sample p-value is $P(\chi^2 \geq X^2_{obs})$

Example - Known cell probabilities

- Question: Are births uniformly spread out throughout the year?
- To answer this question, the number of births in King County, Washington, from 1968 to 1979 were tabulated by month.
- Under the null, the probability of having a birth on any given day is equally likely
- Thus, over this 10 year period, there are 3653 total days of which 310 are in January

$$\text{Total days} = 365 * 10 + 3 \text{ leap days } = 3653$$

- Thus, in January, you would expect the probability of a birth to be

$$\pi_1^0 = \frac{310}{3653} = 0.08486$$

- The following table tabulates the remaining probabilities

Month	Days	Null Prob π_{j0}	Actual Births n_j	Expected $\mu_j = n \cdot \pi_{j0}$	Squared Deviation
Jan	310	0.084862	13,016	13,633	27.95778
Feb	283	0.077471	12,398	12,446	0.184791
Mar	310	0.084862	14,341	13,633	36.72786
Apr	300	0.082124	13,744	13,194	22.96163
May	310	0.084862	13,894	13,633	4.982064
June	300	0.082124	13,433	13,194	4.34416
July	310	0.084862	13,787	13,633	1.730962
Aug	310	0.084862	13,537	13,633	0.681361
Sept	300	0.082124	13,459	13,194	5.338968
Oct	310	0.084862	13,144	13,633	17.5667
Nov	300	0.082124	12,497	13,194	36.77873
Dec	310	0.084862	13,404	13,633	3.859317
Total	3653	1	$n =$160,654	160,654	$X^2 =$ 163.1143

Testing

- Since we did not have to estimate any distributional parameters, the total number of degrees of freedom (DF) are

$$df = 12 - 1 = 11$$

- Thus, $X^2 = 163.1143 \sim \chi^2(11)$
- The $p-$value is

$$P(\chi^2 \geq 163.1143 | df = 11) \leq 0.0001$$

- Thus, based on this study, we would conclude that births are not equally distributed throughout the year
- The following slide gives some idea of where the deviation from the null occured
- This is a very basic residual analysis

Month	Actual Births	Expected	Ratio	
January	13,016	13633.38	0.954716	–fewer than expect
February	12,398	12445.96	0.996147	
March	14,341	13633.38	1.051903	–more than expected
April	13,744	13193.59	1.041718	
May	13,894	13633.38	1.019116	
June	13,433	13193.59	1.018146	
July	13,787	13633.38	1.011268	
August	13,537	13633.38	0.992931	
September	13,459	13193.59	1.020116	
October	13,144	13633.38	0.964104	
November	12,497	13193.59	0.947202	
December	13,404	13633.38	0.983175	

We see that the actual is within $\pm 5\%$ of the expect. Is this clinically relevant?

Using SAS

- The calculations above are subject to rounding errors if done by hand
- It is best to calculate the test value with as little rounding as possible
- This can be easily done in Excel, but Excel doesn't sound that "professional"
- In PROC FREQ in SAS, you can conduct the test.

```
data one;
input month $ actual;
cards;
January      13016
February     12398
March        14341
April        13744
May          13894
June         13433
July         13787
August       13537
September    13459
October      13144
November     12497
December     13404
;
run;
proc freq data=one order=data; <--- ORDER=DATA
 weight actual;                   is Important
 tables month /chisq testp=(
 0.084861757
0.077470572  <--This list needs to be in the same
0.084861757     order as your data
0.082124281
0.084861757
0.082124281
0.084861757
0.084861757
0.082124281
0.084861757
0.082124281
0.084861757
)
;
run;
```

Selected Output

The FREQ Procedure

Month	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
January	13016	8.10	8.49	13016	8.10
February	12398	7.72	7.75	25414	15.82
March	14341	8.93	8.49	39755	24.75
April	13744	8.56	8.21	53499	33.30
May	13894	8.65	8.49	67393	41.95
June	13433	8.36	8.21	80826	50.31
July	13787	8.58	8.49	94613	58.89
August	13537	8.43	8.49	108150	67.32
Septembe	13459	8.38	8.21	121609	75.70
October	13144	8.18	8.49	134753	83.88
November	12497	7.78	8.21	147250	91.66
December	13404	8.34	8.49	160654	100.00

Chi-Square Test
for Specified Proportions

Chi-Square	163.1143
DF	11
Pr > ChiSq	<.0001

Sample Size = 160654

Example - Calves with pneumonia

- Suppose we have a sample of 156 dairy calves born in Okeechobee County, Florida
- Calves were classified as to whether or not they experienced pneumonia within 60 days of birth
- Calves that did get an infection were then additionally classified as to whether or not they developed a second infection within 2 weeks of the first one's resolution

Primary Infection	Secondary Infection Yes	Secondary Infection No
Yes	30	63
No	–	63

- The "no primary, yes secondary" is know as a structural zero. (i.e., you can't have a secondary infection unless you have a primary infection)
- We want to test the hypothesis that the probability of primary infection was the same as the conditional probability of secondary infection, given the calf got the primary infection.

- Let π_{ab} denote the probability that a calf is classified in row a and column b
- Under the null hypothesis that the secondary infection is independent of the primary, the following probability structure occurs by letter π be the probability of an infection

Primary Infection	Secondary Infection Yes	No
Yes	π^2	$\pi(1-\pi)$
No	–	$(1-\pi)$

- Note that

$$\sum \pi = \pi^2 + \pi - \pi^2 + 1 - \pi = 1$$

and that

$$156 = 30 + 63 + 63$$

- Then the kernel of the likelihood is

$$L^* = \left[\pi^2\right]^{n_{11}} \left[\pi(1-\pi)\right]^{n_{12}} \left[1-\pi\right]^{n_{22}}$$

- with a log likelihood of

$$l^* = n_{11} \log \pi^2 + n_{12} \log\left(\pi - \pi^2\right) + n_{22} \log\left(1-\pi\right)$$

- In order to solve for the MLE of π, namely $\hat{\pi}$, we need

$$\frac{dl^*}{d\pi}$$

- As a reminder, recall

$$\frac{d \log(u)}{dx} = \frac{1}{u} \cdot \frac{du}{dx}$$

where log is log base e (all we will talk about in this class)

$$\frac{dl^*}{d\pi} = \frac{2n_{11}}{\pi} + \frac{n_{12}(1-2\pi)}{\pi(1-\pi)} - \frac{n_{22}}{1-\pi}$$

- Setting equal to zero and getting a common demoninator yields

$$\frac{2n_{11}(1-\pi) + n_{12}(1-2\pi) - n_{22}\pi}{\pi(1-\pi)} = 0$$

...(some math)

$$\begin{aligned}
\hat{\pi} &= \tfrac{2n_{11}+n_{12}}{2n_{11}+2n_{12}+n_{22}} \\
&= \tfrac{2*30+63}{2*30+2*63+63} \\
&= 0.494
\end{aligned}$$

Expected Values

- Thus, given $n = 156$ we would expect

$$\begin{aligned}
&\widehat{\mu_{11}} = \hat{\pi}^2 * n = 0.494^2 * 156 = 38.1\\
&\widehat{\mu_{12}} = (\hat{\pi} - \hat{\pi}^2) * n = 39.0\\
&\text{and}\\
&\widehat{\mu_{22}} = (1 - \hat{\pi}) * n = 78.9
\end{aligned}$$

- and

$$X^2 = \sum_i \sum_j \frac{n_{ij} - \widehat{\mu_{ij}}}{\widehat{\mu_{ij}}}$$

- Which you can calculate by hand if you so desire
- Or, you can use SAS

Multinomial Goodness of Fit in SAS

```
data two;
 input cell count;
 cards;
 1 30
 2 63
 3 63
 ;
proc freq data=two order =data;
      weight Count;
      tables cell / nocum testf=(
38.1
39.0
78.9
);
run;
```

Correct X^2 wrong p-value and degrees of freedom

```
                          Test
Cell     Frequency    Frequency      Percent
------------------------------------------
   1          30          38.1        19.23
   2          63            39        40.38
   3          63          78.9        40.38
     Chi-Square Test
for Specified Frequencies
--------------------------
Chi-Square          19.6955 <--- This is correct
DF                        2 <--- NEEDS TO BE ADJUSTED
Pr > ChiSq           <.0001   on account of estimating
                              estimating pi!!!!
Sample Size = 156
```

The correct degrees of freedom are 3 - 1 (for the constraint) - 1 (for the estimated π) $= 1$. However, p is still less than 0.0001.

Likelihood Ratio Test

The 'kernel' of the multinomial likelihood is $L(\cdot) = \prod_j (\pi_j)^{n_j}$

and as such the kernel under the null is $L(\vec{n}, \pi_j) = \prod_j (\pi_{j0})^{n_j}$.

Under the observed sample using the MLE of $\vec{\pi}$ is $L(\vec{n}, \pi_a) = \prod_j (n_j/n)^{n_j}$,

so the likelihood ratio statistic is written as

$$G^2 = 2\sum_{j=1}^{c} n_j \log(\frac{n_j}{n\pi_{j0}})$$

Here $G^2 \dot{\sim} \chi^2$ with $c-1$ degrees of freedom.

Goodness of Fit [GOF]

These three tests (EMT, X^2 and G^2) are generally classified as GOF tests.

As opposed to inference on a probability, we are not interested in calculating a confidence interval for $\vec{\pi}$.

We can use these test to test the fit of data to a variety of distributions.

Example: GOF for Poisson Data

Suppose the following table represents the number of deaths per year that result from a horse kick in the Prussian army.

We want to know if we can model the data using a Poisson distribution.

	Number of deaths				
	0	1	2	3	4
Deaths per year per corp	0	1	2	3	4
Frequency of Occurrence	144	91	32	11	2

The mean number of deaths per year is

$$\widehat{\lambda} = \frac{0(144) + 1(91) + 2(32) + 3(11) + 4(2)}{280} = \frac{196}{280} = 0.70$$

If the number of deaths were distributed as Poisson with $\lambda = .7$, then

$$P(Y = 0) = \frac{e^{-.7}(0.7)^0}{0!} = 0.4966$$

Thus, given $n = 280$, you would expect $n(0.4966) = 139.048$ deaths.

The following table summarizes the remaining expectations:

	Number of deaths				
	0	1	2	3	>4
Observed Frequency	144	91	32	11	2
Expected Frequency	139.048	97.328	34.076	7.952	1.596

$$
\begin{aligned}
X^2 &= \sum_j \frac{(n_j - \mu_j)^2}{\mu_j} \\
&= (144 - 139.048)^2/139.048 + \cdots + (2 - 1.596)^2/1.596 \\
&= 1.9848; \quad p = .5756 \\
G^2 &= 2\sum_j n_j \log(n_j/\mu_j) \\
&= 2(144 \log(144/139.048) + \cdots + 2 \log(2/1.596)) \\
&= 1.86104; \quad p = .39826[NOTE\ \text{:g}^2 \text{ calculated with the natural log}]
\end{aligned}
$$

Note: The degrees of freedom for these tests are 3 (5 - 1 - 1). 5 is the number of categories and the first "-1" is for the constraint. The second "-1" is for the estimation of λ.

Conclusion: There is insufficient evidence to reject the null hypothesis that the data are Poisson. (i.e., the model fits)

Pearson's in SAS using expected frequencies

- Presently, fitting the likelihood ratio statistic in SAS for a one-way table is not "canned"
- That is, you would need to program the calculations directly
- However, PROC FREQ does allow for the specification of expected counts instead of probabilities as we used previously

```
data one;
input deaths count;
cards;
0   144
1   91
2   32
3   11
4   2
;
proc freq data=one order=data;
 weight count;
 tables deaths /chisq testf=(
 139.048
 97.328
 34.076
  7.952
  1.596
  );
run;
```

```
                         Test
deaths     Frequency    Frequency      Percent
----------------------------------------------
     0         144       139.048        51.43
     1          91        97.328        32.50
     2          32        34.076        11.43
     3          11         7.952         3.93
     4           2         1.596         0.71


     Chi-Square Test
for Specified Frequencies
-------------------------
Chi-Square          1.9848
DF                       4 <-- Just note, this is wrong
Pr > ChiSq          0.7385 b/c we estimated mu

Sample Size = 280
```