The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 3, pp. 499–510

A command for significance and power to test for the existence of a unique most probable category

Bryan M. Fellman
MD Anderson Cancer Center
Houston, TX
bmfellman@mdanderson.org

Joe Ensor
MD Anderson Cancer Center
Houston, TX
joensor@mdanderson.org

Abstract. The analysis of multinomial data often includes the following question of interest: Is a particular category the most populous (that is, does it have the largest probability)? Berry (2001, *Journal of Statistical Planning and Inference* 99: 175–182) developed a likelihood-ratio test for assessing the evidence for the existence of a unique most probable category. Nettleton (2009, *Journal of the American Statistical Association* 104: 1052–1059) developed a likelihood-ratio test for testing whether a particular category was most probable, showed that the test was an example of an intersection-union test, and proposed other intersection-union tests for testing whether a particular category was most probable. He extended his likelihood-ratio test to the existence of a unique most probable category and showed that his test was equivalent to the test developed by Berry (2001, *Journal of Statistical Planning and Inference* 99: 175–182). Nettleton (2009, *Journal of the American Statistical Association* 104: 1052–1059) showed that the likelihood ratio for identifying a unique most probable cell could be viewed as a union-intersection test. The purpose of this article is to survey different methods and present a command, `cellsupremacy`, for the analysis of multinomial data as it pertains to identifying the significantly most probable category; the article also presents a command for sample-size calculations and power analyses, `power_cellsupremacy`, that is useful for planning multinomial data studies.

Keywords: st0348, cellsupremacy, cellsupremacyi, power_cellsupremacy, most probable category, multinomial data, cell supremacy, cell inferiority

1 Introduction

If $Y_1, Y_2, \ldots, Y_k$ are independent Poisson-distributed random variables with means μ_1, μ_2, …, μ_k, then $(Y_1, Y_2, \ldots, Y_k)$, conditional on their sum, is multinomial$(N, p_1, p_2, \ldots, p_k)$, where $p_i = \mu_i / \sum_{\forall_k} \mu_k$ represents the probability of the ith category. Multinomial data are common in biological, marketing, and opinion research scenarios. In a recent study, Price et al. (2011) used data from the 2008 National Health Interview Survey to examine whether 18- to 26-year-old women who are most likely to benefit from catch-up vaccination are aware of the human papillomavirus (HPV) vaccine and have received initial and subsequent doses in the 3-dose series. The study found that the most common reasons for lack of interest in the HPV vaccine were belief that it was not needed (35.9%), not knowing enough about it (17.1%), concerns about safety (12.7%),

and not being sexually active (10.3%). These 4 responses were among the 11 possible response categories to the survey question. Is the belief among respondents that the HPV vaccine was not needed the unique most probable reason for lack of interest in the HPV vaccine? Response to questionnaire-based infertility studies varies, and Morris et al. (2013) noted that different modes of contact can affect response. Results of their study indicated that 59% of the women surveyed preferred a mailed questionnaire, 37% chose an online questionnaire, and only 3% selected a telephone interview as their mode of contact. Is a mailed questionnaire the most preferred mode of contact? Are these results significant? The purpose of this article is to survey different methods and to present a command for the analysis of multinomial data as it pertains to identifying the significantly most probable category; the article also presents a command for sample-size calculations and power analyses that is useful for planning multinomial data studies.

2 Methods

Nettleton (2009) posed the test for the supremacy of a multinomial cell probability as an intersection-union test (IUT). Suppose $\mathbf{X} = (X_1, \ldots, X_k)$ has a multinomial distribution with n trials and the cell probabilities $p_1, \ldots, p_k$. The parameter $\mathbf{p} = (p_1, \ldots, p_k)$ lies in the set $\mathbf{P}$ of vectors of order k, whose components are positive and sum to one. The tested null hypothesis states that a particular cell of interest is not more probable than all others. Suppose the kth cell is the cell of interest; then the hypothesis can be formulated as

$$H_0 \colon \bigcup_{i=1}^{k-1} p_k \leq p_i \quad \text{versus} \quad H_1 \colon \bigcap_{i=1}^{k-1} p_k > p_i$$

which Nettleton (2009) noted can be stated as

$$H_0 \colon p_k \leq \max(p_1, \ldots, p_{k-1}) \quad \text{versus} \quad H_1 \colon p_k > \max(p_1, \ldots, p_{k-1})$$

Nettleton (2009) offered three possible asymptotic IUT statistics: the score test, the Wald test, and the likelihood-ratio test. Suppose $\mathbf{x} = (x_1, \ldots, x_k)$ is a realization of $\mathbf{X} = (X_1, \ldots, X_k)$; then $\widehat{p}_i = x_i/n$ so that $\widehat{\mathbf{p}} = (\widehat{p}_1, \ldots, \widehat{p}_k)$ is the maximum likelihood estimate of $\mathbf{p} = (p_1, \ldots, p_k)$. Each asymptotic IUT statistic is zero unless x_k is greater than $\max(x_1, \ldots, x_{k-1})$. Nettleton (2009) also suggested a test based on the conditional distribution of X_k, given the sum of x_k and m, where $m = \max(x_1, \ldots, x_{k-1})$.

2.1 Score test

The test statistic for the asymptotic score test is

$$T_S = \begin{cases} \frac{n(\widehat{p}_k - \widehat{p}_M)^2}{\widehat{p}_k + \widehat{p}_M} & \text{if } \widehat{p}_k > \widehat{p}_M = \max(\widehat{p}_1, \ldots, \widehat{p}_{k-1}) \\ 0 & \text{otherwise} \end{cases}$$

H_0 is rejected if and only if $T_S \geq \chi^2_{(1),1-2\alpha}$, where $\chi^2_{(1),1-2\alpha}$ represents the $\{100 \times (1-2\alpha)\}$th quantile of the χ^2 distribution with 1 degree of freedom. The approximate

p-value for the test is given by $P_r(\chi^2_{(1)} \geq T_S \mid T_S)/2$, where $\chi^2_{(1)}$ denotes a χ^2 random variable with 1 degree of freedom.

2.2 Wald test

The test statistic for the asymptotic Wald test is

$$T_W = \begin{cases} \frac{n(\widehat{p}_k - \widehat{p}_M)^2}{\widehat{p}_k + \widehat{p}_M - (\widehat{p}_k - \widehat{p}_M)^2} & \text{if } \widehat{p}_k > \widehat{p}_M = \max(\widehat{p}_1, \ldots, \widehat{p}_{k-1}) \\ 0 & \text{otherwise} \end{cases}$$

H_0 is rejected if and only if $T_W \geq \chi^2_{(1),1-2\alpha}$. The approximate p-value for the test is given by $P_r(\chi^2_{(1)} \geq T_W \mid T_W)/2$.

2.3 Likelihood-ratio test

The test statistic for the asymptotic likelihood-ratio test is

$$T_{LR} = \begin{cases} 2\left\{ M \ln\left(\frac{2M}{M+x_k}\right) + x_k \ln\left(\frac{2x_k}{M+x_k}\right)\right\} & \text{if } x_k > M = \max(x_1, \ldots, x_{k-1}) \\ 0 & \text{otherwise} \end{cases}$$

H_0 is rejected if and only if $T_{LR} \geq \chi^2_{(1),1-2\alpha}$. The approximate p-value for the test is given by $P_r(\chi^2_{(1)} \geq T_{LR} \mid T_{LR})/2$.

2.4 Conditional binomial test

The conditional distribution of X_k, given $m + x_k$, where $m = \max(x_1, \ldots, x_{k-1})$, is binomial$(m + x_k, 1/2)$. Thus a p-value for testing the null hypothesis that is valid for all n is $P_r\{X_k \geq x_k \mid x_k + \max(x_1, \ldots, x_k)\}$. The conditional IUT is equivalent to a permutation test, where the p-value is expressed as

$$p\text{-value} = \sum_{x=x_k}^{m+x_k} \left(\begin{array}{c} m + x_k \\ x \end{array} \right) \times 2^{-(m+x_k)}$$

The simulation studies by Nettleton (2009) showed that the conditional IUT based on the binomial distribution yielded a true p-value typically less than the nominal value. Farcomeni (2012) suggested that the exact test (that is, conditional binomial) may be conservative and that the exact significance level may be smaller than the desired nominal level. Farcomeni (2012) suggested using the typical continuity correction for the binomial; namely, he recommended the mid-p value as the p-value of the test.

2.5 Mid-p value test

Using the mid-p value approach, we see that the p-value is

$$p\text{-value} = \left(\begin{array}{c} m + x_k \\ x_k \end{array} \right) \times 2^{-(m+x_k+1)} + \sum_{x=x_k+1}^{m+x_k} \left(\begin{array}{c} m + x_k \\ x \end{array} \right) \times 2^{-(m+x_k)}$$

2.6 Inferiority test

The test for cell supremacy can be formulated as

$$H_0 : p_k \leq \max(p_1, \ldots, p_{k-1}) \quad \text{versus} \quad H_1 : p_k > \max(p_1, \ldots, p_{k-1})$$

One could formulate the test for cell inferiority (that is, a particular cell is least probable) as

$$H_0 : p_k \geq \min(p_1, \ldots, p_{k-1}) \quad \text{versus} \quad H_1 : p_k < \min(p_1, \ldots, p_{k-1})$$

Farcomeni (2012) suggests using the exact test for inferiority where the sum goes from 0 to x_k. That is, the p-value for the conditional IUT for inferiority would be

$$p\text{-value} = \sum_{x=0}^{x_k} \left(\begin{array}{c} m + x_k \\ x \end{array} \right) \times 2^{-(m+x_k)}$$

and the mid-p value adjustment could be stated as

$$p\text{-value} = \left(\begin{array}{c} m + x_k \\ x_k \end{array} \right) \times 2^{-(m+x_k+1)} + \sum_{x=0}^{x_k-1} \left(\begin{array}{c} m + x_k \\ x \end{array} \right) \times 2^{-(m+x_k)}$$

Alam and Thompson (1972) discussed the challenges of testing whether a particular cell is least probable from a design point of view. Nettleton (2009) showed that the likelihood-ratio test statistic could be used to test for the existence of a unique most probable cell. That is, rather than test whether a particular cell chosen a priori is the most probable, one could test whether the largest observed cell was uniquely most probable. The likelihood-ratio test statistic matches the test statistic developed by Berry (2001) and rejects H_0 if and only if $T_{LR} \geq \chi^2_{(1),1-2\alpha}$. The approximate p-value for the test is given by $P_r(\chi^2_{(1)} \geq T_{LR} \mid T_{LR})$, where $\chi^2_{(1)}$ denotes a χ^2 random variable with 1 degree of freedom. That is, the p-value is twice the p-value for the test in which a particular cell chosen a priori is most probable.

2.7 Power

We consider the case of a random variable $\mathbf{X}$~multinomial$(n, p_1, \ldots, p_k)$. Without loss of generality, we will assume that p_k is the maximum among the k cells. Let

$p_M = \max(p_1, \ldots, p_{k-1})$—that is, assume the maximum $p_i; i = 1, 2, \ldots, k-1$ occurs at $i = M$—and consider the test

$$H_0 : p_k = p_M \quad \text{versus} \quad H_1 : p_k > p_M$$

The score test rejects H_0 if

$$T_S \geq \chi^2_{(1),1-2\alpha}$$

and for $x_k > x_M$,

$$T_S = \frac{n\left(\widehat{p}_k - \widehat{p}_M\right)^2}{\widehat{p}_k + \widehat{p}_M} = n\left\{ \frac{\left(\widehat{p}_k - \frac{\widehat{p}_k + \widehat{p}_M}{2}\right)^2}{\frac{\widehat{p}_k + \widehat{p}_M}{2}} + \frac{\left(\widehat{p}_M - \frac{\widehat{p}_k + \widehat{p}_M}{2}\right)^2}{\frac{\widehat{p}_k + \widehat{p}_M}{2}} \right\}$$

where α is the significance level of the test. To evaluate

$$\text{power} = P_r(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_k, p_M \ni p_k > p_M)$$

we need the noncentrality parameter,

$$\lambda = n\left\{ \frac{(p_k - p_0)^2}{p_0} + \frac{(p_M - p_0)^2}{p_0} \right\} = 2n\left\{ \frac{(p_k - p_0)^2}{p_0} \right\}$$

where $p_0 = (p_k + p_M)/2$ (Guenther 1977). For example, consider the random variable

$$\mathbf{X}\text{~multinomial}(n = 50, p_1 = 0.1, p_2 = 0.1, p_3 = 0.1, p_4 = 0.3, p_5 = 0.4)$$

Suppose we wish to test the hypothesis

$$H_0 : p_5 \leq \max(p_1, \ldots, p_4) \quad \text{versus} \quad H_1 : p_5 > \max(p_1, \ldots, p_4)$$

at the $\alpha = 0.05$ significance level. The null hypothesis is rejected if $T_S \geq 2.70554$. Solely based on p_4 and p_5, the noncentrality parameter for testing the 5th cell selected a priori as the most probable cell is

$$\lambda = 100 \times \left\{ \frac{(0.4 - 0.35)^2}{0.35} \right\} \approx 0.71429$$

and the approximate power is

$$\text{power} \approx P_r(\chi^2_{(1),0.71479} \geq 2.70554) \approx 0.21833$$

where $\chi^2_{(1),0.71479}$ is a noncentral χ^2 random variable with a noncentrality parameter of 0.71479 and 1 degree of freedom. The simulation of size 100,000 yielded a power equal to 0.214 for this scenario. The approximation is ignorant of the distribution of the first $k-1$ cells. Because p_4 is three times greater than any other cell probability amount in the first $k-1$ cells, the approximation yields a reasonable result. Now consider the random variable

$$\mathbf{X}\text{~multinomial}(n = 50, p_1 = 0, p_2 = 0, p_3 = 0.3, p_4 = 0.3, p_5 = 0.4)$$

We have a trinomial, and there is strong competition for the maximum among the first $k-1$ cells. Because the cells of a multinomial are not independent, one would expect the distribution of the first $k-1$ cells that affect the power to detect the kth cell to be the most probable. The simulated power for this scenario was 0.087. Thus the approximation of power must consider the impact of the distribution of the first $k-1$ cells. The correlation among the two cells of a multinomial is

$$\rho_{a,b} = -\sqrt{\frac{p_a p_b}{(1-p_a)(1-p_b)}}$$

The power to detect the 5th cell as the most probable is the power that $p_5 > p_4$ and $p_5 > p_3$. Consider approximating the power by

$$\text{power} \approx P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_k, p_M\right)\left\{P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_k, p_N\right)\right\}^{1+\rho_{M,N}}$$

where p_M and p_N represent the maximum and the second largest of the cell probabilities of the first $k-1$ cells, respectively, and $\rho_{M,N}$ represents the correlation between cells M and N. For our example, the approximate power is

$$\begin{aligned}
\text{power} &\approx P_r(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_3 = 0.3) \\
&\quad \times \left\{P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_4 = 0.3\right)\right\}^{1+\rho_{4,3}} \\
&\approx (0.21833)\,(0.21833)^{1-0.42857} \\
&\approx 0.09151
\end{aligned}$$

Applying this form of the approximation to the original example with p_1 through p_3 equal to 0.1 and p_4 equal to 0.3 yields an approximate power of

$$\begin{aligned}
\text{power} &\approx P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_3 = 0.3\right) \\
&\quad \times \left\{P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_3 = 0.1\right)\right\}^{1+\rho_{4,3}} \\
&\approx (0.21833)\,(0.91232)^{1-0.21822} \\
&\approx 0.20322
\end{aligned}$$

Table 1 provides simulations of size 100,000 for several scenarios to investigate the adequacy of our proposed approximation. For each scenario, p_6 is the cell of interest, $\rho_{5,4}$ represents the correlation between the 5th and 4th cell, "Sim." is the simulated power, and "Approx." is our power approximation.

Table 1. Power analysis

Scenario	p_1	p_2	p_3	p_4	p_5	p_6	$\rho_{5,4}$	Subjects	Sim.	Approx.
1	0	0.1	0.1	0.1	0.3	0.4	−0.2182	25	0.137	0.119
2								50	0.214	0.203
3								200	0.520	0.519
4								1000	0.984	0.984
5	0	0	0	0.3	0.3	0.4	−0.4286	25	0.057	0.056
6								50	0.087	0.092
7								200	0.353	0.356
8								1000	0.971	0.974
9	0.0626	0.0625	0.0625	0.0625	0.25	0.5	−0.1491	25	0.413	0.384
10								50	0.664	0.651
11								200	0.994	0.993
12								1000	1.000	1.000
13	0	0	0	0.25	0.25	0.5	−0.3333	25	0.260	0.237
14								50	0.504	0.493
15								200	0.989	0.988
16								1000	1.000	1.000
17	0.05	0.05	0.05	0.05	0.2	0.6	−0.1147	25	0.747	0.698
18								50	0.953	0.935
19								200	1.000	1.000
20								1000	1.000	1.000
21	0	0	0	0.2	0.2	0.6	−0.2500	25	0.631	0.567
22								50	0.915	0.890
23								200	1.000	1.000
24								1000	1.000	1.000
25	0.1	0.1	0.1	0.1	0.2	0.4	−0.1667	25	0.257	0.265
26								50	0.550	0.530
27								200	0.981	0.978
28								1000	1.000	1.000
29	0	0	0.2	0.2	0.2	0.4	−0.2500	25	0.143	0.170
30								50	0.326	0.376
31								200	0.953	0.961
32								1000	1.000	1.000

2.8 Conclusions

Nettleton (2009) suggested that the asymptotic procedures are preferred for moderate to large sample sizes based on simulations, but the IUT based on conditional tests is a useful option when a small sample size casts doubt on the validity of the asymptotic procedures. Our power simulations tend to also suggest that the power approximation works best for moderate to large sample sizes. Scenarios 29–32 present a slightly more complex problem with three cells vying for the top spot among the first cells. For these scenarios, our power approximation yields slightly liberal results because the approximate power is consistently larger than the simulated power. Under this scenario, the power to detect the 6th cell as the most probable is the power that $p_6 > p_5$, $p_6 > p_4$, and $p_6 > p_3$. Thus one could improve the approximation by considering the added competition for supremacy among the first $k-1$ cells. That is, for $n = 200$, the approximate power is

$$\begin{aligned}
\text{power} &\approx P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_4 = 0.2\right) \\
&\quad \times \left\{P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_3 = 0.2\right)\right\}^{1+\rho_{4,3}} \\
&\quad \times \left\{P_r\left(T_S \geq \chi^2_{(1),1-2\alpha} \mid p_5 = 0.4, p_3 = 0.2\right)\right\}^{1+2\rho_{4,3}} \\
&\approx (0.97761)\,(0.97761)^{1-0.25}(0.97761)^{1-0.50} \\
&\approx 0.95032
\end{aligned}$$

which compares favorably with the simulated power. However, we believe that for most real-world problems, considering the impact of the top two cell probabilities among the first $k-1$ cells is sufficient.

3 The cellsupremacy, cellsupremacyi, and power_cellsupremacy commands

3.1 Syntax

`cellsupremacy` *varname* [*weight*]

`cellsupremacyi, counts(`*numlist*`)`

`power_cellsupremacy, freq(`*numlist*`) n(#)` [`simulate dots reps(#) alpha(#)`]

`fweights` is allowed; see [U] **11.1.6 weight**.

3.2 Option for cellsupremacyi

`counts(`*numlist*`)` specifies the cell counts for each category of the variable of interest. `counts()` is required.

3.3 Options for power_cellsupremacy

`freq(`*numlist*`)` specifies the frequency of cells for each category of the variable of interest. `freq()` is required.

`n(#)` specifies the number of observations. `n()` is required.

`simulate` calculates the simulated power and the approximate power. When not specified, only the approximated power is calculated.

`dots` shows the replication dots when using the `simulate` option.

`reps(#)` specifies the number of simulations used to calculate the power. The default is `reps(10000)`.

`alpha(#)` specifies the alpha that is used for calculating the power. The default is `alpha(0.05)`.

3.4 Examples

Suppose we are studying breast cancer and we find that the distribution of `subtypes` is a trinomial distribution with `HER2+`, `HR+`, and `TNBC`. In our data, we find that patients with leptomeningeal disease were more likely to be `HER2+` (45%). We are interested in knowing whether this particular category is the most populous (that is, does it have the largest probability of occurring?). The following example will generate a sample dataset and illustrate the use of the new command to answer this question.

```
. set obs 100
obs was 0, now 100
. generate subtype = "HER2+" in 1/45
(55 missing values generated)
. replace subtype = "HR+" in 46/73
(28 real changes made)
. replace subtype = "TNBC" in 74/100
(27 real changes made)
. tab subtype
```

subtype	Freq.	Percent	Cum.
HER2+	45	45.00	45.00
HR+	28	28.00	73.00
TNBC	27	27.00	100.00
Total	100	100.00	

```
. cellsupremacy subtype

TESTS FOR CELL SUPREMACY
Category HER2+ had the largest observed frequency.
TESTING WHETHER CATEGORY HER2+ SELECTED A PRIORI IS MOST PROBABLE.
Quantity           Score     Wald      LR        Binomial  Mid-P
-------------------------------------------------------------------
Test Statistic     3.9589    4.1221    3.9955
p-value            0.0233    0.0212    0.0228    0.0302    0.0237

TEST FOR THE EXISTENCE OF A MOST PROBABLE CELL
Quantity           LR
-------------------------
Test Statistic     3.9955
p-value            0.0456

TESTS FOR CELL INFERIORITY
Category TNBC had the smallest observed frequency.
TESTING WHETHER CATEGORY TNBC SELECTED A PRIORI IS LEAST PROBABLE.
Quantity           Binomial            Mid-P
-----------------------------------------------
p-value            0.5000              0.4469
```

The p-values for all tests are less than 0.05, which indicates that `HER2+` is the most probable. The test for the existence of a most probable cell is also significant. On the other hand, if we were interested in cell inferiority (least probable), we would not reject our hypothesis because our p-values are approximately 0.50. Below is another example with a slightly different distribution than before.

```
. clear

. set obs 100
obs was 0, now 100

. generate subtype = "HER2+" in 1/45
(55 missing values generated)

. replace subtype = "HR+" in 46/85
(40 real changes made)

. replace subtype = "TNBC" in 86/100
(15 real changes made)

. tab subtype
    subtype |      Freq.     Percent        Cum.
------------+-----------------------------------
      HER2+ |         45       45.00       45.00
        HR+ |         40       40.00       85.00
       TNBC |         15       15.00      100.00
------------+-----------------------------------
      Total |        100      100.00

. cellsupremacy subtype

TESTS FOR CELL SUPREMACY
Category HER2+ had the largest observed frequency.
TESTING WHETHER CATEGORY HER2+ SELECTED A PRIORI IS MOST PROBABLE.
Quantity           Score     Wald      LR        Binomial  Mid-P
-------------------------------------------------------------------
Test Statistic     0.2941    0.2950    0.2943
p-value            0.2938    0.2935    0.2937    0.3323    0.2950

TEST FOR THE EXISTENCE OF A MOST PROBABLE CELL
Quantity           LR
-------------------------
Test Statistic     0.2943
p-value            0.5875

TESTS FOR CELL INFERIORITY
Category TNBC had the smallest observed frequency.
TESTING WHETHER CATEGORY TNBC SELECTED A PRIORI IS LEAST PROBABLE.
Quantity           Binomial            Mid-P
-----------------------------------------------
p-value            0.0005              0.0003
```

Because `HER2+` and `HR+` have similar frequencies, we cannot conclude that `HER2+` is the most probable. In this case, we can conclude that `TNBC` is the least probable cell. The above examples can both be implemented by entering the raw counts `cellsupremacyi 45 28 27` or `cellsupremacyi 45 40 15`, respectively.

To illustrate how to use the `power_cellsupremacy` command to calculate the power of the test, we consider the examples in section 2.7 for testing cell superiority for the random variables,

$$\mathbf{X}\text{~multinomial}(n = 50, p_1 = 0, p_2 = 0, p_3 = 0.3, p_4 = 0.3, p_5 = 0.4)$$

and

$$\mathbf{Y}\text{~multinomial}(n = 50, p_1 = 0.1, p_2 = 0.1, p_3 = 0.1, p_4 = 0.3, p_5 = 0.4)$$

```
. clear
. set seed 339487731
. power_cellsupremacy, simulate freq(0 0 0.3 0.3 0.4) n(50)
Simulations (10000)
N        Simulated Power     Approximate Power
50       0.0898              0.0915
. power_cellsupremacy, simulate freq(0.1 0.1 0.1 0.3 0.4) n(50)
Simulations (10000)
N        Simulated Power     Approximate Power
50       0.2121              0.2032
```

4 Acknowledgment

This research is supported in part by the National Institutes of Health through M. D. Anderson's Cancer Center Support Grant CA016672.

5 References

Alam, K., and J. R. Thompson. 1972. On selecting the least probable multinomial event. *Annals of Mathematical Statistics* 43: 1981–1990.

Berry, J. C. 2001. On the existence of a unique most probable category. *Journal of Statistical Planning and Inference* 99: 175–182.

Farcomeni, A. 2012. Testing supremacy or inferiority of multinomial cell probabilities with application to biting preferences of loggerhead marine turtles. *Communications in Statistics—Theory and Methods* 41: 34–45.

Guenther, W. C. 1977. Power and sample size for approximate chi-square tests. *American Statistician* 31: 83–85.

Morris, M., P. Edwards, P. Doyle, and N. Maconochie. 2013. Women in an infertility survey responded more by mail but preferred a choice: Randomized controlled trial. *Journal of Clinical Epidemiology* 66: 226–235.

Nettleton, D. 2009. Testing for the supremacy of a multinomial cell probability. *Journal of the American Statistical Association* 104: 1052–1059.

Price, R. A., J. A. Tiro, M. Saraiya, H. Meissner, and N. Breen. 2011. Use of human papillomavirus vaccines among young adult women in the United States: An analysis of the 2008 National Health Interview Survey. *Cancer* 117: 5560–5568.

About the authors

Bryan Fellman is a research statistical analyst in the Department of Biostatistics at the University of Texas MD Anderson Cancer Center.

Joe Ensor is a research statistician in the Department of Biostatistics at the University of Texas MD Anderson Cancer Center.