Guidance on Using Multinomial Tests for Differences in Distribution

Note: The Equity Evaluation Memo Series is intended to guide OES' commitment to equity in our evaluation process and efforts toward understanding and reducing barriers to equitable access to federal programs. This series is intended to be an internal guidance document for OES team members.

Some of the more common equity-related descriptive and causal research questions at OES center on drawing comparisons across multiple categories between two samples given their status with respect to some policy outcome or behavior. We may alternatively wish to compare how a benefit was distributed among a sample of beneficiaries relative to a well-defined target or eligible population. When we do this, we may use a multinomial statistical test such as a chi-squared test to draw inferences about whether (1) two sub-samples were drawn from the same population or (2) the sample of beneficiaries of a program reflects the population of eligible individuals. Our inferences may allow us to identify under-served sub-groups or to confirm whether a benefit is equitably distributed among targeted beneficiaries.

When performing a multinomial test, one of two versions may be used. The right choice will depend on the research question being asked.

- When we seek to draw inferences about the equality of two samples, a two-sample multinomial test is recommended. For example, ***a chi-squared test of independence*** would be appropriate.
- When we seek to draw inferences about whether a sample was drawn from a particular population, a one-sample multinomial test is recommended. For example, ***a chi-squared goodness-of-fit test*** would be appropriate.

In the course of using multinomial tests, we have found that it is easy to misapply them. Misapplication is not trivial, because failure to implement the right kind of test can lead to unreliable inferences. In cases where we want to test whether two samples are equivalent, a goodness-of-fit test will have anti-conservative properties. That is, we will reject a true null hypothesis far more frequently than our stated alpha threshold. For instance, if our null hypothesis is that the demographic profile of a sample of beneficiaries is equivalent to that of a population of eligible individuals, an anti-conservative test will lead us to reject the null of equivalence far more often than we should if the null really is true. Conversely, in cases where we want to compare a sample to a well-defined population or expected distribution, a test of independence can be too conservative; it will fail to reject false null hypotheses as often as it should. For example, if we are comparing a sample of beneficiaries and a sample of eligible beneficiaries, and if the null is false, an overly conservative test will be underpowered to detect inequality relative to the correct test. It is

important, therefore, to make sure that we select the appropriate test for our research question, and that we use the correct function or routine in our statistical software.

The following sections illustrate when one or the other test is appropriate and provide guidance on how to implement each in R.

Background on Multinomial Distributions

For some background on multinomial tests, we would use such a test when evaluating *multinomial distributions*. Such a distribution describes the frequencies by which individuals in a sample will fall into one of $k > 1$ discrete categories. Each category has an underlying probability where the probabilities across categories sum to 1. Say for example we have a six sided fair die. The expected multinomial distribution for such a die would assign a probability of 1/6 to each side.

Two-Sample vs. Goodness-of-Fit Multinomial Tests

As noted above, we may use a multinomial test to answer one of two questions:

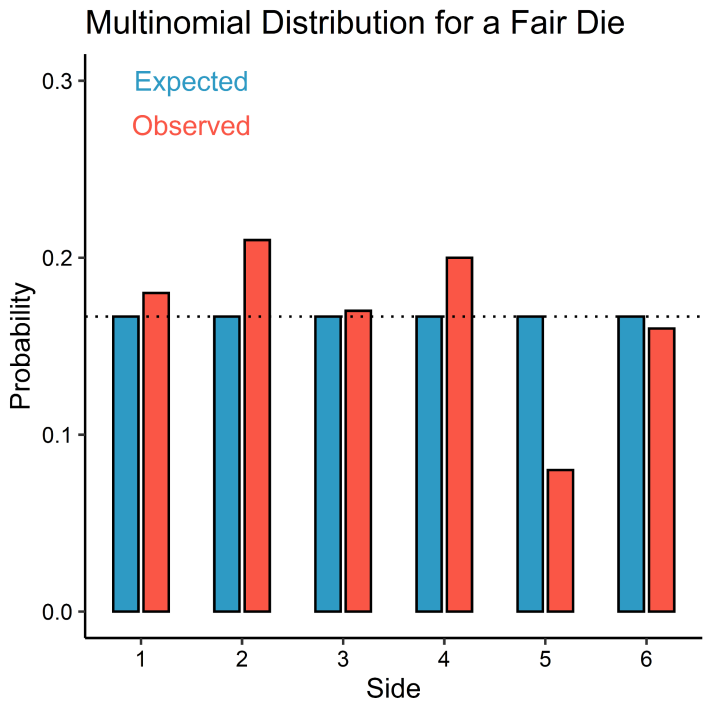
1. Have two samples of interest been drawn from similar populations or multinomial data-generating processes?
2. Has a sample of interest been drawn from a particular well-defined population or multinomial data-generating process?[1]

To illustrate when one or the other test is appropriate consider two kinds of questions we might ask of a six-sided die.

First, say we wanted to test whether a particular six-sided die is fair. To do this, we might roll the die several times and note the proportion of times we roll a 1, 2, 3, etc. Then, we would compare the observed proportions to the expected probabilities to draw inferences about the fairness of the die. Figure 1 illustrates what the expected vs. observed proportions look like after simulating 100 rolls of a fair die.

[1] Other than a population, the expected probabilities from a (weighted) lottery could also be used to characterize a multinomial data-generating process.

Figure 1. Expected vs. Observed Multinomial Distribution for a Fair Six-Sided Die



In this instance, we are interested in comparing a sample of observed die tosses to a well-defined multinomial data-generating process. To test whether the die is fair, we would use a multinomial goodness-of-fit test to determine whether the observed distribution of counts diverge from expectations more than we would expect by chance. One such test is the chi-squared test, which uses the chi-squared test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \,. \qquad (1)$$

For a fair six-sided die, *i* indexes the sides, 1 through 6, *k* denotes the number of sides, E_i denotes the expected number of times we would roll a particular number after *n* rolls, and O_i is the observed number of times that number is rolled.

Using the simulated rolls shown in Figure 1, we get a chi-squared statistic of approximately 6.44. Under the null hypothesis that the die is fair, the expected chi-squared statistic would be approximately 3. As it turns out, the observed chi-squared is not statistically significant ($p = 0.27$). This means that the die we "rolled" in simulations is not statistically distinguishable from a fair die.

Alternatively, say we have two dice and want to compare them to see if they are equivalent to each other. To answer this question, we can no longer use the chi-squared statistic as defined in equation 1. This is because that equation assumes we are comparing *one* random sample to a well-defined (that is, known) multinomial distribution. But, when we seek to test whether *two*

samples have been drawn from the same data-generating process we have to proceed in a different way.

Below is the chi-squared statistic for a two-sample test, or what is also called a test of independence:[2]

$$\chi^2 = \sum_{i=1}^{k} \frac{(K_1 R_i - K_2 S_i)^2}{R_i + S_i}; \text{ where } K_1 = \sum_{i=1}^{k} \sqrt{\frac{S_i}{R_i}} \text{ and } K_2 = \sum_{i=1}^{k} \sqrt{\frac{R_i}{S_i}}. \quad (2)$$

This new formulation takes into account uncertainty in the observed frequencies for sample 1, denoted by R_i, and the observed frequencies for sample 2, denoted by S_i. The values K_1 and K_2 are normalizing constants that ensure that the frequencies for sample 1 are comparable with those in sample 2. Conceptually, this formulation compares the observed values in each sample in a given category to the average of the frequencies in a given category scaled to the size of each sample.

For example, consider Table 1 which shows observed frequencies for two hypothetical samples. How would we obtain the scaled average within categories between samples? We first convert the frequencies to within-sample proportions. For sample 1, this is 0.4 and 0.6 for categories 1 and 2 respectively. For sample 2, this is 0.6 and 0.4. Next, we take the category-wise average of these proportions: 0.5 and 0.5 respectively. Finally, to get the scaled expected frequencies for each cell, we multiply these proportions by the size of each sample. For sample 1, this gives us an expected frequency of 0.5 * 100 = 50 for category 1 and 0.5 * 100 = 50 for category 2. For sample 2, this gives us an expected frequency of 0.5 * 200 = 100 for category 1 and 0.5 * 200 = 100 for category 2. We then use the difference between the observed frequencies and the expected frequencies to calculate the chi-squared statistic for the test of independence.

Table 1. A Hypothetical Comparison of Samples

	Sample 1	Sample 2	*Category Total:*
Category 1	40 (E = 50)	120 (E = 100)	160
Category 2	60 (E = 50)	80 (E = 100)	140
Group Total:	100	200	300

The key difference with this test is that it, unlike a goodness-of-fit test, factors in that the samples being compared are random draws from multinomial distributions of unknown form. Failure to account for this (for example, by using a goodness-of-fit test when a test of independence is appropriate) can yield unreliable inferences, as we discuss in the next section.

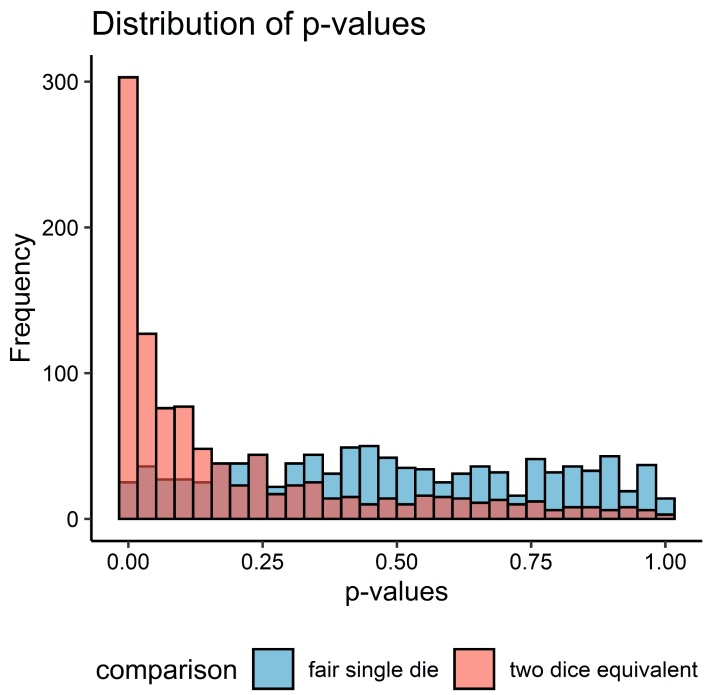
[2] See the technical Appendix for an alternative formulation that also generalizes to the multi-sample case.

Why It's Important to Use the Right Multinomial Test

When correctly applied, multinomial tests should allow us to make good inferences. But, if we misapply these tests, we run the risk of having either an overly conservative or anti-conservative test.

Say we used the goodness-of-fit chi-squared statistic as defined by equation 1 to both perform a test of a die's fairness and a test comparing whether two dice are equivalent. Figure 2 shows the distributions of *p*-values we obtain across multiple iterations of these two different tests. In both cases, the null hypothesis is true.

Figure 2. Frequency of *p*-values Using the One-Sample Formulation for Chi-Squared



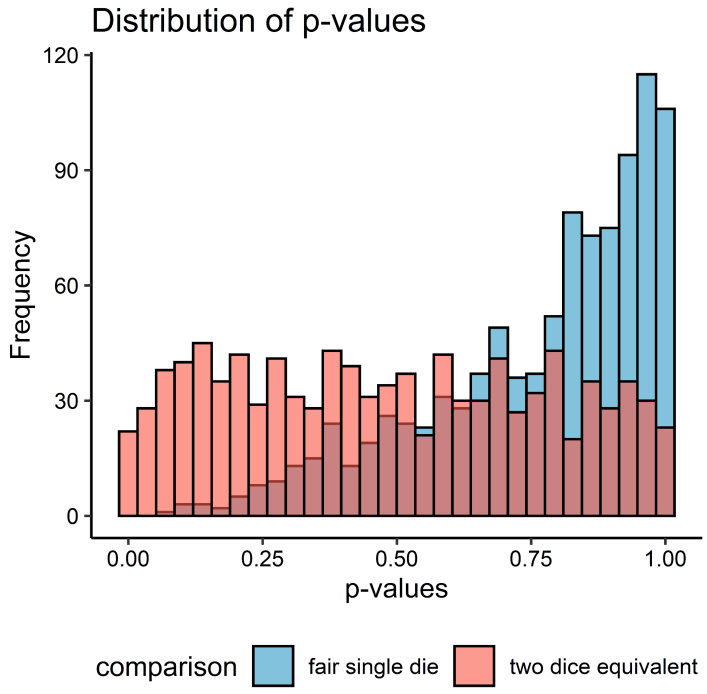
Since the null is true, we should expect to see a fairly uniform distribution of *p*-values. This is what we observe in the case of testing whether a single die is fair – but not in the case of testing whether two dice are equivalent. Instead, we see a disproportionate number of smaller *p*-values clustering at the left side of the figure.

What drives this result is uncertainty in the "referent" group. When we perform a goodness-of-fit test, one of the dice we toss is treated as the reference category that characterizes the expected frequencies against which we compare the observed frequencies of the other die. The one-sample formulation of the chi-squared test fails to take this uncertainty into account when testing the equivalence of two dice. As a result, this test tends to yield a higher number of false positives than expected.

We observe the reverse pathology in *p*-value distributions when we use the two-sample formulation of the chi-squared statistic to perform both of these tests as well. But, this time we observe problems in testing the fairness of a single die. As Figure 3 shows, while we observe a fairly uniform distribution of *p*-values when testing whether two dice are equivalent, the distribution of *p*-values for testing whether a die is fair is skewed to the right. This suggests a disproportionately higher number of large *p*-values than what we would expect under the null hypothesis.

The reason for this is that the two-sample formulation of the chi-squared statistic assumes that the expected frequencies for a fair die are measured with uncertainty. This therefore makes the test overly conservative, rejecting the null hypothesis far less than we would expect if the null were true.

Figure 3. Frequency of *p*-values Using the Two-Sample Formulation for Chi-Squared



An OES Example

OES is currently working on a project related to the Emergency Rental Assistance (ERA) program. One of the goals of this work is to assess whether there are disparities between eligible and recipient populations for ERA. The table below is from the ERA project's analysis plan.

Table 2. An Example of a Demographic Profile Featuring Inner and Outer Cells

	Non-LGBTQ+ persons	LGBTQ+ persons	*Ethnicity Percentage:*
Hispanic or Latino/a	A	B	20%
Not Hispanic or Latino/a	C	D	80%
Gender Percentage:	80%	20%	100%

One of the confirmatory analyses is to assess whether the distribution of demographic categories among recipients is statistically different from that among those eligible for ERA. This will be done using a chi-squared test comparing the observed frequencies in categories in the recipient and eligible groups.

The formulation of the chi-squared statistic will be critical in making sure that the study provides reliable inferences. A key question is, should the demographics of those eligible for ERA be treated as characterizing a well-defined population or as a sample? The answer to this question is pivotal, as Table 2 summarizes. The results are based on a simulation—the details of which are in the Code Appendix—based on testing whether the observed frequencies for four categories (A-D) are equivalent either to those in another sample or correspond to well-defined frequencies that characterize a population.

Table 2. Type I Error Rates

	Two-Sample chi-squared	One-Sample chi-squared
Independence Test is Correct	*0.059*	0.289
Goodness-of-Fit Test is Correct	0.0015	*0.058*

Note: Type I error rates are based on setting the level of the test (the *p*-value threshold for rejecting the null hypothesis) to $\alpha = 0.05$.

The table shows the type I error rates (the proportion of times the null hypothesis is rejected when the null hypothesis is true) for both kinds of chi-square tests (columns) and when one or the other test is the appropriate choice (rows). The conventional threshold of 0.05 has been chosen for rejecting the null. The diagonal of the table highlights the type I error rate when the chosen test is appropriate. Because the level of the test has been set to 0.05, we should expect to reject the null hypothesis when the null is true 5% of the time.

Consistent with Figures 2 and 3, when the recipient and eligible groups should be treated as samples, using the one-sample formulation for chi-squared will have anti-conservative properties. Here, it rejects the null with probability 0.289 rather than near 0.05. Conversely, when the eligible group should be treated as a population, using the two-sample formulation will be too restrictive. Here, it only rejects the null with probability 0.0015, rather than near 0.05.

It is impossible to devise a method to determine the correct test ex ante. Rather, the choice requires an informed judgment about the appropriateness of treating the group of eligible subjects as a sample or as a referent population. In other words:

- Are we interested in whether the group of ERA recipients is equivalent to the ERA eligible group? If yes, then a two-sample test is appropriate.
- Are we interested in whether the group of ERA recipients is a random sample of the ERA eligible group? If yes, then a goodness-of-fit test is appropriate.

How to Implement in R

This section shows how to implement the goodness-of-fit and independence chi-squared tests in R. For Stata users, we recommend reading this guidance produced by Reed College.

To perform chi-square tests in R, we will use `chisq.test()` from base R, which supports both the one-sample and two-sample versions of the chi-square test. It also supports calculating *p*-values asymptotically or via Monte Carlo (MC). As a general rule of thumb, MC p-values are our default recommendation. They are better behaved in cases where there are cells with few observations and should perform just as well as asymptotic p-values otherwise. However, when cells have many observations, asymptotic p-values may be acceptable, especially if MC p-values would be too computationally intensive.

The below code shows how to implement the one-sample, or goodness-of-fit chi-square test in R. For this example, the vector `x` contains observed frequencies for four different strata or categories in a sample of interest, and `p` defines the expected proportions.

```
x <- c(100, 200, 50, 10)   # observed frequencies
p <- c(0.3, 0.2, 0.1, 0.4) # expected distribution

chisq_out <- chisq.test(
  x = x,
  p = p
)
chisq_out
#>
#>  Chi-squared test for given probabilities
#>
```

```
#> data:  x
#> X-squared = 358.29, df = 3, p-value < 2.2e-16
```

To perform inference via Monte Carlo, simply write `simulate.p.value = TRUE`.

```
chisq_out_sim <- chisq.test(
  x = x,
  p = p,
  simulate.p.value = TRUE
)
chisq_out_sim
#>
#>  Chi-squared test for given probabilities with simulated p-value (based
#>  on 2000 replicates)
#>
#> data:  x
#> X-squared = 358.29, df = NA, p-value = 0.0004998
```

To perform a two-sample test, we put the observed frequencies for two samples into a 2-dimensional array. The below shows an example where `x` and `y` are vectors of observed frequencies for two different samples, which are combined into a 2-dimensional array called `xy` where columns denote the sample and rows show frequencies for a given category:

```
x <- c(100, 200, 50, 10)
y <- c(10, 400, 20, 20)
xy <- cbind(x, y)
xy
#>        x   y
#> [1,] 100  10
#> [2,] 200 400
#> [3,]  50  20
#> [4,]  10  20
```

When you give this 2-dimensional array to `chisq.test()`, it knows to implement the two-sample version of the test:

```
chisq_out <- chisq.test(
  x = xy
)
chisq_out
#>
#>  Pearson's Chi-squared test
#>
#> data:  xy
#> X-squared = 148.32, df = 3, p-value < 2.2e-16
```

Again, to obtain p-values via Monte Carlo we write:

```
chisq_out_sim <- chisq.test(
  x = xy,
  simulate.p.value = T
)
chisq_out_sim
#>
#>  Pearson's Chi-squared test with simulated p-value (based on 2000
#>  replicates)
#>
#> data:  xy
#> X-squared = 148.32, df = NA, p-value = 0.0004998
```

Code Appendix

The script below runs the simulation used to produce the results in Table 2.

```
quiet <- function(x) suppressWarnings(x)

# simulation set up
sims <- 2000
cats <- 4
nobs <- 1000

# when a test of independence is appropriate
replicate(
  sims,
  {
    g1 <- as.vector(rmultinom(1, nobs, rep(1 / cats, len = cats)))
    g2 <- as.vector(rmultinom(1, nobs, rep(1 / cats, len = cats)))
    data.frame(
      xx_pval = quiet(
        chisq.test(x = cbind(g1, g2))$p.value
      ),
      xp_pval = quiet(
        chisq.test(x = g1, p = g2, rescale.p = T)$p.value
      )
    ) -> to_return
    list(to_return)
  }
) -> p.values

do.call(
  'rbind', p.values
) -> p.values

apply(
  p.values, 2, function(x) mean(x <= 0.05)
) -> ss_error_rate

# when a goodness-of-fit test is appropriate
replicate(
  sims,
  {
    g1 <- as.vector(rmultinom(1, nobs, rep(1 / cats, len = cats)))
    p  <- as.vector(rep(nobs * (1 / cats), len = cats))
    data.frame(
      xx_pval = quiet(
        chisq.test(x = cbind(g1, p))$p.value
      ),
      xp_pval = quiet(
        chisq.test(x = g1, p = p, rescale.p = T)$p.value
```

```
      )
    ) -> to_return
    list(to_return)
  }
) -> p.values

do.call(
  'rbind', p.values
) -> p.values

apply(
  p.values, 2, function(x) mean(x <= 0.05)
) -> sp_error_rate

# compare type I errors
rbind(
  'sample-sample' = ss_error_rate,
  'sample-population' = sp_error_rate
)
#>                    xx_pval xp_pval
#> sample-sample       0.0590   0.289
#> sample-population   0.0015   0.058
```

Technical Appendix

The formulation for the two-sample case is limited in that it can only accommodate two-sample comparisons. However, there is an alternative formulation that also generalizes to multi-sample cases. The n-sample chi-squared statistic is given as

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}, \tag{3}$$

where *i* indexes *k* > 1 categories and *j* indexes *n* > 1 samples. The set of expected values for the *i*-th category for the *j*-th sample are defined as the outer product of the row and column marginals for a k-by-n matrix of observed frequencies.

The row marginals (or sums) denote the total number of observations per sample:

$$\mathbf{R} = [r_1 \ldots r_j \ldots r_n] : r_j = \sum_{i=1}^{n} O_{i,j}. \tag{4}$$

The column marginals (or sums) denote the total number of observations across samples in a category:

$$\mathbf{S} = [s_1 \ldots s_i \ldots s_k] : s_i = \sum_{j=1}^{k} O_{i,j}. \tag{5}$$

The integer *M* denotes the total number of observations across all samples:

$$M = \sum_{i=1}^{k} \sum_{j=1}^{n} O_{i,j}. \tag{6}$$

Finally, a matrix **E** with *k* rows for each category and *n* columns for each sample denotes the set of expected values which are defined as the outer product of the row and column marginals over *M*:

$$\mathbf{E} = (\mathbf{R} \otimes \mathbf{S}) / M. \tag{7}$$

$E_{i,j}$ then denotes the expected value for category *i* in sample *j*.

We can implement a multi-sample version of the test in R using the same framework that applies to the two-sample case:

Say we have three samples. We would construct a two-dimensional array with rows for each strata and columns for each sample:

```
x <- c(100, 200, 50, 10)
y <- c(10, 400, 20, 20)
z <- c(40, 40, 50, 70)
xyz <- cbind(x, y, z)
xyz
#>         x   y  z
#> [1,] 100  10 40
#> [2,] 200 400 40
#> [3,]  50  20 50
#> [4,]  10  20 70
```

And then we would perform the test like so:

```
chisq_out <- chisq.test(
  x = xyz
)
chisq_out
#>
#>  Pearson's Chi-squared test
#>
#> data:  xyz
#> X-squared = 411.86, df = 6, p-value < 2.2e-16
```