Chi-Square Tests

Overview

In practice, quality professionals sometimes need to collect categorical data to evaluate a process when it is not possible or convenient to collect continuous data. For example, a product may be categorized into two categories such as defective/nondefective or in more than two categories such as excellent, good, fair, and poor. Another example is a finance department that tracks the number of days that invoices are overdue into categories: 15 days or less, 16 to 30 days, 31 to 45 days, or 45 days or more. As a result, the variable of interest is the number of items that fall into each category.

Because of its versatility, chi-square tests are used for many applications that involve categorical data. In the Assistant, we use chi-square tests to:

- Test the goodness-of-fit for a multinomial distribution

  You can use this test to determine whether the data follow the same distribution as in the past. The distribution is defined as a multinomial distribution with a set of historical, or target, percents that define the percent of items that fall into each outcome category. The chi-square test jointly tests whether any percent significantly differs from its respective historical or target percent.

- Test the equality of % defectives for more than 2 groups

  You can use this test to determine whether there is a difference between the percent defectives of different groups. The groups differ by a characteristic of interest such as a product produced by different operators, by different plants, or at different times. The chi-square test jointly tests whether any percent defective significantly differs from any other percent defective.

- Test the association between two categorical variables

  You can use this test to determine whether a categorical outcome variable (Y) is related or associated with another categorical predictor variable (X). The chi-square test jointly

tests whether there is an association between the outcome variable and a predictor variable. In the Assistant, you can perform a Chi-Square Test for Association with a predictor variable (X) that contains two or more distinct values (two or more samples).

For more details on the chi-square test statistic, see Appendix A.

For methods that involve hypothesis testing, it is good practice to ensure that assumptions for the test are satisfied, that the test has adequate power, and that any approximations used to analyze the data produce valid results. For the chi-square tests, the assumptions are inherent to the data collection and we do not address them in data checks.

We focus our attention on power and validity of the approximation methods. The Assistant uses these approximation methods to perform the following checks on your data and reports the findings in the Report Card:

- Sample size
- Validity of test
- Validity of intervals

In this paper, we investigate how these data checks relate to chi-square tests in practice and we describe how we established the guidelines for the data checks in the Assistant.

Data checks

Sample size

Typically, the main objective for conducting a statistical test of hypothesis is to gather evidence to reject the null hypothesis of "no difference". If the samples are too small, the power of the test may not be adequate to detect a difference between the percent defectives that actually exists, which results in a Type II error. It is therefore crucial to ensure that the sample sizes are sufficiently large to detect practically important differences with high probability.

The sample size data check is based on the power of the test. This calculation requires that the user specifies a meaningful difference between an actual population parameter and the hypothesized null value. Because it was too difficult to determine and express this practical difference for Chi-Square Goodness-of-Fit and the Chi-Square Tests for Association, the Assistant only checks the sample size for the Chi-Square % Defective test with more than two samples.

Objective

If the data does not provide sufficient evidence against the null hypothesis, we want to determine whether the sample sizes are large enough for the test to detect practical differences of interest with high probability. Although the objective of sample size planning is to ensure that sample sizes are large enough to detect important differences with high probability, the samples should not be so large that meaningless differences become statistically significant with high probability.

Method

The power and sample size analysis is based on the formulas shown in Appendix B.

Results

When the data does not provide enough evidence against the null hypothesis and you do not specify a practical difference, the Assistant calculates the practical differences that can be detected with an 80% and a 90% probability based on the sample sizes. In addition, if the user provides a particular practical difference of interest, the Assistant calculates sample sizes that yield an 80% and a 90% chance of detecting that difference.

When checking for power and sample size, the Assistant Report Card for the Chi-Square % Defective test for more than two samples displays the following status indicators:

Status	Condition
	The test finds a difference among the % defectives, so power is not an issue. OR Power is sufficient. The test did not find a difference between the % defectives, but the sample is large enough to provide at least a 90% chance of detecting the given difference.
	Power may be sufficient. The test did not find a difference among the % defectives, but the sample is large enough to provide an 80% to 90% chance of detecting the given difference. The sample size required to achieve 90% power is reported.
	Power might not be sufficient. The test did not find a difference among the % defectives, and the sample is large enough to provide a 60% to 80% chance of detecting the given difference. The sample sizes required to achieve 80% power and 90% power are reported.
	The power is not sufficient (< 60%). The test did not find a difference among the % defectives. The sample sizes required to achieve 80% power and 90% power are reported.
	The test did not find a difference among the % defectives. You did not specify a practical difference between the % defectives to detect; therefore, the report indicates the differences that you could detect with 80% and 90% chance, based on your sample sizes and alpha.

Validity of test

The χ^2 test statistic only approximately follows a chi-square distribution. The approximation improves with larger sample sizes. In this section, we evaluate the approximation used to determine the minimum sample size needed for accurate results.

The chi-square approximation to the test statistic is evaluated by examining the impact of small expected cell counts on the Type I error rate (alpha). By using Type I error to evaluate the validity of the test, we develop a rule to ensure that:

- The probability of rejecting the null hypothesis when it is true is small and close to the desired Type I error rate.
- The tail of the null distribution can be reasonably approximated, which is important to accurately calculate the p-value.

Using a standard approach, we defined a small expected cell count as a cell that has an expected cell count less than or equal to 5.

We developed two models to define the proportions under the null hypothesis: the proportions perturbed model and the equal proportion model. For more details, see Appendix C. Both models are used in the simulations referred to later in this paper. The models are used for each

of the chi-square tests with one exception: the proportions perturbed model does not apply to Chi-Square % Defective test for more than two samples.

The Validity of Test data check applies to all chi-square tests in the Assistant. Each data check is described below.

Chi-Square Goodness-of-Fit

Objective

We evaluated the chi-square approximation to the test statistic by investigating the impact of the magnitude and the frequency of the small expected counts on the Type I error rate.

Method

Samples of size n were drawn from a multinomial distribution with the proportions described in the proportions perturbed or equal proportion models (see Appendix C). For each condition, we performed 10,000 Chi-Square Goodness-of-Fit tests with a target significance level of 0.05. For each test, we calculated the actual Type I error as $\frac{Number\ of\ rejected\ tests}{Number\ of\ replicates\ (10000)}$. We defined the range for acceptable Type I error rates from [0.03 – 0.07] and recorded the minimum sample size with a Type I error rate in that range.

Results

The simulation results showed that target cell counts less than 1.25 may lead to incorrect p-values when the percentage of small target cell counts is less than or equal to 50%. Also, target cell counts less than 2.5 may lead to incorrect p-values when the percentage of small target cell counts is greater than 50%. See Appendix D for more details.

When checking the validity of the Chi-Square Goodness-of-Fit test, the Assistant Report Card displays the following status indicators:

Status	Condition
	The minimum target cell count is greater than or equal to 1.25 when the percentage of small target cell counts is less than or equal to 50% OR The minimum target cell count is greater than or equal to 2.5 when the percentage of small target cell counts is greater than 50%. Your sample is large enough to obtain sufficient target counts. The p-value for the test should be accurate.
	If the above conditions do not hold.

Chi-Square Test for Association

Objective

We evaluated the chi-square approximation to the test statistic by investigating the impact of the magnitude and the frequency of the small expected counts on the Type I error rate.

Method

Samples of size n_i are drawn from a multinomial distribution with the proportions defined by the proportions perturbed or equal proportion models (see Appendix C). For simplicity, we chose $n_i = n\ \forall i$. For each condition, we performed 10,000 Chi-Square Tests for Association with a target significance level of 0.05. For each test, we calculated the actual Type I error rate as $\frac{Number\ of\ rejected\ tests}{Number\ of\ replicates\ (10000)}$. We defined the range for acceptable Type I error rates from [0.03 – 0.07] and recorded the minimum sample size with a Type I error rate in that range.

Results

We found that the minimum expected cell count depends on the number of X values and the percentage of small expected cell counts.

- For the proportions perturbed model, when the percentage of small expected cell counts is less than or equal to 50%, the minimum expected cell counts are ≤ 2 and ≤ 1 for number of X values equal to (2 or 3) and (4, 5, or 6) respectively. In addition, when the percentage of small expected cell counts is > 50%, the minimum expected cell counts are ≤ 3 and ≤ 1.5 for number of X values equal to (2 or 3) and (4, 5, or 6) respectively.
- For the equal proportion model, the minimum expected cell count is ≤ 2 when the number of X values equal to (2 or 3) and minimum expected cell count ≤ 1.5 when the number of X values equal to (4, 5, or 6).

For more details, see Appendix E.

When checking the validity of the Chi-Square Test for Association, the Assistant Report Card displays the following status indicators:

Status	Number of X variable values	Condition
	2 or 3	The minimum expected cell count is greater than or equal to 2 when the percentage of small expected cell counts (less than or equal to 5) is less than or equal to 50%. The minimum expected cell count is greater than or equal to 3 when the percentage of small expected cell counts (less than or equal to 5) is greater than 50%.

Status	Number of X variable values	Condition
	4, 5 or 6	The minimum expected cell count is greater than or equal to 1 when the percentage of small expected cell counts (less than or equal to 5) is less than or equal to 50%. The minimum expected cell count is greater than or equal to 2 (for convenience rounded 1.5 to 2) when the percentage of small expected cell counts (less than or equal to 5) is greater than 50%.
	All cases	If above conditions do not hold.

Chi-Square % Defective Test for more than two samples

Objective

We evaluated the chi-square approximation to the test statistic by investigating the impact of the magnitude and the frequency of the small expected counts on the Type I error rate.

Method

We defined the models p = $p_i = p_j \; \forall i, j$ where p = 0.001, 0.005, 0.01, 0.025 and 0.25. Samples of size n_i were drawn from a binomial distribution with the values of p_i described above. For simplicity, we chose $n_i = n \; \forall i$. For each condition, we performed 10,000 Chi-Square % Defective Tests with a target significance level of 0.05. For each test, we calculated the actual Type I error as $\frac{Number\ of\ rejected\ tests}{Number\ of\ replicates\ (10000)}$. We defined the range for acceptable Type I error rates from [0.03 – 0.07] and recorded the minimum sample size with a Type I error rate in that range.

Results

When there are 3 to 6 X values, a minimum expected number of defectives and nondefectives greater than or equal to 1.5 yields a Type I error rate for the test in the interval [0.03, 0.07]. When there are 7 to 12 X values, a minimum expected number of defectives and nondefectives greater than or equal to 1 yields a Type I error rate for the test in the interval [0.03, 0.07].

For more details, see Appendix F.

When checking the validity of the Chi-Square % Defective test for more than two samples, the Assistant Report Card displays the following status indicators:

Status	Number of X values	Condition
	3 to 6	The minimum expected number of defectives and nondefectives is greater than or equal to 1.5.
	7 to 12	The minimum expected number of defectives and nondefectives is greater than or equal to 1.
	All cases	If above conditions do not hold.

Validity of intervals

The comparison intervals in the Chi-Square % Defective for more than two samples and Chi-Square Goodness-of-Fit test are based on the normal approximation. In addition, the individual confidence intervals in the Chi-Square Goodness-of-Fit test are based on the normal approximation. In this section, we evaluate the validity of the normal approximation. According to the general rule found in most statistical textbooks, the approximate confidence interval is accurate if the observed counts are at least 5.

The Validity of intervals data check applies to Chi-Square % Defective for more than two samples and Chi-Square Goodness-of-Fit test.

Chi-Square % Defective for more than two samples

Objective

We wanted to evaluate the general rule for the minimum number of defectives and nondefectives observed in each sample to ensure that the approximate confidence intervals are accurate.

Method

We first define the intervals that are used in the comparison chart. The endpoints of the intervals are defined so that with an overall error rate of approximately $\propto$, any interval that fails to overlap indicates population % defectives that are different. See Appendix G for the formulas used.

The comparison intervals are based on paired comparison confidence intervals. For more details, see the Comparison intervals section in the Assistant White Paper for One-Way ANOVA. We use a normal approximation confidence interval for each pair ($p_i - p_j$) and then use a Bonferroni multiple comparison procedure to control the overall experiment-wise error rate. Therefore, we

only need to evaluate the validity of one of the intervals in the paired comparison procedure to understand the effect of the normal approximation on the comparison intervals.

Results

To evaluate the validity of the normal approximation, we only need to examine how the approximation affects one interval for the difference between % defectives. Therefore, we can simply use the general rule developed for the 2-Sample % Defective case. For more details, see the 2-sample % defective test methods section in the Assistant White Paper for the 2-Sample % Defective test. The simulation results in 2-Sample % Defective test indicate that the accuracy of the approximate confidence interval for the difference between % defectives is generally reliable when samples are sufficiently large --that is, when the observed number of defectives and the observed number of nondefectives in each sample is at least 5.

When checking the validity of the intervals for the Chi-Square % Defective test for more than two samples, the Assistant Report Card displays the following status indicators:

Status	Condition
	All samples have at least 5 defectives and 5 nondefectives. The comparison intervals should be accurate.
	If above condition do not hold.

Chi-Square Goodness-of-Fit

Objective

We wanted to evaluate the general rule for the minimum number of defectives and nondefectives observed in each sample to ensure that the approximate confidence intervals are accurate.

Method

The Assistant's Chi-Square Goodness-of-Fit test includes comparison and individual confidence intervals. We utilize the standard normal approximation intervals for proportions and correct for multiple intervals using the Bonferroni correction (Goodman, 1965). Thus, Bonferroni simultaneous intervals are calculated as follows:

$$p_{iLower} = p_i - Z_{\alpha/2k}\sqrt{\frac{\mathrm{p_i}(1 - \mathrm{p_i})}{\mathrm{N}}}$$

$$p_{iUpper} = p_i + Z_{\alpha/2k}\sqrt{\frac{\mathrm{p_i}(1 - \mathrm{p_i})}{\mathrm{N}}}$$

The endpoints of the intervals are defined so that with an overall error rate of approximately $\propto$, any interval that does not contain the target proportion value indicates that the actual proportion is different from its corresponding target proportion. The individual intervals utilize the same form as the Bonferroni intervals but do not correct for the multiple intervals by using $Z_{\alpha/2}$.

Results

Both of the approaches described above follow a methodology that is similar to the one defined in the Assistant's 2-Sample % Defective test. Therefore, we can use similar rules for the validity of the normal approximation that were developed for that test. For more details, see the 2-sample % Defective test methods section in the Assistant White Paper for the 2-Sample % Defective test. In that paper, we concluded that the comparison intervals and the individual confidence intervals may not be accurate when the sample counts are less than 5.

When checking the validity of the intervals for the Chi-Square Goodness-of-Fit test, the Assistant Report Card displays the following status indicators:

Status	Condition
	All sample counts are at least 5. The intervals should be accurate.
	There are sample counts less than 5.

References

Agresti, A. (1996). An introduction to categorical data analysis. New York, NY: Wiley.

Read, T. & Cressie, N. (1988). Goodness-of-fit statistics for discrete multivariate data. New York, NY: Springer-Verlag.

Fienberg, S. (1980). The analysis of cross-classified categorical data. Cambridge, MA: MIT Press.

Goodman, L. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics, 7*, 247-254.

Appendix A: Chi-square test statistic

The Assistant uses a chi-square test statistic of the form:

$$x^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} = observed counts, as defined in table below:

Case	O_{ij}
Test the goodness of fit for a multinomial distribution	The observed count for the i^{th} outcome is defined as O_{i1}.
Test the equality of more than 2 % defectives	The observed number of defective items and non-defective item for the i^{th} sample is defined as O_{i1} and O_{i2} respectively.
Test the association between two categorical variables	The observed counts for the i^{th} value of the X variable and j^{th} value of the Y variable is defined as O_{ij}.

E_{ij} = Expected count as defined in the table below:

Case	E_{ij}
Test the goodness of fit for a multinomial distribution	$E_{i1} = np_i$ $i = 1, \ldots, k$ (k = number of outcomes) n = sample size p_i = historical proportions $\sum_i p_i = 1$
Test the equality of more than 2 % defectives	$E_{i1} = n_i p$ (for defectives) $E_{i2} = n_i(1-p)$ (for nondefectives) $i = 1, \ldots, k$ (k = number of samples) n_i = i^{th} sample size p = overall proportion defective

Case	E_{ij}
Test the association between two categorical variables	$E_{ij} = \frac{(n_{i.} n_{.j})}{n_{..}}$ $i = 1, \dots, m$ (m = number of X values) $j = 1, \dots, k$ (k = number of Y values) $n_{i.}$ = total count for the i^{th} value of X variable $n_{.j}$ = total count for the j^{th} value of Y variable $n_{..}$ = overall sample size

Appendix B: Power for Chi-Square % Defective test for more than two samples

We use a noncentral chi-square distribution to calculate the power of the test that $p_i = p_j = p \; \forall i, j$. The noncentrality parameter depends on n_i and $p_i \forall i$

where

n_i = the sample size for the i^{th} sample

Each p_i represents an alternative proportion (see the next section in this Appendix, Calculation of Alternative Proportions) calculated from the proportion difference = δ.

We calculate the noncentrality parameter of the chi-square distribution as:

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$O_{i1 = n_i p_i}$

$O_{i2 = n_i (1 - p_i)}$

and calculate the power of the test as

$$\text{Prob}(X \geq x_{1-\alpha} \mid \chi^2)$$

where

X = is a random variable from a noncentral chi-square distribution with noncentrality parameter χ^2.

$x_{1-\alpha}$ = inverse cdf evaluated at $1 - \alpha$ for a central chi-square distribution.

Calculation of alternative proportions

We defined the alternative proportions as follows:

$$p_i = p_c + \frac{n_j}{n_i + n_j} \delta$$

$$p_j = p_c - \frac{n_i}{n_i + n_j} \delta$$

$$p_m = p_c \forall m \neq i, j$$

$$0 < \delta < 1$$

where

$$p_c = \frac{1}{N_T}\sum_{i=1}^{k} n_i \hat{p}_i$$

$\hat{p}_i$ = sample proportion defective items for the i^{th} sample.

NT = total number of observations.

n_i = sample size for i^{th} sample.

For some differences δ , $p_i > 1$ or $p_j < 0$. Therefore, we develop the following rules:

If $p_j < 0$

$$p_i = \delta$$
$$p_j = 0$$
$$p_m = \frac{\delta}{2} \; \forall m \neq i, j$$

If $p_i > 1$

$$p_i = 1$$
$$p_j = 1 - \delta$$
$$p_m = 1 - \frac{\delta}{2} \; \forall m \neq i, j$$

Using the two smallest values of n_i results in the minimum power and using the two largest values of n_i results in the maximum power.

Appendix C: Proportions perturbed model and equal proportion model

Proportions perturbed model

Following Read and Cressie (1988), we define the set of proportions under the null hypothesis as follows:

We choose δ near k - 1 (where k = number of proportions for each sample) and define a set of small p_i as

$$p_i = \frac{(1 - \frac{\delta}{k-1})}{k} \text{ for } i = 1, \ldots, r$$

and the remaining p_i as

$$p_i = \frac{(1 - \sum_{i=1}^{r} p_i)}{(k-r)} \text{ for } i = r+1, \ldots, k$$

The values we used for δ in the simulations are listed in Table 1.

Table 1 δ used in the simulations with resulting small p_i

k	δ	$p_{i=1,\ldots,r}$
3	1.95	0.008
4	2.95	0.004
5	3.90	0.005
6	4.90	0.003

For each k, we varied r = 1,..., k – 1 to change the size of the set of small p_i's. For example, for k = 3, we obtained the following two models described in Table 2.

Table 2 The values of p_i for k = 3 using the proportions perturbed model

r	p1	p2	p3
1	0.008	0.496	0.496
2	0.008	0.008	0.984

Equal proportion model

To obtain a model where 100% of the expected cell counts are small, we use an equal proportion model defined by

$$p_i = \frac{1}{k} \forall\ i$$

Using this model, with a very small sample size, all of the expected cell counts are considered small. With an equal proportion model, the sample sizes need to be very small to achieve a small expected cell count, which likely will not occur in practice.

Appendix D: Validity of test for chi-square goodness-of-fit

For the proportions perturbed model, we plotted the minimum expected cell count needed to achieve a Type I error rate in the interval [0.03, 0.07] against the % of small expected cell counts, as shown Figure 1.

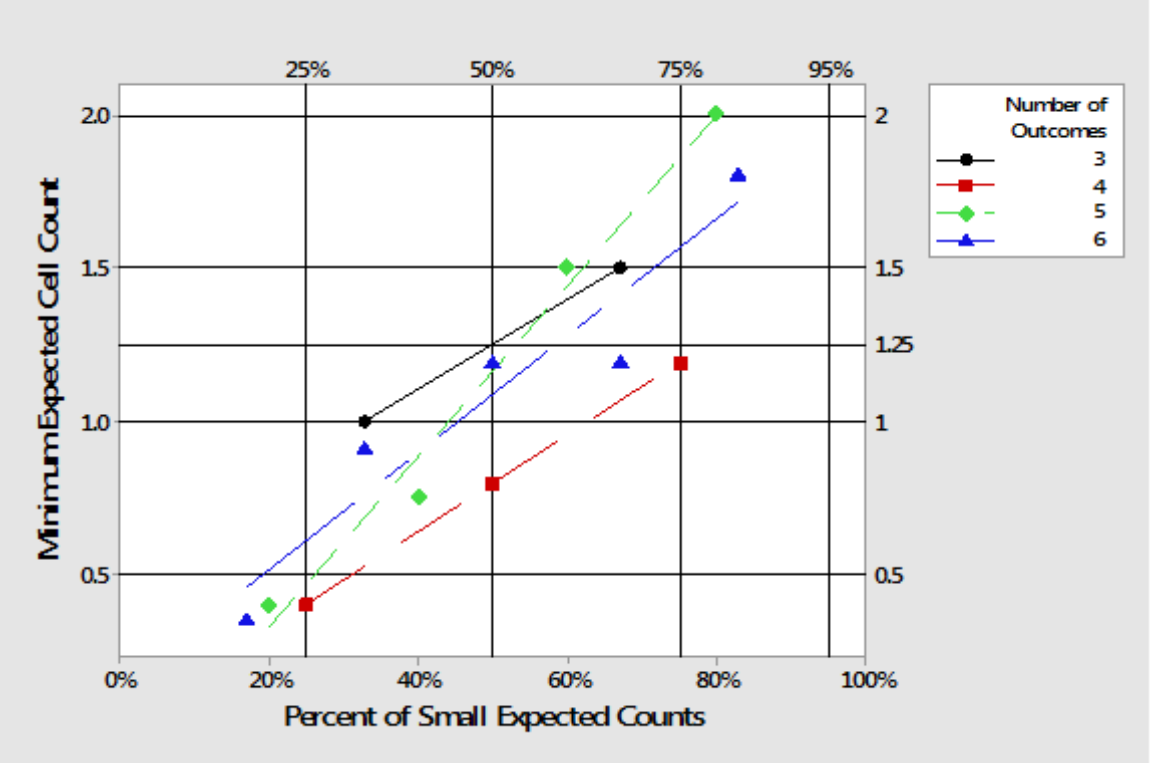

Figure 1 Minimum expected cell counts needed to achieve a Type I error rate in the interval [0.03, 0.07] versus the percent of small expected cell counts.

In Figure 1, when the percent of small expected cell counts is less than 50%, the minimum expected cell counts are less than or equal to 1.25. All minimum expected cell counts are less than or equal to 2. Based on these simulation results, the rules we use in the Assistant Report Card are conservative.

Next we performed the same simulation using the equal proportion model to define the null distribution. Table 4 summarizes the results from the simulation using an equal proportion model.

Table 4 Minimum expected cell count to obtain a Type I error rate in the interval [0.03, 0.07]

k	Minimum expected cell count
3	2.5
4	1.25
5	1
6	1.4

As indicated above, the equal proportion model leads to cases where 100% of the cell counts are small. Table 4 shows that all the minimum expected cell counts are less than or equal to 2.5, which supports the rules we use in the Assistant Report Card.

Appendix E: Validity of test for chi-square test for association

For the proportions perturbed model, we plotted the minimum expected cell count needed to achieve a Type I error rate in the interval [0.03, 0.07] against the % of small expected cell counts for each number of X values, as shown in Figure 2.

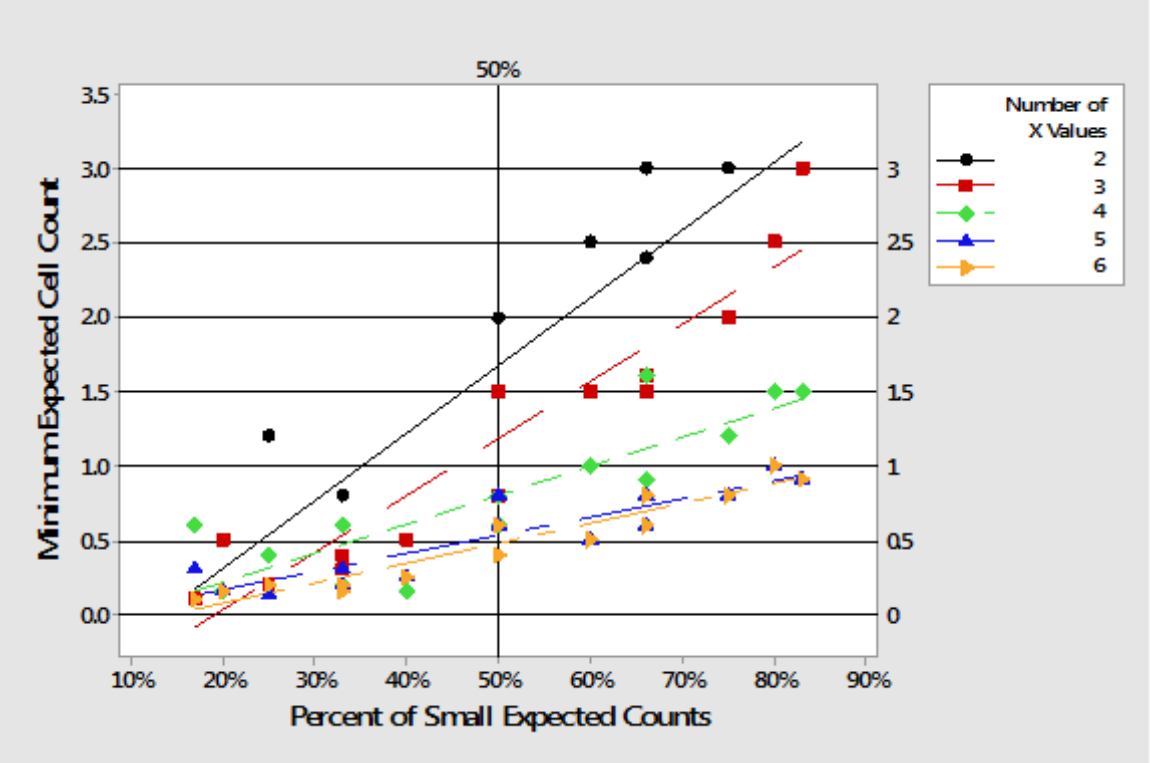

Figure 2 Minimum expected cell counts needed to achieve a Type I error rate in the interval [0.03, 0.07] versus the percent of small expected cell counts.

Figure 2 indicates that the minimum expected cell count depends on the number of X values and the percent of small expected cell counts.

Figure 2 indicates that when the percent of small expected cell counts is ≤ 50%, the minimum expected cell counts are ≤ 2 and ≤ 1 for number of X values equal to 2 or 3 and 4, 5, or 6 respectively. In addition, when the percent of small expected cell counts is > 50%, the minimum expected cell counts are ≤ 3 and ≤ 1.5 for number of X values equal to 2 or 3 and 4, 5, or 6 respectively.

For the equal proportion model, we plotted the minimum expected cell count against the number of X values (m) and the number of Y values (k), as shown in Figure 3.

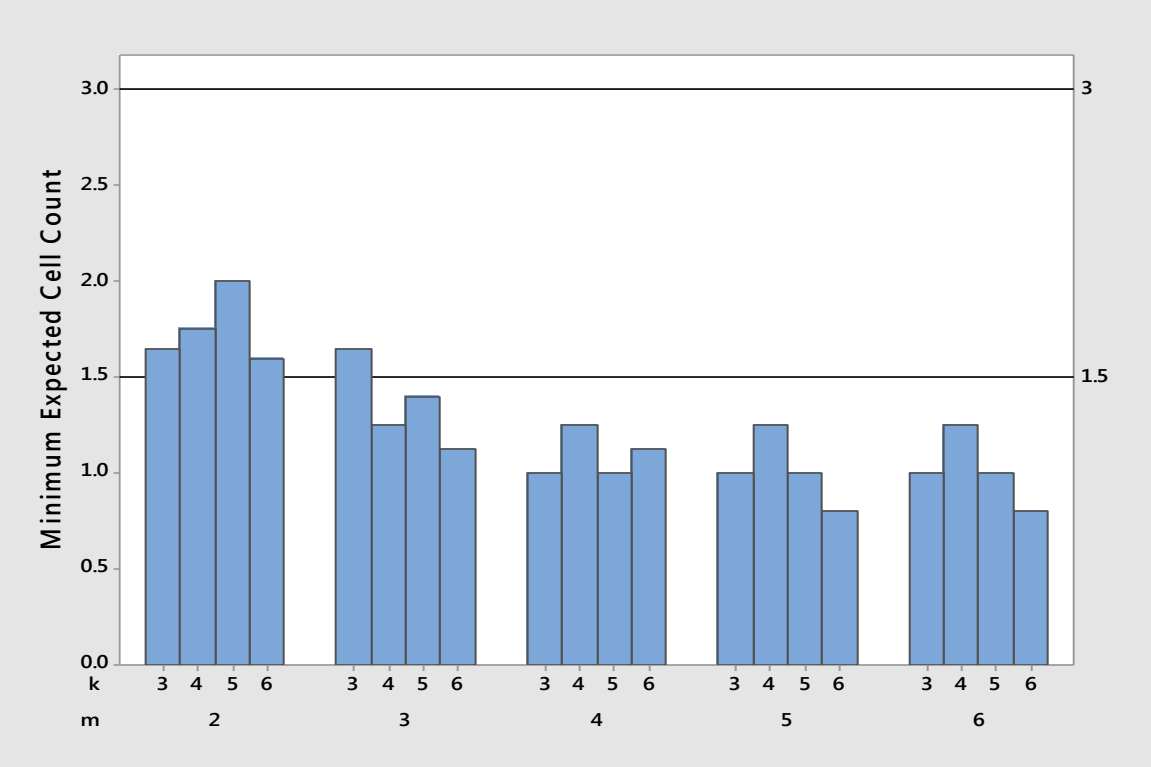

Figure 3 Minimum expected cell count needed to achieve a Type I error rate in the interval [0.03, 0.07] versus the X values (m) and Y values (k)

Figure 3 indicates that the minimum expected cell count is ≤ 2 when the number of X values equal to 2 or 3 and minimum expected cell count ≤ 1.5 when the number of X values equal to 4, 5, or 6. Based on these simulation results, the rules in the Assistant Report Card are conservative.

Appendix F: Validity of test for Chi-Square % Defective for more than two samples

For each p and each m = 3, 4, 5,..., 12, we plotted the minimum expected cell count. The results are displayed Figures 4 and 5.

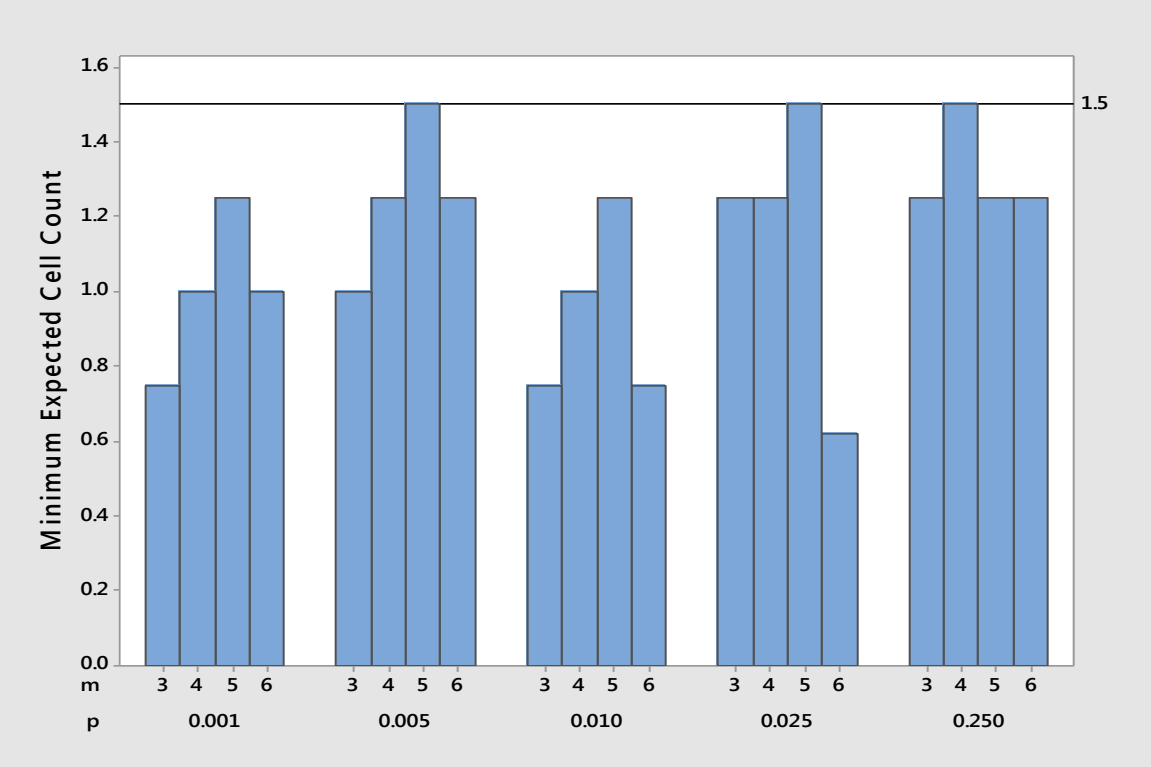

Figure 4 Minimum expected cell count needed to achieve a Type I error rate in the interval [0.03, 0.07] versus the number of X values (m = 3 to 6)

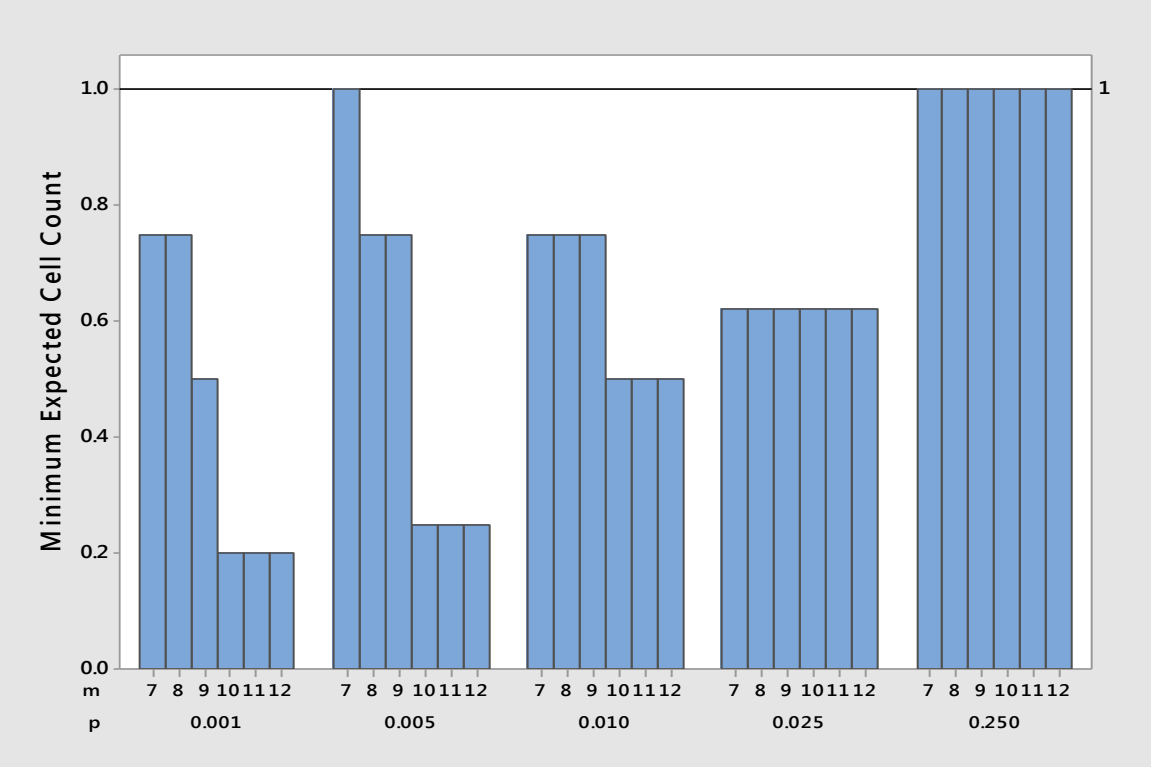

Figure 5 Minimum expected cell count needed to achieve a Type I error rate in the interval [0.03, 0.07] versus the number of X values (m = 7 to 12)

When the number of X values is equal to 3, 4, 5 or 6, an expected cell count greater than or equal to 1.5 yields a Type I error rate for the test in the interval [0.03, 0.07]. When the number of X values is equal to 7, 8, 9,..., 12, an expected cell count greater than or equal to 1 yields a Type I error rate for the test in the interval [0.03, 0.07].

Appendix G: Comparison intervals for Chi-Square % Defective for more than two samples

The lower and upper bound for p_i are defined as follows:

$p_{iLower} = p_i - Z_{\alpha/c}X_i$

$p_{iUpper} = p_i + Z_{\alpha/c}X_i$

where

c = number of comparisons = k (k - 1) /2

where k is the number of samples.

$Z_{\alpha/c}$ = $(1 - \frac{\alpha}{2c})$ percentile for a Normal distribution with mean = 0 and standard deviation = 1

$X_i = ((k-1)\sum_{j \neq i} b_{ij} - \sum\sum_{1 \leq j < l \leq k} b_{jl}) / ((k-1)(k-2))$

where

$$b_{ij} = \sqrt{\frac{p_i(1 - p_i)}{n_i} + \frac{p_j(1 - p_j)}{n_j}}$$