卡方检验

概述

在实际应用中，质量专员有时需要收集类别数据，以在不可能或者不方便收集连续数据时评估过程。例如，一个产品可以分成两类，如不良品/良品或分成两个以上的类别，如优秀、良好、一般和差。再比如跟踪发票逾期天数的财务部门可将发票逾期天数分成以下几类：15 天以内、16 至 30 天、31 天至 45 天，或 45 天以上。因此，所研究的变量是每个类别中包括的项数。

功能全面的卡方检验可进行涉及类别数据的许多应用。在“协助”中，我们用卡方检验：

- 检验多项分布的拟合优度

  您可以使用此检验确定数据是否遵循和过去一样的分布。先将分布定义为多项分布，这里每个项出现的概率可以定义为落入每个类别项的实际比率或历史比率或目标比率。卡方检验可连带检验任何比率是否显著不同于其各自的历史和目标比率。

- 检验 2 组以上不良品率的相等性

  您可以使用此检验确定不同组的不良品率之间是否存在差异。这些组的特征各不相同，例如产品由不同的运营商、不同的工厂或在不同的时间生产。卡方检验可连带检验某个不良品率是否显著不同于任何其它不良品率。

- 检验两个类别变量之间的相关性

  您可以使用此检验确定类别结果变量 (Y) 是否与其他类别预测变量 (X) 相关或关联。卡方检验可连带检验结果变量和预测变量之间是否相关。在“协助”中，您可以用包含两个或多个不同值（两个或两个以上的样本）的预测变量 (X) 执行相关性卡方检验。

有关卡方检验统计量的详细信息，请参见附录 A。

对于涉及假设检验的方法，好的做法是确保满足检验的假设条件，使检验发挥足够的功效，以及任何使用分析数据的近似值都仍然能获得有效的结果。对于卡方检验，这些假设与数据收集有关，我们将不在数据检查中探讨这些问题。

我们将注意力集中在近似方法的功效和有效性上。“协助”将使用这些近似方法，对数据进行以下检查并在报告卡中报告研究结果：

- 样本量

- 检验的有效性
- 区间的有效性

在本书中，我们探讨了如何将这些数据检查与实际应用中的卡方检验关联起来，并介绍了如何为“协助”中的数据检查确立指导方针。

数据检查

样本量

通常情况下，执行统计假设检验的主要目的是收集证据，拒绝“无差异”的原假设。如果样本量太小，检验可能无法检测到不良品率之间切实存在的差异，从而导致 II 类错误。因此，一定要确保样本量足够大，可以高概率地检测到有实际意义的差异。

数据的样本量检查基于检验的功效。这种计算要求用户在实际总体参数和假设原值之间指定一个有意义的差异。由于卡方拟合优度检验和相关性卡方检验极难确定和表示这一实际差异，“协助”只检查两个以上样本的卡方不良品率的样本量。

目标

如果数据没有提供足够的证据来否定原假设，我们要确定样本量是否足够大，以便检验能够高概率地检测到所需的实际差异。尽管计划样本量的目标是确保样本量足够大，可以高概率地检测到重要差异，它们也不应该如此之大，使得无意义的差异高概率地成为显著性差异。

方法

功效和样本量分析基于附录 B 中所示的公式

结果

当数据未提供针对原假设的足够证据且您没有指定有实际意义的差异时，“协助”可以基于现在的样本量计算出以 80% 和 90% 的概率保证能检测到的实际差异。另外，如果用户提供了所需的特定的实际差异，“协助”则会计算出能检测出此差异的机会为 80% 和 90% 的样本量。

当检查功效和样本量时，用于两个以上样本的卡方不良品率检验的“协助报告卡”将显示以下状态指标：

状态	条件
	检验发现到不良品率之间的差异，因此功效不是问题。 或 功效足够。检验没有发现不良品率之间的差异，但是样本大到足以使检测到给定差异至少有 90% 的机会。
	功效可能足够。检验没有发现不良品率之间的差异，但是样本大到足以使检测到给定差异有 80% ～90% 的机会。报告了为实现 90% 的功效所需的样本量。
	功效可能不够。检验没有发现不良品率之间的差异，并且样本大到足以使检测到给定差异有 60% ～80% 的机会。报告了为实现 80% 和 90% 的功效所需的样本量。
	功效不够 (< 60%)。检验没有发现不良品率之间的差异。报告了为实现 80% 和 90% 的功效所需的样本量。

状态	条件
	检验没有发现不良品率之间的差异。您没有指定要检测的不良品率之间的实际差异；因此，该报告将根据样本量和 alpha 值指出您可以有 80% 和 90% 机会所能检测到的差异。

检验的有效性

χ^2检验统计量仅近似遵循卡方分布规则。随着样本量的增大，近似程度会提高。在本节中，我们将用确定为了得到准确结果所需的最小样本量的方法来评估近似性。

通过考查小期望单元格计数对 I 类错误率（alpha 值）的影响来评估检验统计量的卡方近似性。为能使用 I 类错误率来评估检验的有效性，我们研发了一项规则，以确保：

- 拒绝原本成立的原假设的概率小且近似于我们所需的 I 类错误率。
- 能合理地估计原假设分布的尾部，这对精确计算检验的 p 值很重要。

使用标准方法，我们将小期望单元格计数定义为期望单元格计数小于或等于 5。

我们开发了两个模型来定义原假设下的比率：比率扰动模型和等比率模型。有关详细信息，请参见附录 C。两种模型都在模拟中使用，后文会有所介绍。这些模型被用于各个卡方检验，只有一种情况例外：比率扰动模型并不适用于两个以上样本的卡方不良品率检验。

检验有效性的数据检查适用于“协助”中的所有卡方检验。各个数据检查介绍如下。

卡方拟合优度

目标

我们通过研究小期望单元格计数的规模和频率对 I 类错误率的影响，评估了检验统计量的卡方近似性。

方法

大小为 n 的样本来自一个比率根据比率扰动模型或等比率模型定义的多项分布（参见附录 C）。对于每一种情况，我们用 0.05 的目标显著性水平进行了 10,000 次卡方拟合优度检验。对于每一次检验，我们按 $\frac{\text{不合格的检验数}}{\text{仿行数 (10000)}}$ 计算出了实际的 I 类错误率。我们定义了可接受的 I 类错误率的范围，即 [0.03 - 0.07]，并对 I 类错误率落入此范围内的那些检验记录了最小样本量。

结果

模拟结果表明，当小期望目标单元格计数的百分比小于或等于 50% 时，目标单元格计数小于 1.25 可能导致 P 值不正确。另外，当小期望目标单元格计数的百分比大于 50% 时，目标单元格计数小于 2.5 可能导致 p 值不正确。有关详细信息，请参见附录 D。

当检查卡方拟合优度检验的有效性时，“协助报告卡”会显示以下状态指标：

状态	条件
	当小期望目标单元格计数的百分比小于或等于 50% 时，最小期望的目标单元格计数大于或等于 1.25 或 当小期望目标单元格计数的百分比大于 50% 时，最小期望的目标单元格计数大于或等于 2.5。 您的样本足够大，可以获得足够的目标计数。用于检验的 p 值是准确的。
	如果上述条件不成立。

相关性的卡方检验

目标

我们通过研究小期望单元格计数的规模和频率对 I 类错误率的影响，评估了检验统计量的卡方近似性。

方法

大小为 n_i 的样本来自一个其比率根据比率扰动模型或等比率模型定义的多项分布（参见附录 C）。为简单起见，我们选择了 $n_i = n\,\forall i$。对于每一种情况，我们用 0.05 的目标显著性水平进行了 10,000 次相关性卡方检验。对于每一次检验，我们按 $\frac{\text{不合格的检验数}}{\text{仿行数 (10000)}}$ 计算出了实际的 I 类错误率。我们定义了可接受的 I 类错误率的范围，即 [0.03 – 0.07]，并对 I 类误差率落入此范围内的那些检验记录了最小样本量。

结果

我们发现最小期望单元格计数取决于 X 值的个数和小期望单元格计数的百分比。

- 对于比率扰动模型，当小期望单元格计数的百分比小于或等于 50% 时，X 值个数分别等于（2 或 3）和（4、5 或 6）的最小期望单元格计数 ≤ 2 和 ≤ 1。而当小期望单元格计数的百分比大于 50% 时，X 值个数分别等于（2 或 3）和（4、5 或 6）的最小期望单元格计数 ≤ 3 和 ≤ 1.5。
- 对于等比率模型，当 X 值个数等于（2 或 3）时，最小期望单元格计数 ≤ 2，当 X 值个数等于（4、5 或 6）时，最小期望单元格计数 ≤ 1.5。

有关详细信息，请参见附录 E。

当检查相关性卡方检验的有效性时，“协助报告卡”会显示以下状态指标：

状态	X 变量值个数	条件
	2 或 3	当小期望单元格计数（小于或等于 5）的百分比小于或等于 50% 时，最小期望单元格计数大于或等于 2。 当小期望单元格计数（小于或等于 5）的百分比大于 50% 时，最小期望单元格计数大于或等于 3。
	4、5 或 6	当小期望单元格计数（小于或等于 5）的百分比小于或等于 50% 时，最小期望单元格计数大于或等于 1。 当小期望单元格计数（小于或等于 5）的百分比大于 50% 时，最小期望单元格计数大于或等于 2（为方便起见，将 1.5 取整为 2）。
	所有实例	如果上述条件不成立。

两个以上样本的卡方不良品率检验

目标

我们通过研究小期望单元格计数的规模和频率对 I 类错误率的影响，评估了检验统计量的卡方近似性。

方法

我们定义了模型 p = $p_i = p_j \, \forall i,j$，其中 p = 0.001、0.005、0.01、0.025 和 0.25.。大小为 n_i 的样本来自一个值为 p_i 的二项式分布，如上所述。为简单起见，我们选择了 $n_i = n \, \forall i$。对于每一种情况，我们用 0.05 的目标显著性水平进行了 10,000 次卡方不良品率检验。对于每一次检验，我们按 $\frac{\text{不合格的检验数}}{\text{仿行数 (10000)}}$ 计算出了实际的 I 类错误率。我们定义了可接受的 I 类错误率的范围，即 [0.03 – 0.07]，并对 I 类误差率落入此范围内的那些检验记录了最小样本量。

结果

当 X 值在 3 至 6 之间时，最小期望单元格中的不良品和良品数都大于或等于 1 时，其产生的 I 类错误率将落入 [0.03，0.07] 区间中。当 X 值在 7 至 12 之间时，最小期望单元格中的不良品和良品数都大于或等于 1 时，其产生的 I 类错误率将落入 [0.03，0.07] 区间中。

有关详细信息，请参见附录 F。

检查两个以上样本的卡方不良品率检验的有效性时，“协助报告卡”会显示以下状态指标：

状态	X 值个数	条件
	3 至 6	最小期望单元格中的不良品和良品数大于或等于 1.5。
	7 至 12	最小期望单元格中的不良品和良品数大于或等于 1。
	所有实例	如果上述条件不成立。

区间的有效性

两个以上样本的卡方不良品率检验和卡方拟合优度检验中的比较区间基于正态近似分布。此外，卡方拟合优度检验中的各个置信区间也基于正态近似分布。在本节中，我们评估了正态近似分布的有效性。根据大多数统计教材中提供的一般规则，如果观测到的计数至少为 5，则近似置信区间准确。

区间数据检查的有效性适用于两个以上样本的卡方不良品率检验和卡方拟合优度检验。

两个以上样本的卡方不良品率

目标

我们想评估在每个样本中观测到的最小不良品和良品数的一般规则，以确保近似置信区间准确。

方法

我们首先定义在比较图中使用的区间。定义区间的两端，使得整体错误率约为 $\propto$，任何不重叠的区间显示总体不良品率是不同的。有关使用的公式，请参见附录 G。

比较区间基于成对比较置信区间。有关详细信息，请参见《单因素方差分析协助白皮书》中的“比较区间”一节。我们为每对 $(p_i - p_j)$ 使用正态近似置信区间，然后使用 Bonferroni 多重比较程序来控制整个实验的错误率。因此，我们只需要评估成对比较程序中某个区间的有效性，便可以了解正态分布的近似性对比较区间的影响。

结果

为了评估正态近似分布的有效性，我们只需要检查近似性如何影响不良品率之间差值的一个区间。因此，我们可以简单地使用专为双样本不良品率实例制定的一般规则。有关详细信息，请参见《双样本不良品率检验协助白皮书》中的“双样本不良品率检验方法”一节。双样本不良品率检验中的模拟结果表明，当样本量足够大时，对于不良品率之间的差值，近似置信区间的准确度一般比较可靠 – 也就是说，每个样本中观测到的不良品和良品数至少为 5。

当为两个以上样本的卡方不良品率检验检查其区间的有效性时，“协助报告卡”会显示以下状态指标：

状态	条件
	所有样本都具有至少 5 个不良品和 5 个良品。比较区间是准确的。
	如果上述条件不成立。

卡方拟合优度

目标

我们想评估在每个样本中观测到的最小不良品和良品数的一般规则，以确保近似置信区间准确。

方法

“协助”中的卡方拟合优度检验包括比较和单个置信区间。我们利用标准正态近似区间计算比率并使用 Bonferroni 修正法 (Goodman, 1965) 修正多重比较区间。因此，Bonferroni 同时区间计算如下：

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

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

定义区间的两端，使得整体错误率约为 $\propto$，不包含目标比率值的任何区间表示实际比率与对应的目标比率不同。各个区间的形式与 Bonferroni 区间相同，但不使用 $Z_{\alpha/2}$ 修正多重比较区间。

结果

上述两种方法都遵循类似于“协助”的双样本不良品率检验中定义的方法。因此，我们可以使用专为双样本不良品率检验研发的类似规则，来检查正态近似分布的有效性。有关详细信息，请参见《双样本不良品率检验协助白皮书》中的“双样本不良品率检验方法”一节。在本文中，我们得出的结论是，当样本计数小于 5 时，比较区间和单个置信区间可能不准确。

当为卡方拟合优度检验检查其区间的有效性时，“协助报告卡”会显示以下状态指标：

状态	条件
	所有样本计数均至少为 5。区间是准确的。

状态	条件
	有的样本计数少于 5。

参考书

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.

附录 A：卡方检验统计量

“协助”使用以下形式的卡方检验统计量：

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

其中

O_{ij} = 观测到的计数，参见下表中的定义：

实例	O_{ij}
检验多项分布的拟合优度	第 i 个结果观测到的计数被定义为 O_{i1}。
检验超过 2 项的不良品率的相等性	第 i 项样本观测到的不良品和良品项目数分别定义为 O_{i1} 和 O_{i2}。
检验两个类别变量之间的相关性	X 变量的第 i 个值和 Y 变量的第 j 个值的观测计数定义为 O_{ij}。

E_{ij} = 下表中定义的期望数：

实例	E_{ij}
检验多项分布的拟合优度	$E_{i1} = np_i$ $i = 1, \ldots, k$（k = 结果数） n = 样本量 p_i = 历史比率 $\sum_i p_i = 1$
检验超过 2 项的不良品率的相等性	$E_{i1} = n_i p$（对于不良品） $E_{i2} = n_i(1-p)$（对于良品） $i = 1, \ldots, k$（k = 样本个数） n_i = i 样本量 p = 总不良品率
检验两个类别变量之间的相关性	$E_{ij} = \frac{(n_i n_j)}{n_{..}}$ $i = 1, \ldots, m$（m = X 值个数） $j = 1, \ldots, k$（k = Y 值个数） $n_{i.}$ = X 变量的 i 值总计数 $n_{.j}$ = Y 变量的 i 值总计数 $n_{..}$ = 总样本量

附录 B：两个以上样本的卡方不良品率检验的功效

我们使用非中心卡方分布来计算 $p_i = p_j = p\ \forall i,j$ 的检验功效。非中心参数取决于 n_i 和 $p_i \forall i$

其中

n_i = 第 i 个样本的样本量

每个 p_i 代表一个根据比率差值 = δ 计算出的替代比率（参见本附录中的下一节“替代比率计算”）。

卡方分布的非中心参数计算如下：

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

其中

$O_{i1=n_i p_i}$

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

检验功效计算为

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

其中

X = 是以非中心参数为 χ^2 的非中心卡方分布中的随机变量。

$x_{1-\alpha}$ = 由中心卡方分布的逆 cdf 计算得出的 $1-\alpha$ 分位数。

替代比例计算

替代比例定义如下：

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

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

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

$$0 < \delta < 1$$

其中

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

$\hat{p}_i$ = 第 i 个样本项的样本不良品率。

NT = 总观测值个数。

n_i = i 样本的样本量。

对于某些差值 δ，$p_i > 1$ 或 $p_j < 0$。因此我们制定了以下规则：

如果 p_j < 0

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

如果 p_i > 1

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

使用 n_i 的两个最小值将产生最小的功效，使用 n_i 的两个最大值将产生最大的功效。

附录 C：比率扰动模型和等比率模型

比率扰动模型

根据 Read and Cressie (1988)，我们将原假设下的比率组定义如下：

我们选择 δ 近似 k - 1（其中 k = 每个样本中计算比率的个数）并将一组小 p_i 定义为

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

并将其余 p_i 定义为

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

模拟中 δ 使用的值在表 1 中列出。

表 1 δ 在模拟中与生成的小比率 p_i 一起使用

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

对于每个 k 值，我们让 r 变化，r = 1,…, k - 1，以改变一组中取小比率 p_i' s. 的组数的多少 例如，对于 k = 3，我们得到了表 2 中所述的以下两个模型。

表 2 对于 k = 3，使用比率扰动模型的 p_i 的值

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

等比率模型

为了获得一个所有期望单元格的计数都很小的模型，我们使用按以下方式定义的等比率模型

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

使用此模型，是在样本量非常小时才使用的，所有的小期望单元格的计数当然也很小。使用等比率模型，样本量需要非常小，这样才能使得小期望单元格计数也很少，这可能不会在实际应用中发生。

附录 D：卡方拟合优度检验的有效性

对于比率扰动模型，我们根据小期望单元格计数的百分比，绘制了能使其产生的 I 类错误率落入 [0.03，0.07] 区间中的检验之最小期望单元格计数图，如图 1 中所示。

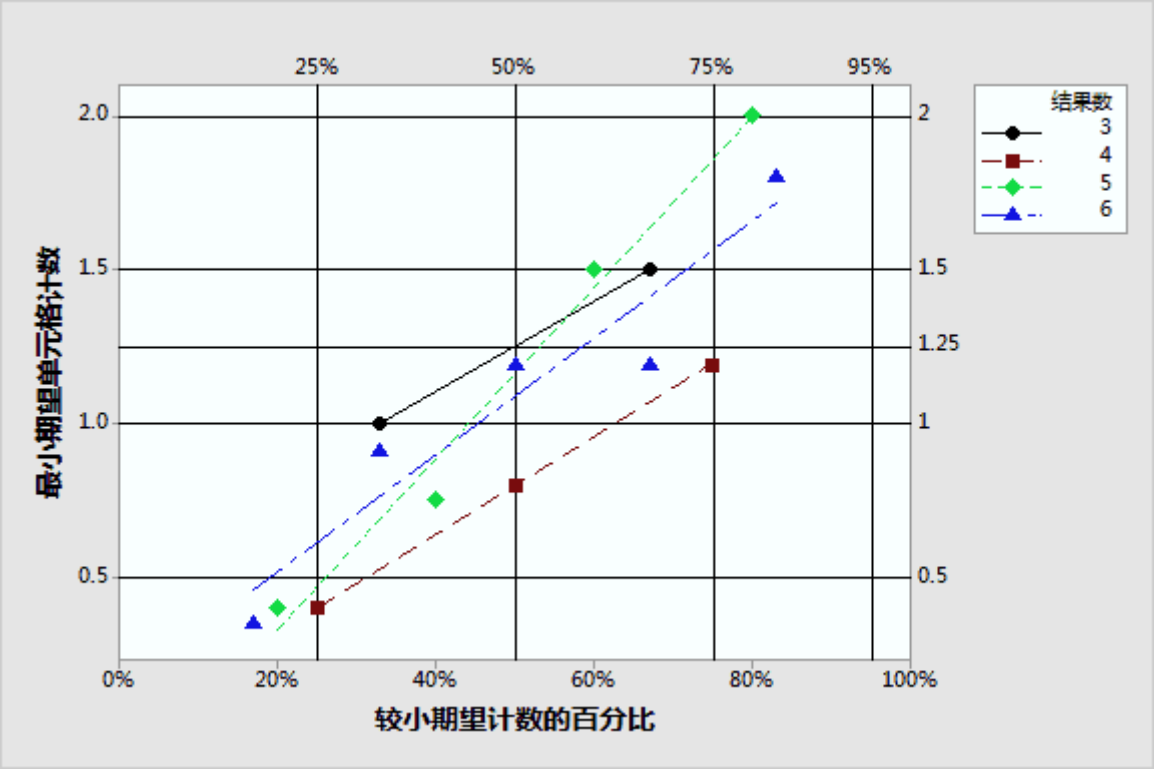

图 1　能使其产生的 I 类错误率落入 [0.03，0.07] 区间中的检验之最小期望单元格计数与小期望单元格计数的百分比。

在图 1 中，当小期望单元格计数的百分比小于 50% 时，最小期望单元格计数小于或等于 1.25。所有的最小期望单元格计数小于或等于 2。根据这些模拟结果，我们在“协助报告卡”中使用的是保守的规则。

接下来，我们进行了相同的模拟，使用等比率模型定义了原假设下的分布。表 4 总结了使用等比率模型的模拟结果。

表 4　I 类错误率落入 [0.03，0.07] 区间中的检验之最小期望单元格计数

k	最小期望单元格计数
3	2.5
4	1.25
5	1

k	最小期望单元格计数
6	1.4

如上所述，等比率模型将导致发生所有单元格计数都小的情况。表 4 显示，所有最小期望单元格计数都小于或等于 2.5，从而支持我们在“协助报告卡”中使用的规则。

附录 E：相关性卡方检验的有效性

对于比率扰动模型，我们根据每个 X 值个数的小期望单元格计数的比例，绘制了最小期望单元格计数图，以使 I 类错误率落入 [0.03，0.07] 区间中，如图 2 中所示。

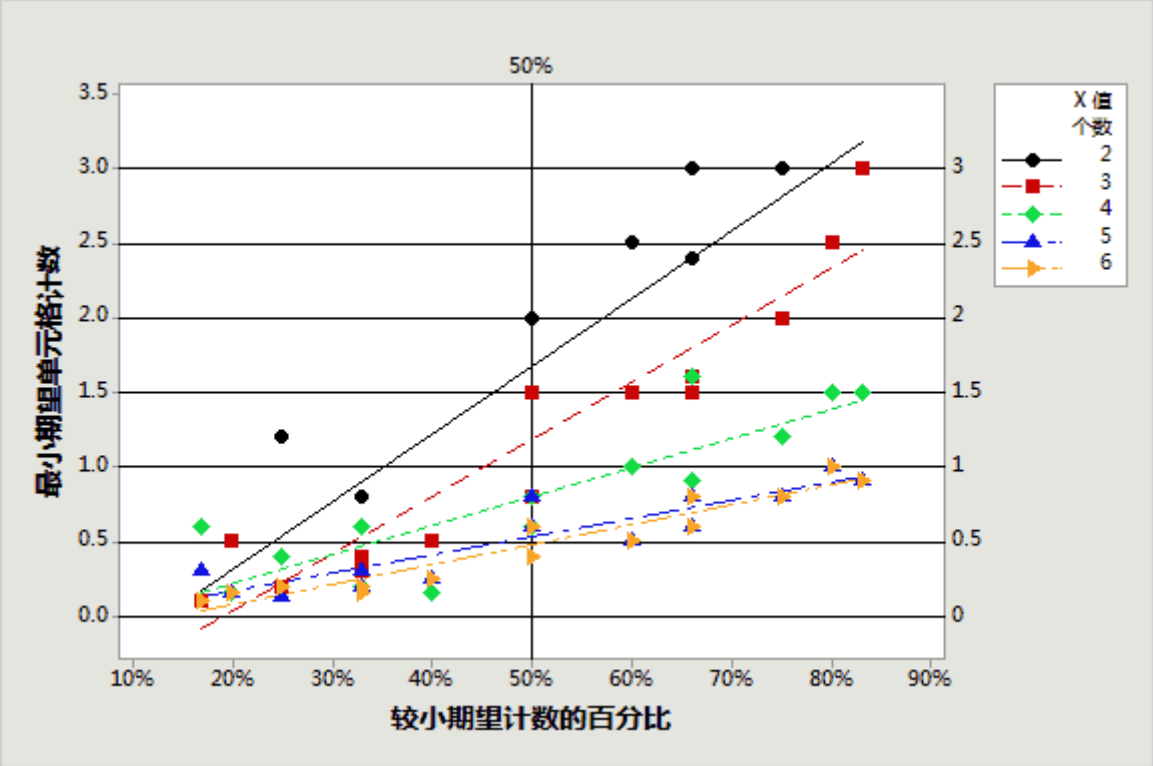

图 2　I 类错误率落入 [0.03，0.07] 区间中的检验所需的最小期望单元格计数与小期望单元格计数的百分比。

图 2 表明，最小期望单元格计数取决于 x 值个数和小期望单元格计数的百分比。

图 2 表明，当小期望单元格计数的比例 ≤ 50% 时，X 值个数分别等于 2 或 3 以及 4、5 或 6 的最小期望单元格计数 ≤ 2 和 ≤ 1。而当小期望单元格计数的比例 > 50% 时，X 值个数分别等于 2 或 3 以及 4、5 或 6 的最小期望单元格计数 ≤ 3 和 ≤ 1.5。

对于等比率模型，我们根据 X 值个数 (m) 和 Y 值个数 (k)，绘制了最小期望单元格计数图，如图 3 中所示。

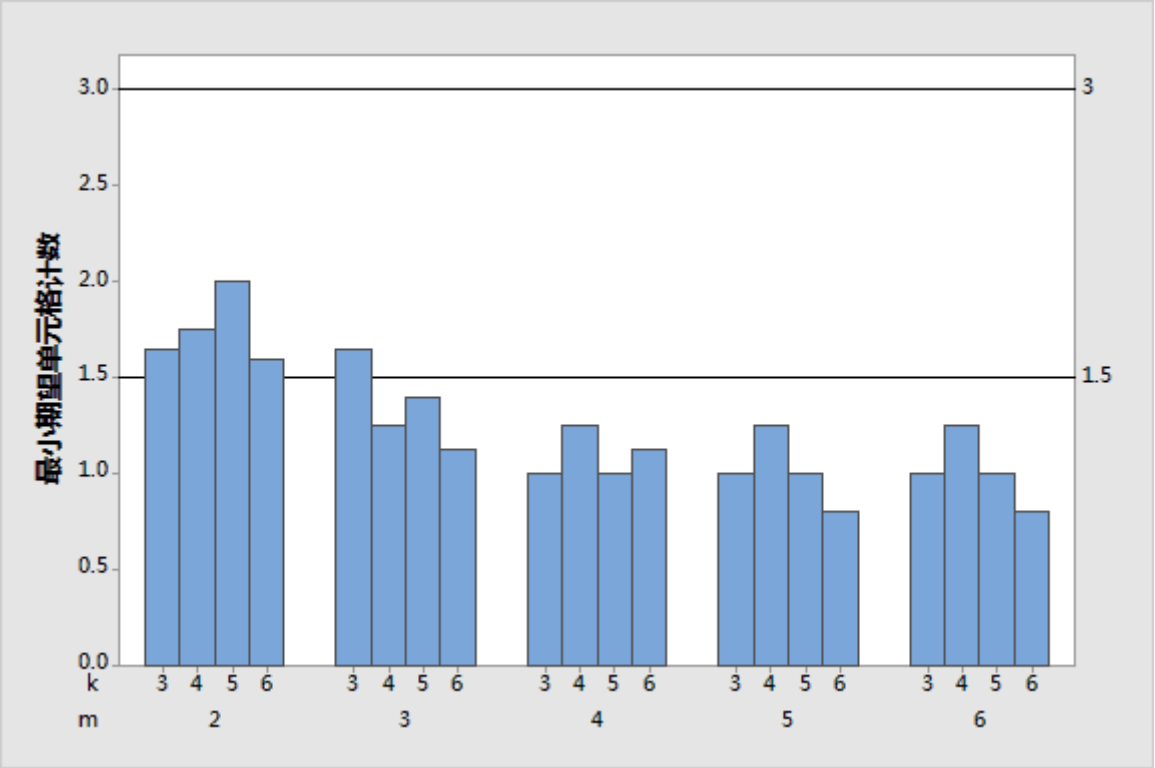

图 3　I 类错误率落入 [0.03，0.07] 区间中的检验所需之最小期望单元格计数与 X 值 (m) 和 Y 值 (k)

图 3 表明，当 X 值个数等于 2 或 3 时，最小期望单元格计数 ≤ 2，当 X 值个数等于 4、5 或 6 时，最小期望单元格计数 ≤ 1.5。根据这些模拟结果，“协助报告卡”中使用的是保守的规则。

附录 F：两个以上样本的卡方不良品率检验的有效性

对于任一 p 值和 3 至 12 之间的 m 值，我们绘制了最小期望单元格计数图。结果显示在图 4 和图 5 中。

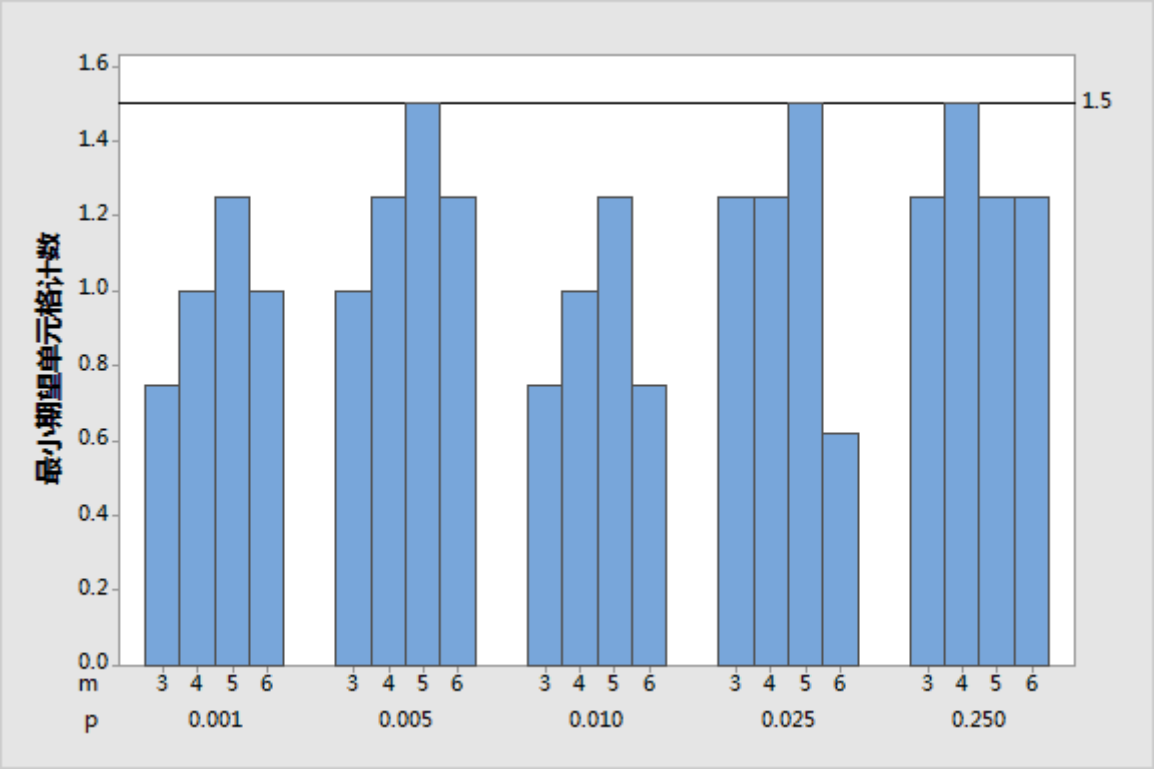

图 4　在 I 类错误率落入[0.03 - 0.07]范围内的那些检验所需的最小期望单元格计数与 X 值个数（m = 3 至 6）

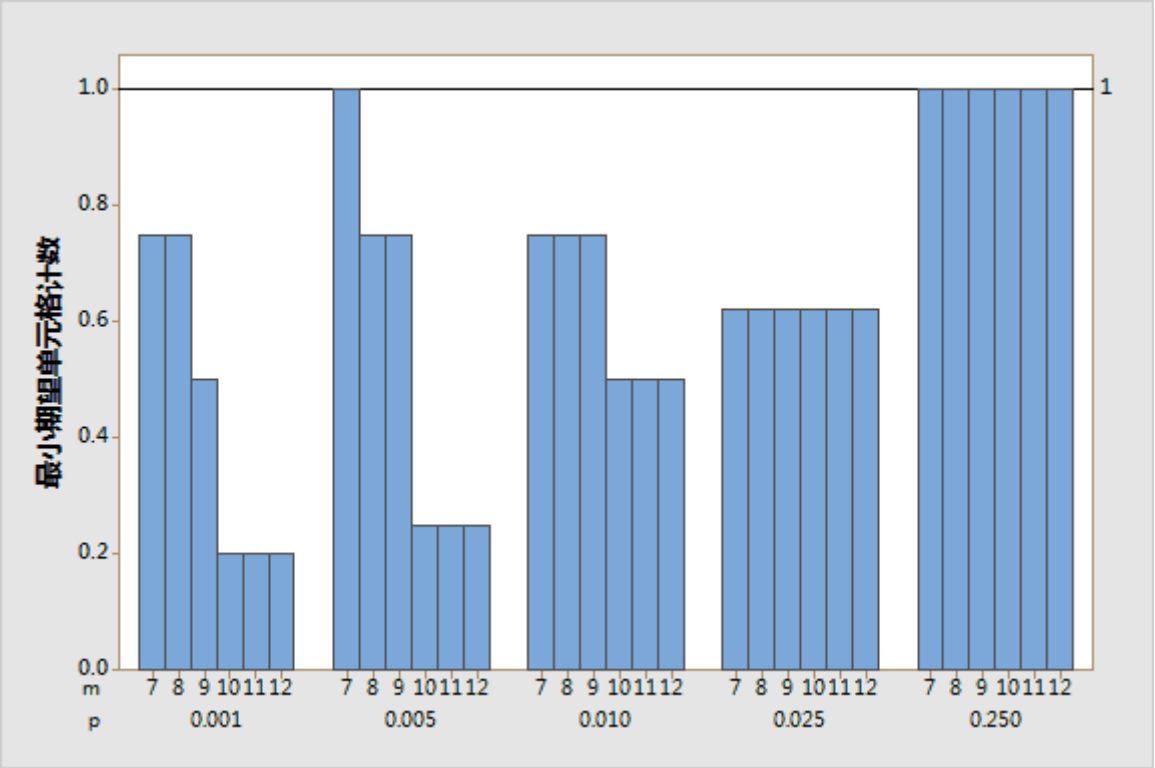

图 5　在 I 类错误率落入[0.03 - 0.07]范围内的那些检验所需的最小期望单元格计数与 X 值个数（m = 7至 12）

当 X 值个数等于 3、4、5 或 6 时，大于或等于 1.5 的期望单元格计数会使生成检验的 I 类错误率落入区间 [0.03，0.07] 中。当 X 值个数在 7 至 12 之间时，大于或等于 1 的期望单元格计数会使生成检验的 I 类错误率落入区间 [0.03，0.07] 中。

附录 G：两个以上样本的卡方不良品率的比较区间

p_i 的下限和上限定义如下：

$p_{i\text{下限}} = p_i - Z_{\alpha/c}X_i$

$p_{i\text{上限}} = p_i + Z_{\alpha/c}X_i$

其中

c = 比较数 = k (k - 1) /2

其中 k 是样本数。

$Z_{\alpha/c}$ = (1 - 均值为 0 标准差为 1 的正态分布的 $\frac{\alpha}{2c}$) 百分位数

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

其中

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