双样本不良品率检验

概述

双比率检验用于确定两个比率之间是否存在显著的差异。在质量分析中，当某个产品或服务有缺陷或无缺陷时，通常使用此检验来确定从两个独立过程中收集的样本的不良品率差异是否显著。

Minitab 协助包括双样本不良品率检验。为此检验收集的数据是两个独立样本各自的缺陷品数，它假设为二项随机变量的观测值。“协助”使用准确方法来计算假设检验结果，因此，实际的类型 I 误差率应接近为此检验指定的显著性水平（alpha 值），而不需要任何进一步的调查。但是，“协助”使用正态近似方法计算不良品率差值的置信区间 (CI)，并使用正态近似检验的理论功效函数执行其功效和样本数量分析。由于这些是近似方法，因此，我们需要评估其准确性。

在本白皮书中，我们将调查近似置信区间在哪些条件下准确。我们还对用于评估双样本不良品率检验的功效和样本数量的方法进行调查，以将近似方法的理论功效与准确检验的实际功效进行比较。最后，我们还检查了在“协助”的“报告卡”中自动执行和显示的以下数据检查，并说明它们对分析结果的影响：

- 置信区间的有效性
- 样本数量

双样本不良品率检验也依赖于其他假设。有关详细信息，请参见附录 A。

双样本不良品率检验方法

置信区间的准确度

虽然“协助”使用 Fisher 精确检验来评估两个样本的不良品率是否存在显著差异，但此差异的置信区间基于正态近似方法。根据大多数统计学教科书中提供的一般规则，如果每个样本中观测到的不良品数和优良品数至少为 5，则此近似置信区间准确。

目标

我们想要评估基于正态近似方法的置信区间在哪些条件下准确。尤其是，我们想要了解与每个样本中不良品数和优良品数有关的一般规则对近似置信区间的准确度的影响。

方法

此公式用于计算两个比率之间的差值的置信区间，并且用于确保其准确度的一般规则如附录 D 中所述。此外，我们介绍了在调查期间制定的较为宽松的修订规则。

我们执行了模拟，以评估各种条件下近似置信区间的准确度。为执行模拟，我们根据几个 Bernoulli 总体生成了各种数量的随机样本对。对于每种类型的 Bernoulli 总体，我们计算了 10,000 个 Bernoulli 样本仿行中每对仿行的两个比率之间的差值的近似置信区间。然后，我们计算了包含这两个比率之间真实差值的 10,000 个区间的比率，称作模拟覆盖概率。如果近似区间准确，则模拟覆盖概率应接近目标覆盖概率 0.95。为评估近似区间的准确度（与每个样本中所需的最小不良品和优良品数的初始和修订规则有关），我们还计算了符合每条规则的 10,000 对样本的百分比。有关详细信息，请参见附录 D。

结果

在样本数量足够大时，即在每个样本中观测到的不良品数和优良品数至少为 5 时，两个比率之间的差值的近似置信区间通常比较准确。因此，我们采用此规则在“报告卡”中进行“置信区间有效性”检查。虽然此规则通常效果不错，但在某些情况下，它过于保守，当两个比率接近 0 或 1 时，它会稍微宽松些。有关详细信息，请参见“数据检查”和附录 D。

理论功效函数的性能

“协助”可执行假设检验，以利用 Fisher 检验比较两个 Bernoulli 总体比率（两个样本中的不良品率）。但是，由于此精确检验的功效函数不容易推导出，因此，必须使用对应的正态近似检验的理论功效函数近似确定此功效函数。

目标

我们想要确定基于正态近似检验的理论功效函数是否适用于在“协助”中评估双样本不良品率检验的功效和样本数量要求。为此，我们需要评估此理论功效函数能否准确反映 Fisher 精确检验的实际功效。

方法

附录 B 中详细介绍了 Fisher 精确检验的方法（包括计算其 p 值）。附录 C 中对基于正态近似检验的理论功效函数进行了定义。根据这些定义，我们执行了相关模拟，以在使用

Fisher 精确检验分析两个样本中的不良品率差值时，估计该检验的实际功效水平（我们将其称作模拟功效水平）。

为执行模拟，我们根据几个 Bernoulli 总体生成了各种数量的随机样本对。对于各个 Bernoulli 总体类别，我们对 10,000 对样本仿行中的每对仿行执行了 Fisher 精确检验。对于每个样本数量，我们计算了检验的模拟功效，以检测作为检验效果显著的 10,000 对样本的一部分的给定差值。为进行比较，我们还根据正态近似检验计算了对应的理论功效。如果近似方法效果不错，则理论和模拟功效水平应该比较接近。有关详细信息，请参见附录 E。

结果

我们的模拟显示，通常，正态近似检验的理论功效函数和 Fisher 精确检验的模拟功效函数几乎相等。因此，“协助”可在执行 Fisher 精确检验时，使用正态近似检验的理论功效函数估计检测实际的重要差值所需的样本数量。

数据检查

置信区间的有效性

由于双样本不良品率检验使用精确检验来评估不良品率差值，因此，其准确度很大程度上受每个样本中不良品和优良品数量的影响。但是，不良品率之间的差值的置信区间基于正态近似方法。当每个样本中的不良品和优良品数量增大时，近似置信区间的准确度也会提高（请参见附录 D）。

目标

我们想要确定样本中的不良品和优良品数量是否足以确保不良品率差值的近似置信区间准确。

方法

我们使用了大多数统计学教科书中提供的一般规则。当每个样本至少包含 5 个不良品和 5 个优良品时，双样本不良品率检验的近似置信区间准确。有关详细信息，请参见上面的双样本不良品率方法部分。

结果

正如双样本不良品率方法部分中汇总的模拟所述，置信区间的准确度取决于每个样本中不良品和优良品的最小数。因此，“协助”会在“报告卡”中显示以下状态指示符，以帮助您评估两个不良品率之间的差值的置信区间的准确度。

状态	条件
	这两个样本都至少有 5 个不良品和 5 个优良品。此差值的置信区间应该准确。
	至少一个样本中的不良品数或优良品数小于 5。此差值的置信区间可能不准确。

样本数量

通常，为收集否定“无差异”原假设的证据，将进行统计假设检验。如果样本太小，检验的功效可能不准确，因此检测不到是否真正存在差值，这将导致类型 II 误差。因此，一定要确保样本数量足够大，以便有较高的概率检测到实际的重要差值。

目标

如果数据没有提供足够的证据否定原假设，则我们想要确定样本数量是否足够大，以便有较高的概率检测到实际的重要差值。虽然计划样本数量的目的是确保样本数量足够大，以便有较高的概率检测到重要差值，但它们不应该大到有较高的概率使无意义的差值变成具有显著的统计意义。

方法

双样本不良品率检验的功效和样本数量分析基于正态近似检验的理论功效函数，该检验可以准确地估计 Fisher 精确检验的实际功效（参见双样本不良品率方法部分的理论功效函数性能中汇总的模拟结果）。理论功效函数可表示为组合样本中不良品率和总不良品率之间的目标差值的函数。

结果

在数据不能提供足够的证据否定原假设时，“协助”将使用正态近似检验的功效函数来计算通过 80% 和 90% 的概率检测到的给定样本数量的实际差值。此外，如果用户提供了相关的特定实际差值，则“协助”将使用正态近似检验的功效函数来计算有 80% 和 90% 的机会检测到差值的样本数量。

为帮助解释结果，“协助”的双样本不良品率检验的“报告卡”会在检查功效和样本数量时显示以下状态指示符：

状态	条件
	此检验发现不良品率之间存在差值，因此，功效不是问题。 或 功效足够。此检验没有发现不良品率之间存在差值，但样本数量足够大，至少有 90% 的机会检测到给定差值（功效 ≥ .90）。
	功效可能足够。此检验没有发现不良品率之间存在差值，但样本数量足够大，有 80% 到 90% 的机会检测到给定差值（.80 ≤ 功效 < .90）。将报告获取 90% 的功效所需的样本数量。
	功效可能不足。此检验没有发现不良品率之间存在差值，但样本数量足够大，有 60% 到 80% 的机会检测到给定差值（.60 ≤ 功效 < .80）。将报告获取 80% 和 90% 的功效所需的样本数量。
	功效不足。此检验没有发现不良品率之间存在差值，并且样本数量不够大，无法提供 60% 的机会检测到给定差值（功效 < .60）。将报告获取 80% 和 90% 的功效所需的样本数量。
	此检验没有发现不良品率之间存在差值。您没有指定要检测的实际差值。根据您的数据，此报告可能会根据您的样本数量和 alpha 值指出有 80% 和 90% 的机会检测到的差值。

参考书

Arnold, S.F. (1990)。*Mathematical statistics*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Casella, G., & Berger, R.L. (1990)。*Statistical inference*. Pacific Grove, CA: Wadsworth, Inc.

附录 A：双样本不良品率检验的其他假设

双样本不良品率检验基于以下假设：

- 每个样本中的数据包括 n 个相异项，每个项归类为不良品或优良品。
- 一个样本中的每个项目是否不良品的概率相同。
- 项目是不良品的可能性不受另一个项目是否为不良品影响。

由于已经为此检验输入了汇总数据（而不是原始数据），因此，无法在“协助”的“报告卡”的数据检查中验证这些假设。

附录 B：Fisher 精确检验

假设我们从 Bernoulli 分布中观测到两个独立随机样本 $X_1,\ldots,X_{n_1}$ 和 $Y_1,\ldots,Y_{n_2}$，以便

$$p_1 = \Pr(X_i = 1) = 1 - \Pr(X_i = 0) \text{ 和 } p_2 = \Pr(Y_j = 1) = 1 - \Pr(Y_j = 0)$$

在以下部分中，我们介绍了用于推算比率之间的差值 $\delta = p_1 - p_2$ 的过程。

公式 B1：Fisher 精确检验和 p 值

可在 Arnold (1994) 中找到 Fisher 精确检验的说明。我们提供了此检验的简要说明。

在进行试验时，设 V 为第一个样本中的成功次数，设 $v = n_1\hat{p}_1$ 为第一个样本中观测到的成功次数。在进行试验时，还设 W 为两个样本中的成功总次数，设 $w = n_1\hat{p}_1 + n_2\hat{p}_2$ 为观测到的成功次数。请注意，$\hat{p}_1$ 和 $\hat{p}_2$ 是 p_1 和 p_2 的样本点估计值。

在原假设 $\delta = p_1 - p_2 = 0$ 条件下，给定 W 的 V 条件分布是概率批量函数的超级几何分布

$$f(v|w) = \frac{\binom{n_1}{v}\binom{n_2}{w}}{\binom{n_1+n_2}{w}}$$

设 $F(v|w)$ 为分布的 c.d.f。则单侧和双侧检验的 p 值为：

- 在检验 $\boldsymbol{\delta < 0}$ 或等效 $\boldsymbol{p_1 < p_2}$ 时

  p 值计算为 $F(v|w)$，其中，v 是 V 的观测值或在第一个样本中观测到的成功次数，w 是 W 的观测值或这两个样本中观测到的成功次数。

- 在检验 $\boldsymbol{\delta > 0}$ 或等效 $\boldsymbol{p_1 > p_2}$ 时

  p 值计算为 $1 - F(v-1|w)$，其中，v 是 V 的观测值或在第一个样本中观测到的成功次数，w 是 W 的观测值或这两个样本中观测到的成功次数。

- 在检验 $\boldsymbol{\delta \neq 0}$ 或等效 $\boldsymbol{p_1 \neq p_2}$ 时

  将根据以下算法计算 p 值，其中，m 是上述超级几何分布的模式。

  - 如果 $v < m$，则 p 值计算为 $1 - F(y-1|w) + F(v|w)$，其中，v 和 w 如上述定义，并且 $y = \min\{k \geq m: f(k|w) \leq f(v|W)\}$
  - 如果 $v = m$，则 p 值为 1.0
  - 如果 $v > m$，则 p 值计算为 $1 - F(v-1|w) + F(y|w)$，其中，v 和 w 如上述定义，并且 $y = \max\{k \leq m: f(k|w) \leq f(v|W)\}$

附录 C：理论功效函数

为比较两个比率（或更具体地说，两个不良品率），我们将按附录 B 中所述使用 Fisher 精确检验。由于此检验的理论功效函数太复杂而无法将其推导出来，因此，我们使用近似功效函数。更具体地说，我们对两个比率使用已知正态近似检验的功效函数，以近似计算 Fisher 精简检验的功效。

双侧检验的正态近似功效函数为

$$\pi(n_1,n_2,\delta)=1-\Phi\left(\frac{-\delta+z_{\frac{\alpha}{2}}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right)+\Phi\left(\frac{-\delta-z_{\alpha/2}\sqrt{p_c(1-p_c)(1/n_1+1/n_2)}}{se}\right)$$

其中，$\delta=p_1-p_2$，

$$se=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

并且 $p_c=(n_1p_1+n_2p_2)/(n_1+n_2)$。

在针对 $p_1>p_2$ 检验 $p_1=p_2$ 时，功效函数为

$$\pi(n_1,n_2,\delta)=1-\Phi\left(\frac{-\delta+z_{\alpha}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right)$$

在针对 $p_1<p_2$ 检验 $p_1=p_2$ 时，功效函数为

$$\pi(n_1,n_2,\delta)=\Phi\left(\frac{-\delta-z_{\alpha}\sqrt{p_c(1-p_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}{se}\right)$$

附录 D：近似置信区间

公式 D1：为两个比率之间的差值计算近似置信区间

基于正态近似方法的 $\delta = p_1 - p_2$ 的渐进 $100(1-\alpha)\%$ 置信区间为：

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

用于评估此近似置信区间的已知一般规则为 $n_1\hat{p}_1 \geq 5$、$n_1(1-\hat{p}_1) \geq 5$、$n_2\hat{p}_2 \geq 5$ 和 $n_2(1-\hat{p}_2) \geq 5$。换言之，如果每个样本中观测到的成功次数和失败次数至少为 5 时，则置信区间准确。

注：在此部分和随后的部分中，我们根据每个样本中的成功次数和失败次数，按最一般的形式表示置信区间的规则。成功是有利事件，失败是有利事件的补充。因此，在双样本不良品率检验的具体环境中，“成功”次数等效于不良品数，“失败”次数等效于优良品数。

公式 D2：近似置信区间的规则

基于正态近似方法且用于置信区间的一般规则为：如果 $n_1\hat{p}_1 \geq 5$、$n_1(1-\hat{p}_1) \geq 5$、$n_2\hat{p}_2 \geq 5$ 和 $n_2(1-\hat{p}_2) \geq 5$，则置信区间准确。即，如果每个样本包含至少 5 次成功（不良品）和 5 次失败（优良品），则区间的实际置信水平等于或近似等于目标置信水平。

由于实际的真实比率未知，因此，可采用成功和失败次数的估计比率（而不是真实比率）表示此规则。但是，在假定真实比率或真实比率已知的理论设置中，此规则可直接用真实比率表示。在这些情况下，可以直接评估预期的成功次数、预期的失败次数、n_1p_1、n_2p_2、$n_1(1-p_1)$ 和 $n_2(1-p_2)$ 对比率之间的差值的置信区间的实际覆盖概率的影响情况。

我们可以通过从两个 Bernoulli 总体（成功概率为 p_1 和 p_2）中抽取较大数量的样本对 n_1 和 n_2 来评估实际覆盖概率。然后，将实际覆盖概率计算为产生置信区间（包含两个比率之间的真实差值）的样本对的相对频率。如果在 $n_1p_1 \geq 5$、$n_2p_2 \geq 5$、$n_1(1-p_1) \geq 5$ 和 $n_2(1-p_2) \geq 5$ 时，实际覆盖概率准确，则按照大数量强定律，在 $n_1\hat{p}_1 \geq 5$、$n_1(1-\hat{p}_1) \geq 5$、$n_2\hat{p}_2 \geq 5$ 和 $n_2(1-\hat{p}_2) \geq 5$ 时，覆盖概率准确。这样，在实际和目标置信水平接近时，有望根据两个 Bernoulli 总体生成较大比率的样本对，以便在此规则有效时，$n_1\hat{p}_1 \geq 5$、$n_1(1-\hat{p}_1) \geq 5$、$n_2\hat{p}_2 \geq 5$ 和 $n_2(1-\hat{p}_2) \geq 5$。在后续模拟中，我们将此规则称作规则 1。

此外，在调查期间，我们发现在许多情况下，如果 $n_1p_1 \geq 5$，并且 $n_2p_2 \geq 5$，或者，如果 $n_1(1-p_1) \geq 5$，并且 $n_2(1-p_2) \geq 5$，则区间的模拟覆盖概率接近目标覆盖概率。这提示我们选择一个更宽松的备选规则，该规则指出，如果 $n_1\hat{p}_1 \geq 5$，并且 $n_2\hat{p}_2 \geq 5$，或者如果 $n_1(1-\hat{p}_1) \geq 5$，并且 $n_2(1-\hat{p}_2) \geq 5$，则近似置信区间准确。在后续模拟中，我们将此修订规则称作规则 2。

模拟 D1：评估近似置信区间的准确度

我们执行了一些模拟，以评估两个比率之间的差值的近似置信区间在哪些条件下准确。我们特别检查了与以下一般规则相关的区间的准确度：

规则 1（初始） $n_1p_1 \geq 5$、$n_2p_2 \geq 5$、$n_1(1-p_1) \geq 5$ 和 $n_2(1-p_2) \geq 5$

规则 2（修订） $n_1\hat{p}_1 \geq 5$ 和 $n_2\hat{p}_2 \geq 5$ 或 $n_1(1-\hat{p}_1) \geq 5$ 和 $n_2(1-\hat{p}_2) \geq 5$

在每个试验中，我们根据按以下比率定义的 Bernoulli 总体对生成了 10,000 对样本：

- A 比率：$\boldsymbol{p_1}$ 和 $\boldsymbol{p_2}$ 都接近 1.0（或接近 0）。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.8$ 和 $p_2 = 0.9$。
- B 比率：$\boldsymbol{p_1}$ 和 $\boldsymbol{p_2}$ 都接近 0.5。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.4$ 和 $p_2 = 0.55$。
- C 比率：$\boldsymbol{p_1}$ 接近 0.5，$\boldsymbol{p_2}$ 接近 1.0。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.4$ 和 $p_2 = 0.9$。

上述比率分类基于推导出近似置信区间所依据的二项分布的 DeMoivre-Laplace 正态近似方法。据说，在 Bernoulli 样本大于 10 并且成功概率接近 0.5 时，这种正态近似方法准确。在成功概率接近 0 或 1 时，通常需要更大的 Bernoulli 样本。

我们将两对总体的样本数量固定为单个值 n（其中 $n = 10, 15, 20, 30, \ldots, 100$）。我们限制了对平衡设计 $(n_1 = n_2 = n)$ 的研究，但没有丢失任何普遍性，原因是，这两个规则都依赖于观测到的成功和失败次数，这些次数可由样本数量和成功比率控制。

为估计这两个总体比率差值的置信区间的实际置信水平（也称作模拟置信水平），我们计算了包含这两个比率的真实差值的 10,000 个区间的比率。每个试验中的目标覆盖概率为 0.95。此外，我们确定了满足这两个规则的条件的 10,000 个样本的百分比。

注：对于一些较小的样本，比率差值的估计标准差为 0。我们考虑了“简并”样本并从试验中丢弃这些样本。结果是，在某些情况下，样本仿行数略小于 10,000。

结果如表 1-11 所示，并以图形方式显示在下图 1 中。

表 1　n=10 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 10$							
类别		**比率 (p)**	$\boldsymbol{np}$	$\boldsymbol{n(1-p)}$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	$\boldsymbol{p_1}$	0.80	8.00	2.00	0.907	0.0	99.1
	$\boldsymbol{p_2}$	0.90	9.00	1.00			
B	$\boldsymbol{p_1}$	0.40	4.00	6.00	0.928	4.4	63.0
	$\boldsymbol{p_2}$	0.55	5.50	4.50			
C	$\boldsymbol{p_1}$	0.45	4.50	5.50	0.919	0.0	48.3
	$\boldsymbol{p_2}$	0.90	9.00	1.00			

表 2　n=15 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 15$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	12.00	3.00	0.938	0.2	100.0
	p_2	0.90	13.50	1.50			
B	p_1	0.40	6.00	9.00	0.914	65.0	97.3
	p_2	0.55	8.25	6.75			
C	p_1	0.45	6.75	8.25	0.93	1.2	86.9
	p_2	0.90	13.50	1.50			

表 3　n=20 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 20$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	16.00	4.00	0.942	1.5	100.0
	p_2	0.90	18.00	2.00			
B	p_1	0.40	8.00	12.00	0.943	92.8	99.9
	p_2	0.55	11.00	9.00			
C	p_1	0.45	9.00	11.00	0.934	4.1	98.2
	p_2	0.90	18.00	2.00			

表 4　n=30 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 30$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	24.00	6.00	0.941	4.3	100.0
	p_2	0.90	27.00	3.00			
B	p_1	0.40	12.00	18.00	0.944	99.7	100.0
	p_2	0.55	16.50	13.50			

$n = 30$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
C	p_1	0.45	13.50	16.50	0.938	7.2	100.0
	p_2	0.90	27.00	3.00			

表 5　n=40 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 40$							
类别		比率 (p)	np	$n(1-p)$	覆盖概率	符合规则 1 的样本百分比	符合规则 2 的样本百分比
A	p_1	0.80	32.00	8.00	0.941	35.1	100.0
	p_2	0.90	36.00	4.00			
B	p_1	0.40	16.00	24.00	0.945	100.0	100.0
	p_2	0.55	22.00	18.00			
C	p_1	0.45	18.00	22.00	0.945	37.7	100.0
	p_2	0.90	36.00	4.00			

表 6　n=50 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 50$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	40.00	10.00	0.942	36.4	100.0
	p_2	0.90	45.00	5.00			
B	p_1	0.40	20.00	30.00	0.944	100.0	100.0
	p_2	0.55	27.50	22.50			
C	p_1	0.45	22.50	27.50	0.935	38.3	100.0
	p_2	0.90	45.00	5.00			

表 7　n=60 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 60$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	48.00	12.00	0.947	72.8	100.0
	p_2	0.90	54.00	6.00			
B	p_1	0.40	24.00	36.00	0.947	100.0	100.0
	p_2	0.55	33.00	27.00			
C	p_1	0.45	27.00	33.00	0.949	73.1	100.0
	p_2	0.90	54.00	6.00			

表 8　n=70 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 70$							
类别		**比率 (p)**	**np**	**n(1 − p)**	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	56.00	14.00	0.939	71.7	100.0
	p_2	0.90	63.00	7.00			
B	p_1	0.40	28.00	42.00	0.945	100.0	100.0
	p_2	0.55	38.50	31.50			
C	p_1	0.45	31.50	38.50	0.944	71.8	100.0
	p_2	0.90	63.00	7.00			

表 9　n=80 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 80$							
类别		**比率 (p)**	np	$n(1-p)$	**覆盖概率**	**符合规则 1 的样本百分比**	**符合规则 2 的样本百分比**
A	p_1	0.80	64.00	16.00	0.947	91.3	100.0
	p_2	0.90	72.00	8.00			
B	p_1	0.40	32.00	48.00	0.947	100.0	100.0
	p_2	0.55	44.00	36.00			

$n = 80$							
类别		比率 (p)	np	$n(1-p)$	覆盖概率	符合规则 1 的样本百分比	符合规则 2 的样本百分比
C	p_1	0.45	36.00	44.00	0.948	91.3	100.0
	p_2	0.90	72.00	8.00			

表 10　n=90 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 90$							
类别		比率 (p)	np	$n(1-p)$	覆盖概率	符合规则 1 的样本百分比	符合规则 2 的样本百分比
A	p_1	0.80	72.00	18.00	0.947	95.18	100.0
	p_2	0.90	81.00	9.00			
B	p_1	0.40	36.00	54.00	0.951	100.0	100.0
	p_2	0.55	49.50	40.50			
C	p_1	0.45	40.50	49.50	0.945	95.2	100.0
	p_2	0.90	81.00	9.00			

表 11　n=100 时，符合规则 1 和规则 2 的样本的模拟覆盖概率和百分比。目标覆盖概率为 0.95。

$n = 100$							
类别		比率 (p)	np	$n(1-p)$	覆盖概率	符合规则 1 的样本百分比	符合规则 2 的样本百分比
A	p_1	0.80	80.00	20.00	0.952	97.7	100.0
	p_2	0.90	90.00	10.00			
B	p_1	0.40	40.00	60.00	0.945	100.0	100.0
	p_2	0.55	55.00	45.00			
C	p_1	0.45	45.00	55.00	0.948	97.7	100.0
	p_2	0.90	90.00	10.00			

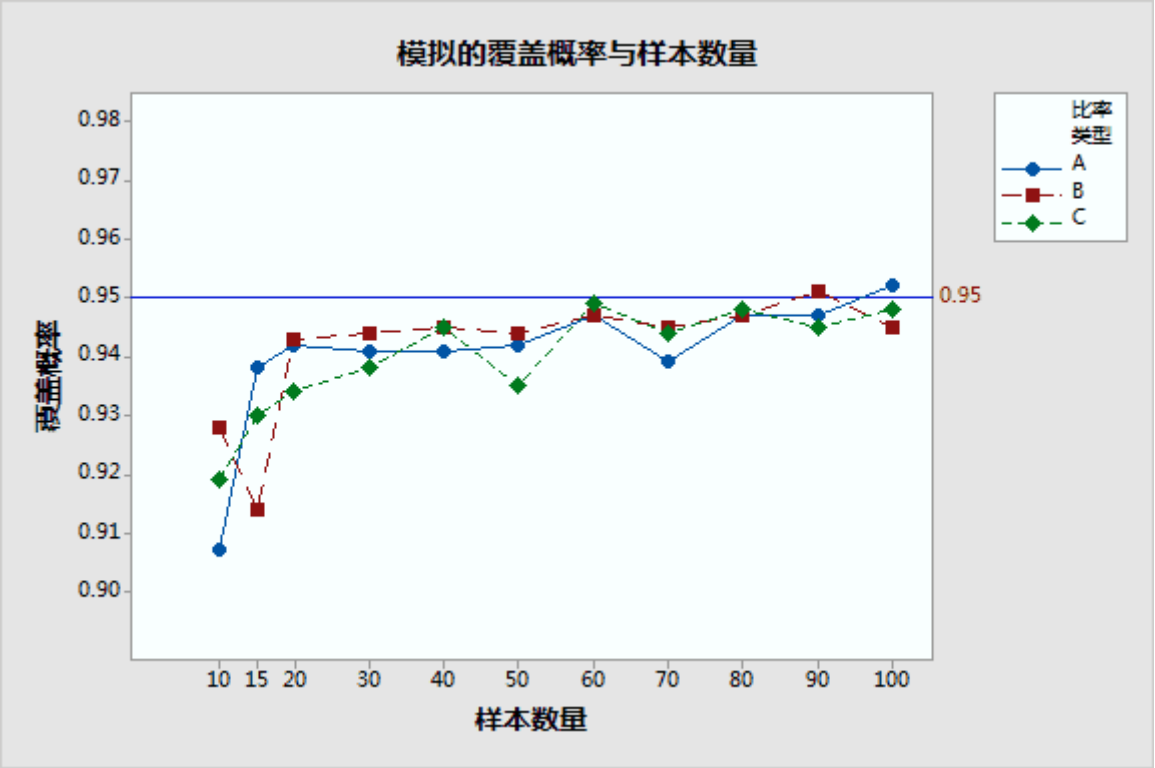

图 1　针对各个 Bernoulli 总体类别的样本数量绘制的模拟覆盖概率。

表 1-11 和图 1 中的结果显示，通常，根据类别 B（这两个比率都接近 0.5 时）中的 Bernoulli 总体生成的样本产生的模拟覆盖概率更稳定，并且更接近目标覆盖概率 0.95。在此类别中，这两个总体中的预期成功和失败次数大于其他类别中的成功和失败次数，即使是小样本也如此。

另一方面，对于根据类别 A（两个总体都接近 1.0 时）或类别 C（一个比率接近 1.0，另一个比率接近 0）中的 Bernoulli 总体对生成的样本而言，样本数量越小，模拟覆盖概率越远离目标值，但预期成功次数（np）或预期失败次数（$n(1-p)$）足够大的情况例外。

例如，在 $n = 15$ 时，考虑根据类别 A 中的 Bernoulli 总体生成的样本。对于每个总体，预期成功次数分别为 12.0 和 13.5，预期失败次数分别为 3.0 和 1.5。即使两个总体的预期失败次数小于 5，模拟覆盖概率也大约为 0.94。这类结果导致我们创建了规则 2，该规则只要求每个样本的预期成功次数*或* 预期失败次数大于或等于 5。

为了更全面地评估规则 1 和规则 2 如何有效评估置信区间的近似度，我们根据试验中的模拟覆盖概率绘制了符合规则 1 的样本百分比和符合规则 2 的样本百分比。图 2 中显示相关绘图。

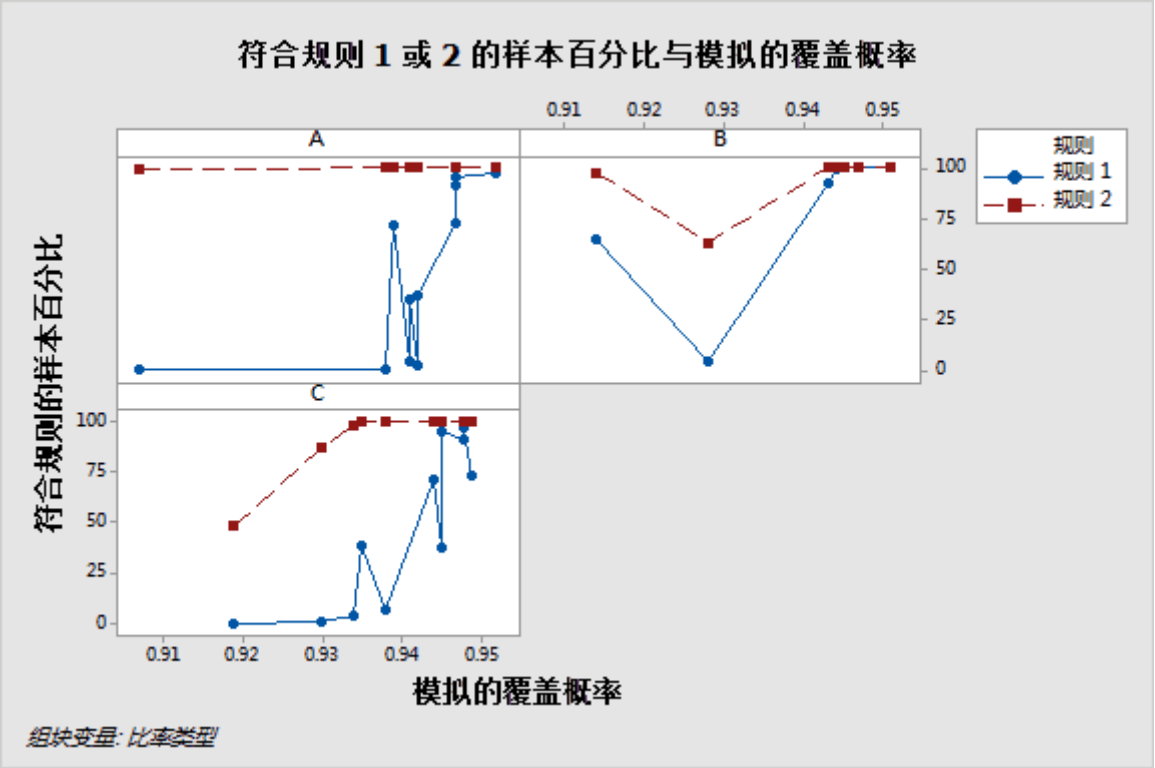

图 2　针对各个 Bernoulli 总体类别，根据模拟覆盖概率绘制的符合规则 1 和规则 2 的样本百分比。

这些图显示，随着模拟覆盖概率接近目标覆盖概率 0.95，满足每个规则要求的样本百分比通常接近 100%。对于根据类别 A 和 C 中的 Bernoulli 总体生成的样本，在样本较小时，规则 1 比较严格，这可通过符合此规则的极低样本百分比来证明，即使模拟覆盖概率接近目标覆盖概率时也是如此。例如，当 $n = 20$ 时，同时样本根据类别 A 中的 Bernoulli 总体生成，则模拟覆盖概率为 0.942（请参见表 3）。但是，符合此规则的样本比率接近 0 (0.015)（请参见图 2）。因此，在这些情况下，此规则可能太保守。

另一方面，对于根据类别 A 中的 Bernoulli 总体生成的小样本，规则 2 不太严格。例如，如表 1 中所示，当 $n = 10$ 时，同时样本根据类别 A 中的 Bernoulli 总体生成，则模拟覆盖概率为 0.907，99.1% 的样本符合此规则。

结论是，在样本数量较小时，规则 1 倾向于过度保守。在样本数量较小时，规则 2 不太保守，可以选作首选规则。但是，规则 1 众所周知且为大家所接受。虽然规则 2 有望成为潜在选择，但在某些情况下，它可能太宽松，如前所述。还可以结合使用这两种规则来利用每个规则的优势；但这种方法需要进一步的调查，才能付诸实践。

附录 E：比较实际功效与理论功效

模拟 E1：使用 Fisher 精确检验评估实际功效

我们设计了一个模拟，以将 Fisher 精确检验的估计实际功效水平（称作模拟功效水平）与基于正态近似检验的功效函数的理论功效水平（称作近似功效水平）进行比较。在每个试验中，我们根据每对 Bernoulli 总体生成了 10,000 对样本。还为每对样本选择了比率，以使这些比率之间的差值为 $p_1 - p_2 = -0.20$。

- A 比率：$\boldsymbol{p_1}$ 和 $\boldsymbol{p_2}$ 都接近 1.0（或接近 0）。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.70$ 和 $p_2 = 0.90$。
- B 比率：$\boldsymbol{p_1}$ 和 $\boldsymbol{p_2}$ 都接近 0.5。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.40$ 和 $p_2 = 0.60$。
- C 比率：$\boldsymbol{p_1}$ 接近 0.5，$\boldsymbol{p_2}$ 接近 1.0。为表示模拟中的这对 Bernoulli 总体，我们使用了 $p_1 = 0.55$ 和 $p_2 = 0.75$。

我们将两对总体的样本数量固定为单个值 n（其中 $n = 10, 15, 20, 30, \ldots, 100$）。我们限制了对平衡设计的研究（$n_1 = n_2 = n$），因为通常会假设两个样本具有相同的数量。我们计算了通过某种功效检测实际的重要差值所需的常用样本数量。

为基于每个模拟结果估计 Fisher 精确检验的实际功效，我们计算了 10,000 个样本对中的一部分，这些样本对的双侧检验在目标显著性水平达到 $\alpha = 0.05$ 时效果显著。为进行比较，我们根据正态近似检验计算了对应的理论功效水平。结果如下表 12 中所示。

表 12　与三种类别的 Bernoulli 总体的近似功效水平比较的 Fisher 精确检验的模拟功效水平。目标显著性水平为 $\alpha = 0.05$。

n	A 比率		B 比率		C 比率	
	$p_1 = 0.70$ $p_2 = 0.90$		$p_1 = 0.40$ $p_2 = 0.60$		$p_1 = 0.55$ $p_2 = 0.75$	
	模拟功效	近似功效	模拟功效	近似功效	模拟功效	近似功效
10	0.063	0.193	0.056	0.14	0.056	0.149
15	0.151	0.271	0.097	0.19	0.101	0.204
20	0.244	0.348	0.146	0.24	0.183	0.259
30	0.37	0.49	0.256	0.338	0.272	0.366
40	0.534	0.612	0.371	0.431	0.381	0.466
50	0.641	0.711	0.477	0.516	0.491	0.556
60	0.726	0.789	0.536	0.593	0.56	0.635

n	A 比率		B 比率		C 比率	
	$p_1 = 0.70$ $p_2 = 0.90$		$p_1 = 0.40$ $p_2 = 0.60$		$p_1 = 0.55$ $p_2 = 0.75$	
	模拟功效	近似功效	模拟功效	近似功效	模拟功效	近似功效
70	0.814	0.849	0.61	0.661	0.649	0.703
80	0.87	0.893	0.66	0.72	0.716	0.76
90	0.907	0.925	0.716	0.77	0.772	0.808
100	0.939	0.948	0.792	0.812	0.812	0.848

表 12 中的结果显示，对于所有三种类别的 Bernoulli 总体（A、B 和 C），近似功效往往大于模拟功效。例如，对于类别 A 中的比率，在近似功效水平为 0.91 时检测到绝对差值-0.20 所需的实际样本数量大约为 90。相比之下，根据近似理论功效函数计算出的对应样本数量估计值大约为 85。因此，根据近似功效函数计算出的样本数量估计值通常略小于获取给定功效水平所需的实际样本数量。

当这些结果显示为功效曲线（如下图 3 中所示）时，您可以更清楚地看到此关系。

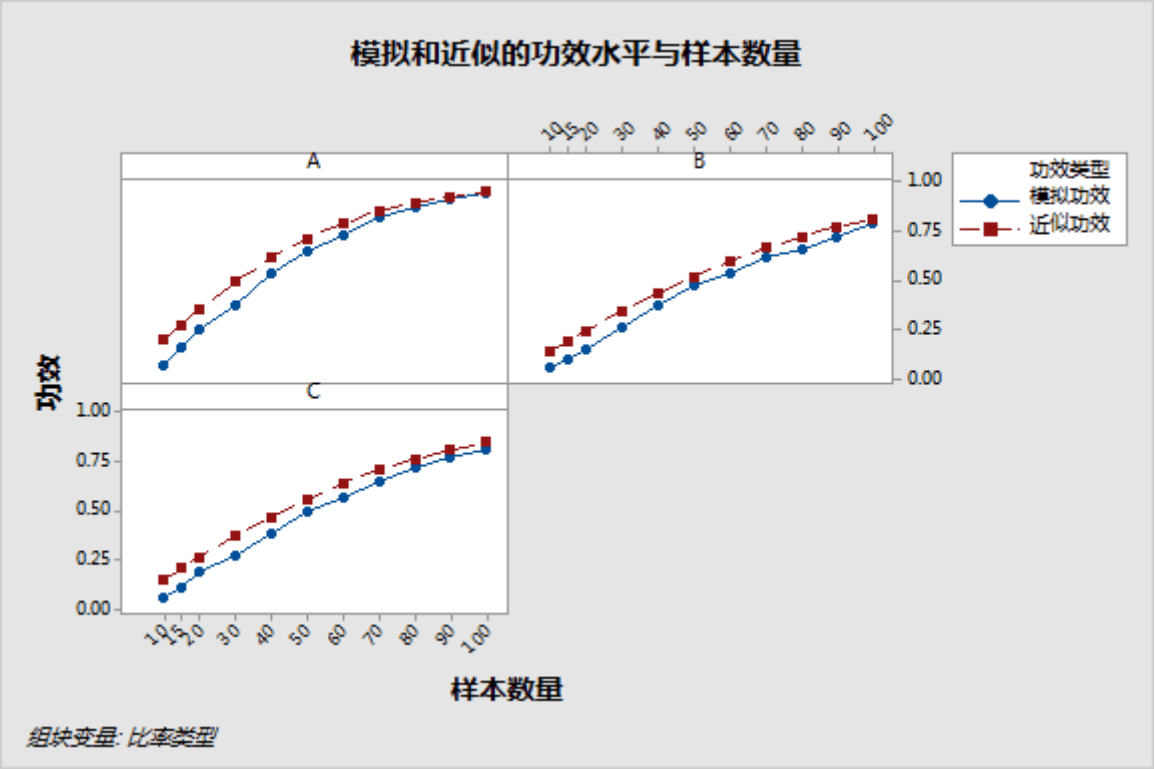

图 3　用于比较两个比率的双侧检验的模拟和近似功效水平图。功效水平根据各个 Bernoulli 总体类别的单个组块中的样本数量绘制而成。

请注意，虽然对于所有三种类别的 Bernoulli 总体（A、B 和 C）而言，模拟功效曲线低于近似功效曲线，但这些曲线之间的差值大小取决于抽取样本所依据的 Bernoulli 总体的真实比

率。例如，当这两个比率接近 0.5（类别 B）时，这两个功效水平通常比较接近。但是，对于与总体类别 A 和 C 关联的总体而言，在样本数量较小时，这两条功效曲线之间的差异更明显。

这些结果表明，通常，正态近似检验的理论功效函数和 Fisher 精确检验的模拟功效函数几乎相等。因此，“协助”可在执行 Fisher 精确检验之前，使用正态近似检验的理论功效函数来估计样本数量。但是，使用近似功效函数计算的样本数量可能略小于获取给定功效所需的实际样本数量，才能检测到这两个比率之间的差值（不良品率）。