

2-Sample % Defective Test

Overview

A test for 2 proportions is used to determine whether two proportions significantly differ. In quality analysis, the test is often used when a product or service is characterized as defective or not defective, to determine whether the percentage of defective items significantly differs for samples collected from two independent processes.

The Minitab Assistant includes a 2-Sample % Defective Test. The data collected for the test are the number of defective items in each of two independent samples, which is assumed to be the observed value of a binomial random variable. The Assistant uses exact methods to calculate the hypothesis test results; therefore, the actual Type I error rate should be near the level of significance (alpha) specified for the test and no further investigation is required. However, the Assistant uses a normal approximation method to calculate the confidence interval (CI) for the difference in % defectives and a theoretical power function of the normal approximation test to perform its power and sample size analysis. Because these are approximation methods, we need to evaluate them for accuracy.

In this paper, we investigate the conditions under which the approximate confidence intervals are accurate. We also investigate the method used to evaluate power and sample size for the 2-Sample % Defective Test, comparing the theoretical power of the approximate method with the actual power of the exact test. Finally, we examine the following data checks that are automatically performed and displayed in the Assistant Report Card and explain how they affect the results of the analysis:

- Validity of CI
- Sample size

The 2-Sample % Defective Test also depends on other assumptions. See Appendix A for details.

2-sample % defective test methods

Accuracy of confidence intervals

Although the Assistant uses Fisher's exact test to evaluate whether the % defectives of the two samples differ significantly, the confidence interval for the difference is based upon the normal approximation method. According to the general rule found in most statistical textbooks, this approximate confidence interval is accurate if the observed number of defectives and the observed number of nondefectives in each sample is at least 5.

Objective

We wanted to evaluate the conditions under which confidence intervals based on the normal approximation are accurate. Specifically, we wanted to see how the general rule related to the number of defectives and nondefectives in each sample affects the accuracy of the approximate confidence intervals.

Method

The formula used to calculate the confidence interval for the difference between the two proportions and the general rule for ensuring its accuracy are described in Appendix D. In addition, we describe a less stringent, modified rule that we developed during the course of our investigation.

We performed simulations to evaluate the accuracy of the approximate confidence interval under various conditions. To perform the simulations, we generated random pairs of samples of various sizes from several Bernoulli populations. For each type of Bernoulli population, we calculated an approximate confidence interval for the difference between the two proportions on each pair of 10,000 Bernoulli sample replicates. Then we calculated the proportion of the 10,000 intervals that contain the true difference between the two proportions, referred to as the simulated coverage probability. If the approximate interval is accurate, the simulated coverage probability should be close to the target coverage probability of 0.95. To evaluate the accuracy of the approximate interval in relation to the original and modified rules for the minimum number of defectives and nondefectives required in each sample, we also calculated the percentage of the 10,000 pairs of samples for which each rule was satisfied. For more details, see Appendix D.

Results

The approximate confidence interval for the difference between two proportions is generally accurate 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. Therefore, we adopted this rule for our Validity of CI check in the Report Card. Although this rule generally performs

well, in some cases it can be overly conservative, and it may be somewhat relaxed when the two proportions are close to 0 or 1. For more details, see the Data Check section and Appendix D.

Performance of theoretical power function

The Assistant performs the hypothesis test to compare two Bernoulli population proportions (% defectives in two samples) using Fisher's test. However, because the power function of this exact test is not easily derived, the power function must be approximated using the theoretical power function of the corresponding normal approximation test.

Objective

We wanted to determine whether the theoretical power function based on the normal approximation test is appropriate to use to evaluate the power and sample size requirements for the 2-Sample % Defective test in the Assistant. To do this, we needed to evaluate whether this theoretical power function accurately reflects the actual power of Fisher's exact test.

Method

The methodology for Fisher's exact test, including the calculation of its p-value, is described in detail in Appendix B. The theoretical power function based on the normal approximation test is defined in Appendix C. Based on these definitions, we performed simulations to estimate the actual power levels (which we refer to as simulated power levels) of Fisher's exact test when it is used to analyze the difference in % defectives from two samples.

To perform the simulations, we generated random pairs of samples of various sizes from several Bernoulli populations. For each category of Bernoulli population, we performed Fisher's exact test on each pair of 10,000 sample replicates. For each sample size, we calculated the simulated power of the test to detect a given difference as the fraction of the 10,000 pairs of samples for which the test was significant. For comparison, we also calculated the corresponding theoretical power based on the normal approximation test. If the approximation works well, the theoretical and simulated power levels should be close. For more details, see Appendix E.

Results

Our simulations showed that, in general, the theoretical power function of the normal approximation test and the simulated power function of Fisher's exact test are nearly equal. Therefore, the Assistant uses the theoretical power function of the normal approximation test to estimate the samples sizes needed to detect practically important differences when performing Fisher's exact test.

Data checks

Validity of CI

Because the 2-Sample % Defective test uses an exact test to evaluate the difference in % defectives, its accuracy is not greatly affected by the number of defectives and nondefectives in each sample. However, the confidence interval for the difference between the % defectives is based on a normal approximation. When the number of defectives and nondefectives in each sample increases, the accuracy of the approximate confidence interval also increases (see Appendix D).

Objective



We wanted to determine whether the number of defectives and the number of nondefectives in the samples are sufficient to ensure that the approximate confidence interval for the difference in % defectives is accurate.

Method

We used the general rule found in most statistical textbooks. When each sample contains at least 5 defectives and 5 nondefectives, the approximate confidence interval for the 2-sample % defective test is accurate. For more details, see the 2-sample % defective method section above.

Results

As shown in the simulations summarized in the 2-Sample % Defective Method section, the accuracy of the confidence interval depends on the minimum number of defectives and nondefectives in each sample. Therefore, the Assistant Report Card displays the following status indicators to help you evaluate the accuracy of the confidence interval for the difference between two % defectives:

Status	Condition
	Both samples have at least 5 defectives and 5 nondefectives. The confidence interval for the difference should be accurate.
	Either the number of defectives or the number of nondefectives in at least one sample is less than 5. The confidence interval for the difference may not be accurate.

Sample size

Typically, a statistical hypothesis test is performed to gather evidence to reject the null hypothesis of “no difference”. If the sample is too small, the power of the test may not be adequate to detect a difference 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.

Objective

If the data does not provide sufficient evidence to reject 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, they should not be so large that meaningless differences become statistically significant with high probability.




Method



The power and sample size analysis for the 2-Sample % Defective test is based upon the theoretical power function of the normal approximation test, which provides a good estimate of the actual power of Fisher’s exact test (see the simulation results summarized in Performance of Theoretical Power Function in the 2-Sample % Defective Method section). The theoretical power function may be expressed as a function of the target difference in % defective and the overall % defective in the combined samples.

Results

When the data does not provide enough evidence against the null hypothesis, the Assistant uses the power function of the normal approximation test to calculate the practical differences that can be detected with an 80% and a 90% probability for the given sample size. In addition, if the user provides a particular practical difference of interest, the Assistant uses the power function of the normal approximation test to calculate sample sizes that yield an 80% and a 90% chance of detection of the difference.

To help interpret the results, the Assistant Report Card for the 2-Sample % Defective Test displays the following status indicators when checking for power and sample size:

Status	Condition
	The test finds a difference between 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 $\geq .90$).
	Power may be sufficient. The test did not find a difference between the % defectives, but the sample is large enough to provide an 80% to 90% chance of detecting the given difference ($.80 \leq \text{power} < .90$). The sample size required to achieve 90% power is reported.
	Power might not be sufficient. The test did not find a difference between the % defectives, and the sample is large enough to provide a 60% to 80% chance of detecting the given difference ($.60 \leq \text{power} < .80$). The sample sizes required to achieve 80% power and 90% power are reported.

Status	Condition
	The power is not sufficient. The test did not find a difference between the % defectives, and the sample is not large enough to provide at least a 60% chance of detecting the given difference (power < .60). The sample sizes required to achieve 80% power and 90% power are reported.
	The test did not find a difference between the % defectives. You did not specify a practical difference to detect. Depending on your data, the report may indicate the differences that you could detect with 80% and 90% chance, based on your sample size and alpha.

References

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.

Appendix A: Additional assumptions for 2-sample % defective test

The 2-Sample % Defective test is based on the following assumptions:

- The data in each sample consist of n distinct items, with each item classified as either defective or not defective.
- The probability of an item being defective is the same for each item within a sample.
- The likelihood of an item being defective is not affected by whether another item is defective or not.

These assumptions cannot be verified in the data checks of the Assistant Report Card because summary data, rather than raw data, is entered for this test.

Appendix B: Fisher's exact test

Suppose that we observe two independent random samples X_1, \dots, X_{n_1} and Y_1, \dots, Y_{n_2} from Bernoulli distributions, such that

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

In the following sections, we describe the procedures for making inferences about the difference between the proportions $\delta = p_1 - p_2$.

Formula B1: Fisher's exact test and p-value

A description of Fisher's exact test can be found in Arnold (1994). We provide a brief description of the test.

Let V be the number of successes in the first sample and $v = n_1 \hat{p}_1$ be the observed number of successes in the first sample when an experiment is performed. Also, let W be the total number of successes in the two samples and $w = n_1 \hat{p}_1 + n_2 \hat{p}_2$ be the observed successes when an experiment is performed. Note that \hat{p}_1 and \hat{p}_2 are the sample point estimates of p_1 and p_2 .

Under the null hypothesis that $\delta = p_1 - p_2 = 0$, the conditional distribution of V given W is the hyper-geometric distribution with the probability mass function

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

Let $F(v|w)$ be the c.d.f of the distribution. Then the p-values for the one-sided and two-sided tests are:

- **When testing against $\delta < 0$ or equivalently $p_1 < p_2$**

The p-value is calculated as $F(v|w)$, where v is the observed value of V or the observed number of successes in the first sample and w is the observed value of W or the observed number of successes in both samples.

- **When testing against $\delta > 0$ or equivalently $p_1 > p_2$**

The p-value is calculated as $1 - F(v - 1|w)$, where v is the observed value of V or the observed number of successes in the first sample and w is the observed value of W or the observed number of successes in both samples.

- **When testing against $\delta \neq 0$ or equivalently $p_1 \neq p_2$**

The p-value is calculated according to the following algorithm, where m is the mode of hypergeometric distribution described above.

- If $v < m$, then the p-value is calculated as $1 - F(y - 1|w) + F(v|w)$, where v and w are as defined above and $y = \min\{k \geq m: f(k|w) \leq f(v|W)\}$
- If $v = m$, then the p-value is 1.0
- If $v > m$, then the p-value is calculated as $1 - F(v - 1|w) + F(y|w)$, where v and w are as defined above and $y = \max\{k \leq m: f(k|w) \leq f(v|W)\}$

Appendix C: Theoretical power function

To compare two proportions (or, more specifically, two % defectives), we use Fisher's exact test, as described in Appendix B. Because a theoretical power function of this test is too complex to derive, we use an approximate power function. More specifically, we use the power function of the well-known normal approximation test for two proportions to approximate the power of Fisher's exact test.

The power function of the normal approximation for the two-sided test is

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

where $\delta = p_1 - p_2$,

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

and $p_c = (n_1p_1 + n_2p_2)/(n_1 + n_2)$.

When testing $p_1 = p_2$ against $p_1 > p_2$ the power function is

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

When testing $p_1 = p_2$ against $p_1 < p_2$ the power function is

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

Appendix D: Approximate confidence intervals

Formula D1: Calculating an approximate confidence interval for the difference between 2 proportions

An asymptotic $100(1 - \alpha)\%$ confidence interval for $\delta = p_1 - p_2$ based on the normal approximation is:

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

A well-known general rule for assessing the reliability of this approximate confidence interval is $n_1\hat{p}_1 \geq 5$, $n_1(1 - \hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$. In other words, the confidence interval is accurate if the observed number of successes and failures in each sample is at least 5.

Note: In this section and the sections that follow, we express the rule for the confidence interval in its most general form, in terms of the number of successes and the number of failures in each sample. A success is the event of interest and a failure is the complement of the event of interest. Therefore, in the specific context of the 2-Sample % Defective Test, the number of “successes” is equivalent to the number of defectives and the number of “failures” is equivalent to the number of nondefectives.

Formula D2: Rules for approximate confidence intervals

The general rule used for confidence intervals based on the normal approximation states that the confidence intervals are accurate if $n_1\hat{p}_1 \geq 5$, $n_1(1 - \hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$. That is, the actual confidence level of the interval is equal to or approximately equal to the target confidence level if each sample contains at least 5 successes (defectives) and 5 failures (nondefectives).

The rule is expressed in terms of the estimated proportions of successes and failures as opposed to the true proportions because in practice the true proportions are unknown. However, in theoretical settings where the true proportions are assumed or known, the rule can be directly expressed in terms of the true proportions. In these cases, one can directly assess how the true expected number of successes and expected number of failures, n_1p_1 , n_2p_2 , $n_1(1 - p_1)$, and $n_2(1 - p_2)$, affect the actual coverage probability of the confidence interval for the difference between the proportions.

We can evaluate the actual coverage probability by sampling a large number of pairs of samples of sizes n_1 and n_2 from the two Bernoulli populations with probability of successes p_1 and p_2 . The actual coverage probability is then calculated as the relative frequency of the pairs of

samples yielding confidence intervals that contain the true difference between the two proportions. If the actual coverage probability is accurate when $n_1p_1 \geq 5$, $n_2p_2 \geq 5$, $n_1(1 - p_1) \geq 5$, and $n_2(1 - p_2) \geq 5$ then by the strong law of large numbers, the coverage probability is accurate when $n_1\hat{p}_1 \geq 5$, $n_1(1 - \hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$. Thus, when the actual and the target confidence level are close, one would expect a large proportion of the pairs of the samples generated from the two Bernoulli populations to be such that $n_1\hat{p}_1 \geq 5$, $n_1(1 - \hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$ if this rule is valid. In the simulation that follows, we refer to this rule as Rule 1.

In addition, in the course of this investigation, in many cases, we noticed that if either $n_1p_1 \geq 5$ and $n_2p_2 \geq 5$ or if $n_1(1 - p_1) \geq 5$ and $n_2(1 - p_2) \geq 5$, then the simulated coverage probability of the interval is near the target coverage. This prompted an alternative and more relaxed rule that states that the approximate confidence intervals are accurate if $n_1\hat{p}_1 \geq 5$ and $n_2\hat{p}_2 \geq 5$, or if $n_1(1 - \hat{p}_1) \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$. In the simulation that follows, we refer to this modified rule as Rule 2.

Simulation D1: Evaluating accuracy of approximate confidence intervals

We performed simulations to evaluate the conditions under which the approximate confidence interval for the difference between two proportions is accurate. In particular, we examined the accuracy of the interval in relation to the following general rules:

Rule 1 (original) $n_1p_1 \geq 5$, $n_2p_2 \geq 5$, $n_1(1 - p_1) \geq 5$, and $n_2(1 - p_2) \geq 5$

Rule 2 (modified) $n_1\hat{p}_1 \geq 5$ and $n_2\hat{p}_2 \geq 5$ OR $n_1(1 - \hat{p}_1) \geq 5$ and $n_2(1 - \hat{p}_2) \geq 5$

In each experiment, we generated 10,000 pairs of samples from pairs of Bernoulli populations defined by the following proportions:

- **A-proportions: both p_1 and p_2 are near 1.0 (or near 0).** To represent this pair of Bernoulli populations in the simulation, we used $p_1 = 0.8$ and $p_2 = 0.9$.
- **B-proportions: both p_1 and p_2 are near 0.5.** To represent this pair of Bernoulli populations in the simulation we used $p_1 = 0.4$ and $p_2 = 0.55$.
- **C-proportions: p_1 is near 0.5 and p_2 is near 1.0** To represent this pair of Bernoulli populations in the simulation, we used $p_1 = 0.4$ and $p_2 = 0.9$.

The classification of proportions above is based on the DeMoivre-Laplace normal approximation to the binomial distribution from which the approximate confidence intervals are derived. This normal approximation is known to be accurate when the Bernoulli sample is larger than 10 and the probability of success is near 0.5. When the probability of success is near 0 or 1, a larger Bernoulli sample is generally required.

We fixed the sample sizes for both pairs at a single value of n , where $n = 10, 15, 20, 30, \dots, 100$. We limited the study to balanced designs ($n_1 = n_2 = n$) without any loss of generality because

both rules depend on the observed number of successes and failures, which can be controlled by the size of the samples and the proportion of successes.

To estimate the actual confidence level of the confidence interval for the difference in the two population proportions (referred to as the simulated confidence level), we calculated the proportion of the 10,000 intervals that contain the true difference between the two proportions. The target coverage probability in each experiment was 0.95. In addition, we determined the percentage of the 10,000 samples for which the conditions under the two rules were satisfied.

Note: For some small samples, the estimated standard error of the difference between the proportions was 0. We considered those samples “degenerate” and discarded them from the experiment. As a result, the number of sample replicates was slightly less than 10,000 in a few cases.

The results are shown in Tables 1-11 and graphically displayed in Figure 1 below.

Table 1 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=10$. The target coverage probability is 0.95.

$n = 10$						
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p_1	0.80	8.00	2.00	0.907	0.0
	p_2	0.90	9.00	1.00		
B	p_1	0.40	4.00	6.00	0.928	4.4
	p_2	0.55	5.50	4.50		
C	p_1	0.45	4.50	5.50	0.919	0.0
	p_2	0.90	9.00	1.00		

Table 2 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=15$. The target coverage probability is 0.95.

$n = 15$						
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p_1	0.80	12.00	3.00	0.938	0.2
	p_2	0.90	13.50	1.50		
B	p_1	0.40	6.00	9.00	0.914	65.0
	p_2	0.55	8.25	6.75		

n = 15						
Category	Proportion (p)	np	n(1 - p)	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
C	p₁	0.45	6.75	8.25	0.930	1.2
	p₂	0.90	13.50	1.50		

Table 3 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for n=20. The target coverage probability is 0.95.

n = 20						
Category	Proportion (p)	np	n(1 - p)	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p₁	0.80	16.00	4.00	0.942	1.5
	p₂	0.90	18.00	2.00		
B	p₁	0.40	8.00	12.00	0.943	92.8
	p₂	0.55	11.00	9.00		
C	p₁	0.45	9.00	11.00	0.934	4.1
	p₂	0.90	18.00	2.00		

Table 4 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for n=30. The target coverage probability is 0.95.

n = 30						
Category	Proportion (p)	np	n(1 - p)	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p₁	0.80	24.00	6.00	0.941	4.3
	p₂	0.90	27.00	3.00		
B	p₁	0.40	12.00	18.00	0.944	99.7
	p₂	0.55	16.50	13.50		
C	p₁	0.45	13.50	16.50	0.938	7.2
	p₂	0.90	27.00	3.00		

Table 5 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=40$. The target coverage probability is 0.95.

$n = 40$							
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2	
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			

Table 6 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=50$. The target coverage probability is 0.95.

$n = 50$							
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2	
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			

Table 7 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=60$. The target coverage probability is 0.95.

$n = 60$						
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p_1	0.80	48.00	12.00	0.947	72.8
	p_2	0.90	54.00	6.00		
B	p_1	0.40	24.00	36.00	0.947	100.0
	p_2	0.55	33.00	27.00		
C	p_1	0.45	27.00	33.00	0.949	73.1
	p_2	0.90	54.00	6.00		

Table 8 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=70$. The target coverage probability is 0.95.

$n = 70$						
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2
A	p_1	0.80	56.00	14.00	0.939	71.7
	p_2	0.90	63.00	7.00		
B	p_1	0.40	28.00	42.00	0.945	100.0
	p_2	0.55	38.50	31.50		
C	p_1	0.45	31.50	38.50	0.944	71.8
	p_2	0.90	63.00	7.00		

Table 9 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=80$. The target coverage probability is 0.95.

$n = 80$							
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2	
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			
C	p_1	0.45	36.00	44.00	0.948	91.3	100.0
	p_2	0.90	72.00	8.00			

Table 10 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=90$. The target coverage probability is 0.95.

$n = 90$							
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2	
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			

Table 11 Simulated coverage probabilities and percentage of samples satisfying Rule 1 and Rule 2 for $n=100$. The target coverage probability is 0.95.

$n = 100$							
Category	Proportion (p)	np	$n(1 - p)$	Coverage Probability	%Samples Satisfying Rule 1	%Samples Satisfying Rule 2	
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			

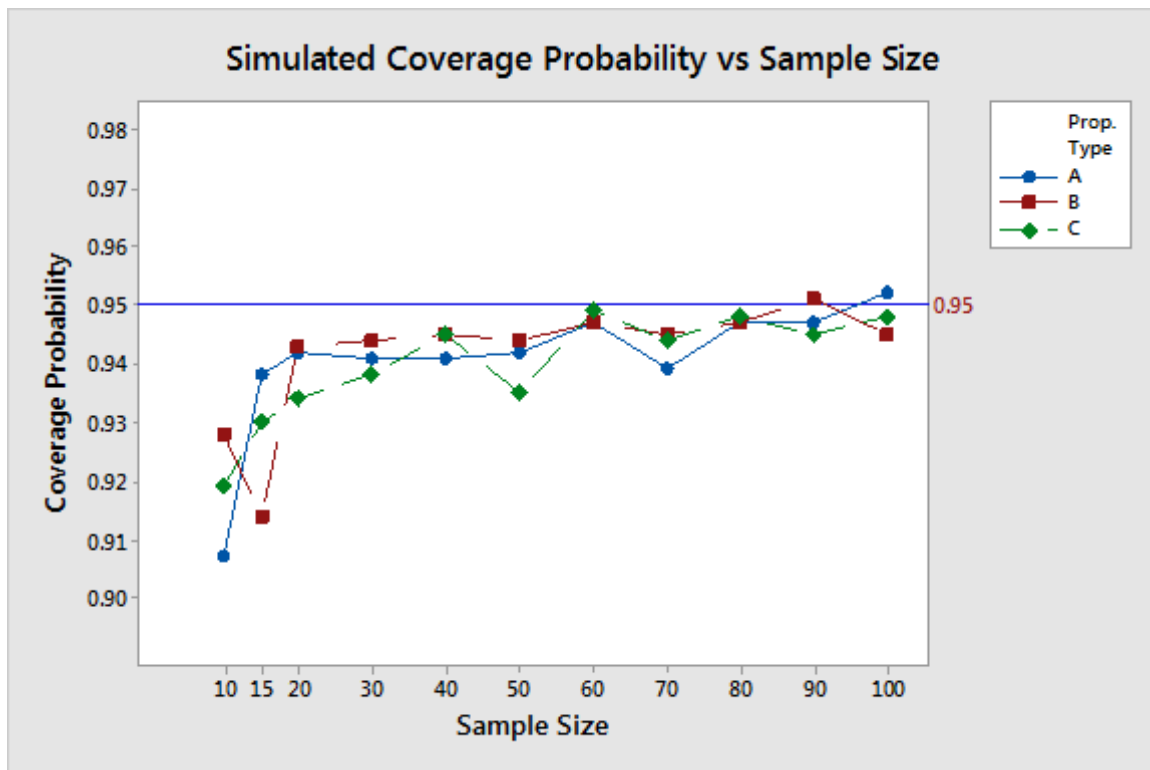


Figure 1 Simulated coverage probabilities plotted against sample size for each category of Bernoulli populations.

The results in Tables 1-11 and Figure 1 show that samples generated from Bernoulli populations in category B (when both proportions are close to 0.5) generally yield simulated coverage probabilities that are more stable and close to the target coverage of 0.95. In this category, the

expected numbers of successes and failures in both populations is larger than in the other categories, even when the samples are small.

On the other hand, for the samples generated from the pairs of Bernoulli populations in category A (when both proportions are near 1.0) or in category C (when one proportion is near 1.0 and the other near 0), the simulated coverage probabilities are off target in the smaller samples, except when either the expected number of successes (np) or the expected number of failures ($n(1-p)$) is large enough.

For example, consider the samples generated from the Bernoulli populations in category A when $n = 15$. The expected numbers of successes are 12.0 and 13.5 and the expected numbers of failures are 3.0 and 1.5 for each population, respectively. Even though the expected number of failures is less than 5 for both populations, the simulated coverage probability is about 0.94. Results such as these led us to create Rule 2, which requires only that *either* the expected number of successes *or* the expected number of failures be greater than or equal to 5 for each sample.

To more fully evaluate how effectively Rule 1 and Rule 2 can assess the approximation for the confidence interval, we plotted the percentage of samples satisfying Rule 1 and the percentage of samples satisfying Rule 2 against the simulated coverage probabilities in the experiments. The plots are displayed in Figure 2.

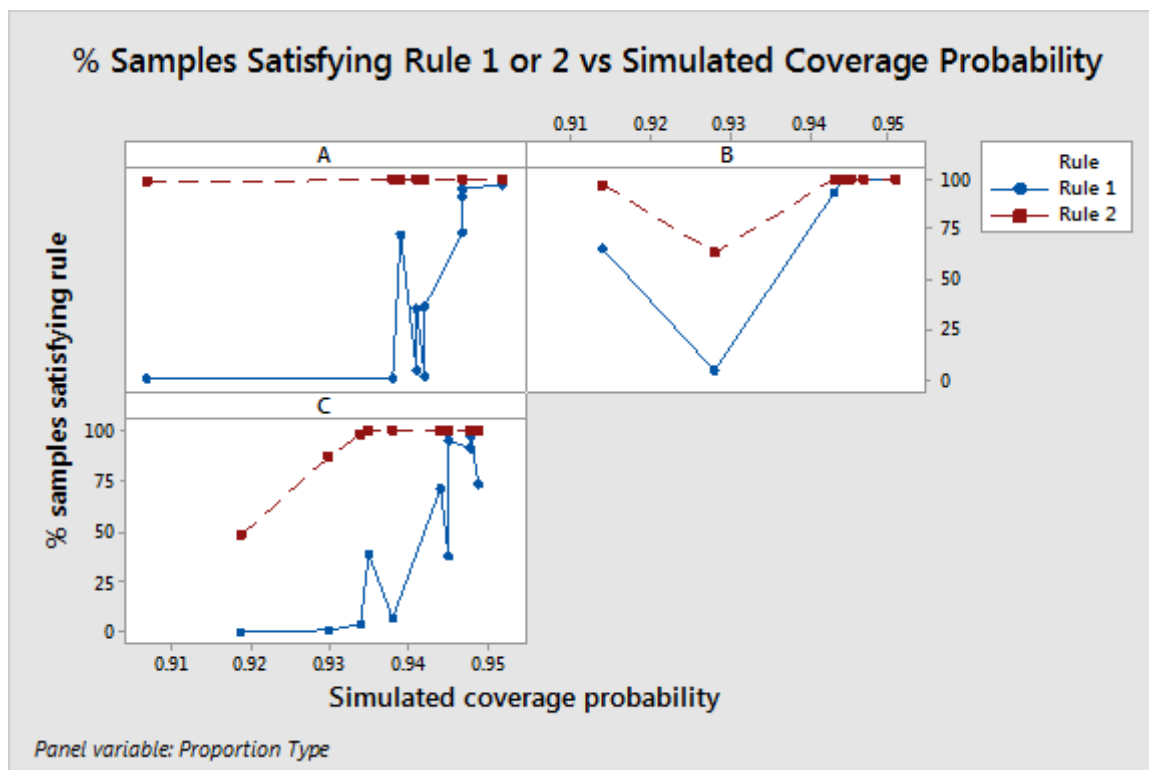


Figure 2 The percentage of samples satisfying Rule 1 and Rule 2 plotted against the simulated coverage probability, for each category of Bernoulli populations.

The plots show that as the simulated coverage probabilities approach the target coverage of 0.95, the percentage of samples that meet the requirements of each rule generally approaches 100%. For samples generated from Bernoulli populations in categories A and C, Rule 1 is stringent when samples are small, as evidenced by the extremely low percentage of samples satisfying the rule, even though the simulated coverage probabilities are close to the target. For example, when $n = 20$ and the samples are generated from the Bernoulli populations in category A, the simulated coverage probability is 0.942 (see Table 3). However, the proportion of samples satisfying the rule is nearly 0 (0.015) (see Figure 2). Therefore, in these cases, the rule may be too conservative.

Rule 2, on the other hand, is less stringent for small samples generated from the Bernoulli populations in category A. For example, as shown in Table 1, when $n = 10$ and the samples are generated from the Bernoulli populations in category A, the simulated coverage probability is 0.907 and 99.1% of the samples satisfy the rule.

In conclusion, Rule 1 tends to be overly conservative when samples are small. Rule 2 is less conservative and may be preferred when the sample sizes are small. However, Rule 1 is well known and well accepted. Although Rule 2 shows promising potential, in some cases it can be too liberal, as shown earlier. One possibility is to combine the two rules to take advantage of each rule's strengths; however, this approach requires further investigation before it can be applied.

Appendix E: Comparison of actual power versus theoretical power

Simulation E1: Estimating actual power using Fisher's exact test

We designed a simulation to compare the estimated actual power levels (referred to as simulated power levels) of Fisher's exact test to the theoretical power levels based on the power function of the normal approximation test (referred to as approximate power levels). In each experiment, we generated 10,000 pairs of samples from pairs of Bernoulli populations. For each pair of samples, the proportions were chosen so that the difference between the proportions was $p_1 - p_2 = -0.20$.

- **A-proportions: both p_1 and p_2 are near 1.0 (or near 0).** To represent this pair of Bernoulli populations in the simulation, we used $p_1 = 0.70$ and $p_2 = 0.90$.
- **B-proportions: both p_1 and p_2 are near 0.5.** To represent this pair of Bernoulli populations in the simulation, we used $p_1 = 0.40$ and $p_2 = 0.60$.
- **C-proportions: p_1 is near 0.5 and p_2 is near 1.0.** To represent this pair of Bernoulli populations in the simulation, we used $p_1 = 0.55$ and $p_2 = 0.75$.

We fixed the sample sizes for both pairs at a single value of n , where $n = 10, 15, 20, 30, \dots, 100$. We limited the study to balanced designs ($n_1 = n_2 = n$) because typically one assumes that the two samples have the same size. We calculated a common sample size needed to detect a practically important difference with a certain power.

To estimate the actual power for Fisher's exact test based on the results of each simulation, we calculated the fraction of the 10,000 sample pairs for which the two-sided test was significant at the target level of significance, $\alpha = 0.05$. Then we calculated the corresponding theoretical power levels based on the normal approximation test for comparison. The results are shown in Table 12 below.

Table 12 Simulated power levels of Fisher’s exact test compared with approximate power levels for the three categories of Bernoulli populations. The target level of significance is $\alpha = 0.05$.

n	A-Proportions		B-Proportions		C-Proportions	
	$p_1 = 0.70$ $p_2 = 0.90$		$p_1 = 0.40$ $p_2 = 0.60$		$p_1 = 0.55$ $p_2 = 0.75$	
	Simulated Power	App. Power	Simulated Power	App. Power	Simulated Power	App. Power
10	0.063	0.193	0.056	0.140	0.056	0.149
15	0.151	0.271	0.097	0.190	0.101	0.204
20	0.244	0.348	0.146	0.240	0.183	0.259
30	0.370	0.490	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	0.560	0.635
70	0.814	0.849	0.610	0.661	0.649	0.703
80	0.870	0.893	0.660	0.720	0.716	0.760
90	0.907	0.925	0.716	0.770	0.772	0.808
100	0.939	0.948	0.792	0.812	0.812	0.848

The results in Table 12 show that the approximate power tends to be higher than the simulated power for all three categories of Bernoulli populations (A, B, and C). For example, for the proportions in category A, the actual sample size required to detect an absolute difference of -0.20 with an approximate power level of 0.91 is about 90. In contrast, the corresponding sample size estimate based on the approximate theoretical power function is about 85. Therefore, the sample size estimate based on the approximate power function is generally slightly smaller than the actual sample size required to achieve a given power level.

You can see this relationship more clearly when the results are displayed as power curves, as shown in Figure 3 below.

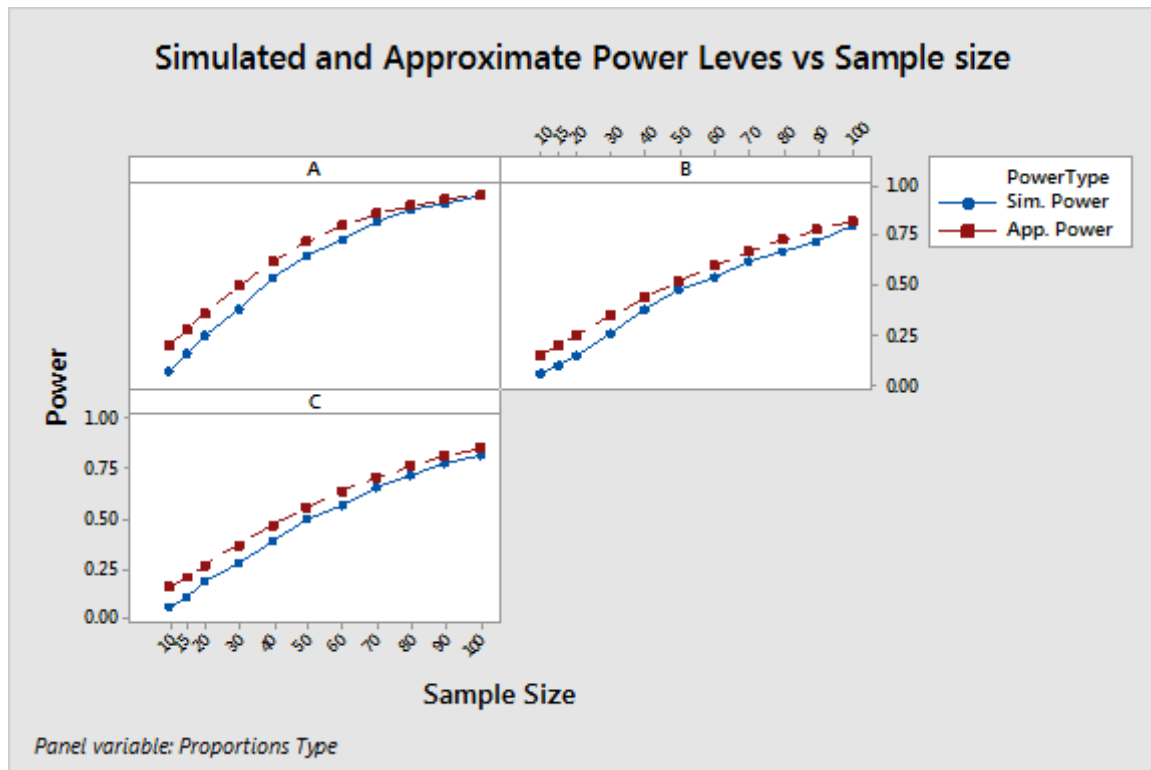


Figure 3 Plots of simulated and approximate power levels of the two-sided test for comparing two proportions. The power levels are plotted against sample size in separate panels for each category of Bernoulli populations.

Notice that although the simulated power curves are lower than the approximate power curves for all three categories of Bernoulli populations (A, B, and C), the size of the difference between the curves depends upon the true proportions of the Bernoulli populations from which the samples are drawn. For example, when the two proportions are near 0.5 (category B), the two power levels are generally close. However, the disparity between the two power curves is more noticeable in small samples for the proportions associated with population categories A and C.

These results show in general, the theoretical power function of the normal approximation test and the simulated power function of Fisher's exact test are nearly equal. Therefore, the Assistant uses the theoretical power function of the normal approximation test to estimate the sample sizes before performing Fisher's exact test. However, the sample sizes calculated using the approximate power function may be a bit smaller than the actual sample sizes required to achieve a given power to detect a difference between the two proportions (% defectives).

© 2015, 2017 Minitab Inc. All rights reserved.

Minitab®, Quality. Analysis. Results.® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries. See minitab.com/legal/trademarks for more information.