Value of small sample sizes in rapid-cycle quality improvement projects

E Etchells,[1,2] M Ho,[3] K G Shojania[1,2]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2015-005094).

[1]Department of Medicine, Sunnybrook Health Sciences Centre and the University of Toronto, Toronto, Ontario, Canada
[2]Centre for Quality Improvement and Patient Safety, University of Toronto, Toronto, Ontario, Canada
[3]Postgraduate Training Program in Internal Medicine, University of British Columbia, Vancouver, Canada

Correspondence to
Dr Edward Etchells, Department of Medicine, Sunnybrook Health Sciences Centre, Rm H469, Toronto, Ontario, Canada M4N 3M5; edward.etchells@sunnybrook.ca

Accepted 27 November 2015
Published Online First 30 December 2015

▶ http://dx.doi.org/10.1136/bmjqs-2015-005076

To cite: Etchells E, Ho M, Shojania KG. *BMJ Qual Saf* 2016;**25**:202–206.

Quality improvement initiatives can become bogged down by excessive data collection. Sometimes the question arises —are we doing an adequate job with respect to a recommended practice? Are we complying with some guideline in at least X% of our patients? The perception that one must audit large numbers of charts may present a barrier to initiating local improvement activities. The model for improvement and its Plan–Do–Study–Act (PDSA) cycles typically require frequent data collection to test ideas and refine the planned change strategy. The perception that data collection must involve many patients can lead to insufficiently frequent PDSA cycles.[1] In this review, we demonstrate the important contributions that small samples can make to improvement projects, including local audits, PDSA cycles and during broader implementation and evaluation.

SMALL SAMPLES FOR DEMONSTRATING LOCAL GAPS IN CARE

Suppose you are a hospital-based clinician who has joined a medication reconciliation working group. Medication reconciliation refers to efforts to avoid unintentional changes to medication regimens at transition points such as hospital admission and discharge.[2] You notice that medication reconciliation did not occur for several patients on your service this week. Your institution sets a target medication reconciliation rate of at least 80%, based on external standards and internal commitments to patient safety. You decide to audit 20 consecutive admissions, and find that only 10 charts (50%) have completed medication reconciliation. You present your findings at the weekly team meeting. Your colleagues tactfully point out that your sample is far too small to draw any meaningful conclusions.

Surprisingly, your sample of 20 consecutive admissions actually provides strong evidence that local performance falls short of your performance target. If your service were actually performing medication reconciliation 80% of the time, a sample of 20 charts would produce an observed reconciliation rate of only 50% (or worse) about three times out of every 1000 similar audits.[3] This probability corresponds to a p value of 0.003, well below the conventional threshold of p=0.05 for statistical significance. In other words, you can confidently reject the null hypothesis that your sample comes from a population in which medication reconciliation occurs at a rate of at least 80%.

This unexpectedly robust result is best understood by going back to high school math class, where students are asked to calculate probabilities related to flipping a fair coin. A fair coin should come up heads 50% of the time (null hypothesis). Suppose you flip the coin 20 times and observe 5 heads. The probability of observing 5 (or fewer) heads in 20 flips of a fair coin, with a 50% chance of coming up heads, is about 2% (p=0.02). A statistician would say that you would reject the null hypothesis of a fair coin. Simply put, someone is probably trying to swindle you! You can do this calculation yourself by going to a free online calculator such as http://vassarstats.net/.[4]

The medication reconciliation audit is analogous to a coin toss. The first difference is that the outcome of heads or tails replaced with successful ('heads') or failed ('tails') medication reconciliation. The second difference is that the expected probability has changed. With a fair coin, the expected probability of 'heads' is 50%. With our medication reconciliation audit, our expected probability of successful medication reconciliation

('heads') is 80%. Each audited chart resembles a toss of the coin and we can equate 'coming up heads' with successful medication reconciliation. Online supplementary appendix 1 shows the exact steps involved in reaching the conclusion that the probability of observing 10 (or fewer) successful reconciliations in 20 charts is about 0.003. With such a low p value, a statistician would say that you can confidently reject the null hypothesis of an 80% rate of medication reconciliation.

The practical implication is that improvement projects do not need large samples to demonstrate a gap in system performance. Table 1 shows the sample size requirements for local quality audits. You can use table 1 in two ways. First, on completing an audit, the table can quickly indicate if your result is statistically significant. For example, if your audit showed an observed system performance of 50% when the desired system performance is 80%, then an audit with a sample size of 12 or more will be statistically significant. Second, you can use this table to plan a sample size for an audit or PDSA cycle. For example, if your 'hunch' is that the observed system performance will be 50%, and you have a desired system performance of 90%, then a sample size as low as 6 will likely suffice (though there is no harm in planning to include a few additional observations to ensure that you have a sample that represents your system's usual performance, as discussed below in 'Can you make reasonable inferences about local system performance? (External validity)' Section).

How is it possible that such small samples permit rejecting the null hypothesis here, while properly designed controlled clinical trials need to enrol hundreds or thousands of patients? One reason is that we are looking at very large differences (eg, 50% vs 80%), whereas clinical trials typically look for much smaller differences. In fact, as shown in table 1, as the observed performance comes closer to the desired target we do require larger sample sizes to show significant differences. For example, you would need an audit sample size of 280 to show that 75% observed performance differed significantly from a desired performance of 80%.

A second reason for the surprisingly small sample sizes shown in table 1 is that clinical researchers want a precise estimate of treatment effect, whereas in local audits, the precision of the estimate of system performance is less important. In our audit, we found that 10/20 (50%) of charts had successful medication reconciliation. How sure are we that the system performance is really 50%? We are not sure at all. Statisticians use 95% CIs to describe the precision of study results (see online supplementary appendix 2 for details). Our audit has a 95% CI that extends from a low of 28% to a high of 72%. In other words, if 100 audits, each of 20 charts, were carried out, 95% of the audits would have a result between 28% and 72%. We would never want a clinical trial to produce a result like this: Drug X cured 50% of patients, but the cure rate could be as low as 28% or as high as 72%. But, for our audit, this result suffices to conclude that our local system performance falls short of 80%. We are less concerned about whether the actual performance is 28% or 72%, because both are unacceptable.

Table 1 Minimum sample sizes required for improvement projects based on observed and desired system performance

Observed system performance (%)	Desired system performance	
	80%	90%
95	26	140
90	70	Not applicable
85	260	180
80	Not applicable	50
75	280	28
70	80	20
66	45	15
60	25	10
50	12	6
40	10	5
20	5	5

The table shows the approximate sample size required to reject the null hypothesis that observed performance (from an audited sample) is consistent with the desired system performance, shown here as being either 80% or 90%. If you wish to calculate an exact p value for your audit or Plan–Do–Study–Act (PDSA) result, follow the steps in online supplementary appendix 1. If you wish to calculate the exact 95% CI for your audit or PDSA result, follow the steps in online supplementary appendix 2. The results shown here all use the conventional two-tailed p value of 0.05.

SMALL SAMPLES CAN MAKE 'RAPID IMPROVEMENT' RAPID

Small samples can also provide useful information in PDSA cycles and other rapid improvement methodologies, not just for simple audits of performance. Inadequate, infrequent data cycles are a common failing in improvement projects that use the PDSA methodology.[1] One reason for inadequate infrequent data cycles may be a tendency to collect too much data in any given cycle.

Suppose that your medication reconciliation audit has stimulated enthusiasm for local improvement. Your team's first change concept consists of a new medication reconciliation form that must be completed by the ordering provider. For your first PDSA cycle, you plan to obtain feedback from users about the form's usability. Your main study measure is whether the clinicians can complete the form without your help. How many clinicians should you study in this cycle?

You can use table 1 to plan your first PDSA. At this early stage you will likely be recruiting friendly highly motivated clinicians (a 'convenience sample') to try out your form. You should aim for at least a 90% success rate for completing the form without any difficulty. You do not want to implement a form that

requires training and personalised support for highly motivated users. Therefore, you will use the third column from table 1 with desired system performance of 90%. Next, you need a hunch about how good you can really expect your form to be in this first go-around. You should be humble, because at early stages nothing works out as intended. Let's estimate that 60% of clinicians will be able to complete the form without personalised help or difficulty. Therefore, a sample size of 10 should be sufficient. In other words, if, as you suspect, only 60% of your convenience sample will compete the form without help, you will only need observations to show that you are not yet at your target of 90% success.

For this first (convenience) sample of 10 volunteer users, 5/10 (50%) completed the form without any input or instructions. The other five became frustrated and gave up. Table 1 tells you that, with an observed success rate of 50% and a desired target of 90%, any audit with a sample of eight or more allows you to confidently reject the null hypothesis that your form is working at a 90% success rate. In other words, your form needs work! If you wish, you can also use the steps in online supplementary appendix 1 to calculate an exact p value (p=0.002) for the probability that you would observe a performance of only 50% if the true performance were 90%. And, online supplementary appendix 2 shows how to calculate the 95% CI for your result: (20%–80%). The quantitative element of the first PDSA cycle is already finished. You should obtain qualitative feedback from your 10 participants (especially the five motivated users who could not complete the form) and make the necessary changes. Then you can start a second PDSA cycle next week.

HANDLE SMALL SAMPLES WITH CARE

We have highlighted the degree to which small sample sizes can drive improvement efforts. Some readers may wonder: surely, there is a catch? (After all, 'There's no such thing as a free lunch.') The catch is simply this: you must handle your small samples with great care. This care is required so that (a) you can have confidence in your results (internal validity) and (b) you can make reasonable inferences about local system performance (external validity). The importance of data quality in larger quality improvement studies has recently been reviewed.[5]

Can you have confidence in your own results? (Internal validity)

You must have an extremely high level of confidence in the data integrity of your small sample. We associate heightened concerns about the integrity of data with large clinical trials. Ironically, the larger the trial, the less it matters if the occasional patient was lost to follow-up, or did not meet strict enrolment criteria. We are not suggesting that standards for the conduct of clinical trials should be relaxed. We simply point out that a trial involving 10 000 patients can tolerate questions about the enrolment of a few specific patients. By contrast, for the small sample sizes we have been discussing, a 'few specific patients' can amount to a large proportion of your sample. One patient represents a substantial contribution to a sample of eight patients. So, the 'catch' (if it can be called that) to using small samples is the need to follow very clear steps for collecting the data.

You can handle your small sample with care by applying five simple steps (see box 1):

1. Define the eligible sample
2. Establish exclusion criteria
3. State your study period
4. Keep a reject log
5. Make data collection complete

We have prepared an example of how to describe a small sample for a medication reconciliation audit in the box 1. First, you should define your eligible sample. For audits, you should aim to enrol *consecutive eligible* patients. Random samples are ideal, but needlessly complex and impractical for most local improvement initiatives. For early PDSA cycles, where the focus shifts to changing provider and system performance, it is practical to use convenience samples. A convenience sample is, essentially, 'whoever you can get'. For example, we used friendly volunteer clinicians for our first PDSA cycle of our medication reconciliation form. However, changes will usually perform better in convenience samples, who are generally highly selected to be motivated and willing to change. Therefore, once your change seems to be working at the desired level, you should conduct an audit using consecutive, unselected providers whenever possible. Of course you could also deliberately sample clinicians who are resistant to change and vocally opposed to your initiative (perhaps this would be called an 'inconvenience sample'.)

Box 1 Example of a carefully handled small sample

Eligible sample: we identified consecutive patients admitted to our inpatient medical service at General Hospital.
Exclusion criteria: we excluded patients who were admitted for <12 h.
Audit period: the audit occurred from Saturday 7 November 2015 at 08:00 h to Sunday 8 November 2015 at 16:00 h.
Reject log: we identified 23 consecutive admitted patients during the audit period. We excluded two patients who were discharged within 12 h, leaving 21 patients for the audit.
Completeness of data collection: we completed data collection for all 20 patients. One chart could not be located.

Second, there will be some patients who should be excluded because the audit or improvement efforts do not apply. In our medication reconciliation audit, we might exclude patients who were admitted for <12 h, because medication reconciliation is not expected to occur during such short admissions. Third, clearly state the start and end times for the audit or cycle. Fourth, keep track of patients who were excluded ('reject log'). In example shown in the box 1, there were 23 potentially eligible patients during the study period, but two were excluded because they were admitted for <12 h. This left exactly 21 patients for the audit.

The paramount concern then becomes completeness of data collection for these 21 patients. Suppose there were actually 21 patients eligible for the audit, but one chart was missing. We found that medication reconciliation occurred in 10/20 patients, but we do not know the one missing result. Therefore, the true results of our audit could have been 10/21 (48%, 95% CI 27% to 69%) or 11/21 (52%, 95% CI 31% to 73%). The incomplete data collection does not substantially alter our interpretation of the audit results, since the 95% CI would not include our target of 80% no matter what the outcome of the audit on the missing chart. By contrast, suppose there were 40 patients eligible for the audit, but 20 charts were missing. We found medication reconciliation in 10/20 of the remaining charts. What is the result of our audit now? The answer is: we don't know. The actual result of our small audit could be as poor as 10/40 (25%, 95% CI 12% to 38%) or as high as 30/40 (75%, 95% CI 62% to 88%). Because of our sloppy methods, we can conclude that our observed system performance is somewhere between 12% and 88%, making the entire exercise useless. Sometimes, the reason for missing charts or other causes of incomplete data may relate to the problem you are trying to solve. Maybe pharmacists have trouble finding charts when attempting to conduct medication reconciliation. We better track down those charts before we attempt to draw conclusions and influence our colleagues!

Can you make reasonable inferences about local system performance? (External validity)

Audits and PDSA cycles are primarily intended to measure and improve local performance. It is important to emphasise that you are not making any assertions about performance on other services or at other institutions. Regardless, you can anticipate criticism that even carefully handled small samples might not be representative of local system performance. For instance, our initial medication reconciliation audit was conducted on patients admitted on a weekend (see box 1). Your colleagues point out that fewer doctors work on weekends, so your sample of 20 charts reflects the performance of only two or three clinicians. They also feel that the workload and decreased support services (such as pharmacists) on weekends mean that your results cannot be generalised to weekday care.

From a statistical point of view, the point about the 20 charts reflecting the care of only a few clinicians raises the issue of 'clustered data."[6] Simply put, most statistical tests assume that measurements are independent. Each flip of a fair coin is independent. It does not matter whether heads came up on the prior flip. By contrast, small samples from a local audit will not be independent if, for instance, the audited charts all involve the same doctor. But, rather than delve into the technical issues involved in handled clustering data, readers interested in improvement can nonetheless appreciate that, if one wants to know how the local system is performing, a sample that reflects the performance of just one doctor will not suffice. You need to consider the degree to which your small sample is representative of local performance. In this case, this means making sure your sample includes charts from as many different doctors as possible. (If you were auditing, say, the use of pressure ulcer preventions strategies, you would similarly need to avoid sampling patients cared for by the same few nurses.)

The point about weekend care differing from weekdays may well be valid. Your colleagues believe that the best medication reconciliation performance will be midweek, when staffing is consistently the highest. Therefore, you decide to conduct a second audit of 20 consecutive patients admitted Tuesday and Wednesday by different medical teams. If the second audit result is similar to the first, you now have strong evidence of a gap in care. If you find excellent performance (100%) on weekdays, you can conclude that the system works well on weekdays, but not on weekends. Improvement efforts can be focused on closing the gap between weekends and weekdays. If the result is indeterminate (eg, 75% performance on weekdays) then another audit of a representative weekday sample can be conducted. In all cases, you have engaged your colleagues in your improvement efforts, and you are gathering useful data to help you understand current gaps and guide change efforts.

In general, you can constructively address criticisms from colleagues about audited samples by:

1. Having an excellent description of your carefully handled sample ('Can you have confidence in your own results? (Internal validity)' Section and box 1)
2. Asking your colleagues to describe why your sample may fail to represent local performance
3. Asking your colleagues to help you conduct another small audit using a sample that addresses their concerns.

SUMMARY

We sought with this review to demonstrate the value of small samples in improvement projects. Small samples can characterise local gaps in care that require improvement and support rapid-cycle improvement.

As you progress through your project and observed performance improves, you may need larger samples to see if you are still below target performance. But, the degree to which you need to know if you are in fact at 70%, 75% or 80% will vary depending on the project. Early on, though, when performance typically falls far short of the desired level, sample sizes of 10–20 observations often suffice. But, you must handle small samples with care, so that you (and anyone you are trying to convince) can be confident in the interpretation and application of the results.

Competing interests None declared.

Provenance and peer review Commissioned; internally peer reviewed.

REFERENCES

1 Taylor MJ, McNicholas C, Nicolay C, *et al*. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290–8.
2 Kwan JL, Lo L, Sampson M, *et al*. Medication reconciliation during transitions of care as a patient safety strategy: a systematic review. *Ann Intern Med* 2013;158(5 Pt 2): 397–403.
3 Altman DG. *Practical statistics for medical research*. New York, NY: Chapman and Hall. 1991:63–71.
4 http://vassarstats.net/ (accessed 26 Nov 2015).
5 Needham DM, Sinopoli DJ, Dinglas VD, *et al*. Improving data quality control in quality improvement projects. *Int J Qual Health Care* 2009;21:145–50.
6 Kerry SM, Bland JM. Sample Size in cluster randomization. *BMJ* 1998;316:549.

Correction: *Value of small sample sizes in rapid-cycle quality improvement projects*

Etchells E, Ho M, Shojania KG. Value of small sample sizes in rapid-cycle quality improvement projects. *BMJ Qual Safe* 2016;25:202–6.

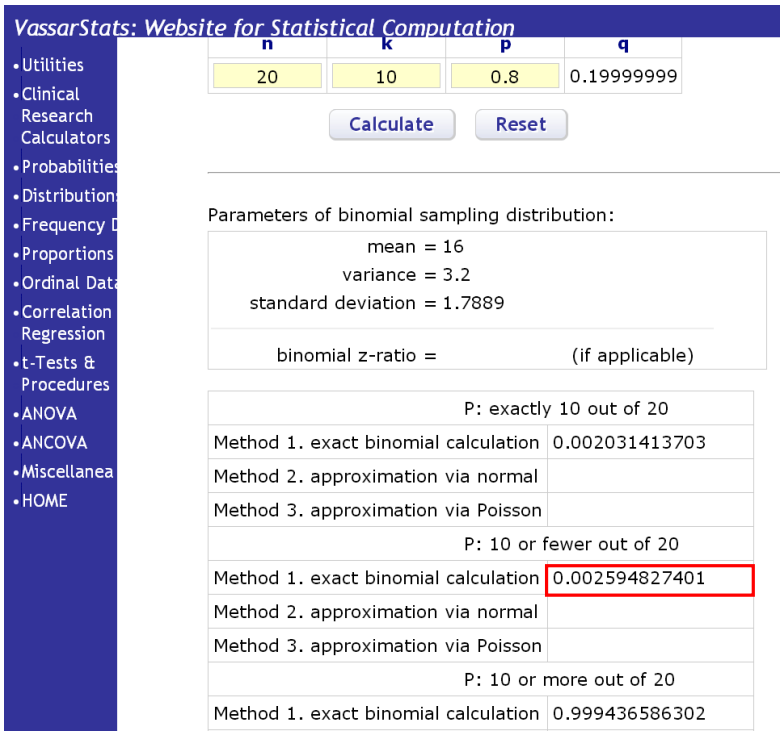
The article has been corrected since it was published online. The authors want to alert readers to the following error identified in the published version. The error is in the last paragraph of the section "Small samples can make 'rapid improvement' Rapid", wherein the minimum sample size has been considered as six instead of eight.

For this first (convenience) sample of 10 volunteer users, 5/10 (50%) completed the form without any input or instructions. The other five became frustrated and gave up. Table 1 tells you that, with an observed success rate of 50% and a desired target of 90%, any audit with a sample of six or more allows you to confidently reject the null hypothesis that your form is working at a 90% success rate.

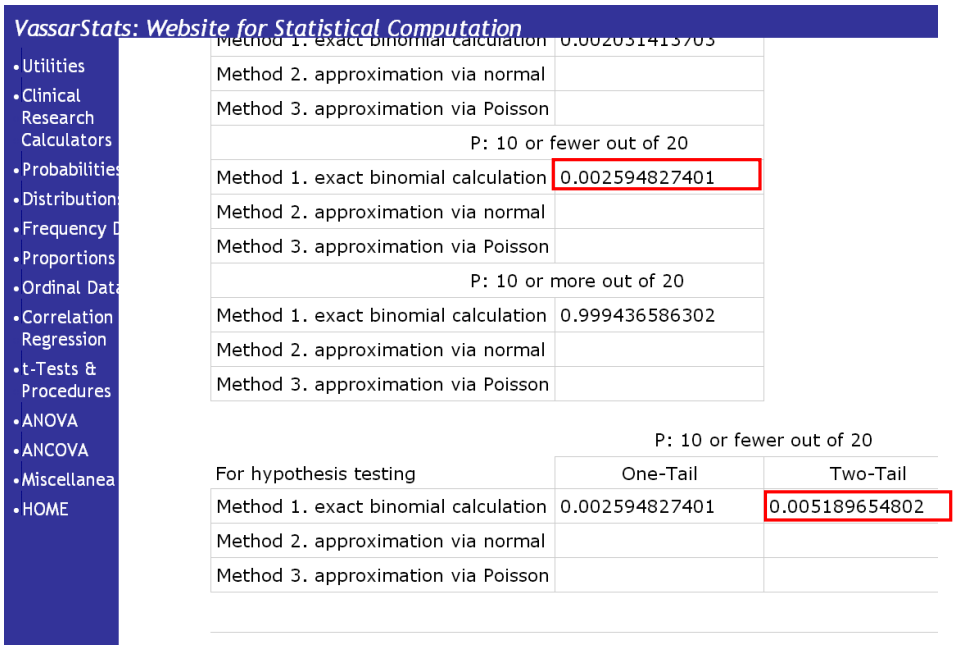
BMJ Qual Saf 2020;**29**:e1. doi:10.1136/bmjqs-2015-005094corr1

APPENDIX 1: Sample Calculation of binomial probability for results of small audit cycle

1. Go to http://vassarstats.net/ (accessed November 25th 2015)
2. From the left hand menu, choose "Probabilities" then "Binomial Probabilities"
3. **Enter value of 'n'.** n is the total number of your sample. For our medication reconciliation audit we had a sample size of 20.
4. **Enter the value for 'k'**, the number of 'successes'. In this case, the number of charts with medication reconciliation completed was 10.
5. **Enter value for p.** p is the expected probability from 0 (= 0%) to 1 (=100%). For quality audits this will be the desired system performance. For our medication reconciliation audit, we had a desired system performance of 80%, so we entered the value 0.8.
6. **Hit Calculate,** producing the following screen



7. The exact probability of observing 10 or fewer successes in 20 trials is about 0.003 (based on the value of 0.00259… in the first red box). For statistical significance we would use the two tailed exact probability of 0.005 (based on the value of 0.00518…. in the second red box, appearing further down the page in screen shot below).

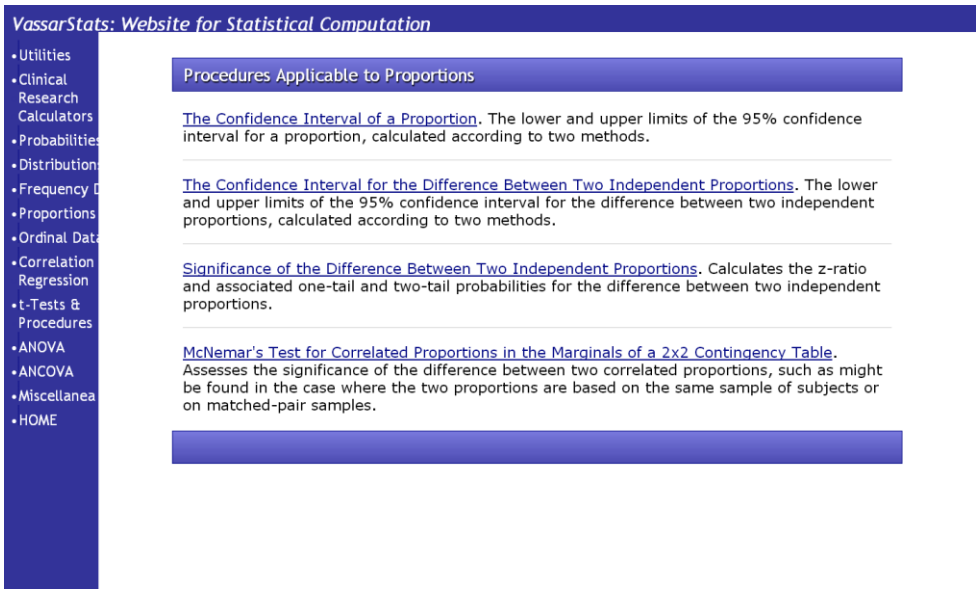


As discussed in the main text, this probability in the second red box (p=0.005) means the following. If local performance were truly at 80% (the target), the chance of observing this performance rate of only 50% (or worse) in 20 observations is only p=0.005 (i.e., 0.5%). Since this probability falls well below the conventional threshold of p=0.05 for statistical significance, you can confidently conclude that you are not operating at 80%.

Appendix 2: Calculating 95% confidence intervals for simple proportions

1. Go to http://vassarstats.net/

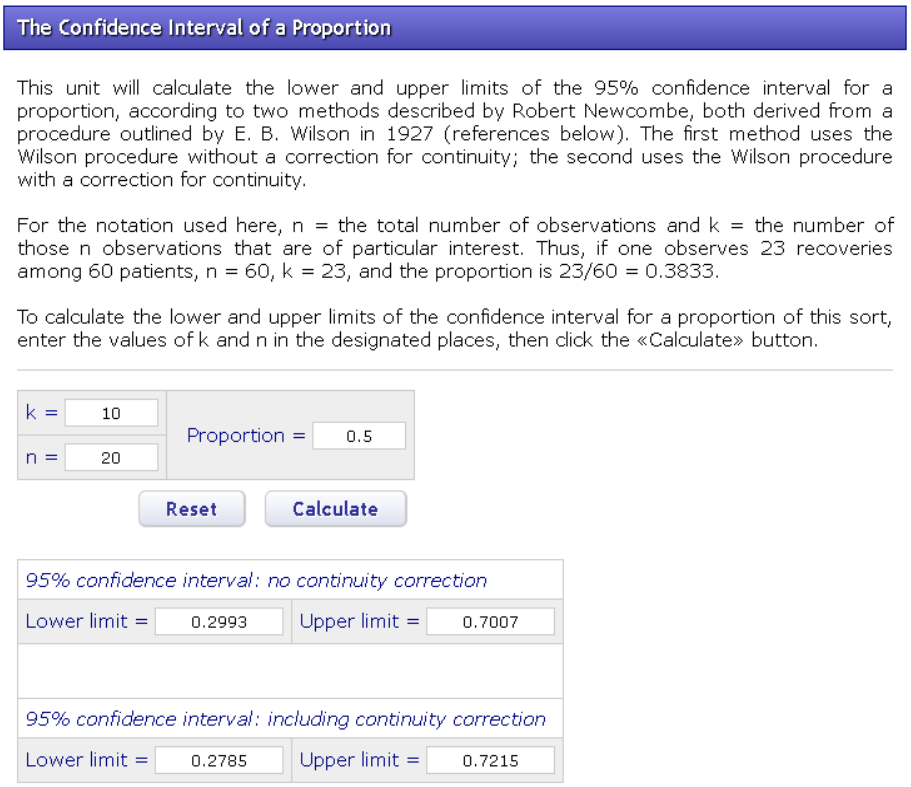
2. Choose **Proportions** from the left handed menu



3. Choose The Confidence Interval of a Proportion from the list of procedures

4. Enter the number of successes (k) and total sample size (n), then hit the calculate button

You will see two answers. Use the second answer ("including continuity correction")



The Confidence Interval of a Proportion

This unit will calculate the lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927 (references below). The first method uses the Wilson procedure without a correction for continuity; the second uses the Wilson procedure with a correction for continuity.

For the notation used here, n = the total number of observations and k = the number of those n observations that are of particular interest. Thus, if one observes 23 recoveries among 60 patients, n = 60, k = 23, and the proportion is 23/60 = 0.3833.

To calculate the lower and upper limits of the confidence interval for a proportion of this sort, enter the values of k and n in the designated places, then click the «Calculate» button.

k =	10	Proportion =	0.5
n =	20		

Reset Calculate

95% confidence interval: no continuity correction			
Lower limit =	0.2993	Upper limit =	0.7007
95% confidence interval: including continuity correction			
Lower limit =	0.2785	Upper limit =	0.7215

Correction: *Value of small sample sizes in rapid-cycle quality improvement projects*

Etchells E, Ho M, Shojania KG. Value of small sample sizes in rapid-cycle quality improvement projects. *BMJ Qual Safe* 2016;25:202–6.

The article has been corrected since it was published online. The authors want to alert readers to the following error identified in the published version. The error is in the last paragraph of the section "Small samples can make 'rapid improvement' Rapid", wherein the minimum sample size has been considered as six instead of eight.

For this first (convenience) sample of 10 volunteer users, 5/10 (50%) completed the form without any input or instructions. The other five became frustrated and gave up. Table 1 tells you that, with an observed success rate of 50% and a desired target of 90%, any audit with a sample of six or more allows you to confidently reject the null hypothesis that your form is working at a 90% success rate.

BMJ Qual Saf 2020;**29**:e1. doi:10.1136/bmjqs-2015-005094corr1