Are continuous measurements always better than binary ones?

Drs. Jan Willem Bikker PDEng

In statistics, it is often said that whenever possible a continuous version of a measurement is preferable to a binary one, at least as far as the statistical properties are concerned.

CQM frequently encounters situations during projects where binary measurements are easy (and customary), and continuous measurements not so easy or requiring investment. For example:

Binary version	Continuous version
Product is leaking yes/no	At what pressure does leaking start?
Users give rating “is this acceptable?” yes/no	Users give rating on 5 point scale with top 2 considered ‘acceptable’
Comparison tests, where users are asked to compare two products. Binary version asks which they prefer	e.g. a 6-point scale; -3 = much prefer left -2 = prefer left -1 = marginally prefer left 1 = marginally prefer right 2 = prefer right 3 = much prefer right

It turns out the advantage of continuous measurements over binary ones is strongest in cases where the percentages are either very close to 0% or very close to 100%. In the examples above, this would be the case for the leaking product, though perhaps not in the other examples.

In this article we first consider the small percentages case and then look at the difference in the information acquired in cases where the percentages are nearer the middle.

Continuous measurements give more information for small percentages

Let’s take the example of mass production. A certain property of a product can be measured and the outcome defined as ‘pass’ or ‘fail’ to indicate the quality of the product. But often a continuous measurement can also be constructed alongside the binary one, with specification limits from which the product’s quality can be assessed. For example, leakage of a product (e.g. a food package). The binary measurement is simply leaks/doesn’t leak. Its continuous counterpart can measure at what pressure difference the product starts to leak. The specification is then a limit on the pressure difference.

When full marks isn’t the full story

Statisticians favour continuous over binary measurement because it yields more useful information. Suppose, for instance, a sample of 100 products shows no failures using the binary measurement.

This might seem good news, but that depends on the underlying process or batch (the 'population' in statistical jargon). If in reality 3% of the products are faulty we would on average expect 3 failures out of 100. However, 1, 2, 4 or 5 failures would still be common, and there's even a reasonable chance that we get 0 failures. So what does it prove if the test reports that 100 out of 100 products were good?

In fact, this can be neatly expressed as follows: with 95% confidence, the true failure rate is between 0 and 3.0% (see appendix for how to compute this). A possible failure rate of 3% is in many practical situations alarmingly high, and typical even in sample sizes as large as 100.

Commented [TC1]: "1, 2 or 3, 5 failures" – was this right??

Commented [TC2]: Check correct: previous version was confusing to me.

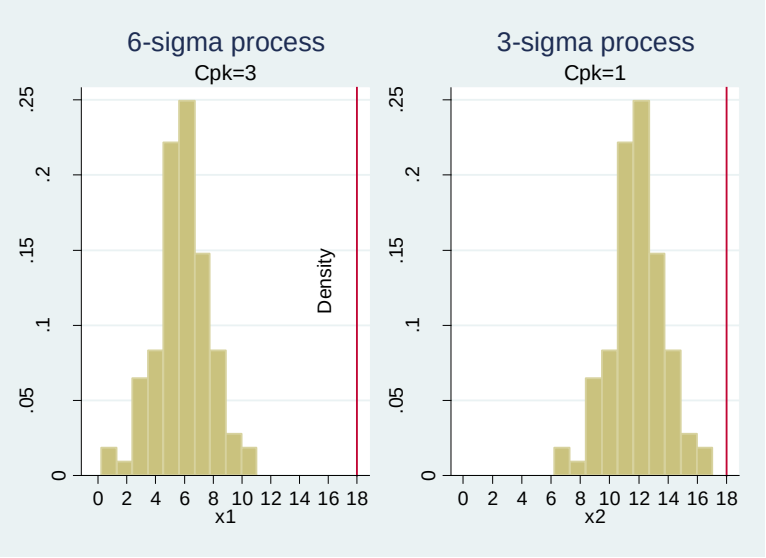
Commented [TC3]: Don't understand what is typical here? – that there will be 3% failure rate? Surely that depends on what you're testing.

Same products, different results

Suppose a continuous equivalent was available for the leakage measurements, and that products with a measured value greater higher than 18 are faulty and less than 18 good.

The following table shows 2 scenarios for x1 and x2, corresponding to a very good and a questionable production process, both having 0 out of 100 failures as before.

In the Six sigma process on the left, the observations x1 are a comfortably large distance from the USL=18 (Upper Specification Limit) and the observations seem to follow a reassuringly normal distribution. While on the right the observations are much closer to the USL.



Tables like these tend to encourage measures such as the sigma-level or Cpk of a process, as opposed to the proportion of failures (i.e. **ppm level** and **reject rate**). Under the additional assumption that the measurements follow a normal distribution, there is a direct relation between the Cpk (or sigma level) and the proportion of failures. Graphically, it's clear that the process on the left is much less likely to identify products as out-of-spec than the one on the right. Thus illustrating the benefit of a continuous measurement.

When percentages are not very small

Now let's take the example of a test to check whether a large majority of users finds a product acceptable. There are two possible protocols:

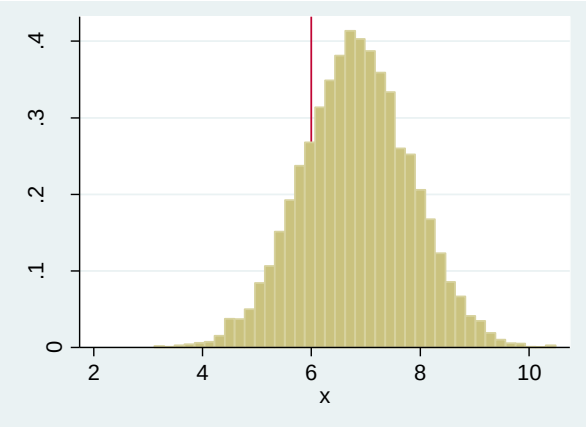
- Binary: each user must answer the question "is the product acceptable?" with "yes" or "no".
- Continuous: users rate the product on a scale from 1-10 (bad to good), with a rating of 6 or higher being 'acceptable'[1].

Setting up our power test

One way to assess the amount of information the protocols will provide is to take an example, calculate confidence intervals of the estimated percentage of 'acceptable', and see how much narrower the confidence intervals are in the continuous version. However, the confidence intervals for the binary case are not symmetric, and the calculations for the continuous case are approximate and non-standard. For this reason, we take another route and compare the power of both protocols.

The goal of the study is to prove statistically that at least 70% of the users find the product acceptable. The power of the study is the probability that your study will be successful. Here success means getting a statistically significant result (i.e. p-value < 5%).

The power depends on the actual percentage of users that finds the product acceptable. In this example we shall assume this actual percentage to be 80% and that the true, unknown distribution is as in the graph: a normal distribution with a mean that 80% of users give a rating of 6 or higher.



We have now the ingredients for a power study for both a binary and continuous test.

[1] In some marketing-style studies, the user rates the product on a 5-point scale where 4 & 5 correspond to 'good' and 'very good', with the focus on the percentage of users scoring 4 or 5. For simplicity we focus in our case on an example with a truly continuous scale.

Power using binary measurements

Each user is asked just to give a yes/no answer as to whether the product is acceptable. The goal is to reject the null hypothesis “p ≤0.7” (one-sided test).

Using 100 to 110 users, the power varies between 69% and 76%. This is a bit vague, but for such binomial tests the power doesn’t always increase strictly as the sample size increases (more details below!).

Power using continuous measurements

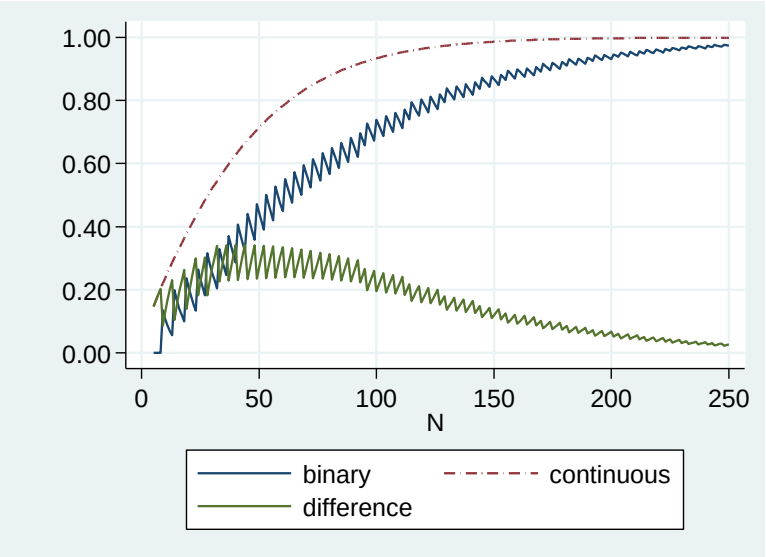
The statement “p ≥0.7” corresponds to the mean of the normal distribution lying a bit to the right of the ‘rating=6’ line. The amount is 0.52 standard deviations because, according to the normal distribution, 70% of the observations are below mean +0.52 standard deviations. We therefore have the following hypothesis test:

H0: mean is ≤ 6 + 0.52 standard deviations
H1: mean is > 6 + 0.52 standard deviations
The quantity 0.52 is in some disciplines known as the *effect size*.

The goal is to reject H0. A power calculation using a one-sided t-test gives for N=100 a power of 93% and for N=110 power=95%. This is higher than the 69-76% in the binary case.

More details on the power

In a plot, the respective powers of the binary and continuous versions are as shown in the graph below: the zigzag pattern for the binary case may be surprising. Intuitively one might think that not all p-values can occur because of the discrete nature of the test (consider outcomes of 80 of 100, 81 of 100, 81 of 101,…).



We see that the continuous case has more power than the binary case, and in our example that the difference can be as large as 30%.

Broadening the test

This is just one example of a null hypothesis and an alternative hypothesis. But how does this work out in general? To investigate, we did the same exercise for a range of combinations. The results are summarized in the next plot.

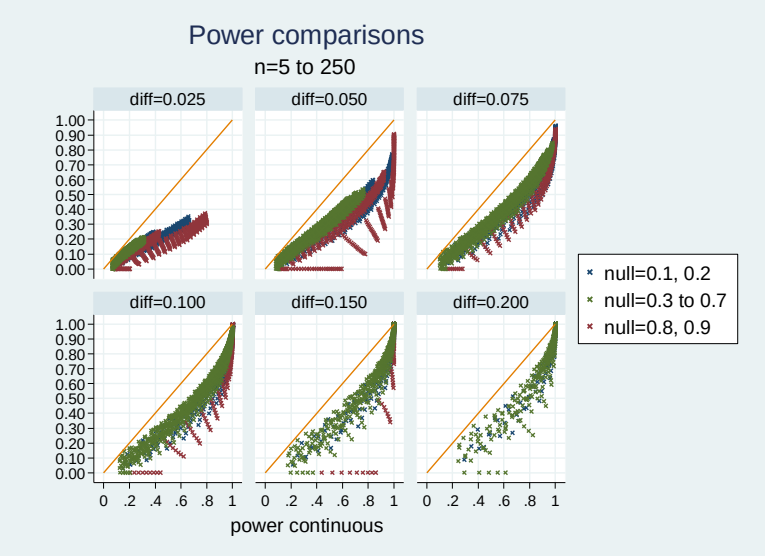
The range of situations is as follows.

- N=5 to 250 (in steps of 1)
- Null hypothesis: fractions 0.10, 0.20, 0.30,…0.90.
- Difference in fraction of null and alternative hypotheses of 0.025, 0.05, 0.075, 0.10, 0.15, 0.20 (as long as null+difference is below 1)
- One-sided hypothesis tests as in the example, for binary tests and equivalent continuous tests.

In the tables below:

- The vertical axis shows the power of the binary test, the horizontal axis shows the power of the continuous test.
- The panels correspond to the differences between the null and alternative fractions (small differences are more difficult to prove).
- The colours indicate the null hypothesis fraction; small, medium, large.
- The green symbols correspond to null fractions close to the middle 0.50; red and blue to more extreme fractions. Each symbol corresponds to a particular sample size N.

- The diagonal lines indicate equal power; the vertical distance between the point and this line is the difference in power.



The plot shows that the green points are mostly 10% to 30% below the diagonal line. The blue and red points go lower and correspond to slightly more extreme fractions.

Conclusions

The general recommendation to use continuous measurements rather than binary ones is clearly vindicated in the case of extreme fractions.

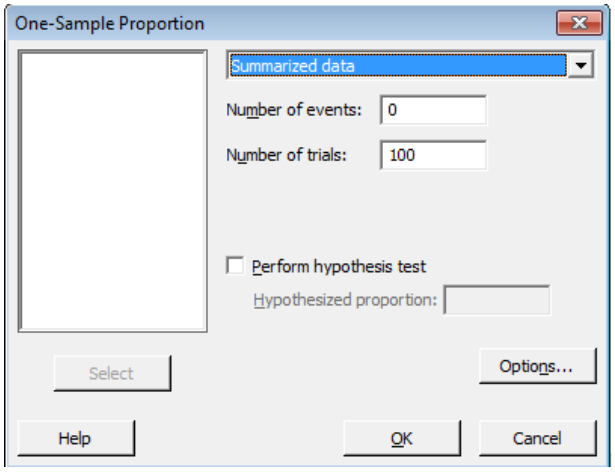
For fractions close to the middle value of 0.5, continuous measurements are still better but not enormously.

When the quality of the information is expressed using the power of a test, we see that the power is 10% to 30% lower for a large range of practical cases where the fractions are between 0.3 and 0.7.

We hope this result may help in practical situations where the pragmatic benefits of binary measurements need to be weighed against the statistical advantages of continuous measurements.

Appendix – binary measurements in Minitab

In Minitab 17, under menu Stat, Basic statistics, 1 Proportion…, you will find the following dialogue box. In this example the drop-down menu at the top was set to “Summarized data”.



The output looks like this:

Test and CI for One Proportion

```
Sample  X    N  Sample p  95% Upper Bound
1       0  100  0.000000         0.029513
```

A so-called one-sided 95% confidence interval of the true proportion is 0 to 0.030.

Note that other statistical software might give different output by default. Stata, for instance, chooses to handle the lower bound differently and reports a 97.5% confidence interval, 0 to 0.036, using the rationale that there should be 2.5% uncertainty on either side of the bound. Of course, the same results can still be obtained using different settings.

For more information, please contact CQM
T +31 40 750 23 23
info@cqm.nl