Lunchtime Webinar Series: Advanced Sampling Part 2

Applying Sampling techniques
Agenda

• Review of Part 1 Webinar
• How to apply sample size calculation for variable data
• How to apply sample size calculation for attribute data
• Summary of lessons learnt on sampling application
When we work with data in most of the cases we work with samples. However, our goal is to **draw conclusions about populations not samples.**

Therefore, we need to make sure that the **sample is representative** - properly representative of the population so that it can be used to accurately determine population characteristics.

Before we sample we need to understand what we are looking for – we call it **delta (δ)** – **critical difference**.

Whenever we make conclusions about population(s) based on sample(s) **there is the risk that our conclusion might be wrong.**
There are two types of errors we can make:

1. Detecting what we were looking for although this does not truly exist – we call it type I error or false detection. The probability of making this error is denoted as $\alpha$.

2. Failing to detect what we were looking for although this truly exist – we call it type II error or missed opportunity. The probability of making this error is denoted as $\beta$.

Beta risk can be calculated before data is collected, alpha risk can only be calculate after data is collected.
Sample size for variable data

Let’s start with simple example.

Let’s produce two samples of data:

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Distribution</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Sample size for variable data
Sample size for variable data

In the example, the statistical hypothesis are the following:

- **Null hypothesis**: There is no difference in population means
- **Alternative hypothesis**: There is a difference in population means

The P-value of 0.626 shows that we do not have enough evidence to reject the null hypothesis and conclude that means are different!

We failed to detect a difference in means although it was truly there. Why?

Let’s check power of our test......
Sample size for variable data

With a sample size of 10
We tried to detect a difference of 0.5
Power of our test is below 20%!
Sample size for variable data

Let’s check what is the sample size required to detect the difference in means of 0.5 with an alpha risk of 5% and a beta risk of 10% (power of 90%)

We need sample size of 86.
Sample size for variable data

If we have budget to collect 86 data points from each sample – then the problem is solved. But if we don’t? What shall we do? Let’s change some of the conditions...

Reducing alpha risk did not work – we need to have an even bigger sample.
Sample size for variable data

Increasing alpha risk above 5% is usually risky. What else we can do?

Let’s try to increase beta risk.

We need less data (64 versus 86)
Sample size for variable data

What else we can do?

Let’s try to increase delta – the minimum difference that we want to reliably detect.

We need even less data.
Sample size for variable data

As we see in the examples, sample size determination is usually a question of finding a compromise between the:

• Amount of data we can afford to collect
• Acceptable alpha and beta risk
• Difference that we are looking for

Let’s imagine that in our case we have a budget to collect 40 data points in each sample.

*What are our options?*
Sample size for variable data

We can either try to compromise on power...

Then we have a 60% chance to detect a difference of 0.5.
Sample size for variable data

...or compromise on critical difference...

Then we can reliably detect a difference of about 0.7.
Sample size for variable data

... or we can find a balance and compromise on both.

With power of 80% we can detect a difference of a bit more than 0.6.
Conclusions: Sample size for variable data

If we want to compromise on alpha risk it will add even more complexity.

We apply the same logic for other statistical tests.

The only thing that we have to know for means testing is standard deviation.

If we can not estimate the standard deviation of our data then we usually have to collect pre-sample just to estimate it.
Let’s again start with a simple example.

Let’s produce four samples of data:

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Distribution</td>
<td>Binomial</td>
<td>Binomial</td>
<td>Binomial</td>
<td>Binomial</td>
</tr>
<tr>
<td>Number of trials</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Event probability</td>
<td>0.5</td>
<td>0.45</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Sample size for attribute data

The business case is the following:

*We want to find out if proportion of complete document sets prepared is the same for experienced and new employees.*

**Null hypothesis:** The proportions are the same

**Alternative hypothesis:** The proportions are not the same

Let’s do 2 proportion tests for the two scenarios (column C1 and C2 are for the first scenario and columns C3 and C4 are for the second scenario)
Sample size for attribute data

We failed to detect difference!
Sample size for attribute data

Let’s check the power of our tests...
Sample size for attribute data

**Conclusion:** In both scenarios our power was too low to detect the difference we were looking for.

Although in both scenarios we were trying to find the difference of 0.05 we had a different power for the difference of 0.05 between 0.5 and 0.45, and the difference of 0.05 between 0.1 and 0.05

Most likely this will be reflected in sample sizes required to detect this difference in both scenarios.

Let’s see...
Sample size for attribute data
Conclusions:
Sample size for attribute data

As we see in both scenarios the sample sizes required are very different.

**Conclusion:** it is not that important that we are looking for a difference of 0.05, it is much more important if we are looking for a difference between 0.5 and 0.45, or between 0.1 and 0.05.

As for variable data, we can try to find a balance between alpha and beta risk, and the difference we are looking for.

In both scenarios we need to have very large samples (hundreds) to detect the difference that we are looking for.
Summary

Alpha, beta and delta have an impact on sample size. Make sure you think them carefully before you start collecting data.

Sample size calculation is done for particular statistical tests you want to apply.

The sample size needed for attribute data is usually much larger than for variable data.

How to deal with the situation that your calculated sample size is so large that it is impossible or too difficult to collect enough data will be covered in next webinar.
Learn more!

**Courses & Workshops**

Learn more statistical analysis tools in the following courses:

- Six Sigma Green Belt
- Six Sigma Black Belt
- Tool Master (advanced statistical analysis tools)

Full Training Schedule: [http://www.bmgi.com/training](http://www.bmgi.com/training)

**Free Tools, Templates & eLearning**

Visit our Open Access Website for more materials and free learning on Lean Six Sigma, Strategy execution, Change & Innovation.

Open Access: [http://www.bmgi.org](http://www.bmgi.org)
Questions & Answers