CÁLCULO DA DIMENSÃO DA AMOSTRA – PARTE I

ricardo.anselmo.castro@tecnico.ulisboa.pt

ABSTRACT

O artigo pretende operacionalizar os conceitos do cálculo da dimensão de uma amostra, mediante dois cenários distintos: o primeiro de quando existe algum conhecimento do processo (dados históricos), e o segundo de quando não existem quaisquer registos.

Palavras-chave: erro amostral, média, desvio-padrão.

PROBLEMA

No decorrer de um projeto Six Sigma, o líder de projeto depara-se, mais cedo ou mais tarde, com a pergunta: «qual a dimensão necessária da amostra, de modo a ter um determinado nível de confiança e exatidão pretendido?". Esta necessidade pode surgir tanto durante a fase measure, como nas fases analyse ou improve.

A resposta à pergunta é sempre um compromisso entre o esforço de recolher uma quantidade elevada de dados e o ganho que daí decorrerá para o projeto. Assume-se nesta altura que já se procedeu à validação do sistema de medição e que se encontrou um modo de recolher os dados de forma aleatória, isto para que a representatividade da amostra esteja de certo modo protegida. Mas, em função da natureza do projeto e do cumprimento de prazos para a sua conclusão, haverá estratégias mais adequadas do que outras. Vamos assim calcular a dimensão da amostra para a *baseline* de dois processos distintos: 1) indústria, onde se vai trabalhar com dados históricos, e 2) serviços, onde se vai trabalhar com novos dados.

DIREÇÃO DA SOLUÇÃO

1) Indústria, com dados históricos

Considere-se uma empresa onde um dos seus objetivos é o de reduzir os custos energéticos numa secção de transformação de matérias-primas. Foi avaliado os custos por equipamento e chegou-se facilmente à conclusão que o Moinho 6 é aquele que mais pesa nos custos anuais. O Y do negócio é euros, e o y do projeto é KWh. Naturalmente que existe uma relação (muito direta) entre o y e o Y. Pretende-se caraterizar o processo, ou seja, calcular a baseline do y. Como fazer?

Sendo o consumo em KWh uma variável contínua, estaremos interessados em calcular uma tendência central e uma dispersão do processo. Tipicamente, uma média e um desvio-padrão. Existem cerca de 20 registos em histórico (KWh consumido por cada lote produzido). Vamos desenhar os dados para entender que tipo de distribuição se trata e se o processo está minimamente estabilizado (o sistema de medição foi validado e por isso podemos confiar nos dados):

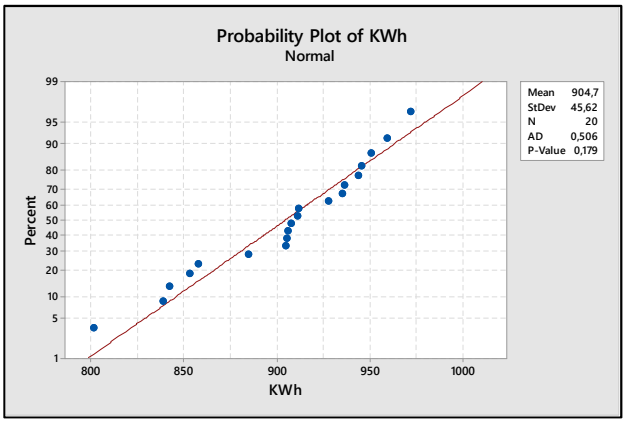

Fig. 1: Normal probability plot, com um p-value de 0,179.

Os dados seguem aproximadamente uma distribuição normal e são estatisticamente previsíveis (não há causas especiais presentes no processo). Se queremos estimar o valor real da média de KWh consumido por lote precisamos, antes de mais, estimar o desvio-padrão.

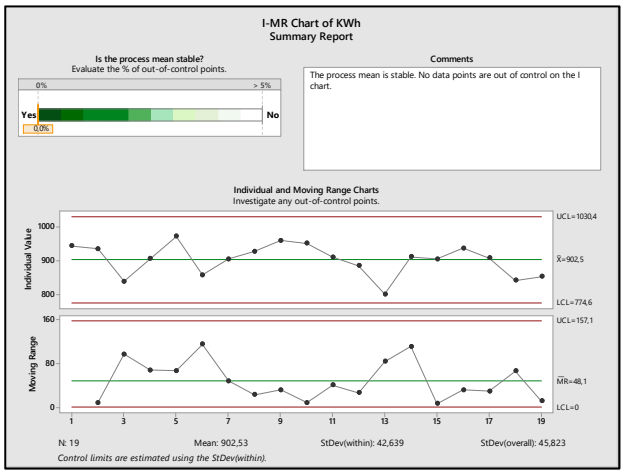

Fig. 2: SPC I-MR dos 20 registos ao longo do tempo.

Estimativa do desvio-padrão

Será que os 20 registos de histórico são suficientes para se estimar o valor do desvio-padrão, com o nível de confiança e exatidão pretendidos? A resposta pode estar na figura 3: sempre que se adiciona mais um registo observa-se o impacto que o mesmo terá na nova estimativa do desvio padrão. Se a variabilidade aumentar de modo significativo, então precisaremos de mais amostras. Se pelo contrário o cálculo da estimativa do desvio-padrão convergir, significa que estamos cada vez mais perto do seu valor real.

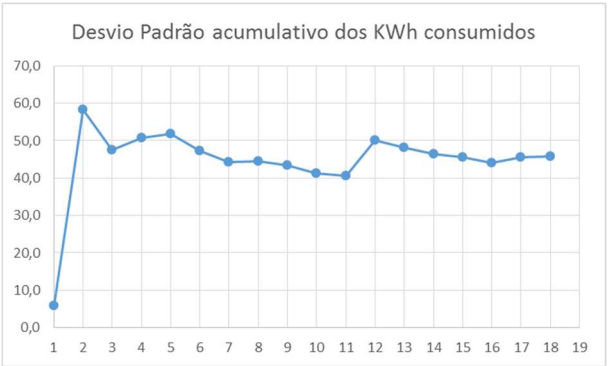

Fig. 3: Evolução da estimativa do desvio-padrão ao longo do tempo (sempre que se adiciona mais uma amostra).

Vemos que a partir da amostra 13, a estimativa parece começar a convergir e, como tal, será adequado utilizar o valor 45KWh como a melhor estimativa que temos do desvio-padrão, relativamente a este processo.

Estimativa da média

Por razões de negócio, sabe-se que não se quer ficar a mais de 36 KWh (cerca de 5€) de distância do valor real da população. Esta distância tem que ver com o erro amostral. Para um nível de confiança de 95% observa-se que é necessário recolher 9 amostras.

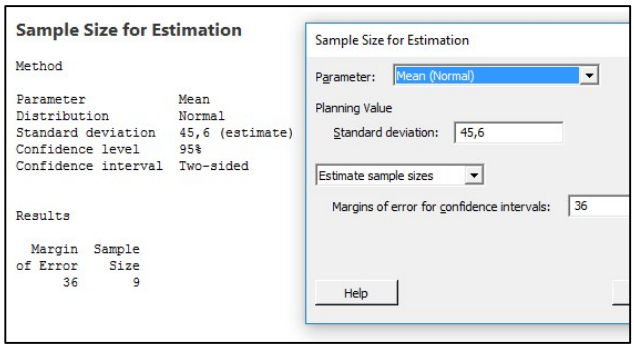

Fig. 4: Dimensão da amostra necessária para um delta de 36KWh e um nível de confiança de 95%.

O Minitab indica-nos a necessidade de recolher 9 amostras, mas como já tínhamos 20 registos, não precisamos de recolher mais informação. Conclui-se então que as figuras 2 e 5 caraterizam a *baseline* deste processo (μ = 902KWh e σ = 46KWh).

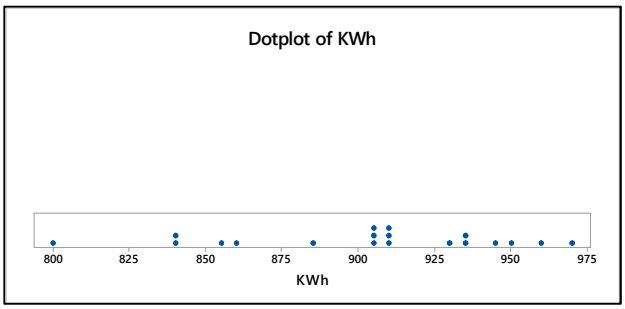

Fig. 5: Dotplot do consumo energético do moinho 6, em KWh.

* Vale ainda a pena referir o seguinte: imagine-se um processo onde se registam os valores mensais de uma determinada caraterística, ao longo de dois anos (dados históricos). Sobre esses valores calculam-se os limites de controlo e se conclui, admitamos, que os dados são estatisticamente previsíveis. Pode-se perguntar: serão 24 (meses) observações suficientes para se considerar que a amostra é representativa, sem fazer qualquer cálculo? Poderemos usar sem preocupações a média e o

desvio-padrão dessas 24 observações? Se confiamos na fiabilidade do sistema de medição e se no curto prazo não existem alterações previstas ao processo, a resposta é sim. Repare-se que se deixou passar tempo suficiente sobre o processo – um horizonte temporal de dois anos – o que parece ser um prazo mais do que razoável para se trabalhar com o valor médio e o desvio-padrão dessas 24 observações. A questão: «qual a dimensão mínima da amostra?» deixa de fazer sentido. A representatividade está assegurada e as observações passadas são suficientes para inferir sobre os parâmetros da população.

2) Serviços, com novos dados

Imagine-se agora a situação onde não existem dados passados. Nesta situação, a equipa não tem outra alternativa senão proceder a um plano de recolha de (novos) dados.

Estimativa do desvio-padrão

No que diz respeito a esta estimativa, a equipa até poderá desenvolver o mesmo raciocínio apresentado na figura 3, mas para o processo em questão, pode ser que o número de amostras seja demasiado elevado, até que os valores comecem a convergir. Para se evitarem surpresas desagradáveis sugere-se a seguinte boa prática: se não existe a mais pequena ideia de qual será o valor do desvio-padrão, uma forma para o podermos estimar é dizer que:

$$\hat{\sigma} = R/5$$

R representa a amplitude do processo, ou seja, qual o valor mínimo e máximo que já foi observado neste processo. Para se evitarem casos extremos é conveniente que os valores mínimos e máximos se refiram ao percentil $p_{0,01}$ e $p_{0,99}$ respetivamente. Numa distribuição normal perfeita 1/6 da amplitude corresponde ao próprio valor do desvio-padrão. Como «coeficiente de segurança» é normal dividir-se a amplitude não por 6, mas por 5. Considere, por hipótese, os números seguintes:

$$\hat{\sigma} = (120 - 10)/5 = 22$$

Em função da cadência do processo, 22 pode ser um número pequeno ou grande. Se pequeno, estamos em condições de começar a desenhar o gráfico 3 e de verificar se o valor do desvio-padrão converge. Toda a análise subsequente será idêntica à apresentada para o estudo de caso na indústria.

Contudo, se 22 for considerado um número grande, pelo tempo que teríamos de deixar decorrer (imagine-se a cadência nas vendas do automóvel Ferrari, numa cidade como Lisboa) deverão ser equacionadas outras alternativas. Naturalmente que a Ferrari bem conhecerá o seu volume de vendas mensal, mas para efeitos de raciocínio imagine-se que tal não era conhecido. Uma outra métrica mais útil e de recolha de dados mais rápida para o projeto (considerando que o objetivo deste seria ajudar a aumentar as vendas) poderia ser o número de visitantes por dia, nos concessionários Ferrari de uma região. Sem visitas não há vendas e se as visitas aumentarem, provavelmente haverá mais negócios a serem fechados. Muito provavelmente, a estimativa do desvio-padrão para o número de visitas diárias será obtida mais rapidamente.

CONCLUSÃO

A dimensão da amostra é um tema que é muitas vezes levantado pelo experimentador, mas serão poucas as situações que realmente se parte para o seu cálculo. Neste artigo mostrou-se uma alternativa de como estimar o desvio padrão, seja pelas amplitudes, seja pelo método de convergência. Este cálculo é importante, pois permitirá definir a dimensão da amostra mínima, para um determinado intervalo de confiança e nível de exatidão. Pretende-se que esta informação possa ser usada em termos práticos pelo praticante de projetos de melhoria contínua, mais especificamente o Black Belt.

REFERÊNCIAS

[1] Castro, Ricardo A. (2012) *Lean Six Sigma – para qualquer negócio*. 3.ª edição, IST Press.

* Este artigo contou com a colaborção do aluno de LSSBB, 5.ª edição (IST) Nuno Ferreira, que se disponibilizou a fornecer os dados apresentados.

Ricardo Anselmo de Castro é coordenador do Programa de Especialização de Lean Six Sigma Black Belt, do Instituto Superior Técnico, e tem dois livros publicados na mesma área.

doctorflow.net
https://tecnicomais.pt/diploma-de-formacao-avancada/lean-six-sigma-black-belt