

Sample size done differently

Ruud van Lieshout BEng MSc CIRM

“... we need at least 80 subjects in our sample otherwise the sample won't be representative...”! This is a remark I often hear while designing user studies. I've always wondered why 80? When I ask for details about this the only thing I hear is that it has been a way of working for many years. I think that what they say is based on the fact that they like to encompass most of the population variation in their sample. I checked with my customer this morning and that is exactly why they say 80 subjects, because they like to see the variation between subjects in their sample as well. But can we understand this from a statistical point of view? Well, I think tolerance intervals can provide an answer.

The definition is : Let $L < U$ be two statistics i.e., quantities calculated from the data. Then $[L, U]$ is called a $100\beta\%$ tolerance interval at confidence level $100(1-\alpha)\%$ if $\Pr(F(U)-F(L) \geq \beta) \geq 1-\alpha$, or if, with high probability, at least a given large part of the distribution will be enclosed between L and U . Typical values for α and β are $\alpha=0.05$ and $\beta=0.95$.

An example of a 95% tolerance interval at confidence level 95% assuming a normal distribution, sample mean=10, sample standard deviation=1, sample size $n=100$ can be found to be [7.766, 12.234] using Minitab.

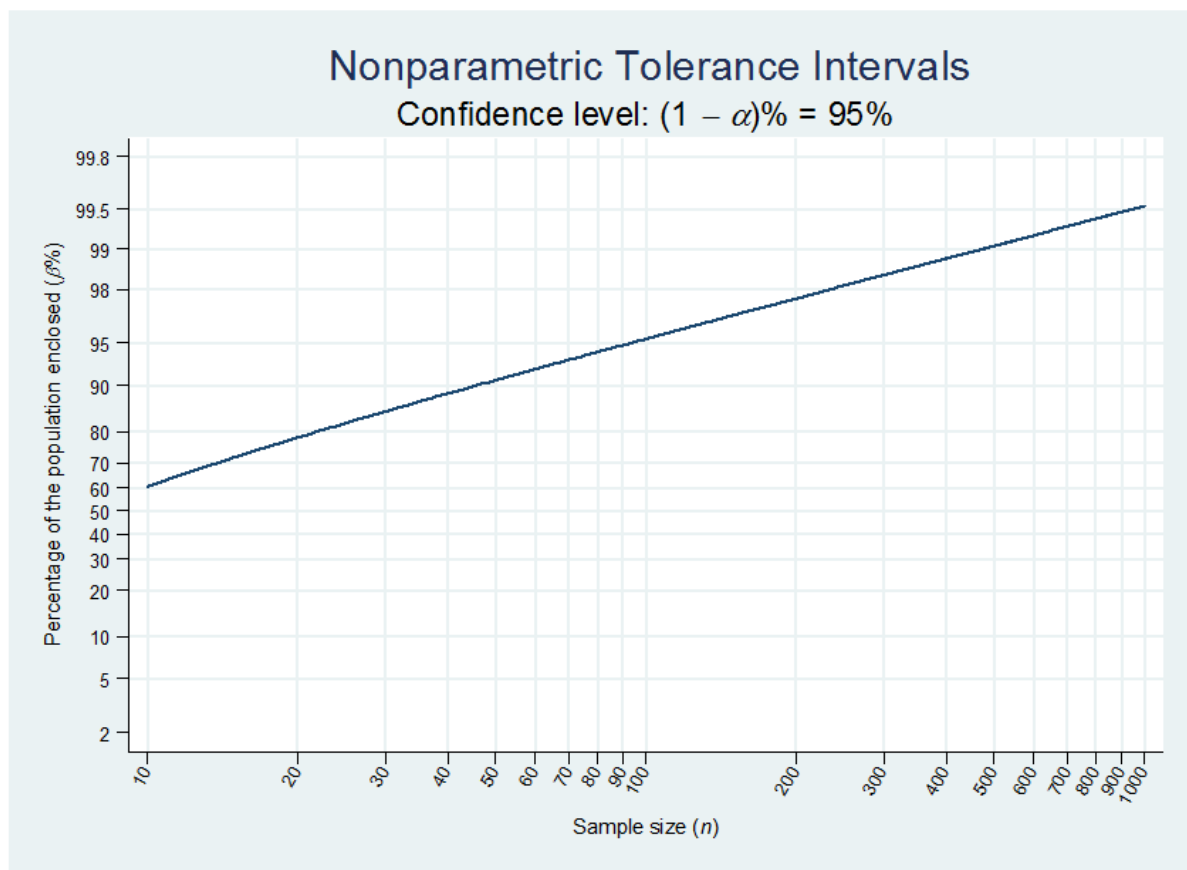
Tolerance intervals however come in two flavours i.e. parametric, like in the example above where we have assumed normality, and non-parametric ones where no specific distribution is assumed. Let's focus on non-parametric tolerance intervals. Say, we take L to be the minimum of the sample and U its sample maximum. We would like see that between this sample minimum and sample maximum a large part of the population is located because if that is so then we will almost include the entire population variation. The question now is how large should my sample be to make this happen? That is, to be able to state that the interval made up from this sample minimum and the sample maximum contains, with high probability (95%, say), at least 95% of the population. Using sample minimum and sample maximum the following relation¹ holds between α , β and sample size n .

This can be solved iteratively but a good approximation can be found here² and be written as:

The solution is shown in the graph below. From this graph it follows that if $\alpha=1-0.95=0.05$ en $\beta=0.95$ roughly a sample size of $n=90$ is needed. This is rather close to the 80 subjects from the rule of thumb.

¹ Mood A. and F. Graybill, Introduction to the theory of statistics, pp 515-516

² <http://www.itl.nist.gov/div898/handbook/prc/section2/prc255.htm>



For more information, please contact CQM

T +31 40 750 23 23

info@cqm.nl

Copyright © 2015 by CQM B.V.

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, without the prior written permission of the publisher.
