Ask a statistician:

Help! Is my sample size big enough?

Anne Beeston writes: **I am planning a study and need some advice on sample size to allow the determination of statistically significant differences in two groups.**

Deirdre Toher writes: Our reader asked about a design with two independent groups of flowers, Specifically, she wants to know if her proposed design is likely to detect a statistically significant difference in microbial load between the group that receives the treatment and the group that does not (called the "control" group). Statistical significance is typically determined by a *p*-value of less than 0.05, which is the probability of collecting the observed data, or something more extreme, assuming the null hypothesis is true. The null hypothesis here would be a version of "no difference between the treatment and control groups", with the definition of the difference related to the statistical assumptions made and the associated statistical test chosen.

Irrespective of the proposed study design, your friendly statistician will need to know the plausible ranges of values of the variation in your data before determining the ideal sample size. This is one reason why pilot studies are conducted – to learn more about the potential spread of data. We also need to know what size of a difference between your groups is considered interesting. If you are looking for a difference of 0.2 then your sample size would need to be much larger than if, for the same type of data, a difference of greater than 2 was of interest.

Sample size calculations require a number of different elements to be specified: the power (the probability of correctly rejecting the null hypothesis); and the maximum acceptable Type I error rate (the probability of the incorrectly rejecting the null hypothesis). These can vary depending on the specific context of the study and the question of interest.

Her proposal is, for both groups: (1) to collect 40 dried flowers; (2) to grind these up and mix together to hopefully achieve a homogenised sample; and (3) obtain three 5-gram samples from the ground mixture. Using these three samples, the microbial load of the dried flowers will then be measured, and the treatment and control groups compared.

The plan to mix together the ground-up flowers is because previous testing demonstrated that the variability between flowers is very high. But by grinding and mixing the researcher stands to lose important insights about variability. Unless the flowers will be ground up in the end application, then a rethink is required.

However, sticking with the plan as originally described, the reader is proposing what could, at best, be considered a sample size of six. Putting aside the debate about whether *p*-values are appropriate or otherwise (see page 38 of the April 2017 issue for more on that), is it possible to achieve a statistically significant result with only three observations in each group?

An important concept in statistics is that making assumptions about the distribution of our data increases statistical power – basically adding information to the problem, much like increasing the size of our dataset. The assumption that the data follow the normal distribution would allow us to use the more powerful independent samples *t*-test. But without being able to refer back to earlier studies, we cannot assess normality based on a sample size so small. With only three observations, we can't visually assess symmetry let alone identify outliers. Instead, we may opt for the Mann-Whitney (M-W) test as this uses only the ranks of the observations and does not require the same assumptions about the distribution shape as a *t*-test does.

But again, we encounter issues: even if all the observations of the microbial load in the treatment group are less than the smallest observation in the control group, the resulting Mann-Whitney test statistic is *U*=0, *p*=.100 (see box for calculation using hypothetical data). With three samples in each group, given the most extreme situation possible, it is impossible to find a statistically significant result.

How big a sample size would be big enough? Well that largely depends on the degree of separation between the two groups and whether our reader is interested in quantifying the size of the difference rather than just the presence or absence of a difference.

Mann Whitney Test Statistic calculation using hypothetical extreme data:

Microbial Load	Group	Rank of value	Sum of Ranks
125	Treatment	1	6
154	Treatment	3	
145	Treatment	2	
175	Control	5	15
185	Control	6	
164	Control	4	

When calculating the Mann-Whitney test statistic the actual values of the microbial load do not matter, only the relative ranks matter.

$$U = \min(U_1, U_2)$$

Where

$$U_1 = n_1 \times n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 3 \times 3 + \frac{3 \times 4}{2} - (1+2+3) = 9 + 6 - 6 = 9$$

$$U_2 = n_1 \times n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 3 \times 3 + \frac{3 \times 4}{2} = 9 + 6 - 15 = 0$$

So $U = \min(9,0) = 0$