Charge questions 1.2 and 1.3 – Comments by Richard Smith

I'd like to comment on a few points that were raised on day 1 where I felt the discussion we had was incomplete.

Charge Question 1.2: two statistical points here, about quantile tests and retrospective power analysis (I'll take them in that order).

Quantile test: in fact the description on pages 8-32 and 8-33 is perfectly clear, but you have to refer to the appendix for tables I.7 through I.10, and then back to the main text for the principal reference, which is Gilbert (1987). I know Gilbert's book, but I don't have it on my bookshelf, so I'd have to take it on trust that what he says is correct. However, I don't think the tables themselves are particularly complicated – I'd expect a user without comprehensive statistical training to be able to apply this procedure.

I recall there were some comments about the power of this test. I don't see anything here about power of the test, but I believe it would be possible to construct such tables for an alternative hypothesis of the following structure: given that the test quantile is exceeded with probability α in the reference (background) sample, what is the probability that the test would detect an exceedance probability of Kα in the survey sample, for some given K>1? I don't think it would be difficult to construct a table to do that. At worst, you could do a simulation.

I think a more relevant question for this committee might be "when should you use this test?" (and if so how, e.g. what value of α?). The question we're being asked is whether it's "technically appropriate and discussed adequately and clearly?" With the caveat I just made, I'd give a yes to "discussed adequately and clearly," but the question about "technically appropriate" doesn't have a simple yes or no answer – the question is not whether the test should be used, but when and how.

I have a similar comment about "retrospective power analysis". First, let me point out that there is no section headed "retrospective power analysis" anywhere in the entire document, so I don't know what procedure we're being asked to evaluate. However, there is a discussion of it on pages 8-47 and 8-48. Let me summarize this as follows. Suppose I conduct a test that the mean level of a contaminant is x against the alternative that it is >x, and suppose the result of the test is to accept the null hypothesis. I can then ask: what would be level y>x at which the test would have rejected the null hypothesis at a particular power? Then ask yourself whether you would be satisfied with a contamination level of y. If y is still at a level you'd consider safe, all well and good. But if it's above that level, you need to go back and do a new test with greater sensitivity.

That's what I understand by the phrase "retrospective power analysis". But again, that's not what the CQ asks. "Is the inclusion and proposed implementation … technically appropriate…?" But the manual doesn't tell us anything about when or how we are supposed to use this procedure, or how to interpret the results, so the question doesn't seem to be well-posed.

Perhaps I could throw in another point here, one that might perhaps be more amenable to committee discussion. It's typical, in power calculations, to work toward a power in the region of 80-90%. That could be quite acceptable when the calculation is intended, for example, to guide the sample size for a clinical trial. But if we are using this for classifying a potential nuclear waste site, shouldn't the power be much higher, something like 99.9%? With such an eminent group of experts in our zoom-room, maybe that's the sort of point we could usefully discuss.

Charge Question 3. Since I was having a hard time following the afternoon's discussion, I went back to it afterwards and tried to draw my own conclusions.

Let me just focus on the part that reads "Is the proposed calculation of measurement uncertainty consistent with the concept of Measurement Quality Objectives?" The first thing I did was to look up where in the document there is any discussion of "measurement uncertainty", and I found that the only section that actually described this concept is Section 6.4 (not one of the sections cited in the charge question). I then searched through Section 6.4 to see whether there was any reference to "Measurement Quality Objectives", but there isn't. Therefore, my answer to the question is "no". It didn't require any statistical expertise for me to answer that question, just simple use of a text editor.

However, let me go a little bit more into what Section 6.4 actually says. After a little preliminary discussion, 6.4.1 is about "Systematic and Random Uncertainties." It's important that the text makes this distinction, since beginners in statistics often ignore the "systematic" part altogether. However, the test doesn't say how one should distinguish these two types of uncertainty in practice. Therefore, I don't find this very helpful.

A common analogy made in elementary statistics texts is the "bull's eye". Suppose you are shooting arrows at a target. One shooter may consistently aim at the correct target, but with so much variability that the arrows land all over the place. That is an example of high variability but low bias. Another shooter may always shoot her arrows in the same place, but it isn't the correct position on the target. That's an example of low variability but high bias. Perhaps some illustration along these lines would help the user understand the distinction.

The next section 6.4.2 is about "statistical counting uncertainty". I stared at formula (6-17) and wondered what on earth was going on. Then I remembered someone during the meeting said something about Poisson processes, and the penny dropped. If X is the number of counts in time t from a Poisson process with rate λ, then a suitable estimator of λ is X/t, and its variance is $\mathrm{Var}(X)/t^2=\lambda t/t^2=\lambda/t$. Substitute λ by its estimator, and we get X/t^2 as the estimated variance of the estimate. Do that twice, once for the gross counts and again for the background, then the variance of the estimated difference is the sum of the variances, and its standard deviation (σ_n) is the square root of that. That makes sense to me, but how is someone without any statistical knowledge supposed to figure that out? And if they are not given any explanation why or how the formula works, how will they know when to use it?

Now let's go on to formula (6-18), about which there was quite a bit of discussion during the meeting, apparently centering on whether it was reasonable to expect engineers to know calculus. Let me pass over that point and comment on the formula itself. In fact this is a well-known formula in statistics, commonly known as the “delta method”. There are two key features of its applicability: (i) the function y being evaluated is sufficiently close to linear in its range of uncertainty that a first-order Taylor expansion is an adequate approximation, (ii) that the error terms are independent, or at least uncorrelated. In fact there are ways around (ii) if you care about that point, but the approximate local linearity of the function is essential to the validity of the formula. That’s the calculus point the engineer needs to understand, not the meaning of $\partial y/\partial x$ (which, in most specific instances where this formula is applied, can be simplified to something explicit).

My point about these formulas is this: as a professional statistician, I can back-calculate from the formula and figure out what the writer really meant. But what is someone with no statistical training supposed to do? That’s why I feel the manual needs to be much more explicit about when these formulas are applicable, so as to guard against their inappropriate use by someone who has no idea what is going on.

Finally, let me comment on the “confidence interval” discussion (section 6.4.4). The very first sentence here says “measurement uncertainty is used interchangeably with the term standard deviation” but surely this contradicts section 6.4.1, where the very careful distinction between systematic and random uncertainty is made (standard deviation is only applicable to the latter, not the former). Apart from that, the discussion in this section is very low-level. I teach confidence intervals in my freshman-level course at UNC, but even at that level I go into far more detail than this, for example showing how the normal probability table is used and discussing when it’s appropriate to switch to a t-table. I honestly don’t see how this is useful to a professional engineer!