The Neyman-Scott 'Paradox'

The original statement of this so-called paradox can be found in J. Neyman and E.L. Scott, *Econometrica*, **16**, 1 (1948) so it's hardly new, but I only came across it recently and thought I'd share it here.

Consider pairs of measurements (x_i, y_i) where $i = 1...N$. These are modelled as independent and identically distributed normal random variables, i.e.

$$x_i \sim N(\mu_i, \sigma^2) \;\; y \sim N(\mu_i, \sigma^2).$$

In other words each pair (x_i, y_i) has a different (unknown) mean μ_i but the pairs all have a common variance σ^2. Our task is to estimate σ^2 from N pairs of measurements treating the μ_i as nuisance parameters. This may seem a slightly weird problem, but it's not really, it's just that each pair of measurements is subject to an unknown systematic offset.

First let's try a Maximum Likelihood Estimator. The likelihood is just the probability of the data given a specific model, which is easy to construct:

$$L = P(\{x_i, y_i\}\,|\mu_i, \sigma^2),$$

so that

$$L = \frac{1}{(2\pi)^N \sigma^{2N}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right]\right\}.$$

The Maximum Likelihood Estimator is the choice of model parameters that maximizes the probability of getting the observed data. If that sounds like a strange criterion to adopt, that's because it is. Nevertheless we continue by differentiating the above with respect to the parameters σ^2 and μ_i and requiring the derivative to be zero to find the maximum. This task, as usual, is simplified by first taking the log of the likelihood and setting derivatives of that to zero instead:

$$\ln L = -N\ln(2\pi) - N\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right].$$

From this we get

$$\frac{\partial \ln L}{\partial \mu_i} = \frac{x_i + y_i - 2\mu_i}{\sigma^2} = 0,$$

and

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\left[(x_i-\mu_i)^2 + (y_i-\mu_i)^2\right] = 0.$$

After a little bit of work solving these simultaneous equations we find that the MLE estimators are, for the variance,

$$\hat{\sigma^2} = \frac{1}{4N}\sum_{i=1}^{N}(x_i - y_i)^2,$$

and for the means,

$$\hat{\mu_i} = \frac{1}{2}(x_i + y_i).$$

The estimator for the means is entirely in accord with intuition, but we can immediately see a problem with the estimator for the variance. Regardless of how many data pairs we have, i.e. no matter how large the value of N, the expectation value of the estimator of the variance is actually $\sigma^2/2$ i.e. off by a factor two from the 'true' answer!

Now let's look at a Bayesian way to approach this problem. This does not maximize the likelihood but instead maximizes the *posterior probability* (i.e. the probability of the parameters given the data) instead. In this case we can choose a uniform (i.e. noninformative) prior on the μ_i and whatever we like as a prior on σ^2, say $P(\sigma^2)$. The posterior probability then just takes the form

$$P\left(\mu_i, \sigma | \left\{x_i, y_i\right\}\right) \propto L \times P(\sigma^2)$$

I'll leave the details as an exercise to the reader, but marginalizing over the nuisance parameters μ_i means constructing

$$P\left(\sigma | \left\{x_i, y_i\right\}\right) = \int \int \cdots \int P\left(\mu_i, \sigma | \left\{x_i, y_i\right\}\right) d\mu_1 d\mu_2 \ldots d\mu_N,$$

which turns out to be

$$P(\sigma^2 | \left\{x_i, y_i\right\}) = \prod \frac{1}{\sigma\sqrt{\pi}} \exp\left[-\frac{(x_i - y_i)^2}{4\sigma^2}\right] \times P(\sigma^2)$$

Taking logs gives us:

$$\ln P(\sigma^2 | \left\{x_i, y_i\right\}) = \mathrm{constant} - \frac{\mathrm{N}}{\sigma} + \frac{1}{4\sigma^4} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left[(\mathrm{x_i} - \mathrm{y_i})^2\right] + \ln \mathrm{P}(\sigma^2).$$

This of course depends on the prior through the last term, but note that for sufficiently large N this term becomes irrelevant as the previous two terms increase with N whereas the prior-dependent term doesn't. Assuming we have a lot of measurements we can thus ignore the prior term; remember that in the MLE case the value of N was irrelevant.

This Bayesian analysis gives the posterior probability distribution of the parameter required, not just and estimate of it, but to get something to compare with the previous result we can maximize the log of the posterior probability:

$$\hat{\sigma^2} = \frac{1}{2N} \sum_{i=1}^{N} (x_i - y_i)^2.$$

Lo and behold the troublesome factor of two that appeared in the MLE estimator has vanished!

Like most paradoxes the Neyman-Scott paradox isn't really a paradox at all. It's just a demonstration that Bayesian reasoning (when done correctly) is consistent because it uses only the laws of probability, which is something other approaches cannot guarantee. Maximum likelihood estimators do give sensible answers in many situations, but why take the chance?