Estimating error from data: one page of bliss

Suppose we have a quantity y that can be measured (e.g., a voltage), but we do not know anything about the "true" value y_{true} or uncertainty σ_y on a given measurement. The following is a reliable approach to estimating both.

If you take N identical measurements y_n, the mean is

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

No surprises there. We typically assume each of the individual measurements y_n is drawn from a distribution having standard deviation σ_y about the "true" value y_{true}, but (again) know neither of these ahead of time. However, sticking the symbol in as placeholders for now, we can propagate the error on the formula for the mean (reacall that, for a sum, uncertainties add in quadrature), finding

$$\begin{aligned}\sigma_{\bar{y}} &= \frac{1}{N}\sqrt{\sum_{n=1}^{N}\sigma_y^2} \\ &= \frac{1}{N}\sqrt{N\sigma_y^2} \\ &= \frac{\sigma_y}{\sqrt{N}}.\end{aligned}$$

Notice the assumption that all σ_y's are identical. More importantly, notice the key difference between $\sigma_{\bar{y}}$ and the per-point uncertainty σ_y. The more we average, the less uncertainty we have, and this improves as $\sqrt{N}$ (or $\sqrt{\text{your patience}}$). Importantly, the error bar on the *mean* is *not* just the standard deviation of the data set, a fact *often* gotten wrong.

We still need an estimate of the per-point uncertainty σ_y, though. If we knew y_{true}, the best estimate would be from "the usual" variance

$$\sigma_y^2 \approx \frac{1}{N}\sum_{n=1}^{N}\left(y_n - y_{true}\right)^2.$$

However, we only have our measured value $\bar{y}$, *not* y_{true}, and, if you think about it (see below), $\bar{y}$ is actually behaving like a "fit parameter" to this particular data set, meaning we have $N-1$ degrees of freedom, and a reduced chi squared

$$\chi_r^2 = \frac{1}{N-1}\sum_{n=1}^{N}\frac{\left(y_n - \bar{y}\right)^2}{\sigma_y^2}.$$

Since we know the *average* value of χ_r^2 should be 1, we can use this to make a reasonable estimate of σ_y^2 by setting $\chi_r^2 = 1$ and solving for σ_y :

$$\sigma_y \approx \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(y_n - \bar{y}\right)^2}.$$

Weird? It's $N-1$ in the denominator, not N.[1] The reason I use ≈ here is because you have a finite number of data points – if you took another set of N points, σ_y would have a different value, because all the measured values would be different. Stated briefly, there is uncertainty on the uncertainty, so make N big enough to not worry about this! Take a look at the "Chi2 Distribution Playtime.py" script to get an intuition for how many points you should take to estimate σ_y and how much you can trust this estimate (edit the value of "DOF").

[1] If you have data that drifts, you can also calculate the line of best fit and substitute that in for $\bar{y}$, using $N-2$ degrees of freedom (two fit parameters) instead of $N-1$.

Practice Time

1. Suppose we have N measurements y_n as discussed above, and we fit the data to a constant. Minimize χ^2 (analytically) to find the fit value. Do you recognize the result?

2. The file "fake-data.txt" contains two columns of data: measurement number and measured value (y_n), drawn from a Gaussian distribution of "hidden" width.

   (a) Use a fitter to fit this data to a constant (you can use "Fit.py"). Does the fit value and uncertainty match the independently calculated mean $\bar{y}$ and uncertainty $\sigma_{\bar{y}}$ for the data set?

   (b) Adjust the assumed uncertainty until $\chi_r^2 = 1$. What value achieves this? Is it close to the "true" value of 7? How do you quantify "close"? Hint: play with the "Chi2 Distribution Playtime.py"

   (c) Try including a slope in your model – perhaps something was drifting with time while you performed your measurements – and do the same. Is *this* estimate "far" from 7?

   (d) Open the text file and look at the "reality" entry, which shows the function to which noise was added. Comparing with (b) and (c)'s results.