Math and General Physics Question 22

Let's describe our measurements as $\{x_i\}$ with uncertainty $\{\sigma_i\}$.

The mean:

$$\bar{x} = \frac{1}{N}\sum_i x_i \tag{1}$$

$$\bar{x} = \frac{\sum_{i=1}^N x_i \sigma_i^{-2}}{\sum_{i=1}^N \sigma_i^{-2}} \tag{2}$$

The median:

$$\mathrm{med}(x) = x_{N/2} \tag{3}$$

The standard deviation:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2} \tag{4}$$

Outliers: can decide these by looking at how many standard deviations away from the mean - 1 away is 68 percent of data, 2 away is 95.5 percent of data, 3 away is 99.7 percent of data. However, this doesn't work if there are many outliers or one that particularly skews the mean. Could use the median as the basis of comparison, but this will still skew if there are too many outliers.

Could also do a Bayesian analysis with a mixture model for outliers. (see Hogg, Bovy and Lang, 2010) Assume data are representative of a true value b and have a flat prior I. To your Bayesian model, add N binary integers q_i, which are 0 if x_i is an outlier and 1 otherwise. To fully parametrize the outliers, we also need the mean and variance of the outlier distribution, Y_b, V_b.

The likelihood for a Gaussian model for outliers becomes:

$$\mathcal{L} = p(\{x_i\}_{i=1}^N | b, \{q_i\}_{i=1}^N, Y_b, V_b | I) \tag{5}$$

$$\mathcal{L} = \prod_{i=1}^N \left[\frac{1}{\sqrt{2\pi\sigma_{yi}^2}}\exp\left(-\frac{[y_i - b]^2}{2\sigma_{yi}^2}\right)\right]^{q_i} \tag{6}$$

$$\times \left[\frac{1}{\sqrt{2\pi[V_b + \sigma_{yi}^2]}}\exp\left(-\frac{[y_i - Y_b]^2}{2[V_b + \sigma_{yi}^2]}\right)\right]^{[1-q_i]} \tag{7}$$

If P_b is the prior probability of finding outliers,

$$p(\{q_i\}_{i=1}^N | P_b, I) = \prod_{i=1}^N [1 - P_b]^{q_i} P_b^{[1-q_i]} \tag{8}$$

We then marginalize over unknown quantities, P_b, Y_b, V_b:

$$p(b|\{y_i\}_{i=1}^N, I) = \int d\{q_i\}_{i=1}^N dP_b dY_b dV_b p(b, \{q_i\}_{i=1}^N, P_b, Y_b, V_b, I), \tag{9}$$

where an integral over $d\{q_i\}_{i=1}^N$ means a sum of all 2^N possible settings and all other integrals are over their prior support.

This is a lot of computation, so instead use mixture model with likelihood

$$\mathcal{L} \propto \prod_{i=1}^N \left[\frac{1 - P_b}{\sqrt{2\pi\sigma_{yi}^2}}\exp\left(-\frac{[y_i - b]^2}{2\sigma_{yi}^2}\right)\right]^{q_i} \tag{10}$$

$$\times \left[\frac{P_b}{\sqrt{2\pi[V_b + \sigma_{yi}^2]}}\exp\left(-\frac{[y_i - Y_b]^2}{2[V_b + \sigma_{yi}^2]}\right)\right]^{[1-q_i]} \tag{11}$$

Challenges are choosing your priors, but once done you can sample the marginalized posterior probability distribution to find the most likely b value.