Probability and Curve Fitting

Let's face it: There are too many variables here, all of which have similar names and definitions. There are 3 different standard deviations in play! Hopefully this helps:

For a single measurement...

The "Gaussian" or "Normal" Probability Density Function

$$p(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

The probability that you'll draw a number between y and $y + dy$ from a Normal distribution. If you draw a million numbers from a normal distribution and make a histogram, it will look Gaussian.

The Gaussian distribution special: The Central Limit Theorem says that the more random factors that contribute to any outcome, the more Gaussian the probability distribution becomes.

μ: True Mean

The peak of a Gaussian distribution that you are drawing from. You often can't know the true mean. If you had a perfect model for your data, the model would predict the true mean. The definition of expectation value says that for the distribution above, $\langle y \rangle = \mu$

σ: True Standard Deviation

The *one-sided width* of a Gaussian distribution that you are drawing from. 68% of the time, you will draw a number within $\pm 1\sigma$ of the true mean μ.

$$\int_{-\sigma}^{+\sigma} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\sigma)^2}{2\sigma^2}} dy \approx 0.68$$

You often can't know the true standard deviation. Even a model of your system can't predict this because it has to do with the net effect of all measurement errors.

σ^2: True Variance

Variances are just squares of standard deviations. These often have nicer mathematical properties than the standard deviation. For example, if you draw several random numbers and sum them together, the variance of the sum is the sum of the variances of each one. This is not true for the standard deviations. Physicists report σ rather than σ^2 because σ of a thing has the same units as the thing itself.

For a sample of $N \approx 5$ measurements with the same settings...

We'll make each measurement $N \approx 5$ times. For example, at *each* knob position, we'll measure the intensity of light $N \approx 5$ times to get $N \approx 5$ measurements: $\{y_1, y_2, \ldots y_N\}$.

$\bar{y}$: Sample Mean

The average of the $N \approx 5$ samples:

$$\bar{y} = (y_1 + y_2 + \cdots + y_N)/N = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

This is an *unbiased estimate of the true mean* μ: given $N \approx 5$ measurements, forming their mean is the best you can do—it won't overestimate or underestimate μ.

This becomes *a single data point* in a plot. (The point's error bars are described below.)

Math note: If you draw N samples and calculate their mean, and you do that procedure a bazillion times, the mean of these *sample means* will converge on the true mean: $\langle \bar{y} \rangle = \mu$.

s^2: Sample Variance

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$$

This is an *unbiased estimate of the true mean variance* σ^2 in the same sense as above: given $N \approx 5$ measurements, computing s^2 is the best you can do—on average, it won't overestimate or underestimate σ.

Math note: If you draw $N \approx 5$ samples and calculate their s^2 a bazillion times, the mean of these bazillion sample variances will converge on the true variance: $\langle s^2 \rangle = \sigma^2$.

s: Sample Standard Deviation

Square root of s^2 from above. Unfortunately, this is *not* an unbiased estimator of σ. If you do this procure for a bazillion groups of $N \approx 5$ samples, the average of these bazillion s calculations won't converge to σ. It'll be too small on average, even if you correctly use $N - 1$ in the formula above: $\langle s \rangle < \sigma$. There's no simple unbiased estimator of σ itself.

$\sigma_{\bar{y}}$: Sample Standard Deviation of the Mean; (or "Standard Error of the Mean")

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{N}} \qquad \text{or estimated from the sample standard deivation as} \qquad = \frac{s}{\sqrt{N}}$$

A mean of N samples (a sample mean $\bar{y}$) gets closer to the true mean μ as you increase your sample size N (say from 5, to 10, to 100). Increasing N gives you get a better estimate of the true mean μ. How much better? Better (narrower) by $\sqrt{N}$.

If you draw N samples and form their sample mean $\bar{y}$, and you do that procedure a bazillion times to make a histogram, the one-sided width of that histogram would be $\sigma_{\bar{y}}$. Around 68% of sample means $\bar{y}$ will be within $\pm 1\sigma_{\bar{y}}$ of the true mean μ.

It's as if $\bar{y}$ was drawn from a normal distribution of mean μ and standard deviation $\sigma_{\bar{y}}$.

This $\sigma_{\bar{y}}$ becomes the *one-sided size of the error bar* of a data point in a plot.

Curve Fitting

You have a model where you can turn a knob to some value x and predict what you'd expect for a measurement of, say, the intensity of light there y. The model is some function $f(x)$. For a linear model, $f(x) = ax + b$. The goal is to find the best parameters like a and b.

You'll sweep through, say, $M \approx 9$ knob positions $N \approx 5$ times. You'll end up with $N \approx 5$ independent measurements for each of $M \approx 9$ knob positions. This is $NM \approx 45$ numbers.

For each of the $M \approx 9$ independent-variable positions $\{x_1, x_2, \ldots x_M\}$, you will compute:

- The sample mean $\bar{y}_m$ (each of the $M \approx 9$ sample means is the average of $N \approx 5$ independent measurements with the knob at that particular x_m). These $M \approx 9$ means are your $M \approx 9$ data points on the plot.

- The standard error of the mean $\sigma_{\bar{y}}$ estimated from data. These are the $M \approx 9$ error bars attached to each point. They go $+\sigma_{\bar{y}_m}$ above the point to $-\sigma_{\bar{y}_m}$ below.

At each knob position x_m, we assume the data is drawn at random from a normal distribution centered on the "true" values: $y_m^{(\text{true})} = f(x_m) = a^{(\text{true})} + b^{(\text{true})} x_m$.

$$p(y_m) = \frac{1}{\sqrt{2\pi\sigma_m}} e^{-\frac{(y_m - f(x_m))^2}{2\sigma_m^2}}$$

The variance σ_m^2 of each measurement is determined by a combination of independent errors that lead to a spread in measurements at each particular knob position x_m. We don't know σ_m^2 a priori, but can estimate it as the sample variance s_m^2.

Each sample mean $\bar{y}_m$ (each data point on the plot) is also drawn from a normal distribution centered on the same "true" mean value, but with a narrower width $\sigma_{\bar{y}_m} = \frac{\sigma_m}{\sqrt{N}}$.

$$p(\bar{y}_m) = \frac{1}{\sqrt{2\pi\sigma_{\bar{y}_m}}} e^{-\frac{(\bar{y}_m - f(x_m))^2}{2\sigma_{\bar{y}_m}^2}}$$

If we plot the data points with error bars on top of the true model line or curve, around 68% of the points should be within the two-sided error bar of the model.
We assume that each data point is statistically independent, meaning the joint probability

$$p(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_M) \quad = \quad p(\bar{y}_1)p(\bar{y}_2)\ldots p(\bar{y}_M) \quad = \quad \left[\prod_{m=1}^{M} \left(\frac{1}{\sqrt{2\pi\sigma_{\bar{y}_m}}}\right)\right] e^{-\chi^2/2},$$

where

$$\chi^2 \equiv \sum_{m=1}^{M} \left(\frac{\bar{y}_m - f(x_m)}{\sigma_{\bar{y}_m}}\right)^2$$

is a measure of how far away the points are from the fit (in terms of standard deviations).

The best fit is where the parameters (a and b in the linear case) maximize the joint probability. Since the giant product in square brackets does not depend on the fit parameters, the maximum joint probability happens when the parameters minimize χ^2. For a general function with lots of parameters, a computer can search for the parameter values that minimize χ^2. For the special case of a linear fit where $f(x) = ax + b$, you can take the derivative of χ^2 with respect to a and b, set them each to 0, and solve exactly for the best a and b using linear algebra. The 1-sigma width σ_a and σ_b can then be computed by propagation of error.

χ^2 of a Fit, the χ^2 Distribution, and the Probability to Exceed

Once you have the best-fit parameters, found by minimizing χ^2 (shifting the fit line as close to the data points as possible), you can look at the value of χ^2 itself for those parameters:

$$\chi^2_{\text{fit}} \equiv \sum_{m=1}^{M} \left(\frac{\bar{y}_m - y_m^{\text{fit}}}{\sigma_{\bar{y}_m}} \right)^2$$

Each term in the sum is the squared number of standard deviations between the data point and the fit line. In other words, it's the distance-squared from the point to the line in units of standard deviations (or equivalently in numbers of error bars).

One could also consider the χ^2 for the true model, assuming you know the true parameters

$$\chi^2_{\text{true}} \equiv \sum_{m=1}^{M} \left(\frac{\bar{y}_m - y_m^{\text{true}}}{\sigma_{\bar{y}_m}} \right)^2 .$$

For the true model, each point should be, on average, one standard deviation away. So χ^2_{true} should be approximately M, the number of data points. This makes $\chi^2_{\text{true}}/M \approx 1$. If you look at a distribution of χ^2_{true}, it should have a mean of M.

The distribution of χ^2_{true} itself is called "the χ^2 distribution with M degrees of freedom."[1] Because you've divided by the standard error in each term, this is the same distribution as the sum of the squares of M standard normal random variables (each with $\mu = 0$ and $\sigma = 1$).

However, since you don't know the true model, and since you used the data itself to find the best-fit parameters that minimized χ^2, your fit is typically going to be a little closer to your particular set of points than the true function would be. (Imagine fitting M points with a polynomial of degree M: It would go exactly through all of the points!)

We'll call the number of fit parameters P. For a line with a slope and intercept, $P = 2$. It can be shown that *each fit parameter effectively removes one degree of freedom*: The distribution of χ^2_{fit} that you'd get by doing a bazillion fits to a bazillion data sets is the "χ^2 distribution with $k = M - P$ degrees of freedom." It's the distribution that you'd get if you summed only k samples drawn from the standard normal distribution.

You can look up how likely your particular value of χ^2_{fit} is, given k. In particular we ask, "What's the probability that I'd find a χ^2_{fit} value *greater* than the one I actually found if I did the experiment again using the same number of data points with the same uncertainties?" This is called the **Probability to Exceed (PTE)**, sometimes written $P_>$, whose distribution is flat between 0 and 1: 98% of the time, you should get a $P_>$ value between 0.01 and 0.99.

Just like the denominator of the sample variance needed $N - 1$ instead of N, χ^2_{fit} has a mean of $k = M - P$ instead of M. This means that

$$\frac{\chi^2_{\text{fit}}}{k} = \frac{\chi^2_{\text{fit}}}{M - P} \approx 1 \qquad \text{just like} \qquad \frac{\chi^2_{\text{true}}}{M} \approx 1$$

We look to χ^2_{fit}/k as a partial measure of "goodness of fit." For a good fit, this is around 1.

If $\chi^2_{\text{fit}}/k \gg 1$, it means that the points were too many standard errors away from the fit. Either the error bars were too small or the function isn't a good model for the data.

If $\chi^2_{\text{fit}}/k \ll 1$, it means that the points are too close to the fit. Either the error bars were too big or you used too many (often non-physical) parameters in the fit (called overfitting).

[1] `https://en.wikipedia.org/wiki/Chi-square_distribution`