Error bars

It's good practice to give some indication of uncertainty or expected variability in experimental results. You will need to report experimental results in your MLPR coursework. Many of you will also write up experimental results in dissertations this year, and you will also want to know how seriously to take numbers that you measure in your future work.

We will discuss some different "standard deviations" that you might see reported, or want to report, including "standard errors on the mean".

Standard errors on a mean

Imagine we are taking a series of experimental measurements $\{x_n\}_{n=1}^{N}$. These could be all sorts of things: for example the times taken for different runs of a program that you're testing, or the weights of different members of a gym.

We will assume that the measurements are taken independently from some unknown distribution. For example, the test conditions for the computer program are stable, and the times depend on independent random choices. Or the gym members were selected at random from a population of interest.

We assume the distribution has a finite mean μ and variance σ^2, which are unknown. We can estimate these from N independent samples:

$$\begin{aligned}\mu \approx \bar{x} &= \frac{1}{N}\sum_{n=1}^{N} x_n\\ \sigma^2 \approx \hat{\sigma}^2 &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\bar{x})^2.\end{aligned}$$

The $(N-1)$ in the estimator for the variance, rather than N, is a small detail you don't need to worry about for this course.[1]

The estimator $\bar{x}$ is itself a random variable. If we gathered a second dataset and computed its mean in the same way, we would get a different $\bar{x}$. For some datasets $\bar{x}$ will be bigger than the underlying true mean μ, sometimes it will be smaller. The mean of $\bar{x}$ is the right answer μ. That is, $\bar{x}$ is an unbiased estimator.

Using the rules of expectations and variances (separate note available), we can estimate the variance of $\bar{x}$. We assume here that the observations are independent:

$$\begin{aligned}\mathrm{var}[\bar{x}] &= \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{var}[x_n]\\ &= \frac{1}{N^2}N\sigma^2 = \sigma^2/N \approx \hat{\sigma}^2/N.\end{aligned}$$

A "typical" deviation from the mean is given by the standard deviation (*not* the variance). So we write:

$$\mu = \bar{x} \pm \hat{\sigma}/\sqrt{N},$$

to give an indication of how precisely we think we have measured the mean of the distribution with our N samples. Some papers might report ± two standard deviations.

If the distribution over observations has finite mean and variance, then for large N the Central Limit Theorem (CLT) tells us that $\bar{x}$ will be approximately Gaussian distributed close to its mean. We *don't* assume that the data are Gaussian distributed. We just note that the sum, and so average, of many values will be approximately Gaussian. With that interpretation, we expect the estimate $\bar{x}$ to be within 1 standard error $\sigma/\sqrt{N}$ about 2/3 of the time, and within 2 standard errors about 95% of the time.

[1] The $N-1$ normalization makes the variance estimator unbiased and is what the Matlab/Octave `var` function does by default. NumPy's `np.var` requires the option `ddof=1` to get the unbiased estimator. However, if N is small enough that this difference matters, you need to be more careful about the statistics than we are in this note.

Care: we don't evaluate the 'true' standard error $\sigma/\sqrt{N}$, but an approximation of it, $\hat{\sigma}/\sqrt{N}$. Moreover, the CLT will only be accurate for large N, and can't be trusted several standard deviations away from the mean. In other words, the "error bar" $\mu = \bar{x} \pm \hat{\sigma}/\sqrt{N}$ gives some indication of what means are plausible. Giving this statement is better than just stating $\mu \approx \bar{x}$, but more concrete statistical statements would require additional analysis.

Application to test set errors

The average test set loss,

$$L_{\text{test}} = \frac{1}{M}\sum_{m=1}^{M} L(y_m, f(x_m)) = \frac{1}{M}\sum_{m=1}^{M} L_m,$$

is an estimate of the generalization error, the average loss we would see if we could gather an infinite test set with the same distribution. How wrong might this estimate be?

We don't and can't assume that the individual losses, $L(y, f(x))$, are Gaussian-distributed. For example, when performing classification we might report the 0–1 loss, which is zero when we are correct and one when we make an error. In this example, the distribution over the losses is a Bernoulli distribution.

However, we can compute the empirical mean and variance of any set of losses, and report an estimate of the mean, with a standard error bar.

Before taking such an error bar on test performance too seriously we would need to think about whether the theory above applies. Are the test cases independent? If the loss isn't bounded, is it likely to have a finite variance we can reasonably estimate?

Reliability of a method?

A standard error on the test set loss indicates how much a particular fitted model might deviate from the quoted performance on future data. It doesn't say whether the machine learning method, if re-run on new data, will work well in future.

Readers of a paper may also be interested in how variable the performance is when the model is fitted on different data (or, for methods later in the course, with different random initial settings). We could report the standard deviation of the models' performances (not a standard error on the mean) to indicate how much a future fit will typically vary from average performance.

There is often more than one standard deviation one might report, and papers are sometimes not clear on whether "error bars" are reporting the variability of a population or a standard error on a mean. Always try to be clear precisely what "error bar" you are reporting and why.

Which method is better?

If we fit two methods, A and B, we can get a test set loss and standard error for both. If these two error bars overlap, it is common—but wrong—to conclude that we can't tell whether A is better than B.

As an example, if A had a loss of 0.1 less than B on every single test case in a test set of 1,000 cases, I would be incredibly confident that A was better than B. The two standard errors for the methods' performance could be larger (for example 0.5). It is possible to be sure of the ordering of two models, without knowing how good either is very precisely.

To test model A against model B we could do a simple paired comparison. Construct the difference in losses on each test case:

$$\delta_m = L(y_m, f(x_m; B)) - L(y_m, f(x_m; A)).$$

If the mean of the δ's is several standard errors greater than zero, we would report that A is the better model.