General Themes

- Basic mathematical approaches to data and models are successful because of fundamental principles, not because of accidents. What are those principles?
- Use appropriate mathematical objects to represent data and models: Measurements in the lab are discrete and finite, but mathematical concepts and models typically involve continua and infinities. So the mathematical constructs should allow for a smooth transition between large but finite and infinite, and a smooth transition between sampled and continuous
- Find natural coordinates based on the intrinsic features of the problem. These will typically lead to more incisive analyses, and descriptions of the data that suggest generalization, not the specific features of the experiment, measuring devices, etc.
- We will make counter-factual assumptions (e.g., "linearity"). We make them because they are extremely useful starting points, and because there are natural ways to relax them. There seems to be a meta- principle: principled methods based on broad assumptions that are known to be only approximately correct work much better than non-principled methods based on ad hoc assumptions.

Why is the field of statistics still an active one?

It's obvious that one needs statistics: to describe experimental data in a compact way, to compare datasets, to ask whether data are consistent with a model. But why is this a hard problem (or at least, why are people still making advances)? We will be focusing on multivariate data (e.g., time series, images), but the reason that "statistics" is nontrivial emerges even when we look at univariate data. Of course multivariate data does not make things better.

Toy example: estimating the mean

To set it up: we suppose that there is a collection of possible outcomes (an "ensemble" Ω, or a set of possible measurements associated with their probabilities). We would like to estimate the true mean of Ω, that is, $\mu = E(x)$.

$E(x)$ means the expected value of x, which we may also write as $\langle x \rangle_\Omega$, to emphasize the dependence on the ensemble Ω. For discrete ensembles, $\langle x \rangle_\Omega = \sum_{x \in \Omega} x p(x)$, where $p(x)$ is the probability of drawing x from Ω. For continuous ensembles, $\langle x \rangle_\Omega = \int_\Omega x p(x) dx$, where $p(x)\Delta x$ is the probability of drawing a value between x and $x + \Delta x$ from Ω.

We now draw N values from Ω, say x_1, …, x_N. Our problem is to craft an "estimator" function, say, $\hat{\mu} = \hat{\mu}(x_1, ..., x_N)$, to provide an estimate of μ. The obvious choice is the

sample mean. This is also known as the "plug-in" estimator, $\hat{\mu}_{plugin} = \frac{1}{N}\sum_{i=1}^{N} x_i$, since it is "plugging in" the measured values into the formula for the mean. But it is not the only choice.

There are some clearly silly choices: (a) a fixed, *a priori* guess, independent of the data (b) throw out the even-numbered measurements, and take the sample mean of the rest, (c) take the sample mean, and add a fixed number (say, 7), (d) take the sample mean, and add a number that depends on N, say, $1/N$, (e) choose one value from the data.

Why do we know that these choices are silly? Conversely, what properties do we want an estimator to have?

Desirable properties of an estimator

A good estimator should be unbiased (i.e., it should not systematically over- or under-estimate), and should converge to the correct value as one has more and more data. That is, as N grows, we would like the expected value of the estimator. $E(\hat{\mu})$ to converge to the correct answer, μ. This has two components: "bias" and "consistency."
An unbiased estimator gives the correct value, on average, when applied to a finite dataset. That is, the bias is the difference between the expected value of the estimator for a sample size N, and the true value, $E(\hat{\mu}) - \mu$, where the expectation is taken over all datasets of size N.

A consistent estimator is one that eventually converges to the correct answer. One way to formalize this is $\lim_{N\to\infty} E\left(\left(E(\hat{\mu}) - \mu\right)^2\right) = 0$: i.e., the mean-squared scatter of the estimator around the correct value eventually goes to 0 as more data are accumulated.
Occasionally, one is interested in alternative definitions of consistency – for example, bounding the largest possible error, rather than bounding the expected mean-squared error. Or bounding the largest possible error that can occur with probability > 0.

Both the lack-of-bias and the consistency conditions can be phrased in stricter forms, in which the speed of approach to 0 is specified, typically K/N for some constant K.
More *efficient* estimators converge more rapidly, i.e., have a smaller *K*. One can't typically expect to do better than a bias that decreases as K/N, for some *K*.

How can one evaluate (theoretically) the performance of a statistic? One does a thought experiment in which one draws multiple sets of N values from Ω, calculates the estimator of interest, and compares it with the true value. At a second level, one could postulate a family of ensembles, say $\Omega(\beta)$, one for each value of an unknown parameter β (or, more typically, a set of parameters β_k), and then see how sensitive the above analysis is to knowing β. That is, how strongly does the merit of the statistic depend on knowing, precisely, the form of the distribution? Sometimes one can do this analytically.

One can imagine a situation in which an estimator performs wonderfully for a particular family of ensembles, but performs terribly for ensembles that are not in this family – i.e., an estimator that is highly sensitive to model error.

Back to the toy example

For Gaussian ensembles, the plug-in estimator for the mean is unbiased, and consistent, and it is the most efficient estimator (K is as small as possible.) But the Gaussian ensemble is the *only* ensemble for which this is true. For ensembles that are not Gaussian, there is always a more efficient estimator of the mean, i.e., one for which the convergence to the true value is faster or more certain. The plug-in estimator remains useful because (a) it is unbiased, (b) it is consistent, (c) it is simple, (d) its properties are simple to calculate, (e) often the improvements conferred by other estimators are fragile, i.e., they are highly sensitive to the assumed shape of the distribution Ω.

Here is an important, practical example of a situation in which the plug-in estimator for the mean can be improved on:

Say the ensemble Ω is known to have the following structure: most of the values come from a Gaussian distribution (mean and variance unknown), but a small fraction, say ε, of the values are corrupted by a large measurement error. That is, ε of the values consist of samples from the underlying Gaussian, to which a large quantity M is either added or subtracted – or even, that the data are replaced by a large quantity $\pm M$. (Think of this as modeling a typographical error, or a rare but very large artifact.) Now, construct an estimator in two steps. First, throw out some fraction $f > \varepsilon$ of the extreme values. Then, take the mean of the remaining values. This is known as a "trimmed mean" estimator.

Since the first step gets rid of much of the variability, the resulting estimator converges faster than the plug-in estimator. If, say, we choose $f = 2\varepsilon$, then nearly all of the time, all outliers will have been eliminated. The trimmed mean is an example of a "robust" estimator (one that is relatively insensitive to outliers). The median is the limiting case of the trimmed mean (discard all but one of the measurements as extreme).

In this example, Ω has heavy tails (platykurtotic). You can think of other examples in which the heavy tail extends in only one direction (Ω is skewed), or has known shape, or even, in which Ω has light tails (leptokurtotic). These lead to other estimators that all do better than the plug-in estimator, each in its own domain. The trimmed mean estimator is worse than the plug-in estimator for a Gaussian, but not much worse. So it is often a good choice. And one often does it even without thinking ("let's throw out that experiment, it's an outlier.")

Note that the above example is *extremely simple* in that we are only considering univariate quantities. Additionally, we are attempting to estimate a quantity – the mean – for which the plug-in estimator is unbiased and consistent. Neither of these are typically the case – for most statistics, even for Gaussian ensembles, the plug-in estimator is biased

(for example, the sample variance). The plug-in estimator for the variance is $\frac{1}{N}\sum\left(x_i - \hat{\mu}_{plugin}\right)^2$, but this is biased; an unbiased estimator is $\frac{1}{N-1}\sum\left(x_i - \hat{\mu}_{plugin}\right)^2$.

In general, there is usually a tradeoff between bias and consistency (i.e., you can optimize the estimator for one, or for the other, but not for both.)

In sum, one reason that statistics is not trivial is that for estimators, one does not have "one size fits all."

Another factor is that computational burden is, in fact, relevant. Prior to computers, usable statistics nearly always were those that one could determine confidence limits, etc. analytically (Gaussian, Poisson). Now more computationally-intensive approaches are practical – e.g., resampling approaches such as the bootstrap and the jackknife. But computational practicality remains a severe consideration, and resampling approaches are not foolproof.

As a rule of thumb, for estimators that are simple (such as a plug-in estimator), it is easier to determine their sensitivity to the choice of the ensemble, and this sensitivity is usually less severe than for a more complex estimator.

A simple situation in which it's not obvious how to construct a good estimator

Say that you know that the ensemble Ω contains only a finite number of different values, and you want to estimate the number of different values. What can you do with a finite sample? Obviously you know that the number of different values in Ω is no less than the number of different values in your sample. But it might be larger, and there's no reasonable way of estimating how much larger it can be unless you know something about the distribution of probabilities in Ω -- especially the distribution of the very low probabilities. With a model of this, you can develop a more sophisticated estimator, but only if you can believe the model.

This example indicates the basic difficulty that confronts estimating "information" from experimental .

Optional homework

(answers at the end of this document)

Q1. Consider the above "silly" estimators for the mean. Which are unbiased? Which are consistent?

Q2. Construct a class of distributions for which the following estimator of the mean is unbiased, and also more efficient than the plug-in estimator: choose the highest and

lowest values of the observations x_i, and take their mean. That is,

$\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$. This is a kind of opposite strategy to the trimmed mean. Justify your claim (analytically or via simulation).

A few principles

We want to be able to describe our data in a way that has meaning to others.

Some units are more natural than others (time: seconds rather than bins, space: cm (or deg) rather than pixels.

Images are obviously multivariate, but so are time series.

Often the origin is arbitrary (e.g., time).

With multidimensional datasets, it is even more of an issue: what direction should the axes point? Is there a grounded notion of "orthogonal" axes? Often at first glance there might seem to be: values at each pixel, or samples at each point in time. But when one looks at the biology or physics, one recognizes that each sample (say, at a specific pixel or time point) reflects underlying "causes" at nearby locations or times as well (blur, filtering). So there is no strong reason that the obvious coordinates (the samples) are the natural coordinates. Blurring and filtering amount to linear transformations on the data; these are inevitable, so we might as well recognize this at the outset.

This leads to notions of:

- natural coordinates (e.g., Fourier analysis)
- data-driven coordinates (e.g., principal components analysis)
- coordinate-free descriptions (e.g., information theory)

And this (especially the notion of natural coordinates) leads us to focus on symmetries of the system. Translation in time is the paradigm.

Symmetries are often only approximate. But it is usually better to use a principled approach that is approximate, than an unprincipled one.

Plans and options

Symmetry in the abstract (group theory)
Multivariate measurements in the abstract (vector spaces and their symmetries)
Implications of symmetry of the independent variable (how groups act on vector spaces)
Natural coordinates
Fourier analysis
    linear systems, filters

noise and variability
Intrinsic symmetries of a vector space, and data-driven coordinates
Principal components analysis
Independent components analysis
Entropy, information, and data analysis
Graph-theoretic approaches
Point processes

Answers to optional homework

Q1. Silly estimators for the mean: which are unbiased, which are consistent?

(a) a fixed, *a priori* guess, independent of the data: biased, inconsistent
(b) throw out even-numbered measurements, and take the sample mean of the rest: unbiased, consistent (but inefficient)
(c) sample mean, plus a fixed number: biased, inconsistent
(d) sample mean plus $1/N$: biased, consistent (but inefficient)
(e) choose one value from the data: unbiased, inconsistent (no improvement as the amount of data increases)

Q2. Construct a class of distributions for which the following estimator of the mean is unbiased, and also more efficient than the plug-in estimator: choose the highest and lowest values of the observations x_i, and take their mean. That is, $\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$. This is a kind of opposite strategy to the trimmed mean. Justify your claim (analytically or via simulation).

Answer: Consider the class of "binary" distributions, i.e., those that contain only two values, a_0 and a_1 (with $a_0 < a_1$), each of which is drawn with a probability of 0.5. So the true mean is $\mu = \frac{a_0 + a_1}{2}$. Most of the time, after a large number of draws *N*, the sample minimum will be a_0 and the sample maximum will be a_1, so the estimator $\hat{\mu} = \frac{1}{2}\left(\min\{x_i\} + \max\{x_i\}\right)$ will be exact. To see how fast this estimator converges to the true mean, we note that the sample minimum and maximum will be accurate after *N* except for the fraction of trials in which the same value is drawn on each sample. This happens $2/2^N = 1/2^{N-1}$ of the time, since it requires that all of the $N-1$ draws beyond the first draw are matched to the first draw. So this estimator converges exponentially. Note that the mean-squared error of standard "plug-in" estimator decreases only like $1/N$.

Note also that if we applied this estimator to a distribution with tails, such as a Gaussian, it would be inconsistent – and in fact it would get worse and worse as we collected more data.