V. Intro to Statistical Inference

For the last few lectures of this class, we will apply what we have learned about probability to some classical **statistical inference problems**. Many laypeople don't understand the distinction between probably and statistics. The dictionary defines a statistic as "a number that represents a piece of information" or "a quantity (as the mean of a sample) that is computed from a sample". The basic idea is that probability deals with models for how random data is generated. Statistics deals with the inverse of this — given data, how can we learn a probability model? The basic framework is that we observe a sequence of random variables

$$X = (X_1, X_2, \dots, X_N)$$

and we would like to extract some information about their distribution just by looking at the data.

Much of classical statistics focuses on different approaches to **parameter estimation**. We assume that the data $(X_1, \dots, X_N)$ has a joint pdf that depends on an *unknown* parameter θ. We want to use the observations to make a reasonable estimate of the value of θ.

Applications of this come up all over the place. Some examples include:

- What proportion of the voters in Florida will vote Republican in 2016?
- A decent model for the number of packets that arrive at a router in a one second interval is a Poisson distribution with parameter λ, where λ quantifies the frequency of arrivals. Of course, this parameter is different for different routers. How do I estimate this parameter at a particular router?
- The variation in height and weight of a certain population might give me clues about their environment. How do we estimate the variance of a normal random variable given a series of independent samples?

After introducing some basic concepts in parameter estimation, we will look at how well we can estimate the mean and the variance of a distribution given a number of independent samples. We will then look at two general frameworks for parameter estimation: *maximum likelihood*, where the general principle is to find the parameter that would make the observations the most likely, and *Bayesian estimation*, which trades off the knowledge gleaned from the observations against our a priori knowledge of what the likely values of the parameter might be.

Parameter Estimation

Given observations $(X_1, X_2, \ldots, X_N)$ an **estimator** is a **random variable** which is a function of the data:

$$\widehat{\Theta}_N = g(X_1, \ldots, X_N),$$

for some known function $g(\cdot)$. The estimator maps the N observations into a single random variable; we make the dependence of $\widehat{\Theta}$ on N explicit (in the subscript) because we will often be interested in how this estimate improves as N gets bigger.

Example: Suppose that we have a coin that turns up heads with unknown probability θ. We would like to estimate θ from a series of observations

$$X_i = \begin{cases} 1 & \text{coin comes up heads} \\ 0 & \text{tails} \end{cases} \quad \Rightarrow \quad p_{X_i}(k;\theta) = \begin{cases} \theta & k = 1 \\ 1-\theta & k = 0 \end{cases}$$

We know the form of the pmf of the X_i, but not the underlying parameter θ — that the pmf depends on this unknown parameter is made explicit with the notation $p_{X_i}(k;\theta)$. A completely reasonable thing to do is to estimate θ using

$$\widehat{\Theta}_N = \frac{X_1 + X_2 + \cdots + X_N}{N}.$$

While we are not treating the underlying parameter θ as being random (it is just unknown), the estimate $\widehat{\Theta}_N$ is random, since it depends on the data X_i which are random variables.

There are many ways we might try to estimate a parameter θ given the data, and some of these ways will be better than others. But there are a few definitions which will help us classify and estimators behavior.

Terminology for estimators

- The **estimation error** of $\widehat{\Theta}_N$ is defined as
$$\text{error}(\widehat{\Theta}_N) = \widehat{\Theta}_N - \theta.$$
Of course, the error is also a random variable.
- The **bias** of $\widehat{\Theta}_N$ is the expected value of the estimation error:
$$\text{bias}(\widehat{\Theta}_N) = \mathrm{E}[\text{error}(\widehat{\Theta}_N)] = \mathrm{E}[\widehat{\Theta}_N] - \theta.$$
An estimator is called **unbiased** if $\mathrm{E}[\widehat{\Theta}_N] = \theta$.
- An estimator is **consistent** if it "converges" to θ as N gets large, no matter what θ is. By "converge", we mean that for *every* $\epsilon > 0$,
$$\lim_{N\to\infty} \mathrm{P}\left(|\widehat{\Theta}_N - \theta| > \epsilon\right) = 0.$$
- One useful way to quantify the performance of an estimator is using the **mean-square error** (MSE):
$$\mathrm{E}[(\text{error}(\widehat{\Theta}_N))^2] = \mathrm{E}[(\widehat{\Theta}_N - \theta)^2].$$
With a little algebraic manipulation, we can rewrite the MSE in terms of the bias and the variance of $\widehat{\Theta}_N$:
$$\mathrm{E}[(\widehat{\Theta}_N - \theta)^2] = \text{bias}^2(\widehat{\Theta}_N) + \text{var}(\widehat{\Theta}_N)$$

Take a minute to think about the difference between bias and variance in an estimator. Both contribute to the total error. Which is better? A dirty secret of statistics is that modifying an estimator to improve on one of these metrics often makes is perform worse on the other metric (this is called the "bias-variance tradeoff").

Sample mean

Two of the most fundamental problems are estimating the mean and variance of a distribution from a data set $X_1, \ldots, X_N$. The most obvious way to do this, and also the best way in almost every situation, is using the sample mean, which we have seen a few times already:

$$M_N = \frac{X_1 + \cdots + X_N}{N}.$$

Suppose that the true underlying mean of the X_i is $\mathrm{E}[X_i] = \mu$. It is clear that M_N is unbiased, since

$$\mathrm{E}[M_N] = \frac{\mathrm{E}[X_1] + \cdots \mathrm{E}[X_N]}{N} = \frac{N\mu}{N} = \mu.$$

We also know that the variance of M_N goes down as N gets bigger. With $\sigma^2 = \mathrm{var}(X_i)$,

$$\mathrm{var}(M_N) = \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_N)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

A direct result of this (which is the exact content of the weak law of large numbers, see the Technical Details section) is that M_N is *consistent*:

$$\lim_{N\to\infty} \mathrm{P}\left(|M_N - \mu| > \epsilon\right) = 0 \quad \forall \epsilon > 0.$$

Sample variance

Recall the definition of the variance of a random variable X:

$$\mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

Given i.i.d. observations $X_1, X_2, \ldots, X_N$ with mean μ and variance σ^2, one way we might estimate the underlying variance is to subtract the mean,

$$\widetilde{X_i} = X_i - \mu,$$

and then form

$$S_N^2 = \frac{\widetilde{X_1}^2 + \widetilde{X_2}^2 + \cdots + \widetilde{X_N}^2}{N}.$$

As above, this will give an unbiased and consistent estimator for the variance σ^2.

Of course, this requires that we know the true mean μ. When we do not know the mean, we can modify the procedure above by replacing μ with the *data dependent* sample mean M_N, forming an estimator from the sample variance

$$\begin{aligned}\bar{S}_N^2 &= \frac{(X_1 - M_N)^2 + (X_2 - M_N)^2 + \cdots + (X_N - M_N)^2}{N} \\ &= \frac{1}{N}\sum_{i=1}^{N}(X_i - M_N)^2.\end{aligned}$$

Is $\bar{S}_N^2$ an unbiased estimate of the variance? Let's check. First, note that we have

$$\begin{aligned}\mathrm{E}[\bar{S}_N^2] &= \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i - M_N)^2\right] \\ &= \mathrm{E}\left[\frac{1}{N}\sum_i X_i^2 - 2M_N \cdot \frac{1}{N}\sum_i X_i + M_N^2\right] \\ &= \mathrm{E}\left[\frac{1}{N}\sum_i X_i^2 - M_N^2\right].\end{aligned}$$

Next, using the facts that

$$\mathrm{E}[X_i^2] = \sigma^2 + \mu^2, \qquad \mathrm{E}[M_N^2] = \frac{\sigma^2}{N} + \mu^2,$$

we have

$$\mathrm{E}[\bar{S}_N^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2 = \frac{(N-1)}{N}\sigma^2.$$

So the estimate $\bar{S}_N^2$ of the variance is biased but asymptotically unbiased as $N \to \infty$.

To get an unbiased estimator, we can simply scale the sum differently:

$$\widehat{S}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - M_N)^2 = \frac{N}{N-1}\bar{S}_N^2.$$

The matlab command that calculates variance actually defaults to this unbiased estimator instead of the more obvious sample variance (though you can change its normalization when you call the function). It is easy to check that $\mathrm{E}[\widehat{S}_N^2] = \sigma^2$.

Technical Details: Weak Law of Large Numbers

In this section, we show that under very mild conditions, the sample mean converges to the true mean. The only condition is that the underlying distribution has finite variance.

We start by stating the main result precisely. Let X be a random variable with pdf $f_X(x)$, mean $\mathrm{E}[X] = \mu$, and variance $\mathrm{var}(X) = \sigma^2 < \infty$. We observe *samples* of X labeled $X_1, X_2, \ldots, X_N$. The X_i are independent of one another, and they all have the same distribution as X. We will show that the sample mean formed from a sample of size N:

$$M_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N),$$

obeys

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \;\; \leq \;\; \frac{\sigma^2}{N\epsilon^2},$$

where $\epsilon > 0$ is an arbitrarily small number. In the expression above, M_N is the only thing which is random; μ and σ^2 are fixed underlying properties of the distribution, N is the amount of data we see, and ϵ is something we can choose arbitrarily.

Notice that no matter how small ϵ is, the probability on the right hand side above goes to zero as $N \to \infty$. That is, for any fixed $\epsilon > 0$,

$$\lim_{N\to\infty} \mathrm{P}\left(|M_N - \mu| > \epsilon\right) = 0.$$

This result is follows from two simple but important tools known as the *Markov* and *Chebyshev* inequalities.

Markov inequality

Let X be a random variable that only takes positive values:

$$f_X(x) = 0, \quad \text{for } \ x < 0, \quad \text{or} \quad F_X(0) = 0.$$

Then

$$\boxed{\mathrm{P}\left(X \geq a\right) \ \leq \ \frac{\mathrm{E}[X]}{a} \quad \text{for all } a > 0.}$$

For example, the probability that X is more than 5 times its mean is 1/5, 10 times the mean is 1/10, etc. And this holds for **any distribution**.

The Markov inequality is easy to prove:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x) \ \mathrm{d}x \\
&\geq \int_a^\infty x f_X(x) \ \mathrm{d}x \\
&\geq \int_a^\infty a f_X(x) \ \mathrm{d}x \\
&= a \cdot \mathrm{P}\left(X \geq a\right)
\end{aligned}$$

and so $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$.

Again, this is a very general statement in that we have assumed nothing about X other than it is positive. The price for the generality is that the bound is typically very loose, and does not usually capture the behavior of $\mathrm{P}\left(X \geq a\right)$. We can, however, cleverly apply the Markov inequality to get something slightly more useful.

Chebyshev inequality

The main use of the Markov inequality turns out to be its use in deriving other, more accurate deviation inequalities. Here we will use it to derive the **Chebyshev inequality**, from which the weak law of large numbers will follow immediately.

Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then

$$\boxed{\mathrm{P}\left(|X-\mu| > c\right) \;\leq\; \frac{\sigma^2}{c^2} \quad \text{for all } c > 0.}$$

The Chebyshev inequality follows immediately from the Markov inequality in the following way. No matter what range of values X takes, the quantity $|X-\mu|^2$ is always positive. Thus

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) \;\leq\; \frac{\mathrm{E}[|X-\mu|^2]}{c^2} = \frac{\sigma^2}{c^2}.$$

Since squaring $(\cdot)^2$ is monotonic (invertible) over positive numbers,

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) = \mathrm{P}\left(|X-\mu| > c\right) \;\leq\; \frac{\sigma^2}{c^2}.$$

We now have a bound which depends on the mean and the variance of X; this leads to a more accurate approximation of the probability.

The weak law of large numbers (WLLN)

We now turn to the behavior of the the sample mean

$$M_N = \frac{X_1 + X_2 + \cdots + X_N}{N},$$

where again the X_i are iid random variables with $\mathrm{E}[X_i] = \mu$ and $\operatorname{var} X_i = \sigma^2$. We know that

$$\mathrm{E}[M_N] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_N]}{N} = \frac{N\mu}{N} = \mu,$$

and since the X_i are independent,

$$\operatorname{var}(M_N) = \frac{\operatorname{var}(X_1) + \operatorname{var}(X_2) + \cdots + \operatorname{var}(X_N)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

For any $\epsilon > 0$, a direct application of the Chebyshev inequality tells us that

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) \;\le\; \frac{\sigma^2}{N\epsilon^2}.$$

The point is that this gets arbitrarily small as $N \to \infty$ no matter what ϵ was chosen to be. We have established, in some sense, that even though $\{M_N\}$ is a sequence of random numbers, it converges to something deterministic, namely μ.

WLLN: Let $X_1, X_2, \ldots$ be iid random variables as above. For **every** $\epsilon > 0$, we have

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) = \mathrm{P}\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \epsilon\right) \longrightarrow 0,$$

as $N \to \infty$.

One of the philosophical consequences of the WLLN is that it tells us that probabilities can be estimated through **empirical frequencies**. Suppose I want to estimate the probability of and event A occurring related to some probabilistic experiment. We run a series of (independent) experiments, and set $X_i = 1$ if A occurred in experiment i, and $X_i = 0$ otherwise. Then given $X_1, \ldots, X_N$, we estimate the probability of A in a completely reasonable way, by computing the percentage of times it occurred:

$$p_{\text{empirical}} = \frac{X_1 + \cdots + X_N}{N}.$$

The WLLN tells us that

$$p_{\text{empirical}} \to \mathrm{P}\left(A\right), \quad \text{as } N \to \infty.$$

This lends some mathematical weight to our interpretation of probabilities as *relative frequencies*.