STAT 801 Lecture 10

Goals of Today's Lecture:

- Discuss Monte Carlo techniques
- Introduce statistical inference

Monte Carlo

Given random variables $X_1, \ldots, X_n$ whose joint density f (or distribution) is specified and a statistic $T(X_1, \ldots, X_n)$ whose distribution you want to know. To compute something like $P(T > t)$:

1. Generate $X_1, \ldots, X_n$ from the density f.
2. Compute $T_1 = T(X_1, \ldots, X_n)$.
3. Repeat N times getting $T_1, \ldots, T_N$.
4. Estimate $p = P(T > t)$ as $\hat{p} = M/N$ where M is number of repetitions where $T_i > t$.
5. Estimate accuracy of $\hat{p}$ using $\sqrt{\hat{p}(1-\hat{p})/N}$.

Notice accuracy inversely proportional to $\sqrt{N}$. There are a number of tricks to make the method more accurate (but they only change the constant of proportionality – the SE is still inversely proportional to the square root of the sample size).

Transformation

If X is to have cdf F and you generate U which is Uniform[0,1] then you can take $X = F^{-1}(U)$. For example X exponential makes $F(x) = 1 - e^{-x}$ and $F^{-1}(u) = -\log(1-u)$. Compare this with the method I showed for generating exponentials.

Acceptance Rejection

If you can't easily calculate F^{-1} but you know f you can try the acceptance rejection method. Find a density g and a constant c such that $f(x) \le cg(x)$ for each x and G^{-1} is computable or you otherwise know how to generate observations $W_1, W_2, \ldots$ independently from g. Generate W_1. Compute $p = f(W_1)/(cg(W_1)) \le 1$. Generate a uniform[0,1] random variable U_1 independent of all the Ws and let $Y = W_1$ if $U_1 \le p$. Otherwise get a new W and a new U and repeat until you find a $U_i \leq f(W_i)/(cg(W_i))$. You make Y be the last W you generated. This Y has density f.

Markov Chain Monte Carlo

In the last 10 years the following tactic has become popular, particularly for generating multivariate observations. If $W_1, W_2, \ldots$ is an (ergodic) Markov chain with stationary transitions and the stationary initial distribution of W has density f then you can get random variables which have the marginal density f by starting off the Markov chain and letting it run for a long time. The marginal distributions of the W_i converge to f. So you can estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.

There are now many versions of this technique including Gibbs Sampling and the Metropolis-Hastings algorithm. (The technique was invented in the 1950s by physicists: Metropolis et al. One of the authors of the paper was Edward Teller "father of the hydrogen bomb".)

Importance Sampling

If you want to compute

$$\theta \equiv E(T(X)) = \int T(x)f(x)dx$$

you can generate observations from a different density g and then compute

$$\hat{\theta} = n^{-1} \sum T(X_i)f(X_i)/g(X_i)$$

Then

$$\begin{aligned} E(\hat{\theta}) &= n^{-1} \sum E\left\{T(X_i)f(X_i)/g(X_i)\right\} \\ &= \int [T(x)f(x)/g(x)]g(x)dx \\ &= \int T(x)f(x)dx \\ &= \theta \end{aligned}$$

Variance reduction

Consider the problem of estimating the distribution of the sample mean for a Cauchy random variable. The Cauchy density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

We generate $U_1, \ldots, U_n$ uniforms and then define $X_i = \tan^{-1}(\pi(U_i - 1/2))$. Then we compute $T = \bar{X}$. Now to estimate $p = P(T > t)$ we would use

$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$

after generating N samples of size n. This estimate is unbiased and has standard error $\sqrt{p(1-p)/N}$.

We can improve this estimate by remembering that $-X_i$ also has Cauchy distribution. Take $S_i = -T_i$. Remember that S_i has the same distribution as T_i. Then we try (for $t > 0$)

$$\tilde{p} = [\sum_{i=1}^{N} 1(T_i > t) + \sum_{i=1}^{N} 1(S_i > t)]/(2N)$$

which is the average of two estimates like $\hat{p}$. The variance of $\tilde{p}$ is

$$(4N)^{-1} Var(1(T_i > t) + 1(S_i > t)) = (4N)^{-1} Var(1(|T| > t))$$

which is

$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$

Notice that the variance has an extra 2 in the denominator and that the numerator is also smaller – particularly for p near 1/2. So this method of variance reduction has resulted in a need for only half the sample size to get the same accuracy.

Regression estimates

Suppose we want to compute

$$\theta = E(|Z|)$$

where Z is standard normal. We generate N iid $N(0,1)$ variables $Z_1, \ldots, Z_N$ and compute $\hat{\theta} = \sum |Z_i|/N$. But we know that $E(Z_i^2) = 1$ and can see easily that $\hat{\theta}$ is positively correlated with $\sum Z_i^2/N$. So we consider using

$$\tilde{\theta} = \hat{\theta} - c(\sum Z_i^2/N - 1)$$

Notice that $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) = Var(\hat{\theta}) - 2cCov(\hat{\theta}, \sum Z_i^2/n) + c^2 Var(\sum Z_i^2/N)$$

The value of c which minimizes this is

$$c = \frac{Cov(\hat{\theta}, \sum Z_i^2/n)}{Var(\sum Z_i^2/N)}$$

and this value can be estimated by regressing the $|Z_i|$ on the Z_i^2!

Statistical Inference

Definition: A **model** is a family $\{P_\theta; \theta \in \Theta\}$ of possible distributions for some random variable X. (Our data set is X, so X will generally be a big vector or matrix or even more complicated object.)

We will assume throughout this course that the true distribution P of X is in fact some P_{θ_0} for some $\theta_0 \in \Theta$. We call θ_0 the true value of the parameter. Notice that this assumption will be wrong; we hope it is not wrong in an important way. If we are very worried that it is wrong we enlarge our model putting in more distributions and making Θ bigger.

Our goal is to observe the value of X and then guess θ_0 or some property of θ_0. We will consider the following classic mathematical versions of this:

1. Point estimation: we must compute an estimate $\hat{\theta} = \hat{\theta}(X)$ which lies in Θ (or something close to Θ).

2. Point estimation of a function of θ: we must compute an estimate $\hat{\phi} = \hat{\phi}(X)$ of $\phi = g(\theta)$.

3. Interval (or set) estimation. We must compute a set $C = C(X)$ in Θ which we think will contain θ_0.

4. Hypothesis testing: We must choose between $\theta_0 \in \Theta_0$ where $\Theta_0 \subset \Theta$.

5. Prediction: we must guess the value of an observable random variable Y whose distribution depends on θ_0. Typically Y is the value of the variable X in a repetition of the experiment.

There are several schools of statistical thinking with different views on how these problems should be done. The main schools of thought may be summarized roughly as follows:

- **Neyman Pearson**: A statistical procedure is evaluated by its long run frequency performance. Imagine repeating the data collection exercise many times, independently. The quality of a procedure is measured by its average performance when the true distribution of the X values is P_{θ_0}.

- **Bayes**: In this school of thought we treat θ as being random just like X. We compute the conditional distribution of what we don't know given what we do know. In particular we ask how a procedure will work on the data we actually got – no averaging over data we might have got.

- **Likelihood**: This school tries to combine the previous 2 by looking only at the data we actually got but trying to avoid treating θ as random.

For the next several weeks we do only the Neyman Pearson approach, though we use that approach to evaluate the quality of likelihood methods.

Likelihood

Suppose you toss a coin 6 times and get Heads twice. If p is the probability of getting H then the probability of getting 2 heads is

$$15p^2(1-p)^4$$

This probability, thought of as a function of p, is the **likelihood** function for this particular data.