STAT 801 Lecture 9

Reading for Today's Lecture: ?
Goals of Today's Lecture:

- State Slutsky's Theorem.
- Introduce the δ method.
- Discuss Monte Carlo techniques

Central Limit Theorems

1. If $Y_1, Y_2, \cdots$ are iid in R^p with mean μ and variance covariance Σ then $n^{1/2}(\bar{Y} - \mu)$ converges in distribution to $MVN(0, \Sigma)$.

2. If for each n we have a set of independent mean 0 random variables $X_{n1}, \ldots, X_{nn}$ and $E(X_{ni}) = 0$ and $Var(\sum_i X_{ni}) = 1$ and
$$\sum E(|X_{ni}|^3) \to 0$$
then $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lyapunov central limit theorem.

3. As in the Lyapunov CLT but replace the third moment condition with
$$\sum E(X_{ni}^2 1(|X_{ni}| > \epsilon)) \to 0$$
for each $\epsilon > 0$ then again $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lindeberg central limit theorem. (Lyapunov's condition implies Lindeberg's.)

4. There are extensions to random variables which are not independent. Examples include the m-dependent central limit theorem, the martingale central limit theorem, the central limit theorem for mixing processes.

5. Many important random variables are not sums of independent random variables. We handle these with Slutsky's theorem and the δ method.

Slutsky's Theorem: If X_n converges in distribution to X and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n$ converges in distribution to $X + c$. More generally, if $f(x, y)$ is continuous then $f(X_n, Y_n) \Rightarrow f(X, y)$.

Warning: the hypothesis that the limit of Y_n be constant is essential.

Definition: We say Y_n converges to Y in probability if

$$P(|Y_n - Y| > \epsilon) \to 0$$

for each $\epsilon > 0$.

The fact is that for Y constant convergence in distribution and in probability are the same. In general convergence in probability implies convergence in distribution. Both of these are weaker than almost sure convergence:

Definition: We say Y_n converges to Y almost surely if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1\,.$$

The delta method: Suppose a sequence Y_n of random variables converges to some y a constant and that if we define $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X. Suppose that f is a differentiable function on the range of Y_n. Then $a_n(f(Y_n) - f(y))$ converges in distribution to $f'(y)X$. If X_n is in R^p and f maps R^p to R^q then f' is the $q \times p$ matrix of first derivatives of components of f. Usually $a_n = n^{1/2}$.

Example: Suppose $X_1, \ldots, X_n$ are a sample from a population with mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. Then

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, \mu_4 - \sigma^4)$$

where $\Rightarrow$ is notation for convergence in distribution. For simplicity I define $s^2 = \overline{X^2} - \bar{X}^2$.

We take Y_n to be the vector with components $(\overline{X^2}, \bar{X})$. Then Y_n converges to $y = (\mu^2 + \sigma^2, \mu)$. Take $a_n = n^{1/2}$. Then

$$n^{1/2}(Y_n - y)$$

converges in distribution to $MVN(0, \Sigma)$ with

$$\Sigma = \left[\begin{array}{cc} \mu_4 - \sigma^4 & \mu_3 - \mu(\mu^2 + \sigma^2) \\ \mu_3 - \mu(\mu^2 + \sigma^2) & \sigma^2 \end{array}\right]$$

Define $f(x_1, x_2) = x_1 - x_2^2$. Then $s^2 = f(Y_n)$ and the gradient of f has components $(1, -2x_2)$. This leads to

$$n^{1/2}(s^2 - \sigma^2) \approx n^{1/2}(1, -2\mu) \begin{bmatrix} \overline{X^2} - (\mu^2 + \sigma^2) \\ \bar{X} - \mu \end{bmatrix}$$

which converges in distribution to the law of $(1, -2\mu)Y$ which is $N(0, a^t \Sigma a)$ where $a = (1, -2\mu)^t$. This boils down to $N(0, \mu_4 - \sigma^2)$.

Remark: In this sort of problem it is best to learn to recognize that the sample variance is unaffected by subtracting μ from each X. Thus there is no loss in assuming $\mu = 0$ which simplifies Σ and a.

Special case: if the observations are $N(\mu, \sigma^2)$ then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Our calculation has

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, 2\sigma^4)$$

You can divide through by σ^2 and get

$$n^{1/2}(\frac{s^2}{\sigma^2} - 1) \Rightarrow N(0, 2)$$

In fact $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution and so the usual central limit theorem shows that

$$(n-1)^{1/2}[(n-1)s^2/\sigma^2 - (n-1)] \Rightarrow N(0, 2)$$

(using mean of χ^2_1 is 1 and variance is 2). Factoring out $n-1$ gives the assertion that

$$(n-1)^{1/2}(s^2/\sigma^2 - 1) \Rightarrow N(0, 2)$$

which is our δ method calculation except for using $n-1$ instead of n. This difference is unimportant as can be checked using Slutsky's theorem.

Monte Carlo

The last method of distribution theory that I will review is Monte Carlo simulation. Suppose you have some random variables $X_1, \ldots, X_n$ whose joint distribution is specified and a statistic $T(X_1, \ldots, X_n)$ whose distribution you want to know. To compute something like $P(T > t)$ for some specific value of t we appeal to the limiting relative frequency interpretation of probability: $P(T > t)$ is the limit of the proportion of trials in a long sequence of trials in which T occurs. We use a (pseudo) random number generator to generate a sample $X_1, \ldots, X_n$ and then calculate the statistic getting T_1. Then we generate a new sample (independently of our first, say) and calculate T_2. We repeat this a large number of times say N and just count up how many of the T_k are larger than t. If there are M such T_k we estimate that $P(T > t) = M/N$.

The quantity M has a Binomial($N, p = P(T > t)$) distribution. The standard error of M/N is then $p(1-p)/N$ which is estimated by $M(N - M)/N^3$. This permits us to guess the accuracy of our study.

Notice that the standard deviation of M/N is $\sqrt{p(1-p)}/\sqrt{N}$ so that to improve the accuracy by a factor of 2 requires 4 times as many samples. This makes Monte Carlo a relatively time consuming method of calculation. There are a number of tricks to make the method more accurate (though they only change the constant of proportionality – the SE is still inversely proportional to the square root of the sample size).

Generating the Sample

Most computer languages have a facility for generating pseudo uniform random numbers, that is, variables U which have (approximately of course) a Uniform$[0,1]$ distribution. Other distributions are generated by transformation:

Exponential: $X = -\log U$ has an exponential distribution:

$$P(X > x) = P(-\log(U) > x) = P(U \le e^{-x}) = e^{-x}$$

Random uniforms generated on the computer sometimes have only 6 or 7 digits or so of detail. This can make the tail of your distribution grainy. If U were actually a multiple of 10^{-6} for instance then the largest possible value of X is $6\log(10)$. This problem can be ameliorated by the following algorithm:

- Generate U a Uniform[0,1] variable.
- Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.
- If $U \le \epsilon$ remember that the conditional distribution of $Y - y$ given $Y > y$ is exponential. You use this by generating a new U' and computing $Y' = -\log(U')$. Then take $Y = Y' - \log(\epsilon)$. The resulting Y has an exponential distribution. You should check this by computing $P(Y > y)$.

Normal: In general if F is a continuous cdf and U is Uniform[0,1] then $Y = F^{-1}(U)$ has cdf F because

$$P(Y \leq y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y)$$

This is almost the technique in the exponential distribution. For the normal distribution $F = \Phi$ (Φ is a common notation for the standard normal cdf) there is no closed form for F^{-1}. You could use a numerical algorithm to compute F^{-1} or you could use the following Box Müller trick. Generate U_1, U_2 two independent Uniform[0,1] variables. Define $Y_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$ and $Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$. Then you can check using the change of variables formula that Y_1 and Y_2 are independent $N(0,1)$ variables.

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 $f(x) \leq cg(x)$ 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 $p = f(W_1)/(cg(W_1)) \leq 1$. Generate a uniform[0,1] random variable U_1 independent of all the Ws and let $Y = W_1$ if $U_1 \leq p$. 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

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

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

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

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 NEXT