

ECON 623/2nd Half

Answers to Problem Set 1

John Rust

Fall 2010

I. Convergence Exercises

A. Consider the sequence $\{x_n\}$ where

$$x_n = \left(1 + \frac{x}{n}\right)^n. \quad (1)$$

Does this sequence have a limit? If so, what is the limit (prove your results for full credit).

Answer Yes, this is nothing but the formula for continuous compounding of interest with x as the product of the interest rate and the time the money is held, so the limit is $\exp\{x\}$. You can take logs and use L'Hôpital's rule to show this:

$$\log(x_n) = n \log\left(1 + \frac{x}{n}\right) = \frac{\log(1 + \frac{x}{n})}{\frac{1}{n}}. \quad (2)$$

Note that as $n \rightarrow \infty$ the numerator and denominator of the fraction on the right hand side of equation (2) both tend to 0. So we can use L'Hôpital's rule to find the limit by differentiating both the numerator and denominator and taking the limit of the ratio of these derivatives. It is not hard, using a little calculus, to show this limit is x . Thus, if $\lim_{n \rightarrow \infty} \log(x_n) = x$, then since \log is a continuous function, it follows that $\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} \exp\{\log(x_n)\} = \exp\{x\}$ by the Continuous Mapping Theorem.

B. Consider the sequence of random variables $\{\tilde{x}_n\}$ where

$$\tilde{x}_n = \left(1 + \frac{x + \tilde{u}_n/n}{n}\right)^n, \quad (3)$$

where $\{\tilde{u}_n\}$ are *IID* $U(0,1)$ random variables. Does this sequence converge? If so, in what sense(s): almost surely?, with probability 1? in probability?, in distribution?, uniformly with probability 1?

Answer The sequence $\{x_n\}$ is INID (Independently but Non-Identically Distributed) due to the presence of the n which changes the distribution of the x_n depending on the value of n . However intuitively, for large n , $x + \tilde{u}_n/n$ will be close to x and thus the \tilde{x}_n should be close to the deterministic sequence x_n given in part I-A above, and thus we would expect this sequence converges almost surely and in probability to the same limit, $\exp\{x\}$. We can formalize this by using the same argument using L'Hôpital's Rule as above to show that $\log(\tilde{x}_n)$ converges almost surely to x . Then the Continuous Mapping Theorem implies that $x_n = \exp\{\log(x_n)\}$ converges almost sure to $\exp\{x\}$, and therefore it also converges in probability to $\exp\{x\}$ and the distribution of \tilde{x}_n converges to a distribution that places probability 1 on the single value $\exp\{x\}$, so the sequence converges in distribution as well. The notion of "convergence

uniformly with probability 1" is not relevant unless we think of x as an argument and thus think of $\tilde{x}_n(x)$ as random functions of x . In that case, for each fixed x , the argument above indicates that $\tilde{x}_n(x)$ converges *pointwise* to $\exp\{x\}$. Now if we restrict x to lie in a compact set, say $x \in [0, 1]$ it is not hard to show that the $\tilde{x}_n(x)$ functions are continuously differentiable functions of x and their derivatives are uniformly bounded by e for all n and all $x \in [0, 1]$ with probability 1 (take the derivative of $\log(\tilde{x}_n(x))$ with respect to x and show it is less than or equal to 1 for all $x \in [0, 1]$). Thus, the collection of random functions $\{\tilde{x}_n(x)\}$ is *stochastically equicontinuous* and by the stochastic generalization of the Ascoli-Arzelà Theorem, it follows that $\tilde{x}_n(x)$ converges uniformly with probability 1 to $\exp\{x\}$ for $x \in [0, 1]$.

C. Consider the sequence of random variables $\{\tilde{x}_n\}$ where

$$\tilde{x}_n = \tilde{u}, \quad n = 1, 2, \dots \quad (4)$$

where \tilde{u} is a $U(0, 1)$ random variable. Does this sequence converge? If so, in what sense(s): almost surely?, with probability 1? in probability? in distribution? uniformly with probability 1? Are the elements of the $\{\tilde{x}_n\}$ IID?

Answer It converges almost surely to the random variable \tilde{u} . The elements are not IID.

D. Consider the sequence of random variables $\{\tilde{x}_n\}$ where

$$\tilde{x}_n = (1/n)\tilde{u} + (1 - 1/n)\tilde{u}_n, \quad n = 1, 2, \dots \quad (5)$$

where \tilde{u} is a $U(0, 1)$ random variable and $\{\tilde{u}_n\}$ are IID $U(0, 1)$ random variables and \tilde{u} is independently distributed from any of the \tilde{u}_n random variables. Is $\{\tilde{x}_n\}$ an IID sequence? If not, why is it not IID? Compute the covariance between \tilde{x}_j and \tilde{x}_{j+t} . What happens to this covariance as $t \rightarrow \infty$? for fixed j ? What happens as $j \rightarrow \infty$ for fixed t ? Does the sequence $\{\tilde{x}_n\}$ converge almost surely? with probability 1? in probability? in distribution? uniformly with probability 1?

Answer $\{\tilde{x}_n\}$ is not a IID sequence. $cov(\tilde{x}_j, \tilde{x}_{j+t}) = cov((\frac{1}{j}\tilde{u} + (1 - \frac{1}{j})\tilde{u}_j, \frac{1}{j+t}\tilde{u} + (1 - \frac{1}{j+t})\tilde{u}_{j+t}) = \frac{1}{j(j+t)}var(\tilde{u}) = \frac{1}{12j(j+t)} \neq 0$ since $cov(\tilde{u}, \tilde{u}_j) = 0$, $cov(\tilde{u}, \tilde{u}_{j+t}) = 0$, and $cov(\tilde{u}_j, \tilde{u}_{j+t}) = 0$. Therefore $\lim_{j \rightarrow \infty} cov(\tilde{x}_j, \tilde{x}_{j+t}) = 0$ and $\lim_{t \rightarrow \infty} cov(\tilde{x}_j, \tilde{x}_{j+t}) = 0$. The sequence $\{\tilde{x}_n\}$ converges in distribution to $U(0, 1)$ since $\frac{1}{n}\tilde{u} \xrightarrow{p} 0$, $\tilde{u}_n \xrightarrow{d} U(0, 1)$ and $\frac{1}{n}\tilde{u}_n \xrightarrow{p} 0$.

E. For the sequence of random variables in part C above, suppose we formed the normalized quantity

$$\bar{U}_N = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{x}_n - 1/2) \quad (6)$$

Would a Central Limit Theorem be applicable to the sequence $\{\bar{U}_N\}$ and would this converge in distribution to a Gaussian distribution? If not, does $\{\bar{U}_N\}$ converge in distribution to something else, or not converge in distribution at all?

Answer $\{\bar{U}_N\}$ does not converge since \tilde{x}_n is highly dependent. Rewrite $\bar{U}_N = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{x}_n - 1/2) = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{u} - 1/2) = \sqrt{N}(\tilde{u} - 1/2)$, then $Var\bar{U}_N = NVar\tilde{u} \rightarrow \infty$ as $N \rightarrow \infty$.

F. Do the same question as in part E. above except use the sequence $\{\tilde{x}_n\}$ where $\tilde{x}_n = \tilde{u}_n$ where $\{\tilde{u}_n\}$ is an *IID* $U(0, 1)$ sequence.

Answer $\{\bar{U}_N\}$ converges by CLT since \tilde{x}_n is *IID* and $E\tilde{x}_n^2 < \infty$. Again $\bar{U}_N = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{x}_n - 1/2) = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{u}_n - E\tilde{u}_n) \rightarrow^d N(0, \text{Var}(\tilde{u}_n))$, where $\text{Var}(\tilde{u}_n) = \frac{1}{12}$.

G. Do the same question as in part E. above except use the sequence $\{\tilde{x}_n\}$ defined in part D. above.

Answer Since $E(\tilde{x}_n) = E(\frac{1}{n}\tilde{u} + (1 - \frac{1}{n})\tilde{u}_n) = \frac{1}{2n} + \frac{1}{2} - \frac{1}{2n} = \frac{1}{2}$, we can rewrite $\bar{U}_N = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{x}_n - 1/2) = \frac{1}{\sqrt{N}} \sum_{n=1}^n (\tilde{x}_n - E\tilde{x}_n)$. We can apply a CLT for stationary ergodic processes since dependence decreases as shown in part E. and $\bar{U}_N \rightarrow^d N(0, \text{Var}(\tilde{x}_n))$, where $\text{Var}(\tilde{x}_n) = \text{Var}(\frac{1}{n}\tilde{u} + (1 - \frac{1}{n})\tilde{u}_n) = \frac{1}{n^2}\text{Var}(\tilde{u}) + \frac{n^2 - 2n + 1}{n^2}\text{Var}(\tilde{u}_n) \rightarrow \frac{1}{12}$.

H. Let $\{F_N(x)\}$ be the sequence of functions given by

$$F_N(x) = \frac{1}{N} \sum_{i=1}^N I\{\tilde{u}_i \leq x\} \quad (7)$$

where $\{\tilde{u}_i\}$ are *IID* $U(0, 1)$ random variables. Does the sequence $\{F_N(x)\}$ converge with probability 1? almost surely? in probability? in distribution? or uniformly with probability 1? If so, specify what it converges to in each case.

Answer The sequence $\{F_N(x)\}$ converges uniformly with probability 1 over a compact / convex set $x \in [0, 1]$. $\Pr\{\lim_{N \rightarrow \infty} [\sup_x |F_N(x) - F(x)|] = 0\} = \Pr\{\lim_{N \rightarrow \infty} [\sup_x |\frac{1}{N} \sum I\{\tilde{u}_i \leq x\} - x|] = 0\} = 1$.

I. Let $\{\bar{B}_N(x)\}$ be a sequence where $\bar{B}_N(x)$ is given by

$$\bar{B}_N(x) = \sqrt{N}(F_N(x) - x) = \frac{1}{\sqrt{N}} \sum_{n=1}^N [I\{\tilde{u}_n\} - x] \quad (8)$$

Does the sequence $\{\bar{B}_N(x)\}$ converge with probability one? almost surely? in probability? in distribution? or uniformly with probability 1? If you find it converges in any of these cases, specify what it converges to (i.e. the limiting quantity that this sequence converges to and whether the limiting quantity is deterministic or random) to receive full credit.

Answer Since \tilde{u}_n is *IID* $I\{\tilde{u}_n\}$ is *IID* Bernoulli distributed with mean $F(x)$ and variance $F(x)(1-F(x))$ such that its average $F_N(x)$ has $EF_N(x) = F(x)$ and $\text{Var}F_N(x) = \frac{1}{N}F(x)(1-F(x))$. By CLT $\bar{B}_N(x)$ converges to $\bar{B}_N(x) = \sqrt{N}(F_N(x) - F(x)) \rightarrow^d N(0, F(x)(1-F(x)))$. For \tilde{u}_n distributed $U(0, 1)$ we have $\bar{B}_N(x) = \sqrt{N}(F_N(x) - x) \rightarrow^d N(0, x(1-x))$.

J. Generate a sample of size $N = 1000$ from the $U(0, 1)$ and plot the $B_N(x)$ for $N = 1000$ and $x \in [0, 1]$. Then do the same when $N = 100$ and $N = 10$ and plot all three B_N functions on the same graph. What do you notice about these functions as N gets larger?

- K. for K *non-random* points (x_1, x_2, \dots, x_K) where $x_j \in (0, 1)$ and $x_1 < x_2 < \dots < x_{K-1} < x_K$, what does the $K \times 1$ vector sequence $\{\bar{B}_N\}$ converge to, where

$$\bar{B}_N = (\bar{B}_N(x_1), \bar{B}_N(x_2), \dots, \bar{B}_N(x_K))' \quad (9)$$

where the $\bar{B}_N(x)$ function is defined in part I above?

Answer Note that the covariance of (i, j) element entering into the summands of the \bar{B}_N vector is $[I\{x \leq x_i - F(x_i)\}][I\{x \leq x_j - F(x_j)\}]$ and the expectation of this, the covariance, is $F(\min(x_i, x_j)) - F(x_i)F(x_j)$. Note that this gives the formula for the diagonal terms that you already derived as a special case $\text{var}(I\{x \leq x_i\} - F(x_i)) = E([I\{x \leq x_i\} - F(x_i)][I\{x \leq x_i\} - F(x_i)]) = F(x_i) - F(x_i)F(x_i)$. This is the covariance matrix of the asymptotic Normal distribution for the B_N vector and the covariance functional for the limiting Brownian Bridge process. By Kolmogorov's theorem, since the finite dimensional distributions of B are all Gaussian, it follows that the Brownian bridge is a mean 0 Gaussian process with covariance function $\text{cov}(B(x_i), B(x_j)) = F(\min(x_i, x_j)) - F(x_i)F(x_j)$.

- L. Suppose the function $f(x|\theta)$ is a) a Borel measurable function of $x \in R^1$ and a continuously differentiable function of θ in a compact set $\Theta \subset R^K$. Suppose $\{\tilde{x}_n\}$ are IID random variables and that $f(x|\theta) < M$ for some constant $M < \infty$ for all $x \in R^1$ and all $\theta \in \Theta$. Consider the quantity $\bar{f}_N(\theta)$ given by

$$\bar{f}_N(\theta) = \frac{1}{N} \sum_{n=1}^N f(\tilde{x}_n|\theta). \quad (10)$$

Does $\bar{f}_N(\theta)$ converge almost surely? with probability 1? in probability? in distribution? uniformly with probability 1? If your answer to any of the above is yes, specify what the limiting quantity is and whether it is deterministic or random. If you find it converges uniformly with probability 1 to something, sketch a proof for why it converges uniformly with probability 1 to this limiting quantity.

Answer Since $\{\tilde{x}_n\}$ are IID $\bar{f}_N(\theta)$ converges almost surely (with probability one) to by SLLN $\bar{f}_N(\theta) \xrightarrow{a.s} Ef(x|\theta) = \int f(x|\theta)f(x|\theta)dx$. Since $f(x|\theta)$ continuously differentiable function of θ in a compact set Θ we have $L = \sup_{\theta \in \Theta} \left| \frac{\partial \bar{f}_N(\theta)}{\partial \theta} \right| < \infty$. By Taylor expansion: $\bar{f}_N(\theta') = \bar{f}_N(\theta) + \frac{\partial \bar{f}_N(\theta)}{\partial \theta} \Big|_{\theta=\tilde{\theta}} (\theta' - \theta)$ such that $\tilde{\theta}$ between θ and θ' . Further $|\bar{f}_N(\theta') - \bar{f}_N(\theta)| = \left| \frac{\partial \bar{f}_N(\theta)}{\partial \theta} \Big|_{\theta=\tilde{\theta}} (\theta' - \theta) \right| \leq \left| \frac{\partial \bar{f}_N(\theta)}{\partial \theta} \Big|_{\theta=\tilde{\theta}} \right| |\theta' - \theta| \leq L |\theta' - \theta|$. Hence $\bar{f}_N(\theta)$ is Lipschitz continuous. A Lipschitz continuous function defined on a compact space is equicontinuous and by the Arzela-Ascoli theorem an equicontinuous function defined on a compact space and bounded converges uniformly with probability one. Since $\bar{f}_N(\theta)$ is bounded, $f(x|\theta) < M$, it converges uniformly w.p.1 to $Ef(x|\theta)$.

- M. Consider the histogram estimator $\bar{H}_N(x)$ of a sequence of random variables $\{\tilde{x}_n\}$ at a point x

$$\bar{H}_N(x) = \frac{1}{N} \sum_{n=1}^N I\{x - 1/N \leq \tilde{x}_n \leq x + 1/N\} \quad (11)$$

Does $H_N(x)$ converge almost surely? with probability 1? in probability? in distribution? uniformly with probability 1? If your answer to any of the above is yes, specify what the limiting

quantity is and whether it is deterministic or random. If you find it converges uniformly with probability 1 to something, sketch a proof for why it converges uniformly with probability 1 to this limiting quantity.

Answer It converges pointwise to the probability density function of \tilde{x}_n : $\overline{H}_N(x) \xrightarrow{a.s.} F(x + 1/N) - F(x - 1/N) \rightarrow 0$, as $N \rightarrow \infty$

N. Consider the histogram estimator $\overline{H}_N(x)$ of a sequence of random variables $\{\tilde{x}_n\}$ at a point x

$$\overline{H}_N(x) = \frac{1}{N} \sum_{n=1}^N I\{x - a \leq \tilde{x}_n \leq x + b\}, \quad (12)$$

where a and b are two fixed *constants*. Does $H_N(x)$ converge almost surely? with probability 1? in probability? in distribution? uniformly with probability 1? If your answer to any of the above is yes, specify what the limiting quantity is and whether it is deterministic or random. If you find it converges uniformly with probability 1 to something, sketch a proof for why it converges uniformly with probability 1 to this limiting quantity.

Answer We know that $\frac{1}{N} \sum_{n=1}^N I\{x - a \leq \tilde{x}_n \leq x + b\} = \frac{1}{N} \sum_{n=1}^N I\{\tilde{x}_n \leq x + b\} - \frac{1}{N} \sum_{n=1}^N I\{\tilde{x}_n \leq x - a\}$. Since $\frac{1}{N} \sum_{n=1}^N I\{\tilde{x}_n \leq x + b\} \xrightarrow{a.s.} F(x + b)$ and $\frac{1}{N} \sum_{n=1}^N I\{\tilde{x}_n \leq x - a\} \xrightarrow{a.s.} F(x - a)$, we have $\overline{H}_N(x) \xrightarrow{a.s.} F(x + b) - F(x - a)$.

O. Consider the sequence $\{\overline{f}_N(\hat{\theta}_N)\}$ where \overline{f}_N is the function defined in part L above, and $\hat{\theta}_N$ is given by

$$\hat{\theta}_N = \theta^* + \tilde{u}_N/N \quad (13)$$

where $\{\tilde{u}_N\}$ is an *IID* sequence of $K \times 1$ $U(0, 1)$ random vectors (i.e. each component of the vector u_N is a $U(0, 1)$ random variable and the component random variables are distributed independently of each other). Further assume that a point $\theta^* \in \Theta$ is chosen so that with probability 1, we have $\hat{\theta}_N \in \Theta$ for all N . Does $\{\overline{f}_N(\hat{\theta}_N)\}$ converge almost surely? with probability 1? in probability? in distribution? or uniformly with probability 1? If the answers to any of these are yes, specify what the limiting quantity is and sketch an argument/proof of how you can show that the limiting quantity is what you claim it to be.

Answer $\left| \overline{f}_N(\hat{\theta}_N) - f(\theta^*) \right| = \left| \overline{f}_N(\hat{\theta}_N) - f(\hat{\theta}_N) + f(\hat{\theta}_N) - f(\theta^*) \right| \leq \left| \overline{f}_N(\hat{\theta}_N) - f(\hat{\theta}_N) \right| + \left| f(\hat{\theta}_N) - f(\theta^*) \right| \xrightarrow{a.s.} 0$ since $\overline{f}_N(\theta) \rightarrow f(\theta)$ and $\hat{\theta}_N \xrightarrow{a.s.} \theta^*$ so that $\overline{f}_N(\hat{\theta}_N) \rightarrow f(\theta^*)$. Therefore $\overline{f}_N(\hat{\theta}_N)$ converges uniformly with probability one to $f(\theta^*)$.

P. Show how your answer to part O above changes if instead we define $\hat{\theta}_N$ by

$$\hat{\theta}_N = \underset{\theta \in \Theta}{\operatorname{argmax}} \overline{f}_N(\theta). \quad (14)$$

What is the limiting quantity in this case?

Answer As we proved that $\overline{f}_N(\theta)$ converges uniformly with probability one to $f(\theta)$ on a compact set and if $\hat{\theta} \xrightarrow{a.s.} \theta^*$ then $\overline{f}_N(\hat{\theta}_N) \rightarrow f(\theta^*)$.

II. If we have N *IID* observations, $\{X_1, \dots, X_N\}$ from an unknown distribution $F(x)$ and we are interested in estimating its mean, $\mu = \int xF(dx)$.

A. Is the *sample mean*

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i \tag{15}$$

a non-parametric estimator of μ ?

Answer Yes the sample mean is a non-parametric estimator since we do not have to make any assumptions about the functional form of the distribution $F(x)$ to implement this estimator, and as long as the distribution has finite mean and variance, the sample mean will be consistent estimator of the true mean, $\mu = \int xF(dx)$.

B. Use the law of large numbers and the central limit theorem to determine the *asymptotic properties* of $\hat{\mu}$. Under what conditions is $\hat{\mu}$ a *consistent estimator* of μ ? What is the asymptotic distribution of $\hat{\mu}$?

Answer Under the IID assumption, we can apply the Law of Large Numbers and the Central Limit Theorem to determine the asymptotic properties of $\hat{\mu}$. The LLN implies that $\hat{\mu}$ converges to μ almost surely (i.e. with probability 1 as $N \rightarrow \infty$), and the CLT implies that

$$\sqrt{N}(\hat{\mu} - \mu) \Rightarrow N(0, \sigma^2),$$

where $\sigma^2 = \text{var}(X_i) = \int (x - \mu)^2 F(dx)$. These results will hold as long as μ and σ^2 are finite. (Check the assumptions required to prove the LLN and CLT).

C. Suppose that the true distribution $F(x)$ is a Normal distribution with mean μ and variance σ^2 . What is the *exact finite sample distribution* of $\hat{\mu}$?

Answer Since sums of normal random variables are normal, we have

$$\hat{\mu} \sim N(0, \sigma^2/N)$$

where the notation is a shortcut for “distributed as a Normal random variable with mean zero and variance σ^2/N .”

D. Suppose $F(x)$ is a *Cauchy distribution*. Can you determine the exact finite sample distribution of $\hat{\mu}$ in this case? If so (or even if not) can you determine whether the sample mean $\hat{\mu}$ is a consistent estimator of μ in this case?

Answer We can use characteristic functions to show that for each N $\hat{\mu}$ has the same distribution as \tilde{X}_1 i.e. it is the same Cauchy distribution as each of the observations. To see this, note that the density of a Cauchy distribution with location parameter μ and scale parameter σ is

$$f(x|\mu, \sigma) = \frac{1}{\pi\sigma \left(1 + \frac{(x-\mu)^2}{\sigma^2}\right)},$$

and the characteristic function for a Cauchy distribution is

$$\Psi_{\tilde{X}_1}(t) = E\{\exp(it\tilde{X})\} = \int_{-\infty}^{\infty} \left[\frac{\exp\{itx\}}{\pi\sigma \left(1 + \frac{(x-\mu)^2}{\sigma^2}\right)} \right] dx = \exp\{i\mu t - \sigma|t|\}.$$

Now using the fact that the characteristic function of a *sum of IID* random variables is the product of the individual characteristic function (i.e.

$$E\{\exp(it \sum_{j=1}^N \tilde{X}_j)\} = \prod_{j=1}^N E\{\exp(it \tilde{X}_j)\} = [\Psi_{\tilde{X}}(t)]^N.$$

In this case of the Cauchy we have the characteristic function of the sum is

$$[\phi_{\tilde{X}}(t)]^N = \exp\{iN\mu t - N\sigma |t|\}.$$

Dividing the sum by N just amounts recaling, and you can easily work out that if \tilde{X} is a Cauchy with parameters (μ, σ) , then for any scalar λ we have $\lambda\tilde{X}$ is a Cauchy with parameters $(\lambda\mu, \lambda\sigma)$ since its characteristic function is

$$\Psi_{\lambda\tilde{X}}(t) = E\{\exp(it\lambda\tilde{X})\} = \Psi_{\tilde{X}}(\lambda t) = \exp\{i\lambda\mu t - \lambda\sigma |t|\}$$

So substituting $\lambda = \frac{1}{N}$, we find that the characteristic function of $\hat{\mu}$ is

$$\Psi_{\hat{\mu}}(t) = \Psi_{\tilde{X}}(t)$$

which implies that for each N $\hat{\mu}$ is distributed as a Cauchy with (μ, σ) . So the finite sample distribution of the sample mean is a Cauchy with parameters (μ, σ) . Clearly this cannot be a consistent estimator for μ since for every N , and even in the limit $N \rightarrow \infty$, we have $\hat{\mu}$ is always a Cauchy distribution. A consistent estimator must converge to μ with probability 1, but this estimator does not converge to anything with probability 1, it is always a Cauchy random variable, and so it only converges in distribution, but not almost surely. Can you determine what distribution $\hat{\mu}$ converges to?

- E. Suppose that instead of trying to estimate μ we are interested in estimating the distribution $F(x)$ itself. How could we estimate this distribution non-parametrically?

Answer We can estimate $F(x)$ using the *empirical distribution function* $\hat{F}(x)$ given by

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^N I\{X_i \leq x\}.$$

- F. Use the law of large numbers and the central limit theorem to determine the *asymptotic properties* of the estimator $\hat{F}(x)$ given by

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^N I\{X_i \leq x\} \tag{16}$$

Under what conditions is $\hat{F}(x)$ a consistent estimator of $F(x)$ and what is its asymptotic distribution?

Answer The conditions to establish consistency and asymptotic normality of the empirical distribution are significantly weaker than the conditions necessary to establish consistency and asymptotic normality of the sample mean. The reason is that the random variables entering the sum defining $\hat{F}(x)$, $I\{\tilde{X}_i \leq x\}$, are Bernoulli random variables and thus bounded, (in particular, they take only two possible values, 0 and 1), even if the random variables \tilde{X}_i have no finite moments, such as the Cauchy distribution. The $B_i(x) = I\{X_i \leq x\}$. Clearly we have

$$E\{B_i(x)\} = F(x)$$

$$\text{var}\{B_i(x)\} = F(x)(1 - F(x)).$$

The latter results follows because for a Bernoulli random variable which takes on the value $B_i(x) = 1$ with probability

$$p = F(x) = \Pr\{I\{\tilde{X}_i \leq x\} = 1\} = \Pr\{\tilde{X}_i \leq x\}.$$

It follows that the mean of the Bernoulli is just $p = F(x)$ and the variance of the Bernoulli is $p(1 - p) = F(x)(1 - F(x))$, since we have

$$\text{var}(B_i(x)) = E\{(B_i(x) - p)^2\} = E\{[B(x_i)^2]\} - [E\{B(x_i)\}]^2 = p(1 - p) = F(x)(1 - F(x)).$$

Thus, $\hat{F}(x)$ is an unbiased estimator of $F(x)$ and it is asymptotically normal, with asymptotic variance $\sigma^2(x) = F(x)(1 - F(x))$. In summary, the LLN implies that $\hat{F}(x)$ converges to $F(x)$ with probability 1, and that

$$\sqrt{N}[\hat{F}(x) - F(x)] \rightarrow N(0, F(x)[1 - F(x)])$$

under very weak conditions, which are just that $\{X_i\}$ IID sample from the distribution F , but F is not required to have any finite moments for this to hold.

- G. Suppose $F(x)$ is known to be a normal distribution, but with an unknown mean μ and variance σ^2 . What is a reasonable *parametric estimator* of the quantity $F(x)$ at some point x ?

Answer Since a normal distribution is entirely determined by the two parameters (μ, σ) , we can estimate just these two parameters by their “sample analogs” the sample mean and standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, and these estimators turn out to be the same as the maximum likelihood estimators of these parameters as we will see below. Then the natural “parametric estimator” of $F(x)$ would be

$$F(x, \hat{\theta}) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)$$

where Φ is the standard normal CDF, where $\hat{\theta}' = (\hat{\mu}, \hat{\sigma})'$.

- H. Compare the *asymptotic variance* of the estimator of $F(x)$ in part 6 with the estimator you determined in part 7. Which of these is a more *efficient estimator*?

Answer We would expect the parametric estimator $F(x, \hat{\theta})$ given above to be more efficient than the non-parametric estimator (the empirical CDF, $\hat{F}(x)$), since the parametric estimator uses the additional prior information that the data are draws from a normal distribution. Thus, we

would expect that the asymptotic variance of $F(x, \hat{\theta})$ to be lower than the asymptotic variance of $\hat{F}(x)$, which is just

$$\Phi\left(\frac{x-\mu}{\sigma}\right)\left[1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right] \quad (17)$$

in this case (since the true distribution is $N(\mu, \sigma)$ whose CDF is $\Phi\left(\frac{x-\mu}{\sigma}\right)$). To compute the asymptotic distribution of $F(x, \hat{\theta})$ we need to apply the delta theorem. In its general form, it states the following: **[Theorem]** *Suppose that $\sqrt{N}[\hat{\theta} - \theta] \rightarrow N(0, \Sigma)$ and that $H(\theta)$ is a continuously differentiable function of θ . Then we have*

$$\sqrt{N}[H(\hat{\theta}) - H(\theta)] \rightarrow N(0, \nabla H(\theta)\Sigma\nabla H(\theta)'),$$

where $\nabla H(\theta)$ is the gradient of $H(\theta)$ with respect to θ .

- **Proof.** Expand $H(\theta)$ in a Taylor series about θ to get

$$H(\hat{\theta}) = H(\theta) + \nabla H(\tilde{\theta})(\hat{\theta} - \theta),$$

where $\tilde{\theta}$ is a point on the line segment joining $\hat{\theta}$ and θ . Since $\sqrt{N}(\hat{\theta} - \theta)$ converges in distribution, it is easy to see that $\hat{\theta}$ itself must converge with probability 1 to θ . So the continuous mapping theorem implies that $\nabla H(\tilde{\theta})$ converges in probability to $\nabla H(\hat{\theta})$ as

$$\sqrt{N}(H(\hat{\theta}) - H(\theta)) = \nabla H(\tilde{\theta})\sqrt{N}(\hat{\theta} - \theta)$$

the continuous mapping theorem and the continuity of $\nabla H(\theta)$ in θ imply that the right hand side of the above equation converges in distribution to

$$\nabla H(\tilde{\theta})\sqrt{N}[\hat{\theta} - \theta] \rightarrow N(0, \Omega)$$

where Ω is given by

$$\Omega = \nabla H(\theta)\Sigma\nabla H(\theta)'$$

It follows that $\sqrt{N}(H(\hat{\theta}) - H(\theta))$ has the same asymptotic distribution, and thus the asymptotic covariance matrix of $H(\hat{\theta})$ is Ω . Now consider the application at hand: we have $H(\theta) = \Phi(x, \theta) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, where $\theta' = (\mu, \sigma)$. So in this case $\nabla H(\theta)$ is given by

$$\nabla H(\theta)' = \begin{bmatrix} \frac{\partial}{\partial \mu} \Phi\left(\frac{x-\mu}{\sigma}\right) \\ \frac{\partial}{\partial \sigma} \Phi\left(\frac{x-\mu}{\sigma}\right) \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \\ -\frac{(x-\mu)}{\sigma^2} \phi\left(\frac{x-\mu}{\sigma}\right) \end{bmatrix}.$$

Now the only remaining issue is to determine the asymptotic covariance matrix Σ of $\hat{\theta} = (\hat{\mu}, \hat{\sigma})'$. This is not too hard to do, either directly, by applying the CLT to the formulas for $\hat{\mu}$ and $\hat{\sigma}$, or by using the fact that there are the maximum likelihood estimators of the true parameters (μ, σ) . We know that asymptotically, if $\hat{\theta}$ is a maximum likelihood estimator that

$$\sqrt{N}(\hat{\theta} - \theta) \rightarrow N(0, [I(\theta)]^{-1}),$$

where $I(\theta)$ is the information matrix given by

$$I(\theta) = E \left\{ \begin{bmatrix} \frac{\partial}{\partial \theta} \log f(\tilde{x}|\theta) \\ \frac{\partial}{\partial \theta} \log f(\tilde{x}|\theta)' \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial \theta} \log f(\tilde{x}|\theta) \\ \frac{\partial}{\partial \theta} \log f(\tilde{x}|\theta)' \end{bmatrix} \right\}.$$

In this case, $\log f(\tilde{x}|\theta)$ is given by

$$\log f(\tilde{x}|\theta) = \frac{-1}{2} \log(2\pi) - \log(\sigma) - \frac{1}{2} \left[\frac{x - \mu}{\sigma} \right]^2.$$

Calculating the derivatives of this with respect to μ and σ and taking expectations, we get

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}$$

so it follows that $\Sigma = [I(\theta)]^{-1} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}$. Now, using this formula for Σ and the above for $\nabla H(\theta)$, we can derive the “sandwich formula” for Ω in this case

$$\Omega = \nabla H(\theta) \Sigma \nabla H(\theta)' = \phi \left(\frac{x - \mu}{\sigma} \right)^2 \left[1 + \frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]. \quad (18)$$

It may not be immediately evident, comparing the formulas for the asymptotic variances of the maximum likelihood estimator of $F(x)$ given in the equation (18) above to the formula in (17), which of the two is larger? Intuition would suggest that the MLE should have a smaller asymptotic variance since it is supposed to be “asymptotically efficient” and makes use of all available prior information, which in this case is the fact that the data $\{\tilde{X}_i\}$ are normally distributed. The empirical CDF does not require any such assumption, and thus does not exploit the prior information that the data are normally distributed. So intuition suggests that for any x and for any (μ, σ) the asymptotic variance of the MLE should be lower than the asymptotic variance of the empirical CDF. To show this, first note that x and (μ, σ) enter both of these asymptotic variances via the standardized value z given by

$$z = \frac{x - \mu}{\sigma}.$$

Thus, it suffices to show that $\phi(z)^2 [1 + z^2/2] < \Phi(z) [1 - \Phi(z)]$ for all z . I leave the mathematical proof of this to you.

- I. Suppose $F(x)$ is a Cauchy distribution. Does knowing this fact affect your ability to consistently estimate $F(x)$ using either the estimator $\hat{F}(x)$ in part 6, or some other “parametric estimator” where you make use of your prior information that $F(x)$ is a Cauchy distribution?

Answer The answer is no. As we discussed in the answer to part 6, the empirical CDF $\hat{F}(x)$ still has the same distribution in the Cauchy case, even though the mean and variance of Cauchy distribution do not exist. Furthermore, since the parameters of the Cauchy are $\theta = (\mu, \sigma)'$ the same parametric estimator works in this case, i.e., $F(x, \hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimator of θ . One can use the delta theorem to calculate the asymptotic distribution of the parametric $F(x, \hat{\theta})$ just as in the normal case, and the same reasoning leads to the conclusion that the parametric estimator has lower asymptotic covariance matrix than the non-parametric estimator.

- J. Suppose instead of just trying to estimate $F(x)$ at a single point x , we want to estimate it at a *vector of points* (x_1, x_2, \dots, x_m) (note: the lower case x_i are *fixed points* and should

be distinguished from the *observations* which are upper case letters, e.g. X_j). Consider the $m \times 1$ vector estimator $(\hat{F}(x_1), \dots, \hat{F}(x_m))$. Using a multivariate version of the central limit theorem, describe the asymptotic distribution of this vector estimator of F at these m points (x_1, \dots, x_m) .

Answer We apply a multivariate version of the central limit theorem, stacking the empirical CDF estimator at the m points as $m \times 1$ vector we denote as $\hat{F}(x)$, where x is not interpreted as the $(m \times 1)$ vector of points $x = (x_1, \dots, x_m) \in \mathfrak{R}^m$. $\hat{F}(x)$ is now just a normalized sum of vectors of indicator functions and it is easy to see that

$$\sqrt{N}[\hat{F}(x) - F(x)] \rightarrow N(0, \Sigma)$$

where Σ is the $(m \times m)$ covariance matrix of the vector of indicators functions

$$\Sigma = \text{var}(I(\tilde{x} \leq x)) = \text{var} \begin{pmatrix} I(\tilde{x} \leq x_1) \\ I(\tilde{x} \leq x_2) \\ \dots \\ I(\tilde{x} \leq x_{m-1}) \\ I(\tilde{x} \leq x_m) \end{pmatrix}.$$

It is easy to see that i^{th} diagonal entry of Σ is just $F(x_i)[1 - F(x_i)]$ since this is just the “marginal” asymptotic variance of $\hat{F}(x_i)$, the empirical CDF evaluated at the single point x_i which we calculated above. So the only thing to worry about is the covariance between two indicators,

$$\Sigma_{ij} = \text{cov}(I(\tilde{x} \leq x_i), I(\tilde{x} \leq x_j)).$$

But recall that the covariance of two random variables is defined as

$$\text{cov}(\tilde{X}, \tilde{Y}) = E\{\tilde{X}\tilde{Y}\} - E\{\tilde{X}\}E\{\tilde{Y}\}.$$

We know that for the two Bernoulli's $B(x_i)B(x_j)$ is just another Bernoulli that takes the value 1 if both the events happen, i.e. if $\tilde{x} \leq x_i$ and $\tilde{x} \leq x_j$. But both of these will happen if and only if $\tilde{x} \leq \min(x_i, x_j)$. Thus we conclude that

$$E\{B(x_i)B(x_j)\} = E\{I(\tilde{x} \leq x_i)I(\tilde{x} \leq x_j)\} = E\{\tilde{x} \leq \min(x_i, x_j)\}$$

since the product of the two Bernoulli's $B(x_i)B(x_j)$ is just another Bernoulli that takes the value 1 if both the events happen, i.e. if $\tilde{x} \leq x_i$ and $\tilde{x} \leq x_j$. But both of these will happen if and only if $\tilde{x} \leq \min(x_i, x_j)$. Thus we conclude that

$$E\{B(x_i)B(x_j)\} = E\{I(\tilde{x} \leq x_i)I(\tilde{x} \leq x_j)\} = E\{\tilde{x} \leq \min(x_i, x_j)\} = F(\min(x_i, x_j))$$

or

$$\Sigma_{ij} = F(\min(x_i, x_j)) - F(x_i)F(x_j).$$

K. **Extra credit, harder question.** Can you write a formula for the *exact finite-sample distribution* of $\hat{F}(x)$ where x is a single (scalar) point in the support of $F(x)$?

Answer Note that $N\hat{F}(x)$ is a sum of independent Bernoulli random variables. Sum of N independent Bernoulli random variables have a binomial distribution (you can show this easily using characteristic functions, for example). Thus $\hat{F}(x)$ will take one of the $N+1$ values $(0, \frac{1}{N}, \frac{2}{N}, \dots, \frac{N-1}{N}, 1)$ with probability given by

$$\Pr\{\hat{F}(x) = \frac{j}{N}\} = \binom{N}{j} F(x)^j [1 - F(x)]^{N-j}$$

III. Consider *parametric estimation* by the method of *maximum likelihood*. Consider the general case where $\{X_1, \dots, X_N\}$ are N IID observations from a *density* $f(x|\theta)$ where θ is a $K \times 1$ vector of unknown parameters to be estimated.

A. Define in the general case, the *maximum likelihood estimator* of θ .

Answer The likelihood function is defined $f(x_1, x_2, \dots, x_N|\theta) = \prod_{i=1}^N f(x_i|\theta) = L(\theta|x)$ and the log likelihood function is $\ln L(\theta|x) = \sum_{i=1}^N \ln f(x_i|\theta)$. The maximum likelihood estimator:

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} L(\theta|x)$$

or equivalently

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \ln L(\theta|x)$$

The FOC are: $\frac{\partial \ln L(\hat{\theta}|x)}{\partial \theta} = 0 \Rightarrow \sum_{i=1}^N \frac{f'(x_i|\hat{\theta})}{f(x_i|\hat{\theta})} = 0 \Rightarrow \hat{\theta}$.

B. Suppose that θ is the 2×1 vector $\theta = (\mu, \sigma)$ and $f(x|\theta)$ is given by

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty. \quad (19)$$

Can you provide explicit formulas for the maximum likelihood estimators of μ and σ in this case?

Answer For the log likelihood $\ln L(\mu, \sigma|x) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^N \frac{(x_i - \mu)^2}{\sigma^2}$ the FOCs

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}_N$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \hat{\mu})^2 = 0 \Rightarrow \hat{\sigma} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}_N)^2}.$$

C. Using the law of large numbers and central limit theorem, can you describe the *asymptotic properties* of the maximum likelihood estimator $\hat{\theta}$ in the general case? What “regularity conditions” do you need to impose on $f(x|\theta)$ in order to establish these properties?

Answer Regularity conditions: a) The first three derivatives of $\ln f(x_i|\theta)$ with respect to θ are continuous and finite for almost all x_i and for all θ ; b) the conditions necessary to obtain the expectations of the first and second derivatives of $\ln f(x_i|\theta)$ are met; c) For all values of θ , $|\partial^3 \ln f(x_i|\theta) / \partial \theta_i \partial \theta_j \partial \theta_k|$ is less than a function that has a finite expectation. Under regularity conditions we have: 1) $\operatorname{plim} \hat{\theta} = \theta$; 2) $\hat{\theta} \sim^d N(0, \{I(\theta)^{-1}\})$ where $I(\theta) = -E[\partial^2 \ln L / (\partial \theta \partial \theta')]$ 3) $\hat{\theta}$ is asymptotically efficient and achieves the Cramer-Rao lower bound 4) The maximum likelihood estimator of $\gamma = c(\theta)$ is $c(\hat{\theta})$ is $c(\theta)$ is continuous and continuously differentiable.

D. Now consider the specific case in part 2, where $f(x|\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ . Can you be more specific about the asymptotic distribution in this case?

Answer By WLLN $\hat{\mu} \rightarrow \mu$ and CLT $\hat{\mu} \rightarrow N(\mu, \frac{\sigma^2}{N})$ respectively $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $\hat{\sigma}^2 \xrightarrow{d} N(\sigma^2, \frac{2\sigma^4}{N})$.

E. What if $f(x|\theta)$ is a *Cauchy distribution* given by

$$f(x|\theta) = \frac{1}{\pi [1 + (x - \theta)^2]}. \quad (20)$$

Can you write down what the maximum likelihood estimator is for θ and determine its asymptotic properties in this case? In particular, is the maximum likelihood estimator consistent and asymptotically normal?

Answer $\hat{\theta} = \operatorname{argmax} \ln L(\theta|x)$ where $\ln L(\theta|x) = \sum_{i=1}^N \ln \left(\frac{1}{\pi[1+(x_i-\theta)^2]} \right) = -\sum_{i=1}^N \ln [\pi(1 + (x_i - \theta)^2)]$.
FOC: $\frac{\partial \ln L(\theta)}{\partial \theta} = \sum_{i=1}^N \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} = 0$. There are multiple roots exist in this equation. When N goes large, typically it is impossible to find an analytical solution. Numerical methods should be used. The location parameter of the Cauchy distribution satisfies the regularity conditions the MLE is consistent and asymptotically normal.

IV. Consider the one unknown parameter regression model

$$y_t = a^* + \varepsilon_t \quad (21)$$

where $\{\varepsilon_t\}$ is an *i.i.d* $U[-1, 1]$ random variable, i.e. the error terms are uniformly distributed on the interval $[-1, 1]$. a^* is the unknown parameter to be estimated.

1. What is the OLS estimator of the parameter a , and what is its asymptotic distribution?

Answer $\hat{a} = \frac{1}{N} \sum_{i=1}^T y_t$, so $a \sim^a N(a^*, \frac{1}{3T})$

2. What is the maximum likelihood estimator for a and describe its properties. Discuss whether the maximum likelihood estimator is well-defined or not, i.e. whether there is generally a unique value of \hat{a}_{mle} that maximizes the likelihood. Sketch what the likelihood looks like in an example case.

Answer The likelihood $L(a|y_1, \dots, y_T) = \prod_{i=1}^T f(y_i|a) = \frac{1}{2^T}$ for $y_t \in [-1, 1]$ and zero otherwise. The likelihood is flat and the parameter a is unidentified, the likelihood is only positive for $y_{(T)} - 1 \leq a \leq y_{(1)} + 1$, where $y_{(T)} = \max(y_1, \dots, y_T)$ and $y_{(1)} = \min(y_1, \dots, y_T)$.

3. Consider the estimator $\hat{a} = \max(y_1, \dots, y_T) - 1$. Is this a consistent estimator of the true parameter a ? Derive the asymptotic distribution of this estimator and compare it to the OLS estimator. Is this estimator unbiased? Is it consistent?

Answer $\Pr(\hat{a} \leq x) = \Pr(\max\{y_1, \dots, y_T\} - 1 \leq x) = [F(x + 1)]^T = \left[\frac{1}{2}(x + 2 - a^*)\right]^T$ with support $\hat{a} \in [a^* - 2, a^*]$. To show \hat{a} is consistent we need to show

$$\lim_{T \rightarrow \infty} \Pr(|\hat{a} - a^*| > \varepsilon) = 0.$$

since $\Pr(|\hat{a} - a^*| > \varepsilon) = 1 - \Pr(|\hat{a} - a^*| \leq \varepsilon) = 1 - [\Pr(\hat{a} - a^* \leq \varepsilon) - \Pr(\hat{a} - a^* \geq -\varepsilon)] = 2 - \Pr(\hat{a} - a^* \leq \varepsilon) - \Pr(\hat{a} - a^* \leq -\varepsilon) = 2 - \left[\frac{1}{2}(a^* + \varepsilon + 2 - a^*)\right]^T - \left[\frac{1}{2}(a^* - \varepsilon + 2 - a^*)\right]^T \rightarrow 0$ as $T \rightarrow \infty$ for $\varepsilon > 0$ therefore $\hat{a} \rightarrow a^*$.

4. Derive the asymptotic distribution of the estimator suggested in part 3 above. Which estimator is preferable in terms of asymptotic efficiency, the OLS estimator or the estimator in part 3? How do both of these estimators compare to the maximum likelihood estimator?

Answer $T(\hat{a} - a^*) \rightarrow^d e^{x/2}$ because $\Pr(T(\hat{a} - a^*) \leq x) = \Pr(\hat{a} \leq \frac{x}{T} + a^*) = [F(\frac{x}{T} + a^* + 1)]^T = [\frac{1}{2}(2 + \frac{x}{T})]^T = [1 + \frac{x/2}{T}]^T \rightarrow e^{x/2}$ as $T \rightarrow \infty$. This MLE is superconvergent, but it has a downward bias.

V. Derive the asymptotic distribution for $\hat{\beta}_N$ based on N observations (y_i, X_i) , $i = 1, \dots, N$ where β is the *maximum likelihood* estimator in the model

$$y_i = X_i\beta^* + u_i \quad (22)$$

where $\{u_i\}$ are *i.i.d. double exponential* random variables, i.e. their density is $f(u) = \exp\{-|u|/\sigma\}/2\sigma$ for $u \in \mathbb{R}^1$. Compare the asymptotic distribution of the maximum likelihood estimator for this model with the OLS estimate and derive formulas for the asymptotic variance covariance for β in each case. Which estimator is more efficient?

Answer The log likelihood function when the errors are double exponential is:

$$\ln L(\beta, \sigma) = -N \ln(2\sigma) + \frac{1}{\sigma} \sum_{i=1}^N |y_i - X_i\beta|.$$

The $\hat{\beta}_{MLE} = \operatorname{argmin}_{\beta} \sum_{i=1}^N |y_i - X_i\beta|$ with the FOC $\sum_{i=1}^N \operatorname{sign}(y_i - X_i\hat{\beta}) X_i = 0$ is the LAD estimator with the asymptotic distribution

$$\sqrt{N}(\hat{\beta}_{MLE} - \beta^*) \rightarrow N\left(0, \left(\frac{1}{2f(0)}\right)^2 \left(\frac{X'X}{N}\right)^{-1}\right)$$

where $f(0) = 1/(2\sigma)$, so

$$\sqrt{N}(\hat{\beta}_{MLE} - \beta^*) \rightarrow N\left(0, \sigma^2 \left(\frac{X'X}{N}\right)^{-1}\right)$$

while the OLS

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta^*) \rightarrow N\left(0, 2\sigma^2 \left(\frac{X'X}{N}\right)^{-1}\right).$$

Therefore MLE more efficient.

VI. Derive the asymptotic distribution of the *sample median* $\operatorname{med}(\tilde{y}_1, \dots, \tilde{y}_N)$ of N *i.i.d.* random variables $(\tilde{y}_1, \dots, \tilde{y}_N)$ where the density of these random variables, $f(y)$ is symmetric about $y = 0$ and satisfies $f(0) > 0$. **Hint:** Use the fact that the median can be defined as the 0.5 *quantile* of the empirical distribution function, $F_N(\operatorname{med}(\tilde{y}_1, \dots, \tilde{y}_N)) = 0.5$ where

$$F_N(x) = \frac{1}{N} \sum_{i=1}^N I\{\tilde{y}_i \leq x\} \quad (23)$$

Let $F(y)$ be the CDF of the \tilde{y} random variables, i.e. the integral of the density function f . For any $\theta \in (0, 1)$ define μ_θ as the solution to $F(\mu_\theta) = \theta$. If the random variable \tilde{y} has a density and if $f(y) > 0$, then its CDF is strictly increasing and hence is invertible, so we can also define $\mu_\theta = F^{-1}(\theta)$, where F^{-1} is the *inverse* of the CDF F . Or more generally, even if F has jumps (discontinuities) we can define

$$\mu_\theta = \inf\{y | F(y) \geq \theta\}. \quad (24)$$

Via this definition we can also define the *sample quantiles* $\hat{\mu}_\theta$ as

$$\hat{\mu}_\theta = \inf\{y | F_N(y) \geq \theta\} \quad (25)$$

and the *median* $\text{med}(\tilde{y}_1, \dots, \tilde{y}_N)$ is just $\hat{\mu}_\theta$ for $\theta = 0.5$. Clearly, the result I am asking you to prove about the asymptotic distribution is just a special case of the following theorem that characterizes the asymptotic distribution of *all* sample quantiles μ_θ for $\theta \in (0, 1)$ (not that we do not include $\theta = 0$ or $\theta = 1$ since these are known as the *sample extremes* and the asymptotic distribution of the sample extremes is generally different — it converges at a different rate than \sqrt{N} and to a non-normal limiting distribution, generally to one of the three types of *extreme value* distributions).

Theorem: Let $(\tilde{y}_1, \dots, \tilde{y}_N)$ be *IID* draws from a CDF F with continuous density f . If f satisfies $f(\mu_\theta) > 0$ we have

$$\sqrt{N}(\hat{\mu}_\theta - \mu_\theta) = \sqrt{N}(F_N^{-1}(\theta) - F^{-1}(\theta)) \implies N(0, \sigma^2) \quad (26)$$

where

$$\sigma^2 = \frac{\theta(1-\theta)}{f(\mu_\theta)^2} \quad (27)$$

To prove this result (which gives you the proof of the asymptotic distribution of the sample median as a special case) first note that as per discussion in the class, The Central Limit Theorem for *IID* random variables implies that for any x in the support of F we have:

$$\sqrt{N}(F_N(x) - F(x)) \implies N(0, \gamma^2),$$

where $\gamma^2 = F(x)[1 - F(x)]$. The following lemma, a slightly modified version of a lemma from R. J. Serfling, (1980) *Approximation Theorems of Mathematical Statistics* Wiley, New York, provides some basic properties of the *quantile function* $F^{-1}(\theta)$:

Lemma 1: Let F be a CDF. The quantile function $F^{-1}(\theta)$, $\theta \in (0, 1)$ is non-decreasing and left continuous, and satisfies:

1. $F^{-1}(F(x)) \leq x$, $-\infty < x < \infty$
2. $F(F^{-1}(\theta)) \geq \theta$, $0 < \theta < 1$
3. If F is strictly increasing in a neighborhood of $\mu_\theta = F^{-1}(\theta)$ we have: $F(F^{-1}(\theta)) = \theta$ and $F^{-1}(F(\mu_\theta)) = \mu_\theta$.
4. $F(x) \geq \theta$ if and only if $x \geq F^{-1}(\theta)$.

Let $x = \mu_\theta = F^{-1}(\theta)$. Using result 3 of Lemma 1 we have:

$$\sqrt{N}(F_N(\mu_\theta) - F(\mu_\theta)) = \sqrt{N}(F_N(F^{-1}(\theta)) - F(F^{-1}(\theta))) \implies N(0, \theta(1 - \theta)).$$

Since empirical CDF's have jumps of size $1/N$ (unless more than one of the $\{\tilde{X}_i\}$'s take the same value), then we can bound the maximum difference between θ and $F(F^{-1}(\theta))$ in Lemma 1-2 as follows:

Lemma 2: Let $(\tilde{X}_1, \dots, \tilde{X}_N)$ be a random sample from a CDF F and suppose that in this sample each \tilde{X}_i happens to be distinct, so that by reindexing we have $\tilde{X}_1 < \tilde{X}_2 < \dots < \tilde{X}_{N-1} < \tilde{X}_N$. Then for all $\theta \in (0, 1)$ we have:

$$|F_N(F_N^{-1}(\theta)) - \theta| \leq \frac{1}{N}.$$

Now, using the property of *stochastic equicontinuity* from the theory of *empirical processes* (see D. Andrews, (1996) *Handbook of Econometrics* (vol. 4) for an accessible introduction and definition of stochastic equicontinuity), we have the result given above is unaffected if we replace μ_θ by a consistent estimate $\hat{\mu}_\theta$:

$$\sqrt{N}(F_N(\hat{\mu}_\theta) - F(\hat{\mu}_\theta)) = \sqrt{N}(F_N(F_N^{-1}(\theta)) - F(F_N^{-1}(\theta))) \implies N(0, \theta(1 - \theta)).$$

Now note that part 2. of Lemma 1 implies that

$$\sqrt{N}(F_N(F_N^{-1}(\theta)) - F(F_N^{-1}(\theta))) \geq \sqrt{N}(\theta - F(F_N^{-1}(\theta))).$$

However since the true CDF F has a density, the probability of observing duplicate $\{\tilde{X}_i\}$'s is zero, so Lemma 2 implies that with probability 1 we have:

$$\sqrt{N}(F_N(F_N^{-1}(\theta)) - F(F_N^{-1}(\theta))) = \sqrt{N}(\theta - F(F_N^{-1}(\theta))) + O_p(1/\sqrt{N}),$$

which implies that:

$$\sqrt{N}(\theta - F(F_N^{-1}(\theta))) \implies N(0, \theta(1 - \theta)).$$

I have now laid out most of the key elements of the proof. See if you can complete the argument to establish the asymptotic distribution (**hint:** you can appeal to the *Delta Theorem* to establish the asymptotic distribution as a nonlinear transformation of a normalized sum of independent random variables).

VII Suppose f and g are two density functions with a unidimensional argument x . Define the *Kullback-Leibler distance* between f and g as $D(f, g)$ given by

$$D(f, g) = - \int \log \left(\frac{f(x)}{g(x)} \right) g(x) dx \tag{28}$$

Show that if $f(x) = g(x)$ for g -almost all x , then $D(f, g) = 0$. Further, show that if $f(x) \neq g(x)$ for a set of x that has positive probability under the density g , then $D(f, g) > 0$. **Hint:** use *Jensen's Inequality*.

[Answer] $D(f, g) = - \int \log \left(\frac{f(x)}{g(x)} \right) g(x) dx = -E \left\{ \log \left(\frac{f(x)}{g(x)} \right) \right\} \geq - \log E \left\{ \frac{f(x)}{g(x)} \right\} = - \log \int \frac{f(x)}{g(x)} g(x) dx = - \log \int f(x) dx = - \log 1 = 0$. If $f(x) = g(x)$ then $D(f, g) = - \int \log \left(\frac{f(x)}{g(x)} \right) g(x) dx = - \int \log \left(\frac{g(x)}{g(x)} \right) g(x) dx = \int \log(1) g(x) dx = 0$.

- A. Now suppose that the *true data generating process* i.e. the true distribution from which a random sample of data that we observe, (x_1, \dots, x_n) has a density g , which is unknown to us as the econometrician. However suppose we *assume* that the data generating process is some parametric model, $f(x|\theta)$ that depends on a $K \times 1$ vector of parameters θ and we restrict θ to a compact parameter space Θ and assume the usual regularity conditions apply (in particular assume that f is a continuously differentiable function of θ and that $\log(f(\tilde{x}|\theta))$ has finite expectation). We say that our assumed model is *misspecified* if there is no $\theta \in \Theta$ for which $f(x|\theta) = g(x)$ for g -almost all x (i.e. for any $\theta \in \Theta$ there is a set of x with positive probability under the true data generating distribution g for which $f(x|\theta) \neq g(x)$). Use this result above to show that for each $\theta \in \Theta$ we have $D(f(\cdot|\theta), g(\cdot)) > 0$.

Answer $D(f(x|\theta), g(x)) = -\int \log \frac{f(x|\theta)}{g(x)} g(x) dx$. If the model is misspecified then for any $\theta \in \Theta$ $f(x|\theta) \neq g(x)$ so that $\log \frac{f(x|\theta)}{g(x)} \neq 0$ and $\exists x$ s.t. $g(x) > 0$ so that $D(f(x|\theta), g(x)) \neq 0 \rightarrow D(f(x|\theta), g(x)) > 0$

- B. Now show that is the model is misspecified but if we try to estimate the misspecified model by the method of maximum likelihood, the maximum likelihood estimator $\hat{\theta}_N$ will converge to a parameter $\theta^* \in \Theta$ that satisfies

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} D(f(\cdot|\theta), g(\cdot)), \quad (29)$$

i.e. maximum likelihood will converge to the parameter θ^* that *minimizes the Kullbeck-Leibler distance between the true data generating process g and the parametric model $f(\cdot|\theta)$* . **Hint:** use the results above, or refer to Halbert White, (1982) "Maximum Likelihood Estimation of Misspecified Models" *Econometrica*.

Answer Define the quasi maximum likelihood estimator:

$$\hat{\theta}_{QMLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \log f(x_i|\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} l_N(\theta)$$

Under the assumptions: 1) $E \log f(x_i|\theta)$ exists and is finite; 2) $E \log f(x_i|\theta)$ continuous on Θ ; 3) $\log f(x_i|\theta)$ satisfies the ULLN then

$$\frac{1}{N} \sum_{i=1}^N \log f(x_i|\theta) \xrightarrow{w.p.1} E[\log f(x_i|\theta)]$$

therefore

$$\hat{\theta}_{QMLE} \xrightarrow{w.p.1} \underset{\theta \in \Theta}{\operatorname{argmax}} E[\log f(x_i|\theta)] = \underset{\theta \in \Theta}{\operatorname{argmax}} \int \log f(x_i|\theta) g(x) dx$$

But $-D(f(x|\theta), g(x)) = \int \log \frac{f(x|\theta)}{g(x)} g(x) dx = E[\log f(x|\theta)] - E[\log g(x)]$. Since the latter term is not a function of θ

$$\hat{\theta}_{QMLE} \xrightarrow{w.p.1} \underset{\theta \in \Theta}{\operatorname{argmin}} D(f(x|\theta), g(x)) = \theta^*$$

- C. Show that if the model is misspecified, the asymptotic distribution of $\hat{\theta}_N$ is under very weak conditions normally distributed but with a covariance matrix $\Omega(\theta^*)$ given by

$$\Omega(\theta^*) = [H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1} \quad (30)$$

where

$$H(\theta) = E \{ \partial^2 \log(f(x|\theta)/\partial\theta\partial\theta') \}, \quad (31)$$

and

$$I(\theta) = E \{ [\partial \log(f(x|\theta)/\partial\theta)] [\partial \log(f(x|\theta)/\partial\theta')] \} \quad (32)$$

Note that $I(\theta)$ is the usual *Information Matrix* of maximum likelihood and the inverse of this is the *Cramer-Rao lower bound*.

Answer By Taylor expansion of the FOC $\frac{\partial l_N(\hat{\theta}_{QMLE})}{\partial\theta} = 0$ around θ^* we have

$$\hat{\theta}_{QMLE} - \theta^* = - \left[\frac{\partial^2 l_N(\tilde{\theta})}{\partial\theta\partial\theta'} \right]^{-1} \left[\frac{\partial l_N(\theta^*)}{\partial\theta} \right]$$

where $\tilde{\theta}$ between $\hat{\theta}_{QMLE}$ and θ^* , then the asymptotic distribution is

$$\sqrt{N} \left(\hat{\theta}_{QMLE} - \theta^* \right) = - \left[\frac{\partial^2 l_N(\tilde{\theta})}{\partial\theta\partial\theta'} \right]^{-1} \left[\sqrt{N} \frac{\partial l_N(\theta^*)}{\partial\theta} \right]$$

First show

$$\sqrt{N} \frac{\partial l_N(\theta^*)}{\partial\theta} \rightarrow^d N(0, I(\theta^*)).$$

Next show $\frac{\partial l_N(\tilde{\theta})}{\partial\theta\partial\theta'} \rightarrow^{w.p.1.} \frac{\partial l(\theta^*)}{\partial\theta\partial\theta'}$ by ULLN, continuous mapping and using the fact that $\tilde{\theta} \rightarrow^{w.p.1.} \theta^*$. So that

$$\sqrt{N} \left(\hat{\theta}_{QMLE} - \theta^* \right) \rightarrow^d N(0, [H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1})$$

- D. Show that when the model is *correctly specified* that we have

$$H(\theta^*) = -I(\theta^*), \quad (33)$$

and therefore the asymptotic covariance matrix of the maximum likelihood estimator is

$$\Omega(\theta^*) = I^{-1}(\theta^*), \quad (34)$$

the usual Cramér-Rao lower bound formula. Thus, the maximum likelihood estimator is asymptotically normal even when the model one is estimating is misspecified, but *the maximum likelihood estimator only asymptotically attains the Cramér-Rao lower bound only when the model is correctly specified*.

Answer When the model is correctly specified the FOC becomes:

$$\int \frac{\partial \log f(x|\theta^*)}{\partial \theta} f(x|\theta^*) dx = 0$$

and by differentiating

$$\int \frac{\partial^2 \log f(x|\theta^*)}{\partial \theta \partial \theta'} f(x|\theta^*) + \frac{\partial \log f(x|\theta^*)}{\partial \theta} \frac{\partial \log f(x|\theta^*)}{\partial \theta'} f(x|\theta^*) dx = 0$$

or

$$E \frac{\partial^2 \log f(x|\theta^*)}{\partial \theta \partial \theta'} + E \frac{\partial \log f(x|\theta^*)}{\partial \theta} \frac{\partial \log f(x|\theta^*)}{\partial \theta'} = 0$$

so that $H(\theta^*) + I(\theta^*) = 0 \Rightarrow H(\theta^*) = -I(\theta^*)$. Since

$$\Omega(\theta^*) = [H(\theta^*)]^{-1} I(\theta^*) [H(\theta^*)]^{-1} = [-I(\theta^*)]^{-1} I(\theta^*) [-I(\theta^*)]^{-1} = [I(\theta^*)]^{-1}$$

VIII Apply your findings in problem VII above to this specific problem. Suppose that you are trying to estimate the single parameter θ in the model

$$\tilde{y} = \theta + \epsilon \tag{35}$$

but believe the ϵ is from a *Cauchy distribution* with density $f(x) = 1/\pi(1+x^2)$, for $x \in R^1$. However suppose that the true data generating process is $\epsilon \sim N(0, \sigma^2)$, i.e. the ϵ random variables are normally distributed with mean 0 and variance σ^2 .

- A. If you estimate θ by maximum likelihood using the incorrect assumption that the error terms ϵ are from a Cauchy distribution when they are really from a $N(0, \sigma^2)$ distribution, will your maximum likelihood estimator be consistent?

Answer The assumed (but incorrect) density function for an observation \tilde{y} is $f(y|\theta) = 1/\pi(1+(y-\theta)^2)$, so the log likelihood for the Cauchy model is

$$\mathcal{L}_N(y_1, \dots, y_N|\theta) = \frac{1}{N} \sum_{i=1}^N \log f(y_i|\theta) = -\log(\pi) - \frac{1}{N} \sum_{i=1}^N \log(1+(y_i-\theta)^2). \tag{36}$$

Assuming the standard regularity conditions for uniform almost sure convergence are satisfied, $\mathcal{L}_N(\theta)$ will converge to its expectation uniformly with probability 1 as $N \rightarrow \infty$. The limit is

$$E\{\log f(\tilde{y}|\theta)\} = -\log(\pi) - \int_{-\infty}^{+\infty} \log(1+(y-\theta)^2) \phi(y|\theta^*, \sigma^2) dy, \tag{37}$$

where $\phi(y|\theta^*, \sigma^2)$ is a normal density with mean θ^* and variance σ^2 , which is the true density from which the data are actually generated. Due to uniform convergence of $\mathcal{L}_N(\theta)$ to $E\{\log f(\tilde{y}|\theta)\}$, it follows from the uniform convergence results we studied in class that $\hat{\theta}_N$, the maximum likelihood estimator, and thus the value of θ that maximizes $\mathcal{L}_N(\theta)$ converges to θ_o , the value of θ that maximizes $E\{\log f(\tilde{y}|\theta)\}$. We can write $\tilde{y} = \theta^* + \sigma\tilde{\epsilon}$ where $\tilde{\epsilon} \sim N(0, 1)$. Therefore we can write

$$\theta_o = \underset{\theta}{\operatorname{argmax}} \int_{-\infty}^{+\infty} -\log(1+(\sigma\epsilon + \theta^* - \theta)^2) \phi(\epsilon|0, 1) d\epsilon \tag{38}$$

Taking derivatives, interchanging the derivative operator and the integral operator, we can show that the first order condition for θ_o is

$$\begin{aligned} 0 &= \frac{d}{d\theta} \int_{-\infty}^{+\infty} -\log(1 + (\sigma\epsilon + \theta^* - \theta)^2) \phi(\epsilon|0, 1) d\epsilon \\ &= 2 \int_{-\infty}^{+\infty} \frac{\sigma\epsilon + \theta^* - \theta}{1 + (\sigma\epsilon + \theta^* - \theta)^2} \phi(\epsilon|0, 1) d\epsilon. \end{aligned} \quad (39)$$

Note that $g(\epsilon) = \sigma\epsilon/[1 + (\sigma\epsilon + \theta^* - \theta)^2]$ is an *odd function* whereas $\phi(\epsilon|0, 1)$, the standard normal density, is an *even function*. It follows that

$$\int_{-\infty}^{+\infty} g(\epsilon)\phi(\epsilon|0, 1) = \int_{-\infty}^{+\infty} \frac{\sigma\epsilon}{1 + (\sigma\epsilon + \theta^* - \theta)^2} \phi(\epsilon|0, 1) d\epsilon = 0, \quad (40)$$

since the product of an odd and even function is odd, and the expectation exists (since it is not hard to show that $E\{|g(\epsilon)|\} < \infty$ using the Lebesgue dominated convergence theorem). Thus the first order condition for θ_0 reduces to

$$0 = (\theta^* - \theta_o) \int_{-\infty}^{+\infty} \frac{1}{1 + (\sigma\epsilon + \theta^* - \theta)^2} \phi(\epsilon|0, 1) d\epsilon. \quad (41)$$

Since the integral in the expression above can be shown to be positive, it follows that the only solution to the first order condition for maximizing $E\{\log f(\tilde{y}|\theta)\}$ in θ is the solution $\theta_o = \theta^*$. That is, even though we are doing maximum likelihood of a misspecified model in this case (since we are assuming that the \tilde{y} are from a Cauchy distribution centered at θ when they are really from a normal distribution centered at θ^*), the misspecification does not affect the consistency of the maximum likelihood estimator and it still converges to the true value θ^* with probability 1 as $N \rightarrow \infty$. This property, i.e. that maximizing a misspecified likelihood still results in a consistent estimator of the parameters θ^* , holds in certain class of problems, especially for *location problems* where as in this case we are interested in estimating parameters that affect the centering of the distribution but not other parameters governing the *shape* of the distribution. The general term *quasi maximum likelihood estimator* (QMLE) refers to the consistent estimation of parameters of interest using a likelihood function that could be misspecified in some respects, but for which we can show is still consistent for the *parameters of interest* θ^* .

- B. If your misspecified maximum likelihood estimator is not consistent, what θ^* does it converge to asymptotically?

Answer

$$\hat{\theta}_{QMLE} \xrightarrow{w.p.1} \operatorname{argmin} D(f(x|\theta), g(x)) = \theta^*$$

- C. Compute the asymptotic variance of the maximum likelihood estimator in the misspecified case. How does this variance compare to the variance of the maximum likelihood estimator if we use the correct specification for the errors terms, i.e. where OLS is the MLE and the asymptotic variance of the MLE $\hat{\theta}_N$ is σ^2 ?

Answer The asymptotic variance is: $\Omega(\theta^*) = [H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1}$ where

$$I(\theta^*) = E \left\{ \frac{d}{d\theta} \log f(\tilde{y}|\theta^*) \right\} = 4 \int_{-\infty}^{+\infty} \left[\frac{y - \theta^*}{1 + (y - \theta^*)^2} \right]^2 \phi(y|\theta^*, \sigma^2) dy, \quad (42)$$

and

$$H(\theta^*) = E \left\{ \frac{d^2}{d\theta^2} \log f(\tilde{y}|\theta^*) \right\} = -2 \int_{-\infty}^{+\infty} \left[\frac{1 - (y - \theta^*)}{1 + (y - \theta^*)^2} \right] \phi(y|\theta^*, \sigma^2) dy, \quad (43)$$

where $\phi(y|\theta^*, \sigma^2)$ is the normal density with mean θ^* and variance σ^2 , which again, is the true distribution that the data are generated by. Using the change of variables $(y - \theta)/\sigma = \epsilon$ where $\epsilon \sim N(0, 1)$ we can rewrite $I(\theta^*)$ and $H(\theta^*)$ as follows

$$I(\theta^*) = 4 \int_{-\infty}^{+\infty} \left[\frac{\sigma\epsilon}{1 + \sigma^2\epsilon^2} \right]^2 \phi(\epsilon|0, 1) d\epsilon, \quad (44)$$

and

$$H(\theta^*) = -2 \int_{-\infty}^{+\infty} \left[\frac{1 - (\sigma\epsilon)^2}{[1 + (\sigma\epsilon)^2]^2} \right] \phi(\epsilon|0, 1) d\epsilon. \quad (45)$$

Notice that $I(\theta^*)$ and $H(\theta^*)$ are only functions of σ but not θ^* . We know that by the Cramer-Rao lower bound, maximum likelihood estimation of a *correctly* specified parametric probability model results in the smallest asymptotic covariance matrix, $I^{-1}(\theta^*)$. In the case where the data are normally distributed with unknown mean θ^* we know that the maximum likelihood estimator of θ is just the sample mean, and the asymptotic variance of $\sqrt{N}(\bar{Y}_N - \theta^*)$ is σ^2 , which equals $I^{-1}(\theta^*)$ in this case. However for the problem at hand, where we are estimating θ using a misspecified maximum likelihood estimator based on the incorrect assumption that the data have a Cauchy distribution, the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta}_N - \theta^*)$ must be strictly larger than σ^2 since the maximum likelihood estimator of the location parameter of the (misspecified) Cauchy probability distribution is a consistent but *inefficient* estimator of θ^* . Its asymptotic covariance matrix is as noted above, $[H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1}$. When $\sigma^2 = 1$ we can calculate this numerically (using Gaussian quadrature to calculate $H(\theta^*)$ and $I(\theta^*)$) to get:

$$\begin{aligned} H(\theta^*) &= -0.688547 \\ I(\theta^*) &= 0.622947 \end{aligned} \quad (46)$$

This implies that the asymptotic variance of $\sqrt{N}[\hat{\theta}_N - \theta^*]$ is given by

$$[H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1} = 1.3140, \quad (47)$$

which is over 30% higher than the Cramer-Rao lower bound, $\sigma^2 = 1$, which would be the asymptotic variance of the correctly specified maximum likelihood estimation, which is just the sample mean in this case. Note that if we ignored the misspecification of the likelihood and used the simpler formula for the asymptotic variance of the MLE, $[I(\theta^*)]^{-1}$, we would get a much higher value of the asymptotic variance,

$$[I(\theta^*)]^{-1} = 1.6053. \quad (48)$$

This would be an incorrect value because it ignores the misspecification of the likelihood function. Thus, this also illustrates the importance of using the White correction or “sandwich formula” for the asymptotic variance of the estimator, $[H(\theta^*)]^{-1}I(\theta^*)[H(\theta^*)]^{-1}$ instead of the simpler Cramer-Rao lower bound formula $[I(\theta^*)]^{-1}$ which is only valid when the model is correctly specified (which is almost never the case). There is a subdirectory for this problem where we put copies of the code used to calculate the integrals in equations (44) and (45) above using it Gaussian quadrature. The Gaussian quadrature rule relies on an $J \times 1$ vector of *quadrature abscissae* q_a and *quadrature weights* q_w , where J is a value you can choose (with higher J resulting in higher accuracy). I used the inverse-probability transform method to calculate the integrals, so for example, to calculate an integral such as

$$E\{g(\tilde{\epsilon})\} = \int_{-\infty}^{+\infty} g(\epsilon)\phi(\epsilon|0,1)d\epsilon \quad (49)$$

we first do a change of variables $u = \Phi(\epsilon)$ where \tilde{u} will be a $U(0,1)$ random variable when $\tilde{\epsilon}$ is a $N(0,1)$ random variable as we discussed in class. Then taking inverses, we can write $\epsilon = \Phi^{-1}(u)$ and substitute this into the integral above to write it as

$$E\{g(\tilde{\epsilon})\} = \int_{-\infty}^{+\infty} g(\epsilon)\phi(\epsilon|0,1)d\epsilon = \int_0^1 g(\Phi^{-1}(u)) du. \quad (50)$$

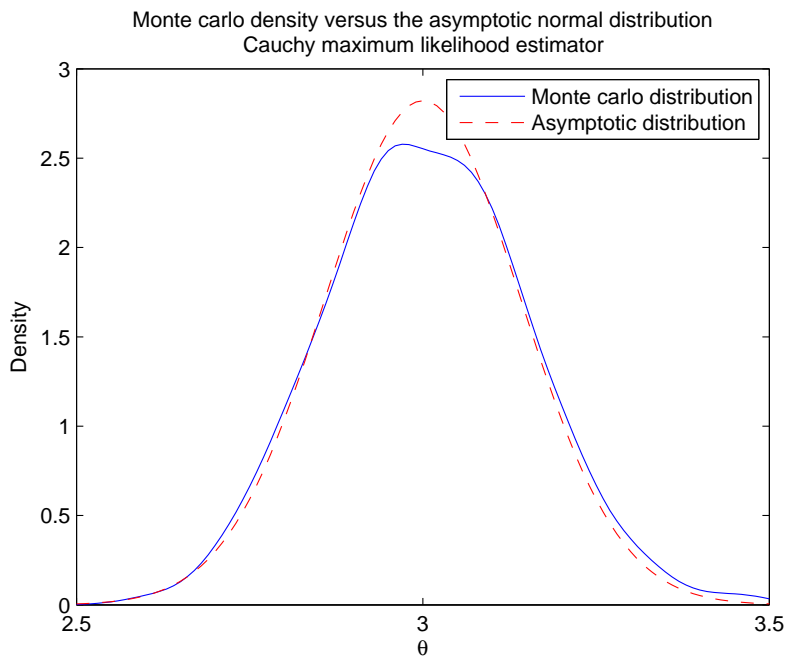
Now to compute this latter integral, we employ the quadrature rule. The quadrature rule approximates an integral over the $[0,1]$ interval as a weighted sum

$$\int_0^1 h(u)du \simeq \sum_{j=1}^J q_w(j)h(q_a(j)), \quad (51)$$

where $q_w(j)$ and $q_a(j)$ are the quadrature weights and abscissae, respectively. Gaussian quadrature is highly accurate, since it is designed so that with J weights and abscissae it can *exactly integrate* all polynomials of degree $2J - 1$ or lower (i.e. the quadrature rule generates the correct answer when the function $h(u)$ is a polynomial function of degree $2J - 1$ or lower). Thus, with $J = 15$, the value used to make the calculations above, Gaussian quadrature can exactly integrate all polynomials of degree 29 or lower. The routine `quadpoints.m` generates these quadrature weights and abscissae. The main program that calculates the integrals in equations (44) and (45) above is called `cauchy-information-hessian.m` and it calls the built-in Matlab function `norminv` to compute $\Phi^{-1}(u)$, and it calls the functions `cauchy-hessian.m` and `cauchy-information.m` to evaluate the expectations inside the integrals in equations (44) and (45) above.

- D. Conduct a Monte Carlo study by assuming that $\theta^* = 3$ and generating Cauchy random variables and numerically estimating θ with samples of size $N = 100$. Generate 500 separate, independent samples of size $N = 100$ and plot the empirical distribution of the maximum likelihood estimates. Is the sample mean of these 500 estimates approximately equal to $\theta^* = 3$? Do you see evidence of a finite sample bias? Also plot the empirical distribution of these 500 estimated values of $\hat{\theta}_N$ and compare them to the theoretical normal distribution with the correct asymptotic variance in this correctly specified case. Does the empirical distribution

Figure 1: Comparison of the Monte Carlo and Theoretical Asymptotic Distribution of the Cauchy MLE



match up well with the predicted normal asymptotic distribution of the Maximum likelihood estimator? **Hint:** from Wikipedia, the CDF of a Cauchy is given by

$$F(x) = \frac{1}{\pi} \arctan(x/\sigma) + \frac{1}{2}, \quad (52)$$

and the inverse of this is

$$F^{-1}(p) = \sigma \tan(\pi(p - 1/2)). \quad (53)$$

Answer Figure 1 above compares the kernel-smoothed distribution of the 500 maximum likelihood estimates of θ from our monte carlo study. These were computed by the Matlab program `cauchy-monte-carlo-study.m` which generated 100 observations from the Cauchy distribution with location parameter $\theta^* = 3$ and estimated this parameter by maximum likelihood 500 times, each time generating a new data set. The figure plots the kernel-density estimate of the distribution of $\hat{\theta}$ and compares it to the theoretical normal approximation to this distribution. This latter distribution is $\hat{\theta} = \theta^* + \sqrt{I^{-1}(\theta^*)} \tilde{\epsilon}$ where $\tilde{\epsilon} \sim N(0, 1)$ and $I(\theta^*) = 0.5$, which you can verify by repeating the calculations part C, but where the true data generating process is a Cauchy distribution. We see that the normal approximation to the monte carlo distribution (i.e. the finite sample distribution of $\hat{\theta}$) is very good and the two distributions are very close to each other, even with a relatively small number of observations, $N = 100$. Recall from the solution to Problem II-D above, if we tried to estimate θ using the sample mean, this estimator would not even be consistent when the true distribution of the data is Cauchy. It is reassuring that even though the Cauchy is a distribution for which no moments (not even the mean!) exist, there is still an estimator for θ that behaves well even in finite samples. Indeed we see little bias in the maximum likelihood estimates of θ^* from this monte carlo exercise.

E. Repeat part D, but use ϵ 's that are drawn from a $N(0, 1)$ distribution. Are the maximum likelihood estimates significantly biased? Plot the empirical distribution of the 500 maximum likelihood estimates against the “misspecification consistent” asymptotic normal distribution for the misspecified maximum likelihood estimator (i.e. the normal distribution with the variance $\Omega(\theta^*)$ given in part C of problem VII above). Does the normal distribution provide a good approximation to the empirical distribution of your 500 maximum likelihood estimates of θ ? Try doing the same thing but without using the “White correction” for the variance, i.e. instead of using $\Omega(\theta^*) = [H(\theta)]^{-1}I(\theta^*)[H(\theta)]^{-1}$ use $\Omega(\theta^*) = [I(\theta^*)]^{-1}$. How well does this normal distribution approximate the empirical distribution of your $\hat{\theta}_N$'s?

Answer Figure 2 above plots the kernel-smoothed monte carlo distribution of the estimated $\hat{\theta}$'s, which are again estimated by maximum likelihood, but now under the situation where the true data generating process is normal rather than Cauchy. Thus, we are in a situation where we are estimating a misspecified model since the distribution of the data are normal rather than Cauchy, whereas we have written a likelihood function assuming that the data are distributed according to the Cauchy distribution. We see that even though the likelihood is misspecified, provided we use the corrected covariance matrix to approximate the exact finite sample distribution of $\hat{\theta}$, the finite sample distribution of $\hat{\theta}$ is very close to its asymptotic normal approximation, even with only $N = 100$ observations. It seems like the price in terms of efficiency loss is very small relative to the benefit of the *robustness* of the Cauchy maximum likelihood estimator. Unlike the sample mean, which is not even a consistent estimator when the data are drawn from the Cauchy distribution, the Cauchy maximum likelihood estimator behaves well in small samples regardless of whether the data are drawn from a normal or a Cauchy distribution. Thus if you wanted to be protected in the worst case where the data really are Cauchy, the price in terms of greater inefficiency in the estimator when the data are not drawn from a Cauchy distribution does not appear to be high. Then why isn't the Cauchy maximum likelihood estimator used more frequently and in preference to the sample mean?

IX. Prove that the *conditional expectation* $E\{\tilde{y}|\tilde{X}\}$ is the *best predictor* of a random variable \tilde{y} using *any measurable function of \tilde{X}* $f(\tilde{X})$. **Hint: you need to show that for any measurable function $f(x)$**

$$E\left\{[\tilde{y} - f(\tilde{X})]^2\right\} \geq E\left\{[\tilde{y} - E\{\tilde{y}|\tilde{X}\}]^2\right\}. \quad (54)$$

Hint: use the Law of Iterated Expectations to show that the inequality above holds, and additionally using the “orthogonality property” of the residual $\tilde{\epsilon}$ given by

$$\tilde{\epsilon} = \tilde{y} - E\{\tilde{y}|\tilde{X}\}. \quad (55)$$

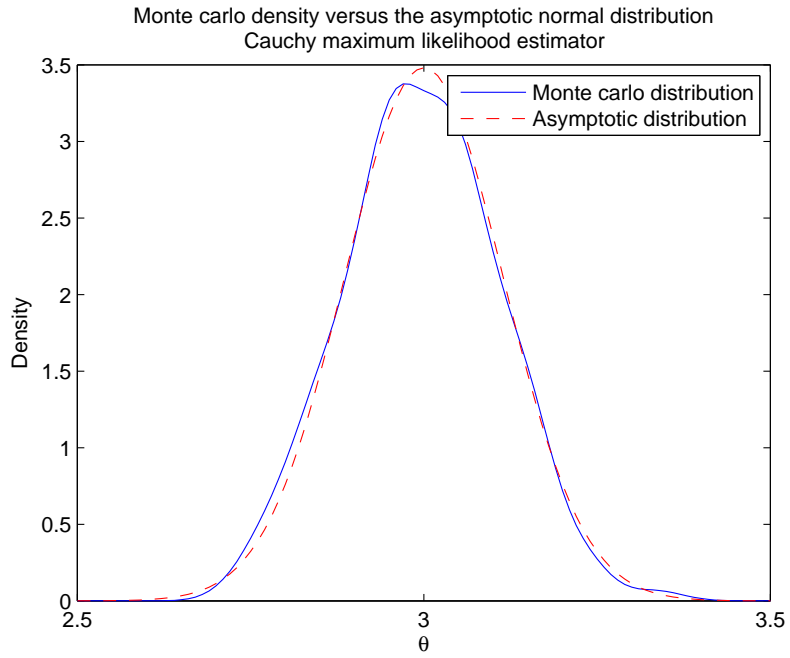
Answer

$$\begin{aligned} E\left\{[\tilde{y} - f(\tilde{X})]^2\right\} &= E\left\{[\tilde{y} - E(\tilde{y}|\tilde{x}) + E(\tilde{y}|\tilde{x}) - f(\tilde{X})]^2\right\} \\ &= E\left\{[\tilde{y} - E(\tilde{y}|\tilde{x})]^2 + 2[\tilde{y} - E(\tilde{y}|\tilde{x})][E(\tilde{y}|\tilde{x}) - f(\tilde{X})] + [E(\tilde{y}|\tilde{x}) - f(\tilde{X})]^2\right\} \\ &= E\{[\tilde{y} - E(\tilde{y}|\tilde{x})]^2\} + 2E\{[\tilde{y} - E(\tilde{y}|\tilde{x})][E(\tilde{y}|\tilde{x}) - f(\tilde{X})]\} + E\{[E(\tilde{y}|\tilde{x}) - f(\tilde{X})]^2\} \end{aligned}$$

By law of iterated expectation the middle term becomes:

$$2E\{[\tilde{y} - E(\tilde{y}|\tilde{x})][E(\tilde{y}|\tilde{x}) - f(\tilde{X})]\} = 2E\{E[[\tilde{y} - E(\tilde{y}|\tilde{x})][E(\tilde{y}|\tilde{x}) - f(\tilde{X})]|\tilde{X}]\}$$

Figure 2: Comparison of the Monte Carlo and Theoretical Asymptotic Distribution of the Cauchy MLE



$$= 2E\{[E(\tilde{y}|\tilde{x}) - E(\tilde{y}|\tilde{x})][E(\tilde{y}|\tilde{x}) - f(\tilde{X})]\} = 2E\{0[E(\tilde{y}|\tilde{x}) - f(\tilde{X})]\} = 0$$

$$\text{So } E\{[\tilde{y} - f(\tilde{X})]^2\} = E\{[\tilde{y} - E(\tilde{y}|\tilde{x})]^2\} + E\{[E(\tilde{y}|\tilde{x}) - f(\tilde{X})]^2\} \geq E\{[\tilde{y} - E(\tilde{y}|\tilde{x})]^2\}$$