Columbia University
Statistical Analysis and Time Series

IEOR-4709

A. Capponi **Problem Set #3** Issued: February 14, 2024
Spring 2025 Due: **BEFORE CLASS** March 5, 2025

Note: Please put the number of hours that you spent on this homework set on top of the first page of your homework. The CA in charge of grading this homework is Boxuan Li. The TA in charge of this homework is Kishore Kuppusamy.

Ex. 1.
Let $X_1, X_2, \ldots, X_n$ be i.i.d. according to a Pareto distribution with density $f_\theta(x) = \theta c^\theta x^{-(\theta+1)}$, where $\theta > 0$ and $0 < c < x$. We want to test the hypothesis $H_0 : \theta = \theta_0$ vs the alternative hypothesis $H_1 : \theta \neq \theta_0$. Following the example shown in the class for testing whether the mean of a Gaussian distribution with unknown variance is θ_0, do the following:

- Write down the critical region of the likelihood ratio test.
- Find the threshold k in the definition of the critical region, so that the significance level of the test is α.

Solution

- The log-likelihood is

$$\begin{aligned} l(\theta; X_1, \ldots X_n) &= \log(\prod_{i=1}^{n} \theta c^\theta x_i^{-(\theta+1)}) \\ &= n\log(\theta) + \theta(n\log(c) - \sum_{i=1} \log(X_i)) - \sum_i \log(X_i) \end{aligned}$$

The MLE for θ is

$$\hat{\theta} = \frac{n}{\sum_i \log(X_i) - n\log(c)}.$$

The likelihood ratio test

$$\begin{aligned} \Lambda &= 2l(\hat{\theta}; X_1, \ldots, X_n) - 2l(\theta_0; X_1, \ldots, X_n) \\ &= 2n\log(\hat{\theta}) - 2\hat{\theta}(\sum \log(X_i) - n\log(c)) \\ &- 2n\log(\theta_0) + 2\theta_0(\sum \log(X_i) - n\log(c)) \\ &= 2n\bigg(\log(\frac{\hat{\theta}}{\theta_0}) - (1 - \frac{\theta_0}{\hat{\theta}})\bigg). \end{aligned}$$

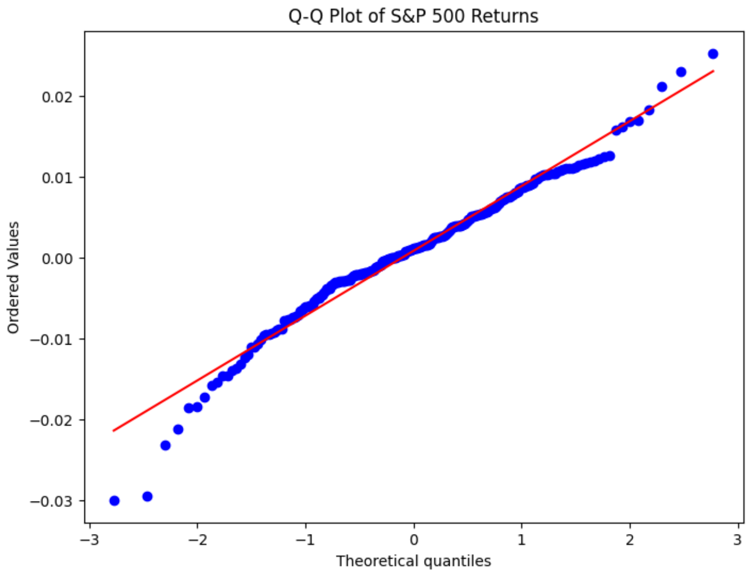
The critical region is $C_1 = \{x : \Lambda(x) \geq k\}$.

- We can use the result derived in lecture 9 that $\Lambda \Rightarrow \chi_1^2$. We can then use the proper quantile of the χ_1^2 distribution to determine the critical region which achieves the desired significance level α. In this way, $k = F^{-1}(1-\alpha)$, where F is the cdf of χ_1^2.

Ex. 2.
Go on Yahoo Finance. Download the time series of historical data of the S&P 500 from Feb 16, 2024 to Feb 16, 2025. Compute the returns of the S&P 500. Do a Q-Q plot of the empirical quantiles versus theoretical quantiles, assuming that the theoretical distribution of the returns is Gaussian with some mean μ and variance σ^2. Based on the plot, is there enough evidence to conclude that the returns are Gaussian distributed? Explain.

Solution The normal Q-Q plot is as below, where the red line is Q-Q line plotted as a reference: if all the points concentrate on this line then the sample has normal distribution.

We can see from the plot that points in the middle are on the red line, but points on the right and left tail evidently deviate from the red line. Therefore S&P 500 returns follow a distribution that has more density on the right end than normal distribution and a smaller density on the left end. These returns have more extreme values on the positive side than a normal distribution does.



Ex. 3.
Consider estimating the distribution function $P(X \leq x)$ at a fixed point x based on a sample $X_1, \ldots, X_n$ from the distribution of X. An estimator is $\frac{1}{n}\sum_i \mathbf{1}_{X_i \leq x}$. If it is known that the true underlying distribution is Gaussian with mean θ and variance 1, another possible estimator is $\Phi(x - \bar{X})$. Calculate the relative efficiency of these estimators, i.e., the ratio of their asymptotic variances. **You may find it useful to use the delta method to find the asymptotic distribution of $\Phi(x - \bar{X})$.**

Solution

If we have two estimator sequences that converge to normal distribution at rate $\sqrt{n}$, their relative efficiency is defined as the ratio of their asymptotic variances. Hence, we only need to find the asymptotic distributions of $\tilde{p} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \leq t}$ and $\hat{p} = \Phi(t - \bar{X})$.

Let us begin by considering $\tilde{p} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \leq t}$. Since

$$\begin{aligned} \mu(\theta) &= E[1(X_i \leq t)] \\ &= P(X_i \leq t) \\ &= \Phi(t - \theta) \end{aligned}$$

and

$$\begin{aligned} \sigma^2(\theta) &= Var[1(X_i \leq t)] \\ &= P(X_i \leq t) - P(X_i \leq t)^2 \\ &= \Phi(t - \theta)[1 - \Phi(t - \theta)]. \end{aligned}$$

By the CLT, we have that $\frac{1}{\sqrt{n}}\sum_{i=1}^n [\mathbf{1}_{X_i \leq t} - \mu(\theta)]$ converges to a Gaussian in distribution with mean zero and variance $\sigma^2(\theta)$.

Now consider $\hat{p} = \Phi(t - \bar{X})$. Since $X_1, \ldots, X_n$ are i.i.d $N(\theta, 1)$, by the CLT we have that $\sqrt{n}(\bar{X} - \theta) \Rightarrow N(0, 1)$ in distribution. Now, we can use delta method to find the asymptotic distribution of $\hat{p}$.

Define $f : x \to \Phi(t - x)$. Then the derivative of f w.r.t x evaluated around θ is

$$f'_x(\theta) = \left[\frac{\partial \Phi(t - x)}{\partial x}\right]_{x=\theta} = -\phi(t - \theta),$$

where $\phi(\cdot) = \Phi'(\cdot) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\cdot)^2}$, the p.d.f of the standard normal distribution. If we take $Z \sim N(0, 1)$, then by the delta method,

$$\sqrt{n}(\hat{p} - \mu(\theta)) = \sqrt{n}(f(\bar{X}) - f(\theta)) \xrightarrow{d} Z f'_x(\theta) = -Z\phi(t - \theta) = N(0, \phi(t - \theta)^2),$$

where the last equality is intended to be in distribution. Thus the asymptotic relative efficiency at θ is

$$\frac{\phi(t - \theta)^2}{\Phi(t - \theta)[1 - \Phi(t - \theta)]}.$$

Notice that $\phi(x)$ and $\Phi(x)$ are log concave functions, so they are quasi-concave w.r.t x, i.e. $\frac{\partial \phi^2(x)}{\partial x^2}$ and $\frac{\partial \Phi^2(x)}{\partial x^2} \leq 0$ around 0. (the tails with $|x| > 1$ are convex but we know they are small, so max will not be achieved there) Therefore, by taking the second derivative, we get

$$\frac{-2\phi(t-\theta)\left.\frac{d^2}{d\xi_1^2}\phi(\xi_1)\right|_{\xi_1=t-\theta}-2\left.\frac{d}{d\xi_1}\phi(\xi_1)\right|_{\xi_1=t-\theta}^2+\frac{4\cdot(2\Phi(t-\theta)-1)\phi(t-\theta)\left.\frac{d}{d\xi_1}\Phi(\xi_1)\right|_{\xi_1=t-\theta}\left.\frac{d}{d\xi_1}\phi(\xi_1)\right|_{\xi_1=t-\theta}}{(\Phi(t-\theta)-1)\Phi(t-\theta)}+\frac{\left(2\Phi(t-\theta)\left.\frac{d^2}{d\xi_1^2}\Phi(\xi_1)\right|_{\xi_1=t-\theta}+2\left.\frac{d}{d\xi_1}\Phi(\xi_1)\right|_{\xi_1=t-\theta}^2-\left.\frac{d^2}{d\xi_1^2}\Phi(\xi_1)\right|_{\xi_1=t-\theta}-\frac{2(2\Phi(t-\theta)-1)^2\left.\frac{d}{d\xi_1}\Phi(\xi_1)\right|_{\xi_1=t-\theta}^2}{(\Phi(t-\theta)-1)\Phi(t-\theta)}\right)\phi^2(t-\theta)}{(\Phi(t-\theta)-1)\Phi(t-\theta)}}{(\Phi(t-\theta)-1)\,\Phi(t-\theta)}$$

it can be seen that the first term in the numerator is positive using the fact that the ϕ is concave near 0.

Consider the last term in the numerator. If we replace $t-\theta$ by x, then the first and third part in the parenthesis $2(\Phi(t-\theta)^2\frac{d\Phi(x)}{dx^2}-\frac{d^2}{dx^2}\Phi(x)=\frac{d\phi(x)}{dx}(2\Phi(x)-1)<0$, since if $x<0$, $\frac{d\phi(x)}{dx}>0$, then $2\Phi(x)-1<0$, and if $x>0$, then $\frac{d\phi(x)}{dx}<0, 2\Phi(x)-1>0$.Moreover, if we look at the second and the fourth terms in the parenthesis, since $\frac{d}{d(x)}\Phi(x)=\phi(x)$, then $2\frac{d}{d(x)}\Phi(x)^2-\frac{2(2\Phi(x)-1)^2\frac{d}{dx}\Phi(x)}{(\Phi(x)-1)-\Phi(x)}=\frac{2\phi(x)^2(\Phi(x)-1)\Phi(x))-2(2\Phi(x)-1)^2\phi(x)^2}{(\Phi(x)-1)\Phi(x)}$. Notice the numerator can be simplified to $\frac{2\phi(x)^2(\Phi(x)-1)(1-\Phi(x))}{(\Phi(x)-1)\Phi(x)}<0$ (since $\Phi(x)-1<0$). Therefore, the numerator of the last term <0, and thus the last term >0.

For the third term in the numerator, $2\Phi(X)-1$ and $\phi'(x)$will have the opposite sign and $\frac{d\Phi}{dx}=\phi(x)>0$, and since its denominator is negative, so the third term is also positive. Lastly, we know that the second term is negative, but since $\frac{d\phi(x)}{dx}$ is small near 0 (the density function is almost flat at 0), we can verify that the second term is always smaller than the first term. Therefore the numerator is always positive near 0, and the denominator is always negative since $\Phi(x)<1$. Hence $\frac{\partial^2}{\partial(t-\theta)}\phi^2(x)/\Phi(x)(1-\Phi(x))<0$, which means the function is concave around 0.

One can also simply plot the function and verify that $\frac{\phi(x)^2}{\Phi(x)[1-\Phi(x)]}$ is a concave function with respect to x around 0, with max value ≈ 0.64 at $x=0$, or equivalently, at $t=\theta$.

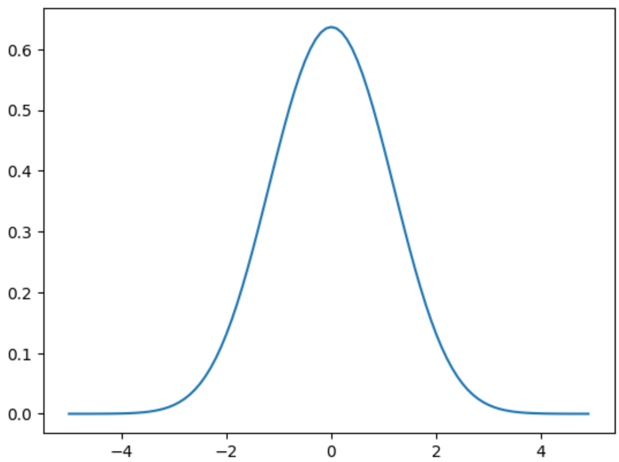


Figure 1: $\frac{\phi(x)^2}{\Phi(x)(1-\Phi(x))}$

As a result, the second estimator is asymptotically more efficient since it always has a smaller asymptotic variance than the first estimator (as the ratio is always smaller than 1). The second estimator is better as it uses all X_i in the sample.

Ex. 4.

A popular model in market microstructure postulates that the total number of trades, ignoring the case of no trade, within a minutes for a particular stock follows a geometric distribution $p(x)=P(X=x)=p^{x-1}(1-p)$, $x=1,2,...$ The following table contains data of all the trade frequencies within 134 minutes with at least one trade. Do the following

- Find the MLE for p.
- Test whether the geometric distribution fits the data.

N. Trades	Frequency
1	46
2	32
3	21
4	11
5	5
6	7
7	4
8	3
9	1
10	2
11	1
12	1

Solution In the chart, it gathers the frequency of minutes which have the same number of trades in a minute. In order to get MLE of $\hat{P}$, we need to separate the minutes in one category and treat them as i.i.d. samples x_i from the geometric distribution, with $n = 134$. So the likelihood function is

$$L(X|P) = \prod_{i=1}^{n} P^{x_i-1}(1-P).$$

Take its logarithm,

$$\log L(X|P) = \sum_{i=1}^{n}(\log(1-P) + (x_i - 1)\log P) = n\log(1-P) + (\sum_{i=1}^{n} x_i - n)\log P.$$

Use the first order condition to compute $\hat{P}$, we have

$$\begin{aligned} & \tfrac{\partial \log L}{\partial P} = -\tfrac{n}{1-P} + \tfrac{\sum_{i=1}^{n} x_i - n}{P} = 0 \\ \rightarrow \quad & nP = (1-P)(\textstyle\sum_{i=1}^{n} x_i - n) \\ \rightarrow \quad & P = \tfrac{\sum_{i=1}^{n} x_i - n}{\sum_{i=1}^{n} x_i} = 1 - \tfrac{n}{\sum_{i=1}^{n} x_i} = 1 - \tfrac{1}{\bar{X}}, \end{aligned}$$

so the MLE $\hat{P} = 1 - \frac{1}{\bar{X}} = 0.6546$ from the data file.

Then we compute the value $E_j = n(1-\hat{P})\hat{P}^{j-1}$, $j = 1, ..., 12$ to test whether the geometric distribution fits the data. Here is the list of values E_j and O_j, where E_{12} represents $nP(X \geq 12)$.

Then we can compute the test statistic,

$$W(X) = 2\sum_{i=1}^{12} O_j \log(\frac{O_j}{E_j}) = 6.01.$$

N. Trades	O_j	E_j
1	46	46.28
2	32	30.30
3	21	19.83
4	11	12.98
5	5	8.50
6	7	5.56
7	4	3.64
8	3	2.38
9	1	1.56
10	2	1.02
11	1	0.67
12	1	0.44

Here, the number of degrees of freedom is d.f. = 12-1-1=10. The p-value is the probability that a χ^2_{10} random variable takes a value greater than $W(X)$, which is equal to 0.814 in this case, meaning that there is no evidence against H_0, i.e. no evidence against the geometric distribution fitting the data.

Alternatively, we can also calculate the Pearson's statistic

$$W(X) = \sum_{i=1}^{12} \frac{(E_j - O_j)^2}{E_j} = 4.50.$$

and the p-value is equal to 0.922, which also shows there is no evidence to reject H_0