Geneva Graduate Institute, Econometrics I

Problem Set 2 Solutions

Francesco Casalena

Fall 2024

Problem 1

Suppose you have data on the height of n female adults living in Switzerland – $\{x_i\}_{i=1}^n$ – whereby the observations in your sample are independent. Based on that, you want to estimate the average height of female adults in the whole population (i.e. the whole of Switzerland). Let this parameter of interest be denoted by θ. You can write your observations as

$$x_i = \theta + u_i \ , \quad \text{with} \quad \mathbb{E}[u_i|\theta] = 0 \ ,$$

i.e. the height of an individual i, x_i, is given by the true average height θ plus some noise u_i around it. Note that this is just another way of writing $\mathbb{E}[x_i|\theta] = \theta$. In addition, you assume that this noise u_i is Normally distributed with some known variance σ^2: $u_i \sim N(0, \sigma^2)$. Note that this – combined with the equation for x_i above – is just another way of writing $x_i \sim N(\theta, \sigma^2)$.

1. Define and derive the Maximum Likelihood (ML) estimator of θ, $\hat{\theta}_{ML}$.

Hint: You first need to derive the likelihood, i.e. the distribution of your data $\{X_i\}_{i=1}^n$ conditional on θ, $p(x|\theta)$, based on the distribution of a single observation X_i, $p(x_i|\theta)$.

Solution:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|x) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(x_i|\theta)$$

Where $p(x_i|\theta)$ is the pdf we assume has generated the data. In this case, it's Normal. We can therefore substitute it into the Likelihood function to obtain:

$$\mathcal{L}(\theta|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\left(\frac{x_i - \theta}{\sigma}\right)^2\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \theta)^2}{\sigma^2}\}$$

Take logs to have the log-Likelihood function:

$$l(\theta|x_i) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2$$

Take the first derivative w.r.t. θ and set it equal to zero:

$$\frac{\partial l(\theta|x)}{\partial \theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta) = 0$$

Solve for θ to obtain:

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \bar{X}$$

2. Set up a Likelihood Ratio (LR) test of size $\alpha = 0.05$ for testing $\mathcal{H}_0 : \theta = \theta_0$ against $\mathcal{H}_1 : \theta \neq \theta_0$, i.e. determine the test-statistic $T(X)$ and the corresponding critical value c_α.

Hint: Note that if $Y \sim N(\mu, v)$, then $(Y - \mu)/\sqrt{v} \sim N(0, 1)$ and $(Y - \mu)^2/v \sim \chi^2_1$.

Solution:

The likelihood ratio test statistic is given by:

$$
\begin{aligned}
T(X) = \frac{p(x|\hat{\theta}_{ML})}{p(x|\theta_0)} &= \frac{((2\pi\sigma^2)^{-n/2})\exp\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\hat{\theta}_{ML})^2}{\sigma^2}\}}{((2\pi\sigma^2)^{-n/2})\exp\{-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\theta_0)^2}{\sigma^2}\}} = \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\hat{\theta}_{ML})^2-(x_i-\theta_0)^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[x_i^2-2x_i\hat{\theta}+\hat{\theta}^2-x_i^2+2x_i\theta_0-\theta_0^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2}\left[-2\hat{\theta}\underbrace{\sum_{i=1}^{n}x_i}_{n\hat{\theta}}+n\hat{\theta}^2+2\theta_0\underbrace{\sum_{i=1}^{n}x_i}_{n\hat{\theta}}-n\theta_0^2\right]\} \\
&= \exp\{-\frac{1}{2\sigma^2}\left[-2n\hat{\theta}^2+n\hat{\theta}^2+2n\hat{\theta}\theta_0-n\theta_0^2\right]\} = \\
&= \exp\{-\frac{1}{2\sigma^2/n}\left[-\hat{\theta}^2+2\hat{\theta}\theta_0-\theta_0^2\right]\} = \\
&= \exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\}
\end{aligned}
$$

We accept the null hypothesis if:

$$
\begin{aligned}
\varphi_{LR} &= \mathbf{1}\{T(x) < c\} = \\
&= \mathbf{1}\left\{\exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\} < c\right\} = \\
&= \mathbf{1}\left\{\frac{1}{\sigma^2/n}(\hat{\theta}-\theta_0)^2 < \tilde{c}\right\}
\end{aligned}
$$

Where $\tilde{c} = 2\log(c)$.

To find $\tilde{c}$ (or, equivalently, c) we set the probability of rejection given that the null is true (i.e. the probability of false rejection) equal to our desired test-size α:

$$
\begin{aligned}
\alpha &\equiv \mathbb{P}\left[\text{reject}\mathcal{H}_0|\mathcal{H}_0\text{is true}\right] = \\
&= \mathbb{P}\left[\varphi(x) = 0|\theta_0\right] = \\
&= \mathbb{P}\left[T(x) > c|\theta_0\right] = \\
&= \mathbb{P}\left[\exp\{\frac{1}{2\sigma^2/n}(\hat{\theta}-\theta_0)^2\} \geq c|\theta_0\right] = \\
&= \mathbb{P}\left[\frac{1}{\sigma^2/n}(\hat{\theta}-\theta_0)^2 \geq \tilde{c}|\theta_0\right]
\end{aligned}
$$

Note that the estimator for theta is normally distributed centered about the true mean: $\hat{\theta} \sim N(\theta, \sigma^2/n)$. Hence, under the null hypothesis: $\hat{\theta} \sim N(\theta_0, \sigma^2/n)$. This in turn implies that:

$$
\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{(\hat{\theta}-\theta_0)^2}{\sigma^2/n} \sim \chi_1^2
$$

Therefore: $\alpha = \mathbb{P}[W \geq \tilde{c}]$, where $W \sim \chi^2_1$. Rearranging, we get $\mathbb{P}[W \leq \tilde{c}] = 1-\alpha$. Thus, $\tilde{c}$ is the $100(1-\alpha)$-th percentile of χ^2_1. The 95th percentile of a chi-squared distribution with 1 degree of freedom is 3.84. Hence, for a size $\alpha = 5\%$-test, we take the critical value $\tilde{c}_{\alpha=0.05} = 3.84$.

3. Suppose $\sigma^2 = 6$ and you observe $n = 4$ observations, $x_1 = 178$, $x_2 = 161$, $x_3 = 168$ and $x_4 = 172$. Based on this data, can you reject $\mathcal{H}_0 : \theta = \theta_0 = 175$ (i.e. that the average height of female adults in Switzerland is 175cm)?

Solution: First compute your ML estimator:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{178 + 161 + 168 + 172}{4} = 169.75$$

Compute the test statistic by plugging in the numbers and compare it to the $\alpha = 0.05$ critical value of a chi-squared distribution with 1 degree of freedom:

$$T_{LR}(X) = \frac{(169.75 - 175)^2}{6/4} = 18.375 > 3.84 = \tilde{c}_{0.05}$$

Since the test statistic is larger than the 5% critical value, we reject the null hypothesis at the 5% significance level. (Actually, we are able to reject the null hypothesis also at the 1% critical value, which has critical value $\tilde{c}_{0.01} = 6.63$).

4. Now let's suppose you could only find the test-statistic for the LR test, $T(X)$, but not the critical value c_α. Do so numerically, i.e.
   (a) For $m = 1 : M$, with $M = 1000$,
   - draw a sample $\{x_i^m\}_{i=1}^n \sim N(\theta_0, \sigma^2)$, setting $\theta_0 = 175$, $\sigma^2 = 6$ and $n = 4$,
   - compute $T(x^m)$.

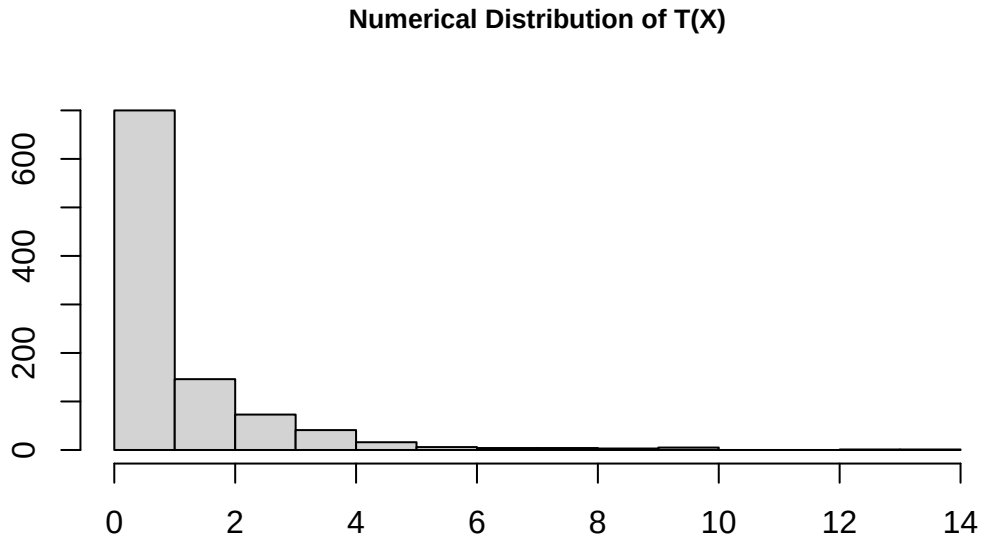
   Plot a histogram of $\{T(x^m)\}_{m=1}^M$. This is your numerical approximation of the distribution of $T(X)$ under $\mathcal{H}_0$.
   (b) Sort your draws $\{T(x^m)\}_{m=1}^M$ from lowest to largest and take the $M(1-\alpha)$th draw. This is your numerical approximation of c_α, the $100(1-\alpha)$th quantile of the distribution of $T(X)$.

Is the value you get close to the true, analytically obtained c_α? What do you expect to happen if you take a larger value for M? Does your conclusion from the previous exercise change if you set up your test numerically as opposed to analytically?

Solution:

```
rm(list = ls())
set.seed(2024)

M <- 1000 #total simulations
v <- numeric(length(M)) #here we store our draws

n <- 4 #set sample size
theta_0 <- 175 #set mean under H_0
stdev <- sqrt(6) #set variance of distribution

for (i in 1:M) {
  x <- rnorm(n, mean = theta_0, sd = stdev) #draw from pdf characterised above
  theta_hat <- mean(x) #compute ML estimator
  test_stat <- ((theta_hat-theta_0)/(stdev/sqrt(n)))^2 #compute LR test statistic
  v[i] <- test_stat #save the computed T(x)
}
```

```r
#plot histogram
hist(v,
     main = "Numerical Distribution of T(X)",
     cex.main = 0.8,
     xlab = " ",
     ylab = " ")
```




```r
#find critical value at 5% significance
c005 <- quantile(v, probs = 0.95)

print(c005)
```

```
##      95%
## 3.627377
```

The numerically computed $c_{0.05}$ is 3.627. The one computed analytically is equal to $ c_{0.05} = 3.84$. The two are relatively close, but by increasing the total number of simulations M the distance between the two decreases. Overall, our results from the previous exercise would not change, as we would still reject the null hypothesis at 5%.

5. Based on your LR-test, find a (general) expression for the 95% confidence interval for θ_0, $C(X)$. How does that interval look like if you apply it to your particular data? Is $\theta_0 = 175$ in that interval? Explain why it should (not) be.

Solution:

In the case of the LR test, the acceptance region is given by:

$$T(X) = \left(\frac{\hat{\theta}_{ML} - \theta_0}{\sigma/\sqrt{n}} \right)^2 \leq \tilde{c}_{0.05} = 3.84$$

With the following algebraic manipulations we can obtain:

$$\begin{aligned}\left(\frac{\hat{\theta}_{ML}-\theta_0}{\sigma/\sqrt{n}}\right)^2 \leq & 3.84\\ -\sqrt{3.84} \leq & \frac{\hat{\theta}_{ML}-\theta_0}{\sigma/\sqrt{n}} \leq \sqrt{3.84}\\ -\sigma\sqrt{3.84} \leq & \hat{\theta}_{ML}-\theta_0 \leq \sigma\sqrt{3.84}\\ -\frac{\sigma}{\sqrt{n}}\sqrt{3.84} \leq & \hat{\theta}_{ML}-\theta_0 \leq \frac{\sigma}{\sqrt{n}}\sqrt{3.84}\end{aligned}$$

By computing $\sqrt{c_{0.05}} = \sqrt{3.84} = 1.96$ we are able to obtain the following CI:

$$C(X) = \left[\hat{\theta}_{ML} - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\theta}_{ML} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

By plugging in the actual numbers $\hat{\theta} = 169.75$, $\sigma = \sqrt{6}$ and $n = 4$ we obtain:

$$C(X) = [167.35, 172.15]$$

Clearly, $\theta_0 = 175$ does not lie in the CI. It should not be in there, as by definition an $(1-\alpha)100$ CI is the set of values of θ_0 by which we are not rejecting the null hypothesis that $\theta = \theta_0$ at the α significance level. Since we rejected $\mathcal{H}_0$ that $\theta = 175$, it does not lie in the interval.

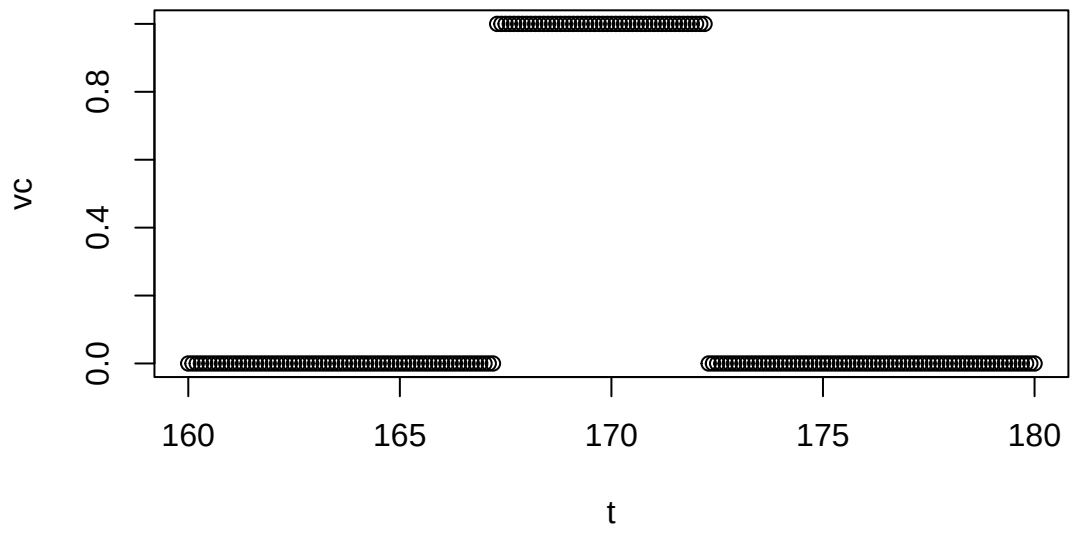
6. Let's again suppose you were not able to analytically set up the LR test and, based on it, find $C(X)$. Find the confidence interval $C(x)$ for your sample numerically as follows. First, fix a grid $\mathcal{T}$ of values for θ_0, $\mathcal{T} = 160 : 0.1 : 180$, and create a vector vc of the same dimension as $\mathcal{T}$. Then, for each $\theta_0 \in \mathcal{T}$,
   (a) repeat the numerical procedure from above to find $c_\alpha(\theta_0)$, the (numerical approximation of the) critical value for a size $\alpha = 0.05$ test for testing $\mathcal{H}_0 : \theta = \theta_0$.
   (b) compute the LR-test-statistic $T(x;\theta_0)$ for your sample x. If $T(x;\theta_0) < c_\alpha(\theta_0)$, then $\theta_0 \in C(x)$ and you record a 1 in the corresponding entry in vc, otherwise $\theta_0 \notin C(x)$ and you record a 0.

Illustrate your $C(x)$ using a scatter plot: put $\mathcal{T}$ on the x-axis and, for each value $\theta \in \mathcal{T}$, have a one on the y-axis if θ is in $C(x)$ and a zero otherwise. How does your $C(x)$ compare to the one obtained analytically?

Solution:

```
#compute our mean from previous exercise
theta_hat <- 169.75

#create grid of total values of theta_0 to test
t <- seq(160, 180, by = 0.1)
#here we store the values of theta_0 that lie in the CI
vc <- numeric(length(t))
#here we store critical values
c005 <- numeric(length(t))

#select total number of random draws
M <- 1000

#for every value of theta_0 in the grid draw M random
#samples and use them to compute the critical value
for (j in seq_along(t)) {
  #each point on the grid is theta_0
  theta_0 <- t[j]
```

```r
  #here we store our test statistics
  #computed from M different random samples
  v <- numeric(length(M))
  #compute the test statistic for the specific point
  #of the grid t
  test_stat0 <- ((theta_hat-theta_0)/(stdev/sqrt(n)))^2

  #get critical value for each point by simulating
  #M samples and computing the distributions
  for (i in 1:M) {
    x <- rnorm(n, mean=t[j], sd=stdev)
    test_stat <- ((mean(x)-theta_0)/(stdev/sqrt(n)))^2
    v[i] <- test_stat
  }
  c005[j] <- quantile(v, probs=0.95)

  #select or reject t-stat according to critical value
  if ( test_stat0 <= c005[j]) {
    vc[j] <- 1
  }
  else {
    vc[j] <- 0
  }
}

#plot the scatter representing the CI
plot(t, vc)
```




```r
#show bounds of the computed CI
CI95 <- c(t[which(vc == 1)[1]], t[which(vc == 1)[length(which(vc == 1))]])
print(CI95)
```

```
## [1] 167.3 172.2
```

The confidence interval computed numerically is $C(X) = [167.4, 172.1]$, which is very close to the one that we found analytically. (Note that if we used a finer grid, for instance increasing by 0.05, our result would have been even more precise).