

Problem Sheet 2

MATH50011
Statistical Modelling 1

Week 3

Lecture 5 (Asymptotic Normality)

1. Prove that if X_1, X_2, \dots converges in probability to X and h is a continuous function, then $h(X_1), h(X_2), \dots$ converges in probability to $h(X)$.
2. Suppose that X_1, \dots, X_n are iid with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.
 - (a) Show that S_n^2 is a consistent estimator of σ^2 . Assume that all required higher order moments of X_i exist and are finite.
 - (b) Use the result in (a) to show that
$$T_n = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \right) \rightarrow_d N(0, 1).$$
3. Suppose that X_1, \dots, X_n are iid strictly positive random variables with $E(\log X_i) = \mu$ and $\text{Var}(\log X_i) = \sigma^2$. Use the delta method to derive the asymptotic normality of the geometric mean $G_n = (\prod_{i=1}^n X_i)^{1/n}$.
4. (**Challenging**) Suppose that X_1, \dots, X_n are iid Uniform($0, \theta$) and define $T_n = \max(X_1, \dots, X_n)$. Find a sequence $a_n = n^k$ for some k such that $a_n(T_n - \theta) \rightarrow_d Z$. What is the distribution of Z ?
5. Does $\sqrt{n}(T_n - \theta) \rightarrow_d N(0, \sigma^2)$ imply that T_n is consistent for θ ? If yes, prove this. Otherwise, provide a counterexample.

Lecture 6 (Maximum Likelihood)

6. Find the MLE for estimating θ based on a random sample X_1, \dots, X_n from the following distributions
 - (a) Bernoulli(θ); (see Example 8)

- (b) Poisson(θ);
(c) Exponential(θ);
7. For the distributions in 6(a-c), find Z such that $\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_d Z$.
8. **(Challenging)** For the distributions in 6(a) and 6(b), find the MLE $\hat{\nu}_n$ of $\nu = g(\theta) = P_\theta(X_1 = 0)$ and show that $\sqrt{n}(\hat{\nu}_n - \nu) \rightarrow_d Z$. Find the distribution of Z in each case.
9. Suppose that we wish to estimate θ based on a random sample X_1, \dots, X_n of Bernoulli(θ) random variables. However, we are only able to obtain a random sample $(Y_i, R_i), \dots, (Y_n, R_n)$ where the R_i 's are iid Bernoulli(p_0) for known p_0 and $Y_i = R_i X_i$ for $i = 1, \dots, n$. Derive the MLEs $\hat{\theta}_a$, $\hat{\theta}_b$ and $\hat{\theta}_c$ for θ based on
- (a) The full data distribution of the X_i 's;
 - (b) The marginal distribution of the Y_i 's;
 - (c) The joint distribution of the (Y_i, R_i) 's.
10. **(Challenging)** Let T_n and U_n be estimators of θ such that

$$\begin{aligned}\sqrt{n}(T_n - \theta) &\rightarrow_d N(0, \sigma_T^2) \\ \sqrt{n}(U_n - \theta) &\rightarrow_d N(0, \sigma_U^2).\end{aligned}$$

The *asymptotic relative efficiency* of T_n with respect to U_n is σ_T^2/σ_U^2 .

Find the asymptotic distributions of the MLEs in 9(b) and 9(c) and calculate the asymptotic relative efficiency of $\hat{\theta}_b$ to $\hat{\theta}_c$. Which of the MLEs do you prefer for estimating θ ? Quantify the loss in efficiency of your preferred estimator to $\hat{\theta}_a$ that is based on the (unobserved) X_i 's. Explain.

R lab: Consistency of the sample median

This exercise is intended to reinforce concepts through use of the R software package.

We consider an example of a simulation study. The simulation consists of 1000 independent replications of the following process:

- i. Generate X_1, \dots, X_n iid $N(\mu, 1)$ with $\mu = 0$;
- ii. Estimate μ by calculating the sample median m_n ;
- iii. Record whether the statement $|m_n - \mu| < \epsilon$ is *true* or *false*.

Lastly, the proportion of times/1000 that the statement in iii. is *true* is calculated.

11. In R, the code below implements the simulation study for $n = 10$ and $\epsilon = 0.1$.

```

set.seed(50011)
result <- logical(length = 1000)
n <- 10
epsilon <- .1
for(i in 1:1000){
  X <- rnorm(n, mean = 0)
  m <- median(X)
  result[i] <- abs(m - 0) < epsilon
}
mean(result)

```

Note that the command `set.seed(50011)` ensures that you obtain the same results each time you run this set of commands.

Type the above commands into your R console (or write a script) and then:

- (a) Explore how the value of `mean(result)` changes by increasing the value of `n` in this code to, e.g. $n = 30, 50, 100, 200, 500, 1000$.
- (b) Referring to the results of your experimentation, comment on whether the sample median appears to be consistent for μ in this setting.