Problem Sheet 3 Solutions

MATH50011
Statistical Modelling 1

Week 4

Lecture 7 (Proof of MLE Consistency and Asymptotic Normality)

1. In the lecture notes, we saw that MLEs are asymptotically normal and sketched a proof of this (subject to regularity conditions). Many other estimators are also the solutions to estimating equations.

   Let $X_1, \dots, X_n$ be i.i.d. real-valued random variables and suppose that we wish to estimate the value of $\theta_0 \in \mathbb{R}$ defined as the unique $E[\psi(X_1, \theta)] = 0$ for a twice continuously differentiable function $\psi : \mathbb{R}^2 \to \mathbb{R}$. Define $\hat{\theta}_n$ as the unique solution to $\sum_{i=1}^n \psi(X_i, \theta) = 0$.

   Sketch a proof that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, \sigma^2(\theta_0))$ and find an expression for $\sigma^2(\theta_0)$. At what steps in your proof sketch do you need additional assumptions required to justify an operation?

Solution. We largely follow the steps for the proof of asymptotic normality of the MLE. We begin with a first-order Taylor expansion/mean value theorem so that

$$0 = \sum_{i=1}^n \psi(X_i, \hat{\theta}_n) = \sum_{i=1}^n \psi(X_i, \theta_0) + \sum_{i=1}^n \frac{\partial}{\partial \theta} \psi(X_i, \theta)\bigg|_{\theta = \tilde{\theta}_n} (\hat{\theta}_n - \theta_0)$$

for some $\tilde{\theta}_n$ between $\hat{\theta}_n$ and θ_0. Let

$$\dot{\Psi}(\tilde{\theta}_n) = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \psi(X_i, \theta)\bigg|_{\theta = \tilde{\theta}_n}.$$

After some rearranging, we have that

$$-\dot{\Psi}(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) = \frac{1}{n} \sum_{i=1}^n \psi(X_i, \theta_0).$$

The RHS above has mean zero by assumption. Provided

$$Var\{\psi(X_i, \theta_0)\} = E\{\psi(X_i, \theta_0)^2\} = A(\theta_0)$$

exists and is finite, we can apply the central limit theorem to find

$$-\dot{\Psi}(\tilde{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i, \theta_0) \to_d N(0, A(\theta_0)).$$

If we have that $-\dot{\Psi}(\tilde{\theta}_n) \to_p B(\theta_0) = E\left\{ \frac{\partial}{\partial\theta}\psi(X_i, \theta)\Big|_{\theta=\theta_0} \right\}$ where $0 < B(\theta_0) < \infty$, then by Slutsky's lemma

$$\frac{-\dot{\Psi}(\tilde{\theta}_n)}{B(\theta_0)} B(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, A(\theta_0)).$$

Hence, we also have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, B^{-2}(\theta_0)A(\theta_0))$$

so that $\sigma^2(\theta_0) = B^{-2}(\theta_0)A(\theta_0)$.

Note that:

- A 2nd-order Taylor expansion could also be used at the beginning of the proof, which then requires that the higher-order term converges to zero in probability.
- The assumption that $-\dot{\Psi}(\tilde{\theta}_n)$ converges in probability could be stated in similar equivalent ways (such as we did for the second partial derivatives of the log-likelihood in the lecture notes).
- For a regular parametric model, with $\psi(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta)$ we would have $A(\theta_0) = B(\theta_0) = I_f(\theta_0)$ so that $\sigma^2(\theta_0) = I_f(\theta_0)^{-1}$ as expected.

2. In the notation the R lab question (see below), define the one-step estimator

$$\hat{\theta}_n^{(1)} = T_n + I_n(T_n)^{-1}U_n(T_n).$$

Suppose that T_n is asymptotically normal and that $I_n(T_n)$ is consistent for the Fisher information $I_f(\theta)$. Show that

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_0) \to_d N(0, I_f(\theta_0)^{-1}).$$

Hint: use a first-order Taylor expansion of $U_n(\theta)$.

Solution. Recall that

$$U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(X_i)$$
$$I_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i)$$

Using the Taylor expansion with mean value remainder we have

$$U_n(\theta_0) = U_n(T_n) + \frac{\partial}{\partial\theta}U_n(\theta)\bigg|_{\theta=\tilde{T}_n}(\theta_0 - T_n) = U_n(T_n) - I_n(\tilde{T}_n)(\theta_0 - T_n)$$

for some $\tilde{T}_n$ between θ_0 and T_n. Rearranging this, we have

$$\begin{aligned} U_n(\theta_0) =& I_n(\tilde{T}_n)T_n + U_n(T_n) - I_n(\tilde{T}_n)\theta_0 \\ =& I_n(T_n)T_n + U_n(T_n) - I_n(T_n)\theta_0 \\ &+ [I_n(\tilde{T}_n) - I_n(T_n)](T_n - \theta_0). \end{aligned}$$

Moreover, we can then express this as

$$\sqrt{n}\{T_n + I_n(T_n)^{-1}U_n(T_n) - \theta_0\} = I_n(T_n)^{-1}\sqrt{n}U_n(\theta_0) - I_n(T_n)^{-1}[I_n(\tilde{T}_n) - I_n(T_n)]\sqrt{n}(T_n - \theta_0).$$

Observe that $I_n(T_n)^{-1} \to_p I_f(\theta_0)^{-1}$, $\sqrt{n}U_n(\theta_0) \to_d N(0, I_f(\theta_0))$, and $\sqrt{n}(T_n - \theta_0) \to N(0, \sigma^2(\theta_0))$ for some $\sigma^2(\theta_0)$. Further, under mild regularity conditions (like the sequence of functions $(I_n)_{n\in\mathbb{N}}$ converges uniformly to I_f or the functions I_ns are $K-$Lipschitz continuous with $K \in (0, \infty)$) we have $I_n(\tilde{T}_n) - I_n(T_n) \to_p 0$. Then, we obtain the RHS converges to $N(0, I_f(\theta_0)^{-1})$ by Slutsky's lemma. Hence, we have

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_0) = \sqrt{n}\{T_n + I_n(T_n)^{-1}U_n(T_n) - \theta_0\} \to_d N(0, I_f(\theta_0)^{-1}).$$

Lecture 8 (Confidence Intervals)

3. Dr. Jetson asked a random sample of 10000 UK households whether or not they own a robotic vacuum cleaner. She finds that 1300 of the households own a robotic vacuum and the other 8700 do not. Based on this data, she estimates that 13% of UK households own a robotic vacuum with a 95% confidence interval of 12.3% to 13.7%. Dr. Jetson tells you that

   > "There is a 95% probability that between 12.3% and 13.7% of UK households own a robotic vacuum cleaner."

   What is the main problem with the above statement? Provide a correct description of the confidence interval suitable for a non-statistician.

   Solution. The main problem with Dr. Jetson's statement is that it tries to make a probability statement based on the fixed interval and fixed proportion of UK households owning a robotic vacuum cleaner. The 0.95 probability refers to a property of the random confidence intervals. If Dr. Jetson repeated her survey many times and constructed 95% confidence intervals for the proportion each time, approximately 95% of the resulting intervals would contain the true proportion.

4. A random sample of 11 components in a factory is collected. The length in cm of each component is recorded below

   3.26 1.76 1.63 1.79 2.43 0.88 0.99 1.12 4.56 2.11 2.73

   Assume that the lengths are normally distributed with mean μ and variance σ^2. Construct a 99% confidence interval for μ.

   Solution. As pivotal quantity we use $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ which is t_{n-1} distributed, where $n = 11$, where $\bar{x}$ and s^2 are the observed sample mean and variance. Using a table (or a calculator/computer), this implies

   $$P\left(\left|\frac{\bar{x}-\mu}{s^2/\sqrt{n}}\right| \leq k\right) = 0.99,$$

   where $k = 3.169$. Hence, a 99% confidence interval for μ is $(\bar{x} - s/\sqrt{n}k, \bar{x} + s/\sqrt{n}k) = (1.07, 3.16)$ (using $\bar{x} = 2.114545$ and $s^2 = 1.201427$).

5. Let $Y_1, \ldots, Y_n$ be i.i.d. Exp(λ), where $\lambda > 0$ is unknown.

   (a) Show that $2\lambda \sum_{i=1}^n Y_i$ has a χ^2-distribution with $2n$ degrees of freedom;

   (b) Derive a $(1-\alpha) \times 100\%$ confidence interval for λ;

(c) Using the following observations, compute a 95% confidence interval for λ.

1.04 1.39 0.1 2.04 4.73 0.89 0.51 0.89 0.66 0.93

(Note: for $X \sim \chi^2_{20}$, $P(X \leq 9.59) = 0.025$ and $P(X \leq 34.17) = 0.975$.)

Solution. Note that $2\lambda Y_i$ is Exp($\frac{1}{2}$), which is χ^2_2. Since the Y_i's are independent, $2\lambda \sum Y_i$ is the sum of n independent χ^2_2 random variables and is therefore distributed as χ^2_{2n}. Hence,

$$P(c_1 < 2\lambda \sum Y_i < c_2) = 1 - \alpha,$$

where $0 < \alpha < 1$ and where $P(X < c_1) = P(X > c_2) = \frac{\alpha}{2}$ and $X \sim \chi^2_{2n}$. A $1 - \alpha$ confidence interval for λ is thus given by $\left(\frac{c_1}{2\sum y_i}, \frac{c_2}{2\sum y_i}\right)$.

Hence, the confidence interval for the data based on $\sum Y_i$ is: $(0.36, 1.3)$

6. Find an approximate 95% confidence interval for the odds that a randomly selected UK household owns a robotic vacuum based on the data in Question 3. (Hint: use the delta method.)

Solution. A point estimate for the odds will be $0.13/0.87 = 0.1494253$.

Using the delta method, we know that with $Odds(p) = p/(1-p)$ and $Odds'(p) = 1/(1-p)^2$ the standard error for the odds will be

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n(1-\hat{p})^4}} = \sqrt{\frac{0.13(1-0.13)}{10000(1-0.13)^4}} = 0.004763577$$

which leads to an approximate 95% confidence interval with limits

$$0.1494253 \pm 1.96 \times 0.004763577$$

or $(0.1400887, 0.1587619)$.

Alternatively, since the odds are an invertible function of the probability, we can apply $p/(1-p)$ to the limits of the original interval and still have a valid 95% confidence interval. This leads to the very similar interval estimate

$$(0.1402509, 0.1587486)$$

7. Use the Bonferroni correction to find a 95% confidence region for (μ, σ^2) based on a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ distribution. Apply your result to construct a 95% confidence region for (μ, σ^2) based on the data in Question 4.

Solution. We will make use of example 27 from the lecture notes for a normal random sample with both μ and σ unknown. To obtain a 95% confidence region for (μ, σ^2) with the Bonferroni correction, we will construct two-sided (1-0.05/2)100% = 97.5% confidence intervals for each parameter.

Let $\bar{X}$ and S^2 be the sample mean and variance of the X_is.

The 97.5% confidence interval for μ has the form

$$\bar{X} \pm t_{n-1,0.9875}\frac{S}{\sqrt{n}} \implies \left(\bar{X} - t_{n-1,0.9875}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,0.9875}\frac{S}{\sqrt{n}}\right)$$

for $t_{n-1,0.9875}$ the value such that the $P(T_{n-1} \leq t_{n-1,0.9875}) = 0.9875 = 1 - 0.025/2$.

The 97.5% confidence interval for σ^2 has the form

$$\left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$$

where $P(\chi^2_{n-1} \leq c_1) = 0.0125$ and $P(\chi^2_{n-1} \leq c_2) = 0.9875$.

We have, by the Bonferroni correction, that

$$\left(\bar{X} - t_{n-1,0.9875}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,0.9875}\frac{S}{\sqrt{n}}\right) \times \left(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1}\right)$$

is a 95% confidence region for (μ, σ^2).

We use $\bar{x} = 2.114545$, $s^2 = 1.201427$, $n = 11$ to find that $t_{10,0.9875} = 2.633767$, $c_1 = 2.707213$, and $c_2 = 22.55825$. This results in the 95% confidence region (simultaneous confidence intervals)

$$(1.2441242.984967) \times (0.5325889, 4.4378752) \approx (1.2, 3.0) \times (0.5, 4.4).$$

R lab: One-Step Estimators

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

In the notes, we saw that numerical methods can facilitate maximisation of the (log) likelihood. In this lab, we illustrate how a simple one-step update to an initial estimator can lead to an accurate approximation of the MLE. The step we take is based on Newton's method.

Suppose that $X_1, \ldots, X_n$ are iid with pdf $f_\theta(x)$. Define

$$U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(X_i)$$
$$I_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i)$$

The one-step estimator is defined as $\hat{\theta}_n^{(1)} = T_n - I_n(T_n)^{-1}U_n(T_n)$, where T_n is an initial estimator of θ. If T_n is an asymptotically normal estimator of θ, then

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta) \to_d N(0, I_f(\theta)^{-1}).$$

You will prove this in the next problem sheet.

8. In this exercise, you will implement a simulation study to explore the behavior of the one-step estimator for the location parameter θ of the Cauchy(θ) distribution with pdf

$$f_\theta(x) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]} \quad -\infty < x < \infty, -\infty < \theta < \infty.$$

Note that $f_\theta(x)$ is symmetric about θ. However, $E_\theta(X)$ does not exist for the Cauchy distribution so the sample mean would be an awful estimator here. Instead, we will use the sample median as an initial estimator of θ.

After drawing $X_1, \dots, X_n$ i.i.d. Cauchy(θ), the sample median $\hat{m}_n$ will be computed and stored as an initial estimator. The values of $U_n(\hat{m}_n)$ and $I_n(\hat{m}_n)$ are then computed and used to construct a one-step estimator $\hat{\theta}_n^{(1)}$ based on $\hat{m}_n$. This experiment will be independently replicated a total of 1000 times, so that we can approximate the sampling distributions of $\hat{m}_n$ and $\hat{\theta}_n^{(1)}$.

The R code below implements the simulation study for $n = 10$ and $\theta = 0$.

```
set.seed(50011)
result.m <- logical(length = 1000)
result.t1 <- logical(length = 1000)
n <- 10
theta <- 0
for(i in 1:1000){
X <- rcauchy(n, location = 0)
m <- median(X)
U <- NULL
I <- NULL
t1 <- m - U/I
result.m[i] <- sqrt(n)*(m-theta)
result.t1[i] <- sqrt(n)*(t1-theta)
}
hist(result.m, freq=FALSE)
hist(result.t1, freq=FALSE)
```

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 an R script and then:

(a) Derive expressions for $U_n(\hat{m}_n)$ and $I_n(\hat{m}_n)$ in terms of X and m. Use your expressions to replace the appropriate NULL definitions in the for loop.

Solution. Taking derivatives of the log-likelihood we find

$$U_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i) = \sum_{i=1}^{n} \frac{2(X_i - \theta)}{1 + (X_i - \theta)^2}$$

and

$$I_n(\theta) = -\frac{\partial}{\partial\theta} U_n(\theta) = -\sum_{i=1}^{n} 2\frac{1 - (X_i - \theta)^2}{[1 + (X_i - \theta)^2]^2}.$$

In the code above, we can assign

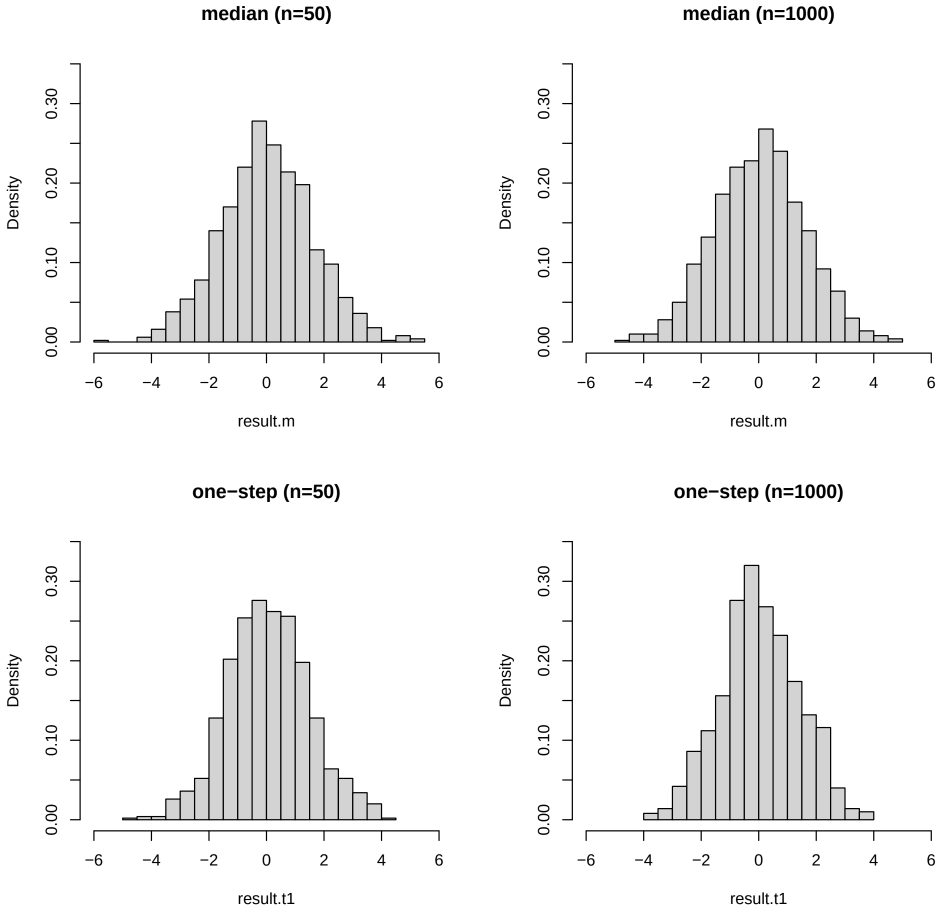
```
U <- 2*sum((X-m)/(1+(X-m)^2))
I <- -2*sum((1-(X-m)^2)/(1+(X-m)^2)^2)
```

(b) Comment on why it is reasonable to store the values of $\sqrt{n}(\hat{m}_n - \theta)$ and $\sqrt{n}(\hat{\theta}_n^{(1)} - \theta)$ instead of $\hat{\theta}_n^{(1)}$ and $\hat{m}_n$.

Solution. We are concerned about convergence of the scaled and centered sequences. We can always solve for the estimators based on the stored values.

(c) Explore how each histogram changes by increasing the value of `n` in this code to, e.g. $n = 30, 50, 100, 200, 500, 1000$. You might also compare other, say numerical, summaries (e.g. mean, variance, quantiles).

Solution. The histograms you generate should suggest that the median has a sampling distribution with slightly higher spread. See below for examples for $n = 50$ and $n = 1000$. In particular, for $n = 1000$, all of the bins for the one-step estimator are contained in the interval [-4,4] whereas the histogram for the median extends beyond this interval.



(d) Referring to your results from (c), comment on whether you prefer the sample median or one-step estimator for estimating θ in this setting.

Solution. Using the histograms from the 1000 simulation experiments as approximations to the sampling distribution, both histograms appear to be centered near zero for large sample sizes. However, the one-step estimator is less variable in larger sample sizes. We would prefer the one-step estimator based on these observations.

Challenge Do your simulations provide evidence that $\sqrt{n}(\hat{\theta}_n^{(1)} - \theta)$ convergences in distribution to a $N(0, I_f(\theta)^{-1})$ random variable? Explain your answer using appropriate graphical and/or numerical evidence.

Solution. The histogram for $n = 1000$ may not look immediately like a normal distribution (it is not quite symmetric or bell-shaped), but this may be due to Monte Carlo error. We can also overlay the density of a normal distribution to help with our assessment.

See Figure 1 below for 9 independent replications of the experiment. Repeating the experiment more than 1000 times (changing the definition of the loop and results vector) could be necessary to get a reliable picture of the sampling distribution. In Figure 2, we see that by increasing the number of replications of the experiment to 9000 there is much less variability between each histogram (and each histogram fits fairly well to the $N(0, I(\theta)^{-1})$ density).

Simulation studies such as this one are common in statistical research. The Stochastic Simulations module in Year 3 explores such ideas in greater detail.

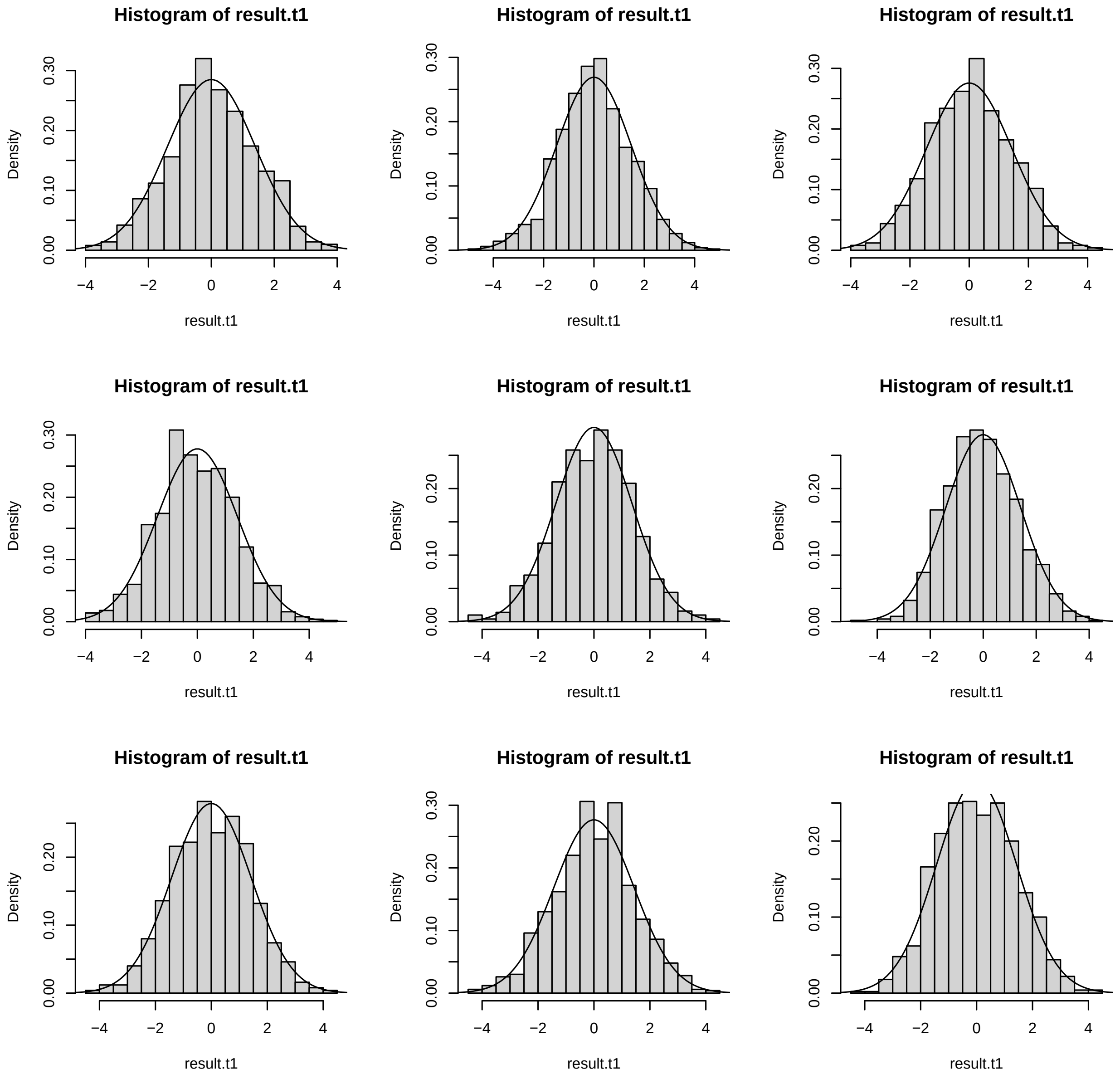


Figure 1: Nine histograms generated from independent runs with 1000 replications of the $n = 1000$ Cauchy one-step estimator.

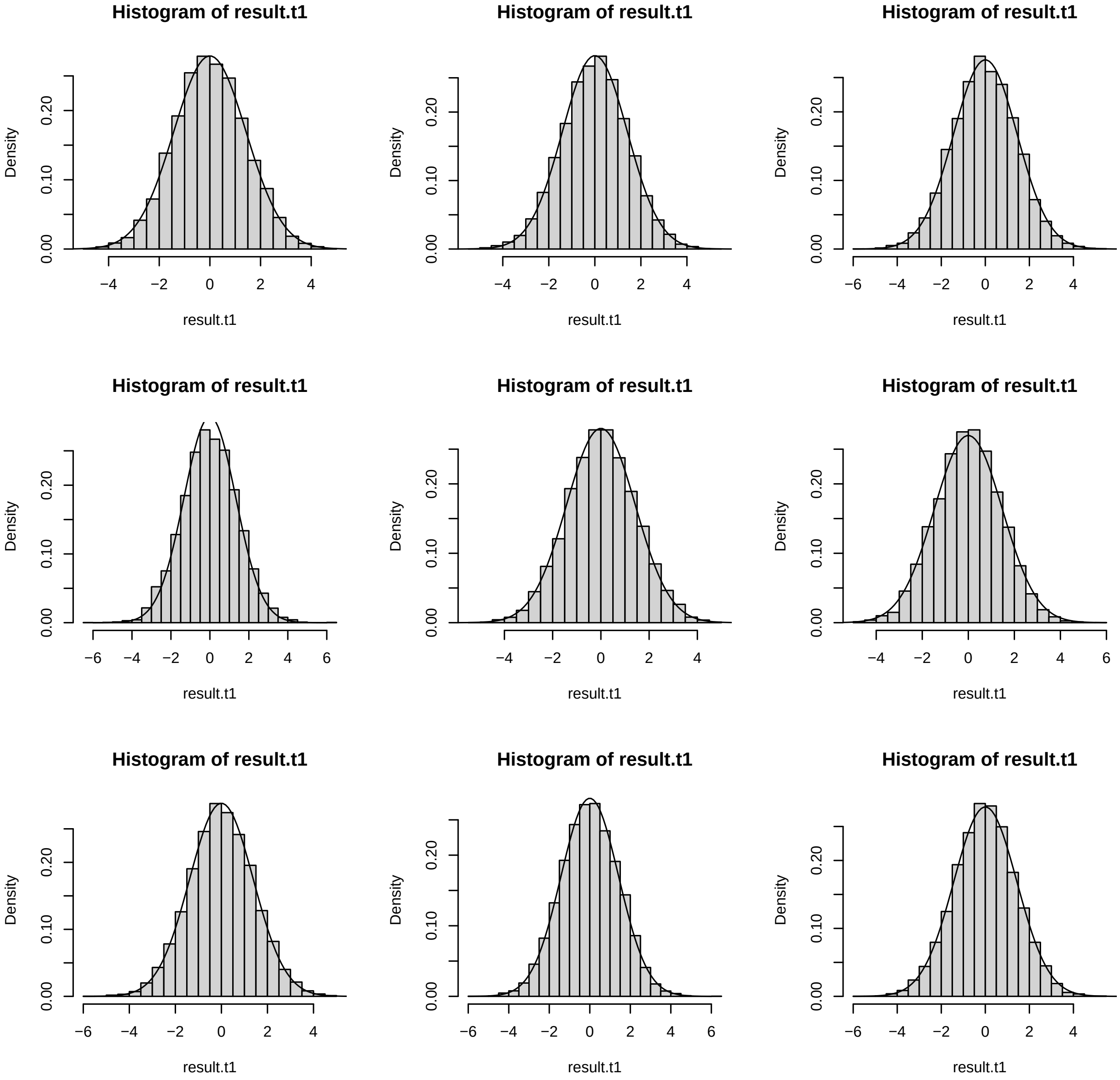


Figure 2: Nine histograms generated from independent runs with 9000 replications of the $n = 1000$ Cauchy one-step estimator.