

Statistical analysis for rounded data

Zhidong Bai

joint work with Baozue Zhang and Shurong Zheng

Applied Probability & Statistics
National University of Singapore

Outline

- Backgrounds
- Model and Problem
- Our Estimation Procedure
- Two important Lemmas in Proof
- Consistency and Asymptotic Normality
- Simulation results and Q-Q plots
- An Example

Backgrounds (Model)

As a simple example, we consider the rounded data $\{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n\}$ which are from

$\{x_1, x_2, \dots, x_n\} \stackrel{i.i.d.}{\sim} N(\mu_0, \sigma^2)$. The parameters μ_0 and σ^2 are estimated by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (\tilde{x}_i - \hat{\mu})^2.$$

Let Rejec-Prob denote the rejection probability

$$P_{\mu_0} \left(\frac{\sqrt{n} |\hat{\mu} - \mu_0|}{\hat{\sigma}} \geq t_{n-1, 0.975} \right)$$

where $t_{n-1, 0.975}$ is the 97.5% quantile of t-distribution with the degree of $n - 1$.

Backgrounds (Table 1)

	(μ_0, σ^2)	$(2.15, 0.25^2)$	$(2.25, 0.25^2)$	$(2.35, 0.25^2)$
n=10000	$(\hat{\mu}, \sqrt{MSE})$	(2.076, 0.074)	(2.157, 0.093)	(2.274, 0.076)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.017)	(0.135, 0.073)	(0.200, 0.137)
	Rejec-Prob	1.0	1.0	1.0
n=200	$(\hat{\mu}, \sqrt{MSE})$	(2.076, 0.077)	(2.157, 0.096)	(2.274, 0.082)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.024)	(0.135, 0.075)	(0.200, 0.138)
	Rejec-Prob	0.92	0.91	0.67
n=20	$(\hat{\mu}, \sqrt{MSE})$	(2.076, 0.097)	(2.158, 0.123)	(2.274, 0.125)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.057)	(0.135, 0.093)	(0.200, 0.145)
	Rejec-Prob	0.19	0.36	0.16
n=10	$(\hat{\mu}, \sqrt{MSE})$	(2.076, 0.116)	(2.158, 0.149)	(2.274, 0.161)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.079)	(0.136, 0.111)	(0.200, 0.154)
	Rejec-Prob	0.43	0.18	0.20

Backgrounds (Table 2)

	(μ_0, σ^2)	$(2.85, 0.25^2)$	$(2.75, 0.25^2)$	$(2.65, 0.25^2)$
n=10000	$(\hat{\mu}, \sqrt{MSE})$	(2.924, 0.074)	(2.843, 0.093)	(2.726, 0.076)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.017)	(0.135, 0.073)	(0.200, 0.137)
	Rejec-Prob	1.0	1.0	1.0
n=200	$(\hat{\mu}, \sqrt{MSE})$	(2.924, 0.077)	(2.843, 0.096)	(2.726, 0.082)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.024)	(0.135, 0.075)	(0.200, 0.138)
	Rejec-Prob	0.92	0.91	0.67
n=20	$(\hat{\mu}, \sqrt{MSE})$	(2.923, 0.096)	(2.842, 0.123)	(2.725, 0.125)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.080, 0.057)	(0.136, 0.093)	(0.200, 0.145)
	Rejec-Prob	0.18	0.36	0.16
n=10	$(\hat{\mu}, \sqrt{MSE})$	(2.923, 0.116)	(2.842, 0.148)	(2.725, 0.160)
	$(\hat{\sigma}^2, \sqrt{MSE})$	(0.079, 0.079)	(0.136, 0.110)	(0.200, 0.155)
	Rejec-Prob	0.43	0.17	0.20

Backgrounds (Results)

From Tables 1 and 2, we can see

- When the sample size n is larger than 20, the estimation results **is not improved**. The estimates $\hat{\mu}$ and $\hat{\sigma}^2$ are **inconsistent**.
- As $n \rightarrow +\infty$, it is easily to **reject**

$$H_0 : \mu = \mu_0 \text{ v.s. } H_1 : \text{not } H_0.$$

- If μ_0 satisfies $m < \mu_0 < m + 0.5$ for any given integer m , then $\hat{\mu}$ **underestimates** μ_0 ; if $m + 0.5 < \mu_0 < m + 1$, then $\hat{\mu}$ **overestimates** μ_0 .

Model and Problem

We mainly consider statistical inference of **rounded data** from the following **MA(p)** and **AR(p)** models which are as follow

$$X_t = c + \epsilon_t + \phi_1 \epsilon_{t-1} + \cdots + \phi_p \epsilon_{t-p} \quad (1)$$

and

$$X_{t+1} = c + \phi_1 \cdot X_t + \cdots + \phi_p X_{t-p+1} + \epsilon_{t+1} \quad (2)$$

where $\epsilon_t \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ for $t = 1, \cdots, n$, $\phi = (\phi_1, \cdots, \phi_p)$.

Without loss of generality, assume that only rounded data

$$\tilde{\mathbf{X}} = (\tilde{X}_1, \cdots, \tilde{X}_n)$$

can be **observed** from $\mathbf{X} = (X_1, \cdots, X_n)$ which are **unobservable**.

Model and Problem

Make statistical inference for unknown parameters (c, ϕ, σ^2) based on the rounded data

$$\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_n).$$

and the above MA(p) and AR(p) models. We can obtain some properties of estimates $(\hat{c}, \hat{\phi}, \hat{\sigma}^2)$ based on our estimation procedure:

- consistency $(\hat{c}, \hat{\phi}, \hat{\sigma}^2) \xrightarrow{a.s.} (c, \phi, \sigma^2)$
- asymptotic normality

Our Estimation Procedure

Define k subsets of the rounded data as follows:

$$(1) \tilde{X}_1 \cdots \tilde{X}_{p+1} \tilde{X}_{k+1} \cdots \tilde{X}_{k+p+1} \cdots \tilde{X}_{(m-1)k+1} \cdots \tilde{X}_{(m-1)k+p+1};$$

$$(2) \tilde{X}_2 \cdots \tilde{X}_{p+2} \tilde{X}_{k+2} \cdots \tilde{X}_{k+p+2} \cdots \tilde{X}_{(m-1)k+2} \cdots \tilde{X}_{(m-1)k+p+2};$$

.....

$$(k) \tilde{X}_k \cdots \tilde{X}_{k+p} \tilde{X}_{2k} \cdots \tilde{X}_{2k+p} \cdots \tilde{X}_{mk} \cdots \tilde{X}_{mk+p};$$

where $m = \lceil (n - p)/k \rceil$. Let

$$p_i \triangleq P(\tilde{X}_j - 0.5 \leq X_j < \tilde{X}_j + 0.5, j = 1, \dots, p + 1).$$

Our Estimation Procedure

Three steps:

- **Step** (1). Note that

$$(\tilde{X}_1 \cdots \tilde{X}_{p+1})(\tilde{X}_{2p+2} \cdots \tilde{X}_{3p+2}) \cdots (\tilde{X}_{(m-1)(p+1)+1} \cdots \tilde{X}_{m(p+1)})$$

form a sample of m iid. $p + 1$ -dimensional random vectors. Denote by $n_{\mathbf{i}}$ the frequency of \mathbf{i} in this **sub-sample**. Then based on this sub-sample, the MLE of parameters can be obtained by

$$\text{maximizing } \sum n_{\mathbf{i}} \log p_{\mathbf{i}}.$$

Denote the MLE by $(\hat{c}_1, \hat{\sigma}_1^2, \hat{\phi}_1)$;

Our Estimation Procedure

- **Step** (2). Similarly, by the j -th subset of the sample, we can construct an MLE $(\hat{c}_j, \hat{\sigma}_j^2, \hat{\phi}_j)$, $j = 2, \dots, p + 1$.
- **Step** (3). Take the averages of the MLE's as our estimators of parameters, *i.e.*

$$\hat{c} = \frac{\sum_{i=1}^{p+1} \hat{c}_i}{(p+1)}, \quad \hat{\phi} = \frac{\sum_{i=1}^{p+1} \hat{\phi}_i}{(p+1)}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{p+1} \hat{\sigma}_i^2}{(p+1)}.$$

Two Important Lemmas

Lemma 1. If X_1, X_2, \dots, X_n is a sample of an $AR(1)$ model with autoregressive coefficient ϕ ($|\phi| < 1$) and normally distributed innovations, then

$$|g(\mathbf{x}_1, \mathbf{x}_2) - g(\mathbf{x}_1)g(\mathbf{x}_2)| \leq K g(\mathbf{x}_1)g(\mathbf{x}_2)|\phi|^k \quad (3)$$

$$\cdot \exp \left\{ \frac{(1 - \phi^2)|\phi|^k}{2\sigma^2(1 + \phi^{2k})} \left((x_{i_2} - \mu)^2 + (x_{i_3} - \mu)^2 \right) \right\} \\ \cdot \left(1 + |x_{i_2} - \mu|^2 + |x_{i_3} - \mu|^2 \right),$$

where $g(\mathbf{x}_1, \mathbf{x}_2)$, $g(\mathbf{x}_1)$ and $g(\mathbf{x}_2)$ are the **joint densities** of $(X_{i_1}, X_{i_1+1}, \dots, X_{i_2}, X_{i_3}, X_{i_3+1}, \dots, X_{i_4})$, $(X_{i_1}, X_{i_1+1}, \dots, X_{i_2})$ and $(X_{i_3}, X_{i_3+1}, \dots, X_{i_4})$, respectively. Also, $k = i_2 - i_1$. Here K is an absolute constant depending on k and ϕ only.

Two Important Lemmas (Continue)

Furthermore, for $\ell > 2$,

$$\begin{aligned} |g(\mathbf{x}_1, \dots, \mathbf{x}_\ell) - \prod_{t=1}^{\ell} g(\mathbf{x}_t)| &\leq K \sum_{t=1}^{\ell-1} |\phi|^{k_t} g_t(\mathbf{x}_1, \dots, \mathbf{x}_\ell) \\ &\cdot \exp \left\{ \frac{(1 - \phi^2) |\phi|^{k_t}}{2\sigma^2 (1 + \phi^{2k_t})} \left((x_{2t} - \mu)^2 + (x_{2t+1} - \mu)^2 \right) \right\} \\ &\cdot \left[1 + |x_{2t} - \mu|^2 + |x_{2t+1} - \mu|^2 \right] \end{aligned}$$

where $\mathbf{x}_t = (x_{i_{2t-1}}, \dots, x_{i_{2t}})$, $k_t = i_{2t+1} - i_{2t}$,
 $g_t(\mathbf{x}_1, \dots, \mathbf{x}_\ell) = g(\mathbf{x}_1, \dots, \mathbf{x}_t)g(\mathbf{x}_{t+1}) \cdots g(\mathbf{x}_\ell)$ and K is the same as in (3).

Two Important Lemmas (Continue)

Lemma 2. Suppose the assumptions of Lemma 1 hold with $|\phi| < 1$. Assume that f is a k -dimensional measurable function such that $E f(X_i) = 0$ and $E f(X_i) f(X_i)^T = \gamma_0$ exists. If $\mathbf{V} = \gamma_0 + \sum_{j=1}^{\infty} [\gamma_j + \gamma_j^T]$ exists and is positive definite, then, with $Z_i = f(X_i)$,

$$\frac{1}{\sqrt{n}}(Z_1 + \cdots + Z_n) \xrightarrow{L} N(\mathbf{0}, \mathbf{V})$$

where $\gamma_j = E f(X_1) f(X_{1+j})^T = \gamma_{-j}^T$.

Consistency and Asymptotic Normality

- **Theorem 1.** The estimates $(\hat{c}, \hat{\phi}, \hat{\sigma}^2)$ obtained by proposed estimation procedure based on $\tilde{X}_1, \dots, \tilde{X}_n$ are **consistent**.
- **Theorem 2.** Under some conditions, then the AMLE $(\hat{c}, \hat{\phi}, \hat{\sigma}^2)$ are asymptotically multivariate normally distributed, that is,

$$\sqrt{\frac{n}{m}} \begin{pmatrix} \hat{c} - c \\ \hat{\phi} - \phi \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \sim N(0, \mathbf{I}^{-1}(\theta) \mathbf{V}_p \mathbf{I}^{-1}(\theta))$$

where $I(\theta)$ and \mathbf{V}_p are given in the proof.

Simulation Setup

Simulation Setup:

- Simulation Model: AR(1) model

$$X_t = \phi X_{t-1} + \epsilon_t$$

where $(\epsilon_1, \dots, \epsilon_n) \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$.

- Rounded data are as follow $\tilde{X}_1, \dots, \tilde{X}_n$.
- Parameter configuration, parameter estimates and their MSEs are listed in [Table 3](#).
- The [Q-Q plots](#) of $\hat{\phi}$ based on different parameter are given.

Simulation Results

Table 3. Simulation Results ($\sigma^2 = 1.0$)

	$n(m)$	$\hat{\phi}$ (MSE)	$\hat{\sigma}^2$ (MSE)
$\phi = 0.3$	500(5)	0.300(0.0026)	0.999(0.0012)
	500(10)	0.298(0.0028)	0.9892(0.0015)
$\phi = 0.5$	500(5)	0.499(0.0018)	0.990(0.0012)
	500(20)	0.464(0.0027)	1.002(0.0015)
	1000(10)	0.492(0.0010)	0.991(0.0008)
	1005(15)	0.493(0.0009)	0.991(0.0009)
	1000(20)	0.488(0.0012)	0.995(0.0007)

Simulation Results

Table 3. (Continue)

$\phi = 0.5$	2000(20)	0.493(0.0005)	0.993(0.0003)
	10000(20)	0.499(0.0001)	0.998(0.00005)
	1000(5)	0.738(0.0007)	0.988(0.0008)
	1000(10)	0.732(0.0009)	0.987(0.0009)
$\phi = 0.75$	1000(20)	0.717(0.0017)	0.998(0.0008)
	1000(40)	0.670(0.0072)	1.018(0.0014)
	2000(5)	0.744(0.0003)	0.990(0.0005)
$\phi = 0.9$	1000(5)	0.842(0.0064)	0.972(0.0038)
	1000(10)	0.804(0.013)	0.989(0.0045)

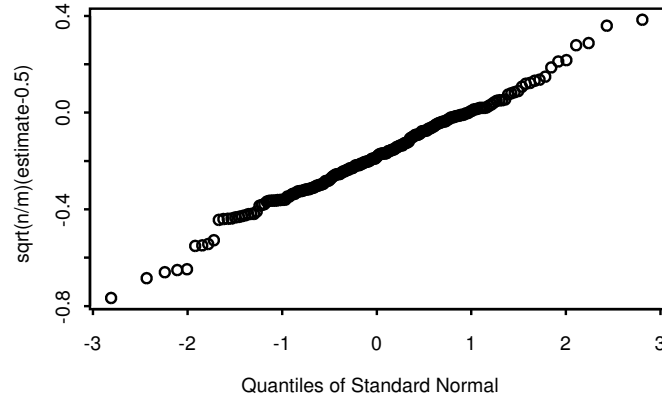
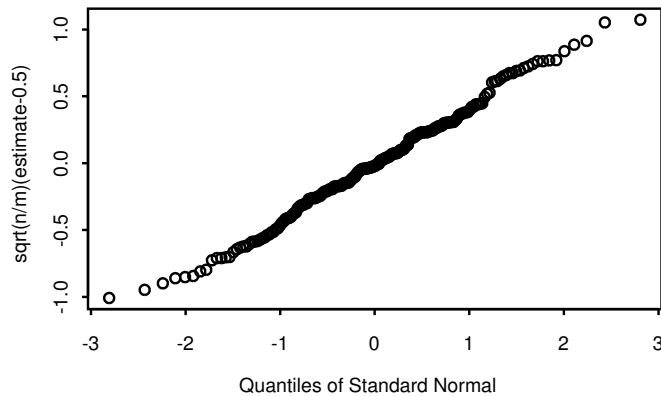
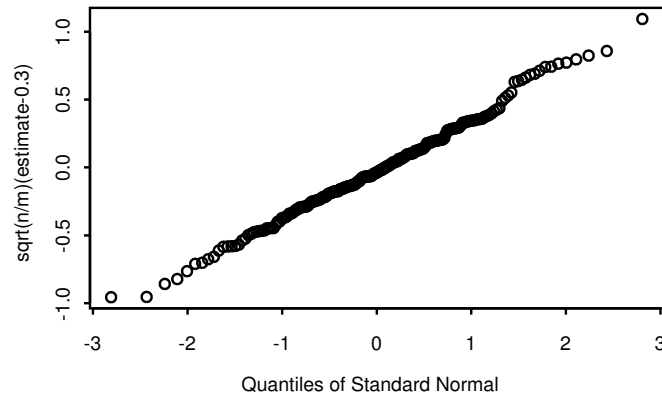
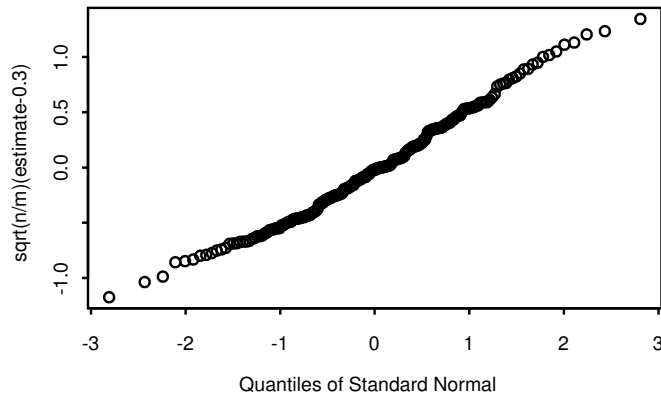
Q-Q plots 1 – 4 for (ϕ, n, m)

up-left: $(0.3, 500, 5)$;

up-right: $(0.3, 500, 10)$

down-left: $(0.5, 500, 5)$;

down-right: $(0.5, 500, 20)$



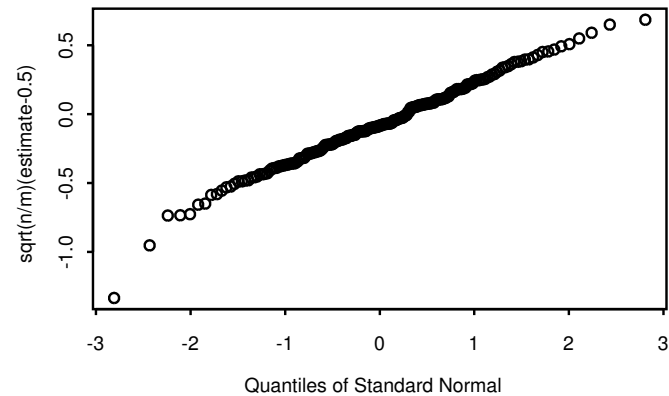
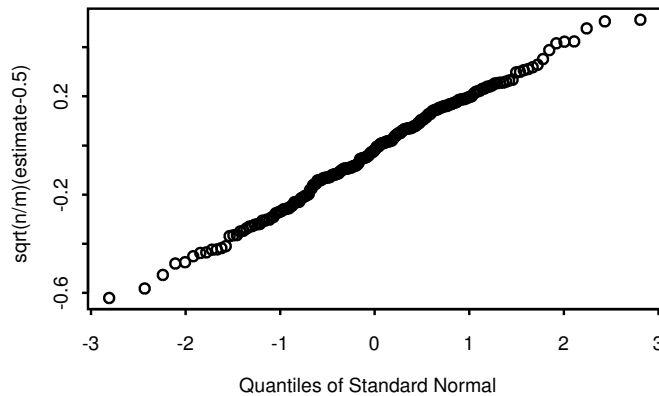
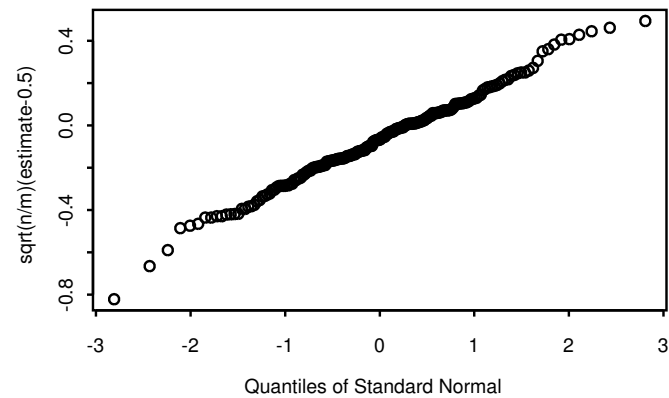
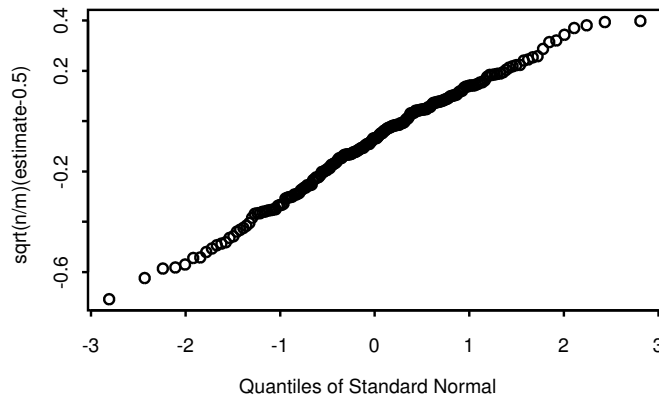
Q-Q plots 5 – 8 for (ϕ, n, m)

up-left: $(0.5, 1000, 20)$;

down-left: $(0.5, 10000, 20)$;

up-right: $(0.5, 2000, 20)$

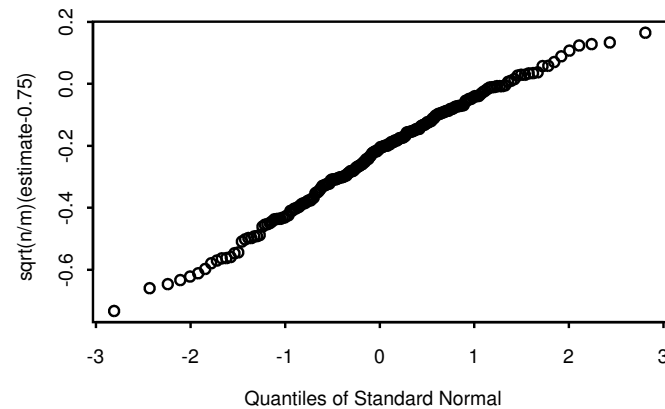
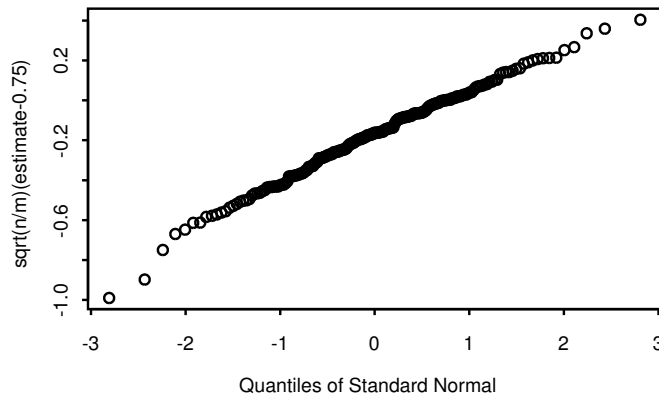
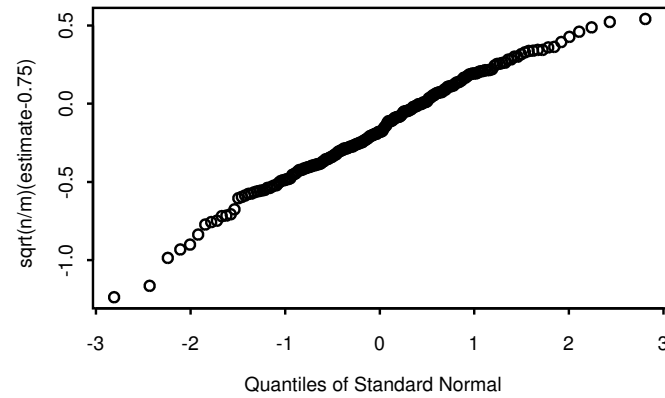
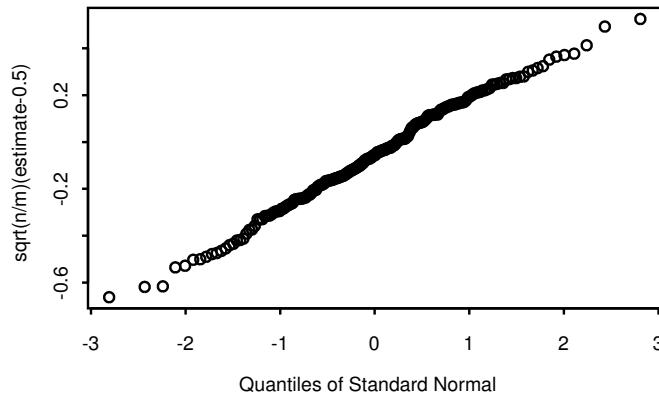
down-right: $(0.5, 1000, 10)$



Q-Q plots 9 – 12 for (ϕ, n, m)

up-left: $(0.5, 1005, 15)$;
down-left: $(0.75, 1000, 10)$;

up-right: $(0.75, 1000, 5)$
down-right: $(0.75, 1000, 20)$



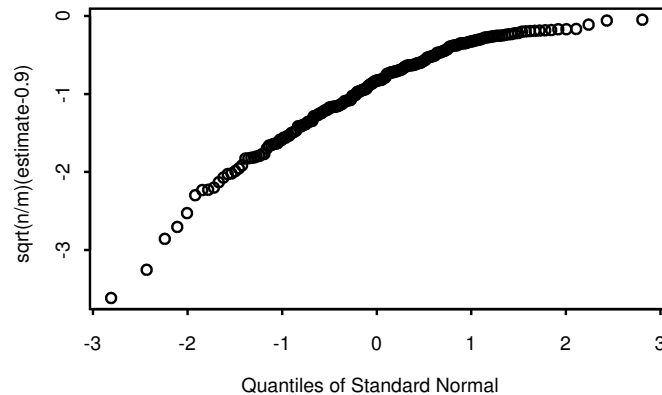
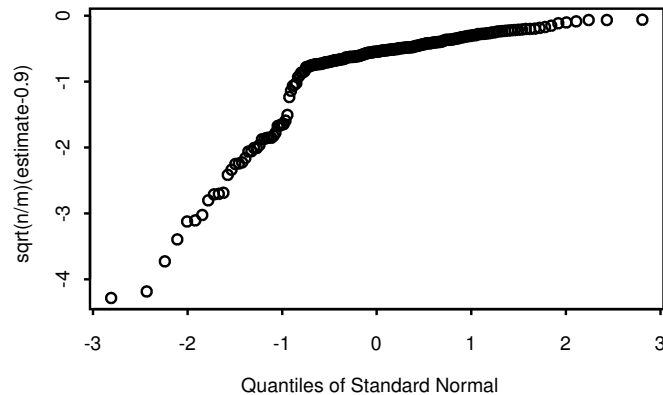
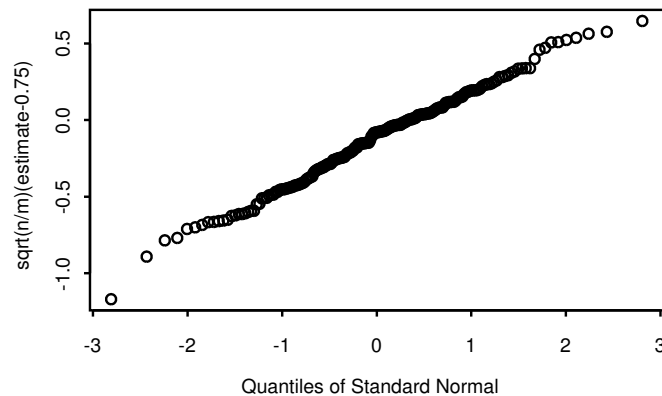
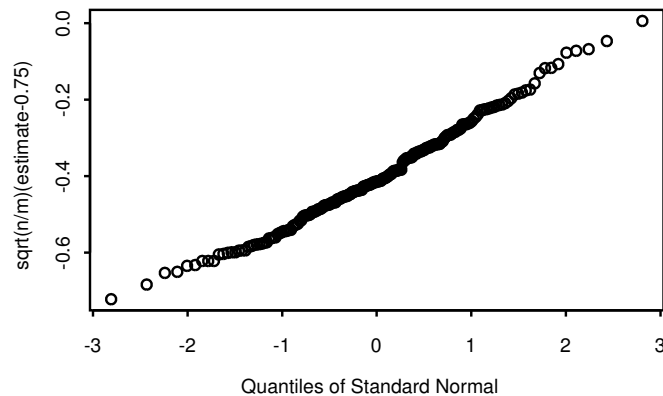
Q-Q plots 13 – 16 for (ϕ, n, m)

up-left: $(0.75, 1000, 40)$;

down-left: $(0.9, 1000, 5)$;

up-right: $(0.75, 2000, 5)$

down-right: $(0.9, 1000, 10)$



An Example

- Origin of Data: 226 chemical temperature readings per minute take values at most to one digit after decimal point from Table Series C from the book *Time Series Analysis (Forecasting and Control)* (Box *et al.* 1994, Third Edition, P544).

Data: (x_1, \dots, x_{226}) .

- This book (P189) suggests that the data satisfy the following AR(1) model

$$\nabla x_t = \phi \nabla x_{t-1} + \epsilon_t$$

where ∇ denotes the first difference notation, that is,

$$\nabla x_t = x_t - x_{t-1}.$$

An Example (Continue)

Results of Example:

- Estimate of ϕ from this book: $\hat{\phi} = 0.8$.
- Estimate of ϕ from our estimation procedure: $\tilde{\phi} = 0.716$.
- Difference of $\hat{\phi}$ and $\tilde{\phi}$: 0.084.
- Reason for the difference: whether to use conventional methods to directly analyze the rounded data or not.

Thank you!