

时间序列分析 第二次作业

辛柏瀛 2020111753

2023-03-13

第一题

对下列 ARIMA 模型，求 $E(\nabla Z_t)$ 和 $Var(\nabla Z_t)$.

$$Z_t = 5 + 2Z_{t-1} - 1.7Z_{t-2} + 0.7Z_{t-3} + a_t - 0.5a_{t-1} + 0.25a_{t-2}.$$

$$\nabla Z_t = 5 + \nabla Z_{t-1} - 0.7 \nabla Z_{t-2} + a_t - 0.5 a_{t-1} + 0.25 a_{t-2}$$

$$\text{记作: } Y_t = 5 + Y_{t-1} - 0.7 Y_{t-2} + a_t - 0.5 a_{t-1} + 0.25 a_{t-2} \quad (*)$$

可认为 $Y_t \sim \text{ARMA}(2,2)$. 其中 $|\phi_2| + |\phi_1| < 1$, $|a_2| < 1$. 是平稳的

$$(*) \text{ 左右同取期望: } \mu_y = 5 + \mu_y - 0.7 \mu_y \Rightarrow \mu = \frac{50}{7} \Rightarrow E(\nabla Z_t) = \frac{50}{7}.$$

将 Y_t 作中心化处理方便起见重新记为 y_t . 由二阶矩性质知不改变其二阶矩

$$(*) \text{ 左右同乘 } y_t \text{ 取期望: } \gamma_0 = 5\mu_y + \gamma_1 - 0.7\gamma_2 + EA_t y_t - 0.5 EA_{t-1} y_t + 0.25 EA_{t-2} y_t \quad ①$$

$$\sim \text{ 同乘 } y_{t-1} \sim : \gamma_1 = 5\mu_y + \gamma_0 - 0.7\gamma_1 + EA_t y_{t-1} - \frac{1}{2} EA_{t-1} y_{t-1} + \frac{1}{4} EA_{t-2} y_{t-1} \quad ②$$

$$\sim \text{ 同乘 } y_{t-2} \sim : \gamma_2 = 5\mu_y + \gamma_1 - 0.7\gamma_0 + EA_t y_{t-2} - \frac{1}{2} EA_{t-1} y_{t-2} + \frac{1}{4} EA_{t-2} y_{t-2} \quad ③$$

$$\text{其中 } EA_j y_k = 0 \quad (j > k > 0); \quad EA_{t-j} y_k = \sigma_a^2 = EA_{t-1} y_{t-1} = EA_{t-2} y_{t-2}$$

$$EA_{t-1} y_t = \sigma_a^2 - 0.5 \sigma_a^2 = 0.5 \sigma_a^2 = EA_{t-2} y_{t-1}; \quad EA_{t-2} y_t = -0.7 \sigma_a^2 + 0.75 \sigma_a^2 + 0.5 \sigma_a^2 \\ = 0.05 \sigma_a^2$$

$$\text{故 } ① \sim ③ \text{ 整理有: } \begin{cases} \gamma_0 = \gamma_1 - 0.7\gamma_2 + \sigma_a^2 - 0.25 \sigma_a^2 + \frac{0.05}{4} \sigma_a^2 \\ \gamma_1 = \gamma_0 - 0.7\gamma_1 - \frac{1}{2} \sigma_a^2 + \frac{1}{8} \sigma_a^2 \\ \gamma_2 = \gamma_1 - 0.7\gamma_0 + \frac{1}{4} \sigma_a^2 \end{cases}$$

$$\Rightarrow \begin{cases} \gamma_0 = 1.4631 \sigma_a^2 \\ \gamma_1 = 0.6989 \sigma_a^2 \\ \gamma_2 = 0.1453 \sigma_a^2 \end{cases}$$

$$\Rightarrow \text{Var}(\nabla Z_t) = 1.4631 \sigma_a^2$$

第二题

考虑两个模型: 模型 A($Z_t = 0.9Z_{t-1} + 0.09Z_{t-2} + a_t$) 和模型 B($Z_t = Z_{t-1} + a_t - 0.1a_{t-1}$).

(a) 识别每个模型为 ARIMA 形式，即确定 p, d, q 及参数 $\phi_i's$ 和 $\psi_j's$ 的取值。

(b) 从哪方面看两个模型是不同的？

(c) 从哪方面看两个模型是相似的？(比较 MA 展式的 ψ 系数和 AR 展式的 π 系数)

(a). A: 记 $\phi_1 = 0.9, \phi_2 = 0.09$. 由于 $\phi_2 + \phi_1 < 1, |\phi_2| < 1$.

故 A 为一个平稳 AR(2) $\Rightarrow p=2, d=0, q=0, \phi_1=0.9, \phi_2=0.09$

$$B: \nabla Z_t = a_t - 0.1 a_{t-1}$$

故 B 为 IMA(1,1) $\Rightarrow p=0, d=1, q=1, \theta=0.1$

(b) A 是平稳模型. B 显然不符合平稳条件 (-阶差分后平稳)

(c) ψ : **A** 已知 $Z_t = 0.9 Z_{t-1} + 0.09 Z_{t-2} + a_t$ ①

由 Thm: $Z_t = \mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$ ②

$$\text{将 } ② \text{ 代入 } ①: \mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

$$= 0.9\mu + 0.9\psi_0 a_{t-1} + 0.9\psi_1 a_{t-2} + \dots + 0.09\mu + 0.09\psi_0 a_{t-2} + \dots + a_4$$

对比各系数:

$$\begin{cases} \psi_0 = 1 \\ \psi_1 = 0.9\psi_0 \\ \psi_2 = 0.9\psi_1 + 0.09\psi_0 \\ \dots \end{cases} \Rightarrow \begin{cases} \psi_0 = 1 \\ \psi_1 = 0.9 \\ \psi_k = 0.9^k + 0.09^{k-1}, k \geq 2 \end{cases}$$

B 已知 $Z_t = Z_{t-1} + a_t - 0.1 a_{t-1}$. 同理将 ② 代入

$$\mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

$$= a_t - 0.1 a_{t-1} + \psi_0 a_{t-1} + \psi_1 a_{t-2} + \dots$$

$$\Rightarrow \begin{cases} \psi_0 = 1 \\ \psi_1 = \psi_0 - 0.1 \\ \psi_2 = \psi_1 \\ \dots \end{cases} \Rightarrow \begin{cases} \psi_0 = 1 \\ \psi_k = 0.9, k \geq 2. \end{cases}$$

π :

A: 对于 AR(2): $\pi_1 = 0.9, \pi_2 = 0.09, \pi_k = 0 (k \geq 3)$

B: $\nabla Z_t = a_t - 0.1 a_{t-1}$

$$\Rightarrow a_t = \nabla Z_t + 0.1 a_{t-1} = Z_t + 0.1 Z_{t-1} + 0.1^2 Z_{t-2} + \dots$$

$$\Rightarrow a_t = Z_t + 0.9 Z_{t-1} + 0.09 Z_{t-2} + \dots$$

$$\Rightarrow \pi_0 = 1, \pi_1 = 0.9, \pi_2 = 0.09, \dots$$

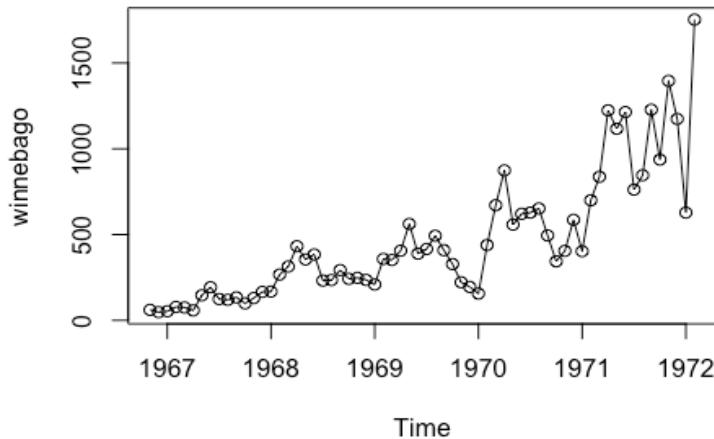
二者取值较为相近。

第三题

数据文件 `winnebago` 包含了 1966 年 11 月至 1972 年 2 月 Winnebago 公司休闲车的每月销售量。

(a) 画出这些数据的时序图并进行阐释;

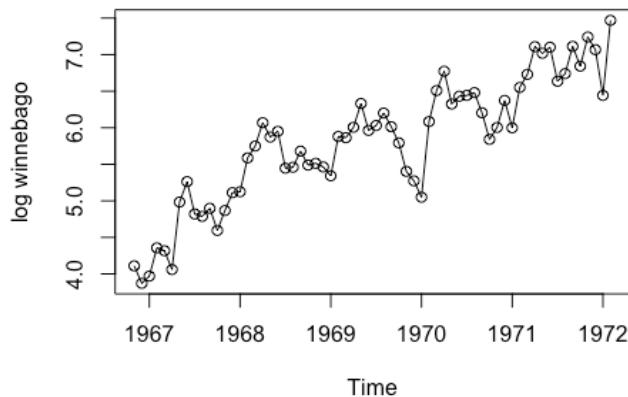
```
# install.packages('TSA')
data("winnebago")
plot(winnebago, type = "o")
```



从时序图上来看，销售量是不平稳的，在 1970 年左右出现逐步提升。随着时间的退役，销售量的均值在逐渐提高，其方差也逐步提高。

(b) 对月度销售量求自然对数，画出对数变换后的时序图，并描述对数变换对这个序列的影响；

```
log_winnebago <- log(winnebago)
plot(log_winnebago, type = "o", ylab = "log winnebago")
```



由 \log 变换后的时序图可以看出，对数变换主要解决了方差不平稳问题，使得整个序列的方差波动近似相同。但对数变换没有解决均值不平稳的问题，整体上还是呈现着向上增长的趋势。

(c) 计算相对变化率 $(Z_t - Z_{t-1})/Z_{t-1}$ 和对数差分 $\nabla \log(Z_t) = \log(Z_t) - \log(Z_{t-1})$, 并进行时序图比较。当数值较小和较大时，两者的对比结果如何？

```

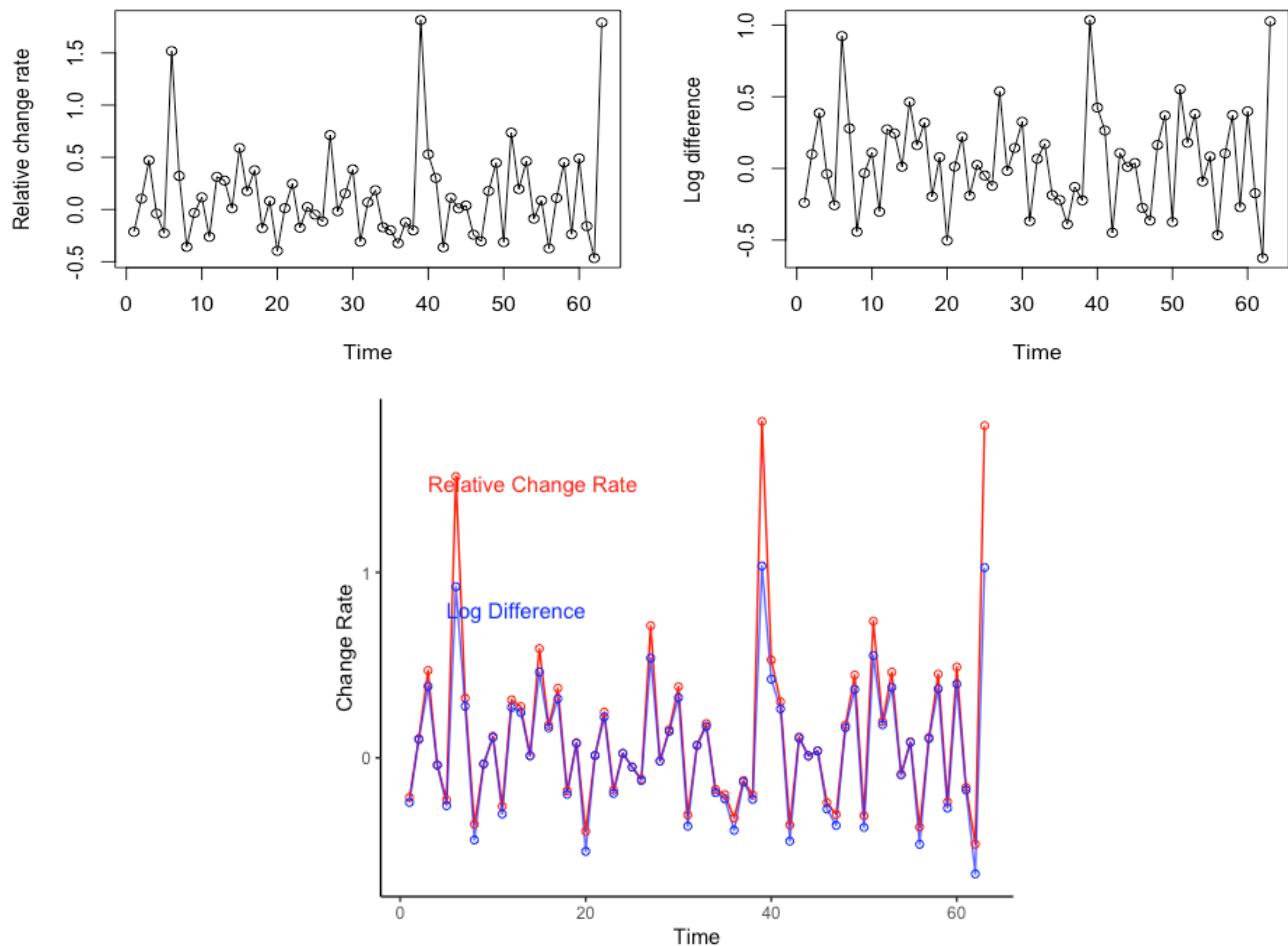
rcr <- 0
logdiff <- diff(log(winnebago))

for (t in 1:63) {
  rcr[t] <- winnebago[t + 1]/winnebago[t] - 1
}

plot(rcr, x = c(1:63), type = "o", ylab = "Relative change rate", xlab = "Time")
plot(logdiff, ylab = "Log difference", x = c(1:63), type = "o", xlab = "Time")

time <- c(1:63)
dat <- as.data.frame(cbind(time, rcr, logdiff))
gph <- ggplot(dat) + geom_line(mapping = aes(y = rcr, x = time), color = "red") +
  geom_line(mapping = aes(y = logdiff, x = time), color = "blue", alpha = 0.7) +
  geom_point(mapping = aes(y = rcr, x = time), color = "red", shape = 1) + geom_p
oint(mapping = aes(y = logdiff,
  x = time), color = "blue", shape = 1) + labs(y = "Change Rate", x = "Time") +
  theme(legend.position = "top") + theme_classic()
gph + annotate("text", label = "Log Difference", color = "blue", x = 12.5, y = 0.8)
+
  annotate("text", label = "Relative Change Rate", color = "red", x = 14.3, y = 1.
48)

```



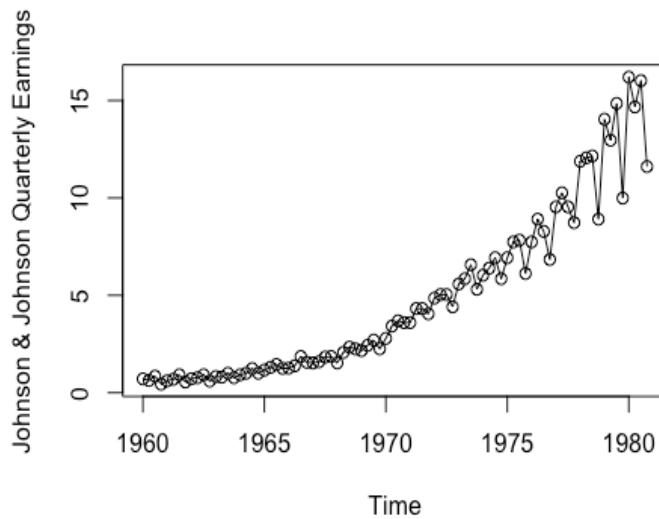
通过上述时间序列图可见，在当变化率较小时，二者较为接近；相反，当变化率较大时，二者差异更为显著，此时通常相对变化率的测定值会略大于对数差分的结果。

第四题

数据集 JJ 包含了强生公司每股收益的季度数据，时间跨度从 1960 年到 1980 年。

(a) 画出数据的时序图并进行阐释；

```
data("JJ")
plot(JJ, type = "o", ylab = "Johnson & Johnson Quarterly Earnings")
```

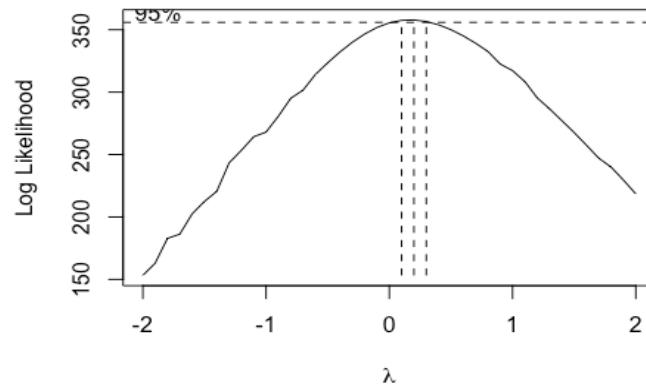


从时间序列图中可以看出，整体数据呈现逐步上升的趋势，并且随着年份的增加方差的波动率也逐步增加。是均值不平稳、方差不平稳的。

(b) 通过使用 R 中的 BoxCox.ar 函数作图，确定对该数据进行幂变换的最优 λ 值；

```
boxcox <- BoxCox.ar(JJ)
```

```
lambda_opt <- boxcox$lambda[which.max(boxcox$loglike)]
```



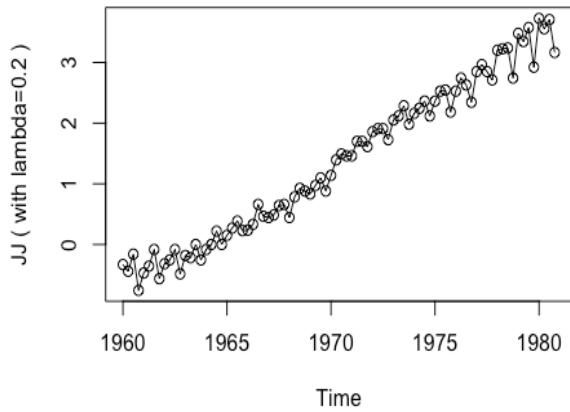
从上方输出可以看出，最优（使得对数似然函数取 max） $lambda = 0.2$. 其对应的 box-cox 变换为：

$$y(0.2) = 5(y^{0.2} - 1)$$

(c) 画出 (b) 中最优变换后的时序图，并判断变换后的序列是否平稳；

```
boxcox_JJ <- 5 * (JJ^0.2 - 1)
```

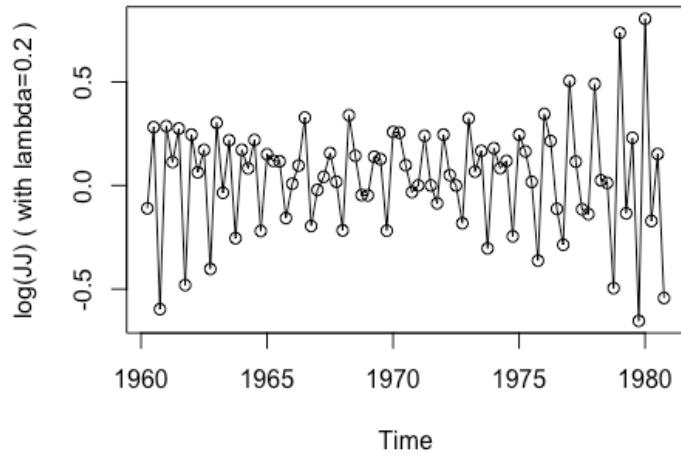
```
plot(boxcox_JJ, type = "o", ylab = "JJ ( with lambda=0.2 )")
```



由时序图可见，该幂变换整体上减少了方差非平稳，但依旧具有较强的均值非平稳。

(d) 对变换后的序列进行差分并画出时序图，由图判断对数差分后的序列是否平稳.

```
boxcox_JJ <- 5 * (JJ^0.2 - 1)
diff_boxcox_JJ <- diff(boxcox_JJ)
plot(diff_boxcox_JJ, type = "o", ylab = "log(JJ) ( with lambda=0.2 )", xlab = "Time")
```



由时序图可见，在经过幂变换、差分变换后，时间序列是基本均值平稳的，但方差仍有一定的“中间小、两头大”的波动趋势。

第五题

序列 $\{Z_t\}$ 及其一阶差分序列的样本 ACF 如表 1 所示，此处样本数量为 100. 基于这些信息，我们可以为该序列识别什么 ARIMA 模型呢？

欲通过ACF识别 $MA(q)$.

观察 Z_t 的 ACF，可发现其不具有截尾特征，因此对 Z_t 的 ACF 求解 Bartlett's 近似
由于进行了阶差分，故当前序列长度为 99.

$$lag_1: \pm \frac{1.96}{\sqrt{99}} = \pm 0.1970$$

$$lag_2: \pm \frac{1.96}{\sqrt{99}} \sqrt{1+2\times(-.49)^2} = \pm 0.2291$$

$$lag_3: \pm \frac{1.96}{\sqrt{99}} \sqrt{1+2\times(-.42)^2+2\times(.18)^2} = \pm 0.2345$$

$$lag_4: \pm \frac{1.96}{\sqrt{99}} \sqrt{1+\dots+2\times(-0.02)^2} = \pm 0.2346$$

$$lag_5: \pm \frac{1.96}{\sqrt{99}} \sqrt{1+\dots+2\times(0.07)^2} = \pm 0.2354$$

$$lag_6: \pm \frac{1.96}{\sqrt{99}} \sqrt{1+\dots+2\times(-0.10)^2} = \pm 0.2370$$

通过与 ACF 比较，可知 当且仅当 $lag=1$ 时 ACF 落于区间以外
故可以考虑识别为 $IMA(1,1)$

第六题

序列 $\{Z_t\}$ 的样本偏自相关系数 (PACF) 如表 2 所示，序列长度为 64. 基于这些信息，我们可以为该序列识别什么 ARIMA 模型呢？

$n=64$ ，由 PACF 的性质

$$\Rightarrow \pm 1.96 / \sqrt{64} = \pm 0.245$$

其中 $lag=1$, $lag=2$ 时 PACF 在区间外，其余 lag 时皆在其内
故考虑 $AR(2)$ 模型。

第七题

假设 $\{X_t\}$ 服从平稳 $AR(1)$ 序列 $X_t = \phi X_{t-1} + a_t$ ，但是我们只能观察到 $Z_t = X_t + N_t$ ，其中 $\{N_t\}$ 是独立于 $\{X_t\}$ 的白噪声测量误差， $\{X_t\}$ 和 $\{N_t\}$ 的方差分别为 σ_X^2 和 σ_N^2 .

(a) 求 $\{Z_t\}$ 的自相关系数函数；

(b) 我们可以为 $\{Z_t\}$ 序列识别什么 ARIMA 模型呢？

(a) 由于 $\hat{z}_t = x_t + n_t$ 且 x_t 平稳

$$\Rightarrow \text{corr}(\hat{z}_t, \hat{z}_{t+k}) = \frac{\rho_k}{1 + \sigma_n^2 / \sigma_x^2}$$

其中 $\rho_k = \text{corr}(x_t, x_{t+k})$

$$\text{而 } x_t \sim AR(1) \Rightarrow \rho_k = \phi^k$$

$$\Rightarrow \text{corr}(\hat{z}_t, \hat{z}_{t+k}) = \frac{\phi^k}{1 + \sigma_n^2 / \sigma_x^2}$$

$$(b) \text{ 由 } \text{corr}(\hat{z}_t, \hat{z}_{t+k}) = \frac{\phi^k}{1 + \sigma_n^2 / \sigma_x^2}$$

可知，随 k 增加， $\{\hat{z}_t\}$ ACF 指数衰减

故可识别为 AR(1) 或 ARMA(1,1)

第八题

模拟 AR(1) 时间序列 $Z_t = 0.7Z_{t-1} + a_t$, 序列长度为 $n = 48$.

(a) 计算该模型在 1 阶和 5 阶滞后处的理论自相关系数;

对于 AR(1), $\rho_k = 0.7^k$

故 $\rho_1 = 0.7, \rho_5 = 0.7^5 \approx 0.1681$

(b) 计算该模型在 1 阶和 5 阶滞后处的样本自相关系数，并将其与理论值进行比较，通过计算

$\text{Var}(r_1) = \frac{1-\phi^2}{n}$ 和 $\text{Var}(r_5) = \frac{1+\phi^2}{n(1-\phi^2)}$ ，量化说明这个比较；

```
set.seed(2023)
dat <- arima.sim(n = 48, list(ar = 0.7))
dat_acf <- acf(dat, plot = FALSE, lag.max = 5)
dat_acf$acf[c(1, 5)]
## [1] 0.6405384206 0.0002979744
```

由(a)问的结果可知，理论上， $\rho_1 = 0.7, \rho_5 = 0.1681$ ，在本次模拟中， $r_1 = 0.6405, r_2 = 0.0003$ ，可以计算其相对误差：

$$e_1 = \frac{\rho_1 - r_1}{\rho_1} = 8.5\%,$$

$$e_2 = \frac{\rho_2 - r_2}{\rho_2} = 99.82\%$$

由此可见，高阶滞后的 ACF 误差更大。代入 $\text{Var}(r_k)$ ，可知

$$Var(r_1) = \frac{1 - 0.7^2}{48} = 0.0106$$

$$Var(r_2) = \frac{1 + 0.7^2}{48 \times (1 - 0.7^2)} = 0.0609$$

故由此可见，高阶滞后的渐进方差更大，波动范围更广。

- (c) 将生成 AR(1) 序列的模拟重复 100 次，并基于每次模拟数据计算 r_1 和 r_5 ，构建 r_1 和 r_5 的抽样分布。比较 r_1 和 r_5 的抽样分布的方差与 $Var(r_1)$ 和 $Var(r_5)$ 对应的渐近方差之间是否接近？

```
set.seed(2023)
acf_1 <- 0
acf_5 <- 0
for (i in 1:100) {
  dat <- arima.sim(n = 48, list(ar = 0.7))
  dat_acf <- acf(dat, plot = FALSE, lag.max = 5)
  acf_1[i] <- dat_acf$acf[1]
  acf_5[i] <- dat_acf$acf[5]
}
acf_1
acf_5

hist(acf_1)
hist(acf_5)

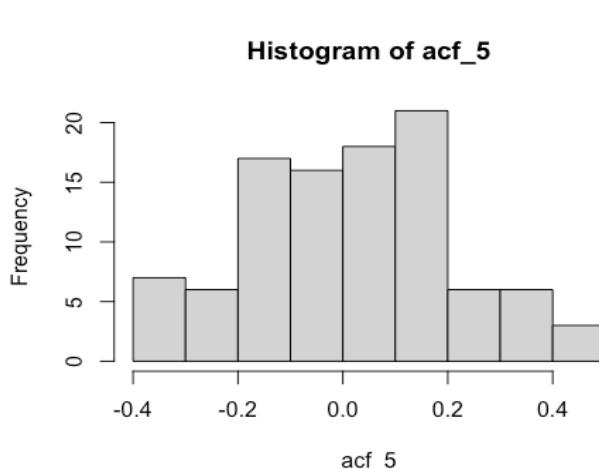
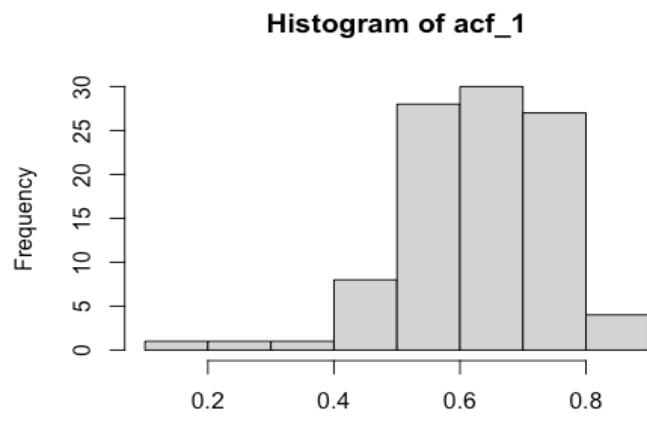
var(acf_1)
var(acf_5)

## [1] 0.6405384 0.5822334 0.6917957 0.5560534 0.6054353 0.7230902 0.4984364
## [8] 0.7125414 0.6694372 0.6628620 0.6419592 0.5187338 0.7357552 0.6839814
## [15] 0.3272536 0.5809915 0.1359561 0.6029947 0.5410253 0.5663550 0.5107038
## [22] 0.5938730 0.6366891 0.7408684 0.6931088 0.7722002 0.5978930 0.5859168
## [29] 0.7180858 0.5754840 0.5872484 0.4547549 0.7846753 0.5764802 0.7936387
## [36] 0.4746278 0.6567334 0.7751190 0.7815672 0.5585815 0.5948824 0.6059773
## [43] 0.6273487 0.6415737 0.4744736 0.6979649 0.7906767 0.5680047 0.5218526
## [50] 0.6675357 0.6049578 0.5520266 0.2199310 0.7493089 0.7281799 0.7261055
## [57] 0.5525481 0.7804387 0.7463051 0.5586934 0.5965218 0.5456493 0.7592010
## [64] 0.6951501 0.6234313 0.8079587 0.4959498 0.7953949 0.4300693 0.6244512
## [71] 0.7074378 0.6904055 0.6804017 0.6213138 0.6787390 0.5672742 0.6721237
## [78] 0.5769656 0.7330926 0.6821166 0.5713160 0.8159862 0.7587514 0.6483683
## [85] 0.6745572 0.6432706 0.8094063 0.7531780 0.7193253 0.5258021 0.5041673
## [92] 0.4959994 0.7379830 0.7474838 0.8004078 0.4681780 0.5486559 0.6881846
## [99] 0.7092223 0.7702439
## [1] 0.0002979744 -0.1592237768 0.1705407133 0.1052677849 -0.0731833182
## [6] 0.3412341154 -0.0751010370 0.1693957648 0.3003877205 -0.1952589442
## [11] -0.0614661919 0.0321085073 -0.1387246238 0.3758816938 -0.1599777130
## [16] 0.1083805805 -0.2196303849 0.1935514999 -0.1581174729 0.0905313557
## [21] -0.2402878322 -0.1945702673 -0.3065275980 -0.1324156987 0.1026100489
```

```

## [26] 0.1499389952 -0.0885065945 -0.1653386952 0.0504322251 0.2103378828
## [31] -0.2034025975 0.0417903459 0.1374553212 -0.1202324286 0.3703054441
## [36] 0.1565225870 0.0678526057 0.4141774324 0.4411559335 0.0455990445
## [41] -0.0990240514 -0.0347098875 -0.1235543491 0.0843281945 -0.0845245318
## [46] 0.0242366613 0.4416614741 -0.3811999321 -0.0749911489 -0.1152471104
## [51] -0.3834229825 -0.1216972273 -0.2737646450 0.0320459461 0.2547391428
## [56] 0.0901402533 0.1038911307 0.1155382598 0.0678681083 -0.0164277588
## [61] 0.2285521115 -0.3366817885 0.1882509194 0.1908668440 -0.0183137599
## [66] 0.3918629761 -0.2740419884 0.2073042543 -0.0972605955 -0.0978791327
## [71] 0.1903692326 -0.1049706035 0.0737487114 0.1286953170 -0.3542269351
## [76] -0.0766982937 -0.1290328971 0.1001838282 0.0702683238 0.2476839050
## [81] -0.1382250181 0.2410783787 0.3136640952 -0.2047592662 0.1433054546
## [86] -0.3498761817 0.0843153676 0.1283530524 0.1021837255 -0.1180049911
## [91] -0.3344423119 -0.1012416208 0.1451400456 -0.0523618714 0.0896745194
## [96] -0.0789727979 -0.0286509783 0.0827306539 0.1325066492 0.0423405038
## [1] 0.01449639
## [1] 0.03777268

```



同样计算相对误差：

$$e_{var_1} = 36.76\%$$

$$e_{var_2} = 37.97\%$$

(d) 相比 $n = 48$ 时, 当 $n = 100$ 时抽样分布的方差与渐近方差之间的接近程度如何?

```
set.seed(2023)
acf_1 <- 0
acf_5 <- 0
for (i in 1:100) {
  dat <- arima.sim(n = 100, list(ar = 0.7))
  dat_acf <- acf(dat, plot = FALSE, lag.max = 5)
  acf_1[i] <- dat_acf$acf[1]
  acf_5[i] <- dat_acf$acf[5]
}

var(acf_1)
var(acf_5)

var1 <- (1 - 0.7^2)/100
var2 <- (1 + 0.7^2)/(1 - 0.7^2)/100
var1
var2

## [1] 0.00452962
## [1] 0.02027745
## [1] 0.0051
## [1] 0.02921569
```

计算 $n = 100$ 的相对误差:

$$e_1^* = 11.1839\%$$

$$e_2^* = 30.5939\%$$

若考虑其绝对误差。其中 1 阶滞后绝对误差为: 5.7×10^{-4} , 5 阶滞后绝对误差: 8.9×10^{-3} . 更可以突出, 当样本量增加, 估计的精度得到大幅提高。