AFFTC-PA-11244

Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests

A
F
F
T
C

Todd Remund
Dr. William Kitto

AIR FORCE FLIGHT TEST CENTER
EDWARDS AFB, CA

July 2011

Approved for public release A: distribution is unlimited.

AIR FORCE FLIGHT TEST CENTER
EDWARDS AIR FORCE BASE, CALIFORNIA
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
21-07-2011

2. REPORT TYPE
ITEA Conference

3. DATES COVERED *(From - To)*
20-07-2011 to 22-07-2011

4. TITLE AND SUBTITLE
Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests

5a. CONTRACT NUMBER
NA

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Todd Remund, Dr. William Kitto

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
812 TSS
307 E. Popson Ave
Edwards AFB, CA 93524

8. PERFORMING ORGANIZATION REPORT NUMBER
AFFTC-PA-11244

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
812 TSS
307 E. Popson Ave
Edwards AFB, CA 93524

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release A: distribution is unlimited.

13. SUPPLEMENTARY NOTES
CA: Air Force Flight Test Center Edwards AFB CA CC: 012100

14. ABSTRACT
Edwards AFB, as a matter of policy, requires statistical rigor be a part of test design and analysis. Statistically defensible methods are used to gain as much information as possible from each test. This requires:

- Statistically defensible methods be identified and applied to each test
- Setting up tests to maximize scope of inference, and
- Determining the power or each test to optimize sample size

This paper demonstrates how Monte Carlo techniques may be applied to aircraft test and evaluation to determine the power of the test and the associated sample size requirements. Traditional methods for determining the power of a test are based on distributional assumptions associated with data. These assumptions may not be appropriate; a distribution-free Monte Carlo technique for power assessment for tests with (possible) serially correlated data is presented. The technique is illustrated with an example from a target location error (TLE) test. Power of the test and appropriate sample sizes are derived using Monte Carlo simulation implemented in R.

15. SUBJECT TERMS
Power, statistics, resampling, Monte Carlo simulation, R, sample size, CEP, CE90, circular error.

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			412 TENG/EN (Tech Pubs)
Unclassified	Unclassified	Unclassified	None	21	**19b. TELEPHONE NUMBER** *(include area code)* 661-277-8615

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests

July 2011

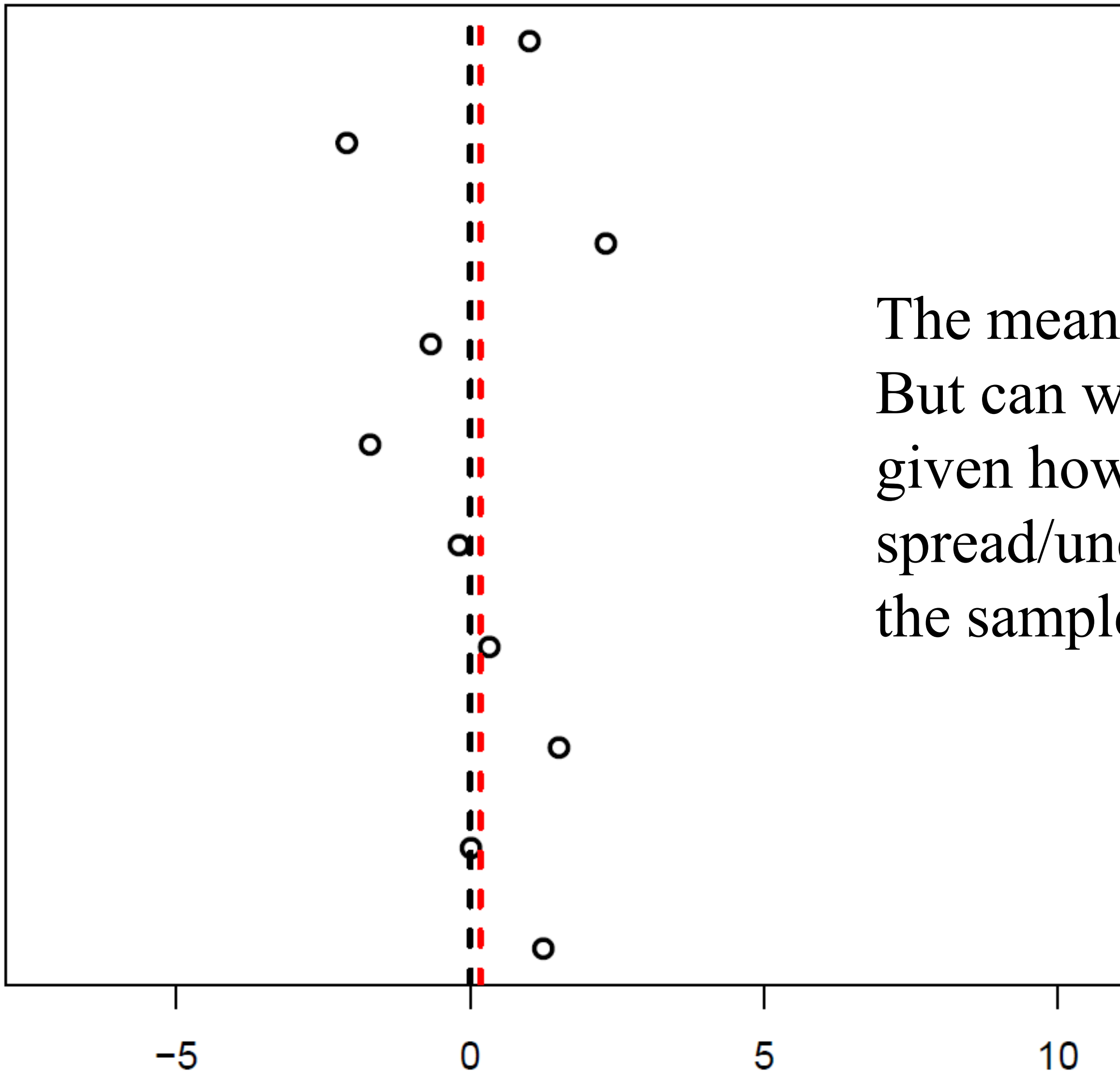
Todd Remund
& William Kitto
412 TW
661-277-6384

Power—what is it?

Truth	Decision	
	H_0: No Change	H_A: Change
H_0: No Change	Correct Negative	False Positive
H_A: Change	False Negative	Correct Positive

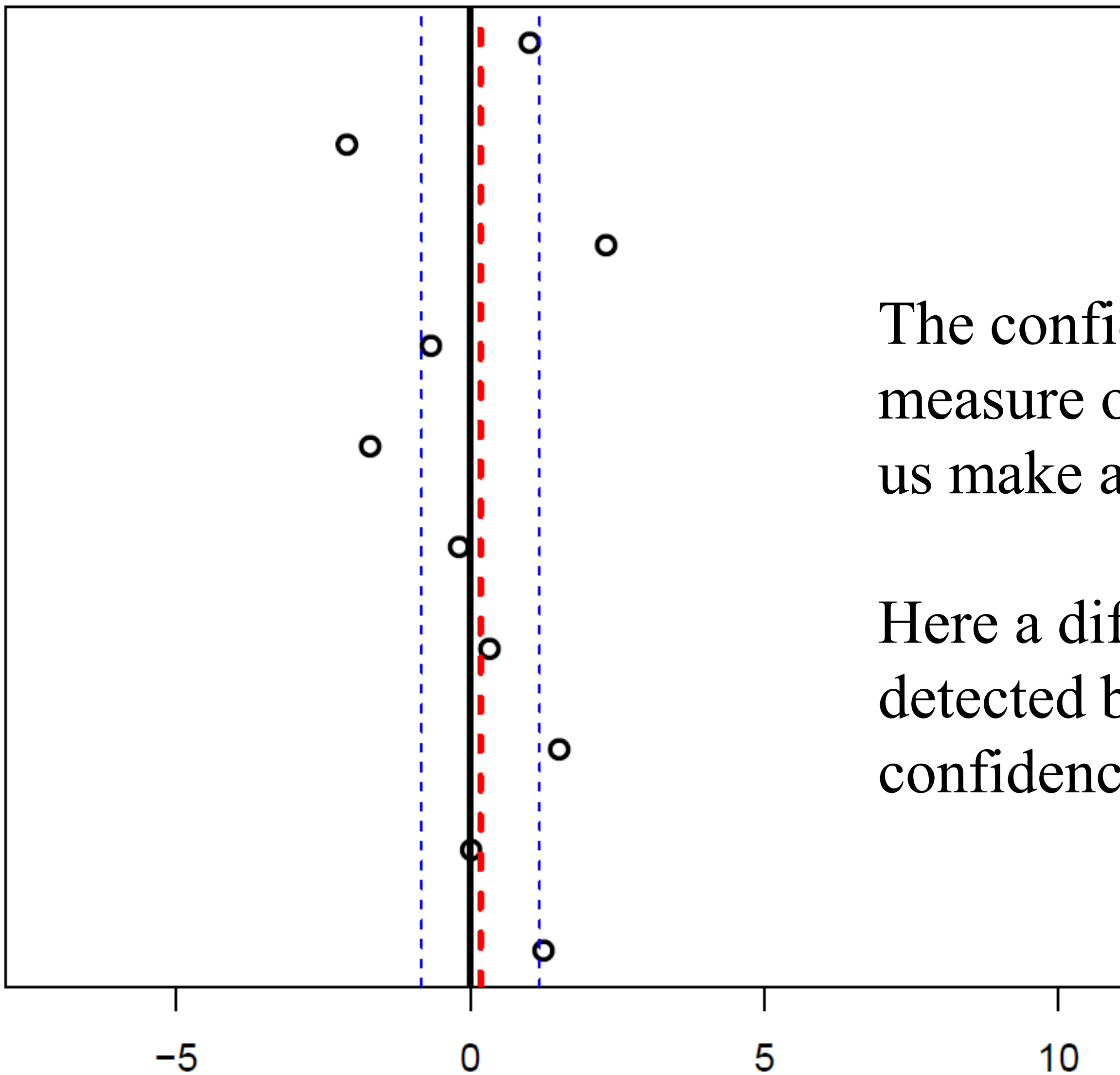
- Power = $\Pr(H_A \mid H_A$ is true$)$
  - Choose sample size, n, to get this
  - Also need to decide what you want to see…stay tuned…
- α = $\Pr(H_A \mid H_o$ is true$)$
  - Choose this number directly
  - Normally 0.05 or 0.1

Uncertainty and its vicissitudes



The mean is never equal to 0. But can we see the difference given how much spread/uncertainty there is in the sample?

Uncertainty and its vicissitudes



The confidence intervals put a measure on uncertainty to help us make a decision.

Here a difference is not detected because the confidence intervals cross zero.

Again…Power—what is it?

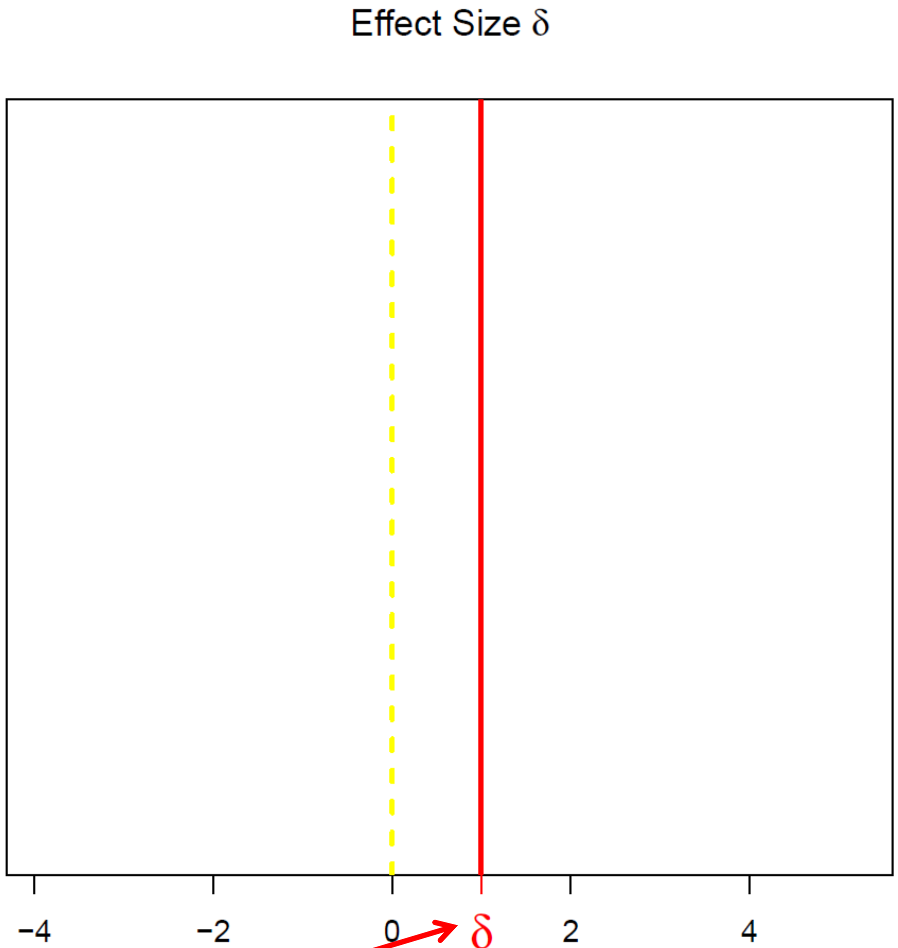
- Power is the proportion of times, in the long run, that our test (t-test, CI) identifies a difference, when it really exists.

- WRT the mean, we need to decide how big of a difference from zero do we care about.
  - This is the effect size, called δ.
- If we choose enough samples the CI shrinks, and it is easier to see a difference.
- But how large of a sample do we need to get to see the δ we want?

Power—via Monte Carlo

- For many applications, such as the one given, power calculations are closed form.
- For other difficult applications, the calculation doesn‟texist.
- Generate Alternate Population
  - With the desired effect
  - With the distribution characteristics needed
- Sample from it repeatedly
  - Each time analyzing the sample and record significance
- Compute the proportion of significant outcomes
  - This is power via Monte Carlo

Setting Up the „Minimal" Alternative (1-sample t-test)

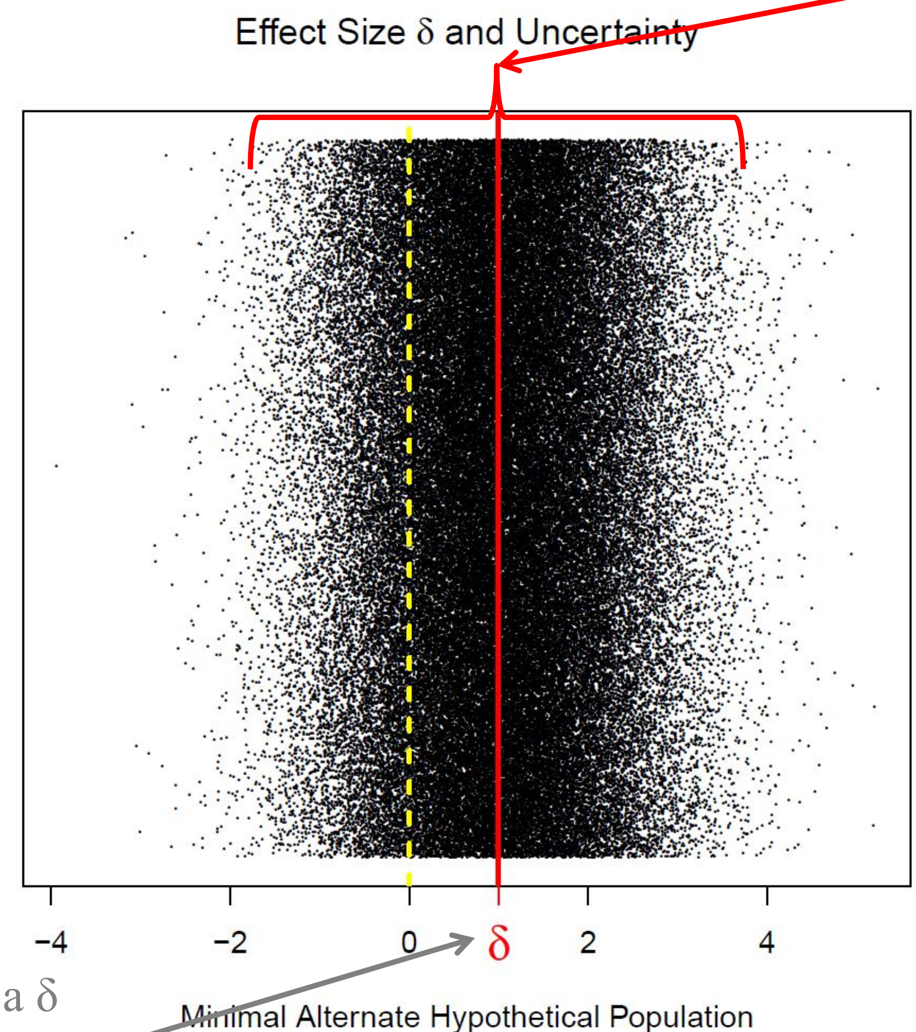


Couch the example in a one-sample t-test for simplicity's sake.

We want to see at least a δ effect size 80% of the time…

Setting Up the „Minimal" Alternative (1-sample t-test)

Couch the example in a one-sample t-test for simplicity's sake.



With this much uncertainty (characteristics of population).

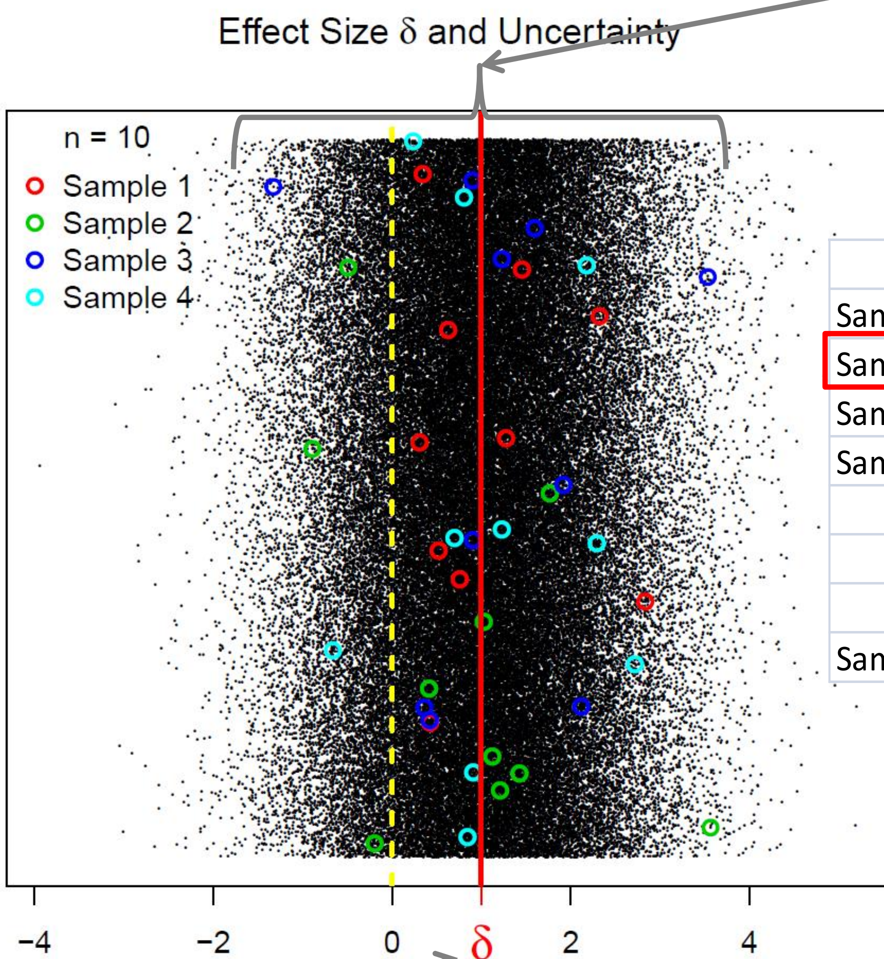
We want to see at least a δ effect size 80% of the time…

Using the Alternate World

Couch the example in a one-sample t-test for simplicity's sake.

Now repeatedly sample from the population and run t-tests each time.

Each sample will have a different estimate for mean and standard deviation.



With this much uncertainty (characteristics of population).

	95% CI	
Sample 1	0.45	1.72
Sample 2	-0.02	1.81
Sample 3	0.25	2.08
Sample 4	0.39	1.85
.	.	
.	.	
.	.	
Sample 1000	0.41	1.59

We want to see a δ effect size 80% of the time…

Power Estimate

- With n=10, st.dev=1, δ=1 for Normal distribution we get—
- Using the Monte Carlo method to calculate power for the one-sample t-test:
  - Power = 80.2%
  - This method has a little variation in the estimate because it is a simulation approach.
- Using conventional methods:
  - Power = 80.31%

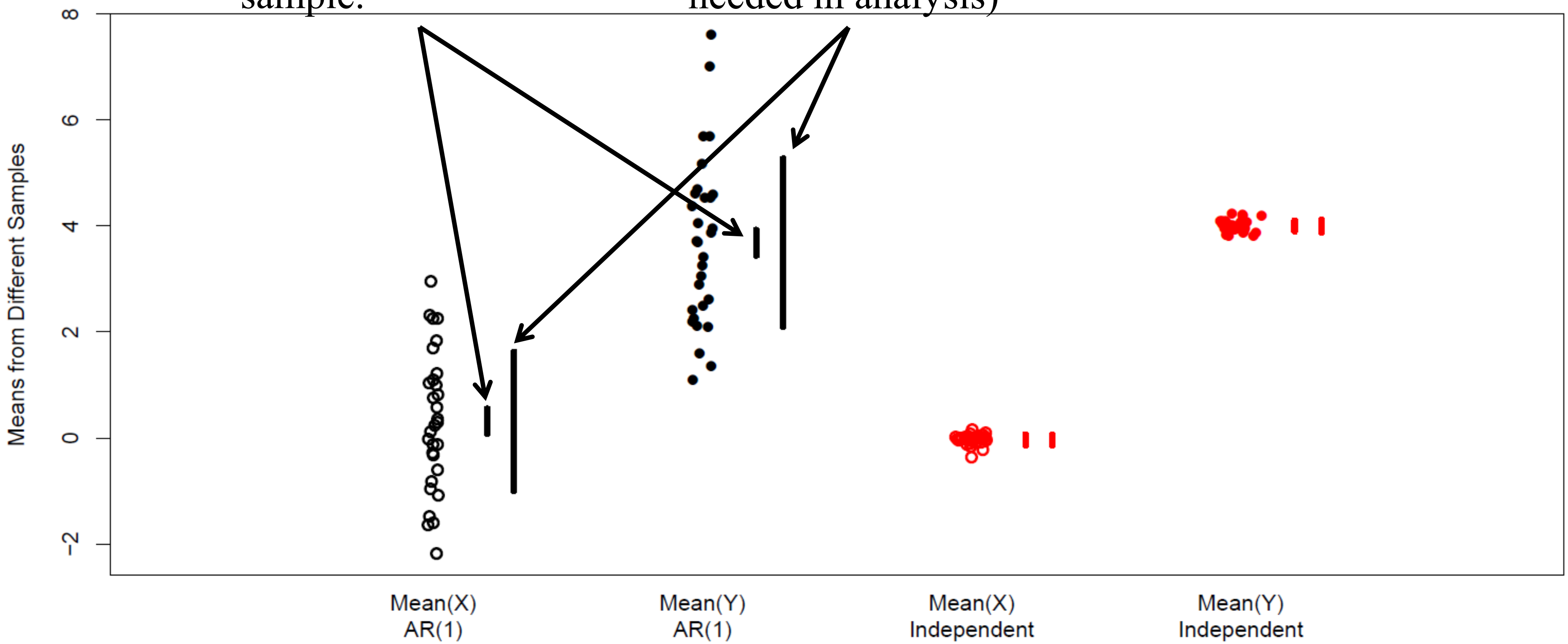
Click through PDF
“SerialMeans.pdf” File

SERIAL CORRELATION

2-Sample t-test w/ Serial Corr.

Incorrect SE of mean from an individual sample.

Actual SE of mean. (An adjustment is needed in analysis)



Need a different version of test.

- **The difference in a regular 2-sample t-test, and an adjusted test is,**
  - **Estimate the autocorrelation, r**
  - **Adjust the SE of the test statistic:**

$$SE(\bar{x}-\bar{y})_{adjusted} = SE(\bar{x}-\bar{y}) * \sqrt{\frac{1+r}{1-r}}$$

$$CI = \bar{x}-\bar{y} \pm z_{1-\alpha/2} SE(\bar{x}-\bar{y})_{adjusted}$$

- **What is the conventional method of computing power for this?**

How do we do it?

1. Create the *minimal alternate hypothetical population MAHP*
2. Take sample of size n from the MAHP
3. Test to see significance with chosen test, (here we"reusing the adjusted CI previous page).
4. Repeat steps 2 and 3 1000, 10000, or more times while recording how many are significant.
5. Find proportion that are significant out of number of repeated loops.

Method: Plug and Play

Question: What ***n*** do we use to detect δ with chosen power?

1. Generate MAHP

__M__inimal __A__lternate __H__ypothetical __P__opulation

Size 100,000 (or something real big)

2. Sample ***n*** values from MAHP

3. Test effect with method

Method 1

4. Record outcome from test (significant or not)

Repeat loop 1K or 10K times

$$Power = \frac{1}{1000}\sum_{i=1}^{1000} v_i$$

Get vector, ***v***, of 0/1, 1 for significant outcome

There may be multiple ways of testing effect.

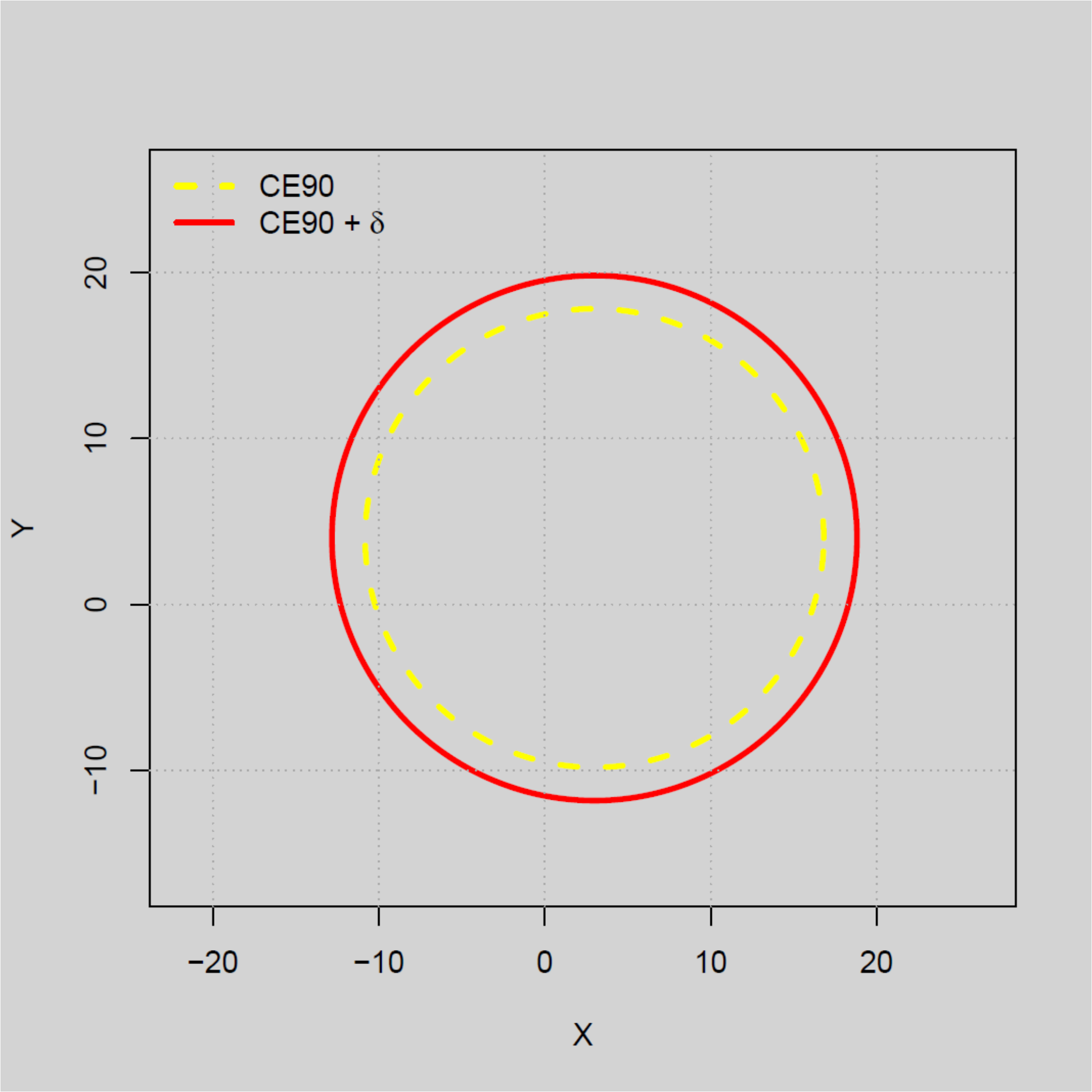
- Method 1
- Method 2
- Method 3

Which is most powerful for given sample size?

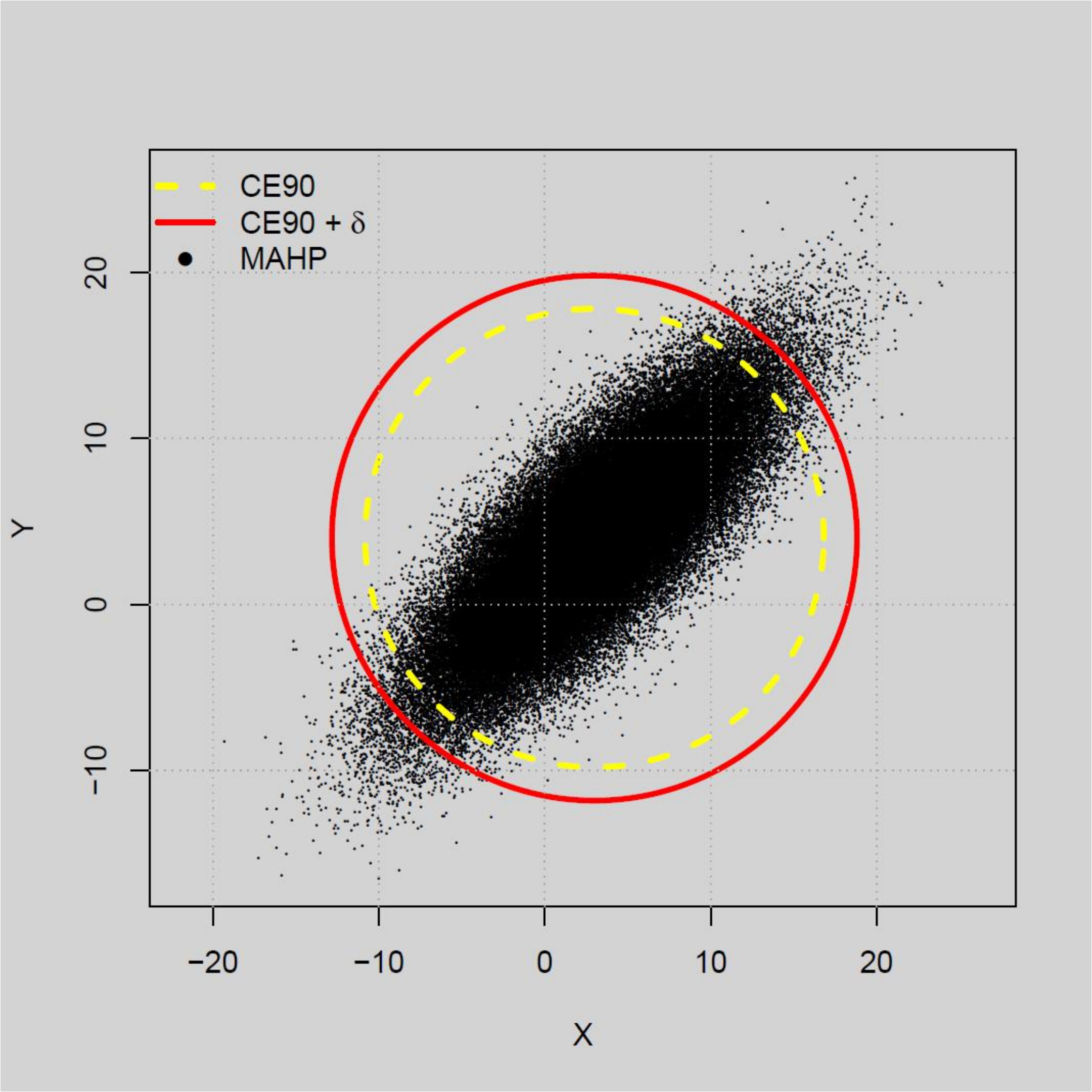
CE90 Power

- We want to know how many runs we need to prove CE90 is meeting spec for a targeting device.
  - How close to spec do you want to be before you are willing to concede that you are no different from spec?
  - We want to have proof of meeting spec if it is at least 2 feet beyond CE90.

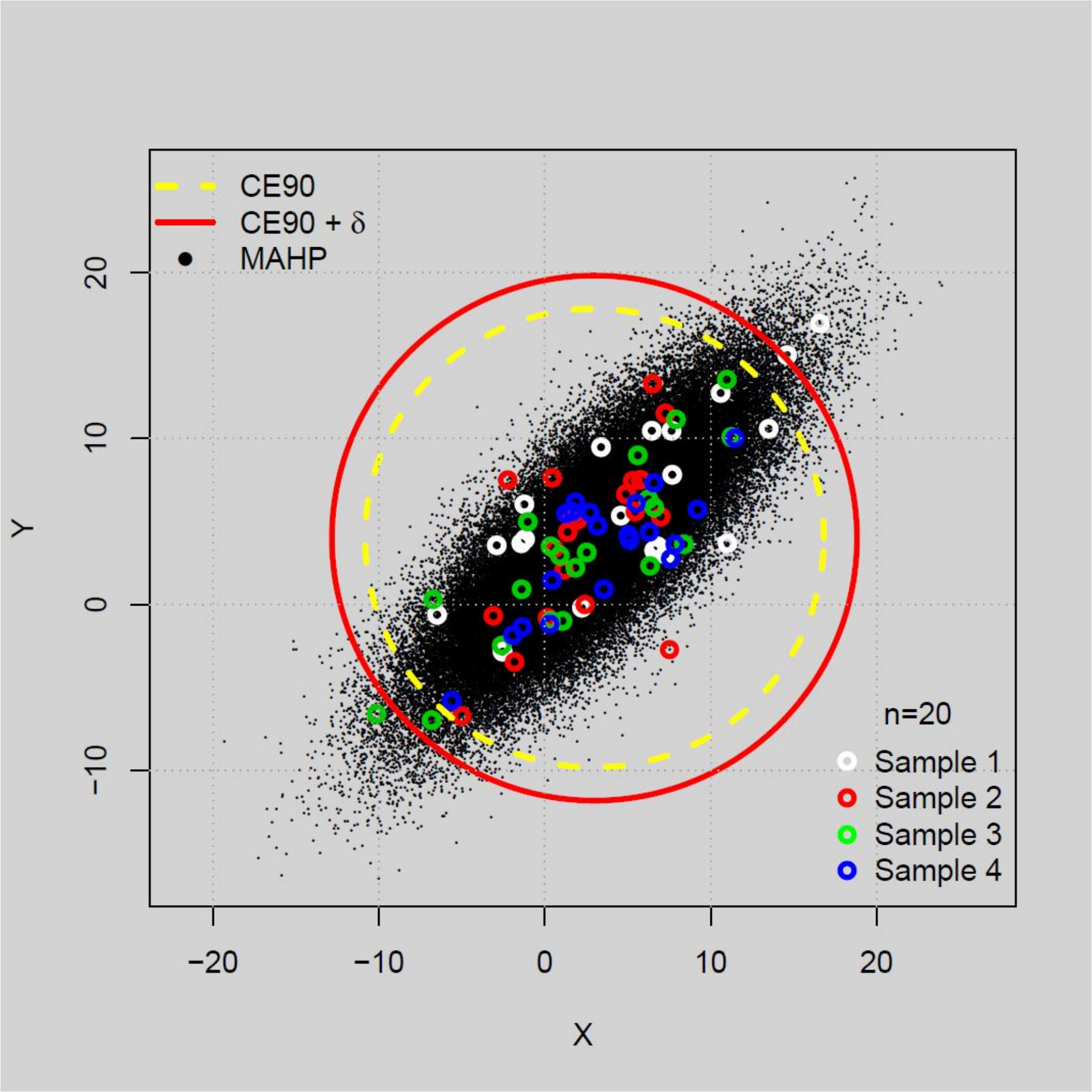
CE90 MAHP



CE90 MAHP



CE90 MAHP



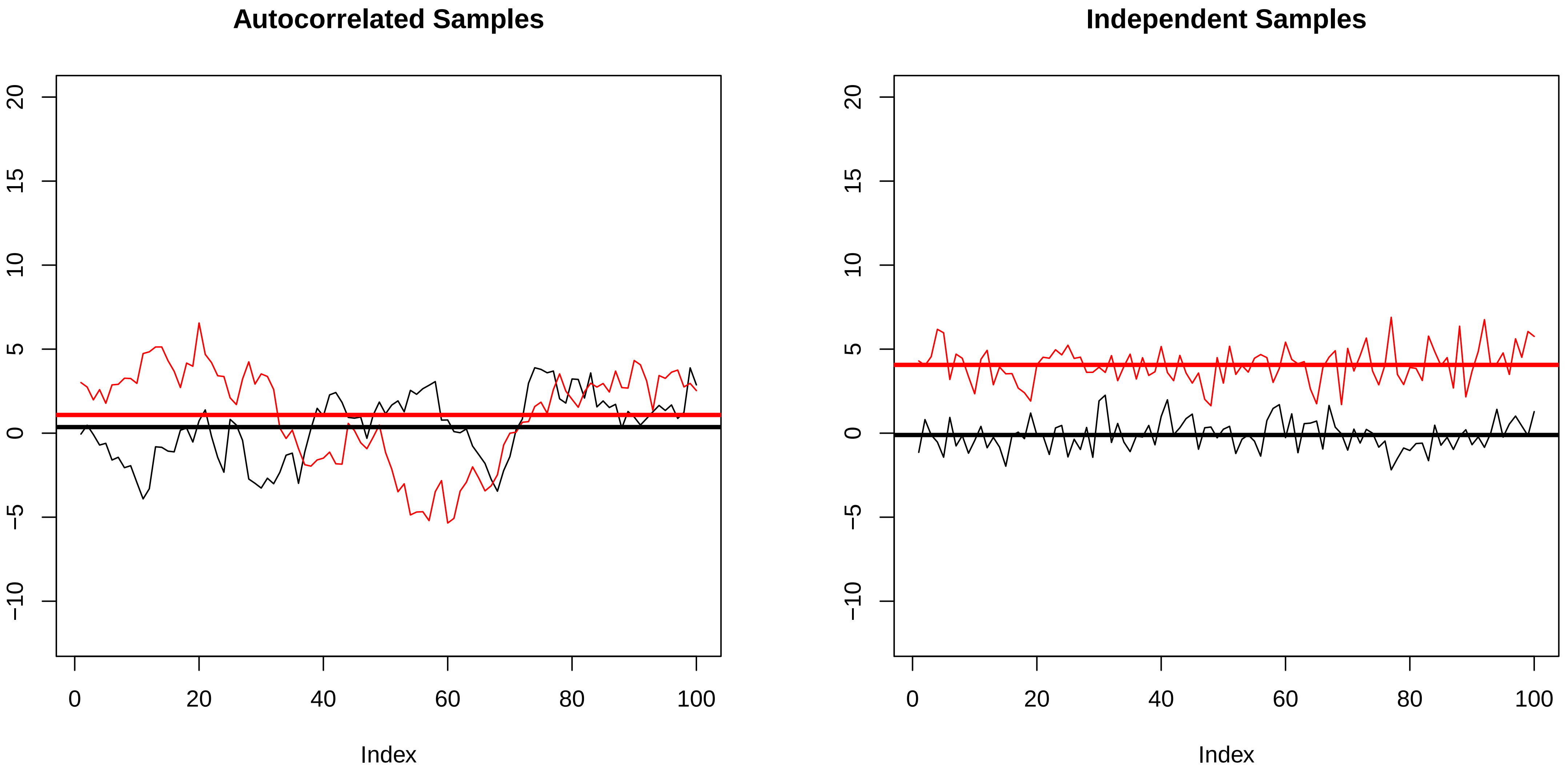
CE90
CE90 + δ
MAHP
n=20
Sample 1
Sample 2
Sample 3
Sample 4
Y
X
20
10
0
−10
−20
−10
0
10
20

CE90 Power

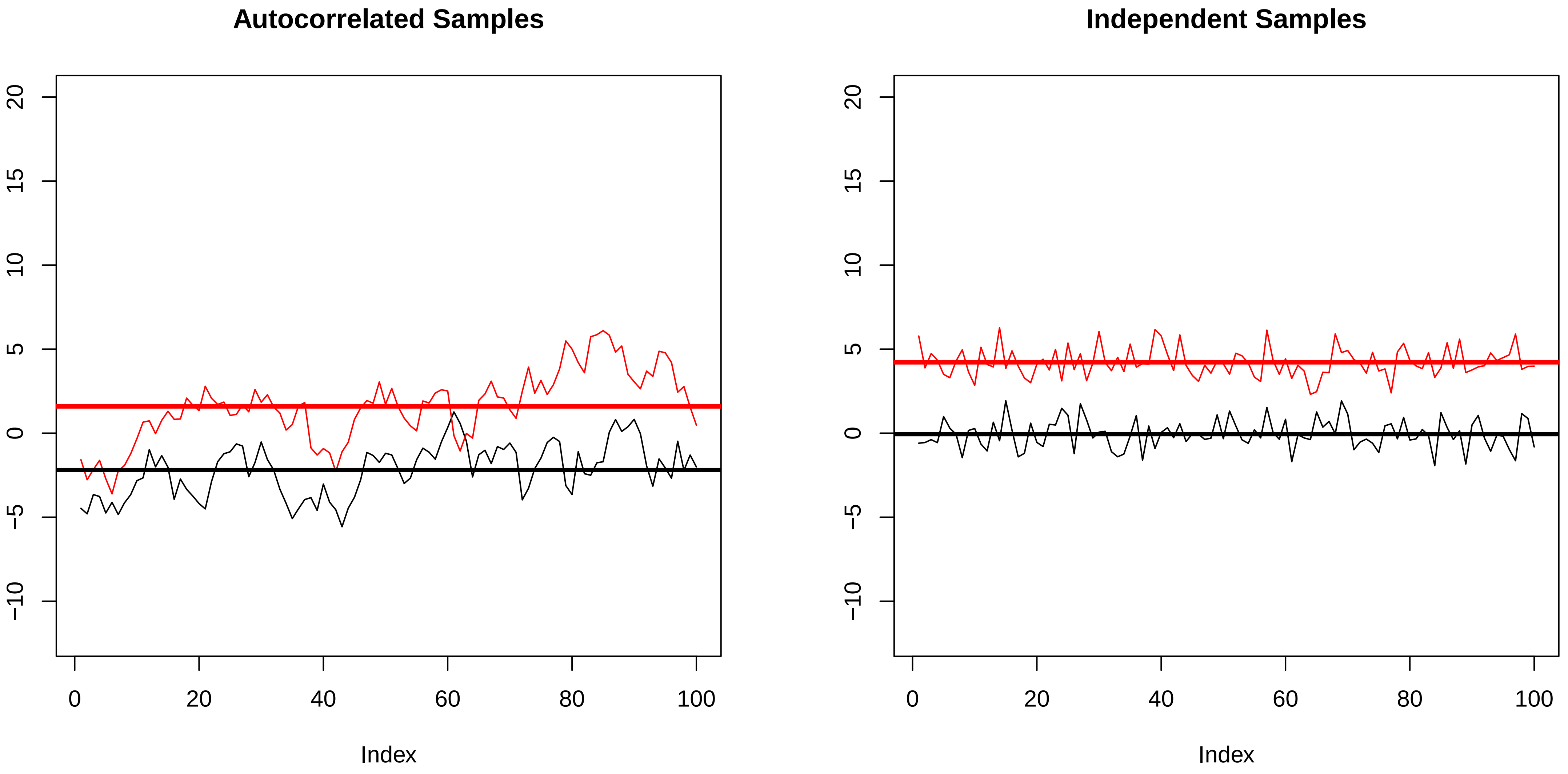
- With δ=2 and sample size of 60,
  - Power = 70%
- This power calculation is only for the specific situations similar to the MAHP. Any different pattern in CE will require a separate power analysis.

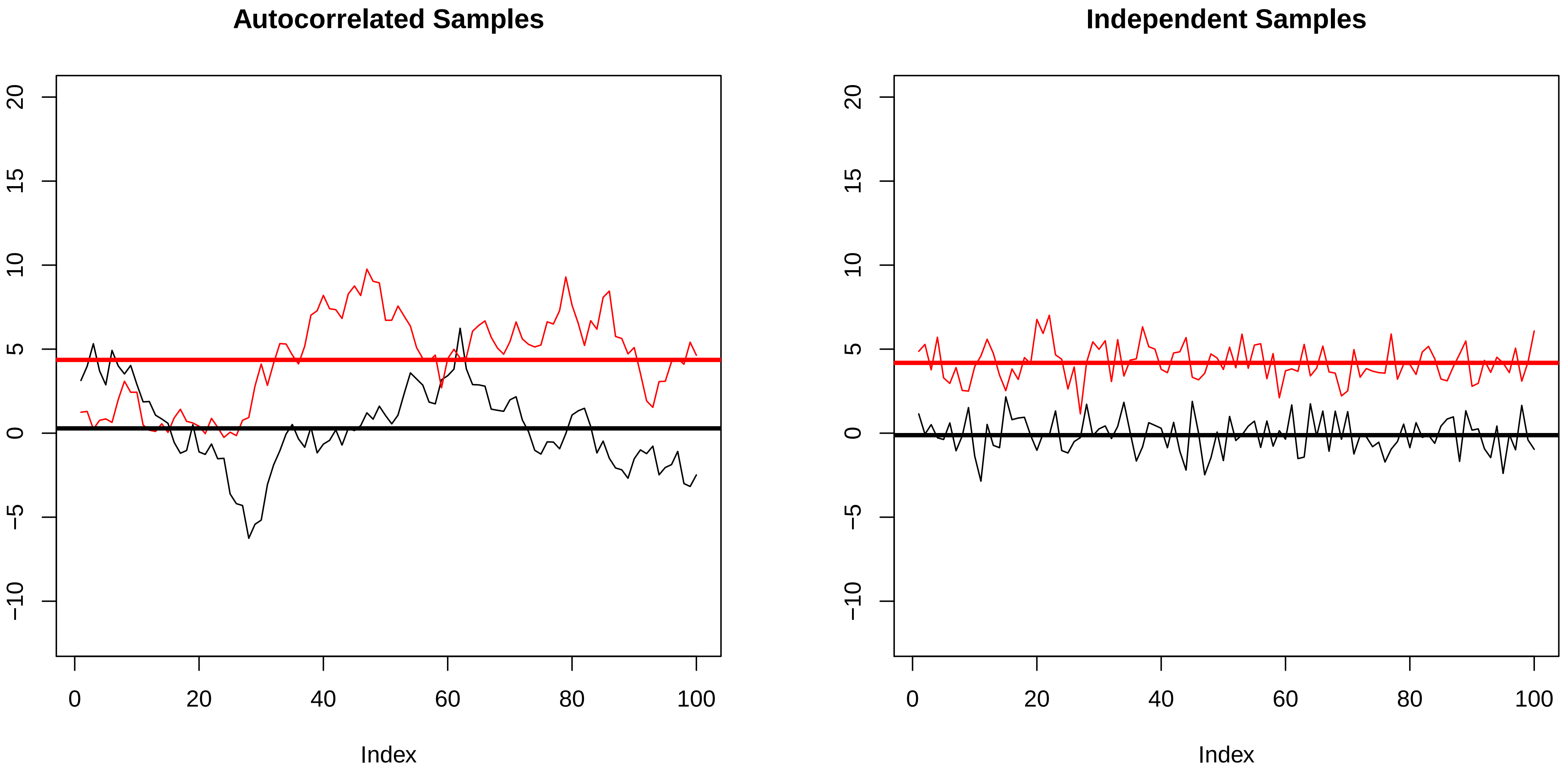
Summary

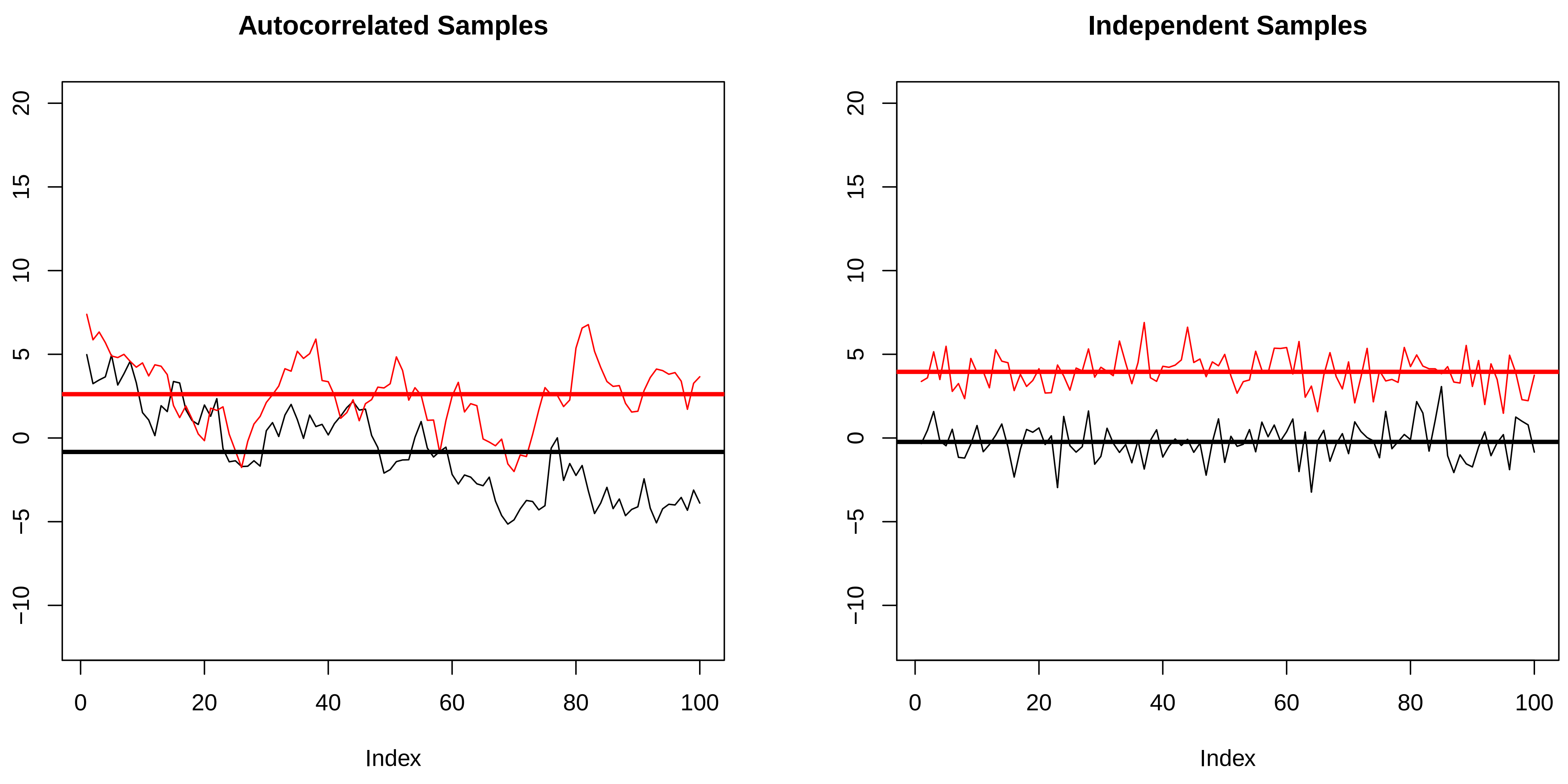
- Monte Carlo power estimation is versatile and can handle most situations.
- It is difficult sometimes to create the MAHP.
- A statistician is likely needed to aid in the process.



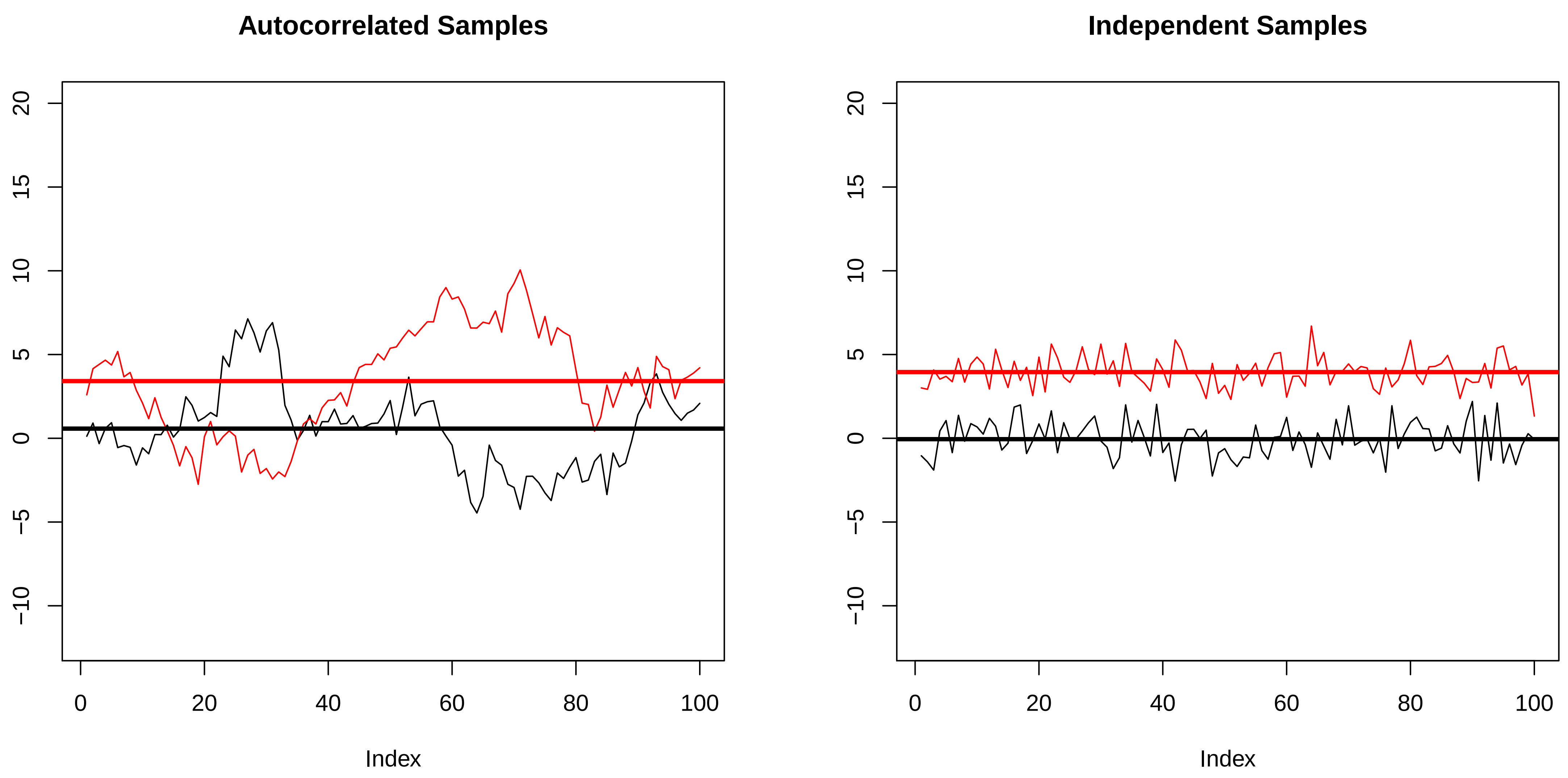
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

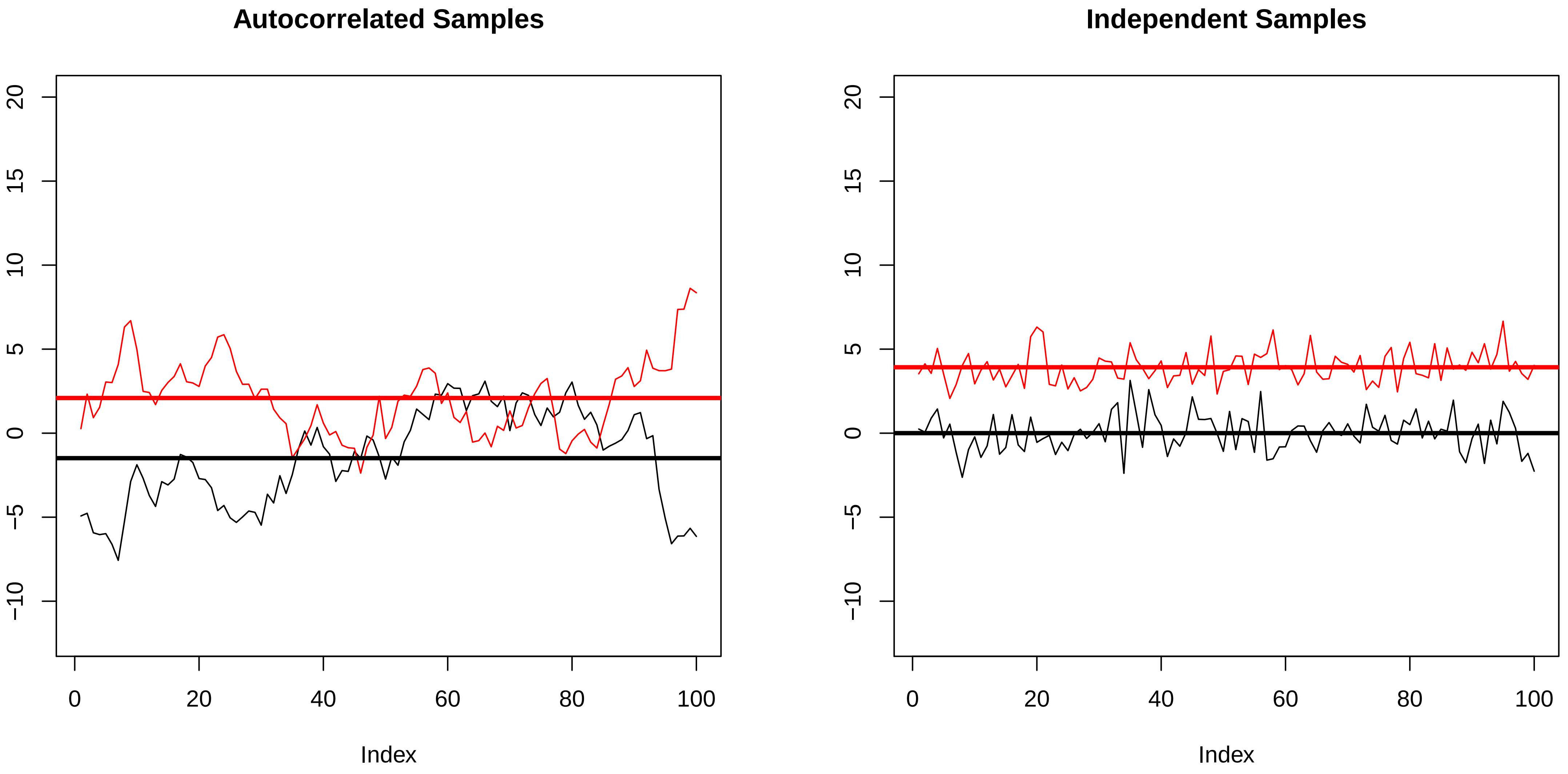

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

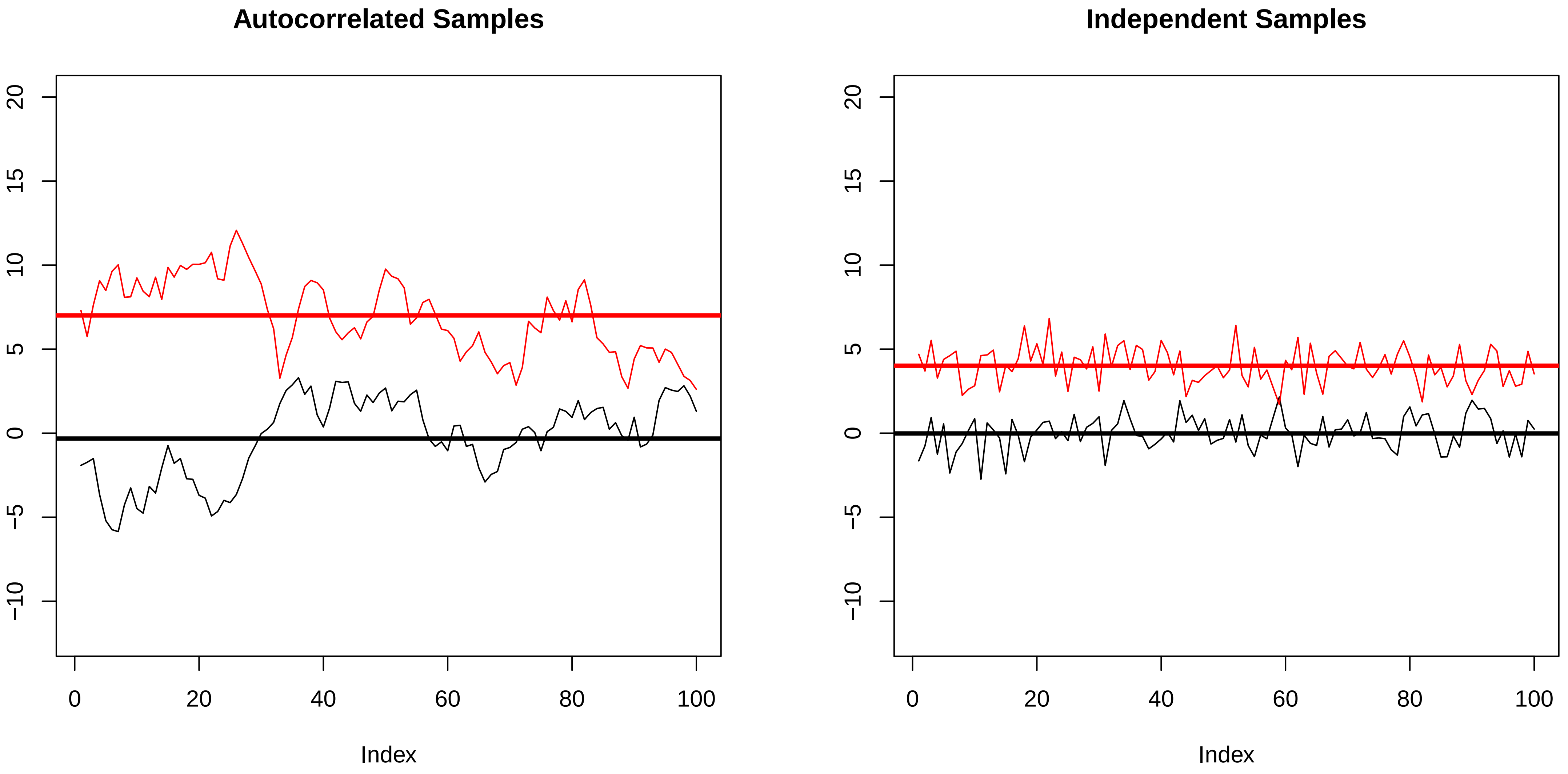


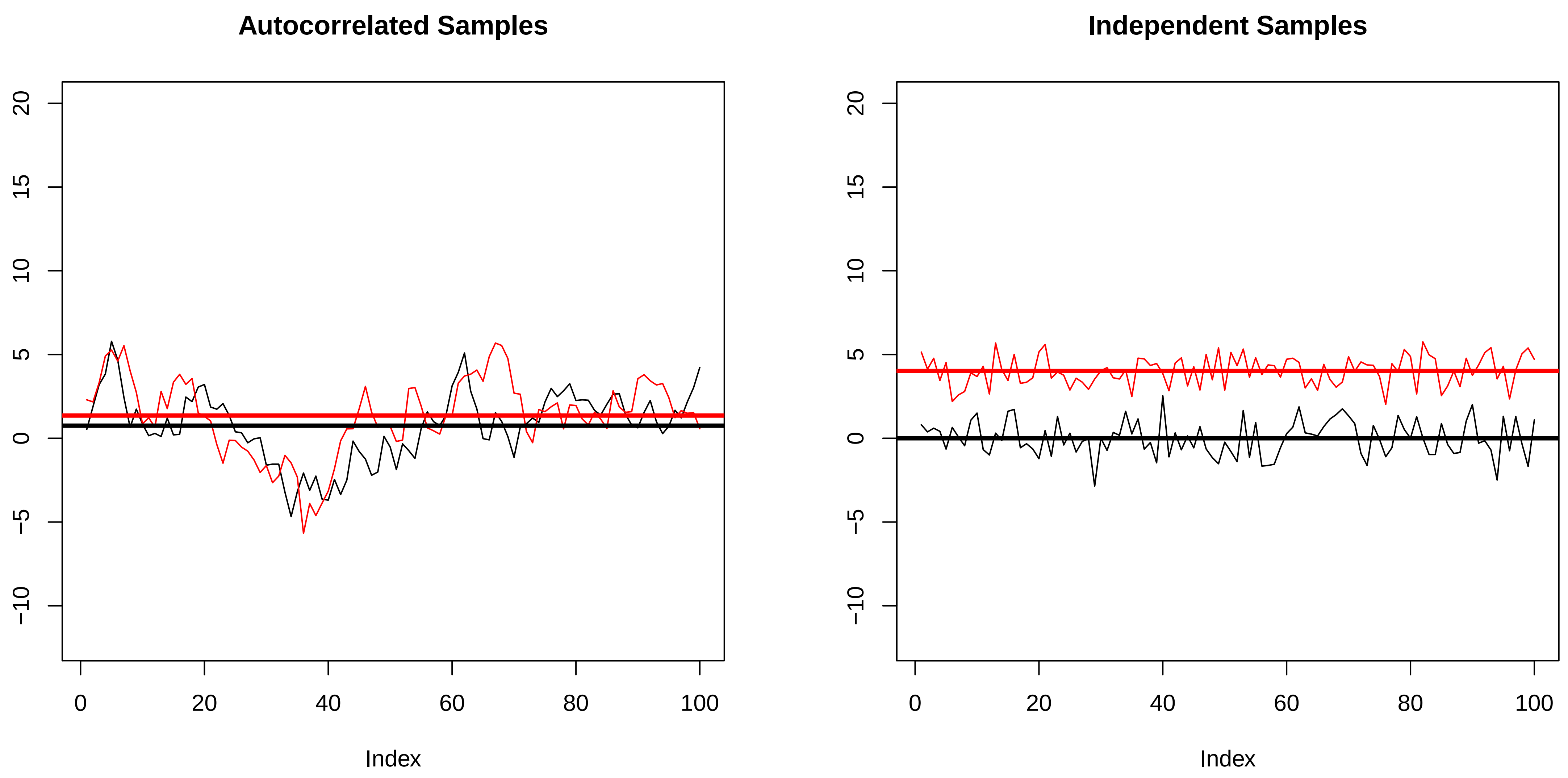
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



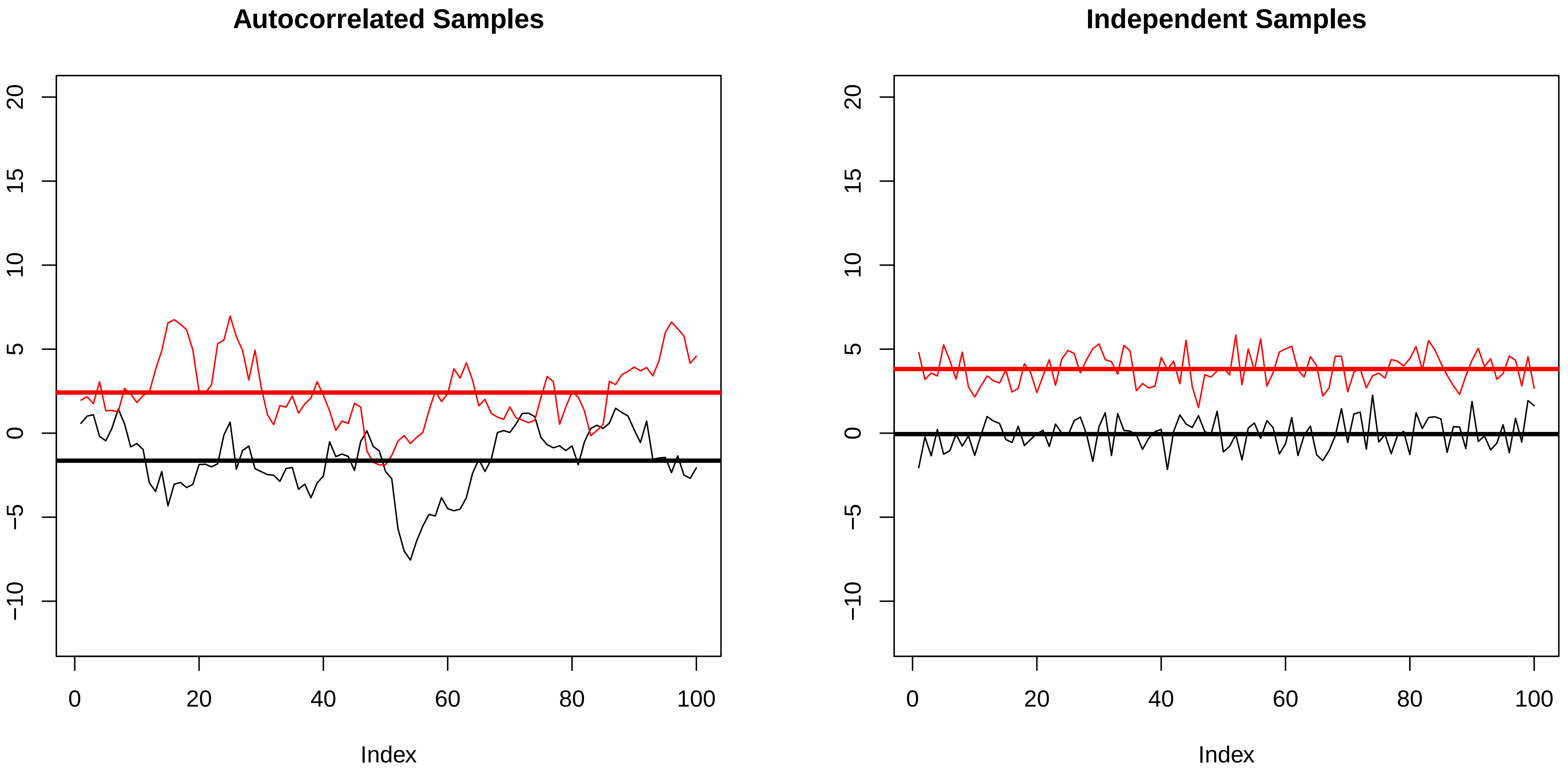
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

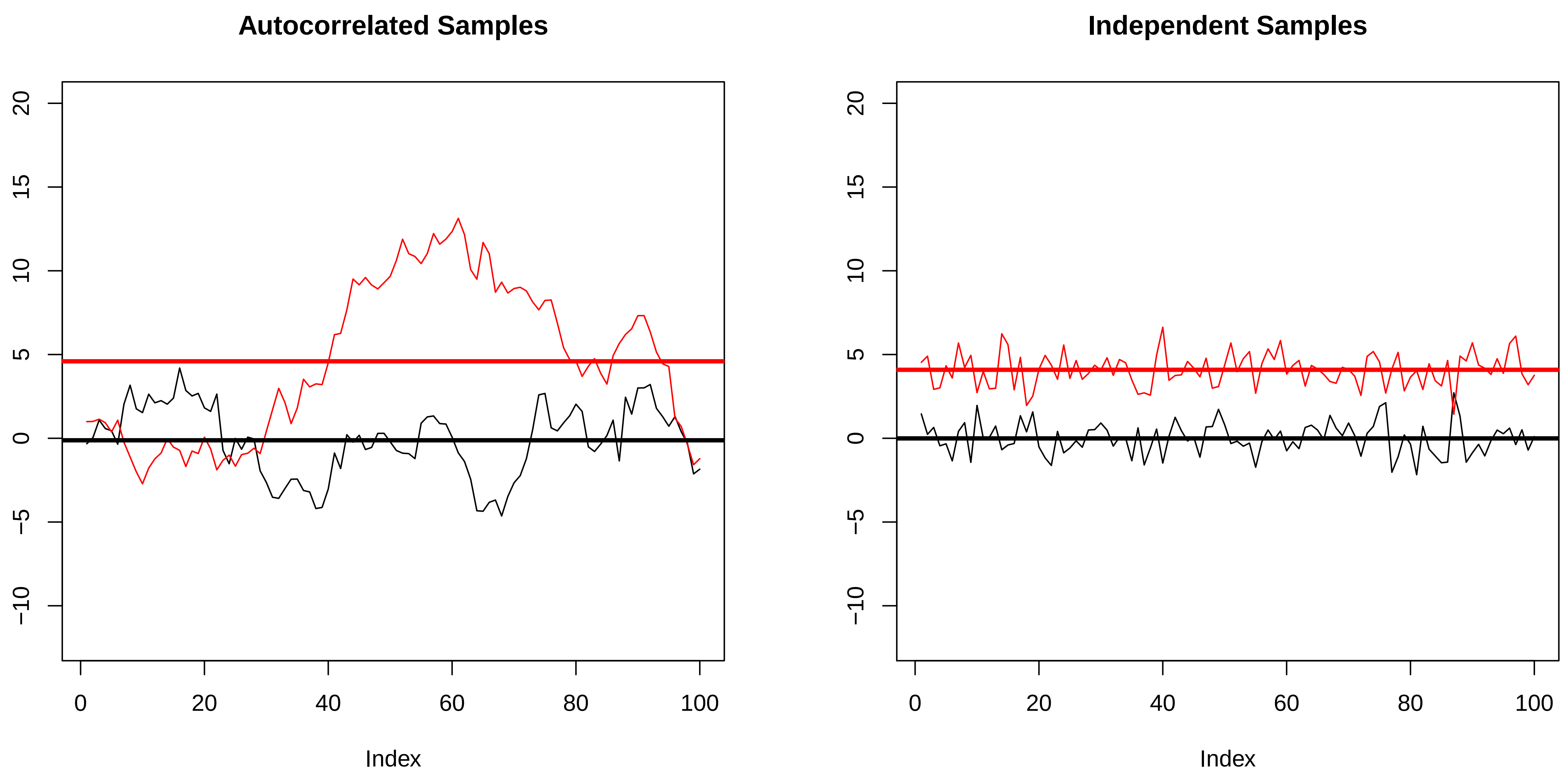

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index



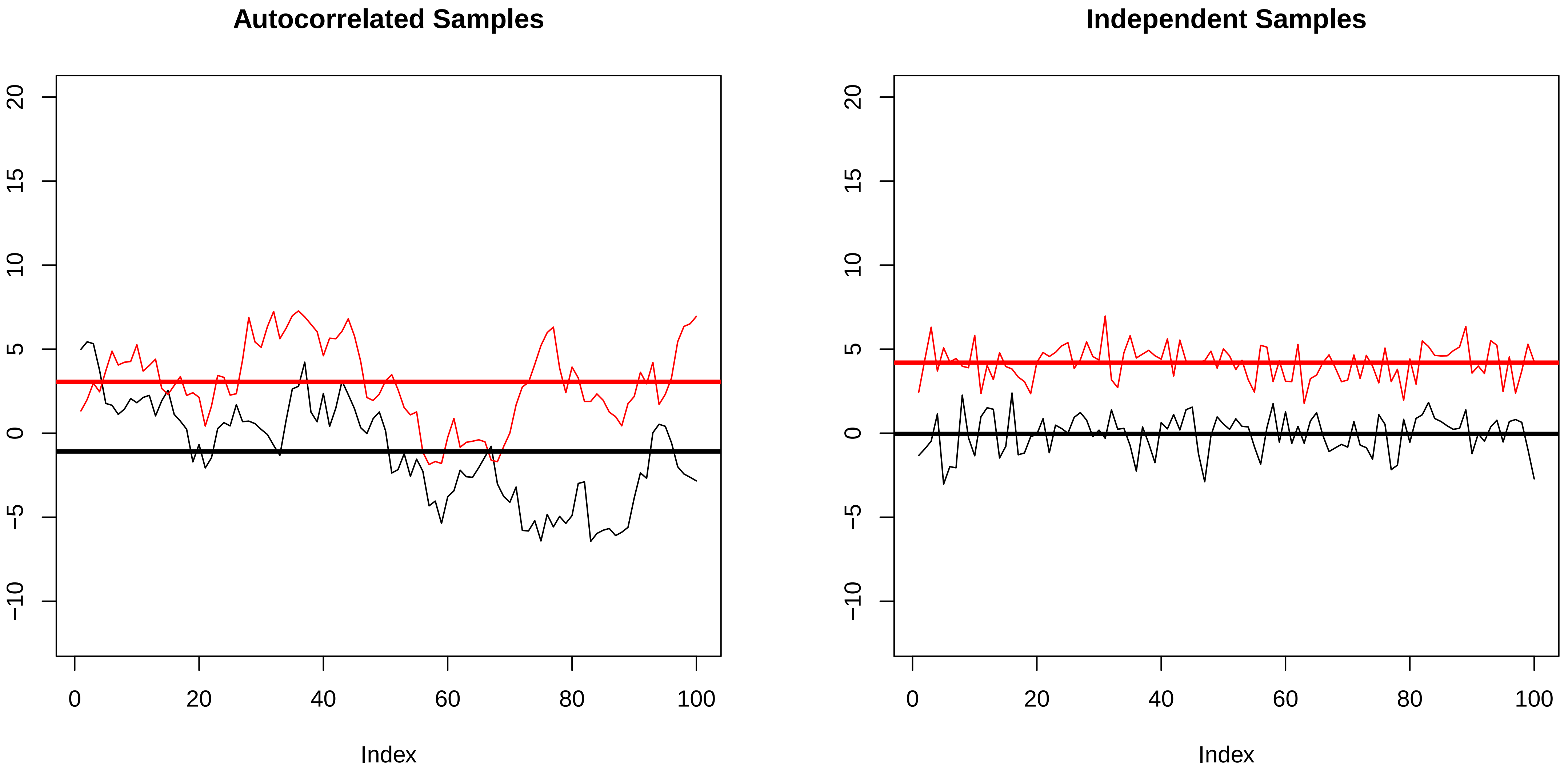
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



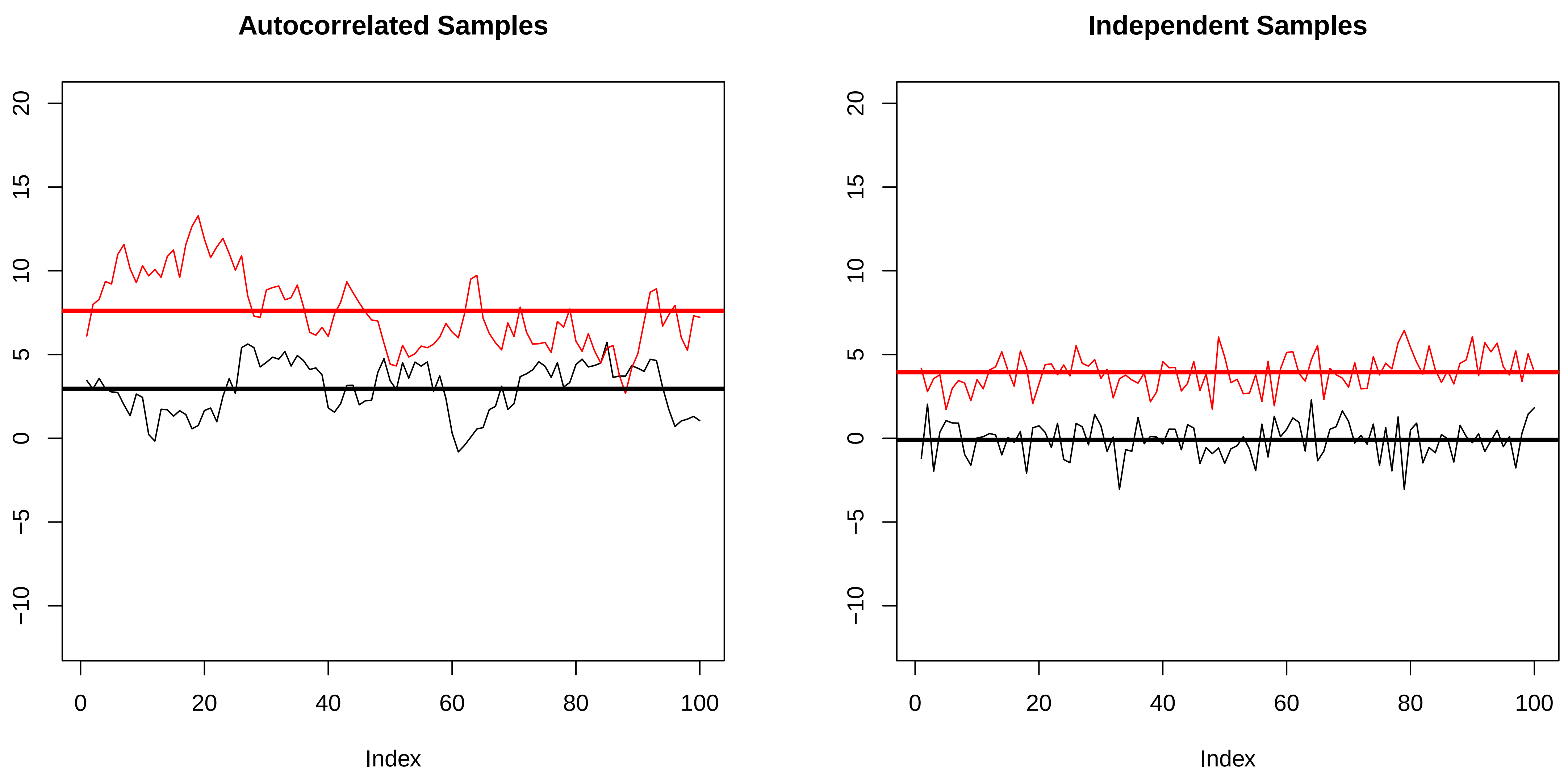
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



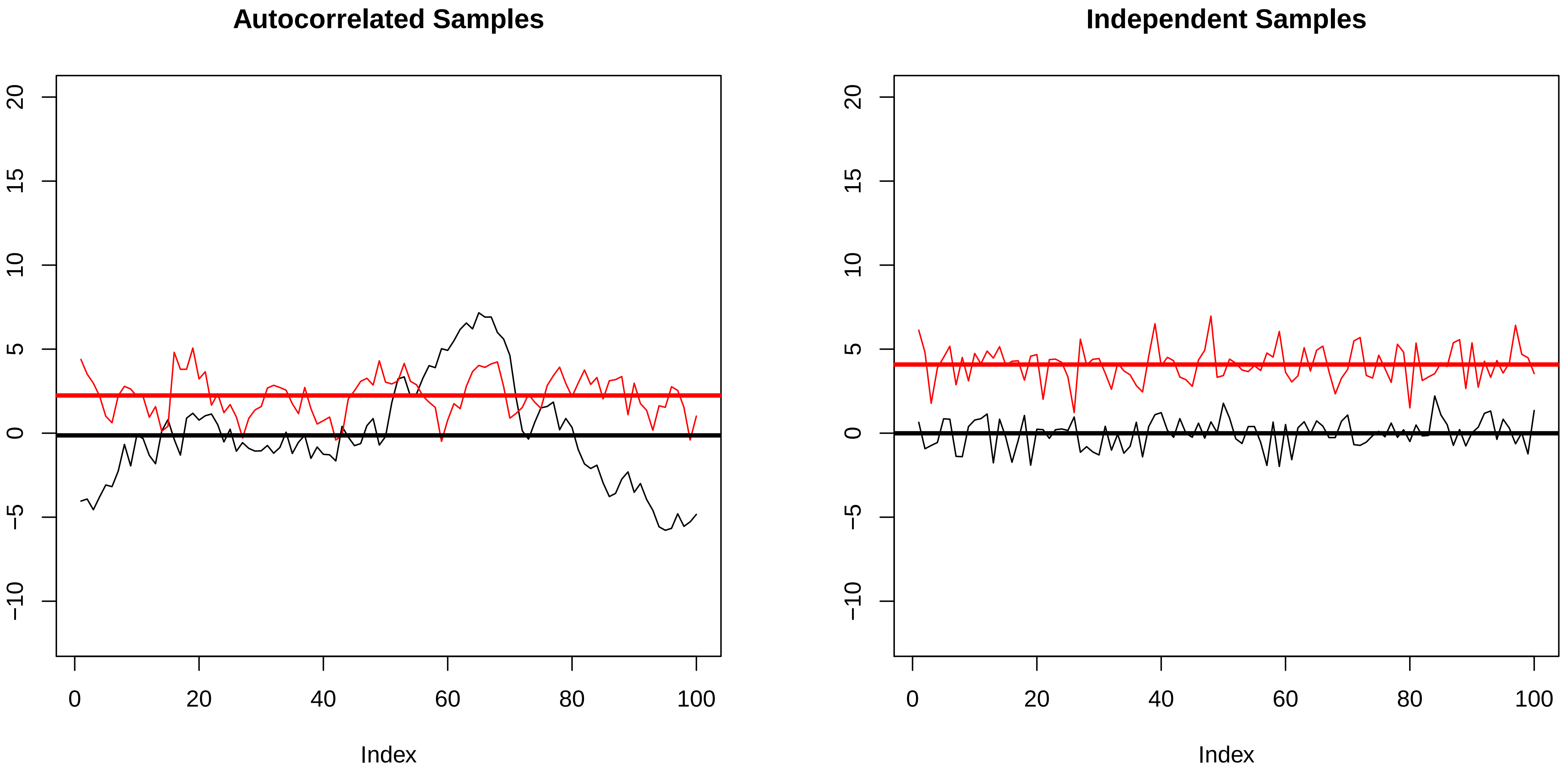
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

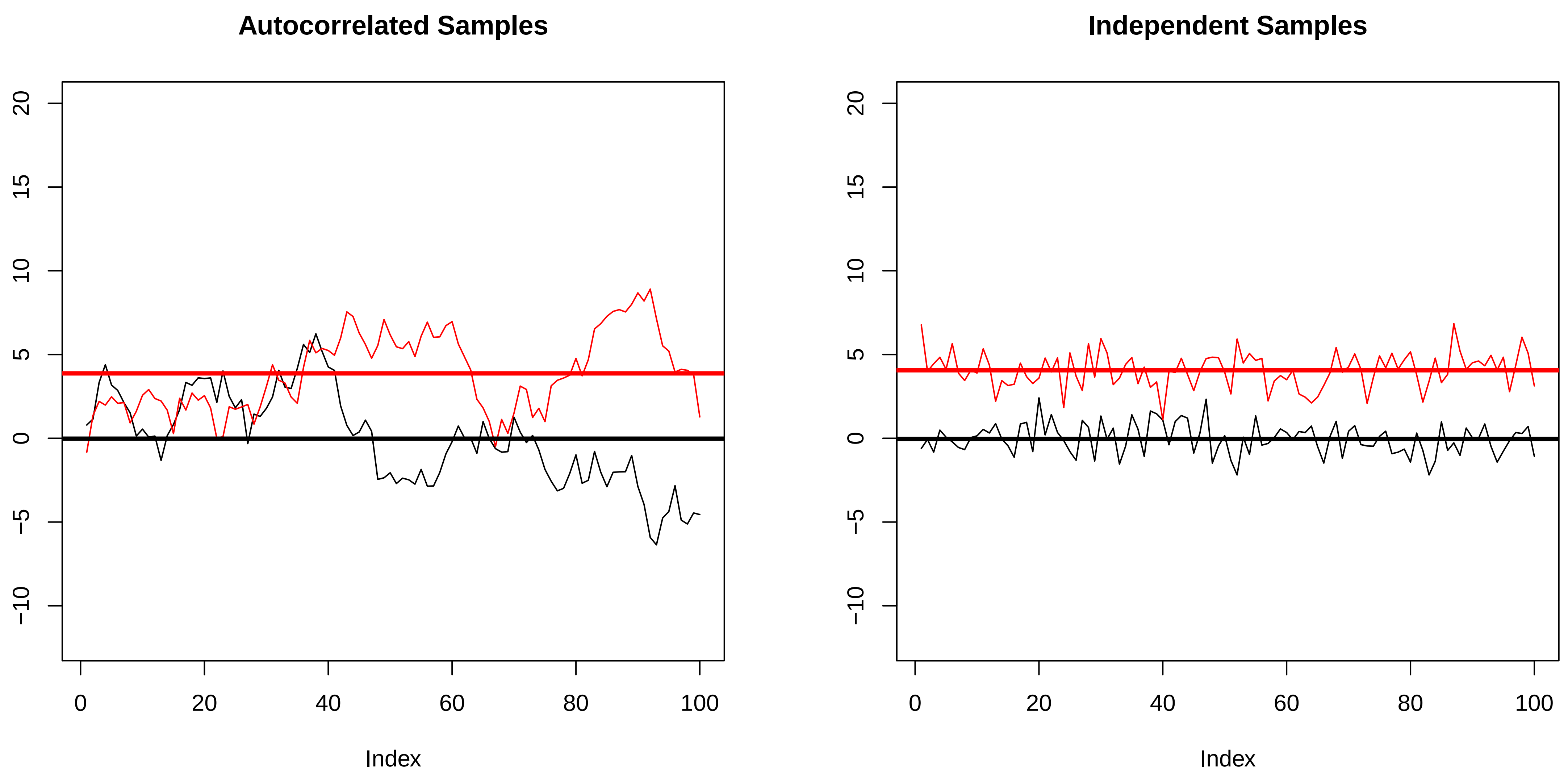


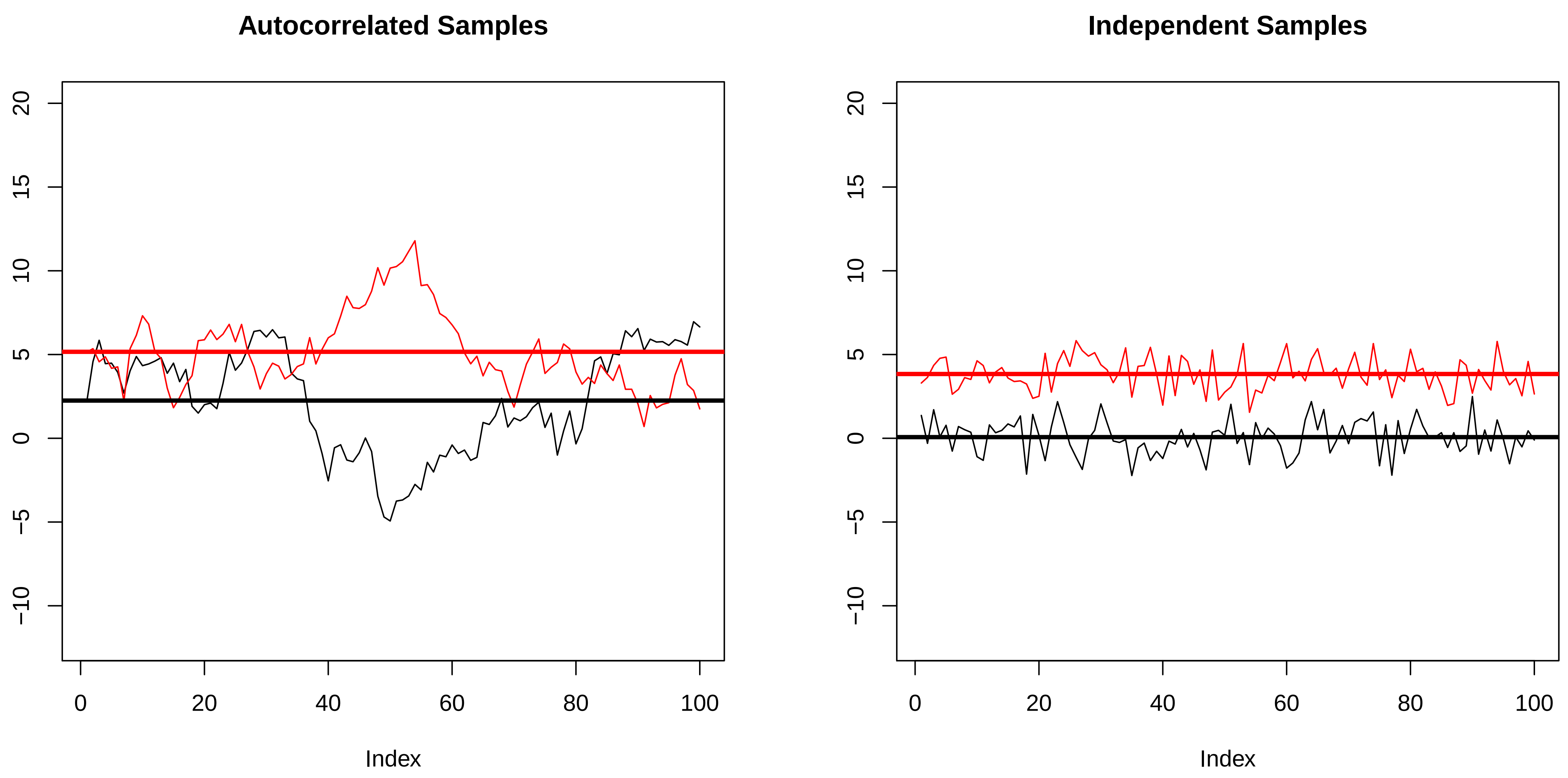
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



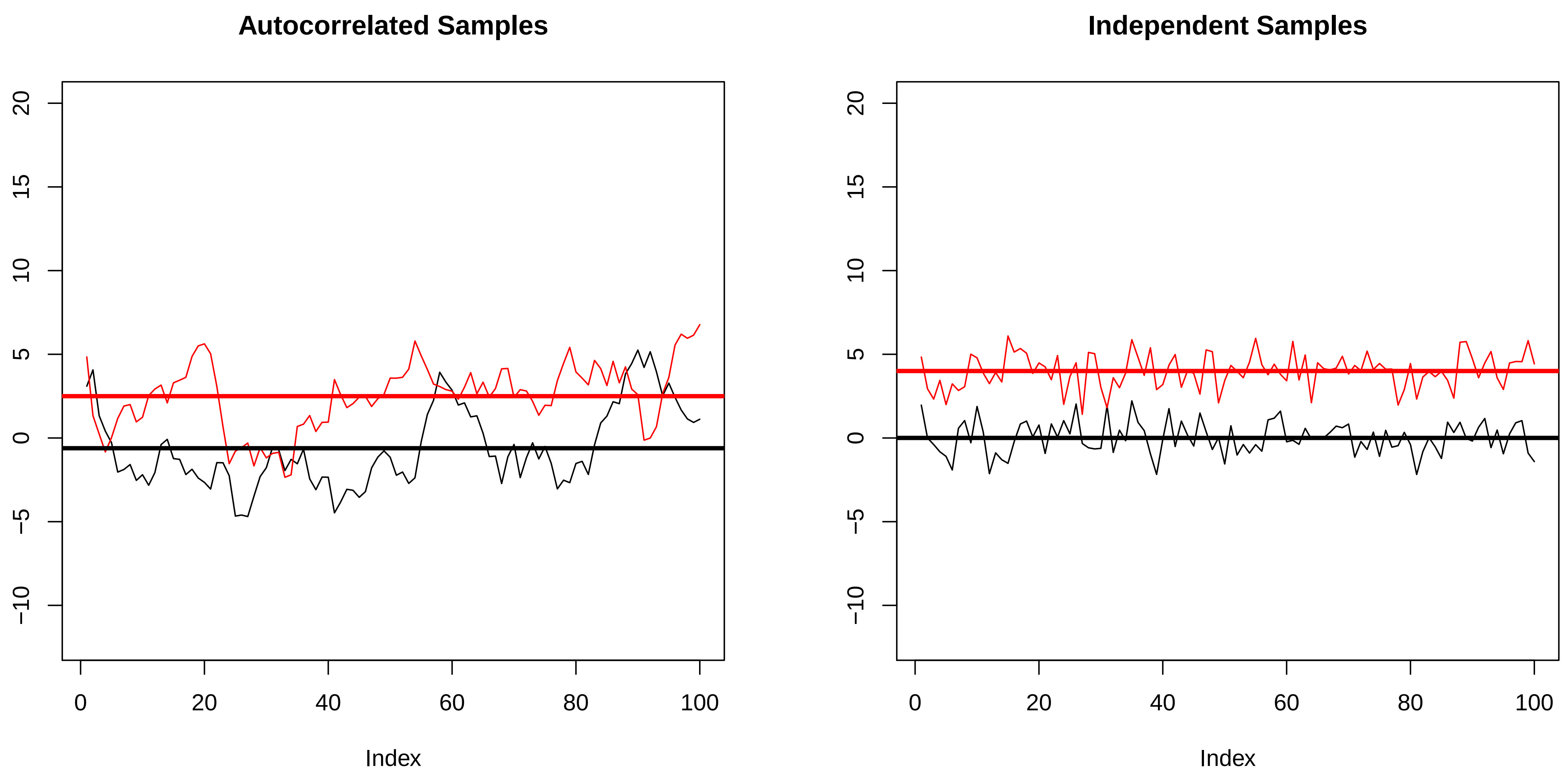
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

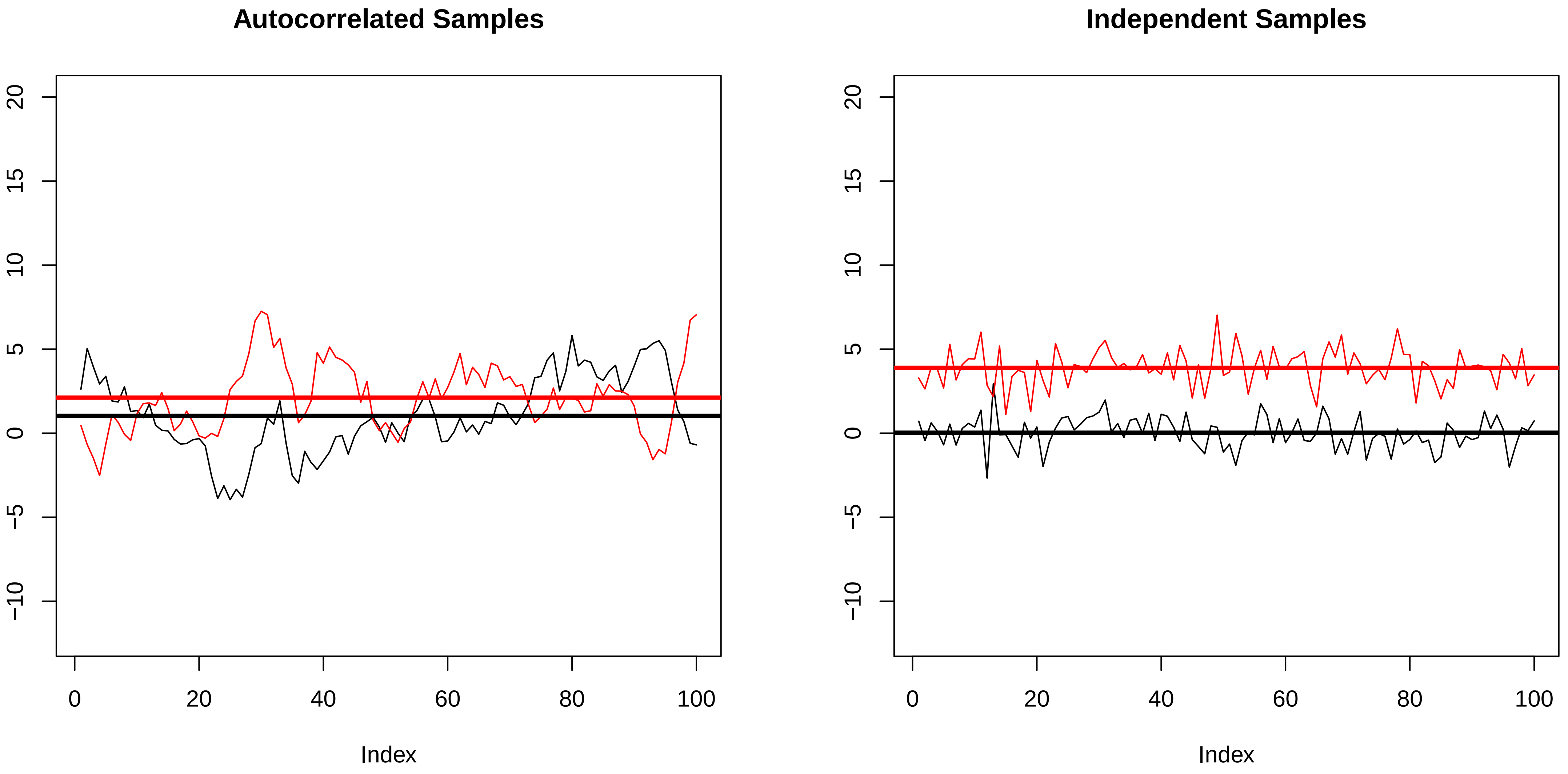

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

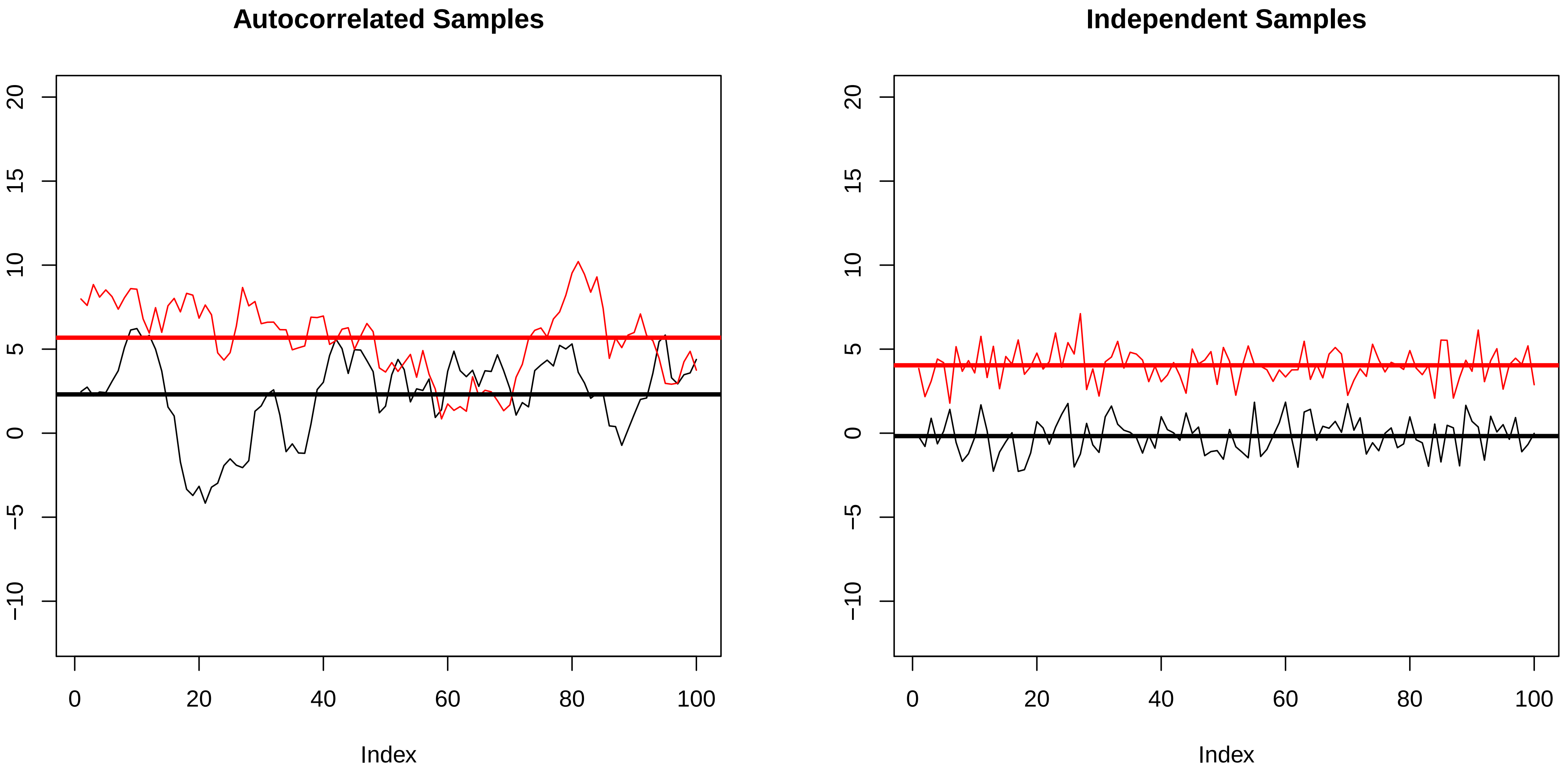

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



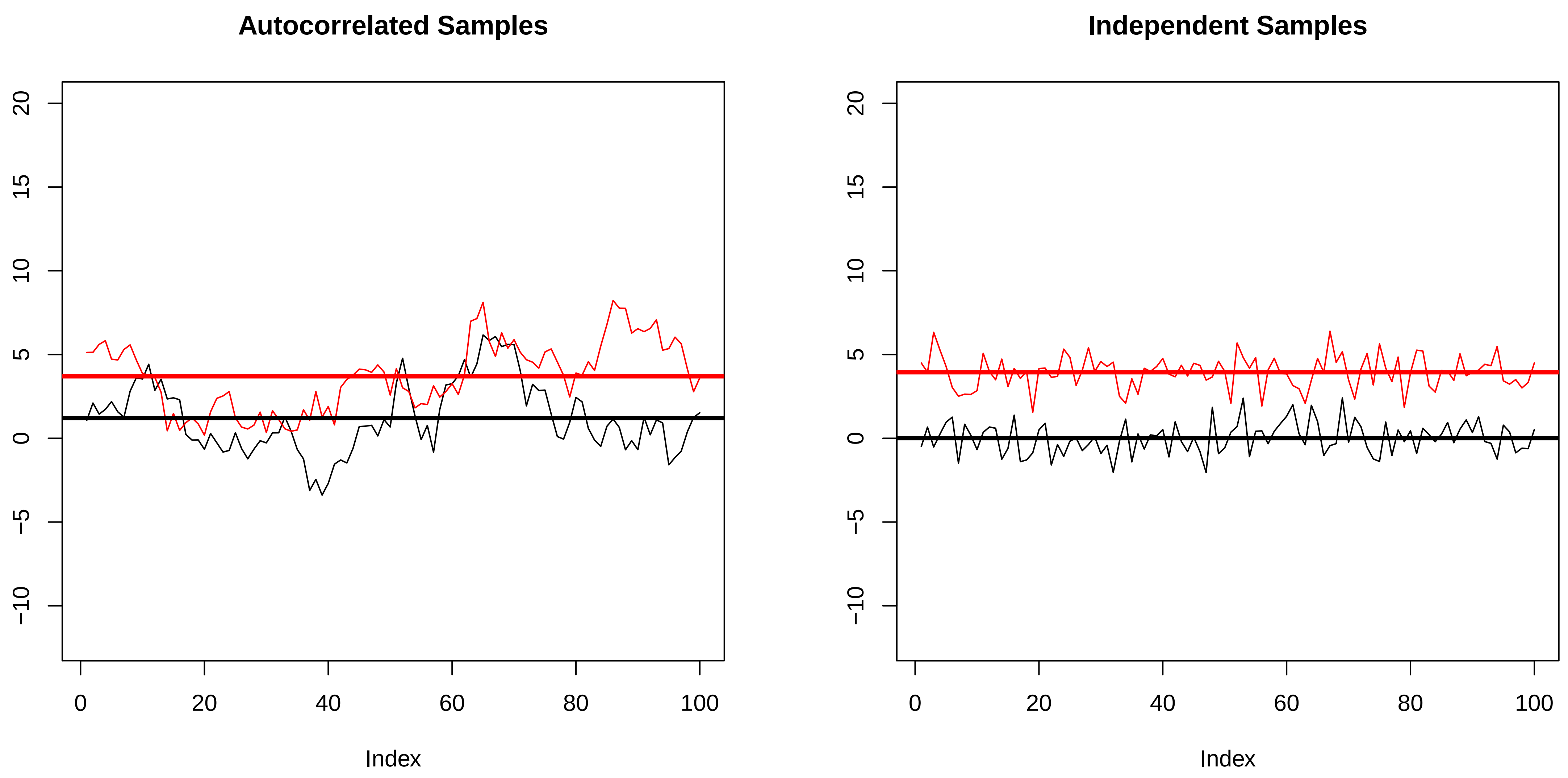
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

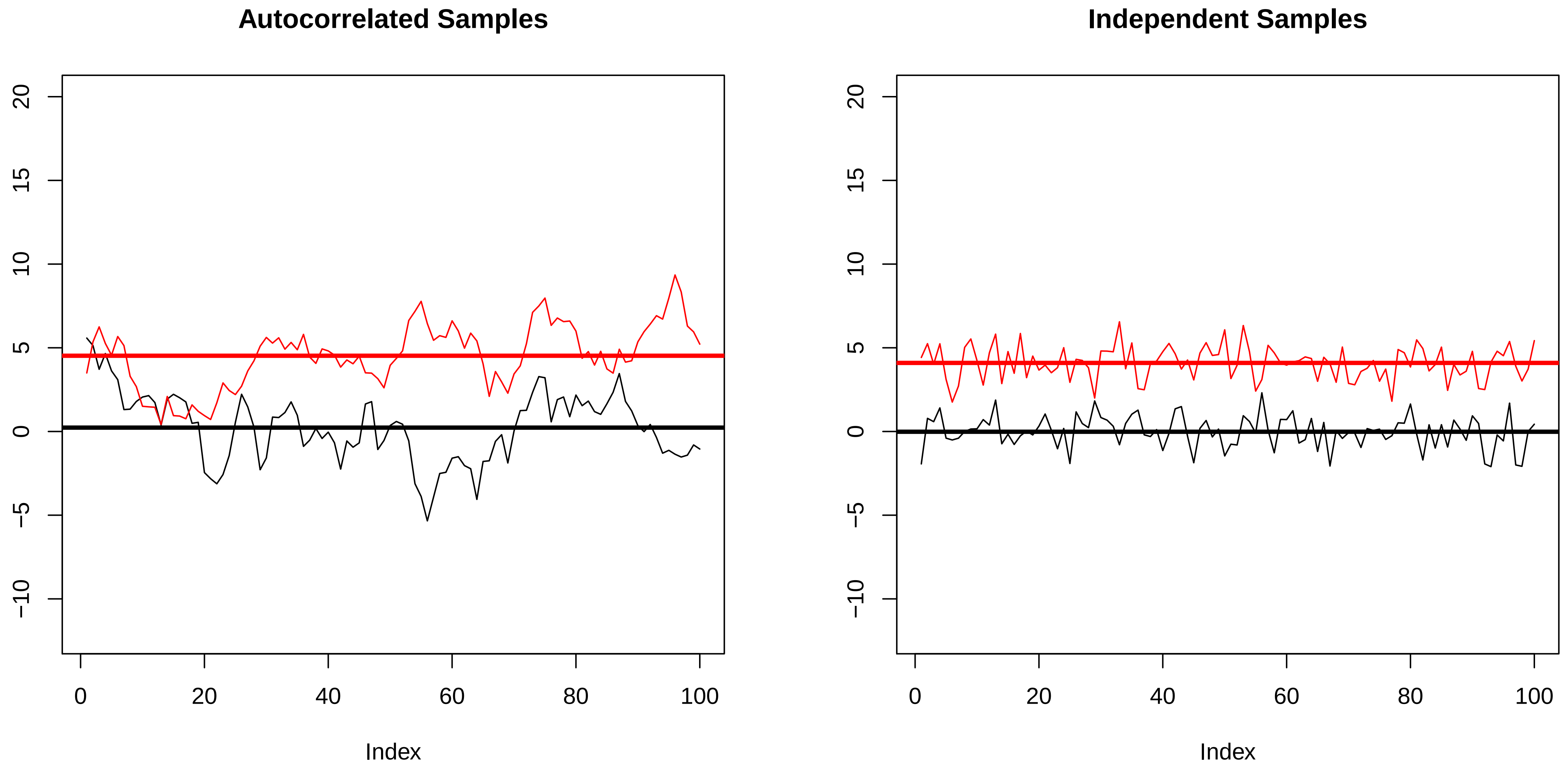

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

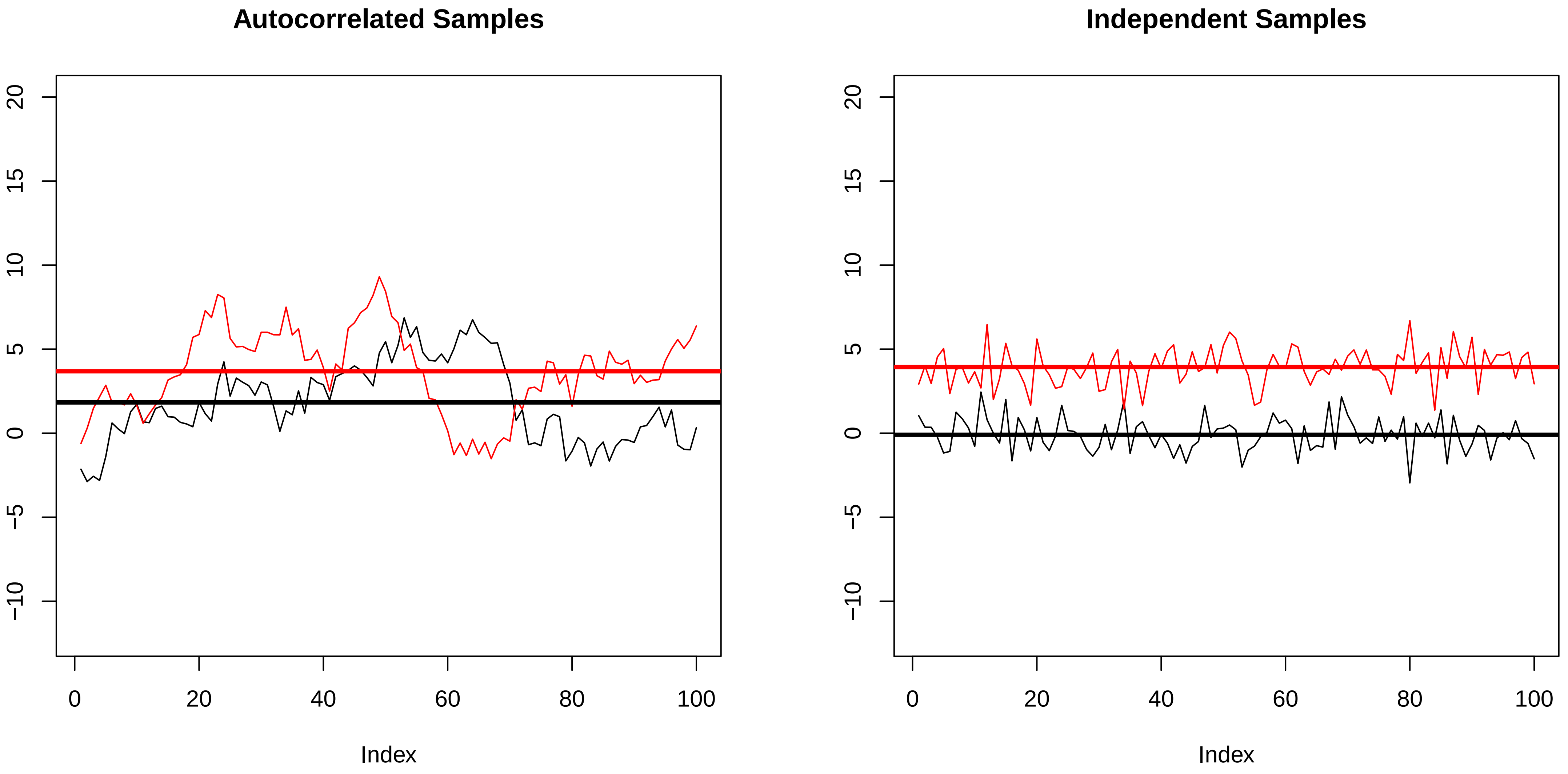

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



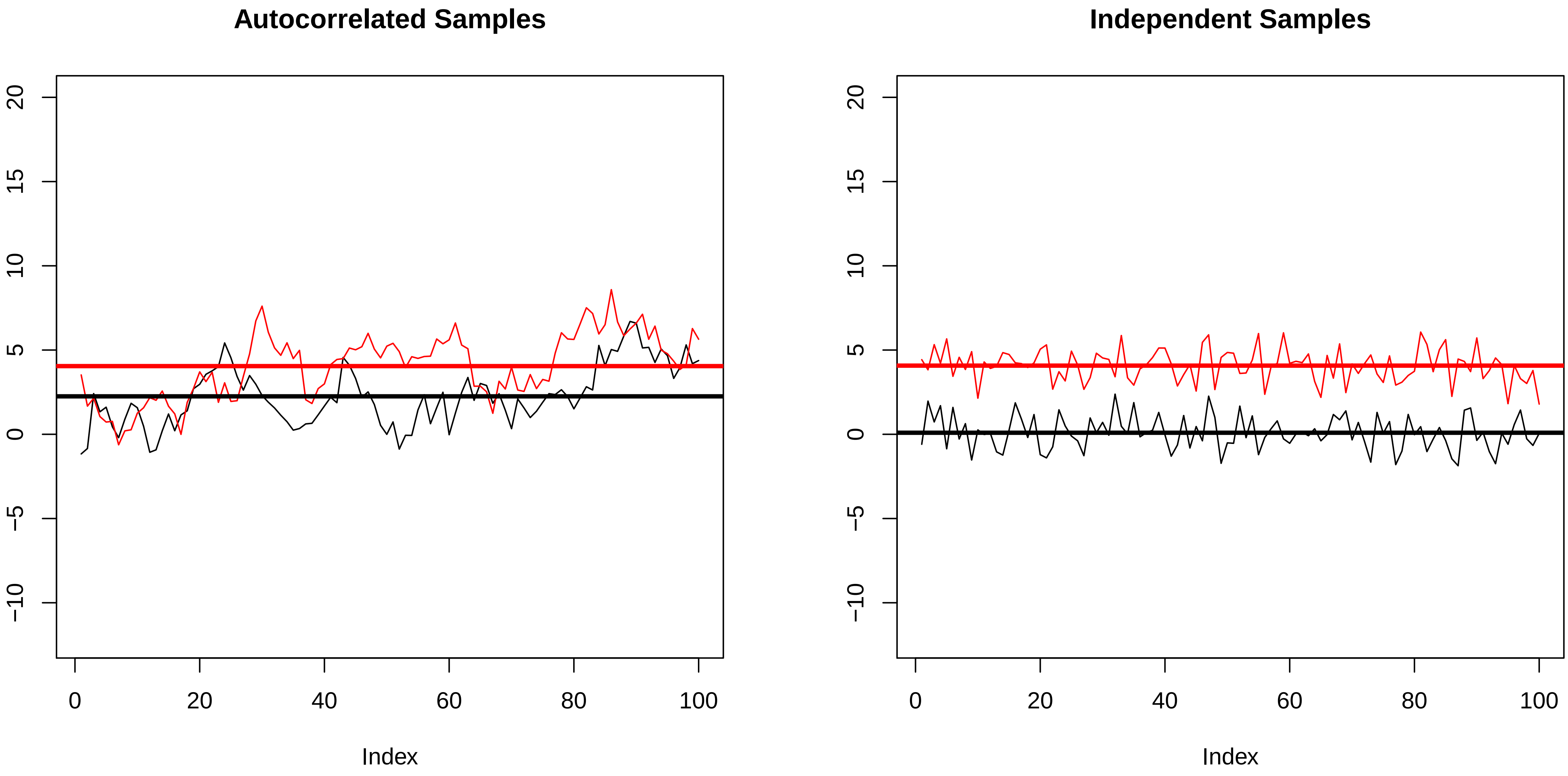
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

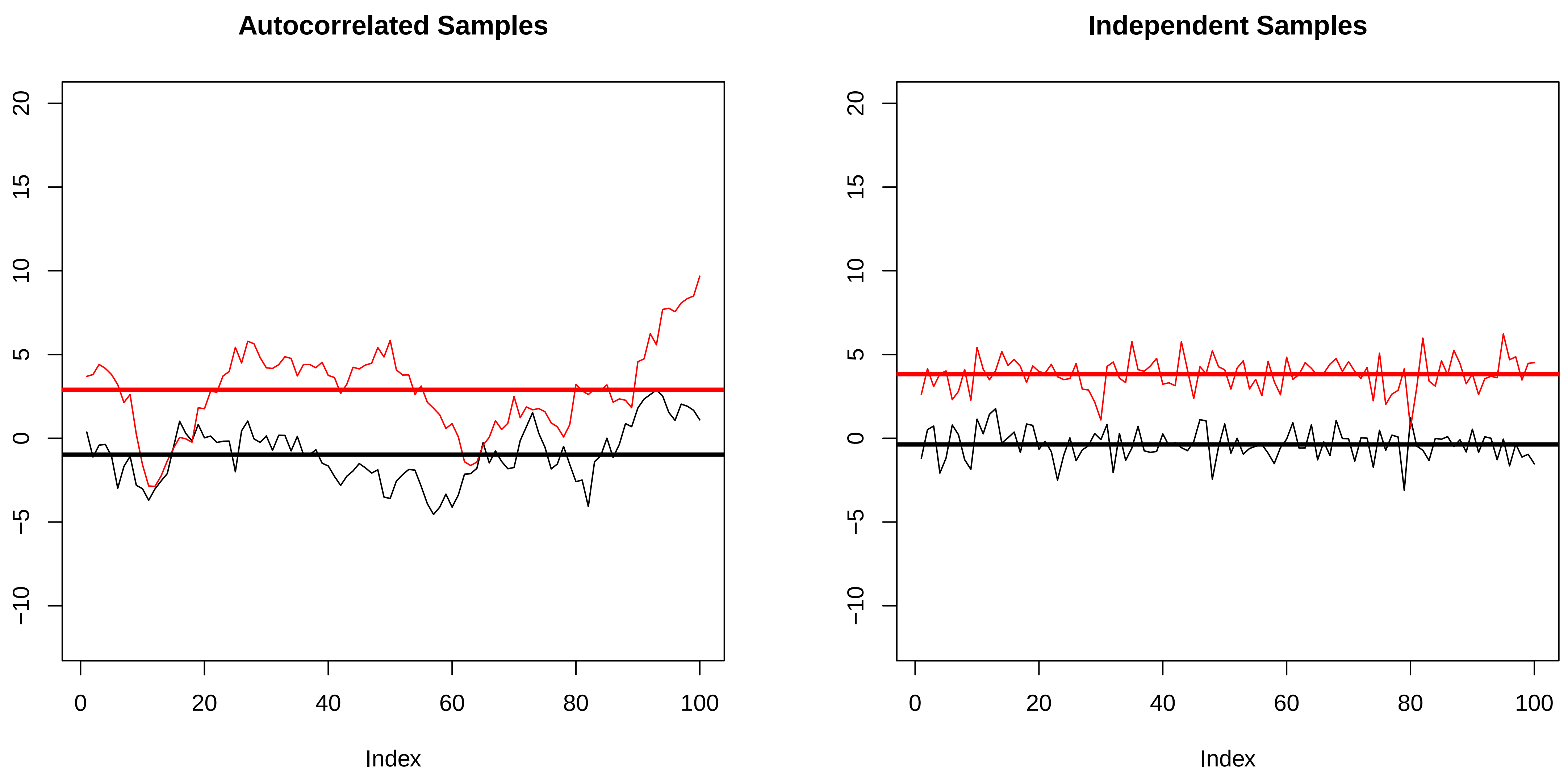

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

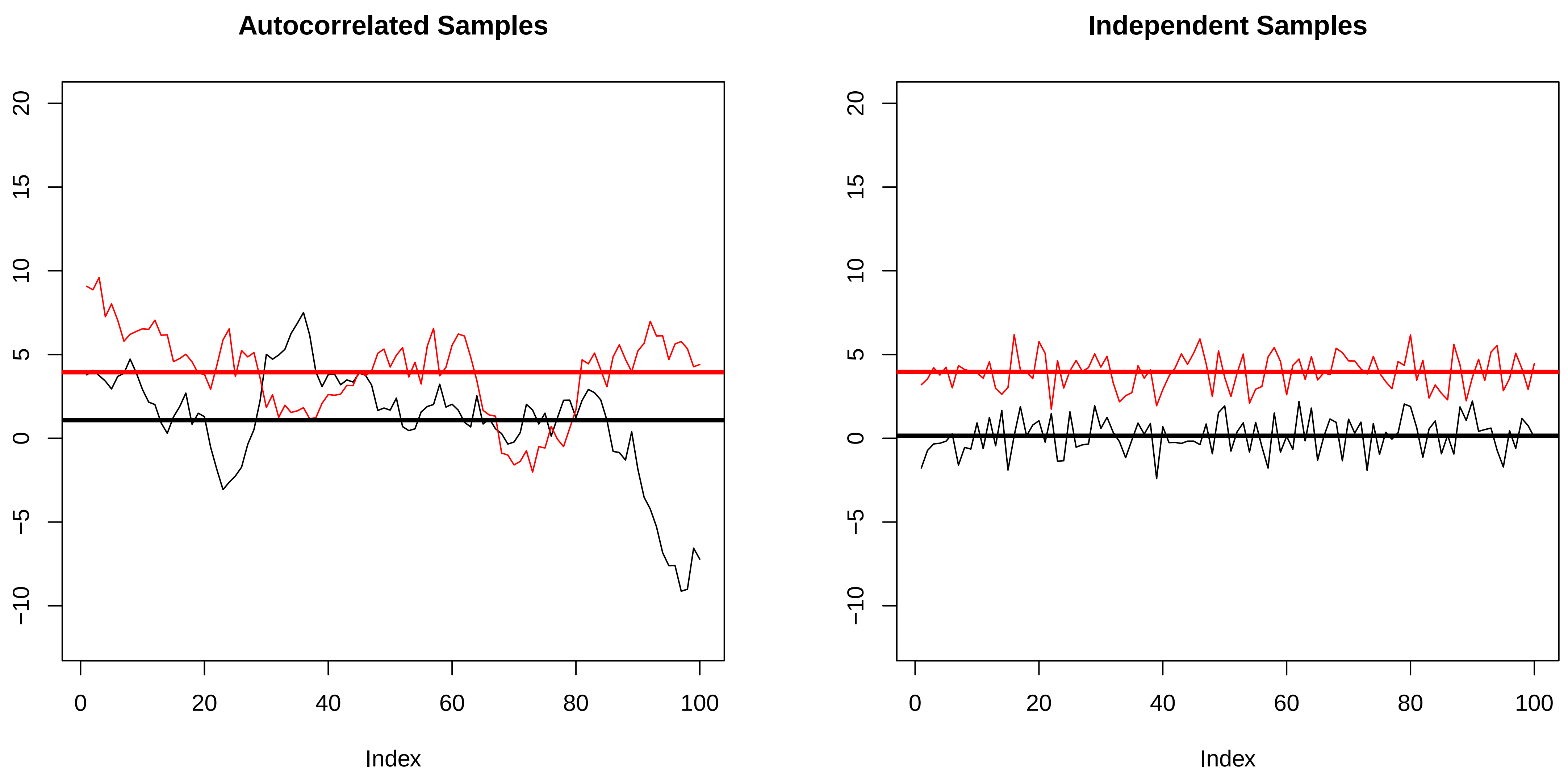


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

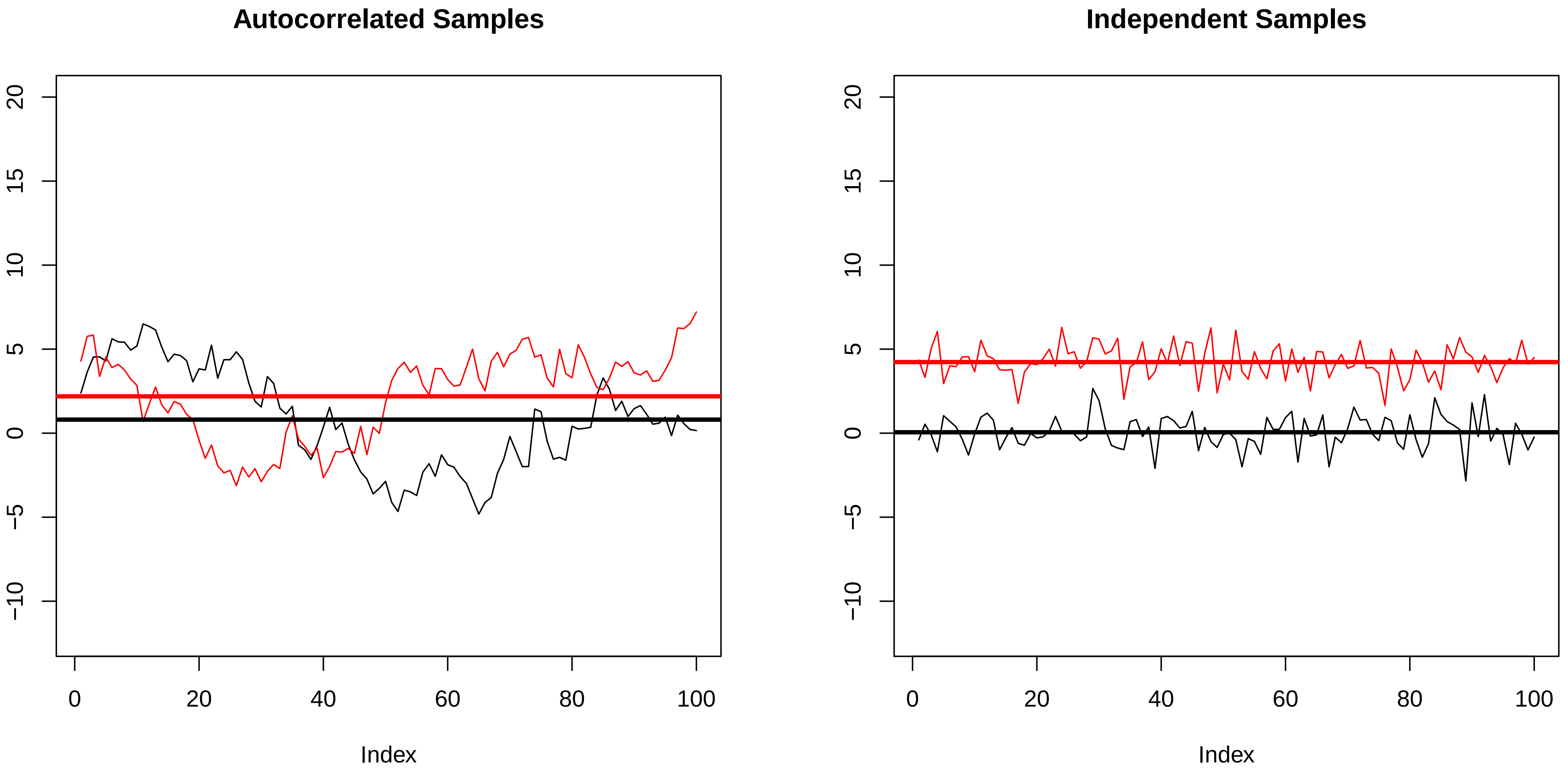

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

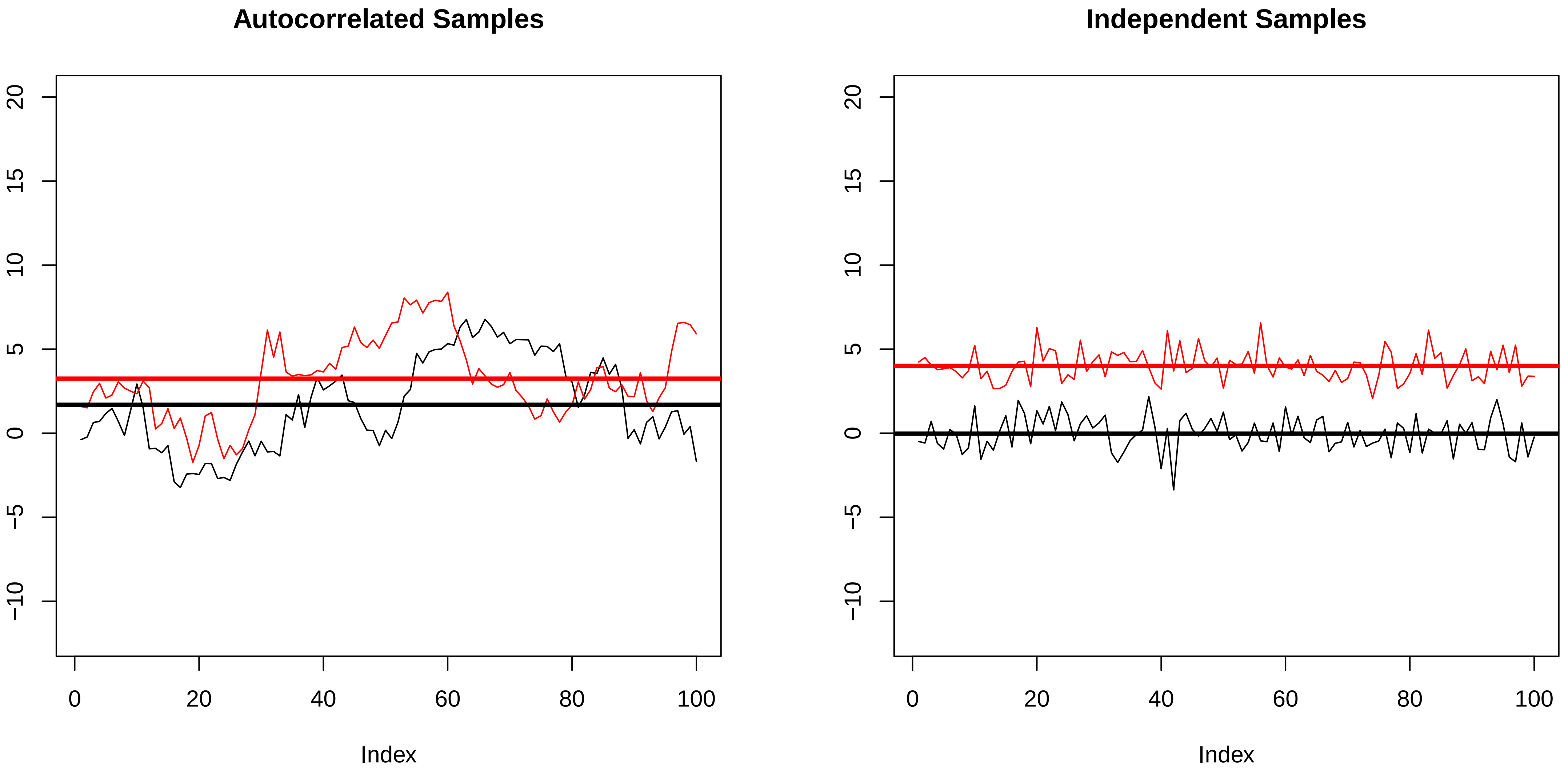


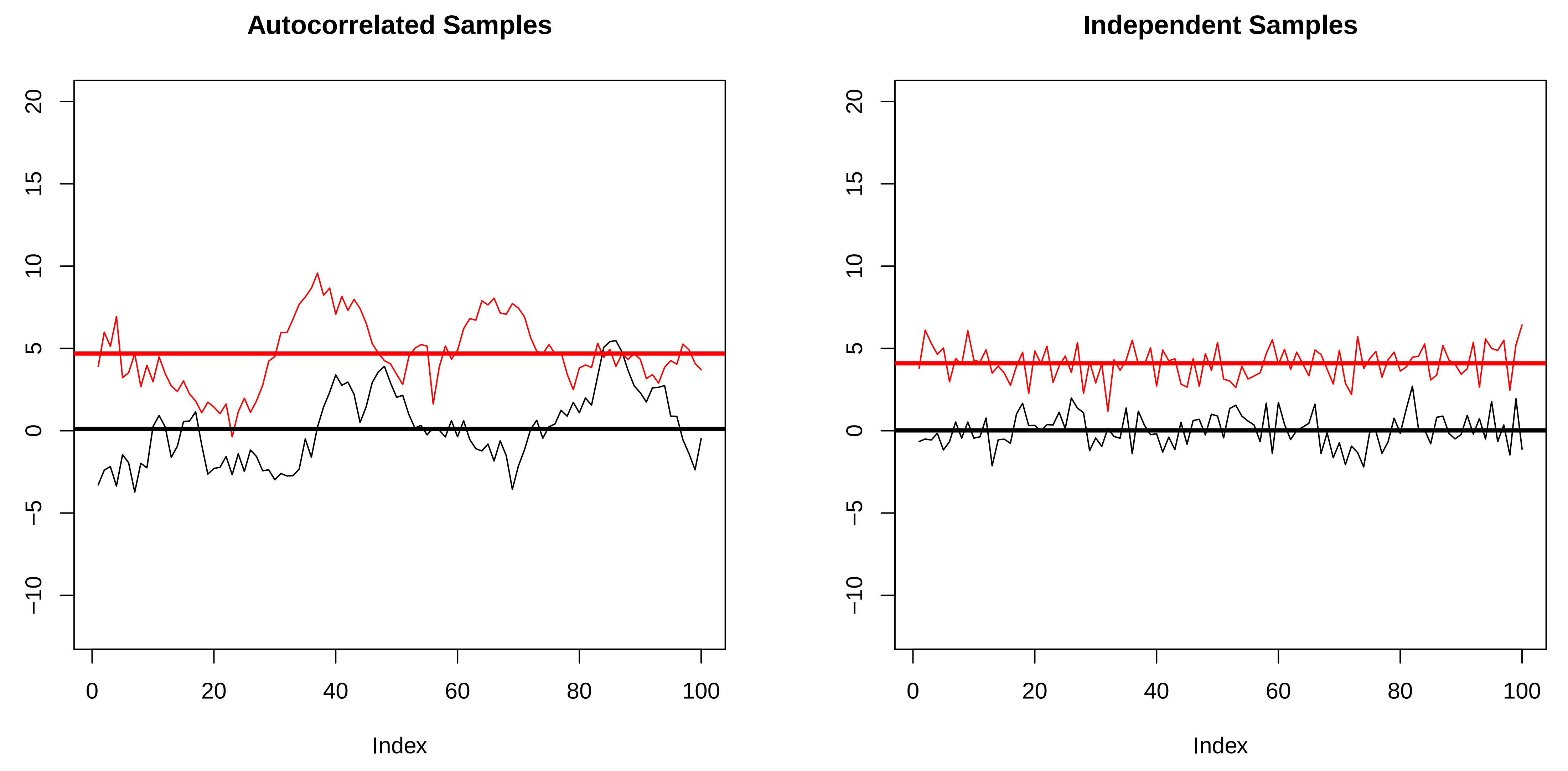
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index



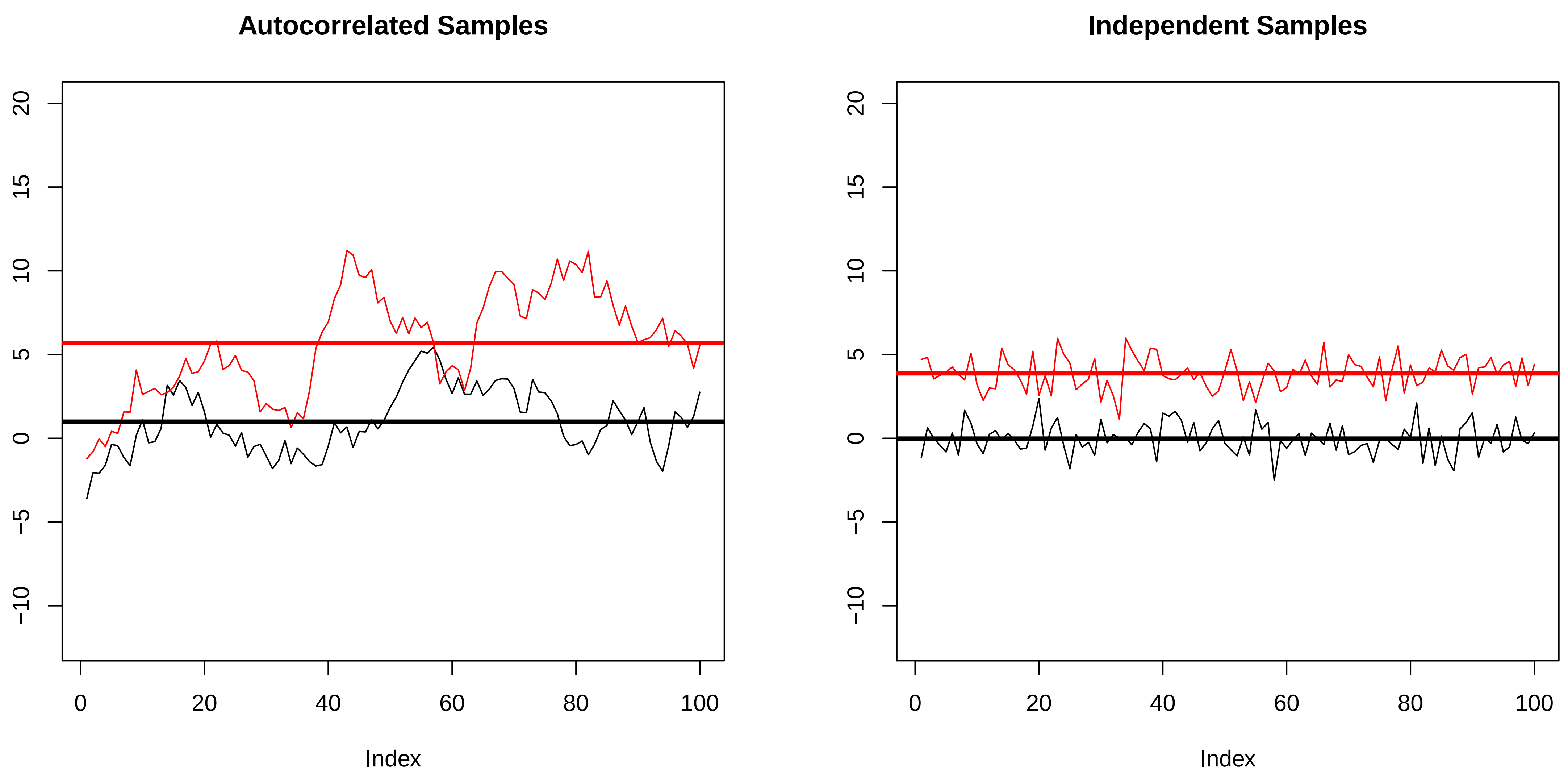
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index

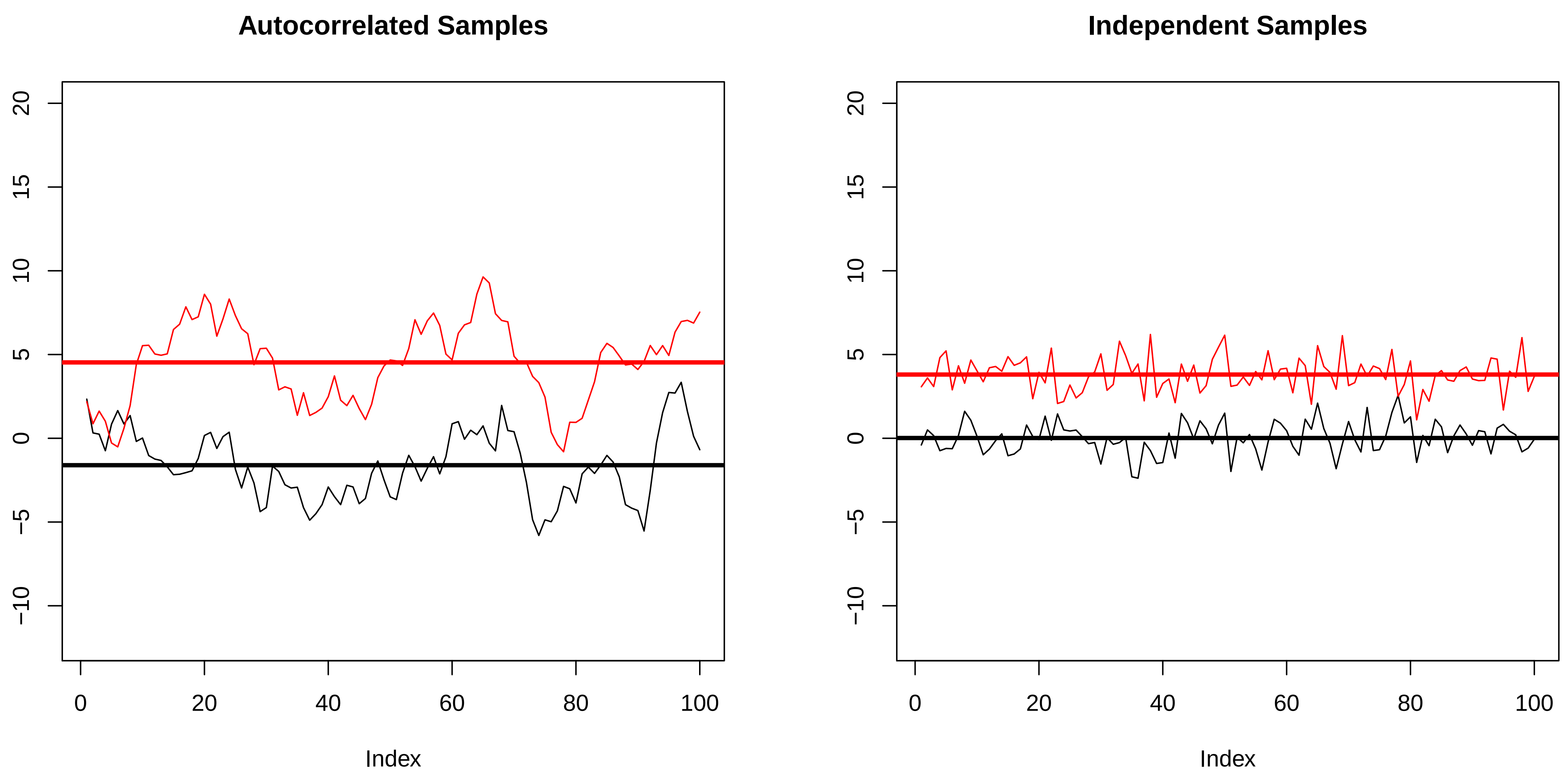

Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index
Index



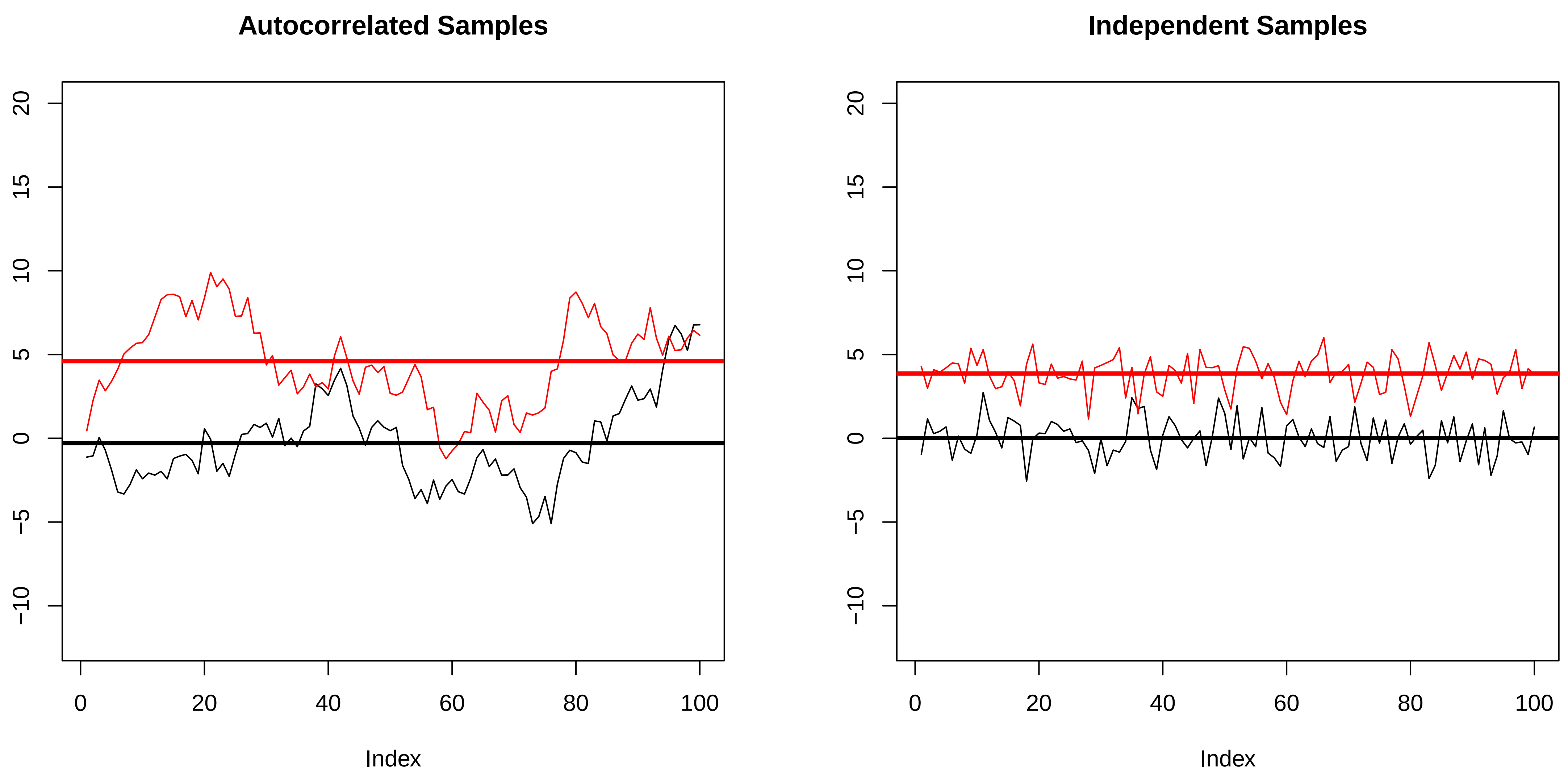
Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index

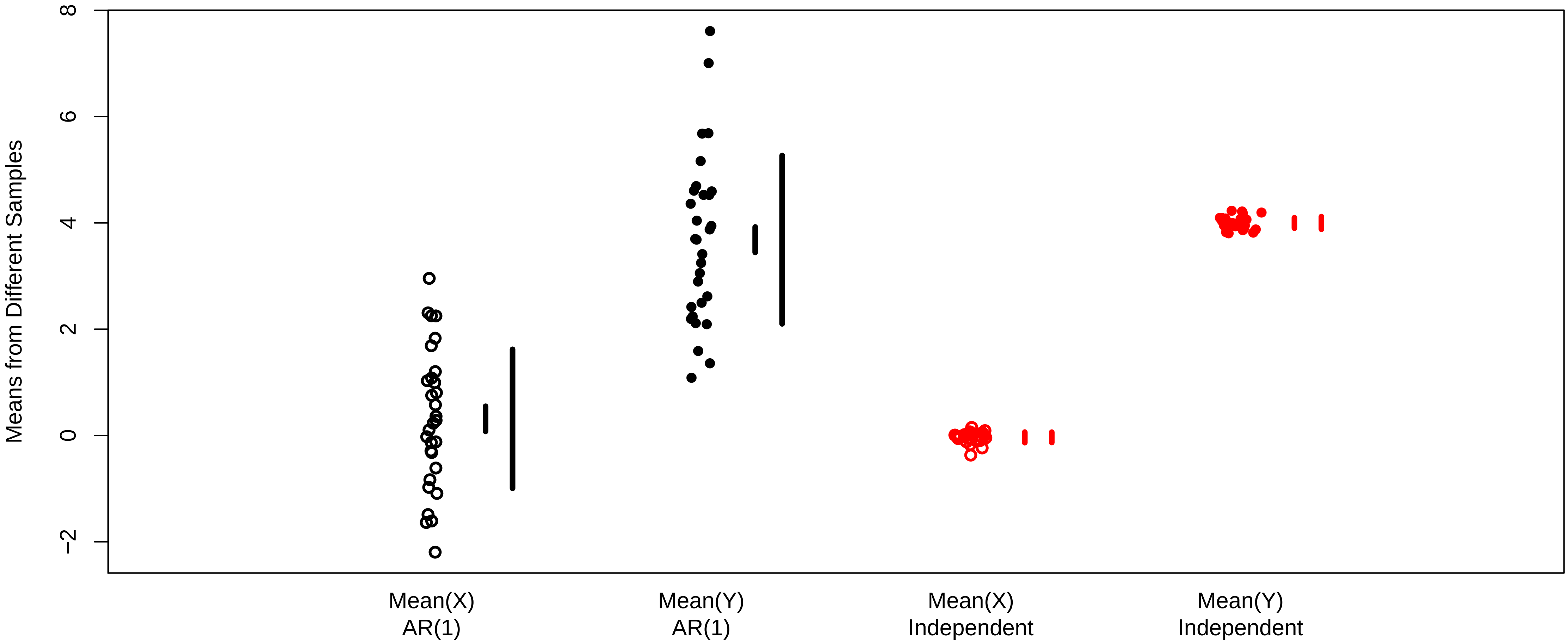


Autocorrelated Samples
Independent Samples
Index
Index



Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index


Autocorrelated Samples
Independent Samples
20
15
10
5
0
−5
−10
0
20
40
60
80
100
Index


Means from Different Samples
8
6
4
2
0
−2
Mean(X)
AR(1)
Mean(Y)
AR(1)
Mean(X)
Independent
Mean(Y)
Independent