

# 單純 回歸分析을 이용한 正規性檢定

## A Normality Test by Using the Simple Regression Analysis

李 昌 鎬\*  
韓 旺 秀\*\*

This paper deals with a normality test to determine whether the data are sampled from normal population or not.

In this paper the property that the mean and variance are independently distributed only for the normal distribution is used as a basis for developing a new test using the simple regression analysis. Considering the mean and variance of a random sample as independent and dependent variables, if it has not the regression relationship we conclude that the data were sampled from the normal distribution.

The Monte-Carlo power study shows that the new test using the simple regression analysis has good power property relative to 6 well-known test methods for 11 distributions.

### 序 論

일반적으로 어떤 統計實驗을 計劃하는 단계에서 당면하게 되는 가장 중요한 문제는 檢査資料들이 어떤 分布를 따를 것인가를 檢定하는 일이다. 그러므로 統計實驗을 計劃하거나 어떤 實驗으로부터의 檢査資料를 分析하기 前에 資料가 어떤 分布를 따를 것인가를 먼저 推定하여야 하는데 實驗을 計劃하는 단계에서는 어떤 分布를 假定할 確實한 指針이 거의 없다.<sup>(1)</sup> 이러한 分布의 推定 및 檢定에 대한 方法으로는 確率用紙나 Histogram 또는 Point Statistic 에 의한 분

布的 推定 및  $X^2$  分布를 이용한 適合度檢定, Kolmogorov - Smirnov 適合度檢定, Cramer-Von Mises 適合度檢定, 標本の Skewness를 이용한 適合度檢定,  $Z$  檢定統計量에 의한 適合度檢定 등이 있다.

Lin 과 Mudholkar에 의해 발표된 새로운 檢定은 正規性を 檢定하는데 있어서, 단지 正規分布일 때만이 平均과 分散이 獨立的으로 分布한다는 성질을 사용하였다. 그러한 새로운 檢定에서는 檢定統計量  $Z$ 의 平均과 分散을 Monte-Carlo Simulation에 의해서 구한 후 檢定을 實施하였으나

\*인하대학교 산업공학과 조교수

\*\*인하대학교 대학원 산업공학과 졸업

本 研究에서는 이와 같은 檢定統計量  $Z$ 를 사용하는 대신에 單純 回歸分析에서 決定係數  $r^2$ 의 信賴性을 檢定하기 위한 檢定統計量  $F_0$ 와 歸無假說  $H_0: b = 0$ 를 檢定하기 위한 檢定統計量  $t_0$ 를 사용하기로 한다. 즉 平均과 分散을 單純 回歸模型의 獨立變數와 從屬變數로 보아 標本平均과 標本分散의 回歸關係 有·無에 따라 그 標本이 正規分布로부터 抽出된 것인지 아닌지를 檢定하고 Lin 과 Mudholkar 와 마찬가지로 이러한 檢定方法의 効率性을 입증하기 위하여 11 種類의 分布에서 시료크기 20 과 30 에 대하여 Monte-Carlo 模擬實驗을 500 번씩 수행하였다. 또 이러한 Monte-Carlo 模擬實驗은 Basic 言語로 Program 했으며 Personal-Computer 를 이용하였다.

[1]  $Z$  檢定統計量에 의한 正規性 檢定<sup>(5)</sup>

Lin 과 Mudholkar 는 “母集團이 正規分布이기 위한 必要하고도 充分한 조건은 그것의 標本平均과 標本分散이 獨立的으로 分布하는 것이다” 라는 特性을 사용하여 비대칭 分布들에 있어서 다른 既存의 檢定方法들 보다는도 우수한 檢出力을 갖는 새로운 檢定方法을 設計하였다. 즉 標本과 平均과 不偏分散  $\bar{x}$  와  $V$ 의 獨立性을 檢定하는 것은 바로 正規性에 대한 適合度의 檢定이 되는 것이다. 여기서  $\bar{x}$  와  $V$ 의 獨立性을 檢定하기 위하여  $n$ 개의 資料  $x_1, x_2, \dots, x_n$ 에서  $n$ 개의 平均과 그것에 해당하는 分散을 다음의 식으로부터 구한다.<sup>(10)</sup>

$$\bar{x}_i = \frac{nx - x_i}{n-1}$$

$$V_i = \frac{1}{n-2} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{n}{n-1} (x_i - \bar{x})^2 \right]$$

$$\bar{x} = \sum_{i=1}^n x_i / n$$

$$i = 1, 2, \dots, n$$

이제는 獨立性을 檢定하는 尺度로써 相關係數

$r$ 을 생각한다. 이러한 相關係數를 根據로한 檢定은 2變數正規分布에 있어서의 獨立性의 檢定에 적절하다. 그러나 分散의 分布는 母集團의 分布가 正規分布를 한다고 할지라도 正規分布에 따르지 않는다. 그러므로 分散의 分布를 근사적으로 正規分布에 따르게 하기 위해서는 Wilson 과 Hilfety(1931)가 제안한 變換  $(V_i)^{1/3}$ 을 통하여 근사적으로 正規分布에 따르게끔 한다. 그러므로 相關係數는

$$r = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{\bar{x}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (\bar{x}_i - \bar{\bar{x}})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$\bar{\bar{x}} = \sum_{i=1}^n \bar{x}_i / n$$

$$y_i = (V_i)^{1/3}$$

$$\bar{y} = \sum_{i=1}^n y_i / n$$

이 된다. 여기서 만약  $r^2$ 이 큰 값을 갖는다면 正規性은 기각될 것이다.

그러나 2變數正規分布로부터 적당數의 標本을 抽出했을 경우에는 상관계수는 正規分布에 따르지 않는다. 이럴 때 Fisher의  $Z$ 變換을<sup>(12)</sup> 사용하게 된다면 檢定統計量  $Z$ 는 標本の 數가 적을 때에도 실제적으로 正規分布에 따르게 된다. 이러한 사실이 바로 檢定統計量  $Z$  즉,

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

를 檢定統計量으로 사용할 수 있게 만들고 만약  $|Z|$ 가 어떤 값  $C$ 보다도 크다면 正規性은 기각될 것이다.

Monte-Carlo Simulation 方法에 의하여  $Z$  檢定統計量의 分布가 Lin 과 Mudholkar 에 의해 研究되었으며 그것은 근사적으로 平均이 0, 分散이  $\sigma_z^2$ 인 正規分布를 따른다는 것이다. 여기서  $Z$ 分布의 分散  $\sigma_z^2$ 은 다음과 같다.

$$\sigma_n^2 = 3/n - 7.324/n^2 + 53.005/n^3$$

또한 Edgeworth 와 Cornish-Fisher 의 근사식을 이용하여

$$P_r(|Z| \geq C) = 2-2\left[\Phi\left(\frac{C}{\sigma_n}\right) - \frac{\gamma_{2n}}{24}\right. \\ \left.\left\{\left(\frac{C}{\sigma_n}\right)^3 - 3\left(\frac{C}{\sigma_n}\right)\right\} \psi\left(\frac{C}{\sigma_n}\right)\right]$$

$\Phi(x)$  : 標準正規累積分布函數

$\psi(x)$  : 標準正規確率密度函數

$$\gamma_{2n} = -11.70/n + 55.06/n^2$$

을 구하였다. 이러한 檢定方法은 正規性を 檢定하는데 있어서 어떠한 檢定方法들 보다도 檢出力이 뛰어나다는 것이 Monte-Carlo Simulation 비교를 통하여 밝혀졌으며 아울러 對稱分布의 非正規性 檢定에 대해서도 檢出力이 좋다는 것이 발표되었다.

## 〔II〕 單純 回歸分析에 의한 正規性 檢定

이 檢定은 어떤 任意標本이 正規母集團으로부터 抽出되었을 必要하고도 充分한 條件은 平均과 分散이 獨立적으로 分布하는 것이라는 특성을 이용하기로 한다. 이 檢定을 수행하기 위해서는 우선  $n$ 個의 資料  $x_1, x_2, \dots, x_n$ 을 얻는다. 이들  $n$ 個의  $x_i$ 로부터  $n$ 個의 平均  $\bar{x}_i$ 와 그것에 해당하는 分散  $V_i$ 을 다음 식으로 부터 얻는다.

$$\bar{x}_i = \frac{n\bar{x} - x_i}{n-1}$$

$$V_i = \frac{1}{n-2} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 - \frac{n}{n-1} (x_i - \bar{x})^2 \right]$$

$$y_i = V_i^{1/3}$$

$$i = 1, 2, \dots, n$$

여기에서  $n$ 個의 平均과 分散의 變換  $\bar{x}_i$ 와  $y_i$ 를 各各 單純 回歸模型의 獨立變數  $x$ 와 從屬變數  $y$ 라 하고 單純 回歸模型을 다음과 같이 表現하기로 한다.

$$y_i = a + bx_i + \varepsilon_i$$

$y_i$  :  $i$  번째 測定된  $y$ 의 값

$a, b$  : 母集團의 回歸係數

$x_i$  :  $i$  번째 주어진 說明變數  $x$ 의 값

$\varepsilon_i$  :  $i$  번째 測定된  $y$ 의 誤差項으로 確率 分布는 平均이 0, 分散이  $\sigma^2$ 인 正規分布에 따르며 다른 誤差項과는 相關關係가 없다.

만약 임의표본의 平均과 分散이 回歸關係가 없다면 獨立變數  $x$ 와 從屬變數  $y$ 는 獨立인 것으로 간주하여 “母集團이 正規分布이기 위한 必要하고도 充分한 條件은 標本平均과 標本分散이 獨立적으로 分布하는 것이다”라는 특성에 부합하므로 그 標本은 正規母集團으로부터 抽出된 標本이라는 結論을 내려준다.

우선  $H_0: r^2 = 0$ 와  $H_1: r^2 \neq 0$ 를 檢定하기 위하여 檢定統計量  $F_0$ 를

$$F_0 = \frac{(n-2) \text{SSR}}{\text{SSE}}$$

$$\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

로부터 구한다. 만약  $F_0$ 의 값이 커서 즉  $F_0 > F_{(1, n-2; 1-\alpha)}$ 이면  $H_0: r^2 = 0$ 를 有意水集  $\alpha$ 에서 기각하고 이 때에는  $y$ 의 變動에  $x$ 가 중요한 영향을 준다고 생각하여 그 任意標本이 正規分布로부터 抽出된 標本이 아니라는 結論을 내려주고 만약 그렇지 않으면  $H_0: r^2 = 0$ 를 採擇하여 이 때의 任意標本  $x_1, x_2, \dots, x_n$ 는 正規分布로부터 抽出된 標本이라는 結論을 내려

준다.

다음으로  $H_0 : b = 0$ 와  $H_1 : b \neq 0$ 를檢定하기 위하여檢定統計量  $t_0$ 를

$$t_0 = \frac{\hat{b}}{\sqrt{\text{Var}(\hat{b})}}$$

$$= \frac{\hat{b}}{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{(n-2) \sum_{i=1}^n (x_i - \bar{x})^2}}}$$

로부터 구한다. 만약  $t_0$ 의 절대값이 自由度  $\phi = n - 2$ 에 해당하는  $t$ 分布表의 값  $t_{(n-2), 1-\frac{\alpha}{2}}$ , 보다 큰 경우 즉  $|t_0| > t_{(n-2), 1-\frac{\alpha}{2}}$ 인 경우 歸無假說  $H_0 : b = 0$ 를 기각하고 이 때에는 두 變數  $x$ 와  $y$ 는 回歸關係가 있다고 보아 任意標本  $x_1, x_2, \dots, x_n$ 은 正規分布로부터 抽出된 標本이 아니라는 結論을 내려주기로 한다. 그러나 만약  $|t_0| < t_{(n-2), 1-\frac{\alpha}{2}}$ 인 경우에는  $H_0 : b = 0$ 를 有意水準  $\alpha$ 에서 採擇하

고 이 때에는 任意標本  $x_1, x_2, \dots, x_n$ 가 正規分布로부터 抽出된 標本이라는 結論을 내려주기로 한다.

### 〔Ⅲ〕 正規性 檢定の Monte-Carlo Simulation 比較

$Z$ 檢定統計量을 사용하여 正規性を 檢定하는 것이 그밖의 檢定統計量을 사용할 때 보다도 더 우수한 檢出力이 보장된다는 것이 Lin과 Mudholkar에 의해 보여졌다. 즉 標本數 20과 30에 대하여 Monte-Carlo Simulation을 各分布마다 1,000번씩 수행하여 그 檢出力이 매우 우수하다는 것을 보였다.<sup>(5)</sup>

本 研究에서는 單純 回歸分析에 의한 正規性檢定の 檢出力과  $Z$ 檢定統計量 및 그밖의 檢定統計量을 사용했을 때의 檢出力을 比較할 목적으로 各分布에서 標本數 20과 30에 대하여 Monte-Carlo Simulation을 500번씩 수행하였다. 이와 같이 單純 回歸分析에 의한 正規性檢定の 檢出力과 Lin과 Mudholkar에 의한 正規性檢定の 結果들을 정리하면 表1과 表2와 같다.

< 表 1 > 正規性에 대한 여러 檢定方法들에 대한 Monte-Carlo Simulation 檢出力 比較\*\*  
( 有意水準 = 0.05, 비대칭분포, 단위 = % )

Population	$\sqrt{\beta_1}$	$\beta_2$	$n$	KS	CM	K	W	$\sqrt{b_1}$	Z	Z*	New Test
Normal (Null case)	0	3.0	20	95.3	95.4	95.1	94.3	94.3	95.1	95.6	80.4
			30	95.7	96.3	96.5	95.7	94.9	95.2	94.4	78.8
Beta(2,1)	-0.57	2.4	20	16.5	22.6	45.7	32.1	11.3	23.0	22.4	54.4
			30	25.1	35.9	63.1	52.4	17.0	36.5	40.8	74.4
Beta(3,2)	-0.29	2.4	20	5.5	5.2	11.9	6.5	2.8	4.4	2.4	15.0
			30	8.2	8.3	16.8	10.7	3.9	6.0	2.2	14.8
Weibull ( $\alpha = 2$ )	0.63	3.3	20	10.2	11.2	12.1	15.6	14.5	15.3	17.4	41.8
			30	13.7	16.0	18.3	23.1	20.7	26.3	25.0	60.4
	6.62	87.7	20	98.4	99.4	99.9	100	98.3	99.8	100	100
			30	100	100	100	100	100	100	100	100
Exponential	2.0	9.0	20	60.1	73.9	83.4	84.2	70.6	82.3	78.7	92.6
			30	77.7	89.8	96.9	96.5	89.4	95.5	95.0	99.4



dholkar에 의해 구해진 Z檢定統計量의 결과치와 비슷함을 알 수 있다. 여기서, 같은 檢定方法을 이용하여 구한 檢出力이 完全히 일치하지 않는 것은 Computer의 기종에 따라서 또는 確率變數의 발생과정에 따라서 차이가 날 수 있기 때문이다. 위의 表1과 表2의 새로운 檢定(New Test)의 檢出力 값은 單純 回歸分析에 의한 正規性檢定の 檢定結果를 정리한 것으로 正規分布일 경우에는 그 檢出力이 다소 떨어지지만 다른 分布들에 있어서의 그 檢出力은 우수함을 알 수 있다.

#### [IV] 結 論

單純 回歸分析에 의하여 正規性檢定을 수행한 결과 表1과 表2에서와 같은 결과를 얻었고 그 檢出力이 사용된 다른 檢定方法들에 비해서 우수하다는 것을 알 수 있었다. 단, 正規分布의 경우에는 正規分布를 正規分布라고 옳게 판단할 確率は 기타 檢定方法들에 있어서의 그것보다는 우수하지 못하나 正規分布가 아닌 것을 正規分布가 아니라고 판단할 確率は 비대칭分布에 있어서는 물론이고 대칭分布의 경우에 있어서도 다른 여러가지 檢定方法들의 檢出力보다도 우수하다는 것을 알 수 있다. 물론 Computer의 기종에 따라서 또는 確率變數의 발생과정에 따라서 차이가 날 수도 있지만 Lin과 Mudholkar의 연구와 본 연구에서의 Z와 Z\*의 결과의 차이

를 고려한다면 檢出力에서 현저한 증가를 가져왔음을 알 수 있다. 그러므로 任意標本의 平均과 分散을 各各 單純 回歸模型의 獨立變數와 從屬變數로 생각하여 標本平均과 標本分散의 回歸關係有·無에 따라 正規性을 檢定하는 것은 우수한 檢定方法이라는 것을 알 수 있다.

本 研究에서는 任意標本의 平均과 分散이 獨立的으로 分布하게 될 때 그 標本은 正規母集團으로부터 抽出된 標本이라는 성질을 이용하여 單純 回歸分析에 의한 正規性檢定을 수행하는데 이 때 任意標本의 分散  $y_i = V_i^{1/3}$ 을 單純 回歸模型의 從屬變數라고 생각하였다. 그러나 標本分散  $V_i$ 을  $y_i = V_i^{1/3}$ 의 形態로 變換하는 方法 이외에도 여러가지 다른 變換도 생각할 수 있는 것이다. 아울러 本 研究에서는 標本數 20과 30에 대해서만 Monte-Carlo Simulation을 수행하여 6種類의 다른 檢定方法들과 그 檢出力을 비교하였으나 그 標本數를 變化시켜 가면서 各 檢定方法들의 檢出力을 비교해 볼 수도 있다.

마지막으로 本 研究가 부족하나마 어떤 檢査資料가 正規分布에 따를 것인가를 檢定하는 正規性 檢定の 문제에 조금이나마 보탬이 되었으면 한다.

끝으로 본 연구의 연구비를 지원해 준 인하대학교 산업과학기술연구소에 감사를 표한다.

#### 要 約

本 研究는 檢査資料가 正規分布에 따를 것인가를 결정하는 正規性檢定을 다룬다.

Lin과 Mudholkar(1980)에 의해 발표된 最近의 檢定方法은 “正規分布일 경우에 平均과 分散이 獨立적으로 分布한다”는 性質을 이용하여 檢定을 수행하였다.

本 研究에서도 이러한 특성을 이용하여 單純回歸分析에 의한 檢定을 시도한다. 이 檢定方法은 任意標本으로부터의 平均과 分散을 各各 單純回歸模型의 獨立變數와 從屬變數로 고려한다. 만약 任意標本의 平均과 分散이 回歸關係가 없다면 獨立인 것으로 간주하여 그 標本은 正規母集團으로

부터 抽出된 標本이라는 結論을 내려준다. 檢出力을 비교할 목적으로 11種類의 대칭·비대칭分布에서 標本の 크기를  $n=20, 30$ 으로 하여 500個의 標本들이 各各 模擬實驗에 사용되었으며 本 研究의 正規性檢定 方法은 Monte-Carlo Simulation 方法을 이용하여 檢出力을 계산함으로써 6個의 다른 檢定方法들과 비교되었다.

正規分布일 경우에는 그 檢出力이 다소 떨어지지만 다른 分布들에 있어서의 그 檢出力은 현저하게 우수하다. 즉 單純回歸分析에 의한 正規性檢定이 다른 檢定方法들에 비해서 우수한 檢出力을 가지는 것으로 나타났다.

## 參 考 文 獻

1. 朴景洙, 信賴度工學, 塔出版社, 1982.
2. 朴聖炫, 回歸分析, 大英社, 1984.
3. Averill M. Law and W. David Kelton, Simulation Modeling and Analysis, McGraw-Hill Book Company, 1982.
4. Babara B. Nelson, "Testing for Normality", Journal of Quality Technology, Vol. 15, No. 3, July 1983.
5. C. Chung Lin and Govind S. Mudholkar, "A Simple Test for Normality against Asymmetric Alternatives", Biometrika, 67, 2, 1980.
6. G. Gorden, System Simulation, 2nd Ed., Prentice-Hall, Inc., 1978.
7. H. Cramér, Mathematical Methods of Statistics, Princeton Univ. Press, 1946.
8. J. D. Gibbons, Nonparametric Statistical Inference, McGraw-Hill, Inc., 1971.
9. Lloyd S. Nelson, "A Simple Test for Normality", Journal of Quality Technology, Vol. 13, No. 1, January 1981.
10. Lloyd S. Nelson, "Combining Statistics from Two Groups and Some Updating Calculations", Journal of Quality Technology, Vol. 10, No. 4, October 1978.
11. Norman L. Johnson and Samuel Kotz, Continuous Univariate Distribution 1, 2 Houghton Mifflin Co., 1970.
12. R.A. Fisher, Statistical Methods for Research Workers, 13th edition. Edinburgh: Oliver and Boyd, 1967.
13. Reuven Y. Rubinstein, Simulation and the Monte-Carlo Method, John Wiley and Sons, 1981.
14. Robert V. Hogg and Allen T. Craig, Introduction to Mathematical Statistics, 4th Ed., Macmillan Publishing Co., Inc., 1978.
15. Ronald H. Randles and Douglas A. Wolfe, Introduction to the Theory of Non-parametric Statistics, John Wiley and Sons, 1979.