연세경제연구 제 ▮권, 제 2호
Yonsei Economic Studies, Vol. ▮, No. 2 (September 1996)

히스토그램에서 급수결정에 대한 시뮬레이션 연구

김상익 · 김영신

히스토그램을 작성할 때 급수를 결정하는 문제에 대한 방법을 시뮬레이션 연구에 의해 고찰하였다. 그 결과 급수 결정에 앞서 표본자료의 왜도(skewness)를 먼저 계산하여 자료분포의 좌우대칭성에 따라 급수결정방법을 달리하여야 하는 것으로 나타났다. 자료의 분포가 좌우대칭형태에 근사하는 경우에는 Scott의 방법이, 비대칭인 경우에는 Freedman-Diaconis [6]방법이 효율적이며, Sturges의 방법도 전반적으로 비교적 우수한 방법으로서 계산의 간편성이라는 장점이 있다.

Ⅰ. 서 론

최근 개인용 컴퓨터의 발전과 더불어 그래픽이나 도표를 이용한 통계적 방법이 주목을 받고 있다. 이러한 현상은 기술통계학의 분야에서 Tukey [15]등 여러 통계학자들의 연구결과로 탐색적 자료분석(exploratory data analysis : EDA)이라는 새로운 주제를 형성하고 있다.

그래픽이나 도표를 이용한 통계분석방법 중 손쉬운 것으로 널리 사용되어 온 방법의 하나는 히스토그램을 이용한 자료의 요약 및 정리 방법이다. 히스토그램은 자료를 몇 개의 계급(class)으로 나누어, 각 계급에 속하는 자료의 수인 돗수(frequency)나 상대돗수(relative frequency)를 막대그래프 형태로 작성하는 방법이다. 히스토그램을 작성할 때 계급의 수인 급수에 따라 자료의 특성과 모집단의 확률밀도함수 형태가 잘 반영되기도 하고 그렇지 않을 수도 있으므로, 급수결정의 문제가 중요하다. 그리고 이러한 급수결정문제는 각 계급의 간격을 등간격으로 똑같이 하는 경우, 자료의 범위를 급수로 나누면 등간격이 결정되므로 급수와 급간의 결정문제는 동일한 문제가 된다.

이러한 급수 혹은 급간 결정에 대한 문제는 일찍이 Sturges [11]에 의해 연구되기 시

건국대학교 상경대학 응용통계학과. 건국대학교 대학원 응용통계학과. 서울시 광진구 모진동 93-1, 133-701.

작하여 초기의 연구로는 Cencov [2], Dixon과 Kronmal [3]의 결과가 있다. 1970년대에 들어 탐색적 자료분석에 대한 연구가 진행되면서 급수결정의 문제는 Tukey [14]등에 의해 재조명되기 시작하여 Doane [4], Beniger와 Robyn [1], Fienberg [5], Wainer와 Thissen [17] 등의 연구가 발표되었다.

특히, 밀도함수추정(density estimation)에서 커넬(kernel)의 윈도우(window)길이를 결정하는 문제와 급간결정의 문제는 유사한 문제가 되어, 밀도함수추정방법을 통한 급간결정방법에 대한 연구도 많이 발표되었다. 이에 대한 결과로는 Scott [8], Freedman과 Diaconis [6], [7], 그리고 Terrell과 Scott [13] 등이 있다.

이러한 급수결정에 대한 많은 연구결과에도 불구하고, 대부분의 통계 패키지를 비롯한 실제자료분석에서는 특별한 방법이 사용되지 않고, 주관적인 판단에 의해 급수를 사전적으로 결정하여 히스토그램을 작성하고 있다. 이러한 사실은 자료의 특징에 따라 급수결정에 대한 여러 방법들간의 효율성에 대한 상호비교연구가 미진한 것도 한 원인으로 판단된다.

본 연구에서는 급수결정에 대한 대표적인 방법들의 기본원리를 고찰하고, 각 방법들을 표본자료의 여러 형태에 따라 시뮬레이션방법에 의해 효율성을 연구하여, 히스토그램을 사용하여 자료분석을 하고자 할 때 하나의 지표를 제시하고자 한다.

다음 제 Ⅱ 절에서는 급수결정방법들간의 효율성의 판단기준으로 사용될 누적평균제곱오차(integrated mean square error : IMSE)를 정의하고, 제 Ⅲ 절에서는 급수결정에 대한 대표적인 방법들을 소개한다. 그리고 시뮬레이션 연구의 방법과 연구결과는 제 Ⅳ 절에 요약되어 있으며 제 Ⅴ 절은 결론으로 이루어져 있다.

Ⅱ. 누적평균제곱오차

확률밀도함수 $f(x)$를 갖는 모집단으로부터 추출된 크기가 n인 표본자료를 $x_1, x_2, \cdots, x_n$ 이라 하고, n개의 표본자료를 k개의 계급 $[t_{j-1}, t_j)$, $j = 1, 2, \cdots, k$, $t_0 < t_1 < \cdots < t_k$로 나누어 히스토그램을 작성한다고 하자. 여기서 $t_0 \le \min\{x_i\}$, $\max\{x_i\} < t_k$이고, $t_{j+1} - t_j = h$로 모든 i에 대하여 동일하다고 하자. 그러면 h는 등간격인 급간이 된다.

그리고 j번째 계급인 $[t_{j-1}, t_j)$의 돗수를 f_j, $j = 1, 2, \cdots, k$라 하면 히스토그램에서 j번째 계급에서의 $f(x)$의 추정량은 다음과 같다.

$$\hat{f}(x) = \begin{cases} \dfrac{f_j}{nh}, & t_{j-1} \le x < t_j, \quad j = 1, 2, \cdots, k \\ 0, & elsewhere \end{cases} \tag{1}$$

모집단분포의 j번째 계급의 확률면적인 P_j는 Talyor 전개를 이용하면 식 (2)와 같다.

$$\begin{aligned} p_j &= \int_{t_{j-1}}^{t_{j-1}+h} f(y)\,dy \\ &= \int_{t_{j-1}}^{t_{j-1}+h} \{f(x) + f'(x)(y-x) + O(h^2)\}dy \\ &= h \cdot f(x) + \frac{1}{2} f'(x)\{h^2 - 2h(x-t_{j-1})\} + O(h^3) \end{aligned} \tag{2}$$

또한, 식 (1)에서 f_j는 n과 식 (2)의 P_j를 모수로 갖는 이항분포에 따르는 확률변수이므로 $\hat{f}(x)$의 평균과 분산은 다음과 같다.

$$\begin{aligned} E[\hat{f}(x)] &= \frac{1}{nh} E[f_j] \\ &= f(x) + \frac{h}{2} f'(x) - f'(x)\,(x-t_{j-1}) + O(h^2) \\ Var[\hat{f}(x)] &= \frac{f(x)}{nh} + O\left(\frac{1}{n}\right) \end{aligned} \tag{3}$$

따라서, $\hat{f}(x)$를 $f(x)$의 추정량으로 사용했을 때, 한 점 x에서 평균제곱오차(mean square error : MSE)는 식 (4)가 된다.

$$\begin{aligned} MSE &= E[(\hat{f}(x) - f(x))^2] \\ &= Var[\hat{f}(x)] + [E[\hat{f}(x)] - f(x)]^2 \\ &= \frac{f(x)}{nh} + \frac{h^2}{4}[f'(x)]^2 + [f'(x)]^2(x - t_{j-1})^2 \\ &\quad - h[f'(x)]^2(x - t_{j-1}) + O\left(\frac{1}{n} + h^3\right) \end{aligned} \tag{4}$$

그리고 식 (4)의 MSE를 적분하면 전 구간에서의 평균제곱오차인 $IMSE$를 구할 수 있다.

$$\begin{aligned} IMSE &= \int (MSE)dx \\ &= \frac{1}{nh} + \frac{h^2}{4}\int f'(x)^2 dx + \int f'(x)^2(x - t_{j-1})^2 dx \\ &\quad - h\int f'(x)^2(x - t_{j-1})\,dx + O\left(\frac{1}{n} + h^3\right) \end{aligned} \tag{5}$$

식 (5)에서 t_{j-1}도 x에 의존하므로, 세 번째 항과 네 번째 항의 t_{j-1}을 변수변환하여 $IMSE$를 재표현하면 다음과 같이 된다.[1)]

1) 변수변환방법과 $IMSE$의 유도과정에 대한 자세한 내용은 Scott [8]를 참조.

$$IMSE = \frac{1}{nh} + \frac{h^2}{12}\int f'(x)^2 dx + O\left(\frac{1}{n} + h^3\right) \quad (6)$$

본 연구에서는 식 (6)의 *IMSE*를 평가기준으로 하여 급수결정방법의 효율성을 비교하고자 한다. 따라서 식 (6)을 최소화시키는 h가 최적급간이 되며, 최적급수 k는 자료의 범위를 최적급수로 나누어서 구하게 된다.[2)]

Ⅲ. 급수결정방법

1. Sturges의 방법

히스토그램에서 급수결정방법으로 Sturges [11]는 자료의 수 n이 2의 power형태, 즉 $2^m = n$인 경우를 가정하여 이항계수(binomial coefficient)를 이용한 방법을 제시하였다.

예를 들어, $n = 16$인 경우 $2^4 = 16$이므로, $(a+b)^4$을 전개했을 경우 이항계수들은 1, 4, 6, 4, 1과 같이 다섯 개가 되어 급수는 $k = 5$로 결정된다. 이러한 경우 첫 번째 계급의 자료돗수는 1, 두 번째 계급의 돗수는 4, 세 번째 돗수는 6이 된다는 것을 의미한다.

일반적으로 n이 2의 power형태인 경우, 이항계수의 수인 급수와 n의 관계는 $2^{k-1} = n$이 되어 급수는 다음과 같이 결정된다.

$$\begin{aligned} k &= 1 + \log_2 n \\ &= 1 + 3.3\log n \end{aligned} \quad (7)^{3)}$$

따라서, Sturges의 방법을 이항계수의 값에서 볼 수 있는 바와 같이 모집단의 분포는 좌우대칭형태를 기본적으로 가정하고 있음을 알 수 있다.

2. Dixon과 Kronmal의 방법

Dixon과 Kronmal [3]은 급수 k를 결정하는데 있어서 반복계산법의 접근방법을 응용하였다. 이 방법에 의하면, 다음의 식 (8)에 의해 급수 k를 결정하여 급간을 구하는 방법을 사용하였을 때 최적급간을 구하는 반복계산과정이 가장 짧아지게 된다.[4)]

2) 급간을 정하는 경우 실제 관측된 자료가 두 개의 계급에 겹치지 않도록 하기 위해 계급범위를 측정단위보다 작은 단위를 가지고 조정하기도 한다. 그리고 k가 정수가 아닌 경우 k값에 가까운 정수값을 선택하여 급수로 한다.

3) 여기서 log는 밑이 10인 상용대수를 의미한다.

4) 최적급간에 대한 정의와, 최적급간을 구하는 반복계산과정에 대한 자세한 내용은 Dixon과 Kronmal [3]을 참조.

$$k = 10 \cdot \log n \tag{8}$$

따라서, Dixon과 Kronmal은 식 (8)에 의해 급수를 결정할 것을 제안하였다.

3. Velleman의 방법

Sturges의 방법이나 Dixon-Kronmal의 방법에서, 급수는 식 (7)과 식 (8)에서 보는 바와 같이 자료의 수 n의 상용대수값인 $\log n$에 비례하게 된다. 따라서, 식 (7)의 비례상수는 3.3이며 식 (8)의 비례상수는 10으로서 두 방법은 유사한 방법이면서도 k값은 무척 다르게 된다. 특히, $n < 30$인 소표본에서 이러한 현상은 두드러지게 된다.

Velleman [16]은 이러한 점을 중시하여 소표본을 중심으로 급수결정에 대한 방법을 고찰하여, 자료의 수 n이 50보다 작은 경우 급수를 다음과 같이 n의 양의 제곱근에 비례하는 형태로 결정할 것을 제안하였다.

$$k = 2 \cdot \sqrt{n} \tag{9}$$

자료의 수가 $n = 100$인 경우 Sturges, Dixon-Kronmal의 방법에 의한 급수는 각각 $k = 7.6$과 $k = 20$이 되며, Velleman의 방법에 의하면 $k = 20$으로 Dixon-Kronmal과 동일하나 Sturges의 방법과 비교하면 크게 된다. 그러나, $n < 100$인 경우 식 (9)에 의해 구한 k는 항상 식 (7)과 식 (8)에 의해 구한 값 사이에 있게 된다.

4. Scott의 방법

Scott [8]의 방법에 의하면 식 (6)의 *IMSE*가 최소가 되는 최적급간 h를 먼저 구하며 급수를 결정하게 된다. 식 (6)을 h에 대하여 미분하여 *IMSE*가 최소가 되는 최적급수 h를 구하면 다음과 같다.

$$h = [6 / \int f'(x)^2 dx]^{1/3} n^{-1/3} \tag{10}$$

모집단의 확률밀도함수 $f(x)$가 정규분포인 경우 식 (10)은 식 (11)과 같이 된다.

$$h = 2.3^{1/3} \pi^{1/6} \sigma n^{-1/3} \tag{11}$$

Scott는 식 (11)의 h를 최적급간으로 사용할 것을 제안하였다. 따라서, 자료의 표준편차를 S라 하고 범위를 R이라 하면 급수는 다음과 같이 결정된다.

$$k = R/(3.49\, Sn^{-1/3}) \tag{12}$$

5. Freedman과 Diaconis방법

최적급간을 구하는데 있어서 Freedman과 Diaconis [6]는 *IMSE* 대신 다음의 식 (13)을 만족하는 h를 최적급간으로 정의하였다.

$$\min_{h} \max_{x} |\hat{f}(x) - f(x)| \tag{13}$$

그리고 식 (13)을 만족하는 최적급간은 다음과 같이 된다.

$$h = c(f) \cdot \left(\frac{\log_e n}{n}\right)^{1/3} \tag{14}$$

여기서 $c(f)$는 확률밀도함수 $f(x)$에 의존하는 상수로서, $f(x)$가 정규분포확률밀도함수인 경우 $c(f)$는 1.66σ가 된다. 따라서, Freedman-Diaconis [6]방법에 의하면 급수는 다음과 같이 결정될 수 있다.

$$k = R/[1.66S(\log_e n/n)^{1/3}] \tag{15}$$

Freedman-Diaconis [7]는 최적급간을 다르게 정의하여, 식 (15)에서 S 대신 자료의 사분위편차(interquartile range) V를 사용하면 식 (16)과 같이 급수가 결정됨을 보였다.

$$k = R/\{(2V)n^{-1/3}\} \tag{16}$$

Scott와 Freedman-Diaconis [6], [7]의 방법들간의 차이는 최적급간 h를 구하는 방법에 기인하며, 세 가지 방법 모두에서 급수는 n의 세제곱근에 비례한다는 사실을 알 수 있다. 이와 같이 $\sqrt[3]{n}$에 비례하는 형태로 급수를 결정하는 방법은 일찍이 Cencov [2]에 의해 제안되었다.

〈표 1〉은 모집단분포가 표준정규분포인 경우 자료의 수(표본크기) n의 변화에 따른 세 가지 방법에 의한 최적급간의 변화를 보여 주고 있다. 〈표 1〉에서 보는 바와 같이 전반적으로 Freedman-Diaconis [6], [7]방법이 Scott의 방법에서보다 최적급간이 짧음을 알 수 있으며(따라서, 급수는 크게 된다), 특히 Scott의 방법과 Freedman-Diaconis [7] 방법에서 최적급간의 비는 3.49/2.698 = 1.3이 되어 Scott의 방법에서 급수는 30% 정도 작게 된다.

〈표 1〉 표본크기변화에 따른 최적급간의 변화

30	1.1231	0.8034	0.8683
50	0.9473	0.7100	0.7324
70	0.8468	0.6524	0.6547
100	0.7519	0.5950	0.5813
150	0.6568	0.5346	0.5078
200	0.5968	0.4949	0.4614
300	0.5213	0.4431	0.4030
400	0.4737	0.4092	0.3662
500	0.4397	0.3845	0.3399

주 : 표준화 정규분포를 가정한 경우로서 $S=1$, $V=1.349$가 된다.

Ⅳ. 몬테-칼로 연구

본절에서는 앞에서 고찰한 급수결정방법들에 대하여 시뮬레이션의 몬테-칼로법(Monte-Calo method)에 의해 효율성을 비교하고자 한다. 히스토그램 작성시 중요한 고려사항은 분포의 좌우대칭성이므로, 본 연구에서는 모집단 확률밀도함수로서 좌우대칭형태는 정규분포 그리고 비대칭형태로는 지수분포, 감마분포, 베타분포를 선정하였다. 또한, 각 확률밀도함수로부터 표본자료를 난수생성법에 의해 추출하였다. 또한, 자료의 크기도 여러 형태로 변화시키면서 고찰하였으며 급수결정방법의 효율성 판단기준으로는 제 Ⅱ 절에서 정의한 *IMSE*를 사용하였다.

1. 연산방법

정규분포인 경우 표준정규분포를 선택하였으며, 지수분포인 경우 모수를 $\theta = 0.5, 1, 2$로 변화시켜 세 가지 확률밀도함수를 선정하였고, 감마분포인 경우 비대칭성을 변화시키기 위해 두 개의 모수 α, β를 $(\alpha = 1, \beta = 1)$, $(\alpha = 1, \beta = 4)$, $(\alpha = 1.5, \beta = 2.0)$, $(\alpha = 2.0, \beta = 4.0)$인 경우 네 가지를 선정하였다. 베타분포인 경우에는 두 개의 모수 α, β를 $(\alpha = 1, \beta = 3)$, $(\alpha = 1.5, \beta = 28.5)$, $(\alpha = 3, \beta = 1)$의 세 가지를 선택하여 비대칭도를 변화시켰다. 몬테-칼로 연구에 사용된 프로그램은 포트란 서브루틴 프로그램인 *IMSL*을 사용하였다.

자료크기는 각각의 확률밀도함수로부터 $n = 30, 50, 70, 100, 150, 200, 300, 400, 500$

까지 변화시켰으며, 표준정규분포와 지수분포의 경우에는 각각 1,000회씩 반복하여 난수를 생성하여[5] 1,000개의 *ISE*(integrated square error)를 계산한 후, 그 평균인 *IMSE*를 산출하였다. 감마분포와 베타분포의 경우에는 각각 500회씩 난수를 반복추출하였다. 식 (17)로 정의되는 *ISE*계산방법은 다음과 같다.

$$ISE = \int \{\hat{f}(x) - f(x)\}^2 dx \tag{17}$$

표본자료를 n개 생성한 후 순서통계량으로 재배열한다. 그리고 $t_0 = \min\{x_i\}$로 하고 각 급수결정방법에 따라 급간 h를 구한 후 k개의 계급 $[t_j, t_j + h)$, $j = 0, 1, 2, \cdots, k$를 구한다. 여기서 $\max\{x_i\}$는 마지막 계급에 속하게 하면서 각 계급에 속한 자료의 수 f_j, $j = 1, 2, \cdots, k$를 구하여 식 (1)에 의해 각 계급에서의 $\hat{f}(x)$를 추정한다.

추정된 $\hat{f}(x)$와 $f(x)$를 이용하여 각 계급에서의 *ISE*값을 적분을 통하여 계산한 후[6] 계산된 계급의 *ISE*값을 더하여 전 구간에서의 *ISE*값인 식 (17)을 구한다.[7]

2. 결과분석

표준정규분포의 경우 계산된 *IMSE*값은 〈표 2〉, 지수분포의 경우에는 〈표 3〉, 감마분포의 경우와 베타분포의 경우에는 각각 〈표 4〉와 〈표 5〉에 정리되어 있다.

〈표 2〉에서 보는 바와 같이 표준정규분포인 경우, 표본크기변화에 따라 전반적으로 Scott의 방법이 가장 우수한 방법으로 평가되며, 이는 *IMSE*를 평가기준으로 사용했을 때 당연한 결과라고 할 수 있다. 또한, Sturges의 방법도 다른 방법과 비교해 볼 때 우수한 방법으로 평가되며, Dixon-Kronmal방법, Velleman의 방법은 비효율적인 방법으로 나타난다.

지수분포인 경우 왜도(skewness)는 2로서 비대칭성이 비교적 심한 경우이다. 〈표 3〉을 고찰해 보면 Freedman-Diaconis [6]방법이 전반적으로 우수하고, 표본크기가 큰 경우에는 Dixon-Kronmal의 방법도 우수한 것으로 판단된다. 또한, Sturges의 방법도 비교적 우수한 방법임을 알 수 있다.

왜도의 값이 $2/\sqrt{\alpha}$로 α값이 커질수록 대칭분포에 근접하는 감마분포의 경우, 〈표 4〉에서 α값이 큰 ③과 ④에서 $n < 100$인 경우 Scott의 방법이 가장 우수하며 표본크기가 큰 경우에는 Freedman-Diaconis [6]방법이 가장 효율적이나, Scott의 방법과 거의 비슷

5) 난수생성에 사용된 서브루틴은 표준정규분포인 경우 RNOF, 지수분포인 경우 RNEXP, 감마분포인 경우 RNGAM, 베타분포인 경우에는 RNBET이다.

6) *ISE*를 구할 때 사용된 적분 서브루틴은 QDAGS와 QDAGI이다.

7) *ISE*를 구할 때 $(-\infty, t_0)$, $[t_k+h, \infty]$의 구간도 고려해 계산하였다.

함을 알 수 있다.

전반적으로 Freedman-Diaconis [6]방법이 효율적인 방법으로 평가되며 표본크기가 큰 경우 Dixon-Kronmal의 방법도 우수하다고 하겠으며, 전반적으로 Sturges의 방법도 비교적 괜찮다고 할 수 있다.

〈표 5〉의 베타분포인 경우를 살펴보면 Scott의 방법과 Freedman-Diaconis [6]의 방법이 우수함을 알 수 있다. 베타분포에서는 두 개의 모수값이 같아짐에 따라 왜도가 0에 근사해 가는 대칭분포를 이루게 된다. 이러한 사실은 α와 β값이 비슷한 ①과 ③의 경우에는 Scott의 방법이 그리고 왜도가 큰 ②의 경우에는 Freedman−Diaconis방법이 효율적인 방법으로 나타나는 점에도 반영이 되고 있다.

종합적으로 살펴볼 때, 모집단분포가 좌우대칭형에 가까운 경우에는 Scott의 방법이 우수한 방법이 되며, 비대칭도가 심할수록 Freedman-Diaconis [6]의 방법이 보다 효율적인 방법이 된다. 그리고 Sturges의 방법도 앞의 두 가지 방법에 비해 전반적으로 효율성은 조금 나쁘지만 비교적 괜찮은 방법으로 평가된다. 또한, 소표본인 경우에도 Velleman의 방법은 효율적인 방법이 되지 못하며 Dixon-Kronmal방법은 대표본인 경우 좌우대칭에 가까운 비대칭분포에서는 우수한 방법이 된다. 그리고 Freedman-Diaconis [7]방법은 우수하지 않은 방법으로 나타난다.

〈표 2〉 표준화정규분포일 때 추정되어진 *IMSE*값

30	0.04960	0.11604	0.08414	0.03586*	0.04251	0.07256
50	0.03329	0.07595	0.06081	0.02788*	0.03109	0.05619
70	0.02264	0.04988	0.04483	0.02222*	0.02264	0.03784
100	0.01535	0.03504	0.03504	0.01531*	0.01845	0.03042
150	0.01294*	0.02758	0.03004	0.01321	0.01424	0.02611
200	0.01156	0.02015	0.02542	0.01155*	0.01213	0.02150
300	0.00888	0.01414	0.01862	0.00815*	0.00853	0.01541
400	0.00754	0.01045	0.01483	0.00711*	0.00740	0.01279
500	0.00767	0.00950	0.01497	0.00623*	0.00769	0.01481

주 : *는 추정되어진 *IMSE*의 최소값임.

〈표 3〉 지수분포일 때 추정되어진 *IMSE*값

30	①	0.18112	0.30888	0.24836	0.18902	0.17202^{+}	0.24902
	②	0.09056	0.15444	0.12418	0.09451	0.08601^{+}	0.12451
	③	0.04528	0.07722	0.06209	0.04726	0.04300^{+}	0.06226
50	①	0.12534	0.19841	0.16878	0.13438	0.12260^{+}	0.17156
	②	0.06267	0.09920	0.08439	0.06719	0.06130^{+}	0.08578
	③	0.03134	0.04960	0.04220	0.03360	0.03065^{+}	0.04289
70	①	0.09641	0.13524	0.12518	0.10380	0.09304^{+}	0.12742
	②	0.04820	0.06762	0.06259	0.05190	0.04652^{+}	0.06371
	③	0.02410	0.03381	0.03129	0.02595	0.02326^{+}	0.03186
100	①	0.07873	0.09715	0.09715	0.08097	0.07268^{+}	0.10026
	②	0.03937	0.04858	0.04858	0.04049	0.03634^{+}	0.05013
	③	0.01968	0.02429	0.02429	0.02024	0.01817^{+}	0.02507
150	①	0.06663	0.06595	0.07153	0.06296	0.05605^{+}	0.07595
	②	0.03331	0.03298	0.03576	0.03148	0.02802^{+}	0.03797
	③	0.01666	0.01649	0.01788	0.01574	0.01401^{+}	0.01899
200	①	0.05891	0.05080	0.05828	0.05040	0.04494^{+}	0.06123
	②	0.02945	0.02540	0.02914	0.02524	0.02247^{+}	0.03062
	③	0.01473	0.01270	0.01457	0.01262	0.01124^{+}	0.01531
300	①	0.05134	0.03563	0.04409	0.03794	0.03429^{+}	0.04608
	②	0.02567	0.01782	0.02205	0.01897	0.01714^{+}	0.02304
	③	0.01284	0.00891	0.01102	0.00949	0.00857^{+}	0.01152
400	①	0.04971	0.02853^{+}	0.03750	0.03143	0.02871	0.03890
	②	0.02485	0.01426^{+}	0.01875	0.01571	0.01436	0.01945
	③	0.01243	0.00713^{+}	0.00937	0.00786	0.00718	0.00973
500	①	0.04513	0.02412^{+}	0.03210	0.02703	0.02489	0.03281
	②	0.02256	0.01206^{+}	0.01605	0.01352	0.01244	0.01641
	③	0.01128	0.00603^{+}	0.00802	0.00676	0.00622	0.00820

주 : ①은 $\theta = 0.5$, ②는 $\theta = 1.0$, ③은 $\theta = 2.0$인 경우임.

$^{+}$는 추정되어진 *IMSE*의 최소값임.

〈표 4〉 감마분포일 때 추정되어진 *IMSE*값

30	①	0.08835	0.15131	0.12392	0.09202	0.08378^{+}	0.12207
	②	0.02209	0.03783	0.03098	0.02283	0.02099^{+}	0.03052
	③	0.02423	0.05480	0.04001	0.01836^{+}	0.02163	0.03926
	④	0.00995	0.02277	0.01682	0.00742^{+}	0.00938	0.01546
50	①	0.06400	0.10172	0.08639	0.06815	0.06270^{+}	0.08778
	②	0.01600	0.02543	0.02160	0.01698	0.01560^{+}	0.02194
	③	0.01585	0.03373	0.02811	0.01283^{+}	0.01525	0.02815
	④	0.00647	0.01401	0.01191	0.00563^{+}	0.00647	0.01122
70	①	0.04870	0.06857	0.06241	0.05201	0.04734^{+}	0.06491
	②	0.01218	0.01714	0.01560	0.01298	0.01185^{+}	0.01623
	③	0.01173	0.02502	0.02280	0.01062^{+}	0.01207	0.02210
	④	0.00533	0.01006	0.00924	0.00496^{+}	0.00526	0.00868
100	①	0.03948	0.04869	0.04869	0.04031	0.03627^{+}	0.04916
	②	0.00987	0.01217	0.01217	0.01006	0.00905^{+}	0.01229
	③	0.00922	0.01746	0.01746	0.00881^{+}	0.00966	0.01705
	④	0.00440	0.00748	0.00748	0.00426^{+}	0.00445	0.00706
150	①	0.03369	0.03273	0.03561	0.03149	0.02804^{+}	0.03745
	②	0.00842	0.00818	0.00890	0.00785	0.00701^{+}	0.00936
	③	0.00726	0.01201	0.01319	0.00714^{+}	0.00759	0.01295
	④	0.00368	0.00501	0.00565	0.00366	0.00358^{+}	0.00536
200	①	0.02974	0.02525	0.02842	0.02532	0.02255^{+}	0.03042
	②	0.07443	0.00631	0.00710	0.00631	0.00562^{+}	0.00760
	③	0.00641^{+}	0.00936	0.01082	0.00647	0.00659	0.01062
	④	0.00333	0.00398	0.00473	0.00314	0.00301^{+}	0.00445
300	①	0.02566	0.01799	0.02194	0.01892	0.01704^{+}	0.02285
	②	0.00642	0.00450	0.00549	0.00472	0.00426^{+}	0.00571
	③	0.00559	0.00673	0.00836	0.00545	0.00541^{+}	0.00819
	④	0.00306	0.00276	0.00363	0.00263	0.00251^{+}	0.00336
400	①	0.02465	0.01435	0.01899	0.01556	0.01424^{+}	0.01947
	②	0.00616	0.00359	0.00475	0.00389	0.00356^{+}	0.00487
	③	0.00527	0.00547	0.00719	0.00492	0.00487^{+}	0.00684
	④	0.00289	0.00217	0.00295	0.00226	0.00211^{+}	0.00276
500	①	0.02211	0.01192^{+}	0.01608	0.01345	0.01234	0.01606
	②	0.00553	0.00308	0.00402	0.00336	0.00298^{+}	0.00401
	③	0.00502	0.00457	0.00614	0.00440	0.00437^{+}	0.00584
	④	0.00285	0.00189	0.00259	0.00207	0.00187^{+}	0.00240

주 : ①은 ($\alpha = 1.0, \beta = 1.0$), ②는 ($\alpha = 1.0, \beta = 4.0$), ③은 ($\alpha = 1.5, \beta = 2.0$), ④는 ($\alpha = 2.0, \beta = 4.0$)인 경우임.
$^{+}$는 추정되어진 *IMSE*의 최소값임.

〈표 5〉 베타분포일 때 추정되어진 *IMSE*값

30	①	0.03388	0.07512	0.05757	0.02393^{+}	0.02703	0.04416
	②	0.04705	0.09996	0.09830	0.04857	0.04439^{+}	0.08074
	③	0.02741	0.06984	0.05560	0.02852^{+}	0.02901	0.04212
50	①	0.02271	0.05146	0.04281	0.01770^{+}	0.01994	0.03213
	②	0.05897	0.05135	0.06651	0.06621	0.03896^{+}	0.06642
	③	0.02295	0.05097	0.04136	0.02073^{+}	0.02401	0.03015
70	①	0.00622	0.01028	0.01252	0.00483^{+}	0.00612	0.00756
	②	0.04381	0.05118	0.04670	0.04523	0.02823^{+}	0.07793
	③	0.01216	0.01953	0.01954	0.00432^{+}	0.01087	0.02497
100	①	0.00306	0.00904	0.00904	0.00266^{+}	0.00267	0.00768
	②	0.02572	0.04418	0.04418	0.02924	0.02509^{+}	0.03756
	③	0.01437^{+}	0.02554	0.02554	0.01649	0.01604	0.01907
150	①	0.00843	0.01802	0.02067	0.00764^{+}	0.00842	0.01403
	②	0.04275	0.07512	0.08407	0.04396	0.04166^{+}	0.07623
	③	0.01832	0.01641	0.02088	0.01312^{+}	0.01431	0.01545
200	①	0.00663	0.01415	0.01707	0.00615^{+}	0.00671	0.01139
	②	0.03802^{+}	0.05957	0.07057	0.03805	0.03927	0.06490
	③	0.01221	0.01424	0.01944	0.01070^{+}	0.01154	0.01317
300	①	0.00483	0.01022	0.01376	0.00459^{+}	0.00492	0.00855
	②	0.03297	0.04134	0.05532	0.03256	0.03226^{+}	0.04876
	③	0.01483	0.00990	0.01411	0.00982^{+}	0.01072	0.01054
400	①	0.00398	0.00801	0.01197	0.00393^{+}	0.00403	0.00725
	②	0.03034	0.03288	0.04529	0.02838^{+}	0.02866	0.04159
	③	0.00860	0.00949	0.01193	0.00848^{+}	0.00850	0.00914
500	①	0.00328^{+}	0.00655	0.01052	0.00330	0.00347	0.00598
	②	0.02935	0.02805	0.03904	0.02615	0.02516^{+}	0.03494
	③	0.00273^{+}	0.00630	0.01070	0.00861	0.00828	0.00807

주 : ①은 ($\alpha = 1.0, \beta = 3.0$), ②는 ($\alpha = 1.0, \beta = 3.0$), ③은 ($\alpha = 1.0, \beta = 3.0$)인 경우임.
$^{+}$는 추정되어진 *IMSE*의 최소값임.

Ⅴ. 맺음말

히스토그램을 작성할 때 우리가 목적하고자 하는 바는 시각적 효과와 함께 모집단 분포의 대략적 형태를 파악함과 아울러 자료의 여러 면모를 탐색하고자 하는데에 있다.

이러한 히스토그램기법은 기술통계학이라는 통계적 영역에서 뿐만 아니라, 생산관리 및 통계적 품질관리 등 통계학의 여러 응용부분에서 가장 손쉬운 방법이면서도 효과적인 방법으로 사용되고 있다. 특히, 통계적 품질관리 분야에서는 일곱 가지 기초수법(seven tools of quality control)으로 먼저 히스토그램의 사용을 강조하고 있다.

이러한 히스토그램을 사용할 때, 중요한 문제인 급수 혹은 급간을 결정하는 방법에 대하여 본 연구에서는 시뮬레이션 연구를 통하여 효율성을 비교하였다. 그 결과 우리는 다음과 같은 현실적인 지표를 유도할 수 있다.

먼저 히스토그램을 작성하기에 앞서, 표본자료의 왜도를 계산하여 왜도값이 0에 가까운 분포로서 좌우대칭성을 보이는 경우 Scott의 방법을 이용하여 급수(혹은 급간)를 결정하는 것이 바람직하며, 왜도값이 분포의 비대칭성을 보여 주는 경우 Freedman-Diaconis [6]방법을 사용하는 것이 바람직하다고 하겠다. Sturges방법을 사용하는 것도 전반적으로 Scott의 방법이나 Freedman-Diaconis [6]방법보다 약간 효율은 떨어지지만 계산의 간편함을 고려하면 추천할 만한 방법으로 판단된다.

그러나, 이러한 시뮬레이션 연구 결과는 다양한 확률밀도함수 중, 대칭형태로는 표준정규분포, 비대칭형태로는 지수분포, 감마분포, 베타분포, 심한 비대칭형태로는 베타분포를 선정하여 제한된 범위내에서 유도된 결론이라는 단점을 갖고 있다. 또한, 효율성을 판단하는 기준도 *IMSE*를 사용한 경우에 국한되어 있다. 따라서, 다양한 형태의 분포와 다른 판단기준에 의한 상호비교는 앞으로 계속 연구되어야 할 과제로 남는다.

❖참 고 문 헌❖

1. Beniger, J. R and Robyn, D. L., "Quantitative Graphics in Statistics: A Brief History," *The American Statistician*, 32, 1978, pp. 1～11.
2. Cencov, N. N., "Evaluation of an Unknown Distribution Density from Observations," *Soviet Mathematics*, 3, 1962, pp. 1559～1562.
3. Dixon, W. J. and Kronmal, R. A., "The Choice of Origin and Scale for Graphs," *Journal of the Association for Computing Machinery*, 12, 1965, pp. 259～261.

4. Donae, D. P., "Aesthetic Frequency Classification," *The American Statistician*, 30, 1976, pp. 181~183.

5. Fienberg, S. E., "Graphical Methods in Statistics," *The American Statistician*, 30, 1979, pp. 165~178.

6. Freedman, D. and Diaconis, P., "On the Maximum Deviation between the Histogram and the Underlying Density," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 58, 1981, pp. 139~167.

7. _____ , "On the Histogram as a Density Estimator : L2 Theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 1981, pp. 453~476.

8. Scoott, D. W., "On Optimal and Data-based Histograms," *Biometrika*, 66, 1979, pp. 605~610.

9. _____ , and Factor, L. E., "Monte Carlo Study of Three Data-based Nonparametric Density Estimators," *Journal of the American Statistical Association*, 76, 1981, pp. 9~15.

10. Silverman, B. W., "Choosing a Window Width When Estimating a Density," *Biometrika*, 65, 1978, pp. 1~11.

11. Sturges, H. A., "The Choice of A Class Interval," *Journal of the American Statistical Association*, 21, 1926, pp. 65~66.

12. Tarter, M. E. and Kronmal, R. A., "An Introduction to the Implementation and Theory of Nonparametric Density Estimation," *The American Statistician*, 30, 1967, pp. 105~112.

13. Terrell, G. R. and Scott., D. W., "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, 1985, pp. 209~213.

14. Tukey, J. W., "Some Graphic and Semigraphic Displays," In T. A. Bancroft (ed.), *Statistical Papers in Honor of G. W. Snedecor*. Ames, 1972, IA : Iowa State University Press.

15. _____ , "Exploratory Data Analysis," 1977, New York : Addison-Wesley.

16. Velleman, P. F., "Interactive Computing for Exploratory Data Analysis I: Display Algorithms," *1975 Proceedings of the Statistical Computing Section*, 1976, Washington, D. C.: American Statistical Association.

17. Wainer, H. and Thissen, D., "Graphical Data Analysis," *Annual Review of Psychology*, 32, 1981, pp. 191~241.