1995 Proceedings of the Autumn Conference, Korean Statistical Society

# 최적 히스토그램의 구현

김 승 환, 전 홍 석[1)]

## 요 약

히스토그램은 분포의 모양에 대한 종합적인 정보를 제공하는 그림기법으로 통계분석에서 많이 사용하는 기본적인 그림기법이다. 하지만, 히스토그램 작성에 있어 계급의 수나, 계급의 폭의 결정이 주관적인 판단에 의존하기 때문에 연구자들은 적절한 히스토그램을 얻기위해서 많은 시행착오를 경험하게 된다. 본 논문에서는 지금까지 알려진 히스토그램의 구현 산법을 범용 통계 소프트웨어인 SAS와 윈도우용 미니탭(Version 9.2) 에서 구현하고 이러한 산법이 최적 히스토그램을 구현할 수 있는가를 연구하였다. 이로써, 사용자는 SAS와 미니탭에서 기본적으로 제공하는 히스토그램 이외에도 여러가지의 다양한 산법에 의한 히스토그램을 간단한 명령어를 이용하여 얻을 수 있다.

주요용어: 차이함수(Discrepancy), AMISE, Kernel 밀도함수 추정법

## 1. 고전적 히스토그램

히스토그램이란 자료로 부터 모집단의 분포를 추측하는데 사용되는 가장 기본적인 그림기법으로 널리 알려져 있고 또한 가장 많이 사용되는 방법이다. 히스토그램을 통하여 모집단의 분포를 적절히 추측하기 위해서는 연구자가 적절한 표본의 갯수, 구간, 계급의 간격을 설정하여야 한다.

이 부분에 대해 소개되어 있는 많은 통계학 책에서는 표본 $n$개가 주어졌을때 히스토그램을 작성하는 방법에 대해 아래와 같이 기술하고 있다.

1. 자료의 최대값과 최소값을 찾아 자료의 범위를 구한다.
2. 자료의 크기를 고려하여 5개에서 15개 정도의 계급의 갯수를 정하고, 적당한 계급구간의 폭을 정한다. 일반적으로, (자료의 범위/계급의 갯수) 보다 조금 큰 숫자를 계급의 폭으로 정한다.
3. 2.에서 정한 계급의 갯수와 폭을 이용하여 서로 중복되지 않는 동일한 계급구간의 폭을 정한다.
4. 각 계급에 속하는 도수와 상대도수를 구한다.
5. 각 계급구간에 도수 혹은 상대도수의 높이로 사각형을 그린다.

[1)] (402-751) 인천광역시 남구 용현동 253, 인하대학교 이과대학 통계학과

위의 방법에서 2.번 부분은 연구자의 주관적인 판단에 의해 결정되기 때문에 적절한 히스토그램을 얻는데 장애가 될 수 있다. 이에 대해 일부의 기초통계학 책에서는 Sturges(1926)에 의해 제안된 식 (1)을 소개하기도 한다.

$$\text{Sturges의 식: } c \approx 1+3.32\log_{10} n,\ n\text{은 표본의 크기} \tag{1}$$

또한, Larson(1975)는 Sturges의 식과 유사한 형태의 아래의 식을 제안하기도 했었다.

$$\text{Larson의 식: } c \approx 1+2.2\log_{10} n,\ n\text{은 표본의 크기} \tag{2}$$

(1)과 (2)식은 표본의 크기에 따라 적절한 구간의 수를 계산하는 식으로 자료분석자에게 어느 정도 기초적인 정보를 제공하고 있지만 식의 구조상 다양한 자료변화에 민감할 수는 없을 것이다.

Scott(1979)는 모집단의 분포가 정규분포를 한다는 가정하에 A-MISE(Asymptotic Mean Integrated Squared Error)를 최소하는 기준으로 유도된 식 (3)을 제안했다.

$$\text{Scott 의 식:}\quad c = \text{Nearest Int}\left(\frac{Ln^{1/3}}{3.49s}\right) \tag{3}$$

$L$: 자료의 범위, $s$: 표준편차, $n$: 표본의 크기

SAS에서는 Scott에 의해 제안된 식 (3)을 사용하고 있다.

## 2. 차이함수(Discrepancy)

히스토그램은 비모수적 밀도함수 추정방법으로 고전적 방법에 속한다. 히스토그램은 밀도함수 추정에서 널리 사용되는 Kernel 밀도함수 추정법과 비교해 볼 때, 계산면에서 간단하지만 통계적으로 유효(efficient)하지 않은 방법이다. [5]. 예를 들어, 모집단의 밀도함수가 평활(smooth)한 모양일 경우, 히스토그램은 사각형의 모양을 가지므로 모집단의 밀도함수에 잘 근사(well approximate)하지 않게 되고, 구간의 시작점과 끝점, 구간 폭의 변화에 따라 그 모양변화가 심하게 나타난다. 이러한 점을 개선하기 위해 히스토그램을 단순히 주어진 구간에서 도수를 구하는 관점이 아닌 사각형의 밀도로 모집단의 밀도를 추정하는 관점으로 생각할 수 있다. 만약, 히스토그램의 구간 시작점과 끝점을 알고 있다고 가정하면 히스토그램은 구간 폭에 의해 결정된다. 또한, 이 구간폭이 등간격인 경우만 국한 한다면 히스토그램의 모양은 구간의 수에 의해 결정된다. 구간의 수가 너무 크거나 작으면 추정의 분산이 커지게 되므로 적절한 구간의 수의 결정은 추정의 분산을 최소화하는 기준으로 결정할 수 있을 것이다.

### 2.1 차이함수(Discrepancy)

* 근사에 의한 차이함수(Discrepancy due to approximation): $\Delta(F, G_\theta)$

근사에 의한 차이함수는 스무드(smooth)한 모양의 연속적인 밀도함수를 사각형의 히스토그램으로 표현하기 때문에 발생하는 차이를 나타내는 것이다. 그러므로, 표본변동에는 전혀 영향을 받지 않는다. Gauss차이함수를 사용하여 $\Delta(F, G_\theta)$를 정의하면 아래와 같이 표현될 수 있다.

$$\Delta(F, G_\theta) = \int (f(x) - g_\theta^{(I)})^2 dx$$

* 추정에 의한 차이함수(Discrepancy due to estimation): $\Delta(G_\theta, G_{\hat{\theta}})$

추정에 의한 차이함수는 일단 사각형의 모양으로 근사된 히스토그램 형태의 밀도함수에 높이를 표본으로 부터 추정하는 과정에서 발생하는 차이를 나타내는 함수이다. 그러므로, 표본의 영향을 받게 된다. 위와 마찬가지로 Gauss차이함수를 사용하여 $\Delta(G_\theta, G_{\hat{\theta}})$를 정의하면 아래와 같이 표현될 수 있다.

$$\Delta(G_{\hat{\theta}}, G_\theta) = \int (g_{\hat{\theta}}^{(I)} - g_\theta^{(I)})^2 dx$$

그러면, 밀도함수와 히스토그램과의 전체 차이를 Gauss 차이함수를 사용하여 표현하면 아래와 같이 된다.

$$\Delta(F, G_{\hat{\theta}}) = \int (f(x) - g_{\hat{\theta}}^{(I)})^2 dx$$

## 3. 최적 히스토그램

이제, 최적히스토그램 모형은 알려진 구간 시작점과 끝점을 알고 있다는 가정하에 위에서 소개한 차이함수를 최소로 하는 모형이라고 정의하자.

### 3.1 Gauss 차이함수기준에 의한 최적 히스토그램

히스토그램을 작성하려면 모든 관측된 값을 포함하는 구간 $(a, a+L)$에서 $I$개의 구간을 정의하여야 한다. 만약, 이들 구간들이 모두 등간격이라고 가정하면 히스토그램 형태의 밀도함수 $g_\theta^{(I)}(x)$는 아래와 같이 정의할 수 있다.

$$g_\theta^{(I)}(x) = 0, \quad x - a \le 0 \; or \; L < x - a$$

$$= \theta_i, \quad \frac{(i-1)L}{I} < x - a \le \frac{iL}{I}, i = 1, 2, \cdots, I,$$

$$where \; \theta_i \ge 0, \theta_I = \frac{I}{L} - \sum_{i=1}^{I-1} \theta_i$$

그러므로, 근사에 의한 차이함수는 아래와 같이 정리될 수 있다.

$$\Delta(F, G_\theta) = \int (f(x) - g_\theta^{(I)})^2 dx = \int f(x)^2 dx - 2\int f(x) g_\theta^{(I)} dx + \int (g_\theta^{(I)})^2 dx$$

위의 식에서 $\int f(x)^2 dx$는 모든 비교하고자 하는 히스토그램에 대해 항상 같은 값을 가지므로 근사에 의한 차이함수를 아래와 같이 정의할 수 있다.

$$\Delta(F, G_\theta) = -2\int_a^{a+L} f(x) g_\theta^{(i)}(x) dx + \int_a^{a+L} (g_\theta^{(i)}(x))^2 dx$$

$$= -2\sum_{i=1}^{I} \pi_i \theta_i + \sum_{i=1}^{I} \frac{\theta_i^2 L}{I}$$

$$where \; \pi_i = F\left(a + \frac{iL}{I}\right) - F\left(a + \frac{(i-1)L}{I}\right)$$

여기서, $n_i$를 $i$ 번째 구간에 포함되는 자료의 수로 정의하면 $n_i$는 모수 $(n, \pi_i)$를 갖는 이항분포를 따르는 확률변수이고 $\theta_i$는 히스토그램 형태 밀도함수의 모수로 $i$ 번째 구간의 히스

토그램 높이이다. $\theta_i$의 추정량으로 최대가능추정값을 사용한다면 $\hat{\theta}_i = \frac{n_i I}{nL}, i = 1,2,\cdots,I$ 를 사용할 수 있다. 그러면 차이함수는 아래와 같이 정의될 수 있다.

$$\Delta(F, G_{\hat{\theta}}) = -2\sum_{i=1}^{I} \pi_i \hat{\theta}_i + \sum_{i=1}^{I} \frac{\hat{\theta}_i^2 L}{I}$$

여기서, 차이함수는 확률변수이므로 각 히스토그램의 차이함수를 구하여 비교하는 것보다는 차이함수의 기대값을 구하여 비교하는 것이 적절하게 된다.

$$E_F\Delta(F, G_{\hat{\theta}}) = -2\sum_{i=1}^{I} \pi_i E\left(\hat{\theta}_i\right) + \sum_{i=1}^{I} \frac{E\left(\hat{\theta}_i^2\right) L}{I}$$

여기서,

$$E\left(\hat{\theta}_i\right) = E\left(\frac{n_i I}{nL}\right) = \frac{I}{L}\pi_i,$$

$$E\left(\hat{\theta}_i^{\,2}\right) = V\left(\hat{\theta}_i\right) + E\left(\hat{\theta}_i\right)^2 = \frac{I^2}{nL^2}(\pi_i + (n-1)\pi_i^2)$$

을 위의 식에 대입하여 풀면

$$E_F\Delta(F, G_{\hat{\theta}}) = \frac{I}{nL}\left(1 - (n+1)\sum \pi_i^2\right) \text{ 이 성립한다.}$$

또한,

$$E\left[\frac{1}{n-1}\left(\sum \frac{n_i^2}{n} - 1\right)\right] = \sum \pi_i^2$$

이므로 차이함수의 기대값 $\sum \pi_i^2$ 의 추정치로 $\frac{1}{n-1}\left(\sum \frac{n_i^2}{n} - 1\right)$를 대입한 식을 최소화하는 히스토그램을 Gauss 기준에서의 최적 히스토그램이라고 정의하기도 하고 이 기준을 $I$로 이차 다항회귀하여 가장 작은 기준을 갖는 $I$를 추정하는 방법을 사용하기도 한다.

$$\text{Gauss기준: } \frac{I}{nL}\left[1 - \frac{n+1}{n-1}\left(\sum_{i=1}^{I} \frac{n_i^2}{n} - 1\right)\right] \tag{4}$$

### 3.2. AIC에 의한 최적 히스토그램

아카이케 정보기준(Akaike Information Criterion; AIC)은 Kullback-Leibler 정보량(Information Quantity)를 이용하여 유도된 모형선택의 기준으로 아래와 같은 값을 최소화하는 모형을 선택하게 된다.

AIC = -2 ( 모형의 최대 로그-가능도 ) + 2 ( 모형의 모수 개수 )

이 기준은 여러개의 모형 중에서 같은 정도의 최대 로그 가능도(Maximum log-likelihood)를 갖는 모형이 있다면 그 중에서 모수의 수가 가장 작은 모형을 선택하겠다는 기준으로

모수절약의 원칙(The law of parsimony)에 입각한 선택기준이다. 이 문제를 히스토그램에 적용시켜 AIC를 최소화하는 히스토그램 모형이 최적 히스토그램이라고 생각할 수 있다.

$x_1, x_2, \cdots, x_n$이 임의의 모집단에서 뽑힌 표본이고, $x_{(1)}$은 자료의 최소값, $x_{(n)}$은 자료의 최대값이라고 하자. 아카이케는 모든 자료들이 포함되는 적절한 구간으로 자료의 측정정도($d$: 측정정도; precision of each measurement)를 고려하여 아래와 같이 제시했다. [6]

$$[x_{(1)} - 0.5d, x_{(n)} + 0.5d]$$

이 구간을 c개의 등간격으로 나누어 초기 범주화(Initial Catcgorization)를 한다. 초기범주의 갯수 c를 아래의 식으로 계산할 것을 제안했다.

$$c = \text{int}[2\sqrt{n} - 1]$$

이러한 초기범주화는 초기범주들을 서로 합동(pooling)시켜 여러가지의 다양한 경우의 히스토그램을 비교하기 위한 방법이다. $i$번째 초기범주안에 자료가 포함될 확률을 $p(i)$, 포함되는 자료의 갯수를 $n(i)$라고 하면 $n(i)$는 다항분포(Multinomial distribution)를 따르고 그의 밀도함수와 로그-가능도(Log-Likelihood)는 아래와 같이 구해진다.

$$\Pr[n(i)|p(i)] = \frac{n!}{\prod_{i=1}^{c} n(i)!} \prod_{i=1}^{c} p(i)^{n(i)}$$

$$l(p(i)) = K_3 + \sum_{i=1}^{c} n(i) \log p(i)$$

$$\text{where } K_3 = \log n! - \log \prod_{i=1}^{c} n(i)!$$

만약, 양쪽 끝의 초기범주들을 $c_1, c_2$개씩 합동(pooling)하고, 나머지 범주들을 등간격으로 $r$개씩 합치는 히스토그램 모형을 생각한다면 이러한 모형은 아래와 같이 표현될 수 있다.

$$Model(c_1, r, c_2):$$

$$p(1) = p(2) = \cdots = p(c_1) = \theta(1)$$

$$p(c_1 + (j-2)r + 1) = \cdots = p(c_1 + (j-1)r) = \theta(j)$$

$$p(c - c_2 + 1) = \cdots = p(c) = \theta\left(\frac{c - c_1 - c_2}{r} + 2\right)$$

$$\text{where } j = 2, \cdots, \frac{c - c_1 - c_2}{r} + 1$$

그러므로, 위의 $Model(c_1, r, c_2)$의 AIC는 아래와 같이 표현될 수 있다.

$$AIC(c_1, r, c_2)$$

$$= (-2)\left[ n'(1) \log \frac{n'(1)}{c_1 n} + \sum_{j=2}^{c'-1} n'(j) \log \frac{n'(j)}{rn} + n'(c') \log \frac{n'(c')}{c_2 n} \right] + 2(c' - 1) \qquad (5)$$

여기서, $n'(k)$는 k번째 계급에 포함된 자료의 수를 의미하고 $c'$은 합동(pooling)이 끝난 후 의 최종 계급의 수를 의미한다. 만약, 계산시에 $n'(k)$=0일 경우 $n'(k)$에 exp(-1)을 치환한다. 이러한 치환은 $k$번째 계급에 속하는 자료가 없을때 $n'(k)\log n'(k)$를 최소화시켜 AIC를 최대로 만들기 위한 것이다.

### 3.3 Average Shifted Histograms

Scott(1985)는 히스토그램의 평활기법으로 Average Shifted Histograms을 제안했다. 이 기법은 M개의 서로 다른 시작점을 갖는 히스토그램을 평균하는 기법으로 여러개의 시작점에서 히스토그램을 얻어 추정된 밀도의 평균값을 사용하므로 히스토그램의 모양이 시작점(Starting point)에 따라 달라지는 문제를 해결했다. M이 증가함에 따라 실제 밀도함수처럼 평활한 결과를 얻을 수 있는 기법으로 일종의 Kernel 밀도함수추정법으로 볼 수 있지만 일반인의 이해가 쉽고 계산시간이 빠른 장점을 가지고 있다.

$$\hat{f}_h(x) = \frac{\sum_j I_{B_j^*}(x) \sum_{k=1-M}^{M-1} w_M(k)\, n_{j+k}}{nh}$$

$h$: 계급의폭, $M$: 히스토그램의 수, (6)

$$w_M(k) = 1 - \frac{|k|}{M},\quad B_Z^* = \left[\frac{zh}{M}, \frac{(z+1)h}{M}\right),$$

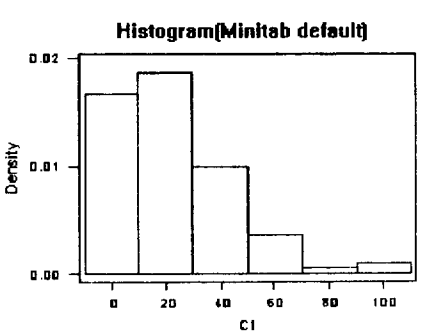
$$n_k = \sum_{i=1}^{n} I_{B_k^*}(X_i)$$

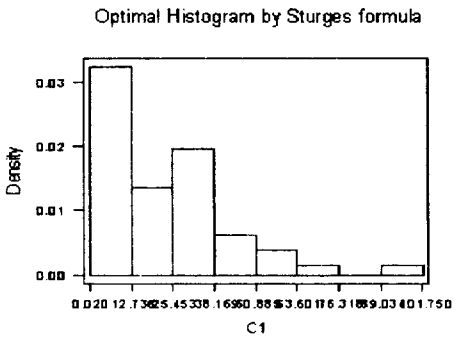
## 4. 히스토그램 산법의 구현

지금까지 알아본 기법을 통계 소프트웨어에서 구현하려면 자동으로 계급의 수를 계산하고 그 값을 히스토그램에 연결할 수 있어야 한다. 또한, 등간격이 아닌 히스토그램은 계급을 나누어 주는 구간하한, 상한값(Cut point)를 계산하고 이를 히스토그램에서 받아들일 수 있어야 한다. 여러 상용 소프트웨어(SAS, Statgraphics, SPSS, MINITAB, S-plus)에 대해 구현 가능성을 검토한 결과 SAS, 윈도우용 미니탭, S-plus에서 가능하다는 결론을 얻었다. 미니탭은 히스토그램 명령에서 구간 경계점(Cut value)을 지정할 수 있으므로 미니탭 매크로 기능을 이용하여 최적의 구간 경계점(Cut value)을 계산하여 히스토그램 명령에 연결하는 방법을 사용하였다. SAS에서는 Proc GCHART를 이용할 수 없기 때문에 SAS/IML, SAS/GRAPH의 PROC GPLOT을 매크로로 연결하여 구현하였다.

## 5. 수치 적용 예

아래의 그림 1.은 지수분포형의 자료를 미니탭에서 구현한 결과이고 그림 2.는 정규분포형의 자료를 SAS에서 구현한 결과이다.



최적 히스토그램의 구현

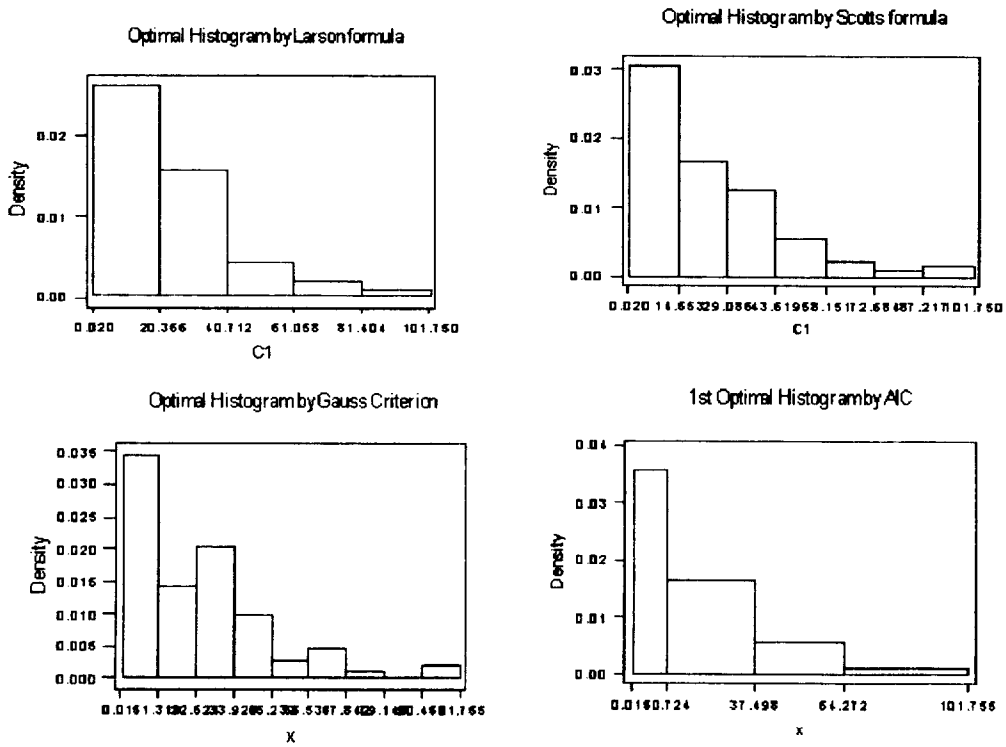
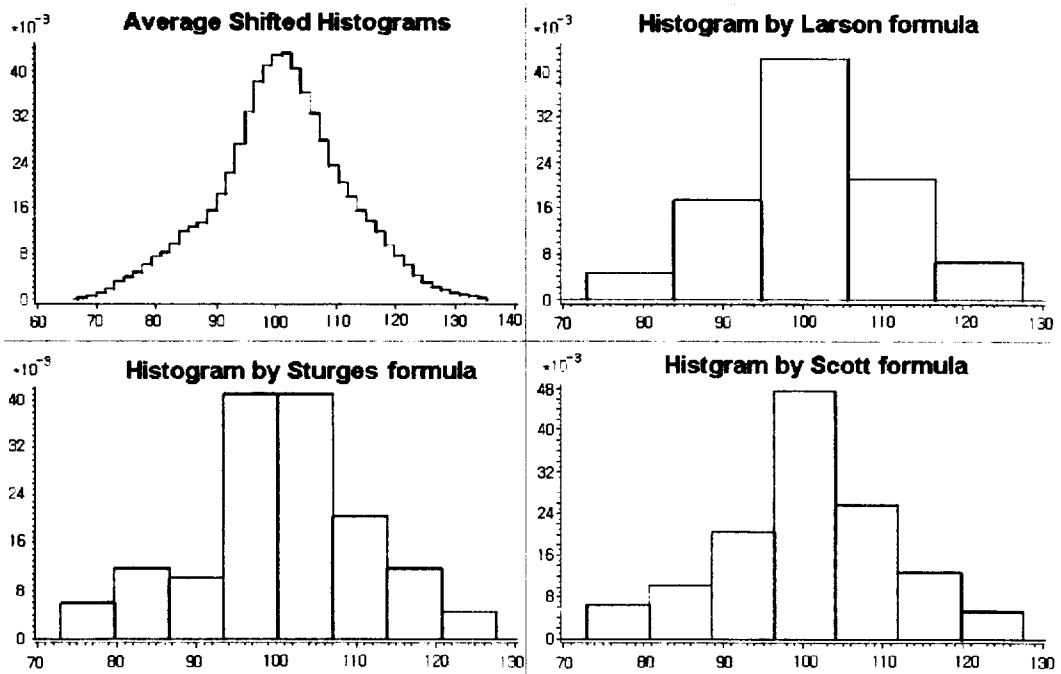



그림 1. 미니탭에서 구현한 히스토그램



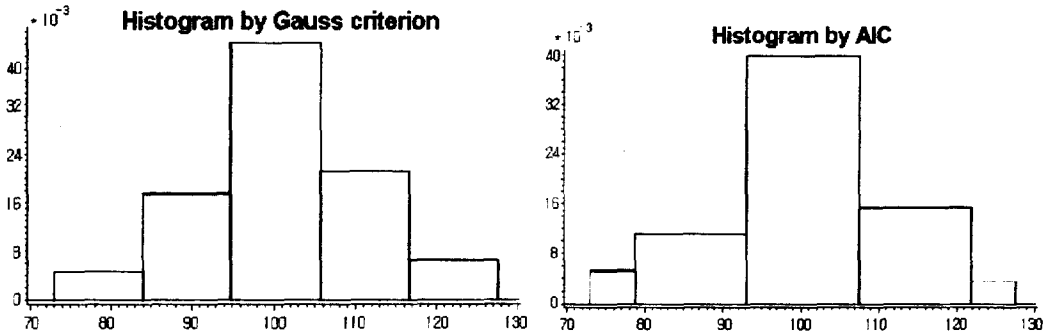



그림 2. SAS에서 구현한 히스토그램

## 6. 결 론

자료로 부터 하나의 히스토그램을 얻는 과정은 주관적인 과정이다. 본 논문에서는 초보적인 통계지식을 가진 사용자가 이러한 주관적인 결정을 하는데 도움을 주기위해 알려져 있는 산법들을 구현하는데 초점을 맞추었다. 본 논문에서 구현한 산법 중 Scott에 의해 제안된 (3) 식의 산법과 Average shifted histograms은 정규성을 가지고 있는 자료에서 다른 산법보다 밀도함수에 가까운 추정을 할 수 있음을 모의실험 결과(그림 3.) 알 수 있었다.

Gauss 기준과 AIC에 의한 방법이 다른 방법에 비해 ISE를 작게하지 못하는 결과를 얻었다. 이는 이 두 방법이 표본변동에 민감하고 모형의 변동에 따른 기준의 변동 폭이 너무 작게 나타나기 때문이 아닌가 추측된다. 또한, 자료의 정규성이 만족되지 않을 경우 여러 방법들이 서로 비슷한 ISE를 가짐을 알 수 있었다. 그러므로, Scott에 의해 제안된 식 (3)이나 (6)식이 히스토그램을 표현하는데는 적절하다고 판단된다.

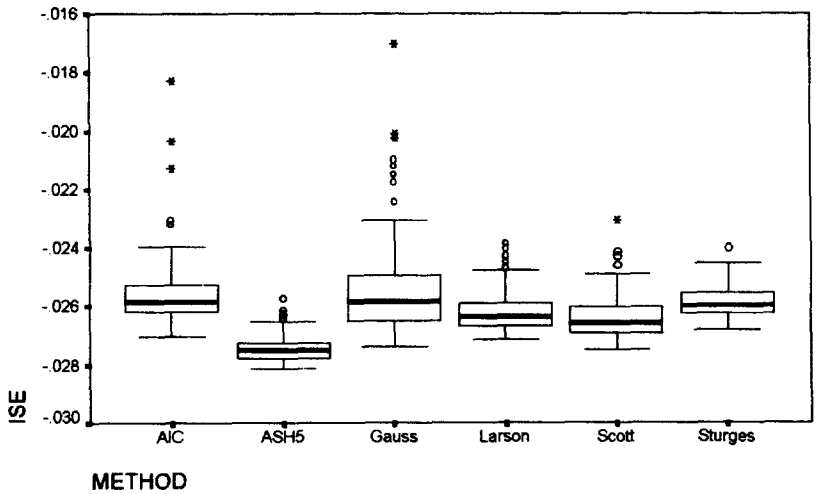



그림 3. 각 방법별 ISE($\int (f(x)-g_{\hat{\theta}}^{(I)}(x))^2 dx$) 비교: 정규성의 경우.

## 참고 문헌

[1] Larson, H. J. (1975). *Statistics: An Introduction.* New York: Wiley.
[2] Linhart & Zucchini (1986). *Model Selection.*
[3] Silverman. B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.
[4] Scott, D. W. (1979). On optimal data-based histograms. *Biometrika*, 66, 605-610.
[5] Scott, D. W. (1985). Average Shifted Histograms. *The Annals of Statistics*, Vol. 13,
[6] Sturges, H. A. (1926). The choice of a class interval. *J. Am. Statist. Assoc.* 21, 65-6. No 3., 1024-1040
[7] Y. Sakamoto, M. Ishiguro, and G. Kitagawa (1986). *Akaike Information Criterion Statistics.* KTK.
[8] Wolfgang Härdle (1991). *Smoothing Techniques with Implementation in S.* Springer-Verlag.