9장. 특수 토픽

강의를 마무리할 때가 되었습니다. 이제까지 여러분들이 익힌 것은 기본적인 몇 가지 개념과 도구입니다. 결코 충분하지는 않지만 앞으로 피가 되고 살이 될 것입니다. 그러나, 시간적·공간적 제약 때문에 수리통계학에서 다룰 필요가 있는 몇 가지 중요한 것들이 누락되는 결과가 되어갑니다. 그 중에 딱 두 가지를 골라 소개하겠습니다. 그러나 정말 잠깐 밖에 할 수 없습니다. 그러므로 여기서는 그냥 들어두는·읽어두는 정도로 만족하십시오. 여러분들 중에 앞으로 통계학과 인연이 이어질 사람들은 행복할 것입니다. 개안(開眼)의 기회가 있을 터이므로.

내가 여기서 고른 두 분야는 로버스트 추론(robust inference)과 재표집 방법(resampling method)입니다. 첫째, 로버스트 추론에서는 위치 모수에 대한 M-추정을 설명하겠습니다. 이것은 개량(改良)된 최대가능도 추정입니다. 더욱 확장하면 선형회귀모형에 대한 M-추정으로 금방 확장 가능하지만 여기서는 거기까지는 다루지 않습니다. 여백을 남겨둬야 하지 않겠습니까? 둘째, 재표집 방법으로서 붓스트랩(bootstrap) 기법과 임의순열 검증(random permutation test)을 맛보이겠습니다. 재표집(再標集, resampling)이란 표본을 모집단으로 간주하고 반복하여 같은 크기의 부표본(副標本, subsample)을 수 없이 재추출하는 것을 말합니다. 컴퓨터를 혹사시키는 계산 집중적(computer-intensive) 방법이지요.

이외에도 수리통계학의 여러 영역들이 남았습니다. 그러나 더 이상 자리가 남지 않았습니다. 그래서 빼뜨릴 수밖에 없습니다만, 그렇다고 해서 여기에 포함되지 않은 토픽들이 덜 중요한 것은 아닙니다. 사실 가장 중요하고 근본적인 것이 남았습니다. 그것은 바로 베이즈 이론(Bayesian theory)입니다. 한마디로 베이즈 이론은 통계적 추론 체계로서 으뜸입니다. 멋진 철학과 논리 체계를 갖고 있지요. 좀 더 고급 수준의 수리통계학 강의에서 베이즈 이론에 푹 빠져 보십시오.

차례 :

9.1* 로버스트 추론

이제까지의 추론은 거의 대부분 특정 확률모형을 바탕에 깔고 하는 것이었습니다. 대표적인 예는 $X_1, \cdots, X_n$ 이 정규분포 $\mathrm{N}(\theta, \sigma^2)$ 으로부터의 임의표본일 때 미지의 파라미터 θ 에 대하여 '어쩌구 저쩌구' 하는 것이었지요. 실제 통계적 문제에서는 어떤 확률분포를 가정할 것인가를 연구자가 정해야 하는데 이것이 쉽지 않습니다. 때문에 분포 가정에 무관하거나 덜 의존적인 추론 방법을 생각하게 됩니다. 분포에 무관한 전자를 비모수적 추론(nonparametric inference)이라고 하고 분포에 덜 의존적인 후자를 로버스트 추론(robust inference)이라고 합니다. 여기서는 '비모수(非母數)'를 빼고 '로버스트 성(强健性 · 剛健性)'을 다룹니다.

임의표본 $X_1, \cdots, X_n$ 각각의 확률분포가

$$f(x;\theta,\sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\theta}{\sigma}\right), \quad -\infty < \theta < \infty, \quad \sigma > 0 \tag{1}$$

의 형태라고 합시다. 여기에서 자연스럽게 θ 는 위치(location)를 나타냅니다. 당분간 σ 를 기지(旣知, known)의 파라미터라고 하겠습니다.

(1)의 부류에 해당하는 첫 번째 예는 앞의 정규분포 $\mathrm{N}(\theta, \sigma^2)$ 입니다. 그 경우엔

$$f_0(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty \tag{2}$$

입니다. 두 번째 예로서 이중 지수분포(double exponential distribution)는

$$f_0(z) = \frac{1}{2} \exp(-|z|), \quad -\infty < z < \infty \tag{3}$$

인 경우입니다. 그밖에 다른 예도 얼마든지 있습니다만, (2)와 (3)의 공통점은

- $f_0(z) = constant \cdot e^{-\rho(z)}, \quad -\infty < z < \infty$ (4)
- $\rho(0) = 0$ 이고 $\rho(z) \geq 0$ 는 $z = 0$ 을 중심으로 대칭
- $\rho(z)$ 가 거의 모든 점에서 미분가능

등이군요.

(4)의 분포에 따라 생성된 임의표본 값을 $x_1, \cdots, x_n$ 이라고 할 때, θ 에 대한 로그 가능도는

$$l(\theta; x_1, \cdots, x_n, \sigma) = -\sum_{i=1}^{n} \rho\left(\frac{x_i - \theta}{\sigma}\right)$$

입니다. 따라서, $\rho'(z) \equiv \psi(z)$ 로 표기하면 최대가능도 방정식은

$$\frac{\partial}{\partial\theta} l\,(\theta\,;x_1,\cdots,x_n,\sigma\,) = \frac{1}{\sigma}\sum_{i=1}^{n}\psi\left(\frac{x_i-\theta}{\sigma}\right) = 0 \tag{5}$$

입니다. 예컨대 정규분포 $\mathrm{N}(\theta,\sigma^2)$의 경우엔

$$\rho\,(z) = \frac{z^2}{2}\,,\quad \psi\,(z) = z \tag{6}$$

이므로 최대가능도 방정식은

$$\frac{1}{\sigma}\sum_{i=1}^{n}\frac{x_i-\theta}{\sigma} = 0$$

이 됩니다. 그리고 이로부터 최대가능도 추정치 $\hat{\theta} = \overline{x}$ 가 나옵니다.

이중지수분포의 경우엔

$$\rho\,(z) = \mid z \mid\,,\ \psi\,(z) = \begin{cases} +1 & \text{if}\, z > 0, \\ -1 & \text{if}\, z < 0 \end{cases} \tag{7}$$

이므로 최대가능도 방정식은

$$\frac{1}{\sigma}\sum_{i=1}^{n} sign\left(\frac{x_i-\theta}{\sigma}\right) = 0$$

이 됩니다. 여기서 $sign\,(z)$는 부호를 나타내므로, 부호의 합이 0이 되기 위하여는 θ 보다 큰 x_i 의 수와 θ 보다 작은 x_i 의 수가 같아야 할 것입니다. 그러므로

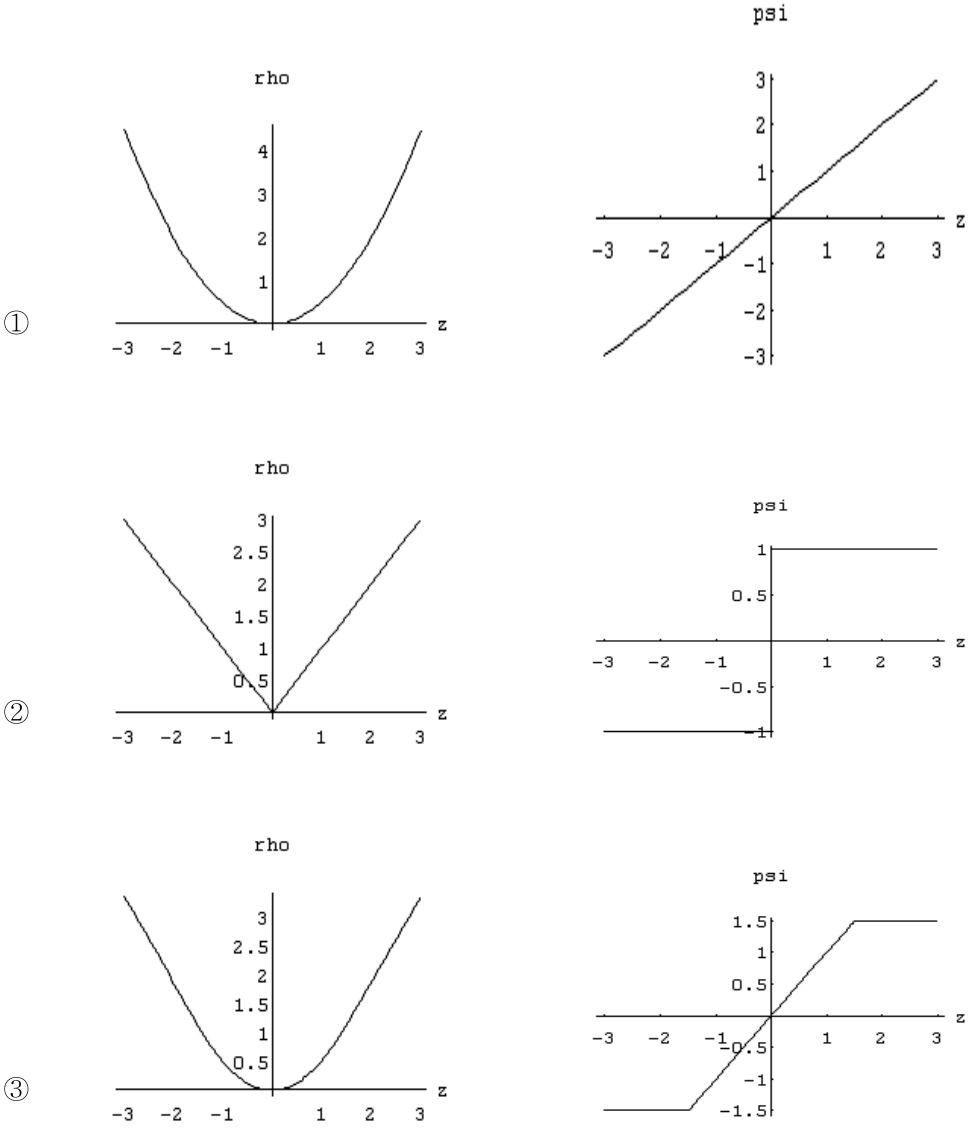
$$\hat{\theta} = \mathrm{median}\,(x_1,\cdots,x_n)\ (=\tilde{x}\,)$$

여야 합니다. $\overline{x}$ 가 정규분포에서는 가장 효율적인 추정량이지만 특이점(outlier)에 지나치게 민감한 반면, $\tilde{x}$ 는 효율성이 다소 떨어지지만 특이점에는 저항적입니다. 따라서 두 추정량을 잘 조화시킴으로써 효율성이 좋으면서도 특이점에 잘 견디는 우수한 추정량을 만들어내자는 것이 로버스트 추정입니다 ('robustness'의 사전적 의미는 강건(强健・剛健)함・튼튼함입니다).

위치 모수 θ에 대한 로버스트 추정에서 가장 대표적인 M-추정(M-estimation)은

$$\psi\,(z) = \begin{cases} +k & \text{if}\ z > +k, \\ z & \text{if}\ -k \le z \le +k, \\ -k & \text{if}\ z < -k \end{cases} \tag{8}$$

로 놓고 (5)를 푸는 것입니다. (8)을 후버(Huber)의 프시(psi) 함수라고 하는데 이것은 (6)과 (7)의 절충입니다. 즉, k가 $+\infty$로 가면 후버의 프시 함수는 정규분포의 프시 함수인 (6)이 되고 k가 0으로 가면 이중지수분포의 프시 함수인 (7)이 됩니다. 후버의 프시함수에서 $k > 0$는 상수로서 1.5 정도로 둡니다. <그림 1>을 보십시오.



<그림 1> 로(rho) 함수와 프시 함수 : ① 정규분포, ② 이중지수분포, ③ M-추정.

※ 참고 : 후버(Huber)의 프시 함수에 대한 로(rho) 함수는 다음과 같습니다.

$$\rho(z) = \int_0^z \psi(u)\,du = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \le k, \\ k|z| - \frac{1}{2}k^2 & \text{if } |z| \ge k. \end{cases}$$

그러면 어떻게 로버스트 추정치를 구할 수 있을까요? 최대가능도 방정식이

$$g(\theta) \equiv \sum_{i=1}^{n} \psi\left(\frac{x_i-\theta}{\sigma}\right) = 0$$

이므로 이것의 1계 미분은

$$g'(\theta) = -\frac{1}{\sigma}\sum_{i=1}^{n} \psi'\left(\frac{x_i-\theta}{\sigma}\right)$$

입니다. θ_0 을 θ 에 대한 초기 추정값이라고 하면, 뉴튼-라프슨 반복식은

$$\theta = \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)} = \theta_0 + \sigma \cdot \frac{\sum_{i=1}^{n} \psi\left(\frac{x_i-\theta_0}{\sigma}\right)}{\sum_{i=1}^{n} \psi'\left(\frac{x_i-\theta_0}{\sigma}\right)}$$

가 됩니다. 후버의 M-추정에서

$$\psi'(z) = \begin{cases} 1 & \text{if } |z| \leqq k, \\ 0 & \text{if } |z| > k \end{cases}$$

이고 장애 파라미터 σ 는 MAD(median of absolute deviations)라고 하는

$$\tilde{\sigma} = \text{median}_{i=1,\cdots,n} |x_i - \tilde{x}| / 0.6745$$

로 대치해 놓습니다 (참고 : $\pm 0.6745\sigma$ 는 정규분포 $N(0,\sigma^2)$ 의 위·아래 4분위수임).

수치 예를 들어 볼까요? 다음은 15명(= n)의 스위스 산부인과 의사가 1년간 시술한 자궁절제수술 수입니다 (크기 순 정렬).

20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86.

이 자료에서 나중의 두 값이 유난히 큽니다. 그러므로 평균(=41.33)이 메디안(=34) 보다 크게 나옵니다. 그리고 MAD는 13.34로 계산됩니다. <표 1>은 k 를 1.5로, 메디안을 초기값으로 하여 후버의 M-추정치를 구하기 위한 SAS/IML 프로그램입니다. 그 결과는

$$\hat{\theta}_M = 37.50$$

입니다. 평균과 메디안의 중간쯤 있음을 볼 수 있습니다.

그 다음 단계의 일은 $\hat{\theta}_M$의 표집분포를 구하는 일이겠습니다. 그것을 알아야 θ 에 관한 신뢰구간을 구할 수 있을 것이기 때문입니다. 다음 절에서 이를 구하는 한 가지 방법을 간략히 설명하겠습니다.

<표 1> 후버의 M-추정치를 구하기 위한 SAS/IML 프로그램

```
/* FileName: robust.iml  */
/* Huber's M-estimation  */

proc iml;
  x = {20 25 25 27 28 31 33 34 36 37 44 50 59 85 86};
  n = ncol(x);
  mean = x[,+]/n;  sd = sqrt((ssq(x)-n*mean*mean)/(n-1));
  y = j(1,n,0);  r = rank(x);
  do i=1 to n;   y[r[i]] = x[i];   end;
  if mod(n,2) ^= 0 then median = y[(n+1)/2];
    else median = (y[n/2] + y[n/2+1])/2;
  x1 = abs(x - median*j(1,n,1));
  r1 = rank(x1);
  do i=1 to n;   y[r1[i]] = x1[i];   end;
  if mod(n,2) ^= 0 then mad = y[(n+1)/2]/0.6745;
    else mad = (y[n/2] + y[n/2+1])/2/0.6745;
  print n mean[format=8.2] sd[format=8.2] median mad[format=8.2];

  k = 1.5;   theta0 = median;   iter=0;   diff=1;
  do while (diff > 0.0001);
   temp = (x-theta0*j(1,n,1))/mad;
   num = temp;  denum = j(1,n,1);
   do i = 1 to n;
     if temp[i] > k then num[i] = k;
       else if temp[i] < -k then num[i] = -k;
     if abs(temp[i]) > k then denum[i] = 0;
   end;
   theta = theta0 + mad*sum(num)/sum(denum);
   diff = abs(theta-theta0);
   iter = iter+1;
   theta0 = theta;
  end;
  print "Huber's M-estimate" iter theta[format=8.2];
quit;
```

9.2* 재표집 방법

미지의 파라미터 θ에 대한 추정량으로 $\hat{\theta} = T(X_1, \cdots, X_n)$을 생각합시다. 통계학에서 핵심적 사항은 $\hat{\theta}$의 확률적 행태를 파악하여 추론에 활용하는 것입니다. $X_1, \cdots, X_n$이 확률분포 $F(x)$로부터의 임의표본이라면 (θ는 $F(x)$의 한 특성치), $\hat{\theta}$의 표집분포를 산출하는 것은 별 문제가 없을 것입니다. 수학적으로 구하는 것이 어려운 경우에는 몬테칼로 방법이 있으니까요. 이 과정을 표현하면 다음과 같습니다.

$$\text{모분포 } F(x) \quad \rightarrow \quad \text{임의표본 } x_1, \cdots, x_n \quad \rightarrow \quad \text{통계량 } T(=\hat{\theta}). \tag{9}$$

통계량 $T(=\hat{\theta})$의 표집분포에서 평균 $E(\hat{\theta}\,;F)$과 분산 $Var(\hat{\theta}\,;F)$를 계산함으로써 추정량 $\hat{\theta}$의 편향과 표준오차를 알게 됩니다. 즉

$$\text{bias}(\hat{\theta}\,;F) = E(\hat{\theta}\,;F) - \theta\,, \quad \text{s.e.}(\hat{\theta}\,;F) = \sqrt{Var(\hat{\theta}\,;F)} \tag{10}$$

그러나 확률분포 $F(x)$의 형태에 대하여 어떤 가정도 할 수 없을 만큼 전혀 모른다면 이와 같은 식으로는 접근할 수 없습니다. 그럼에도 불구하고 $x_1, \cdots, x_n$이 미지의 확률분포 $F(x)$로부터의 임의표본이기 때문에 $F(x)$를

$$\hat{F}(x) = \frac{1}{n}, \quad \text{for } x = x_i,\ i = 1,\cdots,n$$

로 추정하는 것이 가능합니다. 붓스트랩(bootstrap) 방법이란 (9)와 유사한 과정

$$\text{표본분포 } \hat{F}(x) \quad \rightarrow \quad \text{부표본 } x_1^*, \cdots, x_n^* \quad \rightarrow \quad \text{부통계량 } T^*(=\hat{\theta}^*) \tag{11}$$

을 θ에 대한 추론에 활용하는 것입니다. 즉

$$\widehat{\text{bias}}(\hat{\theta}\,;F) = E(\hat{\theta}^*\,;\hat{F}) - \hat{\theta}\,, \quad \widehat{\text{s.e.}}(\hat{\theta}\,;F) = \sqrt{Var(\hat{\theta}^*\,;\hat{F})}\,. \tag{12}$$

이 식은 (9)와 (11)의 대응관계로부터 유도되는 (10)의 귀결입니다. 직관적으로 받아들이기 바랍니다. (12)를 계산하기 위한 구체적 알고리즘은 다음과 같습니다.

붓스트랩 알고리즘(bootstrap algorithm) :

1) 표본분포 $\hat{F}(x)$로부터 부표본(subsample) $x_1^*, \cdots, x_n^*$를 생성시킨다.

2) $x_1^*, \cdots, x_n^*$로부터 부통계량 $T^*(=\hat{\theta}^*) = T(x_1^*, \cdots, x_n^*)$를 산출한다.

3) 단계 1과 2를 B번 반복하여 $\hat{\theta}_{(b)}{}^*$를 얻는다 ($b = 1,\cdots,B$).

4) 다음 식에 의하여 편향과 표준오차를 추정한다.

$$\widehat{\text{bias}}_B(\hat{\theta}\,;F) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)}{}^* - \hat{\theta}\,, \quad \widehat{\text{s.e.}}_B(\hat{\theta}\,;F) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_{(b)}{}^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)}{}^*\right)^2}\,.$$

9.1절에서 다룬 수치 예를 다시 생각하기로 합니다. 위치 모수 θ에 대한 M-추정량이 어느 정도의 변이를 갖는지를 알아보도록 합시다. 한 방법은 방금 설명한 붓스트랩 방법입니다. <표 2>가 이를 위한 SAS/IML 프로그램이고 <그림 2>는 M-추정량의 붓스트랩 분포입니다 ($B = 999$). 붓스트랩 방법으로 산출된 편향 및 표준오차는 다음과 같습니다.

$$\widehat{\text{bias}}_B(\hat{\theta}\,;F) = 37.49 - 37.50 = -0.01, \quad \widehat{\text{s.e.}}_B(\hat{\theta}\,;F) = 5.45\,.$$

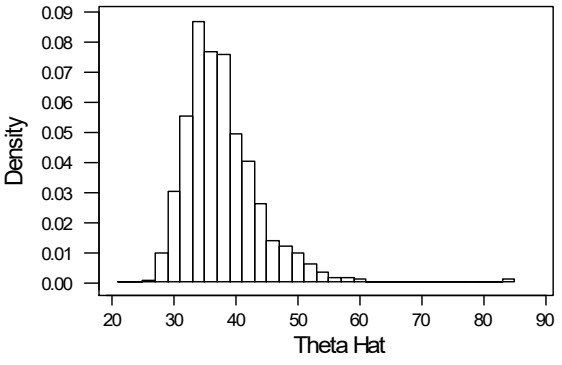
그러므로 $\hat{\theta}$의 편향은 거의 없다고 볼 수 있습니다. 붓스트랩 방법으로 θ에 관한 신뢰구간을 구하는 방법은 여러 가지이지만 초보적인 한 방법은

$$\hat{\theta} \pm 2\widehat{\text{s.e.}}_B = 37.50 \pm 2 \cdot 5.45 = (26.60, 48.40)$$

입니다 (신뢰수준 95%). <그림 2>와 같은 붓스트랩 분포의 상·하위 2.5% 분위수를 활용하는 것도 한 방법입니다. 이 방법으로 구한 θ에 관한 95% 수준의 신뢰구간은

$$(\hat{\theta}^{*(0.025)}, \hat{\theta}^{*(0.975)}) = (29.50, 52.60)$$

입니다. 이 구간에는 붓스트랩 분포의 비대칭성이 일부 반영되어 있습니다.



<그림 2> M-추정량의 붓스트랩 분포

<표 2> M-추정치의 붓스트랩 분포를 구하기 위한 SAS/IML 프로그램

```
/* FileName: boots.iml   */
/* Huber's M-estimation  */

proc iml;
  x = {20 25 25 27 28 31 33 34 36 37 44 50 59 85 86};
  n = ncol(x);
  B = 999;
  theta_M = j(B,1,0);
  xstar = j(1,n,0);

  start boots;
    do i=1 to n;  xstar[i] = x[int(n*ranuni(0))+1];  end;
    mean = xstar[,+]/n;  ;
    y = j(1,n,0);  r = rank(xstar);
    do i=1 to n;   y[r[i]] = xstar[i];   end;
    if mod(n,2) ^= 0 then median = y[(n+1)/2];
      else median = (y[n/2] + y[n/2+1])/2;
    x1 = abs(xstar - median*j(1,n,1));    r1 = rank(x1);
    do i=1 to n;   y[r1[i]] = x1[i];   end;
    if mod(n,2) ^= 0 then mad = y[(n+1)/2]/0.6745;
      else mad = (y[n/2] + y[n/2+1])/2/0.6745;
    k = 1.5;   theta0 = median;   iter=0;   diff=1;
    do while (diff > 0.0001);
     temp = (xstar-theta0*j(1,n,1))/mad;
     num = temp;  denum = j(1,n,1);
     do i = 1 to n;
       if temp[i] > k then num[i] = k;
         else if temp[i] < -k then num[i] = -k;
       if abs(temp[i]) > k then denum[i] = 0;
     end;
     theta = theta0 + mad*sum(num)/sum(denum);
     diff = abs(theta-theta0);
     iter = iter+1;
     theta0 = theta;
    end;
    theta_M[repeat] = theta;
  finish;

  do repeat=1 to B;  run boots;  end;
     print theta_M[format=8.2];
  m1 = theta_M[+,]/B;
  se = sqrt((ssq(theta_M)-B*m1*m1)/(B-1));
     print m1[format=8.2] se[format=8.2];
  theta_0 = j(B,1,0);  r2 = rank(theta_M);
  do k=1 to B;  theta_0[r2[k]] = theta_M[k];  end;
  l=theta_0[(B+1)*0.025];  u=theta_0[(B+1)*0.975];
     print l[format=8.2] u[format=8.2];
quit;
```

붓스트랩 기법과 더불어 임의순열 검증(random permutation test)도 재표집 방법입니다. 여기서는 가장 간단한 예를 들어 보이겠습니다. 이변량 자료

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

에서 두 변수 사이의 상관성에 관한 가설 검증에 관심이 있다고 합시다. 이 때 영가설H_0가 두 변수 사이에 아무런 상관성이 없다는 것인 반면, 대안가설H_1은 음의 상관성이 있다는 것이라고 합시다. 실제로 이 자료에서 피어슨 상관계수 r이 음으로 나왔다면 이것이 대안가설을 지지하는 유의한 증거가 될까요? 이에 대한 대답을 하려면 p-값이 필요하고, p-값을 산출하려면 영가설 하에서의 피어슨 상관계수 r의 분포가 필요하고, 피어슨 상관계수 r의 분포를 구하려면 이변량 자료에 대한 확률분포 가정이 필요합니다. 많은 경우 그것을 이변량 정규분포로 놓습니다만 그것은 편의로 그럴 뿐이지 어떤 확고한 근거가 있는 경우는 별로 없을 것입니다. 그러나 여기서 설명하려는 임의순열 검증(random permutation test)에는 그런 확률분포의 가정이 필요 없습니다. 그 논리 및 절차는 다음과 같습니다.

영가설H_0하에서는 두 변수 X와 Y가 아무런 상관성을 갖지 않습니다. 그러므로 $y_1, y_2, \cdots, y_n$의 한 순열을 $y_1^*, y_2^*, \cdots, y_n^*$라고 할 때

$$(x_1, y_1^*), (x_2, y_2^*), \cdots, (x_n, y_n^*)$$

에서 나오는 상관계수 r^*도 r과 확률적으로 동등합니다. 그런데 이와 같은 순열자료는 모두 $n!$개가 있습니다. 따라서 r^*가 분포를 이루게 됩니다. 그런데 $n!$이라는 숫자가 워낙 크므로 (예컨대 $20! \simeq 2.43 \cdot 10^{18}$) 모든 순열을 다 산출하는 대신 $N(=1000)$개 정도의 순열자료를 임의로 생성시켜 상관계수 r^*의 분포를 구하는 것이 훨씬 쉽고 또 그것으로도 충분합니다. 그리고 분포내에서의 r_0의 위치로써 근사적인 p-값이 나오겠습니다. 즉

$$\text{p-값} = P\{r^* \le r\} \simeq \frac{1}{N} \#\{r^* \le r\}.$$

수치 예를 봅시다. 20명의 아동으로부터 측정된 이변량 자료

X = 15 26 10 9 15 20 18 11 8 20 7 9 10 11 11 10 12 17 11 10

Y = 95 71 83 91 102 87 93 100 104 94 113 96 83 84 102 100 105 121 86 100

에서 변수 X는 첫 단어를 말한 나이(월)이고 변수 Y는 게셀 점수(Gessel score, 일종의 심리 점수)입니다. 두 변수간의 상관계수는 $r = -0.33$ 입니다 ($n = 20$). 말을 더디 시작한 아동일수록 심리적 적응력이 떨어진다고 할 수 있을까요?

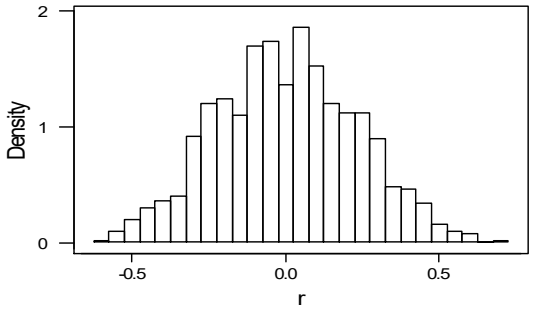
<표 3> 상관계수 r의 임의순열 분포를 구하기 위한 SAS/IML 프로그램

```
/* Random Permutation Test */
/* FileName : permute.iml  */

proc iml;
   Nrepeat = 1000;
   x = {15 26 10 9 15 20 18 11 8 20 7 9 10 11 11 10 12 17 11 10};
   y = {95 71 83 91 102 87 93 100 104 94 113 96 83 84 102 100 105 121 86 100};
   n = ncol(x);
   r = j(Nrepeat,1,0);
   xbar = sum(x)/n;  ybar = sum(y)/n;
   num = sum(x#y) - n*xbar*ybar;
   denum = sqrt((ssq(x)-n*xbar*xbar)*(ssq(y)-n*ybar*ybar));
   r0 = num/denum;
   print n r0[format=8.2];

   u = j(1,n,0);
   count = 0;
   do repeat=1 to Nrepeat;
      do i=1 to n;  u[i] = ranuni(0);  end;
      rnk = rank(u);  y1 = y;
      do i=1 to n;  y1[rnk[i]] = y[i];  end;
      y1bar = sum(y1)/n;
      num = sum(x#y1) - n*xbar*y1bar;
      denum = sqrt((ssq(x)-n*xbar*xbar)*(ssq(y1)-n*y1bar*y1bar));
      r1 = num/denum;
      if r1 <= r0 then count=count+1;
      r[repeat,1] = r1;
   end;

   pvalue = count/Nrepeat;
   print r[format=8.2];
   print Nrepeat pvalue;
quit;
```

<표 3>은 상관계수 r의 임의순열 분포를 구하기 위한 SAS/IML 프로그램이고 <그림 3>은 그것을 그린 것입니다 ($N = 1000$). 그리고 더 자세한 p-값을 계산한 결과는 0.0798입니다 ($N = 10000$). 만약 이변량 자료에 대하여 정규분포를 가정한다면

$$t_0 = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = -1.4832$$

가 영가설 하에서 자유도 $n-2\,(=18)$의 t-분포에서 나오므로 p-값은 0.0777입니다. 이 자료에서는 임의순열 검증의 결과와 아주 근사하군요.



<그림 3> 상관계수 r의 임의순열 분포 ($N = 1000$)

임의순열 검증은 상관계수에 대한 검증 외에도 여러 모평균을 비교하기 위한 분산분석 등에서 정규분포에 기반을 둔 F-검증을 대체하는 강력한 통계적 기법입니다.

9.A* 연습문제

9.1 M-추정을 위한 튜키(Tukey)의 프시(psi) 함수는 다음과 같습니다.

$$\psi(z) = \begin{cases} z\left(1-\left(\frac{z}{B}\right)^2\right)^2 & \text{if } |z| \le B, \\ 0 & \text{if } |z| \ge B. \end{cases}$$

이 프시 함수를 사용하여 9.1절의 자궁절제수술건수의 중심에 대한 M-추정치를 구하시오 (흔히 B는 6으로 놓음). 그리고 튜키의 M-추정량(=bisquare estimate)에 관한 붓스트랩 분포와 편향 및 표준오차의 추정치를 구하세요.

9.2 9.2절의 게셀 점수자료에서, 두 변수간 상관계수 r의 Fisher 변환

$$h(r) = \frac{1}{2}\log_e\frac{1+r}{1-r}$$

의 붓스트랩 분포와 편향 및 표준오차의 추정치를 구하세요. 특히 표준오차의 추정치가 $\frac{1}{\sqrt{n-3}}$에 얼마나 근사하는가를 살펴보세요.

9.B 읽을만한 책

로버스트 추론에 대하여는 다음 책을 읽어보십시오.

- 송문섭 (1996) 「로버스트 통계」 자유아카데미.
- Hogg, R.V. and Craig, A.T. (1995) *Introduction to Mathematical Statistics*, 5th Edition. Prentice Hall. (Section 8.4)
- Li, G. (1985) "Robust Regression," in *Exploring Data Tables, Trends and Shapes* (Edited by D.C. Hoaglin et al.). Wiley. (pp.281-343)

붓스트랩 방법에 대하여는 다음 책을 읽어보십시오.

- 전명식 · 정형철 · 진서훈 (1997) 「붓스트랩 방법의 이해」 자유아카데미.
- Efron, B. and Tibshirani, R. (1995) *Introduction to Bootstrap Methods*.

임의순열 검증에 대하여는 다음 책을 보십시오.

- 허명회 (1997) 「통계적 개념 · 방법 · 응용」 자유아카데미. (p. 78, 87, 124, 145)
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall. (Chapter 6)
- Good, P. (1994) *Permutation Tests*. Springer-Verlag.