Communications of the Korean Statistical Society Vol. 15, No. 3, 2008, pp. 353–365

임상시험에서 중도탈락을 고려한 표본크기의 결정

이기훈[1)]

요 약

임상시험에서 피험자수는 검정가설, 변수값의 분산과 유효차이, 검정력과 유의수준 등에 의해 결정되어진다. 일반적으로 수학적으로 계산된 피험자수에 중도탈락 예상치를 고려한 피험자수를 추가하여 최종적인 실험참가자수를 결정하는데 본 논문에서는 이론적인 계산식에서부터 중도탈락을 고려하여 피험자수를 결정하는 방법을 제안한다. 임상시험에서 많은 자료는 경시적(longitudinal) 형태를 갖고, ITT(intention to treat) 실험의 경우 중도탈락이 생기면 결측값으로 처리하지 않고 탈락직전에 관측된 값을 최종값으로 대체하는 LOCF(last observation carried forward) 방법을 주로 사용한다. 이러한 LOCF 방법은 피험자수 계산에 사용했던 분산과 유효차이 값의 가정에 왜곡을 가져오기 때문에 우리가 원하는 검정력을 보장 받지 못할 수 있다. 본 연구에서는 중도탈락률에 관한 정보를 포함하는 피험자수의 결정식을 제안하고 평균의 동일성 검정 경우에 검정력을 비교하여 이러한 산출방식이 합리적임을 실증하였다.

주요용어: 임상시험; 피험자수; 중도탈락; LOCF; ITT.

1. 서론

통계적인 실험설계를 계획할 때 적당한 피험자수 또는 표본크기(sample size)의 결정은 매우 중요하다. 특히 표본의 수가 실험의 비용에 막대한 영향을 주거나 윤리적 문제를 야기할 수 있는 임상시험(clinical trial)에서는 그 중요성이 매우 강조되어 왔다. 임상시험의 효과에 관한 연구가 학술잡지에 투고될 때 과거에는 충분히 많은 표본의 크기가 최선으로 간주되었지만, 최근에는 실험의 표본크기 산출에 관한 수리적인 근거가 미리 제시되지 않고 행한 시험의 결과는 출판되지 못하는 지경에 이르렀다 (Altman 등, 2000). 일부 학술지는 사전적인 표본수 산출이 계획되지 않은 경우에는 사후검정력(post-experimental power)을 표시하는 걸로 연구자의 편의를 도모한 적도 있지만 Hoeing과 Heisey (2001)에 의하여 사후검정력이 p-값에 의존하는 의미 없는 값이라는 사실이 밝혀진 후로는 표본크기는 실험계획 이전에 반드시 결정되어야 할 사항으로 인식되고 있다.

Altman (1991)에 의하면 충분하지 못한 피험자수는 우리에게 유용한 결과를 얻는데 불충분하기 때문에 경제적으로는 자원의 낭비이고, 윤리적으로는 근거 있는 결과를 얻지도 못하는 불필요한 실험에 환자들을 노출시킨 문제를 갖고 있다고 한다. 또한 통계이론적 측면에서도 적은 표본수로는 귀무가설 (H_0: 흔히 처리효과 없음)을 기각하기 어려우므로

1) (560-759) 전북 전주시 완산구 효자동 3가 1200, 전주대학교 경영학부, 교수. E-mail: khlee@jj.ac.kr

2종의 오류를 크게 한다. 반대로 피험자수가 필요이상으로 많다면 자원의 낭비일 뿐 아니라 과도한 인원에게 효과가 불확실한 처치를 실행해야하는 윤리적 문제를 야기한다. 그리고 통계적으로도 표본수가 많으면 의학적으로는 중요하지 않을 정도의 작은 차이에도 귀무가설을 기각하는 고전적 통계검정의 모순 현상이 나타날 수 있다 (Moher 등, 1994). 그러므로 피험자수는 의학적으로 의미 있는 차이인 유효효과크기(effect size)를 검출할 수 있는 확률이 어느 수준 이상이 될 수 있도록 충분하면서도 과도하지 않은 크기로 결정되어야 한다 (Spiegelhalter와 Freedman, 1986).

적합한 표본크기는 주로 검정력(power of test)을 기준으로 유도하는데, 이를 계산하기 위해 검정통계량, 유의수준, 의미 있는 유효효과차이 등이 결정되어야 하고, 사전 연구(pilot study)에 의하여 기존 치료의 평균, 표준편차, 또는 성공비율 등이 추산 가능해야 한다. 검정통계량의 분포, 대립가설 (H_1: 흔히 처리효과 있음) 아래서의 모수값, 검정의 목적 등이 검정력 계산에 영향을 주는데 계산의 편의상 근사분포나 적당한 가정을 가미하는 경우가 많다 (Machin 등, 1997).

대부분 의학적인 임상시험에서는 자료를 경시적(다시점, longitudinal) 방법으로 수집하게 된다. 이는 연구목적이 경시적 자료분석을 위한 것이 아니고 일반적인 단변량 분석인 경우에도 자료를 다시점으로 수집하고 최종적인 분석에서 마지막 시점에서의 측정자료만을 이용하는 것이 일반적인 임상시험의 분석행태이다. 그런데 일정기간동안 한 개체에서 반복적으로 자료를 수집하게 되면 거의 반드시 일부 실험개체에서 중도탈락(dropout)이 발생하게 된다. 중도탈락에도 불구하고 필요한 피험자수를 확보하기위해 이론적으로 계산된 피험자수에 중도탈락률을 고려한 추가적인 값을 더하여 최종피험자수를 결정하는 것이 일반적이다. 중도탈락한 개체를 최종 분석에 사용할 것인지 제외할 것인지에 따라 각각 ITT(배정된 대로 분석법, intention to treat)와 PP(계획서 순응법, per protocol) 분석법으로 나누어지는데, 엄격한 임상시험에서는 실험을 완전히 종료하지 못하였다 하더라도 최초에 실험에 참여한 환자를 어떤 형태로든 분석에 포함시켜야 한다는 ITT 분석법이 주로 사용된다. 중도탈락된 값을 포함하여 분석을 하려면 관측되지 못한 값을 대체(imputation)하여야 하는데 대부분의 임상시험에서는 치료효과에 대하여 가장 보수적인 판단을 내릴 수 있도록 결측직전의 값을 최종결과값으로 대체하는 LOCF(last observation carried forward) 방법을 사용하도록 하고 있다. 그러나 중도탈락한 개체들은 시험을 완료하지 못하였기 때문에 기대하였던 효과가 미처 나타나지 않은 상태이고 이는 분석자료의 분산을 증대시키는 결과를 가져올 수 있다. 이러한 현상은 통계분석의 검정결과에 영향을 주게 되어 중도탈락을 예상해 일정부분 더해준 피험자수로는 애초에 기대했던 검정력을 확보하지 못할 수 있다. 즉, 중도탈락비율만큼 단순 추가하는 기존의 피험자수 결정방법은 PP 분석에서는 의미가 있지만 ITT 방법에서는 수정되어야 한다. 그래서 본 연구에서는 LOCF 개체를 갖고도 기대하는 검정력을 확보할 수 있는 피험자수 결정 방법에 관하여 제안하고 이를 기존의 방법과 비교하였다.

2장에서는 중도탈락이 전체 자료에 주는 영향을 실제 예제를 통해 살펴보고, 3장에서는 두 집단의 평균비교 검정에서 피험자수 결정, 4장에서는 일원배치분석에서 피험자수 결정에 관하여 제안하고 기존의 방법과 비교하였다. 5장에서는 결론과 향후 연구과제에 관하여

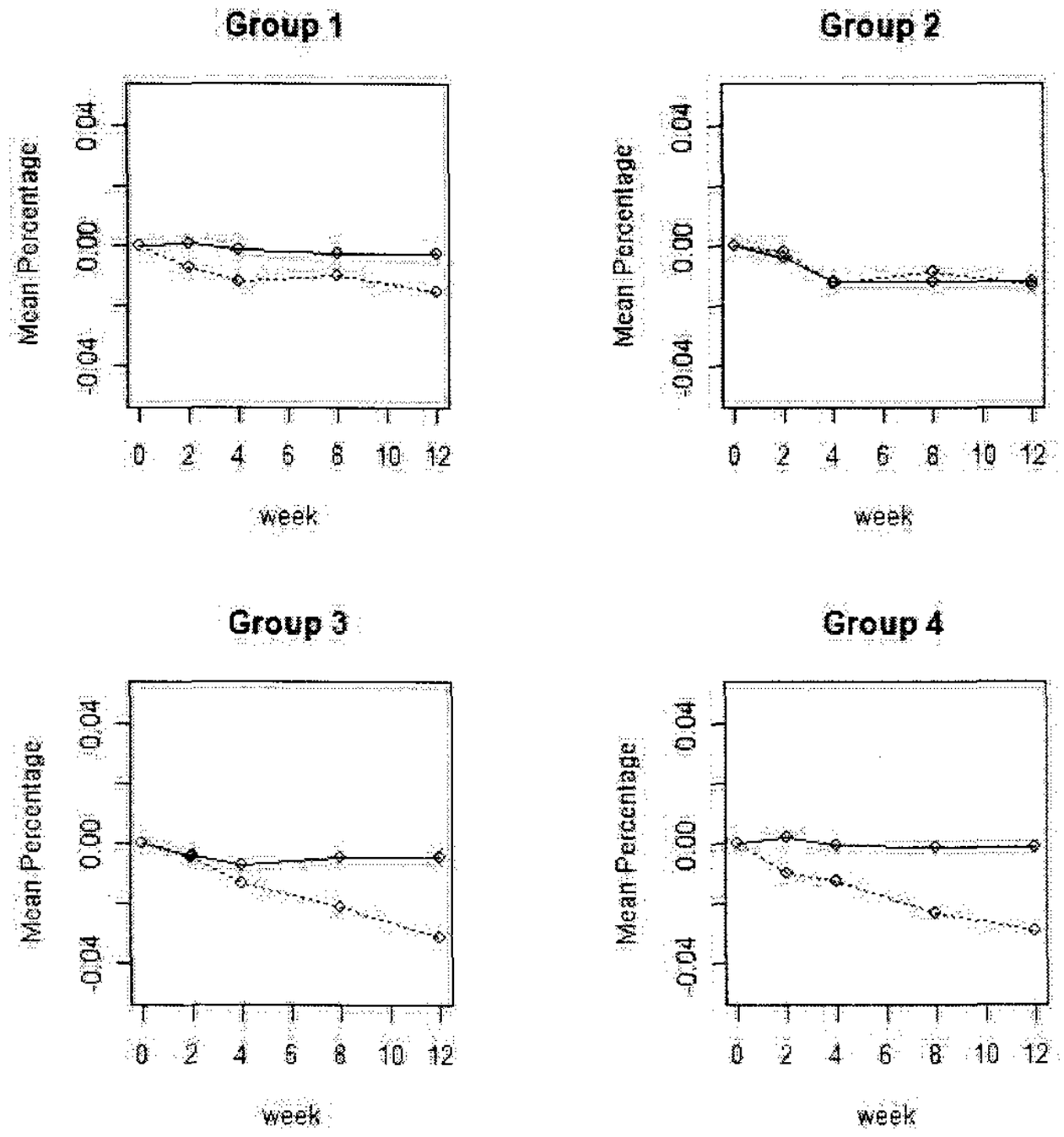

그림 2.1: 투약군 별 탈락개체와 실험완료개체 평균 변화 추이 (중도탈락개체: —, 실험완료개체: ···)

언급하고자 한다.

2. 임상시험에서 중도탈락의 대체: 예제

2005년 3월부터 2006년 3월까지 서울 A병원에서 비만치료제 L제의 체중감소 효과를 측정하고자 124명의 자원자를 4개의 투약군 (위약군, 600㎎, 1200㎎, 1800㎎)으로 나눠 12주간 약을 투여하고 체중변화를 조사한 자료 (Lee, 2006)를 이용하여 중도탈락에 의한 피험자수 부족이 처리효과 분석에 어떤 영향을 미치는지 살펴보겠다. 피험자수 산출에 가정된 모수값으로는 대조군 (위약군)은 평균 3% 의 체중감소와 투여군은 각 4.5%, 5.0%, 5.5% 정도의 체중감소를 예상하였고 각 군의 표준편차는 2.5%로 가정하였다. 유의수준은 5%, 검정력은 80%로 정하고 일원배치분석 모형에서 피험자수를 계산한 결과 각 군에 21명의 피험자가 있을 때 검정력 81.16%를 보장받을 수 있었다. 여기에 중도탈락률을 30%로 가정하여 각 군별로 30명의 피험자수를 확정하고 시험을 진행하였다.

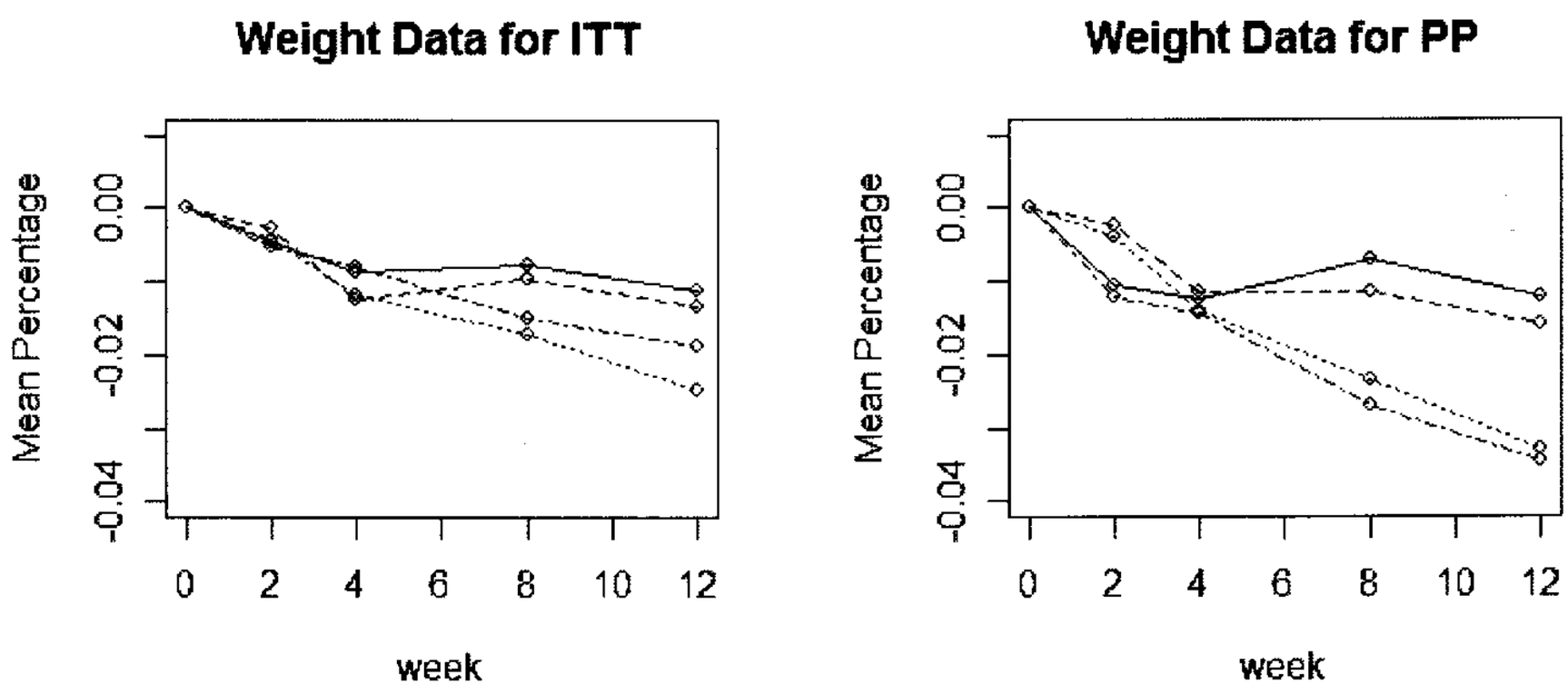

그림 2.2: 실험에 참여한 모든 개체(ITT)와 실험을 완료한 개체(PP)의 투약군별 체중감소율 평균 변화 추이 (위약군: —, 600mg군: - - -, 1200mg군: — · —, 1800mg군: · · ·)

그런데 그 임상시험이 특정 질병의 환자가 치료차 병원 방문시 이루어지는 조사가 아니고 자원자의 자발적인 참여에 의존하였기 때문에 일반적인 다른 임상시험에 비해 중도탈락이 많이 발생하여 최종적으로 67명만이 실험을 마쳤다. Lee (2006)에서 인용한 그림 2.1은 각 투약군 별로 실험을 완료한 개체와 중도탈락한 개체들의 12주간 체중변화율 평균을 도시한 그림이다. 4개의 처리군 중 1200㎎, 1800㎎ 투약군 (group 3, 4)에서 체중감소효과가 어느 정도 확인이 되었다. 그러나 중도탈락한 개체들은 탈락직전의 값들로 대체하여 평균을 구하였기 때문에 체중감소의 효과가 실험완료개체와 뚜렷한 차이를 보이고 있다. 그러므로 LOCF방법으로 탈락개체까지 분석에 포함시킬 경우 체중감소가 과소추정되고 자료의 분산의 증가를 가져올 것이 예상된다.

최종적인 자료로 ITT (124명)와 PP (67명) 기준에 의해 체중변화율의 평균추이를 살펴보면 그림 2.2와 같이 표시된다.

시험을 완료한 개체들만 분석한 PP 기준에서는 피험자수 산출식에서 가정한 것과 같이 대조군과 투약군이 약 2.5%정도의 체중감소 차이를 보였지만 중도탈락자들은 대체한 ITT 기준에서는 그 차이가 약 1%정도로 좁아지고 있다. 두 경우 모두 그림상으로는 처리효과가 존재하는 것을 짐작할 수 있다. 그러나 예상한 평균차이가 존재했던 PP의 경우에도 표본수가 너무 작아 유의한 차이를 검출하는데 실패하였다 ($F(3,63) = 1.8519$, p-값 $= 0.1469$). ITT의 경우는 표본표준편차가 사전연구에서 얻었던 2.5%와 상당한 차이를 보였다. 표본표준편차가 중도탈락이 57%로 많았던 위약군인 경우 5.1%, 중도탈락률이 42%, 33%, 50% 등인 투약군은 각각 3.5%, 3.4%, 2.6% 등으로 측정되었다. 이렇게 중도탈락으로 인해 피험자수 산출식에 사용한 표준편차보다 실제 분석자료의 표준편차가 증가하여 분산분석의 결과는 $F(3,115) = 1.2272$, p-값 $= 0.3031$ 등으로 군 간에 유의한 차이를

판정하지 못하고 있다.

두 경우 모두 처리 효과의 존재를 증명하기에는 부족한 표본수로 인하여 유용한 결과를 얻지 못하는 실험으로 마무리되었다. 이는 중도탈락률이 예상했던 값보다 높게 나온 것도 그 원인이겠지만 중도탈락에 의한 분산의 증가를 고려하지 않은 것도 심각한 실수라고 할 수 있다.

3. 두 집단의 평균비교에서 피험자수 결정

이장에서는 두 집단의 평균을 비교할 때 중도탈락을 고려하여 피험자수를 결정하는 계산식을 유도하도록 하겠다. 표본수가 n으로 같은 두 집단 자료를 각 $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$라 하고 자료를 경시적으로 l회 시점에서 반복 측정하였다고 가정하면 각 개체는 다음과 같이 표현된다.

$$\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{lk}), \qquad \mathbf{y}_k = (y_{1k}, y_{2k}, \ldots, y_{lk}), \quad k = 1, 2, \ldots, n.$$

이때, 각 집단의 k번째 개체의 i시점의 관측값은 다음과 같이 모형화할 수 있다.

$$\begin{aligned} x_{ik} &= \mu_i + \epsilon_{ik} = \mu_0 + \alpha_i + \epsilon_{ik}, \\ y_{ik} &= \mu_i' + \epsilon_{ik}' = \mu_0 + \beta_i + \epsilon_{ik}', \quad i = 0, 1, \ldots, l;\ k = 1, \ldots, n, \end{aligned}$$

여기서 $\epsilon_{ik},\ \epsilon_{ik}' \sim N(0, \sigma^2)$이라 가정한다.

또한, 각 시점에서 중도탈락률을 $d_i\,(i = 1, ..., l)$, 전체 중도탈락률은 $d(= \sum_{i=1}^{l} d_i)$라 표시하고 탈락률은 각 처리에서 동일하다고 가정한다. 여기서 편의상 $1 - d = d_l + 1$로 표시한다. 그러면 관측된 값과 LOCF에 의해 대체된 값으로 이루어진 최종 분석자료값 $X_k^*, Y_k^*\,(k = 1, 2, \ldots, n)$은 $t\,(2 \le t \le l)$시점에서 중도탈락이 발생할 때 다음과 같이 표시할 수 있다.

$$X_k^* = \begin{cases} x_{lk}, & t > l, \\ x_{t-1,k}, & t \le l, \end{cases} \qquad Y_k^* = \begin{cases} y_{lk}, & t > l, \\ y_{t-1,k}, & t \le l, \end{cases} \qquad k = 1, 2, \ldots, n. \tag{3.1}$$

최종자료값의 평균과 분산은 모든 $k\,(k = 1, 2, \ldots, n)$에서 다음과 같이 유도할 수 있다.

$$E(X_k^*) = d_1(\mu_0 + \alpha_0) + \cdots + d_{l+1}(\mu_0 + \alpha_l) = \mu_0 + \sum_{i=1}^{l+1} d_i \alpha_{i-1} = \mu_X^*,$$

$$\begin{aligned} \mathrm{Var}(X_k^*) &= E(X_k^* - \mu_X^*)^2 \\ &= d_1 E(X_0 - \mu_X^*)^2 + d_2 E(X_1 - \mu_X^*)^2 + \cdots + d_{l+1} E(X_l - \mu_X^*)^2 \\ &= d_1 E\left[X_0 - \mu_0 - \alpha_0 - \left(\sum_{i=1}^{l+1} d_i \alpha_{i-1} - \alpha_0\right)\right]^2 + \cdots \\ &\quad + d_{l+1} E\left[X_l - \mu_0 - \alpha_l - \left(\sum_{i=1}^{l+1} d_i \alpha_{i-1} - \alpha_l\right)\right]^2 \end{aligned}$$

$$= \sigma^2 + \sum_{j=1}^{l+1} d_j \left(\sum_{i=1}^{l+1} d_i \alpha_{i-1} - \alpha_{j-1} \right)^2,$$

$$E(Y_k^*) = d_1(\mu_0 + \beta_0) + \cdots + d_{l+1}(\mu_0 + \beta_l) = \mu_0 + \sum_{i=1}^{l+1} d_i \beta_{i-1} = \mu_Y^*,$$

$$\mathrm{Var}(Y_k^*) = \sigma^2 + \sum_{j=1}^{l+1} d_j \left(\sum_{i=1}^{l+1} d_i \beta_{i-1} - \beta_{j-1} \right)^2.$$

두 평균에 대한 동일성 검정의 가설은 다음과 같이 표현할 수 있다.

$$H_0 : \mu_{X_l} = \mu_{Y_l} \ (\alpha_l = \beta_l) \quad vs. \quad H_1 : \mu_{X_l} \neq \mu_{Y_l} \ (\alpha_l \neq \beta_l),$$

여기서, $\mu_{X_l} = \mu_0 + \alpha_l,\ \mu_{Y_l} = \mu_0 + \beta_l,\ k = 1, \ldots, n.$

표본의 수가 충분히 크다고 가정하고 Z-검정법을 사용할 경우 다음과 같이 유의수준 α로 귀무가설을 기각한다.

$$\left| \bar{X}^* - \bar{Y}^* \right| \geq z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}}.$$

이때 검정력 함수는 다음과 같다.

$$\begin{aligned} 1 - \beta &= \Pr(\text{reject } H_0 | \ H_1 \text{ is true}) \qquad (3.2) \\ &= \Pr\left(|\bar{X}^* - \bar{Y}^*| > z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}} \ \middle| \ \mu_{X_l} - \mu_{Y_l} \neq 0 \right) \\ &= 1 - \Pr\left(-z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}} \leq \bar{X}^* - \bar{Y}^* \leq z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}} \ \middle| \ \mu_{X_l} - \mu_{Y_l} \right), \end{aligned}$$

여기서 $\bar{X}^* - \bar{Y}^*$의 분포는 대립가설 하에서 다음과 같이 표준화하여 구한다.

$$Z = \frac{\bar{X}^* - \bar{Y}^* - \left\{ \sum_{j=1}^{l+1} d_j (\alpha_{j-1} - \beta_{j-1}) \right\}}{\sqrt{\left(2\sigma^2 + \sum_{j=1}^{l+1} d_j^2 \left(\alpha_{j-1}^{*2} + \beta_{j-1}^{*2} \right) \right) / n}} \sim N(0,\ 1),$$

여기서, $\alpha_{j-1}^* = \sum_{i=1}^{l+1} d_i \alpha_{i-1} - \alpha_{j-1},\ \beta_{j-1}^* = \sum_{i=1}^{l+1} d_i \beta_{i-1} - \beta_{j-1};\ j = 1, 2, \ldots, l+1.$

검정통계량의 분포에 따라 검정력 (3.2)는 다음과 같이 표시할 수 있다.

$$1 - \Phi\left(\frac{z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}} - \sum d_j (\alpha_{j-1} - \beta_{j-1})}{\sqrt{\frac{2\sigma^2 + \sum d_j^2 (\alpha_{j-1}^{*2} + \beta_{j-1}^{*2})}{n}}} \right) + \Phi\left(\frac{-z_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}} - \sum d_j (\alpha_{j-1} - \beta_{j-1})}{\sqrt{\frac{2\sigma^2 + \sum d_j^2 (\alpha_{j-1}^{*2} + \beta_{j-1}^{*2})}{n}}} \right).$$

윗 식에서 마지막 항의 형태는 중도탈락을 고려하지 않은 일반적인 유도식에서도 n이 커짐에 따라 작은 값이 나와 무시하였다. 중도탈락을 고려한 본 논문에서도 다음 장의 표 3.1의 현실적인 기준을 적용하였을 때 n이 10 정도만 되더라도 마지막항의 값은 0.001보다 작게 되어 편의상 무시하기로 한다. 이에 따라 검정력을 다시 정리하면 다음과 같다.

$$1-\beta \approx 1-\Phi\left(\frac{z_{\frac{\alpha}{2}}\sqrt{\frac{2\sigma^2}{n}}-\sum_{i=1}^{l+1} d_i(\alpha_{i-1}-\beta_{i-1})}{\sqrt{\left(\sigma^2+\sum_{i=1}^{l+1} d_i^2(\alpha_{i-1}^{*2}+\beta_{i-1}^{*2})\right)/n}}\right).$$

표본수 n에 관한 식을 얻기 위해 다음과 같이 표준정규분포의 z-점수를 이용한다.

$$\frac{z_{\frac{\alpha}{2}}\sqrt{\frac{2\sigma^2}{n}}-\sum_{i=1}^{l+1} d_i(\alpha_{i-1}-\beta_{i-1})}{\sqrt{\left(2\sigma^2+\sum_{i=1}^{l+1} d_i^2(\alpha_{i-1}^{*2}+\beta_{i-1}^{*2})\right)/n}} = \Phi^{-1}(\beta) = -z_\beta.$$

그러므로 최종적으로 필요한 군별 표본수는 다음과 같다.

$$n = \left\{z_{\frac{\alpha}{2}}\sqrt{2\sigma^2}+z_\beta\sqrt{2\sigma^2+\sum_{i=1}^{l+1} d_i^2(\alpha_{i-1}^{*2}+\beta_{i-1}^{*2})}\right\}^2 \Big/ \left\{\sum_{i=1}^{l+1} d_i(\alpha_{i-1}-\beta_{i-1})\right\}^2, \quad (3.3)$$

이를 중도탈락을 고려하지 않은 표본수 산출식 (3.4)와 비교하면 분모가 감소하고 분자는 증가하기 때문에 당연히 식 (3.4)에 의한 n^*보다 식 (3.3)의 표본수 n이 큰 값이다.

$$n^* = 2\sigma^2\frac{\left(z_{\frac{\alpha}{2}}+z_\beta\right)^2}{(\alpha_l-\beta_l)^2}. \quad (3.4)$$

그러나 중도탈락률을 고려하여 식 (3.4)에 일정 비율을 곱한 $n^{**}=(1-d)^{-1}n^*$를 최종피험자수로 결정하기 때문에 n과 n^{**}를 비교하는 것이 의미가 있을 것이다. 표 3.1에 특정값에 따른 피험자수 n, n^*, n^{**} 등과 각 검정력을 계산하였다. 반복측정시점의 회수는 $l=5$로 고정하고 각 시점에서 탈락률은 일정하며 $(d_i=d/l,\ i=1,\ldots,l)$ 처리효과는 각 시점에서 일정하게 비례 $(\alpha_i=\delta\cdot i,\ \beta_i=(\delta+\gamma)\cdot i,\ (i=1,\ldots,5))$하는 경우이고 유의수준은 5%, 검정력은 80%로 정하였다.

표 3.1에 의하면 중도탈락 발생으로 인하여 최초에 원하는 검정력을 확보하기 위해서는 기존의 방법보다 약 10% 가량의 표본수가 더 추가되어야 함을 알 수 있다. 예를 들어 $d=0.5$, $\sigma=2.5$ 인 경우 중도탈락이 있을 경우 식 (3.3)의 표본수 n으로는 0.8의 검정력이 보장되지만 기존의 방법인 n^{**}로는 0.76의 검정력이 보장될 뿐이다. 같은 검정력을 얻기 위해서는 n^{**}보다 12%가 많은 피험자수가 필요하다. 모든 경우에 n보다 n^{**}가 많았지만 상대적으로 표본수가 많은 경우에는 5% 정도의 피험자수 추가만이 필요하였다.

표 3.1: 중도탈락을 고려한 표본수 비교 (n: 중도탈락을 포함한 수식으로 유도한 표본수, n^*: 중도탈락을 고려하지 않은 일반식으로 유도한 표본수, n^{**}: 일반식에 의해 구한 표본수에 중도탈락비율만큼 추가한 표본수)

$d = 0.5,\ \delta = 0.5,\ \gamma = 0.5$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
2.5	15.70	31.40	35.01	112%	0.8	0.760
3.0	22.60	45.21	49.15	109%	0.8	0.769
3.5	30.77	61.54	65.83	107%	0.8	0.774
4.0	40.19	80.37	85.08	106%	0.8	0.778
4.5	50.86	101.72	106.87	105%	0.8	0.781
5.5	62.79	125.58	131.23	104%	0.8	0.783

$d = 0.2,\ \delta = 0.5,\ \gamma = 0.5$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
2.5	15.70	19.62	21.36	109%	0.8	0.768
3.0	22.60	28.26	30.29	107%	0.8	0.773
3.5	30.77	38.46	40.83	106%	0.8	0.777
4.0	40.19	50.23	53.00	106%	0.8	0.779
4.5	50.86	63.58	66.79	105%	0.8	0.781
5.0	62.79	78.49	82.20	105%	0.8	0.782

d: 탈락률, $\delta, (\delta + \gamma)$는 각 시점에서 평균변화량, σ: 표준편차

4. 일원배치 분산분석에서 피험자수 결정

3장의 두 집단 평균비교를 여러 개의 평균비교인 일원배치 분산분석으로 확대하면 다음과 같다. 집단의 수가 m이고, 각 집단에서 관측수는 n으로 동일하고, l회 반복측정하였다고 가정하면 각 표본의 경시적 모형은 다음과 같다.

$$x_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu_0 + \alpha_{ij} + \epsilon_{ijk}, \quad i = 0, 1, \ldots, l;\ j = 1, \ldots, m;\ k = 1, \ldots, n,$$

이때, $\epsilon_{ijk} \sim N(0,\ \sigma^2)$이라 가정한다. 또한 각 시점에서 중도탈락률을 $d_i (i = 1, \ldots, l)$, 각 집단에서 전체 중도탈락률은 $d (= \sum_{i=1}^{l} d_i)$로 동일하다고 가정한다. 그러면 식 (3.1)과 유사한 방식으로 관측된 값과 LOCF에 의해 대체된 값으로 이루어진 최종분석자료값 X^*은 다음과 같이 표시된다.

$$X^*_{jk} = \begin{cases} x_{ljk}, & t > l, \\ x_{t-1,jk}, & t \le l, \end{cases} \quad j = 1, 2, \ldots, m;\ k = 1, 2, \ldots, n;\ t = \text{중도탈락시점}.$$

그리고 분산은 3장에서 구한 것과 동일한 형태로 다음과 같다.

$$\mathrm{Var}(X^*_{jk}) = \sigma^2 + \sum_{i=1}^{l+1} d_i \alpha^{*2}_{i-1,j} = \sigma^2_j, \tag{4.1}$$

여기서, $\alpha*_{i-1,j} = \sum_{k=1}^{l+1} d_k \alpha_{k-1,j} - \alpha_{i-1,j},\ j = 1, \ldots, m;\ k = 1, \ldots, n.$ 그리고 여러 집단 평균 동일성 검정의 귀무가설은 다음과 같다.

$$H_0 :\ \mu_{l1} = \mu_{l2} = \cdots = \mu_{lm} = \mu\ (\alpha_{l1} = \cdots = \alpha_{lm}).$$

이러한 귀무가설을 검정할 때 본 연구에서는 Welch (1951) 검정법을 사용하도록 한다. 이는 검정력을 계산함에 있어 식 (4.1)의 제약조건으로 등분산 가정이 위배되어 이분산을 가정한 비중심(non-central) F분포를 고려해야하기 때문이다. 각 집단의 표본평균을 $\bar{x}_j\ (j = 1, \ldots, m)$라 하고 전체 평균을 $\bar{x}$라 하면 Welch 검정법은 다음과 같이 유의수준 α로 귀무가설을 기각한다.

$$\frac{q_w}{(m-1)} \left\{ 1 + \frac{2(m-2)A}{(m^2-1)} \right\}^{-1} \geq\ F(\nu_1, \nu_2;\ 1-\alpha),$$

여기서, $q_w = \sum^m w_j(\bar{x}_j - \bar{x})^2$는 Welch의 검정통계량이고, $w_j = n_j/\sigma_j^2,\ j = 1, \ldots, m,$ $A = \sum^m \{1 - (w_j/\sum w_i)\}^2/(n_j - 1),\ \nu_1 = m - 1,\ \nu_2 = (m^2-1)/3A.$

Welch (1951)와 James (1951)는 각 집단의 분산이 다를 경우의 검정통계량을 제안하고 귀무가설 하에서 이 검정통계량의 분포를 유도하였는데 대립가설 하에서 검정력을 계산하기위해서 Hedges와 Pigott (2001)가 적률(moment)로 근사시킨 방법을 제안하였다. 그러나 실제로 계산해 본 결과 이 방법이 만족할 만한 근사를 보이지 않아 본 논문에서는 Hedges와 Pigott 방법을 수정한 Kulinskaya 등 (2003)이 제안한 좀 더 정밀한 근사방법을 사용하도록 하겠다. 대립가설 하에서 q_w의 분포를 구하기 위해 Kulinskaya 등 (2003)은 q_w를 카이제곱분포를 따르는 변수의 일차함수로 근사시켜서 검정력을 다음과 같이 표현하였다.

$$\text{검정력} = \Pr\left(\chi_f^2 \geq\ \frac{q-b}{c} \right), \tag{4.2}$$

여기서, $q = (m-1)\{1+2(m-2)A/(m^2-1)\}F(\nu_1, \nu_2;\ 1-\alpha),\ b = k_1 - 2k_2^2/k_3,\ c = k_3/4k_2,$ $f = 8k_2^3/k_3^2.$ 그리고 $k_1 = E(q_w) = m-1+\lambda+2\gamma+2\delta,\ k_2 = \mathrm{Var}(q_w) = 2(m-1+2\lambda+7\gamma+14\delta+\upsilon),\ k_3 = E[q_w - E(q_w)]^3 = 8(m-1+3\lambda+15\gamma+45\delta+6\upsilon+2\nu),$ 여기서, λ는 비중심 모수로서 $\lambda = \sum w_j(\mu_j - \mu)^2$이고 $(\mu_j = \mu_{lj},\ j = 1, \ldots, m),\ \gamma = \sum(1 - w_j/\sum w_i)^2/(n_j - 1),$ $\delta = \sum w_j(\mu_j - \mu)^2(1 - w_j/\sum w_i)/(n_j - 1),\ \upsilon = \sum w_j^2(\mu_j - \mu)^4/(n_j - 1),\ \nu = \sum w_j^3(\mu_j - \mu)^6/(n_j - 1)^2$ 등이다.

또한 식 (4.1)에서 분산이 달라지기 때문에 각 집단의 가중치를 다음과 같이 정의한다.

$$w_j = \frac{n_j}{\sigma^2 + \sum d_i \alpha_{i-1,j}^{*2}} = n_j \tau_j, \qquad j = 1, \ldots, m.$$

각 그룹간에 표본수가 n으로 동일하다 $(n_j = n,\ j = 1, \ldots, m)$고 가정하고, 탈락률이 일정하고 평균차이가 등간격이라 가정하면 식 (4.2)에서 유도한 다음과 같은 등식을 통해

n의 값을 구할 수 있다.

$$(m-1)\left\{1+\frac{2(m-2)C}{(n-1)(m^2-1)}\right\}F\left(m-1,\ \frac{(n-1)(m^2-1)}{3C};\ 1-\alpha\right)$$

$$= \left\{m-1+nB+\frac{2C}{(n-1)}+\frac{2nD}{(n-1)}\right\}$$

$$-\frac{\left\{m-1+2nB+7\frac{C}{(n-1)}+14n\frac{D}{(n-1)}+n^2\frac{E}{(n-1)}\right\}^2}{\left\{m-1+3nB+15\frac{C}{(n-1)}+45n\frac{D}{(n-1)}+6n^2\frac{E}{(n-1)}+2n^3\frac{F}{(n-1)}^2\right\}}$$

$$+\frac{\left\{m-1+3nB+15\frac{C}{(n-1)}+45n\frac{D}{(n-1)}+6n^2\frac{E}{(n-1)}+2n^3\frac{F}{(n-1)}^2\right\}}{\left\{m-1+2nB+7\frac{C}{(n-1)}+14n\frac{D}{(n-1)}+n^2\frac{E}{(n-1)}\right\}}\cdot$$

$$\chi^2\left[\frac{\left\{m-1+2nB+7\frac{C}{(n-1)}+14n\frac{D}{(n-1)}+n^2\frac{E}{(n-1)}\right\}^3}{\left\{m-1+3nB+15\frac{C}{(n-1)}+45n\frac{D}{(n-1)}+6n^2\frac{E}{(n-1)}+2n^3\frac{F}{(n-1)}^2\right\}^2};\beta\right],$$

여기서, $B=\sum\tau_j(\mu_j-\mu)^2$, $C=\sum(1-\tau_j/\sum\tau_j)^2$, $D=\sum\tau_j(\mu_j-\mu)^2(1-\tau_j/\sum\tau_i)$, $E=\sum\tau_j^2(\mu_j-\mu)^4$, $F=\sum\tau_j^3(\mu_j-\mu)^6$.

이장에서도 제안하는 n과 기존의 n^{**}를 비교하기위해 표 4.1에 특정값에 따른 피험자수와 검정력을 통계언어 R을 이용하여 계산하였다. 반복측정시점의 회수는 $l=5$로 고정하고 각 시점에서 탈락률은 일정하며 $(d_i=d/l,\ i=1,\ldots,l)$ 처리효과는 각 시점에서 일정하게 비례 $(\alpha_i=\delta\cdot i),\ \beta_i=(\delta+\gamma)\cdot i,\ i=1,\ldots,5)$하는 경우이고 유의수준은 5%, 검정력은 80%로 정하였다.

표 4.1에 의하면 중도탈락에도 불구하고 원하는 검정력을 확보하기 위해서는 기존의 방법보다 약 10 ~ 20% 가량의 표본수가 더 추가되어야 함을 알 수 있다. 예를 들어 $d=0.1,\ \sigma=3$인 경우 중도탈락이 있을 경우 표본수 n으로는 0.8의 검정력이 보장되지만 기존의 방법인 n^{**}로는 0.697의 검정력만이 보장될 뿐이다. 같은 검정력을 얻기 위해서는 n^{**}보다 24%가 많은 피험자수가 필요하다. 상대적으로 표본수가 많은 경우에는 3~7% 정도의 피험자수 증가만이 필요하였다.

5. 결론

본 논문에서는 피험자수를 결정할 때 계산식에 중도탈락을 고려한 크기를 사후적으로 더해주는 기존의 방법이 우리가 원하는 검정력을 보장해주지 못할 수 있음을 지적하고 계산식을 구하는 단계에서부터 중도탈락을 고려할 것을 제안하였다. 특히 여러 평균의 동일성 검정에서 중도탈락이 포함된 표본수 산출식을 유도하고 기존의 산출식과 표본수와 검정력 등을 비교하였다.

표 4.1: 중도탈락을 고려한 표본수 비교 (n: 중도탈락을 포함한 수식으로 유도한 표본수, n^*: 중도탈락을 고려하지 않은 일반식으로 유도한 표본수, n^{**}: 일반식에 의해 구한 표본수에 중도탈락비율만큼 추가한 표본수)

$d = 0.5,\ \delta = 0.5,\ \gamma = 0.5,\ m = 3$
$(\mu_{l1} = 2.5,\ \mu_{l2} = 5,\ \mu_{l3} = 7.5)$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
3	8.44	16.88	20.88	124%	0.8	0.697
5	20.81	41.62	46.53	112%	0.8	0.749
7	39.20	78.40	84.19	107%	0.8	0.769

$d = 0.2,\ \delta = 0.5,\ \gamma = 0.5,\ m = 3$
$(\mu_{l1} = 2.5,\ \mu_{l2} = 5,\ \mu_{l3} = 7.5)$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
3	8.44	10.55	12.51	119%	0.8	0.715
5	20.81	26.01	28.56	110%	0.8	0.757
7	39.20	49.00	52.32	107%	0.8	0.771

$d = 0.5,\ \delta = 0.5,\ \gamma = 0.5,\ m = 5$
$(\mu_{l1} = 2.5,\ \mu_{l2} = 5,\ \mu_{l3} = 7.5,\ \mu_{l4} = 10.0,\ \mu_{l5} = 12.5)$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
10	20.88	41.76	44.16	106%	0.8	0.773
15	44.71	89.42	92.77	104%	0.8	0.783
20	77.99	155.98	160.65	103%	0.8	0.786

$d = 0.2,\ \delta = 0.5,\ \gamma = 0.5,\ m = 5$
$(\mu_{l1} = 2.5,\ \mu_{l2} = 5,\ \mu_{l3} = 7.5,\ \mu_{l4} = 10.0,\ \mu_{l5} = 12.5)$

σ	n^*	n^{**}	n	n/n^{**}	정확한 검정력(n)	실제 검정력(n^{**})
10	20.88	26.10	27.59	106%	0.8	0.773
15	44.71	55.89	58.32	104%	0.8	0.780
20	77.99	97.49	101.26	104%	0.8	0.782

중도탈락된 개체를 탈락직전의 값으로 대체하는 LOCF 시험의 경우에 예비조사에서 추정된 관측값의 평균, 분산 등이 달라짐에 따라 기존 산출방법의 피험자수로는 원하는 검정력을 확보할 수 없을 수 있다. 이에 반하여 제안한 방법은 중도탈락에 의한 평균, 분산 등의 변화를 표본수 산출식에 포함시켜 줌으로써 원하는 검정력을 얻을 수 있었다. 제안한 방법을 고려함으로써 2장에서와 같이 원하는 관측결과를 얻고도 표본수 부족으로 통계학적으로 의미 있는 결론을 얻지 못하는 예를 막을 수 있을 것이다. 그런데 일반적으로 임상시험 연구자들이 피험자수를 산출하기 위하여 PASS나 SAS와 같은 전문통계 패키지를 사용하기 때문에 비통계전문가가 제안한 수식을 이용하여 피험자수를 결정하는데는 무리가 있을 수 있다. 그러나 제안한 산출식에 의한 표본수가 기존의 표본수보다 대략적으로 5%~20% 정도 많기 때문에 이정도 표본수를 추가하는 정도의 주의를 기울이는 것도 의미가 있을 것이다. 향후 중도탈락을 고려한 표본수 산출 프로그램이 작성되어 배포된다면 일

반 연구자들에게 큰 도움이 될 것이다.

참고문헌

Altman, D. G. (1991). *Practical Statistics for Medical Research*, Chapman & Hall/CRC, London.

Altman, D. G., Machin, D., Bryant, T. N. and Gardner, M. J. (2000). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2^{nd} Edition,* London: British Medical Journal.

Hedges, L. V. and Pigott, T. D. (2001). The power of statistical tests in meta-analysis, *Psychological Methods,* **6**, 203–217.

Hoeing, J. M. and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis, *The American Statistician,* **55**, 19–24.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown, *Biometrika,* **38**, 324–329.

Kulinskaya, E., Staudte, R. G. and Gao, H. (2003). Power approximations in testing for unequal means in one-way ANOVA weighted for unequal variances, *Communications in Statistics: Theory and Methods,* **32**, 2353–2371.

Lee, K. H. (2006). A study on one factorial longitudinal data analysis with informative drop-out, *Journal of the Korean Data & Information Science Society,* **17**, 1053–1065.

Machin, D., Campbell, M., Fayers, P. and Pinol, A. (1997). *Sample Size Tables for Clinical Studies, 2^{nd} Edition,* Blackwell Science, London, Edinburgh, Malden and Carlton.

Moher, D., Dulberg, C. S. and Wells, G. A. (1994). Statistical power, sample size and their reporting in randomized controlled trials, *The Journal of the American Medical Association,* **272**, 122–124.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion, *Statistics in Medicine,* **5**, 1–13.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach, *Biometrika,* **38**, 330–336.

[2008년 2월 접수, 2008년 3월 채택]

Sample Size Calculations with Dropouts in Clinical Trials

Ki Hoon Lee[1)]

Abstract

The sample size in a clinical trial is determined by the hypothesis, the variance of observations, the effect size, the power and the significance level. Dropouts in clinical trials are inevitable, so we need to consider dropouts on the determination of sample size. It is common that some proportion corresponding to the expected dropout rate would be added to the sample size calculated from a mathematical equation. This paper proposes new equations for calculating sample size dealing with dropouts. Since we observe data longitudinally in most clinical trials, we can use a last observation to impute for missing one in the intention to treat(ITT) trials, and this technique is called last observation carried forward(LOCF). But LOCF might make deviations on the assumed variance and effect size, so that we could not guarantee the power of test with the sample size obtained from the existing equation. This study suggests the formulas for sample size involving information about dropouts and shows the properties of the proposed method in testing equality of means.

Keywords: Clinical trials; sample size; LOCF; ITT.

1) Professor, School of Business, Jeonju University, Hyoja-dong 3Ga, Wansan-gu, Jeonju, Jeonbuk 560-759, Korea. E-mail: khlee@jj.ac.kr