基礎統計教育教材

1980

經濟企劃院 調査統計局

目　　次

第一編　統　計　入　門

第一章 序 論

우리나라의 統計年鑑에 収録되있는 것을 보면 25가지의 大分類 統計와 326가지의 統計表가 収録되있다. 이러한 모든 統計는 直·間接으로 모두 統計局의 影響을 받지 않을 수 없으며 統計局員들은 이러한 統計의 直接生産者인 同時에 総括者인 것이다.

이러한 尨大한 統計를 生産하기 위해서는 人的, 物的 装備와 이것을 效果的으로 構成 投入하는 機構가 必要하게 된다. 즉 사람과 돈과 組織이 될 것이다. 더 나아가서 有能한 人材와 豊富한 돈과 效率的인 組織이 있어야 바람직한 統計가 生産될 것이다.

1.1 바람직한 統計란?

어떤 統計가 바람직 하겠는가? 誤差없는 統計가 生産되는 것이 바람직 할 것이다. 물론 誤差라는 槪念을 여러가지로 定義가 다르지만 여기서는 標本誤差와 非標本誤差로 区分하여 생각해 보기로 하자. 標本誤差란 部分을 調査하여 全体를 알고자 하는데서 発生하는 必然的인 情報不足에서 発生하는 誤差며, 非標本誤差란 全部를 調査하더라도 생기는 誤差를 말한다. 특히 標本誤差는 調査方法이나 実験設計에 의해 糾明 혹은 統制되지만 非標本誤差는 그렇지 않은 경우가 흔히 있다. 특히 非標本誤差의 問題는 全数調査의 質的인 問題로써 이것은 社会的, 政治的으로 物議를 빚는

경우도 있을수 있다. 이 非標本誤差의 原因으로는 여러가지를 들 수 있을 것이다. 그러나 大体的으로 모든것에서 그 原因을 찾을 수 있다고 하겠다. 예를 들면 「人口 및 住宅센서스」에서 女子의 年齡같은 項目은 우리나라 伝統的으로 女子들은 나이를 正確히 대기를 좀 꺼려하는 慣習이 있다. 이러한 習慣이나 伝統도 非標本誤差의 原因이 되겠다.

바람직한 統計로서는 위에서 設明한 바와같이 우선 誤差가 작아야 할 것이다. 또한 費用이 可及的이면 最小로 드는 것도 빼놓을 수 없는 것이다. 費用이 적을뿐 아니라 時間도 적게들고 最小의 調査로 最大의 效果를 얻을 수 있다면 바람직스러운 統計가 될 것이다.

1.2 어떻게 調査하겠는가?

우선 全部調査를 하느냐 部分調査를 하느냐는 주어진 与件(費用, 時間, 人員…)下에서 可能한 調査가 무엇이냐 또 目的이 어디에 있느냐에 따라 決定하게 될 것이다. 많은 指定統計가 全數調査를 한 것이고 統計年鑑에 収録된 또한 많은 統計가 標本調査를 한 것이다.

標本調査를 할 경우에는 客観的인 調査法(確率調査)과 主観的인 調査法이 있다 하겠다. 前者는 全体母集団에 관한 어떤 推論 혹은 意思決定에 目的을 둔 것이고 後者는 部分이라는 標本自体의

技術 혹은 意思決定으로 끝나게 된다.

客觀的 標本調査는 주어진 与件下에 最善의 目的을 達成하기 위해서는 어떤 原理에 立脚해서 設計, 進行, 分析이 이루어 져야 할 것이다. 이러한 原理를 通称한다면 바로 統計学의 原理라 하겠다.

1.3 資料의 分析

모든 統計表는 調査項目의 組合에 관한 內容을 表示하고 있다고 볼 수 있다. 이 調査項目은 알고자 하는 目的을 具体的으로 表示한 것이라 하겠다. 이 調査項目은 일종의 統計的変数를 나타낸다고 볼 수 있다. 이 変数를 連続変数와 離散変数로 나누어 생각하는 것이 보통이다. 또한 이들 変数의 測定 혹은 調査된 水準을 測定値의 水準 혹은 測定値의 尺度라고 부른다. 測定値의 尺度는

(1) 名目尺度

(2) 順序尺度

(3) 区間尺度

(4) 比率尺度

로 나눈다. 위의 名目尺度와 順序尺度는 非母数統計学 (Non-parametric statistics) 의 大体的인 客体이며 母数統計学은 区間尺度와 比率尺度를 主対象으로 한다.

위의 4가지 尺度로 되어 있는 統計表를 가지고 먼저 簡単히

要約을 하는 方法으로는 代表値와 分散度, 의도, 尖度등 特性値를 計算하거나 推出해 내는 作業일 것이다.

代表値는 여러가지 平均値나 位数등을 들수 있으며 分散度에는 範囲나 分散槪念들이 있고 의도나 尖度는 3次, 4次 積率槪念들로 表示되는 것이 보통이다.

또다른 要約方法으로는 度数分布를 그리거나 度数分布를 函数形態로 導出하는 것이라 하겠다.

度数分布는 確率的인 槪念인 相対度数로서 적절히 表現되는데 確率이라는 것은 반드시 相対度数의 槪念만은 아니다. 예컨데 確率을 相対度数로 보는 見解도 있고 相対度数의 極限槪念으로 보는 見解 또는 測度論 (measure theory) 의 立場으로 또는 믿음의 정도 (degree of belief) 등 多様한 見解가 있다. 하여튼 資料를 要約하는 方法의 하나로 度数分布 혹은 頻度函数도 要約한다면 統計資料의 綜合的인 要約方法이라 하겠다.

어떤 方法으로든 要約된 統計資料를 가지고 마지막 推論分析 혹은 意思決定過程을 客観的으로 妥当한 最適決定을 한다는 것은 普通 古典的 統計学에서는 推定 検定이라한다. 그러나 現代統計学에서는 不確定的現象에 관한 意思決定의 하나로 看做하면 구태여 推定原理와 検定原理가 따로 있는것이 아니고 不確定的인 現象下에서 意思決定原理로 統合된다고 볼 수 있다.

그러나 本講義에서는 편의상 推定 検定을 分離하여 設明하기로 한다. 끝장에서 2変数이상의 分析例로서 回帰分析을 簡単히 設明한다.

第 2 章　統計的　推定

本章에서는 統計的 推定에 대하여 다루게 되는데 1節에서는 推定에 대한 定義에 대하여, 2節에서는 바람직한 推定量 (Good Estimation) 에 관하여, 3節에서는 古典的, 統計的 推定에 대하여 略述하는데 小節로는 点推定과 区間推定을 나누어서 說明하고 4節에서는 Bayesian方法에 대하여 說明하고자 한다.

2.1 推定의 定義

確率変数 X의 分布函数의 형태를 어떤 이유로 해서 알고 있는지 혹은 이럴것이라고 仮定할 수 있는 경우도 그것은 몇개의 未知母数를 包含하고 있는것이 보통이다. 만약 母数의 값을 모두 알고 있다면 分布函数는 確定되므로 推測 統計的 研究의 必要性이 반감된다. 그래서 標本을 利用하여 母集団母数 (未知인) 의 값을 얻을때 이 얻는과정을 推定 (Estimation) 이라 하고 이러한 推定을 통하여 未知인 母集団母数를 推定하기 위하여 標本을 観測하였을때 이 観測에서 얻은 어느 특정한 값을 推定値 (Estimate) 라 하며 推定하는데 있어서 標本이 数字로 特定되기 전에 任意標本 (Random Sample) 과 函数関係를 갖는 것을 推定量 (Estimator) 이라 한다. 가령 任意標本을 $X_1\ X_2 \cdots X_n$ 이라 하면 平均인 $\bar{X}$ 를 이 標本이 抽出된 母集団의 平均인 μ 의 推定量이라

한다. 一般的으로 母集団의 確率密度函数(p·d·f) $f(x,\theta_1 \cdots \theta_K)$의 모양은 알고 있으나 그것은 ℓ ($\leq K$)개의 未知母数 $\theta_1 \cdots \theta_\ell$을 包含하고 있다. 이때 標本 $x_1\ x_2 \cdots x_3$에 관한 ℓ개의 1個函数 $ti(x_1 \cdots x_n)$, $i=1, 2 \cdots \ell$를 취하고 그 값을 θ_i의 推定値로 삼는 것이 点推定(Point Estimation)이다. 예를들면 推定에서 単一한 값을 얻는 標本平均, 標本標準偏差등을 母平均, 母集団偏差등의 点推定이라 한다. 区間推定은 推定値의 区間 〔a,b〕를 지정하고 未知의 母数 θ가 그 속에 包含된다고 하는 형식으로 θ를 推定하려는 것이다. 따라서 推定에는 点推定과 区間推定이 있다.

2.2 바람직한 推定量

推定量의 性質은 小標本(Small Sample) 혹은 有限標本(finite sample)에 대한 것과 大標本(large sample)에 대한 것으로 구분된다. 統計的 推定에 바람직한 推定量은 그 分布가 真正한 母数(true parameter)에 가능한 한 集中되어 있는 것이라 하겠다. 그래서 推定量의 分布에서 그 위치와 모양을 말해 주는 平均과 分散이 推定量의 性質을 평가하는데 대한 対象이 되고 바람직한(좋은) 推定量인가에 대한 決定基準은 小標本인 경우 (1) 不偏性(unbiasedness) (2) 最小分散(minimum variance) 등이고 大標本인 경우 (1) 不偏性 (2) 一致性(consistency)

(3) 有效性(Efficiency)등이다. 그러면 小標本에 대한 推定量의 性質에 대하여 먼저 알아 보기로 한다. 바람직한 推定量의 性質은 標本의 크기를 점점 크게 함에 따라 推定値가 母数값에 수렴하는 성질일 것이다.

a) 不偏性 (Unbiasedness)

만약 $E(t)=\theta$ 라면 $t=t(x_1, x_2 \cdots x_n)$ 는 θ의 不偏推定量(推定値)이다. 이것은 確率変数 t가 推定하려는 母数 θ를 平均으로 갖는 分布에 따른다는 것이다. 예를들면 $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot nE(X) = E(X) = \mu$다. 이것을 보면 標本平均은 母平均의 不偏推定量이 된다. 또 다른 예로서 어떤 統計量에 있어서 標本의 크기 n인 標本分散의 期待値를 살펴보자. 標本分散이 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ 이라 하자.

이때 S^2의 期待値는

$$
\begin{aligned}
E[S^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}[(x_i-\mu) - (\bar{x}-\mu)]^2\right\} \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E(x_i-\mu)^2 - E(\bar{x}-\mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \sigma^2 - \sigma_{\bar{x}}^2 \\
&= \sigma^2 - \frac{\sigma^2}{n}
\end{aligned}
$$

이것은 S^2이 σ^2의 不偏推定値가 아님을 말해 주고 있다. 즉 $\frac{\sigma^2}{n}$라는 偏倚性 (bias)이 생긴다. 따라서 不偏推定値를 구하고자 할때에는 母分散推定에 있어서 標本을 결합하는 方法에 주의하지 않으면 안된다. S^2의 偏倚性을 克服하기 위하여 $\frac{n}{n-1}$을 곱한 것의 期待値를 취하게 되는 것이다. 즉 다음과 같이 된다.

$$E\left[\frac{n}{n-1}S^2\right] = \left(\frac{n}{n-1}\right)\cdot\left(\frac{n-1}{n}\right)\cdot\sigma^2 = \sigma^2$$

S^2가 σ^2의 不偏推定量 $\hat{\sigma}^2$와 같게 되려면 S^2에 $\left(\frac{n}{n-1}\right)$을 곱하여야 한다.

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

따라서 $\hat{\sigma} = S$가 되려면

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

이 되어야 한다. 즉 標本分散 S^2는 母分散 σ^2의 不偏推定量이다.

b) 最小分散 (minimum variance)

不偏性의 性質이 推定에 있어서 하나의 바람직한 性質이기는 하나 그것보다 더 중요한 것은 어떤 의미에서 推定하려는 母数에 접근해 간다는 性質이다.

따라서 두개의 統計量 t, t'가 있을때 t는 θ에 접근해 가는 統計量이고 t'는 그렇지 않으며 또한 t는 偏向性이 있고 t'는 그렇지 않고 不偏推定量이라 하면 母數 θ를 推定할때 推定量으로서 t를 택하게 된다. 어떤 합리적인 접근성의 정의에 대해 두 推定値중에 어느것이 더 母數 θ에 접근하고 있느냐를 決定하는 것이 불가능하기 때문에 θ의 접근 정도 대신에 t의 変異度를 계산한다. 이 変異정도를 계산하는 尺度로서 分散이나 標準偏差를 使用한다. t_1, t_2가 서로 비교하고자 하는 θ의 推定値라 하자. t_1가 t_2보다 良好하려면 모든 가능한 θ의 값에 대하여 $E(t_1-\theta)^2 \leq E(t_2-\theta)^2$이어야 하며 적어도 하나의 θ의 값에 대하여는 부등호가 성립해야 한다. 즉 $\mathrm{Var}(\hat{\theta}) \leq \mathrm{Var}(\hat{\hat{\theta}})$

但 $\hat{\hat{\theta}}$는 $\hat{\theta}$外의 推定量

이렇게 되면 $\hat{\theta}$는 다른 어떠한 推定量중에서도 가장 작은 分散을 갖는 推定量이다. 만약 θ의 推定量인 $\hat{\theta}$가 不偏推定値이면서 最小分散을 갖는다면 이 推定量 $\hat{\theta}$는 母数 θ의 最良不偏推定値가 된다. 만약 推定量 $\hat{\theta}$가 標本観測의 線型函数이고 동시에 最良不偏推定値이면 $\hat{\theta}$는 θ의 最良線型不偏推定量 (the best linear unbiased estimator, blue)이 된다. 예를 들면 標本平均 $\bar{X}$가 標本観測의 線型函数이고 母平均 μ의 最小分散을 가진 不偏推定量이면 이 標本平均 $\bar{X}$는 母平均 μ의 BLUE이다. 大標本 (漸近標本이라고 함)의 경우 推定量의 性質도 小標本의 경우와 비슷하기도 하나 標本의 크기가 커짐에 따라 당해 推定量의 分布가 점점 어느

特定한 分布로 接近하게 되고 이렇게 된 分布를 그 推定量의 漸近分布(asymptotic distribution)라 한다. 漸近分布도 역시 積率(moments)에 의하여 특정지어 지는데 그 중에서도 漸近平均(asymptotic mean)과 漸近分散(asymptotic variance)이 가장 중요한 것이다. 다음과 같은 推定量의 漸近性質이 있다.

a. 漸近的 不偏性(asymptotic unbiasedness)

만약 $\lim_{n\to\infty} E(\hat{\theta}) = \theta$

단, $\lim_{n\to\infty}$ = 標本의 크기 n이 無限大로 접근함에 따른 極限

이면 $\hat{\theta}$는 θ의 漸近的 不偏推定量이 된다. 만약 어느 推定量이 不偏推定値이면 이 推定量은 역시 漸近的 不偏이 된다. 그러나 漸近的 不偏推定量은 반드시 不偏推定値가 되는 것이 아니다. 예를들면 標本分散이 이러하다. n이 無限大로 접근함에 따라 $\frac{n-1}{n}$이 1로 되기 때문이다. 즉

$$E(S^2) = \frac{1}{n} E\left[\sum_{i=1}^{n} (t_i - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n} \sigma^2 - \sigma_{\bar{x}}^2$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

에서 偏倚 $\frac{\sigma^2}{n}$가 0으로 된다.

b) 一致性(consistency)

標本의 크기 n이 無限大에 접근함에 따라 推定量 $\hat{\theta}$의 分布가 한 点으로 収斂하게 될때 이 点을 $\hat{\theta}$의 確率極限(proba-

bility limit)라 하는데 Plim로 表記된다. 즉, 다음과 같은 것을 말한다.

$$\lim_{n\to\infty} P(\theta^* - \varepsilon \leq \hat{\theta} \leq \theta^* + \varepsilon) = 1$$

단 θ^*는 θ와 같을 수도 다를 수도 있는 어느 点. ε = 작을 수 있는데까지 작아진 어떤 仮想的인 陽의 数

만약 $P\lim_{n\to\infty}(|\hat{\theta} - \varepsilon| \leq \varepsilon) = 1$ 또는

$P\lim_{n\to\infty}(|\hat{\theta} - \theta| > \varepsilon) = 0$이 성립하면 이 $\hat{\theta}$를 θ의 一致推定量이라 한다. 하나의 推定量이 不偏인지 아닌지를 아는 것은 비교적 쉬운일이고 두 不偏推定量의 分散을 비교하는 것은 아주 간단한 일이다. 그러나 一致性을 증명하기는 쉬운일이 아니다. 다음 식은 때로는 아주 유용하게 사용된다.

$\hat{\theta} = t(X_1, X_2 \cdots X_n)$을 母数 θ의 推定量이라 할 때

$\lim_{n\to\infty} E(\hat{\theta}) = \theta$이고 $\lim_{n\to\infty} V(\hat{\theta}) = 0$이면 $\hat{\theta}$는 θ의 一致推定量이다.

c) 漸近的 有效性 (asymptotic efficiency)

만약 $\hat{\theta}$가 θ의 一致推定量이면서 다른 어떤 一致推定量보다 가장 작은 分散을 갖고 있고 有限平均과 有限分散을 갖고 漸近的 分布를 하고 또는 推定量이면 이것을 θ의 漸近的 有效推定量이라 한다.

예를 통하여 바람직한 推定量을 알아보자. 전구를 만드는데 있어서 전구에서 抽出한 n개의 標本 ($X_1, X_2 \cdots X_n$)에 K개의

不良品이 들어있다: 즉 $Y=\sum_{i=1}^{n} X_n = K$라 한다. 즉 i번째로 抽出한 전구가 不良品이면 $X_i = 1$, 良品이면 $X_i = 0$, $i = 1, 2 \cdots n$로 나타내면 不良品의 数 Y는 二項分布를 따른다고 한다. 母数(不良比) P의 推定量으로 標本의 不良比 $\hat{p} = \frac{Y}{n}$를 취하면

$E(\hat{p}) = E(\frac{Y}{n}) = \frac{1}{n} np = P$ 이므로 $\hat{p}$는 不偏推定量이다.

$V(\hat{p}) = V(\frac{Y}{n}) = \frac{1}{n^2}(np)(1-p) = p(1-p)/p$

이고 $n \to \infty$이면 $V(\hat{p}) \to 0$ 이므로 $\hat{p}$는 一致推定量이다. 만일 p의 推定量으로 다음과 같이 정의된 p^*를 취한다고 하자. 첫번째 택한 전구가 不良品이면 $P^* = 1$ 그렇지 않으면 $P^* = 0$

$E(P^*) = 1 \cdot P(X=1) + 0 \cdot P(X=0) = P$

$V(P^*) = P(1-P)$이므로 P^*는 不偏推定量이지만 一致推定量은 되지 못한다. 따라서 퍽 좋은 바람직한 推定量은 되지 못한다.

2.3 古典的 推定方法

2.3.1 点推定(point estimation)

点推定이 흔히 알고 있는 보통 推定法인데 이는 確率変数의 観察値에 의한 계산으로 부터 얻어진 하나의 数値가 되며 母数의 近似値로써 取扱하게 되는 것이다. 예컨데 어떤 기계에 의하여 생산되는 연속적인 50개의 部分品중 観察된 比率 P의 点推定이 되는 것이다. 点推定 方法에는 대체로 積率法(method

of moments), 最小自乘法 (least squares method), 最尤法 (maximum likelihood method) 그리고 其他方法 (other methods) 가 있다.

1) 積率法 (method of moments)

대부분의 문제에 있어서 1次의 積率과 2次의 積率에 대해서만 관심을 두는 것이 보통이다. 어떤 몇몇의 문제에서는 4次의 積率까지 使用되며 4次의 積率보다 高次의 積率을 사용하는 경우는 극히 드물다. 이에 대한 한가지 이유로서 抽出実驗을 反復하는 경우 高次의 積率은 아주 不安定하므로 高次積率에서 信賴할 만한 追加的인 情報를 얻을 수 없기 때문이다. 原点에 관한 1次의 積率 m'를 平均 (mean)이라 부르며 보통 $\bar{x}$로 표시한다.

즉 $$\bar{x} = \frac{1}{n}\sum_{i=1}^{h} x_i f_i \qquad (1)$$

단 x_i : i번째의 계급에 대한 계급값

f_i : 그 계급의 絶対度数

h : 계급의 数

n : 絶対度数의 総合

분류되지 않은 자료의 경우 $\bar{x}$는 한 모임의 数의 平均 (average) 으로서 잘 알려진 공식이 된다. 이 공식은 加重平均에 관한 공식이라고도 한다. 그런데 加重平均은 분류된 자료에 適応할 수 있는 공식의 한 종류에 지나지 않는다. x_i 와 f_i 가 그렇게 크지 않거나 계산기를 이용할 수 있으면 $\bar{x}$는 定義式으로 쉽게

계산할 수 있다. 그렇지 않은 경우에는 새로운 変数를 도입하여 간편셈을 함으로써 등간격의 계급으로 분류되어 있는 度数分布表에서는 많은 시간을 절약할 수 있다. 지금 整数値만을 취하는 새로운 変数 μ를 다음과 같이 정의한다.

$$xi = C\mu_i + x_0 \qquad (2)$$

여기서 C는 계급폭이며 x_0은 편하도록 택한 계급값이다. 보통 x_0은 分布의 中心부근에 있는 계급값으로 택하는 것이 계산상 편리하다. 이것을 (1)에 있는 xi에 대입하면

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{h}(C\mu_i + x_0)\cdot fi$$

$$= \frac{1}{n}\sum_{i=1}^{h}(C\mu_i f_i + x_0 f_i)$$

$$= \frac{1}{n}\sum_{i=1}^{h}(\mu_i f_i + \frac{1}{n}\sum_{i=1}^{h} x_0 f_i$$

x_0과 C는 常数이므로 総合記号앞으로 나오게 되어 다음과 같이 된다.

$$\bar{x} = C\cdot\frac{1}{n}\sum_{i=1}^{h}\mu_i\ fi + x_0\frac{1}{n}\sum_{i=1}^{h} fi$$

(1)로부터 C의 係数는 $\bar{\mu}$이며, μ의 정의로부터 x_0의 係数는 1이다. 따라서

$$\bar{x} = C\bar{\mu} + x_0 \qquad (3)$$

$\bar{\mu}$는 비교적 쉽게 계산할 수 있으므로 $\bar{x}$의 값은 계산기의 도움이 없어도 매우 간단하게 구할 수 있다.

變異라 하면 位置測度들레의 資料의 變異를 뜻하는 것이 보통이다. 여기서는 位置의 測度로서 平均을 사용하므로 積率에 의하여 變異의 測度를 얻자면 平均에 관한 積率을 도입할 必要가 있다.

즉 平均에 관한 K次 積率은

$$m_k = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^k \cdot f_i \qquad (4)$$

平均에 관한 2次의 積率 m_2가 變異의 測度로서 사용할 수 있음을 살펴보자. 變異의 測度는 측정의 단위로서 資料와 동일한 것이라야 편리하므로 보통 $\sqrt{m_2}$를 사용하게 된다. 이것을 標準偏差(standard deviation)라 부르며 S로서 표시한다. 즉

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{h}(x_i - \bar{x})^2 f_i} \qquad (5)$$

평균에 관한 2次의 積率 S^2은 경우에 따라서는 變異의 測度로서 標準偏差보다 편리하며 分散(variance)라 불리운다.

(5)식은 계산하는데 불편하므로 간단한 式으로 유도하면 $x_i - \bar{x} = C(\mu_i - \bar{\mu})$이므로 ((2)와 (3)으로부터)

$$\frac{1}{n}\sum_{i=1}^{h}(x_i - \bar{x})^2 f_i$$

$$= \frac{1}{n}\sum_{i=1}^{h} C^2(\mu_i - \bar{\mu})^2 f_i$$

$$= \frac{C^2}{n}\sum_{i=1}^{h}(\mu_i^2 - 2\mu_i\bar{\mu} + \bar{\mu}^2) f_i$$

$$= C^2\left(\frac{\sum_{i=1}^{h} \mu_i^2 f_i}{n} - 2\bar{\mu} \frac{\sum_{i=1}^{h} \mu_i f_i}{n} + \bar{\mu}^2 \frac{\sum_{i=1}^{h} f_i}{n} \right)$$

$$= C^2\left(\frac{\sum_{i=1}^{h} \mu_i^2 f_i}{n} - \bar{\mu}^2 \right)$$

따라서 $S = \sqrt{\frac{\sum_{i=1}^{h} \mu_i^2 f_i}{n} - \bar{\mu}^2}$

분류되지 않은 자료에서는 $S = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}$ 이 (5) 보다 훨씬 편리하다.

예를들어 보자.

아래표는 電話의 通話時間을 반올림하여 秒単位까지 측정한 것을 1,000회 기록한 것이다.

x	f	u	μf	$\mu^2 f$
49.5	6	- 4	- 24	96
149.5	28	- 3	- 84	252
249.5	88	- 2	-176	352
349.5	180	- 1	-180	180
449.5	247	0	0	0
549.5	260	1	260	260
649.5	133	2	266	532
749.5	42	3	126	378
849.5	11	4	44	176
949.5	5	5	25	125
合 計	1,000		257	2,351

平均 $\bar{x} = C\bar{\mu} + x_0$ 이므로

$$\bar{x} = 100 \cdot \left(\frac{257}{1006} \right) + 449.5 = 475.2$$

標準偏差는

$$S = C \cdot \sqrt{\frac{\sum_{i=1}^{m} \mu i^2 f_i}{n} - \bar{\mu}^2}$$ 이므로

$$S = 100 \sqrt{\frac{2351}{1000} - (257)^2} = 151$$

2) 最小自乘法 (least squares method)

주어진 점의 集合에 效果的으로 曲線을 맞추는 문제는 본질적으로는 曲線의 母數 (parameter) 를 效果的으로 推定하는 문제이다. 이러한 母數를 推定하는 方法은 여러가지가 있지만 가장 잘 알려진 것이 다음에 說明하고자 하는 最小自乘法 (least squares method) 이다.

원래 母數의 推定에는 당연히 偏差 (error) 을 수반한다. 그리하여 있을 수 있는 偏差를 되도록 적게하는, 그러한 성격을 밝히는 것이 요구되는 條件이다. 그 中 最小自乘法은 母數에 대한 偏差의 제곱을 最小로 하는 推定値를 얻는 方法이라 할 수 있다. 母數 그 自体는 一般的으로 알 수 없는 母値이지만 어쨌든 偏差의 存在는 豫想되고 그 方向은 正負 (플러스, 마이너스) 의 어느 쪽이든지 나타나기 때문에 理論上 그러한 推定方法이 요구되는 성질이다. 가령 어느 任意標本 $X_1, X_2 \cdots, X_n$이 있는데 각

X_i 는 同一한 平均 μ와 分散 σ^2를 갖는다고 하자. 이에 대한 各各의 実際観測値는 다음과 같은 관계로 표시될 수 있다.

$$X_i = \mu + \mu_i$$

여기서 $E(X_i) = \mu$라는 것은 다 아는 사실이다. 따라서 $E(\mu_i) = 0$은 당연하다. 우리는 μ를 推定하고자 하는데 最小自乘法에 의한 推定은 $\sum_{i=1}^{n} (X_i - \mu)^2$를 가장 작게 할 수 있는 μ의 推定量을 찾으려는 것이다. 즉 어떠한 μ의 값에 대해서도 다음과 같은 관계가 成立한다고 하면

$$\sum_{i=1}^{n} (X_i - \hat{\mu})^2 \leq \sum_{i=1}^{n} (X_i - \mu)^2$$

$\hat{\mu}$는 μ의 最小自乘推定量(least squares estimator)가 된다. 여기서

$$X_i = \hat{\mu} + e_i$$

라고 놓으면 $\hat{\mu}$는 다음과 같이 偏差를 제곱하여 合한 것을 最小化시킴으로써 구해질 수 있다.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (X_i - \hat{\mu})^2$$

$\hat{\mu}$에 대한 식의 一次微分을 0과 같게 놓으면 바로 구해진다.

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\mu}} = -2 \sum_{i=1}^{n} (X_i - \hat{\mu}) = -2 \left[\sum_{i=1}^{n} X_i - n\hat{\mu} \right] = 0$$

$$\sum_{i=1}^{n} X_i = n\hat{u}$$

$$\hat{u} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

이와같이 最小自乘法에 의하여 구한 推定量 $\hat{u}$는 標本平均 $\bar{X}$와 같다.

지금 一例로써 어떠한 두 統計系列의 관계 즉 統計値 X에 대한 Y의 分布를 본다 하고 이를 규정하는 진정한 趨勢直線이 있어서 그것이

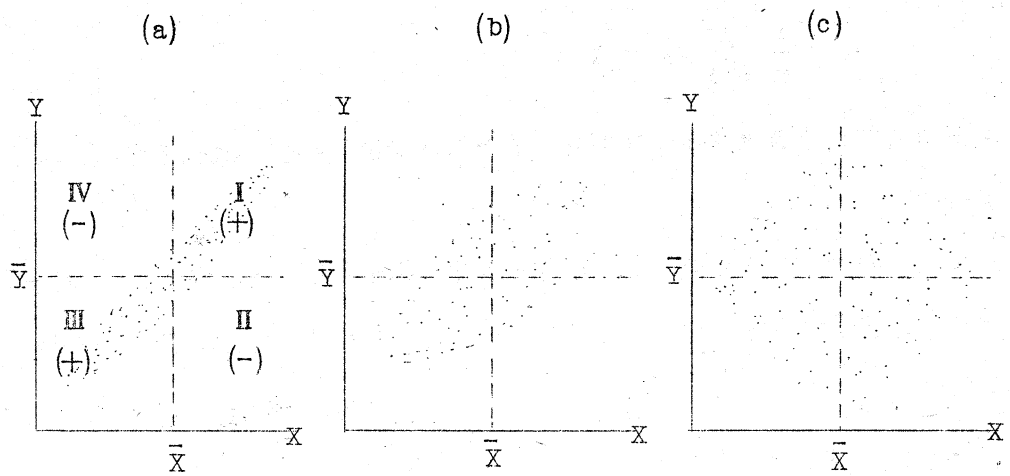
$$Y_i = a + b\ X_i$$

라 하자. 이 식을 真(母) 回帰線(true or population regression line)이다. 여기서 a, b를 구한다는 것은 사실상 불가능하고 X, Y의 標本觀測値(sample observation)를 가지고 a, b를 推定해야 한다. $\hat{a}$를 a의 推定量 $\hat{b}$를 b의 推定量이라 하면 X, Y의 推定된 관계는 $Y_i = \hat{a} + \hat{b}\ X_i + e_i$ (i = 1, 2, 3 … n)가 되고 X, Y의 推定된 回帰線(estimated regression line) 혹은 標本回帰線(sample regression line)을 나타내는 식은 다음과 같다.

$$\hat{Y}_i = \hat{a} + \hat{b}\ X_i \quad (i = 1, 2, 3, \cdots n)$$

標本觀測을 二次元平面위에 표시해 보면 그림에서와 같이 흩어진 点들이 될 것이다.

標本回帰線은 이 점들을 가장 잘 대표하는 線으로 그어져야 한다. 이렇게 그어진 線을 흔히 適合線(fitted line)이라 한다. 이 適合線이 얼마나 잘 그어졌느냐 하는 것은 標本回帰線 즉 平均



<그림> 두 変数간의 관계

에서 点들이 양쪽으로 떨어진 거리 즉 偏差를 最小로 하느냐 하는것이다. 偏差가 両便으로 散在하기 때문에 前提條件에 의하여 偏差의 合은 0이 된다. 그래서 各偏差를 제곱하여 合한 것(自乘合)이 最小가 되면 適合線은 좋게 잘 그어진 것이 된다. 즉 그 偏差의 自乘合이 最小가 되게 하는 것이다. 母数 a, b는 真回帰線의 位置와 기울기이므로 이 自乘合이 最小가 되게 하여 구한 a, b는 가장 좋은 適合線을 구하였다는 것이 되므로 여기서 구한 $\hat{a}$, $\hat{b}$는 真回帰線母数 a, b의 推定量이 된다. 実際標本観測値 즉 흩어진 点들을 나타내는 관계가 위에서 밝힌 바 있는 $Y_i = \hat{a} + \hat{b}\, X_i + e_i$ (1)

이고 그 平均인 標本回帰線을 나타내는 것이 마찬가지로 위에서

말한바와 같이

$$\hat{Y}i = \hat{a} + \hat{b}\,Xi \quad (i = 1, 2, 3 \cdots n) \tag{2}$$

이다. 平均과 各点과의 偏差인 残差(residual)는 (2)를 (1)에서 빼줌으로써 구할 수 있다.

$$ei = Yi - \hat{Y}i = Yi - \hat{a} - \hat{b}\,Xi \tag{3}$$

($\hat{Y}i = \hat{a} + \hat{b}\,Xi$ 를 代入)

이 残差를 제곱하여 合하면 다음과 같이 된다.

$$\sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n} (Yi - \hat{a} - \hat{b}\,Xi)^2$$

우리는 이것을 最小로 해야 한다. 이렇게 하기 위하여 즉 必要條件으로 이 式을 偏微分하여 0과 같게 놓아야 한다. 그런데 우리가 구하고자 하는 것은 $\hat{a}$, $\hat{b}$이므로 $\hat{a}$, $\hat{b}$에 대하여 偏微分을 해야 한다.

$$\frac{\partial \Sigma ei^2}{\partial \hat{a}} = \frac{\partial \Sigma (Yi - \hat{a} - \hat{b}\,Xi)^2}{\partial \hat{a}} = \Sigma 2 (Yi - \hat{a} - \hat{b}\,Xi) \tag{-1}$$

$$= -2\Sigma (Yi - \hat{a} - \hat{b}\,Xi) = 0$$

$$\frac{\partial \Sigma ei^2}{\partial \hat{b}} = \frac{\partial \Sigma (Yi - \hat{a} - \hat{b}\,Xi)^2}{\partial \hat{b}} = \Sigma 2 (Yi - \hat{a} - \hat{b}\,Xi)(-Xi)$$

$$= -2\Sigma Xi (Yi - \hat{a} - \hat{b}\,Xi) = 0$$

이것을 整理하여 놓으면

$$\Sigma Yi = n\hat{a} + \hat{b}\Sigma Xi \tag{4}$$

$$\Sigma Xi\,Yi = \hat{a}\Sigma Xi + \hat{b}\Sigma Xi^2 \tag{5}$$

와 같이 된다. 이 두 式을 標本回帰線의 正規方程式(normal equations)이라 한다. 이 正規方程式을 $\hat{a}$와 $\hat{b}$에 대해서 풀면 다음과 같이 된다.

$$\hat{a} = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n \Sigma X_i^2 - (\Sigma X_i)^2} \tag{6}$$

$$\hat{b} = \frac{n \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n \Sigma X_i^2 - (\Sigma X_i)^2} \tag{7}$$

標本觀測値 그대로를 計算하여 代入하면 위와같은 最小自乘法에 의하여 推定한 $\hat{a}$와 $\hat{b}$의 값이 주어진다. 이것을 偏差 즉 $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$로 보기 위하여 (4)의 各項을 n으로 나누면

$$\bar{Y} = \hat{a} + \hat{b}\bar{X} \tag{8}$$

$$\left(\bar{Y} = \frac{1}{n}\Sigma Y_i \qquad \bar{X} = \frac{1}{n}\Sigma X_i\right)$$

가 된다. 여기에서 最小自乘法에 의하여 推定된 回帰線은 平均点 $(\bar{X}, \bar{Y})$를 지나게 됨을 볼 수 있다.

式 (8)을 式 (2)에서 빼면

$$\hat{Y}_i - \bar{Y} = \hat{a} + \hat{b} X_i - (\hat{a} + \hat{b}\bar{X})$$

$$= \hat{b}(X_i - \bar{X})$$

가 된다. 이것을 다시 쓰면 다음과 같은 回帰線을 나타내는 式이 된다.

$$\hat{y}_i = \hat{b} x_i \tag{9}$$

또한 式 (1)에서 式 (8)을 빼주면 다음과 같다.

$$y_i = \hat{b} x_i + e_i$$

式 (3)에서와 마찬가지로 残差와 그 自乘合은 다음과 같이 구해진다.

$$\epsilon i = yi - \hat{y}i = yi - \hat{b}\, xi$$
$$\Sigma\, \epsilon^2 = \Sigma\, (yi - \hat{b}\, xi)^2 \qquad (10)$$

이것을 微分하여 0으로 놓고 $\hat{b}$에 대하여 풀면 다음과 같이 된다.

$$\frac{\partial \Sigma\, \epsilon i^2}{\partial \hat{b}} = \frac{\partial \Sigma\, (yi - \hat{b}\, xi)^2}{\partial \hat{b}}$$
$$= -2\Sigma\, xi\,(yi - \hat{b}\, xi) = 0$$
$$\hat{b} = \frac{\Sigma\, xi\ yi}{\Sigma\, x_i^2} \qquad (11)$$

式 (8)을 $\hat{a}$에 대하여 풀면

$$\hat{a} = \bar{Y} - \hat{b}\, \bar{x} \qquad (12)$$

가 된다. 式(12)은 標本単純回帰線의 위치를 그리고 式(11)은 標本単純回帰線의 기울기를 測定하는 것이므로 기억해야 할 重要한 公式이다.

예를들면 X, Y에 대한 資料는 1965 ~ 1974年間의 우리나라 1970年 不変価格国民総生産(X)과 民間消費支出(Y)이다.

表를 작성하여 계산된 数字들을 위 公式에 代入해 넣으면 다음과 같은 $\hat{a}$, $\hat{b}$의 数値를 얻게 된다.

$$\hat{a} = \bar{Y} - \hat{b}\, \bar{X} = 183.0 - (0.62)253.8 = 25.64$$

$$\hat{b} = \frac{\Sigma\, xt\ yt}{\Sigma\, xt^2} = \frac{33008.0}{53321.6} = 0.62$$

이것을 흔히 Y를 X에 回帰시켜 $\hat{a}$와 $\hat{b}$를 推定한다고도 한다. 위의 数値를 式 (2)에 넣으면 다음과 같은 回帰線을 나타내는 式이 된다.

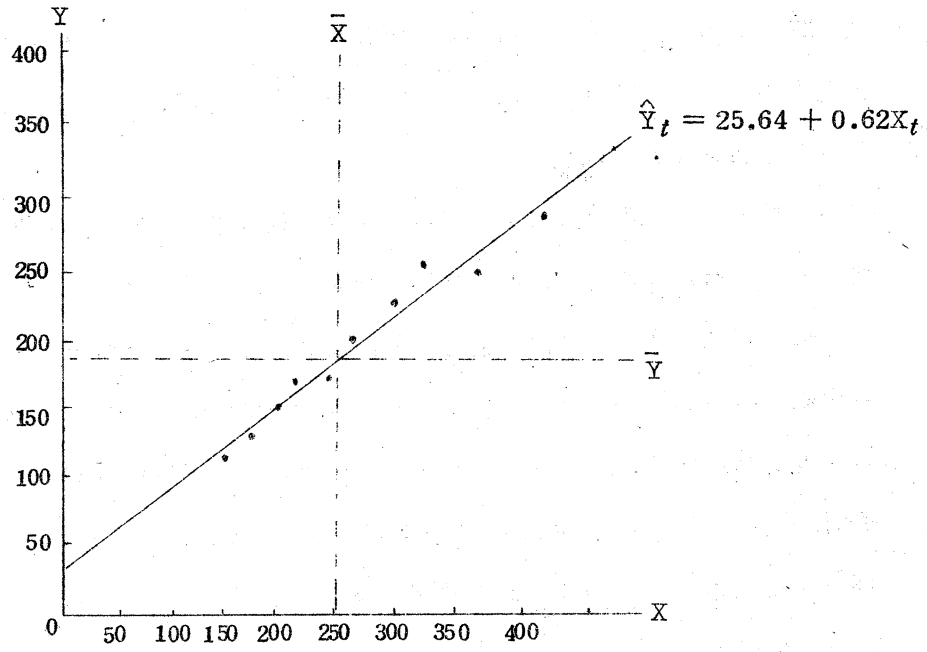
$$\hat{Y}_t = 25.64 + 0.62\ X_t$$

X의 標本観測値를 式에 넣으면 $\hat{Y}$이 計算된다. 이것을 推定 또는 計算된 Y라고도 한다. 本例題의 結果를 나타낸 그림이 아래와 같게 된다.

1965 ~ 1974년간 우리나라 国民総生産(X)과 国民消費支出(Y)
(1970년 不変価格)

(単位 : 100億원)

t	X_t	Y_t	X_t^2	x_t $(X_t-\bar{X})$	y_t $(Y_t-\bar{Y})$	$x_t \cdot y_t$	x_t^2	y_t^2	$\hat{Y}_t$ $(\hat{a}+\hat{b}X_t)$	e_t $(Y_t-\hat{Y}_t)$	e_t^2
1	153	120	23409	-100.8	-63	6350.4	10160.64	3969	120.50	-0.50	0.25
2	172	128	29584	-81.8	-55	4499.0	6691.24	3025	132.38	-4.38	19.18
3	185	140	34225	-68.8	-43	2958.4	4733.44	1849	140.34	-0.34	0.12
4	209	155	43681	-44.8	-28	1254.4	2007.04	784	155.22	-0.22	0.05
5	240	171	57600	-13.8	-12	165.6	190.44	144	174.44	-3.44	11.83
6	259	188	67081	5.2	5	26.0	27.04	25	186.22	1.78	3.17
7	283	208	80089	29.2	25	730.0	852.64	625	201.10	6.90	47.61
8	302	223	91204	48.2	40	1928.0	2323.24	1600	212.88	10.12	102.41
9	352	242	123904	98.2	59	5793.8	9643.24	3481	243.88	-1.88	3.53
10	383	255	146689	129.2	72	9302.4	16692.64	5184	263.10	-8.10	65.61
合計	2538	1830	697466	0	0	33008.0	53321.60	20686		0.06	253.76



<그림> 回帰線実例

3) 最尤法 (maximum likelihood method)

最尤法은 推定量을 구하는데 있어서 보다 일반적인 方法으로 認定받고 있으나 이 方法에 의하여 구해진 推定量은 大標本에서 좋은 性質을 가질 수 있는데 小標本에서는 必然的으로 不偏이 되는 것은 아니다. 最小自乘法은 母数에 대한 統計偏差의 제곱을 최소로 만드는 그러한 推定値를 구하는 方法이었지만 널리 쓰이는 또 하나의 推定法이 바로 最尤法이다. 이 方法은 지목되는 標本이 나타나는 確率을 最大化하는 그러한 母数의 決定法이다. 그것은

말하자면 나타난 現象(統計結果)에 의하여 일정한 分布條件 아래서 그와같은 現象이 일어나는 確率을 最大로 하는 母數의 論理的 推定方法이다. 이는 R.A. Fisher의 공헌이다. 推定方法을 지금부터 說明하기로 한다.

各 相異한 諸母集団에서 역시 각각 相異한 標本이 抽出될 것인데 어느 特定標本은 다른 어느 母集団으로부터 보다 더 어느 일정한 母集団으로부터 抽出되었을 可能性이 더 클 것이다. 그래서 이 特定한 標本은 어느 母集団에 속할 可能性이 제일 크냐 하는 문제가 제기된다. 이러한 着想에 基準을 둔 推定方法이 곧 最尤法이다. 가령 $\theta_1, \theta_2, \theta_3 \cdots \theta_K$라는 母數에 의하여 特徵지어지는 確率分布 $f(x)$를 가진 確率変数 X가 있고 $x_1, x_2, x_3 \cdots xn$등으로 標本을 観測한다고 하자. 그러면 이 観測된 標本을 가장 높은 頻度로 抽出했을 母數의 값을 $\theta_1, \theta_2, \theta_3 \cdots \theta_k$의 最尤推定量(maximum likelihood estimators)이라 한다. 最尤推定量은 그 特定한 標本値의 確率(혹은 確率密度)이 最大가 되게 하는 母數의 값을 말한다. 그래서 最尤推定量을 구하는 것은 곧 $f(x_1, x_2, x_3 \cdots\cdots x_n)$을 最大로 하는 값을 찾는 것이라 할 수 있다. 이상은 母數가 K개 있는 경우였는데 問題의 간략성을 위해 어느 単一母数 θ 하나만을 놓고 보기로 한다. 위의 예에서 그 確率分布 $f(x)$는 未知의 母數 θ에 의하여 特徵지어진다고 하면 Xi는 相互独立이기 때문에 標本의 結合確率分布가 다음과 같이 표시될 수 있다.

$$f(X_1 = x_1,\ X_2 = x_2,\ X_3 = x_3 \cdots\cdots\ X_3 = x_3)$$
$$= f(x_1, x_2, x_3 \cdots\cdots x_3)$$
$$= f(x_1)\ f(x_2)\ f(x_3) \cdots\cdots f(x_n)$$

이들 各個의 限界確率分布는 母数 θ에 依存하고 있으므로 이 母数 θ를 (1)에 포함시키면 다음과 같이 쓸 수 있다.

$$L(x_1, x_2, x_3, \cdots\cdots x_n : \theta)$$
$$= f(x_1 : \theta)\ f(x_2 : \theta)\ f(x_3 : \theta) \cdots\cdots f(x_n : \theta)$$

이와같은 標本의 結合確率分布를 θ의 函数로 看做한 나머지 그 標本에 대한 尤度函数(likelihood function)라 한다.

만약 θ의 모든 값에 대하여 다음과 같은 관계가 성립되면 $\hat{\theta}$는 θ의 最尤推定量이 된다.

$$L(x_1, x_2, x_3 \cdots x_n : \hat{\theta}) \geq L(x_1, x_2, x_3 \cdots x_n : \hat{\hat{\theta}})$$

예를들어 說明해 보자.

平均 μ와 分散 σ^2를 갖고 正規分布를 하고 있는 確率変数 X가 있다고 하자. 任意標本으로 $\{x_1, x_2, x_3 \cdots\cdots x_n\}$가 観測되었다고 할 때 μ와 σ^2의 最尤推定値를 구해보면 여기에서 正規確率函数는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{xi-\mu}{\sigma}\right)^2}$$

또한 尤度函数는

$$L(x_1, x_2 \cdots x_n : \mu, \sigma^2) = \prod_{i=1}^{n} \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \ominus \frac{1}{2\sigma^2} (xi-\mu)^2 \right.$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \ominus -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (xi-\mu)^2$$

과 같이 된다. 이것을 代數式으로 変換시키기 위하여 이것에 log를 취하면 다음과 같이 된다.

$$\log L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (xi-\mu)^2$$

이 式에서 μ와 σ^2가 未知数이므로 이것들에 대한 一次微分을 求하면 다음과 같이 된다.

$$\frac{2(\log^L)}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(xi-\mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (xi-\mu)$$

$$\frac{\partial(\log^L)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (xi-\mu)^2$$

위의 두 式을 같게 놓아 μ의 推定量 $\hat{\mu}$와 σ^2의 推定量 $\hat{\sigma}^2$를 다음과 같이 誘導할 수 있다.

$$\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (xi - \hat{\mu}) = 0$$

$$\sum_{i=1}^{n} (xi - \hat{\mu}) = 0$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} xi = \bar{x}$$

$$-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(xi-\hat{\mu})^2=0$$

$$-n\hat{\sigma}^2+\sum_{i=1}^{n}(xi-\hat{\mu})^2=0 \quad (\text{兩項에 } 2\hat{\sigma}^4\text{을 곱하여})$$

$$-n\hat{\sigma}^2+\sum_{i=1}^{n}(xi-\bar{x})^2=0 \quad (\bar{x}\text{를 } \hat{\mu}\text{에 代入})$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(xi-\bar{x})^2$$

$$\mathrm{E}(\hat{\sigma}^2)=\frac{n-1}{n}\sigma^2$$

위에서 母平均의 推定量은 標本平均과 같으나 分散의 推定量은 標本分散과 같지 않아 不偏이 되지 못한다. 大標本의 경우에는 이 偏倚가 無視될 수 있다. 그래서 小標本인 경우는 수정이 加해져야 한다.

單一方程式(single equation) 推定에서는 일반적인 方法이 못되고 대부분 最小自乘法이 쓰이므로 간단하게 설명하며는 最尤法은 頻度가 가장 높은 標本觀測을 수반하는 母數를 찾는것이다. 그러므로 標本觀測에 대한 尤度函數(likelihood function)를 구하여 이것을 母數에 대하여 極大化해야 한다. Y_i의 分布를 $Yi \sim N(\mathrm{E}(Yi),\sigma_y^2)$이라 仮定한다면 Yi는 이러한 平均과 分散을 가지고 각각 独尤이라 하였다. 그래서 Yi 標本觀測에 대하여 다음과 같은 尤度函數를 놓을 수 있다.

$$\log \mathrm{L}=-\frac{1}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathrm{Y}i-\hat{\mathrm{a}}-\hat{\mathrm{b}}\,\mathrm{X}i)^2$$

단, 平均 $E(Y_i)$는 $\hat{a}+\hat{b}X_i$로 代入

極大化의 一次條件으로 윗 式을 $\hat{a}$, $\hat{b}$에 대하여 偏微分하여 0과 같게 놓으면 다음과 같게 된다.

$$\frac{\partial \log L}{\partial \hat{a}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i-\hat{a}-\hat{b}X_i)(-1)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\hat{a}-\hat{b}X_i) = 0$$

$$\frac{\partial \log L}{\partial \hat{b}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i-\hat{a}-\hat{b}X_i)(-X_i)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i-\hat{a}-\hat{b}X_i) = 0$$

이 두 式을 整理하여 다시 쓰면

$$\sum_{i=1}^{n} Y_i = n\hat{a}+\hat{b}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{a}\sum_{i=1}^{n} X_i + \hat{b}\sum_{i=1}^{n} X_i^2$$

가 된다. 이 두 式은 最小自乘法으로 구한 式과 똑같다. 즉 같은 標本回帰線의 正規方程式이 된다. 이로 미루어 보아 最小自乘法과 最尤法과는 같은 것이라 할 수 있다. 다만 앞에서 보았지만 最尤法에서는 標本의 크기가 작으면 分散의 偏倚性이 있는 점과 分散을 推定하기가 더 힘들다는 것이다.

2.3.2 区間推定 (interval estimation)

앞에서 区間推定 (interval estimation) 에 대하여 간단히 설명하였지만 여기서 자세하게 說明하고자 한다. 어느 주어진 確率로 真正한 母數를 包含하고 있을 点推定에 대한 区間을 信賴区間이라 하고 이 信賴区間의 推定을 区間推定이라 한다. 母集団의 平均과 分散에 대한 信賴区間의 推定을 나누어 설명하고자 한다.

(1) μ의 信賴区間

1) 標準偏差 (σ) 를 알고 있을때의 μ의 信賴区間

한 母集団에서 크기 n의 任意標本을 抽出하였을 때 n이 충분히 크면 標本平均 $\bar{X}$는 正規分布를 한다고 생각할 수 있으므로 이와같은 사실을 利用하여 母平均 μ의 区間推定을 하여보자. 그런데 正規分布를 規定하는 母數는 平均 μ와 標準偏差 σ이다. 그러므로 標本平均 $\bar{X}$를 利用하여 μ에 대한 確率的인 区間推定을 하려면 母標準偏差 σ를 알고 있지 않으면 안된다. 그런데 母標準偏差 σ를 과거의 経験에서 비교적 正確하게 알고 있다고 믿을 수 있는 경우가 없지는 않지만 대부분 σ를 알지 못한다. 그러나 여기서는 문제를 간단히 해결하기 위하여 母集団의 σ를 알고 있다고 생각하기로 한다.

母集団 平均이 μ, 標準偏差가 σ라고 하면 任意標의 크기 n이 충분히 클때는 標本平均 $\bar{X}$는 平均이 μ 標準偏差가 $\frac{\sigma}{\sqrt{n}}$인 正規

分布를 한다. 이것을 標準化 形態로는 다음과 같이 표시한다.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

그러면 $P(Z \leq Z_0) = 0.975$

와 같이 되는 Z_0의 값을 標準正規分布表에서 구하면 $Z_0 = 1.96$

을 얻는다. 즉

$$P\left(\frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \leq 1.96\right) = 0.975$$

여기서 괄호안의 부등호를 고쳐 쓰면 $\bar{X} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$와 같이 되어

$$P\left(\bar{X} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.975$$

와 같이 된다. 마찬가지로

$$P\left(-1.96 \leq \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}\right) = 0.025$$

와 같이 되고, 이때도 괄호안을 고쳐쓰면

$$P\left(\bar{X} \geq \mu - 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.025$$

와 같이 된다. 결국 위의 두가지 사실을 종합해 보면 $\bar{X}$가 $\mu - 1.96 \frac{\sigma}{\sqrt{n}}$와 $\mu + 1.96 \frac{\sigma}{\sqrt{n}}$ 사이에 있는 確率, 즉

$$\mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$$

와 같이 되는 確率이 95 %로 된다.

이제 위의 左側不等式 $\mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X}$ 에서

$$\mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

를 얻고 右側不等式 $X \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}$ 에서

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu$$

를 얻는다. 이와같이 얻은 두개의 不等式을 한데 모아

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

와 같이 쓰고 이것을 母平均 μ에 대한 95% 信頼区間이라 한다. 마찬가지로 하여 99% 信頼区間을 구하면 다음과 같이 된다.

$$\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}$$

위와같은 信頼区間을 자세히 표현하면 両側 信頼限界 (two-sided confidence limits) 라고 한다. 이와같이 말하는 理由는 単側信頼限界 (one-sided confidence limits) 를 필요할 때가 있기 때문이다.

標本의 크기가 클때에는 二項母集頭에서 比率에 대한 区間推定을 하는데 그대로 応用될 수가 있다. 成功의 母比率 P를 가지고 있는 二項母集団에서 크기 n의 標本을 추출하여 成功의 数 r를 얻었다면 標本比率 $\frac{r}{n}$는 P의 点推定値로 되고, 二項分布 理論에 따라 r의 期待値는 np, r의 分散은 npq이므로 $\frac{r}{n}$의 期待値는 $\frac{np}{n} = P$, $\frac{r}{n}$의 分散은 $\sigma_r^2 = \frac{npq}{n}$ 로 된다. 따라서 n이 충분히

를 때는 標本比率 $\frac{r}{n}$는 平均이 P, 標準偏差가 $\sqrt{\frac{pq}{n}}$인 正規分布를 한다. 그러므로 n이 충분히 클 때는 다음과 같은 確率公式을 얻게 된다.

$$P=\left(P-1.96\sqrt{\frac{pq}{n}} \leq \frac{r}{n} \leq P+1.96\sqrt{\frac{pq}{n}} \right) = 0.95$$

2) σ를 모를때의 μ의 信賴区間

위에서는 母平均 μ를 区間推定하는데 母標準偏差 σ를 알고 있다고 하였는데 대부분의 경우 σ를 알지 못한다. 이와같이 σ를 알지 못하는 경우는 標本에서의 觀測値에서 標本標準偏差 S를 계산하여 이것에서 母標準偏差 σ의 推定値로 삼고 σ 대신에 S를 사용하여 区間推定을 한다. 標本의 크기가 충분히 클때 사용하는 것이 좋다. 그래서 標本標準偏差 S를 標準化된 形態인

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

에 σ 대신 代入하여 t-分布를 쓰면

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

단 t_{n-1}=自由度($n-1$)을 가진 t分布

여기서 다음과 같은 確率을 얘기할 수 있다.

$$P\left(-t_{n-1}:\alpha/2 \leq \frac{\overline{X}-\mu}{S/\sqrt{n}} \leq t_{n-1}:\alpha/2 \right) = 1-\alpha$$

($1-\alpha$)는 信賴水準을 가리킨다. 이것을 信賴係数(confidence

coefficient)의 確率이라고 하는 이도 있다. 윗式을 95 % 信頼水準으로 놓고 보면 $\alpha = 0.05$가 되고 윗式은 다음과 같이 쓰여진다.

$$P\left(-t_{n-1\ :0.025} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1\ :0.025} \right) = 0.95$$

이것을 다른 形態로 바꾸어 쓰면

$$\bar{X} - t_{n-1:0.025}\ \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\,n-1:1.025}\ \frac{S}{\sqrt{n}}$$

와 같이 된다. t-分布에서 $\alpha = 0.5$에 따른 自由度 ($n-1$)의 값을 가지고 t값을 찾아 윗式에 代入하면 바로 母平均 μ에 대한 95 % 信頼区間을 얻게 된다. 99 % 信頼区間도 마찬가지로 얻게 된다.

(2) σ의 区間推定

이제까지 母集団의 標本에서 母平均 μ를 区間推定을 하는 것을 説明하였다. 이제 母分散 σ^2를 区間推定하는 問題를 취급하여 보기로 한다. 만약 母集団이 正規分布에 따른다면 x^2-分布表를 이용하여 標本分散 S^2를 기초로 하여 母分散 σ^2을 区間推定을 할 수가 있다. σ^2의 95 % 信頼限界値를 구하기 위하여 $x_{\,0.975^2}$와 $x_{\,0.025^2}$의 값을 x^2-分布表에서 구한다.

$x^2 = f\ S^2/\sigma^2$ 이므로

$$P\left(x_{\,0.975^2} \leq f\ S^2/\sigma^2 \leq x_{\,0.025^2} \right) = 0.95$$

와 같은 관계가 成立하고, 이 관계식에서 다음과 같이 σ^2의 95 % 信頼区間이 얻어진다.

$$\frac{f\,s^2}{\chi_{0.025^2}} \leq \sigma^2 \leq \frac{f\,s^2}{\chi_{0.075^2}}$$

이것이 95 % 信頼区間을 구하기 위한 일반적인 公式이다. 크기 n의 任意標本에서 S^2이 계산된 것이라고 하면 $\chi = X - \bar{X}$로 하였을때 $f\,s^2 = (n-1)S^2 = \Sigma\,\chi^2$와 같이되고 윗式은 다음과 같이 쓸 수가 있다.

$$\frac{\Sigma\,x^2}{\chi_{0.025^2}} \leq \sigma^2 \leq \frac{\Sigma\,x^2}{\chi_{0.975^2}}$$

σ를 좀더 正確히 推定하자면 標本의 크기가 커야된다는 것을 χ^2-分布表를 살펴 봄으로써 알 수 있다.

2.4 Bayesian推定方法

2.4.1 Bayes確率定理

우선 예를 들어 생각해 보자. 흰구슬(W)이 4개, 빨간 구슬(R)이 6개 들어 있는 상자 A가 3개 있고 흰구슬(W)이 2개, 빨간 구슬(R)이 8개 들어 있는 상자 B가 7개 있을때, 任意로 한 상자를 택하여 구슬 하나를 抽出하였을 경우 그것이 상자 A에서 흰구슬이 抽出될 確率이 얼마인가를 생각해보자. 이번에는 딴 사람에 의하여 위와 같은 抽出作業이

실시되고 최종적으로 抽出된 것이 흰구슬(W)이었다는 것만을 알리고, 이것이 A상자에서 抽出되었을 確率이 얼마인가를 물어 봤을 때 이에 대한 대답은 어떻게 될 것인가? 이때에는

$$P(A \mid W)$$

를 구하는 問題로 된다.

$$P(A \mid W) = \frac{P(A, W)}{P(W)}$$

이고, $P(W) = P(A, W) + P(B, W)$

$$P(A, W) = P(A)\ P(W \mid A),$$

$$P(B, W) = P(B)\ P(W \mid B)$$

이므로 윗式은 다음과 같이 된다.

$$P(A \mid W) = \frac{P(A)P(W \mid A)}{P(A)P(W \mid A) + P(B)P(W \mid B)}$$

위의 例의 경우는 $P(W \mid A) = 0.4$, $P(A) = 0.3$, $P(W \mid B) = 0.2$, $P(B) = 0.7$ 이므로

$$P(A \mid W) = \frac{(0.3)\ (0.4)}{(0.3)(0.4) + (0.7)(0.2)} = \frac{12}{36}$$

와 같이 質問에 대한 답을 얻는다.

일반적으로 서로 排反的이고 集合的으로 망라된 事件 $A_1, A_2 \cdots A_n$이 있고 이와같은 事件의 하나가 일어나는 것이 다른 한 事件 B가 일어나는데 필요한 前提條件으로 될때 確率 $P(Ai \mid B)$는 다음과 같게 된다.

$$P(A_i|B) = \frac{P(A_i)\,P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \cdots P(A_n)P(B|A_n)}$$

이것이 Bayes 確率定理라고 한다. 그리고 이것은 英国의 Thomas Bayes의 이름을 따서 命名된 定理이다. 이와 같이 계산된 確率을 事後確率 (posterior probability) 이라고 하며, 근래 소위 不確定的 状況下에서의 意思決定問題에 많이 이용되고 있다. 여기에서 Bayes 定理의 직접적인 応用例를 들어보자.

어느 地方의 한 群内 全体住民의 1 %가 어느 특정한 病에 걸려 있다. 그런데 이 病에 대한 감염여부를 알아보는 새로운 진단방법에 의하면 이 病에 감염되어 있는 사람들의 97 %가 (+) 反応을 나머지 3 %가 (-) 反応을 나타낸다. 한편 감염되어 있지 않은 사람들도 그들중 5 %가 (+) 反応을 나타내고 95 %가 (-) 反応을 나타낸다고 한다. 지금 住民중에서 한 사람을 任意로 抽出하여 이 病에 대한 검진을 하여 본 결과 (+) 反応을 나타냈다고 한다. 그러면 이 사람이 이 病에 감염되어 있을 確率은 얼마인가?

病에 감염되어 있다는 事件을 D, 감염되어 있지 않다는 事件을 N로 표시한다면 위에서 설명한 바는

$P(D) = 0.01$, $P(N) = 0.99$이고

$P(+|D) = 0.97$, $P(-|N) = 0.95$

따라서 $P(-|D) = 0.03$, $P(+|D) = 0.05$

와 같게 된다. 따라서 구하는 답은 Bayes 定理에 의하여

$$P(D \mid +) = \frac{P(D)\,P(+ \mid D)}{P(D)P(+ \mid D) + P(N)P(+ \mid N)}$$

$$= \frac{(0.01)\,(0.97)}{(0.01)\,(0.97) + (0.99)\,(0.05)}$$

$$= 0.16$$

과 같게 된다. 그러므로 $P(N \mid +) = 0.84$ 이것은 病에 걸려 있는 사람의 97 %가 검진결과 (+) 反応을 나타내는 데도 불구하고 검진결과 (+)反応을 나타낸 사람 중 16 %만이 정말로 病에 걸려 있을 것이라는 데 놀라지 않을 수 없다. 그런데 만일 任意로 抽出된 사람이 검진결과 (-) 反応을 나타냈다고 하면. 어떻게 될 것인가? 이 경우는

$$P(D \mid -) = \frac{P(D)\,P(- \mid D)}{P(D)\,P(- \mid D) + P(N)\,P(- \mid N)}$$

$$= \frac{(0.01)\,(0.03)}{(0.01)\,(0.03) + (0.99)\,(0.95)}$$

$$= 0.00032$$

$$P(N \mid -1) = 1 - 0.00032 = 0.99968$$

과 같이 되어 (-)反応을 보인 사람은 거의 틀림없이 건강하다는 것을 알 수 있다. 다시 말하면 (-)反応을 보일 때는 상당히 正確하게 판단을 내릴 수 있다. 그런데 만약 위와같은 검진방법에 의하면 (+)反応을 보이면 항상 病에 감염되어 있는 것으로 판단을 내리고 (-)反応을 보일 때는 언제나 病에 감염되어 있지 않다는 것으로 판정을 내리기로 한다면 이와같은 判定決定方法이

적중될 確率은 얼마나 될까 생각해 보면 옳은 判定을 내릴 確率

$$= P(D, +) + P(N, -)$$

$$= P(D)P(+|D) + P(N)P(-|N)$$

$$= (0.01)(0.97) + (0.99)(0.95)$$

$$= 0.0097 + 0.9405 = 0.9502$$

와 같이 되어 判定이 적중될 確率이 95 %나 된다는 결론을 얻는다. 이와 같이 確率값이 큰 것은 (-) 反応을 보이고 病에 감염되어 있지 않은 사람의 確率이 94.05 %인 것에 기인한다. 실제로 病에 감염되어 있고 (+)反応을 보이고 住民은 全体住民의 겨우 0.97 %이고, 病에 감염되어 있지 않고 (+) 反応을 보이는 住民은 全体住民의 $P(N, +) \times 100\ \% = 4.95\ \%$로서, 人口比例로 보아 後者가 前者의 5倍가 된다. 그러므로 任意로 한 사람을 抽出하였을 때 病에 감염되고 (+)反応을 보이는 사람보다 病에 감염되어 있지 않고 (+) 反応을 보이는 사람이 5倍나 더 잘 抽出될 可能性이 있다는 事実에서 $P(N|+) = 0.84$과 같은 計算이 나오지만, 이와 같이 오진의 原因으로 되는 사람을 검진하게 될 기회가 드물다는 것을 이해하면 옳게 판단을 내리게 되는 確率이 95 %나 된다는 것이 이상하지 않을 것이다.

2.4.2 事前分布와 事後分布

(1) 事前分析 (事前分布)

実際的인 意思決定問題에 있어서는 意思決定者가 여러가지

可能한 自然의 狀態(事件)에 대하여 그 各 狀態가 실현될 수 있는 可能性의 정도에 대하여 어떤 豫測이 있을 것이고 이것이 意思決定에 도움을 줄 것이라고 생각하는 것은 合理的인 것이라고 할 수 있다. 예를 들어 說明하면 P氏는 어떤 것을 발명하여 特許를 얻었다. 銀行은 이것을 企業化하는데 資金을 제공할 것을 이미 약속했다. 한편 조사연구결과 앞으로 5년간은 이 발명으로 利得을 알아 보는데 적당한 期間이라는 것을 알았다. 그의 調査結果에 의하면 만약 製品이 아주 잘 팔리면 앞으로 5년간 ₩800,000,000의 利得이 있고 보통으로 팔리면 ₩200,000,000의 利得이 있으며 보통이하로 팔리면 ₩50,000,000의 損失이 있게 된다. 한편 어느 大企業体에서 이 特許權을 사들일 것을 제안하여 왔다. 이 제안은 잘 팔릴 경우 ₩400,000,000의 配当을 받고 보통으로 팔릴때는 ₩70,000,000을 보통 이하로 팔릴 때는 ₩10,000,000를 배당받게 된다.

이것을 表로 작성하면 아래와 같다.

事 件	행 위	
	A_1 : 기업을 자신이 한다.	A_2 : 特許權을 판다.
θ_1 : 아주 잘 팔린다.	₩ 800,000	400,000
θ_2 : 보통으로 팔린다.	200,000	70,000
θ_3 : 보통이하로 팔린다.	-50,000	10,000

또 各 自然의 狀態가 일어날 수 있는 가능성의 程度 즉 確率이 다음과 같다고 하자 이와같은 確率을 事前確率(prior probability)이라고 한다.

自然의 狀態(事件)	確 率 P(θi)
θ_1 : 아주 잘 팔린다.	0.2
θ_2 : 보통으로 팔린다.	0.5
θ_3 : 잘 팔리지 않는다.	0.3

그러면 行為 A_1에 대한

期待金額値 = (0.2) (₩800,000,000) + (0.5) (₩200,000,000)
+ (0.3) (- ₩ 50,000,000)
= ₩ 245,000,000

行為 A_2에 대한

期待金額値 = (0.2) (₩ 400,000,000)
+ (0.5)(₩ 70,000,000) + (0.3) (₩10,000,000)
= ₩ 118,000,000

와 같이 된다. 이와같은 期待金額値에 기준을 둔다면 行為 A_1 즉 자기가 직접 제품을 생산하기로 결정하게 될 것이다. 이와같이 自然의 狀態에 관한 事前情報를 이용하는 意思決定 分析方法을 事前分析(prior analysis)法이라 한다.

不確定的 狀況下에서 意思決定을 하는데 있어서 유용한 槪念으로 機会損失(opportunity loss)이라는 것이다. 機会損失이라 하는

것은 実際状況에 対応하여 最適의 행위를 취하지 못했기 때문에 생기는 損失을 의미한다. 위에 있는 예에서 실제의 自然状態가 θ_1이라고 할 때 행위 A_1를 택하였다면 이것이 最適의 행위이므로 이때의 機会損失은 ₩0으로 되지만 이때에 행위 A_2를 택하였다면 機会損失은 ₩800,000,000 - ₩400,000,000 = ₩400,000,000으로 된다. 한편 실제 自然의 状態가 θ_3일때 행위 A_1을 택하였다면 機会損失은 ₩10,000,000-(₩-50,000,000) = ₩60,000,000으로 되고 행위 A_2를 택하였다면 ₩0으로 된다. 이와같은 것을 表로 작성하면 아래와 같다. 機会損失은 항상 正의 数로 表示되는 것이며 단순히 損失(loss)이라 한다.

自然의 状態	利得表		機会損失表	
	A_1	A_2	A_1	A_2
θ_1	₩ 800,000	₩ 400,000	₩ 0	₩ 400,000
θ_2	200,000	70,000	0	130,000
θ_3	- 50,000	10,000	60,000	0

이와같은 期待機会損失値를 기준으로 하여도 역시 행위 A_1를 택하는 것이 유리하다. 이와같이 意思決定에 利得에 관한 期待金額値를 使用하거나 또는 損失에 관한 期待値를 사용하거나 최종적으로 意思決定을 하는데 있어서 다를 것이 없다는 것을 알 수 있다. 그런데 機会損失이 期待値를 計算하지 않고 그보다는 하나하나의 機会損失 그 自体에 관심을 갖는 사람이 있다. 이와같은 方法은 어떤 행위에 있어서 일어날 수 있는 가장 큰 損失이 때

로는 치명적인 재기불능의 損失을 가져올 수 있기 때문이다. 행위 A_1을 갖을때 最大機会損失値는 ₩60,000,000이고 행위 A_2는 ₩400,000,000이다. 따라서 可能한 비극적인 最大損失이 最小인 행위는 A_1이라 할 수 있고 이와같이 各 행위에 따르는 最大損失値中에서 最小인 값 ₩60,000,000을 Minimax機会損失値라고 한다. 이러한 Minimax機会損失値를 기준으로 하여 意思決定을 하는 것을 Minimax意思決定法이라 한다.

2) 事後分布 (事後分析)

앞에서 事前確率에 依存하여 어떤 행위에 대한 期待利得을 구하여 意思決定을 하는 方法을 설명하였으나, 여기서는 事前情報 즉 事件이 일어 나는데 대한 事前確率을 이용할 뿐 아니라, 標本抽出 또는 実験을 통하여 얻는 추가적인 情報를 이용하여 各 事件에 대한 事後確率 (posterior probability)을 구하고 이것에 의하여 事後期待利得을 구하여 意思決定을 하는 方法을 설명하고자 한다. 이와 같은 方法을 事後分析 (posterior analysis法이라 하고 이때에 確率論에서의 Bayes公式을 이용하게 된다. 앞에서 発明特許品의 例를 이용하여 설명한다면 우선 発明한 새로운 製品을 試作하여 이에 대한 試験販売를 하여 市場調査를 하여 본다. 그리고 調査結果를 아주 잘 팔리면 x_1으로, 보통으로 팔리면 x_2로 잘 안팔리면 x_3로 表示하기로 한다. 지금 실제 市場調査結果 보통으로 팔려서 x_2라고 하는 結果를 얻었다 하자. 그런데 이와

같은 結果는 어디까지나 試驗調査結果이므로 自然의 狀態를 그대로 반영하고 있다고는 말할 수 없다. 즉 참 상태가 θ_1일 때에도 調査結果 x_2로 되는 일은 있을 수 있는 것이다. 지금 어렇게 되는 確率이 0.1 즉 $P(x_2 \mid \theta_1) = 0.1$, 마찬가지로 真狀態가 θ_2일때 x_2로 되는 確率 $P(x_2 \mid \theta_2) = 0.8$, 真狀態가 θ_3일때 x_2로 되는 確率 $P(x_2 \mid \theta_3) = 0.2$와 같다고 하자. 이와같이 特定한 θi를 條件으로 하고 이 條件下에서 実驗이나 標本抽出에 의하여 어떤 結果를 얻게 되는 確率을 尤度(likelihood)라고 한다. 그리고 이러한 尤度(條件附確率)는 事前確率과 実驗이나 標本抽出結果에서 이론적으로 구해질 수 있는 경우가 많다.

그러면 Bayes公式

$$P(\theta i \mid \chi_2) = \frac{P(\theta i)\ P(x_2 \mid \theta i)}{P(\theta_1)P(x_2 \mid \theta_1) + P(\theta_2)P(x_2 \mid \theta_2) + P(\theta_3)P(x_2 \mid \theta_3)}$$

를 이용하여 다음과 같이 事後確率이 구해진다.

事後確率의 計算

事 件	事前確率 $P_0(\theta i)$	尤 度 $P(x_2 \mid \theta i)$	結合確率 $P_0(\theta i)P(x_2 \mid \theta i)$	事後確率 $P_1(\theta i)=P(\theta i \mid x_2)$
θ_1	0.2	0.1	0.02	0.042
θ_2	0.5	0.8	0.40	0.833
θ_3	0.3	0.2	0.06	0.125
計	1.0		0.48	1.000

여기에서 事後確率은 $P_1(\theta_1) = \frac{0.02}{0.48} = 0.042$

$P_1(\theta_2) = \frac{0.40}{0.48} = 0.833$.

行為 A_1 에 대한 事後期待利得과 事後期待損失

(単位 : ₩1,000)

事 件	事後確率 $P_1(\theta i)$	利 得	損 失	期待利得	期待損失
θ_1	0.042	₩800,000	₩ 0	₩ 33,600	₩ 0
θ_2	0.833	200,000	0	166,600	0
θ_3	0.125	- 50,000	60,000	- 6,250	7,500
	1.000			₩ 193,950	₩ 7,500

$P_1(\theta_3) = \frac{0.06}{0.48} = 0.125$와 같이 구한 것이다. 이와같이 事後確率을 구하면 이것을 利用하여 각 행위에 대한 事後期待利得, 事後期待損失등을 다음과 같이 계산할 수가 있다.

行為 A_2 에 대한 事後期待利得, 事後期待損失

(単位 : ₩1,000)

事 件	事後確率 $P_1(\theta i)$	利 得	損 失	期待利得	期待損失
θ_1	0.042	₩400,000	₩ 400,000	₩ 16,800	₩ 16,800
θ_2	0.833	70,000	130,000	58,310	108,160
θ_3	0.125	10,000	0	1,250	0
	1.000			₩76,360	₩124,960

따라서 행위 A_2에 대한 事後期待利得은 ₩76,360,000, 事後期待損失은 ₩124,960,000과 같게 된다.

2.4.3 Bayesian推定

古典的인 推定方法과 비교하여 說明하고자 한다. 古典的인 統計的 方法에서는 母集団比率 P를 推定하는데 標本比率 $\hat{P}$를 使用한다. 그리고 Bayesian 統計的 方法에서는 推定을 自然의 狀態 P에 대한 하나의 행위로서 규정한다. 또 P가 未知인 경우에는 이것을 하나의 確率変数로 취급한다.

推定하고자 하는 母数의 값을 θ라고 하고 $\hat{\theta}$를 母数 θ의 推定値라고 하자. 그러면 $\hat{\theta}$가 θ와 一致하지 않는 것에서 생기는 損失을 생각할 수가 있을 것이다. 이와같은 損失을 $(\hat{\theta} - \theta)$의 함수로서 정의하는 것이 보통이다. 例를 들면

$$L(\hat{\theta} : \theta) = |\hat{\theta} - \theta|$$

또는

$$L(\hat{\theta} : \theta) = (\hat{\theta} - \theta)^2$$

여기에서 $L(\hat{\theta} : \theta)$는 損失函数를 표시하는 記号이고, 특히 두번째와 같이 정의한 損失函数를 平方誤差損失函数(squared error loss function)라고 한다.

Bayesian統計学者는 우선 $\hat{\theta}$의 모든 可能한 값에 대하여 損失函数의 期待値를 구한다. 이와같이 구한 損失函数의 期待値는 주어진 θ에 대한 條件附, 期待損失値라고 할 수 있으며 이것을

특히 危險値 (Risk) 라고 한다. 여기에서 母数 θ 가 未知일 때는 θ 를 確率変数로 생각하기 때문에 위에 구한 危險値는 θ 에 관한 危險函数 (risk function) 라고 생각할 수 있게 된다.

지금 θ 의 모든 可能한 값에 대하여 事前確率 $P(\theta)$ 가 주어졌다고 생각을 하자. 그러면 이 事前確率을 이용하여 危險値의 期待値를 구할 수 있다. 이것을 Bayes危險値 (Bayes risk) 라고 한다. 그런데 이와 같은 Bayes 危險値는 θ 를 推定하는 方法에 따라 그 값을 달리한다. 모든 推定方法중에서 Bayes危險値를 最小로 하는 推定方法이 있다면 이것이 Bayes推定方法으로 된다.

지금 한 동전의 表面出現 比率 P를 推定하는 문제를 생각한다. n회 동전을 던져서 r회 表面이 나왔다면 古典的인 統計的方法에는 P의 推定値로서 標本比率

$$\hat{P} = \frac{r}{n}$$

를 택한다. 그러나 Bayesian 統計学者는 $L(\hat{P} : P) = (\hat{P} - P)^2$ 와 같은 損失函数를 생각하고, P가 0과 1사이에 均一하게 分布 (이것을 直四角形分布라고 한다.) 되어 있다는 事前確率分布를 부여하여 Bayes 推定方法으로서

$$\hat{P} = \frac{r+1}{n+2}$$

를 얻는다. 여기에서 주목할 만한 것은 n값이 클때는 두가지 方法은 결국 一致한다는 事実이다. 그러나 Bayesian方法에서는

損失函數로서 무엇을 정의하여 사용하느냐 또 事前確率分布를 어떻게 부여하느냐에 따라서 推定方法이 달라진다. 그런데 古典的인 統計学者는 무엇보다도 事前確率分布를 부여하는 그 근거에 대하여 그 不合理性을 문제로 삼는다. 사실 이 점에 대해서는 Bayesian統計学을 신봉하는 사람들도 確信을 가지고 自己主張을 내세울 수 없는 때가 많다. 지금까지는 点推定에 관해서 說明했으나 다음부터는 区間推定에 관해서 說明하기로 한다.

古典的인 統計学理論에 의하면 이와 같은 区間推定에 있어서 母比率 P를 그 区間에 包含하는 確率이 95 %라는 것을 意味한다고 해석하는 것은 잘못이라고 경고하고 있다. 区間推定이 意味하는 것은 区間推定을 되풀이 해서 하면 이와같은 区間이 母數를 包含하게 되는 確率이 95 %라는 것이다. 사실 古典的인 統計学에서는 母數는 어느 特定常数이고 따라서 統計量(推定値)을 條件으로 한 母数 P의 條件附確率같은 것은 생각할 수 없는 것으로 되어 있다. 즉 $P(P \mid \hat{P})$와 같은 것은 생각할 수 없는 것이고 오직 $P(P \mid \hat{P})$만을 생각할 수 있는 것으로 되어 있다. 그러나 Bayesian 統計的 方法은 이와같은 古典的인 方法과는 아주 대조적이다. 母比率 P가 未知일때는 P를 하나의 確率変数로 취급한다. 그러므로 $P(P \mid \hat{P})$와 같은 條件附確率을 서슴치 않고 사용한다. 따라서 P의 事前確率 $P_0(P)$가 주어지면 標本에서 얻은 情報를 기초로 하여 P의 事後確率 $P_1(P)$를 Bayes 公式에

의하여 구할 수 있게 된다. 이와같이 P의 事後分布가 결정되면 確率変数 P를 포함하는 区間, 例를 들면 95 % 信頼区間을 구하는 것은 어렵지 않을 것이며 이것이 古典的인 統計的 方法에 있어서의 信頼区間에 対応하는 것으로 될 것이다. 그런데 재미있는 것은 点推定에 있어서 標本의 크기 n이 대단히 클때는 P가 0과 1사이에서 均一하게 分布한다는 事前確率分布를 부여할 때 古典的 推定方法과 Bayesian 推定方法과는 一致하여 差異가 없다는 것을 알았는데 이러한 관계는 区間推定의 경우에도 성립한다. 그런데 P에 대하여 직사각형분포(均一分布)를 事前確率分布로서 생각하는 것은 P에 대한 어떤 특별한 知識이 없다는 것을 意味하는 것이라 생각할 수 있으므로 이것은 바꾸어 말하면 古典的인 点推定이나 区間推定의 方法은 母数에 대한 合理的인 可能性을 고려하고 있다고는 할 수 없다는 비난을 Bayesian 統計学者들로부터 받게 되는 것이다.

以上에서 論議한 것에서 짐작할 수 있는 바와 같이 古典的인 統計学者와 Bayesian 統計学者들 사이에는 많은 論争이 계속되어 왔고 사용하는 用語도 다르지만 이들 学者들은 다같이 대상으로 하고 있는 自然의 状態에 관한 정보를 수집하기 위하여 標本抽出이나 実験을 하고 이것을 기초로 하여 행위를 決定하는 것과 같은 方法으로 意思決定 問題를 해결하려고 하고 있다. 이 두系統의 学者들은 다같이 意思決定過程에서 주어진 自然의 状態에 대한

標本抽出結果의 條件附確率을 계산한다.

古典的인 統計学者들은 이것을 기초로 하여 意思決定에 따르는 제 1종의 과오 또는 제 2종의 과오의 심각성을 따지게 된다. 그러나 이와 같은 것이 具体的으로 어떻게 決定되어야 하는가에 대하여서는 明確하게 가르쳐 주는 것이 없다. Bayesian 統計学者들은 自然의 状態에 대한 主観的(経験的)인 事前確率을 부여하고 意思決定에 따르는 과오를 나타내는 損失函数를 준비하여 古典的인 의사결정절차를 補完 또는 完成하려는 立場을 취한다고 할 수가 있다. 그러나 이와같은 경우 事前確率과 損失函数가 새로이 論議의 対象으로 제기된다고 할 수가 있다.

古典的 統計学者와 Bayesian統計学者들간에 있어서 중요한 差異点은 事前確率分布를 생각하느냐 안하느냐 하는 点이라고 할 수가 있다. 古典的 統計学者들은 客観的인 確率 또는 相対頻度만이 생각할 수 있는 合理的인 것이라 주장한다. 그들은 主観的 또는 個人的인 確率이 Bayes公式을 이용하여 事後確率을 계산하는데 사용된다는 点이 理解하기 어렵다는 것이다. 그러나 Bayesian 統計学者들은 実際的으로 어떤 문제에 대하여 意思決定을 하는 경우 이미 설명한 바와 같은 Bayesian的 사고의 과정을 거쳐서 하는 것이라고 주장한다. 특히 오직 客観的인

確率만을 이용해야 된다는 견해를 고수한다면 經営이나 経済的인 문제와 관련된 어떤 意思決定問題를 해결하기 어렵다고 주장한다. 그런데 어떻게 事前確率을 부여하느냐 하는 것은 실제적으로 어려운 問題로서 Bayesian 統計的 方法을 신봉하는 사람들도 이 点에 대하여서는 確信이 없는 것이 보통이다. 결국 古典的인 方法과 Bayesian 統計的 方法에 대한 것을 한마디로 비판하고 비교한다는 것은 쉬운 일이 아니다.

第 3 章 統計的 檢定

이 章에서는 不確定性이 介在되어 있는 一部情報에 의하여 合理的인 意思決定을 어떻게 하느냐 하는 것을 다루게 된다. 不確定性을 다루는 데는 이미 알고있는 여러가지 統計量의 確率分布를 活用하게 될 것이다. 第2節에서는 소위 古典的인 統計的 仮説檢定(statistical testing hypothesis)方法을 説明하고, 第3節에서는 어떤 統計量의 確率分布를 仮定하기 어려울 경우에 適用될 수 있는 非母數的 方法(Non-parametric method)을 다루었으며, 또한 仮説檢定에 관한 理論은 하나의 意思決定에 관한 것이라고 할 수가 있다. 그러나 이와같은 仮説檢定에 관한 理論(이것을 伝統的, 古典的 意思決定 方法이라고도 한다)만으로는 実際的인 問題解決에 充分하지 못한 때가 많다. 그러므로 近来에 와서 特히 過去 20餘年間 事業経営, 産業管理, 또는 行政管理上의 意思決定을 돕고 向上시키는데에 必要한 確率과 統計的方法이 크게 要求되어 이와 같은 分野에서 応用되는 統計理論과 方法인 Bayesian 意思決定論(Bayesian decision theory)을 마지막 節에 취급하였다.

3.1 仮説檢定이란?

統計的 決定(statistical decision)을 내리기 위한 方法에서 比較的 単純하게 使用되는 것이 統計的 仮説檢定임을 이미 앞

에서 言及한 바 있지만 이와 같은 統計的 仮説을 設定하는데 있어서는 다음의 例에 의하여 説明하여 보기로 하자.

即, A都市의 T·V 所有世帶 가운데 任意로 300世帶만을 抽出하여 調査한 結果 진공관用 T.V를 가진 世帶가 125이고 電子用 T.V를 가진 世帶가 175라고 하면 이 事実로써 「A都市의 T·V 所有世帶가운데 半数 以上이 電子用 T·V를 가지고 있다」라고 말할 수 있을 것인가? 여기에서 調査対象을 300世帶에서 1,000 或은 2,000世帶로 增加시킨다면 그 結果에 대하여도 電子用 T.V가 半数 以上이다. 라고 決定할 수 있을 것인가 하는 問題가 発生한다.

이러한 경우 電子用 T·V가 1/2以上이라는 仮説을 세우게 되며 이것을 帰無仮説(Null hypothesis)이라고 하여 H_0로 表示하는 것이 普通이며 이에 대하여 標本의 数를 增加시킴에 따라 反対로 진공관用 T·V가 그 差를 좁혀 나아갈지도 모르기 때문에 여기에서도 하나의 仮説로서의 対立仮説(Alternative hypothesis)을 생각할 수 있으며 H_1으로 表示하는 것이 普通이다.

따라서 統計的 仮説検定은 다음 중 하나에 대하여 옳다는 判断을 내리게 될 것이다.

(1) A都市의 T·V의 所有世帶 가운데 1/2以上이 電子用 T·V를 所有하고 있다.

(2) A都市의 T·V 所有世帶 가운데 진공관用 T·V를 所有하고 있는 世帶도 1/2이 된다.

即 (1)은 帰無仮説 H_0를 受落하게 되는 경우이고 (2)는 帰無仮説을 棄却하고 対立仮説 H_1을 採択하는 경우가 되는 것이다.

앞에서 말한바와 같이 統計的 仮説検定은 母集団에서 n個의 標本 $x_1, x_2 \cdots\cdots\cdots\cdots, x_n$을 基礎로 하여 H_0와 H_1의 어느쪽을 採択할 것인가에 있다. 그러므로 標本의 確率分布와 統計量을 適当히 取하게 되면 이와 같은 問題를 解決할 수가 있다.

例를 들어 A都市의 T·V 所有台数의 경우 個別事象의 出現確率이 0.5이고 $n = 300$인 標本集団에서 電子用 T·V (하나의 事象)가 175回 以上 나타날 수 있는 確率이 얼마인가를 求하는 것과 같으므로 正規分布의 平均과 分散을 利用하여

$$\mu = nP = 300 \times 0.5 = 150$$

$$\sigma = \sqrt{npq} = \sqrt{300 \times 0.5 \times 0.5} = 8.66$$

을 計算하고 이것을 面積으로 換算하여 보면

$$Z = \frac{\bar{\chi} - \mu}{\sigma} = \frac{175 - 150}{8.66} = 2.9$$

가 되므로 数値表에서 얻은 2.9 = 0.4981을 가지고 0.5 − 0.4981 = 0.0019가 計算되는 바 電子用 T·V를 所有한 世帯가 175世帯를 초과하는 것은 500回에 1回의 比率을 차지하게 되는 것을 알 수 있다. 그러나 이와 같은 偶然的인 現象인 $P = 1$의 統計的 仮説이 正確하지 못하다고 본다면 이 仮説은 棄却하게 되며 反対로 正確하다고 보면 仮説을 採択하게 되는 것이다.

다시 말하면 이와 같은 判斷(意思決定)들은 어디 까지나 수집된 統計資料의 分析에 의해서 내려지는 것이기 때문에 어느쪽으로 決定이 내리든 간에 거기에는 過誤(잘못)을 저지르게 될 可能性이 항상 있다. 따라서 옳은 仮説을 棄却함으로써 나타나는 過誤를 第1種의 過誤(error of the first kind, or type I error)라 하며 棄却하여야 할 仮説을 取하게 되는 過誤를 第2種의 過誤(error of the second kind, or type II error)라고 한다. 이와 같은 仮説의 採択과 棄却과 이에 따르는 過誤와의 関係를 表示하면 아래 표와 같다.

표 .3.1.1

行 為	自然의 狀態	
	H_0가 真임	H_1이 真임
H_0를 受落	옳은 意思決定	第2種의 過誤
H_0를 棄却	第1種의 過誤	옳은 意思決定

여기에서 第1種의 過誤를 저지르는 確率을 危険率(risk ratio) 或은 有意水準(significant level)이라하며 仮説検定에 있어서의 判斷의 基準으로 삼고있는 것이 普通이다.

이 有意水準의 값으로는 5% 또는 1%로 말하는 것이 보통이다. 그런데 이 有意水準을 작게하면 第2種의 過誤가 커지는 경향이 있다.

또한 標本에서 求하여진 実測値의 判定基準에 따라 仮説을 取하

거나 棄却하게 되는데 이때 棄却하는 領域을 棄却域(critical region)이라고 하여 다음의 그림 3.1.1의 斜線部分에 해당한다.

一般的으로 Neyman－Pearson流의 検定은 標本의 数 n과 有意水準 α를 주어진 것으로 하는 가운데 第2種의 過誤로 β가 最小가 되게하는 方法을 使用하고 있는 바 正規母集団을 構成하는 特性値를

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Xi\,, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

이라고 하면 統計量 T는

$$T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n-1}}$$

로 表現되며 $100\,\alpha\,\%$ 点을 t_0라 할 때

$|t| > t_0$ 이면 H_0를 棄却하고

$|t| < t_0$ 이면 H_0를 採択하게

된다.

이와 같은 検定法에서는 一般的으로 片側検定(one tail-test)과 両側検定(two tail - test)을 取하고 있으며 有意水準 α가 주어진 경우

$$\left.\begin{aligned} P(t > t_0) &= \alpha \\ P(t < t_1) &= \alpha \end{aligned}\right\}$$

는 片側檢定(one tail-test)을 나타내며

$$P(t > t_2) = P(t < t_3) = \frac{\alpha}{2}$$

는 兩側檢定을 表示한다.

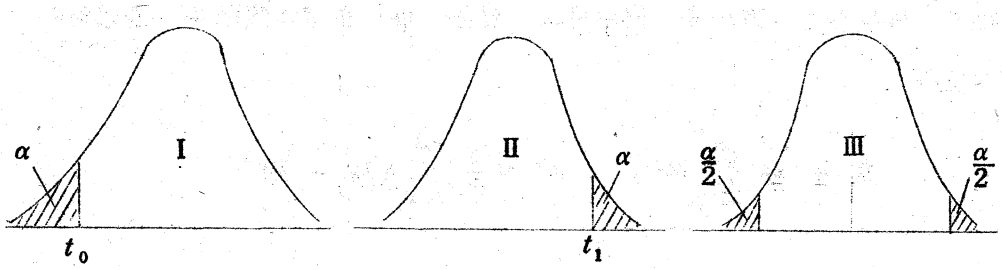


그림 3.1.1

그러므로 有意水準 5%인 경우의 右片側檢定에서는

$t \geq 1.645$ 이면 H_0를 棄却하고

$t < 1.645$ 이면 H_0를 採択하며

左片側檢定에서는

$t \leq -1.645$ 이면 H_0를 棄却하고

$t > -1.645$ 이면 H_0를 採択된다.

또한 兩側檢定에 있어서는

$t > 1.96$과 $t < -1.96$이면 H_0를 棄却하고,

$-1.96 < t < 1.96$ 이면 H_0를 採択하게 된다.

3.2 母数 検定

(1) 平均의 検定

平均의 検定이란 母平均이 주어진 값 μ와 같은가 그렇지 않은가를 検定하는 것이다. 例를들어 서울特別市 中区에 居住하는 住民 100名을 選出하여 나이를 調査한 結果 平均 나이가 30세인 경우, 서울 特別市 全體 市民의 平均나이가 30세라고 할 수 있을 것인가 하는 問題가 제기 되었을 때 간단히 肯定할 수는 없는 것이다. 이것은 母集団과 標本사이에 発生하는 誤差로 하여 母平均과 標本平均이 같을 수도 있으나 다를 수도 있기 때문이다.

(A) 大標本인 경우

任意로 抽出한 10名의 学生에 대한 平均身長이 164cm이다. 이것으로 全国学生의 平均身長이 162cm보다 크다고 할 수 있을 것이다.

이 때의 標準偏差를 5cm라고 하면 母平均이 162cm 이고 確率比率이 $\sigma/\sqrt{n} = 5/\sqrt{100} = 0.5$인 正規分布가 되므로 이를 t로 変換

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{164 - 162}{0.5} = 4$$

를 얻는다. 이 값의 確率은 0.0002가 되어 標本平均 164cm의 学生을 $\mu = 162$cm 이고 $\sigma = 5$cm인 母集団에서 抽出한다는 仮説은 棄却되며 100名의 学生에 대한 平均身長은 全国의 平均身長보다

크다는 것을 알 수 있다.

一般的으로 母平均 μ와 分類 σ^2이 얻어지고 n이 큰 경우에는 有意水準 1%에서는 標準偏差 σ의 2.58倍보다. 큰 경우이 仮説은 棄却되는 것이다.

即 $P\{|\bar{x}-\mu| > 2.58\,\sigma/\sqrt{n}\} = 0.01$이 되므로

① $|\bar{x}-\mu| > 2.58\,\sigma/\sqrt{n}$인 경우에는 H_0는 棄却되고

② $|\bar{x}-\mu \leq 2.58\,\sigma/\sqrt{n}$인 경우에는 H_0는 採択된다.

(B) 小標本인 경우

위의 例를 그대로 利用하여 任意로 抽出한 学生을 10名으로 하고 分散 25㎠에 대하여 検定하면 小標本인 경우의 検定을 説明할 수 있다.

即 이 標本의 確率分布는 $n<30$이므로 正規分布를 이루지 못한데다 母分散이 알려져 있지 않기 때문에 σ에 代値하여 標本標準偏差 S를 利用하면

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

를 얻을 수 있으나 確率変数 t는 標準正規分布를 하지 않을 뿐 아니라 n에 따라서 그 分布状態가 달라진다는 것이 W·S·Gosset (筆名 "Student") 1908이후 R·A· Fisher (1926) 에 의해 完全히 究明되 이와 같은 確率変数 t의 分布를 自由度 $n-1$의 "Student t-分布"라고 한다. 따라서 10名의 学生에 대한 母集団平均 μ_1이 $\mu=162$㎝와 같다고 하는 前提아래 計測하면

$$S^2 = 1/(10-1) \times 10 \times 25 = 27.78$$

$$t = \frac{164 - 162}{\sqrt{27.28}/\sqrt{25}} = 1.897$$

을 얻게 되며 自由度 10 −1 =9인 t 分布는 数値表에서 $t_0 = 3.25$ 보다 큰 確率이 0.01 이므로 $t = 1.897$ 보다 큰 確率은 0.01 보다 큰 것이라는 것을 알 수 있다. 따라서 이 仮説은 棄却할 必要가 없으며 母平均이 162㎝의 母集団에서 抽出한 標本이 10個인 경우에는 標本平均이 164㎝의 数値를 나타낼 수 있다는 것을 보여 준다.

一般的으로 크기가 n인 標本平均 $\bar{x}$가 母平均 μ의 集団에 속하는가 그렇지 않는가를 決定하는데 있어서는 「標本平均 $\bar{x}$의 母平均은 μ이다」라는 仮定아래

① $|t| > t_0$ 이면 H_0를 棄却하고

② $|t| < t_0$ 이면 H_0를 採択한다.

(2) 比率의 検定

比率의 検定은 간단한 例로서 설명해 보기로 한다.

即, 어떤 生産工場에서 A製品을 使用하는 需要者가 80% 되리라고 생각하여 標本으로 100世帯를 抽出하여 調査한 結果 A製品을 使用하는 世帯가 72世帯였다면 帰無仮説은 採択되어질 수 있을까 하는 問題가 된다.

우리는 仮説로써 $H_0 : P = 0.8$을 設定하게 되며 이 때의 有意

水準을 5%라고 하면 標本 x_1, x_2 ……, x_n 에 대한 統計量 Z의 函数는

$$Z = \frac{\bar{x} - np}{\sqrt{np(1-p)}}$$

가 되며 이를 計算하면

$$Z = \frac{72 - 100(0.8)}{\sqrt{100\ (0.8)\ (0.2)}} = -2.00$$

을 얻는다. 그런데 正規分布의 3σ범위는 片側検定의 0.5 − 0.05 = 0.45로서 ± 1.64를 얻게 되므로 다음과 같은 判定을 提示할 수 있다.

即 「72世帶가 A製品을 使用하는 것으로 나타난 Z分布는 −2.00이며 標本에서 얻은 값은 ±1.64이므로 帰無仮説을 기각되지 않을 수 없으며 生産業者 危険을 招来하고 있다」라고 생각할 수 있으므로 이 生産工場으로서의 統計的 決定은 잘못된 것임을 알 수 있다.

(3) 平均値差에 대한 検定

(A) 大標本일 경우

어느 두 地方 A, B에 있어서의 企業體들 사이에 未熟練工에게 支払하는 月平均賃金에 差異가 있는가 하는 것을 알아보기 위하여 이 두 地方에서 각각 100名, 200名을 標本으로 취하여 그들에게 支給되는 賃金을 調査하여 다음과 같은 結果를 얻었다.

지금 A, B 두 地方에서 月平均賃金에 관한 母平均을 각각 μ_1,

地方	平 均	標準偏差	標本의 크기
A	$\bar{x}_1 = ₩18,002$	$S_1 = ₩800$	$n_1 = 100$
B	$\bar{x}_2 = ₩17,042$	$S_2 = ₩900$	$n_2 = 200$

μ_2라하고 母標準偏差를 각각 σ_1, σ_2라고 하자. 그러면 이 問題에서 検定하려는 仮説은 다음과 같이 된다.

帰無仮説 $H_0 : \mu_1 = \mu_2$

対立仮説 $H_1 : \mu_1 \neq \mu_2$

数理統計学에 의하면 平均과 標準偏差가 각각 μ_1, σ_1 그리고 μ_2, σ_2인 두 正規母集団에서 각각 크기 n_1, n_2의 서로 独立인 任意標本을 취하였을 경우의 標本平均을 각각 $\bar{x}_1$, $\bar{x}_2$로 表示한다면 標本平均値差 $\bar{x}_1 - \bar{x}_2$의 分布는 다음과 같이 된다. 즉 $\bar{x}_1 - \bar{x}_2$는 平均이 $\mu_1 - \mu_2$, 標準偏差가 $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$인 正規分布를 한다. 이것은

또

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

로 놓으면 Z는 平均이 0, 標準偏差가 1인 소위 標準正規分布를 한다.

그런데 대개의 경우 두 母集団의 標準偏差 σ_1, σ_2를 알지 못하는 것이 보통일 것이다. 그러나 이러한 경우에도 標本의 크기가

를 때에는 ($n_1 > 30$, $n_2 > 30$), σ_1, σ_2 대신 標本標準偏差 S_1, S_2를 그대로 使用하여도 무방하다. 따라서, 위의 例의 仮説檢定은 H_0 하에서 $\mu_1 - \mu_2 = 0$이므로

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{18,002 - 17,042}{\sqrt{\frac{(800)^2}{100} + \frac{(900)^2}{200}}} = 9.4$$

여기에서 有意水準을 1%로 한다면 數値表에서 $P(|Z| \geq Z_{.01}) = 0.01$로 되는 $Z_{.01}$값으로 2.58을 얻는다. 즉 A, B 두 地方에서의 未熟練工에 대한 賃金差가 있다는 結論을 얻게 된다.

(B) 小標本일 경우

平均이 각각 μ_1, μ_2이고 標準偏差가 각각 σ_1, σ_2인 두 개의 正規母集団에서 標本의 크기가 각각 n_1, n_2 ($n_1 < 30$, $n_2 < 30$)인 独立標本에서의 標本平均値를 $\bar{x}_1$, $\bar{x}_2$ 그리고 σ_1^2, σ_2^2에 대한 不偏推定値를 각각 S_1^2, S_2^2라고 하자. 이때 n_1, n_2가 크지 않으므로

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

가 標準政規分布에 近似的으로 따른다고 할 수는 없는 것이다. 그러면 t -分布를 하는 것일까? 여기에서 만약 $\sigma_1 = \sigma_2 = \sigma$와 같은 仮定이 成立한다면 大標本에서 Z는 즉

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

와 같이 된다. 이와 같은 경우 共同分散 σ^2에 대한 不偏推定値 S^2은 다음과 같이 구한다.

$$S^2 = \frac{(n_1 - 1)\, S_1^2 + (n_2 - 1)\, S_2^2}{n_1 + n_2 - 2}$$

그래서 σ_2에 S^2을 代入한 式

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

는 自由度 $n_1 + n_2 - 2$인 t-分布를 하게 된다.

例. B.1 交信方法과 問題를 解決하는 能力에 관한 研究結果를 보자. 実驗者는 任意로 세 集団을 抽出하고 각 集団에 똑같은 問題를 解決하도록 한다. 그러나 각 集団内에서의 交信方法은 集団別로 다르게 하였다. 즉 첫째 集団은 모든 方法을 다 쓰도록 (all channel pattern) 하였고, 두번째 集団에서는 "차바퀴" 패턴 (wheel pattern)을, 세째 集団에서는 "원" 패턴 (circle pattern)에 의하여 交信토록 하였다. 다음表는 제 2, 제 3 集団에 대한 実驗結果이다.

問題解決에 所要되는 時間 (単位 : 分)

交信패턴	平 均	標準偏差	標本의 크기
차바퀴	$\bar{x}_2 = 19.12$	$S_2 = 3.20$	15
원	$\bar{x}_3 = 29.45$	$S_3 = 5.21$	21

이때 検定하려는 仮説은 다음과 같을 것이다.

帰無仮説 $H_0 : \mu_2 = \mu_3$

対立仮説 $H_1 : \mu_2 \neq \mu_3$

그리고 이와 같은 仮説에 대한 検定은 위의 t 값을 計算하는 것으로 이루어진다. 그런데 이에 앞서서 두 母分散 σ_2^2, σ_3^2의 同一性을 確認할 必要가 있다. 그러나 여기에서는 $\sigma_2^2 = \sigma_3^2 = \sigma^2$이라고 仮定하고 検定節次를 進行하기로 한다.

우선 σ^2에 대한 推定値 S^2을 구한다.

$$S^2 = \frac{(14)(3.20)^2 + (20)(5.21)^2}{15 + 21 - 2} = 20.1836$$

그러므로 帰無仮説下에서 t 값은 다음과 같이 計算된다.

$$t = \frac{|19.12 - 29.45|}{\sqrt{20.1836\ (\frac{1}{15} + \frac{1}{21})}} = 6.8$$

한편 有意水準 5%下에서 自由度 34일 때의 $t_{.05}$의 값은 2.042정도 이므로 결국 帰無仮説 $\mu_2 = \mu_3$는 棄却된다.

(C) $\sigma_1 \neq \sigma_2$일 때의 検定法

두 独立任意標本의 平均値差를 이용하여 母平均値差에 대한 検定을 한다든가 또는 信頼区間의 구하는 데 있어서 두 母集団標準偏差가 서로 같다고 하는 仮定이 의심스러운 때가 많다. 例를 들면 (1) 性質이 다른 두 母集団에 있어서는 이들 各母標準偏差가

서로 같다고 보기 어려운 때가 많다. 한 가지 例로 私立国民学校 学生과 公立国民学校 学生 사이의 어떤 特性을 比較하려고 할 때, 私立学校 学生들 間의 標準偏差와 公立学校 学生들 間의 標準偏差가 같으리라고는 믿기는 어려울 것이다. (2) 두 母集団의 각각의 母平均에 대한 信頼区間을 구하였을 때 区間의 幅이 서로 크게 다른 경우 母平均値間의 差異에 따라서 母標準偏差에도 差異를 나타내는 일이 많다. 따라서 이와 같은 경우 $\sigma_1 = \sigma_2$라고 仮定하기는 어렵다. (3) 母集団의 分布状態가 対称型에서 크게 벗어나 있을 때는 平均μ와 標準偏差σ 사이에 관련성이 농후한 때가 많다. 즉 $\mu_1 = \mu_2$이면 $\sigma_1 > \sigma_2$와 같은 관계가 있기 쉬운 것이다.

그런데 $\sigma_1 \neq \sigma_2$라고 생각될 경우에도 두 独立任意標本에 있어서 標本平均値差 $\bar{x}_1 - \bar{x}_2$의 分散은 역시 $\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}$와 같이 된다. 그러나 이와 같은 경우 $\mu_1 = \mu_2$인 帰無仮説下에서

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

는 理論上 "Student"의 t-分布를 하지 않는 것으로 되어 있다.

이와 같은 경우 $H_0 : \mu_1 = \mu_2$를 検定하는 問題에 대하여 Behrens와 Fisher의 解法이 있고 그 外의 여러가지 方法이 있으나 여기에서는 Cochran의 方法을 소개하기로 한다.

$n_1 = n_2 = n$의 경우 이때에는 $S^2 = \frac{S_1^2 + S_2^2}{2}$

으로 놓으면 위의 式은

$$t' = \frac{\sqrt{n}\,(x_1 - x_2)}{S\sqrt{2}}$$ 이다.

이것은 $\sigma_1 = \sigma_2 = \sigma$일 때의 式과 다를 것이 없는 것이지만 그러나 이 t'의 값은 自由度 $n-1$ ($2n-2$가 아님) 인 t — 分布에 近似하게 分布한다는 것을 利用하여야 한다.

$n_1 \neq n_2$일 경우 例B1에서 $\sigma_1 \neq \sigma_2$라고 仮定하면 우선

$$t' = \frac{|\,19.12 - 29.45\,|}{\sqrt{\frac{(3.20)^2}{15} + \frac{(5.21)^2}{21}}} = 7.35$$

를 구한다. 이때 t'에 대한 有意水準 값은 다음과 같이 計算한다.

$$\frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

여기에서 $w_1 = S_1^2/n_1$, $w_2 = S_2^2/n_2$이고, t_1, t_2는 自由度 $n_1 - 1$, $n_2 - 1$에 対応하는 t 分布表에서의 값 (有意水準 5%) 이다.

이 例에서는

$t_1 = 2.145$, $t_2 = 2.086$

그리고 $w_1 = 0.6827$, $w_2 = 1.2926$

그러면

$$t'_{.05} = \frac{(0.6827)(2.145) + (1.2926)(2.086)}{0.6827 + 1.2926} = 2.1064$$

와 같이 되고 결국

$$t' > t'_{.05}$$

와 같이 되므로 이때에도 帰無仮説은 棄却된다.

(4) 比率의 差에 대한 検定

母比率이 P_1, P_2인 두 母集団이 있다. 이들 두 母集団에서 크기가 각각 n_1, n_2 ($n_1 > 30$, $n_2 > 30$)인 任意標本을 抽出하여 標本比率 P_1, P_2를 얻었다고 하자. 그러면 n_1, n_2가 커지면 이에 따라 P_1, P_2의 分布는 각각 平均이 P_1, P_2, 分散이 P_1Q_1/n_1, P_2Q_2/n_2인 正規分布에 接近하여 간다. 여기에서 $Q_1 = 1 - P_1$, $Q_2 = 1 - P_2$이다. 그러므로 標本比率差 $P_1 - P_2$는 平均이 $P_1 - P_2$, 標準偏差가 $\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$ 인 正規分布에 가까워 진다.

여기에서 다음과 같은 仮説検定을 생각한다.

帰無仮説 $H_0 : P_1 = P_2$

이와같은 경우

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$$

가 近似的으로 標準正規分布를 한다는 것을 利用할 수 있을 것이

라고 생각된다. 그러나 母比率 P_1, P_2는 알지 못한다. 그러므로 H_0下에서 $P_1 = P_2 = P$의 推定値 P를 구하려고 할 것이다. 이때 推定比 P는 다음과 같이 구해진다.

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

그리고 n_1, n_2가 다같이 클 때에는

$$Z = \frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad q = 1 - P$$

가 標準正規分布에 接近하여 간다는 事実에 의하여 H_0를 検定하게 되는 것이다.

例 (4). 1 美国南部에서의 한 選挙에서 白人投票者 중에서 任意로 抽出된 655名중 203名이 民主党 후보에게 投票를 했고 黒人投票者중에서 任意로 抽出된 480名중 173名이 민주당 후보에게 投票를 하였다는 것을 알고 있다. 黒人의 民主党 支持度가 白人의 그것보다 높다고 할 수 있을 것인가?

민주당원에 投票한 標本比率은 각각 다음과 같다.

$$P_1 = \frac{203}{655} = 0.31, \quad P_2 = \frac{173}{480} = 0.36$$

그리고 이때 検定을 위한 仮説은 다음과 같다.

$$H_0 : P_1 = P_2, \quad H_1 : P_1 \neq P_2$$

이 例에서 帰無仮説 $P_1 = P_2 = P$下에서 P의 推定値는

$$P = \frac{203 + 173}{655 + 480} = 0.33$$

와 같이 되므로

$$Z = \frac{|\,0.31 - 0.36\,|}{\sqrt{(0.33)(0.67)\left(\frac{1}{655} + \frac{1}{480}\right)}} = 1.8$$

한편 正規分布表에서 有意水準 5 %에서 $Z_{.05} = 1.96$을 얻으므로 결국 $Z < Z_{.05}$ 와 같이 되어 帰無仮説 $P_1 = P_2$는 棄却되지 않는다.

(5) 分散의 同一性에 관한 検定

$\sigma_1 = \sigma_2$와 같이 仮定하기가 어려울 때는 帰無仮説 H_0 : $\sigma_1 = \sigma_2$, 対立仮説 H_1 : $\sigma_1 \neq \sigma_2$를 세우고 이것을 独立任意標本에서 計算한 標本分散 S_1^2과 S_2^2에 의한 分散比 $F = S_1^2 / S_2^2$ (단 $S_1^2 > S_2^2$일때)를 利用하여 仮説検定을 한다. 統計量 F의 確率分布는 $Z = 1/2 \log_e F$와 같이 変換한 다음에 Z의 分布로서 1924年 R.A. Fisher에 의하여 유도되었다. 이와 같은 Z-分布 대신에 F의 直接的인 確率分布를 사용한 사람은 G.W. Snedecor (Statistical Methods, 1940)이다. 이와 같은 事実에서 "Snedecor의 F-分布"라는 名称이 붙게 되었다. F-分布는 x^2-分布나 t-分布와 같이 統計学理論과 実際面에서 重要한 役割을 하는 分布이다.

例 B.1에서 H_1 : $\sigma_1 \neq \sigma_2$에 대한 H_0 : $\sigma_1 = \sigma_2$를 検定해 보기로 한다.

$$S_1^2 = (3.20)^2 = 10.2400$$

$$S_2^2 = (5.21)^2 = 27.1441$$

을 얻는다. 따라서

$$F = S_2^2/S_1^2 = 2.65$$

와 같은 分散比를 얻고 自由度는 分子 $f_1 = 21 - 1 = 20$, 分母 $f_2 = 15 - 1 = 14$이다. 한편 数値表에서 $f_1 = 20$, $f_2 = 14$에 대한 $F_{.05}$값은 보간법에 의하여 2.86으로 된다는 것을 알 수가 있다.

즉 $F < F_{.05}$, 따라서 이 例에서는 仮説 H_0가 기각되지 않는다

그런데 이 例에서는 検定이 両側検定이지만 많은 경우 対立仮説로서 $\sigma_1 < \sigma_2$와 같이 設定되는 것이 많기 때문에 単側検定을 하는 때가 많다.

3.3 非母数的 検定

(1) χ^2-検定

非母数法에 의한 有意性 検定中 応用範囲가 가장 넓은 것이 χ^2-検定이다. 이 χ^2-検定은 1900年 K.Pearson에 의해서 처음 考察된 것으로 χ^2-分布를 応用하여 몇 가지 属性間에 独立性이 認定되는 가를 検定하고, 実験統計가 理論的인 既知分布에 어느 정도 適合되는가를 検定하는데 使用되고 있다. 以下에서는 먼저 χ^2이 가지는 意味를 説明하고, 適合度의 検定과 独立性의 検定(分割表의 検定)에 대하여 詳述한다.

(A) χ^2의 意味

어떤 標本에 있어서 事象 E_1, E_2……, E_k의 実際觀察度数를 f_1, f_2……, f_k라 하고 期待度数 또는 理論度数를 F_1, F_2 ……, F_k라 하였을 때, 経験値와 理論値와의 差의 自乘을 理論値로 나눈 값의 合計, 即 χ^2은 다음과 같이 定義할 수 있다.

$$\chi^2 = \sum_{i=1}^{k} \frac{(fi - Fi)^2}{Fi}$$

여기에서 觀察度数 fi는 離散変数이므로 위의 식에 의한 χ^2도 離散変数이다. 그리고 위 式에서 $\chi^2=0$이면 理論度数와 觀察度数는 一致하나 $\chi^2>0$이면 両者는 一致하지 않고 χ^2의 값이 클수록 両者의 差가 큼을 意味한다.

(B) 適合度의 検定

適合度의 検定 또는 度数分布表의 適合度 検定이란 経験的인 度数分布가 理論的인 分布에 適合하는가 않는가를 觀察하는 것으로, 이러한 度数分布의 適合度의 検定은 二項分布, 포아슨分布, 正規分布, 其他 任意의 理論的 分布에 대해서도 마찬가지로 検定을 할 수 있고 또 離散分布나 連続分布의 경우에도 할 수 있다.

이 경우 理論的 分布가 주어져 있는 경우(単純仮説)와 이것이 주어지지 않아서 標本觀察의 結果에서 이것을 推定하여 하는 경우(複合仮説)의 두가지가 있으나 後者가 一般的이다. 여기에서는

檢定方法과 二項分布 및 正規分布의 適合度의 檢定에 대해서 간단히 說明하고자 한다.

① 檢定方法

檢定假說을 먼저 세운후 이 假說에 의한 理論度數 F를 求하여 觀察度數 f와 理論度數 F와의 差의 平方을 F로 除하여 檢定統計量 χ^2을 구한다. 다음 自由度 υ를 구한다. 自由度 υ는 階級個數 k, 理論度數를 定하는 既知의 母數를 ℓ이라면 $\upsilon=(k-1)-\ell$이다. 그러나 理論度數 가운데 5以下의 度數가 포함되어 있을 때에는 度數를 適当히 併合하여 모두 5以上 되게 해야 한다. k가 5보다 작으면 理論度數는 相当히 커야 한다. 여기에서 有意水準을 α라 하면 自由度 $(k-1)-\ell$의 χ^2數値表에서 χ^2_α를 구한 다음

$\chi^2>\chi^2_\alpha$ 이면 假說을 棄却

$\chi^2\leq\chi^2_\alpha$ 이면 假說을 採択한다는 方法이다.

② 二項分布에 관한 適合度의 檢定

觀察度數의 総合을 N이라고 할때 確率変数 $x=0,1,2$ ……의 確率을 ${}_nC_x\,p^x q^{n-x}$에 의하여 求하고 여기에 N을 乘하면 確率変数 $x=0,1,2$ ……의 理論度数가 된다. 즉 理論度数 $F_x=N\,{}_nC_x\,p^x q^{n-x}$로 計算된다. 그리고 期待度数(理論度数)中에 5以下가 있으면 이웃칸의 期待度数와 合쳐서 5以上이 되도록 한다.

自由度는 二項分布의 경우 母数는 n, p의 2개이나 p의 값은

未知의 경우가 많고 既知母数는 上述한 바와 같이 n 하나라면 $\ell = 1$ 이므로 $\upsilon = k - 1 - 1 = k - 2$ 이다. 그리고 다음 節次는 検定方法에서 詳述한 것과 같이 한다.

③ 正規分布에 관한 適合度의 検定

正規分布에서는 標準正規変数

$$Z = \frac{x - \bar{x}}{\sigma \text{ (或은 } S)}$$

를 가지고 数値表에서 各級을 確率을 구한 다음 여기에 n을 곱하면 理論度数 F를 구할 수 있으므로 이 理論度数와 観察度数 f를 가지고 위의 検定方法에 따라 計算한다. 그리고 이 경우 既知의 母数는 μ, σ의 2개이므로 自由度는 $\upsilon = k - 1 - 2 = k - 3$이 된다.

(C) 独立性의 χ^2検定

크기 n인 標本에 대하여 要因 x, y에 대하여 아래와 같은 分割表가 주어져 있을 경우 分割表에 나타난 関係가 有意인가 아닌가를 検定하려 할때에 이 2分類標準이 서로 無関係라고 仮定하고 이것을 検定仮説로 하여 χ^2-検定을 하는 것을 말한다. 이때의 分割表는 2개의 質的分類를 組合한 2重分類表이다.

y \ x	x_1	x_2	……	x_j	……	x_l	計
y_1	f_{11}	f_{12}	……	f_{1j}	……	f_{1l}	f_1
y_2	f_{21}	f_{22}	……	f_{2j}	……	f_{2l}	f_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_i	f_{i1}	f_{i2}	……	f_{ij}	……	f_{il}	f_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
y_k	f_{k1}	f_{k2}	……	f_{kj}	……	f_{kl}	f_k
計	$f_{\cdot 1}$	$f_{\cdot 2}$		$f_{\cdot j}$	……	$f_{\cdot l}$	n

検定方法은 위 表에서 觀察度数 f_{ij}에 대한 理論度数 F_{ij}를 다음과 같이 $F_{ij}=f_{i\cdot}f_{\cdot j}/n$을 구한 다음 x와 y가 無関係라는 것을 検定仮説로 하여 検定統計量을 다음 式으로 計算한다.

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij}-F_{ij})^2}{F_{ij}}$$

여기서는 k行 ℓ列 分割表의 경우이므로 自由度 υ는 $(k-1)(\ell-1)$이며 以下는 適合度検定에서의 節次와 같다.

例. 95명을 머리카락의 빛깔과 눈의 빛깔에 따라 分類하여, 다음表를 얻었다. 머리카락의 빛깔과, 눈의 빛깔은 無関係일까?

解. $m=3$, $n=2$이고, $N=95$인 경우이다. 여기에서 仮説은 「頭髮色과 眼色은 서로 無関係이다」이고 理論度数의 計算은

眼色 ＼ 頭髮色	金　髮	黑　髮	計
青　色	32 (24.1)	12 (19.9)	44
茶　色	14 (19.7)	22 (16.3)	36
灰　色	6 (8.2)	9 (6.8)	15
計	52	43	95

青眼金髮　52/95 × 44 = 24.1

茶眼金髮　52/95 × 36 = 19.7 이며 나머지는 다음과 같이 정해진다.

灰眼金髮 = 8.2　　青眼黑髮 = 19.9

茶眼黑髮 = 16.3　　灰眼黑髮 = 6.8

이 理論度数를 해당하는 欄의 ( ) 안에 적어 넣었다. 위의 6 個의 理論度數중에서 自由로 정할 수 있는 것은 2個 뿐이다. 즉 自由度는 2 × 1 = 2 이다.

이 觀察에서의 χ^2값은

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(f_{ij}-F_{ij})^2}{F_{ij}} = 10.67$$

自由度 2일 때의 χ^2-分布의 5%점은 $\chi^2_{.05} = 5.99$ 이므로 仮説은 5% 有意水準에서 棄却된다. 即 두 빛깔 사이에는 서로 関聯性이 있다는 말이다.

① 2×2 分割表의 簡便式

이때에는 다음과 같이 計算이 簡便化 된다.

y \ x	x_1	x_2	計
y_1	f_1	f_2	f_1+f_2
y_2	f_3	f_4	f_3+f_4
計	f_1+f_2	f_2+f_4	$\sum f=n$

即

$$\chi^2=\frac{n\,(f_1f_4-f_2f_3)^2}{(f_1+f_2)\,(f_3+f_4)\,(f_1+f_3)(f_2+f_4)}$$

이때 自由度는 $\upsilon=1$ 이다. 그리고 検定方法은 앞에서와 같다.

2×3 分割表의 경우에도 簡易하게 하는 方法이 있다.

② Yates의 修正係数

그런데 2×2分割表와 같이 階級別 度数가운데에 5 보다 작은 것이 있을 때는 元来 Stirling의 公式을 使用함으로써 誘導된 x^2分布表를 그대로 利用함에는 너무나 誤差가 클것이므로 所謂 Yates의 修正을 하는 例가 많다.

이는 곧 真正한 数値(修正値) x의 数量을 現実的인 度数 n에 対하여 $(n-0.5)<x<(n+0.5)$로 생각하고 処理하는 方法이 된다. 즉 修正値 χ^2은

$$\chi^2(\text{수정치})=\sum_{i=1}^{k}\frac{(|f_j-F_i|-0.5)^2}{F_i}$$

그러나 Yates의 修正은 2×2分割表 以外에는 一般的으로 実施되지 않는다.

③ 比率差의 検定

여기서는 比率의 差에 관한 検定을 x^2一分布를 利用하여서도 할 수 있다는 것을 보여주기로 한다. 우선 例B.1을 다음과 같이 2×2分割表로 정리한다.

	白 人	黒 人	計
민주당 후보에게 투표한 사람수	203	173	376
민주당 후보가 아닌 자에게 투표한 사람수	452	307	759
計	655	480	1,135

이때 다음과 같은 x^2값을 計算하여 検定할 수 있는 것이다.

$$\chi^2 = \sum (f - F)^2 / F$$

단. f는 観測度数 203, 452, 173, 307을 表示하고 F는 帰無仮説下에서 期待되는 理論度数를 나타낸다.

白人이나 黒人이나 민주당 후보에 投票하는 比率은 같다는 仮定下에서의 投票率 P는 위에서 説明한 바와 마찬가지로 表의 合計란을 利用하여 376/1135으로 된다. 따라서 白人 655名中에서 이러한 比率로 민주당 후보에 投票하는 人員数는 655 × 375/1135 = 216.99와 같이 計算된다. 이것이 観測度数 203에 대

하여 기대되는 理論度数(期待度数)인 것이다.

어떤 觀測度数에 대한 期待度数를 구하는 方法은 그가 속해있는 行과 列의 合計를 서로 곱한 다음에 総計로 나누어서 구해진다. 黒人 480名중 민주당 후보에 投票하는 期待度数는 이것이 속해 있는 行의 合計 376과 列의 合計 480을 서로 곱한 다음에 総計 1,135로 나누어서 구해진다. 即 376 × 480 ／ 1135 = 159.01과 같이 計算된다.

아래 表는 f값과 이에 代応하는 F 그리고 $f-F$값을 表示한다.

f 。F 및 $f-F$表

203	173	261.99	159.01	-13.99	13.99
452	307	438.01	320.99	13.99	-13.99
f		F		$f-F$	

그런데 $(f-F)^2$은 어느 것이나 다같이 $(13.99)^2$이라는 것을 알 수가 있다. 2×2 分割表에서는 언제나 이와 같이 된다는 것이 特徴이다. 그러므로 2×2分割表에서의 x^2-값 計算公式을 다음과 같이 고쳐서 使用하는 것이 편리하다.

$$x^2 = (f-F)^2 \sum \frac{1}{F}$$

$$= (13.99)^2 \left(\frac{1}{216.99} + \frac{1}{438.01} + \frac{1}{159.01} + \frac{1}{320.99}\right)$$

$$= 3.2098$$

그러면 이때 自由度는 위表에서 알 수 있는 바와 같이 偏差 $f-F$는 符号를 除外하고는 똑같은 값을 갖는다.

그러므로 Fisher에 의하여 証明된 바와 같이 自由度는 단지 1이다. 따라서 自由度 1인 χ^2-分布表에서 有意水準 5%인 값 3.84와 비교해 보면 $\chi^2 < \chi^2_{.05}$이므로 仮説은 기각되지 않는다. 그런데 이와 같은 경우 χ^2-값의 連続性에 대한 補正이 必要하게 된다. 補正된 χ^2값은 다음과 같이 된다.

$$\chi_c^2 = (|f - F| - 0.5)^2 \sum \frac{1}{F}$$

$$= 2.98$$

(2) 符号検定(The sign test)

어느 特徴을 数量的으로 表示할 수 없을 때가 간혹 있으나 좋다, 나쁘다, 또는 제일 맛이 있다. 두번째로 맛이 있다, 세번째로 맛이 있다, 맛이 없다 등과 같이 等級을 매길 수 있는 경우는 많이 있다. 다음 表는 두 種類의 食品에 대하여 8名의 감정자가 각각 두 음식을 試食한 다음에 맛이 있는 順序에 따라서 1, 2로 順位를 정한 것이다.

여기에서 두 種類의 食品이라는 것은 쇠고기를 가정용 冷藏庫에 0°F로 저장하였다가 만든 작은 쇠고기 파이이고, 또 하나는 0°~15°F의 変温中에 저장하였다가 만든 파이이다. 맛이라고 한 것은 주로 맛 향기에 대하여 감정한 것이다.

쇠고기 파이에 대한 香気의 順位

감 정 자	0 ℉	変温(0° ~ 15 ℉)
A	1	2
B	1	2
C	2	1
D	1	2
E	1	2
F	1	2
G	1	2
H	1	2

이때 検定을 위한 帰無仮説로는 두 가지를 생각할 수 있다. 하나는 0° ~ 15 ℉의 変温中에 쇠고기를 저장하였다가 만든 파이가 0 ℉의 定温中에 저장하였다가 만든 파이와 조금도 다를 것이 없다는 것이고, 또 하나의 생각할 수 있는 帰無仮説은 파이의 향기에는 差異가 있는 것이 確実하지만 감정자의 半인 4名은 0 ℉의 定温에 저장하였다가 만든 파이의 향기를 좋아하고, 나머지 半인 4名은 0° ~ 15 ℉의 変温中에 저장하였다가 만든 파이의 향기를 좋아 한다는 仮説이다. 그러나 이들 두가지중 어떤 것을 帰無仮説로 設定하든간에 0 ℉의 定温中에 저장하였다가 만든 파이가 1번으로 될 確率은 1/2이다. 따라서 이와 같은 仮説下에서는 8名중 4名이 0 ℉에서 저장하였다가 만든 파이에 1번을 매기게

될 것을 기대할 수가 있다. 그러나 실제에 있어서는 이것이 7名과 1名으로 觀測된 것이다. 이와 같은 仮説을 検定하는데 있어서 가장 적절한 方法은 χ^2-検定法이라는 것을 理解하기 어렵지 않을 것이다. 위의 例에서 χ^2값은 다음과 같이 計算된다.

$$\chi^2 = (7-4)^2/4 + (1-4)^2/4 = 4.5$$

이와같은 帰無仮設을 検定하는 경우에는 다음과 같은 便利한 公式을 利用할 수가 있다.

$$\chi^2 = (a-b)^2/n = (7-1)^2/8 = 4.5$$

단, a, b는 두 가지 種類에 대한 個数이고 $n = a + b$이다.

그런데 標本의 크기 n이 작을 때는 위의 公式을 修正한 다음과 같은 公式을 使用하여야 한다.

$$\chi^2 = (|a-b|-1)^2/n = (|7-1|-1)^2/8$$

$$= (6-1)^2/8 = 3.12 \qquad P = 0.078$$

따라서 위의 例의 경우 5% 有意水準下에서 仮説은 棄却되지 않는다.

이와 같은 χ^2-検定은 正規分布를 仮定하였을 때의 対応이 있는 짝지은 実験에 의하여 母平均値差를 比較하는 検定과 対等한 것이라고 할 수 있다. 그리고 위의 例에서 1, 2와 같은 順位番号 대신에 +, −등의 符号를 使用하는 경우가 많으므로 위에서의 検定을 符号検定(the sign test)이라고 한다.

符号検定의 경우 아래表를 利用하면 χ^2값을 計算할 必要없이 直接 検定할 수가 있다.

符号検定에 있어서 各有意水準에 대한 같은 符号数(괄호안의 数字는 실제 確率) <両側 検定用>

표본의 크기	有意水準		
	1 %	5 %	10 %
5			0 (.062)
6		0 (.031)	0 (.031)
7		0 (.061)	0 (.016)
8	0 (.008)	0 (.008)	1 (.070)
9	0 (.004)	1 (.039)	1 (.039)
10	0 (.002)	1 (.021)	1 (.021)
11	0 (.001)	1 (.012)	2 (.065)
12	1 (.006)	2 (.039)	2 (.039)
13	1 (.003)	2 (.022)	3 (.092)
14	1 (.002)	2 (.013)	3 (.057)
15	2 (.007)	3 (.035)	3 (.035)
16	2 (.004)	3 (.021)	4 (.077)
17	2 (.002)	4 (.049)	4 (.049)
18	3 (.008)	4 (.031)	5 (.096)
19	3 (.004)	4 (.019)	5 (.063)
20	3 (.003)	5 (.041)	5 (.041)

※ 괄호안의 숫자는 正確한 有意水準값을 나타낸다.

(3) 測定値間의 差에 順位를 매겨 検定하는 方法(The signed rank test)

対応이 있는 짝지은 標本抽出에 의한 두 母平均値差를 検定하는 t -検定과 対等한 検定方法을 符号検定法이라고 하였으나, 또 한가지 方法으로 Wilcoxon의 符号가 붙은 順位検定法 (the signed rank test)이라는 것이 있다. 이 두 가지 方法은 한 資料에 대하여 다같이 適用할 수 있는 것이지만 여기에서 説明하는 方法이 資料에 따라서는 検定力(検定의 感度)이 높은 것으로 되어 있다.

이 方法은 처음에 짝지어진 각각의 測定値差를 구한다. 그러면 이 差에는 + 또는 -符号가 따를 것이지만 이 符号를 無視하고 差의 絶対値가 작은 것부터 順番으로 1, 2, 3, ……과 같이 番号를 붙인다. 그 다음에 이 番号에 測定値差에 붙어 있는 符号+ 또는 -를 붙인다. 그리고 符号別 番号合計値를 구하고 이 값을 利用하여 検定을 하게 된다.

例. 한 児童心理学者가 어린이가 保育院에 다니는 것이 그들의 社会的 知覚力에 영향을 끼치는가를 알아보기 위하여 다음과 같은 実験을 하였다. 8쌍의 쌍동이를 利用하여 각쌍의 쌍동이 중에서 한 아이를 保育院에 다니게 하고 한 아이는 집에서 놀게 하였다. 그리고 얼마후에 어떤 検査를 거쳐서 다음과 같은 社会的 知覚力에 관한 点数表를 얻었다.

社会的 知覚力에 대한 検査点数

쌍둥이 번 호	保育院에 다닌 아이들의 社会的 知覚力	집에서 놀던 아이들의 社会的 知覚力	差 d	d의 順位	−번호	+번호
1	82	63	19	7		7
2	69	42	27	8		8
3	73	74	-1	-1	1	
4	43	37	6	4		4
5	58	51	7	5		5
6	56	43	13	6		6
7	76	80	-4	-3	3	
8	65	62	3	2		2
				計	4	32

이 表에서 差d에 대하여 +, −符号에 関係없이 우선 差의 크기의 絶対値에 의하여 順位番号를 붙이고 그 다음에 差의 符号를 順位番号에 붙인 것이다. 이 例에서 다음과 같은 帰無仮説을 設定한다.

帰無仮説 H_0 : 社会的 知覚力에 있어서 保育院에 다닌 아이와 집에서 놀던 아이 사이에는 差異가 없다.

이것을 Wilcoxon 検定에서 使用하는 用語로 表現하면 다음과 같이 된다.

H_0 : 負数 d에 対応하는 順位番号의 合計 = 正数 d에 対応하는

順位番号의 合計

H_1 : 負数 d에 対応하는 順位番号의 合計≠正数 d에 対応하는 順位番号의 合計

위의 表에 의하면 一符号의 合計値가 4, +符号의 合計値가 32로 되어 있다. 그런데 이때의 合計値가 작은 것, 즉 4가 検定統計量 T로 된다. 아래 表에 의하면 $n=8$일 때 5% 有意水準下에서의 값은 2로 되어 있다. 그런데 t-検定의 경우

順位의 合에 의한 検定表

合이 이들값과 같거나 작을때 기각한다<両側検定用>

標本의 크기	5% 水準	1% 水準
7	2 (.047)	0 (.016)
8	2 (.024)	0 (.008)
9	6 (.054)	2 (.009)
10	8 (.049)	3 (.010)
11	11 (.053)	5 (.009)
12	14 (.054)	7 (.009)
13	17 (.050)	10 (.010)
14	21 (.054)	13 (.011)
15	25 (.054)	16 (.010)
16	29 (.053)	19 (.009)

※ 괄호안의 숫자는 正確한 有意水準값을 나타낸다.

와는 달리 여기에서는 計算한 検定統計量T의 값이 위표의 값과 같거나 이보다 작을 때 仮定을 기각하게 된다. 따라서 이 경우는 5% 有意水準下에서 仮説은 기각되지 않는다.

이와 같은 問題에서 각 짝의 測定値差의 絶対値 중에서 같은 것이 있을 수 있음은 勿論이다. 例를 들면 위의 例에서 첫째 짝의 差의 값이 13이라면 順位 6의 差와 같게 된다. 이와 같은 경우 6, 7번이라는 番号대신에 6.5, 6.5와 같은 번호값을 使用하게 된다. 또 差의 값이 0일 때는 이 짝을 除外하고 番号를 붙인다. 앞의 例에서 두번째 짝의 差가 0이라고 하면 이것을 除外하고 7번까지만 番号를 붙인다. 즉 이때는 $n=8$이 아니라 $n=7$로 되는 것이다. 또 標本의 크기 n이 16以上일 때는 検定統計量 T가 近似的으로 正規分布를 한다는 것을 利用한다. 即

$$Z = (|\mu - T| - \frac{1}{2}) / \sigma$$

로 놓으면 Z는 近似的으로 標準正規分布에 따른다는 것이다. 단, 위의 式에서 $\mu = n(n+1)/4$, $\sigma = \sqrt{(2n+1)\mu/\sigma}$ 이다.

3.4 Bayesian 意思決定論

(1) 緒 論

意思決定論者 그 중에서도 특히 Bayesian 意思決定論을 다루는 사람들은 前小節에서 説明한 종래의 統計的 方法을 古典的

인 方法이라고 하여 意思決定論的 方法과 구별한다. 그런데 이 두가지 方法은 使用하는 用語에 있어서는 다른 点이 많은 反面에, 問題를 다루는데 있어서나 分析方法에 있어서나 유사한 点이 많은 것이 사실이므로 古典的 方法과 比較하면서 說明하기로 하자.

위에서 말한 差異点과 類似点을 다음과 같은 仮說檢定問題에서 찾아보는 것이 편리할 것이다. 生産品의 不良率 P에 관하여 帰無仮說 $H_0 : P \leq P_0$을, 対立仮說 $H_1 : P > P_0$에 대하여 検定하는 경우를 생각해 보기로 한다. 古典的인 統計的 方法에서는 우선 生産品集団에서 標本을 취하여 그 중에서의 不良品数를 조사하고, 이것을 基礎로 하여 위에서 세운 帰無仮說 H_0를 受落할 것인가, 기각할 것인가 하는 意思決定規則을 만들어 낸다. 이와 같은 경우 第1種의 過誤의 確率의 最大限度로서 α를 定하고, 또 対立仮說 H_1내에서의 P값에 대한 第2種의 過誤를 범하게 되는 確率을 따지게 된다. 이와 같은 것은 다음表3.4.1에 整理해 놓은 바와 같이 우리가 세운 仮說은 自然의 状態에 대하여 記述한 것이고, H_0를 受落 또는 棄却하는 것은 우리의 行為로서 A_1, 또는 A_2와 같이 表記할 수가 있을 것이다. 이와 같은 対応은 古典

表3.4.1 古典的인 統計的 検定과 意思決定論에서의 行為

仮設 : 自然状態	行為 A_1 : H_0를 受落	行為 A_2 : H_0를 기각
H_0가 真이다	過誤 없음	第1種의 過誤
H_1이 真이다	第2種의 過誤	過誤 없음

的인 統計的 方法과 意思決定論과의 類似点을 表示한다. 더구나 이 問題에 대하여 表3.4.2와 같은 損失表를 作成해 보면 이러한 類似点은 더욱 명료해 진다. 이 表에서 θ_1下에서의 行為 A_2의 機会損失을 $L(A_2/\theta_1)$로 表示하고, θ_2下에서의 行為 A_1의 機会損失을 $L(A_1/\theta_2)$로 나타낸 것이다.

表3.4.2 損 失 表

自然의 状態	行 為	
	A_1	A_2
θ_1	0	$L(A_2/\theta_1)$
θ_2	$L(A_1/\theta_2)$	0

이제 두가지 方法間의 差異点을 생각해 보기로 한다. 仮説検定에 있어서는 有意水準 α의 選択이 意思決定方法을 決定짓게 되어 있어, 対立仮説중에서의 選択의 여지가 없게끔 만들어진 것이 特色이다. 즉 이것을 記号로 表示하면 $\alpha = P(A_2/H_0$가 真)인 것에서 알 수가 있다. 따라서 仮説検定에 있어서의 行為간에서의 選択에 대한 주된 判定基準은 이와 같은 型의 誤差가 일어나는 相対頻度이다. 그러면 有意水準 α는 어떻게 選定되는가? 많은 応用例에서 편의상으로 α로서 0.05 또는 0.01를 択하게 되는데, 이러한 경우 問題로 하고 있는 対象에 따라서 어떤 特別한 고려없이 一律的으로 α값이 選定되는 경향이 있다. 그러나 이와 같은 傾向은 仮説検定方法 그 자체의 欠点은 아니다. 이것은 이 方法을 利用하는 사람들의 잘못에 지나지 않으며, 우리는 仮説検定

에 있어서 그 対象에 대한 専門的인 知識을 活用하여 帰無仮説과 対立仮説의 設定에 세심하여야 하며, α의 選定에 있어서 여러가지로 주의 깊은 考察을 하여야 될 것이다. 그런데 바로 이러한 点이 Bayesian 意思決定論의 추종자들에 의해서 批判을 받는 点이다. 그들은 仮説検定이 根拠없는 主観과 判定에 의하여 実施되고 있다고 비난한다. Bayesian統計学者들은 그들의 意思決定節次가 고전적인 仮説検定方法을 論理的으로 発展시킨 것이라고 主張하고 있다. 그들은 自然의 状態(위의 경우는 帰無仮説과 対立仮説)에 대하여 事前確率分布를 부여하고, 그들의 判定에 따르는 損失을 고려하여 意思決定을 한다는 것이다.

(2) 仮説検定과 두 行為問題와의 比較

意思決定을 하는 데 있어서 古典的인 方法과 Bayesian方法과의 比較를 두 行為問題로 説明해 보기로 한다. 한 輸入業者가 어떤 物品을 輸入하는데 있어서, 対象者는 이름과 住所가 収録되어 있는 10,000名이다. 이와 같은 경우 輸入業者의 가능한 두 行為는

A_1 : 物品을 輸入하여 市販한다.

A_2 : 輸入하지 않는다.

여기에서 P를 輸入品을 注文購入하게 될 比率(10,000名중에서의 比率)을 表示하는 것으로 한다. 따라서 이때 注文購入者의 比率 P가 0.05보다 작을 때는 輸入하지 않기로 意思決定을 하게

된다. 이때 損失表는 다음과 같다고 하자.

表 3.4.3 損 失 表 (単位 : ₩ 1,000)

事 件	行 為	
	A_1	A_2
$P \leq 0.05$	₩ 20,000(0.05-P)	0
$P > 0.05$	0	₩ 20,000(P-0.05)

注文購入할 사람 数의 比率을 알기 위하여 명부에서 100名을 임의로 抽出하여 輸入한 物品의 広告를 보내어, 그 反応에 따라서 標本比率 $\hat{P}$를 計算하고, 이것으로 P값을 推測하는 方法을 생각한다.

이때 古典的인 統計的 方法에서의 検定法은 標本比率 $\hat{P}$를 利用하여, $H_0 : \hat{P} \leq 0.05$를 $H_1 : P > 0.05$에 대하여 다음과 같이 한다. 즉 H_0下에서 標本比率 $\hat{P}$의 標準誤差 $\sigma_{\hat{p}}$를

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.05)\ (0.95)}{100}} = 0.0218$$

과 같이 計算하고, $Z = (\hat{p} - 0.05) / \sigma_{\hat{p}}$가 標準正規分布에 따른다는 것을 利用한다. $\alpha = 0.05$로 하였을 때, $Z = 1.65$이므로

$$\hat{P} > 0.05 + (1.65)\ (0.0218) = 0.086$$

일 때는 有意水準 $\alpha = 0.05$下에서 H_0를 棄却하게 되는 것이다.

이 問題에서 比率 $\hat{P}$대신에 注文購入 하겠다는 사람数 x를 가지고 表現하면 100(0.086) =9이므로

① $x > 9$일 때는 H_0를 棄却한다. (輸入販売한다)

② $x \leq 9$ 일 때는 H_0를 受落한다 (輸入販売하지 않는다)

이와 같은 경우 $\alpha = 0.05$ 라고 하는 것은 確率記号로 表示하여

$$P(x > 9 \mid p \leq 0.05) = 0.05$$

라고 하는 것을 뜻한다. 이때의 第2種의 過誤는 $P(x \leq 9 \mid p > 0.05)$ 와 같이 表現된다.

仮説検定의 方法에 따라 意思決定을 하는 데 있어서 좋은 意思決定規則 (위의 例에서는 "$x > 9$ 일 때는 H_0를 棄却한다"는 것이 意思決定規則이다) 은 第1種의 過誤가 작고, 또 第2種의 過誤도 작은 것이라야 된다는 것은 理解하기는 어렵지 않을 것이다. 그러면 여기에서 第1種의 過誤와 第2種의 過誤가 서로 어떤 관계에 있는가 하는 것을 알아보기 위하여 다음과 같은 세가지 意思決定規則을 비교하여 보기로 한다.

① $x > 2$ 일 때 H_0를 棄却한다.

② $x > 7$ 일 때 H_0를 棄却한다.

③ $x > 9$ 일 때 H_0를 棄却한다.

이와 같은 세가지 意思決定規則 (検定節次) 에 대하여 各各의 第1種의 過誤와 第2種의 過誤를 図表로 나타낸 것이 그림 3.4.1 이다.

이 図表에서 알수 있는 것은 第一種의 過誤가 큰 意思決定規則은 第2種의 過誤가 작고, 반대로 第1種의 過誤가 작은 것에 있어서는 第2種의 過誤가 크게 된다는 事実이다. 이러한 現象으

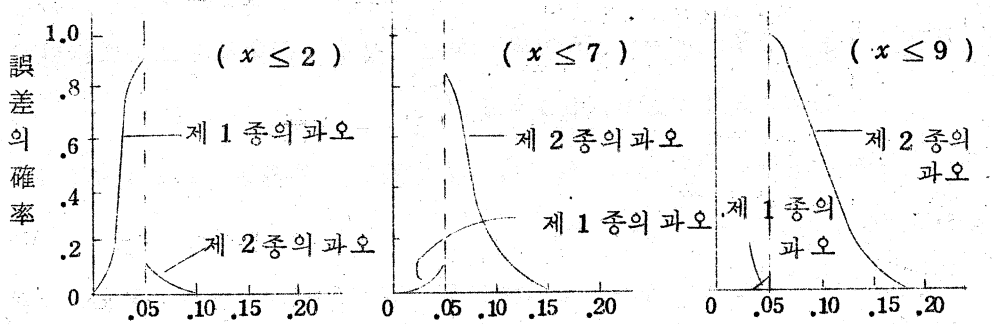


그림 3.4.1 誤差特性曲線

로 보아서 좋은 意思決定規則을 選定한다는 것이 얼마나 어려운 것인가 하는 것을 짐작할 수가 있을 것이다.

이제 이와 같은 問題가 Bayesian意思決定節次에서는 어떻게 解決되는가를 알아 보기로 한다. 우선 図3.4.1에서 $P=0.02$, 0.04, 0.06, 0.08의 경우에 대하여, 各各에 대한 α값을 表示한 表(3.4.4)를 作成하여 두는 것이 必要할 것이다.

表3.4.4 $H_0 : P \leq .05$에 対한 α값

P	意思決定規則		
	$x > 2$	$x > 7$	$x > 9$
0.02	0.5000	0.0002	0.0000
0.04	0.3461	0.0571	0.0054
0.06	0.9545	0.3372	0.1020
0.08	0.9864	0.6443	0.8557

이 表3.4.4와 損失表3.4.2를 利用하여 $x > 2$와 같은 意思決定規則을 使用할 경우, 自然의 狀態가 $P = 0.02$일 때, 期待機会損失額数를 計算하면 다음과 같이 된다.

表3.4.4에 의하면 P (H_0를 棄却 | $P = 0.02$) $= 0.5$, P (H_0를 受落 | $P = 0.02$) $= 0.5$이고 $P = 0.02$일 때의 機会損失은 損失表 3.4.2에서 行為 A_1을 취하였을 때 (H_0를 棄却하였을 때) ₩ 20,000,000 (0.05 − 0.02) = ₩ 600,000이고 行為 A_2를 취하였을 때 (H_0를 受落하였을 때) ₩0이다. 따라서 이때의 $P =$ 0.02라는 條件下에서의 期待機会損失値는 다음과 같이 計算된다.

行 為	P(Ai)	機 会 損 失	期待機会損失
A_1	0.5	₩ 600,000	₩ 300,000
A_2	0.5	0	0
			300,000

$P = 0.04$일 때의 行為 A_1에 대한 機会損失値는 ₩ 20,000,000 (0.05 − 0.04) = ₩ 200,000이므로, 表3.4.4를 利用하여, $P =$ 0.04라고 하는 條件下에서의 期待機会損失値는 다음과 같다.

行 為	P(Ai)	機 会 損 失	期待機会損失
A_1	0.8461	₩ 200,000	₩ 169,220
A_2	0.1539	0	0
			₩ 169,220

그러나 $P=0.06$일 때는 損失表 3.4.3에서 알 수 있는 바와 같이 行為 A_1에 따르는 機会損失은 ₩0이고, 行為 A_2에 따르는 機会損失은 ₩20,000,000(0.06 − 0.05) = ₩200,000이므로 表 3.4.4를 利用하여 이제 條件附 期待機会損失은 다음과 같이 計算된다.

行 為	P (Ai)	機会損失	期待機会損失
A_1	0.9545	0	0
A_2	0,0455	₩ 200,000	₩ 9,100
			₩ 9,100

마찬가지로 $P=0.08$에 대한 條件附 期待機会損失値를 計算하면 ₩ 8,160으로 된다. 以上의 것을 整理하여 表 3.4.5를 얻는다.

表 3.4.5 H_0 : $P = 0.02$일 때의 意思決定規則 $x > 2$에 대한 期待機会損失値

自然의 状態 P	條件附 期待機会損失値
0.02	₩ 300,000
0.04	169,220
0.05	0
0.06	9,100
0.08	8,160

이와 같은 것을 다른 意思決定規則 $x > 7$, $x > 9$에 대하여, 위 表와 같이 整理하면 다음과 같이 된다.

表3.4.6 세가지 意思決定規則에 대한 條件附 期待機会損失値

自然의 狀態 P	意思決定規則		
	$x > 2$	$x > 7$	$x > 9$
0.02	₩ 300,000	₩ 120	₩ 0
0.04	169,220	11,420	1,080
0.06	9,100	132,560	179,600
0.08	8,160	213,420	386,580

條件附 期待損失値는 $P \leq 0.05$ 일 때는 第1種의 過誤로 인한 損失을, $P > 0.05$ 일 때는 第2種의 過誤로 인한 損失을 金額値로 表示한 것이라고 할 수가 있다. 그러므로 以上에서의 考察은 古典的인 仮説検定 方法에 의한 意思決定節次를 "損失"이라고 하는 概念으로 表現하였을 뿐이므로, 本質的으로 Bayesian 方法의 差異点을 나타내는 것이라고는 할 수 없다. 여기에서 Bayesian 統計学者들은 2.4節에서 설명된 바와 같이 自然의 狀態 P의 여러가지 값에 대하여 그 事前確率 $P_0(P)$를 부여한다.

P	$P_0(P)$
0.02	0.10
0.04	0.40
0.06	0.45
0.08	0.05

即, 이와 같은 事前確率을 利用하여 意思決定規則 $x > 2$에 대한 無條件 期待機会損失値를 計算하면 다음과 같이 된다.

表3.4.7 $x > 2$에 대한 無條件期待機会損失値

$P_0(P)$	條件附期待値	期待値
0.10	₩ 300,000	₩ 30,000
0.40	169,220	67,688
0.45	9,100	4,095
0.05	8,160	408
		₩ 102,191

마찬가지로 다른 意思決定規則 $x > 7$, $x > 9$에 대하여 無條件 期待機会損失을 計算하여 위의 結果와 같이 정리하면 다음과 같이 된다.

表3.4.8 $x > 2$, $x > 7$, $x > 9$에 대한 無條件 期待機会損失値

$P_0(P)$	$x > 2$	$x > 7$	$x > 9$
0.10	₩ 30,000	₩ 10	₩ 0
0.40	67,690	4,570	430
0.45	4,100	59,650	80,820
0.05	410	10,670	19,330
	₩ 102,200	₩ 74,900	₩ 100,580

이와 같은 方法에 의하면 세가지 意思決定規則에서 期待損失이 작은 것 即 $x > 7$을 選択할 수 있게 된다. 이와 같이 하는 것이 Bayesian 統計的 方法인 것이다.

第 4 章　相關과　回歸

4.1 相関分析

우리들 日常生活에서 나타나는 経済的이고 自然的인 모든 現象은 統計的 観点에서 본다면 하나의 時系列로 表現될 수 있으며, 그 時系列은 다른 現象의 時系列과의 사이에 하나의 聯関性 (relationship)이 있는 것으로 생각할 수 있다.

即 工場労動者의 賃金과 年齢사이의 관계나 洋品店에 있어서의 雨傘과 日気사이에는 相互依存的인 立場에 놓이게 되는 것을 알 수 있다. 따라서 所得이 낮은 사람은 밍크 코우트를 購入할 수없는 反面, 財産이 많은 사람은 所得以上으로 衣服을 購入할 수도 있다.

이와 같은 相互依存関係는 칼·피어슨(K.Pearson)의 生物測定에서 発展시킨 方法論이며, 이러한 方法에 의하여 얻어지는 情報는 経済豫測과 政策樹立에 매우 価値있는 것으로 評価되기도 한다.

따라서 서로 対応하는 時系列의 統計量을

x_i ($i = 1,2$ ---------- , m)

y_j ($j = 1,2$ ---------- , n)

로 表示하고 x의 값이 決定될 때 y의 값이 決定된다면 이 두 統計量사이의 関係는

$y = f(x)$

라는 函数的 関係로 把握되며 이와 같은 相互聯関性을 相関関係 (correlation)로 認識하는 것이다.

(A) 相関의 数学的 説明

이와 같은 函数的 表示는

①x가 増加함에 따라 y도 一増加하는 경우

②x가 増加함에 따라 y는 減少하는 경우와

③x의 変化에 比例하여 y가 増加하던지 減少하는 경우

④두 変量이 같은 方向이나 反対方向으로 変化하는 傾向이 전혀 나타나지 않은 경우를 들 수 있는 바, 여기에 대하여는 数学的 用語를 使用하여 ⓘ正相関 ⓘ負相関 ⓘ完全相関 ⓘ無相関이라고 한다.

(B) 相関表의 作成

앞에서 言及한 바와 같이 変量 x의 変化에 따라 変量 y가 変化할 때 이를 説明하는 度数分布表를 相関表(correlation table)라고 하는바 다음과 같은 例가 이를 理解할 수 있게 할 것이다.

即 다음의 表4.1.1은 H洋服店의 3年間에 걸친 期別 資料이다. 이 資料를 가로에는 顧客에 대한 度数分布表로서, 세로에는 売上高에 대한 度数分布表를 作成하여 하나의 箱子모양의 度数記入欄에 度数(該当期数)를 記入하면 一般的으로 表4.1.2와 같이 左上에서 右下로 연결되는 対角線에 모이게 되는 特質을 가지고

表4.1.1　　두 変量의　関係

期　　別	顧　　客	売　上　高
1969 / 1	750	268.5 (萬원)
2	700	240.0
3	850	296.0
4	980	312.2
'70/ 1	1,000	328.5
2	900	285.5
3	660	225.0
4	790	274.5
'71/ 1	1,040	309.3
2	920	307.6
3	730	256.7
4	900	315.1

있는 바 이는 두 変量사이에 依存性이 높으면 높을수록 完全対角線을 이루게 된다는 것을 説明한다.

一般的으로 顧客을 x_i, 売上高를 y_i로 文字化 했을 때, 変量 x_i를 基準으로 했을 때의 度数의 合計 g_i와 変量 y_i를 基準으로 했을 때의 度数의 合計 n_j에 의한 다음과 같은 定式을 얻을 수 있다.

$$g_i = f_{i1} + f_{i2} + \cdots\cdots + f_{in} = \sum_{j=1}^{n} f_{ij}$$

$$n_i = f_{1j} + f_{2j} + \cdots\cdots + f_{nj} = \sum_{j=1}^{m} f_{ij}$$

表 4.1.2　　相 関 表 (顧客과 売上高)

顧客 x / 売上高 y		600～700	700～800	800～900	900～1000	1000～1100	計
		650	750	850	950	1,050	
210～230	220	1					1
230～250	240	1					1
250～270	260		2				2
270～290	280		1	1			2
290～310	300			1	1	1	3
310～330	320			1	2		3
計		2	3	3	3	1	12

또한 이와 같은 相関表에 의하여 그래프를 作成할 때에는 直角座標面에 1개 내지 몇 개의 点에 의하여 表示되는 点相関図 (Scatter diagram)를 作成할 수 있는 것도 아울러 일러둔다.

(1) 相関関係의 測定

우리는 相関図를 作成함으로써 두 変量間의 相互依存関係가 어느 程度인가를 대개 把握할 수 있는 바, 이와 같은 性質의 尺度를 相関係数 (Coefficient of Correlation)라 한다.

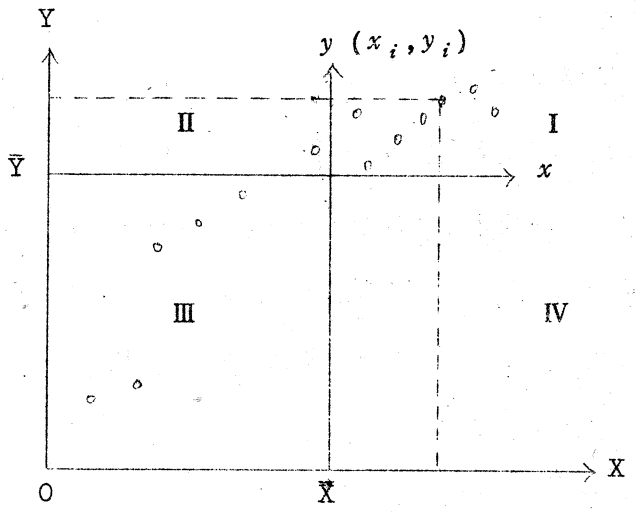
(A) 相関係数의 測定式 導出

따라서 이러한 相互依存性을 하나의 係数로 表示하기 위한 定式을 마련하기 위하여 다음과 같은 説明을 必要로 하게 된다.

即 다음의 도표 4.1.1에서와 같이 各 点은 새로운 座標軸

x , y에 의하여 4상한으로 分割되어지므로 xi , yi에 대한 各 상한의 総和 S를 생각해 보면,

図4.1.1 軸의 変換



$$S = \Sigma_1 xiyi + \Sigma_2 xiyi + \Sigma_3 xiyi + \Sigma_4 xiyi \quad \cdots\cdots (1)$$

(I) (II) (III) (IV)

가 되며 S의 符号는 各 상한의 散布度(dispersion)를 測定하므로서 쉽게 判断할 수 있다. 이와 같이 xi , yi의 総和 S를 測定하면 相関의 程度를 推測할 수 있는 데 相関係数란 이 総和의 平均을 標準化함으로써 얻어질 수 있는 것이다. 다시 말하면 이렇게 하여 얻어진 共分散(Co-Variance)

$$\frac{S}{n} = \frac{1}{n}\Sigma xiyi = \frac{1}{n}\Sigma(X_a - \bar{X})(Y_j - \bar{Y}) \quad \cdots\cdots (2)$$

를 標準偏差 σx , σy로 나누어 줌으로서 標準化가 이루어 진다.

따라서 우리가 求하고자 하는 相関係数 r은

$$r = \frac{\Sigma(Xi-\bar{X})(Y_j-\bar{Y})}{\sqrt{\Sigma(Xi-\bar{X})^2\,\Sigma(Y_j-\bar{Y})^2}} \quad \text{---------- (3)}$$

가 되며 이를 偏差平方合을 求할 必要가 없는 다음과 같은 方法으로도 使用할 수 있다.

$$r = \frac{\Sigma Xi\,Y_j - \frac{1}{n}\Sigma Xi\,Y_j}{\sqrt{\Sigma Xi^2 - \frac{1}{n}(\Sigma Xi)^2}\ \sqrt{\Sigma Y_j^2 - \frac{1}{n}(\Sigma Y_j)^2}} = \frac{\sigma xy}{\sigma x\,\sigma y}$$

여기에서 알아둘 것은 (1) (2) (3) 式에 있어서 使用되는 添数記号 i와 j는 두 変量의 個数가 같은 경우 i로 統一 使用할 수 있으며 이 경우의 (3)式은 다음과 같이 表記되어 진다.

$$r = \frac{\Sigma(Xi-\bar{X})(Yi-\bar{Y})}{\sqrt{\Sigma(Xi-\bar{X})^2\,\Sigma(Yi-\bar{Y})^2}}$$

(B) 相関係数의 決定

이와 같은 相関係数의 範囲는,

$$+1 \geq r \geq -1$$

이 된다. 따라서 r의 絶対値가 갖는값에 따라 相互依存的인 聯関의 程度를 알 수 있는 바, 相関程度의 強弱은 変量 間의 因果関係가 分明히 成立한다고 認定되는 経験的이고 実証的인 検証이 前提로 되었을 때에 局限하여 使用되어야 한다는 点을 附記한다.

一般的으로

① $|r| < 0.3$이면 依存関係가 거의 없다.

② $0.3 < |r| \leq 0.5$ 이면 依存関係가 약간 있다.

③ $0.5 < |r| \leq 0.7$ 이면 依存関係가 確実히 存在한다.

④ $0.9 < |r|$ 이면 強한 依存関係가 있다.

라고 説明하고 있는 데 ③의 경우 相関係数가 0.7 程度이면 依存関係가 分明하다고 하는 説明은 決定係数($R = r^2$) 理論에서 $r^2 \geq 0.5$ 의 性格이 이를 説明하고 있다.

(C) 相関関係의 測定

時系列 相関을 測定하는 데 있어서는 経済的 意味가 갖는 変量의 相互依存関係에 따라 同時相関과 時差相関의 두가지 測定方法이 있다.

지금 두개의 時系列資料에서 나타난 変量을 Xt , Yt라고 하면,

$r(Xt, Yt)$ X와 Y의 同時相関

$r(Xt, Yt-k)$ X와 K時点 前의 Y와의 時差相関

$r(Xt-k, Yt)$ Y와 K時点 前의 X와의 時差相関

으로 表現할 수 있으며 説明의 便宜를 위하여 同時相関의 測定을 하여 보면 表 4.1.3과 같은 計算方法에 의하여

$$\sigma x = \sqrt{14,288} = 119.4$$

$$\sigma y = \sqrt{964.07} = 31.0$$

$$\sigma xy = 3,349.4$$

이므로 相関係数 r는

$$r = \frac{3,349.4}{119.4 \times 31.0} = 0.90$$

表4.1.3 . 相関係数의 計算

	x	y	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - X) \times (y_i - Y)$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$
1	750	268.5	- 101	- 16.4	1,656.4	10,201	268.96
2	700	240.0	- 151	- 44.9	6,779.9	22,761	2,016.01
3	850	296.0	- 1	11.1	- 11.1	1	123.21
4	980	312.2	129	27.3	3,521.7	16,641	745.29
5	1,000	328.5	149	43.6	6,496.4	22,201	1,900.96
6	900	285.5	49	0.6	29.4	2,401	0.36
7	660	225.0	- 191	- 59.6	11,340.9	36,481	3,588.01
8	790	274.5	61	- 10.4	-634.4	3,721	108.16
9	1,040	309.3	189	24.4	4,611.6	35,721	595.36
10	920	307.6	69	22.7	1,566.3	4,761	515.29
11	730	256.7	- 119	- 28.2	3,355.8	14,161	795.24
12	900	315.1	49	30.2	1,479.8	2,401	912.04
計	10,220	3,418.9			40,192.7	171,452	11,568.89
平 均	851	284.9			3,349.4	14,288	964.07

이 되여 두 変量사이에 있어서의 相互依存関係가 強하게 나타나고 있음을 알 수 있다.

(2) 相関係数의 検定

앞에서도 言及한 바와 같이 相関係数의 使用은 実証的이고 分析的인 検証에 依하여서만 可能하므로 여기에서 相関係数의 検定을 必要로 한다.

지금 붉은 주사위와 푸른 주사위를 다섯번 던져 다음과 같이 나왔다고 하자.

붉은주사위	푸른주사위
1	2
4	6
5	4
6	4
2	2

이와 같은 두개의 統計量에 의한 相関係数 r의 計算에서 0.67을 얻었다고 하면 이 두 統計量 사이에 相互依存性이 存在한다고 할 수 있을 것인가의 問題가 提起되며 이 相関係数의 信頼度를 検定해야 한다.

一般的으로 母集団 相関係数 ρ와 n개의 標本에서 얻은 標本 相関係数 r사이에 있어서

$$t_r = \frac{1}{2} \log \frac{1 + r}{1 - r}$$

의 t分布를 利用하게 되는 데 이는 正規分布 $N(\mu, \sigma^2)$와 近似하게 나타나는 바 이라고 놓으면 log가 常用対数로 바뀌어져야

$$\mu = \frac{1}{2} \log \frac{1+\rho}{1-\rho} \qquad \sigma^2 = \frac{1}{n-3}$$

하므로,

$$t_r = \frac{1}{2} \log_e \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+r}{1-r} \quad \text{------(5)}$$

$$\mu = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+\rho}{1-\rho}$$

로 使用되며 만일 母集団 相関係数 ρ가 ρ_o에 같다고 하면 仮説 $H_o : \rho = \rho_o$가 되어져 이에 의하여 仮説検定하게 된다.

그러나 t_r가 $N(\mu, 1/n-3)$으로 나타낼 수 있다는 前提 아래,

$$t = (t_r - \mu_o) \sqrt{n-3} \quad \text{----------(6)}$$

을 計算하는 한편 正規分布의 有意水準 α点의 t_α를 求하여 보면 다음과 같은 検定이 可能해 진다.

① $|t| > t_\alpha$ 이면 H_o를 棄却

② $|t| \leq t_\alpha$ 이면 H_o를 採択

예를 들어 H大学 校의 統計学科 学生88名에 대한 IQ와 数学 成績과의 相関係数는 0.69이고 1年前의 相関係数는 0.75였다면 이 두 現象에 대한 信頼度는 어떠한가에 대하여 検討해야 한다.

이 경우 $n = 88$ $r = 0.69$ $\rho_o = 0.75$ 이므로 仮説로서 $\rho = \rho_o$가 認定되어져야 하며 (5) 式에 의하여,

$$t_r = 1.1513 \log_{10} \frac{1 + 0.69}{1 - 0.69} = 0.843$$

$$\mu_o = 1.1513 \log_{10} \frac{1 + 0.75}{1 - 0.75} = 0.973$$

을 얻고 (6)式에 의하여

$$t = (0.843 - 0.973) \sqrt{88-3} = -1.199$$

를 얻게 되므로 有意水準 5%인 点은,

$$t_{0.05} = 1.96$$

$$| t | = 1.199 < 1.96 = t_{.05}$$

가 되여 H_o는 採択되어 두 相関係数는 같은 것으로 判断되어지는 것이다.

以上과 같은 相関関係는 経済的이고 経営的인 現象에 대한 統計的 分析에 많이 適用되는 方法으로 이러한 変量 間의 관계는 単純히 相関係数의 測定으로 그치는 것이 아니라 回帰分析이라는 方法과 結合하여 定量的인 相互依存関係를 説明해 나갈 때 企業의 成長이나 経済全般的인 与件의 変動, 그리고 生産品에 대한 市場性의 検討와 같은 社会科学 全般에 걸친 内容과 教育水準의 測定이나 生物遺伝의 分析的 方法에도 널리 利用되고 있다는 것을 알아야 하겠다.

4.2 回帰分析

変量 x 를 原因 또는 独立変数, y 를 結果 또는 従属変数라고 하면 2変量의 関係는 一般的으로 다음과 같이 表示할 수가 있다.

$$y = f(x)$$

x에 대한 y의 変化가 어느 정도 規則的으로 일어나는 경우에는 이 函数는 比較的 간단한 数学式으로 表示된다. 가장 간단한 것은 一次直線이다.

$$y = a + bx$$

統計의 問題는 実際의 観察에서 얻어진 2変量 x, y의 関係를 이러한 数学式으로 表現하는 것이다. 이러한 変量사이의 関係를 表示하는 方程式을 回帰方程式(regression equation)이라 한다. 이 方程式이 그리는 線을 回帰線(regression line)이라 하는 바 그것이 直線인 경우에는 直線回帰라 하고 曲線인 경우에는 非直線回帰라 한다.

(1) 直線回帰

回帰直線은 $y = a + bx$로 表示할 수가 있다. a는 $x = 0$일 때의 截片이고 b는 이 直線의 方向係数이다. 이 a,b는 直線의 位置와 方向을 決定하는 数値로서 直線의 方向係数 b를 x에 대한 y의 回帰係数라고 한다. 回帰分析에 있어서는 関係変量 x, y의 観察値에서 a,b등의 값을 計算하는 것이 問題가 된다.

a, b의 값이 決定되면 原因인 変量 x의 값이 주어지면 変量 y의 값은 回帰方程式에서 곧 決定할 수가 있다.

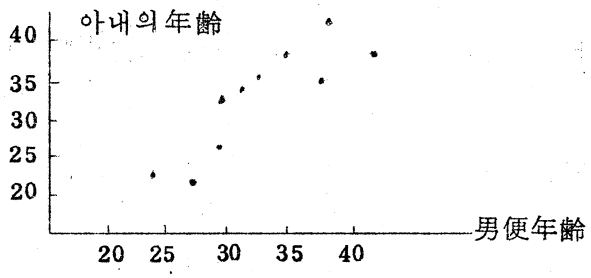
따라서 回帰方程式은 変量 y의 값을 推定하는 式이라고 생각할 수 있다.

身長과 體重, 남편의 結婚年齡과 아내의 結婚年齡사이에는 거의 직선적인 函数関係를 보인다. 예컨대 表4.2.1을 点散図로 表示한 4.2.1図에서 보는 바와 같이 남편의 年齡이 많아짐에 따라 아내의 年齡도 그에 상응하여 많아져 가고 있다. 이 点의 連績에 가장 適合한 直線을 긋는 것이 回帰分析의 問題이지만 観察値와 回帰直線과의 距離d의 自乘의 合計를 最小로 하는 直線은 2線밖에 없다. 여기에 2線이란 하나는 x에 대한 y의 回帰線이고 다른 하나는 y에 대한 x의 回帰線이다.

表4.2.1 夫婦의 結婚年齡

男便(x)	24	26	28	30	32	34	36	38	40	42
아내(y)	21	20	24	27	28	29	33	28	37	33

図4.2.1 夫婦의 結婚年齡



이 2개의 回帰線은 図4.2.2에서 보는 바와 같이 다른 것이 보통이고 다만 2系列 사이에 完全한 函数関係가 있을때에 한해서 일치 된다.

回帰直線의 方程式 $y' = a + bx$로 表示되므로 最小自乗法에 있어서 問題가 되는 것은 傾向線의 計算의 경우와 마찬가지로,

$$\Sigma d^2 = \Sigma (y - y')^2 = \Sigma (y - a - bx)^2$$

의 값을 最小가 되도록 方程式의 定数를 定하는 것이다. 定数 a, b는 다음과 같은 正規方程式에 (normal equation) 주어진 資料를 대입함으로써 計算된다.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

여기서 n은 変量 x, y의 観察数를 가리킨다.

定数 a, b는 또한 正規方程式에서 구해진,

$$a = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

에 의하여 決定할 수도 있다. 여기서 b가 回帰係数인 것이나 이것은 또한 $\bar{x}$, $\bar{y}$를 각각 x, y의 平均値라 할때 다음과 같이 表示할 수가 있다.

$$b = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

表4.2.1에 대한 回帰方程式을 구하면 다음과 같다.

男便의 年齡을 x, 아내의 年齡을 y라 하면 回帰方程式은 다음 式으로 주어진다.

$$y' = a + bx$$

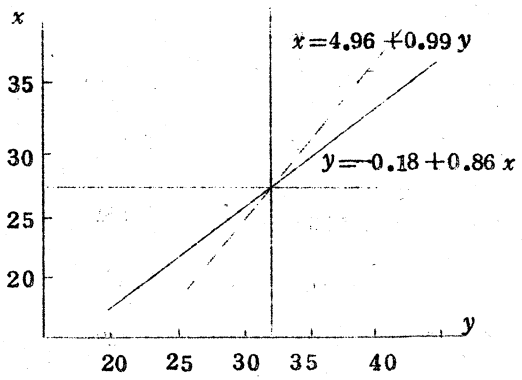


図 4.2.2 2개의 回帰直線

表4.2.1에 대한 正規方程式은 表4.2.3에 따라 다음과 같이 된다.

$$\begin{cases} 282 = 10a + 330b \\ 9,590 = 330a + 11,220b \end{cases}$$

이 方程式을 풀면

$$a = -0.18$$

$$b = 0.86$$

이 된다. 따라서 구하는 回帰方程式은,

$$y' = -0.18 + 0.86x \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1)$$

가 된다. 이것은 x에 대한 y의 回帰方程式이다. 다른 하나의 回帰方程式 즉 y에 대한 x니 그것도 위와 마찬가지로 구해진다.

$$\Sigma x = na' + b'\Sigma x$$

$$\Sigma xy = a'\Sigma x + b'\Sigma y^2$$

$$\begin{cases} 330 = 10a' + 282b' \\ 9,590 = 282a' + 8,238b' \end{cases}$$

이 方程式을 풀면

$$a' = 4.96 \quad , \quad b' = 0.99$$

그러므로 구하는 回帰方程式은

$$x' = 4.96 + 0.99\,y \quad \text{----------(2)}$$

가 된다.

表 4.2.3 直線回帰의 計算

x	y	xy	x^2	y^2
24	21	504	576	441
26	20	520	676	400
28	24	672	784	576
30	27	810	900	729
32	28	896	1,024	784
34	29	986	1,156	841
36	33	1,188	1,296	1,089
38	28	1,064	1,444	784
40	37	1,480	1,600	1,369
42	35	1,470	1,764	1,225
330	282	9,590	11,220	8,238

위의 回帰方程式에서 알 수 있듯이 이 양자는 서로 다르다. 예컨대 (1)식에 의하면 28歲의 男便에 대한 아내의 理論的 年齡은 23.9歲이지만 23.9歲의 아내에 대한 男便의 理論的 年齡은 (2)式에 의하여 28.6歲가 되며 그 사이에 0.6歲라는 差異가 생긴다. 이 差異는 x, y중의 어느 하나가 独立変量으로서 択해

지며 어느 하나가 從屬變數로서 択해지느냐에 의한다. 즉 独立変数로 간주된 系列은 올바른 것이라고 하고, 從屬變數로 看做로 系列에만 誤差가 생길 수 있다고 推定되어 있기 때문이다. 그러나 両 系列사이에 完全한 函数関係가 있을 경우에는 이 2回帰線이 일치함은 물론이다.

回帰直線 $y' = a + bx$ 는 그것과 変量 y 와의 差의 自乘合이 最小가 되도록 그은 直線이다. 最小自乘法에 의해서 適用된 回帰直線은 이동하는 平均値의 線이라고 생각할 수가 있다. 그리고 平均値의 경우와 마찬가지로 回帰線에 대한 変量의 偏差의 合計는 언제나 0이 된다.

1変量의 度数分布의 모양이 平均과 標準偏差에 의해서 代表되는 바와 같이 2変量사이의 関係를 表示하는 경우에도 平均的인 関係를 表示하는 回帰線과 아울러 回帰線에 대한 変量의 変動幅을 생각할 必要가 있다. 回帰線에 대한 変量 y 의 標準偏差는 다음 式으로 주어진다.

$$S_y = \sqrt{\frac{\Sigma (y - y')^2}{n}}$$

또는,

$$S_y = \sqrt{\frac{\Sigma (y - a - bx)^2}{n}}$$

1変量의 度数分布에서 分散度가 적을수록 平均은 変量全體를 잘 대표하는 것과 같이 回帰線에 대한 変量 y 의 分散이 적을수록

回帰線은 変量사이의 関係를 正確하게 대표한다. 따라서 이 경우에는 回帰方程式의 x에 의해서 y를 正確히 推定할 수가 있게 된다.

그러나 回帰線에 대한 変量y의 偏差가 標準偏差 Sy의 3倍 이상이 되는 일은 거의 없다. 表4.2.1에서 回帰線에 대한 아내의 年齡의 標準偏差를 計算하면 다음과 같이 된다.

$$S_y = \sqrt{\frac{\Sigma (y - y')^2}{n}} = 2.03$$

表4.2.4 回帰線에 대한 標準偏差

x	y	y'	$y - y'$	$(y - y')^2$
24	21	20.5	+ 0.5	0.25
26	20	22.2	- 2.2	4.84
28	24	23.9	+ 0.1	0.01
30	27	25.6	+ 1.4	1.96
32	28	27.3	+ 0.7	0.49
34	29	29.1	- 0.1	0.01
36	33	30.8	+ 2.2	4.84
38	28	32.5	- 4.5	20.25
40	37	34.2	+ 2.8	7.84
42	35	35.9	- 0.9	0.81
330	282			41.30

(2) 非直線回帰

男便의 年齡에 대한 아내의 年齡의 回帰線은 直線으로 表示되었지만 回帰線은 원래 相関図상의 각 座標点을 통하는 平均線인 만큼 만일 이 座標点이 曲線的으로 퍼져 있을 경우에는 이를 回帰曲線으로 表示할 必要가 있다. 이제 10명의 社員의 入社試験成績 x와 이들에 대한 每月의 販売成績 y에 대해서 直線相関이 아니라 曲線相関(Curvilinear Correlation)이 成立한다고 하면 回帰曲線은 다음과 같은 正規方程式을 利用하여 이를 計算할 수가 있다.

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^2$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

表4.2.5　　入社試験과　販売成績

	A	B	C	D	E	F	G	H	I	J
入社試験	4	5	6	4	5	6	6	7	9	8
販売成績	5	4	5	6	9	10	9	12	11	9

만약에 x가 1連記号로 주어 졌을 경우에는 原点을 数列의 中央으로 옮기면 $\Sigma x = 0$, $\Sigma x^3 = 0$이 되어 위의 式은 다음과 같이 간단하게 된다.

$$\Sigma y = na + c\Sigma x^2 \cdots$$

$$\Sigma xy = b\Sigma x^2 \cdots$$

$$\Sigma x^2 y = a\Sigma x^2 + C\Sigma x^4$$

이 正規方程式을 풀면,

$$a = \frac{\Sigma y \Sigma x^4 - \Sigma x^2 \Sigma x^2 y}{n\Sigma x^4 - (\Sigma x^2)^2}$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$c = \frac{n\Sigma x^2 y - \Sigma x^2 \Sigma y}{n\Sigma x^4 - (\Sigma x^2)^2}$$

가 된다. 따라서 이 경우에는 回帰曲線의 定数 a , b , c의 計算은 매우 간단하게 된다.

曲線方程式은 一般的으로 整多項式의 形態로 주어진다. 즉,

$$y = a + bx + cx^2 + dx^3 + \cdots\cdots$$

가 그것이다. 이 경우에는 더욱 自由로운 모양의 回帰線을 적합시킬 수가 있다. 일반적으로 k次 整多項式은 $k-1$개의 커브를 갖는 曲線이 된다. 方程式의 定数 a , b , c . . . . 는 傾向直線의 경우와 마찬가지로 最小自乘法에 의해서 変量 y의 回帰線 y'에 대한 偏差의 自乘合計

$$S = \Sigma\{y - (a + bx + Cx^2 + \cdots\cdots)\}^2$$

이 最小로 되도록 정해진 다음과 같은 正規方程式에서 計算된다.

$$\Sigma y = na + b\Sigma x + c\Sigma x^2 + \cdots\cdots$$

$$\Sigma xy = a\Sigma x + b\Sigma y^2 + c\Sigma x^3 + \cdots\cdots$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 + \cdots$$

a , b , c . . . . 의 값은 이와같이 連立方程式에 의해서 구해지는 것이나 回帰方程式이 k 次인 경우에는 正規方程式의 式数는 $k+1$ 개가 된다. 이는 구하고자 하는 定数의 数가 $k+1$ 개가 되기 때문이다. 連立方程式의 式数가 많아지면 方程式을 풀기가 어려워지지만 3 次式 이상의 高次式을 回帰線으로 使用하는 경우는 별로 없다.

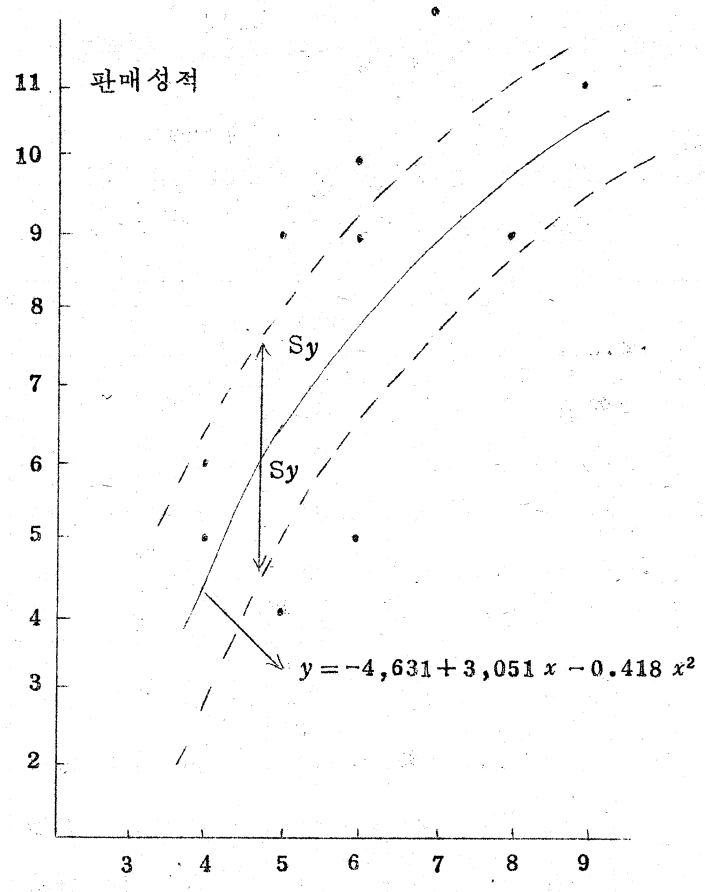
回帰曲線을 表4.2.5에 구해 보면 다음과 같다.

表4.2.6 回帰曲線의 計算

	x	y	xy	x^2y	x^2	x^3	x^4
A	4	5	20	80	16	64	256
B	5	4	20	100	25	125	625
C	6	5	30	180	36	216	1,296
D	4	6	24	96	16	64	256
E	5	9	45	225	25	125	625
F	6	10	60	360	36	216	1,296
G	6	9	54	324	36	216	1,296
H	7	12	84	588	49	343	2,401
I	9	11	99	891	81	729	6,561
J	8	9	72	576	64	512	4,096
合 計	60	80	508	3,420	384	2,610	18,708

$$\begin{cases} 80 = 10a + 60b + 384c \\ 508 = 60a + 384b + 2{,}610c \\ 3{,}420 = 384a + 2{,}610b + 18{,}708c \end{cases}$$

図 4.2.3　　回帰曲線에　대한　標準偏差



이　方程式을　풀면　a = -4,631 , b = 3,051　c = -0.148

그러므로　回帰曲線은　$y' = -4{,}631 + 3{,}051x - 0.148x^2$ 으로 된다.　回帰曲線에　있어서도　直線의　경우와　마찬가지로　標準誤差

Sy를 計算할 수가 있다.

表4.2.5에 대해서 計算한 標準偏差는 다음과 같으며 $y' \pm Sy$를 図示한 것이 図4.2.3이다.

$$Sy = \sqrt{\frac{\Sigma (x - y')^2}{n}}$$

$$= \sqrt{\Sigma \{ y - (-4,631 + 3,051\, x - 0.148\, x^2) \}^2 / 10}$$

$$= \sqrt{36,0311 / 10} = 1.9$$

表4.2.7 回帰曲線에 대한 標準偏差

	y	y'	$y - y'$	$(y - y')^2$	$y' - Sy$	$y' + Sy$
A	5	5.21	0.21	0.0441	3.31	7.11
B	4	6.92	2.92	8.5264	5.02	8.82
C	5	8.35	3.35	11.2225	6.45	10.25
D	6	5.21	0.79	0.6241	3.31	7.11
E	9	6.92	2.08	4.3264	5.02	8.82
F	10	8.35	1.65	2.7225	6.45	10.25
G	9	8.35	0.65	0.4225	6.45	10.25
H	12	9.47	2.53	6.4009	7.57	11.37
I	11	10.84	0.16	0.0256	8.94	12.74
J	9	10.31	1.31	1.7161	8.41	12.21
合 計	80			36.0311		

(3) 多元回帰分析

結果가 되는 從属変数 y에 대하여 原因이 되는 独立変数는 x하나만이 아니라 2개 이상(x_1, x_2. . . .)있는 것이 도리어 보통이다. 예컨대 賃金은 단순히 物価에만 관련이 있는 데 그치지 않고 労動態率, 労動供給, 労動立法, 企業利潤등과도 관련되어 있는 것이다.

따라서 이들 要素의 어느 것이 変化하면 賃金도 많던 적던 반드시 変化하게 마련이다. 結果変数 y에 대하여 独立変数 2개 이상을 생각하는 경우의 回帰分析을 多元回帰分析(multi - variate regression analysis)이라 한다.

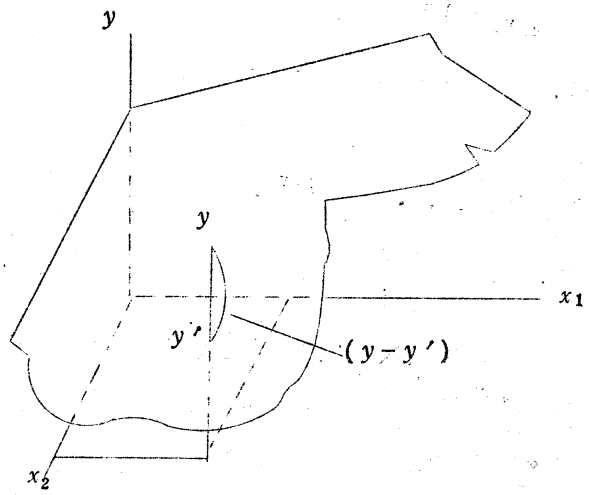
多元回帰의 경우에도 1変数回帰의 경우와 마찬가지로 從属変数 y에 대한 独立変数 x_1, x_2. . . . 의 関係는 直線인 경우도 있고 非直線인 경우도 있다.

独立変数를 x_1, x_2라 하면 回帰方程式은 直線의 경우에는 다음과 같이 된다.

$$y = a + b_1 x_1 + b_2 x_2$$

여기서 a_1, b_1, b_2는 方程式의 定数이고 b_1, b_2는 部分回帰係数(partial regression coefficient)라는 것이다. b_1은 変量 x_2가 一定할때 変量 x_1의 変化로써 일어나는 y의 変化를 나타내고 b_2는 x_1이 일정할 때 x_2의 変化로써 일어나는 y의 変化를 나타낸다.

図4.2.4　　回帰平面



独立変数가 하나인 경우의 回帰方程式 $y = a + bx$가 直線을 표시하는 것과 마찬가지로 独立変数가 2개인 경우의 回帰方程式 $y = a + b_1 x_1 + b_2 x_2$는 x_1, x_2, y의 3軸으로 規定되는 하나의 3次元의 空間에 있어서는 平面을 表示하고 있다. $x_1 = 0$일때 $y = a + b_2 x_2$는 $x_1 = 0$ ($y - x_2$平面)의 直線을 나타내고 $x_2 = 0$일 때 $y = a + b_1 x_1$은 $x_2 = 0$ ($y - x_1$平面)의 直線을 나타내며 또한 $y = 0$인 때 $a + b_1 x_1 + b_2 x_2 = 0$은 $x = 0$ ($x_1 - x_2$平面) 위의 直線을 나타낸다.

이제 x_1, x_2, y가 주어졌다면 이는 x_1, x_2, y의 3軸으로 規定되는 空間의 1点을 나타낸다. x_1, x_2에는 誤差가 없고 y만이 誤差를 가졌다고 보아 $x_1 - x_2$平面위의 点(x_1, x_2)에 垂線을 내리면 구하는 平面은 図4.2.4와 같이 점y'에서 交叉되는 바 이를 回帰平面(regression plane)이라 한다. 말하자면 x_1, x_2 위의 y의 回帰平面은 $y = a + b_1 x_1 + b_2 x_2$로 表示된다. 方程式 $y = a + b_1 x_1 + b_2 x_2$의 定数 a, b, c의 값은 1変数 回帰의 경우와 마찬가지로 変量 y와 그 回帰線

y'에 대한 偏差의 自乘合

$$S = \Sigma \{ y - (a + b_1 x_1 + b_2 x_2) \}^2$$

이 最小가 되도록 정하면 된다.

一般的으로 多元回帰方程式

$$y = a + b_1 x_1 + b_2 x_2 + \cdots\cdots + b_k x_k$$

의 定数 $a, b_1, b_2 \cdots\cdots b_k$의 값은

$$S = \Sigma (y - y')^2$$

$$= \Sigma \{ y - (a + b_1 y_1 + b_2 x_2 + \cdots\cdots + b_k x_k \}^2$$

이 最小가 되도록 하여 얻은 正規方程式에서 計算된다. 正規方程式은 $a, b_1, b_2 \cdots\cdots b_k$ 라는 $(k+1)$元 1次連立方程式으로 주어진다.

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2 + \cdots\cdots + b_k \Sigma x_k$$

$$\Sigma x_1 y = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 \cdots\cdots + b_k \Sigma x_1 x_k \cdots$$

$$\Sigma x_k y = a \Sigma x_k + b_1 \Sigma x_1 x_k + b_2 \Sigma x_2 x_k + \cdots\cdots + b_k \Sigma x_k^2$$

다음 表4.2.8은 独立変数가 2개일 경우의 回帰方程式의 計算을 表示한 것이다. 이를 테면 x, y에 第2의 説明変数 x_2를 추가하여 y를 x_1과 x_2로 説明하려는 回帰分析이다. 回帰方程式은,

$$y = a + b_1 x + b_2 x_2$$

이고 定数計算을 위한 正規方程式은

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$\Sigma x_1 y = a \Sigma x_1 + b \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Sigma x_2 y = a\Sigma x_2 + b_1\Sigma x_1 x_2 + b_2\Sigma x_2^2 \quad \cdots\cdots\cdots \quad (i)$$

이다. 따라서 表4.2.8에서 얻은 結果를 (i)式에 代入하면 다음과 같은 連立方程式이 얻어진다.

表4.2.8 多元回帰의 計算

y	x_1	x_2	$x_1 y$	$x_2 y$	$x_1 x_2$	x_1^2	x_2^2	y^2
1	3	2	3	2	6	9	4	1
2	2	1	4	2	2	4	1	4
2	5	1	10	2	5	25	1	4
4	4	3	16	12	12	16	9	16
6	6	5	36	30	30	36	25	36
15	20	12	69	48	55	90	40	61

$$15 = 5a + 20b_1 + 12b_2$$

$$69 = 20a + 90b_1 + 55b_2$$

$$48 = 12a + 55b_1 + 40b_2$$

이를 풀면,

$$a = -0.24 \quad b_1 = 0.27 \ , \ b_2 = 0.90$$

을 얻는다. 따라서 구하는 回帰方程式은,

$$y' = -0.24 + 0.27 x_1 + 0.90 x_2$$

가 된다.

이 方程式에서는 $b_1 = 0.27$은 x_2가 不変일 때 x_1의 증가에

대해서 y는 平均 0.27만큼 증가함을 표시하고 또 $b_2 = 0.90$은 x_1이 같고 x_2가 変할 경우에 x_2의 증가에 대하여 y가 0.90 증가함을 表示한다.

多元回帰의 경우에도 推定値에 대한 変量의 標準偏差는 1変数回帰의 경우와 마찬가지로 다음과 같이 구해진다.

$$Sy = \sqrt{\frac{\Sigma(y-y')^2}{n}}$$

$$= \sqrt{\Sigma\{y-(0.24+0.27\,x_1+0.90\,x_2)\}^2/5} = 0.75$$

表4.2.9 多元回帰에 대한 標準偏差

y	y'	$y-y'$	$(y-y')^2$
1	2.4	- 1.4	1.96
2	1.2	0.8	0.64
2	2.0	0.0	0.00
4	3.5	0.5	0.25
6	5.9	0.1	0.01
15			2.86

第二編　統計調査方法

第二编　统计调查方法

第 1 章. 統計調査一般

1.1 統計調査의 定義와 種類

統計라고 하면 우리는 곧 우리나라의 総人口, 通貨量, 物価指数, 人口, 出生, 死亡, 生産, 輸出, 貯蓄, 失業 等과 같은 社会現象에 関한 것을 생각하게 된다. 이와같이 統計는 어떤 集団에 関한 事実을 説明하는 것이며 따라서 集団이 아닌 個体(例, 金某人, A企業体 等)에 関한 個別的인 事実을 説明하는 것은 統計가 아니다. 이와 같이 集団에 関한 事実을 説明하는 것은 統計를 作成하기 위하여 우리는 統計調査를 한다.

統計調査라 함은 合目的的인 表示에 依하여 統計集団 또는 部分集団을 形成함을 말하는 것이나 간단히 말하면 統計資料를 蒐集整理하여 統計를 作成하는 節次이다.

이러한 統計調査의 種類를 살펴보면 다음과 같다.

(1) 全数調査와 標本調査

統計調査는 統計集団에 関한 事実을 알려고 하는 것이므로 統計集団을 形成하고 있는 単位全部를 調査하는 것이 所望스러운 일이나 時間, 費用, 人員 等 많은 制約이 있고 또 正確한 調査라는 観点에서도 問題되는 点이 있어 오히려 調査単位의 一部만을 調査하는 경우가 더 많다. 이와같이 調査単位를 全部調査하는 것을 全数調査라고 하고 調査単位의 一部만을 調査하는 것을 標本調査라고 한다.

全数調査와 標本調査는 統計調査의 代表的인 것이며 両調査가 密接한 関係가 있으므로 다음 節에서 詳述하기로 한다.

(2) 第一義 統計調査와 第二義 統計調査

第一義 統計調査는 統計를 作成하는 그 自体를 目的으로 実施하는 調査이며 第二義 統計調査는 統計以外의 目的으로 作成한 記録書類를 利用하여 統計를 만드는 調査이다. 普通 統計調査의 大部分은 前者의 경우이다.

第二義 統計調査의 例를 들면 人口動態調査, 建築許可 統計, 貿易統計와 같은 것이 있다.

第二義 統計調査는 報告 또는 記録의 制度만 있으면 調査의 容易性, 経費, 被調査者의 負担 等을 考慮할 때 第一義 統計調査를 하는 것 보다 쉽게 統計를 作成할 수 있는 것은 明白한 일이다.

反面 第二義 統計調査에는 다음의 例와 같은 欠点이 많다.

戸籍上으로 申告되는 出生, 死亡, 離婚 等의 統計에 있어서 相当한 期間이 経過한 後 発生事項이 申告되며 이를 다시 具体化 하여 統計를 作成할 때까지는 많은 時間을 消費하여 利用에 不便할뿐만 아니라 申告하지 않는 경우가 많이 있어 精度를 低下시킨다.

(3) 静態統計調査와 動態統計調査

統計는 静態統計가 아니면 動態統計에 속한다. 一定한 時点에 있어서의 調査対象의 状態를 調査함으로써 얻는 統計를 静態

統計라 하며 一定한 期間內에 繼続 発生하는 事象을 調査함으로써 얻는 統計를 動態統計라고 한다.

5年 또는 10年에 1回씩 実施하는 人口센서스와 分期別調査인 経済活動 人口調査 또는 每月調査인 人口動態調査, 그리고 鉱工業 센서스와 鉱工業動態調査, 都小売業센서스와 都小売額動態調査 等은 서로 補完하는 調査이며 静態와 動態를 把握하는 면에서 약간 차이가 있다.

統計를 利用할 때에는 両者를 관련시켜 利用하면 그 効用을 높일 수 있다.

1.2 統計調査의 必要性

必要한 統計를 既存資料에 依하여 作成할 수만 있다면 資料蒐集을 위한 時間, 費用, 効力이 節約될 수 있으므로 그위에 좋은 일이 없을 것이다. 그러나 社会現象은 不断히 変化発展하고 있으므로 항상 必要한 既存資料를 갖고 있기는 어려운 일이다.
그뿐 아니라 既存資料란 그것이 統計를 만들기 為하여 蒐集된 것이 아니고 다른 목적에 쓰기 為하여서 蒐集된 것이므로 이를 統計化하는 데는 여러가지 制約이 따르는 것이다. 따라서 基本이 되는 蒐集이 먼저 遂行되지 않으면 안된다.

그런데 統計의 対象이 되는 것은 統計集団이지만 統計調査의 対象이 되는 것은 統計集団 그 自体가 아니고 統計集団을 構成하고

있는 個体(統計単位)이다.

즉 統計는 集団 그 自体에 関한 数量的 知識을 얻는 것이지만 이 統計의 基本資料를 얻는 統計調査는 現実的으로 個個単位를 対象으로 하여 実施되는 것이다.

이때 統計調査의 現実的인 対象이 되는 個個単位를 調査単位라고 한다.

그러나 統計調査가 비록 個個単位를 対象으로 한다고 하더라도 궁극적으로는 集団에 関한 知識(統計)을 얻는데 目的이 있는 것이므로 当初 個体 그 自体에 関한 知識을 얻고자 하는 調査(個別調査)와는 厳密하게 区別된다. 統計調査가 個別調査와 区別되는 点으로,

첫째, 統計調査는 集団에 関한 知識을 얻기 위한 実践的인 過程으로서 그 結果는 반드시 統計로 俱現되는 것이며,

둘째, 集団의 構成単位(調査単位)를 大量으로 観察하는 것이며,

세째, 統計化가 可能하도록 調査単位에 対하여 標識을 부여하여야 하는 것이다.

統計調査는 雑多한 社会現象을 統計作成者의 意図(目的)에 맞추어 統計集団으로 만들어야 하고 統計調査의 目的을 이룰 수 있도록 各 調査単位에 対한 正確한 심사가 이루어져야 하며 이에 뒤따라 当初의 目的에 맞는 統計表를 作成·発表하여야 한다.

統計調査의 全過程을 図示하면 다음과 같다.

企劃 → 準備 → 実査 → 審査·集計 → 統計表作成

統計調査方法은 統計作成者는 勿論 統計를 利用하기만 하는 사람도 統計를 올바로 보고 利用하기 위하여 充分히 理解하여야만 한다.

왜냐하면 統計는 그 目的과 作成者의 立場 및 統計調査方法의 如何에 따라서 얼마든지 다른 結果가 나올 수 있기 때문이다.

統計調査方法은 다음에 調査過程에 따라 詳述하기로 한다.

第2章 · 全數調査와 標本調査

統計調査는 앞에서 말한바와 같이 統計集団에 関한 事実을 알려고 하는 것이므로 統計集団을 構成하고 있는 単位 全部를 調査하는 것이 가장 所望스러운 것이다. 그러나 実際로 調査単位를 全部 調査한다는 것은 時間, 費用, 人員等 많은 制約이 있다.

뿐만아니라 一部만을 調査하고서도 統計集団에 関하여 正確한 知識을 얻을 수만 있다면 오히려 一部만을 調査하는 것이 所望스러운 것이다.

위에서와 같이 調査単位를 全部 調査하는 것을 全数調査라고 하며 調査 単位의 一部만을 調査하는 것을 標本調査라고 한다.

2.1 全数調査

全数調査는 調査対象이라고 생각되는 모든 部分을 全部 調査하는 것을 말한다. 센서스 같은 것은 그 좋은 例이다. 센서스 뿐만아니라 調査対象의 範囲가 한개의 事務室에 불과한 경우라든가 또는 하나의 洞里라든가 한 学校内의 한 学級인 경우라도 그 調査対象의 最終単位를 全部 調査하는 限 이것은 全数調査인 것이다.

그러나 統計調査의 境遇 全数調査는 一般的으로 그 規模가 크다. 여기서 한가지 注意할 것은 어떤 調査対象의 範囲内에 있는 모든 要素를 調査하는 것만이 全数調査가 아니다. 例를 들면 農場만을 全部 調査하면 全数調査가 되는 것이며 農場以外의 家屋이라든가

学校等 그 地域의 모든 것을 調査하여야 全数調査가 되는 것은 아니다.

2.2 標本調査

標本調査는 部分調査라고도 한다. 즉, 이것은 前述한 諸標本抽出法에 따라 調査対象의 一部分을 선출하여 그 全体를 推定하는 調査를 말한다. 이때 調査対象 全体를 母集団이라고 하며 선출된 部分을 標本이라고 한다. 따라서 여기서 問題가 되는 것은 全体를 代表하는 部分을 如何히 抽出하는가 하는 点이다. 많은 調査가 이와 같은 標本調査法에 依하여 이루어지고 있다.

그러나 原則的으로 時間이라든가 費用, 또는 人員이 허락한다면 全数調査가 좋다는 것은 말할 必要가 없는 것이다. 이것은 勿論 精密性이라는 点에 基準을 둔 말이나 調査目的에 비추어 어느 程度의 誤差는 問題視 안되는 경우라든가 또는 調査目的이 많은 사람들 중에 大多数의 어떤 傾向만을 알고자 할 때에는 구태여 많은 人力과 時間을 浪費할 必要가 없는 것이다. 勿論 어느 程度 精密한 知識이 必要한가에 따라서 決定할 問題이며 이 点에 関하여는 앞의 標本調査論에서 説明하였다.

2.3 全数調査와 標本調査의 比較

全数調査와 標本調査는 어느 것이 더 낫다든가 못하다고

한말로 말할 수가 없다. 両者는 各各 長短点이 있으므로 調査의 性格에 비추어서 両者중 하나를 択하여야 한다.

全数調査와 標本調査의 長点을 가려내어 両者의 選択基準을 추려보면 다음과 같다.

(1) 全数調査에는 標本誤差가 없다. 따라서 센서스와 같이 誤差가 전혀 없거나 혹은 그것이 최소한에서 그쳐야할 경우에는 부득이 全数調査를 할 수 밖에 없다.

(2) 事後의 分類가 細分된 각층마다 이루어져야 하고 正確한 資料가 要求될 때에는 標本調査를 하더라도 많은 標本을 必要로 함으로 이 경우 차라리 全数調査를 하는 것이 좋다.

(3) 母集団(統計集団)이 比較的 작은 경우에는 標本調査를 하더라도 推定의 精度를 높이기 위하여서는 全数調査와 거의 対等한 精度의 많은 標本을 推出하여야 한다. 따라서 이때에도 全数調査를 하는 것이 낫다.

(4) 또 어떤 目的을 위하여서는 標本調査는 많은 制約이 있으므로 이 境遇에는 全数調査를 하여야 한다.

(5) 標本調査를 하기 為하여서는 標本抽出 推定, 誤差計算等 보다 専問的인 統計知識이 必要하므로 이러한 知識을 가진 사람을 求하기 힘들때에는 全数調査를 하는 것이 安全하다.

(6) 標本抽出을 위하여는 母集団에 関한 기초자료가 있어야 하는 것인데 이러한 것이 없는 境遇에는 不得已 全数調査를 할 수 밖에 없다.

(7) 全数調査는 調査의 規模가 크기 때문에 巨額의 費用을 必要하게 되는데 이러한 費用의 制約이 있는 경우에는 標本調査를 하면 費用을 節約할 수 있다.

(8) 全数調査는 実査와 集計에 많은 時間이 所要되는데 対하여 標本調査는 이를 短縮 할 수 있으므로 調査 結果를 迅速히 公表하여야할 경우에는 標本調査를 하는 것이 좋다.

(9) 全数調査에서는 많은 調査人員을 必要로 하며 따라서 未熟한 調査員을 使用하지 않을 수 없음에 対하여 標本調査에서는 훨씬 小数의 調査員으로서도 調査가 可能하기 때문에 숙련된 調査員을 使用하지 않으면 안될 경우에도 訓練하기가 容易하다.

(10) 全数調査는 標本誤差가 없는 대신 調査의 大規模性으로 因하여 非標本 誤差가 標本調査에서 보다 크다.
따라서 非標本誤差를 줄이는 것이 重要한 경우에는 標本調査를 하는 것이 좋다.

(11) 그리고 現実的으로 全数調査를 実施할 수 없거나 調査의 性質上 全数調査가 부적당할 경우도 많은데 이러한 경우에는 不得已 標本調査를 하여야 한다. 以上 열거한 것 以外에 標本調査는 標本抽出의 기초자료를 全数調査 結果에 依存하여야 하므로 標本調査를 위하여서도 数年에 一回씩 全数調査를 할 必要가 있고 또 標本調査에서는 推定値 보다 平均이라든가 比率에 더 重点을 주는데 이는 全数調査에서 얻어지는 構造等과 結付됨으로써 더욱 利用価値가 높아지는 것이므로 全数調査와 標本調査는 서로 補完関係에 있다고 할 것이다.

第 3 章　統計調査의　方法

全数調査와　標本調査의　概念과　両調査의　性格을　앞에서　略述하였다. 그러나 이러한 調査를 実施하고자 할 때에는 調査主体와 客体, 調査事項, 調査時期 其外에 地域 및 方法等 여러가지 制約을 받게 된다.　다음에 調査上 一般的으로 要請되는 몇가지 注意事項을 든다.

(1) 合目的性 : 調査対象이 그 目的과 合理的으로 適合되어야 한다.

例를 들면 貨幣의 一般的 購買力을 調査하려면 都売物価를 調査하여야 하고 労動者의 生活程度를 観察하려면 名目賃金 보다 実質賃金의 調査가 더욱 重要하다.

같은 対象의 調査에 関하여서는 目的에 따라 調査의 時期, 場所, 方法等에 差異가 要求될 것은 勿論이다.

(2) 合理性 : 統計対象에 따라서 調査時期·場所 및 그 方法은 科学的인 同時에 経済的이어야 한다.　이를테면 不必要하게 細密한 数学的 調査는 時間과 努力 및 費用의 経済上 반드시 合理的인 方途는 아니다.

(3) 資料의 同質性 : 統計資料의 時間的 및 場所的인 比較 利用度를 向上시키기 위하여는 可及的으로 調査項目, 調査時期, 其他 調査方法을 同質性으로 維持함이 必要하다. (例、國際的 各種 統計基準의 定立과 必要性)

(4) 小間多知의 原則 : 標識의 選定에 있어서 簡単한 調査로서 充分히 実質的인 目的을 達할 수 있도록 最善의 方法을 講究함이 必要하다. 이를 少間多知의 原則이라 한다. 이를테면 道德統計에 있어서 國內의 善行을 일일이 調査하는것 보다 犯罪件数를 調査함이 便利하고 國民所得의 調査에 있어서 收入部門을 일일이 調査合計하는것 보다 消費額, 貯蓄額 等이 支出面에서 이를 把握함이 簡便할 수가 있다.

(5) 外囲的 制約 : 観察対象이 合目的的이라 할지라도 이를 実際調査함이 事実上 不可能할 경우, 法律 또는 社会慣習上 不容認된 경우 또는 経済上 不合理한 경우가 있으니 이러한 外囲的制約은 調査에 있어 미리 留念하여야 한다.

以上에서 列挙한 調査上의 留意点을 念頭에 두고 統計調査를 企劃하여야 함은 勿論이다.

다음에 調査基準 実際調査 調査票의 審査와 集計, 統計表의 作成과 分析等 統計調査의 企劃부터 統計票의 作成까지를 段階的으로 略述한다.

3.1 統計調査의 企劃과 準備

(1) 調査目的

統計調査의 企劃中에서 가장 먼저 하여야 될 일은 調査目的의 設定이다. 調査의 目的이 있으므로서 調査가 実施되는 것

이며 모든 調査過程은 調査目的에 의하여 規制되어야 하는 것이다. 그러므로 調査目的의 設定은 매우 重要하다. 그러나 간혹 調査目的이 不明確하고 抽象的이어서 具体的으로 어떠한 것인지 模糊한 것이 있다.

그뿐 아니라 調査過程에 있어서 調査目的이 忘却되어 調査目的과는 엉뚱한 다른 方向으로 調査가 行하여지기도 한다. 이와같이 調査目的에서 離脫되거나 調査目的이 変質되지 않기 위하여는, 첫째 調査目的을 明確하게 그리고 具体的으로 設定하여야 한다. 調査目的에는 一般的으로 무엇에 関한 知識을 必要로 하며 그 知識은 어디에 利用할 것인가를 明白히 하여야 한다.

둘째, 調査目的과 아울러 생각해야 할 것은 調査目的을 実現하기 위하여 統計的 方法이 採択될 수 있는가 하는 統計調査의 適合性問題이다. 統計調査는 社会現象을 把握하는데 있어 매우 有力한 手段이며 漸次 그 利用度도 높아가고 있지만 모든 分野에서 完全한 手段은 되지 못하고 一定한 限界가 있는 것이다. 統計調査가 그 調査目的을 実現함에 있어 可能하고 適合한 것인가를 判断하기 為하여는 다음 事項을 検討하면 된다.

(가) 調査의 目的으로 하는 知識이 集団에 関한 知識인가

(나) 調査対象이 現実的인 集団이거나 集団化 시킬 수 있는 것인가

(다) 그 集団에 統一된 標識은 賦与할 수 있는가

(라) 그 調査結果가 数量으로 集団을 正確히 表現할 수 있는 것인가

(마) 그리고 그 結果가 客觀性과 普遍性이 있는가

(바) 統計調査가 可能하다 하더라도 現実的으로 隨伴되는 困難 即 大量觀察에 따르는 費用·時間·人員을 克服할 수 있는가

(사) 大量觀察로 因한 얕고 平面的인 知識이 統計目的을 充足시킬 수 있는가

(1) 調査範囲

調査目的이 決定되면 다음에는 調査範囲를 定하는 것이 重要한 일이다. 調査範囲를 定하는 일은 곧 統計集団의 範囲를 定하는 것이며 바꾸어 말하면 現実의 社会現象을 여러가지 要素에 依하여 統計集団化 시키는 것이다. 따라서 調査範囲는 곧 調査集団인 것이다.

調査範囲는 이를 概念的範囲, 時間的範囲 및 場所的範囲로 나누어 볼 수 있다. 概念的範囲라 함은 알려고 하는 社会現象의 属性(標識)을 말하는 것으로서 例컨데 鉱工業, 都売物価等 社会現象의 種類를 限定하고 그 概念을 明確히 規定하여야 하는 것이다.

따라서 鉱工業이라 하더라도 그것이 어떤 規模의 것을 말하는 것인지 明確히 定하여야 하는 것이다. 時間的 範囲는 언제 現在의 社会現象을 調査할 것인가 하는 問題이다. 따라서 静態統計調査에 있어서는 一定한 時点을 그리고 動態統計調査에서는 一定한 期間을 말하는 것이다.

그리고 場所的範囲라 함은 어느 地域에 対한 社会現象을 調査할

것인가 하는 問題이다. 調査範囲는 調査目的에 依하여서 規定되는 것이지만 人員·費用·時間等 与件에 맞추어서 이를 定하지 않으면 안될 것이다.

(3) 調査単位

統計調査에 있어서 調査의 対象이 되는 것이 調査単位이다. 即 必要한 情報를 提供하여 주는 것이 調査単位이다. 따라서 調査를 実施하기 以前에 調査単位를 무엇으로 할 것인가를 定하여야 한다.

그런데 調査単位는 統計集団을 構成하는 単位와 반드시 一致하는 것이 아님을 注意하여야 한다.

例를 들면 人口센서스에 있어서 人口라는 統計集団의 単位는 한 사람 한사람의 個人이지만 調査는 家口를 調査単位로 하여 実施되는 것이다.

이와 마찬가지로 調査単位는 또한 集計単位·分類適用単位·抽出単位 等과도 다르다. 集計単位는 集計에 있어서의 単位이며 分類適用単位는 各種 分類를 適用시키는데 있어서의 単位로서 모두 統計集団의 構成単位와 一致하는 것이다. 그리고 抽出単位는 標本調査에 있어서의 標本을 抽出하는 単位를 말하는 것으로서 전연 別個의 概念인 것이다.

調査単位를 定하는데 있어서 問題가 되는 것은 첫째로, 무엇을 調査単位로 할 것인가 하는것과 둘째로, 調査単位를 全部 調査할

것인가 또는 一部만을 調査할 것인가 하는것이 問題이다.

後者는 곧 全數調査와 標本調査의 問題로서 言及되었으므로 여기서는 調査単位의 決定에 関하여 説明하고자 한다.

調査単位는 統計集団의 構成単位와 一致하는 것이 普通이지만, 때에 따라서는 調査単位를 순전히 調査上의 便宜를 為하여 別途로 定하기도 한다. 調査単位는 가장 正確한 情報를 얻을 수 있고 또한 빠짐없이 情報를 모을 수 있는 最小限의 単位로 하는 것이 普通이지만 鉱工業 調査와 같은 産業調査에 있어서는 一般的으로 事業体를 調査単位로 한다.

(4) 調査項目

調査目的이 確定되면 그 다음에 重要한 것은 調査項目을 決定하는 일이다. 調査項目은 調査目的이 되어 있는 커다란 主題를 여러 次元으로 細分한 것으로 調査標上에 印刷되어 実際로 調査되는 具体的인 事項이다.

調査項目은 調査目的에 関聯되는 必要하고도 充分한 것이 아니면 안되므로 調査項目을 設定하는 데는 細心한 注意와 検討가 必要하다. 調査項目을 設定할 때에는 누구나 興味가 있고 価値가 있을듯 하다고 생각되는 것은 무엇이나 다 包含시키고 싶은 誘惑을 느끼게 되는데 이는 絶対로 避하지 않으면 안된다. 調査項目은 必要하고도 充分한 最小限의 것이 가장 좋은 것이다.

調査項目 設定에 있어서 検討하여야 할 点은 첫째로, 그 項目이 実際로 正確한 資料를 얻을 수 있는 것인가 하는 것과 둘째로, 그 項目이 集計할 수 있는 것인가 세째로, 그 項目이 重点的으로

選定되고 調査項目에 寄与하는 것이 明白한가 네째로, 調査時間, 調査員, 被調査者, 集計費等 現実的인 條件이 具備되어 있는가 等이다.

調査項目에는 그것이 集計되고 分析되는 基本項目과 手段 項目으로서 基本項目을 보다 容易하게 調査할 수 있도록 誘導하는 誘導項目, 調査上의 誤謬를 豫防하고 内容을 照会할 수 있는 対照項目 (Check Item)이 있다.

(5) 調査票

調査票는 調査項目을 一定한 様式으로 配列한 紙面이다. 調査票는 統計調査에 있어서 매우 重要한 用具로서 그것은 첫째로, 調査 할 項目을 明確하게 함으로써 調査項目의 漏落을 防止할 수 있고 둘째로, 各異한 調査員의 観察을 標準化하고 統一시키며 세째로, 観察을 強化하고 正確히 할 수 있는 役割을 한다.

調査票作成에 있어서는 調査票体制, 用紙形式 項目의 配列이 重要한 問題이다.

(가) 体 制

調査票의 크기는 그것이 包含할 項目数와 調査票의 使用方法에 따라서 左右되겠으나 携帯하기 쉽고 取扱保管에 便利한 크기로 하여야 하며 比較的 간단한 調査票로 調査目的을 達成할 수 있는 경우에는 一般的으로 접을 必要가 없는 정도의 크기가 좋다.

調査票의 紙質은 記入, 分類, 集計, 保管等에 便利한 堅固한 것으로서 카드가 좋을 것이다.

그리고 事前 分類가 可能한 것은 調査票의 色을 달리하여 만드는 것도 分類 및 集計를 為하여 便利하다. 調査票의 体制에 있어서 그 크기와 紙質外에 重要한 것은 첫째로, 実査의 方法을 考慮하여야 하는 点이다.

実査의 方法에는 面接調査法·配票調査法·集合調査法·郵便調査法·電話調査法이 있으며 調査票의 記入 方法에는 被調査者가 스스로 記入하는 自計法과 調査者가 記入하는 他計法이 있다. 調査票를 設計할 때에는 위의 어느 方法을 採用한 것인가가 이미 決定되어 있지 않으면 안된다. 그것은 어느 実査方法을 採用할 것인가에 따라서 調査票의 体制가 달라지기 때문이다.

그리고 둘째로, 이와 아울러 생각하여야 할 것은 調査員의 質도 考慮하여야 한다는 点이다. 能力이 미치지 못하는 調査員에게 複雜한 調査票는 절대로 避해야 한다. 調査票는 普通 程度의 調査員도 容易하게 調査할 수 있도록 쉽게 꾸며지는 것이 좋은 것이다.

세째로, 考慮해야 할 것은 集計의 方法이다. 機械集計에 依하고자 할때는 機械集計의 專門家와 協議하여 集計에 便利하도록 調査票를 作成하여야 한다.

(나) 項目配列

調査員은 調査票上의 項目配列 順序에 따라 質問하게 되는 것이므로 質問의 順序가 論理的으로 矛盾이 없어야 하며 応答의

效果를 높이기 為하여 調査項目의 配列에 特히 신경을 쓰지 않으면 안된다. 項目配列은 大体로 다음과 같은 順序에 依하는 것이 좋다.

1) 被調査者가 応答하기 쉬운 것부터 始作할 것.

2) 可能한 限 論理的인 順序에 依하되 難問題는 中間 또는 끝가까이에 配列할 것.

3) 一般的인 項目에서 細部的인 項目으로 展開할 것.

4) 可及的 서로 関聯된 項目이나 集計할 때 함께 必要한 項目은 調査票의 같은 部分에 位置하도록 할 것.

以上에 依하여 項目을 配列하면 다음에 各項目에 一連番号 또는 符号를 붙이는 것을 잊어서는 안된다. 그리고 調査票에는 반드시 符号欄을 設置하여야 한다. 符号欄에는 各種 分類를 表示하는 番号 또는 符号를 記入하는 것이다. 符号欄에는 調査票上의 適切한 位置에 設置하되 可能한 限 集計員이 보기 좋은 곳에 모으거나 当該回答의 가까이에 設置하는 것이 좋을 것이다.

(다) 用語形式

調査票上의 用語나 文句는 그 意味가 完全하게 表現되어야 하며 普通의 知識을 가진 사람이면 明確하게 理解할 수 있는 것이 아니면 안된다. 調査要領書가 別途로 있다고 하더라도 可能한 限 調査票만 가지고서도 調査項目의 意味를 正確히 알 수 있도록 하는 것이 좋다.

調査票의 用語形式에 있어서 特히 注意할 것은 調査項目이 主観

的인 対答을 要求하는 것이어서는 안된다는 것이다. 調査票는 어디까지나 客観的이며 明確하고 簡単한 答을 얻을 수 있는 形式으로 꾸며져야 하는 것이다.

(6) 分 類

統計調査의 結果 蒐集된 調査票는 調査対象에 関한 가장 詳細한 情報이긴 하지만 이를 一定한 方式에 依하여서 整理하고 集計하지 않으면 集団에 関한 知識을 얻을 수는 없다.

統計에 있어서는 一定한 方式에 依한 整理 即 分類가 절대적으로 必要한 것이며 이는 반드시 企劃過程에서 그 基準이 設定되지 않으면 안되는 것이다. 分類라 함은 調査対象을 몇개의 구룹으로 나누는 것으로서 同一한 구룹에 나누어진 것은 同質的인 것으로서 取扱되고 統計数字는 이 구룹에 関하여서만 나타나게 되는 것이다. 따라서 구룹内에 있어서의 異質性은 無視되고 個別的인 特殊한 情報는 捨象되어 버리고 마는 것이다. 이와같이 分類란 調査結果의 複雜한 情報의 集計中에서 必要한 것을 가려내어 簡単한 形態로 整理하는 것을 말하는 것이다.

分類에는 質的分類와 量的分類가 있다. 質的分類란 事業(産業)의 種類나 職業의 種類 或은 性別과 같은 質的인 内容에 関한 分類이며, 量的分類란 사람의 年齢, 事業体의 従業員数나 資本金과 같은 量的인 内容에 関한 分類이다.

위의 어느 것이 든지 事前에 한개 한개의 구룹을 어떤 基準에

依하여서 어느 範囲까지로 하느냐 하는 것을 定하지 않으면 안되는 것이다.

各種 分類를 作成하는데 있어서 다음과 같은 것을 考慮하여야 한다.

(가) 調査의 目的에 따라서 分類를 作成할 것.

(나) 分類된 各 구룹內에 있어서는 모든 単位가 同質的인 것이 되도록 할것.

(다) 구룹과 구룹間은 異質性이 뚜렷이 나타나도록 할 것.

(라) 分類된 単位가 하나도 남김없이 그리고 重複됨이 없이 어떤 구룹엔가 帰属되도록 할 것.

(마) 다른 種類의 調査結果와 比較될 수 있도록 分類를 作成할 것. 따라서 標準分類가 있는 것은 되도록 이를 採用할 것.

(바) 量的分類에 있어서는 合理的인 区間을 設定하되 統計의 一般原則에 따를 것.

(사) 質的分類에 있어서 그 內容이 複雜한 것은 十進分類法에 依하여 小分類·中分類·大分類와 같이 分類하는 것이 便利하다.

(아) 分類가 細分되어 구룹이 過多할 때에는 集団의 內容은 仔細히 볼 수 있으나 全体的 把握에 不便하다. 그反面 分類가 大分되어 구룹이 過少할 때에는 1個의 구룹속에 異質的인 것이 無理하게 区分되기 쉽다.

(자) 2個 以上의 구룹에 重複되어 속하게 되는것. 또는 2個의 구룹中間에 位置하는 것은 그 所属을 明白히 規定하여 두어야

한다.

(7) 調査方法

여기서 말하는 調査方法이란 実査(現地調査)의 方法을 말하는 것인데 이는 調査票設計나 調査対象의 決定과 同時에 定하여져야 하는 것이다.

実査의 方法에는 大体로 다음 다섯가지가 있다.

(가) 面接調査法

調査者가 被調査者를 直接 面接하여 質問과 応答을 通하여 調査하는 方法이다. 이 方法에 依하면 被調査者의 応答을 確実히 얻을 수 있고 調査票의 回答率을 높이며 比較的 길고 複雜한 調査票도 使用할 수 있을뿐 아니라 正確한 調査를 期할 수 있고 補充的인 情報도 얻을 수 있다는 長点이 있는 反面, 많은 調査員과 費用이 所要되고 調査員에 따라서 応答이 달라지기쉬우며 調査員의 不正行為가 発生할 念慮가 있는 点 및 特殊한 階層의 사람은 面接할 수 없다는 点 等의 短点이 있다.

(나) 配票調査法

調査者가 被調査者에게 調査票를 配付하고 一定 期間內에 이를 回収하는 方法이다. 이 方法은 面接調査法에 比하여 追及訪問이 덜 必要하고 調査票의 回収率도 比較的 좋으며 費用도 적게 들지만 被調査者 本人이 記入하지 않거나 本人이 記入한다고 하더라도 調査項目을 誤解하여 記入이 不正確하기 쉬우며 또 內容이

造作되거나 記入漏落이 생길 可能性이 많은 것이다.

(다) 集合調査法

被調査者를 一定한 場所에 集合시켜 同時的으로 調査票에 記入케 하는 方法이다. 이 方法에 依하면 調査의 說明이나 條件을 被調査者 全員에게 対하여 統一시킬 수 있는 利点이 있으며 被調査者를 容易하게 集合시킬 수만 있다면 費用이 적게 들고 簡便하게 調査를 行할 수 있으며 調査員도 小數로 足할 것이다.

그러나 集合 調査法은 特別한 경우를 除外하고는 使用하기 困難할 뿐 아니라 一般的으로 出席率이 나쁜点과 自計式의 一般的 欠陷은 免할 수 없는 点, 被調査者의 차를 無視하게 되는 点, 出席者의 日当이나 交通費를 支給하게 된다면 오히려 費用이 많이 드는 수도 있다는 点 等의 欠点이 있다.

(라) 郵便調査法

調査票를 被調査者에게 郵送하고 이를 다시 郵便에 依하여 回収하는 方法이다. 이 方法은 費用이 적게 들며 넓은 地域에 調査票를 配布할 수 있으며 面接調査가 어려운 特殊한 階層의 사람에게도 調査할 수 있는 長点이 있다. 反面에 調査票의 回収率이 낮고 回収가 어려움에 따라 督促状을 여러번 보내야 하므로 回収에 時間을 要한다.

그리고 配票調査에 있어서와 같이 自計式 一般의 欠点이 있으며 調査忌避를 防止하기 위하여서 調査票가 必然的으로 簡単해지는 等의 短点이 있다.

(마) 電話調査法

이는 調査者가 被調査者에게 電話로 質問하여 調査하는 것이다. 이 方法에 依하면 簡単하고 迅速하게 調査할 수 있으며 費用도 적게 들수 있다. 그러나 電話를 가진 사람만을 対象으로 하여야 하는点, 調査時間이 짧아야 되므로 簡単한 調査以外에는 할 수 없다는 点이 決定的인 短点이다.

위의 여러 方法中에서 集合調査法 等은 制限된 特殊한 調査에만 使用되는 方法이며 一般的으로는 面接調査法·配票調査法·郵便調査法의 順으로 많이 使用되는데 特히 面接調査法은 回収率이 높고 正確하고 詳細한 調査를 期할 수 있다는 長点으로 因하여 가장 널리 使用되고 있다.

따라서 위의 方法 以外에 調査票 記入方法으로서 自計式과 他計式이 있는데 自計式이란 被調査者가 스스로 記入하는 것이고 他計式은 調査員이 記入하는 것이다.

따라서 面接調査法과 電話調査法은 他計式에 依하는 것이며 配票調査法·集合調査法·郵便調査法은 自計式에 依하는 것이다. 이中 他計式은 費用의 問題와 調査員 偏嗜의 問題가 있긴 하여도 自計式에 比하여 훨씬 優越한 것으로 생각되고 있다.

(8) 調査員

調査員은 広義에 있어서는 調査管理者 現地調査員 및 集計員을 含包한 모든 調査従業者를 意味하나 狹義로는 現地 調査員만을

가르킨다.

여기서는 現地 調査員中에서도 特히 面接調査法이 採用되는 경우의 調査員(面接調査員)을 中心으로 하여 說明하고자 한다.

社会現象을 다루는 모든 調査에 있어서는 被調査者가 얼마나 真実한 情報를 提供하여 주는가 하는 被調査者의 協力的 態度와 調査員의 正確하고 熱誠的인 觀察態度가 가장 重要하다 할 것이다. 被調査者는 各己 性別, 年齡, 学歷, 階層, 思想, 性格 等이 다른 各様의 属性과 個性을 가지고 있어서 그들에게서 한결같은 協力을 얻고 그들은 正確히 觀察한다는 것은 容易한 일이 아니다. 調査員은 이와같은 어려운 일을 担当하는 것이며 調査員의 能力과 努力如何에 따라서 그 調査의 成功与否가 左右된다고 하여도 過言이 아니다.

따라서 優殊한 調査員을 確保하고 充分히 訓練시키는 일이 調査를 成功케 하는 하나의 関鍵이라고 할 수 있는 것이다.

(가) 調査員의 資質과 性格

調査員의 適合한 性格으로서는 責任感, 誠実, 明朗, 順応性, 探究心, 忍耐力, 公正性 等이 要求되며 正確한 理解力과 判断力 그리고 調査에 対한 興味等이 要望된다. 그 外의 條件으로서는 健康하고 普通水準 以上의 教養과 知識이 必要하며 調査에 專念할 수 있는 時間的 여유가 있을 것이 要求된다.

調査員의 年齡은 20~30歲가 適合하며 性別은 一般的으로 男性이 適合하나 경우에 따라서는 女性이 適合할 때도 있다. (例 家

計調査) 또 経験이 없는 調査員 보다는 経験이 있는 調査員이 一般的으로 낫다고 할 수 있겠으나 正規의 訓練을 받지 않고 막연 하게 調査員으로서의 経験만을 가지고 있는 것은 반드시 낫다고 할 수 없으며 오히려 調査에 無誠意하고 内容을 造作할 可能性이 있다.

(나) 調査員의 偏嗜

調査員은 그의 性格과 資質에 따라서 一定한 偏嗜를 갖기 쉬운데 이는 調査員의 偏見, 先入観, 思想, 性格等에 基因하는 것으로 同一한 対象의 回答을 듣고도 各異한 두사람의 調査員에 있어서는 各異한 結果가 나타나는 것이다. 調査員에게는 自己의 立場이나 主観 信念에 따라서 事物을 理解하려는 傾向이 있기 때문이다.

調査員의 이러한 偏奇는 調査結果에 커다란 誤差을 가져오기 쉬우므로 調査員을 採用할 때나 調査員을 指導 訓練할 때 特히 이를 避하고 除去하도록 努力하여야 한다.

(다) 調査員의 訓練

調査員의 適合한 資質은 반드시 生来的인 것은 아니며 各種의 訓練에 依하여서 研磨되고 完成될 수 있는 것이다. 勿論, 生来的인 素質도 重要한 것이므로 좋은 調査員을 얻기 為하여서는 知能, 性格等에 対한 「테스트」에 依하여서 미리 選抜하여야 하겠으나 実際問題로서 調査員 採用은 그다지 慎重히 다루어지지 않고 있으므로 特히 調査員의 指導와 訓練을 철저하게 하지 않으면 안

되는 것이다.

調査員의 訓練의 目的은 調査員에게 適合한 資質을 키워주고 調査에 関한 直接的인 知識을 부여하여 面接技術을 研磨시키는데 있는 것이다.

따라서 実際로 調査員 訓練을 為한 講習을 実査直前의 調査準備過程에서 実施하게 되는데 이때에 調査員에게 調査의 目的과 意義를 充分히 認識시키고 아울러 具体的으로 調査項目의 定義, 記入方法, 其他 実査中에 일어날 것으로 豫想되는 여러가지 事項에 関한 処理方法等을 詳細히 指示하여야 한다.

그리고 이러한 具体的인 指示는 반드시 調査要領書나 調査指示書로서 作成되어 調査員에게 交付되어야 한다.

調査員의 訓練方法으로는 実地로 練習調査를 実施하여 그 結果를 1枚씩 慎重하게 検討하여 調査員이 誤解하고 있는 点, 調査票記入方法의 不充分한 点을 指摘하고 注意를 환기하는 方法이 많이 使用되고 있다.

그런데 調査員에 対한 調査要領의 指示는 調査員을 한 場所에 集合시켜 一時에 하는 것이 좋다. 그것은 調査指示者와 時間 및 場所에 따라 指示内容이 달라질 念慮가 있으며 그로 因하여 調査員의 統一된 知識과 観点을 期할 수 없게 되기 때문이다.

(9) 調査区

調査의 範囲가 狭小한 1個地域에 局限되어 있거나 標本調査

인 경우에는 特別히 調査区를 設定할 必要가 없으나 「센서스」와 같이 全國 또는 道와 같은 広大한 地域을 調査의 範囲로 하는 경우에는 한사람의 調査員으로서는 全地域을 担当할 수 없고 여러 사람의 調査員이 地域을 分担하여야 하므로 이를 為한 調査員의 分担地域을 設定할 必要가 있다.

이와같은 各 調査員의 分担区域을 調査区라고 한다. 調査区를 設定하는 目的은 調査地域内에 있는 調査対象을 하나도 漏落하지 않고 또한 重複됨이 없이 捕捉하려는데 있다. 따라서 調査区를 設定할 때에는 漏落된 地域이 없는가 調査区와 調査区間의 경계가 明確한 가를 特히 念頭에 두어야 한다.

한 調査区에는 한 調査員을 配置하는 것이 原則이므로 調査区의 数는 動員 可能한 調査員의 数에 依하여 定하는 것이 普通이다.

그리고 調査区의 設定方法으로는 行政区域에 依하여 市, 郡 単位 또는 里, 洞単位로 하는 方法과 地勢 및 交通에 따라서 決定하는 方法이 있다. 調査区를 設定할 때는 위의 方法中 어느 方法을 択할 것인가를 決定하고 特히 全國的인 「센서스」에서와 같이 많은 調査区를 設定하여야 할 때에는 可及的 크게 拡大된 地図를 準備하여 이에 依하여 区域을 明確히 劃定하는 것이 좋다.

그리고 調査区가 많을 때에는 数個의 調査区를 指導 및 管理할 指導区를 設定하는 것이 調査管理上 効果가 있다.

(10) 調査時期

調査의 企劃이 完了되고 이에 따른 準備가 갖추어지면 調査를 実施하게 되는 것이므로 調査時期를 特別히 考慮할 必要가 없을듯 하나 事実은 調査의 時期가 企劃過程에서 먼저 定하여지고 이에 맞추어 企劃 및 準備日程이 짜여지는 것이 原則이다. 그것은 調査時期가 調査의 進行과 結果에 적지 않은 影響을 주는 때문이다. 만일 調査의 時期가 調査의 時間的 基準과 너무 멀어진 1年 또는 数年前의 事実을 調査한다면 그 調査結果는 内容이 不正確할 뿐 아니라 結果의 利用価値도 낮아지게 될 것이며 또 調査 時期가 調査対象을 찾아가기 어려운 季節이라면 그 調査는 中途에서 一段 保留를 하지 않을 수 없게 될 것이다.

따라서 調査時期는 調査가 可能하고 容易하며 되도록 調査基準時点으로부터 너무 오랜 時日이 지나지 않은 때를 択하도록 하여야 한다.

具体的으로 말하면 調査가 可能하고 容易하기 위하여서는 먼저 그 以前에 企劃의 準備가 되어 있는가 等을 考慮하여야 한다는 것이다.

調査時期는 調査基準時点에서 너무 멀리 択하는 것도 좋지 않지만 調査基準時点에서 너무 近接시켜도 좋지 않은 경우가 있는 것이다.

例를 들면 企業体의 生産 또는 財務活動調査에 있어서는 적어도 企業体가 帳簿整理를 끝내거나 決算을 完了한 後를 調査時期로 択

하여야 할 것이다. 그리고 調査를 年1回 実施하는 경우에는 季節的인 條件을 勘案하여 暴暑酷寒을 避하여야 하며 또 被調査者가 大部分 바쁜때는 択하지 않는 것이 좋다.

実査期間은 위의 調査時期에 一定한 期間을 設定한 것인데 그 期間의 長短은 調査対象数와 調査内容 및 調査員数에 依하여서 決定된다.

実査期間은 짧을수록 좋을 것이므로 費用의 問題가 있겠으나 可能하다면 有能한 調査員을 많이 動員시켜 可及的 実査期間을 短縮시키도록 하는 것이다.

(11) 実査管理, 集計等 計劃

統計調査의 企劃은 調査準備로 부터 結果公表의 最終 段階까지의 全過程에 걸친 事項을 全部 網羅하여 이루어져야 하므로 実査段階 以後에 이루어질 実査管理, 調査票審査, 集計等에 関하여서도 企劃過程에서 미리 計劃되지 않으면 안된다.

実査管理에 関하여서는 正確하고 迅速한 実査의 運営方法을 細密한 部分까지 計劃하여야 하며 調査票審査에 関하여서는 調査票의 内容検討 및 分類를 効果的으로 할 수 있는 方法을 講究하여야 한다.

그리고 集計에 있어서는 集計方法, 集計場所, 集計員, 集計表様式, 公表様式 等에 関하여서 具体的인 計劃이 이루어져야 한다. 特히 集計計劃은 調査項目의 設定과 調査票設計時에 있어 同時에 考慮되

어야 한다. 그럼으로써 集計가 不可能한 것이나 利用上에 意味가 없는 項目은 除外될 수 있는 것이다.

以上의 実査管理, 調査票審査 및 集計에 関한 具体的 事項은 後述하고자 한다.

(12) 被調査者의 協助

実査는 被調査者를 相対로 하는 것이며 被調査者가 積極的으로 調査에 協助하지 않고서는 調査의 目的을 達成할 수 없는 것이므로 事前에 被調査者의 協助를 얻는 方案을 講究하여 두는 것이 主要한 일이다.

実査는 그것이 어떤 種類의 것이든 他人의 時間과 情報를 侵害하는 것이므로 被調査者의 協助를 얻기 為하여서는 무엇보다 먼저 이와같은 侵害를 正当化시키고 相対方에 認識시키는 것이 緊要하다.

이를 위하여서는 이 調査가 被調査者와 社会를 為하여 実施되며 有益한 結果를 가져올 것이라는 点. 그리고 調査実施機関이 権限있는 機関이거나 公的機関인 点 等을 強調하여야 할 것이다.

또한 被調査者가 調査에 非協助的인 理由의 하나가 秘密漏洩에 対한 憂慮에 있으므로 이러한 憂慮를 払拭시키도록 努力하는 것도 重要하다.

따라서 調査員이 実査에 着手하기 前에 被調査者에게 調査의 目的과 趣旨, 調査協助에 関한 간곡한 부탁, 調査内容은 절대로 秘密로 한다는 内容의 調査協助依頼文을 보내는 것이 좋다. 그리고

豫算이 許容하면 이러한 個別的인 協助依賴文外에 新聞, 放送, 포스타等을 通하여 調査의 重要性을 널리 周知시키는 것도 좋다. 이밖에 調査実施機関이 権限이 없거나 公的機関이 아닌 때에는 다른 有力機関의 支援을 받는 것이 매우 効果가 있다.

被調査者의 協助를 얻는 方法으로서 謝礼品을 주는것도 좋은 方法이다.

謝礼品을 주기 위하여서는 莫大한 費用이 所要되지만 被調査者의 協助와 수고에 報答하는 뜻에서 実査後에 謝礼品을 贈呈하는 것이 効果的이다. 또한 実査를 担当한 調査員에게는 調査員의 身分을 証明하고 調査機関을 明示한 調査員 身分証을 発給하여 実査時에 被調査者에 提示케 할 뿐아니라 服装이나 言語를 단정하게 하여 좋은 印象을 주도록 하는 것도 被調査者의 協助를 얻는데 크게 도움이 될 것이다.

(13) 豫備調査

実査에 들어가기 前에 企劃된 内容들이 現実的으로 妥当한가 하는 것을 確認하기 為하여 豫備調査(準備調査 또는 試験調査라고도 한다)를 하는 것이 普通이다.

먼저 試験調査에서는 主로 調査単位, 分類方法, 調査票様式, 調査項目의 定義, 調査項目의 排列, 調査方法, 調査員의 業務量, 被調査者의 協力程度 等과 같이 調査企劃上에 나타난 諸般 事項에 関하여 그 妥当性을 検討 確認하는데 必要하지만 準備調査는 「센서스」와 같이 大規模의 調査에서 豫備名簿나 標本名簿(Sample

List)를 作成하기 위한 実査에 前提的인 作業인 것이다.

따라서 企劃上의 欠陷을 試験하기 위하여 実施하는 豫備調査를 試験調査라 하며 이는 企劃의 完了後에 하는 것이 아니라 企劃途中 即 実査의 前期作業으로 実施되는 것이 通例이다.

3.2 実際調査 (現地調査)

(1) 実査의 重要性

統計調査의 企劃과 準備가 끝나면 実査의 段階로 들어가게 되는데 이 段階에서는 調査員이 被調査者로부터 情報를 얻고 이를 調査票에 記入하며 調査管理者는 実査를 管理하고 記入이 完了된 調査票를 蒐集하게 된다.

実査의 方法에는 面接調査法, 配票調査法, 集合調査法, 郵便調査法, 電話調査法 等이 있음은 既述한 바 있으나 여기서는 面接調査의 경우를 中心으로 하여 説明하기로 한다.

統計調査가 物理現象을 다루는 경우에 있어서는 대개의 경우 調査対象이 下等動物이건 無生物이건 간에 対象과 論議하는 일이란 있을 수 없는 것이다. 그러나 統計調査가 社会現象을 다루는 경우에는 調査対象인 個人 또는 社会集団이 積極的으로 이에 参与하지 않고서는 不可能한 때가 많다. 다시 말하면 社会現象에 関한 調査가 成功하느냐 失敗하느냐의 重要한 関鍵은 被調査者가 얼마나 積極的으로 協力하여 真実한 情報를 提供하여 주느냐에 달

려 있다.

그런데 被調査者는 各己 다른 性格과 特徵을 가지고 있으므로 이러한 被調査者로부터 한결같이 積極的인 協助를 얻고 正確한 調査를 期한다는 것은 결코 容易한 일이 아니다. 社会現象을 다루는 統計調査에 있어서 実査 技術的인 方法이 크게 重要視되는 것은 이 때문이다.

아무리 統計調査의 企劃이 잘 되었다 하더라도 実査過程에서 無能하고 無誠意한 調査員의 熟練되지 못한 調査는 内容이 不正確한 結果를 가져올뿐만 아니라 被調査者의 非協助的인 態度만 助長시킴으로써 将来의 調査마저도 困難하게 만들 우려가 있는 것이다.

대체로 被調査者는 調査를 忌避하거나 真実한 対答을 거부하려는 傾向이 있다.

이것은 자기의 事実上의 秘密이 外部에 漏洩됨으로써 同業者間의 競争에 不利하여 지거나 納税額에 影響을 주지 않을까 하는 疑懼心과 調査에 応하려면 바쁜때의 貴重한 時間이 浪費된다고 생각하기 때문이다. 그러나 우리가 調査目的을 達成하기 위하여는 이러한 被調査者를 우리가 意図하는 대로 이끌고 가서 直実한 対答을 하도록 만들지 않으면 안되는 것이다.

그러므로 이를 위하여서는 무엇보다도 調査員이 훌륭한 資質과 態度 그리고 充分한 知識을 구비하고 誠意있게 調査에 임하여야 하며 또한 그때그때의 状況에 適応하여 適切한 技術的인 面接方法

을 항상 研究하지 않으면 안되는 것이다.

(2) 実査의 管理

調査員에 依한 実査는 調査가 企劃된대로 円滑히 遂行되도록 엄격히 管理되지 않으면 안된다.

먼저 調査員이 実査에 着手하기 前에 調査管理者는 調査員으로부터 実査計劃書 또는 実査日程表를 作成하도록 하여 이를 検討하고 調節하여야 한다. 그리고 이 実査計劃書에 의거 調査員의 実査를 管理·監督하는 것이다.

調査員이 実査를 進行하고 있는 동안에는 調査員이 調査対象을 틀림없이 訪問하여 正確한 調査를 하도록 이를 통제하여야 한다.

調査員中에는 被調査者를 만나지도 않고 卓上에서 調査票를 造作記入하는 者도 있으며 被調査者를 訪問하지도 않고 또는 但1回 訪問하고서는 被調査者의 不在 혹은 調査不能이라고 報告하는 事例가 있으므로 調査管理者는 이러한 일이 発生하는 것을 최대한 防止하여야 하는 것이다. 調査員의 이러한 行為를 統制하는 方法으로는 調査管理者가 不時에 任意의 調査対象을 訪問하거나 調査対象에게 郵便 또는 電話로 調査員의 訪問与否와 態度를 照会하는 等의 方法이 있으며 이 外에 調査票上에 対照項目을 設置하여 調査員을 統制하는 方法도 있다. 卓上造作은 않더라도 他人에게 調査를 依頼하거나 電話 또는 郵便으로 調査하는 것도 이를 防止하여야 한다. 実査도중 調査不能이 나오는 경우에는 調査管理者는

그 事由를 検討하고 再調査를 指示하거나 標本調査인 경우에는 調査対象인 標本을 代替하여야 하는데 이때 注意할 것은 標本의 代替는 調査員이 任意로 하여서는 결코 안되며 調査管理者에 依하여 엄격히 다루어져야 한다는 것이다.

그리고 이와 別途로 調査管理者는 調査不能의 原因을 면밀히 分析하여야 한다. 調査不能에는 여러가지 原因이 있겠으나 대체로 調査対象이 所在不明일 경우, 調査対象이 不適格한 경우, 被調査者가 不在인 경우, 被調査者가 調査에 不応하는 경우 等이 調査不能한 例이다. 調査対象의 所在가 不明한 것은 当初 名簿가 잘못 作成되었거나 名簿가 作成된 後에 調査対象에 変動이 생긴 때문인데 이러한 경우에는 可能한 限 名簿를 修正하고 調査하여야 한다.

그리고 調査対象이 不適格한 경우, 예컨데 調査対象이 調査範囲 밖에 있는 때와 같은 경우에는 대체로 実査를 할 必要는 없겠으나 調査管理者는 반드시 그 事実与否를 確認하여야 한다.

被調査者가 出他하여 不在인 경우에는 결코 調査不能으로 処理하여서는 안된다. 이 때에는 반드시 재차 訪問하여 調査를 遂行하여야 한다.

그런데 가장 困難한 것은 被調査者가 調査에 応하기를 거부하는 경우이다. 그러나 이 경우에도 調査員은 一次 거절 당하였다 하여 調査를 포기하여서는 안되며 재차 誠意있는 努力을 하여야 한다. 그래도 調査를 不応하는 경우에는 調査管理者가 直接 被調査者를 訪問하여 実査를 하여야 한다. 이 外에 調査管理者는

調査員이 実査途中에 実査에 関한 의문점을 問議하여 왔을때 이에 対答하여 주어야 하며, 調査員의 実査上의 各種 隘路点을 해석하여야 한다. 뿐만 아니라 調査員이 불의의 사고로 調査를 遂行할 수 없게 되는 경우에는 期間内에 実査를 完了할 수 있도록 각별한 対策을 講究하여야 한다.

調査員은 実査가 끝나 調査票를 実査管理者에게 提出케 되는데 이에 앞서 調査員으로 하여금 스스로 調査票를 철저히 検討하고 完全히 整理하여 提出케 하여야 한다. 꼭 調査員으로 하여금 매일 調査가 끝났을때 그날의 調査된 調査票에 対하여서는 記入漏落, 記入의 不完全, 記入上의 錯誤를 스스로 発見하여 是正케 하며 또한 글자를 알기 쉽게 하고 記入法의 統一을 기하여 計算한 것이 있으면 이를 検算하고 경우에 따라서는 計算値의 記入을 行하게 하는 等 調査票를 整理하도록 한다. 그리고 調査員이 割当된 調査対象에 対하여서 調査를 全部 끝냈을 때는 지정된 対象에 対하여 빠짐없이 그리고 錯誤없이 調査가 完了되었는가를 스스로 確認케 한 후에 調査票를 提出하게 하여야 한다.

調査員이 記入完了한 調査票를 提出하였을 때는 되도록 現地에서 調査員을 面前에 두고 調査票를 検査하여야 한다. 여기서 調査票上의 記入漏落, 記入不完全, 記入의 錯誤, 調査員의 造作 等을 発見하여 이를 修正케 하거나 再調査를 指示하여야 한다. 이어서 調査員이 割当된 実査를 모두 終了하였을 때에는 所定期日에 一定한 場所에 調査票를 蒐集하고 여기서 調査対象別로 調査票를 点

検하여 調査漏落이나 重複調査 및 調査対象의 同一性 여부 等을 確認하여야 하는 것이다.

(3) 被調査者와의 面接方法

(가) 面接의 準備

調査員은 被調査者와의 面接을 実施하기전에 먼저 面接調査를 가장 効果的으로 할 수 있도록 計劃을 세워야 한다. 即 어떻게 하면 時間과 距離를 短縮하여 被調査者를 訪問할 수 있는가를 研究하여 路程, 日程表를 짜야 하며 調査에 앞서 調査要領書나 調査指示書를 熟読하고 調査票, 調査要領書, 記入道具, 地図等 持参할 物件을 点検하여야 한다.

다음 누구를 面接対象으로 할 것인가를 定하여야 한다.

面接対象으로 가장 適合한 사람은 말할 것도 없이 調査内容에 가장 精通하고 同時에 調査員에게 自由로이 말할 수 있는 位置에 있는 사람이어야 한다.

事業体 内容을 調査하기 위하여서는 그 事業体의 幹部를 面接하지 않으면 안될 것이다.

面接対象이 定하여지면 調査員은 可能하면 그를 訪問하기 前에 그에 関한 知識, 즉 그 사람의 性格, 過去, 및 그가 속하여 있는 部分 社会에 있어서의 慣行 等 豫備知識을 얻도록 하여야 한다.

이와같은 豫備知識은 面接을 円滑하게 이끌고 被調査者의 積極的인 協助를 얻는데 큰 도움이 된다.

다음 調査員은 被調査者에게서 調査할 問題에 関하여 大略的인

知識을 갖추어야 한다.

그리고 狀況이나 條件에 따라 面接進行의 大綱을 事前에 計劃하여 두는것도 面接의 時間을 節約하고 序頭를 效果있게 만들며 要点을 강조할 수 있다는 点에서 반드시 잊어서는 안될 것이다.

끝으로 面接에 앞서 面接場所와 面接時間을 定하는 것도 重要하다.

面接場所와 時間은 調査員의 便宜만에 依하여서 定할 것이 아니라 被調査者가 가장 安楽하게 이야기 할 수 있는 條件을 考慮하여 定하여야 한다.

一般的으로 事業体의 內容을 調査하는 것이라면 面接場所로서는 그 事業体의 事務室이 가장 適当하다.

그것은 必要한 帳簿와 記錄을 即時 參考할 수 있기 때문이다. 그리고 面接時間은 被調査者가 그다지 나쁘지 않고 별다른 일이 없으면 疲勞를 느끼지 않을 때를 択하는 것이 좋다.

可能하면 事前에 場所와 時間을 被調査者와 約束하는 것이 좋을 것이다.

(나) 面接의 場所

面接은 대개 다음과 같은 順序에 따라 進行되어진다.

第1段階：最初로 被調査者와 接触하는 段階로서 먼저 所持한 証明(身分証)이나 명함을 提示하고 人事를 교환하며 自己의 所屬된 調査機関을 밝힌다.

그리고 분위기를 造成하여 가면서 面接의 趣旨와 目的을 相対가

納得할 수 있도록 說明한다. 이때에 所持한 調査協助依頼文이 있으면 이를 주도록 한다.

第2段階 : 여기서는 두사람의 呼吸을 調整하고 親密하고 부드러운 분위기를 만들어 相対方이 親密感을 갖고 積極的으로 協助하도록 努力한다.

第3段階 : 最初의 接触이 끝나고 또 調査員과 被調査者에 親密한 분위기가 造成되면 被調査者를 主役으로 하여 一般的이고 全体的인 이야기를 교환하게 된다.

이 段階에 있어서 이야기 하는 사람은 주로 被調査者이며 調査員은 主로 相対方의 이야기를 関心깊게 들어야 한다.

例를 들면 相対方으로 하여금 一般的인 景気라든가 業界의 動向을 이야기 시키는 것이다.

第4段階 : 이 段階에 있어서는 調査員이 主役이 되어 本格的인 質問을 한다.

調査員은 대체로 주어진 調査票의 順序에 따라 한 項目 한 項目씩 質問하고 그 対答을 調査票에 記入하는 것이다.

第5段階 : 이 段階에서는 相対方이 特히 하고싶은 이야기나 意見을 듣고 이를 参考하여 調査를 補完한다.

또리고 被調査者가 前段階에서 錯誤한 것을 깨닫고 스스로 訂正하여주는 境遇에는 이를 確認하여 調査票의 記入을 訂正한다.

第6段階 : 이 段階에서는 調査가 完全히 이루어졌는가를 確認한다.

即, 調査員이 調査票上에 記入漏落이나 記入不完全, 記入錯誤가 없는가 調査票를 検討하는 것이다.
그리하여 万若 調査票上의 不備를 発見하면 即席에서 이를 補完하여야 한다.

이 段階에서의 調査票検討는 매우 重要한 일인데도 一般的으로 疎忽히 생각되기 쉬운데 여기서 이를 疎忽히 하면 後에 이를 補完하기란 極히 힘든 일이므로 特히 注意를 要한다.

第7段階 : 面接이 끝났으므로 感謝의 뜻을 表示하고 面接場所를 물러 난다.

感謝는 真心으로 表示하여야 하며 설혹 目的을 達成하지 못한 境遇라도 他人의 時間을 多少나마 消費하게 하였으므로 반드시 感謝의 뜻을 表示하여야 한다. 그리고 이때 今後의 継続調査를 確保할 수 있는 素地를 만들도록 親密하고 좋은 印象을 남기도록 하여야 한다.

以上으로 面接의 順序를 段階的으로 区分하여 説明하였으나 이 順序는 유연성이 있는 것이므로 경우에 따라 相当히 変形되지 않으면 안되는 것이다.

그리고 週期的으로 調査되는 継続調査에 있어서는 第1段階와 第2段階는 省略되는 것이 普通이다.

(다) 面接上의 注意

1) 端正하고 素朴한 외모를 갖추어 被調査者에게 좋은 印象을 줄것.

2) 부드러운 態度와 말씨로 自然스럽고 率直하게 이야기 할 수 있는 溫和한 분위기를 만들어 겸손과 礼節을 잃지 않도록 할 것.

3) 自由롭게 이야기 하면서도 真摯性과 沈着性을 나타내어 相対方이 輕蔑感을 갖지 않도록 하며 相対方의 信頼를 얻도록 할 것.

4) 相対方에게 本心에서 우러나오는 따뜻한 理解와 同情을 가지고 接한다.

5) 相対方의 気分, 言外의 말, 익숙지 못한 表現에서 相対方 真意를 把握할 수 있도록 細心한 洞察力을 発揮할 것.

6) 相対方과 論争하거나 相対方을 輕蔑하는 빛을 보이지 말것.

7) 自己의 態度, 質問, 判断이 正当한가를 反省하고 検討할것. 自己의 判断에 너무 確信을 갖는 것은 훌륭한 調査員이 아니다.

8) 忍耐와 責任感을 가질것. 指定된 被調査者를 찾아서 만나지 못한 때에는 몇번이라도 다시 찾아가야 하며, 때로는 相対方에 依하여 参을 수 없는 일을 当하더라도 忍耐할 것. 그리고 真実하지 못한 일이라고 생각되는 것은 안이하게 処理하지 말고 끝까지 真実을 찾아낼 것. 이것은 任務에 忠実한 責任感에서 생기는 것이다.

9) 끊임없는 関心과 研究로서 充分한 知識을 研磨할 것.

10) 相対方과 面接하지도 않고 調査員이 멋대로 調査票를 造作記入하는 일이 결코 없도록 할것.

3.3 調査票의 審査와 集計

(1) 調査票의 審査

調査票의 審査는 実査에 依하여 蒐集된 調査票의 内容을 審査하여 錯誤나 記人漏落을 発見하고 이를 訂正하여 集計할 수 있도록 調査票의 記人内容을 完全하게 하는 作業이다.

調査票審査作業의 内容을 살펴보면 다음과 같다.

(가) 指定된 調査対象에 関하여 調査票가 確保되었는가를 点検하는 것.

(나) 調査票中에 記入漏落된 項目이 없는가 検討하는 것.

(다) 記入이 不完全한 項目을 検出하는 것.

(라) 判読하기 어려운 文字와 数字를 고쳐 쓰는 것.

(마) 調査票記入法을 統一하는것.

(바) 計算錯誤를 発見하여 訂正하는 것.

(사) 調査票記入錯誤를 発見하는 것.

(아) 調査員의 不正記入을 発見하는 것.

(자) 計算値를 記入하는 것.

以上의 것中(사)의 調査票의 記入錯誤는 調査員 또는 被調査者가 不注意에 起因하는 内容上의 錯誤이므로 이를 調査票審査 過程

에서 発見하기는 어려운 것이다. 이는 調査票의 構成이나 調査員의 選択 또는 指導에서 充分히 注意하여 미리 豫防토록 하여야 하나 調査票 審査過程에 있어서도 이러한 錯誤를 発見하도록 最大의 努力을 기울이지 않으면 안되는 것이다. 이를 為하여서는 다음과 같은 方法을 使用할 수가 있다.

첫째로 個個의 調査票中에서 関聯되는 다른 項目과 対照하여 그間에 矛盾이 発見되면 錯誤가 存在하는 可能性이 있다.

둘째로 調査員마다 簡単한 集計를 하여 다른 調査員의 그것과 比較하여 현저한 差異가 있을 경우에는 어느 調査員인가 不注意를 犯했을지도 모른다.

세째는 全体에 対하여서 簡単한 集計를 하여 豫想과 동떨어진 結果가 나올 경우에는 일단 그 項目의 内容을 의심할 수 있을 것이다. 調査票審査에는 実査管理者에 依한 現地審査와 別途調査者에 依한 事後審査의 段階가 있으며 後者는 다시 調査規模에 따라 地方審査와 中央審査의 段階로 나누어 진다.

다만 郵便調査法을 使用할 때에는 審査者에 依한 審査만이 1段階 行하여질 뿐이다.

実査管理者에 依한 審査는 대개 審査期間中 調査員이 記入 完了한 調査票를 提出하였을 때 調査員의 面前에서 한다. 여기서는 記入漏落, 記入不完全, 記入의 錯誤, 調査員의 不正等의 発見에 注力하고 発見된 境遇에는 再調査를 命한다. 同時에 보기 힘든 글자 記入法의 不統一에 注意한다. 現地審査는 実査期間中 毎日하는 것

이 좋으며 적어도 몇번은 하여야 한다. 그리고 調査員이 割当된 実査를 모두 끝마쳤을 때는 調査漏落이나 重複없이 割当対象의 調査가 完了되었는가를 確認하여야 한다.

実査가 모두 完了되고 調査票가 一定한 場所에 蒐集되어 審査員에 依하여서 一斉히 実施되는 調査票 審査에 있어서는 簡単한 集計에 依하여서 記入上의 錯誤나 調査員의 不正을 発見하여 訂正 或은 調査票의 破棄를 하고 또 計算을 検算하며 計算値를 記入하고 境遇에 따라서는 공백난에 回答을 統計的으로 推定하여 메꾸기도 한다.

以上과 같이 調査票 審査는 実査過程에서 発生하는 各種 誤差를 最少限으로 줄일수 있는 最終的인 機会이므로 이를 철저히 실행하지 않으면 안되며 審査를 担当하는 사람은 調査業務에 経験이 있고 이에 精通한 사람이어야 한다.

(2) 符号化(Coding)

調査票 審査에 뒤따르는 作業은 符号化 作業이다. 이것은 調査票의 各 調査項目에 対하여 記入된 回答을 몇개의 구룹으로 分類하고 그 구룹에 対応한 一定한 符号를 各 調査票에 부쳐서 分類 集計에 便利하게 하는 作業이다. 따라서 符号化에 앞서서 分類基準이 定하여져 있지 않으면 안되는데 이 分類基準은 企劃過程에서 미리 決定되는 것이 普通이지만 調査項目에 따라서 調査結果를 보지 않고는 適当한 分類基準을 定할 수 없는 것도 있으므로 이러한 경우에는 審査가 끝난 調査票를 하나 하나 検

討하여 事後에 適当한 分類基準을 定하여야 하는 것이다.

그리고 分類基準이 定하여지면 이에 対応한 符号를 決定하여야 한다. 그런데 符号는 機械集計를 하는 경우와 手集計를 하는 경우에 따라서 若干 相異하다. 機械集計에 있어서의 符号는 1, 2 . 3 · · · · 9 . 0 . X . Y의 10数字, 2文字의 範囲에서 選択하지 않으면 안된다. 이때 1 . 2 . 3 · · ·으로 적는 数字를 順次로 使用하고 9를 「其他」 X를 「不明」, Y를 「非該」하는 式으로 끝 数字와 文字를 特定한 範疇에 該当시키는 것이 普通이다. 手集計에 있어서는 記入이 簡便하고 判読하기 쉬운 것이면 어떤 数字나 文字를 使用하여도 좋으나 一見하여 그 範疇의 内容이 聯想되는 것이면 더욱 좋다. 예컨대 性別에 있어서 男性을m, 女性을 f로 表示하는 것이다.

또 集計에 있어서 가까이 있는 項目과 錯誤하지 않도록 数字가, 나, 다, 大文字, 少文字의 「알파벳」를 섞어서 使用하는 것도 좋은 方法이다. 符号가 決定되면 具体的으로 어디에 무슨 符号를 부치는 가를 細密하게 指示한 符号集을 作成하여 符号記入者에게 나누어 주어 審査가 끝난 調査票에 매장마다 符号를 부치도록 한다. 符号記入者는 符号集에 따라 各 項目의 符号欄에 色鉛筆(色은 統一하여야 한다)로 符号를 記入한다. 이 作業이 끝나면 다른 符号 記入者가 바꾸어서 符号가 제대로 記入되어 있는가 検査하지 않으면 안된다.

符号化 作業은 項目에 따라서는 매우 쉬운것도 있으나 때로는

매우 어려운 것도 있어 아무리 符号集이 細密하게 되어 있다고 하더라도 具体的인 境遇에 어떤 符号를 부쳐야 할지 困難할 때가 많다. 따라서 符号記入者는 相当한 知識과 熟練을 必要로 하며 注意 깊은 作業을 하여야 하는 것이다. 아무리 完全하게 蒐集된 資料라도 符号化를 잘못하면 그 結果는 쓸모없는 것이 되고 마는 것이다.

그런데 以上은 모두 事後符号化의 경우이나 分類基準과 符号가 調査企劃 過程에서 決定되고 그 基準에 따라 分類(구룹化)하기가 매우 쉬운 境遇에는 符号記入을 調査員에게 시키는 境遇가 많은데 이 경우에는 実査를 始作하기 前에 調査員에 対한 철저한 訓練이 있어야 하며 또 事後에 検査가 따라야 할 것이다.

(3) 集計

調査票 審査와 符号化 作業이 끝나면 모든 調査票를 符号에 따라 分類하여야 한다. 集計란 調査票의 各 項目에 記入된 内容은 計算하는 作業을 말하는 것이다.

集計에는 다음과 같은 種類가 있다.

(가) 単純集計와 相関集計

이것은 만들어지는 統計表의 種類에 依한 分類인데 単純集計란 一般的인 度数分布表를 만드는 集計를 말하며 相関集計는 2個以上의 標識에 関하여 度数를 나타낸 相関表를 만드는 集計를 말한다.

(나) 中央集計와 地方集計

中央集計란 調査票를 中央에 送付케 하여 中央에서 一括集計하는 것을 말하며 地方集計는 調査票를 中央에 送付하지 않고 地方調査機関에서 集計하여 作成된 統計表만을 中央에 送付되어 中央에서 各地로부터 収集된 統計表를 取合하여 最終的인 것을 만드는 方法이다. 地方集計의 唯一한 利点은 集計가 迅速하게 끝난다는데 있으나 機械集計가 発達된 現在에는 地方集計를 하여야 할 理由가 없으며 또 地方集計는 正確性을 欠하기 쉬운 致命的인 결함이 있는 것이다.

(다) 手集計와 機械集計

手集計란 統計機械를 使用하지 않고 사람의 손으로써 集計하는 方法이며 機械集計란 一部의 統計機械를 使用하여 集計하는 方法을 말한다. 手集計는 다시 여러가지 方法이 있는데 大体로 調査票를 그대로 集計作業에 使用하는 方法과 調査票의 内容을 다른 集計用 카아드에 転記하여 集計하는 方法으로 나눌 수 있다.

機械集計는 電子計算組織에 依한 集計인데 調査票 한장(또는 一部)마다 펀치 카아드(punch card)를 만들어 이를 펀치에 依하여 分類하고 다시 計算하는 것이다. 手集計와 다른 点은 펀치 카아드의 천공부터 始作하여 카아드의 分類, 카아드의 計算에 걸쳐 一切의 機械에 依하여 이루어지며 따라서 操作이 매우 迅速, 正確하게 行하여진다는 것이다. 機械集計와 手集計와의 優劣은 한마디로 말할 수 없으나 手集計는,

(a) 集計가 単純할 때

(b) 調査対象이 적을 때

(c) 分類가 極히 細分되어 分類된 各 구룹의 調査票 枚数가 적을 때 等에 有利하고,

機械集計는,

(a) 調査対象이 많을 때

(b) 集計가 比較的 複雑할 때

(c) 網羅的인 集計를 할 때

等에 有利하다.

3.4 統計表의 作成과 分析

(1) 統計表의 種類

調査票의 分類 集計가 끝나면 마지막으로 統計表를 作成하여야 한다. 統計表는 一連의 統計調査 業務의 最終 生産物로서 統計의 作成과 利用의 매개물이 되는 것이므로 統計資料의 観察比較, 解釈이 容易하도록 調査에서 収集된 資料(統計数字)를 体系있게 分類하고 簡潔하게 行과 列(欄)로 排列, 整理하여서 다른 説明을 듣지 않더라도 그 表만 보고 調査結果를 把握할 수 있도록 그 構成을 考慮하여야 한다.

統計表의 構成은 統計者의 優劣을 左右하기도 하지만 또한 集計諸表의 全過程을 규제하게 되는 것이므로 企劃過程에서 가장 먼저 考慮되고 決定되어야 하는 것이다. 그런데 統計表는 그 表示方法

形式内容(性質), 統計数字의 加工여부 等에 따라 다음과 같이 分類할 수가 있다.

(가) 文中統計表와 正式統計表

이것은 統計를 表示하는 方法에 依한 分類인데 文中統計表는 文章中 表示한 統計表로서 表題가 없고 文章과 直接 関係되어 있어 文章을 읽지 않으면 統計의 意味를 알 수 없는 統計表이다. 即, 独立性이 없는 統計表이다.

이에 対해서 正式統計表는 表題, 表頭, 表例等 表의 体制를 完備하여 다른 説明을 듣지 않더라도 그 表만 보고 内容을 알 수 있고 여러가지 分析이 可能하도록 作成된 統計表이다. 普通 統計表라고 할 때는 이 正式統計表를 말하는 것이다.

(나) 単純分類表와 相関表

이것은 形式에 依한 分類로서 但 한가지의 分類에 依하여 作成된 統計表를 単純分類表라 하며 数種의 分類를 結合하여 만든 統計表를 相関表라고 한다.

(다) 構造統計表와 系列統計表

이것은 統計表의 内容에 依한 分類로서 構造統計表는 時間과 場所를 一定하게 하고 其他 標識(例 産業分類, 規模, 性別等)에 依하여서 分類하여 놓은 統計인데 이를 普通 度数分布表라고도 한다. 例를 들면 業種別, 規模別, 生産額 또는 性別, 職業別, 従業員数 等의 統計表가 이에 속한다. 이에 対하여 系列統計表는 同種의 統計数字를 時間的 또는 場所的으로 排列한 統計表로서 예컨대 年度別 生産額 또는 道別 従事員数 等의 統計表가 이에 속

한다.

이 中 特히 時系列表는 어떤 現象의 時間的인 変遷을 表示하기 為한 것으로서 生産指数, 物価指数 等과 같이 指数化 하여 表示되는 경우가 많다.

(라) 非加工 統計表와 加工統計表

이것은 統計数字의 加工 与否에 依한 分類로 非加工 統計表는 統計調査 結果 蒐集된 資料의 総和値(合計値)를 아무런 加工없이 그대로 表示한 統計이며 加工統計表는 非加工統計表에 나타난 統計数字(総和値)를 各種 比率, 平均 等으로 加工하여 表示한 統計表이다. 普通 統計調査 過程은 非加工統計表를 作成 公表하는 것으로 一段落되지만 利用者를 為하여서는 보다 一目瞭然하게 알 수 있는 加工統計表를 作成하여 表示하여 주는 것이 親切하다.

例를 들면 各種 指数나 國民所得 等이 加工統計表에 속한다.

(2) 統計表의 構成

統計表(正式統計表)는 実質的 要素가 되는 統計数字를 除外하면 다음과 같은 8個의 部分으로 構成된다.

(가) 表 題 (heading)

(나) 頭 註 (headnote)

(다) 表 頭 (box head)

(라) 表 側 (stub)

(마) 欄 (column)

(바) 行 (Line)

(사) 表 体 (field or body)

(아) 脚 註 (food note)

正式統計表의 構造

表 題 (頭 註)

	表 頭			
表側	-	-	-	-
	-	-	-	-
	-	表 -	体 -	-
	-	-	-	-

脚 註 .

1) 表題는 다시 番号表와 表名으로 成立된다. 表番号는 統計表 表의 番号인데 이것은 한 統計表의 系列中에 있어서의 그 統計表의 関係位置를 表示하는 것이며 또 統計表의 索引도 되는 것이다.

따라서 単独의 統計表에는 番号表를 붙일 必要가 없다. 表名은 表頭의 中心을 이루는 것으로서 統計表 内容의 目録이다. 表名에는 統計의 表示範囲, 分類事項, 地域的範囲, 基準時点 等이 表示되어야 하며 이 네개의 事項이 総合되지 않으면 完全한 目録의 役割을 할 수 없는 것이다.

2) 頭註는 表題와 統計表의 最上位線 사이에 表示되는 注

意事項이나 頭註는 表名을 補充하고 統計表 全体를 理解하는데 必要한 事項으로서 表名과 密接한 関係를 갖는 것이다. 따라서 普通 頭註로서 表示되는 것은 그 統計表의 数字 全体에 関한 単位라든가 統計의 基本時点 또는 統計対象 範囲와 같은 것이다. 頭註가 統計表 全体에 関한 注意事項인데 対하여, 脚註는 表中의 特定한 統計数字 또는 表頭 表側의 意味를 明白히 하거나 資料의 出処를 밝히는데 使用되는 것이다.

脚註는 統計表의 最下位線 아래에 記入한다. 그리고 頭註를 記入할 때는 「註」, 「備考」等을 表示할 必要가 없지만 脚註는 반드시 「註」라고 表示한 다음에 必要事項을 記載하여야 하며 또한 説明을 必要로 하는 数字의 앞(또는 뒤)에는 合符号를 붙여서 註와 連結시켜야 한다.

3) 表頭와 表側은 形式的으로 統計表의 모양을 만들고 実質的으로는 統計系列을 만들어 数字間의 関係를 明白하게 하는 것이다. 表頭는 統計表의 제일 위쪽에 位置하여 各 分類事項을 表示하며 表側은 統計表의 左側에 位置하여 이 亦是 分類事項을 表頭와 表側의 어느쪽에 表示하는 것이 数字간의 関係를 보다 明確하게 나타내는가 그리고 地面을 浪費하지 않고 合理的으로 利用할 수 있는가 하는 것이다.

4) 表体는 統計表中 統計定数가 記入되는 場所를 총칭하는 것으로서 이것은 다시 欄과 行으로 成立된다. 欄은 統計数字가 縦으로 排列되는 것이며 行은 統計数字가 横으로 排列된 것을 말한다.

(3) 統計表 作成上의 注意

統計表는 社会的으로 널리 利用되므로서 비로서 그 価値가 있는 것이므로 먼저 利用者에게 쉽고 便利하도록 作成되어야 한다. 即, 統計表는 統計利用者가 그 統計表를 보고 理解할 수가 없거나 誤解 또는 錯覚을 이르키도록 複雜하게 만들어져서는 안되며 되도록 간결하고 明確하게 作成되어야 한다. 그리고 보다 利用者에게 親切하기 위하여서는 比率, 指数, 平均等 加工統計를 아울러 表示하여 주거나 統計図表化하여 나타내 주는 것도 좋다.

統計表 作成에 있어서 또 한가지 注意하여야 할 事項은 統計表에는 어떤 個人의 秘密이 露出되어서는 안된다는 것이다. 따라서 分類가 細分되므로써 어떤 個人에 関한 事実이 露出될 우려가 있는 것은 技術的으로 적절히 処理하여 個人의 秘密이 나타나지 않도록 하여야 한다.

그리고 特히 誤解하기 쉬운 事項이 있거나 例外的인 事項이 있는 경우에는 반드시 그 内容을 朱書하여야 한다.

統計表의 一般的인 記載方法은 다음과 같다.

(a) 該当事項이 없을때는 「－」로 表示한다.

(b) 未詳인 때는 「・・・」로 表示한다.

(c) 単位未満인 때 即, 表示単位가 千원인데 集計된 金額이 100원 또는 300원이 되었을 때는 「0」으로 表示한다.

끝으로 調査結果를 発表할 때에는 調査目的, 範囲, 調査単位, 調査項目의 定義 調査方法, 標本調査일 때는 標本数, 標本抽出方法等

統計表를 利用하는 사람들이 參考하여야할 事項에 対한 說明書를 統計表의 앞에 붙여야 한다.

(4) 統計表의 分析

統計表가 作成되면 마지막으로 統計表의 分析이 남는다. 統計表의 分析은 우선 統計表를 알기쉽게 꾸미고 說明을 加하여 어떤 結論을 끌어내는 것이다. 먼저 統計表를 알기쉽게 꾸미는데는 統計表의 数字를 加工하여 百分率이나 比率로 나타내는 方法이 있다. 그러나 百分率이나 比率은 아무 数字에 対하여서나 마구 算出하여서는 안된다.

統計表를 分析할 때에는 다음 事項을 檢討하여야 한다.

(가) 가장 重要한 点은 무엇인가

(나) 그밖에 무엇이 表示되어 있는가

(다) 平均은 豫期한 것 또는 다른 資料에 比하여 어떠한가

(라) 各 範疇의 最大数나 最小数는 무엇을 意味하는가 다른것과 比較하여 그 百分率은 무엇을 말하는가

(마) 一般的인 傾向이 있는가 없는가 그것은 어째서 그러한가 例外的인 것은 무엇을 意味하는가

(바) 因果的인 関係가 나타나는가

(사) 이미 알려져 있는 것이나 다른 統計表에 나타나 있는 것과 比較하여 一貫性이 있는가.

(아) 이러한 結果가 나온 것은 標本抽出의 方法이나 調査方法 때문이 아니겠는가

(사) 各 部分의 数字가 다시 検討하기 위하여 調査票를 찾아 보아야 하거나 再集計를 하는 편이 낫지 않을까.

以上과 같이 検討하면서 分析하여 가면 重要한 類似性이나 差異, 連続関係, 因果関係가 発見되어 数字의 限界도 明白하게 될 것이다.

그리고 이와같은 検討와 分析에 이어서 다시 全体的으로 検討하고 分析하지 않으면 안되는데 이때 重要한 것은 이러한 分析이 先入見에 이끌리어 不当하게 強調되거나 軽視되는 일이 있어서는 안된다. 一般的인 傾向과 例外的인 境遇와를 明確히 区別하고 이에 올바른 評価를 加하여야 하는 것이다. 因果関係를 推論하는 경우에도 単純히 数字上의 相関関係가 있는 것만으로 立論하여서는 안되며 넓은 理論的인 考察과 精密한 比較分析이 推論의 前提가 되어야만 한다.

第 4 章　統計調査의　実例 (經濟活動人口調査)

經済活動人口調査는　1957 年부터　1962 年　5 月까지　地方　行政機関을　通하여　労動力調査를　実施하여　毎月　就業者와　失業者의　資料를　蒐集하였다.

그　後　経済開発　5 個年計劃의　樹立과　그　遂行을　為하여　経済活動에　参与하는　人口의　正確한　資料가　切実히　要請되어　経済企劃院　調査統計局에서　過去에　実施한　労動力調査의　諸缺点을　시정하고　보다　正確한　資料를　作成할　目的으로　統計法　第 2 條에　依拠하여　経済活動人口調査를　指定統計　第 4 号로　定하고　1960 年　人口센서스에　根拠를　둔　새로운　標本設計와　専門化된　調査員에　依하여　1962 年　8 月부터　実施하여　왔다.

그러나　時間이　흐름에　따라　調査区内의　特性値가　変하여　1969 年　6 月부터는　1966 年　人口센서스의　調査区에　根拠를　둔　標本에　依하여　再設計　되었고　다시　1970 年　人口　및　住宅센서스가　実施됨에　따라　1972 年　3 月부터는　이에　依한　새로운　調査地域에서　実施하게　되었다.

다음에　経済活動人口調査의　概要, 調査節次, 調査票作成, 調査結果에　対하여　前述한　調査方法에　依拠　紹介함으로써　理解를　돕고자　한다.

4. 1 調査概要

(1) 調査目的

経済人口 및 諸社会与件의 変動에 따르는 国民의 経済活動의 変化를 適期에 正確히 把握하여 雇傭 및 失業의 構造와 変動推移를 分析하고 이에 対한 政策을 樹立하는데, 基礎資料를 提供하는데 그 目的이 있다.

(2) 調査範囲

調査期間을 基準으로 하여 大韓民国 国籍을 가진 常住人口의 1／500에 該当하는 人口中 満14歳 以上者는 모두 調査対象者로 하나 다음 事項에 該当되는 者는 除外한다.

(가) 現役軍人

(나) 刑이 確定된 교도소 수감자

(다) 外国人

常住者라 함은 調査期間을 基準으로 하여 同一한 場所에서 3個月 以上 居住하였거나 居住하려는 期間이 3個月에 達하는 者를 말한다.

但, 航空, 船舶의 船員은 長期 出他中일지라도 調査한다.

(3) 調査項目

本 調査의 調査票는 18個 調査項目으로 되어 있는데 姓名等 6個項目의 人的事項과 1週間의 主된 活動状態等 12個項目의 経済活動事項을 調査한다.

(4) 調査方法

(가) 標本調査

1970年 人口 및 住宅센서스 調査区中 普通調査区에서 1/500에 該当되는 市部 72個 調査区와 郡部 93個 調査区等 総 165個 調査区를 抽出하여 同 調査区内의 全家口 約 12,000 標本家口를 調査하는 標本調査이며 標本抽出에서 除外된 調査区는 落島 閑地調査区, 寄宿舍調査区, 特殊社会施設調査区, 特別調査区等이며 除外된 調査区의 센서스 人口는 約 75萬名에 不過하다.

(나) 週間 및 家口単位調査

経済活動人口調査는 1週間의 調査期間中에 일어나는 経済活動状態를 把握하는 方法으로 家口를 標単位로 調査하고 있다.

調査期間을 1週間으로 定한 理由는 1週間이란 質問時 応答者의 記憶誤差를 적게 하여 正確한 応答을 바랄 수 있고 経常状態의 変動을 充分히 避할 수 있기 때문이다.

(다) 面接調査方法과 配票調査 方法의 併用

経済活動人口調査票는 対象家口를 訪問하여 応答者로부터 얻어지는 정보를 調査員이 記入하는 他計式方法을 使用하며, 就業時間記入票를 応答者 自身이 記入하는 自計式方法을 쓰고 있다.

(5) 調査時期

季節別로 1年에 4번 実施하며 調査対象期間(調査週間)은 指定된 달(3月, 6月, 9月, 12月)의 指定된 1週間으로 하여

実際 調査期間 (実査期間) 은 調査週間 다음의 1 週間으로 한다.

1972 年의 調査時期는 다음과 같이 計劃하고 있다.

	準備調査	調査週間	実地調査
1 / 4 分期 (3 月)	5 日 ~ 11 日	12 日 ~ 18 日	19 日 ~ 25 日
2 / 4 分期 (6 月)	4 日 ~ 10 日	11 日 ~ 17 日	18 日 ~ 24 日
3 / 4 分期 (9 月)	3 日 ~ 9 日	10 日 ~ 16 日	17 日 ~ 23 日
4 / 4 分期 (12 月)	3 日 ~ 9 日	10 日 ~ 16 日	17 日 ~ 23 日

(6) 主要用語定義

(가) 家 口

居住와 家計를 같이 하는 사람의 모임을 家口라 하며 한 사람이라도 別途로 居住하고 独立的인 家計를 이룩하고 있는 境遇에는 하나의 家口로 간주한다.

1 家庭婦, 其他使用人, 店員, 同居人等은 主家口의 家口員으로 包含시킨다.

2 寄房인 同居人으로서 主家口와는 家計를 別途로 하는 境遇와 食事는 같이 하고 있지마는 主家口에게 房賃라든가 食費를 支払하여 家計를 別途로 하는 사람은 主家口의 家口員에 包含시키지 않고 別個의 家口로 한다.

그러나 친척이라든가 主人의 子弟를 돌보아 주고 実費程度의 食費, 房賃를 받고 있는 同居人은 主家口의 家口員으로 한다.

3 学校나 工場等의 寄宿舍等에서 살고 있는 独身者는 各各 単一 家口로 하였으나 이 때 全員이 한 家族처럼 살고 있으면 그 家口一員을 하나의 家口로 한다.

(나) 農 家

農家라 함은 農業을 生業(生計와 営利)으로 하는 家口로서 다음 各 号의 1에 該当하는 家口를 말한다.

1 所有如何를 不問하고 논, 밭, 水源池等의 総面積이 300 坪 以上을 直接 耕作하는 境遇

2 大家畜(소 또는 말)을 1 마리 以上 飼育하는 境遇 (但 運搬用은 除外)

3 中家畜(돼지, 염소, 羊等)을 도합 3 마리 以上 飼育하는 境遇

4 小家畜(토끼等)을 40 마리 또는 닭 30 마리 以上을 飼育하는 境遇

5 高等園芸, 菜蔬, 特用作物等 100 坪 또는 果樹苗圃等을 200 坪 以上 耕作하는 境遇

6 5 郡 以上의 養蜂을 하는 境遇

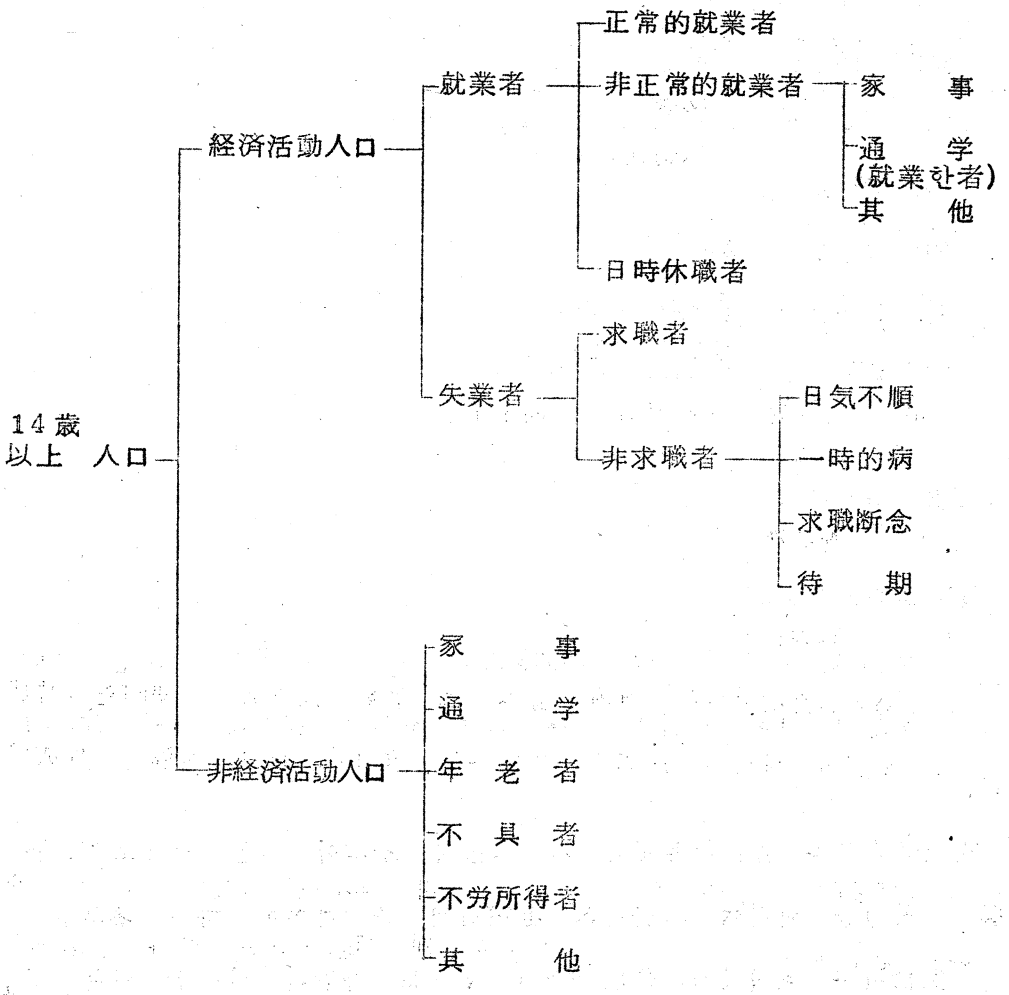
7 蚕種 10 g 以上을 소잠하는 境遇

農業과 他産業을 함께 経営하는 境遇에는 農業収入이 50 % 以上인 境遇에만 農家에 包含시킨다. (但 農業収入面으로 区別하기 困難할 때에는 労動力 投入量으로 区別한다)

(다) 経済活動人口

財貨나 用役을 生産하기 為하여 労動을 提供하는 모든 사람과 提供할 能力이 있는 사람을 말한다.

여기에는 다음과 같은 種類로 分類된다.



(7) 調査結果

経済活動人口調査의 結果表는 다음 9個表로 되어 있으며

經済活動人口年報에 公表한다.

(가) 経済活動人口年報

(나) 年齡階級別 経済活動人口

(다) 非経済活動人口

(라) 年齡階級別 就業者

(마) 産業 및 年齡階級別 就業者

(바) 産業別 就業者

(사) 職業別 就業者

(아) 従事上의 地位別 就業者

(자) 産業 및 就業時間別 就業者

4 . 2 調査節次

(1) 準備調査期間

準備調査期間에는 먼저 標本으로 뽑힌 調査区의 要図를 作成하고 要図에 나타난 居処에 居住하는 家口의 家口員名簿를 作成한다.

調査区要図와 家口員名簿는 本 調査의 基本이 되는 것으로서 要図는 정성을 다하여 깨끗하고 正確하게 作成하여야 하며 名簿는 每 家口마다 빠짐없이 作成하여야 함으로 사람이 살 수 있는 곳은 全部 確認한다.

調査区要図의 作成要領은 먼저 地形 地物의 表示를 補充記入하면

서 境界線을 確認 作成하고 調査区 境界線과 境界線 안쪽에 있는 道路와 建物(또는 居処)을 빠짐없이 口로 그려 넣고 主要建物의 이름이나 地名을 써넣어 位置를 明確히 한 다음 居処의 一連番号를 口안에 調査하기 便利한 順序로 記入한다.

調査区要図와 家口員名簿가 作成되면 調査員은 調査区를 수시로 순회하며 新築建物(居処)과 建物의 撤去等 地形 또는 建物의 変動을 把握하여 要図를 補完 修正하여야 하며 또한 家口員名簿作成後 每 訪問時마다 家口 및 家口員의 変動事項을 把握하여 要図 및 家口員 変動報告書에 記入하여 다음 実査에 漏落이나 重複이 없도록 事前 準備를 하여야 한다.

또한 準備調査期間에 就業時間 記入表를 配付하여 被調査者가 経済活動状況을 事前 記入토록 하여 調査上 便宜와 正確을 期할 수 있도록 한다.

(2) 調査週間

調査週間은 調査対象이 되는 1週間으로서 準備調査期間에 配付한 就業時間記入表를 正確하게 作成하고 있는가의 与否를 確認하고 萬若 記入上의 애로가 있을 境遇 이를 指導하여 正確하게 作成되도록 한다.

또한 就業時間記入 確認中 調査区要図와 家口員의 変動이 있을 境遇 継続하여 補完하고 이를 変動報告書에 記入하여 中央에 報告토록 한다.

(3) 実査期間

実査期間은 調査週間 다음 1週間으로서 各 家口의 満 14歳 以上 人口인 各 사람과 直接面談하여 調査票를 記入하되 準備調査時 配付하여 被調査者가 作成한 就業時間記入票를 比較하여 作成토록 한다.

実査期間에도 調査区의 要図와 家口員名簿를 補完하여 調査의 漏落을 防止하고 次期調査를 便利하게 한다.

調査票의 項目別 作成要領은 다음에 説明토록 한다.

(4) 整理期間

整理期間은 実査가 完了된 後 調査의 正確性을 把握하는 期間으로서 우선 変動報告書의 記入内容을 確認하여 調査의 重複이나 漏落을 検討하고 各 調査票의 内容을 検査하여 調査票作成이 잘 되었는가 살핀 다음 各 調査区別 綜合表를 作成토록 한다.

4.3 調査票作成

(1) 調査員의 態度

(가) 調査員의 身分을 確認할 수 있는 身分証을 반드시 提示한다.

(나) 各 家口의 満 14歳 以上 人口인 各 사람과 直接 面談하여 調査하는 것이 原則이다.

(다) 親切하게 自己紹介를 한 後에 経済活動人口調査의 目的을

說明하고 正確한 答辯을 하여 줄 것을 付託한다.

(라) 調査員이 家口를 訪問하였으나 사람이 없어서 調査를 할 수가 없을 境遇에는 家口員名簿에 表示를 하였다가 再訪問을 하도록 한다.

(마) 調査票記入을 할 때에는 같은 事項 例를 들면 職業이나 産業이 같을 境遇에 「上同」 또는 「〃」으로 記入하지 말고 反復하여 記入하되 該当事項이 없는 난에는 「＼」을 긋는다.

(2) 調査票의 記入要領

經濟活動人口調査는 經濟活動人口를 対象으로 하고 있으므로 家口員中 14歲 以上인 者에 対하여서만 調査토록 한다.

調査項目은 이름等 6種의 人的事項과 活動狀態等 12種의 經濟活動事項이며 第7欄의 「지난 1週日間에 主로 무엇을 하셨읍니까?」에서는 지난 1週間에 主로 한 行為란 平常時에 했던 活動과는 関係없이 指定된 週間 사이에 活動한 行為를 말하는데 9個의 項目中 「1. 일하였음」에 「○」表한 사람은 8欄, 9欄, 10欄은 「＼」을 긋고 11欄부터 質問하며 其外項目의 境遇 8～10欄을 作成토록 한다.

다음에 調査票의 項目을 묻는 方法의 次例를 図表로 나타냈으며 經濟活動人口調査票 作成 例를 添附하여 參考토록 하였다.

4 . 4 調査方法決定의 理論的 基礎

労動力調査方法은 資料蒐集上의 接近法과 概念上의 接近法이 있다.

(1) 資料蒐集上의 接近法

資料蒐集上의 接近法에는 事業体를 調査対象으로 그 事業体에 雇傭되어 있는 被雇傭者와 그와 関聯된 資料를 蒐集하는 事業体 接近法과 家口를 調査対象으로 하여 就業者, 失業者 및 其他 人口学的 特性値를 家口内의 家口員과 関聯하여 資料를 蒐集하는 家口接近法이 있다.

事業体接近法은 被雇傭人의 数를 비롯하여 給与, 賃金, 就業할 計劃, 時間 및 日数, 実際 就業한 時間等 有用한 資料를 同時에 얻을 수 있고 얻어진 資料는 雇傭指数, 賃金(日, 週, 月別)指数, 産業間의 移職率은 얻을 수 있으며 産業, 職業, 性別 就業者의 構成은 調査期間中 就業模型의 変化를 測定할 수 있는 長点이 있으나 調査期間中 한 事業体 以上에 일한 者는 모두 그곳에서 把握되므로 過大로 調査되기 쉬우며 오직 被雇傭者에 限하여 調査되므로 調査의 範囲가 限定되어 있어 雇傭主, 自営業主, 家族従事者, 失業者, 非経済活動人口等은 把握할 수 없는 短点이 있다.

家口接近法은 就業者, 失業者, 非経済活動人口의 特性을 모두 把握할 수 있으며 이를 特性値와 結付하여 家口의 社会的, 経済的, 人口学的 背景을 同時에 研究할 수 있는 長点이 있으며 正確한

給与額이나 賃金率을 算出할 수 없으며 生産量이나 生産性을 把握할 수 없는 短点이 있다.

(2) 概念上의 接近法

調査의 概念上 接近法에는 平常状態에서 주어진 役割이나 機能에 依하여 労動力人口를 調査를 하는 有業者接近法과 어떤 期間內에 実際로 活動한 状態에 依하여 労動力人口를 調査하는 労動力接近法이 있다.

有業者接近法은 職業이 있느냐 없느냐로 就業者와 그 밖에 것으로 区分하기 쉽고 調査員訓練이 容易하며 平常状態에 依하여 얻어진 資料이기 때문에 季節的이나 우발적인 活動에 別로 影響을 입지 않는 長点이 있고 한가지 職業이나 確固不同한 職業이 없는 者들을 定義하기 어려운 同時에 얻어진 結果는 特定한 期間이 없기 때문에 Bench Mark로 使用하기에 不適当하고 처음 求職活動을 한 者를 失業者로 捕捉하기 힘들며 또한 経済的 経済動向을 捕捉할 수 없는 短点이 있다.

労動力接近方法은 짧은 期間동안 일어난 活動을 묻기 때문에 記憶誤謬를 적게 하고 Bench Mark로 提供될 수 있으며 経常状態는 平常状態보다 客観的이고 正確性을 期할 수 있고 失業者, 非経済活動人口를 쉽게 区別할 수 있을뿐 아니라 経常的 経済動向을 빨리 捕捉할 수 있는 長点이 있고 反面에 짧은 期間에 일어난 活動状態를 調査하므로 그 期間中에 우발사건이나 気候変動이 일어났을

때 그 影響이 크게 미치게 되어 継続的으로 調査하여야 한다는 短点을 가지고 있다.

(3) 以上의 두 方法에 対하여 UN에서도 勧告하고 있으나 어느 方法을 択하여야 할지는 調査의 性格이나 그 나라의 経済的, 社会的 与件에 依하여 左右된다.

우리나라의 経済活動人口調査는 就業者, 失業者, 非経済活動人口의 構成的 特性을 모두 把握할 수 있는 家口接近法과 客観的이고 正確性을 期할 수 있는 労動力接近方法에 依拠 実施되고 있다.