標本調査에 있어서의 單純任意抽出法

咸　鍾　郁

1. 序　言

요즘 統計는 모든 分野에서 利用되고 있다. 社會·經濟政策에서는 勿論 農業·醫學 등 利用範圍는 그 어느 것 보다도 넓다고 하겠다.

統計를 作成할 때에는 調査의 目的과 이에 따른 調査方法을 생각하게 된다. 調査方法은 여러 가지로 區分할 수 있지만 特히 調査對象을 全部로 하느냐 一部로 하느냐에 따라 全數調査와 標本調査로 나눌 수 있다.

그런데 全數調査는 調査의 規模에 따라 時間·勞力·費用의 過大로 Bench-mark로 利用되는 人口·農業·産業 등의 센서스에서 一般的으로 利用되며 社會構造의 動態把握이나 學問研究의 基礎資料의 蒐集에는 위이 센서스를 母集團으로 하는 標本調査에 依하여 資料를 얻게 된다.

標本調査를 實施하자면 먼저 標本抽出의 方法을 擇하는 것이 重要하다. 標本抽出方法에는 主觀的 方法에 依한 有意抽出法과 確率的 方法에 의한 任意抽出法이 있는데 後者의 方法이 偏倚性(bias)을 갖지 않는 推定이 可能하고, 母集團에 對한 그 統計量의 分布型이 一定하므로 흔히 利用되는 方法이다.

任意抽出法에는 母集團에서 直接 一定數의 標本을 推出하는 單純抽出法, 母集團을 同質的으로 區分하여 抽出하는 層別抽出法, 母集團의 層間을 同質的으로 하여 抽出하는 集落抽出法, 標本을 2回以上 異時的으로 抽出하는 多回抽出法이 있는데 本稿에서는 任意抽出法中 가장 基礎的이고 簡單한 單純抽出法에 대한 基礎的인 理論을 살펴보고 具體的인 抽出方法을 約述하기로 한다.

2. 可能한 標本의 組合數

確率標本을 理解하기 위하여 가장 簡單한 경우를 들어 母集團과 標本의 關係를 밝혀 보자. 一般的으로 調査項目은 多數이고 그 種類도 屬性·變量의 두 가지가 있다. 그러나 簡單하게 하기 위하여 調査項目이 하나일 때를 생각하여 보자.

母集團은 N個(이것을 母集團의 크기라고 한다)의 單位로 構成되고 이 중에서 n個(이것을 標本의 크기라고 한다)의 單位를 標本으로 抽出한다고 하자. 이 때 變量값을 X라고 한다면 母集團은 N個의 X값으로 되어 있다고 생각할 수 있다.

즉 母集團; $X_1, X_2, X_3, \cdots\cdots, X_N$

標　本; $x_1, x_2, x_3, \cdots\cdots, x_n$

로 構成되며 이 때 x_i는 $X_1, X_2, \cdots, X_N$ 중 어느 한 값이 될 것이며 母集團과 標本의 平均値와 分散은 各各 다음과 같이 定義된다.

母 平 均; $\mu=\frac{1}{N}\sum_{i=1}^{N}X_i$

母 分 散; $\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2$

또는 $S^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\mu)^2$

標本平均; $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

標本分散; $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

이제 크기 N인 母集團에서 크기 n인 標本을 單純任意抽出하면 몇組의 標本이 있을 수 있을 것인가? 組合論에 의하면 ${}_NC_n$組가 있게 된다. 여기서 한 가지 注意하여야 할 것은 이 ${}_NC_n$組라고 하는 것은 N個 중에서 어느 하나를 抽出하고 이 抽出된 것은 除하고 나머지 N−1個 중에서 다시 하나를 抽出하는 것과 같이 할 때의 可能한 方法의 數이다. 다른 方法으로는 N個의

單位 中에서 하나를 抽出하고 다음에 이를 除去하지 않고 다시 原狀으로 되돌려 놓고 再次 N個의 單位中에서 하나를 抽出하는 것과 같은 方法을 n번 反復하여 크기 n의 標本을 얻는 方法이 있다. 이와 같은 方法을 無別限單純任意抽出法 또는 反復을 許容한 單純任意抽出法이라고 하는데 이 때에는 可能한 標本의 組合數는 N^n組이다. 그러므로 前者의 方法을 非復元抽出法이라 하고 後者의 方法을 復元抽出法이라 한다.

3. 標本平均과 標本分散의 期待値

크기 N인 母集團에 크기 n인 標本을 單純任意抽出한다면 ${}_NC_n$組(復元抽出의 경우 N^n組)의 可能한 標本을 얻을 수 있는데 이들 各組의 標本平均과 分散을 求하여 다시 平均하면 各各 母平均과 母分散에 一致한다. 이 때 이 平均値를 期待値라 하며

$$E(\bar{x})=\frac{1}{{}_NC_n}\sum_{i=1}^{{}_NC_n}\bar{x}_i=\mu \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

$$E(s^2)=\frac{1}{{}_NC_n}\sum_{i=1}^{{}_NC_n}s_i^2=S^2=\sigma^2 \cdots\cdots\cdots\cdots (2)$$

로 表示할 수 있다. 또한 이들 標本平均에서부터 標本平均의 期待値 즉 母平均까지의 偏差의 제곱의 期待値(이것은 標本平均의 分散을 뜻한다)를 求하면 다음과 같다.

$$\sigma^2=E\{\bar{x}-E(\bar{x})\}=\frac{1}{{}_NC_n}\sum_{i=1}^{{}_NC_n}(\bar{x}-\sigma^2)$$

$$=\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\left(\text{또는 }\frac{N-n}{N}\cdot\frac{s^2}{n}\right)\cdots (3)$$

以上에서 證明하여 본 몇 가지 事實들은 아주 有用하게 使用되는 것이다.

標本을 抽出하는 것은 標本에서 얻은 값을 가지고 母集團의 값을 推定하고자 하는 것이다. 즉 標本平均 $\bar{x}$는 母平均 μ의 推定値이다. 이 推定値 $\bar{x}$의 期待値는 즉 可能한 모든 標本 ${}_NC_n$組에 대하여 만약 各己의 標本平均을 計算하고 이들을 平均한다면 (1)에서 보는 바와 같이 母平均 μ와 같아진다. 이와 같이 標本推定値의 母數와 같아지는 경우에는 이 推定値를 不偏推定値(Unbiased estimate)라고 하고 이 推定値는 偏倚(Bias)가 없다고 한다.

(1) 및 (2)에 의한 單純任意抽出에서는 標本平均 $\bar{x}$ 및 標本分散 s^2가 各各 母平均 μ 및 母分散 S^2의 不偏推定値임을 알았다. 그런데 推定値는 恒常 不偏하다고는 할 수 없다.

特히 標本標準偏差 s의 期待値는 母標準偏差 S와 一致하지 않는다. 그러나 近似하므로 母標準偏差의 推定値로 使用함에 矛盾이 없다고 할 수 있다. 따라서 推定値는 一般的으로 不偏推定値인 것을 바라나 不偏推定値이어야만 使用할 수 있는 것이 아니고 때에 따라서는 偏倚가 있는 推定値를 使用하는 것이 더 좋을 때도 있으나 여기에서는 다루지 않기로 한다.

4. 標本平均의 分布와 標本誤差

앞에서도 본 바와 같이 크기 N의 母集團에서 標本 n의 可能數는 ${}_NC_n$個이었다. 지금 이들 모든 可能한 標本에 대하여서 모두 調査하여 그 平均値의 分布狀態를 보기로 한다. 이 分布는 母集團의 分布와는 相異할 것이다. 그러니 確率論에 의하면 標本平均의 分布에 대하여 다음의 두 定理가 證明되고 있다. 즉 單純任意標本의 경우에는

A. 母集團의 分布가 正規分布이면 標本平均의 分布도 正規分布가 된다.

B. 母集團의 分布가 어떠하든 n을 크게 하면 標本平均의 分布는 極限에 있어서는 正規分布가 된다.

위의 定理에 의하면 母集團의 分布如何에 不拘하고 標本數 n를 充分히 크게 하면 標本平均의 分布는 近似的으로 正規分布로 보아도 좋다.

그런데 實際問題로서 n는 얼마든지 크게 할 수는 없는 것이므로 n가 어느 程度이면 좋은가가 問題가 되는데 多幸히 變量調査의 경우에는 n가 30以上이면 標本平均의 分布는 正規分布로 볼 수 있다. 이 性質은 대단히 便利한 것으로 標本調査에서는 母集團의 分布型을 一但 考慮할 必要가 없게 된다.

그러면 먼저 平規分布의 性質을 알아 보기로 하자.

어느 統計集團이 平均은 μ이고 標準偏差는 σ인 正規分布를 한다고 하면

A. 이 集團의 分布曲線은 平均値 μ를 中心으로 左右 對稱이고

B. 이 때 標準偏差의 값은 曲線의 모양을 定하고 σ의 값이 클수록 偏平하게 되고 σ의 값이 작을수로 좁고 높게 된다.

C. 平均値 μ를 中心으로 σ만큼의 距離를 取하면 그 區間內에 集團의 總個體中 68.27%가 包含되고 2σ의 거리를 取하면 95.45%, 3σ의 거리를 取하면 99.37% 즉 集團의 모든 個體가 거의 包含되게 된다.

이 제 標本平均의 分布에 위에서 본 正規分布의 性質을 適用하면 즉 모든 可能한 標本에서 얻은 標本平均들은 母平均을 中心으로 2倍의 標本平均의 標準偏差區間을 取할 때 그 區間內에 이들 標本平均中 95% 以上이 包含된다. 따라서 單純任意抽出에 依하여 얻은 어느 하나의 標本平均값 $\bar{x}$와 母平均값 μ와의 差異가 2倍의 標本平均의 標準偏差값보다 적을 確率은 95% 以上이라고 歸納할 수 있겠다.

위에서 標本平均의 標準偏差를 標準誤差(Standard error) 또는 標本誤差(Sampling error)라고 부르며 式(3)을 제곱根으로 얻게 된다.

$$\sigma_{\bar{x}}=\sqrt{\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}}$$

$$\text{또는 } \sqrt{\frac{N-n}{N}\cdot\frac{s^2}{n}} \quad \cdots\cdots\cdots\cdots(4)$$

그런데 母分散 σ^2은 實際에 있어 모르고 있으므로 標本推定値 s^2로 代置하여 쓰게 되며 式(4)는 다시 다음과 같이 된다.

$$s_{\bar{x}}=\sqrt{\frac{N-n}{N}\cdot\frac{s^2}{n}} \quad \cdots\cdots\cdots\cdots(5)$$

이 때 $s_{\bar{x}}$는 標本誤差(또는 標準誤差)의 推定値가 되는 것이다.

따라서 標本에서 計算할 수 있는 것은 $\sigma_{\bar{x}}$가 아니고 $s_{\bar{x}}$인 것이다. 그렇다면 이들에 대하여서도 앞의 理論이 成立하는가 하는 것에 疑問을 갖는 것이 當然한데 實務에 있어서는 一但 成立하는 것으로 보아도 무방하므로 證明없이 받아들이기로 한다.

以上에서 說明한 內容을 다시 要約하여 본다면 즉 平均이 μ, 標準偏差가 σ인 母集團에서 適當한 크기의 標本을 抽出하면 그 標本平均 $\bar{x}$는 平均이 μ이고 標準偏差가 $\sigma_{\bar{x}}$인 正規分布를 한다고 볼 수 있다. 따라서 母平均推定値로 使用된 어느 하나의 標本平均에서 2倍의 標本誤差區間을 取하면 그 區間內에 母平均의 眞値가 位置할 確率이 95% 以上이 된다고 할 수 있다. 이제 이것을 一般式으로 나타내면

$$P_r\{\bar{x}-k\sigma_{\bar{x}}<\mu<\bar{x}+k\sigma_{\bar{x}}\}=0.954 \quad \cdots\cdots(6)$$

($k=2$일 때 즉 標本誤差區間을 取하면)

되고 여기서 k를 信賴係數라 하며 $k=1$일 때에는 68.3%, $k=3$일 때에는 99.7%의 信賴區間을 各各 얻게 된다.

5. 必要한 標本數의 決定

標本數가 크면 클수록 그만큼 調査費用은 많이 들지만 標本誤差는 적어지게 된다. 그런데 實際에 있어서 費用은 制限을 받는 것이기 때문에 이 制限된 豫算內에서 如何히 調査結果의 精度를 높일 수 있는가가 問題된다.

그러나 費用이란 여러 가지 條件으로 調査種類에 따라 달라지기 때문에 測定이 困難하므로 여기서는 費用을 생각하지 않고 調査結果의 精度에 의하여서만 標本數를 決定하는 方法에 대하여 알아 보기로 하자.

例컨대 어떤 地域의 家口當 平均消費支出額을 알고자 할 때 이에 使用될 標本의 크기를 決定키 위하여서는 標本調査의 結果에 의한 推定値가 얼마만큼의 正確性을 가지면 充分한지 또는 얼마만큼의 正確性을 必要로 하는지를 미리 定하여야 한다. 이제 信賴水準 σ에서 推定될 平均消費支出額의 標本誤差의 許容限界를 E 以內로 할려고 한다면

$$k\sigma_{\bar{x}}=k\sqrt{\frac{N-n}{N}\cdot\frac{\sigma^2}{n}}\leqq E \quad \cdots\cdots\cdots\cdots(7)$$

이란 不等式을 만들고 兩邊에 제곱을 取하여 n에 대하여 풀면

$$n\geqq\frac{k^2N\sigma^2}{k^2\sigma^2+NE^2} \quad \cdots\cdots\cdots\cdots(8)$$

을 얻게 된다. 그런데 n을 풀기 위하여서는 반드시 母分散 σ^2를 알고 있어야 하는데 普通 正確히 알지 못한다.

그러므로 過大調査의 結果라든지 또는 他調査의 結果에 의한 推定値를 使用하게 된다.

6. 具體的인 抽出方法

위에서도 說明한 바와 같이 單純任意抽出法이란 母集團의 모든 個體가 標本으로 選出될 確率을 同一하게 갖게 하는 方法으로서 말하자면 제비 뽑는 式의 抽出法이다.

그런데 母集團의 個體의 數가 많을 때에는 제비를 만드는 일만도 相當한 作業量이 되고 또 그것들이 均一하게 만들어지기 어렵다. 그래서 이와 같은 抽籤의 道具로서 統計에서는 亂數表라는 것을 使用한다. 亂數表는 0에서 9까지의 數字를 任意로 配列하여 놓은 것으로 말하자면 0~9의 數字를 抽籤으로 몇 번이고 反復하여 뽑아 그 結果를 記錄한 表라고 생각하면 좋다.

附錄에 있는 亂數表는 다섯 자리의 數로서 配列하여 놓았는데 이는 便宜上 그렇게 한 것 뿐이지 다른 뜻이 있는 것은 아니다.

지금 母集團이 30個의 要素로 構或되어 있다 하자. 이 때에 먼저 이들 30個의 個體에 어떠한 順으로도 좋으니 01, 02, 03, …, 29, 30 과 같이 一連番號를 붙인다. 그리고 亂數表에서 任意의 자리를 골라 거기서부터 시작하여 30 以上의 數가 나오면 그 數에 해당하는 番號의 個體를 標本으로 하면 된다. 가령 附錄의 亂數表을 左上의 처음 두 줄부터 始作하여 아래로 내려 읽기로 한다면 最初로 나오는 數는 02이므로 그 番의 個體를 標本으로 한다.

만약 또 하나의 標本을 抽出코자 한다면 그 밑은 85인데 이는 30보다 큰 數이므로 그냥 건너뛰고 그 다음을 보면 26이니 26番의 個體를 抽出하면 된다.

이와 같은 方法을 繼續하여 나가면 얼마든지 必要한 數의 標本을 뽑을 수 있을 것이다.

다시 母集團의 個體의 數가 300個일 때를 생각하여 보자.

이 때에도 앞에서와 마찬가지로 우선 亂數表의 어느 列, 어느 行부터 始作할 것인가를 定한다. 이의 決定도 全的으로 任意로 決定하는 것이므로 例를 들자면 눈을 감고 鉛筆을 떨어뜨려 그 끝이 맞는 點에서 가장 가까운 두 數를 읽어 처음 數는 列을, 다음 數는 行의 番號로 定하고 그 點부터 始作하여 세 자리씩 읽어 나가는 식으로 하면 된다. 지금 鉛筆 끝이 指示한 點의 數가 95라 한다면 9列 5行을 起點으로 하여 세 줄의 數를 아래로 읽어 나가면서 抽出하게 된다. (이때 옆으로 세 자리씩 읽어 나가도 좋다)

附錄의 亂數表에서 보면 9列 5行을 起點으로 하면 (394), 221, (321), 005, (742), (945), (452), (615), (948), (806), (750), (838), (697), 197, (766), (257), (488), 019……와 같이 된다. 여기에서 ( ) 內의 數는 모두 300을 넘으므로 버리고 221, 005, 197, 019, ……를 取하게 된다. 그런데 위에서 본 바와 같이 세 자리의 數를 읽어 나가자면 300을 넘는 數가 많이 나온다.

이 때에 이들을 全部 버리고 있으면 約 50個의 標本을 抽出하기에도 相當한 時間이 걸린다. 따라서 이를 節約하기 위하여 다음과 같은 方法을 쓰면 좋다. 즉 지금 亂數表에서 나온 數字를 N이라 하면 이를 300으로 나누어 나머지 數를 利用하면 된다. 즉

$$\frac{N}{300}=a+R$$

와 같이 되는데 이 때 이 R는 300보다 작은 數일 것이다. 이렇게 한 結果를 보면 194, 221, 21, 005, 142, 45, 15, 48, 206, 150……과 같이 되어 이들 數에 該當하는 番號의 個體를 標本으로 뽑게 된다.

亂數表의 使用에서 注意하여야 할 일은 언제나 같은 表를 使用하지 않도록 하는 일이다. 始作하는 起點이나 읽는 順序 또는 方向을 任意로 자주 바꾸지 않으면 안된다.

7. 結 言

以上에서 單純任意抽出法의 基礎理論과 具體的인 標本抽出方法을 살펴 보았다.

序言에서 記述한 바와 같이 본 單純任意抽出法은 어디까지나 母集團에 特別한 措置를 加함이 없이 直接的으로 一定數의 標本을 抽出하는 方法이므로 母集團에 있어서 同一性質의 統計 單

(8 page에 계속)