<論 說>

標本誤差에 關하여

咸 鍾 郁

1. 序 言

標本에서 計算한 統計量은 母集團의 母數와는 반드시 一致하지 않는다. 標本에서 計算한 標本의 平均値는 全數調査 즉, 母集團 全體에 對하여 調査한 母集團平均値와는 다소라도 다른 값을 취하는데 이러한 誤差를 標本誤差 또는 抽出誤差라고도 한다. 一般的으로 統計量에 어느 만큼의 誤差가 包含되어 있는가는 確定할 수 없기 때문에 統計量의 誤差가 어떠한 모양으로 나타나는가를 알기 위해서 標本分布 즉, 統計量의 分布를 생각하게 된다. 同一한 母集團에서 一定한 크기의 標本을 몇번이고 抽出하여 平均値나 比率 등의 統計量을 計算하여 보면 標本에 依해서 統計量의 값에는 여러가지 誤差가 생긴다. 이러한 差가 나타나는 것은 統計量이 標本에 의해서 計算되기 때문에 誤差가 包含되어 있으므로 나타나는 것이다. 여기에서 이 統計量의 標本에 依한 變化를 度數分布의 形式으로 정리하여 보면 標本의 誤差가 어떠한 모양으로 나타나는가가 分明히 된다. 標本을 構成하는 單位는 確率變數인데 統計量은 관찰된 標本의 函數이므로 統計量도 確率變數이다.

2. 比率의 標本誤差

母集團의 單位가 두 범주로 分類되어 그 比率이 一定할 때 이 母集團을 二項母集團이라 한다. 또 母集團이 셋 이상의 범주로 나누어 있어도 이것을 二項母集團의 形으로 만들 수 있다. 그러므로 統計에서 二項母集團의 모양으로 취급할 수 있는 경우가 大端히 많다. 母集團에 있어서 어떤 事件의 確率 p가 항상 一定할 때 n回의 관찰에서 이 事件의 出現回數 x는 二項分布를 이룬다. 여기서 x를 n으로 除한 $\frac{x}{n}$는 이 事件의 標本比率이며 관찰의 回數 n이 一定할 때 出現回數 x가 二項分布를 이루면 標本比率 $\frac{x}{n}$도 二項分布를 이룬다. 그리하여 事件의 出現回數 n와 標本比率 $\frac{n}{x}$와의 標本 分布는 二項分布로 주ㅣ다.

二項分布에 있어서 n이 크면 이 分布는

平均値 $m=np$ 分散 $\sigma^2=npq$

의 正規分布에 가까와진다. 一般으로 事件의 出現回數와 標本比率의 標本分布는 다음과 같이 생각한다.

(가) 事件의 出現回數의 경우

母集團에 있어서 事件의 確率 또는 特定한 범주에 속하는 單位의 比率이 p일 경우 크기 n의 관찰標本 가운데에 包含되는 事件의 出現回數(또는 特定한 범주의 單位) x는 平均値 np 分散 npq 따라서 標本偏差 $\sqrt{npq}$의 正規分布를 이룬다.

(나) 標本比率의 경우

위의 標本에서 計算된 標本比率$\frac{x}{n}$는 平均値 p 分散$\frac{pq}{n}$의 正規分布를 이룬다. 이것에서 우리는 標本比率에 包含되는 標本誤差는 標本偏差$\sqrt{\frac{pq}{n}}$의 3倍를 넘는 경우는 거의 없고 標本誤差가 標準偏差의 2倍를 넘는 確率은 5%以下이다. 이 때문에 이 標本比率의 標準偏差를 σ_p로 表示하면 $\sigma_p=\sqrt{\frac{pq}{n}}$인데 이를 特히 標本比率의 標準誤差 (Standard Error of Ratio)라 한다.

3. 有限母集團修正係數

標本比率의 標本誤差의 公式 $\sigma_p=\sqrt{\frac{pq}{n}}$는 母集團이 無限한 것이 전제로 되어 있다. 그러므로 母集團이 有限한 경우에는 標本比率의 標準誤差의 公式은 다음과 같이 修正된다.

$$\sigma_p=\sqrt{\frac{N-n}{N-1}}\cdot\sqrt{\frac{pq}{n}}$$

여기에서 N은 集母圖의 單位數이다. 그리고 $\sqrt{\frac{N-n}{N-1}}$을 有限母集團修正係數(Finite Multiplier of Population)라 한다.

修正係數의 性質에는 다음 세가지 關係가 있는 것을 알 수 있다.

즉, (가) 有限母集團에서의 標本比率의 標準誤差는 無限母集團에서의 같은 크기의 標本에 對한 標準誤差보다 一般的으로 적다. 修正係數는 다음과 같이 근사적으로 變形할 수 있기 때문이다.

$$\sqrt{\frac{N-n}{N-1}}\fallingdotseq\sqrt{1-\frac{n}{N}}=\sqrt{\left(1-\frac{n}{2N}\right)^2-\left(\frac{n}{2N}\right)^2}$$

$$\fallingdotseq 1-\frac{n}{2N}$$

여기에서 N은 母集團의 單位數이고 n은 標本의 크기이므로 $\frac{n}{N}$은 標本의 抽出率이다. 修正係數는 1에서 標本抽出率의 $\frac{1}{2}$을 減한 값이므로 一般的으로 1보다 작은 값이다.

그러나 N이 클수록 抽出率은 작아지고 修正係數는 1에 가까지며 N이 無限大로 되면 즉, 無限母集團인 경우는 1이 되기 때문이다.

(나) 標本의 抽出率이 10%以下이면(실제로는 10%以下의 抽出率이 大部分이다) 有限母集團과 無限母集團과의 修正係數의 差는 大端히 작으므로 有限修正係數를 無視해도 되는 경우가 많다.

抽出率이 10%일 경우 修正係數는 式公에 $\frac{n}{N}$대신 0.1을 대입하여

$$1-\frac{1}{2}\times 0.1=0.95 \text{ 가 되므로}$$

修正係數를 가지고 있지 않는 無限母集團의 경우와도 標準誤差의 差는 5%이며 또 大規模인 標本調査에서는 抽出率이 1%以下가 많은데 이 때에 母集團을 無限으로 생각하여 標準誤差를 計算함으로써 생기는 差는 0.5%以下가 되어 兩者의 差는 大端히 작게 되기 때문이다.

(다) 標本比率의 正確性은 主로 標本의 크기에 依해서 정해지는 것이며 標本抽出率의 영향은 무시해도 좋을 정도로 작다.

4. 平均値의 標本誤差

同一한 母集團에서 같은 크기의 標本을 몇 組든지 抽出하여 그 標本平均 $\bar{x}$를 確率變數로 하여 度數分布表를 만든 것을 標本平均値의 分布(Sample Mean Distribution)라 한다. 이 分布에는 다음과 같은 성질이 있다.

첫째 母集團의 分布의 形이 어떠한 모양을 하고 있던지 이 母集團에서 抽出한 標本에서 計算한 平均値 $\bar{x}$의 標本分布는 正規分布에 가까운 形이 되는 傾向이 있다.

둘째의 성질은 標本平均値의 分布의 中心 즉, 標本平均値의 平均値는 母集團 平均値와 같은데 있는 것이다.

(1) 標本平均値의 平均値와 分散

標本平均値 $\bar{x}$이 分布의 平均値와 分散은 다음과 같다 즉,

母集團平均値가 m 母分散이 σ^2인 無限母集團에서 抽出된 크기 n인 一組의 標本 $x_1, x_2, x_3, \cdots x_n$에 대하여 이 標本平均値를 $\bar{x}=\frac{1}{n}(x_1+x_2+ +x_n)$라 하면

㉮ $\bar{x}$의 平均値는 母集團平均値 m와 같다.

$$E(\bar{x})=m$$

(證明) $E(\bar{x})=\frac{1}{n}E(x_1+x_2+\cdots\cdots+x_n)$

$$=\frac{1}{n}\{E(x_1)+E(x_2)+\cdots\cdots+E(x_n)\}$$

그런데 $x_1, x_2, x_3, \cdots\cdots x_n$은 모두 같은 母集團의 變量이므로

$$E(x_1)=E(x_2)=\cdots\cdots=E(x_n)=m$$

$$\therefore E(\bar{x})=\frac{1}{n}\cdot n\cdot m=m$$

㉯ $\bar{x}$의 分散을 $\sigma_{\bar{x}}{}^2$로 表示하면

$\sigma_{\bar{x}}{}^2=\frac{\sigma^2}{n}$이다.

㉮의 $E(\bar{x})=m$의 關係는 母集團이 有限하건 無限하건 모두 成立한다. 그러나 ㉯의 $\sigma_{\bar{x}}{}^2=\frac{\sigma^2}{n}$은 無限母集團의 境遇에만 成立하고 有限母集團의 경우에는 有限母集團修正係數를 사용하여

$$\frac{N-n}{N-1}\ \frac{\sigma^2}{n} \text{ 으로 表示한다.}$$

(2) 平均値의 標準誤差

위의 標本平均値의 平均値와 分散에 關한 結

果는 母集團의 分布의 形에 關係없이 成立한다. 그리고 이 標本平均値의 分布의 形에 대하여는 다음 두 가지 關係를 알아 두어야 한다.

첫째 母集團平均値m 分散 σ^2의 正規分布로 할 때 이것에서 抽出한 크기 n인 一組의 標本의 平均値의 標本分布는 母平均値m 分散$\frac{\sigma^2}{n}$의 正規分布로 된다.

둘째 그러나 母集團이 正規分布로 되지 않아도 母集團의 平均値 m와 分散 σ^2이 有限値이면 그 分布의 形이 어떤 모양을 하고 있든지 標本平均値의 分布는 標本의 크기 n이 크면 平均値 m 分散$\frac{\sigma^2}{n}$의 正規分布로 가까워지며 n이 크면 클수록 그 가까워지는 정도는 커지며 $n\to\infty$의 極限에 있어서 이것은 正規分布에 完全히 一致하게된다.

標本의 크기 n이 크면 클수록 標本平均値의 分布가 正規分布로 가까와 간다는 것은 數理統計에서는 極히 重要한 의미를 가지고 있다.

母集團分散의 값만 알고 있으면 母集團分散의 形에 關係없이 正規分布를 사용하여 平均値의 標本誤差를 計算할 수 있기 때문이다.

요컨대 平均値m 分散σ^2의 母集團에서 抽出한 크기 n의 標本의 平均値 $\bar{x}$의 標本誤差는

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$$

의 3倍를 넘는 것은 거의 없고 2倍를 넘는 것도 그 確率은 5%以下이다.

$$\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$$

을 平均植의 標準誤差(Standard Error of Mean)라 한다.

위의 平均植의 標準誤差의 公式은 無限母集團의 경우에 사용하는 것이며 有限母集團의 경우에는 標本比率의 경우와 마찬가지로 有限母集團修正係數를 사용하여 平均植의 標準誤差는

$$\sigma_{\bar{x}}=\sqrt{\frac{N-n}{N-1}}\cdot\frac{\sigma}{\sqrt{n}}$$ 이 된다.

5. 小標本理論과 t 分布

標本의 크기 n이 작을 때에는 正規分布가 되지 않고 t 分布가 되기 때문에 t 分布(t—distribution)의 理論을 小標本理論이라 한다. 이제 이것을 고찰해 보자.

위에서 본 바와 같이 標本平均値$\bar{x}$는 母集團平均値가 m 母分散이 σ^2일 때 平均値m 分散 $\frac{\sigma^2}{n}$의 正規分布로 된다.

이제 이것을 標準測定値로 變換하는 公式에 依하여

$$U=\frac{\bar{x}-m}{\sqrt{\frac{\sigma^2}{n}}}=\frac{(\bar{x}-m)\sqrt{n}}{\sigma}$$

로 만들면 이것은 平均値0 分散1인 標準正規分布를 하게 된다. 그런데 여기서 問題가 되는 것은 母集團分散σ^2이 일반적으로 不明하다는 것이다.

이러한 경우에 이 問題를 解決하는 方法은 σ값은 標本에서 計算한 標本標準偏差

$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$을 代用하는 것이다.

이러한 代用을 하게 되면 위의 確率變數 $U=\frac{(\bar{x}-m)\sqrt{n}}{\sigma}$는 正規分布를 이룬다는 理論이 成立하지 않는다.

그러므로 平均値 m 分散 σ^2의 正規母集團에 對해서 $t=\frac{(\bar{x}-m)\sqrt{n-1}}{s}$라는 確率變數 t를 생각하면 t는 標本分布를 이룬다. 이 理論的分布를 自由度(Degree of Freedom) $n-1$의 t分布라 한다.

t分布는 標本의 크기 n이 크게 됨에 따라 그 모양은 正規分布에 가까워진다. t分布에 對해서는 복잡한 계산을 행해야 하는데 實用上은 이에 對한 數値表가 되어있다.

例컨대 $n=11$일 때 自由度 $n-1$은 10이므로 t表에서 確率 $p=0.025$에 對한 數値에서 統計量 t의 絕對値가 2.228을 넘는 確率은 $2\times0.025=0.05$인 것을 알 수 있다.

$$t=\frac{(\bar{x}-m)\sqrt{n-1}}{s}=\frac{\bar{x}-m}{\frac{s}{\sqrt{n-1}}}$$ 이므로

標本平均値의 母集團平均値에 對한 誤差가 $\frac{s}{\sqrt{n-1}}=\frac{s}{\sqrt{10}}$의 2.228 보다 크게 되는 確率은 $2.5\%\times2=5\%$이다.

$\frac{s}{\sqrt{n-1}}$는 母集團의 σ가 不明인 경우에 σ를

(5 page에 계속)