2009/4/22　金融工学教材プリント 1

データの整理・統計

数値データまたは標本とよばれる「数値で表されたデータの集まり」に関する性質を理解することが必要です。統計で処理される数値データ全体の集まりは，「**母集団：*population***」とよばれます。この母集団について，統計処理の対象を母集団のすべてで扱う手法を，「**記述統計**」と言います。あるいは，記述統計は統計データから平均や標準偏差などのデータを特徴づけるものを計算したり，表やグラフを書いて統計データの持っている情報を引き出す方法などを一括した呼称としても使われます。また，母集団の一部のデータを扱い，その統計処理結果を母集団全体に当てはめて考えていくやり方を「**推測統計**」といいます。

一概には言えませんが，両者の決定的な相違は，母集団の大きさです。母集団が小さければ，そのものすべてを対象として，データを特徴づける平均や分散を直接調べることができます。しかし，母集団が余りにも大きくなりすぎたとき，母集団全体を調べることはできなくなります。この場合，重要なことですが，母集団から無作為に標本を抽出して，これをもとに母集団全体の特性を推測するやり方を「推測統計」と言います。

たとえば巨大な母集団を対象として，統計に関して期待する答えが存在したとします。しかし，この統計処理は実際上は不可能ですから，その結果は誰にもわかりません。その意味では「**神様しか知らない**」事案です。この神様しか知らない，母集団にかかわる未知の値を**母数**といい，とくに**母平均**μ，**母分散**σ^2といいます。大きな母集団の統計に関する真の値はとうてい得難いものですが，それが存在することを前提に分析を進めなければ，統計処理は意味を失います。

推測統計では，限られた標本数から標本平均と分散を求め，神様しか知らない真の値である母平均と母分散に近づこうとします。これに対して，標本から計算された標本平均，標本分散などは統計量といい母数ではありません。また，

Faculty of Economics
Osaka Gakuin University

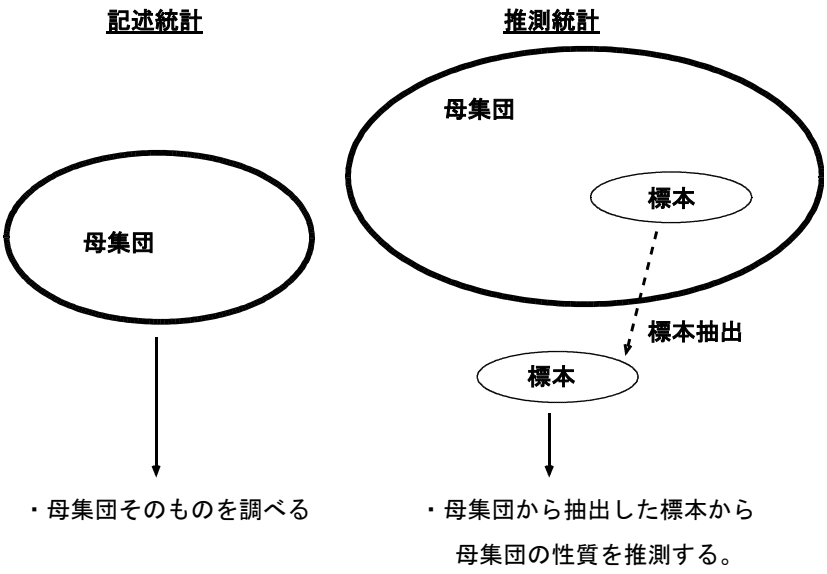
計算式の中に未知の母数が入っているものは，統計量とはいいません。今回の話で，統計量のことをお話ししましたが，標本平均と母平均，標本分散と母分散は厳密に分けて扱われねばなりません。標本分散や標本平均などの統計量が，真の値の母分散や母平均に近いものであるかどうかのチェックが必要です。限られた母集団から収集された標本で計算された標本統計量が，未知の母数とくらべて本当に確からしいのかどうか調べる作業は「検定」といいます。

標本平均　$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$, 標本分散　$s_X^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$

$\frac{1}{n} \cdot \sum_{i=1}^{n} \left(X_i - \mu \right)^2$ は，μ が未知のとき，統計量とはならない。

以上の話を要約すれば，話のイメージは次のような図となります。

記述統計と推測統計のイメージ



Faculty of Economics
Osaka Gakuin University

正規分布の特徴

標準正規分布 X ～ N（0 ，1^2）には正規分布表とよばれる数表が作成されている。まとめとして，X が平均 μ，分散 σ^2，X ～ N（μ, σ^2）ならば，分布について次のような特徴があります。

(1)分布は，x ＝ 0 としての平均 μ を中心に左右対称となる。

(2)分布は，x の値が無限大に近づくにつれて，x 軸に急速に接近する。

(3)分布は，平均値 μ で最も高くなり，左右の裾野へ向かうにつれて低くなる。

次に，最も重要で，このお話の目的でもある確率の話があります。確率変数 X がある値 a と b をとり，それは a ≦ X ≦ b とします。また，その条件を満たす確率を，P({a ≦ X ≦ b})と書きます。このとき，a と b については，次の関係が成立します。

$$P\left(\{a \le X \le b\}\right) = P\left(\left\{\frac{a-\mu}{\sigma} \le X \le \frac{b-\mu}{\sigma}\right\}\right)$$

ここで，μ の中心から左右に標準偏差 1 個（1 σ）だけずれた区間を考えると，それは $\mu-\sigma$ ≦ X ≦ $\mu+\sigma$ となります。正規分布では，X がこの条件を満たす区間にデータが入る確率は決まっていて，それは約 68.3 %となります。

(4) **1 σ の区間確率： P{X ｜ $\mu-\sigma \le X \le \mu+\sigma$} ≒ 0.6826**

次に，μ の中心から左右に標準偏差 2 個（2 σ）だけずれた区間を考えると，それは前と同じに，$\mu-2\ \sigma$ ≦ X ≦ $\mu+2\ \sigma$ と書けます。この区間にデータが入る確率は，約 95.4 %となります。

(5) **2 σ の区間確率： P{X ｜ $\mu-2\ \sigma \le X \le \mu+2\ \sigma$} ≒ 0.9544**

同様に，3 σ は 99.7 %となります。

(6) **3 σ の区間確率： P{X ｜ $\mu-3\ \sigma \le X \le \mu+3\ \sigma$} ≒ 0.9973**

これらの関係は，図 3-7-(b)に整理されています。また，**観測値が確率変動をしていて，母分散とほぼ同じかもしれない標本分散が求まれば，母平均とほぼ同じかもしれない標本平均からの乖離（はなれていること）誤差確率を求めることができます。**

ある株式価格の予想をするとして，次のような架空のデータ処理と統計量を得ることができました。

（Ⅰ） データ・平成 17 年，18 年，19 年の日時データ 500 個。

（Ⅱ） 価格の標本平均・5,000 円，

（Ⅲ） 価格の標本標準偏差・250 円

この株式価格が確率変動していると見なせば，価格変動の確率分布は正規分布で与えられます。標本平均と標本標準偏差が，本当に確からしいものとすれば，変動区間の生起確立を計算することができます。

(4') **1 σ** の変動幅　5,000 ± 1 × 250,　　5,250 ～ 4,750

5,250 円から 4,750 円までの変動幅となる確率は**約 68.3 %**。

(5') **2 σ** の変動幅　5,000 ± 2 × 250,　　5,500 ～ 4,500

5,500 円から 4,500 円までの変動幅となる確率は**約 95.4 %**。

(6') **3 σ** の変動幅　5,000 ± 3 × 250,　　5,750 ～ 4,250

5,750 円から 4,250 円までの変動幅となる確率は**約 99.73 %**。

Faculty of Economics
Osaka Gakuin University