§2-2　記述統計と推測統計

統計的な仮説検証とは、仮説に関連すると考えられるデータを集め、そのデータに対し統計学的な処理をする事で、わかりやすく仮説を検証する方法の事である。その時、集めたデータ(標本、サンプル)から算出される統計的な計算値を統計量といい、統計量を求めることで標本が示す傾向や特性を表そうとする方法を**記述統計**という。

しかし、限られた標本だけを対象とした場合、そこから導かれる傾向や特性といった結果もごく限られた範囲に限定されてしまう。統計的手段を使う事によって、集団全体の性質を知りたい。この時の集団全体、つまり標本を取り出す大元が**母集団**。母集団から抽出した標本によって、母集団の傾向や特性を推測する統計的分析を**推測統計**という。

通常、母集団は非常に大きく、母集団を直接測定する事は困難である。そのため、母集団の傾向や特性を知る為に、母集団から無作為(ランダム)に標本を抽出(サンプリング)する。この時の母集団から集めた標本の数を**標本の大きさ(サンプルサイズ)**という。

2-2-1　記述統計量

記述統計量には大きく分けて、データの中央傾向を知る為の**代表値**と、データの散らばりを知る為の**散布度**、という２つの指標が存在する。これら例として、前者は**平均**、後者は**標準偏差**、がそれぞれ挙げられる。

統計的な指標を選ぶ上で注意しなければならないのが、尺度の種類[1]である。例えば、平均は名義尺度や順序尺度のデータに使用しても意味のある数値は得られないが、間隔尺度や比率尺度のデータに使用すると意味を持つ。

また、標本から求める記述統計量を、母集団における値と区別したい場合は、前者を標本統計量、後者を母数という。

2-2-2　標準化得点(z 値)と偏差値(Z 値)

異なる集団を比較する場合、単純に測定した値のみを比較すると、うまく評

---

[1] ４種類存在し、名義尺度＞順序尺度＞間隔尺度＞比率尺度の順に拡張される。
名義尺度：大小比較が出来ず、事例数を数える尺度。例)性別、血液型、出身地
順序尺度：差は気にせず大小比較のみできる尺度。例)好きか嫌いか、成績評定
間隔尺度：等間隔の目盛りで値の差にのみ意味がある尺度。例)気温
比率尺度：原点からの差や比率を考える尺度。例)身長、体重

価が出来ない事が多い。

例えば、A さんが１回目のテストで７０点、２回目のテストで６４点だったという時、単純に数値だけを比較すると６点下がって成績が落ちたと思えるかもしれないが、実際にはテストの難易度や受講者数が変化しているなど他の条件が変わっている為に点数の変動があったかもしれず、一概には A さんの成績を考察する事は出来ない。このように 2 つのテストで平均や得点のばらつきが異なる場合には、それぞれのデータを平均が０、標準偏差が１になるように標準化を行うと、より公平な条件で成績の比較が出来る。このような標準化を行った得点の事を z 得点、もしくは z 値という。z 値に変換する為には以下の式 2.1 に代入すればよい。

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{\text{観測値} - \text{平均}}{\text{標準偏差}} \quad \text{(式2.1)}$$

この式より A さんの得点を求めると、
１回目のテスト(平均点 72.53 点、標準偏差 10.34)では z=-0.24、
２回目のテスト(平均点 63.72 点、標準偏差 12.49)では z=0.02
となり、負の値から正の値に転じている。この結果は、A さんが１回目のテストでは平均より下に位置していたのが、２回目のテストではほぼ平均に達している事を示している。なので、点こそは下がったものの成績は上がっているとも取れるのである。

このように、z 値に変換すると集団の中での位置が分かるだけでなく、異なる集団の得点の比較も可能になるのである。ただし、この手法を用いると平均以下の得点の z 値はマイナスになるため、間違えやすい・得点を負の値で表示する事になる、といった問題があるため、以下の式 2.2 を使って平均値 50、標準偏差 10 となるような標準値に変換する事が多い。この標準値を**偏差値**といい、正式には Z 値、もしくは Z 得点という。

$$\text{偏差値}(Z) = \text{標準得点}(z) \times \text{標準偏差}(10) + \text{平均値}(50) \quad \text{(式2.2)}$$

この式を使って、A さんの成績を偏差値で見ると、１回目のテストは 47.6、２回目のテストは偏差値 50.2 となり、より分かりやすくなる。

2-2-3　正規分布と標準正規分布

データをいくつかの階級に分け、その階級の中にあるデータの個数を数えた頻度分布が**度数分布**。この頻度分布を棒グラフ状にしたものをヒストグラムといい、背後にある母集団が平均を中心に左右対称のグラフになると**正規分布**と

呼ぶ。この分布の平均を０、標準偏差を１に変換して標準化したものが**標準正規分布**で、平均 0 を中心に±1 標準偏差内に約 68%のデータが入る。母集団が正規分布に従うと仮定される場合には、変換した z 値については、-1.96≦z≦1.96 内に 95%のデータが含まれる。

しかし、実際のデータはきれいな正規分布になっていないので分析前にデータが正規分布に従っているかという調べる必要がある。

2-2-4 平均をモデルにした統計

母集団を代表するようにデータを収集し、モデルを立てどの程度適合するか見る。平均をモデルにした統計では個々の観測値は平均がズレており、それを誤差、残差という。（テストの点数など）ここでは誤差が小さければモデルがデータを良く説明していることになる。母集団を式に当てはめた時の母平均の誤差を標準後さと呼ぶ。

2-2-5 標準誤差と信頼区間

統計分析の関心

→統計のモデルとデータがどれだけ**適合**しているか(グラフの形がどれだけ似ているか)

→そのデータが属している**母集団の傾向を推測する**(母集団すべてを測定することはコスト面から不可能)

→母集団からランダムに取り出したデータがどれだけその母集団の傾向を引き継いでいるかを示す必要がある。(**サンプルデータの中での傾向と母集団の傾向がどれだけ合致しているか**)

→それを示す統計量が「**標準誤差**」と「**信頼区間**」

(1) 標準誤差

母集団のほんの一部である標本(データ)をサンプリング(取り出す)たびに、毎回取り出したデータで平均をとると毎回の平均の値は全く同じにはならずに若干のばらつきを生じる。

ランダムサンプリング：母集団から無作為に標本を抽出すること。無作為に取り出すことでその標本は母集団の傾向をよりよく表す。しかし、実際には、標本の完全無作為抽出は非常に難しい。

例 1）4 択(a.b.c.d)の問題を作成するときに、a とか d の選択肢が正解になる事よりも b とか c のような真ん中の選択肢を正解にすることの方が多い
例 2)A と B という文字を 3 文字ランダムで並べる
→全部で 8 通りあるなかで「AAA」とか「BBB」とかの連続するパターンを意図的に避けがちしてしまう(本当は 8 通り中の 2 通りだから 25%の確率で出るはず)
また、1 回に抽出する標本の大きさ(データの個数)によっても母集団の傾向をどの程度表しているかが変わる。
→取り出したデータが少ないと誤差が大きくなり、多いと誤差が小さくなる。しかし、データをたくさん得るにはコスト(労力)がかかる。

図 2.16 のようにサンプリングする(標本を抽出する)たびに、標本平均の値は変わる(ばらついている)→このように標本平均は全く同じ値にはならない

標準偏差：サンプリングした標本平均の標準偏差(標本平均の値がどの程度ばらついているか)。母集団そのものすべてを測定することはできないので母集団の平均(母平均)はわからない。→何回もサンプリングをして得た標本平均の値の平均値は母平均に近づく

中心極限定理：サンプリングして得た標本平均の分布は、データを多くすればするほど、もとの母集団の形には関係なく、正規分布に近づく
しかし、実際には無限にサンプリングをしてデータを増やすことはできない(なぜなら、それは母集団を直接測定していることとおなじだから)
→式 2.8 を使って**標準誤差(*SE*)**を出す

$$SE = \frac{s}{\sqrt{n}}$$

標準偏差 s が小さければ小さいほど標準誤差が小さい→推定の精度が高くなる
サンプルサイズ n(データの多さ)を大きくすればするほど標準誤差が小さい→推定の精度が高くなる

(2) 信頼区間
標本平均の正確さを表す方法その 2→区間推定
区間推定：母平均が含まれる範囲を推定する事。またその範囲を信頼区間とい

う。得られた標本の平均が、未知の母平均からどの程度ずれている(誤差がある)のかわからない時はこの平均を中心として、「何%の確率でこの区間に簿平均が存在する」という範囲を設定する。
95%信頼区間→z 値が±1.96→こっちの方がよく使う
99%信頼区間→z 値が±2.58→4%しか上がらないのに、上げるためのコストがとても大きい
信頼区間 95%の領域を求める→z 値が 1.96 になる x の値を求める。
この場合、標本平均の信頼区間を求めても意味がないので、簿平均が存在するであろう信頼区間を推定したい。(統計で知りたいのは母集団の性質)
→z 値を求める公式の s には母平均の標準偏差を入れる。

下側信頼限界値：$-1.96 = \frac{x_i - \bar{x}}{s}$を変形して、$x = \bar{x} + (-1.96 \times SE)$

上側信頼限界値：$1.96 = \frac{x_i - \bar{x}}{s}$を変形して、$x = \bar{x} + (1.96 \times SE)$

これを使うと、95%信頼区間は、

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right)$$

と表せる。母平均の標準偏差 s が小さい、もしくはサンプルサイズ n が大きいほど、信頼区間の範囲は狭く推定制度がよい。(サンプルデータが母集団の傾向をよく表している)

小さい標本の場合：サンプリングする標本のサイズが小さいとき(データの数が少ないとき)、正規分布にはならない→自由度によって分布が変わる t 分布を使用する。

$$\left(\bar{x} - t_{n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1}\frac{s}{\sqrt{n}}\right)$$