# 標本を用いた推定のシミュレーションにおける母集団の生成

片山, 雄一
九州大学大学院経済学研究院 : 専門研究員

https://doi.org/10.15017/7377094

出版情報：経済論究. 181, pp.57-71, 2025-07-31. Kyushu Daigaku Daigakuin Keizaigakukai

# 標本を用いた推定のシミュレーションにおける母集団の生成

Generating Population in Simulations of Estimation with Samples

片　山　雄　一[†]
Yuichi Katayama

## 要旨

本論文では，シミュレーションにおける母集団データの生成方法について検討する。標本を用いた分析や推定の結果と，母集団における真の値との比較を行うシミュレーションの枠組みを概説し，分析方法や推定手法を評価するうえで母集団に求められる特性について論じる。これらの特性には，属性の確率分布，平均，分散などが含まれ，単一の属性によって決まる場合と，他の属性との関係によって決まる場合がある。本論文では，基本的な単一属性の値の生成に焦点を当て，母集団生成の方法について述べる。また，検討した生成法を具体的なシミュレーションに適用し，実効性を検証する。

## 1．はじめに

データによる根拠に基づく政策立案もしくは政策評価は，様々な社会問題への対策に有用となりえる。特に，昨今の社会情勢に関する急激な変化は，様々な社会問題の根底にも大きく影響をもたらしている。現代社会において，信頼性の高い根拠を提示することは，今後の社会に向けて，問題提起および問題解決に向けての対策への政策立案に貢献しうる。このような観点から，社会調査によるデータの適切な分析方法や推定手法は，現代の政策科学において重要な役割を担っているといえる。

社会調査では，全数調査を実施することが難しい場合が多く，一般には母集団から抽出された標本を用いて母集団を特徴づけるパラメータを推定する。精度の高い推定結果を得るためには，どのような分析方法や推定手法を用いれば有効かが問題となる。母集団全体を観測することが難しいのが一般的であることから，推定結果を母集団の真の値と照合することは困難であり，分析方法や推定手法の有用性を検証することは現実には不可能である。

しかし，シミュレーションを用いれば，母集団の真の値を用いて分析方法や推定手法の有用性を検証することが可能となり，標本を用いた計算，例えばサンプリングバイアスの補正，がどのくらい正確かを検証できる。本研究では，確率分布に基づく様々な特性の母集団を生成し，分析方法や推定手法の有用性の検証に必要なシミュレーションの枠組みの構築を行なう。シミュレーションの利点は，現実では全体を観測することが困難な母集団の真の値と標本を用いた分析や推定による値が近似しているかを比較できることである。シミュレーションによる分析方法や推定手法の比較においては，母

[†] 九州大学大学院経済学研究院専門研究員

集団の特性を様々に変化させることが可能となる。また，異なる属性や傾向を持つような複雑な構造の母集団や，異常値，外れ値，欠損値を含む母集団も生成できる。さらに，分析方法や推定手法を用いて推定を行なう際に，正規性を仮定できない場合も存在し，従来の分析方法や推定手法の性能の検証および分析方法や推定手法のさらなる発展可能性を探る点においても，本研究での体系的なシミュレーションの意義は大きい。

本論文は，シミュレーションのための母集団の生成について議論している。標本を用いた分析や推定の結果と母集団での真の値の比較を行うシミュレーションの概要について述べ，分析方法や推定手法を評価するためにシミュレーションに用いる母集団に求められる特性を検討する。特性には属性の確率分布や平均，分散などがあり，単一属性のみで特性が決定される場合や他の属性の値との関係の中で決定される場合がある。本論文では，基本的な母集団の生成として単一属性の値の生成を対象とする。また，検討した生成法を具体的なシミュレーションに適用し，その実効性を検証する。

## ２．シミュレーションの概要

シミュレーションを用いることで，母集団全体は明らかとなり，母集団の真の値と標本を用いた分析や推定による値を比較することで分析や推定の精度を評価することができる。これにより，異なる分析方法や推定手法の有用性の評価が可能となり，母集団の特性の違いによる有用性の違いを検証できる。すなわち，シミュレーションはどのような特性の母集団のときにどのような手法が有用かを明らかにする。そのため，シミュレーションでは様々な特性の母集団を生成する必要がある。

分析方法や推定手法の有用性を評価するためのシミュレーションの手順は次のものとなる。

１．設定した母集団の特性に基づいて属性の値を生成する。
２．様々な標本抽出に従って１で生成した母集団から標本を抽出する。
３．評価の対象とする分析方法や推定手法により２の標本を用いて値を計算する。
４．３で求めた値の真の値を母集団を分析することにより計算する。
５．３の値を４の真の値と比較し，３の値の精度などから分析方法や推定手法の有用性を評価する。
６．母集団の特性を変えて１～５を繰り返し，母集団の特性による有用性の違いを検証する。

このようなシミュレーションを行うことで，理論的には有効とされる推定手法が，様々な特性を持つ母集団に対してどの程度有効かを検証できる。現実の母集団は，必ずしも単純な分布や単一の相関構造を持っているわけではない。そのため，本シミュレーションでは，異なる分布特性や相関構造（正の相関，負の相関，非線形関係）を持つ様々な母集団を生成できるようにする。

このような様々な母集団を生成する上で，母集団の最も基本的な特性は属性値の分布である。シミュレーションでは，母集団の個々の属性について分布を仮定し，母集団を生成する。

母集団のデータの大きさを$N$，属性の数を$m$とし，母集団の$i$番目のデータの$j$番目の属性の値を$X_{ij}$ $(i=1, \ldots, N,\ j=1, \ldots, m)$とする。属性の値の分布は独立している場合もあり，他の属性の値と関係性がある場合もある。独立している場合は，確率分布に基づいて$N$個の値を生成する。$k$番目の属性と関係性がある$l$番目の属性の値は，$i$番目のデータの確率分布に基づいて生成された$k$番目

の属性の値 $X_{ik}$ ごとに $X_{il}$ の関係性を設定する。例えば，$k$ 番目の属性の値が大きくなるほどに，対応する $l$ 番目の属性の値の平均が減少するような設定である。この際，負の相関や正の相関などのあらゆる関係性を設定できるようにする。このようにして関係性がある属性の値を生成することで，現実の社会調査等で観察される複雑な相関構造を持つ母集団を模倣することができる。母集団間で異なる相関構造や分布特性（例えば裾の重い分布や多峰性）を持たせることも可能であり，多様な状況下で推定手法を評価することができる。

このように，シミュレーションでは母集団の特性を設定するため，様々な状況における分析や推定の精度を詳細に検証することが可能である。母集団の特性の違いによる有用性の違いを検証することが本研究の目的であるが，母集団においては，確率分布通りに属性の値が分布することは稀であり，さまざまな外的要因により，分布がゆらぐことが一般的である。そのために，シミュレーションにおける母集団生成において属性の値を単純に理論的な確率分布に基づいて生成するだけでなく，確率分布自体にかく乱を加えることにより，分布にゆらぎをあたえる。

本論文では，所要時間の確率分布および所要時間ごとの出向頻度の確率分布に対して，ランダムなかく乱を加える方法を採用した。具体的には，各確率に対して一様分布に従うランダムな値 $\theta$ を加え，全体の合計が１となるように再標準化（正規化）を行うことで，確率分布をかく乱した。加えるランダム値 $\theta$ の上限を設定することで，かく乱の程度を調整できるように設計しており，$\theta$ の上限値を設定をした。これにより，分布の基本的な傾向を保ちつつも，より現実的なばらつきや歪みを持つ母集団を生成することが可能となった。このような設計により，単純な理論モデルでは捉えきれない現実的なデータ特性をシミュレーション上で再現し，推定手法の実践的な有効性や頑健性を検証できるようにしている。

生成した母集団から標本の抽出をシミュレーションする方法は，居住地ベース調査（例：地域住民を対象とした無作為抽出），来街地ベース調査（例：商業施設の来訪者を対象とした抽出），層別抽出（例：年齢層や居住地域ごとに層を分けた上での無作為抽出），非無作為抽出（例：特定条件を満たす者のみを対象とする）などの様々な標本抽出の方法がある。このようにシミュレーションの枠組みは設定を変えることによって様々な事例に適用することができる。標本抽出方法の違いは，得られるデータの性質に影響を与え，推定のバイアスや分散にも差異を生じさせる可能性がある。本研究で構築したシミュレーションの枠組みは，単に特定の調査に限定されるものではない。調査設計段階における最適な標本抽出方法の選択，推定手法の選定，データのバイアス補正方法の評価などを可能とする。

## ３．母集団の特性

母集団の特性には，単一の属性における特性と，複数の属性間における特性が考えられる。単一の属性の特性としては，平均や分散などが挙げられる。一方，複数の属性間の特性としては，相関関係や因果関係が挙げられる。相関関係とは，２つの属性が類似した変動を示す関係であるが，必ずしも一方の属性が他方に影響を与えることを意味するわけではない。それぞれの属性をランダムに独立に

生成すれば，無相関となる。因果関係とは，一方の属性が他方に影響を与える関係である。

任意の特性の母集団を生成するために，単一の属性における特性について検討する。単一の属性の特性は，その属性の $N$ 個の値（$N$ は母集団のデータ数）が持つ特性であり，その分布や集中・拡散の様子を特徴づける様々な統計的指標が存在する。

1．平均

平均は，確率変数の中心傾向を表す最も基本的な指標である（Casella and Berger, 2024）。また，分布の重心に相当する概念であり，データの全体像を単一の数値で要約する役割を果たす。

2．分散

分散は，データの散らばりの程度を数量化する指標である（Mood et al., 1974）。また，分散は外れ値に敏感であり，極端な観測値が存在すると急激に増加する。このため，ロバストな散布度指標として，四分位範囲などの利用も提案されている（Huber and Ronchetti, 2011）。

3．確率分布

単一の属性の本質的な特性は，確率分布に依存する（Wasserman, 2013）。離散型分布では確率質量関数，連続型分布では確率密度関数が対応する。特に，ポアソン分布や指数分布のように，平均と分散が特定の関係を満たす分布では，個別の統計量を自由に設計することが制約される（Johnson et al., 2005）。

4．範囲

範囲はデータの最大値と最小値の差であり，データの広がりを簡便に表す指標である（Mood et al., 1974）。範囲は外れ値に非常に敏感であるため，四分位範囲や中央値絶対偏差など，より頑健性のある指標の導入が提案されている（Huber, 1981）。

これらは，独立していない（自由に決められない）ものがある。例えば，ポアソン分布では，平均と分散が同一になる。平均と分散が一致するような分布は，現実のデータではそれほど見られないので，ポアソン分布を当てはめるケースは限定されている。また，一様分布では平均や分散は範囲から求められる。値の範囲はデータ数次第で，確率分布によっては影響を及ぼす。以下に示すように一様分布の平均は，区間の中央になり，区間 [a，b] 上の一様分布の分散は，$(a-b)^2/12$ となる。

平均

$$
\begin{aligned}
E[X] &= \int_a^b x f(x) dx = \int_a^b \frac{x}{b-a} dx \\
&= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\
&= \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) \\
&= \frac{a+b}{2}
\end{aligned}
$$

分散

$$
\begin{aligned}
V[X] &= \int_a^b x^2 f(x)dx - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a}\left(\frac{b^3-a^3}{3}\right) - \frac{a^2+2ab+b^2}{4} \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{a^2-2ab+b^2}{12} \\
&= \frac{(a-b)^2}{12}
\end{aligned}
$$

複数の属性における特性は，属性間の関係性を特徴づける様々な統計的指標が存在する。

１．相関関係（相関係数）

２つの属性間の関係の強さを定量化する指標として，一般にピアソンの相関係数が用いられる。これは，２つの属性間の線形関係，すなわち直線的な関係の程度を評価するものである。ただし，同じ相関係数の値であっても，非線形な関係を含めると，様々なデータが存在し得る（Matejka and Fitzmaurice, 2017）。ピアソンの相関係数を用いた２つの変数間の相関関係があるデータ生成方法を説明する。ピアソンの相関係数 $r$ は，２つの変数の線形関係の強さと方向を表す指標である。

２．因果関係（回帰係数）

２つの属性間の因果関係があるデータ生成の方法に回帰モデルを用いたデータ生成の方法がある。因果関係とは，一方の属性の値 $x_i$ が他方の属性の値 $y_i$ に直接的な影響を与える関係である。回帰モデルを用いることで，この因果的な影響を回帰係数 $\beta_1$ や切片 $\beta_0$ から定量的に推定することが可能になる。

３．$x$ の値による $y$ の特性の変化

ある属性 $x$ が特定の閾値を超えるか否かによって，別の属性 $y$ が異なる挙動を示す場合に，その境界における $y$ の不連続な変化を用いて因果効果を推定する回帰不連続デザインがある（Imbens and Lemieux, 2008）。また，属性 $x$ の平均値が小さい群と大きい群において，属性 $y$ の平均値が体系的に異なる場合，これは $x$ の水準によって $y$ の分布的特性が変化していることを示唆する。このような変化が観察される場合には, 単なる相関関係や因果関係ではない構造的な関係性を考察する必要があり, 母集団自体を生成するシミュレーションでは可能となる。

本論文では，シミュレーションに用いる母集団の生成において基本的な要素である単一属性の値の生成について検討する。母集団の特性を再現可能な形で生成することは，分析方法や推定手法の妥当性を評価する上で有効となる。属性の値が他の要因から独立して決定される場合に，そのような単一属性に対する確率分布や統計量（平均，分散など）を適切に反映する生成方法の設計と適用について述べる。

## ４．単一の属性の値の生成

母集団のデータ数を $N$ とする。設定された確率分布に従う属性 $x$ の $N$ 個の値を生成する。$x$ の取り得る値は，連続値ではなく，離散値であり，範囲がある。すなわち，最小値と最大値が定められているものとし，値の種類を $n_x$ とする。確率分布による $x$ の生成方法は次のようになる。

一様分布は，$n_x$ 種類の値を，それぞれ生成する。$i$ 番目の値 $(1 \leq i \leq n_x)$ を持つデータ数は，$N/n_x$ である。$N/n_x$ は，一般には整数とはならないため，調整が必要となる。

$N$ を $n_x$ で割った商を $p$，余りを $q$ とする $(0 \leq q < n_x)$。$x$ の値のうち $n_x - q$ 個の値は $p$ 件，$q$ 個の値は $p+1$ 件のデータとなる。$n_x$ 個の値から $q$ 個の値をランダムに選択し，その値のデータの件数を $p+1$ 件にすることで，全体で $N$ 件となるよう調整する。

正規分布に従う $x$ の値は，平均と分散で特徴付けられる。設定された正規分布に従う $N$ 個の実数を，その値に最も近い $x$ の取り得る値で置き換える。例えば，$x$ の取り得る値が，…，15，20，25，…であれば，18.378は20に置き換えられる。生成された値が，$x$ の範囲内であれば適切だが，分散によっては $x$ の範囲外の値を生成することがある。その場合の処理について検討する。

正規分布は，$-\infty$ から $\infty$ までの実数値をとる連続分布であり，与えられた $x$ の範囲に収まらない値があるため，$x$ の取り得る範囲外のデータを除外し，$x$ の分布を切断正規分布とする。除外されたデータの件数を，範囲内の値に応じた件数に従って分配する。

ポアソン分布のカウントデータにおける一般的な特徴として，観測される事象は稀であることが多く，多くの場合で０の件数が多い。このため，データ分布は０付近に右に長い裾を持つ形状を呈することが一般的である。このような分布特性を持つデータには，ポアソン分布が適用されることが多い。

ポアソン分布は，０からの整数値を表すカウントデータを対象とした分布であり，平均が小さくなることから取り得る値は５単位ではなく，１単位とする。また，範囲内の最大値の確率は小さいため，切断は必要ない場合が多い。切断された場合は，分配しなければならない。

以下では，基本的な確率分布である正規分布を特性とする値の生成について検討する。母集団を用いたシミュレーションでは，属性の値には範囲があるのが一般的である。例えば，街頭調査において交通所要時間の最小は０であり，最大が $\infty$ になることはなく，最大値が決まる。最大値をまとめる場合と切断する場合が考えられる。相関関係や因果関係を分析する上で，最大値をまとめることで分析結果に影響を及ぼす可能性がある。切断する場合には，どの程度切断されているかが問題となり，切断されていることに留意する必要がある。切断が大きすぎれば，正規分布のデータとしては不適切である。切断されている部分が小さければ，真の正規分布に近い分布となる。そのため，$x$ の値を生成する際に，範囲外の確率が十分に小さくなる標準偏差であれば，正規分布の母集団として有効であると考える。例えば，正規分布において標準正規分布表から，平均 $-1.96\times$ 標準偏差～平均 $+1.96\times$ 標準偏差の範囲には，全データの約95％が含まれる。すなわち，約５％が切断される。５％を十分に小さいと考えるならば，この性質を利用して，$x$ の範囲が全データの約95％内に収まるような標準偏差を求めることができ，そのような標準偏差を有効と考える。下限上限各2.5％の範囲を切断したデー

タは，厳密には正規分布ではなく，元の正規分布とは異なる切断正規分布になる。切断される件数は全体の５％未満と少ないが，その件数を範囲内の値に確率に応じて分配する必要がある。

類似した概念に95％信頼区間があるが，正規分布における95％信頼区間は，母平均の推定値がその範囲に含まれる確率が95％であることを意味し，正規分布において平均 −1.96× 標準偏差～平均＋1.96× 標準偏差の範囲には全データの約95％が含まれる考え方とは異なる。

有効な標準偏差の最大値は，値の範囲によって異なる。$x$ の範囲を０から120とする。平均 $\mu$，標準偏差 $\sigma$ の正規分布では，95％のデータが $\mu-1.96\times\sigma\sim\mu+1.96\times\sigma$ の範囲に収まる。平均が60の場合，95％の下限と上限は，

$$\mu-1.96\sigma=0$$
$$\mu+1.96\sigma=120$$

であり

$$\sigma=\frac{60}{1.96}\approx 30.61$$

となる。したがって，$\mu=60$ の時，30.61以下の $\sigma$ で生成した正規分布での $x$ の値が０から120のデータは95％を超えることになり，$x$ が実数を取るのであればこれが有効な標準偏差となる。しかし，$x$ の値が連続値ではなく離散値であれば，30.61を超える分散でも95％を超えるデータが範囲内に収まる。

離散値を取る母集団の生成は次のようになる。

１．正規分布に従う属性の値の生成

平均 $\mu$ および標準偏差 $\sigma$ の正規分布に従う母集団のデータを $N$ 個生成する。

２．離散化および切断処理

生成された正規分布に従う $N$ 個の実数を，範囲内の取り得る値に離散化する。範囲外の値は，データから除外し，正規分布を切断する。

３．範囲外のデータの分配

除外した件数分のデータを範囲内の分布に分配する。分配の手順は以下の通りである。

１．切断後のデータの度数分布を計算
２．この度数分布を確率分布とみなし，範囲外データの件数をその確率に従って分配

以上の手順に従って，標準偏差の違いによる切断データの集計のシミュレーションを行う。正規分布に従う属性 $x$ の値は０から120までの５刻みであるとする。２.では，生成された実数値はそれに近い５の倍数に離散化される。０から120の範囲外の値は，データから除外し，正規分布を切断する。０未満のデータ件数と120を超えるデータ件数の合計が範囲外データ数である。平均を60，標準偏差を範囲内データ95％で求めた30.61，それよりも小さい30，大きい40とした。正規分布で生成した $N=10{,}000$ 件のデータの切断データの集計のシミュレーションをそれぞれ100回行った。

表１は，各標準偏差に対する範囲外データ数の平均と最大値，最小値を示したものである。標準偏

差が30では範囲外データ数の平均は５％の500よりも少なく，最大値も467と小さい。標準偏差が40では範囲外データ数の平均は500よりも多く，最小値も大きい。標準偏差30.61では，範囲外データ数の平均は500に近い値であることが期待されるが，シミュレーションではそれよりも少ない411.56で，最大値も467と小さく最小値は369であった。この現象は生成したデータを５の倍数に離散化するために生じる。

表１．切断データの集計結果

| 標準偏差 | 範囲外データ数の平均 | 範囲外データ数の最大値 | 範囲外データ数の最小値 |
|---|---|---|---|
| 30.61 | 411.56 | 467 | 369 |
| 30.00 | 372.54 | 428 | 328 |
| 40.00 | 1181.96 | 1242 | 1104 |

０未満や120を超える範囲外データ数の平均が500を下回ることについて，考察を行なう。シミュレーションでは，正規分布に従うデータを生成し，５の倍数に離散化した後に範囲外データをカウントする。生成したデータを５の倍数に離散化した際に，例えば，以下のような場合が発生することで，カウントが500件にならない可能性がある。

データが−2.3の場合，離散化で０となり，範囲内に収まる。一方，データが121.8の場合は，離散化で120となり，同様に範囲内に収まる。生成された−2.5から122.5のデータは離散化すると０から120の範囲に収まる。すなわち，離散化の前後で切断される値の数は異なり，離散化により切断される値の数が減少する。離散化を考慮した95％のデータが範囲に収まる標準偏差の最大値は，

$$\mu - 1.96\sigma = -2.5$$
$$\mu + 1.96\sigma = 122.5$$

であることから，

$$\sigma = \frac{62.5}{1.96} \approx 31.88$$

となる。

表２から標準偏差が30.61の場合に離散化前の範囲外データの平均値が499.74となり，標準偏差が31.88の場合に離散化後の範囲外データの平均値が500.03となり，標準偏差は30.61ではなく31.88で１万件の５％となることが確認された。有効な切断を５％未満とするならば，範囲から求められる標準偏差30.61以下ではなく，標準偏差31.88以下を有効とすればよいことが示された。

実際のデータは，必ずしも $x$ の取り得る範囲の中央値が平均となるわけではない。切断される範囲が左右対称でない場合を検討する。$x$ の範囲を０から120，平均が60と50，標準偏差31.88の正規分布で生成した $N=10{,}000$ 件のデータの切断データの集計のシミュレーションを100回行なった。

表２．範囲外データの平均値（100回のシミュレーション）

| 平均 | 標準偏差 | 離散化前の範囲外データ | 離散化後の範囲外データ |
|---|---|---|---|
| 60 | 30.61 | 499.74 | 412.15 |
| 60 | 31.88 | 598.50 | 500.03 |

表３．集計結果

| 項目 | 平均 60 | 平均 50 |
|---|---|---|
| 切断前の平均 | 60.01 | 50.00 |
| 切断前の標準偏差 | 31.88 | 31.88 |
| 切断後の平均 | 60.00 | 52.49 |
| 切断後の標準偏差 | 27.75 | 27.47 |
| 範囲外データ数の平均 | 501.94 | 611.42 |
| 範囲外データ数の最大値 | 552.00 | 672.00 |
| 範囲外データ（<0）の平均件数 | 250.48 | 497.57 |
| 範囲外データ（>120）の平均件数 | 251.46 | 113.85 |

表３は，平均が60の場合の０末満の範囲外データの平均件数は250.48であり，平均が50の場合の０末満の範囲外データの平均件数は497.57となることを示している。平均が中央値でないことから切断されるデータが左右対称でない場合，平均に近いほうが切断される件数が多いことから，95％範囲内に収まるような標準偏差は左右対称の時よりも小さくなる。平均が中央値より外れると，標準偏差を小さくする必要がある。

除外されたデータの件数を，範囲内の値の件数の比率で分配する。$i$ 番目の値の確率を $p_i$ とする。$i$ 番目の値に分配される個数 $N_i$ は，切断した件数 $N_c$ を用いて以下のように計算される。

$$N_i = N_c \times p_i \tag{1}$$

しかし，一般に $N_i$ は整数とはならないため，丸め処理が必要となる。単純に丸めると，分配される総数が $N_c$ と一致しない問題が生じる。

分配の処理の具体例として，$N_c$ と分配総数が一致しない場合を示す。

$N_c = 8$ の場合，

| $p_i$ | $N_c \times p_i$ | $N_i$ |
|---|---|---|
| 0.1 | 0.8 | 1 |
| 0.2 | 1.6 | 2 |
| 0.4 | 3.2 | 3 |
| 0.2 | 1.6 | 2 |
| 0.1 | 0.8 | 1 |
| 合計 | | 9 |

この場合，$N_c=8$ に対し，分配総数は９となり，１つ多くなってしまう。

$N_c=11$ の場合，

| $p_i$ | $N_c \times p_i$ | $N_i$ |
|---|---|---|
| 0.1 | 1.1 | 1 |
| 0.2 | 2.2 | 2 |
| 0.4 | 4.4 | 4 |
| 0.2 | 2.2 | 2 |
| 0.1 | 1.1 | 1 |
| 合計 | | 10 |

この場合，$N_c=11$ に対し，分配総数は10となり，１つ少なくなってしまう。

このように，単純な丸め処理では，$N_c$ と $p_i$ が一致しないため，誤差補正の処理が必要となる。

範囲外のデータの取り扱いについて述べてきたが，分配の処理の煩雑さを回避するために，確率分布に基づいて新たに $N$ 件のデータを生成する方法が有効となる。生成する方法は，以下の通りである。

母集団の大きさを $N$ とし，各データ $(i=1,\ \ldots,\ N)$ の属性 $x$ を以下の手順により生成する。

まず，$i$ 番目のデータに対し0.0以上1.0未満の一様分布に従う浮動小数点数の乱数，$U_T^{(i)} \sim \mathcal{U}(0,\ 1)$ を１つ抽出する。この乱数は，(1)式で用いた確率 $p_i$ を累積した属性 $x$ の累積分布関数 $G_T(j)$ に基づいて，属性 $x$ の値を決定するために用いられる。具体的には，以下の条件を満たす最小の整数 $j$ を選択し，それを $i$ 番目のデータの属性 $x$ の値とする。

$$j=\min\left\{j' \in \{0,\ 1,\ \ldots,\ 120\} \mid U_T^{(i)} < G_T(j')\right\}$$

属性 $x$ に対応する $G_T(j)$ の例を以下に示す。

この手順はすべての $i=1,\ \ldots,\ N$ に対して逐次的に適用され，各反復において１つの属性 $x$ の値が必ず一意に決定される。このため，得られる母集団のデータの件数は常に $N$ となる。また，この手法では一様分布からの乱数生成に失敗が生じることはなく，欠損も発生しない設計となっているため，理論的にも実装上もすべての $i$ についてデータ生成が保証される。

表４．属性 $x$ の累積分布関数 $G_T(j)$ の例

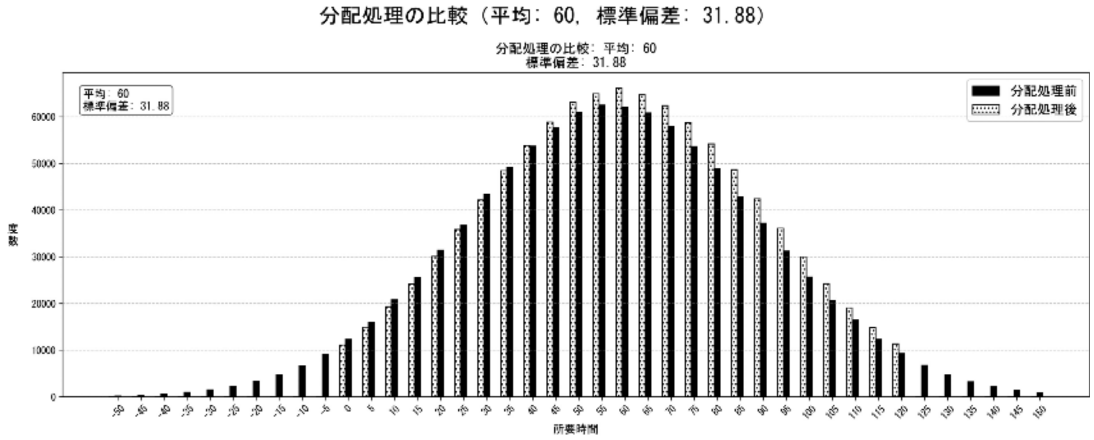
| $j$ | $G_T(j)$ |
|---|---|
| 0 | 0.004 |
| 1 | 0.011 |
| 2 | 0.022 |
| ⋮ | ⋮ |
| 120 | 1.000 |

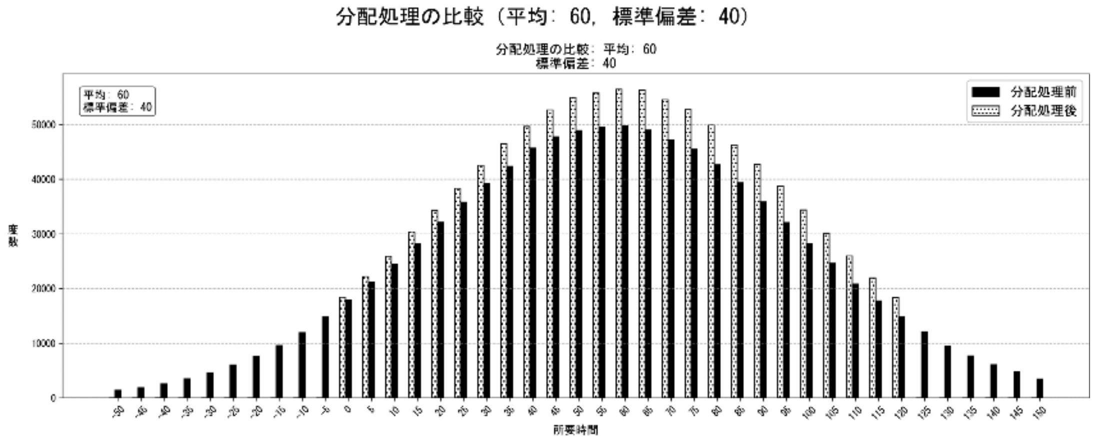
## ５．シミュレーションの例

社会調査における標本抽出による分析を例として，シミュレーションを行なった。具体的には，ある街に人がどのくらいの頻度で出かけるのか，街にたどり着くまでの所要時間との関係を分析することを対象としてシミュレーションする。街頭調査で無作為標本抽出を想定し，その際にサンプリングバイアスがかかるため，標本抽出におけるサンプリングバイアスをシミュレートしている。標本調査による母集団の特性の推定は一般に正規性の仮定に基づいて行なわれる。シミュレーションの枠組みは，この正規性の仮定以外の様々な他の仮定の想定も可能としている。

母集団は，分析の対象となる居住者で，１件のデータを［ID，所要時間（分)］とする。所要時間（分）は普段選択する交通手段による時間であり，０分から120分（５分間隔）とする。本論文では，母集団の単一属性の値の生成を検討しているので，所要時間（分）を正規分布に基づいて100万件の母集団を生成している。また，確率分布にかく乱を与えた単一属性の値の分布を示す。シミュレーションの母集団は，属性の値を確率分布に基づいて生成しているが，完全に確率分布に従うことは現実的ではない。所要時間の確率分布と所要時間ごとの出向頻度の確率分布に乱数を加え，確率分布をかく乱した母集団を生成する。例えば，確率分布に一様なランダムな値を加え，合計が１となるように確率分布を再計算することで実現している。加える値 $\theta$ の上限を変えることで母集団のかく乱のレベルを設定することができる。加える値 $\theta$ の上限を0.05としたものである。

図１は，平均60，標準偏差をそれぞれ31.88，40，50とする正規分布から生成されたデータに対し，累積分布関数を用いた分配による度数分布を示している。標準偏差が31.88の場合，０から120の範囲外にあるデータ数は50142件であったが，単純に丸め処理を行うことで50130件とされ，12件の誤差が生じた。そのため，12件の再分配が必要となる。標準偏差が40の場合は13件，50の場合は14件の誤差が生じた。このように，範囲外データを単純に丸めて分配する方法では，誤差が生じることが確認できる。図２は，平均50，標準偏差を31.88，40，50とした場合の比較を示している。標準偏差が31.88では11件，40では15件，50では９件の誤差となった。これらの誤差は，０から120までを５単位ずつ区切った25区分の度数分布の比率で分配された結果であり，誤差件数の上限は，区分の数である25件となる。また，標準偏差が大きくなるにつれて範囲外データの件数も増加するため，分配されたデータ数が大きくなり，正規分布から乖離する傾向があることが示された。



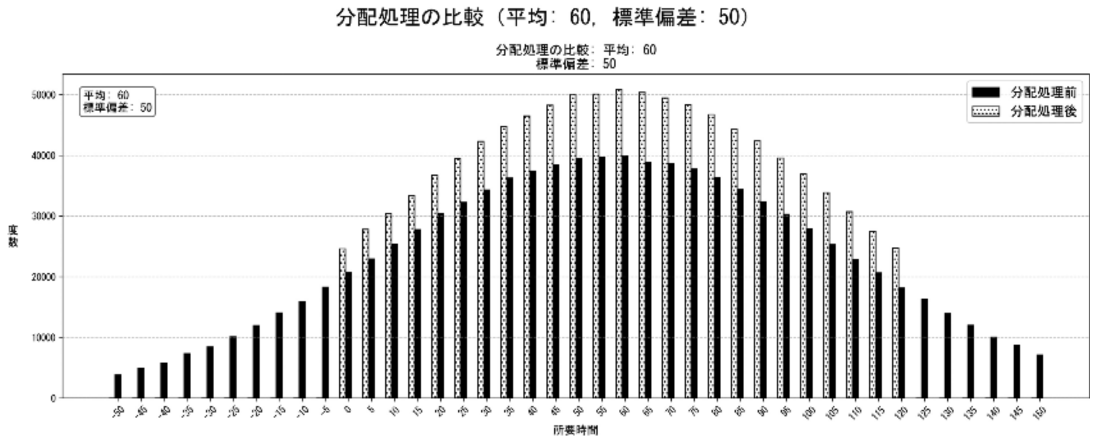




図１．分配処理の比較：平均60で標準偏差31.88，40，50

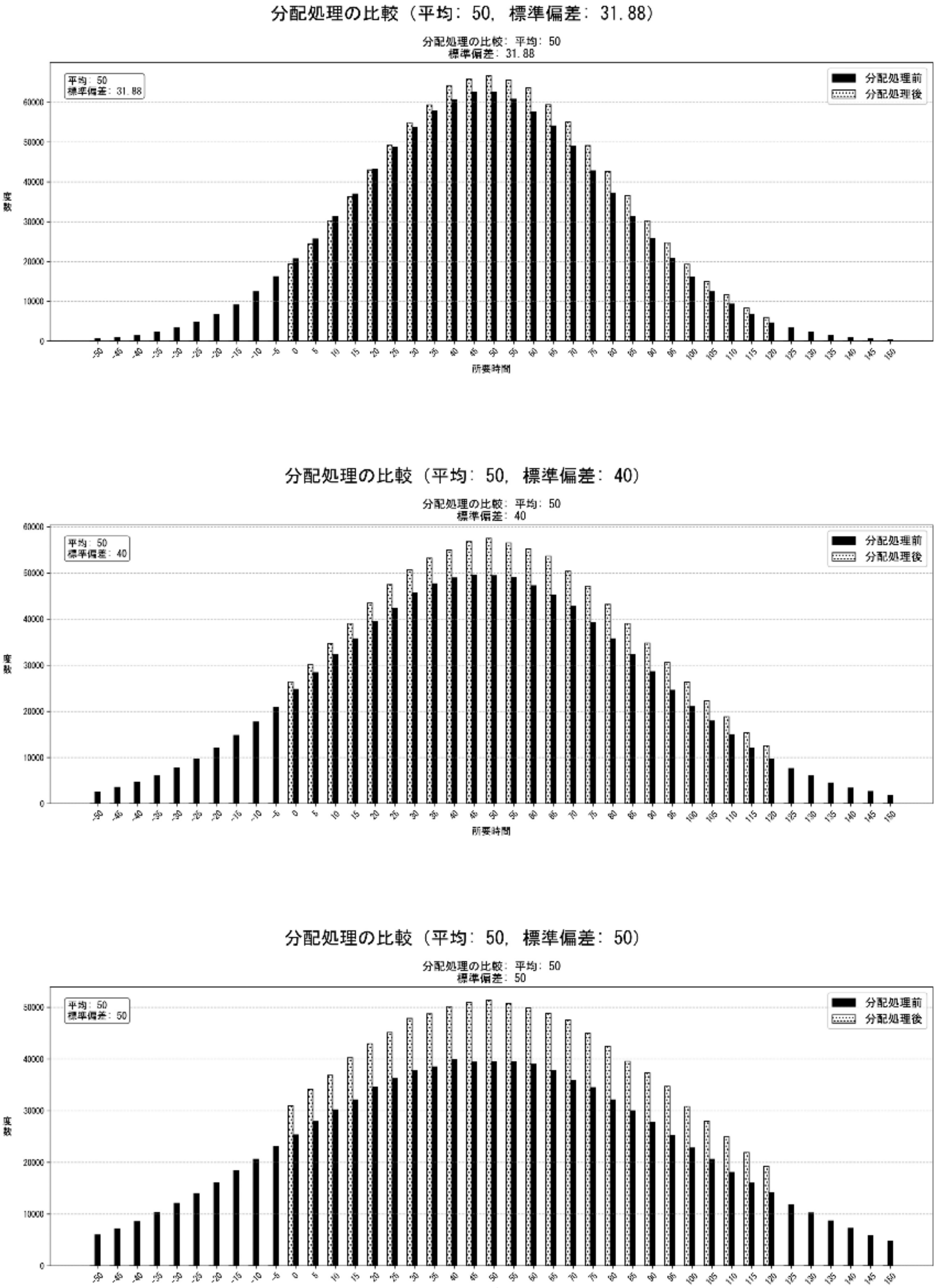



図２．分配処理の比較：平均50で標準偏差31.88, 40, 50

図３は，平均60，標準偏差31.88の正規分布に基づく母集団に対して，かく乱レベルをそれぞれ0.00，0.01，0.02，0.03，0.04，0.05に設定した場合に得られた，100個の母集団のうちの一例を示している。

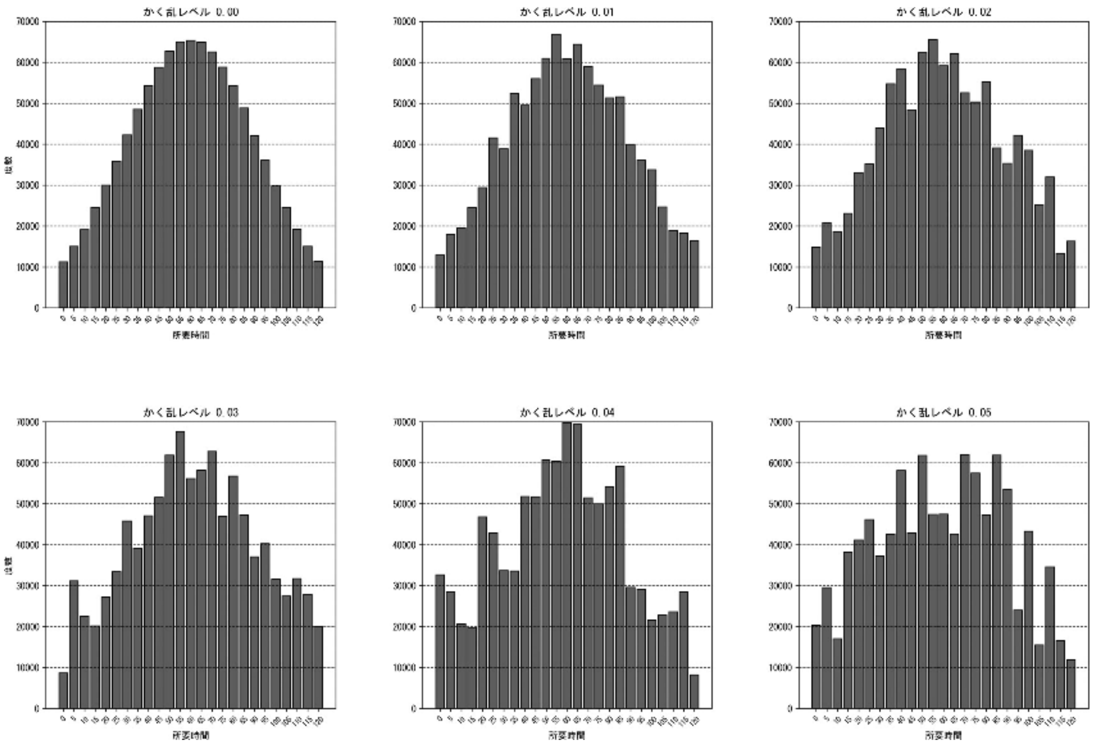


図３．各かく乱レベルの母集団の所要時間

範囲外データの分配に伴う誤差の問題点を明らかにし，累積分布関数を用いた方法により，100万件のデータとなる母集団の生成が可能であることを確認した。所要時間の確率分布に対して，加える値$\theta$の上限を変化させた一様分布に基づく乱数を加えることで，確率分布をかく乱した母集団も生成でき，現実的なばらつきを含むシミュレーション用の母集団が構築された。

## ６．結論と今後の課題

本論文は，シミュレーションのための母集団の生成について述べ，基本的な母集団の生成として単一属性の値の生成を対象とした。標本を用いた分析や推定の結果と母集団での真の値の比較を行うシミュレーションの概要について述べ，分析方法や推定手法を評価するためにシミュレーションに用いる母集団に求められる特性を検討した。特性には属性の確率分布や平均，分散などがあり，単一属性のみで特性が決定される場合や他の属性の値との関係の中で決定される場合がある。今後の課題は，複数の属性についての母集団の生成方法を検討し，シミュレーションによる母集団の真の値を用いて分析方法や推定手法の有用性を検証することである。

## 参考文献

Casella, G., Berger, R. L., Statistical Inference (2nd ed.), CRC press, 2024.

Huber, P. J., Ronchetti, E. M. (2011). Robust statistics. John Wiley & Sons, 2011.

Imbens, G. W., Lemieux, T., “Regression Discontinuity Designs: A Guide to Practice,” Journal of Econometrics, Vol. 142, No. 2, pp. 615-635, 2008.

Johnson, N. L., Kemp, A. W., Kotz, S., Univariate Discrete Distributions (3rd ed.), John Wiley & Sons, 2005.

Mood, A. M., Graybill, F. A., Boes, D. C., Introduction to the Theory of Statistics (3rd ed.), New York, McGraw-Hill, 1974.

Matejka, J., Fitzmaurice, G., “Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing,” Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1290-1294, 2017.

Wasserman, L., All of Statistics: A Concise Course in Statistical Inference, New York, Springer, 2013.