MAT120：数量的推論　第13講ハンドアウト

統計入門

前回の講義では確率を扱い，とくに二項実験と確率分布に焦点を当てました。今回の講義では統計の単元に入り，統計とは何か，なぜ標本抽出が重要なのか，そしてどのようにして偏りがデータに入り込むのかを考えます。また，記述統計の二つの中核的な道具，すなわち中心傾向（真ん中を見つけること）と分散（ばらつきを測ること）も作っていきます。最後に，これらの考え方をハリボーのデータセットに適用し，一つの具体例として見ていきます。次回は正規分布に進み，その次に仮説検定を学びます。

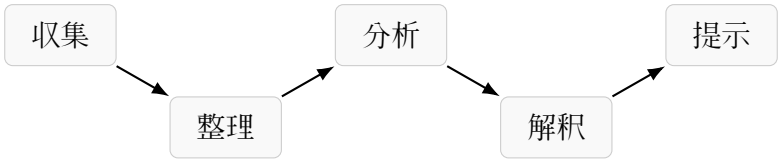
1　統計とは何か

統計学とは，データを数学的に研究する学問です。データとは，ひとまず「現実世界の何かから集められた情報の集まり」と考えてよいでしょう。したがって，統計学の主な仕事は，この情報を調べ，そこにどのような構造があるのかを見つけ出し，その構造を用いて現実世界のさまざまな事柄について意思決定を助けることにあります。統計は科学や産業のいたるところに存在しています。鳥の渡りのパターンから熱の基本理論まで，またマイクロチップの製造工程から経済という複雑に織り込まれた仕組みに至るまで，あらゆるところで統計は使われています。入門レベルにおいて，統計は学ぶべき最も有用で重要な数学的道具の一つだと言ってよいでしょう。

これから3 回の講義を使って，統計の基本概念をいくつか導入していきます。今日はそのいちばん最初として，統計の本質とは何か，そしてデータセットがどのような基本構造を示しうるのか，という問いから始めます。次の講義では，ある巧妙な数学的仕組みに基づいて，統計がどのようにシステムに対する情報に基づいた検定 を行うために使われるのかを見ていきます。

1.1　統計研究の流れ

よい統計研究は，現実世界の文脈と数学的手法とを組み合わせて，あるシステムについて新しい情報を見つけ出します。統計研究の基本的な形の一つは，次のような流れです。



各段階の意味

- **収集** では，現実世界のあるシステムを測定し，そこからデータを集めます。
- **整理** では，データを使いやすい形に並べ直します。
- **分析** では，考え抜かれた数学を用いて，データセットの中の意味ある特徴を記述します。
- **解釈** では，得られた結果が現実世界の文脈で何を意味するのかを説明します。
- **提示** では，結果を他の人にも分かりやすい形で簡潔にまとめます。

この流れが，現実世界に関する考察と数学的道具との洗練された混合であることに注目してください。ここにこそ統計の核心があります。統計は単なる数学理論ではなく，統計的な知識には本質的に現実世界が関わっているのです。

1.2　統計の実際

統計はしばしば退屈で味気ないものに感じられます。なぜなら，どうしても公式中心に見えてしまうからです。実際，統計で使われる多くの公式の背後には深い数学理論がありますが，実践的な統計ではその理論そのものを直接使うことはあまりありません。数学的な洗練さという観点から見ると，統計はとても単純であり，もっともよく使われる公式の多くが基本的な四則

演算しか用いないため，退屈に感じられることがあります。

しかし，まったく同じ理由で，統計は非常に面白くもあります。もっとも基本的な数学操作から出発して，統計はデータの全体的な構造を記述するための巧妙な公式を数多く生み出します。統計の面白さは，これほど強力で有用な道具が，これほど素朴な出発点から生まれてくるところにあります。統計の実用性はいくら強調してもしすぎることはありません。世界は情報でできており，あなたや社会に関する多くの性質はデータの形に変換して分析することができるのです。

1.2.1 記述統計と推測統計

実際には，統計には二つの大きな柱があります。それが記述統計と推測統計です。

- **記述統計** は，集めたデータそのものを記述することを扱います。ここでは，ニュース記事や科学論文で見るようなグラフや図表のような視覚的な記述を思い浮かべがちです。しかし実際には，平均や変動のような数 によってデータを記述することも多く，さらに回帰のようなより高度な手法も含まれます。
- **推測統計** は，データに基づいてどのような結論に到達できるかを扱います。統計研究はしばしば手元の資源によって制約されており，「完全な」データを集めることは高価すぎたり，時間がかかりすぎたり，あるいはそもそも不可能であったりします。そこで推測統計では，少数の情報からより大きなシステムを検討するための巧妙な数学的道具を用います。推測統計には，仮説検定や推定といった手法が含まれます。

1.3 データとは何か

現実には，あるシステムを完全に記述するようなデータセットがいつでも手に入るわけではありません。ですから，私たちは手元にあるデータで満足しなければなりません。統計は，全体の母集団を完全には知らなくても，その母集団について何かを知るための道具を作ります。

私たちのデータの質は，それをどのように集めたかに左右されます。望ましくない効果が分析の中にうっかり組み込まれてしまうこともあり，それを見つけて最小限に抑えるのが私たちの仕事です。

1.3.1 データの種類

データは大きく二つの種類に分けられます。量的データとカテゴリーデータです。

- **量的データ** は数値として表されるデータです。
- **カテゴリーデータ** は数値ではないデータです。

量的データには，分析に役立つ数学的構造があります。一方，カテゴリーデータは，どちらかといえば名前の一覧に近いものです。

1.3.2 量的データの表し方

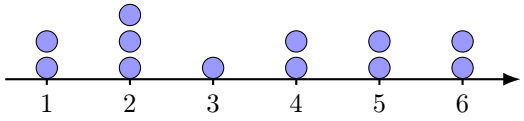
数値データはしばしば$(x_1, x_2, \ldots, x_n)$ という形で書きます。この記法は，これまで見てきた集

合の波括弧{ と} によく似ています。しかし，重要な違いが一つあります。データを書くときには，データの順序 も考えるからです。そのため，データは丸括弧で書きます。これらの括弧は，データがどの順番で並んでいるかも追跡することを表しています。[1]

データの例： $(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$.

同じ値が繰り返し現れる様子を視覚化するのに役立つのが，次のような図です。この図では，数値データを数直線上に置かれた小さな玉として考えます。同じ数値をもつデータが複数あるときには，その上に玉を積み上げていきます。そうすると，高さが頻度を表すことになります。完全を期すために，図の上には常にデータセットそのものも書いておきます。

データ: $(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$



もちろん，上の図はデータの順序そのものを表してはいません。しかし，多くの公式は加法や乗法のような基本的な演算しか使わず，それらは可換なので，順序は必ずしもそれほど重要ではないことが多いのです。

2 標本抽出

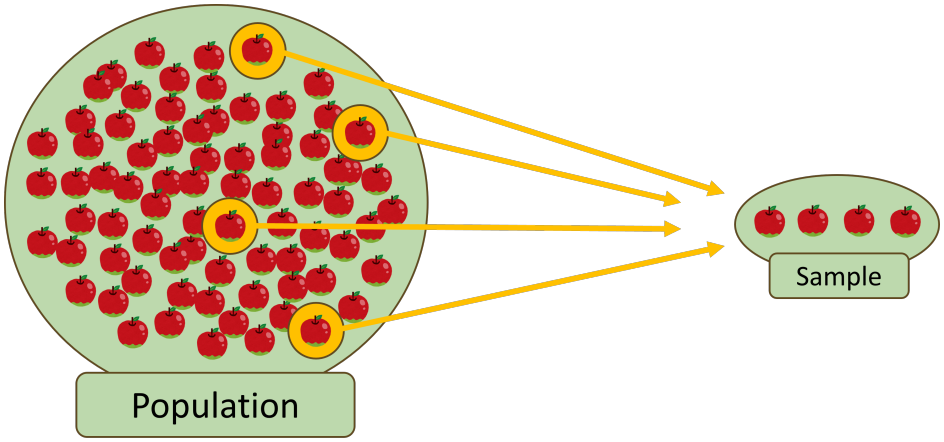
2.1 基本定義

研究したい対象の集まりは，しばしば大きすぎて直接には調べられません。そこで実際には，その中からより小さな部分集合を取り出して，それを調べます。この実践的事情から，次の用語が導かれます。

定義

- **母集団：** 研究したい対象すべての集合。
- **標本：** 母集団から選ばれた部分集合。
- **母数：** 母集団全体についての事実・測定値・量。
- **標本統計量：** 標本から計算される事実・測定値・量。

[1] 実はこの記法は，第7講で線形モデルを扱ったときにも見ました。そこでは方程式の解を「順序つき組」として表し，(x, y) の形で書いていました。データセットの場合は，この記法を一般化しているわけです。



母集団から標本を得る過程を標本抽出（sampling）といいます。標本抽出は統計においてきわめて重要です。というのも，現実には母集団全体にアクセスできることはほとんどないからです。しかし，この過程では十分に注意しなければなりません。というのは，データの取り方によっては，母集団には本当は存在しない物語をうっかり作り出してしまうことがあるからです。統計では，このような誤りを偏り（bias）と呼びます。

2.2　具体例：東京での標本抽出

Jenny が，東京に住む人々の平均給与を知るために統計研究を行いたいとしましょう。Jenny は市役所で働いているわけではないので，東京の母集団についての国勢調査データにはアクセスできません。その代わりに，標本を取らなければなりません。これから，Jenny が取りうるいくつかの標本抽出法を考えます。ただしその前に，重要な点を一つ述べておきます。次の二つの方法を考えてみましょう。

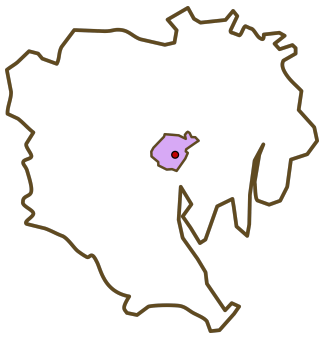
二つの標本抽出法

1. **路上調査：** Jenny が自分の近所を歩き回り，通りで30 人に声をかけて収入を尋ねる。
2. **戸別訪問調査：** Jenny が自分の近所の家を一軒ずつ訪ねて，収入を尋ねる。

一見すると，この二つの方法には大きな違いがないように感じられるかもしれません。しかし，データ収集には人間的要素 があることを忘れてはいけません。たとえば，知らない人が突然家に来てお金のことを尋ねれば，脅威を感じる人もいるかもしれません。そのような状況では，Jenny が獲物を探している泥棒かもしれないと思い，自分の資産を少なく見せるように嘘をつく人もいるかもしれません。すると，結果として得られた標本データの平均給与は，路上調査の平均より低くなるかもしれません。

2.3　ありうる標本抽出法

東京という設定のもとで，Jenny が取りうる標本抽出法として，次の四つを考えることができます。

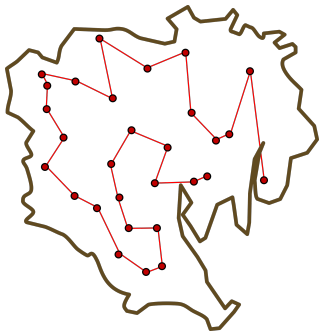


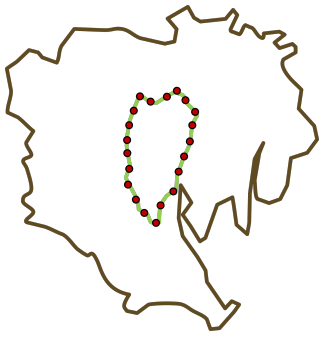
方法1： 千代田区の自宅周辺から30 人を選ぶ。

多様性の欠如： 自分の住む近所から標本を取るのは，たしかに便利です。しかし，その数字は東京全体を表していないかもしれません。地域によって，裕福さや貧しさ，年齢層，職業の種類は異なります。ですから，たとえ30 人に尋ねたとしても，東京全体ではなく，その近所の特徴を測ってしまう危険があります。千代田区の場合，この標本は平均給与を過大評価する可能性が高いでしょう。

方法2： 東京のどこからでも無作為に30 人を選ぶ。

実務上の困難： 都内全域から無作為に人を選べば偏りがなさそうに思えますが，実際にそれを行うには多くの費用と時間がかかります。無作為に選ばれた各地域へ移動するだけでも大変ですし，そこで誰に声をかけるかによっても別の偏りが入り込むかもしれません。それでも，紙の上では偏りがないように見える計画でも，資源の制約によって現実にはうまくいかないことがあります。



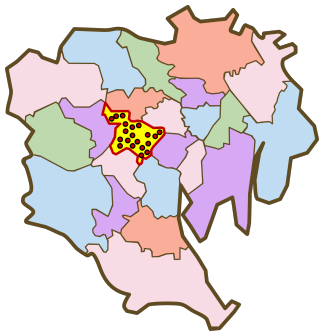


方法3： 山手線に乗り，各30 駅で降りて，それぞれの駅で一人ずつに尋ねる。

アクセス可能性の偏り： 「各駅で一人ずつ」という標本抽出には，東京に住んでいない観光客や外部の人が含まれるかもしれません。また，観光地の駅，通勤の結節点，ビジネス街などに偏る危険もあり，そのためある種の働き手に標本が偏るかもしれません。言い換えれば，この標本は，東京市民のある特定のタイプ，すなわち山手線周辺で働いたり移動したりする人々に偏る可能性があります。

方法4： 東京23 区のうち一つを無作為に選び，その区の中からさらに無作為に30 人を選ぶ。

多様性の欠如： これは東京全域を回るよりは便利であり，Jenny の地元である千代田区だけを使うよりは偏りが少ないですが，それでもなお，この種の標本には多様性の不足が残ります。無作為性は，母集団の多様性が小さな標本の中に完全に表れることを保証しません。無作為性が保証するのは，選び方のルールが偏っていないということだけです。



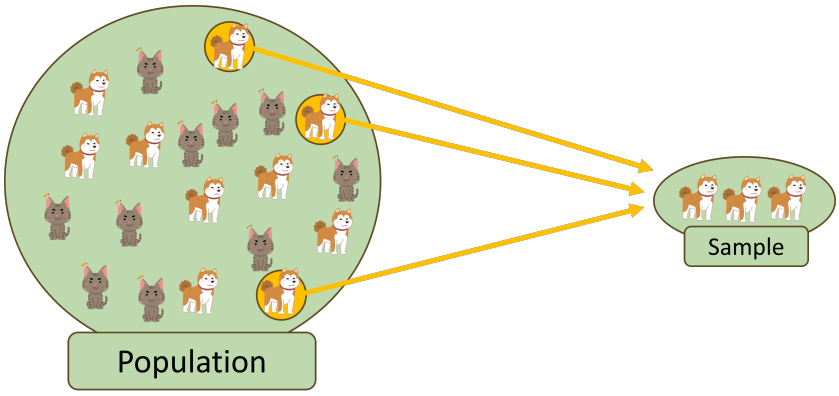
では，ここで最善の標本抽出法とは何でしょうか。実は，それはJenny が何をしたいのかによります。実際には，唯一の「最善の方法」があるとは限りません。標本抽出とは，しばしば資源と誤差とのあいだのバランスであり，ときには妥協しなければならないのです。理想的には，研究者はこれらの偏りを理解しているべきであり，結果を伝えるときにはそれをできるだけ明示的に述べるべきです。

2.4 単純無作為標本

定義：単純無作為標本

大きさn の単純無作為標本とは，母集団の各要素が選ばれる機会を等しくもつように無作為に選ばれた部分集合のことです。同値な言い方をすると，復元しない抽出では，大きさn のすべての部分集合が同じ確率で選ばれます。

選び方が無作為であったとしても，得られた標本が母集団を正直に表さない「はずれ標本」になることはあります。たとえば，ある母集団に猫が10 匹，犬が10 匹いるとして，大きさ3 の無作為標本を取ると，3 匹とも犬になってしまい，猫が過小に表される可能性があります。前節の第四の方法と同じく，単純無作為標本は，どの標本にも特別な選び方のルールがないという意味では偏っていません。しかし，それでも各標本が母集団の多様性を正直に表していることまでは保証しません。



2.5 標本抽出における偏りの例

よくある三つの偏り

- **生存者バイアス：** 「昔の建築はとても美しく，現代の建物はとても醜い。」これは，過去のもののうち，特に優れた例だけが記憶され，保存されているときに起こりえます。
- **選択バイアス：** 「歌舞伎町でみんなに東京は騒がしくて混沌としているかと聞いたら，全員がそうだと答えた。だから東京は騒がしくて混沌としている。」歌舞伎町は東京全体の無作為標本ではありません。
- **資金提供バイアス（利益相反）：** ある研究が，特定の結論によって利益を得る集団から資金提供を受けている場合，結果に偏りが生じることがあります。

以上はすべて標本抽出バイアス の例です。つまり，本来なら母集団には存在しない結論を，標本の選び方のせいでうっかり示してしまうことです。生存者バイアスの例では，話し手は現在

も残っている 過去の建物だけを見て結論を出しています。しかし実際には，醜い建物は取り壊されて新しい重要な計画に置き換えられ，美しい歴史的建築物はその美的価値のために保存されたのかもしれません。もし魔法で1700 年代に戻れたとしても，今なお残っている建物すべてと同じくらい美しい建物ばかりが当時存在したわけでは，おそらくないでしょう。

3　中心傾向

3.1　中心傾向の意味

たくさんのデータを集めたとき，私たちはしばしばそれを一つの代表的な数で要約したいと思います。その一つの方法が，データの「真ん中」がどこにあるかを記述することです。これこそが中心傾向の意図です。すなわち，中心傾向の尺度 とは，データセット全体を入力として受け取り，ある種の「真ん中」を表す一つの数を出力する規則のことです。

実際のところ，データセットの「真ん中」には唯一完全な概念があるわけではありません。むしろ，いくつかの競合する概念があり，文脈によって有用さが異なります。たとえば，データに極端な外れ値（非常に大きい値や非常に小さい値）が含まれると，ある中心傾向の尺度は大きく動きますが，別の尺度はほとんど動かないことがあります。

3.2　総和記号

中心傾向を議論する前に，新しい記法として総和記号を導入しておきます。これは，たくさんの数の和をコンパクトに書く方法です。記号Σ を，大きな和の省略記号として使います。

総和記号

$$\sum_{i=1}^{n} x_i \quad \text{は} \quad x_1 + x_2 + x_3 + \cdots + x_n \text{を意味する。}$$

文字i は添字 であり，どの項を足しているのかを表します。1 とn は，数え始めるところと終わるところを表します。ここでn は，データセット中のデータ点の個数です。

3.3　よくある三つの「平均」：平均値・中央値・最頻値

日常語において，「average（平均）」という言葉は曖昧で，さまざまな意味を持ちます。初等統計でよく使う平均には，平均値，中央値，最頻値の三つがあります。これら三つは，データセットの真ん中を記述する異なる方法であり，それぞれ少しずつ違う問いに答えています。

- **平均値（mean）：** データの「つり合った真ん中」。各データ点を数直線上のおもりだと考えると，平均値はそのつり合い点です。すべての値を使うので，極端な外れ値に敏感です。
- **中央値（median）：** 並べ替えたあとで真ん中にある値。主にデータの順序 に依存するので，外れ値の影響を受けにくいです。データが歪んでいるときには，「典型的な値」として役立つことが多いです。
- **最頻値（mode）：** 最も頻繁に現れる値。これは特にカテゴリーデータ（たとえば「いちばん多いお菓子の種類」）に有用です。数値データでも最頻値は有益なことがありますが，値の重なり方によっては誤解を招くこともあります。

どれを使うべきか（経験則）

- 合計やつり合いが重要なら（たとえば費用を公平に分けるなど），**平均値** が自然です。
- 極端な外れ値を無視した「典型値」がほしいなら（たとえば給与など），**中央値** のほうがよいことが多いです。
- 最もよく現れるカテゴリや値を知りたいなら，**最頻値** を使います。

3.4 中央値

中央値とは，データを二つに分ける値のことです（半分のデータがそれ以下にあり，残り半分がそれ以上にあります）。

Step 1: すべての値を小さい順に並べる。
Step 2: n が奇数なら，真ん中の値（第$(n+1)/2$ 項）を取る。
Step 3: n が偶数なら，第$n/2$ 項と第$(n/2+1)$ 項のちょうど真ん中の数を取る。

たとえば，データセット$(1,1,3,4,5)$ の中央値は3 です。点の数が偶数のデータセット，たとえば$(1,1,3,4,5,5)$ では，中央値は真ん中の二つの値の中間になります（この場合は3.5 です）。

3.5 平均値

標本$x_1, x_2, \ldots, x_n$ に対して，標本平均は$\bar{x}$ と書き，次で定義されます。

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

平均値はつり合い点のように働きます。平均より大きい値は上に引っ張り，平均より小さい値は下に引っ張ります。そして，平均値はそれらがちょうどつり合う点です。データセットの一種の「幾何学的中心」とみなすこともできます。

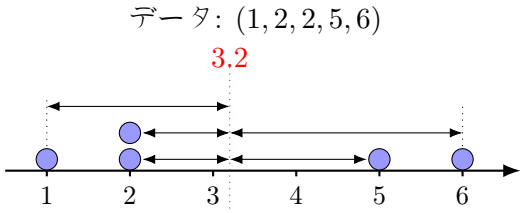
平均値についての便利な事実

データから平均値までのずれを全部足し合わせると，常に0 になります。

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

つまり，平均より「上」にあるデータと「下」にあるデータとは，全体として互いに打ち消し合います。

この事実を実際に見てみるために，上の図で距離を測ってみましょう。



平均値（3.2）の左側にある三本の矢印の長さの和は$1.2 + 1.2 + 2.2 = 4.6$ であり，右側にある二本の矢印の長さの和も$1.8 + 2.8 = 4.6$ になります。

3.6 最頻値

最頻値とは，最も頻繁に現れる値のことです。もし最も頻繁に現れる値が一つだけ なら，その値が最頻値です。複数の値が同じ最大頻度で並ぶときには，データは多峰的（multi-modal）であると言います。教科書によっては，この場合には単に最頻値は存在しない とするものもあります。

演習

データセット$(1, 1, 2, 3, 4, 5)$ を考えなさい。

(a) 平均値はいくつか。
(b) 中央値はいくつか。
(c) 最頻値はいくつか。

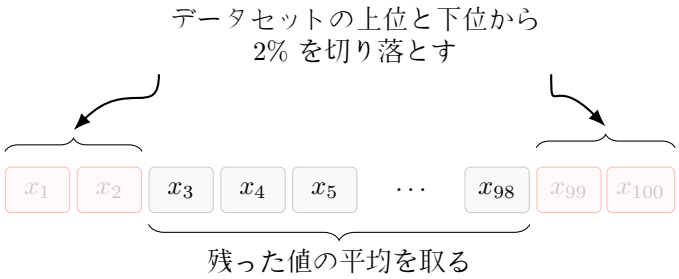
解答

(a) 和は$1 + 1 + 2 + 3 + 4 + 5 = 16$ で，$n = 6$ なので，平均値は$16/6 = 8/3 \approx 2.67$。
(b) 並べ替えた列はすでに$(1, 1, 2, 3, 4, 5)$ です。$n = 6$ は偶数なので，中央値は第3 項と第4 項のちょうど中間であり，$(2 + 3)/2 = 2.5$。
(c) 値1 は二回現れ，他の値は一回ずつしか現れないので，最頻値は（唯一の）1。

3.7 トリム平均

データセットの上位と下位の$x\%$ を切り落として，残りの平均を取ることで「$x\%$ トリム平均」を作ることができます。ここでの意図は，平均値のもつ幾何学的な「つり合い」の考え方を保ちつつ，極端な外れ値の影響を小さくすることにあります。

たとえば，サイズ100 の順序づけられたデータセットがあるとすると，2% トリム平均では両端から2 個ずつ値を取り除き，残った96 個の値で平均を取ります。これはおおよそ次の図のように表されます。

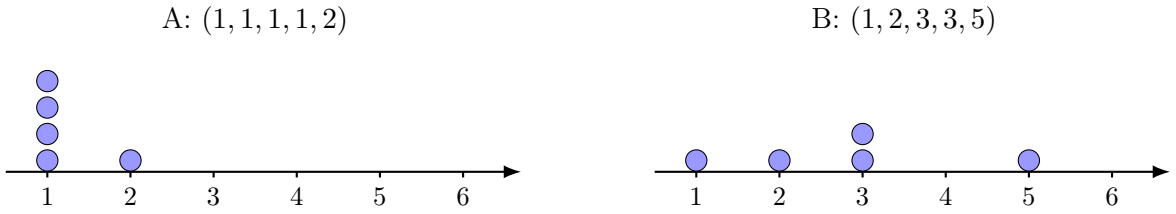


一般に，トリム平均は，データに人間的要素が入る場面でかなり有用です。たとえばスポーツの採点では，審判団が数値評価によって選手を比較することがあります。多くの場合，選手の得点は，最も高い評価と最も低い評価を先に除いてから平均を取るトリム平均で計算されます。これにより結果はより客観的になり，誤りを犯しやすい審判や個人的な好みに左右されやすい審判の影響を小さくできます。

4 変動

中心傾向は，おおまかに言えばデータが「どこにあるか」を教えてくれますが，それがどれほど広がっているかまでは教えてくれません。真ん中の位置はだいたい同じでも，見た目がまったく異なる二つのデータセットを簡単に作ることができます。そこで次の考え方が必要になります。それが変動（variation），あるいはばらつき（spread）です。

たとえば，次の二つを比べてみましょう。



直観的には，データセットB のほうがより広がっています（値が真ん中からより遠くまで散らばっています）。

4.1 範囲

範囲（range）とは，最大値と最小値の差です。

$$\text{範囲} = \max(\text{データ}) - \min(\text{データ}).$$

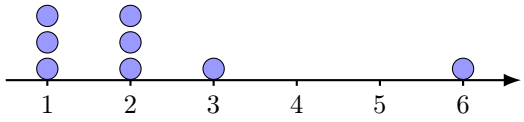
これはばらつきについての手早い最初の記述です。範囲が小さいとばらつきは小さそうであり，範囲が大きいとばらつきは大きそうに見えます。

4.1.1 範囲の問題点

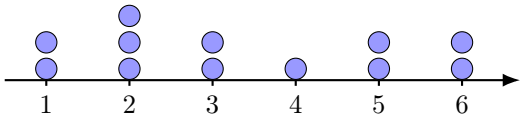
範囲は二つの数（最小値と最大値）しか使わないので，ある意味では，データセットの大部分

を見ていない公式です。また，範囲は外れ値にとても敏感です。極端な値が一つあるだけで範囲はとても大きくなってしまいます。次の二つのデータセットを考えてみましょう。

A: $(1, 1, 1, 2, 2, 2, 3, 6)$



B: $(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$



データセットA には，値6 が外れ値として現れていますが，B にはそうした外れ値はありません。範囲だけで見れば，この二つのデータセットは同じだけ広がっていることになります。しかし，見た目には，A では大多数の点が一か所に固まっており，B ではそうなっていません。ですから，少なくともある意味では，B のほうがA より大きなばらつきをもつべきです。

データのばらつきをもっと丁寧に捉えるには，最小値と最大値だけでなく，すべての データ点を使う統計量が必要です。そこで自然に分散へと進むことになります。

4.2 分散

分散は，範囲よりも包括的で有益なばらつきの尺度です。分散は，平均からの距離の二乗の平均を見ることで，データの広がりを測ります。これらの二乗距離が大きくなるほど，データはより広がっていることになります。

便利な考え方としては次の通りです。

> データ点が平均の近くに集まっていれば，分散は小さい。データ点が平均から遠くにある傾向が強ければ，分散は大きい。

4.2.1 幾何学的な基本アイデア

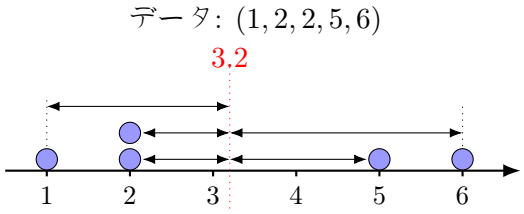
各データ点x について，それが平均$\bar{x}$ からどれだけ離れているかを見ます。これは単に差$(x - \bar{x})$ です。もし生の偏差$(x - \bar{x})$ をそのまま平均しようとすると，正の値と負の値が打ち消し合ってしまい，結果は0 になってしまいます。この望ましくない打ち消しを避けるために，偏差を二乗します。

したがって，基本的な手順は次の通りです。

Step 1: まず平均$\bar{x}$ を求める。
Step 2: 各偏差$(x_i - \bar{x})$ を計算する。
Step 3: 各偏差を二乗する。つまり$(x_i - \bar{x})^2$ を計算する（こうするとすべて正になり，大きな偏差がより強く反映される）。
Step 4: 二乗偏差の平均を取る。

もう一度強調すると，上のStep 3 はとても重要です。一般に，データ点が平均の右側にあるか左側にあるかによって，偏差は正にも負にもなります。

このことを実際に見るために，先ほどの図で距離を測ってみましょう。



すでに見たように，平均の左側にある矢印の和は，右側にある矢印の和と，符号が反対なだけで大きさは同じなので，打ち消し合います。しかし，これらを二乗すると，この打ち消し効果はなくなります。

4.2.2 標本分散の公式

標本$x_1, x_2, \ldots, x_n$ の平均が$\bar{x}$ であるとき，標本分散 はs^2 と書き，次で定義されます。

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

分母はn ではなく$n-1$ です。これは通常，標本を使って母集団のばらつきを推定しているからです。この講義では，主としてこの公式を知り，どう解釈するかを理解していれば十分です。すなわち，s^2 が大きいほど，平均のまわりのばらつきが大きいということです。

実際には，分散はたいてい電卓やコンピュータソフトで計算します。しかし，この公式が何を測っているのかを理解することは大切です。とはいえ，計算の感覚をつかめるように，その手順を簡単に見ておきましょう。

例 データセット$(1, 1, 2, 3)$ を考えます。分散を計算するために，まず平均を求めます。

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

次に，各データ点が平均からどれだけ離れているかを求め，それを二乗します。これは表にするといちばん分かりやすいです。

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

最後に，右端の列の値をすべて足して，$n-1=3$ で割れば分散が得られます。

$$s^2 = \frac{0.5625 + 0.5625 + 0.0625 + 1.5625}{3} = \frac{2.75}{3} \approx 0.917.$$

4.3 標準偏差

標準偏差 とは，分散の平方根です。

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

平方根を取るのは，分散が「二乗単位」で測られているからです（たとえばデータが円なら，分散は円^2 になります）。標準偏差を取ると元の単位に戻るので，平均からの典型的な距離として解釈しやすくなります。このことは次回の講義で非常に重要になります。

4.4 記号に関する重要な注意（標本と母集団）

先に述べたように，推測統計の中心的な問いは，標本統計量から母集団の母数を決定することです。このため，標本統計量と母数とを明確に区別することが非常に大切です。その一環として，標本に関する数学と母集団に関する数学とでは，異なる記号を使うのがよい習慣です。技術的な理由から，標本に対する公式と母集団に対する公式は少し異なり，また非常によく似た概念であっても別の記号を使います。以下にまとめます。

標本統計量と母集団の母数

- **標本統計量：** $\bar{x}$（平均），s^2（分散），s（標準偏差）。
- **母集団の母数：** μ（平均），σ^2（分散），σ（標準偏差）。

4.5 変動係数

異なる尺度をもつデータセットどうしでばらつきを比べたいことがあります。たとえば，標準偏差2 は平均が4 なら大きいですが，平均が200 なら小さいです。そこで，標準偏差を平均で割ることで，ばらつきの大きさを「正規化」することができます。標準偏差と平均の比を取ると，単位のない数が得られます。これはばらつきの尺度であり，変動係数 と呼ばれ，次のように定義されます。

$$\text{母集団: } \mathrm{CV} = \frac{\sigma}{\mu} \qquad \text{and} \qquad \text{標本: } \mathrm{CV} = \frac{s}{\bar{x}}.$$

CV が大きいほど，「値の典型的な大きさに比べてばらつきが大きい」ことを意味します。

4.6 中心傾向と分散を「地図」として見る

中心傾向はデータがどこにあるか を教え，分散（あるいは標準偏差）はそれがどれほど広く散らばっているか を教えます。したがって，平均と標準偏差は，データを地図のように記述する二つの方法です。

$$\text{平均} = \text{中心がどこにあるか}, \qquad \text{標準偏差} = \text{データが中心からどれくらい離れているか}.$$

次回の講義（正規分布）では，この「地図」という考え方がさらに具体的になります。というのも，ベル曲線を使うと，「何標準偏差離れているか」をおおよその百分率に変換できるからです。

5　具体例

よい統計研究は，現実世界の文脈と数学的手法を組み合わせることで，あるシステムについて新しい情報を発見します。ここでは，第1 節の流れを，授業で集めたデータに適用してみましょう。

5.1　ハリボーのデータ（生の標本）

授業では，19 袋の中くらいのサイズのハリボーのお菓子についてデータを集めました。各袋について，種類ごとに何個入っているかを数えました。以下がその生データです。

標本	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
リング	3	7	4	2	6	6	7	3	7	4	2	6	6	4	7	5	5	2	6
ハート	5	4	6	4	5	2	5	7	1	3	6	3	4	6	4	3	2	4	5
ボトル	4	1	4	2	4	2	3	3	4	6	4	0	2	4	4	4	8	1	5
たまご	4	5	3	6	0	6	1	2	6	2	4	4	1	2	1	4	3	7	1
グミベア	7	8	6	12	11	10	10	9	7	11	6	14	15	8	10	9	6	12	7
1袋あたり合計	23	25	23	26	26	26	26	24	25	26	22	27	28	24	26	25	24	26	24

5.2　合計と平均個数

19 袋すべてを通して数えたお菓子の総数は476 個です。1袋あたりのお菓子の平均総数は

$$\frac{476}{19} \approx 25.05$$

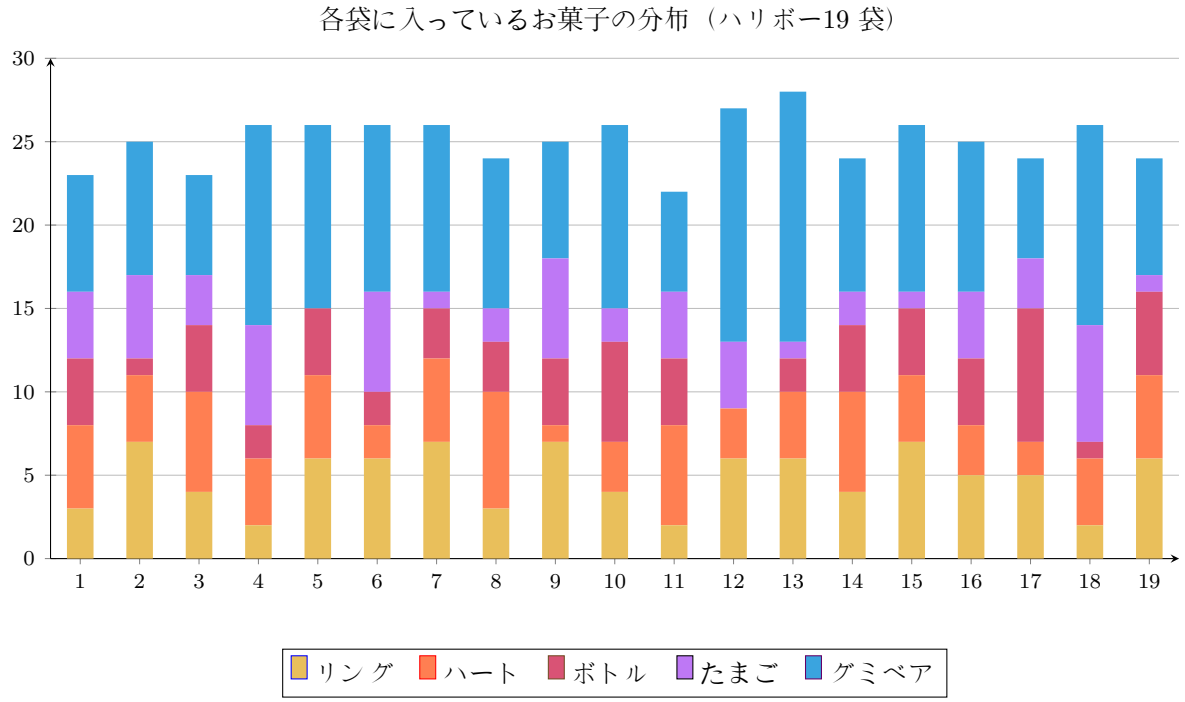
です。

種類ごとの合計と，1袋あたりの平均も同じように計算できます。

種類	**合計**	**1袋あたり平均**	**平均（近似）**
リング	92	92/19	4.84
ハート	79	79/19	4.16
ボトル	65	65/19	3.42
たまご	62	62/19	3.26
グミベア	178	178/19	9.36

5.3　お菓子の分布

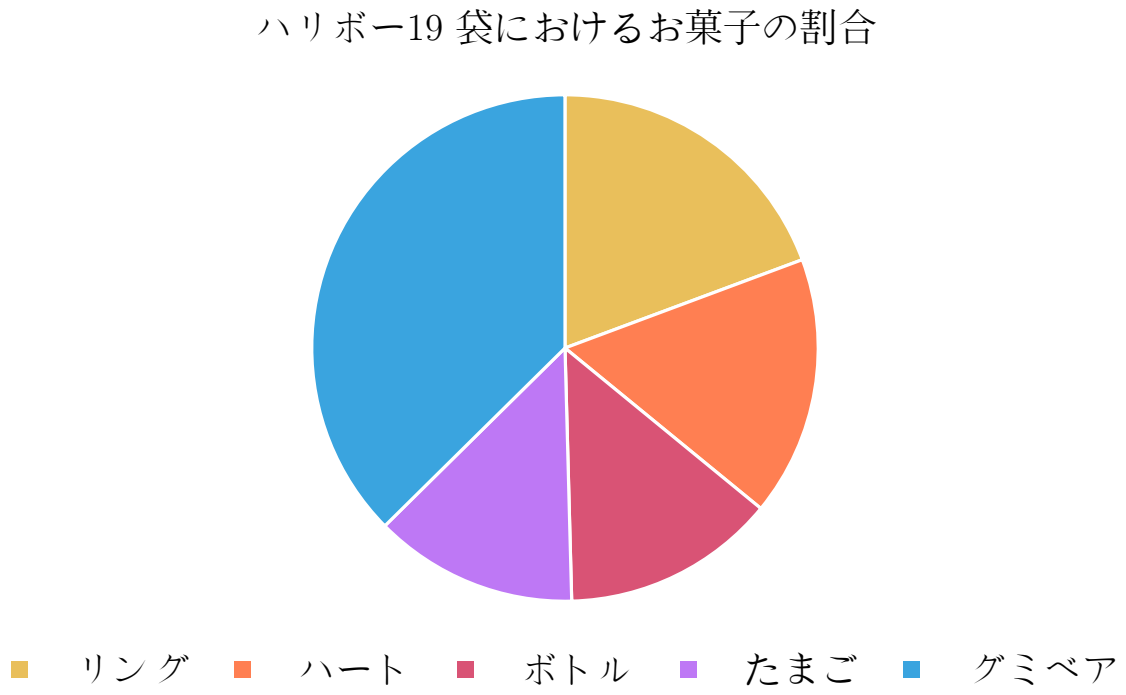
生データを使うと，各袋の中のお菓子の分布を大きなグラフとして描くことができます。

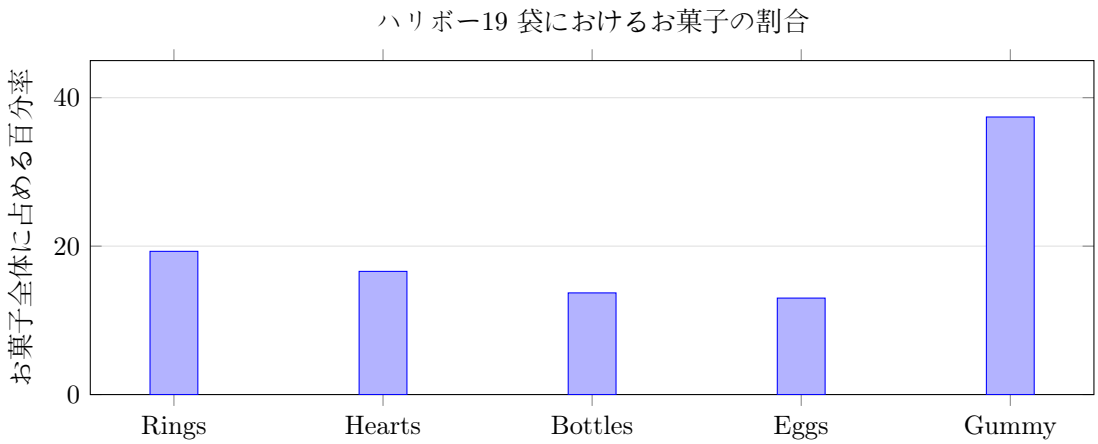


袋の大きさはおおむね25 個前後でかなり一定しているように見えます。しかし，各袋の中のお菓子の内訳はかなり変動しているようです。

5.4　お菓子の割合（相対度数）

種類	割合	百分率（近似）
リング	92/476	19.3%
ハート	79/476	16.6%
ボトル	65/476	13.7%
たまご	62/476	13.0%
グミベア	178/476	37.4%





5.5 ハリボーの個数におけるばらつきと変動

種類	平均	分散	標準偏差	CV
ハート	4.16	2.47	1.57	0.38
たまご	3.26	4.32	2.08	0.64
ボトル	3.42	3.48	1.87	0.55
グミベア	9.36	7.13	2.67	0.29
リング	4.84	3.25	1.80	0.37
1袋あたりのお菓子総数	25.05	2.27	1.51	0.06

1袋あたりのお菓子総数の変動は小さいですが，各種類ごとの変動係数はそれと比べてかなり大

きいです。これは先ほどの観察を裏づけています。すなわち，袋の大きさ自体はかなり一定している一方で，中身の構成はかなり変わるのです。

5.6 Excel でこれらを素早く計算する方法

概念	Excel コマンド
セルの合計	`=SUM(...)`
標本平均	`=AVERAGE(...)`
標本標準偏差	`=STDEV.S(...)`
中央値	`=MEDIAN(...)`
最頻値	`=MODE.SNGL(...)`
変動係数（CV）	`=STDEV.S(...)/AVERAGE(...)`

5.7 結果の解釈（三つの主張）

19 袋のハリボーからなる標本を調べた結果，次の三つが主な発見として示唆されます。

1. この標本では，各種類の割合はそれぞれ20% に近くはありません。とくに，グミベアは他の種類より多いように見えます。
2. 袋の大きさは，おおむね1 袋あたり25 個前後でかなり一定です。
3. お菓子の内訳は袋ごとにかなり変化しています。

5.8 考えられる標本抽出バイアス

よい統計研究であれば，少なくともその研究過程で生じたかもしれない偏りについて考慮するべきです。今回のケースでは，合理的な母集団として，標本抽出の期間中に地元（たとえば日本）で販売されていた中くらいサイズのハリボーの袋全体を考えることができます。そして，そこから19 袋の標本を取りました。以下は，注意すべき偏りの可能性です。

- 19 袋すべてが同じ製造ロットから来ていたかもしれない。
- 各自がデータを正しく記録したかどうかは分からない。
- 値をExcel に入力するときに入力ミスが起こった可能性がある。
- これらのお菓子を作った地元の工場が，他の工場と同じであるかどうかは分からない。
- この特定のハリボーのロットに，局所的な製造上の誤りがあった可能性を排除できない。
- 国によってハリボーの製造基準が違うかどうかも分からないし，たとえ分かっていたとしても，今回の標本がどの国で製造されたものかは記録していない。

5.9 研究の流れの言葉でまとめる

- **収集：** 学生たちが紙に書いたデータ。
- **整理：** Excel。
- **分析：** 平均個数，ばらつき，グラフ。
- **解釈：** 標本が，典型的な袋の大きさやお菓子の構成について何を示唆しているか。

- **提示：** グラフと要約文。

5.10　次にやること

これから，これらの主張を検定し，ハリボーの袋全体という母集団について情報を引き出すための技法を発展させていきます。