医学統計勉強会

東北大学病院循環器内科・東北大学病院臨床研究推進センター　共催

東北大学大学院医学系研究科 EBM 開発学寄附講座

宮田　敏

Absence of evidence is not evidence of absence!

- Carl Sagan -

第１回　基本統計量　― Table 1 を究めよう ―

1.　統計学とは

自然科学、社会科学を問わず実際の現象に取り組む場合、あるいは生産やビジネスの現場でデータに向き合う場合、そこには不確実性や多様性が伴います。例えば、病気の患者にある薬剤を投与したときの効果は、その薬剤の効果だけでなく、患者さんの体調や遺伝的背景、生活習慣など様々な背景因子の影響を受け、事前にその結果を知ることはできません。

しかしこれら不確実な事象には、個々の現象を取り上げれば確かに不確実でも、データに蓄積された過去の経験をもとに何らかの傾向、法則性を見出し、合理的な推論を行うことが可能な場合もあります。そのために、データを収集し解析する方法論が「統計学」である、といえます。データに含まれる不確実性は、確率的事象としてモデル化されます。確率的事象を扱う数学理論が「確率論」になります。すなわち、不確実性や多様性を伴った事象に対して、合理的な推論を行うための方法を提供するのが統計学であり、その理論的枠組みを支えるのが確率論、ということになります。

もし、生命現象あるいは社会現象において関連するすべての要因を制御できれば、不確実性を除かれ、現在の状況と将来の予測を完全に理解できるようになるでしょう。しかし、現実には不確実な現象についてすべての情報を得ることは不可能であり、100%誤りのない判断をすることは困難です。ではどうするか。すべての情報を得ることは無理でも、部分的な情報を集め、それを基に全体を推論することが必要になります。「不確実性」のないところに、統計学は必要ありません。

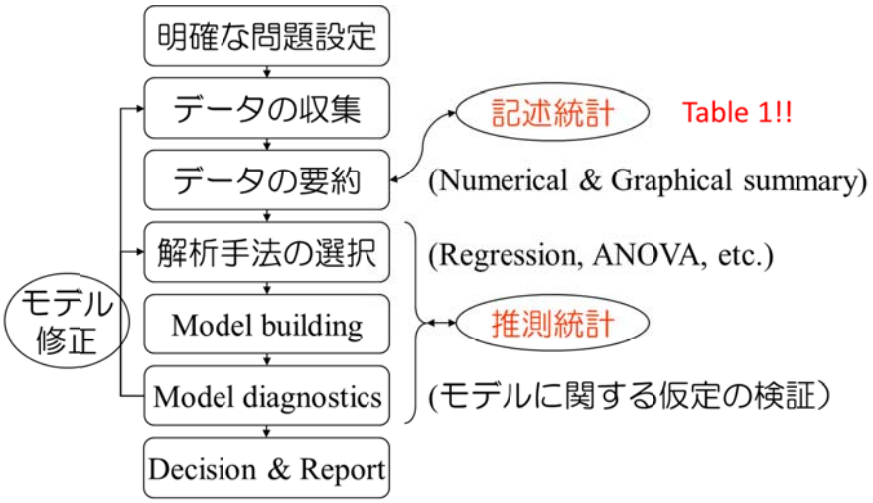
データの解析には、大規模な計算が必要になる場合もあります。また現代の統計学ではデータを可視化（visualization）し、視覚的にデータの特性を捕らえることが必須です。いずれの場合にも、計算機上で統計解析ソフトを利用して解析することが必要です。現在はさまざまな統計解析ソフトが開発されており、標準的な解析はどのソフトでも行えるようになっています。

1．1　データ解析の手順

実際のデータ解析において、興味の対象となる事象に関するすべての情報を得

ることは不可能です。(例えば、与党の支持率を知るために、すべての有権者の意向を確かめるのは現実的ではありません。)　したがって可能な戦略は、興味の対象について部分的な情報を集め、それを基に全体を推論することになります。部分的な情報(データ)から全体の推論を行うわけですから、データの収集は慎重に計画されなければなりません。また、それ以前に推論の目的は何であるのかも、はっきりさせなければなりません。これらを含め、データ解析の手順は以下のフローチャートにまとめられます。

図 1. 1



i. **明確な問題設定** データ解析を始めるためには、いかなる対象について、何を知りたいのか。そのためにはどのようなデータを、どのような方法で集めればよいのかを明らかにする必要があります。そのため解析の目的となる問題を明確に設定する必要があります。それによって、解析の手法も変わってきます。解析の枠組みを決める大切なステップですので、次の節で詳しく解説します。

ii. **データの収集** 解析の目的が定まったら、次は目的に合わせてデータを収集する段階になります。このステップで大切なのは、解析対象から偏りなくデータを集めることです。一概に「偏りなく」データを集める、といっても実は簡単ではありません。

iii. **データの要約** データが収集されても、いきなり解析に移るわけではありません。データの特徴や傾向を大掴みに把握するため、データの要約を行います。次のステップでは解析の方法を選ぶわけですが、そのためにはデータの傾向をつかんでおくことが役に立ちます。

　また、さまざまな解析手法の背後には数学的なモデルがあるわけですが、

モデルは無条件に使えるわけではなく、何らかの前提条件を必要とするのが普通です。データを要約することで、解析しようとするデータがモデルの前提条件を満たしているか吟味することも必要です。

データの要約は、1）データの位置や散らばりを特徴付ける代表値を求める**数量的要約（Numerical Summary）**と、2）各種の図を用いた**視覚的要約（Graphical Summary）**の二つに分けられます。Numerical Summary と Graphical Summary の二つをあわせて**記述統計学（Descriptive Statistics）**と呼ばれます。

iv. **解析手法の選択** 前のステップでデータの大まかな傾向をつかんだあと、解析目的に合わせた手法が選択されます。前述したとおり、解析手法にはその前提となる数学的な条件があり、データがそれを満たさないようなモデルは選択できません。

v. **Model building** このステップで、いよいよデータに解析モデルを当てはめます。この勉強会で取り上げる回帰分析、分散分析、ロジスティック回帰分析、生存時間解析なども、ここでいう統計解析モデルに当たります。

vi. **Model diagnostics（モデル診断）** データに解析モデルを当てはめた後は、解析結果を照らし合わせてモデルの仮定が満たされているか改めて確認する必要があります。この確認作業のことをモデル診断といいます。もしデータがモデルの仮定を満たさないときは、前のステップに戻ってモデルを修正する必要があります。使用する解析モデルを変更することもありますし、データをほかの形に変換することもあります。場合によっては、最初からデータを取り直すこともあります。

vii. **Decision & Report** モデル診断によって、すべての仮定が満たされたことが確認されたら、最終的なモデルの結果を評価し、当初の解析目的にしたがって推測を行います。

1.　2　問題の設定・データ解析のパラダイム

データを解析するとき最初にやるべきことは、そのデータを解析することで何を知りたいのか明確に問題を設定することです。一般に、次の 6 つの概念を明確に定義することで、データ解析の目的を厳密に設定できるといわれています。

i. **母集団（Population）** 解析対象となる個体の集合。もし世論調査で日本の政党の支持率を知りたいのであれば、母集団は日本の有権者の集合になります。病気の患者に対する薬剤の効果を知ることが解析目的なら、その病気に罹る可能性のあるヒトの集合が母集団になります。

ii. **パラメター（Parameter）**母集団を特徴付ける定数。上の世論調査の例であれば、政党支持率がパラメター、薬剤効果であれば、例えば薬の奏功率がパラメターになります。

iii. **標本（Sample）**母集団から抽出された部分。標本が持つ部分的な情報を基に、母集団のパラメターを推測するのがデータ解析の目的です。

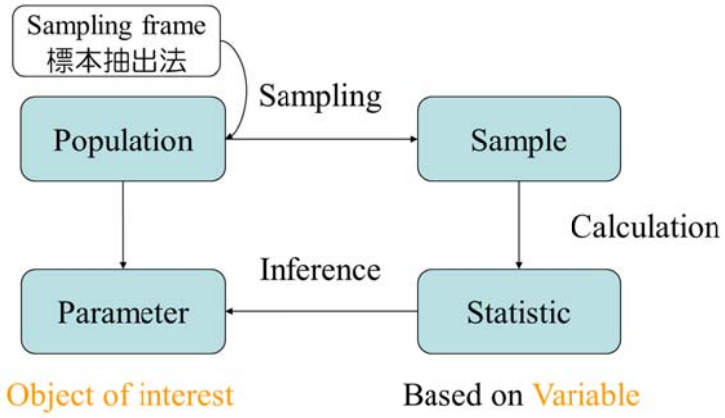
iv. **Sampling frame** 標本として抽出されうる個体の集合。すなわち、標本となる可能性のある個体の集合です。もし母集団の個体すべてに標本となる可能性のない場合、すなわち Sampling frame と母集団が一致しない場合、標本は母集団全体を代表せず解析に偏りが生じます。

v. **変数（Variable）**母集団において、個体間で確率的に異なりうる特性、量。上の世論調査の例であれば、各有権者の各政党への態度（支持･不支持）、薬剤効果の例であれば、薬を投与された各患者の応答性（効果あり･なし）が変数になります。実際に観察された変数の値のことを、**データ**と言います。

vi. **統計量（Statistic）**標本から計算される量。統計量の値によって、パラメターを推測します。

データ解析の目的をもっとも抽象的に定義するのであれば、それは「母集団のパラメターについて何かを知ること」になります。すなわち、解析対象となる母集団の関心のあるパラメターについて推測をすることが、データ解析の目的です。実際には母集団について 100%完全な情報は手に入れることができないので、その一部を標本として抽出し、標本中の個体の変数の値を調べた上で、変数から統計量を計算し、統計量の値からパラメターに関する推測を行うのが、データ解析の流れになります。

図 1. 2



2.　元データの取り扱い

i.　データの形は**長方形**。

データを入力する際は、第一行目に変数名を記入します。多くのソフトウエアは日本語入力に対応していますが、それでも**全角文字は避ける**ほうが無難です。

第二行目以降にデータを記録していきますが、元データにはグラフ等を張り付けたりしません。また、第一列目にはデータの ID を記録します。そうすると、元データは以下のような長方形になるはずです。(横の並びが「行」縦の並びが「列」です。)

systemID	hospitalID	sex	age	height	bodyweight
4	1185645	1	64	173	75.4
11	3329388	1	69	164	72
12	4022624	1	78	155.2	47.2
14	4402536	1	83	159.1	60
22	4862866	2	73	147.6	40.5

ii.　元データは**絶対に改変しない**。

データを解析する際、変数を変換したり新しい変数を定義したりする必要が出てくることがあります。このとき元データを改変して、変換した変数を上書きしたり変数を新たに保存したりしてはいけません。データを改変したときは、必ず**新しいファイル名で保存**しなおします。元のデータを改変した場合、解析を進めるうちに元データが何であったのか分からなくなることがあります。元データがわからなくなれば、**意図せざるデータのねつ造**まであと一歩です。

iii.　患者さんの**個人情報は記載しない**。

個人情報保護の重要性は改めて述べるまでもありませんが、残念ながらいまだに患者さんの名前やカルテ番号など個人に直結するデータを記録したままでデータをやり取りする例が見受けられます。患者さんの個人情報は、データ解析の立場から見れば何の意味もありませんが、万が一外部に流出する、あるいは記録媒体を紛失するなどすれば、研究の中断では済まない問題に発展します。

解析データの**個人情報は削除**する、を徹底する必要があります。

iv.　解析の過程の**詳細なメモ**を残す。

様々な実験の結果を論文にまとめる際、「実験ノート」に詳細を記録することは常識ですがデータ解析でも同じです。解析の過程を記録するのには、次のような意味があります。

- **研究の再現可能性**を確保するため。

  科学的研究においては、第三者の事後的な検証に耐えられるよう研究の再現

が可能でなければなりません。元データと記録メモさえあれば、それ以外に知識のない人でも解析が再現できるような詳細なメモを残す必要があります。

- **備忘録**。自分自身、何をしているのか分からなくなることを防ぐため。
  一日数時間の解析でデータ解析が終わることは、まずありません。一か月、二か月と時間をかけて解析を進めた場合、最初のころに自分が何をしていたのか分からなくなることがあります。データ解析の世界には、「三日後の自分は遠い親戚、一週間後の自分は赤の他人」という言葉があります。赤の他人が見ても、何をしているのか分かるようなメモを心がけます。

2. データ入手時にまずすべきこと

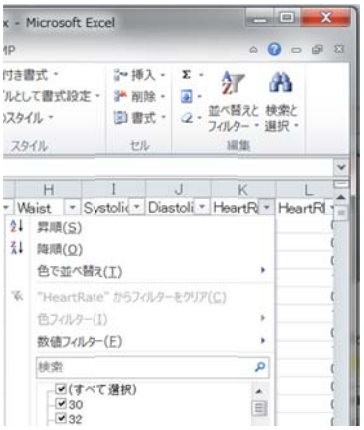
i. データ全体の確認

最初に得られた生データには、往々にして記録ミスや不適切な入力が存在するものです。そういった誤りは適切に修正、削除する必要がありますが、その確認作業を体系的に行うことでミスを減らし時間を節約することができます。

以下の手順は、私が普段行っているものですが参考にして頂ければと思います。

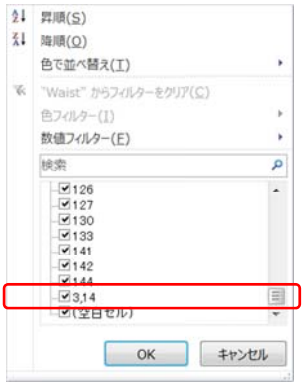
- Excel でデータファイルを開き、「並べ替えとフィルター」→「フィルター」を押してフィルターをオンにする（列見出しに矢印が現れる）

- 列見出しの矢印をクリックして、各列に適用するフィルターが見えるようにする。



- 以下の要領でデータの誤りや異常値の有無を確認する
  - データの範囲：本来、正値しか入らないはずの変数に、負値が入っている等。
  - データが想定の範囲を逸脱する。（小数点の桁間違いで、例えば体重 35.0 kgが 3.5 kgと入力されるなど。）

- 全角文字と半角文字の混在。（“I” と “I” の区別など難しい。前が半角、後ろが全角）
- 質的変数の数字表記：例えば「性別」が、男性→1, 女性→2 で記入されるような場合があるが、間違いの元なので、男性→M, 女性→F のように書き直す。
- 異常な値の検出:例えば小数点とカンマの打ち間違いで、“3.14” が “3,14” となっている場合など。そういった異常値はフィルターの下の方に出る。



- 欠測値の数：データに欠測がある場合、フィルターに「空白セル」と表示される。「空白セル」を選択してフィルターをかければ、欠測値の個数を調べられる。欠測の数が想定より大きかった場合、入力したデータが認識されていない、などの可能性が考えられる。

ii. 以上の確認作業を、すべての変数について行う。どのデータに対して、いかなる修正、削除を行ったか、すべて解析メモに記録する。修正後のデータは新しいファイル名で保存し、これを解析ファイルとする。（元データには手を付けない）

3.　記述統計

データが得られたとき、解析の第一歩はデータを要約し、その分布の大まかな特徴を把握することになります。データの要約（summary）の目的はデータの分布の形状を理解することですが、その方法は 1）数量的なデータの要約（numerical summary）と、2）視覚的なデータの要約（graphical summary）に分けられます。これらを総称して記述統計学といいます。

さて、記述統計の内容について説明する前に、なぜ記述統計によってデータの概要を理解することが重要なのか、今一度考えておきます。

3．1　記述統計の重要性

前述のとおり、記述統計はデータを要約し、データの持つ全体的な特徴、傾向

を表現します。特にデータの分布の位置（location)、分布の広がり(分散、variance)、およびその形状の要約を重視します。なぜこのようなデータの要約が必要なのか、その理由として以下のようなものが考えられます。

1. **適切な解析手法の選択のため**　統計学では、同じ目的に対して複数の解析手法が存在する場合があります。例えば分布の平均を推定する場合でも、分布の形状が左右対称なのか、右左いずれかに歪んでいるのか、あるいはデータの中に質の異なるサブグループが存在するのか、状況によって異なる手法を用いる必要があります。適切な解析方法を選択するためには、データの特徴を把握することが重要になります。
2. **データが公正に誤りなく収集されていることを示すため**　比較対象実験の場合、対照のための条件（例えば、投薬の有無）以外の背景因子には極端な違いがないことが理想です。もし比較群と対照群に違いのある因子があれば、続く解析では介入因子と共に結果への影響を解析しなければなりません。また、例えば比較群と対照群で平均や分散が一致してしまうとか、本来負の値はとらないはずの変数が負の値をとっているとか、異常に欠測値が多いとか、何かデータ収集の誤りを思わせる要素がないことを積極的に明示するのも、記述統計の重要な役割だといえます。

3. 2　数量的なデータの要約

数量的なデータの要約の目的は、分布の形状を特徴付ける統計量を計算し、データの大まかな傾向を理解することです。分布を特徴付ける統計量には、データの位置(中心)を表す量と、データの変動や散らばり（variability, dispersion）を表す量があります。

3. 2. 1　データの位置

i. **平均（mean）**　データの位置(中心)を表す代表値として、最もよく使われるのが平均です。n 個の観測値 $x_1, x_2, \cdots, x_n$ が与えられたとき、平均は以下の式で定義されます。

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

ii. **中央値（median）**　平均に対して、標本の大小の順序に注目し、ちょうど真ん中に来た値でデータの中心を現す代表値に中央値があります。n 個の観測値 $x_1, x_2, \cdots, x_n$ が与えられたとき、これらを大きさの順に並べなおして

$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ としたものを**順序統計量**といいます。つまり、$x_{(1)}$ は最小値、$x_{(n)}$ は最大値になります。順序統計量の概念を用いて、**中央値（Median）**は以下のように定義されます。

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & : \text{n is odd} \\ \left(x_{(n/2)} + x_{((n+1)/2)}\right)/2 & : \text{n is even} \end{cases}$$

つまり中央値とは、標本を大きさ順に並べたとき「真ん中」にくる値です。

iii. **パーセント点（Percentile）**　中央値は、その定義から標本を小さいほうから $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ と並べなおしたとき、50%の順位にある値です。この考え方を拡張して、データの小さいほうから 100×k%の順位にある値を **k-th percentile（パーセント点、百分位点）**といいます。

iv. **四分位点（Quartile）**　とくに、25 パーセント点(25-th percentile)を**第一四分位点(first quartile)**、75 パーセント点(75-th percentile)を**第三四分位点(third quartile)**といいます。50 パーセント点=**第二四分位点(second quartile)** は中央値そのものになります。これら中央値、パーセント点、四分位点は順序統計量を基に定義されており、極端に大きいあるいは極端に小さい異常値に対して影響されにくい性質を持っています。

v. **刈り込み平均（trimmed mean）**　中央値と平均の中間的な概念として、**刈り込み平均（Trimmed mean）**があります。k% trimmed mean は，データから上下 k%を取り除いた後の平均になります。

3. 2. 2 データの広がり

データの中心を現す代表値は、データがどのあたりに分布しているのかその位置を示しています。分布の形状を特徴付けるもうひとつの重要な概念に、データの変動（variability）や散らばり（dispersion）があります。

例えばデータが二つの群に分けられるとき、それぞれの群の平均に意味のある差があるかどうか検討する際、データの散らばりの大きさは重要な役割を果たします。データの散らばりが大きすぎれば、平均のわずかな差はノイズに埋もれてしまいます。平均の差に比べてデータの散らばりが小さければ、よりたやすく平均の差を見出すことができます。

1. **分散（variance)、標準偏差（standard deviation）**　データの散らばりを測る尺度として最もよく用いられるのは、以下に定義する、偏差の二乗(標本と平均との二乗距離)の平均を用いた**分散（variance）**です。また分散の平方根は**標準偏差（standard deviation）**と呼ばれます。

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

2. **四分位点間距離（Inter Quartile Range, IQR）**　データの分布の散らばりの尺度としては、分散と標準偏差は代表的なものです。しかし分散の定義には平均$\bar{x}$が用いられ、各標本x_iの分布の中心からの散らばりは二乗距離$(x_i - \bar{x})^2$で測られます。平均と中央値の関係で見たとおり、$\bar{x}$は極端に大きい、あるいは小さい異常値に影響を受けやすい性質があります。また、同じく異常値があった場合、二乗距離$(x_i - \bar{x})^2$は極端に大きな値をとりやすくなります。したがって分散s^2（およびそれから定義される標準偏差　s）は、やはり異常値に影響されやすいという欠点を持つことになります。s^2は数学的に扱いやすいという利点もありますが、異常値（outlier）に対して影響されにくい（「頑健な」あるいは「ロバスト（robust）な」）散らばりの尺度が必要なこともあります。異常値に対して頑健な散らばりの尺度として用いられるものに、以下の**四分位点間距離（Inter Quartile Range, IQR）**f_sがあります。

$$f_s = \text{第三四分位点} - \text{第一四分位点}$$

第一四分位点（first quartile)、第三四分位点（third quartile)ともに順序統計量を基に定義されますから、f_sは異常値に対し影響されにくい尺度になっています。

標準偏差と標準誤差

以上で述べてきた平均、分散などを用いて、データの数量的な要約が行われます。しかし、論文などで実際に要約を行う際はいくつかの決まったやり方で要約されることが多いようです。論文の中では、しばしば次のような表現を見かけます。

“Continuous variables were expressed as mean ± SD, mean ± SE or median (interquartile range), as appropriate.”

これは、「連続変数（実数値であらわされる変数）は、平均±標準偏差、平均±標準誤差、あるいは中央値（四分位点間距離）の、いずれか適当なもので表現される」ということです。まず新しい概念である**標準誤差**を定義します。

標準誤差　標本平均の標準偏差 $= \frac{s}{\sqrt{n}}$

標準偏差は、観測データ全体の散らばりの大きさを表します。データが正規分布に従う場合は、平均±標準偏差の範囲にデータの 70%弱が分布していると想定できます。これに対して、同じ母集団から何度もサンプル収集を行いその都度標本平均を計算したとき、標本平均の散らばりの大きさはデータ全体の散らばりの大きさよりずっと小さくなると考えられます。この「標本平均の散らばりの大きさ」を測る概念が標準誤差になります。（より正確には、何らかの統計量の標準偏差を標準誤差と言います。特に言及なしに標準誤差というときは、通常上に示したように標本平均の標準偏差（Standard Error of Mean, SEM）を意味します。）

以上を踏まえると、平均±標準偏差、平均±標準誤差、あるいは中央値（四分位点間距離）の使い分けは以下のようになります。

Mean ± SD (Standard deviation)：平均(Mean)を中心に Mean ± SD の範囲に、データ全体の 60～70%が分布している。これは観測データの散らばりを意味するので、データ全体を記述するのに適した表現。

Mean ± SE (Standard error)：同じ母集団から同じサイズの標本を繰り返し採集し、サンプリングのたびに標本平均を計算したとする。このとき平均(Mean)を中心に Mean ± SE の範囲に、標本平均の 60～70%が分布している。標本平均は母集団平均を推定するための推定量であるから、SE は標本平均による母集団平均の推定の精確さ（precision）を測っていることになる。

二群以上を比較するときは、平均の推定を問題にしているので **Mean ± SE が第一選択**。一群の時は，データ全体の散らばりの範囲に興味があれば Mean ±

SD も可能。

Median (IQR): 中央値(Median)を中心に, IQR の範囲にデータ全体の 50%が分布している．観測データ全体の散らばりを記述している点で、平均±標準偏差に対応する概念であることがわかります。

平均±標準偏差を用いるときの注意点として、平均±標準偏差はデータの分布が歪んでいるとき不合理な値をとる可能性があることが挙げられます。

図 2. 1

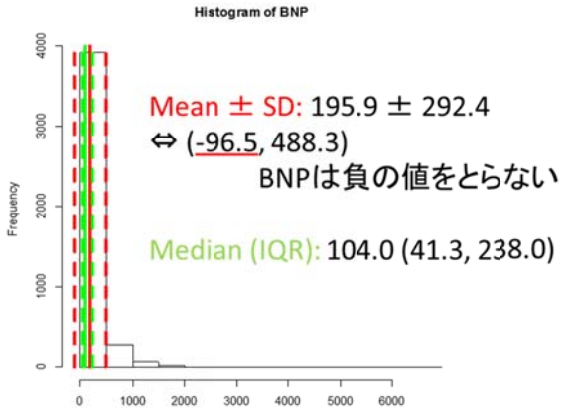


図 2.1 は、ある集団の BNP（brain natriuretic peptide、脳性ナトリウム利尿ペプチド）の分布を示しています。BNP は正の値をとり、右に強くゆがんだ分布を持つことが知られています。図 2.1 のデータの場合、平均 195.9、標準偏差 292.4 ですので平均±SD が 195.9±292.4 であると表記することは、「BNP の値の 60～70%が-96.5 以上 488.3 以下の範囲に分布している」と主張するのと同義です（赤線が平均、赤点線が平均±SD の範囲）。これは BNP が負の値をとる可能性を示唆するもので、ナンセンスであるといわざるを得ません。これに対して IQR は必ずデータの分布する範囲内に収まりますから、このような歪んだ分布に対してデータの散らばりを示すのに適しています（緑線が中央値、緑点線が IQR の上下限）。

平均±標準偏差を使うかどうかは、実際に平均±標準偏差の上限と下限を計算し（慎重を期するのであれば、平均±2×標準偏差）平均±標準偏差の範囲がそのデータの通常の範囲を逸脱しないかどうかで判断します。

3. 3　視覚的なデータの要約

数量的な要約によって、データの分布を特徴付けるさまざまな数値情報を得ることができます。しかし、それによって分布の形状が理解できるとは限りません。分布の形状を把握するには、グラフィカルなデータの要約によって視覚的に分布を捕らえることが有用です。本節では、最も基本的な grahpical summary として、**ヒストグラム**と**ボックスプロット**を取り上げます。

3. 3. 1　ヒストグラム（Histogram）

観測値が得られたとき、標本の範囲（Range）をいくつかの連続する区間（sub-interval）に分割する。この区間を**階級（Class/Bin）**といい、各階級の上限と下限の中間値を**階級値**という。各階級の中に値をとる観測値の個数を**度数(Frequency)**、標本の総数を 1 としたときの各階級の度数の割合（度数/標本数）を**相対度数（relative frequency）**という。横軸に観測値をとり、縦軸に度数もしくは相対度数をとった棒グラフを**ヒストグラム（histogram）**という。もし階級の幅がそれぞれ異なるときは、各階級の上の「長方形」が度数、あるいは相対度数に比例するように、「(長方形の面積)=(階級の幅)×(長方形の高さ)」によって棒グラフの高さを決める。

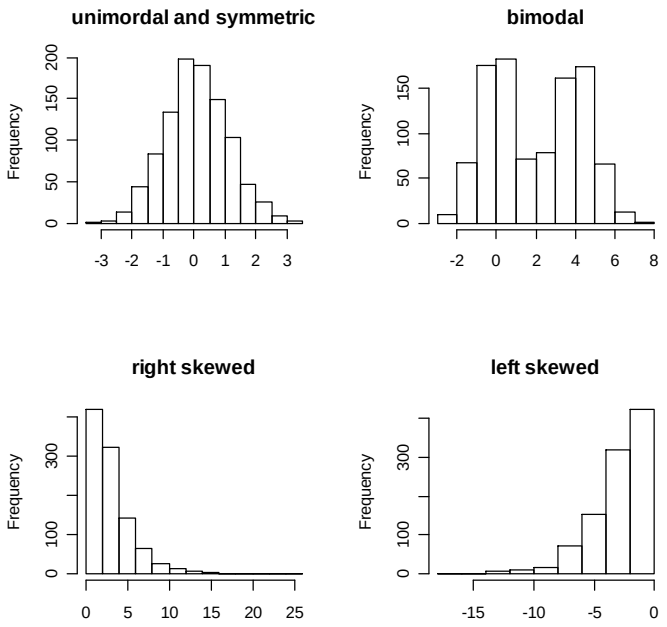
ヒストグラムの階級の数を決めるための方法は、いくつか提案されていますがまだ決定的なものはありません。階級の数 k を決める古典的な方法として、以下の「Sturges の公式」が知られています。

$$k \approx 1 + \log_2 n$$

ヒストグラムの形状

ヒストグラムはデータの分布の形状について、わかりやすい要約を与えてくれます。**unimoda (単峰型)**の分布は、ただひとつのピークを持ちます。これに対し **bimodal(二峰型)**の分布は二つのピークが見られる分布で、データが二つのまったく異なる構造を持つサブグループからなるときなどにおこります。さらに多くのピークを持つ分布は **multimodal(多峰型)**と呼ばれます。分布の対象性に着目すると、まず左右対称なデータによる **symmetric** な分布があげられます。これに対して、分布の右すそが長い分布は**右に歪んだ分布、positively skewed, or right skewed** と呼ばれます。逆に分布の左すそが長い分布は**左に歪んだ分布、negatively skewed, or left skewed** と呼ばれます。それぞれ、代表的な形状のヒストグラムを図示します。

図 2.2



3.　3.　2　ボックスプロット（Box-plot）

ヒストグラムは、分布の全般的な形状を図示するのには適していますが、データの位置や広がりを示す記述統計量を明示することはできません。また、平均値や分散値に大きな影響を与える「はずれ値（Outlier)」を示すこともできません。これらの点を改善する方法として、ボックスプロットがあります。

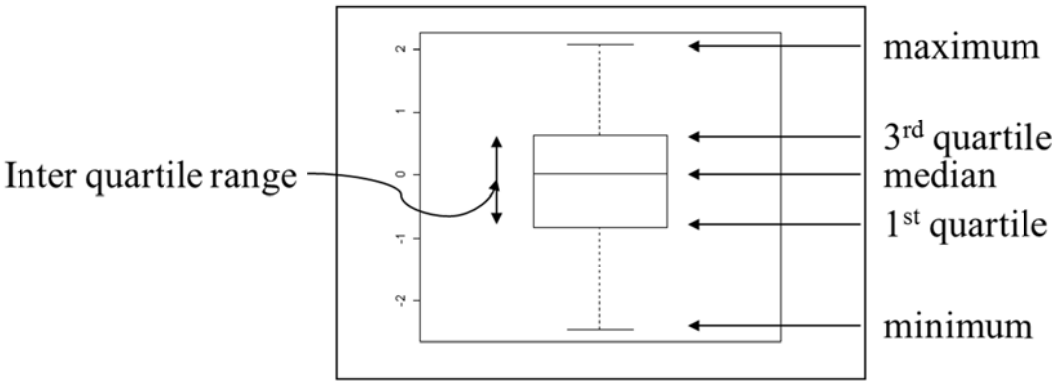
定義　f_s を、データの四分位点間距離（IQR)とする。(第一四分位（first quartile - 1.5 f_s）より小さい観測値、もしくは（第三四分位 + 1.5 f_s）より大きい観測値を**はずれ値（Outlier）**とよぶ。はずれ値は四分位から 3 f_s 以上離れているとき **extreme** であるといい、そうでなければ **mild** であるという。

ボックスプロット（Box-plot）　ボックスプロットは以下の手順で描かれる。1)縦軸に変数値をとり、下限が第一四分位、上限が第三四分位にあたる長方形を描く。2)長方形の中の中央値（Median）にあたる位置に線を描く。3)長方形の上下辺から観測値の最大値、最小値まで「ひげ（whisker)」を描く。ただし、デ

ータの中にはずれ値があるときは、長方形の上下辺から（第一四分位 - 1.5 f_s）

および（第三四分位 + 1.5 f_s）まで「ひげ」を描き、はずれ値は点で表す。

図 2.3

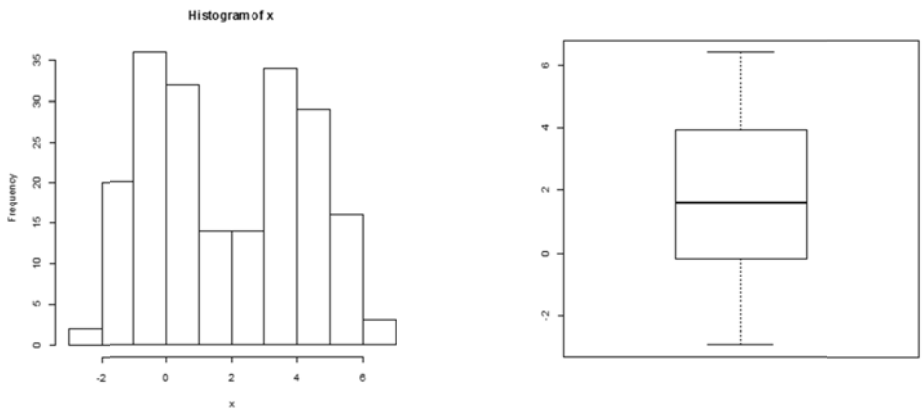


3.　3.　3　ヒストグラムとボックスプロット

本節では､ヒストグラムとボックスプロットという 2 種類の図を紹介しました。この二つがどのような特徴を持つかを示すため、以下の例を考えます。

二峰型のデータ：図 2.4 は同一の二峰型のデータ（ピークを二つ持つデータ）のヒストグラムとボックスプロットを示している。ヒストグラムは、明らかに二峰型の特徴を示しているが、ボックスプロットからは二つのピークを特定することはできていません。

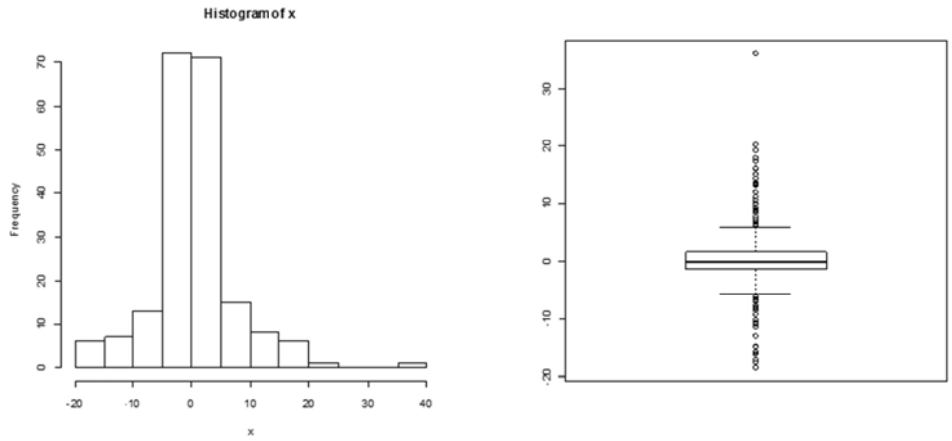
図 2.4



このことから、ヒストグラムはデータの分布の全体的な傾向をとらえるのに向いていることがわかります。

裾の重いデータ：一方図 2.5 は、いわゆる裾の重いデータであって、多数の極端に大きい、あるいは小さい「外れ値」を含んだデータの、ヒストグラムとボックスプロットになります。

図 2.5



ボックスプロットは、その定義から（第三四分位点 ＋1.5IQR）より大きい、もしくは（第一四分位点 －1.5IQR）より小さいデータを「外れ値」として表示するため、データの裾が重い分布の場合極端に大きい（小さい）異常値をとらえるのに適しているといえます。他方ヒストグラムの方は、単峰型のデータのヒストグラムと似ており、すそ野が重いという特徴を十分には捉えていません。このようにヒストグラムとボックスプロットはそれぞれ異なる特徴のデータに適しており、結局両方描くことが必要になります。

最後に、本節で検討した「ピークが二つある」とか「データの裾が重い」といったデータの形状に関する情報は、平均や分散といった数値的なデータの要約ではとらえることができない、という点を強調しておきます。例えば、データの中心を推定するのに平均値と中央値のどちらを使うのか、という判断には、データの分布が左右いずれかの方向に強く歪んでいるかどうか、といった分布の形状に関する情報が必要ですがそれはグラフを使った視覚的なデータの要約によってしか得られないものです。他方、視覚的なデータの解釈は多分に主観的なものですから、数値を用いた客観的な要約で保管してやる必要があります。

結局、数量的な要約と視覚的な要約は、ともに併用する必要があるということになります。

4.　平均・中央値の差の検定

前節では、カテゴリ変数を比較するための検定を考えました。本節では、２群あるいはそれ以上の多群間での、連続変数の平均値、中央値の比較を検討したいと思います。まず、２つのグループの平均の比較（二標本問題）から始めます。

4.　1.　**二標本問題**：二つのグループの平均値の差の検定

$H_0 : \mu_1 = \mu_2$　母集団平均が一定

$H_1 : \mu_1 \neq \mu_2$　母集団平均が異なる

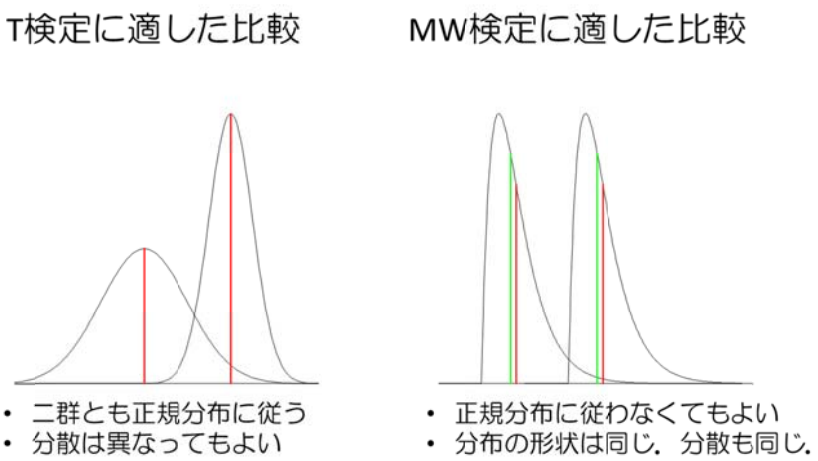
この、二標本問題を検定するための方法は、主として以下の二つです。

Welch's t-test（ウェルチの t 検定）：二群のデータがそれぞれ正規分布に従うと仮定する。二群の分散は等しくなくてもよい。（不等分散）Mean + SD に対応。

Mann-Whitney test, Wilcoxon's rank sum test：二群のデータは任意の同じ形の分布に従う。当然二群の分散は等分散になる。Median (IQR)に対応。

この二つの検定方法でもっとも大きな違いは、t 検定の場合データが正規分布に従う必要があるのに対して、Mann-Whitney (MW) 検定は正規性の仮定を必要としない点にあります。その一方で、Welch's t-test は二群の分散が異なっていてもよいのに対して、MW 検定では分散はおろかデータの分布の形まで等しいことが求められます。

図 4.1



一見、正規性の仮定を必要としない MW 検定の適用範囲の方が広いように見えます。しかし MW 検定が必要とする、二群の分散と分布の形が等しいという仮定はかなり厳しいものであり、どちらを使うかは慎重な判断が必要です。

なお、二標本問題における t 検定には等分散を仮定する検定もありますが、現実に使う場面は多くはないと思います。必ず不等分散を仮定した Welch's t test を使います。また、Mann-Whitney 検定は Wilcoxon's rank sum test とも呼ばれます。これは歴史的な経緯があって二つ名前がついてしまったのですが、理論的には全く同じものですのでどちらの名前を使っても結構です。

4.　2.　三群以上の平均・中央値の差の検定

本節では、二標本問題の拡張として三群以上のグループの平均の比較を検討します。三群以上の比較の場合、検定される仮説は以下の通りです。

$H_0 : \mu_1 = \cdots = \mu_k$　母集団平均が一定

H_1：少なくとも一つの母集団平均が他から異なる

この仮説を検定する方法は、主として以下の二つです。

分散分析（Analysis of Variance, ANOVA）：各群のデータがそれぞれ正規分布に従う。等分散を仮定する。

Kruskal-Wallis test：各群のデータは任意の同じ形の分布に従う。当然各群の分散は等しくなる。

この二つの検定の違いは、分散分析が正規性の仮定を必要とするのに対して、Kruskal-Wallis 検定は正規性の仮定を必要としない点にあります。但し、Kruskal-Wallis 検定も、各群の分布の形が等しいことが必要ですので適用条件が緩いというほどではないと思います。

三群以上の比較において本質的に重要なのは、上記二つの方法のいずれの場合でも、各群の分散が等しいという等分散性の仮定が置かれていることです。一般に対照群と比較群で分散が等しいという条件は、必ず成立するものではありません。もし各群で分散が異なっていた場合には、元データに何らかの変換を

施すことで分散を均等化する（分散を安定化する）必要があります。伝統的には、分散安定化のために対数変換などが用いられてきました。しかし、対数変換でもうまくいかない場合は、さらに進んだ変換（eg. Box-Cox 変換等）を試みる必要があります。これらの変換については、統計解析の専門家にご相談ください。

4.　3. 多重比較（Multiple Comparison）

分散分析の帰無仮説 $H_0 : \mu_1 = \cdots = \mu_k$ が棄却されたとき、「少なくとも一つの母集団平均が他から有意に異なる」という対立仮説を採択することになります。しかしこの対立仮説では、具体的に“どの”母集団平均が他から異なっているかはわかりません。したがって、次の興味はどの母集団平均が異なっているかを調べることになります。この問題を**多重比較（Multiple Comparison）**と呼びます。多重比較では、対立仮説のとり方によっていくつかの場合分けがあります。

Tukey's HSD (Honestly Significant Difference)：すべての対比 $(\mu_i - \mu_j), i \neq j$ についての検定を同時に行う。可能な対比の組み合わせは、k(k-1)/2 通り。

Dunnett の方法：グループの一つがコントロール群である時，コントロール群と他の(k-1)の対照群との比較を同時に行う．

$$H_1 : \mu_1 \neq \mu_2, \mu_1 \neq \mu_3, \cdots, \mu_1 \neq \mu_k$$

可能な対比の組み合わせは（k-1）通り。

Williams の方法：例えばある薬物の効果を考える際、第一群をプラセボ群、第二群以降第 k 群まで順次薬物の投与量を増やした対照群とする。このとき、薬物の効果には以下のような単調性が期待できる場合がある。

$$H_1 : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_k \quad \text{or} \quad H_1 : \mu_1 \geq \mu_2 \geq \cdots \geq \mu_k$$

このとき、上記の対立仮説を検定することで、どの群からプラセボ群と有意に薬効が異なるかを検定することができる。

上記の三つの方法は、いずれも各群のデータが正規分布に従う正規性の仮定を必要とします。正規性の仮定を必要としないノンパラメトリックな検定としては、以下のものが知れています。

	$H_1:\mu_1 \neq \mu_2,\cdots,$ $,\mu_i \neq \mu_j,\cdots,\mu_1 \neq \mu_k$	$H_1:\mu_1 \neq \mu_2,\mu_1 \neq \mu_3,$ $\cdots,\mu_1 \neq \mu_k$	$H_1:\mu_1 \leq \cdots \leq \mu_k$
パラメトリック検定	Tukeyの方法	Dunnetの方法	Williamsの方法
ノンパラメトリック検定	Steel-Swassの方法	Steelの方法	Shirley-Williamsの方法

Take Home Message

1. 統計学とは

我々が観察する現象とそれを記録したデータには、必ず不確実な誤差が伴います。統計学の目的は、この不確実性や多様性を伴った事象に対して、合理的な推論を行うことにあります。不確実性な現象が存在するとき、その対象について 100%完全なすべての情報を得ることは不可能です。しかし、全体の中のある部分をサンプルとして取り出し、そこから全体に対する何らかの傾向、法則性を見いだすことは可能で有り、その方法を提示するのが統計学だといえます。

2. 記述統計

記述統計とは、データを要約し、データの持つ全体的な特徴、傾向を把握するための統計学の分野です。記述統計の目的は、大きく二つに分かれます。

- データの特徴を把握することで、データに適した解析手法を選択する。
- 提示したデータに異常な（通常想定できる範囲を逸脱した）値や、誤りがないことを積極的に示し、データが公正に収集されたことを示す。

記述統計の方法は以下の二通りが有り、併用することで相互補完します。

- 数値的要約：平均、中央値（location）　分散、標準偏差、IQR（scale）
- 視覚的要約：ヒストグラム、ボックスプロット

3. 元データの取り扱い

4. カテゴリデータの要約と比較

カテゴリデータは、度数分布表、分割表にまとめる。分割表は二つのカテゴリデータの水準の組み合わせごとに、度数をまとめた表。二つのカテゴリ変数の「独立性」を検定する方法は、以下の二つ。

- Fisher の直接法（Fisher's exact test）
- χ^2 検定（Yates の連続補正）

正確な p 値を計算できる「Fisher の直接法」を第一選択とします。

4. 平均・中央値の比較

- 二標本問題：Welch’s t-test, Mann-Whitney test
- 三群以上の比較：分散分析、Kruskal-Wallis test
- 多重比較：三群以上の比較で有意差が認められたとき、どの対比において差があるのかを検定する方法。比較の仕方で、各種の方法があります。

以上