1. サンプルから未知の量を推定する

前のチャプターを始める時に，*記述統計*と*推測統計*の決定的な違いを強調しました。第 **??**章で議論したように，記述統計の役割は私たちが*まさに*知りたいものを簡潔に要約することにあると言えます。それに比べて，推測統計学の目的は “私たちがやったことから私たちが知らないことを学ぶ” ことにあります。私たちは確率の基礎を知っていますから，統計的推測の問題についてもうまく考えることができるでしょう。どういうことを学べばいいでしょう?　どうやって学べばいいでしょう? 推測統計の本質にある問いは，伝統的に 2 つの “大きなアイデア” に分割されてきました。推定と仮説検定です。この章のゴール h，この 2 つの大きな課題の前者，推定理論についてですが，まずはサンプリング理論について説明します。というのも，推定理論はサンプリングを理解しなければ意味を成さないからです。結果的にこの章は 2 つのパート，セクション **??** からセクション **??**を通じてサンプリング理論にフォーカスし，セクション **??**と **??**ではサンプリング理論を使って推定を統計的にどう考えるのかを議論します。

1.1 標本，母集団，そして標本抽出

パート の前触れとして，帰納法の謎と，すべて*の*学びには仮定が必要だという事実を強調しました。これが正しいとして，最初にすべきことは，データが意味をなすような一般的な仮定を考えることでしょう。そこで**標本理論**の登場です。確立理論はすべての統計理論を成り立たせる基礎だとすると，標本理論は家を建てる場合の枠組みとでもいえるでしょう。標本理論は，統計的な推論をするときに採用する仮定をたてるときに，かなり大きな役割を果たします。そして統計学者が考えるような “推測をする” ことについて話すとき，私たちが*何から*推論をするのか (標本)，そして*何に対して*推論をするのか (母集団) を明確にする必要があります。

ほとんどすべての状況で，私たちが研究者として手にすることができるのはデータについてのある**標本**です。ある特定の参加者に対して実験をしたのかもしれないし，調査会社が投票意図について

何人かに質問紙調査をしたのかもしれません。このやり方では，データセットは有限で不完全なものにしかなりません。世界中の全員に対して実験したりできませんし，例えば調査会社だってその国の全有権者に電話する時間もお金もないでしょう。記述統計のところで以前議論したときに (第 **??** 章)，この標本だけが我々の関心事でした。標本を記述して，要約して，グラフを描くことだけが目的だったのです。それを変えていくことになります。

1.1.1 母集団を定義する

標本というのは具体的な対象です。データファイルを開いてみれば，そこにはあなたの標本から得たデータがあるはずです。**母集団**は，それに対して，もっと抽象的な概念です。ぼ集団は，あなたが結論を引き出したいと思っている，あるいは標本よりかなりひろく一般化したいと思っている，すべての可能な人，すべての可能な観測の集合を指します。理想的な世界では，研究者はどの母集団に関心があるかを明確にしなければなりません。なぜなら，研究をデザインし，データの仮説検証をすることは，あなたが何か主張したいであろう母集団に依存するからです。

対象となる母集団を明確にするのが難しいこともあります。例えば，この章のはじめにあった “調査会社” の例では，母集団は研究を開始するときの全ての有権者であり，何百万もの人になります。標本は母集団に属する 1000 人ということになります。ほとんどの研究では，状況はそこまでストレートではないのです。典型的な心理学の実験においては，研究対象の母集団はもう少し複雑です。私が参加者 100 人の学部生を使って実験をしたとしましょう。私の目標は，認知的な科学ですから，心がどのようにして働くのかについて知ることです。So, which of the following would count as “the population”: であれば，次のどれが “母集団” としてカウントされるでしょうか。

- Adelaide 大学の心理学コースの学部生全員?
- 世界中のどこかでもいい，あらゆる心理学の学生?
- オーストラリアに今住んでいる人?
- 私の標本と近い年齢のオーストラリア人?
- 今生きてる人なら誰でも?
- 現在，過去，未来にわたって，とにかく人であればよい?
- 地球環境にいる十分な知的操作ができる生命体であればなんでも良い?
- 知的生命体であればなんでも良い?

これらはぞれぞれ心的過程を持つ実際のグループを定義するもので，いずれも認知科学者である私に取っては興味のある対象であり，私の興味関心に対してどれが正しい母集団なのかははっきりさせることはできません。別の例として，前置きのところで話したウェルズリー・コッカーゲームを考えてみましょう。このときの例は，ウェルズリーが 12 勝 0 敗という特殊な流れがありました。母集団はどれでしょう?

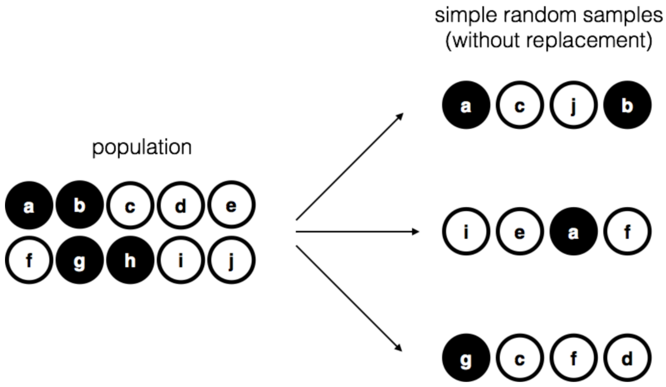


Figure1.1　母集団からの非復元単純ランダムサンプリング

- ウェルズリーとコッカーが，シーズン中に到達する全ての結果
- もしウェルズリーとコッカーが残りの人生の間ずっとゲームをしていたら得られるであろう全ての結果
- もしウェルズリーとコッカーが永遠に生きて世界の終わりまでずっとゲームをし続けていたら得られるであろう全ての結果
- 無数の並行世界において，ウェルズリーとコッカーのペアが同じ 12 回のゲームをそれぞれの宇宙でやっていたとしたら得られるであろう，全ての結果

もう一度言いますが，何が母集団なのかというのは，はっきりしないんです。

1.1.2　単純無作為標本

母集団をどのように定義するかに関係なく，重要なポイントは，サンプルは母集団の部分集合であり，目的はサンプルについての知識を使って母集団の特徴に関する推論を引き出すことです。両者の関係はどんな標本が選択するかという*手続き*に依存します。この手続きは**サンプリング法**に関係しており，なぜそれが問題になるかを理解することが重要です。

話を簡単にするために，10 個のチップが入った袋を想定してみましょう。各チップには重複しない文字が印字されているので，10 個のチップはそれぞれ区別することができます。またこのチップは，黒と白の 2 色に分けられます。このチップのセットが我々の興味がある*母集団*であり，図 **??**の左側にそれが描画されています。この図を見て貰えば分かるように，4 枚の黒いチップと 6 枚の白いチップに分かれているのですが，現実世界と同じように，袋の中を見ないとこれを知ることはできません。ここで次のような “実験” をすると考えてみましょう:　袋を振って，目を閉じて，4 枚のチップを復路に戻すことなくとりだすのです。最初に取り出したのが a チップ (黒) で，次が c チッ

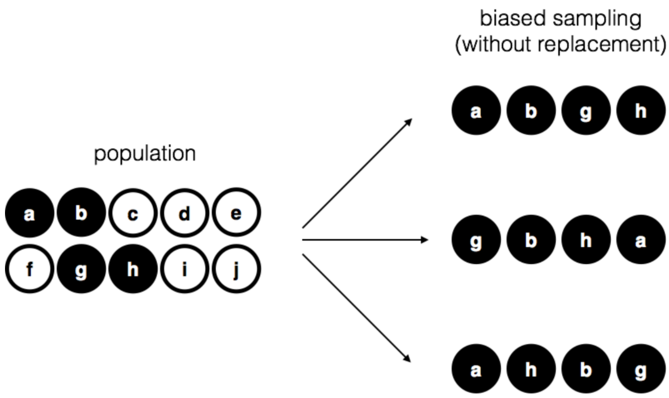


Figure1.2　有限母集団から復元なしの偏ったサンプリングをする

プ (白)，続いて j(白)，最後に b(黒) だったとします。もし望むなら，これらのチップを全て袋に戻し，実験を繰り返すことができます。それを表したのが図 **??**の右側です。毎回結果は違うことになりますが，いずれも手続きは同じです。同じ手続きげも異なる結果になるということを，我々は*確率過程*といいます[*1]。しかし，チップを取り出す前にバッグを振るので，全てのチップが選択される確率は同じぐらいだと考えることができます。母集団に含まれるどのメンバーも等確率で選出されうる手続のことを，**単純無作為抽出**といいます。取り出したチップを元の袋にもどさないというのは，同じことを 2 回観察することはないということですし，このような場合のことを*非復元*抽出をしたといいます。

この抽出手続の重要性に気づいて守るために，別の方法でこの実験をしたらどうなるか考えてみましょう。わたしの 5 歳になる息子が袋を開けて，4 つの黒いチップを取り出そうとしたとします。非復元で，です。この*偏った*サンプリング方法を図 **??**に表してみました。ここで，4 つが黒で白いチップが 0 である時の証拠となる価値を考えてみましょう。これは明らかにサンプリング方法に依存すると思いませんか? サンプリング方法が黒いチップだけを選ぶように偏っていて，その結果，標本が黒いチップだけだったというのでは，母集団についての情報が何も得られません! これが理由で，統計家はデータセットが単純無作為抽出であることを好みますし，そうであるからこそデータ分析がぐっと簡単なものになるのです。

第三の方法は注目に値します。今回，私たちは目を閉じて，袋を振って，チップを取り出します。しかし今回は，取り出したものを記録してから，そのチップを袋に戻すのです。そしてまた目を閉じて，袋を振って，チップを取り出します。この手順を 4 回繰り返します。このやり方でできたデータ

[*1] ランダムであることについてのより適切な数学的定義は，本当に技巧的でこの本の範囲を超えてしまいます。ここではそこまで技巧的にならず，プロセスをくりかえていて毎回違う答えが出る場合はいつでも，確率的な要素を持ったプロセスであるということにします。

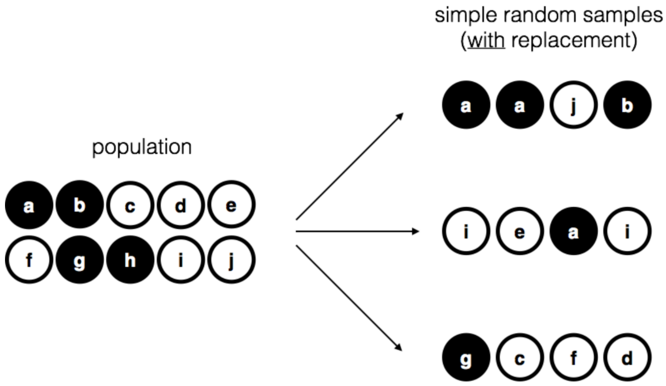


Figure1.3　有限母集団から復元ありの単純無作為抽出

セットは，これもまた単純無作為抽出ですが，取り出した後すぐ袋にチップを戻していますので，**復元**サンプリングといわれます。このやり方と最初の方法との違いは，同じ母集団メンバーを複数回観測することがあるかどうかで，図 **??**に表してみました。

私の経験上，ほとんどの心理学実験は非復元サンプリングをしているようです。というのも，同じ人が実験に 2 回参加することが許されてないからです。しかしほとんどの統計理論は，復元ありの単純無作為標本からデータができていることを仮定しています。現実的にこの違いはほとんど問題になりません。興味のある母集団が大きければ (10 個以上の中身があれば!)，復元あり，なしの違いは気にする必要がないぐらい小さいものです。これに対して，単純無作為サンプリングと，偏ったサンプリングの間の違いは決して見過ごせるものではありません。

1.1.3　ほとんどのサンプルは単純無作為標本ではない

先ほど示した考えられる母集団リストをざっとみればわかるように，興味のある母集団から単純無作為標本を得るのはほとんど不可能です。私が実験する時に，もし参加者がアデレード大学の心理学コースの学生から無作為抽出されたものであるということが明らかになったとしたら，ちょっとした奇跡だとおもうでしょう。一般化したい対象に比べて，圧倒的に狭い母集団でしかなくてもです。他のサンプリング方法に関する議論を深く議論するのは本書の範囲を超えますが，より重要なものをいくつか示しておきますので，知っておいてください。

- 層別サンプリングあなたの考える母集団が，いくつかの下位集団，すなわち*層*になっている (あるいはそう考えられる) としてみましょう。たとえば，いくつかの異なるサイトを通じて研究を走らせている場合などです。母集団全体からランダムに標本を取ってくる代わりに，別々の無作為標本をそれぞれの層から集めてくることになります。層別サンプリングは単純無作為抽出よりも簡単なことがあります。特に母集団が既にいくつかの層に分割されているときはそうです。いくつかの下位集団が滅多にない場合は特に，単純無作為抽出よりも効率的

です。例えば統合失調症を研究する時，母集団を二つの層 (統合失調症患者と，そうでない患者)*[2]に分けて，それ俺の群から同数のサンプルを取れば良いでしょう。人を無作為に選べば，統合失調症の人がほとんど標本に含まれず，あなたの研究の役に立たないものになるでしょう。この特殊なやり方である層別サンプリングは，*オーバーサンプリング*とも呼ばれます。というのも，めずらしい群を過剰に代表させようとしているからです。

- *スノーボールサンプリング*は "隠れた"，あるいはアクセスしにくい母集団からサンプリングする時に特に有用な技術で，特に社会科学で一般的な方法です。例えば，研究者がトランスジェンダーの人たちから意見を聞きたいと思ったとします。研究チームはトランスジェンダーの人たちの連絡を数人しか知らず，研究はその参加者にお願いするところから始まります (第一段階)。調査が終わる頃，参加者は他に調査に協力してくれそうな人に連絡を取ってくれないか，と頼まれます。第二段階では，この新しい連絡先が調査対象になります。このプロセスが，十分なデータがられるまで続くのです。スノーボールサンプリングの大きな利点は，他の方法では得ることが不可能な状況でもデータが得られることです。統計的な意味では，この方法の主な問題点として標本がかなりランダムから外れていることであり，無作為でないやり方をどう扱っていいかは難しい問題になります。一方現実的な意味では，うまくやらないとこの方法は非倫理的になるということです。というのも，隠れている人々は理由があって隠れているのですから。この問題に注目するために，私はトランスジェンダーの人を選びました。注意しないと，暴露されたくない人を暴露してしまうことになるかもしれず (それはとても，とても悪いことです)，ミスをしたわけでないとしても人の社会的ネットワークを使って人を研究するということは，侵入的ではあるのです。コンタクトを取る*前に*インフォームド・コンセントを得るのはとても難しいですし，その人たちにコンタクトを取って，"やあ，君の研究をしたいんだけど" という単純な行為さえ傷つけてしまうことが少なくありません。社会的ネットワークは複雑なものですから，データが手に入るからといって，常に使える手法というわけではありません。
- *コンビニエンス・サンプリング*はその言葉の響き以上のものでも以下のものでもありません。このサンプルを選ぶ方法は，研究者にとっては便利なものですが，興味のある母集団から無作為に選ぶものではありません。スノーボールサンプリングはコンビニエンス・サンプリングの一種ですが，他にもいろいろなものがあるのです。心理学におけるよくある例ですが，研究が心理学コースの学部生い頼っているところがあります。このサンプルは一般に，2 つの意味でランダムではありません。まず，心理学コースの学生に頼っているということは，' あなたのデータがある単一のサブグループに制限されているということです。次に，学生は普通どんな研究に参加するか選んでいるので，そのサンプルは心理学学生自身に選択されたサブセットになっており，ランダムに選択されたサブセットではないことになります。実際には，ほとん

*[2]現実的には単純ではありません。"統合失調症" と "統合失調症ではない" のような二分割できる明確な基準はありません。しかしこれは臨床心理学のテキストではありませんので，私があちこちでやってる単純化には目をつむってください。

どの研究は何らかの形でのコンビニエンス・サンプルです。これは時に厳しい制約になりますが，いつもそうだというわけではありません。

1.1.4 もし単純無作為抽出ができなかったら，どれほどの問題が?

オウケイ，実際のデータ収集ではステキな単純無作為抽出ができないかもしれない，ということでした。それのどこに問題が? ちょっと考えれば，データが単純無作為抽出でないときにどんな問題に*なりえる*かがわかると思います。図 **??** と **??**の違いを考えてみてください。しかし，思ったほど悪くはありません。偏ったサンプルの中には，それほど問題にならないタイプのものもあるのです。例えば，層別サンプリングを使う時はどんなバイアスがあるのかはっきりわかっているわけです。研究効率を*あげる*ために自分でそのバイアスをうんだわけですから。そしてそういう時は，統計的な手法であなたが作り出したバイアスを補正することができます (この本では扱わない技術ですけど!) ですからこういう時は，それほど問題になりません。

もっと一般的にいえば，無作為抽出は目的に対する手段であって，目的そのものではないことを忘れないようにすることが重要です。コンビニエンス・サンプリングのような，偏りがあることがわかっている場合を考えてみましょう。そのサンプリング手法に含まれる偏りは，そこから間違った結論を引き出したときに限って問題になるのです。その観点からいえば，*あらゆる*側面において標本が無作為でなければならないとはいえず，関心対象である心理学的に関係のある現象について無作為さが必要なのだと私は思うのです。私がワーキングメモリの容量について研究しているとしましょう。研究 1 では，今生きている人間から無作為抽出することもできますが，唯一の例外として月曜日生まれの人だけしか集められません。研究 2 では，オーストラリアの人からしか無作為抽出できないとします。私は研究結果を今生きているあらゆる人間に一般化したいと思っています。どちらの研究がマシでしょうか? 答えは，明らかに研究 1 ですよね。どうしてかって? それは “月曜日生まれ” ということがワーキングメモリの容量に興味深い関係があるとはとても思えないからです。それに比べて，“オーストラリア人である” ということが問題になるかもしれない，ということについてはいくつか思い当たるふしがあります。オーストラリアは豊かな工業国家であり，十分に教育システムが発達しています。そういう教育システムの中で成長した人の人生経験は，ワーキングメモリの容量のテストを設計した人と似たような経験をたくさんしているでしょう。この共有された経験というのがどうやって “テストをうける” のかについて，似たような信念を形成しやすくするかもしれませんし，心理学的な実験がどういうものかについて共有された仮定があるかもしれません。こうしたことが，実際に影響するかもしれないのです。例えば “テストを受ける” というスタイルはオーストラリア人の実験参加者に，同じような環境で育っていない人に比べて，かなり抽象的なテストの要素に限定的な注意を向けるやり方を教えたかもしれません。このことがワーキングメモリの容量が何であるか，ということを考えるにあたって，誤解させるように導いてしまうかもしれないのです。

この議論には 2 つのポイントがあります。第一に，あなたが研究をデザインするとき大事なのは，何が母集団なのかに注意を払わなければいけないこと，そしてその母集団に適したやり方でサンプルをとることに注力すべきということです。実際には，みなさんは普通 “便利なサンプル” をとりたくなるでしょう (例えば，心理学教員が，データを集めるのが最もコストが低く，財源が金で溢れかえっているわけではないという理由で心理学の学生からサンプリングするなど) が，もしそうするのなら，少なくともこの槍か t がどんな危険を孕んでいるのかについて，しっかり考えてみるべきでしょう。第二に，あなたが誰かの研究について，人類全体からの無作為標本ではなくコンビニエンス・サンプリングをしているからという理由で批判するとしたら，少なくともそのことが*どれほど*結果を歪めたのかについてのしっかりした理論を提示する礼儀があるだろう，ということです。

1.1.5　母集団パラメータと標本統計

オーケイ。無作為標本の方法論的んな問題は少し横に置いて，違う問題に目を向けてみましょう。ここまで私たちは科学者のいうところの母集団について考えてきました。心理学者にとって，母集団とはひとの集団ということになるでしょう。環境学者にしてみれば，母集団がクマの集団になるかもしれません。ほとんどの場合において，科学者が考える母集団というのは現実世界に存在する具体的な何かです。しかし，統計学者は少し変わったひとたちなのです。一方では，彼らは科学者と同じように現実的な科学と現実世界のデータに興味が*ある*のです。他方で，彼らは数学者が考えるような純粋に抽象的な領域を操作しようとも思っています。その結果，統計的な理論は母集団をどのように定義するかについて，少し抽象的なものになる傾向があります。心理学の研究者が，我々の抽象的で理論的な概念でもって具体的な測定ができるようにする (セクション **??**) のと同じやり方で，統計学者はそれがどう働くかがわかっている数学的対象の用語で “母集団” の概念を操作可能にするのです。これについては，第 **??**章で既に触れているのです。それは確率分布と呼ばれるものです。

アイデアは本当に単純です。IQ スコアについて考えてみましょう。心理学者にとって，興味ある母集団とは IQ スコアを持っている実際の人による集団です。統計学者は母集団を図 **??**a に描かれているような確率分布として操作的に定義することで，これを “単純化” します。IQ テストは平均 IQ が 100 で，標準偏差が 15 の，正規分布に従うようにデザインされています。この値は母集団全体の特徴であり，**母集団パラメータ**として参照することができます。すなわち，我々は母平均 μ が 100 で，母標準偏差 σ が 15 ということができます。

ここで，私が実験しようとしているとしましょう。私は 100 人をランダムに選びだして，IQ テストを実施することで母集団からの単純無作為標本を得ます。私のサンプルが次のような数字から構成されているとしましょう:

```
106 101 98 80 74 ... 107 72 100
```

これらの IQ スコアそれぞれは，平均 100 で標準偏差 15 の正規分布から得られた標本です。ですから私が得たこの標本のヒストグラムを描けば，図 **??**b のようになるでしょう。ご覧になれば分かる

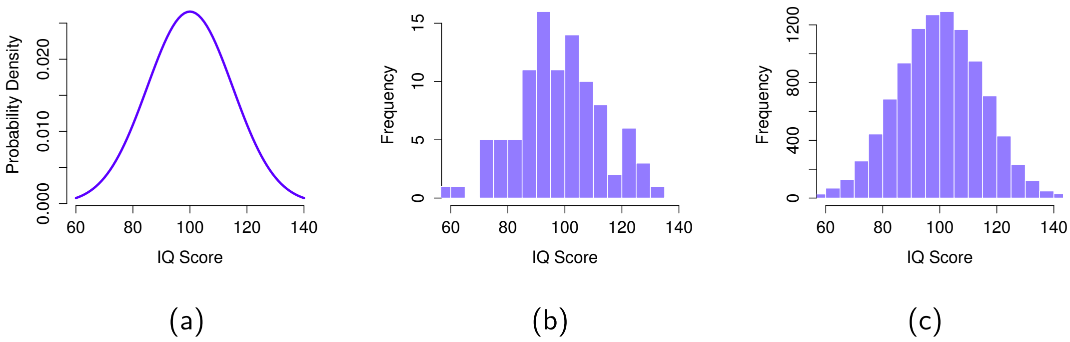


Figure1.4　IQ スコアの母集団分布 (パネル a) と，そこからランダムに抽出された二つの標本。パネル b には 100 の，パネル c には 10,000 人のサンプルが観測されています。

ように，このヒストグラムは図 **??**a にある本当の母集団をみくらべると，　だいたい正しい形をしていますが，非常に荒い近似でしかないことがわかります。標本の平均を計算すると，母集団の平均である 100 にかなり近い数字を得るでしょうが，ピッタリ同じとはいきません。今回の場合，私の標本の IQ スコアの平均は 98.5 で，IQ スコアの標準偏差は 15.9 でした。この**標本統計量**は，私のデータセットの特徴であり，真の母集団の値にかなり近くはありまあすが，同じものではありません。一般に，標本統計量は自分のデータセットから計算できるもので，母集団パラメータはあなたが知りたいと思っているもの，です。この章の後半で，母集団パラメータを標本統計量から推測する方法について話します (セクション **??**) し，その推定にどの程度確信を持てるかを表す方法について議論します (セクション **??**) が，その前に知るべきサンプリング理論についてのいくつかの概念があります。

1.2 大数の法則

前セクションでは，架空の IQ 実験例で，サンプルサイズ $N = 100$ ということでした。真の母集団平均が 100 で，標本平均が 98.5 ですから，まあまあ妥当な近似として勇気づけられる結果でしたね。多くの科学研究において，この正確さのレベルは完璧に受け入れられるものですが，状況が違えばもっと正確さが欲しいと思うかも知れません。もしもっと母集団パラメータに近い標本統計量が欲しいと思えば，何をすれば良いでしょうか?

その答えは明らかに，もっとデータを集めるということになるでしょう。もっと大掛かりな実験をして，今度は 10,000 人から IQ のスコアを得たとします。この実験の結果は JASP を使ってシミュ

レーションできます。IQsim.jasp ファイルが JASP のデータファイルです。このファイルには，mean = 100 と sd = 15 の正規分布する母集団から，10,000 点の無作為標本を得たものが入っています。ところで，これは JASP の新しい変数を作る機能から，R コード rnorm(10000, 100, 15) でもって作ったものです。この大サンプルのヒストグラムと密度プロットは，小サンプルのものよりも真の母集団分布によりよい近似を見せています。標本統計量にもこれが反映されています。大サンプルの IQ の平均は 100.107 で標準偏差は 14.995 です。この値は今や，真の母集団の値にかなり近くなっています。図 **??**をみてください。

こんなことを言うとちょっと馬鹿馬鹿しく感じるんですが，ここで皆さんに汲み取ってもらいたいのは，より大きなサンプルがあればより良い情報をもたらしてくれますよ，と言うことです。馬鹿馬鹿しく感じると言うのは，わざわざ言う必要がないほど明らかなことだからですね。事実，Jacob Bernoulli という確率理論の始祖の一人がこのアイデアを 1713 年に定式化したとき，彼もこれにちょっと変な感じを覚えたわけです。この直感を共有していることを，彼は次のように表現しています。

> *最も愚かな男であっても，本質的な直感によって，あるいは自分自身で，何の指示がなくても (これは驚くべきことだが)，観測が増えれば増えるほど目的の不明確さがより少なくなっていくことについては，確信を持っている。*(**Stigler1986**)

たしかに，この表現は少しばかり人を見下したような感じですが (性差別的であることは言うまでもないですが)，彼の主たるポイントは正しいのです。より多くのデータがあれば，より良い答えができる，というのは全く明らかなことです。問題は，なぜそうなのか，と言うことです。驚かないでほしいのですが，私たちが共有しているこの直感が正しいことがわかり，統計家はこれを**大数の法則**と呼んでいます。大数の法則は数学的な法則で，多くの異なる標本統計量に適用されますが，最も単純に考えるならば，平均 average に関する法則だということになります。標本平均は平均にかんする統計量の例として最もわかりやすいもので (だって平均 mean というのは... 平均 average のことですから)，これでみてみましょう。大数の法則が入っていることを標本平均に応用する時は，標本がより多く手に入れば，標本平均が真の母平均に近づいていくということを言ってることになります。あるいは，もう少し正確に言うならば，標本サイズが無限大に "近づく" ($N \rightarrow \infty$ と書きます), とき，標本平均ガボ平均に近づく ($\bar{X} \rightarrow \mu$)，ということです。[*3]

大数の法則が正しいことを証明しろとは言いませんが，統計理論の中で最も重要なツールの一つであることは間違いありません。大数の法則は，より多くのデータが私たちを真実に導いてくれる，という信念を正当化するのに使えます。それぞれのデータセットについて計算している標本統計量

[*3]技術的には，大数の法則は独立した量の平均として記述される標本統計に関するものです。標本平均はまさにこれにあたります。しかし他の多くの標本統計量も，ある種の平均として記述することができます。例えば標本分散はある種の平均であり，それは大数の法則にし違います。いかし，標本の最小値は，平均の形で描くことができないので，大数の法則に支配されないものです。

は間違っていますが，大数の法則は，より多くのデータを集めれば，それらの標本統計量は真の母集団パラメータにどんどん近づいていくことを教えてくれます。

1.3

標本分布と中心極限定理

大数の法則はとても強力なツールですが，私たちの全ての問いに応えてくれるのに十分というわけではありません。特に，それが与えてくれるのは "長期保証" でしかないのです。長期というのは，我々が何とかしてデータの収集を無限に続けられれば，大数の法則は標本統計量が正しくなることを保証してくれる，と言うことです。しかしジョン・メイヤード・ケインズが経済学の文脈で言った有名な言葉にあるように，長期保証は実際の人生においてあまり役立つものではありません。

> *長期保証は現在の問題を考える上でミスリーディングを招く。長期的に見れば，我々はみな死んでしまうのだから。経済学者はこれをあまりにも簡単に，あまりにも役に立たないタスクを設置した。荒天の季節に彼らが言えるのは，いずれ嵐は去るし，海も穏やかさを取り戻すということだけだ。*(**Keynes1923**)

経済学の例にあるように，心理学や統計学にも同じことが言えます。標本平均を計算する時に，*最終的に*正しい答えに到達することを知っていると言うだけでは，十分ではありません。十分に大きなデータセットを持っていると母平均の正確な値になることを知っていても，*実際のデータセットの*サンプルサイズが $N = 100$ でしかないときには，悲しい慰めにしかなりません。現実では，より控えめなデータセットから計算された標本平均の振る舞いについて，知っておかなければなりません!

1.3.1 平均の標本分布

このことを心に留めおいて，私たちの研究がいずれ標本サイズ 10,000 に到達するだろうという考えを捨て，もっと控えめな実際の実験について考えることにしましょう。今回は，$N = 5$ のサンプルをとって，IQ スコアを測定したとしましょう。前と同じように，JASP でこの実験をシミュレートします。`rnorm` 関数を変更して，`IQsim` というデータ列を作りました。`IQsim` ラベルの横にある f_x をダブルクリックすると，JASP は '計算列' ダイアログを表示し，そこには R コードで `rnorm10000,100,15` と書いてあるでしょう。今回は被験者 5 人分だけでいいので，10000 を 5 に変えて '列を計算する' とするだけです (図 **??**をみてください。)。JASP が私のために 5 つの数字を生成してくれました (あなたの値はきっと違うものになっているでしょう)。便宜上，数字は整数に丸めてあります。

```
124 74 87 86 109
```

Table1.1　IQ 実験の再現，毎回標本サイズは $N = 5$ です。

	1 人目	2 人目	3 人目	4 人目	5 人目	標本平均
再現 1 回目	124	74	87	86	109	96.0
再現 2 回目	91	125	104	106	109	107.0
再現 3 回目	111	122	91	98	86	101.6
再現 4 回目	98	96	119	99	107	103.8
再現 5 回目	105	113	103	103	98	104.4
再現 6 回目	81	89	93	85	114	92.4
再現 7 回目	100	93	108	98	133	106.4
再現 8 回目	107	100	105	117	85	102.8
再現 9 回目	86	119	108	73	116	100.4
再現 10 回目	95	126	112	120	76	105.8

....

今回のサンプルにおける IQ の平均は 96 ちょうどになります。驚くことはないですが，これは先ほどの実験よりも正確さの面で劣ります。次にこの実験を**再現する**ことにしたと思ってください。つまり，私がこの手続きをできるだけ同じように繰り返し，新しく 5 人のサンプルを取って IQ を測定したとします。もう一度 JASP を使って，この手続きによる結果をシミュレートし，5 つの数字を生成しましょう。

```
91 125 104 106 109
```

今回，IQ の平均は 107 になりました。もしこの実験を 10 回繰り返したら，表 **??**にあるような結果を得て，標本平均が毎回の再現実験ごとに変化することがわかります。

このやり方をずっと続けっとしましょう。この “5 つの IQ スコア” の再現実験を，何度も何度もするのです。この実験を繰り返すたびに，標本平均を記録していきます。時間が経つにつれて，新しいデータセットを蓄積していきます。毎回の実験が 1 つのデータポイントを生むのです。私のデータセット例では，最初の 10 回の標本平均が表 **??**にありますが，次のようにデータが始まっています。

```
96.0 107.0 101.6 103.8 104.4 ...
```

これを 10,000 回繰り返して，ヒストグラムを書いたらどうでしょう。まさにそれをしたのが，図 **??** にあります。この図を見るとわかるように，5 つの IQ スコアの平均は，普通 90 から 110 の間に入るようです。しかしより重要なこととして強調すべき点は，私たちがこの再現実験を何度も何度も繰り返すと，最終的には標本平均の*分布*を得られると言うことです! この分布は統計学において特別な

名前を持っていて，**平均の標本分布**といいます。

標本分布は統計学におけるもう 1 つの重要な理論的アイデアあり，小さいサンプルの振る舞いを理解するのに欠かせないものです。例えば，私が最初に行った "5 つの IQ スコア" 実験では，標本平均は 96 でした。図 **??**にある標本分布が教えてくれることは，この "5 つの IQ スコア" 実験はそれほど正確ではないということです。実験を繰り返したとき，標本分布が教えてくれるのは標本平均が 80 から 120 の間のどこかにあるのかなあ，と想像できます。

1.3.2 どんな標本統計量にも標本分布は存在する!

標本分布を考える時に覚えておいてほしいことは，あなたが注目しようとしている*あらゆる*標本統計量について標本分布があるということです。たとえば，またしても "5 つの IQ スコア" 実験を繰り返して，IQ スコアの最大値を書き出したとしましょう。これをするとデータセットは次のようになります:

```
124 125 122 119 113 ...
```

これを繰り返すと，かなり変わった標本分布が得られます。言うならば*最大値の標本分布*です。5 つの IQ スコアの最大値の標本分布は，図 **??**に示しました。おどろくなかれ，5 人をランダムに取り出して，IQ 最大値の人を見つけ出したら，その人は IQ の平均より大きくなるでしょう。ほぼ毎回，IQ が 100 から 140 の範囲で測定されたひとと一緒になってしまうでしょう。

1.3.3 中心極限定理

ここまでくると，標本分布が何なのかについてのちょっとした感覚を掴んでもらったと思います。特に，平均の標本平均 g どんなものかについて。このセクションでは，平均の標本分布がサンプルサイズによってどのように変わるかについて説明していきましょう。直感的には，あなたは既に答えの一部を知っているはずです。観測度数が少ない時は，標本平均はそれほど正確ではありません。小サンプルの実験を繰り返し，平均を何度も計算すると，結構異なる答えを得ることになります。言い換えると，標本分布は非常に幅広いのです。大サンプルの実験を繰り返し，平均を何度も計算すると，同じような答えを得るでしょうし，標本分布は非常に狭くなるでしょう。このことは図 **??**で見ることができます。より大きなサンプルサイズをもてば，より幅の狭い標本分布を得ることが f できるのですね。この効果を評価するためには標本分布の標準偏差を計算すればよく，この数字は**標準誤差**と呼ばれています。統計量の標準偏差は，SE と表記されることが多いです。また標本*平均*の標準誤差に興味があることが多いですから，SEM と書くことがあります。図を見たらわかるように，サンプルサイズ N が増加すると SEM は減少していくのです。

さて，それは話の一部にすぎません。これまで見逃してきたことが，ここにはあります。ここまで提示してきた例は全て，“IQ スコア” 実験に基づくものでした。というのも IQ スコアは，母集団分布が正規分布であると化したので，ほぼ正規分布に従うだろうと考えられるからです。もし正規分布じゃなかったら? 平均の標本分布に何が起こるでしょうか? これについて躍るべきことに，母集団分布がどんな形であっても，N が増えれば平均の標本分布はどんどん正規分布に近づいていくのです。これを理解してもらうために，シミュレーションしてみましょう。そのために，図 **??**のヒストグラムで示したような “傾いた” 分布から始めましょう。黒い線で示されたベルカーブと三角形のヒストグラムを比較すればわかるように，母集団分布は正規分布とは全く似ていないものになっています。次に，たくさんの実験結果をシミュレートします。各実験では，$N = 2$ のサンプルを取ってきて，その標本平均を計算します。図 **??**b のプロットは，これらの標本平均ヒストグラムです (つまり $N = 2$ の平均の標本分布です)。今回，ヒストグラムは ∩ 型の分布をしています。これはまだ正規分布の形ではありませんが，図 **??**a にある母集団分布よりは近づいています。サンプルサイズを $N = 4$ に増やしてみると，平均の標本分布はかなり正規分布に近くなり (図 **??**c)，サンプルサイズが $N = 8$ に至るとほとんど完璧に正規分布の形になります。言い方を変えると，サンプルサイズが小さすぎなければ，平均の標本分布は正規分布に近づいていくのです。あなたの考える母集団分布がどんな形であっても!

これらの図に基づいて，平均の標本分布について以下のような主張に対する根拠を手に入れたと言えるかもしれません。

- 標本分布の平均は母集団の平均と同じ。
- 標本分布の標準偏差 (つまり標準誤差) はサンプルサイズが増えると小さくなる。
- 標本分布の形状はサンプルサイズが増えると正規分布になる。

実は，これらの主張は正しいだけでなく，統計学においてこれら 3 つ全てを証明した有名な定理があり，それが**中心極限定理**なのです。特に，中心極限定理は，もし母集団分布が平均 μ で標準偏差 σ であれば，平均の標本分布も平均 μ で，平均の標準誤差は

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

となるということを示しています。母標準偏差 σ がサンプルサイズ N の平方根で割られているので，SEM はサンプルサイズが大きくなるとどんどん小さくなります。これはまた，標本分布の形状が正規分布になることも教えてくれます[*4]

[*4] いつものように，私はここで少し手抜きをしています。中心極限定理は，このセクションで述べていることよりももう少し一般的な含みを持っています。統計学のもっと入門的なテキストのように，私は中心極限定理が成立するある状況について議論してきました。すなわち，同じ分布からそれぞれ独立して生じた事象について平均を取っている，と言うことです。しかし，中心極限定理はこれよりも広い範囲をカバーしています。例えば，“U-統計” と呼ばれるものがありますが，そこに含まれるのは全て中心極限定理を満たしますから，大きなサンプルサイズであれば正規分布になります。平均はそうした統計量の一つですが，唯一のものではないのです。

この結果はあらゆることについて便利なものです。これはなぜ大きな実験が小さな実験よりも信頼できるのかを教えてくれますし，標準誤差の明示的な公式を与えてくれますから，より大きな実験がどのていど信頼できるのかについて教えてくれるのです。また，なぜ正規分布が*正規*なのかも教えてくれます。実際の実験では，私たちが測定したいと思っていることの多くが異なる量の平均を取りますし (例えば，IQ で測定されるような “一般的な” 知能は，多くの “特別な” 技術や能力の平均であるでしょう)，そうした時に平均化された量が正規分布に従うことになります。この数学的な法則によって，正規分布は実際のデータにおいて何度も何度も出てくることになるのです。

1.4 母数の推定

前のセクションでお見せした全ての IQ の例においては，*母*集団におけるパラメータを事前に知っていたのでした。知能の測定を最初に学ぶ学部学生はみな，IQ スコアが平均 100 で標準偏差 15 に*定義されている*ことを学びます。しかし，それはちょっと嘘です。IQ スコアの真の*母*集団平均が 100 であることをどうやって知り得るのですか? そう，私たちがこれを知っているのは，テストを設計した人たちがとても多くのサンプルを取って，その標本平均が 100 になるようにスコアリングルールを “仕組んだ” からです。もちろんこれは悪いことではなくて，心理学的測定を設計する上で重要なことでもあります。しかし，この理論的な平均が 100 であるというのが，このテスト設計者がテストを作る時に使った*母*集団に合うようにしたものである，ということを心に留めておくことはじゅうようです。良いテスト設計者は，多くの異なる母集団 (年齢層や国籍が異なる集団など) に適用できるように “テストの基準” を提供しようとします。

これは便利なことではありますが，もちろんほぼ全ての研究プロジェクトにおいて，テストを基準化するのに使ったのとは異なる*母*集団を調査することになります。たとえば，鉛製錬所がある南オーストラリアの工業都市ポートピリーで、低レベルの鉛中毒が認知機能に及ぼす影響を測定しようとしたとしましょう。あなたはポートピリー在住の人の IQ スコアを，製鉄所のある南オーストラリア

の工業都市ワイアラのサンプルと比較するでしょう[*5]どちらの町について考えるにしても，単に真のIQ 母集団平均が 100 であると*想定する*のは意味がありません。私の知る限り，南オーストラリアの工業地帯に’自動的に適用できる，正規化されたデータを作っている人はいません。私たちはデータのサンプルから母集団を*推定する*必要があるのです。どうやればいいでしょう?

1.4.1 母平均を推定する

ポートピリーにいって，100 人の地元民に IQ テストを受けてもらったと考えてください。この人たちの IQ スコアの平均は，$\bar{X} = 98.5$ であることがわかったとします。ポートピリーの住民全体の，真の平均 IQ はどうなるでしょう? 明らかに，この質問に対する答えは知りようがないのです。97.2 かもしれないし，103.5 かもしれないわけです。私たちのサンプリングは網羅的なものではないのですから，明確な答えを与えることができません。とはいえ，“一番ましな推測” をすることに狙いを定めたら，それは 98.5 ではないかと私なら言うでしょう。これが統計的推測のエッセンスです。つまり，一番ましな推測をする，ということ。

今回の例で未知の母数を推定するのは，直接的なやり方でした。標本平均を計算して，これで**母平均の推定**としたわけです。とてもシンプルで，次のセクションではこの直感的な答えの統計学的な正当化を与えようと思います。しかし，私がここでやろうとしていることは，標本統計とお数の推定値は概念的に異なるものだということを，あなたにしっかり理解してもらうことなのです。推定値が母集団についての推測であるのに対して，標本統計はあなたのデータを記述するものです。心にとどめておいてほしいのは，統計学者はこれらに言及する時違う書き方をするということ。たとえば，真の母平均が μ と書き表されるとすると，母平均の推定値を $\hat{\mu}$ とします。これに対して，標本平均は $\bar{X}$ や m と書いたりします。しかし，単純無作為抽出したとき，母平均の推定値は標本平均と同じです。標本平均が $\bar{X} = 98.5$ のとき，母平均の推定値もまた，$\hat{\mu} = 98.5$ なのです。この表現をはっきりさせておくために，かんたんな表を用意しました。

[*5] もしあなたが*実際に*この問題を扱おうとするなら，私がここでやろうとしていることよりもっと注意深くやらなければならないことに注意してください。ワイアラとポートピリーの IQ スコアだけを比較して，その違いがすべて鉛中毒のせいだというわけにはいかないのです。もし 2 つの都市の違いが製油所の違いだけにあったとしても (長期的はそうでなかったとしても)，人々が既に汚染が認知的な障害につながると*信じている*という事実を説明する必要があります。第 **??**章を思い出してください，つまりポートピリーサンプルとワイアラのサンプルでは需要の影響が異なることを意味しています。言い換えるなら，人が実際に違いがあると*考えている*ことが，あなたのデータに幻の差を作り出したのかもしれないのです。もし研究者の一団がポートピリーに白衣で現れて IQ テストを始めたとすると，地元の人たちがあなた方がやろうとしていることについて全く何も気づかないとは思えませんし，多くの人があなたがやろうとしていることに腹を立てることさえあり得るでしょう。そういう人たちはテストに協力的ではないでしょう。ポートピリー以外の人はもうすこしうまくやろうと思うでしょう。自分達の街が悪くみられるのは嫌でしょうから。この動機の違いがワイアラではより弱いでしょう。人は “鉛中毒” の概念と同じ意味での “鉄鉱石中毒” の概念がないからです。心理学というのは*難しい*のです。

記号	これは何?	コレが何だかわかってる?
$\bar{X}$	標本兵器	もちろん，データから計算できますから
μ	母平均	おそらく決して知り得ないでしょう
$\hat{\mu}$	母平均の推定値	イエス，単純無作為標本の標本平均と一致します

1.4.2 母標準偏差を推定する

ここまで，推定はとてもシンプルで，あなたはなぜ私がこんなに面倒なサンプリング理論についての話を読ませせようとしているのか不思議に思っているかもしれません。母平均を推定したもの (すなわち $\hat{\mu}$) は，標本統計量 (すなわち $\bar{X}$) と同じことがわかっています。残念ながら，これが常に真だとはならないのです。これを見るために，**母標準偏差の推定**，つまり $\hat{\sigma}$ をどうするか考えてみましょう。今回は推定のために何を使いましょうか? まず思いつくのは，平均を推定した時と同じようにやあることで，標本標準偏差を推定値に使うことですよね。これはほとんどあっているのですが，厳密には違います。

なぜでしょうか。観測度数が 1 しかないサンプルを集めてきたと思ってください。この例では，母集団の真の値について全く直感が働かない時の標本を考えるのがよく，全くのフィクションの例を使うことにしましょう。私のシューズの*クロミュランス*を測定することにしたとします。私のシューズのクロミュランスが 20 であることがわかりました。これが私の標本です。

20

サンプルサイズが $N = 1$ でしかありませんが，これはまったく正当なものです。標本平均は 20 で，というのもこの標本のどれもが標本平均と同じだからです (当たり前ですが!)。そして標本標準偏差は 0 です。*標本*の記述として，これは全く正しいことです。というのも標本が一度数しかなく，標本の散らばりがないからです。標本標準偏差が $s = 0$ となるのは，ここでは正しい答えです。しかし*母集団の*標準偏差を推定するとき，これでは意味がありませんね? たしかに，あなたも私も“クロミュランス”が何なのか全くわかっていませんが，データについては知っているのです。*標本*について全く分散を読み取ることができない唯一の理由，それはいかなる分散も見ることができないほど小さなサンプルだからです! ですから，サンプルサイズ $N = 1$ のときの正しい答えは，“さっぱりわからない”というのが正しいように*思えます*。

注意してほしいのは，標本平均と母平均の時に，同じような洞察をすることはないということです。母平均の最適な推定をしなければならない時，母平均が 20 であると推測することは，全く無意味なものではありません。もちろん，あなたはおそらくこの推定に十分な確信を持てないでしょう。というのも，たった 1 つの観測しかしてないからですが，しかしベストな推定をしてことに変わりはないのです。

この例を少し拡張してみましょう。2 回目の観察をしたと思ってください。今やシューズのクロミュランスデータセットは $N = 2$ になり，サンプルが次のようになったとします。

20，22

今回，私たちのサンプルは*少しばかり*大きくなり，ある程度の分散を認められるようにはなりました。すなわち，なんらかの分散を観測するのに最低限必要な数が，観測度数 2 ということです! この新しいデータセットによって，標本平均は $\bar{X} = 21$ であり，標本標準偏差は $s = 1$ になりました。母集団についてどういうことがわかるでしょう? 繰り返しますが，母集団の平均については標本平均がベストな推定値なのです。標準偏差についてはどうでしょう? これはもう少し複雑です。標本標準偏差は，たった 2 つの観測に基づいて行われますが，もしあなたが私と同じような人であれば，たった 2 件の観測度数だけでは母集団の真の変動を明らかにするには，“十分な変動” とは言えないと直感的にわかるでしょう。推定が*間違っている*というだけでなく，たった 2 つの観測から推測するのはある程度間違っていることが予想されますね。このエラーについての不安は*システマチック*なものです。特に，標本標準偏差が母標準偏差よりも小さくなるのでは，と予想できます。

この直感は正しいのですが，もう少しよいデモンストレーションをすることができます。この直感を数学的に証明できるというのも事実ですが，これは正しい数学的知識がなければあまり役に立ちません。そのかわりに，ある実験の結果をシミュレートしてみましょう。それを踏まえて，IQ 研究の話に戻りましょう。真の母平均が 100 で母標準偏差が 15 だとします。まず，$N = 20$ の IQ スコアで実験を計画し，標本標準偏差を計算したとします。これを何度も繰り返し，標本標準偏差のヒストグラムをプロットすると，*標準偏差の標本分布*を得ることができます。これを図 **??**にプロットしました。真の母標準偏差は 15 ですが，*標本の*標準偏差の平均は 8.5 にしかなりません。図 **??**b にあるのは，これとは少し違って平均の標本分布をプロットしているのですが，そこには母平均が 100，標本平均の平均も 100 であることが示されています。

このシミュレーションを拡張してみましょう。$N = 2$ に限定せず，サンプルサイズを 1 から 10 まで変えて実験を繰り返したとします。標本平均の平均と，標本標準偏差の平均をサンプルサイズの関数としてプロットした結果が，図 **??**に示してあります。左の図 (パネル a) には標本平均の平均を，右の図 (パネル b) には標準偏差の平均をプロットしています。二つの図は*平均でみると*全く違うようで，標本平均の平均は母平均に等しくなります。これは**不偏推定量**で，だからこそ標本平均が母平均の最も良い推定値になりうる理由なのです[*6]右の図は母平均からずれており，標本標準偏差 s は母標準偏差 σ よりちいさくなっています。

これが**偏った推定量**なのです。言い換えると，母標準偏差 σ についての “最も良い推定” である $\hat{\sigma}$

[*6] ここで少し隠し事をしていることを打ち明けておきましょう。不偏性は推定量についての望ましい特徴ですが，偏り以外にも重要なことがあります。しかし，あらゆる細部にまでわたって議論するのは，この本の範疇を超えてしまいます。ですから，ここには複雑な背景が隠されていることを示唆するにとどめます。

をしたいという時は，標本標準偏差 s より少し大きめにして推定するべきなのです。

この系統的なバイアスを修正するのは，実にシンプルなやり方でできます。ここにどうするかかしめしました。表分偏差を追いかける前に，分散を見てみましょう。セクション **??**を思い出してほしいのですが，標本標準偏差は標本平均からの偏差を二乗したものの平均として定義されたのでした。このように。

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

標本分散 s^2 は母分散 σ^2 にバイアスのかかった推定量です。しかし，結局のところ，ちょっとした変換を行うだけで不偏推定量に変えることができます。やるべきことは，N ではなく $N-1$ で割るようにするだけです。そうすると，次のような式になりますね。

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

これが母分散 σ の不偏推定量です。さらに，これがセクション **??**で持ち上がった問いに対する最終的な答えになります。なぜ JASP はちょっと違う分散の答えを返すのか? それは JASP が $\hat{\sigma}^2$ を計算していたから，というわけです。同じような話は，標準偏差にも関わります。母標準偏差の推定値にするには，N ではなく $N-1$ で割ったものを使うのです。

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

そして JASP の標準偏差に組み込まれている関数は，s ではなく $\hat{\sigma}$ の計算をしているわけです。[a]

[a]オーケイ，私はここでもあることを隠しています。奇妙で直感に反することですが，$\hat{\sigma}^2$ は σ^2 の不偏推定量ですから，それの平方根をとった $\hat{\sigma}$ は σ の不偏推定量だと思うでしょう。本当かな? 実はそうではないのです。それだと $\hat{\sigma}$ に実際少しのバイアスが存在するのです。これはまさにおかしなことですよね。すなわち，$\hat{\sigma}^2$ は母分散 σ^2 の不偏推定量なのですが，その平方根を取った $\hat{\sigma}$ は母標準偏差 σ の推定量としてバイアスがあるなんて。おかしいなあ，おかしいぞ，本当かな? ではなぜ $\hat{\sigma}$ はバイアスがあるんでしょう?これについての技術的な答えは，“(平方根のような) 非線形変換は，期待値と一致しないから” ですが，数理統計のコースを取ってない人にとってはよくわからないでしょうね。幸いにして，実践的な目的のためにはこれは大きな問題にならないのです。そのバイアスはとても小さくて，実際には $\hat{\sigma}$ を使っても問題ありません。ときどき数学というのは煩わしいだけのものですね。

最後にひとつ。実際には，多くの人が $\hat{\sigma}$(つまり $N-1$ で割ったもの) を*標本*標準偏差といいがちです。技術的には，これは正しくありません。*標本*標準偏差は s(つまり N で割ったもの) に一致するべきです。概念的にも，計算上も，同じことではありません。一方は標本の特徴の一つであり，他方は母集団の特徴を推定したものです。しかし，実際に応用する時に我々が本当に気にしているのは，母集団のパラメータですから，人は報告する時に $\hat{\sigma}$ を s よりも使ってしまうのです。もちろん正しい数字を報告していることにはなります。ただちょっと，“標本標準偏差” が “母標準偏差の推定値” よりも短くて言いやすいので，文章化する時にちょっと正確でない用語を使ってしまいがちな

のですね。大問題ではないですし，実践上同じようなことを私もしてしまうことがあります。とはいえ，この 2 つの*概念*を区分しておくことは重要だと思うのです。“標本からわかること” と “それがやってきた母集団について推測したこと” を混乱させて使うことは良いことではないですね。あなたが s と $\hat{\sigma}$ について考えるのは同じことだと思った瞬間，まさにその間違いを犯し始めていると思ってください。

さてこのセクションを閉じるにあたって，これらの点をはっきりさせた一組の表を示しておきましょう。

記号	これは何?	これはわかるもの?
s	標本標準偏差	イエス，ローデータから計算できます。
σ	母標準偏差	たぶん永遠にわからないでしょう
$\hat{\sigma}$	母標準偏差の推定値	イエス，でも標本標準偏差と同じものではありません

記号	これは何?	これはわかるもの?
s^2	標本分散	イエス，ローデータから計算できます
σ^2	母分散	たぶん永遠にわからないでしょう
$\hat{\sigma}^2$	母分散の推定値	イエス，でも標本分散と同じものではありません

1.5 信頼区間を推定する

> *統計が意味するのは，あなたが確信していると言う必要がないこと* – 読み人知らず。*7

この章のポイントは，データの標本に基づいて母集団のパラメータを推測するのに必要な，統計学がよって立つサンプリング理論の基礎を大まかに説明することです。この議論からわかるのは，サンプリング理論が必要とする理由の 1 つが，あらゆるデータセットがある種の不確実さを残しており，我々の推定が完全に正確なものになることは決してないからです。この議論のなかで忘れられてきたことの 1 つが，我々の推定に伴う不確実さの量を*評価する*試みです。心理学の学生の IQ は 115 です，と推測するだけでは不十分なのです (ええ，この数字は私が作りました)。この推測がもってい

*7 この詩はあちこちの T シャツやウェブサイト，いくつかの学術論文でさえ言及しているのを見かけるのですが (たとえば，http://www.amstat.org/publications/jse/v10n3/friedman.html)，私は原典を見つけることができませんでした。

る不確実さの程度を表現する，なんらかの言い方が必要ですよね。たとえば，本当の平均は 109 から 121 の間に 95% の確率であります，というような言い方ができたらいいかもしれないですね。これが平均の**信頼区間**というやつです。

標本分布を理解するときに，平均の信頼区間を構成するのはごく簡単なことです。その仕組みはこうです。真の母平均 μ と，母標準偏差が σ だとしましょう。N 人の参加者を募った研究が終わったところで，この参加者から計算した IQ の平均が $\bar{X}$ だとします。中心極限定理 (セクション **??**) により，平均の標本分布は正規分布に近似することがわかっていますね。また，セクション **??** の議論を思い出すと，正規分布の 95% 確率は，真の平均のまわり 2 標準偏差の中にあることもわかっています。

もう少し正確に言いますと，正規分布の 95% の確率は真の平均まわり 1.96 標準偏差のなかにあるのです。次に，標本分布の標準偏差は標準誤差，とくに平均の標準誤差は SEM というのでした。これらを併せて考えると，標本平均 $\bar{X}$ は，母平均の 1.96 標準誤差の中で観測される確率が 95% だということになります。

数学的には，このように表記します。

$$\mu - (1.96 \times \text{SEM}) \leqslant \bar{X} \leqslant \mu + (1.96 \times \text{SEM})$$

ここで SEM は $\sigma/\sqrt{N}$ に一致し，これが真であると 95% の確信をもって言えるのです。しかし私たちが本当に興味を持っている問いに対する答えにはなっていませんね。上の式は母集団パラメータがわかっている時に，標本平均に何が期待できるかを教えてくれるにすぎません。私たちが*知りたい*のは，これの逆の働きです。私たちが知りたいのは，特定の標本を観測することができた時に，母集団パラメータについて何が言えるかなのです。しかし，これはそれほど難しいことではありません。高校数学をちょっと使って，この式を次のように書き換えてみましょう。

$$\bar{X} - (1.96 \times \text{SEM}) \leqslant \mu \leqslant \bar{X} + (1.96 \times \text{SEM})$$

これが教えてくれることは，母平均 μ を 95% の確率で含む値の範囲です。We refer to this range as a **95% confidence interval**, denoted CI_{95}. 要するに，N が十分大きければ (平均の標本分布が正規分布であると信じるのに十分であれば)，この式を書いて 95% 確信区間とすることができるのです。

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}}\right)$$

もちろん，1.96 という数字に特別な意味はありません。これはあなたが 95% 信頼区間を使おうと思った時にかけるための数字，というだけです。もし 70% 信頼区間が欲しければ，マジックナンバー 1.96 ではなく 1.04 をつかうでしょう。

1.5.1　式中のちょっとした間違い

いつものように，私は嘘をつきました。95% 信頼区間を算出する上の式は近似的に正しいのですが，議論の中で重要な詳細を見落としていたのです。平均の標準誤差，SEM に使った式をよく見ると，あなたは真の母標準偏差 σ を知っていなければなりません。そう，セクション **??** にあったように，私たちは真の母集団パラメータについて*知り得ない*という事実に悩まされるのです。真の値 σ を知らないのですから，その代わりに母標準偏差の推定値 $\hat{\sigma}$ を使うしかありません。これは真っ当なやり方に見えますが，そうすることでマジックナンバーを計算する時に正規分布ではなく t 分布のパー線タイルを使う必要があるのです。そしてその答えは，サンプルサイズに依存します。N がとても大きいときは，t 分布でも正規分布でも同じ値になります。そう，1.96 ですね。しかし N が小さいときは，t 分布を使うともっと大きな数字，2.26 になります。

何が起こったのか，それはそんなにミステリアスなことではありません。大きな値になるということは信頼区間が広がるということで，これは真の μ の値がどうなっているかについて我々はより不確実なことしかわからないのだ，ということです。正規分布の代わりに t 分布を使うときは，より大きな数字になるので，より不確実だということです。なぜ他の不確実さを持ってこないといけないのでしょう? ふむぅ，母標準偏差の推定値 $\hat{\sigma}$ が何かおかしいに違いありません! もしそれが間違っていれば，平均の標本分布が実際にどんな形になっているかについて，少しばかり確信が減ることになり，この不確実さがより広い信頼区間に反映されることになると言えるでしょう。

1.5.2　信頼区間の解釈

信頼区間について難しいのは，それが*意味する*ものは何かを理解することです。人が初めて信頼区間に触れたとき，最初の直感は決まって，“信頼区間の内側に真の平均が 95% の確率で存在する” と思わせてくるのです。これはシンプルで，“95% の確実さ” という言葉が意味する常識的な意味を表していると思われます。残念ながら，これは少し正確ではないのです。この直感的定義は，母平均の値に関するあなたの個人的な*信念*に深く関わっていますね。私は 95% の確信があると言いましたが，それは私の信念だからです。日々の生活ではこれで問題ないのですが，もしセクション **??**に戻って考えれば，個人的な信念についての表現はベイズの発想であることに気づくでしょう。しかし信頼区間はベイズのツールではないのです。*この章の他のところでもそうなのですが，信頼区間は頻度主義者のツールで，頻度主義者の手法を使うのであれば，ベイズ的解釈を与えるのは適切ではありませんね。頻度主義者のツールを使うのであれば，頻度主義者に適した解釈にしましょう!*

オーケイ，答えが正しくないというのであれば，正解はなんでしょう? ここで頻度主義的確率とはなんだったかを思い出してみましょう。“統計的な表現” が唯一許されるのは，連続的な事象についてで，頻度主義者は異なる事象の頻度を数え上げるのでした。この見方からすると，95% の信頼区間の解釈は，なんらかの反復についてのものでなければなりません。今回の場合はもし私がこの実験を何度も何度も繰り返して，毎回 95% 確信区間の計算を続けるなら，その区間の 95% は真の平

均を含んでいる，ということになります。もっと一般的に言えば，全信頼区間の 95% は真の母平均を含むこの手続によって構成されたものです。この考え方を図 **??**に示しました。ここでは “10 人の IQ スコア” 実験についての 50 の信頼区間 (上の図) と，“25 人の IQ スコア” 実験についての 50 の信頼区間 (下の図) を示しています。ちょっとした幸運がみられて，私がシミュレートした 100 回の反復実験を通じて，ちょうど 95 回が真の平均を含んでいました。

ここでの根本的な違いは，米事案は確立の言葉を母平均に対していうのですが (つまり，母平均についての我々の不確実さについての言及)，そうした言い方は頻度主義的な確率の解釈のもとでは許されないのです。なぜなら母集団を “繰り返し” たりしないのですから! 頻度主義的に言えば，母平均は固定されたもので確率的な表現はあり得ないのです。しかし信頼区間は反復可能です。実験を繰り返すことができるのですから。ですから頻度主義者は (確率変数が) 真の平均を含む信頼区間について確率的に話すことができますが，その話は（反復イベントではない）真の母平均が，信頼区間のなかに入る確率についてのことなのです。

これは少し衒学的に見えることはわかりますが、重要なことです。問題になるのは，解釈が数学的な違いにつながるからです。。信頼区間に代わるベイジアンのやりかたは，確信区間として知られています。ほとんどの場合，確信区間は信頼区間と同じようになりますが，時々まったく違うものになることがあります。約束しますが，ベイジアン的なものの見方は第 **??**章で論じます。

1.5.3 JASP における信頼区間の計算

JASP の今のバージョンでは，平均の信頼区間を計算するのに ‘記述統計’ の機能の一部として (まだ) 簡単にできるように実装されていません。ですが，‘記述統計’ に S.E. のチェックボックスがありますね。Mean, so you can use this to calculate the lower 95% confidence interval as: つまり，これを使って 95% 信頼区間の下限を次のように計算できます。`Mean - (1.96 * S.E. Mean)` そして 95% の上限は `Mean + (1.96 * S.E. Mean)` です。

95% 信頼区間は心理学における事実上の標準です。例えば，`IQsim.jasp` ファイルを読み込んで，‘記述統計’ から平均と S.E. を選択しまし，シミュレートされた平均 IQ の信頼区間を計算すると，次のようになります。

Lower 95% CI = 100.107 - (1.96 * 0.150) = 99.813

Upper 95% CI = 100.107 + (1.96 * 0.150) = 100.401

今回のシミュレートされた標本データは N=10,000 で，IQ の平均スコアは 100.107，95% 信頼区間は 99.813 から 100.401 になります。幸い，これの解釈はとても簡単にできますね。‘記述統計’ オプションの一部として信頼区間を計算するような，簡単な方法がまだ JASP にはありませんが，ほんのちょっとしたて計算で求めることができます。

どうように，JASP で信頼区間をプロットするのも，(まだ)JASP の ‘記述統計’ オプションは対応していません。ですが，第 **??**章で，ある統計的な検定について学んでいく中で，データ分析の一部

として信頼区間をプロットすることができるようになっていきます。これは本当にかっこいいことで，これについては後ほど説明するとしましょう。

1.6 要約

この章で私は主に 2 つのトピックスについて話しました。この章の半分はサンプリング理論についてのもので，残りの半分は母集団パラメータの推定にサンプリング理論をどう使うかということについてでした。このセクションでは，次のようにまとめておきましょう。

- *標本，サンプリング，母集団についての基本的な考え方 (Section* **??***)*
- *サンプリングの統計理論: 大数の法則 (セクション* **??***)，サンプリング分布と中心極限定理 (セクション* **??***)。*
- *平均と標準偏差を推定する (セクション* **??***)*
- *信頼区間を推定する (セクション* **??***)*

いつものように，この章でカバーできていない，サンプリングと推定についてのトピックスもたくさんありますが，心理学コースの導入としてはちょうどいい分量だと思っています。ほとんどの実践的研究者にとって，これ以上の理論が必要になることはありません。この章で触れていないある大きな問題としては，無作為標本のサンプルが手に入らなかった時のことです。この状況を乗りこなすための統計理論はいろいろあるのですが，この本の範疇を大きく超えてしまいます。

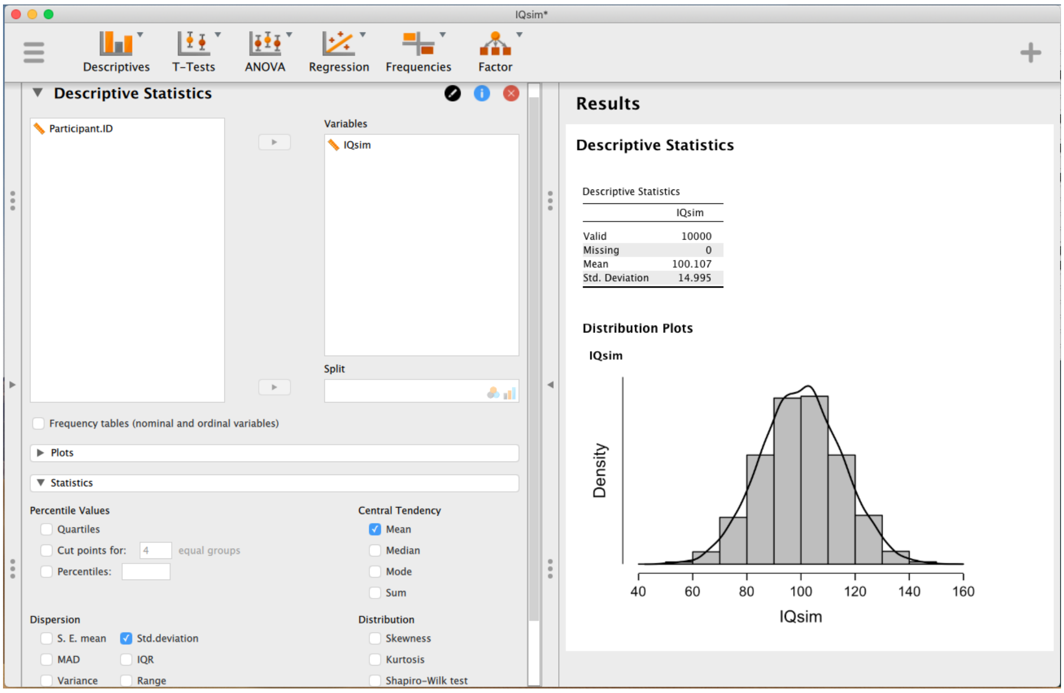


Figure1.5　JASP を使って正規分布から無作為抽出した例

. . . .

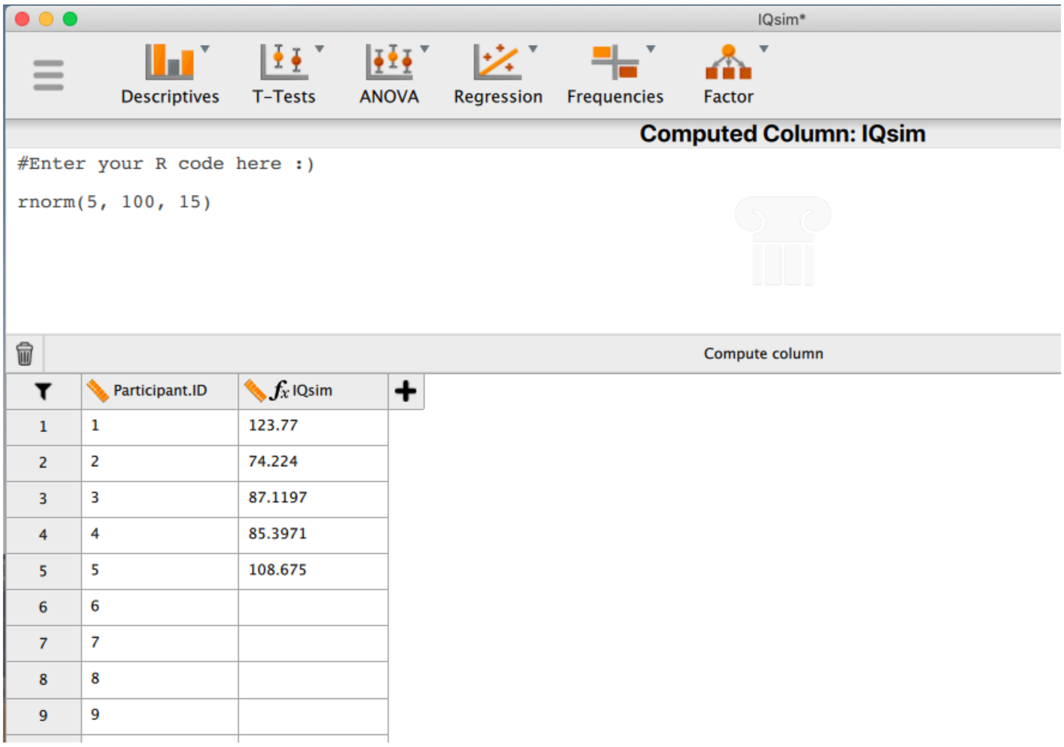


Figure1.6　JASP を使って，$\mu = 100$ で $\sigma = 15$ の正規分布から 5 つのランダムなサンプルを取り出す。

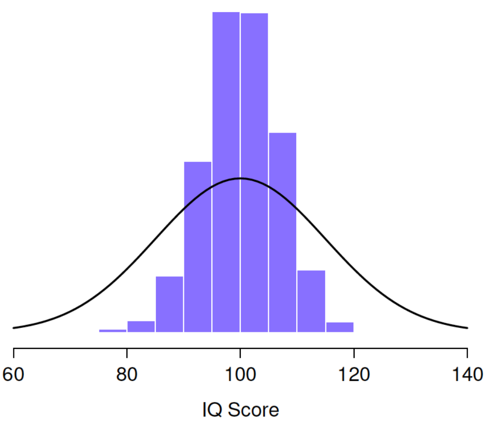


Figure1.7　“5 つの IQ スコア実験” の平均による標本分布。あなたが 5 人を無作為に取り出し，その IQ の*平均*を計算すると，数字はだいたい 80 から 120 の間に入るでしょう。ごく稀に，IQ が 120 より大きいとか 80 より低い人もいるかもしれませんが。比較のために，IQ スコアの母集団分布を黒い線で表現しました。

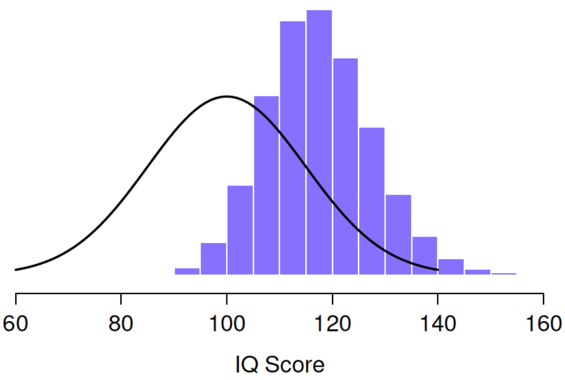


Figure1.8　“5 つの IQ 実験” における*最大値*の標本分布です。5 人を無作為に選び出した標本で，IQ スコアが最も高い人を選び出すと IQ が 100 から 140 の間にいるひとがほとんどでしょう。

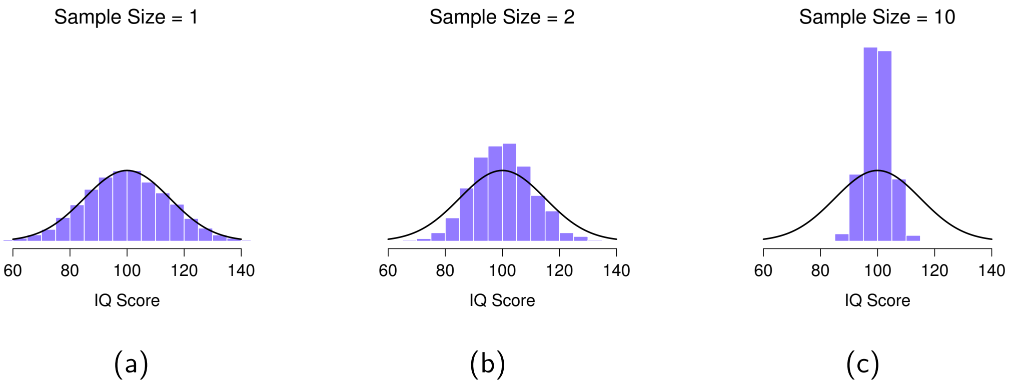


Figure1.9　平均の標本分布がいかにサンプルサイズに依存しているかを描いたもの。各パネルには，IQ データから 10,000 サンプル作り出した例と，各データセットにおける平均 IQ を計算したもの。これらのプロットのヒストグラムが示すのは，この平均の分布です (すなわち，平均の標本分布です)。一人ひとりの IQ スコアは平均が 100, 標準偏差が 15 の正規分布から取り出したもので，それは黒い点線で描画しています。パネル a では，データセットには一人分のデータしかなく，各標本の平均はその人の IQ スコアです。結果として，平均の標本分布は亜 IQ スコアの母集団分布と同じになります。しかしこのサンプルサイズを 2 にあげると，どのサンプル平均も一人の IQ スコアよりも母平均に近づくので，ヒストグラム (標本分布です) は母集団分布よりも少し狭くなります。今度はサンプルサイズを 10 にしてみましょう (パネル c)。すると標本平均は母平均周りでかなり狭く分布する傾向がみてとれます。

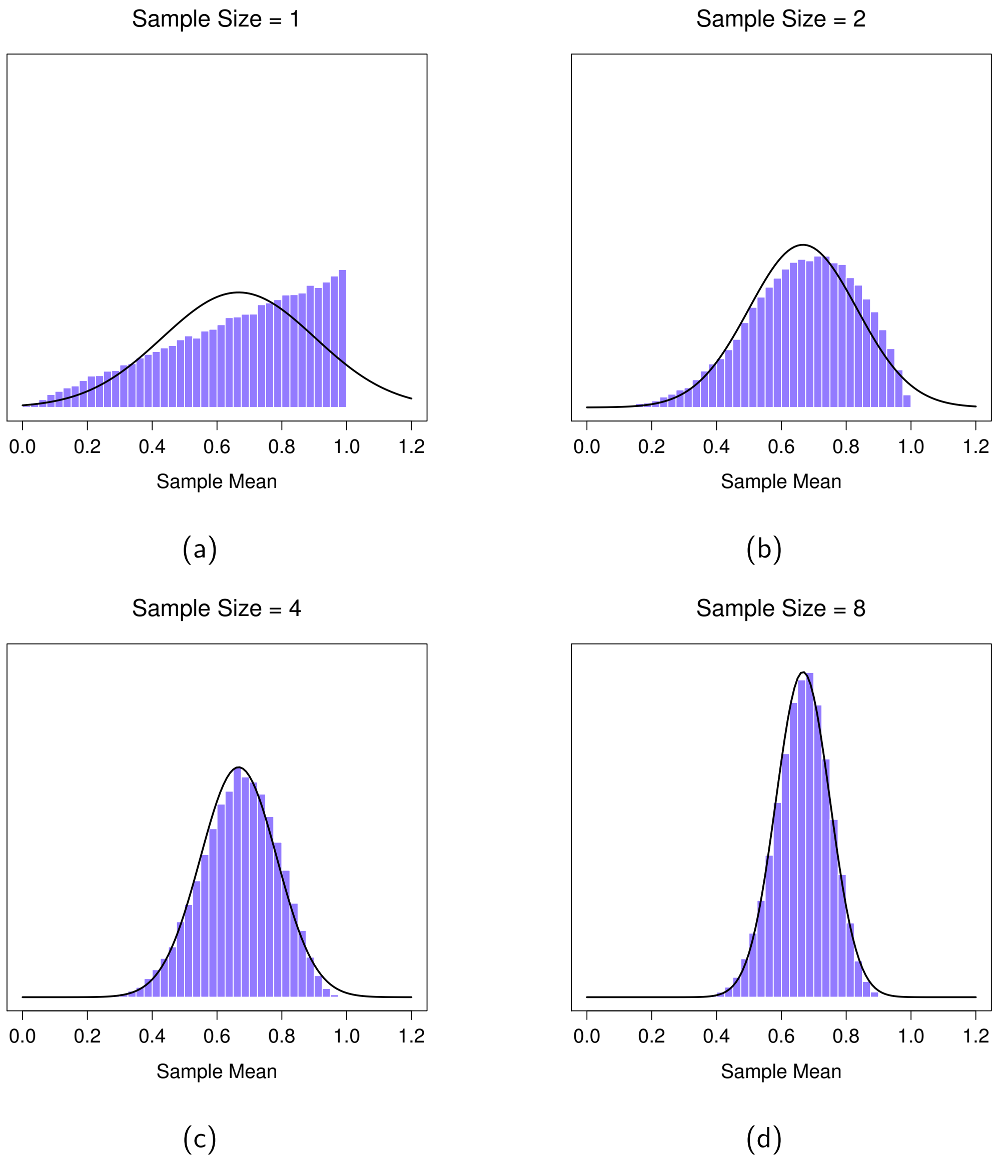


Figure1.10　中心極限定理のデモンストレーション。パネル a では，正規分布でない母集団分布を示しています。パネル b から d はパネル a に示した分布から，それぞれ標本サイズ 2,4,8 のサンプルを取った平均の標本分布を示しています。ご覧のとおり，元の母集団分布が正規分布でなくても，平均の標本分布は正規分布に近づきます。サンプルの度数が 4 のときでも。

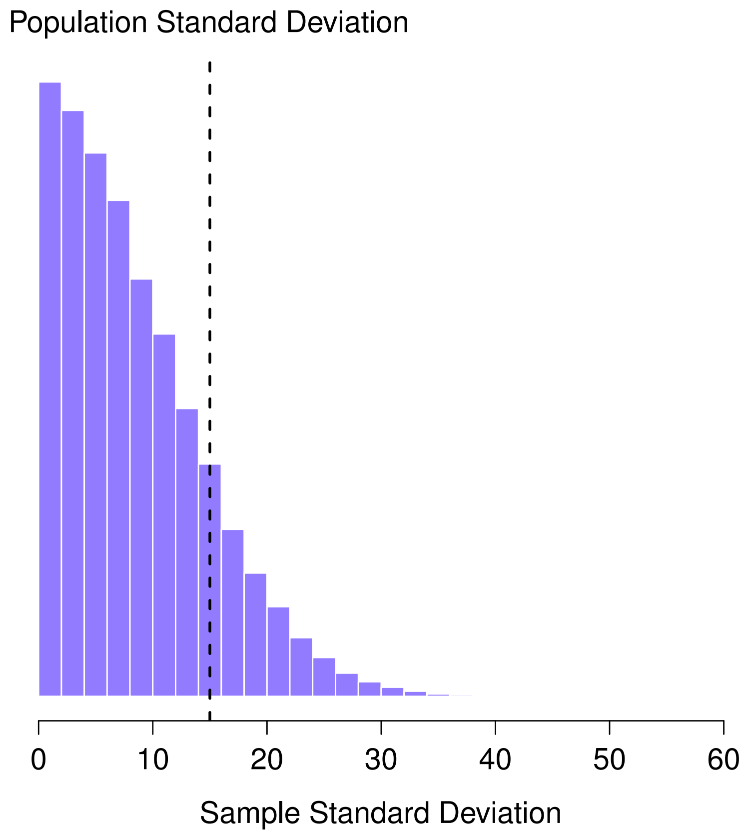
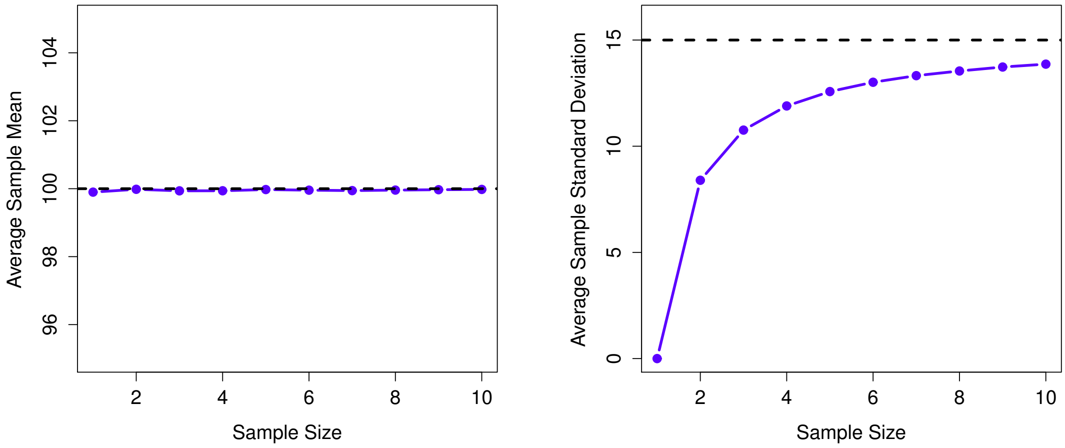


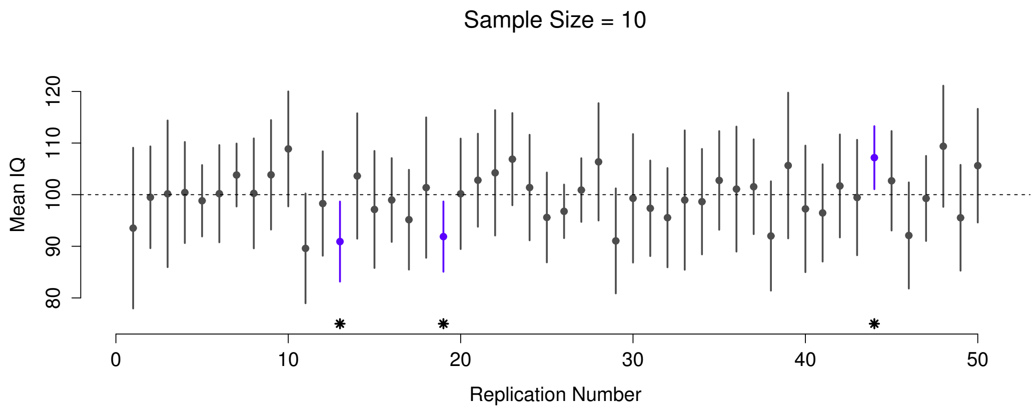
Figure1.11　“2 つの IQ スコア” についての標本標準偏差の標本分布。真の母標準偏差は 15(点線) ですが，ヒストグラムから見て取れるように，実験のほとんどがそれより小さい標準偏差を示しています。平均的には，この実験では標本標準偏差は 8.5 にしかなりません。これは真の値を下回っています! 言い換えれば,，標本標準偏差は母標準偏差の推定値に*基*づいているのです。



(a) (b)

Figure1.12　標本平均は母平均の不偏推定量であるのに対し (パネル a)，標本標準偏差は母標準偏差から偏りのある推定量になっている (パネル b) ことを描いた図。この図を書くために，観測度数 1 のデータセットを 10,000 点シミュレーションで作り，観測度数 2 についても 10,000 点，同様にサンプルサイズ 10 まで順に増やしていったものです。それぞれのデータセットは偽の IQ データからできており，真の母平均が 100，標準偏差が 15 であるようにしています。*平均については*，サンプルサイズに関わらず標本平均は 100 になります (パネル a)。しかし，標本標準偏差は，サンプルサイズが小さい時は顕著ですが，一貫して小さめになります (パネル b)。

(a)



(b)

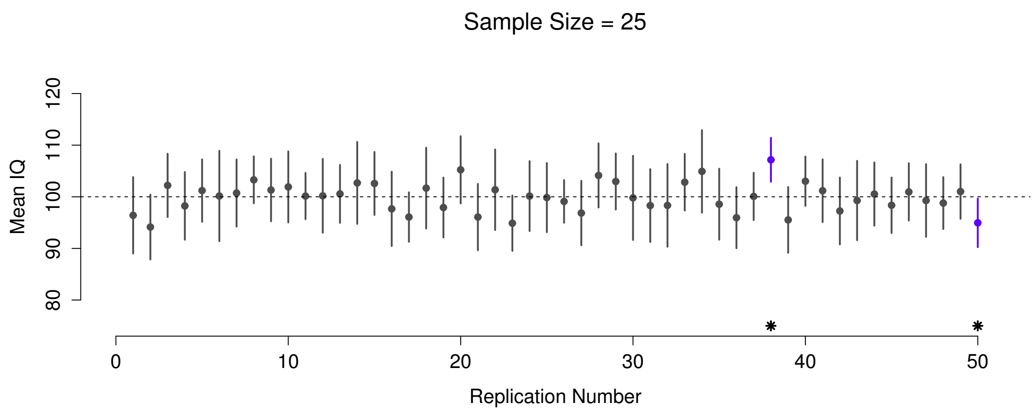


Figure1.13　95% 信頼区間。上の図 (パネル a) では，10 人の IQ を測定した実験を 50 回反復した例を示しています。点線は標本平均で，線は 95% 信頼区間です。50 回中 47 回の信頼区間が真の平均 (すなわち 100) を含んでいましたが，3 回はそうでなかったことをアスタリスクマークで示しています。下の図 (パネル b) では同じようなシミュレーションで，ここでは 25 人の IQ を測定した実験を反復したシミュレーションを示しています。