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 1.1 からセクション 1.3 を通じてサンプリング理論にフォーカスし，セクション 1.4 と　1.5 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?

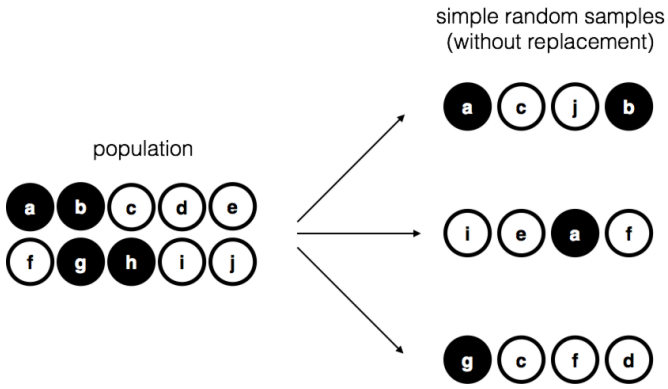


Figure1.1　母集団からの非復元単純ランダムサンプリング

- 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 1.1 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

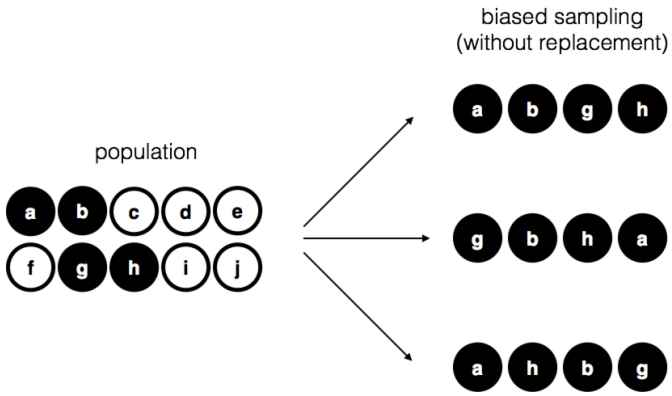


Figure1.2　有限母集団から復元なしの偏ったサンプリングをする

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 1.1 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 1.2 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。

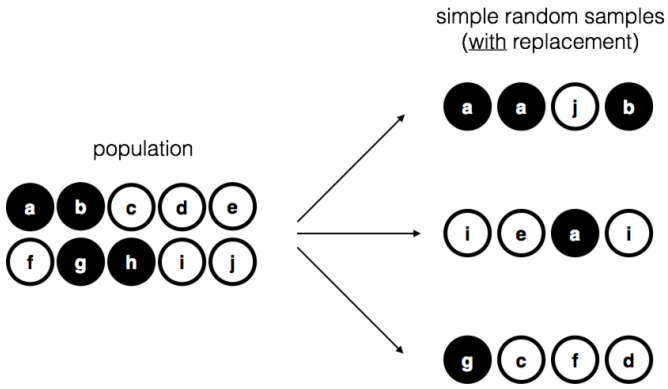


Figure1.3　有限母集団から復元 *あり*の単純無作為抽出

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 1.3 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 1.1 と 1.2 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

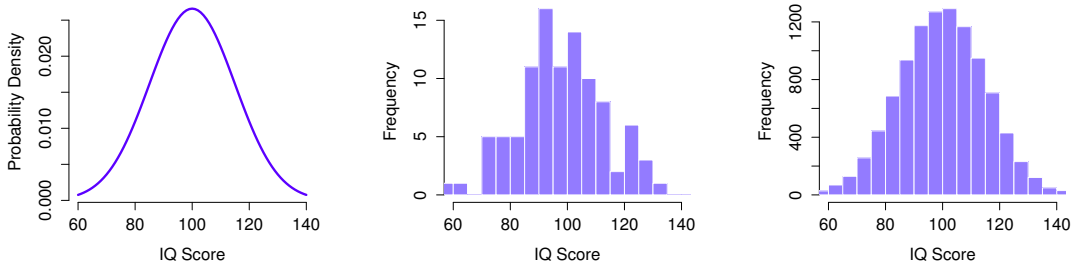
オーケイ。無作為標本の方法論的んな問題は少し横に置いて，違う問題に目を向けてみましょう。ここまで私たちは科学者のいうところの母集団について考えてきました。心理学者にとって，母集団とはひとの集団ということになるでしょう。環境学者にしてみれば，母集団がクマの集団になるかもしれません。ほとんどの場合において，科学者が考える母集団というのは現実世界に存在する具体的な何かです。しかし，統計学者は少し変わったひとたちなのです。一方では，彼らは科学者と同じように現実的な科学と現実世界のデータに興味が *あるのです*。他方で，彼らは数学者が考えるような純粋に抽象的な領域を操作しようとも思っています。その結果，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 1.4a 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 1.4b のようになるでしょう。ご覧になれば分かる



(a) (b) (c)

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 1.4a 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 1.4) 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 1.5) 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 1.5 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 1.6 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 1.1 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 1.1 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 1.7 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 1.7 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 Sampling distributions exist for any sample statistic!

One thing to keep in mind when thinking about sampling distributions is that *any* sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time I replicated the “five IQ scores” experiment I wrote down the largest IQ score in the experiment. This would give me a data set that started out like this:

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

Doing this over and over again would give me a very different sampling distribution, namely the *sampling distribution of the maximum*. The sampling distribution of the maximum of 5 IQ scores is shown in Figure 1.8. Not surprisingly, if you pick 5 people at random and then find the person with the highest IQ score, they're going to have an above average IQ. Most of the time you'll end up with someone whose IQ is measured in the 100 to 140 range.

1.3.3 The central limit theorem

At this point I hope you have a pretty good sense of what sampling distributions are, and in particular what the sampling distribution of the mean is. In this section I want to talk about how the sampling distribution of the mean changes as a function of sample size. Intuitively, you already know part of the answer. If you only have a few observations, the sample mean is likely to be quite inaccurate. If you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow. You can see this visually in Figure 1.9, showing that the bigger the sample size, the narrower the sampling distribution gets. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the **standard error**. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample *mean*, we often use the acronym SEM. As you can see just by looking

at the picture, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something I've been glossing over so far. All my examples up to this point have been based on the "IQ scores" experiments, and because IQ scores are roughly normally distributed I've assumed that the population distribution is normal. What if it isn't normal? What happens to the sampling distribution of the mean? The remarkable thing is this, no matter what shape your population distribution is, as N increases the sampling distribution of the mean starts to look more like a normal distribution. To give you a sense of this I ran some simulations. To do this, I started with the "ramped" distribution shown in the histogram in Figure 1.10. As you can see by comparing the triangular shaped histogram to the bell curve plotted by the black line, the population distribution doesn't look very much like a normal distribution at all. Next, I simulated the results of a large number of experiments. In each experiment I took $N = 2$ samples from this distribution, and then calculated the sample mean. Figure 1.10b plots the histogram of these sample means (i.e., the sampling distribution of the mean for $N = 2$). This time, the histogram produces a $\cap$-shaped distribution. It's still not normal, but it's a lot closer to the black line than the population distribution in Figure 1.10a. When I increase the sample size to $N = 4$, the sampling distribution of the mean is very close to normal (Figure 1.10c), and by the time we reach a sample size of $N = 8$ it's almost perfectly normal. In other words, as long as your sample size isn't tiny, the sampling distribution of the mean will be approximately normal no matter what your population distribution looks like!

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean.

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the **central limit theorem**. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ and the standard error of the mean is

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

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling

distribution becomes normal.[*4]

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us *how much* more reliable a large experiment is. It tells us why the normal distribution is, well, *normal*. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, "general" intelligence as measured by IQ is an average of a large number of "specific" skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

1.4 Estimating population parameters

In all the IQ examples in the previous sections we actually knew the population parameters ahead of time. As every undergraduate gets taught in their very first lecture on the measurement of intelligence, IQ scores are *defined* to have mean 100 and standard deviation 15. However, this is a bit of a lie. How do we know that IQ scores have a true population mean of 100? Well, we know this because the people who designed the tests have administered them to very large samples, and have then "rigged" the scoring rules so that their sample has mean 100. That's not a bad thing of course, it's an important part of designing a psychological measurement. However, it's important to keep in mind that this theoretical mean of 100 only attaches to the population that the test designers used to design the tests. Good test designers will actually go to some lengths to provide "test norms" that can apply to lots of different populations (e.g., different age groups, nationalities etc).

This is very handy, but of course almost every research project of interest involves looking at a different population of people to those used in the test norms. For instance, suppose you wanted to measure the effect of low level lead poisoning on cognitive functioning in Port Pirie, a South Australian industrial town with a lead smelter. Perhaps you decide that you want to compare IQ scores among people in Port Pirie to a comparable sample in Whyalla, a South Australian industrial

[*4] As usual, I'm being a bit sloppy here. The central limit theorem is a bit more general than this section implies. Like most introductory stats texts I've discussed one situation where the central limit theorem holds: when you're taking an average across lots of independent events drawn from the same distribution. However, the central limit theorem is much broader than this. There's a whole class of things called "U-statistics" for instance, all of which satisfy the central limit theorem and therefore become normally distributed for large sample sizes. The mean is one such statistic, but it's not the only one.

town with a steel refinery.[*5] Regardless of which town you're thinking about, it doesn't make a lot of sense simply to *assume* that the true population mean IQ is 100. No-one has, to my knowledge, produced sensible norming data that can automatically be applied to South Australian industrial towns. We're going to have to **estimate** the population parameters from a sample of data. So how do we do this?

1.4.1 Estimating the population mean

Suppose we go to Port Pirie and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of Port Pirie? Obviously, we don't know the answer to that question. It could be 97.2, but it could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless, if I was forced at gunpoint to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess.

In this example estimating the unknown poulation parameter is straightforward. I calculate the sample mean and I use that as my **estimate of the population mean**. It's pretty simple, and in the next section I'll explain the statistical justification for this intuitive answer. However, for the moment what I want to do is make sure you recognise that the sample statistic and the estimate of the population parameter are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often different notation to refer to them. For instance, if the true population mean is denoted μ, then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted $\bar{X}$ or sometimes m. However, in simple random samples the estimate of the population mean is identical to the sample mean. If I observe a sample mean of $\bar{X} = 98.5$ then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

[*5] Please note that if you were *actually* interested in this question you would need to be a *lot* more careful than I'm being here. You *can't* just compare IQ scores in Whyalla to Port Pirie and assume that any differences are due to lead poisoning. Even if it were true that the only differences between the two towns corresponded to the different refineries (and it isn't, not by a long shot), you need to account for the fact that people already *believe* that lead pollution causes cognitive deficits. If you recall back to Chapter **??**, this means that there are different demand effects for the Port Pirie sample than for the Whyalla sample. In other words, you might end up with an illusory group difference in your data, caused by the fact that people *think* that there is a real difference. I find it pretty implausible to think that the locals wouldn't be well aware of what you were trying to do if a bunch of researchers turned up in Port Pirie with lab coats and IQ tests, and even less plausible to think that a lot of people would be pretty resentful of you for doing it. Those people won't be as co-operative in the tests. Other people in Port Pirie might be *more* motivated to do well because they don't want their home town to look bad. The motivational effects that would apply in Whyalla are likely to be weaker, because people don't have any concept of "iron ore poisoning" in the same way that they have a concept for "lead poisoning". Psychology is *hard*.

Symbol	What is it?	Do we know what it is?
$\bar{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes, identical to the sample mean in simple random samples

1.4.2 Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to identical to the corresponding sample statistic (i.e. $\bar{X}$). However, that's not always true. To see this, let's have a think about how to construct an **estimate of the population standard deviation**, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intuitions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the *cromulence* of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample:

20

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20 and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the *sample* this seems quite right, the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the *population* standard deviation it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data. The only reason that we don't see any variability in the *sample* is that the sample is too small to display any variation! So, if you have a sample size of $N = 1$ it *feels* like the right answer is just to say "no idea at all".

Notice that you *don't* have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess because you have only the one observation to work with, but it's still the best guess you

can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this:

```
20, 22
```

This time around, our sample is *just* large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean. If forced to guess we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is *wrong*, after all with only two observations we expect it to be wrong to some degree. The worry is that the error is *systematic*. Specifically, we suspect that the sample standard deviation is likely to be smaller than the population standard deviation.

This intuition feels right, but it would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. First I'll conduct an experiment in which I measure $N = 2$ IQ scores and I'll calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the *sampling distribution of the standard deviation*. I've plotted this distribution in Figure 1.11. Even though the true population standard deviation is 15 the average of the *sample* standard deviations is only 8.5. Notice that this is a very different result to what we found in Figure 1.9b when we plotted the sampling distribution of the mean, where the population mean is 100 and the average of the sample means is also 100.

Now let's extend the simulation. Instead of restricting ourselves to the situation where $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the results shown in Figure 1.12. On the left hand side (panel a) I've plotted the average sample mean and on the right hand side (panel b) I've plotted the average standard deviation. The two plots are quite different: *on average*, the average sample mean is equal to the population mean. It is an **unbiased estimator**, which is

essentially the reason why your best estimate for the population mean is the sample mean.[*6] The plot on the right is quite different: on average, the sample standard deviation s is *smaller* than the population standard deviation σ. It is a **biased estimator**. In other words, if we want to make a "best guess" $\hat{\sigma}$ about the value of the population standard deviation σ we should make sure our guess is a little bit larger than the sample standard deviation s.

[*6] I should note that I'm hiding something here. Unbiasedness is a desirable characteristic for an estimator, but there are other things that matter besides bias. However, it's beyond the scope of this book to discuss this in any detail. I just want to draw your attention to the fact that there's some hidden complexity here.

The fix to this systematic bias turns out to be very simple. Here's how it works. Before tackling the standard deviation let's look at the variance. If you recall from Section **??**, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

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

The sample variance s^2 is a biased estimator of the population variance σ^2. But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N-1$ rather than by N. If we do that, we obtain the following formula:

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

This is an unbiased estimator of the population variance σ. Moreover, this finally answers the question we raised in Section **??**. Why did JASP give us slightly different answers for variance? It's because JASP calculates $\hat{\sigma}^2$ not s^2, that's why. A similar story applies for the standard deviation. If we divide by $N-1$ rather than N our estimate of the population standard deviation becomes:

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

and when we use JASP's built in standard deviation function, what it's doing is calculating $\hat{\sigma}$, not s.[a]

[a] Okay, I'm hiding something else here. In a bizarre and counter-intuitive twist, since $\hat{\sigma}^2$ is an unbiased estimator of σ^2, you'd assume that taking the square root would be fine and $\hat{\sigma}$ would be an unbiased estimator of σ. Right? Weirdly, it's not. There's actually a subtle, tiny bias in $\hat{\sigma}$. This is just bizarre: $\hat{\sigma}^2$ is an unbiased estimate of the population variance σ^2, but when you take the square root, it turns out that $\hat{\sigma}$ is a biased estimator of the population standard deviation σ. Weird, weird, weird, right? So, why is $\hat{\sigma}$ biased? The technical answer is "because non-linear transformations (e.g., the square root) don't commute with expectation", but that just sounds like gibberish to everyone who hasn't taken a course in mathematical statistics. Fortunately, it doesn't matter for practical purposes. The bias is small, and in real life everyone uses $\hat{\sigma}$ and it works just fine. Sometimes mathematics is just annoying.

One final point. In practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N-1$) as the *sample* standard deviation. Technically, this is incorrect. The *sample* standard deviation should be equal to s (i.e., the formula where we divide by N). These aren't the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s. This is the right number to report, of course. It's just that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated

population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, I think it's important to keep the two *concepts* separate. It's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that.

To finish this section off, here's another couple of tables to help keep things clear.

Symbol	What is it?	Do we know what it is?
s	Sample standard deviation	Yes, calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes, but not the same as the sample standard deviation

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

1.5 Estimating a confidence interval

> *Statistics means never having to say you're certain*
>
> – Unknown origin[*7]

Up to this point in this chapter, I've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with a some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to *quantify* the amount of uncertainty that attaches to our estimate. It's not enough to be able guess that, say, the mean IQ of undergraduate

[*7] This quote appears on a great many t-shirts and websites, and even gets a mention in a few academic papers (e.g., http://www.amstat.org/publications/jse/v10n3/friedman.html, but I've never found the original source.

psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem (Section 1.3.3) that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution Section **??** that there is a 95% chance that a normally-distributed quantity will fall within about two standard deviations of the true mean.

To be more precise, the more correct answer is that there is a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean. Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean.

Mathematically, we write this as:

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

where the SEM is equal to $\sigma/\sqrt{N}$ and we can be 95% confident that this is true. However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean given that we know what the population parameters are. What we *want* is to have this work the other way around. We want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

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

What this is telling is is that the range of values has a 95% probability of containing the population mean μ. We refer to this range as a **95% confidence interval**, denoted CI_{95}. In short, as long as N is sufficiently large (large enough for us to believe that the sampling distribution of the mean is normal), then we can write this as our formula for the 95% confidence interval:

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

Of course, there's nothing special about the number 1.96. It just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I would have used 1.04 as the magic number rather than 1.96.

1.5.1 A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ. Yet, in Section 1.4 I stressed the fact that we don't actually *know* the true population parameters. Because we don't know the true value of σ we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the percentiles of the t-distribution rather than the normal distribution to calculate our magic number, and the answer depends on the sample size. When N is very large, we get pretty much the same value using the t-distribution or the normal distribution: 1.96. But when N is small we get a much bigger number when we use the t distribution: 2.26.

There's nothing too mysterious about what's happening here. Bigger values mean that the confidence interval is wider, indicating that we're more uncertain about what the true value of μ actually is. When we use the t distribution instead of the normal distribution we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like, and this uncertainty ends up getting reflected in a wider confidence interval.

1.5.2 Interpreting a confidence interval

The hardest thing about confidence intervals is understanding what they *mean*. Whenever people first encounter confidence intervals, the first instinct is almost always to say that "there is a 95% probability that the true mean lies inside the confidence interval". It's simple and it seems to capture the common sense idea of what it means to say that I am "95% confident". Unfortunately, it's not quite right. The intuitive definition relies very heavily on your own personal *beliefs* about the value of the population mean. I say that I am 95% confident because those are my beliefs. In everyday life that's perfectly okay, but if you remember back to Section **??**, you'll notice that talking about personal belief and confidence is a Bayesian idea. However, confidence intervals are *not* Bayesian tools. Like everything else in this chapter, confidence intervals are *frequentist* tools, and if you are going to use frequentist methods then it's not appropriate to attach a Bayesian interpretation to them. If you use frequentist methods, you must adopt frequentist interpretations!

Okay, so if that's not the right answer, what is? Remember what we said about frequentist probability. The only way we are allowed to make "probability statements" is to talk about a sequence of events, and to count up the frequencies of different kinds of events. From that perspective, the interpretation of a 95% confidence interval must have something to do with replication. Specifically, if we replicated the experiment over and over again and computed a 95% confidence interval for each replication, then 95% of those *intervals* would contain the true mean. More generally, 95% of all confidence intervals constructed using this procedure should contain the true population mean. This idea is illustrated in Figure 1.13, which shows 50 confidence intervals constructed for a "measure 10 IQ scores" experiment (top panel) and another 50 confidence intervals for a "measure 25 IQ scores" experiment (bottom panel). A bit fortuitously, across the 100 replications that I simulated, it turned out that exactly 95 of them contained the true mean.

The critical difference here is that the Bayesian claim makes a probability statement about the population mean (i.e., it refers to our uncertainty about the population mean), which is not allowed under the frequentist interpretation of probability because you can't "replicate" a population! In the frequentist claim, the population mean is fixed and no probabilistic claims can be made about

it. Confidence intervals, however, are repeatable so we can replicate experiments. Therefore a frequentist is allowed to talk about the probability that the *confidence interval* (a random variable) contains the true mean, but is not allowed to talk about the probability that the *true population mean* (not a repeatable event) falls within the confidence interval.

I know that this seems a little pedantic, but it does matter. It matters because the difference in interpretation leads to a difference in the mathematics. There is a Bayesian alternative to confidence intervals, known as *credible intervals*. In most situations credible intervals are quite similar to confidence intervals, but in other cases they are drastically different. As promised, though, I'll talk more about the Bayesian perspective in Chapter **??**.

1.5.3 Calculating confidence intervals in JASP

As of this edition, JASP does not (yet) include a simple way to calculate confidence intervals for the mean as part of the 'Descriptives' functionality. But the 'Descriptives' do have a check box for the S.E. Mean, so you can use this to calculate the lower 95% confidence interval as:

`Mean - (1.96 * S.E. Mean)` , and the upper 95% confidence interval as:

`Mean + (1.96 * S.E. Mean)`

95% confidence intervals are the de facto standard in psychology. So, for example, if I load the `IQsim.jasp` file, check mean and S.E mean under 'Descriptives', I can work out the confidence interval associated with the simulated mean IQ:

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

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

So, in our simulated large sample data with N=10,000, the mean IQ score is 100.107 with a 95% CI from 99.813 to 100.401. Hopefully that's clear and fairly easy to interpret. So, although there currently is not a straightforward way to get JASP to calculate the confidence interval as part of the variable 'Descriptives' options, if we wanted to we could pretty easily work it out by hand.

Similarly, when it comes to plotting confidence intervals in JASP, this is also not (yet) available as part of the 'Descriptives' options. However, when we get onto learning about specific statistical tests, for example in Chapter **??**, we will see that we can plot confidence intervals as part of the data analysis. That's pretty cool, so we'll show you how to do that later on.

1.6

Summary

In this chapter I've covered two main topics. The first half of the chapter talks about sampling theory, and the second half talks about how we can use sampling theory to construct estimates of the population parameters. The section breakdown looks like this:

- Basic ideas about samples, sampling and populations (Section 1.1)
- Statistical theory of sampling: the law of large numbers (Section 1.2), sampling distributions and the central limit theorem (Section 1.3).
- Estimating means and standard deviations (Section 1.4)
- Estimating a confidence interval (Section 1.5)

As always, there's a lot of topics related to sampling and estimation that aren't covered in this chapter, but for an introductory psychology class this is fairly comprehensive I think. For most applied researchers you won't need much more theory than this. One big question that I haven't touched on in this chapter is what you do when you don't have a simple random sample. There is a lot of statistical theory you can draw on to handle this situation, but it's well beyond the scope of this book.

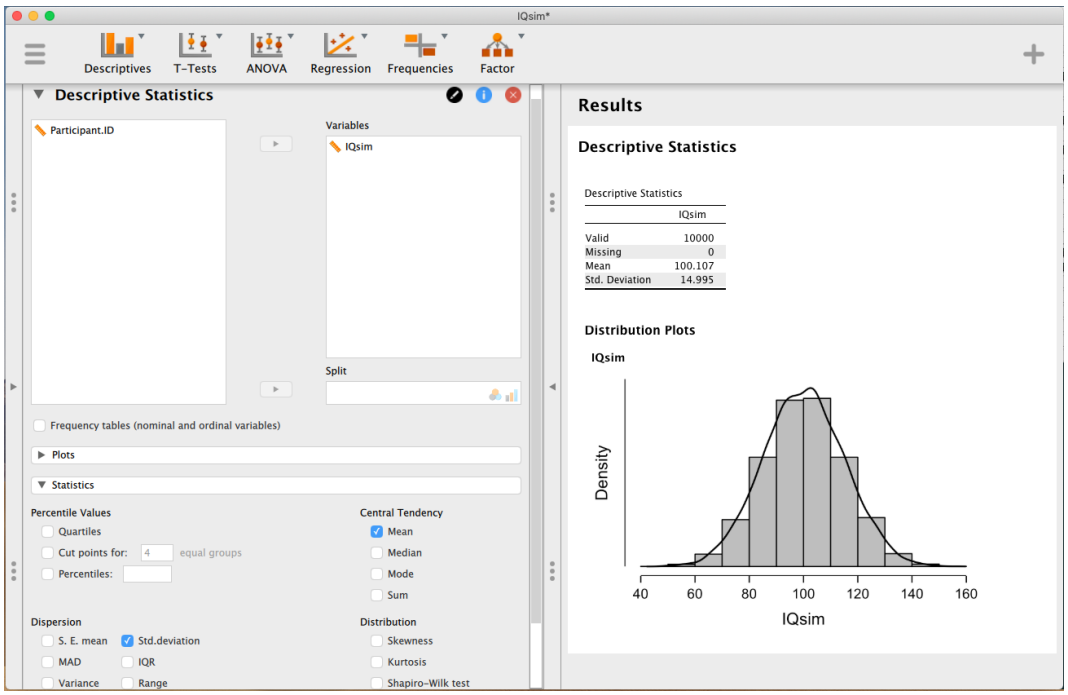


Figure1.5　JASP を使って正規分布から無作為抽出した例

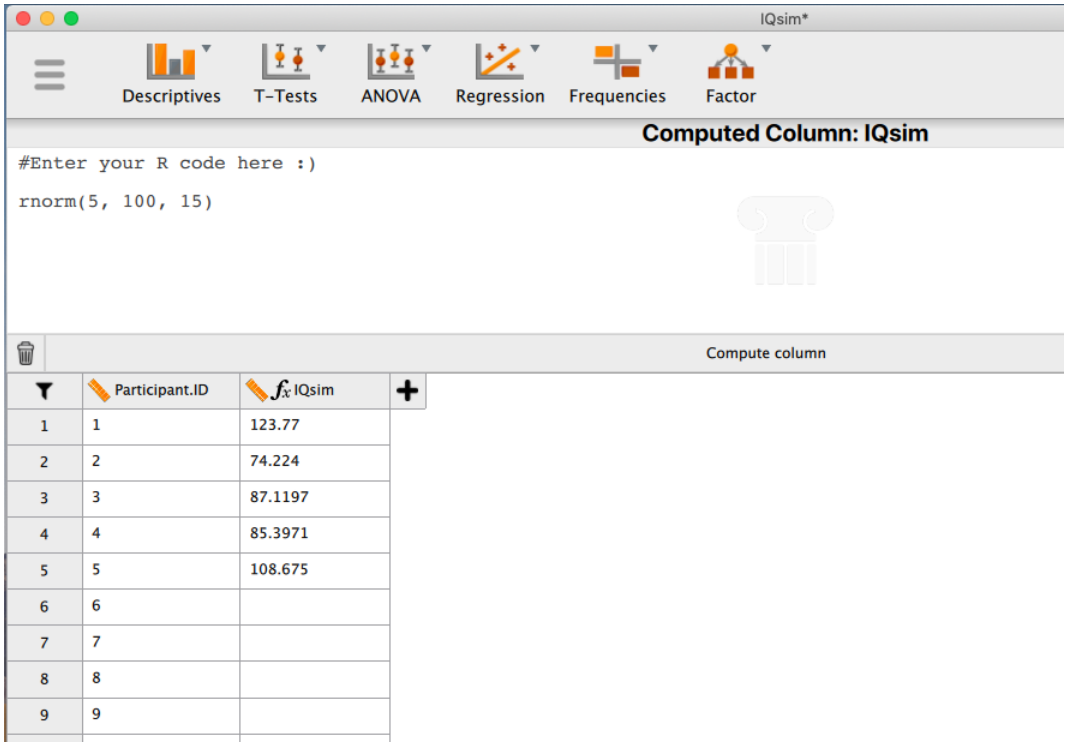


Figure1.6　JASP を使って，$\mu = 100$ で $\sigma = 15$ の正規分布から 5 つのランダムなサンプルを取り出す。

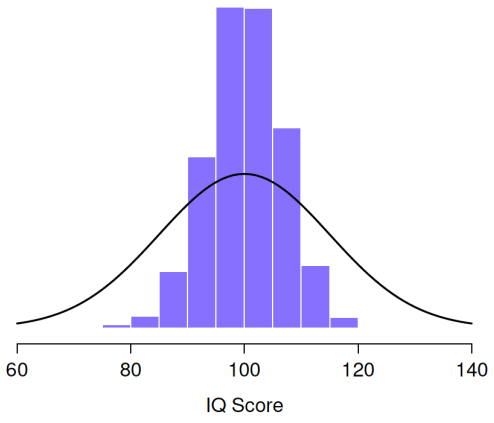


Figure1.7　The sampling distribution of the mean for the "five IQ scores experiment". If you sample 5 people at random and calculate their *average* IQ you'll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

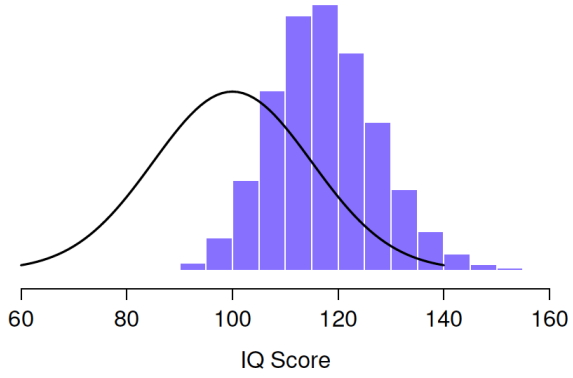


Figure1.8 The sampling distribution of the *maximum* for the "five IQ scores experiment". If you sample 5 people at random and select the one with the highest IQ score you'll probably see someone with an IQ between 100 and 140.

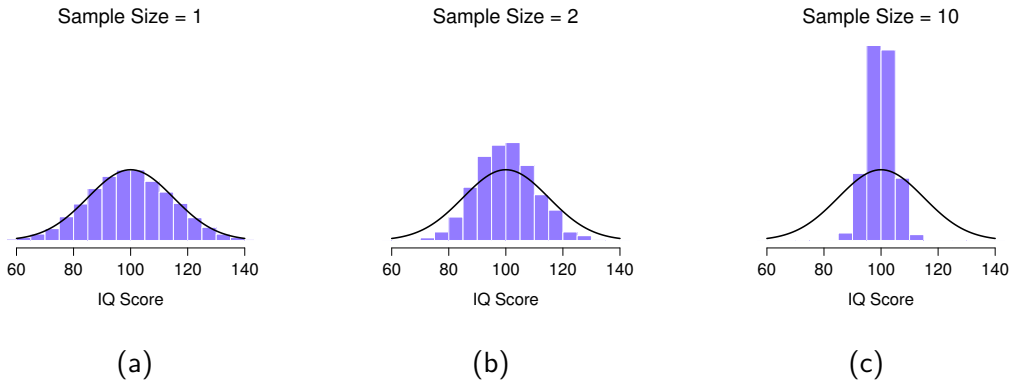


Figure1.9 An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean). Each individual IQ score was drawn from a normal distribution with mean 100 and standard deviation 15, which is shown as the solid black line. In panel a, each data set contained only a single observation, so the mean of each sample is just one person's IQ score. As a consequence, the sampling distribution of the mean is of course identical to the population distribution of IQ scores. However, when we raise the sample size to 2 the mean of any one sample tends to be closer to the population mean than a one person's IQ score, and so the histogram (i.e., the sampling distribution) is a bit narrower than the population distribution. By the time we raise the sample size to 10 (panel c), we can see that the distribution of sample means tend to be fairly tightly clustered around the true population mean.

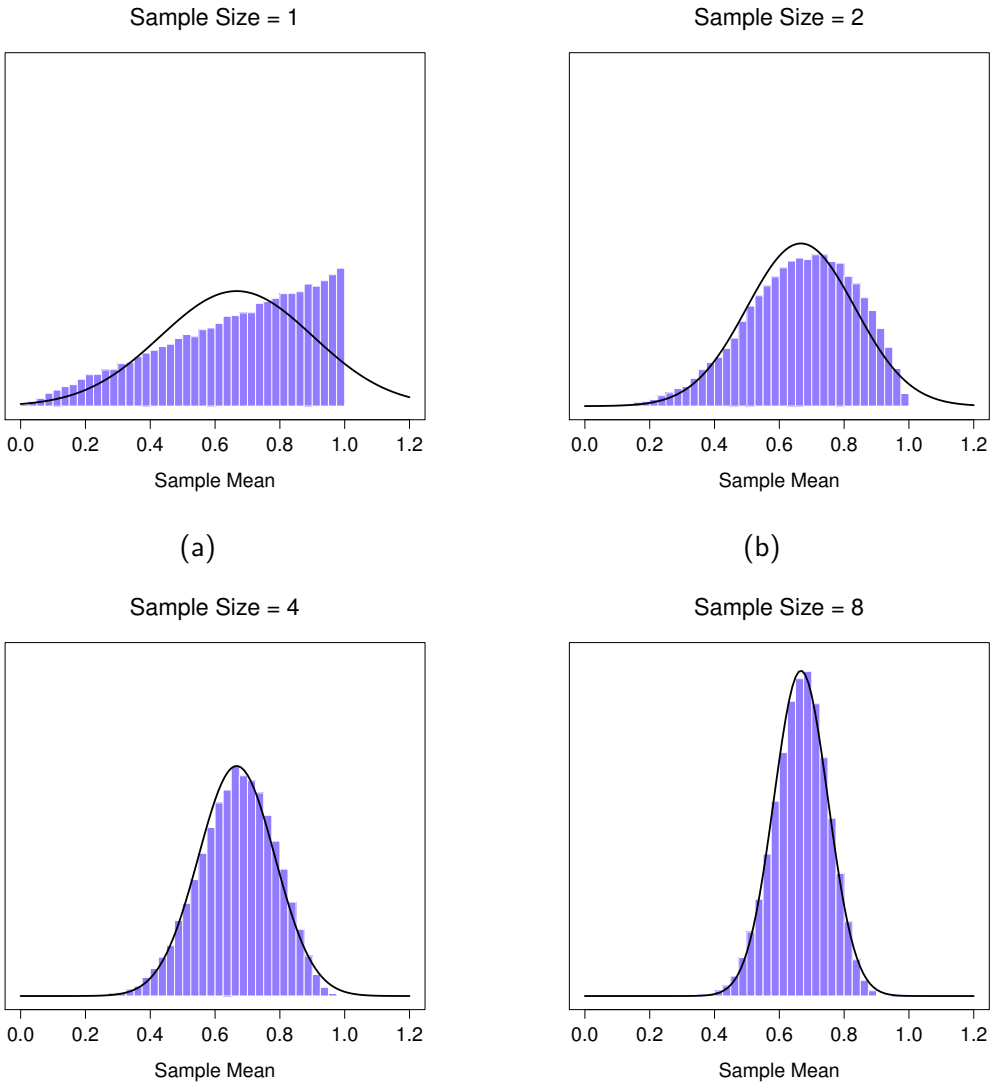


Figure1.10 A demonstration of the central limit theorem. In panel a, we have a non-normal population distribution, and panels b-d show the sampling distribution of the mean for samples of size 2,4 and 8 for data drawn from the distribution in panel a. As you can see, even though the original population distribution is non-normal the sampling distribution of the mean becomes pretty close to normal by the time you have a sample of even 4 observations.

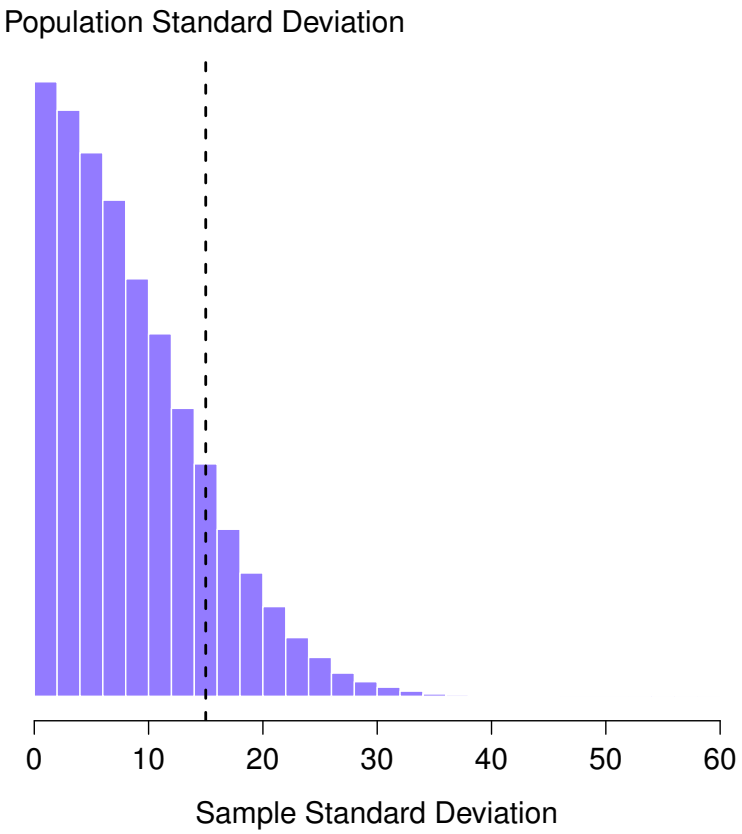
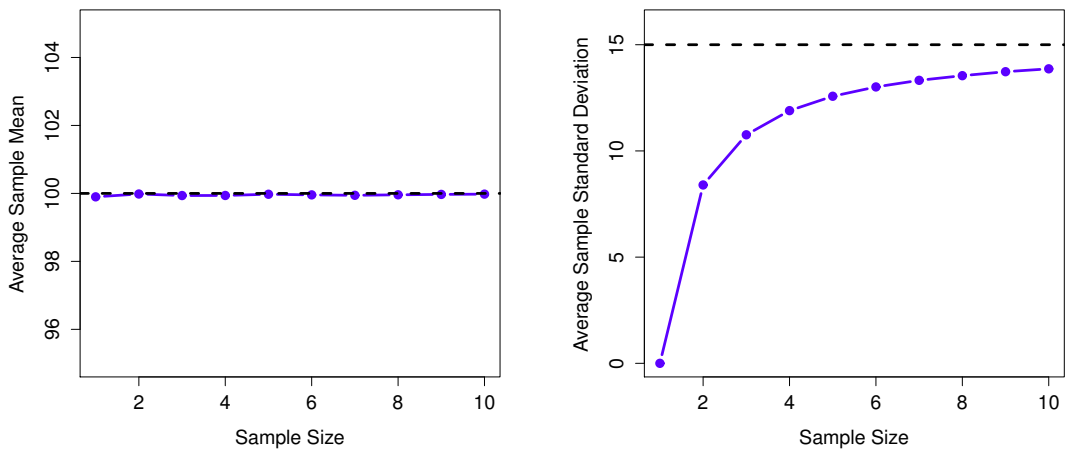


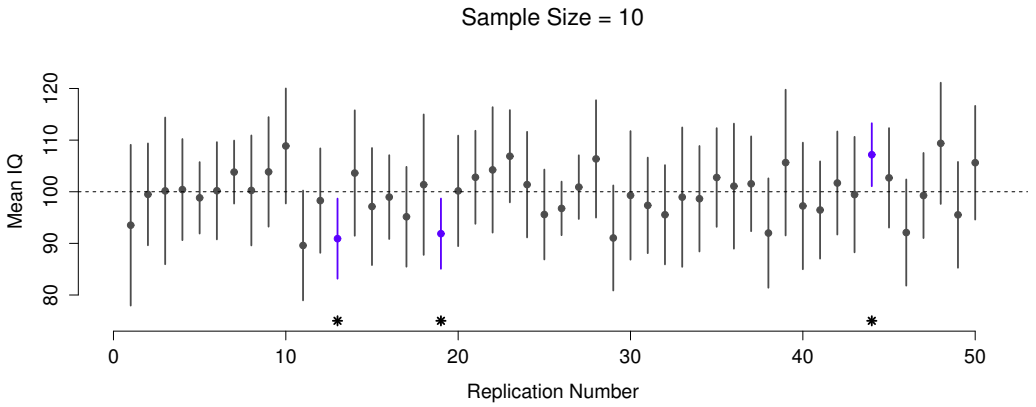
Figure1.11 The sampling distribution of the sample standard deviation for a "two IQ scores" experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a *biased* estimate of the population standard deviation.



(a) (b)

Figure1.12 An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each, 10,000 more with 2 observations, and so on up to a sample size of 10. Each data set consisted of fake IQ data, that is the data were normally distributed with a true population mean of 100 and standard deviation 15. *On average*, the sample means turn out to be 100, regardless of sample size (panel a). However, the sample standard deviations turn out to be systematically too small (panel b), especially for small sample sizes.

(a)



(b)

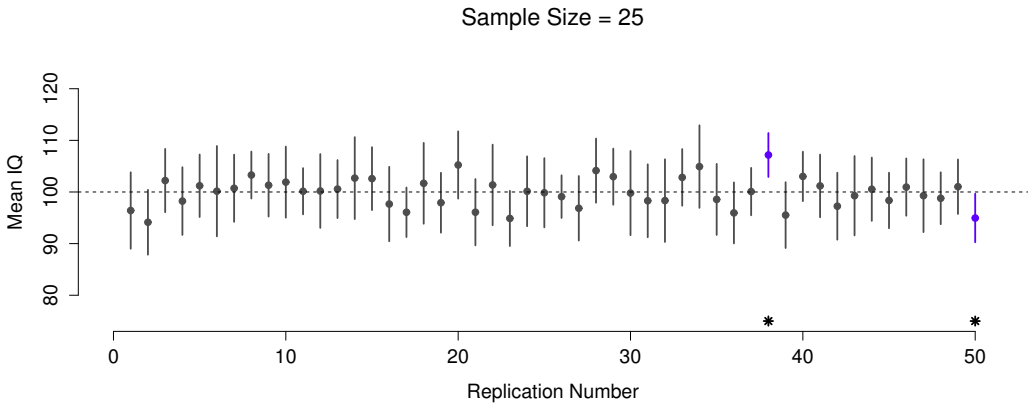


Figure1.13 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true mean (i.e., 100), but the three intervals marked with asterisks do not. The lower graph (panel b) shows a similar simulation, but this time we simulate replications of an experiment that measures the IQs of 25 people.