

Learning Statistics with JASP

A Tutorial for Psychology Students
and Other Beginners

Danielle J. Navarro
David R. Foxcroft
Thomas J. Faulkenberry



Learning Statistics with JASP: A Tutorial for Psychology Students and Other Beginners

(Version $\frac{1}{\sqrt{2}}$)

Danielle Navarro
University of New South Wales
d.navarro@unsw.edu.au

David Foxcroft
Oxford Brookes University
david.foxcroft@brookes.ac.uk

Thomas J. Faulkenberry
Tarleton State University
faulkenberry@tarleton.edu

<http://www.learnstatswithjasp.com>

Overview

Learning Statistics with JASP covers the contents of an introductory statistics class, as typically taught to undergraduate psychology students. The book discusses how to get started in JASP as well as giving an introduction to data manipulation. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, correlation, *t*-tests, regression, ANOVA and factor analysis. Bayesian statistics is covered at the end of the book.

Citation

Navarro, D.J., Foxcroft, D.R., & Faulkenberry, T.J. (2019). *Learning Statistics with JASP: A Tutorial for Psychology Students and Other Beginners*. (Version $\frac{1}{\sqrt{2}}$).

日本語版について

この本は **JASP ユーザの会** の有志が翻訳を担当しています。日本語の翻訳メンバーは下記の通りです (50 音順)。

- 紀ノ定保礼 (静岡理工科大学)
- 国里愛彦 (専修大学)
- 小杉考司 (専修大学；代表者。連絡先は kosugi@psy.senshu-u.ac.jp です)
- 小林穂波 (関西学院大学)
- 五島光
- 竹林由武 (福岡県立医科歯科大学)
- 徳岡大 (高松大学)
- 難波修史 (国立研究開発法人理化学研究所)
- 北條大樹 (東京大学)
- 平川真 (広島大学)
- 武藤拓之 (京都大学 こころの未来研究センター)
- 山根嵩史 (川崎医療福祉大学)

この本は完全にオープンソースです。つまり、あなたが望む方法で自由に改変することができます (ただし著者に適切なクレジットを与える限りにおいて、です。ライセンス条項を確認してください)。

最新バージョン

この本は、翻訳の進捗に合わせて隨時コンパイルされ、バージョンアップしていきます。最後にコンパイルされたのは 2021 年 11 月 22 日です。

This book is published under a Creative Commons BY-SA license (CC BY-SA) version 4.0. This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license - CC BY-SA.

<https://creativecommons.org/licenses/by-sa/4.0/>

The JASP-specific revisions to the original book by Navarro and Foxcroft were made possible by a generous grant to Tom Faulkenberry from the Tarleton State University Center for Instructional Innovation. Also, many thanks to Kristen Bowman for creating the beautiful front and back cover art for the book.

Table of Contents

第 $1/\sqrt{2}$ 版に向けた前書き

素晴らしい姉妹本 “jamovi で学ぶ統計” や “R で学ぶ統計” のアレンジ版, “JASP で学ぶ統計” をご紹介できて嬉しく思っています。このバージョンは Dani Navarro と David Foxcroft による素晴らしい前作の上に成り立っています。前作に投入された努力がなければ、この品質は達成できなかっただでしょう。このアレンジ版を出そうと思ったとき、私はシンプルな目標を持っていました。私は Navarro と Foxcroft のテキストを私自身の授業で使いたかったのですが、直ぐにはそうでないと思ったのは、jamovi ではなく JASP を使っていたからです。どちらも素晴らしいツールなのですが、私は JASP のほうがちょっとばかり好きなのです。というのも、jamovi がプロジェクトとして独立する前から JASP を使っていたからです。ですから、この本を世界の JASP ユーザに提供できることを嬉しく思っています。

2019 年夏、このオープン教育リソース (Open Educational Resource, OER) の執筆ための助成金を与えてくれたタールトン州立大学の Center for Instructional Innovation に感謝します。私の未来の学生 (そして世界中の学生諸君) に向けて、高品質な統計テキストを (おそらく未来永劫) 100/

(誤字脱字などを見つけるなどして) このテキストをよくしてくれる読者を待っています。もし何か貢献できると思ったら私にメールを送ってください (あるいは Githubg ページに参加してフォークしてください)。やっちゃおうぜ!

Thomas J. Faulkenberry

July 12, 2019

バージョン 0.70 に向けた前書き

バージョン 0.65 から今回へのアップデートでは、いくつかの新しい分析が導入されました。ANOVA の章では、反復測定 ANOVA や共分散分析 (ANCOVA) のセクションも追加しました。因子分析やそれに関連する技術の章も導入しました。この新しい要素のスタイルは、本書の他の章と一貫していますが、目の肥えた読者は少し概念的で実用的な説明が強調され、代数的要素が減っていることに気づくかもしれませんね。このことが良いことかどうかわかりませんが、代数については少し後で追加するかもしれません。しかしそれは、私が統計を理解して教えるときの両方のアプローチ、そして私がコース内で教えているとき学生から受け取ったフィードバックを反映したものです。これに合わせて、私も本の残りの部分に目を通し、代数の一部を箱や枠に入れて分離してみました。これらが重要でないとか、役に立たないというのではなく、学生の中にはこれらを読み飛ばしたいと思うかもしれないのに、独立したパートにすることでそうした読者の役に立てばと思うのです。

このバージョンについて、私の学生や同僚、特に Wakefield Morys-Carter から多くのコメントやフィードバックを受けたことに感謝しています。また世界中のみなさんから提案や修正をいただきま

したことに感謝しています！新しい試みの一つとして、この本のサンプルデータファイルが jamovi にアドオンモジュールとして読み込むことができるようになった、というのがあります。Jonathon Love がこれを援助してくれたことに感謝します。

David Foxcroft
February 1st, 2019

バージョン 0.65 に向けた前書き

本書は Danielle Navarro による ‘R で学ぶ統計’ の応用で、統計的なソフトウェアや分析例を jamovi に置き換えたものです。R はパワフルな統計プログラミング言語ですが、教師や学生が統計学習の最初に選択するものではありません。教師や学生によっては、ポイントしてクリックするだけで分析できるタイプのソフトウェアを好みますし、それこそ jamovi でできることです。jamovi は R を使う上で二つの側面だけに狙いを定めています。ポイントしてクリックする、グラフィカルユーザインターフェイス (GUI) と、多くの機能を組み合わせた関数を提供しており、SPSS や SAS のような方法を R でプログラミングする方法を提供しています。重要なことは、jamovi はいつもフリーでオープンであること、それが中心的な価値の一つであることです。jamovi は科学コミュニティによって作られ、科学コミュニティのためのものなのですから。

このバージョンでは、多くの人に下書きを読んでもらって、幾つもの提案や訂正をもらいました。特に Dr David Emery and Kirsty Walter に感謝しています。

David Foxcroft
July 1st, 2018

バージョン 0.6 にむけた前書き

この本は 2015 年にバージョン 0.5 をリリースしてからそれほど大きく変わってはいません。それでも、前より変わったと言った方がいいと思います。私は 2016 年にアデレードからシドニーに移動し、UNSW での経験はアデレードの頃に比べて変わってしまったので、こちらにきてから取り組む機会がずいぶん減ってしまったのです。実際に振り返ってみると、少し奇妙な感じもします。ちょっとコメントすると…

- 奇妙なことですが、この本では一貫して私の性別を間違えていますが、これについては私自身に責任があると思います(笑)。12ページにこのことについて言及した短い脚注があります。現実の生活では、私はジェンダー多様性を認める活動をしていて、この2年ほどはほとんどshe/herの代名詞を使っています。しかし私は、面倒くさがりなので、この本での文章を訂正しようとは思ってません。
- バージョン0.6にむけて、私はそれほど大きく変更せず、指摘してもらったタイプミスやその他の間違いぐらいの、いくつかのマイナーチェンジだけにしました。ただ、セクション14.4で触れている **lsr** パッケージ(これはもうメンテナンスされていません)に含まれる etaSquared関数に関する問題については注目してもらいたいと思います。この関数は、本書のようなシンプルな例ではうまく機能するのですが、見つけ切ってはないのですけど確かにバグがあるんです!ですから、これについては注意しておいてください。
- 最も大きな変更はライセンスで、私はこれをクリエイティブ・コモンズライセンス(特にCC BY-SA 4.0)のもとでリリースすることにし、誰でも利用できるように全てのソースファイルを GitHub レポジトリに置きました。

おそらく **tidyverse** を使ったバージョンを誰かが書いてくれると思うのですが… これな近々 R にとってもっと重要なトピックになってくるでしょう。:-) では。Danielle Navarro

バージョン0.5にむけた前書き

今年もまたアップデートです。今回のアップデートは、本書の理論セクション全体に関わるものです。第9,10,11章は書き直しました。よくなっているといいんですが。同時に、17章も全体的に新しくして、ベイズ統計にフォーカスしました。この変更によって本書は大きく改良されたと思います。私は常に、推測統計全体についての事実が従来型の観点から描かれていることに不満を感じていました。私もすでにベイズ流のデータ分析を自分の仕事に取り入れているのに、です。本書のどこかにベイズの手法を入れることで、本全体として良くなつたなと思えるようになりました。今回のアップデートでは他にもやりたいことがいくつかあったのですが、私はいつも授業の締め切りに追われているので、アップデートが後回しになってしまいます! Dan Navarro

February 16, 2015

バージョン0.4に向けた前書き

前回の前書きを書いてから一年経ってしまいました。今回はいくつか重要な変更点があります。第3,4章は RStudio の特徴について書くのを抑えたので、読みやすくなりましたが、第12,13章はカイ二乗検定と t 検定を実行するための lsr パッケージの新しい関数を使うようにしたので、補正に関する議論が lsr パッケージの新しい関数を参照するように対応させました。バージョン 0.4 の電子版では、内部参照（すなわち、セクションごとの実際のハイパーリンクです）が改良されています。これはバージョン 0.3.1 から導入されたものです。あちこちに新しいことを入れていますが、多くは誤字脱字の修正（タイプを見つけてくれたひと全てに感謝します！）で、バージョン 0.3 と 0.4 が全体的に全く違うというようなことはありません。この 12 ヶ月の間、もっと中身を充実させたいと思ってきました。反復測定 ANOVA や混合モデルについての議論がないのは、全く心苦しいところです。言い訳になりますが、進捗が出ないのは私の二人目の子供が 2013 年の初めに生まれたからで、私は昨年のほとんどを生活の維持に注力したからです。結果的に、この本のように報酬のないプロジェクトは、実際に私に賃金が支払われる仕事に追いやられることになりました。今は状況が幾分改善されましたので、バージョン 0.5 ではもっと前進できるでしょう。

私を驚かせたことの一つは、この本を入手するためのダウンロード数です。ウェブサイトの基本的なトラッキング情報を、数ヶ月前から入手できたのですが、（明らかなロボットツールを除いて）この本は一日平均 90 回ダウンロードされています。これには勇気づけられます。少なくとも何人かはこの本が便利だと思ってくれてるってことですから！

Dan Navarro

February 4, 2014

バージョン 0.3 に向けた前書き

心の中では本当にこの本を出版したくないと思ってるんです。完成してないんですから。

私がこういうときは、その言葉通りなんです。参考文献はまた十うんではないし、章のようやくはセクションタイトルのリストに過ぎないし、索引はないし、読者向けの練習問題はないし、構成は最適とは言えないし、トピックのカバーしている範囲は私の好みを十分に反映していません。さらに、内容的に満足していないところや、書き直さないといけない図もあり、矛盾点や誤字脱字を直す時間も十分にありませんでした。言い換えると、この本は未完成なんです。もし授業の締め切りや数週間後に予定されている赤ちゃんの存在がなければ、私は本当に後悔しなかったと思います。

つまり、もしあなたが大学での教材を探しているとか、Ph.D. の学生さんで R を勉強する方法を探しているとか、統計学の一般的な興味を持っているという人であれば、注意が必要だよと言いたい

のです。あなたが見ているのは最初の下書きで、あなたの目的に沿ったものではないかもしれませんからです。もし出版にお金がかかり、インターネットが周りにない世界であれば、こんな形で公開することは決して考えられないでしょう。この本に\$80 のお金を出す人がいるかと思うと（これは出版社が販売するにあたって、小売価格を申し出してくれたのです），ちょっと申し訳なく思います。しかし今は 21 世紀で、フリーで私のウェブサイトに PDF を乗せることができ、プリント・オン・デマンドサービスでハードコピーを配布すれば、出版社の教科書の半額ですみます。そして私の罪悪感を和らげるため、シェアしたいと思います！覚えておいてほしいのですが、次のサイトからみなさんは無料でソフトコピー（電子版）入手できますし、安価なハードコピーもオンラインで入手できます。

Soft copy: <http://www.compcogscisydney.com/learning-statistics-with-r.html>

Hard copy: www.lulu.com/content/13570633

とは言え、渓谷はまだ残っています：あなたが見ているのは、作業中のバージョン 0.3 です。もしいつの日かバージョン 1.0 になれば、この仕事に責任を持って、これは誰にでも使って欲しいテキストですと言いたいです。そのとき、私はおそらくインターネットに恥ずかしげもなく後悔し、道具として活用するでしょう。しかしその日が来るまでは、私ははっきりした態度は持てずに、この仕事についてアンビバレントな状態にあるというほかありません。

これを踏まえてですが、この本を強くお勧めするあるグループがあります。2013 年の学部生向け研究法 (DRIP と DRIP-A) を受講する心理学の学生です。あなたにとって、この本は理想出来なものになるでしょう。というのも、あなたの統計に関する講義に合わせて書かれたものだからです。もしこのノートによる欠点が発覚した場合、直ぐにそのコメントを適用して問題を修正することができます。効果的なことに、あなたのクラスに特化されたテキストを使うことができますし、それは無料で（電子版）あるいは手数料だけで（紙版）利用できるのです。さらに良いことに、このノートはすでに検証済みです。このノートのバージョン 0.1 は 2011 年のクラスですでに使われていて、バージョン 0.2 は 2012 年のクラスで使われたのです。そして今あなたが見ているのは、新しく改良されたバージョン 0.3 というわけです。このノートがチタンにメッキされたスティックだというつもりはありません—あなたが学生評価フォームでそう言いたいと思ったかもしれません、そのときはどうぞそうしてください—というのも、実際そこまでではないからです。しかし既に何年間か検証されてきていて、うまく機能してきたんだということは言っておきたいと思います。とはいえ、何か問題が生じたときにはわたしたちが直ぐに対応しますし、少なくとも教師の先生方のうち少なくとも一人は隅々までこの本を読んでいることは間違いないのです。

さてそれはさておき、この本が目指しているものが何なのかについて述べておきましょう。中心にある考えは、心理学を学ぶ人に向けて作られた統計の導入的教科書であること、です。ですから、類似の本にあなたが期待するような標準的トピックスはカバーしています。：研究デザイン、記述統計、仮説検証の理論、 t 検定、 χ^2 検定、ANOVA、回帰などです。しかし、いくつかの章では R の統計パッケージに言及しています。データの操作やそのほかのスクリプト、プログラミングなんかについての章も。さらにいえば、この本の目次を見たらお気づきになると思いますが、心理学の学生に統

計を教える際、これまで無視されてきたようなことも多く含まれています。ベイジアンか頻度主義か、という分断は確率の章で議論されますが、ネイマンとフィッシャーの仮説検定に関する不一致も扱います。確率と密度の違いについても説明します。ANOVA のアンバランスデザインにおける平方和の計算式、タイプ I, II, III の扱いについても触れます。またエピローグを見ていただければ、私が追加したかったもっと発展的な要素について明らかになるでしょう。

このアプローチを取る理由は全くシンプルなものです。すなわち、学生が乗りこなせるように、そして楽しめるように、したかったのです。最近の数年間は、心理学の学部生が R を習得するのにほとんど苦労しないことに驚かされています。それは全く簡単だというわけではないですし、成績をつける基準を設定するときは少し優しめにする必要はありますが、最終的にはそこに到達できます。同様に、統計的な考え方で現れる複雑で曖昧な表現を受け入れることに対しても、学生さんはそれほど問題を感じないようです。評価基準が適切に設定されていて、それが提示されている場合は。ですから学生が習得できるのに、教えないわけにいかないでしょう？ その潜在的な能力はとても魅力的です。もし彼らが R を学べば、おそらく最大で最も包括的な統計ツールライブラリである CRAN にアクセスできるということでもあるのですから。そしてもし確率理論の詳細について学べば、オーソドックスな帰無仮説検定からベイジアンの方法に乗り換えようと思ったとき、より乗り換えが容易になります。さらに、データ解析技術を学ぶときに高価で独自仕様になっているソフトウェアに捉われることなく、仕事に持って行くこともできます。

残念ながら、この本は全ての問題を解決する決定打ではありません。私の作業は作業中で、いつかは便利な道具になってくれると思います。数あるものの中の一つになる、そう思います。R を使う統計の基本的な導入をしようとする類書はたくさんあります。そして私の本が優れていると考えるほど、私は傲慢ではありません。でも、他の本よりも私はこの本を気に入っていますし、もしかしたら他の人もそう思ってくれるかもしれませんね。

Dan Navarro

January 13, 2013

Part I.

Background

1. Why do we learn statistics?

*"Thou shalt not answer questionnaires
 Or quizzes upon World Affairs,
 Nor with compliance
 Take any test. Thou shalt not sit
 With statisticians nor commit
 A social science"*

– W.H. Auden^{*1}

1.1

On the psychology of statistics

To the surprise of many students, statistics is a fairly significant part of a psychological education. To the surprise of no-one, statistics is very rarely the *favourite* part of one's psychological education. After all, if you really loved the idea of doing statistics, you'd probably be enrolled in a statistics class right now, not a psychology class. So, not surprisingly, there's a pretty large proportion of the student base that isn't happy about the fact that psychology has so much statistics in it. In view of this, I thought that the right place to start might be to answer some of the more common questions that people have about stats.

A big part of this issue at hand relates to the very idea of statistics. What is it? What's it there for? And why are scientists so bloody obsessed with it? These are all good questions, when you

^{*1}The quote comes from Auden's 1946 poem *Under Which Lyre: A Reactionary Tract for the Times*, delivered as part of a commencement address at Harvard University. The history of the poem is kind of interesting: <http://harvardmagazine.com/2007/11/a-poets-warning.html>

think about it. So let's start with the last one. As a group, scientists seem to be bizarrely fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It's a kind of article of faith among scientists – and especially social scientists – that your findings can't be trusted until you've done some stats. Undergraduate students might be forgiven for thinking that we're all completely mad, because no-one takes the time to answer one very simple question:

Why do you do statistics? Why don't scientists just use common sense?

It's a naive question in some ways, but most good questions are. There's a lot of good answers to it,^{*2} but for my money, the best answer is a really simple one: we don't trust ourselves enough. We worry that we're human, and susceptible to all of the biases, temptations and frailties that humans suffer from. Much of statistics is basically a safeguard. Using "common sense" to evaluate evidence means trusting gut instincts, relying on verbal arguments and on using the raw power of human reason to come up with the right answer. Most scientists don't think this approach is likely to work.

In fact, come to think of it, this sounds a lot like a psychological question to me, and since I do work in a psychology department, it seems like a good idea to dig a little deeper here. Is it really plausible to think that this "common sense" approach is very trustworthy? Verbal arguments have to be constructed in language, and all languages have biases – some things are harder to say than others, and not necessarily because they're false (e.g., quantum electrodynamics is a good theory, but hard to explain in words). The instincts of our "gut" aren't designed to solve scientific problems, they're designed to handle day to day inferences – and given that biological evolution is slower than cultural change, we should say that they're designed to solve the day to day problems for a *different world* than the one we live in. Most fundamentally, reasoning sensibly requires people to engage in "induction", making wise guesses and going beyond the immediate evidence of the senses to make generalisations about the world. If you think that you can do that without being influenced by various distractors, well, I have a bridge in London I'd like to sell you. Heck, as the next section shows, we can't even solve "deductive" problems (ones where no guessing is required) without being influenced by our pre-existing biases.

1.1.1 **The curse of belief bias**

People are mostly pretty smart. We're certainly smarter than the other species that we share

^{*2}Including the suggestion that common sense is in short supply among scientists.

the planet with (though many people might disagree). Our minds are quite amazing things, and we seem to be capable of the most incredible feats of thought and reason. That doesn't make us perfect though. And among the many things that psychologists have shown over the years is that we really do find it hard to be neutral, to evaluate evidence impartially and without being swayed by pre-existing biases. A good example of this is the **belief bias effect** in logical reasoning: if you ask people to decide whether a particular argument is logically valid (i.e., conclusion would be true if the premises were true), we tend to be influenced by the believability of the conclusion, even when we shouldn't. For instance, here's a valid argument where the conclusion is believable:

All cigarettes are expensive (Premise 1)
Some addictive things are inexpensive (Premise 2)
Therefore, some addictive things are not cigarettes (Conclusion)

And here's a valid argument where the conclusion is not believable:

All addictive things are expensive (Premise 1)
Some cigarettes are inexpensive (Premise 2)
Therefore, some cigarettes are not addictive (Conclusion)

The logical *structure* of argument #2 is identical to the structure of argument #1, and they're both valid. However, in the second argument, there are good reasons to think that premise 1 is incorrect, and as a result it's probably the case that the conclusion is also incorrect. But that's entirely irrelevant to the topic at hand; an argument is deductively valid if the conclusion is a logical consequence of the premises. That is, a valid argument doesn't have to involve true statements.

On the other hand, here's an invalid argument that has a believable conclusion:

All addictive things are expensive (Premise 1)
Some cigarettes are inexpensive (Premise 2)
Therefore, some addictive things are not cigarettes (Conclusion)

And finally, an invalid argument with an unbelievable conclusion:

All cigarettes are expensive (Premise 1)
Some addictive things are inexpensive (Premise 2)
Therefore, some cigarettes are not addictive (Conclusion)

Now, suppose that people really are perfectly able to set aside their pre-existing biases about what is true and what isn't, and purely evaluate an argument on its logical merits. We'd expect 100%

of people to say that the valid arguments are valid, and 0% of people to say that the invalid arguments are valid. So if you ran an experiment looking at this, you'd expect to see data like this:

	conclusion feels true	conclusion feels false
argument is valid	100% say "valid"	100% say "valid"
argument is invalid	0% say "valid"	0% say "valid"

If the psychological data looked like this (or even a good approximation to this), we might feel safe in just trusting our gut instincts. That is, it'd be perfectly okay just to let scientists evaluate data based on their common sense, and not bother with all this murky statistics stuff. However, you guys have taken psych classes, and by now you probably know where this is going.

In a classic study, **Evans1983** ran an experiment looking at exactly this. What they found is that when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you'd hope:

	conclusion feels true	conclusion feels false
argument is valid	92% say "valid"	
argument is invalid		8% say "valid"

Not perfect, but that's pretty good. But look what happens when our intuitive feelings about the truth of the conclusion run against the logical structure of the argument:

	conclusion feels true	conclusion feels false
argument is valid	92% say "valid"	46% say "valid"
argument is invalid	92% say "valid"	8% say "valid"

Oh dear, that's not as good. Apparently, when people are presented with a strong argument that contradicts our pre-existing beliefs, we find it pretty hard to even perceive it to be a strong argument (people only did so 46% of the time). Even worse, when people are presented with a weak argument that agrees with our pre-existing biases, almost no-one can see that the argument is weak (people got that one wrong 92% of the time!).^{*3}

If you think about it, it's not as if these data are horribly damning. Overall, people did do better

^{*3}In my more cynical moments I feel like this fact alone explains 95% of what I read on the internet.

than chance at compensating for their prior biases, since about 60% of people's judgements were correct (you'd expect 50% by chance). Even so, if you were a professional "evaluator of evidence", and someone came along and offered you a magic tool that improves your chances of making the right decision from 60% to (say) 95%, you'd probably jump at it, right? Of course you would. Thankfully, we actually do have a tool that can do this. But it's not magic, it's statistics. So that's reason #1 why scientists love statistics. It's just *too easy* for us to "believe what we want to believe". So instead, if we want to "believe in the data", we're going to need a bit of help to keep our personal biases under control. That's what statistics does, it helps keep us honest.

1.2

The cautionary tale of Simpson's paradox

The following is a true story (I think!). In 1973, the University of California, Berkeley had some worries about the admissions of students into their postgraduate courses. Specifically, the thing that caused the problem was that the gender breakdown of their admissions, which looked like this:

	Number of applicants	Percent admitted
Males	8442	44%
Females	4321	35%

Given this, they were worried about being sued!*⁴ Given that there were nearly 13,000 applicants, a difference of 9% in admission rates between males and females is just way too big to be a coincidence. Pretty compelling data, right? And if I were to say to you that these data *actually* reflect a weak bias in favour of women (sort of!), you'd probably think that I was either crazy or sexist.

Oddly, it's actually sort of true. When people started looking more carefully at the admissions data they told a rather different story (**Bickel1975**). Specifically, when they looked at it on a department by department basis, it turned out that most of the departments actually had a slightly *higher* success rate for female applicants than for male applicants. The table below shows the admission figures for the six largest departments (with the names of the departments removed for privacy reasons):

*⁴Earlier versions of these notes incorrectly suggested that they actually were sued. But that's not true. There's a nice commentary on this here: <https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html>. A big thank you to Wilfried Van Hirtum for pointing this out to me.

Department	Males		Females	
	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Remarkably, most departments had a *higher* rate of admissions for females than for males! Yet the overall rate of admission across the university for females was *lower* than for males. How can this be? How can both of these statements be true at the same time?

Here's what's going on. Firstly, notice that the departments are *not* equal to one another in terms of their admission percentages: some departments (e.g., A, B) tended to admit a high percentage of the qualified applicants, whereas others (e.g., F) tended to reject most of the candidates, even if they were high quality. So, among the six departments shown above, notice that department A is the most generous, followed by B, C, D, E and F in that order. Next, notice that males and females tended to apply to different departments. If we rank the departments in terms of the total number of male applicants, we get **A>B>D>C>F>E** (the "easy" departments are in bold). On the whole, males tended to apply to the departments that had high admission rates. Now compare this to how the female applicants distributed themselves. Ranking the departments in terms of the total number of female applicants produces a quite different ordering **C>E>D>F>**A>B****. In other words, what these data seem to be suggesting is that the female applicants tended to apply to "harder" departments. And in fact, if we look at Figure 1.1 we see that this trend is systematic, and quite striking. This effect is known as **Simpson's paradox**. It's not common, but it does happen in real life, and most people are very surprised by it when they first encounter it, and many people refuse to even believe that it's real. It is very real. And while there are lots of very subtle statistical lessons buried in there, I want to use it to make a much more important point: doing research is hard, and there are *lots* of subtle, counter-intuitive traps lying in wait for the unwary. That's reason #2 why scientists love statistics, and why we teach research methods. Because science is hard, and the truth is sometimes cunningly hidden in the nooks and crannies of complicated data.

Before leaving this topic entirely, I want to point out something else really critical that is often overlooked in a research methods class. Statistics only solves *part* of the problem. Remember that we started all this with the concern that Berkeley's admissions processes might be unfairly biased against female applicants. When we looked at the "aggregated" data, it did seem like the university was discriminating against women, but when we "disaggregate" and looked at the

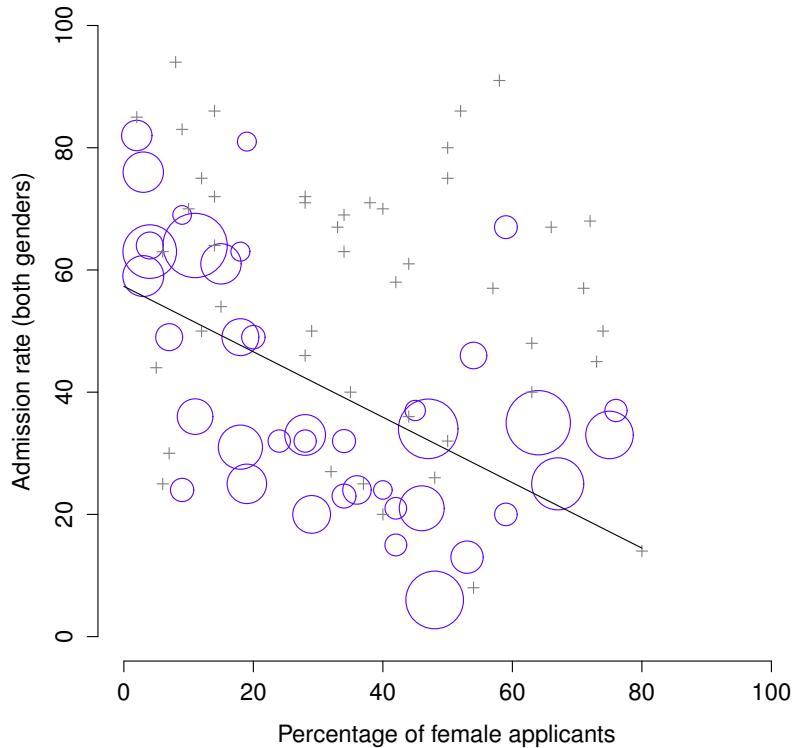


Figure1.1 The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from [Bickel1975](#). Circles plot departments with more than 40 applicants; the area of the circle is proportional to the total number of applicants. The crosses plot departments with fewer than 40 applicants.

.....

individual behaviour of all the departments, it turned out that the actual departments were, if anything, slightly biased in favour of women. The gender bias in total admissions was caused by the fact that women tended to self-select for harder departments. From a legal perspective, that would probably put the university in the clear. Postgraduate admissions are determined at the level of the individual department, and there are good reasons to do that. At the level of individual departments the decisions are more or less unbiased (the weak bias in favour of females at that level is small, and not consistent across departments). Since the university can't dictate which departments people choose to apply to, and the decision making takes place at the level of the department it can hardly be held accountable for any biases that those choices produce.

That was the basis for my somewhat glib remarks earlier, but that's not exactly the whole story, is it? After all, if we're interested in this from a more sociological and psychological perspective, we might want to ask *why* there are such strong gender differences in applications. Why do males tend to apply to engineering more often than females, and why is this reversed for the English department? And why is it the case that the departments that tend to have a female-application bias tend to have lower overall admission rates than those departments that have a male-application bias? Might this not still reflect a gender bias, even though every single department is itself unbiased? It might. Suppose, hypothetically, that males preferred to apply to "hard sciences" and females prefer "humanities". And suppose further that the reason for why the humanities departments have low admission rates is because the government doesn't want to fund the humanities (Ph.D. places, for instance, are often tied to government funded research projects). Does that constitute a gender bias? Or just an unenlightened view of the value of the humanities? What if someone at a high level in the government cut the humanities funds because they felt that the humanities are "useless chick stuff". That seems pretty *blatantly* gender biased. None of this falls within the purview of statistics, but it matters to the research project. If you're interested in the overall structural effects of subtle gender biases, then you probably want to look at *both* the aggregated and disaggregated data. If you're interested in the decision making process at Berkeley itself then you're probably only interested in the disaggregated data.

In short there are a lot of critical questions that you can't answer with statistics, but the answers to those questions will have a huge impact on how you analyse and interpret data. And this is the reason why you should always think of statistics as a *tool* to help you learn about your data. No more and no less. It's a powerful tool to that end, but there's no substitute for careful thought.

1.3 _____

Statistics in psychology

I hope that the discussion above helped explain why science in general is so focused on statistics. But I'm guessing that you have a lot more questions about what role statistics plays in psychology, and specifically why psychology classes always devote so many lectures to stats. So here's my attempt to answer a few of them...

- **Why does psychology have so much statistics?**

To be perfectly honest, there's a few different reasons, some of which are better than others. The most important reason is that psychology is a statistical science. What I mean

by that is that the “things” that we study are *people*. Real, complicated, gloriously messy, infuriatingly perverse people. The “things” of physics include objects like electrons, and while there are all sorts of complexities that arise in physics, electrons don’t have minds of their own. They don’t have opinions, they don’t differ from each other in weird and arbitrary ways, they don’t get bored in the middle of an experiment, and they don’t get angry at the experimenter and then deliberately try to sabotage the data set (not that I’ve ever done that!). At a fundamental level psychology is harder than physics.^{*5}

Basically, we teach statistics to you as psychologists because you need to be better at stats than physicists. There’s actually a saying used sometimes in physics, to the effect that “if your experiment needs statistics, you should have done a better experiment”. They have the luxury of being able to say that because their objects of study are pathetically simple in comparison to the vast mess that confronts social scientists. And it’s not just psychology. Most social sciences are desperately reliant on statistics. Not because we’re bad experimenters, but because we’ve picked a harder problem to solve. We teach you stats because you really, really need it.

- **Can’t someone else do the statistics?**

To some extent, but not completely. It’s true that you don’t need to become a fully trained statistician just to do psychology, but you do need to reach a certain level of statistical competence. In my view, there’s three reasons that every psychological researcher ought to be able to do basic statistics:

- Firstly, there’s the fundamental reason: statistics is deeply intertwined with research design. If you want to be good at designing psychological studies, you need to at the very least understand the basics of stats.
- Secondly, if you want to be good at the psychological side of the research, then you need to be able to understand the psychological literature, right? But almost every paper in the psychological literature reports the results of statistical analyses. So if you really want to understand the psychology, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.
- Thirdly, there’s a big practical problem with being dependent on other people to do all your statistics: statistical analysis is *expensive*. If you ever get bored and want to look up how much the Australian government charges for university fees, you’ll notice something interesting: statistics is designated as a “national priority” category, and so the fees are much, much lower than for any other area of study. This is because

^{*5}Which might explain why physics is just a teensy bit further advanced as a science than we are.

there's a massive shortage of statisticians out there. So, from your perspective as a psychological researcher, the laws of supply and demand aren't exactly on your side here! As a result, in almost any real life situation where you want to do psychological research, the cruel facts will be that you don't have enough money to afford a statistician. So the economics of the situation mean that you have to be pretty self-sufficient.

Note that a lot of these reasons generalise beyond researchers. If you want to be a practicing psychologist and stay on top of the field, it helps to be able to read the scientific literature, which relies pretty heavily on statistics.

- **I don't care about jobs, research, or clinical work. Do I need statistics?**

Okay, now you're just messing with me. Still, I think it should matter to you too. Statistics should matter to you in the same way that statistics should matter to *everyone*. We live in the 21st century, and data are *everywhere*. Frankly, given the world in which we live these days, a basic knowledge of statistics is pretty damn close to a survival tool! Which is the topic of the next section.

1.4

Statistics in everyday life

*"We are drowning in information,
but we are starved for knowledge"*

– Various authors, original probably John Naisbitt

When I started writing up my lecture notes I took the 20 most recent news articles posted to the ABC news website. Of those 20 articles, it turned out that 8 of them involved a discussion of something that I would call a statistical topic and 6 of those made a mistake. The most common error, if you're curious, was failing to report baseline data (e.g., the article mentions that 5% of people in situation X have some characteristic Y, but doesn't say how common the characteristic is for everyone else!). The point I'm trying to make here isn't that journalists are bad at statistics (though they almost always are), it's that a basic knowledge of statistics is very helpful for trying to figure out when someone else is either making a mistake or even lying to you. In fact, one of the biggest things that a knowledge of statistics does to you is cause you to get angry at the newspaper or the internet on a far more frequent basis. You can find a good example of this in Section 3.1.5. In later versions of this book I'll try to include more anecdotes along those lines.

There's more to research methods than statistics

So far, most of what I've talked about is statistics, and so you'd be forgiven for thinking that statistics is all I care about in life. To be fair, you wouldn't be far wrong, but research methodology is a broader concept than statistics. So most research methods courses will cover a lot of topics that relate much more to the pragmatics of research design, and in particular the issues that you encounter when trying to do research with humans. However, about 99% of student *fears* relate to the statistics part of the course, so I've focused on the stats in this discussion, and hopefully I've convinced you that statistics matters, and more importantly, that it's not to be feared. That being said, it's pretty typical for introductory research methods classes to be very stats-heavy. This is not (usually) because the lecturers are evil people. Quite the contrary, in fact. Introductory classes focus a lot on the statistics because you almost always find yourself needing statistics before you need the other research methods training. Why? Because almost all of your assignments in other classes will rely on statistical training, to a much greater extent than they rely on other methodological tools. It's not common for undergraduate assignments to require you to design your own study from the ground up (in which case you would need to know a lot about research design), but it *is* common for assignments to ask you to analyse and interpret data that were collected in a study that someone else designed (in which case you need statistics). In that sense, from the perspective of allowing you to do well in all your other classes, the statistics is more urgent.

But note that "urgent" is different from "important" – they both matter. I really do want to stress that research design is just as important as data analysis, and this book does spend a fair amount of time on it. However, while statistics has a kind of universality, and provides a set of core tools that are useful for most types of psychological research, the research methods side isn't quite so universal. There are some general principles that everyone should think about, but a lot of research design is very idiosyncratic, and is specific to the area of research that you want to engage in. To the extent that it's the details that matter, those details don't usually show up in an introductory stats and research methods class.

Part II.

Describing and displaying data with JASP

2. JASP 入門

ロボットは良く働く。

—Roger Zelazny^{*1}

この章では、JASP の入門方法について説明します。JASP をダウンロードしてインストールする方法について簡単に説明しますが、この章のほとんどでは、JASP ユーザーインターフェースの使用方法の入門に焦点を当てます。この章の目標は、統計の概念を学ぶことではありません。そうではなく、JASP の仕組みとソフトと快適にやり取りする方法について学びます。これを行うために、データセットと変数を検討することに時間を費やします。そうすることで、JASP での作業がどのようなものかを少し感じることができます。

ただし、詳細に進む前に、JASP を使用する理由について少し説明することには少なからず価値があります。本書を読んでいるということは、あなたにはすでに JASP を使用する理由があるのでしょうただ、その理由が「統計の授業で使用しているから」である場合、なぜ教授が授業で JASP を使用することを選択したのかについて少し説明する価値があります。もちろん、他の人々がなぜ JASP を選択するのかは本当のところ知らないので、私が使う理由について話します。

- 当たり前のことですが、手動で行うよりもコンピューターで統計を行うことは、速く、簡単で、強力であるということは述べる価値があります。コンピューターは頭を使わない反復作業に優れており、統計計算の多くは頭を使わない反復作業です。ほとんどの人にとって、鉛筆と紙で統計計算を行う唯一の理由は、学習のためです（新しい概念を学ぶ時は専門家でさえこれを行います）。私の授業では、そのようにいくつかの計算を行うことを時々提案しますが、その唯一の真の価値は教育です。自分でいくつか計算することは、統計の「感覚」を得るために役立ちますので、一度行う価値があります。しかし、一度だけです！
- 従来のスプレッドシート（例えば、Microsoft Excel）で統計を行うことは、一般的には長期的に見ると良くない考えです。多くの人はそれらに馴染みがあると感じるかもしれません、スプレッドシートでは、分析できる範囲が非常に限られています。スプレッドシートを使用し

^{*1}Source: *Dismal Light* (1968).

て実際のデータ分析を行う習慣を身につけた場合、非常に深い穴に掘り込まれることになります。

- プロプライエタリ・ソフトウェア^{*2}を避けることは、とても良い考えです。購入できる商用パッケージはたくさんあります。私が好きなものもあれば、そうでないものもあります。通常、商用パッケージは外観の体裁がとても良く、一般に非常に強力です（スプレッドシートよりもはるかに強力です）。しかし、非常に高価です。通常、企業は「学生版」（本物の一部が使えない版）を非常に安く販売し、その後、びっくりするような価格で完全版の「教育版」を販売しています。また、驚愕するほど高い値段で、商用ライセンスを販売しています。ここでのビジネスモデルは、学生時代にあなたを引き込んで、現実の世界に出かけるときに彼らのツールに依存したままにすることです。しゃくにさわるからといって彼らを責めるのは難しいですが、個人的には、避けることができるなら、何千ドルも払いたくはありません。そして、あなたはそれを避けることができます。JASP のような、オープンソースで無料のパッケージを利用すれば、法外なライセンス料を支払う必要がなくなります。

これらが JASP を使用する主な理由です。ただし、欠点がないわけではありません。JASP は、比較的新しいため^{*3}、それをサポートする教科書やその他のリソースがあまりありません。私たちがよく陥ってしまういくつかの迷惑な癖がありますが、全体的には長所が短所を上回っていると思います。これまでに出会った他のどの選択肢よりもそうです。

2.1 _____

JASP のインストール

さて、セールストークは十分でしょう。始めましょう。他のソフトウェアと同じように、JASP はコンピューターにインストールする必要があります。幸いなことに、JASP はオンラインで無料で配布されており、JASP ホームページからダウンロードできます。

<https://jasp-stats.org/>

ページの上方で、「ダウンロード」という見出しをクリックします。次に、Windows ユーザー、Mac ユーザー、および Linux ユーザー用の個別のリンクが表示されます。関連リンクをたどると、読んで字のとおりのオンラインの説明があります。この原稿の執筆時点では、JASP の現在のバージョンは 0.9.2.0 ですが、通常は数か月ごとに更新されるので、おそらく新しいバージョンが必要になります。

^{*2} 訳注 ソフトウェアの配布者が、ソフトの使用・改変・複製などを制限しているソフトウェア

^{*3} これが執筆された 2019 年 5 月

ます。^{*4}

2.1.1 JASP の起動

いずれにせよ、使用しているオペレーティングシステムに関係なく、JASP を開いて、起動させましょう。JASP の初回起動時に、図 2.1 のようなユーザーインターフェイスが表示されます。

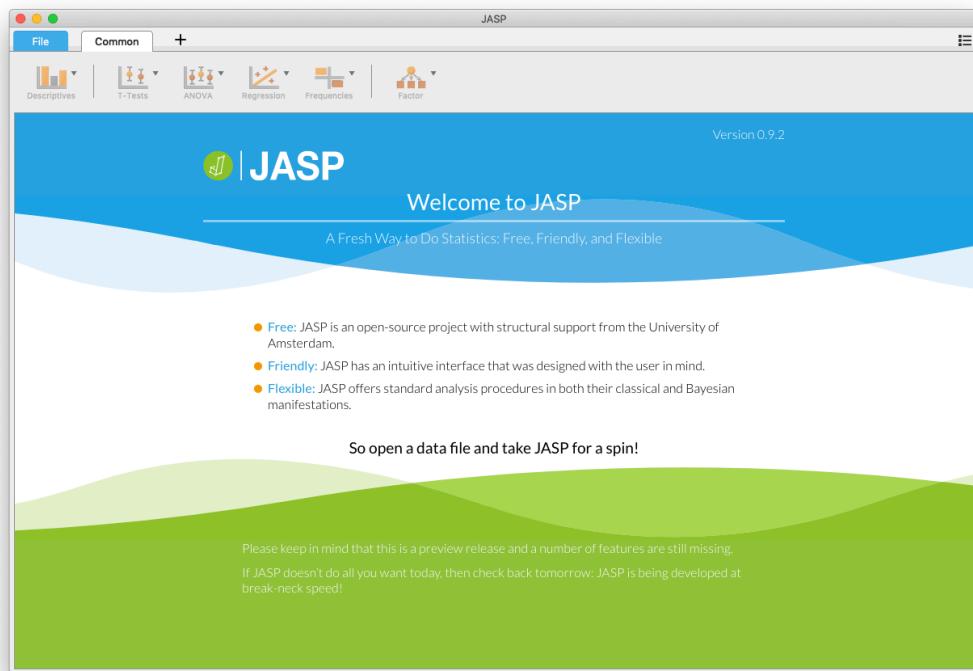


Figure2.1 起動時の JASP

他の統計ソフトウェアの使用経験がある場合、データの入力を開始する場所がないことに少しがっかりするかもしれません。これは JASP 開発者側の意図的な決定です。彼らの哲学は、ユーザーが最も快適なエディターを使用できるようにすることです^{*5}。したがって、JASP にデータを読み込むため上で推奨される方法は、CSV ファイル (.csv) を読み込むことです。CSV ファイルは、スプレッドシートプログラムで作成（と開くことが）できるテキストベースのデータ形式です。これについて

^{*4}この本でやる作業とは違って JASP は頻繁に更新されます。実際、この本の執筆中に何度かアップグレードがありましたが、この本の内容に大きな違いはありませんでした。

^{*5}この重要な問題についての議論が、<https://jasp-stats.org/2018/05/15/data-editing-in-jasp/> にあるので、参照ください

の詳細は、このあとすぐに説明します。

2.2

分析

分析は、上にあるいくつかのボタンから選択できます。分析を選択すると、特定の分析のための「options panel」が表示されます。あなたは、分析のさまざまな部分にさまざまな変数を割り当てたり、さまざまなオプションを選択できます。同時に、分析結果は右側の「Results panel」に表示され、オプションを変更するとリアルタイムで更新されます。

分析を正しく設定したら、オプションパネルの右上にある「OK」ボタンをクリックして、分析オプションを閉じることができます。これらのオプションに戻りたい場合は、結果をクリックすることができます。このようにして、あなた（または同僚）が以前に作成した分析に戻ることができます。

特定の分析が不要になった場合は、結果のコンテキストメニューで削除できます。特定の結果のヘッダー（もしくは、▼）をクリックしてメニューを表示して、「Remove Analysis」を選ぶと、分析を削除できます。しかし、これについては後で詳しく説明します。まず、JASP にいくつかのデータを入れてみましょう。

2.3

JASPへのデータ読み込み

データ分析を行う時に、私たちに関係があると思われるファイル形式がいくつかあります。この本の観点から特に重要なのは 2 つです:

- *jasp files* は、拡張子が.jasp のファイルです。これは、JASP がデータ、変数、および分析を保存するために使用する標準的なファイル形式です。
- コンマ区切り (CSV) ファイルは、拡張子が.csv のファイルです。これは、一般的な古いテキストファイルであり、さまざまなソフトウェアプログラムで開くことができます。csv ファイルは非常にシンプルなので、csv ファイルにデータを保存するのにかなりよく使われます。

2.3.1 CSV ファイルからデータをインポートする

かなり広く使用されているデータ形式の 1 つは、地味な「カンマ区切り」ファイルです。CSV ファ

The screenshot shows a Microsoft Excel spreadsheet titled 'booksales'. On the left, there is a table with four columns: Month, Days, Sales, and Stock.Levels. The first row contains column headers. Rows 2 through 13 contain data for each month from January to December. The right side of the screen displays the raw CSV text data, which includes the header 'Month,Days,Sales,Stock.Levels' followed by 13 rows of data, each separated by a comma. The data shows varying values for Days, Sales, and Stock.Levels across the months.

Month	Days	Sales	Stock.Levels
January	31	0	high
February	28	100	high
March	31	200	low
April	30	50	out
May	31	0	out
June	30	0	high
July	31	0	high
August	31	0	high
September	30	0	high
October	31	0	high
November	30	0	high
December	31	0	high

Month,Days,Sales,Stock.Levels
January,31,0,high
February,28,100,high
March,31,200,low
April,30,50,out
May,31,0,out
June,30,0,high
July,31,0,high
August,31,0,high
September,30,0,high
October,31,0,high
November,30,0,high
December,31,0,high

Figure2.2 booksales.csv のデータファイル。左側は、スプレッドシートソフトを使用してファイルを開きました。ファイルが基本的にテーブルであることを示しています。右側は、同じファイルが標準のテキストエディター（Mac のTextEdit）で開きました。ファイルがどのようにフォーマットされているか示しています。テーブルへの記入は、コンマで区切られます。

イルとも呼ばれ、通常は拡張子.csvを持ちます。CSV ファイルは、昔からある単にシンプルなテキストファイルであり、保存されるのは基本的に单なるデータのテーブルです。これを Figure 2.2 に示します。この図は、私が作成した booksales.csv というファイルを示しています。ご覧のとおり、各行は 1 ヶ月間の書籍販売データを表します。最初の行には実際のデータは含まれませんが、変数の名前があります。

CSV ファイル（あなたが作成したファイルか、誰かが提供したファイル）があれば、左上隅にある「File」タブをクリックして「Open」を選択し、表示されたオプションから選択をすることで、JASP でファイルを開けます。最も一般的には、「Computer」を選択してから「Browse」を選択します。これにより、あなたが使っているオペレーティングシステムに特有のファイルブラウザが開きます。Mac を使用している場合は、ファイルの選択に使用する通常の Finder ウィンドウのように見えるでしょう。Windows では、エクスプローラーウィンドウのように見えます。Mac での表示例は、Figure

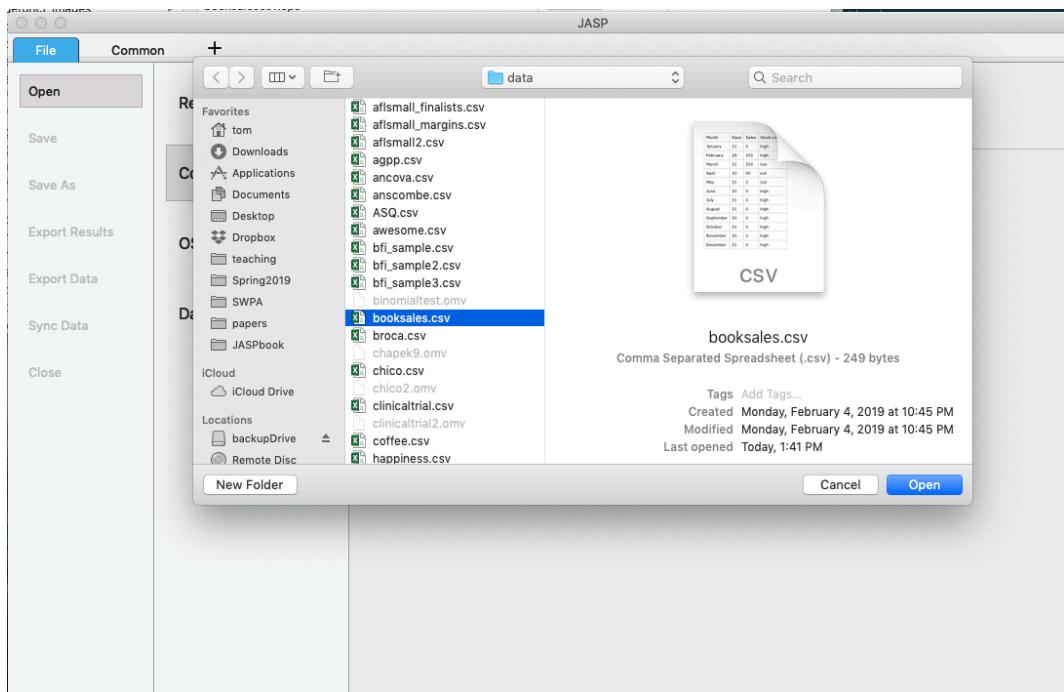


Figure2.3 JASP がインポートする CSV ファイルを選択するように求める Mac 上のダイアログボックス。Mac ユーザーはこれをすぐに理解すると思います。これは、Mac があなたにファイルを探す時に要求する一般的な方法です。Windows ユーザーにはこれは表示されませんが、代わりに、ファイルを選択するときに Windows がいつも出してくる通常のエクスプローラーウィンドウが表示されます。

~??に示されています。あなたはきっと自分のコンピュータに慣れているでしょうから、インポートしたい csv ファイルを見つけるのに問題はないはずです！ 必要なものを見つけて、「Open」ボタンをクリックしてください。

2.4

The spreadsheet

Once loaded into JASP, data is represented in a spreadsheet with each column representing a ‘variable’ and each row representing a ‘case’ or ‘participant’ .

2.4.1 Variables

The most commonly used variables in JASP are ‘Data Variables’ , which contain data loaded from a CSV file. Data variables can be one of three measurement levels, which are designated by the symbol in the header of the variable’ s column.

Nominal variables are for categorical variables which are text labels, for example a column called Gender with the values Male and Female would be nominal. So would a person’ s name. Nominal variable values can also have a numeric value. These variables are used most often when importing data which codes values with numbers rather than text. For example, a column in a dataset may contain the values 1 for males, and 2 for females. It is possible to add nice ‘human-readable’ labels to these values with the variable editor (more on this later).

Ordinal variables are like Nominal variables, except the values have a specific order. An example is a Likert scale with 3 being ‘strongly agree’ and -3 being ‘strongly disagree’ .

Scale variables are variables which exist on a continuous scale. Examples might be height or weight. This is also referred to as ‘Interval’ or ‘Ratio scale’ .

Note that when opening a data file JASP will try and guess the variable type from the data in each column. In both cases this automatic approach may not be correct, and it may be necessary to manually specify the variable type with the variable editor.

2.4.2 Computed variables

Computed Variables are those which take their value by performing a computation on other variables. Computed Variables can be used for a range of purposes, including log transforms, z-scores, sum-scores, negative scoring and means.

Computed variables can be added to the data set with the ‘+’ button in the header row of the data spreadsheet. This will produce a dialog box where you can specify the formula using either R code or a drag-and-drop interface. At this point, I simply want you to know that the capability exists, but describing how to do it is a little beyond our scope right now. More later!

2.4.3 Copy and Paste

As a final note, we will mention that JASP produces nice American Psychological Association (APA) formatted tables and attractive plots. It is often useful to be able to copy and paste these, perhaps into a Word document, or into an email to a colleague. To copy results, click on the header of the object of interest and from the menu select exactly what you want to copy.

Selecting “copy” copies the content to the clipboard and this can be pasted into other programs in the usual way. You can practice this later on when we do some analyses. Also, if you use the \LaTeX document preparation system, you can select “Copy special” and “ \LaTeX code”; doing so will place the \LaTeX syntax into your clipboard.

2.5

Changing data from one measurement scale to another

Sometimes you want to change the variable level. This can happen for all sorts of reasons. Sometimes when you import data from files, it can come to you in the wrong format. Numbers sometimes get imported as nominal, text values. Dates may get imported as text. ParticipantID values can sometimes be read as continuous: nominal values can sometimes be read as ordinal or even continuous. There’s a good chance that sometimes you’ll want to convert a variable from one measurement level into another one. Or, to use the correct term, you want to **coerce** the variable from one class into another.

In 2.4 we saw how to specify different variable levels, and if you want to change a variable’s measurement level then you can do this in the JASP data view for that variable. Just click the check box for the measurement level you want – continuous, ordinal, or nominal.

2.6

Quitting JASP

There’s one last thing I should cover in this chapter: how to quit JASP. It’s not hard, just close the program the same way you would any other program. However, what you might want to do before you quit is save your work! There are two parts to this: saving any changes to the data set, and saving the analyses that you ran.

It is good practice to save any changes to the data set as a *new* data set. That way you can always go back to the original data. To save any changes in JASP, select ‘Export Data’ from the ‘File’ tab, click ‘Browse’ and navigate to the directory location in which you want to save the file, and create a new file name for the changed data set.

Alternatively, you can save *both* the changed data and any analyses you have undertaken by saving as a .jasp file. To do this, from the ‘File’ tab select ‘Save as’, click ‘Browse’ to navigate to the directory location in which you want to save the file, and type in a file name for this .jasp

file. Remember to save the file in a location where you can find it again later. I usually create a new folder for specific data sets and analyses.

2.7

Summary

Every book that tries to teach a new statistical software program to novices has to cover roughly the same topics, and in roughly the same order. Ours is no exception, and so in the grand tradition of doing it just the same way everyone else did it, this chapter covered the following topics:

- Section 2.1. We downloaded and installed JASP, and started it up.
- Section 2.2. We very briefly oriented to the part of JASP where analyses are done and results appear, but then deferred this until later in the book.
- Section 2.3. We saw how to load data files (formatted as .csv files) in JASP.
- Section 2.4. We spent more time looking at the spreadsheet part of JASP, and considered different variable types, and briefly mentioned how to compute new variables.
- Section 2.5. And saw that sometimes we need to coerce data from one type to another.
- Section 2.6. Finally, we looked at good practice in terms of saving your data set and analyses when you have finished and are about to quit JASP.

We still haven't arrived at anything that resembles data analysis. Maybe the next Chapter will get us a bit closer!

3. 記述統計

新しいデータを手に入れたときはいつでも、最初にやるべきことの一つは、データを簡単にまとめ、その傾向を理解しやすくする方法を見つけることです。これこそ記述統計の全てです（この反対は推測統計です）。実際、多くの人が“統計”という言葉を、記述統計の同義語だと思っています。この章で話そうとしているのがそれなのですが、詳細に入る前に、なぜ記述統計が必要なのかという感覚を掴んでもらいたいと思います。そうするためにまず、`aflsmall_margins` ファイルを開いて、ファイルの中にある変数を見てみましょう。

このように、一つの変数 `afl.margins` しかありません。この章ではこの変数に注目しますので、これが何なのか少し説明します。この本に含まれるデータセットとは違って、これは実際に得たデータであり、オーストラリアのフットボールリーグ (AFL) に関するデータです^{*1} 変数 `afl.margins` は、2010 年シーズンのホームゲーム、アウェイゲーム含めた全 176 ゲームの得点差 (獲得点数) が含まれています。

このアウトプットから、このデータが何を言おうとしているのか掴み取るのは簡単ではありません。“データを眺めている”だけでは、データを理解するのに全く効果的ではないのです。このデータが何を言おうとしているか、それを掴み取るために、記述統計を計算する必要があり（この章で扱います）、わかりやすい図を描くことです（第 4 章で扱います）。二つのやり方のうち、記述統計の方がより簡単なのですが、私たちが見ようとしているデータがどんなものなのかのイメージを掴むために、この `afl.margins` データのヒストグラムをお見せすることにしましょう。図 3.2 を見てください。どうやってヒストグラムを描くかについては、セクション 4.1 で説明しますから。今は、ヒストグラムを見てそれが `afl.margins` データを正しく理解する方法であることがわかってもらえば結構です。

^{*1} オーストラリア人ではない人にむけた注意：AFL はオーストラリアのルールで行われるフットボール競技です。この章を読むためにオーストラリアのルールを調べる必要は全くありません。

	afl.margins	
1	56	
2	31	
3	56	
4	8	
5	32	
6	14	
7	36	
8	56	
9	19	
10	1	
11	3	
12	104	
13	43	
14	44	

Figure3.1 JASP が aflsmall_margins.csv ファイルを開いて変数を見せて いるスクリーンショット

.....

3.1 _____

傾向の測定

図 3.2 で示したようなデータの絵を描くというのは、データがどうなっているのかの“要点”をもたらす優れた方法です。データをいくつかの単純な“集約された”統計量に凝縮してみることが、特に便利です。いろんな場面で、まず計算してもらいたいのは**中心傾向**についての測定です。すなわち、あなたのデータの“平均”や“真ん中”がどのあたりにあるんのかを捉えて欲しいのです。最もよく使われる三つの数字は、平均値、中央値、最頻値です。これを順番に説明していくので、その後でそれぞれがどういうときに便利なのかをみていきましょう。

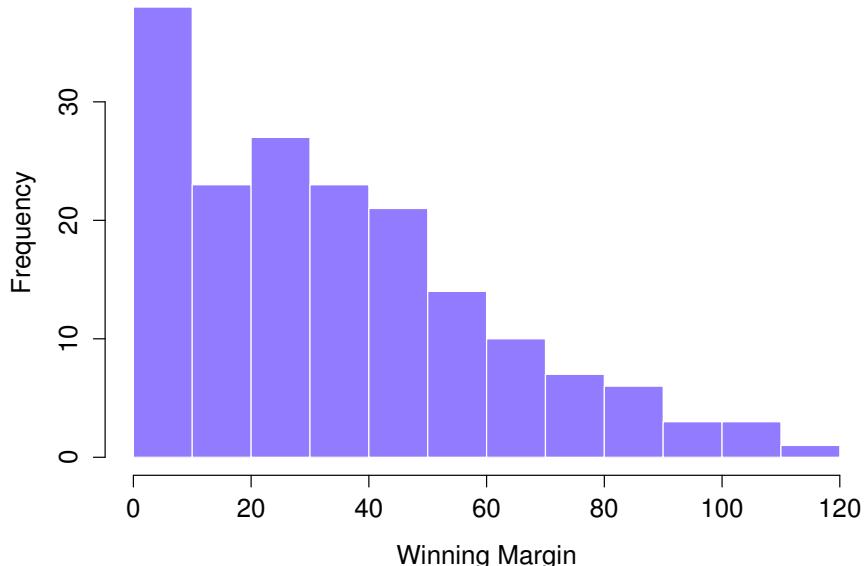


Figure3.2 2020 年の AFL 得点差データ (変数 `afl.margins`) のヒストグラム。ご想像の通り、より大差がつくゲームはより少ないので見て取れます。

3.1.1 平均値

観測値のセットの**平均値**は、普通の、昔ながらの平均値です。全ての値を足し上げて、足した値の数で割ります。最初の 5 つの AFL の得点差は、56,31,56,8,32 ですが、これらの平均値を計算するには単に次のようにするだけです。

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36.60$$

もちろん、この平均の定義は誰にとっても新しいものではないでしょう。アベレージ (すなわち平均値) は、日常生活でもよく使われていますから、みなさんにとってもなじみ深い物でしょう。平均の概念についてはみなさん理解しているでしょうから、この計算を表記するために統計学者が使う数学的表記法について説明する機会とさせてもらって、その後で JASP でどのように計算するか紹介することにしましょう。

最初に導入する表記法は N です。これは平均するときの観測度数の数を表すのに使います (今回の場合は $N = 5$ です)。つぎに、観測値そのものについてのラベルをつけます。これには伝統的に X が用いられ、具体的にそのどれを指し示すのかについて、添字を使います。つまり、 X_1 とすれば最

初の観測値, X_2 とすれば 2 番目の観測値, 以下同様に X_N までいきます。あるいは, 同じことをもう少し抽象的に表現するために, X_i で i 番目の観測値を指すことにします。表記法についてはっきりさせるために, 以下の表では `afl.margins` 変数にある 5 つの観測について, 数学的表記法と対応する実際の値の関係をリストアップしています。

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

オウケイ、では平均の式を書いてみましょう。伝統的に、平均を表すのに \bar{X} を使います。平均の計算は以下の式で表現できます。

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{N-1} + X_N}{N}$$

この式はまったく正しいのですが、ちょっとばかり長ったらしいので、総和記号である Σ を導入してこれを短縮しましょう^a ここでは最初の 5 つの観測について足しあわせをしたいわけですから、長い書き方ですと $X_1 + X_2 + X_3 + X_4 + X_5$ となります。ここで総和の記号を使ってこれを次のように短縮します。

$$\sum_{i=1}^5 X_i$$

文字通り、これは「1 から 5 までの全ての i について、 X_i の値を足し合わせる」と読みます。しかしその意味は基本的に「最初の 5 つの観測値を足す」、です。どちらにせよ、これは平均を使うための記号として使われ、次のように書きます。

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

正直なところ、この数学的な表記法が平均の概念を明確にするのに役立つとは思えません。実際には、私が言葉で言ったのと同じことを書き出しているだけです。すなわち、全ての数字を足しあわせて、足した項の数で割る、です。しかし詳細に書き込んだ本当の理由はこれではありません。私のゴールは、誰もがこの本を読むときに使われるであろう記号について、はっきりと理解しておいてもらうことがあります。 \bar{X} は平均、 Σ は総和、 X_i は i th 番目の観測値で N は観測の総数、ということをね。これらの記号は再利用されるので、みなさんがこれを使った式を「読む」ことができるよう、さらに「多くのものを足しあわせて別のもので割る」と言えるように理解してもらうことが重要なのです。

^a 総和に対して Σ を使うのは、勝手に決めたわけではありません。これはギリシア文字シグマの大文字で、アルファベットで言う S のアナロジーだからです。同様に、全ての積を示すための記号もあって、それは “products”(総積) と呼ばれるので文字としては Π を使います(ギリシアのパイの大文字で、これはアルファベットの P のアナロジーだからです)。

3.1.2 JASP での平均の計算

数学の話はここまで。計算してくれる魔法の箱はどうやって手に入れたらいいでしょうか？ 観測値の数が大きな数字になったら、コンピュータを使って計算させるのが何より簡単です。全てのデータを使って平均の計算をするために、JASP を使いましょう。最初のステップは ‘記述’ のボタンをクリックして、次に ‘記述統計’ をクリックしてください。それから変数 `afl.margins` をハイライト

させて、‘右矢印’をクリックしてそれを‘変数ボックス’に移します。するとすぐに画面の右側に表が現れます。そこには‘記述’についての情報があります。図 3.3 を見てください。

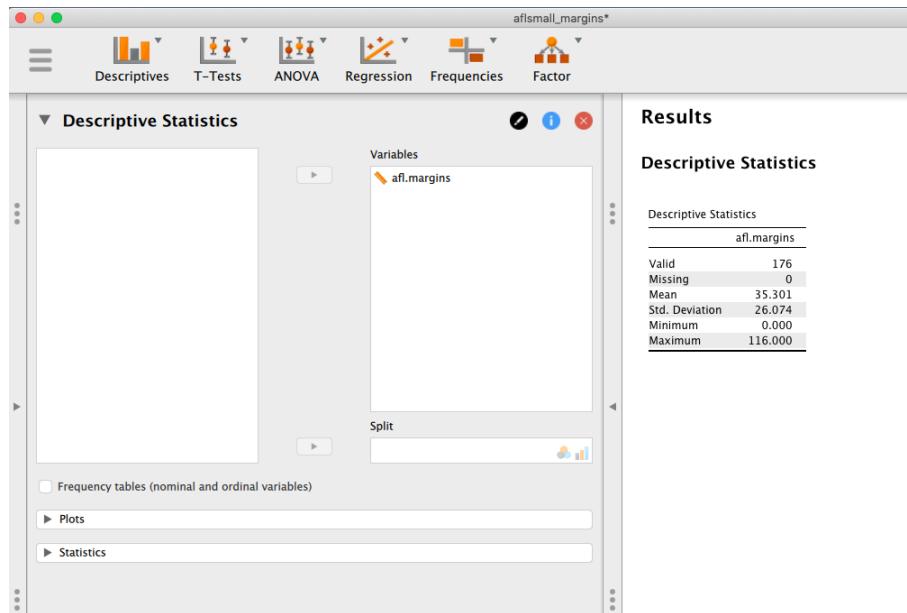


Figure3.3 AFL における 2010 年得点差データ (変数 `afl.margins`) のデフォルトで示される記述統計

図 3.3 に見て取れるように、変数 `afl.margins` の平均値は 35.301 です。他の情報として、観測度数の総数 ($N=176$) や、欠損値の数 (ありません)、変数の中央値、最小値、最大値も含まれています。

3.1.3 中央値

中心化傾向の二つ目の測度としてよく使われるのは、**中央値**です。この説明は平均よりも簡単です。変数セットの中央値というのは、ちょうど真ん中の値という意味です。AFL データの最初の 5 つの値、56,31,56,8,32 に興味があると思ってください。これらの数字の中央値を探すために、これを昇順に並べ替えます。

8, 31, **32**, 56, 56

見てみると、これら 5 つの観測値の中央値は 32 ですね。並べ替えたリストの真ん中にあるからです (より分かりやすくするために、太字にしました)。簡単なことです。でも 5 つでなくて 6 つの観測値に興味があったらどうしましょう? シーズン 6 番目のゲームが得点差 14 点だったとすると、並べ替

えリストは今や次のようにになります。

8, 14, **31**, **32**, 56, 56

そして真ん中の数字はふたつあって、31と32になります。中央値は、この二つの数字の平均値として定義されるので、31.5になります。前と同じで、数字がもっとたくさんあると人の手でやるのはとても難しくなります。実際には、もちろん、誰も真ん中の値を探すためにデータを並べ替えるなんてことはしません。コンピュータを使って、この面倒な作業をやらせるのです。JASPはお願いしたら中央値を出してくれます；単に‘統計’をクリックして、ドロップダウンメニューから‘中心化傾向’メニューの‘中央値’を選んでください。結果は自動的に中央値を含むものにアップデートされ、JASPは[afl.margins](#)変数の中央値が30.500であるとレポートしてくれます。

3.1.4 平均値か中央値か？その違いは？

平均値と中央値の計算方法を知ることは、このお話の一部に過ぎません。あなたはそれがデータの何についてものを言い、それらを使うときに何が仄めかされることになるのかを理解する必要があります。図3.4にそれを描いてみました。平均は、データセットの“重心”的なもので、中央値はデータの“真ん中の値”です。これが意味することは、あなたがこれらのどちらかを使うときに、データの種類が何であって、それで何をやろうとしているのかに関わってきます。ざっくりいうと、

- データが名義尺度水準であれば、平均値も中央値も使うべきではありません。平均値も中央値も数字が割り当てられた値に意味がある、という考え方には依存しているからです。
- データが順序尺度水準であれば、平均値よりも中央値を使う方が良いでしょう。中央値はあなたのデータの順序情報（すなわち、どの数字が大きいか）にだけ関わり、正確な数字には依存しないからです。これこそあなたのデータが順序尺度水準である状況でしょう。それに対して平均は、正確な量的値が観測対象に割り当てられているときに使われる所以、順序尺度データには適していないのです。
- 間隔尺度あるいは比率尺度水準のデータであれば、どちらでも一般的に受け入れられます。どちらを選ぶかは、あなたが何をしたいかによります。平均値はデータの全ての情報を使用します（あなたが大量のデータを持っているときには便利です）。が、極端な、外れ値には敏感です。

最後のパートを少し拡張しましょう。一つの結論として、平均値と中央値の間の体系的な違いは、ヒストグラムが非対称であるとき（歪んでいるとき；セクション3.3を参照）に現れます。これは図3.4に描かれています。中央値は（右図）、ヒストグラムの“ボディ”近くにありますが、平均値（左図）は“尻尾”（極端な値があるところ）に引っ張られています。わかりやすい例を示すために、ボブ（年収\$50,000）、ケイト（年収\$60,000）、ジェーン（年収\$65,000）が席についていると思ってください。テーブルの平均値は\$58,333で、中央値は\$60,000です。ここにビルが座ります。彼の年収は（\$100,000,000）です。年収の平均値は\$25,043,750に跳ね上がりますが、中央値は\$62,500にあが

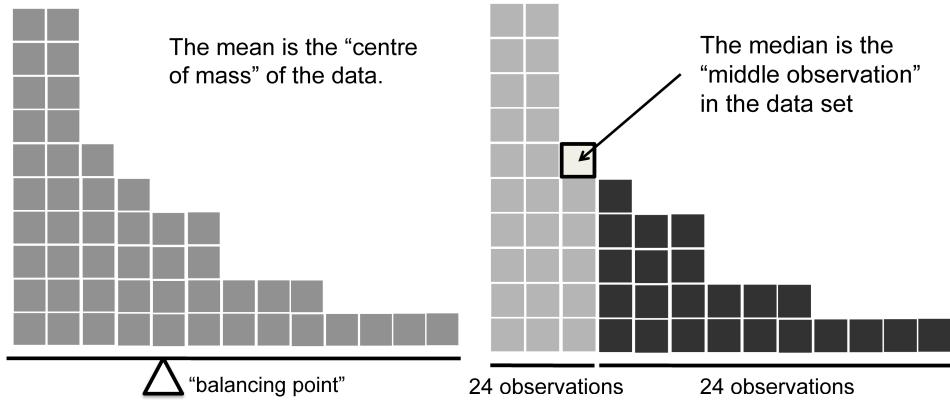


Figure 3.4 平均値と中央値の違いをどう解釈するかについてのイラスト。平均値は基本的にデータセットの“重心”です。データのヒストグラムが固体物だと考えたら、そのバランスを取る点(シーソーみたいに)が平均値です。それに対して、中央値は真ん中の観測で、それより小さいデータが半分、それより大きいデータが半分あるということです。

るだけです。席についている人の全体的な年収に興味があるおなら、平均が正しい答えになるでしょう。しかし典型的な年収の人が知りたいのであれば、中央値がより良い選択肢になるのです。

3.1.5 現実的な例

平均値と中央値の違いについて、何故注意を払うべきなのかの感覚を得るために、現実生活での例で考えてみましょう。私は科学的・統計的知識の足りないジャーナリストを馬鹿にする傾向があるのですが、信頼すべきところは信頼すべきだと思っています。これは2010年9月24日のABCニュース^{*2}になった、ある素晴らしい論文です。

コモンウェルス銀行の上級幹部がこの数週間、世界各地を訪問し、オーストラリアの住宅価格と所得に対する主要な価格の比率が、類似国と比較してどのように優れているかを示すプレゼンテーションを行いました。“住宅価格はこの5.6年、実質的に横ばい状態である”と銀行トレーディング部門のチーフエコノミスト Carig James は言っています。

これはおそらく、住宅ローンを抱えている人や、住宅ローンを希望している人、家賃を払っている人、オーストラリアの住宅市場でここ数年続いていることに全く気がついていない人にとっては、大きな驚きではないでしょうか。元の論文に戻ってみましょう。

CBA(コモンウェルス銀行のこと)は、グラフ、数字、国際比較などで住宅の運命が決まると信じている人と戦ってきました。プレゼンテーションの中には、オーストラリアの家賃は収入に比べて割高であるという議論を、銀行が否定しているとされています。オーストラリアにおいて、世帯主の価格に対する住宅価格は大都市において5.6、全国的には4.3であり、他の多くの先進国と同じぐらいであるとしています。また、サンフランシスコとニューヨークではこの比率は7,

^{*2}www.abc.net.au/news/stories/2010/09/24/3021480.htm

オークランドでは 6.7, バンクーバーでは 9.3 にもなります。

もっとびっくりなニュースです! だけど、この論文は次のように見立てています。

アナリストの多くは、これは銀行によってミスリーディングな図、比較がなされたからだと言います。CBA の資料 4 ページ目をみて、グラフや表の下に書いてある情報ソースをみたら、国際比較の追加的なソースがあることに気づくでしょう—人口動態学についての。コモンウェルス銀行が人口動態学の情報を使ってオーストラリアの住宅価格・収入比率の分析をしていたとすると、その実態は 5.6 とか 4.3 ではなく 9 近くになります。

うーむ、かなりの違いがありますね。一方では 9 といい、他方では 4-5 だ、と言っています。この違いを区分して、本当の値はこの間にあるんだとでもしたほうがよいでしょうか? 全く違います! 正しい答えと、間違った答えがあるような状態なのです。人口動態学は正しく、コモンウェルス銀行は間違っています。論文では次のように指摘しています。:

コモンウェルス銀行の住宅価格対収入の図には明らかな問題があり、平均年収と住宅価格の中央値を比較しているのです(人口動態学の図は収入の中央値と価格の中央値の比較をしているのに)。中央値は真ん中にある点で、極端に高いあるいは低い値を効率よくカットしますが、平均値は年収や資産価値については高所得者が含まれるので高くなる傾向があります。別の言い方をすれば、コモンウェルス銀行の図は Ralph Norris の数百万ドルにも及ぶ給料を収入が話に入れ、かれの(間違いなく)高価な住宅価格は図の中に入れないようにしているので、住宅価格はオーストラリアの中級ぐらいの年収と比較することになります。

これ以上いうことはありません。人口動態学的に計算した比率の方が正しいのです。銀行がやったやり方は間違っています。なぜ数字に得意なはずの銀行がこのような基本的なミスをしたのかというと… 彼らが何を考えていたのかは分からないので、ここまでにしましょう。しかしこの論文が以下の事実についての注意を促しています。関係があるかどうかわかりませんが。

オーストラリア最大の住宅業界牽引者であるコモンウェルス銀行は、住宅価格の上昇については最大級の興味を持っています。住宅ローンや多くの中小企業向けローンの担保として、オーストラリアの住宅の大部分を事実上所有しています。

むにやむにや。

3.1.6 最頻値

サンプルの最頻値は、とても単純です。それは最も頻度が多い値、なのです。AFL の別の変数を使ってこれを説明してみましょう。決勝で最も多くプレーしている選手は誰でしょう? `aflsmall_finalists` ファイルを開いて、`afl.finalists` 変数をみてみましょう。図 3.5 がそれです。この変数には全 400 チームの、1987 年から 2010 年までの間に開催された 200 回の決勝戦情報が載っています。

我々がやるべきことは、全 400 試合を読み通して、決勝戦リストに出てくるチームの名前を数え上げ、**度数分布表**を作ることです。しかしこれは頭を使わない退屈な作業で、まさにコンピュータが得意とするような作業ですね。だから JASP を使いましょう。「記述」の下にある「記述統計」の、`afl.finalists` 変数を選び「変数」ボックスに移し、「度数分布表」と書かれた小さなチェックボック

The screenshot shows the JASP interface with the title bar 'aflsmall_finalists'. The menu bar includes Descriptives, T-Tests, ANOVA, Regression, Frequencies, and Factor. A toolbar below the menu has icons for Descriptives, T-Tests, ANOVA, Regression, Frequencies, and Factor. The main window displays a table with 13 rows and 2 columns. The first column contains row numbers 1 through 13. The second column contains team names: Hawthorn, Melbourne, Carlton, Melbourne, Hawthorn, Carlton, Melbourne, Carlton, Hawthorn, Melbourne, Melbourne, Hawthorn, and Melbourne.

	afl.finalists
1	Hawthorn
2	Melbourne
3	Carlton
4	Melbourne
5	Hawthorn
6	Carlton
7	Melbourne
8	Carlton
9	Hawthorn
10	Melbourne
11	Melbourne
12	Hawthorn
13	Melbourne

Figure3.5 aflsmall_finalists.csv ファイルに修められた変数の JASP スクリーンショット

スをクリックします。すると図 3.6 のようなものが得られるでしょう。

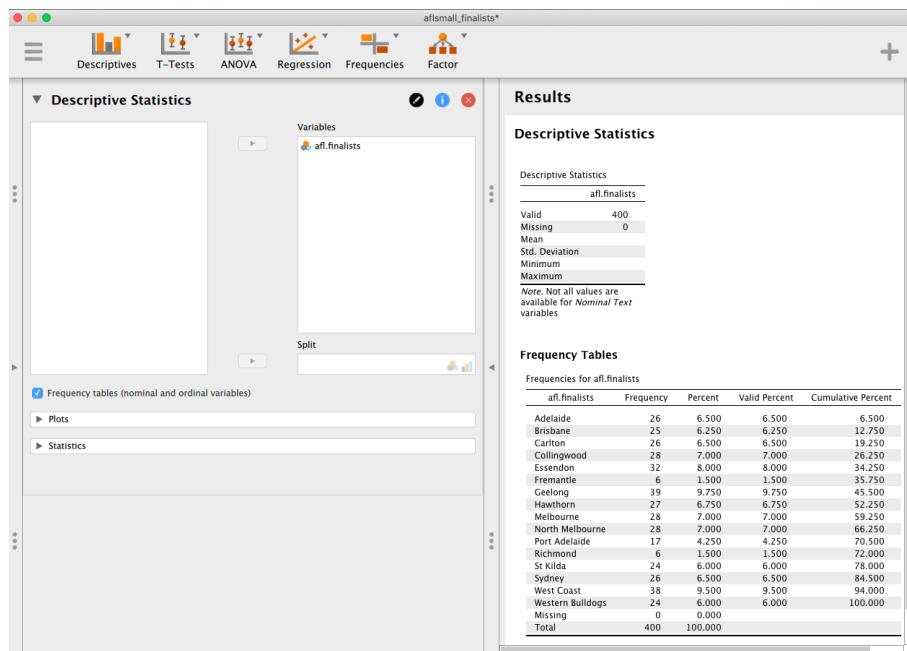


Figure3.6 afl.finalists 変数の度数分布表を示した JASP スクリーンショット

さて度数分布表を入手したわけですが、これをみると 24 年間ずっと、Geelong が他のどのチームよ

りも多く決勝に進んでいることがわかります。ですから `afl.finalists` データの最頻値は "Geelong" だということになります。Geelong(39 回決勝進出) が 1987 年から 2010 年の間で他のどのチームよりも多く決勝に進んでいるのです。また, '記述統計' の表では平均値, 中央値, 最大値, 最小値が計算されていないのも注目です。なぜなら `afl.finalists` 変数は名義的な文字変数であって, これらの値を計算する意味がないからです。

最後に最頻値に関するポイントをもう一つ。名義尺度のデータを持っていたら最頻値を計算するのが最もよくあるケースです。というのも, 平均や最頻値はこの種の変数には向いていないからですが, 順序, 間隔, 比率尺度水準の変数の最頻値を知りたいという時もあります。例えば, `afl.margins` 変数にもどってみましょう。この変数は明らかに比率尺度水準(もしピンとこないのなら, もう一度セクション ?? を読んでみてください)であり, あなたが知りたいのはこの中心に関する測度であれば平均値や中央値を求めるところです。しかしこんなことを考えてみてください: あなたの友達が賭けようぜと言ってきて, ランダムにフットボールのゲームを選ぶとします。誰がプレイするのかを知らずに, 正確な得失点差を推測しないといけないです。正しく当てられたら 50 ドルもらえます。でなければ 1 ドル失います。ほとんど正解に近かった, という残念賞はないものとします。正確に点差を推測しなければならないのです。この賭けをする時, 平均や中央値は全くあなたの役に立ちません。最頻値にかけるべきです。`afl.margins` 変数の最頻値を JASP で計算するには, データセットに戻って '記述' - '記述統計' 画面から, '統計量' と書いてあるセクションを拡大してください。'最頻値' のチェックボックスをクリックして, '記述統計量' テーブルにある最頻値をみます。図 3.7 にあるやつです。そうすると, 2010 年のデータでは 3 点差に賭けるべきだということがわかります。

3.2

変動性の指標

ここまで話してきた統計の話は, 中心化傾向に関するものでした。つまり, そこでの話はデータの "真ん中" とか "代表的な" 値についてでした。しかし, 中心化傾向は計算したい要約統計量の唯一の種類, というわけではありません。計算したい第二のものとして, データの **変動性** があります。つまり, どれぐらいデータが "散らばっているか"? とか, どれぐらい平均や中央値から観測値が "遠くにある" 傾向があるか? というものです。ここでは, データが間隔あるいは比率尺度水準で得られていると考えますから, `afl.margins` データを例に使い続けましょう。このデータを使うことで, 散らばりの指標としていくつかのものを示すことにし, その長所と短所も見ていくことにしましょう。

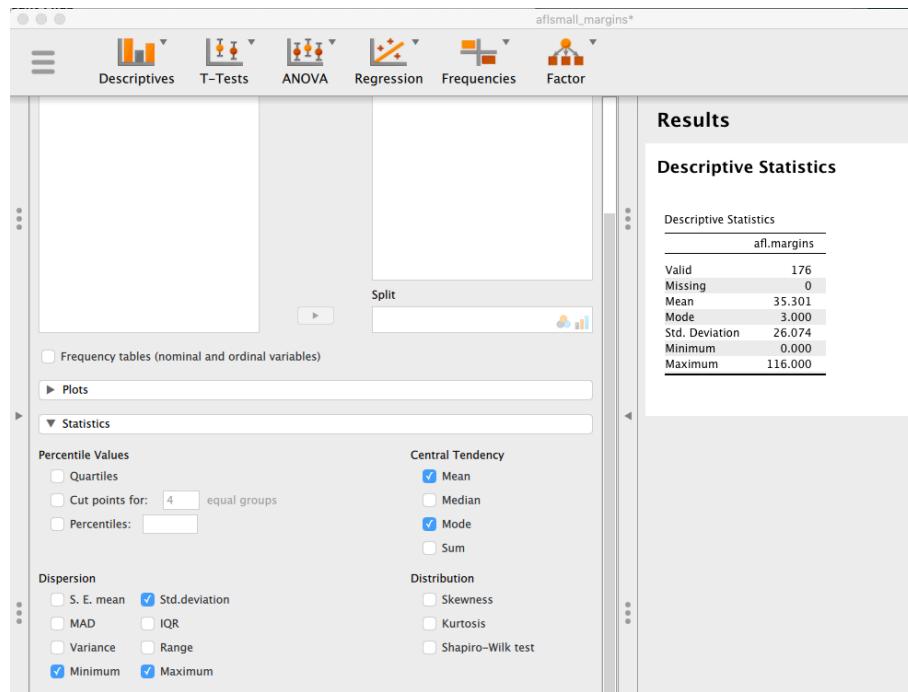


Figure3.7 afl.margins 変数の中央値を示す JASP 画面

3.2.1 範囲

変数の範囲はとてもシンプルなものです。最大値から最小値を引いたもののこと指します。AFL 得点差データの最大値は 116 で最小値は 0 でした。“変動”を表す量として範囲は最も単純なものです、最も悪いものもあります。要約統計量を頑健なものにするために、平均について議論していたことを思い出してください。もしデータセットの中に一つ二つ変な値があると、我々の統計量はそ うしたデータに角に影響されないようにしたいところです。

例えば、変数が極端な外れ値を持っていたとします。

$$-100, 2, 3, 4, 5, 6, 7, 8, 9, 10$$

範囲が頑健な値ではないことは明らかですよね。変数の範囲は 110 になりますが、外れ値を除くとたったの 8 になります。

3.2.2 四分位範囲

四分位範囲 (interquartile range,IQR) は範囲に似ていますが、最大値と最小値の差を使うのではなく、25 パーセンタイルと 75 パーセンタイルの差を使います。パーセンタイルをまだ知らないかも

されませんが、データの 10 パーセンタイルというのはある点 x よりも小さいのがデータの 10% になるような点 x のこと、という意味です。実は、既にこの考え方は出てきています。データの中央値とは、50 パーセンタイルのことですから！JASP では、簡単に 25,50,75 パーセンタイルを見つけることができます。‘記述’の‘記述統計’から‘統計量’の画面にある‘四分位’チェックボックスをクリックするだけです。

Descriptive Statistics

Descriptive Statistics	
<hr/> afl.margins <hr/>	
Valid	176
Missing	0
Mean	35.301
Mode	3.000
Std. Deviation	26.074
Minimum	0.000
Maximum	116.000
25th percentile	12.250
50th percentile	30.500
75th percentile	51.500

Figure3.8 afl.margins 変数の四分位を示す JASP のスクリーンショット

驚くには值しませんが、図 3.8において 50 パーセンタイルは中央値と同じになっています。そして、 $50.50 - 12.75 = 37.75$ ですから、2010 年の AFL 得点差データの四分位範囲は 37.75 ということになります。範囲の解釈は明らかですが、IQR の解釈の仕方はそこまで明らかだというわけではないですね。これは次のように考えるのが最も単純な方法でしょう。すなわち、四分位範囲はデータの“中半分”の範囲だというものです。つまり、データの一つの四分位が 25 パーセンタイル点で、もう一つの点が 75 パーセンタイル点ですから、この二つの間にデータの“中半分”が位置していることになります。IQR はこの中半分をカバーする範囲なのです。

3.2.3 平均絶対偏差

二つの尺度、範囲と四分位範囲をみてきましたが、どちらもデータのパーセンタイルをみて、データの散らばりを測ろうとするアイデアに基づいています。しかし、これだけがこの問題唯一の解決策ではありません。別のアプローチとして、意味のある参照点（ふつう平均値や中央値ですが）を選び、その参照点からの“典型的な”偏差を報告する、ということがあります。“典型的な”偏

差、というのは何を意味しているでしょう？普通これは偏差の平均値や中央値を指します。実際、ここからは二つの尺度が導かれます。“平均絶対偏差”(平均値からの)と，“中央値絶対偏差”(中央値からの)，です。私がこれまでみてきたところ、中央値に基づく尺度が統計的に使われているようで、そちらの方が優れているようです。しかし正直に言って、心理学でこれらが使われてきたのをあまりみたことがありません。平均に基づく尺度の方が、心理学ではよく出てきます。このセッションでは前者について最初説明しますが、その後で2番目についても触れていきます。

前のパラグラフではちょっと抽象的だったかもしれません、平均からの**平均絶対偏差**についてもう少しゆっくりみていきましょう。この尺度が便利なことの一つに、この名前が実際にどうやって計算するのかを表している、ということがあります。AFLの得点差データについて、もう一度最初の5ゲームをみてみると、得点差は56, 31, 56, 8, 32でしたね。ここで計算はある参照点(今回は平均)からの偏差を見るものですから、最初にするべきことは平均つまり \bar{X} を計算することです。最初の5ケースでは、平均は $\bar{X} = 36.6$ になりました。次のステップは各観測値、 X_i を偏差のスコアに変換することです。これは観測値 X_i と平均 \bar{X} の差を計算することができます。つまり、偏差スコアの定義は $X_i - \bar{X}$ となるのです。今回のサンプルにおける最初の観測値は、 $56 - 36.6 = 19.4$ になります。オーケイ、十分シンプルですね。このプロセス、次のステップはこれらの偏差を絶対偏差にすることです。これは負の値を正の値にすることでできます。数学的には-3の絶対値を $|-3|$ と書き、 $|-3| = 3$ とします。この絶対値を使うのは、平均よりも高かったのか低かったのかを気にしないということであり、興味は平均にどれくらい近かったのかというだけだということです。このプロセスができるだけ明白にするために、下の表では、5つの観測値すべてについてこれらの計算を示しています。

用語:	どのゲームで	値	平均偏差	絶対偏差
表記:	i	X_i	$X_i - \bar{X}$	$ X_i - \bar{X} $
	1	56	19.4	19.4
	2	31	-5.6	5.6
	3	56	19.4	19.4
	4	8	-28.6	28.6
	5	32	-4.6	4.6

さてデータセットの各観測値について絶対偏差を計算できたので、これらのスコアの平均を計算しましょう。次のようになります。

$$\frac{19.4 + 5.6 + 19.4 + 28.6 + 4.6}{5} = 15.52$$

はいおしまい。これら5つのスコアについて、平均絶対偏差は15.52でした。

ところで、この簡単な例はこれでおしまいですが、少し話が残っています。まず、数学的な定式化をしておくべきです。しかしこれをしようとすると、平均絶対偏差についての数学的表記が必要です。腹立たしいことに、“平均絶対偏差”と“中央値絶対偏差”はどちらも同じ頭文字 (MAD) ので、曖昧になってしまいますから、平均絶対偏差に何か別の表現を考えないといけないでしょう。やれやれ。*average absolute deviation* を短くして、AAD とすることにしましょう。これでもいくらか曖昧な表記ですが、計算は次のように書くことができます。

$$\text{aad}(X) = \frac{1}{N} \sum_{i=1}^N |X_i - \bar{X}|$$

3.2.4 分散

平均絶対偏差は使いでありますが、変動の尺度として最適というわけではありません。純粋に数学的な観点からは、絶対偏差よりも二乗した偏差の方が好ましい理由があります。これを使うと分散とよばれる尺度を手に入れることになります。それは本当にステキな統計的特徴を持っているのですが、それは横に置いておくとして^{*3}、今から取り上げるとても大きな心理学的欠陥も持っていることを説明します。データセット X の分散は $\text{Var}(X)$ と表記されますが、もっと一般的には s^2 と書きます（その理由はすぐにわかります）。

観測されたデータセットの分散を計算する式は次の通りです。

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

ご覧の通り、基本的には平均絶対偏差で使ったものと同じ形をしていますが、違うのは“絶対偏差”的かわりに“偏差平方”を使っているところです。このため、分散は“平均偏差平方”とも言われます。

さて、基本的な概念を手に入れましたので、具体例でみてみましょう。もう一度、AFL ゲームの最初の 5 つのデータを使います。前回同じアプローチをした時に習って、次のような表にしてみました。

^{*3}えーっと、ちょっとだけ何が最高にクールなのか、“クール”の定義をしてから説明してみましょう。分散は加算的なのです。その意味はこんな感じです。私が二つの変数 X と Y を持っていて、それらの分散がそれぞれ $\text{Var}(X)$ と $\text{Var}(Y)$ だとしましょう。ここで新しい変数 Z を、二つの和、 $Z = X + Y$ で定義したとします。そうすると、 Z の分散は $\text{Var}(X) + \text{Var}(Y)$ になるのです。これがとても便利な特徴なのですが、このセクションで私が説明しようとする他の尺度にはないものなのです。

用語: 表記:	どのゲームで 表記: <i>i</i>	値 X_i	平均偏差 $X_i - \bar{X}$	偏差平方 $(X_i - \bar{X})^2$
	1	56	19.4	376.36
	2	31	-5.6	31.36
	3	56	19.4	376.36
	4	8	-28.6	817.96
	5	32	-4.6	21.16

最後の列には全ての偏差平方が入っていますので、この平均を取れば良いのです。手計算する、つまり電卓を使うと、この分散の値が 324.64 であることがわかります。興奮してきたでしょう? このとき、多分あなたの考えに火がついた問題(すなわち、324.64 の分散って本当に平均なのか?)は横に置いて、JASP でこれをどう計算するかみてみましょう。というのも、これで奇妙なことが明らかになるからです。

まず最初の 5 行だけを含んだ新しいデータを読み込みます。ファイル `aflsmall_margins_first5.csv` を読み込んでください。次に‘統計’メニューの‘記述’-‘記述統計’をクリックし、‘分散’チェックボックスをクリックします(‘ばらつき’グループの中にあるのがわかると思います)。手計算した値(324.64)と同じ数字になりましたか? いや、ちょっと待って、あなたは全く違う答えを手にしたではありませんか(405.800)!おかしいなあ。JASP は壊れてるの? タイポですか? 何が起こってる?

起こった通りのこと、答えは no です。タイポではなく、JASP が間違っているわけでもありません。現に、JASP がここで何をしているのかを説明するのはとても簡単なのですが、JASP がなぜそれをしたのか、というのはちょっと説明に苦労します。ですから“何が起ったのか”から始めましょう。JASP は上で示したのとは少し違う数式を評価したのです。偏差平方の平均を計算したのではありません。平均はデータ点の数 N で割りますが、JASP は $N - 1$ で割ったのです。

言い換えると、JASP は次の式を使って計算したのです。

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

これが何をやったかです。本当に知りたいのは、なぜ JASP が N ではなく $N - 1$ で割ったのか、ですよね。結局のところ、分散は偏差平方の平均なのですよね? だったら N で割るべきじゃないか、サンプルの実際の観測数でね。全くその通りです。しかし、第 6 章で論じるように、“サンプルを記述する”ことと“サンプルのもとになった母集団を推測すること”とのあいだにはちょっとした違いがあるのです。ここまででは、この差の区別をしてきませんでした。あなたが表現したいのがサンプルなのか、母集団の推測するものなのかどうかにかかわらず、平均は同じように計算できたのです。しかし分散や標準偏差、そのほかの尺度ではそうならないのです。私が最初に説明したこと(つまり、

N で割ることによる実際の平均)は、標本の分散を計算することを想定したものでした。しかしほとんどの場合、標本そのものに興味をもってるわけではないでしょう。むしろ、その標本は世界について何かを伝えるために存在しているはずです。そうであれば、あなたが実際に計算したいのは“標本統計量”ではなくて、“母集団の母数”を推定するためのものになるはずです。しかしこの話は、少し先走りすぎています。今は、JASP がすることをただ信じて、第 6 章で推定について論じるときまでこの問題をおいておくことにしましょう。

最後にもう一つ。このセクションはちょっとした推理小説のようになっていました。先ほど分散の式を示し、JASP では“ $N - 1$ ”でやっていること、そしてなぜそうするのかのヒントを書きましたが、最も大事なことは触れていなかったのです。みなさんは分散をどういうものだと理解していますか?記述統計は記述することだけを目的としていますが、今のところ分散は意味不明な数字しかありません。残念なことに、分散の解釈について人間味のない説明しかできない理由は、それがそもそも人間味のないものだからです。これが分散について最も深刻な問題点です。分散は本当は変動を表現する基本的な量であるというある種の美しい数学的特性はあるのですが、現実的に他者との会話に使いたいと思うときには全く役に立たないです。分散は元の変数に関しては全く意味のない数字になります! 全ての数字は二乗されてしまうので、それは何も意味しないことになるのです。これは大問題だ。例えば、以前示した表について言うと、ゲーム 1 における点差は“376.36 ポイントの二乗分、平均より高い”と言うことになります。これはまったく馬鹿馬鹿しい表現ではないですか。計算した分散の 324.64 の時も同じことがいえます。多くのフットボールゲームを見てきましたが、誰も“ポイントの二乗分”なんて言ってるのを聞いたことがありません。これは測定の実際の単位ではなく、分散は意味のない単位を持っているので、人間にとて全く意味のないことになるのです。

3.2.5 標準偏差

オーケイ、分散を使う理由は分かってもらえたとしましょう。説明はしませんが、分散は数学的に良い特性持っていますからね。でもあなたが人間で、ロボットでないなら、データと同じ単位を持っている(つまり二乗した値ではないもの)尺度を使う方がいいと思うでしょう。じゃあどうしましょう? 答えは簡単です! 分散の平方根を取れば良くって、これは**標準偏差**として知られています。“偏差平方平均の根”，つまり RMSD とも呼ばれます。これで問題がスッキリ解決しました。だれも“分散は 324.68 ポイントの二乗”ということの意味を理解することはできませんが，“標準偏差 18.01 ポイント”は簡単に理解できます。元の単位で表現されているんですから。

標準偏差は分散の正の平方根に等しいので、次の式を見ても驚かないと思います。

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

JASP では、‘分散’のチェックボックスと同じセクションに‘標準偏差’のチェックボックスもあります。図 3.8 をみると、JASP は `afl.margins` の標準偏差を `26.074` と答えてくれています。標準偏差はとてもよく使われる所以、チェックするのがデフォルトになっていますが、あなた自身で選んでみてください!!

しかし、分散についての議論でお気づきかもしれません、JASP は実際にはこれとちょっと違ったやり方で計算します。分散を見るだけなら、JASP は N ではなく $N - 1$ で割る方で計算するのです。

第 6 章で再びこのトピックに触れるとき意味がわかると思いますが、この新しい量を $\hat{\sigma}$ (“シグマ・ハット”と読みます) とし、次のように定式化します。

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

標準偏差を解釈するのも少し複雑です。標準偏差は分散から導出されています。そして分散は人にとてあまり意味のない量になっていますから、標準偏差は単純な解釈では済みません。結果的に、私たちのほとんどはちょっとした経験則を用いています。一般的に、平均から標準偏差 1 つぶん離れたところにデータの 68% が含まれ、データの 95% は平均から標準偏差 2 つ分離れたところに 99.7% が、平均から標準偏差 3 つ分離れたところに含まれる、ということが期待できます。このルールはほとんどの場合うまく当てはまりますが、多少の例外はあります。これがちゃんと計算できるのはヒストグラムが対称的で“ベル型”になっているという仮定に基づいています^{*4}。図 3.2 にある AFL の得点差ヒストグラムを見ると、この経験則は私たちのデータに合っているとは思えません! しかし大まかに合っているのです。AFL データの 65.3% が実際に平均から 1 標準偏差の範囲にあります。This is shown visually in Figure 3.9. このことは、図 3.9 に視覚的に示されています。

3.2.6 どの尺度を使いましょうか?

いくつかの範囲についての尺度を紹介してきました。範囲、IQR、平均絶対偏差、分散、標準偏差

^{*4}厳密にいうと、この仮定はデータが正規分布にしたがっているということで、この重要な概念については第 5 章で議論することになります。またこのことは本書で何度も何度も出てきます。

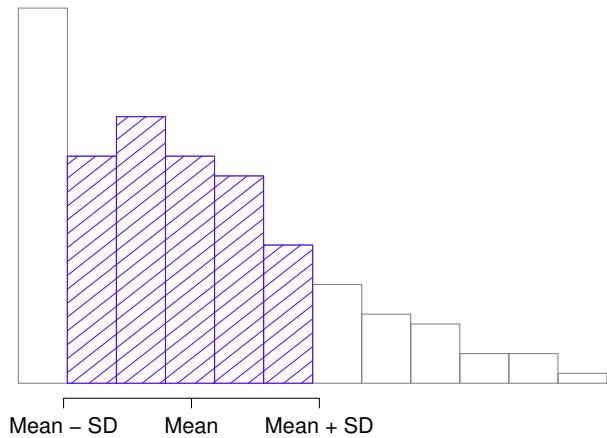


Figure3.9 AFL 得点差データについての標準偏差を描いたもの。色がついているヒストグラムの箇所は平均から 1 標準偏差のなかに入ったデータの数を表しています。今回は 65.3% のデータセットがこの範囲内に入り、次のメイントピックスである“約 68%”のルールに近い結果になっています。

です。そしてその長所と短所についてもみてきました。簡単にまとめておきましょう。

- 範囲 データのちらばり全体を見ます。外れ値に弱く、データの極端な部分を見たいという理由がない場合はあまり使われることはありません。
- 四分位範囲 データの“真ん中あたり”がある場所を教えてくれます。多少、外れ値に強くて中央値を含んでいます。これはよく使われます。
- 平均絶対偏差平均から観測度数が“平均的に”どれくらい離れているかを教えてくれます。解釈しやすいのですが、いくつかの小さな問題点があって（ここでは触れていませんが）、そのせいで統計家は標準偏差ほど魅力を感じていません。時々使われますが、それほど頻度はありません。
- 分散 平均偏差の二乗の平均です。数学的にはエレガントで、平均周りの散らばりを描写するにはたぶん“正しい”方法なのですが、データと同じ単位を使っていないので意味不明な数字になります。数学的なツール以外の用途はほとんどありませんが、非常に多くの統計技法の中に“埋もれて”います。
- 標準偏差分散の平方根です。これは数学的にも非常にエレガントで、データと同じ単位で表現されていますから、解釈も簡単です。平均が中心化傾向の尺度として使われる時は、これが基本です。散らばりの尺度の中で最もポピュラーなものになります。

まとめると、IQR と標準偏差が簡単で、データのばらつきを報告するのに最もよく使われる二大尺度、ということになります。しかし他のものが使われることもあります。この本に載せたのは、わずかではありますがみなさんがどこかで出会うかもしれませんからです。

3.3

歪度と尖度

みなさんが心理学の文献で見かけるかもしれませんる記述統計量が、あと二つあります。歪度と尖度です。実践上はどちらもこれまで話してきた中心化傾向や変動性の尺度ほど、使われるものではありません。歪度はちょっと大事なので見かけることはあるかもしれません、私は科学的レポートの中で尖度を目にしたことはありません。

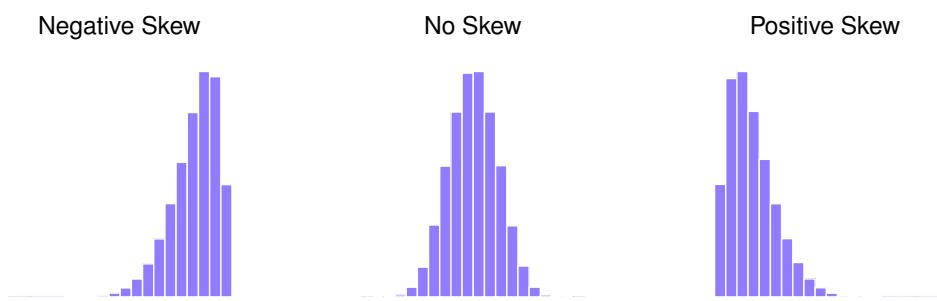


Figure 3.10 歪度のイメージ。左側は歪度が負 (歪度 = -.93), 真ん中は歪みなし (実際ほとんどありません。歪度 = -.006), そして右が正の歪度 (歪度 = -.006) をもつデータです。

.....

歪度の方が面白いので、こちらから話を始めましょう。歪度は基本的に非対称性の尺度で、図を書いてみれば理解は簡単です。図 3.10 にあるように、データに極端に小さな値（下の裾が上の裾よりも“長い”）を持っていて、極端に大きな値はそれほど持っていない（左図）場合、このデータは負の歪度をもつといいます。一方、極端に大きな値が小さい値より大きく多くあるようであれば（右図）、このデータは正の歪度をもつといいます。これが歪度の背後にある考え方です。平均よりも大きな値が相対的に多くあれば、分布は正、すなわち右に歪んでおり、裾も右に寄っています。負、すなわち左への歪みはその逆です。対称的な分布をしていれば、歪み度は 0 です。正に歪んだ分布の歪度は正の値であり、負の値は負の歪み分布だと言えます。

データセットの歪みについての定式化は次のとおりです。

$$\text{skewness}(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^N (X_i - \bar{X})^3$$

ここで N は観測度数の数であり、 \bar{X} は標本平均、 $\hat{\sigma}$ は標準偏差（ただし “ $N - 1$ で割ったバージョン”）です。

ありがたいことに、JASP で歪度の計算することができます。'記述' - '記述統計量' の下にある '統計量' チェックボックスのオプションがそれです。変数 `afl.margins` について、その歪度を計算すると [0.780](#) です。この歪度の推定値を歪度の標準誤差で割れば、このデータがどれほど歪んでいるかの指標を得ることができます。経験的に行って、小さいサンプルでは ($N < 50$)、この値が 2 以下であればそれほど歪みは大きくなく、2 以上であればデータが統計的な分析をするに許される限界を超えて歪んでいる、と考えるのが目安です。これは経験則に過ぎず、この解釈にはっきりした共通見解があるわけではないことに注意してください。ということで、この分析をすると AFL の得点差データはちょっと歪んでいることになります ($0.780 / 0.183 = 4.262$ で、これは明らかに 2 より大きいです)。

時々つかわれる最後の尺度は、実際に使われることは非常に稀なのですが、データセットの尖度です。簡単にいえば、尖度は“尖っているかどうか”的尺度で、図 [3.11](#) にその状況を示しています。慣例によって、“正規分布”(黒い線) は尖度ゼロであり、データセットの尖り具合はこのカーブに比べて相対的に評価されます。

この図にあるように、左のデータはそれほど尖っておらず、尖度は負でこのデータは緩く尖った *platykurtic* データだと言われます右図はとても尖っており、尖度は正でこのデータは尖度の大きい *leptokurtic* データだと言われます。一方、真ん中のデータはちょうどいいぐらいの尖度で、これは中程度の尖度 *mesokurtic* と呼ばれ、尖度はゼロです。下の表にこれをまとめました。

一般的な言い方	専門的な言い方	尖度の値
“かなりフラット”	platykurtic	負
“ちょうどいいぐらい”	mesokurtic	ゼロ
“とても尖っている”	leptokurtic	正

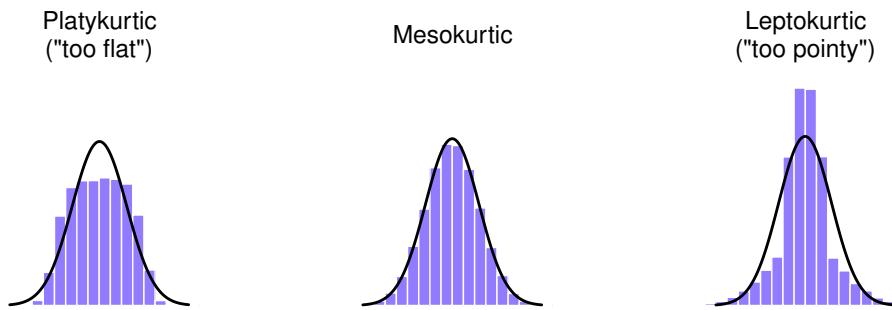


Figure 3.11 尖度の図。左側は“緩く尖った”データセット（尖度 = -.95）であり、これが意味するのはこのデータセットは“かなりフラット”だということです。真ん中の図は“中程度の尖り”をもったデータセット（尖度はほとんど 0）であり、これが意味するのはこのデータの尖度がちょうどいい感じであるということです。最後に、右側の図ですが、“尖度の大きい”データセット（尖度 = 2.12）であり、このデータセットは“とても尖っています”。尖度は正規分布（黒い線）と比べて評価されていることに注意してください。

尖度の式は既に見た分散や歪度の式とかなり似ています。分散が偏差の二乗、歪度が偏差の三乗であったのに対し、尖度は四乗になっています。^a

$$\text{kurtosis}(X) = \frac{1}{N\hat{\sigma}^4} \sum_{i=1}^N (X_i - \bar{X})^4 - 3.$$

^aこの“-3”については正規分布の尖度がゼロになるように統計家が付け加えたものです。“-3”を式の最後に引っ付けておくのはちょっと馬鹿みたいですが、こうすることの数学的な理由があるのです。

大事なのは、JASP で尖度を計算するには歪度の下のチェックボックスをクリックするだけだということで、そうすると尖度の値 **0.101** がその標準誤差 **0.364** と共に表示されます。歪度をその標準誤差で割ったのと同じように計算すると、この値は 2 より小さい (**0.101 / 0.364 = 0.277**) ことがわかります。これは AFL の得点差データの尖度がちょうどいいぐらいだったことを意味しています。

3.4 _____

グループごとの記述統計

よくあることのひとつとして、記述統計量があるグループ変数ごとに分割してみたいと思うことがあります。JASP ではすごく簡単にできます。例えば、ある `clin.trial` データについて、`therapy` のタイプごとに記述統計量を見たいなと思ったとしましょう。これは今まで見せていない、新しいデータセットです。このデータセットは `clinicaltrial.csv` ファイルにあって、第 ?? 章でよく使うようになります（このデータの詳細についてはその時に説明します）。読み込んで、見てみましょう。

	ID	drug	therapy	mood.gain	
1	1	placebo	no.therapy	0.5	
2	2	placebo	no.therapy	0.3	
3	3	placebo	no.therapy	0.1	
4	4	anxifree	no.therapy	0.6	
5	5	anxifree	no.therapy	0.4	
6	6	anxifree	no.therapy	0.2	
7	7	joyzepam	no.therapy	1.4	
8	8	joyzepam	no.therapy	1.7	
9	9	joyzepam	no.therapy	1.3	
10	10	placebo	CBT	0.6	
11	11	placebo	CBT	0.9	
12	12	placebo	CBT	0.3	
13	13	anxifree	CBT	1.1	
14	14	anxifree	CBT	0.8	
15	15	anxifree	CBT	1.2	
16	16	joyzepam	CBT	1.8	
17	17	joyzepam	CBT	1.3	
18	18	joyzepam	CBT	1.4	

Figure3.12 `clinicaltrial.csv` ファイルにある変数を写した JASP スクリーンショット

三つのドラッグがあるのがわかりますね。プラセボと、“anxifree”と“joyzepam”と呼ばれるものです。そしてそれぞれに 6 人割り当てられています。そして 9 人が認知行動療法 (CBT) を受けています。

て、9人が心理療法は何も受けていない状態です。そして `mood.gain` 変数の‘記述’をみてみると、ほとんどの人が気分の向上(平均 = 0.88)を示していますが、この尺度が何なのかわからないまでは、それ以上のことは言えません。でも、それはそれでわるくないのです。全体的には何か勉強になった気になります。

さて、さらに他の記述統計量を見て行きましょう。こんどはセラピーのタイプごとに分けて。JASPで‘統計量’オプションから標準偏差、歪度、尖度にチェックを入れます。同時に、`therapy` 変数を‘分割’ボックスに入れます。すると図 3.13 のような結果が得られます。

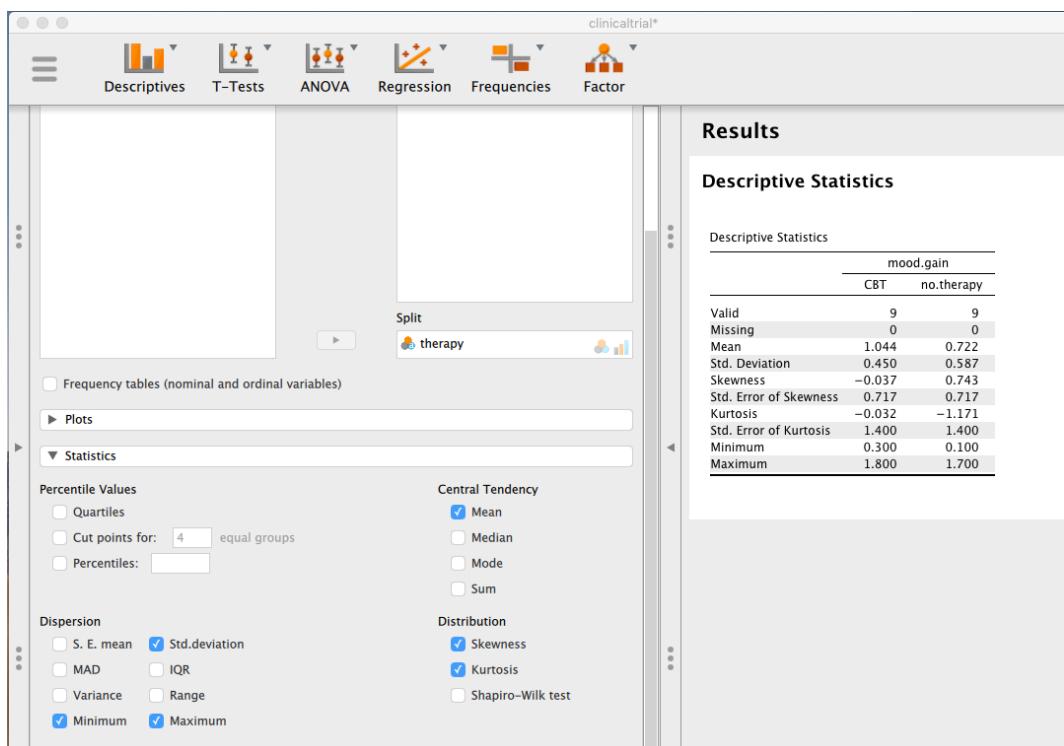


Figure 3.13 セラピータイプごとに分割した記述統計量を示した JASP のスクリーンショット

3.5

標準得点

私の友人が“不機嫌さ”を測定するための新しい質問紙を作ろうとしているとしましょう。この調査票は 50 の質問からなり、不機嫌かどうかについて答えるものとします。大きなサンプルをとって

(仮に百万人ぐらいとったとしましょう!), このデータが正規分布しており, 50 問中 17 点が平均不機嫌スコアで, 標準偏差が 5 だとしましょう。これに比べて, 私の得点は 50 問中 35 点だったとします。私はどれぐらい不機嫌なんでしょうね? これについて考える一つの方法は, 私は 35/50 が不機嫌なのだから, 70% ぐらい不機嫌だと考えることです。しかしちょっと考えてみれば, おかしい気もしますよね。もし私の友人が, その質問紙を少し違った捉え方で答えていたとしたら, その問い合わせ本当に問うていることに比べて, 全体的な分布が簡単に上がったり下がったりしてしまいます。ですから, 私が 70% 不機嫌だというのは, 調査票の質問セットに応じて変わることになります。とても良い質問項目であったとしても, これではあまり意味のある表現にはなりません。

これについての良いやり方の一つは, 私の不機嫌の程度を周りの人と比べることです。驚くべきことに, 私の友人は 1,000,000 人のサンプルを持っていて, その中でたった 159 人だけが私と同じ程度の不機嫌さ (本当ははそなことありませんよ) であれば, 私はトップ 0.016% の不機嫌度ということになります。このほうが, ロウデータを解釈しようとする時にはより意味があるのではないでしょうか。この考え方は, 私の不機嫌さの程度を人の全体的な不機嫌分布にあわせて記述しようとするものであり, 標準化がしようとしているのはまさにこれなのです。これを正しくやる方法の一つは, さっきやって見せたように, パーセンタイルで表現することです。しかし問題があるのは, この方法だと “トップが寂しい” ということです。私の友人が集めたサンプルが 1000 人に過ぎなかったとしましょう (これでもまだ新しい質問紙を検証するためには大きいサンプルですが)。そして今回, 平均が 50 問中 16 点で標準偏差が 5 だったとします。問題は, このサンプルでは私と同じぐらいの不機嫌度を持っている人が一人もいないということです。

しかし, 全てが失われたわけではありません。もう一つのアプローチとして, 私の不機嫌スコアを **標準スコア** に変換するのです。これは z -スコアとも言われています。標準スコアは私の不機嫌スコアが平均から標準偏差いくつ分上にあるかを表すのです。これを “数学っぽく” いうと, 標準偏差は次のように計算できます。：

$$\text{標準スコア} = \frac{\text{ロースコア} - \text{平均}}{\text{標準偏差}}$$

実際数学的には, z スコアについての式は次のようになります。

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

不機嫌さデータに戻っていようと, ダニーの生の不機嫌さデータを標準化された不機嫌スコアに変換することができます。

$$z = \frac{35 - 17}{5} = 3.6$$

この値を解釈するときに, セクション 3.2.5 で触れた, 平均から 3 標準偏差範囲にだいたい 99.7%

が入るという、概算を思い出してください。ですから、私の不機嫌さスコアを z スコアにして 3.6 になったということは、実際私はかなり不機嫌状態にあるということです。実際この推論からいくと、私は全体の 99.98% の人よりも不愉快なのです。そうですよね。

ロースコアをより大きな母集団に広げて解釈することを許すとするなら（そしてそれによって任意の尺度の変数を意味のあるものにするなら）、標準スコアは第二の便利な機能を持っていると言えます。標準スコアはロースコアができないような状況でも互いに比較することができます。たとえば、私の友人が 24 項目からなる外向性を測る別の質問紙を持っていたとしましょう。この尺度が全体的に、平均が 13 で標準偏差 4 であり、私のスコアが 2 だとたつとしましょう。想像の通り、私の外向性のロースコア 2 を、不機嫌さ質問紙のロースコア 35 と比較するのは意味がありません。この二つの変数のロースコアは基本的に違うもので、いわばりんごとオレンジを比較するようなものです。

標準スコアではどうでしょう？ これはちょっと事情が違います。標準スコアを計算すると、不機嫌さは $z = (35 - 17)/5 = 3.6$ 、外向性は $z = (2 - 13)/4 = -2.75$ となります。この二つの数字は相互に比較することができます^{*5}私はほとんどの人の中では外向性が低く ($z = -2.75$)、不機嫌さが高い ($z = 3.6$) のです。しかし私のハズレ具合は外向性よりも不機嫌さの方が大きいといえます。3.6 が 2.75 よりも大きな数字だからです。それぞれの標準化スコアはその観測値がその母集団においてどのあたりに落ちるのかを示すので、全く異なる変数についても標準スコア同士を比較することができるのです。

3.6 _____

要約

基礎統計量を計算することは、あなたが実際にデータを取ったとき真っ先にすべきことの一つであり、記述統計量は推測統計よりも単純で理解しやすいので、他の統計の教科書と同じように私も記述統計から説明しました。この章では、以下のトピックスについて議論しました。

- 中心化傾向の指標 一般的に、中心化傾向はデータがどのあたりにあるのか教えてくれます。典型的に報告される指標は次の三つでしょう；平均値、中央値、最頻値です（セクション Section 3.1）。
- 変動の指標 それに対して、変動の指標はデータがどのように“散らばっているか”を教えてくれます。鍵になる指標としては、次のものがあるでしょう；範囲、標準偏差、四分位範囲です（セクション Section 3.2）。

^{*5}いくつかの注意は必要です。変数 A についての 1 標準偏差が、変数 B の 1 標準偏差と“ある意味”対応しているとは言えないからです。二つの変数に関する z スコアが意味のある比較ができるかどうかを決めるには、常識をはたらかせねばなりません。

- 歪みと尖りの指標 変数の分布が非対称さの指標 (歪度) と、尖り具合 (尖度) もみてきました (セクション 3.3)。
- JASP で群ごとに変数の要約をする この本では JASP でデータ分析をすることに焦点化していますから、異なるサブグループそれぞれについて記述統計量を計算するにはどうするかについても触れました (セクション 3.4)。
- 標準化スコア z -スコアはちょっと変わった野獣です。これは記述統計量とはちょっと違いますし、推測統計の話でもありません。これについてはセクション 3.5 で触れました。この章も理解してもらえたと思います。また後で触れることになります。

次の章では、どうやって絵を描くのかについての話題に移りたいと思います! 誰だって可愛い絵が好きですもんね? しかしその前に、重要な点を抑えておきたいと思います。統計の伝統的な入門コースは、記述統計について小さな配分しかせず、1,2回授業で触れる程度です。授業時間のほとんどの時間は、推測統計学に使われます。というのも、そこが本当に大変なところだからです。それはそれで意味があるのですが、良い記述統計量を選択するという、日々の重要な実践を覆い隠してしまいます。このことを覚えておいて欲しいのです…

3.6.1 エピローグ: 良い記述統計量とは記述的である!

一人の死は悲劇である。

数百万の死は統計である。

– Josef Stalin, Potsdam 1945

$950,000 - 1,200,000$

– ソ連における弾圧の死者数,
1937-1938 (**Ellman2002**)

スターリンの悪名高き、数百万人の死に関する統計の特性についての引用は、少し考えてみる必要があります。彼の主張意図は明らかに、個々人の死は我々の心に触れ、無視することはできないけれども、非常に多くの死については理解できないし、結果的に単なる統計であって、無視してしまうことも簡単である、というところにあります。スターリンは、半分は正しいと思います。統計というのは抽象化であり、個々人の経験を超えた出来事の記述であり、可視化されにくいものです。百万人の死が“本当に”どういうことなのかを想像できる人はほとんどいませんが、一人の死は簡単に想像できますし、孤独な死は悲劇の感情を呼び起こし、Ellman の冷たい統計的記述の感覚が失われたように感じます。

これはそんなに簡単な話ではないのです。数字がなければ、数えなければ、何が起こったのかの記述がなければ、われわれは本当に何が起こったのかを理解する機会すらもてず、この失われた感覚を

呼び起こす機会さえ持つことができません。そして実際には、私はこれを気持ちの良い土曜日の朝に腰掛けながら書いており、世界の半分そしてこれまでの人生でずっと、ソ連の強制収容所から離れたところにいるのですが、Ellman の推定値とスターリンの引用を書く時には鈍い恐怖がズッシリ胃にきて、寒気を覚えます。スターリン主義の弾圧は私の経験を超えたところにありますが、統計データと結びつき、そこに記録された個人史を思うと、私の理解を完全に超えているとはいえないません。なぜなら、Ellman の数字は私たちに教えてくれるからです。2年以上のスターリンの弾圧によって、私の住んでいる街に今生きている全ての男性、女性、子供たちと同じ数の人が消え去ってしまったのだ、ということを。この死の一つ一つに、独自の物語があって、それぞれの悲劇があって、その幾らかは私たちにも知られています。ですから、注意深く選ばれた統計量を見ながら、残虐行為のスケールに焦点化していきましょう。

統計家と科学者の最初の仕事である、データを集めて要約し、何が起ったのかを聴衆に知らせる数字を見つけてくるというのは、簡単なことではないのです。これは記述統計の仕事ですが、数字だけを使って何が言えるかはその仕事ではありません。あなたはデータアナリストであり、統計ソフトパッケージではないのです。あなたがすべきはこれらの統計量を取り出して、記述に持っていくことです。あなたがデータを分析するとき、数字のコレクションをリストアップするだけでは十分ではありません。忘れてはいけないのは、あなたは人間の、聴衆を相手にコミュニケーションしようとしているということです。数字は重要ですが、あなたの聴衆が理解できるような意味のあるストーリーと一緒にでなければなりません。あなたはフレーミングについて考える必要がある、ということです。文脈について考えなければなりません。あなたの統計量が要約した、一つ一つの出来事について考えなければなりません。

4. グラフを描く

何よりもまず、データを見せろ

—Edward Tufte^{*1}

データを可視化することは、データを分析しようとするものにとって最も重要な課題です。これが重要なのは、二つの異なる、しかし相互に関係し合う理由によります。まず，“提示するグラフ”を描くこととは、あなたのデータをスッキリと提示し、読者にとってあなたが言いたいことを簡単に理解させるために視覚的に訴えかけるようにすることです。同じぐらい、あるいはもっと重要なことは、グラフを描くことであなた自身がデータを理解できるようになることです。そのために，“探索的なグラフ”を描くことは、あなたがいざ分析しようとしているデータについて理解するのを助けることになるのが重要なのです。このことは当たり前のようでもありますが、私はこれを人に何回言ったかわからないほどです。

この章の重要さを示すために、優れたグラフというものがいかに有用なのかを示す典型例から始めたいと思います。そのために、図 4.1 に最も有名なデータの可視化の例の一つを示しています。これは 1854 年、John Snow によるコロナの死亡者数の地図です。この図はその単純さにおいて、非常にエレガントだといえます。背景として、われわれは見る人の方向性を示すストリートマップを持っている、というのがあります。地図上には多数の小さな点があり、それぞれがコロナの発祥地点を表しています。大きな文字は水のポンプの位置を示していて、その名前ラベルがついています。この図をちょっと見ただけでも、アウトブレイクの源は Broad Street ポンプを中心にしていることが明らかです。このグラフを見て、Dr.Snow はポンプからハンドルを取り除き、500 人以上を殺したアウトブレイクを終わらせたのです。これが、良いデータの可視化の力です。

この章の目標は二つあります。まず、データを分析したり表示したりするとき、私たちがよく使うグラフについて説明し、続いてこれらのグラフを JASP で作成するにはどうすれば良いかを示します。このグラフそのものは、直接的なものなので、この章のある側面は非常にシンプルだと言えるでしょう。人がよく困惑するのは、グラフをどうやって作るかを学ぶとき、特に良いグラフをどうやっ

^{*1}この言葉の原典は、Tufte の本『量的情報を可視化する』です。

て作れば良いかを学ぶときです。幸い、JASPでのグラフの書き方は、あなたがグラフの見え方にそれほどこだわらなければ、かなりシンプルなものです。私がこれをいうことの意味は、JASPのデフォルトのグラフがかなり良いものだということで、ほとんどの場合すっきりとクオリティの高いグラフィックを提供できるということです。しかし、標準的でない図を描きたいとあなたが思ったとき、あるいは図にかなり特殊な変更を加える必要があるとき、JASPのグラフィック関数は発展的な仕事や詳細な編集にはまだ向いていないということはあります。

Snow's cholera map of London

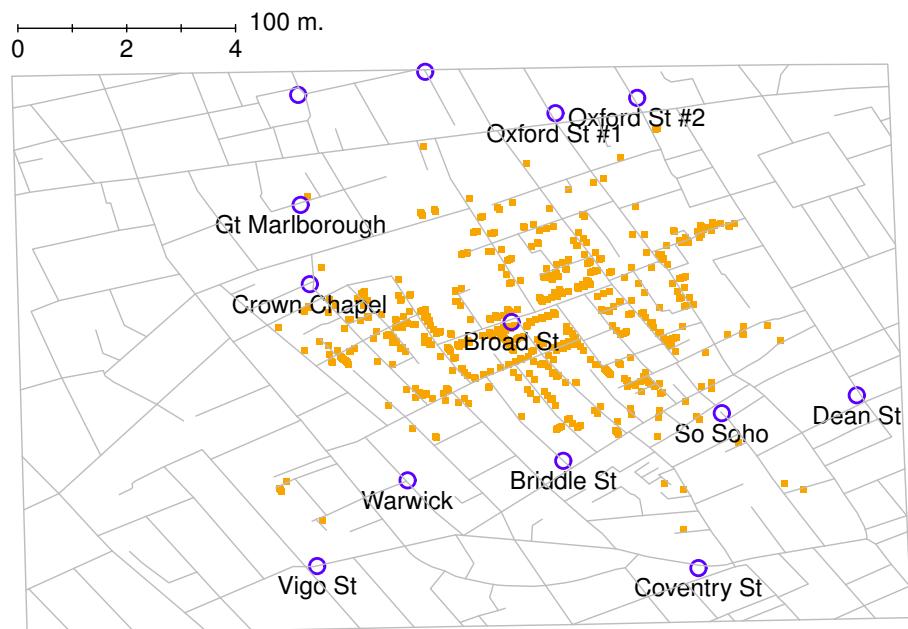


Figure4.1 John Snow のスタイリッシュなコロナマップのオリジナル。小さな各点はコロナ発生点で、大きな円は井戸の位置を示しています。このプロットが明らかにしたように、コロナのアウトブレイクは Broad St のポンプを中心にしていることがわかります。

ヒストグラム

普通のヒストグラムの話から始めましょう。ヒストグラムは最もシンプルで最も一般的な、データ可視化手法の一つです。あなたが間隔尺度水準、あるいは比率尺度水準のデータ（例えば、第3章の `afl.margins` データなど）を持っていて、その辺図宇野全体的な印象を掴みたいと思った時に、ヒストグラムは有効です。ヒストグラムがどんなものかは、ほとんどの人が知っていると思います。広く使われていますからね。でも完璧を期すために、しっかり説明しておきます。あなたがすべきことは、あり得る値をビン幅に分割し、各区間にに入る観測度数の数を数え上げることだけです。この数のことを頻度とかビンの密度といい、それが垂直に伸びるバーとして表示されます。AFLの勝利数データでは、得点が10点未満だったゲームが33ゲームあり、これが以前示した第3の図3.2中、左端のバーの高さとして表されています。以前のグラフはJASPの能力を超えた、Rの発展的プロットパッケージの力を使って描かれていました。しかしJASPもそれに近いことをしてくれます。JASPでのヒストグラムの描画はとても簡単です。「記述」-「記述統計」メニューの下にある「プロット」をひらき、「分布のプロット」チェックボックスをクリックしたのが、図4.2に示されています。JASPのデフォルトでは、y軸が「度数」とラベルされていて、x軸が変数名になっています。ビンは自動的に選択されます。度数が表示されますが、実際の数字はそれほど問題にならないことに注意してください。むしろ、われわれが本当に興味を持っているのは、分布の形状からくる印象なのです。それが正規分布しているのか、それが尖っていたり歪んでいたりしないか？私たちの第一印象は、ヒストグラムから作られるのです。

JASPの特徴を一つ付け加えるなら、「密度」曲線をこのヒストグラムの上に書き加えられるというところです。これをするには「プロット」の下にある「密度を表示」のチェックボックスをクリックしてください。これが図4.3に示されているプロットです。密度プロットは連続した区間や時系列全体をカバーする分布を可視化します。この図は、プロットされた値にカーネルスムージングを使ったヒストグラムの一種で、ノイズを除去した平滑化によって分布をよりスムーズにしたものです。密度プロットのピークは、区間中の値がどこに集中しているかを示してくれています。ヒストグラムの上に密度プロットを描くことの利点は、分布の形をわかりやすくすることにあります。なぜならこれはビン（ヒストグラムで使われている各バー）の数に影響されないからです。たった4つのビンしかないヒストグラムは、20のビンをもつヒストグラムに比べて分布の形をうまく表現できません。でも密度プロットでは、そういう問題が生じません。

この画像はプレゼンテーション用のグラフィック（例えばレポートに入れるもの）にするには、かなり修正する必要がありますが、データを描画する分にはかなりいい仕事をしてくれます。実際、ヒストグラムや密度プロットの強みは（適切に使えば）、データの全体的な広がりを表示し、それがどんな形をしているのかについてかなり良い直感を与えてくれることです。ヒストグラムの欠点は、コンパクトさに欠けるところです。他のプロットと違って、20から30ものヒストグラムを一つの図に詰め込んで人に説明するのはとても難しいのです。そしてもちろん、データが名義尺度水準であれば

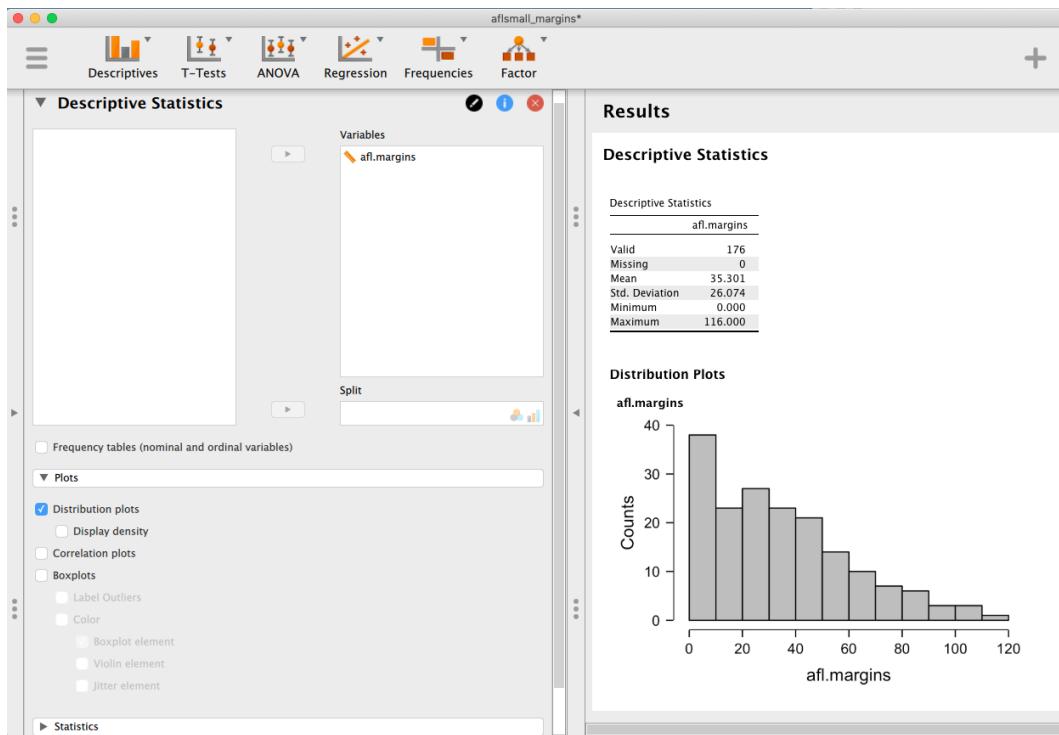


Figure4.2 ‘分布のプロット’ オプションによって作られたヒストグラムを描いた JASP のスクリーンショット

ヒストグラムは適用できません。

4.2

ボックスプロット

ヒストグラムの代わりになるのは、**ボックスプロット**、別名“箱ヒゲ図”と呼ばれるものです。ヒストグラムのように、間隔あるいは比率尺度水準のデータに適しています。ボックスプロットの背後にある考え方とは、中央値、四分位範囲、データの幅を単純に示して見せようというものです。ボックスプロットによる表現は非常にコンパクトで、特にデータ分析の探索的な段階でデータがどんなものかを理解しようとする時の手法としてとてもポピュラーなものになっています。ではそれがどういうものか、`afl.margins` のデータを例にしてみていきましょう。

ボックスプロットがどんなものかを見るために、まず描いてみるのがいいでしょう。‘ボックスプロット’をクリックすれば、右下に図 4.4 のようなものが示されると思います。デフォルトでは、JASP は最も基本的なボックスプロットを示します。このプロットを見れば、そこから何がわかるか

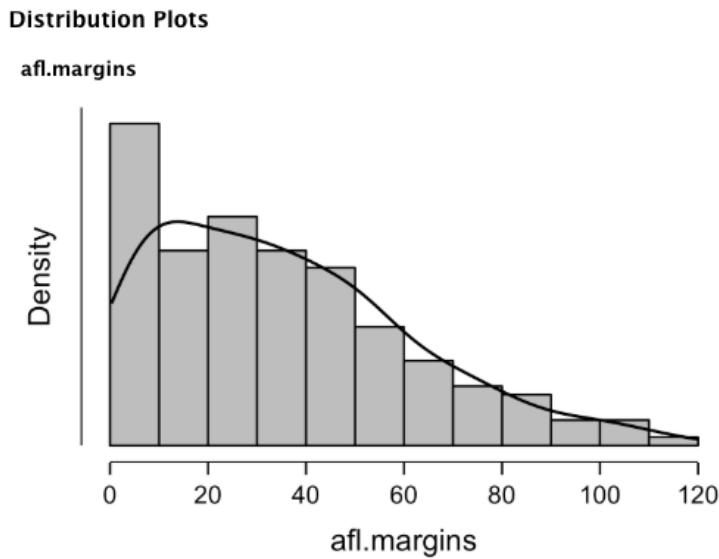


Figure4.3 afl.margins 変数の JASP による密度プロット

一目瞭然です。箱の中心にある太い線が中央値です。箱の幅は 25 パーセンタイルと 75 パーセンタイルの幅になっています。そして “ひげ” の部分はある限界値を超えない最も極端なデータポイントです。デフォルトでは、この限界値は四分位範囲 (IQR) の 1.5 倍で、下限は [25 パーセンタイル点 - \(1.5*IQR\)](#)、上限は [75 パーセンタイル点 + \(1.5*IQR\)](#) になっています。この範囲の外に入る点は、髭でカバーできないので円あるいは点で示され、これは一般的に外れ値とよばれます。私たちの AFL 勝率データでは、二つの観測点がこの範囲の外に落ちており、この観測データは点で表されています（上限は 107 で、スプレッドシートのデータをみると 2 件これより大きいもの、108 と 116 があり、それぞれの点が打たれています）。

4.2.1 Violin plots

伝統的なボックスプロットのバリエーションとして、バイオリンプロットというのがあります。バイオリンプロットはボックスプロットに似ていますが、異なる値におけるデータのカーネル確率密度も表示してくれます。典型的には、バイオリンプロットはデータの中央値と、標準的なボックスプロットと同じような四分位範囲を示すボックスも同時に示します。JASP では、この種の機能は ‘バイオリンの要素’ と ‘ボックスプロット要素’ のチェックボックスをチェックすることができます。図 4.5 では、データ点もプロットしました（これは ‘Jitter 要素’ のチェックボックスを選択することで、

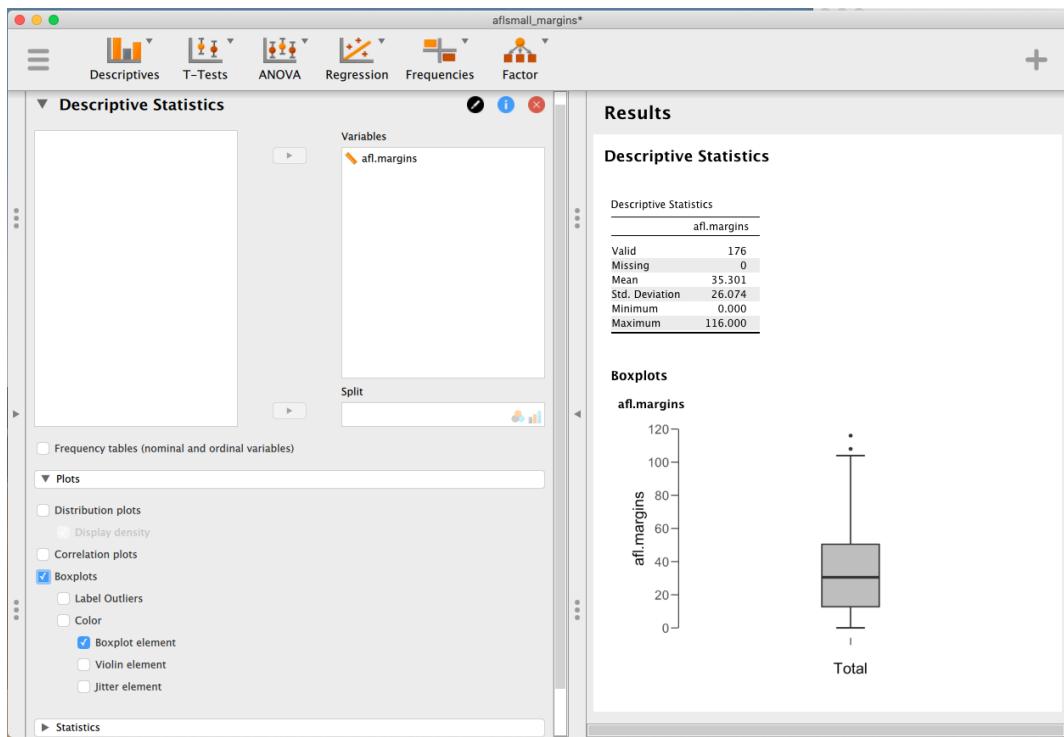


Figure4.4 JASPによるafl.margins変数のボックスプロット

プロットに実際のデータ点を追加します)。

4.2.2 複数のボックスプロットを描画する

最後にもう一つだけ。複数のボックスプロットを一度に書くにはどうしたらいいでしょう？ 例えば、2010年のAFL勝率データだけでなく、1987年から2010年までの各年度のボックスプロットを個別に描きたいと思ったとしましょう。これをするためには、まずデータを見つけなければなりません。このデータはaflsmall12.csvファイルにあります。ではJASPに読み込んで、みてみましょう。これはちょっと大きなデータセットであることがわかると思います。ここには4296ゲームとその変数が含まれています。JASPで**勝率**変数についてのボックスプロットを描く時に、**年度**ごとに分けたいですね。それをするためには、**年度**変数を名義尺度水準の変数に変換し、**年度**にわたってボックスの‘分割’をします。

その結果が図4.6です。このバージョンのボックスプロットは、年度ごとに分割されており、ヒストグラムよりボックスプロットを使った方がいいこともあるのはなぜか、ということがすぐにわかりますね。これを見ると、データの詳細に入り込まなくとも年度ごとにどうなっているか、わかりやすくなっています。もしこのスペースに24個のヒストグラムを詰め込もうとしたら、何が起こるか考

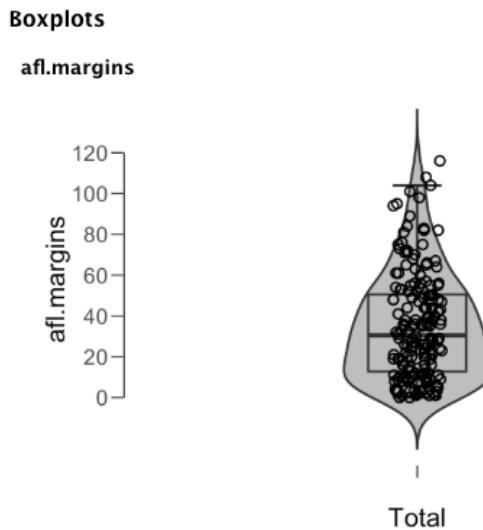


Figure4.5 JASPにおける[afl.margins](#)変数のバイオリンプロットにボックスプロットとデータ点も重ねてみました

.....

えてみてください。そんなことをしても、読者が何かを学べるとは思いませんけどね。

4.3 _____

JASPで画像を保存する方法

ちょっと待って、と思ってるかもしれませんね。JASPでいい図が欠けてもそれを保存したり友達に送り、私のデータがいかに素晴らしいかを語れないようでは意味がありません。図を保存するにはどうしたらいいでしょう？簡単です。プロットの上部、横についている三角形をクリックして、「名前をつけて画像を保存’を選ぶだけです。いくつかのフォーマットを選んで保存することができ、選択できる形式は‘png’, ‘pdf’, ‘eps’, ‘tif’があります。これらのフォーマットで友達に画像を送ったり、(もしかするとさらに重要なことには)それらを課題や論文に含めることができます。

4.4 _____

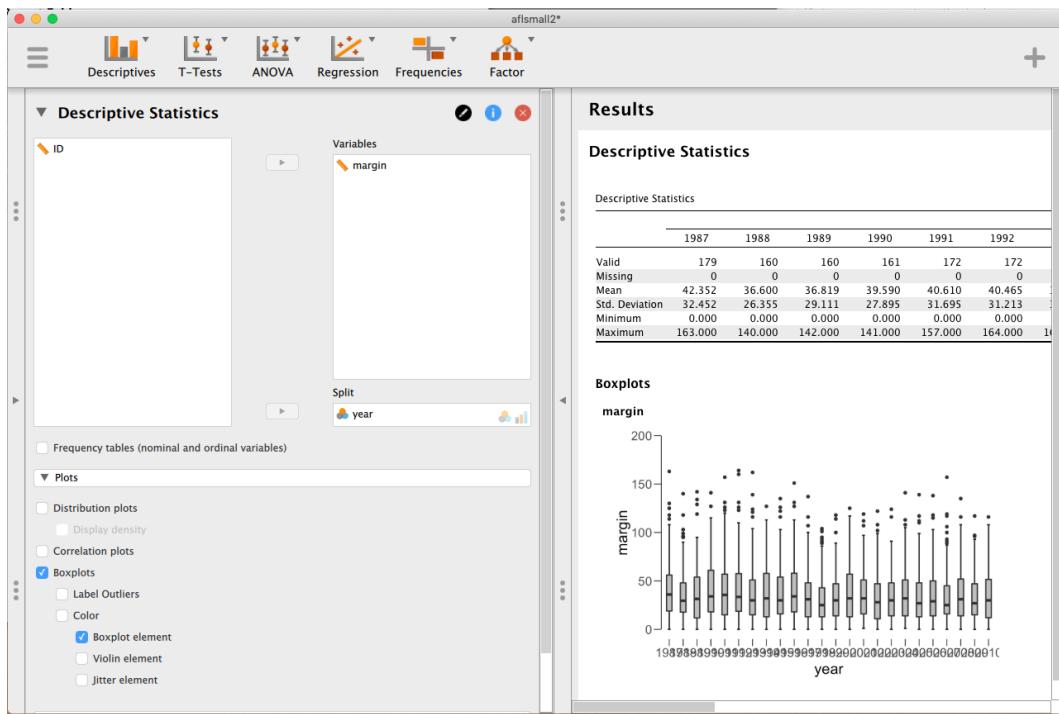


Figure4.6 JASPにおける複数ボックスプロット。aflsmall12 データセットにおける、年度変数ごとの 勝率

要約

多分私は単純な心の持ち主なんですが、絵が好きなんです。新しい論文を書き始めるとき、まず私がすることはどっしり座ってどんな絵を描こうかなと考えるのです。頭の中で、その論文は実際にストーリーに沿った一連の図を思い浮かべています。残りは飾りに過ぎません。私がここで伝えたいことは、人間の視覚システムはとても強力なデータ分析ツールであるということです。図は正しい情報を与え、大量の情報を瞬時に読者に伝えることができます。“百聞は一見にしかず”という諺の通りです。そう考えると、この章はこの本の中で最も重要なものの一つではないかと思うのです、本章で扱ったのは次のとおりです。：

- 一般的なプロット。この章のほとんどは統計学者が好んで使う標準的なグラフを紹介しました。:ヒストグラム(セクション 4.1)とボックスプロット(セクション 4.2)です。
- 画像の保存。重要なことで、あなたの描いた図を出力することについても言及しました(セクション 4.3)

最後にひとつ。JASP は美しいグラフを提供してくれますが、プロットの編集はできるようにはなっていません。もっと発展的なグラフやプロットの可能性を引き出すには、R におけるパッケージ

ジを使うことでさらに強力に進めることができます。最も有名な廟 g システムの一つは、`ggplot2` パッケージ (<http://ggplot2.org/> を参照) によって提供されています。これは“グラフィクスの文法”(Wilkinson2006). という考え方に基づいているのです。それは初心者向けではありません。それを使い始める前に、まず R の全体像を掴む必要がありますし、コツを掴むのには少し時間がかかります。ですが、準備ができたら学ぶ価値はあります。それは本当にパワフルでよりスッキリしたシステムなのですから。

Part III.

Statistical theory

Prelude to Part IV

Part IV of the book is by far the most theoretical, focusing as it does on the theory of statistical inference. Over the next three chapters my goal is to give you an introduction to probability theory (Chapter 5), sampling and estimation (Chapter 6) and statistical hypothesis testing (Chapter 7). Before we get started though, I want to say something about the big picture. Statistical inference is primarily about *learning from data*. The goal is no longer merely to describe our data but to use the data to draw conclusions about the world. To motivate the discussion I want to spend a bit of time talking about a philosophical puzzle known as the *riddle of induction*, because it speaks to an issue that will pop up over and over again throughout the book: statistical inference relies on *assumptions*. This sounds like a bad thing. In everyday life people say things like “you should never make assumptions”, and psychology classes often talk about assumptions and biases as bad things that we should try to avoid. From bitter personal experience I have learned never to say such things around philosophers!

On the limits of logical reasoning

The whole art of war consists in getting at what is on the other side of the hill, or, in other words, in learning what we do not know from what we do.

– Arthur Wellesley, 1st Duke of Wellington

I am told that quote above came about as a consequence of a carriage ride across the country-side.*² He and his companion, J. W. Croker, were playing a guessing game, each trying to predict what would be on the other side of each hill. In every case it turned out that Wellesley was right and Croker was wrong. Many years later when Wellesley was asked about the game he explained that “the whole art of war consists in getting at what is on the other side of the hill”. Indeed, war is not special in this respect. All of life is a guessing game of one form or another, and getting by on a day to day basis requires us to make good guesses. So let’s play a guessing game of our own.

Suppose you and I are observing the Wellesley-Croker competition and after every three hills you and I have to predict who will win the next one, Wellesley or Croker. Let’s say that W refers to a Wellesley victory and C refers to a Croker victory. After three hills, our data set looks like this:

*²Source: <http://www.bartleby.com/344/400.html>.

WWW

Our conversation goes like this:

you: Three in a row doesn't mean much. I suppose Wellesley might be better at this than Croker, but it might just be luck. Still, I'm a bit of a gambler. I'll bet on Wellesley.

me: I agree that three in a row isn't informative and I see no reason to prefer Wellesley's guesses over Croker's. I can't justify betting at this stage. Sorry. No bet for me.

Your gamble paid off: three more hills go by and Wellesley wins all three. Going into the next round of our game the score is 1-0 in favour of you and our data set looks like this:

WWW WWW

I've organised the data into blocks of three so that you can see which batch corresponds to the observations that we had available at each step in our little side game. After seeing this new batch, our conversation continues:

you: Six wins in a row for Duke Wellesley. This is starting to feel a bit suspicious.
I'm still not certain, but I reckon that he's going to win the next one too.

me: I guess I don't see that. Sure, I agree that Wellesley has won six in a row, but I don't see any logical reason why that means he'll win the seventh one.
No bet.

you: Do you really think so? Fair enough, but my bet worked out last time and I'm okay with my choice.

For a second time you were right, and for a second time I was wrong. Wellesley wins the next three hills, extending his winning record against Croker to 9-0. The data set available to us is now this:

WWW WWW WWW

And our conversation goes like this:

you: Okay, this is pretty obvious. Wellesley is way better at this game. We both agree he's going to win the next hill, right?

me: Is there really any logical evidence for that? Before we started this game, there were lots of possibilities for the first 10 outcomes, and I had no idea which one to expect. WWW WWW WWW W was one possibility, but so was WCC CWC WWC C and WWW WWW WWW C or even CCC CCC CCC C. Because I had no idea what would happen so I'd have said they were all equally likely. I assume you would have too, right? I mean, that's what it *means* to say you have "no idea", isn't it?

you: I suppose so.

me: Well then, the observations we've made logically rule out all possibilities except two: WWW WWW WWW C or WWW WWW WWW W. Both of these are perfectly consistent with the evidence we've encountered so far, aren't they?

you: Yes, of course they are. Where are you going with this?

me: So what's changed then? At the start of our game, you'd have agreed with me that these are equally plausible and none of the evidence that we've encountered has discriminated between these two possibilities. Therefore, both of these possibilities remain equally plausible and I see no logical reason to prefer one over the other. So yes, while I agree with you that Wellesley's run of 9 wins in a row is remarkable, I can't think of a good reason to think he'll win the 10th hill. No bet.

you: I see your point, but I'm still willing to chance it. I'm betting on Wellesley.

Wellesley's winning streak continues for the next three hills. The score in the Wellesley-Croker game is now 12-0, and the score in our game is now 3-0. As we approach the fourth round of our game, our data set is this:

WWW WWW WWW WWW

and the conversation continues:

you: Oh yeah! Three more wins for Wellesley and another victory for me. Admit it, I was right about him! I guess we're both betting on Wellesley this time around, right?

me: I don't know what to think. I feel like we're in the same situation we were in last round, and nothing much has changed. There are only two legitimate possibilities for a sequence of 13 hills that haven't already been ruled out, WWW WWW WWW WWW C and WWW WWW WWW WWW W. It's just like I said last time. If all possible outcomes were equally sensible before the game started, shouldn't

these two be equally sensible now given that our observations don't rule out either one? I agree that it feels like Wellesley is on an amazing winning streak, but where's the logical evidence that the streak will continue?

you: I think you're being unreasonable. Why not take a look at *our* scorecard, if you need evidence? You're the expert on statistics and you've been using this fancy logical analysis, but the fact is you're losing. I'm just relying on common sense and I'm winning. Maybe you should switch strategies.

me: Hmm, that is a good point and I don't want to lose the game, but I'm afraid I don't see any logical evidence that your strategy is better than mine. It seems to me that if there were someone else watching our game, what they'd have observed is a run of three wins to you. Their data would look like this: YYY. Logically, I don't see that this is any different to our first round of watching Wellesley and Croker. Three wins to you doesn't seem like a lot of evidence, and I see no reason to think that your strategy is working out any better than mine. If I didn't think that WWW was good evidence then for Wellesley being better than Croker at *their* game, surely I have no reason now to think that YYY is good evidence that you're better at *ours*?

you: Okay, now I think you're being a jerk.

me: I don't see the logical evidence for that.

Learning without making assumptions is a myth

There are lots of different ways in which we could dissect this dialogue, but since this is a statistics book pitched at psychologists and not an introduction to the philosophy and psychology of reasoning, I'll keep it brief. What I've described above is sometimes referred to as the riddle of induction. It seems entirely *reasonable* to think that a 12-0 winning record by Wellesley is pretty strong evidence that he will win the 13th game, but it is not easy to provide a proper logical justification for this belief. On the contrary, despite the *obviousness* of the answer, it's not actually possible to justify betting on Wellesley without relying on some assumption that you don't have any logical justification for.

The riddle of induction is most associated with the philosophical work of David Hume and more recently Nelson Goodman, but you can find examples of the problem popping up in fields as diverse as literature (Lewis Carroll) and machine learning (the "no free lunch" theorem). There really is something weird about trying to "learn what we do not know from what we do know". The critical

point is that assumptions and biases are unavoidable if you want to learn anything about the world. There is no escape from this, and it is just as true for statistical inference as it is for human reasoning. In the dialogue I was taking aim at your perfectly sensible inferences as a human being, but the common sense reasoning that you relied on is no different to what a statistician would have done. Your “common sense” half of the dialog relied on an implicit *assumption* that there exists some difference in skill between Wellesley and Croker, and what you were doing was trying to work out what that difference in skill level would be. My “logical analysis” rejects that assumption entirely. All I was willing to accept is that there are sequences of wins and losses and that I did not know which sequences would be observed. Throughout the dialogue I kept insisting that all logically possible data sets were equally plausible at the start of the Wellesley-Croker game, and the only way in which I ever revised my beliefs was to eliminate those possibilities that were factually inconsistent with the observations.

That sounds perfectly sensible on its own terms. In fact, it even sounds like the hallmark of good deductive reasoning. Like Sherlock Holmes, my approach was to rule out that which is impossible in the hope that what would be left is the truth. Yet as we saw, ruling out the impossible *never* led me to make a prediction. On its own terms everything I said in my half of the dialogue was entirely correct. An inability to make any predictions is the logical consequence of making “no assumptions”. In the end I lost our game because you did make some assumptions and those assumptions turned out to be right. Skill is a real thing, and because you believed in the existence of skill you were able to learn that Wellesley had more of it than Croker. Had you relied on a less sensible assumption to drive your learning you might not have won the game.

Ultimately there are two things you should take away from this. First, as I’ve said, you cannot avoid making assumptions if you want to learn anything from your data. But second, once you realise that assumptions are necessary it becomes important to make sure you *make the right ones!* A data analysis that relies on few assumptions is not necessarily better than one that makes many assumptions, it all depends on whether those assumptions are good ones for your data. As we go through the rest of this book I’ll often point out the assumptions that underpin a particular statistical technique, and how you can check whether those assumptions are sensible.

5. 確率への招待

[神] は私たちに黄昏だけを与えたもうた … 確率の。

– John Locke

本書のここまででは、実験デザインにおける鍵となる概念のいくつかを紹介し、またデータセットをどのように要約することができるかについてお話ししてきました。多くの人にとっては、それが統計の全てです。すなわち、全ての数字を集め、平均値を計算し、図を書いて、それら全てをレポートのあちこちに配置することが。切手の収集のようなかんじでしょうか。ただし使うのは数字ですが。しかし、統計学はそれ以上の範囲をカバーするものです。実際、記述統計は統計学の最も小さい領域の一つで、最も影響力のないところでしかありません。統計におけるもっと広大で有用な領域とは、データについての推論ができるような情報を提供してくれるものです。

あなたが統計的に考えることを始めたら、統計はデータから推論を導き出す助けになるし、至る所で使われている例を目にすることになるでしょう。例えば、新聞 Sydney Morning Herald 誌の 2010 年 10 月 30 日に、次のような記事が掲載されていました。

選挙結果に対して、“私は大変な仕事を抱えている”と首相はコメントしました。彼女の政府が今やこれまでにない支持率の低い労働党であり、予備選挙での支持率が 23 パーセントしかなかったのです。

この種の発言は新聞や日々の生活にあっても特に目立つものではないですが、それが何を言わんとしているのかを考えてみましょう。調査会社が調査を実施するときは、彼らには余裕があるので非常に大きな調査を企画するのが普通です。私は面倒くさがりやなので元の調査を調べなかつたのですが、調査会社がニューサウスウェールズ (NSW) の有権者からランダムに 1000 人を集め、そのうち 230 人 (23%) がオーストラリア労働党 (ALP) に投票するつもりだと答えたとします。2010 年の選挙では、オーストラリア選挙委員会は NSW で 4,610,795 人が投票した、と公表していますから、残る 4,609,795 人 (有権者の約 99.98%) の意見がどうだったか、私たちにはわかりません。調査会社に対して誰も嘘をついていないと仮定しても、我々が 100% の自信を持って言えることは、眞の ALP 予備選挙有権者は $230/4610795$ (約 0.005%) から $4610025/4610795$ (約 99.83%) の間のど

こかにいる、ということだけです。それでは調査会社、新聞、その読者が、ALP の予備選挙の支持率が 23% に過ぎないと正当化する根拠は一体どこにあるのでしょうか？

答えはかなりはっきりしています。もし私が 1000 人の人を無作為に呼んできて、そのうち 230 人が ALP に投票するつもりだと答えたとすると、実際に ALP に投票するつもりの人たち全体のうちの、この 230 人だけということはあり得そうにありません。言い方を変えると、調査会社が集めてきたデータはもっと大きい母集団の代表であることを、我々は想定するのです。さて、どの程度代表しているでしょう？ 本当に ALP 予備投票が 24% であれば私たちは驚くのでしょうか。29% なら？ それとも 37% のとき？ ここまでくると、日々の直感は少し崩れていきます。

もし 24% であっても誰も驚かないでしょうが、37% であればみんな驚くでしょう。しかし、29% になりそうだと言うのは少し厳しい気がします。数字を見て推測するだけでなく、もう少し強力なツールが必要です。

推測統計学がこの種の問題に応えるために私たちに必要なツールであり、この種の問い合わせ科学的営みの中心にあるので、統計学や科学的手法についてのあらゆる入門コースの大半を占めているのです。しかし、統計的推論の理論は**確率理論**の頂点の上に作られています。ということで、今から確率理論を学ぶことにしましょう。確率理論についての議論は、基本的にバックグラウンドを細かく見ていくようなものです。この章にはそれほど統計の話は出てきませんし、この本の他の章ほど数学的な詳細を深く理解する必要もありません。ですが、確率理論は統計を深いところから支える支柱ですから、その基礎をカバーしておくことに価値があるのです。

5.1

確率と統計はどうちがうの？

確率理論の話を始める前に、確率と統計学の関係についてちょっと触れておきましょう。この二つの学問は密接な関係にありますが、全く同じものではありません。確率理論は“偶然の原理”です。それは数学の分野の一つで、異なる種類の出来事がどの程度の頻度で生じるのかを教えてくれるもので、例えば、次のような問いは確率理論を使って答えることができるものです。

- 公平なコインが 10 回連続で表になる確率はどれぐらいですか？
- 6 面サイコロを二回振った時、二つとも 6 が出るのはどれぐらいありえることですか？
- 完全にシャッフルされたデッキからカードを 5 枚引いた時、全てハートのカードになることはどれぐらいありえることですか？くじを引いて当たりが出る確率はどれぐらいでしょう？

これらの質問は、いずれも一般的にありふれたものであることに注意してください。どの場合でも、“世界の真理”が分かっている時に、“どんな種類の出来事が”生じるのだろうか、という形に

なっています。最初の質問では、私はコインが公平である、つまり毎回のコイントスで表が出るのは 50% の確率であると知っているのです。第二の質問では、私はサイコロで 6 が出る確率は 1/6 であることを知っているのです。第 3 の質問では、私はデッキがうまくシャッフルされていることを知っているのです。第 4 の質問では、私はくじが特定のルールに従うことを知っているのです。気づきましたね。決定的な点は、確率的な問いは世界について既知の **モデル** から始まり、私たちはそのモデルを使ってなんらかの計算をするのです。そのモデルはかなりシンプルにできます。例えば、コイントスの例では私たちはモデルを次のように書くことができます。

$$P(\text{heads}) = 0.5$$

これは「表が出る確率は 0.5」と読むことができます。後で見るようすに、0% から 100% の範囲にある比率の数字と同様に、確率は 0 から 1 の数字になります。この確率モデルを使って最初の問い合わせるのですが、私はこれから起こることを実際には知りません。この質問がいうように、10 回表が出るかもしれません。でも、3 回しか出ないかもしれません。これが大事なところなのです。確率理論では、モデルはわかっていますが、データはわからないのです。

それが確率なのです。では統計学とは何でしょう？統計学の問い合わせは、その周りにあって別の働きをするものです。

統計学では、私たちは世界の真理について知りません。私たちが持っているのはデータだけであり、世界の真理について学びたいことはデータから得られるのです。統計的な問い合わせは次のようになることが多いようです。

- もし友達がコイントスを 10 回やって表が 10 回出たとしたら、彼は私をからかっているんじゃないだろうか？
- もしデッキの上から 5 枚カードを取り出して、それが全部ハートだったら、そのデッキがシャッフルされていた可能性はどれぐらい？
- もし宝くじ主催者の配偶者がくじに当選したら、その宝くじがイカサマだった可能性はどれぐらい？

この時、私たちが知っているのはデータだけですね。私が知っていることは、友達が 10 回コイントスをして、全部表であったことだけです。そして私が**推論**したいことは、実際に公平なコインが連續して 10 回表になったのだと結論づけて良いかどうか、あるいは私の友達が私をからかっていると疑って良いかどうかです。ここでのデータは次のようにになります。

表 表 表 表 表 表 表 表 表

そして、私がやろうとしていることは、どの“世界についてのモデル”を信用するべきか、ということです。もしコインが公平なものであれば、私が受け入れるべきモデルのひとつは表が出る確率が 0.5、つまり $P(\text{heads}) = 0.5$ であるというものです。もしコインが公平なものでなければ、表が出

る確率は 0.5 ではないことになるので、それを我々は $P(\text{heads}) \neq 0.5$ と書くでしょう。言い換えると、統計的な推論の問題は、どっちの確率モデルが正しいかということです。これで明らかのように、統計的な問い合わせは確率の問い合わせと同じではないのですが、お互い深く関係し合っているのです。だから統計理論のよい入門書は、確率についての議論とそれがどのように機能するか、というところから始めるのです。

5.2 _____

確率が意味するものは？

いくつかの質問から始めましょう。“確率”とはなんでしょう？あなたは少し驚くかもしれませんが、これについて統計学者や数学者が（ほとんど）同意してくれるのは確率のルールであって、それが本当は何を意味するのかについては、ほとんど同意が得られません。私たちは“偶然 (chance)”とか，“ありそう (likely)”とか，“ありえる (possible)”とか，“たぶん (probable)”という言葉を大変便利に使うので、この問題に答えるのが難しそうだと言われても奇妙に感じます。しかし実生活の中でも、会話がうまくいってないように感じてそれから距離を置いていたけど、（多くの日常的な概念について）あなたがそれが何なのか本当は分かっていなかったことがあとでわかる、という経験をしたことがあるでしょう。

さて始めましょうか。私が二台のロボットチーム、アーセナルとミランが戦うサッカーゲームに賭けたいとします。いろいろ考えて、アーセナルが 80% の確率で勝つだろう、と判断したとします。さてこれは何を意味しているのでしょうか？ これには三つの可能性があります。

- これはロボットチームなので、私は何度も試合を試すことができ、実際そうすると 10 ゲーム中 8 回アーセナルが勝つだろう、ということ。
- どんなゲームであっても、このゲームに賭けるときは、\$1 をミランに賭けたら \$5 戻ってくる（つまり、私の \$1 は賭けに勝つと \$4 儲けさせてくれる）し、アーセナルには \$4 を賭けないと \$5 戻ってこない（つまり、私の \$4 プラス \$1 の儲け），という時に初めて“公平な”賭けが成立する、ということ。
- アーセナルの勝利について、私の主観的な“信念”とか“自信”は、ミランの勝利についての信念より 4 倍強い、ということ。

どれも微妙ですねえ。しかし、いずれも同じ、ではないし、全ての統計学者がこれら全てを認めているわけでもないのです。その理由は、そこには統計学的イデオロギーの違いがあるからで（本當です！），どれを認めるかによって、この表現のいくつかは無意味だと不適切だと言いたくなるでしょう。このセクションでは、ここにある二つの大きなアプローチについて簡単な導入を行います。

アプローチは一つだけということは決してありませんが、その二つは大きな流れなのです。

5.2.1 頻度主義者の観点から

確率への2大アプローチのうち最初のものは、統計学においてより支配的な考え方であり、**頻度主義者の観点**と言われます。それは**長期的な頻度**として確率を定義するからです。私たちが何度も何度もコインフリップをし続けるとしましょう。定義から、このコインは $P(H) = 0.5$ です。私たちは何を観測するでしょう？一つの可能性として、最初の20回は次のようになったとします。

T, H, H, H, H, T, T, H, H, H, H, T, H, H, T, T, T, T, T, H

この例では20回中11回(55%)が面を向いています。今度は、私がずっとコインフリップをしながら表が出る回数(ここでは N_H としておきましょう)を数え、その最初から N 回目までで確率を N_H/N として毎回計算していくとしましょう。そうすると、次のようにになります(これを書くために、私は文字通りコインフリップをしたんですよ！)：

コインフリップの回数	1	2	3	4	5	6	7	8	9	10
表が出た回数	0	1	2	3	4	4	4	5	6	7
割合	.00	.50	.67	.75	.80	.67	.57	.63	.67	.70

コインフリップの回数	11	12	13	14	15	16	17	18	19	20
表が出た回数	8	8	9	10	10	10	10	10	10	11
割合	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55

この流れの最初のうちは、表が出る割合は広く変動し、最初は.00ですが.80まで上昇します。続けていくと、“正しい”答えである.50に徐々に近づいていくような印象を持つ人もいるかもしれません。これがごく簡単にいうところの、頻度主義者の確率の定義です。公平なコインを何度もフリップしつづけ、 N が大きくなれば(無限に近づけば、というのを $N \rightarrow \infty$)と書きますが)、表が上になる割合は50%に収束して行きます。ここには数学的に注意が必要なちょっとしたテクニックがありますが、内容的にはこれが頻度主義者が確率を定義する方法です。残念ながら、私はコインを無限回フリップしたことではないですし、無限回コインをフリップするのに耐えうる無限の精神力も持ち合わせていません。でも私はコンピュータを持っているし、コンピュータはこの辛い作業を厭わないのです。ですから、私はコンピュータに、1000回コインフリップのシミュレーションをやるようにお願いし、 N_H/N が N の増加とともにどうなるかの図を描いてもらいました。実際には、私はそれがまぐれではないことを確認するために4回繰り返しました。その結果が Figure 5.1 です。ご覧いただいたように、表が観測される割合は最終的に変動するのをやめ、落ち着いて行きます。そうなると、

最終的に落ち着いた数字が表の出る確率、ということになります。

頻度主義者の確率の定義は、いくつかの望ましい性質を持っています。まず、それは客観的です。出来事の確率は世界に根ざしたものである必要があります。確率の言葉が意味を持つのは、それが物理的な宇宙の中で生じる（一連の）出来事について言及しているからです^{*1}第二に、曖昧さがありません。同じ一連の出来事を観測した人は誰でも、その出来事の確率を計算しようとすれば、確実に同じ答えに到達します。

しかし、望ましくない特徴もあります。まず、無限の連続というのは物理的な世界にはありません。あなたがポケットからコインを取り出して、コインフリップをし始めたと思ってください。コインが着地するたびに、それは地面に衝撃を与えます。毎回の衝撃で、コインは少しづつ欠けていきます。最終的に、コインは破壊されてしまうかもしれません。だから、“無限の”コインフリップの連続、というのが意味のある概念で客観的なものであるとして、意味のあるフリをするべきかどうか、疑問に思うかもしれません。私たちは出来事の“無限の連続”が物理的な宇宙において現実的なものであるということはできません。なぜなら、物理的な宇宙は無限の何かを許容しないからです。もっと厳密にいうと、頻度主義者の定義は対象が狭いのです。日々の生活の中で人が確率に割り当てて満足していることはたくさんあるのですが、（理論の中でも）仮想的な出来事の連続に割り当てられないものがたくさんあります。例えば、気象学者がテレビで「2048年、12月2日のアデレードで雨が降る確率は60%です」といったとしても、私たちはこれを喜んで受け入れます。ですが、これを頻度主義者の用語でどうやって定義するのかはっきりしません。アデレードは一つしかない街ですし、2048年12月2日も一回しかありません。またこれは無限の出来事の連続はありません、一度きりのことです。頻度主義者の確率は、一度しか起きない出来事について確率でものをいうことを本気で禁じます。頻度主義者の物の見方からすると、明日は雨が降るか降らないかのどちらかです。一度きりの繰り返しのない出来事に“確率”は付随させられないのです。では、頻度主義者がこれを回避するために使う、非常に巧妙なトリックがあることを指摘しなければなりません。一つの可能性として、気象学者が言わんとしているのは、“私が60%の偶然で雨が降る日々、というカテゴリーがある、もしそうした日だけみてこの予測をしたとすると、その日のうちの60%が実際に雨が降るのです”，とまあこういうようなことにすることです。これは非常に奇妙で、直観に反したものですが、頻度主義者がこのような言い方をするのを実際に目にすることでしょう。そしてこの本の後でもこのことがきっと出てくるでしょう（Section 6.5 をみてください）。

5.2.2 ベイジアンの観点から

確率についてのベイジアンの観点は、時に主観的な観点だと言われ、統計学者の中ではマイノリ

^{*1}これはもちろん、頻度主義者が仮説的な発言ができないことを意味するものではありません。単に、もしあなたが確率について表明したいことがあれば、それは潜在的に観測しうる一連の出来事についての言葉を、違う結果の相対的な頻度と共に、表現しなおせるものでなければならないということです。

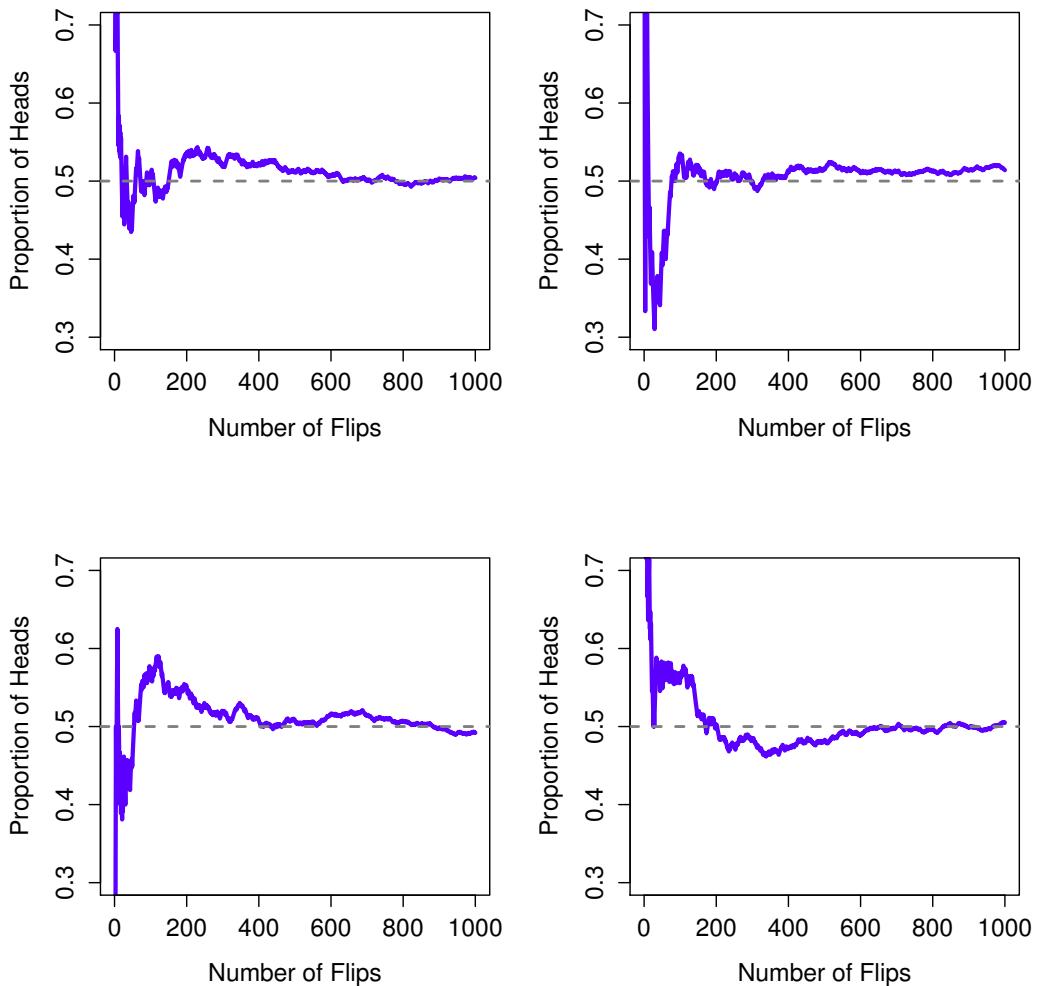


Figure5.1 頻度主義者の確率がどのように働くかの図。コインフリップを何度も何度も繰り返すと、表が出る割合は徐々に落ちていき、真の確率である 0.5 に収束していきます。各パネルが示すのは、四つの異なるシミュレーション実験の結果です。それぞれのケースについて、コインを 1000 回フリップして表が出る割合を追跡し続けたものです。どのケースも最終的にちょうど 0.5 になりはしませんが、もしこれを無限回実験し続けるようにすれば、最終的にはそうなるはずです。

ティでしたが、この数十年の間にかなり牽引力を持ってきました。ベイズ主義的精神は多くの楽しみ方があるので、“これぞ”ベイジアンの観点、と正確に言い切るのは難しいものがあります。主観的確率について考えるもっとも一般的な方法は、出来事についての確率を、出来事の真実に対して知的かつ合理的なエージェントが割り振る**信念の程度**として定義することです。この観点からいくと、確率は現実世界に存在していないことになり、むしろ人やその他の知的存在の思考や仮定の中に存在することになります。

しかし、このアプローチにしたがって仕事をすると、私たちはなんらかの方法で“信念の程度”を操作可能なものにする必要があります。いろいろな方法がありますが、一つのやり方としは、“合理的なギャンブル”的用語で定式化することです。明日雨が降る確率が60%だ、と私が信じているとします。ここで誰かが、もし明日雨が降ったら私に5\$あげるけど、雨が降らなかったら5\$よこせ、という賭けを申し出されたとします。もちろん、私の立場からすれば、これはちょっといい話です。しかし一方で、もし私が明日雨が降る確率は40%だと考えていたのなら、この賭けをするのは悪い手ということになります。つまり、私たちは“主観的確率”的本質を、私がこの賭けを受け入れるかどうかという用語で操作したことになります。

ベイジアンアプローチの利点と欠点はなんでしょう？主な利点は確率を割り当てたいどんな出来事にも適用できるということです。繰り返しのある出来事に限定する必要はありません。(多くの人にとっての)主な欠点は、私たちは完全に客観的にはなれないということです。確率を割り当てるということは、関連する信念の程度についての実態を特定する必要があります。その実態は人間、エイリアン、ロボット、統計学者、誰のものでもいいのです。しかし物事を信じる知的なエージェントがそこにいなければなりません。多くの人にとってはこれが不満のタネになります。幾分曖昧になりますからね。ベイジアンアプローチは、そのエージェントが合理的であること(すなわち確率のルールに従うこと)を要求しますが、誰もが自分自身の信念を持つことを許します。私はコインが公平なものであると信じられるし、あなたは必ずしもそうでなくていいのです。私もあなたも合理的なままで。頻度主義者の観点は、二人の観察者が同じイベントに異なる確率を割り当てるのを許しません。もしそういう事態が生じたら、二人のうちどちらかが間違っているのです。ベイジアンの観点ではこの問題が生じることを止めません。異なる背景知識を持つ二人の観測者が、同じ出来事に対して合法的に異なる信念を持ちうるのです。簡単にいって、頻度主義者の観点が時折見識が狭すぎるよう見える(確率を割り当てる出来事の多くを許してくれない)のに対し、ベイジアンの観点は時折懐が広すぎる(観察者間であまりにも多くの異なる状態を許す)のです。

5.2.3 違いは何？誰が正しいの？

さて、異なる二つの観点をそれぞれ見てきたわけですが、この二つを比較してみることにしましょう。このセクションの最初に提示した、ロボットサッカーゲームの例に戻りましょうか。この三つの表現について、頻度主義者やベイジアンがどう考えると思いますか？頻度主義者がいう正しい確率の定義に当てはまるのは、どの表現でしょう？ベイジアンが選ぶのはどの表現でしょう？頻度主義者

やベイジアンにとって、意味を持たない表現はどれでしょう？もしあなたが二つの観点を理解できたなら、これらの問い合わせにどう答えたらいいかわかるはずです。

オーケイ、あなたは両者の違いを理解していて、その上でどちらが正しいのか、迷っているんですね？正直にいうと、どちらが正しい答えなのか私も知りません。言えることは、頻度主義者のように一連の出来事を考えることは数学的に間違えているわけではないし、ベイジアンの合理的エージェントの信念で定義するのも数学的に間違えているわけではない、ということです。実際、深く掘り下げていくと、ベイジアンと頻度主義者は多くの点で合意できることがあります。多くの頻度主義者の方法は、ベイジアンの合理的なエージェントがするであろう意思決定と同じことを言います。多くのベイジアンの方法は、頻度主義者の良い特徴をも持っています。

ほとんどの部分において、私は現実主義的ですから、私は信頼できるあらゆる統計的な手法を使います。結局、この本での説明は、ベイズ的手法の方が好ましいようになっているかもしれません。しかし私は頻度主義的な方法について、基本的に反対の立場にないのです。誰しもそこまで満足しているわけではありません。例えば、R. フィッシャー卿のことを考えてみます。彼は 20 世紀の統計学者の巨人で、ベイジアンのあらゆることに対する猛烈な敵の一人であり、ベイジアンの確率について、その統計の数学的基礎に関する論文を“より精緻な統計的概念への発展を阻むジャンブル”(Fisher 1922b) といったぐらいです。一方、心理学者の P. ミールは、頻度主義的方法に傾倒すると、あなたを“夢見る乙女の楽しい長旅だが科学的な成果を残すことのない、納得はするけど不毛な知的探索”(Meehl 1967) に連れていくのだ、と言ったりしています。聞いたことがあるかもしれません、統計の歴史はエンターテインメント性を欠きません。

どちらにせよ、私はベイジアンの観点が好きですが、統計分析の多数派は頻度主義的アプローチを基盤にしています。私の理由はプラグマティックなものです。この本のゴールは心理学における典型的な学部統計教育の領域をざっとカバーしていますので、ほとんどの心理学者が使っているような統計的ツールを理解したいと思うのなら、頻度主義者の方法を掴み取る必要があるでしょう。その努力は無駄にならないと約束します。あなたがもしベイジアンの観点に切り替えたいと思うのなら、“オーソドックスな” 頻度主義者の観点で書かれた本を一冊は読み通すべきです。とはいえ、私はベイジアンの観点を全く無視するわけではありません。今までそしてこれからも、私はベイジアンの観点からコメントを追加するでしょうし、Chapter ?? ではより深い内容を掘り下げていきたいと思います。

5.3

確率の理論の基礎

ベイジアンと頻度主義者の思想的な議論はさておき、確率が従うべきルールについてはほとんど同意がとれています。これらのルールに到達するには様々な異なる道があります。もっとも一般的に使

われるアプローチは、20世紀の最も優れたロシアの数学者、アンドレイ・コルモゴロフによって基礎が作られたものです。詳細に立ち入ることはしませんが、それがどのようなものか、ちょっとした感覚をお伝えしようと思います。そのためには、私は私のズボンについて話さなければなりません。

5.3.1 確率分布入門

私の人生における困った真実の一つは、私がズボンを5本しか持っていないということです。ジーンズのものが3つ、スーツの下が一つ、トレーニングウェアのズボンが一つ、です。さらに悲しいことに、私はそれに名前をつけています。わたしはそれを、 X_1, X_2, X_3, X_4, X_5 と呼んでいます。本当にそうなんです。だから私はミスター想像力、と呼ばれています。さてある日、私がそのズボンの一つを取り出して履きました。ズボンを二つ同時に身につけようとするほど愚かではないですし、トレーニングのおかげでズボンを履かずに外に出ることはもうなくなりました。この状況を確率理論の言葉を使って表現するなら、それぞれのズボン(つまり各 X)のことは、**根元事象**といいます。根源事象のキーポイントは、私たちが観測するとき(例えば、私がズボンを身につけようとするとき)はいつでも、結果は一つ、そしてその出来事のどれか一つでしかない、ということです。言ったように、私はいつもズボンを1着しか身につけませんから、私はこの制約を満たしていることになります。同様に、あらゆる確率事象のセットのことを、**標本空間**といいます。確かに、これを“衣装部屋”と呼ぶ人もいるかもしれません、それは確率の用語で私のズボンについて語ることを拒否しているからです。残念。

オーケイ、私たちは今や標本空間(衣装部屋)を手にしたわけで、それは可能な根元事象(ズボン)から出来上がっているので、この各要素である事象に**確率**を割り振っていきたいと思います。事象 X に対して、事象の確率 $P(X)$ は0から1の間の数字です。より大きな $P(X)$ の値は、その事象がより生じやすいことを意味します。そう例えば、もし $P(X) = 0$ なら、事象 X は生じ得ない(つまり、私は決してそのズボンを履かない)ことを意味します。あるいは、もし $P(X) = 1$ なら、事象 X は確実に生じる(つまり、私はいつもそのズボンを履く)ことを意味します。その間にある確率の数字が意味するのは、私は時々それらのズボンを履くということです。たとえば、もし $P(X) = 0.5$ なら、私は二回に一回そのズボンを履く、ということを意味します。

ここまできたら、ほとんど終わったようなものです。最後にやらなければならないことは、“いつも生じるなにか”を認識する必要があるということです。私がズボンを履く時はいつも、本当にズボンをちゃんと履いているのです(おかしなことを言ってるようですが、正しいですよね?)。確率の言葉で幾分古臭い表現になりますが、根元事象の確率を足し合わせると1になる、ということです。これは**確率の総和の法則**として知られているもので、誰もが本当に気にしているわけではありません。

より大事なことは、これらの必然性が満たされたなら、私たちが手にしたのは**確率分布**である、ということです。例えば、ここに確率分布の例があります。

どのズボン?	ラベル	確率
hline 青いジーンズ	X_1	$P(X_1) = .5$
灰色ジーンズ	X_2	$P(X_2) = .3$
黒いジーンズ	X_3	$P(X_3) = .1$
黒いスーツ	X_4	$P(X_4) = 0$
黒いトレーニングウェア	X_5	$P(X_5) = .1$

各事象は 0 から 1 の間の確率についての数字を持っていて、全ての確率を足し合わせると 1 になります。驚きました。この分布を可視化するために、棒グラフを書くこともできます。図 5.2 を見てください。さて、ここにきて、私たちはすべて成し遂げたようです。すでにあなたは確率分布の何たるかを学びましたし、私はズボン全体に注目したグラフの作り方を見つけ出したのです。私たちの勝利です！

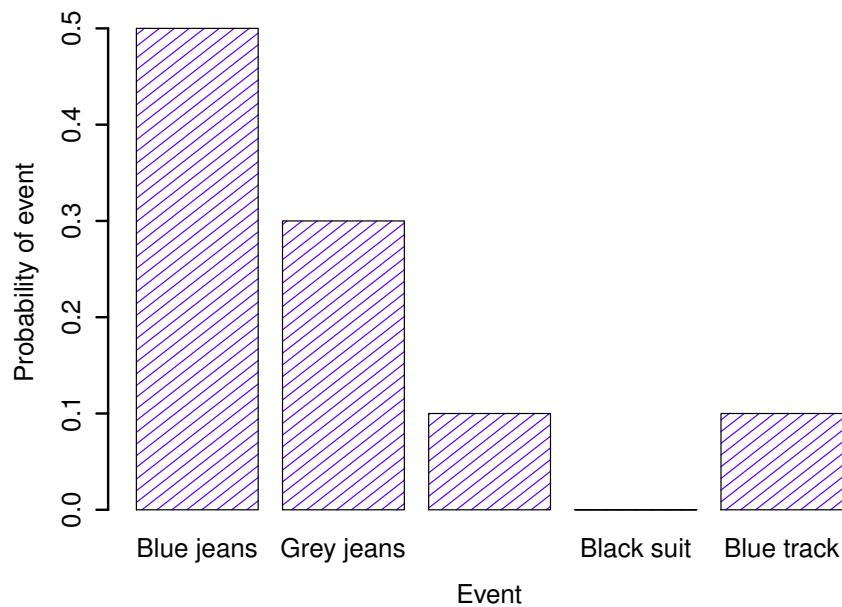


Figure5.2 “ズボン” 確率分布の視覚的記述. ここには 5 つの“根元事象” がって、それは私の持っている 5 本のズボンに対応しています。各事象はなんらかの生起確率を持っています。この数字は 0 から 1 の間の値です。これらの確率すべてを足すと 1 になります。

もう一つ指摘しておかなければならぬことがあります。それは、確率理論は**非根元的事象**についても、根元事象と同じように語ることを許してくれるということです。この考え方を表現する最も簡単な方法として、例をあげましょう。ズボンの例では、私がジーンズを履く確率を完全に適切な方法

Table5.1 確率が満たすべきいくつかの基礎的ルール。この本でこの後お話しする分析を理解するため、これらのルールを知っておく必要は必ずしもありませんが、もう少し深く確立理論を理解しておきたいのなら、重要なことです。

英語で	表記	式
not A	$P(\neg A)$	$= 1 - P(A)$
A or B	$P(A \cup B)$	$= P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B)$	$= P(A B)P(B)$

で参照できるのでした。このシナリオの中ですと、適当な出来事の一つとして実際に生じうる根元事象である限り，“ダンがジーンズを履く”という事象が生じたということができます。この場合，“青いジーンズ”，“黒いジーンズ”，“灰色ジーンズ”が該当します。数学用語で私たちが“ジーンズ”事象を E と定義する時、それは根元事象 (X_1, X_2, X_3) のセットに対応します。これらの根源事象のどれが生じても、 E が生じたと言っても良いでしょう。 E の定義をこのように書き下したとして、確率 $P(E)$ について言及するなら、ちょっと直接的すぎますが、単にこれらを数え上げればいいですね。この場合ですと、

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

とすることであり、青、灰色、黒のジーンズの確率がそれぞれ.5,.3,.1なので、私がジーンズを履く確率は.9だ、ということができます。

この時点で、あなたが非常に明白でシンプルだと思うかもしれません。それは正しいです。私たちがここでやったのは、いくつかの常識的な考え方にある基本的な数学のラップをかけただけ、なのです。とはいえ、これらの単純な始まりから、とてもなく強力な数学的道具を作り上げることができます。この本では決してその詳細にまで立ち入りませんが、そのほかの確率が守るべきルールについてはリストにして、表 5.1 に示しています。これらのルールは私がすでに上で述べたシンプルな仮定から導出できますが、この本のどこにもこのルールを適用するところがありませんので、やらないでおきます。

5.4

二項分布

あなたの想像通り、確率分布は非常に様々に変化しますし、分布の範囲は広大な範囲に及びます。ですが、それら全てが同程度に重要だということではありません。実際、この本の内容の大部分は、5つの分布のどれかに依存しています。その5つとは、二項分布、正規分布、 t 分布、 χ^2 (“カイ二乗”) 分布、 F 分布です。次のいくつかのセクションで私がやろうとしているのは、これら5つ全てにい

ての簡単な導入です。特に二項分布と正規分布に注目していきます。私は二項分布から始めようと思います。これが5つの中では最もシンプルですから。

5.4.1 二項分布の導入

確率の理論は偶然のゲームがどのようにになっているのかを記述しようとする試みから始まりました。ですから、私たちの二項分布についての議論は、サイコロをふったりコインをフリップするお話をするのがよいでしょう。単純な“実験”を想像してみてください。私の小さなあったかい手には、6面サイコロが20個握り締められています。各サイコロの一つの面にはドクロの絵が書いてあって、残りの5つの面には何も書いていないものとします。20個のサイコロ全てを転がした時、ちょうど4つのドクロが出る確率はどれくらいでしょう？サイコロにいかさまがないとすると、サイコロのドクロのある面が上を向く確率が $1/6$ であることがわかります。これを言い換えると、一つのサイコロについてドクロの出る確率は約 $.167$ であるということです。これで私たちの問い合わせに答えるには十分な情報ですね。ではどうなるかみてみましょう。

Table 5.2 二項分布と正規分布の式。私たちはこの本で他にこの式を使うことは本当にはないのですが、より発展的な話に進むためにはちょっと重要なので、ここで文章の邪魔にならない表の形で示しておくのが良いと考えました。二項分布の式の中で、 $X!$ とあるのは階乗関数(つまり、1から X までの全ての数字を掛け合わせたもの)であり、正規分布の‘exp’は指数関数を表します。もしこれらの式があなたにとってあまりわかりやすいものでなかったとしても、そんなに恐れることはありません。

Binomial	Normal
$P(X \theta, N) = \frac{N!}{X!(N-X)!} \theta^X (1-\theta)^{N-X}$	$p(X \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$

普通、名前とか表記法について少し説明するものですね。ここでは N を、実験におけるサイコロを振った回数としますが、これはサイズ・パラメータといわれ、二項分布ではよく参照されるものになります。私たちは θ を、一つのサイコロについてドクロの目が出る確率を表すために使います。この量は、二項分布では普通成功確率と呼ばれます。^{*2} 最後に、 X はその名の通り、私たちのやる実験においてサイコロをふった時に得たドクロの数を意味します。 X の実際の数字は偶然によるものなので、これを確率変数といいます。どの場合でも、これら全ての用語と表記を手に入れたのですから、この問題をもう少し正確に記述することができるようになったわけです。私たちが計算したい数

^{*2}“成功”という言葉がちょっと曖昧なことに注意してください。アウトカムが実際は望ましくないものであったとしても、このようにいいます。もし θ が、バスの事故における怪我をした乗員の数を表す確率であったとしても、私はこれを成功確率と言います。私はバスの乗員が怪我をすることを望んでいるわけではありません！

字は $X = 4$ の確率ですから、 $\theta = .167$ で $N = 20$ ですね。計算したいと思っていることの一般的な“数式”は次のように書くことができます。

$$P(X | \theta, N)$$

そしてここでは $X = 4, \theta = .167, N = 20$ という特別なケースに興味があるわけです。

この問題を解決する議論にうつる前に、表記について一つだけ言及しておきたいことがあります。私が X がパラメータ θ と N による二項分布からランダムに生成されるという時、私は次のような表記をします。

$$X \sim \text{Binomial}(\theta, N)$$

はいはい、あなたが何を言いたいかはわかりますよ。表記法、表記法、表記法。誰がそんなものを気にするんだ、ってことですよね。表記法の話のためにここにいる読者はほとんどいないでしょうから、私は二項分布をどうやって使うのかという話に進んだ方がいいのかもしれませんね。二項分布の式は表 5.2 に書いておきましたから、それを楽しんでくれた読者もいるかもしれませんが、ほとんどの人はそんなに注意深く見なかったでしょう。この本に数式は必要ないですし、それ以上詳細について語るつもりもありませんから。その代わり、二項分布がどんなものかをあなたにお見せしたいと思います。

つまり、図 5.3 は私たちのサイコロ実験で有り得る全ての X の値についてのに二項分布の確率をプロットしたものであり、 $X = 0$ (ドクロが出ない) から $X = 20$ (全部ドクロ) までの全てについてプロットしたことになります。これは基本的な棒グラフで、私が図 5.2 に示した“ズボンの確率”と違うところはありません。横軸は起こりうる全ての事象であり、縦軸はそれらの各事象の確率だと読むことができます。ですから、20 回のうち 4 回ドクロが出るのは大体 0.20(すぐにわかるのですが、正確な答えは 0.2022036) です。言い換えると、あなたがこの実験を繰り返すと、そのうち 20% の偶然性でそうなると期待できます。

二項分布がどのように変化するかの感覚を掴んでもらうために m 、 θ と N の値を変えてから、サイコロを転がす代わりにコインフリップをやったらどうなるか想像してみましょう。今度は、私の実験は公平なコインを繰り返しフリップすることにし、私が興味を持っているアウトカムはコインが表を向いた回数だと考えます。このシナリオだと、成功確率は $\theta = 1/2$ になります。コインを $N = 20$ 回フリップするつもりだとしましょう。この実験では、成功確率を変えましたが、実験の回数は同じなわけです。こうすると私たちの二項分布はどうなるのでしょうか？ そう、図 5.4 が示すように、こうすることの主な効果は分布全体を動かすことになる、と思いますよね。オッケー、じゃあコインを $N = 100$ 回フリップしたらどうなりますか？ そう、この場合は図 5.4b のようになりますね。この分布が示すのは、大まかな中心傾向ですが、確率的な結果におけるちょっとした散らばりもあるのです。

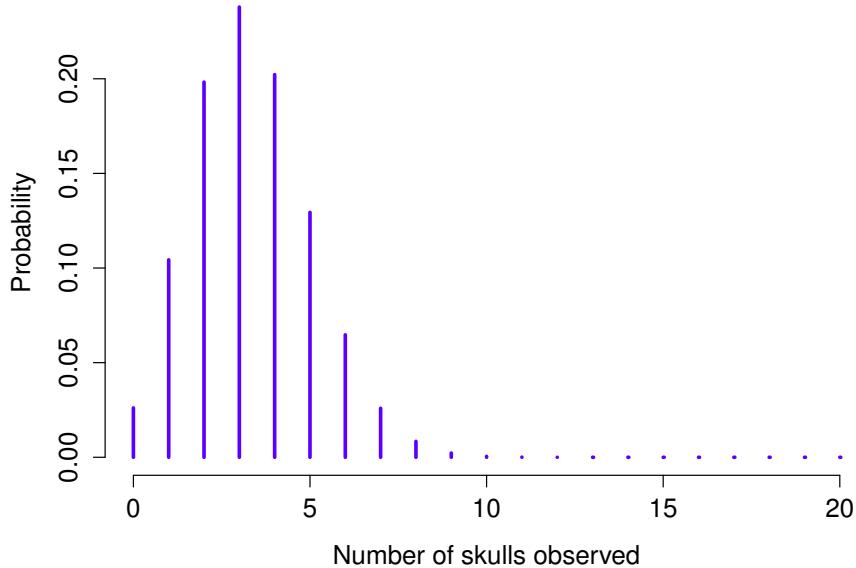


Figure 5.3 サイズパラメータが $N = 20$ で成功確率 $\theta = 1/6$ の二項分布。縦のバーそれぞれがそのアウトカムの確率を表しています(すなわち、値 X の確率)。これは確率分布なので、それぞれの確率は 0 から 1 の間にに入る数字であり、バーの高さを足し合わせると 1 にならなければなりません。

5.5 _____

正規分布

二項分布は概念的に最も簡単な分布でしたから理解しやすかったと思いますが、それが最も重要な分布だったかと言われるとそうではありません。その名誉ある称号は**正規分布**に贈られます。正規分布は“ベルカーブ”や“ガウス分布”とも言われます。正規分布は二つのパラメータをつかって表されます。すなわち、分布の平均を表す μ と、分布の標準偏差を表す σ です。

私たちがよく使う表記法で、変数 X が正規分布に従うというときは、次のように表します。

$$X \sim \text{Normal}(\mu, \sigma)$$

もちろん、これは単なる表記法に過ぎません。これは正規分布そのものについて、なんら面白いことを教えてくれるものではありません。二項分布の時のように、私はこの本に正規分布の数式を含めてはいます。というのも、統計学を学ぶどんな人にとっても、少なくともそれを目にしておくこ

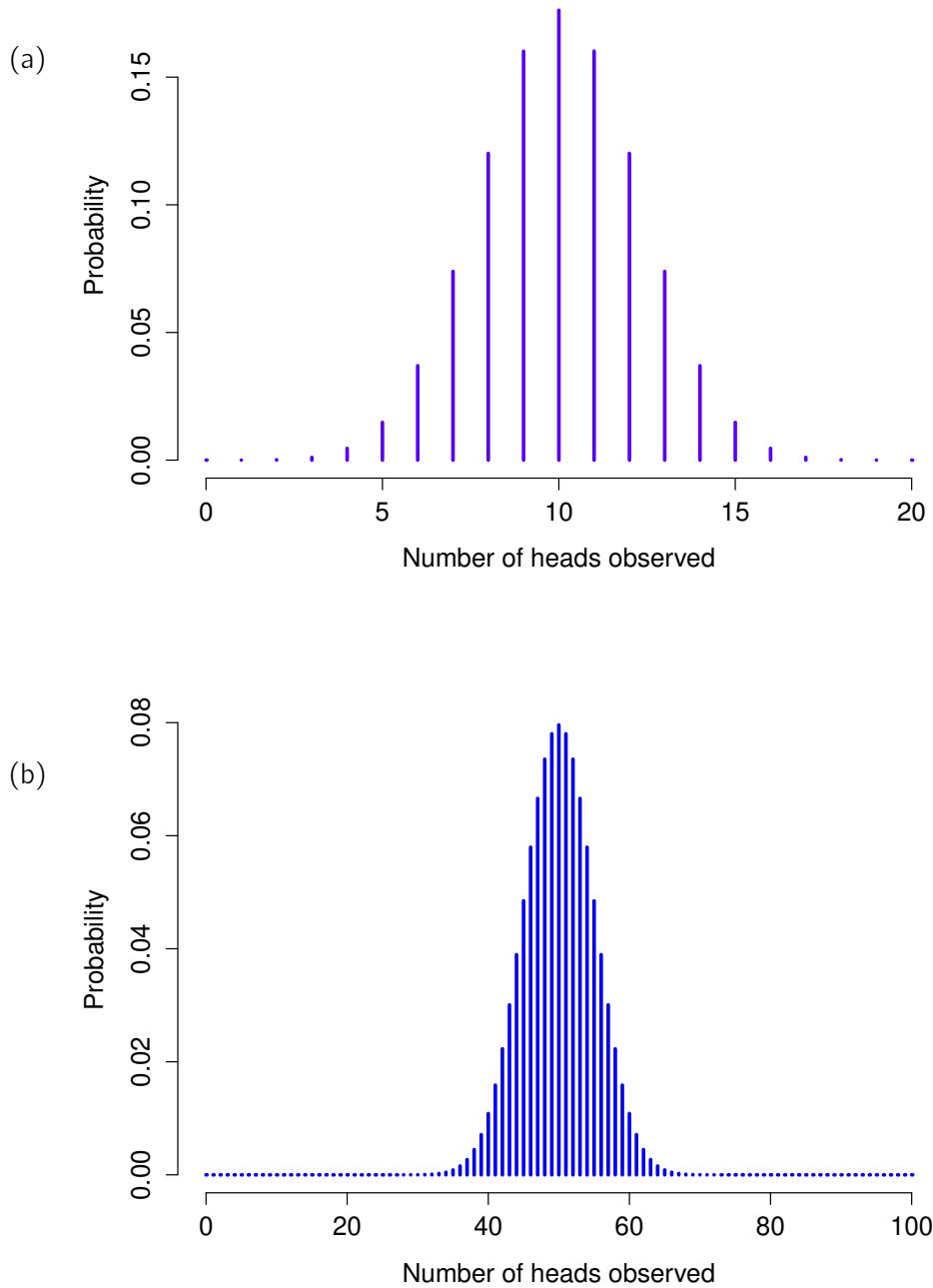


Figure 5.4 私が公平なコインをフリップするシナリオについての二項分布で、想定される確率は $\theta = 1/2$ とします。パネル a では、コインを $N = 20$ 回、パネル b では $N = 100$ 回フリップしたものです。

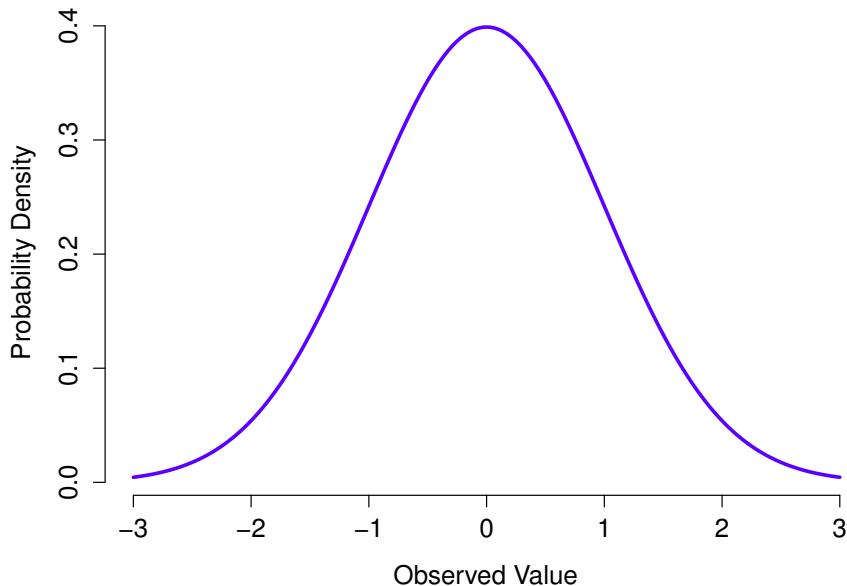


Figure 5.5 平均 $\mu = 0$ で標準偏差が $\sigma = 1$ の正規分布。 x -軸はある変数の値に対応しており、 y -軸はその値を我々がどの程度観測しやすいかを教えてくれます。ですが、 y 軸にあるのは“確率密度”であって“確率”ではないことに注意してください。ここには連続的分布ならではの、微妙でちょっと腹立たしい特徴があつて、 y 軸の振る舞いはちょっと奇妙なのです。すなわち、このカーブの高さは、実際には x の値を観測する確率を表しているわけではないのです。一方で、このカーブの高さは、どの x の値がより生じやすいか（より高いほうがそうなのですが）をあなたに教えてくれるもののです。（この面倒な詳細については、Section 5.5.1 をみてください）

とは重要だと考えるからなのですが、これは入門書でもあるのでそこにフォーカスすることはせず、表 5.2 の中に入れておくに止めておきます。

数学的側面に注目する代わりに、正規分布に従う変数が意味することの感覚を掴んでみましょう。そのために、図 5.5 にある、平均 $\mu = 0$ と標準偏差 $\sigma = 1$ の正規分布プロットを見てみましょう。“ベルカーブ”という名前の由来がわかると思います。そう、ベルみたいに見えますよね。二項分布を描いたときのプロットとは違って、図 5.5 にある正規分布の図では“ヒストグラムのような”バーの代わりにスムーズなカーブが描かれていることに注意してください。これは曖昧な選択を表しているのではなく、二項分布が離散的だったのに対し、正規分布は連続的なのです。例えば、前のセクションでやったサイコロを転がす例では、ドクロが 3 つ、4 つの確率を得ることはできましたが、3.9 個のドクロを考える、というのは不可能です。前のセクションで私が描いた図は、このことを反映していました。図 5.3 では、例えば、バーは $X = 3$ や、 $X = 4$ に位置することはありませんが、その

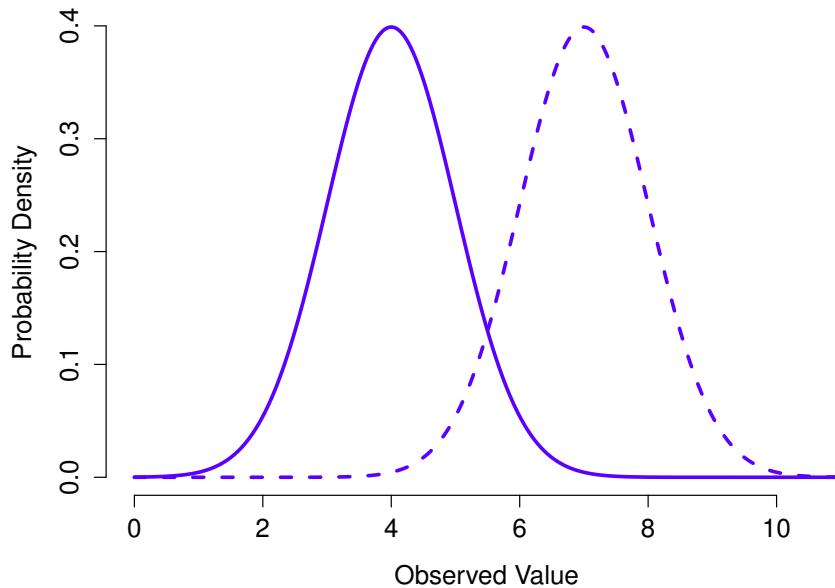


Figure 5.6 正規分布の平均を変えたら何が起こるかを描いたもの。実線は平均 $\mu = 4$ の正規分布。点線は平均 $\mu = 7$ の正規分布を表しています。どちらも、標準偏差は $\sigma = 1$ です。はたして、二つの分布は同じ形をしており、点線が右側にずれています。

間には何もありません。連続的な量というのは、この制限に当てはまらないのです。例えば、天気のことについて考えてみましょう。ある快適な春の日の温度は、23度でも、24度でも、23.9度でも、そのほかどんな間の数字でもあります。というのも、気温というのが連続変数だからです。ですから、正規分布で春の気温を記述するのがまあ適当だろう、ということになります^{*3}

これを念頭において、正規分布がどのような動きをするのか直観的につかめるかどうか、見てみましょう。まず、分布のパラメータ周りで遊んでみた時に、何が起こるかみてみたいと思います。そのため、図 5.6 に標準偏差が同じで平均が異なる正規分布をプロットしました。あなたが想像した通り、全ての分布は同じ“幅”をもっています。違いはそれらが右、あるいは左にシフトすることだけです。そのほかの特性については全て同じです。それに対して、平均を一定にしたまま標準偏差を大きくしていくと、分布の頂点は同じ場所のままですが、分布がどんどん幅広くなることが図 5.7 にみてとれますね。しかし注意して欲しいのは、分布の幅を広くした時に、頂上の高さが縮小して

^{*3} 実際には、正規分布はとても便利なので、変数が現実的に連続的でない場合であっても、それを使う傾向があります。十分なカテゴリー数があれば（例えば、質問紙におけるリッカースケールなどです）、正規分布をその近似として適用するというのが標準的な実践例になっています。あなたが思っているよりも、実戦ではそれでうまくいくのです。

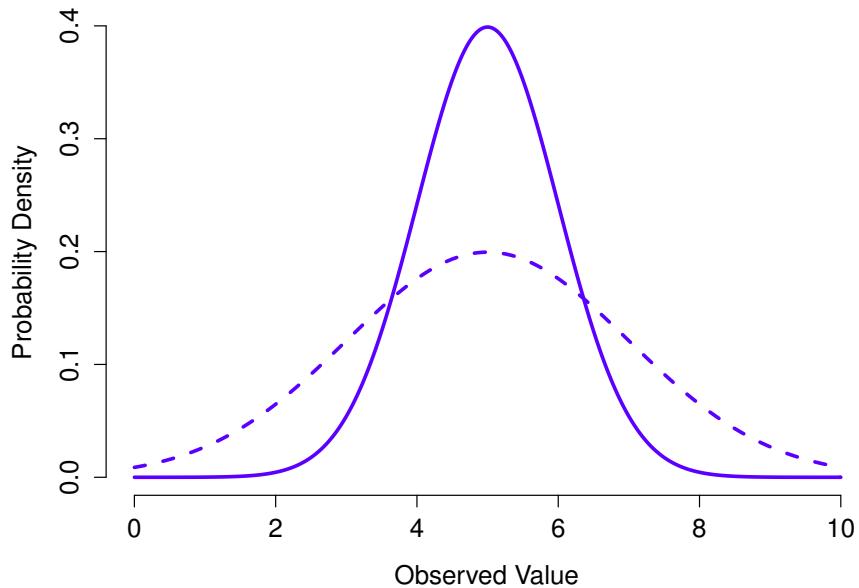


Figure 5.7 正規分布の標準偏差を変えた時に何が起こるかを描いたもの。どちらの分布も平均が $\mu = 5$ ですが、標準偏差が異なります。実線は標準偏差 $\sigma = 1$ の分布で、点線は標準偏差 $\sigma = 2$ の分布です。結果として、分布は同じスポットに“中心化”していますが、点線が実線にくらべて幅が広くなっています。

いくことです。これは起こるべくして起こることです。というのも離散的な二項分布のを描いた時に、バーの高さの合計が 1 になったと同じように、正規分布のカーブの下の領域を合計したものも 1 にならなければならないのです。次に進む前に、正規分布の重要な特徴をもう一つ、指摘しておきたいと思います。具体的な平均と標準偏差がどんな値であるかにかかわらず、平均周りの 1 標準偏差の間に全体の 68.3% が含まれるということです。同様に、平均周りの 2 標準偏差の間に全体の 95.4% が、平均周りの 3 標準偏差の間に全体の 99.7% が含まれます。このことは図 5.8 に描かれています。

5.5.1 Probability density

正規分布に関する議論について、私が触れていないことがあります。一部の入門書では、それは完全に省略されています。多分そうした方がいいのです。その“触れていないこと”というのは、統計

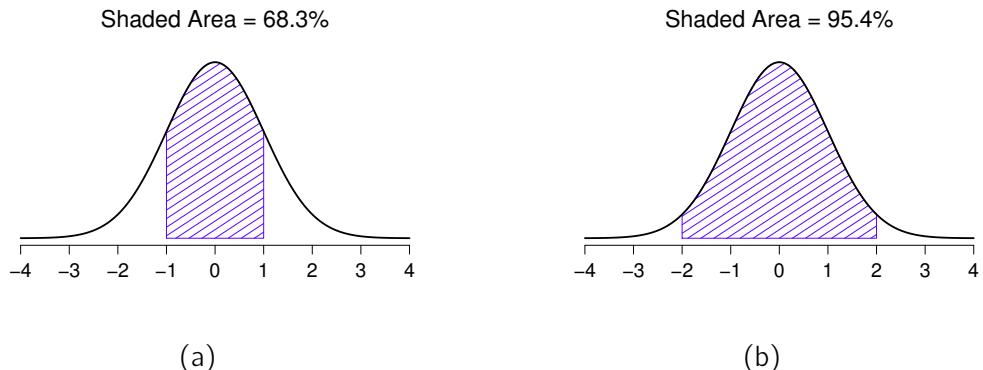


Figure5.8 カーブの下の面積は観測値が特定の範囲で得られる確率を教えてくれます。点線の正規分布は、平均 $\mu = 0$ で標準偏差 $\sigma = 1$ です。影のついた領域は、二つの重要なケースにおける“カーブの下の面積”です。パネル a では、平均周りの 1 標準偏差の中に観測値が得られる確率が 68.3% であることを見てとることができます。パネル b では、平均周り 2 標準偏差の中に観測値が得られる 確率が 95.4% であることを見てとることができます。

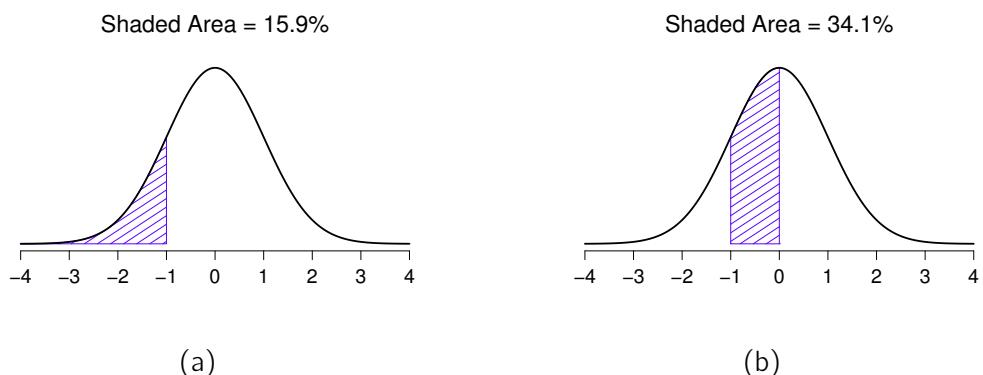


Figure5.9 “カーブの下の領域”についてのさらに二つの例です。平均より 1 標準偏差下より小さい観測値が得られる確率は 15.9% で (パネル a), 平均と平均より 1 標準偏差下の間に観測値が得られる確率は 34.1% です (パネル b)。この二つの数字を足し合わせると, $15.9\% + 34.1\% = 50\%$ になることがわかりますね。正規分布するデータでは、平均より下の観測値が得られる確率は、50% になります。そしてもちろん、このことは平均以上の観測値が得られる確率が 50% になることも意味します。

学において応用される確かに歪んだ基準に照らしてみても、奇妙で直感に反するように思えるからです。幸いにも、基本的な統計学を実行する上では、そこまで深いレベルの理解は必要になるようなことではありません。というより、基本的な領域を越えようとしたときになって初めて、そのことが重要になってくるのです。ですから、もし意味が分からなくてもそれほど恐れることはないですが、その要点を遵守することだけは心がけてください。

正規分布についての議論を通じて、一つか二つ、ちょっと意味が分からぬことがありますね。おそらく気付いたと思うのですが、図の y 軸には“確率密度”というラベルがあります。密度、ではなく。また、私が正規分布の式を書いたときに、 $P(X)$ の代わりに $p(x)$ を使っていることに気づいた人もいるかもしれません。

後々わかるのですが、ここで示されているのは実際には確率ではなく、それ以外の何かなのです。その何かを理解するためには、 X が連続変数であるということが本当は何を意味しているのかについて、少し考える時間を見る必要があります。外の気温について話をしているとしましょう。温度計はそれが 23 度あることを教えてくれますが、私は本当はそうではない、と知っています。ぴったり 23 度ではないのです。23.1 度なのかもしれません。もちろんそれが本当かどうかわからず。というのも実際には 23.09 度かもしれないのですから。しかし私が思うに… というわけです。わかりますね。本当に連続的な量に伴うトリッキーな考え方、あなたは正確にそれがどれぐらいであるかを決して知ることができない、ということです。

では、これが確率について考えるときに何をもたらすか、考えてみましょう。明日の最大気温が平均 23、標準偏差 1 の正規分布からサンプルとして得られるとしましょう。気温が正確に 23 度になる確率はどれくらいでしょう？その答えは“ゼロ”，あるいは“ほとんどゼロになるゼロに近い数字”になるでしょう。何故そうなるのかですって？それは無限に小さいダーツのために、ダーツの矢を投げようとしているようなものだからです。あなたがどれほど優れた腕前の持ち主でも、決して当たることはないでしょう。実生活においては、あなたが決して 23 度ちょうどの値を得ることがないのです。それはいつだって、23.1 度とか、22.99998 度とか、そんな感じになっているはずです。言い換えると、気温がちょうど 23 度になる確率について語るということは、全く無意味だということです。日常用語で、私はあなたに外の気温は 23 度だと言ったりします。でもそのあとで実は 22.9998 度だったということが分かっても、あなたは私を嘘つき呼ばわりしたりしないでしょう。日常用語での“23 度”というのは普通、“22.5 度から 23.5 度の間のどこか”ぐらいの意味しかないので。ですから、ちょうど 23 度である確率について尋ねることがそれほど意味のあることではないとしても、気温が 22.5 度から 23.5 度の間、あるいは 20 度と 30 度の間、もしくはそれ以外の範囲について、確率を問うことは意味があるので。

この議論のポイントは、私たちが連続変数について議論しているとき、特定の値についての確率について言及するのは意味がない、ということを明らかにしておくことです。私たちが話すことができることは、ある値についての確率は常に特定の範囲を持った値についてなのです。あなたが必要とする特定の範囲についての確率を見つけるためには、“カーブの下の領域”を計算しなければなりません。

ん。このことは既にみてきた通りで、図 5.8 の影がついた領域が表しているのは本当の確率です（例えば図 5.8a は平均周りの 1 標準偏差の観測値が得られる確率を表しています）。

オーケー、これでストーリーの一部が説明されます。私は連続的な確率分布をどのように理解すれば良いかについて（例えば、カーブの下の領域というのが鍵です），少しばかり説明してきました。しかし $p(x)$ についての数式で実際に表していたのは何でしょう？ $p(x)$ が確率を表していないことは明らかですが、ではそれは何でしょう？ $p(x)$ で表される量の名前は、**確率密度**で、先ほどの図で書いてあったカーブの高さに対応するものです。密度そのものは、それだけでは意味がありませんが、カーブの下の領域が本当の確率として常に理解できるように“工夫された仕掛け”なのです。正直にいうと、今あなたが知っておくべきことがそれです*4。

5.6 _____

そのほかの便利な分布

正規分布は統計学で最もよく使われる分布ですが（その理由については少し触れましたが）、二項分布もいろいろな目的のために使える便利なものです。しかし統計学の世界は確率分布で埋め尽くされていて、中にはふと通りがかりに出会うものがあります。特にこの本では 3 つの分布が出てきます。 t 分布、 χ^2 分布、そして F 分布です。それぞれの数式を提示しようとは思いませんし、そこまで詳細に語るつもりもないのですが、ちょっとした図をお見せしましょう。

*4 ちょっとした計算を知っている人のために、もう少し正確な説明をしておきます。確率は非負で総和が 1 になるのと同じで、確率密度も非負で積分すると 1 にならなければなりません（積分は全てのとりうる値 X に対して行われます）。 X が a と b の間に落ちる確率を計算するためには、該当する範囲の密度関数に対しての積分、 $\int_a^b p(x) dx$ を定義します。この計算を覚えていない、あるいは習ったことがないと言うのでも心配しなくて結構です。この本ではそれは必要ないんですから。

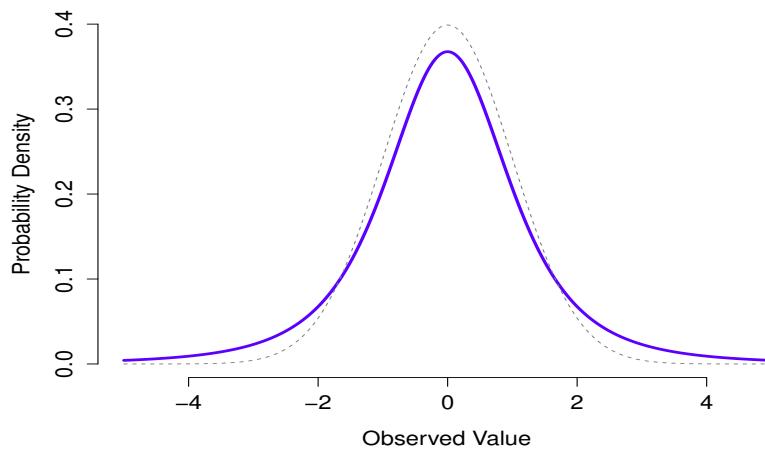


Figure 5.10 自由度 3 の t 分布 (実線)。正規分布に似ているようですが、全く同じというわけではありません。比較のために、標準正規分布を点線でプロットしました。

- **t 分布**は連続分布で正規分布によく似ています。図 5.10 を参照してください。 t 分布の“尻尾”は正規分布よりも“重い”(つまり、より外れた値まで広がっている)ことに注意してください。両者の間には重要な違いがあります。この分布は、データが実際は正規分布に従っていても、その平均と標準偏差がわからないという時に現れるものです。この分布については、第 ?? 章でまた触れることになります。
- **χ^2 分布**は様々な場面で出てくるもう一つの分布です。私たちがこの分布に出会う状況は、カテゴリカルなデータ分析 (第 8 章参照) ですが、実際に至るところで見ることができるものの一つです。数学的な意味を掘り下げていきますが (嫌いな人なんていませんよね?), なぜ χ^2 分布が至る所で見られるのかについての主たる理由は、正規分布する変数がたくさんあれば、その変数を二乗して足し合わせる (この手続きは“平方和 (sum of squares)”といいますが) と、その合計が χ^2 分布に従うからです。このことが便利であると気づくことが多いことに驚くでしょう。ともかく、ここでは χ^2 分布がどんな形なのかを見ておくことにしておきましょう。: 図 5.11.
- **F 分布**は χ^2 分布に少し似ていて、二つの χ^2 分布を比較する必要があるときに出でてきます。確かに、正気の人間で誰がそんなことをしたがるのかと思えますが、実際のデータ分析においてはとても重要であることがわかります。 χ^2 の話をした時に、“平方和”を使う際の大変な分布だと言ったことを覚えていませんか? そうです、もしあなたが二つの異なる“平方和”を比較したいと思ったら、おそらく F 分布について話をしなければならなくなるでしょう。もちろん私は平方和について、まだ何の例も挙げていませんが、第 ?? 章で触れることになります。

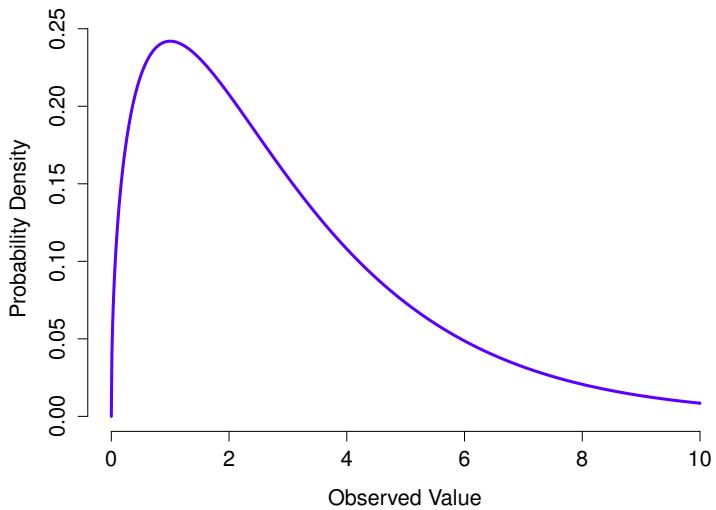


Figure5.11 自由度 3 の χ^2 分布。観測値は 0 より大きくなければなりませんし、この分布は少し歪んでいることに注意が必要です。そこにカイ二乗分布の特徴があります。

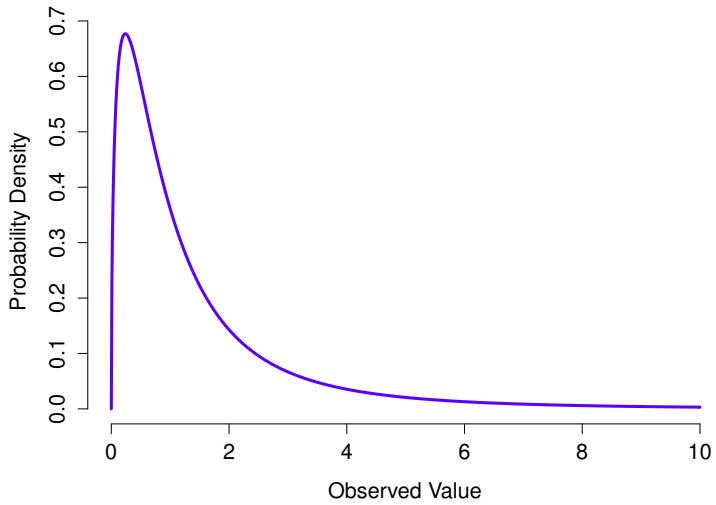


Figure5.12 自由度 3 と 5 の F 分布。定性的にいうなら、カイ二乗分布に少し似ているように見えますが、一般的には全く似ていません。

その時 F 分布について説明していくことになるでしょう。そうそう、図 5.12 も見ておいてくださいね。

さて、このセクションを切り上げる時間がきたようです。ここでは三つの新しい分布を見ました。すなわち χ^2 分布、 t 分布、そして F 分布です。これらは全て連続分布で、何も正規分布に密接に関係しています。ここでの主な目的は、これらの分布が全て互いに、また正規分布と深いレベルで関係していることを理解することです。この本の後の方で、正規分布しているデータ、あるいは少なくとも正規分布していると仮定できるデータを扱っていきます。ここで知っておいて欲しいことはそれだけです。もしあなたのデータが正規分布していると仮定するなら、データ分析を始める時にあちこちで χ^2 分布や t 分布、 F 分布が顔を出してきても、驚くことはありません。

5.7

要約

この章では確率について考えてきました。確率が何を意味するのか、なぜ統計学者はそれが意味することに同意できないのかについて論じました。確率が従わなければならないルールについても話しました。そして確率分布の概念を導入し、統計学者がよく使うより重要な確率分布をいくつか導入するのに、この章のかなりの部分を費やしました。セクションごとに分解すると、次のようになっています。

- 確率理論と統計 (セクション 5.1)
- 頻度主義者とベイズ主義者それぞれの確率の見方 (セクション 5.2)
- 確率理論の基礎 (セクション 5.3)
- 二項分布 (セクション 5.4), 正規分布 (セクション 5.5), そのほかの分布 (セクション 5.6)

あなたの想像通り、私の取材した範囲は網羅的ではありません。確率理論は数学の中でも大きな分野であり、統計学やデータ分析への応用からは全体的に別れたものになっています。ですから、このテーマで書かれた本は何千とあるし、大学では一般に確率論を専門に扱う複数のクラスを提供しています。標準的な確率分布についての解説作業という“単純な”ことでさえも、大きなトピックになってしまふのです。この章で私は 5 つの確率分布を紹介しましたが、私の本棚には 45 章からなる“統計的分布”(Evans2000) という本があって、そこにはもっとたくさんの確率分布が含まれています。あなたにとっては幸運なことかもしれません、必要なのはこのごく一部です。表に出て実世界でデータ分析をするときに、このたくさんの確率分布を知っておく必要はありませんし、この本にあるような分布を必要とすることもないと思いますが、他にも多くの確率分布があることを知っておいて損はありません。

この最後の点から考えると、この章全体がちょっとした余談みたいになりますね。学部生用の心理学のクラスで統計をやる場合はほとんど、この内容については素早く通り過ぎるものですが（私がそうしていることも自覚しています）、より専門的なクラスではこの領域の基本的な基礎をおさらいすることを“忘れて”しまわれるすることがよくあります。大学心理学者のほとんどは確率と確率密度の違いを知りませんし、ベイジアンと頻度主義者の確率の間の違いに気付いた人も最近までほとんどいませんでした。しかし、私はこれらを応用の前に知っておくことが重要だと考えています。例えば、私たちが推測的推論をするときに“許される”言い方についてのルールがたくさんあり、それらの多くは恣意的で奇妙なものに見えます。ところが、ベイジアンと頻度主義者の違いがあることを理解すれば、すぐにそれらが意味をなすのです。同様に、?? 章では t 検定について説明しますが、もしあなたが t 検定の数理を理解したいと思うのなら、 t 分布がどういう見え方をするものなのかを知っていると役立つことでしょう。そういう気付きを得てくれるよう、願っています。

6. Estimating unknown quantities from a sample

At the start of the last chapter I highlighted the critical distinction between *descriptive statistics* and *inferential statistics*. As discussed in Chapter 3, the role of descriptive statistics is to concisely summarise what we *do* know. In contrast, the purpose of inferential statistics is to “learn what we do not know from what we do”. Now that we have a foundation in probability theory we are in a good position to think about the problem of statistical inference. What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two “big ideas”: estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but I’m going to witter on about sampling theory first because estimation theory doesn’t make sense until you understand sampling. As a consequence, this chapter divides naturally into two parts Sections 6.1 through 6.3 are focused on sampling theory, and Sections 6.4 and 6.5 make use of sampling theory to discuss how statisticians think about estimation.

6.1

Samples, populations and sampling

In the prelude to Part III I discussed the riddle of induction and highlighted the fact that *all* learning requires you to make assumptions. Accepting that this is true, our first task to come up with some fairly general assumptions about data that make sense. This is where **sampling theory** comes in. If probability theory is the foundations upon which all statistical theory builds, sampling theory is the frame around which you can build the rest of the house. Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about “making inferences” the way statisticians think about it we need to be a bit more explicit about what it is that we’re drawing inferences *from* (the sample) and what it is that we’re

drawing inferences *about* (the population).

In almost every situation of interest what we have available to us as researchers is a **sample** of data. We might have run experiment with some number of participants, a polling company might have phoned some number of people to ask questions about voting intentions, and so on. In this way the data set available to us is finite and incomplete. We can't possibly get every person in the world to do our experiment, for example a polling company doesn't have the time or the money to ring up every voter in the country. In our earlier discussion of descriptive statistics (Chapter 3) this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarising and graphing that sample. This is about to change.

6.1.1 Defining a population

A sample is a concrete thing. You can open up a data file and there's the data from your sample. A **population**, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about and is generally *much* bigger than the sample. In an ideal world the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses with the data does depend on the population about which you want to make statements.

Sometimes it's easy to state the population of interest. For instance, in the "polling company" example that opened the chapter the population consisted of all voters enrolled at the time of the study, millions of people. The sample was a set of 1000 people who all belong to that population. In most studies the situation is much less straightforward. In a typical psychological experiment determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students at the University of Adelaide?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

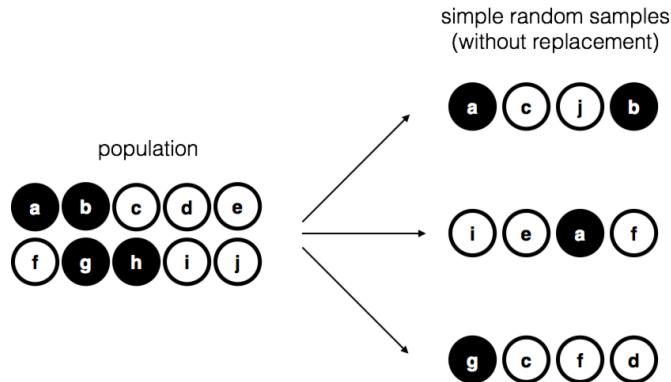


Figure6.1 Simple random sampling without replacement from a finite population

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest. As another example, consider the Wellesley-Croker game that we discussed in the prelude. The sample here is a specific sequence of 12 wins and 0 losses for Wellesley. What is the population?

- All outcomes until Wellesley and Croker arrived at their destination?
- All outcomes if Wellesley and Croker had played the game for the rest of their lives?
- All outcomes if Wellseley and Croker lived forever and played the game until the world ran out of hills?
- All outcomes if we created an infinite set of parallel universes and the Wellesely/Croker pair made guesses about the same 12 hills in each universe?

Again, it's not obvious what the population is.

6.1.2 Simple random samples

Irrespective of how I define the population, the critical point is that the sample is a subset of the population and our goal is to use our knowledge of the sample to draw inferences about the properties of the population. The relationship between the two depends on the *procedure* by which the sample was selected. This procedure is referred to as a **sampling method** and it is important to understand why it matters.

To keep things simple, let's imagine that we have a bag containing 10 chips. Each chip has a

unique letter printed on it so we can distinguish between the 10 chips. The chips come in two colours, black and white. This set of chips is the population of interest and it is depicted graphically on the left of Figure 6.1. As you can see from looking at the picture there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the a chip (black), then the c chip (white), then j (white) and then finally b (black). If you wanted you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 6.1. Each time you get different results but the procedure is identical in each case. The fact that the same procedure can lead to different results each time we refer to as a *random* process.^{*1} However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did *not* put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag and decided to pull out four black chips without putting any of them back in the bag. This *biased* sampling scheme is depicted in Figure 6.2. Now consider the evidential value of seeing 4 black chips and 0 white chips. Clearly it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips then a sample that consists of only black chips doesn't tell you very much about the population! For this reason statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis *much* easier.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample **with replacement**. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure 6.3.

^{*1}The proper mathematical definition of randomness is extraordinarily technical, and way beyond the scope of this book. We'll be non-technical here and say that a process has an element of randomness to it whenever it is possible to repeat the process and get different answers each time.

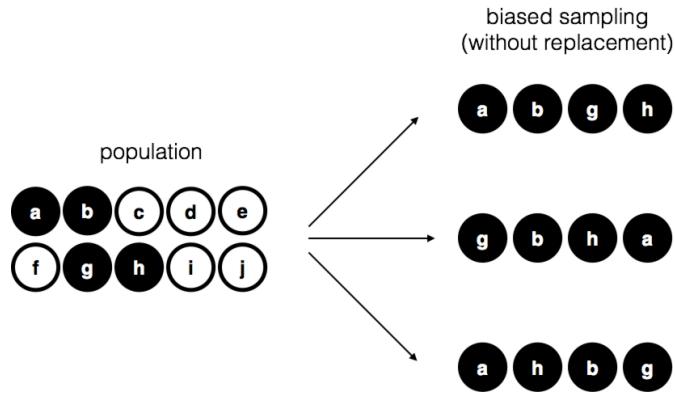


Figure6.2 Biased sampling without replacement from a finite population

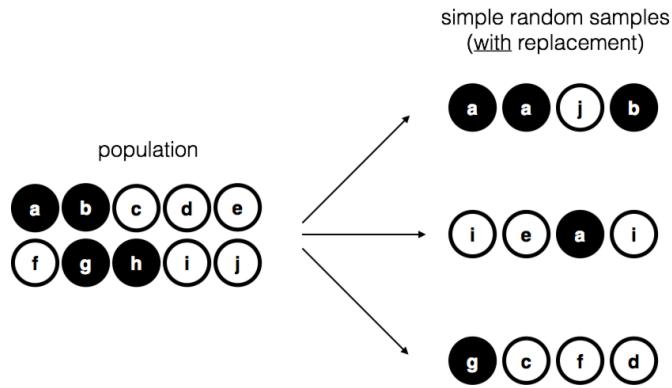


Figure6.3 Simple random sampling *with* replacement from a finite population

In my experience, most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample *with* replacement. In real life this very rarely matters. If the population of interest is large (e.g., has more than 10 entities!) the difference between sampling with- and without- replacement is too small to be concerned with. The difference between simple random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

6.1.3 Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost

impossible to obtain a simple random sample from most populations of interest. When I run experiments I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalise to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones.

- *Stratified sampling*. Suppose your population is (or can be) divided into several different sub-populations, or *strata*. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient than simple random sampling, especially when some of the sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two^{*2} strata (schizophrenic and not-schizophrenic) and then sample an equal number of people from each group. If you selected people randomly you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of stratified sampling is referred to as *oversampling* because it makes a deliberate attempt to over-represent rare groups.
- *Snowball sampling* is a technique that is especially useful when sampling from a “hidden” or hard to access population and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey the participants are asked to provide contact details for other people who might want to participate. In stage 2 those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this issue. If you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that

^{*2}Nothing in life is that simple. There's not an obvious division of people into binary categories like “schizophrenic” and “not schizophrenic”. But this isn't a clinical psychology text so please forgive me a few simplifications here and there.

mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent *before* contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.

- *Convenience sampling* is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects. First, reliance on undergraduate psychology students automatically means that your data are restricted to a single sub-population. Second, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students and not a randomly selected subset. In real life most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

6.1.4 **How much does it matter if you don't have a simple random sample?**

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it *can* matter if your data are not a simple random sample. Just think about the difference between Figures 6.1 and 6.2. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually *know* what the bias is because you created it deliberately, often to *increase* the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, and not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in *every* respect, we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am

able to sample randomly from the Australian population. I want to generalise my results to the population of all living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being “born on a Monday” has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why “being Australian” might matter. Australia is a wealthy, industrialised country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to “take a test”, a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, “test taking” style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials much more than people who haven’t grown up in a similar environment. This could therefore lead to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. First, when designing your own studies, it’s important to think about what population you care about and try hard to sample in a way that is appropriate to that population. In practice, you’re usually forced to put up with a “sample of convenience” (e.g., psychology lecturers sample psychology students because that’s the least expensive way to collect data, and our coffers aren’t exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be. Second, if you’re going to criticise someone else’s study because they’ve used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to *how* this might have distorted the results.

6.1.5 **Population parameters and sample statistics**

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample, let’s consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist a population might be a group of people. To an ecologist a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world. Statisticians, however, are a funny lot. On the one hand, they *are* interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalise our abstract theoretical ideas in terms of concrete measurements (Section ??), statisticians operationalise the concept of a “population” in terms of mathematical objects that they know how

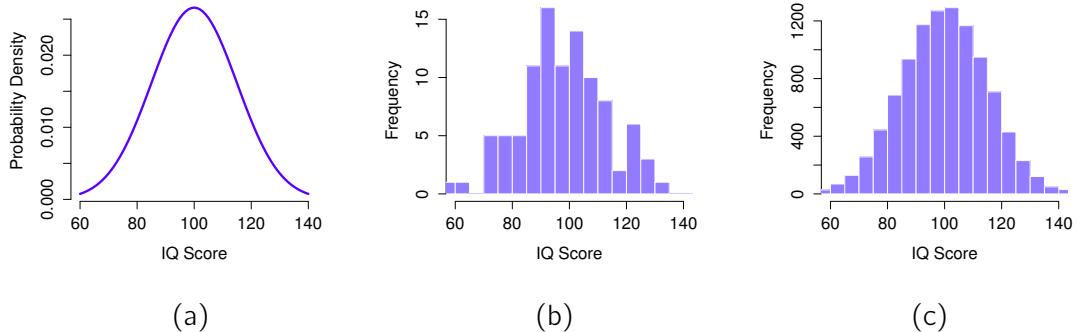


Figure 6.4 The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and in panel c we have a sample of 10,000 observations.

.....

to work with. You've already come across these objects in Chapter 5. They're called probability distributions.

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist the population of interest is a group of actual humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure 6.4a. IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the **population parameters** because they are characteristics of the entire population. That is, we say that the population mean μ is 100 and the population standard deviation σ is 15.

Now suppose I run an experiment. I select 100 people at random and administer an IQ test, giving me a simple random sample from the population. My sample would consist of a collection of numbers like this:

106 101 98 80 74 ... 107 72 100

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample I get something like the one shown in Figure 6.4b. As you can see, the histogram is *roughly* the right shape but it's a very crude approximation to the true population distribution shown in Figure 6.4a. When I calculate the mean of my sample, I get a number that is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in my sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These **sample statistics** are properties of my data set, and although they are

fairly similar to the true population values they are not the same. In general, sample statistics are the things you can calculate from your data set and the population parameters are the things you want to learn about. Later on in this chapter I'll talk about how you can estimate population parameters using your sample statistics (Section 6.4) and how to work out how confident you are in your estimates (Section 6.5) but before we get to that there's a few more ideas in sampling theory that you need to know about.

6.2

The law of large numbers

In the previous section I showed you the results of one fictitious IQ experiment with a sample size of $N = 100$. The results were somewhat encouraging as the true population mean is 100 and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQs of 10,000 people. We can simulate the results of this experiment using JASP. The [IQsim.jasp](#) file is a JASP data file. In this file I have generated 10,000 random numbers sampled from a normal distribution for a population with `mean = 100` and `sd = 15`. By the way, I did this entirely within JASP computing a new variable using the R code `rnorm(10000, 100, 15)`. A histogram and density plot shows that this larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics. The mean IQ for the larger sample turns out to be [100.107](#) and the standard deviation is [14.995](#). These values are now very close to the true population. See Figure 6.5

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli, one of the founders of probability theory, formalised this idea back in 1713 he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal

(Stigler1986)

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct. It really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is... an average), so let's look at that. When applied to the sample mean what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean. Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$), the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).^{*3}

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set the sample statistics that we calculate from it will be wrong, but the law of large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

6.3

Sampling distributions and the central limit theorem

The law of large numbers is a very powerful tool but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a "long run guarantee". In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life.

^{*3}Technically, the law of large numbers pertains to any sample statistic that can be described as an average of independent quantities. That's certainly true for the sample mean. However, it's also possible to write many other sample statistics as averages of one form or another. The variance of a sample, for instance, can be rewritten as a kind of average and so is subject to the law of large numbers. The minimum value of a sample, however, cannot be written as an average of anything and is therefore not governed by the law of large numbers.

*[The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again. (**Keynes1923**)*

As in economics, so too in psychology and statistics. It is not enough to know that we will eventually arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my *actual* data set has a sample size of $N = 100$. In real life, then, we must know something about the behaviour of the sample mean when it is calculated from a more modest data set!

6.3.1 Sampling distribution of the mean

With this in mind, let's abandon the idea that our studies will have sample sizes of 10,000 and consider instead a very modest experiment indeed. This time around we'll sample $N = 5$ people and measure their IQ scores. As before, I can simulate this experiment in JASP by modifying the `rnorm` function that was used to generate the `IQsim` data column. If you double-click on the f_x label beside `IQsim`, JASP will open up the 'Computed Column' dialog, which contains the R code `rnorm(10000, 100, 15)`. Since I only need 5 participant IDs this time, I simply need to change 10000 to 5 and then click 'Compute column' (see Figure 6.6). These are the five numbers that JASP generated for me (yours will be different!). I rounded to the nearest whole number for convenience:

124 74 87 86 109

The mean IQ in this sample turns out to be exactly 96. Not surprisingly, this is much less accurate than the previous experiment. Now imagine that I decided to **replicate** the experiment. That is, I repeat the procedure as closely as possible and I randomly sample 5 new people and measure their IQ. Again, JASP allows me to simulate the results of this procedure, and generates these five numbers:

91 125 104 106 109

This time around, the mean IQ in my sample is 107. If I repeat the experiment 10 times I obtain the results shown in Table 6.1, and as you can see the sample mean varies from one replication to the next.

Now suppose that I decided to keep going in this fashion, replicating this "five IQ scores" exper-

Table 6.1 Ten replications of the IQ experiment, each with a sample size of $N = 5$.

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

.....

iment over and over again. Every time I replicate the experiment I write down the sample mean. Over time, I'd be amassing a new data set, in which every experiment generates a single data point. The first 10 observations from my data set are the sample means listed in Table 6.1, so my data set starts out like this:

96.0 107.0 101.6 103.8 104.4 ...

What if I continued like this for 10,000 replications, and then drew a histogram. Well that's exactly what I did, and you can see the results in Figure 6.7. As this picture illustrates, the average of 5 IQ scores is usually between 90 and 110. But more importantly, what it highlights is that if we replicate an experiment over and over again, what we end up with is a *distribution* of sample means! This distribution has a special name in statistics, it's called the **sampling distribution of the mean**.

Sampling distributions are another important theoretical idea in statistics, and they're crucial for understanding the behaviour of small samples. For instance, when I ran the very first "five IQ scores" experiment, the sample mean turned out to be 96. What the sampling distribution in Figure 6.7 tells us, though, is that the "five IQ scores" experiment is not very accurate. If I repeat the experiment, the sampling distribution tells me that I can expect to see a sample mean anywhere between 80 and 120.

6.3.2 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 6.8. 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.

6.3.3 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 6.9, 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 6.10. 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 6.10b 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 6.10a. When I increase the sample size to $N = 4$, the sampling distribution of the mean is very close to normal (Figure 6.10c), 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

*⁴As 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.

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.

6.4

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

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?

6.4.1 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 population 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 use 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

^{*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*.

handy table:

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

6.4.2 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 be 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 6.11. 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 6.9b 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 6.12. 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 3.2, 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 3.2. 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

^aOkay, 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

6.5

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 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

^{*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.

uncertainty that attaches to our estimate. It's not enough to be able guess that, say, the mean IQ of undergraduate 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 6.3.3) that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution Section 5.5 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:

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

where the SEM is equal to σ/\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:

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

What this is telling us 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.

6.5.1 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 6.4 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.

6.5.2 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 5.2, 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 6.13, 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 ??.

6.5.3 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:

$$\text{Lower 95\% CI} = 100.107 - (1.96 * 0.150) = 99.813$$

$$\text{Upper 95\% CI} = 100.107 + (1.96 * 0.150) = 100.401$$

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.

6.6 _____

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 [6.1](#))
- Statistical theory of sampling: the law of large numbers (Section [6.2](#)), sampling distributions and the central limit theorem (Section [6.3](#)).
- Estimating means and standard deviations (Section [6.4](#))
- Estimating a confidence interval (Section [6.5](#))

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.

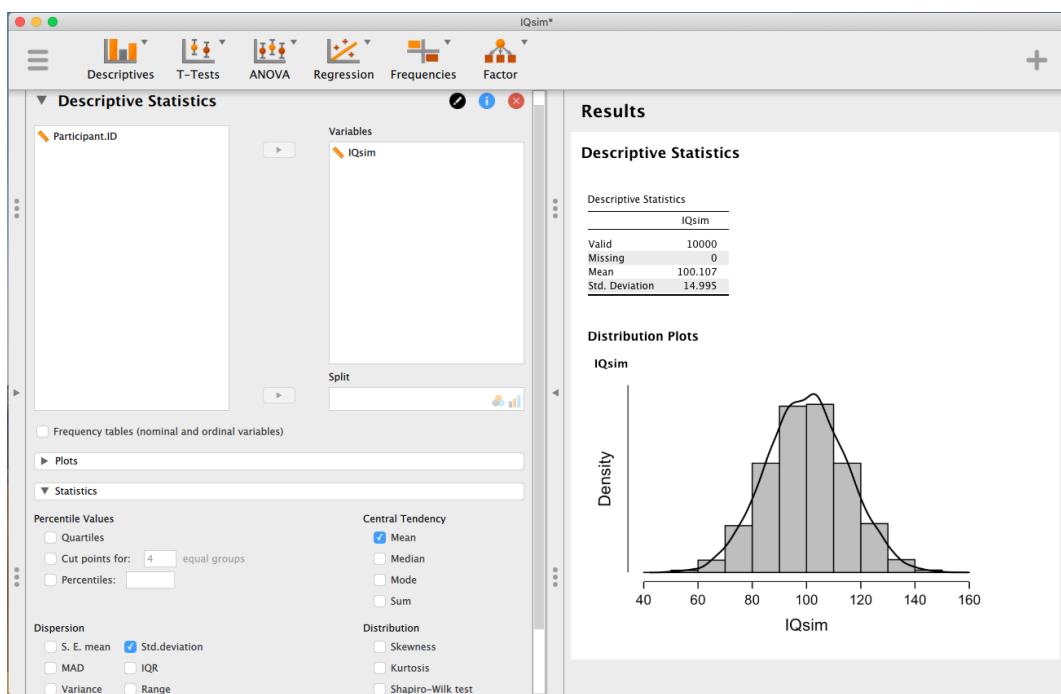


Figure6.5 A random sample drawn from a normal distribution using JASP

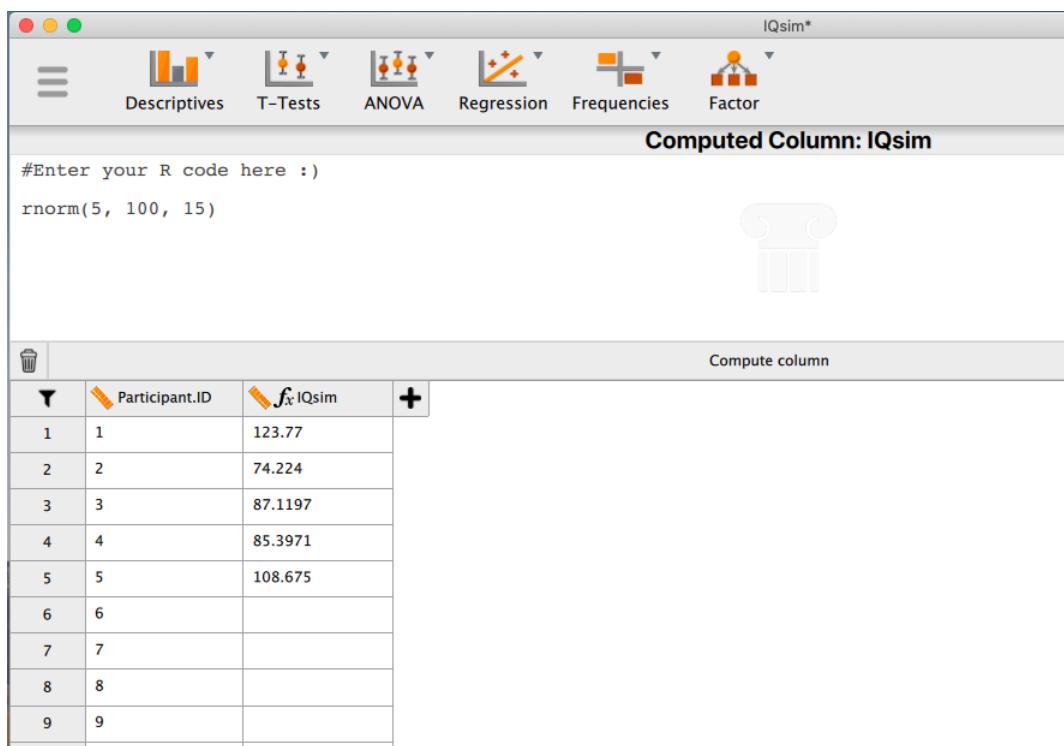


Figure6.6 Using JASP to draw a random sample of 5 from a normal distribution with $\mu = 100$ and $\sigma = 15$.

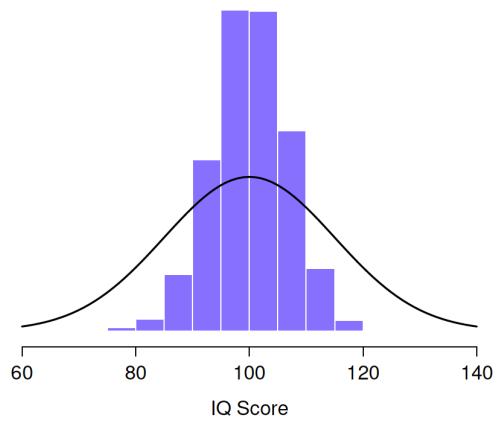


Figure6.7 The sampling distribution of the mean for the “five IQ scores experiment”. If you sample 5 people at random and calculate their *average* IQ you’ll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

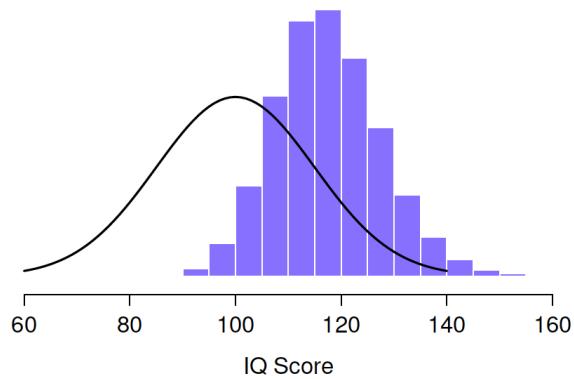


Figure6.8 The sampling distribution of the *maximum* for the “five IQ scores experiment”. If you sample 5 people at random and select the one with the highest IQ score you’ll probably see someone with an IQ between 100 and 140.

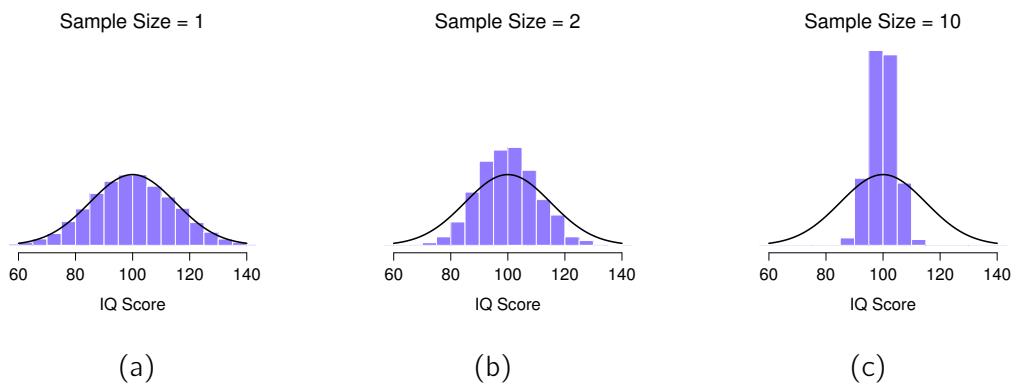
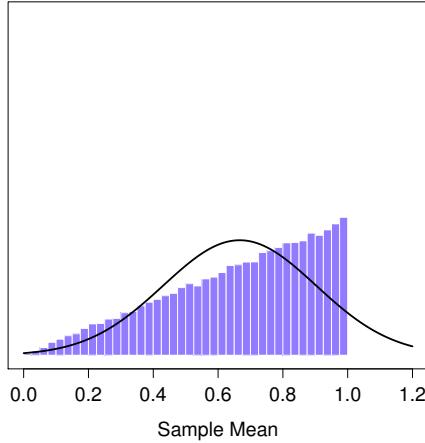


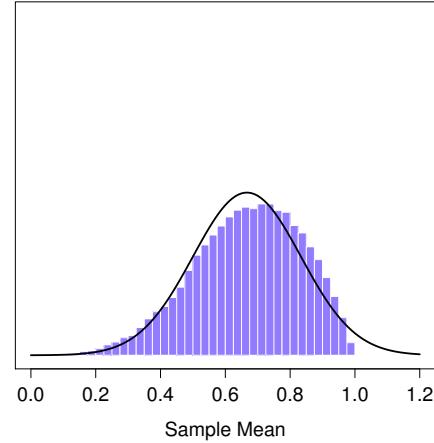
Figure 6.9 An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling 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.

Sample Size = 1



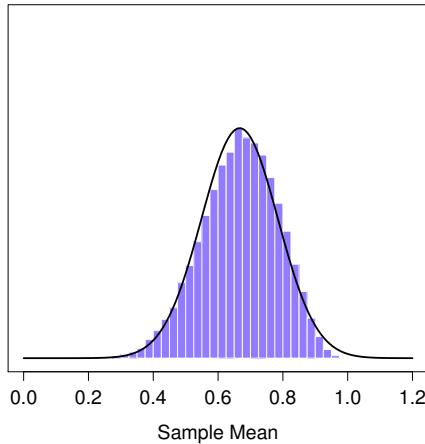
(a)

Sample Size = 2



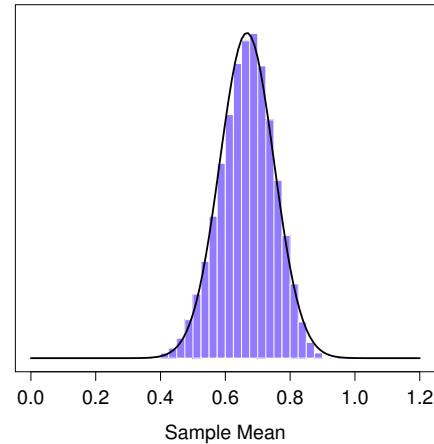
(b)

Sample Size = 4



(c)

Sample Size = 8



(d)

Figure 6.10 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.

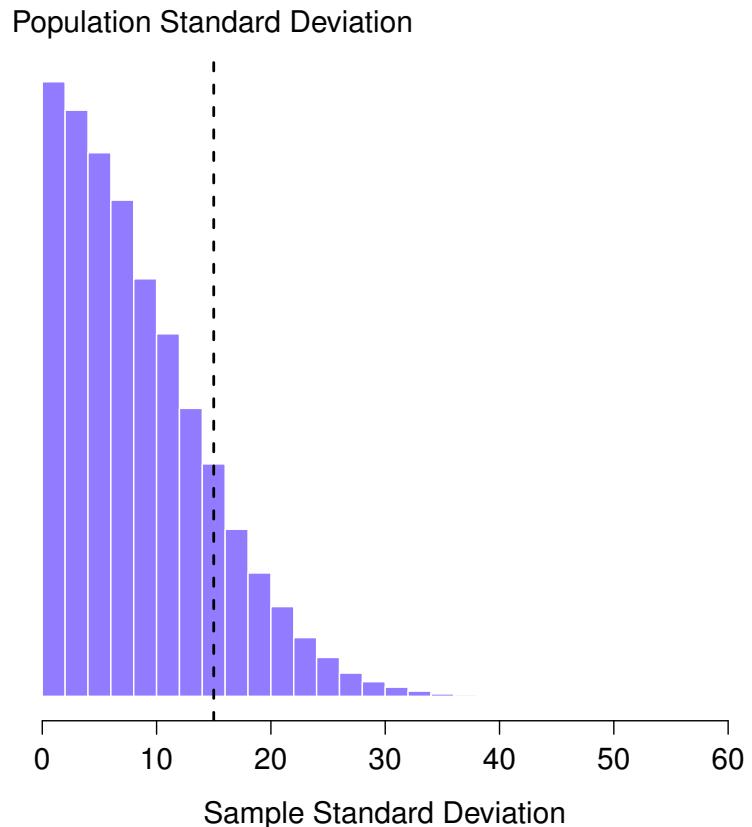


Figure 6.11 The sampling distribution of the sample standard deviation for a “two IQ scores” experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this. 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.

.....

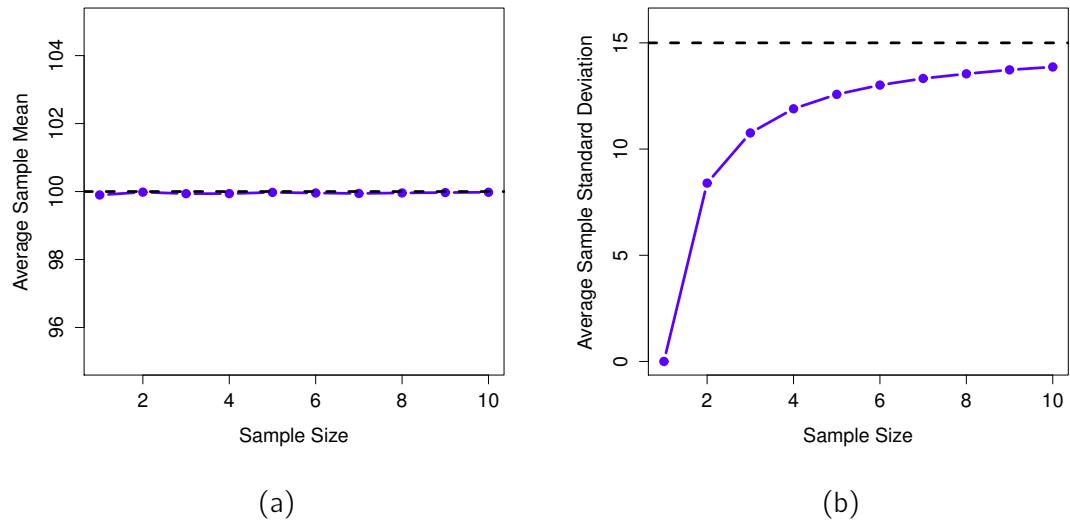


Figure 6.12 An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each, 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.

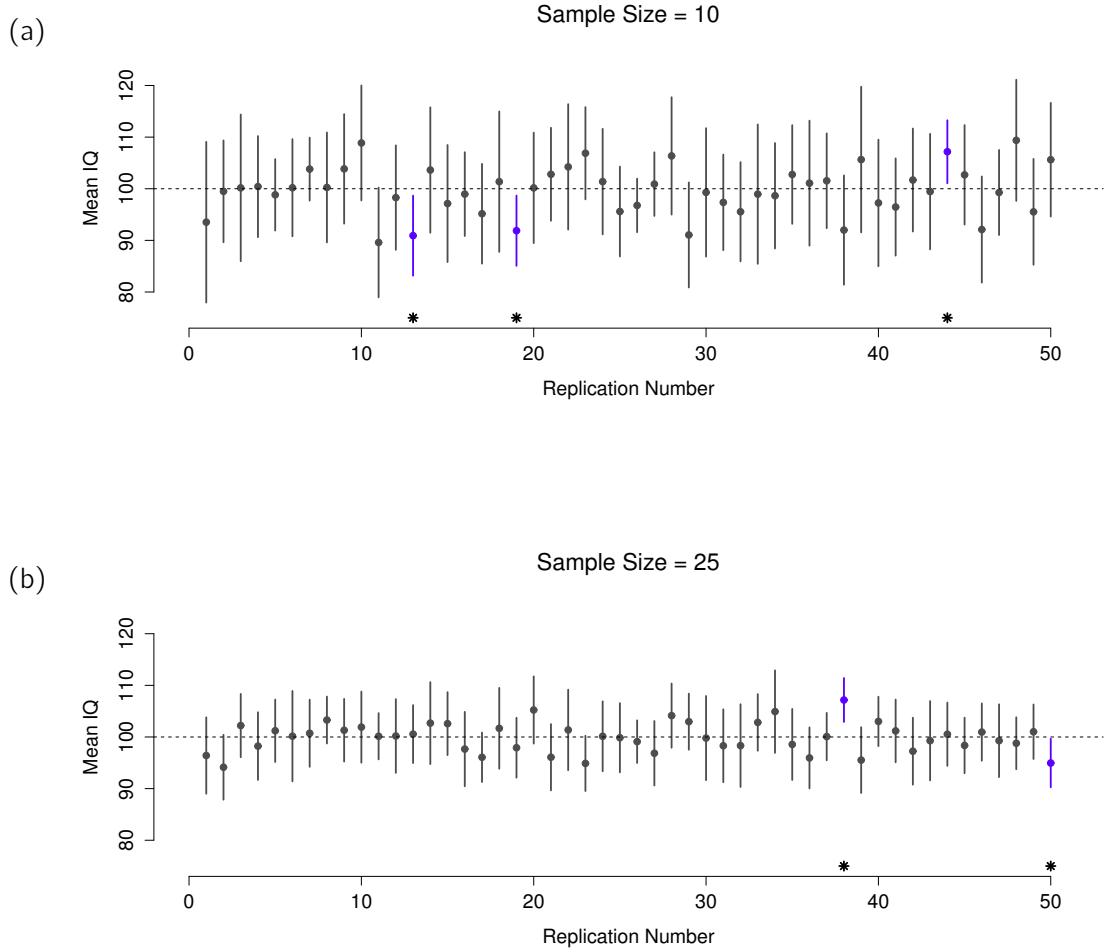


Figure 6.13 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true 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.

7. Hypothesis testing

The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience. This process, however, has no logical foundation but only a psychological one. It is clear that there are no grounds for believing that the simplest course of events will really happen. It is an hypothesis that the sun will rise tomorrow: and this means that we do not know whether it will rise.

– Ludwig Wittgenstein^{*1}

In the last chapter I discussed the ideas behind estimation, which is one of the two “big ideas” in inferential statistics. It’s now time to turn our attention to the other big idea, which is *hypothesis testing*. In its most abstract form, hypothesis testing is really a very simple idea. The researcher has some theory about the world and wants to determine whether or not the data actually support that theory. However, the details are messy and most people find the theory of hypothesis testing to be the most frustrating part of statistics. The structure of the chapter is as follows. First, I’ll describe how hypothesis testing works in a fair amount of detail, using a simple running example to show you how a hypothesis test is “built”. I’ll try to avoid being too dogmatic while doing so, and focus instead on the underlying logic of the testing procedure.^{*2} Afterwards, I’ll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

^{*1}The quote comes from Wittgenstein’s (1922) text, *Tractatus Logico-Philosophicus*.

^{*2}A technical note. The description below differs subtly from the standard description given in a lot of introductory texts. The orthodox theory of null hypothesis testing emerged from the work of Sir Ronald Fisher and Jerzy Neyman in the early 20th century; but Fisher and Neyman actually had very different views about how it should work. The standard treatment of hypothesis testing that most texts use is a hybrid of the two approaches. The treatment here is a little more Neyman-style than the orthodox view, especially as regards the p value.

7.1

A menagerie of hypotheses

Eventually we all succumb to madness. For me, that day will arrive once I'm finally promoted to full professor. Safely ensconced in my ivory tower, happily protected by tenure, I will finally be able to take leave of my senses (so to speak) and indulge in that most thoroughly unproductive line of psychological research, the search for extrasensory perception (ESP).^{*3}

Let's suppose that this glorious day has come. My first study is a simple one in which I seek to test whether clairvoyance exists. Each participant sits down at a table and is shown a card by an experimenter. The card is black on one side and white on the other. The experimenter takes the card away and places it on a table in an adjacent room. The card is placed black side up or white side up completely at random, with the randomisation occurring only after the experimenter has left the room with the participant. A second experimenter comes in and asks the participant which side of the card is now facing upwards. It's purely a one-shot experiment. Each person sees only one card and gives only one answer, and at no stage is the participant actually in contact with someone who knows the right answer. My data set, therefore, is very simple. I have asked the question of N people and some number X of these people have given the correct response. To make things concrete, let's suppose that I have tested $N = 100$ people and $X = 62$ of these got the answer right. A surprisingly large number, sure, but is it large enough for me to feel safe in claiming I've found evidence for ESP? This is the situation where hypothesis testing comes in useful. However, before we talk about how to *test* hypotheses, we need to be clear about what we mean by hypotheses.

7.1.1 Research hypotheses versus statistical hypotheses

The first distinction that you need to keep clear in your mind is between research hypotheses and statistical hypotheses. In my ESP study my overall scientific goal is to demonstrate that clairvoyance exists. In this situation I have a clear research goal: I am hoping to discover evidence for ESP. In other situations I might actually be a lot more neutral than that, so I might say

^{*3}My apologies to anyone who actually believes in this stuff, but on my reading of the literature on ESP it's just not reasonable to think this is real. To be fair, though, some of the studies are rigorously designed, so it's actually an interesting area for thinking about psychological research design. And of course it's a free country so you can spend your own time and effort proving me wrong if you like, but I wouldn't think that's a terribly practical use of your intellect.

that my research goal is to determine whether or not clairvoyance exists. Regardless of how I want to portray myself, the basic point that I'm trying to convey here is that a research hypothesis involves making a substantive, testable scientific claim. If you are a psychologist then your research hypotheses are fundamentally *about* psychological constructs. Any of the following would count as **research hypotheses**:

- *Listening to music reduces your ability to pay attention to other things.* This is a claim about the causal relationship between two psychologically meaningful concepts (listening to music and paying attention to things), so it's a perfectly reasonable research hypothesis.
- *Intelligence is related to personality.* Like the last one, this is a relational claim about two psychological constructs (intelligence and personality), but the claim is weaker: correlational not causal.
- *Intelligence is speed of information processing.* This hypothesis has a quite different character. It's not actually a relational claim at all. It's an ontological claim about the fundamental character of intelligence (and I'm pretty sure it's wrong). It's worth expanding on this one actually. It's usually easier to think about how to construct experiments to test research hypotheses of the form "does X affect Y?" than it is to address claims like "what is X?" And in practice what usually happens is that you find ways of testing relational claims that follow from your ontological ones. For instance, if I believe that intelligence *is* speed of information processing in the brain, my experiments will often involve looking for relationships between measures of intelligence and measures of speed. As a consequence most everyday research questions do tend to be relational in nature, but they're almost always motivated by deeper ontological questions about the state of nature.

Notice that in practice, my research hypotheses could overlap a lot. My ultimate goal in the ESP experiment might be to test an ontological claim like "ESP exists", but I might operationally restrict myself to a narrower hypothesis like "Some people can 'see' objects in a clairvoyant fashion". That said, there are some things that really don't count as proper research hypotheses in any meaningful sense:

- *Love is a battlefield.* This is too vague to be testable. Whilst it's okay for a research hypothesis to have a degree of vagueness to it, it has to be possible to operationalise your theoretical ideas. Maybe I'm just not creative enough to see it, but I can't see how this can be converted into any concrete research design. If that's true then this isn't a scientific research hypothesis, it's a pop song. That doesn't mean it's not interesting. A lot of deep questions that humans have fall into this category. Maybe one day science will be able to construct testable theories of love, or to test to see if God exists, and so on. But right now

we can't, and I wouldn't bet on ever seeing a satisfying scientific approach to either.

- *The first rule of tautology club is the first rule of tautology club.* This is not a substantive claim of any kind. It's true by definition. No conceivable state of nature could possibly be inconsistent with this claim. We say that this is an unfalsifiable hypothesis, and as such it is outside the domain of science. Whatever else you do in science your claims must have the possibility of being wrong.
- *More people in my experiment will say "yes" than "no".* This one fails as a research hypothesis because it's a claim about the data set, not about the psychology (unless of course your actual research question is whether people have some kind of "yes" bias!). Actually, this hypothesis is starting to sound more like a statistical hypothesis than a research hypothesis.

As you can see, research hypotheses can be somewhat messy at times and ultimately they are *scientific* claims. **Statistical hypotheses** are neither of these two things. Statistical hypotheses must be mathematically precise and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the "population"). Even so, the intent is that statistical hypotheses bear a clear relationship to the substantive research hypotheses that you care about! For instance, in my ESP study my research hypothesis is that some people are able to see through walls or whatever. What I want to do is to "map" this onto a statement about how the data were generated. So let's think about what that statement would be. The quantity that I'm interested in within the experiment is $P(\text{"correct"})$, the true-but-unknown probability with which the participants in my experiment answer the question correctly. Let's use the Greek letter θ (theta) to refer to this probability. Here are four different statistical hypotheses:

- If ESP doesn't exist and if my experiment is well designed then my participants are just guessing. So I should expect them to get it right half of the time and so my statistical hypothesis is that the true probability of choosing correctly is $\theta = 0.5$.
- Alternatively, suppose ESP does exist and participants can see the card. If that's true people will perform better than chance and the statistical hypothesis is that $\theta > 0.5$.
- A third possibility is that ESP does exist, but the colours are all reversed and people don't realise it (okay, that's wacky, but you never know). If that's how it works then you'd expect people's performance to be *below* chance. This would correspond to a statistical hypothesis that $\theta < 0.5$.
- Finally, suppose ESP exists but I have no idea whether people are seeing the right colour or the wrong one. In that case the only claim I could make about the data would be that the probability of making the correct answer is *not* equal to 0.5. This corresponds to the

statistical hypothesis that $\theta \neq 0.5$.

All of these are legitimate examples of a statistical hypothesis because they are statements about a population parameter and are meaningfully related to my experiment.

What this discussion makes clear, I hope, is that when attempting to construct a statistical hypothesis test the researcher actually has two quite distinct hypotheses to consider. First, he or she has a research hypothesis (a claim about psychology), and this then corresponds to a statistical hypothesis (a claim about the data generating population). In my ESP example these might be:

Dani's **research** hypothesis: "ESP exists"
Dani's **statistical** hypothesis: $\theta \neq 0.5$

And a key thing to recognise is this. *A statistical hypothesis test is a test of the statistical hypothesis, not the research hypothesis.* If your study is badly designed then the link between your research hypothesis and your statistical hypothesis is broken. To give a silly example, suppose that my ESP study was conducted in a situation where the participant can actually see the card reflected in a window. If that happens I would be able to find very strong evidence that $\theta \neq 0.5$, but this would tell us nothing about whether "ESP exists".

7.1.2 Null hypotheses and alternative hypotheses

So far, so good. I have a research hypothesis that corresponds to what I want to believe about the world, and I can map it onto a statistical hypothesis that corresponds to what I want to believe about how the data were generated. It's at this point that things get somewhat counter-intuitive for a lot of people. Because what I'm about to do is invent a new statistical hypothesis (the "null" hypothesis, H_0) that corresponds to the exact opposite of what I want to believe, and then focus exclusively on that almost to the neglect of the thing I'm actually interested in (which is now called the "alternative" hypothesis, H_1). In our ESP example, the null hypothesis is that $\theta = 0.5$, since that's what we'd expect if ESP *didn't* exist. My hope, of course, is that ESP is totally real and so the *alternative* to this null hypothesis is $\theta \neq 0.5$. In essence, what we're doing here is dividing up the possible values of θ into two groups: those values that I really hope aren't true (the null), and those values that I'd be happy with if they turn out to be right (the alternative). Having done so, the important thing to recognise is that the goal of a hypothesis test is *not* to show that the alternative hypothesis is (probably) true. The goal is to show that the null hypothesis is (probably) false. Most people find this pretty weird.

The best way to think about it, in my experience, is to imagine that a hypothesis test is a criminal

trial^{*4}, *the trial of the null hypothesis*. The null hypothesis is the defendant, the researcher is the prosecutor, and the statistical test itself is the judge. Just like a criminal trial, there is a presumption of innocence. The null hypothesis is *deemed* to be true unless you, the researcher, can prove beyond a reasonable doubt that it is false. You are free to design your experiment however you like (within reason, obviously!) and your goal when doing so is to maximise the chance that the data will yield a conviction for the crime of being false. The catch is that the statistical test sets the rules of the trial and those rules are designed to protect the null hypothesis, specifically to ensure that if the null hypothesis is actually true the chances of a false conviction are guaranteed to be low. This is pretty important. After all, the null hypothesis doesn't get a lawyer, and given that the researcher is trying desperately to prove it to be false *someone* has to protect it.

7.2

Two types of errors

Before going into details about how a statistical test is constructed it's useful to understand the philosophy behind it. I hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but I should now be explicit. Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is never possible. Sometimes you're just really unlucky. For instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence for a conclusion that the coin is biased, but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we *always* have to accept that there's a chance that we made a mistake. As a consequence the goal behind statistical hypothesis testing is not to *eliminate* errors, but to *minimise* them.

At this point, we need to be a bit more precise about what we mean by "errors". First, let's state the obvious. It is either the case that the null hypothesis is true or that it is false, and our test

^{*4}This analogy only works if you're from an adversarial legal system like UK/US/Australia. As I understand these things, the French inquisitorial system is quite different.

will either retain the null hypothesis or reject it.^{*5} So, as the table below illustrates, after we run the test and make our choice one of four things might have happened:

	retain H_0	reject H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

As a consequence there are actually *two* different types of error here. If we reject a null hypothesis that is actually true then we have made a **type I error**. On the other hand, if we retain the null hypothesis when it is in fact false then we have made a **type II error**.

Remember how I said that statistical testing was kind of like a criminal trial? Well, I meant it. A criminal trial requires that you establish “beyond a reasonable doubt” that the defendant did it. All of the evidential rules are (in theory, at least) designed to ensure that there’s (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant, as the English jurist William Blackstone famously said, it is “better that ten guilty persons escape than that one innocent suffer.” In other words, a criminal trial doesn’t treat the two types of error in the same way. Punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same. The single most important design principle of the test is to *control* the probability of a type I error, to keep it below some fixed probability. This probability, which is denoted α , is called the **significance level** of the test. And I’ll say it again, because it is so central to the whole set-up, a hypothesis test is said to have significance level α if the type I error rate is no larger than α .

So, what about the type II error rate? Well, we’d also like to keep those under control too, and we denote this probability by β . However, it’s much more common to refer to the **power** of the test, that is the probability with which we reject a null hypothesis when it really is false, which is $1 - \beta$. To help keep this straight, here’s the same table again but with the relevant numbers added:

^{*5}An aside regarding the language you use to talk about hypothesis testing. First, one thing you really want to avoid is the word “prove”. A statistical test really doesn’t *prove* that a hypothesis is true or false. Proof implies certainty and, as the saying goes, statistics means never having to say you’re certain. On that point almost everyone would agree. However, beyond that there’s a fair amount of confusion. Some people argue that you’re only allowed to make statements like “rejected the null”, “failed to reject the null”, or possibly “retained the null”. According to this line of thinking you can’t say things like “accept the alternative” or “accept the null”. Personally I think this is too strong. In my opinion, this conflates null hypothesis testing with Karl Popper’s falsificationist view of the scientific process. Whilst there are similarities between falsificationism and null hypothesis testing, they aren’t equivalent. However, whilst I personally think it’s fine to talk about accepting a hypothesis (on the proviso that “acceptance” doesn’t actually mean that it’s necessarily true, especially in the case of the null hypothesis), many people will disagree. And more to the point, you should be aware that this particular weirdness exists so that you’re not caught unawares by it when writing up your own results.

	retain H_0	reject H_0
H_0 is true	$1 - \alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1 - \beta$ (power of the test)

A “powerful” hypothesis test is one that has a small value of β , while still keeping α fixed at some (small) desired level. By convention, scientists make use of three different α levels: .05, .01 and .001. Notice the asymmetry here; the tests are designed to *ensure* that the α level is kept small but there’s no corresponding guarantee regarding β . We’d certainly *like* the type II error rate to be small and we try to design tests that keep it small, but this is typically secondary to the overwhelming need to control the type I error rate. As Blackstone might have said if he were a statistician, it is “better to retain 10 false null hypotheses than to reject a single true one”. To be honest, I don’t know that I agree with this philosophy. There are situations where I think it makes sense, and situations where I think it doesn’t, but that’s neither here nor there. It’s how the tests are built.

7.3

Test statistics and sampling distributions

At this point we need to start talking specifics about how a hypothesis test is constructed. To that end, let’s return to the ESP example. Let’s ignore the actual data that we obtained, for the moment, and think about the structure of the experiment. Regardless of what the actual numbers are, the *form* of the data is that X out of N people correctly identified the colour of the hidden card. Moreover, let’s suppose for the moment that the null hypothesis really is true, that ESP doesn’t exist and the true probability that anyone picks the correct colour is exactly $\theta = 0.5$. What would we *expect* the data to look like? Well, obviously we’d expect the proportion of people who make the correct response to be pretty close to 50%. Or, to phrase this in more mathematical terms, we’d say that X/N is approximately 0.5. Of course, we wouldn’t expect this fraction to be *exactly* 0.5. If, for example, we tested $N = 100$ people and $X = 53$ of them got the question right, we’d probably be forced to concede that the data are quite consistent with the null hypothesis. On the other hand, if $X = 99$ of our participants got the question right then we’d feel pretty confident that the null hypothesis is wrong. Similarly, if only $X = 3$ people got the answer right we’d be similarly confident that the null was wrong. Let’s be a little more technical about this. We have a quantity X that we can calculate by looking at our data. After looking at the value of X we make a decision about whether to believe that the null hypothesis is correct, or to reject the null hypothesis in favour of the alternative. The name for this thing that we calculate to guide our

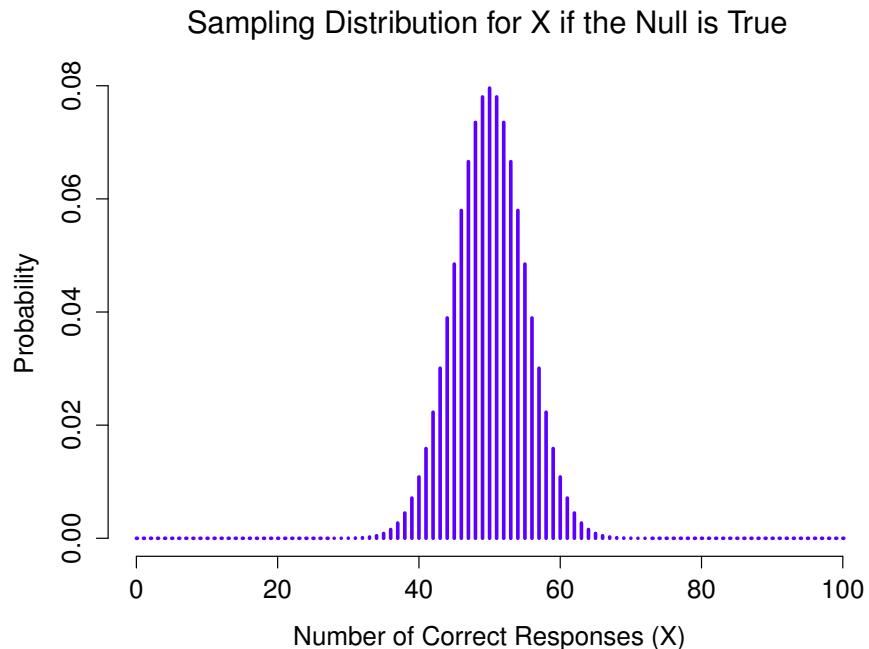


Figure 7.1 The sampling distribution for our test statistic X when the null hypothesis is true. For our ESP scenario this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is $\theta = .5$, the sampling distribution says that the most likely value is 50 (out of 100) correct responses. Most of the probability mass lies between 40 and 60.

choices is a **test statistic**.

Having chosen a test statistic, the next step is to state precisely which values of the test statistic would cause us to reject the null hypothesis, and which values would cause us to keep it. In order to do so we need to determine what the **sampling distribution of the test statistic** would be if the null hypothesis were actually true (we talked about sampling distributions earlier in Section 6.3.1). Why do we need this? Because this distribution tells us exactly what values of X our null hypothesis would lead us to expect. And, therefore, we can use this distribution as a tool for assessing how closely the null hypothesis agrees with our data.

How do we actually determine the sampling distribution of the test statistic? For a lot of hypothesis tests this step is actually quite complicated, and later on in the book you'll see me being slightly evasive about it for some of the tests (some of them I don't even understand myself). However, sometimes it's very easy. And, fortunately for us, our ESP example provides

us with one of the easiest cases. Our population parameter θ is just the overall probability that people respond correctly when asked the question, and our test statistic X is the *count* of the number of people who did so out of a sample size of N . We've seen a distribution like this before, in Section 5.4, and that's exactly what the binomial distribution describes! So, to use the notation and terminology that I introduced in that section, we would say that the null hypothesis predicts that X is binomially distributed, which is written

$$X \sim \text{Binomial}(\theta, N)$$

Since the null hypothesis states that $\theta = 0.5$ and our experiment has $N = 100$ people, we have the sampling distribution we need. This sampling distribution is plotted in Figure 7.1. No surprises really, the null hypothesis says that $X = 50$ is the most likely outcome, and it says that we're almost certain to see somewhere between 40 and 60 correct responses.

7.4 _____

Making decisions

Okay, we're very close to being finished. We've constructed a test statistic (X) and we chose this test statistic in such a way that we're pretty confident that if X is close to $N/2$ then we should retain the null, and if not we should reject it. The question that remains is this. Exactly which values of the test statistic should we associate with the null hypothesis, and exactly which values go with the alternative hypothesis? In my ESP study, for example, I've observed a value of $X = 62$. What decision should I make? Should I choose to believe the null hypothesis or the alternative hypothesis?

7.4.1 Critical regions and critical values

To answer this question we need to introduce the concept of a **critical region** for the test statistic X . The critical region of the test corresponds to those values of X that would lead us to reject null hypothesis (which is why the critical region is also sometimes called the rejection region). How do we find this critical region? Well, let's consider what we know:

- X should be very big or very small in order to reject the null hypothesis.
- If the null hypothesis is true, the sampling distribution of X is $\text{Binomial}(0.5, N)$.
- If $\alpha = .05$, the critical region must cover 5% of this sampling distribution.

Critical Regions for a Two-Sided Test

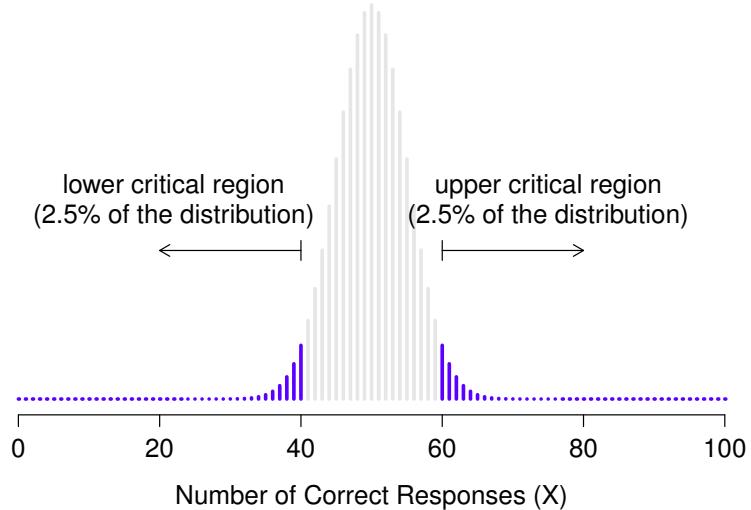


Figure 7.2 The critical region associated with the hypothesis test for the ESP study, for a hypothesis test with a significance level of $\alpha = .05$. The plot shows the sampling distribution of X under the null hypothesis (i.e., same as Figure 7.1). The grey bars correspond to those values of X for which we would retain the null hypothesis. The blue (darker shaded) bars show the critical region, those values of X for which we would reject the null. Because the alternative hypothesis is two sided (i.e., allows both $\theta < .5$ and $\theta > .5$), the critical region covers both tails of the distribution. To ensure an α level of $.05$, we need to ensure that each of the two regions encompasses 2.5% of the sampling distribution.

.....

It's important to make sure you understand this last point. The critical region corresponds to those values of X for which we would reject the null hypothesis, and the sampling distribution in question describes the probability that we would obtain a particular value of X if the null hypothesis were actually true. Now, let's suppose that we chose a critical region that covers 20% of the sampling distribution, and suppose that the null hypothesis is actually true. What would be the probability of incorrectly rejecting the null? The answer is of course 20%. And, therefore, we would have built a test that had an α level of 0.2. If we want $\alpha = .05$, the critical region is only *allowed* to cover 5% of the sampling distribution of our test statistic.

As it turns out those three things uniquely solve the problem. Our critical region consists of the most *extreme values*, known as the **tails** of the distribution. This is illustrated in Figure 7.2. If

we want $\alpha = .05$ then our critical regions correspond to $X \leq 40$ and $X \geq 60$.^{*6} That is, if the number of people saying “true” is between 41 and 59, then we should retain the null hypothesis. If the number is between 0 to 40, or between 60 to 100, then we should reject the null hypothesis. The numbers 40 and 60 are often referred to as the **critical values** since they define the edges of the critical region.

^{*6}Strictly speaking, the test I just constructed has $\alpha = .057$, which is a bit too generous. However, if I’d chosen 39 and 61 to be the boundaries for the critical region then the critical region only covers 3.5% of the distribution. I figured that it makes more sense to use 40 and 60 as my critical values, and be willing to tolerate a 5.7% type I error rate, since that’s as close as I can get to a value of $\alpha = .05$.

At this point, our hypothesis test is essentially complete:

1. (1) we choose an α level (e.g., $\alpha = .05$);
2. (2) come up with some test statistic (e.g., X) that does a good job (in some meaningful sense) of comparing H_0 to H_1 ;
3. (3) figure out the sampling distribution of the test statistic on the assumption that the null hypothesis is true (in this case, binomial); and then
4. (4) calculate the critical region that produces an appropriate α level (0-40 and 60-100).

All that we have to do now is calculate the value of the test statistic for the real data (e.g., $X = 62$) and then compare it to the critical values to make our decision. Since 62 is greater than the critical value of 60 we would reject the null hypothesis. Or, to phrase it slightly differently, we say that the test has produced a statistically **significant** result.

7.4.2 A note on statistical “significance”

Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.

– Attributed to G. O. Ashley^{*7}

A very brief digression is in order at this point, regarding the word “significant”. The concept of statistical significance is actually a very simple one, but has a very unfortunate name. If the data allow us to reject the null hypothesis, we say that “the result is *statistically significant*”, which is often shortened to “the result is significant”. This terminology is rather old and dates back to a time when “significant” just meant something like “indicated”, rather than its modern meaning which is much closer to “important”. As a result, a lot of modern readers get very confused when they start learning statistics because they think that a “significant result” must be an important one. It doesn’t mean that at all. All that “statistically significant” means is that the data allowed us to reject a null hypothesis. Whether or not the result is actually important in the real world is a very different question, and depends on all sorts of other things.

7.4.3 The difference between one sided and two sided tests

There’s one more thing I want to point out about the hypothesis test that I’ve just constructed.

^{*7}The internet seems fairly convinced that Ashley said this, though I can’t for the life of me find anyone willing to give a source for the claim.

If we take a moment to think about the statistical hypotheses I've been using,

$$\begin{aligned} H_0 : \theta &= .5 \\ H_1 : \theta &\neq .5 \end{aligned}$$

we notice that the alternative hypothesis covers *both* the possibility that $\theta < .5$ and the possibility that $\theta > .5$. This makes sense if I really think that ESP could produce either better-than-chance performance *or* worse-than-chance performance (and there are some people who think that). In statistical language this is an example of a **two-sided test**. It's called this because the alternative hypothesis covers the area on both "sides" of the null hypothesis, and as a consequence the critical region of the test covers both tails of the sampling distribution (2.5% on either side if $\alpha = .05$), as illustrated earlier in Figure 7.2.

However, that's not the only possibility. I might only be willing to believe in ESP if it produces better than chance performance. If so, then my alternative hypothesis would only covers the possibility that $\theta > .5$, and as a consequence the null hypothesis now becomes $\theta \leq .5$

$$\begin{aligned} H_0 : \theta &\leq .5 \\ H_1 : \theta &> .5 \end{aligned}$$

When this happens, we have what's called a **one-sided test** and the critical region only covers one tail of the sampling distribution. This is illustrated in Figure 7.3.

7.5

The *p* value of a test

In one sense, our hypothesis test is complete. We've constructed a test statistic, figured out its sampling distribution if the null hypothesis is true, and then constructed the critical region for the test. Nevertheless, I've actually omitted the most important number of all, **the *p* value**. It is to this topic that we now turn. There are two somewhat different ways of interpreting a *p* value, one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but I think that's a bit of a shame. To my mind, Neyman's version is cleaner and actually better reflects the logic of the null hypothesis test. You might disagree though, so I've included both. I'll start with Neyman's version.

Critical Region for a One-Sided Test

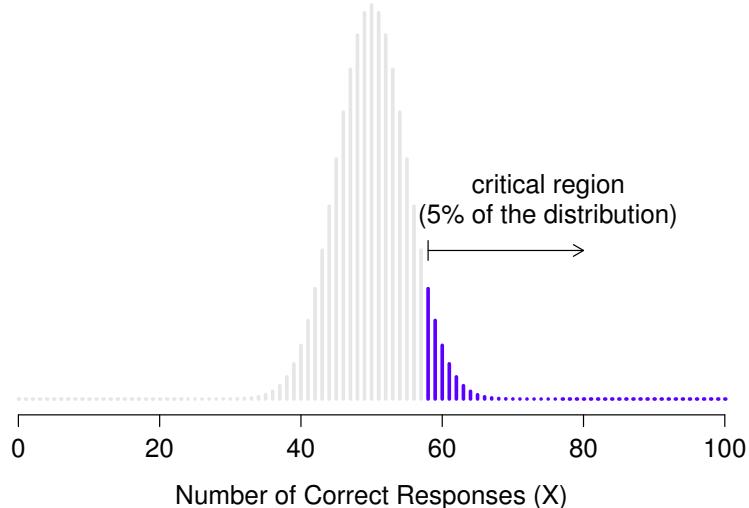


Figure 7.3 The critical region for a one sided test. In this case, the alternative hypothesis is that $\theta > .5$ so we would only reject the null hypothesis for large values of X . As a consequence, the critical region only covers the upper tail of the sampling distribution, specifically the upper 5% of the distribution. Contrast this to the two-sided version in Figure 7.2.

7.5.1 A softer view of decision making

One problem with the hypothesis testing procedure that I've described is that it makes no distinction at all between a result that is "barely significant" and those that are "highly significant". For instance, in my ESP study the data I obtained only just fell inside the critical region, so I did get a significant effect but it was a pretty near thing. In contrast, suppose that I'd run a study in which $X = 97$ out of my $N = 100$ participants got the answer right. This would obviously be significant too but by a much larger margin, such that there's really no ambiguity about this at all. The procedure that I have already described makes no distinction between the two. If I adopt the standard convention of allowing $\alpha = .05$ as my acceptable Type I error rate, then both of these are significant results.

This is where the p value comes in handy. To understand how it works, let's suppose that we ran lots of hypothesis tests on the same data set, but with a different value of α in each case. When we do that for my original ESP data what we'd get is something like this

Value of α	0.05	0.04	0.03	0.02	0.01
Reject the null?	Yes	Yes	Yes	No	No

When we test the ESP data ($X = 62$ successes out of $N = 100$ observations), using α levels of .03 and above, we'd always find ourselves rejecting the null hypothesis. For α levels of .02 and below we always end up retaining the null hypothesis. Therefore, somewhere between .02 and .03 there must be a smallest value of α that would allow us to reject the null hypothesis for this data. This is the p value. As it turns out the ESP data has $p = .021$. In short,

p is defined to be the smallest Type I error rate (α) that you have to be willing to tolerate if you want to reject the null hypothesis.

If it turns out that p describes an error rate that you find intolerable, then you must retain the null. If you're comfortable with an error rate equal to p , then it's okay to reject the null hypothesis in favour of your preferred alternative.

In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values. And as a consequence it has the effect of "softening" our decision process. For those tests in which $p \leq \alpha$ you would have rejected the null hypothesis, whereas for those tests in which $p > \alpha$ you would have retained the null. In my ESP study I obtained $X = 62$ and as a consequence I've ended up with $p = .021$. So the error rate I have to tolerate is 2.1%. In contrast, suppose my experiment had yielded $X = 97$. What happens to my p value now? This time it's shrunk to $p = 1.36 \times 10^{-25}$, which is a tiny, tiny^{*8} Type I error rate. For this second case I would be able to reject the null hypothesis with a lot more confidence, because I only have to be "willing" to tolerate a type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject.

7.5.2 The probability of extreme data

The second definition of the p -value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how, when I constructed the critical region, it corresponded to the *tails* (i.e., extreme values) of the sampling distribution? That's not a coincidence, almost all "good" tests have this characteristic (good in the sense of minimising our type II error rate, β). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are least likely to be observed if the

^{*8}That's $p = .0000000000000000000000000000136$ for folks that don't like scientific notation!

null hypothesis is true. If this rule is true, then we can define the p -value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong.

7.5.3 A common mistake

Okay, so you can see that there are two rather different but legitimate ways to interpret the p value, one based on Neyman's approach to hypothesis testing and the other based on Fisher's. Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is *absolutely and completely wrong*. This mistaken approach is to refer to the p value as "the probability that the null hypothesis is true". It's an intuitively appealing way to think, but it's wrong in two key respects. First, null hypothesis testing is a frequentist tool and the frequentist approach to probability does *not* allow you to assign probabilities to the null hypothesis. According to this view of probability, the null hypothesis is either true or it is not, it cannot have a "5% chance" of being true. Second, even within the Bayesian approach, which does let you assign probabilities to hypotheses, the p value would not correspond to the probability that the null is true. This interpretation is entirely inconsistent with the mathematics of how the p value is calculated. Put bluntly, despite the intuitive appeal of thinking this way, there is no justification for interpreting a p value this way. Never do it.

7.6 _____

Reporting the results of a hypothesis test

When writing up the results of a hypothesis test there's usually several pieces of information that you need to report, but it varies a fair bit from test to test. Throughout the rest of the book I'll spend a little time talking about how to report the results of different tests (see Section ?? for a particularly detailed example), so that you can get a feel for how it's usually done. However, regardless of what test you're doing, the one thing that you always have to do is say something about the p value and whether or not the outcome was significant.

The fact that you have to do this is unsurprising, it's the whole point of doing the test. What might be surprising is the fact that there is some contention over exactly how you're supposed to do it. Leaving aside those people who completely disagree with the entire framework underpinning null hypothesis testing, there's a certain amount of tension that exists regarding whether or not to report the exact p value that you obtained, or if you should state only that $p < \alpha$ for a significance

level that you chose in advance (e.g., $p < .05$).

7.6.1 The issue

To see why this is an issue, the key thing to recognise is that p values are *terribly* convenient. In practice, the fact that we can compute a p value means that we don't actually have to specify any α level at all in order to run the test. Instead, what you can do is calculate your p value and interpret it directly. If you get $p = .062$, then it means that you'd have to be willing to tolerate a Type I error rate of 6.2% to justify rejecting the null. If you personally find 6.2% intolerable then you retain the null. Therefore, the argument goes, why don't we just report the actual p value and let the reader make up their own minds about what an acceptable Type I error rate is? This approach has the big advantage of "softening" the decision making process. In fact, if you accept the Neyman definition of the p value, that's the whole point of the p value. We no longer have a fixed significance level of $\alpha = .05$ as a bright line separating "accept" from "reject" decisions, and this removes the rather pathological problem of being forced to treat $p = .051$ in a fundamentally different way to $p = .049$.

This flexibility is both the advantage and the disadvantage to the p value. The reason why a lot of people don't like the idea of reporting an exact p value is that it gives the researcher a bit *too much* freedom. In particular, it lets you change your mind about what error tolerance you're willing to put up with *after* you look at the data. For instance, consider my ESP experiment. Suppose I ran my test and ended up with a p value of .09. Should I accept or reject? Now, to be honest, I haven't yet bothered to think about what level of Type I error I'm "really" willing to accept. I don't have an opinion on that topic. But I *do* have an opinion about whether or not ESP exists, and I *definitely* have an opinion about whether my research should be published in a reputable scientific journal. And amazingly, now that I've looked at the data I'm starting to think that a 9% error rate isn't so bad, especially when compared to how annoying it would be to have to admit to the world that my experiment has failed. So, to avoid looking like I just made it up after the fact, I now say that my α is .1, with the argument that a 10% type I error rate isn't too bad and at that level my test is significant! I win.

In other words, the worry here is that I might have the best of intentions, and be the most honest of people, but the temptation to just "shade" things a little bit here and there is really, really strong. As anyone who has ever run an experiment can attest, it's a long and difficult process and you often get *very* attached to your hypotheses. It's hard to let go and admit the experiment didn't find what you wanted it to find. And that's the danger here. If we use the "raw" p -value, people will start interpreting the data in terms of what they *want* to believe, not what the data are

Table 7.1 A commonly adopted convention for reporting p values: in many places it is conventional to report one of four different things (e.g., $p < .05$) as shown below. I've included the "significance stars" notation (i.e., a * indicates $p < .05$) because you sometimes see this notation produced by statistical software. It's also worth noting that some people will write *n.s.* (not significant) rather than $p > .05$.

Usual notation	Signif. stars	English translation	The null is...
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.	Rejected
$p < .001$	***	The test was significant at all levels	Rejected
.....			

actually saying and, if we allow that, why are we even bothering to do science at all? Why not let everyone believe whatever they like about anything, regardless of what the facts are? Okay, that's a bit extreme, but that's where the worry comes from. According to this view, you really *must* specify your α value in advance and then only report whether the test was significant or not. It's the only way to keep ourselves honest.

7.6.2 Two proposed solutions

In practice, it's pretty rare for a researcher to specify a single α level ahead of time. Instead, the convention is that scientists rely on three standard significance levels: .05, .01 and .001. When reporting your results, you indicate which (if any) of these significance levels allow you to reject the null hypothesis. This is summarised in Table 7.1. This allows us to soften the decision rule a little bit, since $p < .01$ implies that the data meet a stronger evidential standard than $p < .05$ would. Nevertheless, since these levels are fixed in advance by convention, it does prevent people choosing their α level after looking at the data.

Nevertheless, quite a lot of people still prefer to report exact p values. To many people, the advantage of allowing the reader to make up their own mind about how to interpret $p = .06$ outweighs any disadvantages. In practice, however, even among those researchers who prefer exact p values it is quite common to just write $p < .001$ instead of reporting an exact value for

small p . This is in part because a lot of software doesn't actually print out the p value when it's that small (e.g., SPSS just writes $p = .000$ whenever $p < .001$), and in part because a very small p value can be kind of misleading. The human mind sees a number like $.0000000001$ and it's hard to suppress the gut feeling that the evidence in favour of the alternative hypothesis is a near certainty. In practice however, this is usually wrong. Life is a big, messy, complicated thing, and every statistical test ever invented relies on simplifications, approximations and assumptions. As a consequence, it's probably not reasonable to walk away from *any* statistical analysis with a feeling of confidence stronger than $p < .001$ implies. In other words, $p < .001$ is really code for "as far as *this test* is concerned, the evidence is overwhelming."

In light of all this, you might be wondering exactly what you should do. There's a fair bit of contradictory advice on the topic, with some people arguing that you should report the exact p value, and other people arguing that you should use the tiered approach illustrated in Table 7.1. As a result, the best advice I can give is to suggest that you look at papers/reports written in your field and see what the convention seems to be. If there doesn't seem to be any consistent pattern, then use whichever method you prefer.

7.7

Running the hypothesis test in practice

At this point some of you might be wondering if this is a "real" hypothesis test, or just a toy example that I made up. It's real. In the previous discussion I built the test from first principles, thinking that it was the simplest possible problem that you might ever encounter in real life. However, this test already exists. It's called the *binomial test*, and it's implemented by JASP as one of the statistical analyses available when you hit the 'Frequencies' button. To test the null hypothesis that the response probability is one-half $p = .5$,^{*9} and using data in which $x = 62$ of $n = 100$ people made the correct response, available in the `binomialtest.jasp` data file, we get the results shown in Figure 7.4.

Right now, this output looks pretty unfamiliar to you, but you can see that it's telling you more or less the right things. Specifically, the p -value of 0.02 is less than the usual choice of $\alpha = .05$, so you can reject the null. We'll talk a lot more about how to read this sort of output as we go along, and after a while you'll hopefully find it quite easy to read and understand.

^{*9}Note that the p here has nothing to do with a p value. The p argument in the JASP binomial test corresponds to the probability of making a correct response, according to the null hypothesis. In other words, it's the θ value.

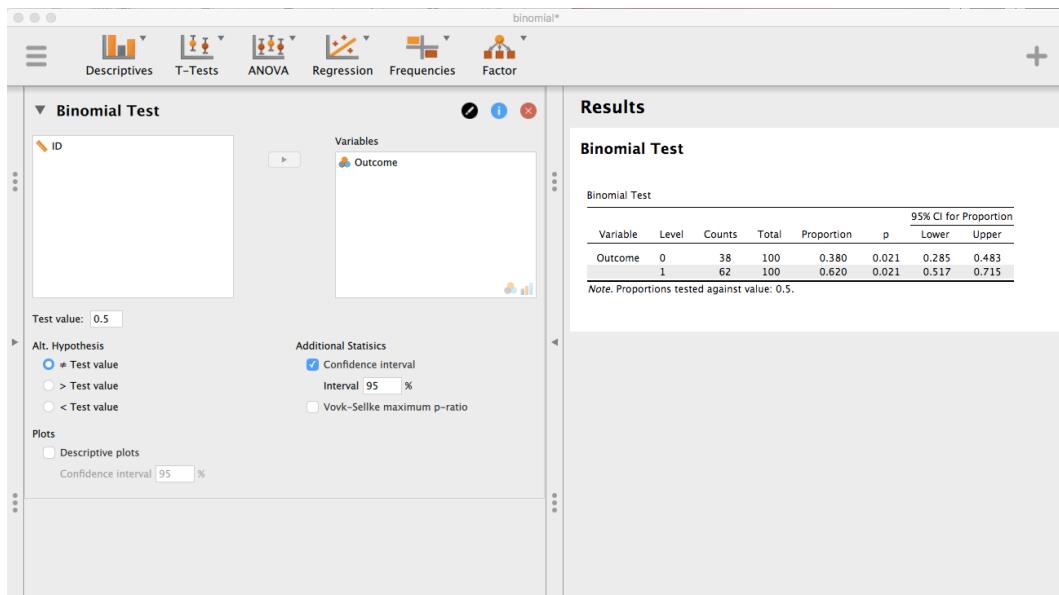


Figure 7.4 Binomial test analysis and results in JASP

7.8

Effect size, sample size and power

In previous sections I've emphasised the fact that the major design principle behind statistical hypothesis testing is that we try to control our Type I error rate. When we fix $\alpha = .05$ we are attempting to ensure that only 5% of true null hypotheses are incorrectly rejected. However, this doesn't mean that we don't care about Type II errors. In fact, from the researcher's perspective, the error of failing to reject the null when it is actually false is an extremely annoying one. With that in mind, a secondary goal of hypothesis testing is to try to minimise β , the Type II error rate, although we don't usually *talk* in terms of minimising Type II errors. Instead, we talk about maximising the *power* of the test. Since power is defined as $1 - \beta$, this is the same thing.

7.8.1 The power function

Let's take a moment to think about what a Type II error actually is. A Type II error occurs when the alternative hypothesis is true, but we are nevertheless unable to reject the null hypothesis. Ideally, we'd be able to calculate a single number β that tells us the Type II error rate, in the

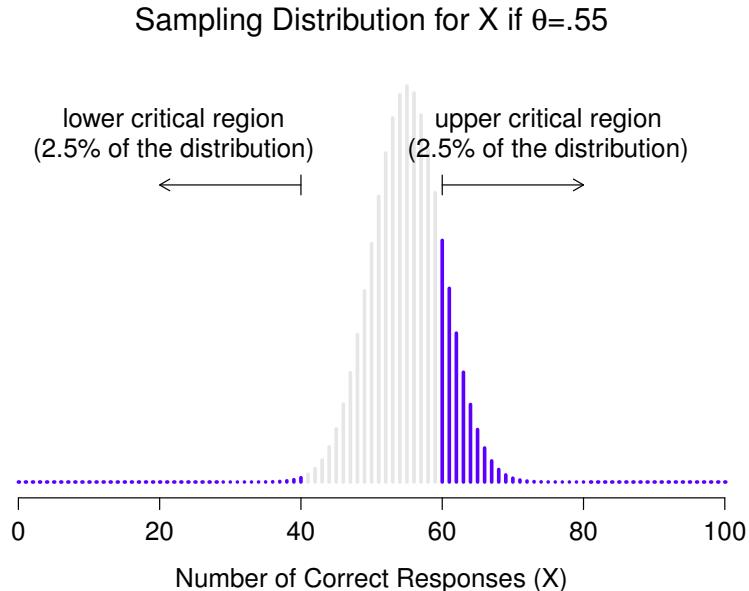


Figure 7.5 Sampling distribution under the *alternative* hypothesis for a population parameter value of $\theta = 0.55$. A reasonable proportion of the distribution lies in the rejection region.

same way that we can set $\alpha = .05$ for the Type I error rate. Unfortunately, this is a lot trickier to do. To see this, notice that in my ESP study the alternative hypothesis actually corresponds to lots of possible values of θ . In fact, the alternative hypothesis corresponds to every value of θ except 0.5. Let's suppose that the true probability of someone choosing the correct response is 55% (i.e., $\theta = .55$). If so, then the *true* sampling distribution for X is not the same one that the null hypothesis predicts, as the most likely value for X is now 55 out of 100. Not only that, the whole sampling distribution has now shifted, as shown in Figure 7.5. The critical regions, of course, do not change. By definition the critical regions are based on what the null hypothesis predicts. What we're seeing in this figure is the fact that when the null hypothesis is wrong, a much larger proportion of the sampling distribution falls in the critical region. And of course that's what should happen. The probability of rejecting the null hypothesis is larger when the null hypothesis is actually false! However $\theta = .55$ is not the only possibility consistent with the alternative hypothesis. Let's instead suppose that the true value of θ is actually 0.7. What happens to the sampling distribution when this occurs? The answer, shown in Figure 7.6, is that almost the entirety of the sampling distribution has now moved into the critical region. Therefore, if $\theta = 0.7$, the probability of us correctly rejecting the null hypothesis (i.e., the power of the test) is much larger than if $\theta = 0.55$. In short, while $\theta = .55$ and $\theta = .70$ are both part of the alternative

hypothesis, the Type II error rate is different.

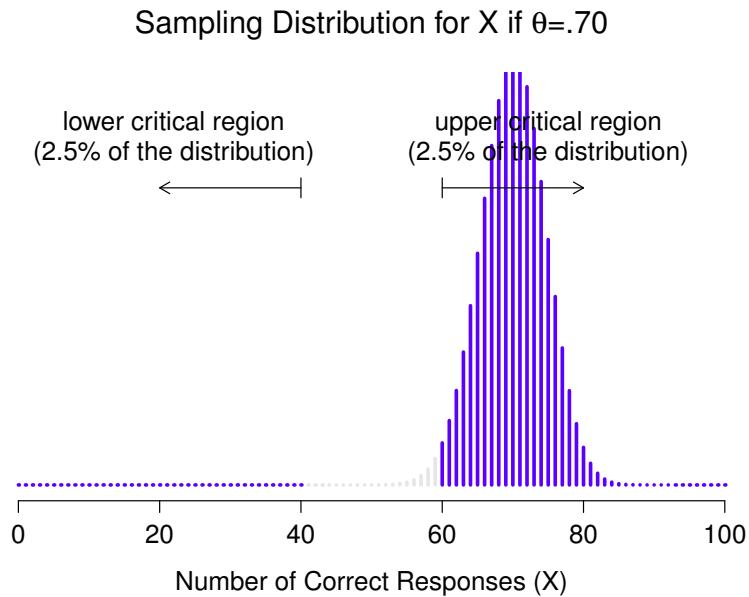


Figure 7.6 Sampling distribution under the *alternative* hypothesis for a population parameter value of $\theta = 0.70$. Almost all of the distribution lies in the rejection region.

What all this means is that the power of a test (i.e., $1 - \beta$) depends on the true value of θ . To illustrate this, I've calculated the expected probability of rejecting the null hypothesis for all values of θ , and plotted it in Figure 7.7. This plot describes what is usually called the **power function** of the test. It's a nice summary of how good the test is, because it actually tells you the power ($1 - \beta$) for all possible values of θ . As you can see, when the true value of θ is very close to 0.5, the power of the test drops very sharply, but when it is further away, the power is large.

7.8.2 Effect size

Since all models are wrong the scientist must be alert to what is importantly wrong.

It is inappropriate to be concerned with mice when there are tigers abroad

– George Box (Box 1976)

The plot shown in Figure 7.7 captures a fairly basic point about hypothesis testing. If the true state of the world is very different from what the null hypothesis predicts then your power

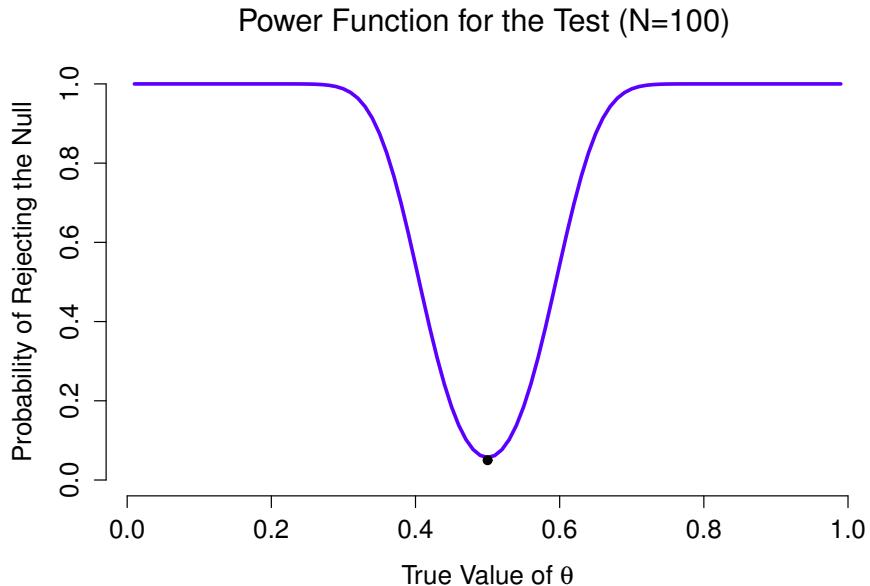


Figure 7.7 The probability that we will reject the null hypothesis, plotted as a function of the true value of θ . Obviously, the test is more powerful (greater chance of correct rejection) if the true value of θ is very different from the value that the null hypothesis specifies (i.e., $\theta = .5$). Notice that when θ actually is equal to $.5$ (plotted as a black dot), the null hypothesis is in fact true and rejecting the null hypothesis in this instance would be a Type I error.

will be very high, but if the true state of the world is similar to the null (but not identical) then the power of the test is going to be very low. Therefore, it's useful to be able to have some way of quantifying how "similar" the true state of the world is to the null hypothesis. A statistic that does this is called a measure of **effect size** ([Cohen 1988](#); [Ellis 2010](#)). Effect size is defined slightly differently in different contexts (and so this section just talks in general terms) but the qualitative idea that it tries to capture is always the same. How big is the difference between the *true* population parameters and the parameter values that are assumed by the null hypothesis? In our ESP example, if we let $\theta_0 = 0.5$ denote the value assumed by the null hypothesis and let θ denote the true value, then a simple measure of effect size could be something like the difference between the true value and null (i.e., $\theta - \theta_0$), or possibly just the magnitude of this difference, $\text{abs}(\theta - \theta_0)$.

Why calculate effect size? Let's assume that you've run your experiment, collected the data, and gotten a significant effect when you ran your hypothesis test. Isn't it enough just to say that you've gotten a significant effect? Surely that's the *point* of hypothesis testing? Well, sort of.

Table 7.2 A crude guide to understanding the relationship between statistical significance and effect sizes. Basically, if you don't have a significant result then the effect size is pretty meaningless because you don't have any evidence that it's even real. On the other hand, if you do have a significant effect but your effect size is small then there's a pretty good chance that your result (although real) isn't all that interesting. However, this guide is very crude. It depends a lot on what exactly you're studying. Small effects can be of massive practical importance in some situations. So don't take this table too seriously. It's a rough guide at best.

	big effect size	small effect size
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed
.....		

Yes, the point of doing a hypothesis test is to try to demonstrate that the null hypothesis is wrong, but that's hardly the only thing we're interested in. If the null hypothesis claimed that $\theta = .5$ and we show that it's wrong, we've only really told half of the story. Rejecting the null hypothesis implies that we believe that $\theta \neq .5$, but there's a big difference between $\theta = .51$ and $\theta = .8$. If we find that $\theta = .8$, then not only have we found that the null hypothesis is wrong, it appears to be very wrong. On the other hand, suppose we've successfully rejected the null hypothesis, but it looks like the true value of θ is only $.51$ (this would only be possible with a very large study). Sure, the null hypothesis is wrong but it's not at all clear that we actually *care* because the effect size is so small. In the context of my ESP study we might still care since any demonstration of real psychic powers would actually be pretty cool^{*10}, but in other contexts a 1% difference usually isn't very interesting, even if it is a real difference. For instance, suppose we're looking at differences in high school exam scores between males and females and it turns out that the female scores are 1% higher on average than the males. If I've got data from thousands of students then this difference will almost certainly be *statistically significant*, but regardless of how small the p value is it's just not very interesting. You'd hardly want to go around proclaiming a crisis in boys education on the basis of such a tiny difference would you? It's for this reason that it is becoming more standard (slowly, but surely) to report some kind of standard measure of effect size along with the results of the hypothesis test. The hypothesis test itself tells you whether you should believe that

^{*10}Although in practice a very small effect size is worrying because even very minor methodological flaws might be responsible for the effect, and in practice no experiment is perfect so there are always methodological issues to worry about.

the effect you have observed is real (i.e., not just due to chance), whereas the effect size tells you whether or not you should care.

7.8.3 Increasing the power of your study

Not surprisingly, scientists are fairly obsessed with maximising the power of their experiments. We want our experiments to work and so we want to maximise the chance of rejecting the null hypothesis if it is false (and of course we usually want to believe that it is false!). As we've seen, one factor that influences power is the effect size. So the first thing you can do to increase your power is to increase the effect size. In practice, what this means is that you want to design your study in such a way that the effect size gets magnified. For instance, in my ESP study I might believe that psychic powers work best in a quiet, darkened room with fewer distractions to cloud the mind. Therefore I would try to conduct my experiments in just such an environment. If I can strengthen people's ESP abilities somehow then the true value of θ will go up^{*11} and therefore my effect size will be larger. In short, clever experimental design is one way to boost power, because it can alter the effect size.

Unfortunately, it's often the case that even with the best of experimental designs you may have only a small effect. Perhaps, for example, ESP really does exist but even under the best of conditions it's very very weak. Under those circumstances your best bet for increasing power is to increase the sample size. In general, the more observations that you have available, the more likely it is that you can discriminate between two hypotheses. If I ran my ESP experiment with 10 participants and 7 of them correctly guessed the colour of the hidden card you wouldn't be terribly impressed. But if I ran it with 10,000 participants, and 7,000 of them got the answer right, you would be much more likely to think I had discovered something. In other words, power increases with the sample size. This is illustrated in Figure 7.8, which shows the power of the test for a true parameter of $\theta = 0.7$ for all sample sizes N from 1 to 100, where I'm assuming that the null hypothesis predicts that $\theta_0 = 0.5$.

Because power is important, whenever you're contemplating running an experiment it would be pretty useful to know how much power you're likely to have. It's never possible to know for sure since you can't possibly know what your real effect size is. However, it's often (well, sometimes) possible to guess how big it should be. If so, you can guess what sample size you need! This idea

^{*11}Notice that the true population parameter θ doesn't necessarily correspond to an immutable fact of nature. In this context θ is just the true probability that people would correctly guess the colour of the card in the other room. As such the population parameter can be influenced by all sorts of things. Of course, this is all on the assumption that ESP actually exists!

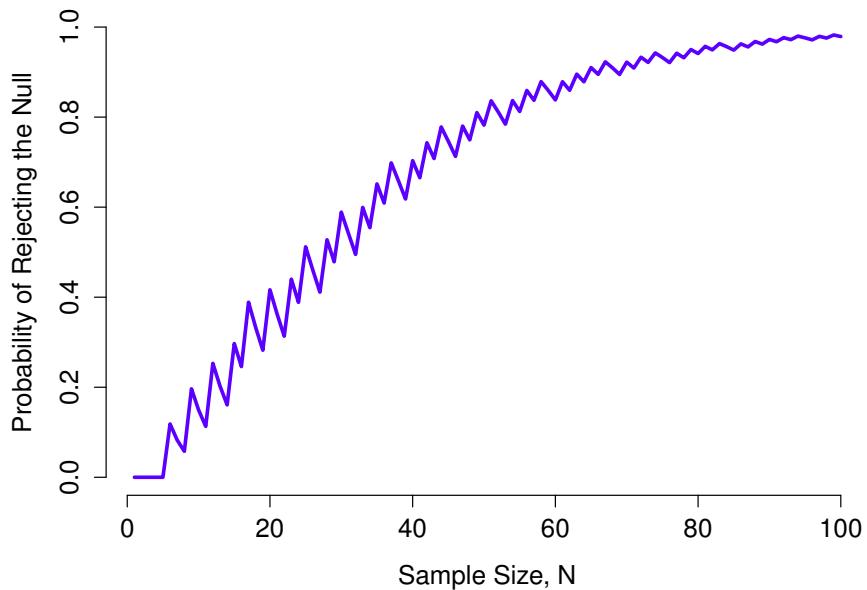


Figure 7.8 The power of our test plotted as a function of the sample size N . In this case, the true value of θ is 0.7 but the null hypothesis is that $\theta = 0.5$. Overall, larger N means greater power. (The small zig-zags in this function occur because of some odd interactions between θ , α and the fact that the binomial distribution is discrete, it doesn't matter for any serious purpose).

is called **power analysis**, and if it's feasible to do it then it's very helpful. It can tell you something about whether you have enough time or money to be able to run the experiment successfully. It's increasingly common to see people arguing that power analysis should be a required part of experimental design, so it's worth knowing about. I don't discuss power analysis in this book, however. This is partly for a boring reason and partly for a substantive one. The boring reason is that I haven't had time to write about power analysis yet. The substantive one is that I'm still a little suspicious of power analysis. Speaking as a researcher, I have very rarely found myself in a position to be able to do one. It's either the case that (a) my experiment is a bit non-standard and I don't know how to define effect size properly, or (b) I literally have so little idea about what the effect size will be that I wouldn't know how to interpret the answers. Not only that, after extensive conversations with someone who does stats consulting for a living (my wife, as it happens), I can't help but notice that in practice the *only* time anyone ever asks her for a power analysis is when she's helping someone write a grant application. In other words, the only time any

scientist ever seems to want a power analysis in real life is when they're being forced to do it by bureaucratic process. It's not part of anyone's day to day work. In short, I've always been of the view that whilst power is an important concept, power *analysis* is not as useful as people make it sound, except in the rare cases where (a) someone has figured out how to calculate power for your actual experimental design and (b) you have a pretty good idea what the effect size is likely to be.^{*12} Maybe other people have had better experiences than me, but I've personally never been in a situation where both (a) and (b) were true. Maybe I'll be convinced otherwise in the future, and probably a future version of this book would include a more detailed discussion of power analysis, but for now this is about as much as I'm comfortable saying about the topic.

7.9

Some issues to consider

What I've described to you in this chapter is the orthodox framework for null hypothesis significance testing (NHST). Understanding how NHST works is an absolute necessity because it has been the dominant approach to inferential statistics ever since it came to prominence in the early 20th century. It's what the vast majority of working scientists rely on for their data analysis, so even if you hate it you need to know it. However, the approach is not without problems. There are a number of quirks in the framework, historical oddities in how it came to be, theoretical disputes over whether or not the framework is right, and a lot of practical traps for the unwary. I'm not going to go into a lot of detail on this topic, but I think it's worth briefly discussing a few of these issues.

7.9.1 Neyman versus Fisher

The first thing you should be aware of is that orthodox NHST is actually a mash-up of two rather different approaches to hypothesis testing, one proposed by Sir Ronald Fisher and the other proposed by Jerzy Neyman (**Lehmann2011**). The history is messy because Fisher and Neyman were real people whose opinions changed over time, and at no point did either of them offer "the definitive statement" of how we should interpret their work many decades later. That said, here's a quick summary of what I take these two approaches to be.

First, let's talk about Fisher's approach. As far as I can tell, Fisher assumed that you only had

^{*12}One possible exception to this is when researchers study the effectiveness of a new medical treatment and they specify in advance what an important effect size would be to detect, for example over and above any existing treatment. In this way some information about the potential value of a new treatment can be obtained.

the one hypothesis (the null) and that what you want to do is find out if the null hypothesis is inconsistent with the data. From his perspective, what you should do is check to see if the data are “sufficiently unlikely” according to the null. In fact, if you remember back to our earlier discussion, that’s how Fisher defines the p -value. According to Fisher, if the null hypothesis provided a very poor account of the data then you could safely reject it. But, since you don’t have any other hypotheses to compare it to, there’s no way of “accepting the alternative” because you don’t necessarily have an explicitly stated alternative. That’s more or less all there is to it.

In contrast, Neyman thought that the point of hypothesis testing was as a guide to action and his approach was somewhat more formal than Fisher’s. His view was that there are multiple things that you could *do* (accept the null or accept the alternative) and the point of the test was to tell you which one the data support. From this perspective, it is critical to specify your alternative hypothesis properly. If you don’t know what the alternative hypothesis is, then you don’t know how powerful the test is, or even which action makes sense. His framework genuinely requires a competition between different hypotheses. For Neyman, the p value didn’t directly measure the probability of the data (or data more extreme) under the null, it was more of an abstract description about which “possible tests” were telling you to accept the null, and which “possible tests” were telling you to accept the alternative.

As you can see, what we have today is an odd mishmash of the two. We talk about having both a null hypothesis and an alternative (Neyman), but usually^{*13} define the p value in terms of extreme data (Fisher), but we still have α values (Neyman). Some of the statistical tests have explicitly specified alternatives (Neyman) but others are quite vague about it (Fisher). And, according to some people at least, we’re not allowed to talk about accepting the alternative (Fisher). It’s a mess, but I hope this at least explains why it’s a mess.

7.9.2 Bayesians versus frequentists

Earlier on in this chapter I was quite emphatic about the fact that you *cannot* interpret the p value as the probability that the null hypothesis is true. NHST is fundamentally a frequentist tool (see Chapter 5) and as such it does not allow you to assign probabilities to hypotheses. The null hypothesis is either true or it is not. The Bayesian approach to statistics interprets probability as a degree of belief, so it’s totally okay to say that there is a 10% chance that the null hypothesis is true. That’s just a reflection of the degree of confidence that you have in this hypothesis.

^{*13}Although this book describes both Neyman’s and Fisher’s definition of the p value, most don’t. Most introductory textbooks will only give you the Fisher version.

You aren't allowed to do this within the frequentist approach. Remember, if you're a frequentist, a probability can only be defined in terms of what happens after a large number of independent replications (i.e., a long run frequency). If this is your interpretation of probability, talking about the "probability" that the null hypothesis is true is complete gibberish: a null hypothesis is either true or it is false. There's no way you can talk about a long run frequency for this statement. To talk about "the probability of the null hypothesis" is as meaningless as "the colour of freedom". It doesn't have one!

Most importantly, this *isn't* a purely ideological matter. If you decide that you are a Bayesian and that you're okay with making probability statements about hypotheses, you have to follow the Bayesian rules for calculating those probabilities. I'll talk more about this in Chapter ??, but for now what I want to point out to you is the p value is a *terrible* approximation to the probability that H_0 is true. If what you want to know is the probability of the null, then the p value is not what you're looking for!

7.9.3 Traps

As you can see, the theory behind hypothesis testing is a mess, and even now there are arguments in statistics about how it "should" work. However, disagreements among statisticians are not our real concern here. Our real concern is practical data analysis. And while the "orthodox" approach to null hypothesis significance testing has many drawbacks, even an unrepentant Bayesian like myself would agree that they can be useful if used responsibly. Most of the time they give sensible answers and you can use them to learn interesting things. Setting aside the various ideologies and historical confusions that we've discussed, the fact remains that the biggest danger in all of statistics is *thoughtlessness*. I don't mean stupidity, I literally mean thoughtlessness. The rush to interpret a result without spending time thinking through what each test actually says about the data, and checking whether that's consistent with how you've interpreted it. That's where the biggest trap lies.

To give an example of this, consider the following example (**Gelman2006**). Suppose I'm running my ESP study and I've decided to analyse the data separately for the male participants and the female participants. Of the male participants, 33 out of 50 guessed the colour of the card correctly. This is a significant effect ($p = .03$). Of the female participants, 29 out of 50 guessed correctly. This is not a significant effect ($p = .32$). Upon observing this, it is extremely tempting for people to start wondering why there is a difference between males and females in terms of their psychic abilities. However, this is wrong. If you think about it, we haven't *actually* run a test that explicitly compares males to females. All we have done is compare males to chance (binomial test was significant) and compared females to chance (binomial test was non significant). If we want to

argue that there is a real difference between the males and the females, we should probably run a test of the null hypothesis that there is no difference! We can do that using a different hypothesis test,^{*14} but when we do that it turns out that we have no evidence that males and females are significantly different ($p = .54$). Now do you think that there's anything fundamentally different between the two groups? Of course not. What's happened here is that the data from both groups (male and female) are pretty borderline. By pure chance one of them happened to end up on the magic side of the $p = .05$ line, and the other one didn't. That doesn't actually imply that males and females are different. This mistake is so common that you should always be wary of it. The difference between significant and not-significant is *not* evidence of a real difference. If you want to say that there's a difference between two groups, then you have to test for that difference!

The example above is just that, an example. I've singled it out because it's such a common one, but the bigger picture is that data analysis can be tricky to get right. Think about what it is you want to test, why you want to test it, and whether or not the answers that your test gives could possibly make any sense in the real world.

7.10 _____

Summary

Null hypothesis testing is one of the most ubiquitous elements to statistical theory. The vast majority of scientific papers report the results of some hypothesis test or another. As a consequence it is almost impossible to get by in science without having at least a cursory understanding of what a p -value means, making this one of the most important chapters in the book. As usual, I'll end the chapter with a quick recap of the key ideas that we've talked about:

- Research hypotheses and statistical hypotheses. Null and alternative hypotheses. (Section 7.1).
- Type 1 and Type 2 errors (Section 7.2)
- Test statistics and sampling distributions (Section 7.3)
- Hypothesis testing as a decision making process (Section 7.4)
- p -values as “soft” decisions (Section 7.5)
- Writing up the results of a hypothesis test (Section 7.6)
- Running the hypothesis test in practice (Section 7.7)
- Effect size and power (Section 7.8)
- A few issues to consider regarding hypothesis testing (Section 7.9)

^{*14}In this case, the Pearson chi-square test of independence (Chapter 8)

Later in the book, in Chapter ??, I'll revisit the theory of null hypothesis tests from a Bayesian perspective and introduce a number of new tools that you can use if you aren't particularly fond of the orthodox approach. But for now, though, we're done with the abstract statistical theory, and we can start discussing specific data analysis tools.

Part IV.

Statistical tools

8. カテゴリカルデータの分析

仮説検定に関する基本的なことを学んだうえで、今度は心理学でよく使われる検定について見ていきましょう。では、どこから始めればよいのでしょうか。全ての教科書がスタート地点に関する合意を持つわけではないのですが、ここでは“ χ^2 検定”（この章では、“カイ二乗 (にじょう)chi-square”と発音します^{*1}）と“t-検定”（Chapter ??）から始めます。これらの検定は科学的実践において頻繁に使用されており、“回帰”（Chapter ??）や“分散分析”（Chapter ??）ほど強力ではないですがそれよりはるかに理解しやすいものとなっています。

“カテゴリカルデータ”という用語は“名義尺度データ”的別名に過ぎません。説明していないことではなく、ただデータ分析の文脈では、“名義尺度データ”よりも“カテゴリカルデータ”という言葉を使う傾向があるのです。なぜかは知りません。なんにせよ、**カテゴリカルデータの分析** はあなたのデータが名義尺度の際に適用可能なツールの集合を指示しています。しかし、カテゴリカルデータの分析に使用できるツールには様々なものがあり、本章では一般的なツールの一部のみを取り上げます。

8.1

The χ^2 (カイ二乗) 適合度検定

χ^2 適合度検定は、最も古い仮説検定の一つです。この検定は世紀の変わり目に Karl Pearson 氏が考案したもので（Pearson1900）、Ronald Fisher 氏によっていくつかの修正が加えられました（Fisher1922）。名義尺度変数に関する観測度数分布が期待度数分布と合致するかどうかを調べます。例えば、ある患者グループが実験的処置を受けており、彼・彼女らの状況が改善されたか、変化がないか、悪化したかを確認するために健康状態が評価されたとします。各カテゴリー（改善、変化なし、悪化）の数値が、標準的な処置条件で期待される数値と一致するかどうかを判断するために、適

*1 また“カイ二乗 (じじょう)chi-squared”とも呼ばれる