

# 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](mailto:d.navarro@unsw.edu.au)

David Foxcroft  
Oxford Brookes University  
[david.foxcroft@brookes.ac.uk](mailto:david.foxcroft@brookes.ac.uk)

Thomas J. Faulkenberry  
Tarleton State University  
[faulkenberry@tarleton.edu](mailto: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](mailto: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随時コンパイルされ、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 18 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っ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](http://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、 $\chi^2$  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ができます。

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<sup>\*1</sup>

### 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

---

<sup>\*1</sup>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,<sup>\*2</sup> 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

---

<sup>\*2</sup>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"	<b>46% say "valid"</b>
argument is invalid	<b>92% say "valid"</b>	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!).<sup>\*3</sup>

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

---

<sup>\*3</sup>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!\*<sup>4</sup> 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):

---

\*<sup>4</sup>Earlier 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 ?? 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

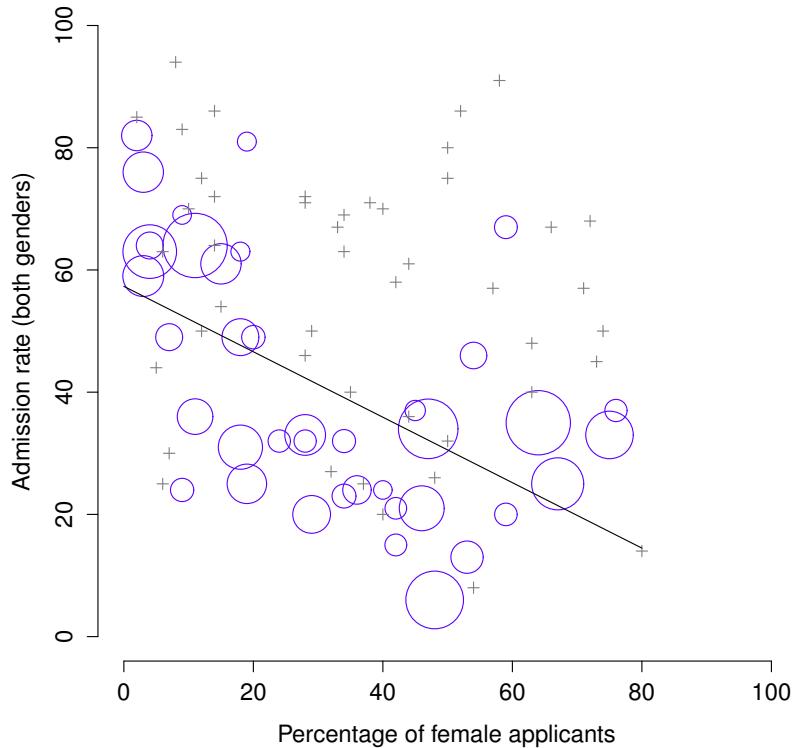


Figure1.1 The Berkeley 1973 college admissions data. This figure plots the admission rate for the 85 departments that had at least one female applicant, as a function of the percentage of applicants that were female. The plot is a redrawing of Figure 1 from **Bickel1975**. Circles plot departments with 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.<sup>\*5</sup>

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

---

<sup>\*5</sup>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 ???. 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<sup>\*1</sup>

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

---

<sup>\*1</sup>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<sup>\*2</sup>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<sup>\*3</sup>、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。

---

<sup>\*2</sup> 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

<sup>\*3</sup> これが執筆された 2019 年 5 月

ます。<sup>\*4</sup>

### 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 ??のようなユーザーインターフェイスが表示されます。

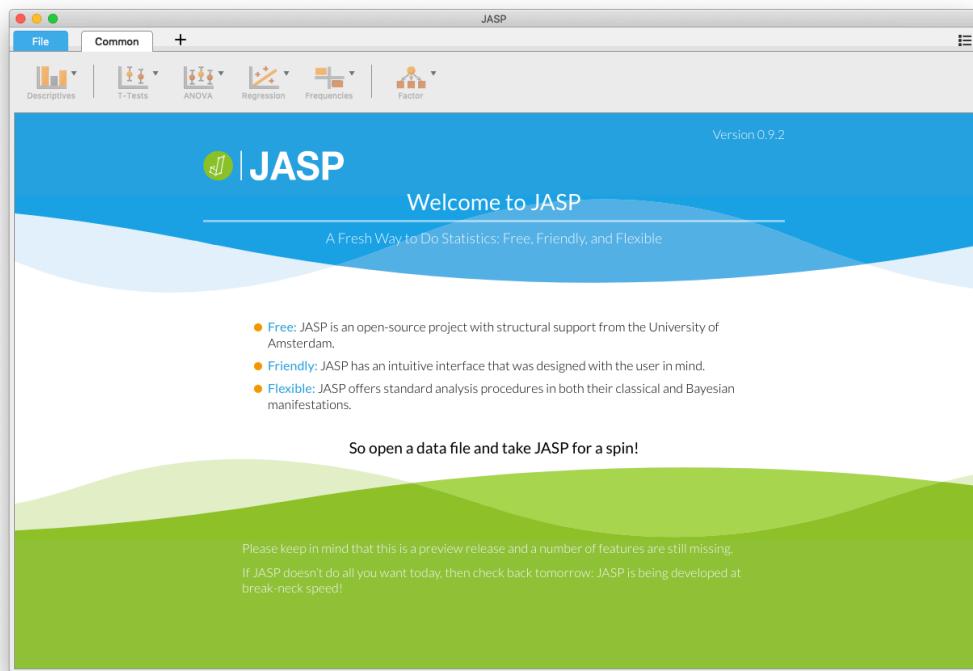


Figure2.1 起動時の JASP

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<sup>\*5</sup>。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

<sup>\*4</sup>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<sup>5</sup>、この本の内容に大きな違いはありませんでした。

<sup>\*5</sup>この重要な問題についての議論が<sup>3</sup>、<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 data starts from row 1 and continues to row 13. The first row contains column headers: 'Month', 'Days', 'Sales', and 'Stock.Levels'. The subsequent rows contain data for each month: January (31 days, 0 sales, high stock), February (28 days, 100 sales, high stock), March (31 days, 200 sales, low stock), April (30 days, 50 sales, out of stock), May (31 days, 0 sales, out of stock), June (30 days, 0 sales, high stock), July (31 days, 0 sales, high stock), August (31 days, 0 sales, high stock), September (30 days, 0 sales, high stock), October (31 days, 0 sales, high stock), November (30 days, 0 sales, high stock), and December (31 days, 0 sales, high stock). The right side of the screen displays the raw CSV text data, which is identical to the spreadsheet content, showing the header 'Month,Days,Sales,Stock.Levels' followed by the same 13 rows of data.

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単なるデータのテーブルです。これを Figure ??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

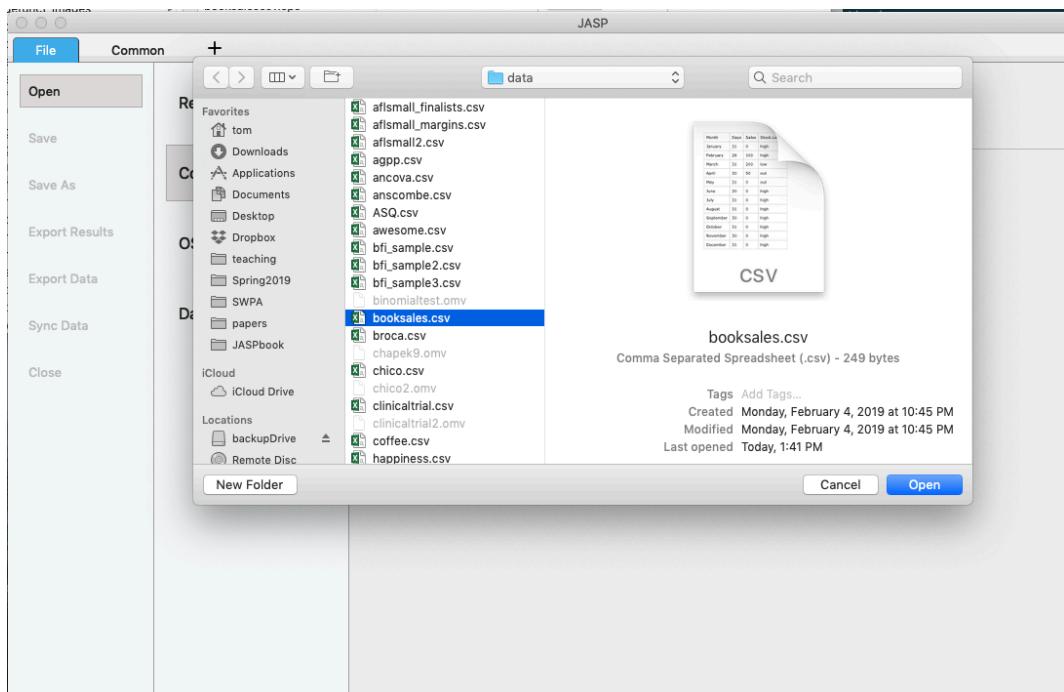


Figure2.3 JASP がインポートする CSV ファイルを選択するように求める Mac 上のダイアログボックス。Mac ユーザーはこれをすぐに理解すると思います。これは、Mac があなたにファイルを探す時に要求する一般的な方法です。Windows ユーザーにはこれは表示されませんが、代わりに、ファイルを選択するときに Windows がいつも出してくる通常のエクスプローラーウィンドウが表示されます。

~??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  $\text{\LaTeX}$  document preparation system, you can select “Copy special” and “ $\text{\LaTeX}$  code”; doing so will place the  $\text{\LaTeX}$  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 ?? 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 ???. We downloaded and installed JASP, and started it up.
- Section ???. 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 ???. We saw how to load data files (formatted as .csv files) in JASP.
- Section ???. 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 ???. And saw that sometimes we need to coerce data from one type to another.
- Section ???. 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<sup>\*1</sup> 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 ?? 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 ?? 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 ?? 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。

---

<sup>\*1</sup> 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 ??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きますので、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。

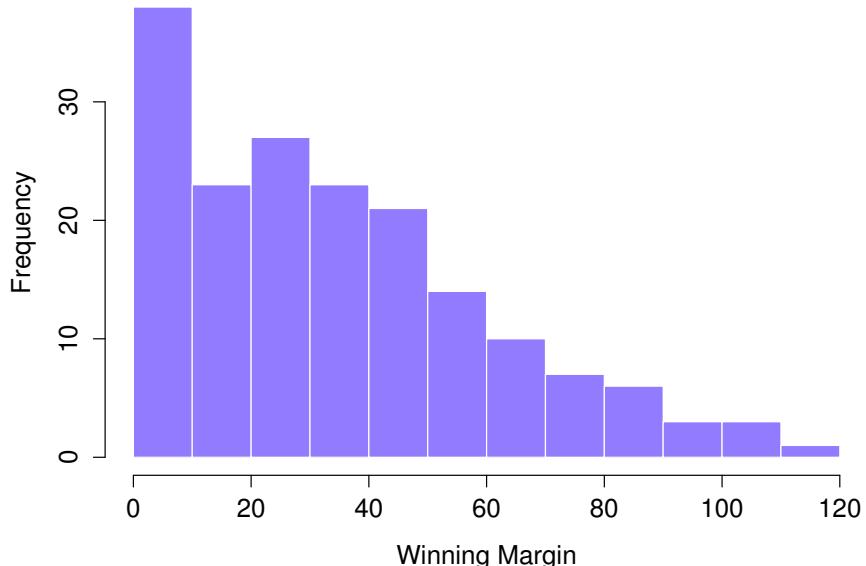


Figure3.2 2020 年の AFL 得点差データ (変数 `afl.margins` ) のヒストグラム。ご想像通り、より大差がつくゲームはより少ないので見て取れます。

### 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  $\Sigma$  を導入してこれを短縮しましょう<sup>a</sup> 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、 $\Sigma$  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。

---

<sup>a</sup> 総和に対して  $\Sigma$  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  $\Pi$  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??を見てください。

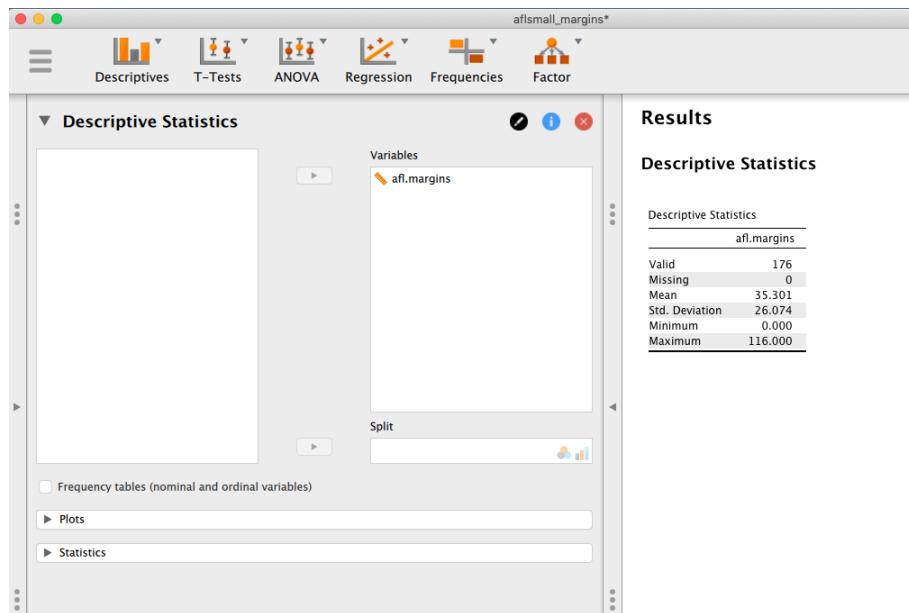


Figure3.3 AFLにおける2010年得点差データ(変数[afl.margins](#))のデフォルトで示される記述統計

.....

図??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ね。

### 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??にそれを描いてみました。平均は、データセットの“重心”的な値で、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、という考え方によると、
- 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 ??を参照）に現れます。これは図 ??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。

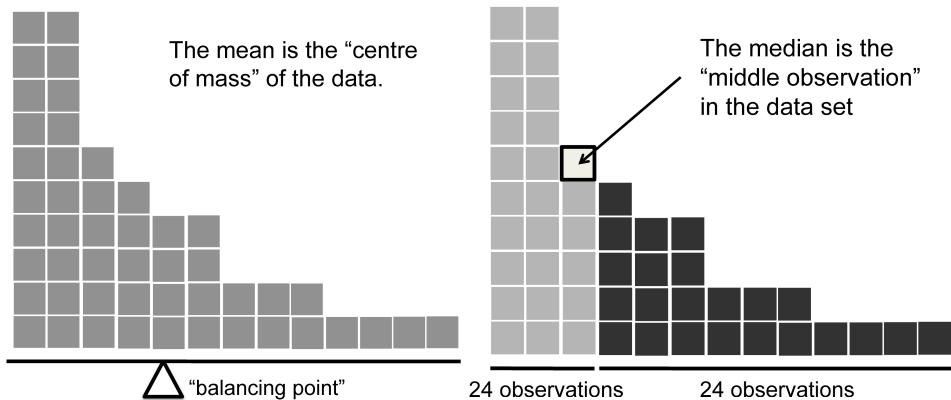


Figure 3.4 平均値と中央値の違いをどう解釈するかについてのイラスト。平均値は基本的にデータセットの“重心”です。データのヒストグラムが固体物だと考えたら、そのバランスを取る点(シーソーみたいに)が平均値です。それに対して、中央値は真ん中の観測で、それより小さいデータが半分、それより大きいデータが半分あるということです。

.....

### 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<sup>\*2</sup>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、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

<sup>\*2</sup>[www.abc.net.au/news/stories/2010/09/24/3021480.htm](http://www.abc.net.au/news/stories/2010/09/24/3021480.htm)

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。

むにゃむにゃ。

### 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 ?? 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スをクリックします。すると図 ?? のようなものが得られるでしょう。

さて度数分布表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

JASP interface showing a table with 13 rows. The first column contains row numbers (1-13) and the second column contains categorical values: Hawthorn, Melbourne, Carlton, Melbourne, Hawthorn, Carlton, Melbourne, Carlton, Hawthorn, Melbourne, Melbourne, Hawthorn, 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

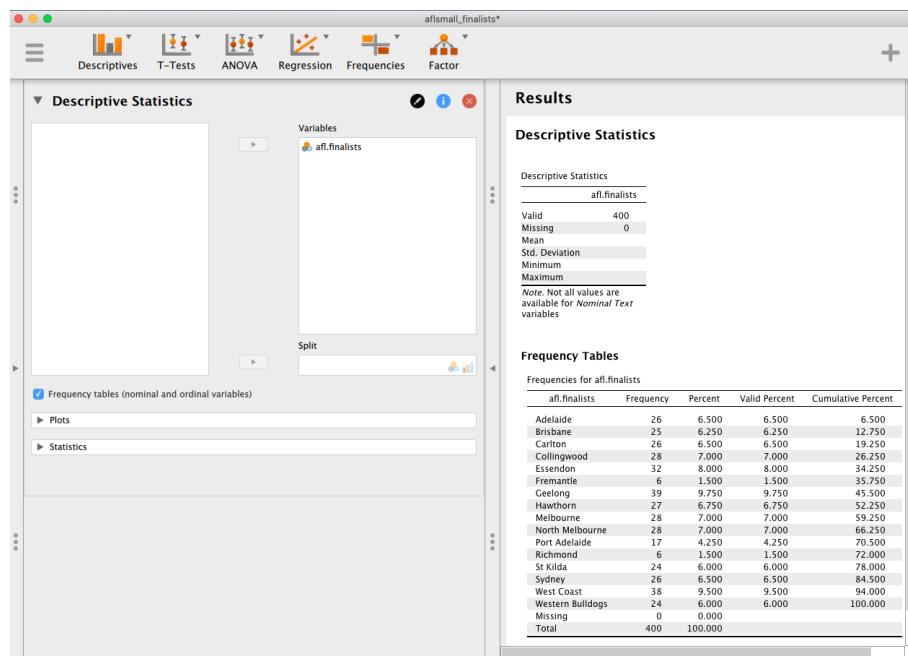


Figure3.6 afl.finalists 変数の度数分布表を示した JASP スクリーンショット

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 ?? 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。

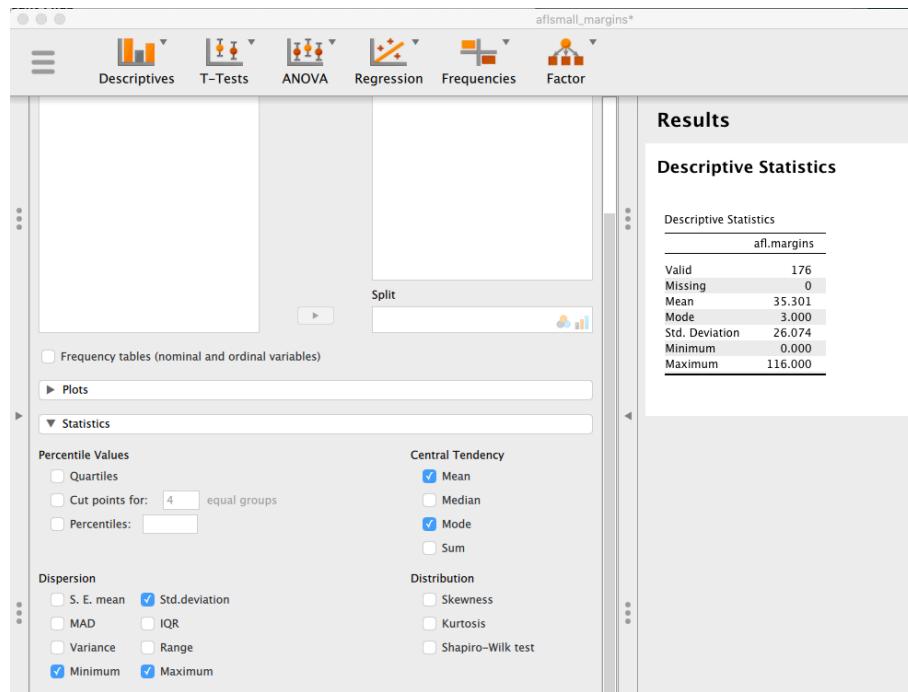


Figure3.7 afl.margins 変数の中央値を示す JASP 画面

### 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

驚くには値しませんが、図??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、というのがあります。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き、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というやれやれ。*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<sup>\*3</sup>、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”のかわりに“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。

---

<sup>\*3</sup>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 ?? 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 ?? 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??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 ??章で再びこのトピックに触れると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<sup>\*4</sup>。図 ??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 ??. このことは、図 ??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

<sup>\*4</sup>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?? 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。

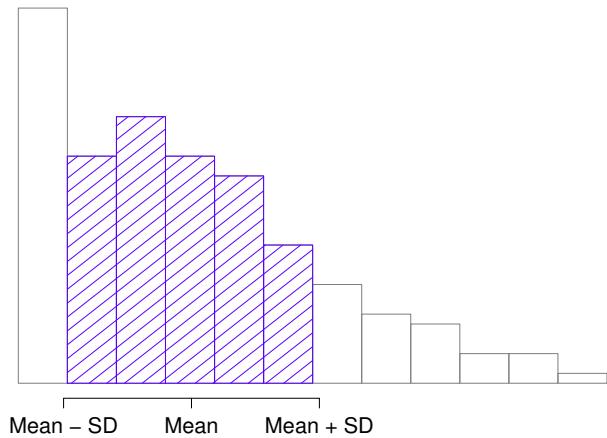


Figure3.9 AFL 得点差データについての標準偏差を描いたもの。色がついているヒストグラムの箇所は平均から 1 標準偏差のなかに入ったデータの数を表しています。今回は 65.3% のデータセットがこの範囲内に入り、次のメイントピックスである“約 68%”のルールに近い結果になっています。

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。

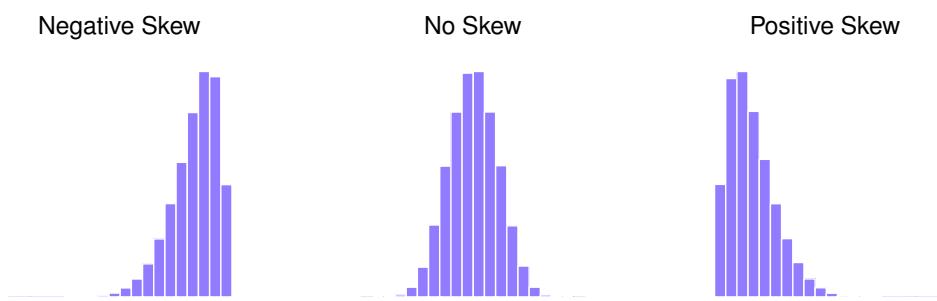


Figure 3.10 歪度のイメージ。左側は歪度が負 (歪度 = -.93), 真ん中は歪みなし (実際ほとんどありません。歪度 = -.006), そして右が正の歪度 (歪度 = -.006) をもつデータです。

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 ??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 ?? 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

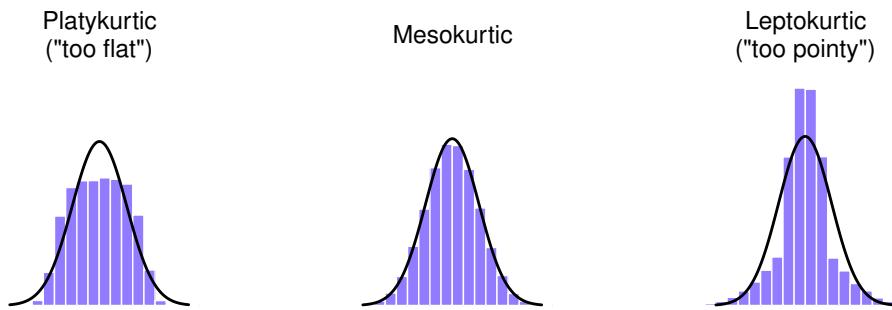


Figure 3.11 尖度の図。左側は“緩く尖った”データセット（尖度 = -.95）であり、これが意味するのはこのデータセットは“かなりフラット”だということです。真ん中の図は“中程度の尖り”をもったデータセット（尖度はほとんど 0）であり、これが意味するのはこのデータの尖度がちょうどいい感じであるということです。最後に、右側の図ですが、“尖度の大きい”データセット（尖度 = 2.12）であり、このデータセットは“とても尖っています”。尖度は正規分布（黒い線）と比べて評価されていることに注意してください。

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。<sup>a</sup>

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

<sup>a</sup>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??のような結果が得られます。

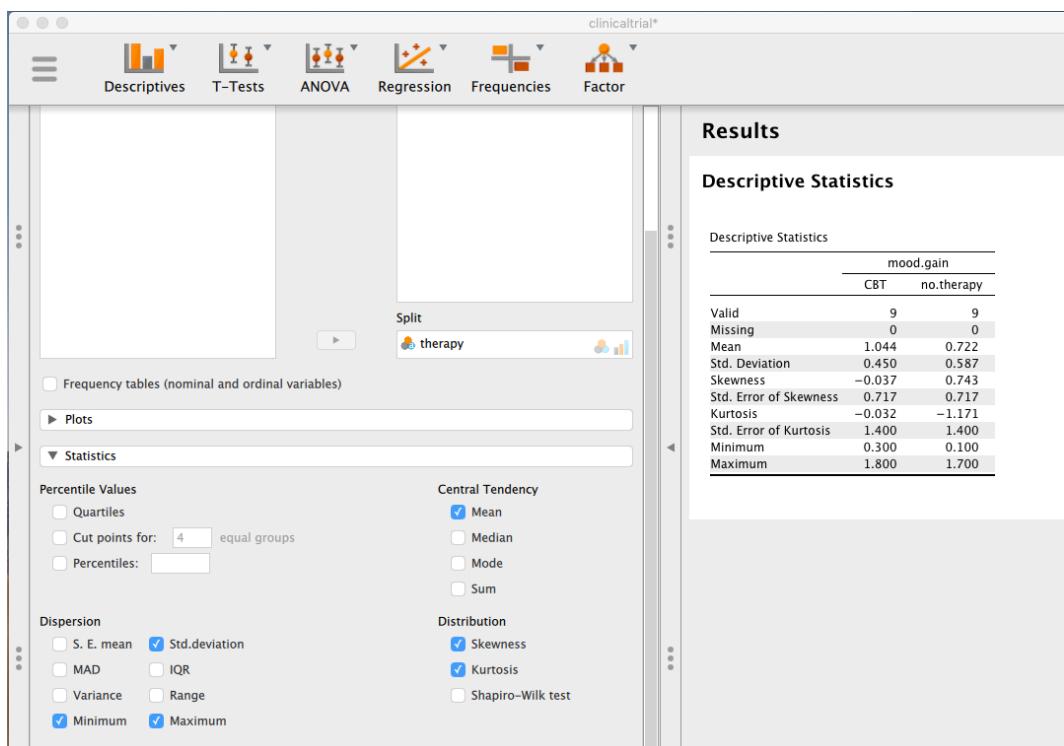


Figure3.13 セラピータイプごとに分割した記述統計量を示した JASP のスクリーンショット

### 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 ??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<sup>\*5</sup>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 ??）。
- 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 ??）。

<sup>\*5</sup>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 ??)。
- 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 ??)。
- 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 ??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があります。

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<sup>\*1</sup>

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。そのためには、図 ??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、ということがあります。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

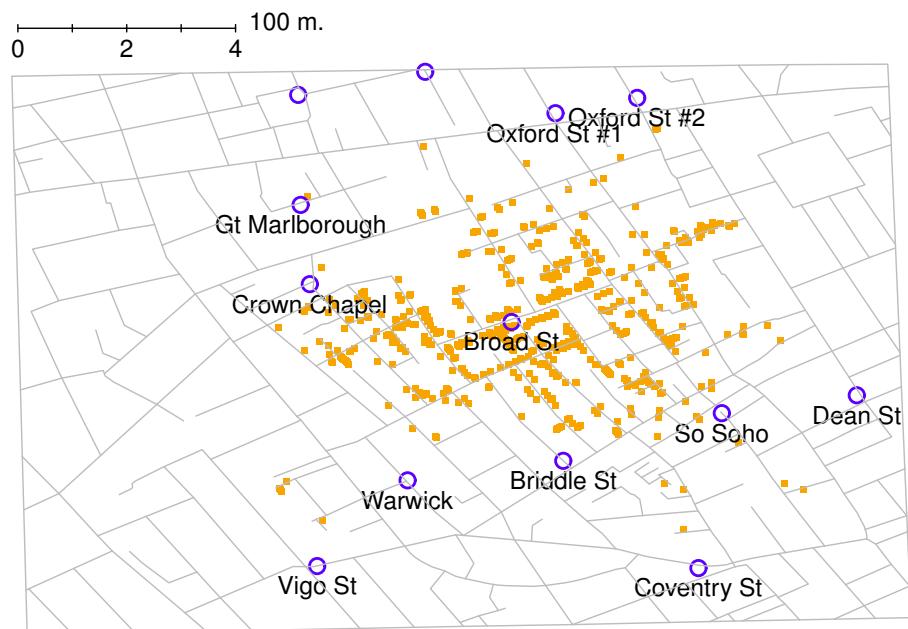


Figure4.1 John Snow のスタイリッシュなコロナマップのオリジナル。小さな各点はコロナ発生点で、大きな円は井戸の位置を示しています。このプロットが明らかにしたように、コロナのアウトブレイクは Broad St のポンプを中心にしていることがわかります。

## 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 ?? 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 ?? の 図 ?? 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 ?? 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 ?? 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ました。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

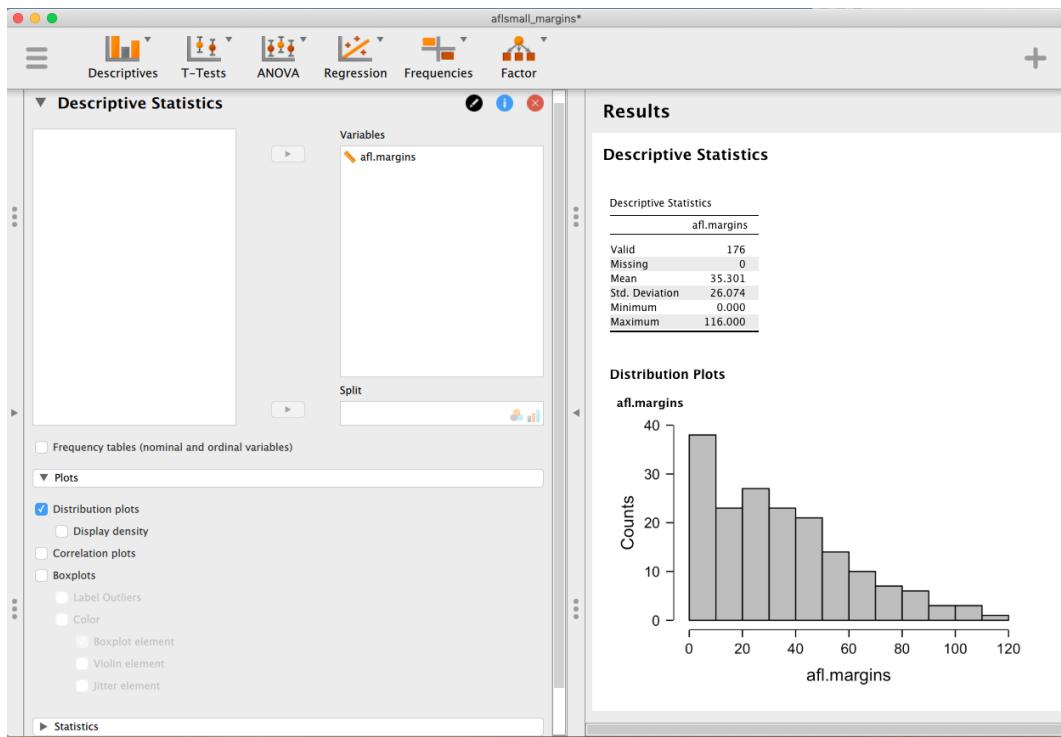


Figure4.2 ‘分布のプロット’ オプションによって作られたヒストグラムを描いた JASP のスクリーンショット

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??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一目瞭

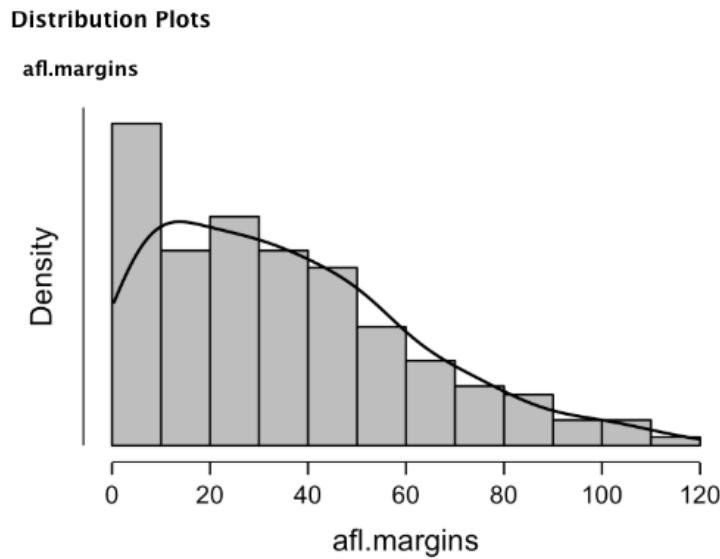


Figure4.3 afl.margins 変数の JASP による密度プロット

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  $25 \text{ パーセンタイル点} - (1.5 * \text{IQR})$  、上限は  $75 \text{ パーセンタイル点} + (1.5 * \text{IQR})$  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 ??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、

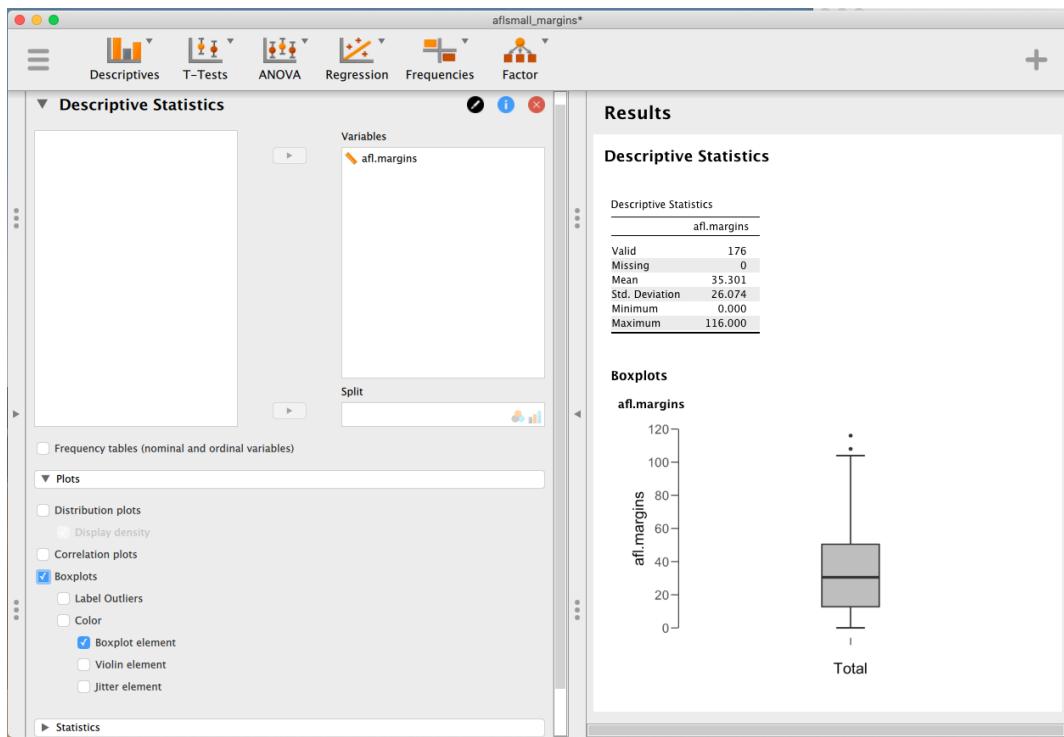


Figure4.4 JASPによるafl.margins変数のボックスプロット

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??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

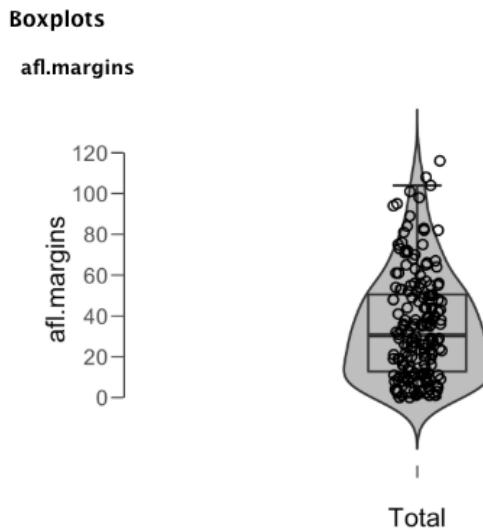


Figure4.5 JASPにおける[afl.margins](#)変数のバイオリンプロットにボックスプロットとデータ点も重ねてみました

.....

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 \_\_\_\_\_

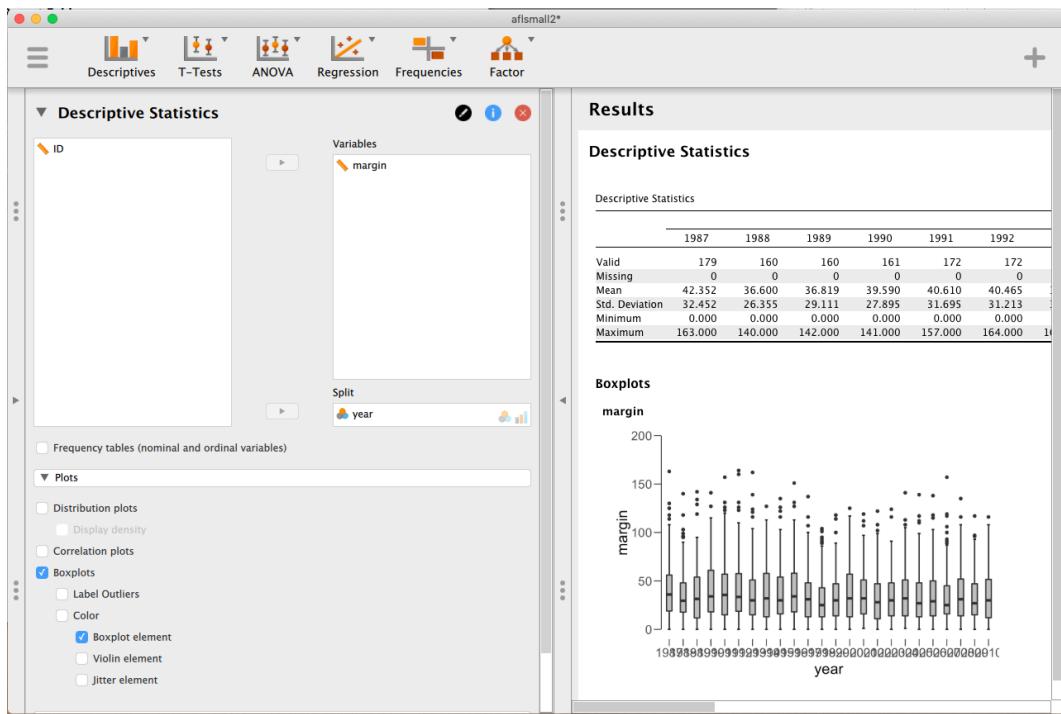


Figure4.6 JASPにおける複数ボックスプロット。aflsmall12データセットにおける、年度変数ごとの勝率

## 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 ??) とボックスプロット (セクション ??) 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 ??)

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 ??), sampling and estimation (Chapter ??) and statistical hypothesis testing (Chapter ??). 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.\*<sup>2</sup> 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:

---

\*<sup>2</sup>Source: <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,T,H,H,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 ??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<sup>\*1</sup>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 ??を見てください）。

### 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

<sup>\*1</sup>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。

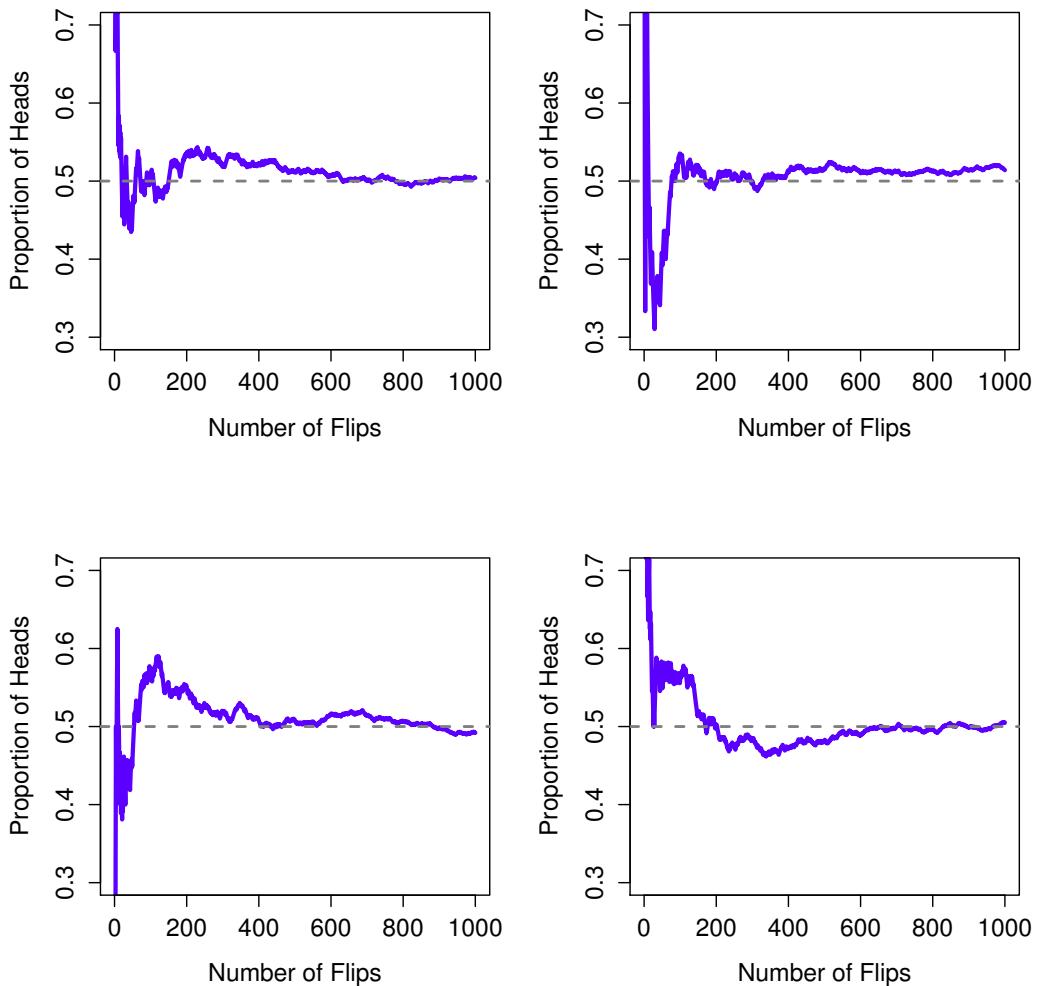


Figure5.1 頻度主義者の確率がどのように働くかの図。コインフリップを何度も何度も繰り返すと、表が出る割合は徐々に落ちていき、真の確率である 0.5 に収束していきます。各パネルが示すのは、四つの異なるシミュレーション実験の結果です。それぞれのケースについて、コインを 1000 回フリップして表が出る割合を追跡し続けたものです。どのケースも最終的にちょうど 0.5 になりますが、もしこれを無限回実験し続けるようにすれば、最終的にはそうなるはずです。

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 ??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！

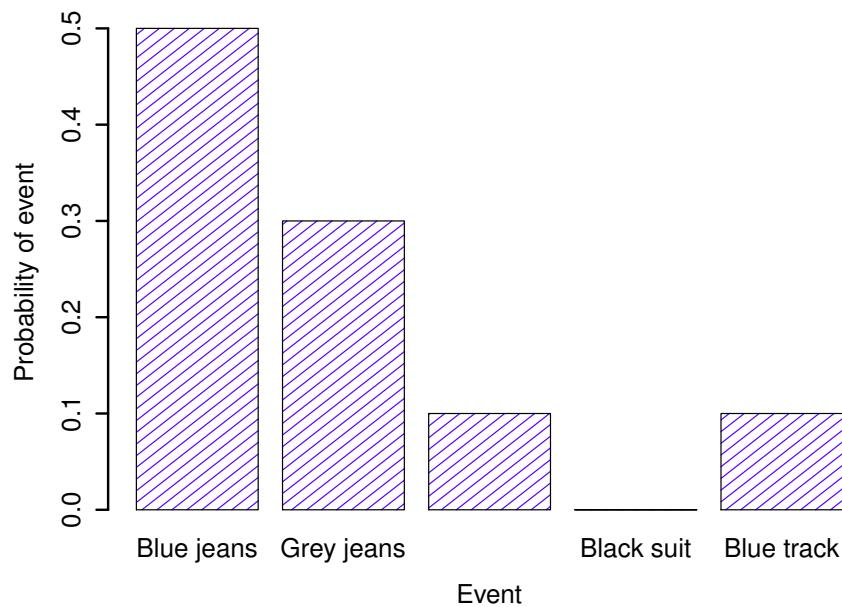


Figure5.2 “ズボン” 確率分布の視覚的記述. ここには 5 つの“根元事象” がって、それは私の持っている 5 本のズボンに対応しています。各事象はなんらかの生起確率を持っています。この数字は 0 から 1 の間の値です。これらの確率すべてを足すと 1 になります。

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 ??に示してあります。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、 $\chi^2$  (“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 $\theta$ 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。<sup>\*2</sup> 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

---

<sup>\*2</sup>“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 $\theta$ 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  $\theta$  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 ??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、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 ??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 ??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  $m$ 、 $\theta$  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 ??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 ??b 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。

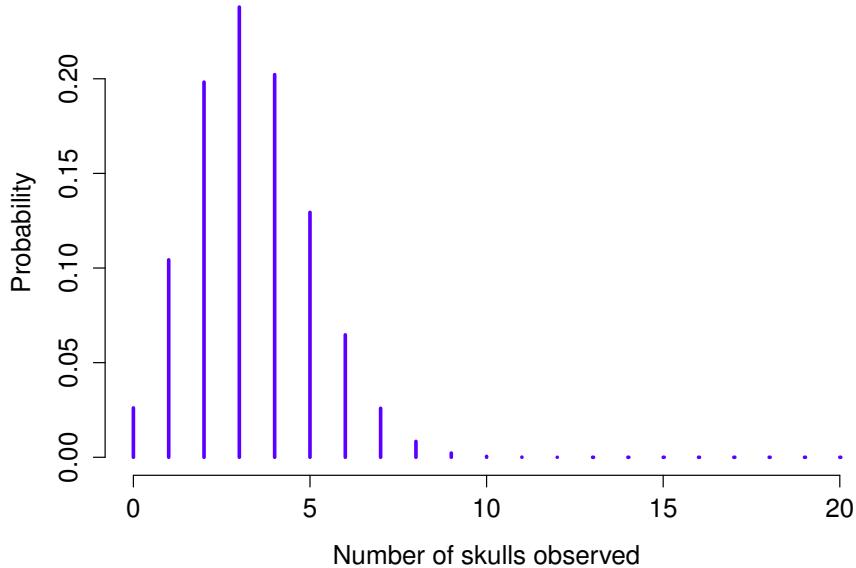


Figure 5.3 サイズパラメータが  $N = 20$  で成功確率  $\theta = 1/6$  の二項分布。縦のバーそれぞれがそのアウトカムの確率を表しています(すなわち、値  $X$  の確率)。これは確率分布なので、それぞれの確率は 0 から 1 の間にに入る数字であり、バーの高さを足し合わせると 1 にならなければなりません。

## 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  $\mu$  と、分布の標準偏差を表す  $\sigma$  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

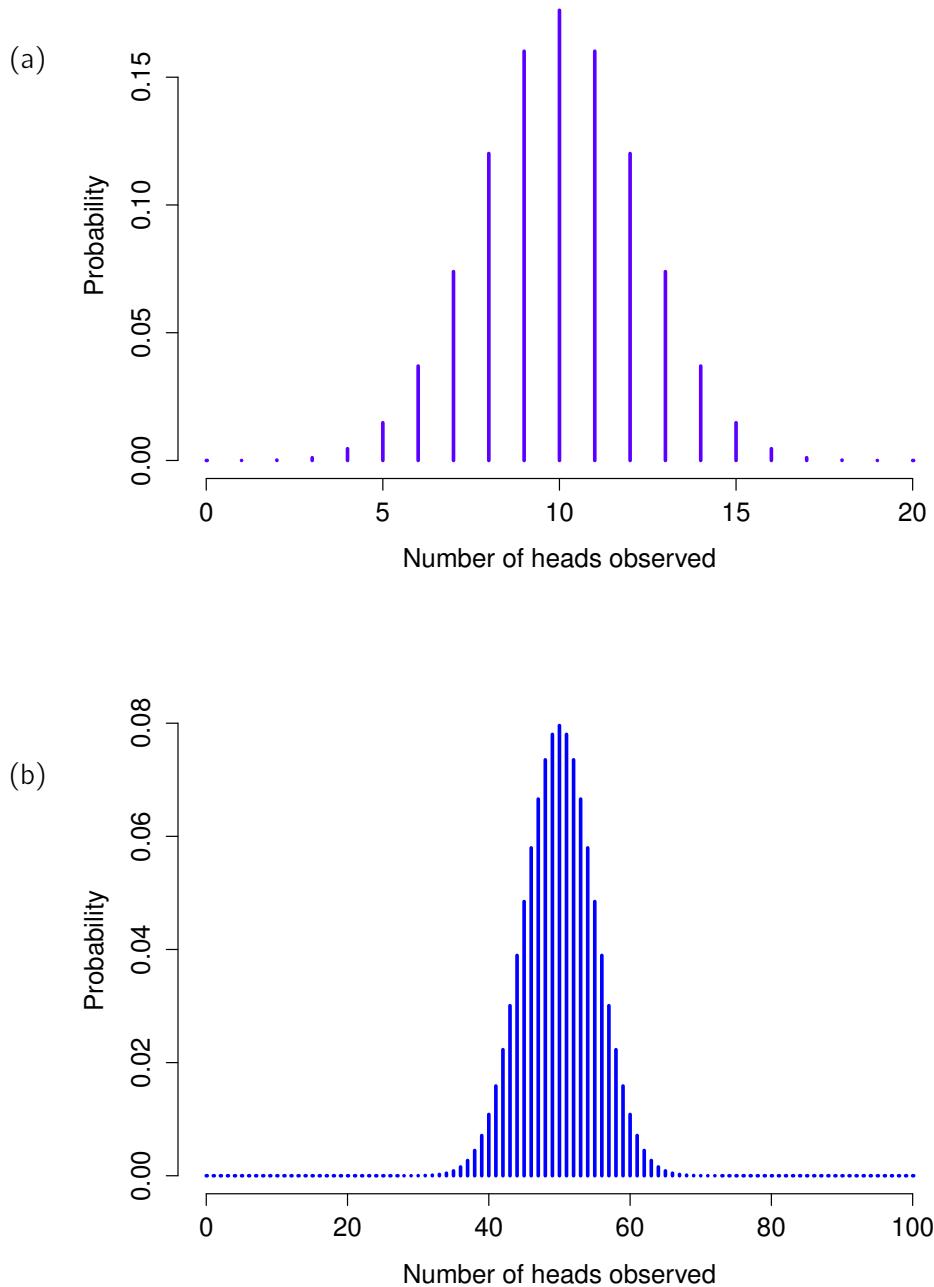


Figure 5.4 私が公平なコインをフリップするシナリオについての二項分布で、想定される確率は  $\theta = 1/2$  とします。パネル a では、コインを  $N = 20$  回、パネル b では  $N = 100$  回フリップしたものです。

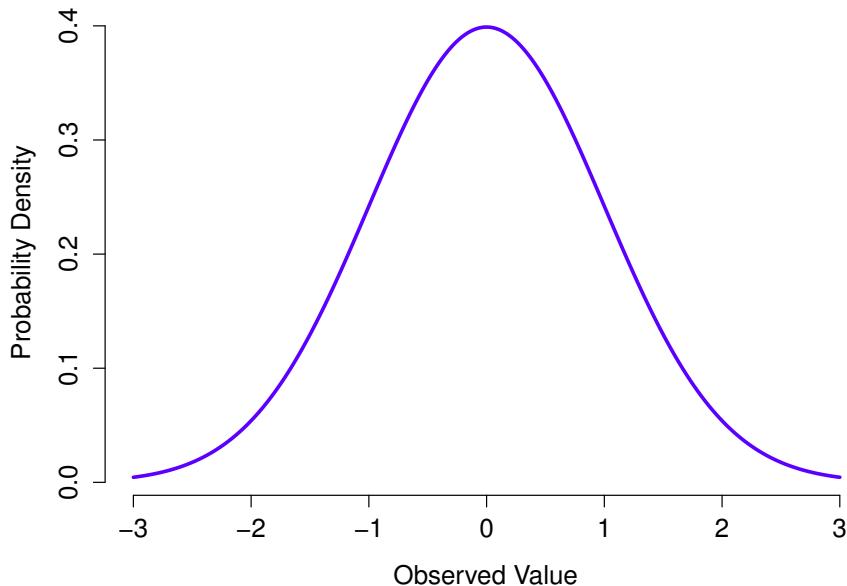


Figure5.5 平均  $\mu = 0$  で標準偏差が  $\sigma = 1$  の正規分布。 $x$ -軸はある変数の値に対応しており、 $y$ -軸はその値を我々がどの程度観測しやすいかを教えてくれます。ですが、 $y$  軸にあるのは“確率密度”であって“確率”ではないことに注意してください。ここには連続的分布ならではの、微妙でちょっと腹立たしい特徴があつて、 $y$  軸の振る舞いはちょっと奇妙なのです。すなわち、このカーブの高さは、実際には  $x$  の値を観測する確率を表しているわけではないのです。一方で、このカーブの高さは、どの  $x$  の値がより生じやすいか（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 ??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 ??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 ??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 ??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たのです。図 ??では、例えば、バーは  $X = 3$  や、 $X = 4$  に位置することはありましたが、その間に

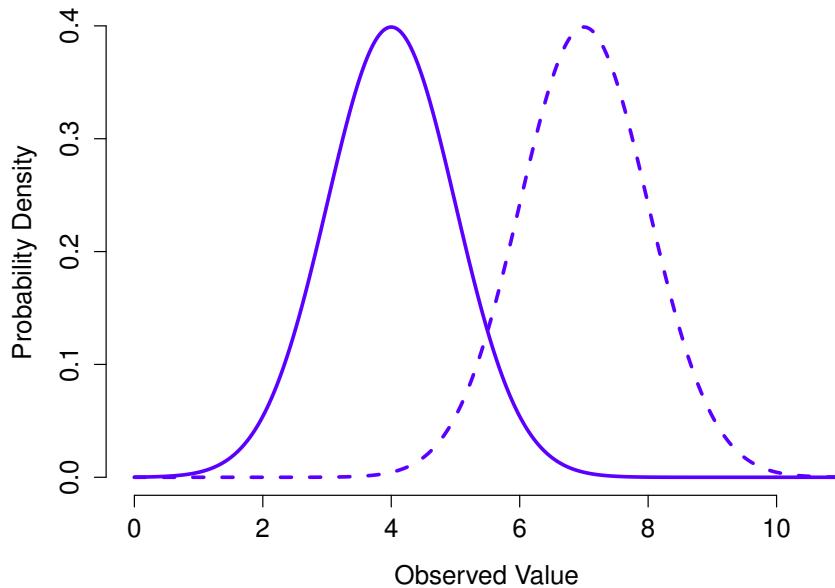


Figure 5.6 正規分布の平均を変えたら何が起こるかを描いたもの。実線は平均  $\mu = 4$  の正規分布。点線は平均  $\mu = 7$  の正規分布を表しています。どちらも、標準偏差は  $\sigma = 1$  です。はたして、二つの分布は同じ形をしており、点線が右側にずれています。

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當だろう、ということになります<sup>\*3</sup>

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 ??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あるが、分布がどんどん幅広くなることが図 ??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い

---

<sup>\*3</sup> 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。

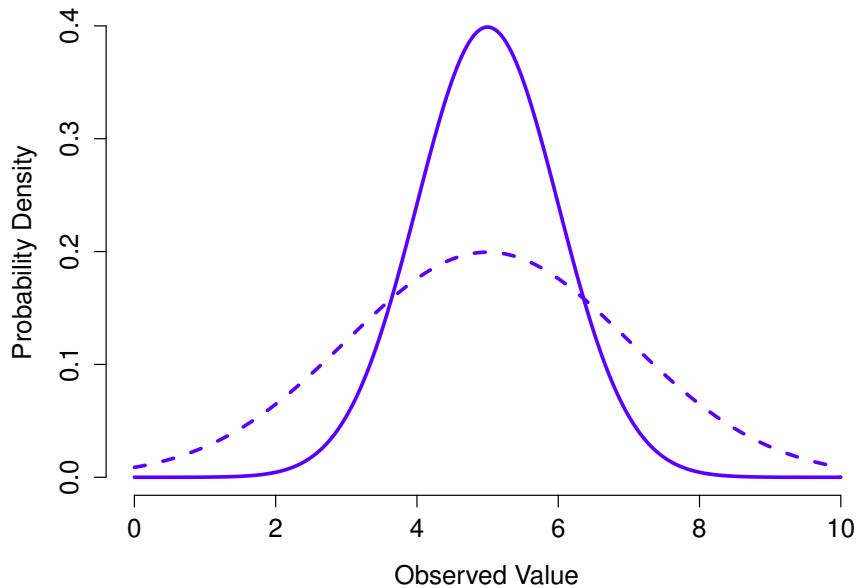


Figure 5.7 正規分布の標準偏差を変えた時に何が起こるかを描いたもの。どちらの分布も平均が  $\mu = 5$  ですが、標準偏差が異なります。実線は標準偏差  $\sigma = 1$  の分布で、点線は標準偏差  $\sigma = 2$  の分布です。結果として、分布は同じスポットに“中心化”していますが、点線が実線にくらべて幅が広くなっています。

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 ?? 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

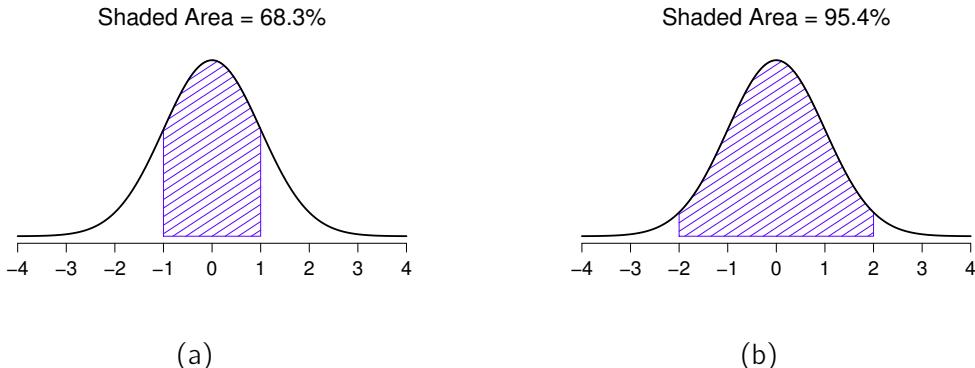


Figure5.8 カーブの下の面積は観測値が特定の範囲で得られる確率を教えてくれます。点線の正規分布は、平均  $\mu = 0$  で標準偏差  $\sigma = 1$  です。影のついた領域は、二つの重要なケースにおける“カーブの下の面積”です。パネル a では、平均周りの 1 標準偏差の中に観測値が得られる確率が 68.3% であることを見てとることができます。パネル b では、平均周り 2 標準偏差の中に観測値が得られる 確率が 95.4% であることを見てとることができます。

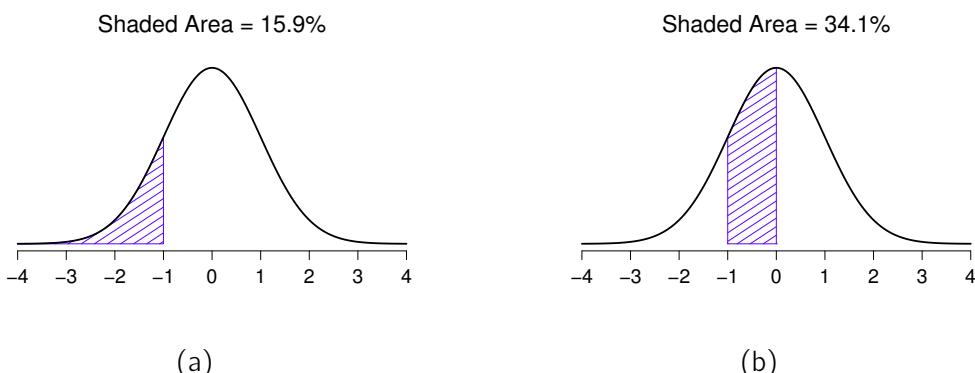


Figure5.9 “カーブの下の領域”についてのさらに二つの例です。平均より 1 標準偏差下より小さい観測値が得られる確率は 15.9% で (パネル a), 平均と平均より 1 標準偏差下の区間に観測値が得られる確率は 34.1% です (パネル b)。この二つの数字を足し合わせると,  $15.9\% + 34.1\% = 50\%$  になることがわかりますね。正規分布するデータでは, 平均より下の観測値が得られる確率は, 50% になります。そしてもちろん, このことは平均以上の観測値が得られる確率が 50% になることも意味します。

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す。

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 ??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 ??a 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、 $\chi^2$  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。

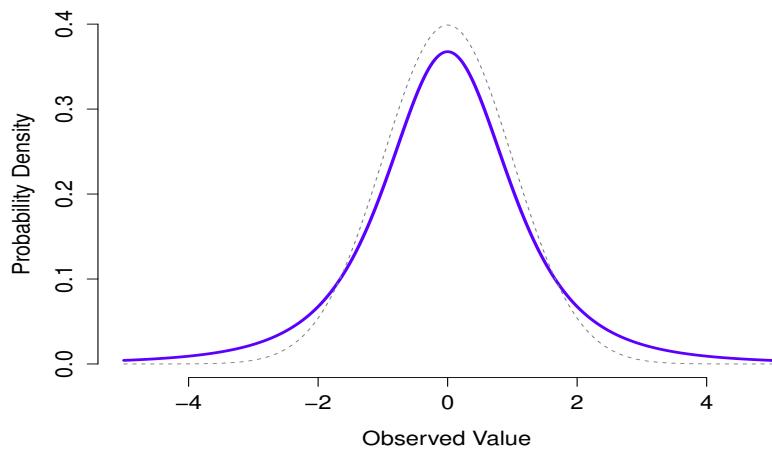


Figure 5.10 自由度 3 の  $t$  分布 (実線)。正規分布に似ているようですが、全く同じというわけではありません。比較のために、標準正規分布を点線でプロットしました。

- **$t$  分布**は連続分布で正規分布によく似ています。図 ??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。
- **$\chi^2$  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 ??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  $\chi^2$  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  $\chi^2$  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  $\chi^2$  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 ??.
- **$F$  分布**は  $\chi^2$  分布に少し似ていて、二つの  $\chi^2$  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。 $\chi^2$  の話をした時に、“平方和”を使う際の大事な分布だと言ったことを覚えていましたか? 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。

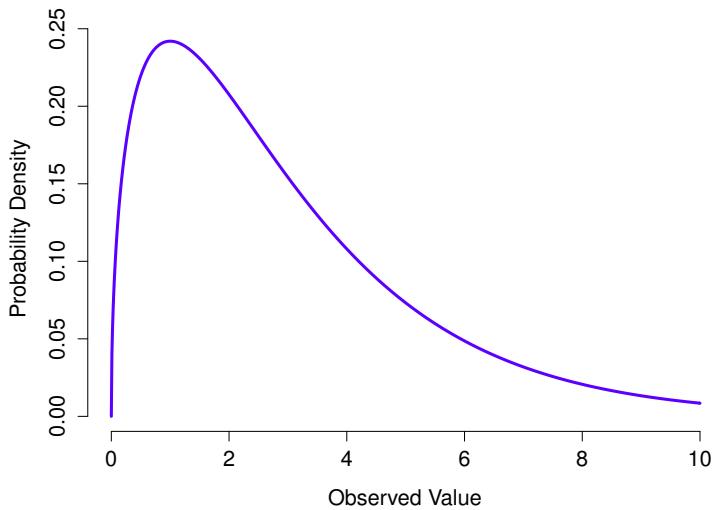


Figure5.11 自由度 3 の  $\chi^2$  分布。観測値は 0 より大きくなければなりませんし、この分布は少し歪んでいることに注意が必要です。そこにカイ二乗分布の特徴があります。

---

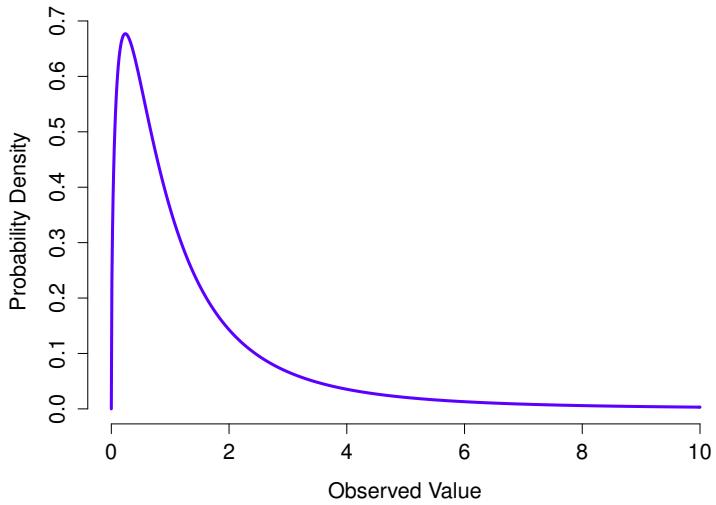


Figure5.12 自由度 3 と 5 の  $F$  分布。定性的にいうなら、カイ二乗分布に少し似ているように見えますが、一般的には全く似ていません。

---

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 ?? 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  $\chi^2$  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  $\chi^2$  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 ??)
- 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 ??)
- 確率理論の基礎 (セクション ??)
- 二項分布 (セクション ??), 正規分布 (セクション ??), そのほかの分布 (セクション ??)

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 ??, 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 ?? through ?? are focused on sampling theory, and Sections ?? and ?? 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 ??) 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?

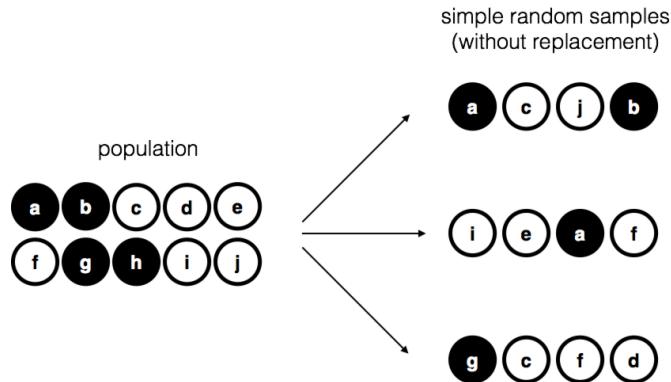


Figure6.1 Simple random sampling without replacement from a finite population

---

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 ???. 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 ???. 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.\*<sup>1</sup> 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 ???. 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 ???.

---

\*<sup>1</sup>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.

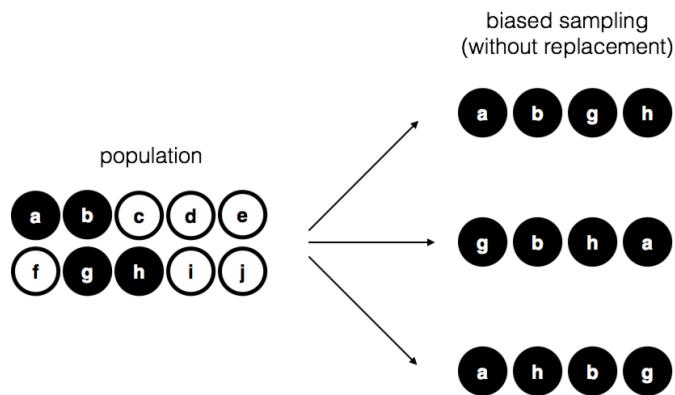


Figure6.2 Biased sampling without replacement from a finite population

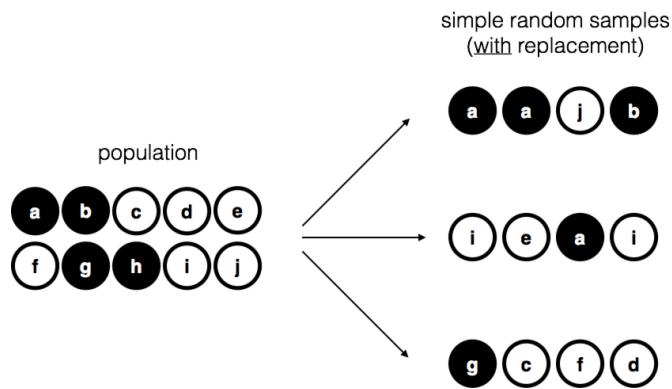


Figure6.3 Simple random sampling *with* replacement from a finite population

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<sup>\*2</sup> 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

---

<sup>\*2</sup>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 ?? and ?? . 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

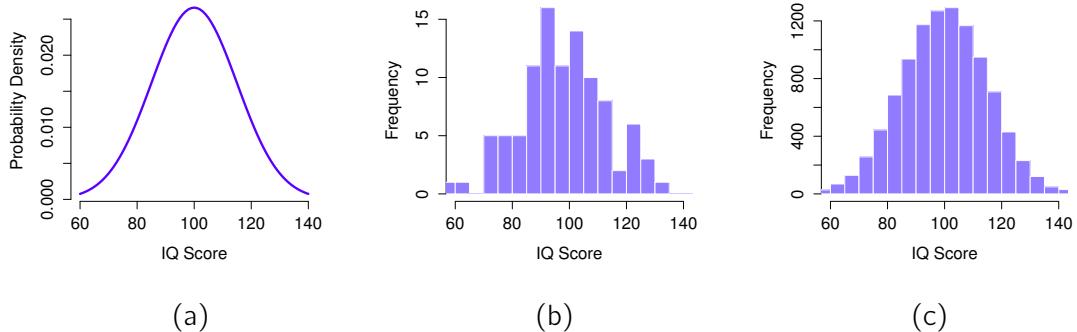


Figure 6.4 The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and in panel c we have a sample of 10,000 observations.

.....

to work with. You've already come across these objects in Chapter ???. 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 ??a. 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  $\mu$  is 100 and the population standard deviation  $\sigma$  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 ??b. 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 ??a. 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 ??) and how to work out how confident you are in your estimates (Section ??) 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 ??

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 ( $\bar{X} \rightarrow \mu$ ).<sup>\*3</sup>

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.

---

<sup>\*3</sup>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 ??). 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 ??, 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-

Table6.1 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 ??, 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 ?? . 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 ?? 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 ???. 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 ??, 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 ???. 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 ??b 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 ??a. When I increase the sample size to  $N = 4$ , the sampling distribution of the mean is very close to normal (Figure ??c), 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  $\mu$  and standard deviation  $\sigma$ , then the sampling distribution of the mean also has mean  $\mu$  and the standard error of the mean is

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

Because we divide the population standard deviation  $\sigma$  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.\*<sup>4</sup>

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

---

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

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.<sup>\*5</sup> 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  $\mu$ , 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

---

<sup>\*5</sup>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
$\mu$	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 ???. 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 ??b 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 ???. 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.<sup>\*6</sup> The plot on the right is quite different: on average, the sample standard deviation  $s$  is *smaller* than the population standard deviation  $\sigma$ . 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  $\sigma$  we should make sure our guess is a little bit larger than the sample standard deviation  $s$ .

---

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

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

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

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

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

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

---

<sup>a</sup>Okay, I'm hiding something else here. In a bizarre and counter-intuitive twist, since  $\hat{\sigma}^2$  is an unbiased estimator of  $\sigma^2$ , you'd assume that taking the square root would be fine and  $\hat{\sigma}$  would be an unbiased estimator of  $\sigma$ . 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  $\sigma^2$ , but when you take the square root, it turns out that  $\hat{\sigma}$  is a biased estimator of the population standard deviation  $\sigma$ . 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
$\sigma$	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
$\sigma^2$	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<sup>\*7</sup>

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

---

<sup>\*7</sup>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  $\mu$  and the standard deviation is  $\sigma$ . 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 ??) that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution Section ?? that there is a 95% chance that a normally-distributed quantity will fall within about two standard deviations of the true mean.

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

Mathematically, we write this as:

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

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

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

What this is telling us is that the range of values has a 95% probability of containing the population mean  $\mu$ . We refer to this range as a **95% confidence interval**, denoted  $\text{CI}_{95}$ . 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  $\sigma$ . Yet, in Section ?? I stressed the fact that we don't actually *know* the true population parameters. Because we don't know the true value of  $\sigma$  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  $\mu$  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 ??, 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 ??, 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 ??)
- Statistical theory of sampling: the law of large numbers (Section ??), sampling distributions and the central limit theorem (Section ??).
- Estimating means and standard deviations (Section ??)
- Estimating a confidence interval (Section ??)

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.

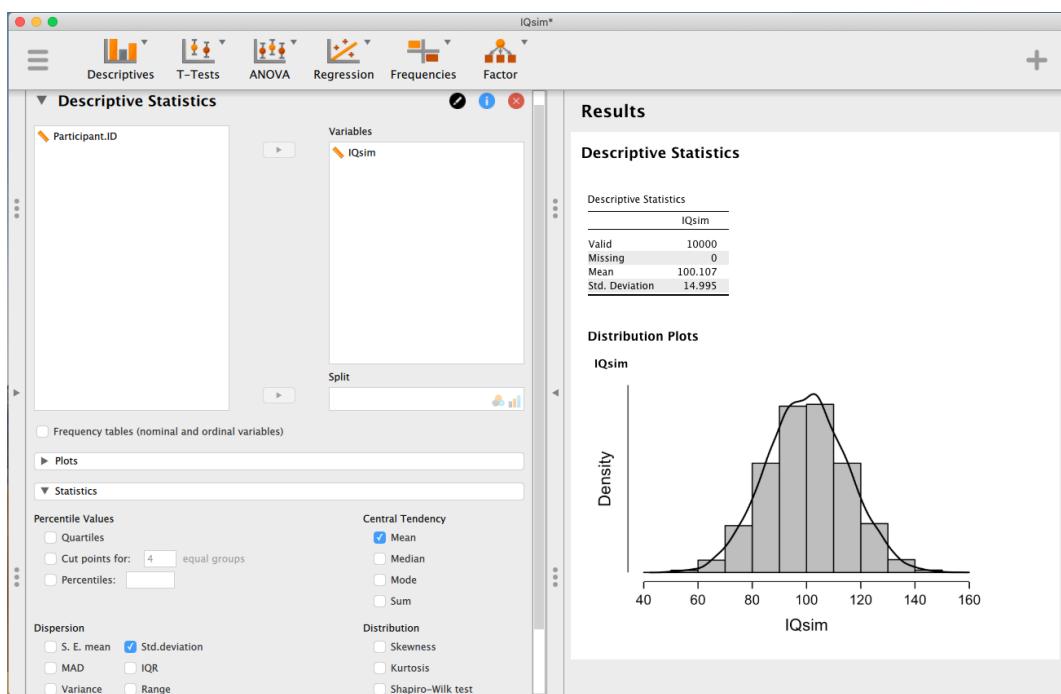


Figure 6.5 A random sample drawn from a normal distribution using JASP

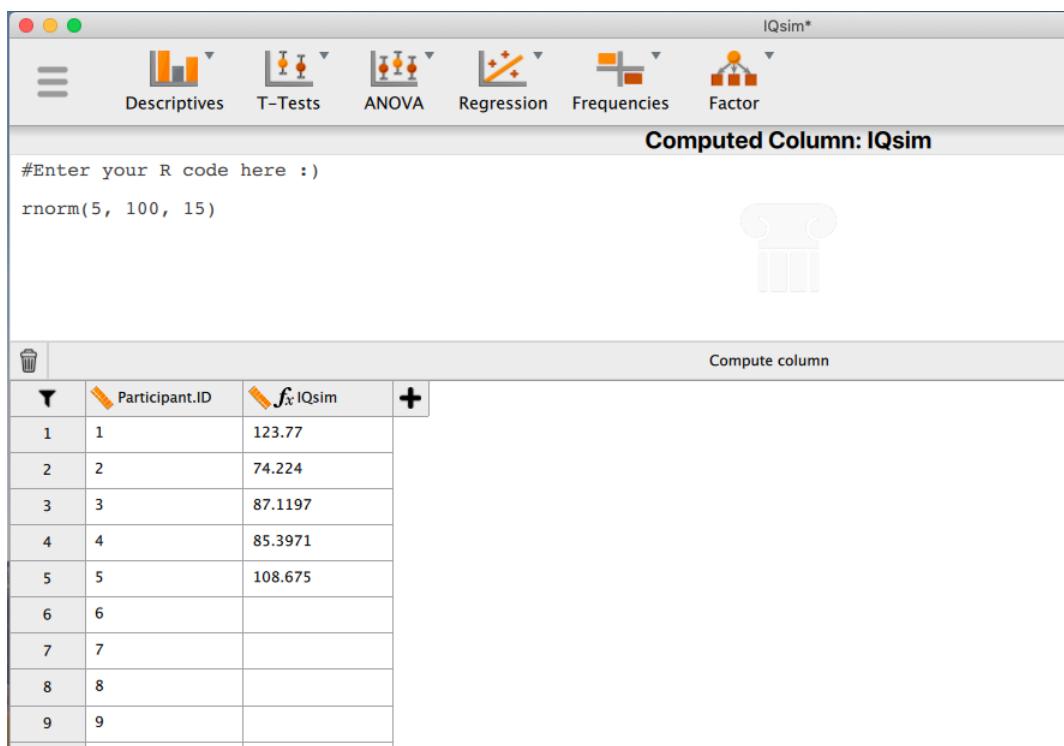


Figure6.6 Using JASP to draw a random sample of 5 from a normal distribution with  $\mu = 100$  and  $\sigma = 15$ .

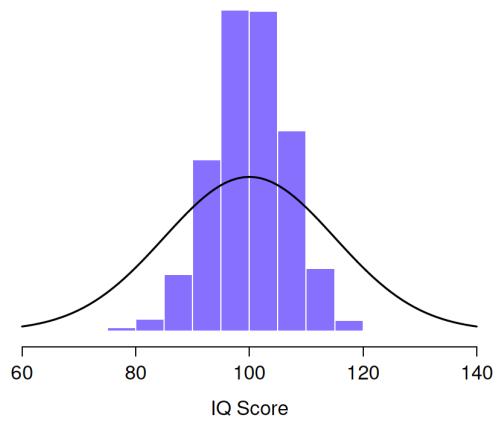


Figure6.7 The sampling distribution of the mean for the “five IQ scores experiment”. If you sample 5 people at random and calculate their *average* IQ you’ll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison,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

---

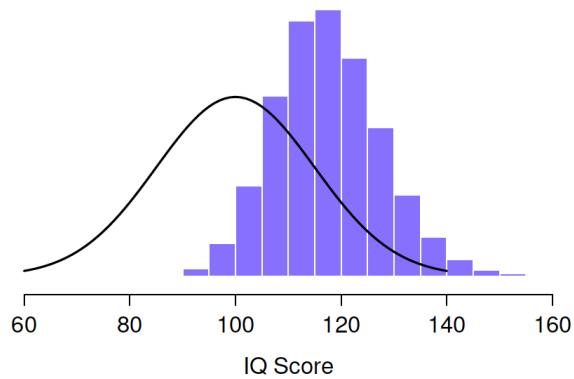


Figure6.8 The sampling distribution of the *maximum* for the “five IQ scores experiment”. If you sample 5 people at random and select the one with the highest IQ score you’ll probably see someone with an IQ between 100 and 140.

---

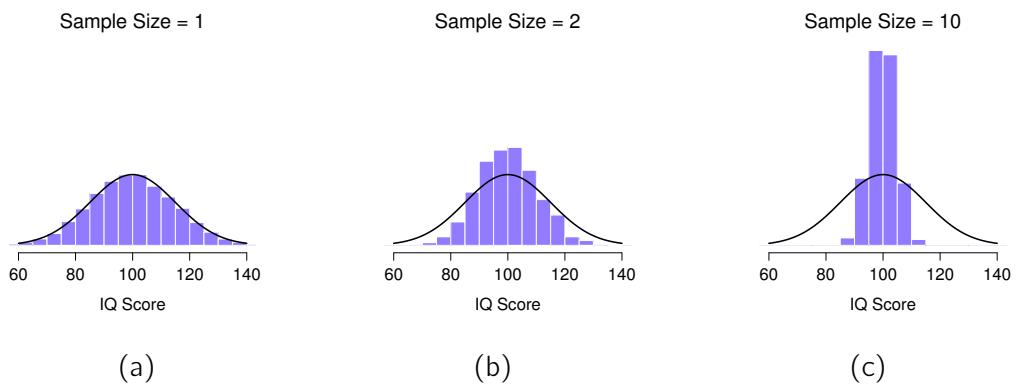
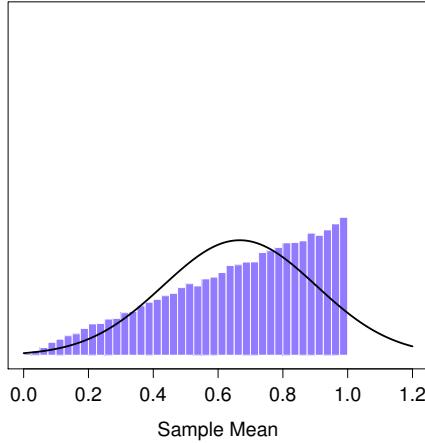


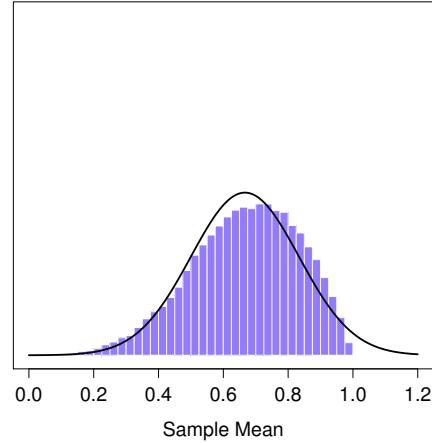
Figure 6.9 An illustration of the how sampling distribution of the mean depends on sample size. In each panel I generated 10,000 samples of IQ data and calculated the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling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



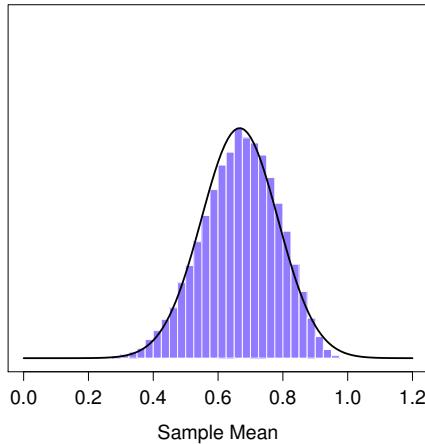
(a)

Sample Size = 2



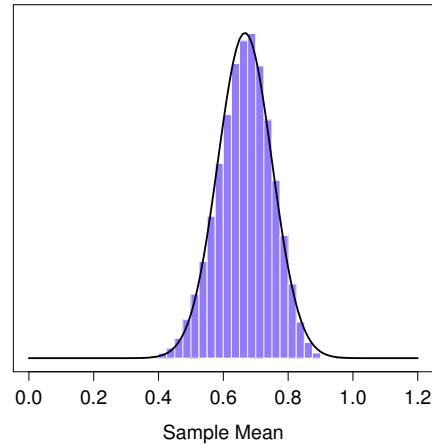
(b)

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.

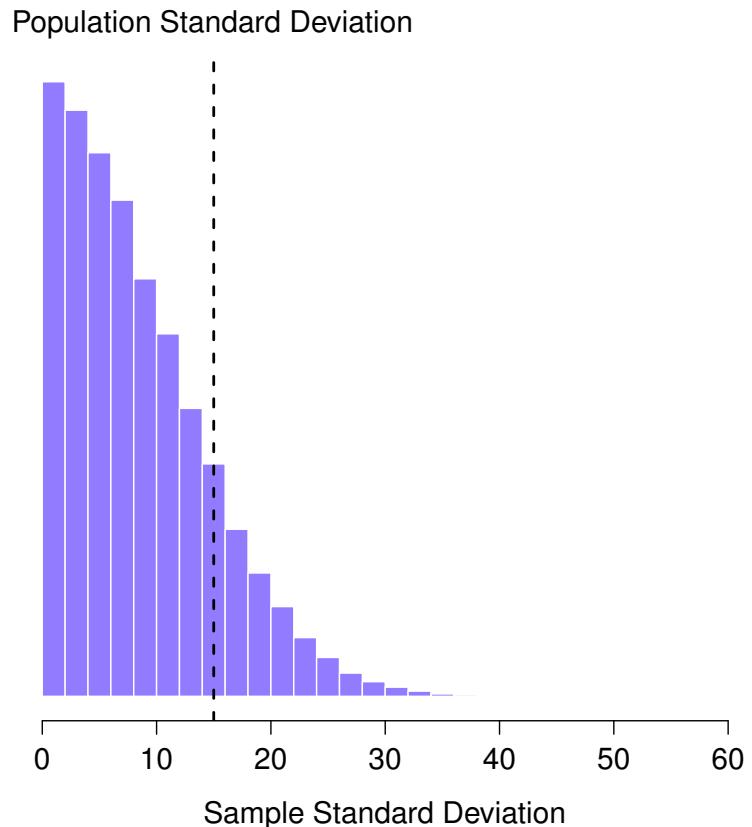


Figure 6.11 The sampling distribution of the sample standard deviation for a “two IQ scores” experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram the vast majority of experiments will produce a much smaller sample standard deviation than this.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

.....

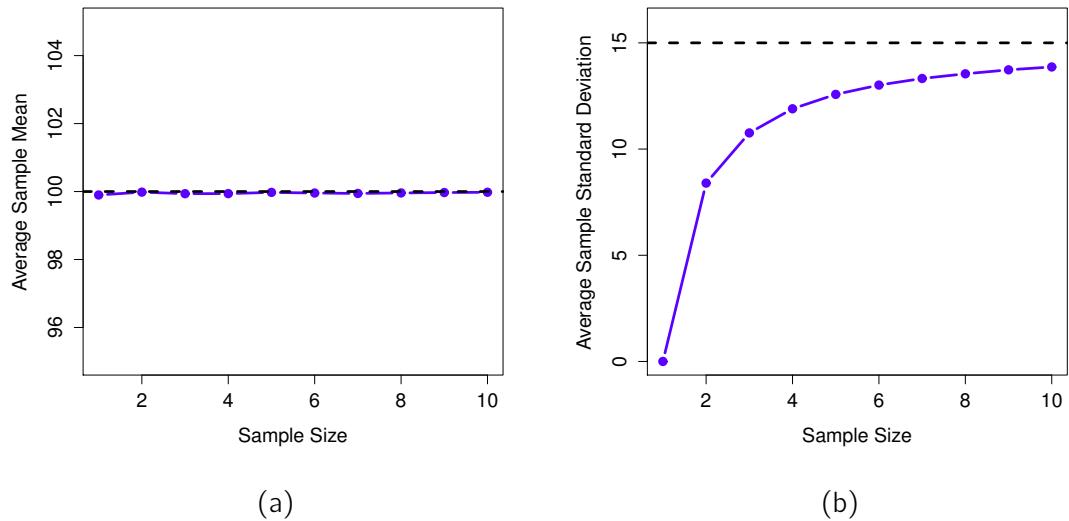


Figure 6.12 An illustration of the fact that the sample mean is an unbiased estimator of the population mean (panel a), but the sample standard deviation is a biased estimator of the population standard deviation (panel b). For the figure I generated 10,000 simulated data sets with 1 observation each,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

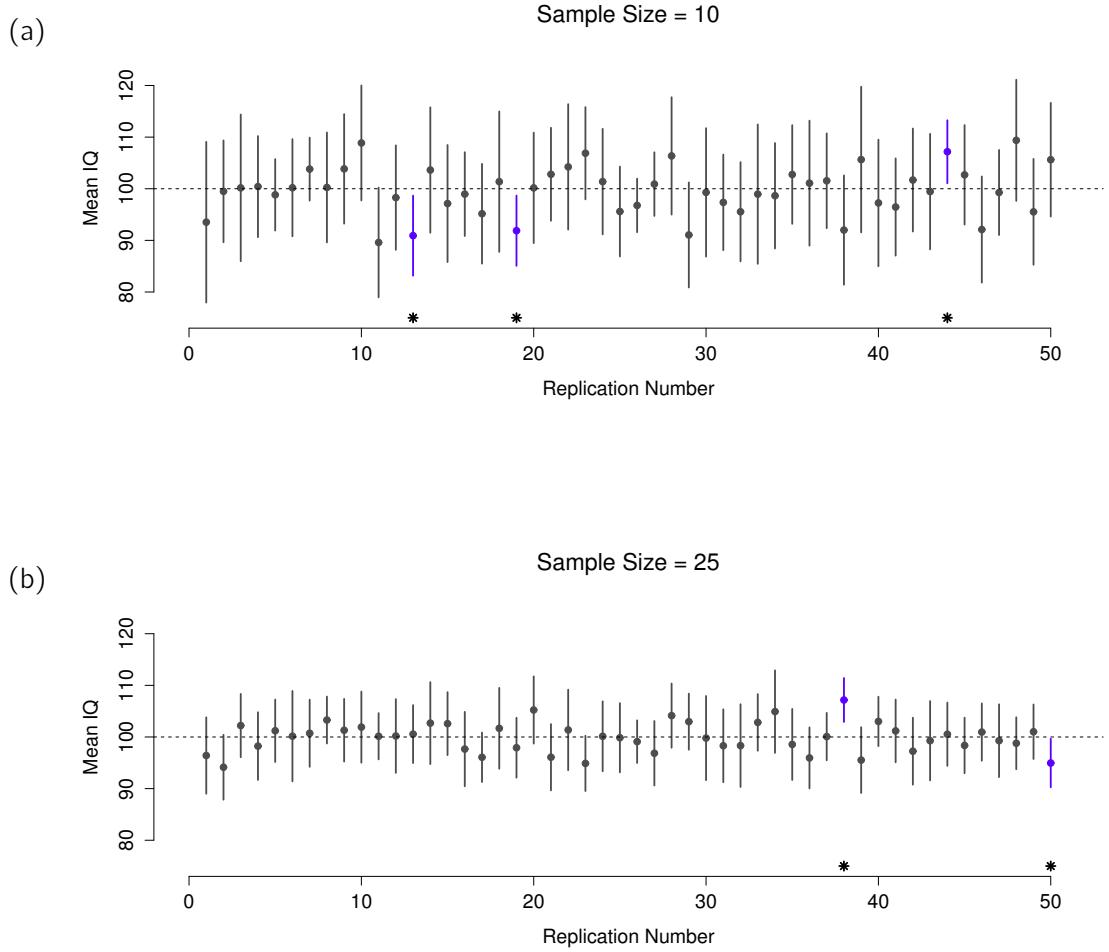


Figure 6.13 95% confidence intervals. The top (panel a) shows 50 simulated replications of an experiment in which we measure the IQs of 10 people. The dot marks the location of the sample mean and the line shows the 95% confidence interval. In total 47 of the 50 confidence intervals do contain the true 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<sup>\*1</sup>

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.<sup>\*2</sup> Afterwards, I’ll spend a bit of time talking about the various dogmas, rules and heresies that surround the theory of hypothesis testing.

---

<sup>\*1</sup>The quote comes from Wittgenstein’s (1922) text, *Tractatus Logico-Philosophicus*.

<sup>\*2</sup>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).<sup>\*3</sup>

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

---

<sup>\*3</sup>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  $\theta$  (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  $\theta$  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<sup>\*4</sup>, *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

---

<sup>\*4</sup>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.<sup>\*5</sup> 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  $\alpha$ , 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  $\alpha$  if the type I error rate is no larger than  $\alpha$ .

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  $\beta$ . 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:

---

<sup>\*5</sup>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)	$\alpha$ (type I error rate)
$H_0$ is false	$\beta$ (type II error rate)	$1 - \beta$ (power of the test)

A “powerful” hypothesis test is one that has a small value of  $\beta$ , while still keeping  $\alpha$  fixed at some (small) desired level. By convention, scientists make use of three different  $\alpha$  levels: .05, .01 and .001. Notice the asymmetry here; the tests are designed to *ensure* that the  $\alpha$  level is kept small but there’s no corresponding guarantee regarding  $\beta$ . 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

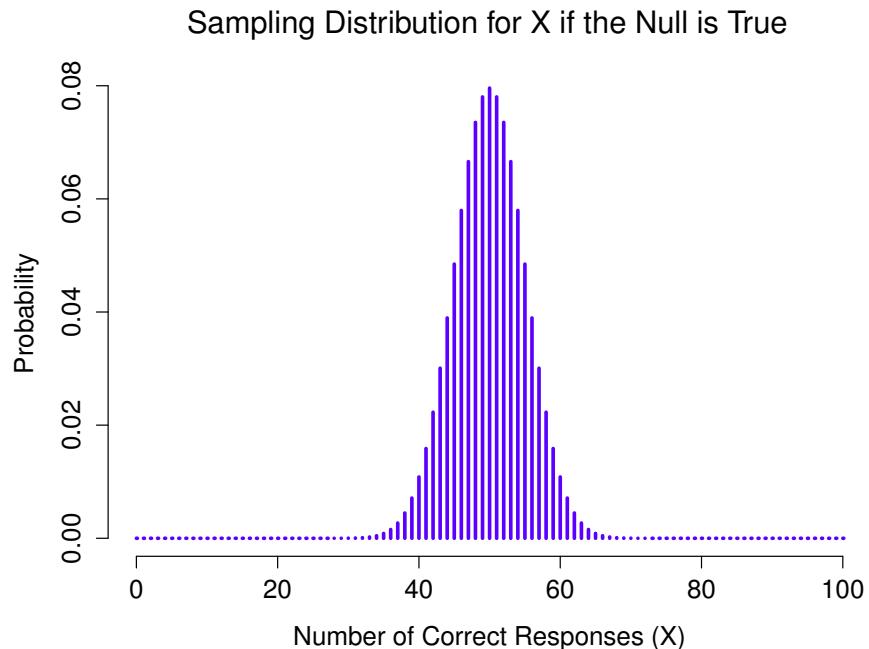


Figure 7.1 The sampling distribution for our test statistic  $X$  when the null hypothesis is true. For our ESP scenario this is a binomial distribution. Not surprisingly, since the null hypothesis says that the probability of a correct response is  $\theta = .5$ , the sampling distribution says that 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 ??). 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  $\theta$  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 ??, 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 ?? . 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

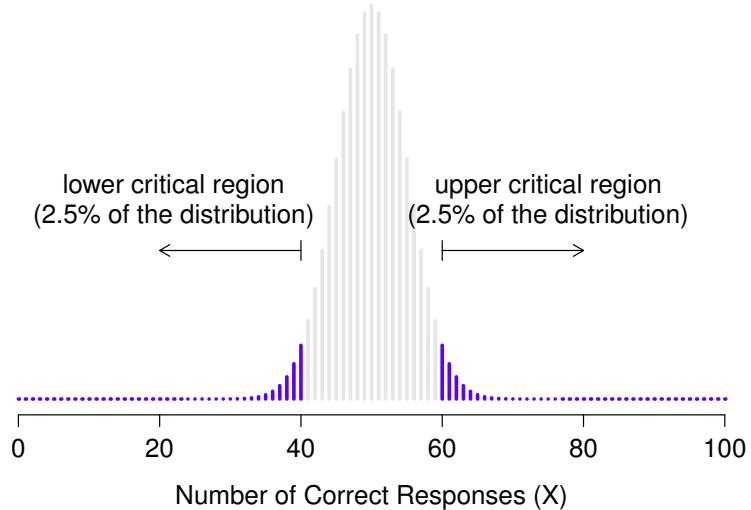


Figure 7.2 The critical region associated with the hypothesis test for the ESP study, for a hypothesis test with a significance level of  $\alpha = .05$ . The plot shows the sampling distribution of  $X$  under the null hypothesis (i.e., same as Figure ??). The grey bars correspond to those values of  $X$  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  $\alpha$  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  $\alpha$  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 ??.

we want  $\alpha = .05$  then our critical regions correspond to  $X \leq 40$  and  $X \geq 60$ .<sup>\*6</sup> 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.

---

<sup>\*6</sup>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  $\alpha$  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  $\alpha$  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<sup>\*7</sup>

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.

---

<sup>\*7</sup>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 ??.

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 ??.

7.5

---

## 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

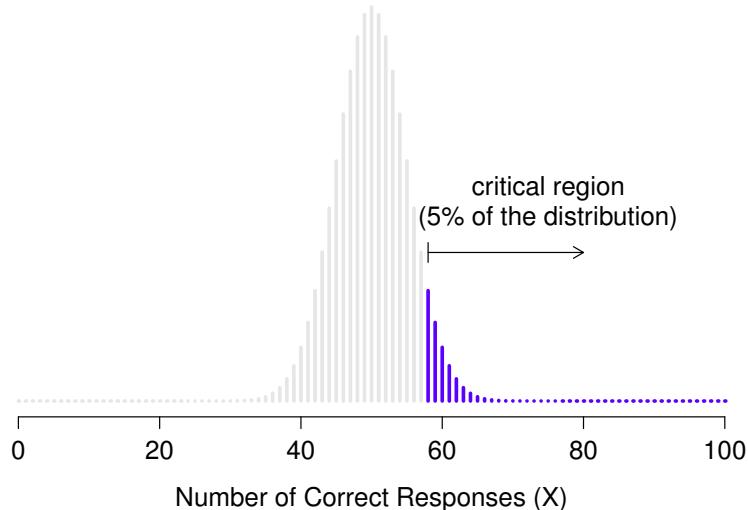


Figure 7.3 The critical region for a one sided test. In this case, the alternative hypothesis is that  $\theta > .5$  so we would only reject the null hypothesis for large values of  $X$ . As a consequence, the critical region only covers the upper tail of the sampling distribution, specifically the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

.....

#### 7.5.1 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  $\alpha$  in each case. When we do that for my original ESP data what we'd get is something like this

Value of $\alpha$	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  $\alpha$  levels of .03 and above, we'd always find ourselves rejecting the null hypothesis. For  $\alpha$  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  $\alpha$  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 ( $\alpha$ ) 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  $\alpha$  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<sup>\*8</sup> 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,  $\beta$ ). 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

---

<sup>\*8</sup>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  $\alpha$  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  $\alpha$  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  $\alpha$  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  $\alpha$  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 ???. 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  $\alpha$  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 ???. 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  $p = .5$ ,<sup>9</sup> 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 ??.

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.

---

<sup>9</sup>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  $\theta$  value.

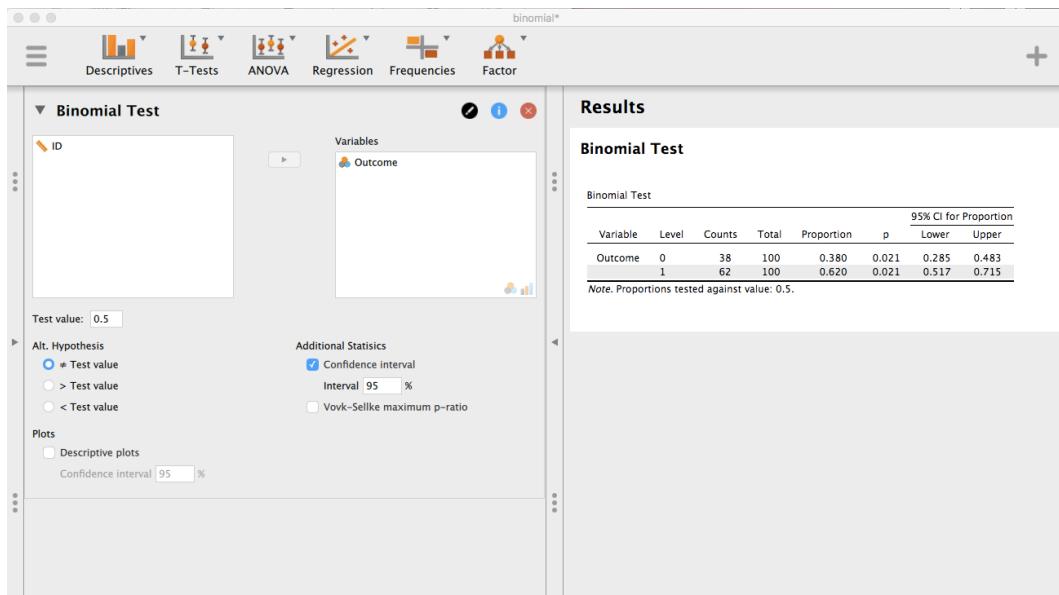


Figure 7.4 Binomial test analysis and results in JASP

## 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  $\beta$ , 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  $\beta$  that tells us the Type II error rate, in the

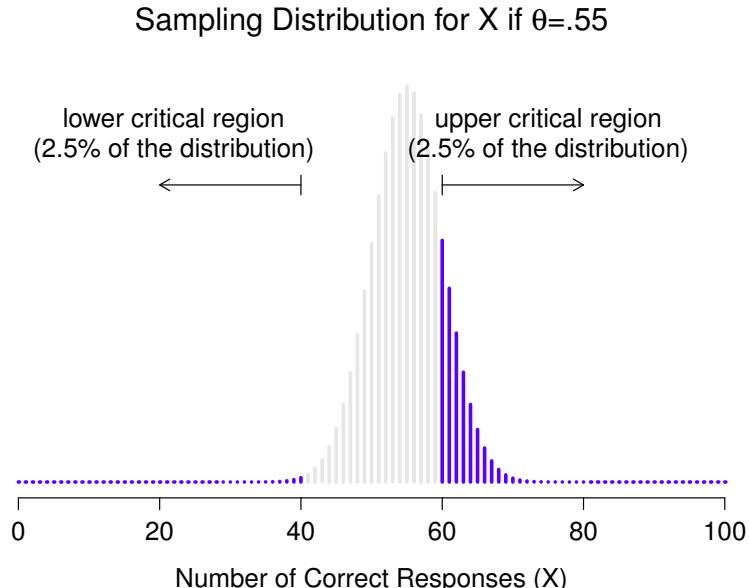


Figure 7.5 Sampling distribution under the *alternative* hypothesis for a population parameter value of  $\theta = 0.55$ . A reasonable proportion of the distribution lies in the rejection region.

---

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  $\theta$ . In fact, the alternative hypothesis corresponds to every value of  $\theta$  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 ???. 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  $\theta$  is actually 0.7. What happens to the sampling distribution when this occurs? The answer, shown in Figure ???, 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.

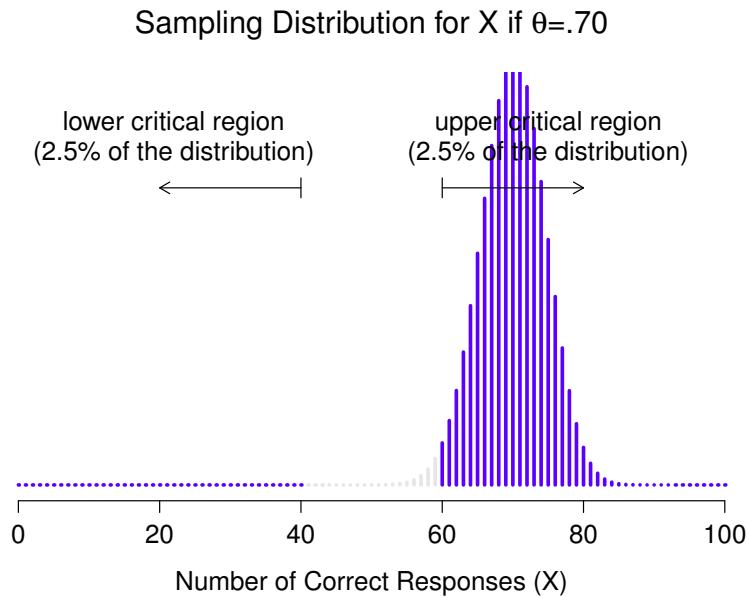


Figure7.6 Sampling distribution under the *alternative* hypothesis for a population parameter value of  $\theta = 0.70$ . Almost all of the distribution lies in the rejection region.

---

What all this means is that the power of a test (i.e.,  $1 - \beta$ ) depends on the true value of  $\theta$ . To illustrate this, I've calculated the expected probability of rejecting the null hypothesis for all values of  $\theta$ , and plotted it in Figure ???. 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  $\theta$ . As you can see, when the true value of  $\theta$  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 (Box1976)

The plot shown in Figure ?? 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 will

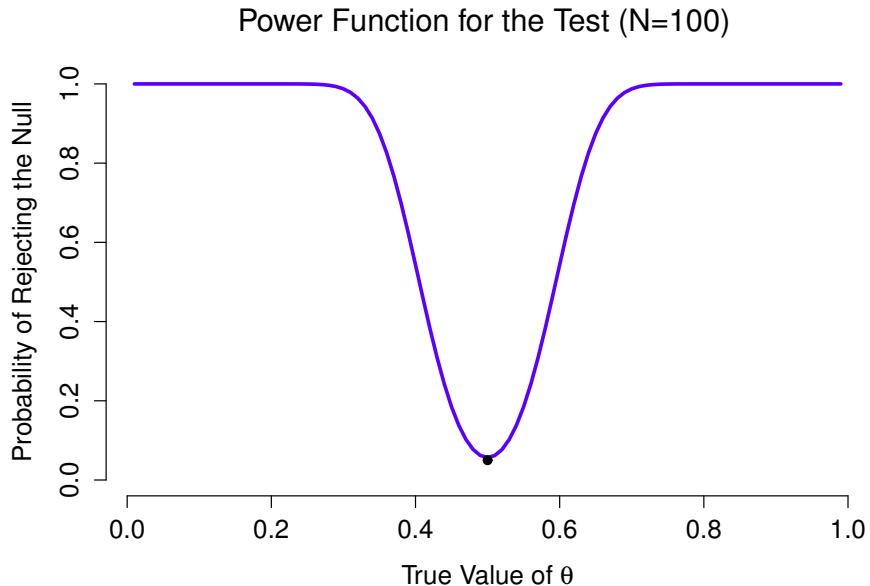


Figure 7.7 The probability that we will reject the null hypothesis, plotted as a function of the true value of  $\theta$ . Obviously, the test is more powerful (greater chance of correct rejection) if the true value of  $\theta$  is very different from the value that the null hypothesis specifies (i.e.,  $\theta = .5$ ). Notice that when  $\theta$  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.

---

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  $\theta$  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  $\theta$  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<sup>\*10</sup>, 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

---

<sup>\*10</sup>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  $\theta$  will go up<sup>\*11</sup> 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 ??, 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

---

<sup>\*11</sup>Notice that the true population parameter  $\theta$  doesn't necessarily correspond to an immutable fact of nature. In this context  $\theta$  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!

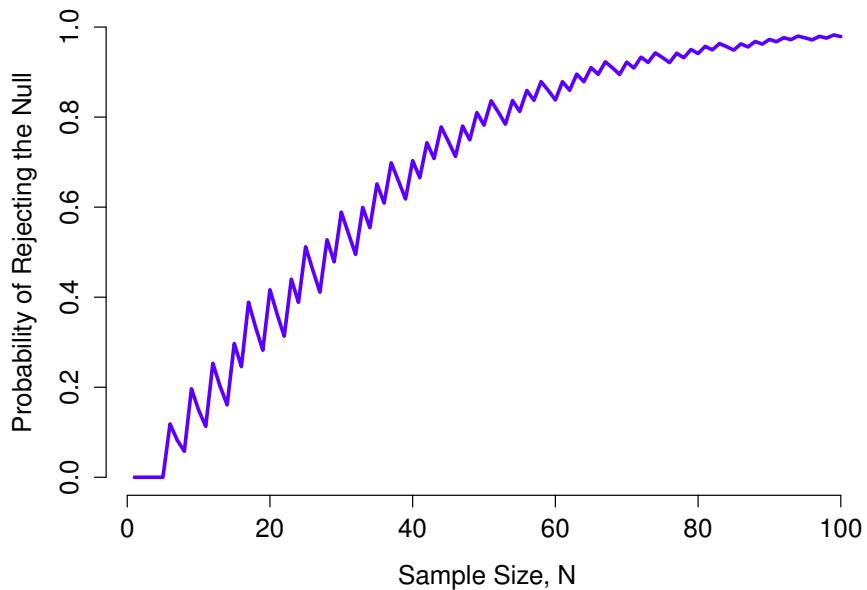


Figure 7.8 The power of our test plotted as a function of the sample size  $N$ . In this case, the true value of  $\theta$  is 0.7 but the null hypothesis is that  $\theta = 0.5$ . Overall, larger  $N$  means greater power. (The small zig-zags in this function occur because of some odd interactions between  $\theta$ ,  $\alpha$  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.<sup>\*12</sup> 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

---

<sup>\*12</sup>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<sup>\*13</sup> define the  $p$  value in terms of extreme data (Fisher), but we still have  $\alpha$  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 ??) 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.

---

<sup>\*13</sup>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, <sup>\*14</sup> 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 ??).
- Type 1 and Type 2 errors (Section ??)
- Test statistics and sampling distributions (Section ??)
- Hypothesis testing as a decision making process (Section ??)
- $p$ -values as “soft” decisions (Section ??)
- Writing up the results of a hypothesis test (Section ??)
- Running the hypothesis test in practice (Section ??)
- Effect size and power (Section ??)
- A few issues to consider regarding hypothesis testing (Section ??)

---

<sup>\*14</sup>In this case, the Pearson chi-square test of independence (Chapter ??)

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“ $\chi^2$  検定”（この章では、“カイ二乗 (にじょう)chi-square”と発音します<sup>\*1</sup>）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 $\chi^2$ (カイ二乗) 適合度検定

$\chi^2$  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

合性検定は適用できます。もう少し、心理学を交えて考えてみましょう。

### 8.1.1 カードデータ

何年にもわたる多くの研究が、人が完全にランダムにふるまおうとすることの難しさを示しています。ランダムに「行動」しようとしても、我々はパターンや構造に基づいて考えてしまします。そのため、「ランダムになにかをしてください」と言われたとしても人々が実際に行なうことはランダムなものにはなりません。結果として、人のランダム性（あるいは非ランダム性）に関する研究は、我々が世界をどのように捉えているのかについての深遠な心理学的問いを数多く投げかけます。このことを念頭に置いて、非常に簡単な研究について考えてみましょう。シャッフルされたカードのデッキを想像して、このデッキの中から「ランダムに」一枚のカードを頭の中で選ぶようにお願いしたとします。一枚目のカードを選んだ後、二枚目のカードを心の中で選択してもらいます。二つの選択に関して、注目するのは選ばれたカードのマーク（ハート、クラブ、スペード、ダイアモンド）です。これをたとえば  $N = 200$  にやってもらうよう依頼した後、選択されようとしたカードが本当にランダムに選ばれているかどうかをデータを確認して調べてみましょう。データは `randomness.csv` に入っています、JASPで開くと3つの変数が表示されるでしょう。変数 `id` は各参加者に対する一意識別子であり、二つの変数 `choice_1` と `choice_2` は参加者が選択したカードのマークを意味しています。

今回は、参加者の選んだ最初の選択肢に注目してみましょう。‘Descriptives’ - ‘Descriptive Statistics’ の下にある `Frequency tables` オプションを選択して、選択された各マークの数をカウントしてみましょう。以下が得られたものです:

	clubs	diamonds	hearts	spades
	35	51	64	50

That little frequency table is quite helpful. Looking at it, there's a bit of a hint that people *might* be more likely to select hearts than clubs, but it's not completely obvious just from looking at it whether that's really true, or if this is just due to chance. So we'll probably have to do some kind of statistical analysis to find out, which is what I'm going to talk about in the next section.

Excellent. From this point on, we'll treat this table as the data that we're looking to analyse. However, since I'm going to have to talk about this data in mathematical terms, it might be a good idea to be clear about what the notation is. In mathematical notation, we shorten the human-readable word “observed” to the letter  $O$ , and we use subscripts to denote the position of the observation. So the second observation in our table is written as  $O_2$  in maths. The relationship between the English descriptions and the mathematical symbols are illustrated below:

label	index, $i$	math. symbol	the value
clubs, ♣	1	$O_1$	35
diamonds, ♦	2	$O_2$	51
hearts, ♥	3	$O_3$	64
spades, ♠	4	$O_4$	50

Hopefully that's pretty clear. It's also worth noting that mathematicians prefer to talk about general rather than specific things, so you'll also see the notation  $O_i$ , which refers to the number of observations that fall within the  $i$ -th category (where  $i$  could be 1, 2, 3 or 4). Finally, if we want to refer to the set of all observed frequencies, statisticians group all observed values into a vector<sup>\*2</sup>, which I'll refer to using boldface type as  $\mathbf{O}$ .

$$\mathbf{O} = (O_1, O_2, O_3, O_4)$$

Again, this is nothing new or interesting. It's just notation. If I say that  $\mathbf{O} = (35, 51, 64, 50)$  all I'm doing is describing the table of observed frequencies (i.e., [observed](#)), but I'm referring to it using mathematical notation.

### 8.1.2 The null hypothesis and the alternative hypothesis

As the last section indicated, our research hypothesis is that "people don't choose cards randomly". What we're going to want to do now is translate this into some statistical hypotheses and then construct a statistical test of those hypotheses. The test that I'm going to describe to you is **Pearson's  $\chi^2$  (chi-square) goodness-of-fit test**, and as is so often the case we have to begin by carefully constructing our null hypothesis. In this case, it's pretty easy. First, let's state the null hypothesis in words:

$$H_0: \text{All four suits are chosen with equal probability}$$

Now, because this is statistics, we have to be able to say the same thing in a mathematical way. To do this, let's use the notation  $P_j$  to refer to the true probability that the  $j$ -th suit is chosen. If the null hypothesis is true, then each of the four suits has a 25% chance of being selected. In other words, our null hypothesis claims that  $P_1 = .25$ ,  $P_2 = .25$ ,  $P_3 = .25$  and finally that  $P_4 = .25$ . However, in the same way that we can group our observed frequencies into a vector  $\mathbf{O}$  that summarises the entire data set, we can use  $\mathbf{P}$  to refer to the probabilities that correspond to

---

<sup>\*2</sup>A vector is a sequence of data elements of the same basic type

our null hypothesis. So if I let the vector  $\mathbf{P} = (P_1, P_2, P_3, P_4)$  refer to the collection of probabilities that describe our null hypothesis, then we have:

$$H_0: \quad \mathbf{P} = (.25, .25, .25, .25)$$

In this particular instance, our null hypothesis corresponds to a vector of probabilities  $\mathbf{P}$  in which all of the probabilities are equal to one another. But this doesn't have to be the case. For instance, if the experimental task was for people to imagine they were drawing from a deck that had twice as many clubs as any other suit, then the null hypothesis would correspond to something like  $\mathbf{P} = (.4, .2, .2, .2)$ . As long as the probabilities are all positive numbers, and they all sum to 1, then it's a perfectly legitimate choice for the null hypothesis. However, the most common use of the goodness-of-fit test is to test a null hypothesis that all of the categories are equally likely, so we'll stick to that for our example.

What about our alternative hypothesis,  $H_1$ ? All we're really interested in is demonstrating that the probabilities involved aren't all identical (that is, people's choices weren't completely random). As a consequence, the "human friendly" versions of our hypotheses look like this:

$$H_0: \quad \text{All four suits are chosen with equal probability}$$

$$H_1: \quad \text{At least one of the suit-choice probabilities } \textit{isn't} 0.25$$

and the "mathematician friendly" version is:

$$H_0: \quad \mathbf{P} = (.25, .25, .25, .25)$$

$$H_1: \quad \mathbf{P} \neq (.25, .25, .25, .25)$$

### 8.1.3 The "goodness-of-fit" test statistic

At this point, we have our observed frequencies  $\mathbf{O}$  and a collection of probabilities  $\mathbf{P}$  corresponding to the null hypothesis that we want to test. What we now want to do is construct a test of the null hypothesis. As always, if we want to test  $H_0$  against  $H_1$ , we're going to need a test statistic. The basic trick that a goodness-of-fit test uses is to construct a test statistic that measures how "close" the data are to the null hypothesis. If the data don't resemble what you'd "expect" to see if the null hypothesis were true, then it probably isn't true. Okay, if the null hypothesis were true, what would we expect to see? Or, to use the correct terminology, what are the **expected frequencies**. There are  $N = 200$  observations, and (if the null is true) the probability of any one of them choosing a heart is  $P_3 = .25$ , so I guess we're expecting  $200 \times .25 = 50$  hearts, right? Or, more specifically, if we let  $E_i$  refer to "the number of category  $i$  responses that we're expecting if the null is true", then

$$E_i = N \times P_i$$

This is pretty easy to calculate. If there are 200 observations that can fall into four categories, and we think that all four categories are equally likely, then on average we'd expect to see 50 observations in each category, right?

Now, how do we translate this into a test statistic? Clearly, what we want to do is compare the *expected* number of observations in each category ( $E_i$ ) with the *observed* number of observations in that category ( $O_i$ ). And on the basis of this comparison we ought to be able to come up with a good test statistic. To start with, let's calculate the difference between what the null hypothesis expected us to find and what we actually did find. That is, we calculate the "observed minus expected" difference score,  $O_i - E_i$ . This is illustrated in the following table.

		♣	♦	♥	♠
expected frequency	$E_i$	50	50	50	50
observed frequency	$O_i$	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0

So, based on our calculations, it's clear that people chose more hearts and fewer clubs than the null hypothesis predicted. However, a moment's thought suggests that these raw differences aren't quite what we're looking for. Intuitively, it feels like it's just as bad when the null hypothesis predicts too few observations (which is what happened with hearts) as it is when it predicts too many (which is what happened with clubs). So it's a bit weird that we have a negative number for clubs and a positive number for hearts. One easy way to fix this is to square everything, so that we now calculate the squared differences,  $(O_i - E_i)^2$ . As before, we can do this by hand:

		♣	♦	♥	♠
expected frequency	$E_i$	50	50	50	50
observed frequency	$O_i$	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0
squared difference	$(O_i - E_i)^2$	225	1	196	0

Now we're making progress. What we've got now is a collection of numbers that are big whenever the null hypothesis makes a bad prediction (clubs and hearts), but are small whenever it makes a good one (diamonds and spades). Next, for some technical reasons that I'll explain in a moment, let's also divide all these numbers by the expected frequency  $E_i$ , so we're actually calculating the *scaled* squared difference,  $\frac{(E_i - O_i)^2}{E_i}$ . Since  $E_i = 50$  for all categories in our example, it's not a very interesting calculation, but let's do it anyway:

		♣	♦	♥	♠
expected frequency	$E_i$	50	50	50	50
observed frequency	$O_i$	35	51	64	50
difference score	$O_i - E_i$	-15	1	14	0
squared difference	$(O_i - E_i)^2$	225	1	196	0
scaled sq. diff.	$(O_i - E_i)^2/E_i$	4.50	0.02	3.92	0.00

In effect, what we've got here are four different "error" scores, each one telling us how big a "mistake" the null hypothesis made when we tried to use it to predict our observed frequencies. So, in order to convert this into a useful test statistic, one thing we could do is just add these numbers up. The result is called the **goodness-of-fit** statistic, conventionally referred to either as  $\chi^2$  (chi-square) or GOF. We can calculate it as  $4.50 + 0.02 + 3.92 + 0.00 = 8.44$ .

If we let  $k$  refer to the total number of categories (i.e.,  $k = 4$  for our cards data), then the  $\chi^2$  statistic is given by:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Intuitively, it's clear that if  $\chi^2$  is small, then the observed data  $O_i$  are very close to what the null hypothesis predicted  $E_i$ , so we're going to need a large  $\chi^2$  statistic in order to reject the null.

As we've seen from our calculations, in our cards data set we've got a value of  $\chi^2 = 8.44$ . So now the question becomes is this a big enough value to reject the null?

#### 8.1.4 The sampling distribution of the GOF statistic

To determine whether or not a particular value of  $\chi^2$  is large enough to justify rejecting the null hypothesis, we're going to need to figure out what the sampling distribution for  $\chi^2$  would be if the null hypothesis were true. So that's what I'm going to do in this section. I'll show you in a fair amount of detail how this sampling distribution is constructed, and then, in the next section, use it to build up a hypothesis test. If you want to cut to the chase and are willing to take it on faith that the sampling distribution is a  **$\chi^2$  (chi-square) distribution** with  $k - 1$  degrees of freedom, you can skip the rest of this section. However, if you want to understand *why* the goodness-of-fit test works the way it does, read on.

Okay, let's suppose that the null hypothesis is actually true. If so, then the true probability that an observation falls in the  $i$ -th category is  $P_i$ . After all, that's pretty much the definition

of our null hypothesis. Let's think about what this actually means. This is kind of like saying that "nature" makes the decision about whether or not the observation ends up in category  $i$  by flipping a weighted coin (i.e., one where the probability of getting a head is  $P_j$ ). And therefore we can think of our observed frequency  $O_i$  by imagining that nature flipped  $N$  of these coins (one for each observation in the data set), and exactly  $O_i$  of them came up heads. Obviously, this is a pretty weird way to think about the experiment. But what it does (I hope) is remind you that we've actually seen this scenario before. It's exactly the same set up that gave rise to the binomial distribution in Section ???. In other words, if the null hypothesis is true, then it follows that our observed frequencies were generated by sampling from a binomial distribution:

$$O_i \sim \text{Binomial}(P_i, N)$$

Now, if you remember from our discussion of the central limit theorem (Section ??) the binomial distribution starts to look pretty much identical to the normal distribution, especially when  $N$  is large and when  $P_i$  isn't *too* close to 0 or 1. In other words as long as  $N \times P_i$  is large enough. Or, to put it another way, when the expected frequency  $E_i$  is large enough then the theoretical distribution of  $O_i$  is approximately normal. Better yet, if  $O_i$  is normally distributed, then so is  $(O_i - E_i)/\sqrt{E_i}$ . Since  $E_i$  is a fixed value, subtracting off  $E_i$  and dividing by  $\sqrt{E_i}$  changes the mean and standard deviation of the normal distribution but that's all it does. Okay, so now let's have a look at what our goodness-of-fit statistic actually *is*. What we're doing is taking a bunch of things that are normally-distributed, squaring them, and adding them up. Wait. We've seen that before too! As we discussed in Section ??, when you take a bunch of things that have a standard normal distribution (i.e., mean 0 and standard deviation 1), square them and then add them up, the resulting quantity has a chi-square distribution. So now we know that the null hypothesis predicts that the sampling distribution of the goodness-of-fit statistic is a chi-square distribution. Cool.

There's one last detail to talk about, namely the degrees of freedom. If you remember back to Section ??, I said that if the number of things you're adding up is  $k$ , then the degrees of freedom for the resulting chi-square distribution is  $k$ . Yet, what I said at the start of this section is that the actual degrees of freedom for the chi-square goodness-of-fit test is  $k - 1$ . What's up with that? The answer here is that what we're supposed to be looking at is the number of genuinely *independent* things that are getting added together. And, as I'll go on to talk about in the next section, even though there are  $k$  things that we're adding only  $k - 1$  of them are truly independent,

and so the degrees of freedom is actually only  $k - 1$ . That's the topic of the next section.<sup>\*3</sup>

### 8.1.5 Degrees of freedom

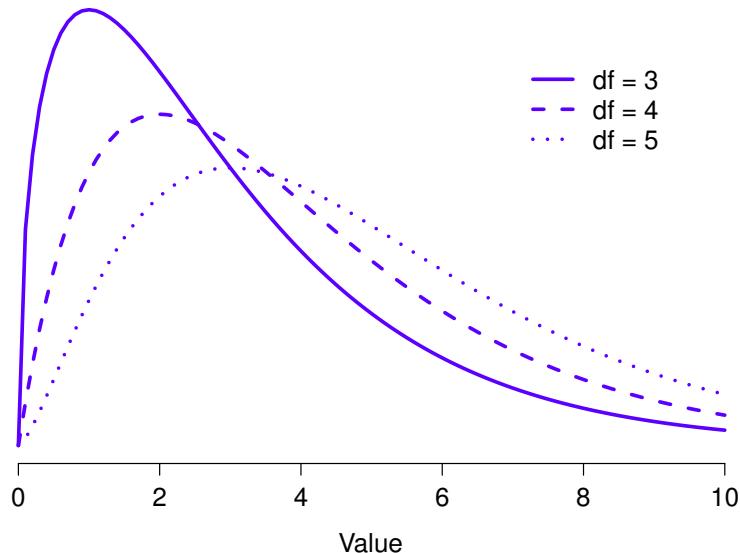


Figure 8.1  $\chi^2$  (chi-square) distributions with different values for the “degrees of freedom”.

When I introduced the chi-square distribution in Section ??, I was a bit vague about what “**degrees of freedom**” actually *means*. Obviously, it matters. Looking at Figure ??, you can see that if we change the degrees of freedom then the chi-square distribution changes shape quite substantially. But what exactly *is* it? Again, when I introduced the distribution and explained its relationship to the normal distribution, I did offer an answer: it’s the number of “normally distributed variables” that I’m squaring and adding together. But, for most people, that’s kind of abstract and not entirely helpful. What we really need to do is try to understand degrees of freedom in terms of our data. So here goes.

The basic idea behind degrees of freedom is quite simple. You calculate it by counting up the

---

<sup>\*3</sup>If you rewrite the equation for the goodness-of-fit statistic as a sum over  $k - 1$  independent things you get the “proper” sampling distribution, which is chi-square with  $k - 1$  degrees of freedom. It’s beyond the scope of an introductory book to show the maths in that much detail. All I wanted to do is give you a sense of why the goodness-of-fit statistic is associated with the chi-square distribution.

number of distinct “quantities” that are used to describe your data and then subtracting off all of the “constraints” that those data must satisfy.<sup>\*4</sup> This is a bit vague, so let’s use our cards data as a concrete example. We describe our data using four numbers,  $O_1$ ,  $O_2$ ,  $O_3$  and  $O_4$  corresponding to the observed frequencies of the four different categories (hearts, clubs, diamonds, spades). These four numbers are the *random outcomes* of our experiment. But my experiment actually has a fixed constraint built into it: the sample size  $N$ .<sup>\*5</sup> That is, if we know how many people chose hearts, how many chose diamonds and how many chose clubs, then we’d be able to figure out exactly how many chose spades. In other words, although our data are described using four numbers, they only actually correspond to  $4 - 1 = 3$  degrees of freedom. A slightly different way of thinking about it is to notice that there are four *probabilities* that we’re interested in (again, corresponding to the four different categories), but these probabilities must sum to one, which imposes a constraint. Therefore the degrees of freedom is  $4 - 1 = 3$ . Regardless of whether you want to think about it in terms of the observed frequencies or in terms of the probabilities, the answer is the same. In general, when running the  $\chi^2$  (chi-square) goodness-of-fit test for an experiment involving  $k$  groups, then the degrees of freedom will be  $k - 1$ .

#### 8.1.6 Testing the null hypothesis

The final step in the process of constructing our hypothesis test is to figure out what the rejection region is. That is, what values of  $\chi^2$  would lead us to reject the null hypothesis. As we saw earlier, large values of  $\chi^2$  imply that the null hypothesis has done a poor job of predicting the data from our experiment, whereas small values of  $\chi^2$  imply that it’s actually done pretty well. Therefore, a pretty sensible strategy would be to say there is some critical value such that if  $\chi^2$  is bigger than the critical value we reject the null, but if  $\chi^2$  is smaller than this value we retain the null. In other words, to use the language we introduced in Chapter ?? the chi-square goodness-of-fit test is always a **one-sided test**. Right, so all we have to do is figure out what this critical value is. And it’s pretty straightforward. If we want our test to have significance level of  $\alpha = .05$  (that is,

---

<sup>\*4</sup>I feel obliged to point out that this is an over-simplification. It works nicely for quite a few situations, but every now and then we’ll come across degrees of freedom values that aren’t whole numbers. Don’t let this worry you too much; when you come across this just remind yourself that “degrees of freedom” is actually a bit of a messy concept, and that the nice simple story that I’m telling you here isn’t the whole story. For an introductory class it’s usually best to stick to the simple story, but I figure it’s best to warn you to expect this simple story to fall apart. If I didn’t give you this warning you might start getting confused when you see  $df = 3.4$  or something, (incorrectly) thinking that you had misunderstood something that I’ve taught you rather than (correctly) realising that there’s something that I haven’t told you.

<sup>\*5</sup>In practice, the sample size isn’t always fixed. For example, we might run the experiment over a fixed period of time and the number of people participating depends on how many people show up. That doesn’t matter for the current purposes.

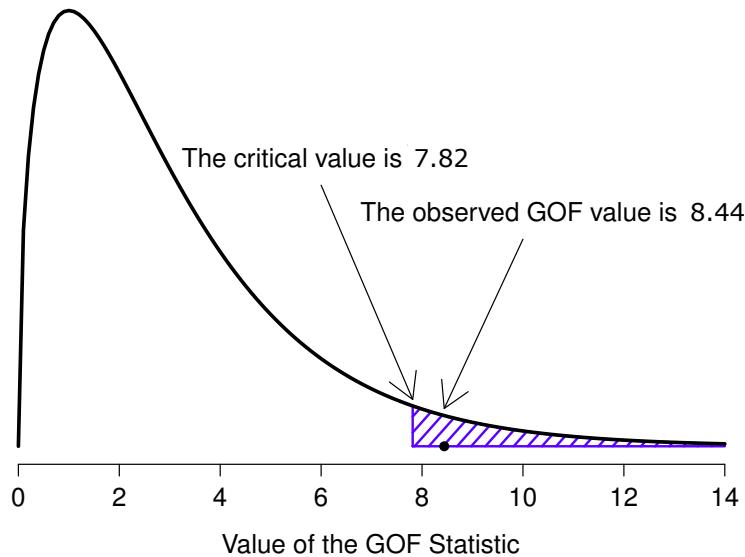


Figure8.2 Illustration of how the hypothesis testing works for the  $\chi^2$  (chi-square) goodness-of-fit test.

.....

we are willing to tolerate a Type I error rate of 5%), then we have to choose our critical value so that there is only a 5% chance that  $\chi^2$  could get to be that big if the null hypothesis is true. This is illustrated in Figure ??.

Ah but, I hear you ask, how do I find the critical value of a chi-square distribution with  $k - 1$  degrees of freedom? Many many years ago when I first took a psychology statistics class we used to look up these critical values in a book of critical value tables, like the one in Figure ?? . Looking at this Figure, we can see that the critical value for a  $\chi^2$  distribution with 3 degrees of freedom, and  $p=0.05$  is 7.815.

So, if our calculated  $\chi^2$  statistic is bigger than the critical value of 7.815, then we can reject the null hypothesis (remember that the null hypothesis,  $H_0$ , is that all four suits are chosen with equal probability). Since we actually already calculated that before (i.e.,  $\chi^2 = 8.44$ ) we can reject the null hypothesis. And that's it, basically. You now know "Pearson's  $\chi^2$  test for the goodness-of-fit".

<b>Degrees of Freedom</b>	<b>Probability</b>								
	<b>0.95</b>	<b>0.90</b>	<b>0.70</b>	<b>0.50</b>	<b>0.30</b>	<b>0.10</b>	<b>0.05</b>	<b>0.01</b>	<b>0.001</b>
<b>1</b>	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
<b>2</b>	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
<b>3</b>	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
<b>4</b>	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
<b>5</b>	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
<b>6</b>	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
<b>7</b>	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
<b>8</b>	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124
<b>9</b>	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
<b>10</b>	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	<b>Non-significant</b>						<b>Significant</b>		

Figure8.3 Table of critical values for the chi-square distribution

.....

Lucky you.

#### 8.1.7 Doing the test in JASP

Not surprisingly, JASP provides an analysis that will do these calculations for you. From the main ‘Analyses’ toolbar select ‘Frequencies’ – ‘Multinomial Test’. Then in the analysis window that appears move the variable you want to analyse (`choice_1` across into the ‘Factor’ box. Also, click on the ‘Descriptives’ check box so that you see the expected counts in the results table. When you have done all this, you should see the analysis results in JASP as in Figure ???. No surprise then that JASP provides the same expected counts and statistics that we calculated by hand above, with a  $\chi^2$  value of 8.44 with 3 d.f. and  $p=0.038$ . Note that we don’t need to look up a critical  $p$ -value threshold value any more, as JASP gives us the actual  $p$ -value of the calculated  $\chi^2$  for 3 d.f.

#### 8.1.8 Specifying a different null hypothesis

At this point you might be wondering what to do if you want to run a goodness-of-fit test but your null hypothesis is *not* that all categories are equally likely. For instance, let’s suppose that someone had made the theoretical prediction that people should choose red cards 60% of the

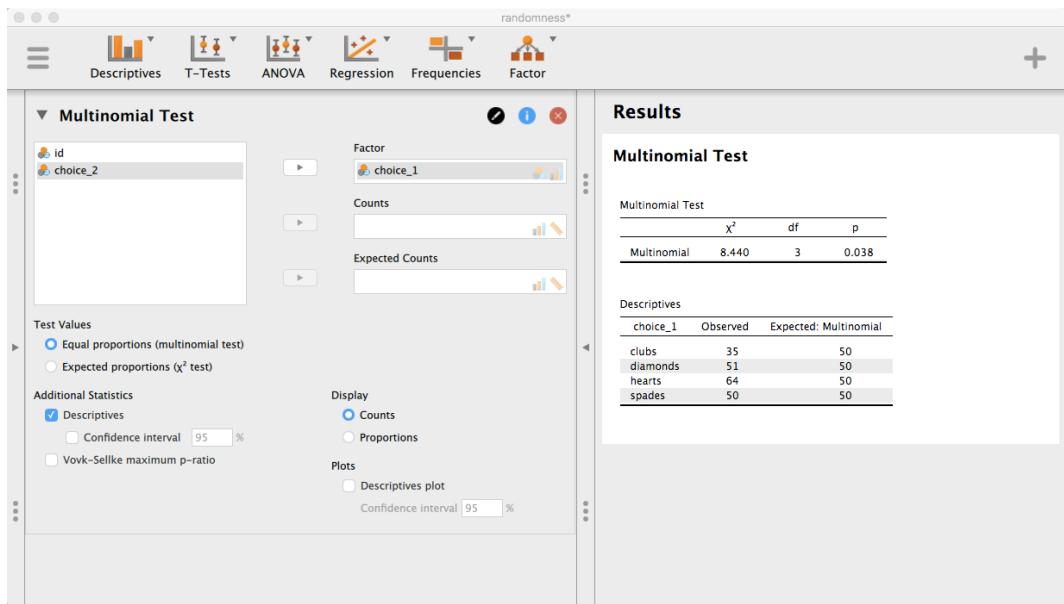


Figure 8.4 A  $\chi^2$  goodness-of-fit test in JASP, with table showing both observed and expected frequencies.

time, and black cards 40% of the time (I've no idea why you'd predict that), but had no other preferences. If that were the case, the null hypothesis would be to expect 30% of the choices to be hearts, 30% to be diamonds, 20% to be spades and 20% to be clubs. In other words we would expect hearts and diamonds to each appear 60 times (30% of 200 is 60), and spades and clubs to each appear 40 times (20% of 200 is 40). This seems like a silly theory to me, but nonetheless, it's pretty easy to test this explicitly specified null hypothesis with the data in our JASP analysis. In the analysis window (see Figure ??) you can click the radio button for 'Expected Proportions ( $\chi^2$  test)'. When you do this, there are options for entering different expected counts for the variable you have selected, in our case this is `choice_1`. Change the counts to reflect the new null hypothesis, as in Figure ??, and see how the results change.

The expected counts are now:

	$\clubsuit$	$\diamondsuit$	$\heartsuit$	$\spadesuit$
expected frequency $E_i$	40	60	60	40

and the  $\chi^2$  statistic is 4.742, 3 d.f.,  $p = 0.192$ . Now, the results of our updated hypotheses and the expected frequencies are different from what they were last time. As a consequence our  $\chi^2$  test statistic is different, and our  $p$ -value is different too. Annoyingly, the  $p$ -value is .192, so we can't reject the null hypothesis (look back at section ?? to remind yourself why). Sadly, despite

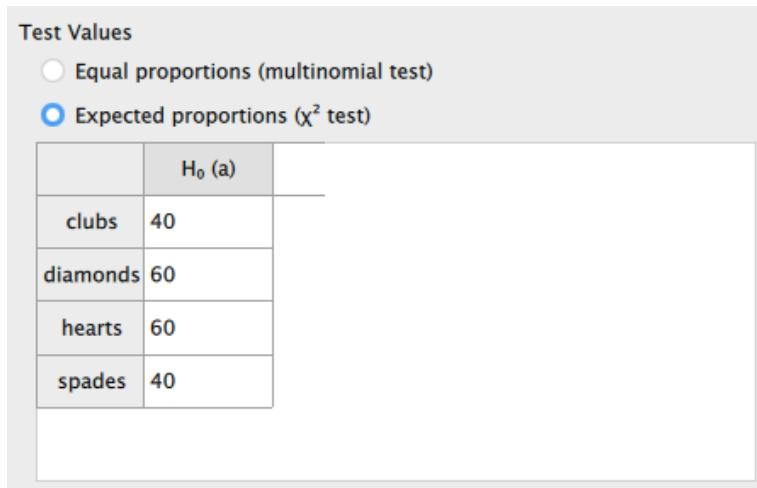


Figure8.5 Changing the expected proportions in the  $\chi^2$  goodness-of-fit test in JASP

.....

the fact that the null hypothesis corresponds to a very silly theory, these data don't provide enough evidence against it.

#### 8.1.9 How to report the results of the test

So now you know how the test works, and you know how to do the test using a wonderful JASP-flavoured magic computing box. The next thing you need to know is how to write up the results. After all, there's no point in designing and running an experiment and then analysing the data if you don't tell anyone about it! So let's now talk about what you need to do when reporting your analysis. Let's stick with our card-suits example. If I wanted to write this result up for a paper or something, then the conventional way to report this would be to write something like this:

Of the 200 participants in the experiment, 64 selected hearts for their first choice, 51 selected diamonds, 50 selected spades, and 35 selected clubs. A chi-square goodness-of-fit test was conducted to test whether the choice probabilities were identical for all four suits. The results were significant ( $\chi^2(3) = 8.44, p < .05$ ), suggesting that people did not select suits purely at random.

This is pretty straightforward and hopefully it seems pretty unremarkable. That said, there's a few things that you should note about this description:

- *The statistical test is preceded by the descriptive statistics.* That is, I told the reader

something about what the data look like before going on to do the test. In general, this is good practice. Always remember that your reader doesn't know your data anywhere near as well as you do. So, unless you describe it to them properly, the statistical tests won't make any sense to them and they'll get frustrated and cry.

- *The description tells you what the null hypothesis being tested is.* To be honest, writers don't always do this but it's often a good idea in those situations where some ambiguity exists, or when you can't rely on your readership being intimately familiar with the statistical tools that you're using. Quite often the reader might not know (or remember) all the details of the test that you're using, so it's a kind of politeness to "remind" them! As far as the goodness-of-fit test goes, you can usually rely on a scientific audience knowing how it works (since it's covered in most intro stats classes). However, it's still a good idea to be explicit about stating the null hypothesis (briefly!) because the null hypothesis can be different depending on what you're using the test for. For instance, in the cards example my null hypothesis was that all the four suit probabilities were identical (i.e.,  $P_1 = P_2 = P_3 = P_4 = 0.25$ ), but there's nothing special about that hypothesis. I could just as easily have tested the null hypothesis that  $P_1 = 0.7$  and  $P_2 = P_3 = P_4 = 0.1$  using a goodness-of-fit test. So it's helpful to the reader if you explain to them what your null hypothesis was. Also, notice that I described the null hypothesis in words, not in maths. That's perfectly acceptable. You can describe it in maths if you like, but since most readers find words easier to read than symbols, most writers tend to describe the null using words if they can.
- A "stat block" is included. When reporting the results of the test itself, I didn't just say that the result was significant, I included a "stat block" (i.e., the dense mathematical-looking part in the parentheses) which reports all the "key" statistical information. For the chi-square goodness-of-fit test, the information that gets reported is the test statistic (that the goodness-of-fit statistic was 8.44), the information about the distribution used in the test ( $\chi^2$  with 3 degrees of freedom which is usually shortened to  $\chi^2(3)$ ), and then the information about whether the result was significant (in this case  $p < .05$ ). The particular information that needs to go into the stat block is different for every test, and so each time I introduce a new test I'll show you what the stat block should look like.<sup>\*6</sup> However the general principle is that you should always provide enough information so that the reader could check the test results themselves if they really wanted to.
- *The results are interpreted.* In addition to indicating that the result was significant, I

---

<sup>\*6</sup>Well, sort of. The conventions for how statistics should be reported tend to differ somewhat from discipline to discipline. I've tended to stick with how things are done in psychology, since that's what I do. But the general principle of providing enough information to the reader to allow them to check your results is pretty universal, I think.

provided an interpretation of the result (i.e., that people didn't choose randomly). This is also a kindness to the reader, because it tells them something about what they should believe about what's going on in your data. If you don't include something like this, it's really hard for your reader to understand what's going on.<sup>\*7</sup>

As with everything else, your overriding concern should be that you *explain* things to your reader. Always remember that the point of reporting your results is to communicate to another human being. I cannot tell you just how many times I've seen the results section of a report or a thesis or even a scientific article that is just gibberish, because the writer has focused solely on making sure they've included all the numbers and forgotten to actually communicate with the human reader.

#### 8.1.10 A comment on statistical notation

*Satan delights equally in statistics and in quoting scripture*

– H.G. Wells

If you've been reading very closely, and are as much of a mathematical pedant as I am, there is one thing about the way I wrote up the chi-square test in the last section that might be bugging you a little bit. There's something that feels a bit wrong with writing " $\chi^2(3) = 8.44$ ", you might be thinking. After all, it's the goodness-of-fit statistic that is equal to 8.44, so shouldn't I have written  $X^2 = 8.44$  or maybe  $GOF = 8.44$ ? This seems to be conflating the *sampling distribution* (i.e.,  $\chi^2$  with  $df = 3$ ) with the *test statistic* (i.e.,  $X^2$ ). Odds are you figured it was a typo, since  $\chi$  and  $X$  look pretty similar. Oddly, it's not. Writing  $\chi^2(3) = 8.44$  is essentially a highly condensed way of writing "the sampling distribution of the test statistic is  $\chi^2(3)$ , and the value of the test statistic is 8.44".

In one sense, this is kind of silly. There are *lots* of different test statistics out there that turn out to have a chi-square sampling distribution. The  $X^2$  statistic that we've used for our goodness-of-fit test is only one of many (albeit one of the most commonly encountered ones). In a sensible, perfectly organised world we'd *always* have a separate name for the test statistic and

---

<sup>\*7</sup>To some people, this advice might sound odd, or at least in conflict with the “usual” advice on how to write a technical report. Very typically, students are told that the “results” section of a report is for describing the data and reporting statistical analysis, and the “discussion” section is for providing interpretation. That’s true as far as it goes, but I think people often interpret it way too literally. The way I usually approach it is to provide a quick and simple interpretation of the data in the results section, so that my reader understands what the data are telling us. Then, in the discussion, I try to tell a bigger story about how my results fit with the rest of the scientific literature. In short, don’t let the “interpretation goes in the discussion” advice turn your results section into incomprehensible garbage. Being understood by your reader is *much* more important.

the sampling distribution. That way, the stat block itself would tell you exactly what it was that the researcher had calculated. Sometimes this happens. For instance, the test statistic used in the Pearson goodness-of-fit test is written  $X^2$ , but there's a closely related test known as the *G-test*<sup>\*8</sup> (**Sokal1994**), in which the test statistic is written as  $G$ . As it happens, the Pearson goodness-of-fit test and the *G-test* both test the same null hypothesis, and the sampling distribution is exactly the same (i.e., chi-square with  $k - 1$  degrees of freedom). If I'd done a *G-test* for the cards data rather than a goodness-of-fit test, then I'd have ended up with a test statistic of  $G = 8.65$ , which is slightly different from the  $X^2 = 8.44$  value that I got earlier and which produces a slightly smaller *p*-value of  $p = .034$ . Suppose that the convention was to report the test statistic, then the sampling distribution, and then the *p*-value. If that were true, then these two situations would produce different stat blocks: my original result would be written  $X^2 = 8.44, \chi^2(3), p = .038$ , whereas the new version using the *G-test* would be written as  $G = 8.65, \chi^2(3), p = .034$ . However, using the condensed reporting standard, the original result is written  $\chi^2(3) = 8.44, p = .038$ , and the new one is written  $\chi^2(3) = 8.65, p = .034$ , and so it's actually unclear which test I actually ran.

So why don't we live in a world in which the contents of the stat block uniquely specifies what tests were ran? The deep reason is that life is messy. We (as users of statistical tools) want it to be nice and neat and organised. We want it to be *designed*, as if it were a product, but that's not how life works. Statistics is an intellectual discipline just as much as any other one, and as such it's a massively distributed, partly-collaborative and partly-competitive project that no-one really understands completely. The things that you and I use as data analysis tools weren't created by an Act of the Gods of Statistics. They were invented by lots of different people, published as papers in academic journals, implemented, corrected and modified by lots of other people, and then explained to students in textbooks by someone else. As a consequence, there's a *lot* of test statistics that don't even have names, and as a consequence they're just given the same name as the corresponding sampling distribution. As we'll see later, any test statistic that follows a  $\chi^2$  distribution is commonly called a "chi-square statistic", anything that follows a *t*-distribution is called a "*t*-statistic", and so on. But, as the  $\chi^2$  versus  $G$  example illustrates, two different things with the same sampling distribution are still, well, different.

As a consequence, it's sometimes a good idea to be clear about what the actual test was that you ran, especially if you're doing something unusual. If you just say "chi-square test" it's not actually clear what test you're talking about. Although, since the two most common chi-square tests are the goodness-of-fit test and the independence test (Section ??), most readers with stats training can probably guess. Nevertheless, it's something to be aware of.

---

<sup>\*8</sup>Complicating matters, the *G-test* is a special case of a whole class of tests that are known as *likelihood ratio tests*. I don't cover LRTs in this book, but they are quite handy things to know about.

## The $\chi^2$ test of independence (or association)

*GUARDBOT 1:* *Halt!*

*GUARDBOT 2:* *Be you robot or human?*

*LEELA:* *Robot...we be.*

*FRY:* *Uh, yup! Just two robots out roboting it up! Eh?*

*GUARDBOT 1:* *Administer the test.*

*GUARDBOT 2:* *Which of the following would you most prefer?*

*A: A puppy, B: A pretty flower from your sweetie,  
or C: A large properly-formatted data file?*

*GUARDBOT 1:* *Choose!*

– Futurama, “Fear of a Bot Planet”

The other day I was watching an animated documentary examining the quaint customs of the natives of the planet *Chapek 9*. Apparently, in order to gain access to their capital city a visitor must prove that they’re a robot, not a human. In order to determine whether or not a visitor is human, the natives ask whether the visitor prefers puppies, flowers, or large, properly formatted data files. “Pretty clever,” I thought to myself “but what if humans and robots have the same preferences? That probably wouldn’t be a very good test then, would it?” As it happens, I got my hands on the testing data that the civil authorities of *Chapek 9* used to check this. It turns out that what they did was very simple. They found a bunch of robots and a bunch of humans and asked them what they preferred. I saved their data in a file called `chapek9.csv`, which we can now load into JASP. As well as the `ID` variable that identifies individual people, there are two nominal text variables, `species` and `choice`. In total there are 180 entries in the data set, one for each person (counting both robots and humans as “people”) who was asked to make a choice. Specifically, there are 93 humans and 87 robots, and overwhelmingly the preferred choice is the data file. You can check this yourself by asking JASP for Frequency Tables, under the ‘Descriptives’ - ‘Descriptive Statistics’ button. However, this summary does not address the question we’re interested in. To do that, we need a more detailed description of the data. What we want to do is look at the `choices` broken down by `species`. That is, we need to cross-tabulate the data. In JASP we do this using the ‘Frequencies’ - ‘Contingency Tables’ button, moving `species` into the ‘Columns’ box and `choice` into the ‘Rows’ box. This procedure should produce a table similar to this:

	Robot	Human	Total
Puppy	13	15	28
Flower	30	13	43
Data	44	65	109
Total	87	93	180

From this, it's quite clear that the vast majority of the humans chose the data file, whereas the robots tended to be a lot more even in their preferences. Leaving aside the question of *why* the humans might be more likely to choose the data file for the moment (which does seem quite odd, admittedly), our first order of business is to determine if the discrepancy between human choices and robot choices in the data set is statistically significant.

### 8.2.1 Constructing our hypothesis test

How do we analyse this data? Specifically, since my *research* hypothesis is that "humans and robots answer the question in different ways", how can I construct a test of the *null* hypothesis that "humans and robots answer the question the same way"? As before, we begin by establishing some notation to describe the data:

	Robot	Human	Total
Puppy	$O_{11}$	$O_{12}$	$R_1$
Flower	$O_{21}$	$O_{22}$	$R_2$
Data	$O_{31}$	$O_{32}$	$R_3$
Total	$C_1$	$C_2$	$N$

In this notation we say that  $O_{ij}$  is a count (observed frequency) of the number of respondents that are of species  $j$  (robots or human) who gave answer  $i$  (puppy, flower or data) when asked to make a choice. The total number of observations is written  $N$ , as usual. Finally, I've used  $R_i$  to denote the row totals (e.g.,  $R_1$  is the total number of people who chose the flower), and  $C_j$  to denote the column totals (e.g.,  $C_1$  is the total number of robots).<sup>\*9</sup>

So now let's think about what the null hypothesis says. If robots and humans are responding

---

<sup>\*9</sup>A technical note. The way I've described the test pretends that the column totals are fixed (i.e., the researcher intended to survey 87 robots and 93 humans) and the row totals are random (i.e., it just turned out that 28 people chose the puppy). To use the terminology from my mathematical statistics textbook (**Hogg2005**), I should technically refer to this situation as a chi-square test of homogeneity and reserve the term chi-square test of independence for the situation where both the row and column totals are random outcomes of the experiment. In the initial drafts of this book that's exactly what I did. However, it turns out that these two tests are identical, and so I've collapsed them together.

in the same way to the question, it means that the probability that “a robot says puppy” is the same as the probability that “a human says puppy”, and so on for the other two possibilities. So, if we use  $P_{ij}$  to denote “the probability that a member of species  $j$  gives response  $i$ ” then our null hypothesis is that:

$H_0$ : All of the following are true:

$$\begin{aligned} P_{11} &= P_{12} \text{ (same probability of saying “puppy”),} \\ P_{21} &= P_{22} \text{ (same probability of saying “flower”), and} \\ P_{31} &= P_{32} \text{ (same probability of saying “data”).} \end{aligned}$$

And actually, since the null hypothesis is claiming that the true choice probabilities don’t depend on the species of the person making the choice, we can let  $P_i$  refer to this probability, e.g.,  $P_1$  is the true probability of choosing the puppy.

Next, in much the same way that we did with the goodness-of-fit test, what we need to do is calculate the expected frequencies. That is, for each of the observed counts  $O_{ij}$ , we need to figure out what the null hypothesis would tell us to expect. Let’s denote this expected frequency by  $E_{ij}$ . This time, it’s a little bit trickier. If there are a total of  $C_j$  people that belong to species  $j$ , and the true probability of anyone (regardless of species) choosing option  $i$  is  $P_i$ , then the expected frequency is just:

$$E_{ij} = C_j \times P_i$$

Now, this is all very well and good, but we have a problem. Unlike the situation we had with the goodness-of-fit test, the null hypothesis doesn’t actually specify a particular value for  $P_i$ . It’s something we have to estimate (Chapter ??) from the data! Fortunately, this is pretty easy to do. If 28 out of 180 people selected the flowers, then a natural estimate for the probability of choosing flowers is 28/180, which is approximately .16. If we phrase this in mathematical terms, what we’re saying is that our estimate for the probability of choosing option  $i$  is just the row total divided by the total sample size:

$$\hat{P}_i = \frac{R_i}{N}$$

Therefore, our expected frequency can be written as the product (i.e. multiplication) of the row total and the column total, divided by the total number of observations:<sup>\*10</sup>

$$E_{ij} = \frac{R_i \times C_j}{N}$$

---

<sup>\*10</sup>Technically,  $E_{ij}$  here is an estimate, so I should probably write it  $\hat{E}_{ij}$ . But since no-one else does, I won’t either.

Now that we've figured out how to calculate the expected frequencies, it's straightforward to define a test statistic, following the exact same strategy that we used in the goodness-of-fit test. In fact, it's pretty much the *same* statistic.

For a contingency table with  $r$  rows and  $c$  columns, the equation that defines our  $X^2$  statistic is

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

The only difference is that I have to include two summation signs (i.e.,  $\sum$ ) to indicate that we're summing over both rows and columns.

As before, large values of  $X^2$  indicate that the null hypothesis provides a poor description of the data, whereas small values of  $X^2$  suggest that it does a good job of accounting for the data. Therefore, just like last time, we want to reject the null hypothesis if  $X^2$  is too large.

Not surprisingly, this statistic is  $\chi^2$  distributed. All we need to do is figure out how many degrees of freedom are involved, which actually isn't too hard. As I mentioned before, you can (usually) think of the degrees of freedom as being equal to the number of data points that you're analysing, minus the number of constraints. A contingency table with  $r$  rows and  $c$  columns contains a total of  $r \times c$  observed frequencies, so that's the total number of observations. What about the constraints? Here, it's slightly trickier. The answer is always the same

$$df = (r - 1)(c - 1)$$

but the explanation for *why* the degrees of freedom takes this value is different depending on the experimental design. For the sake of argument, let's suppose that we had honestly intended to survey exactly 87 robots and 93 humans (column totals fixed by the experimenter), but left the row totals free to vary (row totals are random variables). Let's think about the constraints that apply here. Well, since we deliberately fixed the column totals by Act of Experimenter, we have  $c$  constraints right there. But, there's actually more to it than that. Remember how our null hypothesis had some free parameters (i.e., we had to estimate the  $P_i$  values)? Those matter too. I won't explain why in this book, but every free parameter in the null hypothesis is rather like an additional constraint. So, how many of those are there? Well, since these probabilities have to sum to 1, there's only  $r - 1$  of these. So our total degrees of freedom is:

$$\begin{aligned} df &= (\text{number of observations}) - (\text{number of constraints}) \\ &= (rc) - (c + (r - 1)) \\ &= rc - c - r + 1 \\ &= (r - 1)(c - 1) \end{aligned}$$

Alternatively, suppose that the only thing that the experimenter fixed was the total sample size  $N$ .

That is, we quizzed the first 180 people that we saw and it just turned out that 87 were robots and 93 were humans. This time around our reasoning would be slightly different, but would still lead us to the same answer. Our null hypothesis still has  $r - 1$  free parameters corresponding to the choice probabilities, but it now *also* has  $c - 1$  free parameters corresponding to the species probabilities, because we'd also have to estimate the probability that a randomly sampled person turns out to be a robot.<sup>\*11</sup> Finally, since we did actually fix the total number of observations  $N$ , that's one more constraint. So, now we have  $rc$  observations, and  $(c-1) + (r-1) + 1$  constraints. What does that give?

$$\begin{aligned} df &= (\text{number of observations}) - (\text{number of constraints}) \\ &= rc - ((c-1) + (r-1) + 1) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

Amazing.

### 8.2.2 Doing the test in JASP

Okay, now that we know how the test works let's have a look at how it's done in JASP. As tempting as it is to lead you through the tedious calculations so that you're forced to learn it the long way, I figure there's no point. I already showed you how to do it the long way for the goodness-of-fit test in the last section, and since the test of independence isn't conceptually any different, you won't learn anything new by doing it the long way. So instead I'll go straight to showing you the easy way. After you have run the test in JASP ('Frequencies' - 'Contingency Tables'), all you have to do is look underneath the contingency table in the JASP results window and there is the  $\chi^2$  statistic for you. This shows a  $\chi^2$  statistic value of 10.722, with 2 d.f. and  $p$ -value = 0.005.

That was easy, wasn't it! You can also ask JASP to show you the expected counts - just click on the check box for 'Counts' - 'Expected' in the 'Cells' options and the expected counts will appear in the contingency table. And whilst you are doing that, an effect size measure would be helpful. We'll choose Cramer's V, and you can specify this from a check box in the 'Statistics' options, and it gives a value for Cramer's V of 0.244. We will talk about this some more in just a moment.

This output gives us enough information to write up the result:

Pearson's  $\chi^2$  revealed a significant association between species and choice ( $\chi^2(2) = 10.7$ ,  $p < .01$ ). Robots appeared to be more likely to say that they prefer flowers,

---

<sup>\*11</sup>A problem many of us worry about in real life.

but the humans were more likely to say they prefer data.

Notice that, once again, I provided a little bit of interpretation to help the human reader understand what's going on with the data. Later on in my discussion section I'd provide a bit more context. To illustrate the difference, here's what I'd probably say later on:

The fact that humans appeared to have a stronger preference for raw data files than robots is somewhat counter-intuitive. However, in context it makes some sense, as the civil authority on Chapek 9 has an unfortunate tendency to kill and dissect humans when they are identified. As such it seems most likely that the human participants did not respond honestly to the question, so as to avoid potentially undesirable consequences. This should be considered to be a substantial methodological weakness.

This could be classified as a rather extreme example of a reactivity effect, I suppose. Obviously, in this case the problem is severe enough that the study is more or less worthless as a tool for understanding the difference preferences among humans and robots. However, I hope this illustrates the difference between getting a statistically significant result (our null hypothesis is rejected in favour of the alternative), and finding something of scientific value (the data tell us nothing of interest about our research hypothesis due to a big methodological flaw).

#### 8.2.3 Postscript

I later found out the data were made up, and I'd been watching cartoons instead of doing work.

8.3 \_\_\_\_\_

### The continuity correction

Okay, time for a little bit of a digression. I've been lying to you a little bit so far. There's a tiny change that you need to make to your calculations whenever you only have 1 degree of freedom. It's called the "continuity correction", or sometimes the **Yates correction**. Remember what I pointed out earlier: the  $\chi^2$  test is based on an approximation, specifically on the assumption that the binomial distribution starts to look like a normal distribution for large  $N$ . One problem with this is that it often doesn't quite work, especially when you've only got 1 degree of freedom (e.g., when you're doing a test of independence on a  $2 \times 2$  contingency table). The main reason for this is that

the true sampling distribution for the  $X^2$  statistic is actually discrete (because you're dealing with categorical data!) but the  $\chi^2$  distribution is continuous. This can introduce systematic problems. Specifically, when  $N$  is small and when  $df = 1$ , the goodness-of-fit statistic tends to be "too big", meaning that you actually have a bigger  $\alpha$  value than you think (or, equivalently, the  $p$  values are a bit too small).

**Yates1934** suggested a simple fix, in which you redefine the goodness-of-fit statistic as:

$$\chi^2 = \sum_i \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

Basically, he just subtracts off 0.5 everywhere.

As far as I can tell from reading Yates' paper, the correction is basically a hack. It's not derived from any principled theory. Rather, it's based on an examination of the behaviour of the test, and observing that the corrected version seems to work better. You can specify this correction in JASP from a check box in the 'Statistics' options, where it is called ' $\chi^2$  continuity correction'.

## 8.4

---

### Effect size

As we discussed earlier (Section ??), it's becoming commonplace to ask researchers to report some measure of effect size. So, let's suppose that you've run your chi-square test, which turns out to be significant. So you now know that there is some association between your variables (independence test) or some deviation from the specified probabilities (goodness-of-fit test). Now you want to report a measure of effect size. That is, given that there is an association or deviation, how strong is it?

There are several different measures that you can choose to report, and several different tools that you can use to calculate them. I won't discuss all of them but will instead focus on the most commonly reported measures of effect size.

By default, the two measures that people tend to report most frequently are the  $\phi$  statistic and the somewhat superior version, known as Cramér's  $V$ .

Mathematically, they're very simple. To calculate the  $\phi$  statistic, you just divide your  $X^2$  value by the sample size, and take the square root:

$$\phi = \sqrt{\frac{X^2}{N}}$$

The idea here is that the  $\phi$  statistic is supposed to range between 0 (no association at all) and 1 (perfect association), but it doesn't always do this when your contingency table is bigger than  $2 \times 2$ , which is a total pain. For bigger tables it's actually possible to obtain  $\phi > 1$ , which is pretty unsatisfactory. So, to correct for this, people usually prefer to report the  $V$  statistic proposed by **Cramer1946**. It's a pretty simple adjustment to  $\phi$ . If you've got a contingency table with  $r$  rows and  $c$  columns, then define  $k = \min(r, c)$  to be the smaller of the two values.

If so, then **Cramér's V** statistic is

$$V = \sqrt{\frac{X^2}{N(k - 1)}}$$

And you're done. This seems to be a fairly popular measure, presumably because it's easy to calculate, and it gives answers that aren't completely silly. With Cramér's V, you know that the value really does range from 0 (no association at all) to 1 (perfect association).

## 8.5

---

### Assumptions of the test(s)

All statistical tests make assumptions, and it's usually a good idea to check that those assumptions are met. For the chi-square tests discussed so far in this chapter, the assumptions are:

- *Expected frequencies are sufficiently large.* Remember how in the previous section we saw that the  $\chi^2$  sampling distribution emerges because the binomial distribution is pretty similar to a normal distribution? Well, like we discussed in Chapter ?? this is only true when the number of observations is sufficiently large. What that means in practice is that all of the expected frequencies need to be reasonably big. How big is reasonably big? Opinions differ, but the default assumption seems to be that you generally would like to see all your expected frequencies larger than about 5, though for larger tables you would probably be okay if at least 80% of the expected frequencies are above 5 and none of them are below 1. However, from what I've been able to discover (**Cochran1954**) these seem to

have been proposed as rough guidelines, not hard and fast rules, and they seem to be somewhat conservative (**Larntz1978**).

- *Data are independent of one another.* One somewhat hidden assumption of the chi-square test is that you have to genuinely believe that the observations are independent. Here's what I mean. Suppose I'm interested in proportion of babies born at a particular hospital that are boys. I walk around the maternity wards and observe 20 girls and only 10 boys. Seems like a pretty convincing difference, right? But later on, it turns out that I'd actually walked into the same ward 10 times and in fact I'd only seen 2 girls and 1 boy. Not as convincing, is it? My original 30 *observations* were massively non-independent, and were only in fact equivalent to 3 independent observations. Obviously this is an extreme (and extremely silly) example, but it illustrates the basic issue. Non-independence "stuffs things up". Sometimes it causes you to falsely reject the null, as the silly hospital example illustrates, but it can go the other way too. To give a slightly less stupid example, let's consider what would happen if I'd done the cards experiment slightly differently. Instead of asking 200 people to try to imagine sampling one card at random, suppose I asked 50 people to select 4 cards. One possibility would be that *everyone* selects one heart, one club, one diamond and one spade (in keeping with the "representativeness heuristic"; Tversky & Kahneman 1974). This is highly non-random behaviour from people, but in this case I would get an observed frequency of 50 for all four suits. For this example the fact that the observations are non-independent (because the four cards that you pick will be related to each other) actually leads to the opposite effect, falsely retaining the null.

If you happen to find yourself in a situation where independence is violated, it may be possible to use the nonparametric tests, such as the McNemar test or the Cochran test. Similarly, if your expected cell counts are too small, check out the Fisher exact test. At present, JASP does not implement these tests, but check back later! For now, we'll just mention that these tests exist, but describing them is beyond the scope of this book.

## 8.6

---

### Summary

The key ideas discussed in this chapter are:

- The  $\chi^2$  (chi-square) goodness-of-fit test (Section ??) is used when you have a table of observed frequencies of different categories, and the null hypothesis gives you a set of

“known” probabilities to compare them to.

- The  $\chi^2$  (chi-square) test of independence (Section ??) is used when you have a contingency table (cross-tabulation) of two categorical variables. The null hypothesis is that there is no relationship or association between the variables.
- Effect size for a contingency table can be measured in several ways (Section ??). In particular we noted the Cramér’s  $V$  statistic.
- Both versions of the Pearson test rely on two assumptions: that the expected frequencies are sufficiently large, and that the observations are independent (Section ??). Various nonparametric tests can be used for certain kinds of violations of independence or count assumptions.

If you’re interested in learning more about categorical data analysis a good first choice would be **Agresti1996** which, as the title suggests, provides an *Introduction to Categorical Data Analysis*. If the introductory book isn’t enough for you (or can’t solve the problem you’re working on) you could consider **Agresti2002**, *Categorical Data Analysis*. The latter is a more advanced text, so it’s probably not wise to jump straight from this book to that one.

## 9. 二つの平均の比較

---

第 ?? 章では、アウトカムと予測変数のどちらもが名義尺度である場合について説明しました。この世には実際にそのような場面がたくさんあるので、カイ二乗検定が非常に広く使用されていることがよくわかるでしょう。しかしながら、アウトカムが間隔尺度以上であり、ある群のアウトカム変数の平均値が他の群の平均値よりも高いかどうかに关心をもつ場合がそれ以上に多いかもしれません。例えば、心理学者は、子どものいる親は子どもがいない人よりも不安のレベルが高いかどうかを、あるいは音楽を聴くことで（音楽を聴かないと比べると）ワーキングメモリー容量が低下するかどうかを知りたいと思うかもしれません。医療の場面では、新薬が血圧を上昇あるいは低下するかどうかを知りたいと思うかもしれません。農業の科学者は、オーストラリアの原生植物にリンを加えることでそれら植物が死滅するかどうかを知りたいかもしれません。<sup>\*1</sup>これらの全ての例で、アウトカム変数はまさに連続尺度、間隔尺度、あるいは比率尺度の変数であって、予測変数は 2 値の“グループ化”変数です。言い換えるならば、2 群の平均を比較したいのです。

平均を比較する問題は、通常  $t$  検定を用いて答えることができます。 $t$  検定には、解きたい問題に応じて亜型が存在します。そこで、本章では  $t$  検定の種類の違いに焦点を当てて説明します：1 標本の  $t$  検定は ?? で、独立した標本の  $t$  検定は ?? と ?? で、対応のある標本の  $t$  検定は ?? で述べられます。続いて、片側検定（セクション ??）について説明した後で、 $t$  検定の標準化された効果量の指標である Cohen の  $d$  について簡単に説明します（セクション ??）。章の後半では、 $t$  検定の前提と、その前提を逸脱した場合の対処の仕方に焦点をあてます。それらの役立つ事項について論じる前に、まずは  $z$  検定についての理解を深めることから始めます。

### 9.1

---

<sup>\*1</sup>私用の庭で行った非公式な実験によると、リンの付加するとオーストラリア原生植物は死滅することが分かりました。オーストラリアの植物は地球上の他のどの場所よりもリンの濃度が低い地域に適応しています。なので、もし、家を買って庭に外国産の植物と一緒に原生植物を植えたい場合には、それらを分けて植えなければいけません。ヨーロッパの植物にとって栄養となるものがオーストラリアの植物にとって毒になるのです。

## The one-sample z-test

In this section I'll describe one of the most useless tests in all of statistics: the **z-test**. Seriously – this test is almost never used in real life. Its only real purpose is that, when teaching statistics, it's a very convenient stepping stone along the way towards the *t*-test, which is probably the most (over)used tool in all statistics.

### 9.1.1 The inference problem that the test addresses

To introduce the idea behind the *z*-test, let's use a simple example. A friend of mine, Dr. Zeppo, grades his introductory statistics class on a curve. Let's suppose that the average grade in his class is 67.5, and the standard deviation is 9.5. Of his many hundreds of students, it turns out that 20 of them also take psychology classes. Out of curiosity, I find myself wondering if the psychology students tend to get the same grades as everyone else (i.e., mean 67.5) or do they tend to score higher or lower? He emails me the `zeppo.csv` file, which I use to look at the grades of those students in JASP (stored in the variable `x`):

```
50 60 60 64 66 66 67 69 70 74 76 76 77 79 79 79 81 82 82 89
```

Then I calculate the mean in 'Descriptives' - 'Descriptive Statistics'. The mean value is 72.3.

Hmm. It *might* be that the psychology students are scoring a bit higher than normal. That sample mean of  $\bar{X} = 72.3$  is a fair bit higher than the hypothesised population mean of  $\mu = 67.5$  but, on the other hand, a sample size of  $N = 20$  isn't all that big. Maybe it's pure chance.

To answer the question, it helps to be able to write down what it is that I think I know. Firstly, I know that the sample mean is  $\bar{X} = 72.3$ . If I'm willing to assume that the psychology students have the same standard deviation as the rest of the class then I can say that the population standard deviation is  $\sigma = 9.5$ . I'll also assume that since Dr Zeppo is grading to a curve, the psychology student grades are normally distributed.

Next, it helps to be clear about what I want to learn from the data. In this case my research hypothesis relates to the *population* mean  $\mu$  for the psychology student grades, which is unknown. Specifically, I want to know if  $\mu = 67.5$  or not. Given that this is what I know, can we devise a hypothesis test to solve our problem? The data, along with the hypothesised distribution from which they are thought to arise, are shown in Figure ???. Not entirely obvious what the right answer is, is it? For this, we are going to need some statistics.

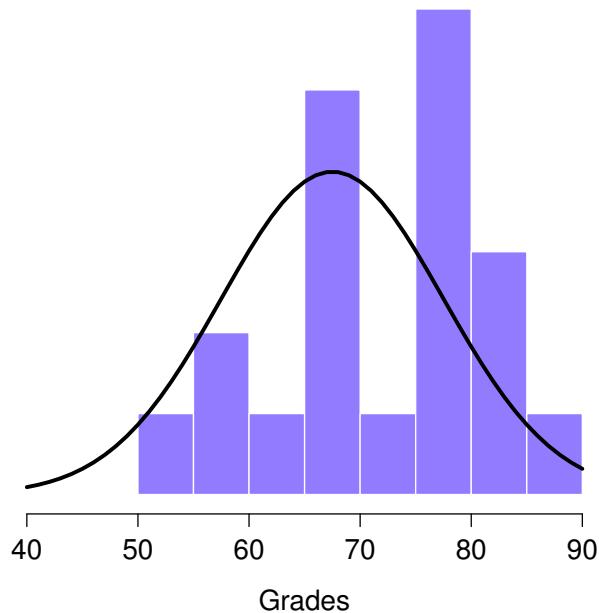


Figure 9.1 The theoretical distribution (solid line) from which the psychology student grades (bars) are supposed to have been generated.

### 9.1.2 Constructing the hypothesis test

The first step in constructing a hypothesis test is to be clear about what the null and alternative hypotheses are. This isn't too hard to do. Our null hypothesis,  $H_0$ , is that the true population mean  $\mu$  for psychology student grades is 67.5%, and our alternative hypothesis is that the population mean *isn't* 67.5%. If we write this in mathematical notation, these hypotheses become:

$$\begin{aligned} H_0 : \mu &= 67.5 \\ H_1 : \mu &\neq 67.5 \end{aligned}$$

though to be honest this notation doesn't add much to our understanding of the problem, it's just a compact way of writing down what we're trying to learn from the data. The null hypothesis  $H_0$  and the alternative hypothesis  $H_1$  for our test are both illustrated in Figure ???. In addition to providing us with these hypotheses, the scenario outlined above provides us with a fair amount of background knowledge that might be useful. Specifically, there are two special pieces of information that we can add:

1. The psychology grades are normally distributed.
2. The true standard deviation of these scores  $\sigma$  is known to be 9.5.

For the moment, we'll act as if these are absolutely trustworthy facts. In real life, this kind of absolutely trustworthy background knowledge doesn't exist, and so if we want to rely on these facts we'll just have to make the *assumption* that these things are true. However, since these assumptions may or may not be warranted, we might need to check them. For now though, we'll keep things simple.

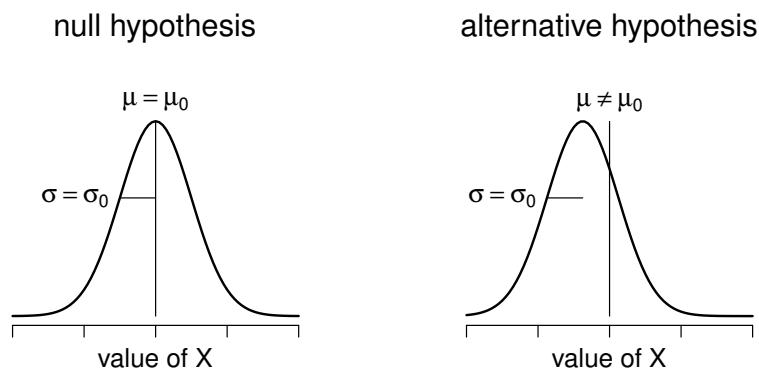


Figure 9.2 Graphical illustration of the null and alternative hypotheses assumed by the one sample z-test (the two sided version, that is). The null and alternative hypotheses both assume that the population distribution is normal, and additionally assumes that the population standard deviation is known (fixed at some value  $\sigma_0$ ). The null hypothesis (left) is that the population mean  $\mu$  is equal to some specified value  $\mu_0$ . The alternative hypothesis is that the population mean differs from this value,  $\mu \neq \mu_0$ .

The next step is to figure out what we would be a good choice for a diagnostic test statistic, something that would help us discriminate between  $H_0$  and  $H_1$ . Given that the hypotheses all refer to the population mean  $\mu$ , you'd feel pretty confident that the sample mean  $\bar{X}$  would be a pretty useful place to start. What we could do is look at the difference between the sample mean  $\bar{X}$  and the value that the null hypothesis predicts for the population mean. In our example that would mean we calculate  $\bar{X} - 67.5$ . More generally, if we let  $\mu_0$  refer to the value that the null hypothesis claims is our population mean, then we'd want to calculate

$$\bar{X} - \mu_0$$

If this quantity equals or is very close to 0, things are looking good for the null hypothesis. If this

quantity is a long way away from 0, then it's looking less likely that the null hypothesis is worth retaining. But how far away from zero should it be for us to reject  $H_0$ ?

To figure that out we need to be a bit more sneaky, and we'll need to rely on those two pieces of background knowledge that I wrote down previously; namely that the raw data are normally distributed and that we know the value of the population standard deviation  $\sigma$ . If the null hypothesis is actually true, and the true mean is  $\mu_0$ , then these facts together mean that we know the complete population distribution of the data: a normal distribution with mean  $\mu_0$  and standard deviation  $\sigma$ . Adopting the notation from Section ??, a statistician might write this as:

$$X \sim \text{Normal}(\mu_0, \sigma^2)$$

Okay, if that's true, then what can we say about the distribution of  $\bar{X}$ ? Well, as we discussed earlier (see Section ??), the sampling distribution of the mean  $\bar{X}$  is also normal, and has mean  $\mu$ . But the standard deviation of this sampling distribution  $\text{se}(\bar{X})$ , which is called the *standard error of the mean*, is

$$\text{se}(\bar{X}) = \frac{\sigma}{\sqrt{N}}$$

In other words, if the null hypothesis is true then the sampling distribution of the mean can be written as follows:

$$\bar{X} \sim \text{Normal}(\mu_0, \text{se}(\bar{X}))$$

Now comes the trick. What we can do is convert the sample mean  $\bar{X}$  into a standard score (Section ??). This is conventionally written as  $z$ , but for now I'm going to refer to it as  $z_{\bar{X}}$ . (The reason for using this expanded notation is to help you remember that we're calculating a standardised version of a sample mean, *not* a standardised version of a single observation, which is what a  $z$ -score usually refers to). When we do so the  $z$ -score for our sample mean is

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\text{se}(\bar{X})}$$

or, equivalently

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

This  $z$ -score is our test statistic. The nice thing about using this as our test statistic is that like all  $z$ -scores, it has a standard normal distribution:

$$z_{\bar{X}} \sim \text{Normal}(0, 1)$$

(again, see Section ?? if you've forgotten why this is true). In other words, regardless of what scale the original data are on, the  $z$ -statistic itself always has the same interpretation: it's equal to the number of standard errors that separate the observed sample mean  $\bar{X}$  from the population mean

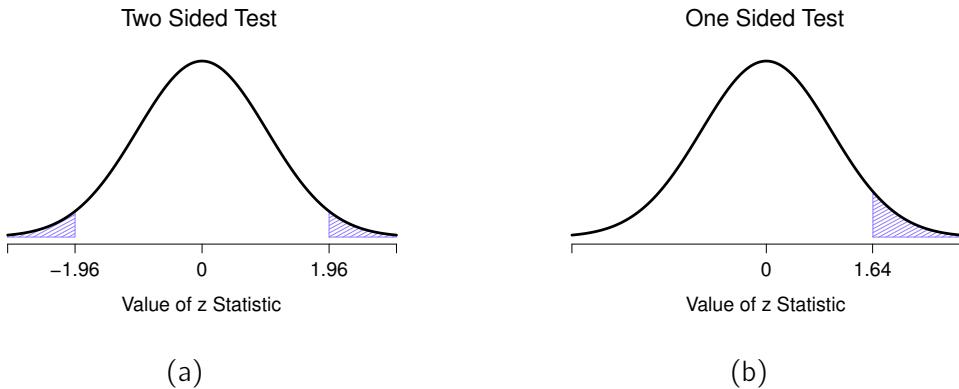


Figure 9.3 Rejection regions for the two-sided  $z$ -test (panel a) and the one-sided  $z$ -test (panel b).

$\mu_0$  predicted by the null hypothesis. Better yet, regardless of what the population parameters for the raw scores actually are, the 5% critical regions for the  $z$ -test are always the same, as illustrated in Figure ???. And what this meant, way back in the days where people did all their statistics by hand, is that someone could publish a table like this:

desired $\alpha$ level	critical $z$ value	
	two-sided test	one-sided test
.1	1.644854	1.281552
.05	1.959964	1.644854
.01	2.575829	2.326348
.001	3.290527	3.090232

This, in turn, meant that researchers could calculate their  $z$ -statistic by hand and then look up the critical value in a text book.

### 9.1.3 A worked example, by hand

Now, as I mentioned earlier, the  $z$ -test is almost never used in practice. It's so rarely used in real life that JASP doesn't have a built in function for it. However, the test is so incredibly simple that it's really easy to do one manually. Let's go back to the data from Dr Zeppo's class. Having loaded the `grades` data, the first thing I need to do is calculate the sample mean, which I've already done (72.3). We already have the known population standard deviation ( $\sigma = 9.5$ ), and the value of the population mean that the null hypothesis specifies ( $\mu_0 = 67.5$ ), and we know the sample

size ( $N=20$ ).

Next, let's calculate the (true) standard error of the mean (easily done with a calculator):

$$\begin{aligned}se(\bar{X}) &= \frac{\sigma}{\sqrt{N}} \\&= \frac{9.5}{\sqrt{20}} \\&= 2.124265\end{aligned}$$

From this, we calculate our  $z$ -score:

$$\begin{aligned}z_{\bar{X}} &= \frac{\bar{X} - \mu_0}{sd(\bar{X})} \\&= \frac{72.3 - 67.5}{2.124265} \\&= 2.259606.\end{aligned}$$

At this point, we would traditionally look up the value 2.26 in our table of critical values. Our original hypothesis was two-sided (we didn't really have any theory about whether psych students would be better or worse at statistics than other students) so our hypothesis test is two-sided (or two-tailed) also. Looking at the little table that I showed earlier, we can see that 2.26 is bigger than the critical value of 1.96 that would be required to be significant at  $\alpha = .05$ , but smaller than the value of 2.58 that would be required to be significant at a level of  $\alpha = .01$ . Therefore, we can conclude that we have a significant effect, which we might write up by saying something like this:

With a mean grade of 73.2 in the sample of psychology students, and assuming a true population standard deviation of 9.5, we can conclude that the psychology students have significantly different statistics scores to the class average ( $z = 2.26$ ,  $N = 20$ ,  $p < .05$ ).

#### 9.1.4 Assumptions of the $z$ -test

As I've said before, all statistical tests make assumptions. Some tests make reasonable assumptions, while other tests do not. The test I've just described, the one sample  $z$ -test, makes three basic assumptions. These are:

- *Normality*. As usually described, the  $z$ -test assumes that the true population distribution is normal.<sup>\*2</sup> This is often a pretty reasonable assumption, and it's also an assumption that we can check if we feel worried about it (see Section ??).
- *Independence*. The second assumption of the test is that the observations in your data set are not correlated with each other, or related to each other in some funny way. This isn't as easy to check statistically, it relies a bit on good experimental design. An obvious (and silly) example of something that violates this assumption is a data set where you "copy" the same observation over and over again in your data file so that you end up with a massive "sample size", which consists of only one genuine observation. More realistically, you have to ask yourself if it's really plausible to imagine that each observation is a completely random sample from the population that you're interested in. In practice this assumption is never met, but we try our best to design studies that minimise the problems of correlated data.
- *Known standard deviation*. The third assumption of the  $z$ -test is that the true standard deviation of the population is known to the researcher. This is just silly. In no real world data analysis problem do you know the standard deviation  $\sigma$  of some population but are completely ignorant about the mean  $\mu$ . In other words, this assumption is *always* wrong.

In view of the stupidity of assuming that  $\sigma$  is known, let's see if we can live without it. This takes us out of the dreary domain of the  $z$ -test, and into the magical kingdom of the  $t$ -test!

## 9.2

---

### The one-sample $t$ -test

After some thought, I decided that it might not be safe to assume that the psychology student grades necessarily have the same standard deviation as the other students in Dr Zeppo's class. After all, if I'm entertaining the hypothesis that they don't have the same mean, then why should I believe that they absolutely have the same standard deviation? In view of this, I should really stop assuming that I know the true value of  $\sigma$ . This violates the assumptions of my  $z$ -test, so in one sense I'm back to square one. However, it's not like I'm completely bereft of options. After

---

<sup>\*2</sup>Actually this is too strong. Strictly speaking the  $z$  test only requires that the sampling distribution of the mean be normally distributed. If the population is normal then it necessarily follows that the sampling distribution of the mean is also normal. However, as we saw when talking about the central limit theorem, it's quite possible (even commonplace) for the sampling distribution to be normal even if the population distribution itself is non-normal. However, in light of the sheer ridiculousness of the assumption that the true standard deviation is known, there really isn't much point in going into details on this front!

all, I've still got my raw data, and those raw data give me an *estimate* of the population standard deviation, which is 9.52. In other words, while I can't say that I know that  $\sigma = 9.5$ , I *can* say that  $\hat{\sigma} = 9.52$ .

Okay, cool. The obvious thing that you might think to do is run a *z*-test, but using the estimated standard deviation of 9.52 instead of relying on my assumption that the true standard deviation is 9.5. And you probably wouldn't be surprised to hear that this would still give us a significant result. This approach is close, but it's not *quite* correct. Because we are now relying on an *estimate* of the population standard deviation we need to make some adjustment for the fact that we have some uncertainty about what the true population standard deviation actually is. Maybe our data are just a fluke . . . maybe the true population standard deviation is 11, for instance. But if that were actually true, and we ran the *z*-test assuming  $\sigma=11$ , then the result would end up being *non-significant*. That's a problem, and it's one we're going to have to address.

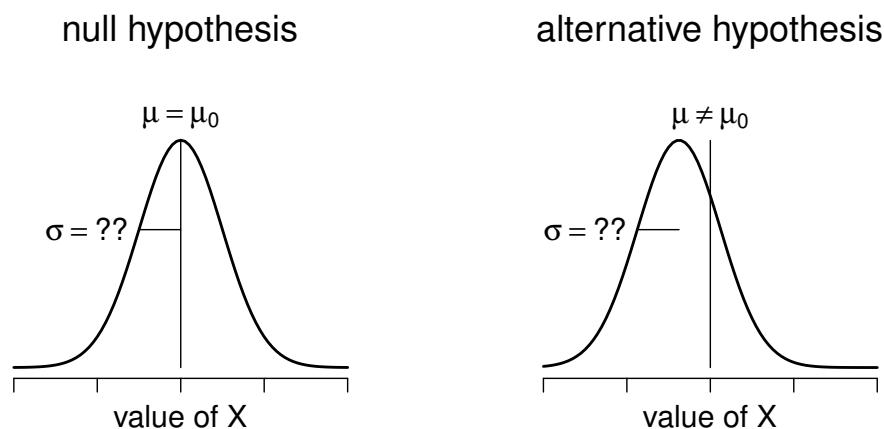


Figure 9.4 Graphical illustration of the null and alternative hypotheses assumed by the (two sided) one sample *t*-test. Note the similarity to the *z*-test (Figure ??). The null hypothesis is that the population mean  $\mu$  is equal to some specified value  $\mu_0$ , and the alternative hypothesis is that it is not. Like the *z*-test, we assume that the data are normally distributed, but we do not assume that the population standard deviation  $\sigma$  is known in advance.

---

### 9.2.1 Introducing the *t*-test

This ambiguity is annoying, and it was resolved in 1908 by a guy called William Sealy Gosset

(**Student1908**), who was working as a chemist for the Guinness brewery at the time (**Box1987**). Because Guinness took a dim view of its employees publishing statistical analysis (apparently they felt it was a trade secret), he published the work under the pseudonym “A Student” and, to this day, the full name of the *t*-test is actually **Student’s *t*-test**. The key thing that Gosset figured out is how we should accommodate the fact that we aren’t completely sure what the true standard deviation is.<sup>\*3</sup> The answer is that it subtly changes the sampling distribution. In the *t*-test our test statistic, now called a *t*-statistic, is calculated in exactly the same way I mentioned above. If our null hypothesis is that the true mean is  $\mu$ , but our sample has mean  $\bar{X}$  and our estimate of the population standard deviation is  $\hat{\sigma}$ , then our *t* statistic is:

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

The only thing that has changed in the equation is that instead of using the known true value  $\sigma$ , we use the estimate  $\hat{\sigma}$ . And if this estimate has been constructed from  $N$  observations, then the sampling distribution turns into a *t*-distribution with  $N - 1$  **degrees of freedom** (df). The *t* distribution is very similar to the normal distribution, but has “heavier” tails, as discussed earlier in Section ?? and illustrated in Figure ?? . Notice, though, that as df gets larger, the *t*-distribution starts to look identical to the standard normal distribution. This is as it should be: if you have a sample size of  $N = 70,000,000$  then your “estimate” of the standard deviation would be pretty much perfect, right? So, you should expect that for large  $N$ , the *t*-test would behave exactly the same way as a *z*-test. And that’s exactly what happens!

### 9.2.2 Doing the test in JASP

As you might expect, the mechanics of the *t*-test are almost identical to the mechanics of the *z*-test. So there’s not much point in going through the tedious exercise of showing you how to do the calculations using low level commands. It’s pretty much identical to the calculations that we did earlier, except that we use the estimated standard deviation and then we test our hypothesis using the *t* distribution rather than the normal distribution. And so instead of going through the calculations in tedious detail for a second time, I’ll jump straight to showing you how *t*-tests are actually done. JASP comes with a dedicated analysis for *t*-tests that is very flexible (it can run lots of different kinds of *t*-tests). It’s pretty straightforward to use; all you need to do is specify ‘T-Tests’ - ‘One Sample T-Test’, move the variable you are interested in ( $x$ ) across into the ‘Variables’ box, and type in the mean value for the null hypothesis (‘67.5’) in the ‘Test value’ box.

---

<sup>\*3</sup>Well, sort of. As I understand the history, Gosset only provided a partial solution; the general solution to the problem was provided by Sir Ronald Fisher.

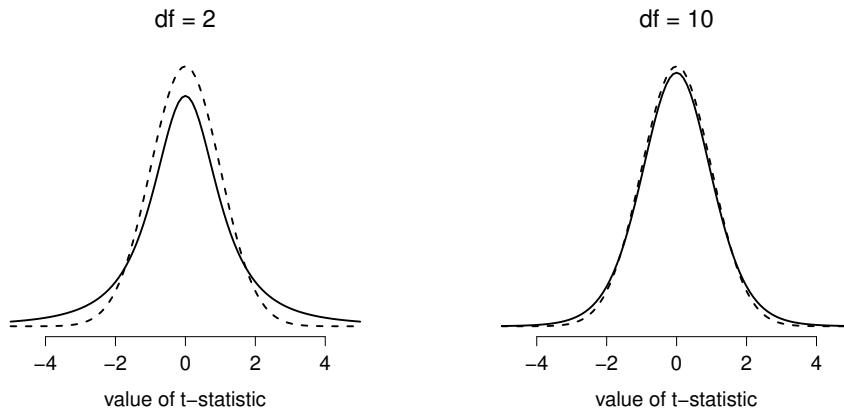


Figure 9.5 The  $t$  distribution with 2 degrees of freedom (left) and 10 degrees of freedom (right), with a standard normal distribution (i.e., mean 0 and std dev 1) plotted as dotted lines for comparison purposes. Notice that the  $t$  distribution has heavier tails (leptokurtic: higher kurtosis) than the normal distribution; this effect is quite exaggerated when the degrees of freedom are very small, but negligible for larger values. In other words, for large  $df$  the  $t$  distribution is essentially identical to a normal distribution.

.....

Easy enough. See Figure ??, which, amongst other things that we will get to in a moment, gives you a  $t$ -test statistic = 2.25, with 19 degrees of freedom and an associated  $p$ -value of 0.036.

It is also easy to calculate a 95% confidence interval for our sample mean. If you select the ‘Location parameter’ and its associated ‘Confidence interval’ option under ‘Additional Statistics’, you’ll see in the JASP output that the ‘Mean difference’ is 4.800 with 95% CI equal to [0.344, 9.256]. This simply means that we are 95% confidence that our estimate of the difference between our sample and the hypothesized mean of 67.5 is between 0.344 and 9.256. If we add these “endpoints” to the hypothesized mean, we get a 95% CI of [67.5+0.344, 67.5+9.256], or said differently, [67.844, 76.800]. If this isn’t clear, don’t worry. We’ll explain a bit more about this in the next section.

Now, what do we *do* with all this output? Well, since we’re pretending that we actually care about my toy example, we’re overjoyed to discover that the result is statistically significant (i.e.  $p$  value below .05). We could report the result by saying something like this:

With a mean grade of 72.3, the psychology students scored slightly higher than the average grade of 67.5 ( $t(19) = 2.25, p < .05$ ); the 95% confidence interval is 67.8 to 76.8.

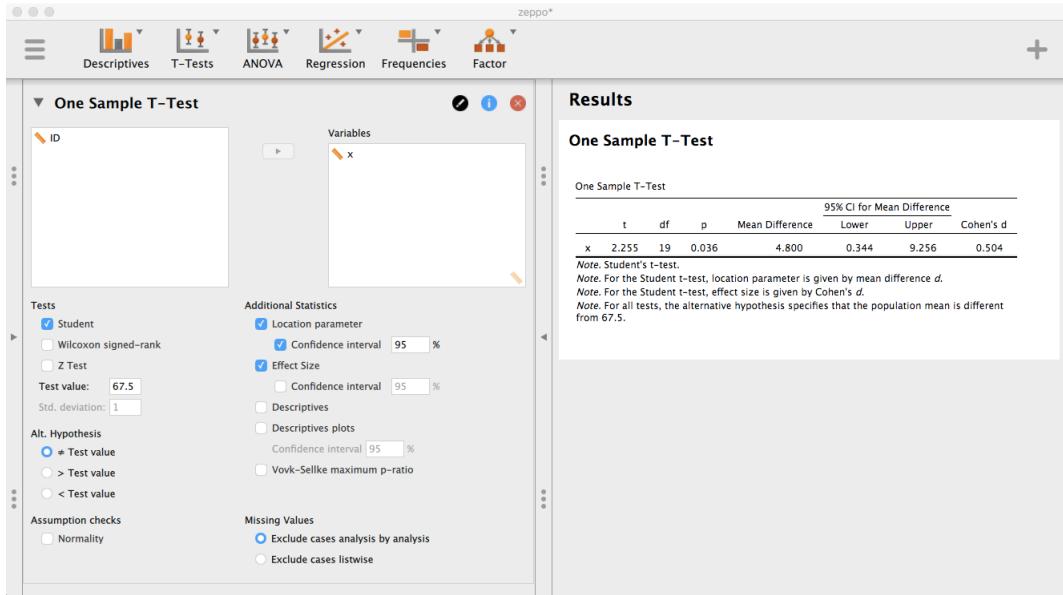


Figure9.6 JASP does the one-sample t-test.

.....

where  $t(19)$  is shorthand notation for a  $t$ -statistic that has 19 degrees of freedom. That said, it's often the case that people don't report the confidence interval, or do so using a much more compressed form than I've done here. For instance, it's not uncommon to see the confidence interval included as part of the stat block, like this:

$$t(19) = 2.25, p < .05, \text{CI}_{95} = [67.8, 76.8]$$

With that much jargon crammed into half a line, you know it must be really smart.\*<sup>4</sup>

### 9.2.3 Assumptions of the one sample $t$ -test

Okay, so what assumptions does the one-sample  $t$ -test make? Well, since the  $t$ -test is basically a  $z$ -test with the assumption of known standard deviation removed, you shouldn't be surprised to see that it makes the same assumptions as the  $z$ -test, minus the one about the known standard deviation. That is

---

\*<sup>4</sup>More seriously, I tend to think the reverse is true. I get very suspicious of technical reports that fill their results sections with nothing except the numbers. It might just be that I'm an arrogant jerk, but I often feel like an author that makes no attempt to explain and interpret their analysis to the reader either doesn't understand it themselves, or is being a bit lazy. Your readers are smart, but not infinitely patient. Don't annoy them if you can help it.

- *Normality*. We're still assuming that the population distribution is normal<sup>\*5</sup>, and as noted earlier, there are standard tools that you can use to check to see if this assumption is met (Section ??), and other tests you can do in its place if this assumption is violated (Section ??).
- *Independence*. Once again, we have to assume that the observations in our sample are generated independently of one another. See the earlier discussion about the z-test for specifics (Section ??).

Overall, these two assumptions aren't terribly unreasonable, and as a consequence the one-sample *t*-test is pretty widely used in practice as a way of comparing a sample mean against a hypothesised population mean.

## 9.3

---

### The independent samples *t*-test (Student test)

Although the one sample *t*-test has its uses, it's not the most typical example of a *t*-test<sup>\*6</sup>. A much more common situation arises when you've got two different groups of observations. In psychology, this tends to correspond to two different groups of participants, where each group corresponds to a different condition in your study. For each person in the study you measure some outcome variable of interest, and the research question that you're asking is whether or not the two groups have the same population mean. This is the situation that the independent samples *t*-test is designed for.

#### 9.3.1 The data

Suppose we have 33 students taking Dr Harpo's statistics lectures, and Dr Harpo doesn't grade to a curve. Actually, Dr Harpo's grading is a bit of a mystery, so we don't really know anything about what the average grade is for the class as a whole. There are two tutors for the class,

<sup>\*5</sup>A technical comment. In the same way that we can weaken the assumptions of the z-test so that we're only talking about the sampling distribution, we *can* weaken the *t*-test assumptions so that we don't have to assume normality of the population. However, for the *t*-test it's trickier to do this. As before, we can replace the assumption of population normality with an assumption that the sampling distribution of  $\bar{X}$  is normal. However, remember that we're also relying on a sample estimate of the standard deviation, and so we also require the sampling distribution of  $\hat{\sigma}$  to be chi-square. That makes things nastier, and this version is rarely used in practice. Fortunately, if the population distribution is normal, then both of these two assumptions are met.

<sup>\*6</sup>Although it is the simplest, which is why I started with it.

Anastasia and Bernadette. There are  $N_1 = 15$  students in Anastasia's tutorials, and  $N_2 = 18$  in Bernadette's tutorials. The research question I'm interested in is whether Anastasia or Bernadette is a better tutor, or if it doesn't make much of a difference. Dr Harpo emails me the course grades, in the `harpo.csv` file. As usual, I'll load the file into JASP and have a look at what variables it contains - there are three variables, `ID`, `grade` and `tutor`. Not surprisingly, the `grade` variable contains each student's grade. The `tutor` variable is a factor that indicates who each student's tutor was - either Anastasia or Bernadette.

We can calculate means and standard deviations, using the 'Descriptives' - 'Descriptive Statistics' analysis (being sure to split by `tutor`). Here's a nice little summary table:

	mean	std dev	N
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

To give you a more detailed sense of what's going on here, I've plotted histograms (not in JASP, but using R) showing the distribution of grades for both tutors (Figure ??), as well as a simpler plot showing the means and corresponding confidence intervals for both groups of students (Figure ??).

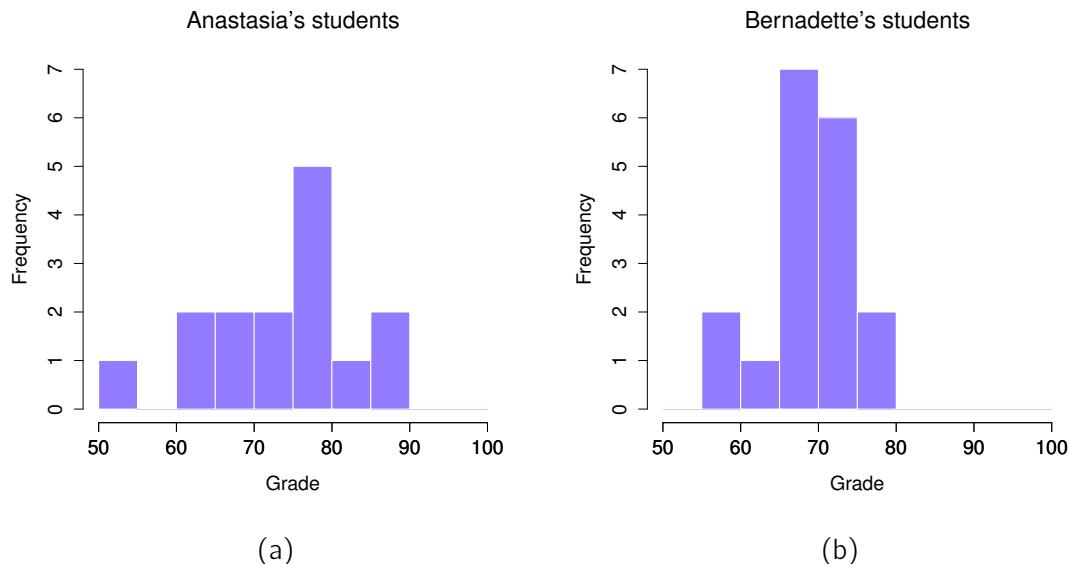


Figure 9.7 Histograms showing the distribution of grades for students in Anastasia's (panel a) and in Bernadette's (panel b) classes. Visually, these suggest that students in Anastasia's class may be getting slightly better grades on average, though they also seem a bit more variable.

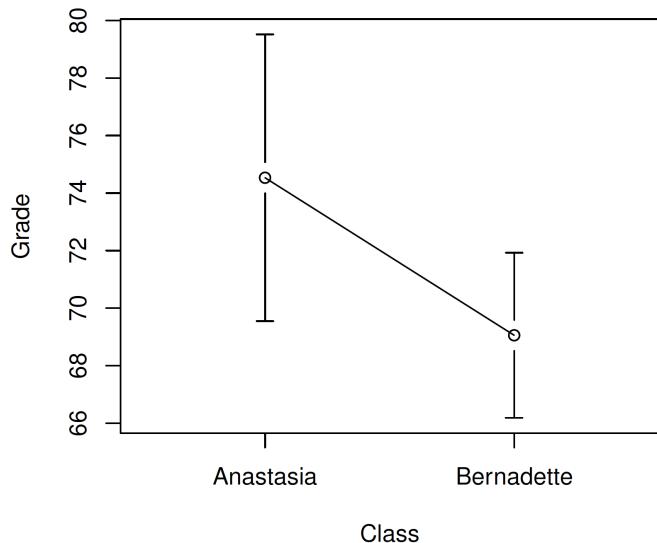


Figure9.8 The plots show the mean grade for students in Anastasia’ s and Bernadette’ s tutorials. Error bars depict 95% confidence intervals around the mean. Visually, it does look like there’s a real difference between the groups, though it’s hard to say for sure.

---

### 9.3.2 Introducing the test

The **independent samples  $t$ -test** comes in two different forms, Student’s and Welch’s. The original Student  $t$ -test, which is the one I’ll describe in this section, is the simpler of the two but relies on much more restrictive assumptions than the Welch  $t$ -test. Assuming for the moment that you want to run a two-sided test, the goal is to determine whether two “independent samples” of data are drawn from populations with the same mean (the null hypothesis) or different means (the alternative hypothesis). When we say “independent” samples, what we really mean here is that there’s no special relationship between observations in the two samples. This probably doesn’t make a lot of sense right now, but it will be clearer when we come to talk about the paired samples  $t$ -test later on. For now, let’s just point out that if we have an experimental design where participants are randomly allocated to one of two groups, and we want to compare the two groups’ mean performance on some outcome measure, then an independent samples  $t$ -test (rather than a paired samples  $t$ -test) is what we’re after.

Okay, so let’s let  $\mu_1$  denote the true population mean for group 1 (e.g., Anastasia’s students),

and  $\mu_2$  will be the true population mean for group 2 (e.g., Bernadette's students),<sup>\*7</sup> and as usual we'll let  $\bar{X}_1$  and  $\bar{X}_2$  denote the observed sample means for both of these groups. Our null hypothesis states that the two population means are identical ( $\mu_1 = \mu_2$ ) and the alternative to this is that they are not ( $\mu_1 \neq \mu_2$ ). Written in mathematical notation, this is:

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_1 : \mu_1 &\neq \mu_2 \end{aligned}$$

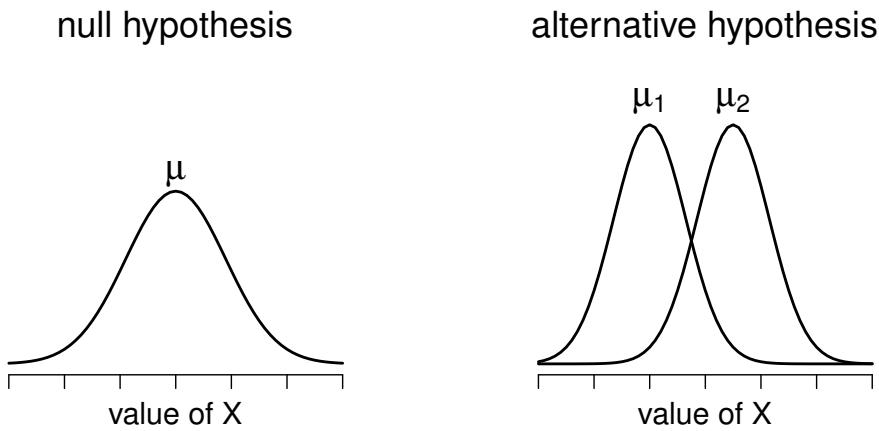


Figure 9.9 Graphical illustration of the null and alternative hypotheses assumed by the Student  $t$ -test. The null hypothesis assumes that both groups have the same mean  $\mu$ , whereas the alternative assumes that they have different means  $\mu_1$  and  $\mu_2$ . Notice that it is assumed that the population distributions are normal, and that, although the alternative hypothesis allows the group to have different means, it assumes they have the same standard deviation.

To construct a hypothesis test that handles this scenario we start by noting that if the null hypothesis is true, then the difference between the population means is *exactly* zero,  $\mu_1 - \mu_2 = 0$ .

---

<sup>\*7</sup>A funny question almost always pops up at this point: what the heck *is* the population being referred to in this case? Is it the set of students actually taking Dr Harpo's class (all 33 of them)? The set of people who might take the class (an unknown number of them)? Or something else? Does it matter which of these we pick? It's traditional in an introductory behavioural stats class to mumble a lot at this point, but since I get asked this question every year by my students, I'll give a brief answer. Technically yes, it does matter. If you change your definition of what the "real world" population actually is, then the sampling distribution of your observed mean  $\bar{X}$  changes too. The  $t$ -test relies on an assumption that the observations are sampled at random from an infinitely large population and, to the extent that real life isn't like that, then the  $t$ -test can be wrong. In practice, however, this isn't usually a big deal. Even though the assumption is almost always wrong, it doesn't lead to a lot of pathological behaviour from the test, so we tend to just ignore it.

As a consequence, a diagnostic test statistic will be based on the difference between the two sample means. Because if the null hypothesis is true, then we'd expect  $\bar{X}_1 - \bar{X}_2$  to be *pretty close* to zero. However, just like we saw with our one-sample tests (i.e., the one-sample z-test and the one-sample t-test) we have to be precise about exactly *how close* to zero this difference should be. And the solution to the problem is more or less the same one. We calculate a standard error estimate (SE), just like last time, and then divide the difference between means by this estimate. So our **t-statistic** will be of the form:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{SE}}$$

We just need to figure out what this standard error estimate actually is. This is a bit trickier than was the case for either of the two tests we've looked at so far, so we need to go through it a lot more carefully to understand how it works.

### 9.3.3 A “pooled estimate” of the standard deviation

In the original “Student t-test”, we make the assumption that the two groups have the same population standard deviation. That is, regardless of whether the population means are the same, we assume that the population standard deviations are identical,  $\sigma_1 = \sigma_2$ . Since we're assuming that the two standard deviations are the same, we drop the subscripts and refer to both of them as  $\sigma$ . How should we estimate this? How should we construct a single estimate of a standard deviation when we have two samples? The answer is, basically, we average them. Well, sort of. Actually, what we do is take a *weighed* average of the *variance* estimates, which we use as our **pooled estimate of the variance**. The weight assigned to each sample is equal to the number of observations in that sample, minus 1.

Mathematically, we can write this as

$$\begin{aligned} w_1 &= N_1 - 1 \\ w_2 &= N_2 - 1 \end{aligned}$$

Now that we've assigned weights to each sample we calculate the pooled estimate of the variance by taking the weighted average of the two variance estimates,  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$

$$\hat{\sigma}_p^2 = \frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}$$

Finally, we convert the pooled variance estimate to a pooled standard deviation estimate, by taking the square root.

$$\hat{\sigma}_p = \sqrt{\frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}}$$

And if you mentally substitute  $w_1 = N_1 - 1$  and  $w_2 = N_2 - 1$  into this equation you get a very ugly looking formula. A very ugly formula that actually seems to be the “standard” way of describing the pooled standard deviation estimate. It’s not my favourite way of thinking about pooled standard deviations, however. I prefer to think about it like this. Our data set actually corresponds to a set of  $N$  observations which are sorted into two groups. So let’s use the notation  $X_{ik}$  to refer to the grade received by the  $i$ -th student in the  $k$ -th tutorial group. That is,  $X_{11}$  is the grade received by the first student in Anastasia’s class,  $X_{21}$  is her second student, and so on. And we have two separate group means  $\bar{X}_1$  and  $\bar{X}_2$ , which we could “generically” refer to using the notation  $\bar{X}_k$ , i.e., the mean grade for the  $k$ -th tutorial group. So far, so good. Now, since every single student falls into one of the two tutorials, we can describe their deviation from the group mean as the difference

$$X_{ik} - \bar{X}_k$$

So why not just use these deviations (i.e., the extent to which each student’s grade differs from the mean grade in their tutorial)? Remember, a variance is just the average of a bunch of squared deviations, so let’s do that. Mathematically, we could write it like this

$$\frac{\sum_{ik} (X_{ik} - \bar{X}_k)^2}{N}$$

where the notation “ $\sum_{ik}$ ” is a lazy way of saying “calculate a sum by looking at all students in all tutorials”, since each “ $ik$ ” corresponds to one student.<sup>a</sup> But, as we saw in Chapter ??, calculating the variance by dividing by  $N$  produces a biased estimate of the population variance. And previously we needed to divide by  $N - 1$  to fix this. However, as I mentioned at the time, the reason why this bias exists is because the variance estimate relies on the sample mean, and to the extent that the sample mean isn’t equal to the population mean it can systematically bias our estimate of the variance. But this time we’re relying on *two* sample means! Does this mean that we’ve got more bias? Yes, yes it does. And does this mean we now need to divide by  $N - 2$  instead of  $N - 1$ , in order to calculate our pooled variance estimate? Why, yes

$$\hat{\sigma}_p^2 = \frac{\sum_{ik} (X_{ik} - \bar{X}_k)^2}{N - 2}$$

Oh, and if you take the square root of this then you get  $\hat{\sigma}_p$ , the pooled standard deviation estimate. In other words, the pooled standard deviation calculation is nothing special. It’s not

terribly different to the regular standard deviation calculation.

<sup>a</sup>A more correct notation will be introduced in Chapter ??.

#### 9.3.4 Completing the test

Regardless of which way you want to think about it, we now have our pooled estimate of the standard deviation. From now on, I'll drop the silly  $p$  subscript, and just refer to this estimate as  $\hat{\sigma}$ . Great. Let's now go back to thinking about the bloody hypothesis test, shall we? Our whole reason for calculating this pooled estimate was that we knew it would be helpful when calculating our *standard error* estimate. But standard error of *what*? In the one-sample  $t$ -test it was the standard error of the sample mean,  $\text{se}(\bar{X})$ , and since  $\text{se}(\bar{X}) = \sigma/\sqrt{N}$  that's what the denominator of our  $t$ -statistic looked like. This time around, however, we have *two* sample means. And what we're interested in, specifically, is the difference between the two  $\bar{X}_1 - \bar{X}_2$ . As a consequence, the standard error that we need to divide by is in fact the **standard error of the difference** between means.

As long as the two variables really do have the same standard deviation, then our estimate for the standard error is

$$\text{se}(\bar{X}_1 - \bar{X}_2) = \hat{\sigma} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

and our  $t$ -statistic is therefore

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{se}(\bar{X}_1 - \bar{X}_2)}$$

Just as we saw with our one-sample test, the sampling distribution of this  $t$ -statistic is a  $t$ -distribution (shocking, isn't it?) as long as the null hypothesis is true and all of the assumptions of the test are met. The degrees of freedom, however, is slightly different. As usual, we can think of the degrees of freedom to be equal to the number of data points minus the number of constraints. In this case, we have  $N$  observations ( $N_1$  in sample 1, and  $N_2$  in sample 2), and 2 constraints (the sample means). So the total degrees of freedom for this test are  $N - 2$ .

#### 9.3.5 Doing the test in JASP

Not surprisingly, you can run an independent samples  $t$ -test easily in JASP. The outcome variable for our test is the student `grade`, and the groups are defined in terms of the `tutor` for each class. So you probably won't be too surprised that all you have to do in JASP is go to the relevant

analysis ('T-Tests' - 'Independent Samples T-Test') and move the `grade` variable across to the 'Variables' box, and the `tutor` variable across into the 'Grouping Variable' box, as shown in Figure ??.

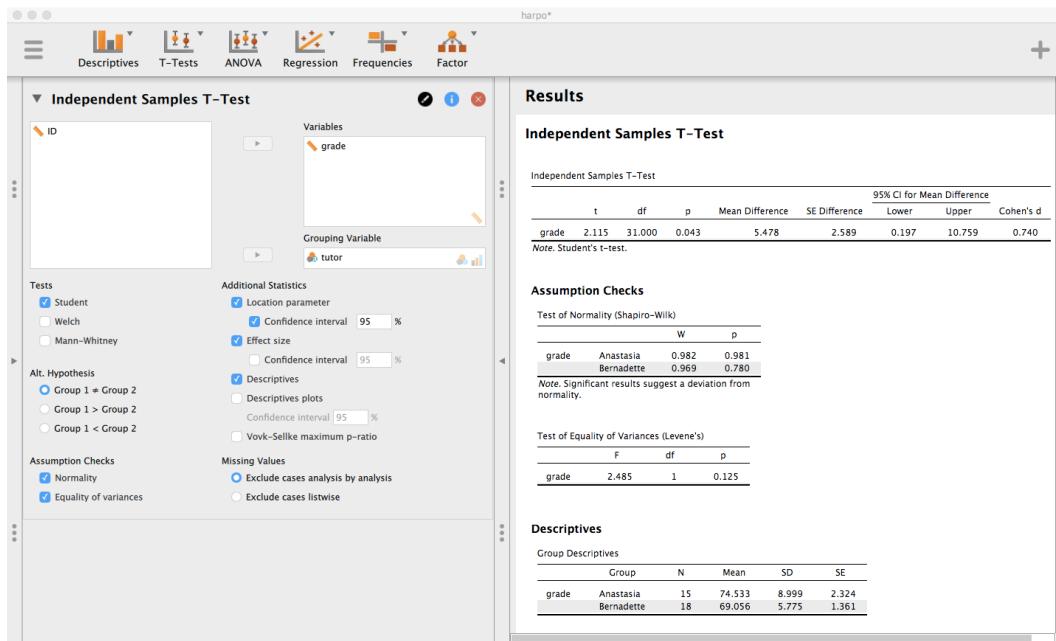


Figure9.10 Independent  $t$ -test in JASP, with options checked for useful results

The output has a very familiar form. First, it tells you what test was run, and it tells you the name of the dependent variable that you used. It then reports the test results. Just like last time the test results consist of a  $t$ -statistic, the degrees of freedom, and the  $p$ -value. The final section reports two things: it gives you a confidence interval and an effect size. I'll talk about effect sizes later. The confidence interval, however, I should talk about now.

It's pretty important to be clear on what this confidence interval actually refers to. It is a confidence interval for the *difference* between the group means. In our example, Anastasia's students had an average grade of 74.533, and Bernadette's students had an average grade of 69.056, so the difference between the two sample means is 5.478. But of course the difference between population means might be bigger or smaller than this. The confidence interval reported in Figure ?? tells you that there's a if we replicated this study again and again, then 95% of the time the true difference in means would lie between 0.197 and 10.759. Look back at Section ?? for a reminder about what confidence intervals mean.

In any case, the difference between the two groups is significant (just barely), so we might write

up the result using text like this:

The mean grade in Anastasia's class was 74.5% (std dev = 9.0), whereas the mean in Bernadette's class was 69.1% (std dev = 5.8). A Student's independent samples  $t$ -test showed that this 5.4% difference was significant ( $t(31) = 2.1, p < .05, Cl_{95} = [0.2, 10.8], d = .74$ ), suggesting that a genuine difference in learning outcomes has occurred.

Notice that I've included the confidence interval and the effect size in the stat block. People don't always do this. At a bare minimum, you'd expect to see the  $t$ -statistic, the degrees of freedom and the  $p$  value. So you should include something like this at a minimum:  $t(31) = 2.1, p < .05$ . If statisticians had their way, everyone would also report the confidence interval and probably the effect size measure too, because they are useful things to know. But real life doesn't always work the way statisticians want it to so you should make a judgment based on whether you think it will help your readers and, if you're writing a scientific paper, the editorial standard for the journal in question. Some journals expect you to report effect sizes, others don't. Within some scientific communities it is standard practice to report confidence intervals, in others it is not. You'll need to figure out what your audience expects. But, just for the sake of clarity, if you're taking my class, my default position is that it's usually worth including both the effect size and the confidence interval.

### 9.3.6 Positive and negative $t$ values

Before moving on to talk about the assumptions of the  $t$ -test, there's one additional point I want to make about the use of  $t$ -tests in practice. The first one relates to the sign of the  $t$ -statistic (that is, whether it is a positive number or a negative one). One very common worry that students have when they start running their first  $t$ -test is that they often end up with negative values for the  $t$ -statistic and don't know how to interpret it. In fact, it's not at all uncommon for two people working independently to end up with results that are almost identical, except that one person has a negative  $t$  value and the other one has a positive  $t$  value. Assuming that you're running a two-sided test then the  $p$ -values will be identical. On closer inspection, the students will notice that the confidence intervals also have the opposite signs. This is perfectly okay. Whenever this happens, what you'll find is that the two versions of the results arise from slightly different ways of running the  $t$ -test. What's happening here is very simple. The  $t$ -statistic that we calculate here

is always of the form

$$t = \frac{(\text{mean 1}) - (\text{mean 2})}{(\text{SE})}$$

If “mean 1” is larger than “mean 2” the  $t$  statistic will be positive, whereas if “mean 2” is larger then the  $t$  statistic will be negative. Similarly, the confidence interval that JASP reports is the confidence interval for the difference “(mean 1) minus (mean 2)”, which will be the reverse of what you’d get if you were calculating the confidence interval for the difference “(mean 2) minus (mean 1)”.

Okay, that’s pretty straightforward when you think about it, but now consider our  $t$ -test comparing Anastasia’s class to Bernadette’s class. Which one should we call “mean 1” and which one should we call “mean 2”. It’s arbitrary. However, you really do need to designate one of them as “mean 1” and the other one as “mean 2”. Not surprisingly, the way that JASP handles this is also pretty arbitrary. In earlier versions of the book I used to try to explain it, but after a while I gave up, because it’s not really all that important and to be honest I can never remember myself. Whenever I get a significant  $t$ -test result, and I want to figure out which mean is the larger one, I don’t try to figure it out by looking at the  $t$ -statistic. Why would I bother doing that? It’s foolish. It’s easier just to look at the actual group means since the JASP output actually shows them!

Here’s the important thing. Because it really doesn’t matter what JASP shows you, I usually try to *report* the  $t$ -statistic in such a way that the numbers match up with the text. Suppose that what I want to write in my report is “Anastasia’s class had higher grades than Bernadette’s class”. The phrasing here implies that Anastasia’s group comes first, so it makes sense to report the  $t$ -statistic as if Anastasia’s class corresponded to group 1. If so, I would write

Anastasia’s class had higher grades than Bernadette’s class ( $t(31) = 2.1, p = .04$ ).

(I wouldn’t actually underline the word “higher” in real life, I’m just doing it to emphasise the point that “higher” corresponds to positive  $t$  values). On the other hand, suppose the phrasing I wanted to use has Bernadette’s class listed first. If so, it makes more sense to treat her class as group 1, and if so, the write up looks like this

Bernadette’s class had lower grades than Anastasia’s class ( $t(31) = -2.1, p = .04$ ).

Because I’m talking about one group having “lower” scores this time around, it is more sensible to use the negative form of the  $t$ -statistic. It just makes it read more cleanly.

One last thing: please note that you *can’t* do this for other types of test statistics. It works for  $t$ -tests, but it wouldn’t be meaningful for chi-square tests,  $F$ -tests or indeed for most of the tests I talk about in this book. So don’t over-generalise this advice! I’m really just talking about  $t$ -tests

here and nothing else!

### 9.3.7 Assumptions of the test

As always, our hypothesis test relies on some assumptions. So what are they? For the Student  $t$ -test there are three assumptions, some of which we saw previously in the context of the one sample  $t$ -test (see Section ??):

- *Normality*. Like the one-sample  $t$ -test, it is assumed that the data are normally distributed. Specifically, we assume that both groups are normally distributed. In Section ?? we'll discuss how to test for normality, and in Section ?? we'll discuss possible solutions.
- *Independence*. Once again, it is assumed that the observations are independently sampled. In the context of the Student test this has two aspects to it. Firstly, we assume that the observations within each sample are independent of one another (exactly the same as for the one-sample test). However, we also assume that there are no cross-sample dependencies. If, for instance, it turns out that you included some participants in both experimental conditions of your study (e.g., by accidentally allowing the same person to sign up to different conditions), then there are some cross sample dependencies that you'd need to take into account.
- *Homogeneity of variance* (also called “homoscedasticity”). The third assumption is that the population standard deviation is the same in both groups. You can test this assumption using the Levene test, which I'll talk about later on in the book (Section ??). However, there's a very simple remedy for this assumption if you are worried, which I'll talk about in the next section.

9.4

---

## The independent samples $t$ -test (Welch test)

The biggest problem with using the Student test in practice is the third assumption listed in the previous section. It assumes that both groups have the same standard deviation. This is rarely true in real life. If two samples don't have the same means, why should we expect them to have the same standard deviation? There's really no reason to expect this assumption to be true. We'll talk a little bit about how you can check this assumption later on because it does crop up in a few different places, not just the  $t$ -test. But right now I'll talk about a different form of the  $t$ -test (**Welch1947**) that does not rely on this assumption. A graphical illustration of what the **Welch**

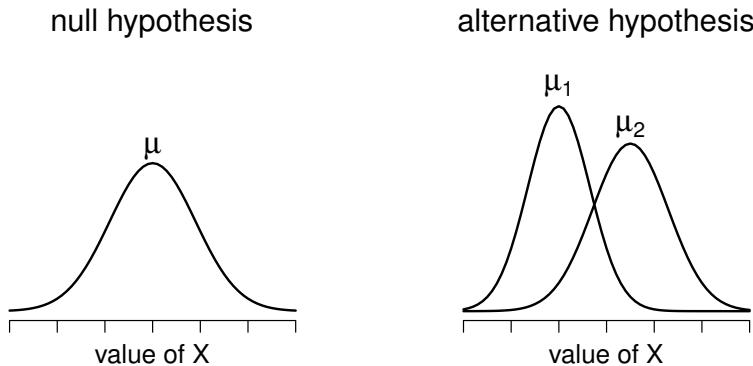


Figure 9.11 Graphical illustration of the null and alternative hypotheses assumed by the Welch  $t$ -test. Like the Student test (Figure ??) we assume that both samples are drawn from a normal population; but the alternative hypothesis no longer requires the two populations to have equal variance.

**$t$  test** assumes about the data is shown in Figure ??, to provide a contrast with the Student test version in Figure ??.

I'll admit it's a bit odd to talk about the cure before talking about the diagnosis, but as it happens the Welch test can be specified as one of the 'Independent Samples T-Test' options in JASP, so this is probably the best place to discuss it.

The Welch test is very similar to the Student test. For example, the  $t$ -statistic that we use in the Welch test is calculated in much the same way as it is for the Student test. That is, we take the difference between the sample means and then divide it by some estimate of the standard error of that difference

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\text{se}(\bar{X}_1 - \bar{X}_2)}$$

The main difference is that the standard error calculations are different. If the two populations have different standard deviations, then it's a complete nonsense to try to calculate a pooled standard deviation estimate, because you're averaging apples and oranges.\*8

---

\*8Well, I guess you can average apples and oranges, and what you end up with is a delicious fruit smoothie. But no one really thinks that a fruit smoothie is a very good way to describe the original fruits, do they?

But you can still estimate the standard error of the difference between sample means, it just ends up looking different

$$se(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

The reason why it's calculated this way is beyond the scope of this book. What matters for our purposes is that the *t*-statistic that comes out of the Welch *t*-test is actually somewhat different to the one that comes from the Student *t*-test.

The second difference between Welch and Student is that the degrees of freedom are calculated in a very different way. In the Welch test, the “degrees of freedom” doesn’t have to be a whole number any more, and it doesn’t correspond all that closely to the “number of data points minus the number of constraints” heuristic that I’ve been using up to this point.

The degrees of freedom are, in fact

$$df = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{(\hat{\sigma}_1^2/N_1)^2/(N_1 - 1) + (\hat{\sigma}_2^2/N_2)^2/(N_2 - 1)}$$

which is all pretty straightforward and obvious, right? Well, perhaps not. It doesn’t really matter for our purposes. What matters is that you’ll see that the “df” value that pops out of a Welch test tends to be a little bit smaller than the one used for the Student test, and it doesn’t have to be a whole number.

#### 9.4.1 Doing the Welch test in JASP

If you tick the check box for the Welch test in the analysis we did above, then this is what it gives you (Figure ??):

##### Independent Samples T-Test ▾

Independent Samples T-Test ▾									
	Test	Statistic	df	p	Mean Difference	SE Difference	95% CI for Mean Difference		
grade	Student	2.115	31.000	0.043	5.478	2.589	0.197	10.759	0.740
	Welch	2.034	23.025	0.054	5.478	2.693	-0.092	11.048	0.724

Figure9.12 Results showing the Welch test alongside the default Student’s t-test in JASP

The interpretation of this output should be fairly obvious. You read the output for the Welch's test in the same way that you would for the Student's test. You've got your descriptive statistics, the test results and some other information. So that's all pretty easy.

Except, except...our result isn't significant anymore. When we ran the Student test we did get a significant effect, but the Welch test on the same data set is not ( $t(23.02) = 2.03, p = .054$ ). What does this mean? Should we panic? Is the sky burning? Probably not. The fact that one test is significant and the other isn't doesn't itself mean very much, especially since I kind of rigged the data so that this would happen. As a general rule, it's not a good idea to go out of your way to try to interpret or explain the difference between a  $p$ -value of .049 and a  $p$ -value of .051. If this sort of thing happens in real life, the *difference* in these  $p$ -values is almost certainly due to chance. What does matter is that you take a little bit of care in thinking about what test you use. The Student test and the Welch test have different strengths and weaknesses. If the two populations really do have equal variances, then the Student test is slightly more powerful (lower Type II error rate) than the Welch test. However, if they *don't* have the same variances, then the assumptions of the Student test are violated and you may not be able to trust it; you might end up with a higher Type I error rate. So it's a trade off. However, in real life I tend to prefer the Welch test, because almost no-one *actually* believes that the population variances are identical.

#### 9.4.2 Assumptions of the test

The assumptions of the Welch test are very similar to those made by the Student  $t$ -test (see Section ??), except that the Welch test does not assume homogeneity of variance. This leaves only the assumption of normality and the assumption of independence. The specifics of these assumptions are the same for the Welch test as for the Student test.

9.5 \_\_\_\_\_

### The paired-samples $t$ -test

Regardless of whether we're talking about the Student test or the Welch test, an independent samples  $t$ -test is intended to be used in a situation where you have two samples that are, well, independent of one another. This situation arises naturally when participants are assigned randomly to one of two experimental conditions, but it provides a very poor approximation to other sorts of research designs. In particular, a repeated measures design, in which each participant is measured (with respect to the same outcome variable) in both experimental conditions, is not suited for

analysis using independent samples *t*-tests. For example, we might be interested in whether listening to music reduces people's working memory capacity. To that end, we could measure each person's working memory capacity in two conditions: with music, and without music. In an experimental design such as this one, each participant appears in *both* groups. This requires us to approach the problem in a different way, by using the **paired samples *t*-test**.

#### 9.5.1 The data

The data set that we'll use this time comes from Dr Chico's class.<sup>\*9</sup> In her class students take two major tests, one early in the semester and one later in the semester. To hear her tell it, she runs a very hard class, one that most students find very challenging. But she argues that by setting hard assessments students are encouraged to work harder. Her theory is that the first test is a bit of a "wake up call" for students. When they realise how hard her class really is, they'll work harder for the second test and get a better mark. Is she right? To test this, let's import the `chico.csv` file into JASP. The `chico` data set contains three variables: an `id` variable that identifies each student in the class, the `grade_test1` variable that records the student grade for the first test, and the `grade_test2` variable that has the grades for the second test.

If we look at the JASP spreadsheet it does seem like the class is a hard one (most grades are between 50% and 60%), but it does look like there's an improvement from the first test to the second one.

If we take a quick look at the descriptive statistics, in Figure ??, we see that this impression seems to be supported. Across all 20 students the mean grade for the first test is 57%, but this rises to 58% for the second test. Although, given that the standard deviations are 6.6% and 6.4% respectively, it's starting to feel like maybe the improvement is just illusory; maybe just random variation. This impression is reinforced when you see the means and confidence intervals plotted in Figure ??a. If we were to rely on this plot alone, looking at how wide those confidence intervals are, we'd be tempted to think that the apparent improvement in student performance is pure chance.

Nevertheless, this impression is wrong. To see why, take a look at the scatterplot of the grades for test 1 against the grades for test 2, shown in Figure ??b. In this plot each dot corresponds to the two grades for a given student. If their grade for test 1 (*x* co-ordinate) equals their grade for test 2 (*y* co-ordinate), then the dot falls on the line. Points falling above the line are the

---

<sup>\*9</sup>At this point we have Drs Harpo, Chico and Zeppo. No prizes for guessing who Dr Groucho is.

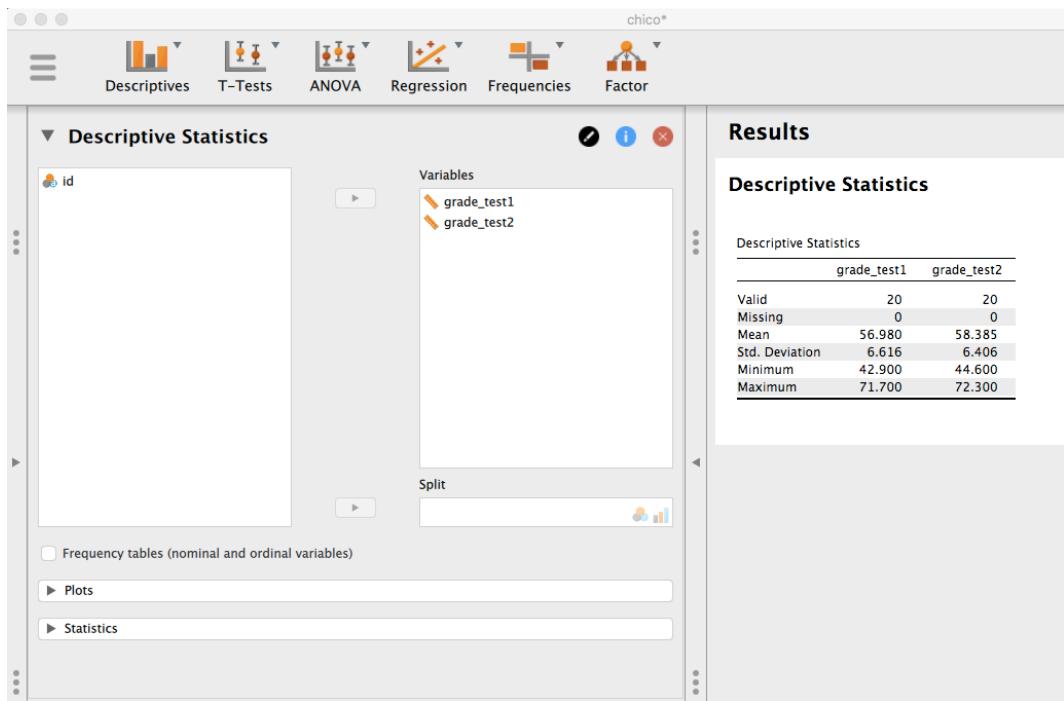


Figure9.13 Descriptives for the two grade\_test variables in the chico data set

.....

students that performed better on the second test. Critically, almost all of the data points fall above the diagonal line: almost all of the students *do* seem to have improved their grade, if only by a small amount. This suggests that we should be looking at the *improvement* made by each student from one test to the next and treating that as our raw data. To do this, we'll need to create a new variable for the *improvement* that each student makes, and add it to the `chico` data set. The easiest way to do this is to compute a new variable. In JASP, click on the “+” at the right-most side of the data columns, name the variable *improvement*, and select the “R” button. After you click the ‘Create column’ button, you can enter the R code `grade_test2 - grade_test1` (see Figure ??).

Once we have computed this new *improvement* variable we can draw a histogram showing the distribution of these improvement scores, shown in Figure ??c. When we look at the histogram, it's very clear that there *is* a real improvement here. The vast majority of the students scored higher on test 2 than on test 1, reflected in the fact that almost the entire histogram is above zero.

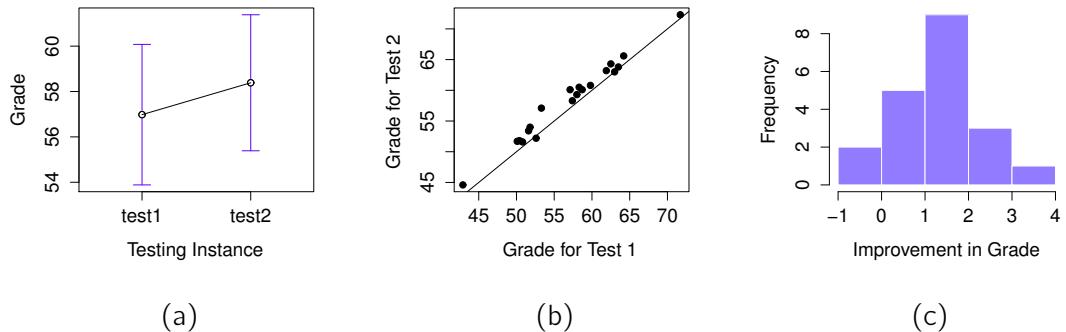


Figure 9.14 Mean grade for test 1 and test 2, with associated 95% confidence intervals (panel a). Scatterplot showing the individual grades for test 1 and test 2 (panel b). Histogram showing the improvement made by each student in Dr Chico's class (panel c). In panel c, notice that almost the entire distribution is above zero: the vast majority of students did improve their performance from the first test to the second one

### 9.5.2 What is the paired samples $t$ -test?

In light of the previous exploration, let's think about how to construct an appropriate  $t$  test. One possibility would be to try to run an independent samples  $t$ -test using `grade_test1` and `grade_test2` as the variables of interest. However, this is clearly the wrong thing to do as the independent samples  $t$ -test assumes that there is no particular relationship between the two samples. Yet clearly that's not true in this case because of the repeated measures structure in the data. To use the language that I introduced in the last section, if we were to try to do an independent samples  $t$ -test, we would be conflating the **within subject** differences (which is what we're interested in testing) with the **between subject** variability (which we are not).

The solution to the problem is obvious, I hope, since we already did all the hard work in the previous section. Instead of running an independent samples  $t$ -test on `grade_test1` and `grade_test2`, we run a *one-sample*  $t$ -test on the within-subject difference variable, `improvement`. To formalise this slightly, if  $X_{i1}$  is the score that the  $i$ -th participant obtained on the first variable, and  $X_{i2}$  is the score that the same person obtained on the second one, then the difference score is:

$$D_i = X_{i1} - X_{i2}$$

Notice that the difference scores is *variable 1 minus variable 2* and not the other way around, so if we want improvement to correspond to a positive valued difference, we actually want "test 2" to be our "variable 1". Equally, we would say that  $\mu_D = \mu_1 - \mu_2$  is the population mean for this

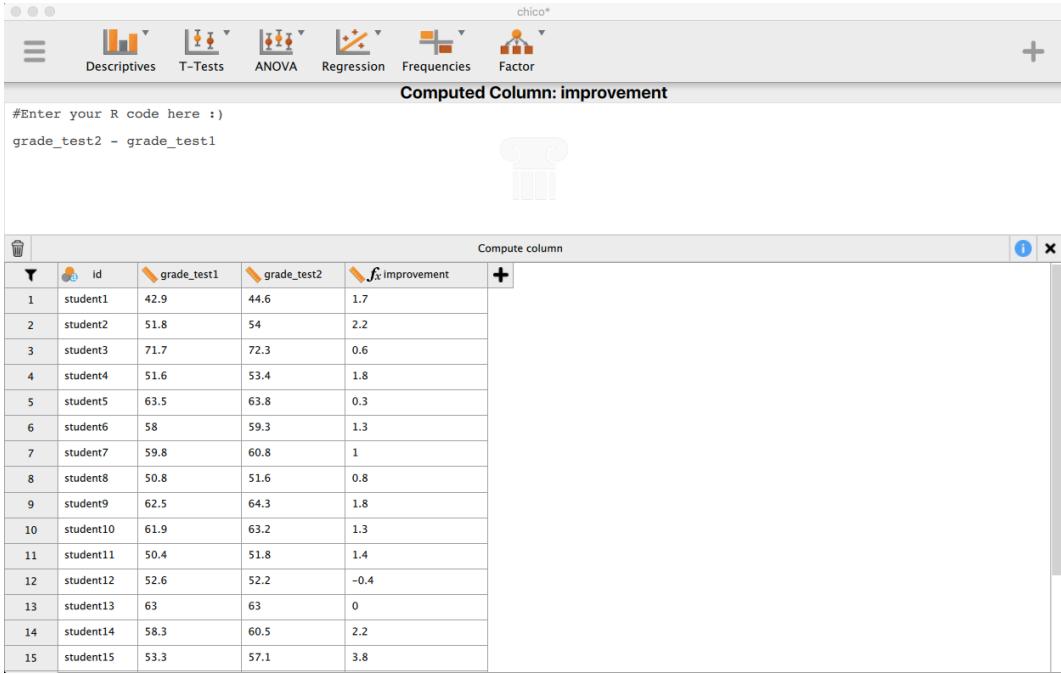


Figure9.15 Using R code to compute an improvement score in JASP.

.....

difference variable. So, to convert this to a hypothesis test, our null hypothesis is that this mean difference is zero and the alternative hypothesis is that it is not

$$\begin{aligned} H_0 : \mu_D &= 0 \\ H_1 : \mu_D &\neq 0 \end{aligned}$$

This is assuming we're talking about a two-sided test here. This is more or less identical to the way we described the hypotheses for the one-sample  $t$ -test. The only difference is that the specific value that the null hypothesis predicts is 0. And so our  $t$ -statistic is defined in more or less the same way too. If we let  $\bar{D}$  denote the mean of the difference scores, then

$$t = \frac{\bar{D}}{\text{se}(\bar{D})}$$

which is

$$t = \frac{\bar{D}}{\hat{\sigma}_D / \sqrt{N}}$$

where  $\hat{\sigma}_D$  is the standard deviation of the difference scores. Since this is just an ordinary, one-sample  $t$ -test, with nothing special about it, the degrees of freedom are still  $N - 1$ . And that's it. The paired samples  $t$ -test really isn't a new test at all. It's a one-sample  $t$ -test, but applied to the difference between two variables. It's actually very simple. The only reason it merits a

discussion as long as the one we've just gone through is that you need to be able to recognise *when* a paired samples test is appropriate, and to understand *why* it's better than an independent samples *t* test.

### 9.5.3 Doing the test in JASP

How do you do a paired samples *t*-test in JASP? One possibility is to follow the process I outlined above. That is, create a “difference” variable and then run a one sample *t*-test on that. Since we've already created a variable called `improvement`, let's do that and see what we get, Figure ??.

#### One Sample T-Test

One Sample T-Test							
	t	df	p	Mean Difference	95% CI for Mean Difference		Cohen's d
					Lower	Upper	
improvement	6.475	19	< .001	1.405	0.951	1.859	1.448

*Note.* Student's t-test.  
*Note.* For the Student t-test, location parameter is given by mean difference *d*.  
*Note.* For the Student t-test, effect size is given by Cohen's *d*.

Figure9.16 Results showing a one sample *t*-test on paired difference scores

.....

The output shown in Figure ?? is (obviously) formatted exactly the same was as it was the last time we used the ‘One Sample T-Test’ analysis (Section ??), and it confirms our intuition. There's an average improvement of 1.4 points from test 1 to test 2, and this is significantly different from 0 ( $t(19) = 6.48, p < .001$ ).

However, suppose you're lazy and you don't want to go to all the effort of creating a new variable. Or perhaps you just want to keep the difference between one-sample and paired-samples tests clear in your head. If so, you can use the JASP ‘Paired Samples T-Test’ analysis. As you will see, the numbers are identical to those that come from the one sample test, which of course they have to be given that the paired samples *t*-test is just a one sample test under the hood.

## One sided tests

When introducing the theory of null hypothesis tests, I mentioned that there are some situations when it's appropriate to specify a *one-sided* test (see Section ??). So far all of the *t*-tests have been two-sided tests. For instance, when we specified a one sample *t*-test for the grades in Dr Zeppo's class the null hypothesis was that the true mean was 67.5%. The alternative hypothesis was that the true mean was greater than *or* less than 67.5%. Suppose we were only interested in finding out if the true mean is greater than 67.5%, and have no interest whatsoever in testing to find out if the true mean is lower than 67.5%. If so, our null hypothesis would be that the true mean is 67.5% or less, and the alternative hypothesis would be that the true mean is greater than 67.5%. In JASP, for the 'One Sample T-Test' analysis, you can specify this by clicking on the '> Test Value' option, under 'Alt. Hypothesis'. When you have done this, you will get the results as shown in ??.

### One Sample T-Test ▾

One Sample T-Test

t	df	p	Mean Difference	95% CI for Mean Difference			Cohen's d
				Lower	Upper	∞	
x 2.255	19	0.018	4.800	1.119	∞	0.504	

Note. Student's t-test.

Note. For the Student t-test, location parameter is given by mean difference *d*.

Note. For the Student t-test, effect size is given by Cohen's *d*.

Note. For all tests, the alternative hypothesis specifies that the mean is greater than 67.5.

Figure9.17 JASP results showing a 'One Sample T-Test' where the actual hypothesis is one sided, i.e. that the true mean is greater than 67.5%

Notice that there are a few changes from the output that we saw last time. Most important is the fact that the actual hypothesis has changed, to reflect the different test. The second thing to note is that although the *t*-statistic and degrees of freedom have not changed, the *p*-value has. This is because the one-sided test has a different rejection region from the two-sided test. If you've forgotten why this is and what it means, you may find it helpful to read back over Chapter ??, and Section ?? in particular. The third thing to note is that the confidence interval is different too: it now reports a "one-sided" confidence interval rather than a two-sided one. In a two-sided confidence interval we're trying to find numbers *a* and *b* such that we're confident that, if we were

to repeat the study many times, then 95% of the time the mean would lie *between*  $a$  and  $b$ . In a one-sided confidence interval, we're trying to find a single number  $a$  such that we're confident that 95% of the time the true mean would be *greater than*  $a$  (or less than  $a$  if you selected ' $<$  Test Value' in the 'Alt. Hypothesis' section).

So that's how to do a one-sided one sample  $t$ -test. However, all versions of the  $t$ -test can be one-sided. For an independent samples  $t$  test, you could have a one-sided test if you're only interested in testing to see if group A has *higher* scores than group B, but have no interest in finding out if group B has higher scores than group A. Let's suppose that, for Dr Harpo's class, you wanted to see if Anastasia's students had higher grades than Bernadette's. For this analysis, in the 'Alt. Hypothesis' options, specify that 'Group 1 > Group2'. You should get the results shown in Figure ??.

### Independent Samples T-Test

Independent Samples T-Test								
	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference		Cohen's d
						Lower	Upper	
grade	2.115	31.000	0.021	5.478	2.589	1.087	$\infty$	0.740

*Note.* Student's t-test.  
*Note.* For all tests, the alternative hypothesis specifies that group *Anastasia* is greater than group *Bernadette*.

Figure9.18 JASP results showing an 'Independent Samples T-Test' where the actual hypothesis is one sided, i.e. that Anastasia's students had higher grades than Bernadette's

.....

Again, the output changes in a predictable way. The definition of the alternative hypothesis has changed, the  $p$ -value has changed, and it now reports a one-sided confidence interval rather than a two-sided one.

What about the paired samples  $t$ -test? Suppose we wanted to test the hypothesis that grades go *up* from test 1 to test 2 in Dr Chico's class, and are not prepared to consider the idea that the grades go down. In JASP you would do this by specifying, under the 'Alt. Hypotheses' option, that `grade_test2` ('Measure 1' in JASP, because we copied this first into the paired variables box)  $>$  `grade_test1` ('Measure 2' in JASP). You should get the results shown in Figure ??.

Yet again, the output changes in a predictable way. The hypothesis has changed, the  $p$ -value has changed, and the confidence interval is now one-sided.

### Paired Samples T-Test

Paired Samples T-Test

	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference		Cohen's d
						Lower	Upper	
grade_test2 - grade_test1	6.475	19	< .001	1.405	0.217	1.030	$\infty$	1.448

Note. Student's t-test.

Note. All tests, hypothesis is measurement one greater than measurement two.

Figure9.19 JASP results showing a 'Paired Samples T-Test' where the actual hypothesis is one sided, i.e. that grade\_test2 ('Measure 1') > grade\_test1 ('Measure 2')

.....

9.7 \_\_\_\_\_

### Effect size

The most commonly used measure of effect size for a *t*-test is **Cohen's *d*** (**Cohen1988**). It's a very simple measure in principle, with quite a few wrinkles when you start digging into the details. Cohen himself defined it primarily in the context of an independent samples *t*-test, specifically the Student test. In that context, a natural way of defining the effect size is to divide the difference between the means by an estimate of the standard deviation. In other words, we're looking to calculate *something* along the lines of this:

$$d = \frac{(\text{mean 1}) - (\text{mean 2})}{\text{std dev}}$$

and he suggested a rough guide for interpreting *d* in Table ???. You'd think that this would be pretty unambiguous, but it's not. This is largely because Cohen wasn't too specific on what he thought should be used as the measure of the standard deviation (in his defence he was trying to make a broader point in his book, not nitpick about tiny details). As discussed by **McGrath2006**, there are several different versions in common usage, and each author tends to adopt slightly different notation. For the sake of simplicity (as opposed to accuracy), I'll use *d* to refer to any statistic that you calculate from the sample, and use  $\delta$  to refer to a theoretical population effect. Obviously, that does mean that there are several different things all called *d*.

My suspicion is that the only time that you would want Cohen's *d* is when you're running a *t*-test, and JASP has an option to calculate the effect size for all the different flavours of *t*-test it provides.

Table9.1 A (very) rough guide to interpreting Cohen's  $d$ . My personal recommendation is to not use these blindly. The  $d$  statistic has a natural interpretation in and of itself. It re-describes the difference in means as the number of standard deviations that separates those means. So it's generally a good idea to think about what that means in practical terms. In some contexts a "small" effect could be of big practical importance. In other situations a "large" effect may not be all that interesting.

$d$ -value	rough interpretation
about 0.2	"small" effect
about 0.5	"moderate" effect
about 0.8	"large" effect

### 9.7.1 Cohen's $d$ from one sample

The simplest situation to consider is the one corresponding to a one-sample  $t$ -test. In this case, this is the one sample mean  $\bar{X}$  and one (hypothesised) population mean  $\mu_0$  to compare it to. Not only that, there's really only one sensible way to estimate the population standard deviation. We just use our usual estimate  $\hat{\sigma}$ . Therefore, we end up with the following as the only way to calculate  $d$

$$d = \frac{\bar{X} - \mu_0}{\hat{\sigma}}$$

When we look back at the results in Figure ??, the effect size value is Cohen's  $d = 0.504$ . Overall, then, the psychology students in Dr Zeppo's class are achieving grades (mean = 72.3%) that are about 0.5 standard deviations higher than the level that you'd expect (67.5%) if they were performing at the same level as other students. Judged against Cohen's rough guide, this is a moderate effect size.

### 9.7.2 Cohen's $d$ from a Student's $t$ test

The majority of discussions of Cohen's  $d$  focus on a situation that is analogous to Student's independent samples  $t$  test, and it's in this context that the story becomes messier, since there are several different versions of  $d$  that you might want to use in this situation. To understand why there are multiple versions of  $d$ , it helps to take the time to write down a formula that corresponds to the true population effect size  $\delta$ . It's pretty straightforward,

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where, as usual,  $\mu_1$  and  $\mu_2$  are the population means corresponding to group 1 and group 2 respectively, and  $\sigma$  is the standard deviation (the same for both populations). The obvious way

to estimate  $\delta$  is to do exactly the same thing that we did in the  $t$ -test itself, i.e., use the sample means as the top line and a pooled standard deviation estimate for the bottom line

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_p}$$

where  $\hat{\sigma}_p$  is the exact same pooled standard deviation measure that appears in the  $t$ -test. This is the most commonly used version of Cohen's  $d$  when applied to the outcome of a Student  $t$ -test, and is the one provided in JASP. It is sometimes referred to as Hedges'  $g$  statistic (**Hedges1981**).

However, there are other possibilities which I'll briefly describe. Firstly, you may have reason to want to use only one of the two groups as the basis for calculating the standard deviation. This approach (often called Glass'  $\Delta$ , pronounced *delta*) only makes most sense when you have good reason to treat one of the two groups as a purer reflection of "natural variation" than the other. This can happen if, for instance, one of the two groups is a control group. Secondly, recall that in the usual calculation of the pooled standard deviation we divide by  $N - 2$  to correct for the bias in the sample variance. In one version of Cohen's  $d$  this correction is omitted, and instead we divide by  $N$ . This version makes sense primarily when you're trying to calculate the effect size in the sample rather than estimating an effect size in the population. Finally, there is a version based on **Hedges1985**, who point out there is a small bias in the usual (pooled) estimation for Cohen's  $d$ . Thus they introduce a small correction by multiplying the usual value of  $d$  by  $(N - 3)/(N - 2.25)$ .

In any case, ignoring all those variations that you could make use of if you wanted, let's have a look at the default version in JASP. In Figure ?? Cohen's  $d = 0.740$ , indicating that the grade scores for students in Anastasia's class are, on average, 0.74 standard deviations higher than the grade scores for students in Bernadette's class. For a Welch test, the estimated effect size is the same (Figure ??).

### 9.7.3 Cohen's $d$ from a paired-samples test

Finally, what should we do for a paired samples  $t$ -test? In this case, the answer depends on what it is you're trying to do. JASP assumes that you want to measure your effect sizes relative to the distribution of difference scores, and the measure of  $d$  that you calculate is:

$$d = \frac{\bar{D}}{\hat{\sigma}_D}$$

where  $\hat{\sigma}_D$  is the estimate of the standard deviation of the differences. In Figure ?? Cohen's  $d = 1.45$ , indicating that the time 2 grade scores are, on average, 1.45 standard deviations higher than the time 1 grade scores.

This is the version of Cohen's  $d$  that gets reported by the JASP 'Paired Samples T-Test' analysis. The only wrinkle is figuring out whether this is the measure you want or not. To the extent that

you care about the practical consequences of your research, you often want to measure the effect size relative to the *original* variables, not the *difference* scores (e.g., the 1 point improvement in Dr Chico's class over time is pretty small when measured against the amount of between-student variation in grades), in which case you use the same versions of Cohen's  $d$  that you would use for a Student or Welch test. It's not so straightforward to do this in JASP; essentially you have to change the structure of the data in the spreadsheet view so I won't go into that here.

## 9.8 \_\_\_\_\_

### Checking the normality of a sample

All of the tests that we have discussed so far in this chapter have assumed that the data are normally distributed. This assumption is often quite reasonable, because the central limit theorem (Section ??) does tend to ensure that many real world quantities are normally distributed. Any time that you suspect that your variable is *actually* an average of lots of different things, there's a pretty good chance that it will be normally distributed, or at least close enough to normal that you can get away with using  $t$ -tests. However, life doesn't come with guarantees, and besides there are lots of ways in which you can end up with variables that are highly non-normal. For example, any time you think that your variable is actually the minimum of lots of different things, there's a very good chance it will end up quite skewed. In psychology, response time (RT) data is a good example of this. If you suppose that there are lots of things that could trigger a response from a human participant, then the actual response will occur the first time one of these trigger events occurs.<sup>\*10</sup> This means that RT data are systematically non-normal. Okay, so if normality is assumed by all the tests, and is mostly but not always satisfied (at least approximately) by real world data, how can we check the normality of a sample? In this section I discuss two methods: QQ plots and the Shapiro-Wilk test.

#### 9.8.1 QQ plots

One way to check whether a sample violates the normality assumption is to draw a “**QQ plot**” (Quantile-Quantile plot). This allows you to visually check whether you're seeing any systematic violations. In a QQ plot, each observation is plotted as a single dot. The x co-ordinate is the theoretical quantile that the observation should fall in if the data were normally distributed (with mean and variance estimated from the sample), and on the y co-ordinate is the actual quantile

---

<sup>\*10</sup>This is a massive oversimplification.

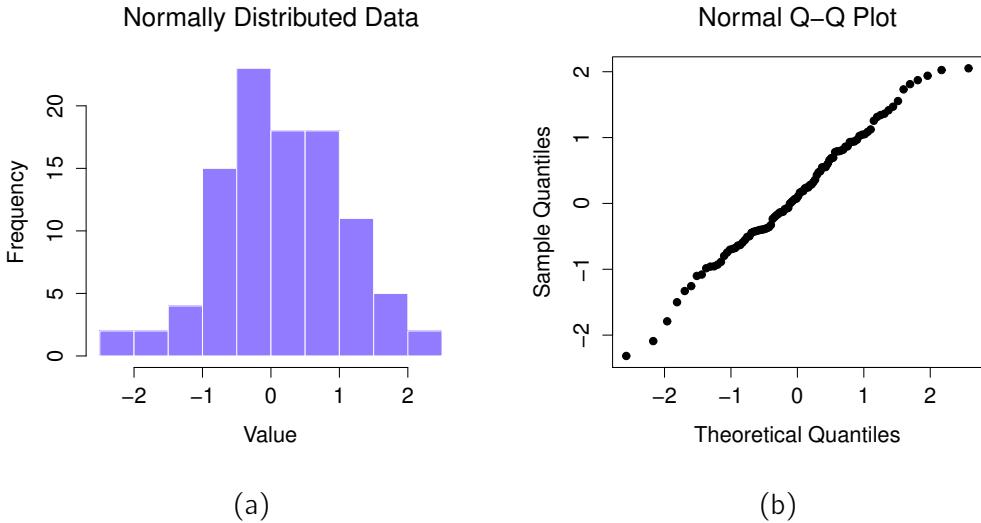


Figure 9.20 Histogram (panel a) and normal QQ plot (panel b) of `normal.data`, a normally distributed sample with 100 observations. The Shapiro-Wilk statistic associated with these data is  $W = .99$ , indicating that no significant departures from normality were detected ( $p = .73$ ).

---

of the data within the sample. If the data are normal, the dots should form a straight line. For instance, let's see what happens if we generate data by sampling from a normal distribution, and then drawing a QQ plot. The results are shown in Figure ???. As you can see, these data form a pretty straight line; which is no surprise given that we sampled them from a normal distribution! In contrast, have a look at the two data sets shown in Figure ???. The top panels show the histogram and a QQ plot for a data set that is highly skewed: the QQ plot curves upwards. The lower panels show the same plots for a heavy tailed (i.e., high kurtosis) data set: in this case the QQ plot flattens in the middle and curves sharply at either end.

### 9.8.2 Shapiro-Wilk tests

QQ plots provide a nice way to informally check the normality of your data, but sometimes you'll want to do something a bit more formal and the **Shapiro-Wilk test** ([Shapiro1965](#)) is probably

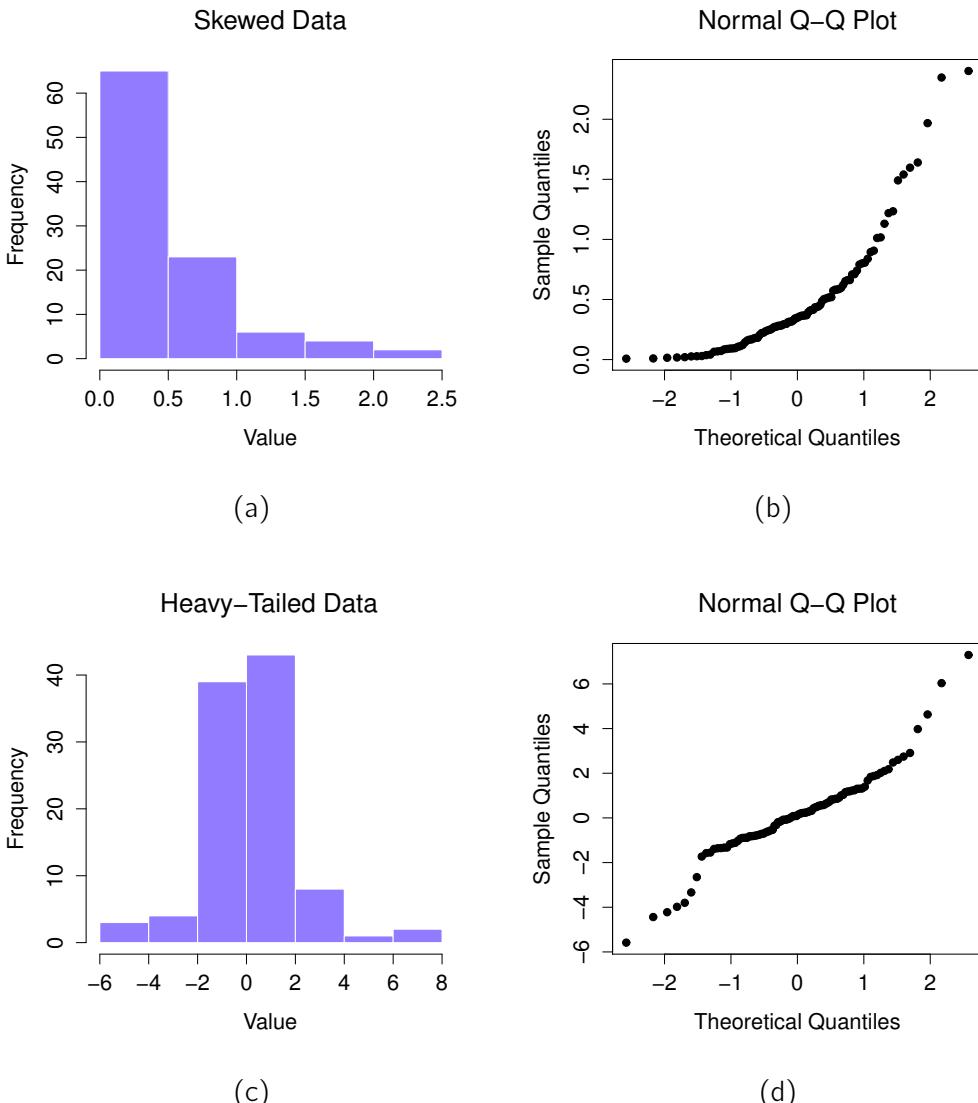


Figure 9.21 In the top row, a histogram (panel a) and normal QQ plot (panel b) of the 100 observations in a skewed data set. The skewness of the data here is 1.94, and is reflected in a QQ plot that curves upwards. As a consequence, the Shapiro-Wilk statistic is  $W = .80$ , reflecting a significant departure from normality ( $p < .001$ ). The bottom row shows the same plots for a heavy tailed data set, again consisting of 100 observations. In this case the heavy tails in the data produce a high kurtosis (2.80), and cause the QQ plot to flatten in the middle, and curve away sharply on either side. The resulting Shapiro-Wilk statistic is  $W = .93$ , again reflecting significant non-normality ( $p < .001$ ).

what you're looking for.<sup>\*11</sup> As you'd expect, the null hypothesis being tested is that a set of  $N$  observations is normally distributed.

The test statistic that it calculates is conventionally denoted as  $W$ , and it's calculated as follows. First, we sort the observations in order of increasing size, and let  $X_1$  be the smallest value in the sample,  $X_2$  be the second smallest and so on. Then the value of  $W$  is given by

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

where  $\bar{X}$  is the mean of the observations, and the  $a_i$  values are ... mumble, mumble ... something complicated that is a bit beyond the scope of an introductory text.

Because it's a little hard to explain the maths behind the  $W$  statistic, a better idea is to give a broad brush description of how it behaves. Unlike most of the test statistics that we'll encounter in this book, it's actually *small* values of  $W$  that indicate departure from normality. The  $W$  statistic has a maximum value of 1, which occurs when the data look "perfectly normal". The smaller the value of  $W$  the less normal the data are. However, the sampling distribution for  $W$ , which is not one of the standard ones that I discussed in Chapter ?? and is in fact a complete pain in the arse to work with, does depend on the sample size  $N$ . To give you a feel for what these sampling distributions look like, I've plotted three of them in Figure ?? . Notice that, as the sample size starts to get large, the sampling distribution becomes very tightly clumped up near  $W = 1$ , and as a consequence, for larger samples  $W$  doesn't have to be very much smaller than 1 in order for the test to be significant.

To get the Shapiro-Wilk statistic in JASP  $t$ -tests, check the option for 'Normality' listed under 'Assumption checks'. In the randomly sampled data ( $N = 100$ ) we used for the QQ plot, the value for the Shapiro-Wilk normality test statistic was  $W = 0.99$  with a  $p$ -value of 0.69. So, not surprisingly, we have no evidence that these data depart from normality. When reporting the results for a Shapiro-Wilk test, you should (as usual) make sure to include the test statistic  $W$  and the  $p$  value, though given that the sampling distribution depends so heavily on  $N$  it would probably be a politeness to include  $N$  as well.

<sup>\*11</sup>Either that, or the Kolmogorov-Smirnov test, which is probably more traditional than the Shapiro-Wilk. Although most things I've read seem to suggest Shapiro-Wilk is the better test of normality, the Kolmogorov-Smirnov is a general purpose test of distributional equivalence that can be adapted to handle other kinds of distribution tests. In JASP the Shapiro-Wilk test is preferred.

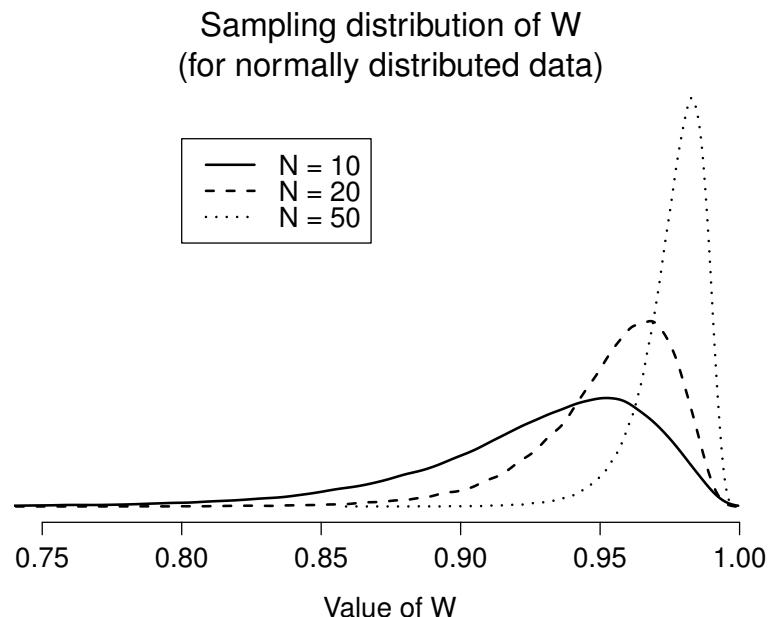
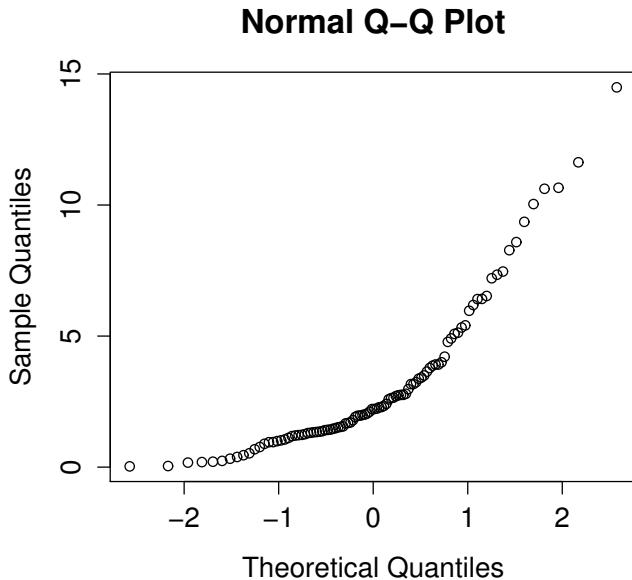


Figure 9.22 Sampling distribution of the Shapiro-Wilk  $W$  statistic, under the null hypothesis that the data are normally distributed, for samples of size 10, 20 and 50. Note that *small* values of  $W$  indicate departure from normality.

### 9.8.3 Example

In the meantime, it's probably worth showing you an example of what happens to the QQ plot and the Shapiro-Wilk test when the data turn out to be non-normal. For that, let's look at the distribution of our AFL winning margins data, which if you remember back to Chapter ?? didn't look like they came from a normal distribution at all. Here's what happens to the QQ plot:



And when we run the Shapiro-Wilk test on the AFL margins data, we get a value for the Shapiro-Wilk normality test statistic of  $W = 0.94$ , and  $p\text{-value} = 9.481\text{e-}07$ . This is clearly a significant departure from normality!

9.9

---

### Testing non-normal data with Wilcoxon tests

Okay, suppose your data turn out to be pretty substantially non-normal, but you still want to run something like a  $t$ -test? This situation occurs a lot in real life. For the AFL winning margins data, for instance, the Shapiro-Wilk test made it very clear that the normality assumption is violated. This is the situation where you want to use Wilcoxon tests.

Like the  $t$ -test, the Wilcoxon test comes in two forms, one-sample and two-sample, and they're used in more or less the exact same situations as the corresponding  $t$ -tests. Unlike the  $t$ -test, the Wilcoxon test doesn't assume normality, which is nice. In fact, they don't make any assumptions about what kind of distribution is involved. In statistical jargon, this makes them **nonparametric tests**. While avoiding the normality assumption is nice, there's a drawback: the Wilcoxon test is usually less powerful than the  $t$ -test (i.e., higher Type II error rate). I won't discuss the Wilcoxon tests in as much detail as the  $t$ -tests, but I'll give you a brief overview.

### 9.9.1 Two sample Mann-Whitney U test

I'll start by describing the **Mann-Whitney U test**, since it's actually simpler than the one sample version. Suppose we're looking at the scores of 10 people on some test. Since my imagination has now failed me completely, let's pretend it's a "test of awesomeness" and there are two groups of people, "A" and "B". I'm curious to know which group is more awesome. The data are included in the file `awesome.csv`, and there are two variables apart from the usual `ID` variable: `scores` and `group`.

As long as there are no ties (i.e., people with the exact same awesomeness score) then the test that we want to do is surprisingly simple. All we have to do is construct a table that compares every observation in group A against every observation in group B. Whenever the group A datum is larger, we place a check mark in the table:

		group B				
		14.5	10.4	12.4	11.7	13.0
group A	6.4	.	.	.	.	.
	10.7	.	✓	.	.	.
	11.9	.	✓	.	✓	.
	7.3	.	.	.	.	.
	10.0	.	.	.	.	.

We then count up the number of checkmarks. This is our test statistic,  $W$ .<sup>\*12</sup> The actual sampling distribution for  $W$  is somewhat complicated, and I'll skip the details. For our purposes, it's sufficient to note that the interpretation of  $W$  is qualitatively the same as the interpretation of  $t$  or  $z$ . That is, if we want a two-sided test then we reject the null hypothesis when  $W$  is very large or very small, but if we have a directional (i.e., one-sided) hypothesis then we only use one or the other.

In JASP, if we run an 'Independent Samples T-Test' with `scores` as the dependent variable, and `group` as the grouping variable, and then under the options for 'tests' check the option for 'Mann-Whitney', we will get results showing that  $U = 3$  (i.e., the same number of checkmarks as shown above), and a p-value = 0.05556.

---

<sup>\*12</sup>Actually, there are two different versions of the test statistic that differ from each other by a constant value. The version that I've described is the one that JASP calculates.

### 9.9.2 One sample Wilcoxon test

What about the **one sample Wilcoxon test** (or equivalently, the paired samples Wilcoxon test)? Suppose I'm interested in finding out whether taking a statistics class has any effect on the happiness of students. My data is in the [happiness.csv](#) file. What I've measured here is the happiness of each student *before* taking the class and *after* taking the class, and the *change* score is the difference between the two. Just like we saw with the *t*-test, there's no fundamental difference between doing a paired-samples test using *before* and *after*, versus doing a one-sample test using the *change* scores. As before, the simplest way to think about the test is to construct a tabulation. The way to do it this time is to take those change scores that are positive differences, and tabulate them against all the complete sample. What you end up with is a table that looks like this:

		all differences									
		-24	-14	-10	7	-6	-38	2	-35	-30	5
positive differences	7	.	.	.	✓	✓	.	✓	.	.	✓
	2	.	.	.	.	.	.	✓	.	.	.
	5	.	.	.	.	.	.	✓	.	.	✓

Counting up the tick marks this time we get a test statistic of  $W = 7$ . As before, if our test is two sided, then we reject the null hypothesis when  $W$  is very large or very small. As far as running it in JASP goes, it's pretty much what you'd expect. For the one-sample version, you specify the 'Wilcoxon signed-rank' option under 'Tests' in the 'One Sample T-Test' analysis window. This gives you Wilcoxon  $W = 7$ ,  $p\text{-value} = 0.037$ . As this shows, we have a significant effect. Evidently, taking a statistics class does have an effect on your happiness. Switching to a paired samples version of the test won't give us a different answer, of course; see Figure ??.

## 9.10 \_\_\_\_\_

### Summary

- A one sample *t*-test is used to compare a single sample mean against a hypothesised value for the population mean. (Section ??)
- An independent samples *t*-test is used to compare the means of two groups, and tests the null hypothesis that they have the same mean. It comes in two forms: the Student test (Section ??) assumes that the groups have the same standard deviation, the Welch test

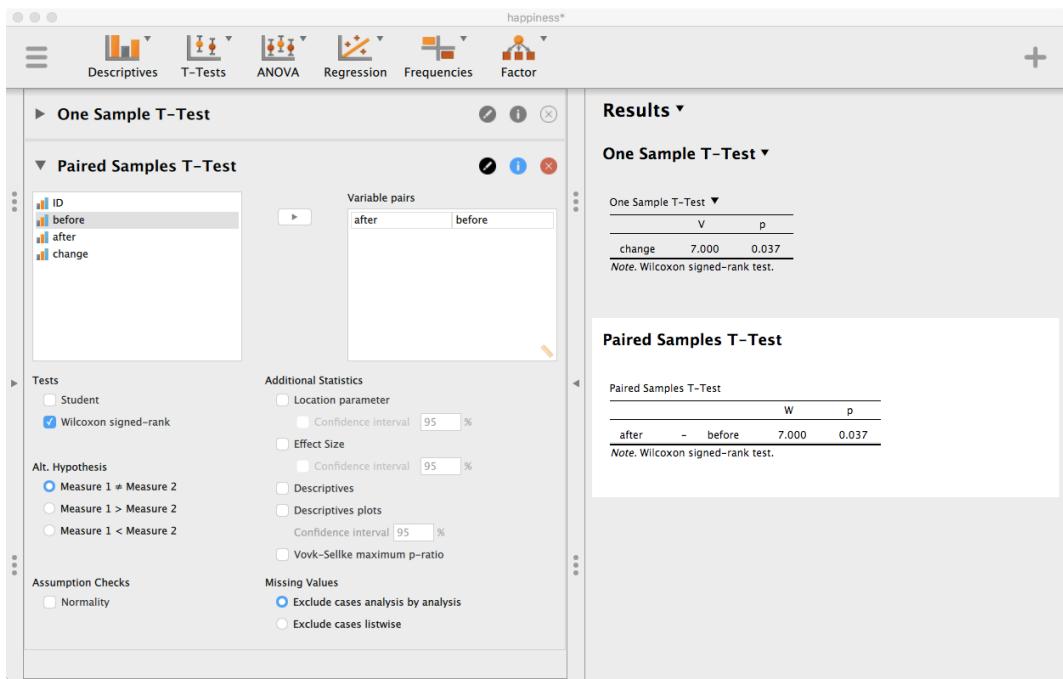


Figure 9.23 JASP screen showing results for one sample and paired sample Wilcoxon non-parametric tests

(Section ??) does not.

- A paired samples  $t$ -test is used when you have two scores from each person, and you want to test the null hypothesis that the two scores have the same mean. It is equivalent to taking the difference between the two scores for each person, and then running a one sample  $t$ -test on the difference scores. (Section ??)
- One sided tests are perfectly legitimate as long as they are pre-planned. (Section ??)
- Effect size calculations for the difference between means can be calculated via the Cohen's  $d$  statistic. (Section ??).
- You can check the normality of a sample using QQ plots (not currently available in JASP) and the Shapiro-Wilk test. (Section ??)
- If your data are non-normal, you can use Mann-Whitney or Wilcoxon tests instead of  $t$ -tests. (Section ??)



## 10. Correlation and linear regression

---

The goal in this chapter is to introduce **correlation** and **linear regression**. These are the standard tools that statisticians rely on when analysing the relationship between continuous predictors and continuous outcomes.

### 10.1

---

#### Correlations

In this section we'll talk about how to describe the relationships *between* variables in the data. To do that, we want to talk mostly about the **correlation** between variables. But first, we need some data.

##### 10.1.1 The data

Table 10.1 Descriptive statistics for the parenthood data.

variable	min	max	mean	median	std. dev	IQR
Dan's grumpiness	41	91	63.71	62	10.05	14
Dan's hours slept	4.84	9.00	6.97	7.03	1.02	1.45
Dan's son's hours slept	3.25	12.07	8.05	7.95	2.07	3.21

---

Let's turn to a topic close to every parent's heart: sleep. The data set we'll use is fictitious, but based on real events. Suppose I'm curious to find out how much my infant son's sleeping habits affect my mood. Let's say that I can rate my grumpiness very precisely, on a scale from 0 (not at

all grumpy) to 100 (grumpy as a very, very grumpy old man or woman). And lets also assume that I've been measuring my grumpiness, my sleeping patterns and my son's sleeping patterns for quite some time now. Let's say, for 100 days. And, being a nerd, I've saved the data as a file called `parenthood.csv`. If we load the data into JASP we can see that the file contains four variables `dan.sleep`, `baby.sleep`, `dan.grump` and `day`. Note that when you first load this data set JASP may not have guessed the data type for each variable correctly, in which case you should fix it: `dan.sleep`, `baby.sleep`, `dan.grump` and `day` can be specified as continuous variables, and `ID` is a nominal(integer) variable.

Next, I'll take a look at some basic descriptive statistics and, to give a graphical depiction of what each of the three interesting variables looks like, Figure ?? plots histograms. One thing to note: just because JASP can calculate dozens of different statistics doesn't mean you should report all of them. If I were writing this up for a report, I'd probably pick out those statistics that are of most interest to me (and to my readership), and then put them into a nice, simple table like the one in Table ??.\*<sup>1</sup> Notice that when I put it into a table, I gave everything "human readable" names. This is always good practice. Notice also that I'm not getting enough sleep. This isn't good practice, but other parents tell me that it's pretty standard.

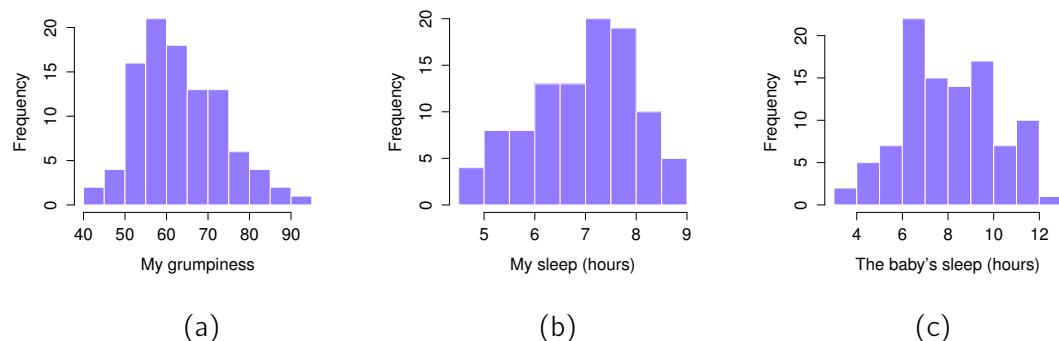


Figure10.1 Histograms for the three interesting variables in the `parenthood` data set.

### 10.1.2 The strength and direction of a relationship

We can draw scatterplots to give us a general sense of how closely related two variables are.

---

\*<sup>1</sup>Actually, even that table is more than I'd bother with. In practice most people pick *one* measure of central tendency, and *one* measure of variability only.

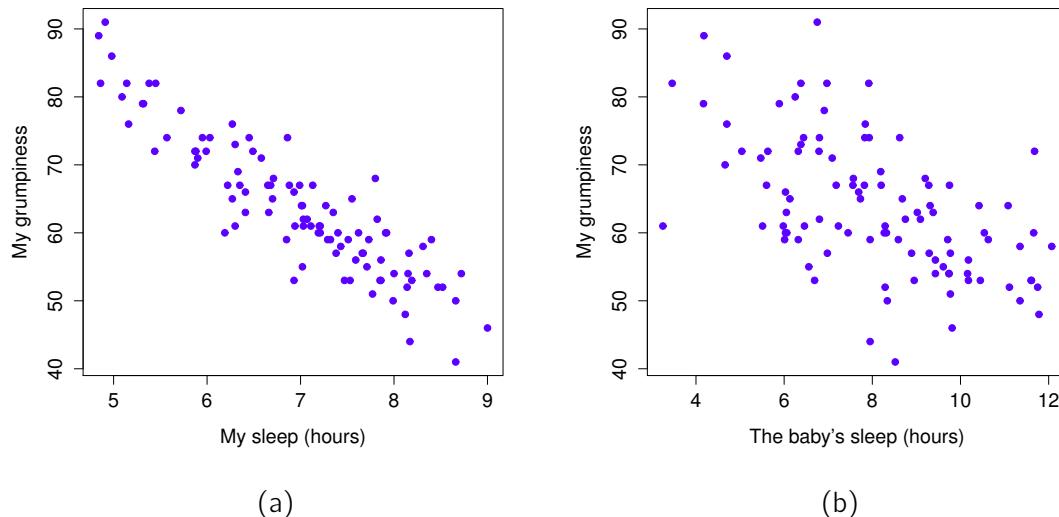


Figure 10.2 Scatterplots showing the relationship between `dan.sleep` and `dan.grump` (left) and the relationship between `baby.sleep` and `dan.grump` (right).

Ideally though, we might want to say a bit more about it than that. For instance, let's compare the relationship between `dan.sleep` and `dan.grump` (Figure ??, left) with that between `baby.sleep` and `dan.grump` (Figure ??, right). When looking at these two plots side by side, it's clear that the relationship is *qualitatively* the same in both cases: more sleep equals less grump! However, it's also pretty obvious that the relationship between `dan.sleep` and `dan.grump` is *stronger* than the relationship between `baby.sleep` and `dan.grump`. The plot on the left is "neater" than the one on the right. What it feels like is that if you want to predict what my mood is, it'd help you a little bit to know how many hours my son slept, but it'd be more helpful to know how many hours I slept.

In contrast, let's consider the two scatterplots shown in Figure ???. If we compare the scatterplot of "`baby.sleep v dan.grump`" (left) to the scatterplot of "`baby.sleep v dan.sleep`" (right), the overall strength of the relationship is the same, but the direction is different. That is, if my son sleeps more, I get *more* sleep (positive relationship, right hand side), but if he sleeps more then I get *less* grumpy (negative relationship, left hand side).

### 10.1.3 The correlation coefficient

We can make these ideas a bit more explicit by introducing the idea of a **correlation coefficient** (or, more specifically, Pearson's correlation coefficient), which is traditionally denoted as  $r$ . The

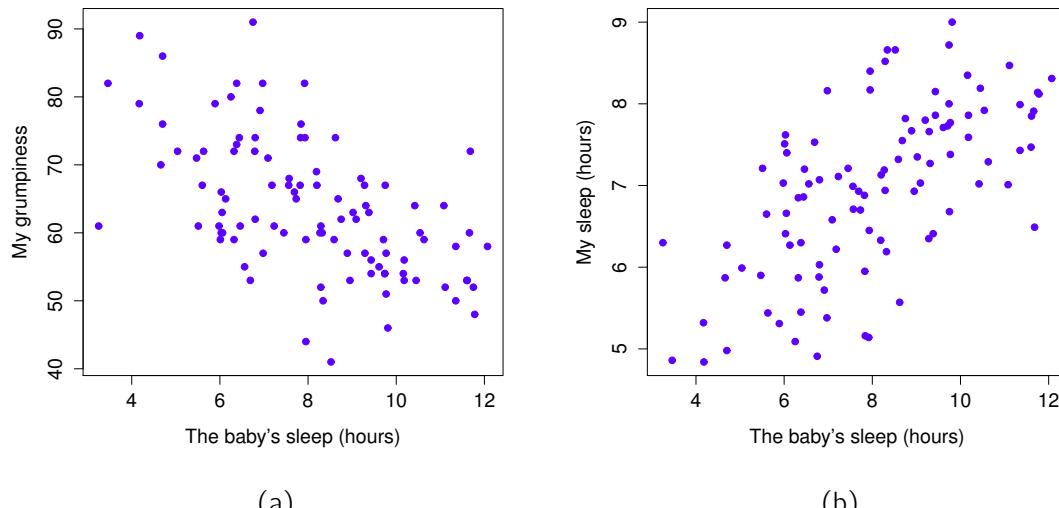


Figure 10.3 Scatterplots showing the relationship between `baby.sleep` and `dan.grump` (left), as compared to the relationship between `baby.sleep` and `dan.sleep` (right).

.....

correlation coefficient between two variables  $X$  and  $Y$  (sometimes denoted  $r_{XY}$ ), which we'll define more precisely in the next section, is a measure that varies from  $-1$  to  $1$ . When  $r = -1$  it means that we have a perfect negative relationship, and when  $r = 1$  it means we have a perfect positive relationship. When  $r = 0$ , there's no relationship at all. If you look at Figure ??, you can see several plots showing what different correlations look like.

The formula for the Pearson's correlation coefficient can be written in several different ways. I think the simplest way to write down the formula is to break it into two steps. Firstly, let's introduce the idea of a **covariance**. The covariance between two variables  $X$  and  $Y$  is a generalisation of the notion of the variance and is a mathematically simple way of describing the

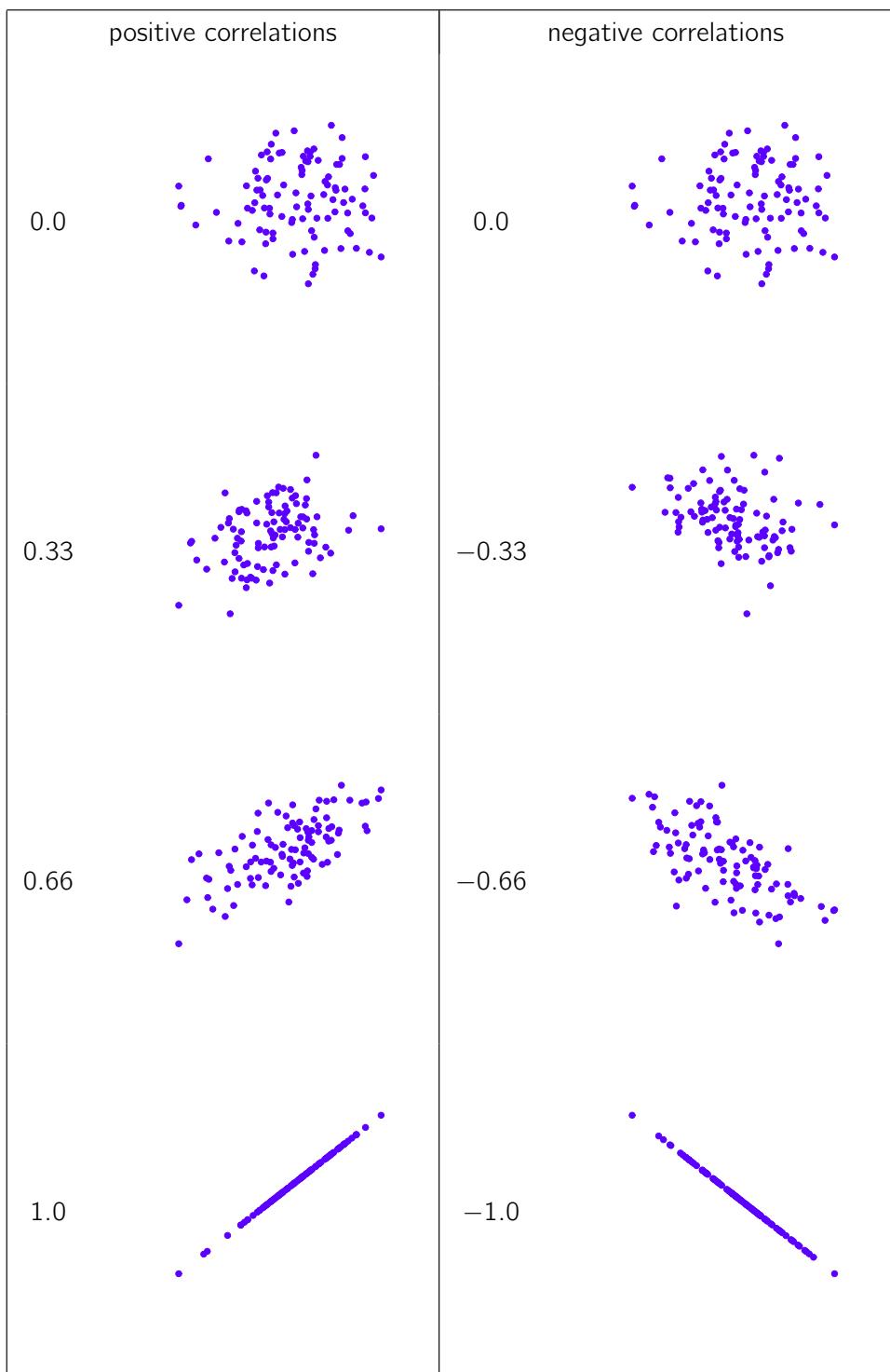


Figure10.4 Illustration of the effect of varying the strength and direction of a correlation. In the left hand column, the correlations are 0, .33, .66 and 1. In the right hand column, the correlations are 0, -.33, -.66 and -1.

relationship between two variables that isn't terribly informative to humans

$$\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Because we're multiplying (i.e., taking the “product” of) a quantity that depends on  $X$  by a quantity that depends on  $Y$  and then averaging<sup>a</sup>, you can think of the formula for the covariance as an “average cross product” between  $X$  and  $Y$ .

The covariance has the nice property that, if  $X$  and  $Y$  are entirely unrelated, then the covariance is exactly zero. If the relationship between them is positive (in the sense shown in Figure ??) then the covariance is also positive, and if the relationship is negative then the covariance is also negative. In other words, the covariance captures the basic qualitative idea of correlation. Unfortunately, the raw magnitude of the covariance isn't easy to interpret as it depends on the units in which  $X$  and  $Y$  are expressed and, worse yet, the actual units that the covariance itself is expressed in are really weird. For instance, if  $X$  refers to the `dan.sleep` variable (units: hours) and  $Y$  refers to the `dan.grump` variable (units: grumps), then the units for their covariance are “hours  $\times$  grumps”. And I have no freaking idea what that would even mean.

The Pearson correlation coefficient  $r$  fixes this interpretation problem by standardising the covariance, in pretty much the exact same way that the z-score standardises a raw score, by dividing by the standard deviation. However, because we have two variables that contribute to the covariance, the standardisation only works if we divide by both standard deviations.<sup>b</sup> In other words, the correlation between  $X$  and  $Y$  can be written as follows:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

---

<sup>a</sup>Just like we saw with the variance and the standard deviation, in practice we divide by  $N - 1$  rather than  $N$ .

<sup>b</sup>This is an oversimplification, but it'll do for our purposes.

By standardising the covariance, not only do we keep all of the nice properties of the covariance discussed earlier, but the actual values of  $r$  are on a meaningful scale:  $r = 1$  implies a perfect positive relationship and  $r = -1$  implies a perfect negative relationship. I'll expand a little more on this point later, in Section ???. But before I do, let's look at how to calculate correlations in JASP.

#### 10.1.4 Calculating correlations in JASP

Calculating correlations in JASP can be done by clicking on the ‘Regression’ – ‘Correlation Matrix’ button. Transfer all four continuous variables across into the box on the right to get the

output in Figure ???. Notice that each correlation (denoted ‘Pearson’s  $r$ ’) is paired with a  $p$ -value. Clearly, something is being tested here, but ignore it for now. We’ll talk more about that soon!

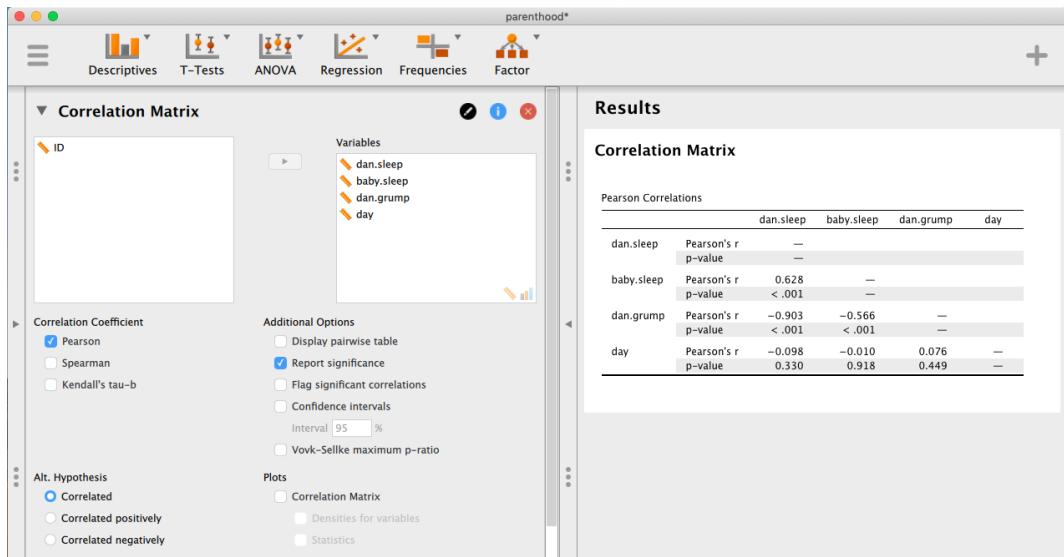


Figure10.5 A JASP screenshot showing correlations between variables in the parenthood.csv file

#### 10.1.5 Interpreting a correlation

Naturally, in real life you don’t see many correlations of 1. So how should you interpret a correlation of, say,  $r = .4$ ? The honest answer is that it really depends on what you want to use the data for, and on how strong the correlations in your field tend to be. A friend of mine in engineering once argued that any correlation less than .95 is completely useless (I think he was exaggerating, even for engineering). On the other hand, there are real cases, even in psychology, where you should really expect correlations that strong. For instance, one of the benchmark data sets used to test theories of how people judge similarities is so clean that any theory that can’t achieve a correlation of at least .9 really isn’t deemed to be successful. However, when looking for (say) elementary correlates of intelligence (e.g., inspection time, response time), if you get a correlation above .3 you’re doing very very well. In short, the interpretation of a correlation depends a lot on the context. That said, the rough guide in Table ?? is pretty typical.

However, something that can never be stressed enough is that you should *always* look at the

Table 10.2 A rough guide to interpreting correlations. Note that I say a *rough* guide. There aren't hard and fast rules for what counts as strong or weak relationships. It depends on the context.

Correlation	Strength	Direction
-1.0 to -0.9	Very strong	Negative
-0.9 to -0.7	Strong	Negative
-0.7 to -0.4	Moderate	Negative
-0.4 to -0.2	Weak	Negative
-0.2 to 0	Negligible	Negative
0 to 0.2	Negligible	Positive
0.2 to 0.4	Weak	Positive
0.4 to 0.7	Moderate	Positive
0.7 to 0.9	Strong	Positive
0.9 to 1.0	Very strong	Positive

.....

scatterplot before attaching any interpretation to the data. A correlation might not mean what you think it means. The classic illustration of this is “Anscombe’s Quartet” (**Anscombe 1973**), a collection of four data sets. Each data set has two variables, an  $X$  and a  $Y$ . For all four data sets the mean value for  $X$  is 9 and the mean for  $Y$  is 7.5. The standard deviations for all  $X$  variables are almost identical, as are those for the  $Y$  variables. And in each case the correlation between  $X$  and  $Y$  is  $r = 0.816$ . You can verify this yourself, since I happen to have saved it in a file called `anscombe.csv`.

You’d think that these four data sets would look pretty similar to one another. They do not. If we draw scatterplots of  $X$  against  $Y$  for all four variables, as shown in Figure ??, we see that all four of these are *spectacularly* different to each other. The lesson here, which so very many people seem to forget in real life, is “*always graph your raw data*” (Chapter ??).

#### 10.1.6 Spearman’s rank correlations

The Pearson correlation coefficient is useful for a lot of things, but it does have shortcomings. One issue in particular stands out: what it actually measures is the strength of the *linear* relationship between two variables. In other words, what it gives you is a measure of the extent to which the data all tend to fall on a single, perfectly straight line. Often, this is a pretty good approximation to what we mean when we say “relationship”, and so the Pearson correlation is a good thing to

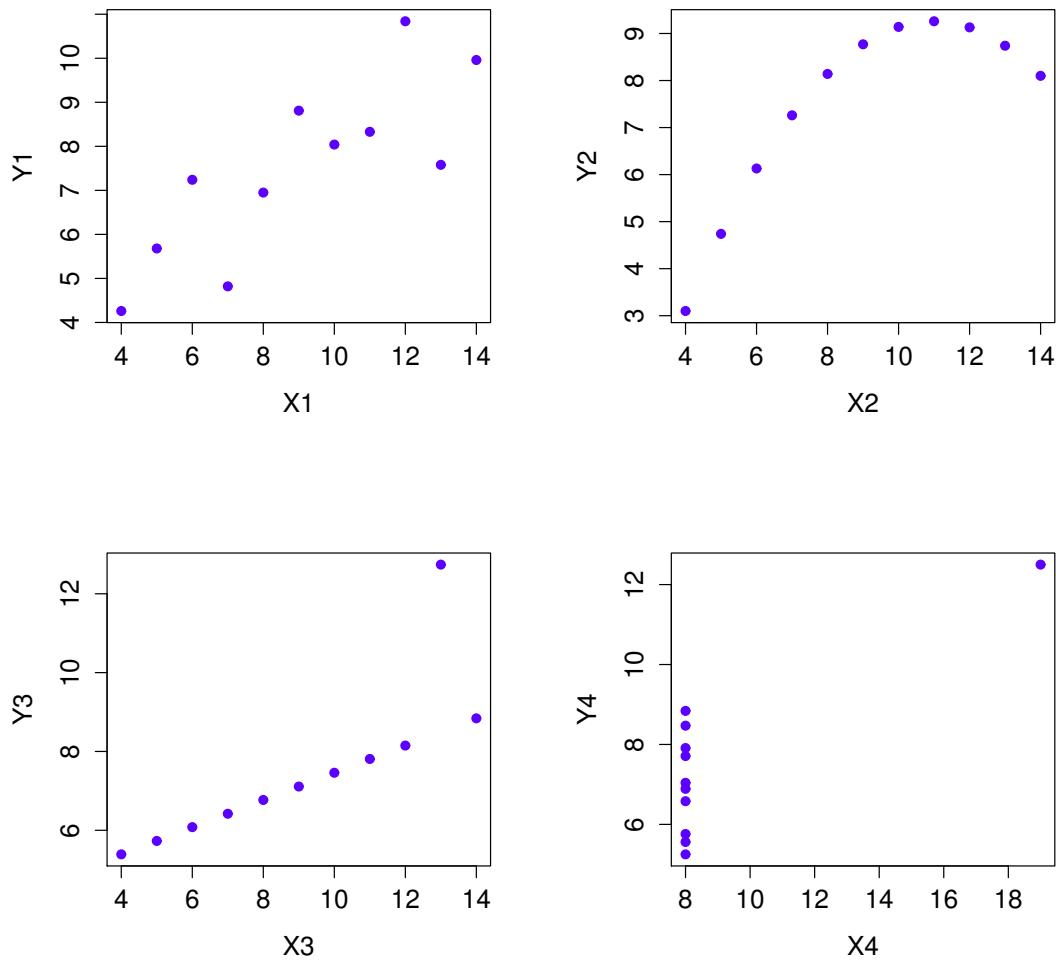


Figure 10.6 Anscombe's quartet. All four of these data sets have a Pearson correlation of  $r = .816$ , but they are qualitatively different from one another.

.....

calculate. Sometimes though, it isn't.

One very common situation where the Pearson correlation isn't quite the right thing to use arises when an increase in one variable  $X$  really is reflected in an increase in another variable  $Y$ , but the nature of the relationship isn't necessarily linear. An example of this might be the relationship between effort and reward when studying for an exam. If you put zero effort ( $X$ ) into learning a subject then you should expect a grade of 0% ( $Y$ ). However, a little bit of effort will cause a massive improvement. Just turning up to lectures means that you learn a fair bit, and if you just

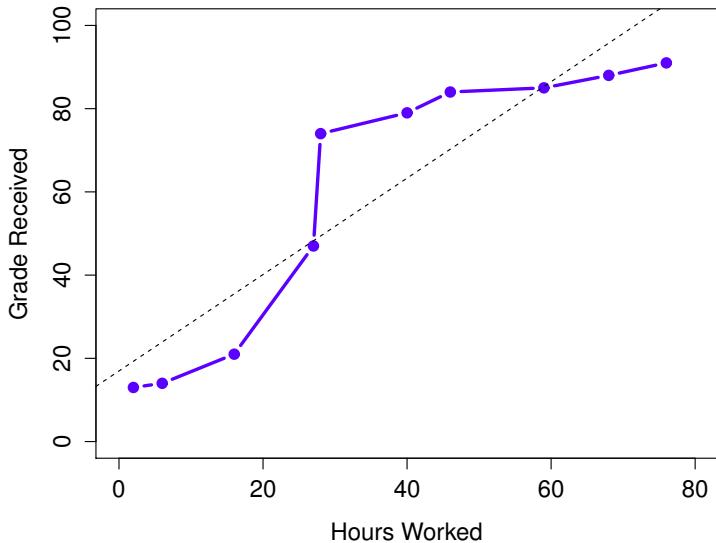


Figure 10.7 The relationship between hours worked and grade received for a toy data set consisting of only 10 students (each circle corresponds to one student). The dashed line through the middle shows the linear relationship between the two variables. This produces a strong Pearson correlation of  $r = .91$ . However, the interesting thing to note here is that there's actually a perfect monotonic relationship between the two variables. In this toy example, increasing the hours worked always increases the grade received, as illustrated by the solid line. This is reflected in a Spearman correlation of  $\rho = 1$ . With such a small data set, however, it's an open question as to which version better describes the actual relationship involved.

turn up to classes and scribble a few things down your grade might rise to 35%, all without a lot of effort. However, you just don't get the same effect at the other end of the scale. As everyone knows, it takes *a lot* more effort to get a grade of 90% than it takes to get a grade of 55%. What this means is that, if I've got data looking at study effort and grades, there's a pretty good chance that Pearson correlations will be misleading.

To illustrate, consider the data plotted in Figure ??, showing the relationship between hours worked and grade received for 10 students taking some class. The curious thing about this (highly fictitious) data set is that increasing your effort *always* increases your grade. It might be by a lot or it might be by a little, but increasing effort will never decrease your grade. If we run a standard Pearson correlation, it shows a strong relationship between hours worked and grade received, with a correlation coefficient of [0.91](#). However, this doesn't actually capture the observation that

increasing hours worked *always* increases the grade. There's a sense here in which we want to be able to say that the correlation is *perfect* but for a somewhat different notion of what a "relationship" is. What we're looking for is something that captures the fact that there is a perfect **ordinal relationship** here. That is, if student 1 works more hours than student 2, then we can guarantee that student 1 will get the better grade. That's not what a correlation of  $r = .91$  says at all.

How should we address this? Actually, it's really easy. If we're looking for ordinal relationships all we have to do is treat the data as if it were ordinal scale! So, instead of measuring effort in terms of "hours worked", let's rank all 10 of our students in order of hours worked. That is, student 1 did the least work out of anyone (2 hours) so they get the lowest rank (rank = 1). Student 4 was the next laziest, putting in only 6 hours of work over the whole semester, so they get the next lowest rank (rank = 2). Notice that I'm using "rank = 1" to mean "low rank". Sometimes in everyday language we talk about "rank = 1" to mean "top rank" rather than "bottom rank". So be careful, you can rank "from smallest value to largest value" (i.e., small equals rank 1) or you can rank "from largest value to smallest value" (i.e., large equals rank 1). In this case, I'm ranking from smallest to largest, but as it's really easy to forget which way you set things up you have to put a bit of effort into remembering!

Okay, so let's have a look at our students when we rank them from worst to best in terms of effort and reward:

	rank (hours worked)	rank (grade received)
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

Hmm. These are *identical*. The student who put in the most effort got the best grade, the student with the least effort got the worst grade, etc. As the table above shows, these two rankings are identical, so if we now correlate them we get a perfect relationship, with a correlation of [1.0](#).

What we've just re-invented is **Spearman's rank order correlation**, usually denoted  $\rho$  to dis-

tinguish it from the Pearson correlation  $r$ . We can calculate Spearman's  $\rho$  using JASP simply by clicking the 'Spearman' check box in the 'Correlation Matrix' screen.

## 10.2

---

### Scatterplots

**Scatterplots** are a simple but effective tool for visualising the relationship between *two* variables, like we saw with the figures in the section on correlation (Section ??). It's this latter application that we usually have in mind when we use the term "scatterplot". In this kind of plot each observation corresponds to one dot. The horizontal location of the dot plots the value of the observation on one variable, and the vertical location displays its value on the other variable. In many situations you don't really have a clear opinions about what the *causal* relationship is (e.g., does A cause B, or does B cause A, or does some other variable C control both A and B). If that's the case, it doesn't really matter which variable you plot on the x-axis and which one you plot on the y-axis. However, in many situations you do have a pretty strong idea which variable you think is most likely to be causal, or at least you have some suspicions in that direction. If so, then it's conventional to plot the cause variable on the x-axis, and the effect variable on the y-axis. With that in mind, let's look at how to draw scatterplots in JASP, using the same `parenthood` data set (i.e. `parenthood.csv`) that I used when introducing correlations.

Suppose my goal is to draw a scatterplot displaying the relationship between the amount of sleep that I get (`dan.sleep`) and how grumpy I am the next day (`dan.grump`). The way in which we can use JASP to get this plot is to use the 'Plots' option under the 'Regression' - 'Correlation Matrix' button, giving us the output shown in Figure ???. Note that JASP draws a line through the points, we'll come onto this a bit later in Section (??). Plotting a scatterplot in this way also allow you to specify 'Densities for variables', which adds a histogram and density curve showing how the data in each variable is distributed. You can also specify the 'Statistics' option, which provides an estimate of the correlation along with a 95% confidence interval.

## 10.3

---

### What is a linear regression model?

Stripped to its bare essentials, linear regression models are basically a slightly fancier version of the Pearson correlation (Section ??), though as we'll see regression models are much more

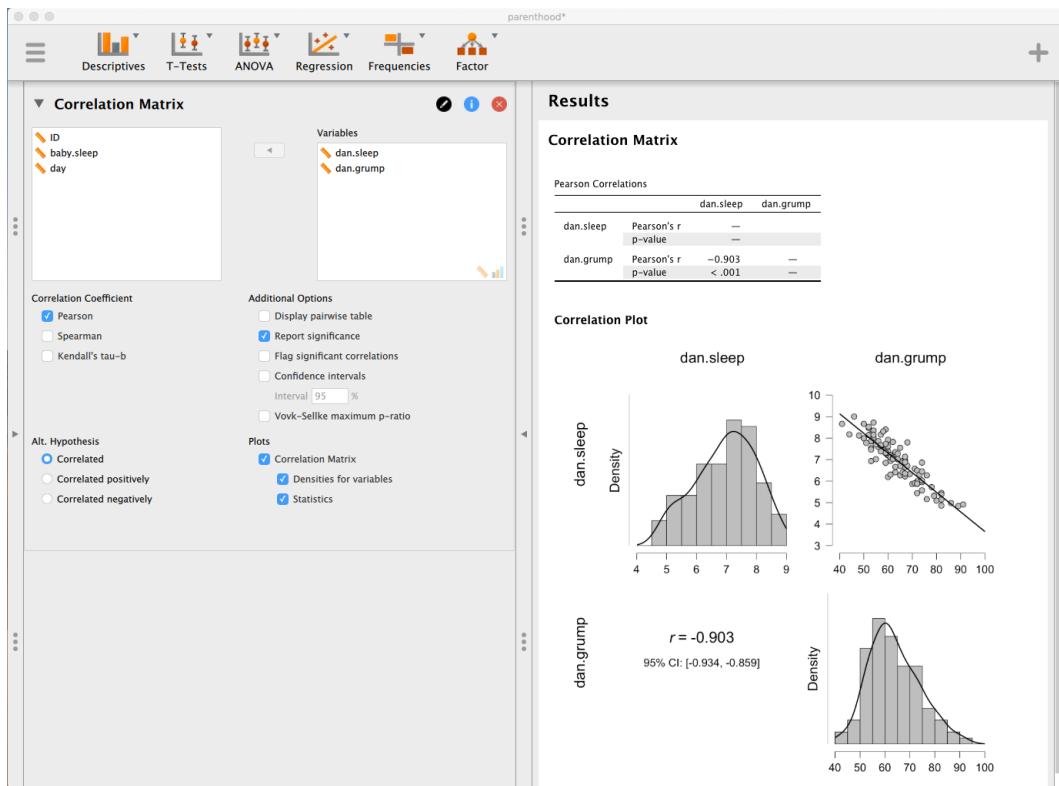


Figure 10.8 Scatterplot via the ‘Correlation Matrix’ method in JASP

powerful tools.

Since the basic ideas in regression are closely tied to correlation, we’ll return to the `parenthood.csv` file that we were using to illustrate how correlations work. Recall that, in this data set we were trying to find out why Dan is so very grumpy all the time and our working hypothesis was that I’m not getting enough sleep. We drew some scatterplots to help us examine the relationship between the amount of sleep I get and my grumpiness the following day, as in Figure ??, and as we saw previously this corresponds to a correlation of  $r = -.90$ , but what we find ourselves secretly imagining is something that looks closer to Figure ??a. That is, we mentally draw a straight line through the middle of the data. In statistics, this line that we’re drawing is called a **regression line**. Notice that, since we’re not idiots, the regression line goes through the middle of the data. We don’t find ourselves imagining anything like the rather silly plot shown in Figure ??b.

This is not highly surprising. The line that I’ve drawn in Figure ??b doesn’t “fit” the data very well, so it doesn’t make a lot of sense to propose it as a way of summarising the data, right? This

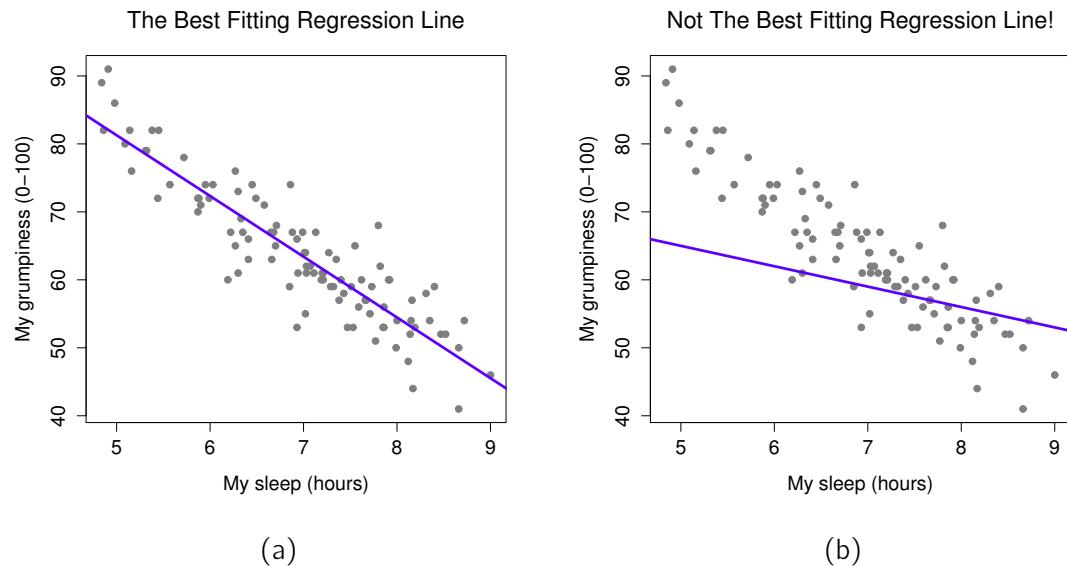


Figure 10.9 Panel a shows the sleep-grumpiness scatterplot from Figure ?? with the best fitting regression line drawn over the top. Not surprisingly, the line goes through the middle of the data. In contrast, panel b shows the same data, but with a very poor choice of regression line drawn over the top.

is a very simple observation to make, but it turns out to be very powerful when we start trying to wrap just a little bit of maths around it. To do so, let's start with a refresher of some high school maths. The formula for a straight line is usually written like this

$$y = a + bx$$

Or, at least, that's what it was when I went to high school all those years ago. The two *variables* are  $x$  and  $y$ , and we have two *coefficients*,  $a$  and  $b$ .<sup>\*2</sup> The coefficient  $a$  represents the *y-intercept* of the line, and coefficient  $b$  represents the *slope* of the line. Digging further back into our decaying memories of high school (sorry, for some of us high school was a long time ago), we remember that the intercept is interpreted as “the value of  $y$  that you get when  $x = 0$ ”. Similarly, a slope of  $b$  means that if you increase the  $x$ -value by 1 unit, then the  $y$ -value goes up by  $b$  units, and a negative slope means that the  $y$ -value would go down rather than up. Ah yes, it's all coming back to me now. Now that we've remembered that it should come as no surprise to discover that we use the exact same formula for a regression line. If  $Y$  is the outcome variable (the DV) and  $X$  is

---

<sup>\*2</sup>Also sometimes written as  $y = mx + b$  where  $m$  is the slope coefficient and  $b$  is the intercept (constant) coefficient.

the predictor variable (the IV), then the formula that describes our regression is written like this

$$\hat{Y}_i = b_0 + b_1 X_i$$

Hmm. Looks like the same formula, but there's some extra frilly bits in this version. Let's make sure we understand them. Firstly, notice that I've written  $X_i$  and  $Y_i$  rather than just plain old  $X$  and  $Y$ . This is because we want to remember that we're dealing with actual data. In this equation,  $X_i$  is the value of predictor variable for the  $i$ th observation (i.e., the number of hours of sleep that I got on day  $i$  of my little study), and  $Y_i$  is the corresponding value of the outcome variable (i.e., my grumpiness on that day). And although I haven't said so explicitly in the equation, what we're assuming is that this formula works for all observations in the data set (i.e., for all  $i$ ). Secondly, notice that I wrote  $\hat{Y}_i$  and not  $Y_i$ . This is because we want to make the distinction between the *actual data*  $Y_i$ , and the *estimate*  $\hat{Y}_i$  (i.e., the prediction that our regression line is making). Thirdly, I changed the letters used to describe the coefficients from  $a$  and  $b$  to  $b_0$  and  $b_1$ . That's just the way that statisticians like to refer to the coefficients in a regression model. I've no idea why they chose  $b$ , but that's what they did. In any case  $b_0$  always refers to the intercept term, and  $b_1$  refers to the slope.

Excellent, excellent. Next, I can't help but notice that, regardless of whether we're talking about the good regression line or the bad one, the data don't fall perfectly on the line. Or, to say it another way, the data  $Y_i$  are not identical to the predictions of the regression model  $\hat{Y}_i$ . Since statisticians love to attach letters, names and numbers to everything, let's refer to the difference between the model prediction and that actual data point as a *residual*, and we'll refer to it as  $\varepsilon_i$ .<sup>\*3</sup> Written using mathematics, the residuals are defined as

$$\varepsilon_i = Y_i - \hat{Y}_i$$

which in turn means that we can write down the complete linear regression model as

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

## 10.4

---

### Estimating a linear regression model

Okay, now let's redraw our pictures but this time I'll add some lines to show the size of the residual for all observations. When the regression line is good, our residuals (the lengths of the

---

<sup>\*3</sup>The  $\varepsilon$  symbol is the Greek letter epsilon. It's traditional to use  $\varepsilon_i$  or  $e_i$  to denote a residual.

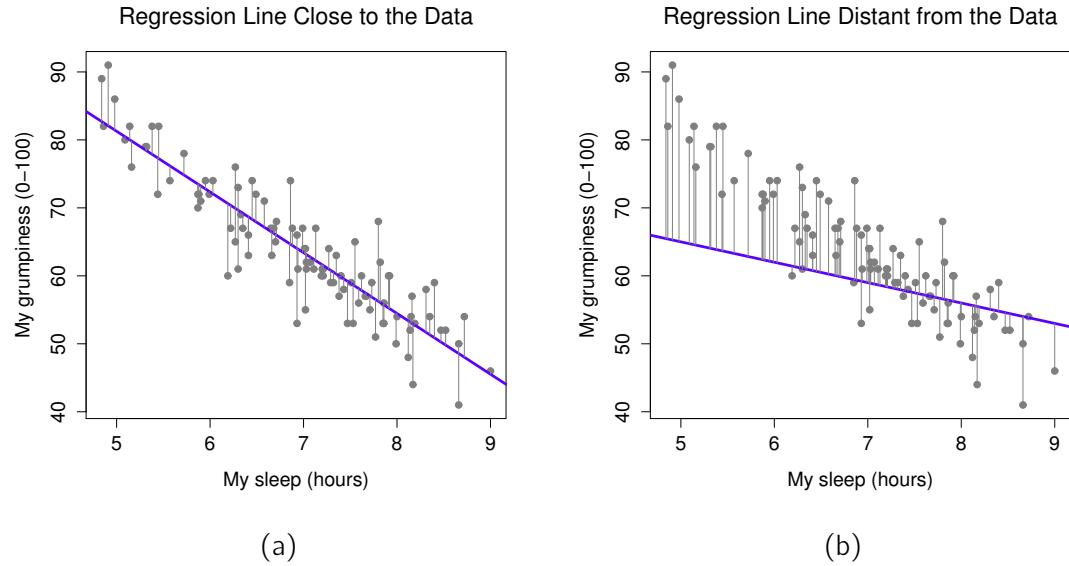


Figure 10.10 A depiction of the residuals associated with the best fitting regression line (panel a), and the residuals associated with a poor regression line (panel b). The residuals are much smaller for the good regression line. Again, this is no surprise given that the good line is the one that goes right through the middle of the data.

solid black lines) all look pretty small, as shown in Figure ??a, but when the regression line is a bad one the residuals are a lot larger, as you can see from looking at Figure ??b. Hmm. Maybe what we “want” in a regression model is *small* residuals. Yes, that does seem to make sense. In fact, I think I’ll go so far as to say that the “best fitting” regression line is the one that has the smallest residuals. Or, better yet, since statisticians seem to like to take squares of everything why not say that:

The estimated regression coefficients,  $\hat{b}_0$  and  $\hat{b}_1$ , are those that minimise the sum of the squared residuals, which we could either write as  $\sum_i(Y_i - \hat{Y}_i)^2$  or as  $\sum_i \varepsilon_i^2$ .

Yes, yes that sounds even better. And since I've indented it like that, it probably means that this is the right answer. And since this is the right answer, it's probably worth making a note of the fact that our regression coefficients are *estimates* (we're trying to guess the parameters that describe a population!), which is why I've added the little hats, so that we get  $\hat{b}_0$  and  $\hat{b}_1$  rather than  $b_0$  and  $b_1$ . Finally, I should also note that, since there's actually more than one way to estimate a regression model, the more technical name for this estimation process is **ordinary least squares (OLS) regression**.

At this point, we now have a concrete definition for what counts as our “best” choice of regression coefficients,  $\hat{b}_0$  and  $\hat{b}_1$ . The natural question to ask next is, if our optimal regression coefficients are those that minimise the sum squared residuals, how do we *find* these wonderful numbers? The actual answer to this question is complicated and doesn’t help you understand the logic of regression.\*<sup>4</sup> This time I’m going to let you off the hook. Instead of showing you the long and tedious way first and then “revealing” the wonderful shortcut that JASP provides, let’s cut straight to the chase and just use JASP to do all the heavy lifting.

#### 10.4.1 Linear regression in JASP

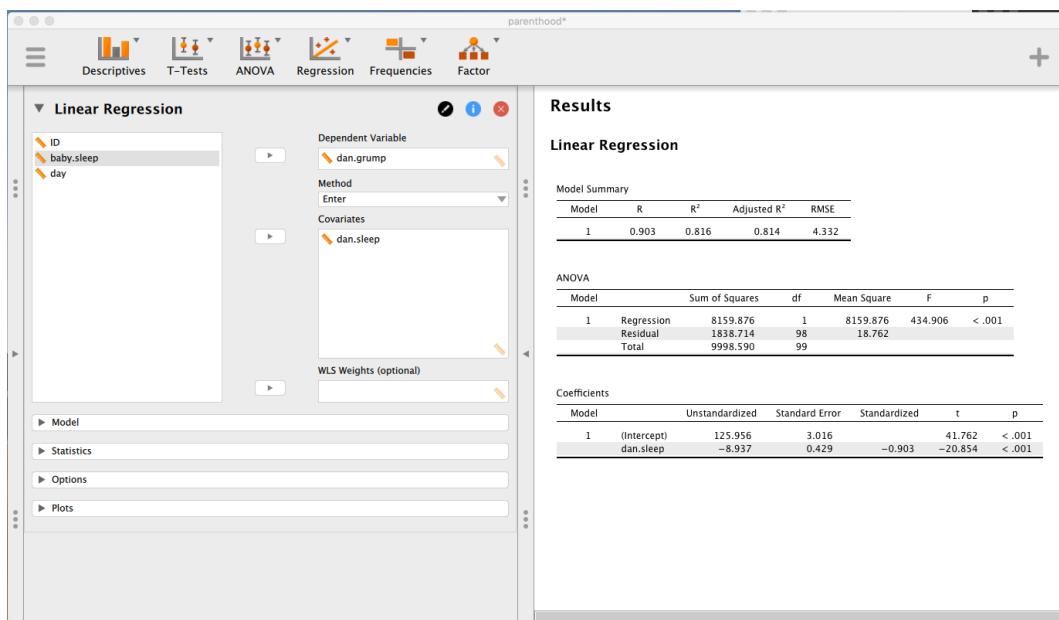


Figure 10.11 A JASP screenshot showing a simple linear regression analysis.

To run my linear regression, open up the ‘Regression’ - ‘Linear Regression’ analysis in JASP, using

---

\*<sup>4</sup>Or at least, I’m assuming that it doesn’t help most people. But on the off chance that someone reading this is a proper kung fu master of linear algebra (and to be fair, I always have a few of these people in my intro stats class), it *will* help you to know that the solution to the estimation problem turns out to be  $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ , where  $\hat{\mathbf{b}}$  is a vector containing the estimated regression coefficients,  $\mathbf{X}$  is the “design matrix” that contains the predictor variables (plus an additional column containing all ones; strictly  $\mathbf{X}$  is a matrix of the regressors, but I haven’t discussed the distinction yet), and  $\mathbf{y}$  is a vector containing the outcome variable. For everyone else, this isn’t exactly helpful and can be downright scary. However, since quite a few things in linear regression can be written in linear algebra terms, you’ll see a bunch of footnotes like this one in this chapter. If you can follow the maths in them, great. If not, ignore it.

the `parenthood.csv` data file. Then specify `dan.grump` as the ‘Dependent Variable’ and `dan.sleep` as the variable entered in the ‘Covariates’ box. This gives the results shown in Figure ??, showing an intercept  $\hat{b}_0 = 125.956$  and the slope  $\hat{b}_1 = -8.937$ . In other words, the best-fitting regression line that I plotted in Figure ?? has this formula:

$$\hat{Y}_i = 125.956 + (-8.937 X_i)$$

#### 10.4.2 Interpreting the estimated model

The most important thing to be able to understand is how to interpret these coefficients. Let’s start with  $\hat{b}_1$ , the slope. If we remember the definition of the slope, a regression coefficient of  $\hat{b}_1 = -8.94$  means that if I increase  $X_i$  by 1, then I’m decreasing  $Y_i$  by 8.94. That is, each additional hour of sleep that I gain will improve my mood, reducing my grumpiness by 8.94 grumpiness points. What about the intercept? Well, since  $\hat{b}_0$  corresponds to “the expected value of  $Y_i$  when  $X_i$  equals 0”, it’s pretty straightforward. It implies that if I get zero hours of sleep ( $X_i = 0$ ) then my grumpiness will go off the scale, to an insane value of ( $Y_i = 125.96$ ). Best to be avoided, I think.

## 10.5

---

### Multiple linear regression

The simple linear regression model that we’ve discussed up to this point assumes that there’s a single predictor variable that you’re interested in, in this case `dan.sleep`. In fact, up to this point every statistical tool that we’ve talked about has assumed that your analysis uses one predictor variable and one outcome variable. However, in many (perhaps most) research projects you actually have multiple predictors that you want to examine. If so, it would be nice to be able to extend the linear regression framework to be able to include multiple predictors. Perhaps some kind of **multiple regression** model would be in order?

Multiple regression is conceptually very simple. All we do is add more terms to our regression equation. Let’s suppose that we’ve got two variables that we’re interested in; perhaps we want to use both `dan.sleep` and `baby.sleep` to predict the `dan.grump` variable. As before, we let  $Y_i$  refer to my grumpiness on the  $i$ -th day. But now we have two  $X$  variables: the first corresponding to the amount of sleep I got and the second corresponding to the amount of sleep my son got. So we’ll let  $X_{i1}$  refer to the hours I slept on the  $i$ -th day and  $X_{i2}$  refers to the hours that the baby slept on that day. If so, then we can write our regression model like this:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \varepsilon_i$$

As before,  $\varepsilon_i$  is the residual associated with the  $i$ -th observation,  $\varepsilon_i = Y_i - \hat{Y}_i$ . In this model, we now have three coefficients that need to be estimated:  $b_0$  is the intercept,  $b_1$  is the coefficient associated with my sleep, and  $b_2$  is the coefficient associated with my son's sleep. However, although the number of coefficients that need to be estimated has changed, the basic idea of how the estimation works is unchanged: our estimated coefficients  $\hat{b}_0$ ,  $\hat{b}_1$  and  $\hat{b}_2$  are those that minimise the sum squared residuals.

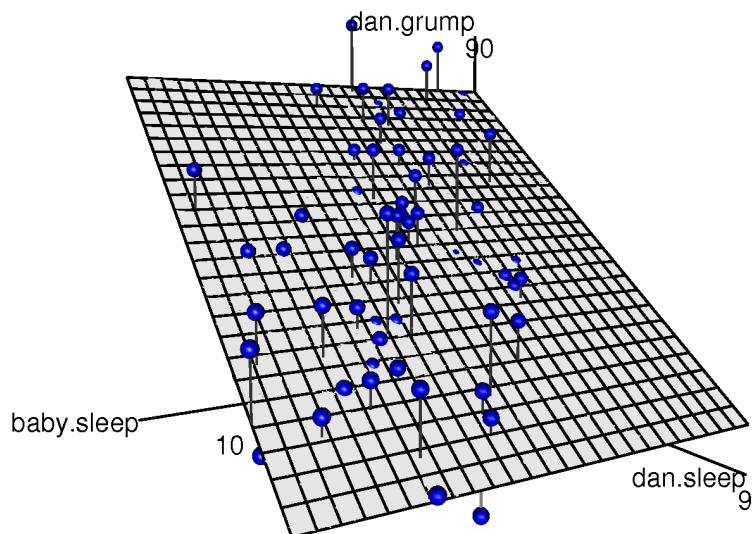


Figure 10.12 A 3D visualisation of a multiple regression model. There are two predictors in the model, `dan.sleep` and `baby.sleep` and the outcome variable is `dan.grump`. Together, these three variables form a 3D space. Each observation (dot) is a point in this space. In much the same way that a simple linear regression model forms a line in 2D space, this multiple regression model forms a plane in 3D space. When we estimate the regression coefficients what we're trying to do is find a plane that is as close to all the blue dots as possible.

### 10.5.1 Doing it in JASP

Multiple regression in JASP is no different to simple regression. All we have to do is add additional variables to the 'Covariates' box in JASP. For example, if we want to use both `dan.sleep` and `baby.sleep` as predictors in our attempt to explain why I'm so grumpy, then move `baby.sleep` across into the 'Covariates' box alongside `dan.sleep`. By default, JASP assumes that the model should include an intercept. The coefficients we get this time are:

(Intercept)	<code>d</code> an. <code>s</code> leep	<code>b</code> aby. <code>s</code> leep
125.966	-8.950	0.011

The coefficient associated with `dan.sleep` is quite large, suggesting that every hour of sleep I lose makes me a lot grumpier. However, the coefficient for `baby.sleep` is very small, suggesting that it doesn't really matter how much sleep my son gets. What matters as far as my grumpiness goes is how much sleep *I* get. To get a sense of what this multiple regression model looks like, Figure ?? shows a 3D plot that plots all three variables, along with the regression model itself.

### 10.5.2 Formula for the general case

The equation that I gave above shows you what a multiple regression model looks like when you include two predictors. Not surprisingly, then, if you want more than two predictors all you have to do is add more  $X$  terms and more  $b$  coefficients. In other words, if you have  $K$  predictor variables in the model then the regression equation looks like this

$$Y_i = b_0 + \left( \sum_{k=1}^K b_k X_{ik} \right) + \varepsilon_i$$

## 10.6

### Quantifying the fit of the regression model

So we now know how to estimate the coefficients of a linear regression model. The problem is, we don't yet know if this regression model is any good. For example, the regression model that we constructed in section ?? *claims* that every hour of sleep will improve my mood by quite a lot, but it might just be rubbish. Remember, the regression model only produces a prediction  $\hat{Y}_i$  about

what my mood is like, but my actual mood is  $Y_i$ . If these two are very close, then the regression model has done a good job. If they are very different, then it has done a bad job.

#### 10.6.1 The $R^2$ value

Once again, let's wrap a little bit of mathematics around this. Firstly, we've got the sum of the squared residuals

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

which we would hope to be pretty small. Specifically, what we'd like is for it to be very small in comparison to the total variability in the outcome variable

$$SS_{tot} = \sum_i (Y_i - \bar{Y})^2$$

While we're here, let's calculate these values ourselves, not by hand though. I have constructed a JASP file called `parenthood_rsquared.jasp`, which you can open from the book's data folder. You'll notice that this data file has 5 variables; two of them are the original `dan.sleep` and `dan.grump` variables that we've already been using. The other three are *calculated* variables:

1. `Y.pred` is the predicted value of grumpiness using the regression equation. It is calculated using the formula '`125.97 + (-8.94 * dan.sleep)`'.
2. `resid` is a measure of the residual error  $\varepsilon_i = Y_i - \hat{Y}_i$ , which represents the difference between our *predicted* value of grumpiness and our *actual* value of grumpiness. It is calculated using the formula '`dan.grump - Y.pred`'.
3. `sq.resid` is the square of the residual, and is calculated using the formula '`resid^2`'.

Since  $SS_{res}$  is the sum of these squared residuals, we can use JASP to find the sum of the `sq.resid` column. Simply click 'Descriptives' - 'Descriptive Statistics' and move `sq.resid` to the 'Variables' box. You'll then need to select 'Sum' from the 'Statistics' options below. This should give you a value of '`1838.714`'.

Wonderful. A big number that doesn't mean very much. Still, let's forge boldly onwards anyway and calculate the total sum of squares as well. That's also pretty simple. Let's calculate  $SS_{tot}$  similarly. This time, you'll need to create a new computed column. Click the "+" symbol to start. For 'Name', let's type '`sq.resid2`' (you'll see why in a minute). Be sure to select the "R" button, then click 'Create Column'. For your R code, type the following (see Figure ??):

```
(dan.grump - mean(dan.grump))^2
```

Then click ‘Compute column’. This will produce a column of values that are themselves residuals, but they are residuals (errors) against a *really bad* predictive model; that is, the model that just predicts grumpiness using the *mean* of all grumpiness values. To find  $SS_{tot}$ , we need to compute the sum of `sq.resid2`, just like we did above.

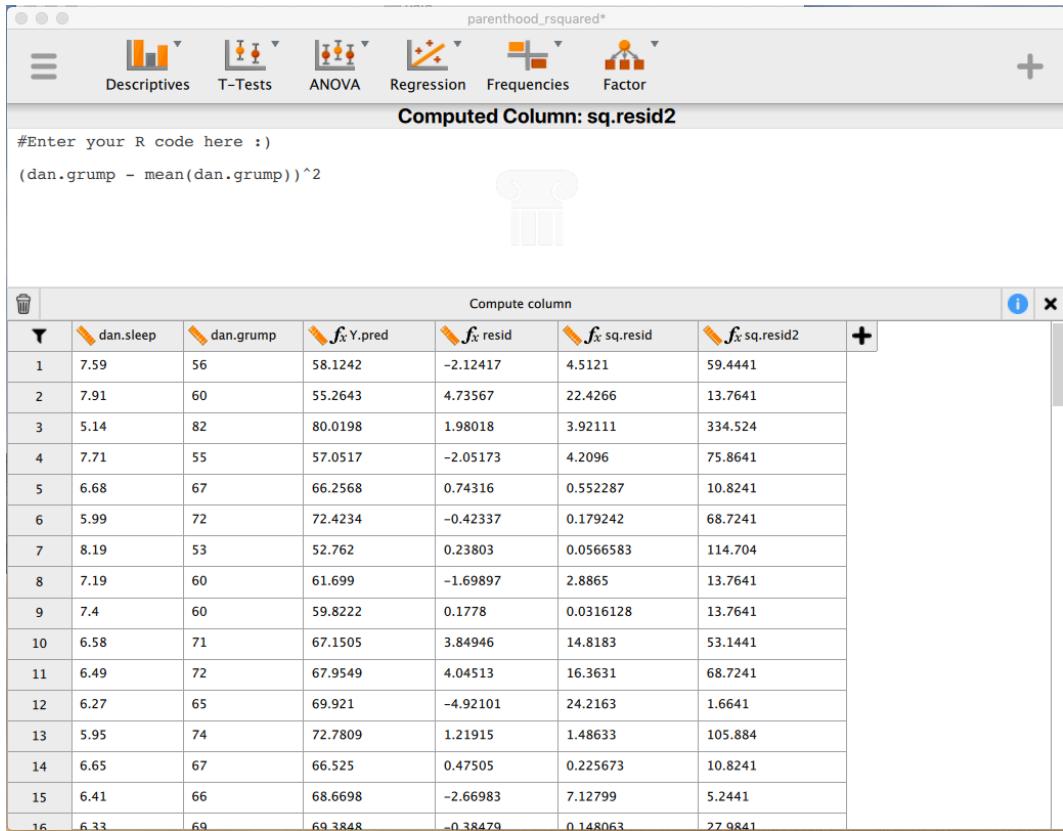


Figure10.13 A JASP screenshot showing a residual computation in R.

This should give you a value of ‘9998.590’. Hmm. Well, it’s a much bigger number than the last one, so this does suggest that our regression model was making good predictions (that is, it has greatly reduced the residual error compared to the model that uses the mean as a single predictor). But it’s not very interpretable.

Perhaps we can fix this. What we’d like to do is to convert these two fairly meaningless numbers into one number. A nice, interpretable number, which for no particular reason we’ll call  $R^2$ . What we would like is for the value of  $R^2$  to be equal to 1 if the regression model makes no errors in predicting the data. In other words, if it turns out that the residual errors are zero. That is, if  $SS_{res} = 0$  then we expect  $R^2 = 1$ . Similarly, if the model is completely useless, we would like  $R^2$

to be equal to 0. What do I mean by “useless”? Tempting as it is to demand that the regression model move out of the house, cut its hair and get a real job, I’m probably going to have to pick a more practical definition. In this case, all I mean is that the residual sum of squares is no smaller than the total sum of squares,  $SS_{res} = SS_{tot}$ . Wait, why don’t we do exactly that? The formula that provides us with our  $R^2$  value is pretty simple to write down,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

and equally simple to calculate by hand:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ &= 1 - \frac{1838.714}{9998.590} \\ &= 1 - 0.1839 \\ &= 0.8161. \end{aligned}$$

The  $R^2$  value, sometimes called the **coefficient of determination**<sup>5</sup> has a simple interpretation: it is the *proportion* of the variance in the outcome variable that can be accounted for by the predictor. So, in this case the fact that we have obtained  $R^2 = .8161$  means that the predictor (`dan.sleep`) explains 81.61% of the variance in the outcome (`dan.grump`).

Naturally, you don’t actually need to do all these computations by hand if you want to obtain the  $R^2$  value for your regression model. It turns out that JASP gives you this by default! Take a look at Figure ?? again; notice that in the top table labeled ‘Model Summary’, the value of  $R^2$  is already there!

#### 10.6.2 The relationship between regression and correlation

At this point we can revisit my earlier claim that regression, in this very simple form that I’ve discussed so far, is basically the same thing as a correlation. Previously, we used the symbol  $r$  to denote a Pearson correlation. Might there be some relationship between the value of the correlation coefficient  $r$  and the  $R^2$  value from linear regression? Of course there is: the squared correlation  $r^2$  is identical to the  $R^2$  value for a linear regression with only a single predictor. In other words, running a Pearson correlation is more or less equivalent to running a linear regression

---

<sup>5</sup>And by “sometimes” I mean “almost never”. In practice everyone just calls it “ $R$ -squared”.

model that uses only one predictor variable.

### 10.6.3 The adjusted $R^2$ value

One final thing to point out before moving on. It's quite common for people to report a slightly different measure of model performance, known as "adjusted  $R^2$ ". The motivation behind calculating the adjusted  $R^2$  value is the observation that adding more predictors into the model will *always* cause the  $R^2$  value to increase (or at least not decrease).

The adjusted  $R^2$  value introduces a slight change to the calculation, as follows. For a regression model with  $K$  predictors, fit to a data set containing  $N$  observations, the adjusted  $R^2$  is:

$$\text{adj. } R^2 = 1 - \left( \frac{SS_{res}}{SS_{tot}} \times \frac{N - 1}{N - K - 1} \right)$$

This adjustment is an attempt to take the degrees of freedom into account. The big advantage of the adjusted  $R^2$  value is that when you add more predictors to the model, the adjusted  $R^2$  value will only increase if the new variables improve the model performance more than you'd expect by chance. The big disadvantage is that the adjusted  $R^2$  value *can't* be interpreted in the elegant way that  $R^2$  can.  $R^2$  has a simple interpretation as the proportion of variance in the outcome variable that is explained by the regression model. To my knowledge, no equivalent interpretation exists for adjusted  $R^2$ .

An obvious question then is whether you should report  $R^2$  or adjusted  $R^2$ . This is probably a matter of personal preference. If you care more about interpretability, then  $R^2$  is better. If you care more about correcting for bias, then adjusted  $R^2$  is probably better. Speaking just for myself, I prefer  $R^2$ . My feeling is that it's more important to be able to interpret your measure of model performance. Besides, as we'll see in Section ??, if you're worried that the improvement in  $R^2$  that you get by adding a predictor is just due to chance and not because it's a better model, well we've got hypothesis tests for that.

10.7

## Hypothesis tests for regression models

So far we've talked about what a regression model is, how the coefficients of a regression model are estimated, and how we quantify the performance of the model (the last of these, incidentally, is basically our measure of effect size). The next thing we need to talk about is hypothesis tests.

There are two different (but related) kinds of hypothesis tests that we need to talk about: those in which we test whether the regression model as a whole is performing significantly better than a null model, and those in which we test whether a particular regression coefficient is significantly different from zero.

#### 10.7.1 Testing the model as a whole

Okay, suppose you've estimated your regression model. The first hypothesis test you might try is the null hypothesis that there is *no relationship* between the predictors and the outcome, and the alternative hypothesis that *the data are distributed in exactly the way that the regression model predicts*.

Formally, our "null model" corresponds to the fairly trivial "regression" model in which we include 0 predictors and only include the intercept term  $b_0$ :

$$H_0 : Y_i = b_0 + \varepsilon_i$$

If our regression model has  $K$  predictors, the "alternative model" is described using the usual formula for a multiple regression model:

$$H_1 : Y_i = b_0 + \left( \sum_{k=1}^K b_k X_{ik} \right) + \varepsilon_i$$

How can we test these two hypotheses against each other? The trick is to understand that it's possible to divide up the total variance  $SS_{tot}$  into the sum of the residual variance  $SS_{res}$  and the regression model variance  $SS_{mod}$ . I'll skip over the technicalities, since we'll get to that later when we look at ANOVA in Chapter ???. But just note that

$$SS_{mod} = SS_{tot} - SS_{res}$$

And we can convert the sums of squares into mean squares by dividing by the degrees of freedom.

$$MS_{mod} = \frac{SS_{mod}}{df_{mod}}$$

$$MS_{res} = \frac{SS_{res}}{df_{res}}$$

So, how many degrees of freedom do we have? As you might expect the  $df$  associated with the model is closely tied to the number of predictors that we've included. In fact, it turns out that  $df_{mod} = K$ . For the residuals the total degrees of freedom is  $df_{res} = N - K - 1$ .

Now that we've got our mean square values we can calculate an  $F$ -statistic like this

$$F = \frac{MS_{mod}}{MS_{res}}$$

and the degrees of freedom associated with this are  $K$  and  $N - K - 1$ .

We'll see much more of the  $F$  statistic in Chapter ??, but for now just know that we can interpret large  $F$  values as indicating that the null hypothesis is performing poorly in comparison to the alternative hypothesis. In a moment I'll show you how to do the test in JASP the easy way, but first let's have a look at the tests for the individual regression coefficients.

#### 10.7.2 Tests for individual coefficients

The  $F$ -test that we've just introduced is useful for checking that the model as a whole is performing better than chance. If your regression model doesn't produce a significant result for the  $F$ -test then you probably don't have a very good regression model (or, quite possibly, you don't have very good data). However, while failing this test is a pretty strong indicator that the model has problems, *passing* the test (i.e., rejecting the null) doesn't imply that the model is good! Why is that, you might be wondering? The answer to that can be found by looking at the coefficients for the multiple regression model we have already looked at in section ?? above, where the coefficients we got were:

(Intercept)	dan.sleep	baby.sleep
125.966	-8.950	0.011

I can't help but notice that the estimated regression coefficient for the `baby.sleep` variable is tiny (0.011), relative to the value that we get for `dan.sleep` (-8.950). Given that these two variables are absolutely on the same scale (they're both measured in "hours slept"), I find this illuminating. In fact, I'm beginning to suspect that it's really only the amount of sleep that / get that matters in order to predict my grumpiness.

We can re-use a hypothesis test that we discussed earlier, the  $t$ -test. The test that we're interested in has a null hypothesis that the true regression coefficient is zero ( $b = 0$ ), which is to be tested against the alternative hypothesis that it isn't ( $b \neq 0$ ). That is:

$$\begin{aligned} H_0 : b &= 0 \\ H_1 : b &\neq 0 \end{aligned}$$

How can we test this? Well, if the central limit theorem is kind to us we might be able to guess that the sampling distribution of  $\hat{b}$ , the estimated regression coefficient, is a normal distribution with mean centred on  $b$ . What that would mean is that if the null hypothesis were true, then the sampling distribution of  $\hat{b}$  has mean zero and unknown standard deviation. Assuming that we can come up with a good estimate for the standard error of the regression coefficient,  $se(\hat{b})$ , then we're in luck. That's *exactly* the situation for which we introduced the one-sample  $t$ -test way back in Chapter ???. So let's define a  $t$ -statistic like this

$$t = \frac{\hat{b}}{se(\hat{b})}$$

I'll skip over the reasons why, but our degrees of freedom in this case are  $df = N - K - 1$ . Irritatingly, the estimate of the standard error of the regression coefficient,  $se(\hat{b})$ , is not as easy to calculate as the standard error of the mean that we used for the simpler  $t$ -tests in Chapter ???. In fact, the formula is somewhat ugly, and not terribly helpful to look at.<sup>\*6</sup> For our purposes it's sufficient to point out that the standard error of the estimated regression coefficient depends on both the predictor and outcome variables, and it is somewhat sensitive to violations of the homogeneity of variance assumption (discussed shortly).

In any case, this  $t$ -statistic can be interpreted in the same way as the  $t$ -statistics that we discussed in Chapter ???. Assuming that you have a two-sided alternative (i.e., you don't really care if  $b > 0$  or  $b < 0$ ), then it's the extreme values of  $t$  (i.e., a lot less than zero or a lot greater than zero) that suggest that you should reject the null hypothesis.

### 10.7.3 Running the hypothesis tests in JASP

To compute all of the statistics that we have talked about so far, all you need to do is make sure the relevant options are checked in JASP and then run the regression. Fortunately, these options are usually selected by default. As you can see in Figure ???, we get a whole bunch of useful output.

The 'Coefficients' at the bottom of the JASP analysis results shown in ?? provides the coefficients of the regression model. Each row in this table refers to one of the coefficients in the regression model. The first row is the intercept term, and the later ones look at each of the predictors. The columns give you all of the relevant information. The first column (labeled 'Unstandardized') is the actual estimate of  $b$  (e.g., 125.966 for the intercept, and -8.950 for the

---

<sup>\*6</sup>For advanced readers only. The vector of residuals is  $\epsilon = \mathbf{y} - \mathbf{X}\hat{b}$ . For  $K$  predictors plus the intercept, the estimated residual variance is  $\hat{\sigma}^2 = \epsilon'\epsilon/(N - K - 1)$ . The estimated covariance matrix of the coefficients is  $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ , the main diagonal of which is  $se(\hat{b})$ , our estimated standard errors.

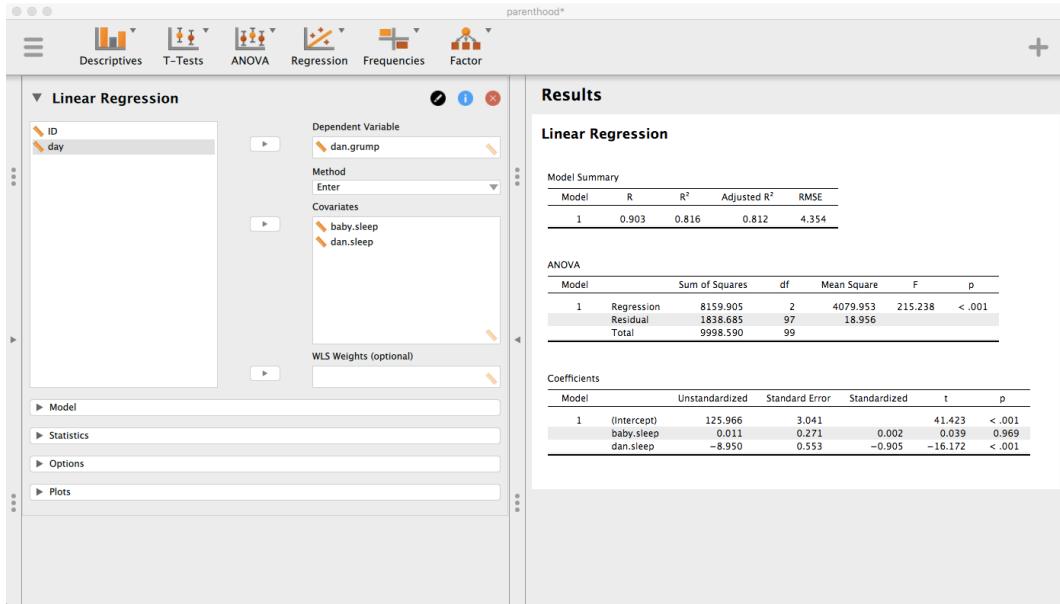


Figure 10.14 A JASP screenshot showing a multiple linear regression analysis, including relevant hypothesis tests.

`dan.sleep` predictor). The second column is the standard error estimate  $\hat{\sigma}_b$ . The third column provides a ‘Standardized’ regression coefficient; more about this in Section ???. The fourth column gives you the  $t$ -statistic, and it’s worth noticing that in this table  $t = \hat{b}/\text{se}(\hat{b})$  every time. Finally, the last column gives you the actual  $p$ -value for each of these tests.\*<sup>7</sup>

The only thing that the coefficients table itself doesn’t list is the degrees of freedom used in the  $t$ -test, which is always  $N - K - 1$  and is listed in the table in the middle of the output, labelled ‘ANOVA’. We can see from this table that the model performs significantly better than you’d expect by chance ( $F(2, 97) = 215.238, p < .001$ ), which isn’t all that surprising: the  $R^2 = .816$  value indicate that the regression model accounts for 81.6% of the variability in the outcome measure. However, when we look back up at the  $t$ -tests for each of the individual coefficients, we can see that the `baby.sleep` variable seems to have no significant effect. All the work in this model is being done by the `dan.sleep` variable. Taken together, these results suggest that this regression model is actually the wrong model for the data. You’d probably be better off dropping the `baby.sleep` predictor entirely. In other words, the simple regression model that we started with

\*<sup>7</sup>Note that, although JASP has done multiple tests here, it hasn’t done any sort of correction for multiple comparisons. These are standard one-sample  $t$ -tests with a two-sided alternative. If you want to make corrections for multiple tests, you need to do that yourself.

is likely the better model.

## 10.8

---

### Regarding regression coefficients

Before moving on to discuss the assumptions underlying linear regression and what you can do to check if they're being met, there's two more topics I want to briefly discuss, both of which relate to the regression coefficients. The first thing to talk about is calculating confidence intervals for the coefficients. After that, I'll discuss the somewhat murky question of how to determine which predictor is most important.

#### 10.8.1 Confidence intervals for the coefficients

Like any population parameter, the regression coefficients  $b$  cannot be estimated with complete precision from a sample of data; that's part of why we need hypothesis tests. Given this, it's quite useful to be able to report confidence intervals that capture our uncertainty about the true value of  $b$ . This is especially useful when the research question focuses heavily on an attempt to find out *how* strongly variable  $X$  is related to variable  $Y$ , since in those situations the interest is primarily in the regression weight  $b$ .

Fortunately, confidence intervals for the regression weights can be constructed in the usual fashion

$$\text{CI}(b) = \hat{b} \pm (t_{crit} \times \text{se}(\hat{b}))$$

where  $\text{se}(\hat{b})$  is the standard error of the regression coefficient, and  $t_{crit}$  is the relevant critical value of the appropriate  $t$  distribution. For instance, if it's a 95% confidence interval that we want, then the critical value is the 97.5th quantile of a  $t$  distribution with  $N - K - 1$  degrees of freedom. In other words, this is basically the same approach to calculating confidence intervals that we've used throughout.

In JASP we can display confidence intervals by selecting 'Confidence intervals' from the 'Statistics' menu in our regression model dialog. The default is 95% CI, but we could easily choose something different, say 99%, if that is what we decided on.

### 10.8.2 Calculating standardised regression coefficients

One more thing that you might want to do is to calculate “standardised” regression coefficients, often denoted  $\beta$ . The rationale behind standardised coefficients goes like this. In a lot of situations, your variables are on fundamentally different scales. Suppose, for example, my regression model aims to predict people’s IQ scores using their educational attainment (number of years of education) and their income as predictors. Obviously, educational attainment and income are not on the same scales. The number of years of schooling might only vary by 10s of years, whereas income can vary by 10,000s of dollars (or more). The units of measurement have a big influence on the regression coefficients. The  $b$  coefficients only make sense when interpreted in light of the units, both of the predictor variables and the outcome variable. This makes it very difficult to compare the coefficients of different predictors. Yet there are situations where you really do want to make comparisons between different coefficients. Specifically, you might want some kind of standard measure of which predictors have the strongest relationship to the outcome. This is what **standardised coefficients** aim to do.

The basic idea is quite simple; the standardised coefficients are the coefficients that you would have obtained if you’d converted all the variables to z-scores before running the regression.\*<sup>8</sup> The idea here is that, by converting all the predictors to z-scores, they all go into the regression on the same scale, thereby removing the problem of having variables on different scales. Regardless of what the original variables were, a  $\beta$  value of 1 means that an increase in the predictor of 1 standard deviation will produce a corresponding 1 standard deviation increase in the outcome variable. Therefore, if variable A has a larger absolute value of  $\beta$  than variable B, it is deemed to have a stronger relationship with the outcome. Or at least that’s the idea. It’s worth being a little cautious here, since this does rely very heavily on the assumption that “a 1 standard deviation change” is fundamentally the same kind of thing for all variables. It’s not always obvious that this is true.

Leaving aside the interpretation issues, let’s look at how it’s calculated. What you could do is standardise all the variables yourself and then run a regression, but there’s a much simpler way to do it. As it turns out, the  $\beta$  coefficient for a predictor  $X$  and outcome  $Y$  has a very simple

---

\*<sup>8</sup>Strictly, you standardise all the *regressors*. That is, every “thing” that has a regression coefficient associated with it in the model. For the regression models that I’ve talked about so far, each predictor variable maps onto exactly one regressor, and vice versa. However, that’s not actually true in general and we’ll see some examples of this in Chapter ???. But, for now we don’t need to care too much about this distinction.

formula, namely

$$\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$$

where  $\sigma_X$  is the standard deviation of the predictor, and  $\sigma_Y$  is the standard deviation of the outcome variable  $Y$ . This makes matters a lot simpler.

To make things even simpler, JASP computes the  $\beta$  coefficients by default, as you can see the third column of the ‘Coefficients’ table in Figure ???. This clearly shows that the `dan.sleep` variable has a much stronger effect than the `baby.sleep` variable. However, this is a perfect example of a situation where it would probably make sense to use the original coefficients  $b$  rather than the standardised coefficients  $\beta$ . After all, my sleep and the baby’s sleep are *already* on the same scale: number of hours slept. Why complicate matters by converting these to z-scores?

## 10.9

---

### Assumptions of regression

The linear regression model that I’ve been discussing relies on several assumptions. In Section ?? we’ll talk a lot more about how to check that these assumptions are being met, but first let’s have a look at each of them.

- *Normality*. Like many of the models in statistics, basic simple or multiple linear regression relies on an assumption of normality. Specifically, it assumes that the *residuals* are normally distributed. It’s actually okay if the predictors  $X$  and the outcome  $Y$  are non-normal, so long as the residuals  $\varepsilon$  are normal. See Section ??.
- *Linearity*. A pretty fundamental assumption of the linear regression model is that the relationship between  $X$  and  $Y$  actually is linear! Regardless of whether it’s a simple regression or a multiple regression, we assume that the relationships involved are linear.
- *Homogeneity of variance*. Strictly speaking, the regression model assumes that each residual  $\varepsilon_i$  is generated from a normal distribution with mean 0, and (more importantly for the current purposes) with a standard deviation  $\sigma$  that is the same for every single residual. In practice, it’s impossible to test the assumption that every residual is identically distributed. Instead, what we care about is that the standard deviation of the residual is the same for all values of  $\hat{Y}$ , and (if we’re being especially paranoid) all values of every predictor  $X$  in the model.
- *Uncorrelated predictors*. The idea here is that, in a multiple regression model, you don’t want your predictors to be too strongly correlated with each other. This isn’t “technically” an assumption of the regression model, but in practice it’s required. Predictors that are too

strongly correlated with each other (referred to as “collinearity”) can cause problems when evaluating the model.

- *Residuals are independent of each other.* This is really just a “catch all” assumption, to the effect that “there’s nothing else funny going on in the residuals”. If there is something weird (e.g., the residuals all depend heavily on some other unmeasured variable) going on, it might screw things up.
- *No “bad” outliers.* Again, not actually a technical assumption of the model (or rather, it’s sort of implied by all the others), but there is an implicit assumption that your regression model isn’t being too strongly influenced by one or two anomalous data points because this raises questions about the adequacy of the model and the trustworthiness of the data in some cases. See Section ??.

## 10.10 \_\_\_\_\_

### Model checking

The main focus of this section is **regression diagnostics**, a term that refers to the art of checking that the assumptions of your regression model have been met, figuring out how to fix the model if the assumptions are violated, and generally to check that nothing “funny” is going on. I refer to this as the “art” of model checking with good reason. It’s not easy, and while there are a lot of fairly standardised tools that you can use to diagnose and maybe even cure the problems that ail your model (if there are any, that is!), you really do need to exercise a certain amount of judgement when doing this. It’s easy to get lost in all the details of checking this thing or that thing, and it’s quite exhausting to try to remember what all the different things are. This has the very nasty side effect that a lot of people get frustrated when trying to learn *all* the tools, so instead they decide not to do *any* model checking. This is a bit of a worry!

In this section I describe several different things you can do to check that your regression model is doing what it’s supposed to. It doesn’t cover the full space of things you could do, but it’s still much more detailed than what I see a lot of people doing in practice, and even I don’t usually cover all of this in my intro stats class either. However, I do think it’s important that you get a sense of what tools are at your disposal, so I’ll try to introduce a bunch of them here. Finally, I should note that this section draws quite heavily from the **Fox2011** text, the book associated with the `car` package that is used to conduct regression analysis in R. The `car` package is notable for providing some excellent tools for regression diagnostics, and the book itself talks about them in an admirably clear fashion. I don’t want to sound too gushy about it, but I do think that **Fox2011**

is well worth reading, even if some of the advanced diagnostic techniques are only available in R and not JASP.

#### 10.10.1 Three kinds of residuals

The majority of regression diagnostics revolve around looking at the residuals, and by now you've probably formed a sufficiently pessimistic theory of statistics to be able to guess that, precisely because of the fact that we care a lot about the residuals, there are several different kinds of residual that we might consider. In particular, the following three kinds of residuals are referred to in this section: "ordinary residuals", "standardised residuals", and "Studentised residuals". There is a fourth kind that you'll see referred to in some of the Figures, and that's the "Pearson residual". However, for the models that we're talking about in this chapter the Pearson residual is identical to the ordinary residual.

The first and simplest kind of residuals that we care about are **ordinary residuals**. These are the actual raw residuals that I've been talking about throughout this chapter so far. The ordinary residual is just the difference between the fitted value  $\hat{Y}_i$  and the observed value  $Y_i$ . I've been using the notation  $\varepsilon_i$  to refer to the  $i$ -th ordinary residual, and darn it, I'm going to stick to it. With this in mind, we have the very simple equation

$$\varepsilon_i = Y_i - \hat{Y}_i$$

This is of course what we saw earlier, and unless I specifically refer to some other kind of residual, this is the one I'm talking about. So there's nothing new here. I just wanted to repeat myself. One drawback to using ordinary residuals is that they're always on a different scale, depending on what the outcome variable is and how good the regression model is. That is, unless you've decided to run a regression model without an intercept term, the ordinary residuals will have mean 0 but the variance is different for every regression. In a lot of contexts, especially where you're only interested in the *pattern* of the residuals and not their actual values, it's convenient to estimate the **standardised residuals**, which are normalised in such a way as to have standard deviation 1.

The way we calculate these is to divide the ordinary residual by an estimate of the (population) standard deviation of these residuals. For technical reasons, mumble mumble, the formula for this is

$$\varepsilon'_i = \frac{\varepsilon_i}{\hat{\sigma} \sqrt{1 - h_i}}$$

where  $\hat{\sigma}$  in this context is the estimated population standard deviation of the ordinary residuals, and  $h_i$  is the "hat value" of the  $i$ th observation. I haven't explained hat values to you yet (but

have no fear,<sup>a</sup> it's coming shortly), so this won't make a lot of sense. For now, it's enough to interpret the standardised residuals as if we'd converted the ordinary residuals to z-scores. In fact, that is more or less the truth, it's just that we're being a bit fancier.

---

<sup>a</sup>Or have no hope, as the case may be.

The third kind of residuals are **Studentised residuals** (also called "jackknifed residuals") and they're even fancier than standardised residuals. Again, the idea is to take the ordinary residual and divide it by some quantity in order to estimate some standardised notion of the residual.

The formula for doing the calculations this time is subtly different

$$\varepsilon_i^* = \frac{\varepsilon_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_i}}$$

Notice that our estimate of the standard deviation here is written  $\hat{\sigma}_{(-i)}$ . What this corresponds to is the estimate of the residual standard deviation that you *would have obtained* if you just deleted the  $i$ th observation from the data set. This sounds like the sort of thing that would be a nightmare to calculate, since it seems to be saying that you have to run  $N$  new regression models (even a modern computer might grumble a bit at that, especially if you've got a large data set). Fortunately, some terribly clever person has shown that this standard deviation estimate is actually given by the following equation:

$$\hat{\sigma}_{(-i)} = \hat{\sigma} \sqrt{\frac{N - K - 1 - \varepsilon_i'^2}{N - K - 2}}$$

Isn't that a pip?

Before moving on, I should point out that you don't often need to obtain these residuals yourself, even though they are at the heart of almost all regression diagnostics. Most of the time the various options that provide the diagnostics, or assumption checks, will take care of these calculations for you. Even so, it's always nice to know how to actually get hold of these things yourself in case you ever need to do something non-standard.

#### 10.10.2 Three kinds of anomalous data

One danger that you can run into with linear regression models is that your analysis might be disproportionately sensitive to a smallish number of "unusual" or "anomalous" observations. I discussed this idea previously in Section ?? in the context of discussing the outliers that get

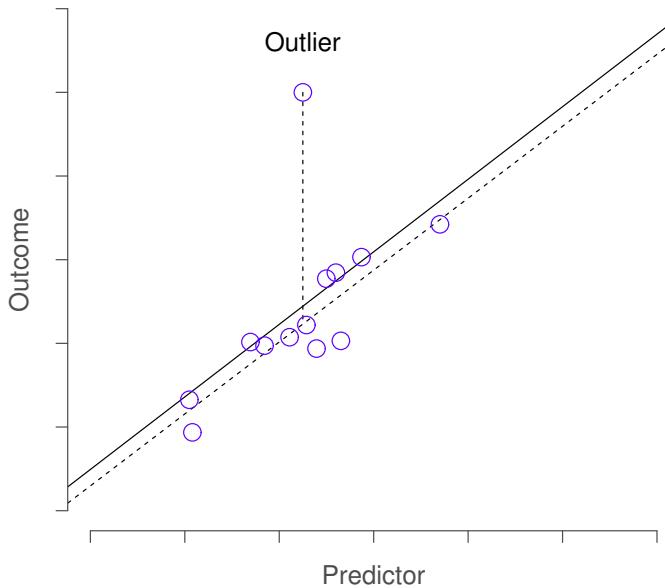


Figure 10.15 An illustration of outliers. The dotted lines plot the regression line that would have been estimated without the anomalous observation included, and the corresponding residual (i.e., the Studentised residual). The solid line shows the regression line with the anomalous observation included. The outlier has an unusual value on the outcome (y axis location) but not the predictor (x axis location), and lies a long way from the regression line.

automatically identified by the boxplot option under ‘Exploration’ – ‘Descriptives’, but this time we need to be much more precise. In the context of linear regression, there are three conceptually distinct ways in which an observation might be called “anomalous”. All three are interesting, but they have rather different implications for your analysis.

The first kind of unusual observation is an **outlier**. The definition of an outlier (in this context) is an observation that is very different from what the regression model predicts. An example is shown in Figure ???. In practice, we operationalise this concept by saying that an outlier is an observation that has a very large Studentised residual,  $\varepsilon_i^*$ . Outliers are interesting: a big outlier *might* correspond to junk data, e.g., the variables might have been recorded incorrectly in the data set, or some other defect may be detectable. Note that you shouldn’t throw an observation away just because it’s an outlier. But the fact that it’s an outlier is often a cue to look more closely at

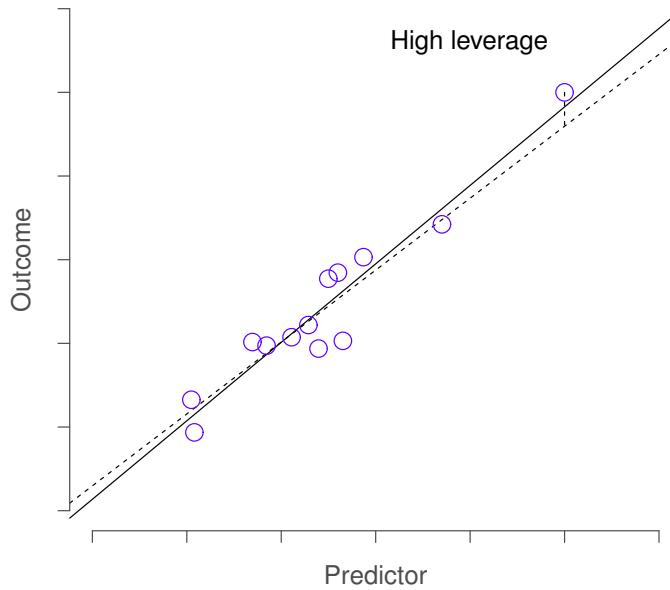


Figure 10.16 An illustration of high leverage points. The anomalous observation in this case is unusual both in terms of the predictor (x axis) and the outcome (y axis), but this unusualness is highly consistent with the pattern of correlations that exists among the other observations. The observation falls very close to the regression line and does not distort it.

.....

that case and try to find out why it's so different.

The second way in which an observation can be unusual is if it has high **leverage**, which happens when the observation is very different from all the other observations. This doesn't necessarily have to correspond to a large residual. If the observation happens to be unusual on all variables in precisely the same way, it can actually lie very close to the regression line. An example of this is shown in Figure ???. The leverage of an observation is operationalised in terms of its *hat value*, usually written  $h_i$ . The formula for the hat value is rather complicated<sup>\*9</sup> but its interpretation is not:  $h_i$  is a measure of the extent to which the  $i$ -th observation is "in control" of where the

---

<sup>\*9</sup>Again, for the linear algebra fanatics: the "hat matrix" is defined to be that matrix  $H$  that converts the vector of observed values  $\mathbf{y}$  into a vector of fitted values  $\hat{\mathbf{y}}$ , such that  $\hat{\mathbf{y}} = H\mathbf{y}$ . The name comes from the fact that this is the matrix that "puts a hat on  $\mathbf{y}$ ". The hat value of the  $i$ -th observation is the  $i$ -th diagonal element of this matrix (so technically I should be writing it as  $h_{ii}$  rather than  $h_i$ ). Oh, and in case you care, here's how it's calculated:  $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ . Pretty, isn't it?

regression line ends up going.

In general, if an observation lies far away from the other ones in terms of the predictor variables, it will have a large hat value (as a rough guide, high leverage is when the hat value is more than 2-3 times the average; and note that the sum of the hat values is constrained to be equal to  $K + 1$ ). High leverage points are also worth looking at in more detail, but they're much less likely to be a cause for concern unless they are also outliers.

This brings us to our third measure of unusualness, the **influence** of an observation. A high influence observation is an outlier that has high leverage. That is, it is an observation that is very different to all the other ones in some respect, and also lies a long way from the regression line. This is illustrated in Figure ???. Notice the contrast to the previous two figures. Outliers don't move the regression line much and neither do high leverage points. But something that is both an outlier and has high leverage, well that has a big effect on the regression line. That's why we call

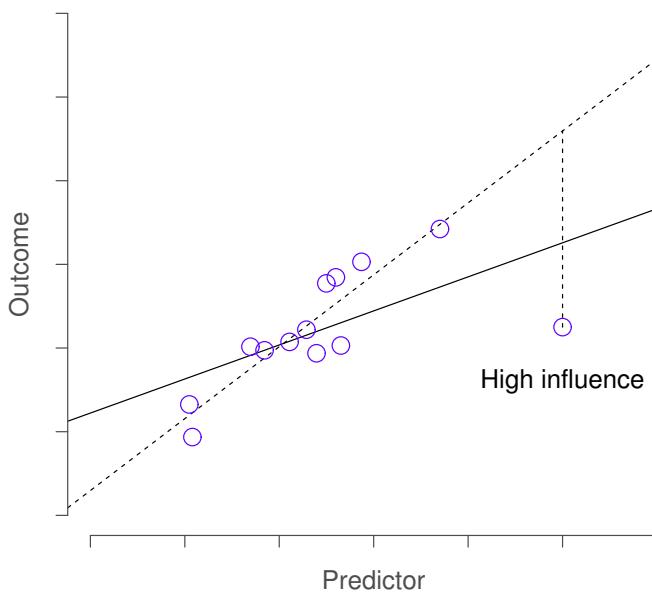


Figure10.17 An illustration of high influence points. In this case, the anomalous observation is highly unusual on the predictor variable (x axis), and falls a long way from the regression line. As a consequence, the regression line is highly distorted, even though (in this case) the anomalous observation is entirely typical in terms of the outcome variable (y axis).

these points high influence, and it's why they're the biggest worry. We operationalise influence in terms of a measure known as **Cook's distance**.

$$D_i = \frac{\varepsilon_i^{*2}}{K+1} \times \frac{h_i}{1-h_i}$$

Notice that this is a multiplication of something that measures the outlier-ness of the observation (the bit on the left), and something that measures the leverage of the observation (the bit on the right).

In order to have a large Cook's distance an observation must be a fairly substantial outlier *and* have high leverage. As a rough guide, Cook's distance greater than 1 is often considered large (that's what I typically use as a quick and dirty rule).

In JASP, information about Cook's distance can be calculated by clicking on 'Casewise diagnostics' under the 'Statistics' menu. There are two ways to visualize these data. First, you can select 'All' to see Cook's distance for each case (i.e., row of data); see Figure ???. Alternatively, you can opt to display *only* those cases for which Cook's distance is greater than some threshold; the default in JASP is 1. In either case, you can see that we have no data cases that are beyond this threshold.

Casewise Diagnostics						
Case Number	Std. Residual	dan.grump	Predicted Value	Residual	Cook's Distance	
1	-0.497	56.000	58.140	-2.140	0.002	
2	1.104	60.000	55.292	4.708	0.017	
3	0.464	82.000	80.045	1.955	0.005	
4	-0.477	55.000	57.060	-2.060	0.001	
5	0.168	67.000	66.281	0.719	0.000	
6	-0.095	72.000	72.407	-0.407	0.000	
7	0.053	53.000	52.773	0.227	0.000	
8	-0.393	60.000	61.700	-1.700	0.001	
9	0.047	60.000	59.797	0.203	0.000	
10	0.890	71.000	67.148	3.852	0.003	
11	0.959	72.000	68.001	3.999	0.027	
12	-1.139	65.000	69.912	-4.912	0.008	

Figure10.18 JASP output showing Cook's distance for each case/row of data

An obvious question to ask next is, if you do have large values of Cook's distance what should you do? As always, there's no hard and fast rule. Probably the first thing to do is to try running the regression with the outlier with the greatest Cook's distance<sup>\*10</sup> excluded and see what happens to the model performance and to the regression coefficients. If they really are substantially different,

<sup>\*10</sup>although currently there isn't a very easy way to do this in JASP, so a more powerful regression program such as the `car` package in R would be better for this more advanced analysis

it's time to start digging into your data set and your notes that you no doubt were scribbling as you ran your study. Try to figure out *why* the point is so different. If you start to become convinced that this one data point is badly distorting your results then you might consider excluding it, but that's less than ideal unless you have a solid explanation for why this particular case is qualitatively different from the others and therefore deserves to be handled separately.

#### 10.10.3 Checking the normality of the residuals

Like many of the statistical tools we've discussed in this book, regression models rely on a normality assumption. In this case, we assume that the residuals are normally distributed. The first thing we can do is draw a QQ-plot in JASP via the 'Plots' - 'Q-Q plot standardized residuals' option. The output is shown in Figure ??, showing the standardised residuals plotted as a function of their theoretical quantiles according to the regression model.

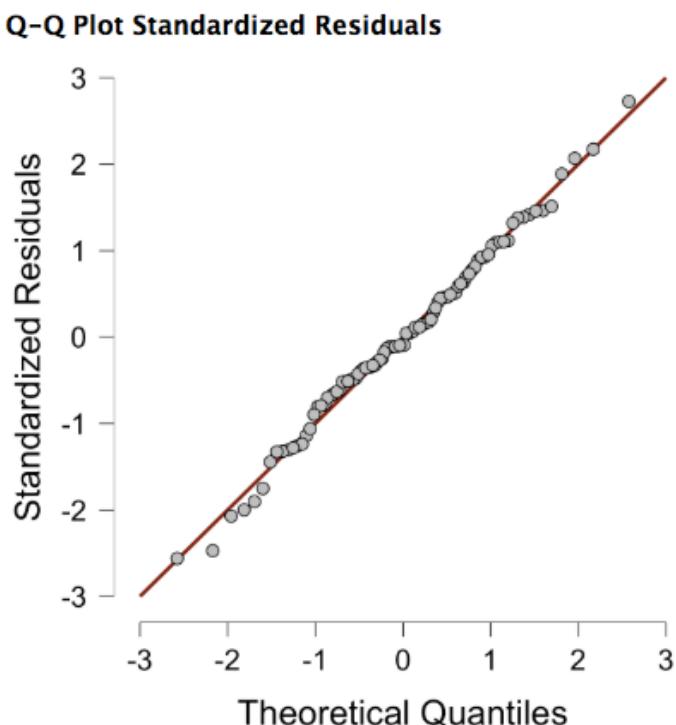


Figure 10.19 Plot of the theoretical quantiles according to the model, against the quantiles of the standardised residuals, produced in JASP.

Another thing we should check is the relationship between the fitted values and the residuals

themselves. We can get JASP to do this using the various ‘Residuals Plots’ choices, each of which provides a scatterplot for the predictor variables, the outcome variable, and the fitted values against residuals; see Figure ???. In these plots we are looking for a fairly uniform distribution of ‘dots’, with no clear bunching or patterning of the ‘dots’. Looking at these plots, there is nothing particularly worrying as the dots are fairly evenly spread across the whole plot. There may be a little bit of non-uniformity in plot (b), but it is not a strong deviation and probably not worth worrying about.

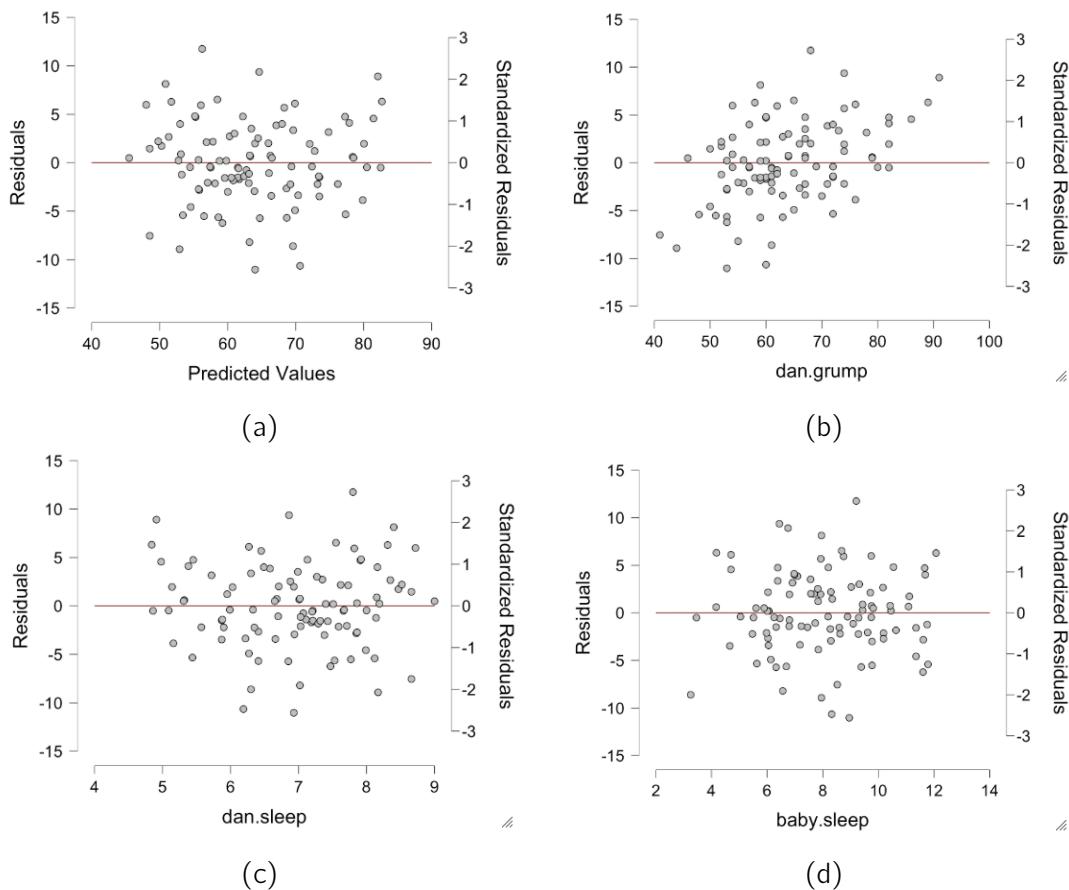


Figure 10.20 Residuals plots produced in JASP

If we were worried, then in a lot of cases the solution to this problem (and many others) is to transform one or more of the variables. Transformations are beyond the scope of this text.

## 10.11

---

## Model selection

One fairly major problem that remains is the problem of “model selection”. That is, if we have a data set that contains several variables, which ones should we include as predictors, and which ones should we not include? In other words, we have a problem of **variable selection**. In general, model selection is a complex business but it’s made somewhat simpler if we restrict ourselves to the problem of choosing a subset of the variables that ought to be included in the model. Nevertheless, I’m not going to try covering even this reduced topic in a lot of detail. Instead, I’ll talk about two broad principles that you need to think about, and then discuss one concrete tool to help you select a subset of variables to include in your model. First, the two principles:

- It’s nice to have an actual substantive basis for your choices. That is, in a lot of situations you the researcher have good reasons to pick out a smallish number of possible regression models that are of theoretical interest. These models will have a sensible interpretation in the context of your field. Never discount the importance of this. Statistics serves the scientific process, not the other way around.
- To the extent that your choices rely on statistical inference, there is a trade off between simplicity and goodness of fit. As you add more predictors to the model you make it more complex. Each predictor adds a new free parameter (i.e., a new regression coefficient), and each new parameter increases the model’s capacity to “absorb” random variations. So the goodness of fit (e.g.,  $R^2$ ) continues to rise, sometimes trivially or by chance, as you add more predictors no matter what. If you want your model to be able to generalise well to new observations you need to avoid throwing in too many variables.

This latter principle is often referred to as **Ockham’s razor** and is often summarised in terms of the following pithy saying: *do not multiply entities beyond necessity*. In this context, it means don’t chuck in a bunch of largely irrelevant predictors just to boost your  $R^2$ . Hmm. Yeah, the original was better.

In any case, what we need is an actual mathematical criterion that will implement the qualitative principle behind Ockham’s razor in the context of selecting a regression model. As it turns out there are several possibilities. The one that I’ll talk about is the **Akaike information criterion (Akaike1974)**; currently, it is not part of JASP’s standard output, but it is quite easy to compute with the model fit data that JASP does produce.

In the context of a linear regression model, the AIC for a model that has  $n$  observations and  $K$  predictor variables (not including the intercept) can be computed<sup>a</sup> as

$$AIC = n \ln(SS_{res}) + 2K$$

---

<sup>a</sup>Strictly speaking, this formula is not completely correct. Akaike's original definition was in terms of something called a *maximum likelihood estimate* for the model, and as such, there are some other terms that appear in the computation. However, many of them don't depend on the model, and given that the purpose of the AIC is to *compare models*, these terms will be present in all models and will mathematically "wash out". Thus, I am presenting a 'bare bones' version of the formula that is sufficient for our purposes.

Here's the basic principle behind using AIC for model comparison: the smaller the AIC value, the better the model performance. If we ignore the low level details it's fairly obvious what the AIC does. On the left we have a term that decreases as the model predictions get better; on the right we have a term that increases as the model complexity increases. The best model is the one that both fits the data well (small  $SS_{res}$ , left hand side) and uses as few predictors as possible (small  $K$ , right hand side). In short, this is a simple mathematical implementation of Ockham's razor.

Let's demonstrate how AIC can be used to compare the two regression models we have computed in this chapter. Consider first the regression model with only one predictor: `dan.sleep` (see Figure ??). In this model, we have  $n = 100$  observations,  $K = 1$  predictor, and  $SS_{res} = 1838.714$ . Thus,

$$\begin{aligned} AIC_1 &= n \ln(SS_{res}) + 2K \\ &= 100 \ln(1838.714) + 2(1) \\ &= 753.68 \end{aligned}$$

Now consider the second model using two predictors: `dan.sleep` and `baby.sleep` (see Figure ??). In this model, we have  $n = 100$  observations,  $K = 2$  predictors, and  $SS_{res} = 1838.685$ . This gives us

$$\begin{aligned} AIC_2 &= n \ln(SS_{res}) + 2K \\ &= 100 \ln(1838.685) + 2(2) \\ &= 755.68. \end{aligned}$$

Since  $AIC_1 < AIC_2$ , this tells us that Model 1 is the better fit, which confirms our intuitions. Adding `baby.sleep` doesn't add much to the model fit, but it increases model complexity. AIC balances these two requirements; the penalty for adding an additional parameter is not outweighed by the meager improvement in model fit.

10.12 \_\_\_\_\_

## Summary

- Want to know how strong the relationship is between two variables? Calculate a correlation (Section ??).
- Drawing scatterplots (Section ??).
- Basic ideas in linear regression and how regression models are estimated (Sections ?? and ??).
- Multiple linear regression (Section ??).
- Measuring the overall performance of a regression model using  $R^2$  (Section ??).
- Hypothesis tests for regression models (Section ??)
- Calculating confidence intervals for regression coefficients and standardised coefficients (Section ??).
- The assumptions of regression (Section ??) and how to check them (Section ??).
- Selecting a regression model (Section ??).



## 11. 平均を比較する（一元配置 ANOVA）

---

この章では心理統計の中で最も広く使われているツールの一つ，“分散分析”として知られている手法を紹介します。普通は ANOVA と呼ばれています。基本的な技術はロナルド・フィッシャー卿によって 20 世紀初頭に開発されたもので、不便な用語を使っているのも彼によるところです。ANOVA という用語は、二つの意味で若干ミスリーディングです。第一に、この技術名は分散といっていますが、ANOVA は平均の違いを検証することに興味があるのです。第二に、いくつかの異なる手法が全て ANOVA として引用されていますが、中には関係が薄いものもあるのです。この本の後の方では、全く違う状況に適用される異なる手法の ANOVA に出会すことになりますが、この章の目的は最も単純な形式の ANOVA についてだけ考えることにします。ここでは異なる群について観測しており、これらの群の違いが関心のある結果変数において差があるかどうかに興味があります。これは**一元配置 ANOVA** と呼ばれる問い合わせ方になります。

この章の構造は次のようになっています。セクション ?? ではこの章を通じて例として利用することになる、架空データについて紹介します。データの導入が終われば、一元配置 ANOVA が実際どのように働くのか、そのメカニクスを記述し（セクション ??），それから JASP でどのように実行するかに注目していきます（セクション ??）。この二つのセクションが、この章の中核となります。この章の残りの箇所では、ANOVA を実行するときには避けられない重要なトピックスの範囲について議論します。例えば効果量をどう計算するのかとか（セクション ??），事後検定や多重比較の補正（セクション ??），ANOVA が依存している仮定（セクション ??）などです。私たちはまた、これらの仮定をチェックする方法や、もし仮定が満たされなかつたら何ができるのかについても論じます（セクション ?? から ??）。それから、反復測定 ANOVA についてセクションもカバーしていきます（セクション ?? から ??）。この章の終わりには、ANOVA とほかの統計的ツールとの関係についても少し紹介します（Section ??）。

11.1 \_\_\_\_\_

## データセットについて

あなたが *Joyzepam* と呼ばれる新しい抗うつ剤をテストする臨床試験に参加するとしましょう。薬の性能を公平に検証するために、この試験では 3 つの異なる薬を含めて検証することにします。一つはプラセボ、もう一つは既に抗うつ/抗不安剤として知られている *Anxifree* です。最初のテストでは、抑うつを抑えるのに変化があるのかを検証するために 18 人の参加者が集められました。薬は心理学的セラピーと一緒に用いられることがありますから、あなたの研究では 9 人が認知行動療法 (CBT) を受けている、残りの 9 人は受けていないという状況にあります。参加者の処遇はランダムに割り当てられているので（もちろん二重盲検法で、です）、3 種類の薬それぞれについて、3 人が CBT を受け、3 人がセラピーを受けない人ということになります。心理学者は各薬の処方後 3 ヶ月たったあとで、個々人の気分を査定し、個々人の気分が全体的に改善されたかどうかを -5 点から +5 点の尺度で評定してもらいました。この研究デザインのデータファイルを `clinicaltrial.csv` にアップロードしています。データセットには三つの変数、`drug, therapy, mood.gain` が含まれています。

この章の目的として、私たちが本当に関心を持っているのは、`mood.gain` に対して `drug` の効果があるかどうかです。最初にやるべきことは、記述統計を計算していくつかのグラフを描くことです。第 ?? 章で、これを JASP でどのようにするか、‘Descriptives’ – ‘Descriptive Statistics’ の中の ‘Split’ ボックス’ を使う方法をお見せしました。その結果を図 ?? に示してあります。

プロットがはっきり示しているように、*Joyzepam* 群は *Anxifree* やプラセボ群よりも大きな改善が見られています。*Anxifree* 群は統制群よりも大きな気分の向上がみられます、その差は大きくありません。私たちが知りたい答えは、これらの差が“本当に”あるものかどうか、あるいはそれがただの偶然なのか?ということです。

### 11.2

---

## ANOVA のしくみ

私たちの臨床試験データに与えられた問い合わせるために、一元配置分散分析をすることになります。まずは一から統計的ツールを組み立てる難しい方法を見せるところからはじめ、JASP に組み込まれているかっこいい ANOVA 関数にアクセスできなくても計算できることを示します。注意深く読み解いて欲しいと思います。そして ANOVA がどういう仕組みなのかを本当に理解するために、1,2 回はこの長い道のりに挑戦してみてください。あなたがこのやり方を掴み取ったら、なにがあっても二度と同じやり方でやらなくても構いませんから。

前のセクションで私が示した実験デザインは、三つの異なる薬による気分の変化の平均を比較することに興味があるということを、強く示していました。つまり、私たちがやろうとしている分析は *t*

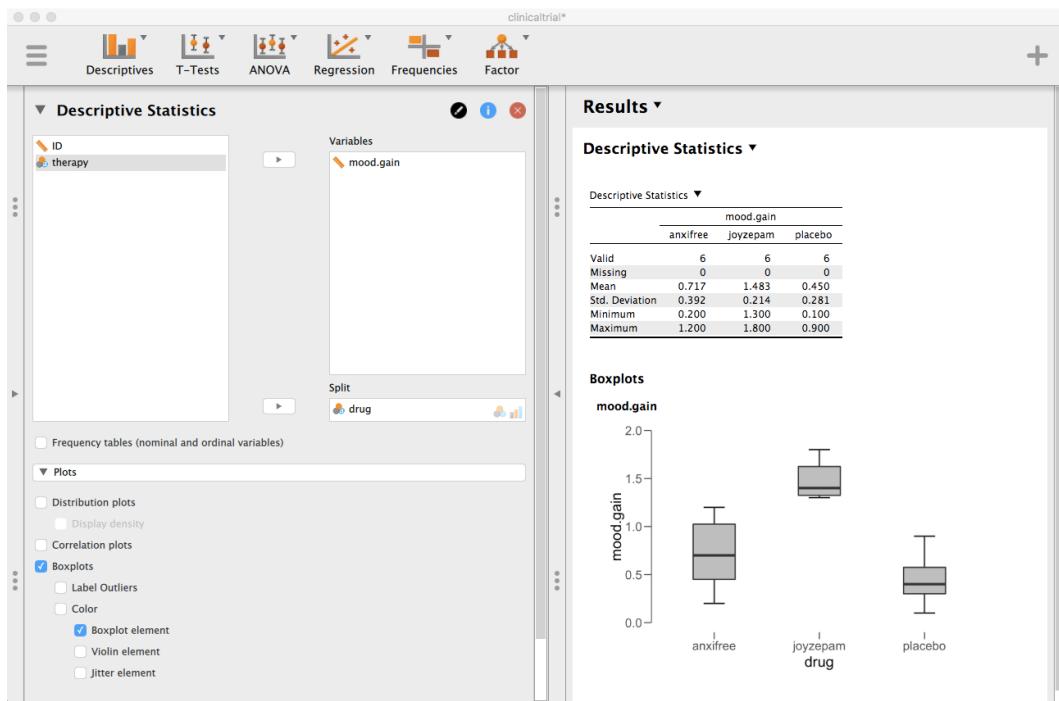


Figure11.1 記述統計と mood gain について薬の設定ごとに分けて描いたボックスプロットの JASP のスクリーンショット。

検定(第 ??章)に似ていますが、二つ以上のグループが含まれていることになります。ここで  $\mu_P$  を プラセボによって作られた気分変容の母平均を示しているとしましょう。そして  $\mu_A$  と  $\mu_J$  がそれぞれ Anxitfree と Joyzepam、この二つの薬による平均だとします。それから(すこし悲観主義ですが)私たちが検証しようとしている帰無仮説は、これら三つの母平均が同じだというものです。つまり二つの薬のどちらも、プラセボに比べてもなんの効果もない、というものです。この帰無仮説は次のように書くことができます。

$$H_0 : \text{次の式が正しい;} \quad \mu_P = \mu_A = \mu_J$$

結果的に、私たちの対立仮説は、三つの異なる処遇の少なくとも一つは、他とは異なるということになります。これを数学的に書くと少しトリッキーに見えます。というのも(この後議論するように)、帰無仮説が間違っている時でもほんのちょっとした違いしかないからです。対立仮説は次のように書くことができます。

$$H_1 : \text{次の式が正しくない;} \quad \mu_P \neq \mu_A = \mu_J$$

帰無仮説は私たちがこれまで見てきた検定のどれと比べても、トリッキーな感じ満載です。どうしたらよいでしょうか? この章のタイトルからして、“分散分析をする”というのが妥当な推測になりますが、“分散を分析する”のが平均についてなんらかの有用な知識を得る助けになるのが何故か、と

いうのがあまりはっきりしません。実際のところ、これが、人が初めて ANOVA に出会うときに感じる最大の概念的な困難なのです。ANOVA がどのように働くかをみるためには、分散について話を始めるのが最も良いことを私は見つけました。実際、分散を記述する数式を使った、ある種の数学ゲームをプレイしながら話をしたいと思います。すなわち、分散の周りで遊ぶところから始めると、これが興味の対象である平均にとって便利なツールになるということがわかります。

### 11.2.1 $Y$ の分散についての二つの式

まず、いくつかの表記法を導入するところから始めましょう。ここではグループの総数を表すのに、 $G$  を使います。三つの薬に関するデータの場合は、 $G = 3$  群あることになります。次に、全体のサンプルサイズを  $N$  とします；つまり、私たちのデータセットには全部で  $N = 18$  人いることになります。同様に、第  $k$  番目の群にいる人は、 $N_k$  と書くことにします。私たちの仮想的臨床検査の場合、サンプルサイズは三つの群全てにおいて  $N_k = 6$  です。<sup>\*1</sup> 最後に、結果変数を  $Y$  と書くことにします。私たちの例では、 $Y$  は気分の変化です。特に、第  $k$  群の第  $i$  番目のメンバーに生じた気分変化は、 $Y_{ik}$  と書きます。同様に、この実験における 18 人全員について、気分変化の平均を撮ったものを  $\bar{Y}$  とし、第  $k$  群における 6 名の気分変化の平均は、 $\bar{Y}_k$  とします。

さて、これで表記法が出揃ったので、式を書き始めることができます。始めるにあたって、セクション ??で使われていた分散の式を思い出しましょう。私たちが記述統計を行っていた、あの懐かしい日々に戻って。 $Y$  の標本分散は次のように書くのでした。

$$\text{Var}(Y) = \frac{1}{N} \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

この数式はセクション ??の分散の式と、ほとんど同じように見えます。違いは今回二つの添字があることだけです：私は群ごと（すなわち、 $k$  の値）に足し合わせて、それから群内の人（すなわち  $i$  の値）について足し合わせました。これは本当に表面的な違いです。もしここで、サンプルにおけるある人  $p$  の結果変数の値として  $Y_p$  を使っていたら、添字は一つだけになったでしょう。二つの添字を使っている唯一の理由は、私が人を群に分けたからであり、群の中の人の番号を割り当てたからです。

ここで具体的な例を見るとわかりやすいでしょう。 $N = 5$  人の人がそれぞれ  $G = 2$  群に分類された、次の表を見てみましょう。適当に“イケてる”グループを群 1、“イケてない”グループを群 2 としましょう。ここには 3 人のイケてる人 ( $N_1 = 3$ ) と、2 人のイケてない人 ( $N_2 = 2$ ) がいます。

---

<sup>\*1</sup> 全ての群で同じ数の観測がなされている場合、その実験デザインは“バランス（のとれている）”デザインだといいます。バランスはこの章のトピックである、一元配置 ANOVA の時にはそれほど問題になりません。もっと複雑な ANOVA をやろうとすると、より重要な問題になってきます。

名前	人	群	群番号	群の中の番号	ダサさ
	$p$		$k$	$i$	$Y_{ik}$ or $Y_p$
Ann	1	cool	1	1	20
Ben	2	cool	1	2	55
Cat	3	cool	1	3	21
Dan	4	uncool	2	1	91
Egg	5	uncool	2	2	22

ここでは、2種類の異なるラベリング技術を使われていることに注意してください。“人”変数は  $p$  で表ますので、サンプルの中の  $p$  番目の人のダサさを  $Y_p$  として表現することができます。例えば、この表では Dan は 4 番目なので、 $p = 4$  とすることができます。さて、この“Dan”なる人物のダサさ  $Y$  について話すとき、彼がどんな人であったとしても、彼のダサさを  $Y_p = 91$  で  $p = 4$  である、という参照をすることができます。しかし、Dan を参照する方法はこれだけではありません。もう一つの方法として、Dan が“イケてない”グループ ( $k = 2$ ) に所属しており、イケてない群 ( $i = 1$ ) のリストの最初の人だということもできます。ですから同じように Dan のダサさを参照するのに、 $Y_{ik} = 91$  で  $k = 2$ かつ  $i = 1$  ということもできるのです。

言い換えると、各対象者  $p$  が一つの組み合わせ  $k$  に対応しているので、上で挙げた式は実質的に元の分散の式と同じ、つまり次のようにになります。

$$\text{Var}(Y) = \frac{1}{N} \sum_{p=1}^N (Y_p - \bar{Y})^2$$

どちらの式でも、サンプルにおける全ての観測例を足し上げることになります。ほとんどの場合、より単純な表記である  $Y_p$  という書き方をします。 $Y_p$  という式は二つの中で明らかに単純な方の書き方ですね。しかし、ANOVA をするときは各対象者がどちらの群に所属しているかを保持した書き方であることが重要になるため、 $Y_{ik}$  をつかってこれを書き表すこともあるのです。

### 11.2.2 分散から平方和へ

オウケイ、分散がどのように計算されるかを大体掴んだところで、**平方の総和**と呼ばれるものを定義しましょう。表記は  $SS_{tot}$  です。これはとても単純です。分散を計算する時には平均偏差の二乗を平均するわけですが、その代わりにそれを単に足し合わせます。

ですので平方の総和の式は、分散の式とほとんど同じです。

$$SS_{tot} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

ANOVA の文脈で分散を分析することについて話をするときは、実際の分散ではなく平方の総和をつかって実行していることになります。平方の総和を使う利点の一つは、それを異なる二種類の変動に分解することができる点です。

まず、群内平方和について話しましょう。そこでは群平均から個々人がどれほどずれているかを見ることができます。

$$SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

ここで  $\bar{Y}_k$  は群平均です。例えば、 $\bar{Y}_k$  は  $k$  番目の薬を与えられた人による、気分変容の平均値です。ですから、実験に参加した全員の平均と個々人を比較するのではなく、おなじ群にいる人の平均と比較していることになります。結果的に、 $SS_w$  の値は平方和の総和よりも小さくなります。というのも、そこには群の違い、すなわち薬が人の気分に与える影響の違いがあるかどうかを完全に除外しているからです。

次に、群の違いだけを捉えた変動を記述する第 3 の表記を定義しましょう。これは全体平均  $\bar{Y}$  と群平均  $\bar{Y}_k$  の間のずれを見ることになります。

この変動の大きさを評価するために、群間平方和を計算することになります。

$$\begin{aligned} SS_b &= \sum_{k=1}^G \sum_{i=1}^{N_k} (\bar{Y}_k - \bar{Y})^2 \\ &= \sum_{k=1}^G N_k (\bar{Y}_k - \bar{Y})^2 \end{aligned}$$

実験におけるすべての人の中にある全変動  $SS_{tot}$  を示すことはそれほど難しくありません。実際には、群間の変動  $SS_b$  と群内の変動  $SS_w$  を足しあわせるのですすなわち、

$$SS_w + SS_b = SS_{tot}$$

イエイ。

さて、何がわかったのでしょうか？結果変数に伴う変動全体 ( $SS_{tot}$ ) は“異なる群の標本平均の違いに伴う変動”( $SS_b$ ) と、“その残りの変動”( $SS_w$ ) を足し合わせたものに切り分けられるということです。

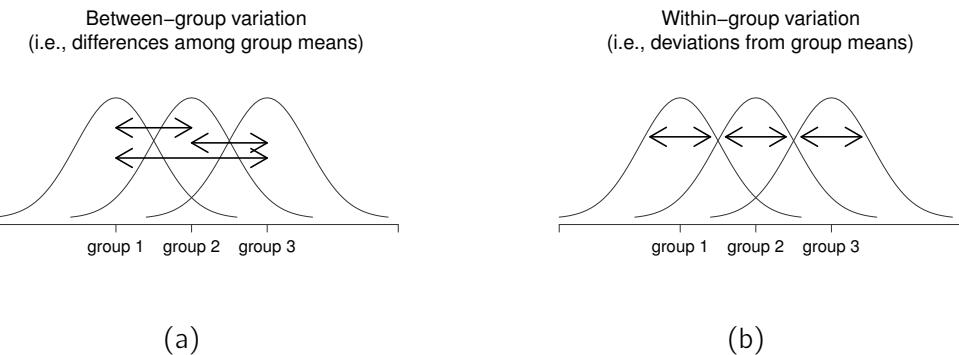


Figure 11.2 “群間” 変動 (パネル a) と “群内” 変動 (パネル b) を図示したものです。左の図では、矢印が群間平均の差を示しています。右図では矢印が群内の変動を強調しています。

す。<sup>\*2</sup>では母平均が群ごとに異なっているかどうかを検証するには、どうすれば良いのでしょうか？ フムウ。待てよ。ちょっとまってください。今思うに、これこそまさに私たちが探していたものです。もし帰無仮説が真であるとすれば、全ての平均はそれぞれほとんど同じものになると思えませんか？ そしてそれが意味するのは、 $SS_b$  が非常に小さい、少なくとも“それ以外の全ての変動”である  $SS_w$  よりは小さいことが期待できるでしょう。ムムッ。仮説検定が始まる予感がします。

### 11.2.3 平方和から F-検定へ

最後のセクションでみたように、ANOVA の背後にある本質的なアイデアは  $SS_b$  と  $SS_w$  という二つの平方和の値を比較するところにあります。群間変動の  $SS_b$  が群内変動  $SS_w$  に比べて大きい時、違う群の母平均は互いに等しいとは言えない、という推論をする根拠をもつことになります。これを実際の帰無仮説検定に変換するために、“ちょっといじくり回す” ことが必要です。最初にお見せするのは、検定統計量として何を計算するのかであり、それは **F 比**です。それからなぜ私たちがそんな風にするのかを理解してもらおうと思います。

平方和の値を F 比に変換するために最初にしなければならないのは、 $SS_b$  と  $SS_w$  に関する**自由度**を計算することです。普通は、自由度はある計算に関わるユニークな“データ点”的の数から、満たす必要のある“制約”的の数を引いたものに対応しています。私たちが計算している群内変動は、群平均 ( $G$  個の制約) の周りにある個々人の観測による変動 ( $N$  データ点) です。対して私たちが興味のある群間変動は、全体平均 (1 個の制約) の周りにある群平均 ( $G$  データ点) の変動です。つまり、こ

---

<sup>\*2</sup> $SS_w$  は ANOVA では誤差、すなわち  $SS_{error}$  と表されることもあります。

この自由度は次のようにになります。

$$\begin{aligned} df_b &= G - 1 \\ df_w &= N - G \end{aligned}$$

オーケー、とても単純ですね。次にすることは、平方和を“平均平方”に変換することで、これは自由度で割ることで計算できます。

$$\begin{aligned} MS_b &= \frac{SS_b}{df_b} \\ MS_w &= \frac{SS_w}{df_w} \end{aligned}$$

最後に  $F$  比を計算するために、群間平均平方を群内平均平方で割ります。

$$F = \frac{MS_b}{MS_w}$$

非常に一般的な意味で、 $F$  統計量の背後にあるものは直感的にわかります。 $F$  の値がより大きいことは、群間変動が群内変動よりも相対的に大きいことを意味します。結果的に、 $F$  の値が大きいことは、帰無仮説に対立するより大きな証拠を得たことになります。しかし、 $F$  がどれぐらい大きければ、実際に  $H_0$  を棄却できるのでしょうか？これを理解するためには、ANOVA が何であるか、平均平方とは何なのかを、もう少し深く理解しなければなりません。次のセクションではこの詳細について説明していきますが、実際に検定が何をしているかに興味のない読者のために、ここではそれを省略しましょう。帰無仮説検定を完成させるために、帰無仮説が真である時の  $F$  の標本分布について知らなければなりません。驚くなれ、帰無仮説のもとでの  $F$  統計量の標本分布は  $F$  分布です。第 ?? 章での  $F$  分布についての議論を思い出してもらいたいのですが、 $F$  分布は二つのパラメータを持っていて、それが二つの自由度に対応しています。最初の自由度  $df_1$  は群間の自由度  $df_b$ 、第二の自由度  $df_2$  は群内の自由度  $df_w$  です。

一元配置分散分析に含まれる、全ての重要な数字の要約を、計算に使った数式とともに、表 ?? に示しました。

#### 11.2.4 データのためのモデルと $F$ の意味

ANOVA の根本的なレベルでは、二つの異なる統計モデルである  $H_0$  と  $H_1$  が競合します。帰無仮説と対立仮説をこのセクションの冒頭で論じた時、これらのモデルが実際に何を表すのかについては少し不完全なまま紹介しました。今からそれを撤回して詳述していきますが、たぶんあなたはそんなことをする私のことが嫌いになるでしょう。思い出してもらいたいのですが、私たちの帰無仮説は全ての群平均は相互に等しいというものでした。もしそうであるなら、結果変数  $Y_{ik}$  について考える自然な方法は、一つの母平均  $\mu$  に、その変動を加えたものとして、個々のスコアを記述

Table11.1 分散分析の中に含まれる重要な全ての数字は，“標準的な”分散分析表の中に組み込まれます。この形式では全ての数字 ( $p$  値と呼ばれる奇妙な数式とコンピュータがないと計算できないものを除いて) が示されています。

	自由度 (df)	平方和 (SS)	平均平方 (MS)	F-統計量	$p$ -値
between groups	$df_b = G - 1$	$SS_b = \sum_{k=1}^G N_k (\bar{Y}_k - \bar{Y})^2$	$MS_b = \frac{SS_b}{df_b}$	$F = \frac{MS_b}{MS_w}$	[complicated]
within groups	$df_w = N - G$	$SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$MS_w = \frac{SS_w}{df_w}$	-	-
.....					

することです。この変動というのは普通  $\varepsilon_{ik}$  で表され、伝統的にそれは観測に伴う誤差とか**残差**と呼ばれています。でも気を付けてください。“有意差”という言葉を見た時と同じように、“エラー(誤差)”には統計における専門的な意味を持っていて、日常的な定義とは異なります。日常用語で言う“エラー”は、ちょっとした失敗と言う意味ですが、統計的にはそうではありません(少なくとも、必ずしもそうではありません)。それを考慮すると、“残差”的方が“エラー(誤差)”よりも良い言葉かもしれません。統計学ではどちらの言葉も、“残っている変動”，すなわちこのモデルが説明できない“あまりもの”を意味しています。

いずれにせよ、帰無仮説を統計モデルとして書き表すと、こんな感じになります。

$$Y_{ik} = \mu + \varepsilon_{ik}$$

ここで(後ほど議論する)ある仮説を立てます。それは残差の値  $\varepsilon_{ik}$  は正規分布に従うというもので、その平均は 0、標準偏差  $\sigma$  はすべての群を通じて同じであるというものです。第 ?? 章で導入した表記法を使うと、この仮定は次のように書くことができます。

$$\varepsilon_{ik} \sim \text{Normal}(0, \sigma^2)$$

対立仮説  $H_1$  についてはどうでしょうか? 帰無仮説と対立仮説の唯一の違いは、各群は異なる母平均を持ちうるということです。ですから、実験における第  $k$  番目の群の母平均を  $\mu_k$  とするなら、 $H_1$  に対応する統計モデルは次のようになります。

$$Y_{ik} = \mu_k + \varepsilon_{ik}$$

ここでもう一度、誤差項が平均 0、標準偏差  $\sigma$  の正規分布に従うと仮定します。すなわち、対立仮説もまた次のように書けるわけです。

オウケイ、さあ  $H_0$  と  $H_1$  を支える統計モデルについて記述してきましたが、そろそろ平均平方の値が何を測っているのか、そして  $F$  の解釈にそれがどう関係するのかを、もう少しありうべき段階に来ました。照明を初めてあなたを退屈がらせることはしませんが、群内平均平方である  $MS_w$  が、誤差分散  $\sigma^2$  の推定量（専門的な意味は、第 ?? 章を参照）とみなすことができる事がわかります。群間平均平方  $MS_b$  も推定量ですが、それが推定しているのは誤差分散に加えて、群平均間の真の差分についての量も入っています。もしこの量を  $Q$  と書くなら、 $F$  統計量は基本的に次のようになっています<sup>a</sup>。

$$F = \frac{\hat{Q} + \hat{\sigma}^2}{\hat{\sigma}^2}$$

ここで、もし帰無仮説が正しければ、真の値は  $Q = 0$  ですし、もし対立仮説が正しければ  $Q > 0$  でしょう（Hays 1994）。つまり、少なくとも  $F$  の値は 1 よりも大きくならなければ、帰無仮説を棄却する可能性がないのです。これは、 $F$  値が 1 よりも小さくなることがないことを意味するのではないことに注意してください。これが意味するのは、帰無仮説が真であれば、 $F$  比の標本分布の平均は 1<sup>b</sup> であり、安全に帰無仮説を棄却するためには  $F$  の値は 1 より大きくなければならないということです。標本分布についてもう少し正確にいうと、帰無仮説が正しい時、 $MS_b$  も  $MS_w$  も残差  $\varepsilon_{ik}$  の分散推定量であることに注意してください。もしこれらの残渣が正規分布に従っていたとする、 $\varepsilon_{ik}$  の分散の推定値がカイ二乗分布に従うのではないか、と考えるかもしれません。というのも（セクション ?? で論じたように）、それこそカイ二乗分布そのものだからです。すなわち、正規分布するものを二乗して足し合させたものだからです。そして、 $F$  分布は（もう一度、その定義に戻りますが）二つの  $\chi^2$  分布の比を取ったものであり、それが私たちの標本分布なわけです。明らかに、私は多少これをいうときに大袈裟に表現していますが、大まかに言って、本当に私たちの標本分布からきたものになります。

<sup>a</sup> 第 ?? 章まで読めば、要因の水準  $k$  における“操作の影響”がどのように  $\alpha_k$  の値で定義されるかを目にするでしょう（セクション ??）。そこでは、 $Q$  は重みづけられた操作の影響の平方、つまり  $Q = (\sum_{k=1}^G N_k \alpha_k^2) / (G - 1)$  となっていることがわかります。

<sup>b</sup> あるいは、もし正確さにこだわるのであれば、 $1 + \frac{2}{df_2 - 2}$ .

### 11.2.5 実例

ここまで議論はかなり抽象的で、少し理論的な話でしたので、今度は実際の例を見ることでもう少し有用な話をみていくと思います。そのために、私がこの章のために導入した臨床試験データに戻りましょう。最初に計算した記述統計量からわかるのは、群平均でした。つまり、気分の向上はプラセボで 0.45 高、Anxitriptaline で 0.72、Jozepam で 1.48 でした。それを念頭に置いて、1899 年のよ

うなパーティーをしましょう<sup>\*3</sup>，そして紙と鉛筆で計算をするのです。最初の五人についてだけ計算することにします。だって地獄の 1899 年じゃないし，私はとっても怠け者ですから。群内平方和， $SS_w$  を計算するところから始めるとしましょう。まず，計算しやすくするために便利な表を作るこ

群 <i>k</i>	アウトカム $Y_{ik}$
プラセボ	0.5
プラセボ	0.3
プラセボ	0.1
anxitfree	0.6
anxitfree	0.4

この段階で表に含めることができるのは，ロウデータそのものでしかありません。つまり群化変数（すなわち `drug`）と結果変数（すなわち `mood.gain`）を各人についてのものだけ，です。結果変数が私たちの式表現によると， $Y_{ik}$  に対応していることに注意してください。次の計算ステップはこの研究の一人におけるひとりについて，対応する群平均  $\bar{Y}_k$  に書き下すことです。これは少し繰り返しになりますが、記述統計を行う際にグループの平均値を計算しているので、特に難しいことではありません。

群 <i>k</i>	アウトカム $Y_{ik}$	群平均 $\bar{Y}_k$
プラセボ	0.5	<b>0.45</b>
プラセボ	0.3	<b>0.45</b>
プラセボ	0.1	<b>0.45</b>
anxitfree	0.6	<b>0.72</b>
anxitfree	0.4	<b>0.72</b>

さて，これらを書き出したので，再び一人一人について，対応する群平均からのズレを計算する必要があります。つまり， $Y_{ik} - \bar{Y}_k$  の引き算をしたいのです。そうした上で，それらを二乗します。そうすると，ここにあるような数字を得ることができます。：

群 <i>k</i>	結果 $Y_{ik}$	群平均 $\bar{Y}_k$	群平均からの偏差 $Y_{ik} - \bar{Y}_k$	偏差の平方 $(Y_{ik} - \bar{Y}_k)^2$
プラセボ	0.5	0.45	<b>0.05</b>	<b>0.0025</b>
プラセボ	0.3	0.45	<b>-0.15</b>	<b>0.0225</b>
プラセボ	0.1	0.45	<b>-0.35</b>	<b>0.1225</b>
anxitfree	0.6	0.72	<b>-0.12</b>	<b>0.0136</b>
anxitfree	0.4	0.72	<b>-0.32</b>	<b>0.1003</b>

最後のステップは結構ストレートです。群内平方和を計算するために，全ての観測に対して偏差平

<sup>\*3</sup>あるいは，もう少し正確に，“1899 年の，友達もいないし 1899 年にはなんの意味もなかった計算をするしかないことをしましょう。だって ANOVA は 1920 年代になるまで出てこないんですから”

方を足し合わせて行きます。

$$\begin{aligned} SS_w &= 0.0025 + 0.0225 + 0.1225 + 0.0136 + 0.1003 \\ &= 0.2614 \end{aligned}$$

もちろん、実際にはデータセットの中にある全 18 人分についてこの計算をした、正しい答えが欲しいわけです。最初の 5 人分だけじゃなくてね。そうしたければ、神とペンを使って計算し続けてもらってもいいんですけど、それはちょっと面倒ですよね。その代わりとして、専用のスプレッドシート・アプリである OpenOffice や Excel を使うと随分と楽に計算できます。是非ご自身でやってみてください。そうすると、群内平方和の値が 1.39 になるはずです。

オーケイ。群内分散  $SS_w$  の計算が終わったので、群間平方和、 $SS_b$  に向き合う時が来ました。この計算は非常によく似ています。大きな違いは、全ての観測に対して群平均  $\bar{Y}_k$  と観測値  $Y_{ik}$  の差分を計算する代わりに、全体平均  $\bar{Y}$ (この回は 0.88 ですが) と群平均  $\bar{Y}_k$  の差を全ての群に対して計算するところです。

群 <i>k</i>	群平均 $\bar{Y}_k$	全体平均 $\bar{Y}$	偏差 $\bar{Y}_k - \bar{Y}$	偏差平方 $(\bar{Y}_k - \bar{Y})^2$
placebo	0.45	0.88	-0.43	0.19
anxitfree	0.72	0.88	-0.16	0.03
joyzepam	1.48	0.88	0.60	0.36

ただし、群間の計算をするときは、この偏差平方にその群に含まれる観測度数  $N_k$  をかけなければなりません。というのも、ある群における観測はどれも(すべての  $N_k$  について)，群間の差を作るのに貢献しているからです。ですから、プラセボ群に 6 人いて、プラセボ群平均の全体平均からの差が 0.19 であれば、全体としてこの 6 人が関わった群間分散は  $6 \times 0.19 = 1.14$  になります。ということで、計算のための表を少し拡張しなければなりません。

群 <i>k</i>	...	偏差平方 $(\bar{Y}_k - \bar{Y})^2$	サンプルサイズ $N_k$	重み付き偏差平方 $N_k(\bar{Y}_k - \bar{Y})^2$
placebo	...	0.19	6	1.14
anxitfree	...	0.03	6	0.18
joyzepam	...	0.36	6	2.16

そして、今やこの研究における全ての群に関する群間平方和が、“重み付き偏差平方”を足し合せることによって得られるわけです。

$$\begin{aligned} SS_b &= 1.14 + 0.18 + 2.16 \\ &= 3.48 \end{aligned}$$

みてきたように、群間の計算は短かったです。さて、平方語の計算は  $SS_b$  and  $SS_w$ 、になりましたので、ANOVA の残りはたいしたことありません。次のステップは自由度の計算です。私たちのデータは  $G = 3$  の群、そして  $N = 18$  人の観測度数をもっていますから、全体の自由度は簡単な引き算で算出できます。

$$\begin{aligned} df_b &= G - 1 = 2 \\ df_w &= N - G = 15 \end{aligned}$$

次に、群内分散と群間分散それぞれについて、平方和と自由度を計算したのですから、前者を後者で割ることで平均平方が計算できますね。

$$\begin{aligned} MS_b &= \frac{SS_b}{df_b} = \frac{3.48}{2} = 1.74 \\ MS_w &= \frac{SS_w}{df_w} = \frac{1.39}{15} = 0.09 \end{aligned}$$

ほとんど終わったようなもんです。この平均平方は、我々の興味ある検定統計量である  $F$  値を計算するのに使われます。群間  $MS$  の値を群内  $MS$  の値で割ることによってその数字が得られます。

$$F = \frac{MS_b}{MS_w} = \frac{1.74}{0.09} = 19.3$$

イヤッホーイ！ 興奮してきましたね？さてこの検定統計量を手に入れたわけですから、最後のステップはこの検定が有意な結果を出してくれたかどうかを見極めることです。第 ?? 章で議論したように、“昔”に戻って統計のテキストを開いたり、後ろの方のセクションに載ってる大きな表のペラペラめくると、特定のアルファの値（帰無仮説が棄却される範囲）、たとえば 0.05, 0.01, 0.001 など、に対応した、自由度 2 と 15 の  $F$  値の閾値を見つけることができます。

そうすると、アルファが 0.001 のときの  $F$  の閾値は 11.34 であることがわかりました。これは私たちが計算した  $F$  よりも小さいので、 $p < 0.001$  と言うことができます。しかしこれは昔ながらのやり方であって、今では賢い統計ソフトウェアがあなたのために正確な  $p$  値を計算してくれますよ。実際、正確な  $p$  値は 0.000071 になります。さて、タイプ I のエラー率について極めて保守的であったとしても、我々は帰無仮説を棄却するのに十分な保証が得られたと言えるでしょう。この時点で、基本的に終了です。計算を終えたらば、これら全ての数字を表 ?? のようにして、分散分析表にまとめるのが伝統的なやり方です。我々の臨床試験データについての、分散分析表は次のようになります。

	自由度	平方和	平均平方	$F$ -統計量	$p$ -値
群間	2	3.48	1.74	19.3	0.000071
群内	15	1.39	0.09	-	-

今日では、こうした表を自分で作る理由がないように思うかもしれません、ほとんどの統計ソフトウェア（JASP もそうです）は ANOVA の結果をこうした表にまとめる傾向があることに気づくと思います。なので、読み方に慣れていた方がいいでしょう。とはいえ、ソフトウェアが完全な分散分析表を出力してくれますが、あなたが書くときに全部を含める理由はあんまりありません。この結果をレポートする標準的な方法は、次のように書くことです。

一言配置の分散分析では、投薬による気分の向上の有意な効果が示された ( $F(2, 15) = 19.3, p < .001$ )。

ふー。短い一文を書くために、たいした苦労をするもんだ。

## 11.3

### ANOVA を JASP で実行する

あなたがこのセクションの最後を読んだときに、特にあなたが私のアドバイスに従って紙とペンで(あるいはスプレッドシートで)計算したのであれば、どんなふうに感じるかは手に取るようにわかります。ANOVA の計算自分でやることは最悪です。手順通りに計算する量がとてもおおいので、ANOVA をしようと思うたびに何度も何度も計算するのは面倒なのです。

#### 11.3.1 あなたの ANOVA のための JASP

あなたの人生を楽にするために、JASP は ANOVA を実行してくれます.... 万歳！ ‘ANOVA’-‘ANOVA’ 分析、とすすみ、`mood.gain` 変数を‘従属変数’ボックスに移動させ、それから `drug` 変数を‘固定効果’ボックスに動かします。こうすると図??のような結果が示されます。<sup>\*4</sup>‘追加オプション’の‘効果量の推定’の下にある、 $\eta^2$  チェックボックスにもチェックを入れたので、結果の表にもそれが反映されています。効果量についてはあとで触れることにします。

#### ANOVA

ANOVA - mood.gain						
Cases	Sum of Squares	df	Mean Square	F	p	$\eta^2$
drug	3.453	2.000	1.727	18.611	< .001	0.713
Residual	1.392	15.000	0.093			

Note. Type III Sum of Squares

Figure11.3 気分の向上と投薬の関係についての JASP による分散分析表

JASP の結果の表には、平方和、自由度、そのほか今は興味のないいくつかの統計量が示されます。しかし、JASP には“群間”とか“群内”といった表示はしていません。その代わりに、もっと意味のある名前がついています。今回の例では、群間分散は投薬がアウトカム変数に及ぼした影響に対

<sup>\*4</sup>JASP の結果は、丸め誤差のあった上の文章で述べたものより正確です。

応しており、群内分散は残差誤差とも呼ばれる“残った”変動分に対応しています。これらの数字を、セクション ??で手計算した数字と比較すると、四捨五入によるズレを除いてほぼ同じであることがわかると思います。群間平方和は  $SS_b = 3.453$  で、群内平方和は  $SS_w = 1.392$ 、そしてそれぞれの自由度は 2 と 15 です。 $F$ -値と  $p$ -値も計算されていて、それも四捨五入によるズレを除いて、先程の長く面倒な方法で計算したものと同じになっていることがわかります。

JASP の結果の表には、平方和、自由度、そのほか今は興味のないいくつかの統計量が示されます。しかし、JASP には“群間”とか“群内”といった表示はしていません。その代わりに、もっと意味のある名前がついています。今回の例では、群間分散は投薬がアウトカム変数に及ぼした影響に対応しており、群内分散は残差誤差とも呼ばれる“残った”変動分に対応しています。これらの数字を、セクション ??で手計算した数字と比較すると、四捨五入によるズレを除いてほぼ同じであることがわかると思います。群間平方和は  $SS_b = 3.453$  で、群内平方和は  $SS_w = 1.392$ 、そしてそれぞれの自由度は 2 と 15 です。 $F$ -値と  $p$ -値も計算されていて、それも四捨五入によるズレを除いて、先程の長く面倒な方法で計算したものと同じになっていることがわかります。

## 11.4

---

### 効果量

ANOVAにおいて効果量を測定するには、二つの異なる方法がありますが、最も一般的に使われているのは  $\eta^2$  (**eta squared**) と偏  $\eta^2$  です。一要因の分散分析では、どちらも同じになりますので、今はとりあえず  $\eta^2$  について説明します。 $\eta^2$  は実際のサンプルについて、次のように定義されます。

$$\eta^2 = \frac{SS_b}{SS_{tot}}$$

これだけです。図 ??の分散分析表についてみてみると、 $SS_b = 3.45$  で  $SS_{tot} = 3.45 + 1.39 = 4.84$  です。ここから、 $\eta^2$  の値は次のように計算できます。

$$\eta^2 = \frac{3.45}{4.84} = 0.71$$

$\eta^2$  の解釈も、実に直接的です。アウトカム変数 (`mood.gain`) において、予測変数 (`drug`) が説明する分散の比率をあらわしているわけです。値が  $\eta^2 = 0$  であれば、両者になんの関係もないことを表しますし、 $\eta^2 = 1$  であれば完璧な関係にあることになります。さらに良いことに、 $\eta^2$  の値は、セクション ??)で説明した  $R^2$  にかなり近いもので、それと同じように解釈することができます。

多くの統計的教科書では  $\eta^2$  を ANOVA における基本的な効果量の測度だと説明していますが、Daniel Lakens がブログで興味深いことを言っています。それによると、 $\eta^2$  はデータ分析業界における最適な効果量測度ではない、なぜなら推定量のバイアスがあるからだ、ということです (<http://daniellakens.blogspot.com.au/2015/06/why-you-should-use-omega-squared.html>)。

ありがたいことに、JASP にはオメガの二乗 ( $\omega^2$ ) という選択肢もついていて、これはよりバイアスが少なく、イータの二乗と並んで使われているものです。

## 11.5

---

### 多重比較と事後の (Post hoc) 検定

2 群以上で ANOVA をして有意な影響をみたときは、実際はどの群がどの群と差があったのか知りたくなるでしょう。投薬の例では、帰無仮説は三つの薬 (プラセボと Anxifree と Joyzepam) が気分に与える影響は同じというものでした。しかし考えてみると、帰無仮説は実際には三つの異なることを同時に主張しているのです。つまり、主張は次の通りです。

- あなたの競争相手である薬 (Anxifree) はプラセボと変わりない (つまり,  $\mu_A = \mu_P$ )。
- あなたの薬 (Joyzepam) はプラセボと変わりない (つまり,  $\mu_J = \mu_P$ )。
- Anxifree と Joyzepam の効果は同じぐらい (i.e.,  $\mu_J = \mu_A$ )。

この三つの主張のどれかが偽であれば、帰無仮説も偽になります。ですから、我々が帰無仮説を棄却するとき、このなかの少なくとも一つは真であることになります。でもどれでしょう？ 三つの命題全てに興味がありますよね。あなたが本当に知りたいのは、あなたの新薬 Joyzepam がプラセボよりも良いはずだ、というものですから、既存の一般的な代替品 (つまり Anxifree) と比べてどの程度の効果があるのかを知っておくのは良いことでしょう。Anxifree がプラセボにくらべてどれぐらい効果があるのかをチェックするのもまた、いいことのはずです。Anxifree は既に他の研究者によって、プラセボに対する効果の検証がしっかり行われているはずですが、あなたの研究でも先行研究と同じ結果が出ることを示すかどうかのチェックをするのも、いいことのはずです。

この三つの異なる命題を使って帰無仮説を特徴付けるとしたら、8 つのあり得た“世界の状態”を区別する必要があります。

可能性:	$\mu_P = \mu_A?$	$\mu_P = \mu_J?$	$\mu_A = \mu_J?$	どの仮説か？
1	✓	✓	✓	帰無仮説
2	✓	✓		対立仮説
3	✓		✓	対立仮説
4	✓			対立仮説
5		✓	✓	対立仮説
6		✓		対立仮説
7			✓	対立仮説
8				対立仮説

帰無仮説を棄却することで、我々は#1 が真である世界を信じることはできないと決めたわけです。次の質問は、残る 7 つの可能性のうち我々が正しいと考えることができるものはどれか？ ということです。この状況に置かれた時にも、データを見るのが助けになります。例えば、図 ?? をみると、Joyzepam はプラセボや Anxifree より良いようですが、Anxifree とプラセボに実質的な違いはないように思えます。しかし、これにはっきりと答えるのは難しくないので、いくつかの検定をして助けてもらうことにします。

#### 11.5.1 “ペアごとの pairwise” t-検定

さて問題解決のために、何をすべきでしょう？ それぞれの平均のペア（プラセボ vs. Anxifree, プラセボ vs. Joyzepam, Anxifree vs. Joyzepam）は既にあるのですから、それぞれに t 検定をしてどうなるかみてみるとどうでしょう？ それを JASP ではするのは簡単です。ANOVA の ‘事後の検定’ オプションへ行き、‘drug’ 変数を右側のアクティブボックスに動かします。するとすぐに、drug の三つのレベルに対して、ペアごとの t-検定全てが表示されます。図 ?? のように。

#### Post Hoc Tests

Post Hoc Comparisons – drug

		Mean Difference	SE	t	P <sub>bonf</sub>	P <sub>holm</sub>
anxitfree	joyzepam	-0.767	0.176	-4.360	0.002	0.001
	placebo	0.267	0.176	1.516	0.451	0.150
joyzepam	placebo	1.033	0.176	5.876	< .001	< .001

Figure11.4 JASP による事後のペアごとの t-検定結果

#### 11.5.2 多重検定のための補正

前のセクションでは、ここで問題についてたくさんの t 検定で対応するというヒントを与えました。これらの分析を実行するときに懸念されるのは、“釣り探検”に出かけてしまったのではないかということです。何か有意になるんじゃないかと期待しながら、理論的なガイダンスなしにたくさんのたくさんの検定を行ってしまいました。このような理論無視の群間比較は、**事後検定 post hoc**

**analysis** と言われます (“post hoc” はラテン語で “after this” という意味です)。<sup>5</sup>

事後の検定をするのはいいのですが、注意が必要なのです。例えば、前のセクションでやったような分析は、それぞれ個別に行われた  $t$ -検定ですが、これは 5% のタイプ I エラー (つまり  $\alpha = 0.5$ ) で三つの検定を行ったことになります。もし私の ANOVA が 10 群について行われていたのだとしたら、わたしは 45 回の “事後の”  $t$  検定をして、ある一つの水準が他の一つと有意に異なるかどうか、それぞれ検証することになります。あなたはそのうちの 2 つか 3 つぐらいが、偶然有意になってしまふかも、と思うかもしれません。第 ?? 章でみたように、帰無仮説検定の背後にある原則は、タイプ I エラーをコントロールしたいというものであったのですが、今や私はたくさんの  $t$  検定を一度に、ANOVA の結果に基づいて判断する目的で実施しており、この時、一連の検定を通じた実際のタイプ I エラーは、完全に制御不能になっています。

この問題についての一般的な解決策は、 $p$  値を調整することです。これは一連の検定全体を通じたエラー発生率をコントロールすることが目的です。(Shaffer1995)。事後の検定をする時は、普通はこのやり方で補正され (いつもではありません)，この手続きは**多重比較の補正**と呼ばれます。時には “同時推論の補正” と言われることもあります。いずれにせよ、この補正のやり方にはいくつかの異なる方法があります。このセクションとセクション ??では、これらのいくつかについて取り上げますが、多くの他のやり方があるんだということは知っておいてください。(Hsu1996)

#### 11.5.3 Bonferroni の補正

最も単純な補正是 **Bonferroni の補正**と呼ばれるもので、それは実際にとーーーーーってもシンプルなものです。(Dunn1961)  $m$  個それぞれの検定をするような事後分析を想像してみましょう。どれかがタイプ I エラーを引き起こす確率の合計が、最大でも  $\alpha$  になることを保証するようにしたいと思っています。<sup>6</sup> このとき、Bonferroni の補正是単に “あなたのものとの  $p$  値を  $m$  で割れ” というだけです。もし元の  $p$  値を表すのに  $p$  と表記するなら、 $p'_j$  を補正した値の書き方として、Bonferroni の補正是次のようにになります。：

$$p' = m \times p$$

ですから、もしあなたが Bonferroni の補正を使おうとするのなら、 $p' < \alpha$  の時に帰無仮説を棄却するようにしてください。この補正の裏にあるロジックはとても直接的ですよね。 $m$  個の異なる検定をするなら、各検定のタイプ I エラーは大きくても  $\alpha/m$  になるはずで、全体的なタイプ I エラーは  $\alpha$  以上になり得ないのでから。これはとても単純なことなので、元の論文で筆者は次のように書いています。

<sup>5</sup> もしあなたが、ある箇所を比較して他のところはしない、という理論的な基盤を持っていれば、話は違ってきます。そういうときは、あなたは “post hoc な” 分析をしようとしているのではなくて、“計画された比較” をしようとしているわけです。こういう状況については本書の後半 (セクション ??) で行いますが、今は話を単純化しておきましょう。

<sup>6</sup> すべての調整法がそうしようとしているわけではない、ということに注意しておきましょう。ここで私が述べているのは、“ファミリーワイズのタイプ I エラー” をコントロールするアプローチというやつです。しかし、他の事後分析では、“偽検出率” のコントロールを目指しているものもあり、ちょっと違うものです

ここで述べた手法はとてもシンプルで一般的なものですから、以前にも誰かが使っていたでしょう。しかし、先行例を見つけることはできませんでしたので、おそらくこの超単純さのせいで、統計学者はいい方法だと気づけなかったのだと思います。  
**(Dunn1961)**

Bonferroni の補正を JASP で使うためには、‘補正’オプションの‘Bonferroni’のチェックボックスをクリックします。そうすると、ANOVA の結果の表の中に、Bonferroni の方法で補正された  $p$  値の列を見つけることができるでしょう (??)。

#### 11.5.4 Holm の補正

Bonferroni の補正がとてもシンプルなものだったわけですが、それが常にベストなものというわけではありません。代わりによく使われるのが、**Holm の補正**というものです (Holm1979)。Holm の補正のアイデアは、あなたが検定を順番にやっていく時に、最も小さい(元の) $p$  値から初めて、最大のものに進んでいくというものです。第  $j$  番目の大きさを持つ  $p$  値は、次のいずれかになります。

$$p'_j = j \times p_j$$

(つまり、最大の  $p$  値は変化させないままにしておいて、二番目に大きな  $p$  値は 2 倍に、三番目に大きな  $p$  値は 3 倍に… というふうにしていきます)。あるいはまた、

$$p'_j = p'_{j+1}$$

どれか一つでも 大きくなったときにこうします。これはちょっと混乱させるような書き方ですので、もう少しゆっくり説明しましょう。Holm の補正が何をするか、というのは次の通りです。まず、 $p$  値を小さいものから大きいものへと、並べ替えてください。一番小さな  $p$  値について、 $m$  倍して終わりです。しかし、他のどれも二段階プロセスを経ていませんね。たとえば、もしあなたが二番目に小さな  $p$  値を動かしたら、それを  $m - 1$  倍しなければなりません。このかけられた数が、あなたが最後に手に入れた補正された  $p$  値よりも大きければ、取っておきましょう。しかし最後のそれよりも小さければ、最後の  $p$  値をコピーします。これがどういう働きをするかを見るために、次の表を見てください。ここには 5 つの  $p$  値についての Holm の補正計算が示されています。

元の $p$	順序 $j$	$p \times j$	Holm $p$
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057
.103	1	.103	.103

これでわかったでしょうか？

少し計算が面倒ですが、Holm の補正はいくつかの良い特性を持っています。Bonferroni よりも良く（つまり、タイプIIエラーがより低く），直感に反してタイプIエラーについては同じなのです。結果として、実践ではよりシンプルな Bonferroni の補正を使う理由がなくなります。常に、より洗練された Holm の補正が効率的だからです。ですから、あなたが多重比較の補正をする時は Holm 法でいくべきでしょう。図??には Bonferroni と Holm の補正された  $p$  値が示されています。

#### 11.5.5 事後検定の記載

最後に、どの群が他と比べて有意に異なっていたかを決める事後分析を行ったら、あなたが書くだろう結果の文章はこんなふうになります。：

事後検定 (Holm の補正された  $p$  を使った) では、Joyzepam が Anxifree ( $p = 0.001$ ) と プラセボ (\$  $p < 0.001$  \$) よりも有意に大きな気分変容をもたらすことが示された。Anxifree はプラセボに比べて良いという証拠は見つからなかった ( $p = .15$ )。

あるいはもし、あなたが  $p < 0.001$  とだけ書くのは嫌だというのであれば、「設定’-‘結果’にいき、‘正確な  $p$  値を表示する’を選択しておけば、正確な  $p$  値を計算することができます。どちらにせよ、あなたが使った Holm の補正による調整済み  $p$  値を使ったことを書いておくことが肝要です。そしてもちろん、既に関係する記述統計量（すなわち、群平均や標準偏差）をどこかに書いてあることを想定しています。だって  $p$  値だけではほとんど情報がないですからね。

## 11.6

---

### 一要因 ANOVA の仮定

あらゆる統計的検定と同じように、分散分析もデータについて、特にその残差についての仮定の上に成り立っています。知っておくべき仮定は次のとおりです。：正規性、分散の均一性、独立性

セクション ??のことを思い出して欲しいのですが、全体を読んでいなくても、せめて斜め読みぐらいはして欲しいのですが、私は ANOVA を支える統計モデルについて次のように説明したのでした。

$$\begin{aligned} H_0 : Y_{ik} &= \mu + \varepsilon_{ik} \\ H_1 : Y_{ik} &= \mu_k + \varepsilon_{ik} \end{aligned}$$

これらの式で、 $\mu$  は一つの全体平均を表していてすべての群を通じて同じものです。また  $\mu_k$  は第

$k$  番目の群の母平均を表しています。ここで注目しなければならないのは、我々のデータが一つの全体平均で表現できる（帰無仮説）のか、異なる群特有の平均値があるのか（対立仮説）ということです。これはもちろん、実際の研究仮説にとって重要なことです！しかし、検定の手続きはいずれも、暗に、残差についてもある仮定を置いていて、そこでは  $\varepsilon_{ik}$  が次のようにになっているのです。

$$\varepsilon_{ik} \sim \text{Normal}(0, \sigma^2)$$

このちょっとした仕掛けがないと、数式がうまく働かないのです。つまり、正確にいうなら、計算して最後に  $F$  統計量を出すことはできますが、 $F$  統計量が実際にあなたが測ろうとしたものちゃんと測っていたかどうか保証できず、 $F$  検定に基づいて引き出したものが間違っていることになるのです。

さて、では残差についての仮定が正しいかどうかをどうやってチェックしたらいいでしょう？ そうですね、上で述べたように、一つの文章には三つの要素がふめ込まれているので、個別に対応することを考えましょう。

- **分散の均一性。** 私たちは母標準偏差（すなわち  $\sigma$ ）について一つの値しか用意していないことに注意してください。各群に個別の値（つまり  $\sigma_k$ ）を考えることもできるのです。これは分散の均一性の仮定として知られています（等分散性ということもあります）。ANOVA は母標準偏差について、すべての群で同じであるという仮定をしているのです。これについてはセクション ??で大々的に論じます。
- **正規性。** 残差は正規分布することが仮定されています。セクション ??で見たように、これは QQ プロットをみることで（あるいは Shapiro-Wilk 検定をすることで）検査できます。ANOVA の文脈におけるこの話は、セクション ??で論じます。
- **独立性。** 独立性の仮定は少しトリッキーです。これが意味することは基本的に、ある残差について知っていても、それは他のどんな残差についてなにも語らないというものです。すべての  $\varepsilon_{ik}$  値は他のどの残差について、どんな“配慮”もしないし、“関係”も持たないことが仮定されています。これを検証する単純明快な方法というのではないですが、この仮定を明らかに満たさない状況というのはあるのです。例えば、もしあなたが反復測定デザインをしているとすると、各被験者は二つ以上の状況に晒されるわけですが、この時独立性は保持されていません。いくつかの観測値関係に何らかの関係があるのは、同じ人に対応しているのですから明らかです！ この時は、反復測定 ANOVA（セクション ??を参照）のような手法を使わなければなりません。

#### 11.6.1 分散の均一性についての仮定をチェックする

分散について予備的な検定を行うことは、波が船の定期便が出向するのに適した状態

かどうかを確認するために、手漕ぎボートで海に出かけるようなものだ。

– George Box (**Box1953**)

諺にあるように、猫の皮を剥ぐ方法は一つではありませんし、分散の均一性の過程を献呈するものいくつかの方法があります(何らかの理由で、誰もそれを口にしませんが)。私がこの話で見たことのある、最も一般的に使われる検定は、**Levene の検定 (Levene1960)** で、これは **Brown-Forsythe 検定 (BrownForsythe1974)** に関わりのある方法です。

標準的な Levene 検定や Brown-Forsythe 検定をやると、どちらであれ検定統計量として  $F$  あるいは  $W$  で表されるものが出てきますが、これは  $Y_{ik}$  の代わりに  $Z_{ik}$  を使うだけで、あとは普通の ANOVA で計算される  $F$  統計量と同じやり方で計算されます。これを念頭において、JASP でこの検定をどうやるかを見ていきましょう。

Levene 検定は本当にシンプルなんです。アウトカム変数として  $Y_{ik}$  があるとしましょう。新しい変数として、 $Z_{ik}$  を定義します。これは群平均からの偏差の絶対値で、次のように定義されます。

$$Z_{ik} = |Y_{ik} - \bar{Y}_k|$$

オーケイ、こうすることで何がいいんでしょう？では  $Z_{ik}$  が実際に何を意味していて、我々は何を検定しようとしているのかを考えていきましょう。 $Z_{ik}$  の値は第  $k$  群における  $i$  番目の観測がその群平均からどの程度離れているかの測度です。そしてここでの帰無仮説は、群が同じ分散を持っている、すなわち群平均からの全体的な偏差が同じであるというものでした。ですから、Levene 検定における帰無仮説は、 $Z$  の母平均が全群で同じであるというものです。ふむう。ところで私たちが今知りたいのは、全部の群平均が同じであるという帰無仮説の統計的な検定でした。どこかでみたことありましたっけ？そうです、これこそ ANOVA で、Levene 検定は新しい変数  $Z_{ik}$  について ANOVA をすることそのものなのです。

Brown-Forsythe 検定のほうはどうでしょうか？何が違うんでしょう？何にもです。Levene 検定との唯一の違いは、変換された変数  $Z$  の作られ方で、これが少し違います。群の平均ではなく群の中央値からの偏差を使うだけです。つまり、Brown-Forsythe 検定は

$$Z_{ik} = |Y_{ik} - \text{median}_k(Y)|$$

ここで  $\text{median}_k(Y)$  は群  $k$  の中央値です。

### 11.6.2 Levene 検定を JASP で行う

オーケイ、ではどうやって Levene 検定をすれば良いでしょうか。本当に簡単なんですが - ANOVA

の下にある‘仮定のチェック’オプション、この‘均一性の検定’チェックボックスをクリックするだけです。そうして結果を見てみると、図??に示しましたが、検定結果が非有意 ( $F_{2,15} = 1.45, p = .266$ ) であることがわかります。ですから、分散の均一性の過程は満たされている、と言えそうです。しかし、外見だけでは騙されます！もしサンプルサイズが大きかったら、ANOVA の頑健性に問題を与えるような分散の均一性の仮定が破られていない時でも、Levene 検定は有意な効果を示しうるのです。これは上の引用にある George Box が指摘したことです。同様に、もしサンプルサイズがとても小さければ、分散の均一性の過程は満たされず、Levene 検定も非有意 (i.e.  $p > .05$ ) になることがあるかもしれません。これが意味するのはつまり、仮定についてのあらゆる統計的検定は、群/カテゴリーごとの平均周りにある標準偏差をプロットしてみないとわからない、ということです…それが似通っている（つまり分散が均一である）かどうかを見ないとね。

### Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
1.450	2.000	15.000	0.266

Figure11.5 JASP における一要因 ANOVA の Levene 検定結果出力

#### 11.6.3 分散の均一性についての仮定を取り除く

今回の例では、分散の均一性の仮定は大丈夫だったことがわかりました。Levene 検定は非有意（標準偏差のプロットも見ながら）で、心配することはないようです。しかし、実際にはこんな幸運ばかりではありませんよね。分散の均一性の仮定が破られた時、どうやって ANOVA を救えば良いのでしょうか？ここで  $t$  検定についての議論を思い出せば、この問題に以前出会っていたことに気づきます。Student の  $t$  検定は等分散を仮定していますが、仮定が成り立たない時の解決策は Welch の  $t$  検定を使う、というものでした。実際、Welch1951 は ANOVA についてのこの問題をどうやって解決するかを示してくれています (the Welch one-way test)。JASP にも One-Way ANOVA 分析が組み込まれています。Welch の補正を組み込むためには、‘仮定のチェック’-‘均一性の補正’の下にある‘Welch’オプションを選択するだけでいいのです。この結果は図 ??に示しています。

ここで何が起こっているかを理解するために、前のセクション ??で最初にやった ANOVA で得ら

ANOVA - mood.gain							
Cases	Homogeneity Correction	Sum of Squares	df	Mean Square	F	p	$\eta^2$
drug	None	3.453	2.000	1.727	18.611	< .001	0.713
drug	Welch	3.453	2.000	1.727	26.322	< .001	0.713
Residual	None	1.392	15.000	0.093			
Residual	Welch	1.392	9.493	0.147			

Note. Type III Sum of Squares

Figure11.6 JASP の一要因 ANOVA の一部としての、 Welch's 均一性の補正

れた数字と比較してみましょう。後戻りする面倒を避けるために、前回の最後に得られた数字を書いておきます。 :  $F(2, 15) = 18.611, p < 0.001$  ですね。これは図??に示された One-Way ANOVA の Homogeneity Correction の下、 None に示されています。

オーケイ、 もとの ANOVA では結果として  $F(2, 15) = 18.6$  が得られてましたが、 Welch の補正をしたら  $F(2, 9.49) = 26.32$  になってますね。言い換えると、 Welch の補正是群内自由度が 15 から 9.49 に減っていて、 結果の F 値が 18.6 から 26.32 に増えています。

#### 11.6.4 正規性の仮定をチェックする

正規性の検定はもう少し直接的です。知っておくべきことはセクション ?? にほとんど書いてあります。やるべきことは、 QQ プロットを描く、 これだけです<sup>\*7</sup> JASP で QQ プロットをするには、「仮定のチェック」に行き、「残差 Q-Q プロット」をチェックします。その結果は図 ?? に示したとおりで、私にはちゃんと正規分布しているように見えます。

### 11.7

#### 正規性の仮定を取り除く

さて正規性のチェックの仕方を見たところで、正規性が破られた時にどうしたらしいか、と疑問を抱くのは当然ですよね。一要因 ANOVA の文脈では、最も簡単な解決策はノンパラメトリックな検定 (すなわち、確率分布のもつ特別な仮定を一切含まないものに立脚したもの) に切り替えることです。ノンパラメトリック検定については以前、第 ?? 章でやりました。二つの群があるときは、

<sup>\*7</sup> Shapiro-Wilk 検定も行うべきですが、これは今の JASP に実装されてません。Shapiro-Wilk 検定が有意でなければ (つまり  $p < .05$  なら)、この正規性の仮定は破られていないことを示しています。しかし、Levene 検定と同様に、もしサンプルサイズが大きければ Shapiro-Wilk の検定で有意になってしまっても、偽陽性である可能性があります。分析にあたって実質的な問題がないような、正規性の仮定が破られていない状況であっても。そして同様に、とても小さいサンプルでは、偽陰性になってしまうかもしれません。だから目で見てわかる WQQ プロットが重要なのです。

Mann-Whitney か Wilcoxon 検定が、あなたの必要としたノンパラメトリックな代替品を提供してくれます。三つ以上の群があるばあいは、**Kruskal-Wallis 順位和検定 (KruskalWallis1952)** を使うことができます。そう、これこそ次に説明しようとしているものです。

#### 11.7.1 Kruskal-Wallis 検定の背後にいるロジック

Kruskal-Wallis 検定はある意味 ANOVA によく似ています。ANOVA では  $Y_{ik}$  から話を始めました。ここでアウトカム変数は第  $k$  群の  $i$  番目の人の意味しています。Kruskal-Wallis 検定では、これら  $Y_{ik}$  の値を全て順位付けして、順序データの分析をするのです。

$R_{ik}$  を第  $k$  群の  $i$  番目のメンバーに与えられた順位だとしましょう。ここで  $\bar{R}_k$  を計算し、第  $k$  群における観測値の平均順位を考えます。

$$\bar{R}_k = \frac{1}{N_k} \sum_i R_{ik}$$

そして全体平均  $\bar{R}$  も計算します。

$$\bar{R} = \frac{1}{N} \sum_i \sum_k R_{ik}$$

これで、全体平均  $\bar{R}$  からの偏差平方を計算することができるようになりました。個々のスコアについてこれを計算する、つまり  $(R_{ik} - \bar{R})^2$  を計算することで、 $ik$  番目の観測値が全体平均の順位からどれだけずれているかについての、“ノンパラメトリックな”測度を手に入れたことになります。次に全体平均から群平均の偏差平方を計算する、つまり  $(\bar{R}_k - \bar{R})^2$  を算出すると、その群が全体平均の順位からどれぐらいずれているかのノンパラメトリックな測度を手に入れたことになります。覚えておいて欲しいのは、いまから ANOVA と同じロジックを辿っていき、以前やったようにここでの順序の平方和を定義することです。まず“全体の順序平方和”を計算します。

$$RSS_{tot} = \sum_k \sum_i (R_{ik} - \bar{R})^2$$

それから“群間順序平方和”を次のように算出します。

$$\begin{aligned} RSS_b &= \sum_k \sum_i (\bar{R}_k - \bar{R})^2 \\ &= \sum_k N_k (\bar{R}_k - \bar{R})^2 \end{aligned}$$

さて、もし帰無仮説が真で群間にどんな差も認められないなら、群間平方和  $RSS_b$  はとても小さくなり、全体順序和  $RSS_{tot}$  よりもグッと小さくなると思われるでしょう。質的には、こ

れは ANOVA で  $F$  統計量を出そうとした時と同じようなものなのですが、技術的な理由から Kruskal-Wallis 検定統計量は普通  $K$  で表され、少し違った方法で計算します。すなわち、

$$K = (N - 1) \times \frac{RSS_b}{RSS_{tot}}$$

として、もし帰無仮説が真なら  $K$  の標本分散は近似的に  $G - 1$  の自由度（ここで  $G$  は群の数です）を持ったカイ二乗分布に従います。より大きな  $K$  の値が出れば、帰無仮説とより一貫性がないということになります。これは一方向検定です。 $K$  が十分大きければ、 $H_0$  を棄却することになります。

### 11.7.2 補足

前のセクションで書いたのは、Kruskal-Wallis 検定の背後にあるロジックです。概念的なレベルでは、この検定がどういう働きをするのかを考えた方がいいでしょう。しかし、純粋に数理的な側面を考えるのは不必要に複雑です。その導出をして見せようとは思いませんが、ちょっとした代数的ごまかし<sup>a</sup>を使って、 $K$  の式が次のように書けることを示しておきましょう。

$$K = \frac{12}{N(N - 1)} \sum_k N_k \bar{R}_k^2 - 3(N + 1)$$

これは、あなたが  $K$  を算出する時にみる数式の最後の形です。この形式は、先ほどのセクションで示したものよりも簡単ですが、全体的に意味をなさないように見えますね。前に示した  $K$  の考え方の方が、順位に基づいた ANOVA のアナロジーとして良いように思えます。しかし最後に得られた検定統計量は、元の ANOVA で使われるものから見ると随分違うものに見える、ということは知っておいて欲しいのです。

いやまた、もっとあります！ なぜいつももっとあるんでしょうね？ 今までの話は、ローデータに紐づけられていない時は、常に真なのです。すなわち、同じ値を持つ変数が二つとない場合は、です。もし同順位があれば、この計算に補正項を入れなければなりません。こうなると、もっとも勤勉な読者でさえも気にしなくなったと思います（あるいは、同順位の項は今すぐ注目しなければ、という意見にはならないと思います）。ですから、さっさとどうやって計算するかを示して、なぜこれがこんな風になるのかというつまらない証明はパスしちゃいましょう。ローデータの度数分布表を作ったとして、 $f_j$  を  $j$  番目の一つしかない数字だとします。ちょっと抽象的ですから、度数分布表の具体例、`mood.gain` をデータセット `clinicaltrials.csv` から取り出しましょう。

0.1	0.2	0.3	0.4	0.5	0.6	0.8	0.9	1.1	1.2	1.3	1.4	1.7	1.8
1	1	2	1	1	2	1	1	1	1	2	2	1	1

この表を見ると、三番目の要素は 2 という数字を持っていることがわかります。これは `mood.gain` が 0.3 というのに対応しているので、二人の気分が 0.3 ポイント上昇したことがわかります。も

うひとつ。先ほど導入した数式的に表現するなら、 $f_3 = 2$  ということですね。イエイ。さて、こうなってくると補正項は次のようにになります。

$$TCF = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

Kruskal-Wallis 統計量の同順位の値は、 $K$  をこの量で割った時に得られます。JASP が計算するのは、この同順位補正版です。やっとこさ、Kruskal-Wallis 検定の理論についての話を終えることができます。Kruskal-Wallis 検定の同順位補正項の計算の仕方を知らない時に感じる不安を取り除けたので、一息つけたのではないかと思います。違います？

<sup>a</sup>jiggery-pokery. 専門用語です。

### 11.7.3 Kruskal-Wallis 検定を JASP で実行する

Kruskal-Wallis 検定が実際にどうなるのかを理解しようとして大変怖い思いをしたわけですが、この検定を実行するのはごく簡単です。というのも、JASP は ANOVA に‘ノンパラメトリック’パートを持っているからです。あなたがすべきことは、グループ変数 `drug` をアクティブボックスに動かすことだけです。そうすると Kruskal-Wallis は図??にあるように、 $\chi^2 = 12.076$ ,  $df = 2$ ,  $p$ -値 = 0.002 であることを示します。

## 11.8

### 反復測定の一要因 ANOVA

一要因反復測定 ANOVA 検定は、三群以上の間の有意な差異を検定するもので、そこでは各群において同じ実験参加者が使われます（あるいは各実験参加者が他の実験群における参加者と密接に関係がある場合です）。このため、各実験群には常に同じ数のスコア（データ点）があることになります。このタイプの実験デザインと分析は、‘対応のある ANOVA’とか‘Within 計画の ANOVA’とも呼ばれます。

反復測定 ANOVA の背後にいるロジックは、独立した ANOVA（‘Between 計画の ANOVA’と呼ばれることもあります）と非常に似ています。前のことを思い出して欲しいのですが、Between 計画の ANOVA では全分散が二つの要素、群間分散 ( $SS_b$ ) と群内分散 ( $SS_w$ ) に切り分けられ、それぞれを対応する自由度で割ることで、 $MS_b$  と  $MS_w$  にして（Table ??参照）、F-ratio を次のようにして計算するのでした。

$$F = \frac{MS_b}{MS_w}$$

反復測定 ANOVA では、F-ratio は同じように計算されますが、独立した ANOVA では分母にくる  $MS_w$  のもとになった群内分散 (SS) ですが、反復測定では  $SS_w$  が二つのパートに分離します。各群で同じ被験者を使いますから、個々人の間にある個人差に伴う分散 ( $SS_{subjects}$  と表されるもの) を、群内分散から取り除くことができるのです。これがどうやって計算されるかと言う、技術的な細部についてはこれ以上分け入ることはしませんが、要するに各被験者が被験者要因の各水準になるということです。この被験者内要因の分散はほかの被験者間要因と同じように計算されます。 $SS_w$  から  $SS_{subjects}$  を引き算することで、 $SS_{error}$  の項はより小さいものになります。

$$\text{Independent ANOVA: } SS_{error} = SS_w$$

$$\text{Repeated Measures ANOVA: } SS_{error} = SS_w - SS_{subjects}$$

この  $SS_{error}$  の項についての変化は、より強い統計的検定を引き出してくれますが、これは  $SS_{error}$  の減少分と誤差項の自由度の減少が相殺しあった上での話 (自由度は  $(n - k)^*8$  から  $(n - 1)(k - 1)$  になります。(独立 ANOVA 計画にはより多くの被験者がいることを思い出して))。

#### 11.8.1 JASP による反復測定 ANOVA

まずデータが必要ですね。Geschwind1972 が言ったように、脳卒中の後に生じる言語障害を正確に把握するためには、ダメージを受けた脳の特定の領域を診断する必要があります。ある研究者は、ブロックの失語症 (脳卒中のあとに一般的に経験される言語障害) を患っている 6 人の患者が経験したある言語障害を特定することに興味があるとします。

Table11.2 3 つの実験課題において成功した試行の数

患者	スピーチ	概念	文法
1	8	7	6
2	7	8	6
3	9	5	3
4	5	4	5
5	6	6	2
6	8	7	4

.....

患者は三つの言語再任課題をやるように言われます。最初の課題 (スピーチ生成) では、患者は一

\*8(n - k) : (被験者の数 - 群の数)

つの単語が実験者によって大声で読み上げられた後、それを反復するように求められました。第二の課題(概念)では、言葉の理解のテストで、患者はたくさんの写真とその名前を対応させるよう求められました。第三の課題(文法)では、正しい言葉の順序についての知識を検証するようなテストで、患者は文法的に正しくない文章を並べ替えるよう求められました。各被験者は全ての課題をやりました。患者が課題をする順番は、患者間でカウンターバランスが取られました。それぞれの課題は10回試行されます。各患者が課題に成功した数が、表??に示されています。これらのデータはbroca.csvファイルにあり、JASPに読み込むことができます。

一元配置反復測定ANOVAをJASPで実行するには、「ANOVA」から「反復測定ANOVA」をクリックし、次のように進めてください(図??を参照。)。

- 反復測定要因名を入力してください。これは全ての被験者に反復された条件を記述するラベルを、選べるようにするためのものです。例えばスピーチ、概念、文法の課題が全ての被験者に課されたのですから、適切なラベルは‘課題’でしょうか。この新しい用意名は、分析における独立変数を意味します。JASPでこれを実行するには、単に‘RM Factor 1’をクリックして、名前を入れるだけです；そこは強調表示されて、新しい名前の入力を待っていると思います—打ち込むだけ！
- 次に‘課題’要因のそれぞれの水準名を入れたいと思います。反復測定要因テキストボックスに、三つの水準を追加する必要があることに注意してください。この三つの水準は、三つの課題を意味しています。: speech, conceptual, and syntaxです。水準に対応するラベルに変えてください。それぞれをクリックして、新しい名前を入力するだけです。
- それから変数を左のボックスに移動させ、‘反復測定のセル’テキストボックスに入れます。変数名がさっそく入力した水準名と合致しているか確認してください。
- 最後に、仮定チェックのオプションの下、‘球面性のチェック’テキストボックスをクリックします(今はひとまず私を信じて！)

反復測定ANOVAのJASPの出力は図??のようになります。結果を見る前に、Mauchlyの球面性の検定をしなければなりません。これは条件間の分散が等しいという仮定を検定するものです(実験条件間の異なるスコアの広がりが、ほとんど同じだという意味です)。図??にあるように、Mauchlyの検定は有意水準が $p = .720$ です。Mauchlyの検定は有意ではなかった(つまり今回の研究例では $p > .05$ だった)ので、分散の間に有意な違いがない(つまり大体等しく、球面性を仮定できる)と結論づけるのが妥当だということになります。

ところで！もしMauchlyの検定が有意( $p < .05$ )であれば、分散間は有意に異なっていることになるので、球面性の仮定が満たされていないことになります。この場合、一要因ANOVAで得られたF値の補正をする必要があります。すなわち、

- “球面性の検定”表のGreenhouse-Geisserの値が $> .75$ であれば、Huynh-Feldtの補正を使う

べきです。

- Greenhouse-Geisser の値が  $< .75$  であれば, Greenhouse-Geisser の補正を使うべきです。

補正された  $F$  値はどちらも, 仮定のチェックオプションの下にある球面性の補正チェックボックスにチェックを入れることで得られます。補正された  $F$  値は図??の結果の表に示されています。

この分析では, Mauchly の球面性検定の  $p$  値は  $p = .720$  (つまり  $p > 0.05$ ) でした。ですから, 球面性の仮定が守られていると考えて,  $F$  値の補正は必要ないことになります。そこで, 球面性補正の出力にある ‘なし’ の反復測定の ‘課題’ の値を使います。 $:F = 6.93, df = 2, p = .013$  で, スピーチ, 理解, 文法という各言語課題の成功回数は有意に異なっている, ということができます ( $F(2, 10) = 6.93, p = .013$ )。

普通, 結果を解釈するために記述統計量をレビューするべきです。‘追加オプション’ のメニューに行き, ‘課題’ を ‘周辺平均’ の下にあるアクションボックスに入れることで, JASP 上でこれらの数値を算出できます。この結果が図??にありますが, 各条件の平均だけでなく 95%CI も示されています。被験者が達成した課題数の平均を比較すると, ブロッカの失語症はスピーチ (平均 = 7.17) と言語理解 (平均 = 6.17) の課題ではそれなりの成績を上げています。しかし, 文法課題ではかなりパフォーマンスが悪く (平均 = 4.33), 事後検定でもスピーチと文法のパフォーマンスには有意な差があります。

## 11.9

---

### ANOVA と Student の $t$ -検定との関係

さて, ANOVA の話を終わらせる前に一つ指摘しておきたいことがあります。多くの人が驚くと思うんですが, 知っておいて損はありません。2群の ANOVA は Student の  $t$  検定と同じものなのです。いや, ほんと。似ているというのではなくて, 実際あらゆる意味で等価なのです。これが常に真であることを証明しようとは思いませんが, 一つ具体的な例を示しましょう。我々の `mood.gain ~ drug` というモデルを, ANOVA で実行する代わりに, `therapy` を予測子として使うことを考えてみましょう。ANOVA を走らせると,  $F$  統計量は  $F(1, 16) = 1.71$  であり,  $p$  値は 0.21 になります。ここには2群しかありませんから, 実際には ANOVA に頼る必要はなく, Student の  $t$  検定をすることにします。こうすることで何が起こるかみてくださいよ。:  $t$ -統計量は  $t(16) = -1.3068$  で  $p$  値は 0.21 になります。不思議なことに,  $p$  値は一致します。もう一度  $p = .21$  になったのです。しかし, 検定統計量はどうでしょう? ANOVA の代わりに  $t$  検定をしたときには, ちょっと違う答えが出ていて,  $t(16) = -1.3068$  でした。しかしこれには, かなり直接的な関係があるのです。 $t$  統計量を二乗すると, 先ほどの  $F$  統計量を得ます。ほら,  $-1.3068^2 = 1.7077$  でしょ。

---

## 要約

この章には多くのことが含まれていますが、まだいくつも取りこぼしたものがあります。わかり切ったことですが、一つ以上のグループ化変数がある状況に興味がある場合の ANOVA をどうやるかについては説明していません。これは第 ?? 章で論じることになります。議論してきた内容について、キートピックスと言えば次のようになるでしょう。

- ANOVA がどのように働くかについての基本的なロジック (セクション ??) と、それを JASP でどう実行するか (セクション ??)。
- ANOVA での効果量をどうやって計算するか (セクション ??)。
- 事後の検定と多重比較の時の補正 (セクション ??)。
- ANOVA の仮定 (セクション ??)
- 分散の均一性の過程をどうやってチェックするか (セクション ??) と、その仮定が破られたときはどうするか (Section ??)。
- 正規性の仮定をどうやってチェックするか (セクション ??) と、その仮定が破られたときにどうするか (Section ??)。
- 反復測定 ANOVA (セクション ??) とそれに等価なノンパラメトリックな Friedman 検定 (セクション ??)。

この本の全てのチャプターについてそうですが、私が依拠しているいくつかの異なるソースがあります。が、最も影響を受けている一冊を挙げるとすれば、**Sahai2000** です。この本は初心者向けではありませんが、ANOVA の背後にある数学的な側面を理解しようとする、少し進んだ読者にとって素晴らしい本だと言えるでしょう。

## Assumption Checks

Q-Q Plot

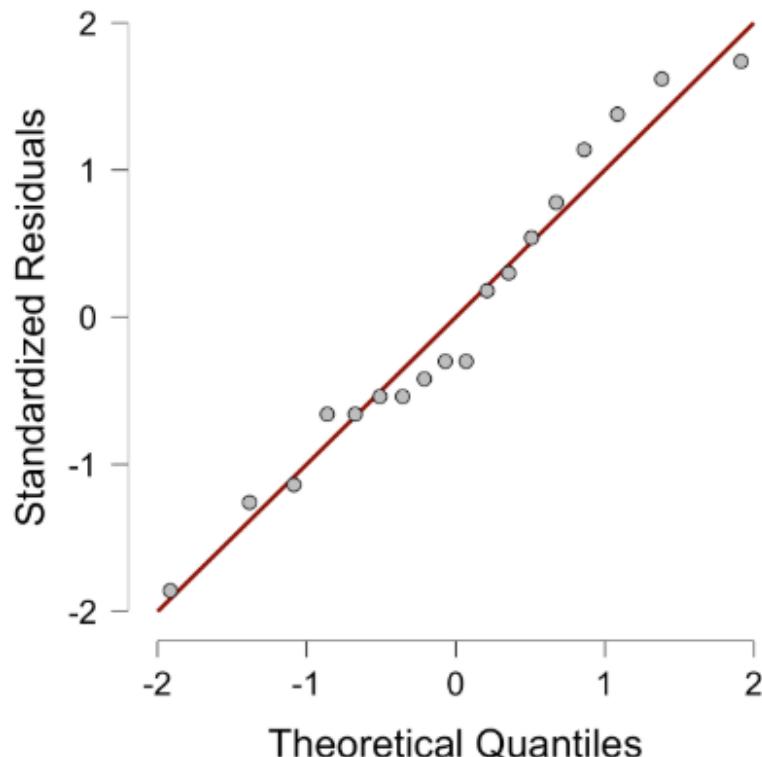


Figure11.7 JASP で作った QQ プロット

---

## Kruskal-Wallis Test

Factor	Statistic	df	p
drug	12.076	2	0.002

Figure11.8 JASP における一要因ノンパラメトリック ANOVA である, Kruskal-Wallis 検定

---

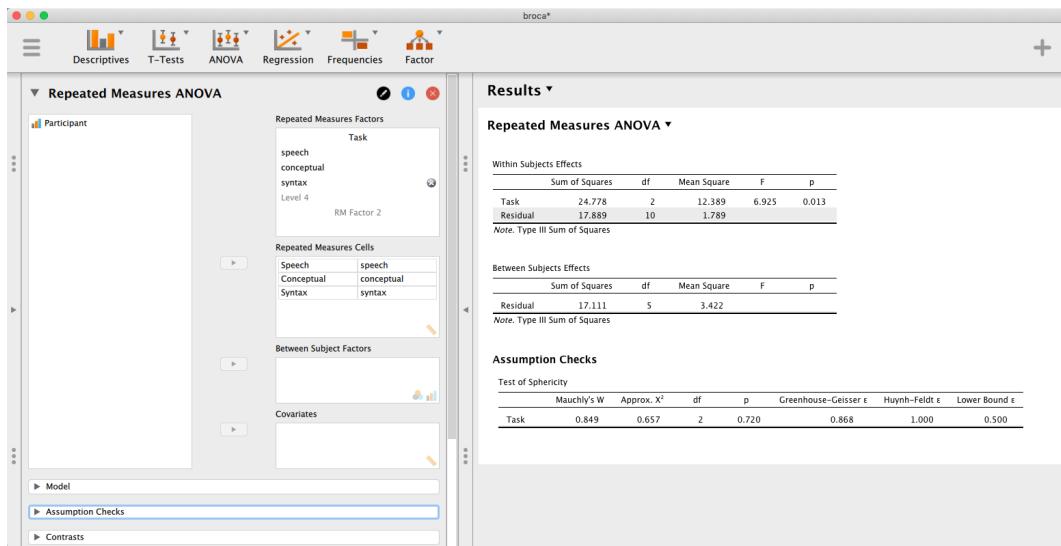


Figure11.9 JASPにおける反復測定 ANOVA

#### Assumption Checks

##### Test of Sphericity

	Mauchly's W	Approx. $\chi^2$	df	p	Greenhouse-Geisser $\epsilon$	Huynh-Feldt $\epsilon$	Lower Bound $\epsilon$
Task	0.849	0.657	2	0.720	0.868	1.000	0.500

Figure11.10 一要因反復測定 ANOVA の出力: Mauchly の球面性テスト

#### Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Task	None	24.778	2.000	12.389	6.925	0.013
	Greenhouse-Geisser	24.778	1.737	14.265	6.925	0.018
	Huynh-Feldt	24.778	2.000	12.389	6.925	0.013
Residual	None	17.889	10.000	1.789		
	Greenhouse-Geisser	17.889	8.685	2.060		
	Huynh-Feldt	17.889	10.000	1.789		

Note. Type III Sum of Squares

Figure11.11 一要因反復測定 ANOVA の出力: 被験者内効果の検定

### Marginal Means

Marginal Means – Task

Task	Marginal Mean	SE	95% CI	
			Lower	Upper
speech	7.167	0.624	5.825	8.509
conceptual	6.167	0.624	4.825	7.509
syntax	4.333	0.624	2.991	5.675

Figure11.12 反復測定 ANOVA の出力: ‘周辺平均’ ダイアログより、記述統計量が示されます。

## 12. 多元配置分散分析

---

ここまでこの章で、我々は多くの統計解析について学んできました。1つの名義的な予測変数を用いて、2つのグループの差 (e.g. *t* 検定, Chapter ??) や、3つ以上のグループの差 (e.g. 一元配置分散分析, Chapter ??) について統計的検定を行う方法について見てきました。回帰分析の章 (Chapter ??) では、複数の量的な予測変数を用いて单一の結果変数を説明するモデルを建てるという、強力で新しいアイデアが紹介されました。回帰モデルを用いることで、例えば、ある生徒のテスト勉強の時間やIQ テストの得点に基づいて、その生徒の読解テストの誤答数を予測することができます。

本章の目的は、複数の予測変数を使用するというアイデアを、分散分析の枠組みへと拡張することです。例えば、我々が、読解テストを用いて3つの異なる学校における生徒の成績を測定しようとしていると考えてみましょう。加えて、我々は、女子と男子が異なる速度で発達している（したがって、成績も平均的に異なることが予想される）と想定しています。各生徒は、彼／彼女らの性別と、所属する学校という2つの異なる変数によって分類されます。我々の目的は、これらのグループ化変数の両方に基づいて、読解テストの成績を分析することです。これを実現するための手法が、いわゆる**多元配置分散分析**です。ここでは2つのグループ化変数があるため、この手法は Chapter ??で登場した一元配置分散分析に対して、二元配置分散分析と呼ばれることがあります。

### 12.1

---

#### 多元配置分散分析 1: 釣合型デザイン, 交互作用なし

分散分析について述べた Chapter ??では、かなり単純な実験計画が想定されていました。各個人は特定のグループに属しており、我々の目的は、いくつかの結果変数について、これらのグループ間で平均値が異なるかどうかを明らかにすることでした。この節では、**多元配置デザイン**と呼ばれる、2つ以上のグループ化変数を持つより広範な実験デザインについて見ていきます。先ほど、こうしたデザインが必要となるような例を1つ挙げました。Chapter ??で登場した別の例では、各個人の経験した**気分の向上**に対する異なる薬の影響に注目しました。この例では薬の有意な効果が見出されま

したが、章の終盤では、それに加えてセラピーの効果を確認するための分析を行いました。セラピーの効果は見出されませんでしたが、同じ結果変数を予測する2つの分析を別個に行つたことに対する若干の懸念があります。おそらく、実際にはセラピーによる気分の向上効果はあるのでしょうか？その効果は薬の効果によって”隠されて”いたために見つけられなかったのではないでしょうか？言い換えれば、我々は薬とセラピーの両方の予測変数を含む、単一の分析を行う必要があります。この分析では、各個人は、投与された薬(3水準の要因)および受けたセラピーの種類(2水準の要因)という2つの要因によって分類されます。こうした分析は $3 \times 2$ 要因デザインと呼ばれます。

JASPの’頻度’ – ’分割表’の分析を用いて薬とセラピーのクロス集計表を作成すると、Figure ??のような表が得られます。

Contingency Tables

drug	therapy		Total
	CBT	no.therapy	
anxitfree	3	3	6
joyzepam	3	3	6
placebo	3	3	6
Total	9	9	18

Figure12.1 薬とセラピーによる分割表

集計表から、2つの要因のあらゆる組み合わせに参加者が属している、すなわちこの分析が完全交差デザインであることだけでなく、各グループに同数の参加者が属していることが分かります。言い換えれば、この分析は釣合型デザインだということです。これは最も単純なケースであるため、この節では、釣合型デザインのデータをどのように分析するかを見ていきます。非釣合型デザインに関する説明はかなり冗長なので、ここでは一旦置いておくことにします。

### 12.1.1 検定したい仮説はどんなものか？

多元配置分散分析は、一元配置分散分析と同様に、母集団の平均値に関する仮説を検定するための手法でしたがって、この分析の仮説が実際にはどのようなものであるかを明確にすることから始めるのが賢明でしょう。しかし、このことについて議論するにあたって、母集団の平均の簡潔な表記法があると非常に便利です。観測値は2つの異なる要因に応じて分類されているため、分析者が関心を持ちうる、非常に多くの平均値があります。これを確かめるために、今回のデザインにおいて計

算可能なあらゆるサンプル平均について考えてみましょう。まず、我々は明らかに、以下のようなグループごとの平均値に関心があります：

%drug	therapy	mood.gain
薬の種類	セラピーの種類	気分の向上
%placebo	no.therapy	0.300000
プラセボ	セラピーなし	0.300000
%anxitfree	no.therapy	0.400000
アンザイフリーセラピーなし		0.400000
%joyzepam	no.therapy	1.466667
ジョイゼパム	セラピーなし	1.466667
%placebo	CBT	0.600000
プラセボ	CBT	0.600000
%anxitfree	CBT	1.033333
アンザイフリーセラピー	CBT	1.033333
%joyzepam	CBT	1.500000
ジョイゼパム	CBT	1.500000

この出力は、2つの要因のあらゆる組み合わせ (e.g., プラセボ群でセラピーなし, プラセボ群でCBTを実施, など) におけるグループごとの平均値のリストになっています。これらの数値に加えて、行と列の平均および全体の平均を、以下のように1つの表で示しておくと便利です：

	セラピーなし	CBT	合計
プラセボ	0.30	0.60	0.45
アンザイフリーセラピー	0.40	1.03	0.72
ジョイゼパム	1.47	1.50	1.48
合計	0.72	1.04	0.88

これらの平均値のそれぞれは、当然ながらサンプル統計量です。これらの値は、我々の研究において行われた特定の観察に依存しています。我々が推定したいのは、これらの値と対応する母集団のパラメータです。すなわち、より広範な母集団の中に存在する真の平均です。これらの母平均も同様に表として整理することができますが、そのためには少々、数学的な表記が必要です。ここでは一般的な表記にしたがって、 $\mu$  を母平均の記号として用います。ただし、表中には様々な平均値があるため、添字を使ってこれらを区別する必要があります。

表記法は次の通りです。この表は2つの要因によって構成されています。各行は要因 A(ここでは薬)のそれぞれの水準に対応し、各列は要因 B(ここではセラピー)のそれぞれの水準に対応します。 $R$  が表中の行数を、 $C$  が列数を表すとき、この分析は  $R \times C$  要因の分散分析と表現することができます。ここでは  $R = 3$ 、 $C = 2$  となります。小文字を使って特定の行と列を表します。したがって、

$\mu_{rc}$  は要因 A の第  $r$  水準 (i.e.  $r$  行目), 要因 B の第  $c$  水準 ( $c$  行目) の母平均を表します。<sup>\*1</sup>母平均は以下のように表すことができます：

	セラピーなし	CBT	合計
プラセボ	$\mu_{11}$	$\mu_{12}$	
アンザイフリー	$\mu_{21}$	$\mu_{22}$	
ジョイゼパム	$\mu_{31}$	$\mu_{32}$	
合計			

さて、残りの組み合わせについてはどうでしょうか？ 例えば、CBT を受けるかどうかに関わらず、今回のような実験においてジョイゼパムを投与される可能性のある（仮想的な）母集団全体の平均的な気分の向上について、どのように記述すれば良いでしょうか。これは”ドット”記法によって表すことができます。先ほどのジョイゼパムの例に関しては、表の第 3 行目の値を平均すれば求められることが分かります。すなわち、2 つのセルの平均値 (i.e.,  $\mu_{31}$  および  $\mu_{32}$ ) を平均化するということです。この平均化の結果は周辺平均と呼ばれ、この場合には  $\mu_{..}$  と表記されます。CBT の周辺平均は、表の第 2 列目についての母平均と対応するため、 $\mu_{.2}$  と表記されます。総平均は、行と列の両方を平均化（周辺化<sup>\*2</sup>）することによって得られる平均値であるため、 $\mu_{..}$  と表記されます。母平均についての完全な表は、以下のように書くことができます：

	セラピーなし	CBT	合計
プラセボ	$\mu_{11}$	$\mu_{12}$	$\mu_{1..}$
アンザイフリー	$\mu_{21}$	$\mu_{22}$	$\mu_{2..}$
ジョイゼパム	$\mu_{31}$	$\mu_{32}$	$\mu_{3..}$
合計	$\mu_{..1}$	$\mu_{..2}$	$\mu_{...}$

この表記法によって、仮説を定式化して表現することが容易になります。以下の 2 点を明らかにすることを目指すと考えてみましょう。まず、薬の選択が気分に何らかの影響を及ぼすか？ 次に、CBT は気分に何らかの影響を及ぼすか？ もちろん、定式化することができる仮説はこれらだけではありません。Section ??において、これらとは別の、非常に重要な仮説の例が示されます。しかし、これらは検定における最も単純な 2 つの仮説であるため、まずはこの 2 つから始めましょう。まず、最初の検定について考えます。もし薬が何の効果も持たないとすると、すべての行平均は同じになる

<sup>\*1</sup>添字を使った表記法の良いところは、その一般化可能性です。もし、この実験に 3 つ目の要因が加わったとしても、単に 3 つ目の添字を追加するだけで済みます。原理的には、添字は実験に加えたい要因の数に応じていくつでも拡張することができますが、本書では 2 つ以上の要因を含む分析を扱うことはほとんどないため、添字が 3 つを超えることはありません

<sup>\*2</sup>技術的には、周辺化は一般的な平均と全く同一ではありません。周辺化は、平均化しようとする様々なイベントの頻度を加味した加重平均です。しかし、釣合型デザインにおいては、すべてのセルの頻度が定義上等しいため、これらは同じ値になります。後に非釣合型デザインについて説明する際に、この計算が非常に頭痛の種になるものだということが分かるでしょう。ですが、今の所は忘れて構いません。

はずですね？ したがって、これが帰無仮説になります。一方で、薬が何らかの効果を持つとすると、行平均は異なるものになることが予想されます。正式には、これらの帰無仮説および対立仮説は、周辺平均の等価性の考え方方に沿って書き表されます：

$$\begin{aligned} \text{帰無仮説, } H_0: & \text{ 行平均が等しい, i.e., } \mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} \\ \text{対立仮説, } H_1: & \text{ 少なくとも 1 つの行平均が異なる} \end{aligned}$$

これらの統計的仮説が、Chapter ??でこれらのデータに対して一元配置分散分析を行った際の仮説と全く同じであることは注目に値します。その際には、プラセボ群の平均的な気分の向上を表す表記として  $\mu_P$  を、2 つの薬のグループ平均を表す表記として  $\mu_A$  と  $\mu_J$  を用い、帰無仮説は  $\mu_P = \mu_A = \mu_J$  で表されました。ここでも同じ仮説について説明しているのですが、複数のグループ化変数を持つより複雑な分散分析においては、より丁寧な表記が必要なため、ここでは帰無仮説は  $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$  と表されます。しかしながら、後述のように、仮説は同じであるものの、2 つ目のグループ化変数が存在することによって、仮説の検定の仕方は微妙に異なります。

もう一方のグループ化変数に話を移して、2 つ目の仮説検定も同様の方法で定式化できることに気付いたとしても、もはや驚かないでしょう。ただし、今度は薬の効果ではなく心理療法に注目するため、帰無仮説は列平均の等価性に対するものになります：

$$\begin{aligned} \text{帰無仮説, } H_0: & \text{ 列平均は等しい, i.e., } \mu_{\cdot 1} = \mu_{\cdot 2} \\ \text{対立仮説, } H_1: & \text{ 列平均は異なる, i.e., } \mu_{\cdot 1} \neq \mu_{\cdot 2} \end{aligned}$$

### 12.1.2 JASP による分析の実行

先ほどの節で説明した帰無仮説と対立仮説には、随分と見覚えがあるように思えます。これらは基本的に、Chapter ??の一元配置分散分析において検定した仮説と同じです。そのため、多元配置分散分析で用いられる仮説の検定も、Chapter ??で登場した  $F$  検定と本質的には同じであると期待しているのではないかでしょうか。平方和 (SS), 平均平方 (MS), 自由度 (df), そして最終的には  $p$  値に変換することのできる  $F$  統計量を参照する方法が、ここでも使えると思っているのではないか？まさにその通りです。そういうわけなので、ここでは前章までとは異なるアプローチを取りたいと思います。本書を通じて、まずは特定の分析の基礎となるロジック（およびある程度の数学的な記述）を説明し、その後に JASP による分析方法の解説を行うアプローチを取ってきました。今回は、これとは逆に、まず JASP でどのように分析を行うかを示します。その理由は、Chapter ??で説明した単純な一元配置分散分析と、この章で使用するより複雑な分散分析との類似点を強調したいからです。

分析しようとしているデータが釣合型の要因計画に対応している場合、分散分析の実施は容易になります。どれほど容易であるかを確認するため、Chapter ??で行った分析を再現することから始め

ましょう。忘れてしまった読者のために、この分析では1つの要因(i.e., 薬)によって結果変数(i.e., 気分の向上)を予測しようとし、Figure ??のような結果を得ている。

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	18.611	< .001
Residual	1.392	15.000	0.093		

*Note. Type III Sum of Squares*

Figure12.2 JASPによる気分の向上を結果変数、薬を予測変数とする一元配置分散分析

加えて、ここではセラピーが気分の向上と関係しているかどうかを知りたいと考えます。Chapter ??で行われた重回帰分析に関する議論を踏まえると、セラピーの変数を2番目の‘固定効果’として加えるだけでこの分析ができると知っても驚かないでしょう。Figure ??を見てください。

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	26.149	< .001
therapy	0.467	1.000	0.467	7.076	0.019
Residual	0.924	14.000	0.066		

*Note. Type III Sum of Squares*

Figure12.3 JASP two way ANOVA of mood.gain by drug and therapy

Figure12.4 JASPによる気分の向上を結果変数、薬およびセラピーを予測変数とする二要因分散分析

先ほどと同様に、この出力も非常に分かりやすくなっています。表の最初の行は、薬の要因に関する群間平方和(SS)と、対応する群間のdfを表しています。平均平方(MS)とF統計量およびp値も示されています。同様に、セラピーの要因に対応する行と、残差(i.e., 群内変動)に対応する行があります。

これらの数値はそれぞれ見覚えのあるものでしょうし、これらの数値の関係もまた、一元配置分散分析を行ったときと変わっていません。平均平方は、SSを対応するdfで割ることによって計算されていることに注意してください。したがって、まだ薬やセラピー、残差については言及していません。

んが、

$$MS = \frac{SS}{df}$$

という関係がここでも成り立ちます。これを確認するために、平方和がどのように計算されるかを気にかける必要はありません。代わりに、JASP が SS を正しく計算してくれたことを信じて、他の数値の意味についても考えてみましょう。まず、**薬**の要因に関して、3.453 を 2 で割ると、平均平方は 1.727 という値になります。**セラピー**の要因に関しては、自由度が 1 しかなく、計算も容易になります：0.467(SS) を 1 で割ると、0.467(MS) が得られます。

$F$  統計量と  $p$  値を見ると、それぞれ 2 つずつあることに気づきます；1 つは**薬**の要因に対応し、もう 1 つは**セラピー**の要因に対応しています。どちらの場合も、 $F$  統計量は要因に対応する平均平方の値を残差に対応する平均平方の値で割ることで計算されます。最初の要因（要因 A；今回の場合は**薬**）を表す省略表記として“A”を、残差を表す省略表記として“R”を用いる場合、要因 A に対応する  $F$  統計量は  $F_A$  で表され、以下のように計算されます：

$$F_A = \frac{MS_A}{MS_R}$$

また、要因 B(i.e., **セラピー**) についても同様の計算ができます。先ほど表中の行数を表す文字としても R を使用しているので、残差の表記として“R”を用いるのは少し紛らわしいですが、 $SS_R$  や  $MS_R$  といった文脈でのみ、“R”を残差を表すものとして使用するので、混乱しないよう願います。ともかく、この式を**薬**の要因に適用すると、要因の平均平方 1.727 を残差の平均平方の値 0.066 で割ることになり、26.149 という  $F$  統計量が得られます。**セラピー**の変数については、0.467 を 0.066 で割ることで、7.076 という  $F$  統計量が計算されます。当然ですが、これらは先ほど JASP が分散分析表で報告した値と同じです。

分散分析表には  $p$  値も含まれています。これもまた、特に目新しいことはありません。2 つの要因のそれについて、要因と結果変数の間に関係は無いという帰無仮説を検定します（これについては後ほど詳しく説明します）。そのため、分散分析を行ったときと（明らかに）同様の方法で、これらの仮説に関する  $F$  統計量を計算しました。これらを  $p$  値に変換するには、帰無仮説（検討している要因の影響はない）のもとでの  $F$  統計量の度数分布である  $F$  分布が必要です。2 つの自由度の値は要因と残差にそれぞれ対応していることにも注目してください。**薬**の要因については、自由度 2 と 14 の  $F$  分布を参照することになります（自由度については後ほど詳しく説明します）。一方、**セラピー**の要因については、自由度 1 と 14 の  $F$  分布を参照することになります。

ここで、このより複雑な要因計画のもとでの分散分析表は、単純な一元配置の分散分析の分散分析表と、ほぼ同様の方法で読み取れることに気づくでしょう。要するに、 $3 \times 2$  要因の多元配置分散分析の結果、**薬**の有意な効果 ( $F_{2,14} = 26.15, p < .001$ ) および**セラピー**の有意な効果 ( $F_{1,14} = 7.08, p = .02$ ) が見出されたことが分かります。あるいは、より専門的で正確な用語を用いると、**薬**と**セラピー**の効果という 2 つの**主効果**があるといえます。現時点では、これらを“主”効果

と呼ぶことはやや冗長に思えますが、これには意味があります。この後、2つの要因の間に“交互作用”があるという可能性を検討するため、通常は主効果と交互作用効果を区別するのです。

### 12.1.3 平方和はどのように計算されるか？

ここまで説明には2つの目的がありました。まず、多元配置分散分析をJASPで実行する方法は、一元配置分散分析とほとんど同じであることを示すことです。唯一の違いは、2つ目の要因の追加です。次に、多元配置分散分析の分散分析表を参照することで、多元配置分散分析の背後にいる基本的なロジックと構造が、一元配置分散分析の背後にいるものと同じであることを示すことです。その感覚を大切にしてください。まさしく、多元配置分散分析は、一元配置分散分析とほとんど同じ方法で構成されているのです。分析の詳細について掘り下げ始めると、この感覚は揺らいできます。得てして、この心地よい感覚は、次第に統計学の教科書の著者に対する恨み辛みへと変わっていきます。

それでは、詳細について見ていくことにしましょう。先ほどの節では、主効果（ここでは薬およびセラピー）に関する仮説検定がF検定であることは説明しましたが、平方和(SS)がどのように計算されるかは示されていませんでした。同様に、自由度(df)の計算方法も説明されていませんが、こちらは比較的単純です。要因Aと要因Bの2つの予測変数があると仮定しましょう。結果変数を $Y$ で表すとき、グループ $rc$ (i.e.,  $r$ は要因Aに対応する行の水準,  $c$ は要因Bに対応する列の水準)に属する $i$ 番目の参加者の反応は $Y_{rci}$ で表すことができます。したがって、 $\bar{Y}$ を用いてサンプル平均を表す場合、同様の表記法でグループ平均、周辺平均、総平均を表すことができます。すなわち、 $\bar{Y}_{rc}$ は要因Aの第 $r$ 水準、要因Bの第 $c$ 水準に対応するサンプル平均を表し、 $\bar{Y}_{r..}$ は要因Aの第 $r$ 水準に関する周辺平均を、 $\bar{Y}_{..c}$ は要因Bの第 $c$ 水準に関する周辺平均を、そして $\bar{Y}_{...}$ は総平均を表します。言い換えれば、サンプル平均は母平均と同様の表で整理することができます。今回のデータでは、以下のようになります：

	セラピーなし	CBT	合計
プラセボ	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1..}$
アンザイフリー	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2..}$
ジョイゼパム	$\bar{Y}_{31}$	$\bar{Y}_{32}$	$\bar{Y}_{3..}$
合計	$\bar{Y}_{..1}$	$\bar{Y}_{..2}$	$\bar{Y}_{...}$

先ほど示したサンプル平均は、 $\bar{Y}_{11} = 0.30$ ,  $\bar{Y}_{12} = 0.60$ などです。今回の例では、**薬**の要因には3つの水準が、**セラピー**の要因には2つの水準があるため、 $3 \times 2$ 要因の多元配置分散分析を実行しようとしていました。より一般的な書き方では、要因A(行方向の要因)がR水準、要因B(列方向の要

因) が  $C$  水準を持ち,  $R \times C$  要因の多元配置分散分析を行うと表現できます。

表記が定まったことで, 2つの要因それぞれの平方和の値を比較的馴染みのある方法で計算することができます。要因 A についての群間の平方和は, (行) の周辺平均  $\bar{Y}_{1..}$ ,  $\bar{Y}_{2..}$  などが総平均  $\bar{Y}..$  との程度異なるかを評価することで計算されます。これには一元配置分散分析と同様の方法が用いられます:  $\bar{Y}_{r..}$  と  $\bar{Y}..$  の平方和の差を計算するのです。具体的には, 各グループに  $N$  人の参加者が属する場合, 以下のように計算されます

$$SS_A = (N \times C) \sum_{r=1}^R (\bar{Y}_{r..} - \bar{Y}..)^2$$

一元配置分散分析と同様に, この数式の中で最も興味深い<sup>a</sup>部分は  $(\bar{Y}_{r..} - \bar{Y}..)^2$  という部分であり, 水準  $r$  についての偏差の 2 乗に関連しています。この式が行なっているのは, 要因の  $R$  水準すべての偏差の二乗を計算し, 足し合わせ, その結果を  $N \times C$  に掛けるという計算です。最後の計算を行う理由は, このデザインでは要因 A において  $r$  水準を持つセルが複数あるためです。実際に, 要因 B のそれぞれの水準に対応する  $C$  通りのセルがあります。例えば, この例では, 薬のアンザイフリーという水準に対応する 2 つの異なるセルがあります: 1 つはセラピーなしのグループ, もう 1 つは CBT のグループです。それだけでなく, これらのセルのそれぞれについて,  $N$  個の観測値があります。したがって, SS の値を「観測値ごと」の群間の平方和を表す量に変換するためには,  $N \times C$  を掛ける必要があるのです。要因 B についての式は,もちろん, いくつかの添え字が異なる点を除いて同じものになります

これらの式が得られたことで, 先ほどの節の JASP の出力と照らし合わせることができます。繰り返しになりますが, こういった計算には専用のスプレッドシートプログラムが役立ちます。

まずは, 薬の主効果について平方和を計算しましょう。各グループについて, 合計  $N = 3$  の参加者がおり,  $C = 2$  の異なる種類のセラピーがあります。見かたを変えると, 特定の薬を投与された  $3 \times 2 = 6$  の参加者がいることになります。スプレッドシートプログラムでこれらの計算を行うと, 薬の主効果に関する平方和の値は 3.45 となります。驚くべきことではありませんが, これは先ほど Figure ??で示した分散分析表における薬の要因の SS と同じ値です。

治療の効果についても, 同様の計算を行うことができます。先ほどと同じく, 各グループには  $N = 3$  の参加者がいますが, 今度は  $R = 3$  の異なる種類の薬があるため, CBT を受けた  $3 \times 3 = 9$  の参加者と, セラピーを受けなかった 9 名の参加者がいます。セラピーの主効果に関する平方和は, 0.47 と計算されます。繰り返しになりますが, 計算結果が Figure ??の分散分析表と同じになることは驚くべきことではありません。

以上が, 2 つの主効果の SS の値を計算する方法です。これらの SS の値は, Chapter ??で一元配置分散分析を行ったときに計算した群間平方和の値と類似しています。ただし, 今回は 2 つの異

なるグループ化変数があることで混乱しやすくなるため、それらを群間の SS 値として捉えることはお勧めできません。F 検定を行うためには、群内平方和も計算する必要があります。回帰分析の章 (Chapter ??) で使用した用語と、そして JASP が分散分析表で出力する用語と合わせるため、群内 SS 値は残差平方和  $SS_R$  で表すことにしましょう。

この文脈において、残差 SS 値について考える最も簡単な方法は、それを結果変数における周辺平均の違いを取り除いた (i.e.,  $SS_A$  および  $SS_B$  を取り除いた) 後の、残りの変動として捉えることです。すなわち、 $SS_T$  というラベルの付いた、平方和の合計の計算から始めることになります。この計算式は、一元配置分散分析の場合とほぼ同じになります。各観測値  $Y_{rci}$  と総平均  $\bar{Y}_{..}$  の差をとり、差の二乗を合計します

$$SS_T = \sum_{r=1}^R \sum_{c=1}^C \sum_{i=1}^N (Y_{rci} - \bar{Y}_{..})^2$$

ここでの「三重総和」は実際以上に複雑に見えます。最初の 2 つの総和は、要因 A のすべての水準 (i.e., 表中の  $r$  のすべての行) および要因 B の全ての水準 (i.e., 表中の  $c$  のすべての列) を合計しています。各  $rc$  の組み合わせは 1 つのグループに対応し、各グループには  $N$  人の参加者が含まれているため、これらの参加者 (i.e., すべての  $i$  の値) を合計する必要があります。つまり、ここで行っているのは、データセット内の全ての観測値 ((i.e.,  $rci$  の全ての組み合わせ) を合計することです。

ここで、結果変数の総合的な変動である  $SS_T$  が明らかになり、その変動のうちどれだけが、要因 A( $SS_A$ ) および要因 B( $SS_B$ ) に起因するかを知ることができます。したがって、残差平方和は 2 つの要因のいずれにも起因しない  $Y$  の変動であると定義されます。言い換えれば、

$$SS_R = SS_T - (SS_A + SS_B)$$

もちろん、残差 SS を直接計算するための公式もありますが、上記のように考えることには、より概念的な意味があります。残差という言葉は、それが変動の残りの部分であることを示しており、上記の式はそれを明確にします。「分散分析モデル」に起因する変動である  $SS_A + SS_B$  を、(回帰分析の章で用いられていたように)  $SS_M$  と表記することも一般的であり、このことから、平方和の総和はモデルの平方和に残差の平方和を加えたものに等しい、という表現がよく使われます。この章の後半において、これは単なる表面的な類似性ではないことが分かります：分散分析と回帰分析が内部で行っていることは、実際に、同じなのです。

いずれにせよ、この式を用いて  $SS_R$  を計算し、JASP の出力した分差分析表と同じ答えが得られることを確認することには、時間を割くだけの価値があるでしょう。繰り返しますが、スプレッドシートを利用すると計算は非常に簡単です。上述の式を用いて SS の総和を算出し (SS の総和 = 4.85 となります)，次に、残差の SS (= 0.92) を求めます。JASP の出力と同じ答えになるはず

です。

<sup>a</sup>訳：「最も退屈な」

#### 12.1.4 自由度はどのように求めるか？

自由度は、一元配置分散分析とほぼ同じ方法で計算されます。ある要因について、自由度は水準数から 1 を引いたものに等しくなります (i.e., 行方向の要因 A については  $R - 1$ , 列方向の要因 B については  $C - 1$ )。したがって、**薬**の要因については  $df = 2$ , **セラピー**の要因については  $df = 1$  となります。後ほど、回帰モデルとしての分散分析モデルの解釈について説明する際に (Section ?? を参照), この数値の算出方法について詳しく説明します。当面の間は、自由度の単純な定義、すなわち、自由度は観測値の数から制約の数を引いたものに等しいという定義を利用できます。このことから、**薬**の要因については、3 つの個別のグループ平均値が観測されていますが、これらは 1 つの総平均によって制約されるため、自由度は 2 となります。残差の自由度の計算方法は、ロジックは似ていますが、全く同じではありません。今回の実験における観測値の総数は 18 です。制約は、総平均に関するものが 1 つ、**薬**の要因の追加のグループ平均に関するものが 2 つ、**セラピー**の要因の追加のグループ平均に関するものが 1 つあるため、この場合の自由度は 14 となります。式で表すと  $N - 1 - (R - 1) - (C - 1)$  となり、 $N - R - C + 1$  のように簡略化されます。

#### 12.1.5 多元配置分散分析と一元配置分散分析

ここまで、多元配置分散分析がどのように行われるかについて見てきました。ここまで経過を、一元配置分散分析の結果と比較することには、時間を割くだけの価値があります。そうすることで、なぜ多元配置分散分析を行わなければならないかが明らかになります。Chapter ??では、まず使用した薬による差異を検討するための一元配置分散分析を実行し、次いでセラピーの違いによる差異を検討するための一元配置分散分析を実行しました。Section ??で述べたように、一元配置分散分析で検定される帰無仮説および対立仮説は、多元配置分散分析で検定される仮説と全く同じです。分散分析表をさらに注意深く見ると、それぞれの分析において、要因に関する平方和の値 (**薬**の要因については 3.45, **セラピー**の要因については 0.92) および自由度の値 (**薬**の要因については 2, **セラピー**の要因については 1) が同じであることが分かります。しかしながら、結果は同じではありません！ 最も注目すべき点は、Section ??において**セラピー**の要因について一元配置分散分析を行った際には、有意な効果は得られなかったことです ( $p$  値は .21 でした)。

一方で、2 要因分散分析における**セラピー**の主効果に着目すると、有意な効果 ( $p = .019$ ) が得られています。これら 2 つの分析は、明らかに同じではありません。

なぜこのようなことが起こるのでしょうか？ その答えは、残差の計算方法を理解することで明らかになります。F 検定の背後にいる考え方とは、特定の要因に起因する変動と、それらで説明できない

変動(残差)の比較であったことを思い出して下さい。セラピーについての一要因分散分析を実行することは、すなわち、薬の効果を無視することになり、薬の要因に由来する変動を残差へと放り込んでしまうことになります！これによって、データは実際以上に煩雑になり、2要因分散分析においては正しく有意な効果が見出されているセラピーの要因が、有意ではなくなってしまいます。

何かの影響を評価しようとするとき、他の重要な何か(e.g., 薬の要因)を無視してしまうと、分析が歪んでしまいます。もちろん、関心のある現象とはまったく関係のない変数は、無視してしまっても問題ありません。実験室の壁の色を記録しておいて、3要因分散分析の結果、その要因が重要でないことが判明した場合には、その無関係な要因を除外した、より単純な2要因分散分析の結果を報告するだけで十分です。重要なのは、実際には差を生じさせる要因を、分析から除外しないことです！

#### 12.1.6 この分析からどのような結果が得られるか？

ここまで説明してきた分散分析モデルは、我々がデータから発見する可能性のあるさまざまなパターンをカバーしています。例えば、2要因分散分析デザインでは、4通りの可能性があります：(a)要因Aの効果のみがある場合、(b)要因Bの効果のみがある場合、(c)要因Aと要因Bの両方の効果がある場合、(d)どちらの要因の効果もない場合です。これら4つの可能性のそれぞれの例が、Figure ??に示されています。

## 12.2

---

### 多元配置分散分析2：釣合型デザイン、交互作用あり

Figure ??に示されている4つのパターンは、いずれも現実的なものです。これらのパターンを生じさせるようなデータセットも非常にたくさん存在します。しかしながら、生じうる結果のパターンはこれで全てではなく、また、ここまで説明してきた分散分析モデルは、あらゆるグループ平均のパターンを網羅しているわけではありません。何故でしょうか？それは、これまでの説明では、薬が気分に影響を与える、セラピーが気分に影響を与える、ということについては議論できますが、両者の交互作用を扱うことができないからです。要因Aと要因Bの交互作用は、要因Aの効果が要因Bの水準に応じて異なる場合には、いつでも生じると言われています。2×2要因の分散分析における、いくつかの交互作用効果の例をFigure ??に示します。より具体的な例を挙げると、アンザイフリーアンドジョイゼパムの作用機序が、全く異なる生理学的メカニズムに依存していると仮定します。ここから、ジョイゼパムがセラピーの有無に関わらず気分に対してほぼ同じ影響をもたらす一方で、アンザイフリーはCBTと組み合わせて投与された場合にはるかに効果的であると考えます。前の章で説明した分散分析では、このアイデアを検討できません。交互作用が生じているかどうかを確かめるには、グループ平均を図示することが有効です。JASPでは、分散分析の「Descriptives Plots」オプションを用いて行うことができます—単に、薬を「Horizontal axis」のボックスに、セラ

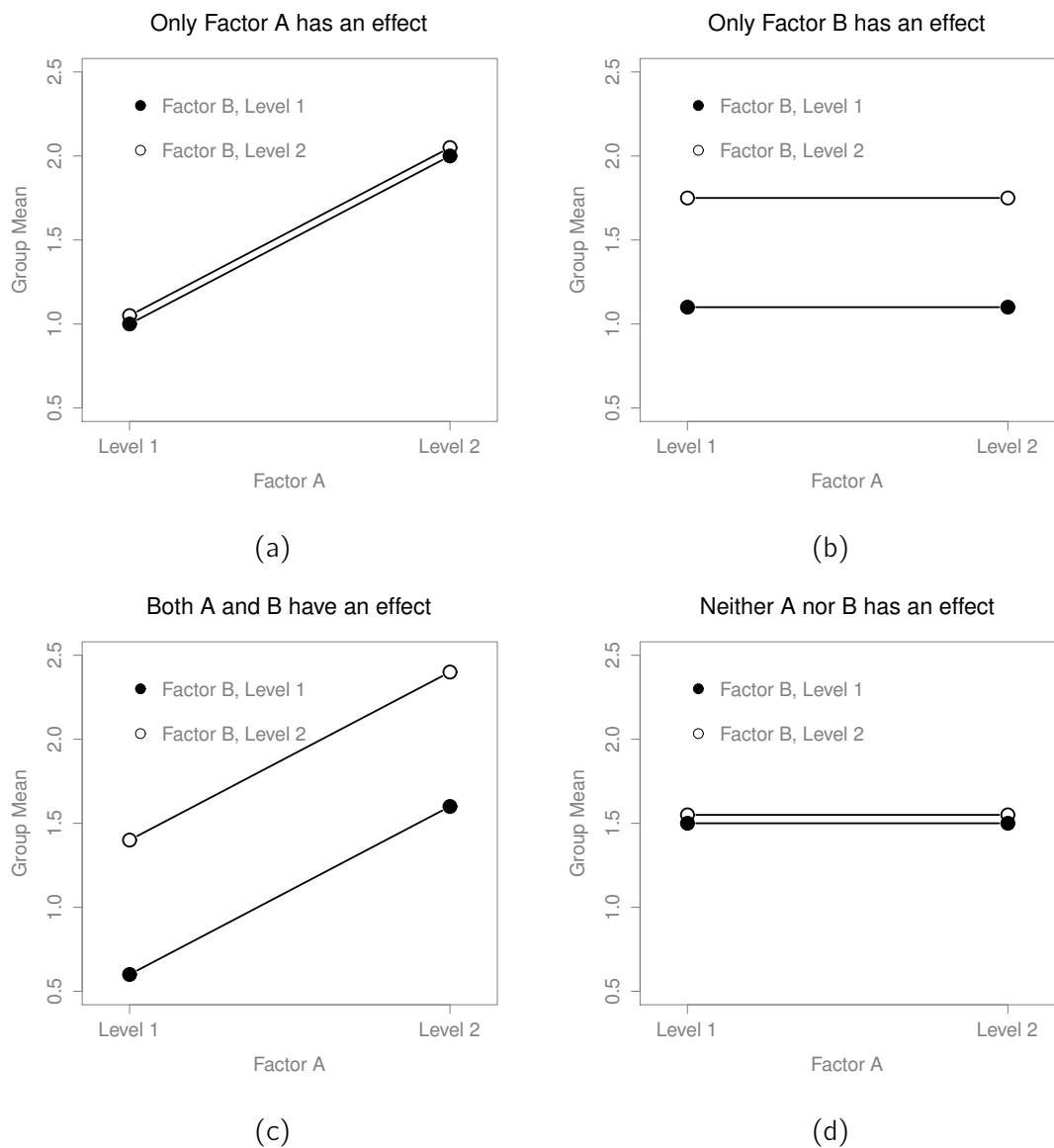


Figure12.5 交互作用のない  $2 \times 2$  要因の分散分析における 4 つの出力。パネル (a) は要因 A の主効果があり、要因 B の主効果がない場合。パネル (b) は要因 A の主効果がなく、要因 B の主効果がある場合。パネル (c) は要因 A, 要因 B のどちらの主効果もある場合。パネル (d) はどちらの要因の主効果もない場合。

ピーを「Separate Lines」のボックスに移動するだけです。Figure ??と同様の図になるはずです。特に注目すべき点は、2本の線が並行ではないということです。ジョイゼパムが投与された場合(中央)のCBTの効果(黒丸の線と白丸の線の差)はゼロに近く、プラセボが用いられた場合のCBTの効果(右側)よりもさらに小さいようです。しかしながら、アンザイフリーが投与されると、CBTの効果はプラセボよりも大きくなります(左側)。この効果は真実でしょうか、それともランダムな変動による単なる偶然なのでしょうか？前章までの分散分析では、この問い合わせに答えることができません。なぜなら、交互作用が存在するというアイデアが含まれていないからです！本章では、この問題点を修正していきます。

### 12.2.1 相互作用とは正確にはどのようなものか？

この節では、交互作用効果という重要なアイデアを紹介します。ここまで見てきた分散分析モデルでは、モデルに含まれる2つの要因(i.e., 薬およびセラピー)にしか着目していませんでした。交互作用を投入することで、モデルに新たな要素が追加されることになります：薬とセラピーの組み合わせです。直観的には、交互作用効果の背後にある考え方は非常に単純です。交互作用は単に、要因Aの影響が、要因Bの水準に応じて変化することを意味しています。しかし、このことは我々のデータに対してどのような意味を持っているのでしょうか？Figure ??に示したいいくつかの図は、それぞれ全く見た目が異なりますが、すべて交互作用効果として扱われます。したがって、この定性的なアイデアを、統計学者が扱うような数学的な記述に変換することは非常に困難です。

結論として、交互作用効果の概念を、対立仮説と帰無仮説の観点から定式化することは困難であり、ましてやこの本の読者の多くは、それほど興味がないと思います。それでも、基本的なアイデアを示しておこうと思います。

まずは、主効果についてもう少し明示的にする必要があります。要因A(今回の分析例では薬)の主効果について考えます。そもそも主効果は、2つの周辺平均 $\mu_{r..}$ が全て等しいという帰無仮説の観点に基づいて定式化されていました。これらの周辺平均の全てが等しいとすると、それらは総平均 $\mu_{...}$ とも等しくなければなりませんね？したがって、ここでは水準 $r$ における要因Aの効果を、周辺平均 $\mu_{r..}$ と総平均 $\mu_{...}$ との差に等しいものとして定義します。

この効果を $\alpha_r$ で表し、以下のように表記します

$$\alpha_r = \mu_{r..} - \mu_{...}$$

ここで、周辺平均 $\mu_{r..}$ の平均値が総平均 $\mu_{...}$ になることと同様の理由で、定義上、すべての $\alpha_r$ の値は合計がゼロになる必要があります。同じように、水準 $i$ における要因Bの効果を、列方向の周

辺平均  $\mu_{..c}$  と総平均  $\mu_{..}$  の差として定義することができます。

$$\beta_c = \mu_{..c} - \mu_{..}$$

繰り返しになりますが、これらの  $\beta_c$  の合計はゼロにならなければなりません。統計学者が  $\alpha_r$  や  $\beta_c$  の値を用いて主効果について説明することを好む理由は、交互作用効果がないということの意味を正確に伝えることができるからです。交互作用がまったくない場合、 $\alpha_r$  および  $\beta_c$  の値を用いて、グループ平均  $\mu_{rc}$  を完全に記述することができます。具体的には以下のようになります

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c$$

これは、グループ平均に関して特別なことは何もない、すなわち、すべての周辺平均が明らかになっても、完全な予測ができないということを意味します。これはまさに、帰無仮説を表しています。対立仮説は、表中の少なくとも一つのグループ  $rc$  において

$$\mu_{rc} \neq \mu_{..} + \alpha_r + \beta_c$$

が成り立つこと、と表現できます。統計学者はしばしば、上記の式をやや異なる形式で表現します。彼らは通常、グループ  $rc$  に関する連づけられた交互作用をいくつかの番号によって定義し、厄介なことに  $(\alpha\beta)_{rc}$  と表現し、そして対立仮説を次のように表します

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c + (\alpha\beta)_{rc}$$

ここで、少なくとも 1 つのグループの  $(\alpha\beta)_{rc}$  は非ゼロです。この表記法はやや見苦しいですが、次の節で説明するように、二乗和の計算方法を説明する際には便利です。

### 12.2.2 交互作用の二乗和の計算

交互作用項  $SS_{A:B}$  の二乗和はどのように計算すれば良いでしょうか？まず、先ほどの節において、実際のグループ平均が周辺平均から予測された値との程度異なるかという観点から、交互作用効果をどのように定義したかについて確認すると良いと思います。もちろん、これらの式はすべて、サンプル統計量ではなく母集団のパラメータに関するものであるため、実際にそれらがどのようなものであるかは分かりません。しかしながら、母平均の代わりにサンプル平均を用いることで、それらを推定することができます。要因 A に関して、水準  $r$  における主効果を推定するための良い方法は、サンプルの周辺平均  $\bar{Y}_{r..}$  とサンプルの総平均  $\bar{Y}_{...}$  の差に着目することです。そこで、これを効果の推定値として用います

$$\hat{\alpha}_r = \bar{Y}_{r..} - \bar{Y}_{...}$$

同様に、水準  $c$  における要因 B の主効果の推定値は、以下のように定義できます

$$\hat{\beta}_c = \bar{Y}_{.c} - \bar{Y}_{..}$$

ここで、2つの主効果の SS の値について説明した式を改めて見てみると、それらの効果に関する項は二乗され足し合わされているということに気づくでしょう！ それでは、交互作用項ではどうなっているでしょうか？ その答えは、以下のように、対立仮説のもとでグループ平均  $\mu_{rc}$  に関する式を再変形することで明らかになります

$$\begin{aligned} (\alpha\beta)_{rc} &= \mu_{rc} - \mu_{..} - \alpha_r - \beta_c \\ &= \mu_{rc} - \mu_{..} - (\mu_{r.} - \mu_{..}) - (\mu_{.c} - \mu_{..}) \\ &= \mu_{rc} - \mu_{r.} - \mu_{.c} + \mu_{..} \end{aligned}$$

そして、母平均の代わりにサンプル統計量を代入すると、グループ  $rc$  における交互作用効果の推定は以下のようになります

$$(\hat{\alpha}\hat{\beta})_{rc} = \bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..}$$

ここで、これらの要因 A における R 水準、および要因 B における C 水準のすべての推定値を足し合わせることで、全体的な相互作用に関する二乗和の式が得られます

$$SS_{A:B} = N \sum_{r=1}^R \sum_{c=1}^C (\bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..})^2$$

各グループについて  $N$  個の観測値があるため、 $N$  が掛けられています。SS の値には、グループ間の変動ではなく、交互作用によって説明される観測値間の変動が反映されていると期待されます。

$SS_{A:B}$  の計算式が準備できたので、交互作用項がモデルの一部であることを認識する必要があります（当然のことですが）。モデルにおける全体の平方和である  $SS_M$  は、関連する 3 つの SS 値の合計  $SS_A + SS_B + SS_{A:B}$  に等しくなります。残差平方和  $SS_T - SS_M$  は、残りの変動、すなわち  $SS_T - SS_M$  として定義されますが、交互作用項があることから、以下のようになります

$$SS_R = SS_T - (SS_A + SS_B + SS_{A:B})$$

結果として、残差平方和  $SS_R$  は、交互作用を含まない分散分析よりも小さくなります。

### 12.2.3 交互作用における自由度

交互作用における自由度の計算は、主効果における計算よりも少しだけ複雑です。まずは、分散分析モデルの全体について考えてみましょう。モデルに交互作用効果が含まれる場合、すべてのグル

プに、独自の平均  $\mu_{rc}$  を持つことが許されます。 $R \times C$  要因の分散分析の場合には、これはモデル中に  $R \times C$  通りの統計量と、総平均はすべてのグループ平均の平均値であるという、たった 1 つの制約があることを意味します。そのため、モデル全体としては  $(R \times C) - 1$  の自由度が必要です。しかし、要因 A の主効果には  $R - 1$  の自由度が、要因 B の主効果には  $C - 1$  の自由度があります。このことは、交互作用に関する自由度が、

$$\begin{aligned} df_{A:B} &= (R \times C - 1) - (R - 1) - (C - 1) \\ &= RC - R - C + 1 \\ &= (R - 1)(C - 1) \end{aligned}$$

という式で表されるように、行の要因と列の要因の自由度の積にすぎないということを意味します。残差の自由度についてはどうでしょうか？ 交互作用項によってある程度の自由度が吸収されるため、残りの自由度は少なくなります。具体的には、交互作用を含むモデルが全部で  $(R \times C) - 1$  の自由度を持ち、1 つの総平均を満たすよう制約されているデータセット内に  $N$  個の観測値があるとき、残差の自由度は  $N - (R \times C) - 1 + 1$ 、あるいは単に  $N - (R \times C)$  となる。

#### 12.2.4 JASP による分散分析の実行

JASP の分散分析モデルに交互作用項を加えることは難しくありません。というよりも、交互作用項は分散分析のデフォルトのオプションであるため、何もする必要はありません。すなわち、例えば **薬とセラピー** という 2 要因の分散分析を実行すると、これらの交互作用項- `drug*therapy` -が自動的にモデルに追加されるということです。<sup>\*3</sup> 交互作用項を含めた分散分析を実行すると、??のような結果が得られます。

結局、今回の分析では、有意な薬の主効果 ( $F_{2,12} = 31.7, p < .001$ ) とセラピーの主効果 ( $F_{1,12} = 8.6, p = .013$ ) が見出されますが、交互作用は有意ではありません ( $F_{2,12} = 2.5, p = 0.125$ )。

#### 12.2.5 結果の解釈

多元配置分散分析の結果の解釈の際には、いくつかの非常に重要なポイントがあります。まずは、仮に（例えば）**薬**の有意な主効果が得られたとしても、どの薬が他と異なるかについては何も分からぬという、一元配置分散分析と同様の問題です。これを明らかにするためには、追加の分析を行う必要があります。Sections ??および??において、この追加の分析のいくつかを紹介します。交互作用効果についても、同様のことがいえます。有意な交互作用がみられたとしても、どのようなパターンの交互作用が存在しているかについては何も分かりません。ここでも、追加の分析を行う必要が

---

<sup>\*3</sup>先ほどの節で説明した、主効果に関する分析を実際に JASP で再現してみた読者は、すでにこの出力を目にしているかもしれません。説明を単純にするため、先述のモデルからは交互作用項を除外しています

あります。

次に、有意な交互作用効果が得られているが、主効果が有意ではないという場合に、解釈が非常にややこしくなるという問題があります。このような結果はときどき生じます。例えば、Figure ??で示されている交差したパターンの交互作用が、実際の結果でも生じる可能性があります。このケースでは、主効果はいずれも有意ではありませんが、交互作用効果が存在します。これは解釈が難しい状況であり、多くの分析者を混乱させます。こうした状況に対して、統計学者が行いがちな一般的なアドバイスは、交互作用が存在する場合、主効果にはあまり注目すべきでないということです。彼らがこのように述べる理由は、主効果の検定は数学的観点から全く正しいのですが、有意な交互作用がみられる場合には、主効果が重要な仮説を検定していることは稀だからです。Section ??を思い出してみると、主効果の帰無仮説は周辺平均が相互に等しいというものであり、周辺平均はいくつかの異なるグループ間の値を平均することで計算されていました。交互作用が有意であるということは、周辺平均を構成するグループが同種ではないということが自明になるため、これらの周辺平均について気にする必要は無くなります。

つまり、こういうことです。改めて、臨床的な例を用いて説明しましょう。 $2 \times 2$ 要因デザインで、2種類の恐怖症の治療法 (e.g., 系統的脱感作法と暴露療法), および2種類の不安軽減薬 (e.g., アンザイフリーとジョイゼパム) の比較を行うと考えてみましょう。そして、治療として脱感作法を行った場合、アンザイフリーには効果がなく、治療として暴露療法を行った場合、ジョイゼパムには効果がないという結果が見出されたと仮定します。いずれの薬も、もう一方の治療においては効果的であるとします。これは古典的な交差パターンの交互作用であり、分散分析を実行すると、**薬**の主効果はありませんが、有意な交互作用が見出されます。さて、主効果がないということは、一体何を意味するのでしょうか？もちろん、2つの異なる治療法を平均すると、アンザイフリーとジョイゼパムの平均的な効果は同じであるということです。しかし、なぜ誰もがこのことを気にかけるのでしょうか？恐怖症の治療において、暴露療法と脱感作法を「平均的に」使用することなどできません。これはあまり意味のない考え方です。暴露療法か脱感作法のどちらかを選ぶことになります。一方の治療法では一方の薬が効果的であり、もう一方の治療法ではもう一方の薬が効果的なのです。ここで重要なのは交互作用であり、主効果はある意味で無関係です。

このような事態はしばしば生じます。主効果は周辺平均の検定であり、交互作用が存在する場合には、周辺平均には注目する必要がなくなることがあります。なぜなら、周辺平均が、本来平均すべきではないものを平均化してしまっていることが、交互作用によって示されるからです！もちろん、交互作用が存在するからといって、主効果が無意味であるとは限りません。多くの場合、大きな主効果と、非常に小さな交互作用が得られます。このとき、「薬 A は一般に薬 B よりも効果的である」(大きな薬の主効果があるため)と主張することができますが、次のような表現を加えて主張を微修正する必要があります。「A と B の薬の効果の差は、治療法によって異なった」。いずれにせよ、ここでの主要なポイントは、有意な交互作用が得られたときには、常に一度立ち止まって、その分析の文脈において、主効果が実際には何を意味しているのかを考えることです。主効果が重要であると

自動的に思い込んでいいません。

## 12.3 \_\_\_\_\_

### 効果量

多元配置分散分析の効果量の計算は、一元配置分散分析で用いられるものとかなり似ています (Section ??を参照)。具体的には、特定の項に対する全体的な効果がどれほど大きいかを測る簡単な方法として、 $\eta^2$  を利用できます。前回同様、 $\eta^2$  はその項に関連づけられた平方和を、総平方和で割ることで定義されます。例えば、要因 A の主効果の効果量を求めるには、以下の式が用いられます：

$$\eta_A^2 = \frac{SS_A}{SS_T}$$

この値が、回帰分析における  $R^2$  とほぼ同様に解釈できるという点も、前回と同じです。<sup>\*4</sup> この値は、要因 A の主効果によって説明することができる結果変数の分散の割合を表しています。したがって、この値は 0(影響なし) から 1(結果変数の変動の全て) の範囲を取ります。さらに、モデル内の全ての項から得られる全ての  $\eta^2$  値の合計は、分散分析モデルの合計の  $R^2$  と等しくなります。例えば、分散分析モデルが完全に適合している場合 (i.e., グループ内の変動が全くない場合！),  $\eta^2$  値は 1 になります。もちろん、このような事態が実生活で生じることは滅多にありません。

ただし、多元配置分散分析を行う際には、偏  $\eta^2$  という 2 つ目の効果量の指標が好んで報告されます。偏  $\eta^2$  ( $\eta_p^2$  もしくは  $\eta_p^2$  と表記される場合もあります) の背後にある考え方とは、特定の項についての効果量 (例えば、要因 A の主効果) を求める場合に、モデル内の他の効果 (e.g., 要因 B の主効果) を意図的に無視するというものです。すなわち、これらの他の全ての項の効果がゼロであると仮定して、 $\eta^2$  値がどうなるかを計算するということです。これは実際には非常に計算が簡単です。分母から他の項に関する平方和を削除すればよいのです。つまり、要因 A の主効果についての偏  $\eta^2$  を求める際には、分母は要因 A の SS 値と残差の合計になります。

$$\text{partial } \eta_A^2 = \frac{SS_A}{SS_A + SS_R}$$

この結果は常に  $\eta^2$  よりも大きな数値となります。私は皮肉屋なので、これが偏  $\eta^2$  の人気の理由だと考えています。偏  $\eta^2$  値もまた 0 から 1 の範囲となり、0 は効果がないことを表します。ただし、偏  $\eta^2$  値の大きさについての解釈は少々ややこしくなります。特に、各項の偏  $\eta^2$  値を直接比較することができない点には注意が必要です！たとえば、グループ内の変動性が全くないとすると、 $SS_R = 0$  となります。このことが意味するのは、すべての項の偏  $\eta^2$  値が 1 になるということです。しかしな

<sup>\*4</sup>この章は、文字 R で様々なものを表現することに関して、新記録を打ち立てているかもしれません。これまでのところ、ソフトウェアパッケージ、平均値の表における行の数、モデルの残差、そして回帰における相関係数を指して、R という文字を使っています。大変申し訳ないと思っています。アルファベットの文字数が十分ではないことは明らかです。我々も、R がそれぞれの文脈において指示しているものを明確にするために、かなりの努力を要しているということを申し添えます

がら、これはモデル内のすべての項が等しく重要である、あるいは、それらの大きさが等しいということではありません。これは、モデル内のすべての項のが、残差の変動と比べて大きな効果量を持つことを意味します。各項について比較することはできません。

このことは、具体例を見てみると分かりやすいでしょう。まず、Figure ??において、交互作用項を含まない分散分析の効果量について見てみましょう：

	eta.sq	partial.eta.sq
%drug	0.71	0.79
薬	0.71	0.79
%therapy	0.10	0.34
セラピー	0.10	0.34

$\eta^2$  の値に着目すると、**薬**の要因が**気分の向上**の分散の 71%(i.e.  $\eta^2 = 0.71$ ) を占めているのに対し、**セラピー**は 10% です。これにより、合計で 19% の変動が考慮されないままとなります(つまり、結果の変動の 19% が残差で構成されます)。全体的に見て、この結果は**薬**には非常に大きな効果<sup>\*5</sup>があり、**セラピー**にはわずかな効果があったことを意味します。次に、Figure ??に示されている偏  $\eta^2$  値を見てみましょう。

**セラピー**の効果はそれほど大きくないため、それを調整しても大きな違いはありません。したがって、**薬**の偏  $\eta^2$  は、 $\rho\eta^2 = 0.79$  とそれほど増加しません。対照的に、**薬**の効果は非常に大きかったため、調整することで大きな違いが生じます。**セラピー**の偏  $\eta^2$  を計算すると、 $\rho\eta^2 = 0.34$  まで上昇していることが分かります。自問しなければならないのは、これらの偏  $\eta^2$  の値が実際に何を意味するのか？ ということです。要因 A の主効果に対する偏  $\eta^2$  を解釈する一般的な方法は、それを要因 A のみを変化させた仮想実験の記述として解釈することです。実際の実験では要因 A と B の両方を変化させましたが、要因 A のみを変化させた実験についても簡単に想像することができます。偏  $\eta^2$  統計量は、そのような実験で得られることが予想される、結果変数の分散の量を表します。ただし、このような解釈は、その他の主効果に関する多くの事項と同様に、有意な交互作用効果が存在する場合には、あまり意味がないということに注意が必要です。

交互作用効果といえば、Figure ??のように、交互作用項を含むモデルの効果量を計算したときに得られるものです。JASP では、単に ‘Additional Options’ - ‘Estimates of effect size’ を選択し、必要な変数を選ぶことで計算されます。見ての通り、主効果の  $\eta^2$  値は変化しませんが、偏  $\eta^2$  値は変化します：

	eta.sq	partial.eta.sq
%drug	0.71	0.84
薬	0.71	0.84

---

<sup>\*5</sup>信じられないほどの大きさです。このデータの不自然さが見えてきましたね！

%therapy	0.10	0.42
セラピー	0.10	0.42
%drug*therapy	0.06	0.29
薬*セラピー		

### 12.3.1 推定グループ平均

多くの場合、分散分析の結果と、それに関連する信頼区間に基づいて、すべてのグループ平均の推定値を報告する必要があります。JASP では、Figure ??にあるように、分散分析の‘Additional Options’ - ‘Marginal Means’の機能を用いてこれを行うことができます。交互作用項薬\*セラピーを、アクションボックスに移動するだけです。実行した分散分析が**飽和モデル** (i.e., 考えられるすべての主効果と交互作用効果を含むモデル) である場合、グループ平均の推定値はサンプル平均とまったく同じになります。重要なのは、信頼区間は、グループごとの個別の標準誤差を使用するのではなく、プールされた標準誤差の推定値を用いるということです。

結果を見ると、セラピーを行わなかったプラセボ群における気分向上の推定平均は 0.300 であり、95% 信頼区間は 0.006 から 0.594 でした。各グループについて信頼区間を計算しても、同じ値にはならないということに注意してください。これは、分散分析モデルが分散の均一性を仮定しているので、プールされた標準偏差の推定値を使用するためです。

## 12.4

---

### 仮定の確認

一元配置分散分析と同様に、多元配置分散分析においても、分散の等質性（すべての群の標準偏差が等しい）、残差の正規性、観測の独立性の 3 つが主要な仮定となります。前の 2 つの仮定については、確認することができます。3 つ目の仮定については、測定値間に何らかの特別な関係性が存在するかどうか、自分自身で評価しなければなりません。例えば、時間を独立変数とする反復測定では、時点 1 と時点 2 の観測変数は同じ人物から測定されているため、関係があります。加えて、飽和モデルを使用していない場合（例えば、交互作用項を省略している場合）には、省略されている項は重要ではないという仮定を置いています。この最後の仮定については、省略された項を含めた分散分析を実行し、それらが有意であるかどうかを確認できるため、チェックすることは比較的容易です。分散の等質性と残差の正規性についてはどうでしょうか？結論からいうと、これらをチェックするのはとても簡単です。一元配置分散分析で行ったチェックの方法となんら変わりません。

#### 12.4.1 分散の等質性

Section ?? で述べたように、異なる群やカテゴリ間で標準偏差のプロットを視覚的に比較し、Levene の検定の結果と一致するかどうかを確認するのは良いアイデアです。Levene の検定の理論については、Section ?? で説明したのでここでは触れません。この検定では、モデルが飽和モデル (i.e., すべての項を含む) であることが期待されています。なぜなら、この検定は主に群内の分散に関係しており、飽和モデル以外について適用しても、実際のところあまり意味がないからです。Levene の検定は、JASP の ‘Assumption Checks’ - ‘Homogeneity tests’ オプションで指定でき、その結果は Figure ?? のようになります。Levene の検定が有意でないということは、標準偏差のプロットの目視による確認との矛盾がなければ、分散の等質性の仮定には違反していないと考えて良いことになります。

#### 12.4.2 残差の正規性

一元配置分散分析と同様に、残差の正規性を簡単な方法で検定できます (Section ??を参照)。しかし、一般的には、QQ プロットを用いて残差を視覚的に調べるのが良いと思います。Figure ?? を見てください。

### 12.5

---

## Analysis of Covariance (ANCOVA)

A variation in ANOVA is when you have an additional continuous variable that you think might be related to the dependent variable. This additional variable can be added to the analysis as a covariate, in the aptly named analysis of covariance (ANCOVA).

In ANCOVA the values of the dependent variable are “adjusted” for the influence of the covariate, and then the “adjusted” score means are tested between groups in the usual way. This technique can increase the precision of an experiment, and therefore provide a more “powerful” test of the equality of group means in the dependent variable. How does ANCOVA do this? Well, although the covariate itself is typically not of any experimental interest, adjustment for the covariate can decrease the estimate of experimental error and thus, by reducing error variance, precision is increased. This means that an inappropriate failure to reject the null hypothesis (false negative or type II error) is less likely.

Despite this advantage, ANCOVA runs the risk of undoing real differences between groups,

and this should be avoided. Look at Figure ??, for example, which shows a plot of Statistics anxiety against age and shows two distinct groups – students who have either an Arts or Science background or preference. ANCOVA with age as a covariate might lead to the conclusion that statistics anxiety does not differ in the two groups. Would this conclusion be reasonable – probably not because the ages of the two groups do not overlap and analysis of variance has essentially “extrapolated into a region with no data” (**Everitt1996**).

Clearly, careful thought needs to be given to an analysis of covariance with distinct groups. This applies to both one-way and factorial designs, as ANCOVA can be used with both.

#### 12.5.1 Running ANCOVA in JASP

A health psychologist was interested in the effect of routine cycling and stress on happiness levels, with age as a covariate. You can find the dataset in the file `ancova.csv`. Open this file in JASP and then, to undertake an ANCOVA, select ‘ANOVA’ - ‘ANCOVA’ to open the ANCOVA analysis window (Figure ??). Highlight the dependent variable ‘happiness’ and transfer it into the ‘Dependent Variable’ text box. Highlight the independent variables ‘stress’ and ‘commute’ and transfer them into the ‘Fixed Factors’ text box. Highlight the covariate ‘age’ and transfer it into the ‘Covariates’ text box. Then click on ‘Additional Options’ - ‘Marginal Means’ and transfer the interaction term `stress*commute` into the action box.

An ANCOVA table is produced in the JASP results window (Figure ??). The F value for the covariate ‘age’ is significant at  $p = .023$ , suggesting that age is an important predictor of the dependent variable `happiness`. When we look at the estimated marginal mean scores, adjustments have been made (compared to an analysis without the covariate) because of the inclusion of the covariate ‘age’ in this ANCOVA. A plot (Figure ??) is a good way of visualising and interpreting the significant effects.

The F value for the main effect ‘stress’ (52.61) has an associated probability of  $p < .001$ . The F value for the main effect ‘commute’ (42.33) has an associated probability of  $p < .001$ . Since both of these are less than the probability that is typically used to decide if a statistical result is significant ( $p < .05$ ) we can conclude that there was a significant main effect of stress ( $F(1, 15) = 52.61, p < .001$ ) and a significant main effect of commuting method ( $F(1, 15) = 42.33, p < .001$ ). A significant interaction between stress and commuting method was also found ( $F(1, 15) = 14.15, p = .002$ ).

In Figure ?? we can see the adjusted, marginal, mean happiness scores when age is a covariate in an ANCOVA. In this analysis there is a significant interaction effect, whereby people with low

stress who cycle to work are happier than people with low stress who drive and people with high stress whether they cycle or drive to work. There is also a significant main effect of stress – people with low stress are happier than those with high stress. And there is also a significant main effect of commuting behaviour – people who cycle are happier, on average, than those who drive to work.

## 12.6

---

### ANOVA as a linear model

One of the most important things to understand about ANOVA and regression is that they're basically the same thing. On the surface of it, you maybe wouldn't think this is true. After all, the way that I've described them so far suggests that ANOVA is primarily concerned with testing for group differences, and regression is primarily concerned with understanding the correlations between variables. And, as far as it goes that's perfectly true. But when you look under the hood, so to speak, the underlying mechanics of ANOVA and regression are awfully similar. In fact, if you think about it, you've already seen evidence of this. ANOVA and regression both rely heavily on sums of squares (SS), both make use of  $F$  tests, and so on. Looking back, it's hard to escape the feeling that Chapters ?? and ?? were a bit repetitive.

The reason for this is that ANOVA and regression are both kinds of **linear models**. In the case of regression, this is kind of obvious. The regression equation that we use to define the relationship between predictors and outcomes *is* the equation for a straight line, so it's quite obviously a linear model, with the equation

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

where  $Y_p$  is the outcome value for the  $p$ -th observation (e.g.,  $p$ -th person),  $X_{1p}$  is the value of the first predictor for the  $p$ -th observation,  $X_{2p}$  is the value of the second predictor for the  $p$ -th observation, the  $b_0$ ,  $b_1$ , and  $b_2$  terms are our regression coefficients, and  $\varepsilon_p$  is the  $p$ -th residual. If we ignore the residuals  $\varepsilon_p$  and just focus on the regression line itself, we get the following formula:

$$\hat{Y}_p = b_0 + b_1 X_{1p} + b_2 X_{2p}$$

where  $\hat{Y}_p$  is the value of  $Y$  that the regression line predicts for person  $p$ , as opposed to the actually-observed value  $Y_p$ . The thing that isn't immediately obvious is that we can write ANOVA as a linear model as well. However, it's actually pretty straightforward to do this. Let's start with a really simple example, rewriting a  $2 \times 2$  factorial ANOVA as a linear model.

### 12.6.1 Some data

To make things concrete, let's suppose that our outcome variable is the `grade` that a student receives in my class, a ratio-scale variable corresponding to a mark from 0% to 100%. There are two predictor variables of interest: whether or not the student turned up to lectures (the `attend` variable) and whether or not the student actually read the textbook (the `reading` variable). We'll say that `attend = 1` if the student attended class, and `attend = 0` if they did not. Similarly, we'll say that `reading = 1` if the student read the textbook, and `reading = 0` if they did not.

Okay, so far that's simple enough. The next thing we need to do is to wrap some maths around this (sorry!). For the purposes of this example, let  $Y_p$  denote the `grade` of the  $p$ -th student in the class. This is not quite the same notation that we used earlier in this chapter. Previously, we've used the notation  $Y_{rci}$  to refer to the  $i$ -th person in the  $r$ -th group for predictor 1 (the row factor) and the  $c$ -th group for predictor 2 (the column factor). This extended notation was really handy for describing how the SS values are calculated, but it's a pain in the current context, so I'll switch notation here. Now, the  $Y_p$  notation is visually simpler than  $Y_{rci}$ , but it has the shortcoming that it doesn't actually keep track of the group memberships! That is, if I told you that  $Y_{0,0,3} = 35$ , you'd immediately know that we're talking about a student (the 3rd such student, in fact) who didn't attend the lectures (i.e., `attend = 0`) and didn't read the textbook (i.e. `reading = 0`), and who ended up failing the class (`grade = 35`). But if I tell you that  $Y_p = 35$ , all you know is that the  $p$ -th student didn't get a good grade. We've lost some key information here. Of course, it doesn't take a lot of thought to figure out how to fix this. What we'll do instead is introduce two new variables  $X_{1p}$  and  $X_{2p}$  that keep track of this information. In the case of our hypothetical student, we know that  $X_{1p} = 0$  (i.e., `attend = 0`) and  $X_{2p} = 0$  (i.e., `reading = 0`). So the data might look like this:

person, $p$	grade, $Y_p$	attendance, $X_{1p}$	reading, $X_{2p}$
1	90	1	1
2	87	1	1
3	75	0	1
4	60	1	0
5	35	0	0
6	50	0	0
7	65	1	0
8	70	0	1

This isn't anything particularly special, of course. It's exactly the format in which we expect to

see our data! See the data file `rtfm.csv`.

### 12.6.2 ANOVA with binary factors as a regression model

Okay, let's get back to talking about the mathematics. We now have our data expressed in terms of three numeric variables: the continuous variable  $Y$  and the two binary variables  $X_1$  and  $X_2$ . What I want you to recognise is that our  $2 \times 2$  factorial ANOVA is *exactly* equivalent to the regression model

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

This is, of course, the exact same equation that I used earlier to describe a two-predictor regression model! The only difference is that  $X_1$  and  $X_2$  are now *binary* variables (i.e., values can only be 0 or 1), whereas in a regression analysis we expect that  $X_1$  and  $X_2$  will be continuous. There's a couple of ways I could try to convince you of this. One possibility would be to do a lengthy mathematical exercise proving that the two are identical. However, I'm going to go out on a limb and guess that most of the readership of this book will find that annoying rather than helpful. Instead, I'll explain the basic ideas and then rely on JASP to show that ANOVA analyses and regression analyses aren't just similar, they're identical for all intents and purposes. Let's start by running this as an ANOVA. To do this, we'll use the `rtfm` data set, and Figure ?? shows what we get when we run the analysis in JASP.

So, by reading the key numbers off the ANOVA table and the mean scores that we presented earlier, we can see that the students obtained a higher grade if they attended class ( $F_{1,5} = 21.6, p = .006$ ) and if they read the textbook ( $F_{1,5} = 52.3, p < 0.001$ ). Let's make a note of those  $p$ -values and those  $F$  statistics.

Now let's think about the same analysis from a linear regression perspective. In the `rtfm` data set, we have encoded `attend` and `reading` as if they were numeric predictors. In this case, this is perfectly acceptable. There really is a sense in which a student who turns up to class (i.e. `attend = 1`) has in fact done "more attendance" than a student who does not (i.e. `attend = 0`). So it's not at all unreasonable to include it as a predictor in a regression model. It's a little unusual, because the predictor only takes on two possible values, but it doesn't violate any of the assumptions of linear regression. And it's easy to interpret. If the regression coefficient for `attend` is greater than 0 it means that students that attend lectures get higher grades. If it's less than zero then students attending lectures get lower grades. The same is true for our `reading` variable.

Wait a second though. *Why* is this true? It's something that is intuitively obvious to everyone who has taken a few stats classes and is comfortable with the maths, but it *isn't* clear to everyone

else at first pass. To see why this is true, it helps to look closely at a few specific students. Let's start by considering the 6th and 7th students in our data set (i.e.  $p = 6$  and  $p = 7$ ). Neither one has read the textbook, so in both cases we can set `reading = 0`. Or, to say the same thing in our mathematical notation, we observe  $X_{2,6} = 0$  and  $X_{2,7} = 0$ . However, student number 7 did turn up to lectures (i.e., `attend = 1`,  $X_{1,7} = 1$ ) whereas student number 6 did not (i.e., `attend = 0`,  $X_{1,6} = 0$ ). Now let's look at what happens when we insert these numbers into the general formula for our regression line. For student number 6, the regression predicts that

$$\begin{aligned}\hat{Y}_6 &= b_0 + b_1 X_{1,6} + b_2 X_{2,6} \\ &= b_0 + (b_1 \times 0) + (b_2 \times 0) \\ &= b_0\end{aligned}$$

So we're expecting that this student will obtain a grade corresponding to the value of the intercept term  $b_0$ . What about student 7? This time when we insert the numbers into the formula for the regression line, we obtain the following

$$\begin{aligned}\hat{Y}_7 &= b_0 + b_1 X_{1,7} + b_2 X_{2,7} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &= b_0 + b_1\end{aligned}$$

Because this student attended class, the predicted grade is equal to the intercept term  $b_0$  *plus* the coefficient associated with the `attend` variable,  $b_1$ . So, if  $b_1$  is greater than zero, we're expecting that the students who turn up to lectures will get higher grades than those students who don't. If this coefficient is negative we're expecting the opposite: students who turn up at class end up performing much worse. In fact, we can push this a little bit further. What about student number 1, who turned up to class ( $X_{1,1} = 1$ ) *and* read the textbook ( $X_{2,1} = 1$ )? If we plug these numbers into the regression we get

$$\begin{aligned}\hat{Y}_1 &= b_0 + b_1 X_{1,1} + b_2 X_{2,1} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ &= b_0 + b_1 + b_2\end{aligned}$$

So if we assume that attending class helps you get a good grade (i.e.,  $b_1 > 0$ ) and if we assume that reading the textbook also helps you get a good grade (i.e.,  $b_2 > 0$ ), then our expectation is that student 1 will get a grade that is higher than student 6 and student 7.

And at this point you won't be at all surprised to learn that the regression model predicts that student 3, who read the book but didn't attend lectures, will obtain a grade of  $b_2 + b_0$ . I won't bore you with yet another regression formula. Instead, what I'll do is show you the following table of *expected grades*:

		read textbook?	
		no	yes
attended?	no	$b_0$	$b_0 + b_2$
	yes	$b_0 + b_1$	$b_0 + b_1 + b_2$

As you can see, the intercept term  $b_0$  acts like a kind of “baseline” grade that you would expect from those students who don’t take the time to attend class or read the textbook. Similarly,  $b_1$  represents the boost that you’re expected to get if you come to class, and  $b_2$  represents the boost that comes from reading the textbook. In fact, if this were an ANOVA you might very well want to characterise  $b_1$  as the main effect of attendance, and  $b_2$  as the main effect of reading! In fact, for a simple  $2 \times 2$  ANOVA that’s *exactly* how it plays out.

Okay, now that we’re really starting to see why ANOVA and regression are basically the same thing, let’s actually run our regression using the `rtfm` data and the JASP regression analysis to convince ourselves that this is really true. Running the regression in the usual way gives the results shown in Figure ??.

There’s a few interesting things to note here. First, notice that the intercept term is 43.5 which is close to the “group” mean of 42.5 observed for those two students who didn’t read the text or attend class. Second, notice that we have the regression coefficient of  $b_1 = 18.0$  for the attendance variable, suggesting that those students that attended class scored 18 points higher than those who didn’t. So our expectation would be that those students who turned up to class but didn’t read the textbook would obtain a grade of  $b_0 + b_1$ , which is equal to  $43.5 + 18.0 = 61.5$ . You can verify for yourself that the same thing happens when we look at the students that read the textbook.

Actually, we can push a little further in establishing the equivalence of our ANOVA and our regression. Look at the  $p$ -values associated with the `attend` variable and the `reading` variable in the regression output. They’re identical to the ones we encountered earlier when running the ANOVA. This might seem a little surprising, since the test used when running our regression model calculates a  $t$ -statistic and the ANOVA calculates an  $F$ -statistic. However, if you can remember all the way back to Chapter ??, I mentioned that there’s a relationship between the  $t$ -distribution and the  $F$ -distribution. If you have some quantity that is distributed according to a  $t$ -distribution with  $k$  degrees of freedom and you square it, then this new squared quantity follows an  $F$ -distribution whose degrees of freedom are 1 and  $k$ . We can check this with respect to the  $t$  statistics in our regression model. For the `attend` variable we get a  $t$  value of 4.65. If we square this number we end up with 21.6, which matches the corresponding  $F$  statistic in our ANOVA.

### 12.6.3 Using contrasts to encode non binary factors

At this point, I've shown you how we can view a  $2 \times 2$  ANOVA into a linear model. And it's pretty easy to see how this generalises to a  $2 \times 2 \times 2$  ANOVA or a  $2 \times 2 \times 2 \times 2$  ANOVA. It's the same thing, really. You just add a new binary variable for each of your factors. Where it begins to get trickier is when we consider factors that have more than two levels. Consider, for instance, the  $3 \times 2$  ANOVA that we ran earlier in this chapter using the `clinicaltrial.csv` data. How can we convert the three-level `drug` factor into a numerical form that is appropriate for a regression?

The answer to this question is pretty simple, actually. All we have to do is realise that a three-level factor can be redescribed as *two* binary variables. Suppose, for instance, I were to create a new binary variable called `druganxitfree`. Whenever the `drug` variable is equal to "`anxitfree`" we set `druganxitfree = 1`. Otherwise, we set `druganxitfree = 0`. This variable sets up a **contrast**, in this case between `anxitfree` and the other two drugs. By itself, of course, the `druganxitfree` contrast isn't enough to fully capture all of the information in our `drug` variable. We need a second contrast, one that allows us to distinguish between `joyzepam` and the placebo. To do this, we can create a second binary contrast, called `drugjoyzepam`, which equals 1 if the drug is `joyzepam` and 0 if it is not. Taken together, these two contrasts allows us to perfectly discriminate between all three possible drugs. The table below illustrates this:

<code>drug</code>	<code>druganxitfree</code>	<code>drugjoyzepam</code>
"placebo"	0	0
"anxitfree"	1	0
"joyzepam"	0	1

If the drug administered to a patient is a placebo then both of the two contrast variables will equal 0. If the drug is Anxitfree then the `druganxitfree` variable will equal 1, and `drugjoyzepam` will be 0. The reverse is true for Joyzepam: `drugjoyzepam` is 1 and `druganxitfree` is 0.

Creating contrast variables is not too difficult to do using JASP's 'Computed Column' functionality. For example, to create the `druganxitfree` variable, you can use some simple R code to create the binary variable in the formula box: `ifelse(drug == 'anxitfree', 1, 0)`. Similarly, to create the new variable `drugjoyzepam` use this code: `ifelse(drug == 'joyzepam', 1, 0)`. Likewise for `therapyCBT`: `ifelse(therapy == 'CBT', 1, 0)`. You can see these new variables in the JASP data file `clinicaltrial2.jasp`. As before, if you click on the " $f_x$ " symbol, you will also be able to see the associated R code.

We have now recoded our three-level factor in terms of two binary variables, and we've already seen that ANOVA and regression behave the same way for binary variables. However, there are

some additional complexities that arise in this case, which we'll discuss in the next section.

#### 12.6.4 The equivalence between ANOVA and regression for non-binary factors

Now we have two different versions of the same data set. Our original data in which the `drug` variable from the `clinicaltrial.csv` file is expressed as a single three-level factor, and the expanded data `clinicaltrial2.omv` in which it is expanded into two binary contrasts. Once again, the thing that we want to demonstrate is that our original  $3 \times 2$  factorial ANOVA is equivalent to a regression model applied to the contrast variables. Let's start by re-running the ANOVA, with results shown in Figure ???. Let's remove the interaction component; remember to go to 'Model' and move the `drug*therapy` term out of the 'Model Terms' box.

Obviously, there are no surprises here. That's the exact same ANOVA that we ran earlier. Next, let's run a regression using `druganxitfree`, `drugjoyzepam` and `CBTtherapy` as the predictors. The results are shown in Figure ??.

Hmm. This will look quite different than what we got last time. Not surprisingly, the regression output prints out the results for each of the three predictors separately, just like it did every other time we conducted a regression analysis. On the one hand we can see that the *p*-value for the `CBTtherapy` variable is exactly the same as the one for the `therapy` factor in our original ANOVA, so we can be reassured that the regression model is doing the same thing as the ANOVA did. On the other hand, this regression model is testing the `druganxitfree` contrast and the `drugjoyzepam` contrast *separately*, as if they were two completely unrelated variables. It's not surprising of course, because the poor regression analysis has no way of knowing that `drugjoyzepam` and `druganxitfree` are actually the two different contrasts that we used to encode our three-level `drug` factor. As far as it knows, `drugjoyzepam` and `druganxitfree` are no more related to one another than `drugjoyzepam` and `therapyCBT`. However, you and I know better. At this stage we're not at all interested in determining whether these two contrasts are individually significant. We just want to know if there's an "overall" effect of drug. That is, what we want JASP to do is to run some kind of "model comparison" test, one in which the two "drug-related" contrasts are lumped together for the purpose of the test. Sound familiar? All we need to do is specify a "null model", which in this case would include the `CBTtherapy` predictor, and omit both of the drug-related variables, as in Figure ???. To do this, click the 'Model' button and check 'Add to null model' for both of the drug terms. You can then assess whether to include `drug` as a predictor in the model by computing AIC

for both models (see Section ??). Remember, smaller is better!

## 12.7

---

### Different ways to specify contrasts

In the previous section, I showed you a method for converting a factor into a collection of contrasts. In the method I showed you we specify a set of binary variables in which we defined a table like this one:

drug	druganxitfree	drugjoyzepam
"placebo"	0	0
"anxitfree"	1	0
"joyzepam"	0	1

Each row in the table corresponds to one of the factor levels, and each column corresponds to one of the contrasts. This table, which always has one more row than columns, has a special name. It is called a **contrast matrix**. However, there are lots of different ways to specify a contrast matrix. In this section I discuss a few of the standard contrast matrices that statisticians use and how you can use them in JASP. If you're planning to read the section on unbalanced ANOVA later on (Section ??), it's worth reading this section carefully. If not, you can get away with skimming it, because the choice of contrasts doesn't matter much for balanced designs.

#### 12.7.1 Treatment contrasts

In the particular kind of contrasts that I've described above, one level of the factor is special, and acts as a kind of "baseline" category (i.e., `placebo` in our example), against which the other two are defined. The name for these kinds of contrasts is **treatment contrasts**, also known as "dummy coding". In this contrast each level of the factor is compared to a base reference level, and the base reference level is the value of the intercept.

The name reflects the fact that these contrasts are quite natural and sensible when one of the categories in your factor really is special because it actually does represent a baseline. That makes sense in our clinical trial example. The `placebo` condition corresponds to the situation where you don't give people any real drugs, and so it's special. The other two conditions are defined in relation to the placebo. In one case you replace the placebo with Anxitfree, and in the other case you replace it with Joyzepam.

The table shown above is a matrix of treatment contrasts for a factor that has 3 levels. But suppose I want a matrix of treatment contrasts for a factor with 5 levels? You would set this out like this:

Level	2	3	4	5
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

In this example, the first contrast is level 2 compared with level 1, the second contrast is level 3 compared with level 1, and so on. Notice that, by default, the *first* level of the factor is always treated as the baseline category (i.e., it's the one that has all zeros and doesn't have an explicit contrast associated with it). In JASP you can change which category is the first level of the factor by manipulating the order of the levels of the variable shown in the 'Data Variable' window (double click on the name of the variable in the spreadsheet column to bring up the 'Data Variable' view).

### 12.7.2 Helmert contrasts

Treatment contrasts are useful for a lot of situations. However, they make most sense in the situation when there really is a baseline category, and you want to assess all the other groups in relation to that one. In other situations, however, no such baseline category exists, and it may make more sense to compare each group to the mean of the other groups. This is where we meet **Helmert contrasts**, generated by the 'helmert' option in the JASP 'ANOVA' - 'Contrasts' selection box. The idea behind Helmert contrasts is to compare each group to the mean of the "previous" ones. That is, the first contrast represents the difference between group 2 and group 1, the second contrast represents the difference between group 3 and the mean of groups 1 and 2, and so on. This translates to a contrast matrix that looks like this for a factor with five levels:

1	-1	-1	-1	-1
2	1	-1	-1	-1
3	0	2	-1	-1
4	0	0	3	-1
5	0	0	0	4

One useful thing about Helmert contrasts is that every contrast sums to zero (i.e., all the columns sum to zero). This has the consequence that, when we interpret the ANOVA as a regression, the intercept term corresponds to the grand mean  $\mu_{..}$  if we are using Helmert contrasts. Compare

this to treatment contrasts, in which the intercept term corresponds to the group mean for the baseline category. This property can be very useful in some situations. It doesn't matter very much if you have a balanced design, which we've been assuming so far, but it will turn out to be important later when we consider unbalanced designs in Section ???. In fact, the main reason why I've even bothered to include this section is that contrasts become important if you want to understand unbalanced ANOVA.

#### 12.7.3 Sum to zero contrasts

The third option that I should briefly mention are “sum to zero” contrasts, called “Simple” contrasts in JASP, which are used to construct pairwise comparisons between groups. Specifically, each contrast encodes the difference between one of the groups and a baseline category, which in this case corresponds to the first group:

1	-1	-1	-1	-1
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

Much like Helmert contrasts, we see that each column sums to zero, which means that the intercept term corresponds to the grand mean when ANOVA is treated as a regression model. When interpreting these contrasts, the thing to recognise is that each of these contrasts is a pairwise comparison between group 1 and one of the other four groups. Specifically, contrast 1 corresponds to a “group 2 minus group 1” comparison, contrast 2 corresponds to a “group 3 minus group 1” comparison, and so on.<sup>\*6</sup>

#### 12.7.4 Optional contrasts in JASP

JASP also comes with a variety of options that can generate different kinds of contrasts in ANOVA. These can be found in the ‘Contrasts’ option in the main ANOVA analysis window, where the following contrast types are listed:

---

<sup>\*6</sup>What's the difference between treatment and simple contrasts, I hear you ask? Well, as a basic example consider a gender main effect, with  $m=0$  and  $f=1$ . The coefficient corresponding to the treatment contrast will measure the difference in mean between females and males, and the intercept would be the mean of the males. However, with a simple contrast, i.e.,  $m=-1$  and  $f=1$ , the intercept is the average of the means and the main effect is the difference of each group mean from the intercept.

Contrast type	
Deviation	Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean)
Simple	Like the treatment contrasts, the simple contrast compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group. By default the first category is the reference. However, with a simple contrast the intercept is the grand mean of all the levels of the factors.
Difference	Compares the mean of each level (except the first) to the mean of previous levels. (Sometimes called reverse Helmert contrasts)
Helmert	Compares the mean of each level of the factor (except the last) to the mean of subsequent levels
Repeated	Compares the mean of each level (except the last) to the mean of the subsequent level
Polynomial	Compares the linear effect and quadratic effect. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect. These contrasts are often used to estimate polynomial trends

## 12.8

---

### Post hoc tests

Time to switch to a different topic. Rather than pre-planned comparisons that you have tested using contrasts, let's suppose you've done your ANOVA and it turns out that you obtained some significant effects. Because of the fact that the  $F$ -tests are "omnibus" tests that only really

test the null hypothesis that there are no differences among groups, obtaining a significant effect doesn't tell you which groups are different to which other ones. We discussed this issue back in Chapter ??, and in that chapter our solution was to run *t*-tests for all possible pairs of groups, making corrections for multiple comparisons (e.g., Bonferroni, Holm) to control the Type I error rate across all comparisons. The methods that we used back in Chapter ?? have the advantage of being relatively simple and being the kind of tools that you can use in a lot of different situations where you're testing multiple hypotheses, but they're not necessarily the best choices if you're interested in doing efficient post hoc testing in an ANOVA context. There are actually quite a lot of different methods for performing multiple comparisons in the statistics literature (**Hsu1996**), and it would be beyond the scope of an introductory text like this one to discuss all of them in any detail.

That being said, there's one tool that I do want to draw your attention to, namely Tukey's "Honestly Significant Difference", or **Tukey's HSD** for short. For once, I'll spare you the formulas and just stick to the qualitative ideas. The basic idea in Tukey's HSD is to examine all relevant pairwise comparisons between groups, and it's only really appropriate to use Tukey's HSD if it is *pairwise* differences that you're interested in.<sup>\*7</sup> For instance, earlier we conducted a factorial ANOVA using the `clinicaltrial.csv` data set, and where we specified a main effect for drug and a main effect of therapy we would be interested in the following four comparisons:

- The difference in mood gain for people given Anxifree versus people given the placebo.
- The difference in mood gain for people given Joyzepam versus people given the placebo.
- The difference in mood gain for people given Anxifree versus people given Joyzepam.
- The difference in mood gain for people treated with CBT and people given no therapy.

For any one of these comparisons, we're interested in the true difference between (population) group means. Tukey's HSD constructs **simultaneous confidence intervals** for all four of these comparisons. What we mean by 95% "simultaneous" confidence interval is that, if we were to repeat this study many times, then in 95% of the study results the confidence intervals would contain the relevant true value. Moreover, we can use these confidence intervals to calculate an adjusted *p* value for any specific comparison.

The `TukeyHSD` function in JASP is pretty easy to use. You simply specify the ANOVA model term that you want to run the post hoc tests for. For example, if we were looking to run post

---

<sup>\*7</sup>If, for instance, you actually find yourself interested to know if Group A is significantly different from the mean of Group B and Group C, then you need to use a different tool (e.g., Scheffe's method, which is more conservative, and beyond the scope of this book). However, in most cases you probably are interested in pairwise group differences, so Tukey's HSD is a pretty useful thing to know about.

hoc tests for the main effects but not the interaction, we would open up the ‘Post Hoc Tests’ option in the ANOVA analysis screen, move the `drug` and `therapy` variables across to the box on the right, and then select the ‘Tukey’ checkbox in the list of possible post hoc corrections that could be applied. This, along with the corresponding results table, is shown in Figure ??

The output shown in the ‘Post Hoc Tests’ results table is (I hope) pretty straightforward. The first comparison, for example, is the Anxifree versus placebo difference, and the first part of the output indicates that the observed difference in group means is .27. The next number is the standard error for the difference, from which we could calculate the 95% confidence interval (you can actually display this 95% CI if you select the ‘Confidence intervals’ box). Then there is a column with the degrees of freedom, a column with the *t*-value, and finally a column with the *p*-value. For the first comparison the adjusted *p*-value is .21. In contrast, if you look at the next line, we see that the observed difference between joyzepam and the placebo is 1.03, and this result is significant (*p* < .001).

So far, so good. What about the situation where your model includes interaction terms? For instance, the default option in JASP is to allow for the possibility that there is an interaction between drug and therapy. If that’s the case, the number of pairwise comparisons that we need to consider starts to increase. As before, we need to consider the three comparisons that are relevant to the main effect of `drug` and the one comparison that is relevant to the main effect of `therapy`. But, if we want to consider the possibility of a significant interaction (and try to find the group differences that underpin that significant interaction), we need to include comparisons such as the following:

- The difference in mood gain for people given Anxifree and treated with CBT, versus people given the placebo and treated with CBT
- The difference in mood gain for people given Anxifree and given no therapy, versus people given the placebo and given no therapy.
- etc

There are quite a lot of these comparisons that you need to consider. So, when we run the Tukey post hoc analysis for this ANOVA model, we see that it has made a *lot* of pairwise comparisons (19 in total), as shown in Figure ?? . You can see that it looks pretty similar to before, but with a lot more comparisons made.

## The method of planned comparisons

Following on from the previous sections on contrasts and post hoc tests in ANOVA, I think the method of planned comparisons is important enough to deserve a quick discussion. In our discussions of multiple comparisons, in the previous section and back in Chapter ??, I've been assuming that the tests you want to run are genuinely post hoc. For instance, in our drugs example above, maybe you thought that the drugs would all have different effects on mood (i.e., you hypothesised a main effect of drug), but you didn't have any specific hypothesis about how they would be different, nor did you have any real idea about *which* pairwise comparisons would be worth looking at. If that is the case, then you really have to resort to something like Tukey's HSD to do your pairwise comparisons.

The situation is rather different, however, if you genuinely did have real, specific hypotheses about which comparisons are of interest, and you *never ever* have any intention to look at any other comparisons besides the ones that you specified ahead of time. When this is true, and if you honestly and rigorously stick to your noble intentions to not run any other comparisons (even when the data look like they're showing you deliciously significant effects for stuff you didn't have a hypothesis test for), then it doesn't really make a lot of sense to run something like Tukey's HSD, because it makes corrections for a whole bunch of comparisons that you never cared about and never had any intention of looking at. Under those circumstances, you can safely run a (limited) number of hypothesis tests without making an adjustment for multiple testing. This situation is known as the **method of planned comparisons**, and it is sometimes used in clinical trials. However, further consideration is out of scope for this introductory book, but at least you know that this method exists!

12.10

---

## Factorial ANOVA 3: unbalanced designs

Factorial ANOVA is a very handy thing to know about. It's been one of the standard tools used to analyse experimental data for many decades, and you'll find that you can't read more than two or three papers in psychology without running into an ANOVA in there somewhere. However, there's one huge difference between the ANOVAs that you'll see in a lot of real scientific articles and the ANOVAs that I've described so far. In real life we're rarely lucky enough to have perfectly balanced designs. For one reason or another, it's typical to end up with more observations in some cells than in others. Or, to put it another way, we have an **unbalanced design**.

Unbalanced designs need to be treated with a lot more care than balanced designs, and the statistical theory that underpins them is a lot messier. It might be a consequence of this messiness, or it might be a shortage of time, but my experience has been that undergraduate research methods classes in psychology have a nasty tendency to ignore this issue completely. A lot of stats textbooks tend to gloss over it too. The net result of this, I think, is that a lot of active researchers in the field don't actually know that there's several different "types" of unbalanced ANOVAs, and they produce quite different answers. In fact, reading the psychological literature, I'm kind of amazed at the fact that most people who report the results of an unbalanced factorial ANOVA don't actually give you enough details to reproduce the analysis. I secretly suspect that most people don't even realise that their statistical software package is making a whole lot of substantive data analysis decisions on their behalf. It's actually a little terrifying when you think about it. So, if you want to avoid handing control of your data analysis to stupid software, read on.

#### 12.10.1 **The coffee data**

As usual, it will help us to work with some data. The `coffee.csv` file contains a hypothetical data set that produces an unbalanced  $3 \times 2$  ANOVA. Suppose we were interested in finding out whether or not the tendency of people to `babble` when they have too much coffee is purely an effect of the coffee itself, or whether there's some effect of the `milk` and `sugar` that people add to the coffee. Suppose we took 18 people and gave them some coffee to drink. The amount of coffee / caffeine was held constant, and we varied whether or not milk was added, so `milk` is a binary factor with two levels, `"yes"` and `"no"`. We also varied the kind of sugar involved. The coffee might contain `"real"` sugar or it might contain `"fake"` sugar (i.e., artificial sweetener) or it might contain `"none"` at all, so the `sugar` variable is a three level factor. Our outcome variable is a continuous variable that presumably refers to some psychologically sensible measure of the extent to which someone is "babbling". The details don't really matter for our purpose. Take a look at the data in the JASP spreadsheet view, as in Figure ??.

Looking at the table of means in Figure ?? we get a strong impression that there are differences between the groups. This is especially true when we look at the standard deviations. Across groups, this standard deviation varies quite a lot.\*<sup>8</sup> Whilst this at first may seem like a straightforward factorial ANOVA, a problem arises when we look at how many observations we have in each

---

\*<sup>8</sup>This discrepancy in standard deviations might (and should) make you wonder if we have a violation of the homogeneity of variance assumption. I'll leave it as an exercise for the reader to double check this using the Levene test option.

group. See the different Ns for different groups shown in Figure ???. This violates one of our original assumptions, namely that the number of people in each group is the same. We haven't really discussed how to handle this situation.

#### 12.10.2 “Standard ANOVA” does not exist for unbalanced designs

Unbalanced designs lead us to the somewhat unsettling discovery that there isn't really any one thing that we might refer to as a standard ANOVA. In fact, it turns out that there are *three* fundamentally different ways<sup>\*9</sup> in which you might want to run an ANOVA in an unbalanced design. If you have a balanced design all three versions produce identical results, with the sums of squares,  $F$ -values, etc., all conforming to the formulas that I gave at the start of the chapter. However, when your design is unbalanced they don't give the same answers. Furthermore, they are not all equally appropriate to every situation. Some methods will be more appropriate to your situation than others. Given all this, it's important to understand what the different types of ANOVA are and how they differ from one another.

The first kind of ANOVA is conventionally referred to as **Type I sum of squares**. I'm sure you can guess what the other two are called. The “sum of squares” part of the name was introduced by the SAS statistical software package and has become standard nomenclature, but it's a bit misleading in some ways. I think the logic for referring to them as different types of sum of squares is that, when you look at the ANOVA tables that they produce, the key difference in the numbers is the SS values. The degrees of freedom don't change, the MS values are still defined as SS divided by df, etc. However, what the terminology gets wrong is that it hides the reason *why* the SS values are different from one another. To that end, it's a lot more helpful to think of the three different kinds of ANOVA as three different *hypothesis testing strategies*. These different strategies lead to different SS values, to be sure, but it's the strategy that is the important thing here, not the SS values themselves. Recall from Section ?? that any particular  $F$ -test is best thought of as a comparison between two linear models. So, when you're looking at an ANOVA table, it helps to remember that each of those  $F$ -tests corresponds to a *pair* of models that are being compared. Of course, this leads naturally to the question of *which* pair of models is being compared. This is

---

<sup>\*9</sup>Actually, this is a bit of a lie. ANOVAs can vary in other ways besides the ones I've discussed in this book. For instance, I've completely ignored the difference between fixed-effect models in which the levels of a factor are “fixed” by the experimenter or the world, and random-effect models in which the levels are random samples from a larger population of possible levels (this book only covers fixed-effect models). Don't make the mistake of thinking that this book, or any other one, will tell you “everything you need to know” about statistics, any more than a single book could possibly tell you everything you need to know about psychology, physics or philosophy. Life is too complicated for that to ever be true. This isn't a cause for despair, though. Most researchers get by with a basic working knowledge of ANOVA that doesn't go any further than this book does. I just want you to keep in mind that this book is only the beginning of a very long story, not the whole story.

the fundamental difference between ANOVA Types I, II and III: each one corresponds to a different way of choosing the model pairs for the tests.

#### 12.10.3 Type I sum of squares

The Type I method is sometimes referred to as the “sequential” sum of squares, because it involves a process of adding terms to the model one at a time. Consider the coffee data, for instance. Suppose we want to run the full  $3 \times 2$  factorial ANOVA, including interaction terms. The full model contains the outcome variable `babble`, the predictor variables `sugar` and `milk`, and the interaction term `sugar*milk`. This can be written as `babble ~ sugar + milk + sugar*milk`. The Type I strategy builds this model up sequentially, starting from the simplest possible model and gradually adding terms.

The simplest possible model for the data would be one in which neither milk nor sugar is assumed to have any effect on babbling. The only term that would be included in such a model is the intercept, written as `babble ~ 1`. This is our initial null hypothesis. The next simplest model for the data would be one in which only one of the two main effects is included. In the coffee data, there are two different possible choices here, because we could choose to add milk first or to add sugar first. The order actually turns out to matter, as we'll see later, but for now let's just make a choice arbitrarily and pick sugar. So, the second model in our sequence of models is `babble ~ sugar`, and it forms the alternative hypothesis for our first test. We now have our first hypothesis test:

Null model: `babble ~ 1`

Alternative model: `babble ~ sugar`

This comparison forms our hypothesis test of the main effect of `sugar`. The next step in our model building exercise is to add the other main effect term, so the next model in our sequence is `babble ~ sugar + milk`. The second hypothesis test is then formed by comparing the following pair of models:

Null model: `babble ~ sugar`

Alternative model: `babble ~ sugar + milk`

This comparison forms our hypothesis test of the main effect of `milk`. In one sense, this approach is very elegant: the alternative hypothesis from the first test forms the null hypothesis for the second one. It is in this sense that the Type I method is strictly sequential. Every test builds directly on the results of the last one. However, in another sense it's very inelegant, because there's a strong asymmetry between the two tests. The test of the main effect of `sugar` (the first test) completely ignores `milk`, whereas the test of the main effect of `milk` (the second test) does

take `sugar` into account. In any case, the fourth model in our sequence is now the full model, `babble ~ sugar + milk + sugar*milk`, and the corresponding hypothesis test is:

Null model: `babble ~ sugar + milk`

Alternative model: `babble ~ sugar + milk + sugar*milk`

Type III sum of squares is the default hypothesis testing method used by JASP ANOVA, so to run a Type I sum of squares analysis we have to select 'Type 1' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options. This gives us the ANOVA table shown in Figure ??.

The big problem with using Type I sum of squares is the fact that it really does depend on the order in which you enter the variables. Yet, in many situations the researcher has no reason to prefer one ordering over another. This is presumably the case for our milk and sugar problem. Should we add milk first or sugar first? It feels exactly as arbitrary as a data analysis question as it does as a coffee-making question. There may in fact be some people with firm opinions about ordering, but it's hard to imagine a principled answer to the question. Yet, look what happens when we change the ordering, as in Figure ??.

The  $p$ -values for both main effect terms have changed, and fairly dramatically. Among other things, the effect of `milk` has become significant (though one should avoid drawing any strong conclusions about this, as I've mentioned previously). Which of these two ANOVAs should one report? It's not immediately obvious.

When you look at the hypothesis tests that are used to define the "first" main effect and the "second" one, it's clear that they're qualitatively different from one another. In our initial example, we saw that the test for the main effect of `sugar` completely ignores `milk`, whereas the test of the main effect of `milk` does take `sugar` into account. As such, the Type I testing strategy really does treat the first main effect as if it had a kind of theoretical primacy over the second one. In my experience there is very rarely if ever any theoretically primacy of this kind that would justify treating any two main effects asymmetrically.

The consequence of all this is that Type I tests are very rarely of much interest, and so we should move on to discuss Type II tests and Type III tests.

#### 12.10.4 Type III sum of squares

Having just finished talking about Type I tests, you might think that the natural thing to do next would be to talk about Type II tests. However, I think it's actually a bit more natural to discuss Type III tests (which are simple and the default in JASP) before talking about Type II tests (which are trickier). The basic idea behind Type III tests is extremely simple. Regardless of which term you're trying to evaluate, run the  $F$ -test in which the alternative hypothesis corresponds to the full ANOVA

model as specified by the user, and the null model just deletes that one term that you're testing. For instance, in the coffee example, in which our full model was `babble ~ sugar + milk + sugar*milk`, the test for a main effect of `sugar` would correspond to a comparison between the following two models:

Null model: `babble ~ milk + sugar*milk`

Alternative model: `babble ~ sugar + milk + sugar*milk`

Similarly the main effect of `milk` is evaluated by testing the full model against a null model that removes the `milk` term, like so:

Null model: `babble ~ sugar + sugar*milk`

Alternative model: `babble ~ sugar + milk + sugar*milk`

Finally, the interaction term `sugar*milk` is evaluated in exactly the same way. Once again, we test the full model against a null model that removes the `sugar*milk` interaction term, like so:

Null model: `babble ~ sugar + milk`

Alternative model: `babble ~ sugar + milk + sugar*milk`

The basic idea generalises to higher order ANOVAs. For instance, suppose that we were trying to run an ANOVA with three factors, `A`, `B` and `C`, and we wanted to consider all possible main effects and all possible interactions, including the three way interaction `A*B*C`. The table below shows you what the Type III tests look like for this situation:

Term being tested is	Null model is <code>outcome ~ ...</code>	Alternative model is <code>outcome ~ ...</code>
<code>A</code>	<code>B + C + A*B + A*C + B*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>B</code>	<code>A + C + A*B + A*C + B*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>C</code>	<code>A + B + A*B + A*C + B*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>A*B</code>	<code>A + B + C + A*C + B*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>A*C</code>	<code>A + B + C + A*B + B*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>B*C</code>	<code>A + B + C + A*B + A*C + A*B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>
<code>A*B*C</code>	<code>A + B + C + A*B + A*C + B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>

As ugly as that table looks, it's pretty simple. In all cases, the alternative hypothesis corresponds to the full model which contains three main effect terms (e.g. `A`), three two-way interactions (e.g. `A*B`) and one three-way interaction (i.e., `A*B*C`). The null model always contains 6 of these 7 terms, and the missing one is the one whose significance we're trying to test.

At first pass, Type III tests seem like a nice idea. Firstly, we've removed the asymmetry that caused us to have problems when running Type I tests. And because we're now treating all terms

the same way, the results of the hypothesis tests do not depend on the order in which we specify them. This is definitely a good thing. However, there is a big problem when interpreting the results of the tests, especially for main effect terms. Consider the coffee data. Suppose it turns out that the main effect of `milk` is not significant according to the Type III tests. What this is telling us is that `babble ~ sugar + sugar*milk` is a better model for the data than the full model. But what does that even *mean*? If the interaction term `sugar*milk` was also non-significant, we'd be tempted to conclude that the data are telling us that the only thing that matters is `sugar`. But suppose we have a significant interaction term, but a non-significant main effect of `milk`. In this case, are we to assume that there really is an "effect of sugar", an "interaction between milk and sugar", but no "effect of milk"? That seems crazy. The right answer simply *must* be that it's meaningless<sup>\*10</sup> to talk about the main effect if the interaction is significant. In general, this seems to be what most statisticians advise us to do, and I think that's the right advice. But if it really is meaningless to talk about non-significant main effects in the presence of a significant interaction, then it's not at all obvious why Type III tests should allow the null hypothesis to rely on a model that includes the interaction but omits one of the main effects that make it up. When characterised in this fashion, the null hypotheses really don't make much sense at all.

Later on, we'll see that Type III tests can be redeemed in some contexts, but first let's take a look at the ANOVA results table using Type III sum of squares, see Figure ??.

But be aware, one of the perverse features of the Type III testing strategy is that typically the results turn out to depend on the *contrasts* that you use to encode your factors (see Section ?? if you've forgotten what the different types of contrasts are).<sup>\*11</sup>

Okay, so if the *p*-values that typically come out of Type III analyses are so sensitive to the choice of contrasts, does that mean that Type III tests are essentially arbitrary and not to be trusted? To some extent that's true, and when we turn to a discussion of Type II tests we'll see that Type II analyses avoid this arbitrariness entirely, but I think that's too strong a conclusion. Firstly, it's important to recognise that some choices of contrasts will always produce the same answers (ah, so this is what is happening in JASP). Of particular importance is the fact that if the columns of our contrast matrix are all constrained to sum to zero, then the Type III analysis will always give the same answers.

---

<sup>\*10</sup>Or, at the very least, rarely of interest.

<sup>\*11</sup>However, in JASP the results for Type III sum of squares ANOVA are the same regardless of the contrast selected, so JASP is obviously doing something different!

### 12.10.5 Type II sum of squares

Okay, so we've seen Type I and III tests now, and both are pretty straightforward. Type I tests are performed by gradually adding terms one at a time, whereas Type III tests are performed by taking the full model and looking to see what happens when you remove each term. However, both can have some limitations. Type I tests are dependent on the order in which you enter the terms, and Type III tests are dependent on how you code up your contrasts. Type II tests are a little harder to describe, but they avoid both of these problems, and as a result they are a little easier to interpret.

Type II tests are broadly similar to Type III tests. Start with a "full" model, and test a particular term by deleting it from that model. However, Type II tests are based on the **marginality principle** which states that you should not omit a lower order term from your model if there are any higher order ones that depend on it. So, for instance, if your model contains the two-way interaction  $A*B$  (a 2nd order term), then it really ought to contain the main effects  $A$  and  $B$  (1st order terms). Similarly, if it contains a three-way interaction term  $A*B*C$ , then the model must also include the main effects  $A$ ,  $B$  and  $C$  as well as the simpler interactions  $A*B$ ,  $A*C$  and  $B*C$ . Type III tests routinely violate the marginality principle. For instance, consider the test of the main effect of  $A$  in the context of a three-way ANOVA that includes all possible interaction terms. According to Type III tests, our null and alternative models are:

Null model:  $\text{outcome} \sim B + C + A*B + A*C + B*C + A*B*C$

Alternative model:  $\text{outcome} \sim A + B + C + A*B + A*C + B*C + A*B*C$

Notice that the null hypothesis omits  $A$ , but includes  $A*B$ ,  $A*C$  and  $A*B*C$  as part of the model. This, according to the Type II tests, is not a good choice of null hypothesis. What we should do instead, if we want to test the null hypothesis that  $A$  is not relevant to our  $\text{outcome}$ , is to specify the null hypothesis that is the most complicated model that does not rely on  $A$  in any form, even as an interaction. The alternative hypothesis corresponds to this null model plus a main effect term of  $A$ . This is a lot closer to what most people would intuitively think of as a "main effect of  $A$ ", and it yields the following as our Type II test of the main effect of  $A$ :<sup>12</sup>

Null model:  $\text{outcome} \sim B + C + B*C$

Alternative model:  $\text{outcome} \sim A + B + C + B*C$

Anyway, just to give you a sense of how the Type II tests play out, here's the full table of tests

---

<sup>12</sup>Note, of course, that this does depend on the model that the user specified. If the original ANOVA model doesn't contain an interaction term for  $B*C$ , then obviously it won't appear in either the null or the alternative. But that's true for Types I, II and III. They never include any terms that you *didn't* include, but they make different choices about how to construct tests for the ones that you did include.

that would be applied in a three-way factorial ANOVA:

Term being tested is	Null model is <code>outcome ~ ...</code>	Alternative model is <code>outcome ~ ...</code>
A	<code>B + C + B*C</code>	<code>A + B + C + B*C</code>
B	<code>A + C + A*C</code>	<code>A + B + C + A*C</code>
C	<code>A + B + A*B</code>	<code>A + B + C + A*B</code>
A*B	<code>A + B + C + A*C + B*C</code>	<code>A + B + C + A*B + A*C + B*C</code>
A*C	<code>A + B + C + A*B + B*C</code>	<code>A + B + C + A*B + A*C + B*C</code>
B*C	<code>A + B + C + A*B + A*C</code>	<code>A + B + C + A*B + A*C + B*C</code>
A*B*C	<code>A + B + C + A*B + A*C + B*C</code>	<code>A + B + C + A*B + A*C + B*C + A*B*C</code>

In the context of the two way ANOVA that we've been using in the coffee data, the hypothesis tests are even simpler. The main effect of `sugar` corresponds to an *F*-test comparing these two models:

$$\begin{aligned} \text{Null model: } & \text{babble} \sim \text{milk} \\ \text{Alternative model: } & \text{babble} \sim \text{sugar} + \text{milk} \end{aligned}$$

The test for the main effect of `milk` is

$$\begin{aligned} \text{Null model: } & \text{babble} \sim \text{sugar} \\ \text{Alternative model: } & \text{babble} \sim \text{sugar} + \text{milk} \end{aligned}$$

Finally, the test for the interaction `sugar*milk` is:

$$\begin{aligned} \text{Null model: } & \text{babble} \sim \text{sugar} + \text{milk} \\ \text{Alternative model: } & \text{babble} \sim \text{sugar} + \text{milk} + \text{sugar*milk} \end{aligned}$$

Running the tests are again straightforward. Just select 'Type 2' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options. This gives us the ANOVA table shown in Figure ??.

Type II tests have some clear advantages over Type I and Type III tests. They don't depend on the order in which you specify factors (unlike Type I), and they don't depend on the contrasts that you use to specify your factors (unlike Type III). And although opinions may differ on this last point, and it will definitely depend on what you're trying to do with your data, I do think that the hypothesis tests that they specify are more likely to correspond to something that you actually care about. As a consequence, I find that it's usually easier to interpret the results of a Type II test than the results of a Type I or Type III test. For this reason my tentative advice is that, if you can't think of any obvious model comparisons that directly map onto your research questions but you still want to run an ANOVA in an unbalanced design, Type II tests are probably a better

choice than Type I or Type III.\*<sup>13</sup>

#### 12.10.6 Effect sizes (and non-additive sums of squares)

JASP also provides the effect sizes  $\eta^2$  and partial  $\eta^2$  when you select these options. However, when you've got an unbalanced design there's a bit of extra complexity involved.

If you remember back to our very early discussions of ANOVA, one of the key ideas behind the sums of squares calculations is that if we add up all the SS terms associated with the effects in the model, and add that to the residual SS, they're supposed to add up to the total sum of squares. And, on top of that, the whole idea behind  $\eta^2$  is that, because you're dividing one of the SS terms by the total SS value, an  $\eta^2$  value can be interpreted as the proportion of variance accounted for by a particular term. But this is not so straightforward in unbalanced designs because some of the variance goes "missing".

This seems a bit odd at first, but here's why. When you have unbalanced designs your factors become correlated with one another, and it becomes difficult to tell the difference between the effect of Factor A and the effect of Factor B. In the extreme case, suppose that we'd run a  $2 \times 2$  design in which the number of participants in each group had been as follows:

	sugar	no sugar
milk	100	0
no milk	0	100

Here we have a spectacularly unbalanced design: 100 people have milk and sugar, 100 people have no milk and no sugar, and that's all. There are 0 people with milk and no sugar, and 0 people with sugar but no milk. Now suppose that, when we collected the data, it turned out there is a large (and statistically significant) difference between the "milk and sugar" group and the "no-milk and no-sugar" group. Is this a main effect of sugar? A main effect of milk? Or an interaction? It's impossible to tell, because the presence of sugar has a perfect association with the presence

---

\*<sup>13</sup>I find it amusing to note that the default in R is Type I and the default in SPSS, JASP, and jamovi is Type III. Neither of these appeals to me all that much. Relatedly, I find it depressing that almost nobody in the psychological literature ever bothers to report which Type of tests they ran, much less the order of variables (for Type I) or the contrasts used (for Type III). Often they don't report what software they used either. The only way I can ever make any sense of what people typically report is to try to guess from auxiliary cues which software they were using, and to assume that they never changed the default settings. Please don't do this! Now that you know about these issues make sure you indicate what software you used, and if you're reporting ANOVA results for unbalanced data, then specify what Type of tests you ran, specify order information if you've done Type I tests and specify contrasts if you've done Type III tests. Or, even better, do hypotheses tests that correspond to things you really care about and then report those!

of milk. Now suppose the design had been a little more balanced:

	sugar	no sugar
milk	100	5
no milk	5	100

This time around, it's technically possible to distinguish between the effect of milk and the effect of sugar, because we have a few people that have one but not the other. However, it will still be pretty difficult to do so, because the association between sugar and milk is still extremely strong, and there are so few observations in two of the groups. Again, we're very likely to be in the situation where we *know* that the predictor variables (milk and sugar) are related to the outcome (babbling), but we don't know if the *nature* of that relationship is a main effect of one or the other predictor, or the interaction.

This uncertainty is the reason for the missing variance. The "missing" variance corresponds to variation in the outcome variable that is clearly attributable to the predictors, but we don't know which of the effects in the model is responsible. When you calculate Type I sum of squares, no variance ever goes missing. The sequential nature of Type I sum of squares means that the ANOVA automatically attributes this variance to whichever effects are entered first. However, the Type II and Type III tests are more conservative. Variance that cannot be clearly attributed to a specific effect doesn't get attributed to any of them, and it goes missing.

## 12.11

---

### Summary

- Factorial ANOVA with balanced designs, without interactions (Section ??) and with interactions included (Section ??)
- Effect size, estimated means, and confidence intervals in a factorial ANOVA (Section ??)
- Checking assumptions in ANOVA (Section ??)
- Analysis of Covariance (ANCOVA) (Section ??)
- Understanding the linear model underlying ANOVA, including different contrasts (Section ?? and ??)
- Post hoc testing using Tukey's HSD (Section ??) and a brief commentary on planned comparisons (Section ??)
- Factorial ANOVA with unbalanced designs (Section ??)

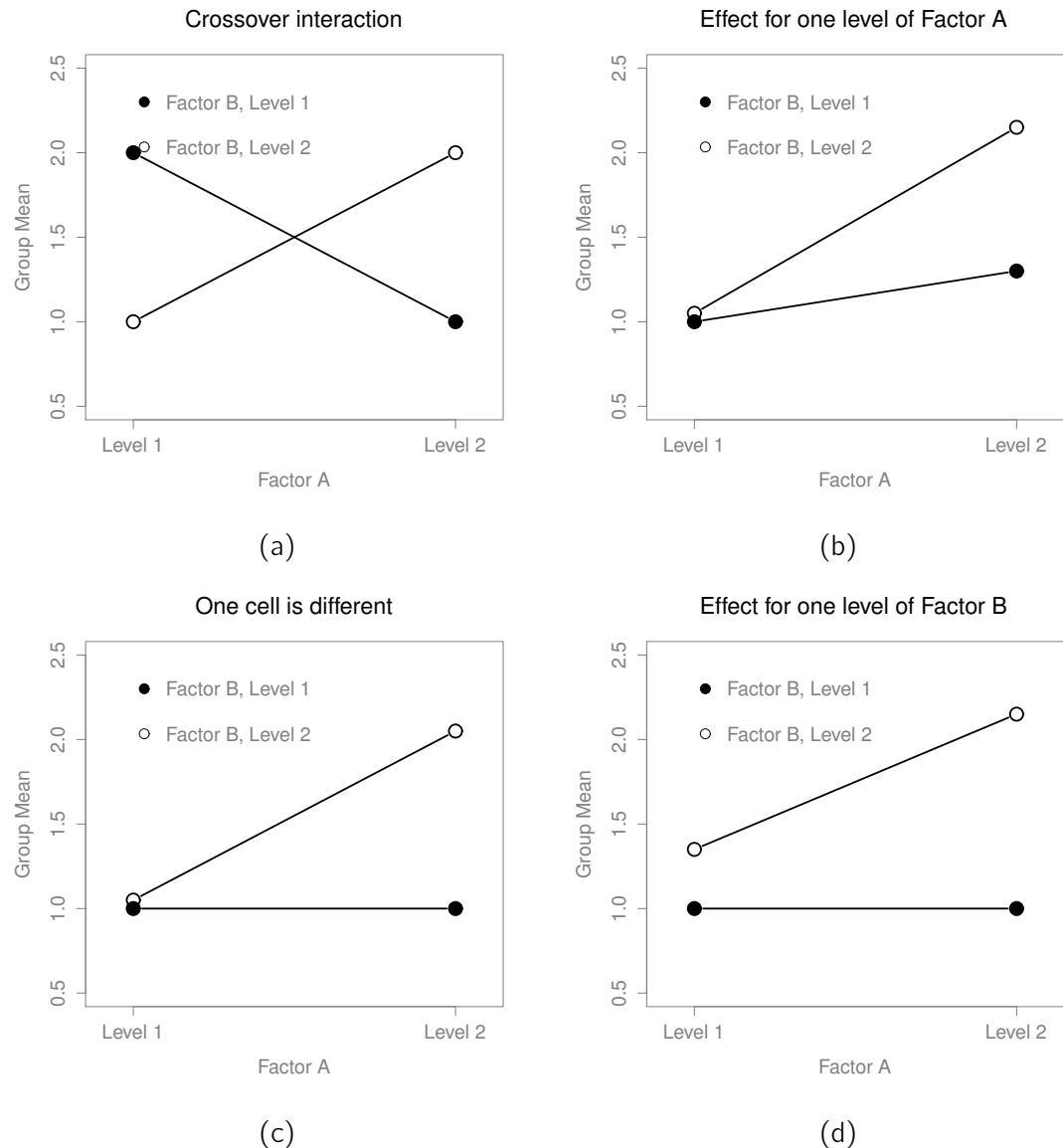


Figure12.6  $2 \times 2$  要因の分散分析における様々な交互作用

## Descriptives ▼

### Descriptives Plot

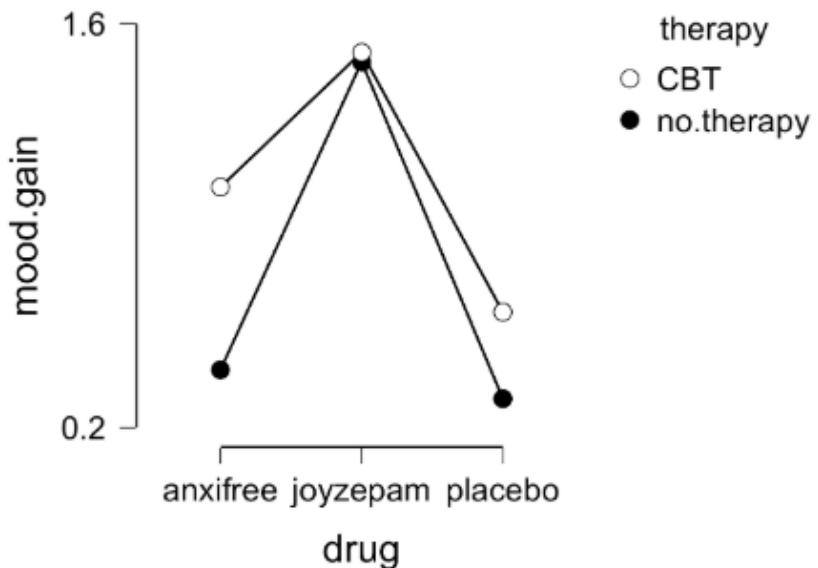


Figure12.7 臨床試験データに対して分散分析の「Descriptives Plot」オプションを使用した際の JASP の出力

### ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	31.714	< .001
therapy	0.467	1.000	0.467	8.582	0.013
drug * therapy	0.271	2.000	0.136	2.490	0.125
Residual	0.653	12.000	0.054		

Note. Type III Sum of Squares

Figure12.8 交互作用項 drug\*therapy を含む、完全な多元配置モデルの出力

## Marginal Means

Marginal Means – drug \* therapy

drug	therapy	Marginal Mean	SE	95% CI	
				Lower	Upper
anxitfree	CBT	1.033	0.135	0.740	1.327
	no.therapy	0.400	0.135	0.106	0.694
joyzepam	CBT	1.500	0.135	1.206	1.794
	no.therapy	1.467	0.135	1.173	1.760
placebo	CBT	0.600	0.135	0.306	0.894
	no.therapy	0.300	0.135	0.006	0.594

Figure12.9 飽和モデルの周辺平均を示す JASP のスクリーンショット, i.e. clinicaltrial データセットの交互作用コンポーネントを含む

## Assumption Checks ▼

### Test for Equality of Variances (Levene's)

F	df1	df2	p
0.206	5.000	12.000	0.954

## Q-Q Plot ▼

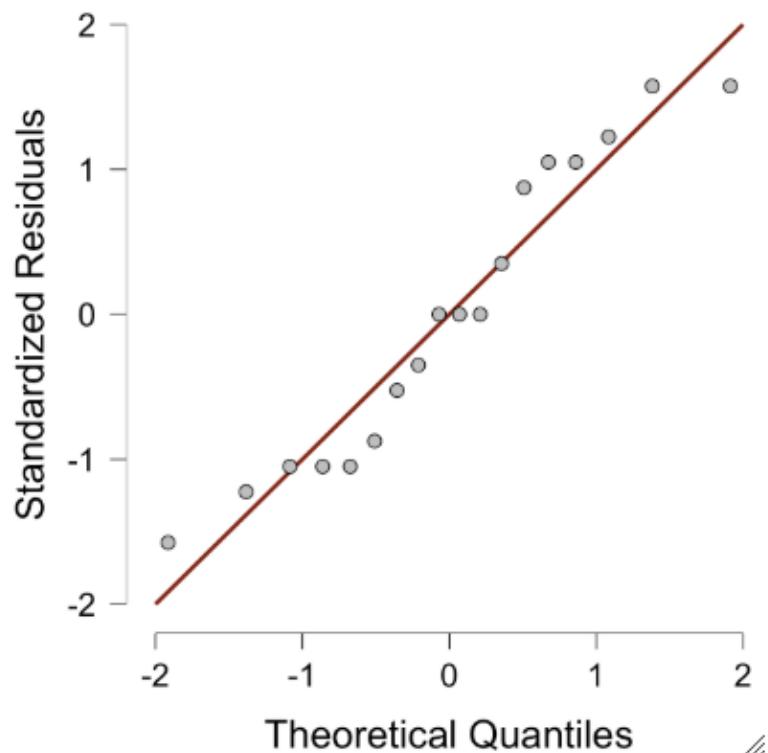


Figure12.10 分散分析モデルの仮定の確認

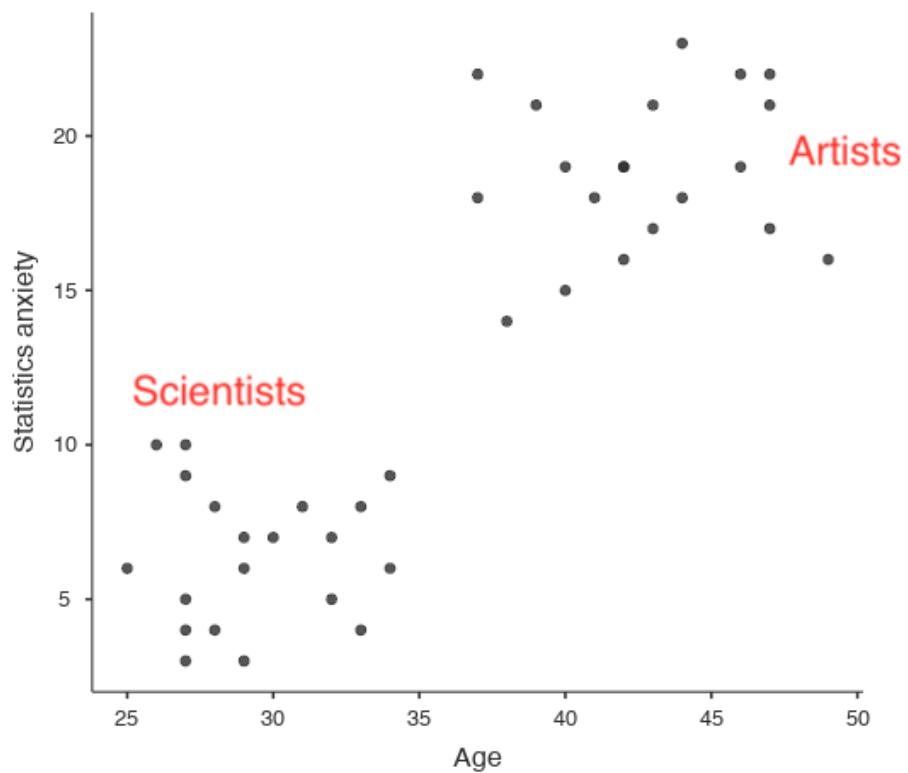


Figure12.11 Plot of Statistics anxiety against age for two distinct groups

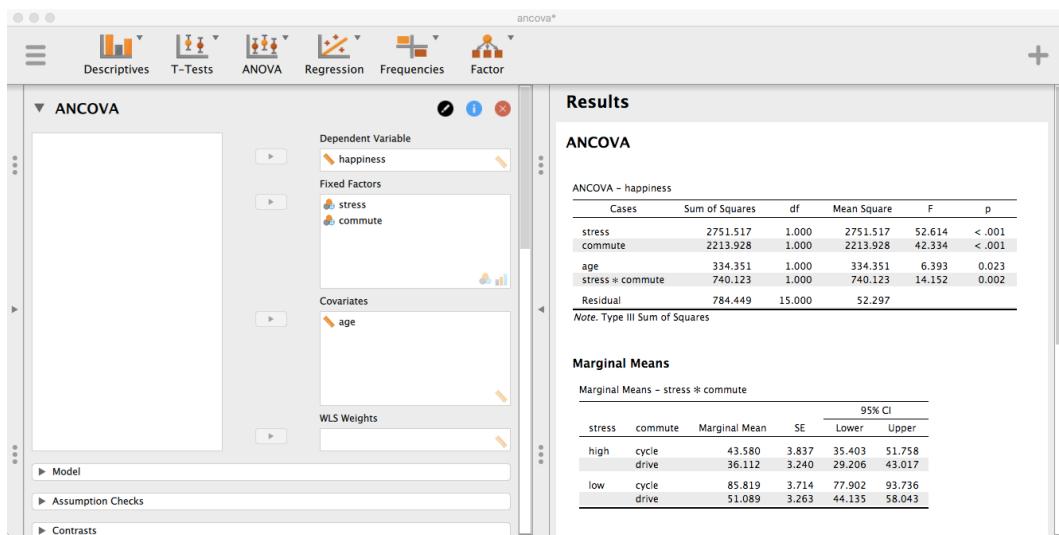


Figure12.12 The JASP ANCOVA analysis window

## Descriptives

### Descriptives Plot

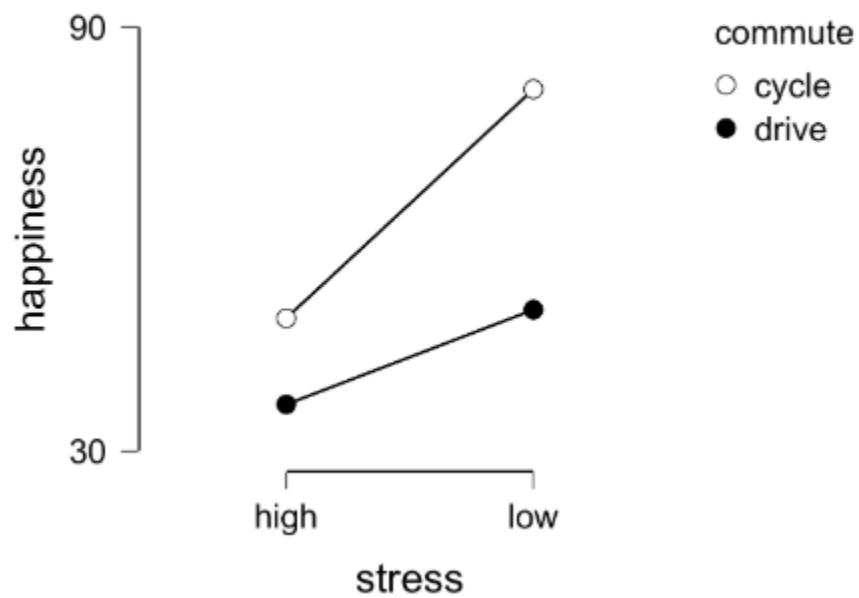


Figure12.13 Plot of mean happiness level as a function of stress and commuting method

#### ANOVA - grade

Cases	Sum of Squares	df	Mean Square	F	p
attend	648.000	1.000	648.000	21.600	0.006
reading	1568.000	1.000	1568.000	52.267	< .001
Residual	150.000	5.000	30.000		

Note. Type III Sum of Squares

Figure12.14 ANOVA of the rtfm.csv data set in JASP, without the interaction term

---

#### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	43.500	3.354		12.969	< .001
	attend	18.000	3.873	0.523	4.648	0.006
	reading	28.000	3.873	0.814	7.230	< .001

Figure12.15 Regression analysis of the rtfm.csv data set in JASP, without the interaction term

---

#### ANOVA - mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	26.149	< .001
therapy	0.467	1.000	0.467	7.076	0.019
Residual	0.924	14.000	0.066		

Note. Type III Sum of Squares

Figure12.16 JASP ANOVA results, without interaction component

---

Coefficients						
Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< .001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure12.17 JASP regression results, with contrast variables druganxifree and drugjoyzepam

## Linear Regression ▾

### Model Summary ▾

Model	R	R <sup>2</sup>	Adjusted R <sup>2</sup>	RMSE
0	0.844	0.713	0.674	0.305
1	0.900	0.809	0.768	0.257

Note. Null model includes druganxifree, drugjoyzepam

### ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	3.453	2	1.727	18.611	< .001
	Residual	1.392	15	0.093		
	Total	4.845	17			
1	Regression	3.921	3	1.307	19.791	< .001
	Residual	0.924	14	0.066		
	Total	4.845	17			

Note. Null model includes druganxifree, drugjoyzepam

### Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
0	(Intercept)	0.450	0.124		3.619	0.003
	druganxifree	0.267	0.176	0.242	1.516	0.150
	drugjoyzepam	1.033	0.176	0.939	5.876	< .001
1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< .001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure12.18 Model comparison in JASP regression, null model 0 vs. contrasts model 1

## Post Hoc Tests ▼

### Post Hoc Comparisons – drug ▼

		Mean Difference	SE	t	Ptukey
placebo	joyzepam	-1.033	0.148	-6.965	< .001
	anxitfree	-0.267	0.148	-1.797	0.206
joyzepam	anxitfree	0.767	0.148	5.168	< .001

### Post Hoc Comparisons – therapy

		Mean Difference	SE	t	Ptukey
CBT	no.therapy	0.322	0.121	2.660	0.019

Figure12.19 Tukey HSD post hoc test in JASP

## Post Hoc Tests ▼

### Post Hoc Comparisons – drug \* therapy ▼

		Mean Difference	SE	t	Ptukey
placebo,CBT	joyzepam,CBT	-0.900	0.191	-4.724	0.005
	anxitfree,CBT	-0.433	0.191	-2.275	0.275
	placebo,no.therapy	0.300	0.191	1.575	0.628
	joyzepam,no.therapy	-0.867	0.191	-4.549	0.007
	anxitfree,no.therapy	0.200	0.191	1.050	0.892
joyzepam,CBT	anxitfree,CBT	0.467	0.191	2.449	0.214
	placebo,no.therapy	1.200	0.191	6.299	< .001
	joyzepam,no.therapy	0.033	0.191	0.175	1.000
	anxitfree,no.therapy	1.100	0.191	5.774	< .001
anxitfree,CBT	placebo,no.therapy	0.733	0.191	3.849	0.022
	joyzepam,no.therapy	-0.433	0.191	-2.275	0.275
	anxitfree,no.therapy	0.633	0.191	3.324	0.053
placebo,no.therapy	joyzepam,no.therapy	-1.167	0.191	-6.124	< .001
	anxitfree,no.therapy	-0.100	0.191	-0.525	0.994
joyzepam,no.therapy	anxitfree,no.therapy	1.067	0.191	5.599	0.001

Figure12.20 Tukey HSD post hoc test in JASP factorial ANOVA with an interaction term

## Results

### Descriptive Statistics

Descriptive Statistics

babble		
	no	yes
Valid	10	8
Missing	0	0
Mean	5.320	4.750
Std. Deviation	0.796	0.962
Minimum	3.900	3.500
Maximum	6.600	5.900

### Descriptive Statistics

Descriptive Statistics

babble			
	fake	none	real
Valid	6	5	7
Missing	0	0	0
Mean	5.033	4.440	5.543
Std. Deviation	0.814	1.038	0.637
Minimum	3.900	3.500	4.600
Maximum	5.900	5.800	6.600

Figure12.21 Descriptives for the coffee.csv data set, separately split by milk and sugar, respectively.

## ANOVA

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
sugar	3.558	2.000	1.779	6.749	0.011
milk	0.956	1.000	0.956	3.628	0.081
sugar * milk	5.944	2.000	2.972	11.277	0.002
Residual	3.162	12.000	0.264		

Note. Type I Sum of Squares

Figure12.22 ANOVA results table using Type I sum of squares in JASP

.....

## ANOVA ▼

ANOVA - babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.444	1.000	1.444	5.479	0.037
sugar	3.070	2.000	1.535	5.824	0.017
milk * sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type I Sum of Squares

Figure12.23 ANOVA results table using Type I sum of squares in JASP, but with factors entered in a different order (milk first)

.....

## ANOVA

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.004	1.000	1.004	3.810	0.075
sugar	2.132	2.000	1.066	4.045	0.045
milk * sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type III Sum of Squares

Figure12.24 ANOVA results table using Type III sum of squares in JASP

## ANOVA

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	0.956	1.000	0.956	3.628	0.081
sugar	3.070	2.000	1.535	5.824	0.017
milk * sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type II Sum of Squares

Figure12.25 ANOVA results table using Type II sum of squares in JASP

Part V.

## **Endings, alternatives and prospects**



## 13. ベイズ統計学

---

現実の問題に関する我々の推論には、想像可能なあらゆる確信の度合いが存在する。最も確からしいものから、道徳的な証拠という最も不確かなものまで。ゆえに、賢人は信念を証拠に釣り合わせる。

– David Hume<sup>\*1</sup>

本書で私がこれまでに紹介してきたアイディアは、頻度主義の立場から見た推測統計学でした。このようなやり方を採用しているのは私だけではありません。実際、心理学の学部生向けに書かれたほぼ全ての教科書では、推測統計学の理論として頻度主義の統計学者の意見が紹介されています。これはひとつの正しいやり方です。私は実用上の理由からこの教育方法を用いてきました。頻度主義の考え方は 20 世紀の大半に渡って統計学の学術領域を席巻しました。この席巻は応用科学者の間で殊更顕著でした。頻度主義の手法は昔も今も心理学者たちの間で使われています。頻度主義の手法は科学論文の至るところで用いられているため、統計学を学ぶ全ての学生は頻度主義の手法を理解しなければならず、さもなくば科学論文の内容を理解できません。しかし残念なことに、少なくとも私の意見では、心理学における現行の統計手法は誤用を招きやすいものであり、頻度主義への依存には批判されるべき点もあります。

この章では、私がそのように考える理由を説明した上でベイズ統計学の紹介をします。ベイズ統計学は伝統的なアプローチよりも一般に優れたアプローチであると私は考えています。

この章は 2 つのパートに分かれています。

セクション ??から ??ではベイズ統計学の全てについて説明します。ベイジアンアプローチが有用である理由だけでなく、基本となる数学の公式もカバーしています。その後で、 $t$  検定のベイズ版を実行する方法を簡単に説明します (セクション ??)。

---

<sup>\*1</sup>[http://en.wikiquote.org/wiki/David\\_Hume](http://en.wikiquote.org/wiki/David_Hume).

## 13.1

---

### 合理的エージェントによる確率的推論

ベイジアンの立場では、統計的推測は信念の修正に他なりません。まずは、この世界についての仮説候補  $h$  の集合について考えてみましょう。どの仮説が真であるかは分かりませんが、どの仮説が正しそうで、どの仮説が正しくなさそうかについての信念を私は多少なり持っています。データ  $d$  を観測したとき、私はこれらの信念を修正しなければなりません。もしデータがある仮説と整合的であれば、その仮説についての私の信念は強められます。もしデータがその仮説と整合的でなければ、その仮説についての私の信念は弱められます。これが全てなのです！このセクションの最後にはベイジアンの推論がどのように作用するかを正確に記述するつもりですが、まずは鍵となるアイディアを紹介するために簡単な例を示したいと思います。以下の推論問題を考えてみましょう。

私は傘を持ち歩いています。あなたは雨が降ると思いますか？

この問題では、私はあなたに一片のデータ ( $d = \text{私は傘を持ち歩いている}$ ) を提示し、雨が降るか否かについてのあなたの信念あるいは仮説を私に教えてくれることを求めていきます。あなたが選ぶことのできる 2 つの選択肢  $h$  は、今日雨が降るか、降らないかです。どうやってこの問題を解決しますか？

#### 13.1.1 事前確率: 以前あなたが信じていたもの

The first thing you need to do is ignore what I told you about the umbrella, and write down your pre-existing beliefs about rain. This is important. If you want to be honest about how your beliefs have been revised in the light of new evidence (data) then you *must* say something about what you believed before those data appeared! So, what might you believe about whether it will rain today? You probably know that I live in Australia and that much of Australia is hot and dry. The city of Adelaide where I live has a Mediterranean climate, very similar to southern California, southern Europe or northern Africa. I'm writing this in January and so you can assume it's the middle of summer. In fact, you might have decided to take a quick look on Wikipedia<sup>\*2</sup> and discovered that Adelaide gets an average of 4.4 days of rain across the 31 days of January. Without knowing anything else, you might conclude that the probability of January rain in Adelaide is about 15%, and the probability of a dry day is 85%. If this is really what you believe about Adelaide rainfall (and now that I've told it to you I'm betting that this really *is* what you believe) then what I have

---

\*2 [http://en.wikipedia.org/wiki/Climate\\_of\\_Adelaide](http://en.wikipedia.org/wiki/Climate_of_Adelaide)

written here is your **prior distribution**, written  $P(h)$ :

Hypothesis	Degree of Belief
Rainy day	0.15
Dry day	0.85

### 13.1.2 尤度: データに関する理論

To solve the reasoning problem you need a theory about my behaviour. When does Dan carry an umbrella? You might guess that I'm not a complete idiot,<sup>\*3</sup> and I try to carry umbrellas only on rainy days. On the other hand, you also know that I have young kids, and you wouldn't be all that surprised to know that I'm pretty forgetful about this sort of thing. Let's suppose that on rainy days I remember my umbrella about 30% of the time (I really am awful at this). But let's say that on dry days I'm only about 5% likely to be carrying an umbrella. So you might write out a little table like this:

Hypothesis	Data	
	Umbrella	No umbrella
Rainy day	0.30	0.70
Dry day	0.05	0.95

It's important to remember that each cell in this table describes your beliefs about what data  $d$  will be observed, *given* the truth of a particular hypothesis  $h$ . This "conditional probability" is written  $P(d|h)$ , which you can read as "the probability of  $d$  given  $h$ ". In Bayesian statistics, this is referred to as the **likelihood** of the data  $d$  given the hypothesis  $h$ .<sup>\*4</sup>

### 13.1.3 データと仮説の同時分布

At this point all the elements are in place. Having written down the priors and the likelihood,

<sup>\*3</sup>It's a leap of faith, I know, but let's run with it okay?

<sup>\*4</sup>Um. I hate to bring this up, but some statisticians would object to me using the word "likelihood" here. The problem is that the word "likelihood" has a very specific meaning in frequentist statistics, and it's not quite the same as what it means in Bayesian statistics. As far as I can tell Bayesians didn't originally have any agreed upon name for the likelihood, and so it became common practice for people to use the frequentist terminology. This wouldn't have been a problem except for the fact that the way that Bayesians use the word turns out to be quite different to the way frequentists do. This isn't the place for yet another lengthy history lesson but, to put it crudely, when a Bayesian says "a likelihood function" they're usually referring one of the *rows* of the table. When a frequentist says the same thing, they're referring to the same table, but to them "a likelihood function" almost always refers to one of the *columns*. This distinction matters in some contexts, but it's not important for our purposes.

you have all the information you need to do Bayesian reasoning. The question now becomes *how* do we use this information? As it turns out, there's a very simple equation that we can use here, but it's important that you understand why we use it so I'm going to try to build it up from more basic ideas.

Let's start out with one of the rules of probability theory. I listed it way back in Table ??, but I didn't make a big deal out of it at the time and you probably ignored it. The rule in question is the one that talks about the probability that *two* things are true. In our example, you might want to calculate the probability that today is rainy (i.e., hypothesis  $h$  is true) *and* I'm carrying an umbrella (i.e., data  $d$  is observed). The **joint probability** of the hypothesis and the data is written  $P(d, h)$ , and you can calculate it by multiplying the prior  $P(h)$  by the likelihood  $P(d|h)$ . Mathematically, we say that

$$P(d, h) = P(d|h)P(h)$$

So, what is the probability that today is a rainy day *and* I remember to carry an umbrella? As we discussed earlier, the prior tells us that the probability of a rainy day is 15%, and the likelihood tells us that the probability of me remembering my umbrella on a rainy day is 30%. So the probability that both of these things are true is calculated by multiplying the two

$$\begin{aligned} P(\text{rainy, umbrella}) &= P(\text{umbrella|rainy}) \times P(\text{rainy}) \\ &= 0.30 \times 0.15 \\ &= 0.045 \end{aligned}$$

In other words, before being told anything about what actually happened, you think that there is a 4.5% probability that today will be a rainy day and that I will remember an umbrella. However, there are of course *four* possible things that could happen, right? So let's repeat the exercise for all four. If we do that, we end up with the following table:

	Umbrella	No-umbrella
Rainy	0.045	0.105
Dry	0.0425	0.8075

This table captures all the information about which of the four possibilities are likely. To really get the full picture, though, it helps to add the row totals and column totals. That gives us this table:

	Umbrella	No-umbrella	Total
Rainy	0.0450	0.1050	0.15
Dry	0.0425	0.8075	0.85
Total	0.0875	0.9125	1

This is a very useful table, so it's worth taking a moment to think about what all these numbers are

telling us. First, notice that the row sums aren't telling us anything new at all. For example, the first row tells us that if we ignore all this umbrella business, the chance that today will be a rainy day is 15%. That's not surprising, of course, as that's our prior.<sup>\*5</sup> The important thing isn't the number itself. Rather, the important thing is that it gives us some confidence that our calculations are sensible! Now take a look at the column sums and notice that they tell us something that we haven't explicitly stated yet. In the same way that the row sums tell us the probability of rain, the column sums tell us the probability of me carrying an umbrella. Specifically, the first column tells us that on average (i.e., ignoring whether it's a rainy day or not) the probability of me carrying an umbrella is 8.75%. Finally, notice that when we sum across all four logically-possible events, everything adds up to 1. In other words, what we have written down is a proper probability distribution defined over all possible combinations of data and hypothesis.

Now, because this table is so useful, I want to make sure you understand what all the elements correspond to and how they written:

	Umbrella	No-umbrella	
Rainy	$P(\text{Umbrella, Rainy})$	$P(\text{No-umbrella, Rainy})$	$P(\text{Rainy})$
Dry	$P(\text{Umbrella, Dry})$	$P(\text{No-umbrella, Dry})$	$P(\text{Dry})$
$P(\text{Umbrella})$		$P(\text{No-umbrella})$	

Finally, let's use "proper" statistical notation. In the rainy day problem, the data corresponds to the observation that I do or do not have an umbrella. So we'll let  $d_1$  refer to the possibility that you observe me carrying an umbrella, and  $d_2$  refers to you observing me not carrying one. Similarly,  $h_1$  is your hypothesis that today is rainy, and  $h_2$  is the hypothesis that it is not. Using this notation, the table looks like this:

	$d_1$	$d_2$	
$h_1$	$P(h_1, d_1)$	$P(h_1, d_2)$	$P(h_1)$
$h_2$	$P(h_2, d_1)$	$P(h_2, d_2)$	$P(h_2)$
$P(d_1)$		$P(d_2)$	

---

<sup>\*5</sup>Just to be clear, "prior" information is pre-existing knowledge or beliefs, before we collect or use any data to improve that information.

#### 13.1.4 ベイズの公式を使って信念を更新する

The table we laid out in the last section is a very powerful tool for solving the rainy day problem, because it considers all four logical possibilities and states exactly how confident you are in each of them before being given any data. It's now time to consider what happens to our beliefs when we are actually given the data. In the rainy day problem, you are told that I really *am* carrying an umbrella. This is something of a surprising event. According to our table, the probability of me carrying an umbrella is only 8.75%. But that makes sense, right? A guy carrying an umbrella on a summer day in a hot dry city is pretty unusual, and so you really weren't expecting that. Nevertheless, the data tells you that it is true. No matter how unlikely you thought it was, you must now adjust your beliefs to accommodate the fact that you now *know* that I have an umbrella.<sup>\*6</sup> To reflect this new knowledge, our *revised* table must have the following numbers:

	Umbrella	No-umbrella
Rainy		0
Dry		0
Total	1	0

In other words, the facts have eliminated any possibility of “no umbrella”, so we have to put zeros into any cell in the table that implies that I’m not carrying an umbrella. Also, you know for a fact that I am carrying an umbrella, so the column sum on the left must be 1 to correctly describe the fact that  $P(\text{umbrella}) = 1$ .

What two numbers should we put in the empty cells? Again, let’s not worry about the maths, and instead think about our intuitions. When we wrote out our table the first time, it turned out that those two cells had almost identical numbers, right? We worked out that the joint probability of “rain and umbrella” was 4.5%, and the joint probability of “dry and umbrella” was 4.25%. In other words, before I told you that I am in fact carrying an umbrella, you’d have said that these two events were almost identical in probability, yes? But notice that *both* of these possibilities are consistent with the fact that I actually am carrying an umbrella. From the perspective of these two possibilities, very little has changed. I hope you’d agree that it’s *still* true that these two possibilities are equally plausible. So what we expect to see in our final table is some numbers that preserve the fact that “rain and umbrella” is *slightly* more plausible than “dry and umbrella”, while still ensuring that numbers in the table add up. Something like this, perhaps?

---

<sup>\*6</sup>If we were being a bit more sophisticated, we could extend the example to accommodate the possibility that I’m lying about the umbrella. But let’s keep things simple, shall we?

	Umbrella	No-umbrella
Rainy	0.514	0
Dry	0.486	0
Total	1	0

What this table is telling you is that, after being told that I'm carrying an umbrella, you believe that there's a 51.4% chance that today will be a rainy day, and a 48.6% chance that it won't. That's the answer to our problem! The **posterior probability** of rain  $P(h|d)$  given that I am carrying an umbrella is 51.4%

How did I calculate these numbers? You can probably guess. To work out that there was a 0.514 probability of "rain", all I did was take the 0.045 probability of "rain and umbrella" and divide it by the 0.0875 chance of "umbrella". This produces a table that satisfies our need to have everything sum to 1, and our need not to interfere with the relative plausibility of the two events that are actually consistent with the data. To say the same thing using fancy statistical jargon, what I've done here is divide the joint probability of the hypothesis and the data  $P(d, h)$  by the **marginal probability** of the data  $P(d)$ , and this is what gives us the posterior probability of the hypothesis *given* the data that have been observed. To write this as an equation <sup>\*7</sup>

$$P(h|d) = \frac{P(d, h)}{P(d)}$$

However, remember what I said at the start of the last section, namely that the joint probability  $P(d, h)$  is calculated by multiplying the prior  $P(h)$  by the likelihood  $P(d|h)$ . In real life, the things we actually know how to write down are the priors and the likelihood, so let's substitute those back into the equation. This gives us the following formula for the posterior probability

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

And this formula, folks, is known as **Bayes' rule**. It describes how a learner starts out with prior beliefs about the plausibility of different hypotheses, and tells you how those beliefs should be revised in the face of data. In the Bayesian paradigm, all statistical inference flows from this one simple rule.

## 13.2

---

<sup>\*7</sup>You might notice that this equation is actually a restatement of the same basic rule I listed at the start of the last section. If you multiply both sides of the equation by  $P(d)$ , then you get  $P(d)P(h|d) = P(d, h)$ , which is the rule for how joint probabilities are calculated. So I'm not actually introducing any "new" rules here, I'm just using the same rule in a different way.

## Bayesian hypothesis tests

In Chapter ?? I described the orthodox approach to hypothesis testing. It took an entire chapter to describe, because null hypothesis testing is a very elaborate contraption that people find very hard to make sense of. In contrast, the Bayesian approach to hypothesis testing is incredibly simple. Let's pick a setting that is closely analogous to the orthodox scenario. There are two hypotheses that we want to compare, a null hypothesis  $h_0$  and an alternative hypothesis  $h_1$ . Prior to running the experiment we have some beliefs  $P(h)$  about which hypotheses are true. We run an experiment and obtain data  $d$ . Unlike frequentist statistics, Bayesian statistics does allow us to talk about the probability that the null hypothesis is true. Better yet, it allows us to calculate the **posterior probability of the null hypothesis**, using Bayes' rule

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

This formula tells us exactly how much belief we should have in the null hypothesis after having observed the data  $d$ . Similarly, we can work out how much belief to place in the alternative hypothesis using essentially the same equation. All we do is change the subscript

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d)}$$

It's all so simple that I feel like an idiot even bothering to write these equations down, since all I'm doing is copying Bayes rule from the previous section.\*<sup>8</sup>

### 13.2.1 The Bayes factor

In practice, most Bayesian data analysts tend not to talk in terms of the raw posterior probabilities  $P(h_0|d)$  and  $P(h_1|d)$ . Instead, we tend to talk in terms of the **posterior odds** ratio. Think of it like betting. Suppose, for instance, the posterior probability of the null hypothesis is 25%, and the posterior probability of the alternative is 75%. The alternative hypothesis is three times as probable as the null, so we say that the *odds* are 3:1 in favour of the alternative. Mathematically, all we have to do to calculate the posterior odds is divide one posterior probability by the other

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

---

\*<sup>8</sup>Obviously, this is a highly simplified story. All the complexity of real life Bayesian hypothesis testing comes down to how you calculate the likelihood  $P(d|h)$  when the hypothesis  $h$  is a complex and vague thing. I'm not going to talk about those complexities in this book, but I do want to highlight that although this simple story is true as far as it goes, real life is messier than I'm able to cover in an introductory stats textbook.

Or, to write the same thing in terms of the equations above

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

Actually, this equation is worth expanding on. There are three different terms here that you should know. On the left hand side, we have the posterior odds, which tells you what you believe about the relative plausibility of the null hypothesis and the alternative hypothesis *after* seeing the data. On the right hand side, we have the **prior odds**, which indicates what you thought *before* seeing the data. In the middle, we have the **Bayes factor**, which describes the amount of evidence provided by the data

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

↑                              ↑                              ↑  
Posterior odds      Bayes factor      Prior odds

The Bayes factor (sometimes abbreviated as **BF**) has a special place in Bayesian hypothesis testing, because it serves a similar role to the *p*-value in orthodox hypothesis testing. The Bayes factor quantifies the strength of evidence provided by the data, and as such it is the Bayes factor that people tend to report when running a Bayesian hypothesis test. The reason for reporting Bayes factors rather than posterior odds is that different researchers will have different priors. Some people might have a strong bias to believe the null hypothesis is true, others might have a strong bias to believe it is false. Because of this, the polite thing for an applied researcher to do is report the Bayes factor. That way, anyone reading the paper can multiply the Bayes factor by their own *personal* prior odds, and they can work out for themselves what the posterior odds would be. In any case, by convention we like to pretend that we give equal consideration to both the null hypothesis and the alternative, in which case the prior odds equals 1, and the posterior odds becomes the same as the Bayes factor.

### 13.2.2 Interpreting Bayes factors

One of the really nice things about the Bayes factor is the numbers are inherently meaningful. If you run an experiment and you compute a Bayes factor of 4, it means that the evidence provided by your data corresponds to betting odds of 4:1 in favour of the alternative. However, there have been some attempts to quantify the standards of evidence that would be considered meaningful in a scientific context. The two most widely used are from **Jeffreys1961** and **Kass1995**. Of the two, I tend to prefer the **Kass1995** table because it's a bit more conservative. So here it is:

Bayes factor	Interpretation
1 - 3	Negligible evidence
3 - 20	Positive evidence
20 - 150	Strong evidence
>150	Very strong evidence

And to be perfectly honest, I think that even the **Kass1995** standards are being a bit charitable. If it were up to me, I'd have called the "positive evidence" category "weak evidence". To me, anything in the range 3:1 to 20:1 is "weak" or "modest" evidence at best. But there are no hard and fast rules here. What counts as strong or weak evidence depends entirely on how conservative you are and upon the standards that your community insists upon before it is willing to label a finding as "true".

In any case, note that all the numbers listed above make sense if the Bayes factor is greater than 1 (i.e., the evidence favours the alternative hypothesis). However, one big practical advantage of the Bayesian approach relative to the orthodox approach is that it also allows you to quantify evidence *for* the null. When that happens, the Bayes factor will be less than 1. You can choose to report a Bayes factor less than 1, but to be honest I find it confusing. For example, suppose that the likelihood of the data under the null hypothesis  $P(d|h_0)$  is equal to 0.2, and the corresponding likelihood  $P(d|h_1)$  under the alternative hypothesis is 0.1. Using the equations given above, Bayes factor here would be

$$BF = \frac{P(d|h_1)}{P(d|h_0)} = \frac{0.1}{0.2} = 0.5$$

Read literally, this result tells us that the evidence in favour of the alternative is 0.5 to 1. I find this hard to understand. To me, it makes a lot more sense to turn the equation "upside down", and report the amount of evidence in favour of the *null*. In other words, what we calculate is this

$$BF' = \frac{P(d|h_0)}{P(d|h_1)} = \frac{0.2}{0.1} = 2$$

And what we would report is a Bayes factor of 2:1 in favour of the null. Much easier to understand, and you can interpret this using the table above.

### 13.3 \_\_\_\_\_

## Why be a Bayesian?

Up to this point I've focused exclusively on the logic underpinning Bayesian statistics. We've talked about the idea of "probability as a degree of belief", and what it implies about how a rational

agent should reason about the world. The question that you have to answer for yourself is this: how do *you* want to do your statistics? Do you want to be an orthodox statistician, relying on sampling distributions and *p*-values to guide your decisions? Or do you want to be a Bayesian, relying on things like prior beliefs, Bayes factors and the rules for rational belief revision? And to be perfectly honest, I can't answer this question for you. Ultimately it depends on what you think is right. It's your call and your call alone. That being said, I can talk a little about why *I* prefer the Bayesian approach.

### 13.3.1 Statistics that mean what you think they mean

*You keep using that word. I do not think it means what you think it means*

– Inigo Montoya, *The Princess Bride*<sup>\*9</sup>

To me, one of the biggest advantages to the Bayesian approach is that it answers the right questions. Within the Bayesian framework, it is perfectly sensible and allowable to refer to “the probability that a hypothesis is true”. You can even try to calculate this probability. Ultimately, isn’t that what you *want* your statistical tests to tell you? To an actual human being, this would seem to be the whole *point* of doing statistics, i.e., to determine what is true and what isn’t. Any time that you aren’t exactly sure about what the truth is, you should use the language of probability theory to say things like “there is an 80% chance that Theory A is true, but a 20% chance that Theory B is true instead”.

This seems so obvious to a human, yet it is explicitly forbidden within the orthodox framework. To a frequentist, such statements are a nonsense because “the theory is true” is not a repeatable event. A theory is true or it is not, and no probabilistic statements are allowed, no matter how much you might want to make them. There’s a reason why, back in Section ??, I repeatedly warned you *not* to interpret the *p*-value as the probability that the null hypothesis is true. There’s a reason why almost every textbook on statistics is forced to repeat that warning. It’s because people desperately *want* that to be the correct interpretation. Frequentist dogma notwithstanding, a lifetime of experience of teaching undergraduates and of doing data analysis on a daily basis suggests to me that most actual humans think that “the probability that the hypothesis is true” is not only meaningful, it’s the thing we care *most* about. It’s such an appealing idea that even trained statisticians fall prey to the mistake of trying to interpret a *p*-value this way. For example, here

---

<sup>\*9</sup><http://www.imdb.com/title/tt0093779/quotes>. I should note in passing that I’m not the first person to use this quote to complain about frequentist methods. Rich Morey and colleagues had the idea first. I’m shamelessly stealing it because it’s such an awesome pull quote to use in this context and I refuse to miss any opportunity to quote *The Princess Bride*.

is a quote from an official Newspoll report in 2013, explaining how to interpret their (frequentist) data analysis:<sup>\*10</sup>

Throughout the report, where relevant, statistically significant changes have been noted. All significance tests have been based on the 95 percent level of confidence.

**This means that if a change is noted as being statistically significant, there is a 95 percent probability that a real change has occurred,** and is not simply due to chance variation. (emphasis added)

Nope! That's *not* what  $p < .05$  means. That's *not* what 95% confidence means to a frequentist statistician. The bolded section is just plain wrong. Orthodox methods cannot tell you that "there is a 95% chance that a real change has occurred", because this is not the kind of event to which frequentist probabilities may be assigned. To an ideological frequentist, this sentence should be meaningless. Even if you're a more pragmatic frequentist, it's still the wrong definition of a  $p$ -value. It is simply not an allowed or correct thing to say if you want to rely on orthodox statistical tools.

On the other hand, let's suppose you are a Bayesian. Although the bolded passage is the wrong definition of a  $p$ -value, it's pretty much exactly what a Bayesian means when they say that the posterior probability of the alternative hypothesis is greater than 95%. And here's the thing. If the Bayesian posterior is actually the thing you *want* to report, why are you even trying to use orthodox methods? If you want to make Bayesian claims, all you have to do is be a Bayesian and use Bayesian tools.

Speaking for myself, I found this to be the most liberating thing about switching to the Bayesian view. Once you've made the jump, you no longer have to wrap your head around counter-intuitive definitions of  $p$ -values. You don't have to bother remembering why you can't say that you're 95% confident that the true mean lies within some interval. All you have to do is be honest about what you believed before you ran the study and then report what you learned from doing it. Sounds nice, doesn't it? To me, this is the big promise of the Bayesian approach. You do the analysis you really want to do, and express what you really believe the data are telling you.

### 13.3.2 **Evidentiary standards you can believe**

If  $[p]$  is below .02 it is strongly indicated that the [null] hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional

---

<sup>\*10</sup><http://about.abc.net.au/reports-publications/appreciation-survey-summary-report-2013/>

line at .05 and consider that [smaller values of  $p$ ] indicate a real discrepancy.

– Sir Ronald **Fisher1925**

Consider the quote above by Sir Ronald Fisher, one of the founders of what has become the orthodox approach to statistics. If anyone has ever been entitled to express an opinion about the intended function of  $p$ -values, it's Fisher. In this passage, taken from his classic guide *Statistical Methods for Research Workers*, he's pretty clear about what it means to reject a null hypothesis at  $p < .05$ . In his opinion, if we take  $p < .05$  to mean there is "a real effect", then "we shall not often be astray". This view is hardly unusual. In my experience, most practitioners express views very similar to Fisher's. In essence, the  $p < .05$  convention is assumed to represent a fairly stringent evidential standard.

Well, how true is that? One way to approach this question is to try to convert  $p$ -values to Bayes factors, and see how the two compare. It's not an easy thing to do because a  $p$ -value is a fundamentally different kind of calculation to a Bayes factor, and they don't measure the same thing. However, there have been some attempts to work out the relationship between the two, and it's somewhat surprising. For example, **Johnson2013** presents a pretty compelling case that (for  $t$ -tests at least) the  $p < .05$  threshold corresponds roughly to a Bayes factor of somewhere between 3:1 and 5:1 in favour of the alternative. If that's right, then Fisher's claim is a bit of a stretch. Let's suppose that the null hypothesis is true about half the time (i.e., the prior probability of  $H_0$  is 0.5), and we use those numbers to work out the posterior probability of the null hypothesis given that it has been rejected at  $p < .05$ . Using the data from **Johnson2013**, we see that if you reject the null at  $p < .05$ , you'll be correct about 80% of the time. I don't know about you but, in my opinion, an evidential standard that ensures you'll be wrong on 20% of your decisions isn't good enough. The fact remains that, quite contrary to Fisher's claim, if you reject at  $p < .05$  you shall quite often go astray. It's not a very stringent evidential threshold at all.

### 13.3.3 The $p$ -value is a lie.

*The cake is a lie.*

– Portal<sup>\*11</sup>

---

<sup>\*11</sup><http://knowyourmeme.com/memes/the-cake-is-a-lie>

Okay, at this point you might be thinking that the real problem is not with orthodox statistics, just the  $p < .05$  standard. In one sense, that's true. The recommendation that **Johnson2013** gives is not that "everyone must be a Bayesian now". Instead, the suggestion is that it would be wiser to shift the conventional standard to something like a  $p < .01$  level. That's not an unreasonable view to take, but in my view the problem is a little more severe than that. In my opinion, there's a fairly big problem built into the way most (but not all) orthodox hypothesis tests are constructed. They are grossly naive about how humans actually do research, and because of this most  $p$ -values are wrong.

Sounds like an absurd claim, right? Well, consider the following scenario. You've come up with a really exciting research hypothesis and you design a study to test it. You're very diligent, so you run a power analysis to work out what your sample size should be, and you run the study. You run your hypothesis test and out pops a  $p$ -value of 0.072. Really bloody annoying, right?

What should you do? Here are some possibilities:

1. You conclude that there is no effect and try to publish it as a null result
2. You guess that there might be an effect and try to publish it as a "borderline significant" result
3. You give up and try a new study
4. You collect some more data to see if the  $p$  value goes up or (preferably!) drops below the "magic" criterion of  $p < .05$

Which would *you* choose? Before reading any further, I urge you to take some time to think about it. Be honest with yourself. But don't stress about it too much, because you're screwed no matter what you choose. Based on my own experiences as an author, reviewer and editor, as well as stories I've heard from others, here's what will happen in each case:

- Let's start with option 1. If you try to publish it as a null result, the paper will struggle to be published. Some reviewers will think that  $p = .072$  is not really a null result. They'll argue it's borderline significant. Other reviewers will agree it's a null result but will claim that even though some null results *are* publishable, yours isn't. One or two reviewers might even be on your side, but you'll be fighting an uphill battle to get it through.
- Okay, let's think about option number 2. Suppose you try to publish it as a borderline significant result. Some reviewers will claim that it's a null result and should not be published. Others will claim that the evidence is ambiguous, and that you should collect more data until you get a clear significant result. Again, the publication process does not favour you.
- Given the difficulties in publishing an "ambiguous" result like  $p = .072$ , option number 3

might seem tempting: give up and do something else. But that's a recipe for career suicide. If you give up and try a new project every time you find yourself faced with ambiguity, your work will never be published. And if you're in academia without a publication record you can lose your job. So that option is out.

- It looks like you're stuck with option 4. You don't have conclusive results, so you decide to collect some more data and re-run the analysis. Seems sensible, but unfortunately for you, if you do this all of your  $p$ -values are now incorrect. *All* of them. Not just the  $p$ -values that you calculated for *this* study. All of them. All the  $p$ -values you calculated in the past and all the  $p$ -values you will calculate in the future. Fortunately, no-one will notice. You'll get published, and you'll have lied.

Wait, what? How can that last part be true? I mean, it sounds like a perfectly reasonable strategy doesn't it? You collected some data, the results weren't conclusive, so now what you want to do is collect more data until the the results *are* conclusive. What's wrong with that?

Honestly, there's nothing wrong with it. It's a reasonable, sensible and rational thing to do. In real life, this is exactly what every researcher does. Unfortunately, the theory of null hypothesis testing as I described it in Chapter ?? *forbids* you from doing this.<sup>\*12</sup> The reason is that the theory assumes that the experiment is finished and all the data are in. And because it assumes the experiment is over, it only considers *two* possible decisions. If you're using the conventional  $p < .05$  threshold, those decisions are:

Outcome	Action
$p$ less than .05	Reject the null
$p$ greater than .05	Retain the null

What *you're* doing is adding a third possible action to the decision making problem. Specifically, what *you're* doing is using the  $p$ -value itself as a reason to justify continuing the experiment. And as a consequence you've transformed the decision-making procedure into one that looks more like this:

---

<sup>\*12</sup>In the interests of being completely honest, I should acknowledge that not all orthodox statistical tests rely on this silly assumption. There are a number of *sequential analysis* tools that are sometimes used in clinical trials and the like. These methods are built on the assumption that data are analysed as they arrive, and these tests aren't horribly broken in the way I'm complaining about here. However, sequential analysis methods are constructed in a very different fashion to the "standard" version of null hypothesis testing. They don't make it into any introductory textbooks, and they're not very widely used in the psychological literature. The concern I'm raising here is valid for every single orthodox test I've presented so far and for almost every test I've seen reported in the papers I read.

Outcome	Action
$p$ less than .05	Stop the experiment and reject the null
$p$ between .05 and .1	Continue the experiment
$p$ greater than .1	Stop the experiment and retain the null

The “basic” theory of null hypothesis testing isn’t built to handle this sort of thing, not in the form I described back in Chapter ???. If you’re the kind of person who would choose to “collect more data” in real life, it implies that you are *not* making decisions in accordance with the rules of null hypothesis testing. Even if you happen to arrive at the same decision as the hypothesis test, you aren’t following the decision *process* it implies, and it’s this failure to follow the process that is causing the problem.\*<sup>13</sup> Your  $p$ -values are a lie.

Worse yet, they’re a lie in a dangerous way, because they’re all *too small*. To give you a sense of just how bad it can be, consider the following (worst case) scenario. Imagine you’re a really super-enthusiastic researcher on a tight budget who didn’t pay any attention to my warnings above. You design a study comparing two groups. You desperately want to see a significant result at the  $p < .05$  level, but you really don’t want to collect any more data than you have to (because it’s expensive). In order to cut costs you start collecting data but every time a new observation arrives you run a  $t$ -test on your data. If the  $t$ -tests says  $p < .05$  then you stop the experiment and report a significant result. If not, you keep collecting data. You keep doing this until you reach your pre-defined spending limit for this experiment. Let’s say that limit kicks in at  $N = 1000$  observations. As it turns out, the truth of the matter is that there is no real effect to be found: the null hypothesis is true. So, what’s the chance that you’ll make it to the end of the experiment and (correctly) conclude that there is no effect? In an ideal world, the answer here should be 95%. After all, the whole *point* of the  $p < .05$  criterion is to control the Type I error rate at 5%, so what we’d hope is that there’s only a 5% chance of falsely rejecting the null hypothesis in this situation. However, there’s no guarantee that will be true. You’re breaking the rules. Because you’re running tests repeatedly, “peeking” at your data to see if you’ve gotten a significant result, all bets are off.

So how bad is it? The answer is shown as the solid black line in Figure ???, and it’s *astoundingly* bad. If you peek at your data after every single observation, there is a 49% chance that you will make a Type I error. That’s, um, quite a bit bigger than the 5% that it’s supposed to be. By way of comparison, imagine that you had used the following strategy. Start collecting data. Every single time an observation arrives, run a *Bayesian t-test* (Section ???) and look at the Bayes factor. I’ll assume that **Johnson2013** is right, and I’ll treat a Bayes factor of 3:1 as roughly equivalent to

---

\*<sup>13</sup>A related problem: <http://xkcd.com/1478/>.

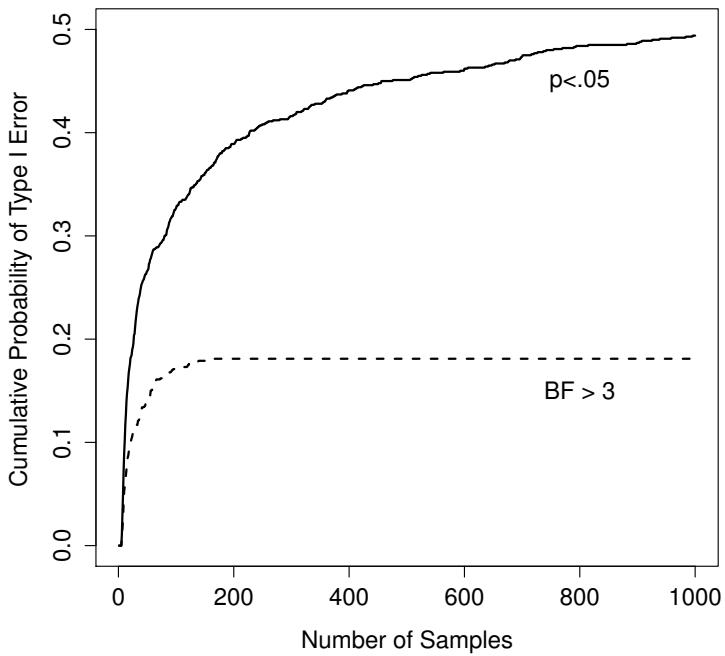


Figure13.1 How badly can things go wrong if you re-run your tests every time new data arrive? If you are a frequentist, the answer is “very wrong”.

.....

a  $p$ -value of .05.<sup>\*14</sup> This time around, our trigger happy researcher uses the following procedure. If the Bayes factor is 3:1 or more in favour of the null, stop the experiment and retain the null. If it is 3:1 or more in favour of the alternative, stop the experiment and reject the null. Otherwise continue testing. Now, just like last time, let’s assume that the null hypothesis is true. What happens? As it happens, I ran the simulations for this scenario too, and the results are shown as the dashed line in Figure ???. It turns out that the Type I error rate is much much lower than the 49% rate that we were getting by using the orthodox  $t$ -test.

In some ways, this is remarkable. The entire *point* of orthodox null hypothesis testing is to control the Type I error rate. Bayesian methods aren’t actually designed to do this at all. Yet, as it turns out, when faced with a “trigger happy” researcher who keeps running hypothesis tests as the data come in, the Bayesian approach is much more effective. Even the 3:1 standard, which

---

<sup>\*14</sup>Some readers might wonder why I picked 3:1 rather than 5:1, given that Johnson2013 suggests that  $p = .05$  lies somewhere in that range. I did so in order to be charitable to the  $p$ -value. If I’d chosen a 5:1 Bayes factor instead, the results would look even better for the Bayesian approach.

most Bayesians would consider unacceptably lax, is much safer than the  $p < .05$  rule.

#### 13.3.4 Is it really this bad?

The example I gave in the previous section is a pretty extreme situation. In real life, people don't run hypothesis tests every time a new observation arrives. So it's not fair to say that the  $p < .05$  threshold "really" corresponds to a 49% Type I error rate (i.e.,  $p = .49$ ). But the fact remains that if you want your  $p$ -values to be honest then you either have to switch to a completely different way of doing hypothesis tests or enforce a strict rule of *no peeking*. You are *not* allowed to use the data to decide when to terminate the experiment. You are *not* allowed to look at a "borderline"  $p$ -value and decide to collect more data. You aren't even allowed to change your data analysis strategy after looking at data. You are strictly required to follow these rules, otherwise the  $p$ -values you calculate will be nonsense.

And yes, these rules are surprisingly strict. As a class exercise a couple of years back, I asked students to think about this scenario. Suppose you started running your study with the intention of collecting  $N = 80$  people. When the study starts out you follow the rules, refusing to look at the data or run any tests. But when you reach  $N = 50$  your willpower gives in... and you take a peek. Guess what? You've got a significant result! Now, sure, you know you *said* that you'd keep running the study out to a sample size of  $N = 80$ , but it seems sort of pointless now, right? The result is significant with a sample size of  $N = 50$ , so wouldn't it be wasteful and inefficient to keep collecting data? Aren't you tempted to stop? Just a little? Well, keep in mind that if you do, your Type I error rate at  $p < .05$  just ballooned out to 8%. When you report  $p < .05$  in your paper, what you're *really* saying is  $p < .08$ . That's how bad the consequences of "just one peek" can be.

Now consider this. The scientific literature is filled with  $t$ -tests, ANOVAs, regressions and chi-square tests. When I wrote this book I didn't pick these tests arbitrarily. The reason why these four tools appear in most introductory statistics texts is that these are the bread and butter tools of science. None of these tools include a correction to deal with "data peeking": they all assume that you're not doing it. But how realistic is that assumption? In real life, how many people do you think have "peeked" at their data before the experiment was finished and adapted their subsequent behaviour after seeing what the data looked like? Except when the sampling procedure is fixed by an external constraint, I'm guessing the answer is "most people have done it". If that has happened, you can infer that the reported  $p$ -values are wrong. Worse yet, because we don't know what decision process they actually followed, we have no way to know what the  $p$ -values *should* have been. You can't compute a  $p$ -value when you don't know the decision making procedure that

the researcher used. And so the reported  $p$ -value remains a lie.

Given all of the above, what is the take home message? It's not that Bayesian methods are foolproof. If a researcher is determined to cheat, they can always do so. Bayes' rule cannot stop people from lying, nor can it stop them from rigging an experiment. That's not my point here. My point is the same one I made at the very beginning of the book in Section ??: the reason why we run statistical tests is to protect us from ourselves. And the reason why "data peeking" is such a concern is that it's so tempting, even for honest researchers. A theory for statistical inference has to acknowledge this. Yes, you might try to defend  $p$ -values by saying that it's the fault of the researcher for not using them properly, but to my mind that misses the point. A theory of statistical inference that is so completely naive about humans that it doesn't even consider the possibility that the researcher might *look at their own data* isn't a theory worth having. In essence, my point is this:

*Good laws have their origins in bad morals.*

– Ambrosius Macrobius<sup>\*15</sup>

Good rules for statistical testing have to acknowledge human frailty. None of us are without sin. None of us are beyond temptation. A good system for statistical inference should still work even when it is used by actual humans. Orthodox null hypothesis testing does not.<sup>\*16</sup>

## 13.4

---

### Bayesian $t$ -tests

An important type of statistical inference problem discussed in this book is the comparison between two means, discussed in some detail in the chapter on  $t$ -tests (Chapter ??). If you can remember back that far, you'll recall that there are several versions of the  $t$ -test. I'll talk a little about Bayesian versions of the independent samples  $t$ -tests and the paired samples  $t$ -test in this

---

<sup>\*15</sup>[http://www.quotationspage.com/quotes/Ambrosius\\_Macrobius/](http://www.quotationspage.com/quotes/Ambrosius_Macrobius/)

<sup>\*16</sup>Okay, I just know that some knowledgeable frequentists will read this and start complaining about this section. Look, I'm not dumb. I absolutely know that if you adopt a sequential analysis perspective you can avoid these errors within the orthodox framework. I also know that you can explicitly design studies with interim analyses in mind. So yes, in one sense I'm attacking a "straw man" version of orthodox methods. However, the straw man that I'm attacking is the one that is used by almost every single practitioner. If it ever reaches the point where sequential methods become the norm among experimental psychologists and I'm no longer forced to read 20 extremely dubious ANOVAs a day, I promise I'll rewrite this section and dial down the vitriol. But until that day arrives, I stand by my claim that default Bayes factor methods are much more robust in the face of data analysis practices as they exist in the real world. Default orthodox methods suck, and we all know it.

section.

#### 13.4.1 Independent samples *t*-test

The most common type of *t*-test is the independent samples *t*-test, and it arises when you have data as in the `harpo.csv` data set that we used in the earlier chapter on *t*-tests (Chapter ??). In this data set, we have two groups of students, those who received lessons from Anastasia and those who took their classes with Bernadette. The question we want to answer is whether there's any difference in the grades received by these two groups of students. Back in Chapter ?? I suggested you could analyse this kind of data using the Independent Samples *t*-test in JASP, which gave us the results in Figure ???. As we obtain a *p*-value less than 0.05, we reject the null hypothesis.

## Results

### Bayesian Independent Samples T-Test

#### Bayesian Independent Samples T-Test

	BF <sub>10</sub>	error %
grade	1.755	7.565e -4

### Independent Samples T-Test

#### Independent Samples T-Test

	t	df	p
grade	2.115	31.000	0.043

*Note.* Student's t-test.

Figure13.2 Bayesian independent Samples *t*-test result in JASP

What does the Bayesian version of the *t*-test look like? We can get the Bayes factor analysis by selecting the 'T-Tests' - 'Bayesian Independent Samples T-Test' option. The dialog is similar to the conventional *t*-test from earlier, so you should already know what to do! For now, just accept

the defaults that JASP provides. This gives the results shown in the table in Figure ???. What we get in this table is a Bayes factor statistic of 1.755, meaning that the evidence provided by these data are about 1.8:1 in favour of the alternative hypothesis.

Before moving on, it's worth highlighting the difference between the orthodox test results and the Bayesian one. According to the orthodox test, we obtained a significant result, though only barely. Nevertheless, many people would happily accept  $p = .043$  as reasonably strong evidence for an effect. In contrast, notice that the Bayesian test doesn't even reach 2:1 odds in favour of an effect, and would be considered very weak evidence at best. In my experience that's a pretty typical outcome. Bayesian methods usually require more evidence before rejecting the null.

#### 13.4.2 Paired samples $t$ -test

Back in Section ?? I discussed the `chico.csv` data set in which student grades were measured on two tests, and we were interested in finding out whether grades went up from test 1 to test 2. Because every student did both tests, the tool we used to analyse the data was a paired samples  $t$ -test. Figure ?? shows the JASP results table for the conventional paired  $t$ -test alongside the Bayes factor analysis. At this point, I hope you can read this output without any difficulty. The data provide evidence of about 6000:1 in favour of the alternative. We could probably reject the null with some confidence!

## 13.5

---

### Summary

The first half of this chapter was focused primarily on the theoretical underpinnings of Bayesian statistics. I introduced the mathematics for how Bayesian inference works (Section ??), and gave a very basic overview of how Bayesian hypothesis testing is typically done (Section ??). Finally, I devoted some space to talking about why I think Bayesian methods are worth using (Section ??).

Then I gave a practical example, a Bayesian  $t$ -test (Section ??). If you're interested in learning more about the Bayesian approach, there are many good books you could look into. John Kruschke's book *Doing Bayesian Data Analysis* is a pretty good place to start (**Kruschke2011**) and is a nice mix of theory and practice. His approach is a little different to the "Bayes factor" approach that I've discussed here, so you won't be covering the same ground. If you're a cognitive psychologist, you might want to check out Michael Lee and E.J. Wagenmakers' book *Bayesian Cognitive Modeling* (**Lee2014**). I picked these two because I think they're especially useful for

## Results ▼

### Paired Samples T-Test

Paired Samples T-Test

		t	df	p
grade_test2	-	grade_test1	6.475	19 < .001

Note. Student's t-test.

### Bayesian Paired Samples T-Test ▼

Bayesian Paired Samples T-Test

		BF <sub>10</sub>	error %
grade_test2	-	grade_test1	5991.577 6.088e -8

Figure13.3 Paired samples T-Test and Bayes Factor result in JASP

.....

people in my discipline, but there's a lot of good books out there, so look around!

## 14. References

---

