1. 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[*1]

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 **??**)。

[*1] http://en.wikiquote.org/wiki/David_Hume.

1.1 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，この世界に関する仮説の候補 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，私は元々持っていた信念を修正しなければなりません。もしある仮説とデータが整合的であれば，その仮説についての私の信念は強められます。もしその仮説とデータが整合的でなければ，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 = 私は傘を持ち歩いている) を提示しています。そして，雨が降るかどうかに関する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？

1.1.1 事前確率: 以前あなたが信じていたもの

あなたが最初にするべきことは，私が傘について話したことを無視して，雨についてあなたが事前に抱いていた信念を書き出すことです。これは重要なことです。新しい証拠 (データ) に照らして信念がどのように修正されたのかについてあなたが正直でありたいのであれば，あなたは絶対に，そのデータが現れる前に信じていたことについて何かを言わねばなりません！ では，あなたは今日雨が降るかどうかについて何を信じているのでしょうか？ 私がオーストラリアに住んでいてること，そしてオーストラリアの多くの地域が暑くて乾燥しているということを，おそらくあなたは知っていることでしょう。私が住んでいるアデレード市は地中海性気候に属しており，南カリフォルニアや南ヨーロッパや北アフリカと非常によく似た気候です。私は 1 月にこの文章を執筆しているので，真夏であることをあなたは仮定できるでしょう。実際に，あなたは Wikipedia[*2]をざっと見て，アデレード市では 1 月の 31 日間に平均 4.4 日雨が降るという情報を見つけたかもしれませんね。他に何も知らないのであれば，アデレード市で 1 月に雨が降る確率はおよそ 15% であり，乾燥した日である確率はおよそ 85% であるとあなたは結論付けることでしょう。もしこれが本当にアデレード市の降水

[*2] http://en.wikipedia.org/wiki/Climate_of_Adelaide

についてあなたが信じていることだとしたら (それを私が今あなたに伝えたのだから，これこそが本当にあなたが信じていることであると私は確信しています)，私がここに書いたことがあなたの**事前分布**ということになります。これを $P(h)$ と書くことにしましょう。

仮説	信念の度合い
雨の日	0.15
乾燥した日	0.85

1.1.2 尤度: データに関する理論

この推論問題を解くためには私の行動についての理論が必要となります。ダンはいつ傘を持っていくのでしょうか？ あなたはおそらく，私が大馬鹿者ではないこと，[*3]そして私が雨の日にだけ傘を持ち歩こうとしていると推察するでしょう。一方で，私に幼い子どもがいることもあなたは知っていますし，私がこの種のことについてかなり忘れっぽいと分かってもあなたは驚きはしないでしょう。

そこで，私が傘のことを思い出せるのは雨が降っている日のうちのおよそ 30% であると仮定してみましょう (私は本当にこれが苦手です)。対照的に，乾燥した日に傘を持っていくことはおよそ 5% しかないということにしましょう。こうすると，以下のような小さな表を作ることができます。

	データ	
仮説	傘あり	傘なし
雨の日	0.30	0.70
乾燥した日	0.05	0.95

この表の各セルは，特定の仮説 h が真であるときにどのようなデータ d が観測されるかについてのあなたの信念を記述したものです。これを覚えておくことが大切です。

この“条件付き確率” は $P(d|h)$ と書き，“h のもとでの d の確率 (the probability of d given h)” と読みます。ベイズ統計学ではこれを，仮説 h のもとでのデータ d の**尤度**と呼びます。[*4]

[*3] これは根拠のない盲信ですが，そういうことにして議論を進めていきましょう。

[*4] うーん，この話はしたくないのですが，統計学者の中には私がここで “尤度” という言葉を使うことに異を唱える人もいるでしょう。“尤度” という言葉は頻度主義統計学では非常に特殊な意味を持っているので，ベイズ統計学における尤度の意味と完全に同じという訳ではない点が問題なのです。私が知る限りでは，ベイジアンにはもともと尤度に対する合意の取れた名称がなかったので，頻度主義の用語を使うことが一般的な慣習となってしまったのです。ベイジアンが頻度主義者とはずいぶん異なる使い方でこの言葉を使っていることが判明したという事実を除けば，このことは問題にはならなかったことでしょう。ここはさらにもうひとつの長い歴史について勉強する場所ではないのですが，大雑把に言えば，ベイジアンが “あるひとつの 尤度関数” と言うときには，この表の行のひとつを指しているのが普通です。頻度主義者が同じことを言うときにも同じ表を参照するのですが，彼らにとっての “あるひとつの 尤度関数” はほとんどいつでも列のひとつを指すのです。この区別は文脈によっては重要ですが，本章の目的においては重要ではありません。

1.1.3 データと仮説の同時分布

これで全ての要素が揃いました。

事前分布と尤度を書き出したので，あなたはベイズ推論に必要な全ての情報を手にしています。次に問題となるのは，この情報を私たちがどのように使うかです。結論から言えば，この状況で使うことのできる非常に簡単な等式があるのですが，重要なのは，なぜ私たちがその式を使うのかをあなたに理解してもらうことです。そこで，より基本的なアイディアからその式を構築することを試みてみたいと思います。

確率論の公式の 1 つから出発することにしましょう。私はその公式をかなり前に Table **??**に掲載したのですが，そのときはあまり大々的に取り上げなかったので，おそらく無視されていたことでしょう。ここで取り上げるのは，*2* つのことが同時に真である確率について語る公式です。先ほどの例でいえば，今日雨が降り (i.e., 仮説 h が真である)，かつ私が傘を持っていく (i.e., データ d が観測される) 確率を計算したいということです。仮説とデータの**同時確率**を $P(d, h)$ と書きます。この同時確率は，事前確率 $P(h)$ と尤度 $P(d|h)$ の掛け算で計算できます。数学的には，

$$P(d, h) = P(d|h)P(h)$$

ということです。では，今日雨が降り，かつ私が傘を持っていく確率はどのくらいでしょうか？ 先ほど議論したように，事前確率が教えてくれるのは雨の日である確率が 15% であるということであり，尤度が教えてくれるのは雨の日に私が傘のことを思い出せる確率が 30% であるということです。したがって，これらが両方とも真である確率は，これら 2 つを掛け合わせて，

$$\begin{aligned} P(\text{雨}, \text{傘あり}) &= P(\text{傘あり}|\text{雨}) \times P(\text{雨}) \\ &= 0.30 \times 0.15 \\ &= 0.045 \end{aligned}$$

と計算することができます。別の言い方をすれば，あなたは実際に何が起こったかを知らされるよりも前に，今日雨が降ってしかも私が傘のことを思い出す確率が 4.5% であると考えることになります。ただしもちろん，起こり得る可能性は *4* 通りありますよね？ なので，このエクササイズを全 4 通りについて繰り返してみましょう。すると，以下のような表ができます。

	傘あり	傘なし
雨の日	0.045	0.105
乾燥した日	0.0425	0.8075

この表には 4 つの可能性のうちどれが尤もらしいかに関する全ての情報が含まれていますが，全体像をきちんと把握するためには，この表に行と列の合計を加えるのが役に立ちます。そうすると次の表が得られます。

	傘あり	傘なし	合計
雨の日	0.0450	0.1050	0.15
乾燥した日	0.0425	0.8075	0.85
合計	0.0875	0.9125	1

この表はとても便利なので，この表に書かれた全ての数字が何を教えてくれるのかについて少し考えてみましょう。まずは，行の和が新しいことを何も語っていないことに注目してください。例えば一行目の和は，傘の問題を無視したときに，今日が雨の日である確率が 15% であることを示しています。もちろんこれは私たちの事前確率なので，何も驚くことはありません。[*5]重要なのは数値そのものではありません。むしろ重要なのは，私たちの計算が理にかなったものであるという自信をこの数値が与えてくれることです。行の和から降水確率が分かるのと同様に，列の和からは，私が傘を持っていく確率が分かります。具体的には，一列目の和は，私が傘を持っていく確率が，平均的には (i.e., 雨の日かどうかを無視すれば) 8.75% であることを示しています。最後に，論理的に可能な 4 つの事象の総和が 1 になることに注目してください。つまり，私たちが書き下したこの表は，データと仮説の全ての可能な組み合わせに対して定義された，適切な (proper) 確率分布であるということになります。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(Umbrella, Rainy)	P(No-umbrella, Rainy)	P(Rainy)
Dry	P(Umbrella, Dry)	P(No-umbrella, Dry)	P(Dry)
	P(Umbrella)	P(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	$P(h_2, d_1)$	$P(h_2, d_2)$	$P(h_2)$
	$P(d_1)$	$P(d_2)$	

[*5]ここではっきりさせておきたいのは，"事前" (prior) 情報とは既存の知識や信念のことであり，その情報を改善するためにデータを収集または使用する前に存在するものです。

1.1.4 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.[*6] 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

[*6] 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?

still ensuring that numbers in the table add up. Something like this, perhaps?

	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 [*7]

$$P(h|d) = \frac{P(d, h)}{P(d)}$$

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.

[*7] 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.

1.2

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.[*8]

1.2.1 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,

[*8] 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.

all we have to do to calculate the posterior odds is divide one posterior probability by the other

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

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 plausibilty 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

$$\underbrace{\frac{P(h_1|d)}{P(h_0|d)}}_{\text{Posterior odds}} = \underbrace{\frac{P(d|h_1)}{P(d|h_0)}}_{\textbf{Bayes factor}} \times \underbrace{\frac{P(h_1)}{P(h_0)}}_{\text{Prior odds}}$$

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.

1.2.2 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

$$\text{BF} = \frac{P(d|h_1)}{P(d|h_0)} = \frac{0.1}{0.2} = 0.5$$

Read literally, this result tells is 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 op evidence in favour of the *null*. In other words, what we calculate is this

$$\text{BF}' = \frac{P(d|h_0)}{P(d|h_1)} = \frac{0.2}{0.1} = 2$$

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.

1.3

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.

1.3.1 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[*9]

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 statstics is forced to repeat that warning. It's because people desperately *want* that to be the correct interpretation. Frequentist dogma notwithstanding,

[*9] 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*.

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 is a quote from an official Newspoll report in 2013, explaining how to interpret their (frequentist) data analysis:[*10]

> 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.

[*10] http://about.abc.net.au/reports-publications/appreciation-survey-summary-report-2013/

1.3.2 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 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.

1.3.3 The p-value is a lie.

> *The cake is a lie.*
> *The cake is a lie.*
> *The cake is a lie.*

The cake is a lie.

– Portal[*11]

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

[*11] http://knowyourmeme.com/memes/the-cake-is-a-lie

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.[*12] 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

[*12] 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.

as a consequence you've transformed the decision-making procedure into one that looks more like this:

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.[*13] 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

[*13] A related problem: http://xkcd.com/1478/.

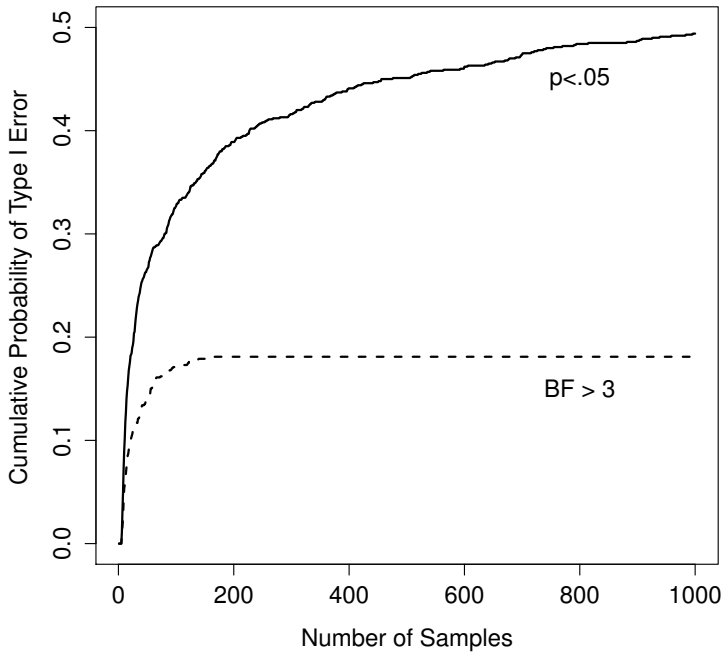


Figure1.1 How badly can things go wrong if you re-run your tests every time new data arrive? If you are a frequentist, the answer is "very wrong".

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 a *p*-value of .05.[*14] 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

[*14] 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.

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 most Bayesians would consider unacceptably lax, is much safer than the $p < .05$ rule.

1.3.4 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 analyis 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[*15]

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.[*16]

1.4 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

[*15] http://www.quotationspage.com/quotes/Ambrosius_Macrobius/

[*16] 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 explictly 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.

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 section.

1.4.1 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_{10}	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.

Figure1.2 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.

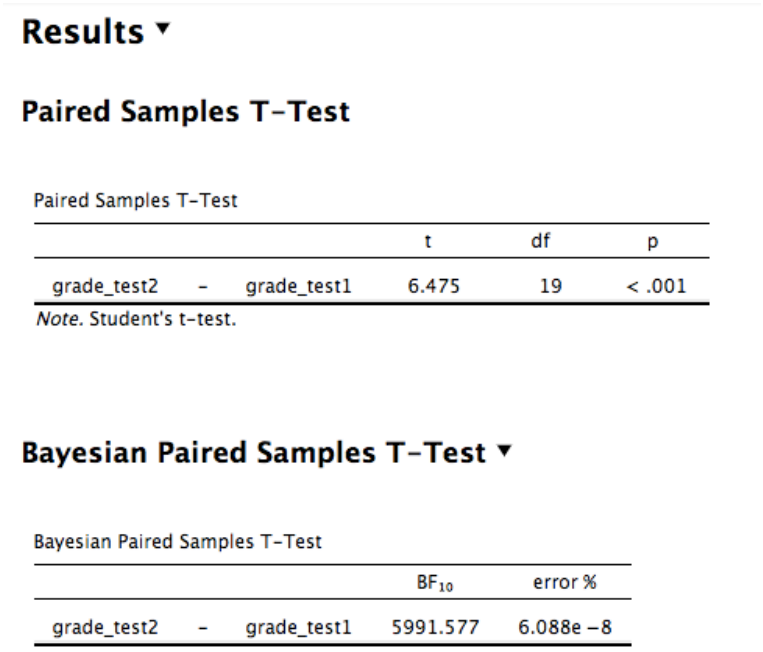
1.4.2 Paired samples *t*-test

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!

1.5 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



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_{10}	error %
grade_test2	-	grade_test1	5991.577	6.088e −8

Figure1.3 Paired samples T-Test and Bayes Factor result in JASP

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 people in my discipline, but there's a lot of good books out there, so look around!