How statistics can tell us what we really want to know

J.P. de Ruiter

Depts. of Computer Science & Psychology
Tufts University

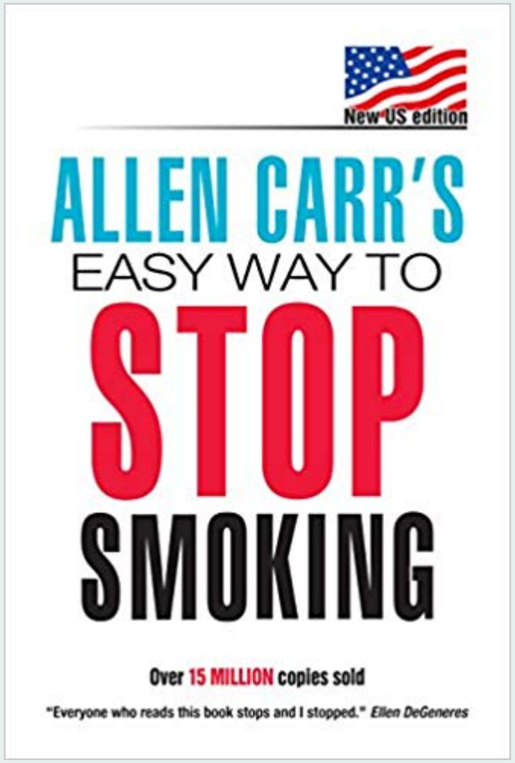
Introduction and background

- My background is in Cognitive Science.
- I am not an expert in statistics.
- I am an expert in being *confused* about statistics.

Let’s be honest

- Bestseller by Alan Carr.
- Thesis: quitting smoking is actually very easy!

- The snag: it isn’t. It’s actually *very* hard.
- But it feels good to believe it. For maybe a day.



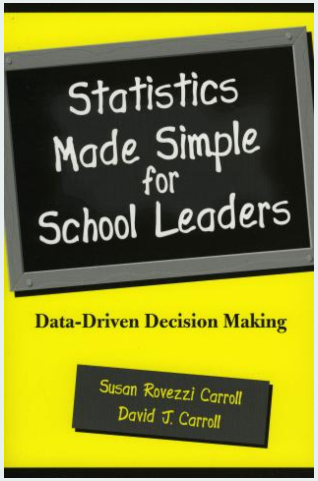
New US edition

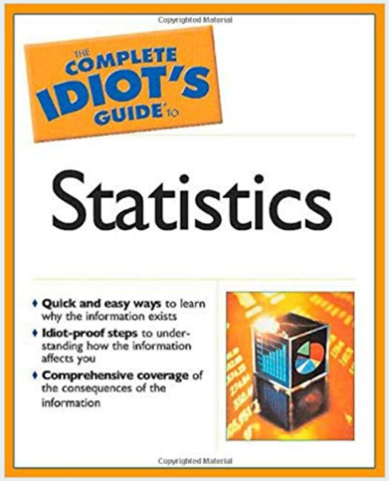
ALLEN CARR’S EASY WAY TO STOP SMOKING

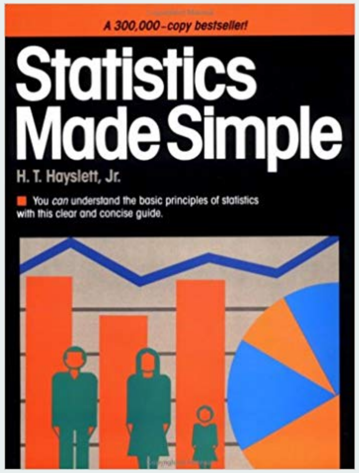
Over 15 MILLION copies sold

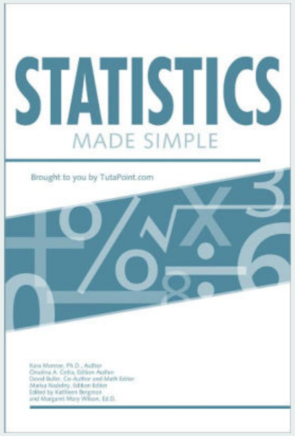
“Everyone who reads this book stops and I stopped.” *Ellen DeGeneres*

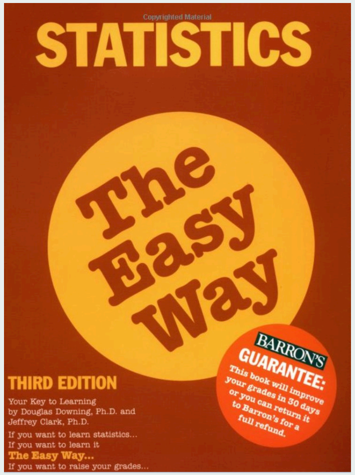
Statistics books (and courses)

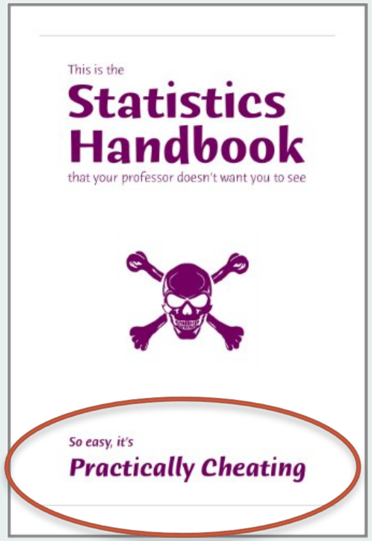

Statistics Made Simple for School Leaders
Data-Driven Decision Making
Susan Rovezzi Carroll
David J. Carroll


Copyrighted Material
THE COMPLETE IDIOT'S GUIDE TO
Statistics
Quick and easy ways to learn why the information exists
Idiot-proof steps to understanding how the information affects you
Comprehensive coverage of the consequences of the information
Copyrighted Material


A 300,000-copy bestseller!
Statistics Made Simple
H. T. Hayslett, Jr.
You can understand the basic principles of statistics with this clear and concise guide.


STATISTICS
MADE SIMPLE
Brought to you by TutaPoint.com
Kara Monroe, Ph.D., Author
Christina A. Cella, Edition Author
David Butler, Co-Author and Math Editor
Marisa Nadolny, Edition Editor
Edited by Kathleen Bergman
and Margaret Mary Wilson, Ed.D.


Copyrighted Material
STATISTICS
The Easy Way
THIRD EDITION
Your Key to Learning
by Douglas Downing, Ph.D. and
Jeffrey Clark, Ph.D.
If you want to learn statistics...
If you want to learn it
The Easy Way...
If you want to raise your grades...
BARRON'S
GUARANTEE:
This book will improve your grades in 30 days or you can return it to Barron's for a full refund.



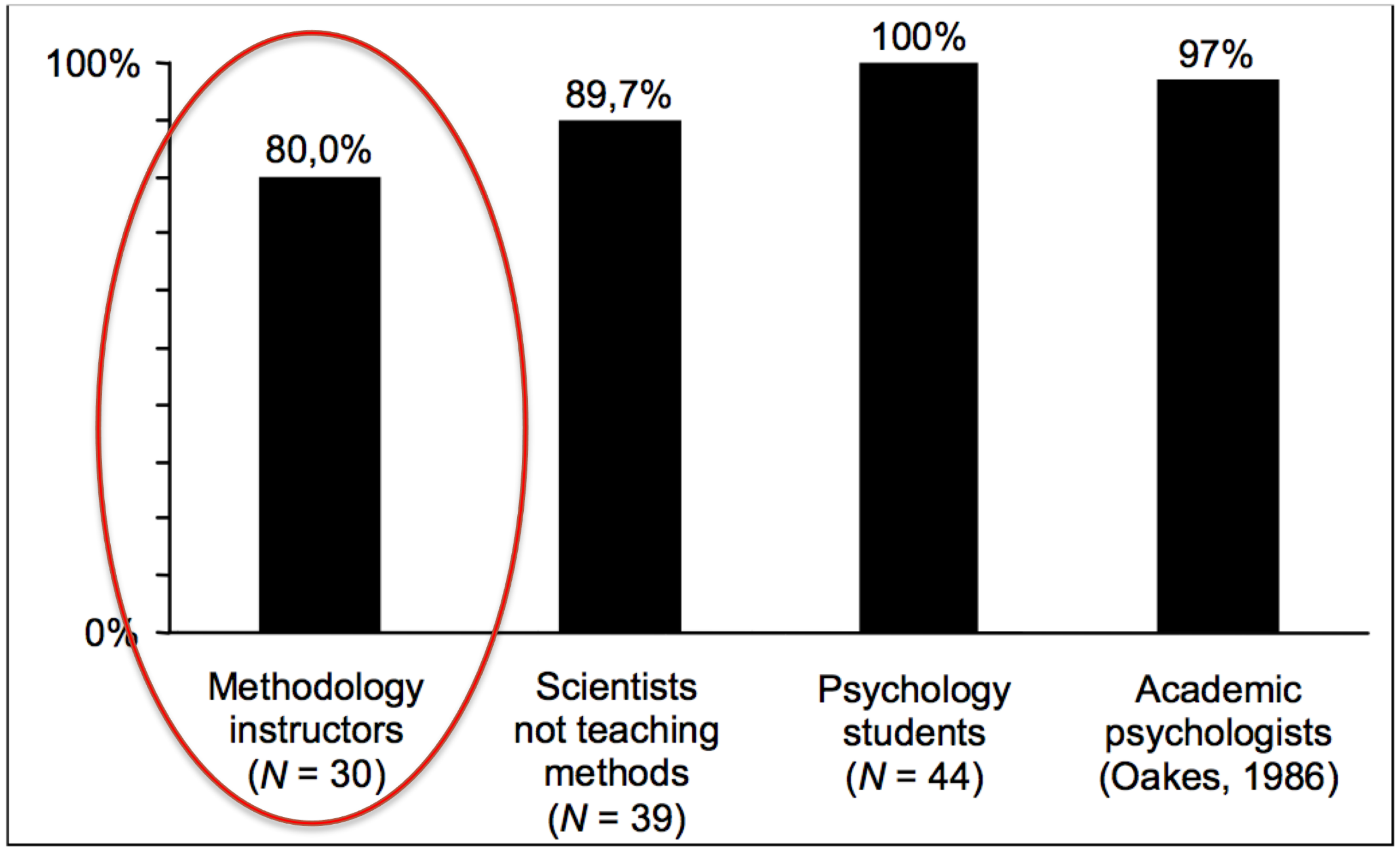
Reality

- Our standard statistical paradigm is *wildly* confusing.
- The most important thing that students seem to remember is that "*Oh yeah it was the other way around or something. Right?*"
- We see frequent and rampant misapplications (researchers) and misunderstandings (reviewers) in the literature.
- There is also empirical evidence that it is confusing.
- In his course, E.J. Wagenmakers offered 20 € to any psychology student who could define significance.
  - 2nd year Psychology students who had completed several methods courses.
  - He tried 7 times, and nobody could do it (correctly).
- Which is not that surprising:

Haller & Krauss 2002

- Asked 6 elementary TRUE/FALSE questions about NHST.
- If you do an experiment, and you get p = 0.01, then:
  1) You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).
  2) You have found the probability of the null hypothesis being true.
  3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
  4) You can deduce the probability of the experimental hypothesis being true.
  5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
  6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.
- Note: all statements are FALSE.

% of people making at least one error



Haller & Kraus, 2002

Results

% error per question

Statement (abbrev.)	Psych. 1986	Psych. 2000	Methods Instr. 2000
H0 is absolutely disproved	15	1	10
Probability of H0 is found	26	36	17
H1 is absolutely proved	13	6	10
Probability of H1 is found	33	66	33
Probability of Type I error	67	86	73
Probability of replication	49	60	37

The Plan

I. Our standard statistical paradigm in more detail. And why it is confusing.

II. What we want from a statistical paradigm.

III. A statistical paradigm that gives us almost everything we want.

IV. Some bad, good, and better news.

PART I

Our standard statistical paradigm in more detail.

And why it is confusing.

A note on ”statistics”

- There is no such thing as “statistics”.
- There is, however, an approach to it that most social scientists use.
- A few paradigms:
  - Null Hypothesis Significance Testing - NHST (Fisher, Neyman-Pearson)
  - Likelihood Statistics (Royall)
  - Information Criterion statistics (Akaike)
  - Bayesian Statistics (Jeffreys, Lindley)
- These are all different ways to support general claims based on collected data.

Why?

- *Why do we do inferential statistics at all*?
  - Why don’t we collect our data, present the relevant informative descriptives (e.g. means), make our claim, and get on with our lives?

- GROUP EXERCISE:
  - please form groups of 4 (or 3)
  - briefly introduce yourselves to each other.
  - Take 5-10 minutes to prepare an official ”statement” that answers this question.
  - Be as <u>precise as possible</u>. Imagine your statement will end up being a central text box in a methods book.
  - Note: “because otherwise the reviewers will reject our paper” is NOT an acceptable answer.

Answers (summarized)

- To generalize from a sample to population.
- To be reasonably sure that our effect wasn't due to chance.
- To have an indication of how sure we are of that.

Detailed NHST example

- Suppose we have measured the height of 4 randomly selected adult men and 4 randomly selected adult women.
- I did this in one of my classes.
- The data in cm:
  - Men: 183, 190, 178, 179
  - Women: 160, 180, 177, 165
- Descriptive statistics:
  - The mean height of the women is 170.5 cm
  - The mean height of the men is 182.5 cm
  - The difference between the groups is 12 cm taller for men.

How to proceed?

- So in this sample, the men are 12.5 cm taller than the women.
  - NOTE: we do not need inferential statistics to make this statement!
  - It is simply true in this group.
- Now how do we establish whether we can generalize this to the population? (let's say the USA here)
- We want to know now what the probability is that men are generally taller than women (and by how much). That's why we collected the data.
- In probability notation: we want to know
  - P(Men are taller|The data), or equivalently
  - P(Men are not taller|The data).
- Problem: early 20th century, this was impossible to compute.

Two heroes of NHST: Gossett and Fisher

- William Gosset (aka “Student”)
  - Beer brewer (Guinness)
  - Developed the t-test
  - Very modest guy

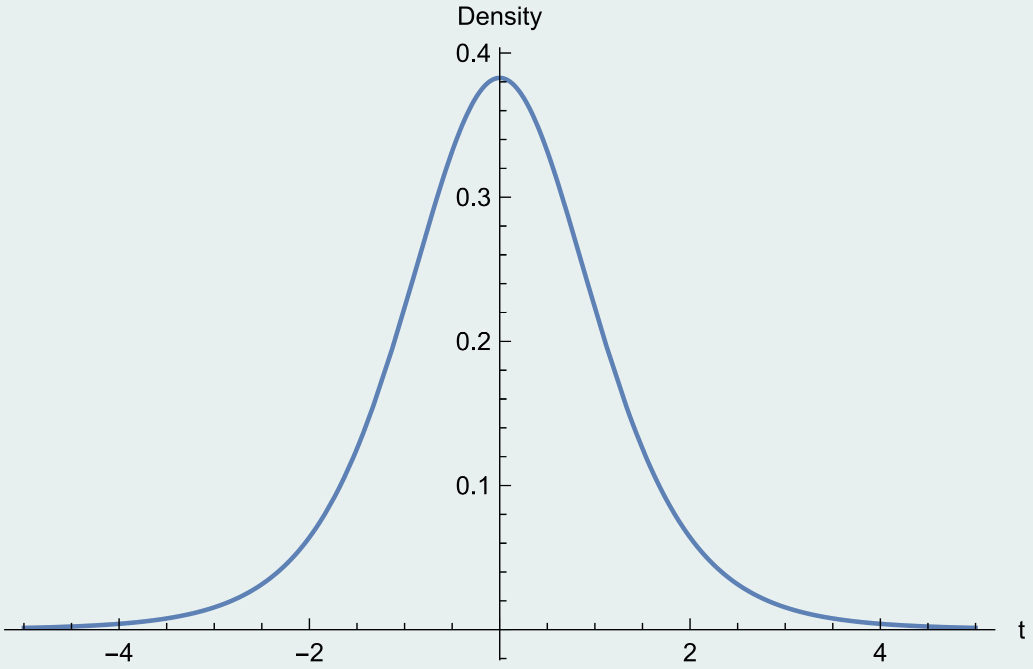
- Sir Ronald Fisher
  - Geneticist and statistician
  - Developed ANOVA and concept of p-values and many other statistical tests and concepts
  - Very confident guy

The procedure

- Formulate a *null hypothesis* that is the negation of what we want to say, in that it assumes there is no such difference. In this case:
  - H0: Men and women are of the same height, or women are taller.
  - H1: Men are taller than women.
- We define a *test statistic*, ***t***, a measure of the *difference* between the two groups.
  - In this case, the difference in means divided by the expected standard deviation of our estimate.
- Now we are going to repeat (in our minds) millions of experiments with 4 men and 4 women heights, randomly sampled, and see what the distribution of that difference measure is, *while assuming H0 is true.*
- *That is, we repeatedly draw 4 x 2 numbers from a normal distribution with mean 176.5 (the mean of the recorded heights in the data) and standard deviation estimated from the data, and then compute the t-difference between the two groups.*

The t-distribution

- Gosset could not simulate that, of course, but was able to mathematically derive the formula for the resulting distribution for groups of size N (in this case 4).
- This is called the *sampling distribution*



Now we compute t for our data

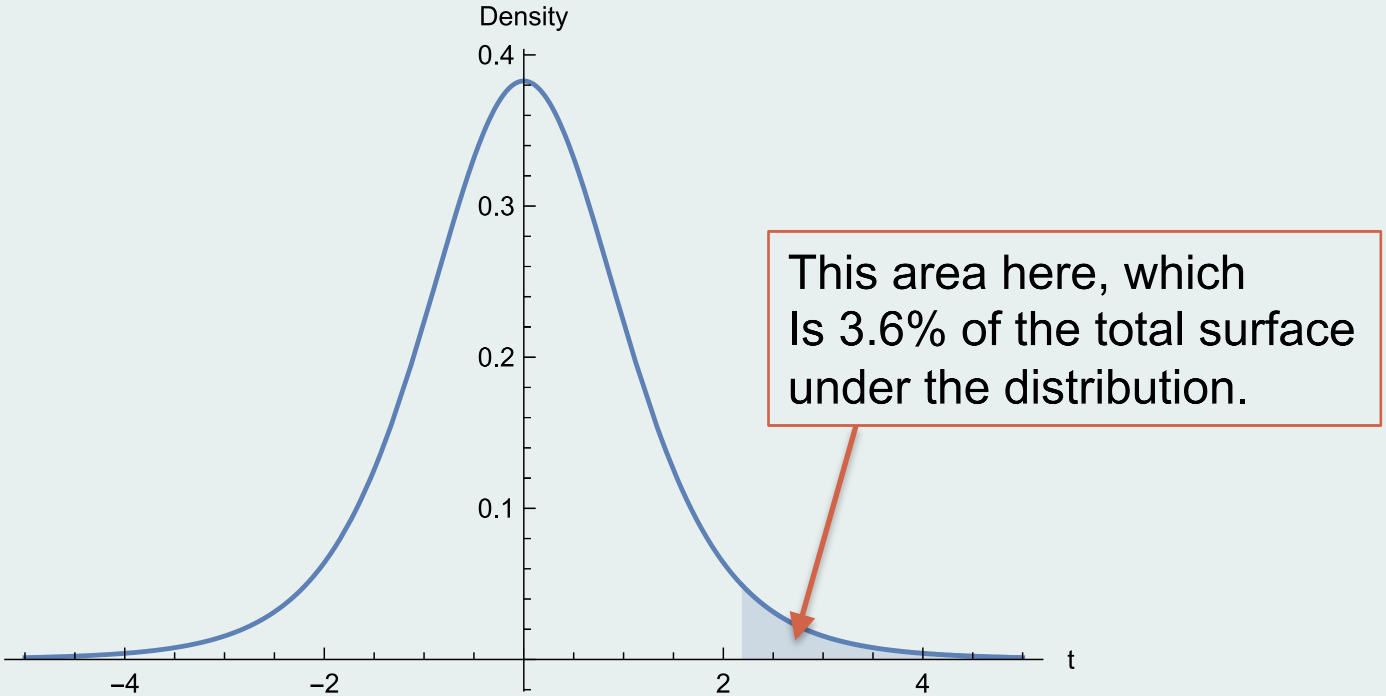
- We know that our difference in means is 12.5 cm
- We estimate the variance from our data, which is

  $\mathrm{S} = \sqrt{\frac{1}{2}}(S_m^2 + S_f^2) = 7.76$

- $t = \frac{\bar{M} - \bar{F}}{S\sqrt{\frac{2}{N}}} = 2.184$

And now comes the trick!

- We computed the test statistic for our height data: 2.184
- And we use our sampling distribution to compute the probability that **our difference** or **even larger differences** would occur **if Ho were true**!



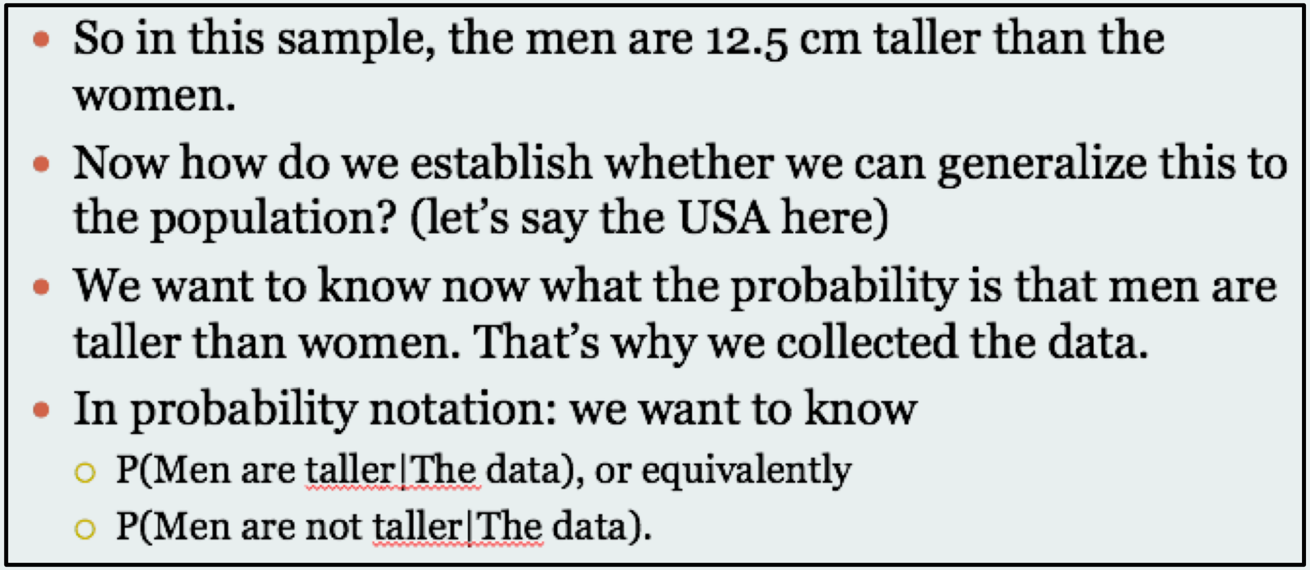
This area here, which
Is 3.6% of the total surface
under the distribution.

Finally: the infamous p-value

- So our (one-sided) p-value is 0.036, which allows us to state that:
  - *The probability that the difference that we found, or an even larger difference, would occur if H0 were true is 0.036.*
  - That is pretty small (smaller than .05, our threshold α)
  - Therefore we conclude that **either** something unlikely has happened, **or** our H0 was false.
  - As it is unlikelier than α, our conventional threshold, we conclude that H0 was false.
  - So we reject H0, and claim that H1 is true: men are taller than women.

Let’s think

- Was this clear? Any questions about this general procedure?
- Do you think this is a convincing line of argumentation?
- Did it answer our original question (repeated below)?



- So in this sample, the men are 12.5 cm taller than the women.
- Now how do we establish whether we can generalize this to the population? (let’s say the USA here)
- We want to know now what the probability is that men are taller than women. That’s why we collected the data.
- In probability notation: we want to know
  - P(Men are taller|The data), or equivalently
  - P(Men are not taller|The data).

What did we compute?

- We originally hoped to get P(H0|Data) or P(H1|Data).
- What we computed was P(Data+|H0).
- These are not the same, even if we ignore the + part.
- I want to make sure we understand that these two probabilities are VERY different, with a VERY clear example:
  - P(I died|I jumped out of a plane) ≈ 1 (but not 1)
  - P(I jumped out of a plane|I died) ≈ 0 (but not 0)

- Very many people believe that the p-value is equal to P(H0|Data).
- It is not. At all.
- This is also the main reason so many people made errors in studies like the one by Haller & Krauss.

Some colorful observations by experts

“What's wrong with [null hypothesis significance testing]? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!”

– Cohen (1994)

“… surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.”

– Rozeboom (1997)

But does it matter in practice?

- Sometimes people (e.g., popular methods teachers in psychology) have the “whatever” position, arguing that yes, for sticklers there is a *theoretical* difference, but in practice it doesn’t matter that much (e.g. Nickerson, 2000).
- This is simply not true, and I will give a few examples of problems that we have using this procedure, and not being fully aware of how it works.
- This is not going to be a comprehensive list. For that, please refer to the many articles written about that. [see references at the end]

PROBLEM 1: WHERE IS H1?

- Example: We want to test if a die is fair or not.
  - H0: Die is fair.
  - H1: Die is loaded (not fair).
- We throw it twice, and both times 6 comes up.
  - $P(2 \times 6 | H0) = 1/36 = 0.03$. This is smaller than .05
  - We conclude that the die is loaded.
- Is this a justified conclusion?
- It is according to the NHST logic.

PROBLEM 1: THE MISSING H1

- The problem, in the words of Sellke et al. (2001):

“Knowing that the data are ‘rare’ under H0 is of little use unless one determines whether or not they are also ‘rare’ under H1.”

- If the die is loaded, such that P(six) is not 1/6 but 1/5, throwing two sixes is *also* rare (1/25 = .04).
- The problem here is that we are tempted to think that if the result is rare under H0, it should be the *opposite of rare* under H1, because H0 and H1 are mutually exclusive.
- But that is a (very common) fallacy! When H0 and H1 are mutually exclusive:
  - ○ P(H0|D) + P(H1|D) = 1 [Always]
  - ○ P(D|H0) + P(D|H1) ≠ 1 [It can be 1, but it generally isn’t.]

PROBLEM 2: Interpreting nonsignificance

- What if we find a p-value that's above .05 (or α)?
- There are only 2 possible outcomes from an NHST test:
  a) $p < \alpha$, hence we have evidence for an effect
  b) $p \geq \alpha$, and then we know nothing.
- If $p > \alpha$, we know that the probability of our data or more extreme data is not as low as α.
- This could have two causes:
  - There is an effect, but we don't have enough data to make it unlikely enough to be NHST-significant.
  - There is no effect.
- We DO NOT KNOW which one is the real cause.

PROBLEM 2: Interpreting nonsignificance

This is weirder than you might think!

- Imagine you do a large and expensive experiment, with lots of controls, and 200 participants.
  - The p-value for your central claim is .09
- WHAT YOU SHOULD FORMALLY DO
  a) Write that we still don't know anything about the claim, other than that it is not significant.
  b) Do another experiment with more participants (you can't use the old data and add some new data to it).

PROBLEM 2: Interpreting nonsignificance

- WHAT PEOPLE ACTUALLY DO
  - Write that .09 is "trending towards marginally significant."
  - Write that with more participants, it's probably going to be significant.
    - We don't know that at all.
  - Write that there was no effect.
    - This is also not a valid conclusion.
  - Run extra participants.
    - Which is not allowed in NHST.
- These are "illegal" but understandable responses.
- Note that there is probably useful information in our data!
  - But we can't use that information for our inference!
- The underlying problem is that NHST has only two possible outcomes:
  a) Yeehaw, we have a significant effect!
  b) We don't know enough to draw any conclusion other than that we do not have significance.
- This is very frustrating.

PROBLEM 3: OPTIONAL STOPPING

- If we have a "marginally significant" result, not too far from .05, but not there yet, we would like to run more participants to "get it significant."
- Colleague at MPI: "I have already reached significance, so you can have the rest of my participants."
- This is NOT ALLOWED in NHST.
- The sampling distribution is based on a pre-defined sampling plan!
- If we deviate from that plan on the basis of "interim p-values" (which are strictly speaking not even defined) then we INFLATE our p-value. We would need to have a sampling plan that takes that into account.
- I don't care that everybody does this: we need to be aware that it gives us more false positives (which substantially contributes to the replication crisis).

PROBLEM 4: p < .05 is weaker than we think

- As we mentioned before, p < .05 does not mean that P(H0) < .05
- We do know that in the long run, if we use NHST with $\alpha = .05$ and **IF** H0 is true, we get around 5% false positives.
- But we don't *know* if H0 is true.
- For instance, if H0 is indeed true, every effect we find would be a false positive. And if H0 is false, none of our effects would be a false positive.
- So the actual "false discovery rate" is normally much higher than .05
- Why?

WHY IS THIS? AN EXAMPLE

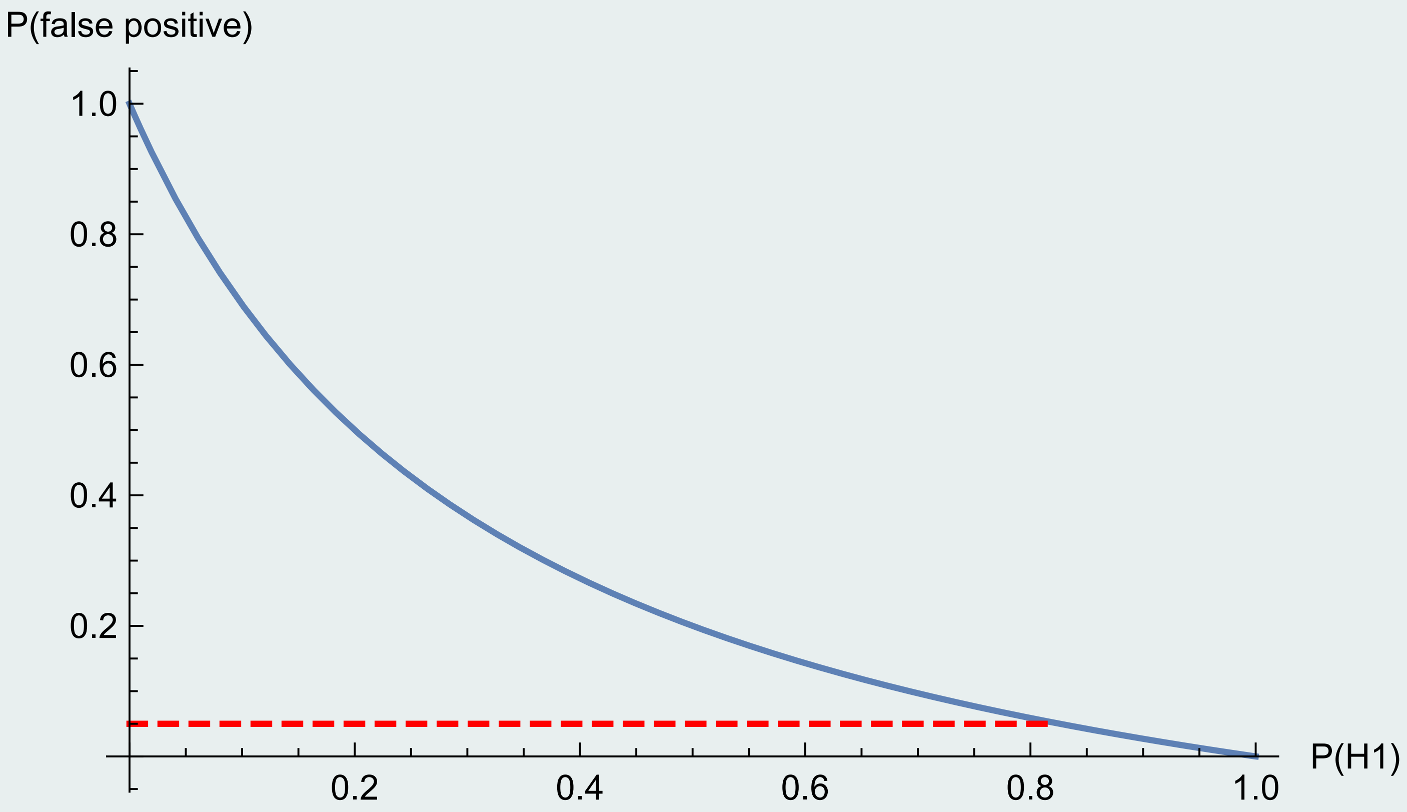
- We run 1000 experiments
- Type I error = 5% (Alpha = .05)
- Type II error = 20% (Power = .8)
- Prevalence P(H1) = 10%

	H0 True	H1 True	Σ
Not reject H0	855	20	875
Reject H0	45	80	125
Σ	900	100	1000

False Discovery Rate = 45 / (45 + 80) = .36

NOT .05!

False Discovery Rate depends on Prevalence



(α = .05, β = .2)

An Exercise in Evidence

Imagine you are in The Netherlands, and your friend want to go to the sauna. You know that in this sauna, there are two types of days.

a) Days for both men & women.

b) Days for women only.

We assume that on days for both men & women, the sauna will have 50% male and 50% female customers. And of course that if it's women-only day, there will be 100% women.

You don't speak Dutch (and nobody in the Netherlands speaks English ☺), but you really want to know what type of day it is because your friend wants to be confident that it's women-only day. So she relies on your scientific help.

So you decide to go to the exit of the sauna and watch who comes out of the sauna, in order to collect evidence about which type of day it is.

Sauna Hypotheses

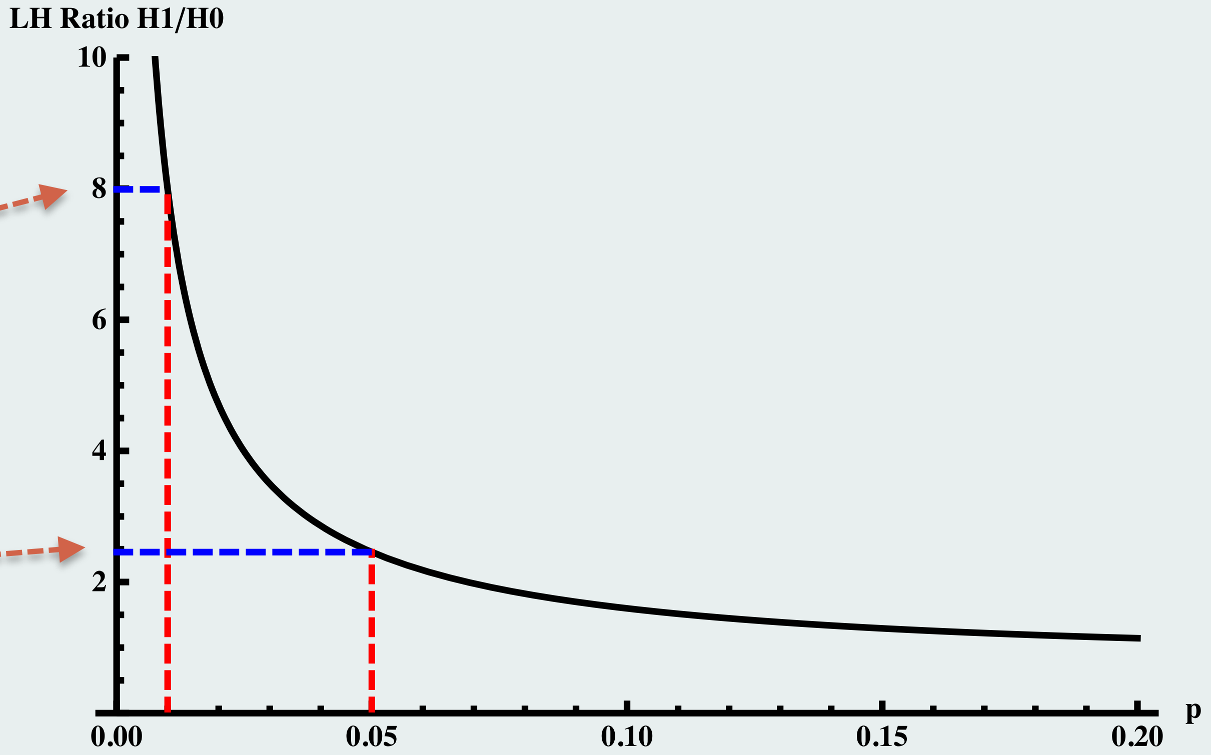
- This way you want to collect evidence for either of the two following hypotheses:
  - H_{mixed}: There is 50% males and 50% females in the sauna today.
  - H_{women}: There are only women in the sauna.
- Note that if a man emerges from the sauna, we are certain that H_{mixed} is true. That's the easy case.
- But if a woman emerges, this could happen under both hypotheses.
- THE QUESTION: after seeing *how many women* emerging from the sauna do you feel really confident that H_{women} is true, i.e., it's a woman-only day?
- You should be confident enough to tell your friend that it's safe to go in.
- There is no correct answer here (scout's honor!). It's an *intuition* exercise.
- Take 5-10 minutes to discuss this in your group. You can report an agreed value or the different values for the members in the group.

Sauna Evidence

- Note that if a woman comes out, the probability that this happens under H_{women} is 1. So P(Woman| H_{women}) = 1. But under H_{mixed} it is ½ because we assume there are 50/50 women and men in there. So the so-called Likelihood Ratio:
- LHR(N) = $\frac{P(N\ women|H_{mixed})}{P(N\ women|H_{women})} = 2^{-N}$
- So 1 woman means LHR = ½, 2 women mean LHR = ¼, etc.
- So the "betting odds" that after 2 women it's women-only day are 4:1
- This is a very natural measure of evidence: the relative probability of the data for both (!) hypotheses.
  - P(H0) : P(H1)

Calibrating p-values

- Sellke, Bayarri & Berger (2001) computed likelihood of H0 : H1 given p-values under for NHST maximally favorable assumptions:



p = .01
3 women

p = .05
slightly more
than 1 woman

To think about

- If you publish something on the basis of p = .05, the probability of H1 being true is *at best* about .7 and the probability of a false positive about .3
- That's an odds ratio of maximally 2.5 : 1
- So more than once every three times you publish something with a p-value close to .05, it's false.
- Those of you who said you wanted to see 5 women (which I think is totally reasonable) would need p-values of 0.0018 to get the same level of evidence.
- This is why many statisticians have argued for at least lowering the standard alpha level from .05 to .005.
- There are also statisticians who are against that, but their arguments are weak.

Summary

- Our standard NHST statistical paradigm centers around the p-value.
  - The p-value is the probability that our data (or more extreme data) occur, assuming H0 is true.
  - So if it is *unlikely* that our data (or more extreme data) would occur when we are *wrong*, we assume we are *right*.
- This is highly confusing.
  - Very many people wrongly assume that the p-value is the probability of H0.
- There are a number of problematic properties of the NHST paradigm.
  - It does not take into account the alternative hypothesis.
  - It does not allow support for the null hypothesis (yeehaw or ?)
  - It does not allow collecting data flexibly and incrementally.
  - It suggests far more evidence than it in fact gives.

Part II: What do we want?

- What we want is the mirror image of what was problematic with NHST.

- We want a statistical method that
  - Is not *massively* confusing
  - Gives us an interpretable measure of evidence
  - Allows us to compare H0 and H1 on equal footing
  - Allows us to collect evidence for H0 as well as H1
  - Allows us to collect data until we have the certainty that we desire, or run out of resources
- The good news: there is such a method.
- The bad news: it is not the standard method, so:
  - It takes some time to learn.
  - There is Resistance from the Status Quo (e.g. reviewers, editors, colleagues).

Part III: Bayesian Statistics

- Bayesian statistics is named after Reverend Thomas Bayes. It is from the mid-18th century, so a lot older than the Frequentist (NHST) statistics by Fisher et al.
- Bayes' friend Richard Price published his work after his death.
- He is the discoverer of *Bayes' Theorem*, which is a (noncontroversial!) formula for conditional probability.
- The person who first applied it systematically to inferential statistics was Pierre Simon Laplace.
- The principal idea of Bayesianism is to have your beliefs be made more accurate by data.

The general idea

- Without going too much into technical detail, Bayesian statistics goes straight to the core of the matter and computes P(Hypothesis|Data).
- As we discussed in part I, computing P(Data|Hypothesis) is usually much easier.
- But with Bayes formula, we can turn the one into the other:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

P(H|D) is the **posterior**, the probability of the Hypothesis AFTER seeing the data D.

P(D|H) is the **likelihood**, the probability of the Data given the Hypothesis

P(H) is the **prior**, the probability of our Hypothesis BEFORE we saw any data.

Belief updating

- So the idea is that we have a belief, which we specify as a probability or probability distribution, and then we obtain data, and use that data to update our belief to get a more accurate belief.
- The belief *before* the updating is the *Prior* belief, the data is the data, and we use these to compute our *Posterior* belief, which is our belief *after* incorporating the new data.
- I will give you a small demo of this procedure using handedness as an example.

[Mathematica Animation Demo of Bayesian Updating]

Waaah the prior!

- Many people are worried about where the prior, the assigned probability of the hypotheses before we see any data, comes from.
- It is our prior belief.
- In fact, it's is not a problem that we incorporate that, it is a solution.
- When Daryl Bem published his (very significant) findings about ESP, did you immediately start believing in ESP?
- If not, why not?
- Probably because you have a strong prior regarding ESP, e.g. because you think it is rather unlikely that the laws of physics are wrong.
- So even if Bem had some evidence, it was not enough to convince you.

The basic idea

- So the formula that Bayesians use is this one:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

- As P(D) is a constant, not dependent on the hypothesis, we can use:
  - Posterior is proportional to Prior times Likelihood

- But there is another form that is much more instructive for us, which is in terms of odds (like the LHR with the sauna).

Posterior Odds = Prior Odds x Likelihood Ratio

The Bayes Factor

$$\underbrace{\frac{P(H_1|D)}{P(H_0|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{Prior odds}} \cdot \underbrace{\frac{P(D|H_1)}{P(D|H_0)}}_{\text{Bayes Factor}}$$

The Bayes Factor is the *relative predictive adequacy* of our competing hypotheses. How much better does one hypothesis predict our data than the other?

Bayes Factor

- The Bayes Factor is in itself a useful measure of evidence. It tells us how much and in which direction (by multiplying) we need to update our belief.
- So if my prior belief that ESP exists is 1:1000000 and someone has evidence that is 100 times more likely under H_{esp} than under $H_{no\ esp}$, my posterior belief will be 100:1 * 1:1000000 = 1:10000
- Given enough experiments like this, I'll eventually end up believing in ESP.
- One advantage of Bayes Factors is that they are not dependent on our Prior beliefs regarding our hypotheses. We can compute and report them, and anyone can use them to update their personal prior.
- But priors can be very important!

Sauna with prior

- Let’s do the sauna exercise again, but now I’m going to give you some additional **prior knowledge**!
- Remember we computed that for every woman emerging from the sauna, the LHR (which his in this case the same as a Bayes Factor) is
  - P(N women come out|H_{mixed}) / P(N women come out|H_{women}) = 2^{-N}
- Now suppose I would give you the additional information that there is only one women-day a week.
  - Would you now require more women to come out before you believe it is women-only day?
  - How many more women?
  - The Prior probability of it being women-day is now not 1:1 (as we implicitly assumed) but 1:6
  - So it should be between 2 (factor 4) and 3 (factor 8) extra women to reach the same level of confidence. Does that match your intuition?

The things to remember

- We do not have the time to go deep into the math here. You just need to remember the following:
  - Bayesian statistics directly compute P(Hypothesis|Data).
  - For this we need to have a specification of our prior knowledge/belief.
    - But we can make this very vague if we don’t know much.
  - We can also ignore the issue of the “model prior”, and compute a *Bayes Factor*, which gives us the relative predictive power of our competing hypotheses.
  - We can report Bayes factors as they are, as they are readily interpretable and non-confusing.
    - E.g., “the data are 8 times more likely under H0 than under H1.”

Part IV: Bad, Good and Better News

- The bad news:
  - Bayes factor and posteriors are very hard to compute, even for mathematicians.
  - Mostly, there are no closed form solutions.
- The good news:
  - Modern developments in numerical estimation + fast computers (i.e., your laptop) have made it possible to compute Bayesian versions of most tests that you know from your "normal" statistics.
  - This includes nonparametric statistics, Chi Square, t-test, ANOVA, Repeated Measures ANOVA, and Multiple Regression.
- The even better news:
  - There is fantastic and free (as in free beer) software for this.
  - The best way to start is with JASP: jasp-stat.org

Advantages of Bayesian Statistics (1)

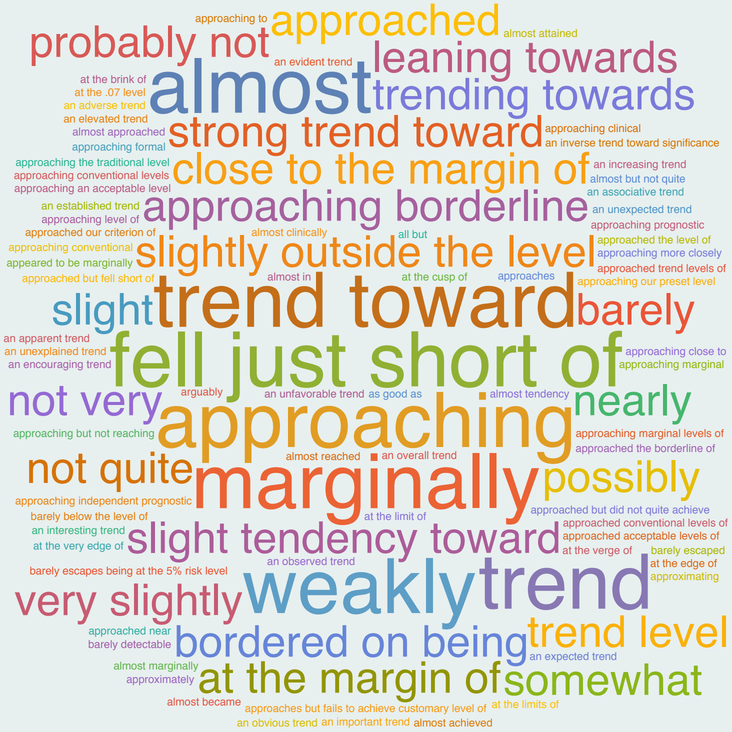
- The reported values are easy to interpret. They are either straight probabilities, or odds ratios.
- It treats H0 and H1 as equals, and takes both into account:
  - Snag: you have to actually specify H1, but that is Good For You™
- It is perfectly possible to collect evidence for H0
  - So the conclusion that two groups are equal on some trait/property can be supported by statistical evidence.
- There are three possible outcomes:
  - Evidence favors H1
  - Evidence favors H0
  - We don't know enough yet to distinguish between H1 and H0.

Advantages of Bayesian Statistics (2)

- There is no problem with optional stopping: we can collect data in any amount and order that we want, until we are satisfied with our evidence or run out of time/money/patience.
  - This advantage cannot be emphasized enough!
  - There are many cases in which we cannot specify a sampling plan ahead of time:
    - Meta-analyses (new studies emerge)
    - Data coming in through natural processes (astronomy, climate science, etc.)
    - Longitudinal studies that run shorter or longer than planned

Easy to report/interpret

- There is no need for verbal gymnastics like
  - “the effect was trying to trend towards marginal significance (p = .073)”

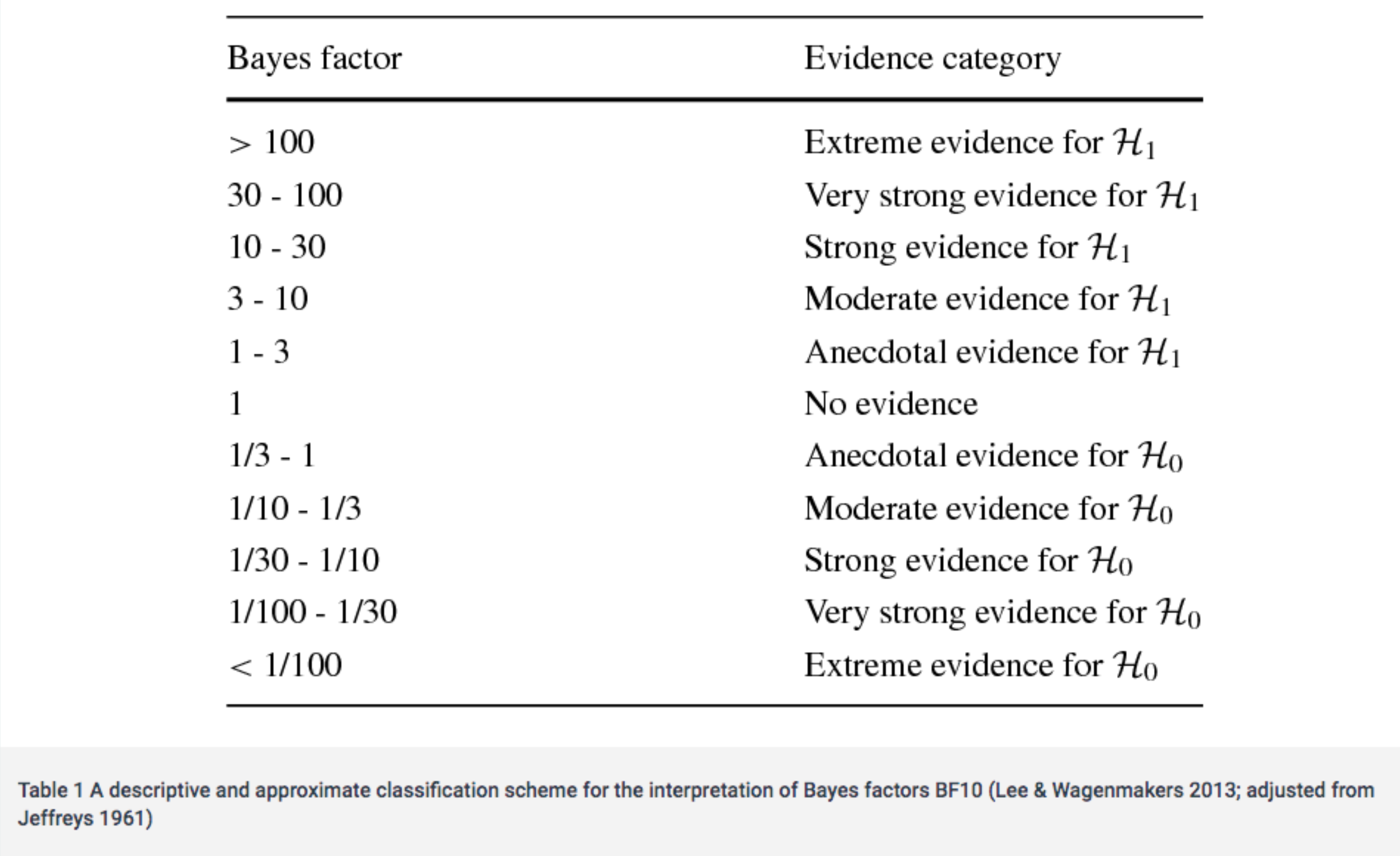


(made from corpus by Matthew Hankins)

There is no cut-off necessary

- One can report a Bayesian Confidence Interval on the parameter of interest:
  - "the difference was with 95% probability in the interval [10.1-13.9 cm]
  - This is NOT possible with "normal" Confidence Intervals, they mean something different.
- Or one can report a Bayes Factor:
  - "The data were 12 times as likely under H1 than under H0."
    - BF10 = 12
  - "The data were 3 times as likely under H0 than under H1."
    - BF01 = 3
  - "The data were about equally likely under H0 as under H1"
    - BF01 ≈ 1
- But if you want verbal labels, there are some, e.g. the ones provided by Jeffreys:

Jeffreys’ classification

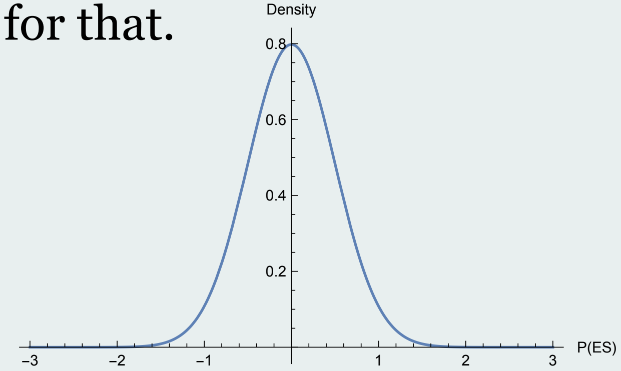


Bayes factor	Evidence category
> 100	Extreme evidence for $\mathcal{H}_1$
30 - 100	Very strong evidence for $\mathcal{H}_1$
10 - 30	Strong evidence for $\mathcal{H}_1$
3 - 10	Moderate evidence for $\mathcal{H}_1$
1 - 3	Anecdotal evidence for $\mathcal{H}_1$
1	No evidence
1/3 - 1	Anecdotal evidence for $\mathcal{H}_0$
1/10 - 1/3	Moderate evidence for $\mathcal{H}_0$
1/30 - 1/10	Strong evidence for $\mathcal{H}_0$
1/100 - 1/30	Very strong evidence for $\mathcal{H}_0$
$< 1/100$	Extreme evidence for $\mathcal{H}_0$

Table 1 A descriptive and approximate classification scheme for the interpretation of Bayes factors BF10 (Lee & Wagenmakers 2013; adjusted from Jeffreys 1961)

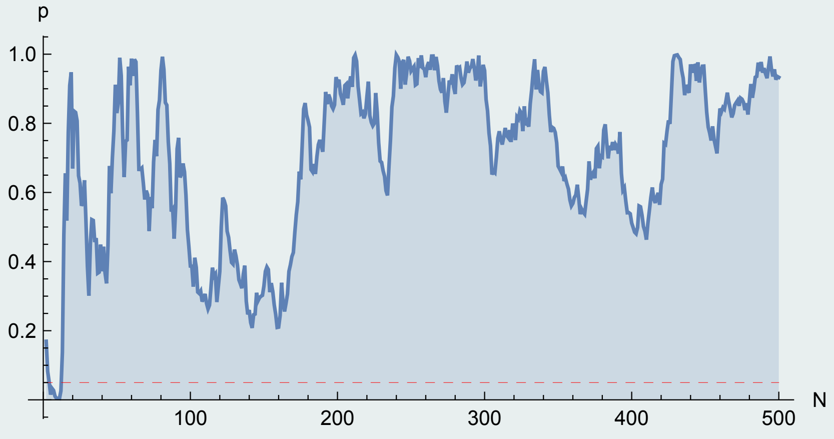
Treating H0 and H1 as equals

- There is no special status for the null hypothesis. Evidence for H0 relative to some (specified) H1 is exactly the same as evidence for H1.
  - Bad news: You'll have to specify a H1 (which is good for you!)
  - Good news: You can do that with distributions, meaning you can express your uncertainty. You don't need to commit to an exact point value for H1.
- E.g.:
  - H0: my effect size (mean / stddev) is 0
  - H1: my effect size is not 0, but a Normal distribution with mean 0 and standard deviation 0.2
  - Note: Standard software has reasonable defaults for that.

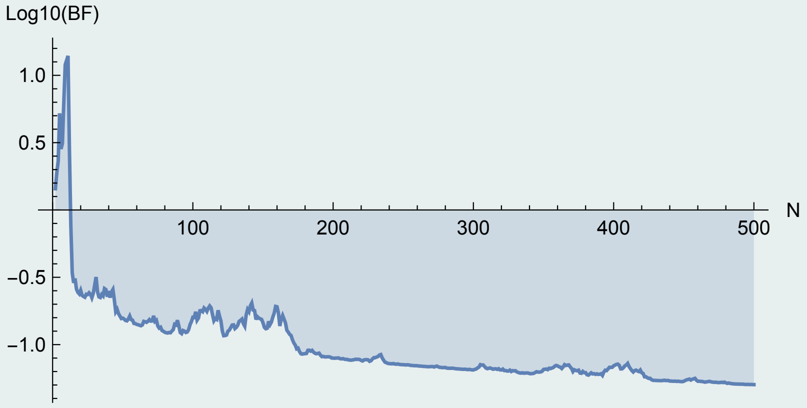


You can have converging evidence for H0

- Imagine we have a steady stream of data about an effect that is actually a null effect. Look what happens if we plot a p-value over the time the data comes in (which is illegal, but we'll do it anyway) and a Bayes Factor (which is perfectly fine).
  - Note: it's the same data.
- This is why we can't use the p-value to claim evidence for H0, but we can with the Bayes Factor.



NHST p-value



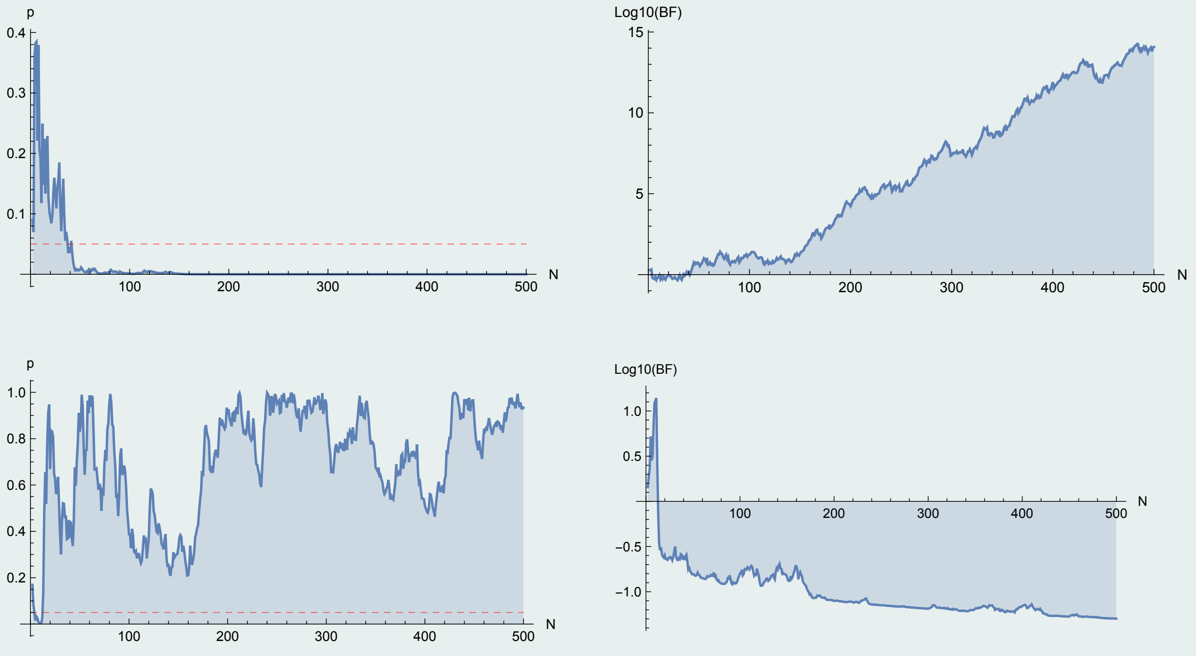
Bayes Factor (Log_{10})

Bayesian stats have *three* possible outcomes

- We have *three* possible outcomes with a Bayes Factor
  - Relative evidence for H1 over H0
    - Bayes Factor BF10 is large
  - Relative evidence for H0 over H1
    - Bayes Factor BF01 is large, or BF10 small (equivalent)
  - We are uncertain
    - Bayes factor hovers around 1
    - We need to get more data.

Bayes Factor converges to H0 or H1

- Both the NHST p-value and the Bayes Factor can detect a real effect.
- It's the null effect that NHST can't do anything useful with



Effect

No effect

P-value

Bayes Factor

Optional Stopping

- With Bayesian statistics, it's perfectly OK to have a flexible sampling plan, e.g.
  - We will collect data from at least 10 participants
  - After that we will collect data, until either:
    - We have conclusive evidence for either H0 or H1 (BF01 or BF01 > 10)
    - We have used up our budget at 100 participants
  - Isn't that great? It's so good that there are still researchers who think it's too good to be true.
- This means that if you work with patients in medical tests, you don't need to make more patients suffer than strictly necessary.
  - There are some arcane procedures for this with NHST (with "alpha budgets" and stuff) but they are complicated to use and report. And you still have all the other disadvantages of NHST…

Pros and Cons

- There are in fact many more problems with NHST that are avoided by Bayesian statistics. See Wagenmakers 2007 for details.
- However, it is fair to say that there is one major disadvantage:
  - It is not very well known, and reviewers are often difficult about it.
- My tip: just do both NHST and Bayes. The reader can choose whether they prefer conclusions like:
  - ”There was no significant effect of word class on RTs (p = .17) so uh, maybe we should collect more data, because our power was too low. Or maybe there was no effect. We just don’t know.”
- Or, alternatively:
  - “The hypothesis that word class has an effect on RTs is only 1.2 times as likely as the null hypothesis that it has no effect, so we clearly need more data.

Things Bayesian Statistics Do **Not** Fix

- As the opponents of Bayesian statistics never tire of pointing out, using Bayesian statistics is **not** a cure against:
  - Bad experimental designs (e.g. confounds)
  - Publication bias (journals only publishing positive results)
  - Violating test assumptions
  - Using the tests in the wrong way
  - Bad or badly formulated theories
  - Fraud
  - Bad weather on Sundays

JASP – SPSS the way God intended

- Statistics program with Graphical UI developed by a team of programmers lead by the tireless and fearless E.J. Wagenmakers.
- Very user-friendly
- Free
- Available for Windows, Mac, and Linux
- Has both “normal” (Frequentist) and Bayesian statistics on board for the same tests, so you can compare the two with only a few clicks.
- Contains beautiful, paste-able (or LaTeX) APA tables and graphics plots in vector format (so no ragged lines in .JPG or .TIFF files).

Literature pointers

- Guidelines for Bayesian testing by the makers of JASP:
  - https://psyarxiv.com/yqxfr/download
- Gentle introductions to Bayesian statistics and some of their advantages in arguing for null effects:
  - https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00781/full
- Article on how to formulate a good prior for your theory
  - https://psyarxiv.com/yqaj4/
- The entire theoretical background for this workshop in one paper:
  - https://www.ejwagenmakers.com/2007/pValueProblems.pdf
- Longer list of advantages of Bayesian statistics over NHST:
  - https://link.springer.com/content/pdf/10.3758%2Fs13423-017-1343-3.pdf
- Article on which articles you need to read if you want to study Bayesian statistics further:
  - https://link.springer.com/article/10.3758%2Fs13423-017-1317-5
- Article on the abuse of power (in the statistical sense):
  - https://www.tandfonline.com/doi/abs/10.1198/000313001300339897

The End

Thank you for your participation!