Prelude

Practical problems with p-values

Problem 1: Conceptual Confusion

What is a p-value?

"The probability of obtaining a test statistic at least as extreme as the one you observed, given that the null hypothesis is true and the intention with which the data were collected is known."

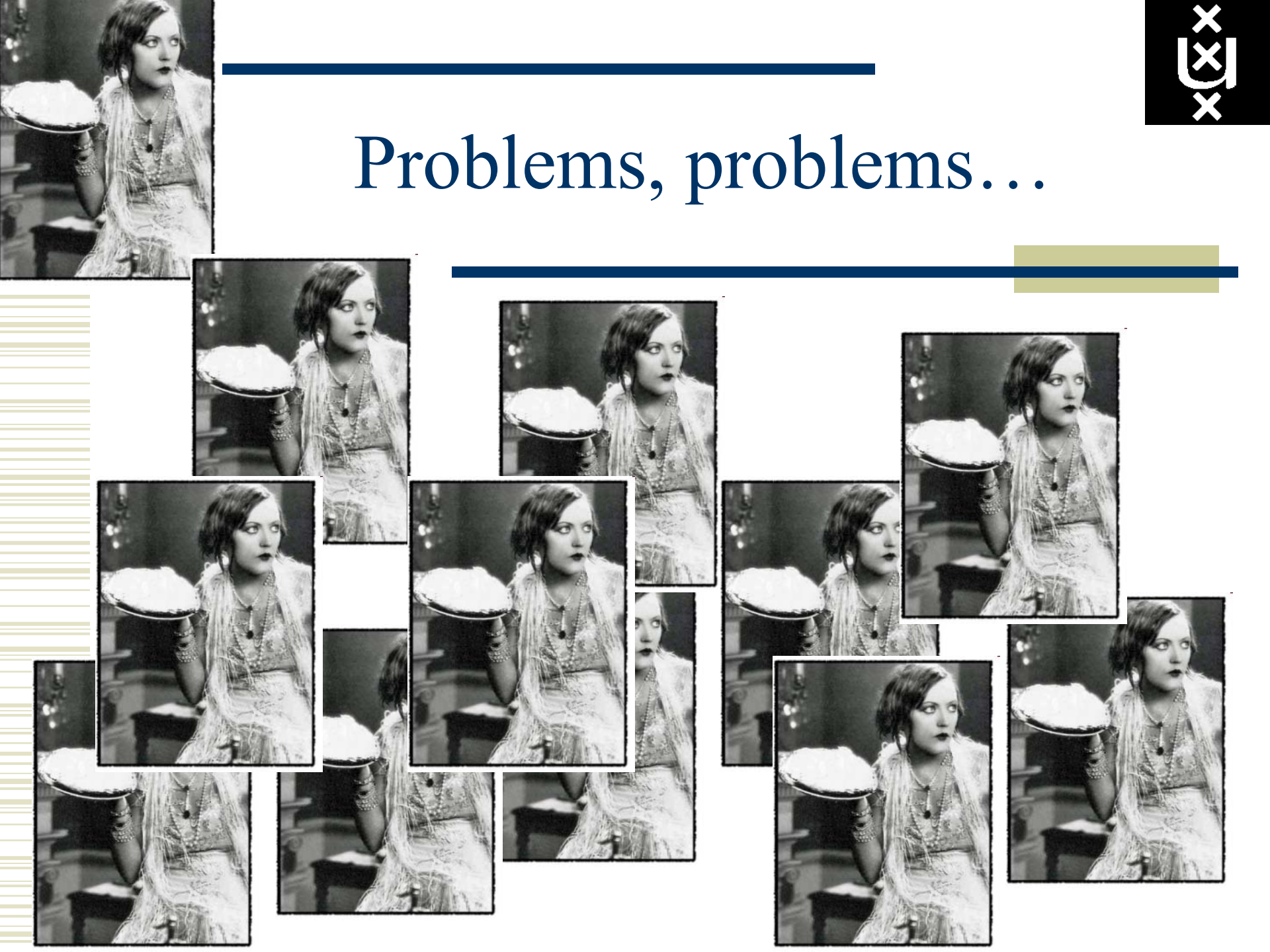
Problem 2: What if p = .001?

- You would like to conclude that H0 is probably false.
- But what about H1? If H1 represents a very small effect, the observed data are also very unlikely under H1.
- And what if H1 is a priori unlikely?

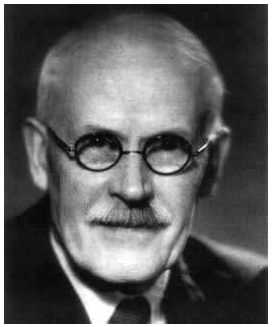
Problem 3: What if p = .35?

- You would like to conclude that H0 is probably true.
- But p-values cannot be used to quantify evidence in favor of H0 – perhaps you did not have enough observations.

Problems, problems…



Bayesian Hypothesis Tests in Practice: The t-Test and a Hierarchical Extension



Eric-Jan Wagenmakers

UNIVERSITEIT VAN AMSTERDAM

Outline Part I

- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions

What is Bayesian Statistics?

"Common sense expressed in numbers"

What is Bayesian Statistics?

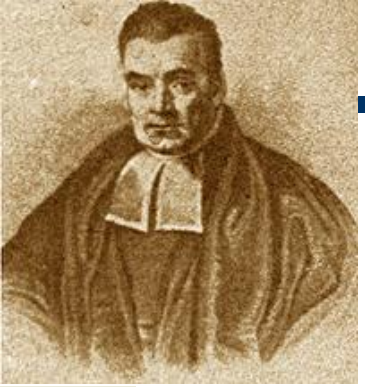
"What you think that classical statistics is"

What is Bayesian Statistics?

"The only good statistics"

[For more background see

Lindley, D. V. (2000). The philosophy of statistics. The Statistician, 49, 293-337.]

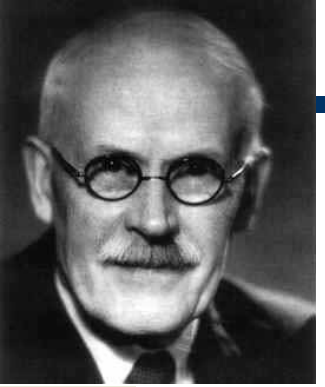


Bayesian Inference in a Nutshell

- In Bayesian inference, uncertainty or degree of belief is quantified by probability.
- Prior beliefs are updated by means of the data to yield posterior beliefs.

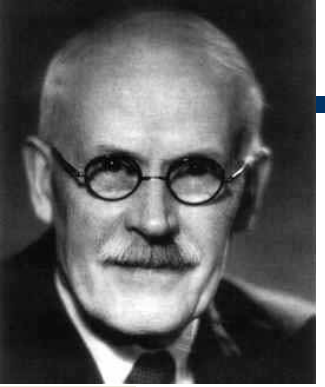
Outline

- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions



Bayesian Model Selection

- Suppose we have two models, M_1 and M_2.
- After seeing the data, which one is preferable?
  - The one that has the highest posterior probability!
  - Compare $P(M_1 \mid D)$ to $P(M_2 \mid D)$.



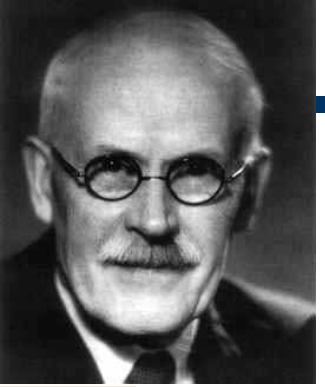
Bayesian Model Selection

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1)}{P(D \mid M_2)} \times \frac{P(M_1)}{P(M_2)}$$

Posterior odds

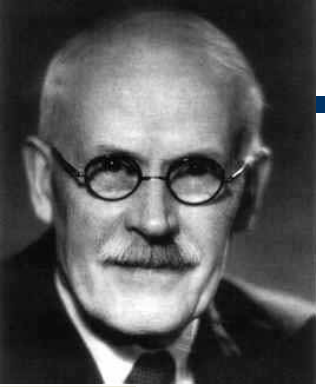
"Bayes factor"

Prior odds



The Bayes Factor

- Is the change from prior to posterior odds brought about by the data.
- Quantifies the evidence for one model versus the other provided by the data.
- If $BF_{12} = 3$, the data are three times more likely to have occurred under M_1 than under M_2.



Guidelines for Interpretation of the Bayes Factor

BF	Evidence
1 – 3	Anecdotal
3 – 10	Substantial
10 – 30	Strong
30 – 100	Very strong
>100	Decisive

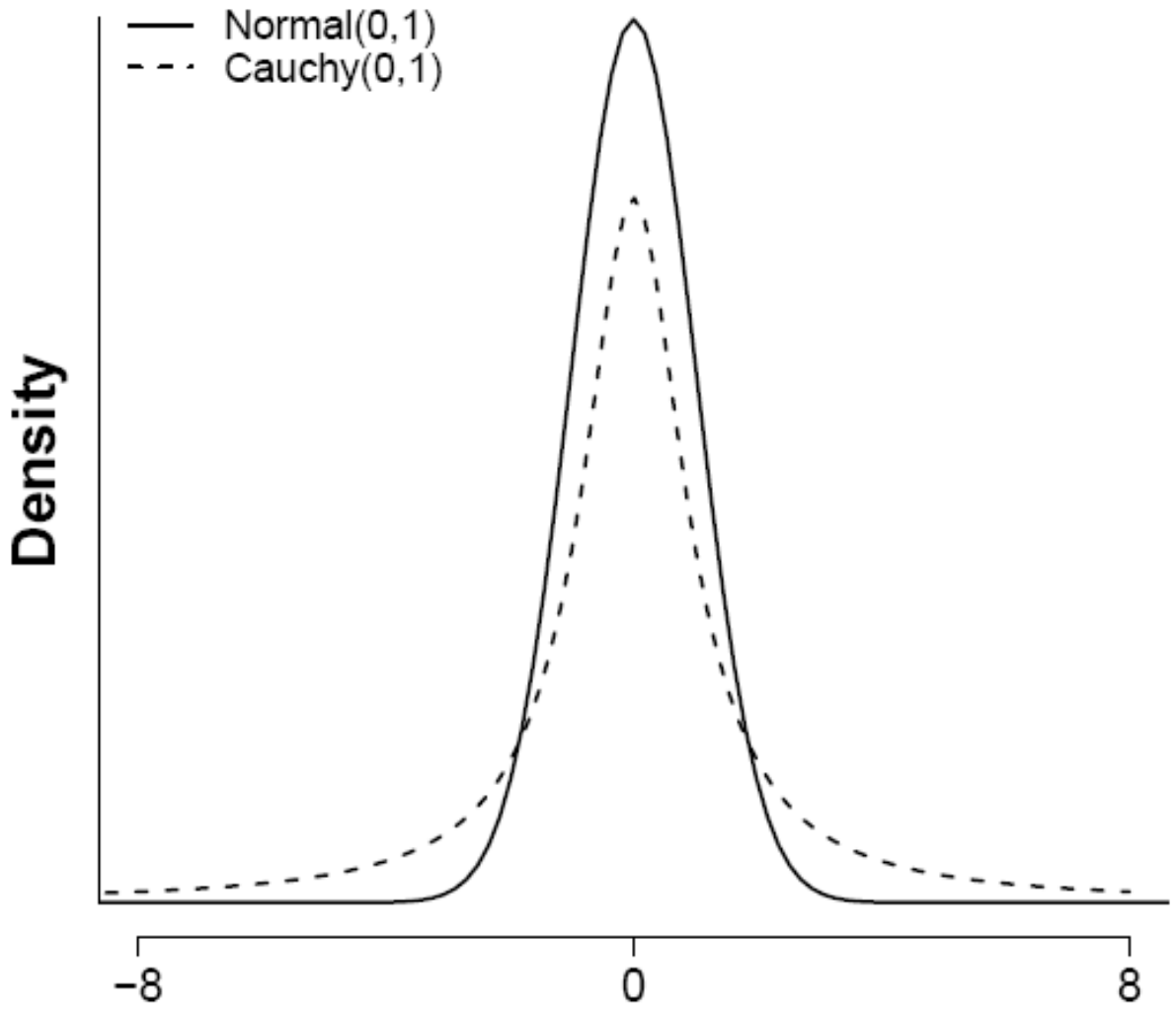
Outline

- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions

A Default Bayesian t-Test

- Based on earlier work, Rouder and colleagues proposed a default or objective Bayesian t-test (sort of an improved BIC).
- Focus is on effect size $\delta = \mu/\sigma$.
- H0: $\delta = 0$
- H1: $\delta \sim$ Cauchy

Cauchy vs. Normal



Outline

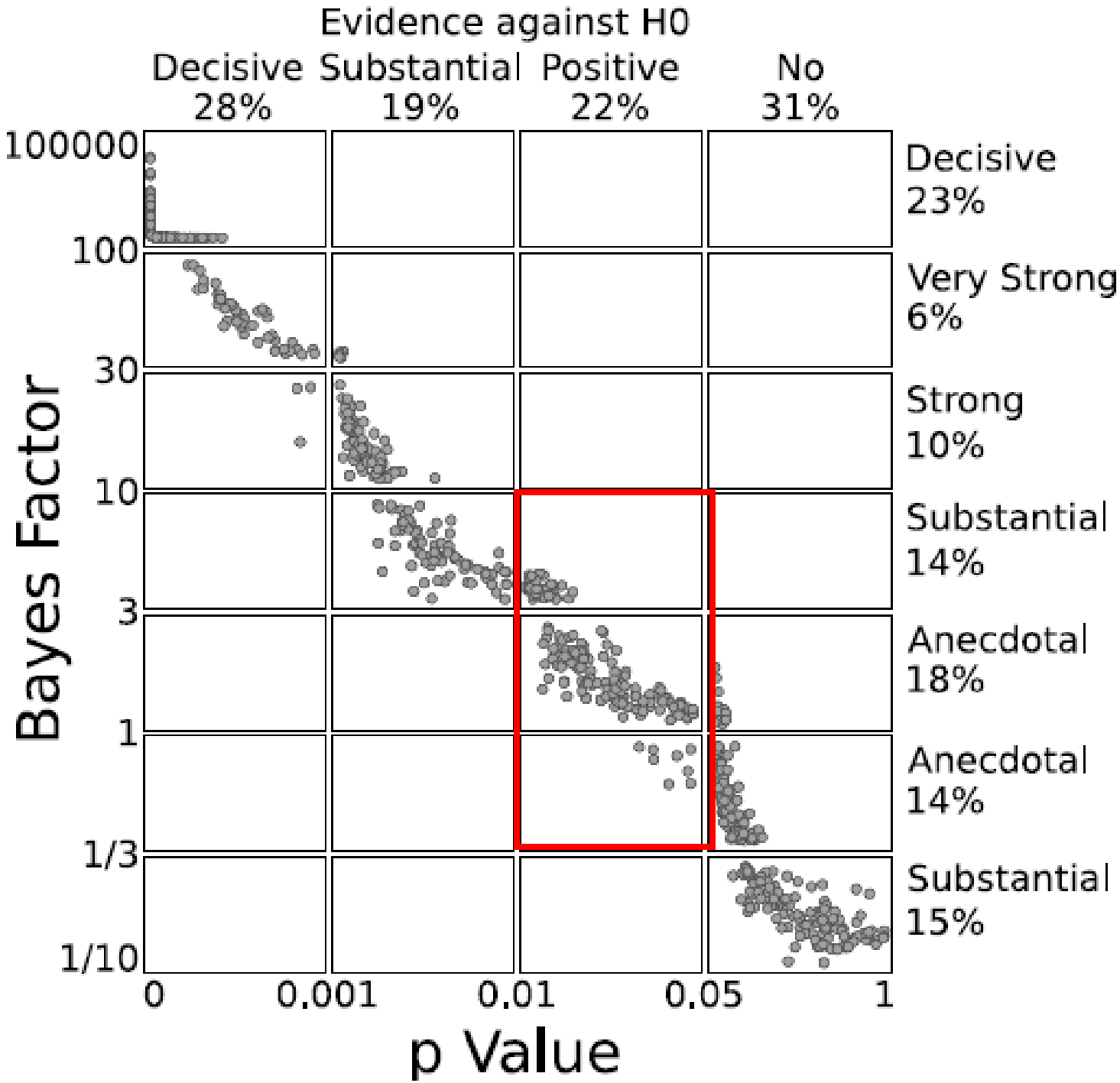
- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions

Empirical Comparison
(Wetzels et al., in press)

- Does the Bayesian t-test (Rouder et al., 2009, PBR) yield the same conclusions as the traditional t-test?
- “We” collected all t-tests reported in the 2007 issues of Psychonomic Bulletin & Review and JEP:LMC.

Empirical Comparison
(Wetzels et al., in press)

- In 252 articles, spanning 2394 pages, “we” found 855 t-tests.
- This translates to an average of one t-test for every 2.8 pages, or about 3.4 t-tests per article.
- For each t-test, we computed the p-value and the Bayes factor…


Evidence against H0
Decisive
28%
Substantial
19%
Positive
22%
No
31%
100000
100
30
10
3
1
1/3
1/10
Bayes Factor
0
0.001
0.01
0.05
1
p Value
Decisive
23%
Very Strong
6%
Strong
10%
Substantial
14%
Anecdotal
18%
Anecdotal
14%
Substantial
15%

Empirical Comparison

- Bayes factors and p-values agree on the direction of the effect: p-values < .05 yield evidence against H0, and p-values > .05 yield evidence against H1.
- Bayes factors and p-values often disagree on the strength of the effect: 70% of p-values in the .01-.05 interval yield evidence that is only “anecdotal”.

Outline

- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions

Feeling the Future

- Dr. Bem published a paper in the *Journal of Personality and Social Psychology* (JPSP).
- Bem reported nine experiments with over 1000 participants.
- Conclusion: people can look into the future, predicting with higher-than-chance accuracy where a picture is going to appear (left or right).

The Colbert Report

Thursday, January 27, 2011

Time-Traveling Porn - Daryl Bem

Stephen uses the power of time-traveling porn to predict the 2012 presidential election, and Daryl Bem discusses his theory of extrasensory pornception. (07:46)

Feeling the Future

- Some things went wrong with this research. Most importantly, the results are based in part on a fishing expedition. Check my website if you want to know the details.
- However, how about the statistics? In eight experiments, Bem computes a t-test and comes up with one-sided p-values < .05.

Feeling the Future

Exp	df	$\|t\|$	p
1	99	2.51	0.01
2	149	2.39	0.009
3	96	2.55	0.006
4	98	2.03	0.023
5	99	2.23	0.014
6	149	1.80	0.037
6	149	1.74	0.041
7	199	1.31	0.096
8	99	1.92	0.029
9	49	2.96	0.002

Feeling the Future
(Wagenmakers et al., 2011, JPSP)

Exp	df	$\|t\|$	p	BF_{01}	Evidence category (in favor of H)
1	99	2.51	0.01	0.61	Anecdotal (H_1)
2	149	2.39	0.009	0.95	Anecdotal (H_1)
3	96	2.55	0.006	0.55	Anecdotal (H_1)
4	98	2.03	0.023	1.71	Anecdotal (H_0)
5	99	2.23	0.014	1.14	Anecdotal (H_0)
6	149	1.80	0.037	3.14	Substantial (H_0)
6	149	1.74	0.041	3.49	Substantial (H_0)
7	199	1.31	0.096	7.61	Substantial (H_0)
8	99	1.92	0.029	2.11	Anecdotal (H_0)
9	49	2.96	0.002	0.17	Substantial (H_1)

The New York Times

Science

Journal's Paper on ESP Expected to Prompt Outrage

By BENEDICT CAREY
Published: January 5, 2011

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

Enlarge This Image

Heather Ainsworth for The New York Times

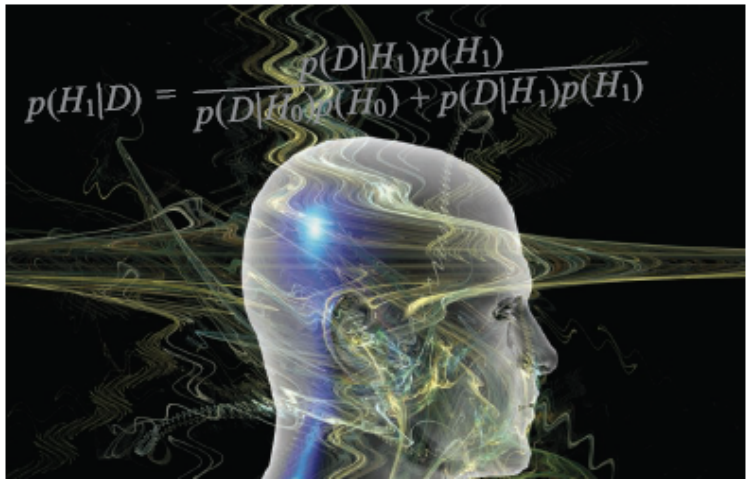
Work by Daryl J. Bem on extrasensory perception is scheduled to be published this year.

The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the paper, to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

STATISTICS

ESP Paper Rekindles Discussion About Statistics

The decision by a top psychology journal to publish a paper on extrasensory perception (ESP) has sparked a lively discussion on blogs and in the mainstream media. The paper's author, Daryl Bem, a respected social psychologist and professor emeritus at Cornell University, argues that the results of nine experiments he conducted with more than 1000 college students provide statistically significant evidence of an ability to predict future events. Not surprisingly, word that the paper will appear in an upcoming issue of the *Journal of Personality and Social Psychology* (*JPSP*) has provoked outrage from pseudoscience debunkers and counteraccusations



prompted him to review the research on the topic. Impressed by what he saw as a number of strong findings, he began experiments of his own.

In the new study, Bem says his goal was to design simple experiments, many of them based on common lab tests used by psychologists. In one experiment, subjects saw two curtains on a computer screen and had to guess which of the two had an erotic picture behind it. In this experiment and seven others, Bem found statistically significant evidence suggesting his subjects had unconscious knowledge of future events. For example, subjects picked the correct curtain about 53% of the time, and a standard statistical test (a *t* test) indicated a p-value of less than .01, well below the .05 threshold typically used to determine statistical significance. Bem says he intentionally stuck to familiar statistical methods: "If you use fancy statistics, people think you're hiding something in the weeds."

that p-values in the .001 to .01 range reflect a true effect only 86% to 92% of the time. The problem is more acute for larger samples, which can give rise to a small p-value even when the effect is negligible for practical purposes, Raftery says.

He and others champion a different approach based on so-called Bayesian statistics. Based on a theory developed by Thomas Bayes, an 18th century English minister, these methods are designed to determine the probability that a hypothesis is true given the data a researcher has observed. It's a more intuitive approach that's conceptually more in line with the goals of scientists, say its advocates. Also, unlike the standard approach, which assumes that each new experiment takes place in a vacuum, Bayesian statistics takes prior knowledge into consideration.

That's important when the effect in question is something like ESP, for which prior knowledge of physics and biology suggests no possible mechanism, says Eric-Jan Wagenmakers, a mathematical psychologist at the University of Amsterdam in the Netherlands. He says the bar should be higher for such "extraordinary claims." Wagenmakers and three colleagues have written a critique of Bem's paper that will appear in the same issue of *JPSP*. It includes a Bayesian reanalysis of Bem's data that concludes that, if any-

Outline

- Bayesian basics
- Bayesian hypothesis tests
- A default Bayesian t-test
- A comparison using 855 published t-tests
- A case study: Feeling the future?
- Interim conclusions

Interim Conclusions

- A Bayesian hypothesis test may give very different results than a classical test. In general, the Bayesian test is not as easily impressed.
- Bayesian hypothesis tests are intuitive and allow you to quantify evidence for H0 versus H1.

Outline Part II

- A default Bayesian t-test
- A hierarchical extension
- Why it can be Bayesian not to be Bayesian

A Default Bayesian t-Test

- *N* participants
- Each participant *i* gets condition A and condition B, and y_i quantifies his or her condition effect.
- The one-sample t-test then assumes

$$y_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2),\ i = 1, \ldots, N.$$

where H0: $\mu = 0$.

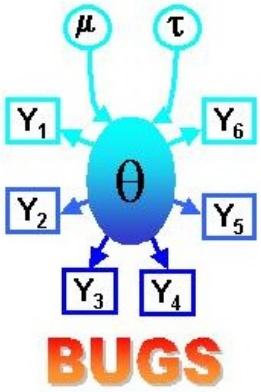
A Default Bayesian t-Test

- Based on earlier work, Rouder and colleagues proposed a default or objective Bayesian t-test.
- Focus is on effect size $\delta = \mu/\sigma$.
- H0: $\delta = 0$
- H1: $\delta \sim$ Cauchy

A Default Bayesian t-Test

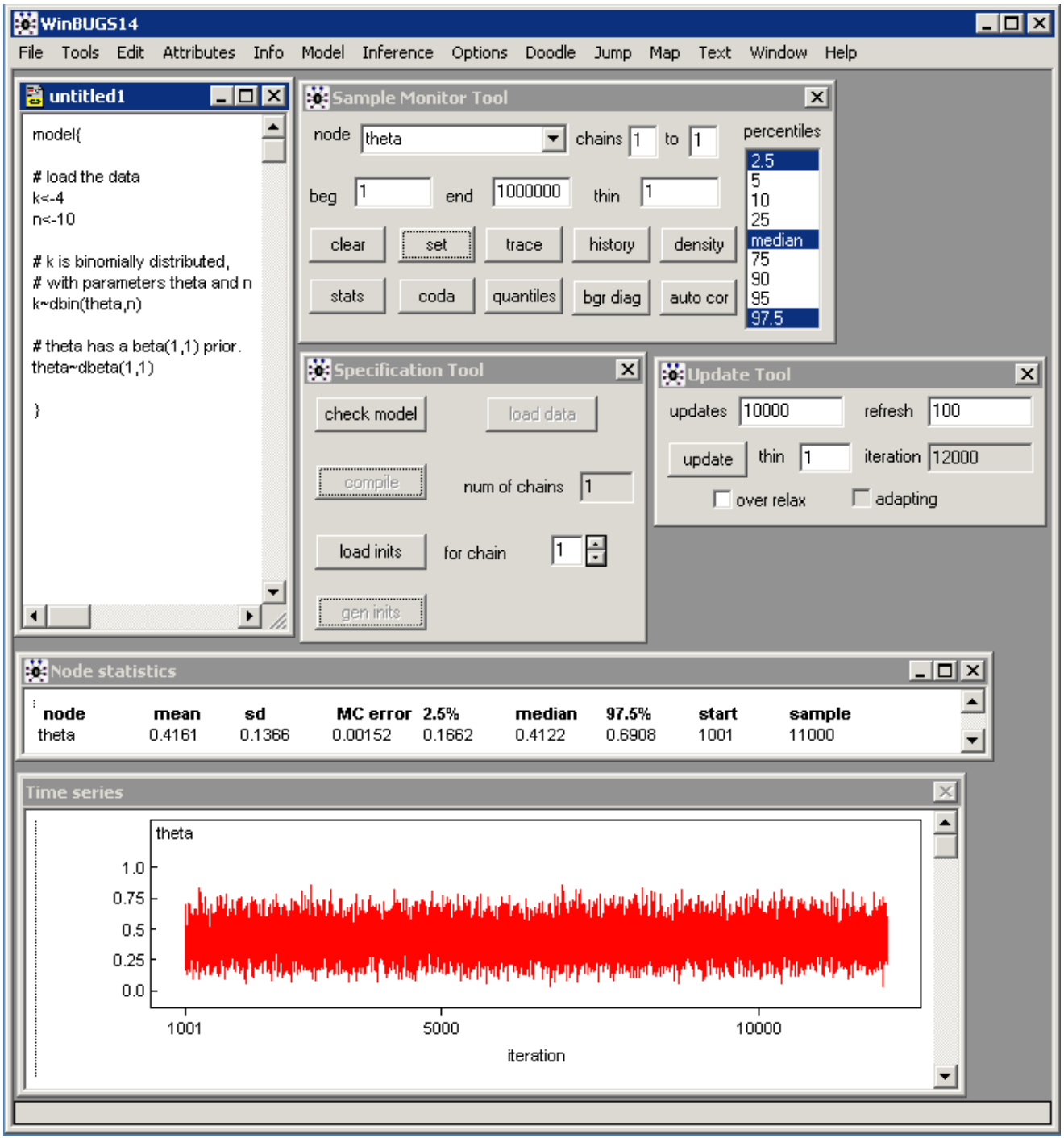
- The resulting Bayes factor is

$$B_{01} = \frac{\left(1+\frac{t^2}{v}\right)^{-(v+1)/2}}{\int_0^\infty (1+Ng)^{-1/2}\left(1+\frac{t^2}{(1+Ng)v}\right)^{-(v+1)/2}(2\pi)^{-1/2}g^{-3/2}e^{-1/(2g)}dg}$$



A Default Bayesian t-Test

- But the test may also be implemented easily in WinBUGS, allowing for flexible extensions (e.g., Wetzels et al., 2009).


WinBUGS14
File Tools Edit Attributes Info Model Inference Options Doodle Jump Map Text Window Help
untitled1
model{
# load the data
k<-4
n<-10
# k is binomially distributed,
# with parameters theta and n
k~dbin(theta,n)
theta has a beta(1,1) prior.
theta~dbeta(1,1)
}
Sample Monitor Tool
node theta chains 1 to 1
beg 1 end 1000000 thin 1
clear set trace history density
stats coda quantiles bgr diag auto cor
percentiles
2.5
5
10
25
median
75
90
95
97.5
Specification Tool
check model load data
compile num of chains 1
load inits for chain 1
gen inits
Update Tool
updates 10000 refresh 100
update thin 1 iteration 12000
over relax adapting
Node statistics
node mean sd MC error 2.5% median 97.5% start sample
theta 0.4161 0.1366 0.00152 0.1662 0.4122 0.6908 1001 11000
Time series
theta
1.0
0.75
0.5
0.25
0.0
1001
5000
10000
iteration

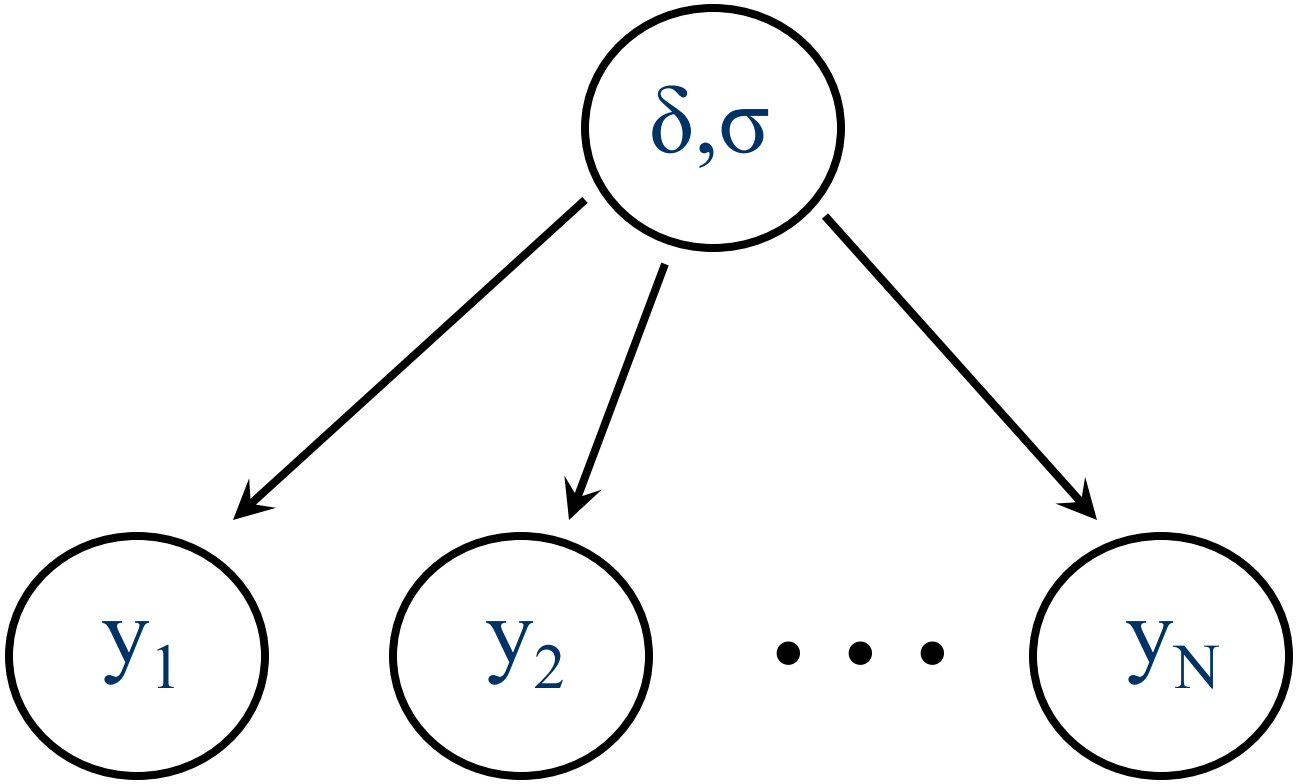
Bayesian Modeling for Cognitive Science
A WinBUGS Workshop

bayescourse@gmail.com

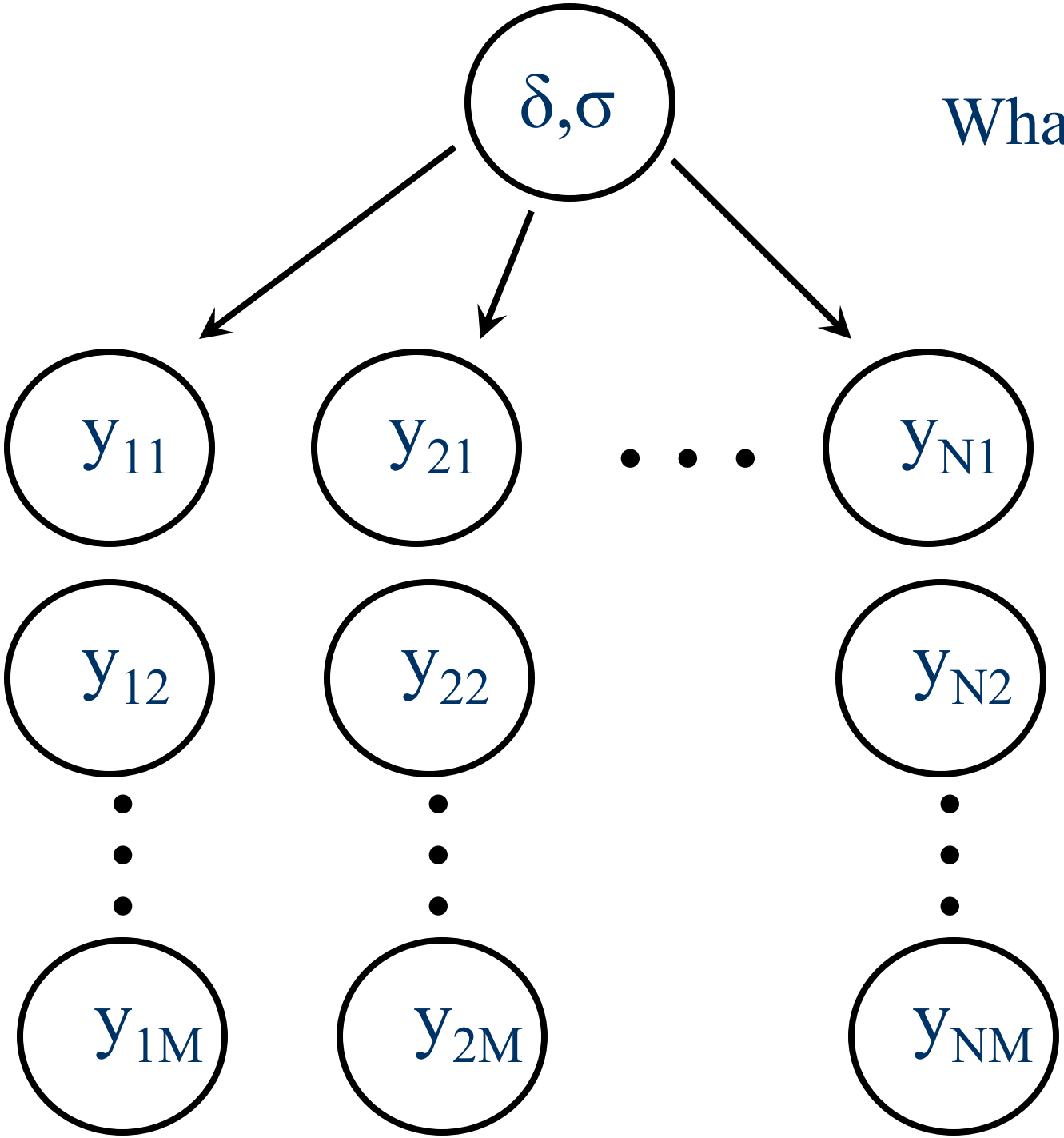
August 22 - August 26, 2011
Amsterdam

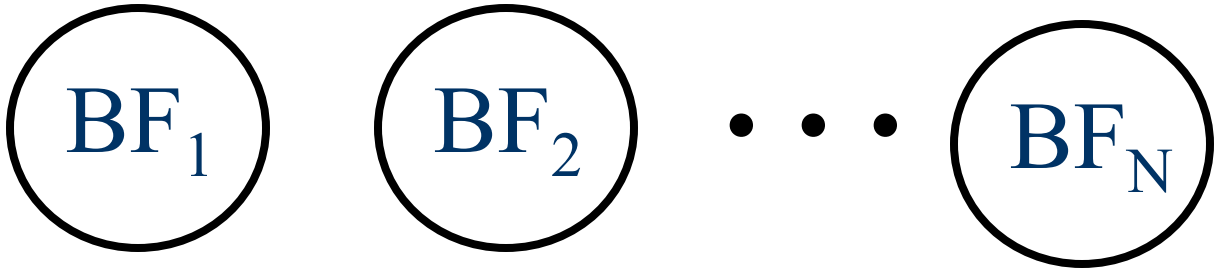
Outline Part II

- A default Bayesian t-test
- A hierarchical extension
- Why it can be Bayesian not to be Bayesian



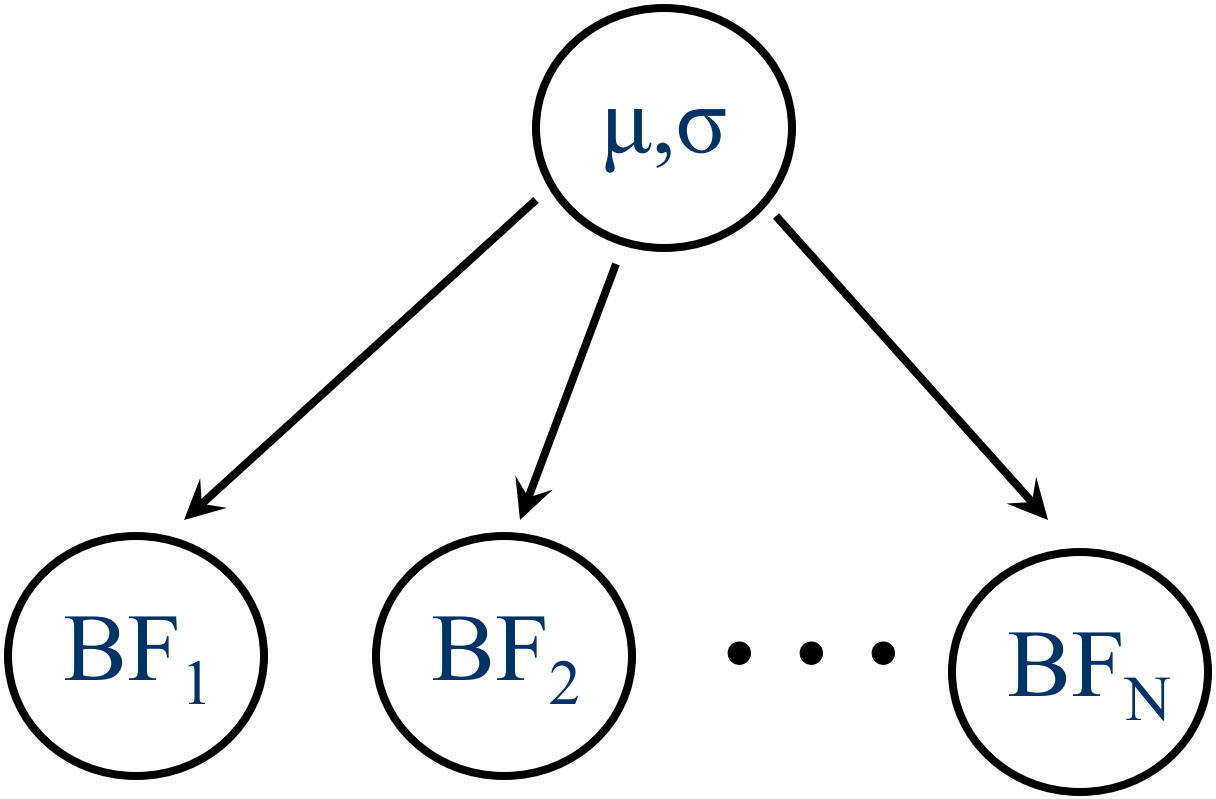
This is the standard application of the t-test


δ,σ
What now?
y11
y21
yN1
y12
y22
yN2
y1M
y2M
yNM

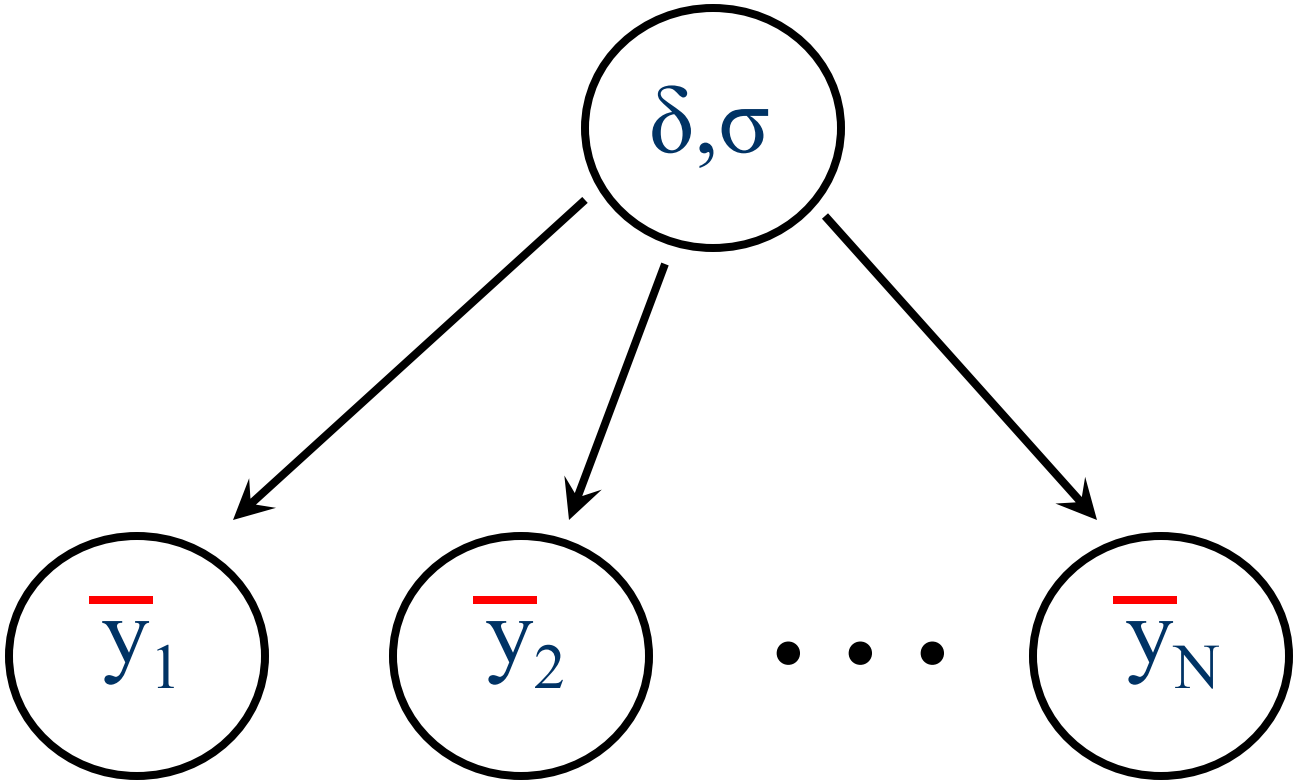


$$= BF_1 \times BF_2 \times \ldots \times BF_N$$

Problem: this analysis ignores the fact that subjects differ from each other.

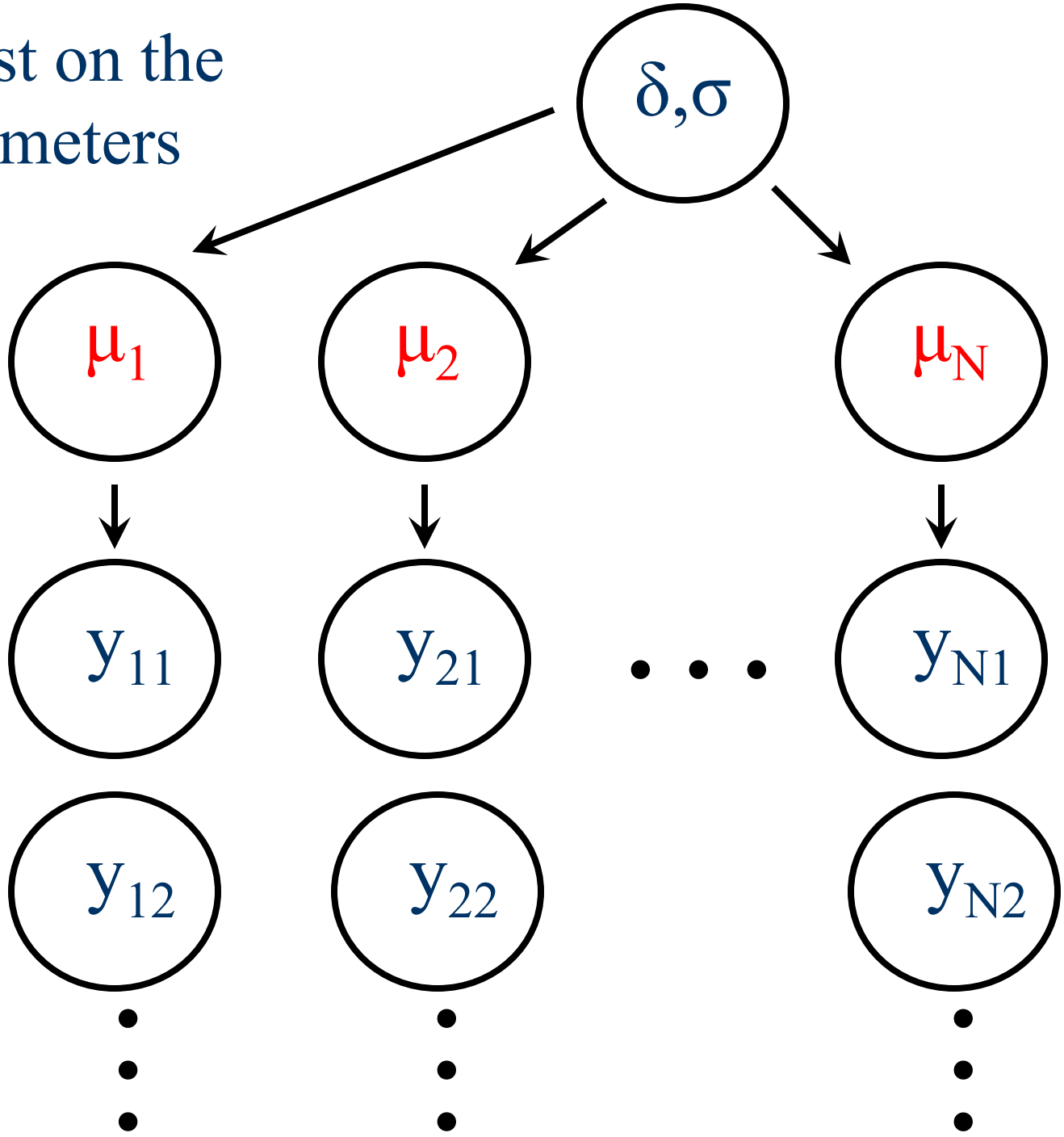


Problem: computes group-average BF, not the total evidence.



Problem: ignores remaining uncertainty in the mean.

A t-test on the parameters



A Critique

- What if the effect is really fixed?
- Or what if you do not know whether your effect is fixed or random?

A Response

- For human participants, effects are almost always random.
- In case of doubt, calculate BF for both fixed effects and random effects model, and weigh these with the posterior probabilities of the fixed effects model being true.
- Note the tension between testing a few participants for a very long time, and testing very many participants for a short time.

Outline Part II

- A default Bayesian t-test
- A hierarchical extension
- Why it can be Bayesian not to be Bayesian

Can it be Bayesian not to be Bayesian?

- Bayesians make decisions that maximize expected utility.
- Confronted with a choice between computing and reporting p-values vs. Bayes factors, what action has the highest expected utility?

Utilities of Bayes factors

- Addresses the question of interest (+++)
- Is theoretically desirable (+)
- Is relatively unknown among practitioners (-)
- Takes time and effort to compute(--)
- …this is particularly true for Bayes factors, where the specification of prior distributions requires considerable care (---)

Utilities of p-values

- Fails to address the question of interest (---)
- Has theoretical problems (-)
- Is well-known among practitioners (+)
- Easy to compute (+++)
- In standard designs, will make the data appear more convincing then they are (+-)

Thanks for Your Attention!