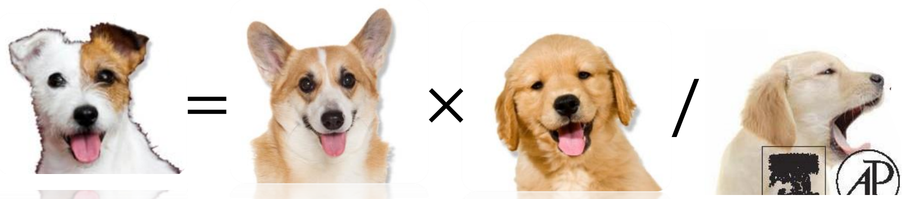
Doing Bayesian Data Analysis



$p(\theta|D) = p(D|\theta) \times p(\theta) \; / \; p(D)$

John K. Kruschke

Outline of Talk:

- Bayesian reasoning generally.
- Bayesian estimation applied to two groups. Rich information.
- The NHST *t* test: perfidious *p* values and the con game of confidence intervals.
- Conclusion: Bayesian estimation supersedes NHST.

Bayesian Reasoning

The role of data is to re-allocate credibility:

Prior Credibility with **New Data**
→ **Posterior Credibility**

via Bayes' rule

Bayesian Reasoning

The role of data is to re-allocate credibility:

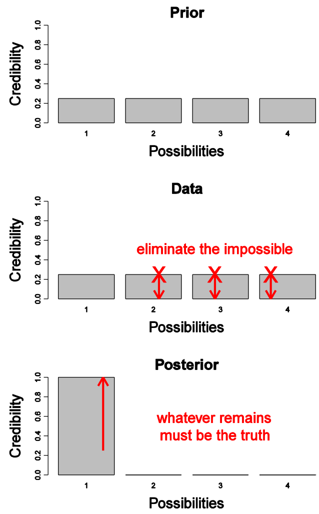
Bayesian reasoning in everyday life is intuitive:

Bayesian Reasoning

The role of data is to re-allocate credibility:

Bayesian reasoning in everyday life is intuitive:

Sherlock Holmes: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890)



Bayesian Reasoning

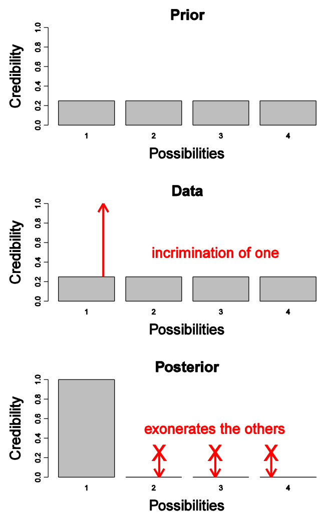
The role of data is to re-allocate credibility:

Bayesian reasoning in everyday life is intuitive:

Sherlock Holmes: "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" (Doyle, 1890)

Judicial exoneration: For unaffiliated suspects, the incrimination of one exonerates the others.

Credibility of the claim that the suspect committed the crime.



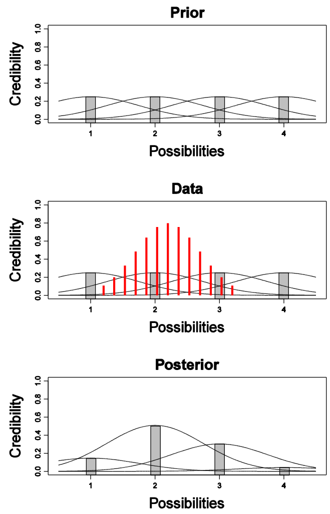
Bayesian Data Analysis

The role of data is to re-allocate credibility:

Bayesian reasoning in data analysis is intuitive:

Possibilities are *parameter values* in a model, such as the *mean* of a normal distribution.

We reallocate credibility to parameter values that are consistent with the data.



Bayesian Data Analysis

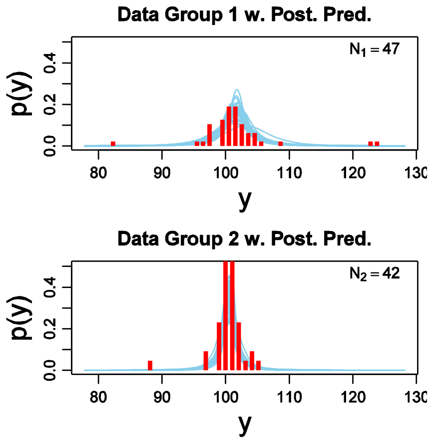
The role of data is to re-allocate credibility:

1. **Define a meaningful descriptive model.**
2. **Establish prior credibility regarding parameter values in the model. The prior credibility must be acceptable to a skeptical scientific audience.**
3. **Collect data.**
4. **Use Bayes' rule to re-allocate credibility to parameter values that are most consistent with the data.**

Robust Bayesian estimation for comparing two groups

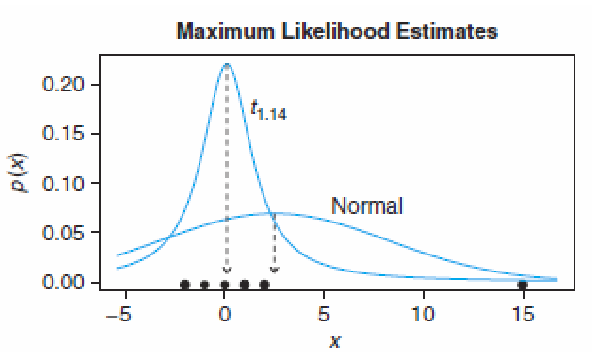
Consider two groups; e.g., IQ of "smart drug" group and of control group.

Step 1: Define a model for describing the data.



 10

Descriptive distribution for data with outliers

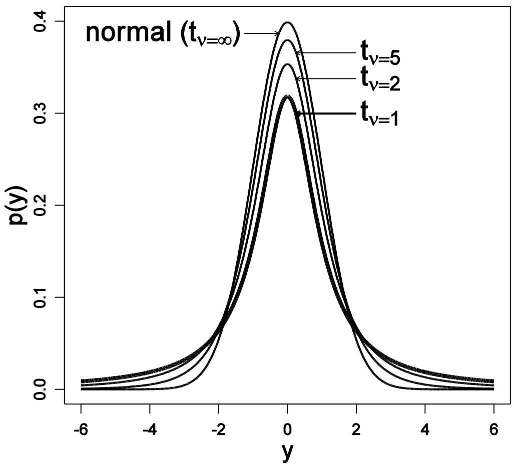


Normal is pulled by outliers, but *t* distribution is not.

t distribution is used here as a description of data, NOT as a sampling distribution for *p* values!

 11

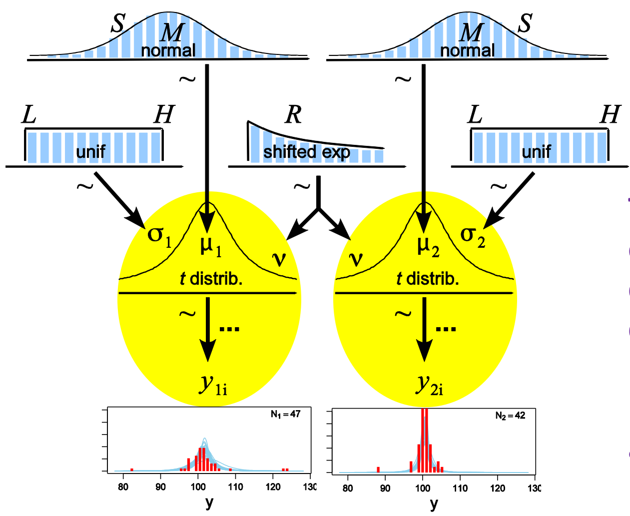
Descriptive distribution for data with outliers



The *t* distribution has normality controlled by the parameter ν.

13

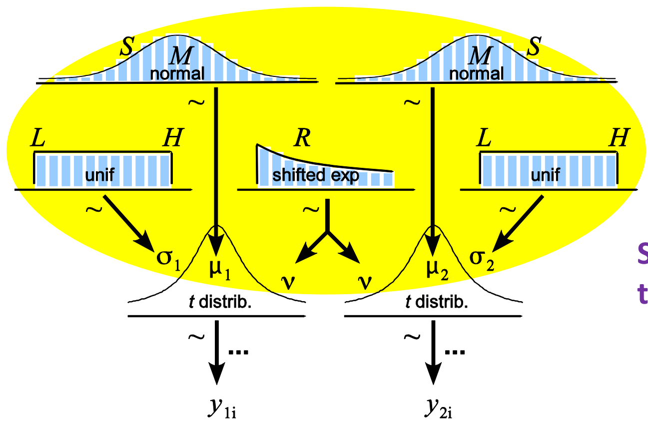
Robust Bayesian estimation for comparing two groups



The data from each group are described by *t* distributions, using five parameters altogether.

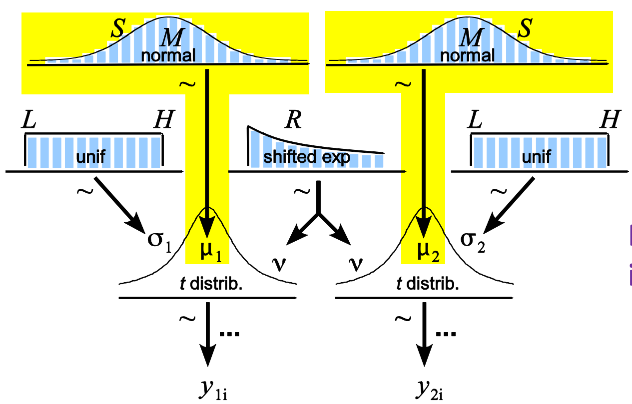
14

Robust Bayesian estimation for comparing two groups



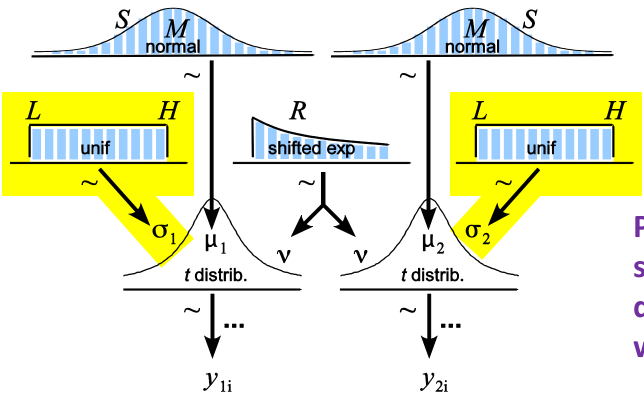
Step 2: Specify the prior.

Robust Bayesian estimation for comparing two groups



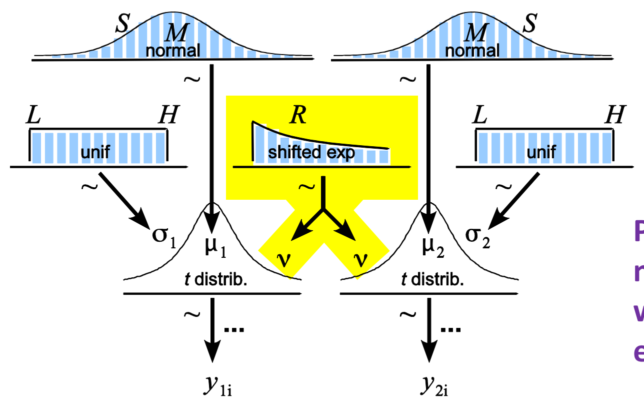
Prior on means is wide normal.

Robust Bayesian estimation for comparing two groups



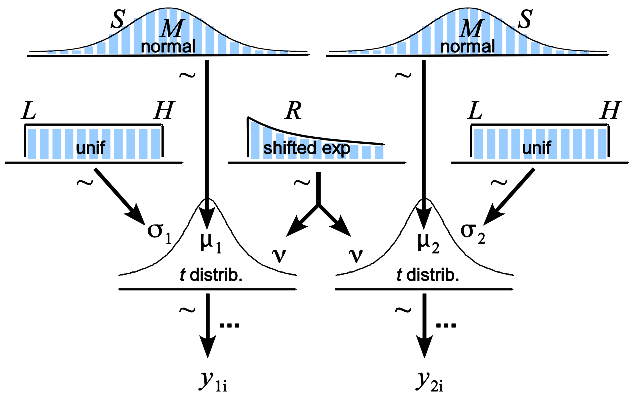
Prior on standard deviations is wide uniform.

Robust Bayesian estimation for comparing two groups



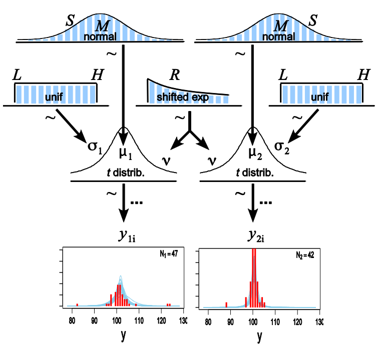
Prior on normality is wide exponential.

Robust Bayesian estimation for comparing two groups



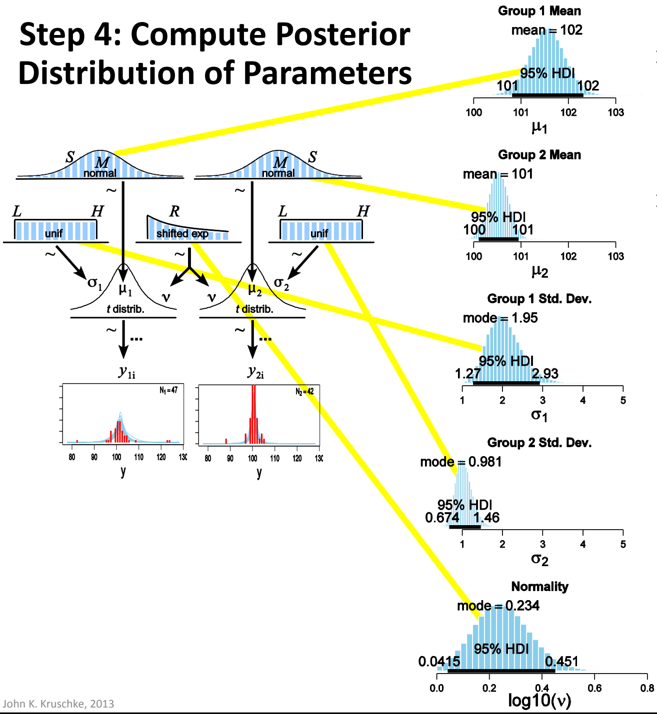
Parameter distributions will be represented by histograms: A huge number of representative parameter values.

Step 3: Collect Data.

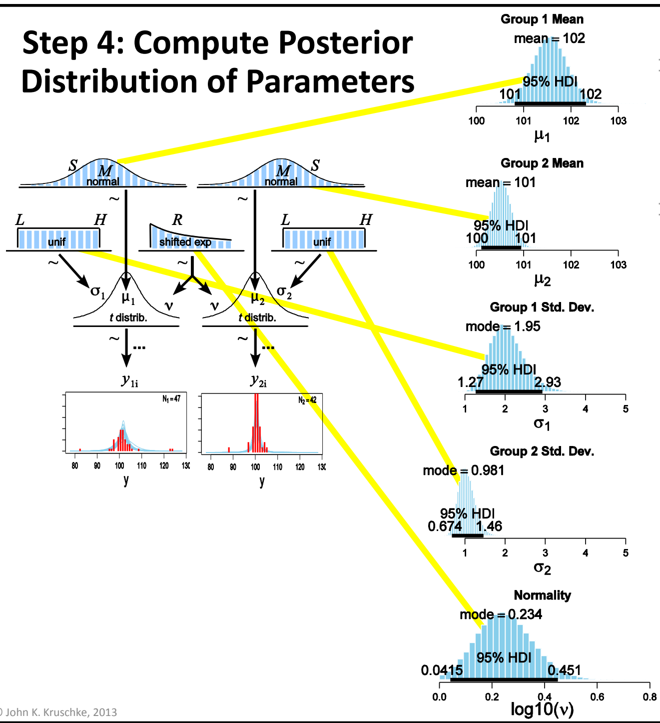


One fixed data set, shown as red histograms.

Step 4: Compute Posterior Distribution of Parameters



Step 4: Compute Posterior Distribution of Parameters



Important: **These are histograms of parameter values from the posterior distribution: A huge number of *combinations* of $\mu_1, \mu_2, \sigma_1, \sigma_2, \nu$ that are jointly credible given the data.**

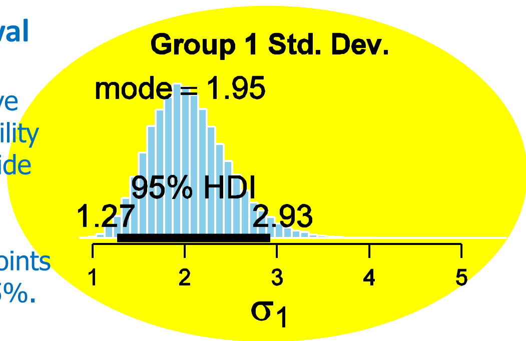
These are *not* data distributions, and *not* sampling distributions from a null hypothesis.

95% HDI:
Highest density interval

Points within the HDI have higher credibility (probability density) than points outside the HDI.

The total probability of points within the 95% HDI is 95%.

Points outside the HDI may be deemed not credible.

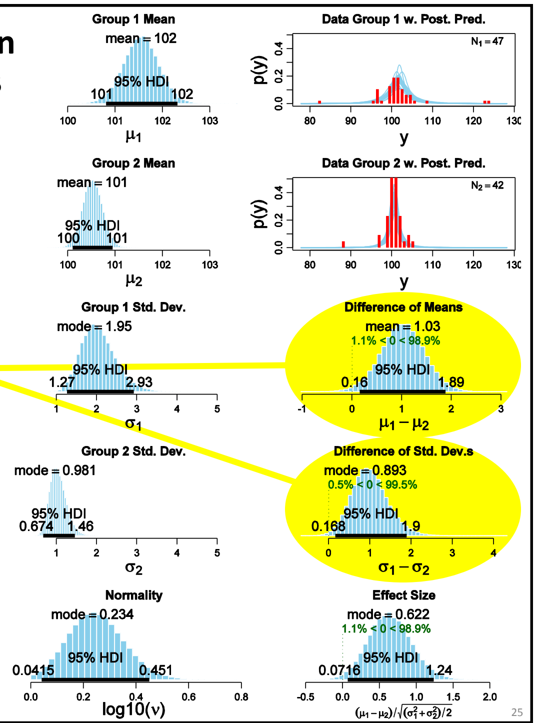


Robust Bayesian estimation for comparing two groups

Differences between groups?
Compute $\mu_1 - \mu_2$ and $\sigma_1 - \sigma_2$ at each of the many credible combinations.

Here, both differences are credibly non-zero.

(NHST would require *two* tests…)

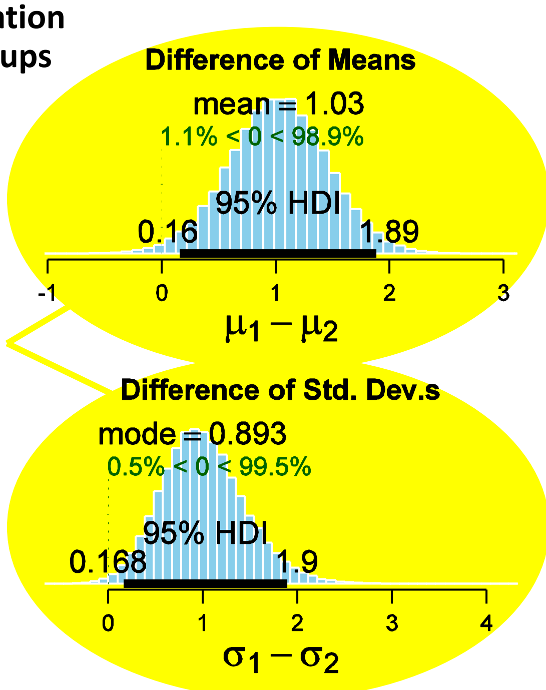


Robust Bayesian estimation for comparing two groups

Differences between groups?
Compute $\mu_1 - \mu_2$
and $\sigma_1 - \sigma_2$
at each of the many credible combinations.

Here, both differences are credibly non-zero.

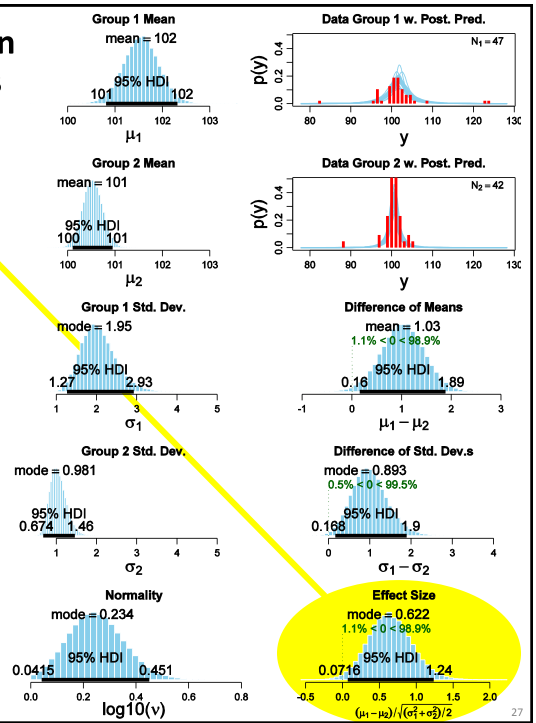
(NHST would require *two* tests…)



 26

Robust Bayesian estimation for comparing two groups

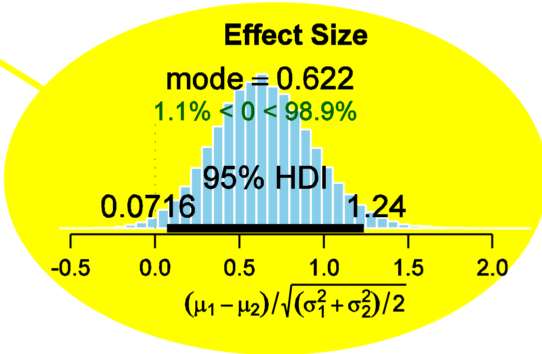
Complete distribution on effect size!



 27

Robust Bayesian estimation for comparing two groups

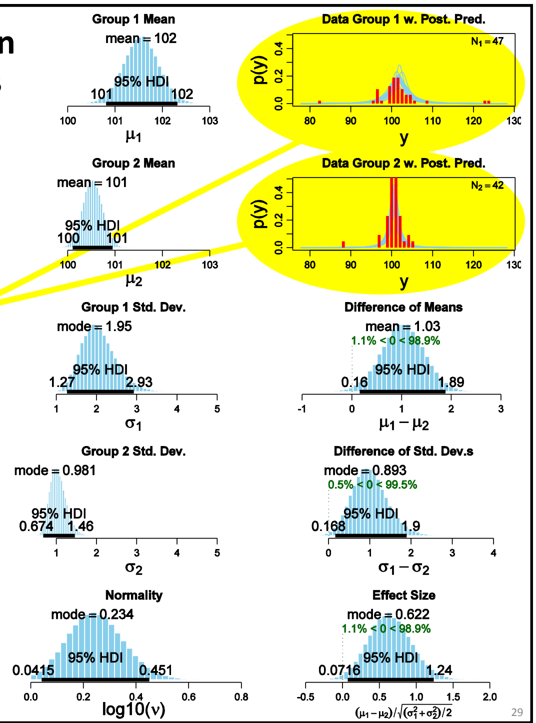
Complete distribution on effect size!



Robust Bayesian estimation for comparing two groups

Are the data described well by the model?

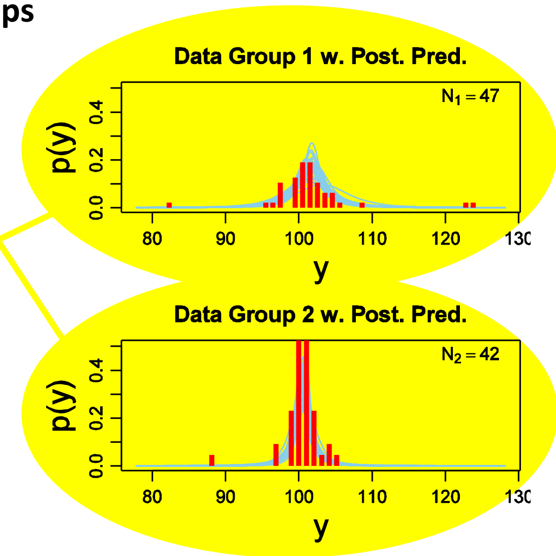
Superimpose a smattering of credible descriptive distributions on data.
= "posterior predictive check"



Robust Bayesian estimation for comparing two groups

Are the data described well by the model?

Superimpose a smattering of credible descriptive distributions on data.
= "posterior predictive check"



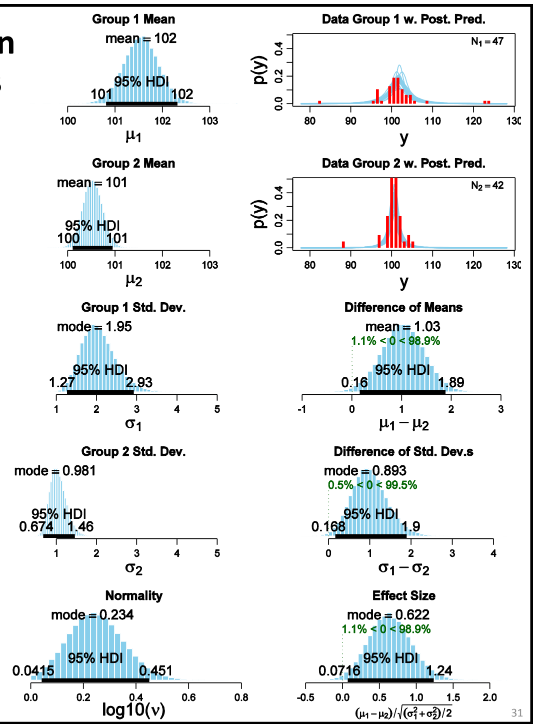
Robust Bayesian estimation for comparing two groups

Summary:

→ Complete distribution of credible parameter values (not merely point estimate with ends of confidence interval).

→ Decisions about multiple aspects of parameters (without reference to *p* values).

→ Flexible descriptive model, robust to outliers (unlike NHST *t* test).



Computer Software:

Packaged for easy use!
Underlying program is never seen.

```
source("BEST.R")   # load the program

# Specify data as vectors (replace with your own data):
y1 = c(101,100,102,104,102,97,105,105,98,…,101)
y2 = c(99,101,100,101,102,100,97,101,104,…,99)

# Run the Bayesian analysis:
mcmcChain = BESTmcmc( y1 , y2 )

# Plot the results of the Bayesian analysis:
BESTplot( y1 , y2 , mcmcChain )
```


An example of a *t* test:

Data:

Group 1: 5.70 5.40 5.75 5.25 4.25 4.74; M1 = 5.18
Group 2: 4.55 4.98 4.70 4.78 3.26 3.67; M2 = 4.32
t = 2.33

Show of hands please:

Who bets that p < .05 ? **Who bets that p > .05 ?**

An example of a *t* test:

Data:
Group 1: 5.70 5.40 5.75 5.25 4.25 4.74; M1 = 5.18
Group 2: 4.55 4.98 4.70 4.78 3.26 3.67; M2 = 4.32
t = 2.33

Show of hands please:

Who bets that p < .05 ? **Who bets that p > .05 ?**

You're right! **You're right!**

Null Hypothesis Significance Testing (NHST)

Consider how we draw conclusions from data:

- Collect data, *carefully insulated from our intentions*.
  - Double blind clinical designs.
  - No datum is influenced by any other datum before or after.
- Compute a summary statistic, e.g., for a difference between groups, the *t* statistic.
- Compute *p* value of *t*. If $p < .05$, declare the result to be "significant."

Null Hypothesis Significance Testing (NHST)

Consider how we draw conclusions from data:

- Collect data, *carefully insulated from our intentions*.
  - Double blind clinical d
  - No datum is influence
- Compute a summary s between groups, the *t* statistic.
- Compute *p* value of *t*. If $p < .05$, declare the result to be "significant."

Value of *p* depends on the intention of the experimenter!

 36

The road to NHST is paved with good intentions.

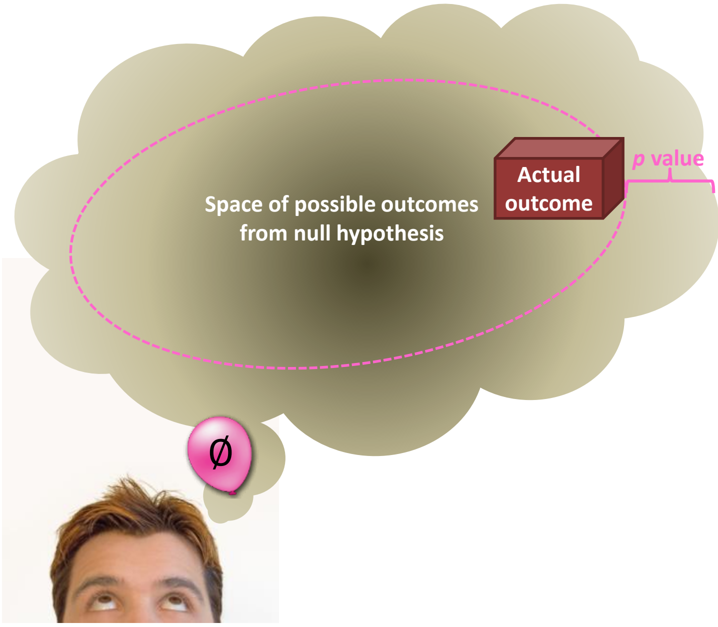
The *p* value is the probability that the actual sample statistic, or a result more extreme, would be obtained from the null hypothesis, *if the* ***intended*** *experiment were repeated ad infinitum.*

$$p \text{ value} = p\left(|t_{\text{null}}| > |t_{\text{act}}|\right)$$

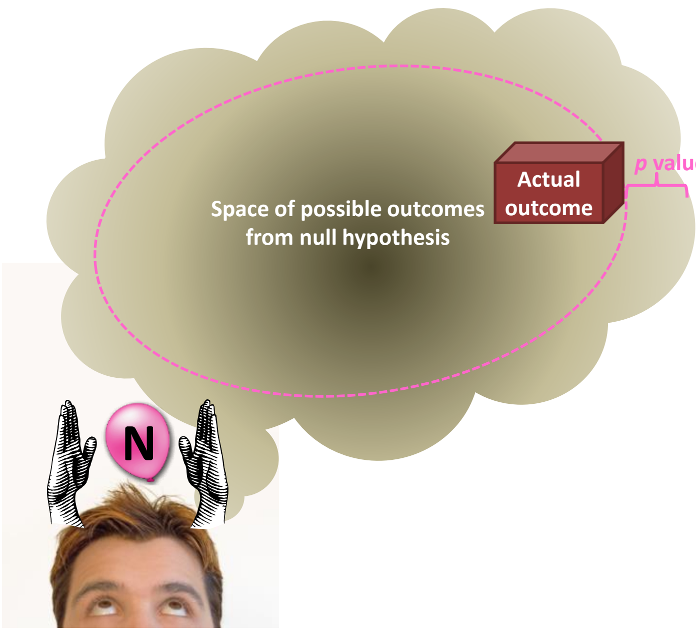
for t_{null} sampled according to the intended experiment

 37

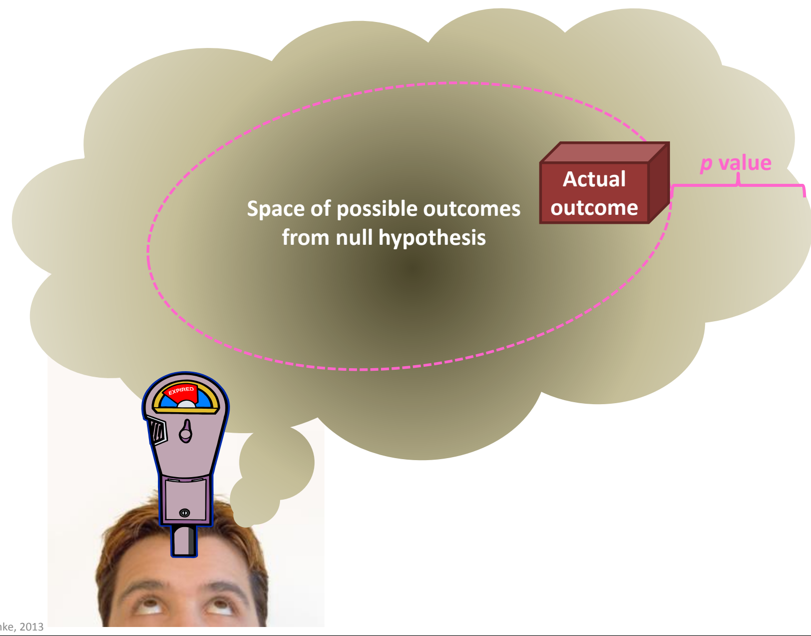
"The" *p* value…



p value for intention to sample until N



p value for intention to sample until Time

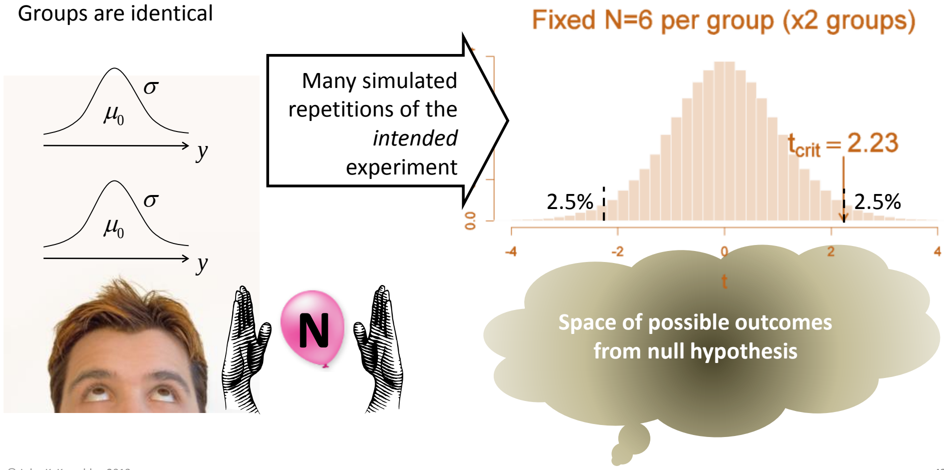


 40

The distribution of *t*

when the intended experiment is repeated many times

Null Hypothesis:
Groups are identical

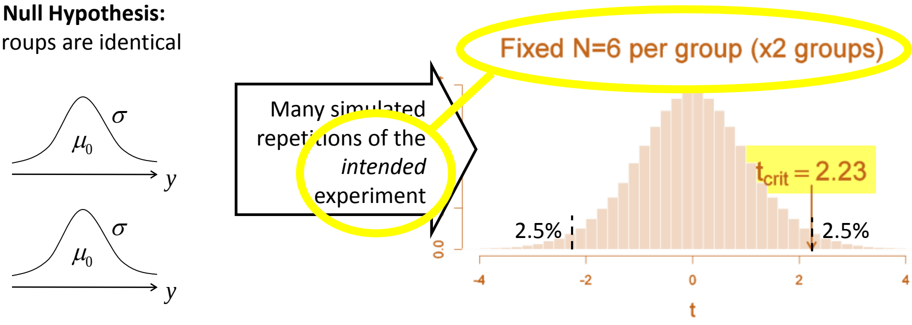


 42

The distribution of *t*

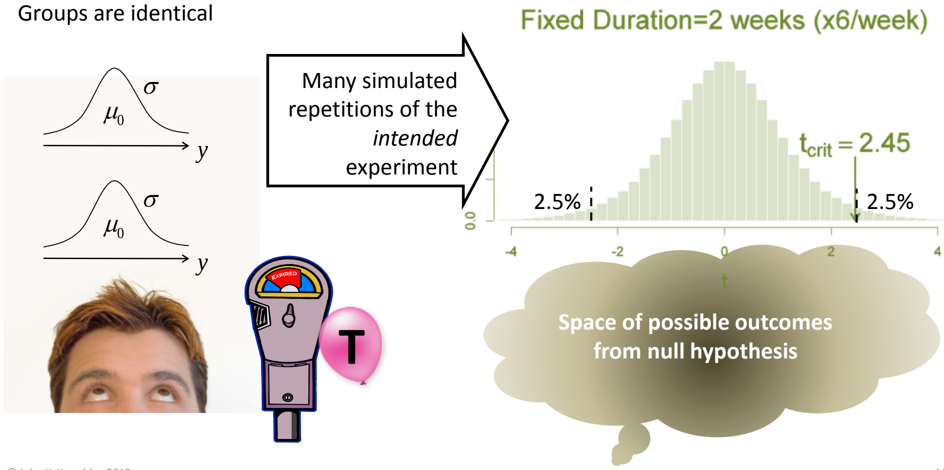
when the intended experiment is repeated many times

Null Hypothesis:
Groups are identical



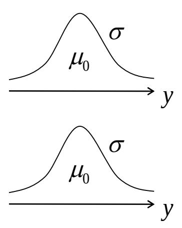
The intention to collect data until the end of the week

Null Hypothesis:
Groups are identical

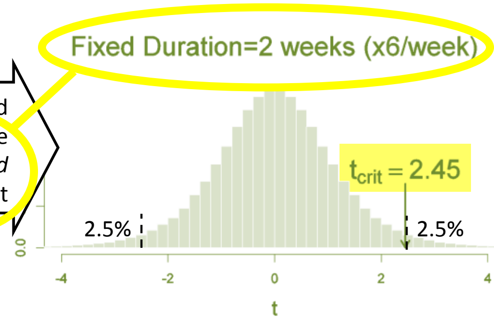


The intention to collect data until the end of the week

Null Hypothesis:
Groups are identical



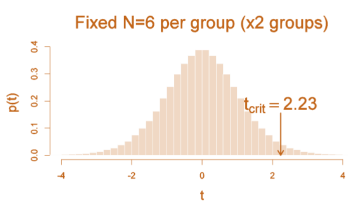
Many simulated repetitions of the *intended* experiment



Two labs collect the same data:

Lab A: Collect data until N=6 per group.

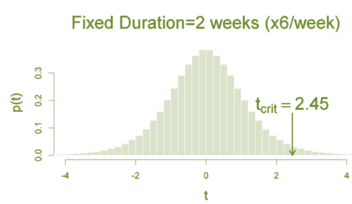
Data:
Group 1: 5.70 5.40 5.75 5.25 4.25 4.74; M1 = 5.18
Group 2: 4.55 4.98 4.70 4.78 3.26 3.67; M2 = 4.32
t = 2.33



Lab A: *Reject* the null.

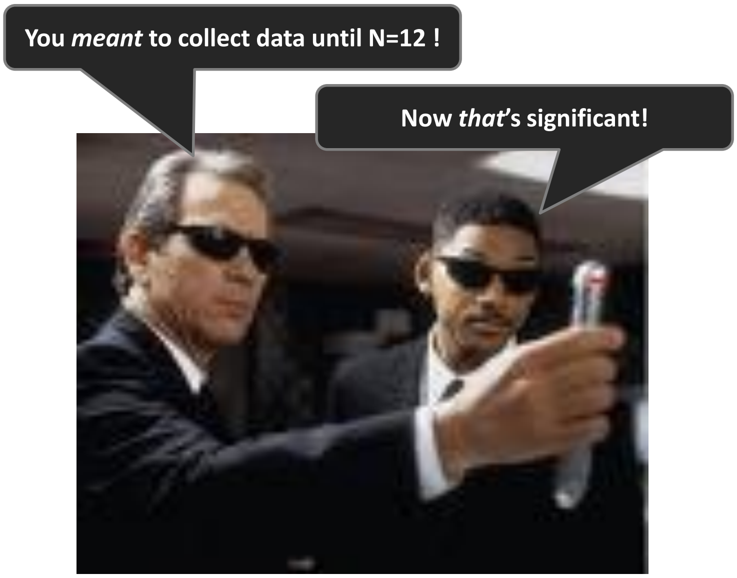
Lab B: Collect data for two weeks.

Data:
Group 1: 5.70 5.40 5.75 5.25 4.25 4.74; M1 = 5.18
Group 2: 4.55 4.98 4.70 4.78 3.26 3.67; M2 = 4.32
t = 2.33



Lab B: Do *not* reject the null.

The *real* use of the Neuralyzer:



Problem is not solved by "fixing" the intention

- All we need to do is decide in advance exactly what our intention is (or use a Neuralyzer after the fact), and have everybody chant a mantra to keep that intention fixed in their minds while the experiment is being conducted. Right?
- Wrong. The data don't know our intention, and the same data could have been collected under many other intentions.

The intention to examine data thoroughly

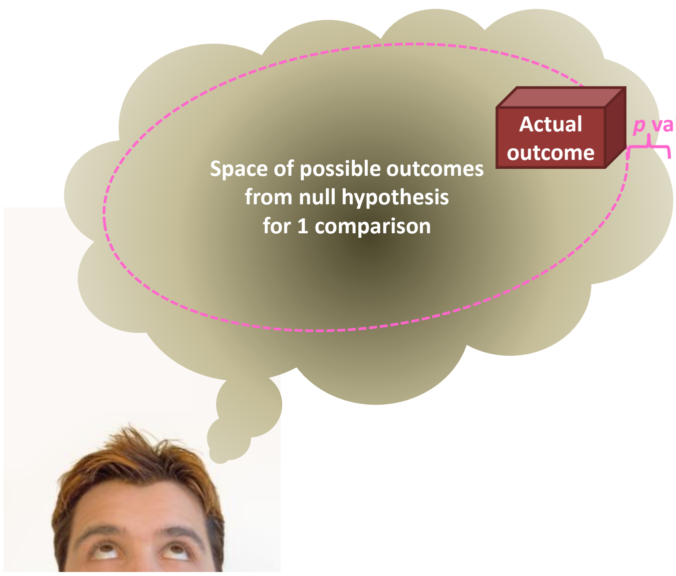
Many experiments involve multiple groups, and **multiple comparisons** of means.

Example: Consider 2 different drugs from chemical family A, 2 different drugs from chemical family B, and a placebo group. Lots of possible comparisons…

Problem: With every test, there is possibility of false alarm! False alarms are bad; therefore, keep the experimentwise false alarm rate down to 5%.

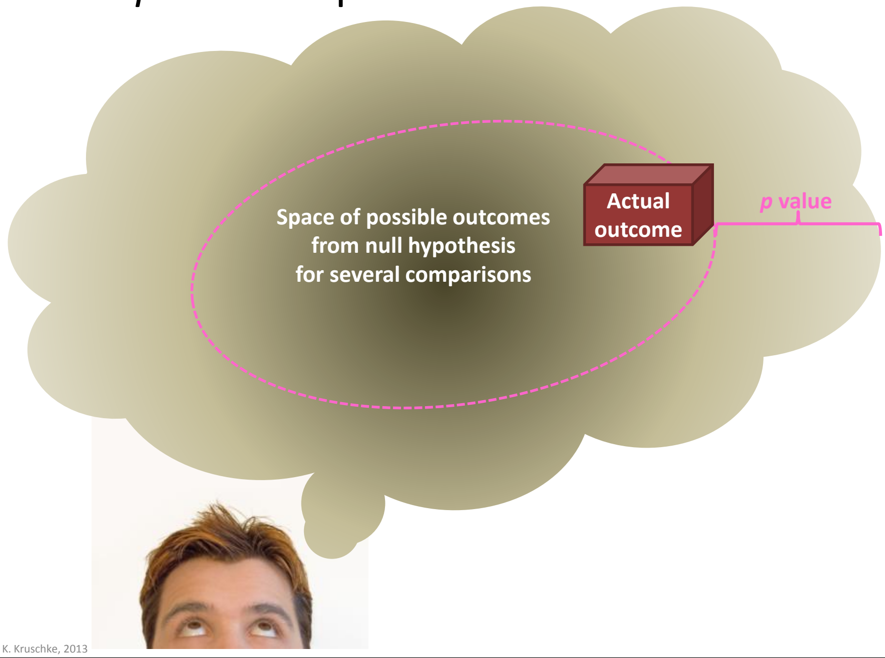
 50

"The" *p* value depends on intended tests:



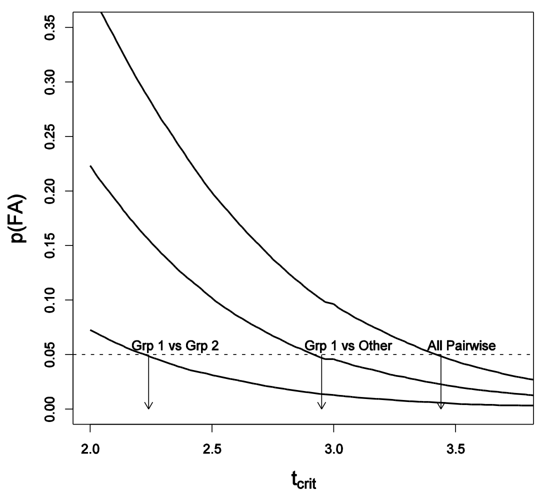
 51

"The" *p* value depends on intended tests:



52

Experimentwise false alarm rate



53

Multiple Corrections for Multiple Comparisons

Begin: Is goal to identify the best treatment?
- Yes: Use **Hsu's method.**
- No: Contrasts between control group and all other groups?
  - Yes: Use **Dunnett's method.**
  - No: Testing all pairwise and no complex comparisons (either planned or post hoc) and choosing to test only some pairwise comparisons post hoc?
    - Yes: Use **Tukey's method.**
    - No: Are all comparisons planned?
      - Yes: Use **Scheffe's method.**
      - No: Is Bonferroni critical value less than Scheffe critical value?
        - Yes: Use **Bonferroni's method.**
        - No: Use Scheffe's method (or, prior to collecting the data, reduce the number of contrasts to be tested).

Adapted from Maxwell & Delaney (2004). Designing experiments and analyzing data: A model comparison perspective. Erlbaum.

Multiple Corrections for Multiple Comparisons

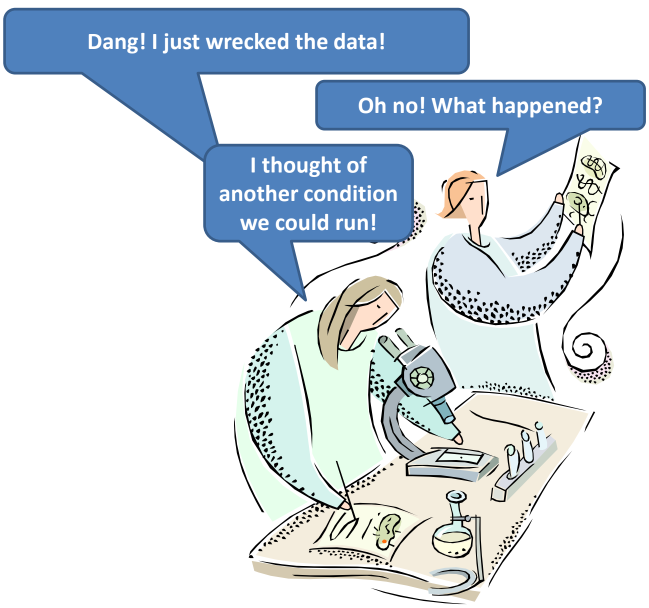
Begin: Is goal to identify the best treatment?
- Yes: Use **Hsu's method.**
- No: Contrasts between control group and all other groups?
  - Yes: Use **Dunnett's method.**
  - No: Testing all pairwise and no complex comparisons (either planned or post hoc) and choosing to test only some pairwise comparisons post hoc?
    - Yes: Use **Tukey's method.**
    - No: Are all comparisons planned?
      - Yes: Use **Scheffe's method.**
      - No: Is Bonferroni critical value less than Scheffe critical value?
        - Yes: Use **Bonferroni's method.**
        - No: Use Scheffe's method (or, prior to collecting the data, reduce the number of contrasts to be tested).

!

Adapted from Maxwell & Delaney (2004). Designing experiments and analyzing data: A model comparison perspective. Erlbaum.

Good intentions make any result *in*significant

- Consider an experiment with two groups.
- Collect data; compute *t* test on difference of means. Suppose it yields $p < .05$
- Now, think thoroughly about all the other comparison groups and other experiment groups you should and could meaningfully run.
- Earnestly intend to run them eventually, and to compare your current results with those results.
- *Poof!* Your current data are no longer significantly different.



Good intentions make many results *significant*

- Consider an experiment with two groups.
- Collect data; compute *t* test on difference of means, using df corresponding to actual N. Suppose $p > .05$, but not by much.
- *You had intended to collect a much larger sample size, but you were unexpectedly interrupted*.
- Use the larger intended N for df in the *t* test.
- *Poof! Your current data are now significantly different!*

? Confidence Intervals provide no confidence ?

General definition of CI:

95% CI is the range of parameter values (e.g., $\mu_1 - \mu_2$) that would not be rejected by $p < .05$

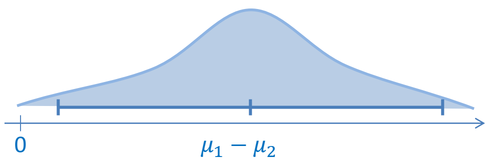
Hence, *the 95% CI is as ill-defined as the p value.*

We see this dramatically in confidence intervals corrected for multiple comparisons.

Confidence Intervals provide no confidence

Confidence intervals provide no distributional information:

We have no idea whether a point at the limit of the confidence interval is any less credible than a point in the middle of the interval.



Confidence Intervals provide no confidence

Confidence intervals provide no distributional information:

We have no idea whether a point at the limit of the confidence interval is any less credible than a point in the middle of the interval.

Implies

vast range for predictions of new data, and "virtually unknowable" power.

NHST autopsy

- *p* values are ill-defined: depend on sampling intentions of data collector. Any set of data has many different *p* values.
- Confidence intervals are as ill-defined as *p* values because they are defined in terms of *p* values.
- Confidence intervals carry no distributional information.

Recall Bayesian estimation for comparing two groups

Summary:

→ Complete distribution of credible parameter values (not merely point estimate with ends of confidence interval)**.**

→ Decisions about multiple aspects of parameters (without reference to *p* values)**.**

→ Flexible descriptive model, robust to outliers (unlike NHST *t* test)**.**