

The (R)Evolution in Statistics: Are We All Becoming Bayesians?

A Discussion for Non-Statisticians

Brani Vidakovic

Georgia Tech

H. Milton Stewart School of Industrial and Systems Engineering, and
The Wallace H. Coulter Department of Biomedical Engineering

- Bayes Rule – Learning About Events

Overview

- Bayes Rule – Learning About Events
- Bayesian Paradigm and Examples

Overview

- Bayes Rule – Learning About Events
- Bayesian Paradigm and Examples
- Why there are so Many Non-Replicable Studies?

Overview

- Bayes Rule – Learning About Events
- Bayesian Paradigm and Examples
- Why there are so Many Non-Replicable Studies?
- Bayesian Alternatives.

Overview

- Bayes Rule – Learning About Events
- Bayesian Paradigm and Examples
- Why there are so Many Non-Replicable Studies?
- Bayesian Alternatives.
- MCMC, WinBUGS, and Meta-Analysis Example

Bayes (1763)

Problem:

Given the number of times in which an unknown event has happened and failed: Required the **chance** that the **probability** of its happening in a single trial lies between any two degrees of probability that can be named.

Bayes (1763)

Problem:

Given the number of times in which an unknown event has happened and failed: Required the **chance** that the **probability** of its happening in a single trial lies between any two degrees of probability that can be named.



Thomas Bayes (1702–1761), from “History of Life Insurance” by Terence ODonnell, 1936.

Bayes' Rule and Bayes' Theorem

Bayes' Rule

Let H_1, H_2, \dots, H_n be set of mutually exclusive events partitioning sample space. Thus, $P(H_1) + \dots + P(H_n) = 1$. Let A can happen under any of H_i with known probabilities $P(A|H_i)$. Then,

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A)},$$

where $P(A) = P(A|H_1)P(H_1) + \dots + P(A|H_n)P(H_n)$.

Bayes' Rule and Bayes' Theorem

Bayes' Rule

Let H_1, H_2, \dots, H_n be set of mutually exclusive events partitioning sample space. Thus, $P(H_1) + \dots + P(H_n) = 1$. Let A can happen under any of H_i with known probabilities $P(A|H_i)$. Then,

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A)},$$

where $P(A) = P(A|H_1)P(H_1) + \dots + P(A|H_n)P(H_n)$.

Learning by Bayes' Rule

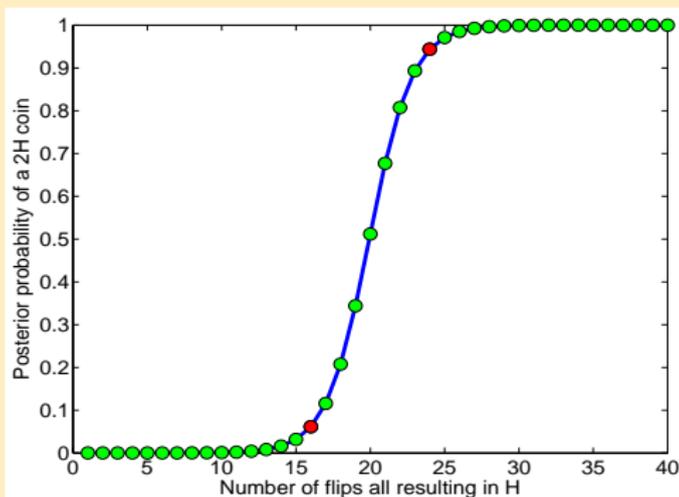
$$P(H_i) \quad \longrightarrow \quad P(H_i|A) \left[= \frac{P(A|H_i)}{P(A)} \times P(H_i) \right].$$

Two-headed Coin

$N = 1,000,000$ coins, 999,999 fair and 1 “two-headed.” A coin is selected at random, flipped n times and in all flips it falls heads up. What is the probability that the two-headed coin was selected?

Two-headed Coin

$N = 1,000,000$ coins, 999,999 fair and 1 “two-headed.” A coin is selected at random, flipped n times and in all flips it falls heads up. What is the probability that the two-headed coin was selected?



Sensitivity, Specificity, and Relatives

Casscells et al. (1978) to 60 students and staff at an elite med school

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, [and true positive rate of 100%], what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

Definitions

D = Disease present, D^c = Disease not present, $+$ = Test positive, $-$ = Test Negative.

$P(+|D)$ = Sensitivity, $P(-|D^c)$ = Specificity, $P(D|+)$ = Positive predictive value (PPV), $P(D^c|-)$ = NPV

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, [and true positive rate of 100%], what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?

The answer to this problem is approximately 2%. Casscells et al. found that only 11 participants gave this answer. The most frequent response was 95%, presumably on the supposition that, because the error rate of the test is 5%, it must get 95% of results correct.

$$P(+|D^c) = 0.05, P(-|D^c) = 0.95, P(+|D) = 1, P(-|D) = 0.$$

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|D^c)P(D^c)} = 1 \times 0.001 / (1 \times 0.001 + 0.05 \times 0.999) = \mathbf{0.0196}$$

Bayes' Theorem

With Bayes' Rule and Inverse Probabilities of Events, no Philosophical Disagreements!

Bayesian – Frequentist Philosophical Disagreement

Disagreements are in the **nature of model parameters** and **use of conditioning**.

- Frequentists/Classical Statisticians: Parameters are **fixed numbers**, Inference involves **optimization**.
- Bayesian Statisticians: Parameters are **random variables**, Inference involves **integration**.

Parameter θ : Elicit prior $\pi(\theta)$, update to the posterior after observing experimental results.

Ingredients in Bayesian Inference

Prior \longrightarrow Posterior

Bayes' Theorem: $\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}$

Ingredients in Bayesian Inference

Prior \longrightarrow Posterior

Bayes' Theorem:
$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal}}$$

- Prior - subjective part of the model
- Likelihood - incorporates experimental data/observations, conditional on data
- Marginal (Prior Predictive) - normalizing constant (“a trouble maker”)

The only coherent way of combining the experimental data and prior information is via the Bayes' Theorem.

- Prior - elicited, conjugate, objective, non-informative, reference, automatic, optimistic/pesimistic, clinical, ...



Posterior Distribution: Ultimate Summary

- Inference conceptually natural and simple.
- Location of the posterior (mean, median, mode) – Bayes' estimators of a parameter.
- Credible sets (Bayesian counterparts of confidence intervals) obtained by percentiles of the posterior.
- Testing hypothesis done by comparing posterior probabilities of competing hypotheses.

Example: Ten Coin Flips – No Heads Up

We are interested in estimating parameter p , probability that possibly biased coin falls heads up.

- Experiment: $n = 10$ flips, $X = 0$ heads observed.

With a Frequentist Hat

- If you are frequentist, only the experiment matters, the estimate of p is $\hat{p} = X/n = 0/10 = 0$. [?!!]

With a Bayesian Hat

- For a Bayesian, the likelihood is conditional on $X = 0$, and since it is Binomial, it is proportional to $p^0(1 - p)^{10}$.
- Under uniform prior the posterior is proportional to $p^0(1 - p)^{10} \times 1$ which is a un-normalized Beta distribution with parameters 1 and 11. The Bayes estimator of p is $1/(1+11) = 1/12$, a more reasonable estimator than the frequentist's 0.

The Likelihood Principle

All information about experiment is in the likelihood. A Bayesian inference **is based on data observed** and not on data that could possibly be observed, or on the manner in which the sampling was conducted.

Example: (Jimmie Savage, 1962, Purdue Symposium)

Suppose a coin is flipped 12 times and 9 heads and 3 tails are obtained. Let p be the probability of heads.

We are interested in testing whether the coin is fair against the alternative that it is more likely to come heads up, or

$$H_0 : p = 1/2 \quad \text{vs.} \quad H_1 : p > 1/2.$$

The p -value for this test is the probability that one observes 9 or more heads if the coin is fair (under H_0).

Savage's Example Cont'd

Consider two scenarios:

Scenario A

Suppose that the number of flips $n = 12$ was decided a priori. Then the number of heads X is binomial and under H_0 (fair coin) the p -value is 0.0730. At a 5% significance level H_0 is **not rejected**

Scenario B

Suppose that the flipping is carried out until 3 tails have appeared. Then, under H_0 , the number of heads Y is a negative binomial and the p -value is 0.0327. At a 5% significance level H_0 is **rejected**.

Thus, two classical tests recommend opposite actions for the same data simply because of how the sampling was conducted.

For a Bayesian – No Difference

In both (A) and (B) the likelihood is proportional to $p^9(1-p)^3$, and for a fixed prior on p there is **no difference in any Bayesian inference**.

Edwards et al. (1963)

“... the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.”

Lack of Reproducibility

The reliability of results from observational studies has been called into question many times in the recent past, with several analyses showing that well over half of the reported findings are subsequently refuted (JNCI, 2007).

The NIH funded randomized clinical trials to follow up exciting results from 20 observational studies. Only 1 replicated.

Bayer Healthcare reviewed 67 in-house attempts at replicating the findings in published research. Less than 1/4 were viewed as having been essentially replicated. Over 2/3 had major inconsistencies leading to project termination (Wadman, 2013).

Amgen publication shows that findings from only 6 out of 53 landmark papers can be replicated by company scientists.

John P. A. Ioannidis, JAMA, 2005, 218-228: Five of 6 highly cited nonrandomized studies were contradicted or had found stronger effects than were established by later studies.

Ioannidis looked at the 49 most famous medical publications from 1990-2003 resulting from randomized trials; 45 claimed successful intervention.

- 7 (16%) were contradicted by subsequent studies
- 7 others (16%) had found effects that were stronger than those of subsequent studies
- 20 (44%) were replicated

50% phase III drug trial failure rates are now being reported
30% of phase III drug trial successes fail to replicate

Why?

- Publication bias
- Significant rewards for positive results with little or no penalty for refuted studies
- Use of egregiously bad statistics. [e.g., If A, B are treatments and C control, $A = C$ & $B \neq C \Rightarrow A \neq B$.]
- Stochastic biases (dependencies, confounding, oversimplified models)
- Failure to properly account for multiple testing.
- Fallacies of p -values

Multiplicity in Testing

Basic research is like shooting an arrow in the air and, where it lands, painting a target, Homer Adkins.



...With widely-used methods, Peter Austin, Ph.D., (Institute for Clinical Evaluative Sciences in Toronto) found that Leos were (significantly) more likely to have gastrointestinal bleeding, while Sagittarians were (significantly) more likely be hospitalized for a broken arm...[Tuma (2007), JNCI 99, 664–668]

Fisher (1926)

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). **Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level.** A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

Matthews (1998)

The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug.

p -values are not error probabilities

A survey reported by Jim Berger

What would you conclude if a properly conducted, randomized clinical trial of a treatment was reported to have resulted in a beneficial response ($p < 0.05$)?

1. Having obtained the observed response, the chances are less than 5% that the therapy is not effective.
2. The chances are less than 5% of not having obtained the observed response if the therapy is effective.
3. The chances are less than 5% of having obtained the observed response if the therapy is not effective.
4. None of the above.

The (most) correct answer is **3**

The question was given to 24 physicians. Half answered incorrectly; all had difficulty distinguishing the subtle differences.

The really correct answer

The chances are less than 5% of having obtained the observed response **or any more extreme** response if the therapy is not effective.

Is accounting for possible not observed data fair?

Jeffreys (1961)

A hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.

NIH Workshop on Reproducibility and Rigor of Research, June 2014

- Rigorous statistical analysis (journals have to have a mechanism to check for statistical accuracy in addition to the regular review process)
- Transparency in Reporting (Standards. Replicates, Statistics, Randomization, Blinding, Power Analysis, Inclusion-Exclusion Criteria)
- Data and Material Sharing (must be made available, if ethically appropriate)
- Consideration of Refutations (Journals obliged to publish refutations, subject to usual standards of quality)
- Establishing Best Practices Guidelines (e.g., precise description of reagents, cell lines, antibodies, animals, etc.)

FDA 2010 Guidelines (for medical devices)

- Valuable prior information is often available for medical devices because of their mechanism of action and evolutionary development.
- Correctly employed Bayesian approaches may be less burdensome.
- Often the use of prior information may alleviate the need for a larger sized trial.
- When an adaptive Bayesian model is applicable, the size of a trial can be reduced by stopping early when conditions warrant.
- The Bayesian approach can sometimes obtain an exact analysis when the frequentist analysis is approximate or too difficult.
- Bayesian approaches to multiplicity problems (multiple endpoints and testing of multiple subgroups) may be advantageous.
- Bayesian methods allow for great flexibility in dealing with missing data.

FDA 2010 Recommendations

In the context of clinical trials, an unlimited look at the accumulated data when sampling is of a sequential nature will not affect the inference. In the frequentist approach, interim data analyses affect type I errors. The ability to stop a clinical trial early is important from the moral and economic viewpoints. Trials should be stopped early due to both futility, to save resources or stop an ineffective treatment, and superiority, to provide patients with the best possible treatments as fast as possible.

Bayesian Handling of Multiplicity and Significance Testing

Multiplicity

Bayesian treatment of multiplicity in testing is ONLY via the prior and is separated from the model/likelihood structure. (In multiple testing for significant gene expressions, if the chip has 10000 genes, each starts with prior probability of $1/10000$ of being expressed).

In GWAS (Genome Wide Association Studies) Wellcome Trust Case Control Consortium proposed a stringent cutoff of $\alpha < 5 \times 10^{-7}$ partially based on Bayesian considerations.

Significance Tests

Bayes' Factors:

Posterior Odds in Favor of H_1 = Prior Odds in Favor of H_1 \times
Bayes' Factor

$$\frac{P(H_1|\text{data})}{P(H_0|\text{data})} = \frac{P(H_1)}{P(H_0)} \times B_{10}$$

Table: Treatment of H_0 according to log-Bayes' factor values: Jeffreys' Scale (Jeffreys, 1961, p. 432)

| Value (Log 10) | Value (Natural Log) |
|---------------------------------------|---------------------|
| $0 \leq \log_{10} B_{10}(x) \leq 0.5$ | poor |
| $0.5 < \log_{10} B_{10}(x) \leq 1$ | substantial |
| $1 < \log_{10} B_{10}(x) \leq 1.5$ | strong |
| $1.5 < \log_{10} B_{10}(x) \leq 2$ | very strong |
| $\log_{10} B_{10}(x) > 2$ | decisive |

MCMC Revolution

- Why Isn't Everyone a Bayesian? ... Brad Efron in American Statistician (1986)

- Instead of finding the posterior, simulate from it.

- Metropolis et al. (1953), Hastings (1970), Geman & Geman (1984), Tanner & Wong (1987), Gelfand & Smith (1990),... ■ Independence in simulated data not critical since Ergodic Theorem for MC holds:

- If θ_i are sampled from a MC with stationary distribution π , then for any function h

$$\frac{\sum_{i=1}^n h(\theta_i)}{n} \rightarrow E_{\pi} h(\theta), \text{ a.s.}$$

- Various strategies to construct MC with where the stationary distribution is the posterior π : Metropolis-Hastings, Gibbs Sampler,

BUGS: Bayesian analysis Using Gibbs Sampling

- A prototype was demonstrated at the 4th Valencia Bayesian meeting in 1991. Only with Gibbs updater.
- The range of other updating methods (Metropolis & family, Slice Sampling, etc) were added after 1996 when the project moved from MRC Cambridge to Imperial College.
- In the mid-1990s WinBUGS, a standalone version of the software was created.
- WinBUGS written in Component Pascal, within Oberon's Rapid Application Development environment known as BlackBox Component Builder (<http://www.oberon.ch/blackbox.html>) [**.odc (oberon document)*]
- In 2004 Andrew Thomas started an open-source version of the software at the University of Helsinki.
- At CDC OpenBUGS 3.2.1 was approved on 6/9/2011 (requestor Antonio Vieira) and WinBUGS 1.4.3 on 7/15/2011 (requestor Elizabeth Zell).

Tamofixen and BC 1/5

OpenBUGS

File Edit Attributes Tools Info Model Inference Doodle Map PKBugs PKHelp Text Window Examples Manuals Help

BCrecurrence

BUGS

BC Recurrence Meta-Analysis

In 14 trials the effect of tamofixen on BC recurrence was meta analyzed.

$rT=c(55, 137,505, 62, 99, 50, 185,186,148,25,223,183,2,129),$
 $nT=c(97, 282,927,123,239,130,311,303,325,79,344,937,2,434),$
 $rC=c(67, 187,590, 74,118, 49, 200,187,178,38,224,185,0,159),$
 $nC=c(100, 306,915,140,236,107,319,307,325,86,350,936,8,449)$

Three different scales LOR, LRR and ARD are considered. A hypothetical new trial probabilities of RR exceeding 1 are found under the three scales for effect assessment (uncomment desired scale in the code).

Parmigianni, G. (2002). Modeling in Medical Decision Practice, Wiley, p. 127
Congdon, P. (2010). Applied Bayesian Hierarchical Methods, Wiley, p. 115

saved

Tamofixen and BC 2/5

```
BCreurrence

model { for (j in 1:J) {
  rT[j]~dbin(pT[j],nT[j]);
  rC[j] ~ dbin(pC[j],nC[j]);
  pC[j] ~ dbeta(a.C,b.C);
  effect[j] ~ dnorm(mu.effect,tau.effect)}

# alternative scales
# log OR scale
  for (j in 1:J) {logit(pT[j]) <- logit(pC[j])+ effect[j]}
  logit(pT.new) <- logit(pC.new)+effect.new

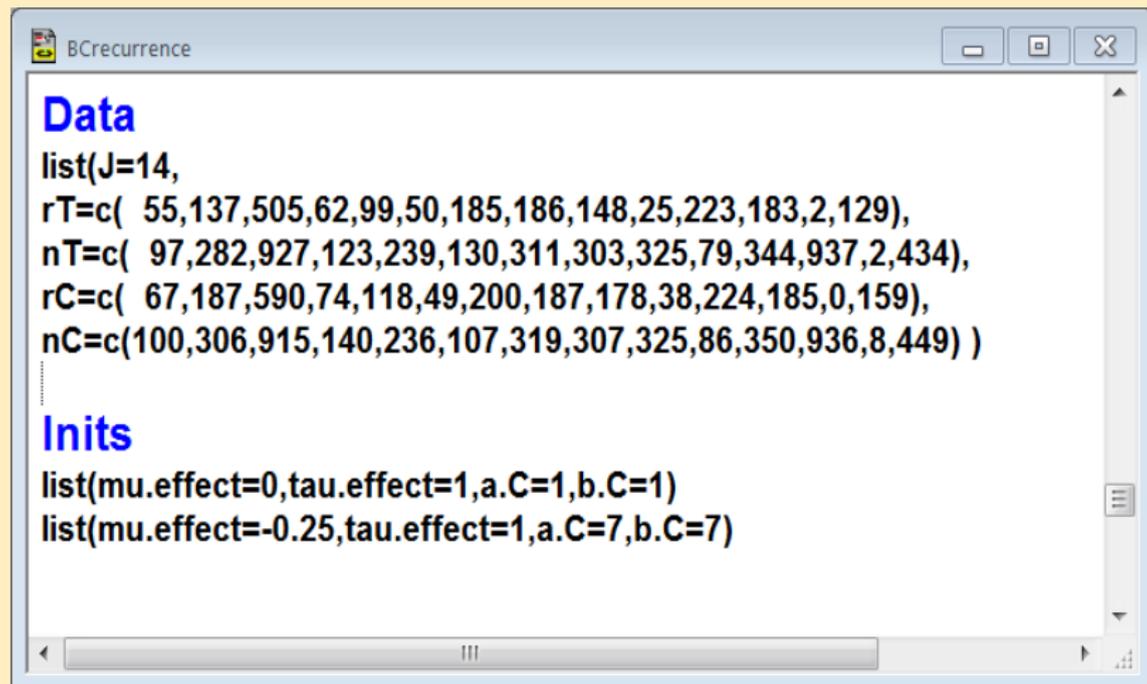
# log RR scale
  # for (j in 1:J) {log(pT[j]) <- log(pC[j])+min(effect[j],-log(pC[j]))}
  # log(pT.new) <- log(pC.new)+effect.new

# absolute risk difference scale
  # for (j in 1:J) {pT[j] <- pC[j]+min(max(effect[j],-pC[j]),(1-pC[j]))}
  # pT.new <- pC.new+effect.new

# predictive relative risk (all models)
effect.new ~ dnorm(mu.effect,tau.effect); pC.new ~ dbeta(a.C,b.C);
RRnew <- pT.new/pC.new; RRnew.above.1 <- step(RRnew-1)

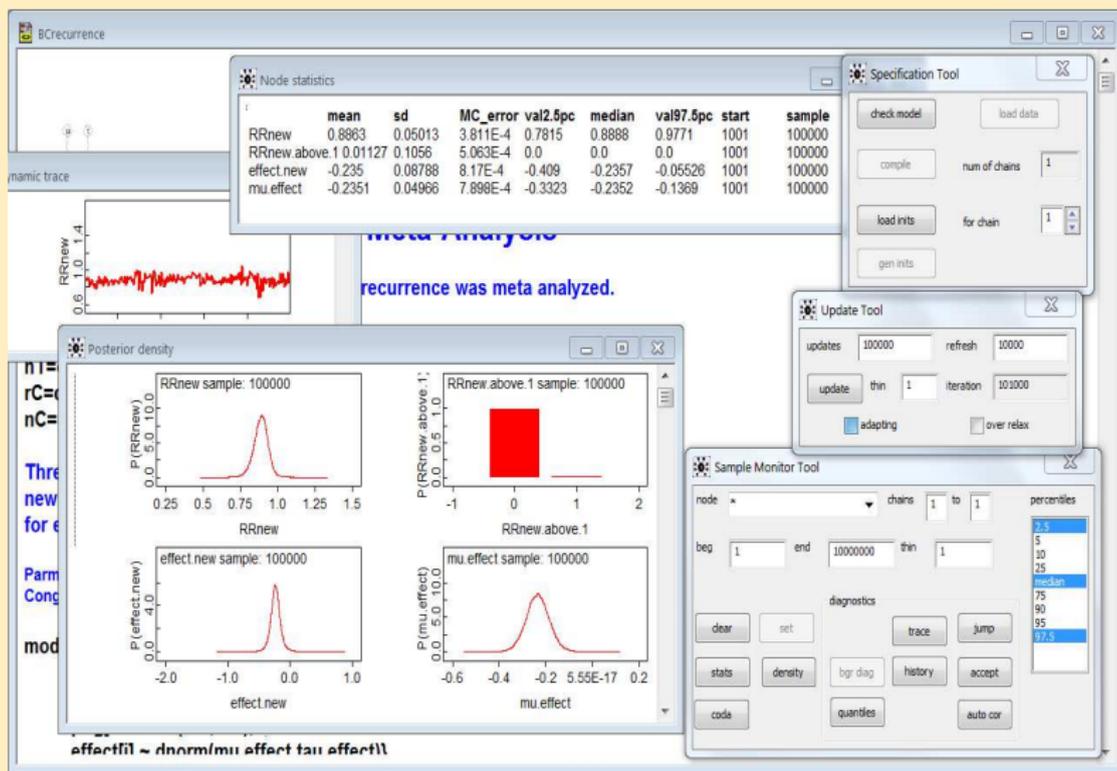
# hyperpriors
a.C ~ dunif(1,100); b.C ~ dunif(1,100)
mu.effect ~ dnorm(0,0.001); tau.effect ~ dgamma(1,0.001)}
```

Tamofixen and BC 3/5

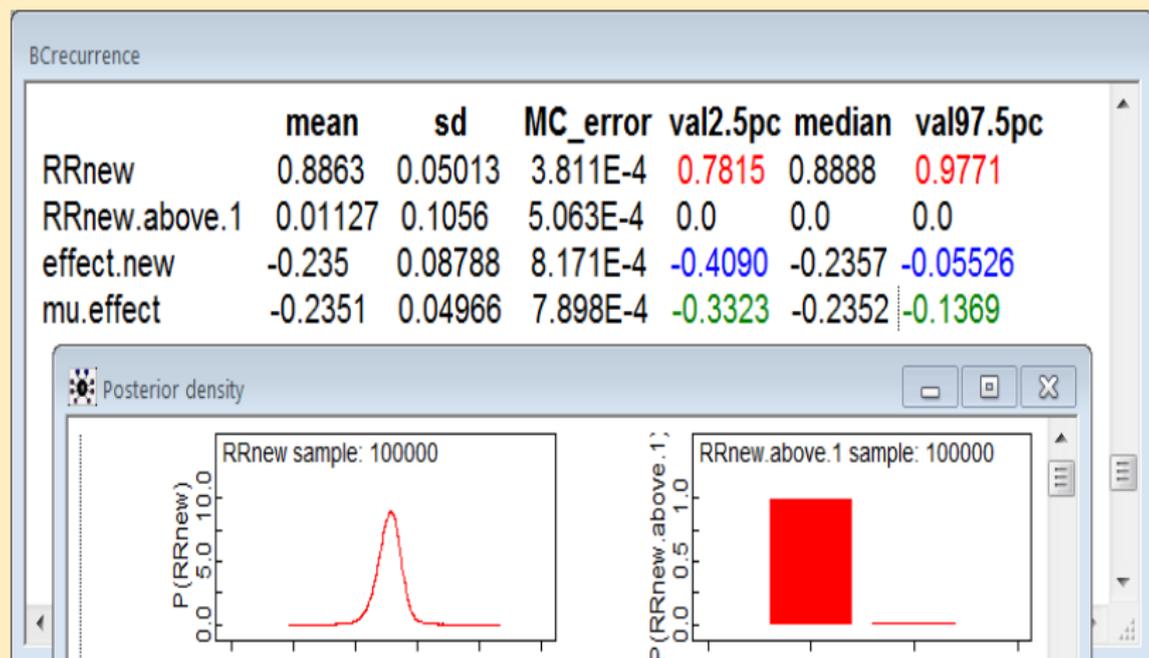


```
BCreurrence
Data
list(J=14,
      rT=c( 55,137,505,62,99,50,185,186,148,25,223,183,2,129),
      nT=c( 97,282,927,123,239,130,311,303,325,79,344,937,2,434),
      rC=c( 67,187,590,74,118,49,200,187,178,38,224,185,0,159),
      nC=c(100,306,915,140,236,107,319,307,325,86,350,936,8,449) )
.....
Inits
list(mu.effect=0,tau.effect=1,a.C=1,b.C=1)
list(mu.effect=-0.25,tau.effect=1,a.C=7,b.C=7)
```

Tamofixen and BC 4/5



Tamofixen and BC 5/5



Conclusions

- Both frequentist and Bayesian approaches have merits
- α -level significance testing and Bayes' factors are connected. More scrutiny needed about significance testing using raw p -values.
- MCMC and modern modeling/computing capabilities make Bayesian approach feasible and attractive
- Readily available WinBUGS/OpenBUGS software allows teaching Bayesian statistics at UG level
- Bayesian revolution: YES; All becoming Bayesians: Probably not; Taking ecumenical point of view: YES

Some References

- Berger, J. (2012). Reproducibility of Science: P-values and Multiplicity, Webinar SBSS.
- Casscells, W., Schoenberger, A. and Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, **299**, 999–1000.
- Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242.
- Fisher, R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of GB*, **33**, 503–513.
- Matthews, R. (1998). Bayesian Critique of Statistics in Health: The Great Health Hoax, Manuscript.
<http://www2.isye.gatech.edu/~brani/isyebayes/bank/pvalue.pdf>
- Vidakovic, B. (2011). *Statistics for Bioengineering Sciences*, Springer Verlag, 753pp.
- Wadman, M. (2013). NIH Mulls Rules for validating key results, *Nature*, **500**, p. 14.