Introduction to Bayesian Inference: Bayes Factor

Tuan V. Nguyen
Genetics Epidemiology of Osteoporosis Lab
Garvan Institute of Medical Research
Introduction to Bayes Factor

- Problem of reproducibility
- P value revisited
- Bayes Factor
- R functions
Lack of reproducibility

- **Observational studies**: out of 20 exciting findings, only 1 was replicated by NIH sponsored clinical trial

- **Basic research and animal models**: Bayer Healthcare reviewed 67 high-profile findings, <1/4 were replicated

- **Randomized clinical trials**: 49 famous trials (1990-2003):
  - 7 (16%) were contradicted by subsequent studies
  - 20 (44%) were replicated
  - 11 (24%) remained unchallenged

Ioannidis et al JAMA 2005; JNCI 2007
Reasons for lack of reproducibility

• Publication bias

• Rewards for "positive" results

• Experimental biases

• **Statistical biases**: Confounding; uncritical use of P-values

• **Bad statistics**: Failure to adjust for
  
  – multiple tests of hypothesis
  
  – multiple looks at data (eg data torture)
  
  – multiple statistical analyses (eg fishing expedition)
Multiple comparisons

- 10,000 compounds were screened for biological activity
- 500 passed the initial screen → studied in vitro
- 25 went to phase I trial (animal trial)
- 1 went to phase II trial (human trial)

"Basic research is like shooting an arrow in the air and, where it lands, painting a target" (Hosmer Adkins)
P-value: a revisit
Test of significance procedure

• Propose a hypothesis; state a null hypothesis \((H_0)\)
• Collect data (eg experiments)
• Calculate P-value

\[
P = \Pr(\text{data} \mid H_0 \text{ is true})
\]

If \(P < 0.05\), the finding is considered "significant"
Test your understanding

<table>
<thead>
<tr>
<th>P-value is ...</th>
<th>In terms of prob ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability that the null hypothesis (no effect) is true</td>
<td>$P(H_0)$</td>
</tr>
<tr>
<td>Probability of the data under H0</td>
<td>$P(x \mid H_0)$</td>
</tr>
<tr>
<td>Probability of H0 under the data</td>
<td>$P(H_0 \mid x)$</td>
</tr>
<tr>
<td>Probability that the data (or more extreme data) if H0 were true</td>
<td>$P(X \geq x \mid H_0)$</td>
</tr>
</tbody>
</table>

P-value = 0.04

Final results of HIV trial
- Infected with HIV
  - 8,197 given vaccine
  - 8,198 given placebo
  - 51 total infected
  - 74 total infected

Total number of people in trial (all HIV-negative men and women aged 18-30) = 15,395
P-VALUE
0.04932
Sample size and "significance"

Four hypothetical studies ...

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample size</th>
<th>Proportion</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>15 (0.75)</td>
<td>0.041</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>114 (0.57)</td>
<td>0.041</td>
</tr>
<tr>
<td>3</td>
<td>2000</td>
<td>1046 (0.525)</td>
<td>0.041</td>
</tr>
<tr>
<td>4</td>
<td>2,000,000</td>
<td>1,001,445 (0.5007)</td>
<td>0.041</td>
</tr>
</tbody>
</table>
Consider this study

**Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand**

Supachai Rerks-Ngarm, M.D., Punnee Pitisuttithum, M.D., D.T.M.H., Sorachai Nitayaphan, M.D., Ph.D., Jaranit Kaewkungwal, Ph.D., Joseph Chiu, M.D., Robert Paris, M.D., Nakorn Premsri, M.D., Chawetsan Namwat, M.D., Mark de Souza, Ph.D., Elizabeth Adams, M.D., Michael Benenson, M.D., Sanjay Gurunathan, M.D., Jim Tartaglia, Ph.D., John G. McNeil, M.D., Donald P. Francis, M.D., D.Sc., Donald Stablein, Ph.D., Deborah L. Birx, M.D., Suparnit Chunsuttiwat, M.D., Chirasak Khamboonruang, M.D., Prasert Thongcharoen, M.D., Ph.D., Merlin L. Robb, M.D., Nelson L. Michael, M.D., Ph.D., Prayura Kunasol, M.D., and Jerome H. Kim, M.D., for the MOPH–TAVEG Investigators*
RESULTS

In the intention-to-treat analysis involving 16,402 subjects, there was a trend toward the prevention of HIV-1 infection among the vaccine recipients, with a vaccine efficacy of 26.4% (95% confidence interval [CI], −4.0 to 47.9; \( P=0.08 \)). In the per-protocol analysis involving 12,542 subjects, the vaccine efficacy was 26.2% (95% CI, −13.3 to 51.9; \( P=0.16 \)). In the modified intention-to-treat analysis involving 16,395 subjects (with the exclusion of 7 subjects who were found to have had HIV-1 infection at baseline), the vaccine efficacy was 31.2% (95% CI, 1.1 to 52.1; \( P=0.04 \)). Vaccination did not affect the degree of viremia or the CD4+ T-cell count in subjects in whom HIV-1 infection was subsequently diagnosed.

CONCLUSIONS

This ALVAC-HIV and AIDSVAX B/E vaccine regimen may reduce the risk of HIV infection in a community-based population with largely heterosexual risk. Vaccination did not affect the viral load or CD4+ count in subjects with HIV infection. Although the results show only a modest benefit, they offer insight for future research. (ClinicalTrials.gov number, NCT00223080.)
How would you interpret this?

P-value = 0.04

Final results of HIV trial
- Infected with HIV

8,197 given vaccine
8,198 given placebo

Total number of people in trial (all HIV-negative men and women aged 18-30) = 16,395
set.seed(666)
n1 = 100;  n2 = 100;
mean1 = 103;  mean2 = 98;
sd = 15;
x1 = rnorm(n1, mean1, sd)
x2 = rnorm(n2, mean2, sd)
x = c(x1, x2)
group = c(rep("A", n1), rep("B", n2))
boxplot(x ~ group)
t.test(x ~ group)
> t.test(x ~ group)

Welch Two Sample t-test

data:  x by group
t = 2.3455, df = 196.13, p-value = 0.02
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.8583085 9.9252430
sample estimates:
mean in group A mean in group B  
  102.0000    96.6082
A surge of $p$-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too)

Joost CF de Winter and Dimitra Dodou

Department of BioMechanical Engineering, Delft University of Technology, Delft, The Netherlands
Figure 1 Number of papers reporting a positive result divided by the total number of papers examined (i.e., papers reporting a positive result + papers reporting a negative result) per publication year, for three scientific disciplines. The figure was created by graphically extracting the data shown in Fanelli’s (2012) figures. The dashed lines represent the results of a simple linear regression analysis.
reporting of significance in the United States, Asia, and Europe and found that the results are too inconsistent to draw conclusions on cross-cultural differences in significance reporting. We argue that the observed longitudinal trends are caused by negative factors, such as an increase of questionable research practices, but also by positive factors, such as an increase of quantitative research and structured reporting.
Multiple tests of hypothesis

- A "significant result" could be a chance finding

- Example:
  - Test 5 hypotheses
  - Each test, we have type I error 5%
  - Probability that one significant result by chance is $1 - (1 - 0.05)^5 = 23\%$

- In general, the prob of obtaining (by chance alone) at least 1 significant result in $k$ tests is $1 - (1 - a)^k$
Traditional statistical rules are a collection of principles and conventions to avoid errors over the long run; they do not tell us how likely our claims are to be true, nor do they easily apply to individual results.
STATISTICAL ERRORS

P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume.
Psychology journal bans P values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of Basic and Applied Social Psychology (BASP) announced that the journal would no longer publish papers containing P values, because the values were too often used to support lower-quality research.

Authors are still free to submit papers to BASP with P values and other statistical measures that form part of ‘null hypothesis significance testing’ (NHST), but the numbers will be removed before publication. “Basic and Applied Social Psychology just went science rogue and banned NHST from their journal. Awesome,” tweeted Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia. But Jan de Ruiter, a cognitive scientist at Bielefeld University in Germany, tweeted: “NHST is really problematic”, adding that banning all inferential statistics is “throwing away the baby with the p-value”.

Bayes Factor
Thomas Bayes
(c. 1702 – April 17, 1761)
7. Hinc & per 3. Art. 4. Si conjectura quæ probabiliter Eventus
cadit inter 0 & $\frac{p}{n}$ probabilitas quæ habeatur sola. est justa, si major quæ $N \times \frac{v}{v - p} x \frac{p}{n}$

$= N \times \frac{p}{n} + 2$, ubi si $h = b = q = 0$, aequatur quæ magis $N \times \frac{p}{n}$. 

8. Si $N$ est ratio cui $L_{x_{-1}}^{\frac{p}{n}} = \frac{p}{n} - \frac{n}{2n^2}$

$= \frac{p}{n}$ quando $n = 1$ est paulum major quam $\frac{v}{v - p} \times \frac{p}{n}$.

9. Sunt $v$ ratio peripheric circuli ad radius

$q \times \frac{p}{n}$ minor $\frac{v}{v - p}$ majus vero quam $N \times \frac{v}{v - p} \times \frac{p}{n}$.
Bayes Theorem: basic fact

\[ P(A | B) = \frac{P(B | A)P(A)}{P(B)} \]
Bayes theorem

\[ P(H \mid D) = P(D \mid H) \times P(H) \]
Bayes theorem -- again

In Bayesian inference, we need 2 key elements: data and prior information.

\[ P(D | H) \times P(H) = P(H | D) \]

D: Data
H: Hypothesis
Bayes theorem – in ratio terms

In Bayesian inference, we need 2 key elements: data and prior information

\[
\frac{P(D \mid H_1)}{P(D \mid H_0)} \times \frac{P(H_1)}{P(H_0)} = \frac{P(H_1 \mid D)}{P(H_0 \mid D)}
\]

D: Data
H: Hypothesis
A more objective way to measure evidence (no need prior information)

$$BF_{10} = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_0)}$$
Question

• Hypotheses
  – $H_0$: there is no effect
  – $H_1$: there is effect

• Do my data (D) favor $H_0$ or $H_1$?

• Answer: Bayes Factor
Bayes factor: A metric of evidence

\[ BF_{10} = \frac{P(data \mid H_1)}{P(data \mid H_0)} \]

<table>
<thead>
<tr>
<th>BF</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;1</td>
<td>Data support ( H_1 ) over ( H_0 )</td>
</tr>
<tr>
<td>&lt;1</td>
<td>Data support ( H_0 )</td>
</tr>
<tr>
<td>1</td>
<td>Data support neither ( H_0 ) nor ( H_1 )</td>
</tr>
</tbody>
</table>
# Interpretation of BF

<table>
<thead>
<tr>
<th>BF</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;100</td>
<td>Decisively favors $H_1$</td>
</tr>
<tr>
<td>30 to 100</td>
<td>Very strong evidence for $H_1$</td>
</tr>
<tr>
<td>10 to 30</td>
<td>Strong evidence for $H_1$</td>
</tr>
<tr>
<td>3 to 10</td>
<td>Substantial evidence for $H_1$</td>
</tr>
<tr>
<td>1 to 3</td>
<td>Weak evidence for $H_1$</td>
</tr>
<tr>
<td>1</td>
<td>No evidence</td>
</tr>
<tr>
<td>0.3 to 1</td>
<td>Weak evidence for $H_0$</td>
</tr>
<tr>
<td>0.1 to 0.3</td>
<td>Substantial evidence for $H_0$</td>
</tr>
<tr>
<td>0.03 to 0.1</td>
<td>Strong evidence for $H_0$</td>
</tr>
<tr>
<td>0.01 to 0.03</td>
<td>Very strong evidence for $H_0</td>
</tr>
<tr>
<td>&lt;0.01</td>
<td>Decisive evidence for $H_0$</td>
</tr>
</tbody>
</table>
Bayes Factor: technical stuff

- Let effect size be $\delta = \mu / \sigma$

$$f(D | \delta, \sigma^2) = \prod_{i=1}^{N} N(y_i | \sigma, \delta)$$

- Under $H_0$, $\delta = 0$, $\sigma$ unknown

- Under $H_1$, $\delta \neq 0$, $\sigma$ unknown, use Jeffrey-Zellner-Siow BF

$$BF = \frac{\left(1 + \frac{t^2}{N-1}\right)^{-N/2}}{(1 + N\tau^2)^{-1/2} \left(1 + \frac{t^2}{(1 + N\tau^2)(N-1)}\right)^{-N/2}}$$

Use R package BayesFactor $\rightarrow$ easier and faster
Bayes Factor for 2 groups

```r
set.seed(666)
n1 = 100;  n2 = 100;
mean1 = 103;  mean2 = 98;
sd = 15;
x1 = rnorm(n1, mean1, sd)
x2 = rnorm(n2, mean2, sd)
x = c(x1, x2)
group = c(rep("A", n1), rep("B", n2))
library(BayesFactor)
dat = data.frame(x, group)
bf = ttestBF(formula = x ~ group, data=dat)
bf
```
Interpretation: The data are 1.98 times more likely under the alternative hypothesis (ie there is an effect) than under the null hypothesis of no effect.
> chains = posterior(bf, iterations=10000)
> dif = chains[,2]
> hist(dif, col="blue", border="white")
> quantile(dif, c(0.025, 0.50, 0.975))

2.5%       50%     97.5%
0.6163938  5.0365395  9.4904096
Larger sample size studies provide stronger evidence?
BF_{01}, data, and sample size

Prior probability of hypothesis, BF, and posterior probability of hypothesis

- **Prior probability** = what do you think the hypothesis is true (before the study is conducted)
- **BF** = evidence
- **Posterior probability** = probability of hypothesis after seeing the evidence
Approximate BF

- Edwards (1963): relationship between BF and test statistic (for t and z test)

- Minimum BF

\[
\min BF = \exp\left(-0.5z^2\right)
\]
A nomogram for $P$ values

Leonhard Held

Abstract

**Background:** $P$ values are the most commonly used tool to measure evidence against a hypothesis. Several attempts have been made to transform $P$ values to minimum Bayes factors and minimum posterior probabilities of the hypothesis under consideration. However, the acceptance of such calibrations in clinical fields is low due to inexperience in interpreting Bayes factors and the need to specify a prior probability to derive a lower bound on the posterior probability.

**Methods:** I propose a graphical approach which easily translates any prior probability and $P$ value to minimum posterior probabilities. The approach allows to visually inspect the dependence of the minimum posterior probability on the prior probability of the null hypothesis. Likewise, the tool can be used to read off, for fixed posterior probability, the maximum prior probability compatible with a given $P$ value. The maximum $P$ value compatible with a given prior and posterior probability is also available.

**Results:** Use of the nomogram is illustrated based on results from a randomized trial for lung cancer patients comparing a new radiotherapy technique with conventional radiotherapy.

**Conclusion:** The graphical device proposed in this paper will enhance the understanding of $P$ values as measures of evidence among non-specialists.
Here, $x$ is the value of the $c_2$-test statistic which has given rise to the observed $P$ value. It can be easily shown that $BF$ decreases with increasing degrees-of-freedom. Perhaps more interestingly, $BF$ is equal to the BS lower bound for normal priors for $\nu = 1$, equals the SBB lower bound for $\nu = 2$, and is equal to the EL lower bound for $\nu \to \infty$. This illustrates that the range of lower bounds on the posterior probability given in Table 1 reflects a large variety of different tests and scenarios.

**Figure 1** A nomogram for $P$ values. The prior probability for the null hypothesis is located on the first axis, the observed $P$ value on the second axis, and the minimum posterior probability on the third axis.
Women’s Health Initiative Study (WHI), JAMA

“A low fat dietary pattern did not result in a statistically significant reduction in invasive breast cancer risk”

Data:

Invasive breast cancer HR 0.91 (0.83 – 1.01), P = 0.07

Breast cancer mortality HR 0.77 (0.48 – 1.22)
# A Comparison of Letrozole and Tamoxifen in Postmenopausal Women with Early Breast Cancer

The Breast International Group (BIG) 1-98 Collaborative Group*

<table>
<thead>
<tr>
<th>Variable</th>
<th>No. of Patients</th>
<th>Hazard Ratio</th>
<th>No. of Events</th>
<th>Hazard Ratio (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary end point</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease-free survival</td>
<td>8010</td>
<td></td>
<td>351</td>
<td>0.81 (0.70–0.93)</td>
<td>0.003</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;65 yr</td>
<td>5143</td>
<td></td>
<td>187</td>
<td>0.82 (0.67–0.99)</td>
<td>0.04</td>
</tr>
<tr>
<td>≥65 yr</td>
<td>2867</td>
<td></td>
<td>164</td>
<td>0.79 (0.64–0.97)</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Here, $x$ is the value of the $c^2$-test statistic which has given rise to the observed $P$ value. It can be easily shown that $BF$ decreases with increasing degrees-of-freedom. Perhaps more interestingly, $BF$ is equal to the BS lower bound for normal priors for $\nu = 1$, equals the SBB lower bound for $\nu = 2$, and is equal to the ELS lower bound for $\nu \to \infty$. This illustrates that the range of lower bounds on the posterior probability given in Table 1 reflects a large variety of different tests and scenarios.

A nomogram for $P$ values

The apparent complexity of the formulae presented in the previous section may be one of the reasons why the proposed calibration of $P$ values has not entered routine scientific research. I therefore suggest to adapt a graphical device, originally developed for diagnostic tests [13], to the setting outlined above. The original Fagan nomogram allows to visually determine the post-test probability for the null hypothesis is located on the first axis, the observed $P$ value on the second axis, and the minimum posterior probability on the third axis.

**Figure 1** A nomogram for $P$ values. The prior probability for the null hypothesis is located on the first axis, the observed $P$ value on the second axis, and the minimum posterior probability on the third axis.
Bayes Factor

• P-value
  – commonly misunderstood
  – affected by sample size and multiplicity of tests

• Bayes Factor: an **objective** metric of evidence
  – "hot" topic of research in genetics and clinical medicine
(Yet another) History of Life as we know it...

Homo Apriorius Pragmaticus
Homo Frequentistus
Homo Sapiens
Homo Bayesianus
Bayes factors in complex genetics

Stephen Sawyer

The past few years have seen tremendous progress in our understanding of the genetics underlying complex disease, with associated variants being identified in dozens of traits. Despite the fact that this growing body of empirical evidence unequivocally shows the necessity for extreme levels of significance and large samples sizes, the reasoning behind these requirements is not always appreciated. As genome-wide association studies reach the limits of their resolution in the search for rarer and weaker effects, the need for appropriate design and interpretation will become ever more important. If the genetic analysis of allow for I results of and the p described European

Bayesian statistical methods for genetic association studies

Matthew Stephens* and David J. Balding‡,§

Bayes Factors for Genome-Wide Association Studies: Comparison with $P$-values

Jon Wakefield*

Toward Evidence-Based Medical Statistics. 2: The Bayes Factor

Steven N. Goodman, MD, PhD
“Half of what doctors know is wrong. Unfortunately we don’t know which half.”

Quoted from the Dean of Yale Medical School, in “Medicine and Its Myths”,