Introduction to Bayesian Inference

Tuan V. Nguyen
Genetics Epidemiology of Osteoporosis Lab
Garvan Institute of Medical Research
Introduction to "Bayesian t-test"

• Bem's experiment
• A session of t-test
• A review of test of significance and P-value
• Introduction to Bayesian inference
Daryl R. Bem, Professor of Psychology, Cornell Univ.

Interested in "psi" phenomenon

Famous for Bem's experiment "Feeling the future" (J Personality and Soc Psychol 2011)
Bem's experiment

- 100 students (50 men, 50 women) exposed to a computer screen with two curtains (random order)
- One curtain has a picture behind it; another is blank
- The picture contains explicit erotic images of consensual couples
- Each participant was shown 36 pairs of curtains (random order)
- Participant was required to select a curtain that he/she felt had the picture behind it
“From the participants’ point of view, this procedure appears to test for clairvoyance. That is, they were told that a picture was hidden behind one of the curtains and their challenge was to guess correctly which curtain concealed the picture. In fact, however, neither the picture itself nor its left/right position was determined until after the participant recorded his or her guess, making the procedure a test of detecting a future event, that is, a test of precognition."
Bem's experiment

- Pictures-to-come were neutral / non-erotic: hit rate 50%
- Pictures-to-come were erotic: hit rate **53%** \((P=0.01)\)
- Among those with "stimulus seeking" character, the hit rate was 58%

**Strong evidence for psi?**
Reproducibility

• Many subsequent experiments could not replicate Bem’s findings

• Part of a larger problem: irreproducibility

• Overreliance on the P-value is part of the problem
t-test of two groups (experiment 1)

```r
set.seed(123)
n1=n2=50
x1 = rnorm(n1, 1.00, 0.12)
x2 = rnorm(n1, 1.03, 0.12)
boxplot(x1, x2, col=c("blue", "red"))
t.test(x1,x2)
```

data:  x1 and x2
t = -1.9766, df = 97.951, p-value = 0.0509
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0870538215  0.0001726889
sample estimates:
mean of x  mean of y
1.004128  1.047569
```r
set.seed(123)
n1=n2=45
x1 = rnorm(n1, 1.00, 0.12)
x2 = rnorm(n1, 1.03, 0.12)
boxplot(x1, x2, col=c("blue", "red"))
t.test(x1,x2)
```

data: x1 and x2
t = -1.1615, df = 86.792, p-value = 0.2486
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07148573  0.01875405
sample estimates:
mean of x mean of y
1.008043  1.034409
The “fragility” of effect: a first look
The “fragility” of effect quantified

Cohen's $d$: 0.2

Interpretation:
- Cohen's $U_0$: 57.93%
- % Overlap: 92.03%
- Probability of Superiority: 55.62%
- Number Needed to Treat: 16.51

http://rpsychologist.com/d3/cohend/
“An empirical assessment of 18 published papers of microarray studies showed that independent analysts could perfectly reproduce the results of only two of the studies…”
“This is one of medicine's dirty secrets: Most results, including those that appear in top-flight peer-reviewed journals, can't be reproduced”
Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah
Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the hypotheses considered.

Factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient

is characteristic of the field targets highly like
or searches for only one true relationship among
and millions of hypothesis to be postulated. Let us all be on the lookout for computational simula
circumscribed fields where only one true relation

PLOSMEDICINE AUGUST 2005 | VOLUME 2 | ISSUE 8 | E124
“Five of the seven largest published studies addressing cancer prognosis did not classify patients better than chance”
Lies, Damned Lies, and Medical Science

Much of what medical researchers conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors—to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.

By David H. Freedman
Test of significance: a revisit
Test of significance procedure

• State the **null hypothesis** \((H_0)\)

• Collect data, and calculate a test statistic (called \(D\))

• Calculate the probability of \(D\) (or more extreme \(D\)) if \(H_0\) is true:

\[ P(D \mid H_0) \]

• Report the exact P-value

This is subsequently called **P value**
How small P should be?

• Statistically significant

• Fisher (Statistical Methods for Research Workers, 1925)

“The value for which $P=0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty”
How small P should be?

... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."...

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 per cent point), or one in a hundred (the 1 per cent point). *Personally, the writer prefers to set a low standard of significance at the 5 per cent 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.*
Problems with P value

- Logical problem: difficult to understand
- Does not tell us about the effect size
- *Does not tell us the likelihood that a hypothesis is true*
- P value is highly dependent on sample size
- *Can be inflated in the presence of multiple tests of hypothesis*
False positive finding can be a problem

- Relationship between intakes of caffeine/coffee/tea and breast cancer overall and in multiple subgroups (50 tests)
- Overall, there was NO association, risk ratio close to 1
- But 4 “significant” differences were found in subgroup analysis:
  - coffee intake was linked to increased risk in those with benign breast disease (P = 0.08)
  - caffeine intake was linked to increased risk of estrogen/progesterone negative tumors and tumors larger than 2 cm (P = 0.02)
  - decaf coffee was linked to reduced risk of BC in postmenopausal hormone users (P = 0.02)

Distribution of p-values from the 50 tests

Likely chance findings!
Hallmarks of a chance finding

- Analyses are exploratory
- Many tests performed, but only a few are significant
- P-values are modest in size (between $P = 0.01$ and $P = 0.05$)
- The pattern of effect sizes is inconsistent
- The P-values are not adjusted for multiple comparisons
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

A SIMPLE idea underpins science: “trust, but verify”. Results should always be subject to
Bayesian idea
• If you go to see a doctor, and you have a +ve test, what do you want to know?

\[ P(\text{+ve test} \mid \text{diseased}) \]

\[ P(\text{diseased} \mid \text{+ve test}) \]

• If you do a study, and you found a “significant” result, what do you want to know?

\[ P(\text{data} \mid \text{no effect}) \]

\[ P(\text{effect} \mid \text{data}) \]
Thomas Bayes

- Rev Thomas Bayes (c. 1702 – April 17, 1761)
- Published only ONE paper (posthumously)
- Famous for Bayes' Theorem
Rev. Thomas Bayes

“An Essay towards solving a Problem in the Doctrine of Chances” published posthumously 1763

I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper...
Bayes Theorem: basic fact

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]
Bayes theorem: another version

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

- **D**: Data
- **H**: Hypothesis

Likelihood

Prior probability of hypothesis

Posterior probability of hypothesis
Bayes Theorem: still another version

\[ \text{Previous data} \times \text{Present data} = \text{Updated data} \]

\[ \text{Prior (prior distribution)} \times \text{Likelihood} = \text{Posterior distribution} \]
A diagnostic problem (hypothetical)

- Incidence of breast cancer (50+ yr): 1%
- Sensitivity = 90%
- Specificity = 95% (false positive = 5%)
- Question: if a woman has a +ve test result, what is her chance of getting cancer?
100,000 women

- Cancer 1000
- Not cancer 99000
100,000 women

Cancer 1000
- 900 +ve
- 100 -ve

Not cancer 99000
- 4950 +ve
- 90050 -ve

P(cancer | +ve) = 900 / (900+4950) = 15.4%
P(+ve | cancer) = 900/1000 = 90%
A research problem (hypothetical)

- Proportion of genetic variants associated with BMD: 0.1%
- Study power = 80%
- False positive = 5%
- Question: *if a significant finding (P = 0.05) is observed, what is the probability that the genetic variant is really associated with BMD*
A Bayesian interpretation of research finding

1000,000 variants

Real effect 1000

800 +ve

200 -ve

No real effect 999,000

49,950 +ve

P(real effect | +ve) = 800 / (800 + 49950) = 1.57%
## Research and Clinical Diagnosis

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence of disease</td>
<td>Null hypothesis</td>
</tr>
<tr>
<td>Presence of disease</td>
<td>Alternative hypothesis</td>
</tr>
<tr>
<td>False positive</td>
<td>Alpha (type I error)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Power (1 – beta)</td>
</tr>
<tr>
<td>Prevalence of disease</td>
<td>Prior probability of hypothesis</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>Posterior probability of hypothesis</td>
</tr>
</tbody>
</table>
Bayesian Analysis
In Bayesian inference, we need 2 key elements: data and prior information

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

\[D: Data\]
\[H: Hypothesis\]
Bayesian view of probability

• **Bayesian view** -- **Subjective**: *probability is an individual person's measure of belief that an event will occur*

• **Classical view** -- **Frequency**: the probability of an event is equal to the ratio of the number of "equipossibilities" (or equiprobable events) favourable to the event in question to the total number of relevant equipossibilities
Prior information on an effect size is required for any Bayesian analysis.

(Just like prevalence of disease in the case of diagnosis)

Prior information is expressed in probability distribution.
Understanding prior information

- **Parameter of interest:** Difference (D)
- **Bayes consider that D is a random variable** (not fixed variable)
- **Average diff = -0.047**

```r
set.seed(123)
n1=n2=50
x1 = rnorm(n1, 1.00, 0.12)
x2 = rnorm(n1, 1.03, 0.12)
boxplot(x1, x2, col=c("blue", "red"))
t.test(x1,x2)
----
t = -1.9766, df = 97.951, p-value = 0.0509
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 
-0.0870538215  0.0001726889
sample estimates:
mean of x mean of y
 1.004128  1.047569
```
Understanding prior information

• A random variable can be represented by a probability distribution.

• A probability distribution reflects YOUR THINKING about the parameter of interest.
Uniform prior

• You have NO IDEA of the effect size (D)
You think $D = 0$, with a variance 0.1, 0.05, 0.025, 0.010
Second element: actual data

- Data are also represented by a probability distribution.
- In t-test, the data are commonly represented by a Normal distribution, with mean $\mu$ and standard deviation $\sigma$:

$$X \sim N(\mu, \sigma)$$
Posterior distribution of effect sizes

\[
\text{Posterior probability of hypothesis} = \text{Likelihood} \times \text{Prior probability of hypothesis}
\]

Let’s try a Bayesian analysis online

http://sumsar.net/best_online/
Posterior distribution of effect sizes

Data group 1

Data here

Result

Click

Nbr of burn-in samples

20000

Nbr of samples

20000

Click to restart!

Data group 2

Trace Plot - Difference of Means

Distribution - Difference of Means

If the 95% Highest Density Interval does not include zero there is a credible difference!
More on posterior distribution

• What is the probability that $D < 0$? Answer: 97.4%
• What is the probability that $D < -0.01$? Answer: 93.3%
• What is the probability that $D \leq -0.02$? Answer: 85.3%
• What is the probability that $D \leq -0.05$? Answer: 38.2%
Bayesian approach

• Data $\times$ Prior information = Posterior information
• No need to calculate P value
• No need to work out sample size
• No need “critical values” of a test statistic
• No worry about “post-hoc test” and Bonferroni’s correction
• Probably will improve the reproducibility of results
set.seed(123)
n1=n2=50
x1 = rnorm(n1, 1.00, 0.12)
x2 = rnorm(n1, 1.03, 0.12)
dat = data.frame(bmd=c(x1,x2), group=rep(c("A", "B"), each=n1))
library(ggplot2)
ggplot(dat, aes(x=bmd, fill=group)) + geom_density(alpha = 0.5)