Bayesian Modeling

José G. Dias, ISCTE-IUL

Bayesian Modeling
Introduction

‘Traditional’ statistics

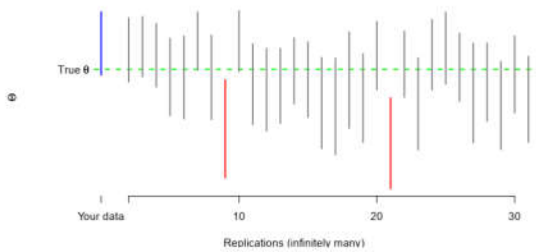
‘Traditional’ statistics

- Parameters (means, standard deviations, regression coefficients) fixed, but unknown
- The only thing that varies is the sample
  - what is a 95% CI?
  - if take 100 samples, 95 of them contain the true value
- ... so take a lot of samples
  - but not really ...
- Rely on asymptotics and CLT
- Let’s call it **frequentist statistics**

The maximum likelihood (MLE)

- ❑ Emphasis on data sample
  - ❑ overall probability multiply all the individual probabilities:
    $p(\boldsymbol{x}|\theta) \;=\; \Pi_i p(x_i|\theta)$
  - ❑ called the likelihood function (joint probability of all observations)
- ❑ MLE – value of the parameter that makes data you observed as likely as possible
- ❑ Intuitive results for standard distributions
  - ❑ normal: $\Sigma_i X_i/n$ for mean (μ) and $\Sigma_i (X_i - \bar{X})^2/n$ for variance (σ^2)
- ❑ More difficult for non-standard distributions

Confidence intervals



- ❑ In almost all **frequentist inference**, confidence intervals take the form $\hat{\theta} \; \pm const \times \; SE$, where the *standard error (SE)* quantifies the 'noise' in some estimate $\hat{\theta}$ of parameter θ
- ❑ The interval traps the truth in 95% of experiments. You ***have to imagine*** repeated experiments, many other datasets, not just the one you have to analyze

Hypothesis testing

nature > social selection > article

Published: 26 February 2015

Psychology journal bans *P* values

Chris Woolston

Nature **519**, 9 (2015) | Cite this article

4199 Accesses | **36** Citations | **1267** Altmetric | Metrics

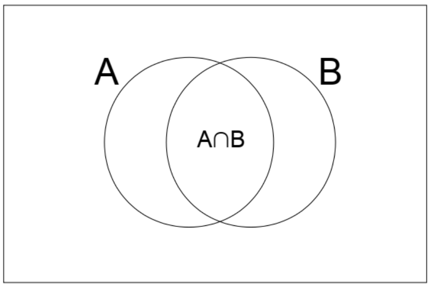
09 March 2015 This story originally asserted that "The closer to zero the P value gets, the greater the chance the null hypothesis is false." P values do not give the probability that a null hypothesis is false, they give the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It is by convention that smaller P values are interpreted as stronger evidence that the null hypothesis is false. The text has been changed to reflect this.

Bayes and his theorem

Rev Thomas Bayes (1701-1761)

Bayes theorem: Conditional Probability

Bayes theorem is about conditional probability

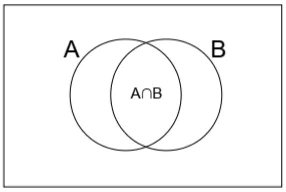


$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B)$$

$$P(A \cap B) = P(B|A)P(A)$$

Bayes theorem: Formula



$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} \\
&= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}
\end{aligned}
$$

Bayes theorem says that if we know $P[A|B]$ we can get $P[B|A]$

Bayes theorem: From events to RV

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

For **discrete random variables** $P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)}$

For **continuous random variables** $f_{x|y}(x|y) = \frac{f_{y|x}(y|x)f_x(x)}{\int f_{y|x}(y|t)f_x(t)\,dt}$

Bayes theorem: History

- 1763 Rev Thomas Bayes gave the first description of the theorem in "An essay toward solving a problem in the doctrine of chance"
- 1774 Pierre Simon Laplace gave more elaborate version of Bayes theorem
- von Mise gave a rigorous proof of Bayes theorem
- What do we mean by ***frequentist/classical statistics***?

The concept of probability

Probabilities: Until now (frequentist)

- Probabilities are **numerical quantities** that measure uncertainty
- Example: "The probability that an even number comes up on a toss of a dice equals 1/2."
- **Justification 1**: Symmetry or exchangeability arguments (equal probability of events):
  p(even) = number of evens rolled / number of possible results
- **Justification 2**: Frequency argument:
  p(even) = long (infinite) run relative frequency

Probabilities: Measures of uncertainty

- **Randomness** creates uncertainty and we already do it in common speech... (often randomness and probability are used exchangeably in speech)
- **Coherence**: probabilities principles of basic axioms of probability theory, best known been the one proposed by Kolmogorov:
  - $0 \leq P(A) \leq 1$
  - if *A* is subset or equal to *B*, then $P(A) \leq P(B)$
  - ...

Probabilities: In practice

- Long (infinite) sequence of independent rolls of the dice
- These involve hypotheticals: physical independence, infinite sequence of rolls, equally likely, mathematical idealizations
- How would either of these justifications apply to
  - If we only roll dice once?
  - What's the probability that Portuguese (men) soccer team wins the next World Cup?
  - How to check the relation between education level and economic development using data from all countries?

Philosophies

Back to Bayesian simplicity

- Bayesian inference can be made transparent
- Inventor of Bayesian inference Pierre-Simon, marquis de Laplace (1749-1827) in *Common sense reduced to computation*

Interpreting Bayesian probabilities

'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. Essentially the notion can only be described by reference to instances where it is used. ***It is intended to express a kind of relation between data and consequence that habitually arises in science and in everyday life****, and the reader should be able to recognize the relation from examples of the circumstances when it arises.*

— Sir Harold Jeffreys, Scientific Inference

It’s all about this!

Probabilities are foundational concept! Probability is not really about numbers; it is about the structure of reasoning.

Glenn Shafer, quoted in Pearl (1988, p. 77)

Different terms for probability have been in use:

- Beliefs
- Credibility
- Plausibilities
- Subjective

Conceptions of probability

Frequentist/classical statistics

- ❑ Probability is long-run relative frequency
- ❑ Probability is a ***property of the world***

Bayesian

- ❑ Probability is a degree of subjective belief
- ❑ A fundamental measure of uncertainty that follow rules probability theory
- ❑ Probability is a property of ***your state of knowledge***

General approach to Bayesian Modeling

A ***Bayesianly justifiable*** analysis is one that "treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes theorem."

Rubin (1984)

Nature of the parameters (θ)

- To the frequentist, it is an unknown constant
- To the Bayesian, since we do not know the value of the parameter, it is a random variable
- That uncertainty is given by a probability distribution, $\pi(\theta)$, that is called the ***prior* distribution**
- Prior because it represents the statistician's belief about θ ***before*** observing the data. Parameters vary randomly (e.g., normal, binomial, Poisson).

Prior distributions

Priors: Subjectivity

- **Subjectivity** is the most frequent objection to Bayesian methods
- The prior distribution influences the result
- Two scientists may arrive at different conclusions using the same data, ***based on the same statistical analysis***
- The influence of the prior goes to zero as the sample size increases (in general)
- A lot of skepticism about priors (choice, assumptions)
- Ideally, based on real prior information

Priors: Specifying priors

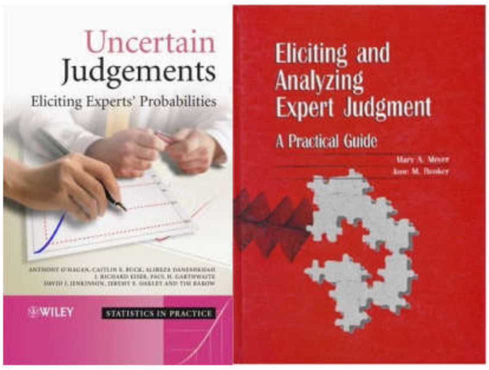
- Priors come from all data *external* to the current study, i.e. everything else
- Let's be frank, we have prior beliefs
  - e.g., correlations or standardized slopes are likely to take any value?
- The same happens with other parts of the model:
  - The distribution of the data normal or Student-t (likelihood)?
  - What to do with missing data or outliers?
  - …

Priors: business as usual

- **Step 1**: Choose a family of distributions (e.g., *N*, *t*, *Beta*)
  - Consistency with substantive knowledge
  - Modeling flexibility/richness
  - Computational ease
  - Ease to interpret
- **Step 2**: Choose a member of the family (e.g., *N*(0,1))
  - Ignorance
  - Theory
  - Previous research
  - Subjective assessment by experts

Priors: (subjective) elicitations

- **A simple possibility**
  - list plausible values
  - weight them by how probable we think they are
  - convert the weights to a probability distribution that sums to one by dividing through by the sum of the weights



- **Subjective ≠ Arbitrary**
  - Can do sensitivity analyses, say, with diffuse priors

Bayesian updating

Bayes theorem as an updating system

- combine what you know with what you observe to update your knowledge

prior + data = update

- combine data **likelihood** ($f(x|\theta)$) with **prior** expectation ($\pi(\theta)$) to update inference on parameters called **posterior** ($\pi(\theta|x)$)

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|t)\pi(t)\,dt}$$

$$\propto f(x|\theta)\pi(\theta)$$

- $f(x)$ drops out as normalizing term (same form, only size changes)

Bayes theorem: Accumulation of evidence

- In other words, if $\pi(\theta)$ approximates our beliefs, then $\pi(\theta|x)$ is optimal to what our posterior (after we have new data) beliefs about θ should be
- Bayes can be used to explore how beliefs should be up-dated given data by someone with no prior information (flat prior)
- **Today's posterior is tomorrow's prior**: as the notation indicates, $\pi(\theta|x)$ might be the prior for the next experiment

Controversies: Bayesian vs. Frequentist

Controversies…

A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule

Stephen Senn, Statistician

The Bayesian approach is "the explicit use of external evidence in the design, monitoring, analysis, interpretation and reporting of a scientific investigation"

Spiegelhalter (2004)

Still…

Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician's toolkit.. . . I am discomforted by this "inferential schizophrenia." Since the Bayesian (B) and frequentist (F) philosophies can differ even on simple problems, at some point decisions seem needed as to which is right. ***I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject****… An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that calibrated Bayes. . . captures the strengths of both approaches and provides a roadmap for future advances.*

Roderick Little, ASA President's Address (2005)

A very long long way …

The ***250-year debate between Bayesians and frequentists*** *is unusual among philosophical arguments in actually* ***having important practical consequences****... The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there's only one way of doing physics but there seems to be at least two ways to do statistics, and they don't always give the same answers… Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist.* ***What's going to happen in the 21st Century?****... I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning.. . .*

Brad Efron, ASA President (2005)

The future!

The only good statistics is Bayesian Statistics

Dennis Lindley in

The Future of Statistics: A Bayesian 21st Century (1975)

A very simple example

Inference about the mean of the normal distribution

- Population is normal: $X \sim N(\mu, \sigma^2)$, with σ^2 is known
- The **likelihood** is $\Pi_i N(X_i|\mu, \sigma^2) \;=\; N(\bar{X}|\,\mu, \sigma^2/n)$
- Other source of information about μ encoded in the **prior** distribution, $\mu \sim N(\mu_\mu, \sigma_\mu^2)$
- Conjugate prior for mean of a normal distribution when the variance is known is also a normal distribution, the **posterior** mean is

$$\mu_{\mu|X} = \frac{\sigma_\mu^{-2}\mu_\mu + n\sigma_\mu^{-2}\bar{X}}{\sigma_\mu^{-2} + n\sigma_\mu^{-2}}$$

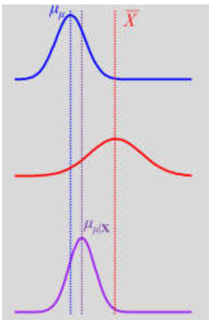
Bayesian weighting

- Posterior is a weighted combination of the likelihood and the prior
- For small data sets, prior is influential
- What about confidence intervals?

Prior: $N(\mu_\mu, \sigma_\mu^2)$

Likelihood: $N(\bar{X}, \sigma_\mu^2/n)$

Posterior: $N(\mu_{\mu|X}, \sigma_{\mu|X}^2)$



$$\mu_{\mu|X} = \frac{\sigma_\mu^{-2}\mu_\mu + n\sigma_\mu^{-2}\bar{X}}{\sigma_\mu^{-2} + n\sigma_\mu^{-2}}$$

Bayesian modeling process

Bayesian computing: No free meals (Efron, 1986)

- This was a very simple example as normal likelihood coupled with a normal prior that turns out to be a normal posterior distribution
- In general, specifying a prior distribution and combining it with the data likelihood will complicate our lives (**computationally intensive**)
- But now just the "subjectivity" remains against the Bayesian approach

Advantages of Bayesian modelling

- A **natural and coherent approach** that synthesizing information over time: characterization as “optimal” by psychology, neuroscience
- **Flexible computational framework** for model estimation, selection and validation that can be adapted to complex situations with no obvious (non-Bayesian) solution (e.g., complex statistical model, estimation of rare events)
- **Efficient approach**, using all available information in a parsimonious way
- Captures additional **uncertainty** in predictions by allowing parameter estimates to vary (e.g., predictions for missing data and predictions)

Modeling process: General approach

1. Exploratory Data Analysis (EDA)
2. Formal model building
    1. Likelihood based methods
    2. Bayesian approaches (built on top of likelihood based approaches)
3. Model checking and validation
4. Model refinement
5. Quantification of uncertainty when stating conclusions

Modeling process: the Bayesian approach

- Let's assume you have research questions, relevant data, and know the nature of the data
- Set the **full probability model**: a joint probability distribution for all observed and unobserved variables that reflect knowledge and how data were collected:

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$$

- Condition on data, **obtain the posterior distribution**:

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x)$$

  Tools: analytic, grid approximation, Markov chain Monte Carlo (Metropolis-Hastings, Gibbs sampling, Hamiltonian), etc.
- **Summarize** results and **examine model fit**, sensitivity of assumptions, reasonable conclusions, etc.

Conclusion

Summary

- **Probability**
  - as generalized logic: ***state of knowledge***, including the model of the situation
  - a way to capture and communicate uncertainty
- **Bayes theorem**
  - a mechanism for probability-based reasoning under uncertainty
  - combine information from prior and data