pumas

Bayesian Workshop: How to use Bayesian methods in Pumas

Jose Storopoli and Mohamed Tarek `{jose.storopoli,mohamed}@pumas.ai`

PumasAI

Outline

1. Pumas

2. Bayesian Statistics
   - 2.1 Recommended References
   - 2.2 What is Bayesian Statistics?
   - 2.3 Frequentist versus Bayesian
   - 2.4 Bayesian Statistics

3. Priors
   - 3.1 Recommended References
   - 3.2 The subjectivity of the Prior
   - 3.3 Types of Priors

4. References

What is Pumas?

Pumas (**P**harmace**U**tical **M**odeling **A**nd **S**imulation) [1] is a suite of tools to perform quantitative analytics of various kinds across the horizontal of pharmaceutical drug development. The purpose of this framework is to bring efficient implementations of all aspects of the analytics in this domain under one cohesive package.

Pumas Features

Pumas 2.3 currently includes:

- Non-compartmental Analysis
- Specification of Nonlinear Mixed Effects (NLME) Models
- Simulation of NLME model using differential equations or analytical solutions
- Deep control over the differential equation solvers for high efficiency
- Estimation of NLME parameters via Maximum Likelihood, Expectation Maximization and Bayesian methods
- Parallelization capabilities for both simulation and estimation
- Mixed analytical and numerical problems
- Simulation and estimation diagnostics for model post-processing
- Interactive model exploration and diagnostics tools through webapps
- Automated report generation for models and non-compartmental analysis
- Global and local sensitivity analysis routines for multi-scale models
- Bioequivalence analysis
- Optimal design of experiments

Bayesian Statistics - Recommended References

- Gelman et al. [2] - Chapter 1: Probability and inference
- McElreath [3] - Chapter 1: The Golem of Prague
- Gelman, Hill, and Vehtari [4] - Chapter 3: Some basic methods in mathematics and probability
- Khan and Rue [5]
- **Probability**:
  - A great textbook - Bertsekas and Tsitsiklis [6]
  - Also a great textbook (skip the frequentist part)- Dekking et al. [7]
  - Bayesian point-of-view and also a philosophical approach- Jaynes [8]
  - Bayesian point-of-view with a simple and playful approach - Kurt [9]
  - Philosophical approach not so focused on mathematical rigor - Diaconis and Skyrms [10]

What is Bayesian Statistics?

Bayesian statistics is a **data analysis approach based on Bayes' theorem** where available knowledge about the parameters of a statistical model is updated with the information of observed data. [2]. Previous knowledge is expressed as a **prior** distribution and combined with the observed data in the form of a **likelihood** function to generate a **posterior** distribution. The posterior can also be used to make predictions about future events.

What changes from Frequentist Statistics?

- **Domain knowledge**:
  - You can incorporate knowledge and insights from previous studies using prior distributions on parameters
- **Epistemic uncertainty**:
  - You can quantify the epistemic uncertainty in the model parameters' values
  - Model identifiability not necessary
  - Works for small and large sample sizes
  - No Gaussian assumptions
- **Conceptually simpler and more general**:
  - Uses probability theory instead of *ad-hoc* methods
  - No p-values, p-hacking and *ad-hoc* assumptions in hypothesis tests

A little bit more formal

- Bayesian Statistics uses probabilistic statements:
  - one or more parameters θ
  - unobserved data $\tilde{y}$
- These statements are conditioned on the observed values of y:
  - $P(\theta \mid y)$
  - $P(\tilde{y} \mid y)$
- We also, implicitly, conditioned on the observed data from any covariate x
- Generally, we are interested in:
  - expected response of a new subject to a drug, e.g. $\mathrm{E}[\hat{y} \mid y]$
  - the probability of drug effect is higher than zero, e.g. $P(\theta > 0 \mid y) \geq 0.95$

Definition of Bayesian Statistics

Definition (Bayesian Statistics)

*The use of Bayes theorem as the procedure to **estimate parameters of interest** θ **or unobserved data** $\tilde{y}$. [2]*

Probability Interpretations

- **Objective** - frequency in the long run for an event:
  - $P(\text{rain}) = \frac{\text{days that rained}}{\text{total days}}$
  - $P(\text{me being elected president}) = 0$ (never occurred)

- **Subjective** - degrees of belief in an event:
  - $P(\text{rain}) =$ degree of belief that will rain
  - $P(\text{me being elected president}) = 10^{-10}$ (highly unlikely)

What is Probability?

Definition (Probability)

We define A is an event and $P(A)$ the probability of event A. $P(A)$ has to be between 0 and 1, where higher values defines higher probability of A happening.

$$
\begin{aligned}
P(A) &\in \mathbb{R} \\
P(A) &\in [0, 1] \\
0 \leq\, &P(A) \leq 1
\end{aligned}
$$

Probability Axioms[i]

- **Non-negativity**: For every A:

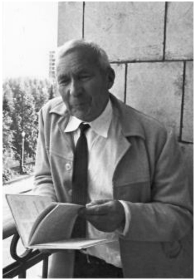
$$P(A) \geq 0$$

- **Additivity**: For every two *mutually exclusive* A and B:

$$P(A) = 1 - P(B) \text{ and } P(B) = 1 - P(A)$$

- **Normalization**: The probability of all possible events $A_1, A_2, \ldots$ must sum up to 1:

$$\sum_{n \in \mathbb{N}} P(A_n) = 1$$



[i] Kolmogorov [11]

Sample Space[ii]

- Discrete

$$\Theta = \{1, 2, \ldots\}$$

- Continuous

$$\Theta \in (-\infty, \infty)$$

[ii] θ domain can be general, not restricted to these domains.

Discrete Sample Space

8 planets in our solar system:

- Mercury - ☿
- Venus - ♀
- Earth - ♁
- Mars ♂
- Jupiter - ♃
- Saturn ♄
- Uranus - ⛢
- Neptune ⯉

Discrete Sample Space[iii]

The planet has a magnetic field

$\theta \in E_1$

The planet has moon(s)

$\theta \in E_2$

The planet has a magnetic field *and* moon(s)

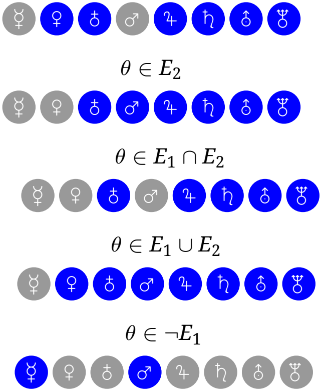
$\theta \in E_1 \cap E_2$

The planet has a magnetic field *or* moon(s)

$\theta \in E_1 \cup E_2$

The planet does *not* have a magnetic field

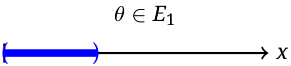
$\theta \in \neg E_1$



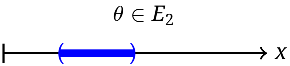
[iii] figure adapted from Michael Betancourt (CC-BY-SA-4.0)

Continuous Sample Space[iv]

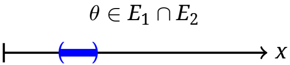
The distance is less than five centimeters



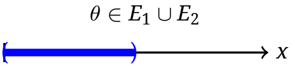
The distance is between three and seven centimeters



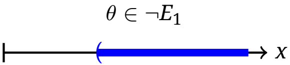
The distance is less than five centimeters
and e between three and seven centimeters



The distance is less than five centimeters
or between three and seven centimeters



The distance is *not* less than five centimeters



[iv] figure adapted from Michael Betancourt (CC-BY-SA-4.0)

Discrete versus Continuous Parameters

Parameters can be continuous, such as: age, height, weight etc.

All probability rules and axioms are valid also for continuous parameters.

The only thing we have to do is to change all sums $\sum$ for integrals $\int$.

Conditional Probability

Definition (Conditional Probability)

Probability of an event occurring in case another has occurred or not.

The notation we use is $P(A \mid B)$, that read as "the probability of observing A given that we already observed B".

$$P(A \mid B) = \frac{\textit{number of elements in } A \textit{ and } B}{\textit{number of elements in } B}$$

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

assuming that $P(B) > 0$.

Caution! Not always $P(A \mid B) = P(B \mid A)$

In the previous example we have the symmetry $P(A \mid K) = P(K \mid A)$, **but not always this is true**[v]

Example (The Pope is catholic)

- $P(\text{pope})$: Probability of some random person being the Pope, something really small, 1 in 8 billion $\left(\frac{1}{8 \cdot 10^9}\right)$
- $P(\text{catholic})$: Probability of some random person being catholic, 1.34 billion in 8 billion $\left(\frac{1.34}{8} \approx 0.17\right)$
- $P(\text{catholic} \mid \text{pope})$: Probability of the Pope being catholic $\left(\frac{999}{1000} = 0.999\right)$
- $P(\text{pope} \mid \text{catholic})$: Probability of a catholic person being the Pope $\left(\frac{1}{1.34 \cdot 10^9} \cdot 0.999 \approx 7.46 \cdot 10^{-10}\right)$
- **Hence**: $P(\text{catholic} \mid \text{pope}) \neq P(\text{pope} \mid \text{catholic})$

[v] More specific, if the basal rates $P(A)$ and $P(B)$ aren't equal, the symmetry is broken $P(A \mid B) \neq P(B \mid A)$

Joint Probability

Definition (Joint Probability)

Probability of two or more events occurring.

The notation we use
is $P(A, B)$, that read as "the probability of observing A and also observing B".

$$
\begin{aligned}
P(A, B) &= \textit{number of elements in } A \textit{ or } B \\
P(A, B) &= P(A \cup B) \\
P(A, B) &= P(B, A)
\end{aligned}
$$

Product Rule of Probability[vi]

Definition (Product Rule)

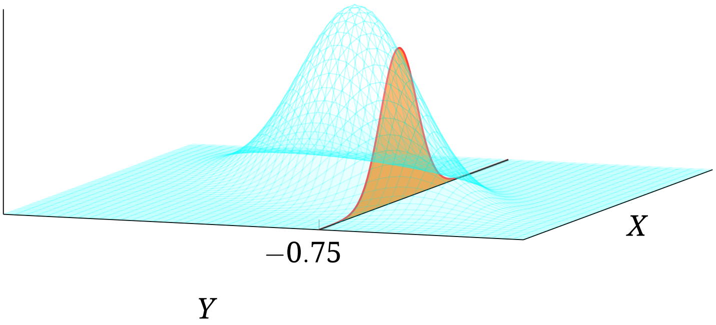
We can decompose a joint probability $P(A, B)$ *into the product of two probabilities:*

$$
\begin{aligned}
P(A, B) &= P(B, A) \\
P(A) \cdot P(B \mid A) &= P(B) \cdot P(A \mid B)
\end{aligned}
$$

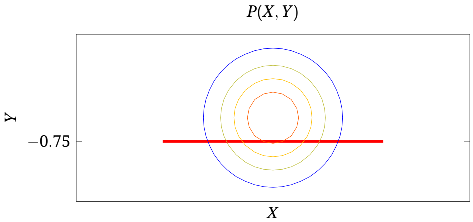
[vi] also called the Product Rule of Probability.

Visualization of Joint Probability versus Conditional Probability

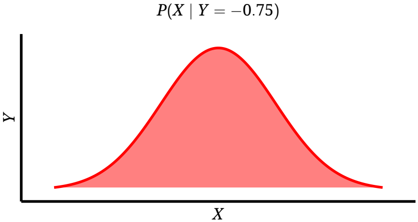
$P(X, Y)$ versus $P(X \mid Y = -0.75)$



Visualization of Joint Probability versus Conditional Probability



$P(X, Y)$



$P(X \mid Y = -0.75)$

Who was Thomas Bayes?

- **Thomas Bayes** (1701 - 1761) was a statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.
- Bayes never published what would become his most famous accomplishment; his notes were edited and published posthumously by his friend **Richard Price**.
- The theorem official name is **Bayes-Price-Laplace**, because **Bayes** was the first to discover, **Price** got his notes, transcribed into mathematical notation, and read to the Royal Society of London, and **Laplace** independently rediscovered the theorem without having previous contact in the end of the XVIII century in France while using probability for statistical inference with census data in the Napoleonic era.

Bayes Theorem

Theorem (Bayes)

Tells us how to "invert" conditional probability:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

Bayes’ Theorem Proof

Remember the following probability identity:

$$P(A, B) = P(B, A)$$
$$P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

OK, now divide everything by $P(B)$:

$$\frac{P(A) \cdot P(B \mid A)}{P(B)} = \frac{P(B) \cdot \quad P(A \mid B)}{P(B)}$$

$$\frac{P(A) \cdot P(B \mid A)}{P(B)} = P(A \mid B)$$
$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

Another Probability Textbook Classic[vii]

Example (Breast Cancer)

How accurate is a **breast cancer** test?

- 1% of women have **breast cancer** (Prevalence)
- 80% of mammograms detect **breast cancer** (True Positive)
- 9.6% of mammograms detect **breast cancer** when there is no incidence (False Positive)

$$P(C \mid +) = \frac{P(+ \mid C) \cdot P(C)}{P(+)}$$

$$P(C \mid +) = \frac{P(+ \mid C) \cdot P(C)}{P(+ \mid C) \cdot P(C) + P(+ \mid \neg C) \cdot P(\neg C)}$$

$$P(C \mid +) = \frac{0.8 \cdot 0.01}{0.8 \cdot 0.01 + 0.096 \cdot 0.99}$$

$$P(C \mid +) \approx 0.0776$$

[vii] Adapted from: Yudkowski - *An Intuitive Explanation of Bayes' Theorem*.

Why Bayes' Theorem is Important?

Idea (We can Invert the Conditional Probability)

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis) \cdot P(hypothesis)}{P(data)}$$

But isn't this the p-value? **NO!**

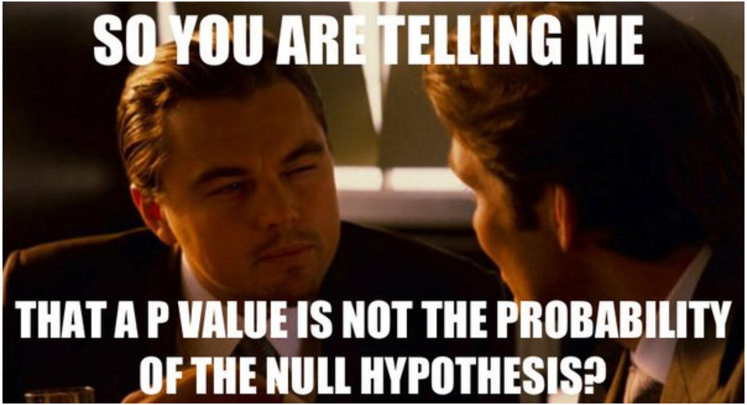
What are p-values?

Definition (p-value)

p-value is the probability of obtaining results at least as extreme as the observed, given that the null hypothesis H_0 is true:

$$P(D \mid H_0)$$

What p-value is **not**!


SO YOU ARE TELLING ME
THAT A P VALUE IS NOT THE PROBABILITY OF THE NULL HYPOTHESIS?

What p-value is **not**!

- **p-value is not the probability of the null hypothesis** - Infamous confusion between $P(D \mid H_0)$ and $P(H_0 \mid D)$. To get $P(H_0 \mid D)$ you need Bayesian statistics.

- **p-value is not the probability of data being generated at random** - No! We haven't stated nothing about randomness.

- **p-value measures the effect size of a statistical test** - Also no... p-value does not say anything about effect sizes. Just about if the observed data diverge of the expected under the null hypothesis. Besides, p-values can be hacked in several ways [12].

The relationship between p-value and H_0

To find out about any p-value, **find out what H_0 is behind it**. It's definition will never change, since it is always $P(D \mid H_0)$:

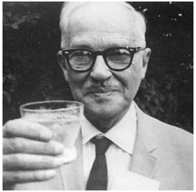
- ***t*-test**: $P(D \mid \text{the difference between the groups is zero})$
- **ANOVA**: $P(D \mid \text{there is no difference between groups})$
- **Regression**: $P(D \mid \text{coefficient has a null value})$
- **Shapiro-Wilk**: $P(D \mid \text{population is distributed as a Normal distribution})$

What are Confidence Intervals?

Definition (Confidence Intervals)

> *A confidence interval of X% for a parameter is an interval* (a, b) *generated by a repeated sampling procedure has probability X% of containing the true value of the parameter, for all possible values of the parameter.*

Neyman [13] (the "father" of confidence intervals)



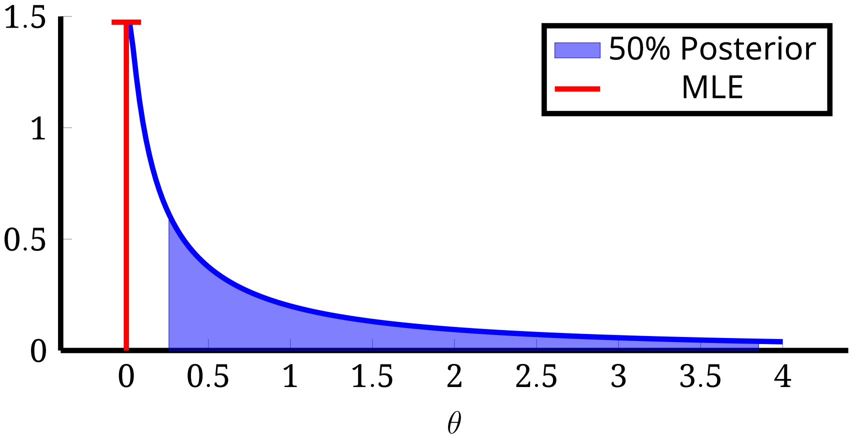
What are Confidence Intervals?

Example (Confidence Intervals of a Public Policy Analysis)

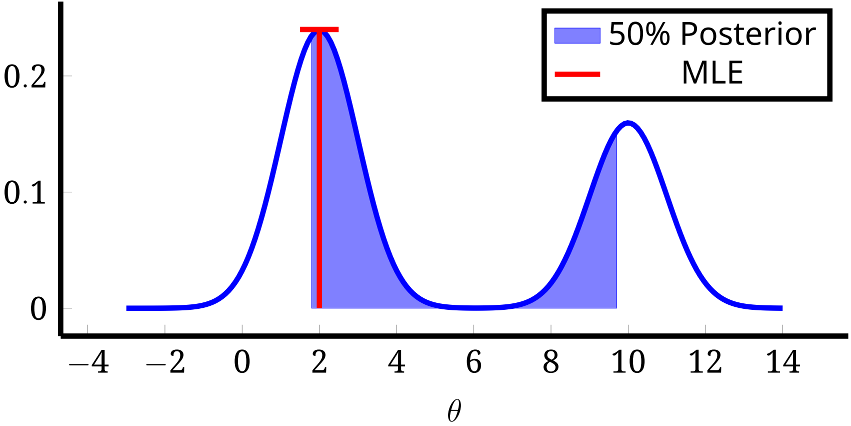
Say you performed a statistical analysis to compare the efficacy of a public policy between two groups and you obtain a difference between the mean of these groups. You can express this difference as a confidence interval. Often we choose 95% confidence. This means that **95 studies out of 100**, that uses the **same sample size and target population**, performing the **same statistical test**, will expect to find a result of the mean difference between groups inside the confidence interval.

Doesn’t say anything about you **target population**, but about you **sample** in an insane process of **infinite sampling** …

Confidence Intervals versus Posterior Intervals



Confidence Intervals versus Posterior Intervals



But why I never see stats without p-values?

We cannot understand p-values if we do no not comprehend its origins and historical trajectory. The first mention of p-values was made by the statistician Ronald Fischer in 1925 [14]:

> *[p-value is a] measure of evidence against the null hypothesis*

- To quantify the strength of the evidence against the null hypothesis, Fisher defended "$p < 0.05$ as the standard level to conclude that there is evidence against the tested hypothesis"
- "We should not be off-track if we draw a conventional line at 0.05"

$p = 0.06$

- Since p-value is a probability, it is also a continuous measure.
- There is no reason for us to differentiate $p = 0.049$ against $p = 0.051$.
- Robert Rosenthal, a psychologist said “surely, God loves the .06 nearly as much as the .05” [15].

But why I never heard about Bayesian statistics?[viii]

> *. . . it will be sufficient . . . to reaffirm my personal conviction . . . that the theory of inverse probability is founded upon an error, and must be wholly rejected.*

Fisher [14]

[viii] *inverse probability* was how Bayes' theorem was called in the beginning of the 20th century

Inside every nonBayesian, there is a Bayesian struggling to get out[ix]

- In his final year of life, Fisher published a paper [16] examining the possibilities of Bayesian methods, but with the prior probabilities being determined experimentally.
- Some authors speculate [8] that if Fisher were alive today, he would probably be a Bayesian.

[ix] quote from Dennis Lindley

Bayes' Theorem as an Inference Engine

Now that you know what is probability and Bayes' theorem, I will propose the following:

$$\underbrace{P(\theta \mid y)}_{\text{Posterior}} = \frac{\overbrace{P(y \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(y)}_{\text{Normalizing Constant}}}$$

- θ – parameter(s) of interest
- y – observed data
- **Priori**: prior probability of the parameter(s) value(s)
- **Likelihood**: probability of the observed data given the parameter(s) value(s)
- **Posterior**: posterior probability of the parameter(s) value(s) after we observed data y
- **Normalizing Constant**[x]: $P(y)$ does not make any intuitive sense. This probability is transformed and can be interpreted as something that only exists so that the result $P(y \mid \theta)P(\theta)$ be constrained between 0 e 1 – a valid probability.

[x] sometimes also called *evidence*.

Bayes’ Theorem as an Inference Engine

Bayesian statistics allows us to **quantify directly the uncertainty** related to the value of one or more parameters of our model given the observed data. This is the **main feature** of Bayesian statistics, since we are estimating directly $P(\theta \mid y)$ using Bayes’ theorem. The resulting estimate is totally intuitive: simply quantifies the uncertainty that we have about the value of one or more parameters given the data, model assumptions (likelihood) and the prior probability of these parameter’s values.

Bayesian vs Frequentist Stats

	Bayesian Statistics	**Frequentist Statistics**
Data	Fixed — Non-random	Uncertain — Random
Parameters	Uncertain — Random	Fixed — Non-random
Inference	Uncertainty regarding the parameter value	Uncertainty regarding the sampling process from an infinite population
Probability	Subjective[xi]	Objective (but with several model assumptions)
Uncertainty	Posterior Interval — $P(\theta \mid y)$	Confidence Interval — $P(y \mid \theta)$

[xi] with highly informative priors.

Priors and Posteriors - Recommended References

- Gelman et al. [2]:
  - Chapter 2: Single-parameter models
  - Chapter 3: Introduction to multiparameter models

- McElreath [3] - Chapter 4: Geocentric Models
- Gelman, Hill, and Vehtari [4]:
  - Chapter 9, Section 9.3: Prior information and Bayesian synthesis
  - Chapter 9, Section 9.5: Uniform, weakly informative, and informative priors in regression

- van de Schoot et al. [17]

Prior Probability

Bayesian statistics is characterized by the use of prior information as the prior probability $P(\theta)$, often just prior:

$$\underbrace{P(\theta \mid y)}_{\text{Posterior}} = \frac{\overbrace{P(y \mid \theta)}^{\text{Likelihood}} \cdot \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(y)}_{\text{Normalizing Constant}}}$$

The subjectivity of the Prior

- Many critics to Bayesian statistics are due the subjectivity in eliciting priors probability on certain hypothesis or model parameter's values.
- Subjectivity is something unwanted in the ideal picture of the scientist and the scientific method.
- Anything that involves human action will never be free from subjectivity. We have subjectivity in everything and science is no exception.
- The creative and deductive process of theory and hypotheses formulations is **not** objective.
- Frequentist statistics, which bans the use of prior probabilities is also subjective, since there is **A LOT** of subjectivity in choosing which model and likelihood function [8, 17]

How to Incorporate Subjectivity

- Bayesian statistics **embraces** subjectivity while frequentist statistics **bans** it.
- For Bayesian statistics, **subjectivity guides our inferences** and leads to more robust and reliable models that can assist in decision making.
- Whereas, for frequentist statistics, **subjectivity is a taboo** and all inferences should be objective, even if it resorts to **hiding and omitting model assumptions**.
- Bayesian statistics also has assumptions and subjectivity, but these are **declared and formalized**

Types of Priors

In general, we can have 3 types of priors in a Bayesian approach [2, 3, 17]:

- **uniform (flat)**: not recommended.
- **weakly informative**: small amounts of real-world information along with common sense and low specific domain knowledge added.
- **informative**: introduction of medium to high domain knowledge.

Uniform Prior (Flat)

Starts from the premise that “everything is possible”. There is no limits in the degree of beliefs that the distribution of certain values must be or any sort of restrictions.

Flat and super-vague priors are not usually recommended and some thought should included to have at least weakly informative priors.

Formally, an uniform prior is an uniform distribution over all the possible support of the possible values:

- **model parameters**: $\{\theta \in \mathbb{R} : -\infty < \theta < \infty\}$
- **model error or residuals**: $\{\sigma \in \mathbb{R}^+ : 0 \leq \theta < \infty\}$

Weekly Informative Prior

Here we start to have “educated” guess about our parameter values. Hence, we don’t start from the premise that “anything is possible”.

I recommend always to transform the priors of the problem at hand into something centered in 0 with standard deviation of 1[xii]:

- $\theta \sim \text{Normal}(0, 1)$ (Andrew Gelman’s preferred choice[xiii])
- $\theta \sim \text{Student}(\nu = 3, 0, 1)$ (Aki Vehtari’s preferred choice)

[xii] this is called standardization, transforming all variables into $\mu = 0$ and $\sigma = 1$.

[xiii] see more about prior choices in the Stan’s GitHub wiki.

References I

1. Rackauckas C, Ma Y, Noack A, Dixit V, Mogensen PK, Byrne S, Maddhashiya S, Santiago Calderón JB, Nyberg J, Gobburu JV, et al. Accelerated predictive healthcare analytics with pumas, a high performance pharmaceutical modeling and simulation platform. 2020
2. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, and Rubin DB. Bayesian Data Analysis. Chapman and Hall/CRC, 2013
3. McElreath R. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC press, 2020
4. Gelman A, Hill J, and Vehtari A. Regression and Other Stories. Cambridge University Press, 2020
5. Khan ME and Rue H. The Bayesian Learning Rule. 2021 Jul 9. Available from: `http://arxiv.org/abs/2107.04562` [Accessed on: 2021 Jul 13]
6. Bertsekas DP and Tsitsiklis JN. Introduction to Probability, 2nd Edition. 2nd edition. Belmont, Massachusetts: Athena Scientific, 2008 Jul 15. 544 pp.
7. Dekking FM, Kraaikamp C, Lopuhaä HP, and Meester LE. A Modern Introduction to Probability and Statistics: Understanding Why and How. Springer, 2010 Oct 19. 504 pp.
8. Jaynes ET. Probability Theory: The Logic of Science. Cambridge university press, 2003

References II

9. Kurt W. Bayesian Statistics the Fun Way: Understanding Statistics and Probability with Star Wars, LEGO, and Rubber Ducks. Illustrated edition. San Francisco: No Starch Press, 2019 Jul 9. 256 pp.

10. Diaconis P and Skyrms B. Ten Great Ideas about Chance. Google-Books-ID: 68iXDwAAQBAJ. Princeton University Press, 2019 Oct 8. 270 pp.

11. Kolmogorov AN. Foundations of the Theory of Probability. Berlin: Julius Springer, 1933

12. Head ML, Holman L, Lanfear R, Kahn AT, and Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol 2015; 13:e1002106

13. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences 1937; 236:333–80

14. Fisher RA. Statistical methods for research workers. Oliver and Boyd, 1925

15. Rosnow RL and Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 1989; 44:1276–84

16. Fisher RA. Some Examples of Bayes' Method of the Experimental Determination of Probabilities A Priori. Journal of the Royal Statistical Society Series B (Methodological). 1962; 24:118–24

References III

17. van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, Vannucci M, Gelman A, Veen D, Willemsen J, and Yau C. Bayesian Statistics and Modelling. Nature Reviews Methods Primers. 2021 Jan 14; 1(1):1–26. DOI: 10.1038/s43586-020-00001-2. Available from: https://www.nature.com/articles/s43586-020-00001-2 [Accessed on: 2021 Feb 15]

Backup Slides