Chapter 2: Background and Motivation

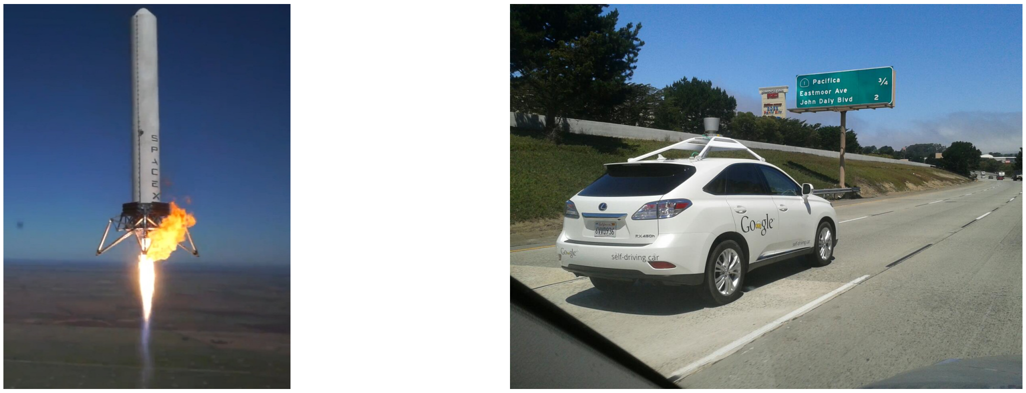
Contents

1 Applications of Bayesian statistics

There are many real-world applications where Bayesian statistics is used in practice. Below are some prominent and/or interesting examples. Of course, one could just as easily come up with a list of applications where non-Bayesian methods are used, but the focus here is on Bayesian statistics.

1.1 Tracking

For vehicle guidance, navigation, and control, it is essential to know the state of the vehicle (location, orientation, velocity) of the vehicle at any given time. Usually, an array of



SpaceX Grasshopper (SpaceX) Google self-driving car

Figure 1: Systems that are likely using Kalman filters or a variant thereof.

This work is licensed under a Creative Commons BY-NC-ND 4.0 International License.

Jeffrey W. Miller (2015). *Lecture Notes on Bayesian Statistics.* Duke University, Durham, NC.

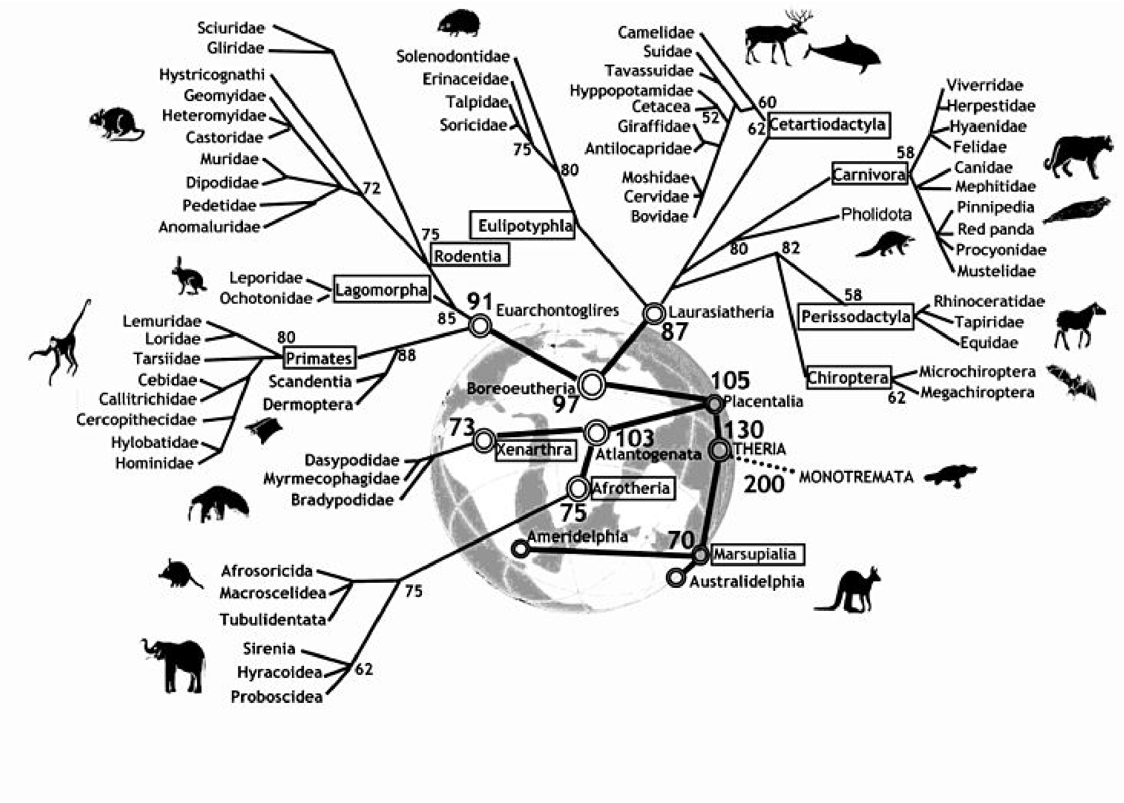


Figure 2: Inferred phylogenetic tree of mammals. (Graphodatsky et al. 2011)

sensors provides various kinds of information of varying quality (e.g., compass, accelerometers, gyroscope, GPS, vision, laser scanner), and this must be combined with knowledge of the vehicle's actions (e.g., wheels, propellors/turbines, rocket engines, ailerons), along with a physical model, in order to infer the state of the vehicle in real-time. In 1960, Rudolf Kalman proposed a solution using a Bayesian time-series model which became known as the Kalman filter. The Kalman filter and its successors have been extraordinarily successful—it is difficult to overstate their importance in the guidance systems of aircraft, spacecraft, and robotics.

1.2 Phylogenetics

Understanding the evolutionary relationships among organisms—that is, the phylogenetic tree—is fundamental in nearly all biological research. Using genetic data from many organisms, along with models of how changes in the genome occur over time, researchers can infer the unknown evolutionary "family tree". Some of the dominant approaches use Bayesian inference (e.g., popular programs include MrBayes and BEAST) and these are widely used throughout biology.

1.3 Political science

In 2008, Nate Silver correctly predicted the U.S. presidential election outcome in 49 out of 50 states, and in 2012, he got all 50 states right. Further, he predicted the U.S. Senate

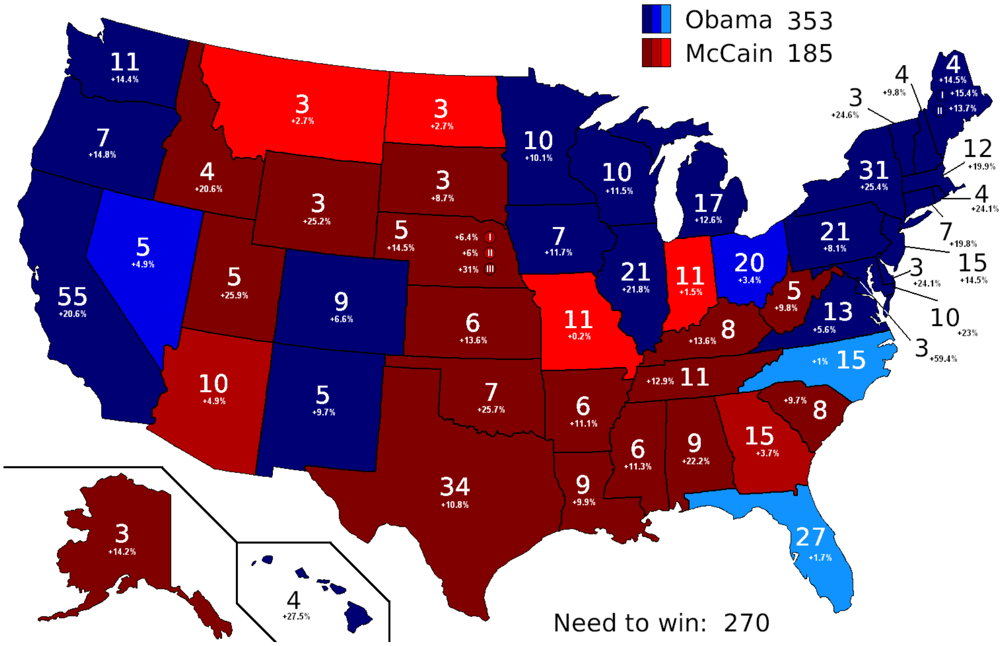


Figure 3: Nate Silver's predictions for the 2008 presidential election.

election outcome correctly in all 35 races in 2008, and in 31 of 33 races in 2012. Silver is an advocate of Bayesian statistics, and although the exact details are secret, it appears that he uses hierarchical Bayesian models to make his forecasts.[1] Bayesian inference is also used extensively by other social science researchers in many other applications.

1.4 Computer science

Spam accounts for the majority of email traffic—typically between 60 to 70% of emails are spam.[2] Yet, due to the sophisticated spam detection algorithms used by email service providers, very little spam gets through to your inbox—and only rarely is real mail classified as spam. For instance, in 2007, Gmail posted the chart in Figure 4, showing that the fraction of spam that gets through is very small indeed.[3] Bayesian models are the most prominent methods for spam detection. A former Microsoft developer who moved to Google reportedly said, "Google uses Bayesian filtering the way Microsoft uses the if statement."[4]

1.5 Search

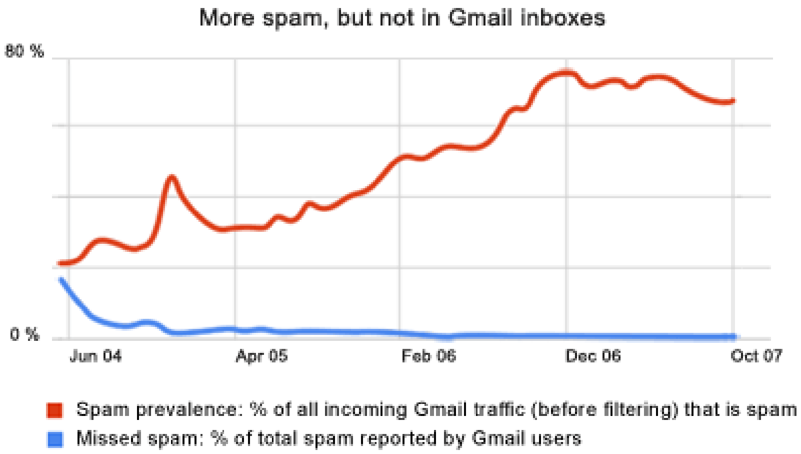
On June 1, 2009, Air France Flight 447 crashed into the Atlantic Ocean, killing all aboard. Despite three intensive searches, the underwater wreckage had still not been found a year

[1] O'Hara, B. (2012). How did Nate Silver predict the US election? The Guardian. http://www.theguardian.com/science/grrlscientist/2012/nov/08/nate-sliver-predict-us-election

[2] Kaspersky Lab, Spam Statistics Reports: Figures, Sources, and Trending Data, 2013–2014. http://usa.kaspersky.com/internet-security-center/threats/spam-statistics-reports-data

[3] Jackson, T. (2007). How our spam filter works. http://gmailblog.blogspot.com/2007/10/how-our-spam-filter-works.html

[4] Joel on Software, Oct 17, 2005. http://www.joelonsoftware.com/items/2005/10/17.html



As the amount of spam has increased, Gmail users have received less of it in their inboxes, reporting a rate less than 1%.

Figure 4

later. French authorities were eventually able to recover the wreckage with the help of a Bayesian search analysis provided by the Metron company (Figure 5). Bayesian search analysis involves formulating many hypothetical scenarios for what happened, constructing a probability distribution of the location under each scenario, and considering the posterior distribution on location given the searches conducted so far. It has also been used to find submarines and ships lost at sea.

1.6 Radiocarbon dating

In archaeology, radiocarbon dating is often used to infer the age of an object. Many times, there is a significant amount of contextual knowledge that can also be brought to bear—for instance, other objects found nearby, where and how deep the object is found, constraints on the order of events, historical records, and other dating techniques. Bayesian methods are now being used to combine such different sources of information in order to improve the accuracy of radiocarbon dating.

1.7 Many other applications as well

This is just a small sampling of real-world applications in which Bayesian methods are actually used in practice.

2 What is Bayesian statistics?

Bayes, in a nutshell

The Bayesian approach can be summarized as follows:

> Assume a probability distribution on any unknowns (this the prior), assume the distribution of the knowns given the unknowns (this is the generating distribution or

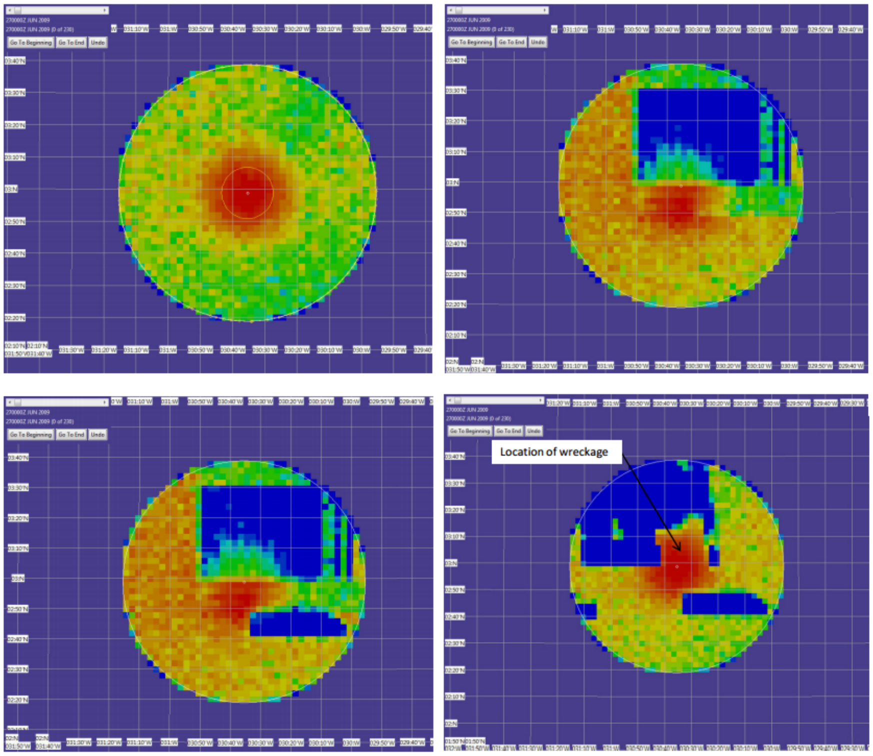


Figure 5: Prior and posteriors (after successive searches) for the location of the wreckage of Air France 447. (Stone et al. 2011)

Figure 6: An archaeological dig. (Axel Hindemith)

likelihood), and then just follow the rules of probability to answer any questions of interest.

An overarching theme of the Bayesian perspective is that uncertainty is quantified with probability distributions. Since essentially all statistical methods involve assuming the form of the generating distribution, it is the prior that distinguishes the Bayesian approach, and makes it possible to just follow the rules of probability.

Bayesian bubble of knowledge

In essence, once distributions have been put on everything, a self-contained "bubble of knowledge" has been constructed in which the answer to any relevant question can—in principle—be answered in a probabilistic manner. The flipside, though, is that this bubble excludes many useful frequentist criteria for evaluating the performance of a procedure. It is important to evaluate Bayesian methods according to these other criteria as well.

What questions of interest often arise?

Here are some recurring examples:

- estimate some unknown parameter or property,
- infer hidden/latent variables or missing data,
- predict future data,
- test a hypothesis, or
- choose among competing models.

In general, answering such questions can be viewed as a problem of choosing the optimal decision—and from our brief study of decision theory, we know that from the Bayesian perspective this consists of minimizing posterior expected loss.

How is this done? What methods are employed?

In order to answer a question of interest, you usually have to get ahold of the posterior in one way or another, and compute one or more posterior expectations (integrals with respect to the posterior density). Three main categories of methods can be distinguished here: exact solution, deterministic approximation, and stochastic approximation.

1. **Exact solution**

   In certain cases, it is computationally feasible to compute the posterior (and posterior expectations) exactly.

   - Exponential families with conjugate priors often enable analytical solutions.
   - Gaussians, in particular, are highly conducive to analytical solutions.
   - For certain graphical models, dynamic programming can provide exact results.

2. **Deterministic approximation**

   Methods include:

   - numerical integration, a.k.a. quadrature/cubature
   - quasi-Monte Carlo (QMC), low discrepancy sequences
   - Laplace's method / Laplace approximation
   - expectation propagation (EP), variational Bayes (VB)

   For low-dimensional integrals, numerical integration and QMC are superior to stochastic approximations. QMC can sometimes perform well in high-dimensional situations as well.

3. **Stochastic approximation**

   For high-dimensional integrals, stochastic approximations are often the only option. The basic idea is that samples from the posterior can be used to approximate posterior expectations. Methods include:

   - Monte Carlo approximation, importance sampling
   - Markov chain Monte Carlo (MCMC) — Gibbs sampling, Metropolis algorithm, Metropolis–Hastings algorithm, slice sampling, Hamiltonian MCMC
   - sequential importance sampling, sequential Monte Carlo, population Monte Carlo
   - approximate Bayesian computation (ABC)

3 What is the difference between the Bayesian and frequentist perspectives?

Once you understand it, the Bayesian approach seems so natural that it is hard to imagine any alternative. In fact, there was no satisfying alternative until the early 1900's, when Karl Pearson, Jerzy Neyman, Egon Pearson, and Ronald Fisher initiated what is now called frequentist statistics.

Roughly, the essential difference between the Bayesian and frequentist perspectives can be described as follows, letting x denote the observed data, and θ denote the unknowns:

- the Bayesian considers only the observed value of x, and treats θ as random,
- the frequentist considers all possible values of x, and treats θ as fixed.

The Bayesian approach is to make the best possible decision given observations x, allowing for uncertainty in θ. The frequentist approach is to use a decision procedure that will have guaranteed performance when used repeatedly, no matter what θ turns out to be.

3.1 Example: Diagnosing celiac disease

Approximately 1 in 100 people is affected by celiac disease, an autoimmune disorder resulting in sensitivity to gluten. Initial diagnosis of the disease is often made using a blood test that measures the level x of a certain antibody. If x is above a certain cutoff point, then a positive diagnosis is made (i.e., disease is present), otherwise, a negative diagnosis is made (i.e., disease not present).[5]

How should this cutoff point be chosen? If the cutoff is too low, there will be too many false positives (incorrectly diagnosing a person as diseased), while if it is too high, there will be too many false negatives (incorrectly diagnosing as undiseased). Given x, we are faced with the following hypothesis testing problem.

Hypothesis 0 (θ_0): Undiseased.
Hypothesis 1 (θ_1): Diseased.

Suppose that from many previous cases, we know the distributions $p(x|\theta_0)$ and $p(x|\theta_1)$ of the antibody level x for undiseased and diseased individuals, respectively.

Bayesian approach

The Bayesian approach is as follows. Since 1 in 100 people is affected, we have $p(\theta_1) = 1/100$ (and thus $p(\theta_0) = 99/100$); this is the prior. Using Bayes' theorem, we can compute the posterior probability of disease for a given individual,

$$p(\theta_1|x) = \frac{p(x|\theta_1)p(\theta_1)}{p(x|\theta_0)p(\theta_0) + p(x|\theta_1)p(\theta_1)}.$$

Quantifying each possible outcome in terms of a loss function ℓ, a diagnosis ($a = \theta_0$ or $a = \theta_1$) is made to minimize the posterior expected loss,

$$\rho(a, x) = \ell(\theta_0, a)p(\theta_0|x) + \ell(\theta_1, a)p(\theta_1|x).$$

Frequentist approach

The usual frequentist approach, on the other hand, is to use a decision procedure that minimizes false negatives, subject to an upper bound on false positives, say, $\alpha = 0.05$. Due to a result called the Neyman–Pearson lemma, this is achieved by choosing $a = \theta_1$ when

$$\frac{p(x|\theta_1)}{p(x|\theta_0)} > c$$

and $a = \theta_0$ otherwise, where $c \geq 0$ is chosen so that the probability of a false positive equals α, i.e.,

$$\mathbb{P}(X \in R_c \mid \theta_0) = \int_{R_c} p(x|\theta_0)dx = \alpha$$

where $R_c = \big\{x : p(x|\theta_1)/p(x|\theta_0) > c\big\}$.

[5] This description is slightly simplified, for illustration purposes.

Comparing the two approaches

So, in the Bayesian approach, the unknown state (diseased or undiseased) is treated as a random variable (since we put a prior on it), and we only consider the observed value of x throughout the analysis. Meanwhile, in the frequentist approach, no prior is used, and the decision procedure (specifically, the choice of c) depends on considering all possible values of the observation x.

The Bayesian approach is optimal under the assumed prior and loss, while the frequentist approach is optimal subject to the chosen bound on false positives.

In a binary decision such as this, it turns out that the two approaches are equivalent, in the sense that for any prior and loss, there is a choice of c for which the Neyman–Pearson procedure coincides with the Bayes procedure—and vice versa, for any c there is a prior and loss for which they coincide.

3.2 Further contrasts between Bayesian and frequentist

Interpretation of probability

From the frequentist perspective, the true value of θ is some unknown but fixed quantity, and it doesn't make sense to speak of θ having a probability distribution—for instance, in the disease diagnosis example of Section 3.1, the patient either has the disease or doesn't, so the probability that the patient has the disease is either 1 or 0.

In order for the Bayesian perspective to make sense, a probability must be interpreted as a subjective level of belief in the truth of a proposition. In contrast, the frequentist interpretation of probability is empirical: the probability of an outcome is the fraction of times it would occur in a sequence of infinitely many trials. (Note: A common misconception is that the frequentist definition only makes sense for trials that have actually occurred many times in the real world, but this is wrong—the trials can be hypothetical.)

Coherence versus calibration

An appealing aspect of the Bayesian approach is its coherence—that is, once the prior has been assumed, no contradictions will arise in the course of doing inference. However, if the prior or likelihood is chosen poorly, this simply means the inferences will be consistently wrong.

Meanwhile, an attractive feature of the frequentist approach is calibration guarantees—that is, we can specify certain performance characteristics, and be guaranteed that they will be met if the procedure is used many times. For instance, in the disease diagnosis example, we specify the fraction of false positives $\alpha = 0.05$, and it is guaranteed that this will be the fraction of false positives occurring in practice (assuming the model is correct).

Frequentist evaluations

From the purely Bayesian perspective, if the prior and likelihood are chosen properly, then the resulting inferences are correct and optimal, and there is nothing more to be said. However, in practice this is not very satisfying, since we often:

- are uncertain about the choice of prior or likelihood,
- employ approximations, and
- want methods which perform well in a wide variety of circumstances.

The frequentist perspective provides empirical and theoretical tools to deal with these issues. Empirically, one can assess performance using cross-validation, test sets, bootstrap, goodness-of-fit tests, and posterior predictive checks. Theoretically, one can provide guarantees of consistency (i.e., convergence to the true value), rates of convergence, and calibration/coverage.

It is a common misconception that the Bayesian and frequentist approaches are mutually exclusive, but this is incorrect. From the frequentist perspective, any procedure can be used—including a Bayesian one—if you can prove that it does what you want.

3.3 Overall recommendation: be pragmatic, not dogmatic

Overall, be pragmatic—that is, use what has been shown to work. As a default approach, the following will serve you well:

> *Design as a Bayesian, and evaluate as a frequentist.*

In other words, construct models and procedures from a Bayesian perspective, and use frequentist tools to evaluate their empirical and theoretical performance. In the spirit of being pragmatic, it might seem unnecessarily restrictive to limit oneself to Bayesian procedures, and indeed, there are times when a non-Bayesian procedure may be preferable to a Bayesian one. However, typically, it turns out that there is no disadvantage in considering only Bayesian procedures—this has been shown formally via the "complete class theorems".

References and supplements

Applications

- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35-45.
- O'Leary, M. A., et al. (2013). The placental mammal ancestor and the postK-Pg radiation of placentals. Science, 339(6120), 662-667.
- Graphodatsky, A. S., Trifonov, V. A., & Stanyon, R. (2011). The genome diversity and karyotype evolution of mammals. Mol Cytogenet, 4(1), 22.
- Stone, L. D., Keller, C. M., Kratzke, T. M., & Strumpfer, J. P. (2011). Search analysis for the underwater wreckage of Air France Flight 447. In Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on (pp. 1-8). IEEE.
- Ramsey, C. B. (2009). Bayesian analysis of radiocarbon dates. Radiocarbon, 51(1), 337-360.

Frequentist and Bayesian

- Kass, R. E. (2011). Statistical inference: The big picture. Statistical Science, 26(1), 1. http://projecteuclid.org/euclid.ss/1307626554
- Jordan, M. I. Are You a Bayesian or a Frequentist? (2009) http://videolectures.net/mlss09uk_jordan_bfway/