Frequentist and Bayesian approach to statistics

Matija Piškorec Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
matija.piskorec@irb.hr

July 12, 2016

Abstract

In this seminar I will present some of the key conceptual and practical differences between frequentist and Bayesian approaches to statistical inference. I will start with a short introductory example of estimating success rate of clinical trials, and continue with parameter inference in the case of linear regression with polynomials. In the end, I will explain the problem of model selection and how it differs in the two approaches.

1 Introduction

The goal of data modeling is to infer joint probability distribution $P(\theta|\mathcal{M},\mathcal{D})$ for our two random variables - data $\mathcal{D}$ and parameters θ, under the assumption of a model $\mathcal{M}$ [1, 2]. Let us assume for now that our model $\mathcal{M}$ is given so that we can temporarily drop the dependence on it in the expression above. In general, we can consider multiple competing models $\mathcal{M}_i$ and evaluate their fit to the data through the process of *model selection* (section 4). We can express this joint probability in two mathematically equivalent ways, either through conditional probability of data $P(\mathcal{D}|\theta)P(\theta)$ or conditional probability of parameters $P(\theta|\mathcal{D})P(\mathcal{D})$. From this we can derive the Bayes theorem:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{1}$$

- $p(\theta|\mathcal{D})$ **posterior** probability
- $p(\mathcal{D}|\theta)$ **likelihood** which gives us probability of data given parameters.
- $p(\theta)$ **prior** probability of parameters before we observe any data.
- $p(\mathcal{D})$ **model evidence** or **marginal likelihood** which servers as a normalizing constant and because it does not depend on the parameters θ it can be ignored in parameter inference.

In principle, Bayes theorem gives as a probabilistically principled way of performing inference on the unknown parameter values θ given observed data $\mathcal{D}$. One of the oldest known usages is by Pierre-Simon Laplace, who improved on Bayes's work and applied the theorem on wide range of problems [3]. However, there are two difficulties in using the Bayes theorem directly for statistical inference. First of these is *technical*, as calculation of a full posterior probability distribution through the Bayes theorem requires a manipulation of high-dimensional integrals. Only for a certain choice of prior and likelihood distributions there is a guarantee that we will obtain a closed form solution. Such is the case, for example, with distributions which are *conjugate* to one another. In general, there is no guarantee that closed form solution even exists, and we have to resort either to direct numerical integration, or to sampling methods such as Markov Chain Monte Carlo (MCMC) [2]. This technical difficulty was alleviated in part by the development of more powerful computational methods in recent decades, and, more earlier, by the development of efficient inference methods which infer *point estimates* rather than the full posterior distribution. These point estimates somehow characterize a *best* solution to the inference, and under certain choice of prior and likelihood distributions there is a guarantee that they have certain beneficial properties - for example, that they correspond to a mode or an average of a posterior distribution.

The other difficulty is a *conceptual* one, as we have to express parameters of our model as probability distributions. The most controversial part of this is the specification of prior - a probability distribution for our parameters before we observed any data at all! In practice, this can actually aid us in our inference, as it allows us to encode prior knowledge into our inference. If priors are reasonable, this will actually make our inference more efficient, as we will need less data to make good inferences because priors already provide part of an explanation. With enough data, any reasonable prior, meaning any prior which assigns a non zero probability to the true parameter values, will be overwhelmed by the weight of data. The downside is that inappropriate priors will introduce bias into our inference. We can try to avoid this by choosing an uninformative prior - a prior which encodes as little information as possible about our parameter (for example, a flat or uniform prior), although this is sometimes impossible and it can have a negative impact on the efficiency of inference [4].

We can now finally highlight some of the main differences between frequentist and Bayesian approaches to inference. If we use the full power of Bayes theorem for inference, including specifying prior distributions for the parameters and obtaining a full posterior distribution for our parameters of interest, we are following a **Bayesian** approach to inference [2, 5]. If we are uncomfortable with the concept of priors and are happy with point estimates for our parameters, we can base our inference solely on the likelihood $p(\mathcal{D}|\theta)$ part of the Bayes theorem. For inference we can use a *likelihood function* $\mathcal{L}(\mathcal{D};\theta)$, which is proportional to the likelihood but itself is not a probability distribution - for example, it does not necessarily integrate to unity! We can optimize likelihood function with efficient numerical optimization methods, giving us a point estimate for our parameters. In that case we are following a **frequentist** approach to inference. The origin of the word *frequentist* is in the fact that, through likelihood, we implicitly consider all possible datasets that could have occurred with this particular combination of model and its parameter values - our dataset is a random variable! However, as we have only one realization of this particular dataset, we can calculate an exact probability of observing this particular dataset under the assumption of a fixed model and its parameters. This allow us to perform a hypothesis testing procedure - we can define a suitable *null model* which states that there is no effect at all, and then calculate the probability of observing a dataset that is as extreme as this one - a *p-value*. If this probability is lower then some prespecified threshold, we can reject the null hypothesis.

So Bayes theorem, being one of the fundamental theorems of probability theory, holds universally regardless whether we choose to follow frequentist or Bayesian approach. The difference is in the interpretation of probability itself. Bayesian approach interprets probabilities as degrees of belief, or knowledge, about unknown parameters considering fixed data. So it is conceptually straightforward in Bayesian approach to express parameters as random variables, and to use full power of Bayesian theorem for inference. On the other hand, frequentist approach interprets probabilities as frequencies of real or hypothetical events which are underlaid by a fixed, although unknown, model and its parameters. In practice, Bayesian analysis involves computation of a full posterior distribution for the parameters of interest, often through numerical integration or a sampling method like Markov Chain Monte Carlo. Frequentist analysis involves calculation of point estimates for the parameters through the likelihood function, often using numerical optimization for finding maximum likelihood solution, and hypothesis testing.

2 Introduction example: Clinical trials

The following example is inspired by discussion on Bayesian clinical trials from [6, 7], and it will serve us to introduce some key differences between frequentist and Bayesian approach to statistical inference. Clinical trial can be modeled as a consecutive execution of a measurement with two possible outcomes - success and failure (of treatment for example). The goal of the analysis is to use information on the observed outcomes and infer the probability of success in general.

Bayesian approach

Bayesian try to model the probability of parameter of interest directly. In this case the most appropriate model is the beta distribution which is parameterized with two parameters α and β, with $\alpha, \beta > 0$. The probability density function is given by:

$$p(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad (2)$$

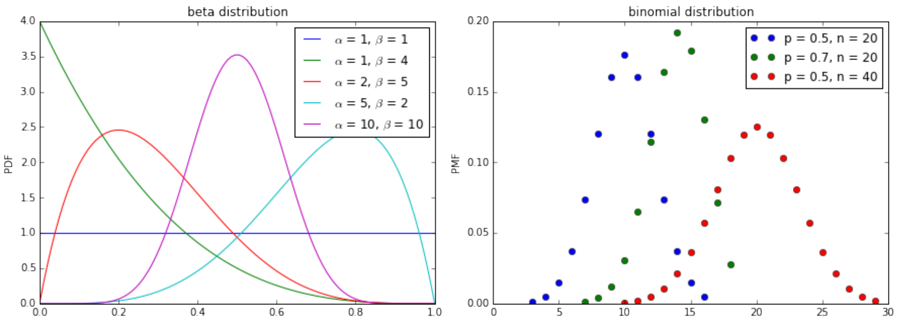

Figure 1: Beta distribution for different values of α and β (left). Binomial distribution for different parameters p and n (right).

Where beta function $B(\alpha, \beta)$ and gamma function $\Gamma(\alpha)$ serve as normalization constants so that probability integrates to 1. Left panel of Figure 1 visualizes beta distribution for several values of α and β. Beta distribution has several properties which make it very convenient to use in Bayesian analysis:

- It is bounded on the domain $[0, 1]$ which makes it ideal to represent probability distribution of parameters that are themselves probabilities, such as the ones in binomial or Bernoulli distributions, as well as percentages and proportions.
- Also, it is *conjugate* to binomial and Bernoulli (as a special case of binomial) distributions. This means that if we express our prior as beta distribution and use binomial or Bernoulli as sampling distributions our posterior will again be a beta distribution. This allows us to iteratively estimate the posterior as we observe each new data point.

In successive binomial trials, where each outcome is either a *success* or a *failure*, parameter α indexes the number of successes and parameter β indexes number of failures, the beta distribution gives us posterior distribution of the parameter p of the binomial distribution. Starting from a flat prior distribution $Beta(1, 1)$ we can iteratively update posterior as we observe each new data point. Figure 2 shows posterior distribution for parameter p after each of the following consecutive outcomes: `SSFSSFS` (where `S` stands for success and `F` for failure). The mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$ and the mode (maximum a posteriori estimate or MAP) $\frac{\alpha-1}{\alpha+\beta-2}$ (for $\alpha, \beta > 1$) which means that in the limit of large number of observations our posterior distribution will be characterized by the frequencies of successes and failures. For our particular seven consecutive outcomes this would yield an estimate of $\hat{p}_{mean} = 0.71$ if estimated from mean or $\hat{p}_{MAP} = 0.80$ if estimated from a mode of a posterior.

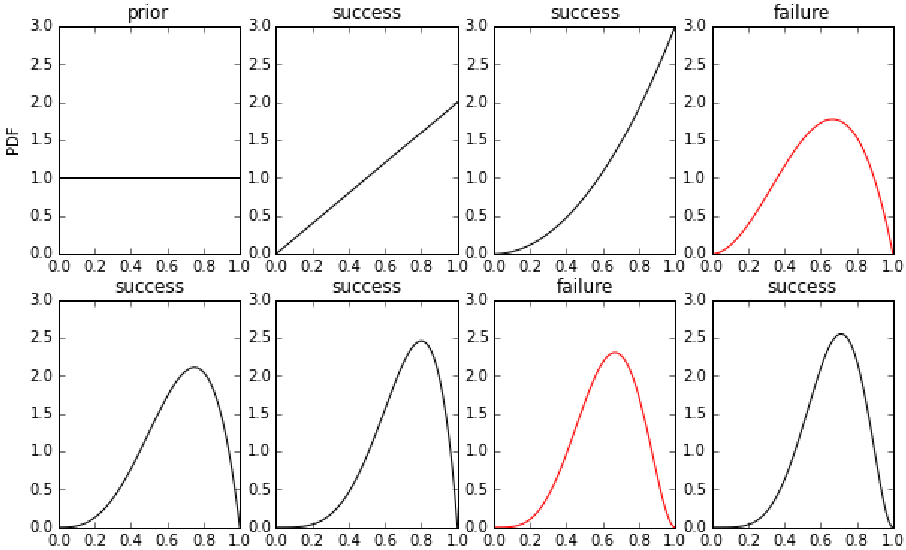

Figure 2: Posterior distribution for parameter p after each of the following consecutive outcomes: `SSFSSFS` (where `S` stands for success and `F` for failure).

Frequentist approach

Frequentists cannot model parameter of interest as a random variable, and so cannot use the beta distribution for inference. However, they can model any quantity derived out of data as a random variable, such as the actual number of successes or failures. For this they can use a binomial distribution, which gives probability of getting exactly k successes in n trials:

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3}$$

Where $\binom{n}{k}$ is a binomial coefficient equal to $\frac{n!}{k!(n-k)!}$, and $p \in [0, 1], n \in \mathbb{N}$. Note that binomial distribution is a discrete distribution, and $P(k; n, p)$ is a *probability mass* while $p(x; \alpha, \beta)$ in the case of beta distribution is a *probability density*. Right panel of Figure 1 visualizes binomial distribution for various parameters p and n, and table bellow list the probabilities for our particular case of six measurements.

For our particular case of seven consecutive measurements `SSFSSFS` (where `S` stands for success and `F` for failure) the respective probabilities for exactly five successful outcomes are $P(X = 5|n = 7, p = 0.25) = 0.001$, $P(X = 5|n = 7, p = 0.5) = 0.164$ and $P(X = 5|n = 7, p = 0.75) = 0.311$. Depending on our assumed probability p the probability of five successful outcomes changes. This is why frequentists are often interested in *null hypothesis testing* - they assume some value of p and then try to reject. In clinical trials, the base value

p_0 (the null hypothesis) is usually a value obtained from a control group and the goal of the hypothesis testing is to show whether the evidence is strong enough to reject this null hypothesis - that is, that the measured effect is statistically significant. Statistical significance is measured as the probability of *a more extreme outcome* then the one actually observed. In our case this means the probability of observing more than five successful outcomes $P(X > 5|n = 7, p) = 1 - F(X = 5|n = 7, p)$, where F is a *cumulative distribution function*. This probability is called *p-value* and in our case it is $p_{0.25} = 0.0013$, $p_{0.5} = 0.0625$ and $p_{0.75} = 0.4450$. The smaller the p-value the more evidence there is that our outcomes are not due to the statistical chance. We can report just the p-value directly or we can choose a *significance level* α which gives us a threshold for rejecting the null hypothesis, the most common being $\alpha = 0.05$ and $\alpha = 0.01$. From our data we can see that we can reject hypothesis $p = 0.25$ with significance level $\alpha = 0.01$, while hypotheses $p = 0.5$ and $p = 0.75$ cannot be rejected with neither of these significance levels.

Let us summarize differences between frequentist and Bayesian approach using the above example:

- Bayesians were required to specify a prior distribution - our knowledge of the parameter before any data is observed. We had chosen flat uniform prior $B(1, 1)$.
- Aside from uniform prior, Bayesians condition only on the data that was actually observed - in this case number of successes α and number of failures β.
- Frequentist have to condition on the quantities related to the experimental design - in this case total number of trials n which is required in the binomial distribution.
- Additionally, frequentists have to condition on the assumed value of the parameter of interest - in this case probability of obtaining a success p. This makes frequentist approach naturally applicable to hypothesis testing where one first assumes some value of a parameter and then seeks statistical evidence to reject or accept this hypothesis.
- While beta distribution gives us an answer about the parameter of interest (probability of each individual success), binomial distribution gives as an answer about the data that we can potentially observe (probability of obtaining an exact number of successes).

Lindley's paradox

A rather striking example of when frequentist and Bayesian approach can lead to a completely different inferences is known as *Lindley's paradox* [8]. It usually arises when one considers two mutually exclusive hypothesis for a parameter of interest - H_0 and H_1, and one of them is very precise while the other is very diffuse. For example, when H_0 is a null hypothesis that a certain proportion $\theta = 0.5$, while alternative hypothesis H_1 is $\theta \neq 0.5$. Examples from literature include inference of the proportion of male versus female births, extra sensory perception (ESP) [9, 10], and statistics of particle collisions [11].

Frequentist would test the hypothesis H_0 by calculating probability of observing the realized proportion in the data, either using a binomial distribution directly or, in case of large number of examples, a normal approximation. Assuming that there really is a slight bias in the data (meaning $\theta \neq 0.5$), hypothesis H_0 would probably be easily rejected with a certain significance level α. Notice that frequentist methodology does not explicitly reference alternative hypothesis H_1 in its analysis, which is one of the reasons for the paradox. Inference with Bayesian methodology calculates the full posterior distribution for H_0 and the observed data $\mathcal{D}$ (which, in our case, is just the observed proportion):

$$P(H_0|\mathcal{D}) = \frac{P(\mathcal{D}|H_0)P(H_0)}{P(\mathcal{D}|H_0) + P(\mathcal{D}|H_1)} \tag{4}$$

As we see we have to include an explicit reference to H_1 through $P(\mathcal{D}|H_1)$. If we assume that both hypothesis are equally likely a priori ($P(H_0) = P(H_1) = 0.5$) we can disregard priors. The question we are answering now is slightly different than in the frequentist case because we are evaluating hypothesis H_0 in the light of alternative hypothesis H_1. The answer might well be that hypothesis H_0 better explains the data than alternative hypothesis H_1, which is in apparent contradiction with the conclusion of the frequentist approach which rejected H_0. This should not surprise us, as the presence of slight bias in the data means that θ is indeed very close to 0.5, and alternative hypothesis $\theta \neq 0.5$ is not informative enough to influence the posterior. This does not mean that frequentist and Bayesian methods are in disagreement, just that they answer different questions and care should be taken when interpreting results of the inference.

Confidence intervals and credibility intervals

I should note here that frequentists and Bayesians calculate two different types of intervals when making inference about unknown parameters. These intervals can coincide under some specific conditions, but in any case their interpretations are very different and care should be taken not to confuse them [12]:

- **Confidence intervals**: In repeated sampling, 90% of realized intervals cover the true parameter θ. In other words (if we interpret *frequency* as *probability*), confidence intervals give us the probability that the interval covers the true value of parameter θ.

- **Credibility intervals**: For these data, there is a 90% probability that the parameter θ is in the interval. In other words, credibility intervals give us the probability that the true value of parameter θ lies within the interval.

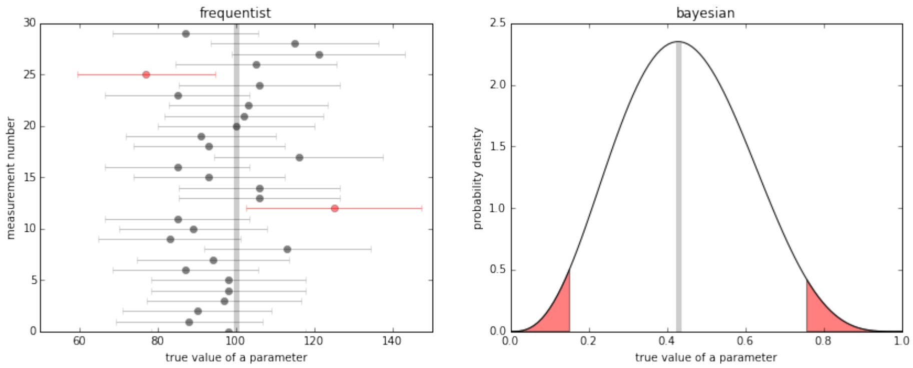

Figure 3: Confidence intervals (left) are defined so that certain proportion of them covers the true parameter, while credibility intervals (right) are defined so that they cover the true parameter with certain probability.

Notice that in the definition of confidence intervals the parameter is *fixed* and interval is *probabilistic*, while in the definition of credibility intervals it is the other way around. It is quite common to interpret confidence intervals as if they express a probabilistic statement about the parameter, while quite contrary, they express a probabilistic statement about the *interval*! Figure 3 highlights difference between confidence and credibility intervals. Confidence intervals are defined so that certain proportion of them covers the true parameter, while credibility intervals are defined so that they cover the true parameter with certain probability.

3 Linear regression

We will show how frequentist and Bayesian approach differ when performing linear regression. Our two main use cases are *parameter fitting*, where we try to find parameters of the best fitting linear model, and *model selection*, where we try to evaluate a fit made by several competing models. Some of these examples were inspired by the similar discussion in [13]. We define our linear model like this:

$$y(x;\theta) = \theta_0 + \theta_1 x \tag{5}$$

$$y \sim \mathcal{N}(y, \sigma^2) \tag{6}$$

Where parameter θ_0 is *intercept* and θ_1 is *slope*. Error of the model is assumed to be normally distributed (Gaussian) with mean zero and standard deviation σ. Left panel of Figure 4 shows an example of 15 data points generated from the above model.

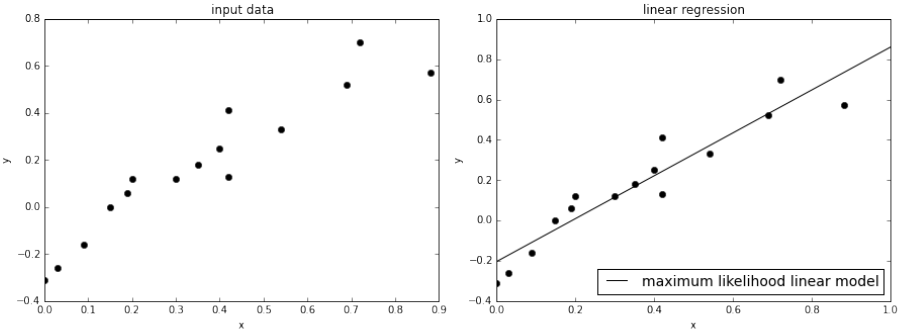

Figure 4: An example data generated from our model (left). Maximum likelihood estimate of linear model for our data (right).

Frequentist approach to linear regression

The probability for each data point is given with the normal distribution:

$$p(x_i, y_i \mid \theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-[y_i - \hat{y}(x_i \mid \theta)]^2}{2\sigma^2}\right] \tag{7}$$

And the likelihood for all N data points (x_i, y_i) is:

$$\mathcal{L}(\{(x_i, y_i)\} \mid \theta, \sigma) = \prod_{i=1}^{N} p(x_i, y_i \mid \theta, \sigma) \tag{8}$$

Due to the small probabilities involved it is much more convenient to express instead a *log-likelihood*, which is proportional to the likelihood:

$$\log \mathcal{L}(\{(x_i, y_i)\} \mid \theta, \sigma) \propto (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i-1}^{N} [y_i - \hat{y}(x_i \mid \theta)]^2\right] \tag{9}$$

The parameters θ_0 and θ_1 can now be estimated by a *maximum likelihood* estimation - that is, by finding parameters that maximize log-likelihood. Because of the monotonicity of logarithm these same parameters also maximize original likelihood function. Because we assumed Gaussian errors on our linear model the resulting log-likelihood is a convex function with a single global minimum which we could find analytically by setting $d \log \mathcal{L}/d\theta = 0$ and solving for θ. This procedure is equivalent to the *least squares* method because we are searching for a solution that minimizes sum of squared errors. We can perform this optimization using Python's SciPy library for scientific computing [15]. Right panel of Figure 4 shows maximum likelihood estimate of linear model for our data.

Bayesian approach to linear regression

Instead of performing maximum likelihood estimate, which is essentially a point estimate, a Bayesian approach is to instead calculate the whole posterior distribution for the parameters of interest. From Bayes theorem (equation 2) our posterior is proportional to the product of likelihood and prior:

$$p(x_i, y_i \mid \theta, \sigma) \propto p(x_i, y_i \mid \theta, \sigma)\ p(\theta, \sigma) \tag{10}$$

Where likelihood $p(x_i, y_i \mid \theta)$ is equivalent to the likelihood in the frequentists case. Care should be taken in the choice of prior on the parameters $p(\theta)$. In general we would like to use an *uninformative prior* in order to influence our posterior as little as possible. Uniform prior is appropriate for θ_0 (intercept) but not for θ_1 (slope) as it biases towards models with higher slope. Uniform prior is also not appropriate for σ as it is not invariant to scaling, so we will use *Jeffreys prior*. Our prior is now this [14]:

$$p(\theta, \sigma) \propto 1/\sigma(1 + \theta_1^2)^{-3/2} \tag{11}$$

In this simple case we could analytically calculate the posterior distribution, but in general case this is not possible. Easier and more straightforward way is to use Markov Chain Monte Carlo [2, 16] to sample from the posterior, and use these samples to characterize posterior distribution. For this we can use Python's library PyMC [1] - an implementation of Metropolis-Hastings sampling method [17]. Figure 5 shows the results for frequentist and Bayesian approach.

We see that the two approaches yielded almost exactly the same solution. This should not be surprising as frequentist methods are sometimes, for good reasons, derived as optimal under certain Bayesian conditions. In this particular case, maximum likelihood estimation under the assumption of Gaussian errors produces the same solution as Bayesian approach with an uniform prior on the slope parameter. The advantage of Bayesian approach is that we obtain the *whole posterior distribution*, not just its mode or mean, and this allow us to estimate uncertainty of our estimate. The left panel of Figure 5 shows the posterior for the parameters θ_0 and θ_1 from which we can extract parameters which fall within first standard deviation from the mean. Right panel of Figure 5 shows the resulting linear models. The uncertainty is lowest in the middle of the data range because this is where majority of linear models have to pass. The highest uncertainty is on the edges of the data range as expected.

4 Model selection

So far we have considered a problem of finding parameters θ while assuming a fixed model $\mathcal{M}$. If the model is not known in advance, and if there are multiple plausible models which can explain the data, our inference methodology needs to be able to evaluate competing model through *model selection*. In this

[1] `https://github.com/pymc-devs/pymc`

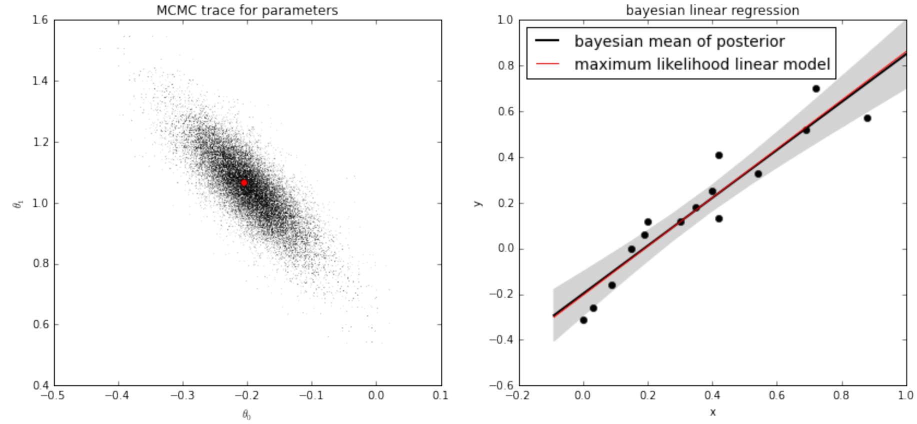

Figure 5: Results of linear regression with frequentist and Bayesian approach. MCMC trace of the parameters θ_0 and θ_1, with best solution marked in red (left). Best fitting lines infered by maximum likelihood (frequentist approach) and as a mean of the posterior distribution (Bayesian approach, right). The uncertainty bands on the right panels correspond to the models that fall within first standard deviation from the mean of the posterior.

section we will demonstrate how we can choose between competing models using frequentists and Bayesian approach. First example is linear regression with polynomials on a two dimensional data, where model selection means selecting an appropriate degree of the polynomial to use in regression. Second example is regression using exponential and logarithmic functions, which is a challenge for frequentist approach because there is no easy way to estimate *model complexity* [22], an issue which Bayesian model selection effectively solves.

4.1 Example 1: Linear regression with polynomials

Figure 6 shows our 15 data points and maximum likelihood linear and quadratic models. We will restrict this example to choosing between a linear and a quadratic model, although methodology is applicable in general. Log likelihood for the linear model is -14.32, and for the quadratic model is -16.12. This would suggest that quadratic model provides a better fit to the data, but this is expected because quadratic model has more degrees of freedom. This is demonstrated on the Figure 6 where we plot log likelihoods for polynomials up to degree nine. The higher the degree of polynomial the better the fit, but this obviously comes at a price of overfitting - model is too complex for the data it should describe and it fits noise along with the underlying pattern. In general, polynomial of degree n will perfectly fit $n+1$ data points, but does not mean that polynomial of highest degree is the best choice for a model. The fit of a model has to be adjusted with the complexity of a model, with more complex models getting more penalty on their log likelihoods. How exactly to do this in

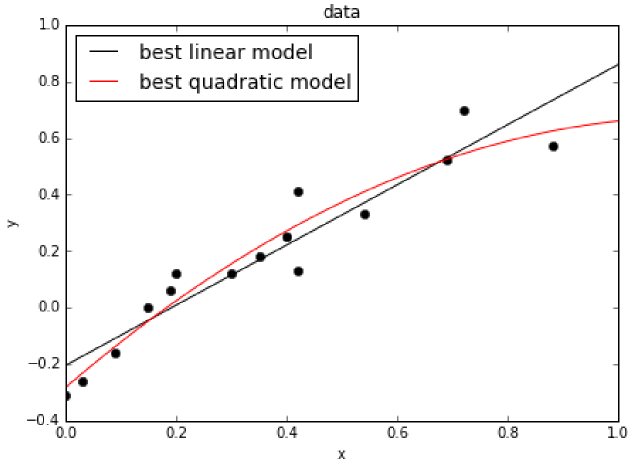

Figure 6: An example data generated from our model and linear and quadratic fit to the data.

frequentist and Bayesian approach is described in the following sections.

Frequentist model selection

Recall that in frequentist interpretation both data and all quantities derived from data are modeled as random variables. In our regression we had defined our likelihood as a sum of Gaussian random variables, and these are distributed according to a χ^2 distribution with certain degree of freedom. So we can use this distribution to evaluate fit of our model to the given data. Left panel of Figure 8 shows a χ^2 distribution with one degree of freedom for the linear model and a χ^2 distribution with two degrees of freedom for the quadratic model. Difference between χ^2 statistics for our two models gives us a probability that we will observe data that favors quadratic model when linear model is true.

Right panel of Figure 8 shows this probability. Assuming the linear model is true, there is a 6% probability that simply by chance we would observe data that favors quadratic model more than the linear. If we choose the significance level $\alpha = 0.05$ and perform null hypothesis testing we will not be able to reject the null hypothesis that data came from the linear model. So in this case there data is ambiguous and there is no strong evidence to suggest data really came from a quadratic model.

Bayesian model selection

Again, Bayesian model selection starts from Bayes theorem [5]. This time, the crucial term is *model evidence* or *marginal likelihood* $p(\mathcal{D}|\mathcal{M}_i)$ that served as a normalizing constant in parameter inference. It can be expressed as a

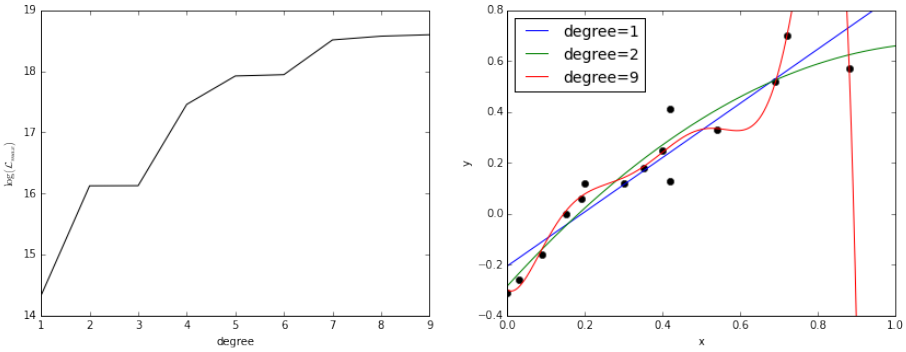

Figure 7: Log likelihoods for polynomials up to the degree nine (left), and the resulting polynomial fits (right). In general, polynomial of a higher degree will always provide a better fit to data, although this does not necessarily mean that it will also provide a better fit for yet *unseen* data.

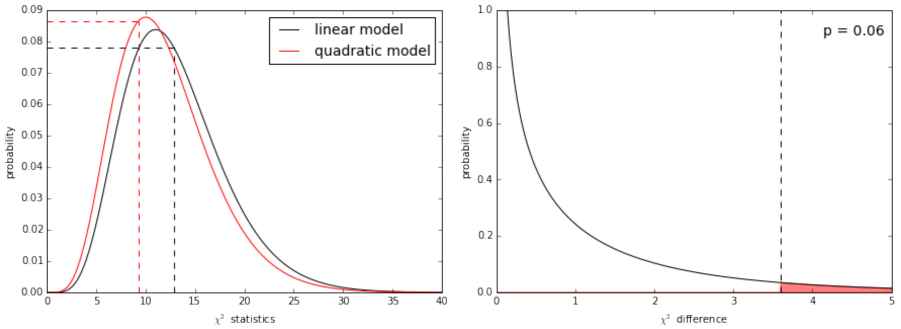

Figure 8: χ^2 distribution with one degree of freedom for the linear model and a χ^2 distribution with two degrees of freedom for the quadratic model (left). Assuming the linear model is true, there is a 6% probability that simply by chance we would observe data that favors quadratic model more than the linear (right).

marginalization of likelihood over all possible parameter values:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta \tag{12}$$

If we view likelihood as a probability distribution over all possible datasets then its marginalization serves as a form of *Occam's razor* [19] which restricts the complexity of models - models which spread their likelihood over too many datasets will assign small probability to each of them, and will be outperformed by simpler models. The ratio of model evidences is called a *Bayes factor* [20] and it tells us how did the prior odds for model changed after data observation:

$$\underbrace{\frac{p(\mathcal{M}_i|\mathcal{D})}{p(\mathcal{M}_j|\mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}}_{\text{prior odds}} \tag{13}$$

Because we are dealing with two very simple models we can perform numerical integration of the likelihood directly for the two models. We will do this with numerical integration routines available in Python's library for scientific computing SciPy [15]. Bayes factor in favor of quadratic model is 1.15, which corresponds to a very weak support to the quadratic model [20]. Similar as in the frequentist approach, we can argue that the evidence is not strong enough to reject the simpler linear model.

4.2 Example 2: Linear regression with exponential and logarithmic models

Let us now consider two models with the same number of parameters, but with different functional forms:

$$\begin{aligned} y &= ax^b + \text{error (Steven's model)} \\ y &= a\ln(x+b) + \text{error (Fechner's model)} \end{aligned} \tag{14}$$

How will we perform model selection here? Just naively employing χ^2 statistics with the given degrees of freedom will provide the same estimate for the two models, and we somehow suspect that this can not be correct. Figure 9 shows data generated from Steven's and Fechner's model and maximum likelihood fits for each model.

In this particular example, log likelihood for Steven's model is around -13, regardless of whether the data was generated from Steven's or Fechner's model. On the other hand, log likelihood for Fechner's model is -53.86 for data generated by Steven's model and -14.95 for data generated by Fechner's model. Table 1 shows model selection using just maximum likelihood estimates for 200 datasets generated by Steven's or Fechner's model. We see that Steven's model is able to fit data generated from Fechner's model even better than the Fechner's model itself! As both models have the same number of parameters these additional degrees of freedom have to be due to the functional form of the model.

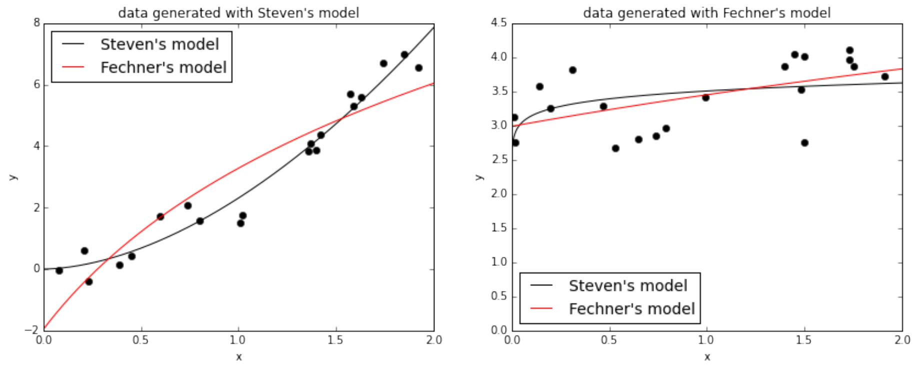

Figure 9: Data generated from Steven's (left) and Fechner's (right) models and maximum likelihood fits.

Table 1: Model selection for Steven's and Fechner's model using maximum likelihood. Numbers represent number of times we selected each model and the actual model from which data were generated.

	data from Steven's	data from Fechner's
Steven's fit	100	46
Fechner's fit	0	54

However, even in this case when models have the same number of parameters, Bayesian model selection using marginal likelihood should still work! Marginalization over the space of parameters implicitly accounts for the expressiveness of models arising either through number of parameters or the functional form of the model. As discussed in [21] and [22], the influence of the functional form of a model can be nonneglibigle for small samples. Table 2 shows model selection using Bayesian model selection for 200 datasets generated by Steven's or Fechner's model. We see that additional degrees of freedom inherent in the Steven's model are accounted for and that we are able to identify underlying model in all cases.

Table 2: Model selection for Steven's and Fechner's model using marginal likelihood. Numbers represent number of times we selected each model and the actual model from which data were generated.

	data from Steven's	data from Fechner's
Steven's fit	100	0
Fechner's fit	0	100

5 Conclusion

This seminar provided a brief introduction to the conceptual and practical differences between Bayesian and frequentist approach to statistical inference. I hope that it is clear by now that Bayesian theorem, being one of the fundamental theorems of probability theory, holds universally regardless of which approach you choose to follow. What is controversial is the usage of the theorem for statistical inference, or “inverse probability” - a term which was in wide use until the beginning of 20^{th} century [23]. The underlying philosophical difference is in the definition of the probability itself - while Bayesians interpret probability as degrees of belief, or knowledge, about unknown parameters, frequentists interpret probability through frequencies of real or hypothetical events. The practical difference is that Bayesians are willing to express models and their parameters as random variables, and to use full power of Bayesian theorem for inference, while frequentist regard model and their parameters as fixed, although unknown, and perform their inference from there. Both approaches have their strengths and weaknesses, with Bayesian being conceptually more straightforward to use although computationally more demanding than the frequentist approach.

Acknowledgment

This review paper has been written for the course: “Statistical methods for data mining” in the first year of postgraduate study 2014./2015. at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.

References

[1] C. M. Bishop, “Pattern Recognition and Machine Learning”, Springer (2007)

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, “Bayesian Data Analysis, 3rd Edition”, Chapman and Hall/CRC (2013).

[3] Stephen M. Stigler, “The history of statistics”. Harvard University press (1986)

[4] Kass, Robert E and Wasserman, Larry, “The selection of prior distributions by formal rules”, *Journal of the American Statistical Association*, 91, 1343-1370 (1996)

[5] T. Ando, “Bayesian Model Selection and Statistical Modeling”, Chapman and Hall/CRC (2010).

[6] Donald R. Berry, “Bayesian clinical trials”, *Nature Reviews Drug Discovery 5*, 27-36 (2006)

[7] Sean R. Eddy, "What is Bayesian statistics?", *Nature Biotechnology 22*, 1177-1178 (2004)

[8] Christian P. Robert, "On the Jeffreys–Lindley's paradox", *Philosophy of Science 81* (2013) `http://arxiv.org/pdf/1303.5973.pdf`

[9] Jose M. Bernardo, "Integrated Objective Bayesian Estimation and Hypothesis Testing", *Bayesian statistics 9* (2010)

[10] Jan Sprenger, "Testing a Precise Null Hypothesis: The Case of Lindley's Paradox", *Philosophy of Science 80* (2013)

[11] Aris Spanos, "Who Should Be Afraid of the Jeffreys-Lindley Paradox?", *Philosophy of Science 80* (2013)

[12] E. T. Jaynes, "Confidence Intervals vs Bayesian Intervals", *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science II*, 175-257 (1976)

[13] Jake VanderPlas, "Frequentism and Bayesianism: A Python-driven Primer", *Proceedings of the* 13^{th} *Python in Science Conference* (SCIPY 2014)

[14] E. T. Jaynes, "Straight line fitting - A Bayesian solution", (1999) `http://bayes.wustl.edu/etj/articles/leapz.pdf`

[15] Eric Jones, Travis Oliphant, Pearu Peterson and others, "SciPy: Open source scientific tools for Python" (2001) `http://www.scipy.org/`

[16] C. Andrieu et. al., "An Introduction to MCMC for Machine Learning", *Machine Learning*, 50, 5-43, (2003)

[17] A. Patil, D. Huard, C.J. Fonnesbeck. "PyMC: Bayesian Stochastic Modelling", *Python Journal of Statistical Software*, 35(4):1-81 (2010)

[18] I. Myung, "The importance of complexity in model selection," *Journal of Mathematical Psychology*, vol. 44, no. 1, pp. 190–204 (2000)

[19] I. Murray and Z. Ghahramani, "A note on the evidence and bayesian occam's razor," Gatsby Computational Neuroscience Unit, University College London, Technical Report, (2005) `http://mlg.eng.cam.ac.uk/zoubin/papers/05occam/occam.pdf`

[20] R. E. Kass and A. E. Raftery, "Bayes factors", *Journal of the American Statistical Association*, vol. 90, no. 430 (1995)

[21] Peter G. Grunwald, "The minium description length principle", The MIT Press (2007)

[22] Jae Myung et. al., "Counting probability distributions: Differential geometry and model selection", PNAS (2000)

[23] Stephen E. Fienberg, "When did Bayesian Inference become Bayesian?", *Bayesian Analysis*, 1 (1), 1–40 (2006)