RESEARCH ARTICLE OPEN ACCESS

Bayesian Estimators for Normal Distribution Parameters, the Frequentist and Bayesian Approaches in Inferential Analysis

Klodiana Bani, Markela Muca

(Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana)

--************************----------------------------------

Abstract:

The three goals of the inferential analysis are: parameter estimation, prediction from data and the model comparison. Usually a parameter of a probability distribution is unknown but determines the property of the distribution, that in the case of the normal distribution are its mean and standard deviation. The "bell curve" of the normal distribution is totally defined by the mean which is its centre and the standard deviation which is its width. For the prediction it is needed the estimation of certain parameters to predict future data. Moreover, the comparison of the models it is related with the selection of the best model among two or more suitable models which explain the data.

The Frequentist inference is based on the long term frequencies but the Bayesian inference is mostly related to the degrees of belief and logical support. Shortly, the overview of the Frequentist means that probabilities are equal to the long term frequencies of an event without attaching them to hypothesis or to any fixed but unknown values, but in contrast with this, for a Bayesian it is possible to use probabilities to represent uncertainty or hypothesis.

In this article, it will be presented the estimation of the normal distribution parameters from the Bayesian inference and at least it will be discussed the comparison of the estimators from the classical and Bayesian analysis from the results obtained from simulations.

Keywords — **Normal distribution, prior distribution, posterior distribution, Bayesian analysis.**

--************************----------------------------------

I. INTRODUCTION

It was reverend Thomas Bayes who proposed Bayesian theory in 1763 and used it for the quantification of binomial distribution by the collected data. Then was Laplace who discovered and named it in 1812 in a generalised form for solving various problems.

Despite its applications, for more than 100 years, the degree of credibility of Bayesian analysis was rejected as vague and subjective and frequencies were accepted only by statisticians.

It was Jeffreys in 1939 ([1]) who rediscovered it and built the modern Bayesian theory in 1961. It was then that the two schools of statistics: Bayesian and Frequentists were distinctly different and set apart. By the 1980s it still remained limited to use due to the needs in the calculations.

Since 1990, it became practical thanks to the rapid developments of hardware and software. The Bayesian techniques, in this way, were applied in various fields of science such as economics, medicine, biology, engineering and so on.

A random variable has normal distribution with expectation θ and variance σ^2 when it distribution is given by formula (1):

$$p(y \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}}, \quad y \in R \qquad (1)$$

This distribution has several important features:

- It is symmetrical according to parameter θ and the mode, median and mean is θ.
- About 95% of the population lies within the range (-1.96σ, 1.96σ).
- The linear combination of random variables with normal densities is also a random variable with normal density. This means that if $X \sim N(\mu,r^2)$ and $Y \sim N(\theta,\sigma^2)$ then $aX+bY \sim N(a\mu+b\theta, a^2r^2+b^2\sigma^2)$.
- The most useful commands in language R for generating normal distribution are: `dnorm`, `rnorm`, `pnorm`, `qnorm`.

II. INFERENTIAL ANALYSIS FOR THE MEAN AND THE CONDITIONING WITH THE VARIANCE

Suppose that we have $Y_1, Y_2, \ldots, Y_n$ indipendent random variables identically distributed with normal distribution $N(\theta,\sigma^2)$. The sample distribution is given by the formula:

$$p(y_1,\ldots, y_n \mid \theta,\sigma^2) = \prod_{i=1}^{n} p(y_i \mid \theta,\sigma^2) =$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(y_i-\theta)^2}{\sigma^2}}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\theta}{\sigma}\right)^2\right\}$$

By splitting the quadratic form under the exponent, it can be seen that $p(y_1,\ldots,y_n \mid \theta,\sigma^2)$ depends on $y_1, y_2, \ldots, y_n$:

$$\sum_{i=1}^{n}\left(\frac{y_i-\theta}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i^2 - 2\frac{\theta}{\sigma^2}\sum_{i=1}^{n} y_i + n\frac{\theta^2}{\sigma^2}.$$

From this equation it can be shown that the two-dimensional statistics $(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2)$ is a sufficient statistic for the pair of parameters (θ, σ^2), from which it derives that the statistics ($\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$) is a sufficient two-dimensional statistic for (θ, σ^2).

The inferential analysis for this bi-parametric model can be divided into two separate parametric problems. According to Carlin ([4]), prior distributions can be built in different ways, mainly from a given value. Let' s first assume that we want to estimate θ when σ^2 is known and for θ will be used a conjugate prior distribution, considering that a prior distribution family is called conjugate if for a given sample the posterior distribution is in the same family of distribution ([10]). For each prior distribution $p(\theta|\sigma^2)$, the posterior distribution will satisfy the equation (2):

$$p(\theta \mid y_1,\ldots,y_n,\sigma^2) \propto p(\theta \mid \sigma^2)\times e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2} \quad (2)$$

$$\propto p(\theta \mid \sigma^2)\times e^{c_1(\theta-c_2)^2}$$

From the equation, we have $p(\theta|\sigma^2)$ to be conjugate then it must contain the quadratic term $e^{c_1(\theta-c_2)^2}$. The simplest family of probability distributions in R that fulfills this condition is the family of normal distributions, which means that if $p(\theta|\sigma^2)$ is a normal distribution and we consider the sample $y_1, y_2, \ldots, y_n$ from this distribution then $p(\theta \mid y_1,\ldots, y_n,\sigma^2)$ is also normal. Assuming that $\theta \sim N(\mu_0,\tau_0^2)$ then the equations are true:

$$p(\theta \mid y_1,\ldots,y_n,\sigma^2) = p(\theta \mid \sigma^2)p(y_1,\ldots,y_n \mid \theta,\sigma^2)/p(y_1,\ldots,y_n \mid \sigma^2)$$

$$\propto p(\theta \mid \sigma^2)p(y_1,\ldots,y_n \mid \theta,\sigma^2)$$

$$\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\times\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}$$

Adding the exponents and not considering -1/2, we have:

$$\frac{1}{\tau_0^2}(\theta^2 - 2\theta\mu_0 + \mu_0^2) + \frac{1}{\sigma^2}(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2)$$

$$= a\theta^2 - 2b\theta + c$$

where $a = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$, $b = \frac{\mu_0}{\tau_0^2} + \frac{\sum_{i=1}^{n} y_i}{\sigma^2}$ and $c = c(\mu_0, \tau_0^2, \sigma^2, y_1,\ldots, y_n)$.

Let we show that $p(\theta \mid y_1,\ldots, y_n,\sigma^2)$ has the form of a normal distribution:

$$p(\theta \mid y_1,...,y_n,\sigma^2) \propto \exp\{a\theta^2 - 2b\theta\}$$
$$= \exp\left\{-\frac{1}{2}a(\theta^2 - 2b\theta/a + b^2/a^2) + \frac{1}{2}b^2/a\right\}$$
$$\propto \exp\left\{-\frac{1}{2}a(\theta - b/a)^2\right\} = \exp\left\{-\frac{1}{2}\left(\frac{\theta - b/a}{1/\sqrt{a}}\right)^2\right\}$$

The function has exactly the same graphical shape with the normal distribution curve where $1/\sqrt{a}$ is playing the role of standard deviation and b/a is the expectation value. While a probability distribution is determined by the shape of its curve, then $p(\theta \mid y_1,..., y_n,\sigma^2)$ is a normal distribution. Marking with μ_n and τ_n^2 the posterior distribution parameters then:

$$\tau_n^2 = \frac{1}{a} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad \mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}.$$

A. The Combination of Information

Conditional probability distributions of parameters μ_n and τ_n^2 are obtained as a combination of parameters μ_0 and τ_0^2 with the sample elements. From the variance of the posterior distribution it results that $\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$ which means the inverse variance of the prior distribution is obtained from the inverse variance of the sample. The inverse variance is called accuracy of the model, so we have:

- $\tilde{\sigma}^2 = 1/\sigma^2$ = accuracy of the sample(it shows how near is y_i with parameter θ)
- $\tilde{\tau}_0^2 = 1/\tau_0^2 =$ accuracy of prior distribution
- $\tilde{\tau}_n^2 = 1/\tau_n^2 =$ accuracy of posterior distribution

It is reasonable to see the accuracy as additional information about the model:

$$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2 \Leftrightarrow$$

(posterior information= prior information+ sample information)

The mean of posterior distribution is given by the formula (3):

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y} \quad (3)$$

Thus, the mean of the posterior distribution is measured from mean of the prior distribution and sample mean. The weight of the sample mean is n/σ^2 which is also the accuracy of the sample mean, also the weight of the prior distribution $1/\tau_0^2$ serves as the accuracy of the prior distribution. If the mean of prior distribution is based on observations by the same population $Y_1, Y_2,\dots,Y_n$ then the variance of the mean of prior observations is $\tau_0^2=\sigma^2/k_0$. In this way the mean of the posterior distribution is written:

$$\mu_n = \frac{k_0}{k_0+n}\mu_0 + \frac{n}{k_0+n}\bar{y}.$$

B. The Prediction

We will consider the prediction of a new observation by a population after we have made the observations $(Y_1=y_1, Y_2=y_2, \dots, Y_n=y_n)$ and we must find the distribution for prediction. It is true that:

$$\{\tilde{Y} \mid \theta,\sigma^2\} \sim N(\theta,\sigma^2) \Leftrightarrow \tilde{Y} = \theta + \tilde{\varepsilon},$$
$$\{\tilde{\varepsilon} \mid \theta,\sigma^2\} \sim N(0,\sigma^2).$$

In other words, accepting that $\tilde{Y}$ has a normal distribution with expectation θ is the same thing as saying that it is given a sum of θ with a normal distributed noise which expectation is 0. Using this result, we can first calculate the mean of the posterior distribution and the variance of:

- $$E(\tilde{Y} \mid y_1, y_2,..., y_n,\sigma^2) = E(\theta + \tilde{\varepsilon} \mid y_1, y_2,..., y_n,\sigma^2)$$
  $$= E(\theta \mid y_1, y_2,..., y_n,\sigma^2) + E(\tilde{\varepsilon} \mid y_1, y_2,..., y_n,\sigma^2)$$
  $$= \mu_n + 0 = \mu_n$$

- $$D(\tilde{Y} \mid y_1, y_2,..., y_n,\sigma^2) = D(\theta + \tilde{\varepsilon} \mid y_1, y_2,..., y_n,\sigma^2)$$
  $$= D(\theta \mid y_1, y_2,..., y_n,\sigma^2) + D(\tilde{\varepsilon} \mid y_1, y_2,..., y_n,\sigma^2)$$
  $$= \tau_n^2 + \sigma^2$$

Since the sum of normal independent variables with normal distributions is also normal then $\tilde{Y} = \theta + \tilde{\varepsilon}$ has a normal distribution.

Thus, the predictive distribution is as in (4):

$$\{\tilde{Y} \mid y_1, y_2,..., y_n,\sigma^2\} \sim N(\mu_n, \tau_n^2 + \sigma^2) \quad (4)$$

C. Example

As an illustration, we will use simulated data with a small sample size from a normal distribution which mean is 1.8. This is the prior information to be used for the calculations of the parameters for prior and posterior distribution of mean and variance respectively. So, we have nine simulated values from the normal distribution *N (1.8, 0.015)*:

1.638164, 1.663346, 1.812662, 1.629400, 1.705748, 1.820818, 1.659060, 1.912620, 1.777257

The population mean is taken $\mu_0=1.8$ and for the variance we suppose that the greater part of probability lies between the double of standard deviation from the sample mean, that is μ_0-2 τ_0>0 or τ_0<1.8/2=0.90. The results are shown in Table 1 for each distribution of the population mean:

TABLE 1
RESULTS OF THE PARAMETERS

Parameters	Sample	Prior	Posterior
Mean	1.735	1.8	1.742
Variance	0.01	0.81	0.01

If $\sigma^2=s^2=0.01$ then $\{\theta \mid y_1, y_2,..., y_n, \sigma^2 = 0.01\} \sim N(1.742, 0.01)$.

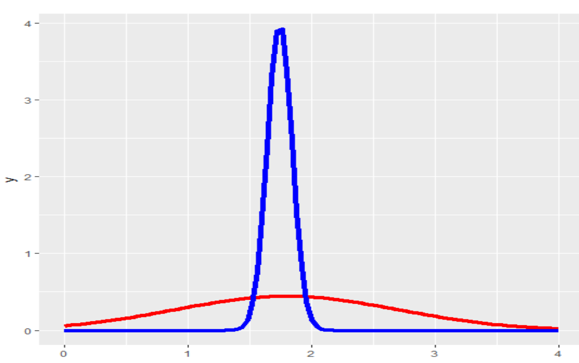

Fig. 1 The prior and the posterior distribution of the population mean.

In the Fig. 1 are shown with the red line the prior distribution and with the blue line the posterior distribution for population mean.

III. INFERENTIAL ANALYSIS OF THE UNKNOWN MEAN AND UNKNOWN VARIANCE

Bayesian inferential analysis for two or more parameters is not very different in the concept of the one with one parameter. For the joint prior distribution $p(\theta,\sigma^2)$ of the parameters θ and σ^2, the finding of the posterior distribution is related to the use of the Bayes rule where usually the conditional distribution can be substituted by the maximum likelihood function ([5]):

$$p(\theta,\sigma^2 \mid y_1, y_2,..., y_n) = \frac{p(y_1, y_2,..., y_n \mid \theta,\sigma^2)p(\theta,\sigma^2)}{p(y_1, y_2,..., y_n)}$$

The procedure begins by finding a family of conjugate prior distributions that makes easy the calculation of posterior distribution. Starting from the conditional probability formula, we get the multiplication formulas and so the joint distribution is written:

$$p(\theta,\sigma^2) = p(\theta \mid \sigma^2)p(\sigma^2)$$

We showed earlier that when σ^2 was known, a prior distribution for θ is the normal distribution (μ_0, τ_0^2). Consider the special occasion when $\tau_0^2 = \sigma^2/k_0$:

$$p(\theta,\sigma^2) = p(\theta \mid \sigma^2)p(\sigma^2)$$
$$= dnorm\ (\theta, \mu_0, \tau_0 = \sigma/\sqrt{k_0}) \times p(\sigma^2)$$

In this case, the parameters μ_0 and k_0 can be interpreted as the mean and the sample size of sample from a previous observation set. For σ^2 we need a prior distribution family to be positively defined in (0, ∞). Such a distribution family is the family of gamma distributions, but unfortunately this distribution family is not conjugate for the variance of a normal variable. However, the family of gamma distributions is conjugate to 1/ σ^2 (the accuracy of σ^2). When it is used such a prior distribution, it is said that σ^2 has a gamma inverse distribution:

accuracy=1/ σ^2~gama(a,b)

variance= σ^2~invers gama(a,b)

For interpretation, instead of parameters *a* and *b* the parameters in the prior distribution will be:

- $E(\sigma^2) = \sigma_0^2 \dfrac{\upsilon_0/2}{\upsilon_0/2-1}$
- D(σ^2) është zbritës në υ_0.
- $\text{mode}(\sigma^2) = \sigma_0^2 \dfrac{\upsilon_0/2}{\upsilon_0/2+1}$, that is why

  mode(σ^2)< σ_0^2<E(σ^2)

In this way, the parameters of the prior distribution (σ_0^2, υ_0) can be interpreted as the variance and the sample size of the prior observations.

D. Inferential Analysis for Posterior Distribution

Suppose we have $Y_1, Y_2, ..., Y_n$ the sample from normal variable $N(\theta, \sigma^2)$ and the prior distribution are: $1/\sigma^2 \sim gama(\frac{\upsilon_0}{2}, \frac{\upsilon_0}{2}\sigma^2)$,

$\theta | \sigma^2 \sim N(\mu_0, \sigma^2 / k_0)$.

As for the joint prior distribution we can write: $p(\theta, \sigma^2) = p(\theta | \sigma^2) p(\sigma^2)$ then even for the posterior distribution it can be done the same:

$$p(\theta, \sigma^2 | y_1, y_2, ..., y_n) =$$
$$= p(\theta | \sigma^2, y_1, y_2, ..., y_n) p(\sigma^2 | y_1, y_2, ..., y_n)$$

The conditional distribution of θ when the sample and σ^2 are given, by replacing $\tau_0^2 = \sigma^2 / k_0$ and $k_n = k_0 + n$ is:

$$\{\theta | y_1, y_2, ..., y_n, \sigma^2\} \sim N(\mu_n, \sigma^2 / k_n) \text{ where}$$
$$\mu_n = \frac{(k_0 / \sigma^2)\mu_0 + (n\sigma^2)\bar{y}}{k_0 / \sigma^2 + n / \sigma^2} = \frac{k_0 \mu_0 + n\bar{y}}{k_n}$$

From this conclusion, if μ_0 is the mean of k_0 prior observations then $E(\theta | y_1, y_2, ..., y_n, \sigma^2)$ is the mean of both k_0 prior observations and the actual sample. The variance $D(\theta | y_1, y_2, ..., y_n, \sigma^2)$ is the ratio of σ^2 to the total number of observations (previous and actual observations). The posterior distribution of σ^2 is taken by integrating from θ:

$$p(\sigma^2 | y_1, y_2, ..., y_n) \propto p(\sigma^2) p(y_1, y_2, ..., y_n | \sigma^2)$$
$$= p(\sigma^2) \int p(y_1, y_2, ..., y_n | \theta, \sigma^2) p(\theta | \sigma^2) d\theta$$

It is taken the result:

$$\{1/\sigma^2 | y_1, y_2, ..., y_n\} \sim gama(\upsilon_n / 2, \upsilon_n \sigma_n^2 / 2)$$
$$\text{where : } \upsilon_n = \upsilon_0 + n$$
$$\sigma_n^2 = \frac{1}{\upsilon_n}\left[\upsilon_0 \sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_n}(\bar{y} - \mu_0)^2\right]$$

This formula gives an interpretation of v_0 as the prior sample size from which is obtained σ_0^2. Since s^2 is the empirical variance of the sample then *(n-1)* s^2 gives the sum of square of the difference of the observations from the sample mean, so $v_0\ \sigma_0^2$ and $v_n\ \sigma_n^2$ are respectively the sum of square of prior and posterior. By multiplying the last equation with v_n it can be said that the sum of posterior squares is equal to the sum of prior squares with the sum of sample squares, while the third term is more difficult to be interpreted. If μ_0 is considered the mean of k_0 prior values with variance σ^2 then $\frac{k_0 n}{k_0 + n}(\bar{y} - \mu_0)^2$ serves as a point estimation for σ^2.

E. Monte Carlo Simulations

For most data analysis it is important to estimate the population mean θ, so it is important to calculate $E(\theta | y_1, y_2, ..., y_n)$ and other numerical characteristics. These ones are determined by the posterior distribution of θ given by the data. As it is known, the conditional distribution of θ provided the data and σ^2 is the normal distribution and the conditional distribution of σ^2 given the data is invers gamma. It can be used the Monte Carlo method to simulate samples of from the joint posterior distribution ([8], [9]), so the simulation of S pair of the parameters would be:

$$\sigma^{2(1)} \sim invers\ gama(\upsilon_n / 2, \sigma_n^2 \upsilon_n / 2),\ \theta^{(1)} \sim N(\mu_n, \sigma^{2(1)} / k_n)$$
$$\vdots \qquad\qquad \vdots$$
$$\sigma^{2(S)} \sim invers\ gama(\upsilon_n / 2, \sigma_n^2 \upsilon_n / 2),\ \theta^{(S)} \sim N(\mu_n, \sigma^{2(S)} / k_n)$$

This is accomplished in R language by using the commands:

```
s2_postsample=1/rgamma(10000,nun/2,s2n*nu
n/2)
teta_postsample=rnorm(10000,mun,sqrt(s2_p
ostsample/kn))
```

This procedure involves the simulation of 10000 pairs representing independent samples from the joint posterior distribution $p(\theta, \sigma^2 | y_1, y_2, ..., y_n)$. Moreover, the simulated values $\{\theta^{(1)}, ..., \theta^{(S)}\}$ represent independent samples from the marginal distribution $p(\theta | y_1, y_2, ..., y_n)$.

The results of MCMC simulation from the example in section II-C are illustrated by the graps:

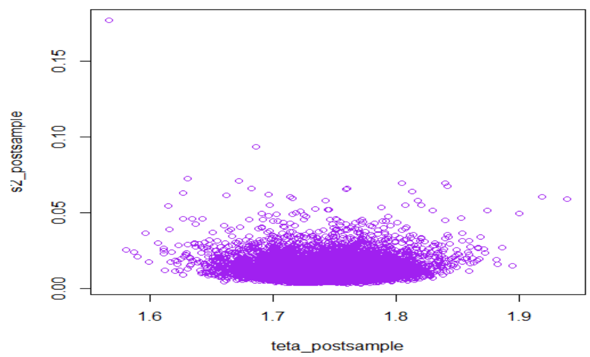

Fig. 2 The joint distribution after MCMC simulations

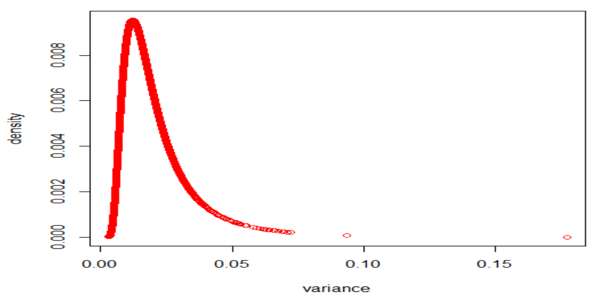

Fig. 3 The marginal distribution of 1/ σ^2 after simulations.

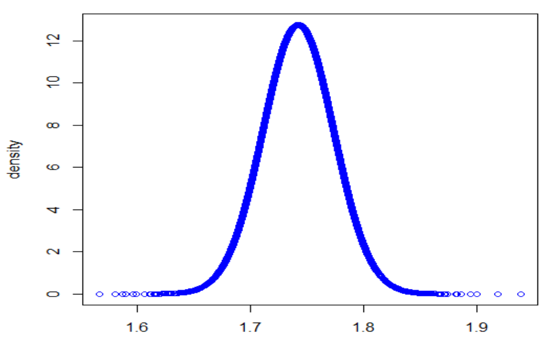

Fig. 4 The marginal distribution of θ after simulations.

A 95% confidence interval for the parameter θ is (1.72, 1.81).

F. Improper Prior

The problem involved is how Bayesian analysis can be used without prior information from the prior distribution. Many authors, from Lindley in 1973 ([6]) and then Kass in 1996([7]), were doubtful in using the improper priors that are not probability distributions, instead of prior distributions. As we refer to the parameters k_0 and v_0 as the prior sample size, it seems as small as these parameters are then the estimation will be more objective. This naturally induces to the thought of what happens to the posterior distribution when k_0 and v_0 are reduced considerably.

The formulas for are:

$$\mu_n = \frac{k_0\mu_0 + n\bar{y}}{k_0 + n}$$

$$\sigma_n^2 = \frac{1}{\upsilon_0 + n}\left[\upsilon_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_0 + n}(\bar{y} - \mu_0)^2\right]$$

When $k_0, \upsilon_0 \to 0$, then we have:

$$\mu_n \to \bar{y}$$

$$\sigma_n^2 \to \frac{n-1}{n}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

These results bring to the following posterior:

$$\{1/\sigma^2 \mid y_1, y_2, ..., y_n\} \sim gama(\frac{n}{2}, \frac{n}{2}\frac{1}{n}\sum(y_i - \bar{y})^2)$$

$$\{\theta \mid y_1, y_2, ..., y_n, \sigma^2\} \sim N(\bar{y}, \sigma^2/n).$$

Marking $\tilde{p}(\theta,\sigma^2) = 1/\sigma^2$ and considering that $p(\theta,\sigma^2 \mid y) \propto p(y \mid \theta,\sigma^2) \times \tilde{p}(\theta,\sigma^2)$, we get the same conditional distribution for θ but a gamma distribution for $1/\sigma^2$ ([11]). From the integration of the joint distribution from σ^2 it comes the result:

$\frac{\theta - \bar{y}}{s/\sqrt{n}} \mid \mathrm{y}_1, \mathrm{y}_2, \ldots \mathrm{y}_\mathrm{n} \sim \mathrm{S(n-1)}$, which means that after the sample is made we have that the unknown parameter is given by a student distribution with n-1 degree of freedom. Meanwhile the conditional distribution $\frac{\bar{Y} - \theta}{s/\sqrt{n}} \mid \theta$ is also with student distribution of n-1 degree of freedom. This means that before the sample is made, the difference of $\bar{Y}$ from the population mean θ is given by a student distribution of n-1 degree of freedom. The difference lies in the fact that before sampling the two parameters $\bar{Y}$ and θ are unknown, but after the sample is made then $\bar{Y} = \bar{y}$ is known and it provides information about the unknown parameter θ.

In this case we are not dealing with the prior probability distribution of (θ, σ^2) which then leads to a posterior distribution of θ that is student distribution with n-1 degree of freedom, so we are not in the case of proper Bayesian analysis. In the limit, theoretical results according to Stein ([2]) show that from a decision – making point of view, any suitable point estimator is a Bayesian estimator or it is the limit of a sequence of Bayesian estimators and each estimator is suitable ([3]).

IV. BIAS AND MEAN SQUARE ERROR OF THE ESTIMATORS

A point estimator of an unknown parameter θ is a function that reflects the data in a single parameter of the parameter space Θ. In the case where the sample is made from a normal distribution and we have the conjugate prior distribution previously considered, the posterior estimation of the mean θ is:

$$\hat{\theta}_{\mathbf{b}}(y_1, y_2, ..., y_n) = E(\theta \mid y_1, y_2, ..., y_n)$$

$$= \frac{n}{k_0+n}\bar{y} + \frac{k_0}{k_0+n}\mu_0 = w\bar{y} + (1-w)\mu_0$$

The elements of the sample for an estimator $\hat{\theta}_{\mathrm{b}}$ refer to its behaviour hypothetically based on repeated surveys or evidence. Let's compare the properties with the mean sample $\hat{\theta}_{\mathrm{e}}(y_1, y_2, ..., y_n) = \bar{y}$ when the exact value of the population mean θ_0 is known:

- $E(\hat{\theta}_{\mathrm{e}} \mid \theta = \theta_0) = \theta_0$, so $\hat{\theta}_{\mathrm{e}}$ is an unbiased estimator of θ_0.
- $E(\hat{\theta}_{\mathrm{b}} \mid \theta = \theta_0) = w\theta_0 + (1-w)\mu_0$, if $\mu_0 \neq \theta_0$, then $\hat{\theta}_{\mathrm{b}}$ is biased.

The bias shows how close is the centre of the sample distribution for a point estimator with the correct value of the parameter. Generally, an unbiased estimator is desired, however the bias does not indicate how far it is an estimator from the correct value. Consider y_1 an unbiased estimator of the population mean, this estimator is further from θ_0 than it is $\bar{y}$. To assess the proximity of an estimator with the correct value θ_0, we use the mean square error (MSE) and if $m = E(\hat{\theta} \mid \theta_0)$ then MSE is:

$$MSE(\hat{\theta}|\theta_0) = E\left[(\hat{\theta}-\theta_0)^2 \mid \theta_0\right] =$$
$$= E\left[(\hat{\theta}-m+m-\theta_0)^2 \mid \theta_0\right] =$$
$$= E\left[(\hat{\theta}-m)^2 \mid \theta_0\right] + 2E\left[(\hat{\theta}-m)(m-\theta_0) \mid \theta_0\right] + E\left[(m-\theta_0)^2 \mid \theta_0\right]$$

While $m = E(\hat{\theta} \mid \theta_0)$, we have $E(\hat{\theta} - m \mid \theta_0) = 0$ therefore the second term is zero, that is:

$$GMK(\hat{\theta} \mid \theta_0) = D(\hat{\theta} \mid \theta_0) + bias^2(\hat{\theta} \mid \theta_0)$$

This means that before the data is collected, the expected distance of an estimator from the correct value depends on the proximity of θ_0 with the distribution centre and by the variance of $\hat{\theta}$. Referring to the comparison of the two estimator $\hat{\theta}_{\mathrm{b}}$ with $\hat{\theta}_{\mathrm{e}}$, $bias(\hat{\theta}_e \mid \theta_0) = 0$ but $\hat{\theta}_{\mathrm{b}}$ has the smallest variability:

$$D(\hat{\theta}_e \mid \theta = \theta_0, \sigma^2) = \frac{\sigma^2}{n}$$

$$D(\hat{\theta}_b \mid \theta = \theta_0, \sigma^2) = w^2 \times \frac{\sigma^2}{n} < \frac{\sigma^2}{n}$$

The mean square errors for the two estimators are:

$$MSE(\hat{\theta}_e \mid \theta_0) = E\left[(\hat{\theta}_e - \theta_0)^2 \mid \theta_0\right] = D(\hat{\theta}_e \mid \theta_0) = \frac{\sigma^2}{n}$$

$$MSE(\hat{\theta}_b \mid \theta_0) = E\left[(\hat{\theta}_b - \theta_0)^2 \mid \theta_0\right]$$
$$= E\left[\{\ w(\bar{y} - \theta_0) + (1-w)(\mu_0 - \theta_0)\ \}^2 \mid \theta_0\right] =$$
$$= w^2 \times \frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \theta_0)^2$$

It is true that $MSE(\hat{\theta}_b \mid \theta_0) < MSE(\hat{\theta}_e \mid \theta_0)$ when

$$(\mu_0 - \theta_0)^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w} = \sigma^2\left(\frac{1}{n} + \frac{2}{k_0}\right).$$

If there are data on the population from which is made the sample, it is easy to find the values μ_0 and k_0 for which the inequality is true. In this case is built a Bayesian estimator with a square mean distance smaller than sample mean.

If we consider μ_0 =100 and σ_0^2=225, then:

$$MSE(\hat{\theta}_e \mid \theta_0) = D(\hat{\theta}_e \mid \theta_0 = 112) = \frac{\sigma^2}{n} = \frac{169}{n}$$

$$MSE(\hat{\theta}_b \mid \theta_0 = 112) = w^2\frac{169}{n} + (1-w)^2 144$$

In Fig. 5 are the graphs for the ratio of MSE for the two estimators for different sample sizes and k_0.

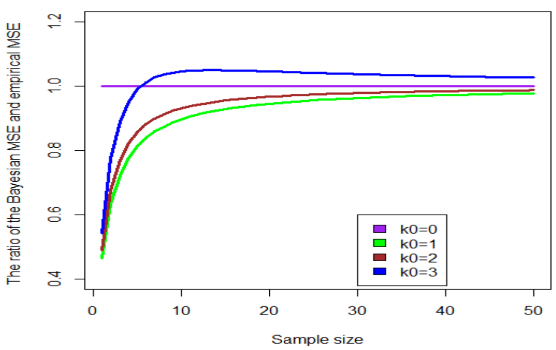

Fig. 5 Bayesian Estimator versus Empirical Estimator MSE

Fig. 5 shows that the MSE for the Bayesian estimator is smaller than the sample mean when k_0 =1, 2 and especially when the sample size is small. When k_0 =3, the MSE is greater for the Bayesian estimator but when the size n increases than it is seen that the bias goes to 0.

The Fig. 6 shows the graphs for different values of k_0 when $n=10$ (small sample size) and for the sample mean.

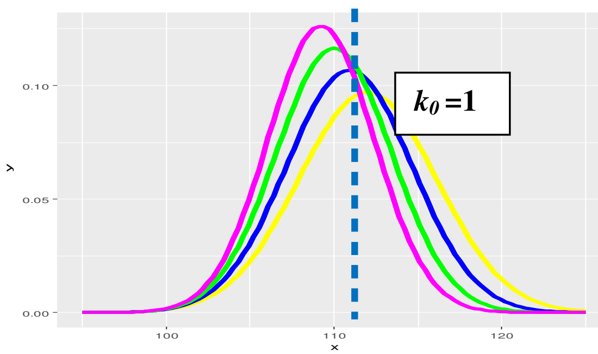

Fig. 6 The graphs of the distributions for the sample mean and the three Bayesian estimators (yellow, blue, green and magenta respectively).

This graph reinforces the fact that when k_0 =1 the Bayesian estimator's graph (the blue curve) is closer the real population mean θ_0=112 (intersected blue line) and its variance is small. This means that this estimator is closer to the true value of the parameter than the sample mean that is the empirical estimator.

CONCLUSIONS

The normal distribution is very important not only for its wide usage in different models, but even for the fact that the sample mean converges to a normal distribution by the Central Limit Theorem.

This probability distribution belongs to the family of exponential distributions where the mean and the empirical variance of the sample are sufficient statistics for its parameters.

The main benefit from the Bayesian inferential analysis is that it allows for small samples to make a better estimation usually starting from prior information. In this way, the normal distribution is determined by the estimators of its parameters and it can be further used in finding various probabilities we are interested in for different applications.

The main difference between Frequentist and Bayesian schemes is in the different ways of defining the probability. The Frequentist statistics (the classic statistics) treats the probability of events and does not quantify the inaccuracy of the true values of parameters. Instead, Bayesian statistics defines the probability distribution over possible values of a parameter that can be useful in different fields of interest.

REFERENCES

[1] Sir Harold Jeffreys, *Theory of Probability*, first published in 1939,Oxford Classic Texts in the Physical Sciences, 2000.

[2] Charles Stein, A *Necessary and Sufficient Condition for Admissibility*, The Annals of Mathematical Statistics, Volume 26, number 3, 1955.

[3] James Berger, *A Robust Generalised Bayes Estimator and Confidence Region for a Multivariate Normal Mean*, The Annals of Statistics, Volume 8, No4, July 1980, p 716-761.

[4] Bradley P. Carlin, Thomas A. Louis, *Bayesian Methods for Data Analysis* (Third ed.). Chapman & Hall/ CRC, Press ISBN 9781584886983, pp27–41, 2008.

[5] Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer. pp. 21–24. ISBN 978-0-387-31073-2, 2006

[6] Dennis.V. Lindley, O. Barndorff-Nielsen, Gustav Elfving, Erik Harsaae, Daniel Thorburn, Anders Hald and Emil Spjötvoll, *The Bayesian Approach*. Scandinavian Journal of Statistic, vol 5, no. 1, pp 1-26, 1978.

[7] Robert E. Kass, Larry Wasserman, *The selection of Prios distribution by formal rules*, Journal of American Statistical Association, Vol 1, No 435, pp 1343-1370, September 1996.

[8] Christian P. Robert and George Casella, *Monte Carlo Statistical Methods*.Springer-Verlag, New York, second edition, 2004.

[9] Christian P. Robert and George Casella, *Introducing Monte Carlo Methods with R*.Use R! Springer-Verlag, New York, 2009.

[10] Christian P. Robert, *The Bayesian Choice. Springer-Verlag, New York,*paperback edition, 2007.

[11] Donald B Rubin, Andrew Gelman, John B. Carlin, Hal Stern, *Bayesian Data Analysis (2nd ed.). Boca Raton: Chapman & Hall/CRC. ISBN 1-58488-388-X. MR 2027492,2003*